Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature, 426, 435-439. DOI: 10.1038/nature02029. [Available at http://www.nature.com].

Language tree divergence times support the Anatolian theory of Indo-European origin

Russell D. Gray & Quentin D. Atkinson

Department of Psychology, University of Auckland, Private Bag 92019,
Auckland 1020, NEW ZEALAND

Languages, like genes, provide vital clues about human history[1,2]. The origin of the Indo-European language family is 'the most intensively studied, yet still most recalcitrant, problem of historical linguistics'[3]. Numerous genetic studies of Indo-European origins have also produced inconclusive results[4,5,6]. Here we analyse linguistic data using computational methods derived from evolutionary biology. We test between two theories of Indo-European origin – the 'Kurgan expansion' and 'Anatolian farming' hypotheses. The former centres on possible archaeological evidence for an expansion into Europe and the near-East by Kurgan horsemen beginning in the sixth millennium BP[7,8]. The latter claims that Indo-European languages expanded with the spread of agriculture from Anatolia around 8,000 to 9,500BP[9]. In striking agreement with the Anatolian hypothesis, our analysis of a matrix of 87 languages with 2,449 lexical items produced an estimated age range for the initial Indo-European divergence of between 7,800BP and 9,800BP. The results were robust to changes in coding procedures, calibration points, rooting of the trees and priors in the Bayesian analysis.

Historical linguists traditionally use the 'comparative method' to construct language family trees from discrete lexical, morphological and phonological data. Unfortunately, whilst the comparative method can provide a *relative* chronology, it cannot provide *absolute* date estimates. A derivative of lexicostatistics, glottochronology, is an alternative, distance-based approach to language tree construction that enables absolute dates to be estimated[10]. Glottochronology uses the percentage of shared 'cognates' between languages to calculate divergence times by assuming a constant rate of lexical replacement or 'glottoclock'. Cognates are words inferred to have a common historical origin because of systematic sound correspondences and clear similarities in form and meaning. Despite some initial enthusiasm, the method has been heavily criticised and is now largely discredited[11,12]. Criticisms of glottochronology, and distance-based methods in general, tend to fall into four main categories: first, by summarizing cognate data into percentage scores, much of the information in the discrete character data is lost, greatly reducing the power of the method to reconstruct evolutionary history accurately[13]; second, the clustering methods employed tend to produce inaccurate trees when lineages evolve at different rates, grouping together languages that evolve slowly rather than languages that share a recent common ancestor[12,14]; third, substantial borrowing of lexical items between languages makes tree-based methods inappropriate; and fourth, the assumption of a strict *glottoclock* rarely holds, making date estimates unreliable[11]. For these reasons historical linguists have generally abandoned efforts to estimate absolute ages. Dixon[15] epitomizes this view with his assertion that, based on the linguistic data, the age of Indo-European 'could be anything – 4,000 years BP or 40,000 years BP are both perfectly possible (as is any date in between)'.

Recent advances in computational phylogenetic methods, however, provide possible solutions to the four main problems faced by glottochronology. First, the problem of information loss that comes from converting discrete characters into distances can be overcome by analysing the discrete characters themselves to find the optimal tree(s). Second, the accuracy of tree topology and branch length estimation can be improved by using explicit likelihood models of evolution. Maximum likelihood methods generally outperform distance and parsimony approaches in situations where there are unequal rates of change[14]. Moreover, uncertainty in the estimation of tree topology, branch lengths and parameters of the evolutionary model can be estimated using Bayesian Markov Chain Monte Carlo[16] (MCMC) methods in which the frequency distribution of the sample approximates the posterior probability distribution of the trees[17]. All subsequent analyses can then incorporate this uncertainty. Third, lexical items that are obvious borrowings can be removed from the analysis, and computational methods such as split decomposition[18], which do not force the data to fit a tree model, can be used to check for non-treelike signals in the data. Finally, the assumption of a strict clock can be relaxed by using rate smoothing algorithms to model rate variation across the tree. The Penalized-Likelihood[19] model allows for rate variation between lineages whilst incorporating a 'roughness penalty' that penalizes changes in rate from branch to branch. This smoothes inferred rate variation across the tree so that the age of any node can be estimated even under conditions of rate heterogeneity.

We applied likelihood models of lexical evolution, Bayesian inference of phylogeny, and rate smoothing algorithms to a matrix of 87 Indo-European languages with 2,449 cognate sets coded as discrete binary characters. This coding was based on Dyen *et al.*'s Indo-European database[20] with the addition of three extinct languages. Examining subsets of languages using split decomposition revealed a strong tree-like signal in the data, and a preliminary parsimony analysis produced a consistency index of 0.48 and retention index of 0.76, well above what would be expected from biological data sets of a similar size[21]. The consensus tree from an initial analysis is shown in figure 1a. The topology of the tree is consistent with the traditional Indo-European language groups[22]. All of these groups are monophyletic and supported by high posterior probability values. Recent parsimony and compatibility analyses have also supported these groupings, as well as a Romano-Germano-Celtic supergroup, the early divergence of Greek and Armenian lineages[23], and the basal position of Tocharian[24]. The consensus tree also reflects traditional uncertainties in the relationships between the major Indo-European language groups. For instance, historical linguists have not resolved the position of the Albanian group and our results clearly reflect this uncertainty (the posterior probability of the Albanian/Indo-Iranian group is only 0.36).

One major advantage of the Bayesian MCMC approach is that any inferences are not contingent upon a specific tree topology. Trees are sampled in proportion to their posterior probability, providing a direct measure of uncertainty in the tree topology and branch-length estimates. By estimating divergence times across the MCMC sample distribution of trees we can explicitly account for variability in the age estimates due to phylogenetic uncertainty, and hence calculate a confidence interval for the age of any node. We estimated divergence times by constraining the age of 14 nodes on each tree in accordance with historically attested events (see supplementary material). We then used penalised-likelihood rate-smoothing to calculate divergence times without the assumption of rate constancy[19]. Another advantage of the Bayesian framework is that prior knowledge about language relationships can be incorporated into the analysis. To ensure that the sample was consistent with well-established linguistic relationships, we filtered the 10,000-tree sample using a constraint tree (see caption, figure 1b). We used the resulting distribution of 3,500 basal divergence time estimates to create a confidence interval for the age of the Indo-European language family (figure 1b).

A key part of any Bayesian phylogenetic analysis is an assessment of the robustness of the inferences. One important potential cause of error is cognacy judgements. In the initial analysis we included all cognate sets in the Dyen *at al.* database[20] in an effort to maximise phylogenetic signal. To assess the impact of different levels of stringency in the cognacy judgements we repeated the analysis with all cognate sets identified by Dyen *et al.* as ‘doubtful’ removed. ‘Doubtful cognates’ (for instance possible chance similarities) could falsely increase similarities between languages and thus lead to an underestimate of the divergence times. Unrecognised borrowing between closely related languages would have a similar effect. Conversely, borrowing between distantly related languages will falsely inflate branch-lengths at the base of the tree and thus increase divergence time estimates. With the doubtful cognates removed, the conservative coding lead to a similar estimate of Indo-European language relationships to that produced using the original coding. The relationships within each of the 11 major groups were unchanged. Only the placement of the weakly supported basal branches differed (see figure 1c). More significantly, the divergence time estimates increased, suggesting that the effects of chance similarities and unrecognised borrowings between closely related languages may have outweighed those of borrowings between distantly related languages. In other words, our initial analysis is likely to have underestimated the age of Indo-European.

The constraint tree used to filter the MCMC sample of trees also contained assumptions about Indo-European history that may have biased the results. We therefore repeated the analyses using a more relaxed set of constraints (see caption, figure 1d). This produced a divergence time distribution and consensus tree almost identical to the original sample distribution (see figure 1d).

Another potential bias lay in the initial coding procedure that made no allowance for missing cognate information. The languages at the base of the tree (Hittite and Tocharian A and B) may appear to lack cognates found in other languages because our knowledge of these extinct languages is limited to reconstructions from ancient texts. This uneven sampling may have increased basal branch-lengths and thus inflated divergence time estimates. We tested this possibility by recoding apparently absent cognates as uncertainties (absent or present) and rerunning the analyses. Whilst divergence time estimates decreased slightly, the effect was only small (see figure 1e).

Finally, although there is considerable support for Hittite (an extinct Anatolian language) as the most appropriate root for Indo-European[22 ,23], rooting the tree with Hittite could be claimed to bias the analysis in favour of the Anatolian hypothesis. We thus reran the analysis using the consensus tree in figure 1 rooted with Balto-Slavic, Greek and Indo-Iranian as outgroups. This *increased* the estimated divergence time from 8,700BP to 9,600, 9,400 and 10,100BP respectively.

The pattern and timing of expansion suggested by the four analyses in figure 1 is consistent with the Anatolian farming theory of Indo-European origin. Radiocarbon analysis of the earliest Neolithic sites across Europe suggest that agriculture arrived in Greece at some time during the ninth millennium BP and had reached as far as Scotland by 5,500BP[25]. Figure 1 shows the Hittite lineage diverging from Proto-Indo-European around 8,700BP, perhaps reflecting the initial migration out of Anatolia. Tocharian, and the Greco-Armenian lineages are shown as distinct by 7,000BP, with all other major groups formed by 5,000BP. This scenario is consistent with recent genetic studies supporting a Neolithic, Near Eastern contribution to the European gene-pool[4,6]. The consensus tree also shows evidence of a rapid period of divergence giving rise to the Italic, Celtic, Balto-Slavic and perhaps Indo-Iranian families, that is intriguingly close to the time suggested for a possible Kurgan expansion. Thus, as Cavalli Sforza *et al.*[26] observed, these hypotheses need not be mutually exclusive.

Phylogenetic methods have revolutionised evolutionary biology over the last 20 years and are now starting to take hold in other areas of historical inference[2,23,24,27,28,29]. The model-based Bayesian framework employed in this paper offers several advantages over previous applications of computational methods to language phylogenies. This approach allowed us to: - identify sections in the language tree that were poorly supported; explicitly incorporate this uncertainty in tree typology and branch length estimates in our analysis; test the possible effects of borrowing, chance similarities, and Bayesian priors on our analysis; and estimate divergence times without the assumption of a strict glottoclock. The challenge of making accurate inferences about human history is an extremely demanding one, requiring the integration of archaeological, genetic, cultural and linguistic data. The combination of computational phylogenetic methods and lexical data to test archaeological hypotheses is a step forward in this challenging and fascinating task.

Method

Data and coding. Data were sourced from Dyen, Kruskal and Black's comparative Indo-European database[20]. The database records word forms and cognacy judgments in 95 languages across the 200 items in the Swadesh word list. This list consists of items of basic vocabulary such as pronouns, numerals and body-parts that are known to be relatively resistant to borrowing. For example, while English is a Germanic language it has borrowed around 50% of its total lexicon from French and Latin. However, only about 5% of English entries in the Swadesh 200 word list are clear Romance language borrowings[1]. Where borrowings were obvious Dyen *et al.* did not score them as cognate, and thus they were excluded from our analysis. 11 of the speech varieties that were not coded by Dyen *et al.* were also excluded. To facilitate reconstruction of some of the oldest language relationships, we added three extinct Indo-European languages, thought to fit near the base of the tree (Hittite, Tocharian A and Tocharian B). Word form and cognacy judgements for all three languages were made on the basis of multiple sources to ensure reliability. Presence or absence of words from each cognate set was coded as '1' or '0' respectively to produce a binary matrix of 2449 cognates in 87 languages.

Tree Construction. Language trees were constructed using a 'restriction site' model of evolution that allows for unequal character-state frequencies and gamma distributed character specific rate heterogeneity *(MrBayes version 2.01[30])*. We used default 'flat' priors for the rate matrix, branch lengths, gamma shape parameter and site-specific rates. The results were found to be robust to changes in these priors. For example, repeating the analyses with an exponential branch-length prior produced a 95% confidence interval for the basal divergence time of between 7,100BP and 9,200BP.

The program was run ten times using four concurrent Markov chains. Each run generated 1,300,000 trees from a random starting phylogeny. On the basis of an autocorrelation analysis only every 10,000th tree was sampled to ensure that consecutive samples were independent. A 'burn-in' period of 300,000 trees for each run was used to avoid sampling trees before the run had reached convergence. Log-likelihood plots and an examination of the post burn-in tree topologies demonstrated that the runs had indeed reached convergence by this time. For each analysis a total of 1,000 trees were sampled and rooted with Hittite. The branch between Hittite and the rest of the tree was split at the root such that half its length was assigned to the Hittite branch and half to the remainder of the tree - divergence time estimates were found to be robust to threefold alterations of this allocation.

Divergence time estimates. 11 nodes corresponding to the points of initial divergence in all of the major language sub-families were given minimum and/or maximum ages based on known historical information (see supplementary material). The ages of all terminal nodes on the tree, representing languages spoken today, were set to zero by default. Hittite and the Tocharic languages were constrained in accordance with estimated ages of the source texts. Relatively broad date ranges were chosen in order to avoid making disputable, *a priori* assumptions about Indo-European history. A likelihood ratio test with the extinct languages removed revealed that rates were significantly non-clocklike (χ^2=787.3, df=82, p<.001). Divergence time estimates were thus made using the semi-parametric, *Penalized-Likelihood* model of rate variation implemented in R8s (*version 1.50*)[19]. The cross-validation procedure was applied to the majority-rule consensus tree (figure 1) to determine the optimal value of the rate smoothing parameter. Step-by-step removal of each of the 14 age constraints on the consensus tree revealed that divergence time estimates were robust to calibration errors. For 13 nodes, the reconstructed age was within 390 years of the original constraint range. Only the reconstructed age for Hittite showed an appreciable variation from the constraint range. This may be attributable to the effect of missing data associated with extinct languages. Reconstructed ages at the base of the tree ranged from 10,400BP with the removal of the Hittite age constraint, to 8,500BP with the removal of the Iranian group age constraint.

1. Pagel, M. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. & Trask, L.) 189-207 (The McDonald Institute for Archaeological Research, Cambridge, 2000).

2. Gray, R.D., & Jordan, F.M. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052-1055 (2000).

3. Diamond, J. & Bellwood, P. Farmers and Their Languages: The First Expansions. *Science* **300**, 597 (2003).

4. Richards, M. *et al.* Tracing European founder lineage in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251-1276 (2000).

5. Semoni, O. *et al*. The genetic legacy of Paleolithic Homo sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**, 1155-1159 (2000).

6. Chikhi, L. Nichols, R. A., Barbujani, G., & Beaumont, M. A. Y genetic data support the Neolithic Demic Diffusion Model. *Proc. Natl. Acad. Sci. USA* **99**, 11008-11013 (2002).

7. Gimbutas, M. The beginning of the Bronze Age in Europe and the Indo-Europeans 3500-2500 B.C. *Journal of Indo-European Studies* **1**, 163-214 (1973).

8. Mallory, J. P. *In search of the Indo-Europeans: Languages, Archaeology and Myth.* (Thames & Hudson, London, 1989).

9. Renfrew, C. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. & Trask, L.) 413-439 (The McDonald Institute for Archaeological Research, Cambridge, 2000).

10. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 453-463 (1952).

11. Bergsland, K. & Vogt, H. On the validity of glottochronology. *Current Anthropology* **3**, 115-153 (1962).

12. Blust, R. in *Time Depth in Historical Linguistics* (eds Renfrew, C., McMahon, A. & Trask, L.) 311-332 (The McDonald Institute for Archaeological Research, Cambridge, 2000).

13. Steel, M.A., Hendy, M.D. & Penny, D. Loss of information in genetic distances. *Nature* **333**, 494-495 (1988).

14. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. in *Molecular Systematics* (eds Hillis, D., Moritz, C. & Mable, B. K.) 407-514 (Sinauer Associates, Inc. Sunderland, Massachusetts, 1996).

15. Dixon, R. M. W. *The Rise and Fall of Language.* (Cambridge University Press, Cambridge, 1997).

16. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M.N., Teller, A.H., & Teller. E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087-1091 (1953).

17. Huelsenbeck, J.P., Ronquist, F., Nielsen, R., & Bollback, J.P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314 (2001).

18. Huson, D.H. SplitsTree: Analyzing and visualizing evolutionary data, *Bioinformatics* **14**, 68-73 (1998).

19. Sanderson, M. *R8s, Analysis of rates of evolution, version 1.50.* (University of California, Davis, 2002).

20. Dyen, I., Kruskal, J. B. & Black, P. FILE IE-DATA1. World Wide Web. Available online: http://www.ntu.edu.au/education/langs/ielex/IE-DATA1 (1997).

21. Sanderson, M. J. & Donoghue, M. J. Patterns of variation in levels of homoplasy. *Evolution*, **43**, 1781-1795 (1989).

22. Gamkrelidze, T. V. & Ivanov, V. V. *Indo-European and the Indo-Europeans. (Trends in Linguistics 80)*. (Mouton de Gruyter, Berlin, 1995).

23. Rexova, K., Frynta, D. & Zrzavy, J. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* **19**, 120-127 (2003).

24. Ringe, D., Warnow, T. & Taylor, A. IndoEuropean and computational cladistics. *Trans. Philol. Soc.* **100**, 59-129 (2002).

25. Gkiasta, M. Russell, T. Shennan, S. & Steele, J. Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* **77**, 45-62 (2003).

26. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The history and geography of human genes.* (Princeton University Press, Princeton, 1994).

27. Holden, C.J. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. Roy. Soc. London* **269**, 793-799 (2002).

30. Barbrook, A. C., Howe, C. J., Blake, N. & Robinson, P. The phylogeny of *The Canterbury Tales*. *Nature* **394**, 839 (1998).

29. McMahon, A. & McMahon, R. Finding families: Quantitative methods in language classification. *Transactions of the Philological Society* **101**, 7-55 (2003).

30. Huelsenbeck, J. P & Ronquist. F. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754-755 (2001).

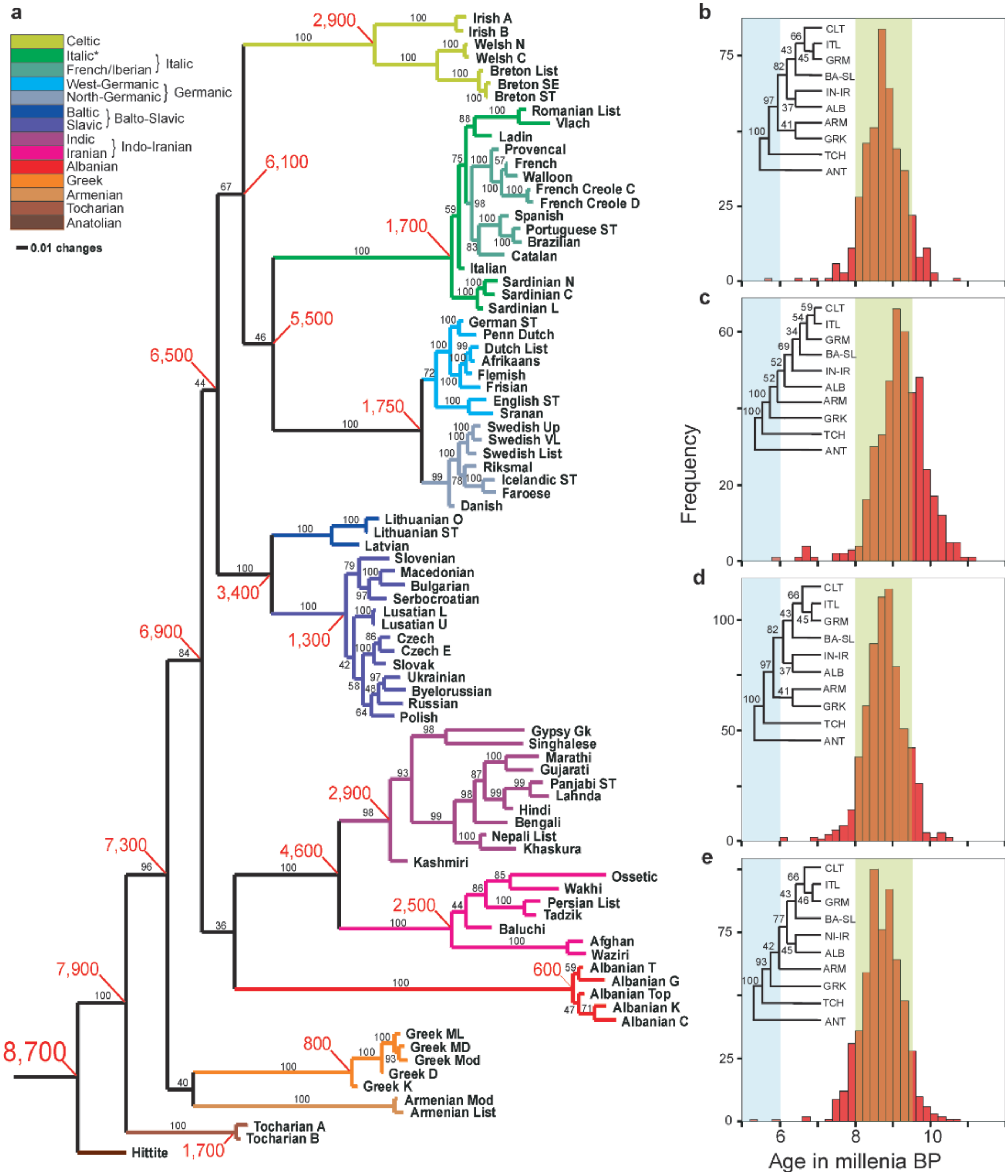
Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank S. Allan, L. Chikhi, M. Corballis, S. Greenhill, J. Hamm, J. Huelsenbeck, G. Nichols, A. Rodrigo, F. Ronquist, M. Sanderson and S. Shennan for useful advice and/or comments on the manuscript.

Correspondence and requests for materials should be addressed to R.G. (e-mail: rd.gray@auckland.ac.nz).

Figure 1 **a** Majority-rule consensus tree based on the MCMC sample of 1,000 trees. The major language groupings are colour coded. Branch-lengths are proportional to the inferred maximum-likelihood estimates of evolutionary change per cognate. Values above each branch (in black) express the Bayesian posterior probabilities as a percentage. Values in red show the inferred ages of nodes in years BP. *Italic also includes the French/Iberian sub-group. Panels **b-e** show the distribution of divergence time estimates at the root of the Indo-European phylogeny for: **b**, initial assumption set using all cognate information and most stringent constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), ((NorthGermanic, WestGermanic), Italic, Celtic)))]; **c**, conservative cognate coding with doubtful cognates excluded; **d**, all cognate sets with minimum topological constraints [(Anatolian, Tocharian, (Greek, Armenian, Albanian, (Iranian, Indic), (Slavic, Baltic), (NorthGermanic, WestGermanic), Italic, Celtic))]; **e**, missing data coding with minimum topological constraints and all cognate sets. Shaded bars represent the implied age ranges under the two competing theories of Indo-European origin – blue for the Kurgan hypothesis and green for the Anatolian farming hypothesis. The relationship between the major language groups in the consensus tree for each analysis is also shown, along with posterior probability values.


a
Celtic
Italic*
French/Iberian
Italic
West-Germanic
North-Germanic
Germanic
Baltic
Slavic
Balto-Slavic
Indic
Iranian
Indo-Iranian
Albanian
Greek
Armenian
Tocharian
Anatolian
0.01 changes
2,900
Irish A
Irish B
Welsh N
Welsh C
Breton List
Breton SE
Breton ST
Romanian List
Vlach
Ladin
Provencal
French
Walloon
French Creole C
French Creole D
Spanish
Portuguese ST
Brazilian
Catalan
Italian
Sardinian N
Sardinian C
Sardinian L
6,100
1,700
5,500
German ST
Penn Dutch
Dutch List
Afrikaans
Flemish
Frisian
English ST
Sranan
1,750
Swedish Up
Swedish VL
Swedish List
Riksmal
Icelandic ST
Faroese
Danish
6,500
Lithuanian O
Lithuanian ST
Latvian
Slovenian
Macedonian
Bulgarian
Serbocroatian
Lusatian L
Lusatian U
Czech
Czech E
Slovak
Ukrainian
Byelorussian
Russian
Polish
3,400
1,300
6,900
Gypsy Gk
Singhalese
Marathi
Gujarati
Panjabi ST
Lahnda
Hindi
Bengali
Nepali List
Khaskura
Kashmiri
2,900
4,600
7,300
Ossetic
Wakhi
Persian List
Tadzik
Baluchi
Afghan
Waziri
2,500
Albanian T
Albanian G
Albanian Top
Albanian K
Albanian C
600
7,900
8,700
Greek ML
Greek MD
Greek Mod
Greek D
Greek K
800
Armenian Mod
Armenian List
Tocharian A
Tocharian B
Hittite
1,700
b
c
d
e
CLT
ITL
GRM
BA-SL
IN-IR
NI-IR
ALB
ARM
GRK
TCH
ANT
Frequency
0
25
50
75
100
20
60
6
8
10
Age in millenia BP