Title: Exploring correlations in genetic and cultural variation across language families in Northeast Asia

Authors
Hiromi Matsumae[a,b,†,‡,*], Peter Ranacher[‡,c,d, *], Patrick E. Savage[e,f,*], Damián E. Blasi[g,h,i], Thomas E. Currie[j], Kae Kognebuchi[k], Nao Nishida[l], Takehiro Sato[m], Hideyuki Tanabe[n], Atsushi Tajima[m], Steven Brown[o], Mark Stoneking[p], Kentaro K. Shimizu[a,b,q], Hiroki Oota[k,r,s,*], and Balthasar Bickel[g,q].

[a] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland.
[b] Kihara Institute for Biological Research, Yokohama City University, 641-12 Maioka-cho, Totsuka-ku, Yokohama, Kanagawa, 244-0813, Japan.
[c] Department of Geography, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland
[d] URPP Language and Space, University of Zurich, Freiestrasse 16, 8032 Zurich, Switzerland
[e] Department of Musicology, Tokyo University of the Arts, 110-8714 Tokyo, Japan
School of Anthropology and Museum Ethnography, University of Oxford, OX2 6PE, Oxford, UK
[f] Faculty of Environment and Information Studies, Keio University, Shonan Fujisawa Campus, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan
[g] Department of Comparative Language Science, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland.
[h] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, 07745 Jena, Germany.
[i] Human Relations Area Files, 755 Prospect Street, New Haven, USA
[j] Human Behaviour & Cultural Evolution Group, Centre for Ecology & Conservation, Department of Biosciences, University of Exeter, Penryn Campus, TR10 9FE, UK.
[k] The Kitasato University Graduate School of Medical Science, Sagamihara, Kanagawa, 252-0374, Japan.
[l] Genome Medical Science Project, Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Chiba, 272-8516, Japan
[m] Department of Bioinformatics and Genomics, Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa, 920-8640 Japan
[n] Department of Evolutionary Studies of Biosystems, School of Advanced Sciences, SOKENDAI (The Graduate University for Advanced Studies), Shonan Village, Hayama, Kanagawa 240-0193, Japan
[o] Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton Ontario L8S 4K1, Canada.
[p] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D04103 Leipzig, Germany.
[q] Center for the Interdisciplinary Study of Language Evolution (ISLE), Plattenstrasse 54, 8032 Zürich, Switzerland.
[r] Kitasato University School of Medicine, Sagamihara, Kanagawa, 252-0374, Japan.
[s] Faculty of Science, the University of Tokyo, Japan.
[†] Current Address: Tokai University School of Medicine. Shimokasuya 143, Isehara, Kanagawa, Japan. 259-1143
[‡] These authors contributed equally to this work.
* Corresponding authors: matsumae.hiromi.g@tokai.ac.jp, psavage@sfc.keio.ac.jp, peter.ranacher@geo.uzh.ch, or hiroki_oota@med.kitasato-u.ac.jp

Abstract

Culture evolves in ways that are analogous to, but distinct from, genomes. Previous studies examined similarities between cultural variation and genetic variation (population history) at small scales within language families, but few studies empirically investigated such parallels across language families using diverse cultural data. We report an analysis comparing culture and genomes from in and around Northeast Asia spanning 11 language families. We extract and summarize the variation in language (grammar, phonology, lexicon), music (song structure, performance style), and genomes (genome-wide SNPs) and test for correlations. We find that grammatical structure correlates with population history. Recent contact and shared descent fail to explain the signal, suggesting relationships that arose before the formation of current families. Our results suggest that grammar might be a cultural indicator of population history, while also demonstrating differences among cultural and genetic relationships that highlight the complex nature of human history.

Introduction

The history of our species has involved many examples of large-scale migrations and other movements of people. These processes have helped shape both our genetic and cultural diversity(*1*). While humans are relatively homogenous genetically, compared to other species, there are subtle population-level differences in genetic variation that can be observed at different geographical scales(*2*). Furthermore, while there are universal features of human behavior (e.g., all known societies have language and music(*3–5*)), our cultural diversity is immense. For example, we speak or sign over 7,000 mutually unintelligible languages(*6*), and for each ethno-linguistic group there tend to be many different musical styles(*7*). Researchers have long been interested in reconstructing the history of global migrations and diversification by combining historical and archaeological data with patterns of present-day biological and cultural diversity. Going back as far as Darwin, many researchers have argued that cultural evolutionary histories will tend to mirror biological evolutionary histories(*8–18*). However, differences in the ways that cultural traits and genomes are transmitted mean that genetic and cultural variation may in fact be explained by different historical processes(*19–22*). Major advances in both population genetics and cultural evolution since the second half of the 20th century now allow us to test these ideas more readily by matching genetic and cultural data(*19*, *23–25*) .

The cultural evolution of language has proven particularly fruitful for understanding past population history(*26–28*). A classic approach involves identifying and analyzing sets of homologous (cognate) words among languages. This lexical approach allows for the reconstruction of evolutionary lineages and relationships within a single language family, such as Indo-European or Austronesian (*26*, *27*, *29*). However, lexical methods cannot usually be applied to multiple language families(*28*) as they do not share robustly identifiable cognates due to a time limit of approximately 10,000 years, after which phylogenetic signals are generally lost(*30–32*). An alternative approach is to study the distribution of features of grammar and phonology, such as the relative order of word classes in sentences or the presence of nasal consonants. While structural data in language tends to evolve too fast to preserve phylogenetic signals of language families(*33*, *34*) and the history of lexica and structure might be partially independent as for example in the emergence of creole languages(*22*), the geographical distribution of language structure often points to contact-induced parallels in the evolution of entire sets of language families beyond their individual time depths(*35*, *36*).

Yet language is only one out of many complex cultural traits that could serve as a proxy for deep history. It has been proposed that music may preserve even deeper cultural history than language(*37–43*). Standardized musical classification schemes (based on features such as

rhythm, pitch, and singing style) can be used to quantify patterns of musical diversity among populations for the sake of comparison with genetic and linguistic differences(*37*, *38*, *41*, *44*). Among indigenous Taiwanese populations speaking Austronesian languages, such analyses revealed significant correlations between music, mitochondrial DNA, and the lexicon(*38*), suggesting that music may indeed preserve population history. However, whether such relationships extend beyond the level of language families remains unknown.

To address this gap, we focus on populations in and around Northeast Asia (Fig. 1). Northeast Asia provides a useful test region because it contains high levels of genetic and cultural diversity – including a large number of small language families or linguistic isolates (e.g., Tungusic, Chukuto-Kamchatkan, Eskimo-Aleut, Yukagir, Ainu, Nivkh, Korean, Japanese)(*35*). Crucially, while genetic and linguistic data throughout much of the world have been published, Northeast Asia is the only region for which published musical data allow for direct matched comparison of musical, genetic, and linguistic diversity(*41*, *45*, *46*).

We here use these matched comparisons to test competing hypotheses about the extent to which different forms of cultural data reflect population history at a level beyond the limits of language families. Specifically, we aim to test whether patterns of cultural evolution are significantly correlated with patterns of genetic evolution (population history), and if so, whether music or language (lexicon(*47*), grammar(*48*, *49*), or phonology(*50–52*)) would show the highest correlation with patterns of genetic diversity, after controlling for the influence of recent contact between languages (spatial autocorrelation) and shared inheritance within individual language families.

Results

We selected all available populations from in and around Northeast Asia (14 populations, encompassing 11 language families/isolates) for which all four sources of data (genome-wide single nucleotide polymorphisms (SNPs), grammars, phonology, and music) were available (Fig. 1, Tables S1-S2, Materials and Methods)(*41*). For genetic data, we newly genotyped 22 Nivkh individuals from Sakhalin Island in Russia using the Illumina Human Omni 2.5-8 BeadChip array (Materials and Methods). First, we investigated the similarity between populations in each of the dimensions of enquiry. For this purpose, we used split networks(*53*), which display multiple sources of similarity in a consistent manner (Fig. 2; and Supporting Information Fig. S10-14, Tables S3-S7). Distance analysis of lexical data resulted in a network topology with an overall star-shaped structure (Fig.2C) (where we exclude Nganasan for lack of data in our source database (*47*)). Exceptions are given by the two pairs of languages that are

related to one another and that stand out as proximate (Even and Evenk both belong to the Tungusic family, while Chuckchi and Koryak both belong to the Chukotko-Kamchatkan family). The results of this distance analysis are consistent with the fact that lexical material is able to detect relationships within language families, but cannot resolve historical relations between families. Therefore, we excluded the lexical data from subsequent analyses.

Distance analyses of grammatical, phonological, genetic, and musical distances reveal potentially more informative structure. Importantly, and in agreement with the claim that language structure does not identify family relationships(*30*, *33*), the clustering emerging from the distances do not generally coincide with languages families, except perhaps for Tungusic (Even and Evenk) in the domain of grammar (with a distance of 17.37 vs. a distance of 21.45 to their next unrelated neighbor Nivkh, Table S6) and Chukotko-Kamchatkan (Chukchi and Koryak) in the domain of phonology (with a distance of 23.45 vs. a distance of 27.81 to the next unrelated neighbor Nganasan). Most of the clustering instead points to inter-family relations: for example, Korean and Japanese are neighbors in the networks based on grammar, SNPs, and music, but not phonology(*54–56*). Buryat and Yakut are close together in SNPs, grammar, and phonology, but not in music (*41*, *57*). The music-based network is consistent with a previous study showing the uniqueness of Ainu music and a distinction of East Asian music from circumpolar music based on cluster analysis of musical components(*41*). Nivkh shows different patterns for each factor. For example, Nivkh is genetically closer to Korean, Japanese, and Buryat than the others and shows the second highest affinity with Ainu in all populations in the distance matrix, reflecting the tree's branch position. However, music, grammar, and phonology do not follow such relationships in Nivkh. Taken together, these results suggest that similarities in the grammatical and phonological structure are not primarily associated with simple vertical descent within language families. Instead, the different cultural features might be associated (between themselves and with genetic history) in the way in which they are transmitted, developed or adopted across families.

We implemented a redundancy analysis (RDA) on the principle components (or coordinates) of the data to explore this idea further (Materials and Methods, Supplementary Information). RDA summarizes the variation in a response variable that can be explained by an explanatory variable and finds directed associations. The RDA analysis reveals two associations that are significant under a permutation test (Fig. 3): grammatical similarity predicts genetic similarity (grammar -> genetics, variance=0.44), and genetic similarity predicts grammatical similarity (genetics -> grammar, variance=0.56).

While distance analysis suggests that both associations reflect deep-time correspondences,

dating back to before the formation of current language families (as identifiably by cognate words), spatial proximity and contact between societies might lead to similar patterns of association that are relatively recent and shallow. In response to this, we evaluated three possible scenarios to explain the signal in the data: 1.) Recent contact scenario: the associations reflect recent and current contact and, hence, can be explained by spatial autocorrelation in the current data; that is, societies that are currently close to each other tend to have similar grammars and population history. 2.) Inheritance scenario: the associations reflect common ancestry. The associations result from vertical descent within the remaining linguistic families for which our sample contains more than one member (Tungusic, Chukotko-Kamchatkan, and Uralic). 3.) Deep-time correspondence scenario: the associations reflect a non-shallow correspondence between language and genetics that cannot be explained by recent contact nor phylogenetic inheritance within known families. To distinguish between the three scenarios we treated spatial proximity and inheritance as potential confounds and carried out a partial RDA to control their effect (Supplementary Information). As societies and languages placed far from the equator tend to display larger spatial ranges(*58*), we represented the territory of each society with areas rather than points and sample random spatial locations from within these areas. The partial RDA reveals strong evidence against the recent contact scenario: spatial proximity fails to explain both associations (Fig. S18). When controlling for spatial autocorrelation (1,000 random samples allowing for the uncertainty of people's locations) the observed explained variance is still greater than that of random permutations (normalized differences between observed and permuted explained variance z > 1 SD in over 99% of spatial samples; Kullback-Leibler Divergence KLD > 2.5; Fig. S16-18). When controlling for both recent contact and phylogenetic inheritance of language in partial RDA, still both associations show stronger evidence than the other relationships (Fig. 4, and S19-21). Of them, genetics -> grammar association is stronger (z > 1 SD in 92.8% of samples, KLD > 1.5) than grammar -> genetics (z > 1 SD in 75.2% of samples; KLD = 1.1).

To evaluate the robustness of the associations, we furthermore performed three types of sensitivity analyses. First, we varied the number of principal components (or coordinates) passed to the RDA and, thus, the amount of variance in both the response and the predictor. Different thresholds of how much variance a component needs to explain in order to be included (10%, 15%, 18%) show little effect on the results, although the association genetics -> grammar is far more robust (all z > 1 SD in at least 91% of samples, all KLD > 1.5) than its inverse (z > 1 SD between 74% and 99% of samples, KLD between 1.2 and 1.4; Fig. S22-S23). Second, we varied the language sample passed to the RDA. While most languages have little to no effect on the

signal, this is not true for Ainu, as removing Ainu from the analysis weakens the support for the associations of grammar and genetics (z > 1 SD in only 25-31%, KLD < 0.2, when controlling for spatial proximity and inheritance; Fig. S25-S27). Third, in the partial RDA, some spatial samples happen to explain the variance in the response better than others (lower tail of observed adjusted R^2 in Fig. S16, S17, S19, and S20). The sampled locations that produce R^2 estimates in the 20% upper and lower tails of the total do not cluster in specific regions in the genetics -> grammar association (Fig. S28-S29), while its inverse tends towards some clustering in the polygons of Buryat, Evenki, and Selkup (Fig. S30-S31). This might be indicative of recent language contact.

To summarize, we found significant correlations between genetics and grammar by the basic RDA using the complete set of genomes, music, and language in Northeast Asia. The partial RDA controlling for geography and linguistic inheritance as well as sensitivity analyses suggest that the genetics -> grammar association is stronger than its inverse, and the relationships may trace back to earlier relationships between languages before the recent contacts and inheritance.

Discussion

We have simultaneously explored the relations among genetic, linguistic, and musical data beyond the level of language families for the first time after Cavalli-Sforza's works(*9*, *12*). We find remarkable evidence for the relationships between population history and grammatical similarity while genomes and grammar might be evolved by different forces.

A possible interpretation of our findings is that the relationship between grammar and population history was exceptionally well preserved over the recent contact beyond language families whether or not evolutionary mechanisms of grammar are the same as genomes. Population genetics detect gene flows between populations beyond phylogenetic relationships. Our dataset covers a phylogenetically-broad range of populations; three lineages to the present-day East Eurasian (Ainu, East Asian, and Northeast Asian) and one to North American (Greenlandic Inuit) (*59*, *60*), including gene flows beyond the lineages, such as Japanese-Ainu(*61*) and Buryat-Yakut(*57*). While the evolutionary forces that influence population history are fairly well understood, determining to what extent the genetic relationships of particular populations reflect shared ancestry vs. prehistoric contact in culture is still challenging. Moreover, the evolutionary processes that influence culture and language are under debate(*62*) but can obviously be very different from those influencing genomes. For example, cultural replacement and language shift can occur within a single generation due to colonization or other sociopolitical factors. Our results removing the influence of the proximity in cultural similarities give support

to the notion that these different data reveal different historical patterns, yet show that some cultural features still can preserve relationships extending even beyond the boundaries of language families. Therefore, grammar was more conservative than the lexicon because the lexical distances are not associated with genetic distances across language families in our dataset (Fig. 2). This contrasts with expectations from historical linguistics(*33*) and also from recent findings that suggest that grammar evolves faster than the lexicon in Austronesian(*63*) and also shows rapid evolution in Indo-European(*64*). The statistical power to detect a signal is weakened when Ainu was removed in the sensitivity analysis (Fig. S24-27). While this might suggest a special position of Ainu in the larger context, we need larger samples of languages and populations inside and outside of Northeast Asia to resolve this question.

The genetics --> grammar correlation is stronger and more robust than its inverse (Fig. S15), indicating that grammatical variation preserves genetic variation better than vice versa. When controlling only for spatial autocorrelation, both are equally strong (Fig. S18), but when controlling for both spatial autocorrelation and common descent (Fig. S21), only genetics --> grammar remains robust. Furthermore, the grammar -> genetics correlation reveals more spatial clustering effects than its inverse. Together, these observations suggest that common descent and contact affect grammatical and genetic variation differently: while they can strongly affect variation in grammar, this is not the case for genes. Genetic variation is more predictable and happens - to some degree - independently of spatial neighborhood and family. This difference between grammar and genes makes it easier to predict grammatical variation from genetics than its inverse.

Interestingly, our results are qualitatively different from the only previous study to quantitatively compare genetic, linguistic, and musical relationships(*38*). Among indigenous Austronesian-speaking populations in Taiwan, music was significantly correlated with genetics but not language, while we find here that music is not robustly associated with either language or genetics. However, there are several methodological differences that might underlie these differences. In particular, the two studies looked at different types of data (genome-wide SNPs, structural linguistic features, and both group and solo songs here vs. mitochondrial DNA, lexical data, and only group songs previously). Further research with larger samples and different types of data may help to elucidate general relationships among language, music, and genetics.

The recent studies highlight Northeast Asian populations as one of major genetic components of basal East Eurasians(*23*, *65*). However, our knowledge about relationships between culture and local population history is limited in Northeast Asia. In addition to revealing an association between genetic and grammatical patterns, our results also reveal complex

dissociations in which these data reflect different local histories. For example, while previous studies suggest specific genetic and cultural relationships between Korean and mainland Japanese populations(*56*, *66*) or posit a shared origin (*67–69*), our findings support similarities in SNPs, music, and grammar, but not in lexicon and phonology (Fig. 2, Supporting Information, Tables S3-S7). Even the Ainu show particular genetic similarity to the Japanese, their music clusters most closely with that of the Koryak (Fig. 2, Tables S3-S5). This may reflect different levels of genetic, linguistic, and musical exchange at different points of history. Musical patterns may reflect more recent cultural diffusion and gene flow from the Okhotsk and other "circumpolar" populations that interacted with the Ainu from the north within the past 1,500 years(*70*), as we previously proposed in our "triple structure" model of Japanese archipelago history(*41*). Newly genotyped Nivkh samples showed the closeness to Ainu in SNPs but not in others (Fig. 2A), suggesting that historical relationships in the coastal region of Northeast Asia. Nivkh might be a key population connecting Ainu and other Northeast Asians, however, the population history of Nivkh is not well understood. Therefore, further analyses are necessary to investigate in more detail of local population history in Northeast Asia including Nivkh.

In conclusion, we have demonstrated a relationship between grammar and genome-wide SNPs across a variety of diverse Northeast Asian language families. Our results suggest that grammatical structure may reflect population history more closely than other cultural (including lexical) data, but we also find that different aspects of genetic and cultural data reveal different aspects of our complex human histories. In other words, cultural relationships cannot be completely predicted by human population histories. Alternative interpretations of such mismatches would be historical events (e.g., language shift in local history) or culture-specific evolution independent from genetic evolution. Future analyses of such relationships at broader scales using more explicit models should help improve our understanding of the complex nature of human cultural and genetic evolution.

Materials and Methods

Experimental Design

Selection of populations in this study

We selected 14 populations for which matching musical (Cantometrics/CantoCore), genetic (genome-wide SNP) and linguistic (grammatical/phonological features) data were available (Tables S1-S2 and Fig. 1). These represented a subset of 35 Northeast Asian populations whose musical relationships were previously published and analyzed in detail(*41*). Linguistically, these 13 populations fall into 11 language families/isolates(*71*) (*72*). Korean, Ainu, Nivkh and

Yukaghir are language isolates. Buryat, Japanese, Yakut, West Greenland Inuit are the sole representatives in our sample of the Mongolic, Japonic, Turkic, and Eskimo-Aleut language families, respectively. The remaining languages are classified into three language families: Koryak and Chukchi are Chukotko–Kamchatkan languages; Even and Evenk are Tungusic languages; and Selkup and Nganasan are Uralic languages.

Music Data

All music data and metadata are detailed in our previous report of circumpolar music(*41*). For the present analysis, we used a subset of 14 of the original 35 populations with matching genetic and linguistic data; these 14 populations are represented by 264 audio recordings of traditional songs. Each song was analyzed manually by P.E.S. using the same 41 classification characters used in (30) (from Cantometrics (29) and CantoCore (35)).

Genetic Data

Nivkh DNA samples from the Horai Collection: We used the DNAs of Nivkh maintained by the Asian DNA Repository Consortium (ADRC). The DNA samples were originally collected in Sakhalin, Russia by Dr. Satoshi Horai in the 2000s and were kept at 4℃ in Sokendai. We genotyped 32 Nivkh individuals (14 females and 18 males) with the Illumina Omni 2.5-8 BeadChip Array at the National Center for Global Health and Medicine (Table_S8_SampleID_Nivkh.xlsx). Two DNA samples were removed because of their poor quality. We selected 2,246,124 sites for SNPs with a call rate greater than 95%. Using PLINK(*73*), we performed a Hardy-Weinberg equilibrium test to exclude sites with $P<10^{-6}$, resulting in 2,246,123 sites. Then we calculated inbreeding coefficients using sites with Minor Allele Frequency (MAF) > 0.01, confirming that none of the cousin equivalents exceeded F = 0.0625. Using the same threshold of MAF, we found kinship between 12 pairs (involving 14 individuals) with PI_HAT >0.125 (third-degree relative or closer). Eight samples were removed; 22 individuals thereby passed the quality control and kinship tests. Then we carried out strand checks between the Illumina Human Omni 2.5-8 BeadChip SNPs and JPT+CHB in 1,000 Genomes using BEAGLE 4.0(*74*). In the Nivkh data, 2,041,779 sites passed the strand check and 114,077 sites were flipped using PLINK. After the strand check, all sites that did not have an allele match were removed. We converted the Illumina unique IDs to rsIDs.

Merging Nivkh and public data: Publicly-available genome-wide SNP array data for 14 populations, including three Nivkh individuals (Table S1)(*56*, *75–79*), were obtained and curated as follows. As several genotyping platforms were used, to avoid discordancy of alleles on +/- strands we used the strand check utility in BEAGLE 4.0(*80*) for a dataset of Ainu against JPT and CHB in 1000 Genomes using BEAGLE 4.0 (*74*). To obtain shared SNPs among

different platforms, genotype datasets including our Nivkh data were merged into a single dataset in PLINK file format by PLINK 1.9(*81*).

Removing outlier individuals: We manually removed outlier individuals from the merged dataset based on results of PCA and ADMIXTURE(*82–84*). Finally, we used 15 individuals of Nivkh (13 individuals from our data and 2 individuals from public data) in the population genetics analysis (Table S1, S8). The final merged genotype dataset included 245 individuals and 37,093 SNPs (total genotyping rate was 0.999). The merged dataset in PLINK format was converted to Genepop format using PGDSpider(*85*).

Language data

Lexical data: Lexical distances between populations were provided by Søren Wichmann on September 13, 2014 using version 16 of the ASJP (Automated Similarity Judgment Program) database(*47*). This program automatically calculates a matrix of pairwise distances between languages by comparing phonetic similarities across 40 categories of basic vocabulary ("Swadesh lists"). Distances are calculated as LDND ("Levenshtein Distance Normalized Divided") values, which corrects for both differences in word lengths and the possibility of chance similarities between languages (*86*). This analysis did not attempt to identify and remove loanwords.
Grammar and phonology data: We combined data on grammatical and phonological traits from AUTOTYP (*52*), WALS(*48*), the ANU Phonotactics database(*50*), and PHOIBLE(*51*) and extracted a set of 21 grammar and 84 phonological features with coverage over 80% in each language, and in most cases 100% (Supporting Information Section 2).

Statistical Analysis

In contrast to population history, standardized methods for modeling cultural evolution across different types of data are not yet established. Therefore, we matched population history to cultural similarities to analyze both genetic and cultural data in a common framework. We obtained distance matrices representing differences between populations/languages for a subsequent comparative analysis using the following procedures for music and language, because musical and linguistic (grammatical and phonological) data have different data structures.

Genetic analysis

To estimate population differentiations, pairwise F_{st} values between populations were calculated with Genepop version 4.2(*87*). Pairwise F_{st} is the proportion of the total genetic variance due to between population differences, and is a convenient measure because it does not depend on the actual magnitude of the genetic variance. In other words, genetic markers that evolve slowly are

expected to have the same F_{st} value as markers that evolve more rapidly, because the total variance is decomposed into within population and between population components.

Music analysis

A previously published matrix of pairwise distances among all 283 songs was calculated using normalized Hamming distances(*88*) to calculate the weighted average similarity across all 41 musical features(*41*). This distance matrix was then used to compute a distance matrix of pairwise musical ϕ_{st} values among the 14 populations using Arlequin(*89*) and the lingoes function of the ade4 package in R(*90*). φ_{st} is analogous to F_{st} but takes into account distances between individual items, making it more appropriate for analysis of cultural diversity(*44*, *88*, *90*). Further details concerning the calculations can be found elsewhere (*90*).

Language analysis

Grammar and phonology data: In contrast to songs and individual genotypes, language data do not represent individuals for each population. In view of the fact that the data are partly numerical and partly categorical, we used a balanced mix of Principle Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) to calculate differences between languages (*91*) (Supplementary Information Section 3). Empty values were imputed using the R package missMDA (*92*).

Comparative analysis of music, SNPs, and language structure

PCoA for SNPs and music: We performed a principal coordinate analysis (PCoA) on the distance matrices of pairwise F_{st} for SNPs and pairwise ϕ_{st} for music (F_{st} and ϕ_{st} matrices are available from github, Supporting Information Section 3)(*93*). Similar to a PCA, a PCoA produces a set of orthogonal axes whose importance is measured by eigenvalues (Fig S2-S5). However, in contrast to the PCA, non-Euclidean distance matrices can be used. Heatplots of PCo and PC were visualized by ggplot2 in R (Fig S6-9) (*94*).

Split network graphs: Distances were visualized using the SplitsTree neighbornet algorithm (version 4, (*95*)) and are reported in detail in Supporting Information Tables S3-S7. In order to control for multicollinearity, we used PCA/MCAs and PCoAs as input rather than the raw data.

Geographic distances: The geographical polygons were taken from the Ethnologue(*96*), supplemented by a hand-drawn polygon estimate for Ainu. In view of the mobility of speakers over time, we sampled 1,000 random locations from within the polygons and used these for assessing correlations. The random point samples were generated in PostGIS https://postgis.net/ (Supporting Information Section 2.4). For each of the 1,000 samples, we computed the spherical distance between all random locations, which we store in a distance matrix. Then we perform a distance-based Moran's eigenvector map analysis (dbMEM) to decompose the spatial structure

of each of the resulting 1,000 distance matrices (Supporting Information Section 3.3)(*97*). Similar to a PCoA, dbMEM reveals the principal coordinates of the spatial locations from which the distance matrix was generated. We only return those eigenfunctions that correspond to positive spatial autocorrelation.

(Partial) Redundancy Analysis: Redundancy analysis (RDA) was carried out to explore the linear relationship between SNPs, grammar, phonology and music. Partial redundancy analysis (pRDA) was used to control for spatial dependence (Supporting Information Section 5). (p)RDA is an alternative to the traditionally used Mantel test, which was found to yield severely underdispersed correlation coefficients and a high false positive rate in the presence of spatially correlated data (*98*). RDA performs a regression of multiple response variables on multiple predictor variables (*99*), while pRDA also allows to control for the influence of confounders. Redundancy analysis yields an adjusted coefficient of determination (adjusted R^2), which captures the variation in the response that can be explained by the predictors. We compare the observed adjusted R^2 values against a distribution under random permutations (Fig. 4, Fig. S15-27). To assess robustness, we z-normalize the difference between observed and permuted adjusted R^2 and report the proportion of samples for which the observed adjusted R^2 is one standard deviation larger than the permuted ($z > 1$ SD). Moreover, we compute the Kullback-Leibler divergence (KLD) between the distribution of observed adjusted R^2 and permuted adjusted R^2. The KLD allows to assess the overall divergence of the two distributions, $z>1$ SD reports the proportion of samples with a strong positive difference. (p)RDA and subsequent analyses were performed in R using the vegan package (64).

Acknowledgements:

Funding: Funding supports for this work were provided by URPP Evolution in Action, University of Zurich to H.M., K.K.S. and B.B., the URPP Language and Space, University of Zurich to P.R., the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI Grant-in-aid for Scientific Research on Innovative Areas #4903(Evolinguistics), Number JP18H05080 and JP20H05013 to H.M.; MEXT KAKENHI Number 16H06469 to K.K.S.; the NCCR Evolving Language, Swiss NSF Agreement #51NF40_180888 to B.B. and K.K.S., MEXT scholarship, MEXT KAKENHI Number 19KK0064, and startup grants from the Keio Research Institute at SFC, Keio Gijuku Academic Development Fund, and Keio Global Research Institute to P.E.S.

General: We thank the Asian DNA Repository Consortium for the use of Ainu SNPs data and URPPs Evolution in Action and Language and Space University of Zurich for support for an

interdisciplinary workshop "Frontiers of early human expansion in Asia: linguistic and genetic perspectives on Ainu, Japan and the North Pacific Rim" to help study linguistic and genetic histories in North Asia. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics, Japan. We thank Søren Wichmann for providing the ASJP distance matrix and Brigitte Pakendorf, Johanna Nichols, Juha Janhunen, Ekaterina Gruzdeva, Anna Bugaeva, Anna Berge, Dagmar Jung, and Marcelo Sánchez, and Chiara Barbieri for discussion.

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

Data and material availability: Data and analysis code are available at https://github.com/derpetermann/music_languages_genes. All raw data were previously published and can be accessed from the original publications and/or public data repositories, with the exception of the Ainu and Nivkh DNA data which was provided by the Asian DNA Repository Consortium. All data and analysis code needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Author Contributions: P.E.S., H.M., H.O., B.B. and K.K.S. initially designed the research with advice from M.S. and S.B.. K.K. N.N. and H.T. genotyped Nivkh DNA samples. T.S. A.T. H.O. and H.M. analyzed genetic data. P.R., D.E.B., B.B., and T.E.C. analyzed language data. P.E.S., H.M. and P.R. analyzed music data. H.M., D.E.B, P.R, and B.B designed and implemented the statistical analysis. H.M, P.E.S., P.R., M.S., B.B., K.K.S., and D.E.B. wrote the manuscript. P.R., D.E.B. and B.B wrote the supporting information.

References

1. M. Pagel, R. Mace, The cultural wealth of nations. *Nature*. **428**, 275–278 (2004).
2. J. Novembre *et al.*, Genes mirror geography within Europe. *Nature*. **456**, 98–101 (2008).
3. D. E. Brown, *Human Universals.* (Temple University Press, 1991).
4. P. E. Savage, S. Brown, E. Sakai, T. E. Currie, Statistical universals reveal the structures and functions of human music. *Proc. Natl. Acad. Sci.* **112**, 8987–8992 (2015).
5. S. A. Mehr *et al.*, Universality and diversity in human song. *Science*. **366** (2019).
6. H. Hammarström, R. Forkel, M. Haspelmath, Glottolog 3.3 (2018), (available at http://glottolog.org).
7. B. Nettl, *The study of ethnomusicology : thirty-three discussions*.
8. C. Darwin, The descent of man, and selection in relation to sex. Volume 1. London: J. Murray (1871).
9. L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, J. Mountain, Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 6002–6 (1988).
10. E. Bortolini *et al.*, Inferring patterns of folktale diffusion using genomic data. *Proc. Natl. Acad. Sci.* **114**, 9140–9145 (2017).
11. R. R. Sokal, Genetic, geographic, and linguistic distances in Europe. *Proc. Natl. Acad. Sci.* **85** (1988).
12. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The history and geography of human genes* (Princeton University Press, 1994).
13. P. Diamond, J. Bellwood, Farmers and Their Languages: The First Expansions. *Science (80-.).* **300**, 597–603 (2003).
14. J. S. Lansing *et al.*, Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 16022 (2007).
15. K. Hunley *et al.*, Genetic and Linguistic Coevolution in Northern Island Melanesia. *PLoS Genet.* **4**, e1000239 (2008).
16. S. A. Tishkoff *et al.*, The genetic structure and history of Africans and African Americans. *Science*. **324**, 1035–44 (2009).
17. N. Creanza *et al.*, A comparison of worldwide phonemic and genetic variation in human populations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1265–72 (2015).
18. G. Longobardi *et al.*, Across language families: Genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol.* **157**, 630–640 (2015).
19. P. J. Richerson, R. Boyd, *Not by genes alone : how culture transformed human evolution* (University of Chicago Press, 2005).
20. R. M. MacCallum, M. Mauch, A. Burt, A. M. Leroi, Evolution of music by public choice. *Proc. Natl. Acad. Sci.* **109**, 12081–12086 (2012).
21. D. E. Blasi, S. Wichmann, H. Hammarström, P. F. Stadler, M. H. Christiansen, Sound–meaning association biases evidenced across thousands of languages. *Proc. Natl. Acad. Sci.* **113**, 10818–10823 (2016).
22. D. E. Blasi, S. M. Michaelis, M. Haspelmath, Grammars are robustly transmitted even during the emergence of creole languages. *Nat. Hum. Behav.* **1**, 723–729 (2017).
23. David Reich, *Who we are and how we got here: Ancient DNA and the new science of the human past* (Pantheon, 2018).
24. J. P. Henrich, *The secret of our success : how culture is driving human evolution, domesticating our species, and making us smarter* (Princeton University Press, 2016).
25. S. C. Levinson, R. D. Gray, Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn. Sci.* **16**, 167–173 (2012).
26. R. D. Gray, a J. Drummond, S. J. Greenhill, Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*. **323**, 479–83 (2009).

27. W. Chang, C. Cathcart, D. Hall, A. Garrett, Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language (Baltim).* **91**, 194–244 (2015).
28. A. Kozintsev, On Certain Aspects of Distance-based Models of Language Relationships, with Reference to the Position of Indo-European among other Language Families. *J. Indo-European Stud.* **46**, 1–264 (2018).
29. S. Xu *et al.*, Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc. Natl. Acad. Sci.* **109**, 4574–4579 (2012).
30. D. A. Ringe Jr, `Nostratic' and the Factor of Chance. *Diachronica*. **12**, 55–74 (1995).
31. R. Gray, Pushing the Time Barrier in the Quest for Language Roots. *Science (80-.).* **309**, 2007–2008 (2005).
32. M. Pagel, Q. D. Atkinson, A. S Calude, A. Meade, Ultraconserved words point to deep language ancestry across Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8471–6 (2013).
33. J. Nichols, in *The Comparative Method reviewed*, M. Durie, M. Ross, Eds. (Oxford University Press, Oxford, 1996), pp. 39–71.
34. S. J. Greenhill *et al.*, Evolutionary dynamics of language systems. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8822–E8829 (2017).
35. J. Nichols, *Linguistic diversity in space and time* (University of Chicago Press, 1999).
36. B. Bickel, J. Nichols, Oceania, the Pacific Rim, and the theory of linguistic areas. *Annu. Meet. Berkeley Linguist. Soc.* **32**, 3–15 (2006).
37. A. Lomax, American Association for the Advancement of Science., *Folk song style and culture* (Transaction Books, 1978).
38. S. Brown *et al.*, Correlations in the population structure of music, genes and language. *Proc. R. Soc. B-Biological Sci.* **281**, 20132072 (2014).
39. E. Callaway, Music is in our genes. *Nat. News* (2007).
40. H. Pamjav, Z. Juhász, A. Zalán, A comparative phylogenetic study of genetics and folk music. *Mol. Genet. ….* **287**, 337–349 (2012).
41. P. E. Savage *et al.*, How 'circumpolar' is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. *Ethnomusicol. Forum* (2015).
42. V. Grauer, Sounding the depths: Tradition and the voices of history. CreateSpace (2011).
43. P. E. Savage, Cultural evolution of music. *Palgrave Commun.* **5(16)** (2019).
44. P. Savage, E. Merritt, T. Rzeszutek, S. Brown, CantoCore: a new cross-cultural song classification scheme. *Anal. Approaches to World Music* (2012).
45. P. E. Savage, Alan Lomax's Cantometrics Project: A comprehensive review. *Music Sci.* **1** (2018).
46. P. E. Savage, in *Oxford Handbook of Music and Corpus Studies*, D. Shanahan, A. Burgoyne, I. Quinn, Eds. (Oxford University Press).
47. S. Wichmann, Eric W. Holman, Cecil H. Brown, The ASJP Database (version 16) (2014).
48. M. S. Dryer, M. Haspelmath, The World Atlas of Language Structures Online (2013), (available at http://wals.info/).
49. B. Bickel *et al.*, The AUTOTYP typological databases. Version 0.1.0 (2017), (available at https://github.com/autotyp/autotyp-data/tree/0.1.0).
50. M. Donohue, R. Hetherington, J. McElvenny, V. Dawson, World Phonotactis Database (2013).
51. S. Moran, D. McCloy, R. Wright, PHOIBLE Online (2014).
52. B. Bickel *et al.*, The AUTOTYP typological databases. Version 0.1.0 (2017).
53. D. H. Huson, D. Bryant, Application of Phylogenetic Networks in Evolutionary Studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
54. K. Hanihara, Dual structure model for the population history of the Japanese. *Nichibunken Japan Rev.* (1991).

55. S. Nakagome *et al.*, Model-based verification of hypotheses on the origin of modern Japanese revisited by Bayesian inference based on genome-wide SNP data. *Mol. Biol. Evol.* **32**, 1533–1543 (2015).
56. T. Jinam *et al.*, The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J. Hum. Genet.* **57**, 787–95 (2012).
57. I. Pugach *et al.*, The Complex Admixture History and Recent Southern Origins of Siberian Populations. *Mol. Biol. Evol.* **33**, 1777–1795 (2016).
58. M. C. Gavin, J. R. Stepp, Rapoport's Rule Revisited: Geographical Distributions of Human Languages. *PLoS One*. **9**, e107623 (2014).
59. T. Gakuhari *et al.*, Jomon genome sheds light on East Asian population history. *bioRxiv*, 579177 (2019).
60. H. McColl *et al.*, The prehistoric peopling of Southeast Asia. *Science (80-.).* **361**, 88–92 (2018).
61. T. Jinam *et al.*, The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. *J. Hum. Genet.* **57**, 787–95 (2012).
62. A. Mesoudi, *Cultural evolution : how Darwinian theory can explain human culture and synthesize the social sciences* (University of Chicago Press, 2011).
63. S. J. Greenhill *et al.*, Evolutionary dynamics of language systems. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E8822–E8829 (2017).
64. M. Widmer, S. Auderset, P. Widmer, J. Nichols, B. Bickel, {NP} recursion over time: evidence from {Indo-European}. *Language (Baltim).* **93**, 1–36 (2017).
65. C. Jeong *et al.*, The genetic history of admixture across inner Eurasia. *Nat. Ecol. Evol.* **3**, 966–976 (2019).
66. M. A. HUGO Pan-Asian SNP Consortium *et al.*, Mapping human genetic diversity in Asia. *Science*. **326**, 1541–5 (2009).
67. N. Tranter, in *The Languages of Japan and Korea*, N. Tranter, Ed. (Routledge, 2012), pp. 3–23.
68. J. Whitman, in *The Languages of Japan and Korea*, N. Tranter, Ed. (Routledge, 2012), pp. 24–37.
69. M. Robbeets, *Diachrony of verb morphology: Japanese and the Transeurasian languages* (DE GRUYTER MOUTON, 2015).
70. T. Sato *et al.*, Origins and genetic features of the Okhotsk people, revealed by ancient mitochondrial DNA analysis. *J. Hum. Genet.* **52**, 618–627 (2007).
71. T. Dallas, "Ethnologue : Languages of the World." (1999), (available at https://www.ethnologue.com/).
72. Harald Hammarström, Sebastian Nordhoff, LangDoc: Bibliographic Infrastructure for Linguistic Typology. *Oslo Stud. Lang.* **3**, 31–43 (2011).
73. C. C. Chang *et al.*, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*. **4** (2015), doi:10.1186/s13742-015-0047-8.
74. S. R. Browning, B. L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
75. M. Rasmussen, Y. Li, S. Lindgreen, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* (2010).
76. I. Lazaridis *et al.*, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513**, 409–413 (2014).
77. S. A. Fedorova *et al.*, Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol. Biol.* **13**, 127 (2013).

78. G. R. Abecasis *et al.*, An integrated map of genetic variation from 1,092 human genomes. *Nature*. **491**, 56–65 (2012).
79. S. A. Fedorova *et al.*, Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol. Biol.* **13**, 127 (2013).
80. B. L. Browning, S. R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* **84**, 210–23 (2009).
81. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–75 (2007).
82. A. L. Price *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* (2006), doi:10.1038/ng1847.
83. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. **12**, 246 (2011).
84. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).
85. H. E. L. Lischer, L. Excoffier, PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*. **28**, 298–299 (2011).
86. D. Bakker *et al.*, Adding typology to lexicostatistics: A combined approach to language classification. *Linguist. Typology*. **13**, 169–181 (2009).
87. F. Rousset, genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–6 (2008).
88. T. Rzeszutek, P. E. Savage, S. Brown, The structure of cross-cultural musical diversity. *Proc. Biol. Sci.* **279**, 1606–12 (2012).
89. L. Excoffier, H. E. L. Lischer, Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–7 (2010).
90. S. Bougeard, S. Dray, Supervised Multiblock Analysis in R with the ade4 Package. *J. Stat. Softw.* **86**, 1–17 (2018).
91. S. Lê, J. Josse, F. Husson, FactoMineR: An Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
92. J. Josse, F. Husson, missMDA : A Package for Handling Missing Values in Multivariate Data Analysis. *J. Stat. Softw.* **70**, 1–31 (2016).
93. S. Dray, P. Legendre, P. R. Peres-Neto, Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM) (2006), doi:10.1016/j.ecolmodel.2006.02.015.
94. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2009).
95. D. H. Huson, D. Bryant, Application of Phylogenetic Networks in Evolutionary Studies, doi:10.1093/molbev/msj030.
96. T. Dallas, "Ethnologue : Languages of the World." (1999).
97. D. Borcard, P. Legendre, All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Modell.* **153**, 51–68 (2002).
98. P. Legendre, M.-J. Fortin, D. Borcard, Should the Mantel test be used in spatial analysis? *Methods Ecol. Evol.* **6**, 1239–1247 (2015).
99. Van Den Wollenberg, Arnold L, Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika*. **42**, 207–219 (1977).

Figure Legends

Figure 1. Geographic areas of 14 languages/populations. Because some of the areas overlap in space, they are plotted in two separate maps.

Figure 2. Neighbornet networks of the populations based on dimensionality-reduced distance matrices in SNPs, Lexicon, Grammar, Phonology, and Music (see Materials and Methods). Colors indicate language families: Selkup and Nganasan belong both to Uralic; Even and Evenk to Tungusic; Koryak and Chukchi to Chukotko-Kamchatkan.

Figure 3. Redundancy Analysis (RDA) between four pairs of factors (genetics, grammar, music, and phonology). Variance in the response explained by each explanatory variable; * indicates a significant association ($p \leqq 0.05$).

Figure 4. The larger the KLD (dark color), the more the distribution of the observed R^2 differs from that of the random permutations. KLD=0 (transparent color) suggests that there is no difference between the distribution of the observed and the permuted R^2.

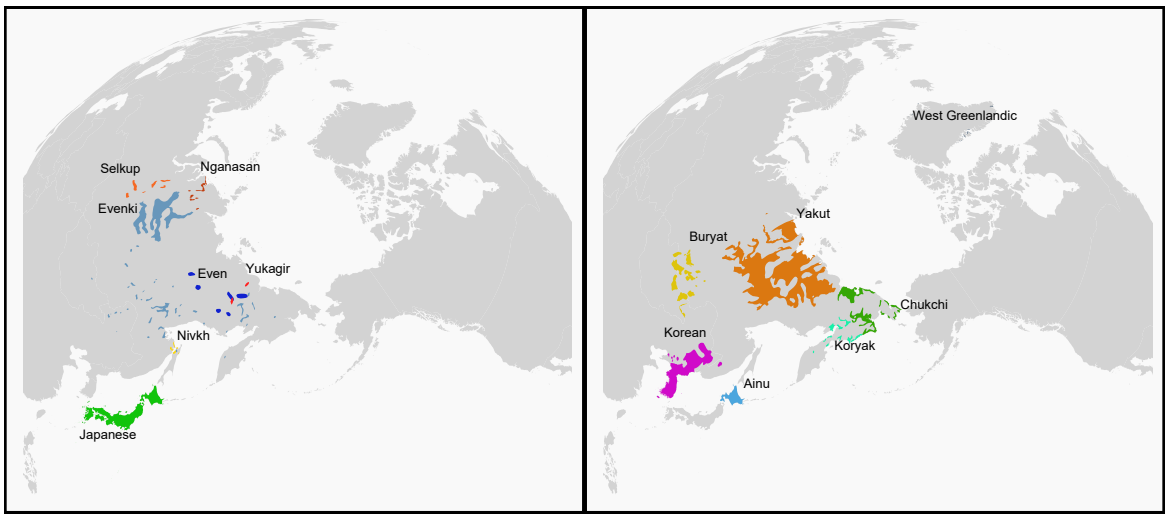
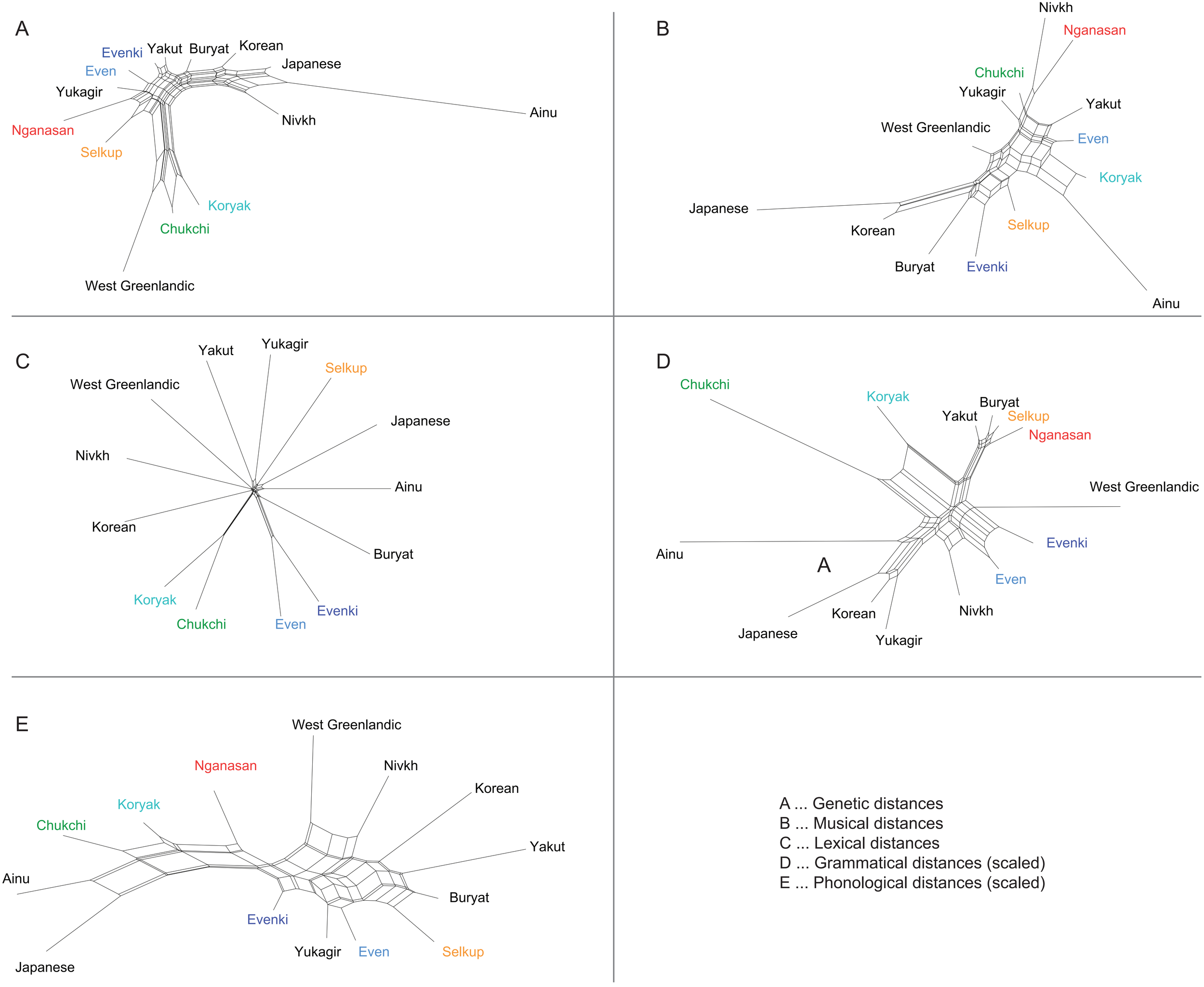


Figure 1: Geographical locations of the fourteen languages. Language polygons are plotted on two panels because they partly overlap in spatial distributions. Similarly-colored pairs of languages belong to the same family: Even and Evenki belong to the Tungusic family, Selkup and Nganasan to the Uralic family, and Koryak and Chukchi to the Chukotko-Kamchatkan family.



A ... Genetic distances
B ... Musical distances
C ... Lexical distances
D ... Grammatical distances (scaled)
E ... Phonological distances (scaled)

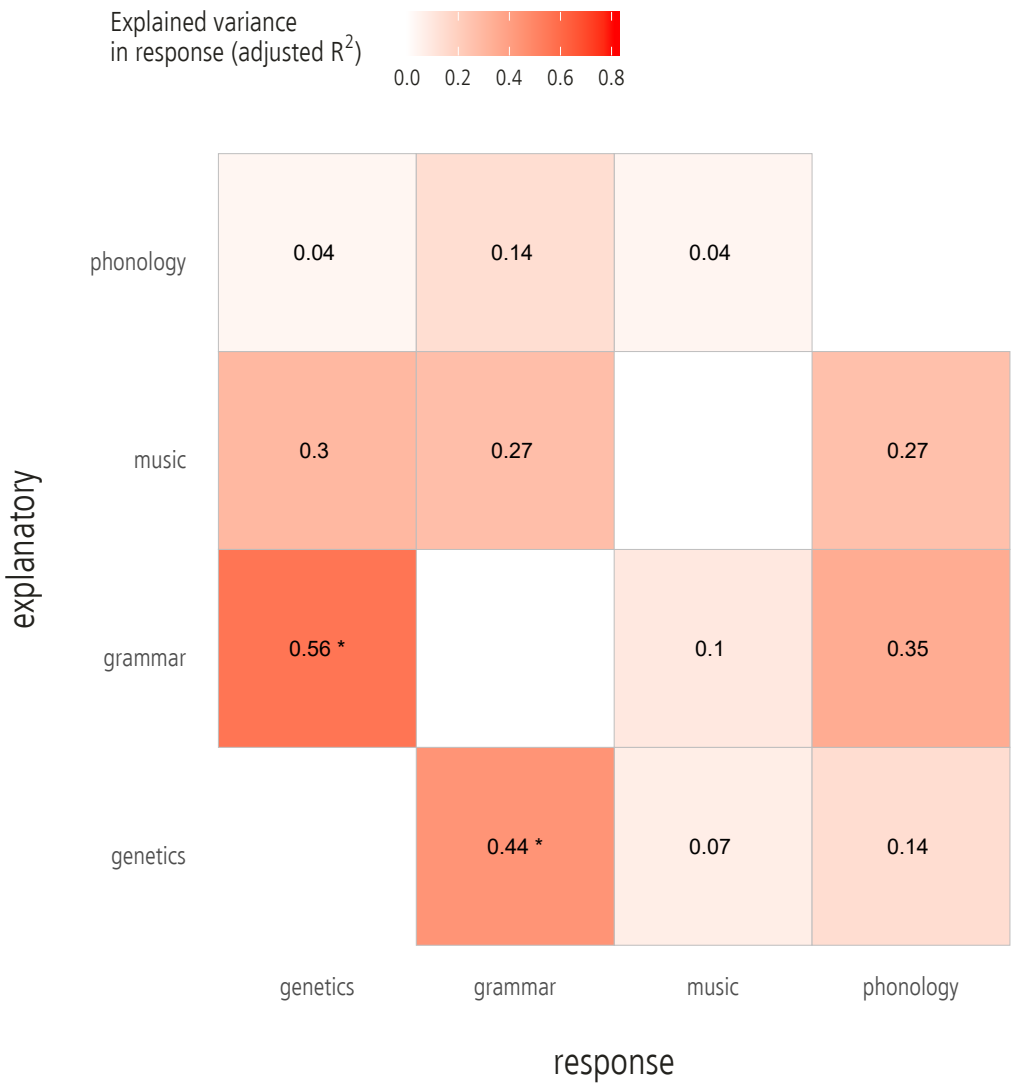


Figure 3: Redundancy Analysis (RDA) between four pairs of factors (genetics, grammar, music, and phonology). Variance in the response explained by each explanatory variable; * indicates a significant association $(p \leq 0.05)$.

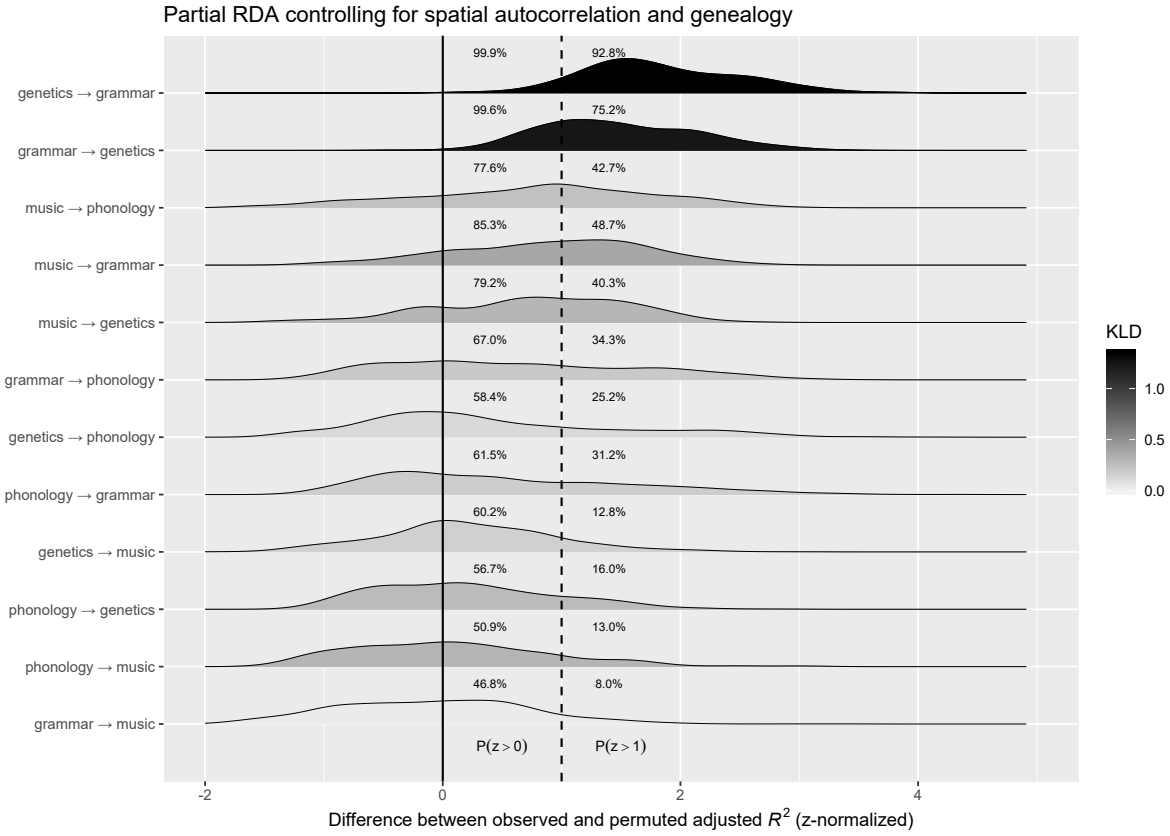


Figure 4: Partial RDA controlling for spatial autocorrelation and genealogy: Densities of the difference between observed and permuted adjusted R^2 (z-normalized) in the partial RDA. All input components contribute at least 15% to the explained variance. Numbers right to the solid and dashed line show the proportion of samples with a positive difference ($P(z > 0$ SD) and a strong positive difference ($P(z > 1$ SD), respectively. Grey shading reflects the Kullback-Leibler divergence (KLD) between the observed and permuted adjusted R^2. The larger the KLD (dark color), the more the distribution of the observed R^2 differs from that of the random permutations. KLD=0 (transparent color) suggests that there is no difference between the distribution of the observed and the permuted R^2.