The Tell-Tale Genome

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, IMF-CSIC, Spanish National Research Council, 08001 Barcelona, Spain; [b]Department of Humanities, Universitat Pompeu Fabra, 08005 Barcelona, Spain; [c]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, Via Ugo Bassi 58/B, 35131 Padova, Italy; [f]McDonald Institute for Archaeological Research, CB2 3ER Cambridge, UK; [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; [i]Department of Anthropology, Durham University, DH1 3LE Durham, UK

This manuscript was compiled on May 24, 2017

Observable patterns of cultural variation are consistently intertwined with demic movements, cultural diffusion, and adaptation to different ecological contexts (Cavalli-Sforza and Feldman 1981; Boyd and Richerson 1985). The quantitative study of gene-culture co-evolution has focused in particular on the mechanisms responsible for change in frequency and attributes of cultural traits, on the spread of cultural information through demic and cultural diffusion, and on detecting relationships between genetic and cultural lineages. Here, for the first time, we make use of worldwide whole-genome sequences (Pagani et al. 2016) 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) 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 ∼4000 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 we propose, the one entailing cultural diffusion biased by linguistic differences is the most plausible one. Additionally, we identify 15 tales which are more likely to be predominantly transmitted through population movement and replacement, and locate putative focal areas for a set of tales which are spread worldwide.

Cultural diffusion | Demic diffusion | Folktales | Whole-genome sequences

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 and cultural diffusion are not mutually exclusive conditions, rather opposite extremes of a continuous gradient whose intermediate and composite positions 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 processes (IBD; [21]). Under an IBD model, individuals or groups which are spatially closer to each other are expected to be more similar than individuals or groups that are located further apart. A positive correlation between cultural dissimilarity and geographic distance between samples is therefore used to infer processes of cultural transmission of non-adaptive information without population replacement [8, 17]. On the other hand, observed genetic distance is the composite result of serial founder events (SFE), 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 single out the relative effect of demic processes on the distribution of cultural variants [8].

In a recent study Creanza and colleagues [17] investigated the process responsible for the observed global distribution of (phonetic) linguistic variability by comparing it to genetic and geographic distances. The authors found high correlation between genetic and geographic distances at a worldwide scale, while linguistic distances were spatially autocorrelated only within a range of ∼10000 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, while it is considerably lower at smaller geographic distances (below ∼6000 km for the present dataset), possibly because of more recent and short-range population movements (Figure 1 top, yellow line). It is worth remembering that global trends have been forming over the past ∼40000 years, while most cultural traditions are likely to have evolved more recently. This 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, to 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 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.

The 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 analyze 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 an 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 have 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 their geographical proximity, suggesting 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 Europe. Ross and colleagues [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 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].

Significance Statement

This paper presents unprecedented evidence on the transmission mechanism underlying the spread of a broad cross-cultural assemblage of folktales in Eurasia and Africa. For the first time, 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 to identify individual stories which are more likely to be transmitted through population movement and replacement. The results provide a novel, empirical solution to operate with linguistic barriers and highlight the impossibility to disentangle 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.

Overview of the present study

In the present study we focus on 596 folktales comprising "Animal Tales” and “Tales of Magic” [27], typed as present (1) or absent (0) in 33 populations (DatasetMAIN) for which whole- genome sequences are available and exhibiting presence of at least five folktales (Figure 1 b; SI Appendix Section 1; Dataset S1 Table S2-I, Table S2-II, Table S2-3.1-3, Table S2-6). 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 (Dataset S1 Table S2-4,Table S2-5). We measure and compare the fit of a number of alternative models comprising: a) 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); b) population movement and admixture between demes (demic process) as a substantial additional driver of folktale transmission; and c) 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 [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. We use this parameter combination to generate alternative models whose fitness 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 further analysis on the diffusion of individual or smaller packages of tales, we infer potential focal areas - intended as a

Conceived the study: EB, LP, JJT; Collected data: SGdS; Analysed data: EB, LP, SS, JJT; Contributed to interpretation of results: EB, LP, ERC, MS, GM, MM, DP, DL, JJT; Wrote manuscript: EB, LP, ERC, CB, JJT

The authors declare no conflicts of interest.

[1] EB and LP contributed equally to this work.

[2] To whom correspondence should be addressed. E-mail: eugenio.bortolini@imf.csic.es; jamie.tehrani@durham.ac.uk

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 asses the impact of ethnolinguistic boundaries on both genetic and folktale variability. To do this, we assign each population to an ethnolinguistic group (see Materials and Methods section; SI Appendix section 5; tDataset S1 able SI2-9.1). Our analysis yielded Φ_{ST}= 0.036 (p<0.001) for genetic distance matrix while Φ_{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 consisted 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 (See Materials and Methods for details). We find that the parameter combinations that resulted in the highest correlation between genetic-geographic distances (intensity=0.1; radius=1500 km) and between folktale-geographic distances (intensity=0.3; radius=3000 km) implies 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 SI2-3.4) and folktales (folktaleL; Table SI2-4.2) respectively. By using raw and corrected distance matrices, we define alternative models as: a) biased cultural diffusion (folktaleL∼geographic); b) demic diffusion (folktale∼genetic); and c) biased demic diffusion (folktaleL∼geneticL).

Assessing models of folktale transmission. We set to test for deviations from the null model of cultural diffusion due to IBD focusing only on Eurasian populations (DatasetEurasia, N=30) to control for the effect of the Out of Africa expansion on genetic distance (See S1 Section 3 for further details). We explore the relationship between our genetic, folktale, and geographic distance matrices using SpaceMix [28] (SI1 section 3). We note that, when transformed into pseudo-spatial coordinates, folktale distances tend to match actual geographic coordinates better than genetic distances (Fig.1c and SI1 Fig. 3.1). The role of geography and of 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 (SI1 section 4).

We then asses the goodness of fit of all 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] (Table 1; See Materials and Methods and SI1 sections 2,6 for details). 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 could be due 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; Fig. SI Appendix 1 Section 7; Table SI1-7.1). Above 4000 km language-biased cultural diffusion presents with the highest fit at all bins, followed by language-biased demic diffusion. Under 4000 km folktale distance exhibits stronger dependence from genetic distance than from geographic distance. This is particularly visible under 2000 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 offer 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 Mann-Whitney-Wilcoxon test. We focus on the 308 folktales that are present in at least five populations and run two separate tests, the first considering all pairs of populations (Table S2-7.1), and a second considering only those within a conservative geographic range of 6000 km (Fig.1a; Table S2-7.2). After Bonferroni correction, 15 out of the 308 analyzed folktales (4.9%; Table S2-8.1, S2-8.2) 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 S2-8.3). Focal areas were generated for the 19 most widespread folktales, which follow four main trends (SI Appendix Section

8). 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 ATU155, "The Ungrateful Snake Returned to Captivity", Figure S1-8-I 1; ATU313 or "The Magic Flight", Fig.3), while others probably started their journey from Caucasus (Figure 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; Figure S1-8-I 18), while a last group could have spread from Africa (Figure S1-8-I 17), 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 demonstrate 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 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 on cultural evolutionary processes at a cross-cultural and cross-continental scale. Our results provide new 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$4000 km, cultural diffusion biased by linguistic barriers exhibits the highest correlation at all geographic bins. At small geographic bins ($< 4000km$) 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 which are genetically closer than populations not exhibiting those tales. Looking for smaller packages of tales or for 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 Section 8;[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/anti-conformism, heterophily, content-dependent biases [5, 23, 30]) or cultural barriers different from linguistic ones, whose chronology may be independently ascertained.

Materials and Methods

Dataset description. Folktale data were sourced from the Aarne Thompson Uther catalogue (ATU; [27]).The present dataset comprises "Animal Tales" (ATU 1-299), and "Tales of Magic" (ATU 300-749). Of the 198 societies in which the tales were recorded, 73 matched available genetic data (Table S2-1). Of these, 33 populations exhibiting at least 5 folktales were selected (Table S2-2.1, Figure 1 b). Each population is described by a string listing the presence (1) or absence (0) 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 non-coding 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 (Table SI2-3.1b). Folktale distance between population pairs was calculated as asymmetric Jaccard distance [40] (Table S2-4.1a). 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; Table S2-5.1a) 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. In order 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 (Tables SI2-3.5 and SI2-4.3). Euclidean representation of geographic distance (Table SI2-5.1) 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].

Analysis of MOlecular VAriance (AMOVA). To implement AMOVA [29] in our analysis, each population was assigned to an ethnolinguistic group derived from Ethnologue (Table SI2-9.1), and we used the function amova in the package pegas [47] in R. Significance values are obtained through permutation (1000 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 at cumulative geographic scales. Measures were obtained through: a) Pearson's product-moment correlation coefficient using the function cor.test in R; b) bias-corrected distance correlation [33] using the function dcor.ttest in the package energy in R [48]; and 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 pseudo-coordinates 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 increases 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 (1500, 0.1) and (3000, 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), $Dgen'_{ij} = Dgen_{ij} * (1 - f_G)$ if $d_{ij} \leqslant d_G$, and $Dfolk' = Dfolk * (1 - f_F)$ if $d_{ij} \leqslant d_F$.

Data availability and codes. R scripts and related commands used to generate all the results described in the paper are available in Supplementary Appendices. Folktale and geographic data, as well as genetic distances, are also available in Supplementary Appendices. Genetic data used to run SpaceMix are taken from [24] (www.ebc.ee/free_data).

ACKNOWLEDGMENTS. The authors are grateful to Adrian Timpson, Anne Kandler, Dugald Foster, Jeremy Kendal, Rachel Kendal, Simon Greenhill, and two anonymous reviewers for their comments and useful suggestions. EB is supported by SimulPast Consolider Ingenio project (CSD2010-00034) funded by the Spanish Ministry of Science and Innovation. LP is supported by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.16-0030 and No. 2014-2020.4.01.15-0012)

1. Currie T, Greenhill S, Gray R, T. H, Mace RR (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. *Proceedings of the Royal Society B: Biological Sciences* (282):20150061.
3. da Silva S, Tehrani J (2016) Comparative phylogenetic analyses uncover the ancient roots of indo-european folktales. *Royal Society Open Science* (3):150645.
4. Cavalli-Sforza LL, Feldman MW (1981) *Cultural Transmission and Evolution. A Quantitative Approach.* (Princeton University Press).
5. Boyd R, Richerson PJ (1985) *Culture and the Evolutionary Process.* (University of Chicago Press).
6. Collard M, Shennan SJ, Tehrani J (2006) Branching, blending and the evolution of cultural similarities and differences among human populations. *Evolution and Human Behavior* 27:169–184.
7. Ackland GJ, Signitzer M, Stratford K, Cohen MH (2007) Cultural hitchhiking on the wave of advance of beneficial technologies. *Proceedings of the National Academy of Sciences* 104(21):8714???8719–8714???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(8):e6747–e6747.
9. Gray RD, Bryant D, Greenhill SJ (2010) On the shape and fabric of human history. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365(1559):3923–3933.
10. Fort J (2012) Synthesis between demic and cultural diffusion in the neolithic transition in europe109. *Proceedings of the National Academy of Sciences of the United States of America* 109:18669–?18673.
11. Lycett SJ (2015) Cultural evolutionary approaches to artifact variation over time and space: basis, progress, and prospects. *Journal of Archaeological Science* 56:21–31. Scoping the Future of Archaeological Science: Papers in Honour of Richard Klein.
12. Ammerman AJ, Cavalli-Sforza LL (1984) *The Neolithic Transition and the Genetics of Populations in Europe.* (Princeton: Princeton University Press).
13. Renfrew C (1992) Archaeology, genetics and linnguistic diversity. *Man* (27(3)):445–478.
14. Renfrew C (2001) From molecular genetics to archaeogenetics. *Proceedings of the National Academy of Sciences* (98(9)):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. *Proceedings of the National Academy of Sciences* 106(42):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(8):e1000491–e1000491.
17. Creanza N et al. (2015) A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences* 112(5):1265–1272.
18. Haak W et al. (2015) Massive migration from the steppe was a source for indo-european languages in europe. *Nature* 522(7555):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. *Journal of Archaeological Science* 43:289–298.
20. Fort J (2015) Demic and cultural diffusion propagated the neolithic transition across different regions of europe. *Journal of the Royal Society 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. U S A* 102(44):15942–7.
23. Ross RM, Greenhill SJ, Atkinson QD (2013) Population structure and cultural geography of a folktale in europe. *Proceedings of the Royal Society of London B: Biological Sciences* 280(1756):20123065–20123065.
24. Pagani L et al. (2016) Genomic analyses inform on migration events during the peopling of eurasia. *Nature* 538(7624):238–242.
25. Grimm W (1884) *Preface to Children's and Household 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. *Evolution and Human Behavior* 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.* (Helsinki: Suomalainen Tiedeakatemia).
28. Bradburd GS, Ralph PL, Coop GM (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* 67(11):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(2):479–491.
30. Shennan S, Crema E, Kerig T (2015) Isolation-by-distance, homophily, and "core" vs. "package" cultural evolution models in neolithic europe. *Evolution and Human Behavior* (36(2)):103–109.
31. Huson D, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2):254–267.
32. Pearson K (1895) Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 58:240–242.
33. Sz?kely G, Rizzo M (2013) The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* 117:193 – 213.
34. Sz?kely G, Rizzo M (2013) Partial distance correlation with methods for dissimilarities. *ArXiv e-prints* p. arXiv:1310.2926v3.
35. Sz?kely G, Rizzo ML (2016) *Partial Distance Correlation*, eds. Cao R, González Manteiga W, Romo J. (Springer International Publishing, Cham), pp. 179–190.
36. Bottigheimer RB (2009) *Fairy tales : a new history.* (Excelsior Editions/State University of New York Press, Albany, N.Y.), pp. vii, 152 p.
37. Bottigheimer RB (2014) *Magic tales and fairy tale magic : from Ancient Egypt to the Italian Renaissance*, Palgrave Historical Studies in Witchcraft and Magic. pp. vii, 208 pages.
38. Thompson S (1977) *The folktale.* (University of California Press).
39. Propp VI (1968) *Morphology of the folktale*, Publications of the American Folklore Society Bibliographical and special series. (University of Texas Press, Austin,), 2d edition, pp. xxvi, 158 p.
40. Jaccard P (1901) Etude comparative de la distribution florale dans une portion des alpes et del jura. *Bulletin del la Societe Vaudoise des Sciences Naturelles* 37:547:579.
41. van Etten J (2014) *gdistance: distances and routes on geographical grids.* R package version 1.1-5.
42. Sz?kely G, Rizzo M, Bakirov N (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35(6):2769–2794.
43. Brown P (2016) *mapmisc: Utilities for Producing Maps.* R package version 1.5.0.
44. Pebesma E, Bivand R (2005) Classes and methods for spatial data in r. *R News* 5(2).
45. Bivand R, Pebesma E, G?mez-Rubio V (2013) *Applied Spatial Data Analysis with R, Second edition.* (Springer, NY.).
46. Nychka D, Furrer R, Paige J, Sain S (2016) *fields: Tools for Spatial Data.* R package version 8.3-6.
47. Paradis E (2010) pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics* 26:419–420.
48. Rizzo ML, Szekely GJ (2016) *energy: E-Statistics: Multivariate Inference via the Energy of Data.* R package version 1.7-0.

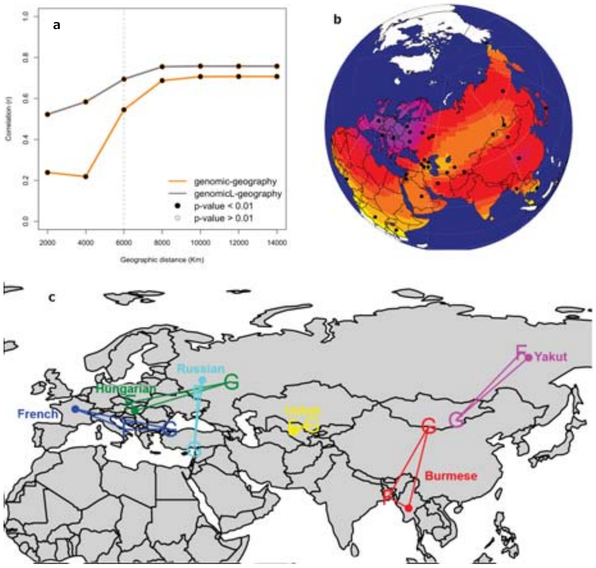


Fig. 1. a) Plot of product-moment correlation values between pairwise genetic distance (both whole genome and biased for linguistic barriers) and pairwise geographic distance over cumulative geographic distance; b) Map showing the spatial distribution of 33 populations comprised in DatasetMAIN. Surface colors represent interpolated richness values (i.e. the number of folktales exhibited by each population). Purple indicates higher values, while yellow indicates lower numbers; c) Example of 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), while the inferred source and direction of possible genetic admixture (G) is mismatched. For example, Burmese and Yakut exhibit quite segregated folktale assemblages, while their putative source of genetic admixture is closer in space. The case of Hungarian is emblematic for its folkloric assemblage rooted in Europe while its putative genetic (and linguistic) source of admixture is located in the Ural region.

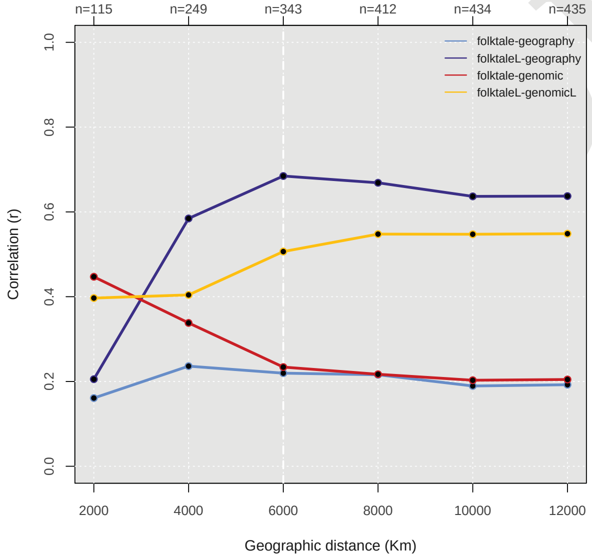


Fig. 2. Comparison of null model of cultural diffusion dictated by IBD (folktale∼geographic; light blue) against all alternative models: demic diffusion (folktale∼genetic; red), language-biased cultural diffusion (folktaleL∼geographic; purple), and language-biased demic diffusion (folktaleL∼geneticL; yellow) over cumulative geographic distance. Product-moment correlation coefficients are calculated at each geographic bin (size=2000 km) with original distance matrices up to 12000 km.

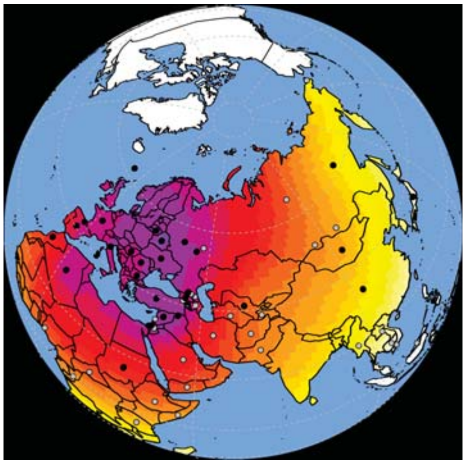
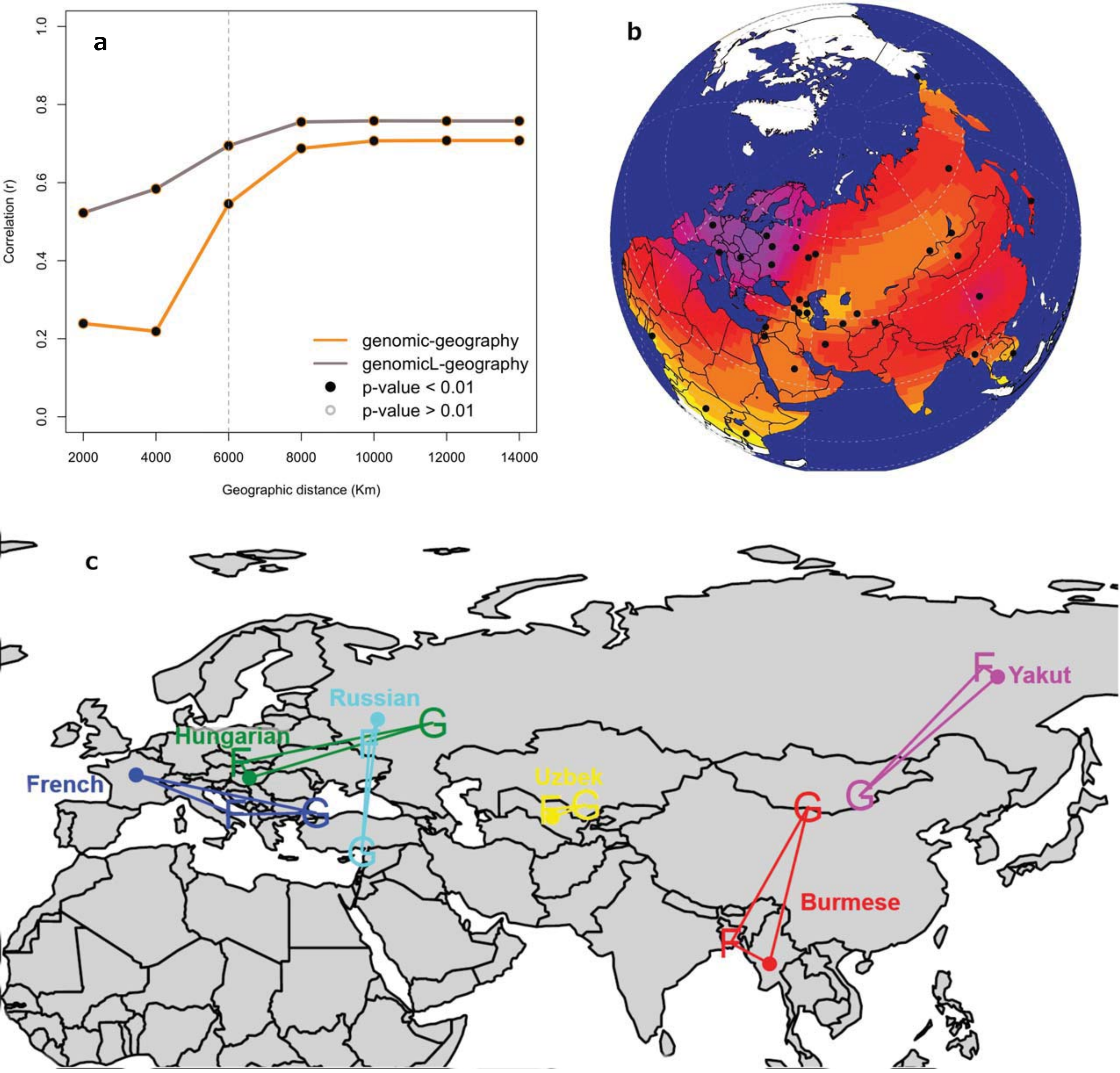


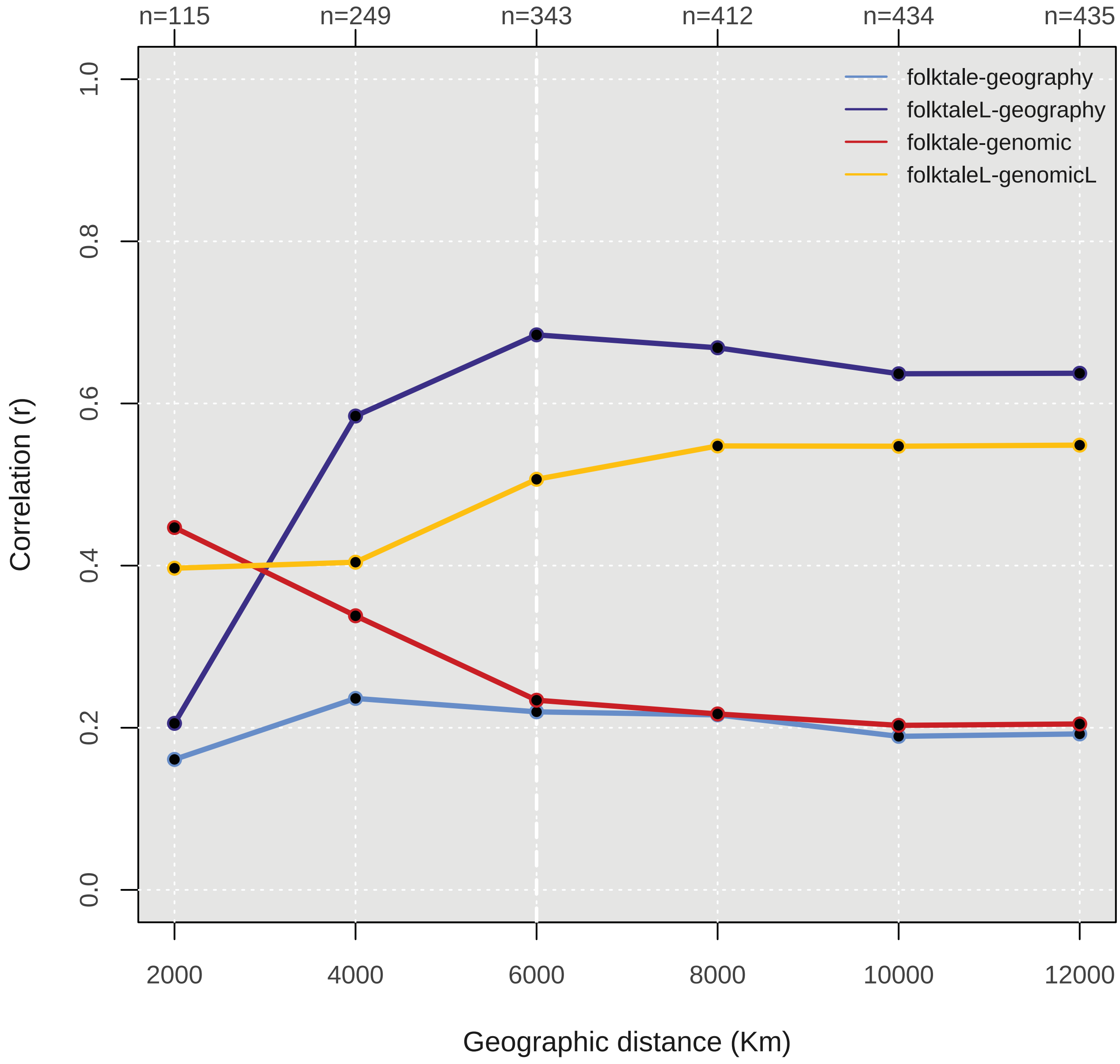
Fig. 3. Possible focal area and dispersion pattern for tale ATU313 "The Magic Flight", one the most popular folktales in the present dataset which may have been additionally spread through population movement and replacement. It is interesting to note how this tale reached locations that are far from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

Table 1. Variable association at a global level

	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
			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

Upper table: 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. Lower table: Results of partial distance correlation for null and alternative models, after Bonferroni correction.


a
1.0
0.8
0.6
0.4
0.2
0.0
Correlation (r)
2000
4000
6000
8000
10000
12000
14000
Geographic distance (Km)
genomic-geography
genomicL-geography
p-value < 0.01
p-value > 0.01
b
c
Russian
Hungarian
French
Uzbek
Burmese
Yakut
F
G


n=115
n=249
n=343
n=412
n=434
n=435
folktale-geography
folktaleL-geography
folktale-genomic
folktaleL-genomicL
Correlation (r)
0.0
0.2
0.4
0.6
0.8
1.0
2000
4000
6000
8000
10000
12000
Geographic distance (Km)