RESEARCH ARTICLE SUMMARY

LANGUAGE EVOLUTION

Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages

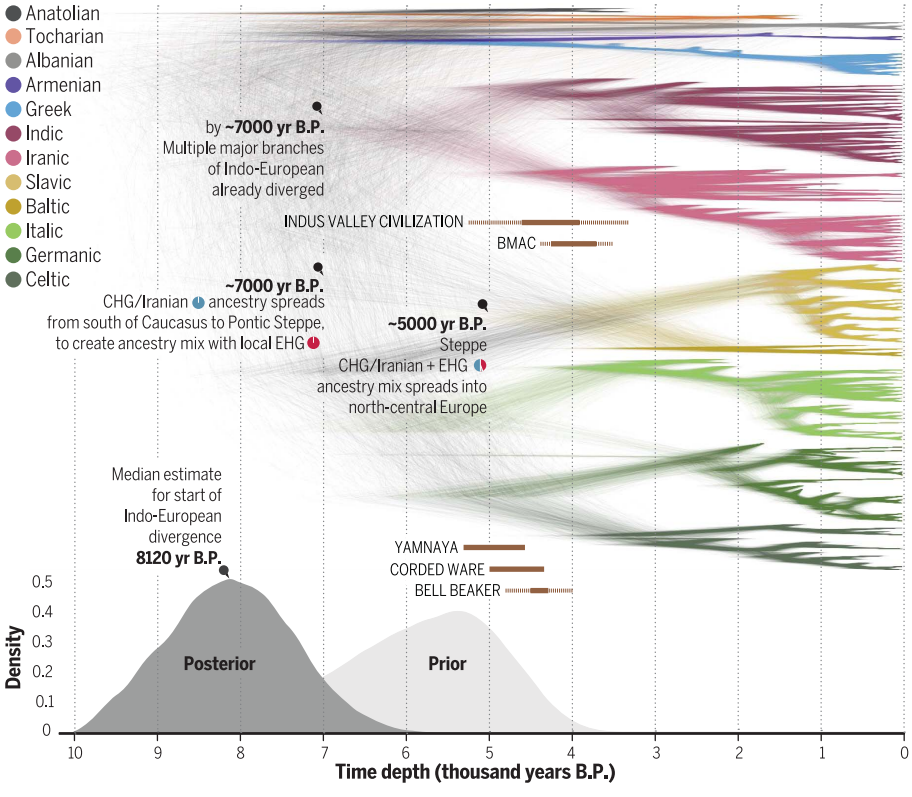
Paul Heggarty *et al*.

INTRODUCTION: Almost half the world's population speaks a language of the Indo-European language family. It remains unclear, however, where this family's common ancestral language (Proto-Indo-European) was initially spoken and when and why it spread through Eurasia. The "Steppe" hypothesis posits an expansion out of the Pontic-Caspian Steppe, no earlier than 6500 years before present (yr B.P.), and mostly with horse-based pastoralism from ~5000 yr B.P. An alternative "Anatolian" or "farming" hypothesis posits that Indo-European dispersed with agriculture out of parts of the Fertile Crescent, beginning as early as ~9500 to 8500 yr B.P. Ancient DNA (aDNA) is now bringing valuable new perspectives, but these remain only indirect interpretations of language prehistory. In this study, we tested between the time-depth predictions of the Anatolian and Steppe hypotheses, directly from language data. We report a new framework for the chronology and divergence sequence of Indo-European, using Bayesian phylogenetic methods applied to an extensive new dataset of core vocabulary across 161 Indo-European languages.

RATIONALE: Previous phylolinguistic analyses have produced conflicting results. We diagnosed and resolved the causes of this discrepancy, two in particular. First, the datasets used had limited language sampling and widesp… coding inconsistency. Second, some anal… enforced the assumption that modern spoken languages derive directly from ancient written languages rather than from parallel spoken varieties. Together, these methodological problems distorted branch-length estimates and date inferences. We present a new dataset of cognacy (shared word origins) across Indo-European. This dataset eliminates past inconsistencies and provides a fuller and more balanced language sample, including 52 nonmodern languages for a denser set of time-calibration points. We applied ancestry-enabled Bayesian phylogenetic analysis to test rather than enforce direct ancestry assumptions.

RESULTS: Few ancient written languages are returned as direct ancestors of modern clades. We find a median root age for Indo-European of ~8120 yr B.P. (95% highest posterior density: 6740 to 9610 yr B.P.). Our chronology is robust across a range of alternative phylogenetic models and sensitivity analyses that vary data subsets and other parameters. Indo-European had already diverged rapidly into multiple major branches by ~7000 yr B.P., without a coherent non-Anatolian core. Indo-Iranic has no close relationship with Balto-Slavic, weakening the case for it having spread via the steppe.

CONCLUSION: Our results are not entirely consistent with either the Steppe hypothesis or the farming hypothesis. Recent aDNA evidence suggests that the Anatolian branch cannot be sourced to the steppe but rather to south of the Caucasus. For other branches, potential candidate expansion(s) out of the Yamnaya culture are detectable in aDNA, but some had only limited genetic impact. Our results reveal that these expansions from ~5000 yr B.P. onward also came too late for the language chronology of Indo-European divergence. They are consistent, however, with an ultimate homeland south of the Caucasus and a subsequent branch northward onto the steppe, as a secondary homeland for some branches of Indo-European entering Europe with the later Corded Ware–associated expansions. Language phylogenetics and aDNA thus combine to suggest that the resolution to the 200-year-old Indo-European enigma lies in a hybrid of the farming and Steppe hypotheses. ■



A DensiTree showing the probability distribution of tree topologies for the Indo-European language family. The time axis shows the estimated chronology of the family's geographical expansion and divergence, calibrated on 52 nonmodern written languages. Annotations add chronological context relative to selected archaeological cultures and expansions of significant ancestry components in the aDNA record. CHG, Caucasus hunter-gatherers; EHG, Eastern (European) hunter-gatherers; BMAC, Bactria-Margiana Archaeological Complex.

All authors and affiliations appear in the full article online.
*Corresponding authors: Paul Heggarty (paul.heggarty@gmail.com); Cormac Anderson (cormacanderson@gmail.com); Denise Kühnert (kuehnert@shh.mpg.de); Russell D. Gray (russell_gray@eva.mpg.de)
Cite this article as P. Heggarty *et al*., *Science* **381**, eabg0818 (2023). DOI: 10.1126/science.abg0818

RESEARCH ARTICLE

LANGUAGE EVOLUTION

Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages

Paul Heggarty[1,2,3]*, Cormac Anderson[3]*, Matthew Scarborough[3,4], Benedict King[3], Remco Bouckaert[5], Lechosław Jocz[6], Martin Joachim Kümmel[7], Thomas Jügel[8], Britta Irslinger[9], Roland Pooth[10], Henrik Liljegren[11], Richard F. Strand[12], Geoffrey Haig[13], Martin Macák[14], Ronald I. Kim[15], Erik Anonby[16,17], Tijmen Pronk[17], Oleg Belyaev[18,19], Tonya Kim Dewey-Findell[20], Matthew Boutilier[21], Cassandra Freiberg[22], Robert Tegethoff[3,7], Matilde Serangeli[7], Nikos Liosis[23], Krzysztof Stroński[24], Kim Schulte[25], Ganesh Kumar Gupta[24], Wolfgang Haak[26], Johannes Krause[26], Quentin D. Atkinson[27,28], Simon J. Greenhill[3,29], Denise Kühnert[30]*, Russell D. Gray[3,27]*

The origins of the Indo-European language family are hotly disputed. Bayesian phylogenetic analyses of core vocabulary have produced conflicting results, with some supporting a farming expansion out of Anatolia ~9000 years before present (yr B.P.), while others support a spread with horse-based pastoralism out of the Pontic-Caspian Steppe ~6000 yr B.P. Here we present an extensive database of Indo-European core vocabulary that eliminates past inconsistencies in cognate coding. Ancestry-enabled phylogenetic analysis of this dataset indicates that few ancient languages are direct ancestors of modern clades and produces a root age of ~8120 yr B.P. for the family. Although this date is not consistent with the Steppe hypothesis, it does not rule out an initial homeland south of the Caucasus, with a subsequent branch northward onto the steppe and then across Europe. We reconcile this hybrid hypothesis with recently published ancient DNA evidence from the steppe and the northern Fertile Crescent.

The Indo-European language family encompasses more than 400 languages (*1*, *2*). These languages are spoken by almost half of the world's population (*2*), and all derive from the same source language: Proto-Indo-European (PIE). For more than 200 years, the origins of Indo-European have been disputed (*3*). The deep link between the widely dispersed Indo-European languages was discovered more than two centuries ago (*4*), but where their common ancestral language was initially spoken, and when and why it spread so far through Eurasia, have remained enigmas ever since. Recent debate has focused on two leading hypotheses. The Steppe hypothesis posits that Indo-European spread out of the Pontic-Caspian Steppe, no earlier than 6500 years before present (yr B.P.), and mostly with horse-based pastoralism from ~5000 yr B.P. (*5*) (Fig. 1B). The farming hypothesis claims that Indo-European dispersed with agriculture out of parts of the Fertile Crescent, beginning as early as ~9500 to 8500 yr B.P. (*6*) (Fig. 1C). Linguistic reconstructions of some PIE lexicon, and ancient contacts with early stages of the Uralic language family, have been widely interpreted as supporting the Steppe hypothesis (*5*, *7*), but the interpretation of these data is controversial (*8*, *9*) (Box 1). In contrast, analyses of Indo-European basic vocabulary using Bayesian phylogenetic methods initially supported the time depth and geographical origin posited by the farming hypothesis (*10*, *11*). Recent papers (*12*–*14*) have challenged those early time-depth estimates, in part because the model used did not allow ancient languages to be directly ancestral to any modern languages. When eight ancient languages were constrained to be directly ancestral, the date estimation for the Indo-European root moved into the time frame of the Steppe hypothesis (*12*). However, a considerable problem with this analysis is that forcing direct ancestry produces date inferences toward the tips of the tree that conflict with the known histories of several branches of Indo-European. The diversification of Romance languages, for example, is inferred to have started only 1000 years ago (*12*), when, in fact, regional differences had begun to arise a millennium earlier, as Roman expansion itself had already led to "great diversity in the Latin that was spoken around the Empire" (*15*). In this study, we investigated, diagnosed, and resolved the problems in data quality that led to these artifacts in dating inferences.

Human ancient DNA (aDNA) is now also reshaping the debate. Results support a substantial influx of genetic ancestry from the Eurasian Steppe ~5000 yr B.P., which could have carried several of the main branches of Indo-European into Europe (*16*–*18*). However, this ancestry signal is less evident in aDNA from Mycenaean Greece (*19*), the Balkans (*20*), and Anatolia (*21*–*23*), casting doubt on whether the Steppe hypothesis can explain the spread of all branches of the family, especially in the eastern Mediterranean and Asia. This fuller aDNA picture "does not support a classical way of looking at the steppe hypothesis" (*24*).

We overcame the limitations of previous linguistic analyses by combining recent advances in Bayesian phylogenetic inference with a far more extensive Indo-European dataset. First, we deployed a sampled ancestor phylogenetic analysis (*25*) that permits but does not force ancient languages to be directly ancestral to modern languages (fig. S5.4). This is achieved by using a birth-death-sampling tree prior (fig. S5.4) in which a branching event in the tree is a "birth" or diversification event, and lineage extinction ("death") events may also occur. Each ancient language covered in our dataset represents an occurrence of "sampling" from the entire diversity of Indo-European languages through time. Rather than assuming that ancient languages were the direct ancestors of their modern relatives, this approach

[1]Departamento de Humanidades, Pontificia Universidad Católica del Perú, 15088 Lima, Peru. [2]Waves Group, Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. [3]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. [4]Department of Nordic Studies and Linguistics, University of Copenhagen, S 2300 København, Denmark. [5]Centre for Computational Evolution, University of Auckland, Auckland 1010, New Zealand. [6]Faculty of Humanities, Jacob of Paradies University, 66-400 Gorzów Wielkopolski, Poland. [7]Seminar for Indo-European Studies, Institut für Orientalistik, Indogermanistik, Ur- und Frühgeschichtliche Archäologie, Friedrich-Schiller-Universität Jena, 07743 Jena, Germany. [8]Center for Religious Studies (CERES), Ruhr University Bochum, 44789 Bochum, Germany. [9]Saxon Academy of Sciences and Humanities, 04107 Leipzig, Germany. [10]Department of Linguistics, Ghent University, 9000 Ghent, Belgium. [11]Department of Linguistics, Stockholm University, 10691 Stockholm, Sweden. [12]Independent scholar, Cottonwood, AZ 86326, USA. [13]Department of General Linguistics, University of Bamberg, 96047 Bamberg, Germany. [14]Independent scholar, 949 74 Nitra, Slovakia. [15]Department of Older Germanic Languages, Faculty of English, Adam Mickiewicz University in Poznań, 60-780 Poznań, Poland. [16]School of Linguistics and Language Studies, Carleton University, Ottawa, ON K1S 5B6, Canada. [17]Leiden University Centre for Linguistics, 2300 RA Leiden, Netherlands. [18]Department of Theoretical and Applied Linguistics, Lomonosov Moscow State University, 119991 GSP-1 Moscow, Russia. [19]Department of Iranian Languages, Institute of Linguistics RAS, Moscow 125009, Russia. [20]Centre for the Study of the Viking Age, School of English, University of Nottingham NG7 2RD, UK. [21]Department of German, Nordic, and Slavic, University of Wisconsin–Madison, Madison, WI 53706, USA. [22]Institut für deutsche Sprache und Linguistik, Sprach- und literaturwissenschaftliche Fakultät, Humboldt-Universität zu Berlin, 10099 Berlin, Germany. [23]Institute of Modern Greek Studies, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece. [24]Faculty of Modern Languages, Adam Mickiewicz University in Poznań, 61-874 Poznań, Poland. [25]Department of Translation and Communication, Jaume I University, 12006 Castelló de la Plana, Spain. [26]Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany. [27]School of Psychology, University of Auckland, Auckland 1010, New Zealand. [28]Centre for the Study of Social Cohesion, University of Oxford, Oxford OX2 6PN, UK. [29]ARC Center of Excellence for the Dynamics of Language, ANU College of Asia and the Pacific, The Australian National University, Canberra, ACT 2600, Australia. [30]Transmission, Infection, Diversification and Evolution Group, Max Planck Institute of Geoanthropology, 07745 Jena, Germany.
*Corresponding author. Email: paul.heggarty@gmail.com (P.H.); cormacanderson@gmail.com (C.A.); kuehnert@shh.mpg.de (D.K.); russell_gray@eva.mpg.de (R.D.G.)

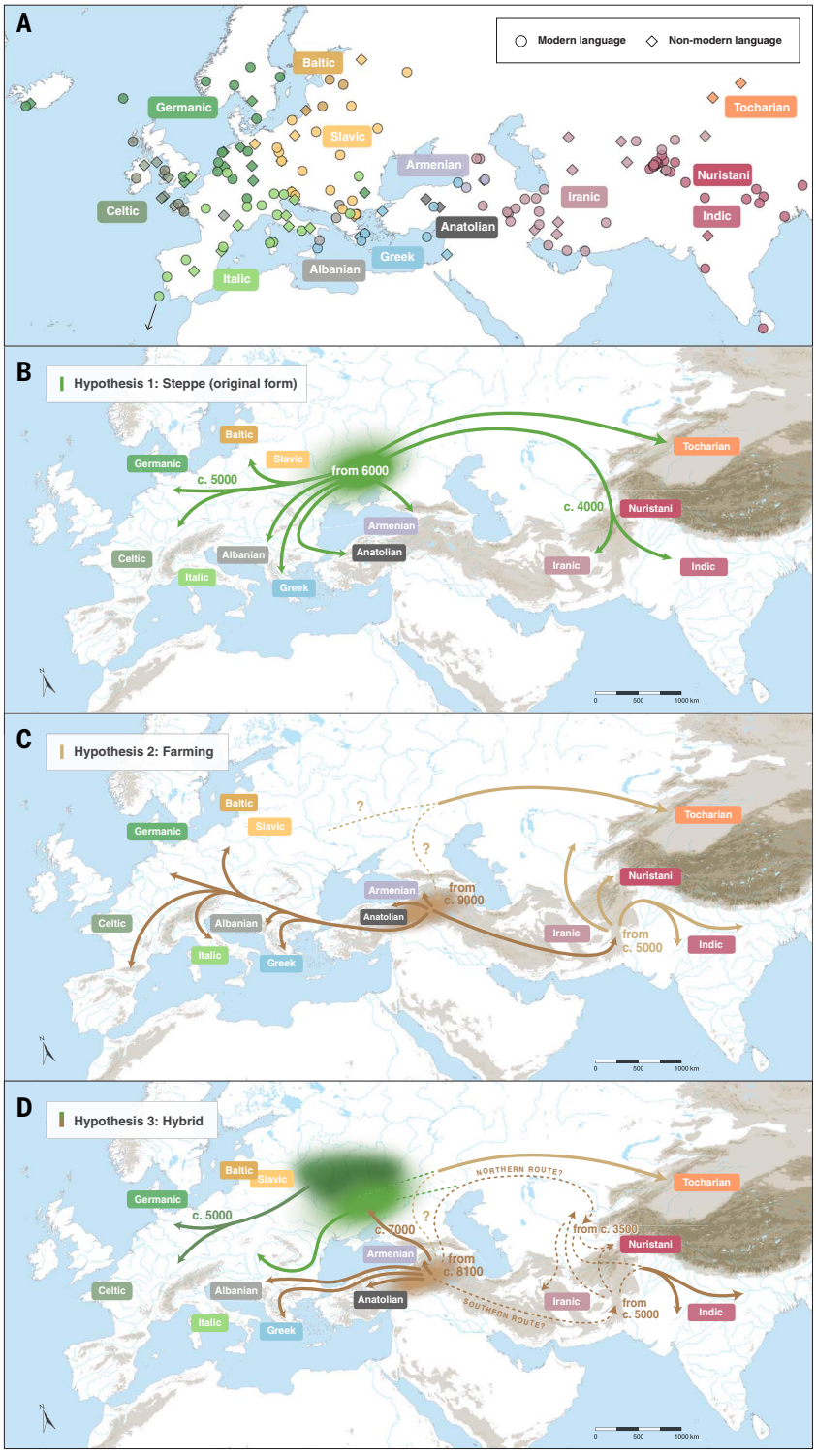


Fig. 1. Indo-European languages through space and time. (**A**) Indo-European languages covered in the IE-CoR database: 109 modern languages (round dots) and 52 nonmodern languages (diamonds). An interactive version is available at https://iecor.clld.org/languages. Colors distinguish the 12 main clades of Indo-European (other potential clades went extinct without sufficient written record). (**B** to **D**) Maps showing alternative hypotheses for the first stages of Indo-European expansion. The hypothesis of an origin in the western steppe (B) contrasts with the hypothesis of an earlier spread with farming (C). The map in (D) shows a hybrid of parts of both hypotheses. Date estimates for the start of divergence within each main clade are given in years before present. Language labels on the hypothesis maps reflect recent end points, not necessarily earlier movements.

estimates from the linguistic dataset itself the relative probability that any ancient language is either a direct ancestor or a sister taxon to its closest modern relatives. The model thus determines from the data whether, for example, the Proto-Romance source of all modern Romance languages goes back directly to the lexicon of written Classical Latin, as constrained by one recent analysis (*12*), or to some slightly different, spoken form of "Vulgar" Latin. To estimate chronology, we used an uncorrel ated relaxed clock to allow different language lineages in the phylogeny to vary in rates of change over time (*26*). Cognacy status also changes much faster in some types of meaning than in others, so we tested different approaches to this, using models of cognate evolution that allow different rates of change for every individual meaning, or for sets of meanings that show similar degrees of divergence in cognacy.

Second, we identified artifacts in previous phylogenetic analyses that result from flaws and inconsistencies in the language datasets used (*27*). To resolve these, we implemented a methodology for encoding cognate data [see supplementary materials (SM) section 2] to maximize consistency across the language dataset and optimize it as input to phylogenetic analysis, creating an entirely new database of Indo-European cognate relationships, named IE-CoR. IE-CoR covers 161 languages, coded by more than 80 specialists on languages of the Indo-European family, to provide much denser and more-balanced sampling both within and between the main subclades of Indo-European. The 52 nonmodern languages in IE-CoR (Fig. 1A) provide a much denser set of date calibrations than earlier databases.

Results

Our main analysis (Fig. 2) produced an estimated date for the root of the Indo-European language family that is too early to be compatible with the Steppe hypothesis: ~8120 yr B.P., with a 95% credible region of 6740 to 9610 yr B.P. [Date estimates are reported here as a median date before present, followed by the 95% credible region (highest posterior density, or HPD), all rounded to the nearest decade, and taking the "present" for modern languages as 2000 CE.] The posterior tree distribution also contained relatively few cases of direct ancestry between language taxa. Of the 52 nonmodern written languages in the IE-CoR database, 27 might theoretically be considered potential candidates to be directly ancestral to more recent languages in their clades. Old English, for example, is potentially ancestral to modern English, and Ancient (Attic) Greek to modern forms of Greek. Figure 3 shows the prior and posterior probabilities for each of these nonmodern languages being a direct ancestor to any later language(s) in its clade

Box 1. Recovering prehistory from languages.

Languages that derive from the same former ancestor language retain signals of that past origin and of their divergence since then. By meticulously comparing the languages within a family, it is possible to reconstruct aspects of their common ancestor language. Much of the PIE sound system (phonology) and word structure (morphology) has been reconstructed, along with hundreds of individual word forms.

Linguistics has other methods to then make inferences about prehistory from such language data. These qualitative methods are often claimed to support the Steppe hypothesis, but each major inference remains disputed.

- Cladistic analysis of selected characters in phonology, morphology, and cognacy yielded no single "perfect phylogeny" (*50*) but was taken to support a node uniting the Indo-Iranic and Balto-Slavic branches (*5*), with putative parallels in aDNA (*49*). This node rested on only three data characters, however. All three are contentious, notably the centum/satem distinction and the "ruki" rule (SM section 7.6.2.1). There is no consensus support for this node in Indo-European linguistics, and our analysis finds little support for it (a posterior probability of just 0.11). We also tested the effect of enforcing this node and found little impact on the root date (Fig. 4, SA6b).
- Apparent ancient loanwords into early stages of the Uralic family (in northern Eurasia) have been argued to originate in the Indo-Iranic branch of Indo-European and thus to point to the steppe as the likely location of such contacts (*5*). However, other and even earlier claimed loanwords, with Caucasian and Semitic languages, are more compatible with an ultimate homeland farther south (*54*).
- Linguistic paleontology assumes that certain word forms reconstructed to PIE denoted particular artifacts, species, and concepts already known to its speakers—most notably the wheel. Reconstruction operates through laws of sound change and can thus be precise and reliable on this level. There are no comparably strict and predictable meaning laws, however, so it is often much more challenging to pinpoint what exact meanings were at specific deep points in time. The same reconstructed word forms have thus been inferred as evidence that PIE speakers either already did know of the wheel (*5*, *65*), or that they did not yet know of it, and that the invention postdated the common ancestor language (*8*, *66*–*68*).

Indo-European origins have remained unresolved because all methods have left scope for interpretation and dispute and have failed to bring consensus on the tree topology, chronology, or homeland. For details, see SM section 2.2.

(see also table S5.2). Our ancestry-enabled analysis finds posterior probabilities >0.01 for only four languages: Classical Armenian (0.50) and three ancient forms of Greek (0.72, 0.39, and 0.31). Only in two of these cases is the posterior probability greater than 50%. We found no support for the higher number of eight direct ancestors enforced in previous analyses (*12*). These results are driven by the cognate data, not our tree prior. In the prior, direct ancestry probabilities ranged from ~42% to 78% for all 27 potential ancestor languages, and the median root date estimate was 5815 yr B.P. (4149 to 8123 yr B.P.). Including the cognate data shifted the root date 2305 years earlier, to our result of a median age of 8120 yr B.P. in the posterior.

This lack of direct ancestry may, at first sight, seem unexpected. Old English is not inferred to be the direct ancestor to modern English, nor is Old Icelandic directly ancestral to modern Icelandic. However, it is important to clarify what a split between lineages represents in phylogenetic analyses of cognate datasets. A split does not just correspond to the major difference between discrete, mutually unintelligible "languages." Rather, lineages must in principle already be split from each other for them to be free to start developing differently. Only once lineages are split can the first difference(s) emerge between them in the predominant lexeme they use, even for just a single meaning in the dataset. So even dialects or registers (written versus spoken) of the "same" language can represent different, parallel sublineages. Thus, ancestry between past written languages and contemporary spoken ones may not be fully direct (SM section 7). A whole language, taken in the broad sense as spanning multiple registers and regional variants, therefore need not correspond just to a single lineage, but may span separate sublineages still very close to each other in the phylogeny. "Latin" as a whole encompassed both written Classical Latin and the spoken ancestor of Romance languages.

In the history of English, the term "Old English" actually refers to a set of various dialects. The IE-CoR Old English data are based on West Saxon, as the best documented of those dialects. As our results correctly reflect, this was not the dialect most directly ancestral to modern English (*28*). Likewise, the Sanskrit of the sacred Vedic texts is not the direct ancestor of modern Indic languages but was a distinct sister dialect. Even the intervening Prākrits of Medieval India "do not derive from Sanskrit" (*29*) and, specifically, "do not go back directly to the dialect which formed the basis of Vedic" (*29*), which stood apart as a "far-western dialect" (*30*). The formal register of a written language typically differs from the contemporaneous spoken language in the predominant usage of different words in a small proportion of the vocabulary, and this specifically includes meanings within the IE-CoR reference set of core lexicon. Even a near-direct ancestor may thus be expected to show some lexical differences with the lineage ancestral to modern spoken languages. For example, modern Romance languages do not derive directly from written Classical Latin (*31*). Instead, "the origins of the Romance languages lie in the (irrecoverable) spoken language … [and] there will always be a mismatch between the Latin sources and the parent of the Romance languages" (*32*). Even one difference, in a single meaning of the 170 in the IE-CoR reference set, logically entails separate sublineages, and that ancestry is not fully direct. In the IE-CoR meaning MOUTH, for example, the Classical Latin *os* was not inherited into any modern Romance languages, and so is not considered the primary term in Proto-Romance. Most Romance languages use cognates derived instead from *bucca* (hence, Italian *bocca*, Spanish *boca*, and French *bouche*, for example), which in colloquial Latin was already used specifically in the meaning MOUTH as early as Cato the Elder (234–149 BCE) (*33*). This one difference is already enough to entail that a phylogenetic analysis of primary lexemes (and thus cognacy states) between Classical Latin and Proto-Romance would correctly return these as separate sublineages, and it is not an isolated example. In practice, "many Classical Latin words do not survive into Romance" (*15*), or survive only sporadically, also in IE-CoR core vocabulary, such as EAT and GO (*15*). Our ancestry-enabled model returns the standard linguistic analysis in this case: that written Classical Latin is not in fact directly ancestral to modern spoken Romance languages. Specifically, in meanings where Classical Latin has a cognate set different to that in all Romance languages, the model correctly identifies which branch is innovating in each case. Even Classical Latin singleton forms are correctly identified as retentions, and the Romance forms as innovations on the ("spoken") branch to Romance (see SM section 6.3). Likewise, written Old Icelandic is not quite directly ancestral to modern spoken Icelandic. This contradicts the assumptions enforced in earlier ancestry-constrained analyses (*12*). Only in four cases were specific historical written languages [Classical Armenian and some forms of Ancient Greek (*34*, *35*)] so close to the ancestor of later languages in their clades as to be nearly indistinguishable in the IE-CoR sample of core vocabulary.

Validation, and robustness analyses

The validity of our results can be evaluated in three ways. First, estimates of lineage split dates can be validated against known historical

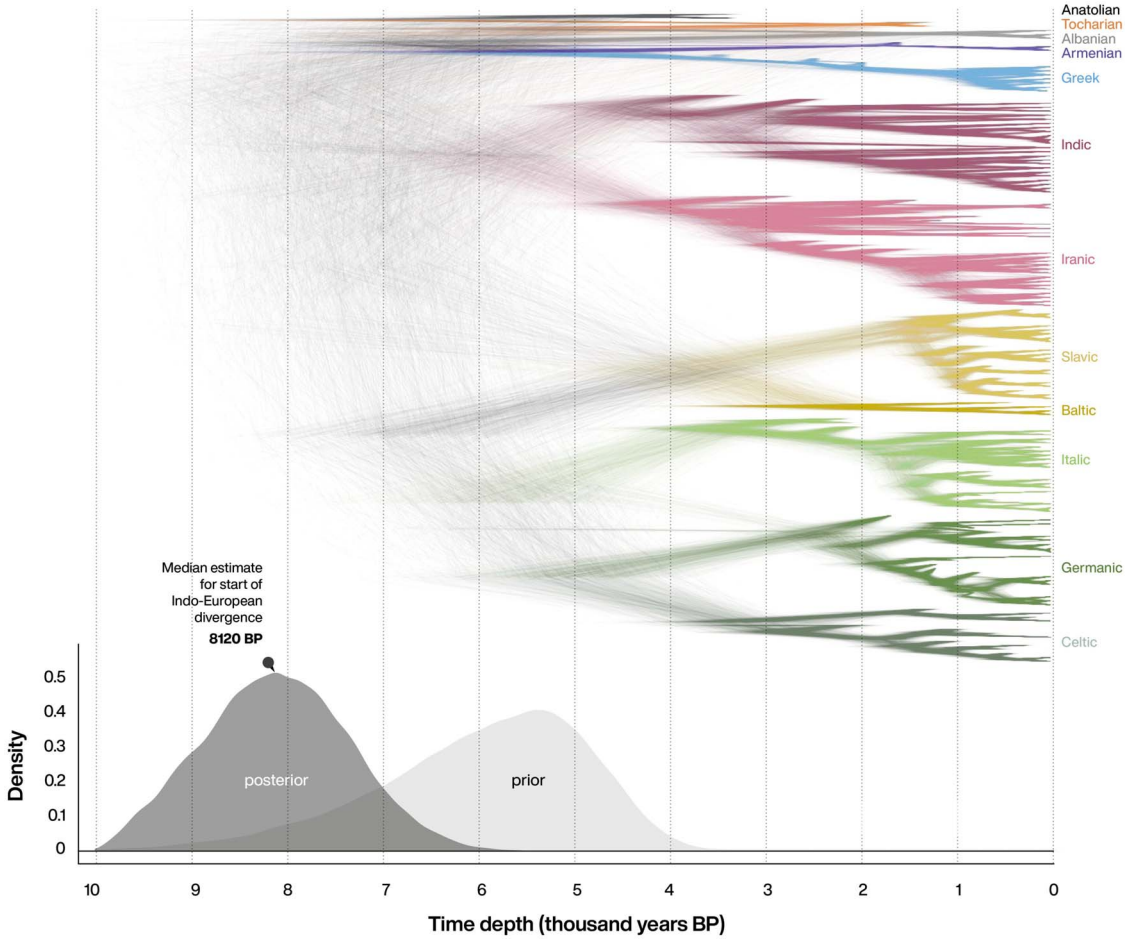


Fig. 2. The posterior probability distribution of trees for the Indo-European family. Distribution visualized using DensiTree (*71*). The time axis shows the estimated chronology of Indo-European expansion. Languages whose tips do not reach the right edge are the 52 nonmodern written languages such as Hittite, Tocharian, Mycenaean Greek, and Old English. These languages were used in the analysis as time calibrations. The two gray curves show the distribution of root date estimates for the tree. The prior is light gray, and the posterior estimate is dark gray.

data. Ancestry constraints used in previous analyses produced lineage split dates far too recent to be compatible with known histories: no divergence among West Norse languages until 1650 CE, none in Romance until 1000 CE, and none in Indic until 100 CE (*12*). These artifacts disappear from the ancestry-enabled analysis in Fig. 2. Icelandic and Faroese, for example, are now dated as splitting from the mainland Scandinavian lineages ~830 CE (470 to 950 CE), closely in line with the first Norse settlement of the Faroes and Iceland in the ninth century. Initial divergence within Romance is accurately dated to the Roman Empire in the first centuries CE. Divergence within Indic is dated to ~4370 yr B.P. (3640 to 5250 yr B.P.), in line with Vedic Sanskrit already being slightly divergent from the lineage(s) ancestral to modern spoken Indic languages (*30*). The inference of an Indo-Iranic split at ~5520 yr B.P. (4540 to 6800 yr B.P.) may, at first glance, seem surprising. Established expectations are for a more recent date, based on the perceived level of similarity between Vedic Sanskrit and Avestan—the earliest known ancient languages in the Indic and Iranic branches, respectively. However, these judgments of linguistic similarity have been largely impressionistic (*36*) rather than quantified. In the precisely defined IE-CoR meanings, Early Vedic and Younger Avestan share only 58.7% cognacy (*37*). This matches the level of cognacy that survives between the most divergent sublineages within the Romance clade, for instance, after roughly two millennia since the spread of the Roman Empire. Early Vedic and Younger Avestan themselves date back to at least the mid-fourth and mid-third millennia before present, respectively. A time depth two millennia earlier (~5520 yr B.P.) for the split between their lineages (Indic versus Iranic) is thus consistent with the 58.7% cognacy overlap between them. More widely, ancient Indo-European languages show close similarities in some aspects of their inflectional morphology (noun declension and verb conjugation) and phonology. These similarities have often been assumed to imply a relatively short time span of divergence since their common ancestor language, but these impressions are also unquantified. Our time-depth estimate implies a long period of relative stability in these aspects, while early Indo-European diverged faster in other respects. Resolving these apparent contrasts in rates of change in different aspects of language (*38*) is a target for future research (see SM section 2.2.3).

Second, our language tree topology can be evaluated against established classifications of Indo-European languages. These classifications identify 10 to 12 main attested subgroups: Anatolian, Tocharian, Albanian, Armenian, Greek, Indic+Iranic, Baltic+Slavic, Germanic, Italic, and Celtic. Our analyses (Fig. 2 and fig. S6.1) returned all of these with 100% posterior probability, including the two widely recognized deeper clades, Indo-Iranic and Balto-Slavic. Beyond this, qualitative methodology in historical linguistics has failed to reach a consensus on how these main branches relate to each other in a higher-order branching, at the earliest stages of Indo-European expansion. Different language data support conflicting tree structures. Classifications are either disputed or fall back on an unstructured rake (*2*). Our analysis, however, does find strong support for specific deep clades—findings that bear directly on interpreting the latest aDNA results across Europe (*16*–*19*, *23*, *39*). Notably, Greek goes with Armenian, while a separate main European clade brings together Germanic, Celtic, and Italic (with Balto-Slavic as next closest). At the root of Indo-European, our results return Anatolian and Tocharian as deeply divergent clades. Support for them

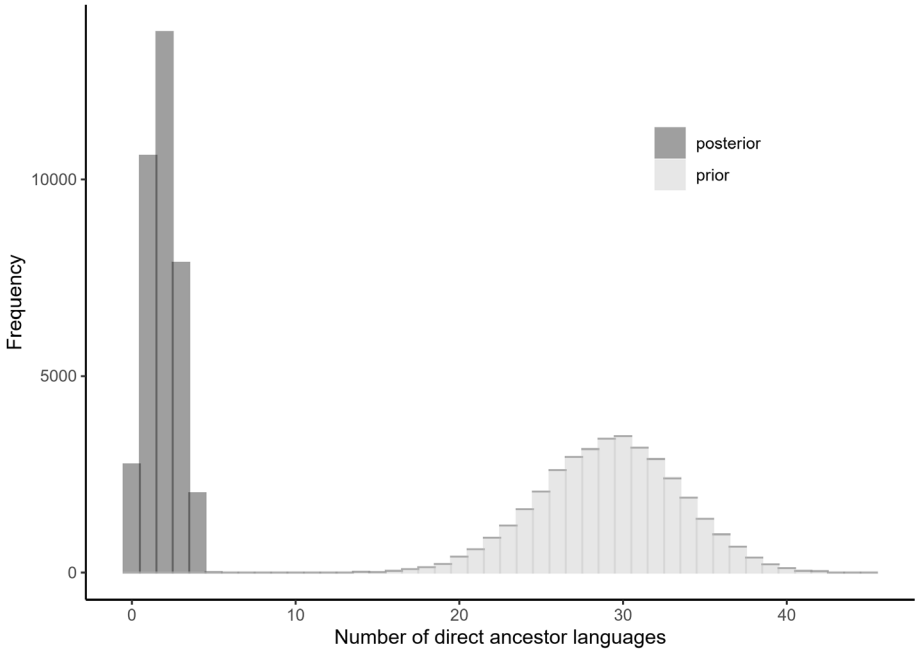


Fig. 3. Histogram of direct ancestry relationships between languages. The IE-CoR database includes 52 nonmodern languages (e.g., Ancient Greek, Classical Latin, and Early Vedic Sanskrit). This histogram shows how many of these 52 languages are returned as directly ancestral to any other language(s) in the dataset. The light-gray distribution shows the prior probability of the number of direct ancestor languages, distributed around a modal value of 28. The dark-gray distribution shows the posterior probability distribution. Only four languages show a posterior probability of being directly ancestral of >0.01%: Classical Armenian (as directly ancestral to modern Armenian) and three historical varieties of Greek [Mycenaean, Ancient Greek (the Attic dialect), and New Testament Greek]. See table S5.2.

forming a joint clade, however, is very limited (a posterior probability of only 25.9%). All three of the deepest clades have <26% support, in line with the lack of consensus among linguists. This may reflect complex "dialect continua" in the early stages of Indo-European (*40*). Toward the tips of the tree, into the historical period when language relationships are most reliably known, our results generally make for a close fit with established classifications, such as the relationships between ancient languages in the Greek clade. Within the major clades, most of the expected subgroups are also returned. In Romance, for example, the Romanian and Sardinian branches are the earliest to split off. Iberian Romance is also returned as a subgroup, as are North, West, and East Germanic; East and West Slavic; and Goidelic and Brythonic Celtic. Finally, we note some parts of our maximum clade credibility (MCC) tree that are not in line with established classifications. The Nuristani languages of the Hindu Kush, for instance, are nested more closely with their Indic neighbors than expected on the basis of other linguistic data, particularly phonology. Within Continental West Germanic, Frisian and historical varieties of German appear misplaced, as do various languages within Southwestern Iranic. The supplement (SM section 8) provides full discussion of unexpected parts of the topology.

Third, we ran a wide range of analyses to test the robustness of our results to alternative approaches. To identify the best-fitting model of cognate evolution, we first compared four models (M1 to M4). Our M1 analysis used a continuous-time Markov chain (CTMC) model for binary data, with gamma rate heterogeneity. Our M2 to M4 analyses all used a binary covarion model, which allows cognates to switch between fast and slow rates at points on the phylogeny, enabling languages to undergo bursts of change. M2 to M4 each used a different site model to accommodate variation in rates of cognate change. M2 used one rate for all meanings, M4 allowed a different rate for every meaning, and M3 was an intermediate, compromise approach using eight different mutation rates, according to the number of cognate sets per meaning (in bins of 1 to 10, 11 to 20, etc.). As shown in Fig. 4 (M1 to M4), results for the estimated time depth of Indo-European were similar across all four models. To identify which model performed best, we used path sampling to estimate the marginal log likelihood of each analysis (*41*). The best-performing model was M3—the binary covarion with binned rates (see table S5.4)—so we took this as our main analysis, for which we report the results here.

To further test the robustness of our results, we continued with this best-fitting model, M3, but varied the analysis in a series of other respects: our sensitivity analyses SA1 to SA10 (Fig. 4). In SA1, we addressed two particularly uncertain date calibrations. Vedic Sanskrit and Avestan are among the oldest languages in IE-CoR and thus offer especially deep calibration points. Their dating is controversial, however, because no original manuscripts survive. We therefore reran our main (M3) model with these two deep calibrations removed. The effect on the root date for Indo-European was negligible: just 94 years (1.16%) older, at 8214 yr B.P. (6785 to 9571 yr B.P.; Fig. 4, SA1). We also repeated the main analysis with the dataset adjusted to an alternative handling of one type of horizontal transmission (parallel loanwords) between language taxa (Fig. 4, SA2). Again, the effect on the root age estimate was minimal: 7934 yr B.P. (6487 to 9455 yr B.P.), that is, 186 years (2.29%) younger.

We further tested the robustness of our results to conditioning on the root (the first branching event), rather than on the origin (the beginning of the root branch) as in previous analyses (*13*, *42*). This led to a median root age 690 years (8.52%) older, with more uncertainty: 8812 yr B.P. (6648 to 11,419 yr B.P.; Fig. 4, SA3). Counting discrete language taxa is complex, given the clinal nature of the distinction between language and dialect, so we also tested alternative values for the prior distribution on the sampling probability at present (Fig. 4, SA4). In the main analysis, we assumed an underlying present-day language diversity of between 400 and 600 languages across Indo-European (*1*, *2*). Varying this assumption does not substantially affect the root age (8120 yr B.P.). Assuming 200 to 400 languages present today gives a root age of 8064 yr B.P. (6582 to 9585 yr B.P.), or 56 years (0.69%) younger (Fig. 4, SA4a). Assuming 600 to 800 languages gives 8177 yr B.P. (6838 to 9595 yr B.P.), or 57 years (0.70%) older (Fig. 4, SA4b). For some ancient languages, the surviving text corpora contain limited data, potentially biasing the analyses. We therefore ran a further sensitivity analysis (Fig. 4, SA5) without the 10 languages most affected by missing data; this gave a root date just 2 years (0.02%) younger, confirming that our main analysis is robust to the high proportions of missing data in such languages.

Our topologies are based on the data type most tractable for estimating chronology: cognacy in core vocabulary (*27*, *38*). Established language classifications are based mainly on phonology and morphology, however. Evolutionary histories need not coincide exactly on these different levels of language. Where our cognacy trees most depart from established classifications (for the Nuristani languages, Southwestern Iranic, and within West Germanic; see SM section 7.1), we tested the effect of applying lower-order clade constraints

to enforce a topology in line with uncontroversial phonological and morphological criteria (Fig. 4, SA6a). This moved the median Indo-European root date 804 years earlier (9.90% older). Separately, we applied higher-order constraints on the deepest relationships between all primary branches of Indo-European, to enforce a topology taken to support the Steppe hypothesis (*5*) (Fig. 4, SA6b). This moved the root date estimate 444 years earlier (5.47% older), further away from the steppe chronology.

With previous Indo-European datasets, enforcing ancestry constraints led to substantially younger root age estimates, enough to bring them into the time range predicted by the Steppe hypothesis (*12*). To test the impact of enforcing direct ancestry on our new IE-CoR dataset, we implemented three different ancestry-constrained analyses (SM section 7.5). In our main analysis, only four languages had >0.01 support for being direct ancestors. Enforcing those as ancestry constraints, and even adding the next (Old English, with support at only 0.0024), had minimal effect on the root date distribution, shifting the median estimate later by just 46 years (0.57% younger) (Fig. 4, SA7b, and table S7). If, contrary to our findings, written Classical Latin is nonetheless constrained to be directly ancestral to spoken Romance, the median root date moves later by 331 years (4.08% younger; Fig. 4, SA7a), to 7889 yr B.P.; but within Romance, the first splits to Romanian and Sardinian are then too late to be compatible with historical and linguistic indications (SM section 6.5). Even if we constrain all 27 IE-CoR languages remotely conceivable as direct ancestors, the root shifts later only by 506 years (6.23% younger), to 7614 yr B.P. (6239 to 9182 yr B.P.; Fig. 4, SA7c). Therefore, with the IE-CoR dataset, ancestry constraints do not lead to radically younger root ages.

This robustness to ancestry constraints is driven by the greater consistency of IE-CoR compared with the earlier Indo-European Lexical Cognacy (IELex) dataset (*11*, *12*). To confirm this, we took the "broad" (*12*) subset of IELex with its associated clade constraints (*12*) and applied to it our main, ancestry-enabled analysis model and tree prior, with (SA8b) and without (SA8a) the eight suggested ancestry constraints (*12*). This confirmed that with IELex, unlike with our IE-CoR dataset, enforcing direct ancestry does move the median root date estimate into a far more recent time frame, younger by 3632 years (42.1%), from 8629 yr B.P. (Fig. 4, SA8a) to 4997 yr B.P. (Fig. 4, SA8b). This contrast in the IELex dataset being far more sensitive to ancestry constraints than our IE-CoR dataset is explained by comparing the terminal branch lengths to the putative ancestor languages in the ancestry-enabled analyses for each dataset (fig. S7.8).

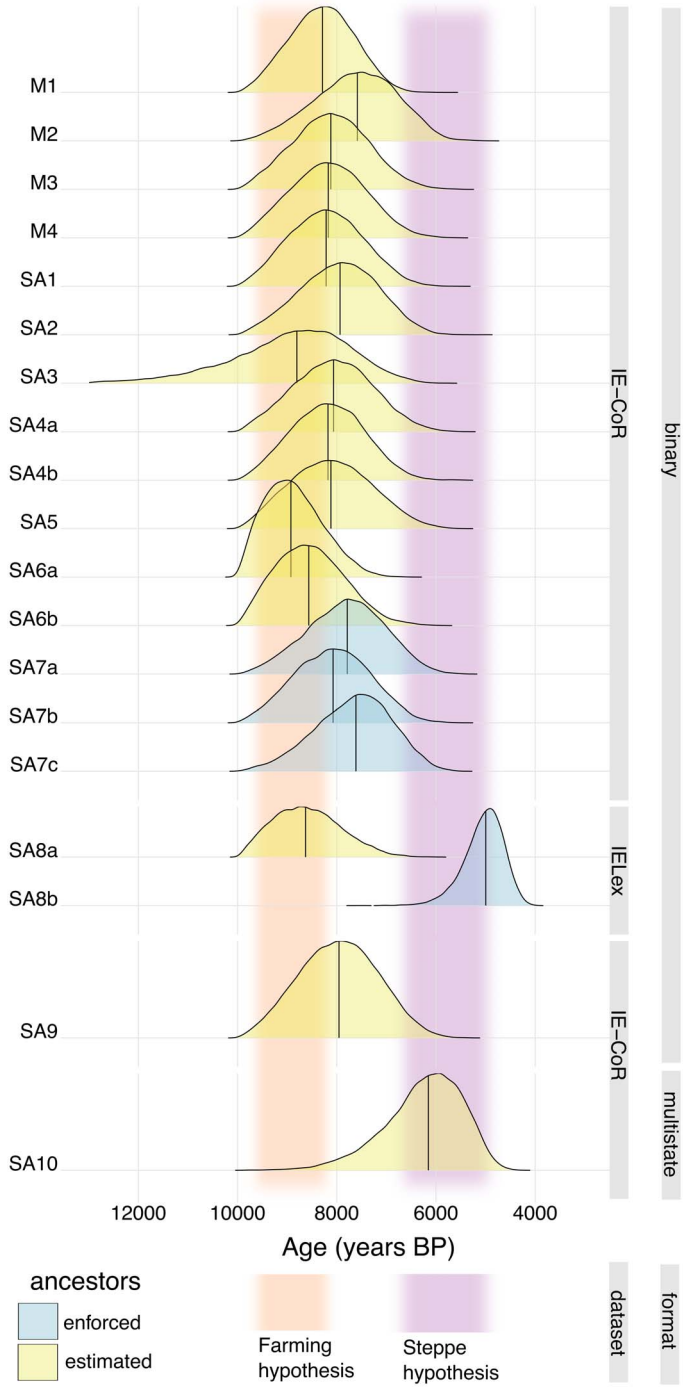


Fig. 4. Posterior probability distributions of the estimated age of Indo-European compared across all four models tested (M1 to M4), and all 10 sensitivity analyses (SA1 to SA10) as robustness tests. M1: Continuous-time Markov chain (CTMC) model for binary data, with gamma rate heterogeneity. M2: Binary covarion model for binary data, with a single joint mutation rate for all meanings. M3: Binary covarion model with eight different mutation rates, according to the number of cognate sets per meaning (in bins of 1 to 10, 11 to 20, and so on). M4: Binary covarion model with a distinct mutation rate for each of the 170 IE-CoR meanings. SA1: With tip calibrations for Early Vedic and Younger Avestan removed. SA2: With parallel loans not excluded but rather coded as unique cognate sets. SA3: With the prior conditioned on the most recent common ancestor, not the origin. SA4a: With a sampling probability assuming 200 to 400 modern languages. SA4b: With a sampling probability assuming 600 to 800 modern languages. SA5: With 10 poorly attested languages removed. SA6a: With targeted lower-order clade constraints. SA6b: With higher-order clade constraints following the Ringe topology (*5*). SA7a: With an ancestry constraint for Latin only. SA7b: With ancestry constraints for the five languages with a >0 posterior probability of being ancestral. SA7c: With all 27 remotely possible ancestry constraints. SA8a: Using the "broad" subset (*12*) of the IELex database with ancestors enabled but not enforced. SA8b: Using the "broad" subset (*12*) of the IELex database with ancestry enforced. SA9: With 57 meanings removed, those for which ancestral state reconstruction (on analysis M3) showed polymorphism per meaning at the root. SA10: Using a multistate model of cognate evolution. All sensitivity analyses SA1 to SA9 are based on model M3, the best-performing model.

These terminal branches are far longer (in some cases by >3000 years) with the IELex "broad" dataset than with IE-CoR. This excess branch length is caused by large numbers of excess entries in the IELex database, representing not just the primary word for a given meaning in any one language but one or more additional words similar in meaning (i.e., near synonyms)

although not the primary term (*27*). In IELex, these near synonyms had been entered highly inconsistently across the different languages (see fig. S1.4 and SM section 1.4). In a phylogenetic analysis, these excess entries equate to additional gains (or losses) in cognate evolution. Where constraints force branch lengths to zero (i.e., direct ancestry), the artifactual gains or losses that would have fallen on these long terminal branches are instead pushed to occur above the constrained ancestor language, after its time calibration. This in turn inflates the estimates of rates of change across the tree [from a median of 0.0055 (0.0046–0.0066) to 0.0132 (0.0119–0.0145) changes per cognate set per thousand years], and these faster rate estimates result in younger root age estimates (*12*). With IE-CoR data, free of excess synonyms, results are much more robust to adding or removing ancestry constraints. A young age estimate for Indo-European resulted only from enforcing inappropriate ancestry constraints on a problematic dataset.

The artifacts that arise from excess synonyms are part of a wider methodological issue. Lexical evolution is multistate, but most phylogenetic analysis methods take input data in binary format. IE-CoR follows strict protocols to ensure data consistency very close to a target of only the single primary cognate set present per meaning per language. (IE-CoR can and does admit cases of absolute synonymy in meaning and usage, but these are rare.) To test for the impact of polymorphism, we used ancestral state reconstruction to identify any meanings for which our main covarion model did in fact "reconstruct" more than one cognate set per meaning at the root. In SA9, we reran the main analysis but with these "root polymorphism" meanings excluded, leaving a remaining subset of 113 of the original 170 IE-CoR meanings. The effect on the root age was minimal: just 255 years (3.11%) younger, at 7955 yr B.P. (6427 to 9436 yr B.P.; Fig. 4, SA9).

A more radical alternative is to switch to a different phylogenetic model that does directly take multistate characters as its input data, rather than binary ones. We devised a multistate model and applied it to the IE-CoR dataset, as SA10. This model did return notably younger root date estimates: 2057 years (25.1%) younger, at 6153 yr B.P. (4926 to 7884 yr B.P.; Fig. 4, SA10), and thus within the range of the original Steppe hypothesis (*5*). This contrast results in particular from a difference in how the models handle polymorphism. Our main binary covarion model does in effect admit polymorphism per meaning, where supported by the data (typically over a period of transition from one word to another as the primary term for a given meaning). For analysis SA10, however, the multistate model required an assumption of no polymorphism at any stage in the tree. In multiple respects, the results indicate that this represents a relatively serious model misspecification. Assessed against established classifications for the Indo-European family, the topology and the chronology (relative and absolute) of the multistate tree are far more unexpected and problematic than the tree from the main binary covarion model. For example, the multistate model returns Tocharian as a late branch, deeply nested within the Indo-European tree together with Albanian, and fails to distinguish East from West Slavic correctly.

Furthermore, in almost all cases where language splits can be historically dated, the multistate model seriously underestimates the time depth of those splits, compressing the chronology across the board. As a further qualitative performance benchmark, we used ancestral state reconstruction in BEAST2 to identify any innovations inferred on the terminal branch to each ancient language. The covarion model returned expected results, pinpointing cognate sets unique to individual language taxa. The multistate model failed to return many of these as innovations, clearly indicating a model misspecification and revealing why the multistate model underestimates time depths. We therefore retain our main results from the binary covarion model (see SM section 7.10 for details and further reasons).

Interpretation

Our robust support for a root date estimate of ~8120 yr B.P. (6740 to 9610 yr B.P.) has major implications for the origins of the Indo-European family, the prehistory of Eurasia, and the interpretation of the latest aDNA results. The Indo-European question centers on where the PIE ancestor language was originally spoken, before any of its first branches diverged outward. The main rival theories are named and defined by where they place that ultimate homeland: the Steppe hypothesis or the Anatolian hypothesis (see Boxes 2 and 3).

Ancient DNA findings do support major expansions into north-central Europe out of not just the Pontic-Caspian Steppe (*16*) but also the Forest Steppe (*39*), dated to between 5000 and 4500 yr B.P. and associated with the Corded Ware culture (*16*). Our results show full support (100% posterior probability) for some of the main European branches of Indo-European remaining in a deep common clade until approximately this time depth. Germanic and Celtic are estimated to have diverged from

Box 2. Linguistics, archaeology, and genetics.

Although Indo-European is a linguistic concept, it was principally archaeologists who set out and developed the best-known competing theories on its origins: the Steppe hypothesis (*7*, *65*, *69*) and the Anatolian, or "farming," hypothesis (*6*, *70*). Most recently, aDNA has brought revolutionary new results and perspectives and can provide chronological constraints and estimates for the magnitude of migratory events in the past.

Linguistics, archaeology, and genetics use very different data and methods, however. Their different, partial records of the past can complement each other, but correlating them is not straightforward. "Cultures" inferred from the archaeological record do not match one-to-one with languages. Similarly, both matches and mismatches can arise between linguistic and genetic lineages, because languages can spread either demically or culturally (see SM section 2.1.2) (*9*). Findings in one discipline thus do not constitute proof or direct support of those in another but can be less or more compatible with competing hypotheses for Indo-European prehistory.

Speakers of Indo-European languages do not form a genetically homogeneous population. There is no single, consistent genetic profile from Iceland to Bangladesh. Realistically, only some partial ancestry component may be common to all or most speakers of Indo-European languages through time and space. Current debate boils down to which of two potential "tracer dyes" makes for the best fit with (Proto-)Indo-European.

- The ancestry profile of Yamnaya culture populations on the Pontic-Caspian Steppe spread widely during the Bronze Age, from ~5000 yr B.P. This profile is a roughly equal (ad)mixture of two earlier ancestries: the Eastern (European) hunter-gatherer (EHG) ancestry originally dominant in Pontic-Caspian and the Caucasus hunter-gatherer (CHG)/Iranian Neolithic ancestry that admixed into the Pontic-Caspian from ~7000 yr B.P.
- This CHG component alone is an alternative candidate for the Indo-European tracer dye. It is first found south of the Caucasus but from ~7000 yr B.P. onward also reached the Pontic-Caspian Steppe. Unlike EHG, the CHG component was also high in Anatolia at the time of the Hittites, who spoke the Anatolian branch of Indo-European, and remains high among speakers of the Indo-Iranic branch to this day.

However, these ancestry components are themselves not static singular entities. Rather, they represent momentary snapshots in time in prehistory, each emerging from preceding forms, and mixtures thereof. Genetic ancestry is fluid and clinal, and a matter of resolution, and therefore challenging to track—and relate to language lineages—unambiguously over many millennia.

each other ~4890 yr B.P. (3720 to 6190 yr B.P.), and Italic from them somewhat earlier, ~5560 yr B.P. (4230 to 6980 yr B.P.). Balto-Slavic is less closely associated with these three, splitting earlier, ~6460 yr B.P. (5040 to 7940 yr B.P.).

The Albanian, Greek, Armenian, and Anatolian branches, however, all separate from this main European clade much deeper in the tree—with mean age estimates long before "steppe" ancestry spread into Europe. So, in both chronology and phylogeny, this expansion from the steppe appears as a secondary phase that carried only some branches of Indo-European into Europe. This is consistent with aDNA findings in other regions that do not support the predictions of the hypothesis that all Indo-European originated on the steppe (*43*). Currently, aDNA evidence does not support a migration from the steppe through the Balkans into Anatolia (*20*, *22*), where traces of steppe ancestry are conspicuously absent in the Bronze Age (*21*–*23*). Steppe ancestry is also largely absent in ancient Greek Early Bronze Age individuals, who instead carry some Early European farmer–like ancestry, and ~25% Caucasus hunter-gatherer/Iranian-like ancestry (*19*, *44*). [The latter was first reported as maximized in hunter-gatherers from the South Caucasus (*45*) and early herders/farmers in northwestern Iran (*46*, *47*), particularly the Zagros, hence the label "CHG/Iranian."] Steppe ancestry up to 50% is attested in Greece only after ~4000 yr B.P. in Middle and Late Bronze Age (Mycenaean) individuals (*19*), with an admixture date estimate of ~4600 to 4000 yr B.P. Ancient Armenians carry predominantly a mix of mostly CHG/Iranian-like (40 to 60%) and Anatolian Neolithic-like ancestry (20 to 40%) and receive only a late contribution of steppe ancestry during the Late Bronze Age, ~3500 to 3000 yr B.P. [as indicated by the appearance of ~15% Eastern (European) hunter-gatherer (EHG) ancestry], which drops to low proportions at ~2000 yr B.P. (*44*, *46*, *48*).

Steppe ancestry, in the form of a mix of EHG+CHG/Iranian-like ancestry, thus did not reach Greece and Armenia until long after the population movements into northern and central Europe out of the Pontic-Caspian Steppe and Forest Steppe ~5000 yr B.P. In our phylogenetic results, Greek and Armenian show no close relationship to the main branches in Europe that plausibly fit with expansion from the steppe: Germanic-Italic-Celtic and possibly Baltic-Slavic. Earlier, however, during the Chalcolithic and Eneolithic periods ~6500 to 5500 yr B.P., CHG/Iranian-like ancestry had already spread across Anatolia, the Caucasus, northern Mesopotamia, and southeastern Europe and had also come to form an integral part of the genomic landscape in the North Pontic region during the Steppe Eneolithic. This expansion of CHG/Iranian-like ancestry represents an alternative candidate for spreading early branches of Indo-European in these regions.

Results from aDNA research thus cannot be fully reconciled with the idea that PIE, and all branches, ultimately originated on the steppe. Recent interpretations of the aDNA record (*5*, *49*) nonetheless continue to follow a recent formulation of the Steppe hypothesis (*5*) that keeps the steppe as the ultimate homeland and posits a corresponding tree topology (*5*, *50*, *51*), albeit one that does not command linguistic consensus. In particular, in this hypothesis, Indo-Iranic, the major eastern branch of Indo-European, was one of the last two main branches to emerge, out of a final major clade with Balto-Slavic. Our results contradict this in both chronology and tree topology. Indo-Iranic branches off early, ~6980 yr B.P. (5650 to 8400 yr B.P.), and support for a common clade with Balto-Slavic is minimal, with a posterior probability of only 12.3%. Recent aDNA data from Central and South Asia have sought to trace movements of people into Western and South Asia by migrations southward from the steppe. However, for the period 4300–3700 yr B.P., samples from the Bactria-Margiana Archaeological Complex (BMAC) do not yet attest to any such southward migration (*49*). Steppe ancestry is not found until ~3500 yr B.P., in the Gandhara Grave Culture in northern Pakistan, and only at limited proportions (*49*). The interpretation that this ancestry can be identified with the first Indo-Iranic dispersal into South Asia (*49*) is not straightforwardly compatible with our earlier date for the separation of Indo-Iranic from the rest of Indo-European (~6980 yr B.P.). We also find that Indic and Iranic had diverged from each other already by ~5520 yr B.P. (4540 to 6800 yr B.P.). To reconcile this with a steppe origin would require an alternative scenario in which Indic and Iranic split from each other approximately two millennia before entering South Asia and Western Asia.

Our analysis indicates that the Indo-European family began with a series of major branching events in relatively quick succession. From ~8120 yr B.P. (6740 to 9610 yr B.P.) to 6140 yr B.P. (4540 to 7880 yr B.P.), Indo-European had split into seven branches (see Table 1 and fig. S6.1), long before "steppe" ancestry spread into Europe and the Altai. These seven include the Anatolian, Greco-Armenian, and Indo-Iranic branches, for which aDNA shows little or no genetic influx from the steppe at ~5300 to 4900 yr B.P.—that is, at time depths early enough to match our estimated split times. Ancient DNA does, however, indicate a spread of CHG/Iranian ancestry in the opposite direction, from south of the Caucasus into the steppe at ~7000 to 6200 yr B.P. (*48*), which created the diagnostic "steppe" mix of ancestries that would later also enter Europe, ~5000 to 4500 yr B.P. This CHG/Iranian component is found first south of the Caucasus, including in the north to northeastern arc of the Fertile Crescent, among early farmers on the flanks of the Zagros Mountains in western Iran (*47*). The same CHG/Iranian (*48*) ancestry component also admixes heavily (by ~5000 yr B.P.) (*22*, *23*) into the region where languages of the Anatolian branch are first documented. CHG/Iranian is the dominant ancestry in ancient Armenia and Iran, in BMAC, and in most present-day populations who speak languages of the Iranic branch. It is also a major ancestry component among speakers of the Indic branch, particularly in regions furthest from the Dravidian-speaking (i.e., non–Indo-European) south of India. Thus, it is the CHG/Iranian ancestry component that most strongly connects the past populations who potentially spoke the branches of Indo-European in Europe and south (and east) of the Caucasus. Our earlier date estimates for the separation of Indo-Iranic from other Indo-European languages (*49*, *52*) are in line with this scenario.

Together, our linguistic results and the aDNA data are fully compatible with neither the Steppe hypothesis (Fig. 1B) nor the farming hypothesis (Fig. 1C). Instead, we propose a hybrid hypothesis (Fig. 1D) in which Indo-European languages spread out of an initial homeland south of the Caucasus, in the northern Fertile Crescent (Box 3). Only one major branch spread northward onto the steppe and then across much of Europe. This proposal matches parts of an existing alternative "South Caucasus" hypothesis (*53*–*55*), but the tree topology differs. The first migration phases are substantially earlier, and the main migration to the steppe follows a different route, through the Caucasus rather than through Central Asia. Crucially, south of the Caucasus is where aDNA first locates the only ancestry component found at high proportions in populations (past and present) associated with both Indo-Iranic and the main European branches of Indo-European. This genetic ancestry also emerged in southeastern Europe during the Late Chalcolithic/Early Bronze Age and predated the spread of "steppe" ancestry. (The Paleo-Balkan branches of Indo-European were formerly spoken in this region, but too few records survive to include them in our dataset.) Our hybrid hypothesis posits that out of this homeland south of the Caucasus, from ~8120 yr B.P., PIE began to diverge as early migrations split it into multiple early branches. One of these branches could have taken Indo-Iranic eastward far earlier than the Steppe hypothesis presumes, but in line with the linguistic chronology in Fig. 3, in which Indo-Iranic emerged as a distinct branch in the early phases of Indo-European divergence. Another main branch reached the steppe directly northward through the Caucasus ~7000 to 6500 yr B.P., compatible with one current interpretation of the aDNA record (*48*). The steppe became

Box 3. What's in a name? Shifting perceptions of the Steppe hypothesis.

The Indo-European question centers on where the common PIE ancestor language was originally spoken, before any of its first branches diverged outward. The main rival theories are named and defined by where they place that ultimate homeland: the Steppe hypothesis (*5*) contrasts with both the Anatolian hypothesis (*6*) and a lesser-known Armenian hypothesis (*53*, *54*).

In the Steppe hypothesis, all branches of Indo-European ultimately go back to migrations out of the Pontic-Caspian Steppe. By definition, this has included a steppe origin for the Anatolian and Tocharian branches (*5*).

Other hypotheses do recognize a prominent role for the steppe, as a staging post for some branches of Indo-European heading either westward (*54*)—or eastward, in Renfrew's variant B (*6*). Nonetheless, these hypotheses reject the idea that all branches originated on the steppe. They instead posit that Indo-European owes its full scale and diversity to cultural and demographic developments not just on the Pontic-Caspian Steppe but ultimately to earlier, deeper causes in lands farther south, in the southern Caucasus or northern Fertile Crescent.

Early aDNA results did support one "massive migration" out of the steppe, into parts of Europe, although it was qualified as "a" source for "at least some" Indo-European languages "in Europe" (*16*). As the aDNA record has grown, interpretations have continued to hold back from identifying the steppe as the source of all branches, notably Indo-Iranic (*45*) and especially Anatolian (*21*, *23*, *24*).

Anatolian is often hypothesized as first to branch off from the rest of the family, followed by Tocharian. There is no full linguistic consensus on this, but "Anatolian first" has led to alternative names and qualifications that can cloud the homeland issue. If (only) extant or Late Indo-European emerged from the steppe, whereas extinct Anatolian and/or Tocharian did not, then strictly the steppe was not the original homeland. Even if the family is rebaptized "Indo-Anatolian" (*23*)—which reflects neither its geographic coverage nor a particular branching order—this does not change the basic question of where the original homeland of the family as a whole was. The relatedness of Anatolian within the family is not in doubt, so if it (or any other branches) did not originate on the steppe, then Indo-European origins lie not in the Steppe hypothesis proper but rather in some form of hybrid hypothesis.

a secondary homeland for the later Yamnaya- and Corded Ware–related expansions into parts of Europe and north-central Asia.

Our results do not directly identify by which route Indo-Iranic spread eastward, so it remains possible that this branch spread through the steppe and Central Asia, looping north around the Caspian Sea (Fig. 1D). Recent interpretations of aDNA argue for this (*49*, *52*), but some aspects of their scenario are not easy to reconcile with our linguistic findings. For example, Indo-Iranic is an early independent branch in our analyses, with no close relationship to Balto-Slavic (see Box 1 and SM section 7.6.2.1), so that argument in favor of a northern route falls away. Genetically, the ancestry of Indo-Iranic speakers also derives much more heavily from south of the Caucasus and from Neolithic Iran than from the Bronze Age steppe (*16*) (see Box 2). Previous interpretations of aDNA from one individual from the Indus Periphery sought to exclude a direct eastward route on the basis of the degree and timing of Anatolian admixture (*49*, *52*), but these have been superseded by methodological and analytical refinements, which no longer exclude this scenario entirely (*56*). More parsimonious geographically, at least, would be a route for Indo-Iranic directly eastward out of a South

Table 1. Estimated time depths of the 12 main well-attested clades of Indo-European and higher-order clades with high posterior probability support. All date estimates are given in years before present, meaning before 2000 CE. The "time depth as independent clade" dates for [Balto-Slavic] + [Italic + Germanic + Celtic], Indo-Iranic, Greco-Armenian, Anatolian, Tocharian, and Albanian are merely indicative, based on splits with <50% posterior support. Date estimates shown are the height_median and height_95%_HPD values in the MCC tree file; see also fig. S6.1.

Major clade (with high posterior probability support)	Time depth as independent clade (split from rest of Indo-European)			Time depth of divergence within clade (between languages attested)		
	Median (yr B.P.)	Posterior probability	95% HPD (yr B.P.)	Median (yr B.P.)	Posterior probability	95% HPD (yr B.P.)
(Proto-)Indo-European	–	–	–	8116	1	6735–9613
[Balto-Slavic] + [Italic + Germanic + Celtic]	6981	0.24	5645–8395	6465	0.63	5036–7944
[Italic + Germanic + Celtic]	6465	0.63	5036–7944	5564	1	4231–6984
Indo-Iranic	6981	0.24	5645–8395	5520	1	4535–6796
Greco-Armenian	6135	0.49	4540–7882	5310	0.86	3999–6930
Balto-Slavic	6465	0.63	5036–7944	3663	1	2531–5034
Anatolian	6932	0.26	5403–8613	4618	1	3857–5620
Indic	5520	1	4535–6796	4366	1	3640–5253
Iranic	5520	1	4535–6796	4110	1	3464–4894
Italic	5564	1	4231–6984	3431	1	2771–4286
Greek	5310	0.86	3999–6930	3364	1	3218–3609
Celtic	4889	0.87	3718–6193	3205	1	2515–3963
Baltic	3663	1	2531–5034	2439	1	1526–3484
Germanic	4889	0.87	3718–6193	2337	1	1931–2865
Tocharian	6932	0.26	5403–8613	1828	1	1495–2315
Armenian	5310	0.86	3999–6930	1578	1	1485–1851
Slavic	3663	1	2531–5034	1493	1	1222–1837
Albanian	6135	0.49	4540–7882	1067	1	468–1882

Caucasus homeland through the Iranian Plateau, south of the Caspian (Fig. 1D).

Ancient DNA provides evidence of past population expansions over the same broad contexts in time and space that saw the Indo-European languages diverge and spread. These aDNA data suggest that the steppe did play a major role in spreading some of the European branches, but they also confirm that (at least) the Anatolian branch did not originate there. This thus points to an ultimate homeland for the Indo-European family south of the Caucasus (*23*). The obvious remaining question is whether all branches other than Anatolian came from the steppe, or only some. For some branches, the past population expansions and admixture events detected in aDNA, and hypothesized as having spread those forms of Indo-European, had only limited genetic impact. Our Bayesian phylogenetic analyses show that those candidate population expansions also postdate the linguistic divergences. Ancient DNA and linguistic phylogenetics thus combine to suggest that the resolution to the 200-year-old Indo-European enigma lies in a hybrid of both the farming and Steppe hypotheses.

Methods summary

Linguistic methodology

The IE-CoR database stores data on cognate relationships (shared word origin) between 161 Indo-European languages, in a reference set of 170 basic meanings. Across these languages and meanings, IE-CoR has a total of 25,918 individual lexeme entries. These lexemes are analyzed into 5013 cognate sets. The linguistic data and supporting citations can be explored and downloaded at iecor.clld.org.

Databases used in previous phylogenetic analyses have been undermined by a series of identifiable failings. To solve these, IE-CoR introduces a series of innovations in the methodology of database design, data collection, and the coding of language data, for both vertical (cognate) and horizontal (loanword) transmission. First, in coverage of language taxa, IE-CoR sampling provides denser coverage of the Indo-European family: 161 languages, as opposed to 24 (*51*), 84 (*57*), 87 (*10*), 103 (*11*), and 52, 82, or 94 (*12*) languages in previous databases [for a comparative table, see table 1 in (*27*)]. Sampling is also more balanced across all main branches of the Indo-European family and fills in gaps in the geographical coverage of previous databases. IE-CoR does now cover, for example, extinct Iranic languages of the steppe and Central Asia, the Nuristani branch of Indo-Iranic languages, and Gaulish as a representative of ancient Continental Celtic. Coverage also prioritizes nonmodern languages (52 in IE-CoR), to provide deeper phylogenetic signal and a fuller range of calibration points for the chronological estimation.

The linguistic data in previous databases were encoded essentially by a single linguist (*51*, *57*) and have been criticized for poor data quality (*58*). IE-CoR coordinated more than 80 specialists in the languages and branches concerned. Past database methodology also led to datasets being inconsistently coded. In particular, some languages were encoded with a proliferation of synonymous lexeme entries. This created wide disparities in the number of cognate sets present per language (fig. S1). These disparities can skew the estimations of branch lengths, rates of evolution, and chronology in phylogenetic outputs (*27*) (SM section 1.4). IE-CoR applies a strict and low 5% tolerance limit for multiple synonymy, as well as a new methodology to minimize scope for data inconsistency across all coders, languages, and meanings. Data coding procedures follow explicit new consistency protocols for both lexeme determination in each language and cognate determination between languages. The IE-CoR set of 170 reference meanings was itself optimized, first with reference to quantitative analyses of worldwide stability and borrowability of lexical meanings (*59*), and secondly by applying the same IE-CoR consistency protocols to systematize the (re)definitions of all meanings, to give a narrower and unambiguous specification of the exact target sense of each. Finally, loanwords are instances of horizontal transmission between languages and thus a potential confound to phylogenetic analyses. IE-CoR introduces a methodology to address inadequacies in how previous datasets have analyzed loanwords. In particular, new data structures distinguish the different consequences, for phylogenetic analysis, when loan events either give rise to independent cognate sets of their own or drive parallel changes across multiple, already divergent languages. This database methodology is presented in full in the supplement (SM section 3).

Phylogenetic analysis

We use Bayesian phylogenetic inference (*60*) to estimate root ages and how many ancient languages are "sampled ancestors" (i.e., directly ancestral to modern ones). For details on the method, see (*61*). Specific details for the application to cognate data can be found in the supplementary materials of analogous previous work (*11*, *62*). Previous phylogenetic analyses of cognate data have assumed that no language in the dataset was directly ancestral to any other (*10*, *11*, *63*). Forcing the opposite assumption—that many ancient languages were directly ancestral—returned significantly different root estimates (*12*) as well as untenable clade age estimates in known historical cases. In this study, we employed a method that uses reversible jump proposals during the Markov chain Monte Carlo run, allowing ancient languages to switch from being ancestral to nonancestral, and vice versa (*25*). In this approach, the posterior probability that an ancient language is ancestral is the proportion of the posterior sample in which it is ancestral. The actual proportion does not necessarily fit the assumption that it is either zero (*10*, *11*, *63*) or 1.

Following earlier work (*11*, *62*, *63*), we used the covarion model (*64*) as a substitution model, and an uncorrelated relaxed clock with a log-normal distribution (*26*). We used path sampling (*41*) to a range of setups for the substitution model and obtained the best fit when the 170 IE-CoR meanings were binned by the number of cognate sets per meaning, and each bin was associated with a different mutation rate (fig. S5.3). The tree prior was parameterized by the quotient of a diversification rate and an extinction rate, the extinction rate itself, a sampling proportion through time, and a sampling probability at present (*12*). Together, these parameters drive the process that generates the tree, leading to older or younger trees, and more or fewer sampled ancestors. We assumed that the diversification rate γ and the extinction rate Δ are of the same order of magnitude (log-normal prior distribution with mean 0 and standard deviation 1 applied to the quotient γ/Δ). We applied a highly conservative Exp(0.2) prior distribution on the extinction rate, which translates to an average time to lineage extinction of 5000 years.

To estimate the sampling proportion, three time periods need to be considered: the time before 4400 yr B.P., when no ancient languages are sampled, where the sampling proportion is zero; the time after the youngest nonmodern language, after which the sampling proportion is also zero; and the time between those two boundaries, when ancient languages were indeed sampled. This "ancient sampling proportion" is bound by an uninformative uniform prior distribution between 0 and 1. The sampling probability at present (what proportion of all contemporary languages are actually covered in the IE-CoR database) is bound by an informative beta distribution ([109,400]), which assumes that the modern languages in our dataset are a subset of about 400 to 600 contemporary Indo-European languages. We also assumed that the origin—the start of the branch above the root of the tree—does not exceed 10,000 yr B.P., as an upper bound on the beginning of divergence between Indo-European languages.

REFERENCES AND NOTES

1. D. M. Eberhard, G. F. Simons, C. D. Fennig, Eds., *Ethnologue: Languages of the World* (SIL International, ed. 25, 2022).
2. H. Hammarström, R. Forkel, M. Haspelmath, S. Bank, Glottolog 4.7 (Max Planck Institute for Evolutionary Anthropology, 2022); https://glottolog.org/.
3. J. Diamond, P. Bellwood, Farmers and their languages: The first expansions. *Science* **300**, 597–603 (2003). doi: 10.1126/science.1078208; pmid: 12714734

4. L. Campbell, W. J. Poser, *Language Classification: History and Method* (Cambridge Univ. Press, 2008).
5. D. W. Anthony, D. Ringe, The Indo-European homeland from linguistic and archaeological perspectives. *Annu. Rev. Linguist.* **1**, 199–219 (2015). doi: 10.1146/annurev-linguist-030514-124812
6. C. Renfrew, *Archaeology and Language: The Puzzle of Indo-European Origins* (Jonathan Cape, 1987).
7. J. P. Mallory, *In Search of the Indo-Europeans* (Thames & Hudson, 1989).
8. J. Clackson, "The origins of the Indic languages: the Indo-European model" in *Perspectives on the Origin of Indian Civilization*, A. Marcantonio, G. N. Jha, Eds. (D.K. Printworld, 2013), pp. 259–287.
9. P. Heggarty, "Prehistory through language and archaeology" in *The Routledge Handbook of Historical Linguistics*, C. Bowern, B. Evans, Eds. (Routledge, 2015), pp. 598–626.
10. R. D. Gray, Q. D. Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439 (2003). doi: 10.1038/nature02029; pmid: 14647380
11. R. Bouckaert *et al*., Mapping the origins and expansion of the Indo-European language family. *Science* **337**, 957–960 (2012). doi: 10.1126/science.1219669; pmid: 22923579
12. W. Chang, C. Cathcart, D. Hall, A. Garrett, Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* **91**, 194–244 (2015). doi: 10.1353/lan.2015.0005
13. T. Rama, Three tree priors and five datasets: A study of Indo-European phylogenetics. *Lang. Dyn. Chang.* **8**, 182–218 (2018). doi: 10.1163/22105832-00802005
14. A. M. Ritchie, S. Y. W. Ho, Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. *J. Lang. Evol.* **4**, 108–123 (2019). doi: 10.1093/jole/lzz005
15. J. Clackson, G. Horrocks, *The Blackwell History of the Latin Language* (Wiley, 2007).
16. W. Haak *et al*., Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015). doi: 10.1038/nature14317; pmid: 25731166
17. I. Olalde *et al*., The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555**, 190–196 (2018). doi: 10.1038/nature25738; pmid: 29466337
18. I. Olalde *et al*., The genomic history of the Iberian Peninsula over the past 8000 years. *Science* **363**, 1230–1234 (2019). doi: 10.1126/science.aav4040; pmid: 30872528
19. I. Lazaridis *et al*., Genetic origins of the Minoans and Mycenaeans. *Nature* **548**, 214–218 (2017). doi: 10.1038/nature23310; pmid: 28783727
20. I. Mathieson *et al*., The genomic history of southeastern Europe. *Nature* **555**, 197–203 (2018). doi: 10.1038/nature25778; pmid: 29466330
21. P. B. Damgaard *et al*., 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018). doi: 10.1038/s41586-018-0094-2; pmid: 29743675
22. E. Skourtanioti *et al*., Genomic history of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus. *Cell* **181**, 1158–1175.e28 (2020). doi: 10.1016/j.cell.2020.04.044; pmid: 32470401
23. I. Lazaridis *et al*., The genetic history of the Southern Arc: A bridge between West Asia and Europe. *Science* **377**, eabm4247 (2022). doi: 10.1126/science.abm4247; pmid: 36007055
24. M. Price, Finding the first horse tamers. *Science* **360**, 587 (2018). doi: 10.1126/science.360.6389.587; pmid: 29748263
25. A. Gavryushkina, D. Welch, T. Stadler, A. J. Drummond, Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLOS Comput. Biol.* **10**, e1003919 (2014). doi: 10.1371/journal.pcbi.1003919; pmid: 25474353
26. A. J. Drummond, S. Y. W. Ho, M. J. Phillips, A. Rambaut, Relaxed phylogenetics and dating with confidence. *PLOS Biol.* **4**, e88 (2006). doi: 10.1371/journal.pbio.0040088; pmid: 16683862
27. P. Heggarty, Cognacy databases and phylogenetic research on Indo-European. *Annu. Rev. Linguist.* **7**, 371–394 (2021). doi: 10.1146/annurev-linguistics-011619-030507
28. E. Finegan, "English" in *The World's Major Languages*, B. Comrie, Ed. (Routledge, ed. 2, 2009), pp. 59–85.
29. R. Lazzeroni, "Sanskrit" in *The Indo-European Languages*, A. G. Ramat, P. Ramat, Eds. (Routledge, 1998), pp. 98–124.
30. C. P. Masica, *The Indo-Aryan Languages* (Cambridge Univ. Press, 1991).
31. J. N. Adams, *The Regional Diversification of Latin 200 BC - AD 600* (Cambridge Univ. Press, 2007). doi: 10.1017/CBO9780511482977
32. J. Clackson, "Latin as a source for the Romance languages" in *The Oxford Guide to the Romance Languages*, A. Ledgeway, M. Maiden, Eds. (Oxford Univ. Press, 2016), pp. 3–13.
33. S. N. Dworkin, "Lexical stability and shared lexicon" in *The Oxford Guide to the Romance Languages*, A. Ledgeway, M. Maiden, Eds. (Oxford Univ. Press, 2016), pp. 577–587.
34. P. Mackridge, "Modern Greek" in *A Companion to the Ancient Greek Language*, E. J. Bakker, Ed. (Wiley, 2010), pp. 564–587.
35. G. Horrocks, *Greek: A History of the Language and its Speakers* (Wiley, 2009).
36. P. Sims-Williams, Genetics, linguistics, and prehistory: Thinking big and thinking straight. *Antiquity* **72**, 505–527 (1998). doi: 10.1017/S0003598X00086932
37. IE-CoR Database, Cognacy overlap between Early Vedic and Younger Avestan (2023); https://iecor.clld.org/values?sSearch_3=Avestan:+Younger,Vedic:+Early.
38. S. J. Greenhill, P. Heggarty, R. D. Gray, "Bayesian phylolinguistics" in *The Handbook of Historical Linguistics, Volume II*, R. D. Janda, B. D. Joseph, B. S. Vance, Eds. (Wiley-Blackwell, 2020), pp. 226–253.
39. L. Papac *et al*., Dynamic changes in genomic and social structures in third millennium BCE central Europe. *Sci. Adv.* **7**, eabi6941 (2021). doi: 10.1126/sciadv.abi6941; pmid: 34433570
40. A. Garrett, "Convergence in the formation of Indo-European subgroups: phylogeny and chronology" in *Phylogenetic Methods and the Prehistory of Languages*, P. Forster, C. Renfrew, Eds. (McDonald Institute for Archaeological Research, 2006), pp. 139–151.
41. G. Baele *et al*., Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29**, 2157–2167 (2012). doi: 10.1093/molbev/mss084; pmid: 22403239
42. C. Zhang, T. Stadler, S. Klopfstein, T. A. Heath, F. Ronquist, Total-evidence dating under the fossilized birth–death process. *Syst. Biol.* **65**, 228–249 (2016). doi: 10.1093/sysbio/syv080; pmid: 26493827
43. P. Heggarty, "Indo-European and the ancient DNA revolution" in *Talking Neolithic: Proceedings of the workshop on Indo-European origins held at the Max Planck Institute for Evolutionary Anthropology, Leipzig, December 2-3, 2013*, G. Kroonen, J. P. Mallory, B. Comrie, Eds., vol. 65 of *Journal of Indo-European Studies Monograph Series* (Institute for the Study of Man, 2018), pp. 120–173.
44. F. Clemente *et al*., The genomic history of the Aegean palatial civilizations. *Cell* **184**, 2565–2586.e21 (2021). doi: 10.1016/j.cell.2021.03.039; pmid: 33930288
45. E. R. Jones *et al*., Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015). doi: 10.1038/ncomms9912; pmid: 26567969
46. I. Lazaridis *et al*., Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424 (2016). doi: 10.1038/nature19310; pmid: 27459054
47. F. Broushaki *et al*., Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499–503 (2016). doi: 10.1126/science.aaf7943; pmid: 27417496
48. C.-C. Wang *et al*., Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat. Commun.* **10**, 590 (2019). doi: 10.1038/s41467-018-08220-8; pmid: 30713341
49. V. M. Narasimhan *et al*., The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019). doi: 10.1126/science.aat7487; pmid: 31488661
50. D. A. Ringe, T. Warnow, A. Taylor, Indo-European and computational cladistics. *Trans. Philol. Soc.* **100**, 59–129 (2002). doi: 10.1111/1467-968X.00091
51. L. Nakhleh, D. Ringe, T. Warnow, Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* **81**, 382–420 (2005). doi: 10.1353/lan.2005.0078
52. V. Shinde *et al*., An ancient Harappan genome lacks ancestry from Steppe pastoralists or Iranian farmers. *Cell* **179**, 729–735. e10 (2019). doi: 10.1016/j.cell.2019.08.048; pmid: 31495572
53. T. V. Gamkrelidze, V. V. Ivanov, *Indoevropejskij jazyk i indoevropejcy: Rekonstrukcija i istoriko-tipologičeskij analiz prajazyka i protokultury* [The Indo-European language and the Indo-Europeans: A Reconstruction and Historical-Typological Analysis of a Proto-Language and a Proto-Culture] (Tbilisi Univ. Press, 1984).
54. T. V. Gamkrelidze, V. V. Ivanov, *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and a Proto-Culture* (De Gruyter Mouton, 1995).
55. T. V. Gamkrelidze, V. V. Ivanov, Indo-European homeland and migrations: Half a century of studies and discussions [In Russian with English summary]. *J. Lang. Relatsh.* **9**, 109–136 (2013).
56. R. Maier *et al*., On the limits of fitting complex models of population history to *f*-statistics. *eLife* **12**, e85492 (2023). doi: 10.7554/eLife.85492; pmid: 37057893
57. I. Dyen, J. B. Kruskal, P. Black, An Indoeuropean classification: A lexicostatistical experiment. *Trans. Am. Philos. Soc.* **82**, iii–132 (1992). doi: 10.2307/1006517
58. A. M. S. McMahon, R. McMahon, *Language Classification by Numbers* (Oxford Univ. Press, 2005).
59. U. Tadmor, "Loanwords in the world's languages: findings and results" in *Loanwords in the World's Languages: A Comparative Handbook*, M. Haspelmath, U. Tadmor, Eds. (De Gruyter Mouton, 2009), pp. 55–75. doi: 10.1515/9783110218442.55
60. R. Bouckaert *et al*., BEAST 2: A software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **10**, e1003537 (2014). doi: 10.1371/journal.pcbi.1003537; pmid: 24722319
61. A. J. Drummond, R. R. Bouckaert, *Bayesian Evolutionary Analysis with BEAST* (Cambridge Univ. Press, 2015).
62. R. R. Bouckaert, C. Bowern, Q. D. Atkinson, The origin and expansion of Pama-Nyungan languages across Australia. *Nat. Ecol. Evol.* **2**, 741–749 (2018). doi: 10.1038/s41559-018-0489-3; pmid: 29531347
63. R. D. Gray, A. J. Drummond, S. J. Greenhill, Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009). doi: 10.1126/science.1166858; pmid: 19164742
64. C. Tuffley, M. Steel, Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63–91 (1998). doi: 10.1016/S0025-5564(97)00081-3; pmid: 9401352
65. D. W. Anthony, *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World* (Princeton Univ. Press, 2007).
66. R. Coleman, A CA book review—*Archaeology and Language: The Puzzle of Indo-European Origins*. Colin Renfrew. *Curr. Anthropol.* **29**, 437–468 (1988).
67. E. Bryant, *The Quest for the Origins of Vedic Culture: The Indo-Aryan Migration Debate* (Oxford Univ. Press, 2001). doi: 10.1093/0195137779.001.0001
68. P. Heggarty, "Why Indo-European? Clarifying cross-disciplinary misconceptions on farming vs. pastoralism" in *Talking Neolithic: Proceedings of the workshop on Indo-European origins held at the Max Planck Institute for Evolutionary Anthropology, Leipzig, December 2-3, 2013*, G. Kroonen, J. P. Mallory, B. Comrie, Eds., vol. 65 of *Journal of Indo-European Studies Monograph Series* (Institute for the Study of Man, 2018), pp. 69–119.
69. M. Gimbutas, "Proto-Indo-European culture: the Kurgan culture during the 5th to the 3rd millennia B.C." in *Indo-European and Indo-Europeans*, G. Cardona, H. M. Koenigswald, A. Senn, Eds. (Univ. of Pennsylvania Press, 1970), pp. 155–198.
70. P. Bellwood, *First Farmers: The Origins of Agricultural Societies* (Blackwell, 2005).
71. R. R. Bouckaert, J. Heled, DensiTree 2: Seeing Trees Through the Forest. bioRxiv 012401 [Preprint] (2014). https://doi.org/10.1101/012401.
72. P. Heggarty, C. Anderson, H.-J. Bibiko, CLDF dataset derived from Heggarty, Paul & Anderson, Cormac & Scarborough, Matthew's "Indo-European Cognate Relationships database project" (IE-CoR) from 2019, version 1.0, Zenodo (2023); https://doi.org/10.5281/zenodo.8089434.
73. P. Heggarty *et al*., Guide to Supplementary Data and Results for Heggarty *et al*. (2023) in Science, version 1.0, Zenodo (2023); https://doi.org/10.5281/zenodo.8147476.

ACKNOWLEDGMENTS

The IE-CoR database was developed as a collaborative enterprise by a consortium of contributors who provided language data by making lexeme determinations for individual languages and/or cognacy determinations between languages. We thank all contributors to the IE-CoR database. The large majority of cognacy determinations at the broad and deep-time Indo-European level were made by M.Sc., with substantial contributions by B.I., R.P., and C.F. Cognacy determinations within specific branches of Indo-European were principally made by L.J. (Slavic), M.Sc. (Greek and ancient Italic), M.J.K. (mostly Iranic), T.J. (Iranic), C.A. (mostly Celtic), H.L. (Hindu-Kush Indic), R.F.S. (Nuristani), R.P. (Indic), G.H. (Iranic), R.T. (Indic), U. Geupel (Albanian), M.M. (Armenian), R.I.K. (Tocharian), A. Falileyev (Celtic), E.A. (Iranic), T.P. (Baltic), O.B. (Ossetic), T.K.D.-F. (Germanic), and M.B. (Germanic). Other contributors who made lexeme determinations for multiple languages in a given branch are: M.Se. (Anatolian), N.L. (Modern Hellenic), K.Sc. (Romance), B.I. (Celtic), N. Williams (Cornish), M. Findell (Germanic), S. Loi (Sardinian), P. Markus (Indic), G.K.G. (Indic), R.P. (Indic), N. Sims-Williams (Iranic),

R. Izadifar (Iranic), and S. Adibifar (Iranic). In some cases, a language expert made lexeme determinations for a single language (listed in alphabetical order by surname): G. Abete, P. Atanasov, E. Baiwir, M.-R. Bastardas, A. Benkato, L. Bevevino, G. Cadorini, L. Cheveau, C. Christodoulou, M. de Vaan, J. Delorme, S. Dworkin, C.F., M. Gheitasi, H. Hammarström, S. Hewitt, A. A. Khan, M. K. Khan, L. Khokhlova, D. Kim, C. Lewin, B. Lushaj, P. Mahmoudveysi, M. Mahommadirad, S. Mersch, J. Mock, B. Moustafa, F. Nemati, M. Nourzaei, P. Ó Muircheartaigh, M. Ourang, H. Pagan, T. Palmer, K. Rehman, G. Rhys, M. Zaman Sagar, L. Steensland, M. Taheri-Ardali, M. Talebi, S. Tittel, A. Verkerk, A. Versloot, P. Videsott, N. Vuletić, M. Widmer, and A. Zeini. The basic relational database structure for IE-CoR was inherited from the LexDB system and the IELex website developed by M. Dunn. The IE-CoR dataset was produced using a database creation system programmed by J. Runge and H.-J. Bibiko, to enter and analyze language data, perform cognate determination, and export nexus and calibration files. The IE-CoR database visualization website at https://iecor.clld.org was programmed primarily by H.-J. Bibiko, within the Cross-Linguistic Linked Data (CLLD) framework developed by R. Forkel. We thank A. Gavryushkina for advice on sampled ancestor prior probabilities and ancestry constraints. We thank M. O'Reilly for the preparation of the figures. Finally, we thank A. Garrett and W. Chang at the Department of Linguistics, University of California, Berkeley, for extensive comments and discussion of this research. **Funding:** This research was funded by the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany). From 11 September 2021 to 10 September 2022, P.H. was funded by the ERC Starting Grant "Waves" (ERC758967). E.A. and G.H. were partially funded by an Alexander von Humboldt Research Fellowship for Experienced Researchers (2016-2018, grant No. 3.1-CAN-1164714-HFST-E). E.A. was also partially funded by a Social Sciences and Humanities Research Council of Canada (SSHRC) Insight Development Grant (2015-2017, grant No. 430-2015-00031). **Author contributions:** R.D.G. initiated and coordinated the study. P.H. and C.A. designed the IE-CoR database and data collection methodology and coordinated the linguistic coding team. M.Sc. oversaw all determination of cognacy at the deep Indo-European level. C.A., M.Sc., L.J., M.J.K., T.J., B.I., R.P., H.L., R.F.S., G.H., M.M., R.I.K., E.A., T.P., O.B., T.K.D.-F., M.B., C.F., R.T., M.Se., N.L., K.St., K.Sc., and G.K.G. were major contributors to the 25,918 lexeme and 5013 cognate determinations in the IE-CoR database. R.B., B.K., S.J.G., and D.K. conducted the phylogenetic analyses, with input from R.D.G., Q.D.A., P.H., and C.A. W.H. and J.K. advised on the aDNA data. P.H., R.D.G., D.K., B.K., and C.A. wrote the text. All authors commented on the manuscript. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** The full IE-CoR cognate dataset for Indo-European languages used in this paper can be viewed and explored through our database app at https://iecor.clld.org. The full CLDF dataset of IE-CoR 1.0 can be freely downloaded at https://doi.org/10.5281/zenodo.8089434 (*72*). The .nexus and .xml data files used as input to each of the phylogenetic analyses are available within the supplementary data and results files, which are available online at https://share.eva.mpg.de/index.php/s/E4Am2bbBA3qLngC and at https://doi.org/10.5281/zenodo.8147476 (*73*)—see the Guide to Supplementary Data and Results Files and Online Resources .pdf file in the supplementary materials. Further details on how to reproduce our results are given in section 5.5 of the supplementary materials, on the pipeline from the raw IE-CoR data tables to the phylogenetic results reported here. From the raw IE-CoR data tables, we first exported a data file in the nexus format required as input to widely used quantitative and phylogenetic analysis software. This was done using the export script [make_nexus.py], written by H.-J. Bibiko and available at: https://github.com/lexibank/iecor/blob/master/iecorcommands/make_nexus.py. The Bayesian phylogenetic analysis software used in this paper, BEAST version 2.6.5, is available at www.beast2.org. Other specific code used is the BEAST2 sampled-ancestors package, available at https://github.com/CompEvol/sampled-ancestors. Sensitivity analysis SA7 used the additional AncestryConstraint.java code written by D.K., available at https://github.com/CompEvol/sampled-ancestors/blob/master/src/sa/evolution/tree/AncestryConstraint.java. Sensitivity analysis SA10 used a BEAST2 add-on package written by B.K. to implement a multistate model, the code for which is available at https://github.com/king-ben/ConceptModels. The input .xml files include the matrix of binary-encoded language cognate set data, date calibrations, the setup of the prior distributions, and the random seeds used in the analyses. Also available in the supplementary data and results files are the log files for all analysis runs, and the resulting posterior tree distributions. For full details, see the Guide to Supplementary Data and Results Files and Online Resources .pdf file in the supplementary materials.

SUPPLEMENTARY MATERIALS

Submitted 27 January 2021; accepted 8 June 2023
10.1126/science.abg0818

Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosaw Jocz, Martin Joachim Kmmel, Thomas Jgel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Mack, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroski, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Khnert, and Russell D. Gray

Science, **381** (6656), eabg0818.
DOI: 10.1126/science.abg0818

Editor's summary

Languages of the Indo-European family are spoken by almost half of the world's population, but their origins and patterns of spread are disputed. Heggarty *et al*. present a database of 109 modern and 52 time-calibrated historical Indo-European languages, which they analyzed with models of Bayesian phylogenetic inference. Their results suggest an emergence of Indo-European languages around 8000 years before present. This is a deeper root date than previously thought, and it fits with an initial origin south of the Caucasus followed by a branch northward into the Steppe region. These findings lead to a "hybrid hypothesis" that reconciles current linguistic and ancient DNA evidence from both the eastern Fertile Crescent (as a primary source) and the steppe (as a secondary homeland). —SNV