Mitochondrial DNA Tracks the Spread of Linguistic Structures Through Unrelated Languages in Africa and Eurasia

Author: Jeremy Collins[1,2*]

Affiliations:

[1]Radboud University, Nijmegen, The Netherlands.

[2]International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands.

*Correspondence to JeremyCollinsMPI@qq.com

Abstract: Language families are groupings of languages based on shared vocabulary innovations, which are useful for demonstrating cultural relatedness of populations, and hence overlap to some extent with the distribution of genetic markers. This paper shows that abstract properties of languages such as grammatical features or phonemes also correlate strongly with genetic markers, after controlling for known language families. Linguistic structures can remain unchanged in language families over thousands of years, but are also crucially borrowed across language families in ways corresponding to gene flow between populations. Genetic variation can explain why unrelated languages share particular structural properties, and also provide evidence for events beyond the reach of current linguistic reconstruction, such as the arrival of languages in North Africa from Eurasia in the early Neolithic.

One Sentence Summary: Mitochondrial DNA haplogroups can predict structural properties of unrelated languages, and push reconstruction of language history beyond the limits of established language families.

Main Text: Genes and languages can often travel together, when people migrate to new regions and bring languages with them. Y chromosome and mitochondrial DNA haplogroups in particular are therefore often compared in their distribution to language families *(1)*. This paper investigates a different side of language history: the way that structural properties of languages can be conserved in language families, and also travel between them. Abstract properties can travel between languages in a way analogous to gene flow between populations. For example, Vietnamese became a tonal language (using pitch to distinguish different words) under the influence of Chinese, despite the fact that they are unrelated (*2*). Similarly, some Bantu languages acquired click sounds from neighbouring San languages (*3*). This process, called 'language contact', raises the question of whether grammatical and phonological properties travel with genes. There is some evidence that linguistic structures are stable and distinctive enough over time to be informative about population history *(4)*. The question is whether the distribution of these structures in unrelated languages can be explained by the movement of people between populations, perhaps of specific lineages that can be tracked by Y chromosome or mitochondrial DNA.

Consider how languages vary in the complexity of consonant clusters, from Georgian which allows words with several consecutive consonants such as *mtsvane*, to Japanese which mostly only allows words with a single consonant in a sequence (apart from 'n') such as *hajimaru* 'begin'. Rules for how many consonants are allowed can travel between languages; for example, Chinese or Malay speakers in Singapore learn English and simplify some consonant clusters

such as *lisp* to *lis* (*5*), because Chinese and Malay lack these consonant clusters. Conversely, some second-language learners can complexify the language that they are learning. Moroccan Arabic has developed complex consonant clusters where Standard Arabic does not have them, such as by changing Standard Arabic *kitabtu* to *ktebt*, and this is likely to be under the influence of the original local Berber languages which themselves allow very complex consonant clusters (*6*).

When a population adopts a new language, as in the above cases, they may therefore change that language according to the structures of the language that they used to speak. This process is called 'substrate influence'. This means that mitochondrial DNA haplogroups might be especially likely to be predictors of syllable structures, because of the finding by Forster and Renfrew that languages are often brought by a small group of men into a new region, making expanding language families correlate with Y chromosome haplogroups, while mtDNA haplogroups typically reflect populations already in the region, who would speak a 'substrate' language *(1)*. In cases such as English being adopted by Chinese speakers in Singapore, then it is the languages of the substrate population (Chinese and Malay) which are predicted to influence the syllable structure of the superstrate language (English). Syllable structures may sometimes be a trace of languages that used to be spoken in regions before the arrival of new languages, and hence show another side of linguistic history resembling that shown by mitochondrial DNA.

If particular syllable structure rules are found across several language families, then they may correlate with the distribution of specific mitochondrial DNA haplogroups. This hypothesis was tested on 73 populations in Africa and Eurasia up to the Pacific, from twenty-six different language families. Population frequencies of 252 haplogroups were collected from genetics literature on these populations *(7-20)*, ranging from the most general, such as haplogroup M which is found everywhere outside of Africa, to the most specific haplogroups for which there is data such as Bc41b. Three syllable structure features were tested (how many consonants are allowed at the beginning of a syllable, how many at the end, and how many tones are used), using data from an independently coded database *(21)*. These three syllable features were tested in linear regressions with all 252 haplogroups. The way to correct for multiple testing was to take the p-values of all the linear regressions and find how many p-values were in between 0 and 0.05 compared with p-values at other intervals (e.g. between 0.4 and 0.45). If there are genuine correlations between consonant cluster complexity and haplogroups above that expected by multiple testing, then there should be many more below p-values below 0.05 than in other intervals. This pattern holds for the actual data, and does not hold for a random arrangement of the data (figures S1-S2). All three syllable features show more significant correlations than expected by chance; for example, coda complexity correlated significantly ($p<0.05$) with 77 haplogroups, compared with 28 by the interval with the second greatest number of correlations ($0.3<p<0.35$). Mixed effects models were then used with language family as a random intercept, and in all three cases there were more significant correlations than expected by chance. A different way of avoiding multiple testing was to use a Mantel test of overall genetic distance between populations and differences in syllable structures, the results of which were significant for coda complexity ($r=0.35$, $p=0.001$) and number of tones ($r=0.2$, $p=0.03$) but not onset complexity ($p=0.17$); using partial Mantel tests one can also simultaneously control for language family and geographic distance, for which the results are still significant (coda complexity $r=0.14$, $p=0.003$; number of tones $r=0.18$, $p=0.008$). Genetic distance in fact correlates with

these two features better than language families do (r=0.14 for coda complexity and r=0.15 for number of tones).

Having established that there are real correlations between syllable structures and mtDNA haplogroups, the strongest correlations can be chosen to see whether particular population movements may be responsible for the spread of particular structures. The strongest correlation is between coda complexity and haplogroup HV (R^2=0.34, p=3.8e-08), a correlation that is significant after a Bonferroni correction for the 756 tests done. Haplogroup HV is a large haplogroup found in Europe, North Africa, and decreasing in frequency in eastern Eurasia (Fig. 1). This distribution is similar to the distribution of languages that allow complex consonant clusters in the coda of syllables (Fig. 2).

Why might the distributions in the two maps be similar? The first reason is because of known relatedness of languages: for instance, English and German both have complex syllables, and speakers of these languages have high frequencies of haplogroup HV. This shows that syllable structure is fairly stable within language families such as Indo-European; but the question is whether this correlation holds beyond that expected because of known language families. In order to answer this, a mixed effects model with haplogroup HV frequency and random intercepts for language family was used, which predicted coda complexity significantly better than a model with just random intercepts for language family (χ^2=5.39, p=0.02). A second method was to randomly sample one language per known language family and retest the correlation between haplogroup HV and coda complexity, which was significant in 49% of the random samples. Haplogroup HV also correlates with other syllabic features such as the complexity of syllable onsets (R^2=0.19, p=7.4e-05) and inversely with the number of tonal distinctions (R^2=0.13, p=0.013), and both are similarly significant in mixed effects models ($p<0.05$).

This result mainly shows that complex consonant clusters have been traveling between language families in Eurasia such as Indo-European, Uralic, Northwest Caucasian, Basque and Dravidian along a similar path to that of the spread of haplogroup HV, either because these language families have been in contact or in some cases because they may be distantly related. Unexpectedly, complex syllables and haplogroup HV also show a similar distribution within Africa. Haplogroup HV correlates significantly within African languages taken by themselves with coda complexity (R^2=0.67, p=0.015) and inversely with number of tonal distinctions (R^2=0.78, p=0.006), although not with onset complexity (R^2=0.34, p=0.1). On an expanded sample of 13 African languages for which haplogroup HV data was available *(22)*, all three of these correlations with haplogroup HV were significant (e.g. with codas, R^2=0.82, p=2e-05). This suggests that the presence of these complex consonant clusters in North African languages is not a random independent development, but connected to the presence of haplogroup HV. The highest frequencies of HV and the most complex syllables are found in North Africa in Berber populations of Morocco and Algeria. The syllable complexity of Berber languages is not likely to be because of Indo-European influence (such as from Latin or French), because Berber languages allow syllables typically more complex than those found in Indo-European languages (often even without any vowels such as *tkst* 'you feed on'), and these are primarily in indigenous words that are not known to be borrowed or influenced by any Indo-European language *(23)*; nor were these consonant clusters introduced by Arabic, which does not allow them. Instead, Berber languages seem to have a Eurasian affinity, in the same way that Berber populations have high frequencies of Eurasian mtDNA haplogroups (e.g. haplogroup U and haplogroup J, which also

correlate with complex codas). The migration of these maternal lineages into North Africa from the Near East or Iberia during the early Neolithic *(24)* may have brought languages with complex codas from Eurasia in North Africa. Since Berber is a branch of the Afro-Asiatic family, this would either suggest that Afro-Asiatic languages may have originated in the Near East, or that complex syllables came into this branch of Afro-Asiatic by substrate influence. In either scenario, it is linguistic evidence for Neolithic Eurasian migration into North Africa, for which there is clear archeological and genetic evidence, but which has so far been lacking in linguistics *(25)*.

To show that mitochondrial DNA haplogroups correlate with features than syllable structures, mtDNA was then compared with data on phonemes *(26)* and word orders *(27)*, which again both showed more significant correlations than expected by chance (Figs. S7-S8). The word order that showed the most significant correlations with mtDNA was the order of numeral and noun, which may be because numerals have been found to be among the most stable words in the lexicon *(28)*. Numerals such as 'two' and 'five' are cognate across all branches of Indo-European, and this durability of numerals may be why their ordering in relation to the noun (preceding the noun in the phrase 'two people') is also preserved across almost all Indo-European languages, in contrast with other word orders such as that of verb and object, which vary between even closely related languages such as English and German *(27)*. Numeral-noun order correlates in a logistic regression with Eurasian haplogroups such as HV ($p=0.001$) and U ($p=0.0003$) in particular, and is the word order found in Berber languages and related Afro-Asiatic languages in Egypt, providing additional support for their Eurasian affinity (Fig. S9). The correlation with haplogroup U, although not with haplogroup HV, in addition holds up within African languages by themselves ($p=5.9e-06$).

Genetic variation such as that involving mitochondrial DNA haplogroups thus has the power to explain why some languages have more complex consonant clusters than others, and in some cases is a better predictor of these traits than known language families. The results also show that individual structural properties of languages can be distinctive and durable, and can spread in a meme-like way through unrelated languages, in ways closely correlating with the spread of individual genetic variants through neighboring populations. When treated as analogous to genetic variants, the distributions of linguistic structures hold promise for showing the way that populations have been spreading and interacting in prehistory.

References and Notes:

1. P. Forster, C. Renfrew, Mother tongue and Y chromosomes. *Science* **333,** 1390-1391 (2011)
2. N. J. Enfield, Areal linguistics and mainland Southeast Asia. *Annual Review of Anthropology* **34**, 181-206 (2005)
3. C. Barbieri, A. Butthof, K. Bostoen, B. Pakendorf, Genetic perspectives on the origin of clicks in Bantu languages from southwestern Zambia. *Eur. J. Hum. Genet.* **21**, 430-436 (2013)
4. M. Dunn, A. Terrill, G. Reesink, R. A. Foley, S. C. Levinson, Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072-5 (2005)
5. L. Lim, Ed., *Singapore English: A Grammatical Description.* (John Benjamins, Amsterdam, 2004)

6. U. Maas, in J. Owens, Ed., *Arabic as a Minority Language.* (Mouton de Gruyter, New York, 2000) 383-404
7. C. Barbieri *et al.*, Ancient substructure in early mtDNA lineages of Southern Africa. *Am. J. Hum. Genet.* **92(2)**, 285-292 (2013)
8. C. Barbieri *et al.*, Contrasting Maternal and Paternal Histories in the Linguistic Context of Burkina Faso. *Mol. Biol. Evol.* **29(4)**, 1213-1223. (2012)
9. M. Derenko *et al*, Phylogeographic analysis of mitochondrial DNA in Northern Asian populations. *Am. J. Hum. Genet.* **81(5)**, 1025-1041 (2007)
10. A. T. Duggan *et al.*, Maternal History of Oceania from Complete mtDNA Genomes: Contrasting Ancient Diversity with Recent Homogenization Due to the Austronesian Expansion. *Am. J. Hum. Genet.* **94(5)**, 721-733. (2014)
11. S. A. Federova *et al.*, Analysis of Mitochondrial DNA Lineages in Yakuts. *Molecular Biology* **37**, 544-553 (2003)
12. A. Helgason *et al.*, mtDNA and the Islands of the North Atlantic: Estimating the Proportions of Norse and Gaelic Ancestry, *Am. J. Hum. Genet.* **68**, 723-737 (2001)
13. T. Kivisild *et al.*, The Genetic Heritage of the Earliest Settlers Persists Both in Indian Tribal and Caste Populations. *Am. J. Hum. Genet.* **2(2)**, 313-332 (2003)
14. M. S. Peng *et al.*, Tracing the Austronesian Footprint in Mainland Southeast Asia: A Perspective from Mitochondrial DNA. *Mol. Biol. Evol.* **27(10)**, 2417-2430 (2010)
15. S. Plaza *et al.*, Joining the Pillars of Hercules, mtDNA Sequences Show Multidirectional Gene Flow in the Western Mediterranean, *Ann. Hum. Genet.* **67**, 312-328 (2003)
16. L. Quintana-Murci *et al.*, Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845 (2004)
17. E. B. Starikovskaya *et al.*, Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. *Ann. Hum. Genet.* **69**, 67-89 (2005)
18. M. Tanaka *et al.*, Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14**, 1832–1850 (2004)
19. B. Wen *et al.*, Genetic Structure of Hmong-Mien Speaking Populations in East Asia as Revealed by mtDNA Lineages. *Mol. Biol. Evol.* **22(3)**, 725-734 (2005)
20. B. Yunusbayev *et al.*, The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29(1)**, 359-365 (2012)
21. M. Donohue, R. Hetherington, J. McElvenny, V. Dawson, World Phonotactics Database. 2013. Department of Linguistics, The Australian National University (2013) http://phonotactics.anu.edu.au. Accessed 1/10/2014.
22. D. A. Badro *et al.*, Y-Chromosome and mtDNA Genetics Reveal Significant Contrasts in Affinities of Modern Middle Eastern Populations with European and African Populations. *PLoS ONE* **8(1)**, e54616. (2013)
23. R. Ridouane, Syllables without vowels: phonetic and phonological evidence from Tashlhiyt Berber. *Phonology* **25**, 321-359 (2008)
24. H. Ennafaa *et al.*, Mitochondrial DNA Haplogroup H Structure in North Africa. *BMC Genetics* **10**, 8 (2009)
25. P. Bellwood, *First Migrants: Ancient Migration in Global Perspective.* (Wiley, 2013)
26. S. Moran, D. McCloy, R. Wright, Eds. PHOIBLE Online. (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2014)

27. M. S. Dryer, M. Chapters 82-83 and 85-90, in M. S. Dryer, M. Haspelmath, Eds., The World Atlas of Language Structures Online. (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013).
28. M. Pagel M, Q. D. Atkinson, A. Meade, Frequency of word use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717-720 (2007)
29. H. Hammarström, R. Forkel, M. Haspelmath, S. Nordhoff, Glottolog 2.0. (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2014).
30. M. van Oven, M. Kayser, Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30(2)**, E386-E394. (2009) http://www.phylotree.org.

Acknowledgements: Thanks to D. Dediu, M. Dunn, N. Enfield, S. Graham, R. Hunsucker, S. Levinson, P. Muysken, E. Ritten, S. Roberts, and K. Smiet for feedback. This work is supported by the Centre for Language Studies, Radboud University, Nijmegen.

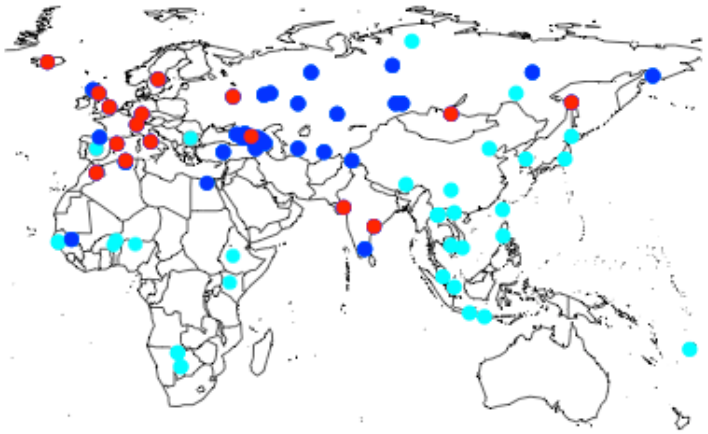

Fig. 1. The spread of complex syllables: languages in red allow three or more consonants in the coda of syllables; languages in dark blue allow two consonants; languages in light blue allow fewer than two consonants.

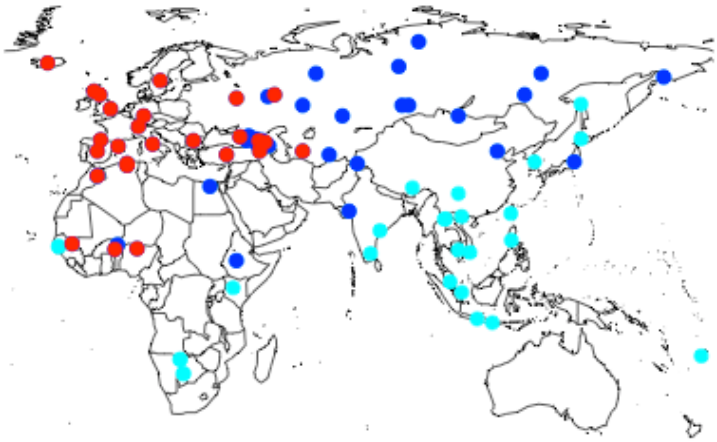

Fig. 2. The spread of haplogroup HV: populations in red have a population frequency of haplogroup HV above 30%; populations in dark blue have a frequency above 0%; and populations in light blue do not have haplogroup HV.