American Journal of Biological Anthropology

WILEY

YEARBOOK OF BIOLOGICAL ANTHROPOLOGY

YEARBOOK ARTICLE OPEN ACCESS

Computational Genomics and Its Applications to Anthropological Questions

Kelsey E. Witt[1] | Fernando A. Villanea[2]

[1]Department of Genetics and Biochemistry and Center for Human Genetics, Clemson University, Clemson, South Carolina, USA | [2]Department of Anthropology, University of Colorado Boulder, Boulder, Colorado, USA

Correspondence: Kelsey E. Witt (kwittdi@clemson.edu)

Received: 24 June 2024 | **Revised:** 14 October 2024 | **Accepted:** 19 December 2024

Keywords: computational genomics | genetic anthropology | population genetics

ABSTRACT

The advent of affordable genome sequencing and the development of new computational tools have established a new era of genomic knowledge. Sequenced human genomes number in the tens of thousands, including thousands of ancient human genomes. The abundance of data has been met with new analysis tools that can be used to understand populations' demographic and evolutionary histories. Thus, a variety of computational methods now exist that can be leveraged to answer anthropological questions. This includes novel likelihood and Bayesian methods, machine learning techniques, and a vast array of population simulators. These computational tools provide powerful insights gained from genomic datasets, although they are generally inaccessible to those with less computational experience. Here, we outline the theoretical workings behind computational genomics methods, limitations and other considerations when applying these computational methods, and examples of how computational methods have already been applied to anthropological questions. We hope this review will empower other anthropologists to utilize these powerful tools in their own research.

1 | Introduction

Genomics has undergone a recent renaissance, with the cost of sequencing a human genome dropping below $1000 for the first time in 2019 (Mullin 2022). This has in turn transformed the way we analyze genomic data. Early genetic analyses focused on small sections of single genes for a small number of individuals because sequencing was an arduous, expensive project. The development of massively parallel sequencing technology (also known as next-generation sequencing or NGS) dramatically improved our capability to sequence large genomic regions by allowing for the simultaneous sequencing of millions of DNA fragments. Continual innovation in NGS processes has resulted in the capability of generating more genomic data at a lower cost. For example, the first commercially-available NGS platform, the Roche 454, was released in 2005 and it generated over 20 million sequenced DNA bases per run. One of the newest NGS platforms, the Illumina Novaseq X, can now generate as many as 16 trillion sequenced DNA bases per run (Illumina 2024). This technological innovation has vastly expanded our ability to sequence genomes. As a result, the amount of public genomic data is larger than ever. Tens of thousands of human genomes have been sequenced, including genome sequencing projects from populations that are historically underrepresented in genomic datasets including the Mexican Biobank (Sohail et al. 2023), the GenomeAsia 100K Project (GenomeAsia 100K Consortium 2019), and an Indian genomic resource, IndiGenomes (Jain et al. 2021). Ancient genome sequencing has also become more affordable and effective, which allows for a more detailed examination of historical populations.

Both authors contributed equally to this work.

In addition, more powerful statistical tools have expanded our options of what questions can be answered by analyzing genomes. An individual's genome carries the instructions responsible for that person's phenotypic variation, but it also carries evidence of the genetic inheritance process itself, which reflects the evolution of their ancestors. There are many avenues to use a genome to explore the evolutionary history of past populations. For example, ancient and modern genomes can be interrogated to examine the frequency trajectory of a gene (Zhang et al. 2021). A single genome carries a record of its ancestral population's expansions and contractions over time (Li and Durbin 2011). Other methods utilize the history of recombination in an individual's genome to study past gene flow events (Hubisz et al. 2020; Kelleher et al. 2019; Speidel et al. 2019). Numerous statistical methods are available to tap into that long demographic and evolutionary history and use individual human genomes to understand more about human evolution on a broad scale.

The difficulty of understanding past events by examining a single human genome lies in disentangling all of these separate forces of evolution, many of which can result in similar changes to genetic variation. If we examine a single evolutionary effect, such as selection or a genetic bottleneck, it is easy to understand and predict what the effect on a human genome would be. However, many of these forces affect the amount and distribution of genetic diversity in similar ways, at least if viewed from a limited portion of the genome. For example, both natural selection in favor of a particular allele and random loss of alleles as a result of a large population bottleneck can result in a genomic region with low genetic diversity, even if natural selection is operating locally and the population bottleneck is affecting the entire genome. Computational tools can help to clarify the population history of these individuals, but these tools use complex statistical methods, which can lead to users running them as a "black box," where results are presented with minimal interpretation about the analytical process. For an uninformed user, it is often difficult to know the limitations or biases of these tools, whether the program has run as expected, and in some cases how to interpret the results. We hope this review can help contextualize the basic theory on which these methods operate.

These computational tools have a lot of exciting applications to anthropology and expanding our understanding of human history and evolution. Historically, human evolution was a topic primarily studied through fossil morphology. Our ability to infer evolutionary history from human genomes empowers us to examine human evolution from a parallel source of information that can complement the morphological work produced by paleoanthropologists. Ancient DNA data, a limited and valuable resource, can complement bioarcheology, as genome data can be used to look more holistically at historical populations, even from a few sequenced genomes. All of these computational tools can also be applied to the study of nonhuman primates. Finally, our understanding of past demographic events and human evolution has health-relevant implications for living individuals today. Multiple studies have linked genetic variants that underwent natural selection in an ancient environment to an increased risk of disease in modern contexts (Sohail 2022; Klunk et al. 2022), and some genetic variants inherited from archaic humans can also impact health-related phenotypes (Koller et al. 2022).

Although many of these new computational tools can be applied to answer anthropological questions, implementation in the anthropological literature has been limited so far. Here, we summarize some of the many tools that have been recently developed to analyze the complexities of sequenced genomes. We provide information about the theoretical underpinnings of the methods, give guidance on how to apply the tools successfully, and share examples of how these tools have been used to answer anthropological questions. We hope that our summary of these computational methods will enable biological anthropologists to incorporate them into their own work and expand the application of these tools to better understand the evolutionary history of our species.

1.1 | Genomic Data Analysis as a High Dimensional Computational Problem

The defining strength of population genetics as a discipline is the precise mathematical modeling of the natural forces that shape genome variation. As these models capture known natural phenomena, they allow for the generation of expectations that can be matched to observed genome data in a unified scientific framework. The difficulty for population geneticists lies in untangling which natural force contributes most to a population of interest, or a region of the genome of interest, during a given time in their evolution.

The base model for genome evolution, the Wright-Fisher model of genetic drift, elegantly describes the random fluctuations of allele frequencies in an idealized population: a population of individuals which does not experience migration, mutation, or selection (Wright 1931; Fisher 1999). In terms of computation, even on paper, this model is very efficient, and can track the change in allele frequencies in large populations over long periods of time. One hundred years after its conception, it is still the basis of the most powerful simulation methods available, including *msprime* and *SLiM* (Messer 2013; Kelleher et al. 2016). Yet, as per Sewall Wright's own admissions from a letter to Ronald Fisher in 1929 (Wright and Fisher 1929), no natural species meets the criteria of an idealized population. Thus, much of this body of work is dedicated to capturing those three other forces: migration, mutation, and selection.

The take-home message from the classical theory of population genetics theory is that variation observed in all genomes is the result of at least four different evolutionary forces: genetic drift, natural selection, migration, and mutation, all acting in different directions at all times. While each one of these processes is elegantly modeled, and models that combine any two forces are simple to conceptualize, the four-way tug-of-war that natural populations continually experience creates some truly chaotic genetic variation (Figure 1). The crux of population genetics is to isolate signals of one natural force at times when it is predominant, over the noise created by the other forces when they are not biologically interesting.

The final piece of the genome evolution puzzle is recombination, which is H.B.S. Haldane's contribution to the new synthesis (Haldane 1919). Recombination ensures that variation between genome positions remains independent given enough physical

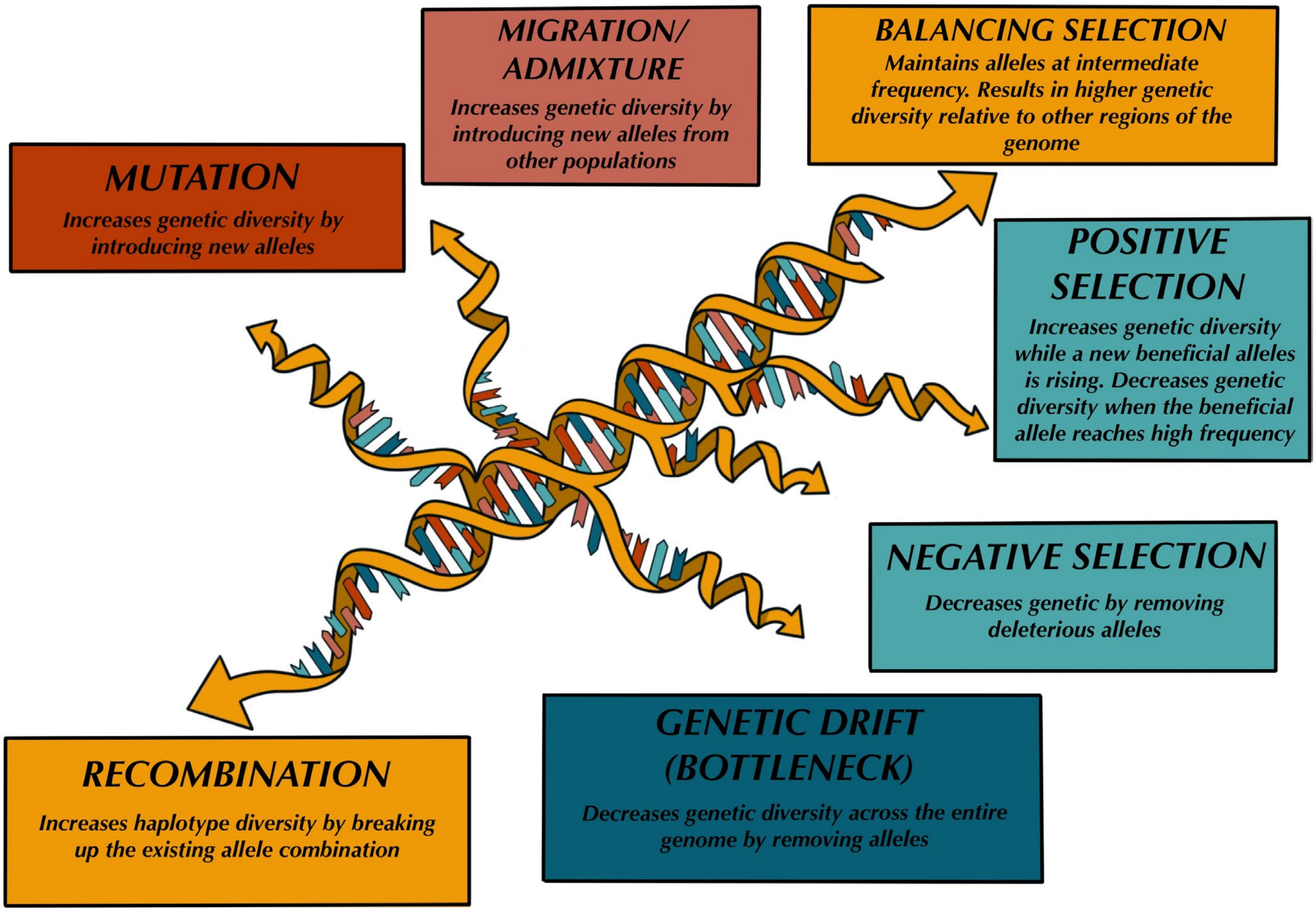
MIGRATION/ ADMIXTURE
Increases genetic diversity by introducing new alleles from other populations
BALANCING SELECTION
Maintains alleles at intermediate frequency. Results in higher genetic diversity relative to other regions of the genome
MUTATION
Increases genetic diversity by introducing new alleles
POSITIVE SELECTION
Increases genetic diversity while a new beneficial alleles is rising. Decreases genetic diversity when the beneficial allele reaches high frequency
NEGATIVE SELECTION
Decreases genetic by removing deleterious alleles
RECOMBINATION
Increases haplotype diversity by breaking up the existing allele combination
GENETIC DRIFT (BOTTLENECK)
Decreases genetic diversity across the entire genome by removing alleles

FIGURE 1 | A visualization of the various forces acting on genetic diversity and whether they contribute to increased or decreased genetic diversity.

distance in a chromosome, or at least that independence can be modeled under a probabilistic model. Because of the likely independence between genome positions sufficiently distant or on different chromosomes, understanding population evolution from genome data turns into a truly monstrous problem. Even the humble genome alignment, a 2-dimensional data object, harbors a record of variation from multiple layers of generative processes, where every variable position may behave independently (or not) given some probabilistic function. Thus, genome data is a high-dimensional computational problem, given its multitude of generative forces that shape variation, even when a genome alignment as a data object is simple.

1.2 | Types of Genetic Data

Publicly accessible genetic and genomic data exist in a multitude of formats. We will briefly discuss the different types of genomic data that is typically available, as this determines which of these programs are applicable to answer particular questions. We also compiled a brief list of publicly accessible genome data sources in Table 1 and a discussion of common data types is summarized in Anderson (2024).

Much of the genetic data publicly available is still fairly short in length and was generated as single-locus targeted sequencing. These sequences are often generated through Sanger sequencing, which can generate hundreds to a few thousand base pairs of continuous sequence data. A variety of genomic regions of interest have been sequenced, but mitochondrial DNA (mtDNA) is the most common, particularly for non-human primates. mtDNA is haploid, it does not recombine, and it is passed only on maternal lines, all contributing to a lower information content than the autosomes. However, portions of the mitochondrial genome mutate very quickly, allowing for the accumulation of variation in shorter historical time spans. These hypervariable regions are the most common form of mtDNA available, as further sequencing of the mitogenome outside these often offers limited additional information due to linkage. Some regions of the mitochondrial genome are also targeted for DNA barcoding, like the cytochrome oxidase subunit 1 gene (COI). Complete mitogenomes are also relatively short sequences—the human reference mitogenome only encompasses ~16,569 base pairs. Short published sequences of the autosomal genome are typically treated as non-recombinant in most analyses due to the low probability of recombination at that scale.

Whole autosomal genome data maximizes genetic information, particularly when structural variation is incorporated in the genome assembly. However, whole genome data is often too rich for computational analyses, as the human reference genome is ~3.1 billion base pairs long. Whole genome data is expensive to sequence, computationally complex to assemble and to implement

TABLE 1 | A brief list of key genetic and genomic data sources.

Database	Relevant publication	Data type	Web page
GenBank (NIH genetic sequence database)	Benson et al. 2012	Various sequences including most single loci and mt data	https://www.ncbi.nlm.nih.gov/genbank/
Thousand Genomes Project	1000 Genomes Project Consortium 2015	Whole genomes aligned to hg19 (GRCh37) reference	http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502
Simons Genome Diversity Panel	Mallick et al. 2016; Watkins et al. 2020	Whole genomes aligned to hg19 (GRCh37) reference	https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data2021
Human Genome Diversity Project	Bergström et al. 2020	Whole genomes aligned to hg38 (GRCh38) reference	https://www.internationalgenome.org/data-portal/data-collection/hgdp
Human Pangenome Reference Consortium	Liao et al. 2023	Whole genomes including structural variants	https://projects.ensembl.org/hprc/
Max Planck Institute Neanderthal and Denisovan Genomes	Various publications	Whole genomes aligned to hg19 (GRCh37) reference	https://www.eva.mpg.de/genetics/genome-projects
Estonia Biocenter Genome Diversity Panel	Pagani et al. 2016	Whole genomes aligned to hg19 (GRCh37) reference	https://evolbio.ut.ee/CGgenomes.html
Allen Ancient DNA Resource	Mallick et al. 2024	SNP array data aligned to hg19 (GRCh37) reference	https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data
GenomeAsia	GenomeAsia Consortium 2019	Whole genomes aligned to hg19 (GRCh37) reference	https://browser.genomeasia100k.org/
LASI-DAD (Diagnostic Assessment of Dementia for the Longitudinal Aging Study in India)	Lee et al. 2019; Kerdoncuff et al. 2024	SNP array data and whole genome imputation with 1000 genomes and TOPMED data, aligned to hg19 (GRCh37) reference	https://dss.niagads.org/datasets/ng00106/

quality controls, and prohibitively expensive to store for large numbers of individuals. Autosomal data is most often captured using techniques that only sequence small variable portions of the genome, as only an estimated ~88 million base pairs in the human genome are variable in living individuals (1000 Genome Project Consortium 2015). Single Nucleotide Polymorphism (SNP) genotyping arrays focus on known variable single base pairs of the genome, often only capturing >1 million variable positions. Data from SNP genotyping arrays is inexpensive to produce, and thus is the most common source of autosomal data; however, it cannot capture novel variation, and thus is best applied to well-characterized populations, which limits its applicability for non-human primates. Finally, RADSeq data and other random shotgun sequencing methods aim to find a compromise for species for which a reference genome has not been completed, or SNP variation is not well characterized (Catchen et al. 2017). RADSeq data, however, does not represent genetic variation evenly across the genome, which limits its use for some analyses and is in some cases only comparable among individuals within a species or even within a single population (Lowry et al. 2017).

1.3 | A Note on Ancient DNA Data

Without doubt, the ability to sequence long-dead individuals revolutionized the biological sciences, with human ancient DNA leading the field due in part to large public interest in questions of human origins and human prehistory. The chemical nature and qualities of ancient molecules are better described in other reviews, such as Slatkin and Racimo (2016) and Orlando et al. (2021), but for the purposes of computational tools, aDNA data can exist as any of the data types described above. However, two qualities of ancient molecules do affect the usefulness of ancient DNA datasets: their rarity and the accumulation of DNA damage that masquerades as true DNA variation. DNA decays very quickly after an organism's death, and the remaining ancient DNA is but a very small portion of a cell's genome (Allentoft et al. 2012). There is little ancient DNA in archaeological remains to start, but it is also fragmented, with sequences of 100 base pairs, and shorter, being typical (Prüfer et al. 2010). In addition, the chemical structure of nucleotides is altered over time by ionizing radiation, oxygen radicals, and hydrolytic damage. A particular problem is the deamination of cytosine into uracil, which can be replaced by thymine during sequencing creating C→T substitutions that can be mistaken for true novel variation. Both laboratory-based techniques and bioinformatic techniques have evolved to minimize these problems, but the resulting genetic data is both less abundant relative to modern genomes and less trustworthy in the case of novel variation. This limits the applicability of some of the methods that we review here: low abundance results in too small sample sizes, sequence lengths might be too short to be informative, or sequence coverage might be too low to resolve heterozygosity. Low sequence coverage, in turn, results in loss of true variation during filtering steps, reduced statistical confidence to call genotypes, and makes downstream computational steps less reliable, such as calling haplotypes by phasing. Yet, if these difficulties can be overcome, ancient DNA is an extremely powerful asset for computational analyses, in particular paired with trustworthy calibration dates of the ancient molecules or associated archaeological evidence. Some of the methods we will discuss here can very naturally combine modern data with ancient data by converting radiocarbon dates and other dating information into an approximation of biological generations.

1.4 | Summary Statistics Reduce Data Dimensionality, but at a Cost

As we have established, genomic data should be treated as high-dimensional data. Historically, few methods use raw genome data as input, as the computational requirements quickly become prohibitive. Instead, most forms of analysis rely on collapsing genomic data into simpler data forms, known as summary statistics (Csilléry et al. 2010). Summary statistics rely on a priori understanding of the natural forces and their effects on genome variation, with the goal of capturing the effects of a particular force (Miró-Herrans and Mulligan 2013). For example, the site frequency spectrum (SFS) collapses all nucleotide variation for a region of the genome—or the entire genome—into a distribution of the allele frequencies, represented as a histogram. As this summary is in itself a collection of many—sometimes hundreds—of values, further summarization is often necessary, and indeed most simpler summary statistics further collapse the SFS: the fixation index (F_{ST}) was designed to quantify population differentiation and structure, the proportion of segregating sites (S) captures mutation rates, and nucleotide diversity (Π) quantifies genetic variation. Composite statistics, such as Tajima's *D* (Tajima 1989), incorporate information from S and Π to test for deviations from the standard neutral coalescent model, driven by demographic history, adaptation, or other non-natural processes.

It is important to remark that the use of summary statistics has an important drawback: all information that is not captured by a particular summary statistic is lost, and no summary statistic perfectly captures the natural forces that shape the genome. This is called *sufficiency*, whether a summary statistic can capture the same information as the full data for the purposes of calculating a parameter of interest (Csilléry et al. 2010). Summary statistics are often handpicked a priori—that is, based on general principles of genome evolution, as opposed to emerging directly from the dataset under observation. This runs the risk of further losing information when there are aspects of particular evolutionary processes that are not captured by a particular summary statistic, which may greatly reduce our ability to understand empirical genomic data. However, as computational power has greatly progressed in the last decade, some genome analyses have been optimized to use complete genomic data, adopt more complex statistics like the SFS as the default summary statistic, or at a minimum, integrate multiple summary statistics.

2 | Applications to Anthropological Genetics

In this review, we discuss a series of programs that apply population genetics theory to solve various biological problems in anthropological genetics, which we also summarize in Table 2 and Figure 2. Further resources for the theory underlying the methods outlined here are available in Table S1.

2.1 | Reconstructions of Human Demography

Anthropology is a historical science, interested in learning about past population events that were meaningful enough to be recorded in oral histories, preserved in archaeological artifacts, or left as visible biological marks in past and present-day populations. As such, using genetic information to gain insight into past population events is perhaps the most logical application of computational tools to anthropology—most explorations of past human demographic events can be motivated, or corroborated, through non-genetic sources of information. Given this motivation, the premier theoretical framework for exploring demographic transitions is the coalescent. Starting from a handful of individuals, with as little as a few hundred base pairs of mitochondrial DNA, the coalescent can reconstruct how these individuals connect back in time through a genealogy, and the deviation of the genealogy from that of an idealized population can predict the timing and magnitude of demographic changes that would be responsible for the skew.

It is important to note here that this approach relies on a specific definition of an idealized population, one which is not affected by genetic drift, migration, or natural selection and one in which mutation and recombination behave with predictable rates. Assuming these conditions apply to a human population under study, we have to further assume that migration is negligible through the time period of interest, and that the region of the genome used is negligibly affected by natural selection. Given these conditions, any deviation from an idealized genealogy must be the result of genetic drift, and as we defined previously, the magnitude of genetic drift is inversely proportional to the effective size of the population. When a population expands or shrinks, the genealogy will deviate in a direction and magnitude corresponding to the effects of genetic drift. While these sets of conditions appear very specific and thus perhaps preclude analyses of most human populations, it is possible to use non-genetic data such as oral histories and the archaeological record to understand whether considerable migration is expected, or if the effects of migration can be ignored. Likewise, we must assume demographic inference will be unaffected by selection by using regions of the genome that are known to be non-functional, or by simply using a large enough portion of the genome and assuming that selection is both rare and cannot act on large portions of the genome all at once.

TABLE 2 | A brief list of new and popular computational tools for the methods discussed in this review. The reference where the method was introduced is listed, as well as the github or resource page where available. Note that this is by no means an exhaustive list, nor does it necessarily represent the "best" tools in each category for a particular project—it is just meant as a starting point for those interested in trying these methods.

Approach	Tool name	Reference	Github/Resource page
Demographic Reconstruction/ Simulation	BEAST2	Bouckaert et al. 2014	https://www.beast2.org/
	fastsimcoal2	Excoffier et al. 2021	http://cmpg.unibe.ch/software/fastsimcoal2/
	SMC++	Terhorst et al. 2017	https://github.com/popgenmethods/smcpp
	CHIMP	Upadhya and Steinrücken 2022	https://github.com/steinrue/chimp
	diCal2	Steinrücken et al. 2019	https://github.com/popgenmethods/diCal2
	ARGweaver-*D*	Hubisz et al. 2020	http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html
	TSinfer	Kelleher et al. 2019	https://github.com/tskit-dev/tsinfer
	Relate	Speidel et al. 2019	https://myersgroup.github.io/relate/
	Msprime	Baumdicker et al. 2022	https://tskit.dev/msprime/docs/stable/intro.html
	SLiM	Haller and Messer 2023	https://messerlab.org/slim/
Test for selection in phylogenies	HyPhy	Kosakovsky Pond et al. 2020	https://stevenweaver.github.io/hyphy-site/methods/selection-methods/
Balancing Selection	BetaScan	Siewert and Voight 2020	https://github.com/ksiewert/BetaScan
Positive Selection	SweepFinder2	DeGiorgio et al. 2016	https://degiorgiogroup.fau.edu/sf2.html
	diploS/HIC	Kern and Schrider 2018	https://github.com/kern-lab/
	ImaGene	Torada et al. 2019	https://github.com/mfumagalli/ImaGene
	rehh	Klassmann and Gautier 2022	https://cran.r-project.org/web/packages/rehh/index.html
	Selscan 2.0	Szpiech 2021	https://github.com/szpiech/selscan
	Disc-PG-GAN	Riley et al. 2024	https://github.com/mathiesonlab/disc-pg-gan.
	Allele frequency time series	Schraiber et al. 2016	https://github.com/Schraiber/selection
	AGES	Akbari et al. 2024	https://reich-ages.rc.hms.harvard.edu/#/
Quantifying population admixture	*D*+	Lopez Fang et al. 2024	https://github.com/LeslyLopezFang/Dplus
	ADMIXTOOLS	Patterson et al. 2012	https://github.com/DReichLab/AdmixTools
	ADMIXTOOLS 2	Maier et al. 2023	https://uqrmaie1.github.io/admixtools/
	Treemix	Pickrell and Pritchard 2012	https://bitbucket.org/nygcresearch/treemix/wiki/Home
Visualizing Admixture	ADMIXTURE	Alexander et al. 2009	https://dalexander.github.io/admixture/
	fastSTRUCTURE	Raj et al. 2014	https://rajanil.github.io/fastStructure/
	ChromoPainter/ fineSTRUCTURE	Lawson et al. 2012	https://people.maths.bris.ac.uk/~madjl/finestructure/

(Continues)

TABLE 2 | (Continued)

Approach	Tool name	Reference	Github/Resource page
Timing admixture	ROLLOFF	Moorjani et al. 2011	https://github.com/priyamoorjani/rolloff
	ALDER	Loh et al. 2013	https://cb.csail.mit.edu/alder/
	GLOBETROTTER	Hellenthal et al. 2014	https://people.maths.bris.ac.uk/~madjl/finestructure/globetrotter.html
Local Ancestry Inference	RFMix	Maples et al. 2013	https://github.com/slowkoni/rfmix
	FLARE	Browning et al. 2023	https://github.com/browning-lab/flare
	Loter	Dias-Alves et al. 2018	https://github.com/bcm-uga/Loter
	Recomb-Mix	Wei et al. 2023	https://github.com/ucfcbb/Recomb-Mix
Archaic Introgression Detection	SPrime	Browning et al. 2018	https://github.com/browning-lab/sprime
	HMMix	Skov et al. 2018	https://github.com/LauritsSkov/Introgression-detection/
	ArchIE	Durvasula and Sankararaman 2019	https://github.com/sriramlab/ArchIE
	Admixfrog	Peter 2020	https://github.com/BenjaminPeter/admixfrog
Migration rates	disperseNN	Smith et al. 2023	https://github.com/kr-colab/disperseNN
	Locator	Battey et al. 2020	www.github.com/kern-lab/locator
Recombination rates	ReLerNN	Adrion et al. 2020	https://github.com/kern-lab/ReLERNN

2.2 | Coalescent-Based Reconstructions of Demographic Trajectory

A valuable aspect of the coalescent for the purposes of anthropological genetics is that we only need to consider the direct ancestors of the individuals for whom we have genomic data, and as genomic data can be a limiting resource, coalescent-based methods are extremely practical (for an expanded summary see Villanea et al. 2020). In terms of computational resources, coalescent-based methods are thrifty for the same reason: a simulation going forward-in-time needs to account for every individual in every lineage, even those whose genetic material will be lost in future generations; a coalescent simulation going backward-in-time is not concerned with lost lineages. Even considering this advantage, coalescent inference is still severely restricted by computational resources. Early coalescent demographic reconstructions by software such as BEAST only became available after 2007 (Drummond and Rambaut) and were limited to strictly linked sites such as mitochondrial sequences. Reconstructions for recombinant sites only became possible with the introduction of Hidden Markov algorithms (Mather et al. 2020), and even then, early tools were limited to comparing a single pair of haploid genomes, usually from the same individual (Li and Durbin 2011), with later tools expanding the method to a handful of genomes (Schiffels and Durbin 2014), and eventually hundreds of genomes (Terhorst et al. 2017). Despite these limitations, coalescent demographic reconstructions are extremely powerful and have yielded valuable insight into human and non-human populations in deep time.

The earliest type of methods that took advantage of these theoretical underpinnings to approximate demographic transitions was the *SIMCOAL* family of methods; starting with *SIMCOAL* (Excoffier et al. 2000), which allows for the simulation of single linked loci (such as mitochondrial DNA or microsatellites). Later, *SIMCOAL 2.0* (Laval and Excoffier 2004) introduced the ability to simulate recombination between loci; *serialSIMCOAL* (Anderson et al. 2005) allowed for sampling at any point in time, which permitted comparisons with ancient DNA samples; *fastSIMCOAL* (Excoffier and Foll 2011) retains the serial sampling capabilities of serialSIMCOAL while also implementing a faster coalescent framework. The newest approach, *fastSIMCOAL2* (Excoffier et al. 2021), infers demographic parameters using SFS and allows for the simulations of long portions of the genome comparable to genomes produced through NGS technology. *fastSIMCOAL2* users can program "historical events" into the simulation at any generation, including population fusion, fission, or other migration patterns, sudden changes in population size, or changes in growth rate over long periods of time. *fastSIMCOAL2* takes these inputs and creates a simulated genome for all individuals along their tree.

The second family of methods to popularize the coalescent as the premier theoretical platform to understand demographic transitions is the *BEAST* family of methods. Drummond et al. (2005) had introduced the Bayesian Skyline Plot to great effect, showing how the method can take genetic data and use it to generate a posterior distribution of population sizes and how they change through time. This methodology was popularized by *BEAST* (Drummond and Rambaut 2007), which allows for the reconstruction of population trajectories backward in time from linked genetic data, primarily mitochondrial genetic data. *BEAST2* (Bouckaert et al. 2014) extends *BEAST* to be easier to integrate with other software

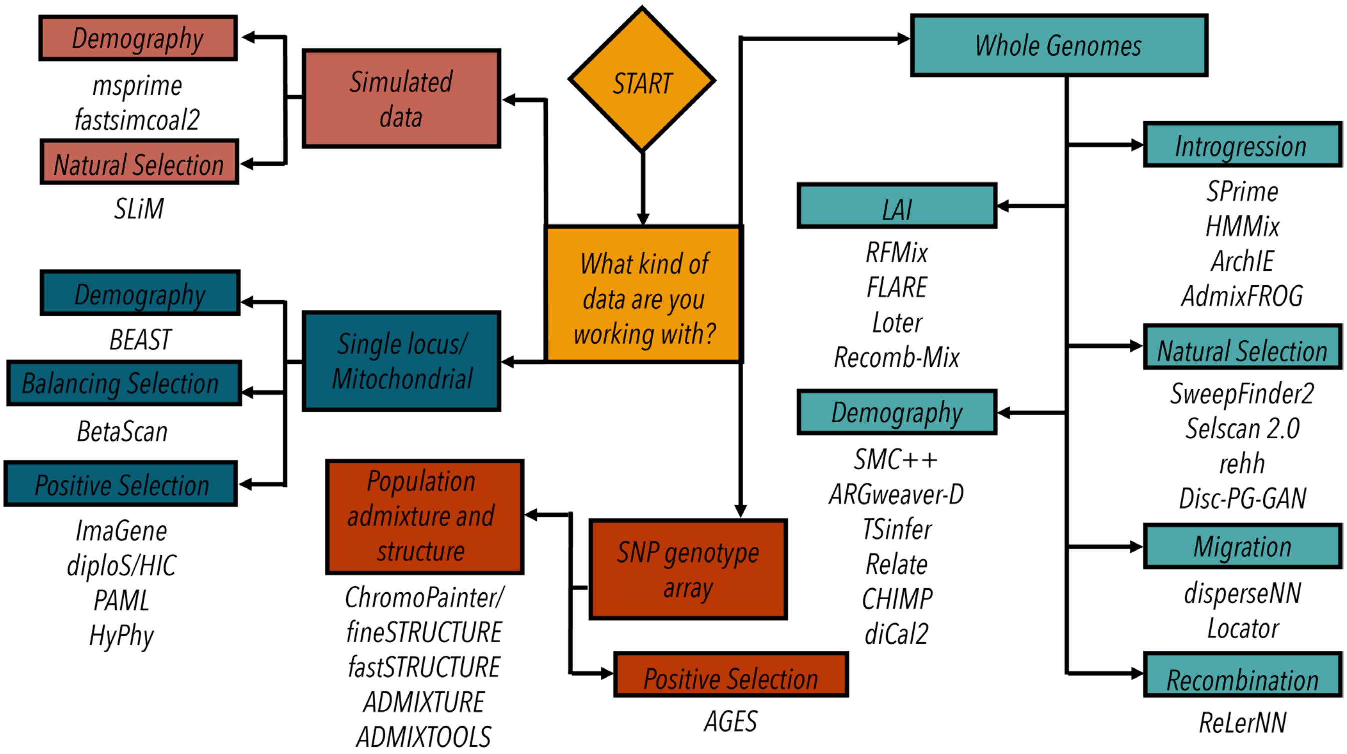

FIGURE 2 | A flow-chart of suggested tools dependent on the form of genetic data available and the biological question of interest.

but maintains the same theoretical underpinning. **BEAST* (Drummond et al. 2012), pronounced "*star BEAST*", expands the *BEAST* functionality for implementation of the multispecies coalescent, allowing for exploring the evolutionary trajectory of related species from multi-gene data, which has relevance for primatology and primate evolution.

2.3 | Approximate Bayesian Computation

As genomic data is large and complex, methods that offer a shortcut into the calculations of likelihoods have been extremely useful as a stop-gap measure, while computational resources catch up with the needs of ever-growing genomic data. Approximate Bayesian computation (ABC)—a model rejection approach-has become widely adopted for its ability to discriminate between complex biological models by approximating likelihood distributions (Csilléry et al. 2010). ABC methods instead sample from the posterior distribution of the parameters by finding simulated values that sufficiently resemble the observed data (Lintusaari et al. 2017). Once again, the power of this approach for population genetics lies in our ability to generate/posit hypotheses through computer simulations based on population genetics theory. ABC rejection algorithms then compare the output of simulations to the empirical data, assigning each hypothesis a probability and generating a distribution of parameters and probabilities akin to a Bayesian posterior distribution (Buzbas and Rosenberg 2015). ABC methods are widely applicable for a variety of data outside genetics, as long as there is a good synthetic data generation system. This includes a wide variety of ecological and evolutionary questions, including trait evolution in phylogenies, and even cultural modeling (Kandler and Powell 2018).

The cornerstone of ABC methods is that generating a simulation that completely captures the evolutionary forces that shaped the empirical data is extremely unlikely, but instead that some simulations will be close and as such the ABC rejection algorithm is based on measuring the discrepancy between an observed and simulated data sets (Lintusaari et al. 2017). To accomplish this, ABC methods typically operate based on handpicked summary statistics, which greatly reduce the computational needs of the rejection algorithm. ABC rejection algorithms calculate the distances between the observed and simulated summary statistics and rank individual simulations from closest to furthest. To create a distribution of summary statistics—and thus approximate a distribution of parameters specified in the simulations—ABC will retain some percentage of the closest simulations, a cut-off referred to as epsilon. In most cases, epsilon is chosen independently of the distance values in the simulations themselves: for example, keeping the closest 1% of all simulations (Beaumont et al. 2002). The choice of epsilon can affect the performance of ABC methods; a small epsilon may not capture an adequate parameter posterior distribution, but a large epsilon will add random variance (large Monte Carlo error). To ameliorate this problem, it is standard practice to conduct an adjustment of parameter distributions based on the variance of distances between summary statistics, often regression-based adjustments, which are summarized in (Beaumont 2019). Some newer solutions for the choice of epsilon which reduce the need for parameter adjustment can be found in Lintusaari et al. (2017). As described previously, handpicked summary statistics based on general principles of population genetics may lack *sufficiency*, that is, do not capture as much information to estimate a parameter as the complete data (Csilléry et al. 2010). However, selecting additional summary statistics often leads to increased distances between simulated and observed data, which in turn requires more simulations, a problem known as the "curse of dimensionality" for ABC (Beaumont 2010). Finally, generating simulations based on a biological model that does not adequately capture an evolutionary process creates model misspecification errors that can cause poor performance (Frazier et al. 2020).

For the purposes of anthropological genetics, ABC has excelled at solving demographic histories, in particular for more recent times where more direct coalescent-based methods lack power (Boitard et al. 2016). Any genetics simulator can be used to generate data for ABC, including some that we review here such as *msprime* and the *SimCOAL* family of simulators. The outputs of these simulations can be reduced to summary statistics and run through any number of software tools designed for ABC rejection. There are also software packages that will bundle genetic data simulation, ABC model rejection, parameter estimation, and regression-based parameter adjustment, such as the R package *ABC* (Csilléry et al. 2012), *DIYABC* (Cornuet et al. 2014), and *BaySICS* (Sandoval-Castellanos et al. 2014). Newer tools include *EasyABC* (Dumoulin et al. 2023), which makes ABC accessible for a variety of ecology and evolution applications in R, while *pyABC* (Schälte et al. 2022) is a similar tool for Python. *PopSizeABC* (Boitard et al. 2016) infers population size histories from a large sample of complete genomes and uses the folded allele frequency spectrum and the average zygotic linkage disequilibrium distance as summary statistics. Finally, ABC hybrids can maximize the potency of other approaches, such as combining ABC with agent-based models (a form of modeling complex biological systems) to filter parameters during the iterative building steps of agent-based models (van der Vaart et al. 2015). Combining ABC with machine learning models can greatly optimize ABC, for example, by using neural networks to construct summary statistics from the simulated data (Jiang et al. 2017; Sanchez et al. 2021) or by using Random Forest treatments to evaluate the power and accuracy of inferences, such as with *DIYACBC Random Forest* (Collin et al. 2021).

2.4 | Markov Chains and Hidden Markov Models

In previous sections, we have highlighted how the main strength of population genetics, and by extension anthropological genetics, is in its deep theoretical background. While theoretical models like the Wright-Fisher model of genetic drift or Kingman's coalescent are powerful, they are also simple; these processes are defined by simple rules, even if the models extend to thousands of individuals over thousands of generations. Because the rules are simple, one particular family of algorithms is extremely well suited for following genetic inheritance over time: the Markov processes.

Markov processes work over sequences of events, or chains, but keep computation costs low by adopting a "memoryless" approach. A first-order Markov process that exists at time t is concerned with predicting some value of interest at a future time $t+1$, but assumes that events at previous times, t–1 for example, have no effect on the future, and thus can be ignored. This ascetic way to consider the future dramatically decreases computational costs. For our purposes as geneticists, as a Markov process moves through the generations, we simply keep a record of any parameter that may be biologically meaningful, such as the size of the population during any time (t). Because the inheritance process is not deterministic but probabilistic, the calculations that drive a change in state during each step are governed by probability distributions determined by biological qualities such as the mutation rate.

There are two particular types of Markov algorithms widely implemented in genetics, Markov-Chain Monte Carlo (MCMC) and Hidden Markov Models (HMM). Explaining the inner mechanics of these algorithms is beyond the scope of this review, but in short, MCMC is a chain of events for which predicting a future state is only informed by the present state (a Markov Chain), and whether a new state is implemented or not is determined by a random drawing of values within the probability distribution (named Monte Carlo after the famous gambling location in Monaco) (Dellaportas and Roberts 2003). HMMs takes this logic one step further, by making a distinction between data we can observe, and the generative process that creates that data, which we cannot directly observe, and is thus *hidden* (Krogh 1998). An HMM applied for genetics will walk through a chain of states, say coalescent events, and will construct a genealogy, but it will not record it, and instead, will only record a value we can observe in sequencing data, such as whether that position in the genome is homozygous or heterozygous (for a detailed description, see Mather et al. 2020). This vastly reduces computational overhead and allows for Markov models to be implemented more widely, say at the scale of the entire human genome. One HMM that is frequently used in population genomic methods utilizes a statistical model first developed by Li and Stephens (2003). This model of linkage across multiple SNPs incorporates an understanding of the recombination map, can be applied to large populations, and forms the basis for many computational genomic methods, including local ancestry inference and other demographic reconstruction methods.

BEAST's, and to some degree fast*SIMCOAL2*'s, main limitation is its incomplete applicability to full genome data, as the complexity of the highly variable recombination rate along the genome obfuscates demographic trajectories and vastly increases the computational overhead. As massively parallel sequencing technology has become the dominant mode for genetic data production generally and for human genomes in particular, an extremely well-annotated genome reference means most available genome data will be composed of full genomes, even for aDNA. In the Ancestral Recombination Graph section, we illustrate the difficulties in folding full genome data into a compact data object, since every recombinant portion of the genome can coalesce differently from its neighboring regions. Given these difficulties, implementing demographic trajectory estimation from full genomes has been a slow process. The Sequentially Markovian Coalescent (SMC) family of methods is well suited for this challenge, although currently at a high computational cost. *PSMC* (Li and Durbin 2011) could only infer the population trajectory from a single unphased diploid genome. However, given the vast amount of information contained in a single genome, the results are still comparable to an analysis of hundreds of individuals using a single-locus method like *BEAST*. *MSMC* (Schiffels and Durbin 2014) expands this approach to up to ten haploid genomes at a time, and SMC++ (Terhorst et al. 2017) allows for the implementation of hundreds of genomes, and additionally does not require phasing the genomes into haploids (similar to PSMC), which was problematic when using SMC methods on ancient genomes in particular. Other notable HMM methods that have been applied to human demography include the diCal family of methods, diCal (Sheehan et al. 2013), diCal-admix (Steinrücken et al. 2018), and the modern iteration diCal2 (Steinrücken et al. 2019), as well as CHIMP (Upadhya and Steinrücken 2022). These methods have been used to great effect when estimating population demographies from difficult-to-obtain genomes, such as the Neanderthal and Denisovan genomes, other ancient genomes, or genomes from endangered species. A more thorough review of coalescent-based HMMs for demography can be found in Spence et al. (2018).

2.5 | Ancestral Recombination Graphs

One data format that has recently become popular for use in population genomic analyses is the Ancestral Recombination Graph (ARG), which catalogs all recombination and coalescence events for a set of individuals. ARGs are very informative because they theoretically contain the history of all the individuals in the dataset, from the present to the time that they coalesce. ARGs are often displayed as a series of trees known as a tree sequence, where each tree represents a region of the chromosome that shares a set of recombination and migration events. Another benefit of ARGs is that they show most of the simultaneous pressures that act to change an individual's genome, including selection, gene flow, and recombination. Because all of these processes can be detected in the ARG, it can be possible to distinguish between the causes of increased or decreased variation, which can be more difficult using other methods. Determining this true history of gene flow, coalescence, and recombination (or even a close approximation) is a computationally-intensive problem, especially as sample sizes increase. Each new recombination or admixture event increases the number of trees that have to be modeled as part of a tree sequence, and so an ARG encompassing even 10 individuals could only be modeled accurately by a tree sequence containing up to hundreds or even thousands of trees. Multiple tools have been developed to explore and derive meaning from ARGs, and each uses a different approach for simplifying the process of inferring an ARG. One method utilizes a "threading" method, which addresses the history of a single individual at a time, conditioning on the history of the other individuals in the dataset (Rasmussen et al. 2014). As this "threading" is used hundreds to thousands of times, the overall simulated ARG slowly approximates the true ARG of the samples as the fit of each individual to the whole is iteratively improved. Other approaches model the ARG as a series of tree sequences, called a *ts* object, which summarizes the genetic tree for each segment of a chromosome being analyzed (Kelleher et al. 2019). This method compresses the data further by storing any branches shared by multiple tree sequences only once, allowing for quicker computational processing, and is the primary storage method for the coalescent simulator *msprime*. Another approach uses the Li and Stephens HMM (as described in the Markov Chain and Hidden Markov Models section) to determine relationships between samples and then turns that relatedness into a series of trees (Speidel et al. 2019).

ARG-based methods have recently become useful for inferring the demographic history of populations as well. As an example, ARGs have been used to identify archaic and super-archaic introgression events (Hubisz et al. 2020) and signals of selection in human populations (Speidel et al. 2019). Other applications of ARGs include simulating evolution for hypothesis testing, correcting for biases in phylogenetic trees, and also getting an extremely detailed view of gene flow events, population divergence, and recombination (Arenas 2013). Continual improvement in the algorithms underlying ARG inference is ongoing, improving the utility of these methods to answer questions in anthropological genetics and other fields (Lewanski et al. 2024). One of the available ARG-based methods is *ARGweaver*, recently improved with the ability to incorporate a demographic model into ARG estimation, known as *Argweaver-D* (Hubisz et al. 2020). Other methods like *tsinfer* and Relate are optimized to run analyses on hundreds of chromosomes, as opposed to the dozens that *Argweaver* can accommodate (Speidel et al. 2019; Kelleher et al. 2019). *Relate* also introduces a test statistic that can detect signals of selection across the ARG (Speidel et al. 2019).

2.6 | Detecting Population Admixture and Local Ancestry Inference

Admixture, or the exchange of genetic material between populations, is one of the most significant sources of novel genetic variation in human populations. Studies of ancient populations worldwide suggest that admixture was common throughout human history, and nearly all people living today have genetic variation deriving from multiple ancestral populations. For the purposes of anthropological genetics, we care about two related research problems: has admixture occurred between two (or more) populations; and if so, can we deconvolute the various ancestries that contribute to a genome. These foci give us access to two sets of related tools: one set that tests for and quantifies the admixture between populations, and another that determines which regions of the genome are associated with which ancestry sources.

Computational tools designed to help identify admixture between populations can be further divided into two groups: those designed to visualize admixture and those designed to work as a test of admixture. Tools designed to visualize population admixture will often be bundled with tools that identify population structure (whether individuals within a sample can be grouped into any number of clusters). For example, chromosome painting tools such as *ChromoPainter* (Lawson et al. 2012) will identify haplotypes within a sample, which can be represented in a colorful graphic for users to visually observe if there is noticeable admixture, but its output can be used by *fineSTRUCTURE* (Lawson et al. 2012) to identify population structure using a clustering algorithm. Similarly, the software packages *ADMIXTURE* (Alexander et al. 2009) and *STRUCTURE* (Hubisz et al. 2009)—and its modern counterpart, *fastSTRUCTURE* (Raj et al. 2014)—will investigate population structure, including inferring the presence of distinct populations, assigning individuals to those populations, and identify migrants and admixed individuals in admixed populations, and quantify the frequency of the distinct haplotypes. These tools can produce the commonly used admixture bar plot, which can be used to visually inspect these features.

Another set of commonly used tools for admixture testing are statistical tests which quantify admixture in a genome by comparing it to genomes originating in populations closely related to the sources of gene flow. The *F*-statistics (Reich et al. 2009), better defined as drift indexes, take advantage of estimations of shared genetic drift between populations, which are proxies for periods of shared evolution in an ancestral population. When the ancestral population splits, the newly formed descendant populations experience genetic drift uniquely; therefore, genetic variation in these new populations will change independently. Admixture between these populations creates a deviation from that expectation, which can be quantified across the genome as a measure of admixture. An excellent review of how the *F* drift indexes work can be found in Peter (2016). Shared drift can also be used to measure the relatedness of two genomes, which is a common use of the *F*-3 statistic, in particular to contextualize the affinities of ancient genomes to modern populations. *F*-statistics can be calculated using *ADMIXTOOLS* (Patterson et al. 2012; Maier et al. 2023), although this requires converting data to a proprietary data format. The R package *ADMIXr* (Petr, Vernot, et al. 2019) simplifies calculating *F*-statistics for commonly used data formats. *F*-statistics can also be used to create admixture graphs or demographic models of population-level coalescence and gene flow that are most consistent with the observed data. Admixture graphs are useful for visualizing the relationships between populations and can incorporate "ghost populations" or ancestral contributors to modern populations that cannot be assigned to any known ancestry sources. Programs that are useful for creating admixture graphs include *Treemix* (Pickrell and Pritchard 2012) and *qpGraph* (Maier et al. 2023).

A closely related statistic used as a test for admixture is Patterson's *D* (Durand et al. 2011), colloquially referred to as the *ABBA-BABA* test. Patterson's *D* tests for genetic variants across the genome which are shared among non-sibling populations, which could only be shared by either gene flow, or recurrent mutation. Patterson's *D* quantifies what proportion of these sites are shared through gene flow as a proxy for the degree of admixture for a population. The software package *Dsuite* can calculate Patterson's *D* as well as the *F*-4 statistic (Malinsky et al. 2021). Patterson's *D* is not well suited for shorter regions of the genome, but *D*+—which further incorporates ancestral shared variation—has been shown to be more powerful for quantifying localized introgression (Lopez Fang et al. 2024). It is also important to remark here that because *F* and *D* statistics use a large number of positions across a genome, it is easy to construct tests for statistical robustness using resampling techniques such as bootstrapping or to identify outlier genomic regions by generating a distribution of the statistic value across the genome. This is a common method for identifying regions of archaic introgression across the human genome. Finally, there are numerous methods available for inferring the timing of admixture events, often using a recombination rate to plot the decay of long haplotypes over time. Some examples are programs such as ALDER (Loh et al. 2013), ROLLOFF (Moorjani et al. 2013), and the R package GLOBETROTTER, which uses the output of ChromoPainter (Hellenthal et al. 2014).

A second set of computational tools are designed to deconvolute the various ancestries found in an individual genome, collectively referred to as "local ancestry inference" (LAI) methods. LAI can have important implications for medical genetics, for example, as genetic variants that cause disease or increase disease risk are unique to specific populations (Martin et al. 2017), which could be targeted by personalized medicine. Yet it is difficult to identify these disease-associated variants admixed individuals, who have recent ancestry from multiple genetically distinct populations. This is an important frontier for which studies of human variation and population history can contribute to medical advances. Yet, admixed populations such as Latin Americans are often excluded from human population evolution studies, because it is difficult to determine which regions of the genome are associated with which ancestry sources.

Partitioning the genome into distinct ancestry sources using LAI has a variety of advantages for population genomics analyses. First, it is a useful tool for understanding the demographic history of a population and identifying how different proportions of an individual's genome derive from different ancestral sources. The amount of genomic sequence from a given ancestry source can be used to infer the amount of gene flow that occurred historically. As mentioned above, determining the ancestral source of a particular genomic variant is another important application of LAI. Some tools have been developed to facilitate the inclusion of admixed populations into genome-wide association studies, a powerful tool to identify disease risk alleles that has often focused on unadmixed individuals (Popejoy and Fullerton 2016; Atkinson et al. 2021). LAI can also identify genetic variants that may have been adaptive in a specific ancestral population historically, but are now primarily found in admixed individuals (Secolin et al. 2019; Witt et al. 2023).

LAI tools use a variety of methods to assign ancestry to genomes of interest, but typically a "reference panel" of genomes from different ancestry sources is compared to an admixed genome, and different sites or regions in the genome are assigned to the different ancestry sources in the reference panel. An early LAI method, *HAP-MIX*, relies on the Li and Stephens model (2003) for inference and has been built on in a number of ways, including the tool *Recomb-Mix* (Wei et al. 2023). Another approach, which is utilized by one of the most commonly-used LAI tools, *RFMix* (Maples et al. 2013), utilizes a Conditional Random Field method, a type of Markov model similar to HMMs that attempts to predict ancestry assignment based on patterns observed in the data. More recent LAI tools even utilize machine learning methods, either in addition to other algorithms or as the primary tool for ancestry classification (Alizadeh et al. 2023; Montserrat et al. 2020).

Multiple tools now exist for LAI, and many of them are able to assign ancestry to admixed genomes with a high rate of accuracy. Tool choice is, therefore, dependent on the specific dataset being analyzed. Different programs require different data and parameter inputs, and so the tool should be chosen based on a specific project's parameters. As an example, *RFMix* (Maples et al. 2013) and *Loter* (Dias-Alves et al. 2018) both require phased genomes, while other tools like *LAMP-LD* (Baran et al. 2012) do not. Some LAI tools require many parameters defining the timing and amount of admixture, while others like *Loter* (Dias-Alves et al. 2018) and *Recomb-Mix* (Wei et al. 2023) can infer them as part of the ancestry inference process. Different programs also have different outputs—some tools like *LAI-Net* and *LAMP-LD*

(Baran et al. 2012) output ancestry sources for genome windows, while others like *Recomb-Mix* (Wei et al. 2023) output the ancestry for specific SNPs.

LAI software performance is also dependent on the complexity of the LAI problem. For example, more recent admixture events are easier to analyze because the ancestry segments will be larger, with fewer segments broken up by recombination. Most tools perform equally well in cases of recent admixture, while in cases of admixture events that occurred 100 generations ago, *RFMix* (Maples et al. 2013) performs very poorly, while *FLARE*, *MOSAIC*, and *AICRF* (Salter-Townshend and Myers 2019; Browning et al. 2023) still maintain a high accuracy. Reference panel selection is also important—accuracy with LAI tools is highest when each ancestry source is represented by a closely-related population in the reference panel. More distant reference populations will negatively impact results, and while some tools like *FLARE* (Browning et al. 2023) and *MOSAIC* (Salter-Townshend and Myers 2019) can infer ancestry from ancestry sources that are not included in the reference panel, accuracy for these ancestry segments is inevitably lower.

2.7 | Archaic Introgression

When anatomically modern humans expanded outside of the African continent, they encountered other hominids, like Neanderthals and Denisovans, that were already living in Eurasia. Several Neanderthal genomes and a single Denisovan genome have now been sequenced to high depth (Green et al. 2010; Meyer et al. 2012; Prufer et al. 2014; Mafessoni et al. 2020), and comparison between archaic and modern human genomes has confirmed that modern humans admixed with Neanderthals and Denisovans. The presence of archaic ancestry in modern human genomes is referred to as archaic introgression. Nearly all humans living today with ancestry from outside of Africa have some amount of archaic ancestry in their genomes (Sankararaman et al. 2014; Sankararaman et al. 2016).

Archaic introgression has been studied in detail and has revealed insights about gene flow between archaic and modern humans, including (1) that multiple Denisovan (and possibly also multiple Neanderthal) populations admixed with humans (Browning et al. 2018; Larena et al. 2021; Villanea and Schraiber 2019; Witt et al. 2022), (2) that the majority of archaic ancestry was removed from modern human populations within 10 generations of admixture (Harris and Nielsen 2016; Petr, Paabo, et al. 2019), producing archaic ancestry “deserts” throughout the genome (Sankararaman et al. 2014; Skov et al. 2020), and (3) that gene flow has occurred between Denisovans and Neanderthals and between Neanderthals and early anatomically modern humans (Slon et al. 2018; Petr, Paabo, et al. 2019; Kuhlwilm et al. 2016; Chen et al. 2020; Hubisz et al. 2020). Much of the focus of the study of archaic introgression is on the identification of archaic genetic variants that have consequences for modern human health and fitness. Many of these variants have increased in frequency in human populations due to positive selection, a process known as adaptive introgression. Perhaps the most famous example of adaptive introgression is in *EPAS1*, a gene which impacts the body's response to hypoxia. Tibetans have a unique variant of *EPAS1* that allows them to thrive at high altitude, and it is very similar to the variant found in Denisovans (Huerta-Sanchez et al. 2014). More recent work has suggested that Tibetans inherited the Denisovan *EPAS1* variant long before they lived at high altitudes, and that the variant was selected for only in the last 10,000 years (Zhang et al. 2021). Many other examples have been identified, including genes related to skin pigmentation (Vernot and Akey 2014; Gittelman et al. 2016), *ABO* blood groups (Villanea et al. 2021), and immune function (Gittelman et al. 2016; Dannemann and Kelso 2017; and Villanea et al. 2023).

Detection of archaic introgression, or ancestry from archaic hominids like Neanderthals and Denisovans, can be considered a special case of LAI. It is, in some ways, a more challenging problem than admixture between modern human groups because gene flow between archaic and modern humans occurred over 40,000 years before present, and so most archaic segments left in the genome are small. However, modern and archaic humans also diverged 500,000 years ago, making the archaic segments retained in the genome easier to identify. Many methods exist to detect archaic introgression (see Racimo et al. 2015 and Ahlquist et al. 2021 for detailed summaries), but most use the sequenced Neanderthal and Denisovan genomes as archaic references and African genomes as an unadmixed reference population. Segments in an individual's genome with high similarity to a sequenced archaic genome and low similarity to an African genome are then inferred to be archaic in origin. Some newer methods like *Sprime* (Browning et al. 2018) and HMM-based methods (Skov et al. 2018) can detect introgression without comparison to an archaic genome, while other methods like *ArchIE* (Durvasula and Sankararaman 2019, 2020) and *IBDMix* (Chen et al. 2020) can identify archaic introgression without comparison to unadmixed outgroups. Another recent method, *Admixfrog*, uses an HMM to detect introgression between populations and a Bayesian model to infer demographic parameters from the data (Peter 2020). It has also been optimized to account for DNA damage and contamination in aDNA. Other methods use summary statistics or ARGs to identify signatures of archaic introgression. These methods allow for the identification of “super-archaic introgression” or introgression from hominids that were more divergent from modern humans than Neanderthals or Denisovans, which has so far been identified in Africans (Durvasula and Sankararaman 2020), Asians and Oceanians (Mondal et al. 2019), and Denisovans (Prufer et al. 2017). The existence of super-archaic introgression in modern humans is still under debate, as human populations whose ancestors may have overlapped in time and space with other fossil hominins have not yet shown evidence for super-archaic introgression (Teixeira et al. 2021), and the patterns in African populations may also be explained by deep structure in Africa (Ragsdale et al. 2023). However, the use of these novel methods have opened up exciting opportunities for understanding human demographic history.

2.8 | Identifying Positive Selection

The study of natural selection has long been a motivating factor in understanding the evolution of species and populations. Discussing all methods designed to identify the directionality and prevalence of natural selection is beyond the scope of this review. Instead, we focus on methods to identify and quantify positive selection, as it represents a history of adaptation in a population. Briefly, if an allele provides a selective advantage, it often leaves a pattern of low genetic diversity surrounding the allele under selection if the beneficial allele rises to a high enough frequency (Smith and Haigh 1974). The size of the region of reduced genetic diversity depends on multiple factors, including the strength of selection and recombination rate in the region. Some of the first signals of positive selection in humans were identified using a targeted approach that focused on specific phenotypes, including lactase persistence (Tishkoff et al. 2007) and malaria resistance (e.g., Ruwende et al. 1995). Now, it is more common to use genome-wide scans using a variety of methods to identify signals of positive selection and then further characterize any signals that arise to infer how the genes with positive signals may have evolved (e.g., Yi et al. 2010; Voight et al. 2006; Sabeti et al. 2007; McVicker et al. 2009; Axelsson et al. 2013; Klunk et al. 2022). How positive selection is detected has been discussed comprehensively in the past (see Nielsen et al. 2005; Vitti et al. 2013), so here we will provide only a brief review.

Positive selection can be studied at the macro-evolutionary scale (between species) and the micro-evolutionary scale (within species). There are two common macro-evolutionary methods to detect positive selection by comparing genetic variation within genes across species: the McDonald-Kreitman test and the HKA (named for Hudson, Kreitman, and Aguade) test (McDonald and Kreitman 1991; Hudson et al. 1987). The first test focuses on synonymous and nonsynonymous mutations, while the second compares within-species genetic variation to across-species genetic divergence. An increase in nonsynonymous mutations or polymorphisms within a single lineage or increased polymorphisms in a particular species or clade would suggest that evolution has occurred in that species or clade. Micro-evolution is detected using several different methods that all leverage the pattern of reduced heterozygosity near a variant under positive selection, or selective sweeps. Comparing synonymous and nonsynonymous mutations between populations can be useful for detecting micro-evolution in addition to macro-evolution. Another commonly-used method focuses on linkage disequilibrium, where positive selection has produced haplotypes that are longer than expected given the recombination rate (Sabeti et al. 2002; Voight et al. 2006; Sabeti et al. 2007; Huff et al. 2010; Liu et al. 2013). Other methods focus on the distribution of allele frequencies across a gene or genome, also known as the site frequency spectrum—an excess of low or high frequency alleles can indicate positive selection, while an excess of intermediate frequency alleles can instead suggest balancing selection (Fu and Li 1993; Fu 1997; Tajima 1989; Tajima 1993; Nielsen et al. 2009). Finally, methods that compare allele frequencies between populations, such as F_{ST}-based methods (Wright 1949; Weir and Cockerham 1984) are especially useful for identifying cases of local adaptation, where positive selection has occurred in some populations but not others (Lewontin and Krakauer 1973; Akey et al. 2002; de Villemereuil and Gaggiotti 2015; Shriver et al. 2004; Yi et al. 2010; Oleksyk et al. 2008; Riebler et al. 2008; Beaumont and Balding 2004; Librado and Orlando 2018; Bonhomme et al. 2010).

Today, it is common for researchers to use multiple tests of selection to confirm selection signals and also account for the differences in how methods detect selection. For example, two LD-based methods, *iHS* and *XP-EHH*, are often used together because they detect selective sweeps at different allele frequencies (e.g., Pickrell et al. 2009), and selection scan software like *selscan 2.0* and the R package *rehh* calculate these and other haplotype-based selection statistics together (Szpiech 2021; Klassmann and Gautier 2022). *XP-EHH* is also often combined with population branch statistic-based methods because the methods use different approaches to identify local adaptation in specific populations (e.g., Yi et al. 2010). Some methods even incorporate multiple of the above methods, such as *hapFLK*, an F_{ST}-based method that also takes linkage disequilibrium into account (Fariello et al. 2013), and the composite likelihood ratio (CLR) test, which looks at population differentiation and allele frequencies (Chen et al. 2010; Kim and Stephan 2002; Nielsen et al. 2005; deGiorgio et al. 2014). While many of these methods have been used extensively with human genomic data, they can often be adapted to other species as well. Projects utilizing Reduced Representation Sequencing methods like RADseq to reduce sequencing costs, and those working with species without high-quality reference genomes, can still use some methods that look for environmental-genetic correlations (e.g., Guillot et al. 2014) or F_{ST}-based methods (see previous paragraph) to identify selection, although additional sequencing may be required in the future to fully understand the selection signals (Manel et al. 2015).

There is also a vast selection of methods designed to capture more elaborate or subtle forms of selective sweeps. This includes detecting ongoing positive selection or incomplete selective sweeps caused by shifting natural selection landscapes, using CLR tests (Vy and Kim 2015) or the haplotype selection program *saltiLASSI* (DeGiorgio and Szpiech 2022). Identifying adaptive introgression is possible with tools such as *Volcano Finder* (Setter et al. 2020) or using convolutional neural networks (Gower et al. 2021). Detecting selection sweeps that occurred in the past or in an ancestral population is possible by using data from three related populations (Racimo 2016). Conversely, a metric known as the singleton density score (SDS) was developed to detect recent positive selection and focuses on the distance between SNPs across individuals (Field et al. 2016). Convergent adaptive evolution can be studied by identifying the increased covariance created by shared selective sweeps (Lee and Coop 2017).

When analyzing genomes for signals of positive selection, interpretation and contextualization of the results is necessary. It is important to keep in mind that positive selection has a signal in the genome that can appear similar to other forces that act on the genome, including population bottlenecks and background selection, which is the removal of deleterious alleles from a genomic region that is highly conserved. Some modern methods that use CLR-based methods like *Sweepfinder* (Kim and Stephan 2002; Nielsen et al. 2005; Huber et al. 2015) explicitly test whether regions with a signal of selection represent positive selection or background selection, and many methods are tested for their robustness against demographic effects (e.g., Huber et al. 2015; Chen et al. 2010; Bonhomme et al. 2010). A recent biological anthropology perspective on selective scans also emphasized the importance of including community engagement in work with human populations, and including diverse groups in genomic studies and as scientific researchers (Hernandez and Perry 2021). They also made the important point that when hypotheses are made about human adaptation based on selection scans, it is important to consider historical biases and how the genetic findings may be misused. Considering the results of a selection scan critically, both from a methodological and an anthropological perspective, is an important part of interpreting the genomic data.

One of the most challenging parts of identifying signals of positive selection is understanding the functional significance of a variant that has been positively selected. Sometimes the allele under selection changes the form and function of a protein (as in the case of lactase persistence or malaria resistance in glucose-6-phosphate dehydrogenase; Tishkoff et al. 2007; Ruwende et al. 1995), or the allele is identified as a structural variant like increased copy number (Villanea et al. 2023; Axelsson et al. 2013; Wroblewski et al. 2023), but often a genomic region with a signal for positive selection shows no obvious genetic mechanism. For example, the *EPAS1* variant in Tibetans has been shown to be under strong selection (Yi et al. 2010; Huerta-Sanchez et al. 2014) and is associated with a phenotype of decreased hemoglobin concentration at altitudes above 4000 m (Beall et al. 2010), but the genetic mechanism underlying this phenotype has yet to be identified despite years of study. To go beyond selection signals to understand how human phenotypes are impacted at the molecular level, collaborations with experts in cell biology, biochemistry, modulation of gene expression, and more are paramount for providing a more comprehensive view on selection and how variants under selection are actually impacting phenotypes.

We also recognize that studying other forms of natural selection is of great importance to anthropological geneticists and would like to briefly touch on three approaches beyond the selection of selective sweep tools we have discussed so far. First, if multi-species data is available for coding gene regions, tests of natural selection based on the accumulation of synonymous versus nonsynonymous mutations (d_N/d_S) can distinguish between negative selection, positive selection, and balancing selection in particular branches of a phylogeny. *PAML* (Yang 2007) is the classic tool for implementing d_N/d_S tests, but other programs such as *HyPhy* (Kosakovsky Pond et al. 2020) are available, and d_N/d_S tests are now even incorporated into the sequence manipulation tool *MEGA* (Kumar et al. 2016). Second, the detection of balancing selection is more difficult than detecting positive selection (for an expanded view see, Fijarczyk and Babik 2015; Bitarello et al. 2023; and Soni and Jensen 2024). Yet, improved methods have been developed in the last decade, including methods using CLR tests (DeGiorgio et al. 2014) or improved summary statistics, such as β (Siewert and Voight 2017), which is implemented in the software *BetaScan2* (Siewert and Voight 2020). Finally, the incorporation of well-dated ancient DNA data with modern data into genetic time series has allowed for the inference of allele frequency trajectories over time. Methods include using single allele frequency time series (Schraiber et al. 2016), the package *diplo-locus* (Cheng and Steinrücken 2023; Fine and Steinrücken 2024), and the *AGES* database (Akbari et al. 2024).

2.9 | A Brief Review of Neural Network Methods

Machine Learning (ML), which aims to develop computer algorithms that improve with experience, is a vast field of research with origins dating back to at least the 1950s (Jordan and Mitchell 2015). ML has applications in nearly all branches of research: from the public release of large language models, such as ChatGPT and other customizable offshoots, to the incorporation of image classification software to expedite museum specimen cataloging, machine learning methods are now virtually everywhere. Even in genetics and genomics specifically, ML has vast applications, which deserve reviews of their own (e.g., Libbrecht and Noble 2015). Molecular anthropology and population genetics at large were some of the earliest adopters of ML approaches; early *unsupervised* ML algorithms such as Principal Component Analysis and other clustering tools are ubiquitous as a way to visualize highly complex genome data (Rosenberg et al. 2002; Novembre et al. 2008). A different form of ML algorithms, *supervised* ML, has only been applied to genomic data as recently as 2016 (Pudlo et al. 2016; Sheehan and Song 2016) as the capabilities to create labels for training data sets required a critical mass of available genomes and better genomic simulation engines that were previously unavailable. While supervised ML is a relatively new paradigm to evolutionary genetics and biological anthropology, it has already been implemented to solve many problems within genomics, including detecting subtle signals of positive selection, deconvoluting ancestries in genomes from admixed populations, and reconstructing complex family genealogies (Schrider and Kern 2018; Finke et al. 2021; Riley et al. 2024). A subcategory of supervised machine learning, *Deep Learning* (DL), includes most of the tools we review here (for an expanded summary of the applications, see LeCun et al. 2015).

The advantages of implementing DL tools are an ongoing topic of scholarly research, yet the prevalence and success of DL tools applied to biological data is self-evident (Flagel et al. 2019). However, one specific advantage of DL tools is the generation of bespoke data summaries. This advantage of some DL methods harks back to the early days of population genetics, and our reliance on handpicked summary statistics. As described in the Summary Statistics section above, a summary statistic may lack sufficiency if the idiosyncrasies of a particular dataset are not taken into account when selecting data summaries, for example for ABC rejection methods. In contrast, every time a DL model is trained, it is creating its own set of summary statistics, incorporating salient data features that are applicable to solving that specific problem. In this sense, every DL model creates a bespoke solution for each biological question, assuming that it is trained under an applicable biological model. Model misspecification is as much a problem for DL as it is for ABC methods, as mentioned previously.

This is a good moment to point out that DL approaches are sometimes labeled derogatorily as "black boxes." The relative obscurity of *why* some data features are selected is sometimes considered a weakness of DL. Some researchers are more interested in the insight generated by understanding the biological meaning of the predictive model, rather than being satisfied with the predictive accuracy of the tool itself (Xu and Jackson 2019). While there is some truth to this argument, as some more complex DL architectures can be impossible to interpret, many articles for supervised learning models in population genetics emphasize model interpretation in addition to model prediction (Novakovsky et al. 2022; Azodi et al. 2020), as observing what features of genomic data have concentrated predictive power can be a source of valuable insight, and we find that fears of the unaccountability of ML methods are often misplaced.

2.10 | Off-the-Shelf Neural Network Methods Applied to Molecular Anthropology Problems

Neural network (NN) methods are a form of DL that have been applied broadly across studies of all evolutionary forces, but the relative novelty of these methods has at times felt like a paradigm shift. Given the relative breakthrough nature of NN models applied to study the forces of evolution, we thought this would warrant its own section rather than grouping these approaches with their traditional counterparts. As a note, the gambit of ML tools applied to population genetics is vastly larger in scope than simply NN methods, with many of these methods equaling or outperforming their NN counterparts, but these are beyond the scope of this review.

As previously suggested, the primary motivation for implementing NN methods is the vastness of genomic data made available by massive parallel sequencing technologies. Public access to hundreds (now thousands) of complete human genomes has created an "embarrassment of riches" situation, where the computational resources needed to apply traditional statistical analyses to such rich data became intractable. Summary statistics are useful for predicting some biologically meaningful parameters, and methods like Approximate Bayesian Computation (ABC) are powerful tools that take advantage of our ability to rapidly simulate genomic data. However, data loss in the process of using these methods is inevitable. Alternative statistical methods which can take advantage of simulated genomes' information in its entirety, at a thrifty computational cost, would provide more powerful alternatives for studying biological processes. Earlier efforts included using ML to refine the subset of simulations fed to ABC analyses, which would improve its estimation power while reducing computational demand. Approximate Bayesian Computation via Random Forest or *ABC-RF* (Pudlo et al. 2016) approaches have shown some success in the field of human evolution, such as predicting super-archaic contributions to the human gene pool (Mondal et al. 2019; Montinaro et al. 2021).

Here, we have reviewed three modes of applying NN to solve population genetics problems, with special interest in human biology. The first approach is the simplest; a NN is trained on sets of genomic data that can be classified with labels. Once the NN is trained, new data can be used as input, which will be classified within the scope of the training data. One example of this approach is *Locator* (Battey et al. 2020), a tool for predicting the geographic origin of individuals from unphased genotype data. The trained NN is then used to predict the geographic location of new samples based on their genotypes. *Locator* can also reveal portions of an individual's genome enriched for ancestry from specific geographic areas. While Locator is primarily employed for wildlife species, including primates, other applications such as ancient individual affinities can be of interest to anthropologists.

The second—and by far the most common—application of NN involves the implementation of genome simulators to produce training data, where the trained NN is then fed empirical data as input for classification or parameter regression. This mode is preferred, as empirical training data is finite, and the process of labeling is labor-intensive. On the other hand, coalescent-based simulators are well-supported by theory (labeling is built into the design process) and computationally thrifty (allowing for the continuous generation of training data as needed). Early adopters of coalescent-based simulation for training NN include ArchIE (Durvasula and Sankararaman 2019), using *ms* as a simulation platform and logistic regression for inference of archaic ancestry in modern human genomes; and Villanea and Schraiber (2019), using *msprime* to simulate Neanderthal introgression into human genomes and an image classification NN to discriminate between demographic models of Neanderthal-human introgression.

More refined applications of simulations as training data for NN include applications for the detection of positive selection: *diploS/HIC* (Schrider and Kern 2018) uses *discoal* to simulate the coalescent process jointly with soft selection sweeps (cases where the beneficial allele is associated with multiple distinct haplotypes, instead of a single haplotype), and a deep convolutional neural network (CNN) approach to classification. *ImaGene* (Torada et al. 2019) uses *msms* to jointly simulate demography and natural selection on a single locus, and a CNN for the detection and quantification of natural selection. Additional applications to population genetics problems include *ReLerNN* (Adrion et al. 2020) using coalescent simulations to train and predict recombination rates directly from a genotype alignment; *disperseNN* (Smith et al. 2023), using *SLiM* for forward-in-time spatial genetic simulations of multiple loci to train a NN to infer the per-generation dispersal distance from a single population.

A final mode of applying NN to solve population genetics problems is the newest and most complex: the iterative training of two NNs, also known as Generative Adversarial Networks or GANs. The intuition of applying a GAN is that simulated data does not originate from empirical data, and thus may not capture some aspects of the "real world." A GAN works by creating two networks that are trained together: a *generator* NN simulates data using a genome simulator as in the previous approach, and a *discriminator* NN, whose job is to distinguish between empirical data and simulated data. If the discriminator NN finds the simulated data a poor fit to empirical data, it will bounce this information back to the generator NN, which will use the information to better generate simulated data, and so on. The goal is for the generator NN to produce simulated data that is indistinguishable from empirical data, and yet it is labeled, and thus useful for classification or parameter regression. Two examples of GANs applied to anthropological problems are using a GAN to infer the parameters of human demographic history from genotype data, using a generator NN based on *msprime*, and a CNN as a discriminator NN (Wang et al. 2021); and using a GAN for human genome scans to detect deviations from neutrality that may be indicative of natural selection (Riley et al. 2024). This last example uses a two-stage approach, first using a GAN between an *msprime* generator and CNN discriminator to detect deviations of neutrality, then further training the discriminator using a small number of forward simulations generated using *SLiM*. These studies showcase the promise of NNs as the premier analysis tool for massive parallel sequencing data, in particular for humans, where non-genetic forms of data can help guide simulation processes and help discriminate between competing evolutionary processes.

3 | Discussion

All the computational tools and techniques mentioned above are a representation of the innovation in computational genomics methods we have seen in the past 10 years. Computational tools are now able to process more genomic data faster than ever, often in ways that can distinguish between the various forces acting on the human genome better than previous methods. Many of these methods can help answer questions of interest to anthropological geneticists, and we think that the application of these tools to new populations, species, and anthropological questions has the potential to generate exciting new results. In a perfect world, the read-the-docs, programming scripts, and software packages should be designed to be user-friendly and well-documented and supported. We note, however, that not every tool we reviewed here lives up to that standard. Poorly documented computational tools and tools that are non-user-friendly can cause significant delays and many frustrations for researchers, and we advocate for tool developers to strive for ease of use. For researchers wanting to add computational genomics methods to their work, we have four primary recommendations.

First, a familiarity with programming languages and computational clusters is essential for the usage of these tools; we suggest that interested users achieve a working understanding of the following languages: *UNIX* is the operating system used for computer clusters and other terminals, which takes command in the UNIX shell language or BASH. Most computational tools we explore here do not have GUIs (graphical user interfaces), so all commands to initialize them are performed directly in the form of BASH commands. In addition, knowledge of basic navigational and summary commands is helpful for data archiving, processing, and storage. Furthermore, computational clusters queue analyses in the form of “jobs” which for convenience are managed by a job scheduler, which takes input in the form of command lines; *SLURM* is particularly common and takes commands in modified BASH, though there are others that may use their own languages. Using a job scheduler allows you to run these analyses in parallel on computational clusters, which are designed to handle computationally-intensive jobs and run without any needed supervision.

While many programming languages were used to develop the analysis tools we explore here, the field of computational biology has generally converged on two very similar object-based languages, Python and R, at least for their interface scripts. For example, while *msprime* is developed and runs in C++, the user command scripts are written in Python, and all the documentation anticipates Python competency. Python is commonly used in genomics broadly, and is especially useful for processing large data files, either to convert genomic data files between formats, or to analyze outputs using built-in graphical and accessory genomic tools. The final language we recommend learning is R, which has numerous statistical and mathematical packages that are useful for genomic analyses and is also useful for processing large data tables (especially mathematically) and producing publication-quality figures. There is a lot of overlap in the post-analysis capabilities of Python and R, so learning R in addition to Python can be left to personal preference and programming needs.

Second, think as critically about your computational tool selection as you would about any other experimental design. This includes a good understanding of how hypothesis testing works for these analyses, how null hypotheses are built into models, and how statistical validity is tested. Each of the above categories of methods has multiple tools available for use, and they differ in the input data required, the number of individuals/sites they can analyze, and the approach they take in producing results. Sometimes your data and its characteristics—or your specific methodological needs—can guide you to the correct tool, while other times there are multiple options that would work perfectly well. There are many articles that compare metrics across multiple tools, such as Schubert et al. (2020) for LAI methods and Vitti et al. (2013) for detecting natural selection. Sometimes, the manuscript linked to a computational tool will also highlight its advantages over other tools and its limitations. Once you have identified a tool that fits your needs, start by reading the *readme* or manual, which will often provide details that can help you narrow down the variables required for your specific project.

Third, get comfortable with an expanded philosophy for arriving at statistical robustness and validity. Proper interpretation of results for Likelihood-based, DL, and Bayesian-based methods requires a solid understanding of basic statistics and evolutionary theory. Many of these tools have built-in metrics to assess precision and robustness that need to be considered in balance with their usefulness. In particular, negative and positive error rates tend to be higher than more traditional methods, but this is considered an acceptable trade-off for access to solutions that would be otherwise computationally intractable. For this reason, we recommend becoming familiar with interpreting precision-recall curves and other representations of statistical power. In addition, it is vital to understand the basics of model fitting, in particular how overfitting is a problem for some of these analyses, and how to test and control for this issue. For ML in particular, it is paramount to understand the metrics for proper training of the algorithms, including interpreting training curves before accepting any results at face value. A particularly useful mindset when implementing these more advanced statistical and computational tools is to accept them as heuristics, more so than providers of definitive answers. The goal of most of these methods is to produce insight into biological processes from data that is extremely noisy, and thus most of these methods result in a *good* answer, but not necessarily the *best* answer. After all, it is impossible to fully realize the evolution of actual populations given our current computational limitations.

A *final* recommendation is that new learners of computational methods take advantage of the resources available to them. For example, many institutions own High Performance Computing clusters (HPC) which are made available to any affiliated researcher and offer workshops in specific programming languages and commonly-used tools, and can often be directly contacted to help with questions related to installing and running software. There are tutorials for learning programming languages online that are available for free, including on places like Youtube. Many commonly-used programs also have associated tutorials or quick-start guides that are either made by the creators of the tool or by other users, who apply the tools extensively. The creators of the tools themselves are often responsive to help with troubleshooting, or have created online communities to support all users of a given tool.

It's a new world for molecular anthropology, and the possibilities are endless. We are excited for future research, which will undoubtedly challenge our current understanding of human evolution—and we are even more excited to see that the ruling paradigm in computational biology is for software and other analysis tools to be publicly available and free to use. We believe that computational and statistical competency is a lower cost of entry for research relative to field- and laboratory-based forms of genetics, and that this trend is democratizing access to research across all institutions and all countries. This increase in free access to scientific resources is exemplified by the public availability of modern and ancient genomes, and the acceptance of preprint services to remove paywalls to exciting new results and methodologies. Through ongoing and important conversations, data sharing is arriving at a point where Indigenous sovereignty and ethical concerns can be balanced with the needs of public access. For this reason, we advocate more than ever for the inclusion of computational competency in the curriculum for all anthropology trainees.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1000 Genomes Project Consortium. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526, no. 7571: 68.

Adrion, J. R., J. G. Galloway, and A. D. Kern. 2020. "Predicting the Landscape of Recombination Using Deep Learning." *Molecular Biology and Evolution* 37, no. 6: 1790–1808.

Ahlquist, K. D., M. M. Banuelos, A. Funk, et al. 2021. "Our Tangled Family Tree: New Genomic Methods Offer Insight Into the Legacy of Archaic Admixture." *Genome Biology and Evolution* 13, no. 7: 1–19.

Akbari, A., A. R. Barton, S. Gazal, et al. 2024. "Pervasive Findings of Directional Selection Realize the Promise of Ancient DNA to Elucidate Human Adaptation." *bioRxiv*.

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. 2002. "Interrogating a High-Density SNP Map for Signatures of Natural Selection." *Genome Research* 12: 1805–1814.

Alexander, D. H., J. Novembre, and K. Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19, no. 9: 1655–1664.

Alizadeh, F., H. Jazayeriy, O. Jazayeri, and F. Vafaee. 2023. "AICRF: Ancestry Inference of Admixed Population With Deep Conditional Random Field." *Journal of Genetics* 102: 49.

Allentoft, M. E., M. Collins, D. Harker, et al. 2012. "The Half-Life of DNA in Bone: Measuring Decay Kinetics in 158 Dated Fossils." *Proceedings of the Royal Society B: Biological Sciences* 279, no. 1748: 4724–4733.

Anderson, C. N., U. Ramakrishnan, Y. L. Chan, and E. A. Hadly. 2005. "Serial SimCoal: A Population Genetics Model for Data From Multiple Populations and Points in Time." *Bioinformatics* 21, no. 8: 1733–1734.

Anderson, E. C. 2024. "Practical Computing and Bioinformatics for Conservation and Evolutionary Genomics." https://eriqande.github.io/eca-bioinf-handbook/.

Arenas, M. 2013. "The Importance and Application of the Ancestral Recombination Graph." *Frontiers in Genetics* 4: 206.

Atkinson, E. G., A. X. Maihofer, M. Kanai, et al. 2021. "Tractor Uses Local Ancestry to Enable the Inclusion of Admixed Individuals in GWAs and to Boost Power." *Nature Genetics* 53: 195–204.

Axelsson, E., A. Ratnakumar, M.-L. Arendt, et al. 2013. "The Genomic Signature of Dog Domestication Reveals Adaptation to a Starch-Rich Diet." *Nature* 495: 360–364.

Azodi, C. B., J. Tang, and S.-H. Shiu. 2020. "Opening the Black Box: Interpretable Machine Learning for Geneticists." *Trends in Genetics* 36, no. 6: 442–455.

Baran, Y., B. Pasaniuc, S. Sankararaman, et al. 2012. "Fast and Accurate Inference of Local Ancestry in Latino Populations." *Bioinformatics* 28, no. 10: 1359–1367.

Battey, C. J., P. L. Ralph, and A. D. Kern. 2020. "Predicting Geographic Location From Genetic Variation With Deep Neural Networks." *eLife* 9: e54507.

Baumdicker, F., G. Bisschop, D. Goldstein, et al. 2022. "Efficient Ancestry and Mutation Simulation With Msprime 1.0." *Genetics* 220, no. 3: iyab229.

Beall, C. M., G. L. Cavalleri, L. Deng, R. C. Elston, Y. Gao, and Y. T. Zheng. 2010. "Natural Selection on *EPAS1 (Hif2alpha)* Associated With Low Hemoglobin Concentration in Tibetan Highlanders." *PNAS* 107, no. 25: 11459–11464.

Beaumont, M. A. 2010. "Approximate Bayesian Computation in Evolution and Ecology." *Annual Review of Ecology, Evolution, and Systematics* 41, no. 1: 379–406.

Beaumont, M. A. 2019. "Approximate Bayesian Computation." *Annual Review of Statistics and Its Application* 6, no. 1: 379–403.

Beaumont, M. A., and D. J. Balding. 2004. "Identifying Adaptive Genetic Divergence Among Populations From Genome Scans." *Molecular Ecology* 13, no. 4: 969–980.

Beaumont, M. A., W. Zhang, and D. J. Balding. 2002. "Approximate Bayesian Computation in Population Genetics." *Genetics* 162, no. 4: 2025–2035.

Benson, D. A., M. Cavanaugh, K. Clark, et al. 2012. "GenBank." *Nucleic Acids Research* 41, no. D1: D36–D42.

Bergström, A., S. A. McCarthy, R. Hui, et al. 2020. "Insights Into Human Genetic Variation and Population History From 929 Diverse Genomes." *Science* 367, no. 6484: eaay5012.

Bitarello, B. D., D. Y. Brandt, D. Meyer, and A. M. Andrés. 2023. "Inferring Balancing Selection From Genome-Scale Data." *Genome Biology and Evolution* 15, no. 3: evad032.

Boitard, S., W. Rodriguez, F. Jay, S. Mona, and F. Austerlitz. 2016. "Inferring Population Size History From Large Samples of Genome-Wide Molecular Data – An Approximate Bayesian Computation Approach." *PLoS Genetics* 12, no. 3: e1005877.

Bonhomme, M., C. Chevalet, B. Servin, et al. 2010. "Detecting Selection in Population Trees: The Lewontin and Krakauer Test Extended." *Genetics* 186, no. 1: 241–262.

Bouckaert, R., J. Heled, D. Kühnert, et al. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLoS Computational Biology* 10, no. 4: e1003537.

Browning, S., B. Browning, Y. Zhou, S. Tucci, and J. Akey. 2018. "Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture." *Cell* 173, no. 1: 1–9.

Browning, S. R., R. K. Waples, and B. L. Browning. 2023. "Fast, Accurate Local Ancestry Inference With FLARE." *American Journal of Human Genetics* 110: 326–335.

Buzbas, E. O., and N. A. Rosenberg. 2015. "AABC: Approximate Approximate Bayesian Computation for Inference in Population-Genetic Models." *Theoretical Population Biology* 99: 31–42.

Catchen, J. M., P. A. Hohenlohe, L. Bernatchez, W. C. Funk, K. R. Andrews, and F. W. Allendorf. 2017. "Unbroken: RADseq Remains a Powerful Tool for Understanding the Genetics of Adaptation in Natural Populations." *Molecular Ecology Resources* 17, no. 3: 362–365.

Chen, H., N. Patterson, and D. Reich. 2010. "Population Differentiation as a Test for Selective Sweeps." *Genome Research* 20: 393–402.

Chen, L., A. B. Wolf, W. Fu, L. Li, and J. M. Akey. 2020. "Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals." *Cell* 180: 677–687.

Cheng, X., and M. Steinrücken. 2023. "Diplo-Locus: A Lightweight Toolkit for Inference and Simulation of Time-Series Genetic Data Under General Diploid Selection." *bioRxiv*.

Collin, F. D., G. Durif, L. Raynal, et al. 2021. "Extending Approximate Bayesian Computation With Supervised Machine Learning to Infer Demographic History From Genetic Polymorphisms Using DIYABC Random Forest." *Molecular Ecology Resources* 21, no. 8: 2598–2613.

Cornuet, J. M., P. Pudlo, J. Veyssier, et al. 2014. "DIYABC v2. 0: A Software to Make Approximate Bayesian Computation Inferences About Population History Using Single Nucleotide Polymorphism, DNA Sequence and Microsatellite Data." *Bioinformatics* 30, no. 8: 1187–1189.

Csilléry, K., M. G. Blum, O. E. Gaggiotti, and O. François. 2010. "Approximate Bayesian Computation (ABC) in Practice." *Trends in Ecology and Evolution* 25, no. 7: 410–418.

Csilléry, K., O. François, and M. G. Blum. 2012. "Abc: An R Package for Approximate Bayesian Computation (ABC)." *Methods in Ecology and Evolution* 3, no. 3: 475–479.

Dannemann, M., and J. Kelso. 2017. "The Contribution of Neanderthals to Phenotypic Variation in Modern Humans." *American Journal of Human Genetics* 101: 578–579.

de Villemereuil, P., and O. E. Gaggiotti. 2015. "A New FST-Based Method to Uncover Local Adaptation Using Environmental Variables." *Methods in Ecology and Evolution* 6, no. 11: 1248–1258.

DeGiorgio, M., C. D. Huber, M. J. Hubisz, I. Hellmann, and R. Nielsen. 2016. "SweepFinder2: Increased Sensitivity, Robustness and Flexibility." *Bioinformatics* 32, no. 12: 1895–1897. https://doi.org/10.1093/bioinformatics/btw051.

DeGiorgio, M., K. E. Lohmueller, and R. Nielsen. 2014. "A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data." *PLoS Genetics* 10, no. 8: e1004561.

DeGiorgio, M., and Z. A. Szpiech. 2022. "A Spatially Aware Likelihood Test to Detect Sweeps From Haplotype Distributions." *PLoS Genetics* 18, no. 4: e1010134.

Dellaportas, P., and G. O. Roberts. 2003. "An Introduction to MCMC." In *Spatial Statistics and Computational Methods*, 1–41. Springer New York.

Dias-Alves, T., J. Mairal, and M. G. B. Blum. 2018. "Loter: A Software Package to Infer Local Ancestry for a Wide Range of Species." *Molecular Biology and Evolution* 35, no. 9: 2318–2326.

Drummond, A. J., and A. Rambaut. 2007. "BEAST: Bayesian Evolutionary Analysis by Sampling Trees." *BMC Evolutionary Biology* 7: 1–8.

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus. 2005. "Bayesian Coalescent Inference of Past Population Dynamics From Molecular Sequences." *Molecular Biology and Evolution* 22, no. 5: 1185–1192.

Drummond, A. J., W. Xie, and J. Heled. 2012. "Bayesian Inference of Species Trees From Multilocus Data Using* BEAST." *Molecular Biology and Evolution* 29: 1969–1973.

Dumoulin, N., F. Jabot, and T. Faure. 2023. "EasyABC: Efficient Approximate Bayesian Computation Sampling Schemes" (Doctoral diss., INRAE).

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. "Testing for Ancient Admixture Between Closely Related Populations." *Molecular Biology and Evolution* 28, no. 8: 2239–2252.

Durvasula, A., and S. Sankararaman. 2019. "A Statistical Model for Reference-Free Inference of Archaic Local Ancestry." *PLoS Genetics* 15: e1008175.

Durvasula, A., and S. Sankararaman. 2020. "Recovering Signals of Ghost Archaic Introgression in African Populations." *Science Advances* 6: eaax5097.

Excoffier, L., and M. Foll. 2011. "Fastsimcoal: A Continuous-Time Coalescent Simulator of Genomic Diversity Under Arbitrarily Complex Evolutionary Scenarios." *Bioinformatics* 27, no. 9: 1332–1334.

Excoffier, L., N. Marchi, D. A. Marques, R. Matthey-Doret, A. Gouy, and V. C. Sousa. 2021. "Fastsimcoal2: Demographic Inference Under Complex Evolutionary Scenarios." *Bioinformatics* 37, no. 24: 4882–4885.

Excoffier, L., J. Novembre, and S. Schneider. 2000. "Computer Note. SIMCOAL: A General Coalescent Program for the Simulation of Molecular Data in Interconnected Populations With Arbitrary Demography." *Journal of Heredity* 91, no. 6: 506–509.

Fariello, M. I., S. Boitard, H. Naya, M. SanCristobal, and B. Servin. 2013. "Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations." *Genetics* 193, no. 3: 929–941.

Field, Y., E. A. Boyle, N. Telis, et al. 2016. "Detection of Human Adaptation During the Past 2000 Years." *Science* 354: 760–764.

Fijarczyk, A., and W. Babik. 2015. "Detecting Balancing Selection in Genomes: Limits and Prospects." *Molecular Ecology* 24, no. 14: 3529–3545.

Fine, A. G., and M. Steinrücken. 2024. "A Novel Expectation-Maximization Approach to Infer General Diploid Selection From Time-Series Genetic Data." *bioRxiv*.

Finke, K., M. Kourakos, G. Brown, et al. 2021. "Ancestral Haplotype Reconstruction in Endogamous Populations Using Identity-By-Descent." *PLoS Computational Biology* 17: e1008638.

Fisher, R. A. 1999. *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. Oxford University Press.

Flagel, L., Y. Brandvain, and D. R. Schrider. 2019. "The Unreasonable Effectiveness of Convolutional Neutral Networks in Population Genetic Inference." *Molecular Biology and Evolution* 36, no. 2: 220–238.

Frazier, D. T., C. P. Robert, and J. Rousseau. 2020. "Model Misspecification in Approximate Bayesian Computation: Consequences and Diagnostics." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 82, no. 2: 421–444.

Fu, Y. X. 1997. "Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection." *Genetics* 147, no. 2: 915–925.

Fu, Y. X., and W. H. Li. 1993. "Statistical Tests of Neutrality of Mutations." *Genetics* 133, no. 3: 693–709.

GenomeAsia 100K Consortium. 2019. "The GenomeAsia 100K Project Enables Genetic Discoveries Across Asia." *Nature* 576: 106–111.

Gittelman, R. M., J. G. Schraiber, B. Vernot, C. Mikacenic, M. M. Wurfel, and J. M. Akey. 2016. "Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments." *Current Biology* 26, no. 24: 3375–3382.

Gower, G., P. I. Picazo, M. Fumagalli, and F. Racimo. 2021. "Detecting Adaptive Introgression in Human Evolution Using Convolutional Neural Networks." *eLife* 10: e64669.

Green, R. E., J. Krause, A. W. Briggs, et al. 2010. “A Draft Sequence of the Neandertal Genome.” *Science* 328, no. 5979: 710–722.

Guillot, G., R. Vitalis, A. le Rouzic, and M. Gautier. 2014. “Detecting Correlation Between Allele Frequencies and Environmental Variables as a Signature of Selection. A Fast Computational Approach for Genome-Wide Studies.” *Spatial Statistics* 8: 145–155.

Haldane, J. B. S. 1919. “The Combination of Linkage Values, and the Calculation of Distances Between the Loci of Linked Factors.” *Journal of Genetics* 8: 299–309.

Haller, B. C., and P. W. Messer. 2023. “SLiM 4: Multispecies Eco-Evolutionary Modeling.” *American Naturalist* 201, no. 5: E127–E139.

Harris, K., and R. Nielsen. 2016. “The Genetic Cost of Neanderthal Introgression.” *Genetics* 203: 881–891.

Hellenthal, G., G. B. Busby, G. Band, et al. 2014. “A Genetic Atlas of Human Admixture History.” *Science* 343, no. 6172: 747–751.

Hernandez, M., and G. H. Perry. 2021. “Scanning the Human Genome for ‘Signatures’ of Positive Selection: Transformative Opportunities and Ethical Obligations.” *Evolutionary Anthropology* 30, no. 2: 113–121.

Huber, C. D., M. DeGiorgio, I. Hellmann, and R. Nielsen. 2015. “Detecting Recent Selective Sweeps While Controlling for Mutation Rate and Background Selection.” *Molecular Ecology* 25, no. 1: 142–156.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard. 2009. “Inferring Weak Population Structure With the Assistance of Sample Group Information.” *Molecular Ecology Resources* 9, no. 5: 1322–1332.

Hubisz, M. J., A. L. Williams, and A. Siepel. 2020. “Mapping Gene Flow Between Ancient Hominins Through Demography-Aware Inference of the Ancestral Recombination Graph.” *PLoS Genetics* 16, no. 8: e1008895.

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. “A Test of Neutral Molecular Evolution Based on Nucleotide Data.” *Genetics* 116: 153–159.

Huerta-Sanchez, E., X. Jin, Asan, et al. 2014. “Altitude Adaptation in Tibetans Caused by Introgression of Denisovan-Like DNA.” *Nature* 512, no. 7513: 194–197.

Huff, C. D., H. C. Harpending, and A. R. Rogers. 2010. “Detecting Positive Selection From Genome Scans of Linkage Disequilibrium.” *BMC Genomics* 11, no. 8: 1–9.

Illumina. 2024. “Novaseq X and Novaseq X Plus Sequencing Systems Specification Sheet.” Accessed October 14 2024. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-x-series-spec-sheet-m-us-00197/novaseq-x-series-specification-sheet-m-us-00197.pdf.

Jain, A., R. C. Bhoyar, K. Pandhare, et al. 2021. “IndiGenomes: A Comprehensive Resource of Genetic Variants From Over 1000 Indian Genomes.” *Nucleic Acids Research* 49: D1225–D1232.

Jiang, B., T. Y. Wu, C. Zheng, and W. H. Wong. 2017. “Learning Summary Statistic for Approximate Bayesian Computation via Deep Neural Network.” *Statistica Sinica* 27: 1595–1618.

Jordan, M. I., and T. M. Mitchell. 2015. “Machine Learning: Trends, Perspectives, and Prospects.” *Science* 349, no. 6245: 255–260.

Kandler, A., and A. Powell. 2018. “Generative Inference for Cultural Evolution.” *Philosophical Transactions of the Royal Society, B: Biological Sciences* 373, no. 1743: 20170056.

Kelleher, J., A. M. Etheridge, and G. McVean. 2016. “Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes.” *PLoS Computational Biology* 12, no. 5: e1004842.

Kelleher, J., Y. Wong, A. W. Wohns, C. Fadil, P. K. Albers, and G. McVean. 2019. “Inferring Whole-Genome Histories in Large Population Datasets.” *Nature Genetics* 51, no. 9: 1330–1338.

Kerdoncuff, E., L. Skov, N. Patterson, et al. 2024. “50,000 Years of Evolutionary History of India: Insights From ∼2,700 Whole Genome Sequences.” *biorXiv*. https://doi.org/10.1101/2024.02.15.580575.

Kern, A. D., and D. R. Schrider. 2018. “diploS/HIC: An Updated Approach to Classifying Selective Sweeps.” *G3: Genes, Genomes, Genetics* 8, no. 6: 1959–1970.

Kim, Y., and W. Stephan. 2002. “Detecting a Local Signature of Genetic Hitchhiking Along a Recombining Chromosome.” *Genetics* 160, no. 2: 765–777.

Klassmann, A., and M. Gautier. 2022. “Detecting Selection Using Extended Haplotype Homozygosity (EHH-Based Statistics in Unphased or Unpolarized Data).” *PLoS One* 17, no. 1: e0262024.

Klunk, J., T. P. Vilgalys, C. E. Demeure, et al. 2022. “Evolution of Immune Genes Is Associated With the Black Death.” *Nature* 611, no. 7935: 312–319.

Koller, D., F. R. Wendt, G. A. Pathak, et al. 2022. “Denisovan and Neanderthal Archaic Introgression Differentially Impacted the Genetics of Complex Traits in Modern Populations.” *BMC Biology* 20: 249.

Kosakovsky Pond, S. L., A. F. Poon, R. Velazquez, et al. 2020. “HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies.” *Molecular Biology and Evolution* 37, no. 1: 295–299.

Krogh, A. 1998. “An Introduction to Hidden Markov Models for Biological Sequences.” In *New Comprehensive Biochemistry*, vol. 32, 45–63. Elsevier.

Kuhlwilm, M., I. Gronau, M. Hubisz, et al. 2016. “Ancient Gene Flow From Early Modern Humans Into Eastern Neanderthals.” *Nature* 530: 429–433.

Kumar, S., G. Stecher, and K. Tamura. 2016. “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.” *Molecular Biology and Evolution* 33, no. 7: 1870–1874.

Larena, M., J. McKenna, F. Sanchez-Quinto, et al. 2021. “Philippine Ayta Possess the Highest Level of Denisovan Ancestry in the World.” *Current Biology* 31, no. 19: 4219–4230.

Laval, G., and L. Excoffier. 2004. “SIMCOAL 2.0: A Program to Simulate Genomic Diversity Over Large Recombining Regions in a Subdivided Population With a Complex History.” *Bioinformatics* 20, no. 15: 2485–2487.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush. 2012. “Inference of Population Structure Using Dense Haplotype Data.” *PLoS Genetics* 8, no. 1: e1002453.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. “Deep Learning.” *Nature* 521: 436–444.

Lee, J., A. B. Dey, P. Khobragade, et al. 2019. “Harmonized Diagnostic Assessment of Dementia for the Longitudinal Aging Study in India (LASI-DAD) Wave 1 Version A Data.” Produced and distributed by the University of Southern California with funding from the National Institute on Aging (R01AG051125, RF1AG055273, U01AG065958). https://doi.org/10.25549/h5wx-ay45.

Lee, K. M., and G. Coop. 2017. “Distinguishing Among Modes of Convergent Adaptation Using Population Genomic Data.” *Genetics* 207, no. 4: 1591–1619.

Lewanski, A. L., M. C. Grundler, and G. S. Bradburd. 2024. “The Era of the ARG: An Introduction to Ancestral Recombination Graphs and Their Significance in Empirical Evolutionary Genomics.” *PLoS Genetics* 20, no. 1: e1011110.

Lewontin, R. C., and J. Krakauer. 1973. “Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms.” *Genetics* 74, no. 1: 175–195.

Li, H., and R. Durbin. 2011. “Inference of Human Population History From Individual Whole-Genome Sequences.” *Nature* 475, no. 7357: 493–496.

Li, N., and M. Stephens. 2003. “Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data.” *Genetics* 165, no. 4: 2213–2233.

Liao, W. W., M. Asri, J. Ebler, et al. 2023. "A Draft Human Pangenome Reference." *Nature* 617, no. 7960: 312–324.

Libbrecht, M. W., and W. S. Noble. 2015. "Machine Learning Applications in Genetics and Genomics." *Nature Reviews Genetics* 16, no. 6: 321–332.

Librado, P., and L. Orlando. 2018. "Detecting Signatures of Positive Selection Along Defined Branches of a Population Tree Using LSD." *Molecular Biology and Evolution* 35, no. 6: 1520–1535.

Lintusaari, J., M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander. 2017. "Fundamentals and Recent Developments in Approximate Bayesian Computation." *Systematic Biology* 66, no. 1: e66–e82.

Liu, X., R. T.-H. Ong, E. N. Pillai, et al. 2013. "Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations." *American Journal of Human Genetics* 92, no. 6: 866–881.

Loh, P. R., M. Lipson, N. Patterson, et al. 2013. "Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium." *Genetics* 193, no. 4: 1233–1254.

Lopez Fang, L., D. Peede, D. Ortega-Del Vecchyo, E. J. McTavish, and E. Huerta-Sánchez. 2024. "Leveraging Shared Ancestral Variation to Detect Local Introgression." *PLoS Genetics* 20, no. 1: e1010155.

Lowry, D. B., S. Hoban, J. L. Kelley, et al. 2017. "Breaking RAD: An Evaluation of the Utility of Restriction Site-Associated DNA Sequencing for Genome Scans of Adaptation."

Mafessoni, F., S. Grote, C. de Filippo, et al. 2020. "A High-Coverage Neandertal Genome From Chagyrskaya Cave." *Proceedings of the National Academy of Sciences of the United States of America* 117, no. 26: 15132–15136.

Maier, R., P. Flegontov, O. Flegontova, U. Isildak, P. Changmai, and D. Reich. 2023. "On the Limits of Fitting Complex Models of Population History to f-Statistics." *eLife* 12: e85492.

Malinsky, M., M. Matschiner, and H. Svardal. 2021. "Dsuite-Fast D-Statistics and Related Admixture Evidence From VCF Files." *Molecular Ecology Resources* 21, no. 2: 584–595.

Mallick, S., H. Li, M. Lipson, et al. 2016. "The Simons Genome Diversity Project: 300 Genomes From 142 Diverse Populations." *Nature* 538, no. 7624: 201–206. https://doi.org/10.1038/nature18964.

Mallick, S., A. Micco, M. Mah, et al. 2024. "The Allen Ancient DNA Resource (AADR) a Curated Compendium of Ancient Human Genomes." *Scientific Data* 11, no. 1: 182.

Manel, S., C. Perrier, M. Pratlong, et al. 2015. "Genomic Resources and Their Influence on the Detection of the Signal of Positive Selection in Genome Scans." *Molecular Ecology* 25, no. 1: 170–184.

Maples, B. K., S. Gravel, E. E. Kenny, and C. D. Bustamante. 2013. "RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference." *American Journal of Human Genetics* 93, no. 2: 278–288.

Martin, A. R., C. R. Gignoux, R. K. Walters, G. L. Wojcik, B. M. Neale, and E. E. Kenny. 2017. "Human Demographic History Impacts Genetic Risk Prediction Across Diverse Populations." *American Journal of Human Genetics* 100: 635–649.

Mather, N., S. M. Traves, and S. Y. W. Ho. 2020. "A Practical Introduction to Sequentially Markovian Coalescent Methods for Estimating Demographic History From Genomic Data." *Ecology and Evolution* 10, no. 1: 579–589.

McDonald, J. H., and M. Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351: 652–654.

McVicker, G., D. Gordon, C. Davis, and P. Green. 2009. "Widespread Genomic Signatures of Natural Selection in Hominid Evolution." *PLoS Genetics* 5, no. 5: e1000471.

Messer, P. W. 2013. "SLiM: Simulating Evolution With Selection and Linkage." *Genetics* 194, no. 4: 1037–1039.

Meyer, M., M. Kircher, M.-T. Gansauge, et al. 2012. "A High-Coverage Genome Sequence From an Archaic Denisovan Individual." *Science* 338, no. 6104: 222–226.

Miró-Herrans, A. T., and C. J. Mulligan. 2013. "Human Demographic Processes and Genetic Variation as Revealed by mtDNA Simulations." *Molecular Biology and Evolution* 30, no. 2: 244–252. https://doi.org/10.1093/molbev/mss230.

Mondal, M., J. Bertranpetit, and O. Lao. 2019. "Approximate Bayesian Computation With Deep Learning Supports a Third Archaic Introgression in Asia and Oceania." *Nature Communications* 10, no. 1: 246.

Montinaro, F., V. Pankratov, B. Yelmen, L. Pagani, and M. Mondal. 2021. "Revisiting the out of Africa Event With a Deep-Learning Approach." *American Journal of Human Genetics* 108, no. 11: 2037–2051.

Montserrat, D. M., C. Bustamante, and A. Ioannidis. 2020. "LAI-Net: Local-Ancestry Inference With Neural Networks." *arXiv*. https://doi.org/10.48550/arXiv.2004.10377.

Moorjani, P., N. Patterson, J. N. Hirschhorn, et al. 2011. "The History of African Gene Flow Into Southern Europeans, Levantines, and Jews." *PLoS Genetics* 7, no. 4: e1001373.

Moorjani, P., N. Patterson, P.-R. Loh, M. Lipson, P. Kisfali, and B. Melegh. 2013. "Reconstructing Roma History From Genome-Wide Data." *PLoS One* 8, no. 3: e58633.

Mullin, E. 2022. "The Era of Fast, Cheap Genome Sequencing Is Here." *WIRED*. https://www.wired.com/story/the-era-of-fast-cheap-genome-sequencing-is-here/.

Nielsen, R., M. J. Hubisz, I. Hellmann, et al. 2009. "Darwinian and Demographic Forces Affecting Human Protein Coding Genes." *Genome Research* 19: 838–849.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. "Genomic Scans for Selective Sweeps Using SNP Data." *Genome Research* 15: 1566–1575.

Novakovsky, G., N. Dexter, M. W. Libbrecht, W. W. Wasserman, and S. Mostafavi. 2022. "Obtaining Genetic Insights From Deep Learning via Explainable Artificial Intelligence." *Nature Reviews Genetics* 24: 125–137.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, and C. Bustamante. 2008. "Genes Mirror Geography Within Europe." *Nature* 456: 98–101.

Oleksyk, T. K., K. Zhao, F. M. De La Vega, D. A. Gilbert, S. J. O'Brien, and M. W. Smith. 2008. "Identifying Selected Regions From Heterozygosity and Divergence Using a Light-Coverage Genomic Dataset From Two Human Populations." *PLoS One* 3, no. 3: e1712.

Orlando, L., R. Allaby, P. Skoglund, et al. 2021. "Ancient DNA Analysis." *Nature Reviews Methods Primers* 1, no. 1: 14.

Pagani, L., D. J. Lawson, E. Jagoda, et al. 2016. "Genomic Analyses Inform on Migration Events During the Peopling of Eurasia." *Nature* 538, no. 7624: 238–242.

Patterson, N., P. Moorjani, Y. Luo, et al. 2012. "Ancient Admixture in Human History." *Genetics* 192, no. 3: 1065–1093.

Peter, B. 2020. "100,000 Years of Gene Flow Between Neanderthals and Denisovans in the Altai Mountains." https://www.biorxiv.org/content/10.1101/2020.03.13.990523v1.

Peter, B. M. 2016. "Admixture, Population Structure, and F-Statistics." *Genetics* 202, no. 4: 1485–1501.

Petr, M., S. Paabo, J. Kelso, and B. Vernot. 2019. "Limits of Long-Term Selection Against Neandertal Introgression." *Proceedings of the National Academy of Sciences of the United States of America* 116: 1639–1644.

Petr, M., B. Vernot, and J. Kelso. 2019. "Admixr—R Package for Reproducible Analyses Using ADMIXTOOLS." *Bioinformatics* 35, no. 17: 3194–3195.

Pickrell, J. K., G. Coop, J. Novembre, et al. 2009. “Signals of Recent Positive Selection in a Worldwide Sample of Human Populations.” *Genome Research* 19: 826–837.

Pickrell, J. K., and J. K. Pritchard. 2012. “Inference of Population Splits and Mixtures From Genome-Wide Allele Frequency Data.” *PLoS Genetics* 8, no. 11: e1002967.

Popejoy, A. B., and S. M. Fullerton. 2016. “Genomics Is Failing on Diversity.” *Nature* 538: 161–164.

Prufer, K., C. de Filippo, S. Grote, et al. 2017. “A High-Coverage Neandertal Genome From Vindija Cave in Croatia.” *Science* 1887, no. 6363: 655–658.

Prufer, K., F. Racimo, N. Patterson, et al. 2014. “The Complete Genome Sequence of a Neanderthal From the Altai Mountains.” *Nature* 505, no. 7481: 43–49.

Prüfer, K., U. Stenzel, M. Hofreiter, S. Pääbo, J. Kelso, and R. E. Green. 2010. “Computational Challenges in the Analysis of Ancient DNA.” *Genome Biology* 11: 1–15.

Pudlo, P., J. M. Marin, A. Estoup, J. M. Cornuet, M. Gautier, and C. P. Robert. 2016. “Reliable ABC Model Choice via Random Forests.” *Bioinformatics* 32, no. 6: 859–866.

Racimo, F. 2016. “Testing for Ancient Selection Using Cross-Population Allele Frequency Differentiation.” *Genetics* 202, no. 2: 733–750.

Racimo, F., S. Sankararaman, R. Nielsen, and E. Huerta-Sanchez. 2015. “Evidence for Archaic Adaptive Introgression in Humans.” *Nature Reviews Genetics* 16: 359–371.

Ragsdale, A. P., T. D. Weaver, E. G. Atkinson, et al. 2023. “A Weakly Structured Stem for Human Origins in Africa.” *Nature* 617: 755–763.

Raj, A., M. Stephens, and J. K. Pritchard. 2014. “fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets.” *Genetics* 197, no. 2: 573–589.

Rasmussen, M. D., M. J. Hubisz, I. Gronau, and A. Siepel. 2014. “Genome-Wide Inference of Ancestral Recombination Graphs.” *PLoS Genetics* 10, no. 5: e1004342.

Reich, D., K. Thangaraj, N. Patterson, A. L. Price, and L. Singh. 2009. “Reconstructing Indian Population History.” *Nature* 461, no. 7263: 489–494.

Riebler, A., L. Held, and E. Stephan. 2008. “Bayesian Variable Selection for Detecting Adaptive Genomic Differences Among Populations.” *Genetics* 178, no. 3: 1817–1829.

Riley, R., I. Mathieson, and S. Mathieson. 2024. “Interpreting Generative Adversarial Networks to Infer Natural Selection From Genetic Data.” *Genetics* 226, no. 4: iyae024.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, et al. 2002. “Genetic Structure of Human Populations.” *Science* 298, no. 5602: 2381–2385.

Ruwende, C., S. C. Khoo, R. W. Snow, et al. 1995. “Natural Selection of Hemi- and Heterozygotes for G6PD Deficiency in Africa by Resistance to Severe Malaria.” *Nature* 376, no. 6537: 246–249. https://doi.org/10.1038/376246a0.

Sabeti, P. C., D. E. Reich, J. M. Higgins, et al. 2002. “Detecting Recent Positive Selection in the Human Genome From Haplotype Structure.” *Nature* 419: 832–837.

Sabeti, P. C., P. Varilly, B. Fry, et al. 2007. “Genome-Wide Detection and Characterization of Positive Selection in Human Populations.” *Nature* 449: 913–918.

Salter-Townshend, M., and S. Myers. 2019. “Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups.” *Genetics* 212, no. 3: 869–889.

Sanchez, T., J. Cury, G. Charpiat, and F. Jay. 2021. “Deep Learning for Population Size History Inference: Design, Comparison and Combination With Approximate Bayesian Computation.” *Molecular Ecology Resources* 21, no. 8: 2645–2660.

Sandoval-Castellanos, E., E. Palkopoulou, and L. Dalen. 2014. “Back to BaySICS: A User-Friendly Program for Bayesian Statistical Inference From Coalescent Simulations.” *PLoS One* 9, no. 5: e98011.

Sankararaman, S., S. Mallick, M. Dannemann, et al. 2014. “The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans.” *Nature* 507, no. 7492: 354–357.

Sankararaman, S., S. Mallick, N. Patterson, and D. Reich. 2016. “The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans.” *Current Biology* 26, no. 9: 1241–1247.

Schälte, Y., E. Klinger, E. Alamoudi, and J. Hasenauer. 2022. “pyABC: Efficient and Robust Easy-To-Use Approximate Bayesian Computation.” *arXiv preprint arXiv:2203.13043*.

Schiffels, S., and R. Durbin. 2014. “Inferring Human Population Size and Separation History From Multiple Genome Sequences.” *Nature Genetics* 46, no. 8: 919–925.

Schraiber, J. G., S. N. Evans, and M. Slatkin. 2016. “Bayesian Inference of Natural Selection From Allele Frequency Time Series.” *Genetics* 203, no. 1: 493–511.

Schrider, D. R., and A. Kern. 2018. “Supervised Machine Learning for Population Genetics: A New Paradigm.” *Trends in Genetics* 34: 301–312.

Schubert, R., A. Andaleon, and H. E. Wheelber. 2020. “Comparing Local Ancestry Inference Models in Populations of Two- and Three-Way Admixture.” *PeerJ* 8: e10090.

Secolin, R., A. Mas-Sandoval, L. R. Arauna, F. R. Torres, T. K. Araujo, and D. Comas. 2019. “Distribution of Local Ancestry and Evidence of Adaptation in Admixed Populations.” *Scientific Reports* 9: 13900.

Setter, D., S. Mousset, X. Cheng, R. Nielsen, M. DeGiorgio, and J. Hermisson. 2020. “VolcanoFinder: Genomic Scans for Adaptive Introgression.” *PLoS Genetics* 16, no. 6: e1008867.

Sheehan, S., K. Harris, and Y. S. Song. 2013. “Estimating Variable Effective Population Sizes From Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach.” *Genetics* 194, no. 3: 647–662.

Sheehan, S., and Y. S. Song. 2016. “Deep Learning for Population Genetic Inference.” *PLoS Computational Biology* 12, no. 3: e1004845.

Shriver, M. D., G. C. Kennedy, E. J. Parra, et al. 2004. “The Genomic Distribution of Population Substructure in Four Populations Using 8,525 Autosomal SNPs.” *Human Genomics* 1, no. 4: 274–286.

Siewert, K. M., and B. F. Voight. 2017. “Detecting Long-Term Balancing Selection Using Allele Frequency Correlation.” *Molecular Biology and Evolution* 34, no. 11: 2996–3005.

Siewert, K. M., and B. F. Voight. 2020. “BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data.” *Genome Biology and Evolution* 12, no. 2: 3873–3877.

Skov, L., M. Coll Macià, G. Sveinbjörnsson, F. Mafessoni, E. A. Lucotte, and K. Stefansson. 2020. “The Nature of Neanderthal Introgression Revealed by 27,566 Icelandic Genomes.” *Nature* 582: 78–83.

Skov, L., R. Hui, V. Shchur, et al. 2018. “Detecting Archaic Introgression Using an Unadmixed Outgroup.” *PLoS Genetics* 14, no. 9: e1007641.

Slatkin, M., and F. Racimo. 2016. “Ancient DNA and Human History.” *Proceedings of the National Academy of Sciences* 113, no. 23: 6380–6387.

Slon, V., F. Mafessoni, B. Vernot, et al. 2018. “The Genome of the Offspring of a Neanderthal Mother and a Denisovan Father.” *Nature* 561: 113–116.

Smith, C. C., S. Tittes, P. L. Ralph, and A. D. Kern. 2023. “Dispersal Inference From Population Genetic Variation Using a Convolutional Neural Network.” *Genetics* 224, no. 2: iyad068.

Smith, J. M., and J. Haigh. 1974. “The Hitch-Hiking Effect of a Favorable Gene.” *Genetical Research* 89: 391–403.

Sohail, M., M. J. Palma-Martinez, A. Y. Chong, et al. 2023. “Mexican Biobank Advances Population and Medical Genomics of Diverse Ancestries.” *Nature* 622: 775–783.

Sohail, M. 2022. “Investigating Relative Contributions to Psychiatric Disease Architecture From Sequence Elements Originating Across Multiple Evolutionary Time-Scales.” *bioRxiv* (2022): 2022–02.

Soni, V., and J. D. Jensen. 2024. “Temporal Challenges in Detecting Balancing Selection From Population Genomic Data.” *G3: Genes, Genomes, Genetics* 14: jkae069.

Speidel, L., M. Forest, S. Shi, and S. R. Myers. 2019. “A Method for Genome-Wide Genealogy Estimation for Thousands of Samples.” *Nature Genetics* 51, no. 9: 1321–1329.

Spence, J. P., M. Steinrücken, J. Terhorst, and Y. S. Song. 2018. “Inference of Population History Using Coalescent HMMs: Review and Outlook.” *Current Opinion in Genetics and Development* 53: 70–76.

Steinrücken, M., J. Kamm, J. P. Spence, and Y. S. Song. 2019. “Inference of Complex Population Histories Using Whole-Genome Sequences From Multiple Populations.” *Proceedings of the National Academy of Sciences* 116, no. 34: 17115–17120.

Steinrücken, M., J. P. Spence, J. A. Kamm, E. Wieczorek, and Y. S. Song. 2018. “Model-Based Detection and Analysis of Introgressed Neanderthal Ancestry in Modern Humans.” *Molecular Ecology* 27, no. 19: 3873–3888.

Szpiech, Z. A. 2021. “Selscan 2.0: Scanning for Sweeps in Unphased Data.” *bioRxiv*. https://doi.org/10.1101/2021.10.22.465497.

Tajima, F. 1989. “Statistical Methods for Testing the Neutral Mutation Hypothesis by DNA Polymorphism.” *Genetics* 123, no. 3: 585–595.

Tajima, F. 1993. “Simple Methods for Testing the Molecular Evolutionary Clock Hypothesis.” *Genetics* 135, no. 2: 599–607.

Teixeira, J. C., G. S. Jacobs, C. Stringer, et al. 2021. “Widespread Denisovan Ancestry in Island Southeast Asia but no Evidence of Substantial Super-Archaic Hominin Admixture.” *Nature Ecology and Evolution* 5: 616–624.

Terhorst, J., J. A. Kamm, and Y. S. Song. 2017. “Robust and Scalable Inference of Population History From Hundreds of Unphased Whole Genomes.” *Nature Genetics* 49, no. 2: 303–309.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, and P. Deloukas. 2007. “Convergent Adaptation of Human Lactase Persistence in Africa and Europe.” *Nature Genetics* 39: 31–40.

Torada, L., L. Lorenzon, A. Beddis, et al. 2019. “ImaGene: A Convolutional Neural Network to Quantify Natural Selection From Genomic Data.” *BMC Bioinformatics* 20, no. Suppl 9: 337.

Upadhya, G., and M. Steinrücken. 2022. “Robust Inference of Population Size Histories From Genomic Sequencing Data.” *PLoS Computational Biology* 18, no. 9: e1010419.

van der Vaart, E., M. A. Beaumont, A. S. Johnston, and R. M. Sibly. 2015. “Calibration and Evaluation of Individual-Based Models Using Approximate Bayesian Computation.” *Ecological Modelling* 312: 182–190.

Vernot, B., and J. Akey. 2014. “Resurrecting Surviving Neandertal Lineages From Modern Human Genomes.” *Science* 343: 1017–1021.

Villanea, F. A., E. Huerta-Sanchez, and K. Fox. 2021. “ABO Genetic Variation in Neanderthals and Denisovans.” *Molecular Biology and Evolution* 38, no. 8: 3373–3382.

Villanea, F. A., A. Kitchen, and B. M. Kemp. 2020. “Applications of Bayesian Skyline Plots and Approximate Bayesian Computation for Human Demography.” *Human Biology* 91, no. 4: 279–296.

Villanea, F. A., D. Peede, E. J. Kaufman, et al. 2023. “The MUC19 Gene in Denisovans: Neanderthals, and Modern Humans: An Evolutionary History of Recurrent Introgression and Natural Selection.” *bioRxiv*. https://doi.org/10.1101/2023.09.25.559202.

Villanea, F. A., and J. G. Schraiber. 2019. “Multiple Episodes of Interbreeding Between Neanderthal and Modern Humans.” *Nature Ecology and Evolution* 3: 39–44.

Vitti, J. J., S. R. Grossman, and P. C. Sabeti. 2013. “Detecting Natural Selection in Genomic Data.” *Annual Review of Genetics* 47: 97–120.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. “A Map of Recent Positive Selection in the Human Genome.” *PLoS Biology* 4, no. 3: e72.

Vy, H. M. T., and Y. Kim. 2015. “A Composite-Likelihood Method for Detecting Incomplete Selective Sweep From Population Genomic Data.” *Genetics* 200, no. 2: 633–649.

Wang, Z., J. Wang, M. Kourakos, et al. 2021. “Automatic Inference of Demographic Parameters Using Generative Adversarial Networks.” *Molecular Ecology Resources* 21, no. 8: 2689–2705.

Watkins, W. S., J. E. Feusier, J. Thomas, C. Goubert, S. Mallick, and L. B. Jorde. 2020. “The Simons Genome Diversity Project: A Global Analysis of Mobile Element Diversity.” *Genome Biology and Evolution* 12, no. 6: 779–794. https://doi.org/10.1093/gbe/evaa086.

Wei, Y., D. Zhi, and S. Zhang. 2023. “Fast and Accurate Local Ancestry Inference With Recomb-Mix.” *bioRxiv*. https://doi.org/10.1101/2023.11.17.567650.

Weir, B. S., and C. C. Cockerham. 1984. “Estimating F-Statistics for the Analysis of Population Structure.” *Evolution* 38: 1358–1370.

Witt, K. E., A. Funk, V. Anorve-Garibay, L. Lopez Fang, and E. Huerta-Sanchez. 2023. “The Impact of Modern Admixture on Archaic Human Ancestry in Human Populations.” *Genome Biology and Evolution* 15, no. 5: evad066.

Witt, K. E., F. Villanea, E. Loughran, X. Zhang, and E. Huerta-Sanchez. 2022. “Apportioning Archaic Variants Among Modern Populations.” *Philosophical Transactions of the Royal Society B* 377: 20200411.

Wright and Fisher. 1929. *Wright x Fisher—Corrections Back and Forth, 1929–30, Part II*, 8–9. Sewall Wright Papers, American Philosophical Society.

Wright, S. 1931. “Evolution in Mendelian Populations.” *Genetics* 16, no. 2: 97–159.

Wright, S. 1949. “The Genetical Structure of Populations.” *Annals of Human Genetics* 15: 323–354.

Wroblewski, T. H., K. E. Witt, S.-B. Lee, et al. 2023. “Pharmacogenetic Variation in Neanderthals and Denisovans and Implications for Human Health and Response to Medications.” *Genome Biology and Evolution* 15, no. 12: evad222.

Xu, C., and S. A. Jackson. 2019. “Machine Learning and Complex Biological Data.” *Genome Biology* 20: 76.

Yang, Z. 2007. “PAML 4: Phylogenetic Analysis by Maximum Likelihood.” *Molecular Biology and Evolution* 24, no. 8: 1586–1591.

Yi, X., E. Huerta-Sanchez, X. Jin, et al. 2010. “Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude.” *Science* 329: 75–78.

Zhang, X., K. Witt, M. Banuelos, et al. 2021. “The History and Evolution of the Denisovan-EPAS1 Haplotype in Tibetans.” *Proceedings of the National Academy of Sciences of the United States of America* 118, no. 22: 1–9.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.