Deconvolving Human Evolutionary History: Using Network-Based Approaches to Better Understand Our Past

Shashwat Deepali Nagar, Andrew B Conley, and I King Jordan, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, United States; IHRC-Georgia Tech Applied Bioinformatics Laboratory, Atlanta, GA, United States; and PanAmerican Bioinformatics Institute, Cali, Colombia

Over the last several years, analyses of ever-accumulating human genome sequence data have made it possible to understand the relationships within and between populations. Evaluating population structure and historical demography can be approximated by a complex web of connections, i.e., a network. This is because, as human populations emerged from Africa, they experienced repeated periods of divergence when they were physically separated followed by convergence (or admixture) when they came back together. Admixture, once considered to be a fairly new aspect of human evolution (owing to the development of intercontinental travel), is increasingly recognized as a common feature that has repeatedly occurred throughout our evolution (Jeong *et al.*, 2019; Lipson *et al.*, 2018; Lorente-Galdos *et al.*, 2019; Reich *et al.*, 2009; Sikora *et al.*, 2019).

This reticulate model of evolution is in stark contrast with the strictly bifurcating perspective held by traditional evolutionary biologists who modeled evolutionary history as a tree—a framework that implicitly forbade the interaction of isolated and differentiated branches. Historically, evolution was visualized as a tree where branches became more and more differentiated over time. Early popularization of this interpretation can be seen from Ernst Haeckel's "Tree of life" where he describes universal common descent of all organisms (**Fig. 1A** and **B**). This view has now evolved and adapted as new data have become available following which the field is moving toward more complex schemes that can accommodate hybridization (**Fig. 1**C). Contrary to a tree model where a node can only have one parent, current evolutionary biologists are using different network-based frameworks to include recent and ancient admixture to assess population history.

Different network architectures can be used to answer specific questions about the histories of populations. For the purposes of this review, we will be focusing primarily on Directed Acyclic Graphs (DAGs) with a few examples of undirected graphs. DAGs are good proxies for modeling relationships between different populations as they can reliably represent relationships between different groups in a temporal fashion. As it is impossible for a population to interact with an ancestral population directly–they can only interact with descendants of an ancestral population in the same temporal space–these relationships can be represented faithfully using DAGs. **Fig. 2** shows the use of DAGs in representing population histories. **U** represents an ancestral population that splits into two populations groups **U′** and **U″** by time $\mathbf{T_1}$ as a result of isolation. By time $\mathbf{T_2}$, **U′** splits into two different populations **A′** and **B′**, while **U″** changes into **D′** simply as a result of the passage of time. During the same temporal slice, we see that **B′** and **D′** admix to form population **C′** in proportions α and $1 - \alpha$, respectively. At time $\mathbf{T_3}$, **C′** results in population **C**, while **A′** becomes **A**, an un-admixed community of **B′** becomes **B** and a part of **D′** becomes **D**. It should be noted that individuals from population **B′** could not have interacted with individuals from either **U′** or **U″** because they exist in different periods of time.

Networks and trees are currently used to answer the following questions:

1. Inferring population closeness—Evaluating which populations in a given set are more similar.
2. Inferring admixture—Assessing whether a particular population descended from multiple ancestral populations.
3. Inferring demography history—Exploring how different populations interact and split to form modern populations.
4. Inferring population structure—Identifying population membership and relationships based only on individual-level data.

The first two approaches in the list above are often inferred using population phylogenies or population trees. These population phylogenies are models where populations are assumed to be related in a tree-like fashion. The branch lengths in these trees corresponds to the extent of genetic differentiation. Three different tree-based statistics to evaluate population closeness were first described by Reich *et al.* (Reich *et al.*, 2009) and later refined and expanded upon by Patterson *et al.* (Patterson *et al.*, 2012). Collectively, these methods are called *f-statistics*. Several authors have commented and helped expand upon the interpretations of

doi:10.1016/B978-0-12-801238-3.11468-0

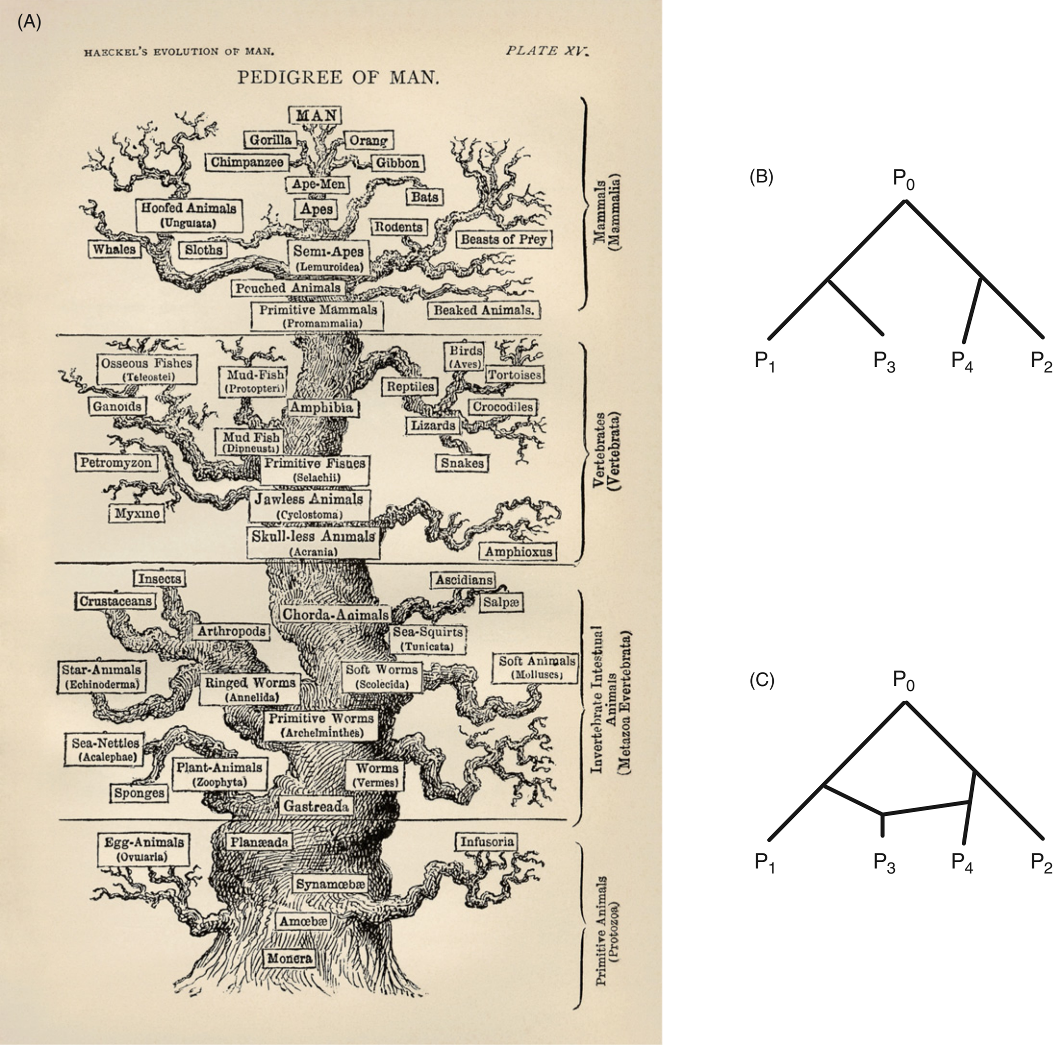

Fig. 1 Different views of evolution. (A) A traditional view of evolution which emphasizes a universal common descent for all living organisms (Ernst Haeckel). (B) A simple bifurcating tree. (C) A reticulate tree with admixture.

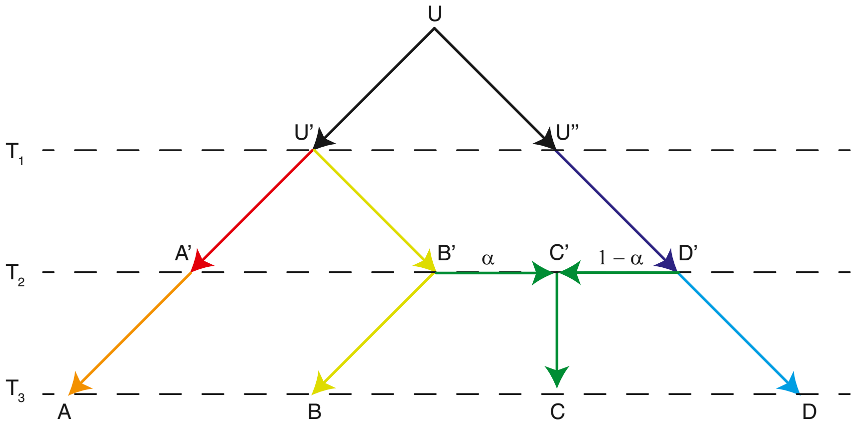

Fig. 2 Schematic showing population divergence and admixture.

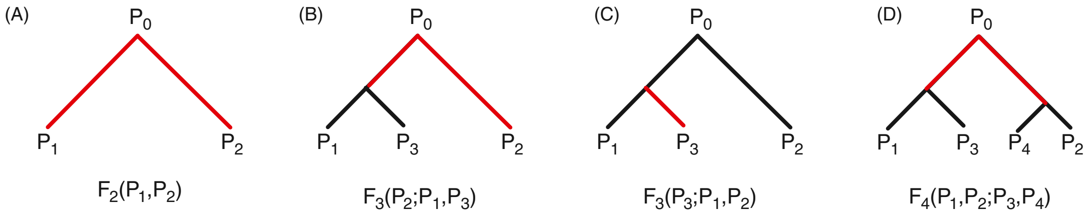

Fig. 3 Visualizing *f-statistics*. *f-statistics* can be visualized as branch lengths in population phylogenies. The relevant branch length is highlighted in *red*. (A) Shows a population phylogeny for F_2, (B) and (C) show population phylogenies for calculating F_3, and (D) represents a population phylogeny required for calculating F_4.

the statistics, most notably in (Peter, 2016; Harris and DeGiorgio, 2017). These statistics are calculated under population phylogeny models—which refers to a null model where populations are related in a tree-like fashion.

Inferring Population Closeness

In the original description by Reich *et al.* (Reich *et al.*, 2009), a distinction is made between empirical quantities called F_2, F_3, and F_4, while the overall statistics are denoted by f_2, f_3, and f_4, respectively. The empirical measurements (F_2, F_3, and F_4) can be understood as branch lengths for two, three, and four populations, respectively for a specific topology calculated for a single biallelic marker. These are then averaged over multiple markers to get different *f-statistics*. The branch lengths can be visualized as shown in **Fig. 3**.

Genetic Drift Calculation

The f_2 statistic can be used to measure the magnitude of genetic drift that has occurred between two populations. We can define genetic drift simply as a change in allele frequency of polymorphisms in the two populations being considered. For a given biallelic genetic marker, F_2 is defined as:

$$F_2(P_1, P_2) = F_2(p_1, p_2) = E\left[(p_1 - p_2)^2\right]$$

where, p_1 and p_2 are the allele frequencies of a single biallelic marker in populations P_1 and P_2, respectively. The final f_2 statistic can be computed by averaging the F_2 values at hundreds of thousands of biallelic polymorphisms across the genome. Since F_2 is interpreted as branch length, there is an additivity property implicit in the definition.

$$F_2(P_1, P_2) = F_2(P_1, P_0) + F_2(P_2, P_0)$$

In most modern population genomic analyses, several populations (on a scale of tens to hundreds) are tested to understand and infer closeness and admixture. A simple way of achieving this is to use the f_2 statistic as a score of dissimilarity. All pairwise f_2 statistics can be computed to obtain a dissimilarity matrix following which a best-fit tree can be obtained. However, the process of obtaining a best-fit tree can be challenging, especially as the number of populations and branches increases. To counter this, the F_3 and F_4 tests provide convenient alternatives that provide simpler tests of closeness that are restricted to trees of size 3 and 4 (for F_3 and F_4, respectively).

Identifying Related Populations

F_3 can be seen as a metric that evaluates the shared portion of the branch from population P_1 to P_3 and population P_2 and P_3 (**Fig. 3**B). Formally, F_3 is defined as:

$$F_3(P_3; P_1; P_2) = F_3(P_3; P_1; P_2) = E[(p_3 - p_1)(p_3 - p_2)]$$

where, p_1, p_2, and p_3 are the allele frequencies of a single biallelic marker in populations P_1, P_2, and P_3, respectively. F_3 can also be expressed in terms of F_2:

$$F_3(P_3; P_1; P_2) = \frac{1}{2}(F_2(P_3, P_1) + F_2(P_3, P_2) - F_2(P_1, P_2))$$

The equation above can be extended to interpret F_2 as any distance metric which can then be used in the different interpretations of F_3.

A simple application of f_3 is to identify the most closely related population from a given set for an unknown population P_X. This was demonstrated in (Raghavan *et al.*, 2014), where the authors used the length of the common branch to identify the most closely related extant population for a new sample discovered in Mal'ta, South-Central Siberia. As can be seen in **Fig. 3**C, if f_3 is calculated for all populations in a given set while keeping the outgroup (P_2) and unknown population (P_1) fixed while cycling through the test populations (P_3 in the figure), the largest f_3 value for F_3(*Outgroup*; *Unknown population*, *Test population*) will help identify the closest test population for the unknown population. This is also called "outgroup f_3." This statistic is used in recent literature to calculate shared genetic drift (Sikora *et al.*, 2019; Tambets *et al.*, 2018; Yang *et al.*, 2017). In a recent publication by Sikora *et al.* (2019), the relationship of three ancient samples and their genetic relatedness to modern-day populations. The statistic points toward a close relatedness between ancient Paleo-Siberians with present-day populations of the Itelmen, Koryaks, and Chukchis, along with Native Americans.

Inferring Admixture

The usual goal for employing f_3 is to test whether population P_3 is admixed. For this interpretation, the null hypothesis is that f_3 is a positive number–the underlying hypothesis being that the data have been generated from a tree that has positive edge lengths. If f_3 is negative, the null hypothesis is rejected and is seen as evidence for admixture. The null model is structured like **Fig. 3**C. This test is often called the 'three-population test' in literature. Several publications use f_3 statistics to support their claim of detecting admixture events (Brucato *et al.*, 2018; Haber *et al.*, 2019; Wang *et al.*, 2019). Recently, Haber *et al.* (2019) use the f_3 statistic to assert that present-day Lebanese Muslims were a result of admixture between Medieval Lebanese, African, and Central/East Asians.

A natural extension of the f-statistics we've seen so far is F_4, which can be formalized as:

$$F_4(P_1, P_2; P_3, P_4) = F_4(p_1, p_2; p_3, p_4) = E[(p_1 - p_2)(p_3 - p_4)]$$

where, p_1, p_2, p_3, and p_4 are the allele frequencies of a single biallelic marker in populations P_1, P_2, P_3, and P_4, respectively. Like F_3, F_4 can also be written in terms of F_2:

$$F_4(p_1, p_2; p_3, p_4) = \frac{1}{2}(F_2(P_1, P_4) + F_2(P_2, P_3) - F_2(P_1, P_3) - F_2(P_2, P_4))$$

The f_4 statistic can be used to evaluate if there is a shared path between 4 different populations, thereby testing whether an unrooted tree with two distinct clusters–(P_1, P_3) and (P_2, P_4) in **Fig. 3**D. If there is no overlap in the drift between members of the two clusters, the f_4 statistic will be 0 (as in the case of $F_4(P_1,P_3; P_2,P_4)$), indicating that they do not share a recent or significant population history.

The f_4 statistic is a powerful tool for inferring admixture, especially when coupled with the f_3 statistic. On identifying a significantly non-zero value for a given topology, common ancestry for the two clusters can be inferred without knowing the direction of admixture. Questions about the direction of admixture can be directly addressed by employing the f_3 statistic. Many recent publications have used the f_4 statistic to understand the demographic history of populations (Lorente-Galdos *et al.*, 2019; Haber *et al.*, 2019; Wang *et al.*, 2019).

Inferring Introgression

For one of the last tree-based methods covered here, we talk about Patterson's D-statistic or the ABBA-BABA test (referred to as the D-statistic going forward) first described in (Green *et al.*, 2010) to test introgression between modern humans and Neanderthals. This test is a powerful test for a deviation from a strict bifurcating evolutionary history.

Simply put, two closely related populations are expected to share almost the same amount of derived alleles with a third similar population, i.e., if populations P_1 and P_2 are closely related and P_3 is a similar population, then the number of derived alleles shared by P_1 and P_3 but not in P_2 and the number of derived alleles common to P_2 and P_3 but not P_1 should be approximately the same. However, if there was any introgression between P_3 and P_1 it would lead to an asymmetry in the shared derived alleles. This expectation forms the basis of the D-statistic. To get information about ancestral and derived alleles, a fourth population P4–which is a common outgroup to all three populations under consideration—is included. This is represented in **Fig. 4**. The test is formally defined as:

$$D = \frac{Count_{ABBA} - Count_{BABA}}{Count_{ABBA} + Count_{BABA}}$$

where ABBA refers to a condition where the derived allele is shared by populations P_2 and P_3, while BABA refers to a situation where the derived allele is shared by populations P_1 and P_3. The D-statistic can range from -1 in a case of extreme introgression

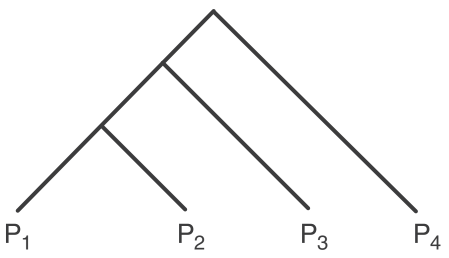

Fig. 4 Schematic for D-statistic calculation.

between P_2 and P_3 to 1 in case of introgression between P_1 and P_3. A value of 0 indicates no introgression. This simple yet powerful approach is often used to detect signatures of introgression (Racimo *et al.*, 2015; Mondal *et al.*, 2016; Feldman *et al.*, 2019). A recent publication by Feldman *et al*. (Feldman *et al.*, 2019), uses the D-statistic to understand the genetic affinity of Anatolian hunter gatherers with other European hunter-gatherer populations to find that the non-basal Eurasian ancestry of ancient Anatolians was derived from a gene pool related to a largely homogeneous gene pool of Later European hunter-gatherers referred to as the "Villabruna cluster."

It should be noted that even though the methods discussed above use tree-based phylogenies to model population histories, the test only work because of the reticulate and complex nature of human evolutionary history. The null hypothesis for each of these tests is that our evolutionary history is tree-like.

Inferring Demography History

From the different tests we've seen so far, it is safe to assert that population groups exchange genes by the process of admixture and simple bifurcating trees are an incorrect representation of the history of a population. To address this, TreeMix (Pickrell and Pritchard, 2012) is an algorithm that accounts for population splits and gene flow while inferring population history. It estimates a maximum likelihood graph based on allele frequencies in sampled populations. The tool optimizes migration weights and branch lengths first, and then searches for optimal graphs.

In the original publication describing TreeMix, the authors demonstrate the tool with two examples – one with human populations and another using dog genomes. For studying human migrations, they used data for 55 modern and archaic human populations from the HGDP (Harvard HGDP-CEPH genotypes) (Cavalli-Sforza, 2005). In doing so, they found some unexpected inferences: (1) that about 16% of Cambodian ancestry can be traced back to a population that is equally related to both Europeans and other East Asians, and (2) an inferred admixture event between the Mozabite (a Berber population from Northern Africa) and southern European populations. They further explored these surprising findings and were able to confirm their findings with the Cambodian population and were able to conclude that the present-day Cambodian population was founded by an admixture event involving southeast Asian populations related to the Dai (Southern China) and a Eurasian population that are only distantly related to present-day populations. However, their findings with the Berber group was not consistent across independent runs of TreeMix. Following this, they interpret these results as pointing to complex patterns of gene flow between northern Africa, southern Europe, and the Middle East.

The group behind TreeMix also used 82 dog breeds or wild canids. Their results are consistent with the known history and genetic bottlenecks known in the establishment of certain dog breeds. One example of this is an inference that bull mastiffs are a result of admixture between bulldogs and mastiffs (which is a known event). However, the group does clarify that on examining the residuals from the model, they found a number of populations that do not fit a strict tree model–saying that the tree model explained 94.7% of the variance in relatedness between breeds, somewhat less than between human populations.

TreeMix is frequently used to infer migration and admixture histories in human populations. In a 2012 publication (Meyer *et al.*, 2012), the authors use this approach to infer genetic exchange between Denisovans and present-day Papuans. More recently, a 2018 publication (McColl *et al.*, 2018) employs this method to infer the prehistoric genetic exchanges that shape Southeast Asia today.

Inferring Population Structure

In a situation where group membership needs to be evaluated for different individuals, their genomic data can be used to unambiguously cluster them into coherent groups whose membership is dictated by genetic features. This approach is a marked departure from the other model-based approaches discussed in this text since it does not rely on an underlying model (and its associated assumptions) to investigate population structure. There are some examples in the literature that use this model-free approach to infer population structure.

An algorithm called NetView (Greenbaum *et al.*, 2016) for evaluating population structure using an undirected network of pairwise genetic similarity of all sampled individuals. This densely connected network is then partitioned using community-detection algorithms. These communities (or densely connected subgraphs) can be equated with population structure.

The algorithm can be equated with other model-free approaches like Principal Component Analysis (PCA) in that it uses genetic distance (or similarity) to model distances between individuals. However, the network-based algorithm facilitates community-detection in a quantitative fashion—something that is not afforded to researchers using PCA without performing additional analyses.

Using NetView, its authors were able to cluster individuals sampled from global populations into different clusters using varying thresholds. Using a low edge-removal threshold, the authors were able to separate African individuals from non-African individuals. On using a medium edge-removal threshold, they were able to distinguish African, Indo-European, and East Asian populations, and upon using a high edge-removal threshold, they were able to distinguish populations from Africa, Europe, India, China, Japan, and Mexico. For readers familiar with PCA analysis of global human populations, these results will be intuitive–however, this network-based method partitions the data without the need for manual/clustering-based partitioning following a traditional PCA.

Another way of using networks to model relationships between individuals is by using Identity-by-Descent (IBD) as similarity metric. This method uses shared haplotypes as a similarity metric compared to absolute similarity used in NetView. This allows for the detection of closer familial structures and pedigrees in the data. This was first proposed and demonstrated in (Gusev *et al.*, 2012). Haplotype-sharing networks were also employed for detecting population structure is demonstrated in a study published by AncestryDNA (Han *et al.*, 2017).

In conclusion, it can be asserted that simple non-reticulate, bifurcating trees are inadequate for modeling evolutionary relationships and that more complex network-based models are needed to understand the deeply convoluted history of a species. The literature discussed here does exactly that and has been seminal in improving our understanding of human evolutionary history.

References

Brucato, N., *et al.*, 2018. The Comoros show the earliest Austronesian gene flow into the Swahili corridor. American Journal of Human Genetics 102 (1), 58–68.

Cavalli-Sforza, L.L., 2005. The human genome diversity project: Past, present and future. Nature Reviews. Genetics 6 (4), 333–340.

Feldman, M., *et al.*, 2019. Late Pleistocene human genome suggests a local origin for the first farmers of Central Anatolia. Nature Communications 10 (1), 1218.

Green, R.E., *et al.*, 2010. A draft sequence of the Neandertal genome. Science 328 (5979), 710–722.

Greenbaum, G., Templeton, A.R., Bar-David, S., 2016. Inference and analysis of population structure using genetic data and network theory. Genetics 202 (4), 1299–1312.

Gusev, A., *et al.*, 2012. The architecture of long-range haplotypes shared within and across populations. Molecular Biology and Evolution 29 (2), 473–486.

Haber, M., *et al.*, 2019. A transient pulse of genetic admixture from the crusaders in the near east identified from ancient genome sequences. American Journal of Human Genetics 104 (5), 977–984.

Han, E., *et al.*, 2017. Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nature Communications 8, 14238.

Harris, A.M., DeGiorgio, M., 2017. Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. Human Biology 89 (1), 21–46.

Jeong, C., *et al.*, 2019. The genetic history of admixture across inner Eurasia. Nature Ecology & Evolution 3 (6), 966–976.

Lipson, M., *et al.*, 2018. Ancient genomes document multiple waves of migration in Southeast Asian prehistory. Science 361 (6397), 92–95.

Lorente-Galdos, B., *et al.*, 2019. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations. Genome Biology 20 (1), 77.

McColl, H., *et al.*, 2018. The prehistoric peopling of Southeast Asia. Science 361 (6397), 88–92.

Meyer, M., *et al.*, 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338 (6104), 222–226.

Mondal, M., *et al.*, 2016. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation. Nature Genetics 48 (9), 1066–1070.

Patterson, N., *et al.*, 2012. Ancient admixture in human history. Genetics 192 (3), 1065–1093.

Peter, B.M., 2016. Admixture, population structure, and F-statistics. Genetics 202 (4), 1485–1501.

Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8 (11), e1002967.

Racimo, F., *et al.*, 2015. Evidence for archaic adaptive introgression in humans. Nature Reviews. Genetics 16 (6), 359–371.

Raghavan, M., *et al.*, 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505 (7481), 87–91.

Reich, D., *et al.*, 2009. Reconstructing Indian population history. Nature 461 (7263), 489–494.

Sikora, M., *et al.*, 2019. The population history of northeastern Siberia since the Pleistocene. Nature 570 (7760), 182–188.

Tambets, K., *et al.*, 2018. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. Genome Biology 19 (1), 139.

Wang, C.C., *et al.*, 2019. Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. Nature Communications 10 (1), 590.

Yang, M.A., *et al.*, 2017. 40,000-Year-old individual from Asia provides insight into early population structure in Eurasia. Current Biology 27 (20), 3202–3208 e9.