Do languages and genes share cultural evolutionary history?

Simon J. Greenhill

Languages and genes tell stories about the past but statistical analysis reveals that these are not always the same.

Since Darwin's "On the Origin of Species," linguists and geneticists have implicitly or explicitly connected languages with genetics. A recent paper by Matsumae *et al.* (*1*) provides a new method for assessing whether different lines of evidence consistently tell similar stories about the past and applies this approach to genetic, linguistic, and music data from 14 Northeast Asian populations.

This work is important because while it may seem easy to discover whether the histories told by languages and genes are similar, it is not. Discerning exactly how human populations evolve presents challenges on multiple levels. What theoretical frameworks make sense? How do we work across disciplines? What methods provide robust answers? Matsumae *et al.* (*1*) address these issues, providing an elegant way to compare what different sources tell us.

To uncover the origins and relationships of human groups around the globe, geneticists use genetic markers (Y chromosome and mitochondrial or autosomal DNA), and linguists use linguistic markers (lexical cognates and grammatical features). An important question then is whether inferences drawn from these different markers are consistent or not: Are they all telling us the same thing or do they each tell us different things about different aspects of human prehistory? Consistency matters because the interpretation of these markers often recruits evidence across disciplines. For example, geneticists often interpret their results in terms of cultural or linguistic groupings and vice versa. However, if languages and genes do not share the same evolutionary history, then interpreting their histories together is problematic.

Why should languages and genes tell us the same story about prehistory? There is only shared history that has generated all the genetic and linguistic variation around the globe. We might expect that the processes and events that caused human populations to split and diverge during prehistory would simultaneously affect those populations' genes and languages. For example, if a group of people moved to settle in a new region, forming a new community, we might expect them to stop talking with and stop intermixing with groups in their homeland (even if only due to the tyranny of distance). Therefore, over time, their languages and genes would then accumulate unique differences from their original population. Furthermore, languages can act as barriers, blocking contact and interaction between people. These barriers would then essentially shoehorn the genetic variation into the boundaries of the linguistic history. However, there are also good reasons to be skeptical of strong matches between language and genes (*2*). A person's genetic makeup is inherited at birth from their parents, but people often speak multiple languages at the same time and can change their language over their life span. Communities can readily change their languages, shifting to others that may be more politically or socially dominant. For example, people in modern day Tenōchtitlan, Mexico City, now tend to speak Mexican Spanish rather than Nahuatl, which was spoken following Spanish conquest, but still show strong indigenous genetic ancestry.

Languages also tend to evolve far more rapidly than genes. The great Austronesian expansion that started spreading across the Pacific from Taiwan to Hawaii 5500 years ago generated more than 1200 very distinct languages (*3*). This rapid rate of change means that linguistic change quickly overwrites and removes any deeper signal. Therefore, genes and languages might be evolving at radically different time scales, and any overlap may just be due to chance.

In the late 1980s, linguists and geneticists debated these issues following the publication of a series of prominent papers. The key study by Cavalli-Sforza *et al.* (*4*) compared the global genetic population tree markers to a global language tree. The authors (*5*) suggested that the genetic and linguistic groupings overlapped significantly, and they argued that these groups must share common origins. Controversy followed. One lengthy critique claimed that Cavalli-Sforza *et al.*'s conclusions were "undermined by analysis of an inadequate database by inappropriate methods and by several conceptual flaws in subsequent interpretations" (*5*). This criticism, of course, was strenuously denied by the authors and debated at length by leading figures in the field.

Why is this a hard debate to put to rest? For a start, there are practical difficulties caused by working across disciplines. It is very difficult to map a linguistic population onto a genetic population. For example, a survey of 100 genetic studies that included Native American populations (*6*) found that 80% of them cite or are influenced by a large-scale "Amerind" classification. However, this classification is highly problematic and rejected by almost all linguistic experts (*2*). This means that the interpretations of the results reported by these studies are flawed, discussing fictional entities that do not exist.

There are technical difficulties too. Many of the previous studies relied on a problematic statistical method—the Mantel test—that is notorious for having very low power and a high false-positive rate (*7*), especially when used to tease apart the effect of genes versus languages while holding other factors like geography constant. This means that published findings about the correlations between languages and genes could well be incorrect.

This is the battlefield that Matsumae *et al.* enter (*2*). They focus on 14 cultures from

Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena 07745, Germany. Email: greenhill@shh.mpg.de

northeast Asia and bring a new weapon and an impressive array of ammunition from these cultures: genetic data, linguistic data including grammar phonology and lexicon, and musical traditions (Fig. 1).

Matsumae *et al.* analyze these data using network methods and redundancy analysis. Interestingly, they find that clustering analyses of each datatype show very different signals of human prehistory. The neighboring Korean and Japanese cultures, for example, cluster together in grammar, genes, and music but not in the lexicon or phonology. In contrast, the two Uralic languages Selkup and Nganasan group together based on their genes, lexicon, and grammar but not their phonology and musical traditions. However, two datatypes do show a striking relationship: The grammar and the genetic markers are strongly correlated with each other. Why? There are three options. Either the shared patterns are caused by recent contact between these populations, or by some of the populations sharing recent history within language families, or they reflect deep historical signal between language families. By controlling for geography and recent history, Matsumae *et al.* claim that the shared patterns are more consistent with deep historical relationships.

There are many implications of this study. The first concerns human prehistory. North Asia is the site of ongoing and vicious disputes about deep linguistic relationships: Are the Korean and Japanese language families linked? Can we connect many of the language families in the region into superfamilies like Altaic? Perhaps Matsumae *et al.* have put their finger on why it is hard to solve these debates: It depends on the data one examines.

Second, we should be very careful when we naïvely link linguistic constructs to genetic histories (and vice versa). The choice of data for investigating different time scales is vitally important—perhaps lexical data could be used for recent time scales while using grammar and genes for deeper time scales? However, more work will be needed to uncover the situations and places appropriate for each dataset.

Last, the biggest implication is the complexity involved in teasing out whether languages and genes tell us the same story. Instead, the question becomes how, why, and when do certain histories cohere, and when do they evolve independently. The methods applied by Matsumae *et al.* provide a promising way forward, but the problem still to be solved is to identify the mechanisms that are shaping these aspects of our history. Matsumae *et al.* are silent on potential mechanisms, but some clues might come from population history, where we might expect that tight couplings of these aspects happen when populations spread rapidly, while uncoupling happens with long-term interactions between populations. More clues come from the characteristics of the datatypes themselves. Some aspects are largely invisible to people—e.g., genes and grammar—and might therefore change in a relatively slow and neutral manner. However, other aspects are very salient to people and can readily be recruited to delimit social groups—e.g., phonology, music, and lexicon—which could act to either speed up their rates of change or fix them as stable cultural components (*8*).

What is clear is that we need more abundant datasets of cultural, linguistic, and genetic data that can be aligned in time and space to provide a rich and multifaceted lens that will allow us to glimpse the full cultural evolutionary history of our species.

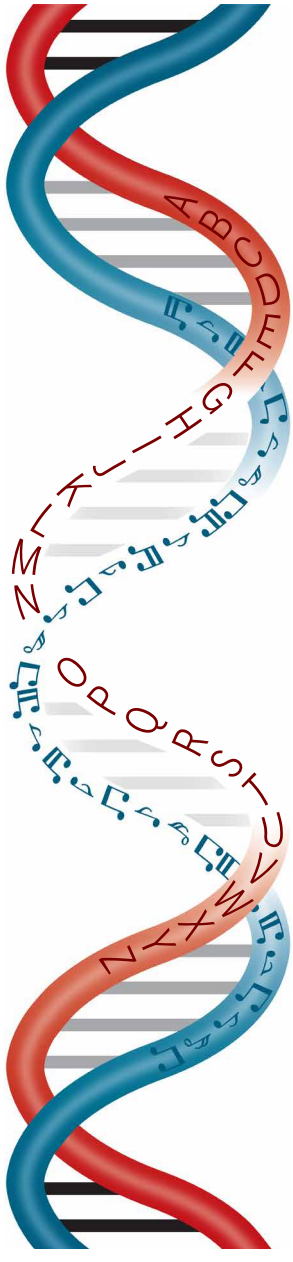

Fig. 1. Datasets of cultural, linguistic, and genetic information will create a rich, multifaceted lens to glimpse the full cultural evolutionary history of our species. Credit: Ashley Mastin/*Science Advances*.

REFERENCES AND NOTES

1. H. Matsumae, P. Ranacher, P. E. Savage, D. E. Blasi, T. E. Currie, K. Kognebuchi, N. Nishida, T. Sato, H. Tanabe, A. Tajima, S. Brown, M. Stoneking, K. K. Shimizu, H. Oota, B. Bickel, Exploring correlations in genetic and cultural variation across language families in northeast Asia. *Sci. Adv.* **7**, abd9223 (2021).
2. L. Campbell, Do languages and genes correlate? *Lang. Dyn. Change* **5**, 202–226 (2015).
3. R. D. Gray, A. J. Drummond, S. J. Greenhill, Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
4. L. L. Cavalli-Sforza, P. M. Menozzi, Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 6002–6006 (1988).
5. R. Bateman, R. O'Grady, V. A. Funk, R. Mooi, W. J. Kress, P. Cannell, D. F. Armstrong, D. Bayard, B. G. Blount, C. A. Callaghan, L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, J. Mountain, J. H. Greenberg, K. Jacobs, Y. Mizoguchi, M. Nunez, R. L. Oswalt, Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language [and comments]. *Curr. Anthropol.* **31**, 1–24 (1990).
6. D. A. Bolnick, B. A. Shook, L. Campbell, I. Goddard, Problematic use of Greenberg's linguistic classification of the Americas in studies of native american genetic variation. *Am. J. Hum. Genet.* **75**, 519–522 (2004).
7. L. J. Harmon, R. E. Glor, Poor statistical performance of the mantel test in phylogenetic comparative analyses. *Evolution* **64**, 2173–2178 (2010).
8. S. J. Greenhill, C.-H. Wu, X. Hua, M. Dunn, S. C. Levinson, R. D. Gray, Evolutionary dynamics of language systems. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E8822–E8829 (2017).

10.1126/sciadv.abm2472

Citation: S. J. Greenhill, Do languages and genes share cultural evolutionary history? *Sci. Adv.* **7**, eabm2472 (2021).

ScienceAdvances

Do languages and genes share cultural evolutionary history?

Simon J. Greenhill

Sci. Adv., 7 (41), eabm2472. • DOI: 10.1126/sciadv.abm2472