Journal of Language Evolution, 2019, 1–22
doi: 10.1093/jole/lzz009
Research article

Best practices in justifying calibrations for dating language families

L. Maurits *,†, **M. de Heer**‡, **T. Honkola**§, **M. Dunn****, **and O. Vesakoski**§

†Department of Geography and Geology, University of Turku, Turku, Finland, ‡Department of Modern Languages, Finno-Ugric Languages, University of Uppsala, Uppsala, Sweden, §Department of Biology, University of Turku, Turku, Finland and **Department of Linguistics and Philology, University of Uppsala, Uppsala, Sweden

*Corresponding author: luke@maurits.id.au

Abstract

The use of computational methods to assign absolute datings to language divergence is receiving renewed interest, as modern approaches based on Bayesian statistics offer alternatives to the discredited techniques of glottochronology. The datings provided by these new analyses depend crucially on the use of calibration, but the methodological issues surrounding calibration have received comparatively little attention. Especially, underappreciated is the extent to which traditional historical linguistic scholarship can contribute to the calibration process via loanword analysis. Aiming at a wide audience, we provide a detailed discussion of calibration theory and practice, evaluate previously used calibrations, recommend best practices for justifying calibrations, and provide a concrete example of these practices via a detailed derivation of calibrations for the Uralic language family. This article aims to inspire a higher quality of scholarship surrounding all statistical approaches to language dating, and especially closer engagement between practitioners of statistical methods and traditional historical linguists, with the former thinking more carefully about the arguments underlying their calibrations and the latter more clearly identifying results of their work which are relevant to calibration, or even suggesting calibrations directly.

Key words: calibration; phylogenetics; Bayesian methods; historical linguistics; divergence timing

1. Introduction

Absolute dating in linguistics faces major methodological challenges and the drastic shortcomings of the earliest approaches (especially glottochronology) have left many practitioners uninterested in or sceptical of all subsequent approaches (McMahon and McMahon 2006). However, modern approaches based on Bayesian phylogenetic inference, which constitute a real and substantial advance beyond glottochronology, are increasingly being applied to language dating problems (Bouckaert, Bowern, and Atkinson, 2018; Bouckaert et al. 2012; Gray and Atkinson 2003; Gray, Drummond, and Greenhill 2009; Grollemund et al. 2015; Honkola et al. 2013; Hruschka et al. 2015; Kitchen et al. 2009; Kolipakam et al. 2018; Lee and Hasegawa 2011; Sagart et al. 2019). At the same time, there is growing interest in the prospect of combining evidence from diverse fields such as linguistics, archaeology, and genetics to enable a holistic study of human history (Haak et al. 2015; Ilumäe et al. 2016; Lang 2018; Tambets et al. 2018). Rigorous language divergence timings will be essential to any such enterprise, as time serves as a

'common currency' between disciplines. The stage is thus set for a renewed discussion of how principled absolute dating can be achieved, tackling rather than shying away from difficult issues.

Modern quantitative approaches to language dating (see e.g. Dunn (2015) or Nichols and Warnow (2008) for summaries) are crucially dependent upon calibrations— the precise specification of prior knowledge about the date of at least one and preferably several points on a phylogenetic tree. This dependence is due to the lack of a natural 'clock' in linguistic change, unlike in, for example, carbon isotope decay and, to a lesser extent, genetic mutation, which allowed archaeology and biology both to turn much earlier to advanced statistical methods for dating (Buck, Cavanagh, and Litton 1996; Kumar 2005; Kumar and Hedges 2016; Ramsey 2009). The process by which calibrations are selected for estimating language family ages is fundamental to the reliability of the age estimates obtained. Despite this, relatively little has been written about the calibration process and best practices for establishing and reporting linguistic calibrations. In contrast, the biology literature has explicitly discussed best practices for justifying fossil calibrations (Parham et al. 2011).

Increased discussion of the linguistic calibration process would offer several benefits. First, clear guidelines on how to make and report good calibrations would enable consistent peer review and prevent the careless use of new dating methods. Secondly, clear and accessible explanations of what calibrations are and how they are used would enable traditional historical linguists to provide or expertly assess them. This is important as establishing reliable calibrations requires a close engagement with the relevant historical linguistic literature and the researchers producing it. Encouraging this kind of close engagement in Bayesian phylogenetic linguistics is a key motivation of this article.

In this article, we aim to promote a well-informed and carefully developed approach to absolute linguistic dating using calibrated Bayesian phylogenetic analyses, emphasising the use of calibrations based on careful, explicit, and transparent arguments. We target a broad audience, consisting primarily of both historical linguists and Bayesian practitioners, who can work together to produce such datings. However, we also aim to reach all historical scientists who may be able to use high-quality linguistic datings to better conceptualise their own work. This includes, among others, archaeologists, population geneticists, and archaeogeneticists. We first familiarise the reader with the essential concepts of these analyses and the calibration procedure before examining key theoretical and practical issues surrounding the derivation of calibrations from different kinds of sources. To provide concrete examples of the explained principles and advocated practices, we include a detailed case study deriving calibrations for the Uralic languages. We reproduce a previous Bayesian analysis of Uralic (Honkola et al. 2013) using these new calibrations in order to demonstrate the substantial impact that revised calibrations can have. However, this demonstrative analysis is not the main focus of the article and we stress that its results should not be taken as an earnest contribution to Uralic studies.

2. Theoretical background

This article is concerned with linguistic phylogenetic analyses performed according to the Bayesian statistical paradigm. Other computational approaches to phylogenetic analysis exist (including maximum likelihood, maximum parsimony, and distance-based approaches—see Felsenstein (2004) for an overview), but the Bayesian approach is by far the best suited to estimating divergence timings and has been used by all recent dating efforts (Bouckaert, Bowern, and Atkinson 2018; Kolipakam et al. 2018; Sagart et al. 2019). An introduction to Bayesian statistics, in general, is well beyond the scope of this article (for a brief formal treatment, see Wasserman (2004)), and is not required to appreciate the important points regarding calibration, which are the primary focus here.

2.1 Timing in Bayesian phylogenetics

In Bayesian phylogenetic analyses, language families are represented by sets of strictly binary-branching trees, where languages are represented as *leaf nodes* and the family's protolanguage as the *root node*. The nodes are connected by *branches*, each of which has a *branch length* associated with it. This is a numerical value representing the amount of time between the *branching events* represented by the nodes at either end of the branch. When the purpose of an analysis is to estimate the age of the family, each of these branch lengths is understood to represent 'real world' time, and is typically given in units of centuries or millennia. The sum of the branch lengths from any extant language to the root of the tree must be the same, and this sum represents the age of the protolanguage. For example, Fig. 1 represents a hypothetical language family, which is 4,000 years old (branch lengths are in units of millennia). The family contains five languages, which are divided into two subfamilies. One subfamily consists of two languages, A and B, whose common ancestor is 3,000 years old, while

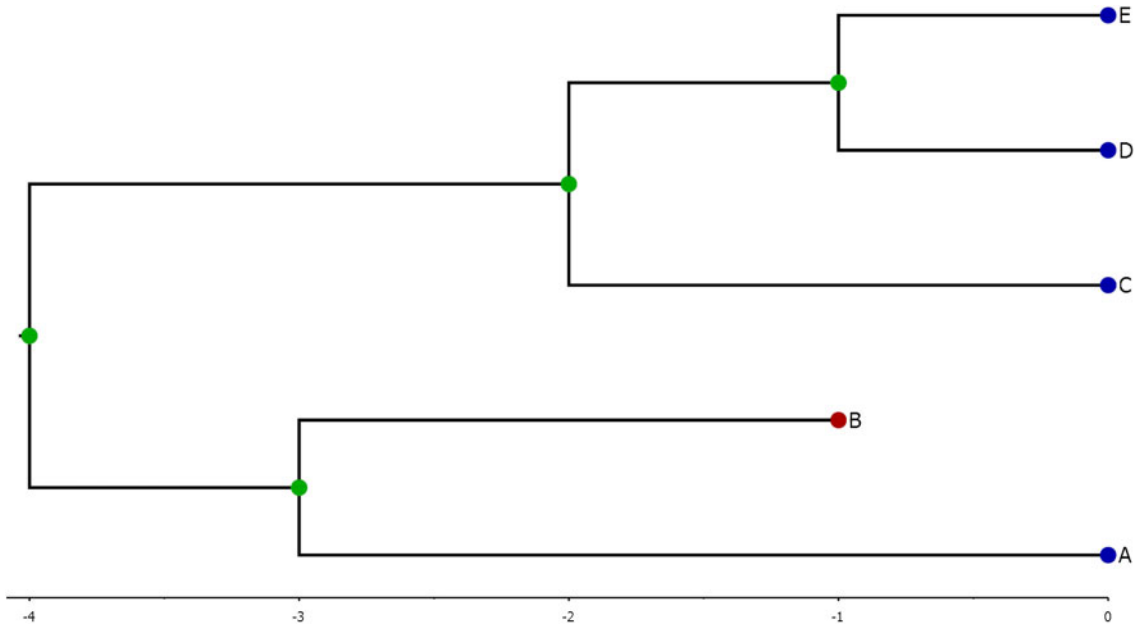

Figure 1. An example of phylogenetic tree, showing five languages with a 4,000-year-old common ancestor. The arrow of time points left to right, and the scale indicates times in millenia before present. Extant languages correspond to the leaf nodes coloured in green, while the single extinct language (B) is represented by a leaf node coloured in red. Ancestral languages, those which are mostly commonly calibrated in analyses, correspond to the interior nodes coloured in blue.

the other subfamily consists of three languages, C, D, and E, whose common ancestor is 2,000 years old. Language B went extinct 1,000 years ago and is not contemporaneous with the other languages. Languages D and E are more closely related to each other than to language C, with a 1,000-year-old common ancestor.

The results of a Bayesian phylogenetic analysis are commonly visualised as a single phylogenetic tree, such as that in Fig. 1. However, this single tree is only a convenient summary of a very large number of trees produced in the course of the analysis— thousands or tens of thousands of trees are typical (Felsenstein 2004; Nascimento, Reis, and Yang 2017). Each tree in this set represents a possible history, represented by a tree *topology* (i.e. branching structure), age of the root, times separating each pair of languages, as well as other possible analytic parameters. These trees are a statistical sample, selected from the set of all possible trees, and output by the MCMC algorithm[1] in proportion to the strength of evidence supporting that particular history. The histories which best explain the data will occur frequently in the set, while poorly fitting histories will be sampled infrequently or not at all. Returning a large set of trees allows the analysis to explicitly convey different degrees of certainty about various aspects of linguistic history.

Assessing the strength of evidence that linguistic data provides for a particular tree is achieved using a probabilistic model of language evolution. These models can be rather complex, but it is not necessary to understand all the details of the models to understand the essentials of the calibration procedure. All models provide a way to quantify how well a proposed tree explains the observed linguistic data. This fit is assessed by comparing the lengths of the branches separating languages with the degree of change in the data for those languages. A tree is considered well-fitting if languages which are separated by short branch distances have more similar linguistic data than those separated by longer distances. The simplest models measure this fit using a so-called *strict clock*, where the expected number of data changes (e.g. cognate class replacements) is directly proportional to the elapsed time. The model estimates a single *clock rate*, with units of, for example, 'expected changes per century'. This is usually a poor fit for linguistic reality, where, for example, variation in sociolinguistic context can cause some branches to accumulate more changes per unit time than others. More advanced models using *relaxed clocks* (Drummond et al. 2006) discard the assumption of a constant proportionality between calendar time and language change, instead estimating a *mean clock*

rate and permitting the rate for each branch in the tree to vary around that mean as the data demands. With either strict or relaxed clock models, rates can also be estimated separately for each feature in the linguistic data, allowing, for example, pronouns or words for body parts to evolve more slowly than adjectives at all points across the tree (Chang et al. 2015; Pagel, Atkinson, and Andrew 2007).

In order to estimate either a strict clock rate or the mean rate for a relaxed clock, the model must be provided with information about the age of some part or parts of the tree. This is what is meant by *calibration*. Without calibration, the clock rate and the age of the tree are free to vary arbitrarily. If every branch were made, say, twice as long, while at the same time the clock rate was halved, the *evolutionary time* separating any two languages remains unchanged: 3,000 years at four expected changes per century is indistinguishable from 6,000 years at two expected change per century. In this way, any age for the tree can be made to fit the data as well as any other. However, if we believe that the protolanguage of some subfamily in the tree is, say, between 1,000 and 2,000 years old, then this prior knowledge and the linguistic data for that subfamily taken together are consistent with a limited range of clock rates. If we consider only these compatible clock rates, then branch lengths cannot vary arbitrarily. Instead, some branch lengths can be 'ruled out' because, using the clock rates implied by the calibration, they would convert to lengths of evolutionary time suggesting much more or much less variation in the data than is actually seen. In this way, the information provided by our calibration about one point on the tree propagates to the entire tree and ensures that we end up with a limited range of plausible ages for the entire family.[2]

2.2 Calibrations—what and where?

We have seen above that calibrations in Bayesian phylogenetics have two essential characteristics: first, they are associated with a *point* on the tree, and secondly, they convey our *prior knowledge about the age* of that point on the tree. We now expand upon these two crucial concepts.

2.2.1 Understanding divergence points on trees

In Bayesian phylogenetic analyses, calibrations are placed on one or more individual nodes of the binary tree. Calibrations are almost always placed on interior nodes of the tree (such as those coloured blue in Fig. 1) and typically the points chosen for calibration are the most recent common ancestors (MRCAs) of some well-established family or subfamily. For example, an analysis of Indo-European languages may place a calibration on the MRCA of the Germanic languages. Such a calibration would typically be referred to as 'a calibration on Proto-Germanic', with 'Proto-Germanic' understood to refer only to the instantaneous point on the tree corresponding to that ancestor.

This terminology differs somewhat from what is common in historical linguistics, where protolanguages are thought of not as instantaneous points in time but as long-lived entities, with Proto-Germanic not necessarily any more fleeting than modern German. This makes perfect sense as protolanguages can undergo significant changes while still maintaining their identity as a single language. With this understanding, protolanguages should strictly speaking not be identified with the tree nodes corresponding to MRCAs, but rather with the entire branch leading to the MRCA node. The point on which calibrations are traditionally placed in Bayesian analyses might be more comfortably thought of by historical linguists as, for example, 'late Proto-Germanic'. However, even this might identify a language which persists for some non-trivial length of time, rather than an instantaneous point. The process of linguistic divergence is a gradual one without clear demarcation points specifying times 'before' and 'after' divergence, as indicated in Fig. 2.

An important and generally overlooked difference between probabilistic models of evolution and the traditional usage of the family tree model in linguistics is that the branching events in a probabilistic model do *not* correspond to the occurrence of the linguistic change(s) defining the split. Rather, the branching event represents

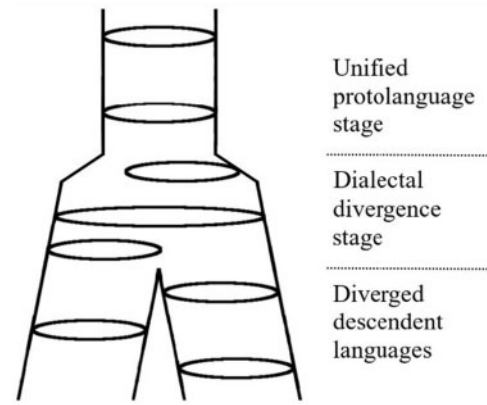

Figure 2. Schematic representation of gradual linguistic divergence, where a single protolanguage diverges into two descendant languages after passing through a protracted dialectal divergence phase where some innovations take place throughout the entire speaker population but others do not.

the time when the evolutionary process becomes *independent* for the two lineages. The first actual linguistic changes in either lineage may not—and typically will not—occur until sometime *after* the branching event.[3] The branching points where independent evolution begins can perhaps best be thought of as corresponding to the point in time where whatever sociolinguistic prerequisites are required for independent changes to occur (e.g. separation/isolation of speaker communities) have been met. This corresponds roughly to the boundary between the protolanguage stage and dialectal divergence stage in Fig. 2.

This raises the question of what sort of evidence should be considered to indicate that a protolanguage has, or has not, 'diverged for calibration purposes'. Given that probabilistic models do not require *any* changes in language data to occur precisely at branching points, calibration times which are argued for on the basis of clear indicators of substantial linguistic divergence, such as the breakdown of mutual intelligibility, are in principle 'too late'. The very earliest indicators of impending divergence, such as the first innovations spreading only through a part of the speaker population, are probably more suitable, though harder to find, evidence. In practice, the amount of uncertainty expressed in a well-justified calibration, combined with the inherent uncertainty in dating that results from using probabilistic models, will probably dominate any uncertainty associated with the difficult issue of deciding which point in the drawn-out divergence process to calibrate against. Regardless, we later advocate an approach to calibration based on using upper and lower bounds, which helps to avoid this very complicated question.

In addition to interior nodes of the tree, calibration points *must* also be placed on the ages of any extinct languages included in the analysis (unless the extinction is very recent relative to the likely age of the family—say within 100 years). Extinct languages correspond to leaf nodes that are closer to the root, such as the one coloured red in Fig. 1. These calibrations (sometimes called 'tip dates') are strictly necessary for the correct interpretation of the extinct language data (see Kitchen et al. (2009) for Bayesian phylogenetic analysis of Semitic languages including several calibrated extinct languages). Here the calibration distribution applies directly to the time of extinction—or, more precisely, to the time at which the linguistic data used for the extinct language was sampled. In practice, the distinction between time of extinction and time of sampling is minimal, as data for extinct languages typically comes from the youngest written source, whose age also provides our estimate of the time of extinction.

2.2.2 Representing beliefs with probability distributions

On the one hand, a formal, statistical approach to dating language families requires that the information we provide about the plausible ages of some point on the tree be put into a form that is precise and explicit. On the other hand, the best understanding of the relevant experts as to the plausible ages often involves substantial uncertainty. This may seem to be a problematic or even impossible combination, but in fact, Bayesian phylogenetics is very well suited to this task. Indeed, one of the great strengths of the Bayesian paradigm is that—unlike all other approaches to tree building—it enables practitioners to specify their degree of certainty in the input to their model and that the outputs, in turn, come with a clear measure of their uncertainty.

This is achieved by representing beliefs about the age of a point in a phylogenetic tree as a *probability distribution*. The *Normal* (or *Gaussian*) distribution, sometimes known as the bell curve, is perhaps the most familiar example of a probability distribution. For present purposes, probability distributions can be thought of as mathematically-defined curves that describe how belief is 'spread out' over a certain range of ages, defined relative to the present (i.e. ages BP[4]). Distributions come in a variety of shapes, and by carefully selecting parameter values the shape of a distribution can be made to represent the views held by the linguistics community about the time at which a protolanguage diverged. For example, when a divergence time is known with a high degree of confidence, it can be represented by a narrow bell curve (Fig. 3a). Alternatively, when there is

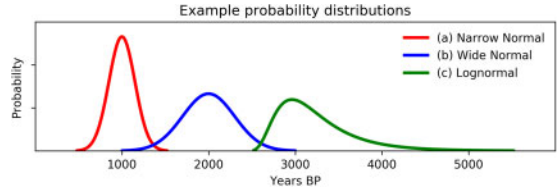

Figure 3. Some examples of probability distribution calibrations. The narrow Normal distribution conveys high confidence in a divergence date of 1000 YBP, with dates from 500 to 1500 YBP being plausible. The wide Normal distribution is a less confident calibration for a divergence date of 2000 YBP, with dates from 1000 to 3000 YBP being plausible. Notice that the peak of the narrow distribution is higher than that of the wide distribution—both distributions have the same total amount of belief to 'spread around', so a narrower (i.e. more certain) distribution necessarily 'piles its belief higher'. The asymmetric lognormal distribution is for a divergence date of 3000 YBP, with slightly younger dates down to 2500 YBP being plausible, but also significantly older dates up to 5000 YBP. Notice that the peaks of the wide Normal and the lognormal distributions have roughly equal heights—despite their different shapes, both represent equally confidence in their respective 'best guesses'.

substantial uncertainty about a time it may be represented by a wider, more diffuse bell curve (Fig. 3b). By carefully choosing the appropriate shape, it is possible to use a probability distribution to convey not only the plausible upper and lower limits on a divergence time, but also which ages within that interval are considered most or least likely. For example, an asymmetric *lognormal distribution* (Fig. 3c) can capture the belief that an event definitely happened before a particular date and most likely not very far in advance of it. Several of the many shapes of distribution which can be used for calibrations are discussed in Supplementary Material.

3. Calibration in practice

3.1 General principles

The first step in deriving a set of calibrations for a Bayesian phylogenetic analysis is to identify the candidate points in the tree on which calibrations could be placed. These are the MRCAs of any subfamilies which are widely accepted as valid genealogical nodes by historical linguists who are experts in the relevant family, as well as any extinct languages included in the analysis, whose calibration is mandatory. Note that nested calibrations are entirely acceptable, for example, it is no problem to calibrate both Proto-Germanic and its subgroup Proto-West-Germanic. The Glottolog database (Hammarström, Forkel, and Haspelmath 2018) is a useful source of conservative genealogical classifications for over 400 language families, with supporting citations to the linguistics literature, and can assist greatly in identifying candidate points. A well-resolved internal structure of a language family is a great help to calibration efforts, and families where extensive borrowing and convergence have made the internal structure so unclear that there are few or even no uncontroversial subfamilies on which calibrations could be placed are not ideal candidates for Bayesian dating. However, so long as recognised subfamilies *do* exist, uncertainty or disagreement as to the nature of their relatedness within the family should not be seen as a reason to avoid Bayesian analysis. On the contrary, Bayesian phylogenetic methods may offer useful insight into precisely this matter.

In general, the more calibrations the better, especially for analyses using a relaxed clock, where the calibrations must inform not only the mean clock rate but also the expected degree of variation about this mean. While calibration of extinct languages is mandatory, it is not necessary to calibrate all other candidate points. Indeed it is rarely possible to construct a high-quality calibration for each candidate point due to a lack of evidence or relevant publications in the literature (especially for less well-studied families) and some candidates may simply not be helpful points to calibrate. Three basic principles for selecting calibration points which we elaborate on below are: (1) prefer calibrations on subfamilies with lots of variation over those with little variation, (2) take care in calibrating subfamilies with unusually faster or slower rates of change, and (3) prefer calibrating subfamilies which are well sampled in the linguistic data to be analysed.

Recall that the mechanism by which calibrations inform estimates of language family age is by helping to estimate the expected rate at which changes in the data occur. Therefore, it is important that there is enough variation in the data for the languages in a calibrated subfamily. Calibration on a small set of languages which have only recently diverged and for which the majority of available data points share the same value has a much reduced ability to inform clock rate estimates. Such a calibration is able to rule out certain rates as 'too fast' to explain the data (if those rates make it highly unlikely that no or few changes would have occurred within the calibrated timespan), but is compatible with arbitrarily slow rates, and thus cannot help to establish an upper bound on family age. Thus, broadly speaking, candidate points which are associated with more variation are better than those associated with less (although see below regarding extreme points). This typically translates into preferring older, larger subfamilies over smaller, younger ones, although this is only a rule of thumb. Calibrating *only* older subfamilies may be counterproductive, though, as older protolanguages are less likely to have written sources or other clear points of evidence, resulting in less certain calibrations.

Care must be taken when calibrating points corresponding to groups of languages which are believed to have undergone a change at unusually fast or slow rates relative to the rest of the family, for example, due to prolonged, intense contact with other families or extreme geographic isolation. If the only calibration in analysis were placed on a subfamily with an unusually fast or slow rate, there is a risk that the language family age will be under- or over-estimated as this uncharacteristic rate is interpreted as the family's norm. This is not to say that such 'extreme points' should not be calibrated. On the contrary, it is important to calibrate them to ensure that they are assigned their believed age, rather than being incorrectly dated on the basis of the more typical rate for the family. Calibrating extreme points is especially important for relaxed clock analyses, as such points are highly informative about the extent of

clock rate variation. However, regardless of whether a strict or relaxed clock is used, every effort should be made to also place calibrations on more typical points to provide counterbalance, ensuring that the family's typical rate is estimated as accurately as possible.

It is rare that a linguistic dataset contains data for every language in a family and careful attention must be paid when a dataset narrowly samples from a large subfamily. As an extreme example, consider an analysis of the Indo-European language family where the only Germanic languages represented in the dataset were Icelandic and Norwegian. A literature review may suggest that Proto-Germanic is around 2,500 years old. However, the MRCA of all the Germanic languages in this analysis (i.e. of Icelandic and Norwegian) is not Proto-Germanic, but in fact the significantly younger Proto-West-Scandinavian. Applying a calibration distribution with a mean of 2,500 years to this point would thus significantly underestimate the rate of change. The addition of data for a single more distantly related Germanic language, say Dutch, would solve this problem and permit the 2,500-year-old calibration to be used; the MRCA of Icelandic, Norwegian, and Dutch is indeed Proto-Germanic. Even Dutch and Icelandic alone would not be problematic, as the requirement is not that all or even most languages in a subfamily be represented in the data, but that as much as possible of the subfamily's genealogical diversity be represented. If a subfamily is only narrowly sampled, such that the MRCA of the sample is much more recent than the MRCA of the entire subfamily, then there are two solutions. Preferably, a calibration should be sought for the MRCA of the actually sampled languages (in this case, Proto-West-Scandinavian). If such a calibration cannot be found, a less informative calibration can be placed on subfamily's 'originate' age. This option is discussed in Supplementary Material (see section *Originate calibrations*).

Once the points to be calibrated have been identified, the next step is to begin research aimed at finding estimated ages for as many of these points as is practical. We propose that a minimum standard for the field be that every calibration used in an analysis where family dating is one of the primary subjects of investigation is backed up by *at least one* citation of a published and peer-reviewed scholarly source, with an explicit and clear statement of the full reasoning behind the complete probability distribution used. The linguistics literature contains many very rough estimates of the ages of various families, which are often stated as loosely as 'about five or six centuries ago' and are presented without any argumentation whatsoever. Even when published by respected authors in well-regarded journals, these claims are not suitable as-is for use in calibration. Instead, calibrations should be based on explicit hypotheses about language history derived from specific, published pieces of data. Such calibrations can be supported or contradicted by future work in the relevant disciplines, and readers and reviewers of analyses using them can make informed decisions about how plausible the calibrations are. These are extremely important criteria for serious research on language divergence timing and they are undermined by vague or implicit arguments and appeals to authority or conventional wisdom. With the emergence of a new generation of quantitative methods for estimating divergence times, we advocate for increased emphasis on explicit argumentation for suggested timings in historical linguistics publications. The rest of this section of the article makes recommendations on how to achieve the 'transparent calibrations' we argue for above.

3.2 Choosing a probability distribution and parameters

We propose the following as a generally applicable process for establishing probabilistic calibrations on protolanguages to be used as a guideline when compiling information and consulting experts. The key is to identify *two* times, based on any of the many possible types of evidence discussed below, to act as upper and lower bounds on the age of the protolanguage. This involves identifying evidence for a time when the language had definitely not begun to diverge (e.g. a written artifact in an early stage of the protolanguage) as well as some independent evidence for a time when the language clearly *had* diverged (e.g. a sound change or a loanword which occurs only in some descendant languages and not others). These two bounds can then be used to determine parameters for a probability distribution such that 95% of the total belief (formally, *probability mass*) is placed between these two bounds. This can be interpreted as informing the model 'I am 95% sure that the true age of this protolanguage lies within this range'. If there is substantial uncertainty in the dating of the evidence, or in whether they reliably indicate bounds on the protolanguage divergence time, it is also possible to put, say, 75% or even 50% of the probability mass between the two endpoints.

Using separate upper and lower bounds to establish the limits of a probability distribution requires more research effort than finding a single point of evidence, but it has significant advantages. Most importantly, a pair of bounds informs not just the position on a timeline

where a distribution should place most of its belief, but also how widely that belief should be spread. When a single item of evidence is used to argue that the mean of a Normal distribution (which positions the peak of the bell curve) should be set to a certain time, it can be very difficult to make a principled choice as to the distribution's standard deviation (which controls the 'spread' or width of the curve). With the bounds-based approach, every parameter of the distribution is derived from the same explicit argument. This approach also allows us to sidestep the very tricky conceptual issue discussed in Section 2.2.1. The time which is actually being calibrated in Bayesian phylogenetics (the time when linguistic evolution becomes independent for two lineages, without any changes necessarily having occurred) is very difficult to reason about and does not leave clear evidence. However, that time is necessarily contained between our two bounds, which correspond to states which *do* leave clear evidence.

When choosing bounds for a calibration, there may be a temptation to 'play it safe' by either defining a very wide calibration interval, or by using a narrow interval but putting only 50% of the probability mass inside it, but either of these provides limited constraint on the range of clock rates, and will only result in a very uncertain estimate of the root age—a principle of 'uncertainty in, uncertainty out' applies. In order to make the calibration as informative as possible, we should endeavour to find the youngest item of evidence for protolanguage unity and the oldest item of evidence for protolanguage disintegration (mathematically speaking, we want the *greatest lower bound* and the *least upper bound*).

Identifying upper and lower bounds for 95% of a distribution's probability does not uniquely specify a distribution—thought must also be given to the *shape* of the distribution (Fig. 3). A sensible and frequently used default option is to use a Normal or Gaussian distribution, with the familiar bell curve shape. Because Normal distributions are symmetric, if the parameters are set so that 95% of the density is between two bounds, the peak of the distribution (or, formally, the *mode*, which conveys our 'best guess' as to the age being calibrated) will be located at the midpoint between those two endpoints. There may be times when we have reason to believe instead that the true age is probably closer to one endpoint than the other. In such situations, an asymmetric probability distribution is appropriate. The lognormal distribution is a common choice, though not the only option. We remind the reader that various probability distributions and situations when they may be useful are described in Supplementary Material.

In some cases (e.g. badly understudied language families), it may be difficult to find both upper and lower bounds for many points in an analysis, especially if the range of candidate calibration points is small. If only an upper bound can be found, then, as a last resort a lower bound of 0 years is always applicable, and similarly the upper bound can always be set to an implausibly high date (say 15,000 years). If using one of these 'extreme bounds', the Normal distribution is not appropriate, as it will assign the highest probability to the midpoint of the interval; whereas if we know only that a divergence is <2,000 years old, it does not necessarily follow that we believe it is more likely to be 1,000 years old than 500 or 1,500 years old. Uniform distributions are a straightforward alternative for this situation, as they do not suppose that any one date within their bounds is more probable than any other. However, uniform distributions have problems of their own and we argue that in situations involving only a single date an exponential distribution may be a more suitable choice (see Supplementary Material for a detailed discussion). Using a single point of evidence to justify a Normal or Lognormal distribution should, as a rule, be avoided and ideally should be done only with an explicit and careful justification for the choice of all the distribution parameters.

An especially challenging calibration situation occurs when there is reason to believe that a language subfamily diverged at approximately the same time as some other event with a known date, but no solid arguments can be made regarding whether divergence occurred before or after the event in question, nor is there any clear idea of how wide an interval may separate the two times. Calibrations based on cases like this must be considered highly speculative; they should be used only as a last resort for studies where the accuracy of the resulting divergence estimate is not critical to the overall research goal. Ideally, they should be used in combination with at least one more principled calibration.

When constructing this kind of calibration, it is sensible to use the known time as the centre of the calibration interval; the real difficulty lies in setting the width of the interval, which must necessarily be arbitrary. Later in this article, we derive three principled calibration distributions for the Uralic language family, the least certain of which has a 95% highest density interval spanning roughly 1,200 years. Guided by the idea that speculative and poorly argued calibrations should not convey more certainty than even the least certain of calibrations which can be justified based on clear evidence, we advise that 1,500 years should be the *minimum* calibration interval width considered for such calibrations.

Of course, there is no good reason to assume that the least certain calibration in our small case study is representative of principled uncertainty about divergence times in general, so this 1,500-year minimum should be critically reassessed over time.

Since the timing of events in the distant past will typically be less certain than more recent events, we reluctantly offer the following as a rule of thumb for rough calibrations derived from a single date: the width of the calibration interval should be equal to half the time of the calibration's midpoint, or 1,500 years, whichever is greater. For example, a linguistic divergence believed to have happened around the time of a historical event dated to 4000 YBP might be calibrated with a Normal distribution centred on 4000 YBP and with 95% of its probability mass between 3,000 and 5000 YBP, defining an interval 2,000 years wide (half of 4,000). However, for a divergence event dated to 2000 YBP, an interval width of 1,000 years (half of 2,000) would be narrower than our recommended minimum width of 1,500 years. Instead, this calibration might be realised with a Normal distribution with 95% of its probability mass between 1250 YBP and 2750 YBP.

3.3 Calibration from written artifacts

Written artifacts are in some ways the ideal source of calibration information for dating language families. Because they are physical objects, their age can often be estimated quite precisely using reliable, objective methods from archeology such as radiocarbon dating—or the age may be clear from text-internal evidence. Furthermore, because the artifact is fundamentally linguistic in nature, there is little difficulty in associating this precise date with a particular language, which can be problematic with non-linguistic archaeological artifacts (discussed below). Unfortunately, written artifacts are quite rare, as most languages do not have, or historically did not have, a written form and were only spoken. Even when written artifacts *do* exist, their application to calibrations is not totally straightforward. The primary difficulty is in deciding how the time at which the artifact was written relates to the time of interest.

To take an illustrative example from a high-profile study, in Bouckaert et al. (2012)'s Indo-European analysis a Normal distribution with mean 1,875.0 and SD 67.0 (95% interval ∼1745–2005 YBP) is used to calibrate Northwest Germanic, on the grounds that the 'Earliest attested North Germanic inscriptions date from 3rd century CE' (∼1710–1810 YBP at time of publication). The rationale for this calibration is unclear, as the inscriptions are in a different language from that whose disintegration is being calibrated: presumably, it is based on an unspoken argument that at least ∼30 and no more than ∼300 years are required to account for known changes between late Northwest Germanic and early North Germanic. This may or may not be defensible based on expert knowledge of Germanic language history, although as long as the argument is unstated it is unlikely to be adequately scrutinised by a reviewer. Regardless of its validity in this case, the approach of narrowly constraining the age of a language based on the age of artifacts containing writing in a descendant language cannot be applied generally, as sufficient knowledge of the tree topology and detailed reconstructions of earlier languages may not be available in many cases. A more limited form of constraint based on descendant languages is possible in some cases and is discussed later.

A more transparent and universally applicable approach to calibration from written artifacts is possible. If an artifact contains, say, written Proto-North-Germanic text (from any stage in the protolanguage's development), then it must necessarily have been written *prior* to the disintegration of Proto-North-Germanic, but in general, there are no guarantees as to whether it was written immediately before disintegration or well in advance of it. As such, the appropriate use of a dateable written artifact is only as a one-sided upper bound on the divergence time of the language in which it is written, that is, the disintegration of the language attested in the written artifact must have happened after the date of the artifact. This should be considered the standard use of written artifacts for calibration, and Normal or log-normal calibrations based only on a single written artifact should be avoided or considered acceptable only if they are accompanied by exceptionally clear and convincing arguments. In general, narrow calibrations can be obtained only by combining a written artifact with some other item of evidence (possibly but not necessarily another written artifact) to provide a complementary bound. Note that this upper bound only approach requires no knowledge of the genealogical structure of the subfamily being calibrated beyond the fact that the language attested in the artifact belongs to the subfamily.

In some circumstances, written artifacts can also be validly used as lower bounds for calibrations. Such an approach is seen in a recent analysis of the Dravidian family (Kolipakam et al. 2018), where the 'South I' subfamily, which includes the Tamil language, is calibrated using a uniform distribution to be at least 2,250 years old on the grounds that Tamil is 'first recorded in a lithic

inscription ... which is dated to c. 254 BCE'. This is a sound approach (the protolanguage from which Tamil descended is necessarily older than Tamil itself) but is possible only because of prior knowledge of Tamil's relatedness to other Dravidian languages. A similar strategy was used by Kitchen et al. (2009), where the divergence of the Semitic languages was calibrated to be older than 4350 YBP, this being the age of 'the earliest known epigraphic evidence of [the Semitic language] Akkadian'. The use of written artifacts to calibrate linguistic analyses is strongly analogous to the use of fossils to calibrate analyses of molecular (e.g. DNA) data and use of lower bound only calibrations as described above is indeed a common practice in the paleontology literature (Barba-Montoya, Dos Reis, and Yang 2017).

In principle, given adequate knowledge of the family structure, a single written artifact may inform two calibrations, by acting as an upper bound for the oldest descendent subfamily and as a lower bound for the youngest subfamily it belongs to. For example, the early North Germanic inscriptions referenced in Bouckaert et al. (2012) could provide both an upper bound on North Germanic divergence and a lower bound on Northwest Germanic—although a tighter upper bound on North Germanic divergence may be possible using the *latest* attested North Germanic inscriptions.

3.4 Calibration from archaeology

Deriving linguistic calibrations from datings in the archaeological literature is often a highly attractive proposition, due to the availability of precise and objective dating techniques. However, except for the special case above where the dated artifacts contain writing, the practice of identifying an archaeologically attested culture (i.e. the group of people who produced a particular set of dateable artifacts) with a linguistic community (i.e. the group of people who spoke a particular protolanguage) is typically neither precise nor objective (but see Rahkonen (2017) for a recent example of using toponymic evidence to attempt a principled identification). We argue that assuming two such groups of people are identical and then using the known age of one to infer the age of the other is methodologically backwards. It would be preferable instead to date both groups by as independent means as possible first and then treat overlapping age estimates as evidence that they may indeed be one and the same. The opposite approach introduces a risk of circularity, where archaeologists and linguists cite each other, each believing that the other field has conclusively dated the community.

Archaeological sources *can* provide viable calibrations in some circumstances. Perhaps the clearest example is the case of the Austronesian language family, whose diversification occurred alongside the geographic expansion of its speakers via a series of 'island hopping' oceanic voyages. Each migration brought a subset of the language community into a new, uninhabited environment, while simultaneously introducing a substantial barrier to linguistic contact with the founding community. That each island was previously uninhabited means the earliest archeological evidence of human presence on an island can be reliably attributed to the speaker group of interest and the barrier to communication means the arrival time actually corresponds very closely to the time of interest, when language evolution becomes independent for two branches of a tree. Thus, it is quite reasonable to use the age of the earliest archaeological finds on an island to directly calibrate the corresponding point on the Austronesian tree. This method was used in a Bayesian analysis of Austronesian (Gray, Drummond, and Greenhill 2009) to place a calibration on the age of 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' (Gray, Atkinson, and Greenhill 2011).

Unfortunately, this pattern of migration into uninhabited territories separated by linguistic barriers is quite rare, making Austronesian an unusually straightforward case for archaeological calibration. Language families in Eurasia, for example, present a much more challenging case. Tens of thousands of years of habitation, migration, language shift, and exchange of material cultures mean that associations between archeological and linguistic cultures are necessarily highly uncertain and, in general, cannot be relied upon for linguistic calibration. Even if it could somehow be determined with certainty that the population responsible for a particular set of archaeological artifacts spoke a language belonging to a particular family, this is only half the battle. It is also important and also difficult to establish how that population's dated arrival to or departure from a particular area relates to the divergence of a particular subfamily.

This is not to say archaeology can contribute nothing of relevance to the calibration of non-Austronesian language families. Directly calibrating language subfamilies from archaeological cultures (as practiced in e.g. Grollemund et al. 2015) should be avoided, but archaeology may still be able to inform calibrations if linguistic evidence suggests that a subfamily diverged before or

after some event which is clearly attested in the archaeological record, for example, the arrival of a particular religion to some part of the world (see Section 3.6 and some calibrations in our Uralic case study for discussion and examples).

Developing a more explicit and detailed account of when and how justifiable linguistic calibrations can be derived from archaeological evidence might be considered a valuable research priority for the field; there are many parts of the world where written artifacts or historical records are sparse or absent and are likely to remain so, but archaeological evidence of human presence can still be found.

3.5 Calibration from historical records

Historical records can often be dated with considerable precision; sometimes even to a particular year, rivaling the precision offered by radiocarbon dating. They may also be more reliably associated with a particular language than non-written archaeological artifacts. These properties make them a useful source of information for linguistic calibrations. However, historical records rarely refer directly to languages or to language divergence and when they do it cannot be taken for granted that the authors have carefully distinguished, for example, languages from ethnicities or dialects from languages. Because of this lack of direct reference, calibrations based on historical records usually require some non-trivial interpretation or hypothesising to relate the content of the records to language history. It is vitally important that this logic is made explicit, as the validity of the calibration cannot be assessed without it. In particular, the nature of this argument determines whether the date(s) in the records constitute an upper or lower bound on the divergence time, or a direct estimate of it. We now consider some concrete examples from the literature.

Analyses of the Indo-European language family provide some examples of challenging calibrations from the historical record. In one analysis of the Indo-European language family (Gray and Atkinson 2003), the Romance languages are calibrated on the grounds that 'the Romance languages probably began to diverge prior to the fall of the Roman Empire' (Gray, Atkinson, and Greenhill 2011). Noting that Dacia was conquered by Rome in 112 AD, while the last Roman troops were withdrawn south of the Danube in 270 AD, Gray and Atkinson 'constrained the age of the node corresponding to the most recent common ancestor of the Romance languages to AD 150–300', or 1700–1850 YBP. A subsequent Indo-European analysis (Bouckaert et al. 2012) refers to the same historical dates of 112 AD and 270 AD, but, without explanation, uses a different calibration with a Normal distribution centred on 2000 YBP and a 95% interval of 1700–2300 YBP.

These calibrations are an example of good practice in one respect: they explicitly state a hypothesis relating the time of Romance divergence to a particular historical event (it happened prior to the fall of the Roman Empire) and presents historical dates claimed to relate to that event (presumably the dates of the Dacian conquest and the troop withdrawal are supposed to bound the 'beginning of the end' for the Romans). This permits the calibration to be critically assessed by linguists and historians. However, the conversion of this hypothesis into a probability distribution appears to have been a cause for confusion. The narrow 150 year wide window, with endpoints very close to the cited historical dates, seems to represent the uncertainty surrounding the timing of the start of the fall of the Roman empire, not of an event which happened 'prior to' that fall—unless the claim is strengthened considerably to 'almost immediately before'. The window used in the later paper is much wider and permits Romance divergence to have begun as much as 500 years earlier than the beginning of the fall. This seems a more realistic calibration, although the 500-year cut-off remains essentially arbitrary. Some degree of arbitrariness is unavoidable in any attempt to turn a claim that a linguistic divergence happened before (or after) some individual point in time into a calibration distribution which is bounded on both sides. Avoiding arbitrary endpoints is a key advantage of deriving calibrations from explicit upper and lower bounds. Another point of evidence, historical or otherwise, establishing a date for clear Romance unity would provide an upper bound and permit a non-arbitrary calibration.

A similar discussion holds for an analysis of Japonic languages (Lee and Hasegawa 2011), in which the divergence of the Kyoto and Tokyo dialects of Japanese are calibrated on the grounds that: 'from historical records, it is clear that … Kyoto has been the political centre of Japan from around 1200 YBP until the Tokugawa military regime … moved the government to the city of Edo (present Tokyo) 407 YBP … following the shift of power, the governing elite, merchants and craftsman settled into the new capital, and their languages … fused with native dialects … to give rise to a distinct dialect'. It is very clear that 407 YBP is an upper bound on the linguistic divergence time and since the date of the move of government is presumably known with high confidence, the primary source of uncertainty comes from the question of how long after the move the divergence began. However, the calibration distribution used is a

wide Normal distribution centred on 407 YBP, which assigns equal belief to dates *before* and after the move. In the absence of an additional point of evidence providing a lower bound, a reversed exponential distribution (see Supplementary Material) parameterised to place 99% of its belief between 0 YBP and 407 YBP would seem a more accurate representation.

Despite their shortcomings, we reiterate that these calibrations should be considered examples of good practice in the way they make an explicit historical argument which is clear enough that their choice of distribution can be critically assessed by third parties, as we have done here. This has not always been the case in the literature. For example, in Gray, Drummond, and Greenhill (2009), the Austronesian subfamily Chamic is calibrated to be between 1800 and 2500 YBP on the grounds that 'speakers of the Chamic language subgroup were described in Chinese records around 1800 years ago and probably entered Vietnam around 2600 years ago' (Gray, Atkinson, and Greenhill 2011). The calibration interval used makes it clear that the date of the Chinese records has been interpreted as an upper bound on the time of Chamic divergence, but this is valid only if the records somehow indicate the existence of a unified Proto-Chamic language. In fact, the records in question appear only to be the first description of the kingdom of Champa and its raiding activities (Thurgood 1999). While it is possible that the entire Chamic family descends from the language of the kingdom of Champa, it is also logically possible that the language spoken in the kingdom of Champa had sister languages which are also ancestors of today's Chamic languages. If there is additional evidence which discounts this possibility it should be explicitly stated as part of the justification for the calibration.

3.6 Calibration from historical linguistic scholarship

While calibration dates for language families without written records are typically sought from extra-linguistic sources such as archaeology and history, there are situations in which historical linguistic analysis in its own right can be brought to bear on the problem, without the risk of circularity.

Perhaps the best-known way that historical linguistics can yield absolute timing information is via *linguistic paleontology*— the reconstruction of vocabulary relating to technological or cultural innovations such as agriculture or religion, as well as features of the natural environment. It may be possible to date these phenomena reliably by archaeological means or through historical records. If a term corresponding to one of these features can be reconstructed to a particular node in a family tree, this suggests that the feature was present prior to the disintegration of the node in question. This can, however, be a hazardous assumption. Reuse of old words in new contexts can lead to wholesale semantic change across a family: for example, a term for a domesticated species may have referred previously to a wild animal; a term for a technology item may have referred to similar but distinct technology items earlier. These concerns are not hypothetical, and there are established cases of reconstructed protolanguages containing words for species which are not archaeologically attested in the relevant speaker areas, while lacking words for species which are attested. For discussion of these problems in the application of linguistic paleontology to Indo-European chronology, see, for example, Krell (1998), Heggarty (2006), and Heggarty and Renfrew (2014). Finally, the process of etymological nativisation,[5] where previous sound changes are applied retrospectively to borrowed words (Aikio 2007), can make words appear to be reconstructable to a time earlier than their actual arrival. For all these reasons, linguistic paleontology must be considered a risky undertaking, and calibrations from alternative methods should be preferred wherever possible.

A different approach to absolute dating via historical linguistics is based on identifying linguistic borrowings from a source which has a long written history or clearer historical record than the recipient language. If borrowings can be identified between languages in the family being calibrated and another family whose chronology is well understood, then known timings for the 'donor' family can imply timings for the recipient. This is because borrowing typically occurs only between contemporary languages. However, loanwords *can* be mediated to one language through another, for example, Romani did not acquire loanwords directly from Arabic but via Persian or Armenian (Matras 2002). Such indirect borrowings could in principle cause 'borrowing through time', resulting in an inaccurate calibration. Thus, calibrations should only be based on loanwords for which such scenarios can be convincingly ruled out or the spread can be argued to have occurred very rapidly. Special care must be taken not to derive calibrations from so-called wandering words, often concentrated in the semantic fields of, for example, technological innovations (Haynie et al. 2014), which are geographically widely diffused loanwords with parallels in various languages and language families, resulting in the original donor language being difficult to identify (Campbell and Mixco 2007). In addition to issues of mediated

borrowing, it is also important to pay close attention to the process by which the timing for the donor family was established. Regardless of the quality of the scholarship identifying the borrowings, if the donor family's tree has been timed by imprecise or unreliable means, then calibrations derived from it will be imprecise or unreliable in kind. Thus, this approach works best when the donor language is attested in dateable written artifacts or when the contact between the two families is known to be associated with a historical event that can be dated by other means. This approach to calibration has not yet been utilised very widely, but we believe it can be quite powerful. Our calibration of the Uralic language family presented below is achieved primarily by this kind of argument.

4. Uralic case study

The Uralic language family is a relatively well-studied family whose genealogical unity is firmly established (for a condensed review of family models, see Syrjänen et al. (2013)). It consists of ∼40 languages spoken in a large area in North-Eastern Europe and Western Siberia, extending from the Eastern side of the Ural Mountains to Fennoscandia in the West, with the exception of Hungarian which is located in central Europe. The first studies considering the relatedness of Uralic languages were conducted in the 18th century and later, with the emergence of the historical-comparative method, the first Uralic family trees were drawn in the 19th century (Donner 1879). The most recent approaches include phylogenetic research on evolution of the family (Honkola et al. 2013; Lehtinen et al. 2014; Syrjänen et al. 2013).

The age and subgrouping of the family have been systematically studied for over a century, but a consensus on the age of the family has yet to emerge. Earlier estimates gave it an age of around 6,000 years or even older (Janhunen 2000; Korhonen 1981), while the 21st century has seen a trend towards revising these estimates downwards, to as little as 4,000 years (Häkkinen 2009; Heikkilä 2014; Kallio 2006b). In general, considerations of Uralic sound history (Janhunen 2000) the timing of loanwords from Indo-European (Joki 1973; Koivulehto 2001), linguistic paleontology (Häkkinen 2009), and results from archaeology (Parpola 2012), have all been used to inform Uralic age estimates. To date, there has only been a single attempt to estimate the age of the family using the Bayesian quantitative framework outlined here (Honkola et al. 2013), arriving at an age estimate for Proto-Uralic of 5300 YBP (95% interval 3330–7500 YBP).

In this section, we present an evaluation and synthesis of information concerning the timing of proto-language divergences published by various experts of Uralic historical linguistics. We aim to increase the accessibility of this large body of scholarship to those interested in dating Proto-Uralic divergence by transparently discussing both suitable and unsuitable lines of evidence for establishing reliable calibrations. The set of calibrations we eventually derive is a substantial improvement upon those used by Honkola et al. (2013). We recommend their use in future Bayesian analyses of the Uralic languages and indeed in any research requiring precise and principled information on Uralic divergence times.

The new calibrations also serve as an illustrative 'worked example' for the principles established above and as a promotion of the philosophy of publishing the explicit and detailed process of formulating calibrations. Such 'transparent calibrations' can easily be reviewed and improved by expert linguists, which would be a very welcome contribution to the newly developing field of language family dating. Publishing them independently of any analysis encourages the reuse of a standard set of calibrations for a given language family, for analyses of different kinds of data (e.g. lexical and typological) or of the same data using different evolutionary models. It also means the calibrations can be evaluated as good or poor matches to the current state of knowledge on their own merits, without any consideration being given to divergence datings or tree topologies which may later be derived from them.

The establishment of calibrations for the Uralic language family is especially challenging due to the sparsity of written records (the oldest sources of written Uralic language date to no later than 800 YBP (Laakso 1991) and the large number of archaeologically attested cultures present in the speaker areas, none of which have been lastingly accepted as being associated with any particular Uralic speaking population (see e.g. Korhonen (1984) for a formerly influential theory linking Proto-Uralic to the Lyalovo culture and Lang (2018) for more recent proposals). However, Uralic languages have borrowed a notable amount of linguistic material from non-Uralic languages belonging to families for which written material and historical documentation reach further back in time than they do for Uralic. Careful analysis of sound changes in the receiving languages and the donor language families (especially Indo-European) allows certain stages in the diversification of the Uralic languages to be associated with stages in the diversification of other families for which absolute times are known with some confidence. Thus, even in the absence of written or

historical records, the Uralic family can be calibrated based on loanword analysis.

As discussed in Section 3, calibration points must be placed on the most recent common ancestor of established subfamilies of the family being dated. Thanks to careful study of the Uralic family with the comparative method of historical linguistics, the internal structure of the family is relatively well understood and there is strong agreement on the identity of several subfamilies. These include the Finnic languages, the Mordvinic languages, the Permic languages, the Saamic languages, the Samoyedic languages, and the Ugric languages. A number of higher-level subfamilies of Uralic have also been proposed, but their validity is not as well-established as any of the subfamilies listed above. One such subfamily, Finno-Saamic, was calibrated in the earlier Bayesian analysis of Uralic, with a 95% age interval of 2000–3000 YBP (Honkola et al. 2013). However, since the Finno-Saamic subfamily is not well-established (Saarikivi and Grünthal 2005), and some phonological studies suggest a single joint protolanguage for the Finnic, Saamic, and Mordvinic languages (Häkkinen 2007), we opt not to consider Finno-Saamic calibration here and limit our focus to undisputed subfamilies.

4.1 Calibration of the Finnic languages

We derive an upper bound for the divergence of Finnic from analysis of Indo-European loanwords. Finnic contains Proto-Scandinavian borrowings from the time of Early Runic, which is attested in dateable written records. The distribution of these loans across Finnic languages and consideration of which sound changes they have participated in suggests that Late Proto-Finnic was still in a unified state around 1500–1800 YBP (Kallio 2015). Similarly, early Slavic borrowings into Finnic suggest that Finnic remained unified around 1200–1600 YBP. The acquisition of these loanwords is estimated from an archaeological dating of the Northern spread of Slavic (Kallio 2006a). These two windows, based on independent evidence, overlap between 1500–1600 YBP, so we take 1550 YBP as a point in time where we can be confident but not certain that Late Proto-Finnic was in a unified state. The slightly older Proto-Scandinavian times based on written records must be considered more reliable than the younger Slavic times based on archaeology and there has also been a suggestion that an early sound change marking the very beginning of minor dialectal differentiation within Proto-Finnic can be dated as early as 1850 YBP (Heikkilä 2014). Therefore, we take 1850 YBP as a highly confident upper bound.

Approximate time for the beginning of Proto-Finnic divergence can be estimated by considering the latest known loanwords which were borrowed into a unified Proto-Finnic, with subsequent loanwords failing to spread across the entire linguistic continuum. For example, some of the last known borrowings into a unified Proto-Finnic are Slavic Christian terms, which can be roughly dated to 1200 YBP. This dating is based on the fact that these terms were borrowed into Slavic from Old High German (Heikkilä 2019; Kallio 2006a; Kiparsky 1975). Since the earliest attestation of Old High German is from 789 AD (Bergmann et al. 2007), these Christian terms cannot have been acquired into Finnic any earlier. Archaeology provides additional evidence for the arrival of Christianity to the Proto-Finnic speaker area at this time, since a Slavic movement into the area is estimated to have started in the eighth century (Kallio 2006a; Purhonen 1998), and religious artifacts, for example, cross-shaped jewelry, tie the archaeological material to the Slavic branch of Christianity specifically (Heikkilä 2019).

Another late borrowing into unified Proto-Finnic is an ethnonymic term for Swedes, borrowed from Scandinavian and whose distribution and phonology suggest that it was in turn borrowed from Finnic into East Slavic. This chain of borrowing suggests a timing around the beginning of the Viking Age (1250 YBP) because later trade routes enabled direct contact between the Vikings and Slavs (Schalin 2014). The Christian terminology from Slavic and the Scandinavian ethnonym together motivate 1225 YBP as a 'best guess' as to the time at which divergence began. Note that our use of archaeological evidence here does not contradict our advice in Section 3.4; we are not making assumptions about which language was spoken by the people producing certain artifacts, but rather about their religious beliefs, which we argue can be significantly more reliably inferred from artifacts.

A clear lower bound for the divergence of Finnic is difficult to find. Many scholars have expressed certainty that the divergence was complete by 1000 AD (Heikkilä 2014; Janhunen 2009; Kallio 2014), but this is apparently based on little more than the notion that Finnic divergence was driven by intensive contact with both Slavic and Viking populations, beginning some centuries before this date (and in the Viking case ending very shortly after it). Further, reconstruction of Finnic speaker areas at around 1000 AD based on various types of evidence (Frog and Saarikivi 2015) suggests that by this time Finnic languages were spoken on both sides of the Gulf of Finland. Increasing Viking and Hanseatic League influence in these waters are believed to have

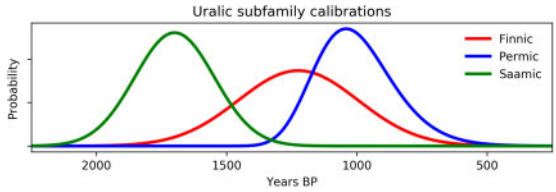

Figure 4. Probability distributions chosen for our three calibration points, on the Finnic, Permic, and Saamic languages. The Finnic and Saamic calibrations are symmetric Normal distribution, while Permic is a 'reversed lognormal' distribution with a longer tail towards younger ages. Note that the wider 95% interval for Permic means that the probability of its mode (roughly 1,250) is lower than that for Finnic.

caused diminishing contact between Finnic speakers on opposite sides of the Gulf, contributing to the separation of the Finnic dialect continuum, supporting the idea this process was underway by 1000 AD. Thus we might take 1000 YBP as a lower bound, although the supporting arguments cannot be considered highly certain. Definitive evidence of Finnic divergence in the form of written records is attested in 1329 AD, by which time a sound change had rendered South-Western dialects of Finnish clearly different from other Finnish dialects and from Estonian (Heikkilä 2016). Since major diversification had happened before this point, we take 670 YBP as a highly confident lower bound.

Thus, we have the best guess at the time of Finnic divergence of 1225 YBP, with upper and lower bounds of very high confidence of 670 YBP and 1850 YBP, and tighter bounds of moderate confidence of 1000 YBP and 1550 YBP. Because the best guess is very close (within 50 years) to the midpoints of both sets of bounds, it is sensible to represent this calibration with a Normal distribution. Calibrating in units of millennia, we set a mean of 1.225 and an SD of 0.23. This corresponds to a mean and mode divergence time of 1230 YBP, with 99% certainty that the age is between 670 YBP and 1850 YBP, and 76% certainty that the age is between 1000 YBP and 1550 YBP (Fig. 4).

4.2 Calibration of the Mordvinic languages

The relatedness of the two Mordvinic languages, Erzya and Moksha, is undisputed by Uralic scholars and this makes their protolanguage a candidate calibration point. However, it is not clear that calibrating Proto-Mordvinic would prove informative. While Erzya and Moksha are today recognised as two distinct but closely related languages, they are sufficiently similar that they have previously been considered as two dialectal variants of the one language by some scholars. This inconsistent view in research history is likely due to repeated migrations and changing contact situations (driven by e.g. the 13th century Mongol-Tatar conquest and later Russian colonisation) having induced periods of linguistic convergence and divergence between the two, stemming from various drivers for mutual contacts (Feoktistov and Saarinen 2005). In light of the similarity between the varieties, it seems probable that there is relatively little variation between Erzya and Moksha in basic vocabulary cognacy, the kind of data most often used for Bayesian dating. As explained in Section 3, small subfamilies without much diversity in the data do not make effective calibration points. Mordvinic is currently not a compelling calibration point for dating Proto-Uralic divergence, so we offer no calibration here.

4.3 Calibration of the Permic languages

Around twenty Volga Bulgar loanwords can be reliably argued to have been borrowed into Proto-Permic, the ancestor of Komi and Udmurt (Rédei and Róna-Tas 1983), establishing that the protolanguage was still unified at the time of the earliest Volga Bulgar contacts. Since Volga Bulgar influence is believed to have begun at the end of the 8th century (Rédei and Róna-Tas 1983), we use 1200 YBP as an upper bound for the calibration.

A lower bound can be based on the lack of Tatar loanwords in Komi, whereas there is a notable number of them in Udmurt—thousands in some dialects (Csúcs 1990). This suggests that the Proto-Permic language had diverged and the Komi speakers had moved North prior to the start of Tatar linguistic influence in the Volga area. The ethnogenesis of the Tatars and their relationship to the Kipchak Turkic groups is complex and difficult to pin down in time, but their emergence has been tied to the development of the Kazan Khanate between 1437 and 1445 AD (Rorlich 1986), that is, around 600 YBP. This is also the midpoint of Csúcs's (1990) estimate of highly intensive Udmurt-Tatar contact between the 14th and 15th centuries, that is, 1300–1500 AD or 500–700 YBP. Thus, we use 600 YBP as a lower bound for the calibration.

The divergence of Permic is often associated in the literature with the establishment of the Volga Bulgar state and the fact that fewer Volga Bulgar words were borrowed into Proto-Permic than into Proto-Udmurt suggests that divergence happened relatively early in the period of Volga Bulgar influence. We, therefore, wish to assign higher confidence to dates closer to the upper bound of our interval than to the lower bound. The Volga Bulgar state was established around 1200 YBP

(Rédei and Róna-Tas 1983) and collapsed following the Mongol invasion around 800 YBP (Agyagási 2012). Thus, we set the mode of our calibration distribution to 1100 YBP, the midpoint of the first half of the state's duration.

The target times for our calibration distribution are a 600 YBP lower bound, a 1200 YBP upper bound, and a most probable date of 1100 YBP. This is a strongly asymmetric calibration—the best guess is much closer to one bound than the other. When the mode of the distribution needs to be closer to the lower bound than the upper bound, lognormal distributions can readily accommodate this, but the opposite case, as seen here, cannot be captured with standard distributions. A uniform distribution between 600 and 1200 YBP is a rough approximation which can be specified in any phylogenetic software, but this discards the information that older ages are more likely. This situation demonstrates the real need for more flexible calibration options in software. Our Permic calibration could be better captured by a triangle distribution or a truncated normal distribution, both of which are uncommon but established probability distributions. In the supplementary material we define a 'reversed lognormal' distribution, which we believe is the most suitable representation. Such a distribution with an offset of 1.85, a mean of −0.18, and an SD of 0.18, allocates 95% of its belief to an interval from 661 to 1263 YBP, with a most likely age of 1041 YBP (Fig. 4).

In our earlier paper (Honkola et al. 2013), the divergence of the Permic languages was calibrated to between 1100 and 1300 YBP but this very narrow interval is an interpretation based on the uneven distribution of Volga Bulgar loanwords in the Permic languages suggesting that the two languages had diverged by the time of the Volga Bulgar state: had they not, Volga Bulgar borrowings into a unified Proto-Permic would be retained with roughly equal frequency in both languages. The argument only establishes that the Permic languages must have diverged *before* contact began with the Volga Bulgar language, and is perfectly compatible with arbitrarily earlier divergences. However, a reconsideration of literature provides a revised timing described above.

4.4 Calibration of the Saamic languages

An upper bound for the divergence of Proto-Saamic can be derived from the study of the Great Saami Vowel Shift: a major reorganisation of the Proto-Saamic vowel system which lead to a significantly different system from that reconstructed for Pre-Saamic and its predecessor. As the shifting phenomenon is extremely regular, it is postulated to have happened among a compact speaker community without significant linguistic differentiation (Aikio 2012). One of the oldest and most prominent sound changes in the Great Saami Vowel Shift has been dated to ca. 2000 YBP, based on consideration of which Germanic loanwords did and did not participate in the change, indicating Saamic unity at that time (Aikio 2006; Heikkilä 2011). Thus we take 2000 YBP as the upper bound for the calibration.

Like Proto-Finnic, the divergence of Proto-Saamic can be tied to an absolute chronology using the stratification of Proto-Scandinavian loanwords, acquired after the Great Saami Vowel Shift during the timeframe of Early Runic (200–500 AD) (Heikkilä 2014). During this period, loanwords were borrowed unevenly into a dialectically diversified Proto-Saami, with more loanwords and borrowed phonological features present in Western than Eastern Saamic (Aikio 2012). These borrowings, together with the distribution of Saamic-specific phonological innovations visible in the borrowings, give evidence that by 1500–1300 YBP, Proto-Saamic was a diverse continuum of dialects, already employing different patterns of etymological nativisation (Aikio 2012). Thus, we use 1400 YBP as a lower bound for the calibration.

As we do not have more specific information based on which we would change the most probable time of divergence towards one or the other end of the 1400–2000 YBP calibration interval, we opt to use a normal distribution centred on 1700 YBP. Thus, specifying the calibration in units of millennia, we use a Normal distribution with a mean of 1.70 and a standard deviation of 0.153, which places 95% of its belief precisely between our two bounds (Fig. 4).

4.5 Calibration of the Samoyedic languages

Divergence estimates of Samoyedic languages are in many cases based on historical data of movements of Turkic tribes in the assumed Proto-Samoyed homeland (Hajdú 1975) and on Yeniseian loanwords in Samoyed (Janhunen 1998; Korhonen and Kulonen 1991). However, these arguments depend on extensive unsupported assumptions about the relationship between various poorly documented tribes and the speakers of certain languages and use unsubstantiated dates. At present, there do not seem to be citable, defensible points of evidence which can be used to reliably calibrate the Samoyedic languages. The literature does contain suggestions of both Turkic (Donner 1924) and Yeniseian loanwords in Samoyedic languages (Hajdú 1953), which leaves the possibility of future calibrations based on more thorough historical linguistic scholarship which is

still is needed to untangle the history of the Samoyedic subgroup. In the earlier Uralic phylogeny of Honkola et al. (2013), the divergence of Samoyedic was calibrated with a uniform distribution between 2000 and 2200 YBP based on the Turkic tribal movements and Yeniseian loanwords mentioned above, but such a narrow interval cannot be justified by such speculative and uncertain evidence.

4.6 Calibration of the Ugric languages

The Ugric subfamily, consisting of Hungarian and the Ob-Ugric languages, Khanty and Mansi, is generally endorsed as a valid genealogical node in Uralic linguistics (Abondolo 1998; Hajdú 1952; Honti 1979, 1997) and has traditionally been considered a relatively old subfamily, based on both interpretations of archaeological evidence (Hajdú 1985; Rédei 1986) and on linguistic evidence suggesting its separation early in the history of the Uralic family. However, there are also scholars sceptical of Ugric unity, emphasising areal contacts which make it difficult to separate inherited features from dispersed ones, especially regarding Ob-Ugric (Aikio 2018; Bakró-Nagy 2013; Gulya 1977). This tension in the literature complicates the task of calibrating Ugric.

Estimates of the absolute age of the divergence of Proto-Ugric have generally been in the vicinity of 2500–3500 YBP (Abondolo 1998), but these are not based on sufficiently reliable arguments to be used as calibrations. Consideration of Iranian borrowings into Proto-Ugric would appear to be the strongest source of timing information for calibrating the subfamily; however, the number of such borrowings is very low and recent research has changed the views on the possible timings of acquisition for some (Häkkinen 2009). Thus, even these lines of evidence must be considered as uncertain. The splitting of Ugric to Hungarian and Ob-Ugric has been dated to older than 3500 YBP based on Iranian periodisation (Korenchy 1972), and this timing is corroborated by changes in the sibilant system shared by all Ugric languages (and also Samoyedic); Iranian borrowings participating in these changes correspond to attested Old Iranian languages, dated to ∼3500 YBP. However, this latter piece of evidence is a new and undeveloped line of study (pers. comm. Holopainen). Possible Turkic borrowings common to all Ugric languages have also been proposed but their status remains highly uncertain (Róna-Tas 1988), so these words cannot provide any additional evidence.

The only indisputable source of timing information for Ugric is the historical attestation of the Hungarian migration to the Carpathian basin, which was settled during the 9th century AD (Fodor et al. 2009). The first known text written in Hungarian dates somewhat later to 1192–1195 AD (Benkő 1980). These dates could be used to provide a lower bound on the originate of Hungarian, but this is unlikely to be a very informative calibration. Old Iranian borrowings acquired separately into Hungarian potentially providing information on the split of Ugric cannot be clearly stratified and dated (Kulonen 1993). In light of the many uncertainties surrounding Ugric timing, and apparently increasing doubt about the group's internal structure or even genealogical validity, we opt not to calibrate the Ugric languages here.

4.7 Summary of Uralic calibrations

After extensive reading of the literature on the Uralic languages and careful consideration of how various points of evidence can be interpreted as upper and lower bounds for divergence events, we have established three calibrations for Bayesian phylogenetic dating of the Uralic language family (Fig. 4). Our previous calibration in Honkola et al. (2013) on the Permic languages was refined substantially, while previous calibrations on Finno-Saamic and Samoyedic were discarded. New calibrations were determined for the Finnic and Saamic subfamilies. Each of the three calibrations proposed here is based on explicit arguments from at least two distinct points of evidence, which determine all parameters of the probability distributions used.

The most promising prospects for additional Uralic calibrations to further improve timing estimates for the family would appear to be better understandings of the history of the Ugric and Samoyedic languages. The presence of Iranian and Turkic loanwords, respectively, in these subfamilies may permit calibrations to be made after further studies on historical phonology and loanword strata. We strongly encourage interested Ugric and Samoyedic specialists to consider these problems.

These three Uralic calibrations will be used in an upcoming Bayesian phylogenetic analysis of the Uralic family, using the recently released UraLex database of basic vocabulary cognacy relationships (Syrjänen et al. 2018) and taking advantage of new modelling developments since Honkola et al.'s (2013) study. However, in order to demonstrate here the substantial impact which critically re-evaluating calibrations can have on analysis, we report the results of repeating the Honkola study with no changes made other than the substitution of our new calibrations (Fig. 5). This caused a considerable change in the posterior distribution of the age of proto-

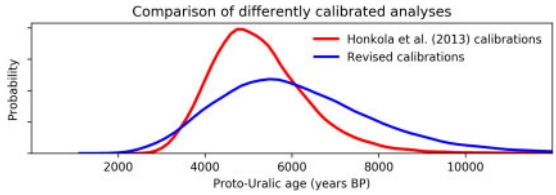

Figure 5. Posterior probability distributions for time of Proto-Uralic divergence based on data and modelling from Honkola et al. (2013), with the original calibrations from that study and those derived here. Note the substantially increased posterior uncertainty.

Uralic: the mean age estimate is increased by 850 years, or by >15% of the previous estimate, while the 95% HPD interval becomes almost 3,000 years wider. This dramatic change in HPD interval width is likely close to a worst-case scenario, as some of the old calibrations used uniform distributions only 200 years wide, conveying very high certainty, whereas the new calibrations, by design, capture the full extent of the uncertainty which must necessarily surround the timing of language divergences from thousands of years ago. We hope the substantial change in the result that comes from this change in calibration highlights the extreme importance of defining calibration distributions in a careful and principled way.

5. Conclusion

The new wave of quantitative historical linguistic analyses striving to rigorously date language families employ a methodology which explicitly accounts for variation in the rate of language evolution throughout time and across the lexicon, making them a substantial advance over previous methods. However, the credibility of the timing estimates these methods yield is entirely dependent upon the credibility of the calibrations that they need as input.

If the modern approaches are to become widely accepted and their findings were taken seriously, it is essential that practitioners take the calibration process seriously, and hold themselves to a high academic standard whereby every calibration is carefully and explicitly justified. Reviewers of articles reporting calibrated Bayesian phylogenies must pay close attention to the calibrations and the arguments underlying them. This requires that the calibration process is widely understood in detail, even by those unfamiliar with the details of, for example, substitution models or clock models, and that the field engages in an open and ongoing discussion about acceptable standards of calibration.

It is also desirable that historical linguists who do not use these new methods themselves nevertheless understand how historical linguistic research is increasingly being used by interdisciplinary researchers to derive calibration intervals. Clear and concise explanations of how certain lines of evidence bearing on the absolute timing of language divergences are comparatively rare in the historical linguistics literature, but the demand for citeable explanations is only likely to increase in the coming years.

To encourage all of these developments, we have endeavoured to explain the calibration procedure in a widely accessible manner, while still paying attention to important technical considerations which have been previously overlooked. We have tried to constructively identify shortcomings in previously published calibrations, including those published by authors of this article. Finally, we have proposed a procedure, along with a demonstrative 'worked example', by which high quality and defensible sets of calibrations may be achieved.

By advocating such a high standard of scholarship with regard to calibrations, we do not wish to discourage the application of Bayesian dating methods to those language families where the unavailability of high-quality evidence makes meeting this standard difficult or perhaps even impossible. Ultimately, approaching the problem of language divergence dating with a framework of explicit statistical models of language change should be considered a substantial advance upon less principled approaches, even in cases where the choice of calibrations cannot be strongly defended. In some cases, reduced rigour may be unavoidable, but only by having clearly defined standards of rigour for the best of cases is it possible for the field to recognise when difficult cases are falling short and by how much, so that the results can be interpreted accordingly.

Finally, we acknowledge that many of the published calibrations we have scrutinised were part of pioneering works, whose authors cannot reasonably have been expected to have anticipated and met future standards of best practice; this article was written to improve the field's future and not to detract from its past.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online.

Acknowledgements

We would like to thank the following Finno-Ugrists for sharing their expertise and references related to the

divergence of different subgroups of the Uralic family: Sirkka Saarinen for Permic and Mordvinic, Sampsa Holopainen for Ugric, Petri Kallio and Santeri Junttila for Finnic, and Mikko Heikkilä for both Finnic and Saamic. We also thank Rogier Blokland for feedback on the entire paper. We thank Niklas Wahlberg for insight into the process by which dates fossils are used to calibrate biological phylogenies and for feedback on an early version of the paper.

Funding

This work was funded (L.M., T.H. and O.V.) by a grant from Kone Foundation (OV-AikaSyyni).

Authors' contributions

L.M. and O.V. planned the article. L.M. and M.D. contributed the technical descriptions in Section 2. All authors contributed to the guidelines and practices recommended in Section 3. M.d.H. and T.H. reviewed historical linguistics literature and consulted with linguists to formulate the calibrations in Section 4. All authors took part in writing the introduction and conclusion and in commenting on all sections of the manuscript.

Notes

1. The Monte Carlo Markov Chain (MCMC) algorithm is a method used in Bayesian phylogenetic analysis to approximate a probability distribution over the space of all possible tree histories. See Dunn (2015) for a brief overview in the context of language phylogeny, or for example, Gilks et al. (1995) for a technical treatment.
2. Bayesian analyses usually also include a component called a 'tree prior', which is another mechanism by which timing information from a calibration point can be propagated to the rest of the tree. We do not discuss this here to keep the explanation easy to follow, and because it is expected that the linguistic data inform the age of the tree (via the clock rate) much more strongly than the tree prior. For a discussion of the influences of tree priors on language dating, see Ritchie and Ho (2019).
3. The actual times at which concrete changes in the data occur are not usually represented at all in these probabilistic models. Rather, the probability calculations integrate over any possible number of changes occuring at all possible points between two nodes.
4. Dating families in 'years before present' has the obvious problem that the age of anything in YBP is not constant as 'the present' moves forward. This problem has been solved in archaeology by explicitly defining 'present' as 1950 AD. This convention does not appear to have been explicitly either copied or rejected in computational historical linguistics thus far, presumably because the typical uncertainty in Bayesian language family datings is much larger than that in radiocarbon dating (often over 1,000 years), making the issue less important. In this article, we take the present to be 2000 AD.
5. Also known variously as historical analogising, correspondence mimicry, borrowing routine, or reclassification.

References

Abondolo, D. (1998) 'Hungarian', in D. Abondolo (ed.) *The Uralic Languages*, pp. 428–56. London: Routledge.

Agyagási, K. (2012) 'Language Contact in the Volga-Kama Area', *Studia Uralo-Altaica*, 49: 21–37.

Aikio, A. (2006) 'On Germanic-Saami Contacts and Saami Prehistory', *Journal de La Société Finno-Ougrienne*, 91: 9–55.

—— (2007) 'Etymological Nativization of Loanwords: A Case Study of Saami and Finnish', in Toivonen I.and Nelson D. (eds.) *Saami Linguistics*, pp. 17–52. Amsterdam & Philadelphia: John Benjamins.

—— (2012) 'An Essay on Saami Ethnolinguistic Prehistory', in R. Grünthal and P. Kallio (eds.) *A Linguistic Map of Prehistoric Northern Europe*, pp. 63–117. Helsinki, Finland: Suomalais-Ugrilaisen Seuran Toimituksia.

—— (2018) 'Notes on the Development of Some Consonant Clusters in Hungarian. *Peri Orthotētos Etymōn – Uusiutuva Uralilainen Etymologia*', *Uralica Helsingiensia*, 11: 91–133.

Bakró-Nagy, M. et al. (2013) 'Uráli Etimológiák a Világhálón', pp. 13–22. *Presented at the Obi-ugor és Szamojéd Kutatások, Magyar Ostörténet Hajdú Péter és Schmidt Éva emlékkonferencia 2012, Pécs*.

Barba-Montoya, J., Dos Reis, M., and Yang, Z. (2017) 'Comparison of Different Strategies for Using Fossil Calibrations to Generate the Time Prior in Bayesian Molecular Clock Dating', *Molecular Phylogenetics and Evolution*, 114: 386–400.

Benkő, L. (1980) *Az Árpád-kor magyar nyelvű szövegemlékei*. Budapest: Akadémiai Kiadó.

Bergmann, R., Moulin, C., and Ruge, N. (2007) *Alt- und Mittelhochdeutsch: Arbeitsbuch zur Grammatik der älteren deutschen Sprachstufen und zur deutschen Sprachgeschichte (7., überarbeitete. Auflage)*. Göttingen: Vandenhoeck & Ruprecht.

Bouckaert, R., Bowern, C., and Atkinson, Q. D. (2018) 'The Origin and Expansion of Pama–Nyungan Languages across Australia', *Nature Ecology & Evolution*, 2: 741–9.

—— et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337/6097: 957–60.

Buck, C. E., Cavanagh, W. G., and Litton, C. D. (1996) *Bayesian Approach to Interpreting Archaeological Data*. Chester: John Wiley & Sons Ltd.

Campbell, L., and Mixco, M. J. (2007) *A Glossary of Historical Linguistics*. Edinburgh, Scotland: Edinburgh University Press.

Chang, W. et al. (2015) 'Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis', *Language*, 91/1: 194–244.

Csúcs, S. (1990) 'A Votják Nyelv Orosz Jövevényszavai 1', *Nyelvtudományi Közlemények*, 72: 323–62.

Donner, K. (1924) 'Zu Den Ältesten Berührungen Zwischen Samojeden Und Türken', *Journal de La Société Finno-Ougrienne*, 40: 3–42.

Donner, O. (1879) *Die gegenseitige Verwandtschaft der finnisch-ugrischen Sprachen*. Hfors.

Drummond, A. J. et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4/5: e88.

Dunn, M. (2015) 'Language Phylogenies', in C. Bowern and B. Evans (eds.) *The Routledge Handbook of Historical Linguistics*, pp. 190–211. London: Routledge.

Felsenstein, J. (2004) *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc.

Feoktistov, A., and Saarinen, S. (2005) *MokšAmordvan Murteet*. Helsinki: Suomalais-ugrilainen Seura.

Fodor, I., Romsics, I., and Csepregi, M. (2009) *Ostörténet és honfoglalás. Magyarország története*, 1. Budapest: Kossuth Kiadó.

Forster, P., and Renfrew, C., eds. (2006) *Phylogenetic Methods and the Prehistory of Languages*. Cambridge: McDonald Institute for Archaeological Research.

Frog, M., and Saarikivi, J. (2015) 'De Situ Linguarum Fennicarum Aetatis Ferreae', *RMN Newsletter*, 9: 64–115.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1995) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Grollemund, R. et al. (2015) 'Bantu Expansion Shows That Habitat Alters the Route and Pace of Human Dispersals', *Proceedings of the National Academy of Sciences of the United States of America*, 112/43: 13296–301.

Gray, R. D., and Atkinson, Q. D. (2003) 'Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin', *Nature*, 426/6965: 435–9.

Gray, R., Drummond, A., and Greenhill, S. (2009) 'Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement', *Science*, 323: 479–83.

Gray, R. D., Atkinson, Q. D., and Greenhill, S. J. (2011) 'Language Evolution and Human History: What a Difference a Date Makes', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366/1567: 1090–100.

Gulya, J. (1977) 'Megjegyzesek az Ugor Őshaza és az Ugor Nyelvek Szétválása Kérdéseiről', in Bartha, A., Czeglédy, K., and Róna-Tas, A. (eds.) *Magyar Őstörténeti Tanulmányok*, pp. 115–21. Budapest: Akadémia Kiadó.

Haak, W. et al. (2015) 'Massive Migration from the Steppe is a Source for Indo-European Languages in Europe', *Nature*, 522: 207–11.

Hajdú, P. (1952) Az ugor kor helyének és idejének kérdéséhez. Nyelvtudományi közlemények 54, pp. 264–69. Budapest.

—— (1953) 'Die Ältesten Berührungen Zwischen Den Samojeden Und Den Jenisseischen Völkern', *Acta Orientalia Academiae Scientiarum Hungaricae*, 3/1/2: 73–101.

—— (1975) *Finno-Ugrian Languages and Peoples*. London: Deutsch.

—— (1985) 'Der Begriff Des Dialekts in Den Uralischen Sprachen', in W. Veenker (ed.) *Dialectologia Uralica: Materialen Des Ersten Internationalen Symposions Zur Dialektologie Der Uralischen Sprachen 4.-7. September 1984 in Hamburg*, pp. 1–15. Wiesbaden: Harrassowitz.

Haynie, H. et al. (2014) 'Wanderwörter in Languages of the Americas and Australia', *Ampersand*, 1: 1–18.

Häkkinen, J. P. (2007) 'Kantauralin Murteutuminen Vokaalivastaavuuksien Valossa', Masters thesis. Helsingin Yliopisto, Helsinki, Finland.

—— (2009) 'Kantauralin Ajoitus ja Paikannus Perustelut Puntarissa', *Suomalais-Ugrilaisen Seuran Aikakauskirja*, 2009/92: 9–56.

Hammarström, H., Forkel, R., and Haspelmath, M. (2018) Glottolog 3.3. Jena: Max Planck Institute for the Science of Human History. Retrieved 25 Oct 2018 <http://glottolog.org>.

Heggarty, P. (2006) 'Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data - and to Dating Language?', in P. Forester and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages. McDonald Institute for Archaeological Research*. Cambridge: McDonald Institute for Archaeological Research.

——, and Renfrew, C. (2014) 'Introduction: Languages' in Renfrew C.and Bahn P. G. (eds.) *The Cambridge World Prehistory*, pp. 19–44. Cambridge: Cambridge University Press.

Heikkilä, M. (2011) 'Huomioita Kantasaamen Ajoittamisesta ja Paikantamisesta Sekä Germaanisia Etymologioita Saamelais-Suomalaisille Sanoille', *Virittäjä*, 115: 68–84.

—— (2014) *Bidrag till Fennoskandiens Språkliga Förhistoria i Tid Och Rum*. Helsinki: Unigrafia.

—— (2016) 'Varhaissuomen Äännehistorian Kronologiasta', *Sananjalka*, 58: 136–58.

—— (2019) Monitieteinen tutkimus kristinuskon tulosta Suomenlahden pohjoispuolelle - kristinuskoon tutustumisen, kristityksi kääntymisen ja kristillisen organisaation rakentamisen ajoitus. Teoksessa Anni Ruohomäki & Anna Sivula (toim.), *Tuhansien vuosien tulokkaat: monen kulttuurin Satakunta*. Satakunta XXXV, 31–71. Harjavalta: Satakunnan Historiallinen Seura.

Honkola, T. et al. (2013) 'Cultural and Climatic Changes Shape the Evolutionary History of the Uralic Languages', *Journal of Evolutionary Biology*, 26/6: 1244–53.

Honti, L. (1979) 'Characteristic Features of Ugric Languages (Observations on the Question of Ugric Unity)', *Acta Linguistica Academiae Scientiarum Hungaricae*, 29: 1–26.

—— (1997) *Az ugor alapnyelv kérdéséhez. Budapesti finnugor füzetek, 7*. Budapest: ELTE Finnugor Tanszék.

Hruschka, D. J. et al. (2015) 'Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution', *Current Biology: CB*, 25/1: 1–9.

Ilumäe, A-M. et al. (2016) 'Human Y Chromosome Haplogroup N: A Non-trivial Time-Resolved Phylogeography that Cuts across Language Families', *American Journal of Human Genetics*, 99: 163–173.

Janhunen, J. (1998) 'Samoyedic', in D. Abondolo (ed.) *The Uralic Languages*, pp. 457–79. London & New York: Routledge.

—— (2000) 'Reconstructing Proto-Uralic Typology Spanning the Millennia of Linguistic Evolution', in Nurk A., Palo T., and Seilenthal T. (eds.) *Congressus Nonus Internationalis Fenno-Ugristarum, 7.–13.8.2000, Tartu. Pars I. Orationes Plenariae & Orationes Publicae*, pp. 59–76. Tartu: Eesti Fennougristide Komitee.

—— (2009) 'Proto-Uralic: What, Where, and When?', in Ylikoski, D. (ed.) *The quasquicentennial of the Finno-Ugrian Society. Suomalais-ugrilaisen seuran toimituksia*, pp. 57–78. Helsinki: Societe finno-ougrienne.

Joki, A. J. (1973) *Uralier Und Indogermanen: Die Älteren Berührungen Zwischen Den Uralischen Und Indogermanischen Sprachen*. Helsinki: Suomalais-ugrilainen seura.

Kallio, P. (2006a) 'A. On the Earliest Slavic Loanwords in Finnic', in J. Nuorluoto (ed.) *The Slavicization of the Russian North: Mechanisms and Chronology. Slavica Helsingiensia 27*, pp. 154–66. Helsinki: University of Helsinki Department of Slavonic and Baltic Languages and Literatures.

—— (2006b) *B. Suomen Kantakielten Absoluuttista Kronologiaa. Virittäjä, 110*, pp. 2–25. Helsinki: Kotikielen seura.

—— (2014) 'The Diversification of Proto-Finnic', in J. Ahola and E. Frog (eds.) *Fibula, Fabula, Fact: The Viking Age in Finland*, pp. 155–68. Helsinki: Studia Fennica Historica 18.

—— (2015) 'The Stratigraphy of the Germanic Loanwords in Finnic', in J. O. Askedal and H. F. Nielsen (eds.) *Early Germanic Languages in Contact*, pp. 23–38. Amsterdam, Philadelphia, PA: North-Western European Language Evolution Supplement Series 27.

Kiparsky, V. (1975) *Russische historische Grammatik 3: Entwicklung des Wortschatzes*. Heidelberg: Carl Winter Universitätsverlag.

Kitchen, A. et al. (2009) 'Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the near East', *Proceedings. Biological Sciences*, 276/1668: 2703–10.

Kulonen, U. (1993) *Johdatus unkarin kielen historiaan*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Kumar, S. (2005) 'Molecular Clocks: Four Decades of Evolution', *Nature Reviews Genetics*, 6: 654–62.

——, and Hedges, S. B. (2016) 'Advances in Time Estimation Methods for Molecular Data', *Molecular Biology and Evolution*, 33/4: 863–9.

Koivulehto, J. (2001) 'The Earliest Contacts Between Indo-European and Uralic Speakers in the Light of Lexical Loans', in C. Carpelan, A. Parpola, and P. Koskikallio (eds.) *Early Contacts between Uralic and Indo-European: Linguistic and Archaeological Considerations. Suomalais-Ugrilaisen Seuran Toimituksia 242*. pp. 235–264. Helsinki: Suomalais-Ugrilainen Seura.

Kolipakam, V. et al. (2018) 'A Bayesian Phylogenetic Study of the Dravidian Language Family', *Royal Society Open Science*, 5/3: 171504.

Korenchy, É. (1972) *Iranische Lehnwörter in den obugrischen Sprachen*. Budapest: Akadémiai Kiadó.

Korhonen, M. (1981) Johdatus lapin kielen historiaan. Suomalaisen Kirjallisuuden Seura, Helsinki.

—— (1984) 'Suomalaisten Suomalais-Ugrilainen Tausta Historiallis-Vertailevan Kielitieteen Valossa', *Suomen Väestön Esihistorialliset Juuret*, 55–71. Helsinki: Societas Scientiarum Fennica.

——, and Kulonen, U. (1991) 'Samojedit', in J. Laakso (ed.) *Uralilaiset Kansat: Tietoa Suomen Sukukielistä ja Niiden Puhujista*, pp. 302–317. Juva: WSOY.

Krell, K. S. (1998) 'Gimbutas' Kurgan-PIE homeland hypothesis: a linguistic critique', in Blench, R. and Spriggs, M. (eds.) *Archaeology and Language II: Archaeological Data and Linguistic Hypotheses*. London: Routledge.

Laakso, J. (1991) 'Itämerensuomalaiset Sukukielemme ja Niiden Puhujat', in Laakso J. (ed.) *Uralilaiset Kansat: Tietoa Suomen Sukukielistä ja Niiden Puhujista*, pp. 49–122. Juva: WSOY.

Lang, V. (2018) *Läänemeresoome Tulemised: Finnic Be-Com ings*. Tartu: Tartu Ülikooli Kirjastus.

Lee, S., and Hasegawa, T. (2011) 'Bayesian Phylogenetic Analysis Supports an Agricultural Origin of Japonic Languages', *Proceedings. Biological Sciences*, 278/1725: 3662–9.

Lehtinen, J. et al. (2014) 'Behind Family Trees: Secondary Connections in Uralic Language Networks. Behind Family Trees: Secondary Connections in Uralic Language Networks', *Language Dynamics and Change*, 4/2: 189–221.

Matras, Y. (2002). *Romani: A Linguistic Introduction*. Cambridge: Cambridge University Press.

McMahon, A. M., and McMahon, R. (2006) 'Why Linguists Don't Do Dates', in P. Forster and C. Renfrew (eds.) *Phylogenetic Methods and the Prehistory of Languages*, pp. 153–60. Cambridge: McDonald Institute for Archaeological Research.

Nascimento, F. F., Reis, M. D., and Yang, Z. (2017) 'A Biologist's Guide to Bayesian Phylogenetic Analysis', *Nature Ecology & Evolution*, 1/10: 1446–54.

Nichols, J., and Warnow, T. (2008) 'Tutorial on Computational Linguistic Phylogeny', *Language and Linguistics Compass*, 2: 760–820.

Pagel, M., Atkinson, Q. D., and Andrew, M. (2007) 'Frequency of Word-Use Predicts Rates of Lexical Evolution Throughout Indo-European History', *Nature*, 449: 717–20.

Parham, J. F. et al. (2011) 'Best Practices for Justifying Fossil Calibrations', *Systematic Biology*, 61/2: 346–59.

Parpola, A. (2012) Formation of the Indo-European and Uralic (Finno-Ugric) language families in the light of archaeology:

Revised and integrated 'total' correlations. *Suomalais-ugrilaisen Seuran Toimituksia*.

Purhonen, P. (1998). 'Kristinuskon Saapumisesta Suomeen: Uskontoarkeologinen Tutkimus', in *Suomen Muinaismuistoyhdistyksen Aikakauskirja 106*. Helsinki.

Rahkonen, P. (2017) 'Onomasticon of Levänluhta and Käldamäki Region', *Journal de la Société Finno-Ougrienne*, 96: 287–316.

Ramsey, C. B. (2009) 'Bayesian Analysis of Radiocarbon Dates', *Radiocarbon*, 51/1: 337–60. doi: 10.1017/S0033822200033865

Rédei, K., and Róna-Tas, A. (1983) 'Early Bulgarian Loanwords in the Permian Languages', *Acta Orientialia Academiae Scientiarum Hungaricae*, 37/1: 3–41.

—— (1986) *Zu Den Indogermanisch-Uralischen Sprachkontakten*. Wien: Verlag der Österreichischen Akad emie der Wissenschaften.

Ritchie, A. M., and Ho, S. Y. W. (2019) 'Influence of the Tree Prior and Sampling Scale on Bayesian Phylogenetic Estimates of the Origin Times of Language Families', *Journal of Language Evolution*, 4/2: 108–23.

Róna-Tas, A. (1988) 'Turkic Influence on the Uralic Languages', in Sinor, D. (ed.) *The Uralic Languages. Description, History and Foreign Influences*, pp. 742–780. Leiden; New York: Brill.

Rorlich, A. (1986) *The Volga Tatars: A Profile in National Resilience*. Washington, DC: Hoover Institution Press.

Sagart, L. et al. (2019) 'Dated Language Phylogenies Shed Light on the Ancestry of Sino-Tibetan', *Proceedings of the National Academy of Sciences*, 116 /21: 10317–22.

Saarikivi, J., and Grünthal, R. (2005) 'Itämerensuomalaisten Kielten Uralilainen Tausta', in J. Vaattovaara, T. Suutari, H. Lappalainen, and R. Grünthal (eds.) *Muuttuva Muoto: Kirjoituksia Tapani Lehtisen 60-Vuotispäivän Kunniaksi*, pp. 111–46. (Kieli; No. 16). Helsinki: Helsingin Yliopiston Suomen Kielen Laitos.

Schalin, J. (2014) Scandinavian-Finnish Language Contact in the Viking Age in the Light of Borrowed Names. *Studia Fennica. Historica*, pp. 399–436.

Syrjänen, K. et al. (2018). Lexibank/uralex: UraLex Basic Vocabulary Dataset (Version v1.0) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.1459402> accessed 25 Oct 2018.

—— et al. (2013) 'Shedding More Light on Language Classification Using Basic Vocabularies and Phylogenetic Methods: A Case Study of Uralic', *Diachronica*, 30/3: 323–52.

Tambets, K. et al. (2018) 'Genes Reveal Traces of Common Recent Demographic History for Most of the Uralic-Speaking Populations', *Genome Biology*, 19/1: 139.

Thurgood, G. (1999) *From Ancient Cham to Modern Dialects: Two Thousand Years of Language Contact and Change*. Honolulu, Hawaii: University of Hawai'i Press.

Wasserman, L. (2004) 'All of Statistics: A Concise Course in Statistical Inference'. *Springer Texts in Statistics*. New York, NY: Springer.