ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/328696049

Bayesian Phylolinguistics

Preprint · November 2018

CITATIONS
0

READS
924

3 authors:

Simon J Greenhill
Max Planck Institute for the Science of Human History
92 PUBLICATIONS **3,170** CITATIONS
SEE PROFILE

Paul Heggarty
Max Planck Institute for the Science of Human History
77 PUBLICATIONS **665** CITATIONS
SEE PROFILE

Russell D Gray
Max Planck Institute for Evolutionary Anthropology
254 PUBLICATIONS **11,508** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

DPLACE View project

Grammar of tool-making in New Caledonian crows View project

All content following this page was uploaded by Simon J Greenhill on 02 November 2018.

The user has requested enhancement of the downloaded file.

BAYESIAN PHYLOLINGUISTICS

Simon J. Greenhill[1,2], Paul Heggarty[1] and Russell D. Gray[1,2,3]

[1.] Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, 07745 Jena, German.
[2.] ARC Centre of Excellence for the Dynamics of Language, Australian National University, Canberra, ACT 0200, Australia.
[3.] School of Psychology, University of Auckland, Auckland 1142, New Zealand.

1. INTRODUCTION

Change is coming to historical linguistics. Big, or at least "big*ish* data" (Gray and Watts 2017), are now becoming increasingly available in the form of large web accessible lexical, typological and phonological databases (e.g. ABVD (Greenhill et al 2008), Chirilla (Bowern 2016), Phoible (Moran 2014), WALS (Haspelmath 2014), Autotyp (Bickel et al 2017) and the soon to be released Lexibank, Grambank, Parabank and Numeralbank http://www.shh.mpg.de/180672/glottobank). This deluge of data is way beyond the ability of any one person to process accurately in their head. The deluge will thus inevitably drive the demand for appropriate computational tools to process and analyze the fast wealth of freely available linguistic information. In this chapter we will briefly describe one such set of computational tools – Bayesian phylogenetic methods – and outline their utility for historical linguistics. We will focus on four main questions: what is Bayesian phylolinguistics, why does this approach typically focus on lexical data, how is it able to estimate divergence dates, and how reliable are the results?

2. WHAT IS BAYESIAN PHYLOGENETICS?

There is a proud tradition of building language family trees in historical linguistics that was popularized by August Schleicher in the nineteenth century (Atkinson and Gray 2005), but dates back to at least the seventeenth century (List el al 2016). These language family trees are typically constructed in a qualitative manner from innovations in lexical, phonological, and morphological characters. However, despite the fact that implicit in the comparative method is some kind of optimization procedure, traditional historical linguists do not use an explicit optimality criterion to select the best tree, nor do they use an efficient computer algorithm to search for the best tree. This is surprising given that the task of finding the best tree, or set of trees, is inherently a combinatorial optimization problem of considerable computational difficulty. For just five languages there are 105 ways of subgrouping them in a rooted bifurcating tree. For 10 languages this number grows to over 34,000,000, and for 100 languages the number of possible trees is greater than a number of atoms in the universe (Felsenstein 1979). Traditional language family trees suffer from two additional limitations – their branching patterns reflect a only a relative chronology and contain no information about the uncertainty of any proposed subgroupings. Bayesian phylogenetic methods provide a useful *supplement* to the comparative method. They enable us to build trees with both explicit estimates of branch lengths and subgrouping uncertainty in an objective repeatable manner. These trees can then be used to evaluate subgrouping hypotheses (Gray el al 2009, Greenhill and Gray 2012), date language divergences (Gray et al 2009, Bouckaert et al 2012), estimate ancestral states, test hypotheses of functional dependencies in linguistic features (Dunn et al 2011), and infer geographic homelands and migration routes (Bouckaert et al 2012, Grolemund et al 2016). So what exactly are Bayesian phylogenetic methods?

Bayesian Phylogenetic methods use Bayes Theorem to make probabilistic inferences about phylogenetic trees and their model parameters. The approach calculates a posterior probability distribution of trees P(A|B) as a function of the prior probability of a tree P(A) and the likelihood of the data (B) given the model of character change. In the analysis of DNA sequences the model specifies the number of parameters to be estimated for the rate/s of nucleotide substitution. They can range from simple models where there is just one rate (the Jukes Cantor model), to more complex models where different types of substitutions have different rates (e.g. the General Time Reversible model where there are six rate parameters to be estimated). The model can also include complexities such as between site rate variation where different nucleotide sites are fitted to a distribution of rates (e.g. a gamma distribution). MCMC (Markov Chain Monte Carlo) methods are typically used to estimate the posterior distribution. In this procedure the search through tree and parameter space starts with a random tree and arbitrary values for the model parameters and branch lengths. A likelihood score is calculated for this tree and set of parameters. Then a new tree and model parameters are proposed. If the likelihood score is better this proposal is accepted. If the likelihood is worse, then the proposal will be accepted with a probability determined by the ratio of likelihood of the proposed tree divided by the existing one. This process is continued for a very large number of steps (generations). The MCMC chain explores tree and parameter space incrementally finding the set of trees that best fit the data given the model. The tree topology, branch lengths and model parameters are saved at intervals so that the progress of the chain can be assessed. After an initial "burn in" period the MCMC chain should end up converging on a region where the trees are sampled in proportion to their posterior probability. In practice, it is sensible to use multiple runs with different random starting trees to check that the analysis has converged on this region. For a recent review of Bayesian phylogenetics in biology that includes a good discussion of some of the practical issues with MCMC searches see Nascimento et al (2017).

A key feature of Bayesian phylogenetic inference is that the result of the analysis is not a single tree but rather a set of trees and their model parameters sampled in proportion to their posterior probability. This set of trees is often summarized in a consensus tree. Consensus trees are normally depicted with the posterior probability of a clade (subgroup) shown on the branch. This number is the percentage of trees in the posterior distribution that contains that branch. Thus we end up not just with a single optimal tree, but rather a set of trees that lets us evaluate the extent the data supports a particular inference (given the model and the prior).

Figure 1 shows how this approach can be adapted for linguistic inferences. First, some data must be selected. To date this has mainly been basic vocabulary. The rationale for this choice will be discussed in the next section. At this point we will simply note that in principle other kinds of linguistic data can be used in phylogenetic analyses and have (see Greenhill et al 2010, Greenhill et al 2017), and that combined analyses are also possible. Given that we have lexical data the next and in many ways most crucial and time-consuming step involves coding the data for cognacy. Here the comparative method is king. We want real cognates not mere "look-a-likes". While considerable progress has been made on automating cognate coding – up to 89% accuracy - these procedures are best viewed as assisting rather than replacing the linguist (List et al 2017). Once the lexical data have been coded for cognacy the cognate sets can be converted into a matrix suitable for phylogenetic analyses (Step 2). The matrix might either be a multistate or binary depending on the model of character change that is going to be selected. In the multistate matrix each semantic slot would be a character (column) with the value for each cell reflecting a cognate

class. In the binary matrix each cognate set has its own column and the value in the cell reflect the presence or absence of that cognate in the language (see figure 1).

Step three involves the specification of the prior. This can include information on the likely distribution of trees, model parameters, branch lengths and the like. One advantage of the Bayesian approach is that information from other sources can be included in the prior to aid the best inference. For example, the age of ancient manuscripts or the timing of known historical events can be used to calibrate the trees to help date divergence times (see section 4). Inferences from other kinds of data (e.g. phonological innovations) could also be included in the prior to aid subgrouping inferences. Step 4 entails selecting a model of character change. Models can range from the simple binary single rate model with strict clock, to models with rate heterogeneity (Yang 1993) and a relaxed clock. There is a tradeoff in model selection. If the model is too simple the inferences may be inaccurate. If it is too complex (over-parameterised) then the excess of parameters to be estimated will inflate the variance associated with each estimate and thus limit the power of the analysis (Burnham and Anderson 1998). In careful Bayesian analyses the performance of different models should be evaluated using Bayes factors to determine the best model. In practice we have found that the covarion model (Penny et al 2001), where sites can switch between fast and slow rates on different parts of the tree, often outperforms simpler models such as the single rate or Dollo (single cognate gain) models. Similarly - and this will be of no surprise to critics of glottochronology - the relaxed clock model out performs the strict clock.

Step 5 is the MCMC search through tree and parameter space. Because each step in the Markov chain is necessarily strongly correlated with the previous step, samples from the chain need to be taken many generations apart (e.g. 1000 or even 10,000 generations). The chain also needs to be run long enough to get well past "burnin", and multiple runs with different random seeds are needed to check for convergence. The final step in the Bayesian analysis is to summarize the tree topology and the key parameters of interest such as divergence dates. While consensus trees are often used to summarize the inferences about the tree topology, it is our experience that linguists often ignore the posterior probabilities on the branches and treat all branches in the consensus tree equally. This is a bad mistake. The real result of the analysis is not a single tree but the posterior set of trees. A key feature of Bayesian phylolinguistics is that this set reveals the strength of support in the data for various subgroupings (given the prior and the model). A useful way to visualize this uncertainty is to plot all the trees from the posterior distribution on top of each other in a "densitree" (Bouckaert 2010). Figure 2 shows a densitree for Central Pacific languages. The tree was constructed from cognate coded basic vocabulary from the Austronesian Basic Vocabulary Database (Greenhill et al 2008). A binary covarion model was used in the analysis. The densitree shows that some subgroups such as Eastern Polynesian are reliably recovered. It also reveals considerable uncertainty in the placement of many languages, perhaps as a consequence of the conflicting signal caused by dialect networks (Gray et al 2010).

3. Why Use "The Lexicon"?

Historical linguists do not generally see the lexicon as their data-set of choice. The lexicon is the level of language most exposed to borrowing, and it is also not immune to chance resemblances that may appear to go back to a common original form, but do not (e.g. Spanish *mucho* and English *much*). The comparative method insists instead on other types of language data, deemed much more conclusive of language relationships: form-to-meaning correspondences across extensive morphological paradigms (e.g. the parallels first identified between the case endings in Sanskrit, Greek and Latin); and repeated, consistent sound correspondences, linked by regular and natural 'laws' of sound change between them. So why have most Bayesian phylogenetic analyses

nonetheless drawn their comparative language data from the lexicon? And can the results be trusted? To answer these questions we first need to clarify what exactly is meant by 'lexical data'. We also need to break this down into three separate issues, about the suitability of the lexicon for different purposes: (1) to establish *whether* languages are related within a family; (2) to recover the *phylogeny* of an already established family; and (3) to research a family's *chronology*.

3.1 The Lexicon for Establishing Relatedness?

Swadesh first drew up his famous lists to target meanings that he assumed were the most stable and resistant to borrowing. In the early heyday of lexicostatistics, this morphed into an optimistic assumption that even a minimal level of apparent matches between the Swadesh lists for two languages could be presumed to be deep shared cognates rather than loanwords, and thereby demonstrated a deep 'stock'-level relationship (see Heggarty 2010: 307-309). Even the core lexicon is by no means entirely borrowing-free, however, so claims to establish relatedness from lexical lists alone have never convinced. Basic vocabulary is still used in exploratory studies, as a fertile hunting ground where deep cognates are most likely to survive. But that has to be followed up by the orthodox comparative method, to either confirm or dismiss whether any apparent matches reflect a relationship of common descent (cognacy), or not. Likewise, a Bayesian phylogeny or 'family tree' is meaningful only for languages that do actually form a family with each other, and only when based on data that are probative of phylogenetic relationships. Such data lie not in lexis *per se*, but in cognate status, which can only be established by the comparative method. So Bayesian phylogenetic methods are not proposed here as a method to try to establish whether languages are related in the first place. On the contrary, they rely on the comparative method to establish that in advance, so as to create the input data they need: cognacy judgements. In any case, a simple lexicostatistical *count* of putative 'matches' has nothing to do with Bayesian phylogenetic analyses and their explicit *modelling* of descent with modification through time.

3.2 Why Use 'The Lexicon' for Phylogeny?

Once the comparative method has already established a family, one can then move on to the second task: recovering the structure of its family tree, i.e. the basic objective of *phylogenetic* analysis. To this end, the comparative method continues to privilege regular sound changes and morphology (especially paradigms, where available), however. It can seem questionable, then, that many Bayesian phylogenetic approaches are described as using 'lexical data' instead. It is easy to misconstrue and overplay, however, the apparent contrast between the comparative method and 'mere' lexical data. For a start, the preference for morphological paradigms belongs only to a best-case wish-list. In reality, extensive paradigms are simply not available in all language types. With isolating languages, the comparative method has to make do without them — hence the greater difficulties and doubts surrounding the deep reconstruction of families such as Sino-Tibetan. Lexical data are universally available (even if somewhat more limited in polysynthetic languages). There are also many under-documented and extinct languages for which lexical data are all we have, and all we ever will have.

As for sound laws, in order to establish that a sound correspondence is regular and repeated, the comparative method requires multiple tokens of that correspondence. Where are those tokens to be found? Firstly, grammatical morphemes are so few in number that most tokens are necessarily found in the lexicon instead. Secondly, tokens of regular sound changes are by definition found mostly in cognates. (Loanwords may mimic *some* past sound changes when phonologically adapted into the borrower language, but otherwise show only those changes that arose thereafter.) Thirdly, cognate tokens are not evenly distributed across the whole lexicon, but

necessarily concentrated in that part of it least susceptible to lexical replacement (whether by loanwords or language-internal semantic shifts). So while in principle there is no special focus on basic vocabulary when searching for tokens of regular sound correspondences, in practice the comparative method finds much of its evidence (more than in any other comparable sector of the vocabulary) in exactly the same stable core of the lexicon that phylogenetic analyses use.

More importantly still, much of the confusion and dismissal of the phylogenetic value in mere 'lexical data' is because that term is in fact a popular misnomer. The crucial comparative linguistic datum used is not lexicon, but cognacy. The 0 and 1 (or multi-state) values that form the input to phylogenetic analyses do not directly represent lexemes at all, or their phonological forms. What the 0s and 1s represent, directly and uniquely, are relationships of cognacy, i.e. common descent. Simply drawing up the Swadesh list for a given language is not enough. The list of lexemes is not the data, and is unusable by the method until each lexeme has been assigned to its correct cognate set *vis-à-vis* the lexemes in all other languages in that family. Only then is there any actual comparative data to be fed into the phylogenetic analysis — because, to repeat, the input data are not lexical forms, but cognacy relationships. So to describe the data as 'lexical' is a misnomer particularly because it implies, by omission, no input from phonology (sound correspondences), morphology, and the comparative method in general. Yet these are, on the contrary, all inescapably implicit in the concept of cognacy itself. The task of assigning cognacy, to create the input data, requires detailed, qualitative, historical linguistic analysis, by the comparative method.

Any cognate set is defined by common descent. Specifically, it is defined and identified by the proto-form from which all its members descend directly (i.e. without borrowing, see below). The cognate set to which English *rain* and German *Regen* belong, for example, is defined by the Proto-Germanic form *regna (www.cobl.info/meaning/rain). Likewise, the fact that Greek *kardía*, Russian *serdce* and Old Irish *cride* are cognate is defined by the Proto-Indo-European form *ḱerd-, from which they all descend (www.cobl.info/meaning/heart). That is, to define cognate sets relies on the comparative method, its sound-change laws and reconstructions. Thousands of such individual cognate sets are what make up the phylogenetic signal that Bayesian approaches make use of.

Cognacy assignment, when properly performed, integrates and rests on all of the data, methodology and findings of orthodox comparative-historical linguistics, not least in phonology and morphology. It is only that methodology that can reliably establish which word-forms are true cognates to each other, even when sound changes have obscured original similarities. Likewise, only the comparative method can tell apart true cognates from loanwords or from chance lookalikes, essentially because those do *not* exhibit the expected regular sound correspondences. The data that Bayesian phylogenetic analysis uses are best not described as just 'lexical data', then, but more accurately as *cognacy data,* for a sample list of precisely defined *lexical* meanings.

Finally, one should be under no illusions as to the many complexities and challenges in preparing cross-linguistic data-sets of cognacy in lexical meanings. Early data-sets such as that by Dyen, Kruskal & Black (1992) have been rightly criticized for many failings; much progress is still being made towards more rigorous and consistent policies for drawing up such databases, specifically as appropriate for the purposes of phylogenetic analysis and chronology estimation. The methodology needed is too complex to do it justice in this chapter, however, and is set out in much more detail in Heggarty *et al*. (in prep.). For a major language family, assigning cognacy across over scores or hundreds of language varieties, ancient and modern, in up to 200 target

meanings for each, is a daunting task. Even for the best-researched language families like Indo-European, it demands painstaking historical linguistic analysis by large teams of language and family experts. That all serves only to re-emphasise the fundamental point: far from being discarded, all the methodology and findings of orthodox historical linguistics are precisely what databases of *cognacy* in basic lexicon are founded upon in the first place.

3.3 Size Matters

There is one other crucial advantage of data in the form of cognacy in lexical meanings: there are enough of them. For Bayesian phylogenetic analysis to work well requires very significant amounts of data in order to estimate all main aspects of the results: the tree structure, branch lengths, model parameters, the corresponding time-depths, and so on. The large quantities of data needed are only really available from cognacy in the lexicon. The experience of Ringe *et al*. (2002), in their (non-Bayesian) search for a "perfect phylogeny" for the Indo-European family, is instructive here. Ringe *et al*. start out with the historical linguist's instinctive preference for data characters in phonology and morphology. But they find only 22 or the former, and 15 of the latter, that they consider validly usable, and retain even "fewer morphological characters" in their follow-up study (Nakhleh *et al*. 2005: 394). This not only makes for a very small data-set, but as Ringe *et al*. (2002: 98) themselves admit, "the worst news is yet to come: the vast majority of our well-behaved monomorphic characters simply define one or more of the ten uncontroversial subgroups of the family, contributing nothing to their higher-order subgrouping". In practice, the main higher-order nodes in their output trees turn out to be supported by just two, one or even none of these phonological and morphological characters. Most of the phylogenetic information ends up being provided by their lexical cognacy characters after all, thanks to the sheer number of them.

A further consideration is that Ringe *et al.*'s phonological characters are effectively all just binary. Their P2 character "full 'satem' development of dorsals", for example, allows values of just present or absent. Moreover, any one language cannot change from one of those states to the other more than once over the entire history of Indo-European. Each lexical meaning, however, typically corresponds to a range of many different cognacy states, effectively multiplying the amount of discriminatory data in that one meaning. For a broad family such as Indo-European, the average meaning has well over ten different cognacy states (for exact statistics per meaning see www.cobl.info/wordlist/Jena175). And in each meaning, changes from one (cognacy) state to another can arise continuously, across the tree, throughout the family's divergence history. A 200-meaning list thus turns into many thousands of cognacy states, and even if not all of them are informative on the higher-order branching, they all still contribute crucially to the estimation of branch lengths and time-depths.

This is also why phylogenetic analyses are better to aim for data-sets of the order of Swadesh's initial 200 meanings, rather than shorter versions slimmed down for other purposes, such as Swadesh's own 100-meaning list or Tadmor *et al*.'s (2010: 238-243) Leipzig-Jakarta list, also of just 100 meanings. This need for sheer quantity of data is another reminder of the scale of the task to draw up such comparative databases.

3.4 Not Cherry-Picking, But 'Safety in Numbers'

Large data-sets, based on a pre-determined list of meanings intended to be applied to any language family, also have the advantage of forcing objectivity. They avoid the danger of individual scholars subjectively cherry-picking data characters, not least in phonology and morphology, that may be unrepresentative and open to interpretation (including, for instance, as to which state was

ancestral). It is striking that historical linguistics has still failed to come to a consensus family tree even for many of the best researched of all language families, including Romance, Germanic, and Indo-European as a whole. In part this is because of how the discipline has essentially proceeded on the disputed points: competing scholars each invoke small subsets of the language data, cherry-picked to argue for one tree structure over another, but which fail to convince while other cherry-picked subsets point to rival analyses. Subjectively *excluding* other selected data, meanwhile, comes with a risk of circularity, in that particular characters may be judged unreliable and excluded precisely because they do not fit with a preconceived idea of what the tree structure should be. Ringe *et al.*'s (2002) data-set is heavily "screened", one of the reasons why they end up with so few informative characters, but they are still frustrated by far more characters incompatible with any tree (at least 16) than supporting the main higher-order nodes (mostly 4 or less). It is not that all language data necessarily reflect phylogeny; of course not. But determining which innovations most likely arose in parallel, for instance, and on which different branches, is precisely one of the results that emerges in any case from a powerful Bayesian phylogenetic analysis — and in the form of a balanced, quantitative assessment rather than a subjective and potentially circular judgement. We shall see shortly below the obvious dangers of cherry-picking data for chronological purposes, too, when trying to second-guess how much or how little language change is 'plausible' within a given time-span. Avoiding such subjectivity is certainly one of the attractions of large data-sets of cognacy across a pre-determined list of lexical meanings.

3.5 Why Use the Lexicon *for Dating?*

Establishing the phylogeny of a language family is one thing; researching its chronology is quite another. The fact that glottochronology was directly based on lexicostatistics has given the lexicon a bad name for this purpose. The basic problem is not the lexicon itself, however, nor the concept of cognacy. Rather, the undoing of glottochronology was its unrealistically strict assumptions, especially of a universal, unvarying rate of change, and its simplistic distance-based *measures*, without any real *model* of linguistic 'descent with modification' through time. Certainly, it is hardly as if historical linguists have enjoyed much success and unanimity in looking to other levels of language, beyond the lexicon, to help in estimating time-depths.

3.6 Dating by Phonetics or Morphosyntax?

In phonology and phonetics, for instance, rates of change and divergence appear to vary even more freely and spectacularly than is usual in the core lexicon. Take for example Latin [akʷam] *water*, [altʊm] *high* and [ad illʊm] *to that*, which in only two millennia have all attrited, in French, to just [o] (distinguished only in spelling as <eau>, <haut> and <au>), alongside other spectacular examples of change such as [kalidʊm]→[ʃo], [kaballos]→[ʃfo] and [habitʊm]→[y] (*hot*, *horses* and *had*). Or one can just as well cherry-pick examples of very slow sound change, even over the many millennia since Proto-Indo-European. Reconstructed *bʰréh₂t(ē)r *brother*, for example, still retains its initial obstruent + rhotic cluster little changed into modern Welsh [braut], Russian [brat], English [bɹʌðəɹ] and 'even' French [fʁɛʁ].

With such great variation in possible rates, arguing just from one-off examples leaves us open to cherry-picking instances where change has been either abnormally fast or abnormally slow, and thus unrepresentative and no proof of anything for dating purposes. Rather, it leaves us open to impressionistic and subjective expressions of conviction, and is what we need to get away from, in favour of an objective, balanced distribution of rates of change, and an explicit historical model.

A keystone in the traditional short chronology for Indo-European, for instance, is the perception that Avestan and Vedic are 'so close' that the divergence between them 'must' be but a matter of a few centuries. Sims-Williams (1998: 126), for example, offers selected sentences that "may be

transposed from the one language into the other merely by observing the appropriate phonological rules". Much the same, though, can be said of selected phrases in Italian and Spanish, for instance: see Heggarty & Renfrew (2014: 545). Change and divergence have been minimal in no end of word pairs, such as Italian [liŋgwa] vs. Spanish [lɛŋgwa] *tongue* (from Latin [liŋgʷa]), [mɔndo]~[mundo] *world*, [θjelo]~[t͡ʃelo] *sky*, etc.. Indeed other cases show no real divergence at all: [kanto]~[kanto] *I sing*, [salta]~[salta] *s/he jumps*, and so on. Here too, simply applying phonological rules can straightforwardly transpose one language to the other, but that hardly proves a time-depth of divergence of just a few centuries, as traditionally envisaged between Avestan and Vedic. For the net divergence between Italian and Spanish to arise, out of Latin, took the same two millennia that in French saw much more radical change.

Morphosyntax, likewise, does not seem to offer a viable data-type for language dating. Starting out from the inflectional case system of Proto-Indo-European, for instance, its modern descendants show an entire gamut of different amounts of net case loss, from total to almost nil. At one extreme, French has lost all traces of the system on nouns, while English retains only its 'Saxon genitive'. Other languages have maintained parts of the system, albeit heavily eroded (e.g. Romanian and German). At the other extreme, meanwhile, most Slav languages still count up to six cases. So over the same time-depth since Proto-Indo-European, different branches and languages have lost (and sometimes, re-created) cases at very different rates and stages in time. Note how Bulgarian, for example, has bucked the trend of Slavic continuity and seen most of its case system collapse since the time of Proto-Slavic. It seems hard to make anything of this for chronological purposes. Take the famous case of Lithuanian, often evoked in admiration at how little some of its case-endings have changed since Proto-Indo-European. This is of no help to decide between the two main hypotheses on the time-depth of Indo-European: little change over six millennia hardly excludes little change over another few millennia, either. A further consideration is that different levels of language can change at very different rates. Within Romance, French is unquestionably an outlier in how much phonetic attrition it has undergone. In lexis, however, French is unremarkable, a typical Romance language with retention and cognacy rates of exactly the same order as most of its sister languages.

Given all this variability, it is no surprise that those historical linguists who dare to make pronouncements on their 'intuitions' of how long a time-span is or is not plausible for a given range of divergence across a family … they do not necessarily agree. Many Indo-Europeanists, seduced by the claims of linguistic palaeontology (likewise subjective and disputed, see Heggarty 2014, Heggarty forthcoming), have taken as read a Steppe hypothesis time-depth of c. six millennia. That assumption has then set their benchmark for intuitions on what rates of divergence are 'plausible', and forced further assumptions on crucial chronological questions such as the time-depth of the Indic-Iranic split that led to Vedic and Avestan. Remove the assumptions, however, and other linguists such as Dolgopolsky (1990: 239) see the Steppe hypothesis time-frame espoused by Mallory (1989) as "utterly unrealistic". Mallory's "dating, which presupposes that Proto-Anatolian, Proto-Indo-Iranian, Greek and other descendant languages could have diverged from each other for a mere 2000 years, is absolutely inconceivable" for Dolgopolsky. Many more linguists would challenge other aspects of Mallory's (1989) chronology, in which Proto-Germanic, for instance, as recently as 2500 years ago, is "a late Indo-European dialect".

3.7 Why Linguistic Dating *Should* Be Possible

Whatever levels of language one looks to, then, cherry-picked examples and impressionistic pronouncements from 'intuition' lead us to a dead-end, a stand-off between opposing subjective positions. So it is perhaps understandable that some linguists steer clear of any idea of dating

from language. "Linguists don't do dates", as McMahon & McMahon (2006) put it, while for Dixon (1997: 47, 49) "What has always filled me with wonder is the assurance with which many historical linguists assign a date to their reconstructed proto-language… How do they know? … It does seem to be a house of cards".

Such pessimism, however, actually loses sight of some basic tenets of historical linguistics. First, it is not entirely mysterious why rates of change can vary; there are often good explanations, both language-internal and language-external. Sub-systems within a language may long be stable, but then once perturbed (sometimes as a knock-on effect of changes on other levels of the language) can change rapidly until they settle into a new stable system, as in the case-system collapse in Bulgarian, or phonological system change in Old Irish. External contexts, too, can vary from relative isolation to punctuations of intense contact. This is widely seen as the explanation for Bergsland & Vogt's (1962) paradigm case of the contrast between high cognate retention in Icelandic alongside bursts of lexical turnover in English, triggered by the Viking and Norman conquests. In short, where rates depart significantly from 'default' expectations, there is often a good reason.

Second, it is clear that language change is unstoppable. So however much the rates may vary, changes do at least build up progressively through time. Indeed in practice many of the same historical linguists who repeat the *de rigueur* disclaimers that glottochronology is not to be trusted do themselves go on to invoke its 'dates', for want of anything better. See Kaufman & Golla (2000: 52), for instance, on the major language families of the Americas. So although glottochronology as a method is firmly discredited, on both of the above grounds there remains a widespread, grudging acceptance that rates of change over time are at least not a complete free-for-all. Or to put it more formally: notwithstanding some much-cited instances of exceptional change or stability, those plausibly represent the extremes of a distribution, even if a broad one, of more usual and more 'modal' rates of change.

That already brings us closer to a potential way of making use of these basic tenets of historical linguistics for a much more sophisticated, objective and realistic approach to language dating. Certainly, with all the discipline's knowledge of language history and how it proceeds, we should be able to do far better than just simplistically totting up numbers up to 100 or 200, as glottochronology did. Indeed, instead we can enlist another fundamental tenet, namely the process of linguistic 'descent with modification' by which any family of languages diverges out of its common ancestor. That points to the applicability, in principle, of phylogenetics (and note, clarifies that it is not glottochronology at all).

3.8 Why Lexical Cognacy for Dating?

Given all those possibilities from Bayesian phylogenetics, it is worth reconsidering which type of language data might be most suitable as input, specifically for chronological purposes. We have already seen some of the problems with other levels of language, and in fact cognate status in the lexicon turns out to have one distinct advantage, specifically for chronological purposes, even over other, form-based data types that can seem more typically what historical linguists would use. For chronology, what is needed is a set of c**omparative data** points that are free to change iteratively and repeatedly, so that their 'clock' does not at some point stop ticking and no longer allow change. This can be a problem with most form-based linguistic data: sounds or morphemes that disappear completely generally cannot then reappear, to start changing again. Take the example of ablaut grades in Indo-European. Many Indo-European roots are found shared across multiple branches of Indo-European, but from one branch to the next the lexemes used can be based on

different ablaut grades. In the meaning FOOT, for example, Germanic the lengthened o-grade of PIE *ped- (hence *foot*), Greek *πούς, ποδός* comes from the (short) o-grade, and Latin *pes, pedis* from the e-grade.

Certainly, while a cognate set is essentially defined by a shared root morpheme, it can be helpfully subdivided into its various 'cognate sub-sets', according to finer morphological or phonological criteria such as different ablaut grades of that root, or the presence of one or other affix additional to the root. Ultimately, though, the ablaut alternation system of Indo-European largely broke down, and lexemes in different branches fossilized their forms, which were no longer free to change any further in this criterion. By the time of Latin, *pes, pedis* was already long fixed as of e-grade, and thereafter all Romance languages inherited that fixed datum. Italian *piede*, or English *foot*, cannot change (back) to any other Indo-European ablaut grade, which ceased being a meaningful system many millennia ago.

From a dating perspective, then, ablaut grade is, unhelpfully, a datum for which the change clock stopped ticking definitively, long ago deep in the history of the Indo-European family. The clock that does remain ticking, meanwhile, is (root) cognate status. Italian and English, like any other Indo-European language, could in principle ultimately switch to a new lexeme for FOOT, in a different (root) cognate set. This ability to iteratively change, for the clock never to stop ticking, is a crucial characteristic that makes cognate status in lexical meanings the data type of choice specifically for the purpose of *dating*. It is also one of the main reasons why it is important most advantageous to analyze cognate status at the level of the root, rather than to use cognate subsets to try to contribute to the chronological estimation.

Note that there are in fact two tasks, then, that it is possible to take somewhat separately: recovering the phylogeny, and dating that phylogeny. The preference for using the cognacy turnover data alone for dating, then, does not limit the language data used to contribute to determining the phylogeny. To that end, one certainly can include finer analysis such as breaking down root cognate sets into their respective subsets, on further morphological and phonological criteria. Indeed, one can make use of morphological and phonological characters that are not iterative, but one-off or sequential changes. These can even be used to constrain the model to return only trees compatible with those characters, but then on that constrained phylogeny one can use only the lexical cognacy data to contribute to the dating. (See Atkinson *et al*. (2005: 209) for an analysis constrained to Ringe *et al.*'s (2002) data, including notably their key phonological and morphological criteria.)

4. How do Bayesian phylogenetic methods actually date language divergences?

Dating is important – it enables us to link linguistic divergences to archaeological and historical events, and build a coherent story about human prehistory. The search for a rigorous methodology to make accurate inferences about timing first led to a set of approaches called lexicostatistics and glottochronology, and then to an almost puritanical rejection of these approaches. More recently, new methods from evolutionary biology – which do not share the fatal shortcomings of glottochronology – have been applied to linguistic questions. Let us start this section with a brief account of this history.

Morris Swadesh (Swadesh 1950, 1952, 1955) proposed a simple computational approach to building language trees called *lexicostatistics*. He argued that the number of cognate words shared between two languages in basic vocabulary was a good indicator of degree of relationship. The logic is simple: successively grouping languages by amount of cognates they shared would give rise

to a family tree with more similar languages grouped together. Swadesh further realised that the degree of similarity shared between two languages was also an approximate measure of how much time had elapsed since they had separated. Making a direct analogy to radioactive decay, Swadesh argued that languages lost similarity at a relatively constant rate over time. These rates were naturally only applicable to a standardized sample of vocabulary that they were calculated from, so Swadesh proposed a series of standardized 100- and 200- item wordlists. For example, if two languages shared *C* cognates and there was a constant "clock" rate *r*, then the age of divergence *t* could be estimated by:

$$t = \log C / 2 \log r$$

Swadesh (1950) initially suggested that the clock rate *r* was 85%/1000 years based on the differences between old and modern English. This figure was later revised to 81%/1000 years based on a survey of rates in the basic vocabulary of historically attested languages (Lees 1953). For example, if two languages shared 81% cognates on a 200 item word list then the estimated age would be 1000 years, while two languages sharing 66% might be expected to have diverged 2000 years ago. This approach became known as *glottochronology*.

Glottochronology was first applied by Swadesh to the Salishan languages (Morris Swadesh 1950). However, the promise of using linguistics to date prehistoric population expansions was so tempting that glottochronology was rapidly adopted by linguists and used on languages around the world (see Hymes (1960) for a review). The findings were widely read outside linguistics with eminent anthropologists carefully discussing the methodology (Kroeber 1955) and arguing that it revolutionized prehistory (Murdock 1964).

However, critics were quick to note some serious methodological flaws – with the key flaw being that language change is not clock-like. Languages can vary substantially in their rates of lexical replacement. For example, a damaging critical study showed that glottochronology would estimate that Old Norse and Icelandic diverged less than 200 years ago. This age is far younger than the historically attested age of 1000 years (Bergsland and Vogt 1962). In fact, rather than the stately 81% loss per 1000 years proposed by glottochronology, these rates varied around 15-20% (Bergsland and Vogt 1962). The finding of substantial rate variation in languages was so damaging to glottochronology that historical linguistics largely rejected quantitative methods. Today, lexicostatistics and glottochronology are seen as textbook examples of bad linguistic methodology, and we are told that "linguists do not do dates" (McMahon and McMahon 2006).

In biology, the scenario started in a similar vein but played out rather differently. Biologists had developed a method to build trees from pairwise similarity (e.g. Sokal and Michener 1958). Divergence times could then be estimated from these trees using a "molecular clock" rate (Zuckerkandl and Pauling 1965), e.g. Kimura (1968) estimated a rate of one base pair change every 1.8 years. Just as in linguistics, biologists were quick to notice the problem of rate variation which could cause incorrect estimates of both the relationships on the family tree (Felsenstein 1978), and the estimated ages (Kirsch 1969).

Unlike historical linguists, who largely rejected computational approaches, biologists developed methods for handling variation in rates. One method, known as *nonparametric rate smoothing* (Sanderson 1997), takes the observed rates of change in the data, and 'smooths' these to fit around calibration information. Another method – currently the best – is known as the *relaxed clock* (in contrast to what might be called the *strict clock* assumption of glottochronology) (Drummond et al. 2006). In the relaxed clock model rates for each branch are drawn from a log-

normal or exponential distribution with parameters estimated from the data. This procedure allows each branch to have its own rate and for the data to inform the analysis about the magnitude of rate variation across the languages.

To *calibrate* the clock, historical information is used to constrain the age of known branches. These calibrations allow the analysis to estimate the clock rates in regions where the timing is known and extrapolate these to infer rates in regions where the timing is unknown. While these calibrations can be simple point estimates (e.g. a given year) it is more common to specify these as probability distributions as this allows uncertainty in the divergence time estimates to be explicitly accounted for (Ho and Phillips 2009). For example, (Birchall, Dunn, and Greenhill 2016) used this approach to date the origin of the Chapacuran languages in South America to around 1,040 years ago. One of the calibrations they used was based on (Meireles 1989) suggestion that Moré and Cojubim speakers were a single population that diverged when the Jesuits arrived in the region in the 1740s and no mention is made of the Cojubim until 1781. (Birchall, Dunn, and Greenhill 2016) implemented this calibration as a log-normal distribution with a 95% probability that Moré and Cojubim diverged between 213 and 723 years ago – so the lower bound was close to the 1780s date, while the upper bound stretched 500 years in the past to allow for the two communities to have diverged somewhat earlier than the Jesuits arrival in case (Meireles 1989) was incorrect.
Given the serious failings of glottochronology, how confident can we be that these methodological advances are accurate? In a simulation test of phylogenetic methods for resolving language history (Greenhill, Currie, and Gray 2009), we evaluated the performance of the rate-smoothing approach. First, we simulated linguistic data on two known phylogenies (described in more detail above). We then calibrated two arbitrarily chosen nodes on each tree and constrained them to be within +- 10% of their true age. We found that the rate smoothing approach was able to estimate a value close to the true root age even when borrowing rates reached up to 15% – the difference between the true age and the estimate age was 2.2% or 6.1% depending on the tree topology. As borrowing increased, the estimated age of the tree decreased. This result shows that these methods are fairly robust to most reasonable levels of borrowing and can reasonably accurately estimate language divergence times. The accuracy and performance of these methods continues to be an ongoing research topic within phylogenetics and performance will likely continue to improve (Drummond and Suchard 2010; Lanfear, Welch, and Bromham 2010; Duchêne, Lanfear, and Ho 2014).

5. How accurate are the results of Bayesian phylolinguistics?

How well do Bayesian phylogenetic methods work at recovering language history? As with any inferential tool it is critical to assess the performance and accuracy. There are two ways we can assess the performance and accuracy of phylogenetic methods on linguistic data. The first way is to validate the phylolinguistic results with results from the comparative method, while the second validates the results of analyses of simulated data where the complete history is known.

5.1 Verification with the comparative method.

Many of the studies using phylogenetic methods discuss how consistent their results are with those of the comparative method. Studies showing strong concordance with the comparative method and their phylogenetically estimated subgroups include Aslian (Dunn et al. 2011), Athabaskan (Sicoli and Holton 2014), Bantu (Grollemund et al. 2015; Grollemund 2012), Chapacuran (Birchall, Dunn, and Greenhill 2016), Huon Peninsula (Greenhill 2015), Indo-European (Bouckaert et al. 2012; Chang et al. 2015), Pama-Nyungan (Bowern and Atkinson 2012), Semitic

(Kitchen et al. 2009), Timor-Alor-Pantar (Robinson and Holton 2012), Tupian (Galucio et al. 2015), Tupi-Guarani (Michael et al. 2015), and Uralic (Syrjänen et al. 2013).

To date perhaps the best comparison between the findings of the traditional comparative method and phylogenetic methods can be found in the Austronesian languages. There are strong hypotheses based about the origin, and subgrouping of these languages derived from the comparative method, and robust dates for many of the main groups derived from archaeology (Diamond and Bellwood 2003; Bellwood and Dizon 2008; Robert Blust 2013; Andrew Pawley 2002; Green 2003; A. M. S. Ko et al. 2014). The Austronesian languages originated in Taiwan where 9 of the 10 major branches are still spoken (Robert Blust 1999; Robert Blust 2013) around 5500 years ago (Robert Blust 2013). The remaining branch – Proto-Malayo-Polynesian – spread south through the Philippines around 1000 years later (Bellwood and Dizon 2008; Robert Blust 1999; Andrew Pawley 2002), through Indonesia and along the coast of New Guinea. By around 3000-3200 years ago the Austronesians can be strongly linked to the development of the Lapita cultural complex in Near Oceania (Sheppard, Chiu, and Walter 2015; Green 2003) which brought a distinctive dentate stamped and red-slipped pottery style, a range of domesticates, and social practices. Finally, around 2800-3000 years ago the Austronesians settled Western Polynesia around 2900 years ago, paused again for 1000 years, and then entered East Polynesia (Wilmshurst et al. 2011).

Importantly, Austronesian is a difficult test case as it is a large family with more than 1200 languages that originated from a rapid population expansion into and across a range of new environments, encountering speakers of very different languages, and undergoing substantial shifts in population size. All of these factors have resulted in huge variation in cognate retention rates between languages (Robert Blust 2000; Greenhill 2015). One of the largest lexicostatistical studies ever conducted (Dyen 1962) on these languages produced bizarre results – concluding that the homeland of the Austronesian language family was in Near Oceania rather than in Taiwan (Greenhill, Drummond, and Gray 2010).

We used phylogenetic methods on 400 Austronesian languages to test between the above 'pulse-pause' settlement scenario and an alternative proposal from genetics. This alternative scenario, the "Slow Boat", suggests the Austronesians originated in Wallacea (the region around modern day Sulawesi) between 13,000 and 17,000 years ago before spreading north into the Philippines and Taiwan, and east into Oceania (Oppenheimer and Richards 2001; Soares et al. 2008). Our results (Gray, Drummond, and Greenhill 2009) showed overwhelming support for the first 'pulse-pause' scenario and none for the 'slow boat' scenario. This support for the pulse-pause scenario was both for the broad tree topology and age of the family – we estimated a mean age of 5230 years (95% highest posterior density interval of 4730-5790 years).

In subsequent papers we carefully evaluated the similarity between the phylogenetic tree and the subgroupings proposed by the comparative method (Greenhill, Drummond, and Gray 2010; Greenhill and Gray 2012). Our analysis did in fact recover most of the key subgroupings identified by the comparative method. Figure 3 visualizes this strong consistency between the linguistically attested subgroupings (based on the classification in the *Ethnologue* (Lewis 2009) and the phylogenetic results. The similarities are overwhelming – and statistically significant (Greenhill, Drummond, and Gray 2010; Greenhill and Gray 2012). We recovered the following groups (working down the tree): Malayo-Polynesian (Dahl 1973), the Philippine 'microgroups' e.g. Bashiic, Central Luzon, Central Philippines, Cordilleran, Manobo, Palawanic, and Subanun (R. Blust 1991), Malayic (Adelaar 1992), Chamic (Thurgood 1999), Celebic (Mead 2003), Greater South Sulawesi (Adelaar 1994), North Sarawak (R. Blust 1974), Eastern Malayo-Polynesian (R. Blust 1978), Central-Eastern Malayo-Polynesian (R. Blust 1978), Bima-Sumba (Robert Blust 2008), Central Maluku

(Collins 1982), Yamdena-North Bomberai (Robert Blust 1993), and South-Halmahera/West New Guinea (R. Blust 1978). We recover the large Oceanic subfamily [Dempwolff (1927); Ross1998], and its Near Oceanic component groups including Temotu (Malcolm Ross and Næss 2007), South-East Solomonic (A. Pawley 1972), Admiralties (M. Ross 1988), Papuan Tip (M. Ross 1988), Meso-Melanesian (M. Ross 1988), and the majority of the North New Guinea languages are strongly grouped together (M. Ross 1988) (some languages from Willaumez are misplaced due to unidentified loan words (Greenhill, Drummond, and Gray 2010)). Remote Oceania is also well resolved with the phylogenies recovering Central Pacific (Grace 1959), Polynesian [Andrew Pawley (1966); Marck2000], Micronesian(Bender et al. 2003), South Vanuatu (Lynch 2001), North and Central Vanuatu [Tryon (1976); Clark2009], New Caledonia and the Loyalties (Ozanne-Rivierre 1992).

To be fair, there are some mismatches with the expected groupings. In total 25 of the 400 languages are misplaced (Greenhill, Drummond, and Gray 2010). Some of these misplacements are due to errors in the data – e.g. the Willaumez languages (Nakanai, Maututu, Lakalai) sit closer to the North New Guinea languages rather than their expected sisters in the Meso-Melanesian subgroup. This misplacement is probably due to unidentified lexical borrowings between the Willaumez languages and their Meso-Melanesian neighbours in west New Britain. However, many of the other misplacements are due to long-standing classification difficulties. For example, it is unclear whether Irarutu should be in the Central Malayo-Polynesian or the South-Halmahera/West New Guinea subgroup (Blust 1993). Another group we do not recover is Central Malayo-Polynesian, however, this group is now thought to be a dialect chain with low internal cohesion that is only supported by overlapping isoglosses (Blust 1993). So while there are some errors in the data – discussed in full in (Greenhill, Drummond, and Gray 2010) – the Bayesian phylogenetic trees do remarkably well at recovering both the known subgroups and the places of uncertainty.

5.2 Validation with simulations

The second way we might validate phylogenetic inference is to simulate data on a known 'true' phylogeny and then analyze the simulated data to see if we recover the 'true' tree. This approach is commonly used in the biological phylogenetics literature to test the effectiveness of these methods. Greenhill et al (Greenhill, Currie, and Gray 2009) applied this logic to test how well Bayesian phylogenetic methods could recover the true tree with lexical cognate data. First, we created two tree topologies: one shaped like Polynesian with an unbalanced, chained topology, and the other shaped like Uto-Aztecan with a well-balanced tree. Then we used a simple 'stochastic Dollo' model of linguistic evolution (Nicholls and Gray 2006; Atkinson et al. 2005) to simulate cognate data. This model started at the root of the tree and worked its way down to the tips evolving new cognate sets to create a new simulated dataset.

One of the major objections to phylogenetic methods in linguistics is that the borrowing of items between languages invalidates the tree (Terrell 1988, Moore 1994). Rather than throwing methods away at the first sign of trouble it is better to assess performance and identify when the method breaks down. To this end we also simulated borrowing events in these data. During the simulation process, traits were randomly selected and copied to another lineage simulating the borrowing of a cognate set between languages. We varied the amount of borrowing from 0% to 50% of the total cognates.

Finally, for each of the simulated datasets, we used Bayesian phylogenetic methods to estimate the posterior probability distribution. For each analysis we then constructed the maximum clade

credibility tree (a single summary tree of the posterior) and then compared how close this estimate was to the 'true' tree. Our results show that when there is no borrowing the analysis finds a tree almost statistically indistinguishable from the true tree. As the amount of borrowing in the data increased so did the distance between the true tree and the recovered tree. What level of borrowing is plausible? A reasonably extreme example of language borrowing is English (leaving aside creoles and mixed languages), which has borrowed more than 60% of its total lexicon from French and Latin (Embleton 1986). However, only 16% of English's basic vocabulary is borrowed (Embleton 1986). Given that most language phylogenies are built from basic vocabulary data, we took 0-20% as the extreme range of *undetected borrowing* in the type of datasets used by language phylogenies. Over this range the distance between the true trees and the recovered trees was very small – the balanced topology showed an average perturbation of at most 0.7%, while the more fragile chained topology showed an average perturbation of 6.8% at most. In summary, this simulation study shows that Bayesian phylogenetic methods are exceptionally good at finding a tree very close to the true history even in the face of realistic levels of undetected borrowing.

Conclusion

The view we have taken here is that Bayesian methods are a powerful *supplement* to traditional linguistic scholarship - not a replacement. In combination with the comparative method, they enable us to estimate uncertainty in subgrouping proposals, quantify branch lengths and infer divergence dates. In the future we expect to see more combined analyses of lexical, phonological and morphological data and the development of increasingly sophisticated "computer assisted" cognate detection and models of cognate evolution. If linguists can exorcise the ghost of lexicostatistics past, then there is an exciting future to embrace.

References

Adelaar, K.A. 1992. Proto-Malayic: A Reconstruction of Its Phonology and Part of Its Morphology and Lexicon. *Canberra: Pacific Linguistics*.

———. 1994. "The Classification of the Tamanic Languages." In *Language Contact and Change in the Austronesian World*, edited by T. Dutton and D. Tryon, 1–42. Berlin: Mouton de Gruyter.

Atkinson, Q. and Gray, R.D. 2005. Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systematic Biology*, 54(4), 513-526.

Atkinson, Q., Nicholls, G., Welch, D., & Gray, R. 2005. From words to dates: water into wine, mathemagic or phylogenetic inference? Transactions of the Philological Society 103(2): p.193–219. http://dx.doi.org/10.1111/j.1467-968X.2005.00151.x

Bellwood, Peter, and E. Dizon. 2008. "Austronesian Cultural Origins: Out of Taiwan, via the Batanes Islands, and Onwards to Western Polynesia." In *Past Human Migrations in East Asia: Matching Archaeology, Linguistics, and Genetics*, edited by A Sanchez-Mazas, R. Blench, M. D. Ross, I. Peiros, and M. Lin, 23–39. London: Routledge.

Bender, B. W., Ward H. Goodenough, F. H. Jackson, J. C. Marck, Kenneth L. Rehg, H. Sohn, S. Trussel, and J. W. Wang. 2003. "Proto-Micronesian Reconstruction." *Oceanic Linguistics* 1 (42): 272–358.

Bergsland, K., & Vogt, H. 1962. On the validity of glottochronology. Current Anthropology 3(2): *p.115–153. www.jstor.org/stable/2739527*

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. (2017). *The AUTOTYP typological databases*. Version 0.1.0 https://github.com/autotyp/autotyp-data/tree/0.1.0

Birchall, Joshua, Michael Dunn, and Simon J. Greenhill. 2016. "A Combined Comparative and Phylogenetic Analysis of the Chapacuran Language Family." *International Journal of American Linguistics* 82 (3): 255–84.

Blust, R. 1974. *The Proto-North Sarawak Vowel Deletion Hypothesis*. University of Hawaii.

———. 1978. "Eastern Malayo-Polynesian: A Subgrouping Argument." In *Second International Conference on Austronesian Linguistics: Proceedings, Fascicle I, Western Austronesian*, edited by S.A. Wurm and L. Carrington, 181–234. Canberra: Pacific: Linguistics.

———. 1991. "The Greater Central Philippines Hypothesis." *Oceanic Linguistics* 30 (73): 129.

Blust, Robert. 1993. "Central and Central-Eastern Malayo-Polynesian." *Oceanic Linguistics* 32: 241–93.

———. 1999. "Subgrouping, Circularity and Extinction: Some Issues in Austronesian Comparative Linguistics." In *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, edited by Zeitoun E. and P J-K. Li, 31–94. Taipei, Taiwan: Symposium Series of the Institute of Linguistics, Academia Sinica.

———. 2000. "Why Lexicostatistics Doesn't Work: The 'Universal Constant' Hypothesis and the Austronesian Languages." In *Time Depth in Historical Linguistics*, edited by C Renfrew, A McMahon, and L Trask, 311–31. Cambridge: McDonald Institute for Archaeological Research.

———. 2008. “Is There a Bima-Sumba Subgroup?” *Oceanic Linguistics* 47 (1): 45–113. doi:10.1353/ol.0.0006.

———. 2013. *The Austronesian Languages*. Revised Ed. Canberra: Asia-Pacific Linguistics.

Bouckaert, R.R. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*. 26(10):1372-1373. doi: 10. 1093/bioinformatics/btq110. Epub 2010 Mar 12.

Bouckaert, Remco R., P. Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, a. J. Drummond, Russell D. Gray, M. a. Suchard, and Quentin D. Atkinson. 2012. “Mapping the Origins and Expansion of the Indo-European Language Family.” *Science* 337 (6097): 957–60. doi:10.1126/science.1219669.

Bowern, Claire, and Quentin D. Atkinson. 2012. “Computational Phylogenetics and the Internal Structure of Pama-Nyungan.” *Language* 88 (4): 817–45.

Bowern C. 2016. Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia. *Lang* Doc Conserv 10:1–44.

Burnham, Kenneth P. & David R. Anderson. 1998. *Model Selection and Inference: A practical information-theoretic approach*. New York: Springer.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. “Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis.” *Language* 91 (1): 194–244. doi:10.1353/lan.2015.0005.

Collins, J.T. 1982. “Linguistic Research in Maluku: A Report of Recent Fieldwork.” *Oceanic Linguistics* 21: 73–146.

Dahl, O.C. 1973. *Proto-Austronesian*. Vol. 15. *Sweden: Scandinavian Institute of Asian Studies Monograph Series.*

Dempwolff, O. 1927. “Das Austronesische Sprachgut in Den Melanesischen Sprachen.” *Folia Ethnoglossica* 3: 32–43.

Diamond, Jared M, and Peter Bellwood. 2003. “Farmers and Their Languages: The First Expansions.” *Science* 300 (5619): 597–603. doi:10.1126/science.1078208.

Dixon, R.M.W. 1997. The Rise and Fall of Languages. *Cambridge: Cambridge University Press.*

Dolgopolsky*, A.* 1990. More about the Indo-European homeland problem. *Mediterranean Language Review* 6–7: p.230–248.

Drummond, Alexei J., and Marc A. Suchard. 2010. “Bayesian Random Local Clocks, or One Rate to Rule Them All.” *BMC Biology* 8 (114): 114. doi:10.1186/1741-7007-8-114.

Drummond, Alexei J., Simon Y. W. Ho, Matthew J. Phillips, and Andrew Rambaut. 2006. “Relaxed Phylogenetics and Dating with Confidence.” *PLOS Biology* 4 (5): e88.

Duchêne, Sebastián, Robert Lanfear, and Simon Y W Ho. 2014. “The Impact of Calibration and Clock-Model Choice on Molecular Estimates of Divergence Times.” *Molecular Phylogenetics and Evolution* 78 (June): 277–89. doi:10.1016/j.ympev.2014.05.032.

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson, and Neele Becker. 2011. “Aslian Linguistic Prehistory: A Case Study in Computational Phylogenetics.” *Diachronica* 28 (3): 291–323. doi:10.1075/dia.28.3.01dun.

Dyen, Isidore. 1962. “The Lexicostatistical Classification of the Malayopolynesian Languages.” *Language* 38: 38–46.

Dyen, I., Kruskal, J.B., & Black, P. 1992. *An Indoeuropean Classification: A Lexicostatistical Experiment*. Philadelphia: American Philosophical Society. www.wordgumbo.com/ie/cmp/iedata.txt

Embleton, Sheila. 1986. *Statistics in Historical Linguistics*. Bochum: Studienverlag Brockmeyer.

Felsenstein, J. 1978. The number of evolutionary trees. *Systematic Zoology* 27: 27-33

Galucio, Ana Vilacy, Sérgio Meira, Joshua Birchall, Denny Moore, Nilson Gabas Júnior, Sebastian Drude, Luciana Storto, Gessiane Picanço, and Carmen Reis Rodrigues. 2015. "Genealogical Relations and Lexical Distances Within the Tupian Linguistic Family." *Boletim Do Museu Paraense Emilio Goeldi: Ciencias Humanas* 10 (2): 229–74. doi:10.1590/1981-81222015000200004.

Grace, G.W. 1959. *The Position of the Polynesian Languages Within the Austronesian (Malayo-Polynesian) Language Family. Memoir 16 of the International Journal of American Linguistics*. Indiana: Indiana University publications in anthropology; linguistics.

Gray, Russell D., Alexei J. Drummond, and Simon J. Greenhill. 2009. "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement." *Science* 323 (5913): 479–83. doi:10.1126/science.1166858.

Gray, R.D., Greenhill, S.J., and Bryant, D. (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society London, B,* 365, 3923-3933.

Gray, R.D. & Watts, J. (2017). Cultural macroevolution matters. *Proceedings of the National Academy of Sciences*, 114 (30) 7846-7852.

Green, Roger Curtis. 2003. "The Lapita Horizon and Traditions-Signature for One Set of Oceanic Migrations." In *Pacific Archaeology: Assessments and Anniversary of the First Lapita Excavation (July 1952)*, edited by C. Sand, 95–120. New Caledonia.

Greenhill, Simon J. 2015. "TransNewGuinea.org: An Online Database of New Guinea Languages." *PLoS ONE* 10 (10): 1–17. doi:10.1371/journal.pone.0141563.

Greenhill, Simon J., and Russell D. Gray. 2012. "Basic Vocabulary and Bayesian Phylolinguistics: Issues of Understanding and Representation." *Diachronica* 29 (4): 523–37. doi:10.1075/dia.29.4.05gre.

Greenhill, Simon J., Thomas E. Currie, and Russell D. Gray. 2009. "Does Horizontal Transmission Invalidate Cultural Phylogenies?" *Proceedings. Biological Sciences / the Royal Society* 276 (1665): 2299–2306. doi:10.1098/rspb.2008.1944.

Greenhill, Simon J., Alexei J Drummond, and Russell D. Gray. 2010. "How Accurate and Robust Are the Phylogenetic Estimates of Austronesian Language Relationships?" *PLoS ONE* 5 (3): e9573. doi:10.1371/journal.pone.0009573.

Greenhill SJ, Blust R, Gray RD (2008) The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics. *Evol Bioinform Online* 4:271–283.

Grollemund, Rebecca. 2012. "Nouvelles Approches En Classification: Application Aux Langues Bantu Du Nord-Ouest." PhD thesis.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. 2015. "Bantu Expansion Shows Habitat Alters the Route and Pace of Human Dispersals." *Proceedings of the National Academy of Sciences of the USA*. doi:10.1073/pnas.1503793112.

Haspelmath, M. 2005. The World Atlas of Language Structures, *Oxford Univ Press, Oxford*.

Heggarty, P. 2010. Beyond lexicostatistics: How to get more out of 'word list' comparisons. *Diachronica* 27(2): p.301–324. http://doi.org/10.1075/dia.27.2.07heg

Heggarty, P. 2014. Prehistory through language and archaeology. In C. Bowern & B. Evans (eds) Routledge Handbook of Historical Linguistics, 598–626. London: Routledge. https://www.academia.edu/3687718

Heggarty, P. forthcoming. Why Indo-European? Clarifying cross-disciplinary misconceptions on farming vs. pastoralism, *Journal of Indo-European Studies — special issue on Indo-European and Farming,* edited by G. Kroonen & B. Comrie.

Heggarty, P., & Renfrew, C. 2014. South and Island South-East Asia: Languages. In C. Renfrew & P. Bahn (eds) The Cambridge World Prehistory, 534–558. *Cambridge: Cambridge University Press*.

Ho, Simon Y. W., and Matthew J. Phillips. 2009. "Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times." *Systematic Biology* 58 (3): 367–80. doi:10.1093/sysbio/syp035.

Hymes, Dell. 1960. "Lexicostatistics so Far." *Current Anthropology* 1 (1): 3–44.

Kaufman, T., & Golla, V. 2000. Language groupings in the New World: their reliability and usability in cross-disciplinary studies. In C. Renfrew (ed) America Past, America Present: Genes and Languages in the Americas and Beyond, 47–57. *Cambridge: McDonald Institute for Archaeological Research*.

Kimura, M. 1968. "Evolutionary Rate at the Molecular Level." *Nature* 217: 624–26.

Kirsch, John A W. 1969. "Serological Data and Phylogenetic Inference: The Problem of Rates of Change." *Systematic Zoology* 18 (3): 296–311.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa, and Connie J Mulligan. 2009. "Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the Near East." *Proceedings of the Royal Society B: Biological Sciences* 270 (1668): 2703–10. doi:10.1098/rspb.2009.0408.

Ko, Albert Min Shan, Chung Yu Chen, Qiaomei Fu, Frederick Delfin, Mingkun Li, Hung Lin Chiu, Mark Stoneking, and Ying Chin Ko. 2014. "Early Austronesians: Into and Out of Taiwan." *American Journal of Human Genetics* 94 (3). The American Society of Human Genetics: 426–36. doi:10.1016/j.ajhg.2014.02.003.

Kroeber, A L. 1955. "Linguistic Time Depth Results so Far and Their Meaning." *International Journal of American Linguistics* 21 (2): 91–104.

Lanfear, Robert, John J Welch, and Lindell Bromham. 2010. "Watching the Clock: Studying Variation in Rates of Molecular Evolution Between Species." *Trends in Ecology & Evolution* 25 (9). Elsevier Ltd: 495–503. doi:10.1016/j.tree.2010.06.007.

Lees, R B. 1953. "The Basis of Glottochronology." *Language* 29 (2): 113–27.

Lewis, Paul M., ed. 2009. *Ethnologue: Languages of the World*. 16th ed. Dallas, Texas: SIL International.

List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Philippe Lopez & Eric Bapteste. 2016. Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics. *Biology Direct* 11(39). 1–17.

Lynch, J. 2001. The Linguistic History of Southern Vanuatu*. Canberra: Pacific Linguistics*.

Mallory, J.P. 1989. In Search of the Indo-Europeans. London: Thames & Hudson.

McMahon, A.M.S., & McMahon, R. 2006. Why linguists don't do dates. In P. Forster & C. Renfrew (eds) Phylogenetic Methods and the Prehistory of Languages, 153–160. *Cambridge: McDonald Institute for Archaeological Research*.

Mead, D. 2003. "Evidence for a Celebic Supergroup." In *Issues in Austronesian Historical Phonology*, edited by John Lynch, 115–41. *Canberra: Pacific Linguistics*.

Meireles, Denise Maldi. 1989. *Guardiães da Fronteira: Rio Guaporé, século XVIII*. Petrópolis: Vozes.

Michael, Lev, Natalia Chousou-polydouri, Keith Bartolomei, Erin Donnelly, Vivian Wauters, and Zachary O Hagan. 2015. "A Bayesian Phylogenetic Classification of Tupi-Guarani." *LIAMES* 15 (2): 1–36.

Moore, J. H. 1994. "Putting Anthropology Back Together Again: The Ethnogenetic Critique of Cladistic Theory." *American Anthropologist* 96 (4): 925–48.

Moran S., McCloy D., Wright R. (2014) *PHOIBLE Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig).

Murdock, George Peter. 1964. "Genetic Classification of the Austronesian Languages: A Key to Oceanic Culture History." *Ethnology* 3 (2): 117–26. doi:10.2307/3772706.

Nakhleh, L., Ringe, D., & Warnow, T. 2005. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. Language 81(2): p.382–420. www.jstor.org/stable/4489897

Nascimento, F.F., Reis, M.d., and Yang, Z. (2017). A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*, 1, 1446–1454

Nicholls, Geoff K., and Russell D. Gray. 2006. "Dated Ancestral Trees from Binary Trait Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (3): 545–66. doi:10.1111/j.1467-9868.2007.00648.x.

Oppenheimer, Stephen J., and Martin B Richards. 2001. "Fast Trains, Slow Boats, and the Ancestry of the Polynesian Islanders." *Science Progress* 84 (3): 157–81.

Ozanne-Rivierre, F. 1992. "The Proto-Oceanic Consonantal System and the Languages of New Caledonia." *Oceanic Linguistics* 31: 191–207.

Pawley, A. 1972. "On the Internal Relationships of Eastern Oceanic Languages." In *Studies in Oceanic Culture History*, edited by R.C. Green and M. Kelly, 13:3–106. Honolulu: Bernice P. Bishop Museum.

Pawley, Andrew. 1966. "Polynesian Languages: A Subgrouping Based on Shared Innovations in Morphology." *Journal of the Polynesian Society* 75 (1): 39–64.

———. 2002. "The Austronesian Dispersal: Languages, Technologies and People." In *Examining the Farming/Language Dispersal Hypothesis*, edited by P Bellwood and Colin Renfrew, 251–73. Cambridge: McDonald Institute for Archaeological Research.

Penny, David, B J McComish, M A Charleston, and Michael D Hendy. 2001. "Mathematical Elegance with Biochemical Realism: The Covarion Model of Molecular Evolution." *Journal of Molecular Evolution* 53 (6): 711–23.

Ringe, D.A., Warnow, T., & Taylor, A. 2002. Indo-European and computational cladistics. Transactions of the Philological Society 100(1): p.59–129. http://dx.doi.org/10.1111/1467-968X.00091

Robinson, Laura C, and Gary Holton. 2012. “Internal Classification of the Alor-Pantar Language Family Using Computational Methods Applied to the Lexicon.” *Language Dynamics and Change* 2: 1–27. doi:10.1163/22105832-20120201.

Ross, M. 1988. *Proto Oceanic and the Austronesian Languages of Western Melanesia*. Canberra: Australian National University; Pacific Linguistics.

Ross, Malcolm, and Åshild Næss. 2007. “An Oceanic Origin for Äiwoo, a Language of the Reef Islands.” *Oceanic Linguistics* 46: 456–98.

Sanderson, Michael J. 1997. “A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy.” *Molecular Biology and Evolution* 14 (12): 1218–31.

Sheppard, Peter J, Scarlett Chiu, and Richard Walter. 2015. “Re-Dating Lapita Movement into Remote Oceania.” *Journal of Pacific Archaeology* 6 (1): 26–36.

Sicoli, Mark A., and Gary Holton. 2014. “Linguistic Phylogenies Support Back-Migration from Beringia to Asia.” *PloS One* 9 (3): e91722. doi:10.1371/journal.pone.0091722.

Sims-Williams, P. 1998. Genetics, linguistics, and prehistory: thinking big and thinking straight. Antiquity 72(277): p.505–527.

Soares, P, J A Trejaut, J H Loo, Catherine Hill, M Mormina, C L Lee, Y.M. Chen, et al. 2008. “Climate Change and Postglacial Human Dispersals in Southeast Asia.” *Molecular Biology and Evolution* 25 (6): 1209–18.

Sokal, Robert R, and Charles D. Michener. 1958. “A Statistical Method for Evaluating Systematic Relationships.” *The University of Kansas Science Bulletin* 38 (22): 1409–38.

Swadesh, M. 1952. “Lexico-Statistic Dating of Prehistoric Ethnic Contacts.” *Proceedings of the American Philosophical Society* 96 (4): 452–63.

———. 1955. “Towards Greater Accuracy in Lexicostatistic Dating.” *International Journal of American Linguistics* 21 (2): 121–37.

Swadesh, Morris. 1950. “Salish Internal Relationships.” *International Journal of American Linguistics* 16 (4): 157–67. doi:10.1086/464084.

Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski, and Niklas Wahlberg. 2013. “Shedding More Light on Language Classification Using Basic Vocabularies and Phylogenetic Methods: A Case Study of Uralic.” *Diachronica* 30 (3): 323–52. doi:10.1075/dia.30.3.02syr.

Tadmor, U., Haspelmath, M., & Taylor, B. 2010. Borrowability and the notion of basic vocabulary. *Diachronica* 27(2): p.226–246. http://dx.doi.org/10.1075/dia.27.2.04tad

Terrell, J. 1988. “History as a Family Tree, History as an Entangled Bank: Constructing Images and Interpretations of Prehistory in the South Pacific.” *Antiquity* 62: 642–57.

Thurgood, G. 1999. From Ancient Cham to Modern Dialects. 28. *Hawaii: Oceanic Linguistics Special Publications*.

Tryon, D. T. 1976. New Hebrides Languages: An Internal Classification. *Canberra: Pacific Linguistics.*

Welch, John J, and Lindell Bromham. 2005. “Molecular Dating When Rates Vary.” *Trends in Ecology & Evolution* 20 (6): 320–7. doi:10.1016/j.tree.2005.02.007.

Wilmshurst, Janet M, Terry L Hunt, Carl P Lipo, and Atholl J Anderson. 2011. “High-Precision Radiocarbon Dating Shows Recent and Rapid Initial Human Colonization of East Polynesia.”

Proceedings of the National Academy of Sciences of the United States of America 108 (5): 1815–20. doi:10.1073/pnas.1015876108.

Yang, Z. 1993. “Maximum-Likelihood Estimation of Phylogeny from Dna Sequences When Substitution Rates Differ over Sites.” *Molecular Biology and Evolution* 10 (6): 1396–1401.

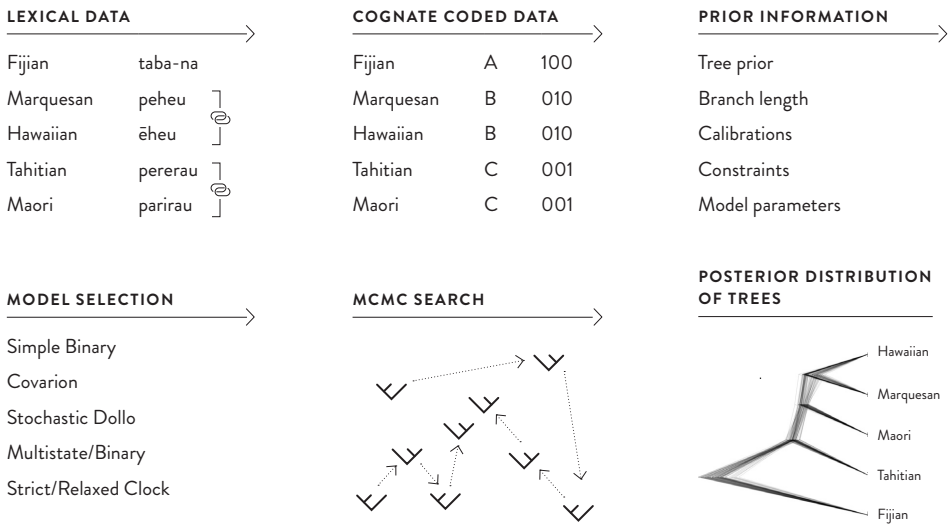
Zuckerkandl, E., and L. Pauling. 1965. “History of Evolutionary Molecules as Documents.” *Journal of Theoretical Biology* 8

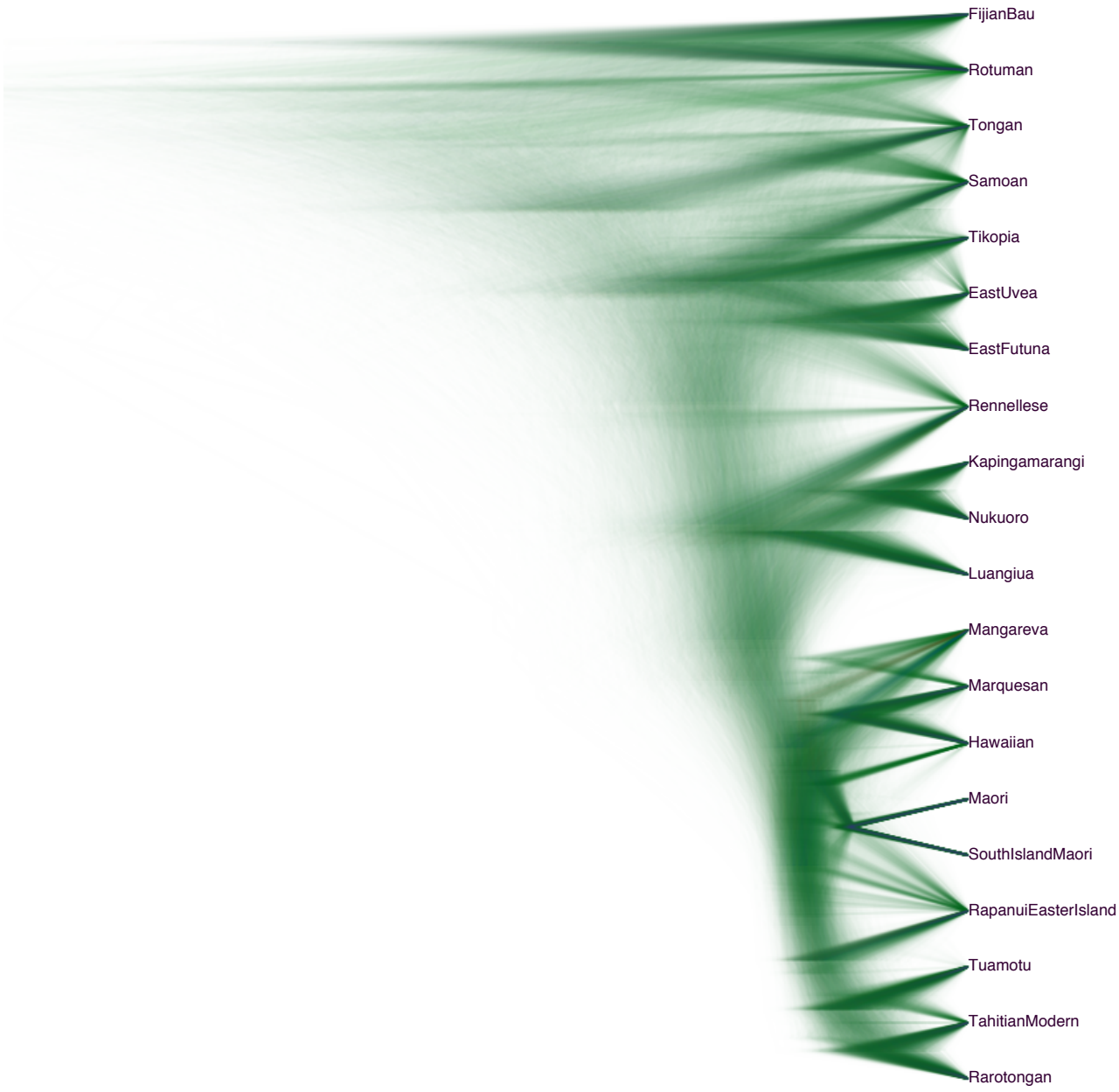
FIGURES

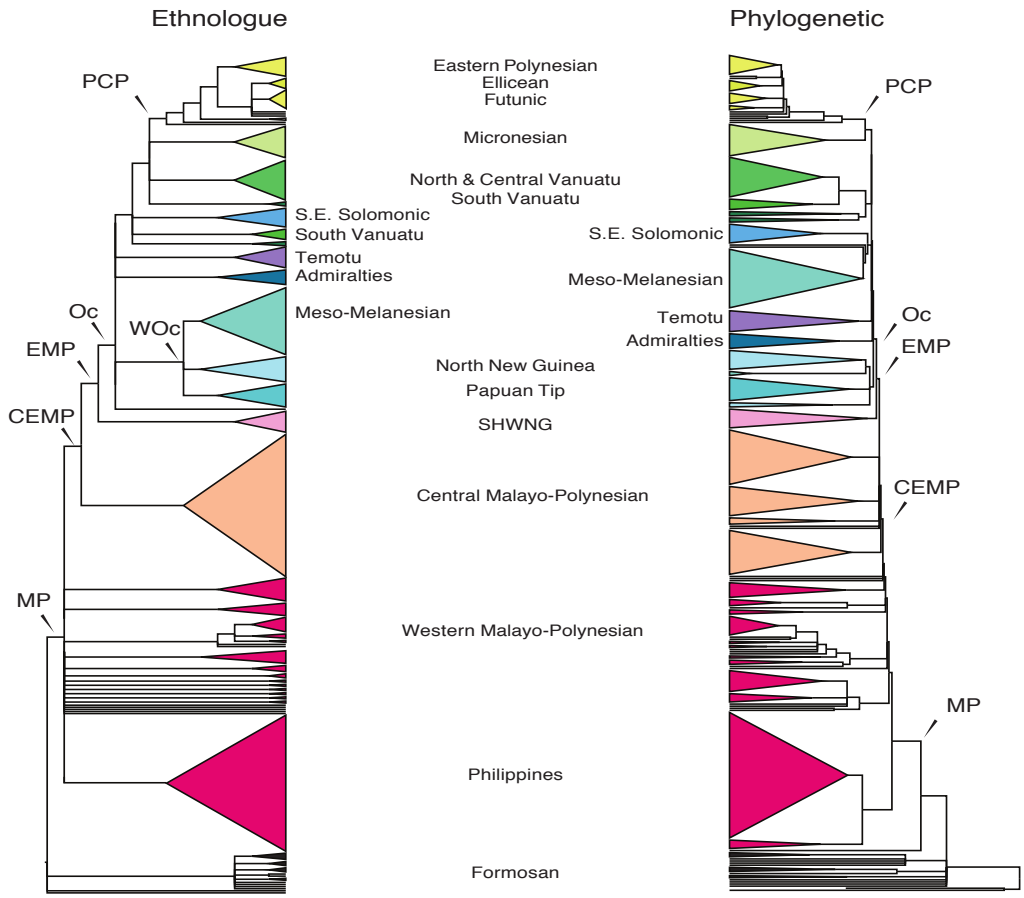
Figure 1. The steps involved in a Bayesian phylogenetic analysis of lexical data. First the lexical data is cognate coded, then the cognate sets are expressed as either a multistate or binary matrix, then the prior information for the Bayesian analysis is specified, then the stochastic model of character change is selected, then multiple MCMC searches are run, and finally the resulting posterior distribution of trees and their associated parameters are summarized.

Figure 2. A densitree of Central Pacific languages constructed from the posterior distribution of trees from a Bayesian analysis of basic vocabulary. Note that the densitree reveals considerable conflicting signal that can not be captured in single tree.

Figure 3. A comparison of the Austronesian phylogeny (Maximum Clade Credibility tree from a Bayesian analysis) from Gray et al. (2009) vs. the "known" subgrouping from Ethnologue. The Bayesian estimate recovers most of the major subgroups in the Ethnologue classification, and the mismatches mainly occur where the putative subgroups are not broadly accepted. Both trees differ enormously from lexicostatistical estimates of Austronesian relationships (see Greenhill and Gray 2012).




FijianBau
Rotuman
Tongan
Samoan
Tikopia
EastUvea
EastFutuna
Rennellese
Kapingamarangi
Nukuoro
Luangiua
Mangareva
Marquesan
Hawaiian
Maori
SouthIslandMaori
RapanuiEasterIsland
Tuamotu
TahitianModern
Rarotongan



View publication stats