Published in final edited form as:
Science. 2012 August 24; 337(6097): 957–960. doi:10.1126/science.1219669.

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], **Alexander V. Alekseyenko**[6], **Alexei J. Drummond**[1,7], **Russell D. Gray**[5], **Marc A. Suchard**[8,9,10], and **Quentin Atkinson**[5,11,*]

[1]Department of Computer Science, University of Auckland, Auckland, New Zealand [2]Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium [3]Max Planck Institute for Psycholinguistics, Post Office Box 310, 6500 AH Nijmegen, The Netherlands [4]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Kapittelweg 29, 6525 EN Nijmegen, The Netherlands [5]Department of Psychology, University of Auckland, Auckland 1142, New Zealand [6]Department of Medicine, New York University, New York, USA [7]Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand [8]Department of Biomathematics, David Geffen School of Medicine at UCLA, Los Angeles, USA [9]Department of Biostatistics, UCLA School of Public Health, Los Angeles, USA [10]Department of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, USA [11]Institute of Cognitive and Evolutionary Anthropology, University of Oxford, 64 Banbury Road, Oxford, OX2 6PN, United Kingdom

Abstract

There are two competing hypotheses for the origin of the Indo-European language family. The conventional view places the homeland in the Pontic steppes approximately 6kya. An alternative hypothesis claims the languages spread from Anatolia with the expansion of farming 8–9.5kya. Here we use 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 between the homeland hypotheses. We find decisive support for an Anatolian 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 in the 9th millennium BP. These results highlight the critical role phylogeographic inference can play in resolving longstanding debates about human prehistory.

*Correspondence to: q.atkinson@auckland.ac.nz.

Supplementary Materials:
Materials and Methods
References (31–63)
Figures S1–S12
Tables S1–S5
Additional Materials:
Google Earth KML - movie showing the expansion of the Indo-European languages through time. Contours on the map represent the 95% highest posterior density distribution of the range of Indo-European.
BEAST XML file – BEAST script from the full relaxed random walk analysis, including age constraints, location data and cognate information.

All data is available as Supporting Online Material.

Model-based methods for Bayesian inference of phylogeny have been applied to comparative basic vocabulary data to infer ancestral relationships between languages (1–3). This work has focussed on the use of sub-grouping and time-depth estimates to test competing hypotheses but lacks explicit geographic models of language expansion. Here we use 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 steppes region north of the Caspian Sea. Whilst 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' semi-nomadic pastoralists beginning in the sixth millennium BP (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 argued to support a steppe homeland (9). However, the reliability of inferences derived from linguistic paleontology and claimed borrowings remain controversial (5, 10). The alternative, 'Anatolian hypothesis' holds that Indo-European languages spread with the expansion of agriculture from Anatolia (in present-day Turkey), beginning 8000–9500 years BP (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 between an Anatolian and steppe Indo-European homeland, we adapted and extended a Bayesian phylogeographic inference framework developed to investigate the origin of virus outbreaks from molecular sequence data (13, 14). We use 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 applying Bayesian phylogenetic methods to linguistic data (1–3), we model language evolution as the gain and loss of 'cognates' (homologous words) through time (18–20). We combine phylogenetic inference with a relaxed random walk (14) (RRW) 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. In order to increase the realism of the spatial diffusion, we extend the RRW process in two important and novel 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 Figure 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 (BFs) (21) comparing the posterior to prior odds ratio of a root location within the hypothesized Anatolian homeland (11) (yellow polygon, Figure 1) with two versions of the steppe hypothesis – the initial proposed Kurgan steppe homeland (6) (dark blue) and a later refined hypothesis (7) (light blue) (Table 1). BFs show strong support for the Anatolian hypothesis under a RRW model. The geographic centroid of the languages considered here falls within the broader Steppe hypothesis (green star, Figure 1), 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 argued 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 ($BF_{SteppeI}$=216; $BF_{SteppeII}$=227) (15).

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 investigate 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 dataset, both analyses still favor an Anatolian origin (Table 1).

The RRW approach avoids internal node assignments over water, but assumes, along the unknown tree branches, the same underlying migration rate across water as 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 100 times less likely than from land to land. 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).

Overall, our results strongly support an Anatolian homeland for the Indo-European language family. The inferred location (Figure 1) and timing (7,116–10,410 BP 95% highest posterior density [HPD]) of Indo-European origin is congruent with the proposal that the family began to diverge with the spread of agriculture from Anatolia 8,000–9,500 years BP (11). In addition, the basal relationships in the tree (Figure 2 inset; Figure S1 & 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 BP (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), South East 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 BP (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 BP, 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. Since we know 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 Supplementary KML file). 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the New Zealand Phylogenetics Meeting and the National Evolutionary Synthesis Center (NESCent), NSF EF-0423641 for fostering collaboration on this project. RB, RG, SG and AJD are partially supported by the Marsden Fund. MAS is partially supported by NIH R01 GM086887 and R01 HG006139. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) under Grant Agreement n°278433-PREDEMICS and ERC Grant agreement no. 260864.

References and Notes

1. Gray RD, Atkinson QD. Nature. Nov 27.2003 426:435. [PubMed: 14647380]
2. Gray RD, Drummond AJ, Greenhill SJ. Science. 2009; 323:479. [PubMed: 19164742]
3. Kitchen A, Ehret C, Assefa S, Mulligan CJ. Proceedings of the Royal Society B-Biological Sciences. 2009; 276:2703.
4. Diamond J, Bellwood P. Science. 2003; 300:597. [PubMed: 12714734]
5. Mallory, JP.; Adams, DQ. The Oxford introduction to Proto Indo European and the Proto Indo European world. Oxford University Press; New York: 2006. p. xxivp. 731
6. Gimbutas, M. Indo-European and Indo-Europeans. Cardona, G.; Hoenigswald, HM.; Senn, A., editors. University of Pennsylvania Press; Philadelphia: 1970. p. 155-97.
7. Gimbutas M. Journal of Indo-European Studies. 1977; 5:277.
8. Koivulehto, J. Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations. Carpelan, C.; Parpola, A.; Koskikallio, P., editors. Vol. 242. SUST (Suomalais-Ugrilaisen Seuran toimituksia); Helsinki: 2001. p. 235-263.
9. Anthony, DW. The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world. Princeton University Press; Princeton, NJ: 2007.
10. Heggarty, P. Phylogenetic methods and the prehistory of languages. Forster, P.; Renfrew, C., editors. McDonald Institute for Archaeological Research; Cambridge: 2006. p. 183-194.
11. Renfrew, C. Archaeology and language: the puzzle of Indo-European origins. J. Cape; London: 1987. p. xivp. 346[8] p. of plates
12. Balter M. Science. 2003; 302:1490. [PubMed: 14645820]
13. Lemey P, Rambaut A, Drummond AJ, Suchard MA. PLoS Computational Biology. 2009; 5
14. Lemey P, Rambaut A, Welch JJ, Suchard MA. Molecular Biology and Evolution. 2010; 27:1877. [PubMed: 20203288]
15. Materials and methods are available as supporting material on Science Online.
16. Ringe D, Warnow T, Taylor A. Transactions of the Philological Society. 2002; 100:59.
17. Dyen I, Kruskal JB, Black P. Transactions of the American Philosophical Society. 1992; 82:1.
18. Nicholls, GK.; Gray, RD. Phylogenetic Methods and the Prehistory of Languages. Clackson, J.; Forster, P.; Renfrew, C., editors. The McDonald Institute for Archaeological Research; Cambridge: 2006. p. 161-172.
19. Alekseyenko A, Lee C, Suchard MA. Systematic Biology. 2008; 57:772. [PubMed: 18853363]
20. Drummond, AJ.; Suchard, MA.; Xie, D.; Rambaut, A. Molecular Biology and Evolution. 2012.
21. Suchard MA, Weiss RE, Sinsheimer JS. Molecular Biology and Evolution. 2001; 18:1001. [PubMed: 11371589]
22. Gkiasta M, Russell T, Shennan S, Steele J. Antiquity. Mar.2003 77:45.
23. Chikhi L. Human Biology. 2009; 81:639. [PubMed: 20504188]
24. Haak W, et al. PLoS Biol. 2010; 8:e1000536. [PubMed: 21085689]
25. Lacan M, et al. Proceedings of the National Academy of Science USA. 2011; 108:9788.
26. von Cramon-Taubadel N, Pinhasi R. Proceedings of the Royal Society B-Biological Sciences. 2011; 278:2874.
27. Glover, I.; Higham, C. The Origins and Spread of Agriculture and Pastoralism in Eurasia. Harris, D., editor. Blackwell; Cambridge: 1996.
28. Holden CJ. Proc Biol Sci. Apr 22.2002 269:793. [PubMed: 11958710]
29. Currie TE, Greenhill SJ, Gray RD, Hasegawa T, Mace R. Nature. 2010; 467:801. [PubMed: 20944739]
30. Kass RE, Raftery AE. Journal of the American Statistical Association. 1995; 90:773.

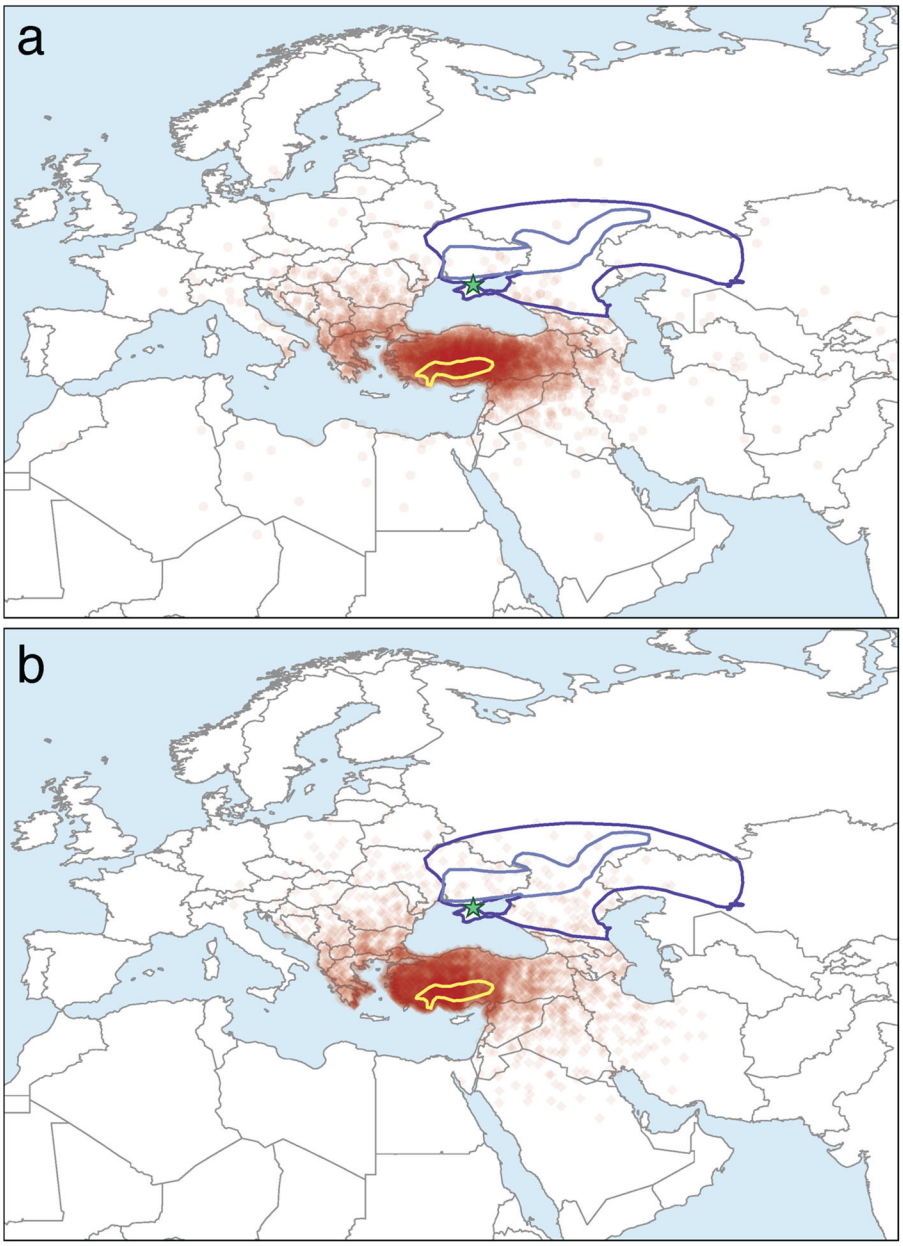

Fig. 1.

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. 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 100 times less likely than movement into land (see Fig. S5 for results under the other landscape based models). The blue polygons delineate the proposed origin area under the Steppes hypothesis – dark blue shows the initial suggested homeland (6) and light blue shows a later version of the Steppes hypothesis (7). 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.

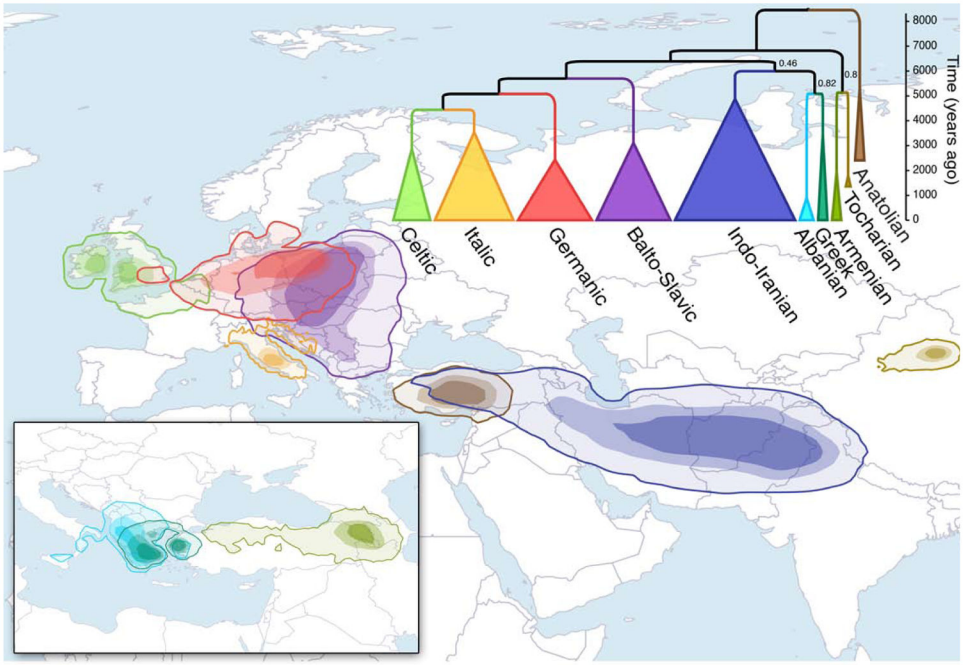

Fig. 2.

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. For clarity, Albanian, Armenian and Greek subfamilies are shown separately (inset). Contours represent the 95% (largest), 75% and 50% HPD regions, based on kernel density estimates (15).

Table 1

Bayes factors comparing support for the Anatolian and Steppes hypotheses

We estimated BFs directly using expectations of a root model indicator function taken over the MCMC samples drawn from the posterior and prior of each hypothesis. BFs greater than 1 favour an Anatolian origin. A BF of 5–20 is taken as substantial support, over 20 as strong support and BFs greater than 100 are considered decisive (30).

	Bayes Factor
RRW: all languages	
Anatolian vs. Steppe I	175
Anatolian vs. Steppe II	159.3
RRW: ancient languages only	
Anatolian vs. Steppe I	1404.2
Anatolian vs. Steppe II	1582.6
RRW: contemporary languages only	
Anatolian vs. Steppe I	12.0
Anatolian vs. Steppe II	11.4
Landscape aware: Diffusion	
Anatolian vs. Steppe I	298.2
Anatolian vs. Steppe II	141.9
Landscape aware: 10x less likely into water	
Anatolian vs. Steppe I	197.7
Anatolian vs. Steppe II	92.3
Landscape aware: 100x less likely into water	
Anatolian vs. Steppe I	337.3
Anatolian vs. Steppe II	161
Landscape aware: Sailor	
Anatolian vs. Steppe I	236
Anatolian vs. Steppe II	111.7