LETTERS
https://doi.org/10.1038/s41562-018-0457-6
nature human behaviour

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]

There are more than 7,000 languages spoken in the world today[1]. It has been argued that the natural and social environment of languages drives this diversity[2–13]. However, a fundamental question is how strong are environmental pressures, and does neutral drift suffice as a mechanism to explain diversification? We estimate the phylogenetic signals of geographic dimensions, distance to water, climate and population size on more than 6,000 phylogenetic trees of 46 language families. Phylogenetic signals of environmental factors are generally stronger than expected under the null hypothesis of no relationship with the shape of family trees. Importantly, they are also—in most cases—not compatible with neutral drift models of constant-rate change across the family tree branches. Our results suggest that language diversification is driven by further adaptive and non-adaptive pressures. Language diversity cannot be understood without modelling the pressures that physical, ecological and social factors exert on language users in different environments across the globe.

Present-day linguistic diversity is non-randomly distributed across the globe, forming patterns at multiple levels. For example, more than 7,000 languages are currently spoken, and these can be classified into a few hundred language families[1]. Each family contains (ideally) all—and only—descendants of a single ancestral protolanguage. Given that languages evolve through time in a manner similar to the evolution of biological species—through splits, extinctions and horizontal exchange—a language family can be approximated by a structured family tree (or phylogeny) that comprises a set of languages spoken by actual human groups occupying geographical space. An intriguing observation is that not only individual languages are non-randomly distributed across the globe; language families are too: some families are huge, spanning vast areas, while others are much more circumscribed. It has been proposed that this patterning reflects ancestral historical events and processes, such as demographic migrations and spreads, or language shift through elite dominance[14]. Additionally, there is an emerging view that language diversification cannot be fully understood except in the wider context of physical, cultural and biological variation[15–17].

A fundamental question, then, is why and how do language family trees unfold? Is linguistic diversification a self-contained process, or do pressures related to geographic and demographic dimensions drive diversification and shape language family trees? The classic view holds that explanations of diversity have to be sought 'first on the basis of recognized processes of internal change'[18]. Here, 'internal' changes are either seen as a 'rather directionless pursuit of individual forms down the branches of the family tree'[19] or as regular phenomena such as sound change and analogy[19]. Internal changes are often associated with the term 'linguistic drift'[20], which is theoretically distinct from 'population drift' (that is, the social or geographic isolation of speaker communities[21]). However, in practice, Sapir[20] argued that both types of drift interact: variation in individual speakers' utterances accumulate and lead to the formation of dialects and, eventually, languages. The prediction of this account is that purely random variation in language usage could give rise to diversity by means of social and geographic isolation, corresponding to 'neutral drift' models in evolutionary biology.

Accounts based on language internal change have come under criticism for underestimating the role of geography and demography. Nichols[2] has shown that language diversity is greater at low latitudes, along coastlines and in mountainous areas, among others. Nettle[3] found evidence for language density being influenced by ecological risk: areas that have longer growing seasons also support a larger number of languages—a finding that is corroborated by more recent statistical analyses[4]. Other studies investigated global linguistic diversity in relation to geographic and demographic data (see Gavin et al.[5] for a review). Predictors of linguistic diversity include latitude[6,7], altitude and rugosity[4], temperature and rainfall[7–12], political complexity, and subsistence strategy[13], as well as island size in the Pacific[11].

However, it is a standard procedure in evolutionary biology to test neutral drift models before further adaptive processes are invoked for explanation. As pointed out in an overview article by Gavin et al.[5], our understanding of linguistic diversification is still rudimentary. The mechanisms of neutral change, movement, contact and selection have not been disentangled yet. Here, we test different evolutionary models by adding a phylogenetic dimension. This allows us to investigate how strong the links between family tree structure and environmental factors are on a global scale. In evolutionary biology, the strength of the association between population-level traits and a given phylogeny is measured using the so-called phylogenetic signal[22,23]. Estimating phylogenetic signals, we test three fundamental hypotheses:

- Independent evolution hypothesis (H_0). 'Internal' linguistic properties and 'external' environmental factors generally evolve independently: there is no link between environmental factors and the shape of language family trees (that is, their values are randomly distributed across the tips of the trees) and phylogenetic signals are close—or equal—to zero);
- Neutral evolution hypothesis (H_1). Internal properties and environmental factors are linked via neutral drift: the values of the environmental factors follow the predictions of a Brownian motion model (that is, a constant-rate random walk along the branches of the family trees) and phylogenetic signals are close to one;

[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

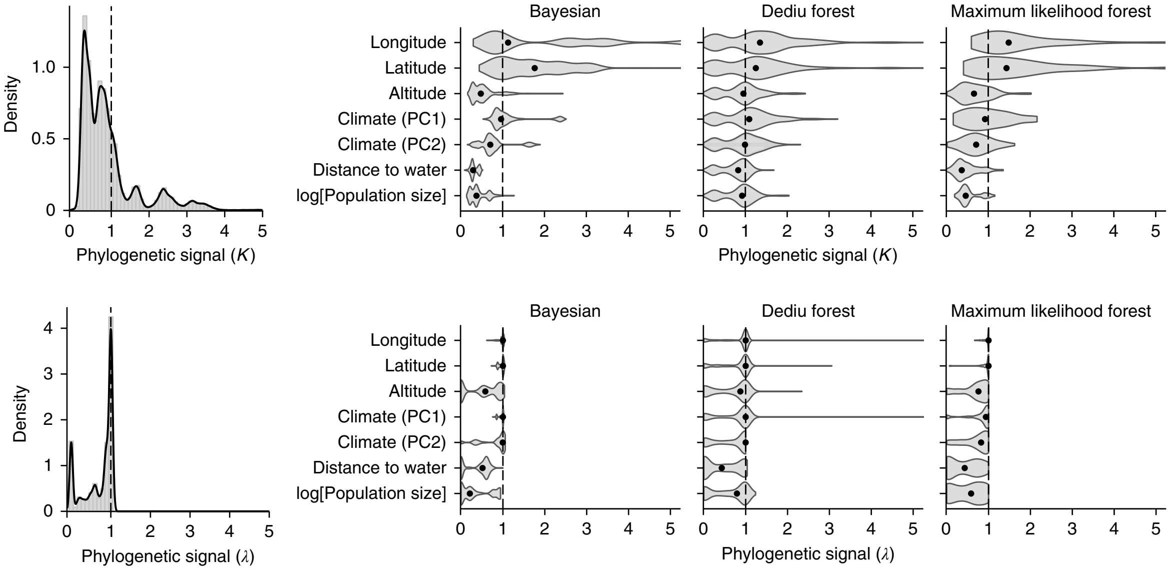

Fig. 1 | Density distributions of phylogenetic signals for *K* and *λ*, and violin plots with distributions per environmental factor. The density distributions shown in the left-hand panels include the phylogenetic signals of all three tree sources and all seven environmental variables. The dashed vertical lines indicate the phylogenetic signal value expected under Brownian motion along the branches of the trees (that is, 1.0). The violin plots to the right show the distributions of *K* and *λ* per environmental factor. Black dots represent median values. The grey transparent areas are density distributions of phylogenetic signal values. The *x* axis is limited to values up to a maximum of five. These plots were produced using the ggplot2 R package[48]. PC, principal component.

Table 1 | Information on the phylogenetic trees used

Tree set	Topology	Branch lengths	Method	Trees	Families
Dediu's forest	W, E, G, A	Lexicon to syntax	Various	351	41
Bayesian trees	No constraints	Cognates	Bayesian	5,801	7
Maximum likelihood forest	No constraints, G	Lexical lists	Maximum likelihood	58	29
			Total	6,210	46

A: AUTOTYP; E, Ethnologue; G, Glottolog; W, WALS.

- Variable evolution hypothesis ($H_{0\text{–}1}$ and H_{1+}). Internal properties and environmental factors are linked via adaptive and nonadaptive processes beyond neutral drift: while phylogenetic signals are significantly higher than zero, they can be either lower or higher than one.

While some recent work has used phylogenetic signal analysis in specific linguistic contexts[24–26], we describe a large-scale analysis of environmental factors for many language families spread across the world.

First, we report the results for 42 tree source subsets (Fig. 1). All median values, upper confidence intervals and *P* values are given in Supplementary Results 1. In this dataset, all environmental factors have median phylogenetic signals significantly higher than 0.1 according to Wilcoxon signed rank tests. This holds for two phylogenetic signal metrics (Blomberg's *K* and Pagel's *λ*) and across three tree sources (see Table 1 and Supplementary Methods 1 for details of the tree sources). The median values range from $\widetilde{\lambda} = 0.21$ to $\widetilde{\lambda} = 1$ and from $\widetilde{K} = 0.3$ to $\widetilde{K} = 1.77$, respectively.

Median phylogenetic signals are also in most cases significantly different from 1.0 (that is, not in the range of 0.9 to 1.1). Some exceptions are the median latitude and longitude *λ* signals (see lower panels on the right of Fig. 1). However, for Blomberg's K (upper panels), longitude and latitude signals are significantly higher than 1.0 across all three tree sources. For example, the median longitude values range from $\widetilde{K} = 1.13$ to $\widetilde{K} = 1.49$ across different tree sources. This suggests that *λ* is at the ceiling for longitude and latitude. Median phylogenetic signals for altitude, population size and distance to water are mostly between 0.1 and 0.9. The results for distances to lakes, rivers and oceans separately are given in Supplementary Results 2. The first principal component of climate has phylogenetic signals close to 1.0. The signal for the second principal component is weaker as it is between 0.1 and 0.9 in some cases (see Supplementary Methods 5 for details of the principal components analyses).

Second, we report phylogenetic signals by family. Plots with median phylogenetic signal values and a table showing the Wilcoxon test results by family subsets are given in Supplementary Results 3. Some families stand out with high median *K* and *λ* signals. Some examples include Atlantic-Congo for longitude ($\widetilde{K} = 7.23$ and $\widetilde{\lambda} = 1$), Uto-Aztecan for latitude ($\widetilde{K} = 2.5$ and $\widetilde{\lambda} = 1$) and climate (principal component 1) ($\widetilde{K} = 1.98$ and $\widetilde{\lambda} = 1$), Sino-Tibetan for altitude ($\widetilde{K} = 1.35$ and $\widetilde{\lambda} = 1$), and Austronesian for population size ($\widetilde{K} = 0.71$ and $\widetilde{\lambda} = 0.93$). Example trees for these families are shown in Fig. 2. Moreover, the environmental factor most strongly reflected on phylogenetic trees can differ between families. For instance, for the Atlantic-Congo family, longitude has the strongest reflection on the family tree, while for the Uto-Aztecan family, latitude does.

Figure 3 gives an overview of the number and percentages of subset median values in our two analyses (by tree source and family) in line

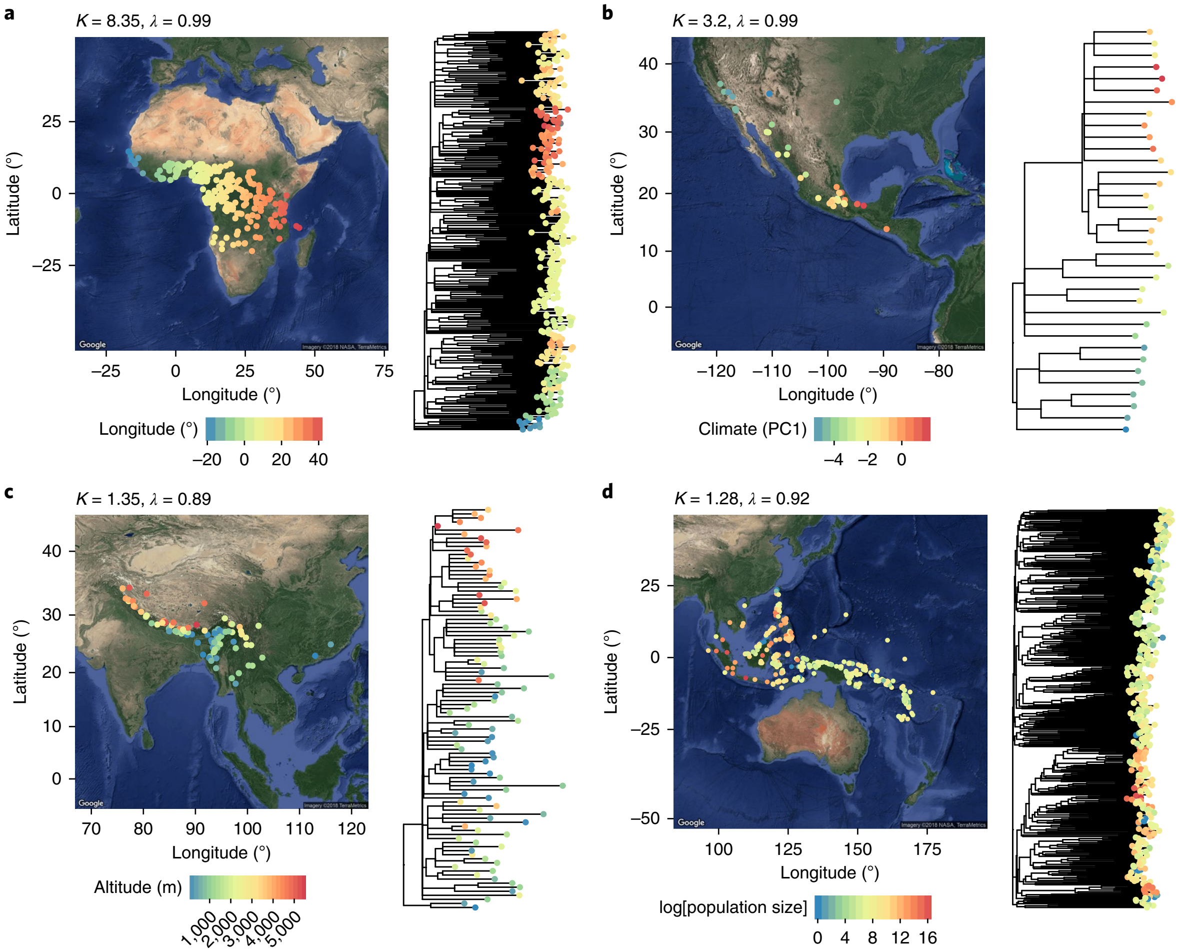

Fig. 2 | Environmental factors reflected on family trees. Colours indicate values of environmental factors from low (blue) to high (red). The phylogenetic trees used here are part of Dediu's forest. **a**, Map of the spread of $n=346$ Atlantic-Congo languages with longitude indicated by colour. The phylogenetic tree to the right was built on the topology from Glottolog and branch lengths were derived from ASJP word lists. **b**, Uto-Aztecan languages ($n=35$) with climate (principal component 1) indicated by colour. The phylogenetic tree was built on the topology from Glottolog and branch lengths were derived from ASJP word lists. **c**, Sino-Tibetan languages ($n=99$) with altitude indicated by colour. The phylogenetic tree was built on the topology from Ethnologue and branch lengths were derived from WALS features. **d**, Austronesian languages ($n=421$) with logged population size indicated by colour. The phylogenetic tree was built on topology from Glottolog and branch lengths were derived from ASJP word lists. Plots were produced using the ggtree and ggmap R packages[49,50]. Map data: Google/NASA/TerraMetrics (**a**,**d**); Google/TerraMetrics (**b**,**c**).

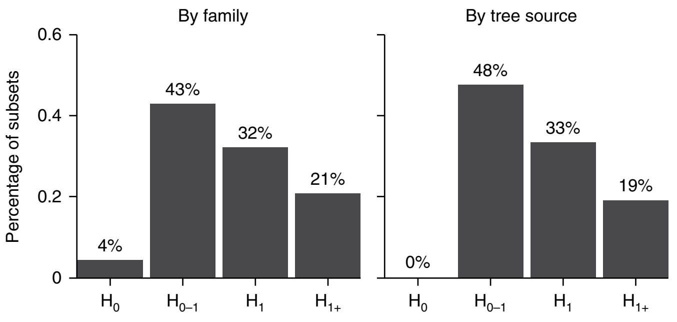

Fig. 3 | Percentages of subsets in line with evolutionary hypotheses. Percentages of subsets by family (left) and tree source (right) in line with the four evolutionary hypotheses of how environmental factors and tree structure are connected (H_0, H_{0-1}, H_1 and H_{1+}).

with the four hypotheses (H_0, H_1, H_{0-1} and H_{1+}). We draw the following conclusions with regards to underlying evolutionary processes:

It is generally very unlikely that the phylogenetic trees and environmental variables we have included in our analyses evolved completely independently. Namely, in our 42 and 140 subsets of more than 85,000 signals, we find only 6 instances where H_0 can be upheld. Examples include altitude λ signals, as reflected on the Tupí-Guaraní ($\widetilde{\lambda} = 0$) and Turkic ($\widetilde{\lambda} = 0$) family trees (see also Supplementary Results 3).

For some subsets (32 and 33%, respectively) neutral drift (H_1) is a possible explanation. In these subsets, median phylogenetic signals are close to 1.0 (that is, 0.9 to 1.1) according to the Wilcoxon test results. Hence, there are examples where the link between environmental factors and family tree structure fits the expectation of constant-rate Brownian motion along the tree branches. For instance, the Arawakan family yields climate (principal component 1) phylogenetic signals close to 1.0, as evidenced by both signal metrics ($\widetilde{\lambda} = 1.01$ and $\widetilde{K} = 1.05$).

Most frequent are median phylogenetic signals that are significantly higher than 0.1 and smaller than 0.9 (in 43 and 48% of the cases, respectively). This is probably due to interactions between different environmental variables, which in our analyses are considered in isolation from each other. For example, the latitude and longitude of a population can change in non-random ways due to dispersals along coastlines and rivers[2], circumvention of uninhabitable lands such as oceans, deserts and high mountain ranges, conquering and colonization of inhabited and uninhabited lands, and migrations due to famines and warfare. In the case of population size, the signal will depend on the minimum group size necessary for a split, as well as the sustainable group size in a particular environment[3]. Also, geographic, demographic, climatic and other pressures—beyond the ones taken into account here—are likely to further shape family trees.

Furthermore, Blomberg et al.[22] give two explanations for $K < 1$ based on computational simulations: (1) measurement error (see Supplementary Discussion 1); and (2) adaptive evolution independent of the phylogeny (that is, homoplasy; also called convergent evolution). This is the case if languages adapt to particular climatic and sociolinguistic environments independent of inheritance. Potential examples of convergent evolution include phonetic changes due to climate[15,27] and morphological changes due to population structure[26,28–30]. However, note that most of the language family trees used here are built using lexical material. Phonetic and morphological adaptations are therefore unlikely to play a predominant role for the overall outcome.

Unfortunately, inferring exact evolutionary mechanisms just from observing $K < 1$ is not possible. Computational simulations have identified different scenarios that can lead to phylogenetic signals below 1.0, including stabilizing and fluctuating selection in combination with differing mutation rates, functional constraints and others[31]. Further analyses are needed to uncover the exact mechanisms yielding reduced phylogenetic signal in particular language families.

In a considerable number of cases (19 and 21%, respectively), phylogenetic signals are significantly higher than 1.1. In fact, these percentages are higher if only K signals are included. As explained in Supplementary Methods 3, λ is very unlikely to exceed 1.0. Blomberg et al.[22] identify two scenarios under which $K > 1$ is to be expected—namely: (1) heterogeneous rate genetic drift (that is, changing rates of genetic drift), with high initial genetic drift that then slows down towards the tips of the family tree; and (2) niche occupancy, meaning that species occupy many different niches early on, adapt to them and diverge, but when all niches are filled, they do not further diverge.

For our data, 'heterogeneous rate genetic drift' translates into varying rates of change for our environmental variables. Latitudes and longitudes of speaker populations might have changed fast initially—close to the root of a given tree—and then slowed down systematically. The geographic locations where languages are spoken today would then still reflect early linguistic splits. The scenario of 'niche occupancy' seems in line with this. Take the example of the Atlantic-Congo tree in Fig. 2. The longitude signal for this tree is exceptionally high ($K = 8.35$), which is also the case for the median signal of the family ($\widetilde{K} = 7.23$) across the different tree sources. Early large-scale migrations such as the Bantu expansion could explain why observed geographical distances are smaller than expected from evolving longitudes along the branches of the tree by constant Brownian motion.

Lateral transfer (that is, borrowing of lexical and structural material) is another mechanism that drives linguistic diversification and convergence[32]. Borrowing increases the similarity between the donor and recipient language, and is more likely in geographic proximity. This can increase the phylogenetic signal of longitudes, latitudes and altitudes. However, in most cases, particular care is taken to exclude potentially borrowed material when building linguistic phylogenies of the kind underlying our analyses.

Furthermore, we can assess which environmental factor has the strongest phylogenetic signal overall. Across the three different tree sources illustrated in Fig. 1 (right panels) a systematic cline emerges for phylogenetic signal strength by environmental factor: longitude/latitude > climate (principal components 1 and 2) > altitude > distance to water ~ population size. However, there is also considerable variance between families (see Supplementary Results 3). For instance, while large families of Africa and Eurasia tend to have stronger longitude signals (Atlantic-Congo, Afro-Asiatic, Altaic, Austroasiatic, Indo-European and Sino-Tibetan), large families of North and South America tend to have stronger latitude signals (Arawakan, Athabaskan–Eyak–Tlingit, Otomanguean, Quechuan, Tupí-Guaraní and Uto-Aztecan). The fact that longitude has generally stronger signals in large African and Eurasian families is probably related to the hypothesis that east–west spreads have played a more important role for human expansions than north–south spreads. The rationale behind this is that climate and vegetation are more similar across different longitudes, and this might facilitate expansions—especially when associated with agriculture[33]. This hypothesis is also tested in another recent quantitative study[34].

Finally, we want to mention potential issues with the phylogenetic signal approach as applied to language family trees: (1) bias through error (that is, imprecisions in the trees and tip values); (2) geographic and population size variation within languages; (3) systematic variance between the tree sources used; and (4) factors beyond geography, climate, distance to water and population size influencing diversification. These problems and caveats are laid out in more detail in Supplementary Discussion 1.

In conclusion, we find that the structure of language family trees generally reflects environmental factors associated with particular language communities. Across more than 6,000 phylogenetic trees of 46 families, this effect is clearly stronger than expected under the null hypothesis of independence between language 'internal' structure and language 'external' environmental factors. Importantly, the links between environmental factors and the structure of family trees often deviate from the predictions of neutral drift, suggesting that there are adaptive and non-adaptive forces rooted in the physical and social environment that affect the evolution of language families. This supports recent claims that pure drift falls short of explaining a considerable proportion of language diversity. Instead, adaptive pressures[16,35] and other non-adaptive processes have to be taken into consideration. The forces further driving diversification potentially include convergent evolution, niche occupancy, heterogeneous rate drift and lateral transfer of lexical and structural material. Language family trees reflect both internal forces and shallower or deeper historical phenomena. In consequence, understanding global linguistic diversity is not possible without analysing the physical and social circumstances of language users.

Methods

Language family trees were collected using three tree sources. First, a database[36] comprising linguistic trees for several hundred language families (here, referred to as 'Dediu's forest') was compiled. It is available via github (https://github.com/ddediu/lgfam-newick). Tree topologies were taken from Ethnologue[37], the World Atlas of

Language Structures (WALS) Online[38], AUTOTYP[39] and Glottolog[1]. These were bare of any branch length information. This information was added using a variety of different methods[36]. We selected a subset of these trees for phylogenetic signal analyses (see Supplementary Methods 1 for details).

Second, Bayesian trees were supplied by the authors of recent phylogenetic studies for a total of seven language families: Arawakan, Austronesian, Bantu, Indo-European, Pama-Nyungan, Tupí-Guaraní and Turkic. The respective studies used cognate data in conjunction with Bayesian phylogenetic methods to derive a collection of high posterior probability trees (Supplementary Methods 1).

Third, trees were derived via the maximum likelihood method applied to Automated Similarity Judgment Program (ASJP) word lists (http://asjp.clld.org/). These were further divided into two sets: maximum likelihood trees with branch lengths and topology inferred (1) with Glottolog topologies as constraints and (2) without constraints. There were a total of 29 family trees for each. Details on data availability are provided online[40].

Note that we only included trees with at least 20 tips; that is, languages. Smaller numbers resulted in low statistical power to detect phylogenetic signal[22,41]. We arrived at a sample of 6,210 trees (see Table 1).

Environmental variables were collected from different online resources. Approximated latitude and longitude information per language was available via Glottolog[1]. We transformed longitudes to run from −25 to 335° instead of the standard −180 to 180°. This is necessary as families in the Pacific (for example, Austronesian) expand across the 180 to −180° line, which distorts the longitude signals. Based on latitude and longitude coordinates, we estimated altitude using the Google Maps Elevation API (https://developers.google.com/maps/documentation/elevation) via the R[42] package rgbif[43]. To analyse phylogenetic signals of climate, we harnessed the first two principal components of a principal components analysis of 19 climatic variables (see Supplementary Methods 5). We also included distance to water (lakes, rivers and oceans) as an environmental variable (Supplementary Methods 6). Population size data were taken from the last openly available version of Ethnologue[37]. We took the natural logarithm of population sizes, otherwise extreme values (for example, for English and Mandarin Chinese) would dominate the signal for the whole family. We arrived at a sample of 6,998 languages (unique ISO 639-3 codes) of 232 Glottolog families for which latitude, longitude, altitude, climatic information, distance to water and population size data were available (see Supplementary Data 3). We chose these variables to reflect different dimensions of the environment. Note that some were mutually correlated (Supplementary Note 1).

There are many methods to estimate phylogenetic signal for continuous variables, and Supplementary Methods 2 discusses the advantages and disadvantages of different metrics. Here, we focus on two in particular: Pagel's λ[44–46] and Blomberg's K[22]. We use the R function phylosig() in the package phytools[47] to calculate both λ and K values. Phylogenetic signals of around 0.1 are in line with H_0. Values close to 1.0 are generally in line with H_1 (but see Revell et al.[31] for some cautionary notes). All other values point to either H_{0-1} or H_{1+}. Note that K can exceed 1.0, while for λ this is unlikely but theoretically possible (see Supplementary Methods 3 and 4 for further details).

The structure of our data is such that we have 6,210 phylogenetic trees of 46 families, 2 phylogenetic signal metrics and 7 environmental variables. We thus obtain 6,210 × 2 × 7 = 86,940 phylogenetic signals. All signals are described in Supplementary Results 4. For statistical analyses, we subset these in two ways: first, 42 subsets by signal metric (2), environmental factor (7) and tree source (3); and second, 644 subsets by signal metric (2), environmental factor (7) and language family (46).

To test statistical significance, we used Wilcoxon signed rank tests, since density distributions of phylogenetic signals by subsets are generally non-normal. This can be visually checked in Fig. 1. We used the R function wilcox.test() to assess whether median phylogenetic signals for these different subsets were significantly higher than zero (≥0.1), close to one (≥0.9) or significantly higher than one (≥1.1). We report percentages of phylogenetic signal median values per subset that are in line with the four hypotheses outlined above (H_0, H_{0-1}, H_1 and H_{1+}). Note that in the file 'wilcoxonResults_families.csv', we give the results of Wilcoxon tests for all 644 family subsets (see Supplementary Data 7). However, in the percentage counts of median values supporting a given hypothesis, we only include 140 families for which there are more than 20 phylogenetic trees, since tests for fewer than 20 data points are biased to yield non-significant results.

Importantly, individual trees and corresponding tip values can give rise to a range of phylogenetic signals, even when simulated with constant-rate Brownian motion. The expected mean value for K under Brownian motion is 1.0 and the median is 0.9 (see Supplementary Methods 4 for a simulation).

Finally, we used Bonferroni correction for multiple testing to adjust P values (see Supplementary Results 1 for further discussion).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Code availability. R code is available in separate files described in the Guide to the Supplementary Information.

Data availability

Data availability is detailed in Supplementary Methods 1. Individual data files are described in Supplementary Data 1–7 in the Guide to the Supplementary Information.

Received: 21 December 2017; Accepted: 24 September 2018;
Published online: 5 November 2018

References

1. Hammarström, H., Forkel, R., Haspelmath, M. & Bank, S. *Glottolog 3.2* (Max Planck Institute for the Science of Human History, 2018); http://glottolog.org
2. Nichols, J. Linguistic diversity and the first settlement of the New World. *Language* **66**, 475–521 (1990).
3. Nettle, D. Using social impact theory to simulate language change. *Lingua* **108**, 95–117 (1999).
4. Coupé, C., Hombert, J.-M., Marsico, E. & Pellegrino, F. in *East Flows the Great River: Festschrift in Honor of Prof. William S-Y. WANG on his 80th Birthday* (eds Peng, G. & Shi, F.) 76–103 (City Univ. Hong Kong Press, Hong Kong, 2013).
5. Gavin, M. C. et al. Toward a mechanistic understanding of linguistic diversity. *Bioscience* **63**, 524–535 (2013).
6. Mace, R. & Pagel, M. A latitudinal gradient in the density of human languages in North America. *Proc. Biol. Sci.* **261**, 117–121 (1995).
7. Collard, I. F. & Foley, R. A. Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evol. Ecol. Res.* **4**, 371–383 (2002).
8. Moore, J. L. et al. The distribution of cultural and biological diversity in Africa. *Proc. Biol. Sci.* **269**, 1645–1653 (2002).
9. Dimmendaal, G. J. Language ecology and linguistic diversity on the African continent. *Lang. Linguist. Compass* **2**, 840–858 (2008).
10. Axelsen, J. B. & Manrubia, S. River density and landscape roughness are universal determinants of linguistic diversity. *Proc. Biol. Sci.* **281**, 20133029 (2014).
11. Gavin, M. C. & Sibanda, N. The island biogeography of languages. *Glob. Ecol. Biogeogr.* **21**, 958–967 (2012).
12. Gavin, M. C. et al. Process-based modelling shows how climate and demography shape language diversity. *Glob. Ecol. Biogeogr.* **26**, 584–591 (2017).
13. Currie, T. E. & Mace, R. Political complexity predicts the spread of ethnolinguistic groups. *Proc. Natl Acad. Sci. USA* **106**, 7339–7344 (2009).
14. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1994).
15. Everett, C., Blasi, D. E. & Roberts, S. G. Climate, vocal folds, and tonal languages: connecting the physiological and geographic dots. *Proc. Natl Acad. Sci. USA* **112**, 1322–1327 (2015).
16. Lupyan, G. & Dale, R. Why are there different languages? The role of adaptation in linguistic diversity. *Trends Cogn. Sci.* **20**, 649–660 (2016).
17. Dediu, D., Janssen, R. & Moisik, S. R. Language is not isolated from its wider environment: vocal tract influences on the evolution of speech and language. *Lang. Commun.* **54**, 9–20 (2017).
18. Welmers, W. E. *African Language Structures* (University of California Press, Berkeley and Los Angeles, 1973).
19. McMahon, A. M. *Understanding Language Change* (Cambridge Univ. Press, Cambridge, 1994).
20. Sapir, E. *Language: An Introduction to the Study of Speech* (Harcourt, Brace & World, New York, 1921).
21. Jones, M. C. & Singh, I. *Exploring Language Change* (Routledge, New York, 2005).
22. Blomberg, S. P., Garland, T. Jr., Ives, A. R. & Crespi, B. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).
23. Symonds, M. R. & Blomberg, S. P. in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology* (ed. Garamszegi, L. Z.) 105–130 (Springer, Heidelberg, 2014).
24. Verkerk, A. Diachronic change in Indo-European motion event encoding. *J. Hist. Linguist.* **4**, 40–83 (2014).
25. Verkerk, A. The correlation between motion event encoding and path verb lexicon size in the Indo-European language family. *Folia Linguist. Hist.* **35**, 307–358 (2014).
26. Bentz, C., Verkerk, A., Kiela, D., Hill, F. & Buttery, P. Adaptive communication: languages with more non-native speakers tend to have fewer word forms. *PLoS ONE* **10**, e0128254 (2015).
27. Everett, C. Evidence for direct geographic influences on linguistic sounds: the case of ejectives. *PLoS ONE* **8**, e65275 (2013).
28. Lupyan, G. & Dale, R. Language structure is partly determined by social structure. *PLoS ONE* **5**, e8559 (2010).

29. Dale, R. & Lupyan, G. Understanding the origins of morphological diversity: the linguistic niche hypothesis. *Adv. Complex Syst.* **15**, 1150017 (2012).
30. Bentz, C. & Winter, B. Languages with more second language speakers tend to lose nominal case. *Lang. Dynam. Change* **3**, 1–27 (2013).
31. Revell, L. J., Harmon, L. J. & Collar, D. C. Phylogenetic signal, evolutionary process, and rate. *Syst. Biol.* **57**, 591–601 (2008).
32. Thomason, S. G. & Kaufman, T. *Language Contact, Creolization, and Genetic Linguistics* (Univ. California Press, Berkeley & Oxford, 1988).
33. Diamond, J. M. *Guns, Germs and Steel: The Fates of Human Societies* (W. W. Norton, New York & London, 1999).
34. Güldemann, T. & Hammarström, H. in *Language Dispersal, Diversification and Contact* (eds Crevels, M. & Muysken, P.) (Oxford Univ. Press, Oxford, 2017).
35. Lupyan, G. & Dale, R. in *Language Structure and Environment* (eds De Busser, R. & LaPolla, R. J.) 289–316 (John Benjamins Publishing Company, Amsterdam, 2015).
36. Dediu, D. Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Lang. Dynam. Change* **8**, 1–21 (2018).
37. Lewis, M. P., Simons, G. F. & Fenning, C. D. *Ethnologue: Languages of the World* 17th edn (SIL International, Dallas, 2013); http://www.ethnologue.com
38. Dryer, M. S. & Haspelmath, M. *The World Atlas of Language Structures Online* (Max Planck Digital Library, 2013); http://wals.info/
39. Nichols, J., Witzlack-Makarevich, A. & Bickel, B. *The AUTOTYP Genealogy and Geography Database 2013 Release* (2013); https://www.spw.uzh.ch/autotyp/
40. Jäger, G. Global-scale phylogenetic linguistic inference from lexical resources. Preprint at http://arxiv.org/abs/1802.06079 (2018).
41. Münkemüller, T. et al. How to measure and test phylogenetic signal. *Methods Ecol. Evol.* **3**, 743–756 (2012).
42. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
43. Chamberlain, S. rgbif: Interface to the Global 'Biodiversity' Information Facility 'API', R package version 0.9.5 (2016); https://CRAN.R-project.org/package=rgbif
44. Pagel, M. Inferring evolutionary processes from phylogenies. *Zool. Scr.* **26**, 331–348 (1997).
45. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
46. Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726 (2002).
47. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
48. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, New York, 2016).
49. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
50. Kahle, D. & Wickham, H. ggmap: spatial visualization with ggplot2. *R. J.* **5**, 144–161 (2013).

Acknowledgements

C.B. and G.J. were funded by the German Research Foundation (DFG FOR 2237; project 'Words, Bones, Genes, Tools: Tracking Linguistic, Cultural, and Biological Trajectories of the Human Past') and the ERC Advanced Grant 324246 EVOLAEMP. D.D. was funded by The Netherlands Organisation for Scientific Research VIDI grant 276-70-022 and the European Institutes for Advanced Study Fellowship Program. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

C.B. was responsible for project inception, statistical and phylogenetic analyses, and writing of the paper. D.D., A.V. and G.J. contributed data, phylogenetic analyses and writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0457-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to C.B.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Christian Bentz

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ □ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

☒ □ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The R code is given as supplementary material and described in Supplementary Methods 4
Data analysis	The R code is given as supplementary material and described in Supplementary Methods 4

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data availability is detailed in Supplementary Data 1-4

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences　□ Behavioural & social sciences　☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study uses quantitative phylogenetic methods applied to linguistic data.
Research sample	The sample consists of 6210 language family trees of overall 46 different language families.
Sampling strategy	We aimed to collect the biggest possible sample of language family trees given online resources and published data.
Data collection	Bayesian trees were collected by Annemarie Verkerk. Trees given in "Dediu's forest" were produced by Dan Dediu. The ML trees were produced by Gerhard Jäger. Information on environmental variables per language were collected by Dan Dediu and Chris Bentz. Details about these data sets are given in supplementary data files.
Timing and spatial scale	NA
Data exclusions	Data exclusions for various reasons are given in Supplementary Data 2.
Reproducibility	NA
Randomization	NA
Blinding	NA

Did the study involve field work?　□ Yes　☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study	
☒	□	Unique biological materials
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
☒	□	Human research participants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging