Inferring patterns of folktale diffusion using genomic data

Eugenio Bortolini[a,b,c,1,2], Luca Pagani[d,e,1], Enrico R. Crema[f], Stefania Sarno[c], Chiara Barbieri[g], Alessio Boattini[c], Marco Sazzini[c], Sara Graça da Silva[h], Gessica Martini[i], Mait Metspalu[d], Davide Pettener[c], Donata Luiselli[c], and Jamshid J. Tehrani[i,2]

[a]Complexity and Socio-Ecological Dynamics Research Group, Department of Archaeology and Anthropology, Institución Milá y Fontanals, Spanish National Research Council (CSIC), 08001 Barcelona, Spain; [b]Department of Humanities, Universitat Pompeu Fabra, 08005 Barcelona, Spain; [c]Laboratory of Molecular Anthropology, Department of Biological, Geological, and Environmental Sciences, University of Bologna, 40126 Bologna, Italy; [d]Estonian Biocentre, 51010 Tartu, Estonia; [e]Department of Biology, University of Padova, 35131 Padua, Italy; [f]Department of Archaeology and Anthropology, University of Cambridge, CB2 3DZ Cambridge, United Kingdom; [g]Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, 07745 Jena, Germany; [h]Institute for the Study of Literature and Tradition, Faculty of Social Sciences and Humanities, New University of Lisbon, 1069-061 Lisbon, Portugal; and [i]Centre for the Coevolution of Biology and Culture, Department of Anthropology, Durham University, DH1 3LE Durham, United Kingdom

Edited by Marc W. Feldman, Stanford University, Stanford, CA, and approved June 13, 2017 (received for review September 1, 2016)

Observable patterns of cultural variation are consistently intertwined with demic movements, cultural diffusion, and adaptation to different ecological contexts [Cavalli-Sforza and Feldman (1981) *Cultural Transmission and Evolution: A Quantitative Approach*; Boyd and Richerson (1985) *Culture and the Evolutionary Process*]. The quantitative study of gene–culture coevolution has focused in particular on the mechanisms responsible for change in frequency and attributes of cultural traits, the spread of cultural information through demic and cultural diffusion, and detecting relationships between genetic and cultural lineages. Here, we make use of worldwide whole-genome sequences [Pagani et al. (2016) *Nature* 538:238–242] to assess the impact of processes involving population movement and replacement on cultural diversity, focusing on the variability observed in folktale traditions ($n = 596$) [Uther (2004) *The Types of International Folktales: A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson*] in Eurasia. We find that a model of cultural diffusion predicted by isolation-by-distance alone is not sufficient to explain the observed patterns, especially at small spatial scales (up to ∼4,000 km). We also provide an empirical approach to infer presence and impact of ethnolinguistic barriers preventing the unbiased transmission of both genetic and cultural information. After correcting for the effect of ethnolinguistic boundaries, we show that, of the alternative models that we propose, the one entailing cultural diffusion biased by linguistic differences is the most plausible. Additionally, we identify 15 tales that are more likely to be predominantly transmitted through population movement and replacement and locate putative focal areas for a set of tales that are spread worldwide.

cultural diffusion | demic diffusion | whole-genome sequences | folktales | Eurasia

Advances in DNA sequencing have opened new ways for exploring the demographic histories of human populations and the relationship between patterns of genetic and cultural diversity around the world. Newly available genome-wide evidence enables us to go beyond the use of linguistic relationship as a measure of common ancestry (1–3) and offers unprecedented support for studying the mechanisms underlying the transmission of cultural information over space and time (4–11) as well as the coevolution of genetic and cultural traits (12–18) across populations.

A key question for research in this area concerns the extent to which patterns of cultural diversity documented in the archaeological and ethnographic records have been generated by demic processes (i.e., the movement of people carrying their own cultural traditions with them) or cultural diffusion (i.e., the transfer of information without or with limited population movement/replacement) (6, 19, 20). Before tackling this question, however, it is critical to note that demic processes and cultural diffusion are not mutually exclusive conditions but rather, are opposite extremes of a continuous gradient, with intermediate and composite positions that more accurately represent empirical reality.

A broadly adopted null model of cultural diffusion draws on the expectation that selectively neutral variants would form geographic clines produced over time by isolation-by-distance (IBD) processes (21). Under an IBD model, individuals or groups that are spatially closer to each other are expected to be more similar than individuals or groups that are located farther apart. A positive correlation between cultural dissimilarity and geographic distance between samples is, therefore, used to infer processes of cultural transmission of nonadaptive information without population replacement (8, 17). However, observed genetic distance is the composite result of serial founder events, long-term IBD, and subsequent migratory events, which imply recent movement and resettling of people (22). A higher correlation between genetic distance and cultural dissimilarity than between culture and geography has, therefore, been proposed as a way to

Significance

This paper presents unprecedented evidence on the transmission mechanism underlying the spread of a broad cross-cultural assemblage of folktales in Eurasia and Africa. State-of-the-art genomic evidence is used to directly assess the relevance of demic diffusion processes, in particular on the distribution of Old World folktales at intermediate geographic scales, and identify individual stories that are more likely to be transmitted through population movement and replacement. The results provide an empirical solution to operate with linguistic barriers and highlight the impossibility of disentangling genetic from geographic relationships at a cross-continental scale, warning against the direct use of extant genetic variability to infer processes of long-range cultural transmission.

Author contributions: E.B., L.P., and J.J.T. designed research; E.B., L.P., A.B., M.S., S.G.d.S., G.M., M.M., D.P., and D.L. performed research; E.B., L.P., S.S., and J.J.T. analyzed data; E.R.C., S.S., C.B., A.B., M.S., G.M., M.M., D.P., D.L., and J.J.T. contributed to the interpretation of results; E.B., L.P., and S.G.d.S. performed data collection; and E.B., L.P., E.R.C., C.B., and J.J.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Genetic data used to run SpaceMix are available at www.ebc.ee/free_data. R scripts and related commands used to generate all of the results described in the paper are available at doi.org/10.5281/zenodo.821360.

[1]E.B. and L.P. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: eugenio.bortolini2@unibo.it or jamie.tehrani@durham.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1614395114/-/DCSupplemental.

single out the relative effect of demic processes on the distribution of cultural variants (8).

In a recent study, Creanza et al. (17) investigated the process responsible for the observed global distribution of (phonetic) linguistic variability by comparing it with genetic and geographic distances. The authors found high correlation between genetic and geographic distances at a worldwide scale, whereas linguistic distances were spatially autocorrelated only within a range of ∼10,000 km. The lack of residual correlation between genetic and linguistic distances up to this spatial scale did not allow the authors to reject their null model and was interpreted as a signal of cultural diffusion being the main driver of the distribution of phonetic variants in human populations.

The use of genetic variability as a plausible proxy to reject cultural diffusion as the sole responsible for the distribution of cultural traits depends on being able to disentangle genetic signals from geography. The high correlation between genetic and geographic distances at a global scale (22) lowers the inferential power of this model. However, this relationship is not constant across different geographic scales. We noted that the correlation obtained between pairwise genetic distances is stronger when measured across all possible population pairs at larger geographic scales, whereas it is considerably lower at smaller geographic distances (below ∼6,000 km for this dataset), possibly because of more recent and short-range population movements (Fig. 1*A*, yellow line). It is worth remembering that global trends have been forming over the past ∼40,000 y, whereas most cultural traditions are likely to have evolved more recently. This claim is supported by previous studies (17) and suggests that the effect of population movements independent from IBD can be identified only within limited geographic scales. At this spatial resolution, events shaping the distributions of genetic and cultural divergence are more likely to occur at the same temporal scale and hence, be more probably causally related.

An additional confounder is the potential effect of linguistic barriers, which might cause departures from a pure IBD model by constraining the exchange of genetic and/or cultural information between demes belonging to different ethnolinguistic groups. Given the relevance that spoken language has on the transmission of folktales and the light but measurable impact that they have for variants of individual tales in Europe (23), ethnolinguistic barriers should also be considered as key components of plausible alternative models to IBD.

Diffusion of Folktales: Investigating Mechanisms of Cultural Transmission in the Genomic Era

Here, we capitalize on the short-range decoupling of genetic and geographic distance to further infer mechanisms of genetic and cultural coevolution by using newly available genomic evidence (24) as an unbiased proxy of population relatedness. To do so, we analyzed the observed distribution of a set of individual folktales in Eurasia, looking for deviations from the null model of cultural diffusion predicted by geographic distance alone. Folktales are a ubiquitous and rigorously typed form of human cultural expression and hence, particularly well-suited for investigating cultural processes at wider cross-continental scale. Researchers since the Brothers Grimm (25) have long theorized about possible links between the spread of traditional narratives and population dispersals and structure but found mixed levels of support for this hypothesis when using indirect evidence for demic processes, such as linguistic relationships among cultures. One recent study suggested that, within the same linguistic family (Indo-European), the distributions of a substantial number of fairy tales were more consistent with linguistic relationships than with their geographical proximity, suggesting that they were inherited from common ancestral populations (3). This finding is confirmed by the relevance that ethnolinguistic boundaries may have for the transmission of variants of individual folktales in

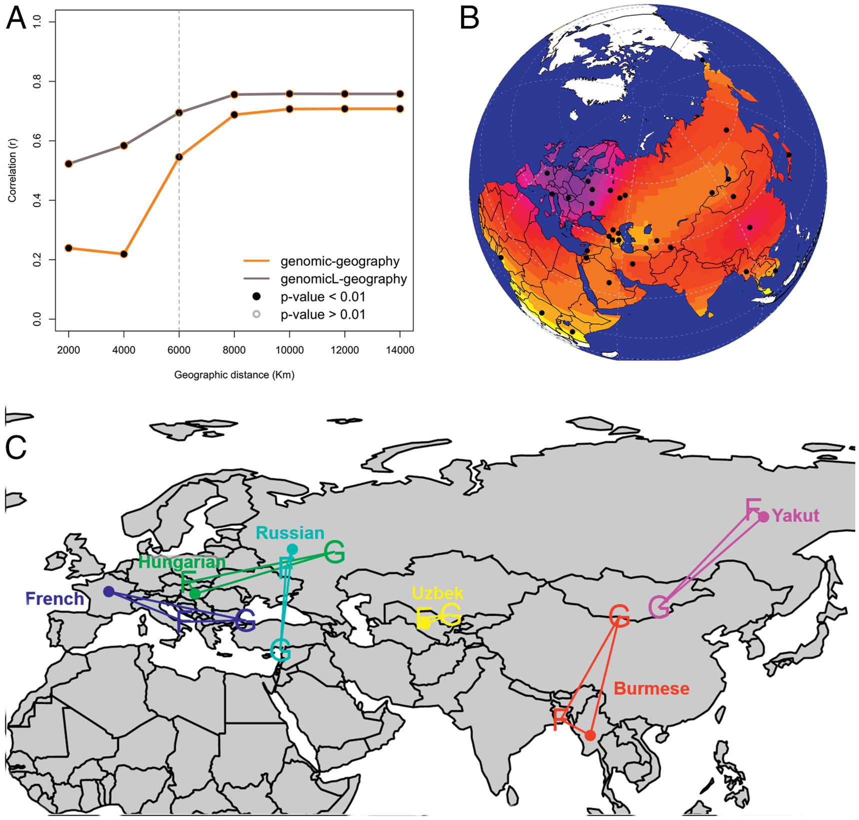


Fig. 1. (*A*) Plot of product–moment correlation values between pairwise genetic distance (both whole genome and biased for linguistic barriers) and pairwise geographic distance over cumulative geographic distance. (*B*) Map showing the spatial distribution of 33 populations in dataset MAIN. Surface colors represent interpolated richness values (i.e., the number of folktales exhibited by each population). Purple indicates higher values, whereas yellow indicates lower numbers. (*C*) Example of a map with SpaceMix results for genetic and folktale distance both projected on standard geographic coordinates. It is evident how, overall, folktale distribution (F) tends to cluster closer to geographic coordinates (dots), whereas the inferred source and direction of possible genetic admixture (G) are mismatched. For example, Burmese and Yakut exhibit quite segregated folktale assemblages, whereas their putative source of genetic admixture is closer in space. The case of Hungarian is emblematic for its folkloric assemblage rooted in Europe, whereas its putative genetic (and linguistic) source of admixture is located in the Ural region.

Europe. Ross et al. (23) have shown that, at population level, geographic distribution explains more variability than ethnolinguistic grouping. At this scale, when controlling for the effect of geography, linguistic boundaries do not show any residual significant relationship with folktale variant distribution, suggesting a possible temporal mismatch between folktale and linguistic traditions. However, when individual folktales are considered, ethnolinguistic identity is a significant predictor. This fact suggests that demes belonging to different ethnolinguistic affiliations may undergo higher costs for the transmission of individual folktales, even when they are closer in space. The simultaneous effect of shared linguistic ancestry and spatial proximity was also documented on the distributions of folktales recorded among Arctic hunter-gatherers (26).

Overview of This Study

In this study, we focus on 596 folktales comprising "animal tales" and "tales of magic" (27) typed as present (one) or absent (zero) in 33 populations (dataset MAIN), for which whole-genome sequences are available and exhibiting presence of at least five folktales (Fig. 1*B*, *SI Appendix*, and Dataset S1 Tables S1-2.1, S1-2.2, S1-2.3, and S1-2.4). Following previous examples (8), we test for deviations from a null model of pure cultural diffusion without population replacement (IBD), in which geographic distance alone is the best predictor of the decreasing number of shared folktales between pairs of populations. We measure and compare the fit of a number of alternative models comprising (*i*) a clinal model, in which populations belonging to different ethnolinguistic groups are less likely to share folktales as predicted by IBD (cultural diffusion with linguistic barriers); (*ii*) population movement and admixture between demes (demic process) as a substantial additional driver of folktale transmission; and (*iii*) a demic process constrained by linguistic barriers.

We test our hypothesis first by visualizing possible mismatches between actual geographic location of each population and the location inferred by applying explicit models accounting for genetic and cultural admixture (population movement with replacement) (28). We quantify the impact of linguistic barriers on both genetic and folktale variability using analysis of molecular variance (AMOVA) (29). We further investigate this by looking for the set of linguistic barrier parameters (intensity and geographic buffer) that maximizes the fit between genetic distance and geographic distance on the one hand and folktale distance and geographic distance on the other hand. We use this parameter combination to generate alternative models, with fitness that is formally assessed at both a global scale and over cumulative geographic distance. Following the assumptions of previous works (8), we develop a method to identify those folktales that—in the whole corpus—may be more likely to have been transmitted through population movement and replacement, supporting the idea that individual tales may have undergone different processes. To provide a starting point for this additional analysis on the diffusion of individual or smaller packages of tales, we infer potential focal areas—intended as a putative proxy for center of origin—of the most popular tales in the dataset.

Results

Effects of Ethnolinguistic Boundaries. We use AMOVA (29) to formally assess the impact of ethnolinguistic boundaries on both genetic and folktale variability, focusing only on Eurasian populations (dataset Eurasia; $n = 30$) to control for the effect of the Out of Africa expansion on genetic distance (*SI Appendix* and Dataset S1, Tables S1-3.1, S1-3.2, S1-3.3, and S1-3.4). We assign each population to an ethnolinguistic group (*Materials and Methods*, *SI Appendix*, and Dataset S1, Tables S1-4.1 and S1-4.2). Our analysis yielded $\Phi_{ST} = 0.036\,(P < 0.001)$ for genetic distance matrix, whereas $\Phi_{ST} = 0.1\,(P < 0.001)$ for distances based on folktale distributions. These results confirm the expected differential impact of intergroup boundaries between genetic and cultural variability and are consistent with previous results obtained for population structure on the transmission of cultural traits (23, 30).

We use this evidence to further investigate the separate effects of linguistic barriers on the flow of genetic and cultural information by focusing on two parameters (i.e., intensity and geographic buffer of the cultural barrier) (details are in *Materials and Methods*). We find that the parameter combinations that resulted in the highest correlation between genetic–geographic distances (intensity = 0.1; radius = 1,500 km) and between folktale–geographic distances (intensity = 0.3; radius = 3,000 km) imply that linguistic barriers have a differential impact of these two kinds of information, and we use this parameter setting to generate two corrected distance matrices for genetics (geneticL) (Dataset S1, Table S1-5.1) and folktales (folktaleL) (Dataset S1, Table S1-5.2), respectively. By using raw and corrected distance matrices, we define alternative models as (*i*) biased cultural diffusion (folktaleL ∼ geographic), (*ii*) demic diffusion (folktale ∼ genetic), and (*iii*) biased demic diffusion (folktaleL ∼ geneticL).

Assessing Models of Folktale Transmission. We set out to test for deviations from the null model of cultural diffusion caused by IBD. We explore the relationship between our genetic, folktale, and geographic distance matrices using SpaceMix (28) (*SI Appendix*). We note that, when transformed into pseudospatial coordinates, folktale distances tend to match actual geographic coordinates better than genetic distances (Fig. 1*C* and *SI Appendix*, Fig. S1-3.1). The role of geography and ethnolinguistic barriers is also confirmed by a NeighborNet (31) based on folktale distances, showing a broad spatial clustering and proximity/reticulation between demes belonging to the same ethnolinguistic group (*SI Appendix*).

We then assess the goodness of fit of all of the alternative models at a global scale by comparing Pearson's product–moment correlation (32), bias-corrected distance correlation (33), and partial distance correlation (34, 35) (Tables 1 and 2; details are in *Materials and Methods* and *SI Appendix*). It is evident how, after Bonferroni correction, all alternative models accounting for ethnolinguistic boundaries perform better than the models that do not consider them. With both product–moment correlation coefficient and bias-corrected distance correlation, the best model is the one representing cultural diffusion with linguistic barriers followed by demic processes constrained by linguistic barriers. With distance correlation, however, the difference between the two models is smaller than with standard correlation coefficient. When the dependence between variables is assessed controlling for a third variable through partial distance correlation, linguistic-biased cultural diffusion remains as good a predictor of folktale variability as IBD. This phenomenon could be due

Table 1. Variable association at a global level

Model	cor	*P*	bcdCor	*P*
Folktale ∼ genetic	0.20	<0.001	0.20	<0.001
Folktale ∼ geographic	0.19	<0.001	0.31	<0.001
Genetic ∼ geographic	0.71	<0.001	0.84	<0.001
FolktaleL ∼ geneticL	0.55	<0.001	0.55	<0.001
FolktaleL ∼ geographic	0.64	<0.001	0.57	<0.001
GeneticL ∼ geographic	0.76	<0.001	0.83	<0.001

Comparison between null model of cultural diffusion predicted by IBD (folktale ∼ geographic) and alternative models [i.e., demic diffusion (folktale ∼ genetic), cultural diffusion biased by linguistic barriers (folktaleL ∼ geographic), and demic diffusion biased by linguistic barriers (folktaleL ∼ geneticL)]. Values refer to Pearson's product–moment correlation (cor) and bias-corrected distance correlation (bcdCor) after Bonferroni correction.

Table 2. Partial distance correlation at a global scale

Model	pdCor	*P*
Folktale ∼ genetic, geographic	−0.11	1.00
Folktale ∼ geographic, genetic	0.26	<0.001
FolktaleL ∼ geneticL, geographic	0.17	<0.001
FolktaleL ∼ geographic, geneticL	0.25	<0.001

Results of partial distance correlation for null (folktale ∼ geographic, genetic) and alternative models [i.e., demic diffusion (folktale ∼ genetic, geographic), cultural diffusion biased by linguistic barriers (folktaleL ∼ geographic, geneticL), and demic diffusion biased by linguistic barriers (folktaleL ∼ geneticL, geographic)] after Bonferroni correction.

to the fact that, at a global scale, correlation between language-corrected genetic distance and geographic distance is higher (Fig. 1) and lowers the residual signal.

Significant deviations from the null model of cultural diffusion predicted by IBD are further investigated over cumulative geographic distance by comparing Pearson's correlation coefficients (Fig. 2 and *SI Appendix*, Table S1-7.1). Above 4,000 km, language-biased cultural diffusion presents with the highest fit at all bins followed by language-biased demic diffusion. Under 4,000 km, folktale distance exhibits stronger dependence from genetic distance than from geographic distance. This relationship is particularly visible under 2,000 km, where the effect of linguistic barriers is the same for genetic and cultural variability.

All results allow us to reject the null model of plain cultural diffusion predicted by IBD and suggest instead that, of all alternative models, the one involving cultural diffusion mitigated by linguistic barriers could be the most plausible one. In addition, as previously pointed out (Fig. 1), results consistently confirm that small geographic scale offers a more efficient disentanglement between possible uncoupled effects of genetic and geographic distances over cultural variables—even after correcting for potential ethnolinguistic barriers.

Uniform Body of Knowledge or Individual Units? Our results show that, when considering the folktales contained in our dataset as a uniform corpus, the null model dictated by IBD could be rejected. Previous results (23), however, have shown that individual tales or smaller groups of tales may be transmitted across populations as partially independent evolutionary units. If a given cultural trait is not transmitted through population movement and replacement, populations that share it should not exhibit significantly lower genetic distance than populations that do not exhibit it (8). To single out folktales that markedly contradict such null hypothesis, we compare the distribution of pairwise genetic distances corrected for ethnolinguistic boundaries among populations sharing a given tale against distances of the remaining pairs of populations using the Mann–Whitney–Wilcoxon test. We focus on 308 folktales that are present in at least five populations and run two separate tests, the first considering all pairs of populations (Dataset S1, Table S1-6.1) and a second considering only those within a conservative geographic range of 6,000 km (Fig. 1*A* and Dataset S1, Table S1-6.2). After Bonferroni correction, 15 of 308 analyzed folktales (4.9%) (Dataset S1, Tables S1-7.1 and S1-7.2) present with significantly lower than expected pairwise genetic distance, hence allowing us to reject our null hypothesis and suggesting that these tales may indeed have spread during events of demic diffusion biased by ethnolinguistic barriers.

Folktale Dispersal and Focal Areas. For a subset of the analyzed folktales, we identify focal areas, representing potential areas of origin and defined as locations that maximize the decay of a given folktale abundance over geographic distance measured with Pearson's correlation coefficient (Dataset S1, Table S1-8.1). Focal areas were generated for the 19 most widespread folktales, which follow four main trends (*SI Appendix*). Some of these tales possibly started to be diffused mostly via cultural transmission from Eastern Europe, with subsequent radial diffusion across Eurasia and Africa [such as Aarne Thompson Uther catalog 155 (ATU155): "The Ungrateful Snake Returned to Captivity" in *SI Appendix*, Fig. S1-8-I *1* or ATU313: "The Magic Flight" in Fig. 3], whereas others probably started their journey from Caucasus (*SI Appendix*, Fig. S1-8-I *6–8*). Examples of the latter are ATU400: "The Man on a Quest for His Lost Wife," ATU480: "The Kind and Unkind Girls," ATU531: "The Clever Horse," and ATU560: "The Magic Ring." Some narrative plots might have originated in northern Asia—such as the famous "Thumbling" (Tom Thumb) (*SI Appendix*, Fig. S1-8-I *18*)—whereas a last group could have spread from Africa (*SI Appendix*, Fig. S1-8-I *17*), such as in the case of ATU670: "The Man Who Understands Animal Language."

Discussion

Using Genetic Evidence to Infer Processes of Cultural Transmission. Our results resonate with broader questions in cultural evolutionary studies, particularly those concerning the mechanisms of cultural transmission over time and space. They show that the use of newly generated, whole-genome sequences offers a unique opportunity for an unbiased assessment of patterns of cultural variation in the ethnographic and archaeological records. Genetic variability has been already interpreted in the past as a direct proxy of the movement of human groups over time and space, and as such, it has been used as a potential marker of demic mechanisms (8, 17).

We show the effect of ethnolinguistic barriers on both genetic and cultural population structure. By introducing an empirical approach, we find that ethnolinguistic identity has a potentially independent and differential impact on genetic and cultural information. More specifically, our results suggest that linguistic

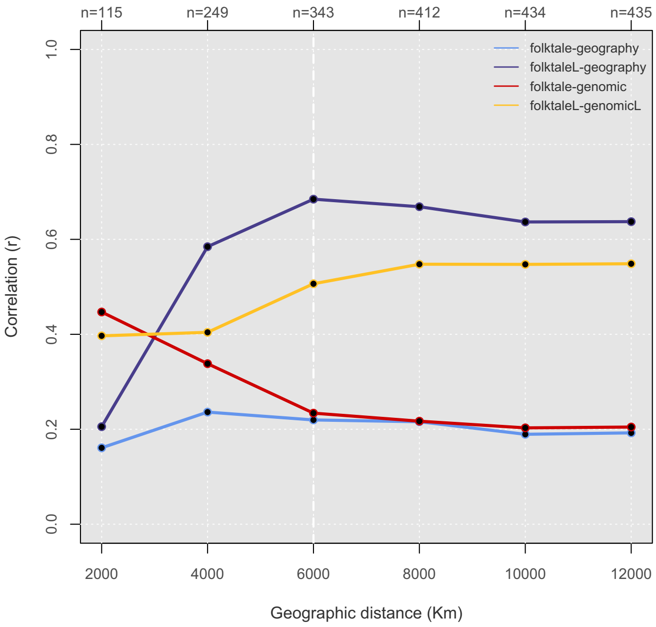


Fig. 2. Comparison of the null model of cultural diffusion dictated by IBD (folktale ∼ geographic; light blue) against all alternative models: demic diffusion (folktale ∼ genetic; red), language-biased cultural diffusion (folktaleL ∼ geographic; purple), and language-biased demic diffusion (folktaleL ∼ geneticL; yellow) over cumulative geographic distance. Product–moment correlation coefficients are calculated at each geographic bin (size = 2,000 km), with original distance matrices up to 12,000 km.

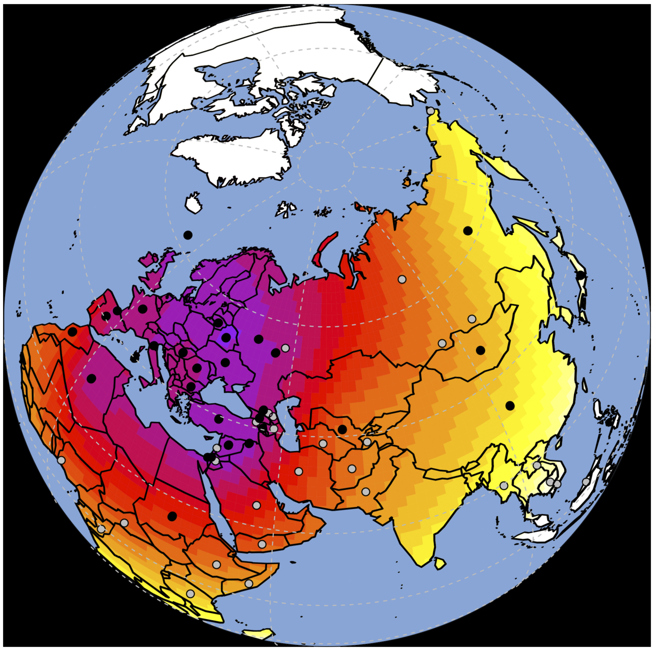


Fig. 3. Possible focal area and dispersion pattern for tale ATU313 "The Magic Flight," one the most popular folktales in this dataset, which may have been additionally spread through population movement and replacement. It is interesting to note how this tale reached locations that are far from its putative origin (such as Japan and southeastern Africa), whereas it was not retained by many populations located in between (gray dots).

barriers may be twice as effective on the diffusion of cultural traits than on population movement and that the decay over geographic distance of such effect is almost two times slower for culture than for genetic information. Nevertheless, this work very explicitly generates a cautionary tale concerning the use of genomic evidence for investigating such events at a cross-continental or global scale, where geographic clines in genetic variability are the result of different processes that can hardly be disentangled and that may present with considerable temporal mismatch with more recent cultural processes.

Cultural Evolutionary Mechanisms of Folktale Transmission. Folktales are a prime example of a universal form of cultural expression linked to various vectors of propagation over generations and across geographic and ethnolinguistic barriers that allows us to address questions of cultural evolutionary processes at a cross-cultural and -continental scale. Our results provide insights on the processes driving the spread of folkloric narratives that go beyond previous studies that were limited to a single language family (3).

By correcting for the presence of ethnolinguistic barriers, we find that the null model of cultural diffusion predicted by IBD alone cannot explain the observed distribution of folktales across Eurasia. Instead, beyond $\sim 4{,}000$ km, cultural diffusion biased by linguistic barriers exhibits the highest correlation at all geographic bins. At small geographic bins ($< 4{,}000$ km), population movements and linguistic barriers may be more relevant than geographic proximity, pointing once again at the possible importance of small-scale processes of cultural transmission for testing more specific hypotheses when using genetic evidence. In addition, processes other than simple cultural diffusion may be more relevant for a smaller group of tales shared by pairs of populations that are genetically closer than populations not exhibiting those tales. Looking for smaller packages of tales or individual tales and their variants can be useful to shed light on the formation process of this vast body of popular knowledge. The long-range patterns detected by our analyses may complement this picture by suggesting a more ancient origin of some of these folktales (*SI Appendix*) (36–39). On a broader level, these results can be used in the future to infer directional trends of cultural dispersal as well as to test for the emergence of systematic social biases [such as prestige bias, conformism/anticonformism, heterophily, and content-dependent biases (5, 23, 30)] or cultural barriers different from linguistic ones, which have a chronology that may be independently ascertained.

Materials and Methods

Dataset Description. Folktale data were sourced from the ATU (27). This dataset comprises animal tales (ATU1–299) and tales of magic (ATU300–749). Of 198 societies in which the tales were recorded, 73 matched available genetic data (Dataset S1, Table S1-1). Of these groups, 33 populations exhibiting at least five folktales were selected (Fig. 1*B* and Dataset S1, Table S1-2.2). Each population is described by a string listing the presence (one) or absence (zero) of any of the included 596 folktales.

Genetic, Folktale, and Geographic Distances. Genetic distances were estimated by the average pairwise distances between two genomes, one from each population, including both coding and noncoding regions to avoid ascertainment biases. Genetic distance for (i, j) pairs of populations represented by more than one genome was calculated as the average of all possible (i, j) pairs of genomes. As a consequence, the diagonal of the genetic distance matrix was not constrained to be zero (Dataset S1, Table S1-3.2). Folktale distance between population pairs was calculated as asymmetric Jaccard distance (40) (Dataset S1, Table S1-3.3). Geographic distance was calculated as pairwise great circle distance with a waypoint located in the Sinai Peninsula to constrain movement of African demes [through the package gdistance in R (41)]. Coordinates (longitude and latitude in decimal degrees) (Dataset S1, Tables S1-9.1 and S1-9.2) identify the assumed center of the area occupied by a given folkloric tradition as defined by the ATU index.

Transformation of Dissimilarities into Euclidean Distances. To perform bias-corrected and partial distance correlation, folktale, genetic, and geographic distances were transformed into their exact Euclidean representations (33, 42). The original folktale and genetic distance matrices were scaled through classic multidimensional scaling using the function cmdscale in R and following the procedure for exact representation (34). Euclidean distances were computed from the obtained number of descriptors (n – 2) using the function dist in R (Dataset S1, Tables S1-10.1 and S1-10.2). Euclidean representation of geographic distance (Dataset S1, Table S1-10.3) was instead obtained by reprojecting the original set of coordinates on a plane using two-point equidistant projection through the functions tpeqd in the package mapmisc (43) and spTransform in the package sp in R (44, 45). Euclidean distance between the new set of coordinates was computed using the function rdist in the package fields in R (46).

AMOVA. To implement AMOVA (29) in our analysis, each population was assigned to an ethnolinguistic group derived from Ethnologue (https://www.ethnologue.com; Dataset S1, Table S1-4.1), and we used the function amova in the package pegas (47) in R. Significance values are obtained through permutation (1,000 iterations).

Variable and Model Comparison. The relationship between original and biased folktale, genetic, and geographic pairwise distance matrices was quantitatively assessed at global scale and cumulative geographic scales. Measures were obtained through (*i*) Pearson's product–moment correlation coefficient using the function cor.test in R, (*ii*) bias-corrected distance correlation (33) using the function dcor.ttest in the package energy in R (48), and (*iii*) partial distance correlation using the function pdcor.test in the package energy in R. In parallel, SpaceMix (28) was used to compute folktale and genetic pseudocoordinates, which were compared with actual geographic coordinates to explore inferred processes of admixture.

Estimating the Effect of Ethnolinguistic Barriers on Genetic and Folktale Distance. We assumed that, if existent, a linguistic barrier would act on pairs of populations that belong to different linguistic families and live within a d geographic distance and artificially increase the actual genetic (*Dgen*) or folktale (*Dfolk*) distance by an intensity factor f. We also assumed that parameters d and f may be different when looking at genetic (d_G, f_G) and folktale (d_F, f_F) distances. We assessed the correlation between geographic and genetic or folktale distances at increasing spatial bins before and after correcting for putative linguistic barriers. Particularly, we chose as best pairs of (d_G, f_G) and (d_F, f_F) those that maximized the above-mentioned correlations. Notably, $f_G = 0$ or $f_F = 0$ (i.e., absence of linguistic barriers) had an equal chance of being picked up as the best values for our parameters. We instead reported (1,500, 0.1) and (3,000, 0.3) as best pairs of genetic and folktale parameters, respectively. To obtain unbiased genetic ($Dgen'$) and folktale ($Dfolk'$) distances, we, therefore, corrected for the effect of linguistic barriers, so that, for populations (i, j), $Dgen'_{ij} = Dgen_{ij} \times (1 - f_G)$ if $d_{ij} \leqslant d_G$ and $Dfolk'_{ij} = Dfolk_{ij}{}^{*}(1 - f_F)$ if $d_{ij} \leqslant d_F$.

Data Availability and Codes. R scripts and related commands used to generate all of the results described in the paper are available at doi.org/10.5281/zenodo.821360. Folktale and geographic data as well as genetic distances are also available in Dataset S1. Genetic data used to run SpaceMix are taken from ref. 24 (www.ebc.ee/free_data).

ACKNOWLEDGMENTS. We thank Adrian Timpson, Anne Kandler, Dugald Foster, Jeremy Kendal, Rachel Kendal, Simon Greenhill, and two anonymous reviewers for their comments and useful suggestions. E.B. is supported by SimulPast Consolider Ingenio Project CSD2010-00034 funded by the Spanish Ministry of Economy, Industry, and Competitiveness. L.P. is supported by the European Union through European Regional Development Fund Projects 2014-2020.4.01.16-0030 and 2014-2020.4.01.15-0012.

1. Currie TE, Greenhill SJ, Gray RD, Hasegawa T, Mace R (2010) Rise and fall of political complexity in island South-East Asia and the Pacific. *Nature* 467:801–804.
2. Mathew S, Perreault C (2015) Behavioural variation in 172 small-scale societies indicates that social learning is the main mode of human adaptation. *Proc Biol Sci* 282:20150061.
3. da Silva S, Tehrani J (2016) Comparative phylogenetic analyses uncover the ancient roots of Indo-European folktales. *R Soc Open Sci* 3:150645.
4. Cavalli-Sforza LL, Feldman MW (1981) *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton Univ Press, Princeton).
5. Boyd R, Richerson PJ (1985) *Culture and the Evolutionary Process* (Univ of Chicago Press, Chicago).
6. Collard M, Shennan SJ, Tehrani J (2006) Branching, blending and the evolution of cultural similarities and differences among human populations. *Evol Hum Behav* 27:169–184.
7. Ackland GJ, Signitzer M, Stratford K, Cohen MH (2007) Cultural hitchhiking on the wave of advance of beneficial technologies. *Proc Natl Acad Sci USA* 104:8714–8719.
8. Pinhasi R, von Cramon-Taubadel N (2009) Craniometric data supports demic diffusion model for the spread of agriculture into Europe. *PLoS One* 4:e6747.
9. Gray RD, Bryant D, Greenhill SJ (2010) On the shape and fabric of human history. *Philos Trans R Soc Lond B Biol Sci* 365:3923–3933.
10. Fort J (2012) Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. *Proc Natl Acad Sci USA* 109:18669–18673.
11. Lycett SJ (2015) Cultural evolutionary approaches to artifact variation over time and space: Basis, progress, and prospects. *J Archaeol Sci* 56:21–31.
12. Ammerman AJ, Cavalli-Sforza LL (1984) *The Neolithic Transition and the Genetics of Populations in Europe* (Princeton Univ Press, Princeton).
13. Renfrew C (1992) Archaeology, genetics and linnguistic diversity. *Man* 27:445–478.
14. Renfrew C (2001) From molecular genetics to archaeogenetics. *Proc Natl Acad Sci USA* 98:4830–4832.
15. Bell AV, Richerson PJ, McElreath R (2009) Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. *Proc Natl Acad Sci USA* 106:17671–17674.
16. Itan Y, Powell A, Beaumont MA, Burger J, Thomas MG (2009) The origins of lactase persistence in Europe. *PLoS Comput Biol* 5:e1000491.
17. Creanza N, et al. (2015) A comparison of worldwide phonemic and genetic variation in human populations. *Proc Natl Acad Sci USA* 112:1265–1272.
18. Haak W, et al. (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207–211.
19. Crema ER, Kerig T, Shennan S (2014) Culture, space, and metapopulation: A simulation-based study for evaluating signals of blending and branching. *J Archaeol Sci* 43:289–298.
20. Fort J (2015) Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe. *J R Soc Interface* 12:20150166.
21. Wright S (1943) Isolation by distance. *Genetics* 28:114–138.
22. Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 102:15942–15947.
23. Ross RM, Greenhill SJ, Atkinson QD (2013) Population structure and cultural geography of a folktale in Europe. *Proc Biol Sci* 280:20123065.
24. Pagani L, et al. (2016) Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* 538:238–242.
25. Grimm W (1884) Preface to children's and household tales. *The Complete Grimm's Fairy Tales* (George Bell, London).
26. Ross RM, Atkinson QD (2016) Folktale transmission in the arctic provides evidence for high bandwidth social learning among hunter-gatherer groups. *Evol Hum Behav* 37:47–53.
27. Uther HJ (2004) *The Types of International Folktales: A Classification and Bibliography. Based on the System of Antti Aarne and Stith Thompson* (Suomalainen Tiedeakatemia, Helsinki).
28. Bradburd GS, Ralph PL, Coop GM (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* 67:3258–3273.
29. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
30. Shennan S, Crema E, Kerig T (2015) Isolation-by-distance, homophily, and "core" vs. "package" cultural evolution models in Neolithic Europe. *Evol Hum Behav* 36:103–109.
31. Huson D, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254–267.
32. Pearson K (1895) Notes on regression and inheritance in the case of two parents. *Proc R Soc Lond* 58:240–242.
33. Székely G, Rizzo M (2013) The distance correlation t-test of independence in high dimension. *J Multivar Anal* 117:193–213.
34. Székely G, Rizzo M (2013) Partial distance correlation with methods for dissimilarities. arXiv:1310.2926v3.
35. Székely G, Rizzo ML (2016) *Partial Distance Correlation*, eds Cao R, González MW, Romo J, (Springer International Publ, Cham, Switzerland), pp 179–190.
36. Bottigheimer RB (2009) *Fairy Tales: A New History* (Excelsior Editions/State Univ of New York Press, Albany, NY), p 152.
37. Bottigheimer RB (2014) Palgrave historical studies in witchcraft and magic. *Magic Tales and Fairy Tale Magic: From Ancient Egypt to the Italian Renaissance* (Palgrave Macmillan, Basingstoke, UK), p 208.
38. Thompson S (1977) *The Folktale* (Univ of California Press, Oakland, CA).
39. Propp VI (1968) *Morphology of the Folktale*, Publications of the American Folklore Society Bibliographical and Special Series (Univ of Texas Press, Austin, TX), 2nd Ed, pp 26–158.
40. Jaccard P (1901) Etude comparative de la distribution florale dans une portion des alpes et del jura. *Bull del la Societe Vaudoise des Sci Nat* 37:547–579.
41. van Etten J (2014) gdistance: Distances and Routes on Geographical Grids (R Package), Version 1.1-5.
42. Székely G, Rizzo M, Bakirov N (2007) Measuring and testing dependence by correlation of distances. *Ann Stat* 35:2769–2794.
43. Brown P (2016) Mapmisc: Utilities for Producing Maps (R Package), Version 1.5.0. Available at https://CRAN.R-project.org/package=mapmisc. Accessed January 18, 2017.
44. Pebesma EJ, Bivand RS (2005) Classes and methods for spatial data in R. *R News* 5:9–13.
45. Bivand R, Pebesma E, Gómez-Rubio V (2013) *Applied Spatial Data Analysis with R* (Springer, New York), 2nd Ed.
46. Nychka D, Furrer R, Paige J, Sain S (2016) Fields: Tools for Spatial Data (R Package), Version 8.3-6. Available at https://CRAN.R-project.org/package=fields. Accessed January 18, 2017.
47. Paradis E (2010) pegas: An R package for population genetics with an integrated–modular approach. *Bioinformatics* 26:419–420.
48. Rizzo ML, Székely GJ (2016) Energy: E-Statistics: Multivariate Inference via the Energy of Data (R Package), Version 1.7-0. Available at https://CRAN.R-project.org/package=energy. Accessed January 9, 2017.

Inferring patterns of folktale diffusion using genomic data

SI Appendix

Eugenio Bortolini, Luca Pagani, Enrico R. Crema, Stefania Sarno, Chiara Barbieri, Alessio Boattini, Marco Sazzini, Sara Graça da Silva, Gessica Martini, Mait Metspalu, Davide Pettener’ Donata Luiselli, Jamshid J. Tehrani

Contents

1 Extended dataset description

Folktale data were sourced from the Aarne Thompson Uther (ATU) Index - a catalogue of over 2,000 distinct “international tale types” distributed among more than 200 cultures [22]. Each international type represents an independent, self-contained storyline comprising a combination of motifs (e.g. specific events, characters, or artefacts) that is recognizably stable across cultures. We constructed a dataset recording the cross-cultural distributions of two groups of folktales: 'Animal Tales' (ATU 1 – 299), which feature non-human protagonists, as typified by Aesop's fables, and 'Tales of Magic' (ATU 300 – 749), which concern beings or objects with supernatural powers, such as fairies, witches or magic rings [22]. We focused on these two genres because they are the most richly documented and most culturally widespread groups of tales in the ATU Index.

73 of the 198 societies in which the tales were recorded could be matched with populations for which whole genome sequences were available (Table S2-I). Of these, 33 (Dataset$_{MAIN}$) were selected based on a threshold of minimum richness (i.e. those exhibiting at least 5 folktales; **Table S2-II**) and the presence of viable genetic proxies. Each population was univocally described by a string listing the presence (1) or absence (0) of any of the included 596 folktales (**Table S2-II**).

In addition to Dataset$_{MAIN}$ we generate an additional subset which is functional to testing explicit hypotheses, i.e. Dataset$_{EURASIA}$ (N=30) which does not include the 3 African population present in Dataset$_{MAIN}$ (Table S1-II, i.e. Congolese, Tanzanian, and West African);

2 Distances

2.1 SNP filtering

The whole genome sequences used in this study were generated, QCed and phased as part of a broader study [21]. The bulk of ~39M SNPs were used to calculate the statistics described below.

2.2 Genetic, Folktale, Ethnolinguistic and Geographic distances

2.2.1 Genetic distance

Genetic distances were estimated by the average pairwise distances between two genomes, one from each population. Genetic distance for (i,j) pairs of populations represented by more than one genome each was calculated as the average of all possible (i,j) pairs of genomes. As a consequence the diagonal of the genetic distance matrix was not constrained to be zero (Table S2-3.1-3).

2.2.2 Folktale distance

Since the original dataset (Table S1-I) and Dataset$_{\mathrm{MAIN}}$ (Table S1-II) comprise binary evidence of presence (1) or absence (0) of a given folktale in a set of worldwide populations, we calculate folktale distance between populations as an asymmetric pairwise Jaccard distance[2]. Symmetric Jaccard distance between population A and population B is calculated as

$$J_\delta(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \qquad (1)$$

in other words as the ratio between the number of differences and the sum of similarities and differences which can be identified by comparing A and B. In the present work, we assume X_{ij} = 0 (absence of the *j*th tale in the *i*th population of dataset *X*) to be the ancestral state. Accordingly, we adopt an asymmetrical coefficient that does not consider absence of the *j*th tale in two sampled populations *i* and *k* (*Xij*+*Xkj* = 0) as an instance of homology (for the substantial limits posed by double zeros to inference in ecology and related disciplines please refer to Legendre and Legendre[3]).

Therefore, in a dataset *X* formed by *n* rows each representing a population univocally described by a string of presence (1)/absence (0) values of *J* folktales, we eliminate double zeros and calculate pairwise folktale distance (*F𝜹*) between population *Xi* and population *Xk* as

$$F_\delta(X_i, X_k) = \frac{\sum_{\substack{j=1 \\ k \neq i}}^{J} \quad [Xij + Xkj = 1]}{\sum_{\substack{j=1 \\ k \neq i}}^{J} \quad [Xij + Xkj = 1] + \sum_{\substack{j=1 \\ k \neq i}}^{J} \quad [Xij + Xkj = 2]} \qquad (2)$$

where square Iverson brackets equal 1 if their internal condition is satisfied and 0 if it is not satisfied. The resulting value is the ratio between the number of inter-population differences, and the sum of inter-population differences and similarities based on the presence of the *jth* folktale in both populations.

2.2.3 Geographic distance

Geographic distances were calculated as pairwise Great Circle Distance using the package *gdistance* in R [42] and by constraining the hypothesised movement of people through one waypoint located in the Sinai Peninsula. Coordinates (longitude and latitude in decimal degrees) expressing the location of each population comprised in Table S2-5 identify the assumed centre of the area occupied by a given folkloric tradition as defined by ATU index.

2.2.4 Euclidean distances

In order to perform bias corrected and partial distance Correlation, folktale, genetic, and geographic distances were transformed into their exact Euclidean representations (as indicated in Szekely et al 2007, 2013). The original folktale and genetic distance matrices were scaled through Classic Multidimensional Scaling using the function cmdscale in R and following the procedure for exact representation presented by Szekely et al. (2013a). Euclidean distances were computed from the obtained n-2 number of descriptors using the function dist in R with method set to “euclidean”. Euclidean representation of geographic distance was instead obtained by reprojecting the original set of coordinates on a plane using two-point equidistant projection through the function spTransform in the package sp in R (Pebesma and Bivand 2005; Bivand et al. 2013) . Actual Euclidean distance between the new set of coordinates was computed using the function radish in the package fields in R (Nychka et al. 2016).

3 SpaceMix

We performed two independent SpaceMix (Bradburd et al. 2013) analyses aimed at retrieving the “genetic” and the “folktale” spatial positions of our Eurasian samples. For the genetic run, we used the 126,554 SNPs of chromosome 22 that were variable in at least one of 30 samples each representing one of our studied populations. Given the high number of available markers we deemed a single chromosome to be sufficient to yield reliable results.

The geographic information was obtained from Table S2-5.1**,** the genetic information was inputted using the count/total option and Spacemix was used with the default parameters:
run.spacemix.analysis(n.fast.reps=10, fast.MCMC.ngen=1e5, fast.model.option="target",
long.model.option="source_and_target", data.type="counts", sample.frequencies = NULL,

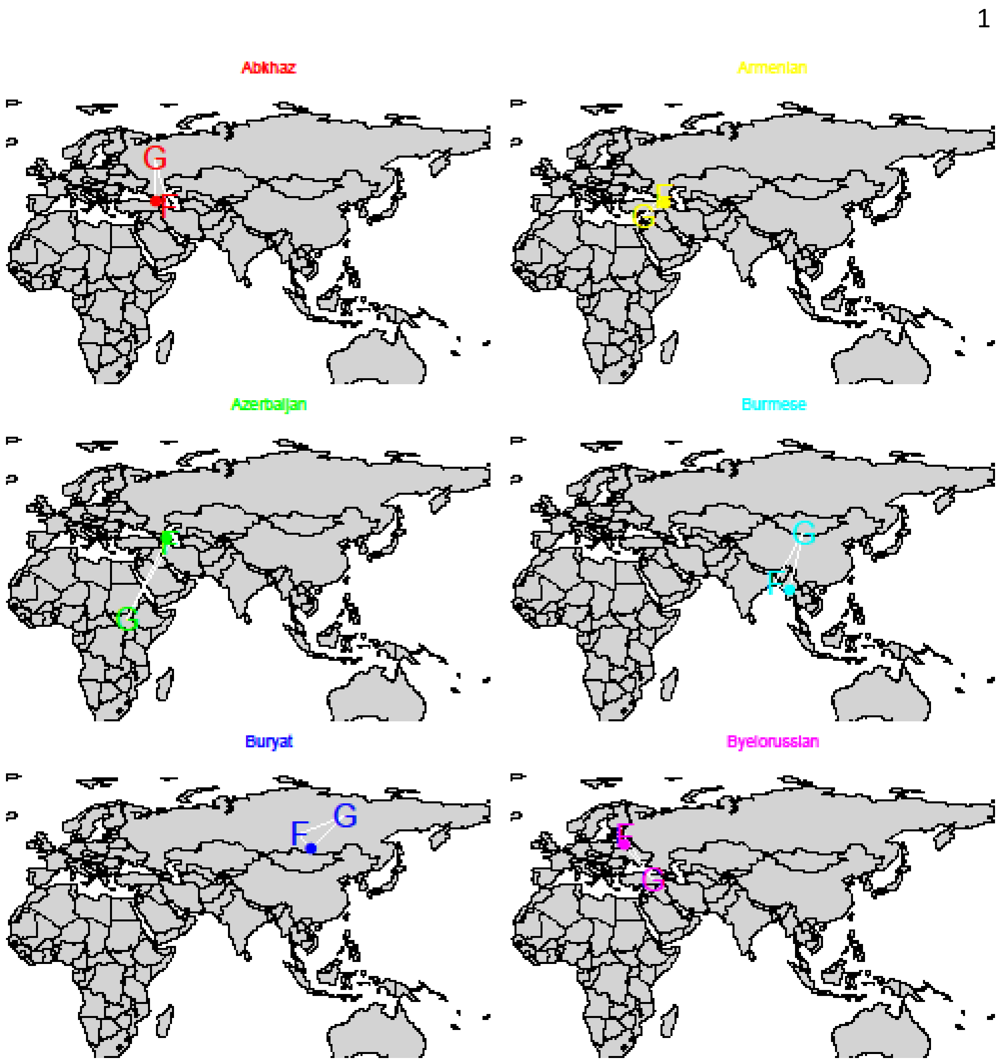
```
mean.sample.sizes = NULL, counts = count, sample.sizes = total, sample.covariance = NULL,
target.spatial.prior.scale = NULL, source.spatial.prior.scale = NULL, spatial.prior.X.coordinates=coord[,1],
spatial.prior.Y.coordinates=coord[,2], round.earth=FALSE, long.run.initial.parameters = NULL,
k=nrow(count), loci=ncol(count), ngen=1e6, printfreq=1e2, samplefreq=1e3, mixing.diagn.freq = 50,
savefreq=1e5, directory = NULL, prefix = "OurPrefix")
```

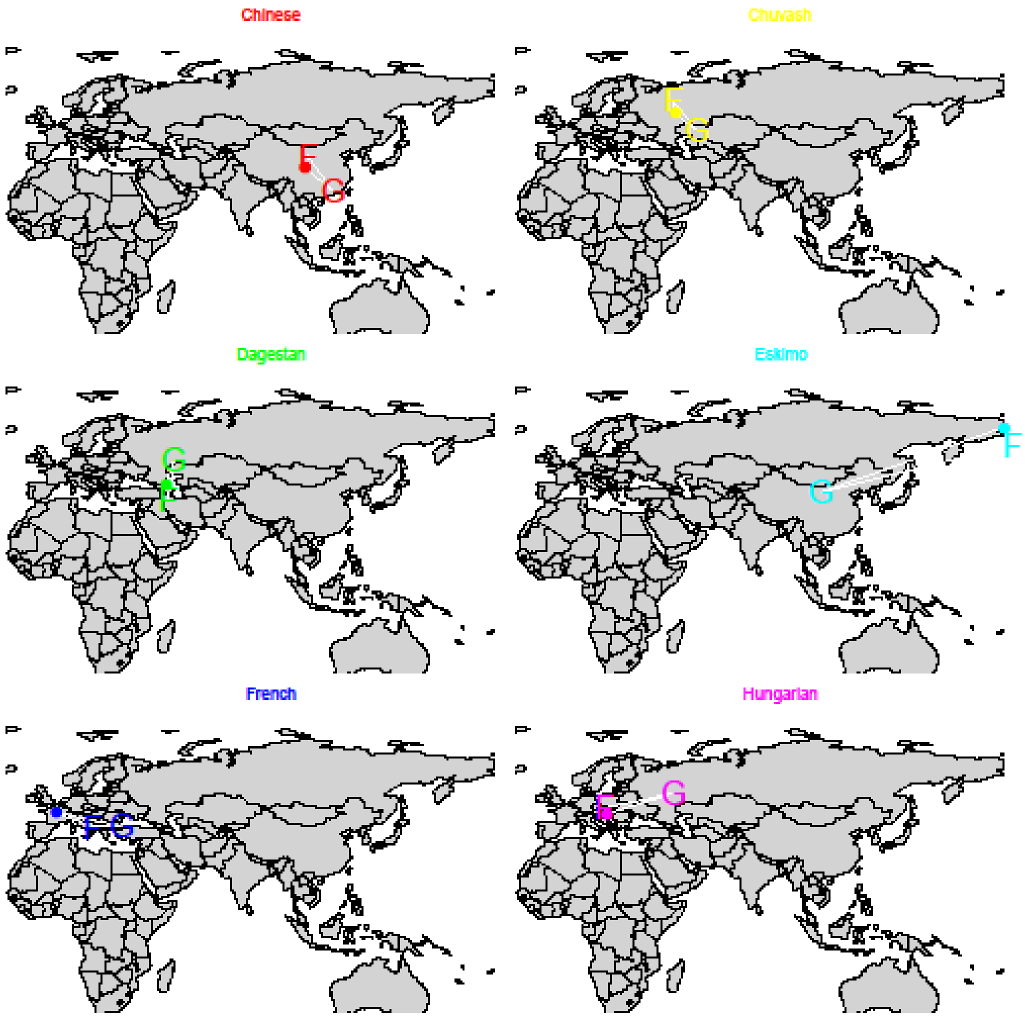
This yielded a “geno-geographic” positioning for each of the 30 samples.

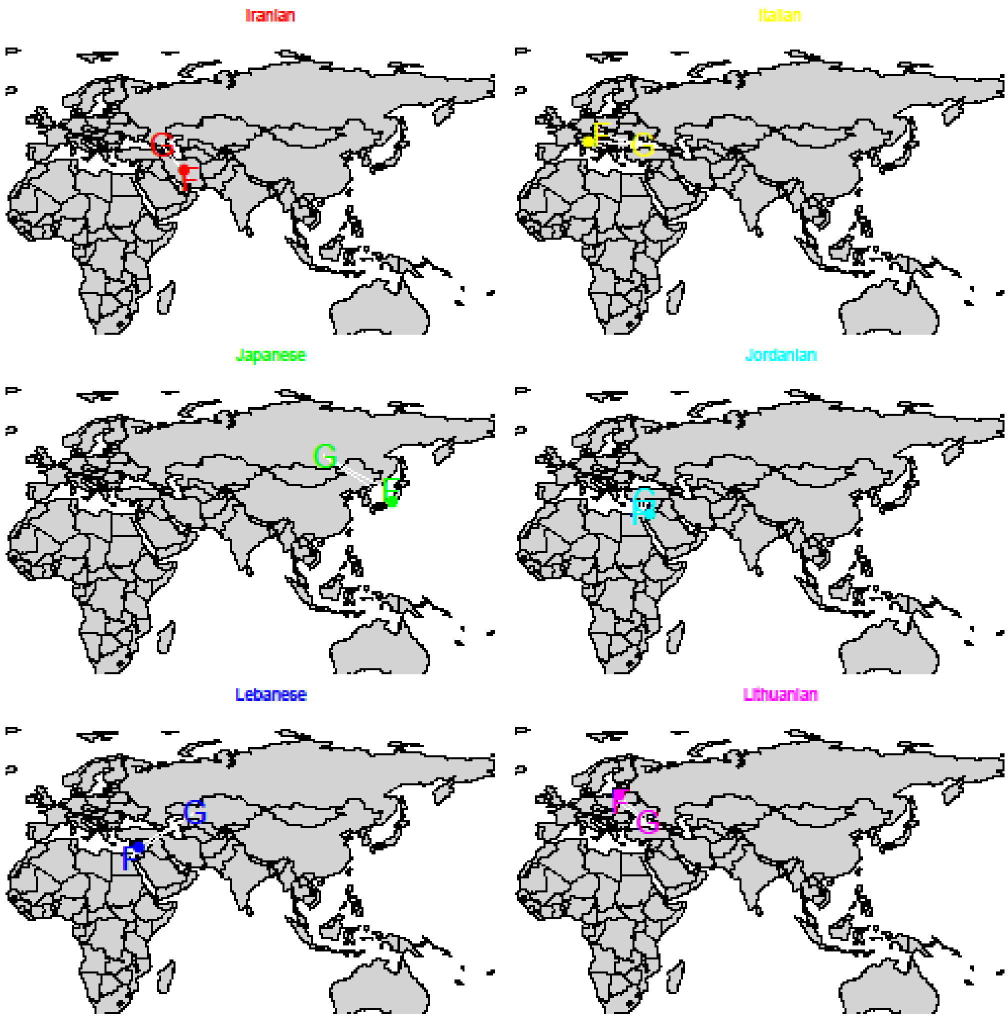
The same procedure was replicated using this time “folktale” information as input data. Particularly we generated a pseudo-genetic file where each population was typed as a single individual and the presence of a given folktale was registered as an homozygous trait.

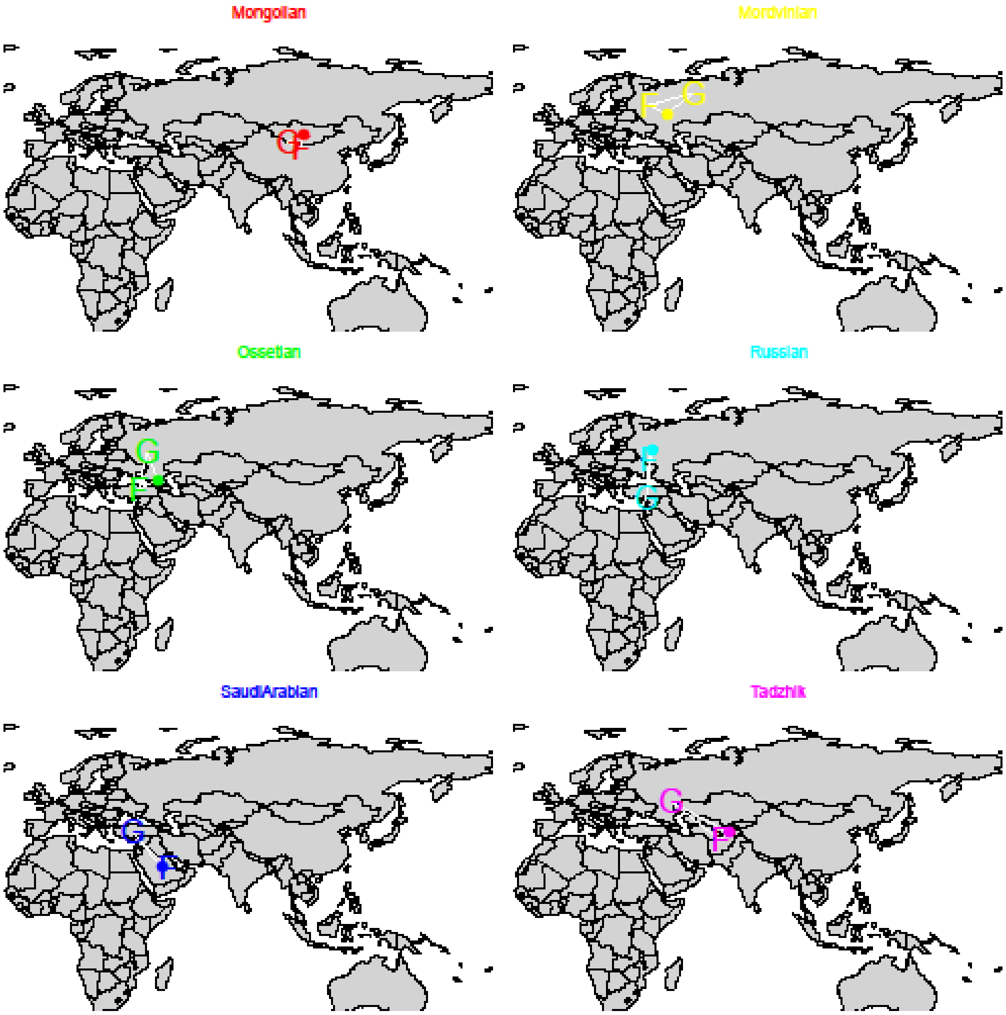
The “geno-geographic” and “folk-geographic” coordinates hence generated were compared with the actual geographic coordinates, denoting a tendency for the folk-geographic coordinates to approximate better than the geno-geographic ones the actual geographic locations of the sampled populations (Figure S1-3.1).

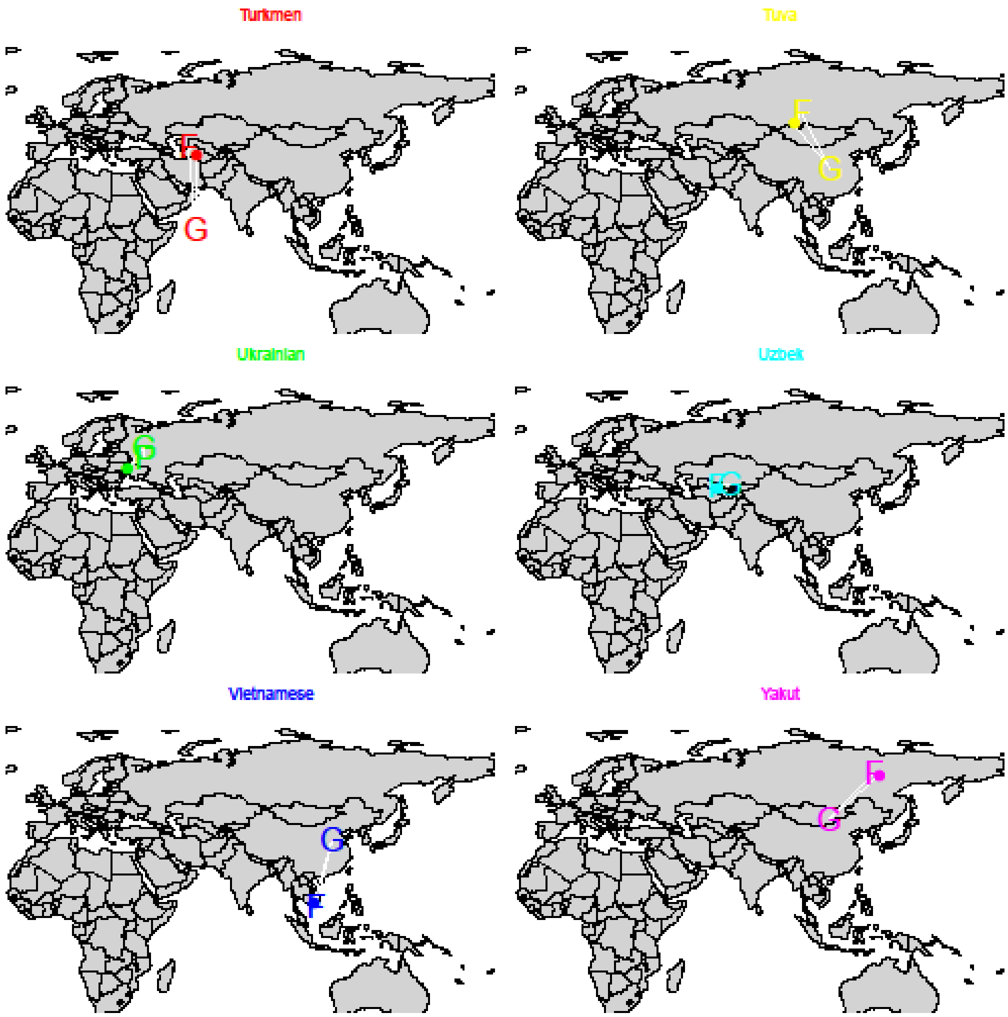
Figure S1-3.1 SpaceMix analyses for each of the Eurasian populations showing the geographic (dot), geno-geographic (G) and folk-geographic (F) coordinates joined by white segments.


1
Abkhaz
G
F
Armenian
F
G
Azerbaijan
F
G
Burmese
G
F
Buryat
F
G
Byelorussian
F
G


Chinese
F
G
Chuvash
F
G
Dagestan
G
F
Eskimo
F
G
French
F
G
Hungarian
F
G


Iranian
G
F
Italian
F
G
Japanese
G
F
Jordanian
G
F
Lebanese
G
F
Lithuanian
F
G

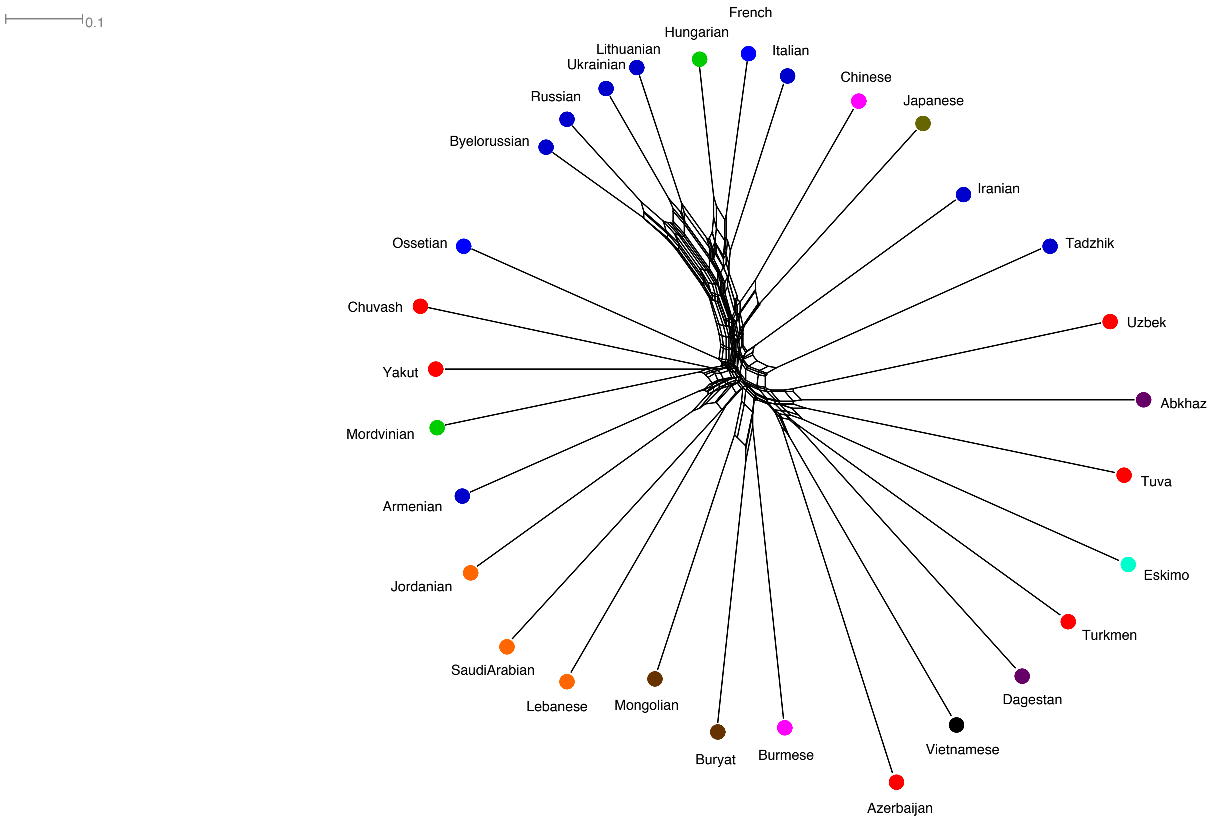

Mongolian
G
F
Mordvinian
F
G
Ossetian
G
F
Russian
F
G
SaudiArabian
G
F
Tadzhik
G
F


Turkmen
F
G
Tuva
F
G
Ukrainian
G
F
Uzbek
F
G
Vietnamese
G
F
Yakut
F
G

4. NeighborNet

A Neighbornet analysis was also carried out to further explore the impact of geography and linguistic ancestry on the distribution of folktales in the present dataset. The analysis yielded the following graph (Fig. S1-4.1), exhibiting a certain degree of spatial clustering in addition to proximity and reticulation among linguistic close relatives (e.g. within the Indo-European family, the Semitic family, and the Mongolian family). Overall, the spatial structure in the dataset seems to be stronger (e.g. the position of Hungarian, Japanese/Chinese). Some language families, notably Turkic, Uralic and Caucasian, are scattered across the network. The degree of reticulation is quite high, as testified by the relevant quartet statistics (delta score = 0.317; Q-residual score = 0.002335), suggesting that cultural admixture processes between demes may have an important role.

Fig. S1-4.1. Neighbornet graph based on folktale distance.Linguistic color code: red = Turkic; blue = Indo-European; pink = Sino-Tibetan; purple = Caucasian; turquoise = Eskimo-Aleut; orange = Semitic; light green = Uralic; dark green = Japonic, brown = Mongolian; black = Austro-asiatic



5. AMOVA

We use Analysis of Molecular Variation (AMOVA; Excoffier et al. 1992) to formally asses the impact of ethnolinguistic boundaries on both genetic and cultural (folktale) variability. To do this, we assigned each population to an ethnolinguistic group (derived from Ethnologue; SI table S2-9.1), and used the function amova in the package pegas in R to run the analysis (Paradis 2010). AMOVA is commonly used in population genetics to assess the degree of population structure in a metapopulation. In other words, it measures the degree of variability existing between predetermined groups as opposed to the amount of variability observed within each group. AMOVA returns a summary statistic (Phi_{ST}) obtained by computing the ratio between intergroup diversity estimate and total diversity in the metapopulation. Although it is derived from the more general class of F_{ST} measures, Phi_{ST} evaluates symmetric distance matrices while F_{ST} is based on correlations between individual variant frequencies. Such measures have already been successfully adopted to investigate co-evolutionary patterns in genetic and cultural datasets (Bell et al. 2009; Rseszutek et al. 2012), in cultural datasets alone (Shennan et al 2015), and in one case on the distribution of folktale variants in Europe (Ross et al. 2013). Results of these works consistently show that average intergroup cultural dissimilarity is stronger than genetic dissimilarity measured on the same set of demes, while PhiST values obtained for cultural markers are usually in a range comprised between 0.02 (musical diversity; Rseszutek et al. 2012) and higher levels for variants of individual folktales (PhiST =0.09; Ross et al. 2013), or personal ornaments (Phi_{ST} =0.109) and pottery (PhiST =0.134) in Neolithic Europe (Shennan et al 2015). Our results confirm this differential impact on genetic variability on the one hand, and cultural variability on the other.

6. Bias-corrected distance correlation and partial distance correlation

Distance correlation is a measure of statistical dependence between two variables which is specifically suited for testing such hypothesis on pairs of symmetric distance matrices. Its value equals zero if and only if the two variables are statistically independent (Székely et al. 2007). In addition: a) the resulting statistics are not bound to linear models. On the contrary, they are sensitive to all types of dependent relationships, including nonlinear and non monotone ones (Székely et al 2007); b) it is not prone to the same problems raised for standard and partial Mantel tests (Guillot and Rousset 2013); and c) it is usually preferred to other measures of nonlinear association when small or practical sample size are concerned (Gorfin et al 2011).

Given two distance or dissimilarity matrices, standard distance correlation performs double centering of each matrix by subtracting row and column average to rows and columns of the original matrix, and adding the grand mean of the distance matrix to the results, so that all columns and rows of the resulting matrices sum to zero. Distance correlation between two such scaled matrices - as in Pearson's Product-moment correlation coefficient is computed by dividing the distance Covariance by the product of the respective distance standard deviations. Distance Covariance is obtained by computing the summed cross-product between the two double-centered matrices and averaging it over squared sample size. Distance correlation is therefore not the correlation between original distances. It is instead based on cross-products between scaled moment obtained by double-cantering the original matrices.

In the present paper, we perform Bias-Corrected Distance Correlation (Szekely and Rizzo 2013) suited for bigger sample size (in the present study we have 435 observation when all pairs are considered over 30 populations, which is exactly the example size provided by the authors and developers of the method), This method corrects for potential limitations of original distance Covariance and distance Correlation measures (Székely et al. 2007) when dimension tends to infinity, and is based on an unbiased estimator or the squared distance population covariance. A suited t-test of independence is offered. In addition, we perform Partial distance correlation (Szekely et al 2013a) to assess the impact of one variable over another, while controlling for the effect of a third variable. For calculating partial distance correlation the standard double-centering used in standard and bias-corrected distance correlation is replaced by a different centering technique named U-centering and based on the demonstration that such transformed distance and dissimilarity matrices have a corresponding U-centered Euclidean representation in Hilbert space (a generalization of Euclidean plane with a finite or infinite number of dimensions; Szekely et al 2013a). Simulation studies confirm that the joint significance test controls type I error rate at its nominal level, and outperforms partial correlation and partial Mantel test in terms of power (Szekely et al 2013a).

7. Exploring association between variables over cumulative geographic distance

Table S1-7.1 Model comparison over cumulative geographic distance. Results report Pearson's product-moment correlation coefficients plotted in Fig.2 and obtained with original distance matrices.

N	bindist	r.folk	p.folk	r.folk(Lw)	p.folk(Lw)	r.folk	p.folk	r.folk(Lw)	p.folk(Lw)
		geo	geo	geo	geo	geno	geno	geno(Lw)	geno(Lw)
115	2000	0.16	0.09	0.21	0.03	0.45	<0.001	0.40	<0.001
249	4000	0.24	<0.001	0.58	<0.001	0.34	<0.001	0.40	<0.001
343	6000	0.22	<0.001	0.68	<0.001	0.23	<0.001	0.51	<0.001
412	8000	0.22	<0.001	0.67	<0.001	0.22	<0.001	0.55	<0.001
434	10000	0.19	<0.001	0.64	<0.001	0.20	<0.001	0.55	<0.001
435	12000	0.19	<0.001	0.64	<0.001	0.20	<0.001	0.55	<0.001

8. Diffusion of most popular tales and estimation of possible focal points from spatial distribution

We identify 19 “most popular” tales, i.e. folktales that are present in at least 30 populations out of 60 Old World populations available in the original presence/absence matrix (Table S2-10, S2-11), which are briefly summarized below:

- **ATU 155 ‘The Ungrateful Snake Returned to Captivity’**: A snake (or another dangerous animal) attacks a man who rescues it, and is punished by other animals.
- **ATU 300 ‘The Dragon Slayer’**: A man rescues a beautiful maiden from a dragon/monster, often with the help of his dogs. Later, he exposes an imposter who claims credit for the deed.
- **ATU 301 ‘The Three Stolen Princesses’**: A man rescues three women from a pit. His companions betray him by leaving him in the pit and stealing the girls. With the aid of a spirit the hero flies up and exposes his companions, marrying the youngest girl.
- **ATU 303 ‘The Twins, Or Blood Brothers’**: A hero rescues a woman and marries her, but is later bewitched. His twin brother sets out to find him, and is mistaken by the woman for her husband. The twin releases his brother, who kills him in a jealous rage, mistakenly believing him to have seduced his wife. The twin is later resuscitated.
- **ATU 313 ‘The Magic Flight’**: A man elopes with the daughter of a demon or king. She uses magical objects to obstruct their pursuers and they escape.
- **ATU 314 ‘Goldener’**: A golden-haired man marries the king’s daughter. He is mocked by his brothers-in-law, but succeeds in completing heroic deeds where they fail and is made the heir.
- **ATU 325 ‘The Magician and his Apprentice’**: A boy is given to a magician to be his apprentice. The boy learns the art of sorcery and frees himself from his master after a battle in which they transform into a succession of different kinds of animals.
- **ATU 400 ‘The Man on a Quest for his Lost Wife’**: A miscellaneous group of stories concerning a man who is separated from his wife during an adventure. When he finds her she is about to marry another man, but he proves his identity to her and they are reconciled.

- **ATU 403 'The Black and the White Bride'**: A girl is to marry the king, but her stepmother tries to kill her and replace her with her own daughter. The girl proves her identity to the king and exposes her stepsister as an imposter.
- **ATU 480 'The Kind and Unkind Girls'**: A girl goes on a journey and is kind to those she encounters. She is rewarded. Her stepmother sends her own daughter on the same journey but she is unkind and gets punished.
- **ATU 531 'The Clever Horse'**: A miscellaneous group of stories in which a young man is helped to complete some near-impossible tasks by a talking horse and marries a princess.
- **ATU 550 'Bird, Horse and Princess'**: Three brothers are sent on a quest by their father to catch a magic bird. The youngest brother succeeds but is betrayed by the other two who try to claim the prize. With the help of a magical animal the hero exposes his brothers.
- **ATU 554 'The Grateful Animals'**: A man helps a series of animals, who reciprocate by helping him to complete a series of near-impossible tasks.
- **ATU 560 'The Magic Ring'**: A boy acquires a magic ring that grants him wishes. He marries a princess, who steals the ring to elope with her lover. The boy recovers a ring and punishes his faithless wife and her lover.
- **ATU 563 'The Table, the Donkey and the Stick'**: A man acquires magical objects from a supernatural being. He is cheated out of them and given plain objects in their place, but manages to recover his possessions and punish the cheat.
- **ATU 613 'The Two Travellers'**: After losing an argument with his companion, a man is blinded. He learns the secrets of birds and recovers his sight as well as gaining new powers. His companion imitates him and is punished by the birds.
- **ATU 670 'The Man Who Understands Animal Languages'**: A snake teaches a man the languages of animals on condition he keeps it a secret. The man's wife nags him to teach her but he refuses after being warned of the consequences by a male animal (usually a rooster).
- **ATU 700 'Thumbling'**: A couple wish for a child and are given a tiny boy through supernatural means. The boy is lost and goes on a series of adventures until he is reunited with his parents.
- **ATU 707 'The Three Golden Children'**: A woman marries a king and gives birth to three children, who are stolen by her jealous sisters. When they grow up the children go on a quest to find their parents and eventually expose the sisters.

One possible explanation for the wide dispersion of these tales is that they spread through the dissemination of written texts, which would allow them to travel much further and in a much shorter period than would be possible solely through the vectors of human dispersal and traditional oral transmission. Such a process would most likely have coincided with the emergence of the fairy tale as a popular literary genre in the sixteenth and seventeenth centuries (Bottigheimer 2014). Although many folktales were incorporated into literary works prior to this period (for example, in medieval romances), the development of new, cheap printing technologies together with the growth of international trade networks and European colonialism would have allowed tales to circulate to much wider audiences than was previously possible (ibid.).

In fact, seven of the tales listed above were published in major fairy tale collections during this period, including Giovanni Francesco Straparola's *Le Piacevoli Notti* in 1550-55 (ATU 314, ATU 325, ATU 670), Giambattista Basile's *Lo cunto de li cunti* in 1634 (ATU 301, ATU 480, ATU 560), Charles Perrault's *Histoires ou contes du temps passé* in 1697 (ATU 480, ATU 700). However, one obvious question raised by the hypothesis that these tales spread via textual transmission is why the other stories contained within these collections did not achieve similarly wide cross-cultural distributions. Secondly, while literary versions have undoubtedly made a major contribution to the modern forms of these tales, there is compelling evidence that they were derived from already well-established and widespread oral traditions, rather than the other way round (Ben-Amos et al. 2010). For example, ATU 301 ‘The Three Stolen Princesses’ occurs in Greek and Indian myths that long predate Basile’s Italian fairy tale of 1634. Similarly, versions of ATU 325 ‘The Magician and His Pupil’, ATU 560 ‘The Magic Ring’ and ATU 670 ‘The Man Who Understands Animal Languages’ appear in Indian and Middle Eastern sources (including the Ramayana and One Thousand and One Nights) that are clearly independent of later European literary versions of these tales (Thompson 1977).

An alternative explanation for the distribution of these tales is that they reflect signatures of demic or cultural diffusion that are more ancient than the other patterns detected in the dataset. In order to characterize these signatures, we sought to identify possible centers of origin and dispersal for the tales. To do so, we assessed the amount of linear correlation between geographic distance from each population exhibiting a given tale and the distribution of the proportion of the remaining populations displaying that tale over the same geographic gradient. More specifically, we binned pairs of populations into fixed intervals of geographic distance (2000 Km), and for each bin we calculated the proportion of populations exhibiting a given folktale. We then calculate linear correlation between the distribution of percentages obtained for each geographic bin and geographic distance from all the populations in the dataset that exhibit that folktale. All the above mentioned analyses have been performed in R. The assumption is that - if we envisage a long-range and ancient diffusion process whose vectors are solely human dispersal and traditional cultural transmission - we expect to obtain a distance-decay patterning with higher percentages indicating potential “origin populations” from which increasingly lower percentages depart forming a clinal trend over geographic distance. To avoid losing information we did not focus on single coordinate pairs exhibiting the highest values for each tale, and plotted instead the distribution of correlation coefficients on a map in order to visually define the most probable area of origin (centres of origin exhibiting the lowest correlation coefficients; Figure S8-I). Probability surfaces were obtained by interpolating correlation coefficients computed for each population using the function producing plate spline interpolation of the *fields* package in R [47].

The resulting potential patterns of diffusion are represented in Figure S6-I. Although each of these patterns presents some specificities, some general trends can be found - as summarized in the main text. In particular, four main multi-directional waves of diffusion can be hypothezised:

1. Potential African origin (e.g. ATU 670)
2. Southward spread from northern Eurasia (e.g. ATU 700)
3. Eastern European origin (e.g. ATU 301, 303, 313)
4. Middle-Eastern/Caucasian origin (e.g. ATU 314, 400, 480, 560)

While further research is needed to verify these patterns (for example, by reconstructing the evolutionary histories of variants of each tale type to test whether they match the dispersal scenarios), the results have significant implications for current understandings about the origins of international folktale traditions. In particular, they suggest a less Euro-centric view of tale origins than traditional "historic-geographic" reconstructions based on the frequency of variants (i.e. the number of versions of a given tale type recorded in each population) and chronology of literary versions. A major problem with this approach is that conclusions about a tale's origins may often be skewed by the strong European bias in both the richness of the folktale and literary records. For example, ATU 300 'The Dragon Slayer' – which has been proposed to be the original archetype storyline from which all fairy tales are derived (Propp 1968)– was previously believed to have originated in medieval western Europe, most likely France, where the earliest known versions were recorded. Our analysis instead suggests that this tale – together with the related tale ATU 313 'The Twins' may have arrived in western Europe from further east, either from the region of modern day Ukraine and Belarus, or of Turkey and Kurdistan. Similarly, whereas folklorists have claimed that ATU 670 'The Man Who Understands Animal Languages' originated in Europe and was transported to Africa through colonialism, our findings reverse the direction of transmission and suggest that the tale probably arose in North Africa.

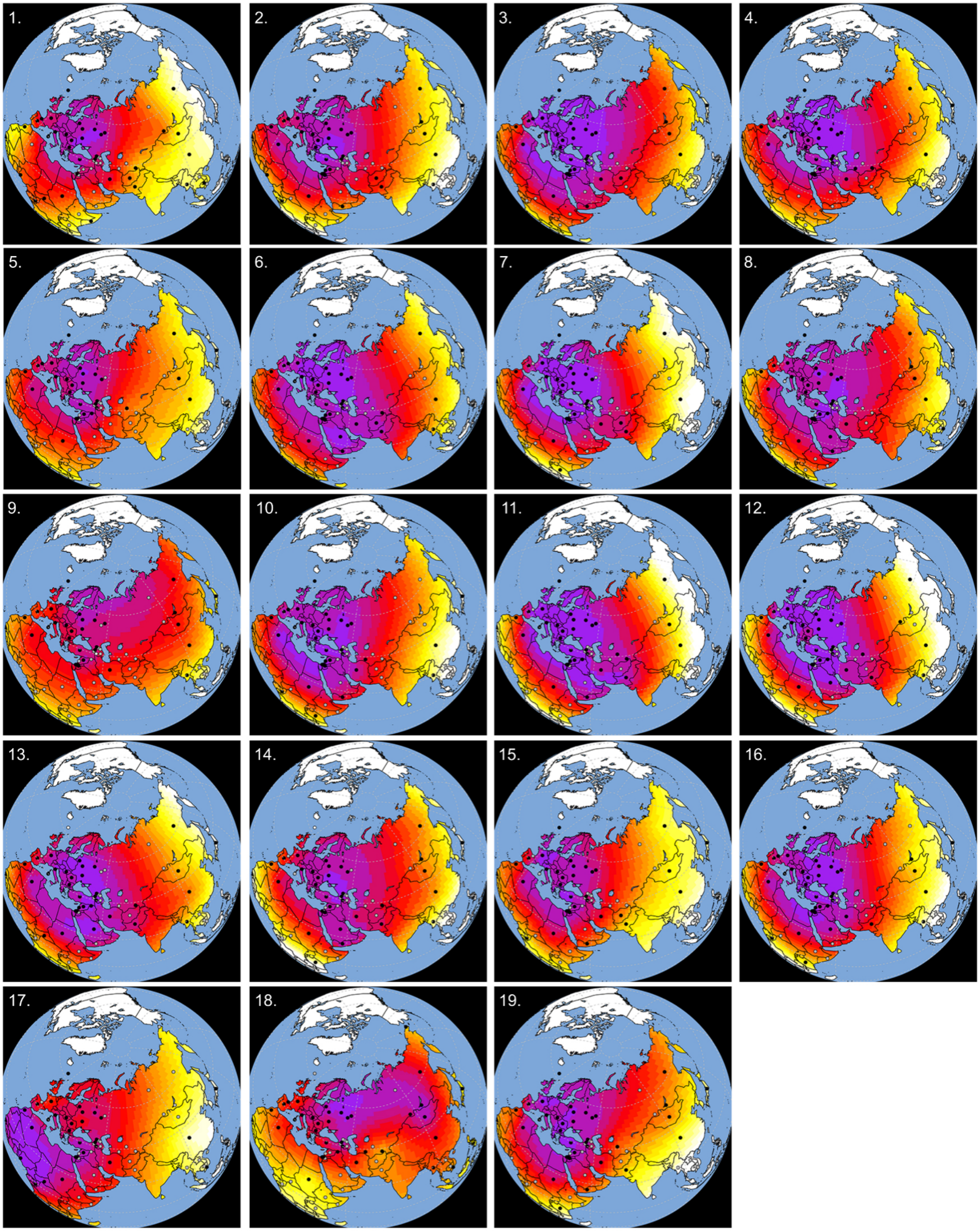

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.

Figure S1-8-I - Plot of the probable areas of origin of the 19 “most popular” tales: Probability surfaces have been obtained interpolating correlation coefficients computed for each population. Grey dots indicate populations that do not exhibit the specific tale of interest. 1) Tale 155; 2) Tale 300; 3) Tale 301; 4) Tale 303; 5) Tale 313; 6) Tale 314; 7) Tale 325; 8) Tale 400; 9) Tale 403; 10) Tale 480; 11) Tale 531; 12) Tale 550; 13) Tale 554; 14) Tale 560; 15) Tale 563; 16) Tale 613; 17) Tale 670; 18) Tale 700; 19) Tale 707.