Title: Exploring deep-time relationships between cultural and genetic evolution in Northeast Asia

Authors

Hiromi Matsumae[a,b,1,†,*], Patrick E. Savage[c,†,*], Peter Ranacher[†,d], Damián E. Blasi[e,f,g,*],Thomas E. Currie[h], Takehiro Sato[i], Atsushi Tajima[i], Steven Brown[j], Mark Stoneking[k], Kentaro K. Shimizu[a,b], Hiroki Oota[l,*], and Balthasar Bickel[e].

[a] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland.
[b] Kihara Institute for Biological Research, Yokohama City University, 641-12 Maioka-cho, Totsuka-ku, Yokohama, Kanagawa, 244-0813, Japan.
[c] Faculty of Environment and Information Studies, Keio University, Shonan Fujisawa Campus, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan
[d] Department of Geography, University of Zurich, Winterthurerstr. 190 CH-8057 Zurich, Switzerland
[e] Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zürich, Switzerland.
[f] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, 07745 Jena, Germany.
[g] Human Relations Area Files, 755 Prospect Street, New Haven, USA
[h] Human Behaviour & Cultural Evolution Group, Centre for Ecology & Conservation, Department of Biosciences, University of Exeter, Penryn Campus, TR10 9FE, UK.
[i] Department of Bioinformatics and Genomics, Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, 13-1 Takara-machi, Kanazawa, Ishikawa, 920-8640 Japan
[j] Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton Ontario L8S 4K1, Canada.
[k] Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D04103 Leipzig, Germany.
[l] Kitasato University School of Medicine, Sagamihara, Kanagawa, 252-0374, Japan.
[1] Current Address: Tokai University School of Medicine. Shimokasuya 143, Isehara, Kanagawa, Japan. 259-1143
* Corresponding authors: matsumae.hiromi.g@tokai.ac.jp, psavage@sfc.keio.ac.jp, damianblasi@gmail.com, or hiroki_oota@med.kitasato-u.ac.jp
† Equal contribution

Abstract

Culture evolves in ways that are analogous to, but distinct from, genetic evolution. Previous studies have demonstrated correlations between genetic and cultural diversity at small scales within language families, but few studies have empirically investigated parallels between genetic and cultural evolution across multiple language families using a diverse range of cultural data. Here we report an analysis comparing cultural and genetic data from 13 populations from in and around Northeast Asia spanning 10 different language families/isolates. We construct distance matrices for language (grammar, phonology, lexicon), music (song structure, performance style), and genomes (genome-wide SNPs) and test for correlations among them. After controlling for spatial autocorrelation and recent contact, robust correlations emerge between genetic and grammatical distances. Our results suggest that grammatical structure might be one of the strongest cultural indicators of human population history, while also demonstrating differences among cultural and genetic relationships that highlight the complex nature of human cultural and genetic evolution.

Significance Statement

Comparing cultural traits to the genetic relationships of populations can reveal the extent to which cultural diversification reflects population history. To date, this approach has been mostly used to compare genetic relationships with the linguistic relationships that hold within language families, thereby limiting time depth to considerably less than 10,000 years. Here, we compare the genetic relationships of 13 populations in and around Northeast Asia to linguistic and musical relationships spanning different language families, thereby probing potential effects of population history at deeper time depths. We find that after controlling for geography, similarities in grammatical relationships reflect genetic relationships, suggesting that grammatical structure captures deep-time population 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, 4)), our cultural diversity is immense. For example, we speak or sign over 7,000 mutually unintelligible languages(5), and for each ethno-linguistic group there tend to be many different musical styles (6). 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(7–17). 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(18–21). 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(18, 22–24) .

The cultural evolution of language has proven particularly fruitful for understanding past population history(25–27). A classic approach involves identifying and analyzing sets of homologous words (cognates) 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 (25, 27, 28). However, lexical methods cannot usually be applied to multiple language families(26) 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(29–31). 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 tends to evolve too fast to preserve phylogenetic signals of language families(32, 33) and the history of lexica and structure might be partially independent as for example in the emergence of creole languages(21), the geographical distribution of language structure often points to contact-induced parallels in the evolution of entire sets of language families at deep time depths(34, 35).

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(36–41). 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(36, 37, 40, 42). Among indigenous Taiwanese populations speaking Austronesian languages, such analyses revealed significant correlations between music, mitochondrial DNA, and the lexicon (37), 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, Korean, Japanese)(34). 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(40).

We 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, and if so, whether music or language (lexicon(43), grammar(44, 45), or phonology(45–47)) would show the highest correlation with patterns of genetic diversity, after controlling for the influence of spatial autocorrelation and recent contact.

Results

We selected 13 populations from in and around Northeast Asia (encompassing 10 language families/isolates) for which all five sources of data (genome-wide single nucleotide polymorphisms (SNPs), grammars, phonology, music and geographic distance) were available (Fig. 1, Tables S1-S2, Materials and Methods)(40).

First, we investigated the similarity between populations in each of the dimensions of enquiry. For this purpose, we used phylogenetic split networks(48), which are capable of displaying multiple sources of similarity in a consistent manner (Fig. 2; and Supplementary Information Fig S10-14, Tables S3-S7). Distance analysis of lexical data resulted in a network topology with an overall star-shaped structure (Fig.2B) (where we exclude Nganasan for lack of data in our source database (43)). Exceptions are given by the two pairs of languages that are related to each other 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). This is consistent with the fact that lexical material is able to detect relationships within language families, but that it cannot resolve historical relations between families. Therefore, we excluded the lexical data from subsequent analyses.

Analyses of grammatical, phonological, genetic and musical distances reveal potentially more informative structure. Importantly, and in agreement with the assumption that language structure does not identify family relationships(29, 32), the clustering emerging from the distances do not generally coincide with languages families, except for the two Tungusic languages (Even and Evenki) in the domain of grammar (but even there the cluster shows more reticulation than in the lexical data). The clustering instead points to inter-family relations: Korean and Japanese cluster together in the networks based on grammar, SNPs, and music, but not phonology(49–51). Relationships between Ainu, Japanese, and Korean in SNP-based distances reflect prehistorical admixture in the Japanese Archipelago (50) but such relationships are not reflected in grammar-based distances. Buryat and Yakut are close together in SNPs, grammar and phonology, but not in music (40, 52). West Greenland Inuit shows no particular similarities to any of the other populations. 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(40).

Taken together, these results suggest that, in contrast to the usual assumption, similarities in grammatical and phonological structure are not primarily associated with simple vertical descent within linguistic 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 societies.

However, spatial proximities and contact between societies might lead to patterns of association that are relatively recent and shallow. Additionally, societies and languages placed far from the equator tend to display larger spatial ranges(53). In response to this, we implemented a partial redundancy analysis (pRDA) that controls for spatial proximity (Materials and Methods). We furthermore controlled large-scale spatial uncertainty by re-sampling locations from within the geographical ranges of each population, and we controlled for the three family relations (Tungusic, Chukotko-Chamkatkan, Uralic) by re-sampling one language at a time from these families.

When comparing the distributions of observed vs. permuted values of adjusted R^2 across the spatial ranges of the populations, we found that grammatical similarity reflects genetic similarity, i.e. our pRDA model successfully predicts genetic similarity by grammatical similarity after controlling for spatial proximity. Both the effect size and the significance of this association are robust across spatial and language-family re-sampling (Fig. 3). The results are furthermore robust against the amount of dimensionality-reduced variation that informs the models (Fig 3, Fig S15-31). The association of grammar with genetics contrasts with all other relationships, which are weak (differences include 0), non-robust (significant in only a few spatial locations) or both. Surprisingly, while grammatical similarity predicts genetic similarity across 82% of sampled locations, the reverse association is not robust against the spatial ranges of populations and holds for only 50% of the sampled locations (and even less when models are based on data components that explain higher or lower proportions of explained variance than in Fig. 3; see Supplementary Information, Figs. S27 and S29). Follow-up analyses show that the 5% of sampled locations with the lowest adjusted R^2 are not clustered in space (Fig. S31), suggesting that weak associations between grammar and genetics occur in random locations and not as systematic artefacts of spatial proximity.

Discussion

We have explored simultaneously the relations among genetic, linguistic, and musical data beyond the level of language families for the first time, thereby probing possible effects of population history at a time depth beyond the barrier of language families (ca. 10,000 years). We found that grammatical similarity predicts genetic similarity, while no other association is both strong and robust (Fig. 3).

One possible interpretation of this association is that grammar preserves population history exceptionally well over deep time, i.e. that the similarities evolved in parallel with population history and that they remained basically the same ever since the genetic population differences emerged several thousand years ago. This scenario predicts that in the region we analyzed, grammar was more conservative than the lexicon because the lexical distances are not associated with genetic distances (Fig. 2). While it is possible that local, possibly quite complex developments in pre-history might confirm this prediction, it contrasts with expectations from historical linguistics (32) and also from recent findings that suggest that grammar evolves faster than the lexicon in Austronesian(33) and shows rapid evolution also in Indo-European(54). A further prediction of the scenario is that, all else being equal, the association of grammar and genetics is symmetrical, i.e. one would expect not only that in grammar predicts genetics, but also that genetics predicts grammar. Yet, this is not what we find.

An alternative scenario that explains this asymmetry assumes that while the grammatical similarity is indeed rooted in deep population history, more recent historical factors may also have contributed to the preservation of this similarity over time. In other words, shared population history may be a necessary but not by itself a sufficient condition for leaving signals in grammatical similarity. A likely additional factor that explains the preservation of the grammatical similarity is language contact, which is well known to shape the distribution of grammatical features over large regions(34, 35, 55). The relevant contact events must have happened in the unknown past since our results are independent of the current spatial proximity between populations and also independent of the exact locations within the current language territories (as controlled by resampling). Furthermore, the contact events are possibly independent of gene flow, i.e. purely linguistic.

Interestingly, our results are qualitatively different from the only previous study to quantitatively compare genetic, linguistic and musical relationships (37). Among indigenous Austronesian-speaking populations in Taiwan, music was significantly correlated with genetics but not language, while we here find 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 may help to understand general relationships among language, music, and genetics.

In addition to revealing an association between genetic and grammatical patterns, our results also reveal complex dissociations in which these data reflect different histories. For example, the Ainu show particular genetic, and to some lesser extent also grammatical, similarity to the Japanese and Korean, while 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. Specifically, the grammatical patterns may reflect a shared history rooted in the prehistorical hunter-gatherer "Jomon" population that inhabited the Japanese archipelago over 15,000 years ago(56), while the 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(57), as we previously proposed in our "triple structure" model of Japanese archipelago history(40). While previous studies suggest specific genetic and cultural relationships between Korean and mainland Japanese populations(50, 58) or posit a shared origin (59–61), our findings support similarities in SNPs, music, and grammar, but not in phonology (Fig. 2, Supporting Information, Tables S3-S7).

While the evolutionary forces that influence genetic history are fairly well understood, determining to what extent the genetic relationships of particular populations reflect shared ancestry vs. prehistoric contact 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 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.

In conclusion, we have demonstrated a relationship between genome-wide SNPs and grammar across a variety of diverse Northeast Asian language families after controlling for spatial and language-family dependencies. Our results suggest that grammatical structure may track population history more closely than other cultural (including lexical) data, but also suggests 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 a 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

Selection of populations in this study

We selected 13 populations for whom 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(40). Linguistically, these 13 populations fall into 10 language families/isolates (63) (64). Korean, Ainu and Yukaghir are language isolates. Buryat, Japanese, Yakut, West Greenland Inuit are the sole representatives in our sample of the Mongolic, Japonic, Turkic, 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 samples and analysis

All music data and metadata are detailed in our previous report of circumpolar music (40). For this analysis, we used a subset of 13 of the original 35 populations who had matching genetic and linguistic data; these 13 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)). A matrix of pairwise distances among all 264 songs was calculated using normalized Hamming distances(65), and this distance matrix among songs was used to compute a distance matrix of pairwise musical ϕ_{st} values among the 13 populations using Arlequin(66) and the lingoes function of the ade4 package in R(67). Further details concerning the calculations can be found elsewhere(42, 65).

Genetic samples and analysis

Publicly-available genome-wide SNP array data for 13 populations (Table S1)(50, 68–71) 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(72) for a dataset of Ainu against JPT and CHB in

1000 Genomes Phase 1 released version 3 downloaded from BEAGLE website: http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/. To obtain shared SNPs among different platforms, genotype datasets were merged into a single dataset in PLINK file format by PLINK 1.9(73). The final merged genotype dataset included 273 individuals and 68,658 SNPs (total genotyping rate was 0.999). The merged dataset in PLINK format was converted to Genepop format using PGDSpider(74) and pairwise F_{st} values between populations were calculated with Genepop version 4.2(75).

Language analysis

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(43). 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 (76). This analysis did not attempt to identify and remove loanwords.

Grammar and phonology data: We combined data on grammatical and phonological traits from AUTOTYP (45), WALS(44), the ANU Phonotactics database(46), and PHOIBLE(47) 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).In view of the fact that the data are partly numerical and partly categorical, we use a balanced mix of PCA and MCA (77). Empty values are imputed using the R package missMDA (78).

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 ϕ_{st} for SNPs and pairwise ϕ_{st} for music (Supporting Information Section 3)(79). 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) (80).

Split network graphs: Distances are visualized using the SplitsTree neighbornet algorithm (version 4, (81)) 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 Ethnologue(63), 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. The random point samples were generated in PostGIS https://postgis.net/(Supporting Information Section 2.3). 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)(82). 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 autocorrelation.

Partial Redundancy Analysis: Partial redundancy analysis (pRDA) was carried out to explore the linear relationship between SNPs, grammar, phonology and music while controlling for spatial dependence (Supporting Information Section 5). pRDA 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(83). pRDA performs a regression of multiple response variables on multiple predictor variables (84). It yields an adjusted coefficient of determination (adjusted R^2), which captures the variation in the response that can be explained by the predictors (Supporting Information). We compare the observed adjusted R^2 values against a distribution under random permutations (Fig. 3, Fig. S15-26). To assess significance, we report the proportion of 1,000 locations for which the difference is significant at a 5% rejection level under random permutations. The pRDA and subsequent analyses were performed in R using the vegan package (64).

Acknowledgements

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, Kae Koganebuchi, Dagmar Jung, Marcelo Sánchez, and Chiara Barbieri for discussion. Funding supports for this work were provided by URPP Evolution in Action, University of Zurich to K.K.S. and B.B., a Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI Grant-in-aid for Scientific Research on Innovative Areas #4903(Evolinguistics), Number JP18H05080 to H.M, by a MEXT scholarship, a Startup Grant from the Keio Research Institute at SFC, an Individual Grant from the Keio Gijuku Academic Development Fund to P.E.S, and MEXT KAKENHI Number 16H06469 to K.K.S.

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.. 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.

References

1. Pagel M, Mace R (2004) The cultural wealth of nations. *Nature* 428(6980):275–278.
2. Novembre J, et al. (2008) Genes mirror geography within Europe. *Nature* 456(7218):98–101.
3. Brown DE (Donald E (1991) *Human universals* (Temple University Press).
4. Savage PE, Brown S, Sakai E, Currie TE (2015) Statistical universals reveal the structures and functions of human music. *Proc Natl Acad Sci* 112(29):8987–8992.
5. Hammarström H, Forkel R, Haspelmath M (2018) Glottolog 3.3. Available at: http://glottolog.org [Accessed September 1, 2018].
6. Nettl B *The study of ethnomusicology□: thirty-three discussions*.
7. Darwin C (1871) *The descent of man, and selection in relation to sex. Volume 1*.
8. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci U S A* 85(16):6002–6.
9. Bortolini E, et al. (2017) Inferring patterns of folktale diffusion using genomic data. *Proc Natl Acad Sci* 114(34):9140–9145.
10. Sokal RR (1988) Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci* 85(5).
11. Cavalli-Sforza LL (Luigi L, Menozzi P, Piazza A (1994) *The history and geography of human genes* (Princeton University Press).
12. Diamond, J. Bellwood P (2003) Farmers and Their Languages: The First Expansions. *Science (80-)* 300(5619):597–603.
13. Lansing JS, et al. (2007) Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci U S A* 104(41):16022.
14. Hunley K, et al. (2008) Genetic and Linguistic Coevolution in Northern Island Melanesia. *PLoS Genet* 4(10):e1000239.
15. Tishkoff SA, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–44.
16. Creanza N, et al. (2015) A comparison of worldwide phonemic and genetic variation in human populations. *Proc Natl Acad Sci U S A* 112(5):1265–72.
17. Longobardi G, et al. (2015) Across language families: Genome diversity mirrors linguistic variation within Europe. *Am J Phys Anthropol* 157(4):630–640.
18. Richerson PJ, Boyd R (2005) *Not by genes alone□: how culture transformed human evolution* (University of Chicago Press).
19. MacCallum RM, Mauch M, Burt A, Leroi AM (2012) Evolution of music by public choice. *Proc Natl Acad Sci* 109(30):12081–12086.
20. Blasi DE, Wichmann S, Hammarström H, Stadler PF, Christiansen MH (2016) Sound–meaning association biases evidenced across thousands of languages. *Proc Natl Acad Sci* 113(39):10818–10823.
21. Blasi DE, Michaelis SM, Haspelmath M (2017) Grammars are robustly transmitted even during the emergence of creole languages. *Nat Hum Behav* 1(10):723–729.
22. Reich D (Of HMS (2018) *Who we are and how we got here□: ancient DNA and the new science of the human past* (Pantheon).
23. Henrich JP (2016) *The secret of our success□: how culture is driving human evolution, domesticating our species, and making us smarter* (Princeton University Press).
24. Levinson SC, Gray RD (2012) Tools from evolutionary biology shed new light on the diversification of languages. *Trends Cogn Sci* 16(3):167–173.
25. Chang W, Cathcart C, Hall D, Garrett A (2015) Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language (Baltim)* 91(1):194–244.
26. Kozintsev A (2018) 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(No. 1 & 2):1–264.
27. Gray RD, Drummond AJ, Greenhill SJ (2009) Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement R. *Science (80-)* 323:1–5.
28. Xu S, et al. (2012) Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proc Natl Acad Sci* 109(12):4574–4579.

29. Ringe Jr DA (1995) `Nostratic' and the Factor of Chance. *Diachronica* 12:55–74.
30. Gray R (2005) Pushing the Time Barrier in the Quest for Language Roots. *Science (80-)* 309(5743):2007–2008.
31. Pagel M, Atkinson QD, S Calude A, Meade A (2013) Ultraconserved words point to deep language ancestry across Eurasia. *Proc Natl Acad Sci U S A* 110(21):8471–6.
32. Nichols J (1996) The Comparative Method as heuristic. *The Comparative Method Reviewed*, eds Durie M, Ross M (Oxford University Press, Oxford), pp 39–71.
33. Greenhill SJ, et al. (2017) Evolutionary dynamics of language systems. *Proc Natl Acad Sci U S A* 114(42):E8822–E8829.
34. Nichols J (1999) *Linguistic diversity in space and time* (University of Chicago Press).
35. Bickel B, Nichols J (2006) Oceania, the Pacific Rim, and the theory of linguistic areas. *Proc Annu Meet Berkeley Linguist Soc* 32S:3–15.
36. Lomax A, American Association for the Advancement of Science. (1978) *Folk song style and culture* (Transaction Books).
37. Brown S, et al. (2014) Correlations in the population structure of music, genes and language. *Proc R Soc B-Biological Sci* 281:20132072.
38. Callaway E (2007) Music is in our genes. *Nat News*. doi:doi:10.1038/news.2007.359.
39. Pamjav H, Juhász Z, Zalán A (2012) A comparative phylogenetic study of genetics and folk music. *Mol Genet …* 287(4):337–49.
40. Savage PE, et al. (2015) How 'circumpolar' is Ainu music? Musical and genetic perspectives on the history of the Japanese archipelago. *Ethnomusicol Forum*. doi:10.1080/17411912.2015.1084236.
41. Grauer V (2011) Sounding the depths: Tradition and the voices of history. CreateSpace. Available at: http://soundingthedepths.blogspot.com/.
42. Savage P, Merritt E, Rzeszutek T, Brown S (2012) CantoCore: a new cross-cultural song classification scheme. *Anal Approaches to World Music*.
43. Wichmann S, Eric W. Holman, Cecil H. Brown (2014) The ASJP Database (version 16). Available at: http://asjp.clld.org/.
44. Dryer MS, Haspelmath M (2013) The World Atlas of Language Structures Online. Available at: http://wals.info/.
45. Bickel B, et al. (2017) The AUTOTYP typological databases. Version 0.1.0. Available at: https://github.com/autotyp/autotyp-data/tree/0.1.0.
46. Donohue M, Hetherington R, McElvenny J, Dawson V (2013) World Phonotactis Database.
47. Moran S, McCloy D, Wright R (2014) PHOIBLE Online.
48. Huson DH, Bryant D (2006) Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol* 23(2):254–267.
49. Hanihara K (1991) Dual structure model for the population history of the Japanese. *Nichibunken Japan Rev* 2:1–33.
50. Jinam T, et al. (2012) 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(12):787–95.
51. Nakagome S, et al. (2015) 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(6):1533–1543.
52. Pugach I, et al. (2016) The Complex Admixture History and Recent Southern Origins of Siberian Populations. *Mol Biol Evol* 33(7):1777–1795.
53. Gavin MC, Stepp JR (2014) Rapoport's Rule Revisited: Geographical Distributions of Human Languages. *PLoS One* 9(9):e107623.
54. Widmer M, Auderset S, Widmer P, Nichols J, Bickel B (2017) {NP} recursion over time: evidence from {Indo-European}. *Language (Baltim)* 93:1–36.
55. Bickel B (2019) Large and ancient linguistic areas. *Language Dispersal, Diversification, and Contact: A Global Perspective*, eds Crevels M, Hombert J-M, Muysken P (Oxford University Press, Oxford).
56. McColl H, et al. (2018) The prehistoric peopling of Southeast Asia. *Science (80-)* 361(6397):88–92.
57. Sato T, et al. (2007) Origins and genetic features of the Okhotsk people, revealed by ancient mitochondrial DNA analysis. *J Hum Genet* 52(7):618–627.

58. HUGO Pan-Asian SNP Consortium MA, et al. (2009) Mapping human genetic diversity in Asia. *Science* 326(5959):1541–5.
59. Tranter N (2012) Introduction: Typology and Area in Japan and Korea. *The Languages of Japan and Korea*, ed Tranter N (Routledge), pp 3–23.
60. Whitman J (2012) The Relationship Between Japanese and Korea. *The Languages of Japan and Korea*, ed Tranter N (Routledge), pp 24–37.
61. Robbeets M (2015) *Diachrony of verb morphology: Japanese and the Transeurasian languages* (DE GRUYTER MOUTON).
62. Mesoudi A (2011) *Cultural evolution□: how Darwinian theory can explain human culture and synthesize the social sciences* (University of Chicago Press).
63. Dallas T (1999) “Ethnologue□: Languages of the World.” Available at: https://www.ethnologue.com/.
64. Harald Hammarström, Sebastian Nordhoff (2011) LangDoc: Bibliographic Infrastructure for Linguistic Typology. *Oslo Stud Lang* 3(2):31–43.
65. Rzeszutek T, Savage PE, Brown S (2012) The structure of cross-cultural musical diversity. *Proc Biol Sci* 279(1733):1606–12.
66. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10(3):564–7.
67. Bougeard S, Dray S (2018) Supervised Multiblock Analysis in R with the ade4 Package. *J Stat Softw* 86(1):1–17.
68. Rasmussen M, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463(7282):757–62.
69. Lazaridis I, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513(7518):409–413.
70. Fedorova SA, et al. (2013) Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. *BMC Evol Biol* 13:127.
71. Abecasis GR, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
72. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84(2):210–23.
73. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–75.
74. Lischer HEL, Excoffier L (2011) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28(2):298–299.
75. Rousset F (2008) genepop’007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour* 8(1):103–6.
76. Bakker D, et al. (2009) Adding typology to lexicostatistics: A combined approach to language classification. *Linguist Typology* 13(1):169–181.
77. Lê S, Josse J, Husson F (2008) FactoMineR: An Package for Multivariate Analysis. *J Stat Softw* 25(1):1–18.
78. Josse J, Husson F (2016) missMDA□: A Package for Handling Missing Values in Multivariate Data Analysis. *J Stat Softw* 70(1):1–31.
79. Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecol Modell* 196(3–4):483–493.
80. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York).
81. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23(2):254–267.
82. Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol Modell* 153(1–2):51–68.
83. Legendre P, Fortin M-J, Borcard D (2015) Should the Mantel test be used in spatial analysis? *Methods Ecol Evol* 6(11):1239–1247.
84. Van Den Wollenberg, Arnold L (1977) Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika* 42(2):207–219.

Figure Legends

Figure 1. Geographic areas of 13 languages/populations. Because some of the areas overlap in space, they are plotted in two separate maps.

Figure 2. Neighbornet networks of 13 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; Koryak and Chukchi to Chukotko-Kamchatkan.

Figure 3. Densities of the difference between observed and permuted adjusted R^2 values in the partial RDA. Each input component contributes at least 15% to the explained variance in the model. Numbers between brackets and grey shading correspond to the proportion of spatial locations (SL) for which the difference between observed and permuted adjusted R^2 is larger than 0 with $p \leq .05$. Results are proportionally similar when analyses are based on input components that contribute more (20%, Fig. S29) and less (10%, Fig. S27) to the explained variance and suggest that grammatical similarity predicts genetic similarity both strongly (large difference to a permutation baseline) and robustly (significantly in most re-sampled locations). No other association is both strong and robust.

Fig. 1

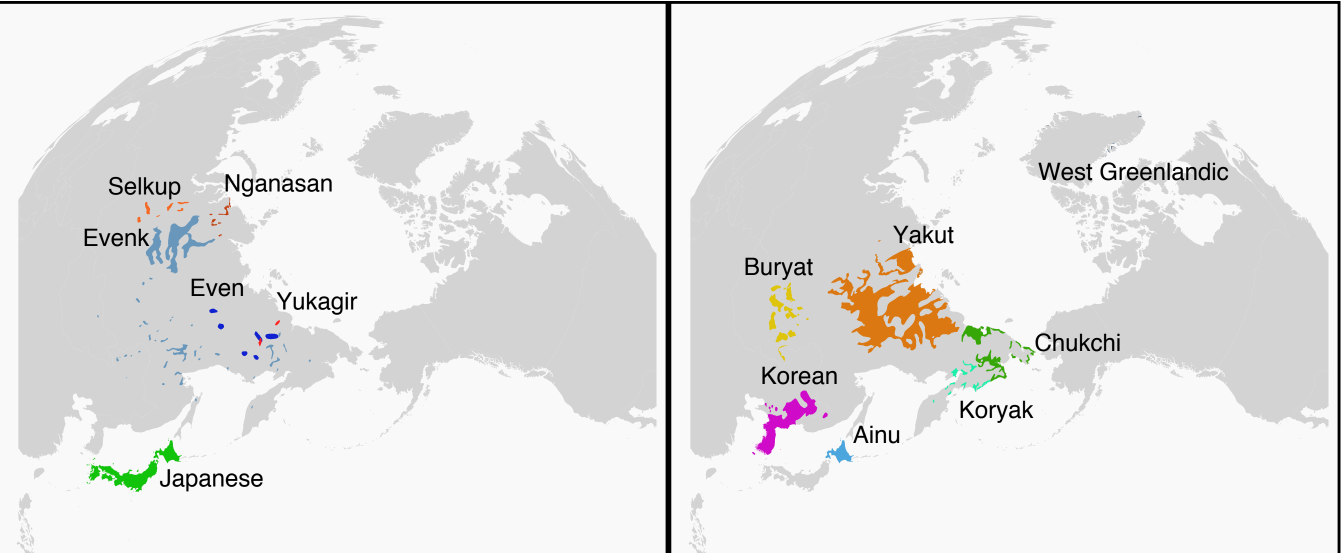
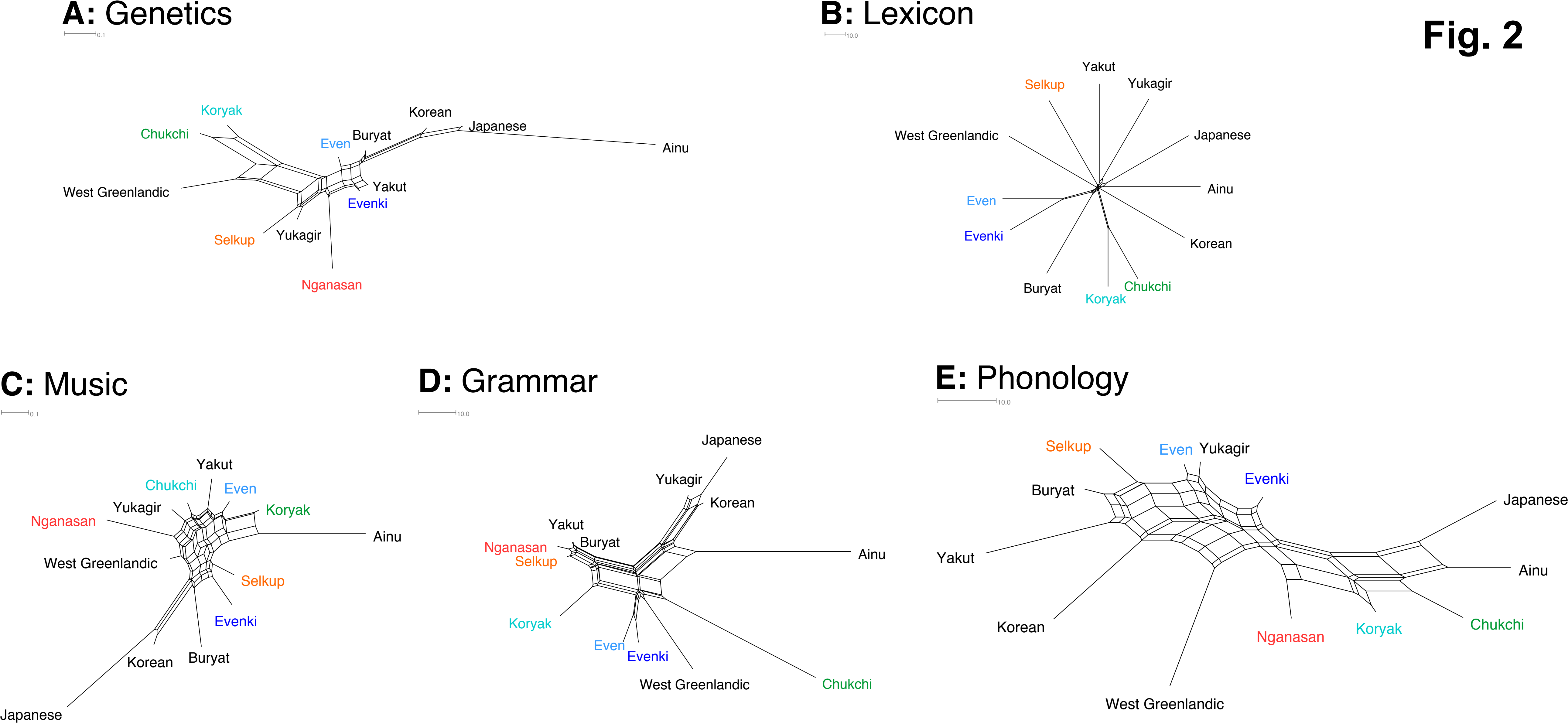

Selkup
Nganasan
Evenk
Even
Yukagir
Japanese
West Greenlandic
Yakut
Buryat
Chukchi
Korean
Koryak
Ainu

Fig. 2


A: Genetics
0.1
Koryak
Chukchi
Korean
Japanese
Even
Buryat
Ainu
West Greenlandic
Yakut
Evenki
Selkup
Yukagir
Nganasan
B: Lexicon
10.0
Yakut
Selkup
Yukagir
West Greenlandic
Japanese
Ainu
Even
Evenki
Korean
Buryat
Chukchi
Koryak
C: Music
0.1
Yakut
Chukchi
Even
Yukagir
Koryak
Nganasan
Ainu
West Greenlandic
Selkup
Evenki
Korean
Buryat
Japanese
D: Grammar
10.0
Japanese
Yukagir
Korean
Yakut
Nganasan
Buryat
Selkup
Ainu
Koryak
Even
Evenki
West Greenlandic
Chukchi
E: Phonology
10.0
Selkup
Even
Yukagir
Evenki
Buryat
Japanese
Yakut
Ainu
Korean
Nganasan
Koryak
Chukchi
West Greenlandic

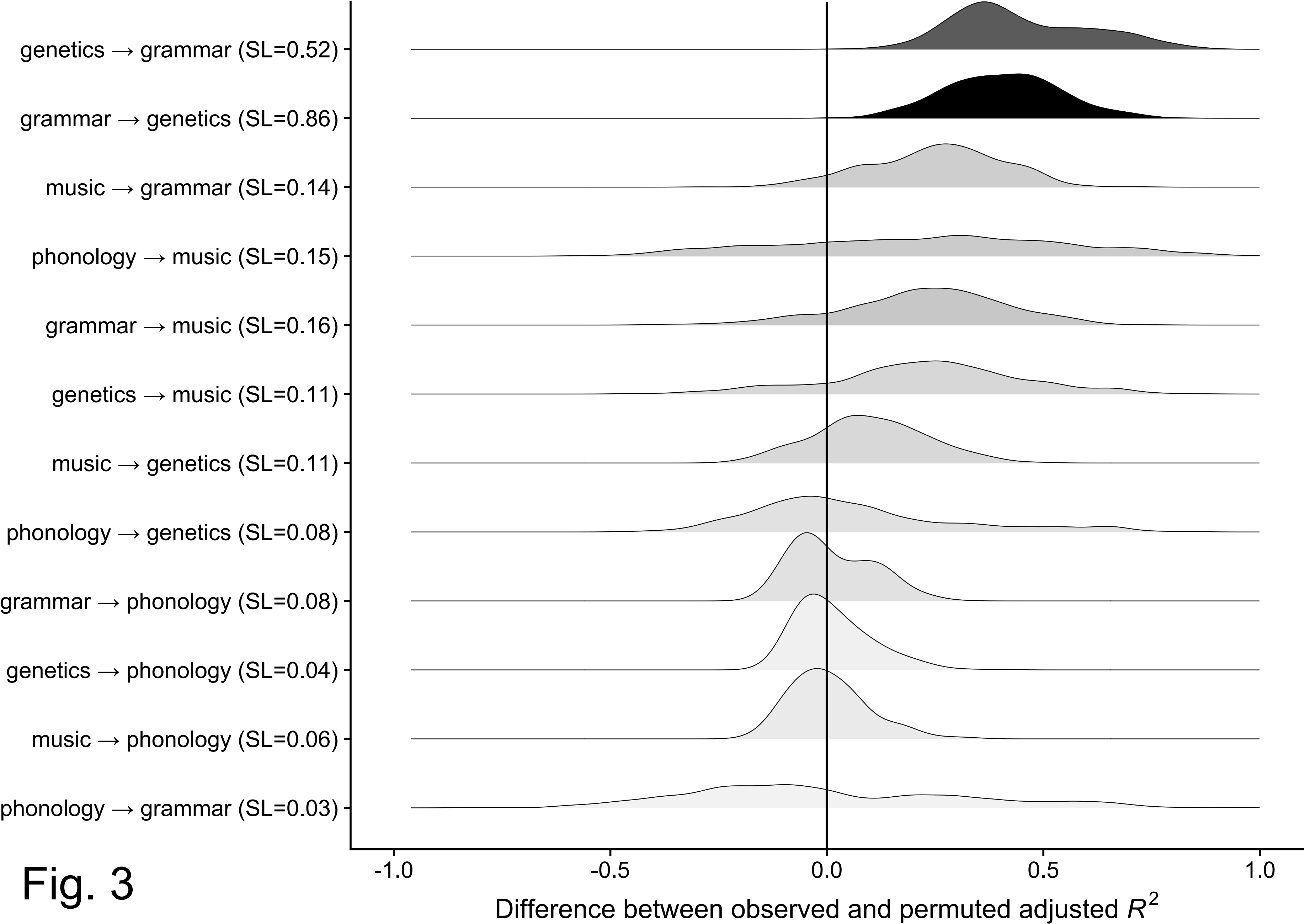

genetics → grammar (SL=0.52)
grammar → genetics (SL=0.86)
music → grammar (SL=0.14)
phonology → music (SL=0.15)
grammar → music (SL=0.16)
genetics → music (SL=0.11)
music → genetics (SL=0.11)
phonology → genetics (SL=0.08)
grammar → phonology (SL=0.08)
genetics → phonology (SL=0.04)
music → phonology (SL=0.06)
phonology → grammar (SL=0.03)
-1.0
-0.5
0.0
0.5
1.0
Difference between observed and permuted adjusted R^2

Fig. 3