Supplementary Online Material.

Supplementary Text:
Section 1 details the language subgrouping of Austronesian presented in Gray et al. (2009). Here we describe the congruence between our results and that of the linguistic comparative method. Section 2 discusses the small number of misplaced languages criticised by Donohue et al. and the potential reasons for these misplacements. Section 3 describes how we constructed a lexicostatistical tree for the section on "issue 6" in the main text. Section 4 provides methodological information about how we calculated the correlations between languages and geography under "issue 7" of the main text.

Supplementary Figures:
1. An example of a likelihood calculation for an unrooted tree demonstrating how changes in cognate sets are inferred probabilistically.
2. Histograms showing the expected amount of borrowing in lexical items according to the World Loanword Database.
3. A Densitree plot of the Polynesian language tree.
4. A Neighbour Net of the Central Pacific languages.
5. Histograms showing quartet distance comparisons between trees.
6. A lexicostatistical tree of 400 Austronesian languages.

1. Congruence between the Bayesian phylogenetic trees and the linguistic comparative method.
Contrary to the claims of Donohue et al, the trees presented in Gray et al. (2009) are highly congruent with the expected linguistic subgroupings. Our trees are rooted with the Formosan languages in Taiwan as most linguists argue (e.g. Blust 1999), and not elsewhere in South-East Asia as claimed by some non-linguists (Oppenheimer & Richards 2001). The Formosan languages do not form a distinct subgroup, but rather correctly inferr multiple first-order branches of Austronesian (Blust 1999). Our results strongly support the Malayo-Polynesian subgroup (posterior probability=1.00, Dahl 1973, Blust 1977) that separates the rest of the family from the Formosan languages. Consistent with Blust (1977, 2009), we show Western Malayo-Polynesian is not a monophyletic group. It has a number of primary branches rather than a single node (see figure 1). The first of these primary branches strongly groups the languages of the Philippines (1.00) into the various Philippine microgroups (Blust 1991): Bashiic (1.00), Cordilleran (1.00), Central Luzon (0.97), Palawanic (0.70), Central Philippines (1.00), Manobo (1.00), Subanun (1.00). Our trees show reasonable support for the subgroups of Malayic (0.78, Adelaar 1992) and Chamic (0.72, Thurgood 1999). However we find no support for grouping these together into Malayo-Chamic (Blust 1994, Thurgood 1999). Other primary branches in Western Malayo-Polynesian are replicated correctly including Celebic (0.99, Sneddon 1993, Mead 2003), Greater South Sulawesi (1.00, Adelaar 1994), and North Sarawak (0.90, Blust 1974).

Next, both the Central Malayo-Polynesian languages and the Eastern Malayo-Polynesian subfamily are grouped into Central-Eastern Malayo-Polynesian with strong support (0.85, Blust 1978). Central Malayo-Polynesian falls into a number of subgroups: a strong grouping of the languages of Timor (0.98), a strongly supported clustering (1.00) of the controversial Bima-Sumba subgroup (0.82, Blust 2008), and

the weakly supported subgroups of Central Maluku (0.54, Collins 1982) and the Yamdena-North Bomberai group (0.56, Blust 1993). There is no support here for a single Central Malayo-Polynesian group consistent with arguments that these languages are descended from a dialect network and show low internal cohesion (Blust 1993, 2009). South-Halmahera/West New Guinea is well supported (0.80, Blust 1978), and branches off as a sister-group to the Oceanic languages. Oceanic and South-Halmahera/West New Guinea are linked correctly into the Eastern Malayo-Polynesian subfamily albeit with weak support (0.58, Blust 1978)

Oceanic is supported perfectly (1.00, Dempwolff 1927, 1938, Pawley 1973, Lynch et al. 2002, Ross, Pawley & Osmond 1998, 2003, 2008, 2011). Inside Oceanic, we show strong support for South-East Solomonic (1.00, Pawley 1972), Temotu (1.00, Ross & Næss 2007), Admiralties (1.00, Ross 1988, Blust 1998, Lynch et al. 2002), and Papuan Tip (1.00, Ross 1988) subgroups. Meso-Melanesian is weakly supported (0.65, Ross 1988). The majority of the North New Guinea languages are strongly grouped together (0.85, Ross 1988). Our results link the North New Guinea and Papuan Tip families (0.96) together as New Guinean Oceanic (Pawley 1978).

We show weak support for a Remote Oceanic (0.69) group containing Central Pacific (1.00, Grace 1959, Geraghty 1983), which, in turn, contains Polynesian (1.00, Pawley 1966, Marck 2000). Our Remote Oceanic group includes the Micronesian languages, which cluster into a Nuclear Micronesian subgroup (1.00, Bender et al. 2003a,b). Also included in our Remote Oceanic group are the subgroups of South Vanuatu (Lynch 2001) and North/Central Vanuatu (Tryon 1976, Clark 1985, Lynch et al. 2002) that are strongly supported (0.99 and 1.00 respectively), and these are grouped with the languages from New Caledonia and the Loyalties to form Southern Oceanic (0.99, Ozanne-Rivierre 1992).

2. Inconsistences in the Bayesian language subgroupings.

Donohue et al. ignore the broad consistencies of our trees and attack the placement of single languages. Perhaps that most problematic aspect of Donohue et al.'s response to our paper is that they treat the consensus trees we reported as a single tree rather than as a visual summary of the distribution of most probable trees. This mistake is a serious misunderstanding and it causes Donohue et al. to make some major misinterpretations of our results. The tree reported in Figure S5 of Gray et al. (2009) is a Majority-Rule consensus tree (Margush & McMorris 1983, Bryant 2003) across the sample of the posterior distribution. Figure S5 is therefore a summary of 4,200 trees. The Majority-Rule method collapses branches that are found in less than 50% of the tree sample and places them in an unresolved cluster within the parent clade. Attending to the posterior probabilities on nodes of consensus trees is absolutely critical for their correct interpretation. It matters a great deal whether a node has a 0.51 or a 0.99 posterior probability. Donohue et al. fail to understand the importance of these posterior probabilities, and remove them from their discussion and figures.

First, Donohue et al. note that Nakanai is placed with the North New Guinea languages instead of with the Meso-Melanesian languages. In our original supplementary material we noted that Nakanai, along with the other Willaumez languages (Lakalai and Maututu) in West New Britain, was placed here in our trees. We suggested that this placement was presumably due to unidentified lexical borrowings between these Willaumez languages and the neighboring languages of

West New Britain belonging to the Meso-Melanesian subgroup. This does not invalidate our support for the North New Guinea subgroup.

The second subgrouping that Donohue et al. take issue with is Meso-Melanesian. They claim that our placement of Mussau with the Meso-Melanesian subgroup invalidates this grouping. Our results indeed place Mussau at the base of most of the Meso-Melanesian languages, followed by the language Vitu. In our figure S5 of Gray et al. (2009) the placement of the label (node 10) for Meso-Melanesian was situated there for clarity. Again, this issue was discussed in the supplement where we note that the placement of the St. Matthias subgroup (represented by Mussau) is only weakly supported. Moreover, the placement of Vitu at the base of this subgroup is not particularly surprising given that the Bali-Vitu lineage is thought to be a primary branch of Meso-Melanesian (Lynch et al. 2002). The slight misplacement of a single language does not mean that Meso-Melanesian is not supported by our results.

Next, Donohue et al. critique two of the recovered subgroups - South Halmahera/West New Guinea and Eastern Malayo-Polynesian - with the same criticism. Both of these subgroups should include the language Irarutu (called by the alternative name Kasira by our language informant). Donohue et al. claim that Irarutu is a language that belongs to the South Halmahera/West New Guinea subgroup. In our results, this language falls to the base of the parent clade. The placement of Irarutu has been problematic for some time (e.g. Anceaux 1961, Blust 1978, 1993, Voorhoeve 1989). For example, Blust (1993, p.272) states that "Irarutu apparently is not a (Central Malayo-Polynesian) language, and shows no known positive evidence of belonging to the (South Halmahera/West New Guinea) group. Its position for the present remains indeterminate". Current opinion at the most only weakly subgroups Irarutu with South Halmahera/West New Guinea, possibly as a first-order subgroup (Voorhoeve 1989). Therefore, rather than incorrect, our analyses are reflecting this classificatory difficulty by placing Irarutu between the Central Malayo-Polynesian and South Halmahera/West New Guinea languages. Voorhoeve (1989) notes that Irarutu has also undergone contact with Koiwai and other Central Malayo-Polynesian languages from the Bomberai peninsula. This contact may be the reason why our trees are more conservative and do not weakly subgroup Irarutu with the South-Halmahera/West New Guinea languages. Once again incorrectly reading the Majority-Rule consensus tree has misled Donohue et al. into thinking that this placement invalidates both the South Halmahera/West New Guinea subgroup and the Eastern Malayo-Polynesian subgroup. It does not.

Next, Donohue et al note that our Central Maluku group includes the languages of Aru. The placement of the two languages Ujir and Ngaibor from Aru in a group together with 13 Central Maluku languages may reflect unidentified borrowings between these two neighboring subgroups in Maluku. Again, this does not invalidate our Central Maluku subgrouping, but suggests that there may be an as-yet unidentified relationship (either descent or contact) between the languages of Maluku.

Donohue et al next question our placement of Koiwai and Kei inside the Yamdena-North Bomberai group. Blust (1993) places Koiwai into the neighboring subgroup of South Bomberai. Our results could suggest a greater subgroup including the Yamdena-North Bomberai with the South Bomberai languages. Alternatively, the placement of Koiwai here may reflect the widespread diffusion of features such as

glide truncation across the Bomberai region (Blust 1993). On first glance, the placement of Kei with these languages is unusual. However, Blust (personal communication, 17/3/2009) has unpublished data suggesting that Kei probably belongs to a slightly larger group that includes Yamdena-North Bomberai. Instead of being incorrect, our results may in fact be confirming yet to be published results.

Donohue et al also criticise our Greater South Sulawesi group for not including the Tamanic languages of Borneo. Specifically, the language Maloh falls to the base of the parent clade, and is not included in this subgroup. Again, Donohue et al's argument here is due to misunderstanding the Majority-Rule consensus tree representation. The analyses here are conservative and are refusing to link Maloh more closely with other languages. Once again, this does not indicate a lack of support for the Greater South Sulawesi language subgroup.

Next Donohue et al note that our trees group the Barito languages (Katingan, Ma'anyan, Merina [Malagasy], Nagaju Dayak, Tunjung) with the North Borneo subgroup. We also noted this in our supplementary material. The most likely sister-clade for the Barito languages is that of Sama-Bajaw (R. Blust, personal communication, 17/3/2009). The placement of Sama-Bajaw (Bajo, Inabaknon, Mapun, Samal (Siasi Dialect), Yakan) in our trees is also unusual (again, as we noted in our supplementary material). We have been reassessing the cognate coding in that area, and have uncovered 17 previously unrecognised loan words in the Sama-Bajaw language Inabaknon. These borrowings are the likely explanation for the minor mismatch between our results and the traditional linguistic subgroupings in this region.

Donohue et al. question the placement of the Sangiric language subgroup (Bantik, Sangil Sarangani Islands, Sangir, Sangir Tabukang Dialect, Tontemboan). Our results place this subgroup as a higher-order grouping within the Western Malayo-Polynesian languages. According to Blust (1991) this language microgroup should be a primary branch of the Philippines family. We already noted this discrepancy in our original supplementary material. The Sangiric languages are located in the Sulawesi region but are definitely Philippines languages (Blust 1991, Sneddon 1993). Our placement of Sangiric as a deeper group within the Western Malayo-Polynesian linkage may either reflect contact-induced change with neighboring Sulawesi languages, although there is little evidence for contact (Blust, pers. Comm. 08/02/2012) Once again, the minor shift of 5 languages from a clade of around 60 Philippines languages does not invalidate our support for the Philippines subgroup.

Next, Donohue et al. note that our Malayo-Sumbawan subgroup should not include Javanese, or the Sumatran languages (e.g. Lampung, Gayo or Batak). As we discussed in our supplementary material, the internal classification of our Malayo-Sumbawan subgroup differs from that proposed by Adelaar (2005a, 2005b). We suggested that these differences could be explained by unidentified borrowings between languages within these subgroups (Adelaar 2005b). In our original paper we suggested that these differences might be explained by unidentified borrowings between languages within these subgroups. For example, Balinese has a number of vocabulary registers and the higher status register is heavily Javanised (Adelaar, 2005a). It is possible that the Balinese word list reflects this Javanised register that may have caused the Javanese

language to be placed inside this subgroup (M. Ross, personal communication, 22/12/2008).

Donohue et al. then move on to critique our results for not linking Malayic and Chamic into Malayo-Chamic (Blust 1994, Thurgood 1999). In fact, our results do group the Malayic and Chamic languages into a higher-order subgroup with moderate posterior probability (0.60). Inside this group there are other languages that do not belong to Malayo-Chamic (e.g. Javanese, see above), but this does not invalidate that group.

Donohue et al. mistakenly chastise us for placing Paiwan inside Malayo-Polynesian. This is false – Donohue et al. have made a basic error in reading the tree – Paiwan correctly falls outside Malayo-Polynesian.

Finally, Donohue et al. note that our results include the disputed Bima-Sumba subgroup. Blust's recent reanalysis of the evidence for this subgroup (Blust 2008) found strong support for Sumba-Hawu, but not for grouping that with Bima. Our results are actually in concordance with this, showing strong support for Sumba-Hawu (0.99), and only moderate support for Bima-Sumba (0.82). Our trees place Bima closer to the Sumba-Hawu languages, than the neighboring languages of Flores-Lembata. Blust (2008) argues that whilst Bima-Sumba might not exist, a higher order subgrouping might exist that includes the Ambon-Timor languages along with the Sumba-Hawu and Bima languages. Our results match this interpretation.

3. Constructing a Lexicostatistical Tree:

To provide a comparison to the phylogenetic tree we computed the pairwise cognate distances for the same cognate data. We then clustered the trees using the UPGMA algorithm (Sokal & Sneath 1963) as implemented in SplitsTree v4.12 (Huson & Bryant 2006). This algorithm is the same as the standard Lexicostatistics clustering approach (Swadesh 1952).

4. Correlations between languages and geography:

To calculate the relationship between languages and geography we took the 400 languages studied in Gray et al. (2009), and obtained geographical locations for all of these from the Austronesian Basic Vocabulary Database (Greenhill et al. 2008). We then calculated the geographical "great-circle" distance between all pairs of languages using the standard Haversine formula (Sinnott 1984). Second, we calculated the phylogenetic distance between each pair of languages using a common measure of distance on a phylogeny – the patristic distance (the sum of the branch-lengths on the shortest path between each language). Third, we calculated the Mantel correlation using a Pearson correlation statistic and 1000 random permutations (R Development Core Team 2011, Oksanen et al. 2010).

References:

Adelaar, K.A. 1992. "Proto-Malayic: a reconstruction of its phonology and part of its morphology and lexicon". Canberra: Pacific Linguistics.

Adelaar, K.A. 1994. "The classification of the Tamanic languages". *Language contact and change in the Austronesian world*. ed. by T. Dutton & D. Tryon, 1-42. Berlin: Mouton de Gruyter.

Adelaar, A. 2005a. "Malayo-Sumbawan". *Oceanic Linguistics*, 44: 357-388.

Adelaar, A. 2005b. "The Austronesian languages of Asia and Madagascar: A historical perspective". *The Austronesian languages of Asia and Madagascar*. ed. by A. Adelaar & N.P. Himmelmann, 1-42. London: Routledge.

Anceaux, J.C. 1961. "The linguistic situation in the islands on Yapen, Kurundu, Nau, and Miosnum, New Guinea". *Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde*, 35. The Hague: Nijhoff.

Bender, W.B., Goodenough, W.H., Jackson, F.H., Marck, J.C., Rehg, K.L., Sohn, H., Trussel, S. & Wang, J.W. 2003. "Proto-Micronesian reconstructions I". *Oceanic Linguistics*, 42:1-110.

Bender, W.B., Goodenough, W.H., Jackson, F.H., Marck, J.C., Rehg, K.L., Sohn, H., Trussel, S. & Wang, J.W. 2003. "Proto-Micronesian reconstructions II". *Oceanic Linguistics*, 42:271-358.

Blust, R. 1974. *The Proto-North Sarawak vowel deletion hypothesis*. Unpublished Ph.D dissertation. University of Hawaii.

Blust, R.A. 1977. "The proto-Austronesian pronouns and Austronesian subgrouping: a preliminary report". *University of Hawai'i Working Papers in Linguistics* 9:1–15.

Blust, R. 1978. "Eastern Malayo-Polynesian: a subgrouping argument". *Second international conference on Austronesian linguistics: proceedings, Fascicle I, Western Austronesian*. ed. by S.A. Wurm & L. Carrington, 181-234. Canberra: Pacific Linguistics.

Blust, R. 1991. "The Greater Central Philippines hypothesis". *Oceanic Linguistics*, 30:73-129.

Blust, R. 1993. "Central and Central-Eastern Malayo-Polynesian". *Oceanic Linguistics*, 32:241-293.

Blust, R. 1994. "The Austronesian settlement of mainland Southeast Asia". *Papers from the Second Annual Meeting of the Southeast Asian Linguistics Society*. ed. by K.L. Adams & T.J. Hudak, 25-83. Arizona: Arizona State University.

Blust, R. 1998. "A note on higher-order subgroups in Oceanic". *Oceanic Linguistics*, 37:182-188.

Blust, R. 1999. "Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics". *Selected papers from the Eighth International Conference on Austronesian Linguistics* vol. 1. ed. by E. Zeitoun & P. Jen-kuei Li, 31–94. Taipei, Taiwan: Symposium Series of the Institute of Linguistics, Academia Sinica.

Blust, R. 2008. "Is there a Bima-Sumba subgroup?" *Oceanic Linguistics*, 47: 45-113.

Blust, R. A. 2009. *The Austronesian languages*. Canberra: Pacific Linguistics.

Bouckaert, R. R. 2010. "DensiTree: making sense of sets of phylogenetic trees". *Bioinformatics* 26, 1372-1373.

Bryant, D. 2003. "A classification of consensus methods for phylogenetics". *Bioconsensus*. ed. by M. Janowitz, F.J. Lapointe, F. McMorris, B. Mirkin, & F. Roberts. 1-21, Providence, Rhode Island: DIMACS-AMS.

Clark, R. 1985. "Languages of north and central Vanuatu: groups, chains, clusters and waves". *Austronesian linguistics at the 15th Pacific Science congress*. ed by A. Pawley, & L. Carrington, 199-236. Canberra: Pacific Linguistics.

Collins, J.T. 1982. "Linguistic research in Maluku: A report of recent fieldwork". *Oceanic Linguistics*, 21:73-146.

Dahl, O.C. 1973. *Proto-Austronesian*. Scandinavian Institute of Asian Studies. Monograph Series No. 15. Sweden: Studentlitteratur.

Dempwolff, O. 1927. "Das austronesische Sprachgut in den melanesischen Sprachen". *Folia Ethnoglossica*, 3:32-43.

Dempwolff, O. 1938. "Vergleichende Lautlehre des austronesischen Wortschatzes, Band 3: Austronesisches Wörterverzeichnis". *Beihefte zur Zeitschrift für Eingeborenen-Sprachen* 19. Berlin: Dietrich Reimer.

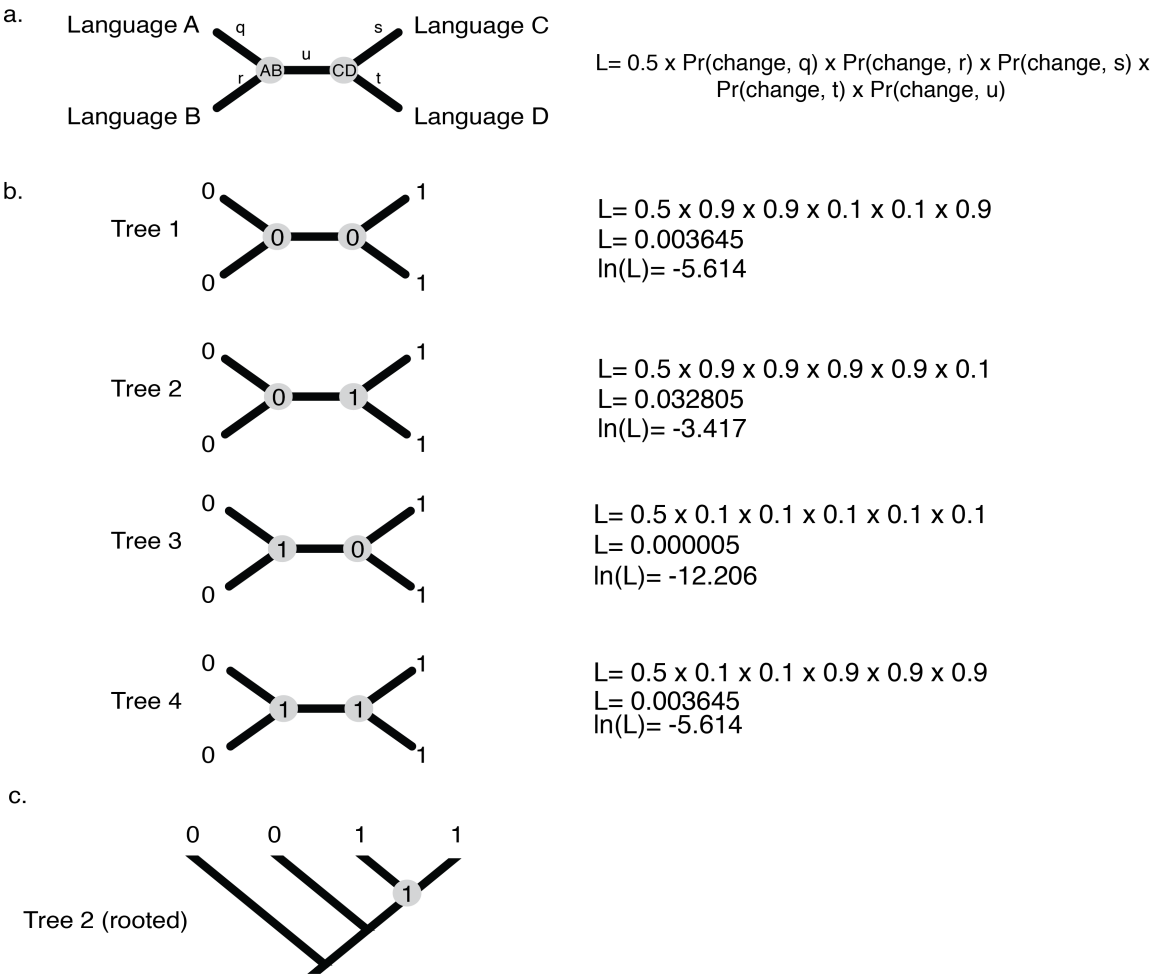
Geraghty, P. A. 1983. *The History of the Fijian Languages*. Oceanic Linguistics Special Publication No.19. Honolulu: University of Hawaii Press

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 and linguistics.

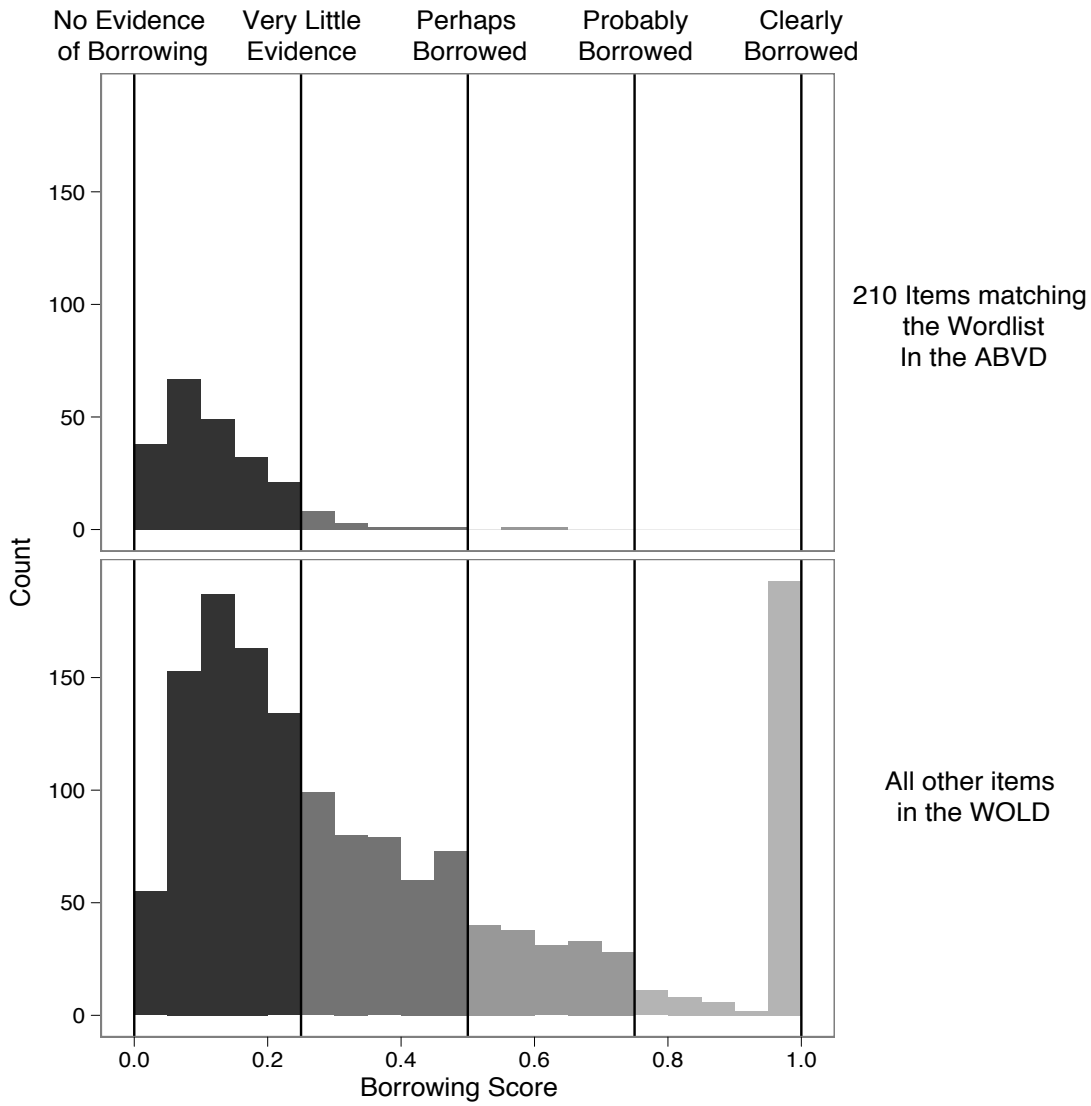
Gray, R. D., Drummond, A.J., & Greenhill, S.J. 2009. "Language phylogenies reveal expansion pulses and pauses in Pacific settlement". *Science* 323: 479-83.

Greenhill, S. J, Blust, R., & Gray, R. D. 2008. "The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics". *Evolutionary Bioinformatics,* 4: 271-283.
Huson D.H., & Bryant, D. 2006. "Application of Phylogenetic Networks in Evolutionary Studies", Molecular Biology and Evolution, 23, 254-267,
Lynch, J. 2001. *The linguistic history of Southern Vanuatu*. Canberra: Pacific Linguistics.
Lynch, J., Ross, M.D. & Crowley, T. ed. 2002. *The Oceanic languages*. Richmond: Curzon Press.
Marck, J. 2000. *Topics in Polynesian language and culture history*. Canberra: Pacific Linguistics.
Margush, T., & McMorris, F.R. 1981. "Consensus n-trees". *Bulletin of Mathematical Biology*, 43:239-244.
Mead, D. 2003. "Evidence for a Celebic supergroup". *Issues in Austronesian historical phonology*. ed. by J. Lynch, 115-141. Canberra: Pacific Linguistics.
Oksanen, J., Blanchet F. G., Kindt, R., Legendre, P., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., & Wagner, H. 2010. "Vegan: Community Ecology Package". R package version 1.17-4.
Oppenheimer, S., & Richards, M. 2001. "Fast trains, slow boats, and the ancestry of the Polynesian islanders". *Science Progress*, *84*, 157-181.
Ozanne-Rivierre, F. 1992. "The Proto-Oceanic consonantal system and the languages of New Caledonia". *Oceanic Linguistics*, 31:191-207.
Pawley, A. 1966. "Polynesian Languages: A subgrouping based on shared innovations in morphology". *Journal of the Polynesian Society*, 75: 39-64.
Pawley, A. 1972. "On the internal relationships of eastern Oceanic languages". *Studies in Oceanic Culture History*. ed. by R.C. Green & M. Kelly, Vol. 3, 3-106. Pacific Anthropological Records No. 13. Honolulu: Bernice P. Bishop Museum.
Pawley, A. K. 1973. "Some problems in Proto-Oceanic grammar". *Oceanic Linguistics*, 12:103-188
Pawley, A. 1978. "The New Guinea Oceanic hypothesis". *Kivung* 13: 99-151.
R Development Core Team. 2011. R: A Language and Environment for Statistical Computing. http://www.r-project.org.
Ross, M. 1988. "Proto-Oceanic and the Austronesian languages of Western Melanesia", Canberra: Pacific Linguistics.
Ross, M. D., Pawley, A., & Osmond, M. eds. 1998. *The lexicon of Proto-Oceanic: Volume 1, Material Culture*. Canberra: Australian National University.
Ross, M. D., Pawley, A., & Osmond, M. eds. 2003. *The lexicon of Proto-Oceanic: Volume 2, The Physical Environment*. Canberra: Australian National University.
Ross, M. D., Pawley, A., & Osmond, M. eds. 2008. *The lexicon of Proto-Oceanic: Volume 3, Plants*. Canberra: Australian National University.
Ross, M. D., Pawley, A., & Osmond, M. eds. 2011. *The lexicon of Proto-Oceanic: Volume 4, Animals*. Canberra: Australian National University.
Ross, M. & Næss, A. 2007. "An Oceanic origin for Äiwoo, the language of the Reef Islands?" *Oceanic Linguistics*, 46: 456–498.
Sinnott, R. W. 1984. "Virtues of the Haversine". *Sky and Telescope* 68: 159–161.
Sneddon, J. N. 1993. "The drift towards final open syllables in Sulawesi languages". *Oceanic Linguistics*, 32:1-44.
Sokal, R.R., Sneath, P.H.A.. 1963. *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.
Swadesh, M. 1952. "Lexico-statistic dating of prehistoric ethnic contacts". *Proceedings of the American Philosophical Society 96*, 453–463.
Thurgood, G. 1999. *From ancient Cham to modern dialects: Two thousand years of language contact and change*. Hawaii: University of Hawaii Press.
Tryon, D.T. 1976. *New Hebrides languages: An internal classification*. Canberra: Pacific Linguistics.
Voorhoeve, C.L. 1989. "Notes on Irarutu". *Irian*, 18: 106-119.

Supplementary Figures:

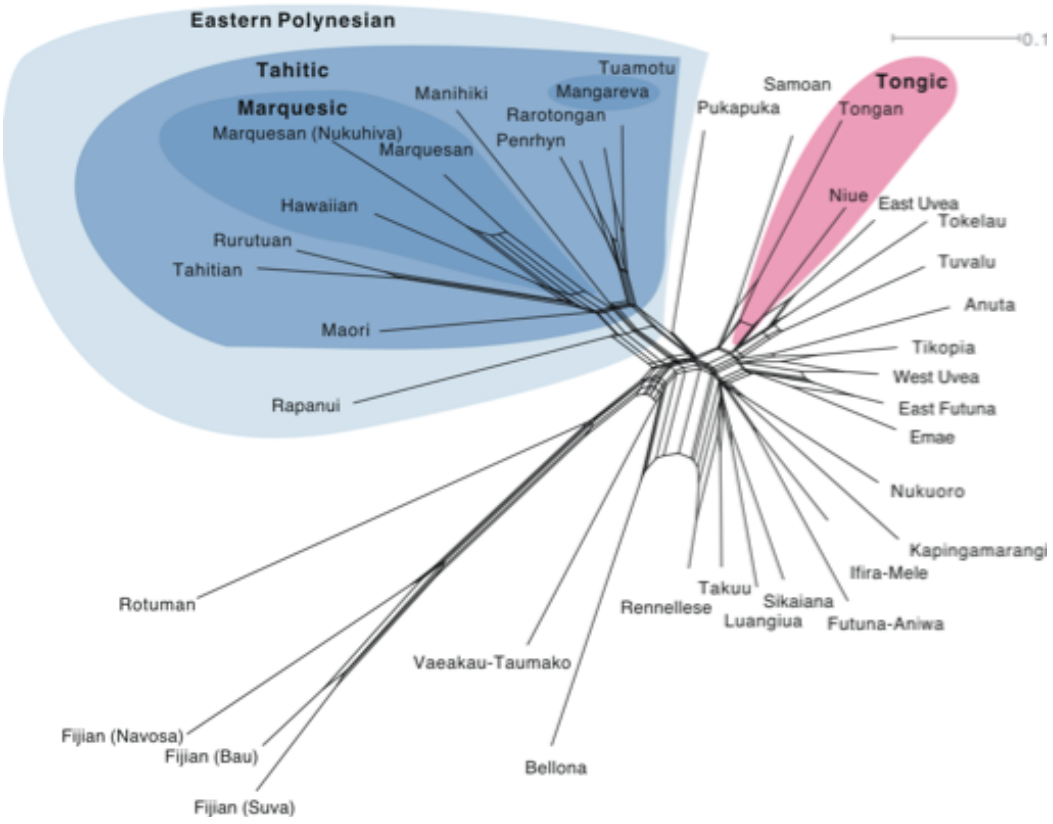


Supplementary Figure 1: Example of a likelihood calculation for an unrooted tree of four languages A, B, C, and D. This example demonstrates how character state changes in each cognate set are inferred probabilistically on every branch on the tree. Languages C and D have innovated a new cognate set in proto-language CD, somewhere along branch *u*. Figure 1a Shows a schematic of an unrooted tree of languages A, B, C, and D with branches *q*, *r*, *s*, *t*, and *u*. The equation for calculating the likelihood of this tree is on the left. The model assumes equal state probabilities such that the probability of the root of the tree is 0.5. The probability of no change along a branch is assumed to be 0.9, whilst the probability of change along a branch is the inverse, 0.1. Figure 1b shows the four possible state assignments of the proto-languages (Trees 1-4) and their likelihood calculations. For example, the likelihood for Tree 1 is the product of the root probability of 0.5, and the probability of no change on branches q, r, and u, and changes on branches s and t. In contrast, the likelihood of Tree 2 is the product of the root probability multiplied by the probability of no changes on branches q, r, s and t, and the probability of one change on branch u. Tree 2 fits the data much better than the other trees. Figure 1c shows Tree 2 rooted with Language A (e.g. based on prior knowledge of the outgroup). Rooting this tree indicates that the cognate set was innovated on the branch leading to the proto-language CD. Thus our methodology does not ignore the distinction between retentions and innovations, but is based on inferring them.

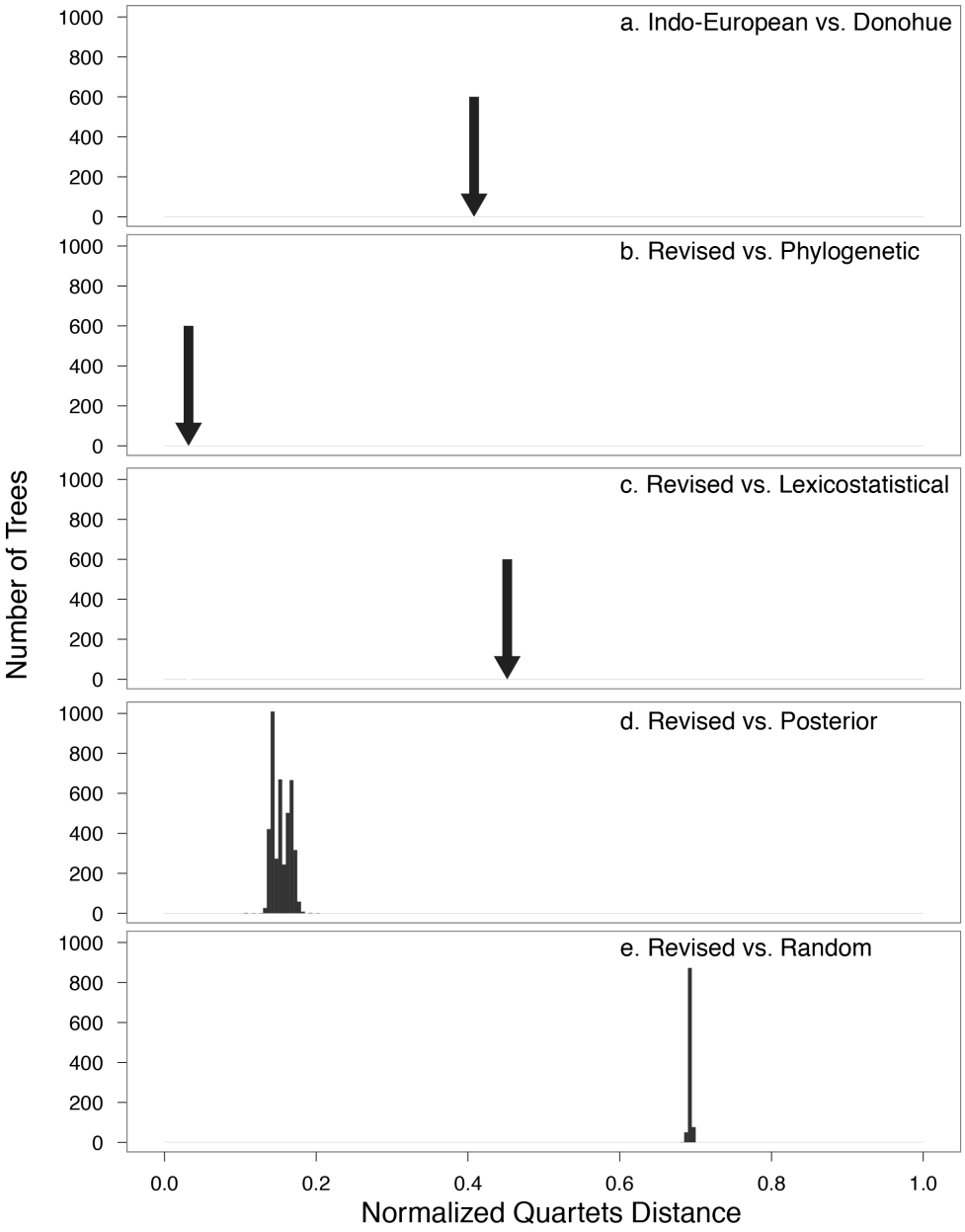


Supplementary Figure 2: Histograms showing the expected degree of borrowing in lexical items according to the World Loanword Database (WOLD). The vast majority of the words in the Austronesian Basic Vocabulary Database fall below the "Very Little Evidence of Borrowing" threshold (top).

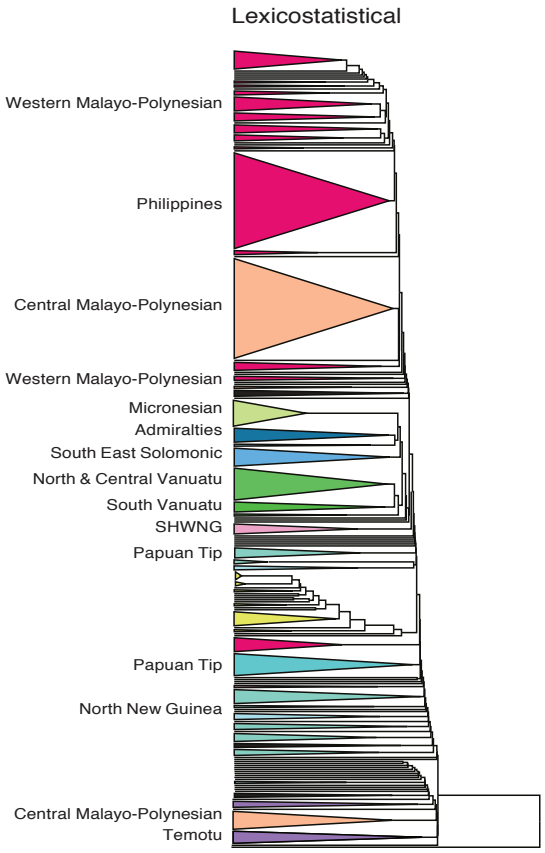
Supplementary Figure 3: Densitree plot (Bouckaert 2010) of the Polynesian languages showing 800 trees from the Posterior probability distribution of Gray et al. (2009). The plot shows marked conflicting signal and uncertainty in many of the language subgroups as would be expected given the dialect networks that occurred when Proto-Polynesian broke up (Geraghty 1983, Marck 2000, Pawley 1967, 2009, 2010, Rensch 1987). Despite Donohue et al.'s claims that our tree misplaces Rapanui with Marquesan and Mangarevan, the Densitree reveals the uncertainty around the subgrouping and shows that Rapanui often falls to the base of Proto-Eastern Polynesian in accordance with Marck (2000). We have highlighted in red one of these tree topologies.



Supplementary Figure 4: NeighbourNet graph inferred from Central Pacific basic vocabulary data -see Gray et al. (2010). Note that this network includes known loan words.



Supplementary Figure 5: Histograms showing the Normalized Quartets Distance. Trees that are closer together will have a score near 0, whilst trees that are maximally different will have a score approaching 1.0. Fig 2a. shows the distance between the Donohue et al. Indo-European tree and the standard Indo-European subgrouping. Fig 2b. shows the distance between the phylogenetic tree in figure S5 of Gray et al. and the revised tree to match the subgrouping differences noted by Donohue et al. Fig 2c shows the distance between the revised tree and the lexicostatistical tree. Fig 2d. shows the distance between this revised tree and the trees found by Gray et al. 2e. shows the distance between the revised tree and a random sample of 1000 trees.



Supplementary Figure 6: A tree of the 400 Austronesian languages built using lexicostatistics. Colors match those used in figure 4. This tree shows the vastly different tree topology found by lexicostatistical methods caused by a failure to handle the substantial rate variation found in Austronesian (e.g. the tree is rooted in Melanesia not Formosan, and Western Malayo-Polynesian is most divergent).