ANTHROPOLOGY

Exploring correlations in genetic and cultural variation across language families in northeast Asia

Hiromi Matsumae[1,2]*†‡, Peter Ranacher[3,4]*‡, Patrick E. Savage[5,6]*, Damián E. Blasi[7,8,9,10,11], Thomas E. Currie[12], Kae Koganebuchi[13], Nao Nishida[14], Takehiro Sato[15], Hideyuki Tanabe[16], Atsushi Tajima[15], Steven Brown[17], Mark Stoneking[18], Kentaro K. Shimizu[1,2,19], Hiroki Oota[13,20,21]*, Balthasar Bickel[7,19]*

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 have empirically investigated these 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 (genetic 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 homogeneous 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*)], our cultural diversity is immense. For example, we speak or sign more than 7000 mutually unintelligible languages (*4*), and for each ethno-linguistic group, there tend to be many different musical styles (*5*). Researchers have long been interested in reconstructing the history of global migrations and diversification by combining historical and archeological 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 (*6*–*9*). However, differences in the ways that cultural traits and genomes are transmitted mean that genetic and cultural variation may be explained by different historical processes (*10*–*15*). 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 (*10*, *16*).

The cultural evolution of language has proven particularly fruitful for understanding past population history (genetic history statistically inferred from genetic variations) (*17*–*19*). A classic approach involves identifying and analyzing sets of homologous (cognate) words among languages. This lexical approach allows the reconstruction of evolutionary lineages and relationships within a single language family, such as Austronesian (*20*) or Indo-European (*17*, *18*). However, lexical methods cannot usually be applied to multiple language families (*19*), 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 (*20*, *21*). An alternative approach is to study the distribution of features of grammar

[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. [2]Kihara Institute for Biological Research, Yokohama City University, 641-12 Maioka-cho, Totsuka-ku, Yokohama, Kanagawa 244-0813, Japan. [3]Department of Geography, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland. [4]URPP Language and Space, University of Zurich, Freiestrasse 16, 8032 Zurich, Switzerland. [5]Faculty of Environment and Information Studies, Keio University, Shonan Fujisawa Campus, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan. [6]Department of Musicology, Tokyo University of the Arts, 110-8714 Tokyo, Japan. [7]Department of Comparative Language Science, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland. [8]Department of Human Evolutionary Biology, Harvard University, Peabody Museum, 5th Floor, 11 Divinity Avenue, Cambridge, MA 02138, USA. [9]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, 07745 Jena, Germany. [10]Linguistic Convergence Laboratory, School of Linguistics, Faculty of Humanities, Higher School of Economics University, 21/4 Staraya Basmannaya Ulitsa, Building 5, Moscow, Russian Federation. [11]Human Relations Area Files, 755 Prospect Street, New Haven, CT, USA. [12]Human Behaviour & Cultural Evolution Group, Centre for Ecology & Conservation, Department of Biosciences, University of Exeter, Penryn Campus, Penryn, Cornwall TR10 9FE, UK. [13]Kitasato University Graduate School of Medical Science, Sagamihara, Kanagawa 252-0374, Japan. [14]Genome Medical Science Project, Research Center for Hepatitis and Immunology, National Center for Global Health and Medicine, Chiba 272-8516, Japan. [15]Department of Bioinformatics and Genomics, Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa 920-8640, Japan. [16]Department of Evolutionary Studies of Biosystems, School of Advanced Sciences, The Graduate University for Advanced Studies, SOKENDAI, Shonan Village, Hayama, Kanagawa 240-0193, Japan. [17]Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada. [18]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D04103 Leipzig, Germany. [19]Center for the Interdisciplinary Study of Language Evolution (ISLE), Plattenstrasse 54, 8032 Zürich, Switzerland. [20]Kitasato University School of Medicine, Sagamihara, Kanagawa 252-0374, Japan. [21]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 113-0033, Japan.
*Corresponding author. Email: matsumae.hiromi.g@tokai.ac.jp (H.M.); psavage@sfc.keio.ac.jp (P.E.S.); peter.ranacher@geo.uzh.ch (P.R.); hiroki_oota@bs.s.u-tokyo.ac.jp (H.O.); balthasar.bickel@uzh.ch (B.B.)
†Present address: Tokai University School of Medicine, Shimokasuya 143, Isehara, Kanagawa 259-1143, Japan.
‡These authors contributed equally to this work.

and phonology, such as the relative order of word classes in sentences or the presence of nasal consonants. Structural data in language tend to evolve too fast to preserve phylogenetic signals of language families (*22*, *23*), and the history of lexica and structure might be partially independent as, for example, in the emergence of creole languages (*12*). However, 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 (*24*, *25*).

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 (*26*–*29*). 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 (*26*, *27*, *29*). Among indigenous Taiwanese populations speaking Austronesian languages, these analyses revealed significant correlations between music, mitochondrial DNA, and the lexicon (*27*), suggesting that music may preserve population history. However, whether these 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, and Japanese). 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 direct matched comparison of musical, genetic, and linguistic diversity (*30*, *31*).

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 (*32*), grammar (*33*, *34*), or phonology (*34*–*36*)] 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; Materials and Methods) (*29*). 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 inquiry. For this purpose, we used split networks (*37*), which display multiple sources of similarity in a consistent manner (Fig. 2, figs. S12 to S16, and tables S2 to S6). Distance analysis of lexical data resulted in a network topology with an overall star-shaped structure (Fig. 2C). Exceptions are given by the three pairs of languages that are related to one another and that stand out as proximate (Even and Evenki both belong to the Tungusic family, Chukchi and Koryak both belong to the Chukotko-Kamchatkan family, and Selkup and Nganasan both belong to the Uralic family) (*4*). 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.

Distance analyses of grammatical, phonological, genetic, and musical distances reveal potentially more informative structure. In agreement with the claim that language structure does not identify family relationships (*20*, *22*), the clustering emerging from the distances does not generally coincide with language families, except for Chukotko-Kamchatkan (Chukchi and Koryak) in genetics and

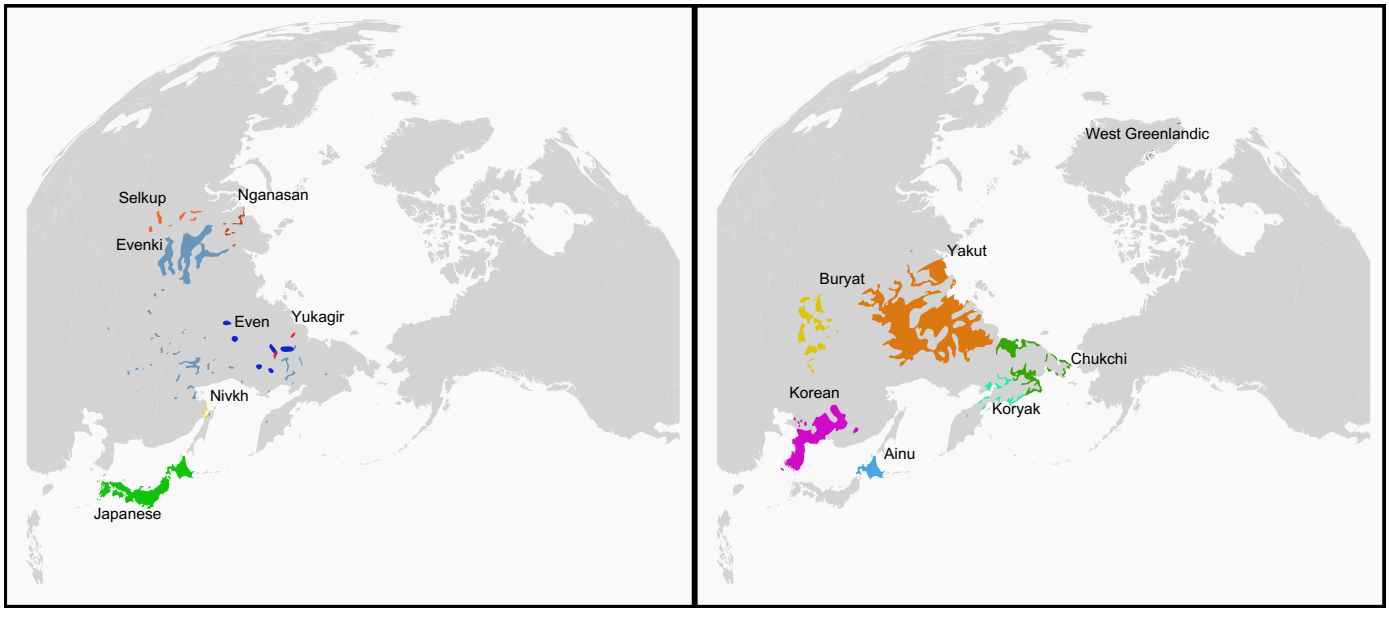

Fig. 1. Geographic areas of 14 languages/populations. Because some of the areas overlap in space, they are plotted in two separate maps.

phonology (where the within-family distance d_{fam} is smaller than the distance d_{nun} to the next unrelated neighbor, relative to the total distance range: genetics $d_{fam} = 0.15 < d_{nun} = 0.26$; phonology $d_{fam} = 0.28 < d_{nun} = 0.36$ (Supporting Information 1, section 4.1), and marginally for Tungusic (Even and Evenki) in grammar ($d_{fam} = 0.22 < d_{nun} = 0.28$). 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 (*38*). Buryat and Yakut are close together in SNPs (*39*), grammar, and phonology, but not in music. 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 (*29*). 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 (table S3), reflecting the tree's branch position. However, music, grammar, and phonology do not follow these relationships in Nivkh.

Together, these results suggest that neither the population history nor the cultural features (other than the lexicon) evolved by simple vertical descent along language families. Instead, apart from the possible case of Chukotko-Kamchatkan, they might have each followed independent trajectories. While this challenges the idea of a unified phylogeny, it leaves open the possibility that some of the features are associated with each other because they trace back to a prehistoric maze of horizontal and vertical transmission. In other words, features might still be associated with each other because they

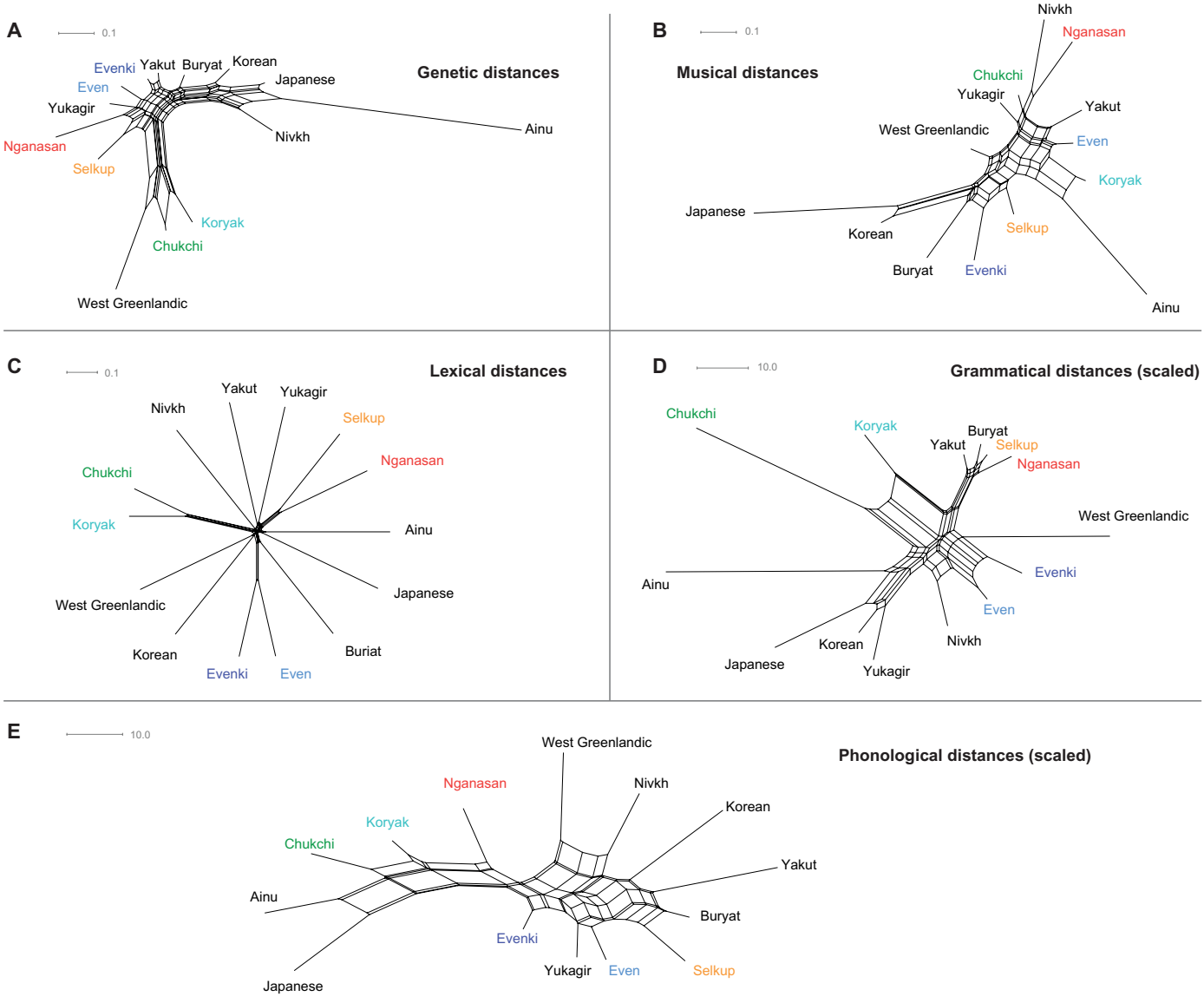

Fig. 2. Neighbornet networks of the populations based on dimensionality-reduced distance matrices in SNPs, lexicon, grammar, phonology, and music (see Materials and Methods). Colors indicate language families: Selkup and Nganasan belong both to Uralic; Even and Evenki to Tungusic; and Koryak and Chukchi to Chukotko-Kamchatkan.

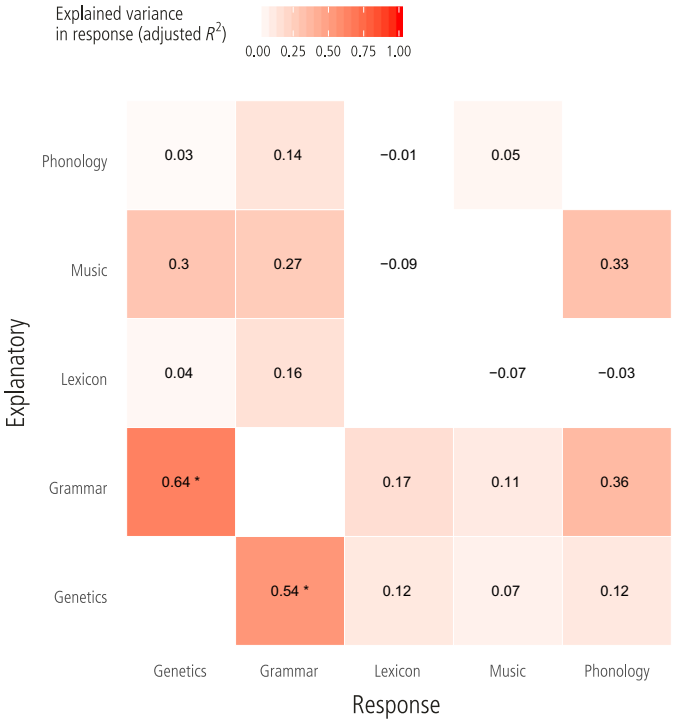

Fig. 3. RDA between five pairs of factors (lexicon, genetics, grammar, music, and phonology). Variance in the response explained by each explanatory variable; * indicates a significant association ($P \leq 0.05$).

were present in the same period(s) and places in which people were in contact and/or were genetically related. To find out whether any such association is still detectable today, we implemented a redundancy analysis (RDA) on the principal components (or coordinates) of the data (Materials and Methods and Supporting Information 1). 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, adjusted R^2 = 0.64), and genetic similarity predicts grammatical similarity (genetics → grammar, adjusted R^2 = 0.54).

While both associations possibly 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. To find out, we evaluated three possible scenarios to explain the signal in the data: (i) 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. (ii) 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). (iii) Deep-time correspondence scenario: The associations reflect a nonshallow correspondence between grammar and genetics that cannot be explained by recent contact or 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 (Supporting Information 1, section 5). As societies and languages placed far from the equator tend to display larger spatial ranges (*40*), we represented the territory of each society with areas rather than points and sampled 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 (figs. S18 to S20). When controlling for spatial autocorrelation (1000 random samples allowing 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 more than 99% of spatial samples; Kullback-Leibler divergence (KLD) > 3; fig. S20 and table S7]. When controlling for both recent contact and phylogenetic inheritance of language in partial RDA, still both associations show stronger evidence than the other relationships ($z > 1$ SD in ≥90% of samples, KLD ≥ 1.5; Fig. 4, figs. S21 to S23, and table S8). Our analysis reveals no other associations at comparable strengths; there are a few weak signals (e.g., grammar, music, and phonology; Fig. 2), but they all disappear once we control for both spatial autocorrelation and genealogy (Fig. 4 and table S8), suggesting that any patterns here are likely to stem from recent contact and family-specific lines of inheritance.

Given the relatively small sample of only 14 groups, we evaluate the robustness of the grammar/genetics associations through 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 to be included (10%, 15%, and 18%) show little effect on the results ($z > 1$ SD in at least 84%, KLD > 1.2; figs. S24 and S25 and table S9). 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 14 to 31%, KLD ≤ 0.2, when controlling for spatial proximity and inheritance; figs. S26 to S29 and tables S10 and S11). 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 figs. S21 and S22). Spatial clusters of locations with low adjusted R^2 might indicate recent language contact (see section 5.4, Supporting Information 1), and clusters with high adjusted R^2 might indicate that systematic outliers influence the signal. We mapped locations in the 0.2 (figs. S30 and S31) and 0.8 percentile (figs. S32 and S33). We find only weak and partial clustering in the high percentile, and none in the low percentile. This suggests that neither recent contact nor systematic outliers explain the signal.

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 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. We

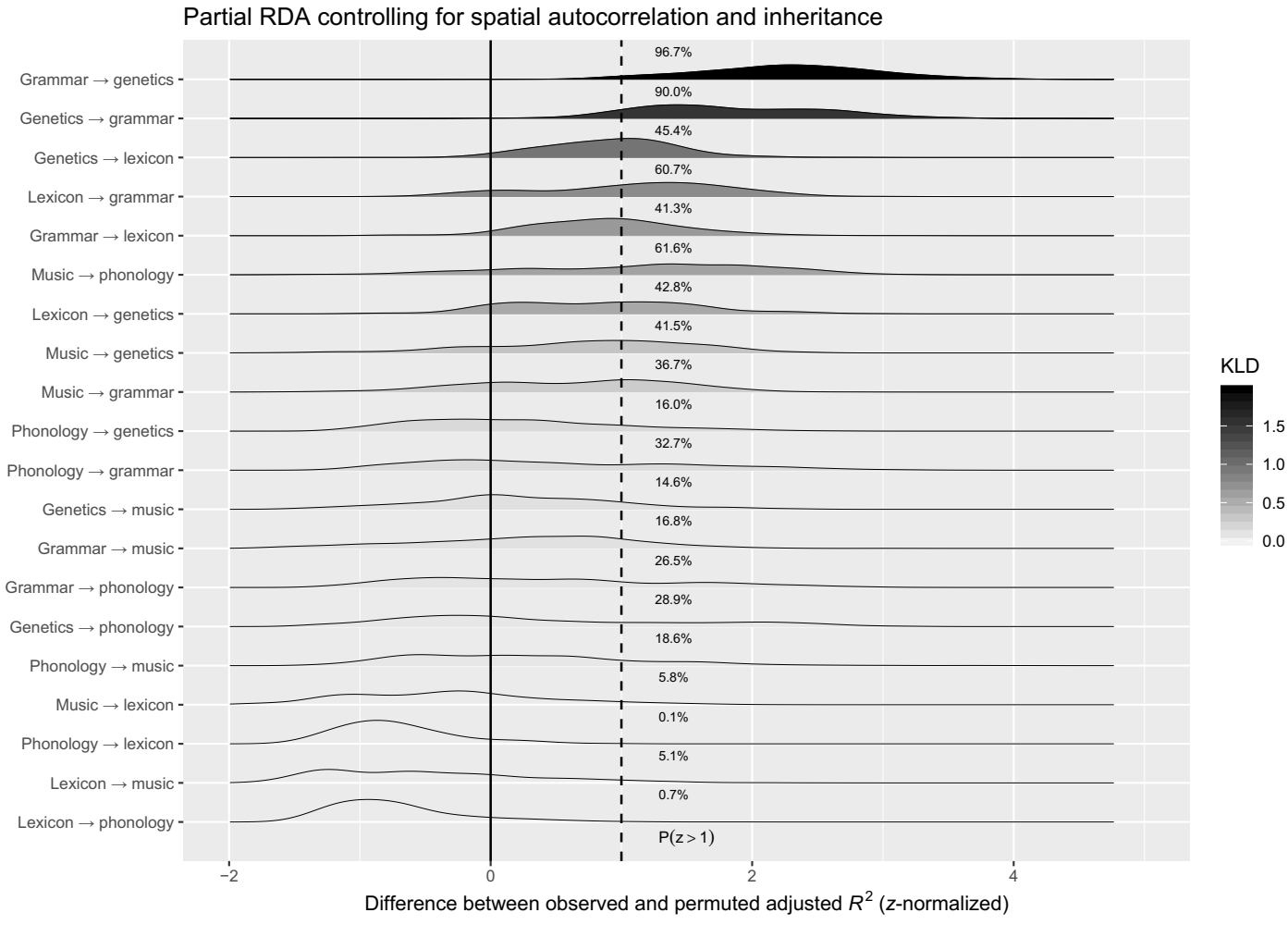

Fig. 4. Partial RDA controlling for spatial autocorrelation and linguistic inheritance: Densities of the difference between observed and permuted adjusted R^2 (*z*-normalized) in the partial RDA. Numbers right to the dashed line show the proportion of samples with a difference of at least one SD. Gray shading reflects the KLD between the observed and permuted adjusted R^2. The KLD is transparent for associations where the *z*-normalized difference is negative for more than 50% of the samples.

find remarkable evidence for the relationships between population history and grammatical similarity, while genomes and grammar might be influenced by different evolutionary forces, such as a difference between mating systems and cultural transmission (*13*).

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, regardless of whether or not the evolutionary mechanisms of grammar are the same as those of 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) (*41*), including gene flows beyond the lineages, such as Japanese-Ainu (*38*) and Buryat-Yakut (*39*). 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 versus prehistoric contact in culture is still challenging. Moreover, the evolutionary processes that influence culture and language are under debate (*14*) but can obviously be very different from those influencing genomes. For example, cultural replacement and language shift can occur even within a single generation due to colonization or other sociopolitical factors, like warfare and cultural expansion (*15*, *42*). 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 can still preserve relationships extending even beyond the boundaries of language families. The similarities in grammar do not arise from simply following the genetic phylogeny (see Fig. 2D, which lacks the Korean-Japanese-Nivkhh-Ainu and Koryak-Chukchi-West Geenlandic clusters in Fig. 2A). Instead, they are likely to reflect a complex interplay of partially independent vertical and horizontal transmission in prehistory.

This pattern is markedly different for the lexicon that traces language families but does not reveal higher-level relationships in our dataset (Fig. 2). This contrasts with expectations from historical linguistics (*22*) and also from recent findings that suggest that grammar evolves faster than the lexicon in Austronesian (*23*) and also shows rapid evolution in Indo-European (*43*); for example,

while English and Hindi preserve many cognate words (*name* versus *nām*, *hand* versus *hāth*, etc.), they differ substantially in word order (verb-medial versus verb-final) and case-marking (invariable nouns versus complex case system). However, these findings bear on grammatical evolution within families, while our approach seeks to unravel a shared history that allows early contact between families. Therefore, our findings are compatible with a scenario where specific traits (e.g., word order) evolved rapidly within families but were repeatedly copied and readapted, yielding a relatively uniform profile over a prehistoric period (*44*) that mirrors the genetic network of the same period.

The statistical power to detect a signal is weakened when Ainu was removed in the sensitivity analysis (figs. S26 to S29 and table S10). While this might suggest a special position of Ainu in the northeast Asian context (*45*), we need larger samples of languages and populations inside and outside of the region to resolve this question.

Our results are qualitatively different from the only previous study to quantitatively compare genetic, linguistic, and musical relationships (*27*). 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 versus 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 (*46*). The high linguistic diversity in northeast Asia may reflect prehistorical relationships with less influence from agricultural populations by geographic barriers, as hypothesized in the previous studies (*24*, *47*). 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, potentially including cultural shift. For example, while previous studies suggest specific genetic and cultural relationships between Korean and mainland Japanese populations (*38*) or posit a shared origin (*48*, *49*), our findings support similarities in SNPs, music, and grammar, but not in lexicon and phonology (Fig. 2 and Supporting Information 1) (*50*). Although the Ainu show particular genetic similarity to the Japanese, their music clusters more closely with that of the Koryak (Fig. 2 and tables S3 and S4). 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 1500 years (*51*), as we previously proposed in our "triple structure" model of Japanese archipelago history (*29*). Newly genotyped Nivkh samples showed the closeness to Ainu in SNPs but not in others (Fig. 2A), suggesting 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. Thus, Neighbornet trees might reflect the relationships linking populations, but further analyses are necessary to investigate, in more detail, the local population history and cultural relationships in northeast Asia including Nivkh. Most pressingly, future research will need a larger sample of societies and a richer coding of their cultural traits.

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 these mismatches would be historical events (e.g., language shift in local history) or culture-specific evolution independent from genetic evolution. Future analyses of these 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 and S13 and Fig. 1). These represented a subset of 35 northeast Asian populations whose musical relationships were previously published and analyzed in detail (*29*). Linguistically, these 14 populations fall into 11 language families/isolates (*4*). Korean, Ainu, Nivkh, and Yukaghir are language isolates. Buryat, Japanese, Yakut, and 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 Evenki are Tungusic languages; and Selkup and Nganasan are Uralic languages. Note that the need to assemble matching genetic, linguistic, and musical data meant that some important populations could not be included (e.g., we had matching musical and genetic data for multiple Ryukyuan populations, but no corresponding grammatical data were available, while for the Aleut genetic and linguistic data were available but not musical). Future research should attempt to collect new data to allow more complete comparisons within and between language families.

Music data

All music data and metadata are detailed in our previous report of circumpolar music (*29*). 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 (*52*)].

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 S. Horai in the 1990s (*53*) and were kept at 4°C 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_S16_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 (*54*), 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 1000 Genomes using BEAGLE 4.0 (*55*). 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) (*38*, *56*–*59*), 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 for a dataset of Ainu against JPT and CHB in 1000 Genomes. 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.

Removing outlier individuals. We manually removed outlier individuals from the merged dataset based on results of principal components analysis (PCA) and ADMIXTURE (*60*–*62*). Last, we used 15 individuals of Nivkh (13 individuals from our data and 2 individuals from public data) in the population genetics analysis (tables S1 and S16). 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 (*63*).

Language data

Lexical data. We measured lexical distances between those words in the ASJP (Automated Similarity Judgment Program) database v. 19 (*32*) that have best coverage in our sample, corresponding to 40 concepts that are attested in at least 74% of all word lists. These correspond to the concepts commonly thought to be most stable over time (*64*) and to best reflect language relatedness, at least as a first approximation (Supporting Information 3) (*65*).

Grammar and phonology data. We combined data on grammatical and phonological traits from AUTOTYP (*34*, *66*), WALS (*33*), the ANU Phonotactics database (*35*), and PHOIBLE (*36*) and extracted a set of 25 grammar and 87 phonological features with coverage more than 80% in each language, and in most cases 100% (Supporting Information 2 and table S13).

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 (*67*). 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 (*68*) to calculate the weighted average similarity across all 41 musical features (*29*). This distance matrix was then used to compute a distance matrix of pairwise musical φ_{st} values among the 14 populations using Arlequin (*69*) and the lingos function of the ade4 package in R (*70*). φ_{st} is analogous to F_{st} but takes into account distances between individual items, making it more appropriate for analysis of cultural diversity (*68*, *70*). Further details concerning the calculations can be found elsewhere (*70*).

Language analysis

Lexical data. For the main analysis, we compute distances in ASJP word alignments weighted by sound correspondence probabilities, a method that provides good first approximations of language relatedness (Supporting Information 3, table S14, and fig. S34) (*65*). For comparability with other ASJP-based work, we also report normalized Levenshtein distances (Supporting Information 3, table S15, and fig. S35).

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 PCA and multiple correspondence analysis (MCA) to calculate differences between languages (Supporting Information 1, section 3) (*71*). Empty values were imputed using the R package missMDA (*72*).

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 1, section 3) (*73*). Similar to a PCA, a PCoA produces a set of orthogonal axes whose importance is measured by eigenvalues (figs. S2 to S6). However, in contrast to the PCA, non-Euclidean distance matrices can be used. Heat plots of PCo and PC were visualized by ggplot2 in R (figs. S7 to S11) (*74*).

Split network graphs. Distances were visualized using the SplitsTree neighbornet algorithm [version 4; (*37*)] and are reported in detail in Supporting Information 1, tables S2 to S6, and figs. S12 to S16. 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 (*75*) via the World Language Mapping System (*76*), supplemented by a hand-drawn polygon estimate for Ainu.

In view of the mobility of speakers over time, we sampled 1000 random locations from within the polygons and used these for assessing correlations. Location samples were always taken from geometries (i.e., polygons on a sphere) and not from a potentially distorted image of these geometries on a map. Location samples were generated in PostGIS https://postgis.net/ (Supporting Information 1, section 2.4). For each of the 1000 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 1000 distance matrices (Supporting Information 1, section 3.3) (*77*). 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) RDA. RDA was carried out to explore the linear relationship between SNPs, grammar, phonology, and music. Partial RDA was used to control for spatial dependence (Supporting Information 1, section 5) (*78*). (Partial) 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 (*79*). RDA performs a regression of multiple response variables on multiple predictor variables (*80*), while partial RDA also allows to control for the influence of confounders. RDA 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 and figs. S18 to S23). 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 SD larger than the permuted ($z > 1$ SD). Moreover, we compute the 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 (*65*).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/7/34/eabd9223/DC1

REFERENCES AND NOTES

1. M. Pagel, R. Mace, The cultural wealth of nations. *Nature* **428**, 275–278 (2004).
2. J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, M. Stephens, C. D. Bustamante, Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
3. D. E. Brown, *Human Universals* (Temple Univ. Press, 1991).
4. H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, Glottolog 4.3 (2020); http://glottolog.org.
5. B. Nettl, *The Study of Ethnomusicology: Thirty-Three Discussions* (University of Illinois Press, 2015).
6. C. Darwin, *The Descent of Man, and Selection in Relation to Sex* (J. Murray, 1871), volume 1.
7. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, 1994).
8. E. Bortolini, L. Pagani, E. R. Crema, S. Sarno, C. Barbieri, A. Boattini, M. Sazzini, S. G. da Silva, G. Martini, M. Metspalu, D. Pettener, D. Luiselli, J. J. Tehrani, Inferring patterns of folktale diffusion using genomic data. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 9140–9145 (2017).
9. R. Boyd, M. Bogerhoff-mulder, W. H. Durham, P. J. Richerson, Are cultural phylogenies possible? *Hum. Nat. Biol. Soc. Sci.* , 355–386 (1997).
10. P. J. Richerson, R. Boyd, *Not by Genes Alone: How Culture Transformed Human Evolution* (University of Chicago Press, 2005).
11. R. M. MacCallum, M. Mauch, A. Burt, A. M. Leroi, Evolution of music by public choice. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12081–12086 (2012).
12. 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).
13. R. D. Gray, S. J. Greenhill, R. M. Ross, The pleasures and perils of darwinizing culture (with Phylogenies). *Biol. Theory* **2**, 360–375 (2007).
14. A. Mesoudi, *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences* (University of Chicago Press, 2011).
15. M. Stoneking, *An Introduction to Molecular Anthropology* (Wiley-Blackwell, 2016).
16. 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).
17. R. D. Gray, A. J. Drummond, S. J. Greenhill, Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
18. W. Chang, C. Cathcart, D. Hall, A. Garrett, Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91**, 194–244 (2015).
19. 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 Eur. Stud.* **46**, 1–264 (2018).
20. D. A. Ringe Jr., `Nostratic' and the factor of chance. *Diachronica* **12**, 55–74 (1995).
21. R. Gray, Pushing the time barrier in the quest for language roots. *Science* **309**, 2007–2008 (2005).
22. J. Nichols, in *The Comparative Method Reviewed*, M. Durie, M. Ross, Eds. (Oxford Univ. Press, 1996), pp. 39–71.
23. S. J. Greenhill, C. H. Wu, X. Hua, M. Dunn, S. C. Levinson, R. D. Gray, Evolutionary dynamics of language systems. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8822–E8829 (2017).
24. J. Nichols, *Linguistic Diversity in Space and Time* (University of Chicago Press, 1999).
25. B. Bickel, J. Nichols, Oceania, the Pacific Rim, and the theory of linguistic areas. *Annu. Meet. Berkeley Linguist. Soc.* **32**, 3–15 (2006).
26. A. Lomax, American association for the advancement of science, in *Folk Song Style and Culture* (Transaction Books, 1978).
27. S. Brown, P. E. Savage, A. M. S. Ko, M. Stoneking, Y. C. Ko, J. H. Loo, J. A. Trejaut, Correlations in the population structure of music, genes and language. *Proc. R. Soc. B Biol. Sci.* **281**, 20132072 (2014).
28. H. Pamjav, Z. Juhász, A. Zalán, E. Németh, B. Damdin, A comparative phylogenetic study of genetics and folk music. *Mol. Genet. Genomics* **287**, 337–349 (2012).
29. P. E. Savage, H. Matsumae, H. Oota, M. Stoneking, T. E. Currie, A. Tajima, M. Gillan, S. Brown, How 'circumpolar' is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. *Ethnomusicol. Forum* **24**, 443–467 (2015).
30. P. E. Savage, Alan Lomax's cantometrics project: A comprehensive review. *Music Sci.* **1**, (2018).
31. P. E. Savage, S. Brown, E. Sakai, T. E. Currie, Statistical universals reveal the structures and functions of human music. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8987–8992 (2015).
32. S. Wichmann, E. W. Holman, C. H. Brown, R. Forkel, T. Tresold, The ASJP Database (version 19) (2020); https://doi.org/10.5281/zenodo.3843469.
33. M. S. Dryer, M. Haspelmath, The world atlas of language structures online (2013); http://wals.info/.
34. B. Bickel, J. Nichols, T. Zakharko, A. Witzlack-Makarevich, K. Hildebrandt, M. Rießler, L. Bierkandt, F. Zúñiga, J. B. Lowe, The AUTOTYP typological databases. Version 0.1.0 (2017); https://github.com/autotyp/autotyp-data/tree/0.1.0.
35. M. Donohue, R. Hetherington, J. McElvenny, V. Dawson, World Phonotactis Database (2013); http://phonotactics.anu.edu.au/ [accessed 20 December 2015]; https://web.archive.org).
36. S. Moran, D. McCloy, R. Wright, PHOIBLE Online (2014); https://phoible.org.
37. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
38. Japanese Archipelago Human Population Genetics Consortium, T. Jinam, N. Nishida, M. Hirai, S. Kawamura, H. Oota, K. Umetsu, R. Kimura, J. Ohashi, A. Tajima, T. Yamamoto, H. Tanabe, S. Mano, Y. Suto, T. Kaname, K. Naritomi, K. Yanagi, N. Niikawa, K. Omoto, K. Tokunaga, N. Saitou, 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**, 787–795 (2012).
39. I. Pugach, R. Matveev, V. Spitsyn, S. Makarov, I. Novgorodov, V. Osakovsky, M. Stoneking, B. Pakendorf, The complex admixture history and recent southern origins of Siberian populations. *Mol. Biol. Evol.* **33**, 1777–1795 (2016).
40. M. C. Gavin, J. R. Stepp, Rapoport's rule revisited: Geographical distributions of human languages. *PLOS ONE* **9**, e107623 (2014).
41. T. Gakuhari, S. Nakagome, S. Rasmussen, M. E. Allentoft, T. Sato, T. Korneliussen, B. N. Chuinneagáin, H. Matsumae, K. Koganebuchi, R. Schmidt, S. Mizushima, O. Kondo, N. Shigehara, M. Yoneda, R. Kimura, H. Ishida, T. Masuyama, Y. Yamada, A. Tajima, H. Shibata, A. Toyoda, T. Tsurumoto, T. Wakebe, H. Shitara, T. Hanihara, E. Willerslev, M. Sikora, H. Oota, Ancient Jomon genome sequence analysis sheds light on migration patterns of early East Asian populations. *Commun. Biol.* **3**, 437 (2020).
42. B. Pakendorf, in *The Routledge Handbook of Historical Linguistics*, C. Bowern, B. Evans, Eds. (Routledge, ed. 1, 2015), pp. 627–641.
43. M. Widmer, S. Auderset, P. Widmer, J. Nichols, B. Bickel, NP recursion over time: Evidence from Indo-European. *Language* **93**, 1–36 (2017).

44. B. Bickel, in *Language Dispersal, Diversification, and Contact*, M. Crevels, P. Muysken, Eds. (Oxford Univ. Press, 2020), pp. 78–101.
45. A. Bugaeva, J. Nichols, B. Bickel, Appositive possession in Ainu and around the Pacific. *Linguistic Typology* (2021); https://doi.org/10.1515/lingty-2021-2079.
46. C. Jeong, O. Balanovsky, E. Lukianova, N. Kahbatkyzy, P. Flegontov, V. Zaporozhchenko, A. Immel, C. C. Wang, O. Ixan, E. Khussainova, B. Bekmanov, V. Zaibert, M. Lavryashina, E. Pocheshkhova, Y. Yusupov, A. Agdzhoyan, S. Koshel, A. Bukin, P. Nymadawa, S. Turdikulova, D. Dalimova, M. Churnosov, R. Skhalyakho, D. Daragan, Y. Bogunov, A. Bogunova, A. Shtrunov, N. Dubova, M. Zhabagin, L. Yepiskoposyan, V. Churakov, N. Pislegin, L. Damba, L. Saroyants, K. Dibirova, L. Atramentova, O. Utevska, E. Idrisov, E. Kamenshchikova, I. Evseeva, M. Metspalu, A. K. Outram, M. Robbeets, L. Djansugurova, E. Balanovska, S. Schiffels, W. Haak, D. Reich, J. Krause, The genetic history of admixture across inner Eurasia. *Nat. Ecol. Evol.* **3**, 966–976 (2019).
47. P. Diamond, J. Bellwood, Farmers and their languages: The first expansions. *Science* **300**, 597–603 (2003).
48. M. Robbeets, *Diachrony of Verb Morphology: Japanese and the Transeurasian Languages* (De Gruyter Mouton, 2015).
49. N. Tranter, *Languages of Japan and Korea* (Routledge, 2012).
50. A. Ceolin, C. Guardiano, M. A. Irimia, G. Longobardi, Formal syntax and deep history. *Front. Psychol.* **11**, 488871 (2020).
51. T. Sato, T. Amano, H. Ono, H. Ishida, H. Kodera, H. Matsumura, M. Yoneda, R. Masuda, Origins and genetic features of the Okhotsk people, revealed by ancient mitochondrial DNA analysis. *J. Hum. Genet.* **52**, 618–627 (2007).
52. P. E. Savage, E. Merritt, T. Rzeszutek, S. Brown, CantoCore: A new cross-cultural song classification scheme. *Anal. Approaches World Music* **2**, 87–137 (2012).
53. V. Gurtsevitch, N. Senyuta, J. Shih, V. Stepina, O. Pavlish, A. Syrtsev, O. Susova, L. Yakovleva, L. Scherbak, M. Hayami, HTLV-I infection among Nivkhi people in Sakhalin. *Int. J. Cancer* **60**, 432–433 (1995).
54. C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
55. 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).
56. M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. P. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, M. Bunce, E. D. Lorenzen, J. Binladen, X. Guo, J. Zhao, X. Zhang, H. Zhang, Z. Li, M. Chen, L. Orlando, K. Kristiansen, M. Bak, N. Tommerup, C. Bendixen, T. L. Pierre, B. Grønnow, M. Meldgaard, C. Andreasen, S. A. Fedorova, L. P. Osipova, T. F. G. Higham, C. B. Ramsey, T. v. O. Hansen, F. C. Nielsen, M. H. Crawford, S. Brunak, T. Sicheritz-Pontén, R. Villems, R. Nielsen, A. Krogh, J. Wang, E. Willerslev, Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
57. I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J. M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J. M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, J. Krause, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
58. S. A. Fedorova, M. Reidla, E. Metspalu, M. Metspalu, S. Rootsi, K. Tambets, N. Trofimova, S. I. Zhadanov, B. Kashani, A. Olivieri, M. I. Voevoda, L. P. Osipova, F. A. Platonov, M. I. Tomsky, E. K. Khusnutdinova, A. Torroni, R. Villems, Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): Implications for the peopling of Northeast Eurasia. *BMC Evol. Biol.* **13**, 127 (2013).
59. 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. De Pristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, G. A. M. Vean, An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
60. A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
61. D. H. Alexander, K. Lange, Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
62. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLOS Genet.* **2**, e190 (2006).
63. H. E. L. Lischer, L. Excoffier, PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* **28**, 298–299 (2011).
64. E. W. Holman, S. Wichmann, C. H. Brown, V. Velupillai, A. Müller, D. Bakker, Explorations in automated language classification. *Folia Linguist.* **42**, 331–354 (2008).
65. G. Jäger, Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data* **5**, 180189 (2018).
66. B. Bickel, T. Zakharko, Recoding of Wals Online (2018); https://github.com/IVS-UZH/WALS-recodings.
67. F. Rousset, genepop'007: A complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8**, 103–106 (2008).
68. T. Rzeszutek, P. E. Savage, S. Brown, The structure of cross-cultural musical diversity. *Proc. Biol. Sci.* **279**, 1606–1612 (2012).
69. 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**, 564–567 (2010).
70. S. Bougeard, S. Dray, Supervised multiblock analysis in R with the ade4 package. *J. Stat. Softw.* **86**, 1–17 (2018).
71. S. Lê, J. Josse, F. Husson, FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
72. J. Josse, F. Husson, missMDA: A package for handling missing values in multivariate data analysis. *J. Stat. Softw.* **70**, 1–31 (2016).
73. S. Dray, P. Legendre, P. R. Peres-Neto, Spatial modelling: A comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol. Model.* **196**, 483–493 (2006).
74. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2009).
75. D. M. Eberhard, G. F. Simons, C. D. Fennig, *Ethnologue: Languages of the World* (SIL International, ed. 24, 2021); https://www.ethnologue.com/.
76. Global Mapping International (GMI), World Language Mapping System (WLMS); http://www.worldgeodatasets.com/language.
77. D. Borcard, P. Legendre, All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model.* **153**, 51–68 (2002).
78. D. Borcard, P. Legendre, Drapeau, Partialling out the Spatial Component of Ecological Variation. *Ecology* **73**, 1045–1055 (1992).
79. P. Legendre, M.-J. Fortin, D. Borcard, Should the Mantel test be used in spatial analysis? *Methods Ecol. Evol.* **6**, 1239–1247 (2015).
80. A. L. van den Wollenberg, Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* **42**, 207–219 (1977).
81. J. Dellert, T. Daneyko, A. Münch, A. Ladygina, A. Buch, N. Clarius, I. Grigorjew, M. Balabel, H. I. Boga, Z. Baysarova, R. Mühlenbernd, J. Wahle, G. Jäger, NorthEuraLex: A wide-coverage lexical database of Northern Eurasia. *Lang. Resour. Eval.* **54**, 273–301 (2020).
82. G. Jäger, Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Lang. Dyn. Chang.* **3**, 245–291 (2013).
83. J. Good, M. Cysouw, Languoid, Doculect, and Glossonym: Formalizing the Notion 'Language'. *Lang. Doc. Conserv.* **7**, (2013).
84. J.-M. List, S. J. Greenhill, R. D. Gray, The potential of automatic word comparison for historical linguistics. *PLOS ONE* **12**, e0170046 (2017).
85. S. J. Greenhill, Levenshtein distances fail to identify language relationships accurately. *Comput. Linguist.* **37**, 689–698 (2011).

Acknowledgments: We thank the Asian DNA Repository Consortium for the use of Ainu SNP 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 S. Wichmann for providing the ASJP distance matrix and B. Pakendorf, J. Nichols, J. Janhunen, E. Gruzdeva, A. Bugaeva, A. Berge, D. Jung, N. Neureiter, M. Sánchez, and C. Barbieri for discussion. **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.; and the HSE

University Basic Research Program, funded by the Russian Academic Excellence Project '5-100' to D.E.B. **Author contributions:** P.E.S., H.M., H.O., B.B., and K.K.S. initially designed the research with advice from T.E.C., M.S., and S.B. K.K., N.N., and H.T. genotyped the Nivkh DNA samples. T.S., A.T., H.O., and H.M. analyzed the genetic data. P.R., D.E.B., and B.B. analyzed the language data. P.E.S., H.M., and P.R. analyzed the 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 Supplementary Materials. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Data and analysis code are available at https://github.com/derpetermann/music_languages_genes. 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 SNP data, which have not been authorized for deposition in public data repositories. We obtained permission to analyze Ainu data from the corresponding authors of the original paper (*38*); future requests to access these data should also be addressed to those authors. Please contact H.M. and H.O. to request the raw Nivkh SNP data, which were provided by the Asian DNA Repository Consortium and are available upon reasonable request with appropriate approval of the human genomic DNA research ethics committee. All data and analysis code needed to reproduce and evaluate the conclusions in the paper are present.

Submitted 15 November 2020
Accepted 20 May 2021
Published 18 August 2021
10.1126/sciadv.abd9223

Citation: H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Koganebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Sci. Adv.* **7**, eabd9223 (2021).

Exploring correlations in genetic and cultural variation across language families in northeast Asia

Hiromi Matsumae, Peter Ranacher, Patrick E. Savage, Damián E. Blasi, Thomas E. Currie, Kae Koganebuchi, Nao Nishida, Takehiro Sato, Hideyuki Tanabe, Atsushi Tajima, Steven Brown, Mark Stoneking, Kentaro K. Shimizu, Hiroki Oota and Balthasar Bickel

Sci Adv **7** (34), eabd9223.
DOI: 10.1126/sciadv.abd9223