Mapping the Origins and Expansion of the Indo-European Language Family

Remco Bouckaert *et al.*

Science **337**, 957 (2012);
DOI: 10.1126/science.1219669

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

octahedron to truncated tetrahedron, but also induced the growth of a second twin boundary along a neighboring {111} face that is about 72° apart from the first (*10*). Thus, the driving force for the growth of the fourth tip of the tetrahedron is likely rapid growth along two adjacent twin boundaries (Fig. 1D). This conclusion is supported by high-magnification STEM images of individual, fully formed tetrahedra that show distinct lines of contrast along their edges, suggesting the presence of twin planes running parallel to the faces of the tetrahedra, and was further confirmed by electron diffraction studies (figs. S10 to S12) (*24*).

Later stages of the growth pathway outlined in Fig. 1 were probed by increasing the silver/gold ratio in the reaction seeded with Au octahedra. When the number of Au octahedral seeds added to the reaction was reduced (effectively increasing the silver/gold ratio), bimetallic particles with truncated bitetrahedral and even decahedral Ag shells formed (fig. S13). We observed a large dispersity in terms of particle shape in this reaction, similar to what occurred when we used pseudo-spherical seeds. These data indicate that tetrahedra continued to develop twin defects such that a five-fold twinned decahedron could form. STEM images from the growth of the pseudo-spherical seeds revealed bimetallic particles with truncated bitetrahedral, bitetrahedral, truncated decahedral, and decahedral shapes (Fig. 1, E to H). We propose that a tetrahedron can develop a third twin plane, causing a change in shape to a bitetrahedron, and then eventually develop a fourth and fifth twin plane, resulting in the growth of a decahedron.

The lack of synthetic procedures for preparing Au analogs for many of the shapes depicted in Fig. 1, C to G, prevented a closer study of these individual growth steps. However, we have previously studied the plasmon-mediated deposition of Ag onto Au decahedral seeds under nearly identical conditions (*2*). We found that Au five-fold twinned decahedra grew into bimetallic 20-fold twinned icosahedra in a manner similar to the transformations outlined in Fig. 1. These data are consistent with the conclusion that, for this synthetic system, multiply twinned particles formed by successive twinning and that decahedra, regardless of whether they comprise Au or Au-core/Ag-shell structures, can transform into icosahedra through this twinning process (Fig. 1, H to J).

These data show that this particle labeling strategy is particularly useful for elucidating growth pathways when crystal twinning is involved. This method of analysis allows for the discrimination of twin defects inherent to the seed particle and those that develop during the growth of a crystal, essentially distinguishing the potential growth pathways of multiply twinned nanoparticles. This work not only provides valuable insight into the growth mechanisms of multiply twinned structures, which will help to more effectively synthesize such particles in the future, but also demonstrates how nanoparticle labels can be used to effectively track and monitor the growth of nanomaterials in much the same manner that fluorescence and isotopic labeling strategies have been used to study molecular materials. We also anticipate that if this method of analysis is combined with in situ TEM observations (*14*, *16*–*18*), an even greater understanding of nanoparticle growth can be obtained.

References and Notes

1. C. Xue, J. E. Millstone, S. Li, C. A. Mirkin, *Angew. Chem. Int. Ed.* **46**, 8436 (2007).
2. M. R. Langille, J. Zhang, C. A. Mirkin, *Angew. Chem. Int. Ed.* **50**, 3543 (2011).
3. Y. Xia, Y. Xiong, B. Lim, S. E. Skrabalak, *Angew. Chem. Int. Ed.* **48**, 60 (2009).
4. W. Niu *et al.*, *J. Am. Chem. Soc.* **131**, 697 (2009).
5. J. Zeng *et al.*, *J. Am. Chem. Soc.* **132**, 8552 (2010).
6. M. L. Personick, M. R. Langille, J. Zhang, C. A. Mirkin, *Nano Lett.* **11**, 3394 (2011).
7. T. K. Sau, C. J. Murphy, *J. Am. Chem. Soc.* **126**, 8648 (2004).
8. B. Nikoobakht, M. A. El-Sayed, *Chem. Mater.* **15**, 1957 (2003).
9. W. Niu, G. Xu, *Nano Today* **6**, 265 (2011).
10. H. Hofmeister, *Z. Kristallogr.* **224**, 528 (2009).
11. Q. Zhang, J. Xie, Y. Yu, J. Yang, J. Y. Lee, *Small* **6**, 523 (2010).
12. B. Wiley, T. Herricks, Y. Sun, Y. Xia, *Nano Lett.* **4**, 1733 (2004).
13. J. G. Allpress, J. V. Sanders, *Surf. Sci.* **7**, 1 (1967).
14. K. Yagi, K. Takayanagi, K. Kobayashi, G. Honjo, *J. Cryst. Growth* **28**, 117 (1975).
15. H. Hofmeister, *Thin Solid Films* **116**, 151 (1984).
16. N. de Jonge, F. M. Ross, *Nat. Nanotechnol.* **6**, 695 (2011).
17. H. Zheng *et al.*, *Science* **324**, 1309 (2009).
18. J. M. Yuk *et al.*, *Science* **336**, 61 (2012).
19. M. Tsuji *et al.*, *Cryst. Growth Des.* **10**, 296 (2010).
20. R. Jin *et al.*, *Science* **294**, 1901 (2001).
21. R. Jin *et al.*, *Nature* **425**, 487 (2003).
22. C. Xue, G. S. Métraux, J. E. Millstone, C. A. Mirkin, *J. Am. Chem. Soc.* **130**, 8337 (2008).
23. X. Wu *et al.*, *J. Am. Chem. Soc.* **130**, 9500 (2008).
24. See supplementary materials on *Science* Online.
25. F. Kim, S. Connor, H. Song, T. Kuykendall, P. Yang, *Angew. Chem. Int. Ed.* **43**, 3673 (2004).
26. J. Zhou *et al.*, *Langmuir* **24**, 10407 (2008).
27. J. Zhang, M. R. Langille, C. A. Mirkin, *J. Am. Chem. Soc.* **132**, 12502 (2010).
28. B. Pietrobon, V. Kitaev, *Chem. Mater.* **20**, 5186 (2008).
29. J. L. Elechiguerra, J. Reyes-Gasga, M. J. Yacamán, *J. Mater. Chem.* **16**, 3906 (2006).
30. D. J. Smith, A. K. Petford-Long, L. R. Wallenberg, J.-O. Bovin, *Science* **233**, 872 (1986).
31. F. Baletto, R. Ferrando, *Rev. Mod. Phys.* **77**, 371 (2005).

Acknowledgments: This material is based on work supported by the U.S. Air Force Office of Scientific Research; the U.S. Department of Defense National Security Science and Engineering Faculty Fellowships Program/Naval Postgraduate School (award N00244-09-1-0012); the Non-equilibrium Energy Research Center, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award DE-SC0000989; and the NSF Materials Research Science and Engineering Center (MRSEC) program (DMR-1121262) at the Materials Research Center of Northwestern University. The microscopy work was performed in the EPIC facility of the NUANCE Center at Northwestern University, which is supported by NSF Nanoscale Science and Engineering Center, NSF MRSEC, the Keck Foundation, the State of Illinois, and Northwestern University. This work was also supported by the U.S. Air Force Office of Scientific Research through National Defense Science and Engineering graduate fellowship 32 CFR 168a (M.L.P.). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the agency sponsors.

Supplementary Materials

www.sciencemag.org/cgi/content/full/337/6097/954/DC1
Materials and Methods
Figs. S1 to S13
References (*32*, *33*)

5 June 2012; accepted 10 July 2012
10.1126/science.1225653

Mapping the Origins and Expansion of the Indo-European Language Family

Remco Bouckaert,[1] Philippe Lemey,[2] Michael Dunn,[3,4] Simon J. Greenhill,[5,6] Alexander V. Alekseyenko,[7] Alexei J. Drummond,[1,8] Russell D. Gray,[5,9] Marc A. Suchard,[10,11,12] Quentin D. Atkinson[5,13]*

There are two competing hypotheses for the origin of the Indo-European language family. The conventional view places the homeland in the Pontic steppes about 6000 years ago. An alternative hypothesis claims that the languages spread from Anatolia with the expansion of farming 8000 to 9500 years ago. We used Bayesian phylogeographic approaches, together with basic vocabulary data from 103 ancient and contemporary Indo-European languages, to explicitly model the expansion of the family and test these hypotheses. We found decisive support for an Anatolian origin over a steppe origin. Both the inferred timing and root location of the Indo-European language trees fit with an agricultural expansion from Anatolia beginning 8000 to 9500 years ago. These results highlight the critical role that phylogeographic inference can play in resolving debates about human prehistory.

Model-based methods for Bayesian inference of phylogeny have been applied to comparative basic vocabulary data to infer ancestral relationships between languages (*1*–*3*). Such studies have focused on the use of subgrouping and time-depth estimates to test competing hypotheses, but they lack explicit geographic models of language expansion. Here, we used two novel quantitative phylogeographic inference tools derived from stochastic models in evolutionary biology to tackle the "most recalcitrant problem in historical linguistics" (*4*)—the origin of the Indo-European languages. The "steppe hypothesis" posits an origin in the Pontic steppe region north of the Caspian Sea. Although the archaeological record provides a number of candidate expansions from this area (*5*), a steppe homeland is most commonly linked to evidence

of an expansion into Europe and the Near East by Kurgan seminomadic pastoralists beginning 5000 to 6000 years ago (*5*–*7*). Evidence from "linguistic paleontology"—an approach in which terms reconstructed in the ancestral "proto-language" are used to make inferences about its speakers' culture and environment—and putative early borrowings between Indo-European and the Uralic language family of northern Eurasia (*8*) are cited as possible evidence for a steppe homeland (*9*). However, the reliability of inferences derived from linguistic paleontology and claimed borrowings remains uncertain (*5*, *10*). The alternative "Anatolian hypothesis" holds that Indo-European languages spread with the expansion of agriculture from Anatolia (in present-day Turkey), beginning 8000 to 9500 years ago (*11*). Estimates of the age of the Indo-European family derived from models of vocabulary evolution support the chronology implied by the Anatolian hypothesis, but the inferred dates remain controversial (*5*, *10*, *12*), and the implied models of geographic expansion under each hypothesis remain untested.

To test these two hypotheses, we adapted and extended a Bayesian phylogeographic inference framework developed to investigate the origin of virus outbreaks from molecular sequence data (*13*, *14*). We used this approach to analyze a data set of basic vocabulary terms and geographic range assignments for 103 ancient and contemporary Indo-European languages (*15*–*17*). Following previous work that applied Bayesian phylogenetic methods to linguistic data (*1*–*3*), we modeled language evolution as the gain and loss of "cognates" (homologous words) through time (*18*–*20*). We combined phylogenetic inference with a relaxed random walk (RRW) (*14*) model of continuous spatial diffusion along the branches of an unknown, yet estimable, phylogeny to jointly infer the Indo-European language phylogeny and the most probable geographic ranges at the root and internal nodes. This phylogeographic approach treats language location as a continuous vector (longitude and latitude) that evolves through time along the branches of a tree and seeks to infer ancestral locations at internal nodes on the tree while simultaneously accounting for uncertainty in the tree.

To increase the realism of the spatial diffusion, our method extends the RRW process in two ways. First, to reduce potential bias associated with assigning point locations to sampled languages, we use geographic ranges of the languages to specify uncertainty in the location assignments. Second, to account for geographic heterogeneity, we accommodate spatial prior distributions on the root and internal node locations. By assigning zero probability to node locations over water, we can incorporate into the analysis prior information about the shape of the Eurasian landmass.

The estimated posterior distribution for the location of the root of the Indo-European tree under the RRW model is shown in Fig. 1A. The distribution for the root location lies in the region of Anatolia in present-day Turkey. To quantify the strength of support for an Anatolian origin, we calculated the Bayes factors (*21*) comparing the posterior to prior odds ratio of a root location within the hypothesized Anatolian homeland (*11*) (Fig. 1, yellow polygon) with two versions of the steppe hypothesis—the initial proposed Kurgan steppe homeland (*6*) and a later refined hypothesis (*7*) (Table 1). Bayes factors show strong support for the Anatolian hypothesis under a RRW model. This model allows large variation in rates of expansion and so is sufficiently flexible to fit the alternative hypothesis if the data support it. Further, the geographic centroid of the languages considered here falls within the broader steppe hypothesis (Fig. 1, green star), indicating that our model is not simply returning the center of mass of the sampled locations, as would be predicted under a simple diffusion process that ignores phylogenetic information and geographic barriers.

Our results incorporate phylogenetic uncertainty given our data and model and so are not contingent on any single phylogeny. However, phonological and morphological data have been interpreted to support an Indo-European branching structure that differs slightly from the pattern we find, particularly near the base of the tree (*16*). If we constrain our analysis to fit with this alternative pattern of diversification, we find even stronger support for an Anatolian origin (in terms of Bayes factors, $BF_{Steppe\ I} = 216$; $BF_{Steppe\ II} = 227$) (*15*).

[1]Department of Computer Science, University of Auckland, Auckland 1142, New Zealand. [2]Department of Microbiology and Immunology, Rega Institute, KU Leuven, 3000 Leuven, Belgium. [3]Max Planck Institute for Psycholinguistics, Post Office Box 310, 6500 AH Nijmegen, Netherlands. [4]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Kapittelweg 29, 6525 EN Nijmegen, Netherlands. [5]Department of Psychology, University of Auckland, Auckland 1142, New Zealand. [6]School of Culture, History & Language and College of Asia & the Pacific, Australian National University, 0200 Canberra, ACT, Australia. [7]Center for Health Informatics and Bioinformatics, New York University School of Medicine, New York, NY 10016, USA. [8]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland 1142, New Zealand. [9]Department of Philosophy, Research School of the Social Sciences, Australian National University, 0200 Canberra, ACT, Australia. [10]Department of Biomathematics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA. [11]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095, USA. [12]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA. [13]Institute of Cognitive and Evolutionary Anthropology, University of Oxford, Oxford OX2 6PN, UK.

*To whom correspondence should be addressed. E-mail: q.atkinson@auckland.ac.nz

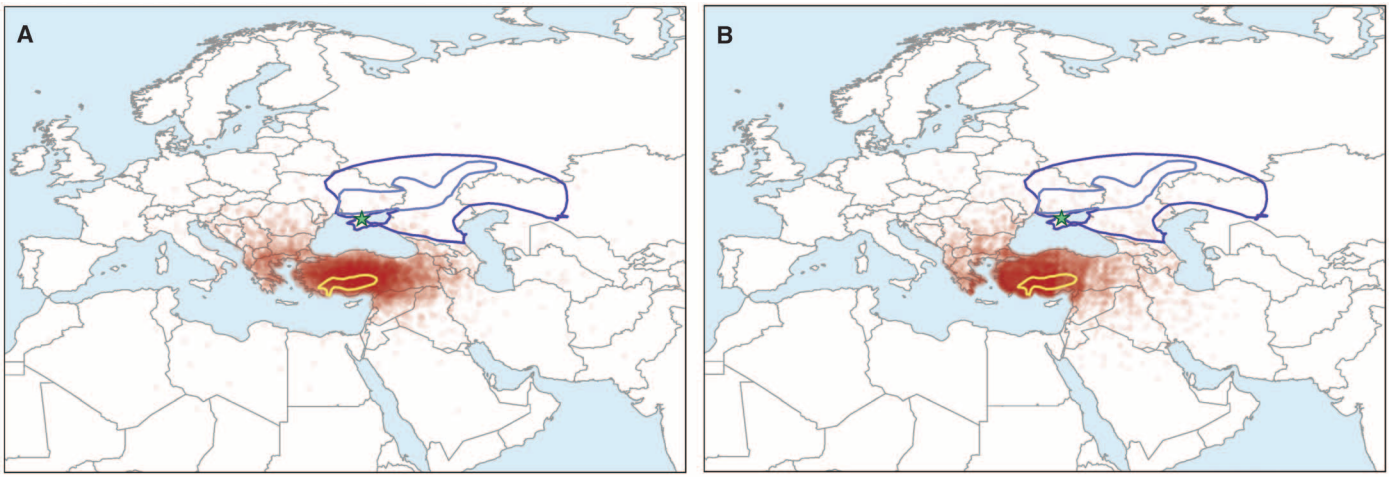

Fig. 1. Inferred geographic origin of the Indo-European language family. (**A**) Map showing the estimated posterior distribution for the location of the root of the Indo-European language tree under the RRW analysis. Markov chain Monte Carlo (MCMC) sampled locations are plotted in translucent red such that darker areas correspond to increased probability mass. (**B**) The same distribution under a landscape-based analysis in which movement into water is less likely than movement into land by a factor of 100 (see fig. S5 for results under the other landscape-based models). The blue polygons delineate the proposed origin area under the steppe hypothesis; dark blue represents the initial suggested Kurgan homeland (*6*) (steppe I), and light blue denotes a later version of the steppe hypothesis (*7*) (steppe II). The yellow polygon delineates the proposed origin under the Anatolian hypothesis (*11*). A green star in the steppe region shows the location of the centroid of the sampled languages.

As the earliest representatives of the main Indo-European lineages, our 20 ancient languages might provide more reliable location information. Conversely, the position of the ancient languages in the tree, particularly the three Anatolian varieties, might have unduly biased our results in favor of an Anatolian origin. We investigated both possibilities by repeating the above analyses separately on only the ancient languages and only the contemporary languages (which excludes Anatolian). Consistent with the analysis of the full data set, both analyses still supported an Anatolian origin (Table 1).

The RRW approach avoids internal node assignments over water, but it does assume, along the unknown tree branches, the same underlying migration rate across water as across land. To investigate the robustness of our results to heterogeneity in rates of spatial diffusion, we developed a second inference procedure that allows migration rates to vary over land and water (*15*). This landscape-based model allows for the inclusion of a more complex diffusion process in which rates of migration are a function of geography. We examined the effect of varying relative rate parameters to represent a range of different migration patterns (*15*). Figure 1B shows the inferred Indo-European homeland under a model in which migration from land into water is less likely than from land to land by a factor of 100. At the other extreme, we fit a "sailor" model with no reluctance to move into water and rapid movement across water. Consistent with the findings based on the RRW model, each of the landscape-based models supports the Anatolian farming theory of Indo-European origin (Table 1).

Our results strongly support an Anatolian homeland for the Indo-European language family. The inferred location (Fig. 1) and timing [95% highest posterior density (HPD) interval, 7116 to 10,410 years ago] of Indo-European origin is congruent with the proposal that the family began to diverge with the spread of agriculture from

Table 1. Bayes factors comparing support for the Anatolian and steppe hypotheses. We estimated Bayes factors directly, using expectations of a root model indicator function taken over the MCMC samples drawn from the posterior and prior of each hypothesis. Bayes factors greater than 1 favor an Anatolian origin. A Bayes factor of 5 to 20 is taken as substantial support, greater than 20 as strong support, and greater than 100 as decisive (*30*).

Phylogeographic analysis	Bayes factor	
	Anatolian vs. steppe I	Anatolian vs. steppe II
RRW: All languages	175.0	159.3
RRW: Ancient languages only	1404.2	1582.6
RRW: Contemporary languages only	12.0	11.4
Landscape aware: Diffusion	298.2	141.9
Landscape aware: Migration from land into water less likely than from land to land by a factor of 10	197.7	92.3
Landscape aware: Migration from land into water less likely than from land to land by a factor of 100	337.3	161.0
Landscape aware: Sailor	236.0	111.7

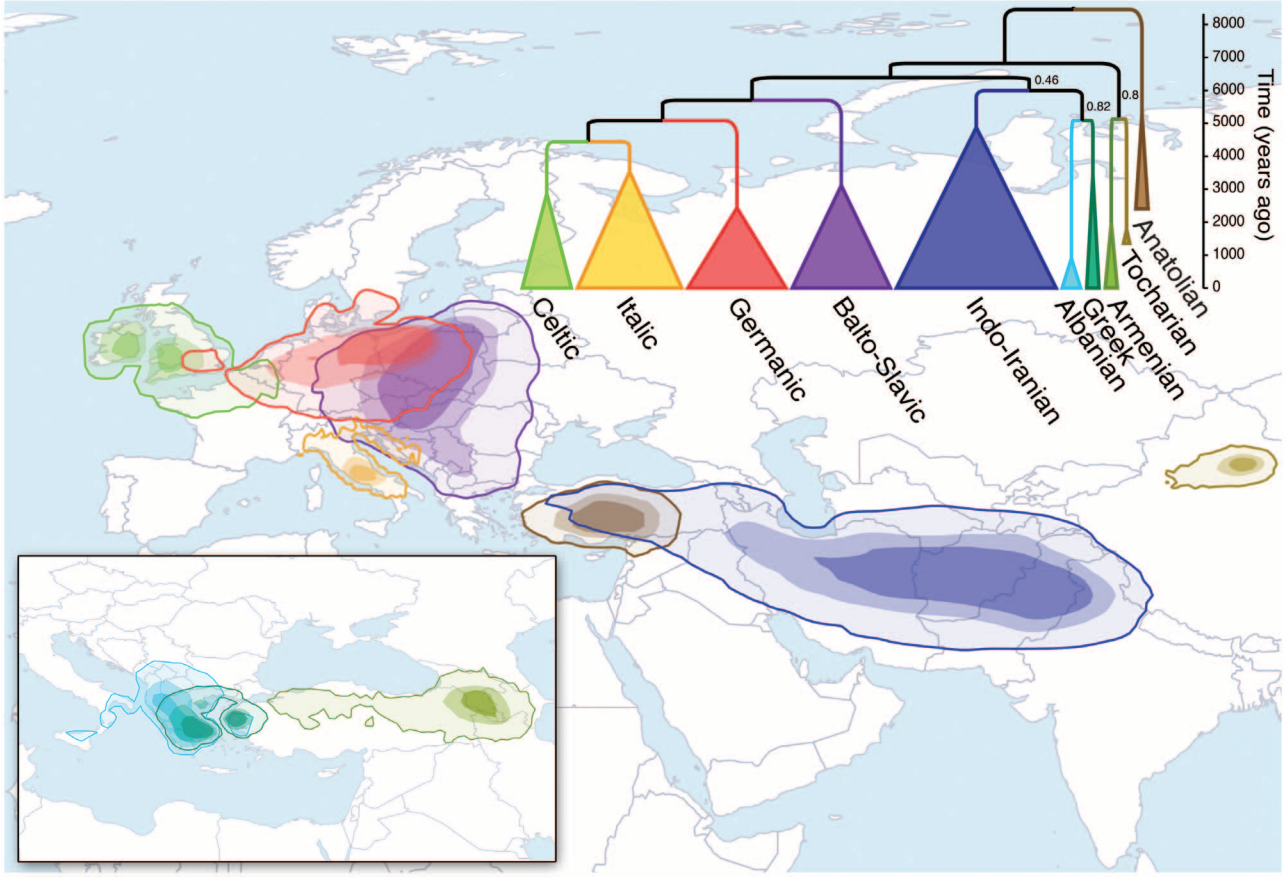

Fig. 2. Map and maximum clade credibility tree showing the diversification of the major Indo-European subfamilies. The tree shows the timing of the emergence of the major branches and their subsequent diversification. The inferred location at the root of each subfamily is shown on the map, colored to match the corresponding branches on the tree. Albanian, Armenian, and Greek subfamilies are shown separately for clarity (**inset**). Contours represent the 95% (largest), 75%, and 50% HPD regions, based on kernel density estimates (*15*).

Anatolia 8000 to 9500 years ago (*11*). In addition, the basal relationships in the tree (Fig. 2, inset, and figs. S1 and S2) and geographic movements these imply are also consistent with archaeological evidence for an expansion of agriculture into Europe via the Balkans, reaching the edge of western Europe by 5000 years ago (*22*). This scenario fits with genetic (*23*–*25*) and craniometric (*26*) evidence for a Neolithic, Anatolian contribution to the European gene pool. An expansion of Indo-European languages with agriculture is also in line with similar explanations for language expansion in the Pacific (*2*), Southeast Asia (*27*), and sub-Saharan Africa (*28*), adding weight to arguments for the key role of agriculture in shaping global linguistic diversity (*4*).

Despite support for an Anatolian Indo-European origin, we think it unlikely that agriculture serves as the sole driver of language expansion on the continent. The five major Indo-European subfamilies—Celtic, Germanic, Italic, Balto-Slavic, and Indo-Iranian—all emerged as distinct lineages between 4000 and 6000 years ago (Fig. 2 and fig. S1), contemporaneous with a number of later cultural expansions evident in the archaeological record, including the Kurgan expansion (*5*–*7*). Our inferred tree also shows that within each subfamily, the languages we sampled began to diversify between 2000 and 4500 years ago, well after the agricultural expansion had run its course. Figure 2 plots the inferred geographic origin of languages sampled from each subfamily under the RRW model. The interpretation of these results is straightforward when all the main branches of a subfamily are represented in the sample. In cases where there are branches not represented, such as Continental Celtic, the inferred time depths and locations may not correspond to the origin of all known languages in a subfamily. Because we know that the Romance languages in our sample are descended from Latin, this group presents a useful test case of our methodology. Our model correctly assigns high posterior support to the most recent common ancestor of contemporary Romance languages around Rome (fig. S3). Using this approach, we may therefore be able to test between more recent origin hypotheses pertaining to individual subgroups. Moreover, by combining the time-depth and location estimates across all internal nodes, we can generate a picture of the expansion of all Indo-European languages across the landscape (fig. S4 and movie S1).

Language phylogenies provide insights into the cultural history of their speakers (*1*–*3*, *28*, *29*). Our analysis of ancient and contemporary Indo-European languages shows that these insights can be made even more powerful by explicitly incorporating spatial information. Linguistic phylogeography enables us to locate cultural histories in space and time and thus provides a rigorous analytic framework for the synthesis of archaeological, genetic, and cultural data.

References and Notes

1. R. D. Gray, Q. D. Atkinson, *Nature* **426**, 435 (2003).
2. R. D. Gray, A. J. Drummond, S. J. Greenhill, *Science* **323**, 479 (2009).
3. A. Kitchen, C. Ehret, S. Assefa, C. J. Mulligan, *Proc. R. Soc. B* **276**, 2703 (2009).
4. J. Diamond, P. Bellwood, *Science* **300**, 597 (2003).
5. J. P. Mallory, D. Q. Adams, *The Oxford Introduction to Proto Indo European and the Proto Indo European World* (Oxford Univ. Press, New York, 2006).
6. M. Gimbutas, in *Indo-European and Indo-Europeans,* G. Cardona, H. M. Hoenigswald, A. Senn, Eds. (Univ. of Pennsylvania Press, Philadelphia, 1970), pp. 155–197.
7. M. Gimbutas, *J. Indo-Europ. Stud.* **5**, 277 (1977).
8. J. Koivulehto, in *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations*, C. Carpelan, A. Parpola, P. Koskikallio, Eds. (Suomalais-Ugrilaisen Seuran Toimituksia, Helsinki, 2001), pp. 235–263.
9. D. W. Anthony, *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World* (Princeton Univ. Press, Princeton, NJ, 2007).
10. P. Heggarty, in *Phylogenetic Methods and the Prehistory of Languages*, P. Forster, C. Renfrew, Eds. (McDonald Institute for Archaeological Research, Cambridge, 2006), pp. 183–194.
11. C. Renfrew, *Archaeology and Language: The Puzzle of Indo-European Origins* (Cape, London, 1987).
12. M. Balter, *Science* **302**, 1490 (2003).
13. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, *PLOS Comput. Biol.* **5**, e1000520 (2009).
14. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, *Mol. Biol. Evol.* **27**, 1877 (2010).
15. See supplementary materials on *Science* Online.
16. D. Ringe, T. Warnow, A. Taylor, *Trans. Philol. Soc.* **100**, 59 (2002).
17. I. Dyen, J. B. Kruskal, P. Black, *Trans. Am. Philos. Soc.* **82**, 1 (1992).
18. G. K. Nicholls, R. D. Gray, in *Phylogenetic Methods and the Prehistory of Languages*, P. Forster, C. Renfrew, Eds. (McDonald Institute for Archaeological Research, Cambridge, 2006), pp. 161–172.
19. A. V. Alekseyenko, C. J. Lee, M. A. Suchard, *Syst. Biol.* **57**, 772 (2008).
20. A. J. Drummond, M. A. Suchard, D. Xie, A. Rambaut, *Mol. Biol. Evol.* **29**, 1969 (2012).
21. M. A. Suchard, R. E. Weiss, J. S. Sinsheimer, *Mol. Biol. Evol.* **18**, 1001 (2001).
22. M. Gkiasta, T. Russell, S. Shennan, J. Steele, *Antiquity* **77**, 45 (2003).
23. L. Chikhi, *Hum. Biol.* **81**, 639 (2009).
24. W. Haak *et al.*, *PLoS Biol.* **8**, e1000536 (2010).
25. M. Lacan *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9788 (2011).
26. N. von Cramon-Taubadel, R. Pinhasi, *Proc. R. Soc. B* **278**, 2874 (2011).
27. I. Glover, C. Higham, in *The Origins and Spread of Agriculture and Pastoralism in Eurasia*, D. Harris, Ed. (Blackwell, Cambridge, 1996), pp. 413–441.
28. C. J. Holden, *Proc. R. Soc. B* **269**, 793 (2002).
29. T. E. Currie, S. J. Greenhill, R. D. Gray, T. Hasegawa, R. Mace, *Nature* **467**, 801 (2010).
30. R. E. Kass, A. E. Raftery, *J. Am. Stat. Assoc.* **90**, 773 (1995).

Acknowledgments: We thank the New Zealand Phylogenetics Meeting and the National Evolutionary Synthesis Center (NESCent) NSF grant EF-0423641, for fostering collaboration on this project. Supported by the Marsden Fund (R.B., R.D.G., S.J.G., and A.J.D.), Rutherford Discovery Fellowships (Q.D.A., A.J.D.), administered by the Royal Society of New Zealand, and by NIH grants R01 GM086887 and R01 HG006139 (M.A.S.). The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 278433-PREDEMICS and European Research Council (ERC) grant agreement 260864.

Supplementary Materials

www.sciencemag.org/cgi/content/full/337/6097/957/DC1
Materials and Methods
Figs. S1 to S12
Tables S1 to S5
References (*31*–*62*)
Movie S1
BEAST input file
NEXUS tree file

26 January 2012; accepted 1 June 2012
10.1126/science.1219669

Assembly of an Evolutionarily New Pathway for α-Pyrone Biosynthesis in *Arabidopsis*

Jing-Ke Weng,[1]* Yi Li,[1] Huaping Mo,[2] Clint Chapple[1]†

Plants possess arrays of functionally diverse specialized metabolites, many of which are distributed taxonomically. Here, we describe the evolution of a class of substituted α-pyrone metabolites in *Arabidopsis*, which we have named arabidopyrones. The biosynthesis of arabidopyrones requires a cytochrome P450 enzyme (CYP84A4) to generate the catechol-substituted substrate for an extradiol ring-cleavage dioxygenase (AtLigB). Unlike other ring-cleavage–derived plant metabolites made from tyrosine, arabidopyrones are instead derived from phenylalanine through the early steps of phenylpropanoid metabolism. Whereas CYP84A4, an *Arabidopsis*-specific paralog of the lignin-biosynthetic enzyme CYP84A1, has neofunctionalized relative to its ancestor, AtLigB homologs are widespread among land plants and many bacteria. This study exemplifies the rapid evolution of a biochemical pathway formed by the addition of a new biological activity into an existing metabolic infrastructure.

As sessile organisms, land plants evolved the ability to synthesize specialized metabolites that are key to their adaptation to terrestrial ecosystems (*1*). The specialized metabolic pathways in plants typically comprise multiple catalytic steps that are spatially and temporally regulated and range from being widespread across land plants examined to date to lineage-specific (*2*). For example, flavonoids are ubiquitous in land plants, but the anticancer drug taxol is made only in certain yew species (*3*). The latter observation, and others like it, suggest that