Across language families: Genome diversity mirrors linguistic variation within Europe

Across Language Families: Genome Diversity Mirrors Linguistic Variation Within Europe

Giuseppe Longobardi,[1,2] Silvia Ghirotto,[3] Cristina Guardiano,[4] Francesca Tassi,[3] Andrea Benazzo,[3] Andrea Ceolin,[1] and Guido Barbujani[3]*

[1]*Department of Language and Linguistic Science, University of York, York, UK*
[2]*Department of Humanities, University of Trieste, Trieste, Italy*
[3]*Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy*
[4]*Department of Communication and Economics, University of Modena-Reggio Emilia, Modena, Italy*

KEY WORDS parametric comparison method; genome-wide diversity; single-nucleotide polymorphisms; human evolutionary history

ABSTRACT *Objectives:* The notion that patterns of linguistic and biological variation may cast light on each other and on population histories dates back to Darwin's times; yet, turning this intuition into a proper research program has met with serious methodological difficulties, especially affecting language comparisons. This article takes advantage of two new tools of comparative linguistics: a refined list of Indo-European cognate words, and a novel method of language comparison estimating linguistic diversity from a universal inventory of grammatical polymorphisms, and hence enabling comparison even across different families. We corroborated the method and used it to compare patterns of linguistic and genomic variation in Europe. *Materials and Methods:* Two sets of linguistic distances, lexical and syntactic, were inferred from these data and compared with measures of geographic and genomic distance through a series of matrix correlation tests. Linguistic and genomic trees were also estimated and compared. A method (Treemix) was used to infer migration episodes after the main population splits. *Results:* We observed significant correlations between genomic and linguistic diversity, the latter inferred from data on both Indo-European and non-Indo-European languages. Contrary to previous observations, on the European scale, language proved a better predictor of genomic differences than geography. Inferred episodes of genetic admixture following the main population splits found convincing correlates also in the linguistic realm. *Discussion:* These results pave the ground for previously unfeasible cross-disciplinary analyses at the worldwide scale, encompassing populations of distant language families. Am J Phys Anthropol 157:630–640, 2015. © 2015 The Authors American Journal of Physical Anthropology Published by Wiley Periodicals, Inc.

Why are humans biologically different, and how did they come to speak different languages? Taken separately, these questions have certainly been faced for millennia now, but it was Charles Darwin (1859) who explicitly put forth the idea of a parallelism between biological evolution and language diversification; Darwin foresaw that a perfect pedigree of human populations would also represent the best possible phylogenetic tree of the world's languages. Indeed, factors isolating populations from each other (such as barriers to migration, or just distance) are expected to promote both biological and cultural divergence, while factors facilitating contacts should have the opposite effect; but gene/language parallelisms might in fact be deeper than that, in many respects (and recently some scholars went as far as claiming a role even for adaptation, not only in biological evolution, but in certain linguistic changes as well (Levinson and Gray, 2012).

Darwin's evolutionary framework was immediately accepted by linguists such as Schleicher (1863); however, it took more than a century for his parallelism intuition to be tried against actual data (Sokal, 1988; Cavalli-Sforza et al., 1988), and to become part of a broader research program (Renfrew, 1987; Cavalli Sforza et al., 1994). The idea is that linguistic diversification caused by demographic processes, mainly population dispersal, would generate parallel patterns of genetic and linguistic variation. That would often be the rule, but where a small group imposes its language upon a larger population (élite dominance: Renfrew, 1992), there might arise a local mismatch between genetic and linguistic diversity, so that already from this exception to the general rule one could detect the occurrence of an important event of language replacement.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: European Research Council; Grant number: ERC-2011-AdvG_295733 Grant (Langelin) (to G.L. and G.B.); Grant sponsor: Italian Ministry for Research and Universities (MIUR) PRIN 2010-2011 (to G.B.) and York Centre for Linguistic History and Diversity (to A.C.).

*Correspondence to: Guido Barbujani, Department of Life Sciences and Biotechnology, University of Ferrara, Via Borsari 46, I-44121 Ferrara, Italy. E-mail: g.barbujani@unife.it

Received 7 January 2015; revised 2 April 2015; accepted 14 April 2015

DOI: 10.1002/ajpa.22758
Published online 8 June 2015 in Wiley Online Library (wileyonlinelibrary.com).

The results of the line of studies above were illuminating on the one hand, but controversial on the other. The case for analogies between linguistic and genetic variation in empirical fact (Barbujani and Sokal, 1990; Barbujani and Pilastro, 1993; Cavalli-Sforza et al., 1994; Sajantila et al., 1995; Poloni et al., 1997; Belle and Barbujani, 2007) and in methods (Ringe et al., 2002; Gray and Atkinson, 2003; McMahon and McMahon, 2003; Heggarty, 2006; Gray et al., 2009; Bouckaert et al., 2012; Berwick et al., 2013), clearly emerged at a regional level; instead, at the larger geographical scale, many such conclusions were received with skepticism, especially by linguists.

One weakness of early approaches was in fact the unavailability of numerical taxonomies of languages to be matched with the biological ones: classical methods have produced chance-proof demonstrations of absolute relatedness for words/languages, but hardly provided quantitative cognacy measures even for subarticulation of acknowledged families. The second reason for skepticism was that solid linguistic relationships have so far been inferred from comparing vocabulary items (words/morphemes) and their sound structures; now, formally identifiable correspondences of such items in sound/meaning (chance-safe etymologies) are known to dissolve with time, while accidental similarities emerge, due to the arbitrariness of lexical variation combined with general constraints on possible phonological systems. Therefore, although the time depth at which these processes disrupt the potential for long-range linguistic classification is far from established (Nichols, 1996; Greenhill et al., 2010), it has been anyway impossible to convincingly infer distant relationships (across evident families) from lexical comparisons. As a consequence, large-scale gene/language comparisons were undermined by scientifically unsupported linguistic classifications and taxonomic procedures (Ringe, 1996; Bolnick et al., 2004; Greenhill, 2011; Ringe and Eska, 2013).

In this article, we try to overcome such problems through two tools recently developed for language comparison: Bouckaert et al.'s (2012) expanded list of Indo-European (henceforth IE) lexical cognates and Longobardi and Guardiano's (2009) Parametric Comparison Method (PCM). We used these new resources to interpret patterns of genome-wide variation in 15 European populations (from three different linguistic families), inferred from autosomal single nucleotide polymorphisms (SNPs) data; the final dataset included 805 individuals, and after data cleaning and integration we had >177,000 SNPs autosomal SNPs for the analysis.

MATERIALS AND METHODS

The linguistic approaches

Computational approaches to phylogenetic linguistics have led to refinements of lists of taxonomic characters available for classifying Indo-European (IE) languages. The latest breakthrough in this domain is Bouckaert et al.'s (2012) IE cognate list, IELex, which summarizes the etymological expert judgments assigned by Dyen et al. (1992), Ringe et al. (2002), along with other sources, for a 207 Swadesh-list (Swadesh, 1952) in several IE languages. This makes available a richer device for quantitative experiments on IE lexical diversification. To take into proper account language-internal polymorphism (synonymy), in Bouckaert et al. (2012) several lexical roots are often listed for the same meaning.

The second tool, the PCM, is a more radical departure from traditional procedures and databases. Languages are increasingly studied by theoretical linguists not merely as lists of words, but also as sets of recursive rules (technically, generative grammars: Chomsky, 1955) combining words into an infinite number of sentences (Chomsky, 1965). Therefore, an alternative to comparison of vocabularies is precisely exploring the phylogenetic potential of grammatical diversity (different rules of (co-)occurrence, order, and interpretation of various classes of words, morphemes, and features: Nichols, 1992; Longobardi, 2003; Guardiano and Longobardi, 2005).

Longobardi and Guardiano's (2009) central hypothesis is that syntactic change, though insightfully shown to be "catastrophic" in specific "local" instances (Lightfoot, 1991), when considered as an overall phenomenon might arise slowly enough to produce retrievable evolution (Longobardi, 2003). If so, generative grammars could encode a historical signal, useful for deeper classifications of languages and populations.

In the PCM, the core grammar of any language is represented as a string of binary symbols, each encoding the value of a syntactic parameter (Chomsky, 1981; Clark and Roberts, 1993; Baker, 2001; Roberts, 2007; Biberauer, 2008). Parameters are drawn from a supposedly universal list, defining a structured variation space within the human capacity often labeled "universal grammar" (UG) or "faculty of language." Therefore, through the PCM, in principle, all languages, no matter how lexically distant, could now be compared, bypassing many problems arising with word collation. Case studies suggested that the chance probability of parametric resemblance can be computed and controlled for (Bortolussi et al., 2011), as well as certain amounts of homoplasy (Longobardi, 2012) and admixture (Longobardi et al., 2013); finally, there is less a priori reason to expect external (e.g. cultural) factors to exert selective pressure on syntax than on lexical items (Guardiano and Longobardi, 2005; Longobardi and Guardiano, 2009; Ringe and Eska, 2013). A proof-of-concept study of gene/language congruence in a small sample of Old-World populations has already shown how correlations can be found between a preliminary set of parametric distances and genetic ones (Colonna et al., 2010).

However, the more sophisticated linguistic samples become, the more pervasive appear internal implications (nonindependence of characters, leading to redundant information). They saliently arise in grammar (Greenberg, 1963; Hawkins, 1983; Baker, 2001; Biberauer, 2008), but also affect vocabularies, once synonymy is encoded, and may have a disruptive potential for calculating taxonomic distances, if ignored. Hence, in this study, to compute plausible distances from Bouckaert et al.'s (2012) list, we propose a specific weighted measurement (see below). Since the PCM has been already originally designed for spelling out hypotheses on, and controlling for, crossparametric implications (Longobardi and Guardiano, 2009; Bortolussi et al., 2011), here, the non-independence of characters is controlled by making explicit hypotheses about implications of syntactic properties and adopting a distance calculation appropriate for them (see below).

Languages and populations

Recently, the PCM has been empirically validated on a set of 26 IE languages (Longobardi et al., 2013), syntactically defined through 56 binary parameters described in the corresponding online support material (https://benjamins.com/#catalog/journals/jhl.3.1.07lon/additional).

Fig. 1. Geographic distribution of the samples considered in this study. Indo-European-speaking populations in blue, populations speaking Finno-Ugric languages (Hungarian, Finnish) and the linguistic isolate (Basque) in red.

The method produced near-perfect taxonomies within IE, and also suggested that diachronic resetting of syntactic parameters is slower than lexical replacement. Thus, to investigate gene-language congruence in Europe, we took the intersection between the languages of this syntactic dataset and those of the much wider sample of Bouckaert et al. (2012), and selected from it a further subset of 12 varieties, for whose speakers genome-wide data are publicly available.

Then, we expanded the analysis to include three non-IE-speaking populations, for which we also found available genomic information and whose languages had been previously analyzed in terms of PCM (Longobardi and Guardiano, 2009). For such languages (Finnish, Hungarian, and Basque), we set the same 56 parameters of Longobardi et al. (2013) (Supporting Information Table 1).

Genetic analyses

Genomic data on 13 populations were found in POPRES (dbGap accession phs000145.v1.p1; Nelson et al., 2008), a public resource for genetic research including 5,886 subjects genotyped at 500,568 loci using the Affymetrix 500K SNP chip. To determine the geographic location that best represents each individual's ancestry, we used a strict criterion of sample selection excluding individuals who reported mixed grandparental ancestry. A Basque (Henn et al., 2012) and a Finnish (1000 Genomes Project Consortium, 2012) sample were then added (Fig. 1). Outliers and individuals showing high levels of genetic similarity, which may point to biological relatedness, were excluded, and all data were merged using PLINK (Purcell et al. 2007). To avoid any ambiguity in strand alignment, we removed from the merged genotype datafile the alleles carrying ambiguities in strand-flipping, namely A/T and C/G polymorphisms. The final dataset comprises 177,949 markers that passed quality filters in all datasets for 805 individuals (minor allele frequency = 0.01, genotyping rate = 98%). Genetic distances (d_{GEN}), i.e., F_{ST} values between pairs of populations (Weir and Cockerham, 1984) were calculated by the 4P software (Benazzo et al., 2015). As a preliminary test, we summarized the genetic structure of the studied populations by ADMIXTURE (Alexander et al., 2009).

Distance matrices

We started by inferring four matrices of pairwise distances between the 12 IE-speaking populations and for the whole set of 15 European populations: geographic (d_{GEO}), genomic (d_{GEN}), and two types of linguistic distances, syntactic (d_{SYN}, from Longobardi and Guardiano, 2009) and lexical (d_{LEX}, from Bouckaert et al., 2012 new word list).

Syntactic distances have been calculated according to the formula already proposed in previous works (Longobardi and Guardiano, 2009, Bortolussi et al., 2011) to account for the existence of parameter values neutralized by system-internal implications: normalized Hamming distance or Jaccard distance (Jaccard, 1901), i.e. the number of differences between two languages divided by the sum of their identities and differences (cf. also Lewandowsky and Winter, 1971). As a consequence, all the pairwise syntactic distances (d_{SYN}) end up falling between 0 and 1.

In calculating d_{SYN}, all differences and identities in each parameter value are computed as having the same taxonomic value and the two values of every parameter are considered equiprobable. We are aware that different parameters may have different phylogenetic weights, as for instance argued in Rigon (2009, 2012). However, since syntactic arguments for parameter hierarchies, diachronic stability, and markedness hypotheses seem to be still under development (Roberts, 2012), we do not have an appropriate theory of parametric weights to rely on. Hence, we decided to consider all parameters and values, as a scientific idealization, of equal import, waiting for stronger evidence on markedness, stability, and related issues: for, if correct phylogenetic results are already attained through less fine-grained characters, it would be curious if they were substantially worsened once theoretically more refined (hence less idealized, in principle more realistic) characters and analyses were employed. The opposite strategy instead would introduce a further source of arbitrariness, casting doubts on the robustness and noncircularity of the taxonomic achievements.

The second type of pairwise linguistic distances, d_{LEX}, is based on lexical comparisons. Such distances were initially estimated as the number of character differences out of the number of all lexical roots expressed at least in one of the two languages compared; again, this way all distances fall between a minimum of 0 and a maximum of 1. However, in our particular dataset, it turned out that almost all values of the resulting matrix were scattered around 0.9, hence scarcely informative and historically not plausible. This is likely to be a natural consequence of the criteria adopted to compute differences: indeed, Swadesh-lists require each meaning to be expressed by at least one lexical root in each of the languages; since polymorphism within the same language is expected to be a marked phenomenon, every lexical root displayed in a language but not in another is likely to predict a different lexical root to express the same meaning in the second language, thus doubling differences. It was necessary to take this into account and to assign differences a weight of 0.5 (rather than 1), so obtaining a more informative distance matrix (d_{LEX}), whose values

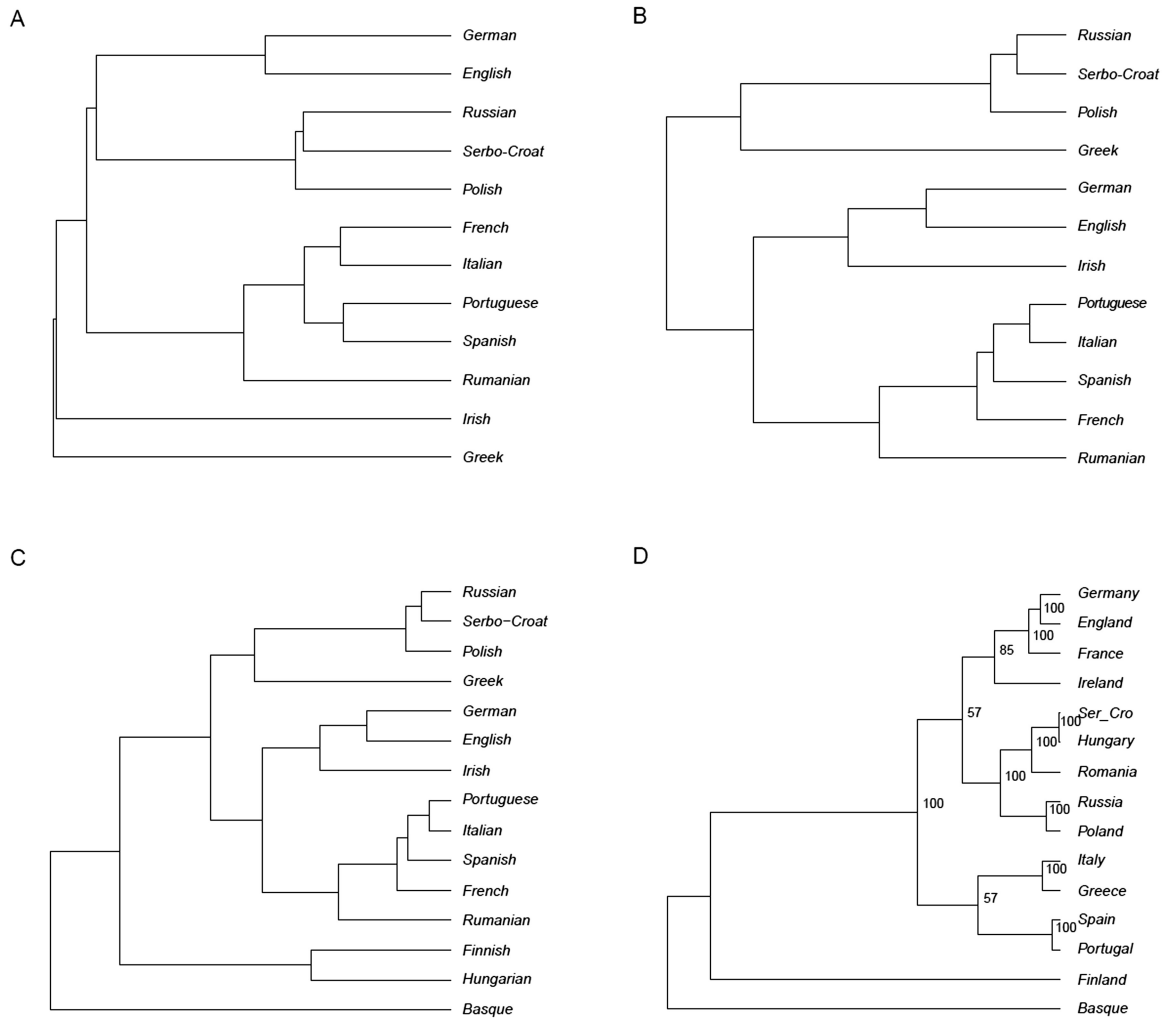

Fig. 2. UPGMA trees summarizing population relationships. Distances inferred from: (**A**) lexical and (**B**) syntactic comparisons among 12 Indo-European-speaking European populations; (**C**) syntactic comparisons among 15 European languages, and (**D**) F_{ST} distances among 15 populations sharing 177,949 SNPs. Lexical distances were estimated from lists of cognate words, amounting to over 6,000 roots (http://ielex.mpi.nl/); syntactic distances were measured over 56 parameters of nominal phrases (http://dx.doi.org/10.1075/jhl.3.1.07lon.additional). In (D), numbers indicate the support of the branching after 100 bootstrap replicates. The matrix perturbation techniques usable to test the robustness of trees (bootstrapping and jackknifing) provide stable topologies, but owing to the small number of characters involved they are only relatively reliable (cf. Longobardi et al., 2013 for more details). Therefore, bootstrapping scores have been only reported here for the genetic tree D.

pattern better on average with those previously obtained (Longobardi et al., 2013) from Dyen et al. (1992). Given that, by definition, only within the same family is it possible to compute some safe rate of common lexical etymologies, for comparison between languages from different families, which accordingly share no etymology, distance 1 was assigned by default. An approximation to the distance between Hungarian and Finnish was tentatively computed from some literature references (Laasko, 2000; Peust, 2013).

Matrix comparisons

Correlations between pairs of these distance matrices were calculated according to the Mantel (1967) procedure, using the *mantel* function of the R *Vegan* package. The significance was empirically estimated over 10,000 permutations. To exclude the potentially confounding effect of some variable, we also ran partial Mantel tests, thus calculating the correlation between two matrices while controlling for (i.e. keeping constant) a third distance matrix. To this end, we used the *mantel.partial* function of the R *Vegan* package. Finally, to compare tree topologies (Steel and Penny, 1993), we calculated the path difference distance between trees using the *treedist* function of the R *phangorn* package, and we generated the 100,000 pairs of random trees for 12 and 15 taxa with the *rtree* function of the R *ape* package.

An improved method to describe population splits and later gene flow

Population structure depends on a number of evolutionary and demographic processes which may be difficult or impossible to summarize in the form of a simple bifurcating tree. Therefore, we also represented genomic variation by a network in which populations may exchange migrants after they have split from their

common ancestors, thus violating the simplistic assumptions of most tree-building models (Pickrell and Pritchard, 2012). The first step in this exercise is the estimation of a maximum-likelihood tree. Populations poorly fitting the tree model are then identified, and migration events involving them are superimposed, so that the tree with the added migration episodes will account for a greater proportion of the overall genetic variance than the simple tree itself. This way, each population may have multiple origins, and the migrational contacts in the descendant populations are highlighted.

RESULTS

First of all, we made sure that the smaller subset of 12 IE languages displays a significant syntax-lexicon correlation, and retains as a plausible phylogenetic structure as that generated from the wider sample of 26 in Longobardi et al. (2013). Thus, for such 12 IE languages/populations, we compared d_{SYN} and d_{LEX} with one another. The two linguistic matrices appeared highly correlated ($r = 0.82$).

To better understand to what extent lexical and syntactic differences mirror each other, we represented the matrices in tree form (Fig. 2A,B), calculated the path difference distance between trees (Steel and Penny, 1993), and compared this value with those obtained in 100,000 pairs of random topologies drawn, with replacement, from the total set of the possible topologies for 12 taxa. No closer match between topologies was observed (hence $P < 10^{-5}$). Notice that the deep branches in the lexical tree in Figure 2A are short: this seems to reflect a phenomenon observed and preliminarily discussed in Longobardi and Guardiano (2009) and Longobardi et al. (2013): while syntactic distances do not show any sign of saturation at least within the domain of IE and European languages, lexical distances capture well the more recent separations within IE subfamilies, but already lose resolution when different subfamilies of IE are compared.

The syntactic and lexical matrices are highly correlated with each other, and showed very similar levels of correlation with genetic distances ($r = 0.49$ and 0.51, respectively), both with high statistical significance, which stands Bonferroni correction for multiple tests (Table 1). Syntactic distances also show a tighter association with geography than their lexical counterparts. Most importantly, the correlations of both lexicon and syntax with genetic distances are higher than between genes and geography ($r = 0.38$). I.e. once precise measurements of linguistic differences are used, language turns out a better predictor of genetic differences than geography in Europe.

Such results arose from already available IE databases. In order to strengthen them, we extended the analysis to the three non-IE languages of Europe mentioned above (i.e. Finnish, Hungarian, and Basque). To do so, we crucially relied on PCM's ability to compare languages even from different families. Recall, indeed, that calculating lexical distances from cognates for languages from different families is an essentially vacuous procedure, since by definition such languages share no common etymologies: hence the theoretically maximal distance must a priori be assigned, so that the result is largely uninformative. A way to overcome this shortcoming was the development of the PCM, precisely because it relies on polymorphic characters which are in principle universal.

The same four matrices and six correlations as above were recalculated for the whole set of 15 populations (Table 2). The correlations between genes and languages, both for syntax (0.60) and lexicon (0.54), remain much higher than between genes and geography (0.30). Actually, the latter correlation further decreases, while the one most significantly rising is that between genes and syntax (from 0.49 to 0.60), confirming that syntax, in Europe, is a better predictor of genomic variation than geography. Indeed, the correlation remains significant even after removing the effects of geography through a partial Mantel test (d_{SYN} vs. d_{GEN} $r = 0.57$), and after Bonferroni correction for multiple tests.

Instead, all the correlations with geography become lower in the 15-unit sample, probably because not all the three linguistic outliers added are also geographical outliers. Also, the correlation of genes and lexicon appears to increase, though hardly significantly, presumably beginning

TABLE 1. Mantel correlations between genetic, geographic, and two kinds of linguistic distances in Indo-European-speaking populations of Europe

Distance matrices	r	P
d_{LEX} d_{GEO} Linguistic (lexical)-geographic	0.206	0.077
d_{LEX} d_{GEN} Linguistic (lexical)-genetic	0.514	0.0001
d_{SYN} d_{GEO} Linguistic (syntactic)-geographic	0.385	0.008
d_{SYN} d_{GEN} Linguistic (syntactic)-genetic	0.491	0.0004
d_{LEX} d_{SYN} Linguistic (lexical)-linguistic (syntactic)	0.822	0.0001
d_{GEN} d_{GEO} Genetic-geographic	0.390	0.011

After Bonferroni correction for multiple tests, these results are significant at the $P = 0.0006$ level.

TABLE 2. Mantel correlations and partial Mantel correlations between matrices of syntactic, lexical, geographic, and genetic distance in 15 populations in Europe

Distance matrices	r	P
d_{GEN} d_{GEO} Genetic-geographic	0.299	0.030
$d_{SYN}d_{LEX}$ Syntactic-lexical	0.850	0.001
d_{SYN} d_{GEO} Syntactic-geographic	0.240	0.039
d_{LEX} d_{GEO} Lexical-geographic	0.084	0.264
d_{SYN} d_{GEN} Syntactic-genetic	0.599	0.001
d_{LEX} d_{GEN} Lexical-genetic	0.537	0.001
d_{GEN} d_{GEO} (d_{SYN}) Genetic-geographic (syntax held constant)	0.200	0.114
d_{GEN} d_{GEO} (d_{LEX}) Genetic-geographic (lexicon held constant)	0.302	0.035
d_{SYN} d_{GEO} (d_{GEN}) Syntactic-Geographic (genetics held constant)	0.079	0.264
d_{LEX} d_{GEO} (d_{GEN}) Lexical-geographic (genetics held constant)	−0.095	0.736
d_{SYN} d_{GEN} (d_{GEO}) Syntactic-genetic (geography held constant)	0.570	0.002
d_{LEX} d_{GEN} (d_{GEO}) Lexical-genetic (geography held constant)	0.538	0.001

After Bonferroni correction for multiple tests, these results are significant at the $P = 0.012$ level.

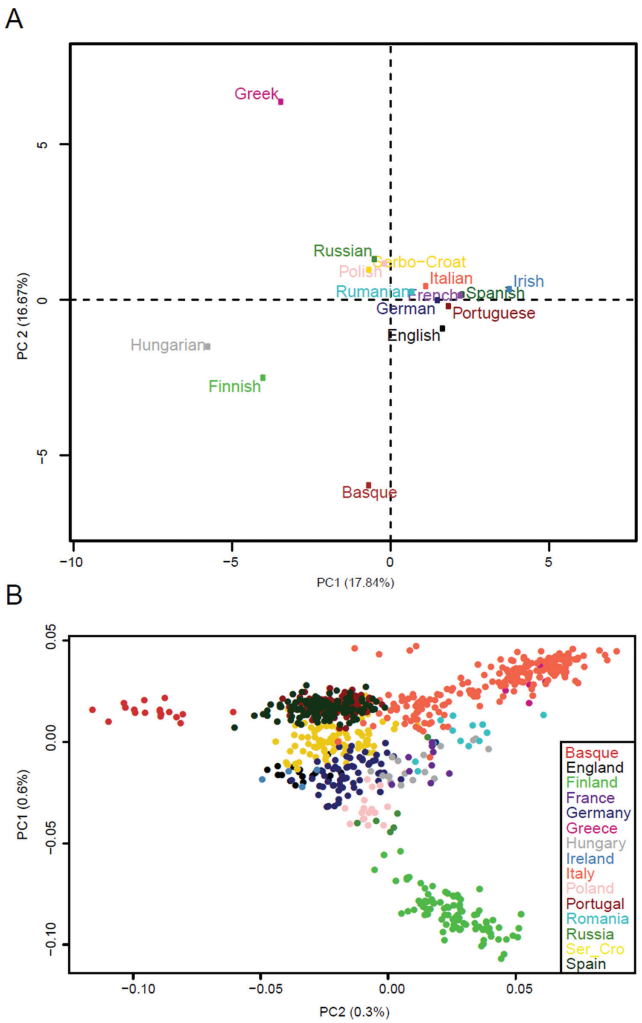

Fig. 3. Projection on two dimensions of the main components (PCA) of linguistic (**A**) and individual genomic (**B**) variation. The linguistic PCA was performed using the *R FactoMineR* program, with neutralized parameter values coded as "NA," whereas the genomic PCA was calculated with the *R SNPRelate* package (Lê et al., 2008). Note that the linguistic scatter diagram accounts for a fraction of the total variance that is >25-fold as large as that accounted for by the genomic scatter diagram.

to suffer from the mentioned saturation of cross-family lexical distances. Thus, the more languages from different families will be added for comparison, the more we expect reliance on the PCM to become crucial. To better understand gene-language congruence at the cross-family European level, we focused in more detail on syntactic distances.

We drew a UPGMA tree and carried out a Principal Component Analysis (PCA) from d_{SYN}. The tree (Fig. 2C) singles out IE and meets all further basic expectations: the deepest nodes first separate Basque, and then the pair of Finno-Ugric languages, from the cluster comprising all the IE varieties. Within this cluster, Romance, Germanic and Slavic form three subclusters; then Greek and Irish, as the only representatives of their subfamilies in this study, occur on separate branches, although close to their geographic neighbors.

In the PCA, the combination of the first two axes, jointly accounting for 34.5% of the variance, separates IE languages (but Greek) from the others: Greek, an IE language without very close relatives, falls anyway opposite to Finnish, Hungarian, and Basque (Fig. 3A). This pattern is largely expected, and the position of Greek as the outlier of IE in our sample agrees with previous computational experiments on lexical datasets (Bouckaert et al., 2012, Gray and Atkinson, 2003).

In short, through syntax, precise comparison and measuring is finally possible even across established linguistic families: the main families and subfamilies of Europe were discriminated by means of just 56 abstract characters, suggested by formal grammatical theory, through standard methods of evolutionary biology, without resorting to methodologically disputable cross-family lexical comparisons.

Then, to synthetically visualize genomic diversity in a parallel way, we drew the corresponding UPGMA tree and carried out a PCA analysis from d_{GEN}. The tree (Fig. 2D) shows that two out of the three linguistic outliers, Finns and Basques, are clearly differentiated also genomically, and connected to the other populations by long independent branches. The rest of the tree mainly reflects geographical distances, and contains all IE-speaking populations, as well as Hungarians, who appear genetically related with their geographical neighbors, Serbs and Rumanians. Again, the path difference distance (Steel and Penny, 1993) was calculated between the syntactic and the genetic tree, and the probability to obtain a closer match between random trees with 15 populations turned out $P < 0.004$. This implies a tight relationship between the tree topologies inferred from syntax and genes, one highly unlikely to have arisen just by chance. The only salient divergence is represented by the position of Hungarians, mostly falling within a large group of Central Europeans. Note that a sharp genomic differentiation of Basques from most other Europeans has not been confirmed in all previous genomic studies (see Laayouni et al., 2010).

Then, we carried out a parallel PCA of the >177,000 SNPs in 805 individuals from the 15 populations representative of the previously considered languages. As expected, given the well-known low levels of cross-population diversity in humans in general (Barbujani and Colonna, 2010) and in Europe in particular (Novembre et al., 2008), the proportion of the overall variance accounted for by the two main axes is much lower (less than 1%) than in the analysis of linguistic data (Fig. 3B), as previously observed. However, the two PCAs are qualitatively similar in several respects, with a main central cluster containing all *IE* speakers along with Hungarians, and with Finns and Basques appearing as outliers though both relatively close to their nearest geographical neighbors (Poles and Spaniards, respectively).

An unsupervised analysis of population structure through ADMIXTURE basically led to the same conclusions as the PCA, and confirmed the peculiar genetic position of Hungarians. Postulating three ancestral genomic clusters for Europe, i.e. as many as the language families in the database ($k = 3$ plot, Fig. 4), such clusters largely correspond to: (i) Basques, (ii) Finns, and (iii) all other Europeans including Hungarians; the Basque sample shows connections with the Spanish and French ones (blue component), and Finns seem to share some ancestry with Northern Europeans (Germans and Poles, orange component). Other analyses, assuming

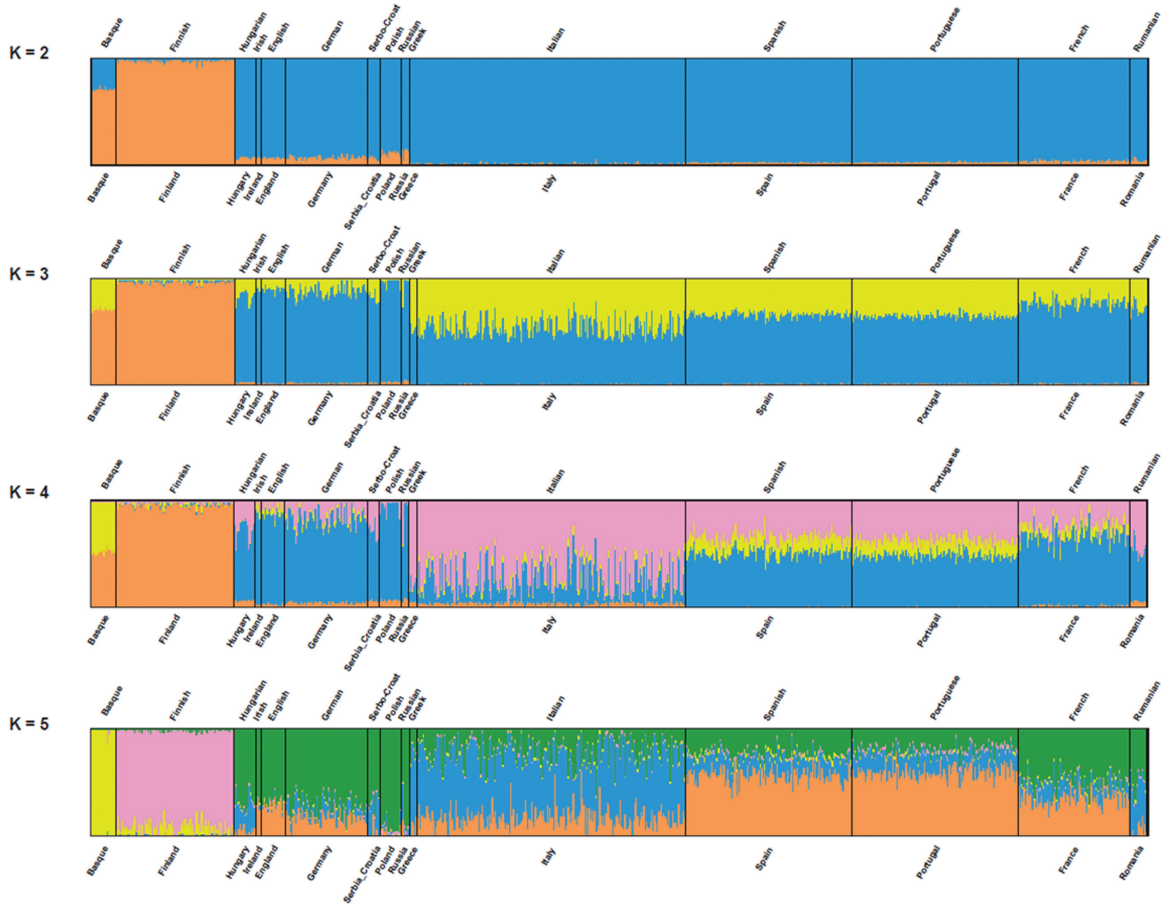

Fig. 4. Unsupervised ancestry-inference analysis based on the software ADMIXTURE. Each individual genotype is represented by a column in the area representing the appropriate population, and colors correspond to the fraction of the genotype that can be attributed to each of the K groups ($2 \leq K \leq 5$) assumed to have contributed to the populations' ancestry.

different numbers of clusters in the genomic data, are also given for completeness of information.

We further investigated the evolutionary relationships between populations by a method designed to identify gene flow episodes after the main population splits (Fig. 5). Indeed, a tree-like representation of genomic (or linguistic, for that matter) relationships disregards the possibility of exchanges occurring after populations separated from their common ancestor. The contribution of migrants to Rumania from Russia (0.43) as well as from Greece is in agreement with the populations' geographical proximity, and their traditionally well-assessed horizontal linguistic connection: the received concept of a Balkan common linguistic area, or *Sprachbund*, has found at least some suggestive correspondence even in the parametric linguistic analysis, for in three parameters Rumanian, the outlier of the Romance branch of the language tree (Fig. 3A), shares a state with Greek in contrast to the rest of Romance, in one also with Bulgarian (Longobardi et al., 2013). The Southern European origin of a fraction of the Hungarians (0.31), instead, is not apparently matched either in the linguistic PCA (Fig. 4A) or tree (Fig. 3A), only finding a loose potential correspondence in one of the 56 syntactic characters, Parameter 7 (DGP), whose Hungarian state might in theory have been borrowed from either German or Rumanian. Relatively recent gene flows, occurring after the main population splits, seem therefore to nicely match at least a fraction of the linguistic variation not immediately representable by classifying languages into families. It is an intriguing conjecture that biological relationships unpredictable by vertical linguistic history might reflect secondary gene flow independently detected by TreeMix.

The peculiar gene-language mismatch of Hungarians was already noticed by Cavalli Sforza et al. (1994), though without the possibility of quantifying cross-family linguistic distances. Now this has become possible thanks to the PCM, and indeed, the genes-syntax correlation recalculated after removing Hungary further rises very significantly ($r = 0.74$; Table 3), while the genes-geography one remains low ($r = 0.28$), confirming Hungary as an exception, in this respect. The skew becomes even sharper in partial Mantel tests (respectively $r = 0.72$ for gene/syntax, with geography held constant, and $r = 0.09$ for gene/geography, with syntax held constant), the sharpest demonstration to date of a language/biology correlation for the core of Europe.

DISCUSSION

Reliable evidence for parallelism of genetic and linguistic change had previously been provided, although

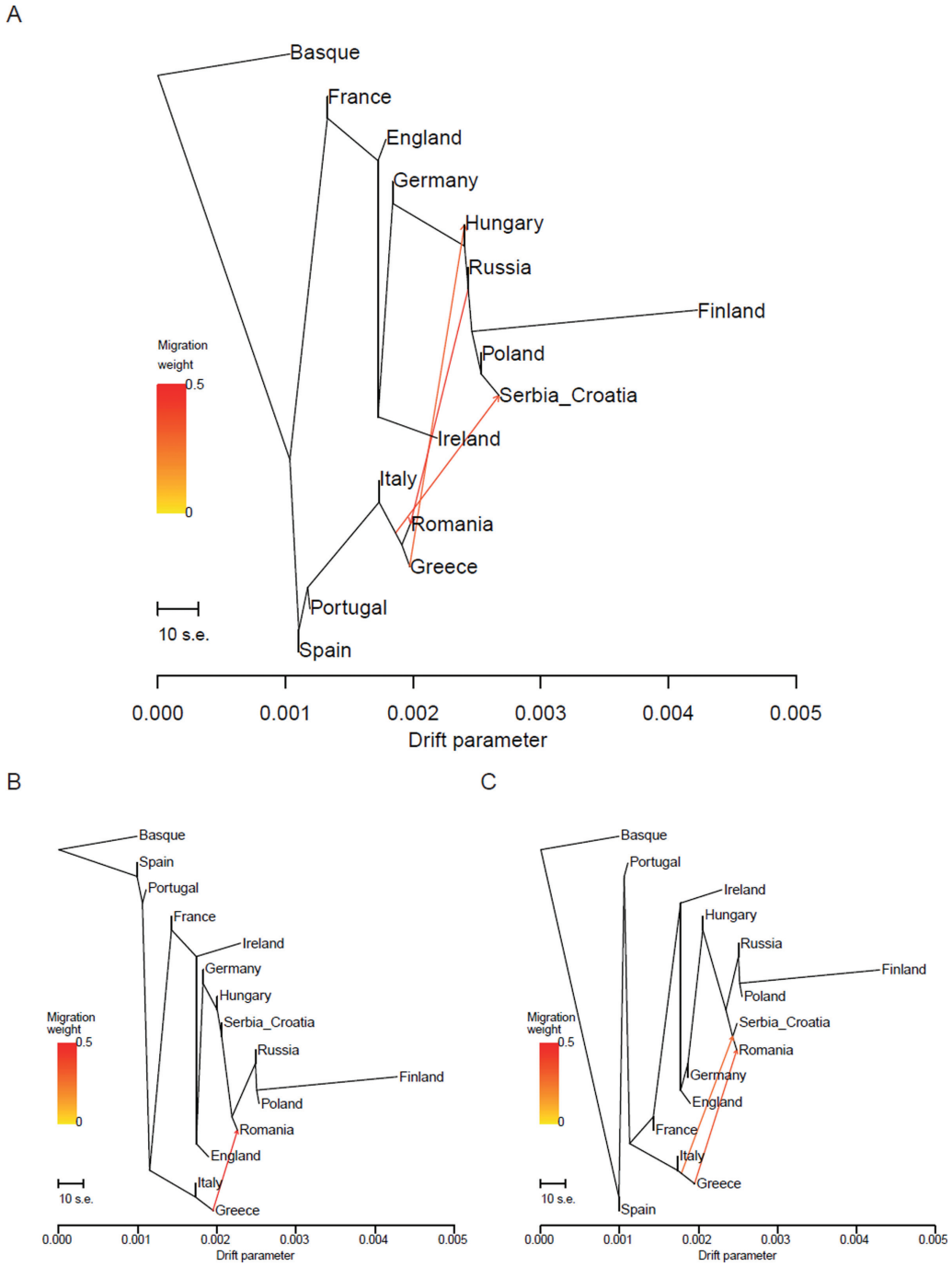

Fig. 5. Maximum-likelihood population trees. The algorithm chosen, TreeMix (*28*), estimates phylogenetic relationships with (**A**) three, (**B**) one, and (**C**) two superimposed migration events after the main population splits.

only on a regional scale (Sokal, 1988) and without formal quantification of language distances. Here, first through a quantitative approach to cognate words (Bouckaert et al., 2012), and then through a syntactic method (the PCM, Longobardi and Guardiano, 2009) designed for comparing languages across separate families, we overcome both limits of previous studies. In particular, through syntax, precise comparison and measuring is finally possible even across established linguistic families: the main families/subfamilies of Europe are discriminated by means of just 56 abstract characters suggested by formal grammatical theory, using standard methods of evolutionary biology and without resorting to unsafe long-range etymologies.

This allowed a series of cross-family correlation tests which reach precise conclusions on a broader continental scale: populations speaking similar languages in Europe tend to resemble each other at the genomic level, thus suggesting that cultural change and biological divergence have proceeded in parallel in Europe, at least

TABLE 3. Mantel correlations and partial Mantel correlations between matrices of syntactic, lexical, geographic, and genetic distance for 14 populations in Europe (after removing Hungary)

Distance matrices	*r*	*P*
d_{GEN} d_{GEO} Genetic-geographic	0.275	0.048
$d_{SYN}d_{LEX}$ Syntactic-lexical	0.850	0.001
d_{SYN} d_{GEO} Syntactic-geographic	0.291	0.026
d_{LEX} d_{GEO} Lexical-geographic	0.152	0.144
d_{SYN} d_{GEN} Syntactic-genetic	0.740	0.001
d_{LEX} d_{GEN} Lexical-genetic	0.687	0.001
d_{GEN} d_{GEO} (d_{SYN}) Genetic-geographic (syntax held constant)	0.093	0.254
d_{GEN} d_{GEO} (d_{LEX}) Genetic-geographic (lexicon held constant)	0.238	0.083
d_{SYN} d_{GEO} (d_{GEN}) Syntactic-geographic (genetics held constant)	0.135	0.178
d_{LEX} d_{GEO} (d_{GEN}) Lexical-geographic (genetics held constant)	−0.053	0.615
d_{SYN} d_{GEN} (d_{GEO}) Syntactic-genetic (geography held constant)	0.717	0.001
d_{LEX} d_{GEN} (d_{GEO}) Lexical-genetic (geography held constant)	0.679	0.001

After Bonferroni correction for multiple tests, these results are significant at the $P = 0.012$ level.

as a rule (for exceptions, also see Bolnick et al., 2004). The partial correlation tests show that populations speaking similar languages also tend to be genetically closer than expected on the sheer basis of their geographic location, so that in Europe language, i.e. basic vocabulary and now, at an even wider scale, syntax, appear to offer a better prediction of genomic distances than geography.

These correlations with independent historical variables provide a new type of evidence for the PCM and in turn for the general biolinguistic approach it is inspired by (Lightfoot, 1999; Di Sciullo and Boeckx, 2011; Berwick et al., 2013), and strengthens the controversial hypothesis that parameters do encode a phylogenetic signal (Lightfoot, 2006). Working out parameter theory against such historical evidence is also a new way of evaluating it with respect to its major critical points, e.g., the learnability issues (Boeckx and Leivada, 2013). Note that the trees in this study were inferred from the distance matrices, since character-based programs seem less suitable for heavily implicational systems (Longobardi et al., 2013). It has also been argued (Heggarty, 2006) that, despite the admitted loss of information, distances may even remedy some shortcomings of parsimony-based character programs in dealing with occasional homoplasies or backmutations. Finally, and most crucially, for the purpose of calculating Mantel correlations between qualitatively and quantitatively very different entities (56 parameters, 178.000 SNPs), distances seem a necessary mediation/conversion.

We could thus move on to a more detailed analysis of population diversity in Europe and of the possible exceptions to the conclusions above. When population relationships were summarized by trees, the main elements of disagreement were represented by the positions of Hungarians and Rumanians, which cluster genetically with speakers of Serbo-Croatian despite being highly differentiated syntactically. These populations all dwelling in Central Europe, it is reasonable to suspect an effect of geographical proximity, enhancing gene flow between neighboring countries.

Using a method that highlights the most significant episodes of genetic exchange after population splits, a likely situation among humans (Barbujani and Colonna, 2010), especially in Europe, we could precisely find evidence of the possibly relevant biological contacts among speakers of IE-subfamilies (from Slavic-speaking areas into Rumania and from Southern Europe into the Balkans) and between Ugric and IE speakers (from the Balkans into Hungary). Pending further investigation, it appears that where biological relationships are not those expected from vertical linguistic history, they are plausibly accounted for by relatively recent gene flow processes independently detected by Treemix.

In particular, concerning the real exception to our congruence pattern, notice that the presence in modern Hungarians of DNA markers currently common in Northern and Central Asia has been interpreted as a consequence of westward gene flow in Medieval times (Csányi et al., 2008; Bíró et al., 2009; Hellenthal et al., 2014); this is obviously connected with historical migrations in the 9th century and with the fact that the current language is closely related to the Ugric-speaking communities along the Ob river. However, the current low frequency of those markers is not what one would expect to observe, had a substantial demographic replacement occurred (Nadasi et al., 2007; Hellenthal et al., 2014). Careful analyses of 10th century ancient DNA in Hungary showed a predominance of European mitochondrial haplotypes in burials attributed to the lower classes, and a high incidence of Asian haplotypes in high-status individuals of that period (Tömöry et al., 2007), which points to the Asian immigrants as representing a social élite, rather than the bulk of the population. The exception to the results of the present study is thus nicely justified in this scenario, suggesting that when a Finno-Ugric language was introduced in Hungary, the genetic buildup of the population changed only in part, thus retaining similarities with its geographic neighbors, an example of the process called élite dominance by Renfrew (1992). On the contrary, the same case cannot be easily made for Basques (Alonso et al., 2005; Rodríguez-Ezpeleta et al., 2010; Young et al., 2011; Martinez-Cruz et al., 2012) or Finns, for whom, to the best of our knowledge, no available evidence suggests a similar model of partial demographic replacement associated with language replacement (Nelis et al., 2009). Thus, the comparative linguistic/genomic analysis, attempted in the present study, seems able to single out and precisely assess these differences in the population histories of the three non-IE members of our sample.

Our results confirm the fruitfulness of importing numerical and biostatistical methods into language phylogenetics (McMahon and McMahon, 2005), but even more of resorting to radically new (Heggarty et al., 2005) and deeper (Longobardi, 2012) levels of taxonomic characters for a thorough reconstruction of both demographic and linguistic history.

In particular, we see good chances to obtain trustworthy taxonomic insights when the PCM is applied to longer-range computations that could not be safely attempted through traditional lexical methods, and we expect to find interesting and illuminating correlations between genetic and linguistic diversity across other continents, contributing to the "New Synthesis" research line (Renfrew, 1987). Sokal (1988) and Cavalli-Sforza et al. (1988) could venture into addressing Darwin's gene-language congruence issue thanks to the theoretical progress of 20th century genetics; along with the availability of broad genomic datasets, the corresponding progress of formal grammatical theory over the past 50 years may now enable us to accurately test the hypothesis on ever larger and more solid grounds.

ACKNOWLEDGMENTS

The authors are indebted to R. Gray and M. Dunn for kindly directing them to the expanded IE database used to infer lexical distances, to all the participants in the international workshop *Advances in Phylogenetic Linguistics* (Ragusa Ibla, July 13–17, 2013), and to two anonymous reviewers for their useful comments.

LITERATURE CITED

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664.

Alonso S, Flores C, Cabrera V, Alonso A, Martín P, Albarrán C, Izagirre N, de la Rúa C, García O. 2005. The place of the basques in the European Y-chromosome diversity landscape. Eur J Hum Genet 13:1293-1302.

Baker M. 2001. The atoms of language. New York: Basic Books.

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

Barbujani G, Pilastro A. 1993. Genetic evidence on origin and dispersal of human populations speaking languages of the nostratic macrofamily. Proc Natl Acad Sci USA 90:4670–4673.

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

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

Benazzo A, Panziera A, Bertorelle G. 2015. 4P: fast computing of basic population genetics statistics from large DNA polymorphism panels. Ecol Evol 5:172–175.

Berwick RC, Friederici AD, Chomsky N, Bolhuis JJ. 2013. Evolution, brain, and the nature of language. Trends Cogn Sci 17:89–98.

Biberauer T, editor. 2008. The limits of syntactic variation. Amsterdam: Benjamins.

Bíró AZ, Zalán A, Völgyi A, Pamjav H. 2009. A Y-chromosomal comparison of the Madjars (Kazakhstan) and the Magyars (Hungary). Am J Phys Anthropol 139:305–310.

Boeckx C, Leivada E. 2013. Entangled parameter hierarchies: problems for an overspecified universal grammar. PLoS One 8:e72357-

Bolnick DA, Shook BA, Campbell L, Goddard I. 2004. Problematic use of Greenberg's linguistic classification of the Americas in studies of native American genetic variation. Am J Hum Genet 75:519–522.

Bortolussi L, Longobardi G, Guardiano C, Sgarro A. 2011. How many possible languages are there? In: Bel-Enguix G, Jiménez-López MD, editors. Biology, computation and linguistics. Amsterdam: IOS Press, p 168–179.

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, Drummond AJ, Gray RD, Suchard MA, Atkinson QD. 2012. Mapping the origins and expansion of the Indo-european language family. Science 337:957–960.

Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The history and geography of human genes. Princeton: Princeton University Press.

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

Chomsky N. 1955. The logical structure of linguistic theory. MS Thesis (published in 1975). New York: Plenum.

Chomsky N. 1965. Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Chomsky N. 1981. Lectures on government and binding. Dodrecht: Foris.

Clark R, Roberts I. 1993. A computational model of language learnability and language change. Linguistic Inquiry 24:299–345.

Colonna V, Boattini A, Guardiano C, Dall'ara I, Pettener D, Longobardi G, Barbujani G. 2010. Long-range comparison between genes and languages based on syntactic distances. Hum Hered 70:245–254.

Csányi B, Bogácsi-Szabó E, Tömöry G, Czibula A, Priskin K, Csősz A, Mende B, Langó P, Csete K, Zsolnai A, Conant EK, Downes CS, Raskó I. 2008. Y-chromosome analysis of ancient Hungarian and two modern Hungarian-speaking populations from the Carpathian basin. Ann Hum Genet 72:519–534.

Darwin C. 1859. On the origin of species. London: John Murray.

Di Sciullo AM, Boeckx C., editors. 2011. The biolinguistic enterprise. New perspectives on the evolution and nature of the human language faculty. Oxford: Oxford University Press.

Dyen I, Kruskal J, Black PJ. 1992. An Indoeuropean classification: a lexicostatistical experiment. Trans Philos Soc 82:1–132.

Gray RD, Atkinson QD. 2003. Language-tree divergence times support the anatolian theory of Indo-European origin. Nature 426:435–439.

Gray RD, Drummond AJ, Greenhill SJ. 2009. Language phylogenies reveal expansion pulses and pauses in pacific settlement. Science 323:479–483.

Greenberg J. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In: Greenberg J, editor. Universals of language. Cambridge, MA: MIT Press. p 73–113.

Greenhill SJ. 2011. Levenshtein distances fail to identify language relationships accurately. Comput Linguist 37:689–698.

Greenhill SJ, Atkinson QD, Meade A, Gray RD. 2010. The shape and tempo of language evolution. Proc Biol Sci 277: 2443–2450.

Guardiano C, Longobardi G. 2005. Parametric comparison and language taxonomy. In: Battlori M, Picallo C, Roca F, editors. Grammaticalization and parametric variation. Oxford: Oxford University Press. p 149–174.

Hawkins J. 1983. Word order universals. New York: Academic Press.

Heggarty P. 2006. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data – and to dating language? In: Clackson J, Forster P, Renfrew C, editors. Phylogenetic methods and the prehistory of languages. Cambridge: McDonald Institute for Archaeological Research. p 183–194.

Heggarty P, McMahon A, McMahon R. 2005. From phonetic similarity to dialect classification: a principled approach. In: Delbecque N, Geeraerts D, van der Auwera J, editors Perspectives on variation: sociolinguistic, historical, comparative. Amsterdam: Mouton de Gruyter. p 43–91. Available at: http://dx.doi.org/10.1515/9783110909579.43.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343:747–751.

Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A,

Bertranpetit J, Bustamante CD, Comas D. 2012. Genomic ancestry of north Africans supports back-to-Africa migrations. PLoS Genet 8:e1002397.
Jaccard P. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull De La Soc Vaudoise Des Sci Nat 37:547–579.
Laasko J. 2000. "Related words" in Finnish and Hungarian. Available at: http://www.helsinki.fi/~jolaakso/f-h-ety.html.
Laayouni H, Calafell F, Bertranpetit J. 2010. A genome-wide survey does not show the genetic distinctiveness of Basques. Hum Genet 127:455–458.
Lê S, Josse J, Husson F. 2008. FactoMineR: an R package for multivariate analysis. J Stat Software 25:1–18.
Levinson SC, Gray RD. 2012. Tools from evolutionary biology shed new light on the diversification of languages. Trends Cogn Sci 16:167–173.
Lewandowsky M, Winter D. 1971. Distance between sets. Nature 234:34–35.
Lightfoot D. 1991. How to set parameters. Cambridge, MA: MIT Press.
Lighfoot D. 1999. The development of language: acquisition, change and evolution. Cambridge, MA: MIT Press.
Lightfoot D. 2006: How new languages emerge. Cambridge: Cambridge University Press.
Longobardi G. 2003. Methods in parametric linguistics and cognitive history. Linguist Variat Yearbk 3:101–138.
Longobardi G. 2012. Convergence in parametric phylogenies: homoplasy or principled explanation? In: Galves C, Cyrino S, Lopes R, Sandalo F, Avelar J, editors. Parameter theory and linguistic change. Oxford: Oxford University Press. p 304–319.
Longobardi G, Guardiano C. 2009. Evidence for syntax as a signal of historical relatedness. Lingua 119:1679–1706.
Longobardi G, Guardiano C, Silvestri G, Boattini A, Ceolin A. 2013. Toward a syntactic phylogeny of modern Indo-European languages. J Histor Linguist 3:122–152.
Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220.
Martínez-Cruz B, Harmant C, Platt DE, Haak W, Manry J, Ramos-Luis E, Soria-Hernanz DF, Bauduer F, Salaberria J, Oyharçabal B, Quintana-Murci L, Comas D. Genographic Consortium 2012. Evidence of pre-roman tribal genetic structure in basques from uniparentally inherited markers. Mol Biol Evol 29:2211–2222.
McMahon A, McMahon R. 2003. Finding families: quantitative methods in language classifying. Trans Philol Soc 101:7–55.
McMahon A, McMahon R. 2005. Language classification by numbers. Oxford: Oxford University Press.
Nadasi E, Gyurus P, Czakó M, Bene J, Kosztolányi S, Fazekas S, Dömösi P, Melegh B. 2007. Comparison of mtdna haplogroups in Hungarians with four other European populations: a small incidence of descents with Asian origin. Acta Biol Hungarica 58:245–256.
Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskácková T, Balascák I, Peltonen L, Jakkula E, Rehnström K, Lathrop M, Heath S, Galan P, Schreiber S, Meitinger T, Pfeufer A, Wichmann HE, Melegh B, Polgár N, Toniolo D, Gasparini P, D'Adamo P, Klovins J, Nikitina-Zake L, Kucinskas V, Kasnauskiene J, Lubinski J, Debniak T, Limborska S, Khrunin A, Estivill X, Rabionet R, Marsal S, Julià A, Antonarakis SE, Deutsch S, Borel C, Attar H, Gagnebin M, Macek M, Krawczak M, Remm M, Metspalu A. 2009. Genetic structure of Europeans: a view from the north–east. PLoS One 4:e5472.
Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH. 2008. The population reference sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83:347–358.
Nichols JA 1992. Linguistic diversity in space and time. Chicago: The University of Chicago Press.
Nichols JA. 1996. The comparative method reviewed: regularity and irregularity in language change. In: Durie M, Ross M, editors. The comparative method as heuristic. New York: Oxford University Press. p 39–71.
Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD. 2008. Genes mirror geography within Europe. Nature 456:98–101.
Peust C. 2013. Towards establishing a new basic vocabulary list (Swadesh list). Available at: http://www.peust.de/peustBasicVocabularyList.pdf, accessed April 27, 2015.
Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8:e1002967.
Poloni ES, Semino O, Passarino G, Santachiara-Benerecetti AS, Dupanloup I, Langaney A, Excoffier L. 1997. Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. Am J Hum Genet 61: 1015–1035.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
Renfrew C. 1987. Archaeology and language: the puzzle of Indo-European origins. London: Jonathan Cape.
Renfrew C. 1992. Archaeology, genetics and linguistic diversity. Man 27:445–478.
Rigon G. 2009. A quantitative approach to the study of syntactic evolution. Doctoral Dissertation. Pisa: Università di Pisa.
Rigon G. 2012. An evolutionary perspective on diachronic syntax. Studi E Saggi Linguistici 50:31–95.
Ringe D. 1996. The mathematics of Amerind. Diachronica 13: 135–154.
Ringe D, Eska J. 2013. Historical linguistics. Cambridge: Cambridge University Press.
Ringe D, Warnow T, Taylor A. 2002. Indo-European and computational cladistics. Trans Philol Soc 100:59–129.
Roberts I. 2007. Diachronic syntax. Oxford: Oxford University Press.
Rodríguez-Ezpeleta N, Alvarez-Busto J, Imaz L, Regueiro M, Azcárate MN, Bilbao R, Iriondo M, Gil A, Estonba A, Aransay AM. 2010. High-density SNP genotyping detects homogeneity of Spanish and French basques, and confirms their genomic distinctiveness from other European populations. Hum Genet 128:113–117.
Roberts I. 2012 On the nature of syntactic parameters: a programme for research. In: Galves C, Cyrino S, Lopez R, Avelar J, editors. Parameter theory and linguistic change. Oxford: Oxford University Press, 319–334.
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, Pääbo S. 1995. Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 5:42–52.
Schleicher A. 1863. Die Darwinsche Theorie und die Sprachwissenschaft. Weimar: Böhlau.
Sokal RR. 1988. Genetic, geographic, and linguistic distances in Europe. Proc Natl Acad Sci USA 85:1722–1726.
Steel MA, Penny P. 1993. Distribution of tree comparison metrics- some new results. Syst Biol 42:126–141.
Swadesh M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. Proc Philos Soc 96:453–63.
Tömöry G, Csányi B, Bogácsi-Szabó E, Kalmár T, Czibula A, Csosz A, Priskin K, Mende B, Langó P, Downes CS, Raskó I. 2007. Comparison of maternal lineage and biogeographic analyses of ancient and modern Hungarian populations. Am J Phys Anthropol 134:354–368.
Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.
Young KL, Sun G, Deka R, Crawford MH. 2011. Paternal genetic history of the basque population of Spain. Hum Biol 83:455–475.