Supporting Online Material for

Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa"

Rory Van Tuyl* and Asya Pereltsvaig

*To whom correspondence should be addressed. E-mail: roryvantuyl@gmail.com

Published 10 February 2012, *Science* **335**, 657-d (2012)
DOI: 10.1126/science.1209176

This PDF file includes:

Supporting Online Material for

Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa"

Rory Van Tuyl and Asya Pereltsvaig
Email – roryvantuyl@gmail.com

Table of Contents

1. Materials and Methods

1.1 Data Source

All data were obtained from the original paper's Supporting Online Material (SOM- Available at: www.sciencemag.org/cgi/content/full/332/6027/346/DC1)

1.2 Computation Method

Computation was performed using standard statistical and plotting functions in Microsoft Excel 2007.

1.3 Mapping with Pearson's Product Moment Correlation Coefficient

As in the original Atkinson paper, we used Pearson's Product Moment Correlation Coefficient between *d*, the distance from a chosen test point to the stated location for a language, and TNPD for that language (s is the standard deviation for the entire data set, $\bar{d}$ and $\overline{TNPD}$ the mean values of the variables, n is the number of data points in the set):

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(d_i - \bar{d})}{s_d} * \frac{(TNPD_i - \overline{TNPD})}{s_{TNPD}} \qquad \text{(Eq. S1)}$$

The point at which R was maximized we called the *Point of Maximum Correlation* or *Point of Origin* [PO]. The distance from the test point to all points in the data set was calculated using spherical trigonometry and was matched to Atkinson's published results as an accuracy check. Peak R was determined by manually altering the Latitude and Longitude of the reference point. Contours of constant R (normalized to R_{max}) were generated from a grid of 88 reference points extending from Longitude 15°W to 35°E and Latitude 30°N to 40°S, with the results shown in Fig. 2A.

The map of African TNPD topography (Fig. 2B) was generated using a grid of 240 test windows of 1000 km diameter spaced at 557km intervals covering Longitude 20°W to 50°E and Latitude 40°N to 35°S, in which the average TNPD of all languages in each window was recorded and plotted (Fig. S1).

The discrepancy between the contours of constant R and contours of constant TNPD is due to an inherent limitation of the Correlation Mapping technique. Because R at every test point on the grid is a function of ALL data points in the space (Eq. S1), locality is lost, and artifacts such as significant values of R in the Atlantic Ocean result. This of course casts doubt on the meaning, if any, of the PO and the contours of R. However, topographical values of TNPD depend on only those languages in the 1000 km vicinity of the test point, so the meaning is clear.

To see what effect TNPD values had on the location of the PO, we substituted random numbers (with the same range of values) for actual TNPD data. Fig. S1 implies that a different geographical distribution of data would have yielded different results.

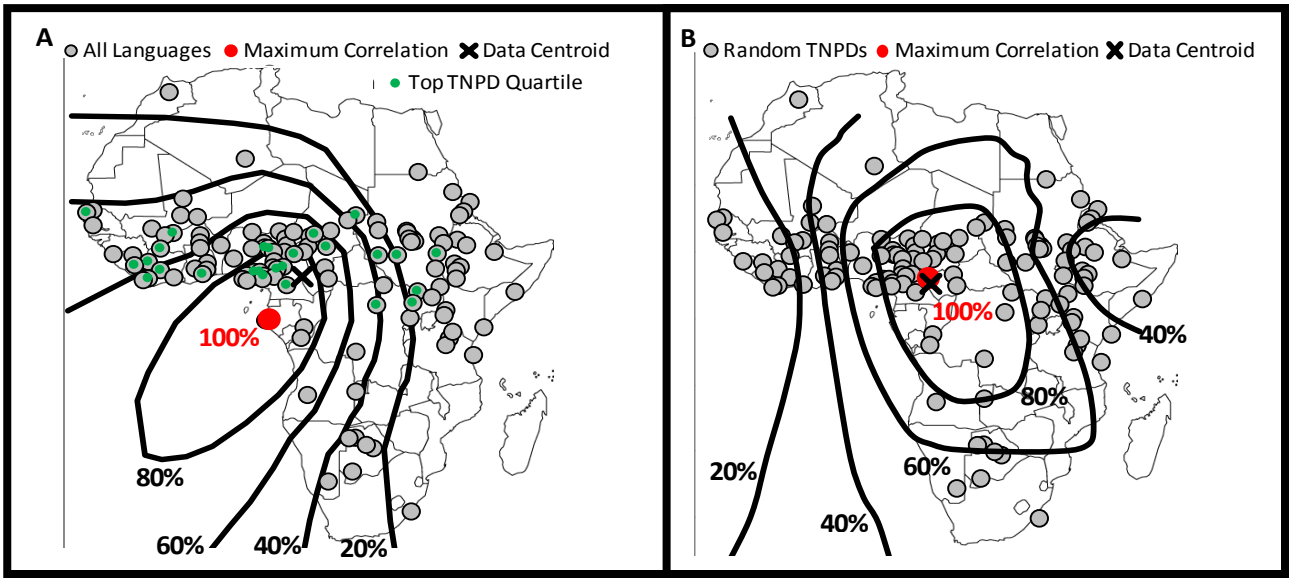

Fig. S1. Map (A) of African languages with contours of constant normalized correlation coefficient centered about a peak value (100%). Maximum correlation is 737 km southwest of the centroid for all data points. Map (B) assigns random numbers for TNPD at locations of actual languages. Maximum correlation is at the centroid for all data points. These maps imply that while TNPD distribution has some effect on correlation contours, a correlation maximum can exist even with random TNPDs, and that a major determinant for the apparent "Point of Origin" location may be nothing more than today's geographic distribution of data points. Distribution in antiquity is unknown.

1.4 Segmented Regression

We adopted the following protocol for linear regression fitting of data: only the expectation function that produces the lowest standard deviation of the residuals should be used. So we used segmented regression continent by continent for the TNPD -vs- distance plots in Fig. 1. The result was a 7% better fit to the data than would have resulted from a linear regression across all continents (Note 4). We concur with Atkinson's finding (his Fig. 1b) that there is a stepwise decline in TNPD between continents. We calculated these steps to be equivalent to 1.1-1.4 vowels ***or*** 0.52-0.66 tones (Table S1).

SFE theory involves a one-way flow of phoneme inventory from predecessor groups to founder groups. As TNPD approaches a minimum at the continental waypoint, a single language emerges from the statistical possibilities at that waypoint to form the founder language for the new continent. Table S1 shows the size of TNPD steps at the continental waypoints:

Waypoint	Location: Latitude/ Longitude	Upstream Intercept	Downstream Intercept	Step in Normalized Units	Step in Equivalent Vowels	Step in Equivalent Consonants	Step in Equivalent Tones
Africa-N. Asia	30N/31E	0.343	-0.049	0.392	1.121	2.956	0.517
N. Asia - N. America	66N/170W	-0.024	-0.433	0.409	1.169	3.083	0.540
N. America - S. America	8N/77.5W	-0.018	-0.520	0.502	1.437	3.787	0.663

Table S1: Step offsets of TNPD between continents amount to 1.1 to 1.4 vowels or 0.52 to 0.66 tones. Consonants play little role in TNPD regression (Note 6).

1.5 Regression Through Waypoints

Atkinson adopted the method of Ramachandran (1) for using waypoints between continents to help approximate the migration paths of prehistoric peoples. The great circle distances between waypoints represent the shortest possible migration path, not the actual path. Ideally, the routes between waypoints should be over land only, but in fact pass over the Arctic Ocean and the Gulf of Mexico (2). The waypoint method measures the "downstream" distances (b_i) with reference to the waypoint, and the "upstream" distances (a_i) with reference to the previous waypoint or, as in the case of Africa, a Point of Origin determined by maximum correlation. This scheme is consistent with the Founder Effect model, where upstream TNPD values are completely independent of downstream values. The slope of the regression fit from PO, when projected forward to the waypoint, calculates the Upstream Intercept, the most probable value of the founder language based on upstream data. Similarly, the downstream data regression, when projected backwards to the waypoint, defines the Downstream Intercept, the most probable value of the founder language based on downstream data. The difference in these intercept values (Table S1) represents the predicted change in phoneme inventory as languages cross the continental waypoint. Because out-of-African languages downstream of the waypoint exhibit a different regression slope than African languages upstream of the waypoint, we deem in

incorrect to calculate a regression line across a waypoint. Causality requires that the downstream data have no effect on the upstream regression slope, since upstream languages evolved at an earlier time. A continuous regression line forces a change to the slope of the regression fit to upstream data, violating this causality constraint.

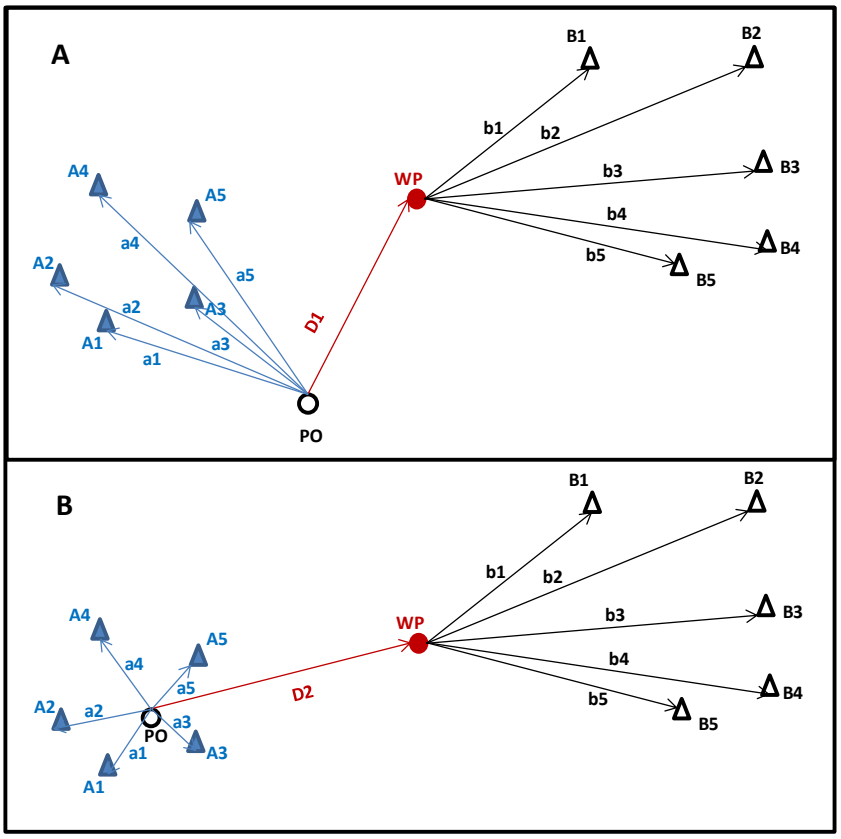

Fig. S2. Diagram showing how repositioning the Point of Origin affects distances to data points differently, depending on whether the data points are upstream (blue) or downstream (black) of a waypoint (WP). The downstream distances (b_i) are referenced to the waypoint and the upstream distances (a_i) to the Point of Origin [PO]. When the PO location is changed, the upstream distances all change by different amounts, depending on the two-dimensional relationship between PO and the points A_i. Consequently, the regression slope and correlation coefficient for upstream points change. But downstream points retain their relative positions, regression slopes and correlation coefficients when the PO moves. The PO-to-B_i distances all shift equally by the amount (D2 – D1). This effect is visible in Fig. 1.

[1] S. Ramachandran *et al.*, *Proceedings of the National Academy of Sciences (USA)*, **102**, 15942 (2005).
[2] Google Earth illustrates this.