Analysis Description for Comment on "Human sound systems are shaped by post-Neolithic changes in bite configuration"

Sergei Tarasov, Josef C. Uyeda

May 29, 2019

All scripts for running the described analyses are located at `https://github.com/sergeitarasov/Tarasov_Uyeda_SupplementaryMaterials_2019`

1 Binomial Causal Models (BCM)

The probability of hunter-gatherers (HG) is non-uniformly distributed across the globe and hence is dependent on geographic area. In turn, geographic areas bear phylogenetic signal for language relatedness due to historical human dispersal or ancestry – specifically, geography largely determines language families. Different language families may have varying potential to produce languages with and without labiodental phonemes (LB) purely due to their ancestry (i.e., the heritable features of language passed through the phylogeny). Based on these facts, it's logical to assume that the probability of hunter-gatherers (θ_H) and labiodentals (θ_L) might be correlated due to their shared dependence on geography, without a direct effect of hunter-gatherer societies driving changes in the probability of labiodentals. Alternatively, the probability of labiodentals can be dependent on both geography and the probability of hunter-gatherers due to the latter's causal affect on bite configuration.

We test these hypotheses by assessing independent and dependent binomial causal models using Bayesian inference. The independent model assumes that even though θ_H and θ_L are dependent on the geography, they must be conditionally independent given that geography is observed. Thus, the distribution of θ_H and θ_L may covary across geographic areas but be independent within an area. The dependent model is similar to the independent one but assumes the dependence of θ_L on θ_H within a geographic area.

1.1 Independent model M_{ind}

Model Assumption: Presence of labiodental phonemes is independent from subsistence within a geographic area. Presence of labiodentals and subsistence are modeled as independent binomial proportions.

Model Specification:

- N_i number of languages per Area i.

- H_i number of languages with the hunter-gathering society per Area i.
- HL_i number of languages where the society is hunter-gathering and labiodental phonemes are present per Area i.
- AL_i number of languages where the society is agricultural and labiodental phonemes are present per Area i.
- θ_{Hi} probability of languages with hunter-gathering societies per Area i.
- θ_{Li} probability of languages with labiodental phonemes present per Area i.

Model Description:

- $H_i \sim Binomial(N_i, \theta_{Hi})$
- $HL_i \sim Binomial(H_i, \theta_{Li})$
- $AL_i \sim Binomial(N_i - H_i, \theta_{Li})$

In this independent model both variables H_{Li} and A_{Li} are generated by the same parameter θ_{Li}.

1.2 Dependent model M_{dep}

Model Assumption: Presence of labiodental phonemes is dependent on subsistence within a geographic area.

Model Specification: Same as the independent model.

Model Description: the description is similar as the the independent model except that the variables HL_i and AL_i are generated by two different parameters θ_{L1i} and θ_{L2i} indicating that probability of languages with labiodental phonemes present in Area i is different between hunter gathering and agricultural societies.

- $H_i \sim Binomial(N_i, \theta_{Hi})$
- $HL_i \sim Binomial(H_i, \theta_{L1i})$
- $AL_1 \sim Binomial(N_i - H_i, \theta_{L2i})$

1.3 Inference

We ran each model in Rstan (*1*) (uniform priors for all theta parameters) to estimate the parameters and marginal likelihoods for the dependent and independent models for each area separately. Bayes factors (BFs) calculated at the logarithmic scale $BF(M_{ind} - M_{dep})$ are given in Table 1.

1.4 Results

As in the original paper (*2*), if the whole world is analyzed as a single area, we obtain significant support for the dependent model. However, if areas are analyzed separately then, in the most cases, the independent model is slightly better. The exceptions are N-C Asia and Africa. The former shows good fit for the dependent model and the latter slightly prefers the dependent model for GMR dataset (Table 1). The majority of languages in Africa belong to Atlantic-Congo family. If we split Africa into two datasets Atlantic-Congo vs. non-Atlantic-Congo then only the former shows positive fit for the dependent model. In sum, we find no global pattern of correlation between labiodentals and hunter-gatherers if we assume their different proportions per geographic area. Instead, the results from (*2*) appear to be driven primarily by among area correlations, which suggests common causes driving both the number of labiodentals and the probability of hunter gathering rather than a direct effect of hunter gathering on labiodentals.

	Area	GMR	AUTOTYP
1	*Africa	-1.475	0.626
2	Papua	1.222	0.402
3	S America	2.447	1.773
4	W and SW Eurasia	0.001	0.000
5	C America	0.290	0.285
6	S/SE Asia	0.235	0.174
7	N America	0.638	1.036
8	Pacific	0.347	0.003
9	Australia	-0.004	0.001
10	*N-C Asia	-5.620	-2.033
11	*Whole World	-213.174	-18.054
12	*Africa: Atlantic-Congo	-1.922	-0.406
13	Africa: non-Atlantic-Congo	0.571	0.715

Table 1: Fit of Binomial causal models. This table shows Log Bayes Factors, calculated as $BF(M_{ind} - M_{dep})$, for data from GMR and AUTOTYP databases. * means positive or strong support for M_{dep}.

2 Linear Regression of across-area variation in labiodentals and subsistence

2.1 Motivation

Even though, labiodental phonemes and subsistence are independent in 8 out of 10 focal geographic areas, the visual observation of the probability of labiodental phonemes (θ_{Li}) and hunter-gatherers (θ_{Hi}) shows a negative correlation. To test whether this correlation is statistically significant we conduct a series of linear regression analyses.

2.2 Methods

We test three regression models using Log Bayes factor BF for GMR and AUTOTYP datasets. Languages from African and N-C Asia were excluded from this and all

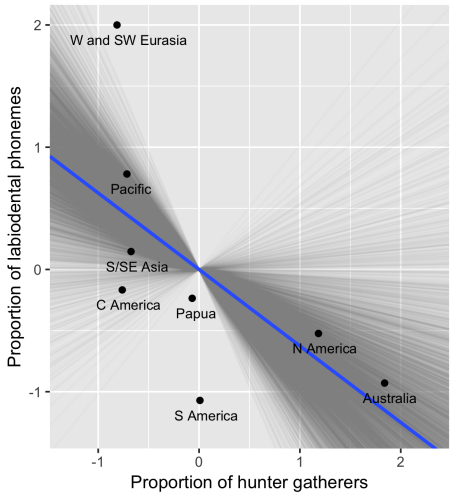

Figure 1: Slope and its posterior distribution of the best-fit regression model R_1 for GMR dataset.

further analyses since they were identified to show dependency between labiodentals and subsistence in the previous analysis, which implies that they have two labiodental proportions – for hunter gatherer and agrarian societies respectively. This precludes them from consistent modeling with languages from other areas that have only one proportion.

The regression models are:

- R_0: $\hat{\theta}_{Li} = \beta_0 + \hat{\theta}_{Hi} + \epsilon$
- R_1: $\hat{\theta}_{Li} = \beta_1 \hat{\theta}_{Hi} + \epsilon$
- R_2: $\hat{\theta}_{Li} = \beta_0 + \beta_1 \hat{\theta}_{Hi} + \epsilon$

Herein, $\hat{\theta}_{Li}$ and $\hat{\theta}_{Hi}$ are normalized per-area empirical probabilities (proportions) of languages with LB and HG societies respectively; they are maximum likelihood estimated of θ_{Li} and θ_{Hi} from the Binomial casual independent model. The empirical probabilities were normalized as $(\theta - mean(\theta))/sd(\theta)$.

2.3 Results

$BF(R_1 - R_0)$ is 3.26 and 2.47 for GMR and AUTOTYP datasets respectively that favors the model R_1 with the slope -0.6 and -0.5 (for GMR and AUTOTYP datasets respectively), thereby supporting negative correlation between $\hat{\theta}_{Li}$ and $\hat{\theta}_{Hi}$ (Fig. 1). Note, the model R_2 was always worse than R_1 given BF.

3 Predictive Posterior Simulations (PPS)

3.1 Motivation

The found negative correlation between θ_H and θ_L may well support the result of (*2*) that change in human diet stimulated the emergence of labiodentals. It is also

known that the number of phonemes declines with a distance from Africa following human dispersion routes due to the serial founder effects (*3*). Thus, we hypothesized that the correlation between θ_H and θ_L might have a similar explanation in which the decrease in the number of labiodental phonemes is causally related to the distance from Africa but not to subsistence, while the correlation between θ_H and θ_L is either a coincidence due to the current distribution of hunter gatherers across the globe or driven by unobserved confounders that drive both LB and HG across geographic space.

To assess these two hypotheses, we perform Bayesian posterior predictive simulations (PPS) for a range of the potential predictors. Our PPS simulations allow assessing the predictors' goodness of fit by simulating per-area probabilities (= proportions) of labiodentals ($\hat{\theta}_{Li}$) from posterior distributions and comparing the simulated and observed per-area probabilities, thereby enabling us to test weather the observed correlation can be the product of the distance.

3.2 Methods

3.2.1 Data

In all linear regression models below, we use a set of similar predictors as in the original paper (*2*). In addition, we add the predictor $distance$ and treat *non-labiodental phonemes* differently. $Distance$ indicates the distance from Africa; it was calculated as suggested in (*3*) but adjusted to match the geographic areas of current datasets (see R scripts). The *Non-labiodental phonemes* predictor indicates *phonemes – (non-labiodental fricatives + labiodentals)*. The full set consists of six predictors: *distance, non-labiodental fricatives, non-labiodental phonemes, subsistence, area, intercept only*. $Pred_j$ denotes one of the six predictors in this set.

3.2.2 Simulations

First, we run six Poisson regressions $Pr(L_n) = Poisson(L_n, \lambda)$, each using one of the six predictors, to sample parameters from the models' posterior distributions. These regressions model the number of labiodentals (L_n) per language n given $Pred_j$; with a lambda parameter that takes the form $log(\lambda) = \beta_0 + \beta_1 Pred_j$.

Given the posterior distribution of each model $p_j(\beta_0, \beta_1|L_n, Pred_j)$, we simulate the predictive posterior distribution (PPD) of labiodentals $\widetilde{L}_n = p_j(\widetilde{L}_n|L_n)$ per language n for each individual model. Next, we convert the PPD of labiodentals to the posterior predictive probabilities (=proportions) of labiodentals $\widetilde{\theta}_{Li}$ per area i as:

$$\widetilde{\theta}_{Li} = \frac{\sum_{n=1,n\in i}^{N_i} I(\widetilde{L}_n)}{N_i}, \tag{1}$$

where $I(\widetilde{L}_n)$ equals 1 if $\widetilde{L}_n > 0$ and equals 0 otherwise; and N_i is the total number of languages in Area i.

The similarity between predictive posterior probability $\widetilde{\theta}_{Li}$ and the observed probability $\hat{\theta}_{Li}$ across all areas is calculated as the average mean squared error μ_{err}:

$$\mu_{err} = \frac{1}{8}\sum_{i=1}^{8} err_i, \tag{2}$$

where err_i is the per-area mean squared error:

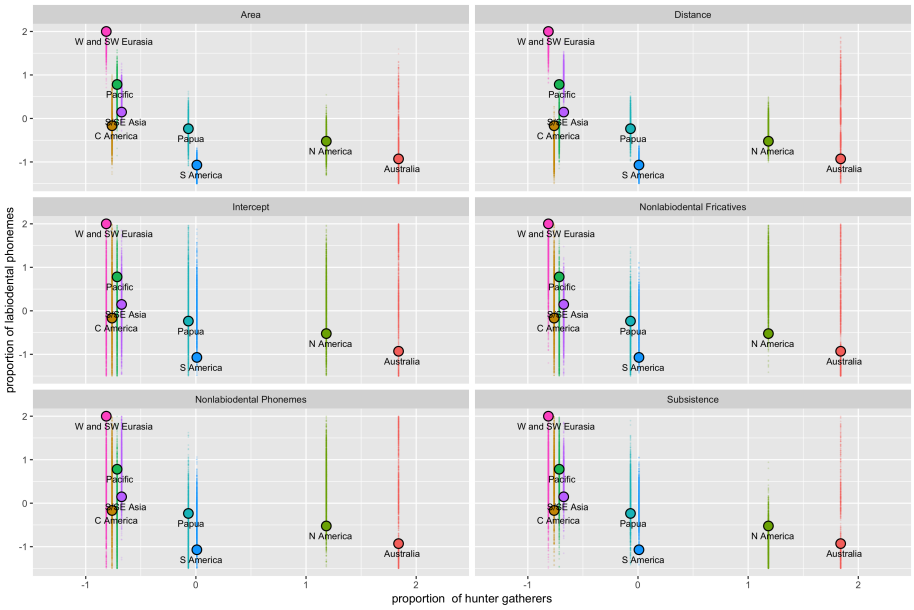

Figure 2: Predictive Posterior Simulations. Observed empirical proportions $\hat{\theta}_{Li}$ (circles) and their posterior predictions $\widetilde{\theta}_{Li}$ (traces) for all six predictors used in PPS simulations.

$$err_i = \frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_{Li} - \widetilde{\theta}_{Li,s})^2, \tag{3}$$

where S is the size of the sample drawn from PPD, and s is the s^th sample.

The average mean squared error μ_{err} is used to assess the goodness of fit of the six alternative predictors – lower value of μ_{err} indicates better fit.

3.3 Results

Mean squared error is lowest for the predictor $area$ (Table 2) that was included as a gold-standard for this analysis since the empirical proportions were estimated by per-area basis prior to running PPS (so, $area$ is the expected best-fit predictor). On the other extreme, mean squared error is the highest for the $intercept$ that bears not relationship of linear dependency. The error for the $distance$ comes second after $area$ and it is significantly smaller than that for the $subsistence$ that indicates a bitter fit.

Fig. 2 demonstrates that data simulated using $distance$ fit the observed data almost as accurately as those simulated using the gold standard $area$. The simulations using $subsistence$ show significantly less accuracy.

4 Poisson Linear Regression (PLR): model comparison

We run Poisson regressions using all combinations of predictors $Pred$ and their second order interactions $Pred$={*Distance, Non-labiodental Fricatives, Non-labiodental*

	Dataset	GMR	AUTOTYP
1	Area	0.143	0.523
2	Distance	0.437	0.824
3	Subsistence	0.861	1.614
4	Nonlabiodental Fricatives	1.045	1.511
5	Nonlabiodental Phonemes	1.444	1.663
6	Intercept	1.724	1.754

Table 2: Average mean squared error μ_{err} from Predictive Posterior Simulations.

Phonemes, Subsistence, Area$\}$. This results in 32767 models which we assess using Akaike information criterion (AIC).

95% of cumulative Akaike weights include 1892 (5.8%) and 4549 (14%) models for GMR and AUTOTYP datasets, respectively. The majority of the best models (GMR=80%, AUTOTYP=83%) include the subsistence predictor only through interactions with other predictors; GMR=7% and AUTOTYP=8% include the the subsistence as a stand-alone predictor, and GMR=6% and AUTOTYP=9% do not include subsistence. dAIC between the best model with and without the subsistence predictor dAIC = AIC(with subsistence)-AIC(without subsistence) is 2.3 (1220.839-1220.167) and -0.98 (309.4692-310.4446) for GMR and AUTOTYP datasets respectively, which in both cases is negligible. All these suggest that adding subsistence as the predictor of labiodentals brings very little additional information is second-order relationships are used.

5 Phylogenetic Analyses

5.1 Model Specification

To test whether phylogeny has branch-specific rates of trait evolution we use a modified version of the Bayesian method proposed by (*4*) and adjust it to the purpose of ancestral character state reconstruction of discrete traits. In this method, branch-specific rates are modeled using Dirichlet process prior (DPP) that assumes that branches of phylogeny are distributed into rate clusters whose number is inferred from data. We perform a joint inference for all ten binary characters (ten labiodental phonetic values) that were usid in (*2*); our modified branch-specific DPP model assumes that all ten characters share the same transition rate matrix Q across phylogeny:

$$Q = \begin{pmatrix} -1 & 1 \\ r & -r \end{pmatrix}.$$

At the same time each tree branch $branch_i$ has its own rate multiplier b_i that is distributed according to DPP: $b_i \sim DPP(\alpha, G_0)$. The probability of seeing character states over $branch_i$ is hence $P(branch_i) = \exp(Qb_i l_i)$, where l_i is the length of $branch_i$. Thus, this implementation allows us to model rate heterogeneity across branches of the tree.

5.2 Inference

We implemented and ran this model in RevBayes (*5*). As in the original paper, we performed inference using two phylogenies of Indo-European languages [named on GitHub as B-dataset (*6*) and C-dataset (*7*)] that included 52 tips. Those phylogenies contain posterior samples of trees; 1000 trees were randomly selected from them for our analyses. The concentration parameter a of DPP was given gamma prior with the expectation of 5 rate categories, while base distribution G_0 was set as $gamma(shape = 4, rate = 8)$; exponential distribution ($\lambda = 5$) was used for the rate parameter r. For each dataset, we ran two runs and assessed the parameter convergence using trace plot; the posterior samples were found converged for both datasets.

5.3 Results

The number of branch-specific rate categories was 1 with the posterior probability 1.0 for both B-dataset and c-dataset.

References

1. B. Carpenter, *et al.*, *Journal of statistical software* **76** (2017).
2. D. E. Blasi, *et al.*, *Science* **363**, eaav3218 (2019).
3. Q. D. Atkinson, *Science* **332**, 346 (2011).
4. T. A. Heath, M. T. Holder, J. P. Huelsenbeck, *Molecular biology and evolution* **29**, 939 (2011).
5. S. Höhna, *et al.*, *Systematic Biology* **65**, 726 (2016).
6. R. Bouckaert, *et al.*, *Science* **337**, 957 (2012).
7. W. Chang, C. Cathcart, D. Hall, A. Garrett, *Language* **91**, 194 (2015).