PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

*Phil. Trans. R. Soc. B* (2011) **366**, 1090–1100
doi:10.1098/rstb.2010.0378

*Review*

# Language evolution and human history: what a difference a date makes

**Russell D. Gray*, Quentin D. Atkinson and Simon J. Greenhill**

*Department of Psychology, University of Auckland, Auckland 1142, New Zealand*

Historical inference is at its most powerful when independent lines of evidence can be integrated into a coherent account. Dating linguistic and cultural lineages can potentially play a vital role in the integration of evidence from linguistics, anthropology, archaeology and genetics. Unfortunately, although the comparative method in historical linguistics can provide a relative chronology, it cannot provide absolute date estimates and an alternative approach, called glottochronology, is fundamentally flawed. In this paper we outline how computational phylogenetic methods can reliably estimate language divergence dates and thus help resolve long-standing debates about human prehistory ranging from the origin of the Indo-European language family to the peopling of the Pacific.

**Keywords:** linguistics; glottochronology; Indo-European; Austronesian; cultural evolution

## 1. INTRODUCTION

Historical inference is hard. Trying to work out what happened 600 years ago is difficult enough. Trying to make inferences about events 6000 years ago may seem close to impossible. As W. S. Holt observed, the study of human history is 'a damn dim candle over a damn dark abyss'. And yet evolutionary biologists routinely make inferences about events millions of years in the past. Our ability to do this was revolutionized by Zuckerkandl & Pauling's [1] insight that molecules are 'documents of evolutionary history'. Molecular sequences have inscribed in their structure a record of their past. Similarities generally reflect common ancestry. Today, computational phylogenetic methods are routinely used to make inferences about evolutionary relationships and processes from these sequences. These inferences are more powerful when independent lines of evidence, such as information from studies of morphology, geology and palaeontology, are brought to bear on a common problem.

Languages, like genes, are also 'documents of history'. A vast amount of information about our past is inscribed in the content and structure of the approximately 7000 languages that are spoken today [2]. Historical linguists have developed a careful set of procedures termed the 'comparative method' to infer ancestral states and construct language family trees [3,4]. Ideally, as Kirch & Green [5] and Renfrew [6] have argued, independent evidence from anthropology, archaeology and human genetics are used to 'triangulate' inferences about human prehistory and cultural evolution. From anthropology comes an understanding of social organization, from archaeology comes an absolute chronology of changes in material culture, and from genetic studies we get information about the sequence of population movements and the extent of admixture. Traditionally, historical linguistics has contributed inferences about ancestral vocabulary and a relative cultural chronology to this synthesis.

While this 'new synthesis' [7] is a worthy aim, it is often very difficult to link the different lines of evidence together. Archaeological remains do not speak. Genes and languages can have different histories or appear spuriously congruent. The one thing that is critically important to successfully triangulating the different lines of evidence together is timing. If archaeological, genetic and linguistic lines of evidence show similar absolute dates for a common sequence of events, then our confidence that a common process is involved would be hugely increased, and the 'damn dark abyss' of human history greatly illuminated. Sadly, the absence of appropriate calibration points and systematic violations of the molecular clock mean that there are large sources of error associated with most genetic dates for human population history [8]. Sadder still, although the comparative method in linguistics can provide a relative chronology, it cannot provide absolute date estimates. In the words of April MacMahon & Rob MacMahon [9] 'linguists don't do dates'. We are not so pessimistic. In what follows we will outline why dating linguistic lineages is a difficult, but not impossible, task.

## 2. DATING DIFFICULTIES

A quick glance at an Old English text, such as the epic poem *Beowulf*, should be enough to convince anyone of two facts. Languages evolve and they evolve rapidly. New words arise and others are replaced. Sounds change, grammar morphs and speech communities

*rd.gray@auckland.ac.nz

One contribution of 26 to a Discussion Meeting Issue 'Culture evolves'.

split into dialects and then distinct languages. Given this linguistic divergence over time, one plausible intuition is that it might be possible to use some measure of this divergence to estimate the age of linguistic lineages in much the same way that biologists use the divergence of molecular sequences to date biological lineages. 'Glottochronology' attempts to do just that. In the early 1950s, a full decade before Zuckerkandl & Pauling introduced the idea of a molecular clock to biology, Swadesh [10,11] developed an approach to historical linguistics termed lexicostatistics and its derivative 'glottochronology'. This approach used lexical data to determine language relationships and to estimate absolute divergence times. Lexicostatistical methods infer language trees on the basis of the percentage of shared cognates between languages—the more similar the languages, the more closely they are related. Cognates are words in different languages that have a common ancestor. In biological terminology they are homologous. Words are judged to be cognate if they have a pattern of systematic sound correspondences and similar meanings. Glottochronology is an extension of lexicostatistics that estimates language divergence times under the assumption of a 'glottoclock', or constant rate of language change. The following formulae can be used to relate language similarity to time along an exponential decay curve:

$$t = \frac{\log C}{2 \log r},$$

where $t$ is time depth in millennia, $C$ is the percentage of cognates shared and $r$ is the 'universal' constant or rate of retention (the expected proportion of cognates remaining after 1000 years of separation). Usually analyses are restricted to the Swadesh word list—a collection of 100–200 basic meanings that are thought to be relatively culturally universal, stable and resistant to borrowing. These include kinship terms (e.g. mother, father), terms for body parts (e.g. hand, mouth, hair), numerals and basic verbs (e.g. to drink, to sleep, to burn). For the Swadesh 200-word list, the retention rate ($r$) was estimated from cases where the divergence date between languages was known from historical records. This rate was found to be approximately 81 per cent.

Unfortunately, this apparently simple and elegant solution to the important problem of dating linguistic lineages encountered some major obstacles [12,13], and thus most historical linguists now view glottochronological calculations with considerable scepticism. The most fundamental obstacle encountered by glottochronology is the fact that languages, just like genes, often do not evolve at a constant rate. In their classic critique of glottochronology, Bergsland & Vogt [12] compared present-day languages with their archaic forms. They found considerable evidence of rate variation between languages. For example, Icelandic and Norwegian were compared with their common ancestor, Old Norse, spoken roughly 1000 years ago. Norwegian has retained 81 per cent of the vocabulary of Old Norse, correctly suggesting an age of approximately 1000 years. However, Icelandic has retained over 95 per cent of the Old Norse vocabulary, falsely suggesting that Icelandic split from Old Norse less than 200 years ago. This is not an isolated example. In a survey of Malayo-Polynesian languages, Blust [13] documented variations in the retention of basic vocabulary driven by factors such as language contact and large changes in population size that ranged from 5 to 50 per cent in the approximately 4000 years from Proto Malayo-Polynesian to the present. Blust argued that these huge differences in retention rates inevitably distorted both the trees obtained by lexicostatistics and the glottochronological dates.

It is ironic that over the past half-century, computational methods in historical linguistics have fallen out of favour while in evolutionary biology computational methods have blossomed. Rather than giving up and saying, 'we don't do dates', computational biologists have developed methods that can accurately estimate phylogenetic trees and divergence dates even when there is considerable lineage-specific rate heterogeneity. Evolutionary biologists today typically use likelihood and Bayesian methods to explicitly model all the substitution events, instead of building trees from pairwise distance matrices [14,15]. The development of these more powerful computational methods has been facilitated by both a spectacular increase in the availability of molecular sequences and dramatic increases in computational power in the past 20 years [16]. The use of all the sequence information and more complex and realistic models of the substitution process mean that likelihood and Bayesian methods outperform the simple clustering methods, especially when rates of molecular change are not constant [17]. In addition to developing methods to build more accurate trees, evolutionary biologists have recently developed methods to obtain more accurate date estimates, even when there are departures from the assumption of a strict molecular clock.

One popular approach pioneered by Sanderson [18,19] involves two steps. First, a set of phylogenetic trees and their associated branch lengths are estimated. In Bayesian phylogenetic analyses of molecular evolution, the branch lengths are proportional to the number of substitutions along a branch given the data, the substitution model and the priors. The second step involves converting the relative branch lengths into time. Calibration points are required to do this. These are places where nodes (branching points) on the trees can be constrained to a known date range. These known node ages are then combined with the branch-length information to estimate rates of evolution across each tree. A penalized-likelihood model is used to allow rates to vary across the tree while incorporating a 'roughness penalty'. The more the rates vary from branch-to-branch, the greater the cost (see [18,19] for more detail). The algorithm allows an optimal value of the roughness penalty to be selected. In this way, the combination of calibrations, branch-length estimates and the rate-smoothing algorithm enables dates to be estimated without assuming a strict clock. An alternative 'relaxed phylogenetics' approach, in which the tree and the dates are simultaneously estimated, has recently been developed

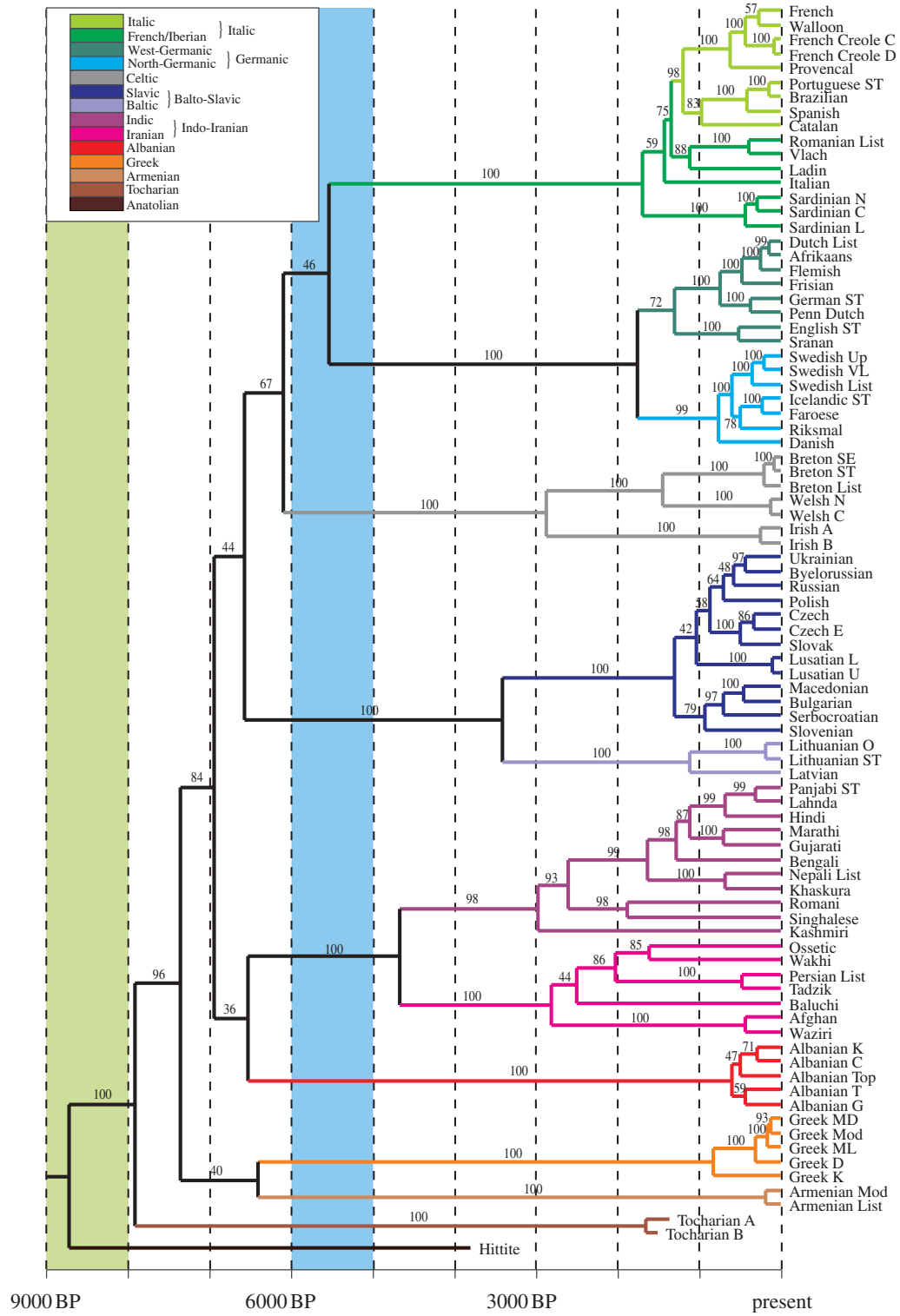


Figure 1. A dated phylogenetic tree of 87 Indo-European languages. The tree is a consensus tree derived from the posterior samples of trees in the Bayesian analyses reported by Gray & Atkinson [32]. The values on the branches are the posterior probability of that clade. The root age of the tree is in the age range predicted by the Anatolian hypothesis. This figure also shows an interesting point that we had noted, but not emphasized, in our initial paper—while the root of the tree goes back around 8700 years, much of the diversification of the major Indo-European subgroups happened around 6000–7000 BP. This means that both the Anatolian and the Kurgan hypotheses could be simultaneously true. There was an initial movement out of Anatolia 8700 years ago and then a major radiation 6000–7000 years ago from southern Russia and the Ukraine. It also means that the intuition shared by many linguists that the Indo-European language family is about 6000 years old could be correct for the vast majority of Indo-European languages, just not the deeper subgroups.

by Drummond *et al.* [20]. In 'relaxed phylogenetics', the assumption of a strict clock can be eased by modelling the rate variation using lognormal or exponential distributions (see [20] for more detail). In the sections that follow, we will explore how these computational phylogenetic methods can be used to illuminate the linguistic and cultural history of people both in Europe and in the Pacific.

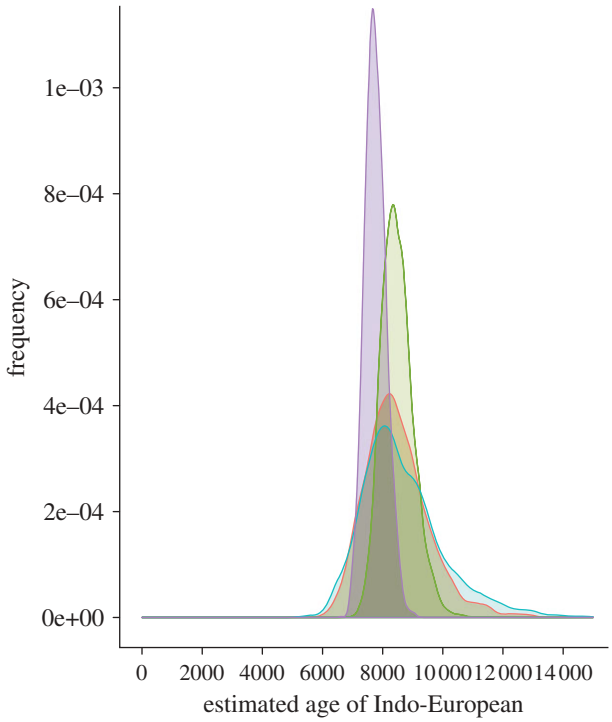


Figure 2. Distributions of the age of Proto Indo-European estimated from the data of Ringe *et al.* [38]. Four different analyses were conducted using the program BEAST. Two analyses assumed equal rates of cognate gain and loss—one with a strict clock (light green) and one with a lognormal relaxed clock (orange). The other two analyses assumed that cognates could only be gained once but lost multiple times (stochastic Dollo). Again one implemented a strict clock (purple) and one used a lognormal relaxed clock (light blue). The date estimates obtained in all four analyses were consistent with the Anatolian hypothesis.

## 3. THE ORIGIN OF THE INDO-EUROPEAN LANGUAGES

The origin of the Indo-European languages has recently been described as 'one of the most intensively studied, yet still most recalcitrant problems of historical linguistics' [21, p. 601]. Despite over 200 years of scrutiny, scholars have been unable to locate the origin of Indo-European definitively in time or place. Theories have been put forward advocating ages ranging from 4000 to 23 000 years, with hypothesized homelands including Central Europe, the Balkans and even India. Mallory [22] acknowledges 14 distinct homeland hypotheses since 1960 alone. He rather colourfully remarks that, 'the quest for the origins of the Indo-Europeans has all the fascination of an electric light in the open air on a summer night: it tends to attract every species of scholar or would-be savant who can take pen to hand' [22, p. 143].

Of all the diverse theories about the origin of Indo-Europeans there are currently two that receive the most attention. The first, put forward by Gimbutas [23,24] on the basis of linguistic and archaeological evidence, links Proto-Indo-European (the hypothesized ancestral Indo-European tongue) with the Kurgan culture of southern Russia and the Ukraine. The Kurgans were a group of semi-nomadic, pastoralist, warrior-horsemen who expanded from their homeland in the Pontic steppes during the fifth and sixth millennia BP, conquering Danubian Europe, Central Asia and India, and later the Balkans and Anatolia. This expansion is thought to roughly match the accepted ancestral range of Indo-European [25]. As well as the apparent geographical congruence between Kurgan and Indo-European territories, there is linguistic evidence for an association between the two cultures. Words for supposed Kurgan technological innovations are consistent across widely divergent Indo-European sub-families. These include terms for 'wheel' (*rotho-, *$k^{W}$(e)$k^{W}$l-o-), 'axle' (*aks-lo-), 'yoke' (*jug-o-), 'horse' (*ekwo-) and 'to go, transport in a vehicle' (*wegh- [14,15]): it is argued that these words and associated technologies must have been present in the Proto-Indo-European culture and that they were likely to have been Kurgan in origin. Hence, the argument goes, the Indo-European language family is no older than 5000–6000 BP. Mallory [22] argues for a similar time and place of Indo-European origin—a region around the Black Sea about 5000–6000 BP (although he and many linguists are more cautious and refrain from identifying Proto-Indo-European with a specific culture such as the Kurgans).

The second theory, proposed by the archaeologist Renfrew [26], holds that Indo-European languages spread, not with marauding horsemen, but with the expansion of agriculture from Anatolia between 8000 and 9500 years ago. Radiocarbon analysis of the earliest Neolithic sites across Europe provides a fairly detailed chronology of agricultural dispersal. This archaeological evidence indicates that agriculture spread from Anatolia, arriving in Greece at some time during the ninth millennium BP and reaching as far as the British Isles by 5500 BP [27]. Renfrew maintains that the linguistic argument for the Kurgan theory is based only on limited evidence for a few enigmatic Proto-Indo-European word forms. He points out that parallel semantic shifts or widespread borrowing can produce similar word forms across different languages without requiring that an ancestral term was present in the proto-language. Renfrew also challenges the idea that Kurgan social structure and technology was sufficiently advanced to allow them to conquer whole continents in a time when even small cities did not exist. Far more credible, he argues, is that Proto-Indo-European spread with the spread of agriculture.

The debate about Indo-European origins thus centres on archaeological evidence for two population expansions, both implying very different timescales—the Kurgan theory with a date of 5000–6000 BP, and the Anatolian theory with a date of 8000–9500 BP. One way of potentially resolving the debate is to look outside the archaeological record for independent evidence, which allows us to test between these two time depths. Does linguistics hold the key? Well, not if linguists do not do dates. However, if we could reliably date the origin of the Indo-European languages, then it would make a huge difference to this 200 year old debate.

We set about this rather daunting task by building on what Darwin dubbed the 'curious parallels' between biological and linguistic evolution (see [28] for an analysis of the history of these parallels). If

languages, like biological species, are also 'documents of history', then perhaps they could be analysed using the same computational evolutionary methods. Maybe the solutions biologists have found to violations of the molecular clock could be used to overcome problems with glottochronology. It requires a large amount of data to estimate tree topology and branch lengths accurately. Our data were taken from the Dyen *et al.* [29] Indo-European lexical database, which contains expert cognacy judgements for 200 Swadesh list terms in 95 languages. Dyen *et al.* [29] identified 11 languages as less reliable and hence they were not included in the analysis presented here. We added three extinct languages (Hittite, Tocharian A and Tocharian B) to the database in an attempt to improve the resolution of basal relationships in the inferred phylogeny. For each meaning in the database, languages were grouped into cognate sets. By restricting analyses to basic vocabulary, such as the Swadesh word list, the influence of borrowing can be minimized. For example, although English is a Germanic language, it has borrowed around 60 per cent of its total lexicon from French and Latin. However, only about 6 per cent of English entries in the Swadesh 200-word list are clear Romance language borrowings [30]. Known borrowings were not coded as cognate in the Dyen *et al.* database. The cognate sets were binary-coded—that is in a matrix a column was set up for each cognate set in which the presence of a cognate for a language was denoted with a '1' and an absence with a '0'. This produced a matrix of 2449 cognate sets for 87 languages. This matrix was analysed in the Bayesian phylogenetics package MRBAYES [31] using a simple model that assumed equal rates of cognate gains and losses to produce a sample of trees from the posterior probability distribution of the trees (the set of trees found in the Markov chain Monte Carlo runs post 'burn in' given the data, model of cognate evolution and priors on variables such as the parameters of the model and branch lengths). In order to infer divergence dates, we needed to calibrate the rates of evolution by constraining the age of nodes on each tree in accordance with historically attested dates. For example, the Romance languages probably began to diverge prior to the fall of the Roman Empire. The last Roman troops were withdrawn south of the Danube in AD 270. Thus, we constrained the age of the node corresponding to the most recent common ancestor of the Romance languages to AD 150–300. We constrained the age of 14 nodes on the trees. The penalized rate-smoothing algorithm was then used to covert the set of trees into dated 'chronotrees' (see [32] for more details on the methods and calibrations used).

Our initial analyses provided strong support for the time-depth predictions of Anatolian hypothesis. The date estimates for the age of Proto Indo-European centred around 8700 BP (figure 1). None of our sample of chronotrees was in the 5000–6000 years BP age range predicted by the Kurgan hypothesis. A key part of any Bayesian phylogenetic analysis is an assessment of the robustness of the inferences. We did our best to try and 'break' the initial result. We examined the impact of altering the branch length priors in our analysis, of throwing out cognates Dyen *et al.* had dubbed 'dubious', of removing some calibrations, of trimming the data to the most stable items and rerooting the trees. None of these changes substantially altered our date estimates of the age of Proto Indo-European. If anything, they often tended to make the distribution older, not younger [32,33].

The response to our paper was rather mixed. While some linguists were positive, many simply failed to understand that the methods we had used were substantially different from traditional glottochronology [34]. A small number of critics raised concerns about the data we had used, the binary coding of the cognate sets, the simple model of cognate evolution and the impact of undetected borrowing. Let us deal with each of these potentially valid concerns in turn.

First, although the cognate coding in the Dyen *et al.* dataset was conducted by experienced linguists, it may well contain some errors [35]. While these errors are likely to be a relatively small proportion of the total data, it is possible that they might have biased our date estimates. It is also possible that the simple stochastic model of cognate evolution we used led to inaccurate results because the model assumed that the rates of cognate gain and loss were equal—an assumption that is not realistic. It is rare for very similar words with similar meanings to be independently invented [36]. A more realistic model would thus allow cognates to be gained only once but lost multiple times. This mirrors the principle in evolution biology known as Dollo's Law, which suggests that once complex structures are lost they are unlikely to be evolved again. While simple models do not necessarily produce inaccurate results [33], in Bayesian analyses it is important to assess the robustness of the conclusions to any model misspecification. For this reason, Geoff Nicholls and R.G. developed a stochastic 'Dollo' model of cognate evolution [37]. We used this model to analyse an independent dataset [38], predominantly comprising ancient Indo-European languages. These analyses of a separate dataset with an entirely different model produced almost identical results to our initial analyses of the Dyen data [37,39]. Not content with this proof of the robustness of our analyses, we recently re-analysed the Ringe *et al.* data using the lognormal relaxed clock and the stochastic Dollo model implemented in the package BEAST [40]. Yet again the date estimates for Proto Indo-European fell into the age range predicted by the Anatolian hypothesis (figure 2). Re-analysing the Dyen *et al.* data with the lognormal relaxed clock and the stochastic Dollo model also produced results that are highly congruent with the initial results of Gray & Atkinson.

If either problems with the data or the model of cognate evolution appear to have biased our results, what about the binary coding of the cognate sets? Evans *et al.* [41] claim that our coding is 'patently inappropriate' because it assumes independence between the cognate sets. Our sets are clearly not independent because one form will often replace another within a meaning class (although some polymorphism does occur). On the surface this is a plausible argument. However, Evans *et al.* provide no argument for why the lack of independence will bias the time-

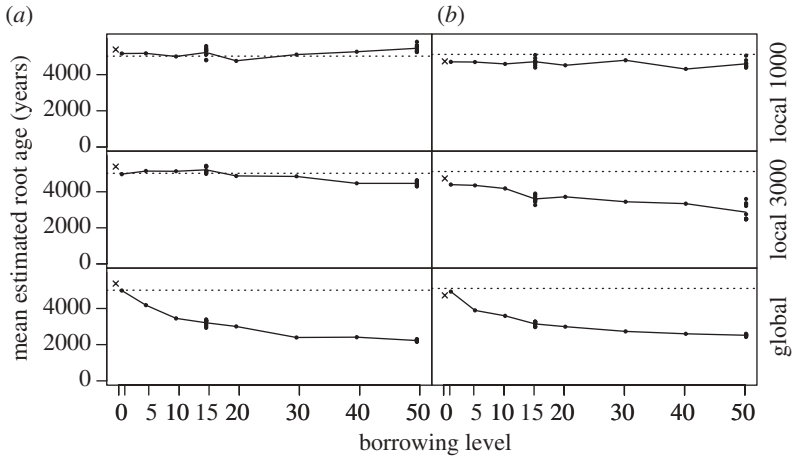


Figure 3. Mean reconstructed root time for each simulation under three borrowing scenarios: (i) local borrowing within 1000 years, (ii) local borrowing within 3000 years, and (iii) global borrowing. Two different tree topologies were used in the simulations: (*a*) tree 1 and (*b*) tree 2. The dotted line marks the true root age and the cross marks the root age under the no borrowing scenario.

depth estimates to be too old (rather than merely underestimating the variance). On the contrary, we have simulated totally dependent cognate evolution and shown that it does not bias the date estimates [39]. Others have found empirically that binary and multi-state codings of the same lexical data produce virtually identical results [42]. Furthermore, Pagel & Mead [43] demonstrated that, at least when the number of states is constant, binary and multi-state-coded data produce trees that differ only in length by a constant proportionality. In other words, the binary and multi-state trees are just scaled versions of one another and therefore the date estimates will not be biased. This result is also likely to hold when the number of states varies (M. Pagel 2010, personal communication).

Removing all the borrowed cognates from a dataset can be difficult. While irregular sound correspondences might make some easy to identify, others may be difficult to detect. Garrett [44] argues that borrowing of lexical terms, or advergence, within the major Indo-European subgroups could have distorted our results. To assess this possibility, we examined the impact of different borrowing scenarios by simulating cognate evolution [45]. The results showed that tree topologies constructed with Bayesian phylogenetic methods were robust to realistic levels of borrowing in basic vocabulary (0–15%). Inferences about divergence dates were slightly less robust and showed a tendency to *underestimate* dates (figure 3). The effect is pronounced only when there is global rather than local borrowing on the tree. This is the least likely scenario we simulated and suggests that if our estimates for the age of Indo-European are biased by undetected borrowing at all, they are likely to be too young, rather than too old.

While all these re-analyses and simulation studies demonstrate the reliability of our estimates for the age of Indo-European, perhaps the most compelling refutation of our critics' arguments comes from the model validation analyses we recently conducted. Nicholls & Gray [46] sequentially removed calibration points from some analyses conducted using the stochastic Dollo model implemented in the program TRAITLAB. We then re-ran the analysis and examined the date estimates of these nodes. If model misspecification meant that our age estimates were systematically too old, then the estimated ages should be systematically older than the known ages of the nodes in the trees that we removed the calibrations from. This was not the case. Overwhelmingly, the estimates were congruent with the known node ages.

## 4. THE AUSTRONESIAN EXPANSION

The Austronesian settlement of the vast Pacific Ocean has been a topic of enduring fascination. It is the greatest human migrations in terms of the distance covered and the most recent. There are two major hypotheses for the Austronesian settlement of the Pacific. The first hypothesis is the 'pulse–pause' scenario [5,47–49]. This scenario argues that the ancestral Austronesian society developed in Taiwan around 5500 years ago. Around 4000–4500 years ago, there was a rapid expansion pulse across the Bashii channel into the Philippines, into Island Southeast Asia, along the coast of New Guinea, reaching Near Oceania by around 3000–3300 years ago [50]. As the Austronesians travelled this route, they integrated with the existing populations in the area (particularly in New Guinea), and innovated new technologies. After reaching Western Polynesia (Fiji, Tonga and Samoa) approximately 3000 years ago, the Austronesian expansion paused for around 1000–1500 years, before a second rapid expansion pulse spread Polynesian languages as far afield as New Zealand, Hawaii and Rapanui.

The second hypothesis of Pacific settlement—the 'slow boat' scenario—argues for a much older origin in Island Southeast Asia [51–53]. According to this scenario, date estimates from mitochondrial DNA lineages suggest that Austronesian society developed around 13 000–17 000 years ago in an extensive network of sociocultural exchange in the Wallacean region around Sulawesi and the Moluccas. Proponents of this scenario propose that the submerging of the Sunda shelf at the end of the last ice-age triggered

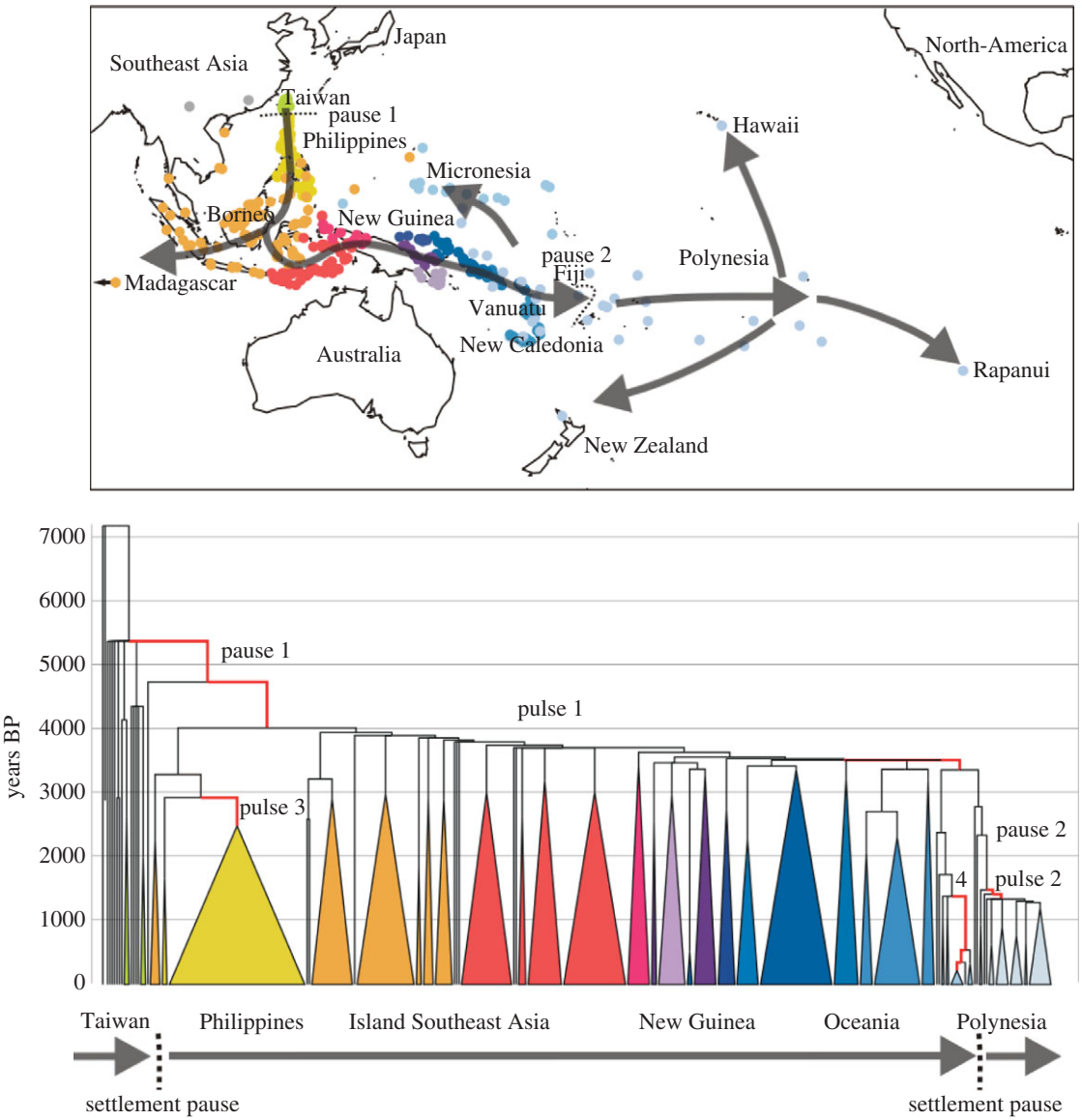


Figure 4. Map and language family tree showing the settlement of the Pacific by Austronesian-speaking peoples. The map shows the settlement sequence and location of expansion pulses and settlement pauses. The tree is rooted with some outgroup languages (Buyang and Old Chinese) at its base. It shows an Austronesian origin in Taiwan around 5200 years ago, followed by a settlement pause (pause 1) between 5200 and 4000 years ago. After this pause, a rapid expansion pulse (pulse 1) led to the settlement of Island Southeast Asia, New Guinea and Near Oceania in less than 1000 years. A second pause (pause 2) occurs after the initial settlement of Polynesia. This pause is followed by two pulses further into Polynesia and Micronesia around 1400 years ago (pulses 2 and 4). A third expansion pulse occurred around 3000–2500 years ago in the Philippines.

the Austronesian expansion [53]. This 'flood' led to a two-pronged movement of people, north into the Philippines and Taiwan, and east into the Pacific. Significantly, they argue that this movement of people was paralleled by the spread of Austronesian languages (i.e. Austronesian genes and languages have a common history). 'The Austronesian languages originated within island Southeast Asia during the Pleistocene era and spread through Melanesia and into the remote Pacific within the past 6000 years' [54, p. 1236].

These two scenarios of Pacific settlement make quite different predictions about the origin, age and sequence of the Austronesian expansion. The pulse–pause model predicts that a phylogenetic tree of Austronesian languages should be rooted in Taiwan and show a chained topology that mirrors the generally eastwards spread of the languages. According to this model, the Austronesian language family should be about 5500 years old. Most boldly, the model predicts that there should be a long pause between the Taiwanese languages and the rest of Austronesian, followed by a rapid diversification pulse and then another long pause in Polynesia. In contrast, the slow boat model predicts that any language family tree should be rooted in the Wallacean region, be between 13 000 and 17 000 years old and have a two-pronged topology with one branch going north to the Philippines and Taiwan and the other eastwards along the New Guinea coast out into Oceania.

Clearly, a robustly dated language phylogeny would be an ideal way to test between the pulse–pause and slow boat models of Austronesian expansion. However, the construction of an accurate, dated language phylogeny for the Austronesian languages provides numerous challenges for any would be language

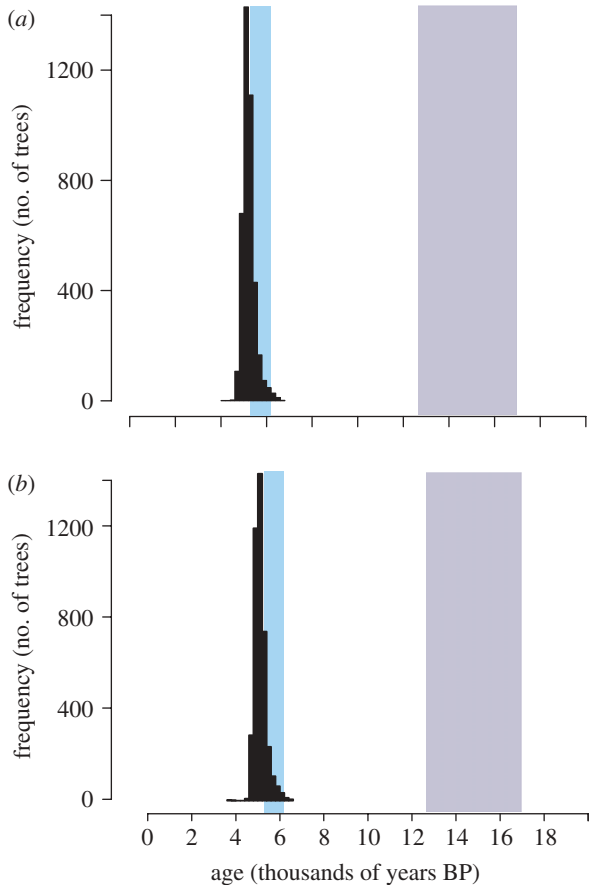


Figure 5. Histograms of the Bayesian phylogenetic estimates for the age of Proto Austronesian. (*a*) Shows the estimated age when all calibrations were used. (*b*) Shows the estimates when only Proto Oceanic and three ancient languages were used as calibrations.

phylogenticist. First, the rapid expansion of the Austronesian family means that it is likely to be difficult to resolve the fine branching structure of the Austronesian language tree as there is little time for the internal branches on the tree to develop numerous shared innovations [55]. Second, as these languages moved across the Pacific, they encountered new environments and the consequent need for new terminology may have increased the rates of language replacement. This acceleration in rates is likely to be exacerbated by the effects of language contact—particularly within Near Oceania [56]. Additionally, many Austronesian languages have small speech communities, which are also likely to speed up the rates of language evolution [57]. The effects of these factors can be seen in the 10-fold variation in cognate retention rates in Austronesian languages [13].

Successful phylogenetic analyses require data with sufficient historical information to resolve the aspects of the phylogeny we are interested in. Over the past 7 years we have compiled a large web-accessible database of cognate-coded basic vocabulary for over 700 Austronesian languages [16]. This database was initially based on 230 language word lists we obtained from Bob Blust, but by placing it on the web we have been able to grow and refine the database and cognate coding with the generous assistance of linguists around the globe. In the 400-language dataset reported in Gray *et al.* [49], the 210 items of basic vocabulary produced a matrix of 34 440 binary-coded cognate sets.

The first prediction we tested with the Bayesian phylogenetic analyses of this data concerned the origin and sequence of Austronesian expansion. Under the pulse–pause scenario, the Austronesians originated in Taiwan and had a single-chained expansion down through the Philippines, through Wallacea, along New Guinea into Near Oceania and Polynesia. In contrast, the slow boat scenario posits a two-pronged expansion from a Wallacean origin. Our set of trees placed the root of trees in Taiwan, and followed it with the sequence predicted by the pulse–pause scenario (figure 4).

The second key prediction of the two Pacific settlement scenarios concerned the age of the expansion. To test this prediction, we estimated the age at the root of our trees. To begin with, we calibrated 10 nodes on the trees with archaeological date estimates and known settlement times. For example, speakers of the Chamic language subgroup were described in Chinese records around 1800 years ago and probably entered Vietnam around 2600 years ago [58]. We can therefore calibrate the appearance of the Chamic node on our tree to between 2000 and 3000 years ago. A second calibration, based on archaeological evidence, constrains the age of the hypothesized ancestral language spoken by all the languages of Near Oceania, Proto Oceanic. The speakers of Proto Oceanic arrived in Oceania around 3000–3300 years ago and brought with them distinctively Austronesian societal organization and cultural artefacts. These artefacts have been identified and dated archaeologically, and include the Lapita adze/axe kits, housing types, fishing equipment (such as the one-piece rotating fishhooks, and one-piece trolling lure), as well as common food plants and domesticated animals from Southeast Asia.

To estimate the age of the Austronesian family without assuming a strict glottoclock, we used the penalized likelihood approach outlined above. The results unequivocally supported the younger age of the pulse–pause scenario, with an origin of the Austronesian family around 5200 years ago (figure 5*a*). Like the Indo-European analyses, the results were robust to assumptions about specific calibration points. For example, when we removed all the calibration points, apart from the Proto Oceanic constraint and the three ancient languages [59], the estimated age of Proto Austronesian was virtually identical (figure 5*b*).

The pulse–pause scenario makes a third key prediction by proposing a sequence of expansion pulses and pauses. Under this scenario, there were two pauses in the great expansion—the first occurred before the Austronesians entered the Philippines around 5000–4000 years ago, and the second occurred after the settlement of Western Polynesia (Fiji, Samoa, Tonga) starting around 2800 years ago. We tested this prediction in two ways. First, we identified the branches on our trees corresponding to these two pauses (figure 4). The length of the branches again represents the number of changes in cognate sets. If these pauses did occur, then those branches should be much longer

than others owing to the increased amount of time for linguistic change. Indeed, the length of these branches was significantly longer than the overall branch-length distribution, providing good evidence that pauses did occur in the predicted locations.

The pulse–pause scenario predicts pulses as well as pauses. If there were expansion pulses in language change, then we would expect to see increases in language diversification rates after the predicted pauses. To test this prediction, we modelled language diversification rate over our set of language trees. This method identified a number of significant increases in language diversification rates (branches coloured red in figure 4). Two of these increases occurred as predicted on the branches just after the two pauses. Intriguingly, we identified some unpredicted pulses as well. The third pulse we identified suggested a more recent population expansion in the Philippines around 2000–2500 years ago as one language subgroup expanded at the expense of others. The fourth pulse occurred in the Micronesian languages and appears to be linked to the second pulse into Polynesia.

What insights can these language dates give us about the great Austronesian expansion? It has been suggested that the first pause might be linked to an inability of the Austronesians to cross the 350 km Bashi channel separating the Philippines from Taiwan [47,48]. Terms for outrigger canoes and sails can only be reconstructed back to the languages occurring after the first pause [47,48]. It seems likely therefore that the invention of the outrigger enabled the Austronesians to cross the channel and spread rapidly across the rest of the Pacific. After travelling 7000 km in 1000 years, what might have caused the Austronesians to stop in Western Polynesia? Expanding into Eastern Polynesia presented the Austronesians with a new range of challenges that would have also required technological or social solutions including: the ability to estimate latitude from the stars, the ability to sail across the prevailing easterly tradewinds, double-hulled canoes for greater stability and carrying capacity, and social strategies for handling the greater isolation [60].

The results reveal the rapidity of cultural spread. The Austronesians travelled—and settled—the 7000 km between the Philippines and Polynesia in around 1000 years. During this relatively short time, the Austronesian culture not only spread, but developed the collection of technologies known as the Lapita cultural complex [5]. This complex includes distinctive and elaborately decorated pottery, adzes/axes, tattooing, barkcloth and shell ornamentation. Our results suggest that either this complex was generated in a very short time-window (four or five generations), or there was substantial post-settlement contact between Near Oceania and the pre-Polynesian society. One possibility is that there is a more complex history in this region. The languages of New Caledonia and Vanuatu show some strikingly non-Austronesian features such as serial verb constructions, and the cultures there show some unusual similarities with some cultures from highland New Guinea—including nasal septum piercings, penis sheathes and mop-like headdresses [61]. It has recently been suggested that one explanation for these similarities might be two waves of settlement into Remote Oceania, with a first wave of Austronesian-speaking settlers being rapidly followed by a second wave of Papuan peoples who had acquired Austronesian voyaging technology [61].

## 5. CONCLUSION

Some scholars are rather sceptical that anything of substance can come out of attempts to 'Darwinize culture'. They concede that some loose analogies can be found, but claim that these are rather superficial and unlikely to yield substantive insights into complex cultural processes. In the words of Fracchia & Lewontin [62, p. 14], Darwinian approaches to culture do not 'contribute anything new except a misleading vocabulary that anesthetises history'. The focus of Fracchia & Lewontin's critique is on selectionist, memetic accounts of historical change. Elsewhere, we have argued that phylogenetic or 'tree thinking' provides another way of Darwinizing culture that does not require a commitment to problematic aspects of memetics such as particulate cultural inheritance and tidy lineages of directly copied replicators [63]. One aspect of phylogenetic inference is the estimation of divergence dates. The accurate estimation of divergence dates is a tricky business. Care needs to be taken to ensure that the calibrations are valid and the inferences are robust to possible model misspecification and undetected borrowing. However, the central theme of this paper has been that when it comes to understanding our past these carefully estimated dates really do make a difference. Robust phylogenetic estimates of linguistic divergence dates give us a powerful tool for testing hypotheses about human prehistory. They enable us to integrate linguistic, archaeological and genetic data, and link major population expansions to innovations in culture such as the development of farming and the invention of the outrigger canoe. In short, a phylogenetic approach to culture illuminates rather than anaesthetizes history.

I would like to thank Lyle Campbell, Kevin Laland, April McMahon, Robert Ross, Andy Whiten and an anonymous referee for their helpful comments on this manuscript. Figures are reprinted with permission from *Proceedings of the Royal Society B* (figure 3) and *Science* (figure 4). This work was funded by Marsden grants from the Royal Society of New Zealand.

## REFERENCES

1 Zuckerkandl, E. & Pauling, L. 1965 Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366. (doi:10.1016/0022-5193(65)90083-4)
2 Evans, N. 2010 *Dying words: endangered languages and what they have to tell us*. Oxford, UK: Blackwell.
3 Durie, M. & Ross, M. 1996 *The comparative method reviewed: regularity and irregularity in language change*. New York, NY: Oxford University Press.
4 Campbell, L. & Poser, W. J. 2008 *Language classification: history and method*. Cambridge, UK: Cambridge University Press.
5 Kirch, P. V. & Green, R. C. 2001 *Hawaiki, Ancestral Polynesia. An essay in historical anthropology*. Cambridge, UK: Cambridge University Press.

6 Renfrew, C. 2002 The 'emerging synthesis': the archaeogenetics of farming/language dispersals and other spread zones. In *Examining the farming/language dispersal hypothesis* (eds P. Bellwood & C. Renfrew). Cambridge, UK: McDonald Institute for Archaeological Research.
7 Renfrew, C. 2010 Archaeogenetics: towards a 'new synthesis'? *Curr. Biol.* **20**, R162–R165. (doi:10.1016/j.cub.2009.11.056)
8 Ho, S. Y. W. & Larson, G. 2006 Molecular clocks: when times are a-changin'. *Trends Genet.* **22**, 79–83. (doi:10.1016/j.tig.2005.11.006)
9 McMahon, A. & McMahon, R. 2006 Why linguists don't do dates: evidence from Indo-European and Australian languages. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 153–160. Cambridge, UK: McDonald Institute for Archaeological Research.
10 Swadesh, M. 1952 Lexicostatistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 452–463.
11 Swadesh, M. 1955 Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Linguist.* **21**, 121–137. (doi:10.1086/464321)
12 Bergsland, K. & Vogt, H. 1962 On the validity of glottochronology. *Curr. Anthropol.* **3**, 115–153. (doi:10.1086/200264)
13 Blust, R. 2000 Why lexicostatistics doesn't work: the 'universal constant' hypothesis and the Austronesian languages. In *Time depth in historical linguistics* (eds C. Renfrew, A. McMahon & L. Trask), pp. 311–332. Cambridge, UK: McDonald Institute for Archaeological Research.
14 Pagel, M. 1999 Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884. (doi:10.1038/44766)
15 Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314. (doi:10.1126/science.1065889)
16 Greenhill, S. J., Blust, R. & Gray, R. D. 2008 The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. *Evol. Bioinformatics* **4**, 271–283.
17 Huelsenbeck, J. P. 1995 Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**, 17–48.
18 Sanderson, M. 2002 R8s, analysis of rates of evolution, version 1.50. See http://ginger.ucdavis.edu/r8s/.
19 Sanderson, M. 2002 Estimating absolute rates of evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109.
20 Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. 2006 Relaxed phylogenies and dating with confidence. *PLoS Biol.* **4**, e88. 699–710. (doi:10.1371/journal.pbio.0040088)
21 Diamond, J. & Bellwood, P. 2003 Farmers and their languages: the first expansions. *Science* **300**, 597–603. (doi:10.1126/science.1078208)
22 Mallory, J. P. 1989 *In search of the Indo Europeans: language, archaeology and myth*. London, UK: Thames and Hudson.
23 Gimbutas, M. 1973 Old Europe c. 7000–3500 BC, the earliest European cultures before the infiltration of the Indo-European peoples. *J. Indo-Eur. Stud.* **1**, 1–20.
24 Gimbutas, M. 1973 The beginning of the Bronze Age in Europe and the Indo-Europeans 3500–2500 BC. *J. Indo-Eur. Stud.* **1**, 163–214.
25 Trask, L. 1996 *Historical linguistics*. New York, NY: Arnold.
26 Renfrew, C. 1987 *Archaeology and language: the puzzle of Indo-European origins*. London, UK: Cape.
27 Gkiasta, M., Russell, T., Shennan, S. & Steele, J. 2003 Neolithic transition in Europe: the radiocarbon record revisited. *Antiquity* **77**, 45–62.
28 Atkinson, Q. & Gray, R. D. 2005 Curious parallels and curious connections: phylogenetic thinking in biology and historical linguistics. *Syst. Biol.* **54**, 513–526. (doi:10.1080/10635150590950317)
29 Dyen, I., Kruskal, J. B. & Black, P. 1997 FILE IE-DATA1. See http://www.ntu.edu.au/education/langs/ielex/IE-DATA1.
30 Embleton, S. 1986 *Statistics in historical linguistics*. Bochum, Germany: Brockmeyer.
31 Huelsenbeck, J. P. & Ronquist, F. 2001 MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754–755. (doi:10.1093/bioinformatics/17.8.754)
32 Gray, R. D. & Atkinson, Q. D. 2003 Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439. (doi:10.1038/nature02029)
33 Atkinson, Q. D. & Gray, R. D. 2006 How old is the Indo-European language family? Progress or more moths to the flame? In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 91–109. Cambridge, UK: McDonald Institute for Archaeological Research.
34 Balter, M. 2003 Early date for the birth of Indo-European languages. *Science* **302**, 1490–1491. (doi:10.1126/science.302.5650.1490a)
35 Johnson, K. 2008 *Quantitative methods in linguistics*. Malden, MA: Blackwell.
36 Campbell, L. 2004 *Historical linguistics: an introduction*, 2nd edn. Edinburgh, UK: Edinburgh University Press.
37 Nicholls, G. K. & Gray, R. D. 2006 Quantifying uncertainty in a stochastic Dollo model of vocabulary evolution. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 161–171. Cambridge, UK: McDonald Institute for Archaeological Research.
38 Ringe, D., Warnow, T. & Taylor, A. 2002 Indo-European and computational cladistics. *Trans. Phil. Soc. B* **100**, 59–129. (doi:10.1111/1467-968X.00091)
39 Atkinson, Q., Nicholls, G. & Gray, R. D. 2005 From words to dates: water into wine, mathemagic or phylogenetic inference? *Trans. Phil. Soc.* **103**, 193–219. (doi:10.1111/j.1467-968X.2005.00151.x)
40 Drummond, A. J. & Rambaut, A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)
41 Evans, S. N., Ringe, D. & Warnow, T. 2006 Inference of divergence times as a statistical inverse problem. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 119–129. Cambridge, UK: McDonald Institute for Archaeological Research.
42 Kitchen, A., Ehret, C., Assefa, S. & Mulligan, C. J. 2009 Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East. *Proc. R. Soc. B* **276**, 2703–2710. (doi:10.1098/rspb.2009.0408)
43 Pagel, M. & Meade, A. 2006 Estimating rates of lexical replacement on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 173–182. Cambridge, UK: McDonald Institute for Archaeological Research.
44 Garrett, A. 2006 Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In *Phylogenetic methods and the prehistory of languages* (eds P. Forster & C. Renfrew), pp. 139–152. Cambridge, UK: McDonald Institute for Archaeological Research.
45 Greenhill, S. J., Currie, T. E. & Gray, R. D. 2009 Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. B* **276**, 2299–2306. (doi:10.1098/rspb.2008.1944)

46 Nicholls, G. K. & Gray, R. D. 2008 Dated ancestral trees from binary trait data and its application to the diversification of languages. *J. R. Stat. Soc. B* **70**, 545–566. (doi:10.1111/j.1467-9868.2007.00648.x)
47 Blust, R. 1999 Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. In *Selected papers from the Eighth International Conference on Austronesian Linguistics* (eds E. Zeitoun & P. J. K. Li), pp. 31–94. Taipei, Taiwan: Academia Sinica.
48 Pawley, A. 2002 The Austronesian dispersal: languages, technologies and people. In *Examining the farming/language dispersal hypothesis* (eds P. Bellwood & C. Renfrew), pp. 251–274. Cambridge, UK: McDonald Institute for Archaeological Research.
49 Gray, R. D., Drummond, A. J. & Greenhill, S. J. 2009 Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483. (doi:10.1126/science.1166858)
50 Spriggs, M. 2010 'I was so much older then, I'm younger than that now': why the dates keep changing for the spread of Austronesian languages. In *A journey through Austronesian and Papuan linguistic and cultural space: papers in honour of Andrew K. Pawley* (eds J. Bowden, N. Himmelmann & M. Ross), pp. 113–140. Canberra, Australia: Pacific Linguistics.
51 Oppenheimer, S. & Richards, M. 2001 Fast trains, slow boats and the ancestry of the Polynesian islanders. *Sci. Prog.* **84**, 157–181. (doi:10.3184/00368500178 3238989)
52 Hill, C. *et al.* 2007 A mitochondrial stratigraphy for island southeast Asia. *Am. J. Hum. Genet.* **80**, 29–43. (doi:10.1086/510412)
53 Soares, P. *et al.* 2008 Climate change and postglacial human dispersals in southeast Asia. *Mol. Biol. Evol.* **25**, 1209–1218. (doi:10.1093/molbev/msn068)
54 Richards, M., Oppenheimer, S. & Sykes, B. 1998 mtDNA suggests Polynesian origins in eastern Indonesia. *Am. J. Hum. Genet.* **63**, 1234–1236. (doi:10.1086/302043)
55 Pawley, A. 1999 Chasing rainbows: implications of the rapid dispersal of Austronesian languages for subgrouping and reconstruction. In *Selected papers from the Eighth International Conference on Austronesian Linguistics*, vol. 1 (eds E. Zeitoun & P. J. K. Li), pp. 95–138. Taipei, Taiwan: Academia Sinica.
56 Ross, M. 1996 Contact-induced change and the comparative method: cases from Papua New Guinea. In *The comparative method reviewed: regularity and irregularity in language change* (eds M. Durie & M. D. Ross), pp. 180–217. New York, NY: Oxford University Press.
57 Nettle, D. 1999 Is the rate of linguistic change constant? *Lingua* **108**, 119–136. (doi:10.1016/S0024-3841(98)00047-3)
58 Thurgood, G. 1999 *From ancient Cham to modern dialects: two thousand years of language contact and change*. Hawaii: University of Hawaii Press.
59 Greenhill, S. J., Drummond, A. J. & Gray, R. D. 2010 How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PLoS ONE* **5**, e9573. (doi:10.1371/journal.pone.0009573)
60 Irwin, G. 1998 The colonization of the Pacific: chronological, navigational and social issues. *J. Polynesian Soc.* **107**, 111–144.
61 Blust, R. 2008 Remote Melanesia: one history or two? An addendum to Donohue and Denham. *Oceanic Linguist.* **47**, 445–459. (doi:10.1353/ol.0.0012)
62 Fracchia, J. & Lewontin, R. C. 2005 The price of metaphor. *History Theory* **44**, 14–29. (doi:10.1111/j.1468-2303.2005.00305.x)
63 Gray, R. D., Greenhill, S. J. & Ross, R. M. 2007 The pleasures and perils of Darwinzing culture (with phylogenies). *Biol. Theory* **2**, 360–375. (doi:10.1162/biot.2007.2.4.360)