MOLECULAR BIOLOGY & GENETICS

Research Article

Reconciling the father tongue and mother tongue hypotheses in Indo-European populations

Menghan Zhang[1†], Hong-Xiang Zheng[1†], Shi Yan[1*], Li Jin[2, 3*]

[1] Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, and Human Phenome Institute, Fudan University, Shanghai, 200438, China

[2] State Key Laboratory of Genetic Engineering, and Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai, 200438, China

[3] Chinese Academy of Sciences Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, SIBS, CAS, Shanghai, 200031, China

† These authors contributed equally to this work.

*Correspondence and requests for materials should be addressed to Shi Yan (yanshi@fudan.edu.cn) and Li Jin (lijin@fudan.edu.cn).

Abstract: In opposite to the Mother Tongue Hypothesis, the Father Tongue Hypothesis states that humans tend to speak their fathers' language, based on a stronger correlation of languages to paternal lineages (Y-chromosome) than to maternal lineages (mitochondria). To reassess these two competing hypotheses, we conducted a genetic-linguistic study of 34 modern Indo-European (IE) populations. In this study, genetic histories of paternal and maternal migrations in these IE populations were elucidated using phylogenetic networks of Y-chromosomal and mitochondrial DNA haplogroups, respectively. Unlike previous studies, we quantitatively characterized the languages based on lexical and phonemic systems, separately. We showed that genetic and linguistic distances are significantly correlated with each other and that both are correlated with geographic distances among these populations. However, when controlling for geographic factors, only the correlation between the distances of paternal and lexical characteristics and between those of maternal and phonemic remained. These unbalanced correlations reconciled the two seemingly conflicting hypotheses.

Keywords: Indo-European populations; Y-Chromosomal haplogroup; mitochondrial DNA haplogroup; Lexical system; Phonemic system

Received: 30-Apr-2018; Revised: 01-Aug-2018; Accepted: 06-Aug-2018

The hypothesis that the language usage follows matrilineal inheritance has been supported by genetic evidence as in the Austronesian-speaking populations and South American Indians [1, 2]. This is called as the Mother Tongue Hypothesis *sensu stricto*. In contrast, on the basis of other findings from genetic and anthropological research [3-9], population geneticists and anthropologists advocate the Father Tongue Hypothesis, which cites that a strong correlation between languages and Y-chromosomes. A global picture of sex-specific transmission of language change, at the population level, has been described by Forster and Renfrew [10]. They summarized that the paternal lines dominate the survivor language in an already-populated region, whereas the maternal lines reflect only the ancient settlement. Therefore, the Father Tongue Hypothesis seems to prevail over the Mother Tongue Hypothesis. However, the controversy between these two hypotheses for IE populations suggests that Y-chromosomal composition in paternal lines may be essential predictor of language but not the only one [10].

In addition, quantified language affiliations such as designation of language families and subgroups [5] and divergence times deduced from the tree [7] have been used to measure linguistic difference in such studies. However, these two types of data, which can be extracted from linguistic documents, have been argued to be coarse estimations of language differences [11]. Such data provide only holistic evolutionary hints of languages without fully considering linguistic compositions, including lexical and phonemic systems, which may portray distinct evolutionary processes. The evolution of lexical systems, such as loss or gain of core vocabulary, can trace language divergence [12]. In comparison, the evolution of phonemic systems is more complicated. Phonemes can change not only the diachronically but also synchronically, such as contact-induced (i.e. phoneme borrowings [13]) or spontaneous evolution (i.e. Great Vowel Shift [14]) . However, some researchers suggest that in contrast to lexical systems, phonemic systems could be more conservative and provide earlier insights into the evolution of languages [15, 16].

Here, we reassessed the correlation between genetic and linguistic characteristics in 34 modern IE populations (Fig. 1a), for which all four types of datasets (lexicon, phonemes, Y-chromosomal composition, and mitochondrial DNA (mtDNA) composition) are available. We assembled compositions of the Y-chromosomal and mtDNA haplogroups or paragroups from the corresponding IE populations, which reflect paternal and maternal lines, respectively (see Supplementary S1.1 and Fig. 1b). These haplogroup or paragroups were defined using stable mutations so that they are all formed already in the Paleolithic Age (over 10,000 year) [17, 18]. For example, the categorization of lineages was not changed during the evolutionary processes of Indo-European languages, therefore representing the mixing process of the ancestral populations. Instead of the formerly used linguistic classification or coalescence time, we utilized two types of linguistic data representing distinct evolutionary processes of language systems (see Supplementary S1.2). The first type was the lexicon of IE languages came from the publicly available Dunn's lexical dataset [19]. The other was phonemic data from PHOIBLE database [20] that contain segment types corresponding to the sound system of the IE language. Although genetic and linguistic characteristics all reflect the ethno-genetic history of IE population divergence and interactions, they portray different evolutionary processes.

Neighbour-Nets were constructed to delineate the differences between 34 IE population groups clustering at the genetic and linguistic levels (Fig. 2). The reticulations within each net reflect conflicting signals against tree-like structures and support incompatible groupings [21]. These structures are likely produced by potential horizontal transmissions between populations or languages such as admixture, and potential parallel evolution in linguistics as well [22]. The Neighbour-Net for Y-chromosomes with substantial reticulations shows complicated relationships among IE populations (Fig. 2a), indicating a substantial historical population contact and admixture among the males. In contrast, the Neighbour-Net for mtDNA in Fig. 2b clearly illustrates an East-West geographic polarization, indicating two major IE populations in matrilineages: Indo-Iranian and European. Due to the limited lexical borrowings in the Dunn's lexical dataset [12], the Neighbour-Net for lexicon thus appeared to better approximate a tree-like structure with fewer reticulations than the phonemic Neighbour-Net. The clustering groups for languages based on lexicon were consistent with traditional linguistic classifications. In contrast, the Neighbour-Net for sound systems showed evidence of a substantial conflicting signal between phonemic characteristics. The network did not accurately recover many attested phylogenetic relationships among IE languages. None of the language groups were monophyletic at phonemic level.

To investigate the relationships between genetic and linguistic characteristics, we performed the Mantel test on the pairwise genetic and linguistic distance matrices of 34 IE populations. Fig. 3a clearly shows that the genetic and linguistic characteristics were strongly correlated with each other. However, these correlations have been argued to be false signals because all these variables could be dependent on geography [23]. In 34 IE languages, all the genetic and linguistic distances indeed had significantly positive relationships with the geographic distances for these Indo-European populations (see Supplementary S2.1).

To exclude the geographic effects, we then adopted the partial Mantel test to re-appraise the relationships between genetics and linguistics in these populations (Fig. 3b). When controlling for the effect of geographic distance of pairwise IE populations, there was no significant correlation between Y-chromosomal and mtDNA distance matrices. It indicated that paternal and maternal lineages had different ethnic histories in IE populations. Similarly, lexical and phonemic systems of IE languages experienced different evolutionary processes because of no correlation between lexical and phonemic distances. In particular, the correlations between the Y-chromosomal and phonemic distance matrices, as well as those between the mtDNA and lexical matrices, were no longer significant. This result therefore suggests that both Y-chromosome–phoneme and mtDNA–lexicon relationships between the IE samples could be sufficiently predicted by their geographic distance. However, the correlation between Y-chromosomal and lexical distances remained significant (partial Mantel r = 0.2042, *p*-value$<10^{-3}$), as did the correlation between mtDNA and phonemic distances (r = 0.4273, *p*-value$<10^{-3}$). In addition, we performed alternative two partial statistical tests to validate the reliability of these observations (see Methods). The results of three partial statistical tests were consistent with each other (Table S1). Such observations of unbalanced correlations, after removing the effect of geography, suggest that the change in lexicon reflects the differences in paternal lines, while phonemic dissimilarity reflects the differences in maternal lines. Moreover, we adopted an alternative lexical dataset provided by Bouckaert *et al.* [24] to validate the statistical results of Mantel and partial Mantel tests, especially for the correlation between Y-chromosomes and lexicon (see Supplementary S1.2 and S2.2). The results obtained from this lexical dataset were consistent with those for Dunn's dataset. In addition, Jackknife resampling approach was used to evaluate the robustness of the correlation between genetics and linguistics (see Supplementary S2.3 and Table S2-S3).

These observations of unbalanced correlation between genetics and linguistics could be explained by population contact and admixture at first. If there is no contact and admixture between the populations or languages, the phylogenies of genetics and linguistics should ideally follow tree-like structures and resemble each other. However, population contacts have long been known to change local population structures and language systems. The causes of such population contacts include marriage between neighbouring populations or between local people and immigrants, such as military conquerors or merchants. Especially, the different performances of female and male dispersal have been confirmed that female lives more locally than male [25-28] (See Supplemantary S2.4). In other words, the immigrants tend to be highly sex-biased with a higher concentration of males [10, 29]. This could be also why we found no significant correlation between paternal and maternal in IE populations, under controlling the geographic effects. When immigration is associated with social prestige such as colonists, the immigrants form a new community that speaks the languages brought with them, while their spouses (usually women) are from the local region. Therefore, the social prestige of male immigrants could reasonably lead to the correlation between the Y-chromosome and languages [30].

The language learning by local women could constitute the reason for unbalanced correlation of mtDNA to lexicon and phonemes. Due to the social prestige of male immigrants, their local spouses have to adopt the language of their husbands and pass it to future generations [6, 10, 15]. This process is second language acquisition and easily develops language fossilization [31]. The language fossilization is a linguistic mechanism that a learner of a second language tends to preserve some linguistic features of the first language, and develops a form of inter-language [31]. Under this circumstance, women can easily replace the lexicon from another [21], but attempt to retain local accents influenced by their native language [32]. In other words, women change to adopt the same word usage as their husbands in daily life but still speak using their own pronunciation. In mixed-language marriages with these male immigrants, women prefer to pass down their inter-languages to offspring [10, 33]. As a result, it yields the correlation between mtDNA and phonemes we observed. Hence, we courageously proposed a hypothetical scenario in Indo-European populations that lexical system of language dominated by their father, while the phonemic system of language determined by their mother.

The co-evolution between genes and languages is asymmetrical in Indo-European populations. Our findings provide strong statistical evidence to reconcile the conflicting Father Tongue and Mother Tongue hypotheses. The population involved in this study are located within a single continent, and all of them speak languages belonging to Indo-European language family. Much of the genetic patterns hence may have its roots in the spread of IE language. Further cross-continental comparison between genetic and linguistic data would provide us more remarkable co-evolutionary processes of population and language. Notably, what we observed from the correlation between linguistics and genetics is macroscopic. The scenario that the mother learns her husband's language and teaches the children is definitely one possible mechanism, which has been elaborated by historical linguist van Driem [30]. In the future, more detailed explorations are warranted into the mechanisms of language change at the micro level, including infant's language acquisition and development from the father and mother, even other social structure. Moreover, the present research paradigm can be extended to other human cultural and social traits [34-36]. On basis of interdisciplinary approaches, there is a significant importance and challenge for us to re-examine several general hypotheses of population and cultural evolution at the global scope.

Methods:

Distance matrices and Neighbour-Net

To delineate the relationships between 34 Indo-European populations and their languages, we applied the Neighbour-Net method [37, 38] to the four datasets of genetic and linguistic properties, respectively. The genetic Neighbour-Nets were calculated from distance matrices on haplogroup frequencies using the Euclidean distance method. According to the linguistic distance matrices used in Creanza et al. [13], we applied Hamming distance matrices [39] on comparing the presence/absence of traits (lexicons and phonemes). Notably, for Bouckaert dataset, each hamming distance of pair-wise languages was calculated ignoring all missing cognate sets in pair-wise languages compared. The linguistic Neighbour-Nets were established with Hamming distance matrices from lexical and phonemic data. In addition, we applied the Orthodromic distance (great circle distance) of two locations for the metric of geographic distance, and transformed the distance (d) into the logarithmic scale following the formula log10(d). The hamming distance for Bouckaert dataset and geographic distance calculation was implemented in Matlab. All network analyses were performed in SplitsTree4 (http://www.splitstree.org/) using default settings.

Mantel test and partial Mantel test

In this paper, we used the Mantel test to detect the relationships between languages and genes, and the partial Mantel test to further study the correlation between languages and genes controlled with geographic effects. All statistical tests were implemented in Matlab® R2015b (MathWorks, Inc.). The Matlab scripts for the

Mantel test and partial Mantel test were provided by Prunier et al. [40] (URL: http://www.jeromeprunier.eg2.fr/5.html).

To validate the credibility of the statistical results, we adopted two alternative partial correlation tests. The first was the linear Pearson's correlation test [41] implemented in Matlab® as the function partialcorr. The other was a modified partial Mantel test that was developed by Smouse et al. [42] to examine the Mantel correlation between two residuals from linear regressions of genes/languages on geographic distance metrics, respectively. Specifically, we designated the three matrices to be compared as A, B and C. The users tested the significance of partial correlation by computing residual matrices from the regressions of A on C and B on C, and then carried out a Mantel test between the two residual matrices with the permutation approach. In this process, we performed the Matlab script of the Mantel test programed by Enrico Glerean (http://becs.aalto.fi/~eglerean/permutations.html). The numbers of permutations in all Mantel or partial Mantel tests were set at 10,000 in this study.

Principal Component Analysis and Procrustes Analysis

We here conducted a series of Principal Component Analyses [43] (PCA) to identify the principal coordinates of each high-dimensional linguistic or genetic data of IE populations. Then, we performed Procrustes Analysis of each genetic and linguistic PCs versus the geographic coordinates of these IE populations. The rationale of Procrustes analysis [44, 45] is to find an optimal transformation of two or more maps that maximize the similarity of the transformed maps, and to score the similarity between two optimally transformed maps. In this study, two maps in comparison are the two-dimensional plot of the first two PCs and the geographic map of latitudes and longitudes of 34 IE populations. A permutation test [46, 47] then can measure the probability that a randomly chosen permutation of the points in any one map produces a greater similarity score than that observed for the actual points in the other map.

Following Wang et al. [48], we calculated a similarity score on the statistic

$t_0 = \sqrt{(1-D)}$, where D is the minimized sum of squared distances in Procrustes Analysis. We then calculated empirical p-value for t0 values over 105 permutations of geographic locations. All computational procedures of PCA, Procrustes Analysis and permutation tests were implemented in Matlab® R2015b (MathWorks, Inc.).

Jackknife resampling method

We performed jackknife resampling approach to evaluate the robustness of the statistical conclusions based on partial Mantel test. In this study, we considered the balance of the samples sizes between Indo-Iranian and European populations, and designed two schemes of Jackknife resampling approach [49-51]:

Scheme I: We sampled all the available Indo-Iranian populations from the dataset, and randomly selected equal amount of populations from the total European populations.

Scheme II: We randomly selected the same number of population samples from the total IE populations in order to compare to the resampling in Scheme I.

Accordingly, we resampled 22 IE populations (11 Indo-Iranian + 11 European for Scheme I, and randomly 22 out of 34 in Scheme II) for Dunn's dataset, and 18 (9 + 9 for Scheme I, and 18/32 for Scheme II) for a new lexical dataset of 207 words by Bouckaert et al. For each resampling scheme, the random selection was repeated for 500 times and thus 500 Jackknife resampled data sets of selected population sample were generated. For each dataset, we re-conducted partial Mantel tests to examine the correlation between these genetic and linguistic data controlling for geographic effects (Y-chromosome & Lexicon; Y-chromosome & Phoneme; mtDNA & Lexicon; mtDNA & Phoneme). The correlation coefficients and p-values were re-calculated. For the correlation coefficients obtained via Jackknife method, we listed the statistic descriptions including median, min, max and 95% confidence intervals in Table S2. And for the distribution of p-values, we calculated quantiles (0.25, 0.50 and 0.75) and counted the number of p-values less than 0.05 or 0.01. We counted the occurrence of p-value<0.05 and <0.01 out of Jackknife 500 replicates to measure the robustness.

Notably, the occurrence was a relative value to compare the results of different partial Mantel tests.

Data availability

All linguistic and genetic data that support the findings of this study are available within the paper and its supplementary information files.

References:

[1] N. J. Fagundes, S. L. Bonatto, S. M. Callegari Jacques *et al.*, Genetic, geographic, and linguistic variation among South American Indians: possible sex influence. *Am J Phys Anthropol.*2002; 117: 68-78.

[2] J. K. Lum, R. L. Cann, J. J. Martinson, and L. B. Jorde, Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. *Am J Hum Genet.* 1998; 63: 613-624.

[3] G. Chaubey, M. Metspalu, C. Ying, R. Mägi, I. G. Romero, P. Soares*, et al.* Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. *Mol Biol Evol*. 2011; **28**: 1013-24.

[4] M. Kayser, Y. Choi, O. M. Van, S. Mona*, et al.*, The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. *Mol Biol Evol 2008*; **25**: 1362-1374.

[5] B. M. Kemp, A. González-Oliver, R. S. Malhi, *et al.*, Evaluating the Farming/Language Dispersal Hypothesis with genetic variation exhibited by populations in the Southwest and Mesoamerica. *Proc Natl Acad Sci U S A.* 2010; 107: 6759-64.

[6] E. S. Poloni, O. Semino, G. Passarino, *et al.*.Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet*. 1997; 61: 1015-35.

[7] R. R. Sokal, Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci U S A*. 1988; 85: 1722-6.

[8] M. Stoneking and F. Delfin. The human genetic history of East Asia: weaving a complex tapestry. *Curr Biol*. 2010; 20: 188-193.

[9] E. T. Wood, D. A. Stover, C. Ehret, *et al.*. Contrasting patterns of Y chromosome and mtDNA variation in Africa: Evidence for sex-biased demographic processes. *Eur J Hum Genet.* 2005; 13: 867-76.

[10] P. Forster and C. Renfrew. Evolution. Mother tongue and Y chromosomes. *Science.* 2011; 333: 1390-1.

[11] A. McMahon and R. McMahon, *Language classification by numbers*: Oxford University Press on Demand, 2005.

[12] S. Nelson-Sathi, J.-M. List, H. Geisler *et al.*. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proc Biol Sci.* 2011; 278: 1794-1803.

[13] N. Creanza, M. Ruhlen, T. J. Pemberton, *et al.*. A comparison of worldwide phonemic and genetic variation in human populations. *Proc Natl Acad Sci U S A.* 2015; 112: 1265-1272.

[14] W. Labov, M. Yaeger, and R. Steiner, *A quantitative study of sound change in progress* vol. 1: US Regional Survey, 1972.

[15] S. G. Thomason, *Language contact*: Edinburgh University Press Edinburgh, 2001.

[16] M. Dunn, A. Terrill, G. Reesink, R. A. Foley, and S. C. Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science.* 2005: 309: 2072.

[17] G. D. Poznik, Y. Xue, F. L. Mendez*, et al.*. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet*. 2016; 48: 593.

[18] Q. Fu, A. Mittnik, P. L. F. Johnson*, et al.*. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol.* 2013; 23: 553-9.

[19] M. Dunn, S. J. Greenhill, S. C. Levinson, and R. D. Gray. Evolved structure of language shows

lineage-specific trends in word-order universals. *Nature.* 2011; 473: 79-82.

[20] S. Moran, D. McCloy, and R. Wright. PHOIBLE online. *Leipzig: Max Planck Institute for Evolutionary Anthropology,* 2014.

[21] S. J. Greenhill, Q. D. Atkinson, A. Meade, and R. D. Gray. The shape and tempo of language evolution. *Proc Biol Sci.* 2010; 277: 2443-50.

[22] T. Warnow, S. N. Evans, D. Ringe, and L. Nakhleh. A stochastic model of language evolution that incorporates homoplasy and borrowing. *Phylogenetic methods and the prehistory of languages,* pp. 75-90, 2006.

[23] Z. H. Rosser, T. Zerjal, M. E. Hurles*, et al.*. Y-Chromosomal Diversity in Europe Is Clinal and Influenced Primarily by Geography, Rather than by Language. *Am J Hum Genet.* 2000; 67: 1526-43.

[24] R. Bouckaert, P. Lemey, M. Dunn, *et al.*. Mapping the origins and expansion of the Indo-European language family. *Science.* 2012; 337: 957-960.

[25] B. Hewlett, J. M. H. V. D. Koppel, and L. L. Cavalli-Sforza. Exploration Ranges of Aka Pygmies of the Central African Republic. *Man.* 1982; 17: 418-430.

[26] I. Nasidze, E. Y. S. Ling, D. Quinque, *et al.*. Mitochondrial DNA and Y-Chromosome Variation in the Caucasus. *Ann Hum Genet.* 2004; 68: 205-21.

[27] B. Wen, H. Li, D. Lu*, et al.*. Genetic evidence supports demic diffusion of Han culture. *Nature.* 2004; 431: 302-305.

[28] S. Lippold, H. Xu, A. Ko*, et al.*. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig Genet.* 2014; 5: 13.

[29] N. Marchi, T. Hegay, P. Mennecier, *et al.*. Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *Am J Phys Anthropol.* 2017; 162: 627-640.

[30] G. Van Driem. Etyma, shouldered adzes and molecular variants. *Methods in Contemporary Linguistics,* 2012.

[31] L. Selinker. Interlanguage. *IRAL - International Review of Applied Linguistics in Language Teaching,* 1972; 10: 209-232.

[32] P. Avery and S. Ehrlich, *Teaching American English Pronunciation: A Textbook and Reference Manual on Teaching the Pronunciation of North American English, Written Specifically for Teachers of English as a Second Language (ESL)*: OUP Oxford, 1992.

[33] C. Renfrew and M. Jones, *Traces of ancestry: studies in honour of Colin Renfrew*: McDonald Inst of Archeological, 2004.

[34] P. E. Savage, S. Brown, E. Sakai, and T. E. Currie. Statistical universals reveal the structures and functions of human music. *Proc Natl Acad Sci U S A.* 2015; 112: 8987-92.

[35] T. E. Currie, S. J. Greenhill, R. D. Gray, *et al*.. Rise and fall of political complexity in island South-East Asia and the Pacific. *Nature.* 2010; 467: 801-804.

[36] J. Watts, O. Sheehan, Q. D. Atkinson, *et al.*. Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature.* 2016; 532: 228-231.

[37] D. Bryant and V. Moulton. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol.* 2004; 21: 255-265.

[38] D. H. Huson and D. Bryant. Application of Phylogenetic Networks in Evolutionary Studies. *Mol Biol Evol.* 2006; 23: 254-267.

[39] R. W. Hamming. Error Detecting and Error Correcting Codes. *Bell Labs Technical Journal.* 1950; 29: 147-160.

[40] J. G. Prunier, B. Kaufmann, S. Fenet*, et al.*. Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: towards a better assessment of functional connectivity using an individual-based sampling scheme. *Mol Ecol*. 2013; 22: 5516-30.

[41] R. A. Fisher. The distribution of the partial correlation coefficient. *Metron.* 1924; 3: 329-332.

[42] P. E. Smouse, J. C. Long, and R. R. Sokal. Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. *Syst Zoo.* 1986; 35: 627-632,.

[43] Jolliffe and Ian. Principal Component Analysis. *Springer Berlin*. 1986; 87: 41-64.

[44] C. Goodall. Procrustes Methods in the Statistical Analysis of Shape. *Journal of the Royal Statistical Society*. 1991; 53: 285-339.

[45] J. C. Gower. Generalized procrustes analysis. *Psychometrika.* 1975; 40: 33-51.

[46] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*: Springer, 2011.

[47] D. A. Jackson. PROTEST: A PROcrustean Randomization TEST of community environment concordance. *Écoscience*. 1995; 2: 297-303.

[48] C. Wang, Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton, J. A. Hardy*, et al.*. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Mol Biol.* 2010; 9: Article 13.

[49] M. H. Quenouille. Problems in Plane Sampling. *Annals of Mathematical Statistics.* 1949; 20: 355-375.

[50] M. H. Quenouille. Notes on Bias in Estimation. *Biometrika.* 1956; 43: 353-360.

[51] J. W. Tukey. Bias and Confidence in Not Quite Large Samples. *Annals of Mathematical Statistics*. 1958: 29: 614.

Acknowledgements: This work was supported by projects at the National Natural Science Foundation of China (31501010, 31401060, 31401062, 31521003 and 91731000), the Special Program for Key Basic Research of the Ministry of Science and Technology, China (2015FY111700), the Postdoctoral Science Foundation of China (2015M570316, 2015T80394 and 2016M601496), and the Science and Technology Committee of Shanghai Municipality (16JC1400500). We thank Prof. Wuyun Pan at Fudan University for his comments. We also thank Nature Research Editing Service for its linguistic assistance during the preparation of this manuscript.

Author contributions: M.Z. and H.Z contributed equally. M.Z., H.Z., S.Y. and L.J. designed the research; M.Z., H.Z, S.Y. performed the research; M.Z., H.Z., S.Y. and L.J. analyzed the results; and M.Z., H.Z., S.Y. and L.J. wrote the paper. The authors declare no conflicts of interest.

Competing financial interests: The authors declare no competing financial interests.

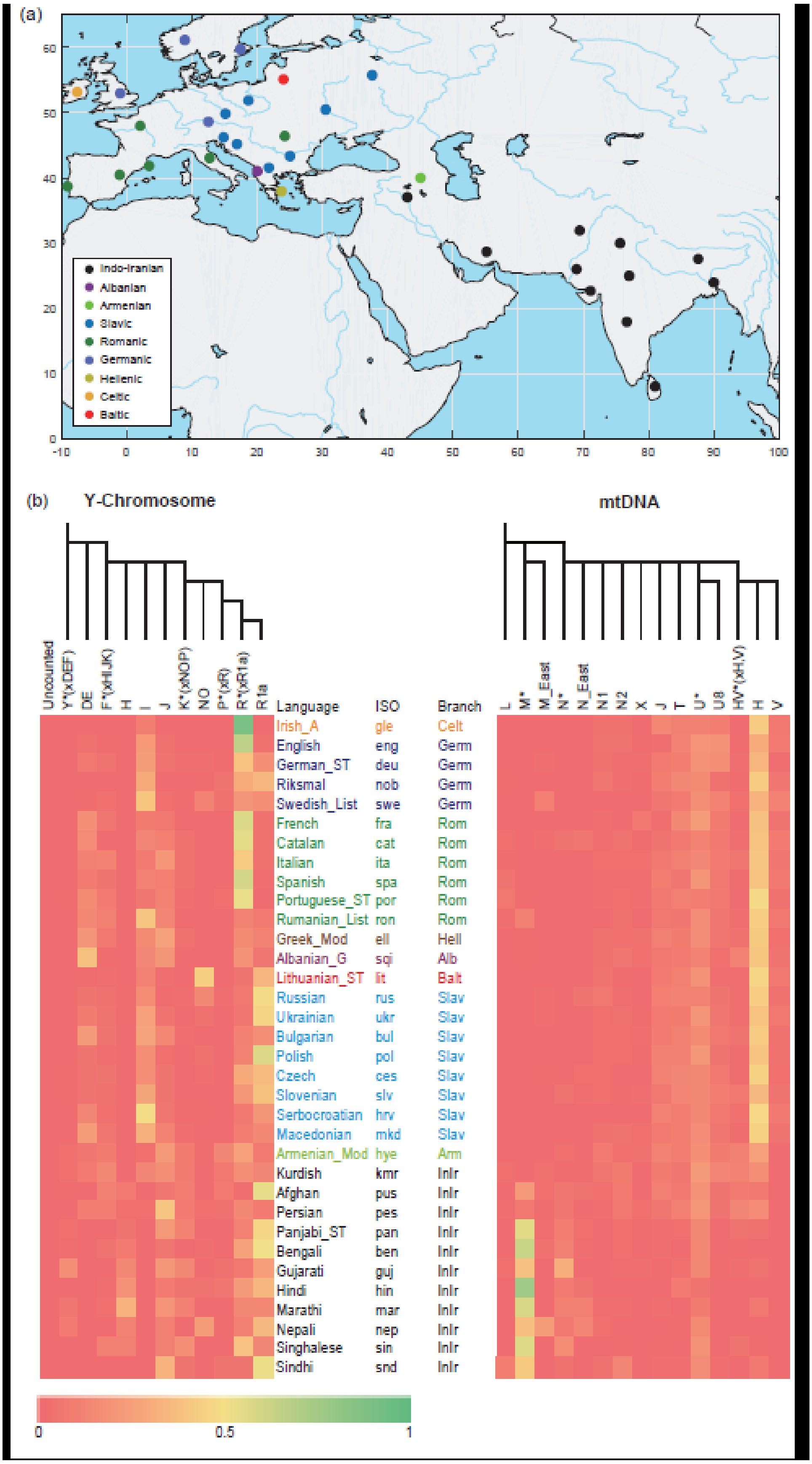
(a)
60
50
40
30
20
10
0
-10
0
10
20
30
40
50
60
70
80
90
100
Indo-Iranian
Albanian
Armenian
Slavic
Romanic
Germanic
Hellenic
Celtic
Baltic
(b)
Y-Chromosome
mtDNA
Uncounted
Y*(xDEF)
DE
F*(xHIJK)
H
I
J
K*(xNOP)
NO
P*(xR)
R*(xR1a)
R1a
Language
ISO
Branch
L
M*
M_East
N*
N_East
N1
N2
X
J
T
U*
U8
HV*(xHV)
H
V
Irish_A gle Celt
English eng Germ
German_ST deu Germ
Riksmal nob Germ
Swedish_List swe Germ
French fra Rom
Catalan cat Rom
Italian ita Rom
Spanish spa Rom
Portuguese_ST por Rom
Rumanian_List ron Rom
Greek_Mod ell Hell
Albanian_G sqi Alb
Lithuanian_ST lit Balt
Russian rus Slav
Ukrainian ukr Slav
Bulgarian bul Slav
Polish pol Slav
Czech ces Slav
Slovenian slv Slav
Serbocroatian hrv Slav
Macedonian mkd Slav
Armenian_Mod hye Arm
Kurdish kmr InIr
Afghan pus InIr
Persian pes InIr
Panjabi_ST pan InIr
Bengali ben InIr
Gujarati guj InIr
Hindi hin InIr
Marathi mar InIr
Nepali nep InIr
Singhalese sin InIr
Sindhi snd InIr
0
0.5
1

Fig. 1. (a) Geographic locations of 34 modern Indo-European populations, colored by language group. (b) The heat maps of Y-chromosomal and mtDNA haplogroup frequencies of 34 Indo-European populations, aligned with the population speaking each language.

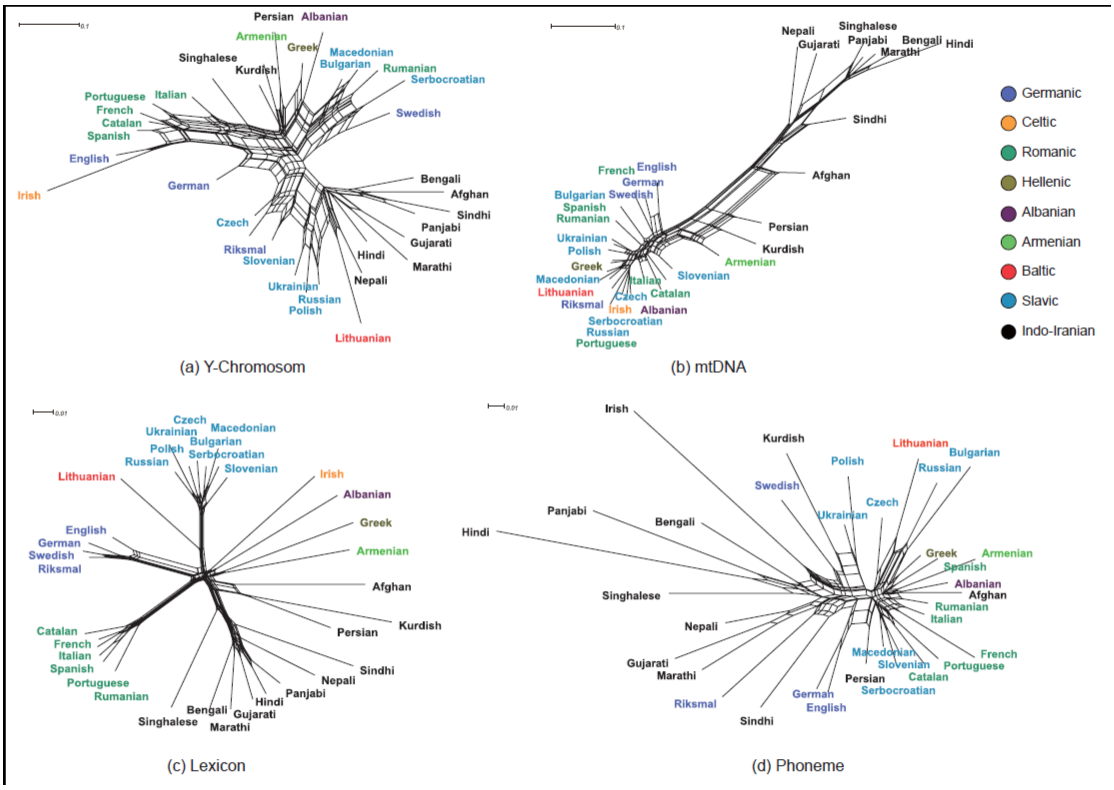

Fig. 2 Neighbour-Nets of 34 Indo-European populations calculated from the Euclidean distance matrices using (a) Y-chromosomal haplogroups and (b) mtDNA haplogroups; Neighbour-Nets of IE languages calculated from the Hamming distance matrices using (c) lexicon and (d) phonemes. The colours in the legend correspond to the language groups.

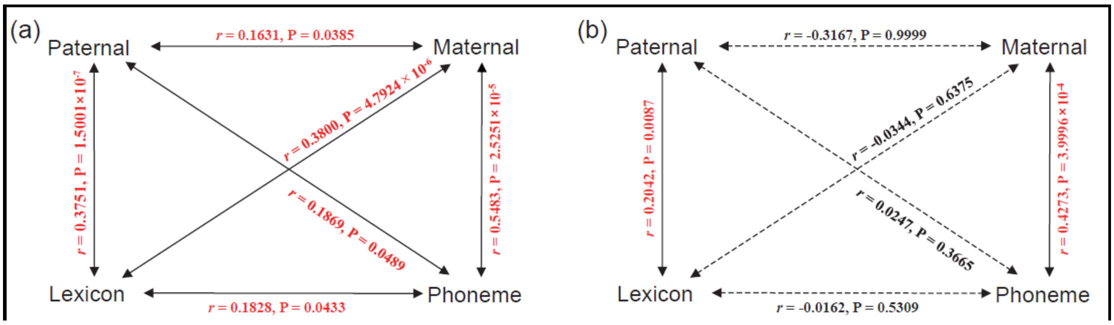

Fig. 3 Mantel correlations between four distance matrices for Y-chromosome, mtDNA, phoneme and lexicon. (a) Mantel correlations; (b) Partial Mantel correlations when controlling for geographic effects. The number of permutations of the Mantel test was set at 10,000. The red text shows significant Mantel correlations. Solid lines represent a *p*-value < 0.05. Dashed lines represent no significance, *p*-value > 0.05.