Original Paper

Human Heredity

Hum Hered 2010;70:245–254
DOI: 10.1159/000317374

Received: February 5, 2010
Accepted after revision: June 16, 2010
Published online: October 14, 2010

Long-Range Comparison between Genes and Languages Based on Syntactic Distances

Vincenza Colonna[a] Alessio Boattini[b] Cristina Guardiano[c] Irene Dall'Ara[a] Davide Pettener[b] Giuseppe Longobardi[d] Guido Barbujani[a]

[a]Dipartimento di Biologia ed Evoluzione, Università di Ferrara, Ferrara, [b]Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Bologna, [c]Dipartimento di Scienze del Linguaggio e della Cultura, Modena, and [d]Laboratorio di Linguistica e Antropologia Cognitiva, Dipartimento di Storia e Culture, Trieste, Italy

Key Words
Demographic history · Genetic distances · Genetic drift · Microsatellite markers · SNPs

Abstract
Objective: To propose a new approach for comparing genetic and linguistic diversity in populations belonging to distantly related groups. ***Background:*** Comparisons of linguistic and genetic differences have proved powerful tools to reconstruct human demographic history. Current models assume on both sides that similarities reflect either descent from common ancestry or the balance between isolation and contact. Most linguistic phylogenies are ultimately based on lexical evidence (roughly, words and morphemes with their sounds and meanings). However, measures of lexical divergence are reliable only for closely related languages, thus large-scale comparisons of genetic and linguistic diversity have appeared problematic so far. ***Methods:*** Syntax (abstract rules to combine words into sentences) appears more measurable, universally comparable, and stable than the lexicon, and hence certain syntactic similarities might reflect deeper linguistic relationships, such as those between distant language families. In this study, we for the first time compared genetic data to a matrix of syntactic differences among selected populations of three continents. ***Results:*** Comparing two databases of microsatellite (Short Tandem Repeat) markers and Single Nucleotides Polymorphisms (SNPs), with a linguistic matrix based on the values of 62 grammatical parameters, we show that there is indeed a correlation of syntactic and genetic distances. We also identified a few outliers and suggest a possible interpretation of the overall pattern. ***Conclusions:*** These results strongly support the possibility of better investigating population history by combining genetic data with linguistic information of a new type, provided by a theoretically more sophisticated method to assess the relationships between distantly related languages and language families.

Introduction

Human geneticists and historical linguists often seek answers to similar questions and face similar problems [1]. Starting from Sokal's [2] analysis of genetic distances in Europe, many anthropological and genetic studies have modeled population history and migration using language similarities as a cue of evolutionary relatedness. Most such studies (but not all: see e.g. [3]) showed that, in

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2010 S. Karger AG, Basel
0001–5652/10/0704–0245$26.00/0

Accessible online at:
www.karger.com/hhe

Guido Barbujani
Dipartimento di Biologia ed Evoluzione
Università di Ferrara
Via Borsari 46, IT–44100 Ferrara (Italy)
Tel. +39 0532 455 312, Fax +39 0532 249 761, E-Mail g.barbujani@unife.it

general, genetic change does parallel language change [4–12], hence both tend in general to reflect the same demographic processes. In turn, comparisons with genetic data might in principle cast some light on population histories where linguistic relationships are unresolved or controversial (e.g. [13–15]), even within the Indo-European family [16], although the linguistic tradition appears reluctant to incorporate genetic information in its analyses [17]. Many objections against this combined approach stem from the unreliability of distant linguistic comparisons of the sort needed to match genetic comparisons. This depends on the fact that safely identifiable similarities of words/morphemes in sound and meaning tend to dissolve within a short time span, sometimes placed around 8,000 ± 2,000 years [18]. Linguistic evolutions over longer time periods may be impossible to reconstruct from lexical comparisons, because of deceiving affinities emerging by sheer chance and the lexicon's inability to provide exact and broad-scope taxonomic (distance) measures. Accordingly, large-scale genetic studies [5] had to resort to very coarse classifications of languages, which are generally controversial among linguists. A recently explored [19] way out of the limits of lexical comparison may then be to focus on other aspects of language, such as the abstract syntactic characters identified within formal cognitive approaches and, more specifically, grammatical parameters of the sort increasingly investigated by theoretical syntax ever since Chomsky's original proposals [20, 21].

Phylogenetic Reconstructions in Linguistics

The basic problem for linguistic comparative methods is how to identify correspondences between languages which are significant against chance and thus call for historical explanation. The taxonomic characters typically used for that purpose are *lexical* entities, broadly understood as including both roots and grammatical morphemes, as well as sound laws connecting them crosslinguistically. Lexical items, when resembling each other in form and meaning (e.g. Eng. *thick*, Germ. *dick*), seem to provide the best proof of common origin.

However, by simply inspecting overall lexical items for sound/meaning similarities, serious problems may arise; many resemblances among words are simply due to chance: e.g. Eng. *much*, *day*, *have* and Spanish *mucho*, *dia*, *haber* are false cognates. By contrast, real cognates (e.g. Eng. *full* and It. *pieno*) may not look alike at all. This led in the 1870s to the establishment of a sophisticated chance-proof method, the *classical* comparative method [18, 22], focusing on more abstract entities, i.e. phonological correspondences ('sound laws') systematically recurrent throughout the lexicon (e.g. Eng. *th-* = Germ. *d-*: *thin, thief, thing* ... = *dünn, Dieb, Dinge* ...). Etymologies supported by such correspondences turn out to provide more reliable evidence of historical relatedness than superficial word resemblances. However, the very condition of its success (looking for very improbable, hence necessarily *rare*, agreements) has restricted the classical comparative method to close languages whose kinship is normally already obvious: for, regular sound correspondences between languages tend to disappear from their vocabularies in a few millennia after separation, leaving little hope of deeply digging into prehistory. Furthermore, languages not proven to be related display no common etymology, virtually by definition, therefore their distances are all automatically flattened on the theoretically admitted maximum, and become uninformative.

The Parametric Comparison Method (PCM) [23, 24] is a new method of language comparison based on the idea that the core grammar of any natural language can in principle be represented by a string of binary symbols (e.g. a series of + and – , along with the special case of a certain amount of null values [19]), each symbol coding the value of a linguistic parameter. Such strings of symbols can be unambiguously collated [25] and language distances and chance probability of agreements precisely measured. Furthermore, since parameters, much like genetic polymorphisms, are drawn from a supposedly universal list, all languages, no matter how distant, could in principle be compared by this method. Data collection in PCM has been tentatively conducted along the lines of the Modularized Global Parametrization program [23], aiming to reconcile relative typological coverage with accuracy and depth of analysis required by generative syntactic theory. Therefore, all data have been checked against native speakers: none could have been drawn solely from existing repertoires, which are all far from sufficiently detailed to set the parameters used.

The first phylogenetic experiments [19] have shown that PCM successfully meets several conceptual (chance probability and other expectations) and empirical (independently well established historical classifications) adequacy criteria and have also allowed the first formal demonstration that the rate of evolution in syntax is between 3 and 4 times slower than that of the corresponding vocabularies, as computed over a standard source [26].

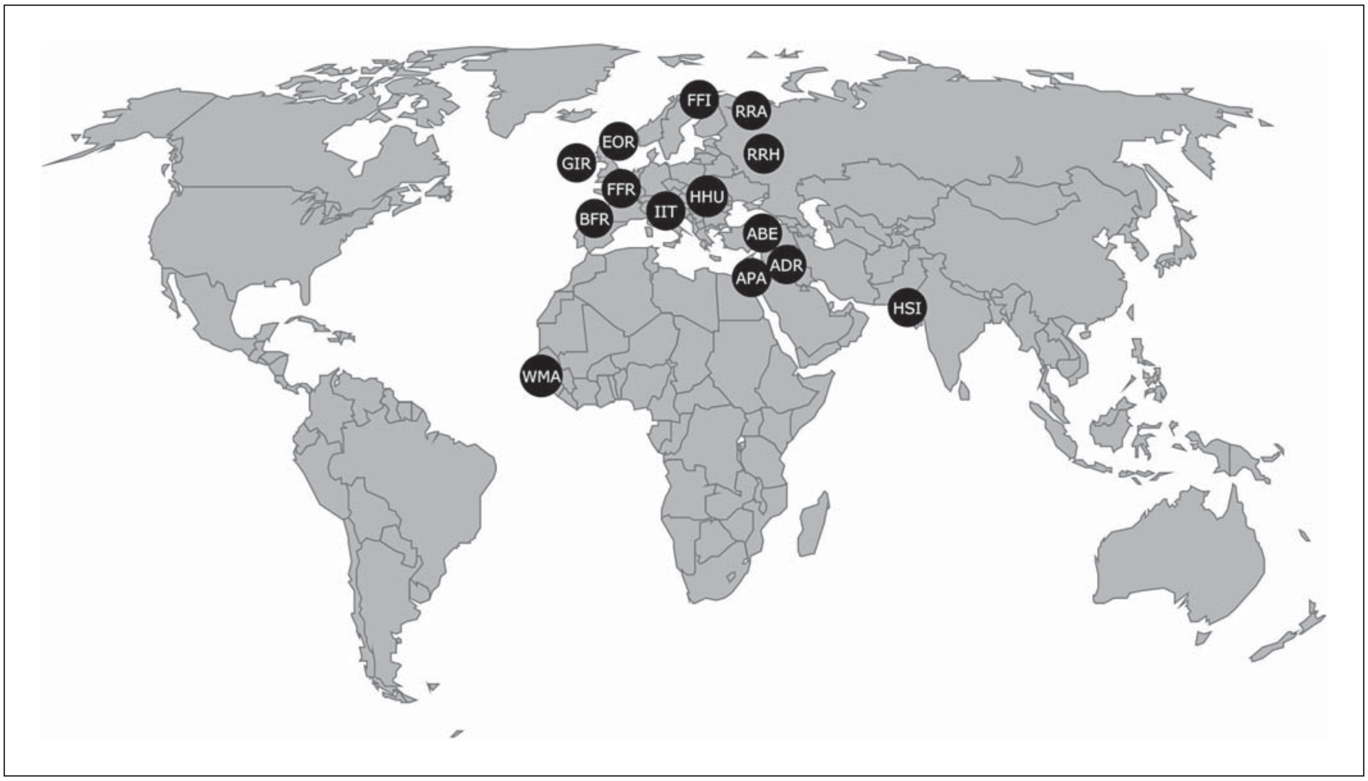

Fig. 1. Populations considered in this study. The three-letter population labels are detailed in table 1.

Material and Methods

Linguistic Distances

The 62 binary parameters used here, slightly adapted from Longobardi and Guardiano [19], describe one specific module of syntax in agreement with Guardiano et al. [23]: that of nominal arguments (roughly entity-denoting expressions such as *John, old John, my best friend, the friend I met yesterday, my friend's new book*, etc.). Their values were initially set in 18 Indo-European languages (Italian, Salentino, Spanish, French, Portuguese, Rumanian, Grico, Greek, English, German, Norwegian, Bulgarian, Serbo-Croatian, Russian, Slovenian, Irish, Welsh, and Hindi) and 6 non-Indo-European languages (Hebrew, Standard Arabic, Wolof, Hungarian, Finnish, and Basque).

According to Longobardi and Guardiano's procedure [19], the relationship between each pair of languages was first represented as an ordered pair of integers, *i* (a count of identities in parameter values) and *j* (a count of differences): when one or both languages of a pair happened to exhibit a null value for a parameter, that parameter was not counted at all for that pair; also, a few empirically uncertain states were counted as null for the purposes of these computations. Then, a measure of linguistic distance, d_{LAN}, was calculated by dividing the number of differences by the total number of parameters considered in the comparison: $d_{LAN} = j/(i + j)$ (essentially a Jaccard or Tanimoto distance [27]). A pair of languages identical for all parameters considered will thus have $d_{LAN} = 0$, a pair differing for all parameters will have $d_{LAN} = 1$, all other cases falling in between.

Genetic Data and Distances

Populations from the Human Genome Diversity Panel (HGDP)-CEPH version 2.0 [28] and from the ALFRED database [29, 30] were selected to try to match the languages compared by Longobardi and Guardiano [19] (fig. 1). Fourteen of the languages found a potential match in the chosen databases (table 1), though with two qualifications. The match between the Wolof language of Senegal (Atlantic subfamily of Niger-Congo) and the Senegalese Mandenka population available is admittedly far from satisfactory, since the latter is constituted by speakers of a different Niger-Congo language, namely one from the Mande subfamily. However, in the only study available (a mitochondrial analysis of Western African populations), the Mandenka appeared very similar to speakers of the Wolof subfamily [31]. The match between the Sindhi population and the Indo-Aryan language sampled (basically a variety of Western Hindi) is also hardly optimal; we can only hope that at the level of analysis of nominal syntax the differences between the two relevant language varieties will turn out relatively minor: from the scanty literature available at least some similarity can be guessed [32]. The genetic data among those publicly available at the HGDP CEPH Genotype Database V2.0 (ftp://ftp.cephb.fr/hgdp_v2/) consist of 784 autosomal Short Tandem Repeat (STR) loci and 3,840 autosomal Single Nucleo-

Table 1. Populations from the HGDP-CEPH and ALFRED databases matching the languages considered in [19]

Linguistic family	Language	Population	HGDP ID	ALFRED sample ID	Pop label	Coordinates	Individuals, n
Niger-Congo	Wolof*	Mandenka	Mandenka	SA001467S	WMA	12N, 12E	24
Basque	Basque	French Basque	French Basque	SA001504K	BFR	43N, 0	24
Uralic	Finnish	Finnish	*na*	SA000018J	FFI	67.5N, 27.5E	35
	Hungarian	Hungarian	*na*	SA002023H	HHU	47N, 19.5E	89
Afro-Asiatic	Arabic	Druze	Druze	SA002257Q	ADR	32N, 35E	48
	Arabic	Palestinian	Palestinian	SA001474Q	APA	32N, 35E	51
	Arabic	Bedouins	Bedouins	SA002254N	ABE	31N, 35E	49
Indo-European	Russian	Russian	Russian	SA001510H	RRH	61N, 40E	25
	Russian	Russian	*na*	SA000019K	RRA	64.32N, 40.34E	47
	Hindi*	Sindhi	Sindhi	SA001479V	HSI	25.5N, 69E	25
	French	French	French	SA001503J	FFR	46N, 2E	29
	English	Orcadian	Orcadian	SA001508O	EOR	59N, 3W	16
	Italian	Italian	North Italian	SA002255O	IIT	46N, 10E	14
	Gaelic	Irish	*na*	SA000057M	GIR	53.5N, 8.5W	113

Information about linguistic family from Ethnologue (http://www.ethnologue.com).
na = Not available. * Further explanation about the language-population match in the text.

tides Polymorphisms (SNPs), typed in a total of 305 subjects belonging to 10 populations (table 1). Of the two Italian populations present in the HGDP database, Tuscans were excluded since 78% of the 3,840 SNPs had >5% missing data. Data from the ALFRED database (http://alfred.med.yale.edu) consist of allele frequency for 671 SNPs typed in 598 subjects belonging to 14 populations (table 1). Genotypic data were inferred from allele frequencies by custom Python scripting and used for subsequent analyses. While all HGDP populations found a match in the ALFRED dataset, there is no match between ALFRED and HGDP polymorphisms except for two SNPs, namely rs1408801 and rs239031.

In order to separate, as far as possible, the DNA sites potentially subject to selection from neutral sites, the position of SNPs with respect to genes and region-surrounding genes were determined according to ENSEMBL homo_sapiens_variation_57_37b database (ftp://ftp.ensembl.org/pub/current_mysql/homo_sapiens_variation_57_37b/) using Biomart [33] (online supplementary table 1; for all online supplementary material, see www.karger.com/doi/10.1159/000317374). Thus, according to their position, SNPs were classified as genic (GEN) when falling in transcribed or regulatory regions, or intergenic (INT) otherwise. In both datasets, the vast majority of SNPs is in genic regions (76% and 78% in ALFRED and HGDP, respectively).

We estimated matrices of Reynolds et al. [34] and Cockerham and Weir F_{ST} [35] pairwise genetic distances using Arlequin version 3.5 software [36] separately for STRs and SNPs. We will refer thereafter to these matrices as d_{GEN}, in general or, when necessary, as d_{SNP} and d_{STR}. Significance of F_{ST}s was assessed by 10,000 permutations. Among several measures of genetic distance, Reynolds's and Cockerham and Weir's statistics are sensitive to the consequences of drift, and do not consider the effects of the mutation rate. We chose them in agreement with the widespread assumption (see e.g. [6, 37]) that associations between linguistic and genetic distances are caused by the common impact of isolation on both variables, leading to a process of language divergence paralleling the effects of genetic drift. Recent simulation studies have also shown that, contrary to other measures of genetic distance which perform better in comparisons between closely-related populations, Reynolds's index is appropriate for comparing genetically distant populations, such as populations of different continents [38]. Specific distance measures exist for both STRs and SNPs (e.g. [39, 40]) and were used in preliminary analyses, but, for reasons of internal consistency, we chose to use the same metrics for both STR and SNP comparisons. Thus, two sets of distances were calculated using SNPs in the GEN and INT category and noticeably the two sets of distances do not significantly differ (Kolmogorov-Smirnov D = 0.527; p = 0.000 and D = 0.543; p = 0.000 for Reynolds and F_{ST} distances, respectively).

Comparisons of Genetic and Linguistic Data

Matrices of geographic distances (d_{GEO}) between all pairs of populations were calculated by custom Python scripting in two ways: we estimated the great circle distances, or alternatively we took into account the likely migrational routes of early humans out of Africa, forcing the distances between populations through obligated waypoints as described in Ramachandran et al. [41]. Spearman nonparametric correlations coefficients (r) between pairs of distance matrices were estimated and their significance was tested according to the Mantel [42] tests procedure as described in Sokal and Rohlf [43]. In addition, because populations close in the geographical space tend to form genetic and linguistic clusters, we calculated partial correlations, i.e. coefficients comparing populations as if all pairs were at the same geographic distance. We did this by estimating the association of d_{GEN} (in fact, d_{SNP} and d_{STR})

and d_{LAN} by partial Mantel tests [44], with d_{GEO} held constant [45]. The significance of the correlation coefficients thus obtained was evaluated by permuting 10,000 times rows and columns of one distance matrix, while keeping the other matrix constant; this way, we generated a distribution of random pseudovalues of the statistics, against which the observed *r* values were compared. All calculations were carried out using the R Vegan package [46].

To graphically represent the genetic similarities between populations we used two nonparametric approaches, namely Multidimensional Scaling (MDS) and Procrustes analysis. For the MDS procedure, we started from the d_{GEN} and d_{LAN} matrices and projected the datapoints, corresponding to populations/languages, in a bi-dimensional space so that the distances between the points approximate the respective degree of dissimilarity [47]. Stress values relative to the MDS configurations were calculated according to Kruskal formula [48] and turned out to be always lower than the cutoff value according to Sturrock and Rocha [49] (that is 21.7 and 13.3% for 2 dimensions and 14 and 10 objects, respectively). The Procrustes method rotates a bi-dimensional set of points to maximum congruence with a target set of points by minimizing the sum of squared differences [50]. This way, once again using the R Vegan package [46], we compared MDS graphs based on genetic and linguistic distances.

Table 2. Mantel correlation between matrices of genetic distances calculated using different SNPs subsets

			ALL		GEN	
			r	p	r	p
ALFRED	Reynolds	GEN	**0.88**	**<0.001**	–	–
		INT	**0.71**	**<0.001**	**0.36**	**0.0302**
	F_{ST}	GEN	**0.88**	**<0.001**	–	–
		INT	**0.71**	**<0.001**	**0.36**	**0.0332**
HGDP	Reynolds	GEN	**1.00**	**<0.001**	–	–
		INT	**0.96**	**<0.001**	**0.94**	**<0.001**
	F_{ST}	GEN	**1.00**	**<0.001**	–	–
		INT	**0.96**	**<0.001**	**0.94**	**<0.001**

ALL = all SNPs; GEN = genic SNPs; INT = intergenic SNPs; r = Spearman correlation coefficient. Statistical significance at the $p < 0.05$ level highlighted in bold type. Data relative to the two datasets are reported.

Table 3. Mantel correlation and partial correlation coefficients, between genetic (STR), linguistic (LAN) and geographic (GEO) distance matrices for populations in the HGDP dataset

	Reynolds		Cockerham and Weir	
	r	p	r	p
STRs, n	784		784	
$r_{STR,LAN}$	**0.68**	**<0.001**	**0.68**	**<0.001**
$r_{STR,GEO}$	**0.70**	**<0.001**	**0.70**	**0.00**
$r_{STR,LAN.GEO}$	**0.69**	**<0.001**	**0.69**	**<0.001**

Linguistic distances based on the Parametric Comparison Method. Genetic distances calculated according to Reynolds and Cockerham and Weir's F_{ST}.

r = Spearman correlation coefficient. Statistical significance at the $p < 0.05$ level highlighted in bold type. All correlations are statistically significant after Bonferroni correction.

Results

Since neither linguistic nor genetic distances were normally distributed (Shapiro S = 0.96; p = 0.003 for d_{LAN}; see online suppl. table 2 for d_{GEN}) the nonparametric Spearman correlation coefficient was chosen for subsequent analyses. In all the analyses here reported, the results were significant after Bonferroni correction for multiple tests, unless otherwise specified.

As expected, the matrices of geographic distances were highly correlated ($r = 0.89$, $p < 10^{-4}$). However, in all comparisons with genes or languages, the correlation coefficients based on great circle distances were lower than those considering the routes of past migrations (online suppl. table 3) and so, for the sake of simplicity, we shall report only the results based on the latter. For both datasets, d_{LAN} showed only a statistically not significant correlation with d_{GEO} ($r_{GEO,LAN} = 0.30$, p = 0.098 for HGDP; $r_{GEO,LAN} = 0.26$, p = 0.077 for ALFRED). Pairs of genetic distance matrices were generally highly correlated ($r \geq 0.71$), with the exception of the matrices of distance between genic and intergenic polymorphisms estimated from the ALFRED dataset, where the correlation, although nominally significant ($p \approx 0.03$) was much lower (table 2).

Genetic distances inferred from the 784 STRs showed a high and consistent correlation with both d_{LAN} and d_{GEN} (table 3). The partial correlation coefficients show that the positive relationship between languages and genes is due only in minimal part to the common correlation with geography.

Two sets of genetic distances were calculated using SNPs (GEN and INT), and the two matrices did not significantly differ (Kolmogorov-Smirnov D = 0.527, p < 0.0001, and D = 0.543, p < 0.0001 for Reynolds and F_{ST} distances, respectively). However, they showed a more complicated pattern of association with other variables,

Table 4. Mantel correlation and partial correlation coefficients, between genetic (SNP), linguistic (LAN) and geographic (GEO) distance matrices for populations in the HGDP and ALFRED datasets

	Reynolds	ALFRED		ALFRED_10		HGPD	
		r	p	r	p	r	p
ALL	SNPs, n	671		671		3,840	
	$r_{SNP,LAN}$	0.20	0.123	**0.47**	**0.017**	**0.56**	**0.005**
	$r_{SNP,GEO}$	0.24	0.099	0.48	0.032	**0.70**	**<0.001**
	$r_{SNP,LAN.GEO}$	0.15	0.179	0.39	0.052	**0.51**	**0.013**
GEN	SNPs, n	509		509		2,992	
	$r_{SNP,LAN}$	0.24	0.098	**0.48**	**0.020**	**0.57**	**0.005**
	$r_{SNP,GEO}$	0.09	0.241	0.23	0.187	**0.70**	**<0.001**
	$r_{SNP,LAN.GEO}$	0.22	0.096	**0.44**	**0.035**	**0.53**	**0.011**
INT	SNPs, n	162		162		848	
	$r_{SNP,LAN}$	0.18	0.141	0.36	0.076	**0.49**	**0.026**
	$r_{SNP,GEO}$	**0.51**	**0.002**	**0.74**	**<0.001**	**0.66**	**0.002**
	$r_{SNP,LAN.GEO}$	0.06	0.329	0.22	0.175	0.40	0.061

ALFRED_10 is a subset of the ALFRED database including only populations also present in the HGDP panel. Linguistic distances based on the Parametric Comparison Method. Genetic distances calculated according to Reynolds using all SNPs in the datasets (ALL) or only those in 'genic' or 'intergenic' regions (GEN and INT, respectively).

r = Spearman correlation coefficient. Statistical significance at the $p < 0.05$ level highlighted in bold type. All correlations are statistically significant after Bonferroni correction.

depending on whether the SNPs considered fall in genic or intergenic chromosome regions (table 4). The ALFRED dataset showed in general a poor level of association between variables and no overall significance after Bonferroni correction; only for intergenic DNA polymorphisms did $r_{SNP,GEO}$ reach significance. Conversely, in the HGDP dataset correlations are high and significant, and particularly so for SNPs in coding regions of the genome.

To better understand whether the clear difference between the correlations estimated from the ALFRED and the HGDP datasets may reflect differences in the populations, or in the markers, considered, we redid all the calculations after removing four of ALFRED's populations, so that the geographical span of the comparisons was the same as in the analyses of the HGDP dataset. This time, we observed a general increase of the correlation coefficients, and in several cases the correlations achieved statistical significance (with the exception of $r_{SNP,LAN}$ for the intergenic polymorphisms, which, however, was based only on 162 DNA sites) (table 4). Two of the four samples thus removed represent Uralic, i.e. non-Indo-European, speakers of Europe (Hungarians and Finns), namely populations likely to disrupt the linear relationship between d_{GEO} and d_{LAN}. In all these analyses, $r_{GEN,LAN}$ appeared to be stronger in genic than in intergenic regions.

To identify which populations departed most from the general pattern of correlation between genetic and linguistic distances, we applied Procrustes analysis to the MDS graph (not reported) obtained from d_{GEN} versus d_{LAN} comparisons. Populations were plotted twice in a two-dimensional graph, based respectively on their linguistic and genetic relationships, and then we calculated the distance between pairs of points representing each population (namely residuals). When the two criteria yield exactly the same result, the distance is zero; when the two criteria are discordant, the distance between pairs of points is proportional to the difference observed between the two classification criteria. Results in figure 2 show residuals for all populations and suggest that the lack of complete correlation is mostly due to the behavior of Arabic-speaking (APA, ADR, ABE; for abbreviations, see table 1) and Basque populations, the most salient being Basques. Indeed, while from a genetic standpoint Basques are very similar to the other European populations, they represent a linguistic outlier, namely a non-Indo-European language isolate in Europe.

In contrast, the two groups for which we had only an approximate correspondence between language and biological population studied, namely the African Mandenka and the Asian Sindhi, show very low residuals (in fact, the lowest in the ALFRED dataset, fig. 2a), which is what one would expect under the hypothesis that there is indeed a parallelism between linguistic and genetic change. Therefore, although not a proof, this finding suggests that the approximation we had to resort to was, at least, a reasonable idealization.

Discussion

In this study we compared populations belonging to four major linguistic phyla [51, 52], namely Afro-Asiatic (Arabic), Niger-Congo (Wolof), Uralic (Finnish and Hungarian) and Indo-European (French, Irish, Italian, English, Russian and Sindhi), and a linguistic isolate, Basques; among Indo-European languages, four different branches were represented, namely Indo-Iranian, Italic, Balto-Slavic and Germanic. Estimating linguistic distances among these samples from lexical comparisons would have been highly problematic, or simply unwarranted [17].

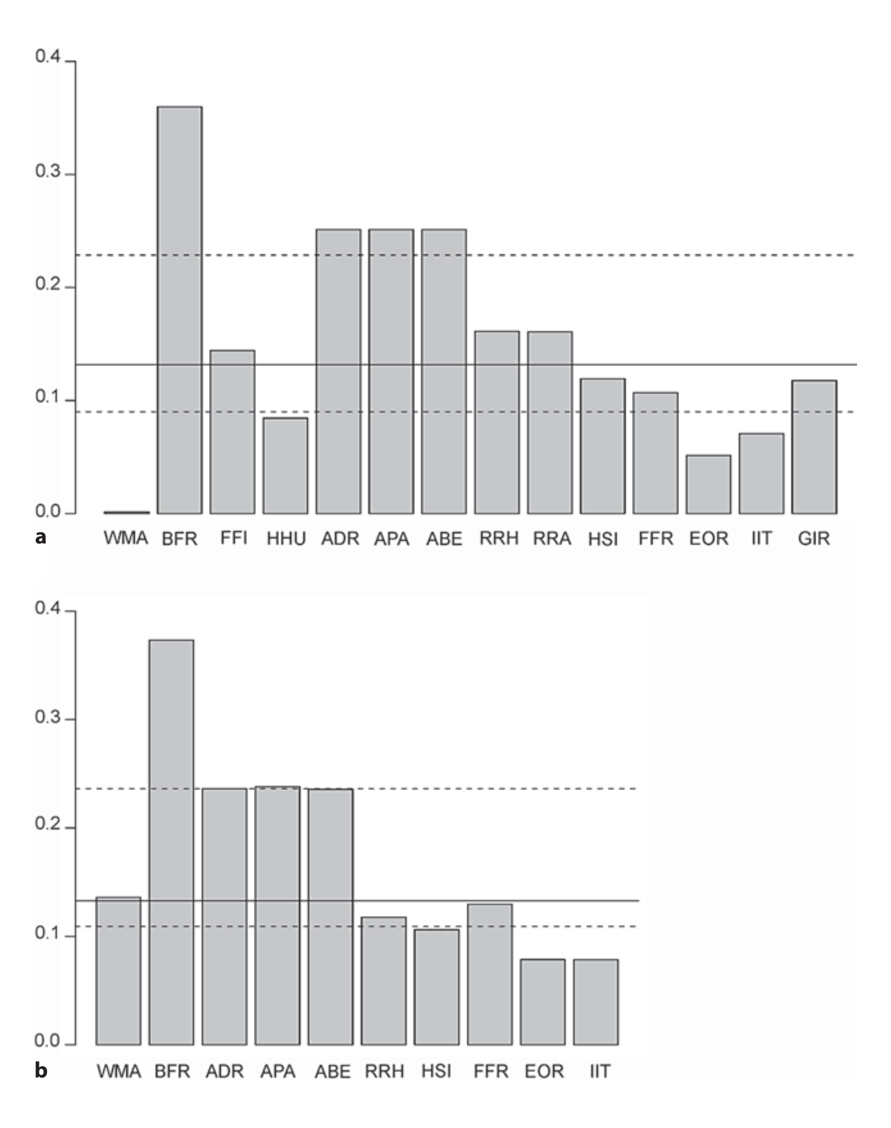

Fig. 2. Procrustes residuals based on Reynolds d_{GEN} and d_{LAN} for the ALFRED (**a**) and HGDP (**b**) datasets. Solid line = median value; dashed lines = 95% confidence interval based on standard deviation. Population labels as in table 1.

The preliminary analyses that we ran show that Longobardi and Guardiano's [19] index of distance based on syntax, besides being grounded in grammatical theory and well matched by traditional linguistic phylogenies, shares some useful empirical properties with other indices that are popular among geneticists. Indeed, this measure shows a broad general correlation with genetic distance, and allows one to identify outlier populations. Because these observations are based on a small set of populations, scattered across a broad subset of the planet's language diversity, these results were not necessarily expected. Indeed, among an estimated total of approximately 5,000 languages spoken today, only for a handful the similarities with others are so loose that they are classified as linguistic isolates, but in this study the isolate, Basque, represented up to 1/8 of the total. In addition, we had cases of different populations speaking the same language, Russian and Arabic in particular ($d_{LAN} = 0$) which has necessarily reduced the correlations, because these populations cannot be genetically identical ($d_{GEN} = 0$). Both these factors were likely to disrupt the simple linear relationship expected between d_{GEN} and d_{LAN}, but in fact they did not. The correlations appeared significant, and generally high, when d_{GEN} were estimated both from SNP and STR markers, which is encouraging for future, larger-scale applications of the methodology. The additional tests we ran show that the association between languages and genes becomes tighter when the effects of geography are held constant, and when the main outlier is removed.

Therefore, the results of our analyses seem robust to a number of possible complications of the study design, hence they clearly suggest that analyses of broader sets of data representing a more detailed sampling of human biological and cultural diversity (crucially including syntactic distances calculated by PCM) will yield interpretable results and indeed allow previously impossible large-scale comparisons.

We do not think we can draw firm general conclusions on human genetic and linguistic diversity at this stage, but some patterns are already evident and, if confirmed by analyses of larger datasets, will need to be better understood. First, most genetic distances (i.e. those calculated from all HGDP data and intergenic SNPs of the ALFRED database) showed the well-known positive correlation with geographic [6, 10, 53] and linguistic distances, in agreement with many previous studies based on lexical comparisons [1, 2, 9, 38]. However, these correlations are now shown to exist over several continents, because for the first time they were estimated on the basis of a robust measure of linguistic distance, suitable for such long-range comparisons.

We observed a slightly closer relationship of syntactic distances with genetic distances inferred from STR, rather than SNP, variation. This seems a rather general conclusion of our study, irrespective of the specific d_{GEN} index estimated. The main difference between the genetic systems considered might lie in the different mutational mechanisms, or different mutation rates, generating different geographical patterns of STR and SNP diversity, the former more closely comparable to the relatively fast-evolving differences in syntax.

We found no substantial differences between SNP markers in genic or intergenic regions, when considering $r_{GEN,LAN}$, even though variation at SNPs mapping in or near coding regions might reflect, to an extent difficult to define in advance, the effects of selection. This is not really surprising, because (a) SNPs falling in coding genome regions are not necessarily subjected to selection, and (b) there is by now rather extensive evidence that, even for loci subjected to selection, selective pressures are generally mild, and do not substantially contribute to determining patterns of genetic diversity over much of the world [54, 55]. In any case, a priority for future, larger studies will be to run preliminary neutrality tests, in the hope of discriminating the effects of selection from those of mutation.

The measures of genetic distance we used throughout the study basically represent population differences as due to genetic drift, disregarding the effects of mutation. We thought this was the proper choice, as it is well known that mutations accumulate across long evolutionary periods, whereas the frequencies of allelic variants, no matter whether they are estimated at the DNA level or at the protein level, fluctuate rapidly because of drift [11, 56]. The latter mechanism seemed to be the most suitable to model population process occurring at the time-scale of linguistic change, which is also deeply affected by population contacts and isolation [57].

Future developments of this analysis will require assemblage of broader linguistic and genetic databases. Once these datasets will be ready for comparison, we plan to get into finer detail in the study of the effects of demographic history on genes and languages. Of course, we know in advance that human population history has been complex, and therefore identifying patterns for specific geographical regions and markers is no guarantee that the same patterns will be observed elsewhere or with other markers [58]. However, both the rule (whether, in which parts of the world, and to what extent, there is indeed a correlation between syntactic differences and genetic diversity) and the exceptions (outlier populations, defined on the basis of syntax or gene pool) will be fruitful and interesting to investigate.

Acknowledgments

This study was partly supported by the Italian Ministry of University PRIN 2008 grant to G.B., and by the 'Strategic Project 2006–2009' grant of the University of Bologna to D.P.; V.C. was supported by an Italian Ministry of University (MIUR) FIRB grant.

References

1 Atkinson QD, Gray RD: Curious parallels and curious connections – phylogenetic thinking in biology and historical linguistics. Syst Biol 2005;54:513–526.

2 Sokal RR: Genetic, geographic, and linguistic distances in Europe. Proc Natl Acad Sci USA 1988;85:1722–1726.

3 Mona S, Grunz KE, Brauer S, Pakendorf B, Castri L, Sudoyo H, Marzuki S, Barnes RH, Schmidtke J, Stoneking M, Kayser M: Genetic admixture history of Eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. Mol Biol Evol 2009;26: 1865–1877.

4 Barbujani G, Sokal RR: Zones of sharp genetic change in Europe are also linguistic boundaries. Proc Natl Acad Sci USA 1990; 87:1816–1819.

5 Belle EM, Barbujani G: Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. Am J Phys Anthropol 2007;133:1137–1146.

6 Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J: Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc Natl Acad Sci USA 1988;85:6002–6006.

7 Lansing JS, Cox MP, Downey SS, Gabler BM, Hallmark B, Karafet TM, Norquest P, Schoenfelder JW, Sudoyo H, Watkins JC, Hammer MF: Coevolution of languages and genes on the island of Sumba, Eastern Indonesia. Proc Natl Acad Sci USA 2007;104: 16022–16026.

8 Penny D, Watson EE, Steel MA: Trees from Languages and Genes Are Very Similar. Systematic Biology 1993;42:382–384.

9 Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L: Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. Am J Hum Genet 1997;61:1015–1035.

10 Quintana-Murci L, Krausz C, Zerjal T, Sayar SH, Hammer MF, Mehdi SQ, Ayub Q, Qamar R, Mohyuddin A, Radhakrishna U, Jobling MA, Tyler-Smith C, McElreavey K: Y-chromosome lineages trace diffusion of people and languages in Southwestern Asia. Am J Hum Genet 2001;68:537–542.

11 Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus ML, Aula P, Beckman L, Tranebjaerg L, Gedde-Dahl T, Issel-Tarver L, DiRienzo A, Paabo S: Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 1995;5:42–52.

12 Torroni A, Schurr TG, Yang CC, Szathmary EJ, Williams RC, Schanfield MS, Troup GA, Knowler WC, Lawrence DN, Weiss KM, et al: Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 1992;130: 153–162.

13 Li H, Wen B, Chen SJ, Su B, Pramoonjago P, Liu Y, Pan S, Qin Z, Liu W, Cheng X, Yang N, Li X, Tran D, Lu D, Hsu MT, Deka R, Marzuki S, Tan CC, Jin L: Paternal genetic affinity between Western Austronesians and Daic populations. BMC Evol Biol 2008;8:146.

14 Nasidze I, Sarkisian T, Kerimov A, Stoneking M: Testing hypotheses of language replacement in the Caucasus: evidence from the Y-chromosome. Hum Genet 2003;112:255–261.

15 Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL: History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol 2007;24:2180–2195.

16 Mallory J: In Search of the Indo-Europeans: Language, Archaeology and Myth. London, Thames and Hudson, 1993.

17 Bateman R, Goddard I, O'Grady R, Funk VA, Mooi R, Kress WJ, Cannell P: Speaking of forked tongues. The feasibility of reconciling human phylogeny and the history of language. Curr Anthropol 1990;31:1–24.

18 Nichols JA: The comparative method as heuristic; in Durie M, Ross M (eds): The Comparative Method Reviewed: Regularity and Irregularity in Language Change. New York, Oxford University Press, 1996, pp 39–71.

19 Longobardi G, Guardiano C: Evidence for syntax as a signal of historical relatedness. Lingua 2009;119:1679–1706.

20 Biberauer T: The Limits of Syntactic Variation. Amsterdam/Philadelphia, John Benjamins, 2008.

21 Chomsky N: Lectures on Government and Binding. Dordrecht, Foris, 1981.

22 Meillet A: La Méthode Comparative en Linguistique Historique. Paris, Champion, 1928.

23 Guardiano C, Longobardi G: Parametric comparison and language taxonomy; in Batllori M, Hernanz ML, Picallo C, Roca F (eds): Grammaticalization and Parametric Variation. Oxford, Oxford University Press, 2005, pp 149–174.

24 Longobardi G: Methods in parametric linguistics and cognitive history. Linguistic Variation Yearbook 2003;3:101–138.

25 Roberts I: Review of Harris A & Campbell L, Historical Syntax in Cross-Linguistic Perspective. RomPh 1998;51:363–370.

26 Dyen I, Kruskal J, Black P: An Indoeuropean classification: a lexicostatistical experiment. Trans Am Phil Soc 1992;82.5:1–132.

27 Deza E, Deza MM: Dictionary of Distances. Amsterdam, Elsevier, 2006.

28 Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL: A human genome diversity cell line panel. Science 2002;296:261–262.

29 Osier MV, Cheung KH, Kidd JR, Pakstis AJ, Miller PL, Kidd KK: ALFRED: an allele frequency database for diverse populations and DNA polymorphisms – an update. Nucleic Acids Res 2001;29:317–319.

30 Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, Kidd KK: ALFRED – the ALlele FREquency Database – update. Nucleic Acids Res 2003;31: 270–271.

31 Cerný V, Hájek M, Bromová M, Cmejla R, Diallo I, Brdicka R: MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. Human Biology 2006;78:9–27.

32 Cole J: Sindhi; in Strazny P (ed): Encyclopedia of Linguistics. New York, Routledge, 2005.

33 Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A: BioMart Central Portal – unified access to biological data. Nucleic Acids Res 2009;37:W23–W27.

34 Reynolds J, Weir BS, Cockerham CC: Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 1983;105:767–779.

35 Cockerham CC, Weir BS: Covariances of relatives stemming from a population undergoing mixed self and random mating. Biometrics 1984;40:157–164.

36 Excoffier L, Lischer HEL: Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour 2010, in press.

37 Sokal RR, Oden NL, Thomson BA: Genetic changes across language boundaries in Europe. Am J Phys Anthropol 1988;76:337–361.

38 Libiger O, Nievergelt CM, Schork NJ: Comparison of genetic distance measures using human SNP genotype data. Hum Biol 2009; 81:389–406.

39 Nei M, Tajima F, Tateno Y: Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J Mol Evol 1983;19:153–170.

40 Rousset F: Equilibrium values of measures of population subdivision for stepwise mutation processes. Genetics 1996;142:1357–1362.

41 Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: 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 2005;102:15942–15947.

42 Mantel N: The detection of disease clustering and a generalized regression approach. Cancer Res 1967;27:209–220.

43 Sokal RR, FJ Rohlf: Biometry. The Principles and Practice of Statistics in Biological Research. New York, W.H. Freeman and Company 1995, pp 815–816.

44 Smouse PE, Long JC, Sokal RR: Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Systematic Zoology 1986;35:627–632.

45 Legendre P, Legendre L: Numerical Ecology. 2nd English ed. Amsterdam, Elsevier, 1998.

46 Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Henry M, Stevens H, Wagner H: Vegan: Community Ecology Package. R package version 1.15–4 http://CRAN.R-project.org/package=vegan (accessed February 3, 2010).

47 Cox TF, Cox MAA: Multidimensional Scaling. New York, Chapman and Hall, 1994.

48 Kruskal JB, Wish M: Multidimensional Scaling. Beverly Hills, California, Sage, 1978.

49 Sturrock K, Rocha J: A multidimensional scaling stress evaluation table. Field Methods 2000;12:49–60.

50 Peres-Neto PR, Jackson DA: How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. Oecologia 2001; 129:169–178.

51 Heine B, Nurse D: African Languages. An Introduction. Cambridge, Cambridge University Press, 2000.

52 Ruhlen M: A Guide to the World's Languages. London, Edward Arnold, 1987.

53 Sokal RR, Oden NL, Legendre P, Fortin MJ, Kim J, Thomson BA, Vaudor A, Harding RM, Barbujani G: Genetics and language in European populations. American Naturalist 1990;135:157–175.

54 Balaresque PL, Ballereau SJ, Jobling MA: Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet 2007;16:R134–R139.

55 Lopez Herraez D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M: Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. PLoS One 2009; 4:e7888.

56 Barbujani G: DNA variation and language affinities. Am J Hum Genet 1997;61:1011–1014.

57 Renfrew C: Before Babel: speculations on the origins of linguistic diversity. Cambridge Archaeological Journal 1991;1:3–23.

58 Barbujani G, Colonna V: Human genome diversity: frequently asked questions. Trends Genet 2010;26:285–295.