

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

Materials and Methods

Figures S1-S15

Tables S1-S9

References

## Table of Contents

<b>1. Preparation of linguistic data .....</b>	<b>3</b>
1.1 Processing the Ruhlen database.....	3
1.2 Modified consonants and modified vowels .....	3
1.3 Removing non-phonemic distinctions: aspirates and dentals .....	3
1.4 Encoding Ruhlen phonemes using Unicode .....	4
1.5 Annotating Ruhlen languages with speaker population sizes and geographic coordinates .....	4
1.6 Filtering of languages for analysis .....	5
1.7 Processing PHOnetics Information Base and LExicon (PHOIBLE) .....	6
1.8 Annotating PHOIBLE languages with speaker population sizes and geographic coordinates .....	7
1.9 Generating binary matrices for both databases .....	7
1.10 Defining geographic regions .....	7
1.11 Population size and phoneme inventory size .....	8
<b>2. Merging of genetic data with linguistic data .....</b>	<b>8</b>
2.1 Microsatellite data.....	8
2.2 Matching genetic populations to Ruhlen database languages.....	9
<b>3. Statistical analyses of linguistic and genetic data .....</b>	<b>9</b>
3.1 Correlations between phonemic, genetic, and geographic distance .....	9
3.2 Regression analyses using individual languages.....	10
3.2.1 Model selection.....	11
3.2.2 Jackknifing over geographic regions .....	12
3.2.3 Including number of neighbors in regression analyses.....	12
3.2.4 Sensitivity of regression results to multiple matches to the same ISO code in the Ruhlen database .....	12
3.3 Allele and phoneme frequency analyses.....	13

## 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](http://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?<sup>w</sup> and k?<sup>w?</sup> appear in the database, but can be considered equivalent, so we standardized these types of duplications into k?<sup>w</sup>).

#### 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 (<sup>h</sup>) only results in a phonemic distinction when it occurs in a language that also has the unaspirated version of the same phoneme. Similarly, dental (ꝝ) 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 `tʃ` 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](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 `tʃ` 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/iso639-2>.

[3/documentation.asp?id=nob](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  $j^{\text{th}}$  phoneme in the  $i^{\text{th}}$  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  $\geq 0$  km and  $<1000$  km were assigned to distance class 1,  $\geq 1000$  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  $\sim 10,000$  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 ( $64^{\circ}\text{N}, 177^{\circ}\text{E}$ ); Cairo, Egypt ( $30^{\circ}\text{N}, 31^{\circ}\text{E}$ ); Istanbul, Turkey ( $41^{\circ}\text{N}, 28^{\circ}\text{E}$ ); Phnom Penh, Cambodia ( $11^{\circ}\text{N}, 104^{\circ}\text{E}$ ); Prince Rupert, Canada ( $54^{\circ}\text{N}, 130^{\circ}\text{W}$ ); and Panama City, Panama ( $8.967^{\circ}\text{N}, 79.533^{\circ}\text{W}$ ). 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.  $AIC = -2 \ln(L(\hat{\theta}|Y)) + 2K$ , 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.  $BIC = -2 \ln(L(\hat{\theta}|Y)) + K \ln(n)$ , where  $n$  is the length of the vector  $Y$ . For a simple linear regression fit using the Ruhlen database,  $BIC - AIC = [-2 \ln(L(\hat{\theta}|Y)) + K \ln(n)] - [-2 \ln(L(\hat{\theta}|Y)) + 2K] = K \ln(n) - 2K = 3 \ln(2004) - 6 = 16.81$ .

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,

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

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,  $\mathbf{AB}$  represents the Mantel correlation between matrices **A** and **B**,  $\mathbf{AB} \bullet \mathbf{C}$  represents the partial Mantel correlation between matrices **A** and **B** controlling for matrix **C**, and  $\mathbf{AB}=0$  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  $\geq 0$  km and  $<1000$  km were assigned to distance class 1,  $\geq 1000$  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  $\sim 10,000$  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  $p < 2.4 \times 10^{-5}$ ) 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  $p < 1.4 \times 10^{-13}$ ). 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  $\log_{10}(\text{population size})$  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  $r=0.71$  ( $p = 4.9 \times 10^{-94}$ ). (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  $r=0.79$  ( $p = 1.24 \times 10^{-36}$ ).

**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](http://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

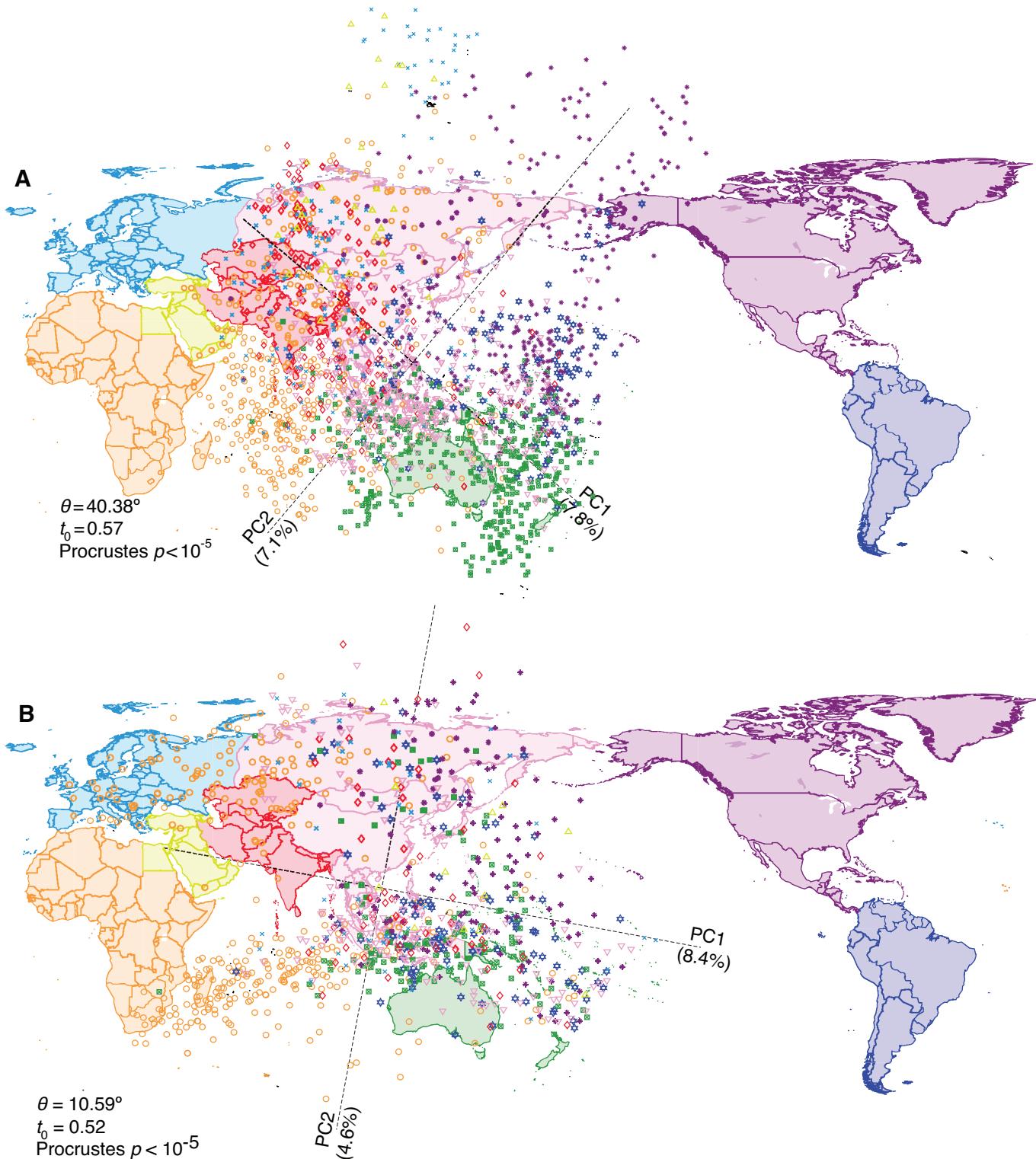


Fig. S2

A

Model	Expectations of the model	Model	Expectations of the model
<pre> A --&gt; B                v      C     </pre>	$AC \neq 0$ $BC \neq 0$ $AB = 0$ $AB \cdot C \neq 0$ $BC \cdot A \neq 0$ $AC \cdot B \neq 0$ $AC \cdot B \geq AC$ $BC \cdot A \geq BC$	<pre> A --&gt; B                v      C     </pre>	$AB \neq 0$ $BC \neq 0$ $AB \geq AC$ $AC \cdot B = 0$ $AB \cdot C \neq 0$ $BC \cdot A \neq 0$ $AB \cdot C \leq AB$ $BC \cdot A \leq BC$ $AB \times BC \approx AC$
<p>All of our Mantel results are consistent with this model</p> <div style="border: 2px solid red; padding: 10px;"> <p><b>Geography (A)</b></p> <pre> A --&gt; B                v      C     </pre> <p>Genes (B) Languages (C)</p> <math>AB \neq 0</math>  <math>AC \neq 0</math>  <math>AB \cdot C \neq 0</math>  <math>AC \cdot B \neq 0</math>  <math>BC \cdot A = 0</math>  <math>AC \cdot B \leq AC</math>  <math>AB \cdot C \leq AB</math>  <math>AB \times AC \approx BC</math> </div>			

B

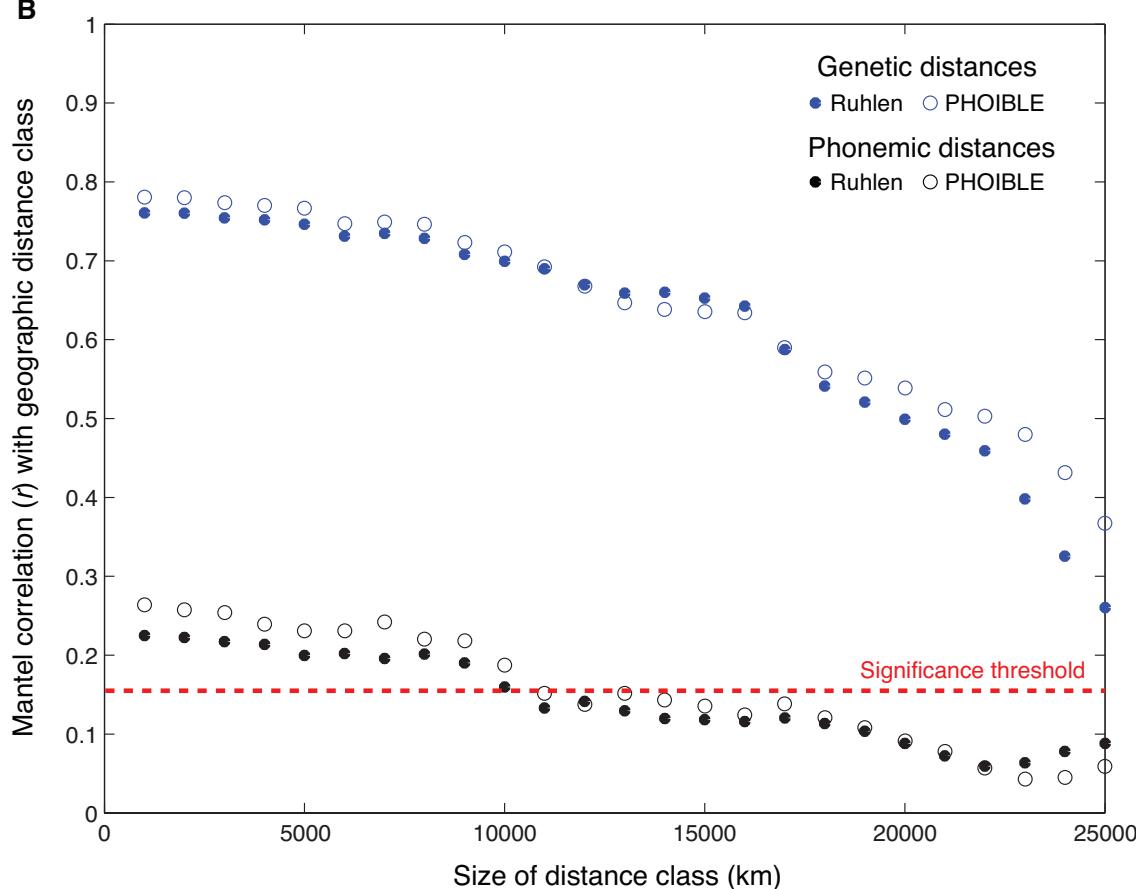


Fig. S3

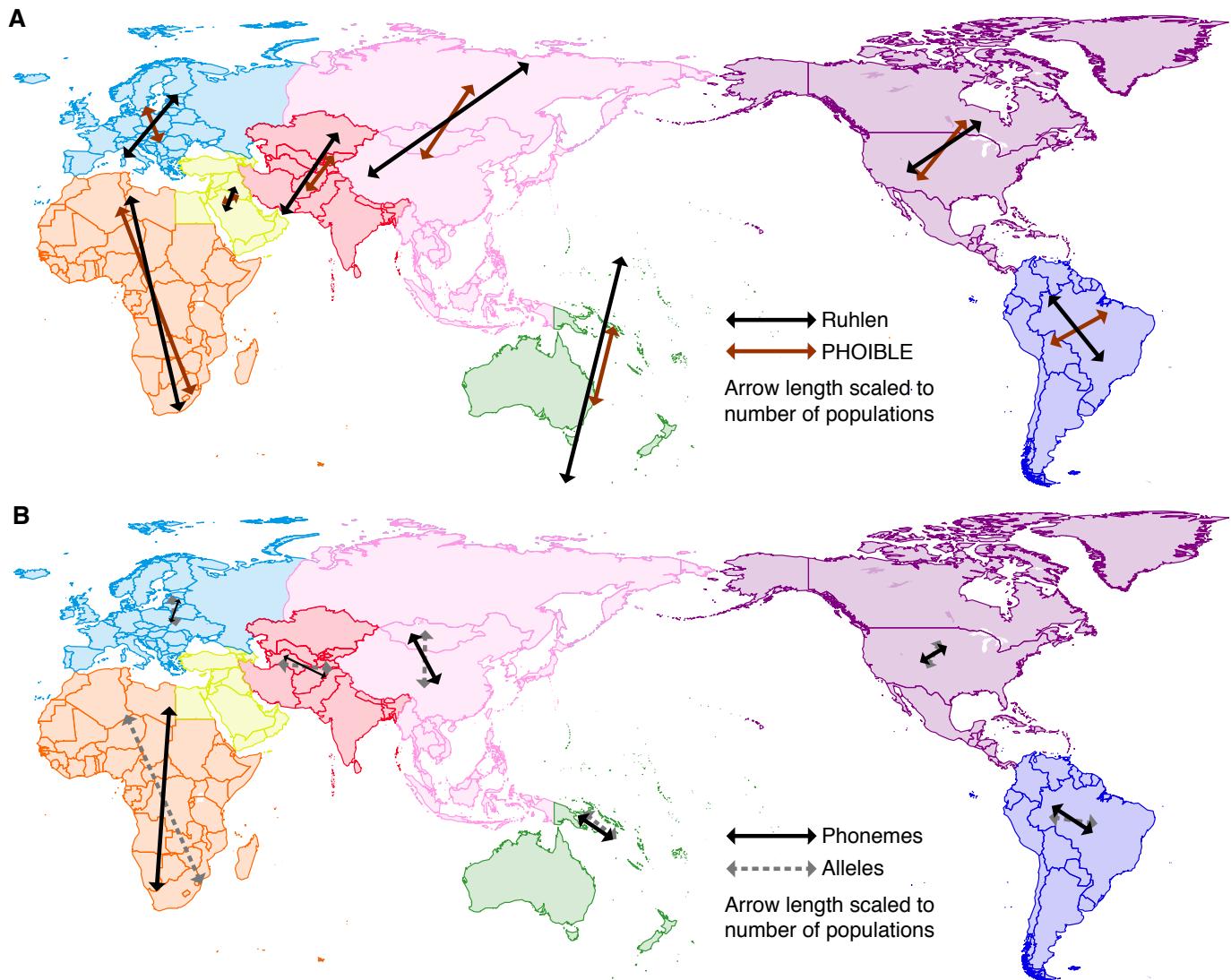


Fig. S4

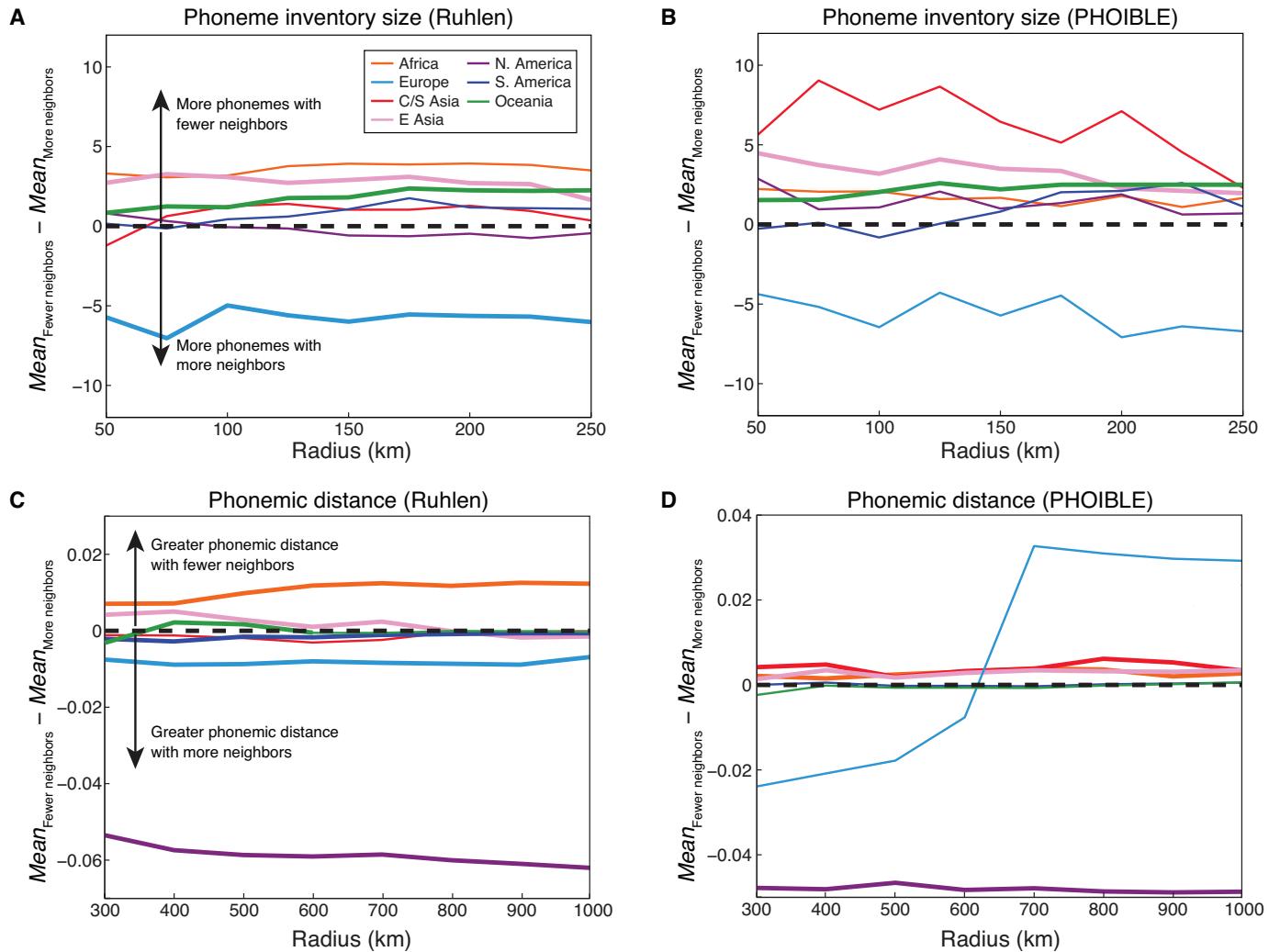


Fig. S5

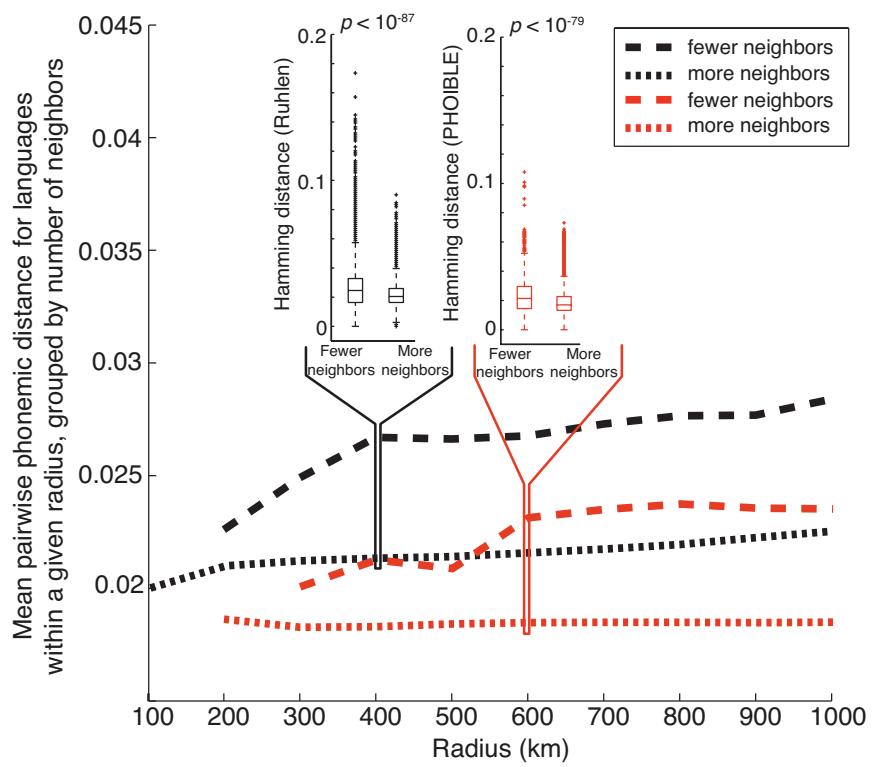


Fig. S6

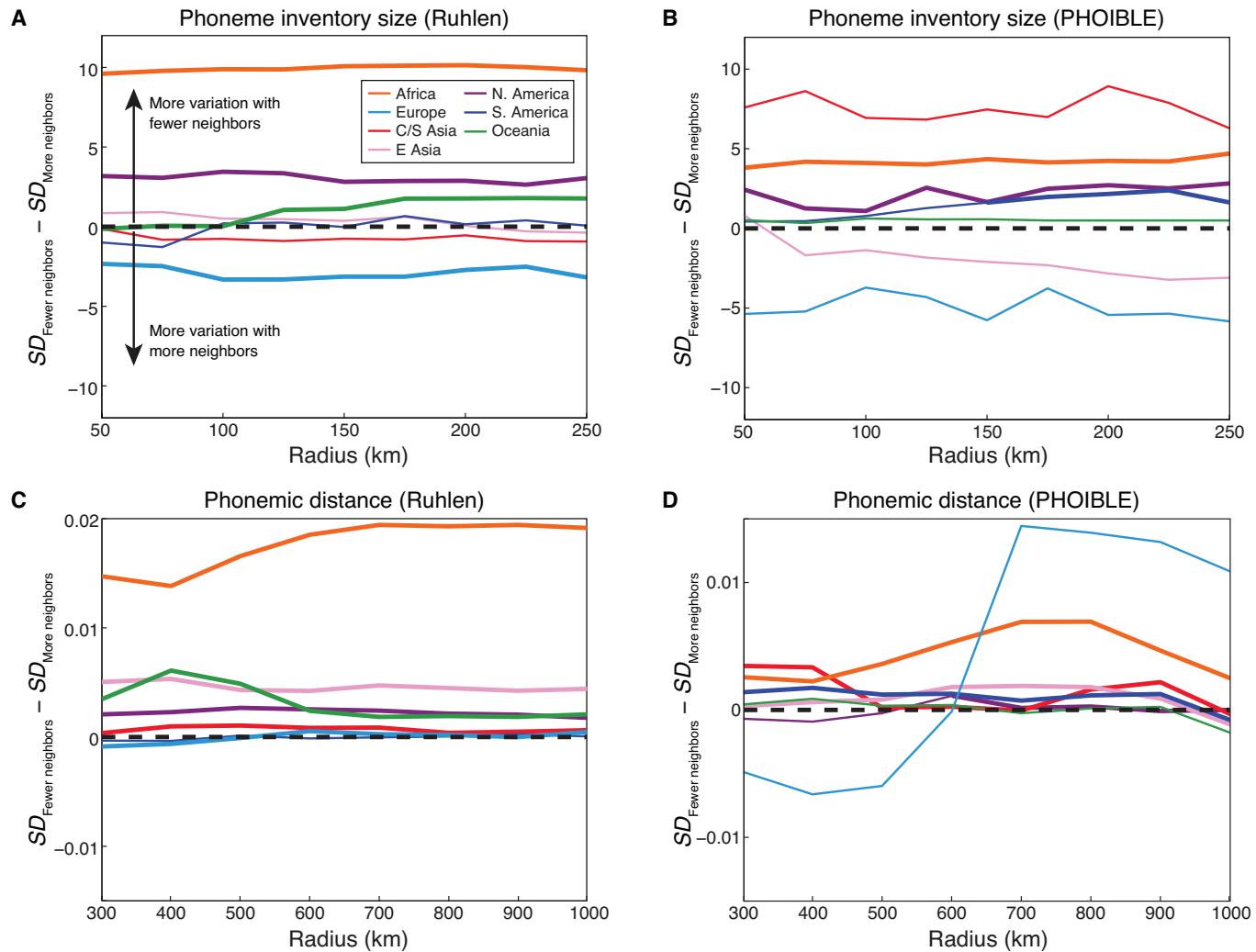


Fig. S7

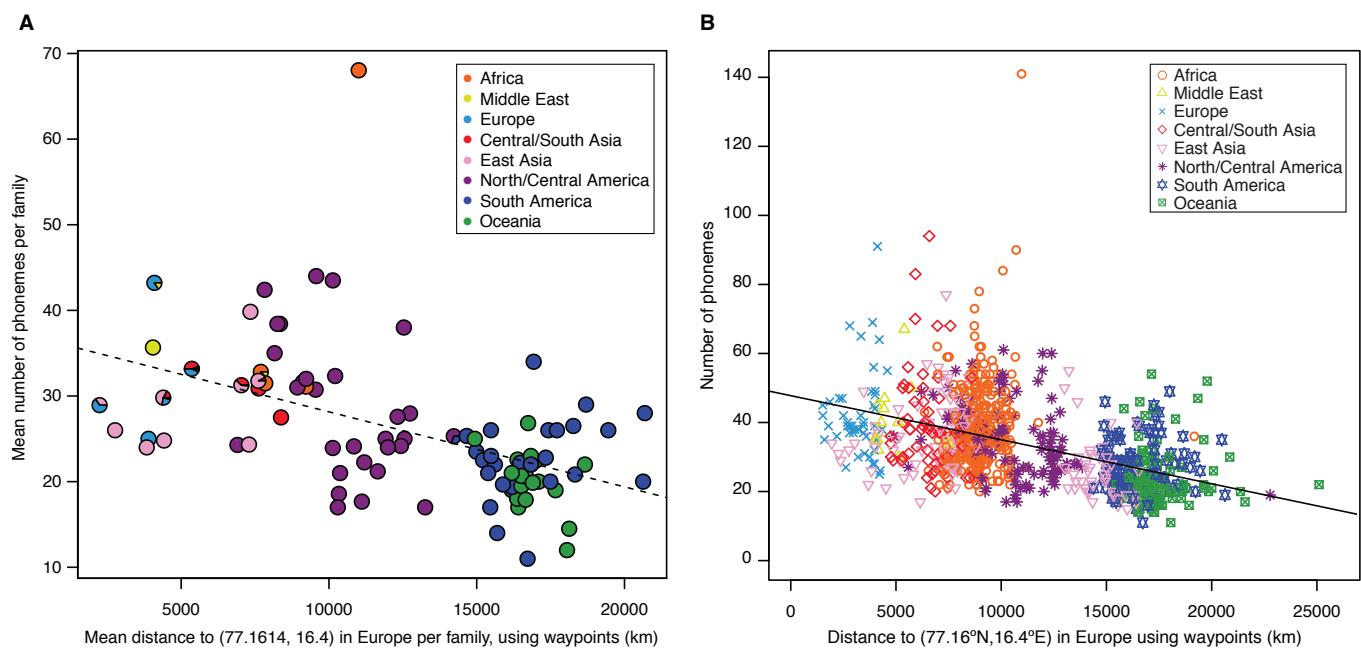


Fig. S8

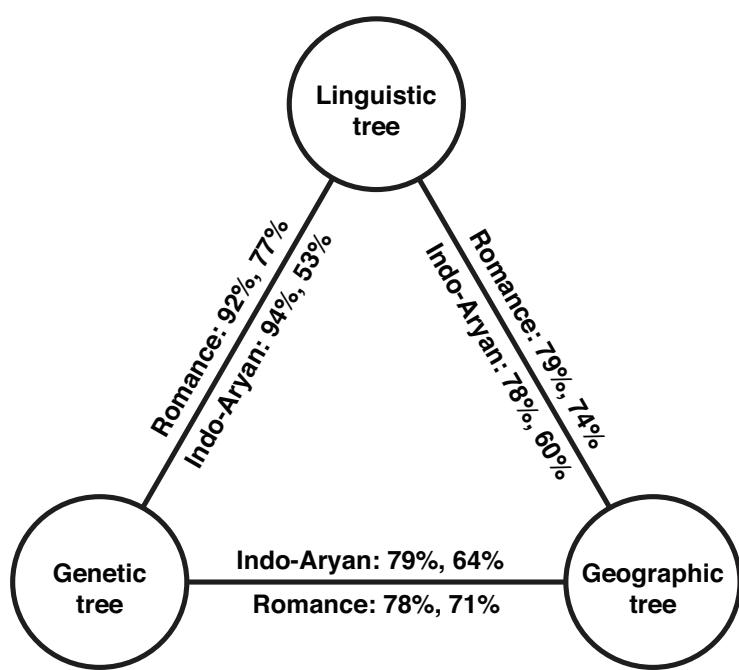


Fig. S9

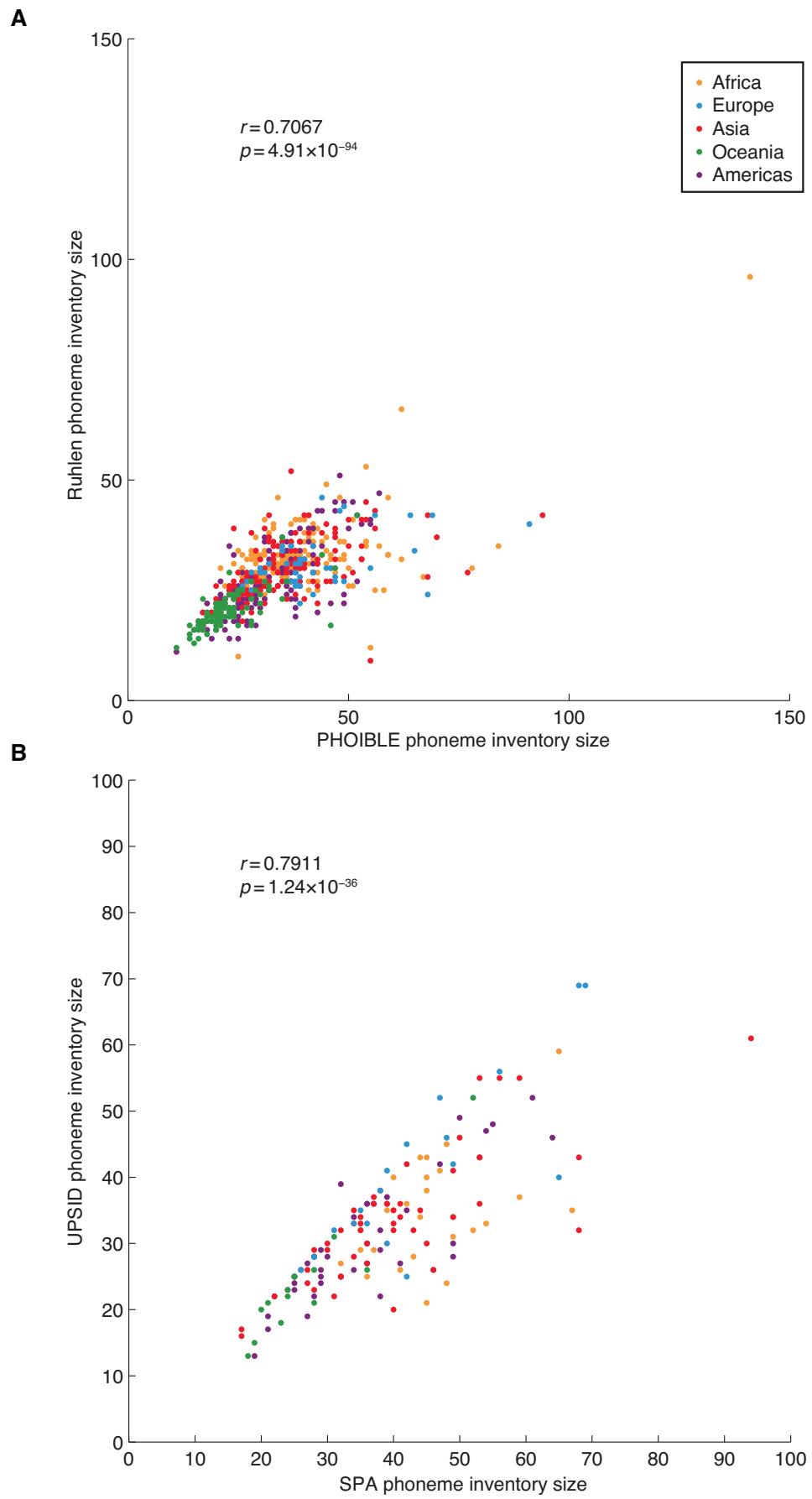


Fig. S10

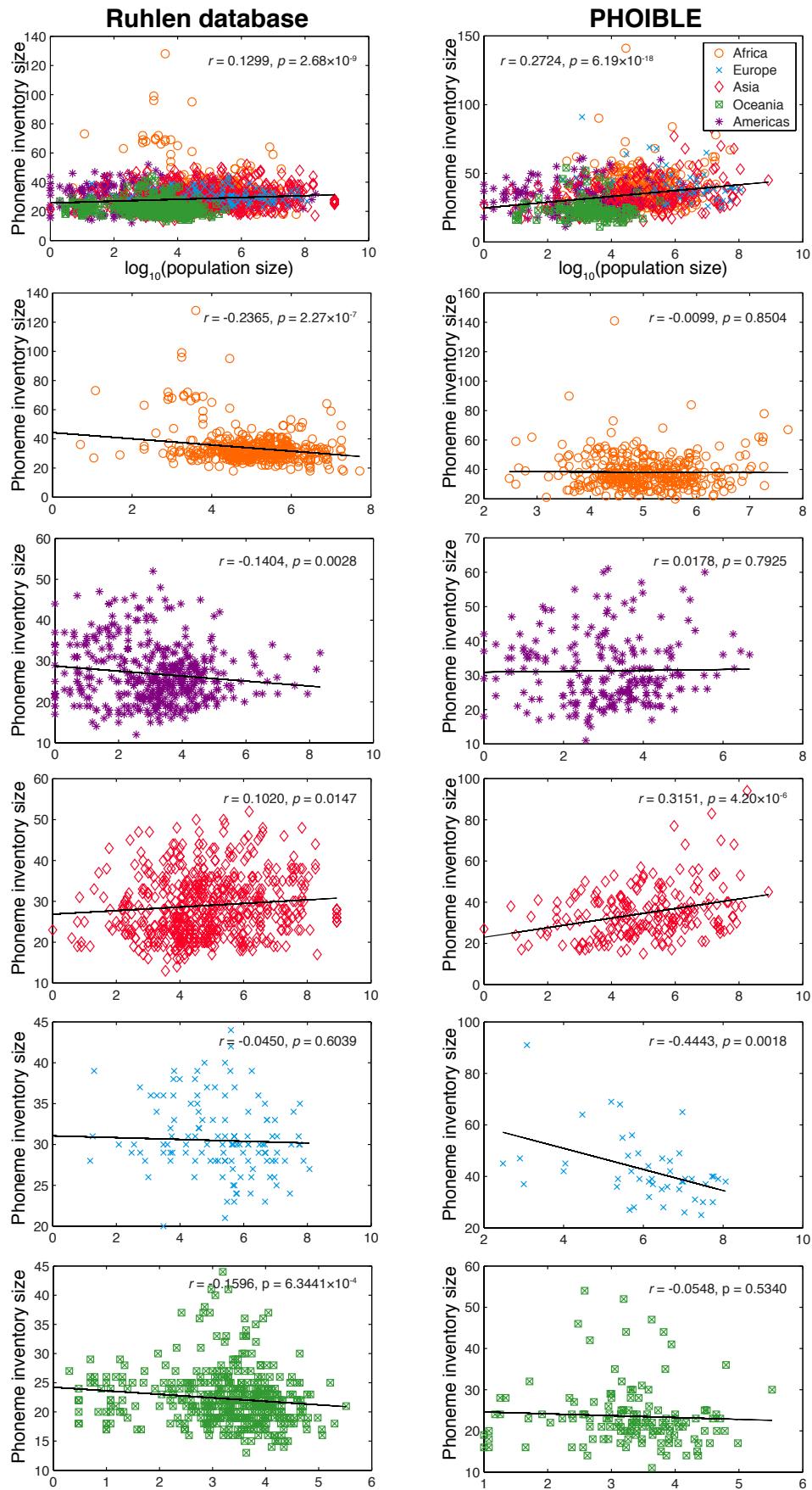


Fig. S11

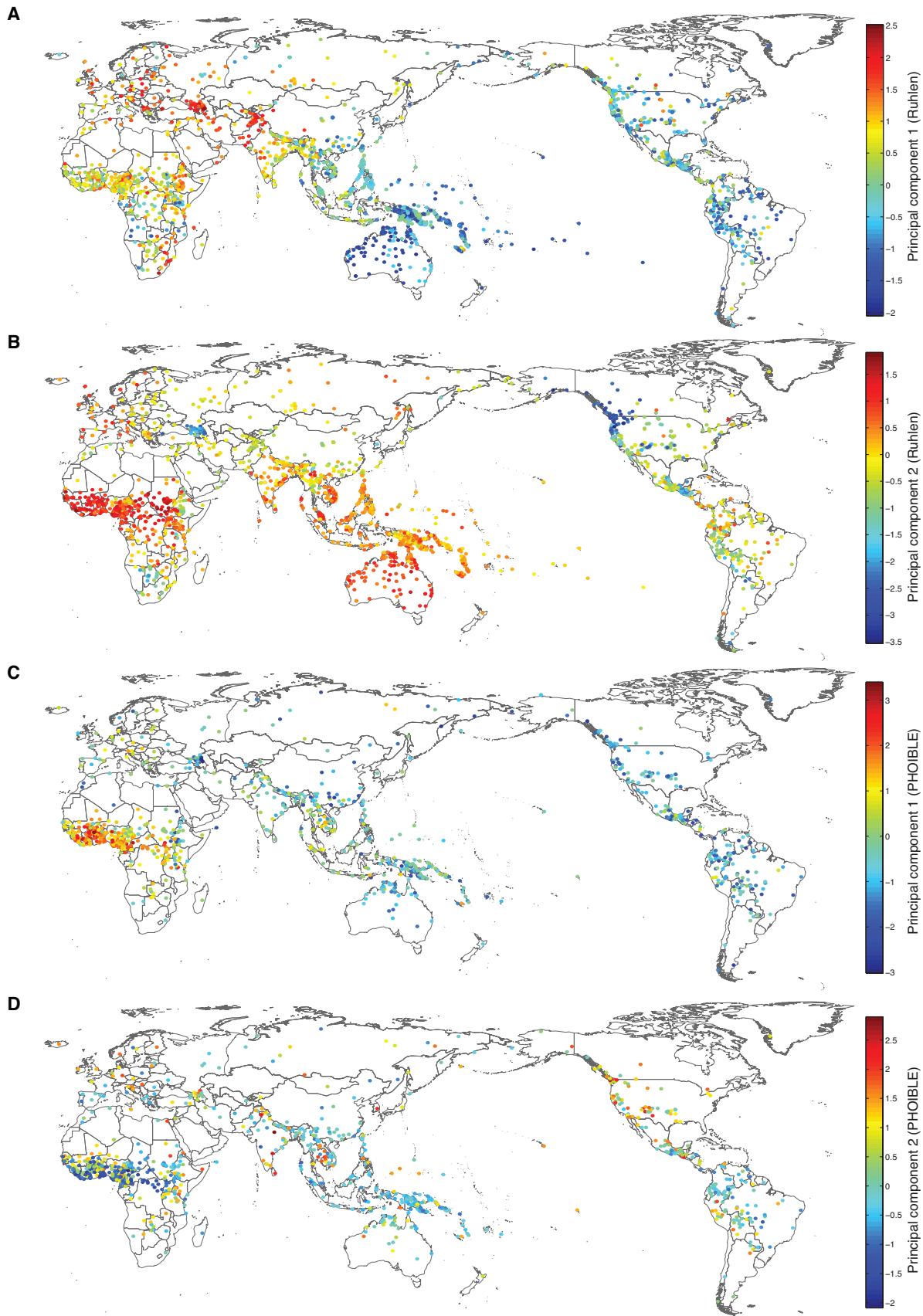


Fig. S12

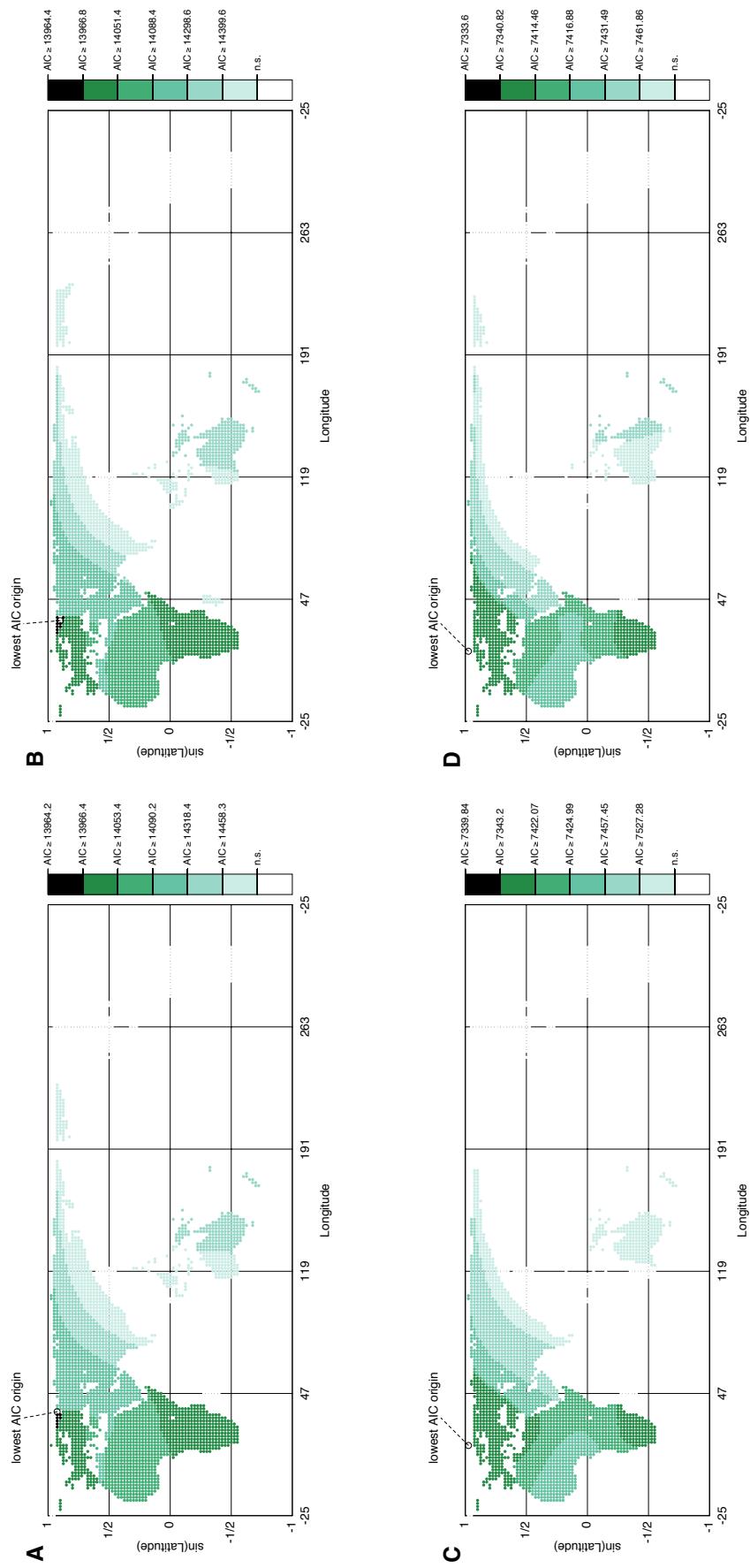


Fig. S13

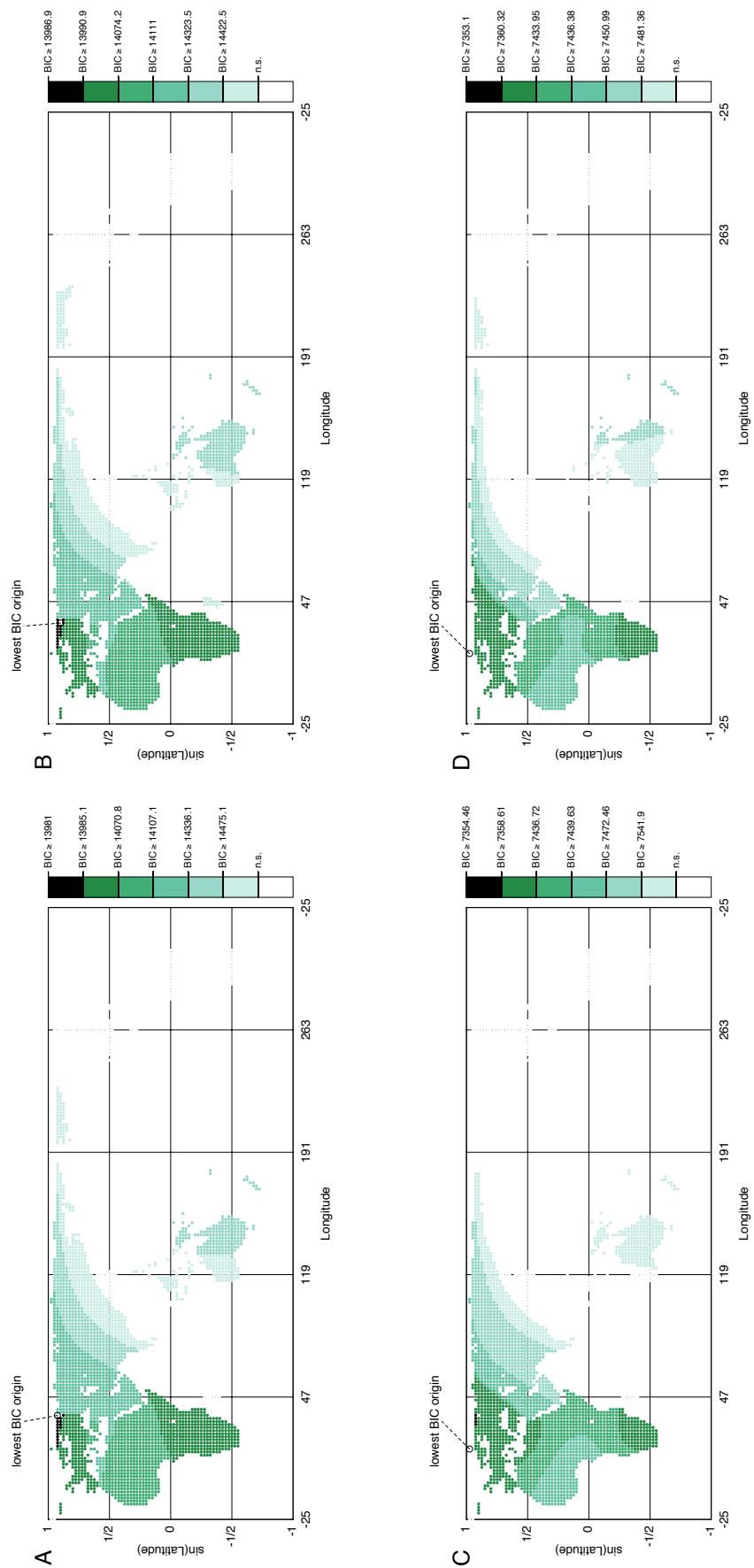


Fig. S14

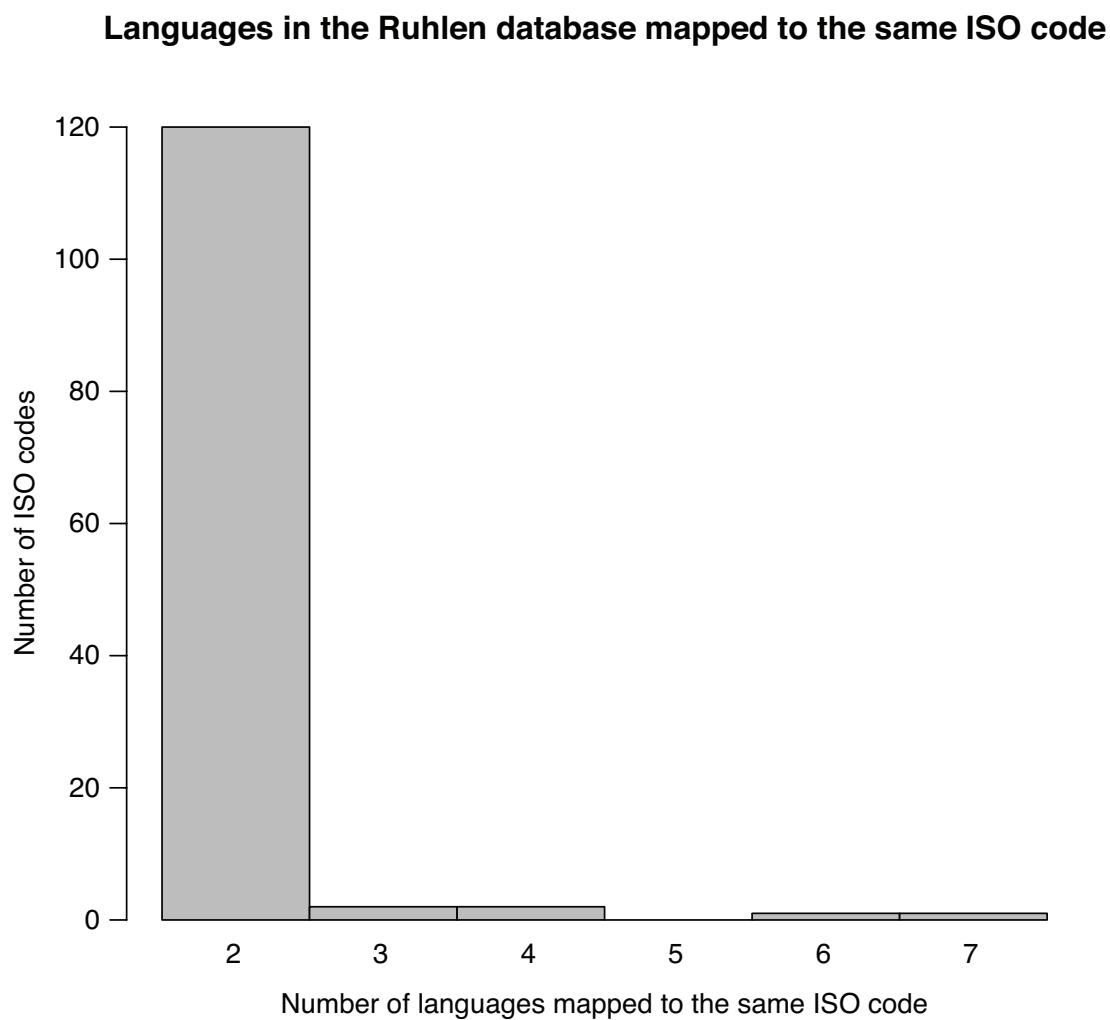
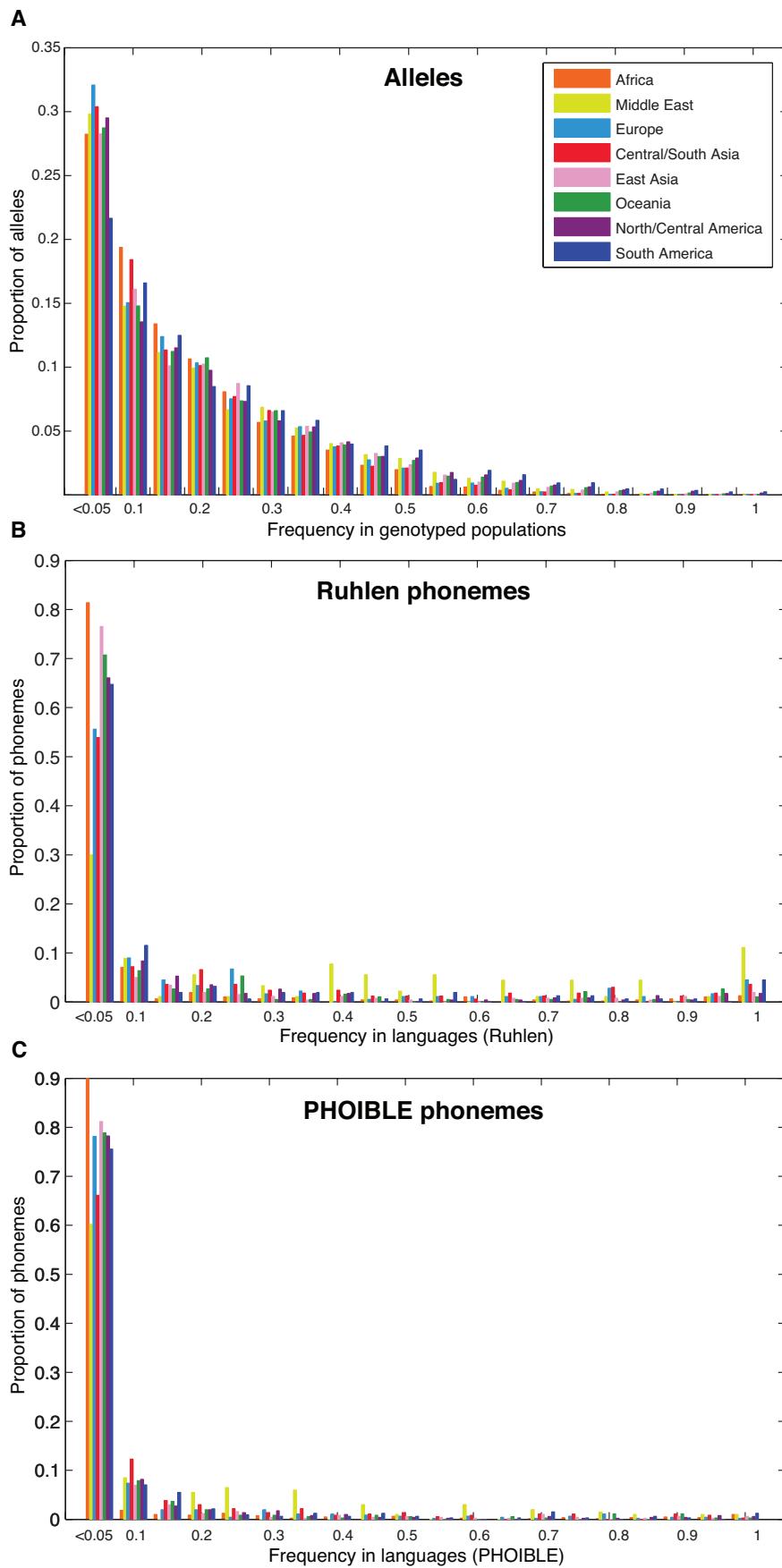


Fig. S15



**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
Worldwide	2082	0.5728	<b>1.00E-05</b>	0.1767	<b>1.00E-04</b>	0.0704	<b>1.00E-04</b>	0.277	<b>1.00E-04</b>
Africa	468	0.5022	<b>1.00E-05</b>	0.3026	<b>1.00E-04</b>	0.2363	<b>1.10E-03</b>	0.5277	<b>1.00E-04</b>
Middle East	32	0.6385	<b>1.00E-05</b>	0.4049	<b>1.00E-04</b>	0.4188	<b>1.00E-04</b>	0.403	<b>1.00E-04</b>
Europe	135	0.5888	<b>1.00E-05</b>	0.4128	<b>1.00E-04</b>	0.3775	<b>1.00E-04</b>	0.1825	<b>1.00E-04</b>
C./S. Asia	166	0.3681	<b>1.00E-05</b>	0.3446	<b>1.00E-04</b>	0.2208	<b>1.00E-04</b>	0.2769	<b>1.00E-04</b>
E. Asia	374	0.4299	<b>1.00E-05</b>	0.2977	<b>1.00E-04</b>	0.12	<b>5.00E-04</b>	0.2471	<b>1.00E-04</b>
N./C. America	305	0.4989	<b>1.00E-05</b>	0.2638	<b>1.00E-04</b>	0.1601	<b>3.00E-04</b>	0.2127	<b>1.00E-04</b>
S. America	147	0.2625	<b>5.00E-05</b>	0.0174	0.37296	-0.0093	0.48545	0.1213	<b>5.30E-03</b>
Oceania	455	0.4996	<b>1.00E-05</b>	0.2919	<b>1.00E-04</b>	0.2758	<b>1.00E-04</b>	0.3143	<b>1.00E-04</b>
Worldwide	968	0.5155	<b>1.00E-05</b>	0.2174	<b>1.00E-04</b>	0.0081	0.31547	0.3264	<b>1.00E-04</b>
Africa	362	0.347	<b>1.00E-05</b>	0.2718	<b>1.00E-04</b>	0.2091	<b>2.90E-03</b>	0.337	<b>1.00E-04</b>
Middle East	13	0.3823	0.3016	0.3499	<b>6.10E-03</b>	0.2033	0.12789	0.2932	0.020298
Europe	47	0.3577	<b>0.0041</b>	0.2824	<b>2.00E-04</b>	0.222	0.012999	0.1133	0.062594
C./S. Asia	58	0.0975	0.8032	0.1732	<b>5.20E-03</b>	-0.0688	0.78482	0.1387	0.019498
E. Asia	136	0.3186	<b>1.00E-05</b>	0.2339	<b>1.00E-04</b>	-0.0055	0.53095	0.2369	<b>1.00E-04</b>
N./C. America	122	0.2281	<b>0.003</b>	0.0458	0.23878	-0.0333	0.60704	0.1789	<b>1.00E-04</b>
S. America	99	0.2661	<b>0.0014</b>	0.2101	<b>1.00E-04</b>	0.16	0.013999	0.131	<b>9.10E-03</b>
Oceania	131	0.3905	<b>1.00E-05</b>	0.1857	<b>3.00E-04</b>	0.1755	0.016898	0.2125	<b>1.00E-04</b>

**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	<b>0.0243</b>	0.2708	<b>3.7×10<sup>-4</sup></b>	0.1801	<b>1.0×10<sup>-4</sup></b>	0.2652	<b>1.0×10<sup>-4</sup></b>
Phonemes vs. genes	0.1592	0.0577	0.3565	<b>1.0×10<sup>-5</sup></b>	0.1571	<b>2.3×10<sup>-3</sup></b>	0.2399	<b>2.0×10<sup>-4</sup></b>
Genes vs. geography	0.6995	<b>1.0×10<sup>-5</sup></b>	0.7851	<b>1.0×10<sup>-5</sup></b>	0.761	<b>1.0×10<sup>-4</sup></b>	0.781	<b>1.0×10<sup>-4</sup></b>

(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	<b>0.01</b>	0.1284	<b>0.009999</b>
Genetic vs. geographic distance controlling for phonemic distance	0.7495	<b>9.9×10<sup>-5</sup></b>	0.7664	<b>9.9×10<sup>-5</sup></b>
Phonemic vs. genetic distance controlling for latitudinal distance	0.1236	<b>9.9×10<sup>-3</sup></b>	0.1793	<b>4.00×10<sup>-4</sup></b>
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	<b>2.00×10<sup>-4</sup></b>	0.6031	<b>2.60×10<sup>-3</sup></b>

**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			Angle between linguistic and genetic axes
Region	Number of populations	Axis	Maximum Mantel $r$	p-value	Axis	Maximum Mantel $r$	p-value	
<b>Ruhlen database (139 populations)</b>								
Africa	62	13°–193°	0.269	<b>0.012</b>	3°–183°	0.455	<b>0.001</b>	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	<b>0.001</b>	3°
E. Asia	13	179°–359°	0.217	0.121	167°–347°	0.352	<b>0.009</b>	12°
N./C. America	8	39°–219°	0.594	<b>0.001</b>	139°–319°	0.544	<b>0.001</b>	78°
S. America	11	95°–275°	0.429	0.056	120°–300°	0.417	<b>0.001</b>	25°
Oceania	16	120°–300°	0.571	<b>0.004</b>	161°–341°	0.435	<b>0.005</b>	41°
<b>PHOIBLE populations (114 populations)</b>								
Africa	55	156°–226°	0.311	<b>0.012</b>	4°–184°	0.494	<b>0.001</b>	27°
Middle East	0	--	--	--	--	--	--	--
Europe	6	172°–352°	0.035	0.307	20°–200°	0.499	<b>0.021</b>	23°
C./S. Asia	13	96°–276°	0.185	0.194	116°–296°	0.228	0.118	20°
E. Asia	14	1°–181°	0.377	<b>0.034</b>	152°–332°	0.198	0.060	29°
N./C. America	6	25°–205°	0.899	<b>0.017</b>	59°–239°	0.146	0.442	34°
S. America	11	98°–278°	0.437	<b>0.018</b>	123°–303°	0.453	<b>0.001</b>	25°
Oceania	9	129°–309°	0.646	<b>0.001</b>	124°–304°	0.539	<b>0.001</b>	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	Qxú	100	Africa
6	Qxú	96	Africa
5	Qxú	95	Africa
29	Xú	76	Africa
40	N huki	72	Africa
32	G wi	71	Africa
30	G jana	69	Africa
17	G labake	68	Africa
24	Danisin	68	Africa
26	Khoe	68	Africa
31	G jana	68	Africa
1160	Xhosa	68	Africa
1	Hadza	67	Africa
27	Buka	67	Africa
28	Handa	67	Africa
18	G labake	65	Africa
33	Naron	63	Africa
37	ǂHó	63	Africa
4	Qxú	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 d'Alene	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	Tlameulula	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	Chulyym	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	Kalispele	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	Haroí	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	Anufo	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	A/C (not rescaled) of model for origin in columns 4 & 5	R <sup>2</sup>
Mean	Ruhlen	1	77.1614	16.4	<b>659.4515</b>	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	<b>563.0373</b>	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)	[:]: long
[u]: velarized	[.]: dental (consonants only)
[v]: pharyngealized	[.]: retroflex (consonants only)
[h]: aspirated	[.]: fortis
[?]: 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	ϐ	281	ϐ	Latin Letter Small Capital Inverted R
633	ɿ	279	ɿ	Latin Small Letter Turned R
\ldblquote	܂	127	܂	Latin Small Letter H With Stroke
\'98	܃	268	܃	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	<i>AIC</i> (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 <sup>1</sup>	2	64.1581	34.4	13964.45
	PHOIBLE <sup>2</sup>	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
Total number of phonemes, excluding clicks and modifications	Ruhlen	1	67.6684	36.2	12756.91
	Ruhlen	2	77.1614	16.4	12758.45

---

<sup>1</sup> 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.

<sup>2</sup> 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 A/C	Longitude of geographic center with lowest A/C	Continent of lowest-A/C center
Africa	Ruhlen	468	77.1614	16.4	Europe
	PHOIBLE	362	77.1614	16.4	Europe
Europe	Ruhlen	135	64.1581	-16	Europe
	PHOIBLE	47	77.1614	16.4	Europe
Middle East	Ruhlen	32	67.6684	36.2	Europe
	PHOIBLE	13	77.1614	16.4	Europe
Central/South Asia	Ruhlen	166	-33.367	27.2	Africa
	PHOIBLE	58	77.1614	16.4	Europe
East Asia	Ruhlen	374	67.6684	36.2	Europe
	PHOIBLE	136	77.1614	16.4	Europe
Oceania	Ruhlen	455	-31.6682	29	Africa
	PHOIBLE	131	42.4542	47	Asia
North/Central America	Ruhlen	305	-31.6682	27.2	Africa
	PHOIBLE	122	67.6684	36.2	Europe
South America	Ruhlen	147	67.6684	36.2	Europe
	PHOIBLE	99	77.1614	16.4	Europe