



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

Quentin D. Atkinson

Science **332**, 346 (2011);

DOI: 10.1126/science.1199295

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of January 11, 2012):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/content/332/6027/346.full.html>

Supporting Online Material can be found at:

<http://www.sciencemag.org/content/suppl/2011/04/13/332.6027.346.DC1.html>

This article **cites 23 articles**, 10 of which can be accessed free:

<http://www.sciencemag.org/content/332/6027/346.full.html#ref-list-1>

This article has been **cited by** 3 articles hosted by HighWire Press; see:

<http://www.sciencemag.org/content/332/6027/346.full.html#related-urls>

This article appears in the following **subject collections**:

Evolution

<http://www.sciencemag.org/cgi/collection/evolution>

ability to control the intricate structure of DNA nano-architectures and create more diverse building blocks for molecular engineering.

References and Notes

- N. C. Seeman, *J. Theor. Biol.* **99**, 237 (1982).
- N. C. Seeman, *Nature* **421**, 427 (2003).
- T. J. Fu, N. C. Seeman, *Biochemistry* **32**, 3211 (1993).
- J. H. Chen, N. C. Seeman, *Nature* **350**, 631 (1991).
- E. Winfree, F. Liu, L. A. Wenzler, N. C. Seeman, *Nature* **394**, 539 (1998).
- H. Yan, S. H. Park, G. Finkelstein, J. H. Reif, T. H. LaBean, *Science* **301**, 1882 (2003).
- W. M. Shih, J. D. Quispe, G. F. Joyce, *Nature* **427**, 618 (2004).
- F. Mathieu *et al.*, *Nano Lett.* **5**, 661 (2005).
- R. P. Goodman *et al.*, *Science* **310**, 1661 (2005).
- F. A. Aldaye, H. F. Sleiman, *J. Am. Chem. Soc.* **129**, 13376 (2007).
- Y. He *et al.*, *Nature* **452**, 198 (2008).
- C. Zhang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10665 (2008).
- C. Lin, Y. Liu, H. Yan, *Biochemistry* **48**, 1663 (2009).
- P. W. K. Rothmund, *Nature* **440**, 297 (2006).
- E. S. Andersen *et al.*, *Nature* **459**, 73 (2009).
- S. M. Douglas *et al.*, *Nature* **459**, 414 (2009).
- Y. Ke *et al.*, *J. Am. Chem. Soc.* **131**, 15903 (2009).
- H. Dietz, S. M. Douglas, W. M. Shih, *Science* **325**, 725 (2009).
- Materials and methods are available as supporting material on *Science* Online.
- P. W. K. Rothmund *et al.*, *J. Am. Chem. Soc.* **126**, 16344 (2004).
- P. J. Hagerman, *Annu. Rev. Biophys. Chem.* **17**, 265 (1988).

Acknowledgments: We thank the AFM applications group at Bruker Nanosurfaces for assistance in acquiring some of the high-resolution AFM images, using Peak Force Tapping with ScanAsyst on the MultiMode 8. This research was partly supported by grants from the Office of Naval

Research, Army Research Office, National Science Foundation, Department of Energy, and National Institutes of Health to H.Y. and Y.L. and from the Sloan Research Foundation to H.Y. Y.L. and H.Y. were supported by the Technology and Research Initiative Fund from Arizona State University and as part of the Center for Bio-Inspired Solar Fuel Production, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award DE-SC0001016.

Supporting Online Material

www.sciencemag.org/cgi/content/full/332/6027/342/DC1
Materials and Methods
SOM Text
Figs. S1 to S39
Tables S1 to S19
Reference 22

18 January 2011; accepted 4 March 2011
10.1126/science.1202998

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

Quentin D. Atkinson^{1,2*}

Human genetic and phenotypic diversity declines with distance from Africa, as predicted by a serial founder effect in which successive population bottlenecks during range expansion progressively reduce diversity, underpinning support for an African origin of modern humans. Recent work suggests that a similar founder effect may operate on human culture and language. Here I show that the number of phonemes used in a global sample of 504 languages is also clinal and fits a serial founder-effect model of expansion from an inferred origin in Africa. This result, which is not explained by more recent demographic history, local language diversity, or statistical non-independence within language families, points to parallel mechanisms shaping genetic and linguistic diversity and supports an African origin of modern human languages.

The number of phonemes—perceptually distinct units of sound that differentiate words—in a language is positively correlated with the size of its speaker population (*I*) in such a way that small populations have fewer phonemes. Languages continually gain and lose phonemes because of stochastic processes (2, 3). If phoneme distinctions are more likely to be lost in small founder populations, then a succession of founder events during range expansion should progressively reduce phonemic diversity with increasing distance from the point of origin, paralleling the serial founder effect observed in population genetics (4–9). A founder effect has already been used to explain patterns of variation in other cultural replicators, including human material culture (10–13) and birdsong (14). A range of possible mechanisms (15) predicts similar dynamics govern-

ing the evolution of phonemes (11, 16) and language generally (17–20). This raises the possibility that the serial founder-effect model used to trace our genetic origins to a recent expansion from Africa (4–9) could also be applied to global phonemic diversity to investigate the origin and expansion of modern human languages. Here I examine geographic variation in phoneme inventory size using data on vowel, consonant, and tone inventories taken from 504 languages in the World Atlas of Language Structures (WALS) (21), together with information on language location, taxonomic affiliation, and speaker demography (Fig. 1 and table S1) (15).

Consistent with previous work (1), speaker population size is a significant predictor of phonemic diversity (Pearson's correlation $r = 0.385$, $df = 503$, $P < 0.001$), with smaller population size predicting smaller overall phoneme inventories (fig. S1A). The same relationship holds for vowel ($r = 0.378$, $df = 503$, $P < 0.001$) and tone ($r = 0.230$, $df = 503$, $P < 0.001$) inventories separately, with a weaker, though still significant, effect of population size on consonant diversity ($r =$

0.131, $df = 503$, $P = 0.003$). To account for any non-independence within language families, the analysis was repeated, first using mean values at the language family level (table S2) and then using a hierarchical linear regression framework to model nested dependencies in variation at the family, subfamily, and genus levels (15). These analyses confirm that, consistent with a founder effect model, smaller population size predicts reduced phoneme inventory size both between families (family-level analysis $r = 0.468$, $df = 49$, $P < 0.001$; fig. S1B) and within families, controlling for taxonomic affiliation {hierarchical linear model: fixed-effect coefficient (β) = 0.0338 to 0.0985 [95% highest posterior density (HPD)], $P = 0.009$ }.

Figure 1B shows clear regional differences in phonemic diversity, with the largest phoneme inventories in Africa and the smallest in South America and Oceania. A series of linear regressions was used to predict phoneme inventory size from the log of speaker population size and distance from 2560 potential origin locations around the world (15). Incorporating modern speaker population size into the model controls for geographic patterning in population size and means that the analysis is conservative about the amount of variation attributed to ancient demography. Model fit was evaluated with the Bayesian information criterion (BIC) (22). Following previous work (5, 6), the set of origin locations within four BIC units of the best-fit location was taken to be the most likely area of origin under a serial founder-effect model.

The origin locations producing the strongest decline in phonemic diversity and best-fit model lie across central and southern Africa (Fig. 2A). This region could represent either a single origin for modern languages or the main origin under a polygenesis scenario. The best-fit model incorporating population size and distance from the origin explains 31% of the variance in phoneme inventory size [correlation coefficient (R) = 0.558, $F_{2,501} = 113.463$, $P < 0.001$] (Fig. 3). Both population size ($r_{\text{population}} = 0.146$, $P = 0.002$)

¹Department of Psychology, University of Auckland, Private Bag 92019, Auckland, New Zealand. ²Institute of Cognitive and Evolutionary Anthropology, University of Oxford, 64 Banbury Road, Oxford OX2 6PN, UK.

*E-mail: q.atkinson@auckland.ac.nz

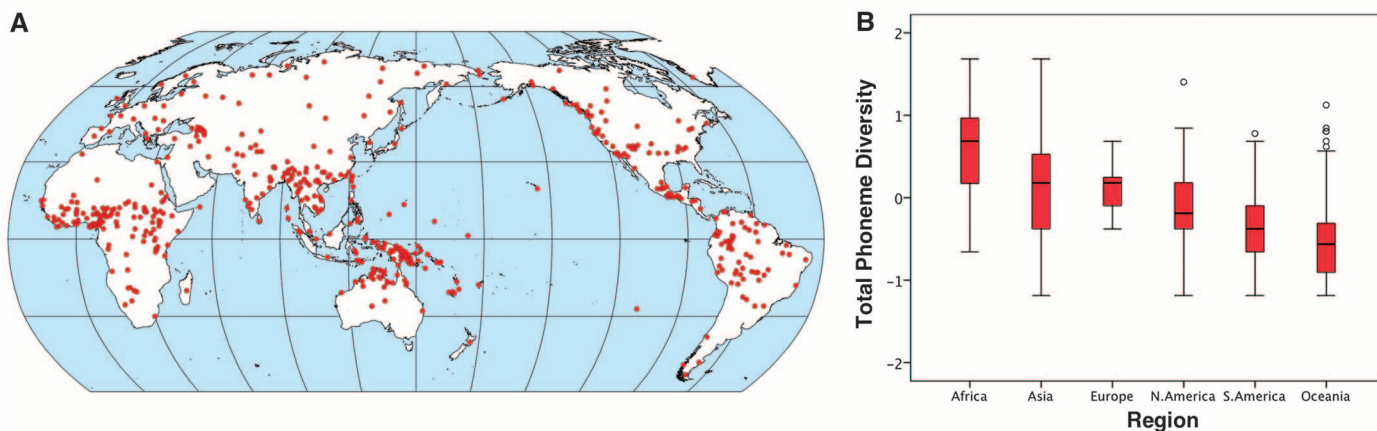


Fig. 1. Language locations and regional variation in phonemic diversity. **(A)** Map showing the location of the 504 sampled languages for which phoneme data was compiled from the WALS database. **(B)** Box plots of overall phonemic diversity by

region reveal substantial regional variation ($\chi^2 = 188.7$, $df = 5$, $P < 0.001$), with the highest diversity in Africa and the lowest diversity in Oceania and South America. The same regional pattern also applies at the language family level (fig. S2).

and distance from origin ($r_{\text{distance}} = -0.438$, $P < 0.001$) are significant predictors in the model. Controlling for population size, distance from origin accounts for 19% of the variance in phonemic diversity ($r = 0.545$, $P < 0.001$). The relationship also holds for vowel ($r = -0.394$, $P < 0.001$), consonant ($r = -0.260$, $P < 0.001$), and tone diversity ($r = -0.391$, $P < 0.001$) separately.

To account for relatedness within families, I repeated the above regressions using mean values across language families (table S2) and under a hierarchical linear model comprising the three taxonomic levels recorded in WALS (15). The hierarchical model results closely matched those of the individual language analysis (fig. S4). Adding an interaction effect did not significantly improve model fit, indicating that the patterns reported here reflect a consistent trend that holds across the globe. The family-level analysis was consistent with the individual language analyses, although the credible region of origin is expanded to include all of Africa (Fig. 2B). Distance from the best fit origin ($r_{\text{distance}} = -0.401$, $P = 0.004$) and population size ($r_{\text{population}} = 0.300$, $P = 0.036$) are both significant predictors and account for 39% of the variance in phonemic diversity between families ($R = 0.627$, $F_{2,47} = 15.190$, $P < 0.001$; fig. S5). As a further test of the robustness of these findings, individual regressions were repeated using partial Mantel tests, which allow for non-independence between data points and avoid assumptions about the statistical distributions underlying the variables of interest (15). The results of this analysis matched the findings reported above (table S3).

To examine the possibility of language polygenesis, distance from a second origin location was added as a predictor to a model incorporating population size and distance from the best-fit origin in Africa. The best-fit models in this anal-

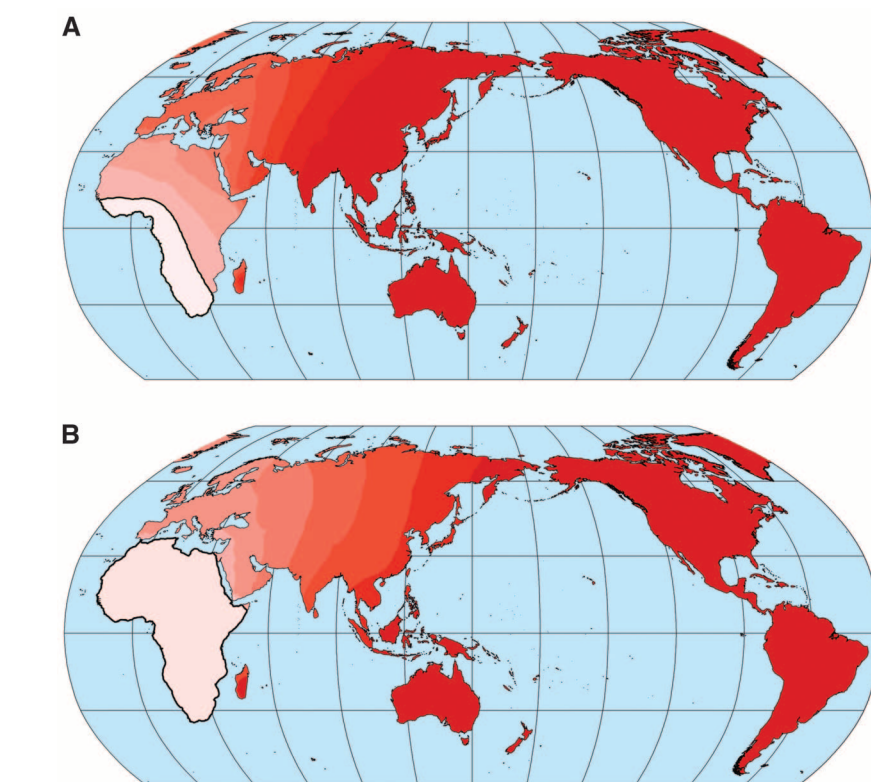
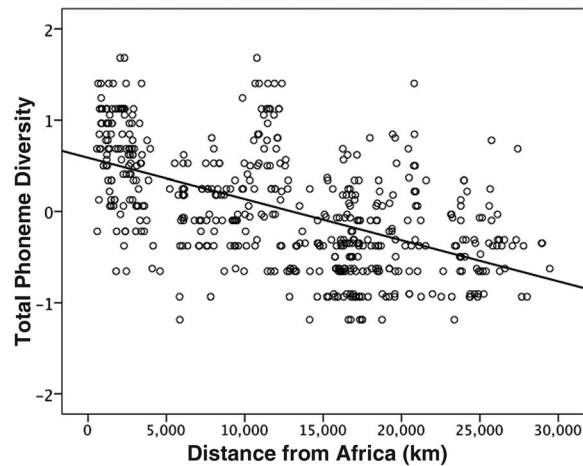


Fig. 2. Likely area of language origin. Maps show the likely location of a single language origin under a founder effect model of phonemic diversity (controlling for population size) inferred from **(A)** individual languages and **(B)** mean diversity across language families. Lighter shading implies a stronger inverse relationship between phonemic diversity and distance from the origin and better fit of the model, as measured by the BIC. The most likely region of origin, comprising those locations within four BIC units of the best-fit origin location, is the area of lightest shading outlined in bold.

ysis did not show a significant negative correlation between distance from a second origin and phonemic diversity. Restricting the analysis to second origin locations that do show an inverse relationship, a region of best fit can be identified in South America (fig. S6). However, this pattern

does not appear under the hierarchical linear model or language family-level analysis; adding a second origin does not improve the fit of either model as measured by the BIC, and all putative second origin locations are within four BIC units. The area identified in the individual-level analysis may

Fig. 3. Phonemic diversity versus distance from the best-fit origin in Africa. A plot of distance from the best-fit origin location in Africa against overall phoneme inventory size is shown. Distance from the origin alone explains 30% of the variation in phonemic diversity (fitted line; $r = -0.545$, $n = 504$ languages, $P < 0.001$) and 19.2% of the variation after controlling for modern speaker population size ($r_{\text{distance}} = -0.438$, $P < 0.001$; $r_{\text{population}} = 0.146$, $P < 0.001$; $R = 0.558$, $F_{2,501} = 113.463$, $P < 0.001$).



be an artefact of slightly higher diversity levels in the Americas than in Oceania, despite the two being comparable distances from Africa, possibly because of a stronger founder effect across the remote Pacific (18–20, 23). When the languages of Oceania are removed from the individual analysis, the effect of distance from an African origin remains ($r_{\text{distance1}} = -0.447$, $P < 0.001$) but there is no significant effect of distance from the secondary origin ($r_{\text{distance2}} = -0.065$, $P = 0.192$). As expected if language spread south with the initial colonization of the Americas, distance from the Bering Strait is inversely correlated with phoneme inventory size within the Americas after controlling for population size ($r_{\text{distance}} = -0.173$, $P = 0.043$).

An ostensibly global cline in phonemic diversity supporting an expansion from Africa could also arise as an artefact of a series of more recent expansions after the Last Glacial Maximum (LGM) into northern Eurasia, the Americas, and the remote Pacific. Languages in these regions show lower average phonemic diversity than in the rest of the world ($t = -6.597$, $df = 503$, $P < 0.001$), which is consistent with a more recent colonization. However, expansion after the LGM does not account for the global cline in phonemic diversity. Distance from Africa remains a significant predictor of phonemic diversity after controlling for colonization since the LGM ($r_{\text{distance}} = -0.401$, $P < 0.001$; $r_{\text{population}} = 0.152$, $P = 0.001$; $r_{\text{LGM}} = 0.032$, $P = 0.419$), as well as when these more recently colonized areas are excluded from the analysis altogether ($r_{\text{distance}} = -0.511$, $P < 0.001$; $r_{\text{population}} = 0.253$, $P < 0.001$) (19).

Demographic factors other than population size may also influence phonemic diversity, particularly those affecting levels of contact and borrowing between groups of speakers. Because neighboring populations at similar points in an expansion are more likely to have similar phonemes and levels of diversity, moderate horizontal transfer between populations can maintain a cline, as has been the case for human genetic diversity (24). However, geographic variation in language diversity (the number of languages per

unit of area), population density (the number of speakers per unit of area), or language area (the total area over which a language is spoken) could affect regional phonemic diversity by increasing contact within and between groups and creating more opportunities to borrow new phonemes. To test whether any such effect could explain the observed global cline in phonemic diversity, these additional measures were included, together with population size and distance from the best-fit origin in Africa, in a regression model predicting phoneme inventory size (15). Controlling for other demographic variables in this way, sub-Saharan Africa remains the most likely area of origin (fig. S7). Distance from the best-fit origin location is a significant predictor at the individual language level ($r_{\text{distance}} = -0.413$, $P < 0.001$), family level ($r_{\text{distance}} = -0.384$, $P < 0.008$), and in the hierarchical linear model [$\beta = -3.419 \times 10^{-5}$ to -2.223×10^{-5} (95% HPD), $P < 0.001$]. The demographic variables are highly correlated and did not show significant independent effects on phonemic diversity. Stepwise regressions indicated that a model incorporating distance from Africa, population size, and (at the individual language level only) language area best explained phoneme inventory size (15).

The single major cline in phonemic diversity is consistent with a linguistic founder effect operating under conditions of rapid expansion from a most likely origin in Africa. This supports a picture of language spread that is congruent with similar analyses of human genetic (4, 5, 7, 8) and phenotypic (6, 9) diversity. Phonemic diversity appears to be highly stable within major language families (15), indicating that, despite the many sociolinguistic processes at work (2, 3), robust statistical patterns in global variation can persist for many millennia and could plausibly reflect a time scale on the order of the African exodus. Outside Africa, the highest levels of phonemic diversity are found in language families thought to be autochthonous to Southeast Asia. This also fits with genetic evidence (25, 26), indicating that Southeast Asia experienced particularly pronounced popula-

tion growth immediately after the African exodus, meaning that languages in this region should have been least affected by population bottlenecks and would have had the most time to recover diversity.

Although distance from Africa explains much less of the variation in phonemic diversity (19%) than in neutral genetic markers (80 to 85%) (5, 7), the effect is comparable to that obtained from analysis of human mitochondrial DNA (18%) (8) or phenotypic data (14 to 28%) (6). To the extent that language can be taken as an example of cultural evolution more generally, these findings support the proposal that a cultural founder effect operated during our colonization of the globe, potentially limiting the size and cultural complexity of societies at the vanguard of the human expansion (10, 11). An origin of modern languages predating the African exodus 50,000 to 70,000 years ago puts complex language alongside the earliest archaeological evidence of symbolic culture in Africa 80,000 to 160,000 years ago (27, 28). Truly modern language, akin to languages spoken today, may thus have been the key cultural innovation that allowed the emergence of these and other hallmarks of behavioral modernity and ultimately led to our colonization of the globe (29).

References and Notes

1. J. Hay, L. Bauer, *Language* **83**, 388 (2007).
2. C. D. Yang, *Knowledge and Learning in Natural Language* (Oxford Univ. Press, Oxford, 2003).
3. W. Labov, *Principles of Linguistic Change: Social Factors* (Blackwell, Oxford, UK, 2001).
4. S. Ramachandran et al., *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15942 (2005).
5. F. Prugnolle, A. Manica, F. Balloux, *Curr. Biol.* **15**, R159 (2005).
6. A. Manica, W. Amos, F. Balloux, T. Hanihara, *Nature* **448**, 346 (2007).
7. J. Z. Li et al., *Science* **319**, 1100 (2008).
8. F. Balloux, L. J. L. Handley, T. Jombart, H. Liu, A. Manica, *Proc. Biol. Sci.* **276**, 3447 (2009).
9. L. Betti, F. Balloux, W. Amos, T. Hanihara, A. Manica, *Proc. Biol. Sci.* **276**, 809 (2009).
10. J. Diamond, *Nature* **273**, 185 (1978).
11. J. Henrich, *Am. Antiq.* **69**, 197 (2004).
12. S. J. Lycett, N. von Cramon-Taubadel, *J. Archaeol. Sci.* **35**, 553 (2008).
13. D. S. Rogers, M. W. Feldman, P. R. Ehrlich, *Proc. Biol. Sci.* **276**, 3835 (2009).
14. A. J. Baker, P. F. Jenkins, *Anim. Behav.* **35**, 1793 (1987).
15. Materials and methods are available as supporting material on Science Online.
16. B. de Boer, *J. Phonetics* **28**, 441 (2000).
17. L. L. Cavalli-Sforza, M. W. Feldman, *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton Univ. Press, Princeton, NJ, 1981).
18. Q. D. Atkinson, R. D. Gray, *Syst. Biol.* **54**, 513 (2005).
19. Q. D. Atkinson, A. Meade, C. Venditti, S. J. Greenhill, M. Pagel, *Science* **319**, 588 (2008).
20. P. Trudgill, *Linguist. Typol.* **8**, 305 (2004).
21. M. Haspelmath, M. S. Dryer, D. Gil, B. Comrie, Eds., *The World Atlas of Language Structures Online* (Max Planck Digital Library, Munich, 2008).
22. K. P. Burnham, D. R. Anderson, *Model Selection and Inference* (Springer, New York, 1998).
23. P. V. Kirch, R. C. Green, *Curr. Anthropol.* **28**, 431 (1987).
24. L. J. L. Handley, A. Manica, J. Goudet, F. Balloux, *Trends Genet.* **23**, 432 (2007).

25. M. G. Palanichamy *et al.*, *Am. J. Hum. Genet.* **75**, 966 (2004).
26. Q. D. Atkinson, R. D. Gray, A. J. Drummond, *Mol. Biol. Evol.* **25**, 468 (2008).
27. C. Henshilwood, F. d'Errico, M. Vanhaeren, K. van Niekerk, Z. Jacobs, *Science* **304**, 404 (2004).
28. C. W. Marean *et al.*, *Nature* **449**, 905 (2007).
29. I. Tattersall, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16018.
30. Thanks to M. Pagel, O. Curry, R. Dunbar, M. Dunn, R. Gray, S. Greenhill, M. Grove, S. Roberts, R. Ross, and S. Schutz for useful advice and/or comments on the manuscript. I declare no competing financial interest.

SOM Text
Figs. S1 to S8
Tables S1 to S4
References

19 October 2010; accepted 4 March 2011
10.1126/science.1199295

Supporting Online Material

www.sciencemag.org/cgi/content/full/332/6027/346/DC1
Materials and Methods

Interplay Between Changing Climate and Species' Ecology Drives Macroevolutionary Dynamics

Thomas H. G. Ezard,^{1,2*} Tracy Aze,³ Paul N. Pearson,³ Andy Purvis¹

Ecological change provokes speciation and extinction, but our knowledge of the interplay among the biotic and abiotic drivers of macroevolution remains limited. Using the unparalleled fossil record of Cenozoic macroperforate planktonic foraminifera, we demonstrate that macroevolutionary dynamics depend on the interaction between species' ecology and the changing climate. This interplay drives diversification but differs between speciation probability and extinction risk: Speciation was more strongly shaped by diversity dependence than by climate change, whereas the reverse was true for extinction. Crucially, no single ecology was optimal in all environments, and species with distinct ecologies had significantly different probabilities of speciation and extinction. The ensuing macroevolutionary dynamics depend fundamentally on the ecological structure of species' assemblages.

The wide-ranging mechanisms that generate and maintain biodiversity have been grouped in many different ways, but a fundamental distinction exists between biotic and abiotic drivers (1–5). If interactions among species are the dominant drivers of evolution, as in the “Red Queen” model (1, 4), then diversification rates among groups of interacting species are expected to show diversity-dependent dynamics with ongoing turnover at equilibrium. A consequence of this mechanism is that diversification rates are expected to decrease as a function of diversity. Conversely, if evolution is driven chiefly by changes in the physical environment—as in the “Court Jester” model (3), named to contrast with the Red Queen—macroevolutionary dynamics should be dominated by cladewide effects of abrupt abiotic perturbations. Although the interplay between these alternative drivers has long been recognized as fundamental for regulating diversity (6, 7), progress toward understanding their interaction has been slow (2). The incompleteness of the fossil record often necessitates temporally and taxonomically coarse paleontological analyses (7–10), whereas molecular phylogenies are restricted to extant species

and therefore offer little insight into extinction (11, 12). To distinguish how interwoven biotic and abiotic processes regulate diversity, high-

resolution data on multiple forcing mechanisms should be allied to paleontological, species-level phylogenies constructed on sufficiently complete fossil records over substantial periods of evolutionary history. This resolution is rare (2, 6), but Cenozoic macroperforate planktonic foraminifera provide a suitable record for testing these hypotheses (13).

Planktonic foraminifera are sexually reproducing protists distributed throughout the world's oceans. The calcium carbonate “shells” (known as “tests”) of dead individuals rain down on the ocean floor and can, under favorable conditions, generate continuous microfossil sequences that span millions of years. The group's usefulness for stratigraphic correlation (14) and paleoclimatic reconstruction (15) has led to extensive documentation of its morphology (14) and depth habitats (13). The phylogenetic relationships within the macroperforate clade of the Cenozoic have recently been revised comprehensively (13), via the application of Simpson's evolutionary species concept (16). Under this concept, each species is intended to represent a single line of descent (16) that begins with a speciation (clado-

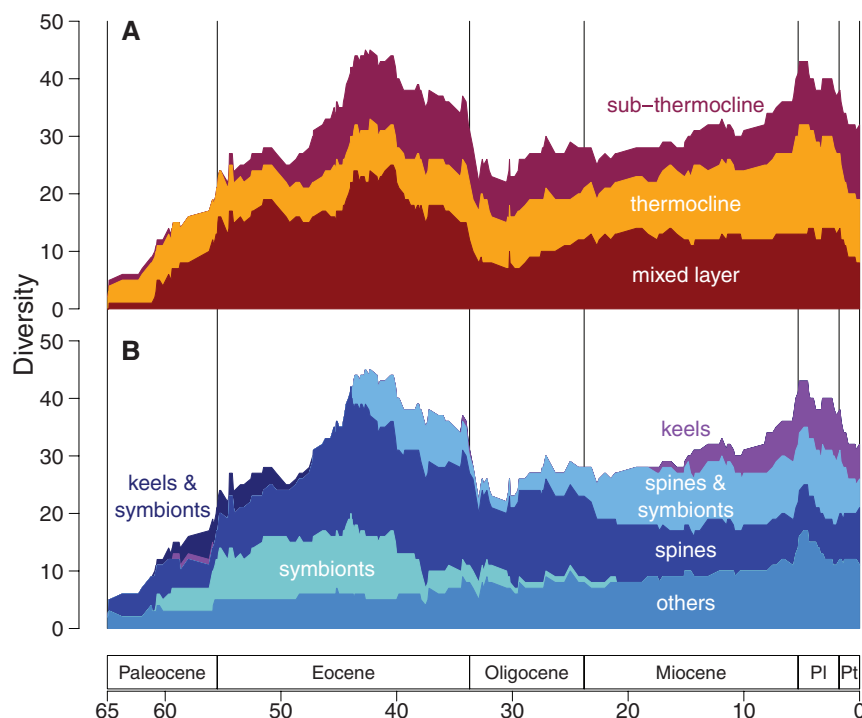


Fig. 1. The relative frequencies of depth habitats (A) and morphologies (B) of macroperforate planktonic foraminifer species across the Cenozoic (13) has fluctuated substantially. Time (million years before the present) is based on the marine geological time scale (14). PI, Pliocene; Pt, Pleistocene.

¹Division of Biology, Silwood Park Campus, Imperial College London, Ascot, Berkshire, SL5 7PY, UK. ²Department of Mathematics, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, GU2 7XH, UK. ³School of Earth and Ocean Sciences, Cardiff University, Cardiff, CF10 3YE, UK.

*To whom correspondence should be addressed. E-mail: t.ezard@surrey.ac.uk