Chapter 45

Transeurasian unity from a population genetic perspective

Choongwon Jeong, Chuan-Chao Wang and Chao Ning

Abstract

Contemporary Transeurasian-speaking populations reside in a wide geographic area encompassing the Eurasian steppe, Northeast China and the Russian Far East, as well as Korean peninsula and the Japanese archipelago. From population genetic studies of contemporary Transeurasian-speakers, it is well known that they are genetically heterogeneous, probably due to historical mixing with non-Transeurasian populations during their migration. Here, we provide an up-to-date overview of the genetic relationship among Transeurasian populations. Specifically, we focus on highlighting i) studies of contemporary populations explicitly taking into account the above stated recent admixture, and ii) paleogenomic studies of ancient genomes to directly recover prehistoric gene pools predating recent admixture events. These studies show an underlying shared genetic substratum among the Transeurasian populations, which is best represented by ancient populations from Northeast China and the Russian Far East, as well as present-day Tungusic-speaking populations.

Keywords: paleogenomics, admixture, Transeurasian, population genetics, Northeast China

45.1 Introduction

The human genome contains an enormous amount of information on the past demographic history of a population to which an individual belongs, in the form of genetic variation. In the past few decades, molecular anthropologists have made great efforts to infer the genetic history of world-wide populations by analyzing multiple types of genetic materials; maternally inherited mitochondrial DNA (mtDNA), paternally inherited Y chromosome, and autosomal and X chromosomal DNA. Autosomes and X chromosome are inherited from both of the parents (except for males inheriting their X chromosome only from mother). They also experience recombination every generation, an exchange of chromosomal segments between two copies of chromosome, while gametes (eggs and sperms) are formed. In contrast, the mtDNA and non-recombining portion of Y chromosome are strictly inherited uniparentally without recombination and have been intensively studied since the mid-1980s to trace human past (Cavalli-Sforza and Feldman 2003; Jobling and Tyler-Smith 2003). In addition to the simple inheritance pattern, they also have a large number of genetic variants tagged with a lineage and evolve faster due to high mutation rate and small effective population size. These features help to generate a strong phylogeographic signal. For example, early studies on the mtDNA and Y chromosomal genetic variation found a pattern matching well with the recent out-of-Africa hypothesis of anatomically modern human origin, but not with the multiregional hypothesis, thus resolving decades-old debate (Cann et al. 1987; Hammer 1995).

Molecular anthropologists have constructed comprehensive haplogroup trees for mtDNA and Y chromosome following a hierarchical nomenclature system by assigning combinations of letters and numbers as haplogroup names (Y Chromosome Consortium

2002; Van Oven and Kayser 2009). A haplogroup represents a lineage within the mtDNA or Y chromosome phylogeny defined by a set of shared mutations. The most striking feature of mtDNA and Y chromosome is their population or region-specific haplogroup distribution. The mtDNA haplogroup H, J, K, T, V, X and U are found mainly in West Eurasia while mtDNA lineage A, B, CZ, D, E, F, and Y are common in East Eurasia. Y chromosomal haplogroup E, G, H, I, J, R and T are prevalent in West Eurasia, but C, D, and O are largely distributed in East Asia (Torroni et al. 2006; Karafet et al. 2008). The more detailed phylogeny can give better resolutions in resolving questions at population or even family-pedigree scale, such as the Mongolian expansion shown by a star-like Y chromosomal haplogroup C3*xC3c (Zerjal et al. 2003), and the early migration of the ancestors of the Aisin Gioro clan from the middle reaches of Amur River to southeastern Manchuria revealed by Y chromosomal lineage C3b1a3a2-F8951 (Wei et al. 2017). More importantly, the genetic patterns in human societies are often influenced by their cultural practices, such as residence patterns and language changes, which can be fortunately inferred by investigating the uniparental markers. For instance, the genetic and language correlation studies over the past few years favor the sex-specific scenario of language change. Some researchers hypothesized that the language change in an already-populated region may require a minimum proportion of male immigrants based on strong association of languages with Y haplogroups but not with maternal mtDNA (Forster and Renfrew 2011), although others do not agree with the language-gene association (Pakendorf 2014a).

Recent advances in genomic technology provide a high-throughput access to genetic variation in the nuclear genome (autosomal and X/Y chromosomal DNA). For technological reasons, single nucleotide polymorphisms (SNPs), a single base pair difference at a certain position of the genome, have become the most widely used type of genetic variation data (Wang et al. 1998; Sachidanandam et al. 2001). A single SNP contains only a very small amount of noisy demographic information. However, each of them importantly contains a small bit of independent information due to recombination events separating them. Therefore, by combining information from millions of SNPs scattered across the genome, one can obtain an accurate portrait of key demographic events of the population of interest. For example, a single diploid genome can be used to reconstruct past changes in the effective population size (N_e), a population genetic parameter reflecting census population size as well as population structure, back to hundreds of thousands of years even prior to the origin of our own species (Li and Durbin 2011). Genome-wide variation data were also used to identify low levels of genetic contribution from archaic hominins to non-Africans, which was unable to be detected by mitochondrial studies (Green et al. 2010; Reich et al. 2010).

Population geneticists have long studied the genetic history of ethnic groups speaking Transeurasian languages ("Transeurasians" for the rest of this chapter), which is composed of Turkic, Mongolic, Tungusic, Koreanic and Japonic language families (Robbeets 2005; Johanson and Robbeets 2010a). Often, these efforts focused on a subset of Transeurasians, often a specific language family, for investigating the signature of genetic admixture, a mixing of genetically distinct populations, which resulted from cultural events of interest. For example, several genomic studies highlighted Turkic and Mongolic populations to characterize and date their admixture (a mixture of two or more divergent gene pools) with nearby western Eurasian populations during their westward expansions (Hellenthal et al. 2014; Yunusbayev et al. 2015). Such expansions were associated with the historical expansions of nomadic empires, such as Genghis Khan's one or earlier Turkic ones (Hellenthal et al. 2014). Another example is the admixture in the contemporary Japanese, who are modeled as mixed descendants of indigenous Jomon hunter-gatherers and immigrating Yayoi farmers, the latter of which brought paddy-field rice farming and metallurgy into the archipelago, probably from the Korean peninsula (Hanihara 1991; Habu 2004; Jinam et al. 2012). On the other hand, several southern Siberian populations, especially Turkic and Yeniseian speaking groups around the Altai-Sayan region, have been highlighted for their potential relationship with Native Americans (Santos et al. 1999; Starikovskaya et al. 2005).

Compared to active studies on questions described above, population geneticists have shown little interest in broad genealogical questions on which many Transeurasian linguists have focused (Starostin et al. 2003; Robbeets 2005; Vovin 2005c): i) whether Transeuarsian languages share a common ancestor and ii) if so, how the branches of Transeurasians are related to each other? This apparent inconsistency in research interest stems from, at least in part, that languages and genes evolve at different rates, making some questions easier to be answered by one discipline than their counterparts by the other. For example, linguists often use words for universally basic concepts, such as simple numerals or names of body parts, to investigate a "deep" relationship between language families assuming that the origin of these basic words is likely to predate the origins of those language families (Crowley and Bowern 2010). The time period going back beyond the last 10,000 years is barely explored because of the fast language evolution. In contrast, the age of shared genetic variants often goes back to hundreds of thousands of years, so that even modern humans and archaic hominins, such as Neandertals and Denisovans, share a large fraction of genetic variants (Green et al. 2010). Therefore, genetic drift for the last few thousand years, which is the temporal range of key linguistic questions, is far too small to provide high statistical power to capture underlying population differentiation. Also, such sharing of common markers, in combination with widespread background gene flow between nearby populations ("isolation-by-distance"), makes it difficult to trace back the origin of a genomic segment in an admixed population beyond recent admixtures between highly divergent populations (Price et al. 2009; Maples et al. 2013).

Methodological innovations are gradually enhancing statistical power to characterize subtle genetic difference between populations. Specifically, novel analytical methods focus on extracting more recent history from two distinct sources; i.e. sharing of low frequency (rare) variants and sharing of a large segment of the genome. Low frequency variants are in principle better for disentangling recent population history because they tend to be young and geographically confined (Gravel et al. 2011). Therefore, sharing of rare variants between populations provides strong evidence for the recent connection between them. Such an approach can be extremely powerful, distinguishing closely related sources of gene flow such as Anglo-Saxons and Vikings (Schiffels et al. 2016), but it also requires a large quantity of genomic resources. Sharing of a large segment of the genome ("haplotype") is a powerful tool to trace recent common ancestry, because recombination constantly breaks such a long shared haplotype every generation (Ralph and Coop 2013). Haplotype sharing has been extensively used in estimating recent common ancestry and effective population size (Browning and Browning 2011). In addition, sharing of shorter haplotypes between distantly related populations is shown to harbor rich information in demographic inference, such as characterizing dates and sources of admixture (Hellenthal et al. 2014). Therefore, applying these novel methods to Transeurasian populations has a great potential to investigate the genetic relationship of these groups in detail.

A growing field of paleogenomics, a study of genomes from ancient biological remains, has made major breakthroughs in understanding our genetic prehistory for the last decade or so (Slatkin and Racimo 2016). More recent demographic events, such as an admixture or a bottleneck (a reduction in population size for a prolonged time period), often complicate inference on those in more distant past. Also, even complex demographic models are still much simpler than the actual demography, capturing only a part of what has happened. Therefore, ancient DNA serves as an extremely powerful tool to overcome these general issues by making a direct observation of our past. For example, studies of prehistoric European genomes provided unequivocal evidence that early Neolithic European farmers were migrants from Anatolia and interbred with descendants of Mesolithic European hunter-gatherers to form later Neolithic populations (Keller et al. 2012; Skoglund et al. 2012). Furthermore, another wave of demic diffusion, associated with Chalcolithic and early Bronze Age populations from the Central Russian Uplands, introduced the third major ancestry into Europe; these three ancestries mixed and formed the gene pools of most contemporary European populations (Lazaridis et al. 2014; Haak et al. 2015). Compared to Europe, there are only a few paleogenomic studies published in Eastern Eurasia so far, most of which are published within a year (Fu et al. 2013; Jeong et al. 2016b; Siska et al. 2017; Yang et al. 2017; Damgaard et al. 2018a; Damgaard et al.

2018b; Jeong et al. 2018a; Lipson et al. 2018; McColl et al. 2018). Extending this short list to regions and time periods directly relevant to the evolution of Transeurasians, such as Northeast China and Mongolia, is expected to happen for the next few years.

In this chapter, we will overview recent population genomic studies of Transeurasian populations, focusing on those that explore the shared genetic substratum of Transeurasian populations and the phylogenetic relationship of these shared ancestries.

45.2 An overview of the genetic structure within Transeurasians

Transeurasians occupy a wide geographic landscape, encompassing the Eurasian steppe and Northeast Asia, as well as Siberia in the further north (Figure 45.1). Five language families that constitute Transeurasians, Turkic, Mongolic, Tungusic, Japonic and Koreanic, are not comingled in this vast region; instead, their geographic distribution is clearly structured. For the Altaic groups, Turkic populations widely distribute from Eastern Europe and Anatolia in the west to the Altai-Sayan region in the east, Mongolic populations reside east of Turkic ones in Mongolia and southern Siberia near the lake Baikal, and Tungusic populations locate further east in Northeast China and the Russian Far East (see Figure 45.1). Koreanic and Japonic populations distribute further south, occupying the Korean peninsula and the Japanese archipelago, respectively (Figure 45.1).

<Insert Figure 45.1 here>

Figure 45.1 A set of Transeurasian populations on the map of Eurasia

Populations included in generating Figure 45.2 are marked here. Tungusic, Mongolic and Turkic populations are marked with red, purple and green colors, respectively. For the full names of populations, please see Figure 45.2.

Interestingly, the genetic structure within Transeurasians mirrors their geographic structure. That is, one would expect that Transeurasians in the west share more alleles with non-Transeurasian speakers of the west (e.g. European populations) than those in the east do. Also, Koreanic and Japonic speakers tend to share more alleles with populations of China and Southeast Asia than the Altaic speakers do. Here, we show the parallel between genes and geography in Transeurasians using the principal component analysis (PCA) of Eurasian populations (Figure 45.2; Jeong et al. 2018b). PCA is a widely used dimension-reduction technique, summarizing a complex genetic structure within a data set into a few orthogonal axis of variation (Patterson et al. 2006). When each individual is plotted against PC1 and PC2 (each individual is marked by a three-letter code), a clear genetic cline of Turkic speaking populations run along the east-west direction on PC1, neighboring Mongolic and Tungusic groups in the further east (Figure 45.2). Turkic populations are also separated along PC2; e.g. among the western-most populations, Tatars and Chuvash are close to Russians and other Northeast Europeans, while Azeri are close to the other Caucasus populations (Figure 45.2). This also matches well with their current geographic locations (Figure 45.1). Likewise, on PC2, Eastern Eurasians are structured as a north-south cline, with north Siberians on top and Southeast Asians on the bottom (Figure 45.2). As expected, Koreans and Japanese are much closer to Southeast Asians than Tungusic speakers are.

<Insert Figure 45.2 here>

Figure 45.2 Top two principal components of 2,077 Eurasian individuals

We used a data set reported by Jeong et al. (2018b). We calculated PCs with 593,124 autosomal SNPs in the Affymetrix HumanOrigins genotyping array (Patterson et al. 2012). PC1 separates Western and Eastern Eurasian populations, with multiple parallel clines running in between. Turkic and Mongolic populations form the middle and the bottom clines, and the Uralic and Yeniseian populations in North Eurasia form the top one. PC2 separates both Western and Eastern Eurasians along the north-south direction. Population names and their three-letter abbreviations are presented at the bottom. Ancient individuals, marked by color-filled symbols, are projected onto PCs using "*lsqproject: YES*" option in the smartpca program. Same symbols and codes are used across Figures 45.1, 45.2 and 45.4.

From this observation, it is clear that Transeurasians harbor extreme genetic heterogeneity that mirrors their geographic location. In short, a Transeurasian population tends to share extra genetic affinity with nearby non-Transeurasian populations that the other Transeurasians at more distant locations do not, likely due to gene flow between nearby populations. Such a pattern of proximity-based allele sharing, often called as "isolation-by-distance", is a universal feature of human genetic structure (Ramachandran et al. 2005; Lao et al. 2008; Novembre et al. 2008; Wang et al. 2012; Pugach et al. 2016).

The genetic heterogeneity in Transeurasians makes it obsolete to take a naive approach for the Transeurasian unity and substructure. First, many Transeurasian populations are genetically much closer to nearby non-Transeurasians than they are to distant Transeurasians. For example, Balkars, a Turkic-speaking group in Caucasus, are genetically very similar to neighboring Caucasian groups such as Adygei, but show almost no genetic similarity with Japanese (Lazaridis et al. 2016). That is, contemporary Transeurasians do not descend from a single homogenous gene pool because they experienced extensive admixture with their non-Transeurasian neighbors. A more relevant genetic question for the Transeurasian unity is whether Transeurasians share a part of their genomes that can trace back to their ancestral gene pool after accounting for admixture. Second, inference on the phylogenetic relationship between various Transeurasian groups is equally confounded by these historical admixtures from disparate sources, because a naive population tree inferred without considering admixture merely reflects an average of the relationship between multiple ancestry components that constitute each Transeurasian group.

Thus, it becomes clear that characterizing the admixture in each Transeurasian group is an essential prerequisite for understanding the Transeurasian genetic history. It also evokes an attention to the genetic profile of their non-Transeurasian neighbors. Therefore, the Transeurasian question can only be sufficiently solved by contextualizing it in the evolution of the Eurasian gene pool particularly during the Holocene.

45.3 A shared genetic substratum among the Altaic populations

45.3.1 Western Eurasian admixture in Turkic populations

Among the Transeurasians, Turkic speaking populations show the highest genetic heterogeneity as well as the largest geographic distribution (Figure 45.1). Historical records suggest that they originally inhabited in Mongolia but gradually moved westward to reach their contemporary locations (Robbeets et al., this volume: Chapter 43). Especially, the rise of the nomadic Turkic Khaganate during the 6th century is often considered as the beginning of their expansion into the current geographic distribution (Robbeets et al., this volume: Chapter 43).

It seems reasonable to assume that these early Turkic-speaking populations were genetically close to contemporary East Asian populations, considering their geographic location as well as historical descriptions. That is, gradually increasing genetic affinity with Western Eurasian populations in Western Turkic groups is likely to be the result of their admixture with nearby Western Eurasians during their expansion (Figure 45.2). Therefore, much effort has been put into testing and characterizing the east-west admixture in Turkic populations.

Genome-wide variation data provide multiple ways of identifying admixture. In case of a recent admixture, where individuals in the admixing population are still heterogeneous in their admixture proportions, a visual summary of the genetic structure using PCA or genetic clustering methods often provides strong qualitative evidence of admixture (Pritchard et al. 2000; Alexander et al. 2009). However, once the admixture proportion becomes homogenized by recombination, typically happening after a few dozens of generations since mixing, more formal tests of admixture are required (van Dorp et al. 2015). The first category of formal tests of admixture uses correlations of allele frequency in multiple populations to test if a given set of populations fits with a tree-like relationship or not (Patterson et al. 2012). One example is three-population or f_3 statistic, defined as a product of the allele frequency difference between the target population and two references, summed over all SNPs in the data (Patterson et al. 2012).

When the two references are good proxies for the true sources of the target population, we expect the allele frequency of the target is between the two references. Therefore the following statistic has a negative value:

$$f_3(Target; Ref_1, Ref_2) = \sum_{i=1}^{n} (p_{target}^{i} - p_{ref1}^{i})(p_{target}^{i} - p_{ref2}^{i})$$

The second category leverages over admixture linkage disequilibrium (LD), which is defined as non-random combination of alleles in two markers generated by admixture: e.g. in two bi-allelic SNPs (with 0 and 1 alleles segregating), a combination of 1/1 alleles has a frequency different from the product of 1 allele frequency in the two SNPs. LD builds up when a new mutation occurs on a certain haplotype background and gradually decays in time due to recombination between two markers. As a result, markers in short distance tend to be in strong LD, and LD exponentially decays over genetic distance. Admixture between distinct gene pools can generate a strong LD extending over megabases, much longer than the usual LD decaying within about 100 kilobases (Reich et al. 2001; Ardlie et al. 2002). Assuming a pulse-like admixture model, one can get an estimate of time since admixture (“admixture date”) by fitting the decay of admixture LD as a function of genetic distance. A few closely related methods utilize this to provide a formal test of admixture as well as an estimate of admixture date (Moorjani et al. 2011; Loh et al. 2013). Importantly, admixture LD is independent of information used by the allele frequency-based methods (e.g. f_3 statistic), thus providing independent evidence of admixture.

Such genome-wide methods have provided unequivocal evidence of admixture in most Turkic populations genetically investigated so far. First, Uygurs from northwestern China, who participated in the Human Genome Diversity Panel (HGDP), show both extremely negative f_3 statistic and clearly exponential decay of admixture LD, with the estimated admixture date around 800 years before present (yBP) (Patterson et al. 2012). Second, a recent study focusing on the demographic history of Siberians includes several Turkic populations, such as Yakuts, Dolgans, Altaians and Tuvans, who all show significant negative f_3 statistics (Pugach et al. 2016). Third, Yunusbayev et al. (2015) report genome-wide data of 22 Turkic populations from a wide geographic range. In most cases, Turkic populations show admixture signals in both allele frequency and LD-based methods. Last, our new study also clearly detects admixture signals in Yugur ("Yellow Uygur"), Kazakh, Kyrgyz, Salar and Western Yugur, with widely varying contributions of Western Eurasian ancestry from about 4% in Yugur, 10% in Tuvans, 15% in Salar, 20% in Altaians, 30–40% in Kazakh and Kyrgyz, to even 50% in Uygurs (Wang et al. 2019).

The admixture LD-based methods date the genetic admixture in individual Turkic populations roughly around 20–40 generations ago, corresponding to 600–1,200 yBP (Yunusbayev et al. 2015). Such estimates should be taken with care, because the model assumes a single instantaneous mixing of two distinct gene pools, while the actual population process is likely to involve continued mixing over multiple generations even when we consider the simplest scenario. Therefore, the genetic estimate of admixture date is an average over multiple generations' admixture and thus may not be a good proxy of the beginning of admixture if admixture happened over a long time period, or in multiple distinct waves. Indeed, a recent study utilizes a novel method to model admixture process more complex than a single-pulse model, and suggests a two-pulse admixture in Uygurs: a younger pulse around 750 yBP and an older one around ~3,750 yBP (Feng et al. 2017). The estimate for the older admixture event (3,750 yBP) is much earlier than previous estimates based on a single-pulse model but fits within the range of dating of mummies that exhibited European features discovered in the Tarim basin 4,000–2,000 yBP (Feng et al. 2017).

The Y chromosomal profile of Turkic speaking populations is consistent with the genomic evidence showing West Eurasian related admixture. The haplogroup R1a1a-M17, which is most frequently observed in eastern Europe, western and central Asia (Underhill et al. 2015), has reached high frequencies in Turkic groups, comprising 2%–6% of different Yakut groups, 5%–10% in Tofalar, around 7%–20% in Tuvans, about 60% in Tatar, Shor and Kyrgyz. The West Eurasian specific mtDNA lineages also reach high frequencies in Turkic speaking groups (Comas et al. 2004).

45.3.2 Genetic evidence for the shared pre-admixture substratum among the Altaic populations

Due to the strong genetic admixture in Turkic populations, the linguistic question of the Altaic unity does not translate well into its naive genetic counterpart: i.e. whether the Altaic-speaking populations form a genetic clade against all nearby non-Altaic speaking ones. Instead, it is necessary to investigate if Turkic, Mongolic and Tungusic-speaking populations share a part of their ancestry, which predates gene flows from non-Altaic neighbors, that forms a clade against the non-Altaic gene pools (Figure 45.3). Therefore, it boils down to testing if Transeurasians without recognizable Western Eurasian admixture are a better proxy to the Eastern Eurasian ancestry in the Altaic populations than non-Transeurasians.

<Insert Figure 45.3 here>

Figure 45.3 A schematic explanation of the effect of admixture on the inference of the genetic relationship between three Transeurasian populations (T_1-T_3) and their neighbors (O_1 and O_2)

(A) The actual genetic history involves a strong gene flow from O_1 to one of the Transeurasian (T_2). Considering only the Transeurasian side of ancestry, all Transeurasians form a clade against the two outgroups, and T_2 and T_3 are more closely related to each other than they are to T_1. (B) A naive reconstruction of the population relationship without considering admixture will fail to recover both the Transeurasian unity (T_2 is closer to O_1 than the other Transeurasians) and the internal relationship between Transeurasians (T_1 and T_3 are closer to each other than they are to T_1).

Hellenthal and colleagues (2014) introduce a new powerful method, GLOBETROTTER, for characterizing admixture using dense genome-wide data. This method uses the pattern of haplotype sharing between individual human genomes to detect signatures of admixture. Specifically, admixture and following recombination will make the probability of the two markers coming from a single donor population to decay as a function of genetic distance (Hellenthal et al. 2014). By modeling haplotype copying from multiple donor populations at once, it can not only provide a strong and robust test of admixture even with a single individual's genome, but also offers a quantitative platform to characterize fine-scale affinity between donor populations and the true unidentified source. It is also able to compare the single-pulse two-way admixture model with more complex ones, such as the multi-way admixture or the multiple admixture models.

When GLOBETROTTER was applied to the published world-wide genetic data, Turkic and Mongolic populations showed remarkably consistent admixture dates, ranging between 1,206 and 1,368 AD, with Mongolic or Tungusic populations in the data set as a substantial donor (Hellenthal et al. 2014). It was interpreted as the signature of the Mongol empire expansion based on the matching date range (Hellenthal et al. 2014).

Our new study extends the GLOBETROTTER-based characterization of the Eastern Eurasian ancestry in Turkic and Mongolic populations to a much bigger set (Jeong et al. 2018b). Consistent with previous publications, we find that most Turkic and Mongolic populations have rather recent admixture dates with the median of 689 yBP (ranging from 309 to 1,104 yBP; Jeong et al. 2018b). Comparable dates are obtained when we apply the admixture LD-based method. Also, Tungusic populations are consistently the best Eastern Eurasian donor in most of Turkic and Mongolic populations. A whole set of Eastern Eurasians from the north-south cline (PC2 in Figure 45.2) is included as potential donors in this analysis; the results thus strongly suggest that the Tungusic populations are the best contemporary proxy of the Eastern Eurasian substratum in Turkic and Mongolic populations.

45.3.3 Paleogenomic evidence for the shared ancestry of the Altaic populations

If the contemporary data-based inference accurately reflects the genomic past of the Altaic populations, we expect to discover ancient populations genetically similar to contemporary Tungusic populations. A recent study reports ancient genomes from the

Russian Far East dating back to 5,700 BC, which are shown to be genetically most similar to contemporary Tungusic populations, such as Ulchi, Negidal and Nanai at the lower Amur River basin, as well as Nivkh people in the nearby Sakhalin island (Siska et al. 2017). Our unpublished ancient genome data from Northeast China and the Russian Far East, collectively covering a long time sequence from Mesolithic (~11,000 yBP) to early Medieval (~1,000 yBP), shows that this genetic profile was widespread in these regions since Mesolithic (Wang et al. 2019; Ning et al. 2019). Although linguistically distinct from nearby Tungusic speakers, Nivkhs show a genetic profile similar to the Tungusic groups in the mainland, with an additional ancestry related to Ainu, who co-inhabits the island (Jeong et al. 2016a). Considering that contemporary Tungusic populations are mostly distributed over Northeast China and the Russian Far East, except for the northern groups Evens and Evenks (Figure 45.1), these results suggest that the Tungusic gene pool has long occupied this region at least for the last 11,000 years.

Three recent paleogenomic studies collectively portrait genetic changes in northern Mongolia and southern Siberia around the Baikal lake from early Neolithic to late Bronze Age (Damgaard et al. 2018a; Damgaard et al. 2018b; Jeong et al. 2018a). Early Neolithic hunter-gatherers from the Baikal region, dated to 5,000–4,000 BC, have a genetic profile that is most similar to ancient individuals from Northeast China and the Russian Far East, i.e. the Tungusic-related profile (Damgaard et al. 2018a). In this region, Upper Paleolithic genomes from Mal'ta and Afontova Gora archaeological sites show a genetic profile markedly different from the Neolithic one (Raghavan et al. 2014; Lazaridis et al. 2016). The Upper Paleolithic gene pool is frequently called "Ancient North Eurasians" (ANE). It is a sister group of the prehistoric European hunter-gatherers and therefore much more distantly related to East Asians (Raghavan et al. 2014). Also, the north-south cline of Eastern Eurasians is at least partially formed by higher ANE-related admixture in northern populations, culminating in Native Americans who derive about 40% of their ancestry from ANE (Raghavan et al. 2014; Lazaridis et al. 2016). Therefore, the early Neolithic Baikal genomes show that there was a massive genetic influx of the Tungusic-related gene pool in this region between the Upper Paleolithic and early Neolithic.

Late Neolithic and early Bronze Age Baikal hunter-gatherers have higher genetic affinity to ANE than the early Neolithic ones do (Damgaard et al. 2018a; Damgaard et al. 2018b). They are modeled to have about 10% ancestry coming from ANE on top of the early Neolithic gene pool (Damgaard et al. 2018a; Jeong et al. 2018a). Late Bronze Age individuals from Khövsgöl aimag in northern Mongolia, dated to 1,400–900 BC, are mostly similar to the early Bronze Age Baikal individuals but with ~7% contribution from Bronze Age western steppe populations, such as those associated with Sintashta culture (Jeong et al. 2018a). Interestingly, one of 20 Khövsgöl individuals is markedly different from the rest but is most closely related to present-day Tungusic/Nivkh populations as well as ancient individuals from the Russian Far East (Jeong et al. 2018a). This shows the presence of gene flow from the Tungusic-related gene pool into Mongolia during Late Bronze Age. The impact of this and later gene flow is also shown by comparing Khövsgöls to Tuvinians, a contemporary population in the neighboring region. Ancient and contemporary populations with ANE-related ancestry are more closely related to Khövsgöl than to Tuvinians, while contemporary East Asians show the opposite pattern (Jeong et al. 2018a).

Taken together, these studies show that a gene pool represented by present-day Tungusic populations has long been present in a large geographic area across Northeast China and the Russian Far East at least since 9,000 BC. This gene pool expanded into the Baikal region and mostly replaced the Upper Paleolithic ANE gene pool before ~5,000 BC. Due to additional genetic influx, Late Bronze Age Khövsgöls were distinct from present-day Mongolic and Turkic-speaking populations, showing extra affinity to ANE. Together with one outlier individual, this difference strongly suggests that additional gene flow from the Tungusic-related gene pool into Mongolia occurred since Late Bronze Age. The ancestral gene pools of present-day Mongolic and Turkic populations are likely to have been formed during this gene flow within the last 3,000 years, as a mixture of immigrant Tungusic-related gene pool and the local populations in Mongolia and the Altai-Sayan region, respectively. Further paleogenomic studies in Mongolia and Northeast China will be critical to provide the genetic overview of more recent cultural groups in Mongolia, such as Xiongnu, Xianbei and Turks.

45.4 Genetic connection between Japonic and Koreanic populations

The dual origin of modern Japanese and Ryukyuans, who constitute the Japonic family of Transeurasian languages, as a mixture of indigenous Jomon hunter-gatherers and immigrant Yayoi rice farmers is now well proven by archaeological, anthropological and genetic data (Hanihara 1991; Jinam et al. 2012). The earliest Yayoi archaeological sites are found in northern Kyushu beginning 300 BC or earlier, and Yayoi people spread both to northeast and to southwest with rice and millet farming technologies as well as metallurgy (Hudson 1999). As a result, populations in the northeastern-most (Ainu in

Hokkaido and Sakhalin) and the southwestern-most (Ryukyuans in Okinawa) regions preserve a higher amount of Jomon ancestry than the mainland Japanese in between, who derive about 20% of their ancestry from Jomon (Jeong et al. 2016a). The genetic study of Jomon and Ainu, the closest modern relative of prehistory Jomon people, suggests that the Jomon/Ainu ancestry is a deep branch of Eastern Eurasians presumably predating the split of Native American and East Asian ancestors (Jeong et al. 2016a; Kanzawa-Kiriyama et al. 2017; McColl et al. 2018). This is also in line with that Jomon/Ainu and even the mainland Japanese have a derived mutation in the *EDAR* (ectodysplasin A receptor) gene in much lower frequency than the rest of Eastern Eurasians (Jeong et al. 2016a). This adenine-to-guanine mutation at the SNP rs3827760 shows an extreme genetic differentiation between continents; it reaches close to 100% frequency in East Asians and Native Americans but extremely rare in the other continental populations (Kamberov et al. 2013). Such a pattern strongly suggests that it arose to high frequency due to positive natural selection in Eastern Eurasians, although the adaptive nature of its phenotypic impact is still not well understood.

Population genetic studies point out Koreans as the best proxy for the ancient Yayoi with regard to modeling the admixture in Japanese (Jinam et al. 2012). For example, contemporary Koreans share more alleles with mainland Japanese than Han Chinese do. Together with the close geographic distance between the Korean peninsula and the Japanese archipelago, it is likely that the Yayoi people arrived in the archipelago via the peninsula. However, considering the absence of paleogenomic data of the Yayoi people and contemporaneous Koreans, it is still unclear if such a genetic affinity is due to the early common ancestry or due to continued migration between the two regions after the Yayoi period.

The affinity between Japonic and Koreanic populations could also be seen from the shared Y chromosomal lineage O2b-M176. This haplogroup has been detected in approximately 30%–40% of Japonic speaking people, about 20%–40% in Koreans and 2%–20% in Tungusic speaking Manchu, Hezhen and Xibo, but nearly absent in all other populations. The Japonic and Koreanic populations also exclusively share a subclade of O2b-M176, which is called O2b1-47z comprising about 20%–30% of Japonic speaking people and 4%–12% of Koreans (Jin et al. 2003; Hammer et al. 2006; Xue et al. 2006; Nonaka et al. 2007; Kim et al. 2011). It is estimated that the time to the most recent common ancestor (T_{MRCA}) of the O2b1-47z clade traces back to 1,900–2,500 BC (Poznik et al. 2016), suggested probably to be one of the ancestral lineages contributing to the Yayoi people. The Japonic people have more than 30% of haplogroup D-M174, which is also found at high frequency in Ainu, Tibetans and Andamanese, but at low frequency in Koreans (Shi et al. 2008). The age of Japonic specific lineage D2-M55 was 14,060–31,050 years (Hammer et al. 2006), much older than that of haplogroup O2b1-47z. Therefore, D2-M55 most likely corresponds to the Jomon period.

45.5 Genetic connection between the Altaic and the Japonic/Koreanic populations

Given the strong genetic link between contemporary Koreanic and Japonic speakers, the linguistic question of broad Transeurasian unity translates into testing the genetic link between Koreanic and Tungusic populations. Again, a naive approach of testing their cladeness against non-Transeurasian populations rejects the idea of Transeurasian unity, because these populations are genetically heterogeneous. Specifically, contemporary Koreans are genetically much closer to Chinese and southeast Asian populations than Tungusic populations are to them, as shown in their intermediate positions along PC2 in our PCA analysis (Figure 45.2).

Consistent with PCA results, formal measures of genetic affinity show contemporary Tungusic speakers to be genetically closer to modern Koreans and Japanese than to Han Chinese or other southern Chinese populations (e.g. Hmong-Mien or Tai-Kadai speakers) (Siska et al. 2017). Since this still holds when modern Tungusic groups are replaced by early Neolithic genomes from the Russian Far East, such an affinity has a deep history (Siska et al. 2017). That is, a part of Korean/Japanese genomes traces its ancestry back to early Neolithic populations in the nearby region, which contemporary Tungusic populations strongly resemble.

It is of great importance to understand when and how the gene flow between Korean/Japanese and the Tungusic-related populations in Northeast China and the Russian Far East happened: does it represent a shared Transeurasian genetic substratum, or reflect a more recent gene flow between adjacent populations? We propose that pre-farming and early farming populations of the Korean peninsula and its nearby region, especially the West Liao River region, are key to resolve this question, based on the Farming/Language Dispersal Hypothesis (Renfrew 1997; Bellwood and Renfrew 2002; Diamond and Bellwood 2003; Bellwood 2005). If the former hypothesis is true, we expect to observe pre-farming Korean populations similar to the Tungusic-related gene pool, as well as a large-scale gene flow, associated with the spread of farming, from a source related to present-day populations in southern China.

<Insert Figure 45.4 here>

Figure 45.4 Top two principal components of 367 Eastern Eurasian individuals

We used the same data set for Figures 45.2 and 45.4. Population names and their three-letter abbreviations are presented at the bottom. Ancient individuals, marked by color-filled symbols, are projected onto PCs using “*lsqproject: YES*” option in the smartpca program.

The densely populated agricultural center in the West Liao River valley and nearby area in Liaoning is a plausible geographic source for such hypothetical migrations. Its early Neolithic Xinlongwa culture has been hypothesized to be a source of a farming population expansion associated with the wide geographic distribution of Transeurasian languages (Robbeets 2017d). Given its geographic, it is also possible that Northeast China itself experienced farming-associated migrations between the Tungusic-related gene pool and one related to Chinese populations further to the south. Archaeogenetic studies of uniparental markers indeed suggest a genetic transition within Northeast China, showing an increase in the genetic affinity with populations from the Central Plain over time (Cui et al. 2013; Zhang et al. 2017).

Our unpublished genome data capture time-dependent changes in the genetic profile of the West Liao River populations in fine resolution (Ning et al. 2019). The genetic profile of early farming populations in this region (3,500–3,000 BC, associated with Hongshan culture) is already intermediate between the Tungusic-related gene pool and more southern populations (Figure 45.4; Ning et al. 2019). Later individuals dated to 2,000–

1,500 BC, associated with Lower Xiajiadian culture with more intensive farming, show even less genetic affinity with the Tungusic-related gene pool. Interestingly, later individuals associated with nomadic pastoralist Upper Xiajiadian culture shows an increased affinity to the Tungusic-related gene pool.

It is currently unknown when and how the genetic profile of Hongshan period individuals formed primarily due to lack of earlier genomes in the West Liao River region, especially those belonging to Xinlongwa and Zhaobaogou cultures. Future genetic studies focusing on them will be critical to understand the socio-cultural context of the genetic change in the West Liao River region: e.g. whether there was a genetic change between Xinlongwa and Hongshan periods and whether the West Liao River population was already relying on farming to a substantial degree when the genetic interaction with the southern one began.

45.6 Discussion

Genetics, archaeology and linguistics provide independent evidence of the relationship of past and contemporary human populations, cultures and languages. Although they have great potential to corroborate each other to reveal human prehistory, findings and implications from one discipline have either stayed within the field due to communication barriers or have been received with misunderstanding.

There are multiple reasons for this situation. Setting aside general difficulty of communicating detailed academic knowledge between disparate disciplines, one main reason is that genes, languages and material cultures do not correspond in one-to-one manner. Numerous examples even among the modern populations show decoupling between these elements. For example, western-most Finno-Ugric populations in Europe, such as Hungarians, are genetically largely indistinguishable with nearby Indo-European populations, while their language is of distinct origin (Lazaridis et al. 2016). Therefore, it is important to understand that languages, genes and material cultures evolve without a complete linkage and therefore show disparate relationships among populations.

Evolutionary paths of genes, languages and material cultures become even more divergent when we consider fine-scale details, because they evolve in different ways and speeds. This fundamental difference also provides each discipline with a varying level of resolution to reconstruct past across time. Because languages evolve much faster than genes, most language families have evolved and diversified in large scale for the last few thousand years and thus their internal structure is arguably more tree-like. In contrast, slow changes in genes make the genetic structure within these populations much more network-like. Also, linguists are relatively confident in defining a subset of linguistic features, such as basic vocabularies or structural features, to explore the relationship between languages and therefore take care of borrowing by contact, a linguistic counterpart of genetic admixture. Due to the slow speed of genetic change, on the other hand, human populations share a large fraction of common genetic variants (The 1000 Genomes Project Consortium 2015). Therefore, once admixture happens, it becomes a nontrivial task to infer the original ancestry of each segment in an admixed genome (“local ancestry deconvolution”) (Price et al. 2009).

Due to these reasons, the straightforward question of Transeurasian unity in linguistics has not been well translated into a corresponding question in genetics. Instead, population genetics of Transeurasians has focused on the topics that genetic data can have a high resolution, such as characterizing an admixture between divergent Eastern and Western Eurasian populations, for which Turkic populations provide an example (Yunusbayev et al. 2015). The relevant question of the Transeurasian unity, i.e. how the pre-admixture genetic substratum in Transeurasians are related to each other and to non-Transeurasian gene pools, had been largely considered out of the resolution of genetic data for decades, because such a small difference between closely related gene pools is easily superseded by a large difference introduced by a gene flow from a distinct outgroup. As reviewed in this chapter, genetic evidence supporting the shared genetic substratum among the Transeurasians is emerging only recently with the aids of new sophisticated computational methods and paleogenomic data. For example, our study based on a comprehensive sampling of present-day Turkic and Mongolic populations across inner Eurasia shows that present-day Tungusic-speaking populations match their Eastern Eurasian ancestry better than non-Transeurasian-speaking Eastern Eurasian populations, e.g. northern Siberians (e.g. Nganasans, Yukagirs, Koryaks and Chukchis) or East Asians further to the south (e.g. Austroasiatic, Austronesian, Hmong-Mien, Sino-Tibetan and Tai-Kadai speaking populations).

Paleogenomics has provided critical information to resolve decades-long debates on European prehistory, such as the agent of Neolithization in Europe (demic diffusion vs. cultural diffusion) or the origin of the Indo-European language (Neolithic Anatolia vs. Bronze Age steppe) (Lazaridis et al. 2014; Haak et al. 2015). With its focus rapidly expanding out of Europe, it is also expected to provide transformative information on the East Asian prehistory.

It is worth highlighting that paleogenomic data work best when provided as a time series in regions of interest. Genetic data are powerful to detect the genetic affinity between populations but have only a poor resolution in reconstructing the temporal dynamics of such affinity, because what we can observe is only the mean across time. A good example is provided by the Central Asian populations in the steppe: genetic dating shows only relatively recent admixture around the time of the Mongolian empire (Hellenthal et al. 2014; Yunusbayev et al. 2015), while paleogenomic data confirms that the east-west admixture in the steppe began several millennia ago in early Bronze Age (Allentoft et al. 2015; Mathieson et al. 2015). Paleogenomic time series data are a key to distinguishing a long-range migration within the limited period from a gradual exchange of genes between nearby population over a long period. While the latter is an omnipresent evolutionary process that shapes the first-degree approximation of human population structure, the former represents unusual prehistoric phenomena what genetics, archaeology and linguistics actually try to identify.

From this perspective, genetics of Transeurasian prehistory needs to focus on a dense paleogenomic sampling of Northeast China, Mongolia and Korea. The West Liao River basin is a home for a sequence of cultures with early adoption of millet cultivation (Shelach 2000; Jin 2002). Data from this region will be able to portrait the nature of ancestral Transeurasian gene pool and its interaction with its Sinitic southern neighbors. Mongolia and Korea will provide further information to understand how the Altaic and Japono-Koreanic populations, were differentiated and more recently how their sub-branches were formed.

Acknowledgment

The research leading to these results has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 646612) granted to Martine Robbeets. Choongwon Jeong is also supported by the Max-Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean. Chuan-Chao Wang is also supported by the National Natural Science Foundation of China (31801040), Nanqiang Outstanding Young Talents Program of Xiamen University (X2123302), and by the Fundamental Research Funds for the Central Universities (ZK1144).

This is a draft version of a chapter that appears in Robbeets, M. and A. Savelyev (eds). *The Oxford Guide to the Transeurasian Languages* (OUP, 2020), see https://global.oup.com/academic/product/the-oxford-guide-to-the-transeurasian-languages-9780198804628. The research leading to these results has received funding from the European Research Council under the Horizon 2020 Program/ ERC Grant Agreement n. 646612 granted to Martine Robbeets.