A comparison of worldwide phonemic and genetic variation in human populations

Nicole Creanza[a], Merritt Ruhlen[b], Trevor J. Pemberton[c], Noah A. Rosenberg[a], Marcus W. Feldman[a,1], and Sohini Ramachandran[d,e,1]

[a]Department of Biology and [b]Department of Anthropology, Stanford University, Stanford, CA 94305; [c]Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada R3E 0J9; and [d]Department of Ecology and Evolutionary Biology and [e]Center for Computational Molecular Biology, Brown University, Providence, RI 02912

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2013.

Contributed by Marcus W. Feldman, December 17, 2014 (sent for review July 16, 2014; reviewed by Quentin D. Atkinson and Keith Hunley)

Worldwide patterns of genetic variation are driven by human demographic history. Here, we test whether this demographic history has left similar signatures on phonemes—sound units that distinguish meaning between words in languages—to those it has left on genes. We analyze, jointly and in parallel, phoneme inventories from 2,082 worldwide languages and microsatellite polymorphisms from 246 worldwide populations. On a global scale, both genetic distance and phonemic distance between populations are significantly correlated with geographic distance. Geographically close language pairs share significantly more phonemes than distant language pairs, whether or not the languages are closely related. The regional geographic axes of greatest phonemic differentiation correspond to axes of genetic differentiation, suggesting that there is a relationship between human dispersal and linguistic variation. However, the geographic distribution of phoneme inventory sizes does not follow the predictions of a serial founder effect during human expansion out of Africa. Furthermore, although geographically isolated populations lose genetic diversity via genetic drift, phonemes are not subject to drift in the same way: within a given geographic radius, languages that are relatively isolated exhibit more variance in number of phonemes than languages with many neighbors. This finding suggests that relatively isolated languages are more susceptible to phonemic change than languages with many neighbors. Within a language family, phoneme evolution along genetic, geographic, or cognate-based linguistic trees predicts similar ancestral phoneme states to those predicted from ancient sources. More genetic sampling could further elucidate the relative roles of vertical and horizontal transmission in phoneme evolution.

cultural evolution | human migration | languages | population genetics

Both languages and genes experience descent with modification, and both are affected by evolutionary processes such as migration, population divergence, and drift. Thus, although languages and genes are transmitted differently, combining linguistic and genetic analyses is a natural approach to studying human evolution (1, 2). Cavalli-Sforza et al. (3) juxtaposed a genetic phylogeny with linguistic phyla proposed by Greenberg (described in ref. 4) and observed qualitative concordance; however, their comparison of linguistic and genetic variation was not quantitative. A later analysis of genetic polymorphisms and language boundaries suggested a causal role for language in restricting gene flow in Europe (5). More recently, population-level genetic data have been compared with patterns expected from language family classifications (2, 6–12). Other studies addressed whether the serial founder effect model from genetics—human expansion from an origin in Africa, followed by serial contractions in effective population size during the peopling of the world (13, 14)—explains various linguistic patterns (15–19).

Past studies are generally asymmetrical in their approaches to the comparison of genes and languages: some focus on genetic analysis and use linguistics to interpret results, and others analyze linguistic data in light of genetic models. Our study directly compares the signatures of human demographic history in microsatellite polymorphisms from 246 worldwide populations (20) and complete sets of phonemes (phoneme inventories) for 2,082 languages; these are the largest available datasets of both genotyped populations and phonemes, the smallest units of sound that can distinguish meaning between words. Languages do not hold information about deep ancestry as genes do, and phoneme evolution is complex: phonemes can be transmitted vertically from parents to offspring or horizontally between speakers of different languages, and phonemes can change over time within a language (21–23). We compare the geographic and historical patterns evident in phonemes and genes to determine the traces of human history in each data type.

Phonemic data were compiled by M.R. (the Ruhlen database); for 2,082 languages with complete phoneme inventories and referenced sources in this database, we annotated each language with geographic coordinates (Fig. 1*A*) and the number of speakers reported (24). We also analyzed PHOIBLE (PHOnetics Information Base and Lexicon) (25), a linguistic database with phoneme inventories for 968 languages. For 139 globally distributed populations in the Ruhlen database (114 in PHOIBLE), we matched each population's genetic data to the phoneme inventory of its native language (20), producing novel "phoneme–genome datasets" that allow joint analysis of genes and languages.

Significance

Linguistic data are often combined with genetic data to frame inferences about human population history. However, little is known about whether human demographic history generates patterns in linguistic data that are similar to those found in genetic data at a global scale. Here, we analyze the largest available datasets of both phonemes and genotyped populations. Similar axes of human geographic differentiation can be inferred from genetic data and phoneme inventories; however, geographic isolation does not necessarily lead to the loss of phonemes. Our results show that migration within geographic regions shapes phoneme evolution, although human expansion out of Africa has not left a strong signature on phonemes.

Author contributions: M.R., M.W.F., and S.R. conceived of the study; N.C., M.W.F., and S.R. designed research; M.R. developed the Ruhlen database; N.C. and S.R. prepared and analyzed linguistic data; T.J.P. and N.A.R. prepared genetic data; N.C. and T.J.P. analyzed genetic data; N.C. merged linguistic data with the Ethnologue and with genetic data, and conducted phylogenetic analyses; N.C., N.A.R., M.W.F., and S.R. wrote the paper with input from all authors.

Reviewers: Q.D.A., University of Auckland; and K.H., University of New Mexico.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: Linguistic data from the Ruhlen database analyzed in this paper are available in Datasets S1–S3.

[1]To whom correspondence may be addressed. Email: mfeldman@stanford.edu and sramachandran@brown.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1424033112/-/DCSupplemental.

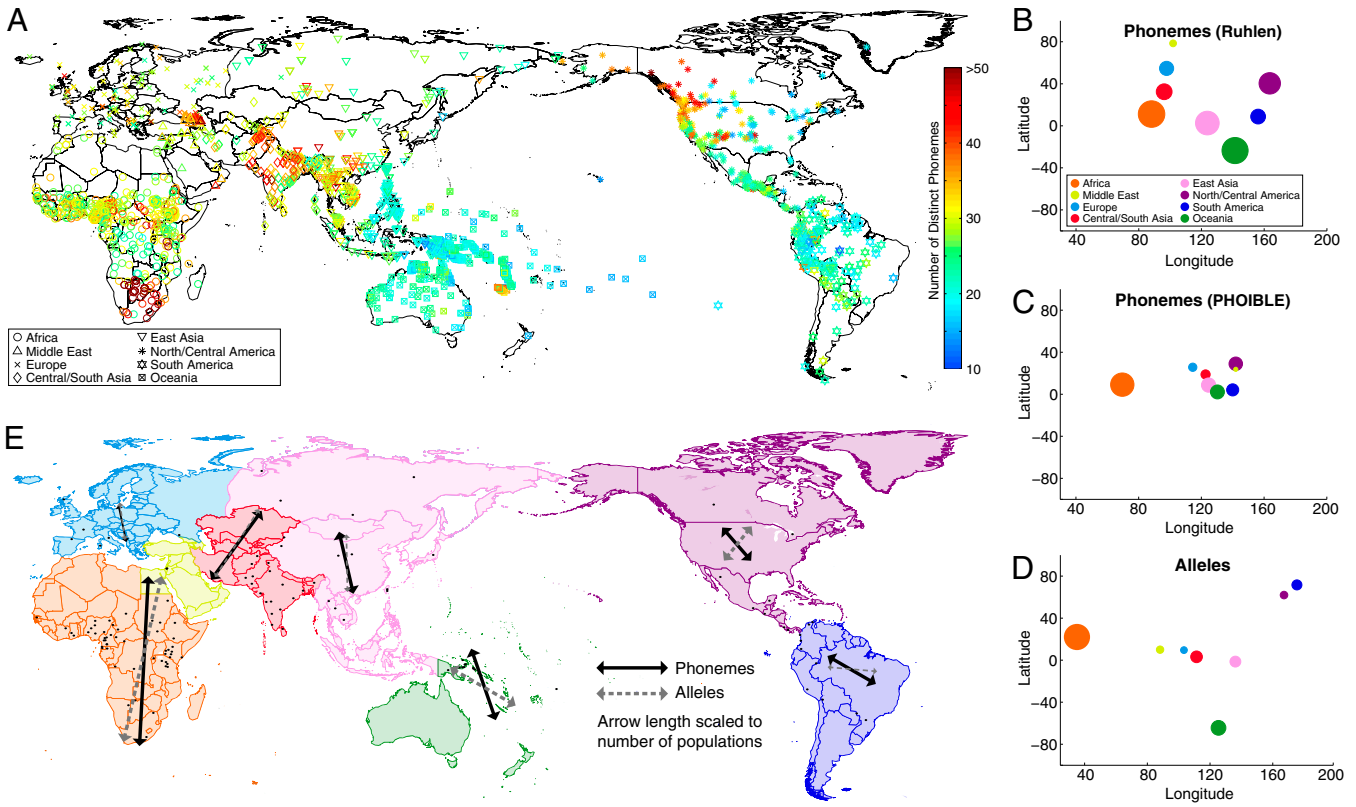

Fig. 1. Procrustes-transformed PCs for all phonemes and regional axes of phonemic and genetic differentiation. (*A*) Locations of 2,082 languages in the Ruhlen database. Phoneme inventory size of each language is indicated by the color bar. We performed Procrustes analyses to compare the first two PCs of phonemic data (*B* and *C*) and genetic data (*D*) to the geographic locations of languages/populations ($P < 10^{-5}$ for all three comparisons after 100,000 permutations). The mean Procrustes-transformed PC values (*B*) for phonemes in the Ruhlen database ($t_0 = 0.57$), (*C*) for phonemes in PHOIBLE ($t_0 = 0.52$), and (*D*) for allele frequencies ($t_0 = 0.69$) are displayed in each geographic region. Circle size corresponds to number of languages (*B* and *C*) or populations (*D*). (*E*) For the Ruhlen phoneme–genome dataset, pairwise geographic distance matrices were projected along different axes (calculated at 1° intervals); within each region, the rotated axis of geographic distance that was most strongly associated (greatest Mantel *r*) with phonemic distance (black arrows) and genetic distance (gray dashed arrows) is shown. Thinner arrows (Europe, East Asia, South America) indicate nonsignificant associations. Black dots indicate population locations for the Ruhlen phoneme–genome dataset. With the exception of North America, axes of phonemic differentiation and genetic differentiation are similar in most regions (North America: 78° difference; other regions: mean difference 16°).

To compare the signatures of human demographic history on genetic variation and phoneme inventories, we used Procrustes analyses to compare principal components (PCs) for both data types with sample geographic locations and determined whether phonemic and genetic distance are more correlated than expected from geographic distance alone. We also developed a new method for identifying regional axes of linguistic and genetic differentiation and tested whether the origin of the human expansion out of Africa can be detected from the geographic distribution of the numbers of phonemes in languages (phoneme inventory sizes). Conflicting predictions exist for the effects of geographic isolation and population contact on language evolution (e.g., refs. 26–29); we tested these by comparing phoneme inventories according to language density at varying radii. We also quantified the extent to which phoneme evolution can be modeled along genetic, geographic, and cognate-based phylogenies. With these joint analyses, we tested whether phonemes and alleles carry signatures of ancient population divergence and recent human migrations, and we identified demographic processes that have different effects on phonemes and alleles.

Results

Global Principal Component Analyses of Phonemic and Genetic Variation. Principal component analysis (PCA) is used to identify axes of variation in high-dimensional datasets (30, 31). To quantify broad similarities between geographic locations of samples (Fig. 1*A*) and PCs of phonemic and genetic data, we used Procrustes analyses (32) for all pairs of data types. We found significant concordance ($P < 10^{-5}$) between the first two PCs of phoneme presence/absence data and geographic locations for 2,082 languages in the Ruhlen database (Procrustes $t_0 = 0.57$) and for 968 languages in PHOIBLE ($t_0 = 0.52$), as well as between microsatellite data and geographic locations of 246 populations ($t_0 = 0.69$) (*SI Appendix*, Fig. S1). The mean values of Procrustes-transformed PCs of both phonemes and alleles corresponded to relative locations of geographic regions (Fig. 1 *B–D*): Africa was most differentiated from the Americas and Oceania, and Eurasian regions had intermediate locations.

Some differences between phonemic and genetic variation are also evident in Fig. 1 *B–D*. For example, the South American genetic sample was more differentiated from all other populations than the North American sample (Fig. 1*D*). In contrast, South American languages were near Oceanic languages in PC-space; on average, languages in both of these regions have small phoneme inventories (Fig. 1 *A–C*). The significant association between PCs and geographic locations for both languages and genes suggests that spatial diffusion has contributed to both phonemic and genetic variation.

Global Comparisons of Phonemic and Genetic Differentiation. To further quantify these associations with geography, we calculated pairwise Mantel correlations between phonemic distance, genetic distance, and geographic distance. Geographic distance and phonemic [Jaccard (33)] distance were significantly associated

for both the Ruhlen database (Mantel $r = 0.18$, $P < 10^{-4}$) and PHOIBLE ($r = 0.22$, $P < 10^{-4}$). The association between phonemic and geographic distance was also significant within all geographic regions except South America in the Ruhlen database and North/Central America in PHOIBLE (*SI Appendix*, Table S1). The phoneme–genome datasets showed a significant association (Mantel r) between phonemic distance and genetic distance (Ruhlen $r = 0.157$, $P = 2 \times 10^{-3}$; PHOIBLE $r = 0.240$, $P = 2 \times 10^{-4}$), between phonemic and geographic distances ($r = 0.18$, $P < 10^{-4}$; $r = 0.27$, $P < 10^{-4}$), and between genetic and geographic distances ($r = 0.76$, $P < 10^{-4}$; $r = 0.78$, $P < 10^{-4}$) (*SI Appendix*, Table S2). Thus, both phonemic and genetic data exhibited significant spatial autocorrelation; samples in geographic proximity were similar to one another, because of shared ancestry, spatial diffusion, or both (34, 35). To test the distance range of this spatial autocorrelation, we partitioned the geographic distance matrix into distance classes (*SI Appendix*). Whereas genetic distance showed spatial autocorrelation worldwide, phonemes were more similar among languages in the same distance class only within a range of ∼10,000 km (*SI Appendix*, Fig. S2*B*); beyond 10,000 km, phoneme inventories within a distance class were not more similar to one another than to those in another distance class.

To identify variables driving correlations between phonemic, genetic, and geographic distance (as in ref. 35), we controlled for each variable in turn with partial Mantel tests (36) (*SI Appendix*, Fig. S2). The partial Mantel correlation between genetic and phonemic distance was not significant when controlling for geographic distance (Ruhlen $r = 0.05$, $P = 0.16$; PHOIBLE $r = 0.05$, $P = 0.17$), suggesting both genetic and phonemic distance between samples can be predicted by their relative geographic locations (*SI Appendix*, Fig. S2 and Table S2). The relationship between geographic and phonemic distance controlling for genetic distance was significant ($r = 0.11$, $P = 0.01$; $r = 0.13$, $P < 0.01$), as was that between geographic and genetic distance controlling for phonemic distance ($r = 0.75$, $P < 10^{-4}$; $r = 0.77$, $P < 10^{-4}$). Through processes including migration and isolation by distance, geographic separation of populations could have led to spatial structuring in both data types, suggesting that geographic distance drives the similarity between genetic and phonemic distance.

These Mantel tests gave similar results within geographic regions, with a notable exception: in Oceania, genetic and phonemic distance were significantly correlated when controlling for geographic distance (Ruhlen $P = 2 \times 10^{-4}$; PHOIBLE $P = 2.6 \times 10^{-3}$) (*SI Appendix*, Table S2). Thus, for Oceanic populations, whose history includes extensive migration over water in the recent past (9), genetic and phonemic distance were more correlated than predicted by geographic distance.

Fine-Scale Geographic Axes of Variation. We developed a novel method to identify the geographic axes that are most closely associated with both phonemic and genetic differentiation. The significant association that we observed between geography and both phonemic and genetic variation (*SI Appendix*, Table S2) does not establish directions of geographic movement that best explain the current geographic distribution of phonemes and alleles. Furthermore, axes of variation determined from PCA do not necessarily represent specific large-scale migrations (37).

To determine fine-scale geographic axes that reflect differentiation between languages, we measured geodesic distance projected along different axes: the latitudinal and longitudinal axes, and the 1° increments between them. Within regions, we calculated Mantel correlations between geographic distance projected along each of these axes and phonemic distance. The axis with the greatest Mantel r identified the direction with the strongest association between geographic distance and phonemic distance (Fig. 1*E* and *SI Appendix*, Fig. S3 and Table S3).

For the phoneme–genome datasets, the rotated geographic axis identified as having the strongest association with phonemic distance was similar to that identified for genetic distance (Fig. 1*E* and *SI Appendix*, Fig. S3), suggesting that similar signatures of the directions of human differentiation within regions can be inferred from human genetic data and phonemic data. The greatest difference (78°) between the axes of differentiation predicted by phonemes and genes for the Ruhlen phoneme–genome dataset was based on eight populations unevenly spread across North America. However, genetic and phonemic axes of differentiation were similar for the six North American populations in the PHOIBLE phoneme–genome dataset (*SI Appendix*, Table S3). Further genotyping in this region will determine whether sparse sampling has driven this result. Our analysis does not specify which population processes, such as migration events, isolation by distance, and cultural diffusion, contribute to these axes of differentiation. Although these global analyses indicate strong associations between languages, genes, and geography, the worldwide patterns can be violated in local areas (e.g., Oceania in *SI Appendix*, Table S2 and North America in Fig. 1*E*).

Geographic Isolation and Neighboring Languages. Geographic isolation and drift could also drive local genetic and linguistic differentiation. Whereas geographic isolation decreases genetic diversity, studies disagree about the impact of isolation and processes analogous to drift on languages (e.g., refs. 26–29 and 38).

Over a series of radial distances, we assessed the effect of geographic isolation on phonemes in each language by comparing the phoneme inventories of each language and its neighbors. For languages that have fewer than or equal to the median number of neighboring languages within a radius of k kilometers ("fewer neighbors"), we observed a small but significant increase in phoneme inventory size as well as significantly higher phonemic distance between geographically close languages for many values of k (Fig. 2); this trend was also observed within Africa, Central/South Asia, East Asia, and Oceania (*SI Appendix*, Fig. S4). In areas with greater language density, phonemes were on average more similar between languages than in areas with fewer neighbors (*SI Appendix*, Fig. S5). In addition, languages with fewer neighbors had significantly higher variance in both phoneme inventory size and phonemic distance (Ansari–Bradley $P < 2 \times 10^{-3}$); this trend was also significant within Africa, Central/South Asia, East Asia, North America, and Oceania (*SI Appendix*, Fig. S6).

Geographic Signal Within and Between Language Families. The analyzed languages did not evolve independently: neighboring languages are often in the same family and related languages might share more phonemes. To address this, we compared phonemic distance with geographic distance to each language, separately for languages in the same language family and in different families. Geographic distance was significantly positively correlated with phonemic distance; this was true both for language pairs within the same family and for language pairs in different families within the same geographic area. Associations significantly different from zero ($P < 10^{-3}$) were positive for 99% of within-family comparisons and 87% of between-family comparisons. There was no significant difference in this relationship for languages in the same and different language families (Wilcoxon $P = 0.22$) (Fig. 3*C*). When two languages were geographically near, they tended to share more phonemes even if they were not closely related, suggesting a relationship between phonemes and geography both within and between language families.

The Signature of Ancient Population Divergence on Genes and Languages. Global genetic and phonemic patterns were not universally concordant: the most genetically polymorphic populations [top fifth percentile for number of microsatellite alleles observed (20)] are all in Africa, whereas the largest phoneme inventories in the Ruhlen database (top 5% of 2,082 languages, corresponding to at least 43 phonemes) (*SI Appendix*, Table S4) were globally distributed, predominantly in Africa (41 languages), Asia (32 languages), and North America (18 languages). Similarly, in PHOIBLE the languages with the most phonemes (top 5% of 968 languages, corresponding to at least 54 phonemes), were mainly in Africa (29 languages), Asia (12 languages), and North

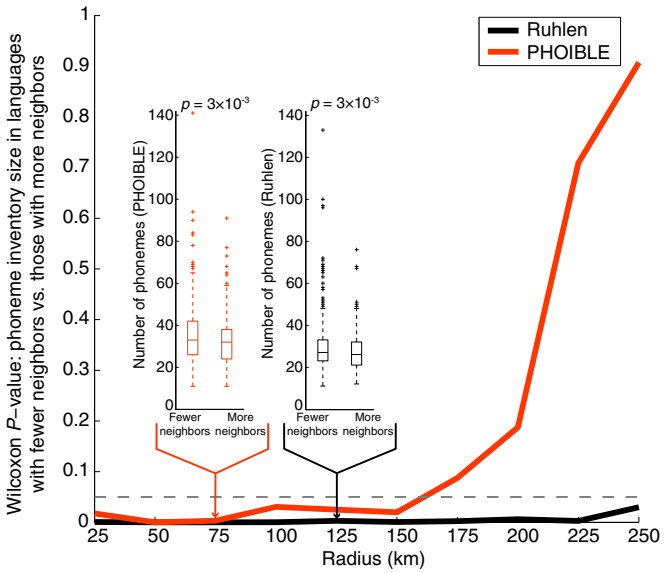

Fig. 2. The effect of geographic isolation on phonemes. Languages with fewer neighbors (less than or equal to median number of neighbors) had significantly more phonemes (Wilcoxon rank-sum test) than languages with more neighbors for all tested radii in the Ruhlen database (black line) and for radii < 175 km in PHOIBLE (red line). Examples are shown as inset boxplots: within a radius of 75 km for languages in PHOIBLE, the median number of neighbors was three languages; we observed slightly but significantly more phonemes in languages with zero to three neighbors than in languages with four or more neighbors (red boxplot inset). Similarly, within a radius of 125 km for languages in the Ruhlen database, there was a small but significant increase in the number of phonemes for languages with the median number of neighbors (8) or fewer (black boxplot inset).

America (7 languages). These distributions suggest that population divergence across large distances might have affected phonemic and genotypic variation differently.

Ancient population divergence is evident in human genetic diversity, which decreases with distance from southern Africa, a signature of the serial founder effect (13, 39, 40). Parallel patterns of decreasing diversity out of Africa have been reported for the partially vertically transmitted human pathogen *Helicobacter pylori* (41) and in human morphometric data (42). Inference of the human expansion out of Africa has also been attempted using categorical phoneme inventories (15), although phonemes are not necessarily lost after a population bottleneck. The conclusions from Atkinson (15) that language expansion followed a serial founder effect out of Africa and that phoneme inventory size was significantly correlated with current speaker population size (as in ref. 43) have both generated much debate (e.g., refs. 16–19, 25, 28, and 44–46). Using both databases of phoneme inventories, we tested whether ancient human population divergence out of Africa left a similar signature on phonemes to that on genes.

To compare the Ruhlen database and PHOIBLE with previous studies (15–18, 25), we regressed phoneme inventory size on geographic distance from 4,210 geographic centers on Earth (2, 13) and tested for a linear decrease in number of phonemes with distance to each center. For both databases, the geographic center with the most support for this model (lowest Akaike Information Criterion, AIC) was in northern Europe (Fig. 3) (Ruhlen 67.6684°, 36.2°; PHOIBLE 77.1614°, 16.4°); the distance between these centers is 1233.5 km. A decrease in number of phonemes with distance from Eurasia has been observed before (16).

Although our analysis identifies a Eurasian center as the best-fit origin, we do not claim that a serial founder effect is an appropriate model for language expansion: phoneme inventory size is a coarse summary statistic, and phoneme loss does not necessarily occur with reduced population size or geographic isolation. Rather, the identified location is roughly equidistant from most languages in Oceania and South America, effectively grouping these regions of generally small phoneme inventory size to produce a significantly negative slope. Furthermore, the 2,082 points in the regression are not independent: many represent closely related languages (Fig. 3*A*). To reduce this dependence, we repeated the regression analysis using the mean or median values for the independent and dependent variables within each language family (Fig. 3*B*). As with individual languages, the best-fit origin was found in Northern Europe for the within-family mean and median values for both the Ruhlen database and PHOIBLE (*SI Appendix*, Fig. S7 and Table S5).

To address the relationship between current speaker population size and phoneme inventory size (25, 28, 44–46), we repeated the regression analysis using speaker population size as an additional independent variable, and we found no statistical support in the Ruhlen database for including it in our regression models ($P = 0.35$). For PHOIBLE, including the base 10 logarithm of speaker population sizes reported by Ethnologue as another independent variable in the regression model produced the same best-fit center as the simple linear regression (67.6684°, 36.2°) and led to a modest but significant increase in the variance

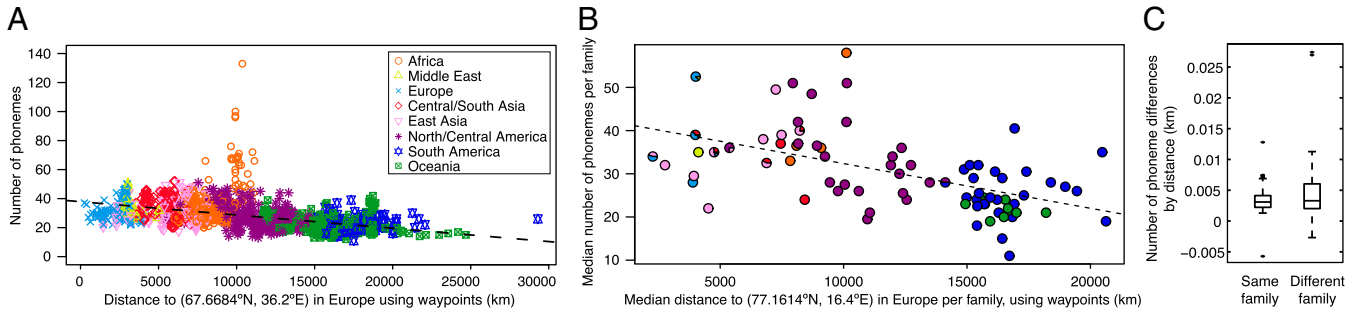

Fig. 3. Best-fit linear regressions of phoneme inventory size on geographic distance. For both databases, the best-fit geographic center was located in northern Europe, roughly equidistant from Oceania and South America, grouping two regions with small phoneme inventories and producing a significantly negative slope. This finding suggests that phonemes do not show a strong signature of ancient population divergence. (*A*) Regression from the best-fit of 4,210 geographic centers on the Earth for languages in the Ruhlen database (see *SI Appendix*, Fig. S7 for PHOIBLE). (*B*) Using the median number of phonemes within each family, the best-fit geographic center for language families in PHOIBLE remained in northern Europe (see *SI Appendix*, Fig. S7 for Ruhlen). Geographic regions are indicated by color as in *A*, but *y*-axis scales differ. (*C*) Phonemic distance increases with geographic distance, even for languages in different families. For significant correlations between phonemic distance and geographic distance, the slope of the regression line for both within-family and between-family comparisons (*y* axis) was positive the vast majority of the time, and the distributions of these slopes were not significantly different from one another (Wilcoxon $P = 0.22$).

explained by the regression (from $r = 0.2082$ in the simple regression to $r = 0.2114$ in the multiple regression, $P = 4.33 \times 10^{-3}$).

Ancestral Character Estimation of Phonemes Along Genetic, Geographic, and Linguistic Phylogenies. In regression analyses, phoneme inventory size did not show a signature of ancient population divergence (Fig. 3), and horizontal transmission between languages could play a role in phoneme evolution (Fig. 3*C*). Linguistic trees are constructed using cognate words predicted to have shared ancestry; similarly, genetic phylogenies assume vertical transmission of alleles. To account for the effect of borrowing between neighboring populations on phoneme distributions, we constructed a tree from geographic distances between languages. To assess the extent to which linguistic, genetic, and geographic relationships each describe phoneme evolution, we used three trees to estimate ancestral phoneme inventories and checked the concordance of these with ancestral phoneme inventories found in the literature (Table 1).

For an Indo-European linguistic tree (47), a genetic tree of Indo-European-speaking populations, and a neighbor-joining tree of the geographic distances between language locations, we estimated the probability of phoneme presence at two internal nodes. Fig. 4 *A*–*C* illustrates the results of ancestral character estimation for an example phoneme, /t/. We then compared these ancestral character estimates to the phoneme inventories of well-studied ancient languages for which primary sources exist: we used Vulgar Latin phonemes to approximate the phoneme inventory ancestral to modern Romance languages (48, 49) and Vedic Sanskrit phonemes to approximate the phoneme inventory ancestral to modern Indo-Aryan languages (50). For phoneme inventories in both databases, the cognate-based phylogeny (47), a geographic tree, and a genetic phylogeny gave similar predictions of the phoneme inventories of Vulgar Latin and Vedic Sanskrit (Table 1). The prediction of phoneme presence/absence with the ancestral character estimation algorithm was consistent with published sources for 67–88% of phonemes. Of the phonemes in published inventories that were accurately predicted by ancestral character estimation, most (53–94%) were predicted by multiple trees (*SI Appendix*, Fig. S8). In addition, each tree gave similar estimates for relative rates of phoneme change (Fig. 4*D*).

Discussion

We have analyzed the largest available datasets of both phoneme inventories and genotyped populations. Across multiple analyses, phonemic and genetic samples showed strong signatures of their geographic location. Phonemic and genetic differentiation also occurred along similar axes, indicating that genetic and linguistic data show similar signatures of human population dispersal within regions. The data types were discordant in two ways: first, although relatively isolated populations lose genetic diversity, their languages might be more susceptible to change than those of populations with many neighbors; second, phonemes might not retain a signature of human expansion out of Africa as genes do.

Differences among populations in both phonemes and allele frequencies were strongly correlated with geographic distance. Furthermore, phonemes showed an association with geographic distance regardless of language classification but did not show the strong signatures of ancient population divergence found in genetic data. This suggests that phoneme inventories are affected by recent population processes and thus carry little information about the distant past (e.g., ref. 23); in contrast to genes, phoneme inventories in our analyses did not follow the predictions of a serial founder effect out of Africa. We also pinpoint where differences between genes and languages occur, both geographically and by characteristics of the surrounding populations. Our findings suggest that geographic isolation has different effects on genes and phonemes. Languages with fewer neighboring languages were more phonemically different from their neighbors than those with more neighbors, and geographically isolated populations may gain phonemes while losing genetic variation. In addition, ancestral phoneme inventories estimated along genetically, geographically, and lexically determined phylogenies produced similar results (Table 1).

We quantified the similarity between phoneme inventories and genetic polymorphisms on a worldwide scale. To guard against spurious correlations between phoneme inventories and geography, we analyzed two databases and repeated the analyses using subsets of the data. The two phoneme databases yielded similar results, giving additional support for our conclusions (51). Geographic distance was a significant predictor of both phonemic distance between languages and genetic distance between populations (*SI Appendix*, Fig. S2 and Table S2). The spatial distribution of populations, via migration and isolation by distance, could have led to geographic structure in both genes and languages; this result alone does not shed light on the existence or extent of any deep historical signal in either data type. The association between genetic variation and phonemic variation was largely explained by the geographic distribution of populations: beyond common signatures of spatial structure in genes and languages, genetic distance was not causally related to phonemic distance. Furthermore, the spatial structuring in genes and languages did not occur on the same scale: genetic samples showed spatial autocorrelation worldwide, but phoneme inventories were spatially autocorrelated only within a range of ∼10,000 km (*SI Appendix*, Fig. S2*B*).

Phonemic distance increased with geographic distance, even for languages that were not classified as belonging to the same language family, that is, without recent shared ancestry (Fig. 3*C*). Nearby languages shared more phonemes than distant ones, suggesting that geographic proximity and opportunities for language contact could lead to phoneme borrowing between languages that do not have recent shared ancestry (21, 22, 27, 28). Relatively isolated languages exhibited more variance in number of phonemes than languages with many neighbors (Fig. 2). This finding supports the hypothesis that more geographically isolated populations, with smaller social networks and fewer second-language learners, may be more likely to undergo sound changes, such as losing or gaining phonemes (27–29, 38).

Geographically isolated languages tended to be more different from their neighbors than languages in regions of high language density (*SI Appendix*, Fig. S5). This finding agrees with Trudgill's hypothesis that isolation can both preserve existing language complexity and lead to spontaneous complexification (28) but is in stark contrast to genetic drift, whereby isolation reduces genetic diversity within populations (13, 52). Contact among speakers of different languages could initiate phoneme change, as borrowed words could introduce phonemes or use existing phonemes in new phonological contexts (22, 27). Long-term contact could promote phoneme sharing between languages (27, 28), perhaps increasing phoneme similarity in areas of high language density but not for isolated languages.

Genetic differentiation between human populations increases with geographic distance (13, 52–54), but the degree of differentiation may vary along different geographic axes (54–56). Within large regions, we computed the geographic axes along which phonemic differentiation was most closely associated with geographic distance between languages; these were consistent with

Table 1. Accuracy of ancestral character estimation for Vulgar Latin and Vedic Sanskrit

Language	Cognate tree	Genetic tree	Geographic tree
Vulgar Latin	71% (88%)	67% (75%)	69% (86%)
Vedic Sanskrit	68% (83%)	72% (77%)	62% (80%)

Using cognate, genetic, and geographic trees of Indo-European populations, ancestral character estimates (63) of phoneme presence/absence were compared with published phoneme inventories for Vulgar Latin and Vedic Sanskrit (48–50); percent accuracy is indicated for the Ruhlen database and PHOIBLE (in parentheses).

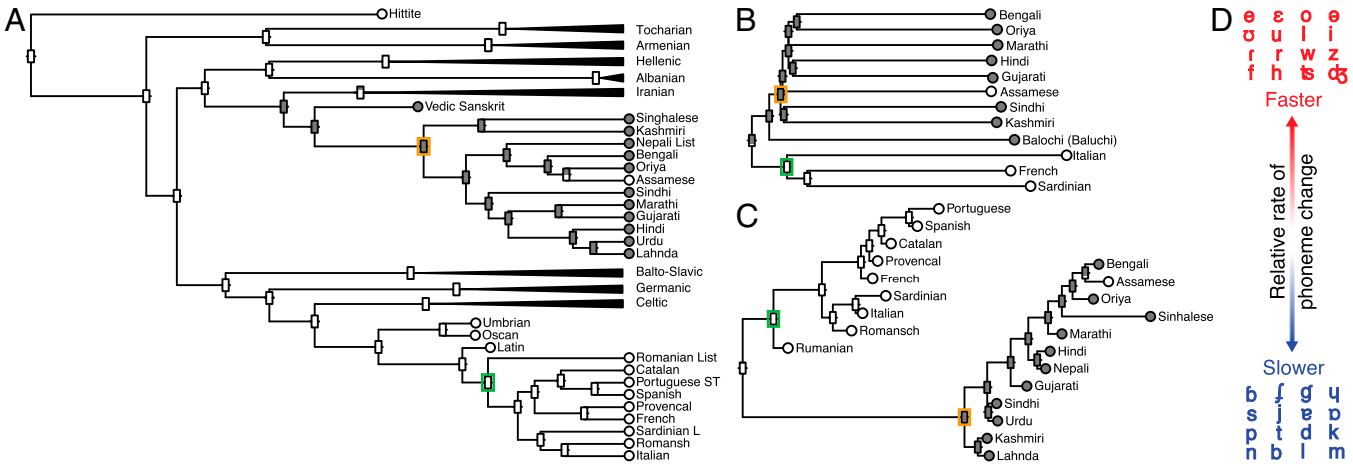

Fig. 4. Estimating ancestral phoneme states with cognate-based, geographic, and genetic trees. (*A*) Phylogeny of Indo-European languages (47) with presence of the phoneme /l/ indicated by gray circles at each tip. Based on the tree topology and branch lengths, the probability of phoneme presence at interior nodes was predicted by ancestral character estimation (63). The amount of gray in the bar at each node represents the probability of phoneme presence, with white representing absence. The green rectangle highlights the low probability (2.84×10^{-3}) of the presence of phoneme /l/ in the ancestor to Romance languages, as shown by the lack of gray at that node. The orange rectangle highlights the probability of /l/ presence in the ancestor to Indo-Aryan languages (~1). (*B*) Phylogeny of Indo-European populations constructed with genetic data from ref. 20. (*C*) Neighbor-joining tree of geographic distances between Indo-European-speaking populations. As in *A*, the presence of /l/ in the language spoken by a given population is indicated in *B* and *C* by gray circles, and the probability of this phoneme's presence at interior nodes (predicted by ancestral character estimation) is shown by the amount of gray at each node. For all three trees, the phoneme /l/ was estimated to be likely absent in the language ancestral to the Romance languages (indicated by a mostly white bar inside each green rectangle) and likely present in the language ancestral to the Indo-Aryan languages (orange rectangle). (*D*) Examples of phonemes in the Ruhlen database were grouped by their relative rate of change from high (red) to low (blue) as predicted by the ancestor character estimation algorithm with all three trees. Predictions of relative rates of phoneme change were consistent among all pairs of the three trees (Spearman's $\rho \geq 0.73$, $P \leq 4.9 \times 10^{-15}$).

axes predicted using microsatellite data (Fig. 1*E* and *SI Appendix*, Fig. S3). This analysis could provide an alternative to PCA for making inferences about human populations. The first two PCs of both allele frequencies and phoneme inventories were significantly associated with geographic locations; however, PCA does not specify the mechanism underlying this association (37) or directly suggest deep historical signal in either data type.

A regression-based analysis of phoneme inventory size (15) concluded that a global sample of 504 languages fit a serial founder effect model of expansion out of Africa (but see refs. 16–19). Using a similar approach, we found that phoneme inventory size decreased with geographic distance from northern Europe (Fig. 3); we do not conclude that this supports an origin for language in Europe for several reasons. Although a population's genetic diversity reflects the number of its founders, the relationship between the number of founders of a population and its language's phonemes is more complex (18, 21, 25, 27, 43–46). Furthermore, only a subset of the model's predictions apply to languages (16), and the mutation rate of phonemes may be high enough that signatures of ancient divergence are erased faster in phonemes than in genes (39, 57). In contrast to previous studies (15, 43), speaker population sizes did not explain a significant proportion of variation in phoneme inventory size (as in ref. 25) (*SI Appendix*, Fig. S9).

Human genetic phylogenies display relationships among populations that reflect the vertical transmission of genes. Cognate-based phylogenies offer an independent linguistic approach to identifying relationships among populations (21, 47). At a timescale over which linguistic inference is possible, we estimated ancestral phoneme states from phoneme inventories using genetic, geographic, or cognate-based phylogenies (Fig. 4). For each tree, our estimates of ancestral phoneme states are consistent (62–88%) (Table 1) with published ones. Differences between estimated and published phoneme inventories could occur because the ancestral character estimation algorithm makes inaccurate assumptions regarding phoneme evolution (such as a constant rate of phoneme change) or because a binary scheme of phoneme presence and absence does not reflect that certain sound changes are more likely than others. In estimating ancestral phoneme inventories, the performance of the genetic phylogeny depends on the distribution of genotyped populations in the language family (Fig. 4*B*). Despite few genetic samples, the genetic, geographic, and linguistic trees predicted roughly similar ancestral phoneme inventories, and this type of analysis could provide an opportunity for future collaboration between linguists and geneticists. Vertical descent from a common ancestor is not an ideal model for phoneme evolution over long timescales; analyses like those in Fig. 4 and Table 1 shed light on the extent to which a vertical model is appropriate for a given dataset.

Our results reflect that both borrowing and vertical transmission influence phoneme distributions among languages; increasing the density of genetic samples is necessary to rigorously estimate the relative roles of these processes in phoneme evolution. Moreover, joint analysis using genetic, geographic, and linguistic phylogenies provides a framework for future applications to data: given genetic or geographic relationships among a set of populations, a subset of information about ancestral languages may be extracted without prior knowledge of linguistic relationships. These joint analyses of genetic and linguistic data yield insight into the effect of evolutionary forces on linguistic traits that could not be achieved by either data type alone.

Materials and Methods

Preparation of Linguistic and Genetic Data. For 2,082 languages, the Ruhlen database has complete phoneme inventories, sources, and a corresponding entry in the Ethnologue database (24); the presence/absence matrix of phonemes in the Ruhlen database is archived at PNAS. PHOIBLE (phoible.org) (25) contains phoneme inventories for 968 languages; 621 could be matched across databases (*SI Appendix*, Fig. S10).

For the Ruhlen database, we annotated languages with an International Organization for Standardization (ISO) 639-3 language code and an ISO 3166-1 alpha-3 country code corresponding to an entry in the Ethnologue, which contained latitude and longitude coordinates and speaker population size estimates. PHOIBLE contains ISO 639-3 codes, geographic coordinates, and phoneme inventories. We encoded the presence of 728 phonemes in 2,082 languages in the Ruhlen database and 1,587 phonemes in 968 languages in PHOIBLE into separate binary matrices for analysis (*SI Appendix*). Unless specified, we performed analyses on both databases.

We also analyzed a dataset of 645 microsatellite loci from several studies (20). Using population names and locations (20), we matched genotyped populations to their native language (*SI Appendix*). For 139 populations in the Ruhlen database and 114 in PHOIBLE, we were able to merge genetic, geographic, and phonemic data (the phoneme–genome datasets).

Principal Components and Procrustes Analyses. For the Ruhlen database and PHOIBLE, we performed PCA on the binary matrices of phonemic data (*SI Appendix*, Fig. S11) along with Procrustes analysis of phoneme PCs versus the geographic coordinates of languages analyzed. Following Wang et al. (32), we calculated a similarity statistic $t_0 = \sqrt{(1 - D)}$, where D is the minimized sum of squared distances after Procrustes analysis. We calculated empirical P values for t_0 values over 10^5 permutations of geographic locations. For eight geographic regions (detailed in *SI Appendix*), we calculated the mean values of the Procrustes-transformed principal components (Fig. 1 *B–D*). For the phoneme–genome datasets, we performed Procrustes comparisons between each pair of data types: phoneme PCs, genetic PCs, and geographic locations.

Correlations Between Phonemic, Genetic, and Geographic Distance. For the Ruhlen database and PHOIBLE, we compared geographic (great-circle with waypoints) and phonemic [Jaccard (33) and Hamming (58)] distance matrices using Mantel tests (P values calculated over 10^4 permutations). In addition, we considered latitudinal and longitudinal distance separately by calculating the absolute value of the difference in latitude and longitude coordinates. For the phoneme–genome datasets (139 populations in Ruhlen and 114 in PHOIBLE), we assembled pairwise geographic, phonemic, and genetic (allele-sharing) distance matrices and performed Mantel tests between each pair of matrices. We then performed partial Mantel tests to compare each pair of distance matrices while controlling for the third. We repeated each test for each region separately. (See *SI Appendix* for further details.)

For each pair of languages, let $\vec{A}$ be the vector connecting their geographic locations. We projected $\vec{A}$ in the direction of a given vector $\vec{B}$ by computing $|\vec{A}|cos(\theta)$, where θ is the angle between $\vec{A}$ and $\vec{B}$. $\vec{B}$ was then rotated at 1° intervals around the unit circle, and the distance between each pair of languages projected in the direction of $\vec{B}$ was recorded in a projected distance matrix. Within each geographic region, we performed Mantel tests between these distance matrices projected in different directions and both genetic and phonemic distance and recorded the direction with the largest Mantel r statistic (Fig. 1*E*, and *SI Appendix*, Fig. S3 and Table S3).

Phoneme Similarity as a Function of Language Density. We performed a series of Wilcoxon rank-sum and Ansari–Bradley tests, comparing the phoneme inventory sizes in languages with less than or equal to the median number of neighbors versus the phoneme inventory sizes in languages with greater than the median number of neighbors. We defined the number of neighboring languages as the number of languages whose geographic location in the Ethnologue database (24) occurs within a certain radius of the focal language's Ethnologue coordinates. We varied radii from 25 km to 250 km in steps of 25 km for this analysis.

We also analyzed Hamming distance between languages, defined as the number of phonemic differences between languages divided by the number of possible phonemes in the database. For each linguistic database, we calculated the pairwise phonemic distance between a focal language and all other languages within a given radius, and we recorded the number of languages neighboring the focal language within that radius. Languages with no neighboring languages within a given radius were excluded. With Wilcoxon rank-sum and Ansari–Bradley tests, we then compared the distribution of phonemic distances from languages with the median number of neighbors or fewer to those with greater than the median number of neighbors, varying radii from 100 km to 1,000 km in steps of 100 km. Note that we could only test phonemic distance at radii with a median number of neighbors greater than or equal to 2.

Phoneme Similarity Within and Between Language Families in PHOIBLE. We compared the relationship between phonemic distance and geographic distance for pairs of languages in the same language family and in different language families. If a given language was classified into a language family by PHOIBLE (25, 59), we performed "within-family comparisons" by calculating both the pairwise geographic distance and the pairwise phonemic distance [Hamming (58) and Jaccard (33)] between that language and other members of the same language family (excluding members of the same language family located more than 10,000 km away). For these within-family comparisons with the given language, we then regressed phonemic distance onto geographic distance and recorded the correlation coefficient, the P value of the correlation coefficient, and the slope of the fitted linear model.

We then performed "between-family comparisons" with the same language using languages in other language families that were within the same geographic radius as the within-family comparisons: either the maximum distance to a member of the same family or 10,000 km, whichever was smaller. For the between-family comparisons, we again regressed phonemic distance onto geographic distance and recorded the correlation coefficient, the P value of the correlation coefficient, and the slope of the fitted linear model. After completing this procedure for all languages, we compared the distribution of regression slopes and correlation coefficients for within-family and between-family comparisons using a Wilcoxon rank-sum test. Because languages in the Ruhlen database were not annotated with this classification system, this analysis was performed only on PHOIBLE.

Regression Analyses. We performed a series of regressions of phoneme inventory size on geographic distance from each of 4,210 centers drawn from the surface of the earth as in ref. 13. One independent variable in all models fitted was geographic distance between languages and each of 4,210 centers, calculated using obligatory waypoints from refs. 13 and 2. In regression analyses, we only used languages with Ethnologue speaker population size greater than 0 (2,004 languages in Ruhlen, 967 in PHOIBLE).

For each linguistic database, let our dependent variable, $\vec{Y}$, be the vector of phoneme inventory sizes across languages with speaker population size > 0. We used two types of model for each database: (*i*) phoneme inventory sizes in $\vec{Y}$ were regressed on geographic distances to a center for each of 4,210 centers, and (*ii*) phoneme inventory sizes in $\vec{Y}$ were regressed on geographic distances to a center and the base 10 logarithm of speaker population size for each of 4,210 centers. We estimated model parameters Θ (regression coefficients, intercept, and residuals) using linear regression of $\vec{Y}$ as a function of geographic distance to a center (and speaker population size).

For model selection, we used AIC. Because values of AIC lie on a relative scale, values were rescaled by subtracting the minimum AIC observed for a given model fit across 4,210 centers. Models with a rescaled AIC ≤ 2 are considered to have equivalent support (60) (*SI Appendix*, Fig. S12).

More detail on regression analyses conducted here, such as jackknifing over geographic regions and using different measures of phoneme inventory size (e.g., eliminating click phonemes) for the dependent variable $\vec{Y}$ are discussed in *SI Appendix* and produced qualitatively similar results to those presented here.

We repeated the regression analyses with languages grouped by Ethnologue language family (Ruhlen database) or family/root (PHOIBLE). For both databases, simple linear regressions (geographic distance to the center as the independent variable) and multiple linear regressions (geographic distance to the center and base 10 logarithm of speaker population size as independent variables) were fitted, and the dependent variable was total phoneme inventory size. We then calculated the mean and median value of the independent and dependent variables within each family (root).

The Ruhlen database has 2,046 languages classified in 98 Ethnologue language families; 36 Ruhlen entries with language families labeled as "Unclassified," "Language Isolate," or "Mixed Language" were excluded from this analysis. PHOIBLE has 949 language classified into 81 language roots; 19 languages listed with unclassified roots (denoted as "UNCL" by PHOIBLE) were excluded from this family-based analysis.

Phylogenetic Analyses. To construct a rooted tree of 246 nonadmixed human populations, we analyzed the 246 microsatellite loci from the MS5339 dataset of Pemberton et al. (20) with chimpanzees as an outgroup. First, we generated allele-sharing genetic distance matrices, bootstrapping over loci 1,000 times using MICROSAT (61). We constructed a consensus neighbor-joining tree (NEIGHBOR; extended Majority Rule CONSENSE) (62). We generated maximum-likelihood estimates for consensus tree branch lengths using CONTML (62), with an allele-sharing distance matrix generated from all 246 loci. This tree was trimmed using the drop.tip function (63) to include only the subset of populations speaking Indo-European languages. For these populations, we also constructed a neighbor-joining tree of geographic distances (using waypoints as in ref. 13) between languages. Branch lengths of the linguistic and geographic trees were rescaled to be comparable to the genetic tree.

We then applied an equal-rates ancestral character estimation algorithm to the Indo-European subset of populations using the ace function in the Analyses of Phylogenetics and Evolution package in R (63) to predict the probability that each phoneme was present at each ancestral node of the tree. For populations with Indo-European languages, we performed this analysis with three phylogenies: the genetic consensus tree, the tree of geographic distances, and a published Bayesian cognate-based linguistic tree of Indo-European languages (47). We tested 728 phonemes in the Ruhlen database and 1,587 phonemes in PHOIBLE and estimated: (*i*) the rate of change of each phoneme along both trees and (*ii*) the ancestral character

states at two nodes, the common ancestor to Romance languages and the common ancestor to Indo-Aryan languages. Most phonemes in each database did not occur in any Indo-European languages and were thus estimated to be absent at all ancestral nodes. For phonemes present in at least one Romance or Indo-Aryan language, we compared the phoneme presence/absence predicted by the ancestral character estimation algorithm with a published phoneme inventory and calculated the percent accuracy by dividing the number of phonemes correctly predicted by the number of phonemes tested.

ACKNOWLEDGMENTS. We thank the two reviewers; Kenichi Aoki, Mike Alfaro, Erika Edwards, Kevin Laland, Stacey Smith, Oana Carja, and Chaolong Wang for helpful discussions; Alyna Khan and Paul O'Leary McCann for informatic support; and Jeffrey C. Long for comments on an earlier version. This work was supported in part by the Morrison Institute for Population and Resource Studies; the Center for Computational, Evolutionary, and Human Genomics at Stanford; the Ruth Landes Memorial Research Fund (to N.C.); the Alfred P. Sloan Foundation (S.R.); and the Pew Charitable Trusts (S.R.).

1. Reich D, et al. (2012) Reconstructing Native American population history. *Nature* 488(7411):370–374.
2. Wang S, et al. (2007) Genetic variation and population structure in Native Americans. *PLoS Genet* 3(11):e185.
3. Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci USA* 85(16):6002–6006.
4. Ruhlen M (1987) *A Guide to the World's Languages* (Stanford Univ Press, Stanford, CA).
5. Barbujani G, Sokal RR (1990) Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc Natl Acad Sci USA* 87(5):1816–1819.
6. Piazza A, et al. (1995) Genetics and the origin of European languages. *Proc Natl Acad Sci USA* 92(13):5836–5840.
7. Rosenberg NA, et al. (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet* 2(12):e215.
8. Hunley KL, Cabana GS, Merriwether DA, Long JC (2007) A formal test of linguistic and genetic coevolution in native Central and South America. *Am J Phys Anthropol* 132(4): 622–631.
9. Friedlaender JS, et al. (2008) The genetic structure of Pacific Islanders. *PLoS Genet* 4(1):e19.
10. Wang S, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet* 4(3):e1000037.
11. Tishkoff SA, et al. (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–1044.
12. Schlebusch CM, et al. (2012) Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* 338(6105):374–379.
13. 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(44):15942–15947.
14. Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. *Curr Biol* 15(5):R159–R160.
15. Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027):346–349.
16. Hunley K, Bowern C, Healy M (2012) Rejection of a serial founder effects model of genetic and linguistic coevolution. *Proc R Soc Lond B Biol Sci* 279(1736):2281–2288.
17. Wang CC, Ding QL, Tao H, Li H (2012) Comment on 'Phonemic diversity supports a serial founder effect model of language expansion from Africa.' *Science* 335(6069):657.
18. Cysouw M, Dediu D, Moran S (2012) Comment on 'Phonemic diversity supports a serial founder effect model of language expansion from Africa'. *Science* 335(6069):657.
19. Maddieson I, Bhattacharya T, Smith DE, Croft W (2011) Geographical distribution of phonological complexity. *Linguist Typol* 15(2):267–279.
20. Pemberton TJ, DeGiorgio M, Rosenberg NA (2013) Population structure in a comprehensive genomic data set on human microsatellite variation. *G3* 3(5):891–907.
21. Campbell L (1998) *Historical Linguistics: An Introduction* (MIT Press, Cambridge, MA).
22. Hoijer H (1948) Linguistic and cultural change. *Language* 24(4):335–345.
23. Hock H, Joseph B (1996) *Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics* (Mouton de Gruyter, Berlin).
24. Lewis MP (2009) *Ethnologue: Languages of the World* (SIL International, Dallas, TX), Vol 16, Available at www.ethnologue.com. Accessed April 15, 2011.
25. Moran S, McCloy D, Wright R (2012) Revisiting population size vs. phoneme inventory size. *Language* 88(4):877–893.
26. Bakker P (2004) Phoneme inventories, language contact, and grammatical complexity: A critique of Trudgill. *Linguist Typol* 8(3):368–375.
27. Trudgill P (2004) Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguist Typol* 8(3):305–320.
28. Trudgill P (2011) Social structure and phoneme inventories. *Linguist Typol* 15(2): 155–160.
29. Dahl Ö (2004) *The Growth and Maintenance of Linguistic Complexity* (Benjamins Publishing, Amsterdam).
30. Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos Mag* 2(11):559–572.
31. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. *Science* 201(4358):786–792.
32. Wang C, et al. (2010) Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Mol Biol* 9:13.
33. Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bull Soc Vaud Sci Nat* 44:223–270.
34. Sokal RR (1979) Testing statistical significance of geographic variation patterns. *Syst Zool* 28(2):227–232.
35. Legendre P (1993) Spatial autocorrelation: Trouble or new paradigm? *Ecology* 74(6): 1659–1673.
36. Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool* 35(4):627–632.
37. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 40(5):646–649.
38. Nettle D (2012) Social scale and structural complexity in human languages. *Philos Trans R Soc Lond B Biol Sci* 367(1597):1829–1836.
39. DeGiorgio M, Jakobsson M, Rosenberg NA (2009) Out of Africa: Modern human origins special feature: Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci USA* 106(38):16057–16062.
40. Henn BM, et al. (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc Natl Acad Sci USA* 108(13):5154–5162.
41. Linz B, et al. (2007) An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445(7130):915–918.
42. Manica A, Amos W, Balloux F, Hanihara T (2007) The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448(7151):346–348.
43. Hay J, Bauer L (2007) Phoneme inventory size and population size. *Language* 83(2): 388–400.
44. Bowern C (2011) Out of Africa? The logic of phoneme inventories and founder effects. *Linguist Typol* 15(2):207–216.
45. Donohue M, Nichols J (2011) Does phoneme inventory size correlate with population size? *Linguist Typol* 15(2):161–170.
46. Dahl Ö (2011) Are small languages more or less complex than big ones? *Linguist Typol* 15(2):171–175.
47. Bouckaert R, et al. (2012) Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097):957–960.
48. Hall RA (1950) The reconstruction of Proto-Romance. *Language* 26(1):6–27.
49. Grandgent CH (1907) *An Introduction to Vulgar Latin* (DC Heath and Company, Boston).
50. Whitney WD (1879) *A Sanskrit Grammar; Including Both the Classical Language, and the Older Dialects, of Veda and Brahmana* (Breitkopf and Härtel, Leipzig, Germany).
51. Roberts S, Winters J (2013) Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE* 8(8):e70902.
52. Jakobsson M, et al. (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451(7181):998–1003.
53. Li JZ, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100–1104.
54. Ramachandran S, Rosenberg NA (2011) A test of the influence of continental axes of orientation on patterns of human gene flow. *Am J Phys Anthropol* 146(4):515–529.
55. Nei M, Roychoudhury AK (1993) Evolutionary relationships of human populations on a global scale. *Mol Biol Evol* 10(5):927–943.
56. Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. *Proc Natl Acad Sci USA* 109(44):17758–17764.
57. DeGiorgio M, Degnan JH, Rosenberg NA (2011) Coalescence-time distributions in a serial founder model of human evolutionary history. *Genetics* 189(2):579–593.
58. Hamming RW (1950) Error detecting and error correcting codes. *Bell Syst Tech J* 29(2): 147–160.
59. Dryer MS, Haspelmath M, eds (2013) *The World Atlas of Language Structures Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany).
60. Burnham KP, Anderson DR (2010) *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (Springer, New York), 2nd Ed.
61. Minch E, Ruiz-Linares A, Goldstein DB, Feldman MW, Cavalli-Sforza LL (1997) MICROSAT, Version 1.5 b. Available at genetics.stanford.edu/hpgl/projects/microsat/. Accessed November 19, 2012.
62. Felsenstein J (2005) *PHYLIP (Phylogeny Inference Package) version 3.6.* Distributed by the author (Department of Genome Sciences, Univ of Washington, Seattle, WA).
63. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20(2):289–290.

SI Appendix: A comparison of worldwide phonemic and genetic variation in human populations

Table of Contents

SI Materials and Methods

1. Preparation of linguistic data

1.1 Processing the Ruhlen database

An introduction to the Ruhlen database and notation used in it is available in typology-descr.pdf, prepared by Merritt Ruhlen and available at ehl.santafe.edu. The Ruhlen database was originally created with Microsoft Notepad's encodings of International Phonetic Alphabet (IPA) characters; we converted it into presence/absence matrices and files with phonemes listed in Unicode.

We made the following removals during processing of the database: we removed marginal phonemes (noted by the source to be either very rare or only occurring in loanwords and denoted in the Ruhlen database within parentheses); we removed any repetitions of phonemes within the same language; we removed extra whitespace between phonemes to facilitate downstream processing; we removed any superfluous punctuation. In seven cases, we made alterations to the raw data based on the written comments in the database; all involved comments mentioning phonemes used in some contexts but not all contexts (i.e., by women or in colloquial speech). In these 7 cases, we added those phonemes used in particular contexts to the inventories in the Ruhlen database. All other comments were simply removed from the database. Files used in analysis, generated after the filtering steps described in sections 1.1-1.6, are archived at PNAS as Datasets S1-S3.

To facilitate comparisons between languages, we standardized the representation and ordering of phonemes to ensure that separate observations of the same phoneme were coded consistently throughout the database (e.g. both $k^{ʔw}$and $k^{wʔ}$ appear in the database, but can be considered equivalent, so we standardized these types of duplications into $k^{ʔw}$).

1.2 Modified consonants and modified vowels

Modifications to consonants and vowels (Table S6) are listed separately from the phonemes modified in the Ruhlen database. For example, if certain phonemes in a language can be prenasalized, the Ruhlen database encodes the presence of prenasalized consonants separately from the individual phonemes. When individual consonants had multiple modifications ("compound modification"), we encoded the presence of the particular compound modification in each language in which it occurs.

1.3 Removing non-phonemic distinctions: aspirates and dentals

Of the modifications in Table S6, aspiration (^{h}) only results in a phonemic distinction when it occurs in a language that also has the unaspirated version of the same phoneme. Similarly, dental (◌̪) only results in a phonemic distinction when the corresponding alveolar phoneme is also present. Thus, we removed aspiration and dental if they made only allophonic but not phonemic distinctions.

Due to the limited occurrence of clicks across the world's languages, aspiration in clicks was treated separately from other fields. As with other phonemes, if aspiration provided a phonemic distinction within clicks in a language, the presence of all aspirated click phonemes was recorded. In only one language with clicks, Xhosa (Ruhlen language number 1160), were aspirated clicks recorded without any corresponding unaspirated clicks; for Xhosa we considered these clicks to be functionally equivalent to unaspirated clicks and removed the aspiration distinction within clicks.

In 7 languages, the only consonant field where aspiration resulted in a phonemic distinction is in clicks. These languages (along with their language number in the Ruhlen database) are: Sandawe (2), Nama (11), !Ora (13), Xû (29), N|amani (38), Southern Sotho (1158), and Swati (1162). In these 7 cases, we removed the aspiration distinction from phonemes in consonant fields other than clicks, but left the distinction as reported in the database for clicks.

In summary, removing non-phonemic distinctions as detailed in this section resulted in 728 phonemes occurring in at least one language across 5736 languages.

1.4 Encoding Ruhlen phonemes using Unicode

We encoded all phonemes found in the Ruhlen database in Unicode form in our files archived at PNAS (Dataset S2). Because the Ruhlen database was originally encoded using Microsoft Notepad, Unicode encodings were translated to utf-8 from Notepad's code points for IPA. To map Notepad's code points to Unicode IPA (see http://www.utf8-chartable.de/unicode-utf8-table.pl for details on code points and names of code points), we had to make changes in the encoding of 5 characters (Table S7) but did not alter the raw data otherwise.

1.5 Annotating Ruhlen languages with speaker population sizes and geographic coordinates

Where possible, each language in the Ruhlen database was annotated with an ISO 639-3 code (a three-letter code for each language that is set by the International Organization for Standardization) and an ISO 3166-1 alpha 3 code (a three-letter code for each country) corresponding to an entry in the Ethnologue database [1]. Also using the Ethnologue database, we annotated the Ruhlen database with geographic coordinates (latitude and longitude) and speaker population size estimates.

To avoid mismatches, we annotated our database by hand with information from the Ethnologue. For a language *l* in the Ruhlen database, we searched the Ethnologue for a corresponding language that matched *l* in name (or alternate/dialect name) as well as country (or geographic region) or language family classification; if both country (geographic region) and language family classification were available, both were used for matching to the Ethnologue. Additional alternate language names and alternate spellings could be found in the titles of the sources listed in both the Ruhlen database and the Ethnologue; when necessary, these were used for matching. The Ethnologue tends to have a separate language entry for each nation where a language is spoken. In such cases, we matched Ruhlen language *l* to the Ethnologue entry located in *l*'s country as specified in the Ruhlen database.

The Ruhlen database often provides more detailed location information than the country in which the language is spoken (e.g., states, provinces, islands, and regions where language *l* is spoken). The Ethnologue database contains maps of many countries with language locations plotted. When more detailed location information was given in the Ruhlen database, we compared this to maps in Ethnologue to confirm language matches or choose the correct match.

If the Ruhlen database had ambiguous or missing location information for a language, we could often match the language to an Ethnologue entry based on name, classification, and/or sources. When such a language is spoken in multiple countries, the Ethnologue entry corresponding to the country with the largest speaker population was selected.

If a language from the Ruhlen database was divided into several dialects in the Ethnologue, we chose the dialect that most closely matched the geographic location given for the language in the Ruhlen database. If the location did not resolve ambiguity between dialects, the dialect with the largest population size was selected. For example, if the Ruhlen database contained one entry for a language spoken on an island, and the Ethnologue splits this language into two dialects on the same island, the Ruhlen entry was matched to the dialect entry in the Ethnologue with the larger speaker population. Similarly, when the Ruhlen database provides entries for multiple dialects of the same language, but only one overarching language entry was given in the Ethnologue, each dialect in the Ruhlen database was matched with the same Ethnologue entry. Note that this type of redundant matching only influences analyses that use geographic coordinates and speaker population sizes (see also Section 3.2.4).

Using the annotation process detailed here, we were able to match 4189 languages of the 5736 Ruhlen entries with ISO 639-3 codes; each of these codes has a corresponding Ethnologue entry with geographic and speaker population size information. Over 97% of the 1547 unmatched Ruhlen entries had no phonemic data.

1.6 Filtering of languages for analysis

Of the 5736 languages in the Ruhlen database, 3508 languages did not have both consonant and vowel data and were therefore excluded from this analysis. Dataset S3 summarizes the presence of phoneme data for all languages, and which languages have sources, in the Ruhlen database.

From the 2228 languages with phonemic data, we removed 4 languages (Ruhlen language numbers: 1548, 1549, 1708, and 2686) from analysis due to incomplete phonemic data. Entries 1548 and 1549 contained information on vowel harmony, but no other vowel data. Entry 1708 had modified consonant data and no other consonant data. Of all consonant fields, entry 2686 only had glides. In addition, two languages, numbers 2681 and 2523, were excluded from further analysis because neither language had sources listed for the phonemic data. (Since researchers might categorize phonemes differently, we only include a language in our analyses if the source of the typological data is referenced.) This left 2222 languages with sources and complete phonemic data for analysis; a spreadsheet indicating the presence of sources and phonemic data in the Ruhlen database is archived at PNAS (Dataset S3).

We then excluded proto-languages, invented languages, pidgins, and creoles from our analysis; this filter removed 74 languages with phonemic data, leaving 2148

languages for analysis. We then excluded 2 languages — Margi (number 1395) and Yele (number 4332) — due to more recent research, which raised concerns that the labial-alveolar double articulations reported in both of these languages might be more accurately transcribed as sequences of phonemes [2, 3].

We matched all but 64 of the remaining 2146 languages to entries in the Ethnologue database; the Ruhlen entries remaining unmatched to Ethnologue entries were excluded. Our final dataset for analysis contained 2082 languages with a complete set of the following information: geographic coordinates and speaker population sizes from Ethnologue, sources for data reported in the Ruhlen database, and phonemic data for 728 phonemes. These data are archived at PNAS (Dataset S1).

1.7 Processing PHOnetics Information Base and LExicon (PHOIBLE)

Data analyzed in Moran [4] and Moran *et al*. [5] is the basis for PHOIBLE and can be accessed at http://phoible.org. The data analyzed in Moran *et al*. is labeled "Phoneme level supplemental data" (`MoranEtAl2012_phonemeData.tab`) and the data analyzed in Moran is labeled "PHOIBLE phoneme level MySQL dump (XML)" (`Moran2012_phonemeData.xml`); we analyzed the latter file. There are two differences between the PHOIBLE MySQL dump and the Phoneme level supplemental data. In the XML version we analyzed, phoneme t͡ʃʰ was absent from Korean, and two ISO codes have changed (`moq` is used instead of `mhz` for the language Mor in Indonesia, and `yue` is used instead of `shn` for the language Cantonese in China; these changes are supported by the sources listed in the original UPSID entries). `MoranEtAl2012_phonemeData.tab` was accessed at http://phoible.org/download, and `Moran2012_phonemeData.xml` was accessed at https://github.com/clld/phoible/blob/master/phoible/static/data/Moran2012_phonemeData.xml.

We made the following adjustments when processing PHOIBLE:

1) We included the phoneme t͡ʃʰ in the Korean inventory.
2) For Cantonese, PHOIBLE has two inventories, inventory_id 19 (inventory from SPA/Crothers) and 642 (inventory from UPSID). In the file `Moran2012_phonemeData.xml`, the ISO code for both of these inventories is `yue`. Using PHOIBLE's hierarchy for choosing which inventory to report when multiple sources contained phoneme data for a language, as detailed in Moran [4], we eliminated the UPSID entry for the language Cantonese.
3) We associated each PHOIBLE entry with a corresponding Ethnologue entry using the ISO codes provided as "language_code_id" in the file `Moran2012_phonemeData.xml`. We then used Ethnologue's geographic coordinates and population size estimates for languages unless otherwise noted.
4) PHOIBLE labels Norwegian (PHOIBLE inventory_id 159) with the ISO code `nob`, which is not an Ethnologue code. For population size and geographic coordinates in our analyses, we use the Ethnologue code `nor` for Norwegian. For more details on codes `nob` and `nor`, see http://www-01.sil.org/iso639-

3/documentation.asp?id=nob and http://www-01.sil.org/iso639-3/documentation.asp?id=nor.

5) The language Sumo in PHOIBLE is associated with the ISO code `ulw`, which is a new code that replaces `sum`. Since this change occurred after the Ethnologue vol. 16 [1] was published, we lacked the Ethnologue location and population size estimates for this language. As a result, we used the location listed in PHOIBLE for Sumo instead of the Ethnologue location.
6) We removed tones when analyzing PHOIBLE, since the Ruhlen database did not have consistent information for tones in language records.

PHOIBLE contains phoneme inventories for 968 languages (once we removed the duplicate Cantonese entry); these languages were included in subsequent analyses.

The encoding of modifications is a difference between the Ruhlen database and PHOIBLE: in PHOIBLE, each modified phoneme is encoded individually, whereas in the Ruhlen database, modifications are encoded separately (Section 1.2). This difference in encoding introduces a discrepancy in number of phonemes between the two databases. In total, 728 distinct phonemes (including modifications) were observed across the 5736 languages in the Ruhlen database. In PHOIBLE, 1587 phonemes were observed across 968 languages.

1.8 Annotating PHOIBLE languages with speaker population sizes and geographic coordinates

PHOIBLE provides an ISO code with each language. These ISO codes matched codes used by the Ethnologue database for all languages except as mentioned above for Norwegian and Sumo languages. With these two modifications, we annotated each PHOIBLE entry with a speaker population size estimate and geographic coordinates using the Ethnologue entry with the corresponding ISO code.

1.9 Generating binary matrices for both databases

We converted each database into separate presence/absence matrices for data analysis. Element $A_{i,j}$ in each presence/absence matrix indicates the presence (1) or absence (0) of the j^{th} phoneme in the i^{th} language. The PHOIBLE matrix we generated indicates the presence or absence of 1587 phonemes in 968 languages. The Ruhlen matrix has dimension 2082 (number of languages analyzed in the Ruhlen database) by 728 (the number of observed phonemes across all languages). The Ruhlen presence/absence matrix (Dataset S1) and a file with corresponding column labels (Dataset S2) is archived at PNAS.

1.10 Defining geographic regions

To compare our analyses of these linguistic databases with previous genetic studies that separated worldwide samples into geographic regions (e.g. Pemberton *et al.* [6], Ramachandran and Rosenberg [7]), we defined regions for languages in the Ruhlen database and PHOIBLE using the United Nations geoscheme (http://unstats.un.org/unsd/methods/m49/m49regin.htm) applied to the Ethnologue

location for each language. We grouped each language into one of the following regions: Middle East (UN "Western Asia" region plus Egypt), Central/South Asia (UN "Central Asia" plus "Southern Asia"), East Asia (UN "Eastern Asia," "Southeastern Asia," and the portion of Russia that is east of the Ural Mountains), Africa (minus Egypt), Europe (including the portion of Russia that is west of the Ural Mountains), Oceania, North America, and Central/South America (as in Fig. 1). The Ethnologue records the country and latitude/longitude point locations for each language; for Russian (located in Russia), the range of the language in question spans two geographic regions (Europe and East Asia). We assigned Russian to Europe due to higher population density of speakers, but the latitude/longitude point location from the Ethnologue appears east of the Ural Mountains.

1.11 Population size and phoneme inventory size

We calculated the correlation between population size and phoneme inventory size for languages in the Ruhlen database and PHOIBLE, and repeated the analysis for languages within Africa, the Americas, Asia, Europe, and Oceania. Overall in the Ruhlen database, population size explains little of the observed variation in phoneme inventory size (r = 0.1299, Fig. S9). In fact, within each tested region except Asia, phoneme inventory size and speaker population size are either uncorrelated or *negatively* correlated (Fig. S9). In PHOIBLE, population size explains slightly more of the observed variation in phoneme inventory size (r = 0.2724, Fig. S9); once again, within each tested region except Asia, phoneme inventory size and speaker population size are either uncorrelated or negatively correlated (Fig. S9).

2. Merging of genetic data with linguistic data

2.1 Microsatellite data

We analyzed a dataset of microsatellite markers that combined data from several studies; the merging of data is described in Pemberton et al. [6]. We used two datasets from Pemberton et al. [6]: (*i*) MS5339, which has genotype data from 246 loci, was used to generate the rooted tree used in phylogenetic analyses; (*ii*) MS5255, which has genotype data from 645 loci, was used for all other analyses. We excluded the Dogon population from our analyses since it was noted that the samples were of lower quality (the sample size was 3 and average missingness was 21.6% across 645 microsatellite genotypes; see also Supplemental Material of [8]). We also excluded data from admixed populations as identified by Pemberton et al. [6] and excluded the Australian population due to missing sampling location information. In total, microsatellite markers from 246 human populations were included in our analyses.

We tested for outlier individuals by generating a matrix of individuals by alleles. A column was assigned for each unique allele of each marker such that matrix entry $A_{i,j}$ was assigned a value of 0, 1, or 2 based on the number of copies of allele j sampled from individual i. We then performed a principal components analysis (PCA) on this matrix and recorded the scores for the first four principal components (PCs) for each individual.

An individual with a score more than six standard deviations from the mean of any of the first four PCs was considered an outlier. None of the individuals met these criteria, so all individuals (except those in excluded populations mentioned above) were considered for further population-level analyses.

2.2 Matching genetic populations to Ruhlen database languages

Using the population names and locations reported by Pemberton et al. [6], we matched as many genetically sampled populations as possible to their native language in the Ruhlen database. When genetic studies provided linguistic information (e.g., Table S1 of Tishkoff *et al*. [8] reported that the San individuals sampled were speakers of the Qxû language), we used this information to match languages to genetic populations. Additional information on HGDP-CEPH samples and populations is available using http://alfred.med.yale.edu [9], including some linguistic information or more specific collection locations for HGDP-CEPH populations.

We assigned 147 languages to populations when there was an exact or nearly exact match between genetic population name and language name (or alternate name) in the Ruhlen database. For genetic populations that remained unmatched, we used the Ethnologue and a literature search to determine whether the population name was associated with a single language.

When a population could not be matched by name alone — for example, when several dialects of a language were present in the Ruhlen database but the genetic population name did not specify a dialect — we consulted Ethnologue's language maps at the latitude and longitude of the genetic data collection site to determine whether a single language assignment could be resolved.

In summary, out of 246 populations for which we assembled microsatellite data for phylogenetic analysis [6], we matched 203 populations to individual language entries in the Ruhlen database. (For another four populations of the 44 unmatched to a Ruhlen-database language, we did not have enough information to assign the population to a single dialect.) Of the 203 populations that matched to a single Ruhlen entry, consonant and vowel data (and sources, Ethnologue locations, and Ethnologue speaker population sizes) were available for 139 populations. The phonemic and genomic data for these populations constitute the phoneme–genome dataset.

3. Statistical analyses of linguistic and genetic data

3.1 Correlations between phonemic, genetic, and geographic distance

For both the Ruhlen database and PHOIBLE, we assembled pairwise matrices of geographic distance (great-circle distance with waypoints, as detailed in section 3.2 and [10]) and phonemic distance (Jaccard [11] and Hamming [12]). For a pair of languages, Jaccard distance equaled the number of phonemic differences divided by the number of phonemes present in at least one of the two languages, and Hamming distance equaled the number of phonemic differences divided by the total number of phonemes in the database. For both databases, a Mantel test [13, 14] comparing phonemic distance to geographic distance was significantly different from zero for both the full set of

populations and the phoneme–genome subset of populations (Tables S1, S2). Similarly, a Mantel test comparing genetic distance to geographic distance was significantly different from zero for both phoneme–genome datasets. These results suggest that both genetic variation and phonemic variation are significantly spatially autocorrelated.

With three distance matrices—here, genetic (allele-sharing) distance, phonemic distance, and geographic distance for the phoneme–genome datasets—partial Mantel tests [15] can give some insight into the possible causal relationships among the three matrices (as in [14]; Fig. S2). The Mantel and partial Mantel correlations between genetic distance, geographic distance, and phonemic distance were consistent with a model in which geographic distance between populations is causally linked to both genetic distance and phonemic distance (Fig. S2). Legendre presents "four possible models of causal relationships involving three matrices, in terms of the expected results of the simple and partial Mantel tests" [14]. In our analysis of genetic distance, geographic distance, and phonemic distance, Mantel and partial Mantel results were best represented by the bottom-left model in Fig. S2A. The spatial dispersal of populations via migration and isolation by distance can lead to geographic structure in both genes and languages; beyond any common signatures in genes and languages due to this spatial structuring, genetic distance correlated with phonemic distance.

Spatial autocorrelation analysis can also be used to predict the range of distances over which two variables are correlated by partitioning the geographic distance matrix into distance classes [16, 17, 18]. In this way, a strong signal of spatial autocorrelation over short distances can be distinguished from spatial autocorrelation over longer distances. We first partitioned the pairwise geographic distance matrix into 1000 km distance classes. Distances ≥ 0 km and <1000 km were assigned to distance class 1, ≥ 1000 km and <2000 km were assigned to distance class 2, and so on. We performed Mantel tests to compare this matrix of distance classes to both genetic and phonemic distances. We then increased the distance class size to 25,000 km in 1000 km increments and repeated the Mantel tests for each distance class size. We found that genetic distance showed significant spatial autocorrelation for all tested distance classes: genetic distance is correlated with geographic distance on a worldwide scale (Fig. S2B). However, phonemes were more similar among languages in the same distance class only within a range of ~10,000 km (Fig. S2B). Beyond this distance, the signal of spatial autocorrelation was not significant. In other words, beyond 10,000 km, phoneme inventories within one distance class were not more similar to one another than to those in another distance class.

3.2 Regression analyses using individual languages

As stated in the main text, we performed regressions of phoneme inventory size from both the Ruhlen and PHOIBLE databases on geographic distance from a center (Figs. 4A-B, S7) using each of 4210 centers drawn from the surface of the earth as described in the Methods of Ramachandran *et al*. [10]. Geographic distances between languages and each center were calculated using obligatory waypoints as in Ramachandran et al. [10], Wang et al. [19], and Ramachandran and Rosenberg [7]. These waypoints are: Anadyr, Russia (64°N, 177°E); Cairo, Egypt (30°N, 31°E); Istanbul, Turkey (41°N, 28°E); Phnom Penh, Cambodia (11°N, 104°E); Prince Rupert, Canada (54°N, 130°W); and Panama City, Panama (8.967°N, 79.533°W). When calculating geographic distances to centers

from certain isolated islands (or groups of islands), using the geographic region assigned by the UN geoscheme would have led us to calculate a putative distance of migration that was very different from the path humans took to get to these locations. These locations were: Hawaii, Malaysia, Indonesia, Madagascar, Philippines, Easter Island (which we classified as "Oceania" for our calculations) and the Falkland Islands (which we classified as part of South America).

Using the Ruhlen database, we also performed regressions using phoneme inventory size, excluding modifications, clicks, and modifications and clicks (Table S8). Modifications are different from other phonemes in the Ruhlen database (Table S6): they are differences in the way a sound is produced that can be applied to multiple phonemes in a language. Modifications were encoded as separate phonemes in Ruhlen, so we tested their impact on the regressions by repeating these tests without them. Clicks only occur in 39 of the 2082 languages used in this analysis; only two of these languages are outside Africa. Excluding clicks and modifications constrained the dataset to phonemes that represent sounds themselves (as opposed to modifications of sounds) and that are not biased toward a specific geographic region (as clicks are). For PHOIBLE, an analysis without modifications was not performed because PHOIBLE encodes modifications differently: each modified phoneme is listed individually in a phoneme inventory.

The results of these regression analyses are shown in Figs. 4, S7, S12, S13 and Tables S5 and S8.

3.2.1 Model selection

In the Ruhlen database, we fit regression models for the 2004 languages that had Ethnologue speaker population sizes greater than 0. We use Akaike's Information Criterion (*AIC*) for model selection. $AIC = -2\ln\left(L\left(\hat{\theta}\middle|Y\right)\right) + 2K$, where $L(\hat{\theta}|Y)$ is the likelihood of the estimated parameters given the data Y and K is the number of estimable parameters in the model [20]. Here, K is the number of regression coefficients plus two, to account for the constant and the residual sum of squares. *AIC* can only be used to compare estimated models when the numerical values of the dependent variable are identical. We can use *AIC* in our analyses to find the origin out of the 4210 tested with most support for a dependent variable (Fig. S12). Note, however, that we must use another measure (e.g., a correlation coefficient) to compare the fit of the best models (e.g., Figs. 4, S7, S12).

To facilitate comparisons with past studies, we also conduct model selection using the Bayesian Information Criterion (*BIC*; Fig. S13). Previous studies have used a wider threshold of four *BIC* units for model selection [e.g. 21, 22, based on 23]. Using this more conservative threshold, we find model selection using *BIC* is equivalent to that using *AIC* (Figs. S12, S13); this is because *BIC* and *AIC*, when calculated for the same dependent variable and dataset, will differ by a constant value. $BIC = -2\ln\left(L\left(\hat{\theta}\middle|Y\right)\right) + K\ln(n)$, where n is the length of the vector Y. For a simple linear regression fit using the Ruhlen database, $BIC - AIC = \left[-2\ln\left(L\left(\hat{\theta}\middle|Y\right)\right) + K\ln(n)\right] - \left[-2\ln\left(L\left(\hat{\theta}\middle|Y\right)\right) + 2K\right] = K\ln(n) - 2K = 3\ln(2004) - 6 = 16.81$.

We also tested whether incorporating speaker population size into the regression model significantly improved prediction of phoneme inventory size. Our null hypothesis is that the model with fewer parameters is the appropriate model for the observed data. Given two models M_0 and M_1, where M_0 is nested within M_1 and M_1 only has one more parameter than M_0, the test statistic,

$$\frac{[(Residual\ SS\ M_0) - (Residual\ SS\ M_1)]}{\left[\frac{(Residual\ SS\ M_1)}{Residual\ df\ M_1}\right]} \sim F_{1,residual\ df\ M_1}$$

where *SS* denotes "sum of squares" and *df* denotes "degrees of freedom". If the ratio is large, then we reject the null hypothesis and conclude that the additional parameter in M_1 significantly improves the model fit.

3.2.2 Jackknifing over geographic regions

Fixing total phoneme inventory size as the dependent variable, we tested the sensitivity of the most-supported origin when each geographic region was excluded from analysis via jackknifing for each database (Table S9).

Of the eight geographic regions in this analysis, the lowest-*AIC* origin was in northern Europe when each geographic region was removed except Central/South Asia (Ruhlen), North/Central America (Ruhlen), and Oceania (both datasets) (Table S9). This reflects that the regression of phoneme inventory size on geographic distance to centers is strongly influenced by the low phoneme inventory sizes observed in both datasets in languages in North/Central America and Oceania.

3.2.3 Including number of neighbors in regression analyses

We also fit regressions for all 4210 centers on land that further included number of neighbors (within either a 100 km radius or a 250 km radius) as an independent variable. Number of neighbors did not significantly improve model fit for the PHOIBLE database phoneme inventory sizes ($p > 0.20$). In the Ruhlen database, including number of neighbors did significantly improve model fit (100 km radius: $p = 0.008259$; 250 km radius: $p = 0.0005355$).

There is no statistical support in either linguistic database for including an interaction term between number of neighbors and Ethnologue speaker population size to improve model fit; for both databases, the interaction term regression coefficient was statistically equivalent to 0 (Ruhlen $p > 0.7$; PHOIBLE $p > 0.4$).

3.2.4 Sensitivity of regression results to multiple matches to the same ISO code in the Ruhlen database

Some languages in the Ruhlen database are matched to the same ISO code (Fig. S14). Languages matched to the same ISO code have the same speaker population sizes and geographic coordinates in our analyses. To see whether this biased any of our regression analyses, we generated 100 replicate samples where just one language was sampled for each ISO code in the Ruhlen database. Across the 100 replicate samples, the origin with

minimum AIC across 4210 centers was (67.6684, 36.2) — the same lowest-AIC origin when analyzing the full Ruhlen database — and the correlation in AIC and r^2 across all models fit compared to the full dataset is >0.99999 for all 100 replicate samples. We conclude that there is no effect of multiple languages being mapped to the same ISO code on any regression results presented here.

3.3 Allele and phoneme frequency analyses

For 645 human microsatellite loci [6], we calculated allele frequencies for each of 246 non-admixed populations genotyped at these markers. Pooling populations within the eight geographic regions (as described in 1.10), we generated a histogram of these allele frequencies with 20 bins. Population allele frequencies of 0 were excluded from this allele frequency spectrum (Fig. S15). For each of the 728 phonemes catalogued across the 2082 languages analyzed here, we calculated the phoneme's frequency in a geographic region. As above, we pooled languages within a geographic region and generated a histogram of phoneme frequencies with 20 bins, excluding phonemes never observed in that region. Thus, the smallest bin of these frequency spectra (frequency<0.05) does not include alleles or phonemes with frequency equal to zero (Fig. S15). The phoneme frequency spectrum is similar to the allele frequency spectrum in that most phonemes and alleles occur at frequencies less than 5% (Fig. S15). This is characteristic of neutral genes as opposed to those under selection. We repeated this analysis for the 968 languages in PHOIBLE (Fig. S15)

Supporting Information: Figure Captions

Fig. S1. Procrustes analysis of phonemes. In the Procrustes analysis, the principal component scores are rotated, scaled, and translated to minimize the sum of the squared Euclidean distances between PC scores and corresponding geographical points [24]. (**A**) We conducted a principal components analysis (PCA) of phonemic data in the Ruhlen database, and we Procrustes-transformed the scores of the first two principal components onto the geographic location for each language. An empirical *p*-value was calculated after 100,000 permutations. (**B**) The same analysis was performed with PHOIBLE.

Fig. S2. Spatial autocorrelation of genes and languages. (**A**) Figure drawn after Legendre 1993, Fig. 6 [14]. Here, **AB** represents the Mantel correlation between matrices **A** and **B**, **AB•C** represents the partial Mantel correlation between matrices **A** and **B** controlling for matrix **C**, and **AB**=0 indicates that the calculated Mantel statistic between matrices **A** and **B** is not significantly different from zero. When all Mantel and partial Mantel results were considered, the bottom left model was the most consistent with our observations, where **A** represents geographic distance, **B** represents genetic distance, and **C** represents phonemic distance. (**B**) To determine the distance over which spatial correlation is evident for genetic distance and phonemic distance, we partitioned the geographic distance matrix into classes. The *x*-axis represents distance class size; for a distance class size of 1000 km, geographic distances ≥ 0 km and <1000 km were assigned to distance class 1, ≥1000 km and <2000 km were assigned to distance class 2, and so on. Using Mantel tests, we compared the distance class matrices to both genetic and phonemic distances for both phoneme–genome datasets (Ruhlen and PHOIBLE). The significance threshold is indicated by a red dashed line. We found that genetic distance showed significant spatial autocorrelation for all tested distance classes (blue dots). However, phonemic distance was correlated with geographic distance within a range of ~10,000 km (black dots). Beyond this distance, the signal of spatial autocorrelation was not significant.

Fig. S3. Axes of phonemic and genetic differentiation. (**A**) Comparison of axes of greatest phonemic differentiation as predicted by 2082 languages in the Ruhlen database (black arrows) and 968 languages in PHOIBLE (dark red arrows). All arrows represent significant associations between phonemic distance and geographic distance in the direction indicated by the arrow. (**B**) Comparison of axes of greatest phonemic differentiation (black arrows) with axes of greatest genetic differentiation (gray dashed arrows) for 114 populations in the PHOIBLE phoneme–genome dataset. In both panels, arrows are scaled to the number of populations compared within each region. Thinner arrows indicate associations that were not statistically significant.

Fig. S4. The effect of geographic isolation on phonemes within regions. Populations in each region were separated into two groups according to their number of neighboring languages: less than or equal to the median number of neighbors and greater than the median number of neighbors. We then compared the languages in these two groups based on their number of phonemes and their phonemic distance to their neighbors. Statistical significance (Wilcoxon rank-sum test) is indicated by bold lines. (**A**) In the Ruhlen

database, languages with fewer neighbors had significantly more phonemes in East Asia and Oceania and significantly fewer phonemes in Europe. (**B**) In PHOIBLE, languages with fewer neighbors had significantly more phonemes in East Asia and Oceania. (**C**) In the Ruhlen database, languages with fewer neighbors had significantly greater phonemic distance to those neighbors in Africa, East Asia, and Oceania (within certain radii). Languages with fewer neighbors had significantly smaller phonemic distance to those neighbors in Europe and North America. (**D**) In PHOIBLE, languages with fewer neighbors had significantly greater phonemic distance to those neighbors in Africa, East Asia, and Central/South Asia. Languages with fewer neighbors had significantly smaller phonemic distance to those neighbors in North America.

Fig. S5. The effect of geographic isolation on phonemic distance to neighboring languages. For all radii greater than 200 km, phonemic (Hamming) distance was significantly greater (Wilcoxon $p < 2.4\times10^{-5}$) for languages with fewer neighboring languages (less than or equal to the median number of neighbors) than for languages with more neighbors (greater than the median number of neighbors). For all radii, the variance in phonemic distance was also significantly greater for languages with fewer neighbors (Ansari-Bradley $p < 1.4\times10^{-13}$). Dashed lines indicate the mean phonemic distance for languages with less than or equal to the median number of neighbors, and dotted lines indicate mean phonemic distance for languages with greater than the median number of neighbors at the indicated radii. Black lines indicate the Ruhlen database and red lines indicate PHOIBLE. Since distance measures require at least two languages for comparison, lines begin at the first radius where the median number of neighbors was at least two so that pairwise comparisons were possible for languages with less than or equal to the median number of neighbors. Inset boxplots show the distributions of phonemic distance values for languages with fewer and more neighbors; Wilcoxon *p*-values are indicated.

Fig. S6. The effect of geographic isolation on phoneme variance within regions. Panels are similar to Figure S4, with statistical significance (Ansari-Bradley test) indicated by bold lines. (**A**) In the Ruhlen database, languages with fewer neighbors had significantly greater variance in phoneme inventory sizes in Africa, North America and Oceania and significantly less variance in Europe. (**B**) In PHOIBLE, languages with fewer neighbors had significantly more phonemes in Africa, North America, and South America. (**C**) In the Ruhlen database, languages with fewer neighbors had significantly more variance in phonemic distance to those neighbors in Africa, East Asia, Central/South Asia, North America, and Oceania. (**D**) In PHOIBLE, languages with fewer neighbors had significantly greater phonemic distance to those neighbors in Africa, East Asia, Central/South Asia, North America (within certain radii), and South America.

Fig. S7. Best-fit linear regressions of phoneme inventory size on geographic distance. As in Fig. 4, we estimated linear decrease in number of phonemes with distance to 4210 geographic centers on the Earth. (**A**) Regression from the best-fit geographic center for language families in the Ruhlen database, using the median number of phonemes within each family. The best-fit geographic center remained in northern Europe when languages were grouped by language family classifications. (**B**) Regression from the best-fit

geographic center for languages in PHOIBLE. As for the Ruhlen database (Fig. 4A), the best-fit geographic center was located in northern Europe.

Fig. S8. Overlap in predicted ancestral phoneme inventories. For phoneme inventories from the Ruhlen database and PHOIBLE, as well as genetic, geographic, and cognate-based linguistic trees, we used an ancestral character estimation algorithm to estimate the phoneme inventories of the ancestor to Romance languages and Indo-Aryan languages. For comparison, we used published phoneme inventories for Vulgar Latin [25, 26] and Vedic Sanskrit [27] to approximate the phoneme inventories ancestral to Romance languages and Indo-Aryan languages, respectively. We then calculated the overlap between our predictions by dividing the number of phonemes in the published inventory whose ancestral presence was correctly predicted with both trees by the number of phonemes in the published inventory correctly predicted with at least one tree. For each comparison, the percent overlap is given first for the Ruhlen database, then PHOIBLE.

Fig. S9. Population size and phoneme inventory size. Scatterplots of number of phonemes against $\log_{10}$(population size) for 2004 languages worldwide (with Ethnologue speaker population size > 0) in the Ruhlen database (left panels) and 967 languages (with speaker population size > 0) in PHOIBLE (right panels). Within-region correlations are shown for Africa, Americas, Asia, Europe, and Oceania. Linear regression lines are shown in black, and the correlation coefficient and *p*-values are displayed on each plot. Whereas the slope of the regression line is weakly positive for the plot containing all languages in each database, the slope is not significantly different from zero or negative for all individual geographic regions considered except for Asia; this pattern exists for both the Ruhlen database and PHOIBLE.

Fig. S10. Correlation of phoneme inventory sizes between databases. (**A**) Of the 968 languages in PHOIBLE, 621 could be matched to languages in the Ruhlen database. The phoneme inventory sizes in these languages show a correlation of r=0.71 ($p = 4.9\times10^{-94}$). (**B**) PHOIBLE synthesizes several databases of phonemes, including the Stanford Phonology Archive (SPA) and UPSID. For 165 languages in PHOIBLE, data from both SPA and UPSID were available; the phoneme inventory sizes in these languages show a correlation of r=0.79 ($p = 1.24\times10^{-36}$).

Fig. S11. Synthetic maps of phoneme principal components. The first ten principal components explained 41.25% of the variance in the RUHLEN database, and the first ten principal components explained 34.66% of the variance in PHOIBLE. (**A**) The first principal component scores for languages in the Ruhlen database are represented by color (indicated on the color bar). The second principal component scores for languages in the Ruhlen database (**B**), as well as the first (**C**) and second (**D**) principal component scores for languages in PHOIBLE, are similarly depicted.

Fig. S12. Model selection based on *AIC* across 4210 centers for linear regressions of phoneme inventory size on geographic distance. The color of each of the 4210 locations (shown as filled circles) indicates either an *AIC* value (see gradient on the right)

or a point where there was no statistical support for a linear relationship between phoneme inventory size and geographic distance to the location (points shown in grey, indicated by "n.s." (not significant) on gradients to the right). In panels **A** and **C**, results of simple linear regressions are shown. In panels **B** and **D**, results of multiple linear regressions, for which independent variables are geographic distance to the center and base-10 logarithm of speaker population size, are shown. In all panels points denoted as "n.s." did not have statistical support for the regression coefficient of geographic distance to the center being different from zero after Bonferroni correction across 4210 tests.

Points with *AIC* within 2 units of the minimum *AIC* observed in each panel are shown in black; these models are considered to have equivalent support to the model with lowest observed *AIC*. In panels C and D, there are no points with models whose *AIC* fall within 2 units of the minimum observed *AIC*. (**A**) The Ruhlen database, simple linear regression. The point with most support is (67.6684, 36.2); seven points have models with equivalent support (shown as filled black circles). 34.80% of points fall in the n.s. category. (**B**) Ruhlen, multiple linear regression. The point with most support is (64.1581, 34.4); nine points have models with equivalent support (shown as filled black circles). 35.8% of points fall in the n.s. category. (**C**) PHOIBLE, simple linear regression. The point with most support is (77.1614, 16.4); no other points have *AIC* within 2 units of the minimum observed *AIC*. 41.40% of points fall in the n.s. category. (**D**) PHOIBLE, multiple linear regression. The point with most support is (77.1614, 16.4); no other points have *AIC* within 2 units of the minimum observed *AIC*. 45.2% of points fall in the n.s. category.

Fig. S13. Model selection based on the Bayesian Information Criterion (BIC) across 4210 centers for linear regressions of phoneme inventory size on geographic distance. Similar to Figure S12, but here the point with the lowest BIC value is shown as an open circle in each panel with a dotted line indicating "lowest BIC origin". Points with BIC within 4 units of the minimum BIC observed in each panel are shown in black; these models are considered to have equivalent support to the model with lowest observed BIC [21, 33, 23]. In panel **D**, there are no points with models whose BIC fall within 4 units of the minimum observed BIC. (**A**) The Ruhlen database, simple linear regression. The point with most support is (67.6684, 36.2); 15 points have models with equivalent support (shown as filled black circles). 34.80% of points fall in the n.s. category. (**B**) Ruhlen, multiple linear regression. The point with most support is (64.1581, 34.4); 16 points have models with equivalent support (shown as filled black circles). 35.8% of points fall in the n.s. category. (**C**) PHOIBLE, simple linear regression. The point with most support is (77.1614, 16.4); four points have models with equivalent support (shown as filled black circles). 41.40% of points fall in the n.s. category. (**D**) PHOIBLE, multiple linear regression. The point with most support is (77.1614, 16.4). There are no other points with models whose BIC fall within 4 units of the minimum observed BIC. 45.2% of points fall in the n.s. category.

Fig. S14. The distribution of number of languages mapped to the same ISO code in the Ruhlen database. There are 126 ISO codes that have multiple languages in the Ruhlen database matched to them. The mode of the distribution is 2.

Fig. S15. Allele and phoneme frequency spectra. The allele frequency spectrum (**A**) and phoneme frequency spectrum of the Ruhlen database (**B**) and PHOIBLE (**C**) show a high proportion of both alleles and phonemes at low frequency (greater than 0 but less than 0.05).

References

1. Lewis MP (2009) *Ethnologue: Languages of the world, Vol 16*. (SIL International, Dallas). Online: http://www.ethnologue.com

2. Ladefoged P, Maddieson I (1996) *The Sounds of the World's Languages*, (Blackwell, Oxford).

3. Maddieson I (1983) The analysis of complex phonetic elements in Bura and the syllable. *Studies in African Linguistics* 14:285–310.

4. Moran S (2012) *Phonetics Information Base and Lexicon*. PhD thesis, University of Washington. Online: www.phoible.org

5. Moran S, McCloy D, Wright R (2012) Revisiting population size vs. phoneme inventory size. *Language* 88(4):877–893.

6. Pemberton TJ, DeGiorgio M, Rosenberg NA (2013) Population structure in a comprehensive genomic data set on human microsatellite variation. *G3* 3(5):891–907.

7. Ramachandran S, Rosenberg NA (2011) A test of the influence of continental axes of orientation on patterns of human gene flow. *Am J Phys Anthropol* 146:515–529.

8. Tishkoff SA, *et al*. (2009) The genetic structure and history of Africans and African Americans. *Science* 324(5930):1035–1044.

9. Rajeevan H *et al.* (2012) ALFRED: an allele frequency resource for research and teaching. *Nucleic Acids Res* 40:D1010–D1015.

10. 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 Nat Acad Sci USA* 102(44):15942–15947.

11. Jaccard P (1908) Nouvelles recherches sur la distribution florale. *Bull Soc Vaudoise Sci Nat* 44:223–270.

12. Hamming RW (1950) Error detecting and error correcting codes. *Bell Syst Tech J* 29(2):147–160.

13. Sokal RR (1979) Testing statistical significance of geographic variation patterns. *Syst Zool* 28(2):227–232.

14. Legendre P (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology* 74(6):1659–1673.

15. Smouse PE, Long JC, Sokal RR (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Syst Zool* 35(4):627–632.

16. Slatkin M, Arter H (1991) Spatial autocorrelation methods in population genetics. *Am Nat* 138(2):499–517.

17. Sokal RR, Jacquez GM, Wooten MC (1989) Spatial autocorrelation analysis of migration and selection. *Genetics* 121(4):845–855.

18. Peakall R, Ruibal M, Lindenmayer DB (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, *Rattus fuscipes*. *Evolution* 57(5):1182–1195.

19. Wang S, *et al*. (2007) Genetic variation and population structure in Native Americans. *PLoS Genet* 3(11):e185.

20. Burnham KP, Anderson DR (2010) *Model selection and multi-model Inference: A practical information-theoretic approach, ed. 2.* (Springer-Verlag, New York).

21. Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027):346–349.

22. Betti L, Balloux F, Amos W, Hanihara T, Manica A (2009) Distance from Africa, not climate, explains within-population phenotypic diversity in humans. *Proc Roy Soc B-Biol Sci*, 276(1658):809-814.

23. Burnham KP, Anderson DR (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociol Method Res*, 33(2), 261-304.

24. Wang C, *et al*. (2010) Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Molec Biol* 9:1–22.

25. Hall RA (1950) The reconstruction of Proto-Romance. *Language* 26(1):6–27.

26. Grandgent CH (1907) *An introduction to Vulgar Latin.* (DC Heath and Company, Boston).

27. Whitney WD (1879) *A Sanskrit grammar; including both the classical language, and the older dialects, of Veda and Brahmana.* (Breitkopf and Härtel, Leipzig).

Fig. S1

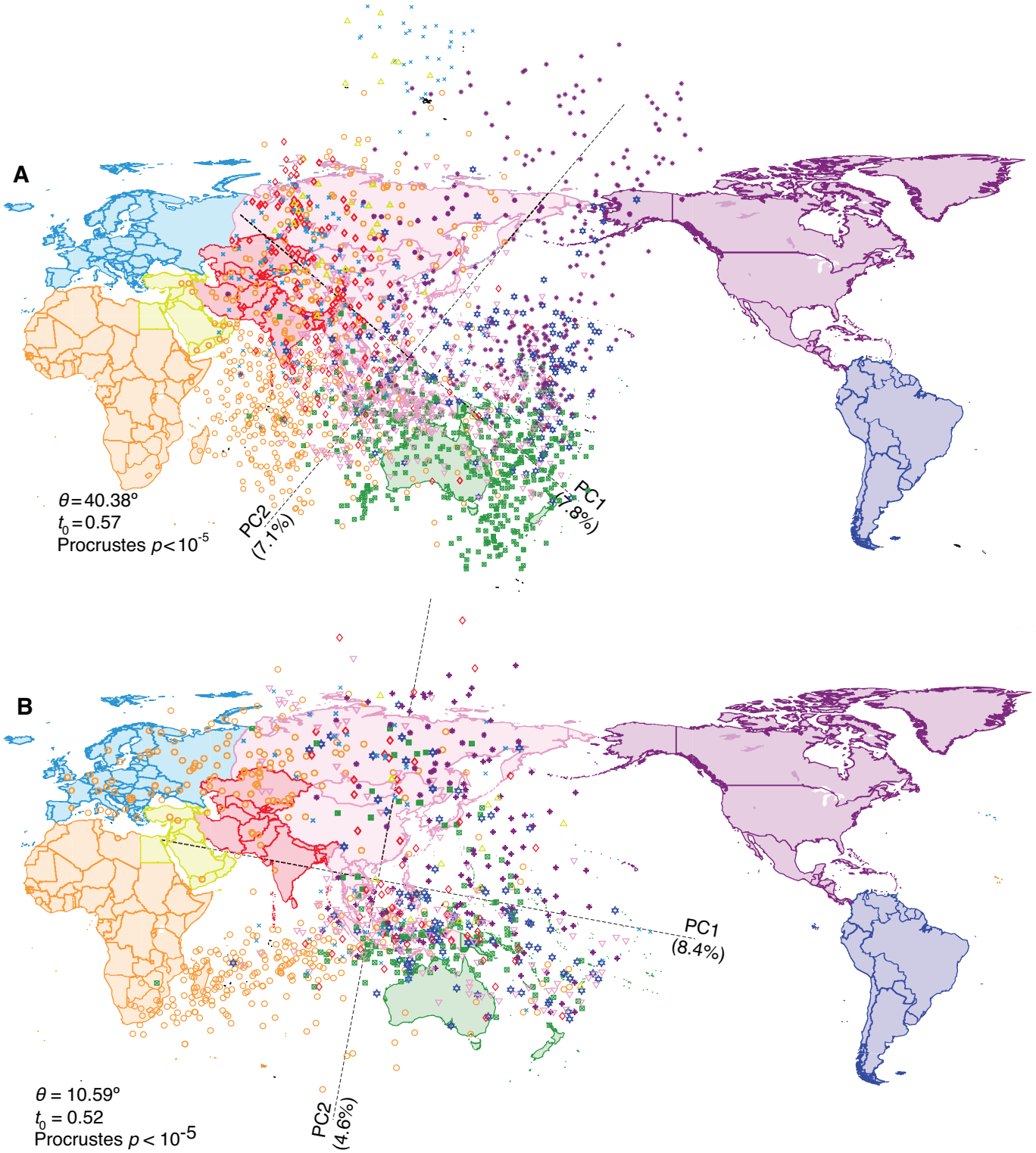
A
θ = 40.38°
t0 = 0.57
Procrustes p < 10-5
PC2 (7.1%)
PC1 (14.8%)
B
θ = 10.59°
t0 = 0.52
Procrustes p < 10-5
PC2 (4.6%)
PC1 (8.4%)

Fig. S2

A

Model	Expectations of the model
A → C ← B	$AC \neq 0$ $BC \neq 0$ $AB = 0$ $AB \cdot C \neq 0$ $BC \cdot A \neq 0$ $AC \cdot B \neq 0$ $AC \cdot B \geq AC$ $BC \cdot A \geq BC$
A → B → C	$AB \neq 0$ $BC \neq 0$ $AB \geq AC$ $AC \cdot B = 0$ $AB \cdot C \neq 0$ $BC \cdot A \neq 0$ $AB \cdot C \leq AB$ $BC \cdot A \leq BC$ $AB \times BC \approx AC$

All of our Mantel results are consistent with this model

Geography (A) → Genes (B), Languages (C)	$AB \neq 0$ $AC \neq 0$ $AB \cdot C \neq 0$ $AC \cdot B \neq 0$ $BC \cdot A = 0$ $AC \cdot B \leq AC$ $AB \cdot C \leq AB$ $AB \times AC \approx BC$
A → B ↔ C	$AB \neq 0$ $AC \neq 0$ $BC \neq 0$ $AB \cdot C \neq 0$ $AC \cdot B \neq 0$ $BC \cdot A \neq 0$

B

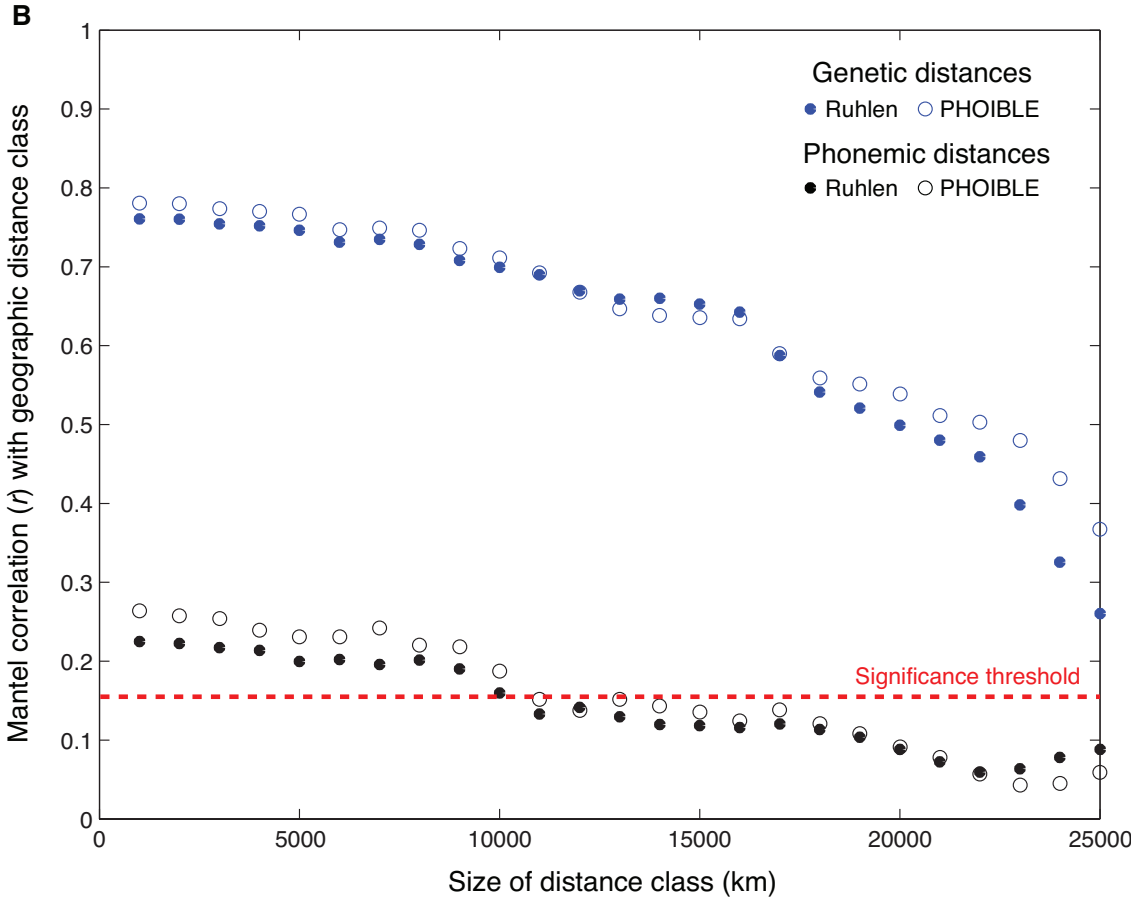

Fig. S3

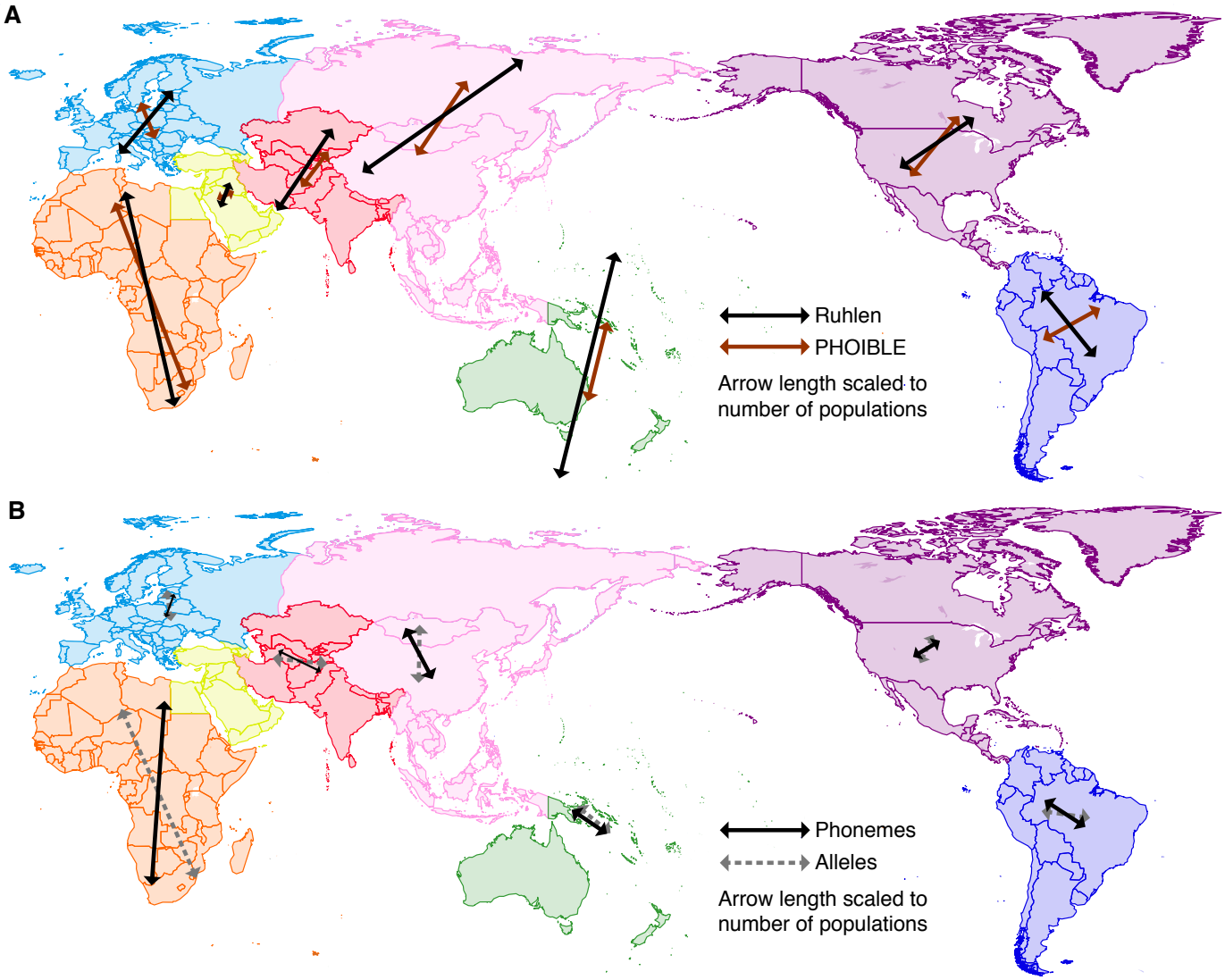
A
Ruhlen
PHOIBLE
Arrow length scaled to number of populations
B
Phonemes
Alleles
Arrow length scaled to number of populations

Fig. S4

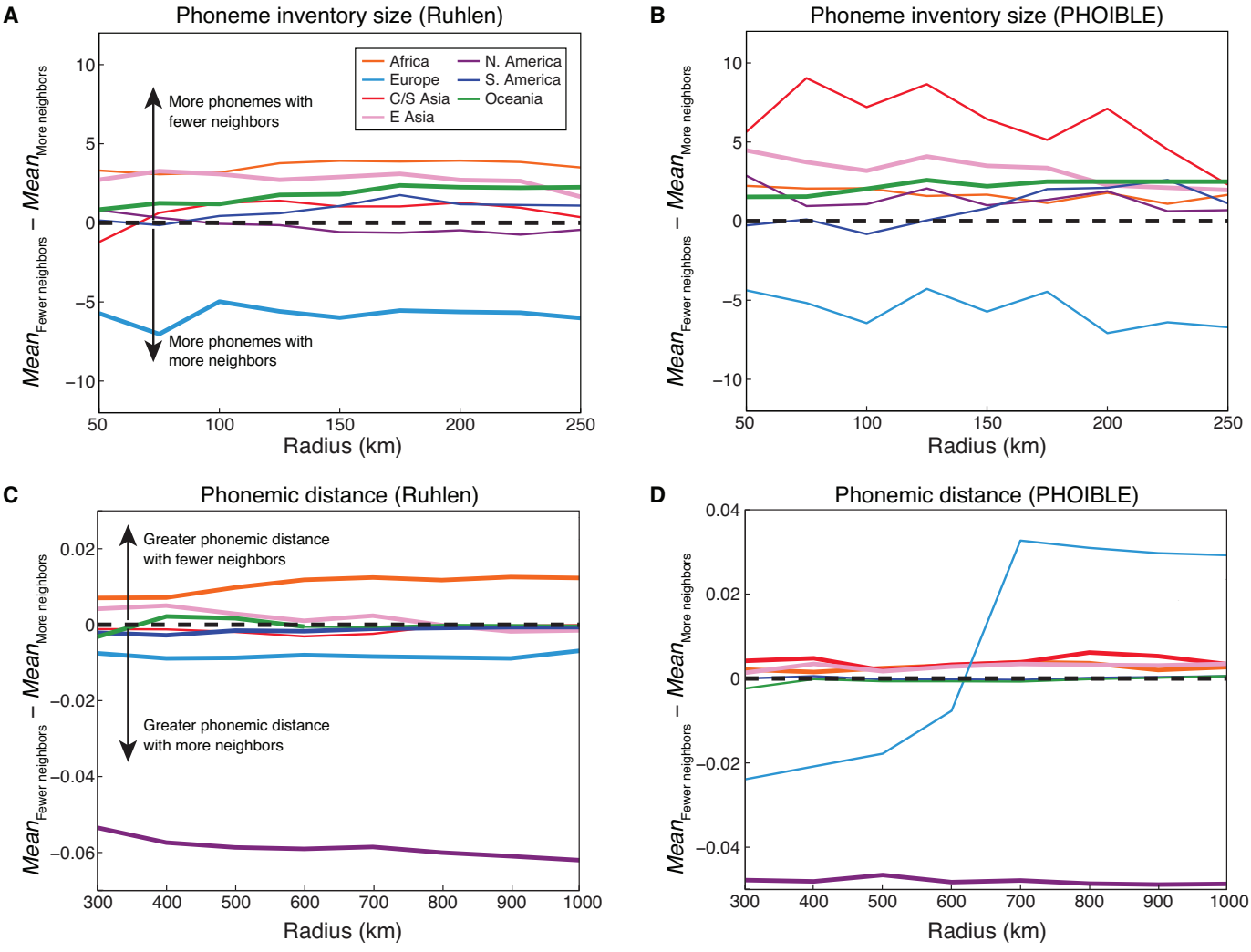
A
Phoneme inventory size (Ruhlen)
Africa
Europe
C/S Asia
E Asia
N. America
S. America
Oceania
More phonemes with fewer neighbors
More phonemes with more neighbors
Mean Fewer neighbors − Mean More neighbors
Radius (km)
B
Phoneme inventory size (PHOIBLE)
C
Phonemic distance (Ruhlen)
Greater phonemic distance with fewer neighbors
Greater phonemic distance with more neighbors
D
Phonemic distance (PHOIBLE)

Fig. S5

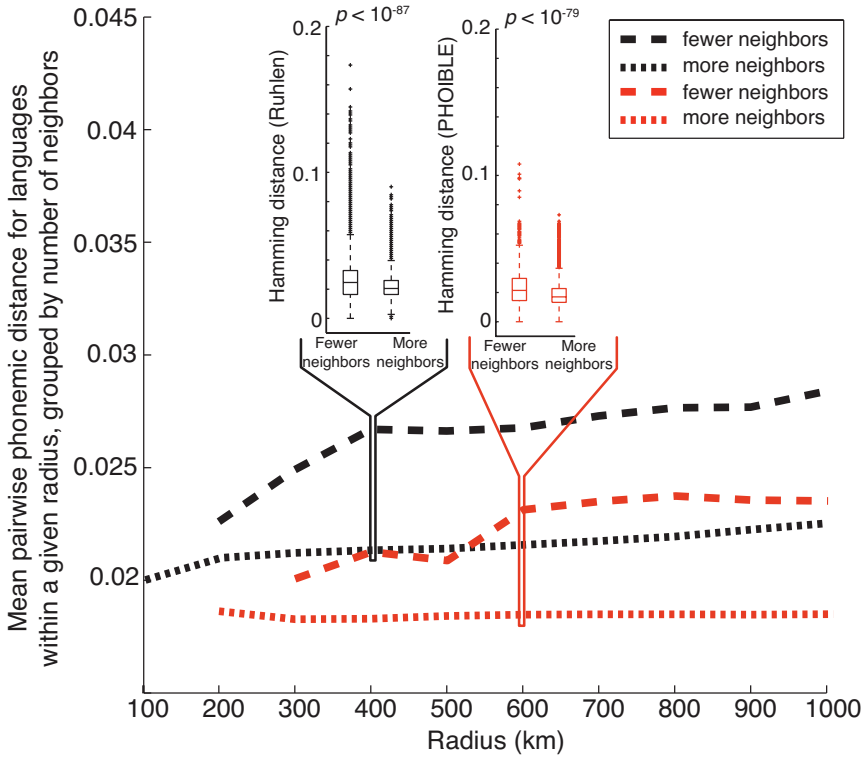
Mean pairwise phonemic distance for languages within a given radius, grouped by number of neighbors
0.045
0.04
0.035
0.03
0.025
0.02
100
200
300
400
500
600
700
800
900
1000
Radius (km)
fewer neighbors
more neighbors
fewer neighbors
more neighbors
Hamming distance (Ruhlen)
p < 10-87
0.2
0.1
0
Fewer neighbors
More neighbors
Hamming distance (PHOIBLE)
p < 10-79
0.2
0.1
0
Fewer neighbors
More neighbors

Fig. S6

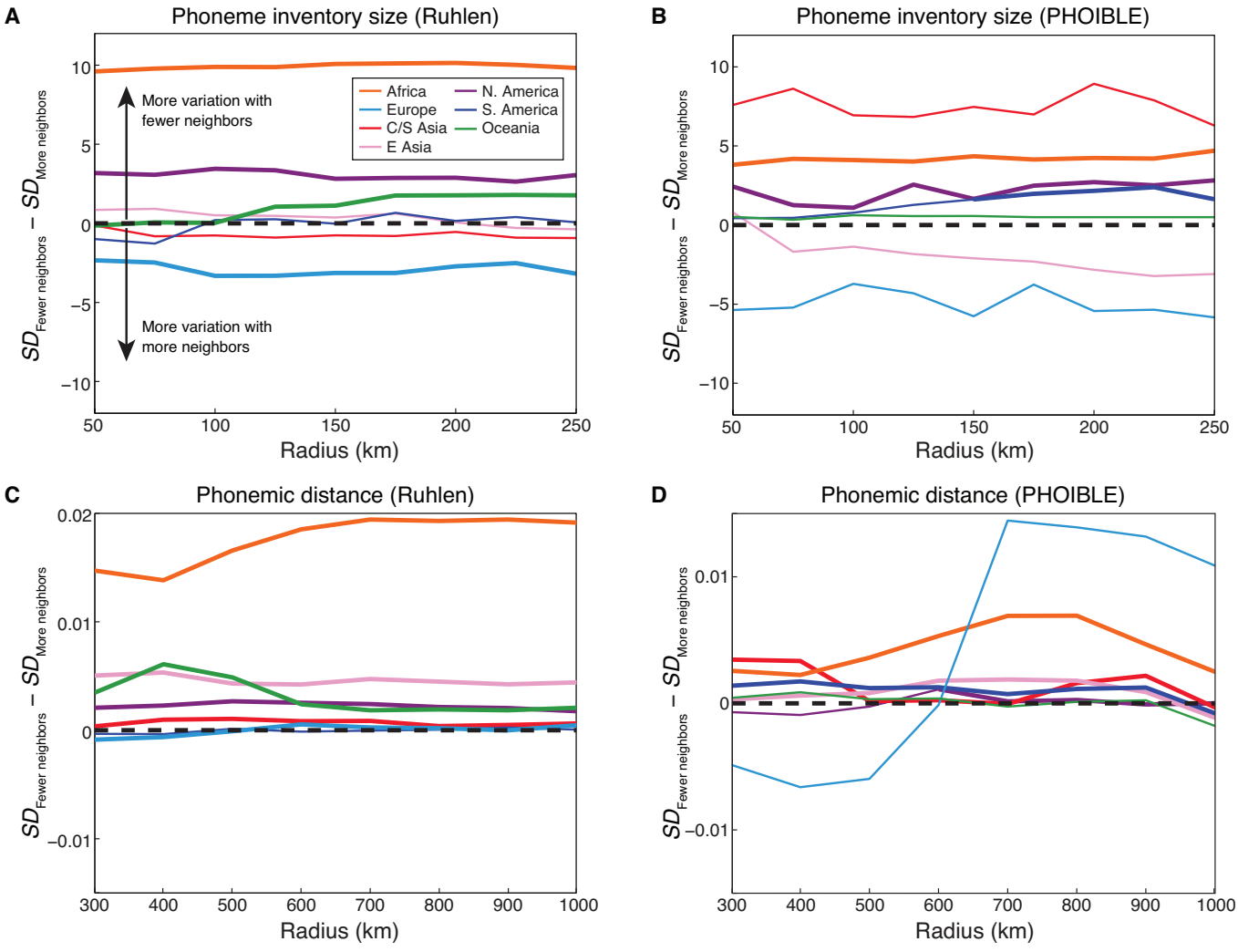
A
Phoneme inventory size (Ruhlen)
More variation with fewer neighbors
More variation with more neighbors
Africa
Europe
C/S Asia
E Asia
N. America
S. America
Oceania
SD Fewer neighbors – SD More neighbors
Radius (km)
B
Phoneme inventory size (PHOIBLE)
C
Phonemic distance (Ruhlen)
D
Phonemic distance (PHOIBLE)

Fig. S7

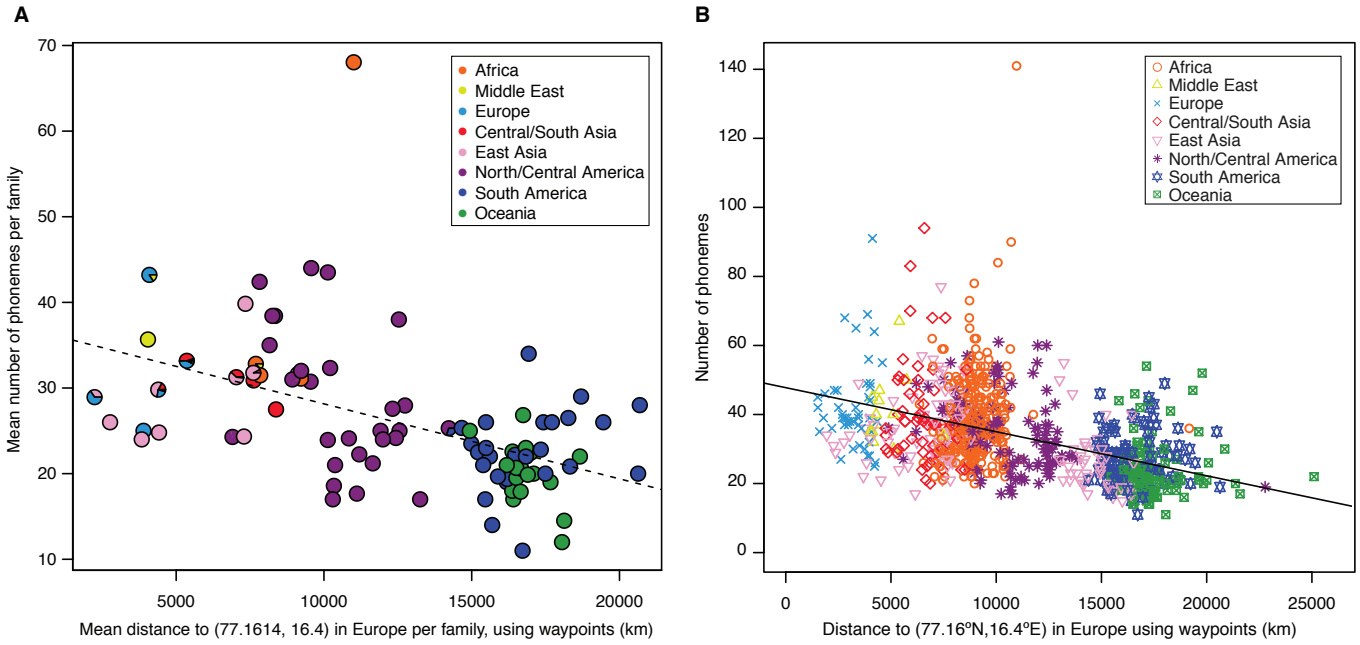
A
Mean number of phonemes per family
Mean distance to (77.1614, 16.4) in Europe per family, using waypoints (km)
Africa
Middle East
Europe
Central/South Asia
East Asia
North/Central America
South America
Oceania
B
Number of phonemes
Distance to (77.16ºN,16.4ºE) in Europe using waypoints (km)

Fig. S8

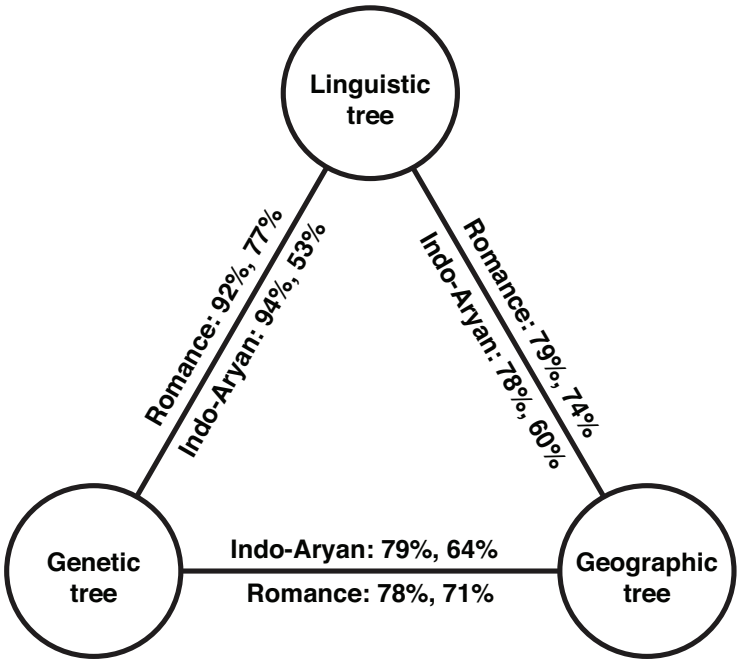
Linguistic tree
Romance: 92%, 77%
Indo-Aryan: 94%, 53%
Romance: 79%, 74%
Indo-Aryan: 78%, 60%
Genetic tree
Indo-Aryan: 79%, 64%
Romance: 78%, 71%
Geographic tree

Fig. S9

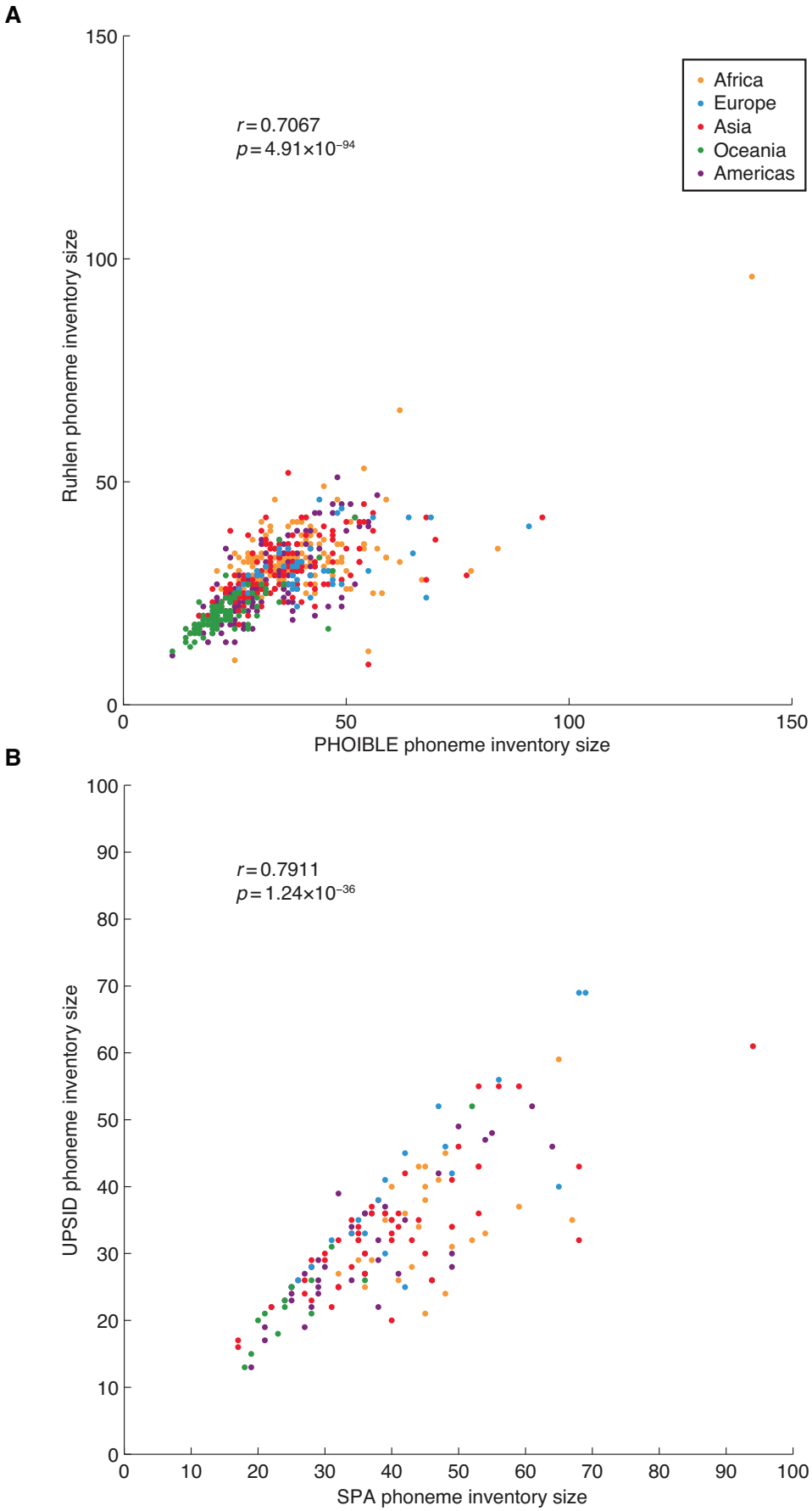
A
150
100
50
0
0
50
100
150
Ruhlen phoneme inventory size
PHOIBLE phoneme inventory size
r = 0.7067
p = 4.91×10^-94
Africa
Europe
Asia
Oceania
Americas
B
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
UPSID phoneme inventory size
SPA phoneme inventory size
r = 0.7911
p = 1.24×10^-36

Fig. S10

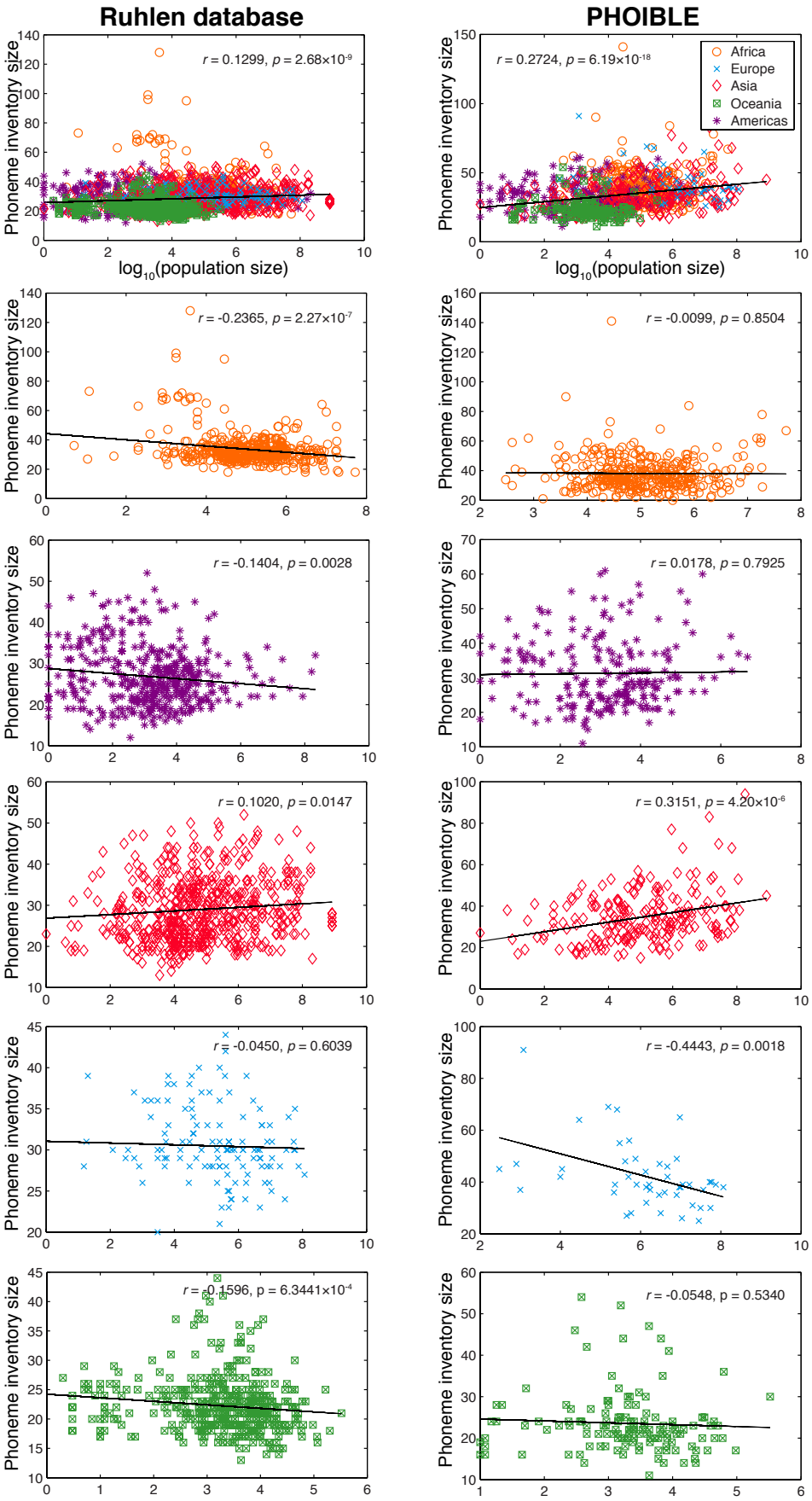
Ruhlen database
PHOIBLE
Phoneme inventory size
log10(population size)
r = 0.1299, p = 2.68×10-9
r = 0.2724, p = 6.19×10-18
Africa
Europe
Asia
Oceania
Americas
r = -0.2365, p = 2.27×10-7
r = -0.0099, p = 0.8504
r = -0.1404, p = 0.0028
r = 0.0178, p = 0.7925
r = 0.1020, p = 0.0147
r = 0.3151, p = 4.20×10-6
r = -0.0450, p = 0.6039
r = -0.4443, p = 0.0018
r = -0.1596, p = 6.3441×10-4
r = -0.0548, p = 0.5340

Fig. S11

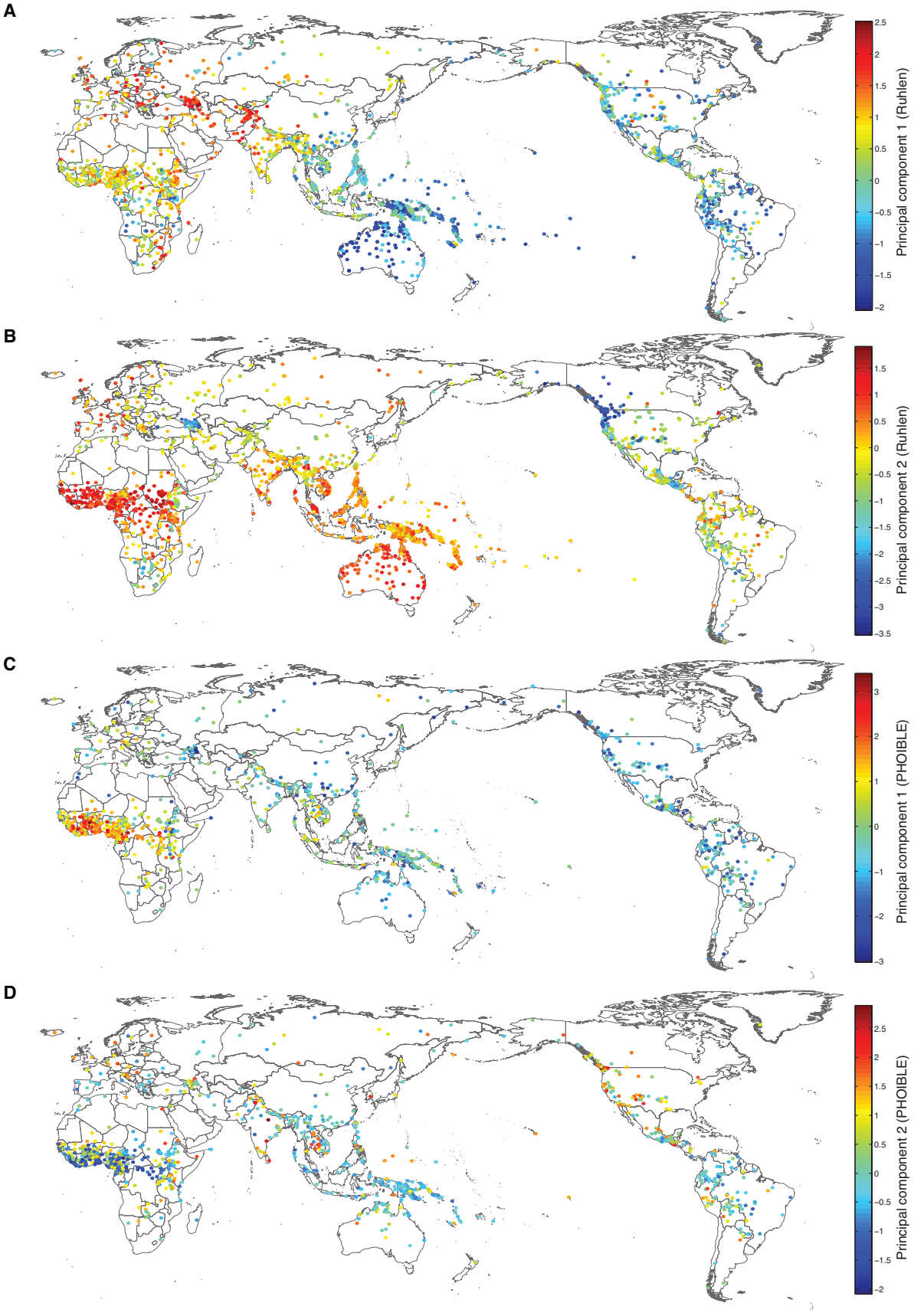
A
Principal component 1 (Ruhlen)
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
B
Principal component 2 (Ruhlen)
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
C
Principal component 1 (PHOIBLE)
3
2
1
0
−1
−2
−3
D
Principal component 2 (PHOIBLE)
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2

Fig. S12

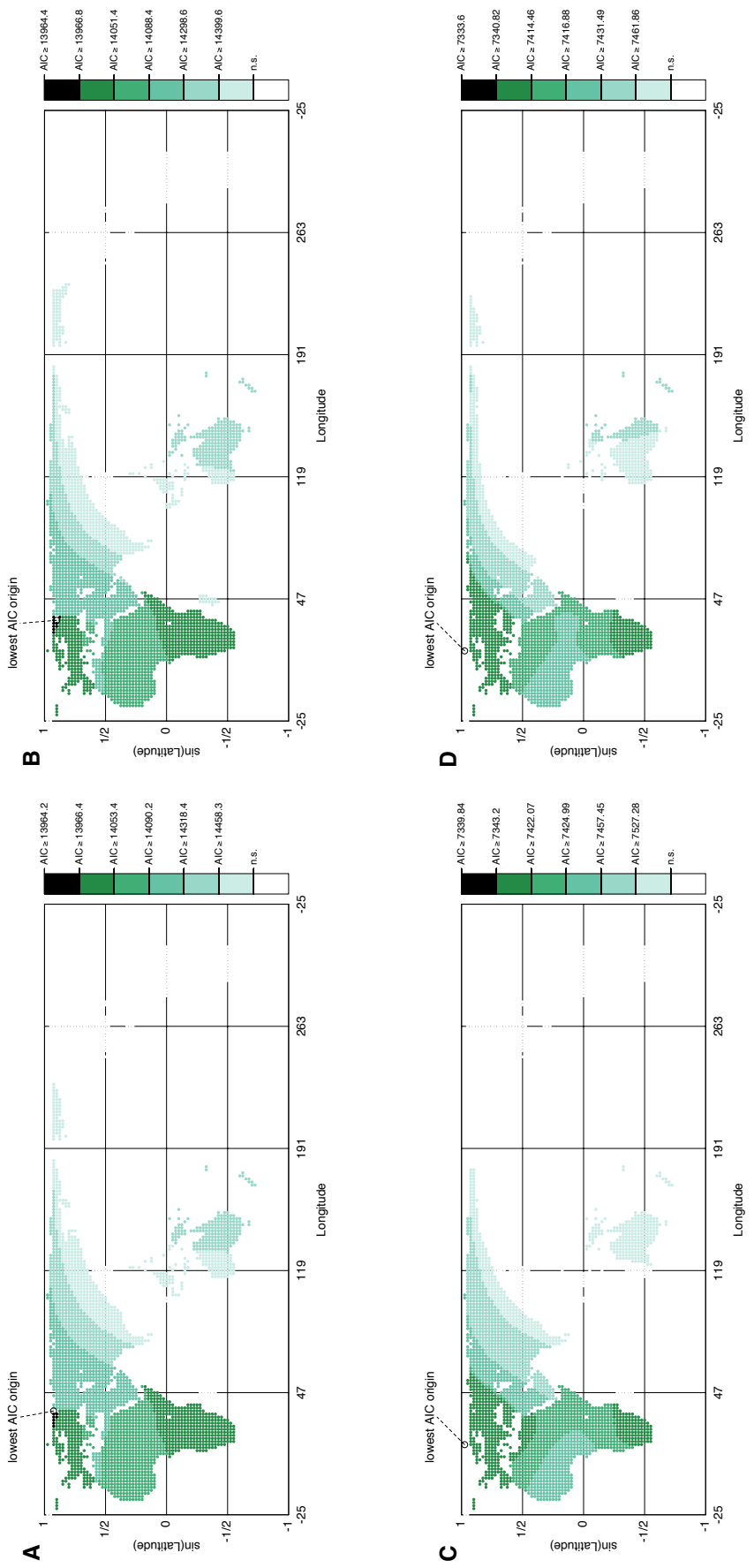
A
B
C
D
lowest AIC origin
sin(Latitude)
Longitude
AIC ≥ 13964.2
AIC ≥ 13966.4
AIC ≥ 14053.4
AIC ≥ 14090.2
AIC ≥ 14318.4
AIC ≥ 14458.3
AIC ≥ 13964.4
AIC ≥ 13966.8
AIC ≥ 14051.4
AIC ≥ 14088.4
AIC ≥ 14298.6
AIC ≥ 14399.6
AIC ≥ 7339.84
AIC ≥ 7343.2
AIC ≥ 7422.07
AIC ≥ 7424.99
AIC ≥ 7457.45
AIC ≥ 7527.28
AIC ≥ 7333.6
AIC ≥ 7340.82
AIC ≥ 7414.46
AIC ≥ 7416.88
AIC ≥ 7431.49
AIC ≥ 7461.86
n.s.

Fig. S13

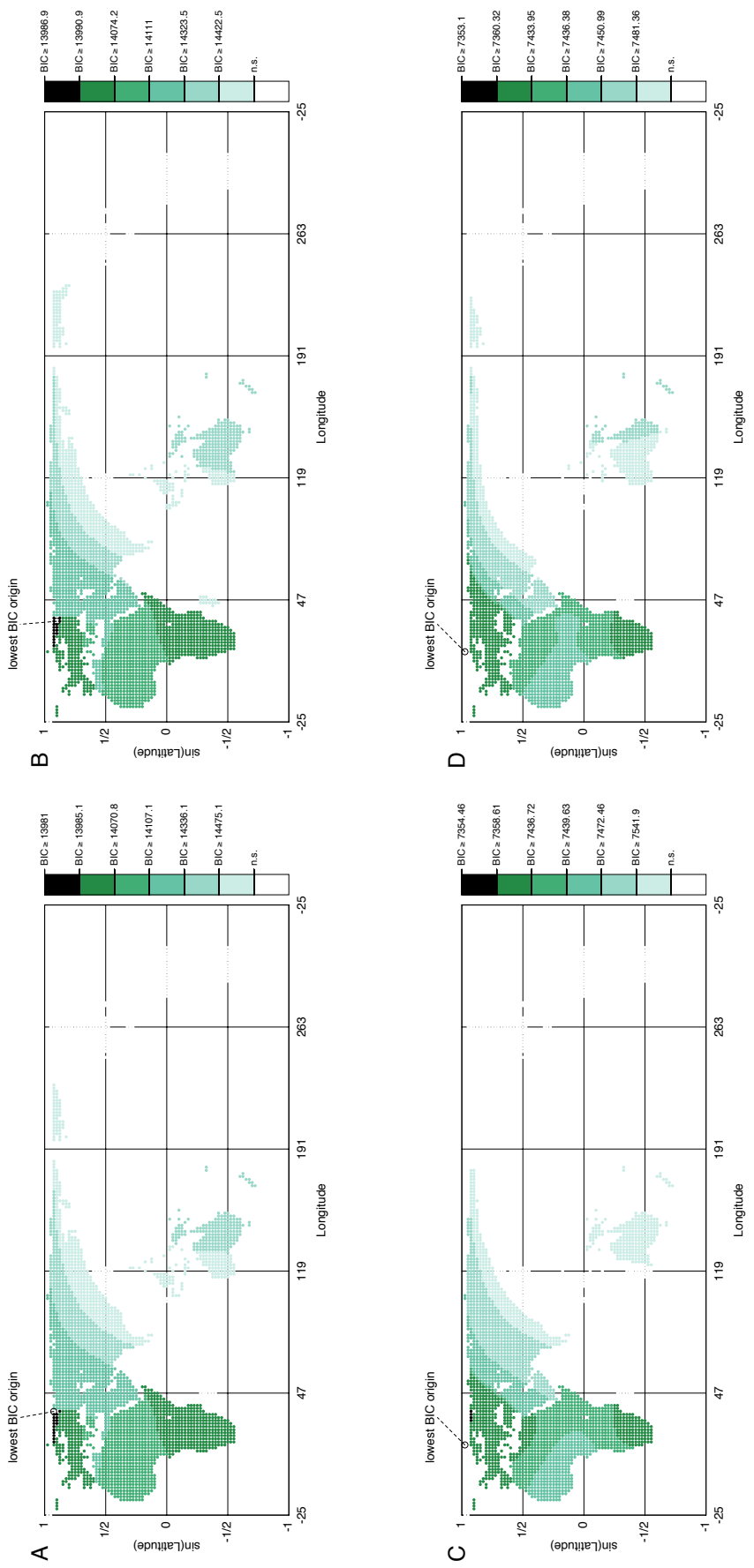
A
B
C
D
lowest BIC origin
sin(Latitude)
Longitude
BIC ≥ 13981
BIC ≥ 13985.1
BIC ≥ 14070.8
BIC ≥ 14107.1
BIC ≥ 14336.1
BIC ≥ 14475.1
BIC ≥ 13986.9
BIC ≥ 13990.9
BIC ≥ 14074.2
BIC ≥ 14111
BIC ≥ 14323.5
BIC ≥ 14422.5
BIC ≥ 7354.46
BIC ≥ 7358.61
BIC ≥ 7436.72
BIC ≥ 7439.63
BIC ≥ 7472.46
BIC ≥ 7541.9
BIC ≥ 7353.1
BIC ≥ 7360.32
BIC ≥ 7433.95
BIC ≥ 7436.38
BIC ≥ 7450.99
BIC ≥ 7481.36
n.s.

Fig. S14

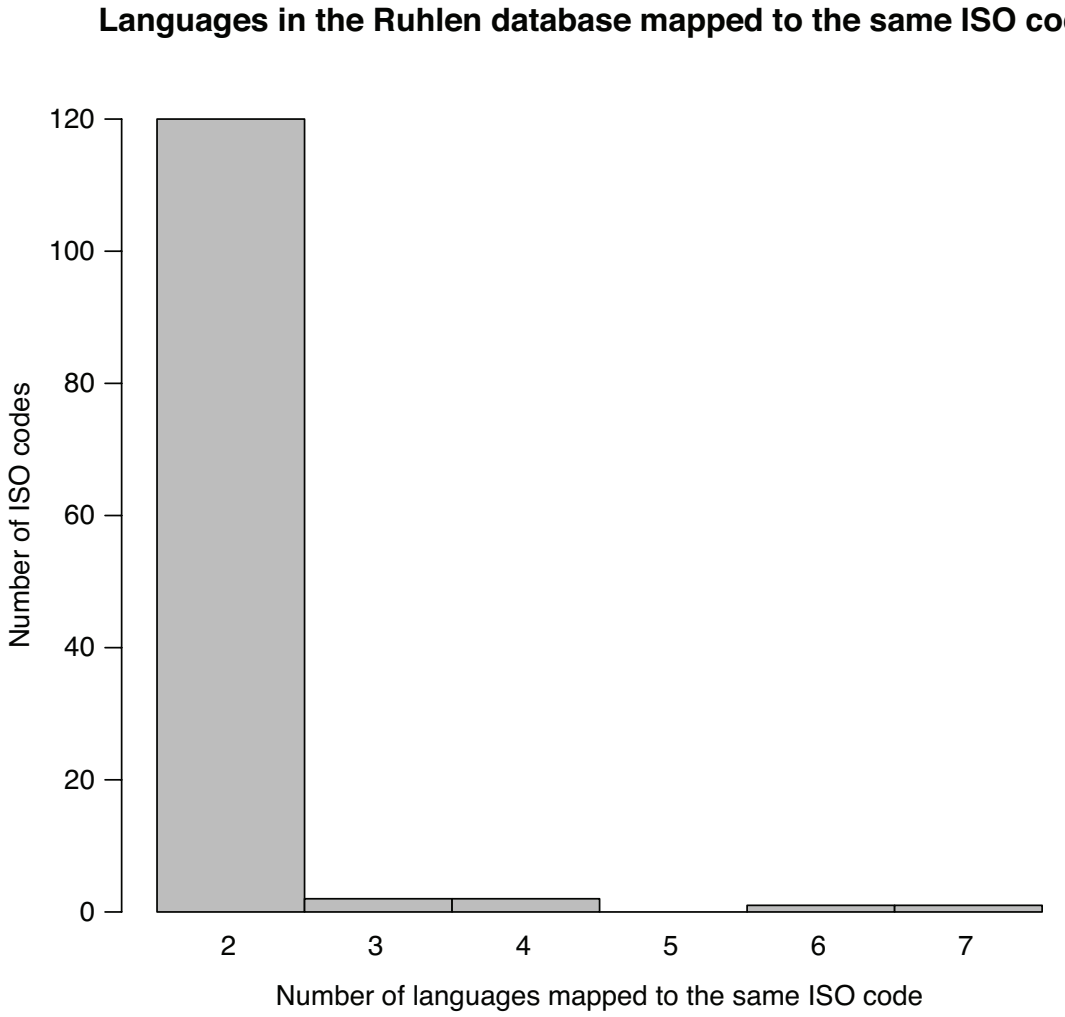
Languages in the Ruhlen database mapped to the same ISO code
Number of ISO codes
120
100
80
60
40
20
0
2
3
4
5
6
7
Number of languages mapped to the same ISO code

Fig. S15

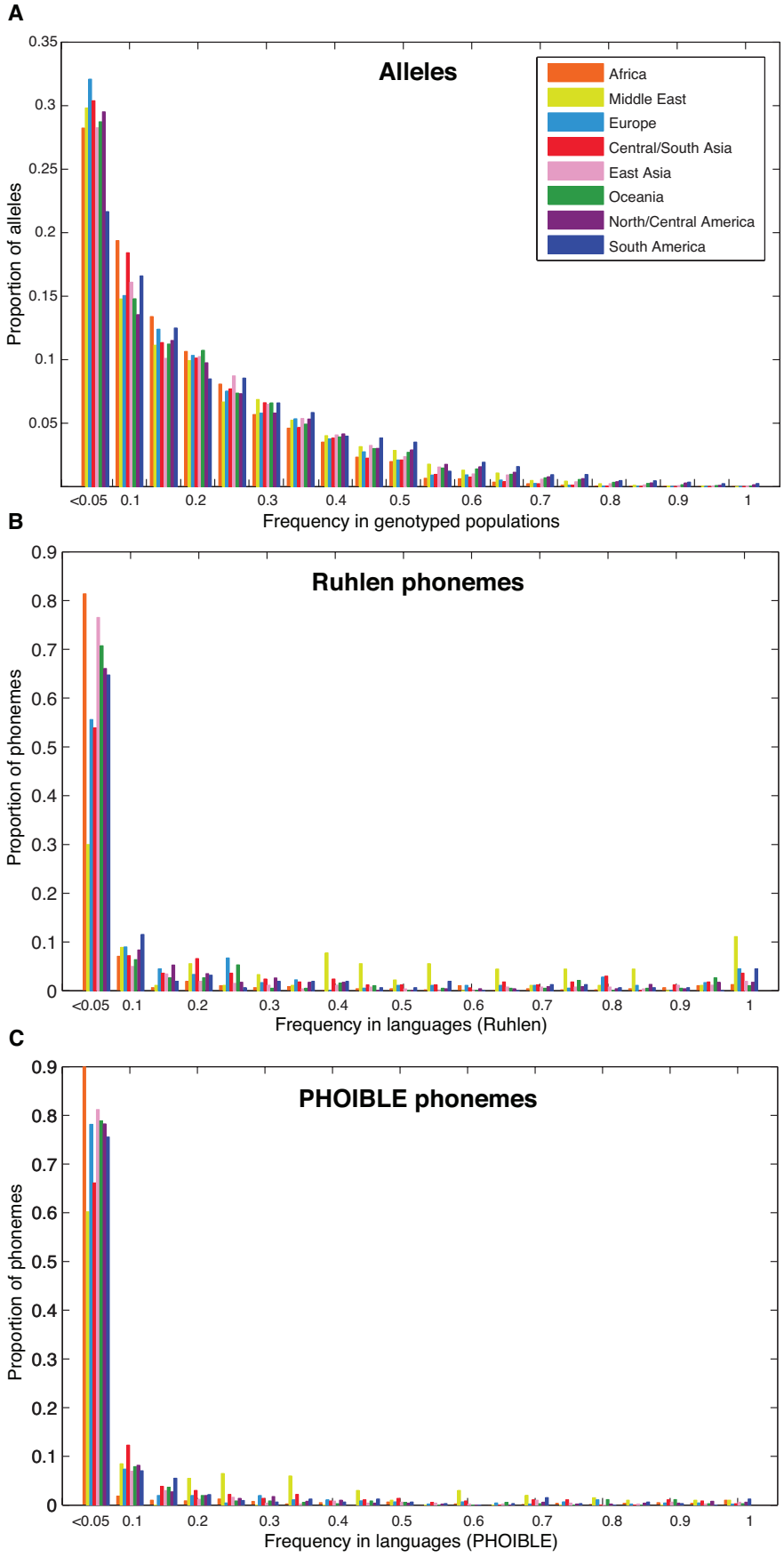
A
Alleles
Africa
Middle East
Europe
Central/South Asia
East Asia
Oceania
North/Central America
South America
Proportion of alleles
0.35
0.3
0.25
0.2
0.15
0.1
0.05
<0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Frequency in genotyped populations
B
Ruhlen phonemes
Proportion of phonemes
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
<0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Frequency in languages (Ruhlen)
C
PHOIBLE phonemes
Proportion of phonemes
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
<0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Frequency in languages (PHOIBLE)

Table S1. Results of Procrustes analyses and Mantel tests for phoneme inventories in the Ruhlen database and PHOIBLE. Procrustes analyses were conducted between phoneme principal components and geographic coordinates for each language (see *Materials and Methods*). Mantel tests were performed with the distance matrices indicated. Empirical *p*-values are reported.

			Language PC vs. geographic location		Phonemic distance (Jaccard) vs. geographic distance		Phonemic distance (Hamming) vs. geographic distance		Phonemic distance (Jaccard) vs. latitudinal distance		Phonemic distance (Jaccard) vs. longitudinal distance	
	Region	Num. langs.	Procrustes t_0	*p*-value	Mantel *r*	*p*-value	Mantel *r*	*p*-value	Mantel *r*	*p*-value	Mantel *r*	*p*-value
Ruhlen database	Worldwide	2082	0.5728	**1.00E-05**	0.1767	**1.00E-04**	0.0704	**1.00E-04**	0.277	**1.00E-04**	0.1709	**1.00E-04**
	Africa	468	0.5022	**1.00E-05**	0.3026	**1.00E-04**	0.2363	**1.10E-03**	0.5277	**1.00E-04**	0.0608	**1.70E-03**
	Middle East	32	0.6385	**1.00E-05**	0.4049	**1.00E-04**	0.4188	**1.00E-04**	0.403	**1.00E-04**	0.2583	**2.00E-04**
	Europe	135	0.5888	**1.00E-05**	0.4128	**1.00E-04**	0.3775	**1.00E-04**	0.1825	**1.00E-04**	0.3626	**1.00E-04**
	C./S. Asia	166	0.3681	**1.00E-05**	0.3446	**1.00E-04**	0.2208	**1.00E-04**	0.2769	**1.00E-04**	0.246	**1.00E-04**
	E. Asia	374	0.4299	**1.00E-05**	0.2977	**1.00E-04**	0.12	**5.00E-04**	0.2471	**1.00E-04**	0.184	**1.00E-04**
	N./C. America	305	0.4989	**1.00E-05**	0.2638	**1.00E-04**	0.1601	**3.00E-04**	0.2127	**1.00E-04**	0.2728	**1.00E-04**
	S. America	147	0.2625	**5.00E-05**	0.0174	0.37296	-0.0093	0.48545	0.1213	**5.30E-03**	0.0042	0.4474
	Oceania	455	0.4996	**1.00E-05**	0.2919	**1.00E-04**	0.2758	**1.00E-04**	0.3143	**1.00E-04**	0.0597	0.0276
PHOIBLE	Worldwide	968	0.5155	**1.00E-05**	0.2174	**1.00E-04**	0.0081	0.31547	0.3264	**1.00E-04**	0.2354	**1.00E-04**
	Africa	362	0.347	**1.00E-05**	0.2718	**1.00E-04**	0.2091	**2.90E-03**	0.337	**1.00E-04**	0.2182	**1.00E-04**
	Middle East	13	0.3823	0.3016	0.3499	**6.10E-03**	0.2033	0.12789	0.2932	0.020298	0.3293	**8.90E-03**
	Europe	47	0.3577	**0.0041**	0.2824	**2.00E-04**	0.222	0.012999	0.1133	0.062594	0.2068	**7.70E-03**
	C./S. Asia	58	0.0975	0.8032	0.1732	**5.20E-03**	-0.0688	0.78482	0.1387	0.019498	0.1351	0.0339
	E. Asia	136	0.3186	**1.00E-05**	0.2339	**1.00E-04**	-0.0055	0.53095	0.2369	**1.00E-04**	0.1951	**1.00E-04**
	N./C. America	122	0.2281	**0.003**	0.0458	0.23878	-0.0333	0.60704	0.1789	**1.00E-04**	0.1809	**7.00E-04**
	S. America	99	0.2661	**0.0014**	0.2101	**1.00E-04**	0.16	0.013999	0.131	**9.10E-03**	0.1702	**2.00E-04**
	Oceania	131	0.3905	**1.00E-05**	0.1857	**3.00E-04**	0.1755	0.016898	0.2125	**1.00E-04**	0.04674	0.23348

Table S2. Results of Procrustes analyses and Mantel tests for phoneme–genome datasets. (**A**) For 139 languages with genetic, phonemic, and geographic data in the Ruhlen database and 114 languages with genetic, phonemic, and geographic data in PHOIBLE, we performed pairwise Procrustes analyses between data types (phonemes, genotypes, and geographic locations). Procrustes similarity values (t_0) and empirical p-values (calculated after 100,000 permutations) are listed. We also performed Mantel tests with the corresponding distance matrices: phonemic (Jaccard) distance, genetic (allele-sharing) distance, and geographic (great-circle) distance.

	Ruhlen database		PHOIBLE		Ruhlen database		PHOIBLE	
	Procrustes t_0	*p*-value	Procrustes t_0	*p*-value	Mantel *r*	*p*-value	Mantel *r*	*p*-value
Phonemes vs. geography	0.1712	**0.0243**	0.2708	**3.7×10^{-4}**	0.1801	**1.0×10^{-4}**	0.2652	**1.0×10^{-4}**
Phonemes vs. genes	0.1592	0.0577	0.3565	**1.0×10^{-5}**	0.1571	**2.3×10^{-3}**	0.2399	**2.0×10^{-4}**
Genes vs. geography	0.6995	**1.0×10^{-5}**	0.7851	**1.0×10^{-5}**	0.761	**1.0×10^{-4}**	0.781	**1.0×10^{-4}**

(**B**) We calculated partial Mantel test results, comparing phonemic and genetic distance matrices while controlling for geographic distance. The association between phonemic and genetic distance is no longer significant when controlling for geographic distance and longitudinal distance (the difference in longitude coordinates) but not latitudinal distance (the difference in latitude coordinates). This finding is consistent for all regions except Oceania (see *Results*).

	Ruhlen database		PHOIBLE	
	Mantel *r*	*p*-value	Mantel *r*	*p*-value
Phonemic vs. genetic distance controlling for geographic distance	0.05363	0.15768	0.05439	0.17038
Phonemic vs. geographic distance controlling for genetic distance	0.1094	**0.01**	0.1284	**0.009999**
Genetic vs. geographic distance controlling for phonemic distance	0.7495	**9.9×10^{-5}**	0.7664	**9.9×10^{-5}**
Phonemic vs. genetic distance controlling for latitudinal distance	0.1236	**9.9×10^{-3}**	0.1793	**4.00×10^{-4}**
Phonemic vs. genetic distance controlling for longitudinal distance	0.07907	0.067993	0.1027	0.033297
Phonemic vs. genetic distance controlling for geographic distance within regions:				
Africa	0.09796	0.14099	0.1254	0.09819
Europe	-0.22	0.69163	-0.5955	0.94451
C./S. Asia	0.1519	0.09899	-0.2405	0.91231
E. Asia	-0.1365	0.75352	-0.00159	0.50655
N./C. America	0.2445	0.19608	-0.2867	0.72533
S. America	-0.07942	0.67473	0.09186	0.34207
Oceania	0.4221	**2.00×10^{-4}**	0.6031	**2.60×10^{-3}**

Table S3. Results of Mantel tests along varied axes. The geographic distance vector connecting each pair of languages was rotated at 1-degree intervals, and the Mantel correlation was calculated between phonemic (Jaccard) distance and each matrix of rotated distances. The axis that maximized the Mantel correlation is shown; in all cases, the Mantel *r* statistic gradually increased as the distance matrix rotation approached this maximized axis.

			Genetic distance			Linguistic distance			
	Region	Number of populations	Axis	Maximum Mantel *r*	*p*-value	Axis	Maximum Mantel *r*	*p*-value	Angle between linguistic and genetic axes
Ruhlen database (139 populations)	Africa	62	13°–193°	0.269	**0.012**	3°–183°	0.455	**0.001**	10°
	Middle East	2	--	--	--	--	--	--	--
	Europe	8	172°–352°	0.159	0.241	168°–348°	0.416	0.128	4°
	C./S. Asia	19	33°–213°	0.205	0.061	36°–216°	0.461	**0.001**	3°
	E. Asia	13	179°–359°	0.217	0.121	167°–347°	0.352	**0.009**	12°
	N./C. America	8	39°–219°	0.594	**0.001**	139°–319°	0.544	**0.001**	78°
	S. America	11	95°–275°	0.429	0.056	120°–300°	0.417	**0.001**	25°
	Oceania	16	120°–300°	0.571	**0.004**	161°–341°	0.435	**0.005**	41°
PHOIBLE (114 populations)	Africa	55	156°–226°	0.311	**0.012**	4°–184°	0.494	**0.001**	27°
	Middle East	0	--	--	--	--	--	--	--
	Europe	6	172°–352°	0.085	0.307	20°–200°	0.499	**0.021**	28°
	C./S. Asia	13	96°–276°	0.185	0.194	116°–296°	0.228	0.118	20°
	E. Asia	14	1°–181°	0.377	**0.034**	152°–332°	0.198	0.060	29°
	N./C. America	6	25°–205°	0.899	**0.017**	59°–239°	0.146	0.442	34°
	S. America	11	98°–278°	0.487	**0.018**	123°–303°	0.453	**0.001**	25°
	Oceania	9	129°–309°	0.646	**0.001**	124°–304°	0.539	**0.001**	5°

Table S4. Languages with large phoneme inventories. (**A**) Language names and locations of the languages in the top 5% of phoneme inventory sizes in the Ruhlen database. (**B**) Language names and locations of the languages in the top 5% of phoneme inventory sizes in PHOIBLE.

A. Ruhlen database

Ruhlen ID	Language name	Phoneme inventory size	Region
38	N/amani	133	Africa
3	Qxû	100	Africa
6	Qxû	96	Africa
5	Qxû	95	Africa
29	Xû	76	Africa
40	N|huki	72	Africa
32	G|wi	71	Africa
30	G||ana	69	Africa
17	G||abake	68	Africa
24	Danisin	68	Africa
26	Kxoe	68	Africa
31	G||ana	68	Africa
1160	Xhosa	68	Africa
1	Hadza	67	Africa
27	Buka	67	Africa
28	Handa	67	Africa
18	G||abake	65	Africa
33	Naron	63	Africa
37	ǂHû	63	Africa
4	Qxû	60	Africa
22	Shua	58	Africa
1161	Zulu	58	Africa
2	Sandawe	53	Africa
23	Shua	53	Africa
1167	Tsonga	53	Africa
1155	Venda	52	Africa
1157	Northern Sotho	52	Africa
1162	Swati	52	Africa
1686	Marathi	52	Central/South Asia
1689	Konkani	52	Central/South Asia
4640	Tlingit	51	North/Central America
2130	Western Tibetan	51	East Asia
2352	Miao	51	East Asia
2358	Miao	51	East Asia
1158	Southern Sotho	50	Africa
1302	Mangbetu	50	Africa
2356	Miao	50	East Asia
25	Deti	49	Africa
2467	Loven	49	East Asia
1636	Kryts	49	Middle East
477	Igbo	48	Africa
2360	Punu	48	East Asia
1163	Ndebele	47	Africa
4670	Chipewyan	47	North/Central America
4676	Carrier	47	North/Central America
2403	Lawa	47	East Asia
1630	Lezgi	47	Europe
1294	Kara	46	Africa
1301	Madi	46	Africa
1387	Tera	46	Africa

1543	Dahalo	46	Africa
4726	Haisla	46	North/Central America
1678	Pashai	46	Central/South Asia
1682	Dumaki	46	Central/South Asia
1688	Marathi	46	Central/South Asia
2044	Toda	46	Central/South Asia
278	Duru	45	Africa
4669	Slave	45	North/Central America
4725	Heiltsuk	45	North/Central America
4757	Coeur dAlene	45	North/Central America
4759	Western Keres	45	North/Central America
4760	Yuchi	45	North/Central America
4908	Eastern Pomo	45	North/Central America
1690	Sindhi	45	Central/South Asia
1700	Bhili	45	Central/South Asia
1712	Awadhi	45	Central/South Asia
1750	Parachi	45	Central/South Asia
2218	Angami	45	Central/South Asia
1977	Ordos	45	East Asia
2126	Central Tibetan	45	East Asia
2357	Miao	45	East Asia
2536	Lakkia	45	East Asia
1618	Axvax	45	Europe
13	!Ora	44	Africa
340	Viri	44	Africa
1304	Mamvu	44	Africa
4753	Columbian	44	North/Central America
4916	Chumash	44	North/Central America
4938	Tlamelula	44	North/Central America
5003	Otomi	44	North/Central America
1705	Hindi	44	Central/South Asia
1715	Maithili	44	Central/South Asia
2132	Magar	44	Central/South Asia
1961	Chulym	44	East Asia
1635	Tsaxur	44	Europe
1844	Scottish Gaelic	44	Europe
1638	Udi	44	Middle East
134	Basari	43	Africa
211	Dagara	43	Africa
4752	Shuswap	43	North/Central America
4755	Kalispel	43	North/Central America
4805	Coos	43	North/Central America
1667	Bashkarik	43	Central/South Asia
1671	Wotapuri	43	Central/South Asia
1701	Gade Lohar	43	Central/South Asia
1703	Hindi	43	Central/South Asia
1741	Wakhi	43	Central/South Asia
2361	Mien	43	East Asia
2504	Mon	43	East Asia
2997	Haroi	43	East Asia
3002	North Raglai	43	East Asia
1601	Ubyx	43	Europe
1634	Rutul	43	Europe
3402	Yuaga	43	Oceania
3423	Iaai	43	Oceania

B. PHOIBLE

Language ID (ISO)	Language name	Phoneme inventory size	Region
ktz	!Xu	141	Africa
hin	Hindi-Urdu	94	Central/South Asia
aqc	Archi	91	Europe
yey	Yeyi	90	Africa
daf	Dan	84	Africa
skr	Siraiki	83	Central/South Asia
ary	Moroccan Arabic	78	Africa
prk	Parauk	77	East Asia
bav	Babungo (grassfields bantu, ring)	73	Africa
pan	Punjabi	70	Central/South Asia
lbe	Lak	69	Europe
bkm	Kom	68	Africa
gle	Irish Gaelic	68	Europe
kru	Kurukh	68	Central/South Asia
tel	Telugu	68	Central/South Asia
arz	Egyptian Arabic	67	Middle East
hun	Hungarian	65	Europe
ndb	Kensei Nsei	65	Africa
rut	Rutul	64	Europe
apd	Arabe	62	Africa
hts	Hadza	62	Africa
ibo	Igbo	62	Africa
tow	Jemez	61	North/Central America
maz	Mazahua	60	North/Central America
zpq	San Bartolomé Zoogocho Zapotec	60	North/Central America
amh	Amharic	59	Africa
dal	Dahalo	59	Africa
fwe	Fwe	59	Africa
xtc	Katcha	59	Africa
aka	Akan	58	Africa
bby	Befang	57	Africa
chp	Chipewyan	57	North/Central America
cko	Anufɔ	57	Africa
jya	Jiarong	57	East Asia
azo	Awing	56	Africa
bam	Bambara	56	Africa
cqd	Hmong	56	East Asia
kas	Kashimiri	56	Central/South Asia
kbd	Kabardian	56	Europe
nla	Ngombale	56	Africa
ace	Acehnese	55	East Asia
bqx	Kambari	55	Africa
kwk	Kwakiutl	55	North/Central America
mlt	Maltese	55	Europe
ote	Otomi	55	North/Central America
dic	Dida	54	Africa
grg	Ma'di	54	Oceania
nmg	Mvumbo	54	Africa

Table S5. Best-fit linear regressions of total phoneme inventory size onto geographic distance, using mean or median values within each language family for total number of phonemes and geographic distance to the center. Geographic centers shown had the lowest rescaled *AIC* across 4210 centers on land for each model fitted for each dataset. The Ruhlen database has 2046 languages classified in 98 Ethnologue language families; 36 Ruhlen entries with language families labeled as "Unclassified", "Language Isolate" or "Mixed Language" were excluded from this analysis. PHOIBLE has 949 language classified into 81 language roots; 19 languages listed with unclassified roots (denoted as "UNCL" by PHOIBLE) were excluded from this analysis. Two types of models were fitted: "1" in the "Number of independent variables" column denotes that the only independent variable in the linear regression was geographic distance to the origin; "2" denotes that a multiple linear regression was fitted, with geographic distance to the origin and base-10 logarithm of current speaker population size as independent variables; all models have an intercept as well. The lowest-*AIC* value observed across models fitted for each database is shown in bold.

Mean or median values per family?	Linguistic dataset	Number of independent variables	Latitude of geographic center with lowest AIC	Longitude of geographic center with lowest AIC	*AIC* (not rescaled) of model for origin in columns 4 & 5	R^2
Mean	Ruhlen	1	77.1614	16.4	**659.4515**	0.2620
Median	Ruhlen	1	77.1614	16.4	660.0656	0.2518
Mean	PHOIBLE	1	77.1614	16.4	579.1587	0.2826
Median	PHOIBLE	1	77.1614	16.4	563.1790	0.2972
Mean	Ruhlen	2	77.1614	16.4	661.4131	0.2623
Median	Ruhlen	2	77.1614	16.4	662.0635	0.2518
Mean	PHOIBLE	2	77.1614	16.4	577.4127	0.3150
Median	PHOIBLE	2	77.1614	16.4	**563.0373**	0.3155

Table S6. The various diacritics used in the Ruhlen database to represent the modifications of basic consonants and vowels. Unless noted, a particular modification applied to both consonants and vowels. Taken from Ruhlen's document typology.pdf, available along with the database at http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl (the listing name is "a global linguistic database").

[ʲ]: palatalized (consonants only)	[˜]: nasalized
[ʷ]: labialized (consonants only)	[ː]: long
[ʷ]: velarized	[̪]: dental (consonants only)
[ˤ]: pharyngealized	[̺]: retroflex (consonants only)
[ʰ]: aspirated	[͈]: fortis
[ˀ]: glottalized	[̥]: voiceless
[ᵍ]: voiced click (clicks only)	[̤]: breathy voice
[̩]: syllabic	[̰]: creaky voice (vowels only)
[ᵐ ⁿ ᵑ]: prenasalized	

Table S7. Unicode conversions. Conversion of Notepad-specific characters to Unicode is detailed below. This conversion was necessary for a handful of characters encoding phonemes in the Ruhlen database.

Notepad encoding	Notepad character	Unicode code point (hexadecimal)	Unicode character	Unicode character name
\f1b\f0	β	03B2	β	Greek Small Letter Beta
\rquote	ɣ	281	ʁ	Latin Letter Small Capital Inverted R
633	ɹ	279	ɹ	Latin Small Letter Turned R
\ldblquote	ħ	127	ħ	Latin Small Letter H With Stroke
\'98	ɨ	268	ɨ	Latin Small Letter I With Stroke

Table S8. Geographic centers with the best-fit linear regressions of phoneme inventory size onto geographic distance. Geographic centers shown had the lowest rescaled *AIC* across 4210 centers on land for each model fit for each dataset (Figure S12). Two models were fit for each dependent variable: "1" denotes that the only independent variable in the linear regression was geographic distance to the origin; "2" denotes a multiple linear regression was fit, with geographic distance to the origin and base-10 logarithm of current speaker population size as independent variables; all models have an intercept as well.

Dependent variable	Linguistic dataset	Number of independent variables	Latitude of geographic center with lowest AIC	Longitude of geographic center with lowest AIC	*AIC* (not rescaled) of model for origin in columns 3 & 4
Total number of phonemes	Ruhlen	1	67.6684	36.2	13964.23
	PHOIBLE	1	77.1614	16.4	7339.84
	Ruhlen[1]	2	64.1581	34.4	13964.45
	PHOIBLE[2]	2	77.1614	16.4	7333.60
Total number of phonemes, excluding tones	PHOIBLE	1	77.1614	16.4	7241.228
	PHOIBLE	2	77.1614	16.4	7339.62
Total number of phonemes, excluding modifications	Ruhlen	1	67.6684	36.2	13828.83
	Ruhlen	2	67.6684	36.2	13827.98
Total number of phonemes, excluding clicks	Ruhlen	1	67.6684	36.2	12938.85
	PHOIBLE	1	77.1614	16.4	7248.50
	Ruhlen	2	67.6684	36.2	12940.85
	PHOIBLE	2	77.1614	16.4	7241.22
Total number of phonemes, excluding clicks and modifications	Ruhlen	1	67.6684	36.2	12756.91
	Ruhlen	2	77.1614	16.4	12758.45

[1] When regressing total number of phonemes, the geographic center in the lowest-*AIC* model using the Ruhlen database and multiple linear regression is 398.74 km away from the lowest-*AIC* center using the Ruhlen database and simple linear regression.

[2] When regression total number of phonemes, the geographic center in the lowest-*AIC* model using PHOIBLE and multiple linear regression is 1233.55 km away from the lowest-*AIC* center using PHOIBLE and simple linear regression.

Table S9. Jackknife analysis of geographic regions. Geographic locations producing the lowest AIC across 4210 fitted models, jackknifing over geographic regions for languages in both linguistic datasets. The dependent variable for all models fit is total number of phonemes; only results for simple linear regressions are shown here. Geographic centers shown had the greatest support out of 4210 possible centers on land, based on lowest rescaled *AIC* within each model fit, excluding languages in the continental region listed in column 1 below.

Continental region excluded	Linguistic dataset used	Number of languages in excluded region	Latitude of geographic center with lowest *AIC*	Longitude of geographic center with lowest *AIC*	Continent of lowest-*AIC* center
Africa	Ruhlen	468	77.1614	16.4	Europe
Africa	PHOIBLE	362	77.1614	16.4	Europe
Europe	Ruhlen	135	64.1581	-16	Europe
Europe	PHOIBLE	47	77.1614	16.4	Europe
Middle East	Ruhlen	32	67.6684	36.2	Europe
Middle East	PHOIBLE	13	77.1614	16.4	Europe
Central/South Asia	Ruhlen	166	-33.367	27.2	Africa
Central/South Asia	PHOIBLE	58	77.1614	16.4	Europe
East Asia	Ruhlen	374	67.6684	36.2	Europe
East Asia	PHOIBLE	136	77.1614	16.4	Europe
Oceania	Ruhlen	455	-31.6682	29	Africa
Oceania	PHOIBLE	131	42.4542	47	Asia
North/Central America	Ruhlen	305	-31.6682	27.2	Africa
North/Central America	PHOIBLE	122	67.6684	36.2	Europe
South America	Ruhlen	147	67.6684	36.2	Europe
South America	PHOIBLE	99	77.1614	16.4	Europe