DISCUSSION NOTES

Basic vocabulary and Bayesian phylolinguistics

Issues of understanding and representation*

Simon J. Greenhill and Russell D. Gray
Australian National University / University of Auckland

Donohue et al.'s critique of our work on the origins and spread of the Austronesian language family is marred by misunderstandings of our approach. We respond to these by noting that our Bayesian phylogenetic approach: (1) distinguishes between retentions and innovations probabilistically, (2) focuses on basic vocabulary not 'the lexicon', (3) eliminates known loanwords, (4) produces results that are congruent with the results of the comparative method and conflict with the scenarios requiring unprecedented amounts of language shift postulated by Donohue et al.

Introduction

Every field has well-known traps. In historical linguistics, researchers are frequently accused of failing to distinguish retentions from innovations, failing to base inferences on regular sound change, and confusing loanwords with cognates (Campbell & Poser 2008). Donohue, Denham & Oppenheimer's (this issue) critique of our recent work on the origins and spread of the Austronesian language family (Gray, Drummond & Greenhill 2009) pushes these buttons. Our study tested different scenarios for the expansion of the Austronesian language family, and found overwhelming support for an Austronesian origin in Taiwan around 5,200 years ago followed by a series of expansion pulses and pauses. We found no evidence for alternative scenarios advanced by Oppenheimer & Richards 2001, Donohue & Denham 2010, and Soares et al. 2011. We outline Bayesian phylogenetic methodology and respond to seven issues raised by Donohue et al.

* We thank Bob Blust, Claire Bowern, Lyle Campbell, Sheila Embleton, Emily Gasser, Anthony Grant, Andy Pawley, Malcolm Ross, Joseph Salmons and two anonymous reviewers for comments and discussion on this paper.

Diachronica 29:4 (2012), 523–537. DOI 10.1075/dia.29.4.05gre
ISSN 0176–4225 / E-ISSN 1569–9714 © John Benjamins Publishing Company

Bayesian phylogenetic inference

In a phylogenetic analysis, the data are treated as a fixed observation and the aim is to find a model that explains the current pattern of data well. The second component of a phylogenetic analysis, therefore, is a model of how these data could have arisen. To model lexical change we make simplifying assumptions about the processes involved. A simple model would allow cognates to be gained and lost at the same rate over time and across lineages. This rate is therefore a *parameter* of the model and its value is estimated as part of the analysis. A more complex model could add a parameter to allow cognates to be gained at a different rate to cognates being lost. Yet more complex models could be constructed to account for the substantial variation in rates across semantic domains by allowing some cognates to change faster than others, mimicking what we know about how rates of change vary by part of speech or semantic field. Another approach would be to allow cognates to switch between faster and slower rates on different branches of the tree using the covarion model (Penny et al. 2001).

Whilst language change is complex, the model's simplicity does not necessarily discredit or invalidate the methodology. Constructing a model is a trade-off between over- / under-fitting parameters (Burnham & Anderson 1998). As more parameters are added to the model, the fit to the data will improve especially if these parameters capture an important aspect of language change. However, as parameters are added, increasingly more data is required to accurately estimate the values of the model parameters. If there are too many parameters for the data to describe adequately, the model is *over-parameterised* and the estimated values of the parameters become unreliable (i.e. sampling error increases). Therefore, our aim is not to construct a complex model that captures every aspect of language change, but rather to construct the simplest model that provides the best estimates of the parameters with finite amounts of data. In our Austronesian analyses the covarion model was the best fitting model.

Given a model, data, and a tree, we can calculate a numeric score that quantifies the fit of the data to that model called the *likelihood*. We use this likelihood value to find the tree(s) that explain the data well (i.e. the best family tree). First we start with a random tree and calculate the likelihood of the data given the model on that tree. Second, we randomly permute the tree in some way, for example, by changing the tree's topology (i.e. the relationships specified by the tree), or by altering the branch lengths, or by modifying the model's rate parameter. In maximum likelihood phylogenetic inference, if the likelihood of this new tree is *worse* than the previous tree, we throw away the new tree, but if the likelihood is *better* we keep the new tree. By repeating this process many times we can attempt to find the single best tree i.e. the *maximum likelihood* tree. Bayesian phylogenetic inference

— the approach we used — extends maximum likelihood to search through the possible trees and parameters and samples the trees in proportion to their *posterior probability* calculated from the tree and the parameter estimates given the data, the model and the initial starting values of the parameters (priors). Greenhill & Gray (2009) provide a more detailed account of this method and its application to linguistics. Bayesian tree construction methods are now the approach of choice in evolutionary biology (Huelsenbeck et al. 2001).

The Bayesian approach returns a set of trees rather than a single tree thus enabling us to measure the statistical uncertainty in our estimates (Huelsenbeck et al. 2000). This means that we can quantify the support for particular subgrouping hypotheses according to their posterior probability ranging from 0 (no support) to 1.0 (complete support). For example, our results show the Oceanic subgroup had a posterior probability of 1.0, 100% of the trees in the sample contained this subgroup. In contrast, the Central Maluku subgroup proposed by Collins (1982) was only supported by 0.54 of the sampled trees, indicating weak support for this proposal. Unfortunately, Donohue et al. treat the consensus tree we reported as a single tree rather than as a visual summary of the posterior distribution of trees, where the posterior probabilities on each node indicate the degree of support.

Issue 1. Failure to distinguish innovations from retentions

Donohue et al. claim that our method does not discriminate between shared retentions and shared innovations. Phylogenetic approaches grew directly out of the cladistics revolution started by Hennig (1966) who clearly distinguished shared retentions ('symplesiomorphies') from shared innovations ('synapomorphies'). Hennig argued that only synapomorphies are diagnostic of monophyletic (i.e. minimal) groups. Phylogenetics, like historical linguistics post Brugmann (1884), has followed that logic ever since. Modern Bayesian phylogenetic methods do not require an *a priori* distinction between retentions and innovations; rather these are inferred probabilistically in the analysis, an outcome rather than an input into the analysis. Similarly, the distinction between innovations and retentions is not a given in the comparative method, but rather an *outcome* of subgrouping hypotheses, as different subgroupings make different claims about innovations.

Here we briefly explain how phylogenetic methods make this distinction, and a more detailed worked example is available in the supplement (Supp. Figure 1). In phylogenetic inference, searches typically return unrooted trees. However, if we can root these trees, then we can infer the *directionality* of character state changes on the tree and thus infer whether a cognate set is a retention or an innovation (again, see supplementary material). Directionality can either be specified

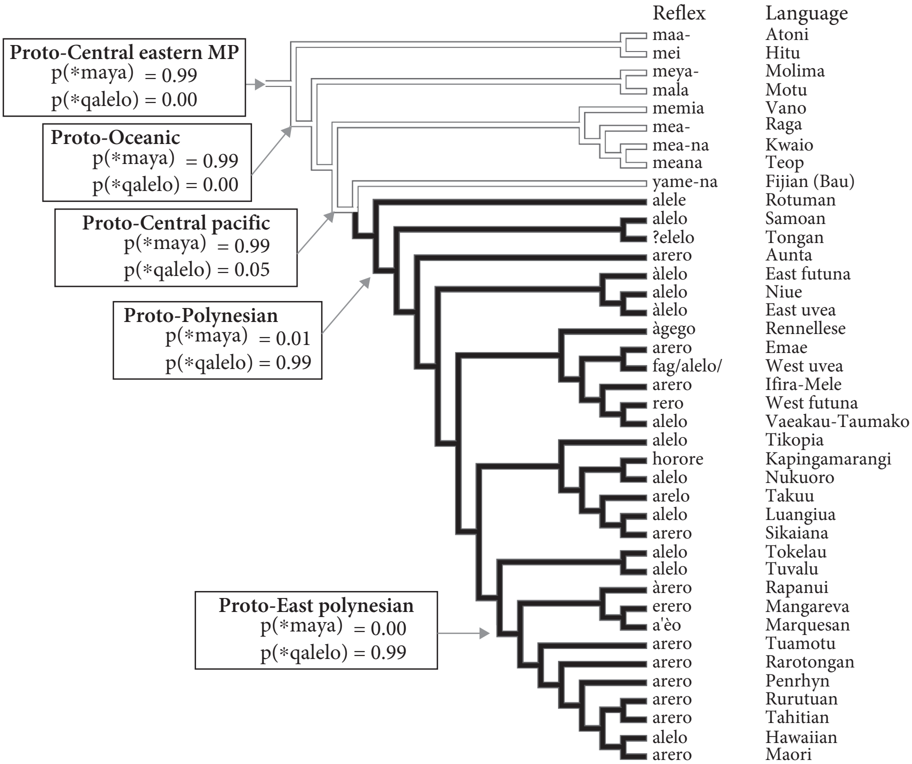

Figure 1. The probabilistic inference of lexical innovations, showing the distribution of glosses of the cognate sets *maya and *qalelo "tongue" on a rooted language tree. When these are fitted onto the tree using a model of state change that allows cognates to be gained and lost at different rates (Lewis 2001), there is a 0.99 probability that *qalelo was innovated on the branch between Fijian and the Polynesian language plus Rotuman. There are very low probabilities that this innovation occurred earlier or later in the tree. These probabilistic inferences fit closely with those of the comparative method (Pawley 2010, Greenhill & Clark 2011).

manually using an outgroup, or by using a model that infers directionality (such as one that has a different rate of cognate gains to losses). We used both in Gray et al. (2009). Far from ignoring the distinction between innovations and retentions, our methodology is based on inferring them (Figure 1 & Supp. Figure 1).

Issue 2. Information about sound changes is completely neglected

Donohue et al. state that information about sound changes is "completely neglected" in our approach, implying we have not followed the comparative method, even likening it to Greenberg's mass comparison. Regular sound changes form the basis of the cognate judgements in our database, which reflect the combined input of Robert Blust, John Lynch, Jeff Marck, Malcolm Ross, Laurent Sagart, and many others (Greenhill et al. 2008). We are grateful for the huge input the database has received from the Austronesian linguistic community and reject the claim that these scholars have "completely neglected" sound change in making cognate judgements.

Is there any sense in which we neglect sound changes? As mentioned above, model-based phylogenetic inference requires strategic choices about the relative complexity of the model. We have chosen to focus on tractable, robust models of cognate evolution rather than modelling all aspects of language change. Our aim was to test hypotheses about the spread and timing of the Austronesian languages. To achieve this aim we chose to use some relatively simple — and hence tractable — models of gains and losses in cognate sets (Greenhill & Gray 2009). These models have been shown to make accurate inferences about linguistic subgroups (Gray & Atkinson 2003, Greenhill et al. 2010a). We decided not to model sound change *within* cognates, nor to simultaneously model both cognate evolution and sound change. Modelling sound change would require a vast increase in the number of parameters to be estimated. Many attempts use a simple model of sound change, the Levenshtein distance (e.g. Holman et al. 2011), but in our investigations this method was highly inaccurate (Greenhill 2011). Indeed, the Levenshtein-based classifications of Austronesian (Petroni & Serva 2008) spectacularly failed to recover anything close to the traditional Austronesian tree and rooted the family in Near Oceania (Greenhill 2011). More complex models of sound change are an interesting area of future research (e.g. Bouchard-Côté et al. 2009), but such models are not required to infer linguistic relationships accurately.

Issue 3. Lexical borrowing overwhelms genealogical signal

Donohue et al. state "lexical items are widely recognised to be the elements of a language most prone to diffusion, and the least reliable (in the absence of regular sound correspondences) to determine phylogenetic relationships". This may be true but it does not follow that *all* lexical items are highly borrowable. Other aspects of language such as morphological paradigms and phonological innovations can provide robust evidence for subgrouping (Durie & Ross 1996), but many of the major Austronesian subgroups are largely defined by lexical innovations (Eastern Malayo-Polynesian shows 56 lexical innovations (Blust 2009), or are well supported by lexical innovations (e.g. Central Pacific, Admiralties, Central Malayo-Polynesian, others). Moreover, we did not analyse 'the lexicon', but rather carefully selected basic vocabulary items appropriate for Austronesian languages

(Blust 1981, 2000) and relatively resistant to borrowing (Embleton 1986). We removed any identified loanwords, and used methods that are robust to the effects of borrowing. Phylogenies built from basic vocabulary in this way are highly congruent with the traditional subgroupings proposed by historical linguistics e.g. Aslian (Dunn et al. 2011b), Austronesian (Gray et al. 2009, Greenhill et al. 2010a), Bantu (Holden 2002), Indo-European (Gray & Atkinson 2003), Japonic (Lee & Hasegawa 2011), Semitic (Kitchen et al. 2009), and Uto-Aztecan (Dunn et al. 2011a).

Donohue et al. write: "Studies of basic vocabulary in some of the Austronesian languages considered by Gray et al. show up to 48% borrowing rates in the basic vocabulary (e.g. Grant 2005)". This implies that 48% is common, but Grant (2005) only identifies 40% borrowing, and only in one language Jarai — a language not included in our analyses. Grant calls this 40% "astoundingly high, and is almost unparalleled in the record of the world's languages" (2005: 54).

The critical issue is not whether there are high levels of borrowing, but whether there are high levels of borrowing in our sample of Austronesian basic vocabulary. Three recent studies shed light on this. The first, Tadmor et al. (2010), surveyed languages for the World Loanword Database project. The 41 languages sampled are biased towards those known to have many loanwords. On average, they showed 24% borrowing across a wide sample of the lexicon with basic vocabulary more resistant to borrowing. The second, by Nelson-Sathi et al. (2011), demonstrates that borrowing levels in the Swadesh basic vocabulary of Indo-European languages is, on average, around 8% (with 7% variation between languages). The third, by Bowern and colleagues (2011), surveyed 122 languages spoken in northwest Amazonia, northern Australia, and California and the Great Basin. These languages are spoken in small-scale hunter-gatherer societies, and in regions commonly thought to have high levels of borrowing. However, these languages show an average of 5% borrowings in basic vocabulary — far less than 48%.

So what are the likely levels of borrowing in our data set? The World Loanword Database (Haspelmath & Tadmor 2009) provides comprehensive borrowability statistics from words in 41 languages. For each word, the database gives a 'Borrowing score' that encodes how likely it is that the word is borrowed in any of those languages. This score ranges from 0 (No evidence for borrowing) to 1 (Clearly borrowed). If the words in our analyses were a poor choice we would expect them to score above 0.5. Instead, plotting the borrowing scores of the 210 items in the Austronesian Basic Vocabulary Database shows that, overwhelmingly, these items score in the 'no borrowing' to 'very little evidence of borrowing ranges' with an average of 0.13 (Supp. Figure 2).

Finally, Donohue et al. fail to mention that in compiling our database we spent considerable time identifying loanwords and that we removed any identified loans in our analyses. Moreover, we did not include languages that were known to have high levels of borrowing. Naturally in a database of this size some loanwords were likely overlooked. However, we have tested the robustness of our methods on different levels of *undetected* borrowing (Greenhill et al. 2009). In a series of simulation studies we showed that the estimates of tree topology and time-depth were very robust despite quite high levels of borrowing between languages, around 20% of basic vocabulary every 1000 years (Greenhill et al. 2009). This is 20% *unidentified* borrowings per 1000 years — removing identified loans in the database from the analysis makes this threshold much higher. When Donohue et al. state that "Gray et al.'s method is as likely to detect the cumulative historical effects of lexical borrowing … as it is to detect historical developments resulting from original differentiation from a proto-language", are they claiming that rates of borrowing in basic vocabulary are at least as high as rates of inheritance?

Issue 4. There are clear discrepancies between the phylogeny and expected language relationships in the placement of individual languages

The major features of our trees are congruent with the results of the comparative method (see Figure 2 and Supp. Text 1). However, Donohue et al. do not discuss this but rather focus in their supplement on the placement of a small number of languages they claim are incorrectly subgrouped. Their identification of these misplacements is nothing new — we discussed these in detail in Gray et al. 2009, and followed that paper with a detailed discussion of these problematic languages (Greenhill et al. 2010a). We provide another thorough analysis of the consistencies and discrepancies in the supplementary material (Supp. Text 1 & 2). To summarise, Donohue et al. critique the placement of 37 of the 400 languages in our trees. Of the 37, three are mistakes from misreading the tree (Bima, Malayo-Chamic, Paiwan). Four of the other putative misplacements reflect long-standing classification difficulties or potential subgroupings (Irarutu, Kei, Maloh, Mussau). This leaves 30 of 400 languages that might be misplaced due to unidentified borrowings or a lack of distinguishing lexical innovations: a grand total of 7.5%. With 400 languages there are 5.8×10^{984} possible rooted bifurcating trees: more trees than there are atoms in the universe. With finite amounts of data it is simply not realistic to expect to recover every single branching point in a tree of 400 languages. Some lack of resolution and minor misplacement of taxa is expected even with large datasets and good models. Pointing out minor misplacements of individual languages does not invalidate the rest of that subgroup. A parallel can be drawn between early classifications of Indo-European that incorrectly placed Armenian within Indo-Iranian (Hübschmann [1875] 1967), a misplacement which did not invalidate either Indo-Iranian or Indo-European.

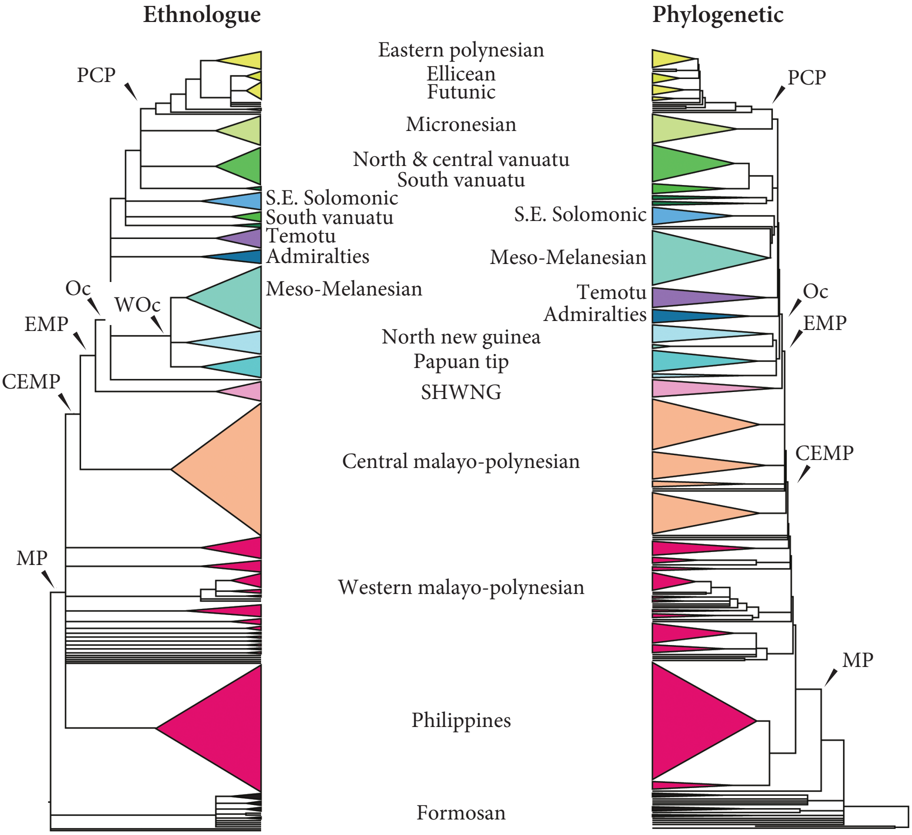

Figure 2. Similarity between the comparative method tree from *Ethnologue* (left) and Gray et al.'s phylogenetic trees (right). Major subgroups are labelled and color-coded. See the supplementary material for details on points of congruence between these two trees and the lack of congruence with a lexicostatistical tree. Abbreviations: CEMP (Central-Eastern Malayo-Polynesian), EMP (Eastern Malayo-Polynesian), MP (Malayo-Polynesian), Oc (Oceanic), PCP (Proto-Central Pacific), SHWNG (South Halmahera-West New Guinea), & WOc (Western Oceanic).

Issue 5. Polynesian subgroupings are "completely scrambled"

Donohue et al. focus on Polynesian as an example of how our method fails. They claim that our results are at odds with the accepted Polynesian subgrouping and that we replicate almost none of the internal subgroups. According to Donohue et al. Polynesian is the "least controversial subgroup within Austronesian" and is "perhaps the most studied and best-understood part of the Austronesian tree". However, the integrity of the Polynesian group itself is not at issue. The issue is the internal classification of the Polynesian languages. Despite the uncontroversial nature of the Polynesian subgroup as a whole (e.g. Marck 2000, Pawley 2010), the internal classification of Polynesian remains an active area of debate (e.g. Pawley 2009, 2010, Wilson 2010). In fact, the only Polynesian subgroups that are relatively uncontroversial are Eastern Polynesian and Central Eastern Polynesian, both of which are replicated in our trees. The proposal for an 'Ellicean' subgroup placing Tuvaluan with the Northern Outliers (Howard 1981, Marck 2000, Pawley 1967) has been severely criticised (Wilson 2010), nor is the Marquesic group widely accepted (e.g. Wilson 2010).

Donohue et al. imply that our subgrouping of Tongan with Samoan reflects ongoing borrowing between these languages, but Pawley (2010) finds very little evidence for this (possibly 4 borrowings on the 200-item Swadesh list). They also state that Tongan and Samoan show "higher than normal percentages of lexical retentions from Proto-Polynesian", but again, this is not so. Tongan and Samoan have retention rates of 41.5% and 38% in basic vocabulary. Tuvaluan (43.5%), Mele-Fila (45%), Sikaiana (45%), and Takuu (43.5%) all have higher retention rates (Pawley 2010). Tongan and Samoan are well within the 'normal' retention ranges found in Western Polynesian (36–43.5%).

The claim that our results "completely scramble" Polynesian is incorrect. Our analyses recovered the Proto Polynesian and Proto Eastern Polynesian nodes with posterior probabilities of 1.0. Often in places where there are apparent points of incongruence between our trees and Marck's (2000), the posterior probabilities are low and some of the set of most probable trees are actually congruent at that point (Supp. Figure 3). For example, while Rapanui is grouped with Mangareva and the Marquesas in 0.56 of the trees, it falls outside this group in a way that is congruent with Marck's tree in the remaining 0.44. That is, according to our data, the support for those two arrangements is about equally probable. Therefore our analyses do not reject Marck's tree but rather highlight conflicting signal in the data. The problem that any tree building method encounters with the Polynesian languages is that many of the subgroups have broken up in a series of interlocking dialect chains (e.g. Geraghty 1983, Marck 2000, Pawley 1967, 2009, 2010). A major advantage of Bayesian methods over simple approaches reporting a single tree is that the conflicting signal caused by the breakup of dialect networks can be reflected in the posterior distribution of trees (Supp. Figure 3). Because of the complex history of Polynesian, we have used phylogenetic network methods to further investigate the conflicting signal in these languages (Gray et al. 2010). Our analyses found that the Polynesian languages have strikingly high levels of conflicting signal, even in their basic vocabulary (Supp. Figure 4).

Issue 6. The degree of similarity between our results and the comparative method is overstated

Donohue et al. present a tree of 21 Indo-European languages that, despite some notable misplacements, replicates the accepted subgroupings with an accuracy of 84%. Donohue et al. use this comparison to argue that the 81% accuracy of Gray et al.'s 400 languages trees is very poor. However the logic of this argument is poor. Comparing the congruence between relatively clock-like small trees vs. congruence between big trees with extreme rate variation is comparing apples to oranges. Tree structures are complicated things to compare: for 21 languages there are 3.2×10^{23} possible rooted bifurcating trees. With 400 languages there are 5.8×10^{984} possible trees. Picking languages one thinks are misplaced is not an appropriate way to compare trees (Steel & Penny 1993). Rather than focusing on individual languages, phylogeneticists have developed a suite of tree comparison metrics to quantify the extent of differences between trees. The quartets distance is a standard tree-comparison metric that measures the number of different combinations of four language subsets in both trees (Steel & Penny 1993). This score is normalized across all quartets in the tree, and will range from 0 with identical trees to 1 for maximally different trees.

To show how inaccurate this comparison is we use the quartets method to quantify the similarity between the trees. First, we modified the original Gray et al. (2009) Austronesian maximum clade credibility tree (a single-tree summary of the posterior tree distribution) to match the 34 misplaced languages in Donohue et al.'s supplement, reflecting the generally-accepted subgroupings. This 'revised' tree is actually very close to the original tree, with a quartets distance of only 0.03 (Supp. Figure 5b). Second, we systematically calculated the normalized quartets distance between the revised tree to each of the 4,200 trees in the Gray et al. posterior probability distribution. Again, the revised tree is actually very similar to the phylogenetic trees with a mean normalized quartets score of 0.15 (Supp. Figure 5d). By comparison, the normalized quartets distance between Donohue's Indo-European tree and accepted IE subgroupings is 0.41 (Supp. Figure 5a), indicating that Donohue et al.'s tree is substantially different from the accepted subgrouping.

The large variation in retention rates found in Austronesian languages led to one of the great failures of lexicostatistics — the tree rooted Austronesian in Island Melanesia rather than Taiwan (Dyen 1962, Greenhill & Gray 2009). We compared the fit of a tree built with lexicostatistical methods (Supp. Fig 6, Supp. Text 3) to the revised Austronesian tree and found that the two were very different at 0.44 (Supp. Figure 5c). This result demonstrates that the Bayesian phylogenetic tree is much closer to the expected phylogeny then the lexicostatistical tree. Finally, if our results completely ‘scrambled’ the expected phylogeny then the distance between the revised tree and a completely random set of trees would be low. To test this, we calculated the distance between the revised tree and 1000 trees generated randomly (R Development Core Team 2011, Paradis et al. 2004). Contrary to the ‘scrambling’ hypothesis, the mean distance between the random trees and the revised tree is actually very large: 0.68 (Supp. Fig 5e). In summary, there is striking congruence between our results and those of the comparative method. The putatively misplaced languages in our trees are only a small subset of the total 400, and the alleged misplacements do not affect our central findings about the Taiwanese rooting and chain-like expansion sequence revealed in our trees.

Issue 7. The trees represent “distance-decay and local borrowing” rather than phylogeny

Donohue et al. claim that our results simply reflect geography, or “distance-decay”, arguing that the effects of local borrowing and distance-decay “undermine [our] whole lexical approach as a means to replicate linguistic families and their subgroups”. However, although closely related languages are often likely to be geographically proximate, the structure of our trees fits the expected Austronesian topology better than a simple distance-decay pattern. If the trees reflected geography rather than genealogy, why is Bima more closely related to the Bird’s Head SHWNG languages (~1900 km away) than its geographical neighbours Bali and Sasak (~300 km away)? Why are the languages of the Philippines more closely related to the languages of Polynesia then to Taiwan? Similarly, if the signal was just geography Maori should group with the languages of New Caledonia rather than Eastern Polynesian languages, and the Chamic languages would be at the base of the trees.

Donohue et al. select examples of borrowing between geographical neighbours, but this only affects a small proportion of the tree (§§4–5, Supp Text 1 & 2). To systematically evaluate the claim that the signal in basic vocabulary data is primarily geographical rather than genealogical we used a Mantel test (Mantel 1967), which estimates the correlation between two distance matrices. The results showed significant correlation between geography and phylogeny in these languages at r=0.3 (p<0.001, see Supp. Text 4). However, the magnitude of this effect is small — a correlation of 0.3 means that variation in geography only explains 9% of the variation in phylogeny. Perhaps 9% is large enough for concern, but languages and geography are often linked not because of borrowing, but because languages tend to be born next to their sister languages. It is incorrect to state that “clearly geography, at least as much as known linguistic subgrouping, is a predictor for the results of [our] clustering”. Geography only predicts 9% of the signal in the tree.

Conclusions

None of the issues raised by Donohue et al. cast doubt on our central findings about the origin, age, expansion sequence and manner of the spread of Austronesian languages. Instead, the points raised are incorrect, reveal misunderstandings of phylogenetic inference, and are often irrelevant to the core issues at stake. To reiterate, quantification of the overall similarity between our trees and those of the comparative method reveals a high level of congruence. The majority of the signal in the basic vocabulary is genealogical, with geography only explaining 9% of the signal. Our approach extends the comparative method by providing a principled way of quantifying support for subgrouping hypotheses, and a robust method for inferring dates. Our analyses support an origin of Austronesian in Taiwan ca. 5,200 years ago and a series of expansion pulses and pauses. These findings are robust to the removal of most archaeological calibrations (Greenhill et al. 2010a), and contrast markedly with theories of massive language shift advanced by Oppenheimer & Richards 2001, Donohue & Denham 2011, Soares et al. 2011.

Computational phylogenetic methods are still relatively new in historical linguistics, and mutual misunderstandings are bound to occur. However, as we noted in Gray et al. (2009), the way forward lies in “the *combined power* of linguistic scholarship, database technologies, and computational phylogenetic methods” (emphasis added). These methods are a powerful supplement to traditional linguistic scholarship not a replacement.

References

Blust, Robert. 1981. Variation in Retention Rate among Austronesian Languages. Paper presented at the Third International Conference on Austronesian Linguistics, Bali, 19–23 January, 1981.

Blust, Robert. 2000. “Why Lexicostatistics Doesn’t Work: The ‘universal’ constant hypothesis and the Austronesian languages”. *Time Depth in Historical Linguistics* ed. by Colin Renfrew, April McMahon & Larry Trask, 311–331. Cambridge: The McDonald Institute for Archaeological Research.

Blust, Robert. 2009. *The Austronesian Languages*. Canberra: Pacific Linguistics.

Bouchard-Côté, Alexandre, Thomas L. Griffiths & Dan Klein. 2009. “Improved Reconstruction of Protolanguage Word Forms”. *Proceedings of Human Language Technologies*, 65–73. Boulder, Colorado: Association for Computational Linguistics.

Bowern, Claire, Patience Epps, Russell D. Gray, Jane Hill, Keith Hunley, Patrick McConvell & Jason Zentz. 2011. “Does Lateral Transmission Obscure Inheritance in Hunter-Gatherer Languages?” *PLoS ONE* 6.e25195.

Brugmann, Karl. 1884. “Zur Frage nach den Verwandtschaftsverhältnissen der indogermanischen Sprachen”. *Internationale Zeitschrift für allgemeine Sprachwissenschaft* 1.226–256.

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

Campbell, Lyle & William J. Poser. 2008. *Language Classification: History and method. Linguistics*. Cambridge: Cambridge University Press.

Collins, James T. 1982. "Linguistic Research in Maluku: A report of recent fieldwork". *Oceanic Linguistics* 21.73–146.

Donohue, Mark & Tim Denham. 2010. "Farming and Language in Island Southeast Asia". *Current Anthropology* 51.223–256.

Durie, Mark & Malcom Ross, eds. 1996. *The Comparative Method Reviewed: Regularity and irregularity in language change.* Oxford: Oxford University Press.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011a. "Evolved Structure of Language Shows Lineage-specific Trends in Word-order Universals". *Nature* 473.79–82.

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson & Neele Becker. 2011b. "Aslian Linguistic Prehistory: A case study in computational phylogenetics". *Diachronica* 28:3.291–323.

Dyen, Isidore. 1962. "The Lexicostatistical Classification of the Malayopolynesian Languages". *Language* 38.38–46.

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

Geraghty, Paul A. 1983. *The History of the Fijian Languages* (= *Oceanic Linguistics Special Publication*, 19). Honolulu: University of Hawaii Press.

Grant, Anthony P. 2005. "The Effects of Intimate Multidirectional Linguistic Contact: In Chamic". *Chamic and Beyond: Studies in mainland Austronesian languages* ed. by Anthony P. Grant & Paul Sidwell, 37–104. Canberra: Pacific Linguistics.

Gray, Russell D. & Quentin D. Atkinson. 2003. "Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin". *Nature* 426.435–439.

Gray, Russell D., David Bryant & Simon J. Greenhill. 2010. "On the Shape and Fabric of Human History". *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* 365.3923–3933.

Gray, Russell D., Alexei J. Drummond & Simon J. Greenhill. 2009. "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement". *Science* 323.479–483.

Greenhill, Simon J. 2011. "Levenshtein Distances Fail to Identify Language Relationships Accurately". *Computational Linguistics* 37.689–698.

Greenhill, Simon J, Robert Blust & Russell D. Gray. 2008. "The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics". *Evolutionary Bioinformatics* 4.271–283.

Greenhill, Simon J. & Ross Clark. 2011. "POLLEX-Online: The Polynesian lexicon project online". *Oceanic Linguistics* 50.551–559. http://pollex.org.nz

Greenhill, Simon J., Thomas E. Currie & Russell D. Gray. 2009. "Does Horizontal Transmission Invalidate Cultural Phylogenies?" *Proceedings of the Royal Society of London. Series B, Biological sciences* 276. 2299–2306.

Greenhill, Simon J., Alexei J. Drummond & Russell D. Gray. 2010a. "How Accurate and Robust are the Phylogenetic Estimates of Austronesian Language Relationships?" *PLoS ONE* 5.e9573.

Greenhill, Simon J. & Russell D. Gray. 2009. "Austronesian Language Phylogenies: Myths and misconceptions about Bayesian computational methods". *Austronesian Historical Linguistics and Culture History: A festschrift for Robert Blust* ed. by K. Alexander Adelaar & Andrew Pawley, 375–397. Canberra: Pacific Linguistics.

Haspelmath, Martin & Uri Tadmor, eds. 2009. *Loanwords in the World's Languages: A comparative handbook*. Berlin: De Gruyter Mouton.

Hennig, Willi. 1966. *Phylogenetic Systematics* (trans. by D. Davis & R. Zangerl). Urbana: University of Illinois Press.

Holden, Claire J. 2002. "Bantu Language Trees Reflect the Spread of Farming across Sub-Saharan Africa: A maximum-parsimony analysis". *Proceedings of the Royal Society of London, B* 269.793–799.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström & Sebastian Sauppe, et al. 2011. "Automated Dating of the World's Language Families Based on Lexical Similarity". *Current Anthropology* 52.841–875.

Howard, Irwin. 1981. "Proto-Ellicean". *Studies in Pacific Languages and Cultures in Honour of Bruce Biggs* ed. by Jim. Hollyman & Andrew Pawley, 101–118. Auckland: Linguistic Society of New Zealand.

Hübschmann, Heinrich [1875] 1967. "On the Position of Armenian in the Sphere of the Indo-European Languages". *A Reader in 19th Century Historical Indo-European Linguistics* ed. by Winfred P. Lehmann, 164–189 (trans. by Winfred P. Lehmann). Bloomington: Indiana University Press.

Huelsenbeck, John P., Bruce Rannala & John P. Masly. 2000. "Accommodating Phylogenetic Uncertainty in Evolutionary Studies". *Science* 288.2349–2350.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen & Jonathan P. Bollback. 2001. "Bayesian Inference of Phylogeny and its Impact on Evolutionary Biology". *Science* 294.2310–2314.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa & 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 of London*, *B* 276.2703–2710.

Lee, Sean & Toshikazu Hasegawa. 2011. "Bayesian Phylogenetic Analysis Supports an Agricultural Origin of Japonic languages". *Proceedings of the Royal Society, B* 278.3662–3669.

Lewis, Paul O. 2001. "A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data". *Systematic Biology* 50.913–925.

Mantel, Nathan A. 1967. "The Detection of Disease Clustering and a Generalized Regression Approach". *Cancer Research* 27.209–220.

Marck, Jeffrey C. 2000. *Topics in Polynesian Language and Culture History*. Canberra: Pacific Linguistics.

Nelson-Sathi, Shijulal, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin & Tal Dagan. 2011. "Networks Uncover Hidden Lexical Borrowing in Indo-European Language Evolution". *Proceedings of the Royal Society B: Biological Sciences* 278.1794–1803.

Oppenheimer, Stephen & Martin Richards. 2001. "Fast Trains, Slow Boats, and the Ancestry of the Polynesian Islanders". *Science Progress* 84.157–181.

Paradis, Emmanuel, Julien Claude & Korbinian Strimmer. 2004. "APE: Analyses of phylogenetics and evolution in R language". *Bioinformatics* 20.289–290.

Pawley, Andrew. 1967. "The Relationships of Polynesian Outlier Languages". *Journal of the Polynesian Society* 76.1–29.

Pawley, Andrew 2009. "Polynesian Paradoxes: Subgroups, wave models and the dialect geography of Proto Polynesian". Paper presented at the Eleventh International Conference on Austronesian Linguistics, Aussois, France, 22–26 June, 2009.

Pawley, Andrew 2010. “Retention and Replacement of Basic Vocabulary in the Central Pacific Languages”. Paper presented at the Eighth International Conference on Oceanic Linguistics, Auckland, 4–9 January, 2010.

Penny, David, Bennet J. McComish, Michael A. Charleston, & Michael D. Hendy. 2001. “Mathematical Elegance with Biochemical Realism: The covarion model of molecular evolution”. *Journal of Molecular Evolution* 53.711–723.

Petroni, Filippo & Maurizio Serva. 2008. “Language Distance and Tree Reconstruction”. *Journal of Statistical Mechanics: Theory and experiment* 2008.P08012.

R Development Core Team. 2011. *R: A language and environment for statistical computing*. http://www.r-project.org.

Soares, Pedro, Teresa Rito, Jean Trejaut, Maru Mormina, Catherine Hill, Emma Tinkler-Hundal & Michelle Braid, et al. 2011. “Ancient Voyaging and Polynesian Origins”. *The American Journal of Human Genetics* 88.1–9.

Steel, Mike A. & David Penny. 1993. “Distribution of Tree Comparison Metrics — Some new results”. *Systematic Biology* 42.126–141.

Tadmor, Uri, Martin Haspelmath & Bradley Taylor. 2010. “Borrowability and the Notion of Basic Vocabulary”. *Diachronica* 27:2.226–246.

Wilson, William. 2010. “New Evidence for Equatorial Outlier-East Polynesian”. Paper presented at the Eighth International Conference on Oceanic Linguistics, Auckland, New Zealand, 4–9 January, 2010.

Authors’ Addresses

Simon J. Greenhill
School of Culture, History & Language
ANU College of Asia and the Pacific
Australian National University
Canberra ACT 0200, Australia

simon.greenhill@anu.edu.au

Russell D. Gray
Department of Psychology
University of Auckland
Auckland 1142, New Zealand

rd.gray@auckland.ac.nz