nature human behaviour

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41562-018-0457-6

In the format provided by the authors and unedited.

The evolution of language families is shaped by the environment beyond neutral drift

Christian Bentz[1,2]*, Dan Dediu[3,4], Annemarie Verkerk[5] and Gerhard Jäger[1,2]

[1]Department of General Linguistics, University of Tübingen, Tübingen, Germany. [2]DFG Center for Advanced Studies: 'Words, Bones, Genes, Tools', University of Tübingen, Tübingen, Germany. [3]Collegium de Lyon, Institut d'Études Avancées, Lyon, France. [4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. [5]Max Planck Institute for Science of Human History, Jena, Germany. *e-mail: chris@christianbentz.de

Supplementary information for: The evolution of language families is shaped by the environment beyond neutral drift

Bentz, C.*[1,2], Dediu, D.[3,4], Verkerk, A.[5] and Jäger, G.[1,2]

[1]Department of General Linguistics, University of Tübingen
[2]DFG Center for Advanced Studies, University of Tübingen
[3]EURIAS fellow, Collegium de Lyon, Institut d'Études Avancées, France
[4]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[5]Max Planck Institute for Science of Human History, Jena, Germany
*corresponding author: chris@christianbentz.de

Contents

Supplementary Results 1: Wilcoxon test results by tree set **3**

Supplementary Results 2: Results for distances to lakes, rivers, and oceans **5**

Supplementary Results 3: Wilcoxon test results and violin plots by language family **6**

Supplementary Results 4: All phylogenetic signals **9**

Supplementary Methods 1: Phylogenetic tree sources **10**
Dediu's forest 10
ML trees 11
Bayesian trees 11

Supplementary Methods 2: Advantages and disadvantages of λ and K **12**

Supplementary Methods 3: Pagel's λ **13**
Example 14

Supplementary Methods 4: Blomberg's K **16**
Example 16

Supplementary Methods 5: Climate and bioclimatic data **20**

Supplementary Methods 6: Distance to water **24**

Supplementary Discussion 1: Problems and caveats **25**

Supplementary Note 1: Correlations between environmental variables **27**

Supplementary Results 1: Wilcoxon test results by tree set

Supplementary Table 1 gives the results of the one-sided Wilcoxon signed rank tests for median values of phylogenetic signals (see Supplementary Data 6 in Guide to SI; filename: `wilcoxonResults.csv`).

There are 42 median phylogenetic signals of subsets by signal metric, tree source, and environmental variable. This corresponds to the facetting of Figure 1 in the main paper (right panels). The results for the 644 subsets by families are given in Supplementary Results 3. N represents the number of trees in the respective subset. The table further gives medians, the upper confidence intervals (CI), as well as p-values of three one-sided Wilcoxon tests with the following null and alternative hypotheses:

- $H_{0.1} \rightarrow$ median phylogenetic signal ≥ 0.1,
  alternative hypothesis: < 0.1;
- $H_{0.9} \rightarrow$ median phylogenetic signal ≥ 0.9,
  alternative hypothesis: < 0.9;
- $H_{1.1} \rightarrow$ median phylogenetic signal ≥ 1.1,
  alternative hypothesis: < 1.1;

To correct for multiple testing, we apply the Bonferroni correction for each hypothesis test ($H_{0.1}$, $H_{0.9}$, $H_{1.1}$) separately. This means that the p-values are multiplied by 42 (tree source analysis) or 644 (language family analysis) respectively. This can yield values > 1. We set these back to 1. Values are further rounded to three digits with R function `round()`.

Using the results of these three statistical tests we can assess the probability of finding the phylogenetic signals that we estimated given the four hypotheses outlined in the main paper, i.e. whether median phylogenetic signals are close to 0.1 (H_0), in between 0.1 and 0.9 (H_{0-1}), around 0.9 to 1.1 (H_1), or higher than 1.1 (H_{1+}). For example, the first row of Supplementary Table 1 gives the median and upper CI of the K value of distance to water on the 5801 Bayesian trees (of 7 different language families), i.e. $\tilde{K} = 0.30$ and $CI = 0.33$. The p-value in the $H_{0.1}$ column is 1, meaning that the probability of finding this data (phylogenetic signal distribution) assuming that the null hypothesis is actually true (i.e. $\tilde{K} \geq 0.1$) is virtually 1. In comparison, the probabilities given the other two null hypotheses, i.e. $\tilde{K} \geq 0.9$, $\tilde{K} \geq 1.1$, are virtually 0. Together these three statistical tests strongly suggest that the actual median value of K for this particular subset and external variable is in between 0.1 and 0.9. The overall percentages of median values in line with the four relevant hypotheses are given in the main paper.

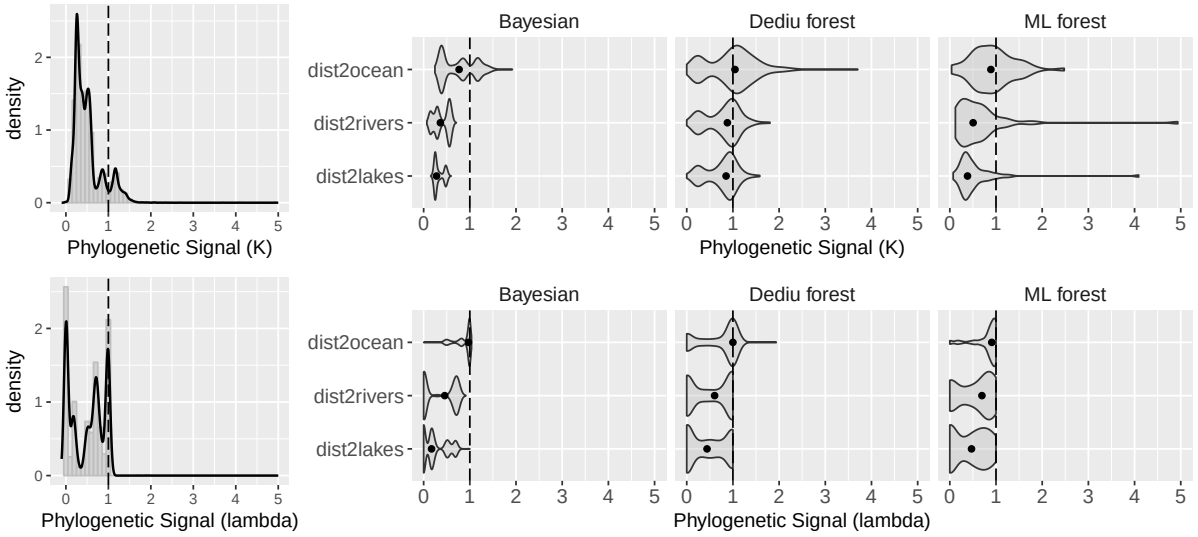
The R code used for these analyses is found in `phyloSignal_pvalues.R` (see R code files in Guide to SI).

Supplementary Table 1: Wilcoxon test results by signal metric (method), tree set, and environmental variable. Median values are separated by method ("K" and "lambda") and ordered from lowest to highest.

Method	Tree.Set	Variable	N	Median	CI	H0.1	H0.9	H1.1
K	Bayesian	dist2water	5801	0.30	0.33	1	0	0
K	ML forest	dist2water	58	0.37	0.53	1	0	0
K	Bayesian	log_popSize	5801	0.38	0.42	1	0	0
K	ML forest	log_popSize	58	0.46	0.58	1	0	0
K	Bayesian	altitude	5801	0.48	0.50	1	0	0
K	ML forest	altitude	58	0.66	0.75	1	0.001	0
K	Bayesian	climate_PC2	5801	0.71	0.72	1	0	0
K	ML forest	climate_PC2	58	0.71	0.80	1	0.019	0
K	Dediu forest	dist2water	351	0.83	0.81	1	0	0
K	Dediu forest	log_popSize	351	0.92	0.91	1	1	0
K	ML forest	climate_PC1	58	0.92	1.09	1	1	1
K	Dediu forest	altitude	351	0.95	0.95	1	1	0
K	Bayesian	climate_PC1	5801	0.96	1.00	1	1	0
K	Dediu forest	climate_PC2	351	0.99	0.97	1	1	0
K	Dediu forest	climate_PC1	351	1.09	1.12	1	1	1
K	Bayesian	longitude	5801	1.13	2.08	1	1	1
K	Dediu forest	latitude	351	1.25	1.30	1	1	1
K	Dediu forest	longitude	351	1.35	1.45	1	1	1
K	ML forest	latitude	58	1.44	1.82	1	1	1
K	ML forest	longitude	58	1.49	2.02	1	1	1
K	Bayesian	latitude	5801	1.77	2.07	1	1	1
lambda	Bayesian	log_popSize	5801	0.21	0.38	1	0	0
lambda	Dediu forest	dist2water	351	0.43	0.50	1	0	0
lambda	ML forest	dist2water	58	0.43	0.49	1	0	0
lambda	Bayesian	dist2water	5801	0.51	0.34	1	0	0
lambda	Bayesian	altitude	5801	0.58	0.50	1	0	0
lambda	ML forest	log_popSize	58	0.58	0.59	1	0	0
lambda	ML forest	altitude	58	0.76	0.82	1	0	0
lambda	Dediu forest	log_popSize	351	0.79	0.63	1	0	0
lambda	ML forest	climate_PC2	58	0.82	0.84	1	0.021	0
lambda	Dediu forest	altitude	351	0.87	0.68	1	0	0
lambda	ML forest	climate_PC1	58	0.94	0.93	1	1	0
lambda	Bayesian	climate_PC2	5801	1.00	0.93	1	1	0
lambda	Bayesian	longitude	5801	1.00	0.99	1	1	0
lambda	Bayesian	latitude	5801	1.00	0.97	1	1	0
lambda	Bayesian	climate_PC1	5801	1.00	0.98	1	1	0
lambda	Dediu forest	longitude	351	1.00	1.00	1	1	0
lambda	Dediu forest	latitude	351	1.00	0.99	1	1	0
lambda	Dediu forest	climate_PC1	351	1.00	0.92	1	1	0
lambda	Dediu forest	climate_PC2	351	1.00	0.76	1	0	0
lambda	ML forest	longitude	58	1.00	1.00	1	1	0
lambda	ML forest	latitude	58	1.00	1.00	1	1	0

Supplementary Results 2: Results for distances to lakes, rivers, and oceans

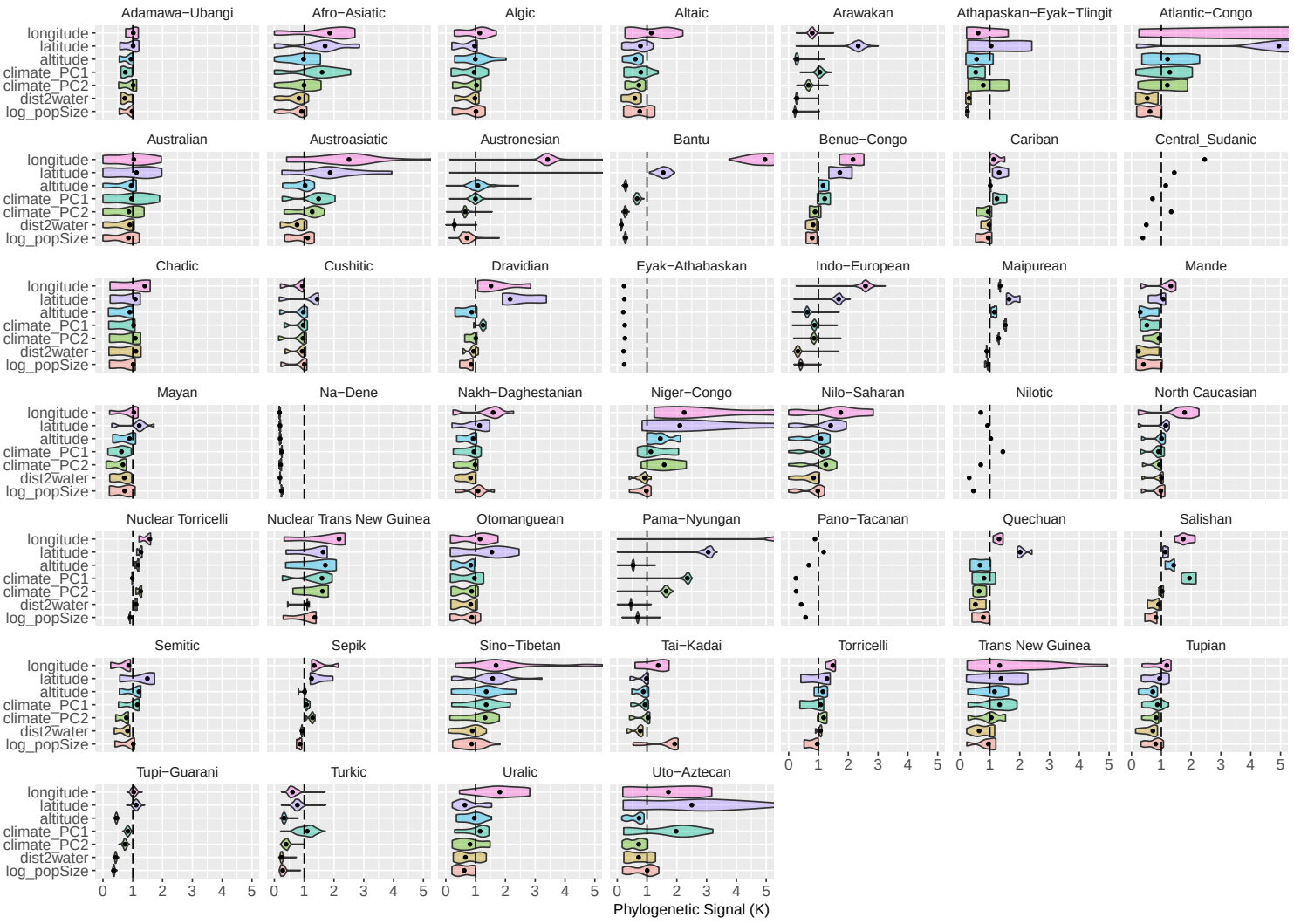
Supplementary Figure 1 gives phylogenetic signal results for distances to lakes, rivers, and oceans separately. Distances to oceans generally have the strongest signals (close to 1.0), followed by distances to rivers, and distances to lakes. The data underlying this plot is found in the file `phyloSignals_water.csv` (Supplementary Data 5 in the Guide to SI). The code to produce this plot is found in `violinPlots_treeSource_3water.R` (see R code files in Guide to SI).



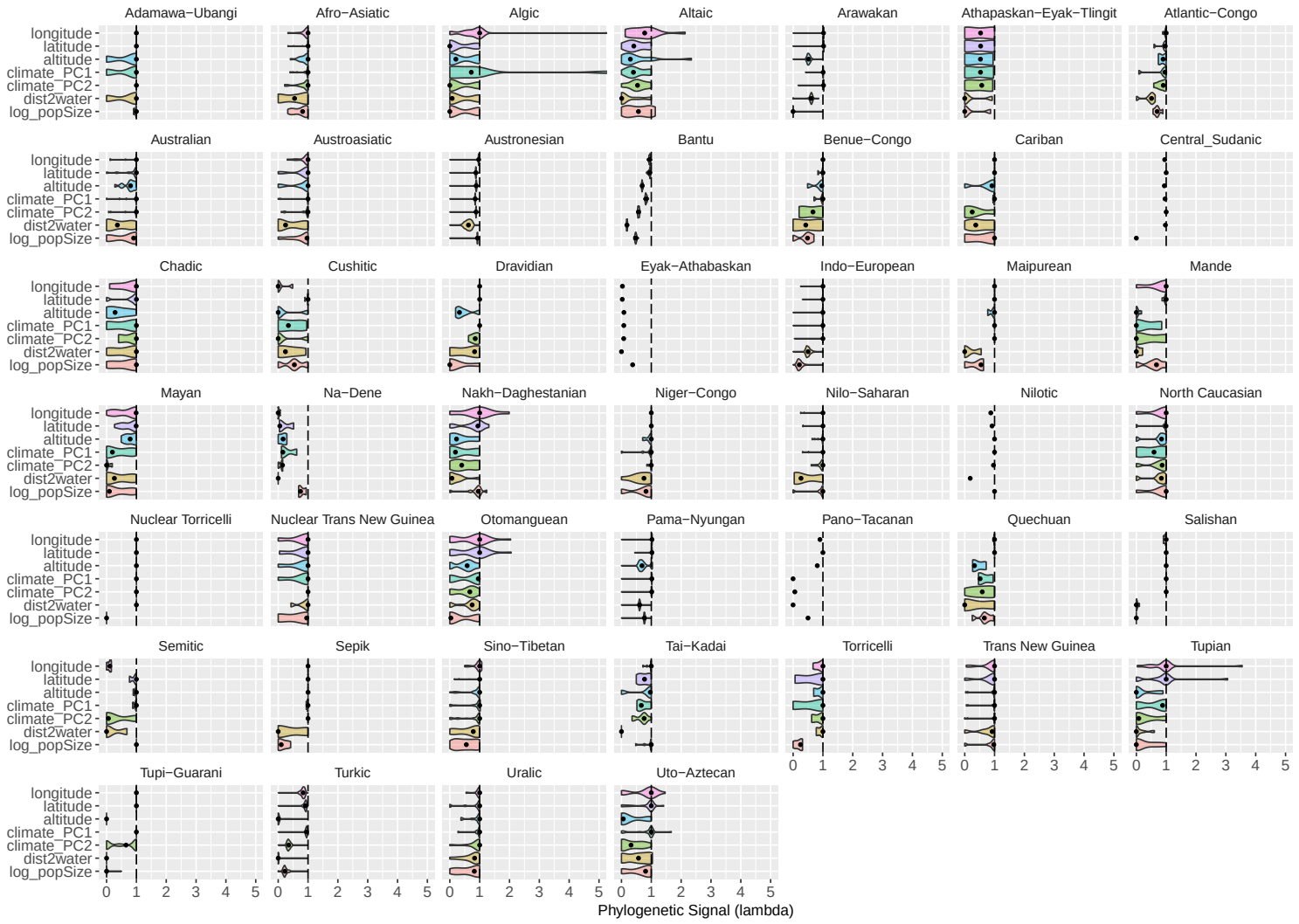
Supplementary Figure 1: Density distributions of phylogenetic signals for K and λ (upper and lower left panel). This includes phylogenetic signals of all three tree sources and three distances to bodies of water (lakes, rivers, oceans). The dashed vertical line indicates the phylogenetic signal value expected under BM along the branches of trees. Violin plots with distributions of Blomberg's K and Pagel's λ per environmental factor are given in the six panels to the upper and lower right. Black dots represent median values. The grey transparent areas are density distributions of phylogenetic signal values. The x-axis is limited to values of a maximum of five. This plot is produced using the `ggplot2` (Wickham and Chang, 2012) R package.

Supplementary Results 3: Wilcoxon test results and violin plots by language family

The code to run one-sided Wilcoxon signed rank tests by families is included in `phyloSignal_pvalues.R` (see R code files in Guide to SI). The results are given in file `wilcoxonResults_families.csv` (Supplementary Data 7 in Guide to SI).
Supplementary Figure 2 and Supplementary Figure 3 give violin plots for K and λ facetted by language families. The code to produce these is in `violinPlots_family.R`.



Supplementary Figure 2: Blomberg's K for environmental factors by language family. Black dots represent median values. The coloured areas are density plots. Only families with > 20 languages are included. The x-axis is limited to values of a maximum of five. The black dashed line indicates the phylogenetic signal value expected under BM, i.e. 1.0.



Supplementary Figure 3: Pagel's λ for environmental factors by language family. Black dots represent median values. The coloured areas are density plots. Only families with > 20 languages are included. The x-axis is limited to values of a maximum of five. The black dashed line indicates the phylogenetic signal value expected under BM, i.e. 1.0.

Supplementary Results 4: All phylogenetic signals

The code in `phyloSignal_v03.R` (see R code files in Guide to SI) is used to calculate phylogenetic signals given a file with Newick trees (e.g. `forestDediu.csv` or `forestML.csv` as found in Supplementary Data 1 and Supplementary Data 2 in the Guide to SI) and a file with environmental information per language (see Supplementary Data 3 in Guide to SI; filename: `glotto3.2_externalData.csv`) as input. Further details of how to use the code are found as comments in the code file. The resulting file with all 86968 phylogenetic signals (86940 without the "world tree") is named `phyloSignals.csv` (see also Supplementary Data 4 in Guide to SI).

Supplementary Methods 1: Phylogenetic tree sources

Dediu's forest

Dediu's forest comprises many thousands of language family trees (see Dediu, 2018, Appendix A). Tree topologies are taken from Ethnologue (Lewis et al., 2013) (147 topologies), WALS (Dryer and Haspelmath, 2013) (214), AUTOTYP (Nichols et al., 2013) (403), and Glottolog (Hammarström et al., 2018) (435). This adds up to 1199 tree topologies.

Branch lengths are added to these by means of methods categorized into three types: a) branch lengths directly reflecting the tree topology (*constant*, *proportional*, and *grafen*), b) both topology *and* branch lengths derived from a distance matrix (i.e. neighbour joining, *nj*), and c) branch lengths derived from a distance matrix and mapped onto a given tree topology by non-linear least squares and a genetic algorithm (*nnls*, *ga*). We exclude the trees with branch lengths derived by methods under a) as these inflate the tree set by adding branch length information predictable by the topology.

Furthermore, the distance matrices used in methods under b) and c) can derive from: vocabulary (*ASJP16*), geographic distances (*great circle* distance), distances based on phonological, grammatical, semantic, and syntactic features from WALS (*gower* and *euclidean* method for distance calculation with and without missing data imputation), distances based on AUTOTYP grammatical features (*gower* method), and distances based on the tree topologies (genetic method, denoted as *mg*). Geographic distances are excluded here, as these would yield tautological analyses when we calculate phylogenetic signals of geographic dimensions. Also, the gower method for WALS features gives similar results to the euclidean distance method. We only use the trees with euclidean branch lengths in order to not unneccesarily inflate the sample with very similar trees. Finally, the genetic method also uses branch lengths based on the tree topology and is also excluded in order to not inflate the tree set.

Taking these considerations into account, we arrive at the following nine branch length methods applied to Ethnologue, AUTOTYP, Glottolog, and WALS topologies: *nj + asjp16*, *nj + autotyp*, *nj + wals(euclidean)*, *nnls + autotyp*, *nnls + asjp16*, *nnls + wals(euclidean)*, *ga + autotyp*, *ga + asjp16*, *ga + wals(euclidean)*.

Note that tree topologies from WALS and AUTOTYP are generally less resolved than topologies from Glottolog and Ethnologue. This is due to conscious decisions to include only firmly established clades. In combination with certain branch length methods (e.g. nnls), these topologies can yield sparsely differentiated phylogenetic trees for certain families. For example, the Algic language family tree based on the WALS topology features Yurok (yur) and Wiyot (wiy) as isolates (within the family) while all the rest of the family (24 languages) form part of the Algonquian clade. Such "underspecified" trees can give rise to extreme phylogenetic signals. The longitude signal for this tree with branch lengths based on ASJP16 is $\lambda = 6$(!). As explained in Supplementary Methods 2, λ is very unlikely to exceed 1.0. However, in this particular case, due to there being only three deep splits in the family tree which are predicted by longitude, λ vastly exceeds 1.0. While such cases of extreme signals exist in our data, the overall results reported in the main paper are not driven by these. Namely, if we remove all trees built on WALS and AUTOTYP topologies, we still get qualitatively similar results for the analyses of median phylogenetic signals.

The Newick trees selected from Dediu's forest are found in `forestDediu.csv` (Sup-

plementary Data 1 in Guide to SI).

ML trees

Maximum likelihood trees are found in file `forestML.csv` (Supplementary Data 2 in Guide to SI). Note, again, that the Newick tree syntax uses commas. These are not supposed to delimit separate columns in the .csv file.

Bayesian trees

Supplementary Table 2 gives information about the sample of Bayesian trees retrieved from published studies. The number of languages included in phylogenetic signal analyses (Lang.) is smaller than the overall number of languages sampled in the respective study (Lang. orig.), as only languages are included for which information about the relevant environmental variables is available. The original sets of posterior trees (Trees orig.) are downsampled to a maximum of 1000 to have similar numbers of posterior trees for statistical analyses. Supplementary Table 3 gives information about tree availability.

Supplementary Table 2: Information on the Bayesian trees.

Family	Lang.*	Lang. orig.†	Trees	Trees orig.	Reference
Arawakan	43	60	1000	9750	Walker and Ribeiro (2011)
Austronesian	333	400	1000	1000	Gray et al. (2009)
Bantu	318	425	100	100	Grollemund et al. (2015)
Indo-European	75	103	1000	12500	Bouckaert et al. (2012)
Pama-Nyungan	151	285	701	701	Bowern and Atkinson (2012)
Tupí-Guaraní	30	32	1000	15000	Michael (2015)
Turkic	25	26	1000	4317	Hruschka et al. (2015)
Total:	**975**	**1331**	**5081**	**43368**	

*Lang.: number of languages included in the phylogenetic signal analyses
† Lang. orig.: languages originally sampled in the respective study

Supplementary Table 3: Information on Bayesian tree availability.

Family	Reference	Cognate data	MCC Tree	Tree sample
Arawakan	Walker and Ribeiro (2011)	supplement	–	upon request
Austronesian	Gray et al. (2009)	github[1]	–	github[1]
Bantu	Grollemund et al. (2015)	supplement	github[2]	upon request
Indo-European	Bouckaert et al. (2012)	supplement	supplement	upon request
Pama-Nyungan	Bowern and Atkinson (2012)	upon request	github[3]	upon request
Tupí-Guaraní	Michael (2015)	upon request	github[4]	upon request
Turkic	Hruschka et al. (2015)	upon request	–	upon request

[1] https://github.com/D-PLACE/dplace-data/tree/master/phylogenies/gray_et_al2009
[2] https://github.com/D-PLACE/dplace-data/tree/master/phylogenies/grollemund_et_al2015
[3] https://github.com/D-PLACE/dplace-data/tree/master/phylogenies/bowern_and_atkinson2012
[4] https://github.com/D-PLACE/dplace-data/tree/master/phylogenies/michael_et_al2015

Supplementary Methods 2: Advantages and disadvantages of λ and K

A recent study by Münkemüller et al. (2012) assessed the statistical properties of phylogenetic signal measures based on simulated data. They test Abouheif's C_{mean}, Moran's I, Pagel's λ, and Blomberg's K for dependence on the size of phylogenies, the resolution of tree structure, branch length information, and the evolutionary models chosen. The findings most relevant for the current study are:

1. Only λ and K are valid metrics for a quantitative comparison across different phylogenies, the values of Abouheif's C_{mean} and Moran's I are not comparable across different trees.

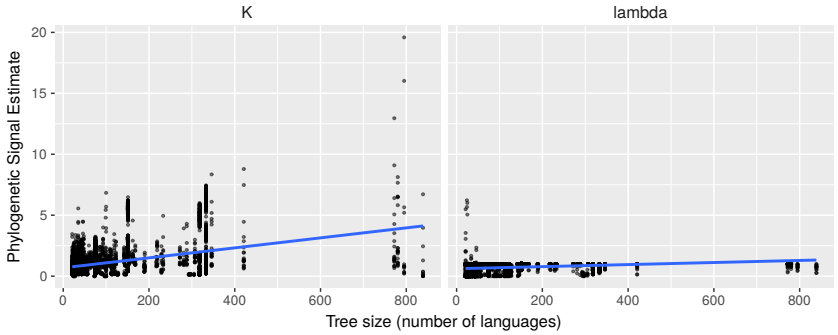
2. The number of species (languages in our case) has an impact on phylogenetic signal estimates – except for Pagel's λ. The mean value of λ does not respond to increasing size of phylogenies, while the mean value of K does (to a small extent). However, we have tested this for our tree samples and the results are given in Supplementary Figure 4. There are small to moderate Pearson correlations between tree size (number of languages) and phylogenetic signal estimates for both λ and K.

3. The number of species has an impact mainly on the variance of estimations. With λ performing the poorest for small phylogenies ($n = 20$). Variation for different simulation runs becomes small once sizes of $n = 50$ are used.

4. The effects of *polytomies* (more than two branches descending from a node) are negligible for both λ and K.

5. Missing branch lengths do influence K, but not λ.

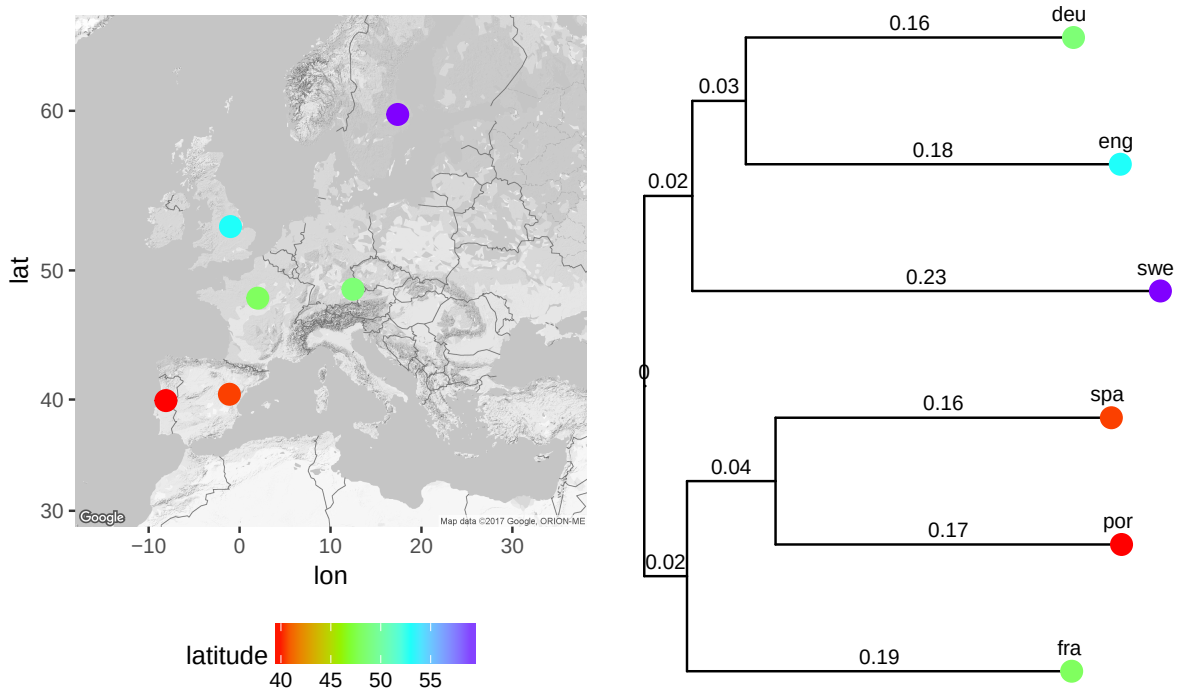
6. Blomberg's K is most sensitive to changes in the underlying evolutionary model. It outperforms the other indices in identifying subtle tree transformations such as random walks with central tendencies in an OU model, or slow-downs and speed-ups in rate of evolution.



Supplementary Figure 4: Relationship between phylogenetic signal estimates and tree size as number of languages for K (left panel) and λ (right panel). Linear regression lines are overlaid. The Pearson correlation is $r = 0.37$ for K, and $r = 0.26$ for λ. The code is found in `phyloSignal_treeSize.R` (see R code files in Guide to SI).

Supplementary Methods 3: Pagel's λ

Assume a phylogenetic tree of six Indo-European languages as given in Supplementary Figure 5. Splits in this tree represent the divergence of two (or more) daughter varieties from a common ancestral language. The bare topology of this tree corresponds to genealogical subgroupings. The upper clade represents the Germanic genus, the lower clade the Romance genus. The branch lengths reflect the estimated amount of language change since the last common ancestor.



Supplementary Figure 5: Six Indo-European languages (deu: German, eng: English, swe: Swedish, spa: Spanish, por: Portuguese, fra: French) of the Germanic and Romance genera on a schematic tree. Colours indicate latitudes (purple: high latitudes, red: low latitudes). Plots are produced using the `ggtree` (Yu et al., 2016) and `ggmap` (Kahle and Wickham, 2013) packages.

The value of the environmental variable, e.g. the approximate latitude where a language is spoken, is represented in Supplementary Figure 5 by the colour of tips. All languages belonging to the Germanic subbranch are approximated to latitudes above 48 degrees, whereas all Romance languages fall below. In this example, the latitude where a language is spoken predicts the genus it belongs to. There are two basic scenarios of how latitude values might have changed over time:

1. Latitude changed completely independently of the structure of the tree. It is simply predicted by the original value plus the summed lengths of branches leading to the language (proportional to time elapsed) multiplied with some random noise. This is a constant rate Brownian motion (BM) process.

2. The latitude changed by means of BM *along* the tree, i.e they reflect shared history (implying shared structural features) of languages. In this case, trait values will

not be phylogenetically independent. To predict the trait values, we have to take the shared history of languages into account.

In order to evaluate whether the first or the second scenario is more likely given the phylogenetic tree and the tip values, a parameter λ is introduced (Freckleton et al., 2002). λ is a multiplier for the off-diagonal elements of variance-covariance matrix, which reflects branch lengths from root to tip for individual languages as well as shared branch lengths for pairs of languages. In the case of $\lambda = 0$ all shared branch lengths (expected covariances) are converted to zero, which is equivalent to the first scenario of BM without any phylogenetic information. Conversely, in the case of $\lambda = 1$, shared branch lengths are left unchanged, in accordance with a BM model that takes phylogeny into account. Freckleton et al. (2002) show that a maximum likelihood approach can be used to find the value for λ that results in the highest likelihood for the actual trait values observed on the tips of the tree.

Example

In the simplified example of Supplementary Figure 5, latitudes are indicated by color. The actual values are: $\text{latitude}_{deu} = 48.65$, $\text{latitude}_{eng} = 53$, $\text{latitude}_{swe} = 59.8$, $\text{latitude}_{spa} = 40.44$, $\text{latitude}_{por} = 39.91$, $\text{latitude}_{fra} = 48$. The constant random walk BM model to estimate a trait value y of a language i is given in Freckleton et al. (2002, p. 713) as

$$y_i = \alpha + \sum_{j=1}^{T_i} \epsilon_{i,j} t_{i,j}, \tag{1}$$

where α is the original trait value, $\epsilon_{i,j}$ is random noise for language i and branch j with mean 0 and variance σ^2, $t_{i,j}$ is the length of branch j of language i, and the summation runs over all branches leading to a language T_i. Under the assumption that languages evolve independently of each other, their trait values will follow a multinormal probability density distribution specified in Freckleton et al. (2002, p. 713). However, to take shared ancestry into account we have to incorporate covariances between pairs of languages y_i and y_j:

$$cov(y_i, y_j) = \sigma^2 t_a, \tag{2}$$

where t_a is the length of branches up to the last common ancestor. Under a BM model the variances and covariances of languages are expected to be proportional to the lengths of branches leading to the tip, and the last common ancestor respectively. For example, the expected variance for German is $0.02+0.03+0.16 = 0.21$, and the expected covariance with English $0.02 + 0.03 = 0.05$. Hence, we need to construct a 6×6 variance-covariance matrix for languages given below (from left to right: deu, eng, swe, spa, por, fra).

$$\mathbf{V} = \begin{pmatrix} 0.21 & 0.05 & 0.02 & 0 & 0 & 0 \\ 0.05 & 0.23 & 0.02 & 0 & 0 & 0 \\ 0.02 & 0.02 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.22 & 0.06 & 0.02 \\ 0 & 0 & 0 & 0.06 & 0.23 & 0.02 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0.21 \end{pmatrix}$$

Diagonal elements are the distances from root to tip per language, and all the off-diagonal elements are lengths of shared branches between languages. The variance-covariance matrix $\mathbf{V}$ can be integrated in the multinormal probability density distribution to account for common ancestry (see Freckleton et al., 2002, p. 714 for the actual equation).

In the next step, we can use maximum likelihood estimation to assess whether the actual values observed on the tips (6 latitude values) are more likely to occur under the independent BM model, or under the BM model integrating $\mathbf{V}$. However, instead of just getting likelihoods for these two alternatives, Pagel's λ is introduced as a multiplier for the off-diagonal elements of $\mathbf{V}$. $\lambda = 0$ is equivalent to the independent BM model, and $\lambda = 1$ to the BM model with covariances taken into account. The values between 0.0 and 1.0 assess "how much" phylogeny is actually needed to render the tip values likely.

Values of $\lambda > 1$ are also possible. In this case, to render the latitude tip values most likely, the expected covariances between languages are bigger than the ones actually observed on the tree. However, there is an important restriction: The off-diagonal elements cannot become bigger than the diagonal elements. This restriction derives from the logical fact that a language cannot share more history with another language than with itself. For example, assume we would consider $\lambda = 5$ as a value to transform $\mathbf{V}$. The expected covariance between German and English would then become $0.05 \times 5 = 0.25$, which is bigger than the expected variance for either German or English. Moreover, remember that the same λ is multiplied with all covariances. Hence, if in a given family there are two languages which have recently split from a common ancestor, and share a high relative covariance, λ will be capped at approximately 1.0.

Supplementary Methods 4: Blomberg's K

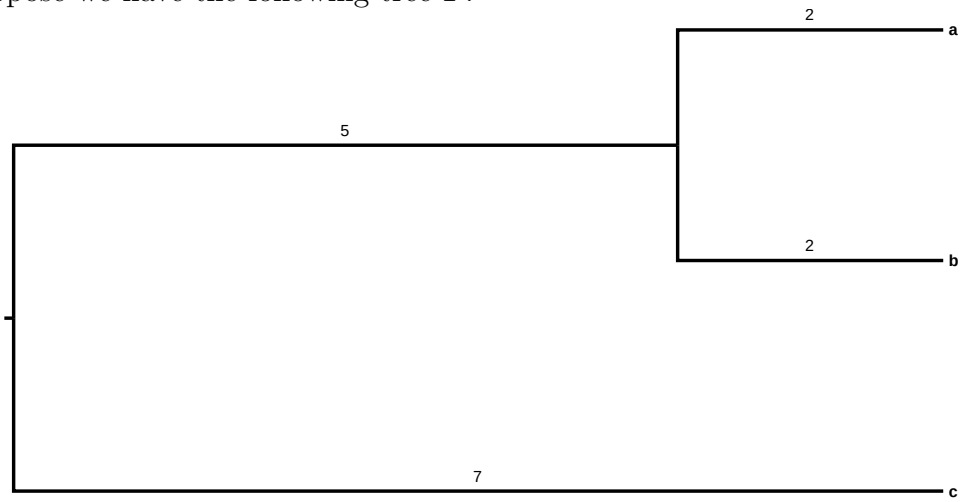
K is a descriptive statistic following a similar rationale as Pagel's λ, i.e. inference of phylogenetic signal based on variance-covariance matrices (Blomberg et al., 2003). First, the mean squared error for the empirical tip values as compared to the "phylogenetically corrected mean", also called *phylogenetically weighted mean* based on independent contrasts (see Felsenstein, 1985 and Symonds and Blomberg, 2014), is calculated. This is denoted MSE_0. Secondly, the mean squared error of tip values – based on the tree topology and branch lengths – is calculated under the assumption of Brownian motion. This is denoted MSE.

The ratio $\frac{MSE_0}{MSE}$ is indicative of how well the emiprical tip values fit the tree under Brownian motion, and lends itself as another measure of phylogenetic signal. If the tree topology and branch lengths accurately predict the empirical trait values, then MSE is relatively small, and the ratio big. However, Blomberg et al. (2003) point out that the ratio shows a dependence on tree size and shape, and is not directly applicable as a phylogenetic signal measure across different trees. Instead, they suggest to use the observed ratio scaled by the ratio expected under a BM model. Thus, they define the K statistic as

$$K = observed\frac{MSE_0}{MSE}/expected\frac{MSE_0}{MSE}. \tag{3}$$

Example

Suppose we have the following tree T:



Furthermore, suppose for some continuous character $\mathbf{X}$ the tip values are

$$\begin{aligned} x_a &= 2 \\ x_b &= 4 \\ x_c &= -10 \end{aligned}$$

Assuming a BM model of character evolution, the expected variance-covariance matrix of the tip values is then

$$\mathbf{V} = \begin{pmatrix} 7 & 5 & 0 \\ 5 & 7 & 0 \\ 0 & 0 & 7 \end{pmatrix}$$

With the covariance between two tips equaling the length of the branch from the root to the last common ancestor of these tips. Computing Blomberg's K for $\mathbf{X}$ and T requires the following steps:

1. Compute the *phylogenetically correct mean* $\hat{x}$ of $\mathbf{X}$ given T.
2. Compute the *mean squared error* MSE_0 of $\mathbf{X}$ in Euclidean space, i.e. under the assumption that $\mathbf{X}$ contains no phylogenetic signal.
3. Compute the *mean squared error* MSE of $\mathbf{X}$ under the assumption that $\mathbf{X}$ evolved along the branches of T according to the BM model. Here the error is defined as the Mahalanobis distance between the components of $\mathbf{X}$ and $\hat{x}$ given $\mathbf{V}$.
4. Compute the expected value $E[{}^{MSE_0}/_{MSE}]$ given $\mathbf{V}$.
5. $K = ({}^{MSE_0}/_{MSE})/E[{}^{MSE_0}/_{MSE}]$.

The phylogenetic mean $\hat{x}$ is calculated according to the generalized least squares method:

$$\hat{x} = \frac{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{X}}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}},$$

where $\mathbf{1}$ is a column vector of length n (n being the number of tips) consisting of 1s. In our example, we have

$$\begin{aligned}
\mathbf{V}^{-1} &= \begin{pmatrix} {}^{7}/_{24} & {}^{-5}/_{24} & 0 \\ {}^{-5}/_{24} & {}^{7}/_{24} & 0 \\ 0 & 0 & {}^{1}/_{7} \end{pmatrix} \\
\hat{x} &= \frac{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{X}}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}} \\
&= \frac{{}^{-13}/_{14}}{{}^{13}/_{42}} \\
&= -3
\end{aligned}$$

Therefore we have

$$\begin{aligned}
MSE_0 &= \frac{(\mathbf{X}-\hat{x})^T(\mathbf{X}-\hat{x})}{2} \\
&= 61.5
\end{aligned}$$

The mean squared error MSE under the assumption of T and a BM model is defined as:

$$MSE = \frac{(\mathbf{X}-\hat{x})^T V^{-1}(\mathbf{X}-\hat{x})}{2}$$

In our example, we have $MSE = 7$. The fact that MSE is substantially smaller than MSE_0 indicates that X carries a phylogenetic signal.

The empirical value of ${}^{MSE_0}/_{MSE}$ is

$$\begin{aligned} {}^{MSE_0}/_{MSE} &= \frac{61.5}{7} \\ &\approx 8.79 \end{aligned}$$

The expected value of ${}^{MSE_0}/_{MSE}$ is calculated as

$$E[{}^{MSE_0}/_{MSE}] = \frac{\text{trace}(V) - \frac{n}{\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1}}}{n-1}$$

For the example, this means

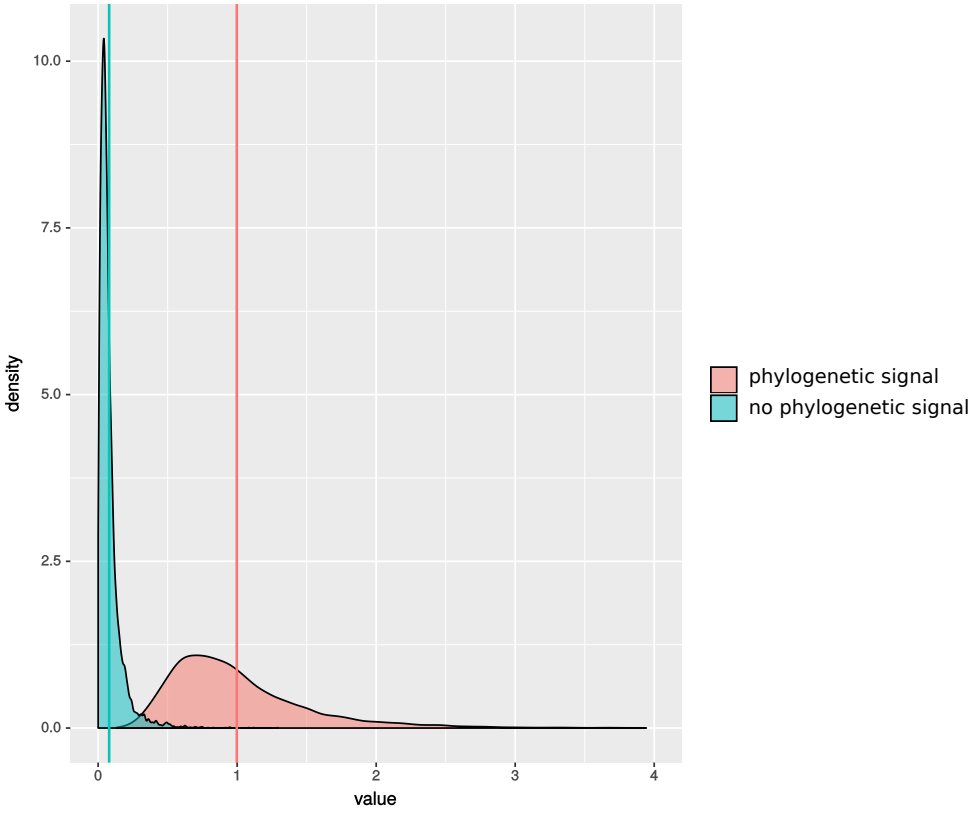
$$\begin{aligned} E[{}^{MSE_0}/_{MSE}] &= \frac{21 - \frac{3}{13/42}}{2} \\ &\approx 5.65 \end{aligned}$$

Therefore

$$\begin{aligned} K &\approx \frac{8.79}{5.65} \\ &\approx 1.55 \end{aligned}$$

Simulation To further illustrate the distributional properties of K for scenarios with and without phylogenetic signal, we generated 100,000 random phylogenies using the function `rphylo()` from the R-package `ape` (Paradis et al., 2004). For each phylogeny, two characters were simulated, one according to a BM model given the phylogeny, and one according to a standard normal distribution independent of the phylogeny. For both characters, Blomberg's K was calculated using the function `phylosig()` from `ape`. Supplementary Figure 6 shows the distribution of K-values obtained this way.

The mean value of K under BM is exactly 1.0. Thus, we expect this mean value for distributions of K-signals if BM is the underlying evolutionary mechanism. Importantly, the mean value of a model without any phylogenetic information is not exactly zero, but rather around 0.1 (0.08 more precisely). We thus cannot expect mean values to be exactly zero when calculating signals for actual language family trees and environmental variables. We take this into account in our statistical analyses by testing whether mean values are significantly higher than 0.1 instead of zero.



Supplementary Figure 6: Density distribution of Blomberg's K with (red) and without (blue) phylogenetic signal of the values on the tips of a given tree. Vertical lines indicate mean values (0.08 and 1.0 respectively). The median values of the distributions are 0.06 and 0.89.

Supplementary Methods 5: Climate and bioclimatic data

We downloaded the *WorldClim* (Version 1.4, March 2018) data (Hijmans et al., 2005), freely available online (http://www.worldclim.org/version1). This consists of raster layers of several variables at different resolutions (for details please see Hijmans et al. 2005 and http://www.worldclim.org/formats1). More precisely, we used R's (R Core Team, 2017) `raster` library (Hijmans, 2017) to automatically download and import the 19 "bio" variables at the 5 minutes resolution ("10 km grids").

These variables are derived from the primary monthly temperature and rainfall variables and are biologically more meaningful (see detailed description at http://www.worldclim.org/bioclim):

- *BIO1* = Annual Mean Temperature
- *BIO2* = Mean Diurnal Range (Mean of monthly (max temp - min temp))
- *BIO3* = Isothermality (BIO2/BIO7) (* 100)
- *BIO4* = Temperature Seasonality (standard deviation *100)
- *BIO5* = Max Temperature of Warmest Month
- *BIO6* = Min Temperature of Coldest Month
- *BIO7* = Temperature Annual Range (BIO5-BIO6)
- *BIO8* = Mean Temperature of Wettest Quarter
- *BIO9* = Mean Temperature of Driest Quarter
- *BIO10* = Mean Temperature of Warmest Quarter
- *BIO11* = Mean Temperature of Coldest Quarter
- *BIO12* = Annual Precipitation
- *BIO13* = Precipitation of Wettest Month
- *BIO14* = Precipitation of Driest Month
- *BIO15* = Precipitation Seasonality (Coefficient of Variation)
- *BIO16* = Precipitation of Wettest Quarter
- *BIO17* = Precipitation of Driest Quarter
- *BIO18* = Precipitation of Warmest Quarter
- *BIO19* = Precipitation of Coldest Quarter

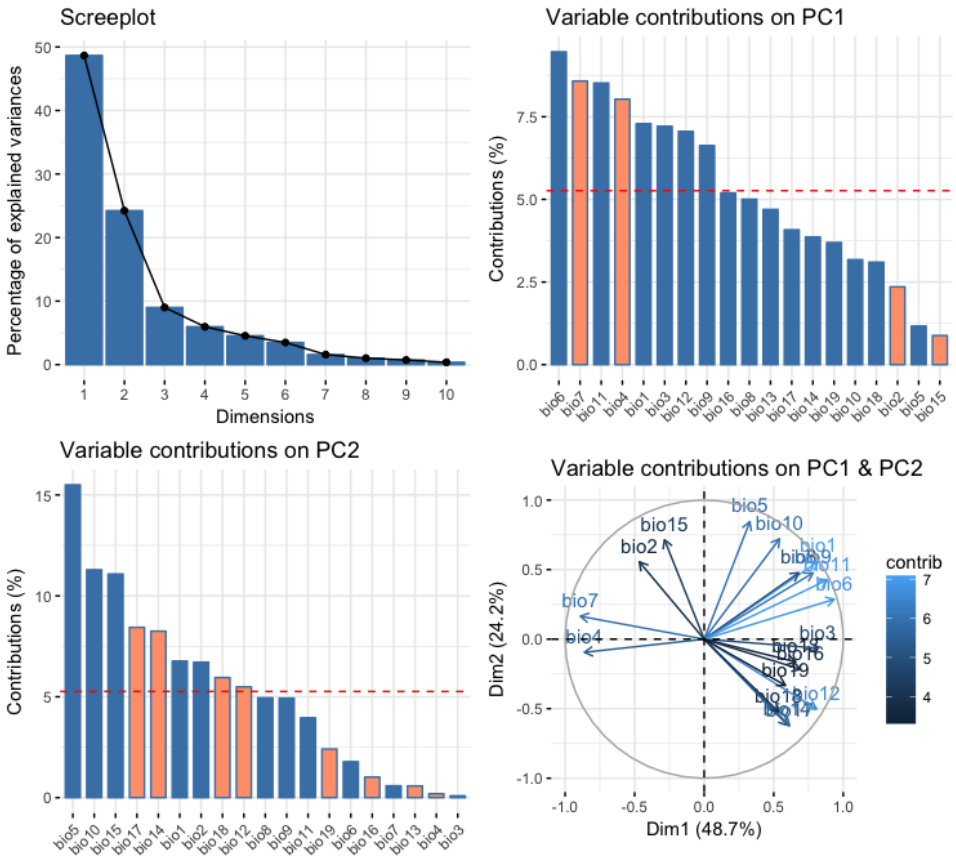
We further retained these data only for the languages with non-missing geographic coordinates in Glottolog 3.2 (freely available at http://glottolog.org/meta/downloads), resulting in 7,913 datapoints.

The inspection of the Pearson correlation matrix (uncorrected for the occurrence of spatial non-independence) between the 19 "bio" variables suggests that there are interesting patterns of correlation, prompting us to conduct a Principle Component Analysis (PCA). This suggests that the first two PCs explain together 72.9% of the variance, and represent meaningful dimensions of the climatic data (see Supplementary Table 4 and Supplementary Figure 7).

The source code is given in the `climate.Rmd` R markdown file (see R code files in Guide to SI).

Supplementary Table 4: First two PCs with their percentage of explained variance, interpretation of high and low values, and attempted overall interpretation.

PC	**% var**	**Loadings**	**High values**	**Low values**	**Interpretation**
PC1	48.7%	+(Min Temperature of Coldest Month) +(Mean Temperature of Coldest Quarter) +(Mean Temperature of Driest Quarter) +(Annual Mean Temperature)	High average and lowest temperatures	Low average and lowest temperatures	**Hot and stable (wetter) climates** ("tropical") vs
		-(Temperature Annual Range) -(Temperature Seasonality) +(Isothermality)	Low variation in temperatures	High variability in temperatures	**Cold and variable (drier) climates** ("temperate"/"cold")
		+(Annual Precipitation)	High precipitation	Low precipitation	
PC2	24.2%	+(Max Temperature of Warmest Month) +(Mean Temperature of Warmest Quarter) +(Mean Diurnal Temperature Range) +(Annual Mean Temperature)	High average and max temperatures, large diurnal temperature range	Low average and max temperatures, low diurnal temperature range	**Hot, dry and variable** ("desert") vs
		-(Precipitation of Driest Quarter) -(Precipitation of Driest Month) -(Precipitation of Warmest Quarter) +(Precipitation Seasonality) -(Annual Precipitation)	Low precipitation and high seasonality	High precipitation and low seasonality	**Cold, wet and stable** ("oceanic")



Supplementary Figure 7: PCA of the "bio" variables. Top-left: scree plot showing the percentage of explained variance by each PC. Top-right and bottom-left: loading of each variable on PC1 and PC2 respectively; blue represents positive and orange negative loadings; dashed red horizontal line is the expected value of the contribution if they were uniform and may be considered as a limit of contribution importance. Bottom-right shows the loadings on both PC1 and PC2 with contribution represented by the shade of blue.

Supplementary Methods 6: Distance to water

We downloaded the *OpenStreetMaps'* "Reduced waterbodies as raster masks" data, available online at http://openstreetmapdata.com/data/water-reduced-raster under the Open Data Commons Open Database License (ODbL; see http://openstreetmapdata.com/info/license), coming as zip archives, each containing several raster files (in GeoTIFF format) at different zoom levels for:

- "ocean" (coastline):
  http://data.openstreetmapdata.com/ocean-raster-reduced-3857.zip
- "lakes" (lakes and other bodies of standing water):
  http://data.openstreetmapdata.com/lakes-raster-reduced-3857.zip
- "rivers" (river and artificial canal areas):
  http://data.openstreetmapdata.com/river-raster-reduced-3857.zip.

We used zoom level 2 for the oceans and 4 for the lakes and rivers, as these represent good trade-offs between precision and computational costs: given that oceans are rather big, using a low zoom does not affect the distances to them too much but it reduces the computational costs drastically, while for lakes and rivers a higher zoom is manageable and captures smaller (but not too small) bodies of water.

Thus, we extracted three raster files (ocean:2, lakes:4 and rivers:4), and for each, we first projected them into the WGS84 datum (from their original Mercator projection) using R's (R Core Team, 2017) `raster` (Hijmans, 2017) package, followed by the extraction of the coordinates of the grid cells marked as "water". Finally, we computed the shortest distance on the WGS84 ellipsoid between the coordinates of the 7,913 languages with non-missing geographic information from Glottolog 3.2 (freely available at http://glottolog.org/meta/downloads), and these "water" cells using the function `geoDist()` in package `geosphere` (Hijmans, 2016).

This process resulted, for each of the 7,913 languages, in three distances to the nearest ocean ("dist2ocean"), lake ("dist2lakes"), and river ("dist2rivers"), to which we also added a general distance to the closest body of water ("dist2water") computed as the minimum of the three distances.

Please note that the zoom level used, especially for lakes and rivers, has the positive side effect of implicitly selecting bodies of water above a certain size. This explains, for example, why some islands are very far from lakes or rivers, as they probably do not support ones large enough to be captured by the raster's resolution level.

The source code is given in the `dist2water.Rmd` R markdown file (see R code files in Guide to SI).

Supplementary Discussion 1: Problems and caveats

Bias through error Imprecisions in linguistic tree topologies, branch lengths, and in mean tip values add noise to phylogenetic signal estimations. Note that a) Münkemüller et al. (2012) illustrate that phylogenetic signal metrics are fairly robust to changes in branch lengths, which often constitute the biggest source of variance between trees, and b) noise will generally lead to lower phylogenetic signals, rather than higher ones (Blomberg et al., 2003). Hence, noise cannot explain findings of high values. Though it could explain systematic differences between signals for latitudes and longitudes, on one hand, and signals for altitude and population size, on the other hand. More noise in altitude and population size data compared to latitude and longitude data could drive systematic differences.

Also, shortened branches towards the root of a tree can inflate phylogenetic signal as measured with K (Revell et al., 2008). It is possible that the general lack of historical linguistic data might lead to less certainty about branch lengths (i.e. language change) towards the roots of trees. These could then be systematically shortened, leading to an upward bias for K and λ.

However, we currently have no clear evidence for either of these biases in our data. We use state-of-the-art trees from different sources to ensure that our results extrapolate across the currently available phylogenetic trees. In future studies, our methodology can be applied to replicate our results with any newly available tree sets.

Variation within languages Phylogenetic signal does not account for variation within species (Kamilar and Cooper, 2013), i.e. languages in our case. Clearly, the same language can be spoken across a fairly wide range of latitudes, longitudes and altitudes, which is especially the case for global languages like Chinese or English. This variation in the environmental factors is not taken into account here. Even global languages are assigned single longitude, latitude and altitude points. This might seem overly coarse, but note that: a) even global languages like Standard English were spoken in a much more confined area when they evolved from their last common ancestor with their closest relatives, and these areas of origin are – on a global scale – are reasonably close to the current location of the standard language given in Glottolog; b) the biggest part of languages in our environmental language data set (83%) is spoken by less than 100,000 thousand speakers, with 42% being spoken by less than 10,000 speakers. Such smaller languages are also fairly limited in terms of geographic spread.

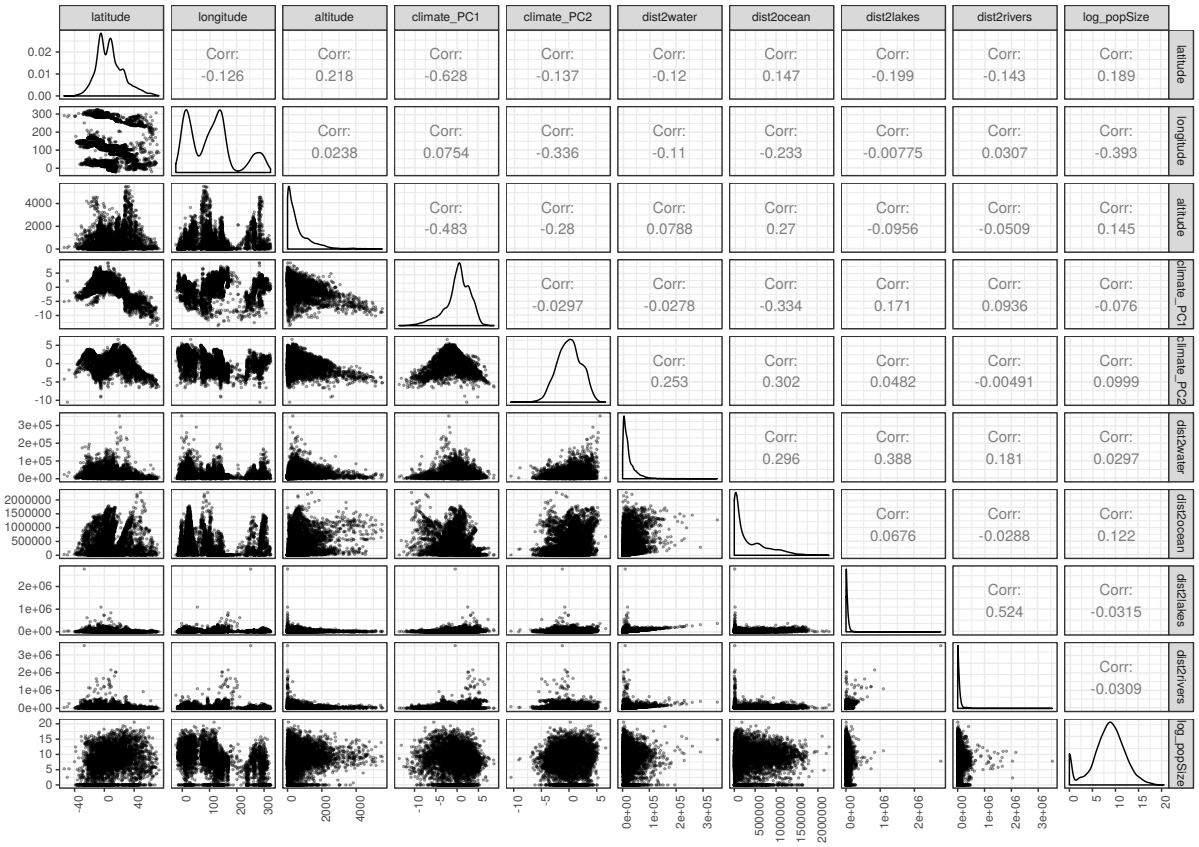
Impact of different tree types We used trees built on different topology and branch length sources. We have not discussed here if there are systematic differences in phylogenetic signals linked to particular accounts of building linguistic family trees, e.g. lexical information versus morphological or syntactic information. This is certainly an important avenue for furture research. On the upside, the main results discussed in this article are robust in the sense that all three sets of trees and tree-building methods yield approximatly the same outcome.

Beyond simple measures of geography and demography Of course, estimating phylogenetic signals of latitudes, longitudes, altitudes, climatic variables, distances to water, and population sizes on language family trees is only a first step towards understanding linguistic diversification. We need to also take into account differences in the

mechanisms by which language families spread: language shift, demographic expansion, and migration (Nichols, 1992, p.372). There are probably no cases in which an entire language family spread through either one of these mechanisms alone. Another aspect to be taken into account are cultural differences between language families, especially the acquisition of technologies that would impact the way language communities could spread, including farming (Diamond and Bellwood, 2003), the domestication of the horse (Clutton-Brock, 1992), and the invention of the outrigger canoe (Gray et al., 2009), see also Richerson et al. (2009).

Supplementary Note 1: Correlations between environmental variables

Correlations between all environmental variables considered here are given in Supplementary Figure 8. Moderate to high correlations are found for climate_PC1 and latitude ($r = -0.63$), climate_PC1 and altitude ($r = -0.48$), and between distance to lakes and distance to rivers ($r = 0.52$). The code for this plot can be found in `scatterPlots_corEnvVar.R` (see R code files in Guide to SI).



Supplementary Figure 8: Correlations between environmental variables. Lower left panels give scatterplots, upper right panels give Pearson correlations. This figure is produced using the package `GGally` (Schloerke et al., 2016).

References

Blomberg, S. P., Garland Jr, T., Ives, A. R., and Crespi, B. (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57(4):717–745.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337:957–960.

Bowern, C. and Atkinson, Q. (2012). Computational phylogenetics and the internal structure of pama-nyungan. *Language*, 88(4):817–845.

Clutton-Brock, J. (1992). *Horse power: a history of the horse and the donkey in human societies.* Harvard University Press, Cambridge: MA.

Dediu, D. (2018). Making Genealogical Language Classifications Available for Phylogenetic Analysis: Newick Trees, Unified Identifiers, and Branch Length. *Language Dynamics and Change*, 8.1.

Diamond, J. and Bellwood, P. (2003). Farmers and their languages: the first expansions. *Science*, 300(5619):597–603.

Dryer, M. S. and Haspelmath, M., editors (2013). *World Atlas of Language Structures online.* Max Planck Digital Library, Munich.

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15.

Freckleton, R. P., Harvey, P. H., and Pagel, M. (2002). Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist*, 160(2):712–726.

Gray, R. D., Drummond, A. J., and Greenhill, S. J. (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, 323:479–483.

Grollemund, R., Branford, S., Bostoen, K., Meade, A., Venditti, C., and Pagel, M. (2015). Bantu expansion shows that habitat alters the route and pace of human dispersals. *Proceedings of the National Academy of Sciences*, 112(43):13296–13301.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S., editors (2018). *Glottolog 3.2.* Jena: Max Planck Institute for the Science of Human History.

Hijmans, R. J. (2016). *geosphere: Spherical Trigonometry.* R package version 1.5-7.

Hijmans, R. J. (2017). *raster: Geographic Data Analysis and Modeling.* R package version 2.6-7.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15):1965–1978.

Hruschka, D. J., Branford, S., Smith, E. D., Wilkins, J., Meade, A., Pagel, M., and Bhattacharya, T. (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, 1(25):1–9.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

Kamilar, J. M. and Cooper, N. (2013). Phylogenetic signal in primate behaviour, ecology and life history. *Phil. Trans. R. Soc. B*, 368(1618):20120341.

Lewis, M. P., Simons, G. F., and Fenning, C. D., editors (2013). *Ethnologue: Languages of the world*. SIL International, Dallas, Texas, 17th edition.

Michael, L. D. (2015). A bayesian phylogenetic classification of tupí-guaraní. *LIAMES*, 15.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., and Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3(4):743–756.

Nichols, J. (1992). *Linguistic diversity in space and time*. University of Chicago Press, Chicago.

Nichols, J., Witzlack-Makarevich, A., and Bickel, B. (2013). The autotyp genealogy and geography database: 2013 release.

Paradis, E., Claude, J., and Strimmer, K. (2004). Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics*, 20(2):289–290.

R Core Team (2017). R: A language and environment for statistical computing.

Revell, L. J., Harmon, L. J., and Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic Biology*, 57(4):591–601.

Richerson, P. J., Boyd, R., and Bettinger, R. L. (2009). Cultural Innovations and Demographic Change. *Human Biology*, 81(2-3):211–235.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Larmarange, J. (2016). *GGally: Extension to 'ggplot2'*. R package version 1.3.0.

Symonds, M. R. and Blomberg, S. P. (2014). A primer on phylogenetic generalised least squares. In Garamszegi, L. Z., editor, *Modern phylogenetic comparative methods and their application in evolutionary biology*, pages 105–130. Springer.

Walker, R. S. and Ribeiro, L. A. (2011). Bayesian phylogeography of the arawak expansion in lowland south america. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1718):2562–2567.

Wickham, H. and Chang, W. (2012). ggplot2: An implementation of the grammar of graphics.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2016). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*.