Title:

Language trees with sampled ancestors support an early origin of the Indo-European languages

One Sentence Summary:

Dated language phylogenies support a new "hybrid hypothesis" for the Indo-European language family.

Authors:

Paul Heggarty[1,2]*, Cormac Anderson[1]*, Matthew Scarborough[1], Benedict King[1], Remco Bouckaert[1,3], Lechosław Jocz[4], Martin Joachim Kümmel[5], Thomas Jügel[6], Britta Irslinger[7], Roland Pooth[8], Henrik Liljegren[9], Richard F. Strand[10], Geoffrey Haig[11], Martin Macák[10], Ronald I. Kim[12], Erik Anonby[13,14], Tijmen Pronk[14], Oleg Belyaev[15,16], Tonya Kim Dewey-Findell[17], Matthew Boutilier[18], Cassandra Freiberg[19], Robert Tegethoff[1,5], Matilde Serangeli[5], Nikos Liosis[20], Krzysztof Stroński[21], Kim Schulte[22], Ganesh Kumar Gupta[21], Wolfgang Haak[234], Johannes Krause[23], Quentin D. Atkinson[1,24], Simon J. Greenhill[1,25], Denise Kühnert[26]*, Russell D. Gray[1,24]*

*= corresponding author

Affiliations:

1. Dept of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.
2. Waves group, Dept of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.
3. Centre for Computational Evolution, University of Auckland, Auckland 1010, New Zealand.
4. Faculty of Humanities, Jacob of Paradies University, Teatralna 25, 66-400 Gorzów Wielkopolski, Poland.

5. Seminar for Indo-European Studies, Institut für Orientalistik, Indogermanistik, Ur- und Frühgeschichtliche Archäologie, Friedrich-Schiller-Universität Jena, Germany.
6. Institute of Empirical Linguistics, Goethe University Frankfurt, Senckenberganlage 31, 60325 Frankfurt am Main.
7. Saxon Academy of Sciences and Humanities, Karl-Tauchnitz-Straße 1, 04107 Leipzig.
8. Dept of Linguistics, Ghent University, Blandijnberg 2, 9000 Ghent, Belgium.
9. Dept of Linguistics, Stockholm University, Universitetsvägen 10 C, Frescati, 10691 Stockholm, Sweden.
10. Independent scholar.
11. Dept of General Linguistics, University of Bamberg, Schillerplatz 17, 96047 Bamberg, Germany.
12. Dept of Older Germanic Languages, Faculty of English, Adam Mickiewicz University in Poznań, Aleja Niepodległości 4, 61-874 Poznań, Poland.
13. School of Linguistics and Language Studies, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada.
14. Leiden University Centre for Linguistics, Postbus 9515, 2300 RA Leiden, Netherlands.
15. Department of Theoretical and Applied Linguistics, Lomonosov Moscow State University, 1st Humanities Building, 1-51 Leninskie Gory, 119991 GSP-1 Moscow, Russia.
16. Department of Typology, Institute of Linguistics RAS, Bolshoi Kislovsky Lane 1/1, 125009 Moscow, Russia.
17. Centre for the Study of the Viking Age, School of English, Trent Building, University of Nottingham, UK.
18. Dept of German, Nordic and Slavic, University of Wisconsin-Madison, Madison, Wisconsin.
19. Institut für deutsche Sprache und Linguistik, Sprach- und literaturwissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin.
20. Institute of Modern Greek Studies, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece.
21. Faculty of Modern Languages, Adam Mickiewicz University in Poznań, Aleja Niepodległości 4, 61-874 Poznań, Poland.
22. Dept of Translation and Communication, Jaume I University, 12006 Castelló de la Plana, Castelló, Spain.
23. Dept of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.
24. School of Psychology, University of Auckland, 23 Symonds St., Auckland 1010, New Zealand.
25. ARC Center of Excellence for the Dynamics of Language, ANU College of Asia and the Pacific, The Australian National University, Canberra, ACT 2600, Australia.

26. Transmission, Infection, Diversification & Evolution Group, Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, 07745 Jena, Germany.

* Correspondence to: Paul.Heggarty@gmail.com (P.H.), cormacanderson@gmail.com (C.A.), kuehnert@shh.mpg.de (D.K.), russell_gray@eva.mpg.de (R.G.)

Abstract

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 c. 9000 BP, while others support a spread with horse-based pastoralism out of the Pontic-Caspian Steppe c. 6000 BP. Here we present an extensive new database of Indo-European core vocabulary that eliminates past inconsistencies in cognate coding. Ancestry-enabled phylogenetic analysis of our new dataset indicates that few ancient languages are direct ancestors of modern clades, and produces a root age for the family of c. 8120 BP. While 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 northwards 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.

For Reviewers: confidential access codes to the IE-CoR database

URL: https://iecor.clld.org

username: cobl2

password: cobl2cobl2

Main Text:

The Indo-European language family contains over 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 over 200 years, the origin of the Indo-European language family has been hotly disputed (*3*). The deep link between the widely dispersed Indo-European languages was discovered over 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 BP, mostly with horse-based pastoralism from c. 5000 BP (*5*) (Fig. 1b). The farming hypothesis claims that Indo-Europeans dispersed with agriculture out of parts of the Fertile Crescent, beginning as early as c. 9500-8500 BP (*6*) (Fig. 1c). Linguistic reconstructions of some Proto-Indo-European 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*) (see SM §8). 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 significant problem with this analysis is that forcing direct ancestry produces date inferences lower down the tree that conflict with the known histories of several branches of Indo-European. The diversification of Romance, for example, is inferred to have started only 1000 years ago (*12*), when in fact regional differences had already begun to arise a millennium earlier, as Roman expansion led to "great diversity in the Latin that was spoken around the Empire" (*15*). Here we investigate, diagnose and resolve the problems in data quality that led to these artifacts.

Ancient human DNA (aDNA) is now also reshaping the debate. Results support a substantial influx of genetic ancestry from the Eurasian Steppe c. 5000 BP, which could have carried several of the main branches of Indo-European in Europe (*16–18*). However, this ancestry signal is less evident in aDNA findings from Mycenaean Greece (*19*), the Balkans (*20*) and Anatolia (*21, 22*), casting doubt on whether the Steppe hypothesis can explain the spread of all branches of the family,

especially in the eastern Mediterranean and in Asia. This fuller aDNA picture "does not support a classical way of looking at the steppe hypothesis" (*23*).

Here we overcome the limitations of previous linguistic analyses by combining recent advances in Bayesian phylogenetic inference with a completely new and far more extensive Indo-European dataset. First, we deploy a sampled ancestor phylogenetic analysis (*24*) that permits but does not force ancient languages to be directly ancestral to modern languages. This analysis is enabled by a birth-death-sampling tree prior in which sampled languages are not removed from the pool of diversifying languages (Fig. S4.2). Rather than assuming that ancient languages were the direct ancestors of their modern relatives, this approach estimates from the linguistic dataset itself the relative probability that any ancient language in our sample 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. Second, we identify artifacts in previous phylogenetic analyses that result from flaws and inconsistencies in the language datasets used. To resolve these, we implement a new methodology for encoding cognate data (see SM §2) to maximize consistency across the language dataset and optimize it as input to phylogenetic analysis, to create an entirely new database of Indo-European cognate relationships (IE-CoR). IE-CoR covers 161 languages, coded by over 80 specialists on languages of the Indo-European family, to provide much denser and more balanced sampling both within and between the main sub-clades of Indo-European. The 52 non-modern languages (Fig. 1a) provide a much denser set of date calibrations than earlier databases.

Results

The analysis (Fig. 2), using a relaxed clock and allowing rates of change to vary by sets of meanings, produced an estimated date for the root of the Indo-European family that is too early to be compatible with the Steppe hypothesis: c. 8120 BP, with a 95% credible region of 6740 to 9610 BP. (Date estimates are reported here as a median date BP, followed by the 95% credible region (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 non-modern languages in the IE-CoR database, 27 can 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. Fig. 3 shows the prior and posterior probabilities for each of these non-modern languages being a direct ancestor to any later language(s) in its clade (see also Table S5.2). Our ancestry-enabled analysis finds non-negligible (>0.01) posterior probabilities for only four languages: Classical Armenian, and three ancient forms of Greek. Only in two of these cases is the posterior probability over 50%. We found no support for the higher number of 8 direct ancestors enforced in previous analyses (*12*). The lexical data drive these results: in our prior, where direct ancestry probabilities ranged from c. 42% to 78% for all 27 potential ancestor languages, the median root date estimate was 5815 BP (4149-8123 BP). The data have driven this 2305 years earlier to our result of a median age of 8120 BP 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 ancestral to modern Icelandic. However, it is important to be clear on what a split between lineages represents in phylogenetic analyses of language wordlists. A split does not just correspond to the major difference between discrete, mutually unintelligible 'languages'. Instead, phylogenetic lineages split as soon as the first difference emerges in the predominant lexeme used for any meaning in the dataset. So even dialects or registers (written vs. spoken) of the 'same' language can split into different taxa, and thus ancestry between these taxa and contemporary languages may not be direct (SM §7).

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 (*25*). 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 Mediaeval India "do not derive from Sanskrit" (*26*), and specifically, "do not go back directly to the dialect which formed the basis of Vedic" (*26*), which stood apart as a "far-western dialect" (*27*). Importantly, 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. So even a near-direct ancestor may 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 (*28*). 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" (*29*). Even just one difference, in a single meaning of

the 170 in the IE-CoR reference set, logically entails a split, and that ancestry is not fully direct. In practice, "many Classical Latin words do not survive into Romance" (*15*), including in IE-CoR meanings such as MOUTH, while others survive only sporadically, in meanings 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. Likewise, written Old Icelandic is not quite directly ancestral to modern spoken Icelandic. This contradicts the assumptions used in earlier ancestry-constrained analyses (*12*). Only in four cases were specific written, historical languages (Classical Armenian and some forms of Ancient Greek (*30, 31*)) so close to the ancestor of later languages in their clades as to be near 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 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 c. 830 CE (470-950 CE), closely in line with the first Norse settlement of the Faroes and Iceland. Initial divergence within Romance is accurately dated to the Roman Empire in the first centuries CE. Divergence within Indic is dated to c. 4370 BP (3640-5250 BP), in line with Vedic Sanskrit already being slightly divergent from the lineage(s) ancestral to modern spoken Indic languages (*27*). The inference of an Indo-Iranic split at c. 5520 BP (4540-6800 BP) 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 (*32*), rather than quantified. In the precisely defined IE-CoR meanings, Early Vedic and Younger Avestan in fact share only 58.7% cognacy (*33*). This matches the level of cognacy that survives between the most divergent sub-lineages 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 BP, respectively. A time-depth two millennia greater (c. 5520 BP) for the split between their lineages (Indic vs. Iranic) is thus consistent with the 58.7% cognacy overlap between them.

Second, the language tree topology can be evaluated against established classifications of Indo-European languages. These traditional classifications identify 10-12 main attested subgroups: Anatolian, Tocharian, Albanian, Armenian, Greek, Indic+Iranic, Baltic+Slavic, Germanic, Italic, and Celtic. Our analyses (Fig. 2 and Fig. S3) returned all of these with 100% posterior probability, including the two widely recognized deeper clades, Indo-Iranic and Balto-Slavic. Beyond this, however, qualitative methodology in historical linguistics has failed to reach a consensus on how these 10 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 ten-way rake (*1*). 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*). 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 forming a joint clade, however, is very limited (a posterior probability of only 25.9%). All three of the deepest clades have less than 26% support, in line with the lack of consensus among linguists. This may reflect complex 'dialect continua' in the early stages of Indo-European (*34*). Towards 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, 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 than expected with their Indic neighbors. Within Continental West Germanic, Frisian and historical varieties of German appear misplaced, as do various languages within Western Iranic. The supplement (SM §7) provides full discussion of unexpected parts of the topology.

Third, we ran a series of sensitivity analyses (SA1 to SA8) to test the robustness of our results to alternative approaches, on eight levels (Fig. 4). 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 re-ran the analysis 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 BP (6785-9571 BP; Fig. 4, SA1). We also repeated the main analysis with 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 BP (6487-9455 BP), i.e. 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, 35*). This led to a median root age older by 690 years (8.52%), with more uncertainty: 8812 BP (6648-11419 BP; Fig. 4, SA3). Counting discrete language taxa is complex, given the clinal nature of the language/dialect distinction, 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 400-600 languages across Indo-European (*1, 2*). Varying this assumption does not affect the root age (8120 BP) significantly. Assuming 200-400 languages present today gives a root age of 8064 BP (6582-9585 BP), i.e. 56 years (0.69%) younger (Fig. 4, SA4a). Assuming 600-800 languages gives 8177 BP (6838-9595 BP), i.e. 57 years (0.70%) older (Fig. 4, SA4b). For some ancient languages, the surviving text corpus is too limited for a full dataset, in potentially biased ways. We therefore ran a further sensitivity analysis (Fig. 4, SA5) without the ten languages most affected. The root date moved just 2 years younger (0.02%), 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 (*36, 37*). Established language classifications are based largely on phonology and morphology, however, and 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, south-western Iranic and within West Germanic, see SM §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 older (9.90%). 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 older (5.47%), however, also further away from the Steppe chronology.

With previous Indo-European datasets, enforcing ancestry constraints led to significantly 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 a new form of ancestry-constrained analysis (SM §6.5). In our main analysis, only four languages had non-negligible (>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 just 46 years (0.57%) younger (Fig. 4, SA7b; Table S6). If, contrary to our findings, written Classical Latin is nonetheless constrained to be directly ancestral to spoken Romance, the median root date moves younger by 331 years (4.08%; Fig. 4, SA7a), to 7889 BP, but within Romance the first splits to Romanian and Sardinian are then too late to be compatible with historical and linguistic indications (SM §6.5). Even if we constrain all 27 IE-CoR languages remotely conceivable as direct ancestors, the root shifts younger only by 506 years (6.23%), to 7614 BP (6239–9182 BP; Fig. 4, SA7c). Therefore, with the new IE-CoR data-set, 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 to the earlier IELex dataset (*11, 12*). To confirm this, we took the 'broad' subset of IELex (*12*) with its associated clade constraints (*12*) and applied to it our main, ancestry-enabled analysis model (M3) and tree prior, with (SA8b) and without (SA8a) the suggested 8 ancestry constraints (*12*). This confirmed that with IELex rather than our new IE-CoR dataset, enforcing direct ancestry does move the median root date estimate much younger, by 3632 years (42.1%), from 8629 BP (Fig. 4, SA8a) to 4997 BP (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. S6.8). 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 excess synonyms in IELex (*37*)(see Fig. S1, SM §1.4), which in the analysis equate to gains/losses in cognate evolution. Where constraints force branch lengths to zero (i.e. direct ancestry), the artifactual gains/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/cognate set/kYr), 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 for Indo-European can be retrieved only by enforcing inappropriate ancestry constraints on a problematic dataset.

Interpretation

Our robust support for a root date estimate of c. 8120 BP (6740-9610 BP) 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 Proto-Indo-European ancestor language was originally spoken, before any of its first branches diverged outwards. The main rival theories are named and defined by where they place that ultimate homeland: the *Steppe* hypothesis, or the *Anatolian* hypothesis (see SM §9).

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 (*38*), dated to 5000-4500 BP 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 each other c. 4890 BP (3720-6190 BP), and Italic from them somewhat earlier, c. 5560 BP (4230-6980 BP). Balto-Slavic is less closely associated with these three, splitting earlier c. 6460 BP (5040-7940 BP).

The Albanian, Greek, Armenian and Anatolian branches, however, all separate from this main European clade much deeper in the tree — long before the expansion of 'steppe' ancestry into Europe. In both chronology and phylogeny then, this expansion 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 (*39*). Currently, aDNA evidence does not support a migration from the Steppe through the Balkans into Anatolia (*20, 22*). If this were the case we would expect to find not only clear traces of steppe ancestry in Anatolia (*21, 22*), but also that Anatolian should branch with other European languages, rather than producing the oldest split date. In addition, steppe ancestry is absent in ancient Greek Early Bronze Age individuals, who instead carry 25% "CHG/Iranian-like" ancestry (40), as do ancient Armenians (*40*). (This ancestry was first reported as being the predominant/main genetic component in samples from hunter-gatherers in the South Caucasus (*41*), and early herders/farmers in north-western Iran (*42, 43*), particularly the Zagros, hence the label "CHG/Iranian".) Steppe ancestry up to 50% is attested in Greece only after c. 4000 BP in Middle and Late Bronze Age (Mycenaean) individuals (*19, 44*), with an admixture date estimate of c. 4600-4000 BP.

Ancient DNA research thus suggests stepping back from assuming that Proto-Indo-European, and all branches, ultimately originated on the Steppe. Interpretations of the aDNA record (*5, 45–47*) have nonetheless continued to follow a recent formulation of the Steppe hypothesis (*5*) that keeps the Steppe as the ultimate homeland, and posits a corresponding tree topology, 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, c. 6980 BP (5650-8400 BP), and support for a common clade with Balto-Slavic is minimal, 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 southwards from the Steppe. However, for the period 4300-3700 BP, samples from the Bactria-Margiana Archaeological Complex (BMAC) do not yet attest to any such southward migration (*47*). Steppe ancestry is not found until c. 3500 BP, in the Gandhara Grave Culture in northern Pakistan, and only at limited proportions (*47*). The interpretation that this ancestry can be identified with the first Indo-Iranic dispersal into South Asia (*47*) is incompatible with our earlier date for the separation of Indo-Iranic from the rest of Indo-European (c. 6980 BP). We also find that Indic and Iranic had diverged from each other already by c. 5520 BP (4540-6800 BP). To reconcile this with a Steppe origin would require an alternative scenario in which Indic and Iranic split from each other c. two millennia before entering Western and South Asia.

Our analysis indicates that the Indo-European family began with a series of major branching events in relatively quick succession (Fig. 3). From c. 8120 BP (6740-9610 BP) to 6140 BP (4540-7880 BP), Indo-European had split into seven branches (see Table 1, Fig. S5), 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 c. 5300-4900 BP, i.e. early enough to match our estimated split times. Ancient DNA does, however, indicate a spread of CHG/Iranian ancestry in the opposite direction, i.e. from south of the Caucasus into the Steppe c. 7000-6200 BP (*40*), which created the diagnostic 'steppe' mix of ancestries that would later also enter Europe, c. 5000-4500 BP. This CHG/Iranian component is found first south of the Caucasus, including in the northern/eastern arc of the Fertile Crescent, among early farmers on the flanks of the Zagros mountains in western Iran (*42*). The same CHG/Iranian (*40*) ancestry component also admixes heavily (by c. 5000 BP) (*22*) into the region where languages of the Anatolian branch are first documented. It is the dominant ancestry in ancient Armenia and Iran, 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 connects the putative speakers of the European branches and those of the Indo-European languages south of the Caucasus. Our earlier date estimates for the separation of Indo-Iranic from other Indo-European languages (*47, 48*) support 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 new 'hybrid' hypothesis (Fig. 1d), in which Indo-European languages spread out of an initial homeland south of the Caucasus, in the northern Fertile Crescent. Only one major branch spread northwards onto the Steppe, and then across much of Europe. This proposal matches parts of an alternative 'South Caucasus' hypothesis (*49–51*), but the tree topology differs. The first migration phases are significantly 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 also formerly spoken in this region, but too few records survive to include them in our dataset.) Our hybrid hypothesis holds that out of this homeland south of the Caucasus, from c. 8120 BP Proto-Indo-European began to diverge as early migrations split it into multiple early branches. One of these took Indo-Iranic eastwards far earlier than the Steppe hypothesis presumes, but in line with the linguistic chronology in Fig. 3. Indo-Iranic emerged as a distinct branch already within the first phase of Indo-European divergence. Another main branch reached the Steppe directly northwards through the Caucasus c. 7000-6500 BP, compatible with one current interpretation of the aDNA record (*40*). The Steppe then became a secondary homeland for the later Corded Ware-related expansions into Europe.

In sum, aDNA 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 some major role, but they also confirm that at least the Anatolian branch did not originate there and point to an ultimate homeland for the Indo-

European family south of the Caucasus instead. This effectively refocuses the Indo-European question: did all branches other than Anatolian come from the Steppe, or only some? For some branches, the potential candidate expansion(s) detected in aDNA had only limited genetic impact. The key contribution from Bayesian language phylogenetics is now to reveal that those past population expansions also came too late, with respect to the language chronology that we report here. 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.

References

1. H. Hammarström, R. Forkel, M. Haspelmath, Glottolog 4.1 (2019), (available at http://glottolog.org).

2. D. M. Eberhard, G. F. Simons, C. D. Fennig, Eds., *Ethnologue: Languages of the World* (SIL International, Dallas, ed. 22, 2019; https://www.ethnologue.com/statistics/family).

3. J. M. Diamond, P. Bellwood, Farmers and their languages: the first expansions. *Science*. **300**, 597–603 (2003).

4. L. Campbell, W. J. Poser, *Language Classification: History and Method* (Cambridge University Press, Cambridge, 2008).

5. D. W. Anthony, D. Ringe, The Indo-European homeland from linguistic and archaeological perspectives. *Annu. Rev. Appl. Linguist.* **1**, 199–219 (2015).

6. C. Renfrew, *Archaeology and Language: The Puzzle of Indo-European Origins* (Jonathan Cape, London, 1987).

7. J. P. Mallory, *In Search of the Indo-Europeans* (Thames & Hudson, London, 1989).

8. J. Clackson, in *Perspectives on the Origin of Indian Civilization*, A. Marcantonio, G. N. Jha, Eds. (D.K. Printworld, Dartmouth, 2013; https://www.academia.edu/9452122), pp. 259–287.

9. P. Heggarty, in *Routledge Handbook of Historical Linguistics*, C. Bowern, B. Evans, Eds. (Routledge, London, 2015; https://www.academia.edu/3687718), 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).

11. R. Bouckaert, P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, Q. D. Atkinson, Mapping the origins and expansion of the Indo-European language family. *Science*. **337**, 957–960 (2012).

12. W. Chang, C. Cathcart, D. Hall, A. Garrett, Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* . **91**, 194–244 (2015).

13. T. Rama, Three tree priors and five datasets: A study of Indo-European phylogenetics. *Language Dynamics and Change*. **8**, 182–218 (2018).

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. *Journal of Language Evolution*. **4**, 108–123 (2019).

15. J. Clackson, G. Horrocks, *The Blackwell History of the Latin Language* (John Wiley & Sons, Oxford, 2011).

16. W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, D. Reich, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*. **522**, 207–211 (2015).

17. I. Olalde, S. Brace, M. E. Allentoft, I. Armit, K. Kristiansen, T. Booth, N. Rohland, S. Mallick, A. Szécsényi-Nagy, A. Mittnik, E. Altena, M. Lipson, I. Lazaridis, T. K. Harper, N. Patterson, N. Broomandkhoshbacht, Y. Diekmann, Z. Faltyskova, D. Fernandes, M. Ferry, E. Harney, P. de Knijff, M. Michel, J. Oppenheimer, K. Stewardson, A. Barclay, K. W. Alt, C. Liesau, P. Ríos, C. Blasco, J. V. Miguel, R. M. García, A. A. Fernández, E. Bánffy, M. Bernabò-Brea, D. Billoin, C. Bonsall, L. Bonsall, T. Allen, L. Büster, S. Carver, L. C. Navarro, O. E. Craig, G. T. Cook, B. Cunliffe, A. Denaire, K. E. Dinwiddy, N. Dodwell, M. Ernée, C. Evans, M. Kuchařík, J. F. Farré, C. Fowler, M. Gazenbeek, R. G. Pena, M. Haber-Uriarte, E. Haduch, G. Hey, N. Jowett, T. Knowles, K. Massy, S. Pfrengle, P. Lefranc, O. Lemercier, A. Lefebvre, C. H. Martínez, V. G. Olmo, A. B. Ramírez, J. L. Maurandi, T. Majó, J. I. McKinley, K. McSweeney, B. G. Mende, A. Modi, G. Kulcsár, V. Kiss, A. Czene, R. Patay, A. Endrődi, K. Köhler, T. Hajdu, T. Szeniczey, J. Dani, Z. Bernert, M. Hoole, O. Cheronet, D. Keating, P. Velemínský, M. Dobeš, F. Candilio, F. Brown, R. F. Fernández, A.-M. Herrero-Corral, S. Tusa, E. Carnieri, L. Lentini, A. Valenti, A. Zanini, C. Waddington, G. Delibes, E. Guerra-Doce, B. Neil, M. Brittain, M. Luke, R. Mortimer, J. Desideri, M. Besse, G. Brücken, M. Furmanek, A. Hałuszko, M. Mackiewicz, A. Rapiński, S. Leach, I. Soriano, K. T. Lillios, J. L. Cardoso, M. P. Pearson, P. Włodarczak, T. D. Price, P. Prieto, P.-J. Rey, R. Risch, M. A. R. Guerra, A. Schmitt, J. Serralongue, A. M. Silva, V. Smrčka, L. Vergnaud, J. Zilhão, D. Caramelli, T. Higham, M. G. Thomas, D. J. Kennett, H. Fokkens, V. Heyd, A. Sheridan, K.-G. Sjögren, P. W. Stockhammer, J. Krause, R. Pinhasi, W. Haak, I. Barnes, C. Lalueza-Fox, D. Reich, The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*. **555**, 190 (2018).

18. I. Olalde, S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, M. Silva, K. Dulias, C. J. Edwards, F. Gandini, M. Pala, P. Soares, M. Ferrando-Bernal, N. Adamski, N. Broomandkhoshbacht, O. Cheronet, B. J. Culleton, D. Fernandes, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, Z. Zhang, J. M. J. Arenas, I. J. T. Moyano, D. C. Salazar-García, P. Castanyer, M. Santos, J. Tremoleda, M. Lozano, P. G. Borja, J. Fernández-Eraso, J. A. Mujika-Alustiza, C. Barroso, F. J. Bermúdez, E. V. Mínguez, J. Burch, N. Coromina, D. Vivó, A. Cebrià, J. M. Fullola, O. García-Puchol, J. I. Morales, F. X. Oms, T. Majó, J. M. Vergès, A. Díaz-Carvajal, I. Ollich-Castanyer, F. J. López-Cachero, A. M. Silva, C. Alonso-Fernández, G. D. de Castro, J. J. Echevarría, A. Moreno-Márquez, G. P. Berlanga, P. Ramos-García, J. Ramos-Muñoz, E. V. Vila, G. A. Arzo, Á. E. Arroyo, K. T. Lillios, J. Mack, J. Velasco-Vázquez, A. Waterman, L. B. de L. Enrich, M. B. Sánchez, B. Agustí, F. Codina, G. de Prado, A. Estalrrich, Á. F. Flores, C. Finlayson, G. Finlayson, S. Finlayson, F. Giles-Guzmán, A. Rosas, V. B. González, G. G. Atiénzar, M. S. H. Pérez, A. Llanos, Y. C. Marco, I. C. Beneyto, D. López-Serrano, M. S. Tormo, A. C. Valera, C. Blasco, C. Liesau, P. Ríos, J. Daura, M. J. de P. Michó, A. A. Diez-Castillo, R. F. Fernández, J. F. Farré, R. Garrido-Pena, V. S. Gonçalves, E. Guerra-Doce, A. M. Herrero-Corral, J. Juan-Cabanilles, D. López-Reyes, S. B. McClure, M. M. Pérez, A. O. Foix, M. S. Borràs, A. C. Sousa, J. M. V. Encinas, D. J. Kennett, M. B. Richards, K. W. Alt, W. Haak, R. Pinhasi, C. Lalueza-Fox, D. Reich, The genomic history of the Iberian Peninsula over the past 8000 years. *Science*. **363**, 1230–1234 (2019).

19. I. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, N. Rohland, S. Pfrengle, A. Furtwängler, A. Peltzer, C. Posth, A. Vasilakis, P. J. P. McGeorge, E. Konsolaki-Yannopoulou, G. Korres, H. Martlew, M. Michalodimitrakis, M. Özsait, N. Özsait, A. Papathanasiou, M. Richards, S. A. Roodenberg, Y. Tzedakis, R. Arnott, D. M. Fernandes, J. R. Hughey, D. M. Lotakis, P. A. Navas, Y. Maniatis, J. A. Stamatoyannopoulos, K. Stewardson, P. Stockhammer, R. Pinhasi, D. Reich, J. Krause, G. Stamatoyannopoulos, Genetic origins of the Minoans and Mycenaeans. *Nature*. **548**, 214–218 (2017).

20. I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, D. Fernandes, M. Ferry, B. Gamarra, G. G. Fortes, W. Haak, E. Harney, E. Jones, D. Keating, B. Krause-Kyora, I. Kucukkalipci, M. Michel, A. Mittnik, K. Nägele, M. Novak, J. Oppenheimer, N. Patterson, S. Pfrengle, K. Sirak, K. Stewardson, S. Vai, S. Alexandrov, K. W. Alt, R. Andreescu, D. Antonović, A. Ash, N. Atanassova, K. Bacvarov, M. B. Gusztáv, H. Bocherens, M. Bolus, A. Boroneanţ, Y. Boyadzhiev, A. Budnik, J. Burmaz, S. Chohadzhiev, N. J. Conard, R. Cottiaux, M. Čuka, C. Cupillard, D. G. Drucker, N. Elenski, M. Francken, B. Galabova, G. Ganetsovski, B. Gély, T. Hajdu, V. Handzhyiska, K. Harvati, T. Higham, S. Iliev, I. Janković, I. Karavanić, D. J. Kennett, D. Komšo, A. Kozak, D. Labuda, M. Lari, C. Lazar, M. Leppek, K. Leshtakov, D. L. Vetro, D. Los, I. Lozanov, M. Malina, F. Martini, K. McSweeney, H. Meller, M. Menđušić, P. Mirea, V. Moiseyev, V. Petrova, T. D. Price, A. Simalcsik, L. Sineo, M. Šlaus, V. Slavchev, P. Stanev, A. Starović, T. Szeniczey, S. Talamo, M. Teschler-Nicola, C. Thevenet, I. Valchev, F. Valentin, S. Vasilyev, F. Veljanovska, S. Venelinova, E.

Veselovskaya, B. Viola, C. Virag, J. Zaninović, S. Zäuner, P. W. Stockhammer, G. Catalano, R. Krauß, D. Caramelli, G. Zariņa, B. Gaydarska, M. Lillie, A. G. Nikitin, I. Potekhina, A. Papathanasiou, D. Borić, C. Bonsall, J. Krause, R. Pinhasi, D. Reich, The genomic history of southeastern Europe. *Nature*. **555**, 197 (2018).

21. P. de B. Damgaard, N. Marchi, S. Rasmussen, M. Peyrot, G. Renaud, T. Korneliussen, J. V. Moreno-Mayar, M. W. Pedersen, A. Goldberg, E. Usmanova, N. Baimukhanov, V. Loman, L. Hedeager, A. G. Pedersen, K. Nielsen, G. Afanasiev, K. Akmatov, A. Aldashev, A. Alpaslan, G. Baimbetov, V. I. Bazaliiskii, A. Beisenov, B. Boldbaatar, B. Boldgiv, C. Dorzhu, S. Ellingvag, D. Erdenebaatar, R. Dajani, E. Dmitriev, V. Evdokimov, K. M. Frei, A. Gromov, A. Goryachev, H. Hakonarson, T. Hegay, Z. Khachatryan, R. Khaskhanov, E. Kitov, A. Kolbina, T. Kubatbek, A. Kukushkin, I. Kukushkin, N. Lau, A. Margaryan, I. Merkyte, I. V. Mertz, V. K. Mertz, E. Mijiddorj, V. Moiyesev, G. Mukhtarova, B. Nurmukhanbetov, Z. Orozbekova, I. Panyushkina, K. Pieta, V. Smrčka, I. Shevnina, A. Logvin, K.-G. Sjögren, T. Štolcová, K. Tashbaeva, A. Tkachev, T. Tulegenov, D. Voyakin, L. Yepiskoposyan, S. Undrakhbold, V. Varfolomeev, A. Weber, N. Kradin, M. E. Allentoft, L. Orlando, R. Nielsen, M. Sikora, E. Heyer, K. Kristiansen, E. Willerslev, 137 ancient human genomes from across the Eurasian steppes. *Nature*, 1 (2018).

22. E. Skourtanioti, Y. S. Erdal, M. Frangipane, F. Balossi Restelli, K. A. Yener, F. Pinnock, P. Matthiae, R. Özbal, U.-D. Schoop, F. Guliyev, T. Akhundov, B. Lyonnet, E. L. Hammer, S. E. Nugent, M. Burri, G. U. Neumann, S. Penske, T. Ingman, M. Akar, R. Shafiq, G. Palumbi, S. Eisenmann, M. D'Andrea, A. B. Rohrlach, C. Warinner, C. Jeong, P. W. Stockhammer, W. Haak, J. Krause, Genomic History of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus. *Cell*. **181**, 1158–1175.e28 (2020).

23. M. Price, Finding the first horse tamers. *Science*. **360**, 587–587 (2018).

24. 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).

25. E. Finegan, in *The World's Major Languages*, B. Comrie, Ed. (Routledge, London, ed. 2, 2009), pp. 59–85.

26. R. Lazzeroni, in *The Indo-European Languages*, A. G. Ramat, P. Ramat, Eds. (Routledge, London, 1998), pp. 98–124.

27. C. P. Masica, *The Indo-Aryan Languages* (Cambridge University Press, Cambridge, 1991).

28. J. N. Adams, *The Regional Diversification of Latin 200 BC - AD 600* (Cambridge University Press, Cambridge, 2014; https://doi.org/10.1017/CBO9780511482977).

29. J. Clackson, in *The Oxford Guide to the Romance Languages*, A. Ledgeway, M. Maiden, Eds. (Oxford University Press, Oxford,

2016), pp. 3–13.

30. P. Mackridge, in *A Companion to the Ancient Greek Language*, E. J. Bakker, Ed. (John Wiley & Sons, Chichester, 2010), pp. 564–587.

31. G. Horrocks, *Greek: A History of the Language and its Speakers* (John Wiley & Sons, Chichester, 2009).

32. N. Sims-Williams, in *The Indo-European Languages*, A. G. Ramat, P. Ramat, Eds. (Routledge, London, 1998), pp. 125–153.

33. IE-CoR Database, IE-CoR cognacy overlap listing for Vedic: Early vs. Avestan: Younger (2020), (available at https://iecor.clld.org/values?sEcho=3&sSearch_2=Avestan:+Younger,Vedic:+Early).

34. A. Garrett, in *Phylogenetic Methods and the Prehistory of Languages*, P. Forster, C. Renfrew, Eds. (McDonald Institute for Archaeological Research, Cambridge, 2006), pp. 139–151.

35. 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).

36. S. J. Greenhill, P. Heggarty, R. D. Gray, in *The Handbook of Historical Linguistics*, R. D. Janda, B. D. Joseph, B. S. Vance, Eds. (Wiley-Blackwell, Hoboken, New Jersey, ed. 2, 2020), vol. 2.

37. P. Heggarty, Cognacy databases and phylogenetic research on Indo-European. *Annu. Rev. Appl. Linguist.* **7**, 371–394 (2021).

38. L. Papac, M. Ernée, M. Dobeš, M. Langová, A. B. Rohrlach, F. Aron, G. U. Neumann, M. A. Spyrou, N. Rohland, P. Velemínský, M. Kuna, H. Brzobohatá, B. Culleton, D. Daněček, A. Danielisová, M. Dobisíková, J. Hložek, D. J. Kennett, J. Klementová, M. Kostka, P. Krištuf, M. Kuchařík, J. K. Hlavová, P. Limburský, D. Malyková, L. Mattiello, M. Pecinovská, K. Petriščáková, E. Průchová, P. Stránská, L. Smejtek, J. Špaček, R. Šumberová, O. Švejcar, M. Trefný, M. Vávra, J. Kolář, V. Heyd, J. Krause, R. Pinhasi, D. Reich, S. Schiffels, W. Haak, Dynamic changes in genomic and social structures in third millennium BCE central Europe. *Sci Adv*. **7** (2021), doi:10.1126/sciadv.abi6941.

39. P. Heggarty, 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. (Journal of Indo-European Studies, 2018; http://jies.org/DOCS/monojpgs/Mon65.html), vol. 65 of *Journal of Indo-European Studies Monograph Series*, pp. 120–173.

40. C.-C. Wang, S. Reinhold, A. Kalmykov, A. Wissgott, G. Brandt, C. Jeong, O. Cheronet, M. Ferry, E. Harney, D. Keating, S.

Mallick, N. Rohland, K. Stewardson, A. R. Kantorovich, V. E. Maslov, V. G. Petrenko, V. R. Erlikh, B. C. Atabiev, R. G. Magomedov, P. L. Kohl, K. W. Alt, S. L. Pichler, C. Gerling, H. Meller, B. Vardanyan, L. Yeganyan, A. D. Rezepkin, D. Mariaschk, N. Berezina, J. Gresky, K. Fuchs, C. Knipper, S. Schiffels, E. Balanovska, O. Balanovsky, I. Mathieson, T. Higham, Y. B. Berezin, A. Buzhilova, V. Trifonov, R. Pinhasi, A. B. Belinskij, D. Reich, S. Hansen, J. Krause, W. Haak, Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions. *Nat. Commun.* **10**, 590 (2019).

41. E. R. Jones, G. Gonzalez-Fortes, S. Connell, V. Siska, A. Eriksson, R. Martiniano, R. L. McLaughlin, M. Gallego Llorente, L. M. Cassidy, C. Gamba, T. Meshveliani, O. Bar-Yosef, W. Müller, A. Belfer-Cohen, Z. Matskevich, N. Jakeli, T. F. G. Higham, M. Currat, D. Lordkipanidze, M. Hofreiter, A. Manica, R. Pinhasi, D. G. Bradley, Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* **6**, 8912 (2015).

42. F. Broushaki, M. G. Thomas, V. Link, S. López, L. van Dorp, K. Kirsanow, Z. Hofmanová, Y. Diekmann, L. M. Cassidy, D. Díez-del-Molino, A. Kousathanas, C. Sell, H. K. Robson, R. Martiniano, J. Blöcher, A. Scheu, S. Kreutzer, R. Bollongino, D. Bobo, H. Davoudi, O. Munoz, M. Currat, K. Abdi, F. Biglari, O. E. Craig, D. G. Bradley, S. Shennan, K. R. Veeramah, M. Mashkour, D. Wegmann, G. Hellenthal, J. Burger, Early Neolithic genomes from the eastern Fertile Crescent. *Science*. **353**, 499–503 (2016).

43. I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, S. Connell, K. Stewardson, E. Harney, Q. Fu, G. Gonzalez-Fortes, E. R. Jones, S. A. Roodenberg, G. Lengyel, F. Bocquentin, B. Gasparian, J. M. Monge, M. Gregg, V. Eshed, A.-S. Mizrahi, C. Meiklejohn, F. Gerritsen, L. Bejenaru, M. Blüher, A. Campbell, G. Cavalleri, D. Comas, P. Froguel, E. Gilbert, S. M. Kerr, P. Kovacs, J. Krause, D. McGettigan, M. Merrigan, D. A. Merriwether, S. O'Reilly, M. B. Richards, O. Semino, M. Shamoon-Pour, G. Stefanescu, M. Stumvoll, A. Tönjes, A. Torroni, J. F. Wilson, L. Yengo, N. A. Hovhannisyan, N. Patterson, R. Pinhasi, D. Reich, Genomic insights into the origin of farming in the ancient Near East. *Nature*. **536**, 419–424 (2016).

44. F. Clemente, M. Unterländer, O. Dolgova, C. E. G. Amorim, F. Coroado-Santos, S. Neuenschwander, E. Ganiatsou, D. I. Cruz Dávalos, L. Anchieri, F. Michaud, L. Winkelbach, J. Blöcher, Y. O. Arizmendi Cárdenas, B. Sousa da Mota, E. Kalliga, A. Souleles, I. Kontopoulos, G. Karamitrou-Mentessidi, O. Philaniotou, A. Sampson, D. Theodorou, M. Tsipopoulou, I. Akamatis, P. Halstead, K. Kotsakis, D. Urem-Kotsou, D. Panagiotopoulos, C. Ziota, S. Triantaphyllou, O. Delaneau, J. D. Jensen, J. V. Moreno-Mayar, J. Burger, V. C. Sousa, O. Lao, A.-S. Malaspinas, C. Papageorgopoulou, The genomic history of the Aegean palatial civilizations. *Cell*. **184**, 2565–2586.e21 (2021).

45. D. A. Ringe, T. Warnow, A. Taylor, Indo-European and computational cladistics. *Trans. Philol. Soc.* **100**, 59–129 (2002).

46. 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).

47. V. M. Narasimhan, N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, S. Mallick, I. Lazaridis, N. Nakatsuka, I. Olalde, M. Lipson, A. M. Kim, L. M. Olivieri, A. Coppa, M. Vidale, J. Mallory, V. Moiseyev, E. Kitov, J. Monge, N. Adamski, N. Alex, N. Broomandkhoshbacht, F. Candilio, K. Callan, O. Cheronet, B. J. Culleton, M. Ferry, D. Fernandes, S. Freilich, B. Gamarra, D. Gaudio, M. Hajdinjak, É. Harney, T. K. Harper, D. Keating, A. M. Lawson, M. Mah, K. Mandl, M. Michel, M. Novak, J. Oppenheimer, N. Rai, K. Sirak, V. Slon, K. Stewardson, F. Zalzala, Z. Zhang, G. Akhatov, A. N. Bagashev, A. Bagnera, B. Baitanayev, J. Bendezu-Sarmiento, A. A. Bissembaev, G. L. Bonora, T. T. Chargynov, T. Chikisheva, P. K. Dashkovskiy, A. Derevianko, M. Dobeš, K. Douka, N. Dubova, M. N. Duisengali, D. Enshin, A. Epimakhov, A. V. Fribus, D. Fuller, A. Goryachev, A. Gromov, S. P. Grushin, B. Hanks, M. Judd, E. Kazizov, A. Khokhlov, A. P. Krygin, E. Kupriyanova, P. Kuznetsov, D. Luiselli, F. Maksudov, A. M. Mamedov, T. B. Mamirov, C. Meiklejohn, D. C. Merrett, R. Micheli, O. Mochalov, S. Mustafokulov, A. Nayak, D. Pettener, R. Potts, D. Razhev, M. Rykun, S. Sarno, T. M. Savenkova, K. Sikhymbaeva, S. M. Slepchenko, O. A. Soltobaev, N. Stepanova, S. Svyatko, K. Tabaldiev, M. Teschler-Nicola, A. A. Tishkin, V. V. Tkachev, S. Vasilyev, P. Velemínský, D. Voyakin, A. Yermolayeva, M. Zahir, V. S. Zubkov, A. Zubova, V. S. Shinde, C. Lalueza-Fox, M. Meyer, D. Anthony, N. Boivin, K. Thangaraj, D. J. Kennett, M. Frachetti, R. Pinhasi, D. Reich, The formation of human populations in South and Central Asia. *Science*. **365**, eaat7487 (2019).

48. V. Shinde, V. M. Narasimhan, N. Rohland, S. Mallick, M. Mah, M. Lipson, N. Nakatsuka, N. Adamski, N. Broomandkhoshbacht, M. Ferry, A. M. Lawson, M. Michel, J. Oppenheimer, K. Stewardson, N. Jadhav, Y. J. Kim, M. Chatterjee, A. Munshi, A. Panyam, P. Waghmare, Y. Yadav, H. Patel, A. Kaushik, K. Thangaraj, M. Meyer, N. Patterson, N. Rai, D. Reich, An ancient Harappan genome lacks ancestry from Steppe pastoralists or Iranian farmers. *Cell*. **179**, 1–7 (2019).

49. T. V. Gamkrelidze, V. V. Ivanov, *Indoevropjskij jazyk i indoevropejcy: Rekonstrukcija i istoriko-tipologieskij 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 University Press, Tblisi, 1984).

50. 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* (Mouton de Gruyter, Berlin, 1995).

51. T. V. Gamkrelidze, V. V. Ivanov, Indo-European homeland and migrations: half a century of studies and discussions. *Journal of*

Language Relationship. **9**, 109–136 (2013).

52. R. Bouckaert, J. Heled, DensiTree 2: Seeing Trees Through the Forest. *bioRxiv*, 012401 (2014).

Acknowledgements

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 by Matthew Scarborough, with significant contributions also by Britta Irslinger, Roland Pooth and Cassandra Freiberg. Within specific branches of Indo-European, cognacy determinations were principally by: Lechosław Jocz (Slavic), Matthew Scarborough (Greek and ancient Italic), Martin Joachim Kümmel (mostly Iranic), Thomas Jügel (Iranic), Cormac Anderson (mostly Celtic), Henrik Liljegren (Hindu-Kush Indic), Richard Strand (Nuristani), Roland Pooth (Indic), Geoffrey Haig (Iranic), Robert Tegethoff (Indic), Ulrich Geupel (Albanian), Martin Macak (Armenian), Ronald Kim (Tocharian), Alexander Falileyev (Celtic), Erik Anonby (Iranic), Tijmen Pronk (Baltic), Oleg Belyaev (Ossetic), Tonya Kim Dewey-Findell (Germanic), and Matthew Boutilier (Germanic). Other contributors who made lexeme determinations for multiple languages in a given branch are: Matilde Serangeli (Anatolian), Nikos Liosis (Modern Hellenic), Kim Schulte (Romance), Britta Irslinger (Celtic), Nicholas Williams (Cornish), Martin Findell (Germanic), Simone Loi (Sardinian), Patrycja Markus (Indic), Ganesh Gupta (Indic), Roland Pooth (Indic), Nicholas Sims-Williams (Iranic), Raheleh Izadifar (Iranic) and Shirin Adibifar (Iranic). In some cases, a language expert made lexeme determinations for a single language (listed by alphabetical order of surname): Giovanni Abete, Petar Atanasov, Esther Baiwir, Maria-Reina Bastardas, Adam Benkato, Lisa Bevevino, Giorgio Cadorini, Loïc Cheveau, Charalambos Christodoulou, Michiel de Vaan, Jérémie Delorme, Stephen Dworkin, Cassandra Freiberg, Mojtaba Gheitasi, Harald Hammarström, Steve Hewitt, Afsar Ali Khan, Muhammad Kamal Khan, Liudmila Khokhlova, Deborah Kim, Christopher Lewin, Borana Lushaj, Parvin Mahmoudveysi, Masoud Mahommadirad, Sam Mersch, John Mock, Baydaa Moustafa, Fatemeh Nemati, Maryam Nourzaei, Peadar Ó Muircheartaigh, Muhammed Ourang, Heather Pagan, Timothy Palmer, Khwaja Rehman, Guto Rhys, Muhammad Zaman

Sagar, Lars Steensland, Mortaza Taheri-Ardali, Mahnaz Talebi, Sabine Tittel, Annemarie Verkerk, Arjen Versloot, Paul Videsott, Nikola Vuletić, Manuel Widmer and Arash Zeini.

The basic relational database structure for IE-CoR was inherited from the LexDB system and the IELex website developed by Michael Dunn. The IE-CoR data set was produced using the new CoR database creation system programmed by Jakob Runge and Hans-Jörg 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 mostly by Hans-Jörg Bibiko, within the CLLD framework developed by Robert Forkel.

We thank Alexandra Gavryushkina for advice on sampled ancestor prior probabilities and ancestry constraints. We thank Michelle O'Reilly for the preparation of the figures. Finally, we thank Andrew Garrett and Will Chang at the Dept of Linguistics, University of California, Berkeley, for extensive comments and discussion of this research.

Funding

This research was funded by the Max Planck Society, through the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany). From 11.09.2021, PH was funded by the ERC Starting Grant 'Waves' (ERC758967).

Author contributions

RG initiated and coordinated the study. PH and CA designed the IE-CoR database and data collection methodology, and coordinated the linguistic coding team. MSc oversaw all determination of cognacy at the deep Indo-European level. CA, MSc, LJ, MJK, TJ, BI, RP, HL, RFS, GH, MM, RIK, EA, TP, OB, MB, CF, RT, MSe, NL, KStr, KSch and GKG were major contributors to the 25,918 lexeme and cognate determinations in the IE-CoR database. RB, BK, SG and DK conducted the phylogenetic analyses, with input from RG, QA, PH and CA. WH and JK advised on the ancient DNA data. PH, RG, DK, BK and CA wrote the text; all authors commented on the manuscript.

Competing interests

All authors declare 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 freely downloaded at https://iecor.clld.org. The .xml data files used as input to each of the phylogenetic analyses are available in the supplementary data and results files online at https://share.eva.mpg.de/index.php/s/Hq5XLC3xga4nBe2 under the password: iecognacy — see the **Guide_to_Suppl_Data_and_Results_Files.pdf**.

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 Denise Kühnert, available at: https://github.com/CompEvol/sampled-ancestors/blob/master/src/beast/evolution/tree/AncestryConstraint.java

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 **Guide_to_Suppl_Data_and_Results_Files.pdf**.

Supplementary Materials:

Suppl_Materials.pdf, including: language database methodology (SM §1, §2), phylogenetic analysis methods (SM §1.2), time calibrations (SM §3), main analysis prior set-up and alternative models (SM §4), results from main analysis (SM §5), results from sensitivity analyses (SM §6), comparison of phylogenetic results with established language classifications (SM §7), and contextualization against other linguistic data (SM §8), archaeology and ancient DNA (SM §9).

Figures S1, S4.1, S4.2, S5.1, S6.8 (within Suppl_Materials.pdf)

Tables S3, S4.4, S5.2, S6, S6.1 (within Suppl_Materials.pdf)

Guide_to_Suppl_Data_and_Results_Files.pdf, including links to the IE-CoR database and to the full supplementary data and results files online.

IECoR_Suppl_Files_Light_Main_M3_Only.zip, including the core data and results files for our main analysis.

Figures & Tables

Figure 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 non-modern 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). Maps (b)-(d) show 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). Map (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.

Figure 2. A DensiTree(*52*) showing the posterior probability distribution of trees for the Indo-European family. The time axis shows the estimated chronology of Indo-European expansion. Languages whose tips do not reach the right edge are the 52 non-modern written languages such as Hittite, Tocharian, Mycenaean Greek and Old English. These languages were used in the analysis as time calibrations. Median age estimates and 95% HPD intervals for major lineages are given to the right of the plot. 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.

Figure 3. Histogram of direct ancestry relationships between languages. The IE-CoR database includes 52 non-modern languages (e.g. Ancient Greek, Classical Latin, 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 4 languages show non-negligible posterior probabilities of being directly ancestral: 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 supplementary Table S5.2.

Figure 4. Posterior probability distributions of the estimated age of Indo-European across all sensitivity analyses, compared to the main analysis (0). 1. With tip calibrations for Early Vedic and Younger Avestan removed. 2. With parallel loans not excluded, but coded as unique cognate sets. 3. With the prior conditioned on the MRCA, not the origin. 4a. With a sampling probability assuming 200-400 modern languages. 4b. With a sampling probability assuming 600-800 modern languages. 5. With ten poorly attested languages removed. 6a. With targeted lower-order clade constraints. 6b. With higher-order clade constraints following the Ringe topology (*5*). 7a. With an ancestry constraint for Latin only. 7b. With ancestry constraints for the 5 languages with > 0.001 posterior probability of being ancestral. 7c. With all 27 remotely possible ancestry constraints. 8a. Using the 'broad' subset (*12*) of the IELex database with ancestors enabled but not enforced. 8b. Using the 'broad' subset (*12*) of the IELex database with ancestry enforced.

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)	
	all date estimates in years 'BP' as *before 2000 AD*			
	median	95% HPD	median	95% HPD
(Proto-)Indo-European	——	——	**8116 BP**	**6735–9613 BP**
[Balto-Slavic] + [Italic + Germanic + Celtic]	6981 BP	5645–8395 BP	6465 BP	5036–7944 BP
[Italic + Germanic + Celtic]	6465 BP	5036–7944 BP	5564 BP	4231–6984 BP
Indo-Iranic	6981 BP	5645–8395 BP	5520 BP	4535–6796 BP
Greco-Armenian	6135 BP	4540–7882 BP	5310 BP	3999–6930 BP
Balto-Slavic	6465 BP	5036–7944 BP	3663 BP	2531–5034 BP
Anatolian	6932 BP	5403–8613 BP	4618 BP	3857–5620 BP
Indic	5520 BP	4535–6796 BP	4366 BP	3640–5253 BP
Iranic	5520 BP	4535–6796 BP	4110 BP	3464–4894 BP
Italic	5564 BP	4231–6984 BP	3431 BP	2771–4286 BP
Greek	5310 BP	3999–6930 BP	3364 BP	3218–3609 BP
Celtic	4889 BP	3718–6193 BP	3205 BP	2515–3963 BP
Baltic	3663 BP	2531–5034 BP	2439 BP	1526–3484 BP
Germanic	4889 BP	3718–6193 BP	2337 BP	1931–2865 BP
Tocharian	6932 BP	5403–8613 BP	1828 BP	1495–2315 BP
Armenian	5310 BP	3999–6930 BP	1578 BP	1485–1851 BP
Slavic	3663 BP	2531–5034 BP	1493 BP	1222–1837 BP
Albanian	6135 BP	4540–7882 BP	1067 BP	468–1882 BP

Table 1: **Estimated time-depths of the twelve main well-attested clades of Indo-European, and higher order clades with high posterior probability support.** Dates grayed out are merely indicative, based on splits with less than 50% posterior support. Date estimates shown are the height_median and height_95%_HPD values in the MCC tree file; see also Figure S5.0.