Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Genetics & Development

Coevolution of languages and genes

Brigitte Pakendorf

The evolution of languages shares certain characteristics with that of genes, such as the predominantly vertical line of transmission and the retention of traces of past events such as contact. Thus, studies of language phylogenies and their correlations with genetic phylogenies can enrich our understanding of human prehistory, while insights gained from genetic studies of past population contact can help shed light on the processes underlying language contact and change. As demonstrated by recent research, these evolutionary processes are more complex than simple models of gene-language coevolution predict, with linguistic boundaries only occasionally functioning as barriers to gene flow. More frequently, admixture takes place irrespective of linguistic differences, but with a detectable impact of contact-induced changes in the languages concerned.

Addresses
Laboratoire Dynamique du Langage, UMR5596, CNRS & Université Lyon Lumière 2, Lyon, France

Corresponding author: Pakendorf, Brigitte (Brigitte.Pakendorf@cnrs.fr)

Current Opinion in Genetics & Development 2014, **29**:39–44

This review comes from a themed issue on **Genetics of human origin**

Edited by **Aida M Andrés** and **Katja Nowick**

http://dx.doi.org/10.1016/j.gde.2014.07.006

Introduction

Ever since Darwin [1] it has been assumed that the genetic and linguistic evolution of humans should be largely correlated [2–4]. The reasons for this lie in the perceived parallelism of genes and languages: genes are obligatorily passed on from parents to their offspring, and in general the first language children learn to speak is that of their parents, so that languages tend to be inherited in a vertical line as well. Similarly, genepools diverge when populations become substructured and increasingly isolated from each other; such reproductive isolation would also lead to communicative isolation and thus linguistic substructuring and the development of new languages out of an erstwhile common ancestor. Furthermore, both genes and languages can retain traces of past demographic events such as contact, detectable via genetic admixture on the one hand and linguistic changes (loanwords or structural borrowings) on the other. Thus, the correlations (as well as the lack thereof) between genetic and linguistic relationships can help shed light on the (pre)history of human populations and enrich our understanding of the processes that shape both genetic and linguistic diversity [5]. This relatively young but burgeoning field of interdisciplinary research can be approached from three different angles of investigation: firstly, the coevolution of genes and languages; secondly, prehistoric population contact and its effect on language evolution and change; and thirdly, the demographic history of language families to shed light on the prehistory of the peoples speaking these languages. I here review each of these approaches in turn for readers with no background in linguistics, covering to a large extent the last five years. Since the review by Pagel [6] focuses largely on mechanisms of language evolution and thus does not cover all the aspects touched upon here, some references to the older background literature are also included. However, I do not cover the question of how human language may have evolved in the first place; for some recent discussion see [7,8].

Coevolution of genes and languages

The investigation of language-gene coevolution was first undertaken in the late 1980s and early 1990s, when sufficient allele frequency data for a large number of human populations had been collected to make such research feasible [9]. Major questions of interest concern the extent to which linguistic differences present barriers to gene flow and thus shape genetic diversity [4], as suggested by the results of early studies [2,3,10,11], as well as whether the coevolution of genes and languages follows a branching model marked by successive splits and isolation, or rather an isolation-by-distance model with decreasing genetic and linguistic exchange over increasing geographical distances [12•]. While some studies have found that genetic structure indeed correlates with linguistic affiliation [4,13], implying that language can represent a barrier to gene flow, others have found that at regional levels gene flow and language contact have erased such patterns of phylogenetic splits [12•,14••].

In order to investigate the coevolution of genes and languages, correlations between matrices of linguistic and genetic distances are most commonly investigated. This poses a methodological problem: while the genetic distances are based on actual empirical data that are tailoured to the markers used, until recently the linguistic distances were arbitrary values assigned on the basis of a (controversial) phylogeny of languages compiled by Ruhlen [15]. In such approaches, languages belonging to different linguistic phyla are assigned the largest value and languages at lower levels in the phylogeny receive smaller values, for example, [4,16]; the validity of results obtained by such comparisons is of course rather questionable. More recent studies have circumvented the methodological problem of defining appropriate linguistic distances by calculating these with empirical data — frequently the number of retained cognate items (i.e. words going back to a common ancestor) in word lists [13,14••,17•].

However, calculating linguistic distances from lexical cognates is restricted to relatively closely related languages, as sound changes and replacement of words reduce the number of detectable cognates with time; it is commonly assumed that the time-depth for the establishment of genealogical relationships based on lexical cognates is 7000–10 000 years [18,19]. Since genetic data are not subject to such temporal constraints, it is necessary to find linguistic measures that are amenable to comparison even across very distinct language families if one wants to investigate gene-language coevolution at a global scale. Structural data, that is abstract grammatical features such as the order of subject, verb, and object or the presence/absence of definite or indefinite articles, have been suggested as potentially more suitable for the investigation of genealogical relationships at deeper time depth [19,20••,21•]. Thus, using such structural features Dunn *et al.* [22•] were able to reconstruct a phylogeny of Papuan languages of Island Melanesia which are unclassifiable using lexical data.

A further such attempt at avoiding the temporal restriction of linguistic data is the 'Parametric Comparison Method' [23–25]. This is based on so-called parameter settings, which are abstract feature values at a varying number of syntactic features supposedly ''. . . predefined by our invariant language faculty, Universal Grammar . . .'' ([24]: 1684). Since this approach assumes that these parameters are part of the innate Universal Grammar, they should be found, and hence comparable, across all languages irrespective of their degree of genealogical relationship [23,24] — making them perfectly comparable to genetic data and avoiding the problems inherent in the use of lexical data mentioned above. Therefore this approach appears to be the ideal solution for the investigation of genetic and linguistic coevolution at a global scale. However, there are several issues that diminish its value. The biggest problem concerns the list of supposedly universal parameters — which is yet to be defined. Even the proponents of the idea admit that ''UG parameters number at least in the hundreds, although we are too far from being able to make precise estimates'' ([24]: 1687), while a survey of the relevant literature was unable to find more than seven parameters that were mentioned by more than one author, none of which were uncontroversial among specialists [26]. In addition, the ascertainment scheme for the parameters is heavily biased: most are proposed on the basis of data from only individual language families or even subfamilies, or at most on a comparison of two very distinct languages, for example, English and Japanese [26]. Furthermore, Longobardi and colleagues propose to restrict their investigation to the domain of nominal arguments, such as Mary, Mary's book, the person I spoke to, etc. [23,24], in order to circumvent the problem of the potentially large number of as yet unidentified parameters for which insufficient data are available across numerous languages. At last count their dataset comprised only 56 parameters, not all of which are independent of each other — restricting the applicability of phylogenetic tools that assume independence of data [25]. Thus their linguistic distance measures are based on a very small set of features ascertained in a small number of languages and taken from a very limited domain of grammar (comparable to a genetic study of global diversity limiting itself to a small number of partly linked SNPs ascertained in only a few populations), which casts considerable doubt on the validity of the results that can be obtained with this method.

While initial analyses of language-gene coevolution simply assessed the degree of correlation between linguistic and genetic diversity, increasingly such studies are used to address specific hypotheses, for example [12•,14••,27•]. For instance, de Filippo *et al.* [14••] used correlations between genetic distances, linguistic distances calculated from cognate lexemes in word lists, and model-based geographic distances to investigate which route would have been followed by the expanding speakers of Bantu languages in sub-Saharan Africa. They find that the data fit a model of a relatively late split of eastern and western Bantu languages, although the signal of this split would have subsequently been diminished by gene flow and language contact (Figure 1). At a broader geographic scale, Hunley *et al.* [27•] undertook a worldwide comparison of the diversity of autosomal microsatellites and that of inventories of distinctive sounds (phonemes) to elucidate whether serial founder events would have shaped both systems equally. With this study they partly addressed the claim by Atkinson [28] that phoneme inventories show evidence of a serial founder event correlating with the Out-of-Africa dispersal of modern humans. In contrast to Atkinson's results, Hunley *et al.* find no evidence for serial founder events in the phoneme data. This different result is most probably due to the very different datasets the studies were based on: Atkinson [28] made use of a published dataset [29] in which phoneme inventories of 504 languages were classified into size bins (e.g. 'small', 'medium', and 'large' for vowel inventories). In contrast, Hunley *et al.* [27•] based their study on 908 phonemes coded as present or absent for 725 languages.