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.

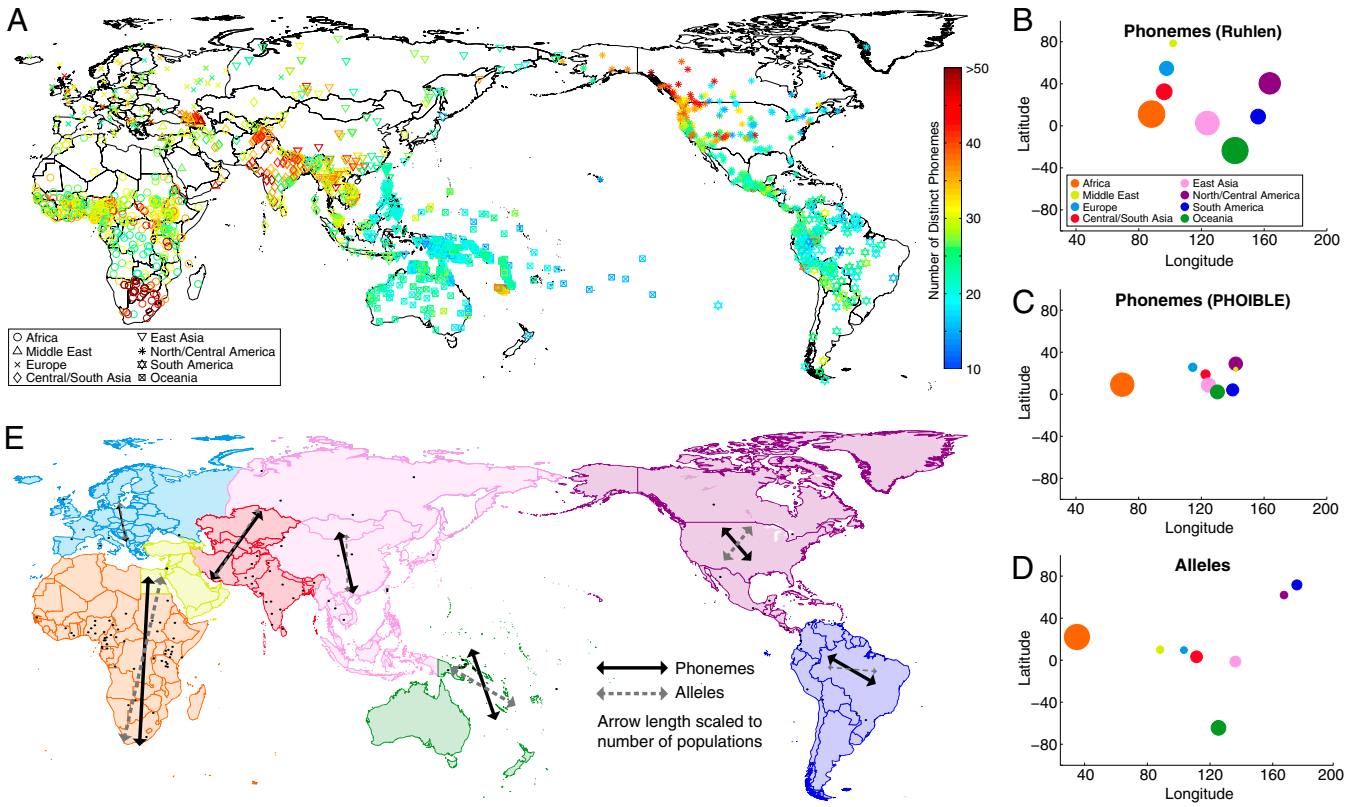

Fig. 1. Procrustes-transformed PCs for all phonemes and regional axes of phonemic and genetic differentiation. (*A*) Locations of 2,082 languages in the Ruhlen database. Phoneme inventory size of each language is indicated by the color bar. We performed Procrustes analyses to compare the first two PCs 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

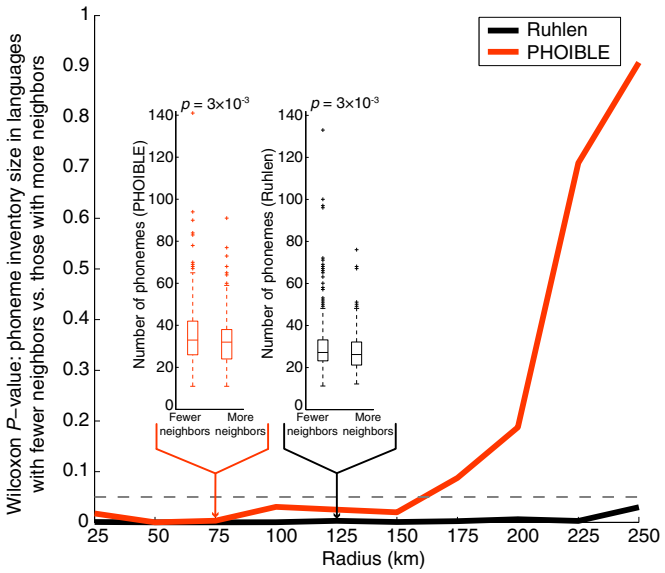

Fig. 2. The effect of geographic isolation on phonemes. Languages with fewer neighbors (less than or equal to median number of neighbors) had significantly more phonemes (Wilcoxon rank-sum test) than languages with more neighbors for all tested radii in the Ruhlen database (black line) and for radii < 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

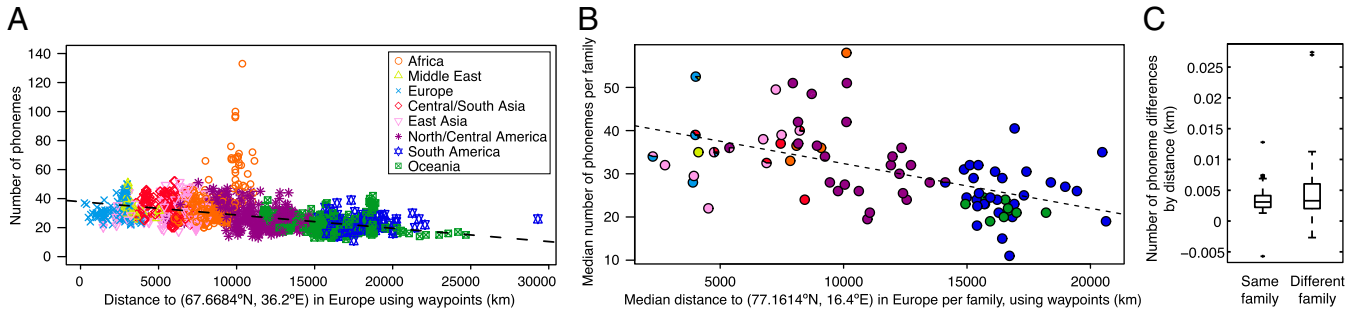

Fig. 3. Best-fit linear regressions of phoneme inventory size on geographic distance. For both databases, the best-fit geographic center was located in northern Europe, roughly equidistant from Oceania and South America, grouping two regions with small phoneme inventories and producing a significantly 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).

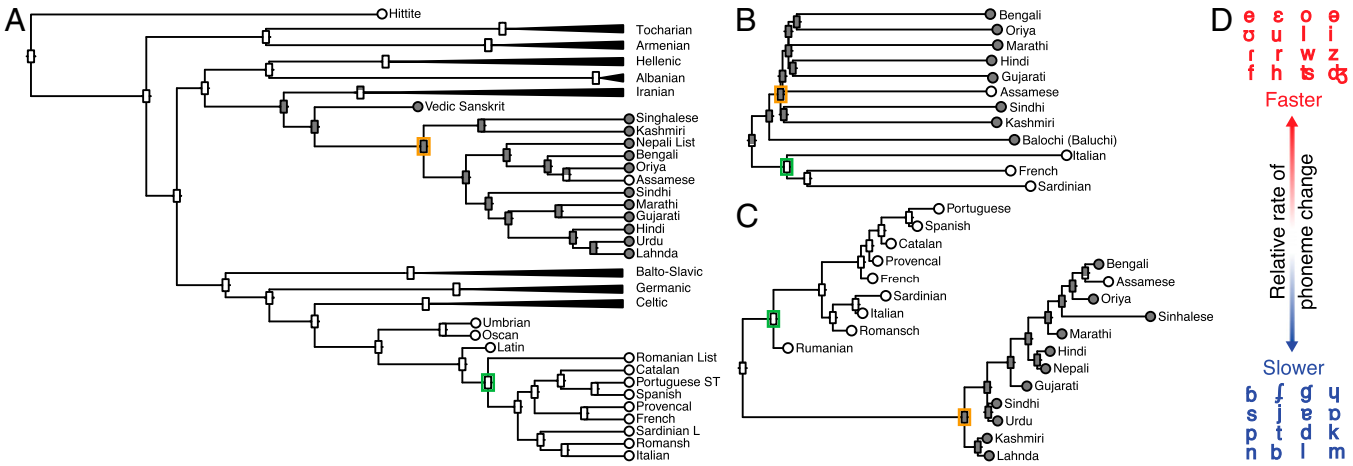

Fig. 4. Estimating ancestral phoneme states with cognate-based, geographic, and genetic trees. (*A*) Phylogeny of Indo-European languages (47) with presence of the phoneme /l/ indicated by gray circles at each tip. Based on the tree topology and branch lengths, the probability of phoneme presence at 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.