COMMENTARY

Reassessment of global gene–language coevolution

Keith Hunley[1]

Department of Anthropology, University of New Mexico, Albuquerque, NM 87131

In *On the Origin of Species*, Darwin proposed that human races and languages evolved in concert following a tree-like history of splits and isolation (1). Linguists and anthropologists have long been skeptical of this idea because historical and ethnographic evidence suggest that group boundaries are fluid and differentially permeable to the movement of peoples and languages. For this reason, as Sapir (2) so eloquently put it, "the history of each is apt to follow a distinctive course." In PNAS, Creanza et al. (3) weigh in on this debate and provide a conceptual and methodological framework for future studies of population genetic and linguistic coevolution.

Modern-day studies of gene–language coevolution trace to a seminal publication in 1988 by Cavalli-Sforza et al. (4), the goal of which was to construct the history of population splits during human evolution. Those authors took Darwin's proposition as given and used a crude language classification to corroborate the population tree (Fig. 1). The study was criticized by linguists for the reason stated above, and for the additional reason that languages change too quickly to permit reconstruction of linguistic relationships at deep time depths.

In the past decade, genomic studies have enhanced our understanding of human origins and dispersals. At the global level, there is a growing consensus that a serial founder effect (SFE) process played an important role in shaping global patterns of neutral genetic diversity. The process entails a series of population splits, movements into unoccupied territory, and isolation. In humans, the SFE process began in Africa and proceeded through Eurasia into the Americas and Oceania. At the within-population level, the process produced a steady decay in genetic diversity with increasing geographic distance from East Africa; at the between-population level, it produced a steady increase in genetic distance with increasing geographic distance (5, 6).

In 2011, Atkinson (7) reignited the debate about global gene–language coevolution by proposing that phoneme inventories in human languages had undergone a parallel SFE process. His conclusion was based on the finding that the number of phonemes in 504 widespread languages decreased linearly with increasing geographic distance from Africa. If true, Darwin's model has been vindicated, and phoneme inventories from thousands of languages, including many that are extinct and dying, have the potential to provide important details about human evolution.

Creanza et al. (3) enter the fray by performing joint and parallel analyses of the most comprehensive genomic and linguistic data available. The data consist of phoneme counts in 2,082 languages (3, 8) and autosomal microsatellite polymorphisms in 246 populations [collated by Pemberton et al. (9)]. Creanza et al.'s (3) methods consist of a series of sophisticated geospatial analyses, including novel analyses of the geographic axes of greatest genetic and linguistic differentiation and quantification of the effects of drift on the phoneme inventories of isolated languages. The results are decisive with respect to Darwin's proposal at the global level, and they provide novel insights

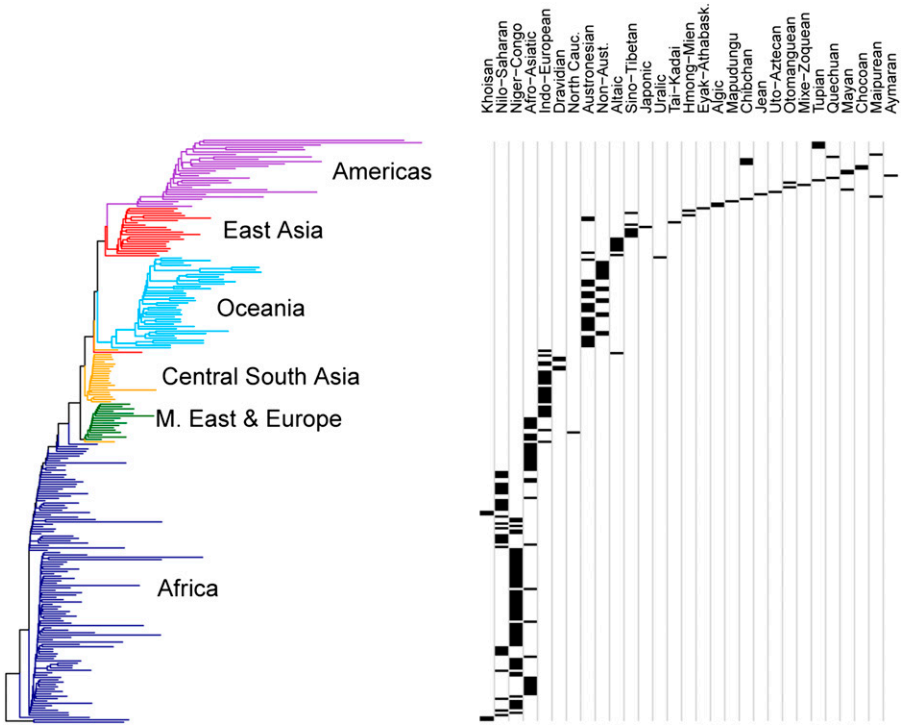

Fig. 1. The population tree was constructed from Nei's minimum genetic distances estimated from autosomal microsatellite polymorphisms in 248 populations (9). Branches are colored according to geographic region. The table shows the language family affiliation of each population taken from the primary classification entry in the Ethnologue (17), with the exception that non-Austronesian languages in Oceania were placed into a single group. The figure corroborates the finding of broad correspondence between genetic and linguistic patterns first identified by Cavalli-Sforza et al. (4). However, because there are no connections between the language families, and no internal structure within them, the correspondence is not a test of Darwin's proposal, and the linguistic pattern does not corroborate the genetic pattern.

Author contributions: K.H. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The author declares no conflict of interest.

See companion article on page 1265 in issue 5 of volume 112.

[1]Email: khunley@unm.edu.

into gene–language coevolution at the regional level.

At the within-group level, in contrast to Atkinson's (7) finding, Creanza et al. (3) find that: (*i*) the phonemic decay is strongest when the origin is located in north Eurasia, not East Africa, and (*ii*) the decay from this origin is not linear, but is instead caused by a dichotomous pattern of high phoneme inventories in Eurasian languages and exceptionally low inventories in South American and Oceanic languages. However, at the regional level, joint analyses by Creanza et al. (3) show that the geographic axes of greatest differentiation are similar for the genetic and phonemic data, suggesting the possibility that languages and populations may have spread together within regions.

At the between-group level, both phonemic and genetic distances are correlated with geographic distance. The correlation for phonemic distances falls off above 10,000 km, whereas the correlation for genetic distances persists across the full geographic range. The persistent correlation of the genetic data is consistent with range expansion predicted by the SFE process, but the falloff in the phonemic correlation is not. Additionally, phonemic distances are correlated with geographic distances whether the phonemes are sampled from the same or from different language families. These results are consistent with an isolation-by-distance process produced by the steady movement of phonemes between neighboring languages.

Two other findings are inconsistent with an SFE process for phonemes. First, in partial Mantel tests, the correlation between genetic and phonemic distances loses significance when geographic distance is controlled, indicating that genetic–phonemic correspondence is purely the result of the correlation of both with geographic distance. Second, Creanza et al. (3) demonstrate that geographically isolated languages have no fewer phonemes than languages with multiple close neighbors. This result suggests that drift acts differently on phoneme levels than on genetic diversity and, therefore, there is no theoretical basis for a phonemic SFE process. These results constitute a rejection of the SFE process for phonemes.

Creanza et al.'s (3) region-level joint analyses leave the door open to the possibility of coevolution at local geographic scales. However, absent a clear theoretical framework for the evolution of phoneme inventories, studies of regional coevolution might do better to concentrate on lexical data. Recently, scholars have refined language classifications using these data in combination with methods from statistical genetics. These studies have improved our understanding of the origin and spread of language families in Oceania, Eurasia, and North America (10–12).

In Oceania, for example, Gray et al. (11) found strong support for a Pulse-Pause model of Austronesian language dispersal from Southeast Asia through Remote Oceania beginning about 5,500 y ago. Aspects of this model are supported by archaeological and genetic data, but these data also reveal a more complex history of interactions between Austronesian speakers and long-resident non-Austronesian speakers in portions of Oceania (13). As a result, there is little correspondence between patterns of population genetic and linguistic diversity in these portions of the region today (14). Most studies of coevolution in other regions also find, at best, weak correlations between patterns of genetic and linguistic diversity (15, 16). For example, of 17 studies reviewed by Barbujani (15), only 6 found a significant correlation between genetic and linguistic distances.

In this regard, in 1921, Sapir provided a compelling description of the myriad ways in which the English language and the so-called English race departed from one another (2). This example, combined with his studies of Native American language and culture, led to his rejection of Darwin's proposal that races and languages evolve in concert. Studies of gene–language coevolution over the past 30 y support this viewpoint. The association between genes and languages is often transient, and the evolution of one may provide limited information about the evolution of the other.

Moving forward, there is still an important place for studies of gene–language coevolution. In recent years, scholars have begun to publish databases of high-quality lexical and grammatical data from hundreds of languages (10, 11). With these data, it is feasible to fit complex evolutionary models using methods from statistical genetics. These models can be jointly fit to genetic data, permitting a more accurate assessment of the degree of gene–language coevolution at different places and times, the evolutionary and social causes of departures in correspondence, and the mechanisms of language transmission in different social and political contexts. This new study by Creanza et al. (3) provides the first step in this endeavor.

1 Darwin C (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* (J. Murray, London).

2 Sapir E (1921) *Language, Race and Culture. Language: An Introduction to the Study of Speech* (Harcourt Brace & Company, New York, NY), pp 207–220.

3 Creanza N, et al. (2015) A comparison of worldwide phonemic and genetic variation in human populations. *Proc Natl Acad Sci USA* 112(5):1265–1272.

4 Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988) Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci USA* 85(16): 6002–6006.

5 Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. *Curr Biol* 15(5): R159–R160.

6 Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 102(44): 15942–15947.

7 Atkinson QD (2011) Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332(6027): 346–349.

8 Moran S, McCloy D, Wright R (2012) Revisiting population size vs. phoneme inventory size. *Language* 88(4):877–893.

9 Pemberton TJ, DeGiorgio M, Rosenberg NA (2013) Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 (Bethesda)* 3(5):891–907.

10 Bouckaert R, et al. (2012) Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097):957–960.

11 Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913):479–483.

12 Wheeler WC, Whiteley PM (2014) Historical linguistics as a sequence optimization problem: The evolution and biogeography of Uto-Aztecan languages. *Cladistics*, 10.1111/cla.12078.

13 Friedlaender JS, et al. (2008) The genetic structure of Pacific Islanders. *PLoS Genet* 4(1):e19.

14 Hunley K, et al. (2008) Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS Genet* 4(10):e1000239.

15 Barbujani G (1991) What do languages tell us about human microevolution? *Trends Ecol Evol* 6(5):151–156.

16 Hunley K, Long JC (2005) Gene flow across linguistic boundaries in Native North American populations. *Proc Natl Acad Sci USA* 102(5):1312–1317.

17 Lewis MP, Simons GF, Fennig CD (2014) *Ethnologue: Languages of the World* (SIL International, Dallas), 17th Ed, Available at www.ethnologue.com. Accessed January 12, 2015.