Chapter 13

Advances in population genetics and language history: how large datasets and ancient DNA changed the picture

Chiara Barbieri, Paul Widmer

Abstract

Since the development of genetic analysis for the study of population history, the discipline has been paired with linguistics to compare human demographic and cultural trajectories, focusing on targeted regional cases and a few global scale studies. This chapter reviews advances in the field of genetics from the past two decades and focuses on different scales of resolution: technological improvements, geographic coverage and time depth of data availability, with the study of modern and ancient DNA. To understand how these advances can integrate multidisciplinary studies of language history, this chapter illustrates possible scenarios which link the history of languages with the history of their speakers. Migration, contact, and isolation are some of the factors in play, which can also contribute to the degree of correspondence between genes and languages.

Key words

Population genetics, global scale, ancient DNA, demography, technological advances, migration, contact, isolation

Bio

Chiara Barbieri is a geneticist working in the Department of Evolutionary Biology and in the Department of Comparative Language Science at the University of Zurich. She leads the group 'Human genetic diversity across languages and cultures'. Her research is situated at the intersection of molecular anthropology and processes of language diversification and contact, both at a global scale and at a regional scale, with a focus in South America.

Paul Widmer holds the chair of Indo-European Studies in the Department of Comparative Language Science at University of Zurich. His research focuses on the cultural context, transmission and description of ancient Indo-European languages, the anthropological, social, and biological drivers of language diversification and change, and language contact.

13.1 Introduction

Technical innovations and emerging trends in linguistics, genetics and archaeology are constantly opening new avenues for multidisciplinary research. While each discipline is "evolving" at a different pace, they have all experienced data driven approaches as a catalyst for innovation. Quantitative, large scale data collections are now available for different linguistic features (Carling 2017; Moran and McCloy 2019; Dryer and Haspelmath 2020; Batsuren, Bella and Giunchiglia 2022; Bickel *et al.* 2022, Skirgård *et al.* 2023) as well as for anthropological fields of research related to material and immaterial culture (Teixidor-Toneu, Jordan and Hawkins 2018; Turchin *et al.* 2018; Aguirre-Fernández *et al.* 2021; Wood *et al.* 2022, Passmore *et al.* 2023). Large quantitative datasets can be used in combination with powerful analytical tools to evaluate patterns and associations of features. Many of these tools come from methods used in the biological sciences, including population genetics (Atkinson and Gray 2005; Mace and Holden 2005; Mesoudi, Whiten and Laland 2006; Tëmkin and Eldredge 2007; Steiner, Stadler and Cysouw 2011; Levinson and Gray 2012).

The increase in data density and cross-continental coverage is gradually building a comprehensive global catalogue of human history and extant diversity. Small-scale local case studies are essential to anchor our knowledge of events and dynamics in time and space: from there, specialists of different historical disciplines can gain a deep understanding of the factors at play and formulate more general hypotheses to test. The vast amount of knowledge accumulated allows us to zoom out and explore patterns at a continental and global scale, with obvious limitations imposed by data coverage. Dense, cross-continental datasets are available for different types of linguistic data including, for example, phonetic data (Moran and McCloy 2019), lexical distances (Wichmann, Brown and Holman 2022), etymologies (like in https://www.wiktionary.org/), and typological features (Skirgård *et al.* 2023). The list of languages mapped in available datasets has increased by orders of magnitude, and continues to grow, in an attempt to effectively represent the diversity of all known and described languages of the world, including ancient, extinct, and reconstructed ones (Moran, Grossman and Verkerk

2021). Standardized resources like Glottolog (Hammarström *et al.* 2022) or Ethnologue (Lewis 2009) provide ever-growing and comprehensive lists of languages and their genealogical classification. Such global linguistic datasets and linguistic mapping have enabled scientists to explore questions of linguistic history (Blasi *et al.* 2019; Hua *et al.* 2019) going beyond the technical limitations of time depth reconstructions - up to 5000 years ago for cultural transmission and up to 7000 to 9000 years ago for language history.

The availability of genetic data has also been increasing in volume and geographic coverage. Population genetic studies have been focusing on many regions of the world, clarifying the demographic dynamics of the past, and providing indirect evidence to utilize for archaeological and linguistic historical reconstructions. One of the main genetic revolutions of this century has been the analysis of ancient DNA (aDNA), which has increased in quality and the quantity of ancient individuals genotyped. Ancient DNA provides a direct anchoring for the evolution of genetic profiles in time and space and can serve as a source for mapping demographic and linguistic trajectories in the past. The combination of large, high-resolution genetic and linguistic datasets opens novel avenues for testing specific hypotheses with explicit models. Nevertheless, some fundamental questions persist. To what extent do linguistic and genetic history align? How deep can we go in reconstructing human history, when genetic and linguistic data are combined?

This chapter is dedicated to recent advances in genetics research across two dimensions of resolution: time and space. The following sections will highlight the potential and limitations of the currently available data for reconstructing human history, and their possible applications in linguistic studies.

13.2 Advances in the field of genetics

13.2.1 Type of genetic data and geographic coverage

Genetic datasets for the study of human history have traditionally focused on two types of genetic marker, each associated with slightly different analytic approaches: uniparental markers and autosomal markers. Uniparental markers comprise mitochondrial DNA (mtDNA) and the non-recombinant portion of the Y chromosome DNA (sometimes referred to as NRY). mtDNA is transmitted without recombination in the maternal line, NRY is transmitted without recombination in the paternal line to males only. Autosomal markers are those in the remaining

set of chromosomes, which is subject to recombination, and is transmitted by virtually all ancestors of an individual.

Uniparental markers have been widely used in the past decades because of their power to reconstruct human history (Underhill and Kivisild 2007; Lippold *et al.* 2014). Their transmission modality is particularly relevant for tracing single lineage movements (Pakendorf and Stoneking 2005; Calafell and Larmuseau 2017) and possible sex-biased patterns (Heyer *et al.* 2012). These are cases when females and males engage in different demographic histories, display different mobility and migration strategies, or represent different effective population sizes (i.e. the fraction of the population who reproduces and passes the genes to the next generation). Standardized laboratory protocols for the analysis of uniparental markers have been accessible at relatively affordable costs and are widely used and developed across disciplines, e.g. in forensic genetics, a discipline that shares with anthropology the interest in understanding global genetic diversity. These shared standardized lab procedures have also been facilitating the production of cross-continental data that could be conveniently analysed against a compatible broad panel of population diversity.

MtDNA used to be analysed initially only for a few key nucleotide mutations (Single Nucleotide Polymorphisms or SNPs, pronounced [snɪp]) defining haplogroups of interest. Haplogroups correspond to major branches of the mtDNA tree, which share mutations from a common ancestor, often have a specific spatiotemporal signature, and are conventionally named with capital letters. Subsequently, it became common practice to sequence a specific block of the molecule, ~400 base pairs (bp) long, corresponding to the first block of the Hypervariable region (sometimes referred to as Hypervariable Segment, conventionally split into two blocks: HVSI and HVSII). Finally, laboratories started to routinely sequence the whole mitochondrial DNA molecule, or mitogenome (Torroni *et al.* 2006; van Oven 2015), which is 16596 bp long in the human reference (Andrews *et al.* 1999; Behar *et al.* 2012).

Y chromosome data was also traditionally analysed for a few key SNP markers defining haplogroups of interest, with the addition of Short Tandem Repeats (STRs, or microsatellites) haplotypes, which reported the number of repetitions of short blocks of DNA in a set of specific positions ("loci"). Only recently, geneticists focused on sequence data to expand the resolution of the Y chromosome variation, and to find new SNP variants (Hallast *et al.* 2015; Jobling and Tyler-Smith 2017). However, the effort required for sequencing the non-recombinant portion of the Y chromosome is much larger than for sequencing the mtDNA genome. This is due to the larger size of the Y chromosome molecule (~50 million bp) and to technical challenges

caused by large fractions of repetitive regions (Bachtrog and Charlesworth 2001). This explains why standardized, comparative data for Y chromosome sequences are still scarce in comparison to mtDNA genomes, despite the shift to "Next Generation Sequencing" (NGS) – or "Massive Parallel Sequencing" (MPS) as it is referred to in forensic genetics – that has allowed scientists to use standard protocols and relatively affordable technologies for DNA sequencing from the early 2010s onward (Kircher and Kelso 2010; Metzker 2010; Goodwin, McPherson and McCombie 2016).

The other type of genetic markers traditionally studied in human genetics are the so-called autosomal markers, a term that covers the whole genomic variation in diploid chromosomes (excluding sex chromosomes). This is the genetic variation inherited from (virtually) all ancestors, and therefore more representative of the history of an individual or a population, in comparison to the uniparental markers. Autosomal markers analysed included single SNPs or sets of microsatellites, in some cases genotyped with standardized kits, again developed in the forensic field (Rosenberg *et al.* 2002; Jakobsson *et al.* 2008; Kayser and de Knijff 2011). The term was then expanded beyond the definition of autosomal markers, to include markers from all the chromosomes and potentially also mtDNA: in this sense, we now talk about genome-wide or genomic data, and not anymore about autosomal data.

A common method to retrieve a large number of SNPs in one sequencing run is the SNP chip. Those chips, or arrays, include probes targeting a defined set of known polymorphic (variable) positions. Several chips have been proposed by laboratories and companies, manufactured to include thousands – or hundreds of thousands – SNP positions. The SNPs included in the chips are known to be polymorphic in populations where full genomes were available and are often relevant for medical studies (Ragoussis 2009). The use of SNP chips creates a bias in our ability to detect new variants in understudied populations, in particular from under-sampled regions of the world (Lachance and Tishkoff 2013; Pugach and Stoneking 2015). The advantage of generating data with available SNP chips relies again on having standardized laboratory protocols and comparative datasets, but the choice of platforms available is still very diverse, each one having different advantages and disadvantages. For a comprehensive study of the genomic variation of a geographic region, or a continent, it is therefore often necessary to merge datasets generated with different SNP chips. The larger the number of SNP chips to be merged, the smaller the number of SNPs that overlap between the chips will be, negatively impacting the analysis. A popular SNP chip used in human history studies is the Human Origins Array (Affymetrix-Axiom), which includes ~600,000 SNPs and

is designed specifically for human history analysis, ascertained for SNPs variable in 11 populations from all continents (Patterson *et al.* 2012).

Finally, the ultimate type of data for a full, high-resolution analysis is the whole genome, consisting of more than 3 billion bases, of which ~84 million have been recognized as SNPs (The 1000 Genomes Project Consortium, Auton *et al.* 2015). A good quality genome, where SNPs can be reliably determined, should be sequenced at high coverage so that each base is sequenced at least 20 or 30 times (conventionally indicated as 20X or 30X coverage). Generating full genomes is still a cost-intensive process, that requires solid bioinformatic expertise. Merging genomes generated with different sequencing technologies and different filtering steps is a particularly delicate task because it may bring subtle batch effects into the downstream population genetics analysis.

Generating global genomic data with comparable and compatible technologies relies on available samples, or DNA diversity panels. Panels of diversity originally built for medical studies counted on many individuals from a few populations representing all the continents, e.g. 11 populations in the HapMap project (Altshuler *et al.* 2010), or 26 populations in the 1000 Genomes project (Auton *et al.* 2015; Byrska-Bishop *et al.* 2022). Other panels have been designed to further expand our knowledge of cross-continental human diversity with a larger number of diverse populations. A widely used genomic panel built for medical and historical studies is the Human Genome Diversity Project HGDP–CEPH, assembled in 2002, which includes 54 linguistically and culturally diverse populations (Bergström *et al.* 2020; Aneli, Birolo and Matullo 2022). The Simons Genome Diversity Project employed a different sampling strategy, with two individuals per population, and generated 300 high-coverage genomes from 142 populations (Mallick et al. 2016). Other endeavors dedicated to mapping global human diversity and history through our DNA included the Genographic Project, launched in 2005, which was reliant on voluntary participation (Behar *et al.* 2007) and the cooperation of numerous laboratories dedicated to sample collection. Completion of the project was undermined by difficulties in coordinating the large number of research teams involved, and frequent backlash from indigenous representatives concerned with biocolonialist practices and opaque ethical standards (Malhi 2009).

The efforts to represent global human diversity to a fine-grained level were also motivated by a problematic underrepresentation of minority groups. As briefly mentioned before, with the problem of ascertainment bias in available SNP chip platforms, our knowledge of genomic variation mostly comes from panels of diversity, or cohorts, centered in Western countries and

European ("white") ancestry. With the scope of inclusivity and more just production of health-related research, the World Medical Association redacted the Declaration of Helsinki in 1964 (WMA — The World Medical Association-Declaration of Helsinki 1964), which provides guidelines for ethical research with human participants (Williams 2008). The guidelines aim to foster access to research while at the same time protecting minorities and vulnerable groups from exploitation. In particular, indigenous communities are often poorly represented in genetic studies and receive the least benefit from scientific research (Hudson *et al.* 2020). The rising attention towards best ethical practices, transparency and participation of minorities and indigenous groups can also be considered an important advance in the field of genetics (Claw *et al.* 2018).

13.2.2 Source of genetic data, modern and ancient DNA

Recruitment of voluntary participants in genetic studies can be performed through medical centres, as cohorts for patients and healthy/control individuals, from citizen science calls, or from anthropologically motivated fieldwork. Cohorts originally recruited for medical studies are also used as a powerful source of analysis for human history studies (Leslie *et al.* 2015). Citizen science approaches, sometimes through participants who buy commercially available kits for ancestry testing, have also been employed in human history research (Bryc *et al.* 2015). Anthropological fieldwork would provide a valuable source of data for the special attention on the genealogy of the participant, often recruited to be representative of a specific ethnolinguistic group in time and space. Individuals involved in the sample would possibly fit the "four grandparents rule", having all four grandparents from the same region and/or ethnolinguistic group. A linguistically informed sampling strategy would possibly give the best assessment of the language spoken by the participants and their parents and grandparents, even if the common situation of multilingual communities poses challenges to a unique population-language assignation.

The most relevant advance in the field of human genetics, nevertheless, revolves around a new source of genetic data: ancient human bones. Technological advances in data generation, from fragmented "low quality" DNA, have been projecting the depth of genetic reconstructions further back into the past (Orlando, Gilbert and Willerslev 2015; Llamas, Willerslev and Orlando 2017). In a way, the analysis of "modern" genetic samples already enables historical reconstructions into the past, and does not just provide a snapshot of the present. With modern

genomic data, in fact, it is possible to read pieces of the genetic history of (virtually) all the ancestors of a group or an individual, while with uniparental markers it is possible to trace the trajectory of single maternal and paternal histories through generations. Missing individuals in the reconstruction of the past are "inferred" from the present-time variation. However, with the analysis of ancient genetic samples our power of reconstructing the past reaches further levels of sophistication: genetic variation is directly placed into the picture, with a precise anchoring in time (dating the biological material with radiocarbon dating techniques, or indirect stratigraphy chronologies), and place (the archaeological site where the sample was initially excavated).

Some limitations of aDNA analysis concern the degradation of the DNA from ancient individuals, which impacts the quality and quantity of data retrieved. One characteristic of degraded fragments of aDNA is that they are shorter than "modern" DNA from fresh biological material (Orlando, Gilbert and Willerslev 2015). Because of this characteristic, an accessible type of genetic marker to analyse is mtDNA, which is a short molecule, present in organisms in many more copies than the chromosome DNA of the nucleus. The focus on ancient mtDNA variation has brought new attention to the tradition of uniparental studies and new motivation to generate comparative datasets of mtDNA genomes from present-time populations. Other types of genetic markers routinely analysed with aDNA are whole genomes, often sequenced at low coverage for degradation issues and overall scarcity of DNA. Ancient genomes are not immediately mergeable with modern genomes, for some of the technical issues described in section 13.2.1. A popular method for the analysis of aDNA is the sequencing of a panel of ~1.24 million SNPs – commonly referred to as the 1,240 K capture (Mathieson *et al.* 2015). The panel includes all the SNPs from the Human Origins array (Patterson *et al.* 2012), plus a selection of other variable SNP positions, assuring compatibility between existing datasets. One of the largest collections of ancient and modern DNA, compatible and standardized either with the 1,240 K or Human Origins array, is the one available in the Allen Ancient DNA Resource (Allen Ancient DNA Resource, version 54.1.p1), which currently includes the published genetic profile of almost 10,000 ancient individuals.

13.2.3 Improvements in resolutions for genetic studies

Geographic resolution representing cross-continental human diversity in the present and the past (with aDNA) is crucial to achieve a good understanding of variation patterns and

population relationships. The global panels of diversity described in section 13.2.1 have been extensively analysed at the full genomic resolution, but still have an incomplete coverage for some regions. Uniparental markers, which have been studied for a long time in a systematic and standardized way, often provide a dense geographic resolution for comparative analysis. High-resolution mtDNA sequence data is available for most regions of the world, either for studies representing population history and diversity, or for the study of single lineages of interest. Global, high-resolution Y chromosome data is available for haplogroups and for microsatellites. Databases built for forensic use represent a useful resource of data (Roewer and Parson 2012). MtDNA is still particularly valuable for studies with aDNA. The comparison of ancient and modern mtDNA has unlocked important findings, for example for understanding population turnover and population structure in the past (e.g. Llamas *et al.* 2016 in the Americas; Posth *et al.* 2016 in Europe).

When looking at geographic coverage and the types of genetic markers available, some discontinuities can be found, which undermine the possibility of matching linguistic and archaeological data in specific regions. There are regions extensively studied for uniparental markers in the past, but poorly represented with high-resolution genomic data in recent publications. This is the pattern found in the Americas, for example, where uniparental data is relatively abundant – particularly from studies in the past decade (Bisso-Machado and Fagundes 2021), but recent genomic data is relatively scant, for the concerns of indigenous representatives advocating for a more transparent participation in the scientific discourse, against the unsatisfactory ethical standards of the past. Some chapters of this book, therefore, deal with a genetic representation of the history of speakers that is either partial for the type of markers analysed (uniparental data genotyped at low resolution), but richer in geographic coverage; or vice-versa, that is partial for the patchy geographic coverage, but richer in genotyping information (high-density SNP data or whole genomes).

Finally, as high-resolution data has become available, both in terms of fine-grained genomic data and global population coverage, one could ask if the production of detailed human genetic diversity actually corresponds to a richer understanding of the details of human history. Much of the knowledge obtained with the first studies of mtDNA and Y chromosome haplogroups remains largely unchanged after the release of more high-resolution genomic data: for instance, our origins in the African continent (Vigilant *et al.* 1991), the patterns and timing of dispersal of the early migrations in the Americas (Schurr and Sherry 2004; Tamm *et al.* 2007), and the patterns and timing of the Austronesian expansion in the Pacific (Kayser *et al.* 2008).

Nevertheless, with genomic datasets it is possible to achieve a more detailed resolution of regional history. Fine-grained reconstruction of demographic changes requires high-density genomic data, which in turn allows the retrieval of rare SNP variants. With this type of data, demographic dynamics can be then reconstructed both in the distant past (Mathieson and McVean 2014; Schiffels and Durbin 2014) and towards the present (Ralph and Coop 2013; Kelleher *et al.* 2019). Ultimately, the role of aDNA in resolving complex episodes in human history is undisputed (Pickrell and Reich 2014; Slatkin and Racimo 2016; Liu *et al.* 2021), and some of these episodes have particular relevance for the linguistics-archaeology triangulations discussed in this book. The next section clarifies how the genetic knowledge accumulated so far can be of meaningful use for linguistic studies, and highlights opportunities and limitations.

13.3 Integrating new genetic findings in linguistic studies

The incremental release of ancient DNA and genome-wide dataset of present-time genetic diversity allowed geneticists to further advance the knowledge of our past, with cases of migration and contact illuminating linguistic and cultural dynamics of change. For example, we could clarify the degree of population turnover in the Austronesian expansion (Lipson *et al.* 2018; Pugach *et al.* 2018), the time depth of genetic structure in sub-Saharan Africa and the Bantu spread (Vicente and Schlebusch 2020; Lipson *et al.* 2022), and, in particular, the complex sequence of large migrations through western Eurasia (Allentoft *et al.* 2015; Lazaridis *et al.* 2022), which is the region that provided most of the available aDNA so far (Olalde and Posth 2020). Some of the most remarkable discoveries from aDNA studies concern the relationships between modern humans and our distant relatives, like Neandertals and Denisovan hominins who lived up to 500,000 years ago (Green *et al.* 2010; Reich *et al.* 2010). However, it is far beyond the scope of this book to examine linguistic insights over such a deep time scale.

Table 13.1. Possible population demographic scenarios that can be reconstructed with genetic data, and suggestions of social and linguistic scenarios that could be tentatively associated. In the third column, illustrative cases from different regions of the world.

Demographic scenario from genetics	Possible social and linguistic scenarios	Examples
Contact between populations	Language contact Lexical borrowing Pattern borrowing	Multiple layers of loanwords in English (Durkin 2014) Shared constructional patterns in the upper Northern Amazon (Gijn *et al.* 2023) Sharing of clicks between Khoisan and Bantu groups (Barbieri *et al.* 2013)
Isolated population with low genetic diversity	Language isolates Language enclaves	Basque (Flores-Bello *et al.* 2021) Mapudungun (Arango-Isaza et al. 2023) German-speaking linguistic isolates from the Eastern Italian Alps (Capocasa *et al.* 2013) Hadza (Lachance *et al.* 2012)
Stratified/admixed population with high genetic diversity	Language shift High amount of (adult) L2 speakers Language mixing Formation of creoles Increased transparency and regularity of form-function mappings	Historically documented shifts to colonial languages in the Americas and in North Africa (to Spanish, Portuguese, Arabic, etc.) (Pena, Santos and Tarazona-Santos 2020; Bird *et al.* 2023) Reduction of inflection and allomorphy (Bentz and Winter 2013; Polinsky 2018) Malta genetic stratification and Maltese as a mixed language (see Barbieri *et al.* 2022) Formation of Krio, an English-lexifier creole and lingua franca in Sierra Leone (Finney 2013)
Genetic homogeneity between populations	Linguistic Areas Phylogenetic relatedness within families	Western Eurasia (Lao *et al.* 2008; Haspelmath 2001) Central Andes (see Chapter 36 in this book)
Genetic barriers between populations, which can lead to a population split	Diversification of two or more languages within and between families	High between-population diversity in the Americas is associated with high language family diversity (Belle and Barbujani 2007; Wang *et al.* 2007)
Population expanding	Imposing cultural and linguistic packages over a region	Bantu migration (Fortes-Lima *et al.* 2023)
Population contracting (bottleneck)	Risk of language extinction	Native American population decimation after European contact (O'Fallon and Fehren-Schmitz 2011)

Population replacement in a region (from aDNA transects and recent admixture)	Language replacement, language shift	Hungarians maintaining the original Uralic language despite genetic replacement (Maróti *et al.* 2022) Shift from Uralic to Slavic in the Suzdal region, Volga-Oka interfluve (Peltola *et al.* 2023) Introduction of West Germanic Languages in Britain with population replacement (Gretzinger *et al.* 2022)

To understand the impact of these advances in linguistic-genetic comparisons, we need to clarify which processes are meaningful to address, which information is useful for our comparisons, and which questions we can ask. We can start by outlining the key role of genetics, which is reconstructing demographic changes and patterns of relatedness. The dynamics between speakers can play various roles in the diversification and diffusion of languages. One process to address is the level of interaction between populations: how strong is the contact between two or more groups? How pronounced is the isolation and genetic distinctiveness of a group? Another demographic aspect is the level of diversity inside a population: how large is the population size – in genetics represented as effective population size, as specified above? How stratified is a population, from genetically homogeneous to genetically very diverse? Finally, another process is the variation of population parameters through time, for example, does the population expand or reduce in size (bottleneck)? Does the stratification within a population lead to a split of two "daughter" populations? When did events of split, merger, migration and contact occur?

These scenarios can impact the evolution of a language through sociolinguistic dynamics. Some possible associations between population genetic scenarios and linguistic (and social) scenarios are listed in Table 13.1, with a few examples to illustrate the possible outcomes. High genetic diversity and traces of genetic admixture in a population are signals of large population size and a previous history of contact. These attributes can be linked with measurements of linguistic diversity (Nichols 1992). While some studies show that language complexity is reduced when the number of speakers is larger (Lupyan and Dale 2010), there is evidence for larger linguistic repertoires in less isolated populations (Trudgill 2002), and for population size not affecting language change (Wichmann and Holman 2009) nor phonemic inventory size (Moran, McCloy and Wright 2012). In some cases, small populations can be multilingual, and multilingual populations can be either diverging or converging (Evans 2018, 2019). The

possible association between population size and rate of language change might therefore be very subtle and difficult to capture for all languages (Greenhill et al. 2018). Genetic distinctiveness between populations can be associated with mating barriers, and with cultural/linguistic barriers, which can also impact language diversification and change (Efrat-Kowalsky *et al.* 2022, White 1997). Language boundaries take on variable forms and can in turn affect genetic structure in a region (Kandler, Unger and Steele 2010). Language barriers that are permeable to gene flow can be affected by enhanced language contact. Finally, relevant insights can be gained by following migration paths and the effects of contact or population replacement. When a region appears genetically homogeneous, this might be the result of long-distance movement of people and/or large population size. These regions might be subjected to linguistic homogeneization and/or linguistic areal effects within and across families (Nichols 1992). Migration might be the ultimate push for language spread, but only if we assume that people carry their language and genes together. If this is the case, population turnover is assumed to be accompanied by language change, or language shift in a region. Language shift can be defined as the change of language through cultural exposure, without substantial genetic contribution from another group, resulting in a mismatch between genetic and linguistic patterning. The topic of overall language-gene correspondence and parallel vertical transmission will be addressed in more depth in section 13.4.

Both uniparental and genomic markers have been successfully used in disentangling demographic and historical scenarios paired with linguistic and archaeological data on regional case studies, and several examples are discussed in Part 3 of this book. Autosomal/genomic datasets are useful in drawing broad patterns of relatedness and are versatile to analyse the relevant case scenarios discussed in Table 1: migration, contact, variation in population size, and replacement. In terms of systematic, broad comparisons, a special note should be made for uniparental markers, which can also be particularly useful for questions related to linguistic diversity, sociolinguistic dynamics, and language acquisition in history – despite their technical limitations. Studies have employed uniparental data to assess generalized sex-biased patterns, exploring maternal or paternal influences in language transmission. Human populations are generally patrilocal (i.e., the couple moves to the place of residence of the male's family), and this results in a higher structure for the Y chromosome, in comparison to a higher homogeneity of the mtDNA, due to the higher mobility of females (see also Chapter 14). Studies of regional (Tambets *et al.* 2018) and global variation detected a tendency to match linguistic patterning with those reconstructed with the Y chromosome, suggesting a stronger case for paternal

inheritance, or "father tongue" (Forster and Renfrew 2011). Another study broke this effect down into different elements of linguistic diversity and found that, while lexical diversity correlates with Y chromosome diversity, phonemic diversity correlates with mtDNA diversity, suggesting different mechanisms and parental roles in language acquisition (Zhang *et al.* 2019). Other interesting uses of uniparental datasets for linguistic study include the focus of preferential linguistic exogamy, i.e. when people from an ethnolinguistic group are encouraged to mate with females from another linguistic group: this social norm characterizes certain Tukanoan-speaking communities in the Colombian Amazon (Arias *et al.* 2018) – see Chapter 37 in this book.

13.4 Do languages and genes correspond globally?

Many of the gene-language parallels outlined in this book would be explained with an association of genes and languages through migrations and population replacements, which implies that people usually carry their genes and their languages through time and space (see also Chapter 14 in this book). But how systematic would these association patterns be in a global gene-language overview? The overall question of gene-language correspondence has a deep history of study and has been tackled with different approaches. It was first envisaged by Charles Darwin in *On the Origin of Species* (Darwin 1859), and then formally tested with human genetic data in the pioneering work of LL. Cavalli-Sforza and R. Sokal in the 80s (Cavalli-Sforza *et al.* 1988; Sokal 1988). This first phase of gene-language studies focused on methodological advances in the analysis of genetic data and spatial differentiation process (Barbujani and Sokal 1990; Cavalli-Sforza, Minch and Mountain 1992; Barbujani and Pilastro 1993; Penny, Watson and Steel 1993; Belle and Barbujani 2007). The linguistic base of these studies consisted of language trees coming from historical linguistics, centered on the Eurasian continent (i.e. relationships between speakers of Indo-European languages and sub-families). As a side note, these genetic and cultural comparison studies sealed a long-lasting partnership between human geneticists and the linguistic classifications, proposed at that time by Joseph Greenberg and Merrit Ruhlen (Greenberg 1987, Ruhlen 1991). Their linguistic references are still commonly used in the field of molecular anthropology – despite some of them being rightly criticized by most linguists (Bolnick *et al.* 2004 – see Chapter 37 in this book).

A second phase of gene-language studies occurred in the last decade with the work of Renfrew, Bellwood and Diamond, which linked the demographic processes carrying both

genes and languages together with massive human expansions. Such expansions would have been triggered by technological advances, like a shift in subsistence from foraging to food production (Renfrew and Bellwood 2002; Diamond and Bellwood 2003 – see Chapter 16 in this book). These studies concentrated on single language families and covered different continents – not only Europe (Lewis *et al.* 2005; Kayser *et al.* 2008; de Filippo *et al.* 2011; Barbieri *et al.* 2014), and in recent years have been drawing further insights from aDNA (Haak *et al.* 2015; Posth *et al.* 2018). In parallel, circumscribed case studies of contact on regional scale reported clear cases of mismatch between languages and genes (Pakendorf 2014). Examples of mismatches (or language shifts) have been contextualized by human genetic studies of different regions of the world (Nasidze and Stoneking 2001; Chaubey *et al.* 2008; Mona *et al.* 2009; Steele and Kandler 2010; Barbieri *et al.* 2011; Pickrell *et al.* 2012).

A third phase in gene-language studies developed from explicit and systematic gene-language comparisons: expanding up to a worldwide scale and including both genetic data and quantitative linguistic data in the process. A first effort to check for worldwide gene-language correspondence with dense datapoint coverage was done by Creanza et al. (2015): using autosomal microsatellite data and a database of phonetic diversity, they found parallel axes of correlations at a regional scale, while controlling for geography. Longobardi et al. (2015) employed a gene-language comparison across language families to prove a general framework hypothesis: the convergence of syntactic data and genetic data on deep time-scale relationships. In other studies, different sources of linguistic variation (e.g., phonemic and lexical) were considered, but only within one language family, typically the Indo-European. In a multidisciplinary study on northeast Asia, genetic, musical, and linguistic cross-family data were compared: the authors found an association between genetic and linguo-morphological (but not lexical) data that holds above spatial autocorrelation effects (Matsumae *et al.* 2021).

In many studies the focus is often put on the matches between genes and languages, disregarding the mismatches as an exception to the norm. Nevertheless, as we have seen, language transmission occurs not only on vertical pathways (from one generation to another), but also by horizontal transfer, within the same generation. To tie the gene-language parallel together, the framework should consider coherent demographic dynamics and the alignment of geographic patterns and time frames (Heggarty 2014). Spatial correspondences can represent a confounding factor, since until the introduction of general literacy, geographic proximity was the major constraint to human contact, both demographic and cultural. The simple ecological model of isolation by distance (IBD) predicts a clinal genetic distance distribution proportional

to geographic distance (Rousset 1997): such spatial autocorrelation is also verified in humans (Sokal, Oden and Thomson 1992; Prugnolle, Manica and Balloux 2005). Concerning time, languages can change faster than the gene pool of a population, and there are limitations to the evolutionary time frame that can be considered to reconstruct linguistic relationships (see section 13.1). As a result, the relationships between populations (genetics) and languages are shaped by demographic events that might have occurred at different times (Levinson and Gray 2012; Pagel 2017).

In a recent paper, the gene-language congruence was tested at a global scale, using a large genome-wide dataset, and focusing on a systematic search for cases of mismatch (Barbieri *et al.* 2022). The database used, named GeLaTo (Genes and Languages Together) is assembled with "modern" DNA from publications which provide enough geographic and cultural information to assign each genetic population to a language – through Glottocodes (Hammarström *et al.* 2022). The type of genome-wide data used is the Human Origins SNP chip described in section 13.2.1, which provides genomic high resolution, compatibility through published datasets, and relatively low ascertainment bias issues (Patterson *et al.* 2012).

The results focused on two levels of mismatches: single population mismatches, and general patterns of mismatches through different language families. At the population level, the study observed how frequently each population is genetically close to a population speaking a language from an unrelated language family. Different analyses indicated that 20% of populations were in mismatch, a value consistent after downsampling over-represented language families. Overall, there is a tendency for single matches of populations genetically closer to their linguistic relatives, but importantly, mismatches are pervasive and ubiquitous in all continents and language families reported. As for the global patterns of mismatch within families, several language families display profiles of genetic cohesiveness, e.g., Indo-European, Atlantic-Congo, Mongolic-Khitan and Sino-Tibetan. Speakers from these families are genetically closer to each other than to speakers of other language families, even at large geographic distances. Language families that are particularly non-genetically cohesive are Turkic, Austroasiatic, Austronesian and Uralic. These overall non-genetically cohesive families can also include genetically-cohesive branches: within Turkic, there is correspondence for the Nuclear Oghuz speakers, and within Austronesian there is correspondence for the Oceanic speakers. The lack of a coherent gene-language picture in Austronesian is explained by the introgression of Papuan ancestry into the more recent Austronesian ancestry, which occurs especially in the regions of Melanesia and Papua (see Chapter 28 of this book).

An analysis of correspondence between linguistic divergence time and genetic divergence time returned substantial matches only within the Indo-European. The proposed chronological origins of major language families are often more recent than the corresponding timings of genetic divergence. This is possibly due to a higher uncertainty in reconstructing events far back into the past, for both genetic and linguistic methods. Genetic divergence timing within a language family tends to be very ancient, with some population pairs diverging > 10,000 years ago: these time frames are not reconcilable with the mostly accepted history of language families known to date. Very ancient divergence times can result from effects of isolation and drift, and from the coexistence of divergent ancestries within the same population. Genetic admixture, defined as the presence of divergent ancestries and lineages within the same population, is a common process in human history (Hellenthal *et al.* 2014; Nielsen *et al.* 2017), and populations should not be considered as representatives of unique "pure" ancestries (Kampourakis and Peterson 2023).

Finally, it is relevant to note how the linguistic and genetic diversity of the same region can change with increased genetic resolution and geographic coverage. For example, a previous study of mtDNA variation in the Caucasus flagged both Turkic-speaking Azerbaijani and Indo-European-speaking Armenians as language shifters, their genetic diversity being more similar to other Caucasian neighboring populations from diverse language families than to other speakers of the same language family (Nasidze and Stoneking 2001; Schönberg *et al.* 2011). With the genome-wide data and higher number of populations in GeLaTo, Azeri Azerbaijani speakers are confirmed as genetically distant from other Turkic speakers, fully in line with the proposed language shift. Armenians, on the other hand, are shown to be genetically more closely related to neighboring speakers of Indo-European languages, and do not stand out as language shifters (Barbieri *et al.* 2022).

13.5 Ancient DNA: how deep can we dig into the past?

In considering genetic history coming from ancient human remains, it is obvious that any type of association with language history is indirect and very tentative. In other words, bones do not speak. Nevertheless, this rich catalogue of genetic history now available from aDNA can still provide valuable insights into the history and diversity of languages. Two different perspectives can be taken with the analysis of aDNA together with modern DNA: first, are present-day

speakers a good representation of the history of their language? And second, do demographic changes through time affect the history of a language or language family?

For the first perspective, aDNA can indirectly be used to assess the stability of a genetic pool through time and space. In the recruitment of the participants to a genetic study, the four grandparents rule allows anchoring the gene-language association for at least two generations (see section 13.2.2). As genetic analysis reconstructs the history of a gene pool further back than two generations, the question arises: what language did the ancestors of this individual (or this population) speak? The degree of continuity between today's speakers and previous inhabitants of the region across a time transect can be associated with a scenario of population stability (Table 13.1). When this is also associated with a degree of material cultural stability (e.g. material culture associated with the archaeological sites remaining substantially unchanged through time) an association between genes, language, and material culture can emerge. This type of evidence should be taken with caution, as language shifts without any genetic turnover are still possible (as seen in section 13.4), and more generally, as the association between people, material culture and language is proved to be very nuanced across regions and cultures of the world. For example, the spread of the 'Bell Beaker complex' over western Europe from ca 2800 BC onwards over Iberia and central Europe was not primarily driven by demographic events; its expansion to Britain, on the other hand, is associated with an almost complete genetic turnover (Olalde *et al.* 2018).

Global history from aDNA reports both extreme cases, with regions characterized by stable occupation and in situ development and regions characterized by drastic genetic turnover. Again, aDNA is key to monitoring and assessing such population shifts (Mourier *et al.* 2012). Stable occupation is often found in sites in the Americas, and could trace back to early regional population structure after the initial migrations in the early Holocene (Lindo *et al.* 2017; Nakatsuka *et al.* 2020). Genetic continuity since the early Iron Age is found in present-day Basque populations of Europe (Flores-Bello *et al.* 2021), marking a genetic distinctiveness of the region which matches the presence of an isolated, non-Indo-European language (see Table 13.1).

Population turnover is described in Oceania with the Austronesian expansion (Lipson *et al.* 2018) and in various transects in Europe, where different ancestries have been diffusing at different times (Haak *et al.* 2015; Lazaridis *et al.* 2022). For example, Great Britain has witnessed a succession of large-scale migrations from the post-Roman early medieval times: an almost complete replacement with genetic lineages from continental northern Europe,

followed by pulses of migrations from a "France related" ancestry matching a Frankish connection in the archaeological record (Gretzinger *et al.* 2022) (Table 13.1). The fact that this migration was involving both women and men, and that the ancient individuals analysed as migrants were not systematically associated with markers of social prestige, suggests that Britain indeed experienced a mass migration. This is in contrast with an alternative hypothesis of elite male migration associated with the introduction of West Germanic language replacing the Celtic or Latin substrates (Schrijver 2013). In the medieval Volga-Oka interfluve, a genetic time transect describes a population turnover from genetic profiles associated with the Uralic speakers, towards genetic profiles associated with the arrival of Indo-European Slavic languages (Peltola *et al.* 2023). The now-extinct Meryans could be a group of Uralic speakers who were inhabiting the region of Suzdal before the population turnover (Frog and Saarikivi 2015).

In particular circumstances, aDNA is extracted from human remains associated with specific cultural packages characterizing the archaeological site. These packages could also incorporate a linguistic identifier and suggest an association with an archaic or extinct language. This works particularly well in the presence of ancient written sources, either from historians of that time who described an association of a certain material culture and a language, or from ancient texts written in a certain language, sometimes on the artifacts themselves (e.g., inscriptions on stones, pottery, etc). When this information is available, we can interrogate aDNA about the demography of these people to infer historical and sociolinguistic dynamics, following the questions already outlined in section 13.3. Were these people diverse, were they coming from a large effective population size, and were they in continuity with previous and/or subsequent people living in the same region, or in other contemporary sites associated with the same material culture?

An informative case study is the one of the Etruscans, a population from central Italy who spoke a (now extinct) non-Indo-European language. According to historians of that time, the ancestors of the Etruscans belonged to the tribe of the Lydians. Because of extreme famine a group of Lydians chosen by lot had to leave their homelands in western Anatolia and settled in Italy (Herodotus, 1.94.5-7). Recently, a large aDNA transect study showed that Etruscans lacked a recent Anatolian-related admixture. They had been assimilating local ancestries, possibly of Italic origin, and have been also sharing genetic profiles with neighbouring Romans (Posth *et al.* 2021). These genetic results would not support Herodotus' narrative of an Anatolian origin of the Etruscan language but would be compatible with local development

with a few sources of admixture (Bonfante and Bonfante 2002). Other extinct languages left a trace in the historical written record. In the Italian peninsula, the Daunians of Apulia were characterized by a distinctive, cosmopolitan ancestry of autochthonous origin and possible Balkan influence, distinct from subsequent genetic ancestries from the Roman empire onwards (Aneli *et al.* 2022). The Daunian language was an Indo-European language related to the Messapic branch of difficult classification. Its possible origin in the Balkans (Matzinger 2005) would match the ancient genetic profile retrieved. Tocharian, known from written remains from Chinese Turkestan dating to the 1st millennium CE, is the sole member of a now extinct branch of Indo-European. Its speakers are possibly associated with ancient populations of the Tarim basin. An ancient DNA study described a Bronze Age Steppe-related migration into the broad region of Xinjiang during the 2nd Millennium BC, associated with the formation of ancestries in the region (Ning *et al.* 2019). Another aDNA study which more specifically analysed Tarim individuals from 2100–1700 BC did not report any traces of Steppe-related ancestry, but only local ancestry: the Tarim individuals are described as genetically isolated groups who adopted a pastoralist and agriculturalist culture (Zhang *et al.* 2021). Their connection to other early Indo-European speakers cannot therefore be traced with precision.

Another case is the one of internal genetic diversity and presence of diverse ancestries. The encounter between diverse groups in the same population can foster language shift and mixing (Table 13.1). Such cosmopolitan genetic contexts can be described for heterogeneous individual profiles through close burials in time and space. Large, cosmopolitan urban contexts have been described for specific archaeological settings, for example, those characterizing the Titicaca region during the Tiwanako period, possibly shaping the formation of the Aymara and/or Quechua languages and the linguistic contact between them (Nakatsuka *et al.* 2020), or those characterizing Rome and its colonies through time (Antonio *et al.* 2019), where Latin was exposed to contact with other languages.

Caveats should be posed in extrapolating population history from ancient remains. One of them is the fact that only a few individuals are retrieved per site, and they might be not representative of the whole population of speakers, or they could be migrants. This confounding effect can be mitigated by an accurate evaluation of the burial context. Objects associated with the human remains can inform us about the status of each individual in a burial site, for example identifying them as ruling elite. The migration of Germanic speaking groups into Britain was described as a massive turnover, with individuals of both sexes and a lack of specific designation of prestige (Gretzinger *et al.* 2022). In other circumstances, migrants can

be elite individuals, genetically distant from the rest of the autochthonous population. For example, the Hungarian population preserved a language from the Uralic family brought by the Magyars, who conquered the Carpathian Basin in the 9th century CE (Longobardi *et al.* 2015; Santos *et al.* 2020). Ancient DNA from elite Magyar individuals links them to an initial migration of speakers from Asia (Tömöry *et al.* 2007), precisely from an early admixture of Mansis, early Sarmatians, and descendants of late Xiongnus (Maróti *et al.* 2022). Because of the relatively small number of migrants, this genetic link got diluted through time, as present-day Hungarians are genetically indistinguishable from their Indo-European-speaking neighbours (Tambets *et al.* 2018; Santos *et al.* 2020; Maróti *et al.* 2022). Other fine-grained analyses of ancient burials are now able to reconstruct relationships between individuals that are informative for aspects of the society like kinship relationships, post-marital residence, sex-biased migrations, and even possible political alliances (Gretzinger *et al.* 2022; Rivollat *et al.* 2022; Villalba-Mouco *et al.* 2022; Skourtanioti *et al.* 2023). While this information is not directly relevant for linguistic reconstruction, it can still give insights into how permeable cultural and linguistic barriers are to gene flow and contact.

13.6 Conclusions

Joint efforts have by now produced large linguistic and cultural data sets which cover a large part of the known spectrum of linguistic and cultural diversity. At the same time, technical improvements, together with increasing efforts to include underrepresented ethnolinguistic groups, are extending the quality and quantity of genetic data available. When combined and analysed with adequate methods, these approaches may give deep insights into human history such as, for example, the constantly evolving distributions of languages and linguistic features in space and time. In terms of geographic coverage, we still see large gaps for example in the Americas, New Guinea, Australia, and sub-Saharan Africa, which might correspond to gaps of knowledge in human genetic history. It is difficult to assess if we are still missing crucial pieces in the puzzle of genetic history reconstruction, or if we are reaching an asymptote in the overall knowledge of existing genetic diversity.

Ancient DNA is providing more depth for projecting linguistic inferences into the past. The demographic connections suggested are necessarily indirect and should be filtered with evidence from other disciplines and robust historical information. The possibility to reconstruct population continuity vs. replacement through time in a given region is particularly valuable

for linguistic studies. It is one of the most simple and informative type of analyses performed with ancient DNA from different time depths, compared against modern DNA. Most of the examples included in section 13.5 come from the widely studied region of Western Eurasia, which sees a fortunate convergence of diverse cultures and ancient written sources, together with resources allocated for genetic, linguistic, anthropological, and archaeological research. Here the availability of aDNA allows for fine-grained analysis over the historical time scale, up to the Bronze-Age migrations. Above that limit, aDNA data becomes sparser, and linguistic associations become weaker, especially in the absence of ancient written texts and because of the limitations of linguistic reconstruction. Further informative examples are expected from regions of the world where research efforts and investments have been less abundant. In this very fecund time for genetic studies, more fine-grained reconstruction of past demography and society will be able to inform linguistic research with new sources of indirect inference, opening unexplored avenues in multidisciplinary studies of human history.

Acknowledgements

We thank Stuart Watson for proofreading the article. Chiara Barbieri was supported by the URPP "Evolution in Action" of the University of Zurich. Chiara Barbieri and Paul Widmer were supported by the NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888, and the SNSF Sinergia project "Out of Asia" (grant number CRSII5_183578).

References

Aguirre-Fernández, G., C. Barbieri, A. Graff, *et al.* (2021). 'Cultural Macroevolution of Musical Instruments in South America', *Humanit Soc Sci Commun* 8: 1-12.

Allen Ancient DNA Resource https://reich.hms.harvard.edu/allen-ancient-dna-resourceaadr-downloadable-genotypes-present-day-and-ancient-dna-data, version 54.1.p1. visited 30 april 2023.

Allentoft, M.E., M. Sikora, K.G. Sjögren, *et al.* (2015.) 'Population Genomics of Bronze Age Eurasia', *Nature* 522: 167-72.

Altshuler, D.M., R.A. Gibbs, L. Peltonen, *et al.* (2010). 'Integrating Common and Rare Genetic Variation in Diverse Human Populations', *Nature* 467: 52-8.

Andrews, R.M., I. Kubacka, P.F. Chinnery, *et al.* (1999). 'Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA', *Nature genetics* 23: 147.

Aneli, S., G. Birolo, G. Matullo (2022). 'Twenty Years of the Human Genome Diversity Project', *Human Population Genetics and Genomics* 2 DOI: 10.47248/hpgg2202040005.

Aneli, S. T. Saupe, F. Montinaro *et al.* (2022). 'The Genetic Origin of Daunians and the Pan-Mediterranean Southern Italian Iron Age Context', *Molecular Biology and Evolution* 39: msac014.

Antonio, M.L., Z. Gao, H.M. Moots *et al.* (2019). 'Ancient Rome: A Genetic Crossroads of Europe and the Mediterranean', *Science* 366: 708-14.

Arango-Isaza *et al.* (2023). 'The Genetic History of the Southern Andes from Present-Day Mapuche Ancestry', *Current Biology* Provisionally Accepted.

Arias, L., C. Barbieri, G. Barreto, *et al.* (2018). 'High-Resolution Mitochondrial DNA Analysis Sheds Light on Human Diversity, Cultural Interactions, and Population Mobility in Northwestern Amazonia', *American Journal of Physical Anthropology* 165: 238-55.

Atkinson, Q.D., R.D. Gray (2005). 'Curious Parallels and Curious Connections – Phylogenetic Thinking in Biology and Historical Linguistics', *Systematic Biology* 54: 513-26.

Auton, A., G.R. Abecasis, D.M. Altshuler, *et al.* (2015). 'A Global Reference for Human Genetic Variation', *Nature* 526: 68-74

Bachtrog, D., and B. Charlesworth (2001). 'Towards a Complete Sequence of the Human Y Chromosome', *Genome Biology* 2: reviews1016.1.

Barbieri, C., D.E. Blasi, E. Arango-Isaza, *et al.* (2022). 'A Global Analysis of Matches and Mismatches between Human Genetic and Linguistic Histories', *Proceedings of the National Academy of Sciences* 119: e2122084119.

Barbieri, C., A. Butthof, K. Bostoen K *et al.* (2013). 'Genetic Perspectives on the Origin of Clicks in Bantu Languages from Southwestern Zambia', *Eur J Hum Genet* 21:430–6.

Barbieri, C., P. Heggarty, L. Castrì, *et al.* (2011). 'Mitochondrial DNA Variability in the Titicaca Basin: Matches and Mismatches with Linguistics and Ethnohistory', *American Journal of Human Biology* 23: 89–99.

Barbieri, C., P. Heggarty, D. Yang Yao, *et al.* (2014). 'Between Andes and Amazon: The Genetic Profile of the Arawak-speaking Yanesha', *American Journal of Physical Anthropology* 155: 600–9.

Barbujani, G., A. Pilastro (1993). 'Genetic Evidence on Origin and Dispersal of Human Populations Speaking Languages of the Nostratic Macrofamily', *Proceedings of the National Academy of Sciences of the United States of America* 90: 4670–3.

Barbujani, G., R.R. Sokal (1990). 'Zones of Sharp Genetic Change in Europe are also Linguistic Boundaries', *Proceedings of the National Academy of Sciences of the United States of America* 87: 1816–9.

Batsuren, K, G. Bella, F. Giunchiglia (2022). 'A Large and Evolving Cognate Database', *Lang Resources & Evaluation* 56: 165–89.

Behar, D.M., S. Rosset, J. Blue-Smith, *et al.* (2007). 'The Genographic Project Public Participation Mitochondrial DNA Database', *PLOS Genetics* 3:e104.

Behar, D.M., M. Van Oven, S. Rosset, *et al.* (2012). 'A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from its Root', *American Journal of Human Genetics* 90: 675–84.

Belle, E.M.S., G. Barbujani (2007). 'Worldwide Analysis of Multiple Microsatellites: Language Diversity Has a Detectable Influence on DNA Diversity', *American Journal of Physical Anthropology* 133: 1137–46.

Bentz, C., B. Winter (2013). 'Languages with More Second Language Learners Tend to Lose Nominal Case', *Language Dynamics and Change* 3:1–27.

Bergström, A., S.A. McCarthy, R. Hui, *et al.* (2020). 'Insights into Human Genetic Variation and Population History from 929 Diverse Genomes', *Science* 367: eaay5012.

Bickel, B., J. Nichols, T. Zakharko, *et al.* (2022). *The AUTOTYP database.* DOI: 10.5281/zenodo.5931509.

Bird, N., L. Ormond, P. Awah, *et al.* (2023). 'Dense Sampling of Ethnic Groups within African Countries Reveals Fine-Scale Genetic Structure and Extensive Historical Admixture', *Science Advances* 9:eabq2616.

Bisso-Machado, R., N.J.R. Fagundes (2021). 'Uniparental Genetic Markers in Native Americans: A Summary of All Available Data from Ancient and Contemporary Populations', *American Journal of Physical Anthropology* 176:445–58.

Blasi, D.E., S. Moran, S.R. Moisik, *et al.* (2019). 'Human Sound Systems are Shaped by Post-Neolithic Changes in Bite Configuration', *Science* 363:eaav3218.

Bolnick, D.A.W., B.A.S. Shook, L. Campbell, *et al.* (2004). 'Problematic Use of Greenberg's Linguistic Classification of the Americas in Studies of Native American Genetic Variation', *American journal of human genetics* 75: 519–22.

Bonfante, G., L. Bonfante (2002). *The Etruscan Language: An Introduction, Revised Editon*. Manchester University Press.

Bryc, K., E.Y. Durand, J.M. Macpherson, *et al.* (2015). 'The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States', *The American Journal of Human Genetics* 96: 37–53.

Byrska-Bishop, M., U.S. Evani, X. Zhao, *et al.* (2022). 'High-coverage Whole-Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios', *Cell* 185: 3426-3440.e19.

Calafell, F., M. Larmuseau (2017). 'The Y Chromosome as the Most Popular Marker in Genetic Genealogy Benefits Interdisciplinary Research', *Human Genetics* 136: 559–73.

Capocasa, M., C. Battaggia, P. Anagnostou, *et al.* (2013). 'Detecting Genetic Isolation in Human Populations: A Study of European Language Minorities', *PLOS ONE* 8:e56371.

Carling, G. (ed.) (2017). *Diachronic Atlas of Comparative Linguistics Online*.

Cavalli-Sforza, L.L., E. Minch, J.L. Mountain (1992). 'Coevolution of Genes and Languages Revisited', *Proceedings of the National Academy of Sciences of the United States of America* 89: 5620–4.

Cavalli-Sforza, L.L., A. Piazza, P. Menozzi, *et al.* (1988). 'Reconstruction of Human Evolution: Bringing Together Genetic, Archaeological, and Linguistic Data', *Proceedings of the National Academy of Sciences of the United States of America* 85:6002–6.

Chaubey, G., M. Metspalu, M. Karmin, *et al.* (2008). 'Language Shift by Indigenous Population: A Model Genetic Study in South Asia', *International Journal of Human Genetics* 8:41–50.

Claw, K.G., M.Z. Anderson, R.L. Begay, *et al.* (2018). 'A Framework for Enhancing Ethical Genomic Research with Indigenous Communities', *Nat Commun* 9:2957.

Creanza, N., M. Ruhlen, T.J. Pemberton, *et al.* (2015). 'A Comparison of Worldwide Phonemic and Genetic Variation in Human Populations', *Proceedings of the National Academy of Sciences of the United States of America* 112:1265–72.

Darwin, C. (1859). *On the Origin of Species*. London: Murray.

Diamond, J., P. Bellwood (2003). 'Farmers and Their Languages: The First Expansions', *Science* 300:597–603.

Dryer, M., M. Haspelmath (2020). cldf-datasets/wals: The World Atlas of Language Structures Online. DOI: 10.5281/zenodo.3731125.

Durkin, P. (2014). *Borrowed Words*. Oxford University Press.

Efrat-Kowalsky, N., P. Ranacher, N. Neureiter, *et al.* (2022). 'Oldest Attested Languages in the Near East Reveal Deep Transformations in Linguistic Landscapes', *Scientific Reports* 2022.

Evans, N. (2018). Did language evolve in multilingual settings? *Biology and Philosophy* 32.10.1007/s10539-018-9609-3.

Evans, N. (2019). Linguistic divergence under contact. In Cennamo M. & Fabrizio C. (eds), *Selected Papers from the 22nd International Conference on Historical Linguistics*. Amsterdam/Philadelphia: John Benjamins. Pp. 563-591.

de Filippo, C., M. Whitten, *et al.* (2011). 'Y-chromosomal Variation in Sub-Saharan Africa: Insights into the History of Niger-Congo Groups', *Molecular Biology and Evolution* 28: 1255.

Finney, M.A. (2013). 'Krio', In: Michaelis SM, Maurer P, Haspelmath M, *et al.* (eds.). *The Survey of Pidgin and Creole Languages. Vol. I: English-Based and Dutch-Based Languages*. Oxford: Oxford University Press, 157–66.

Flores-Bello, A., F. Bauduer, J. Salaberria, *et al.* (2021). 'Genetic Origins, Singularity, And Heterogeneity of Basques', *Current Biology* 31:2167-2177.e4.

Forster, P., C. Renfrew (2011). 'Mother Tongue and Y Chromosomes', *Science* 333:1390–1.

Fortes-Lima, C.A., C. Burgarella, R. Hammarén, *et al.* (2023). 'The Genetic Legacy of the Expansion of Bantu-Speaking Peoples in Africa', 2023.04.03.535432.

Frog, M., J. Saarikivi (2015). 'De Situ Linguarum Fennicarum Aetatis Ferreae: Pars I.', *RMN Newsletter*.

Gijn, R. van, J. Case, M. Bruil, *et al.* (2023). 'Lexically Driven Patterns of Contact in Alignment Systems of Languages of the Northern Upper Amazon', *Open Linguistics* 9, DOI: 10.1515/opli-2022-0224.

Goodwin, S., J.D. McPherson, W.R. McCombie (2016). 'Coming of Age: Ten Years of Next-Generation Sequencing Technologies', *Nat Rev Genet* 17:333–51.

Green, R.E., J. Krause, A.W. Briggs, *et al.* (2010). 'A Draft Sequence of the Neandertal Genome', *Science* 328:710–22.

Greenberg, J.H. (1987). Language in the Americas, Stanford University Press.

Greenhill, S.J., X. Hua, C.F. Welsh, *et al.* (2018). 'Population Size and the Rate of Language Evolution: A Test Across Indo-European, Austronesian, and Bantu Languages', *Frontiers in Psychology* 9:576.

Gretzinger, J., D. Sayer, P. Justeau, *et al.* (2022). 'The Anglo-Saxon Migration and the Formation of the Early English Gene Pool', *Nature* 610:112–9.

Herodotus, 1957. Herodotus in four volumes. Harvard University Press, Cambridge (Mass.).

Haak, W., I. Lazaridis, N. Patterson, *et al.* (2015). 'Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe', *Nature* 522:207–11.

Hallast, P., C. Batini, D. Zadik, *et al.* (2015). 'The Y-chromosome Tree Bursts into Leaf: 13,000 High-Confidence Snps Covering the Majority of Known clades', *Molecular biology and evolution* 32:661–73.

Hammarström, H., R. Forkel, M. Haspelmath, et al. (2022). Glottolog 4.7. Leipzig.

Haspelmath, M. (2001). The European linguistic area: Standard Average European. In Martin Haspelmath (Ed.), Language typology and language universals. (Handbücher zur Sprach- und Kommunikationswissenschaft) (pp. 1492-1510). Berlin: de Gruyter.

Heggarty, P. (2014). 'Prehistory through Language and Archaeology', in The Routledge Handbook of Historical Linguistics. Routledge, 598–626.

Hellenthal, G., G.B.J. Busby, G. Band, *et al.* (2014). 'A Genetic Atlas of Human Admixture History', *Science* 343:747–51.

Heyer, E., R. Chaix, S. Pavard, *et al.* (2012). 'Sex-specific Demographic Behaviours that Shape Human Genomic Variation', *Molecular Ecology* 21:597–612.

Hua, X., S.J. Greenhill, M. Cardillo, *et al.* (2019). 'The Ecological Drivers of Variation in Global Language Diversity', *Nature Communications* 10:2047.

Hudson, M., N.A. Garrison, R. Sterling, *et al.* (2020). 'Rights, Interests and Expectations: Indigenous Perspectives on Unrestricted Access to Genomic Data', *Nat Rev Genet* 21:377–84.

Jakobsson, M., S.W. Scholz, P. Scheet, *et al.* (2008). 'Genotype, Haplotype and Copy-Number Variation in Worldwide Human Populations', *Nature* 451:998–1003.

Jobling, M.A., C. Tyler-Smith (2017). 'Human Y-chromosome Variation in the Genome-Sequencing Era', *Nat Rev Genet* 2017;18:485–97.

Kampourakis, K., E.L. Peterson (2023). 'The Racist Origins, Racialist Connotations, and Purity Assumptions of the Concept of "Admixture" in Human Evolutionary Genetics', *Genetics* 223: iyad002.

Kandler, A., R. Unger, J. Steele (2010). 'Language Shift, Bilingualism and the Future of Britain's Celtic languages', *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 365:3855–64.

Kayser, M., Y. Choi, M. van Oven, *et al.* (2008). 'The Impact of the Austronesian Expansion: Evidence from mtDNA and Y Chromosome Diversity in the Admiralty Islands of Melanesia', *Molecular biology and evolution* 25:1362–74.

Kayser, M., P. de Knijff (2011). 'Improving Human Forensics through Advances in Genetics, Genomics and Molecular Biology', *Nat Rev Genet* 12:179–92.

Kelleher, J., Y. Wong, A.W. Wohns, *et al.* (2019). 'Inferring Whole-Genome Histories in Large Population Datasets', *Nat Genet* 51:1330–8.

Kircher, M., J. Kelso (2010). 'High-throughput DNA Sequencing – Concepts and Limitations', *BioEssays* 32:524–36.

Lachance, J., S.A. Tishkoff (2013). 'SNP Ascertainment Bias in Population Genetic Analyses: Why it is Important, and How to Correct It', *BioEssays* 35:780–6.

Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A.G., Gieger, C., Wichmann, H.-E., Rüther, A., Schreiber, S., Becker, C., Nürnberg, P., Nelson, M.R., Krawczak, M., Kayser, M. (2008). Correlation between Genetic and Geographic Structure in Europe. *Current Biology* 18, 1241–1248. https://doi.org/10.1016/j.cub.2008.07.049

Lazaridis, I., S. Alpaslan-Roodenberg, A. Acar, *et al.* (2022). 'The Genetic History of the Southern Arc: A Bridge between West Asia and Europe', *Science* 377:eabm4247.

Leslie, S., B. Winney, G. Hellenthal, *et al.* (2015). 'The Fine-Scale Genetic Structure of the British Population', *Nature* 519:309–14.

Levinson, S.C., R.D. Gray (2012). 'Tools from Evolutionary Biology Shed New Light on the Diversification of Languages', *Trends in Cognitive Sciences* 16:167–73.

Lewis, C.M., R.Y. Tito, B. Lizárraga, *et al.* (2005). 'Land, Language, and Loci: mtDNA in Native Americans and the Genetic History of Peru', *American Journal of Physical Anthropology* 127:351–60.

Lewis, M.P. (2009). *Ethnologue: Languages of the World*. SIL International Dallas, TX.

Lindo, J., A. Achilli, U.A. Perego, *et al.* (2017). 'Ancient Individuals from the North American Northwest Coast Reveal 10,000 Years of Regional Genetic Continuity', *Proceedings of the National Academy of Sciences of the United States of America* 114:4093–8.

Lippold, S., H. Xu, A. Ko, *et al.* (2014). 'Human Paternal and Maternal Demographic Histories: Insights from High-Resolution Y Chromosome and mtDNA Sequences', *Investigative genetics* 5:13.

Lipson, M., E.A. Sawchuk, J.C. Thompson, *et al.* (2022). 'Ancient DNA and Deep Population Structure in Sub-Saharan African Foragers', *Nature* 603:290–6.

Lipson, M., P. Skoglund, M. Spriggs, *et al.* (2018). 'Population Turnover in Remote Oceania Shortly after Initial Settlement', *Current Biology* 28:1157-1165.e7.

Liu, Y., X. Mao, J. Krause, *et al.* (2021). 'Insights into Human History from the First Decade of Ancient Human Genomics', *Science* 373:1479–84.

Llamas, B., L. Fehren-Schmitz, G. Valverde, *et al.* (2016). 'Ancient Mitochondrial DNA Provides High-Resolution Timescale of the Peopling of the Americas', *Science Advances* 2:1–10.

Llamas, B., E. Willerslev, L. Orlando (2017). 'Human Evolution: A Tale from Ancient Genomes', *Philosophical Transactions of the Royal Society B: Biological Sciences* 372:20150484.

Longobardi, G., S. Ghirotto, C. Guardiano, *et al.* (2015). 'Across Language Families: Genome Diversity Mirrors Linguistic Variation within Europe', *American Journal of Physical Anthropology* 157:630–40.

Lupyan, G., R. Dale (2010). 'Language Structure is Partly Determined by Social Structure', *PLoS ONE* 5:e8559.

Mace, R., C. Holden (2005). 'A Phylogenetic Approach to Cultural Evolution', *Trends in ecology & evolution* 20:116-121.

Malhi, R. (2009). 'Implications of the Genographic Project for Molecular Anthropologists', *International Journal of Cultural Property* 16:193–4.

Mallick, S., H, Li, M. Lipson, *et al.* (2016). 'The Simons Genome Diversity Project: 300 Genomes from 142 Diverse Populations', *Nature* 538:201–6.

Maróti, Z., E. Neparáczki, O. Schütz, *et al.* (2022). 'The Genetic Origin of Huns, Avars, and Conquering Hungarians', *Current Biology* 32:2858-2870.e7.

Mathieson, I., I. Lazaridis, N. Rohland, *et al.* (2015). 'Genome-wide Patterns of Selection in 230 Ancient Eurasians', *Nature* 528:499–503.

Mathieson, I., G. McVean (2014). 'Demography and the Age of Rare Variants', *PLoS Genetics* 10:e1004528.

Matsumae, H., P. Ranacher, P.E. Savage *et al.* (2021). 'Exploring Correlations in Genetic and Cultural Variation across Language Families in Northeast Asia', *Science Advances* 7:9223–41.

Matzinger, J. (2005). 'Messapisch und Albanisch', *International Journal of Diachronic Linguistics and Linguistic Reconstruction* 2:29–54.

Mesoudi, A., A. Whiten, K.N. Laland (2006). 'Towards a Unified Science of Cultural Evolution', *Behavioral and Brain Sciences* 29:329–47.

Metzker, M.L. (2010). 'Sequencing Technologies — the Next Generation', *Nat Rev Genet* 11:31–46.

Mona, S., K.E. Grunz, S. Brauer, *et al.* (2009). 'Genetic Admixture History of Eastern Indonesia as Revealed by Y-Chromosome and Mitochondrial DNA Analysis', *Molecular biology and evolution* 26:1865–77.

Moran, S., E. Grossman, A. Verkerk (2021). 'Investigating Diachronic Trends in Phonological Inventories Using BDPROTO', *Lang Resources & Evaluation* 55:79–103.

Moran, S., D. McCloy (eds.) (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Moran, S., D. McCloy, R. Wright (2012). 'Revisiting Population Size Vs. Phoneme Inventory Size', *Language* 88:877–93.

Mourier, T., S.Y.W. Ho, M.T.P. Gilbert, *et al.* (2012). 'Statistical Guidelines for Detecting Past Population Shifts Using Ancient DNA', *Molecular Biology and Evolution* 29:2241–51.

Nakatsuka, N., I. Lazaridis, C. Barbieri,*et al.* (2020). 'A Paleogenomic Reconstruction of the Deep Population History of the Andes', *Cell* 181:1131-1145.e21.

Nasidze, I., M. Stoneking (2001). 'Mitochondrial DNA Variation and Language Replacements in the Caucasus', *Proceedings of the Royal Society B: Biological Sciences* 268:1197–206.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*.

Nielsen, R., J.M. Akey, M. Jakobsson, *et al.* (2017). 'Tracing the Peopling of the World Through Genomics', *Nature* 541:302–10.

Ning, C., C.-C. Wang, S. Gao, *et al.* (2019). 'Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan', *Current Biology* 29:2526-2532.e4.

O'Fallon, B.D., L. Fehren-Schmitz (2011). 'Native Americans Experienced a Strong Population Bottleneck Coincident with European Contact', *Proceedings of the National Academy of Sciences of the United States of America* 108:20444–8.

Olalde, I., S. Brace, M.E. Allentoft, *et al.* (2018). 'The Beaker Phenomenon and the Genomic Transformation of Northwest Europe', *Nature* 555:190–6.

Olalde, I., C. Posth (2020). 'Latest Trends in Archaeogenetic Research of West Eurasians', *Current Opinion in Genetics & Development* 62:36–43.

Orlando, L., M.T.P. Gilbert, E. Willerslev (2015). 'Reconstructing Ancient Genomes and Epigenomes', *Nat Rev Genet* 16:395–408.

van Oven, M. (2015). 'PhyloTree Build 17: Growing the Human Mitochondrial DNA tree', *Forensic Science International: Genetics Supplement Series* 5:e392–4.

Pagel, M. (2017). 'Q&A: What is Human Language, When Did It Evolve and Why Should We Care?', *BMC Biol* 15:64.

Pakendorf, B. (2014). 'Coevolution of Languages and Genes', *Current opinion in genetics & development* 29:39–44.

Pakendorf, B., M. Stoneking (2005). 'Mitochondrial DNA and Human Evolution', *Annual Review of Genomics and Human Genetics* 6:165–83.

Passmore, S., Barth, W., Greenhill, S.J., *et al.* (2023). Kinbank: A global database of kinship terminology. PLOS ONE 18, e0283218. https://doi.org/10.1371/journal.pone.0283218

Patterson, N., P. Moorjani, Y. Luo, *et al.* (2012). 'Ancient Admixture in Human History', *Genetics* 192:1065–93.

Peltola, S., K. Majander, N. Makarov, *et al.* (2023). 'Genetic Admixture and Language Shift in the Medieval Volga-Oka Interfluve', *Current Biology* 33:174-182.e10.

Pena, S.D.J., F.R. Santos, E. Tarazona-Santos (2020). 'Genetic Admixture in Brazil', *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 184:928–38.

Penny, D., E.E. Watson, M.A. Steel (1993). 'Trees from Languages and Genes are Very Similar', *Systematic Biology* 42:382–4.

Pickrell, J.K., N. Patterson, C. Barbieri, *et al.* (2012). 'The Genetic Prehistory of Southern Africa', *Nature Communications* 3:1143.

Pickrell, J.K., D. Reich (2014). 'Toward a new History and Geography of Human Genes Informed by Ancient DNA', *Trends in Genetics* 30:377–89.

Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.

Posth, C., K. Nägele, H. Colleran, *et al.* (2018). 'Language Continuity Despite Population Replacement in Remote Oceania', *Nature Ecology & Evolution* 2:731–40.

Posth, C., G. Renaud, A. Mittnik, *et al.* (2016). 'Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe', *Current Biology* 26:827–33.

Posth, C., V. Zaro, M.A. Spyrou, *et al.* (2021). 'The Origin and Legacy of the Etruscans Through a 2000-Year Archeogenomic Time Transect', *Science Advances* 7:7673–97.

Prugnolle, F., A. Manica, F. Balloux (2005). 'Geography Predicts Neutral Genetic Diversity of Human Populations', *Current Biology* 15:R159–60.

Pugach, I., A.T. Duggan, D.A. Merriwether, *et al.* (2018). 'The Gateway from Near into Remote Oceania: New Insights from Genome-Wide Data', *Molecular Biology and Evolution* 35:871–86.

Pugach, I., M. Stoneking (2015). 'Genome-wide Insights into the Genetic History of Human Populations', *Investigative Genetics* 6:6.

Ragoussis, J. (2009). 'Genotyping Technologies for Genetic Research', *Annual Review of Genomics and Human Genetics* 10:117–33.

Ralph, P., G. Coop (2013). 'The Geography of Recent Genetic Ancestry across Europe', *PLOS Biology* 11:e1001555.

Reich, D., R.E. Green, M. Kircher, *et al.* (2010). 'Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia', *Nature* 468:1053–60.

Renfrew, C., P. Bellwood (2002). *Examining the Farming/Language Dispersal Hypothesis*. University of Cambridge: McDonald Institute for Archaeological Research.

Rivollat, M., A. Thomas, E. Ghesquière, *et al.* (2022). 'Ancient DNA Gives New Insights into a Norman Neolithic Monumental Cemetery Dedicated to Male Elites', *Proceedings of the National Academy of Sciences* 119:e2120786119.

Roewer, L., W. Parson (2012). 'Internet Accessible Population Databases: YHRD and EMPOP', *Encyclopedia of Forensic Sciences: Second Edition*, 357–64.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, *et al.* (2002). 'Genetic Structure of Human Populations', *Science* 298:2381–5.

Rousset, F. (1997). 'Genetic Differentiation and Estimation of Gene Flow from F-Statistics under Isolation by Distance', *Genetics* 145:1219–28.

Ruhlen, M. (1991). A Guide to the World's Languages: Volume I, Classification. Stanford University Press.

Santos, P., G. Gonzàlez-Fortes, E. Trucchi, *et al.* (2020). 'More Rule than Exception: Parallel Evidence of Ancient Migrations in Grammars and Genomes of Finno-Ugric Speakers', *Genes* 11:1491.

Schiffels, S., R. Durbin (2014). 'Inferring Human Population Size and Separation History From Multiple Genome Sequences', *Nat Genet* 46:919–25.

Schönberg, A., C. Theunert, M. Li, *et al.* (2011). 'High-throughput Sequencing of Complete Human mtDNA Genomes from the Caucasus and West Asia: High Diversity and Demographic Inferences', *European Journal of Human Genetics* 19:988–94.

Schrijver, P. (2013). *Language Contact and the Origins of the Germanic Languages*. Routledge.

Schurr, T.G., S.T. Sherry (2004). 'Mitochondrial DNA and Y Chromosome Diversity and the Peopling of the Americas: Evolutionary and Demographic Evidence', *American Journal of Human Biology* 16:420–39.

Skirgård, H., H.J. Haynie, D.E. Blasi, *et al.* (2023). 'Grambank Reveals the Importance of Genealogical Constraints on Linguistic Diversity and Highlights the Impact of Language Loss', *Science Advances* 9:eadg6175.

Skourtanioti, E., H. Ringbauer, G.A. Gnecchi Ruscone, *et al.* (2023). 'Ancient DNA Reveals Admixture History and Endogamy in the Prehistoric Aegean', *Nat Ecol Evol* 7:290–303.

Slatkin, M., F. Racimo (2016). 'Ancient DNA and Human History', *Proceedings of the National Academy of Sciences of the United States of America* 113:6380–7.

Sokal, R.R. (1988). 'Genetic, Geographic, and Linguistic Distances in Europe', *Proceedings of the National Academy of Sciences of the United States of America* 85:1722–6.

Sokal, R.R., N.L. Oden, B.A. Thomson (1992). 'Origins of the Indo-Europeans: Genetic Evidence', *Proceedings of the National Academy of Sciences of the United States of America* 89: 7669–73.

Steele, J., A. Kandler (2010). 'Language Trees ≠ Gene Trees', *Theory in Biosciences* 129:223–33.

Steiner, L., P.F. Stadler, M. Cysouw (2011). 'A Pipeline for Computational Historical Linguistics', *Language Dynamics and Change* 1:89–127.

Tambets, K., B. Yunusbayev, G. Hudjashov, *et al.* (2018). 'Genes Reveal Traces of Common Recent Demographic History for Most of the Uralic-Speaking Populations', *Genome Biology* 19:139.

Tamm, E., T. Kivisild, M. Reidla, *et al.* (2007). 'Beringian Standstill and Spread of Native American Founders', *PLoS One* 2:e829.

Teixidor-Toneu, I., F.M. Jordan, J.A. Hawkins (2018). 'Comparative Phylogenetic Methods and the Cultural Evolution of Medicinal Plant Use', *Nature Plants* 4:754–61.

Tëmkin, I., N. Eldredge (2007). 'Phylogenetics and Material Cultural Evolution', *Current Anthropology* 48:146–54.

Tömöry, G., B. Csányi, E. Bogácsi-Szabó, *et al.* (2007). 'Comparison of Maternal Lineage and Biogeographic Analyses of Ancient and Modern Hungarian Populations', *American Journal of Physical Anthropology* 134:354–68.

Torroni, A., A. Achilli, V. Macaulay *et al.* (2006). 'Harvesting the Fruit of the Human mtDNA Tree', *TRENDS in Genetics* 22:339–45.

Trudgill, P. (2002). 'Linguistic and Social Typology', In: Chambers J, Trudgill P, Schilling-Estes N (eds.). *The Handbook of Language Variation and Change*. Oxford: Blackwell Pub, 707–28.

Turchin, P., T.E. Currie, H. Whitehouse, *et al.* (2018). 'Quantitative Historical Analysis Uncovers a Single Dimension of Complexity that Structures Global Variation in Human Social Organization', *Proceedings of the National Academy of Sciences* 115:E144–51.

Underhill, P.A., T. Kivisild (2007). 'Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations'.

Vicente, M., C.M. Schlebusch (2020). 'African Population History: An Ancient DNA Perspective', *Current Opinion in Genetics & Development* 62:8–15.

Vigilant, L., M. Stoneking, H. Harpending, *et al.* (1991). 'African Populations and the Evolution of Human Mitochondrial DNA', *Science* 253:1503.

Villalba-Mouco, V., C. Oliart, C. Rihuete-Herrada, *et al.* (2022). 'Kinship Practices in the Early State El Argar Society from Bronze Age Iberia', *Sci Rep* 12:22415.

Wang, S., C.M. Lewis Jr, M. Jakobsson, *et al.* (2007). 'Genetic Variation and Population Structure in Native Americans', *PLoS Genet* 3:e185.

White, N. (1997). Genes, languages and landscapes in Australia. In McConvell, P. and Evans, N. (eds.). *Archaeology and linguistics: Aboriginal Australia in global perspective*, 45-82. Oxford: Oxford University Press.

Wichmann, S., C.H. Brown, E.W. Holman (eds). (2022). The ASJP Database. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Wichmann, S., and E.W. Holman (2009). 'Population size and rates of language change', *Human Biology* 81: 259–74.

Williams, J.R. (2008). 'The Declaration of Helsinki and Public Health', *Bull World Health Organ* 86:650–2.

WMA - The World Medical Association-*Declaration of Helsinki* 1964.

Wood, A.L.C., K.R. Kirby, C.R. Ember, *et al.* (2022). 'The Global Jukebox: A public Database of Performing Arts and Culture', *PLOS ONE* 17:e0275469.

Zhang, F., C. Ning, A. Scott, *et al.* (2021). 'The Genomic Origins of the Bronze Age Tarim Basin Mummies', *Nature* 599: 256–61.

Zhang, M., H.-X. Zheng, S. Yan, *et al.* (2019). 'Reconciling the Father Tongue and Mother Tongue Hypotheses in Indo-European Populations', *National Science Review* 6:293–300.