Supporting Information: Automated reconstruction of ancient languages using probabilistic models of sound change

Alexandre Bouchard-Côté
David Hall
Thomas L. Griffiths
Dan Klein

This Supporting Information describes the learning and inference algorithms used by our system (Section 1), details of simulations supporting the accuracy of our system and analyzing the effects of functional load (Section 2), lists of reconstructions (Section 3), lists of sound changes (Section 4), and analysis of frequent errors (Section 5).

1 Learning and inference

The generative model introduced in the Materials and Methods section of the main paper sets us up with two problems to solve: estimating the values of the parameters characterizing the distribution on sound changes on each branch of the tree, and inferring the optimal values of the strings representing words in the unobserved protolanguages. Section 1.1 introduces the full objective function that we need to optimize in order to estimate these quantities. Section 1.2 describes the Monte Carlo Expectation-Maximization algorithm we used for solving the learning problem. We present the algorithm for inferring ancestral word forms in Section 1.4.

1.1 Full objective function

The generative model specified in the Materials and Methods section of the main paper defines an objective function that we can optimize in order to find good protolanguage reconstructions. This objective function takes the form of a regularized log-likelihood, combining the probability of the observed languages with additional constraints intended to deal with data sparsity. This objective function can be written concisely if we let $\mathbb{P}_\lambda(\cdot), \mathbb{P}_\lambda(\cdot|\cdot)$ denote the root and branch probability models described in the Materials and Methods section of the paper (with transition probabilities given by the above logistic regression model), $I(c)$, the set of internal (non-leaf) nodes in $\tau(c)$, $\mathrm{pa}(\ell)$, the parent of language ℓ, $\mathrm{r}(c)$, the root of $\tau(c)$ and $W(c) = (\Sigma^*)^{|I(c)|}$. The full objective function is then

$$\mathrm{Li}(\lambda, \kappa) = \sum_{c=1}^{C} \log \sum_{\vec{w} \in W(c)} \mathbb{P}_\lambda(w_{c,\mathrm{r}(c)}) \prod_{\ell \in I(c)} \mathbb{P}_\lambda(w_{c,\ell}|w_{c,\mathrm{pa}(\ell)}, n_{c\ell})\mathbb{P}_\kappa(n_{c\ell}|\nu_{c\ell}) - \frac{||\lambda||_2^2 + ||\kappa||_2^2}{2\sigma^2} \qquad (1)$$

where the second term is a standard L^2 regularization penalty intended to reduce over-fitting due to data sparsity (we used $\sigma^2 = 1$) [1]. The goal of learning is to find the value of λ, the parameters of the logistic regression model for the transition probabilities, that maximizes this function.

1.2 A Monte Carlo Expectation-Maximization algorithm for reconstruction

Optimization of the objective function given in Equation 1 is done using a Monte Carlo variant of the Expectation-Maximization (EM) algorithm [2]. This algorithm breaks down into two steps, an E step in which the objective function is approximated and an M step in which this approximate objective function is optimized. The M step is convex and computed using L-BFGS [3] but the E step is intractable [4], in part because it requires solving the problem of inferring the words in the protolanguages. We approximate the solution to this inference problem using a Markov chain Monte Carlo (MCMC) algorithm [5]. This algorithm repeatedly samples words from the protolanguages until it converges on the distribution implied by our generative model. Since this procedure is guaranteed to find a local maximum of the objective, we ran it with several different random initializations of the model parameters. The next two subsections provide the details of these two parts of our system.

1.2.1 E step: Inferring the posterior over string reconstructions

In the E step, the inference problem is to compute an expectation under the posterior over strings in a protolanguage given observed word forms at the leaves of the tree.[1] The typical approach in biological InDel models [6] is to use Gibbs sampling, where the entire string at each node in the tree is repeatedly resampled, conditioned on its parent and children. We will call this method Single Sequence Resampling (SSR). While conceptually simple, this approach suffers from mixing problems in large trees, since it can take a long time for information to propagate from one region of the tree to another [6]. Consequently, we use a different MCMC procedure, called Ancestry Resampling (AR) that alleviates these mixing problems. This method was originally introduced for biological applications [7], but commonalities between the biological and linguistic cases make it possible to use it in our model.

Concretely, the problem with SSR arises when the tree under consideration is large or unbalanced. In this case, it can take a long time for information from the observed languages to propagate to the root of the tree. Indeed, samples at the root will initially be *independent* of the observations. AR addresses this problem by resampling one thin vertical slice of all sequences at a time, called an ancestry (for the precise definition of the algorithm, see [7]). Slices condition on observed data, avoiding mixing problems, and can propagate information rapidly across the tree. We ran the ancestry resampling algorithm for a number of iterations that increased linearly with the number of iterations of the EM algorithm that had been completed, resulting in an approximation regime that could allow the EM algorithm to converge to a solution [8]. To speed-up the large experiments, we also used an approximation in AR. This approximation is based on fixing the value of a set-valued auxiliary variables z_g, where $z_g = \{w_{g,\ell}\}$, and ℓ ranges over the set of all languages (both internal and at the leaves). Conditioning on these variables, sampling $w_{g,\ell}|z_g$ can be done exactly using dynamic programming and rejection sampling.

1.2.2 M step: Convex optimization of the approximate objective

In the M step, we individually update the parameters λ and κ as specified in Equation 1. We show how λ is updated in this section, κ can be optimized similarly. Let $\mathcal{C} = (\omega, t, p, \ell)$ denote local transducer contexts from the space $\mathbf{C} = \{S, I, R\} \times \Sigma \times \Sigma \times L$ of all such contexts. Let $N(\mathcal{C}, \xi)$ be the expected number of times the transition ξ was used in context $\mathcal{C}$ in the preceding E-step. Given these sufficient statistics, the estimate of λ is given by optimizing the expected complete (regularized) log-likelihood $\mathcal{O}(\lambda)$ derived from the original objective function given Equation [1] in the Materials and Methods section of the main paper (ignoring terms that do not involve λ),

$$\mathcal{O}(\lambda) = \sum_{\mathcal{C}\in\mathbf{C}} \sum_{\xi\in\mathscr{S}} N(\mathcal{C}, \xi) \Big[\langle \lambda, f(\mathcal{C}, \xi) \rangle - \log \sum_{\xi'} \exp\{\langle \lambda, f(\mathcal{C}, \xi') \rangle\} \Big] - \frac{||\lambda||^2}{2\sigma^2}.$$

[1] To be precise: the posterior is over both protolanguage strings and the derivations between these strings and the modern words.

We use L-BFGS [3] to optimize this convex objective function. L-BFGS requires the partial derivatives

$$\begin{aligned}\frac{\partial \mathcal{O}(\lambda)}{\partial \lambda_j} &= \sum_{\mathcal{C}\in\mathbf{C}}\sum_{\xi\in\mathscr{S}} N(\mathcal{C},\xi)\Big[f_j(\mathcal{C},\xi) - \sum_{\xi'}\theta_{\mathcal{C}}(\xi';\lambda)f_j(\mathcal{C},\xi')\Big] - \frac{\lambda_j}{\sigma^2}\\ &= \hat{F}_j - \sum_{\mathcal{C}\in\mathbf{C}}\sum_{\xi\in\mathscr{S}} N(\mathcal{C},\cdot)\theta_{\mathcal{C}}(\xi';\lambda)f_j(\mathcal{C},\xi') - \frac{\lambda_j}{\sigma^2},\end{aligned}$$

where $\hat{F}_j = \sum_{\mathcal{C}\in\mathbf{C}}\sum_{\xi\in\mathscr{S}} N(\mathcal{C},\xi)f_j(\mathcal{C},\xi)$ is the empirical feature vector and $N(\mathcal{C},\cdot) = \sum_{\xi} N(\mathcal{C},\xi)$ is the number of times context $\mathcal{C}$ was used. $\hat{F}_j$ and $N(\mathcal{C},\cdot)$ do not depend on λ and thus can be precomputed at the beginning of the M-step, thereby speeding up each L-BFGS iteration.

1.3 Approximate Expectation-Maximization for cognate inference

The procedure for cognate inference is similar: we again operate within the Expectation-Maximization framework, and the M-steps are identical. However, because of different characteristics of the cognate inference problem, we make a few different choices for approximations for the E-step. The Monte Carlo inference algorithm described in the preceding section works exceedingly well for reconstructing words for known cognates. However, a Monte Carlo approach to determining which words are cognate requires resampling the innovation variables $n_{g\ell}$, which is likely to lead to slow mixing of the Markov chain.

We therefore take a different approach when doing cognate inference. Here, we restrict our system to not perform inference over all possible reconstructions for all words, but to only use words that correspond to some observed modern word with the same meaning. The simplification is of course false, but it works well in practice. Specifically, we perform inference on the tree for each gloss using message passing [9], also known as pruning in the computational biology literature [10], where each message $\mu(w_{g\ell})$ has a non-zero score only when $w_{g\ell}$ is one of the observed modern word forms in gloss g. The result of inference yields expected alignments counts for each character in each language along with the expected number of innovations that occur at each language. These expectations can then be used in the convex M-step.

1.4 Ancestral word form reconstruction

In the E step described in the preceding section, a posterior distribution π over ancestral sequences given observed forms is approximated by a collection of samples $X_1, X_2, \ldots X_S$. In this section, we describe how this distribution is summarized to produce a single output string for each cognate set.

This algorithm is based on a fundamental Bayesian decision theoretic concept: Bayes estimators. Given a loss function over strings $\mathrm{Loss} : \Sigma^* \times \Sigma^* \to [0,\infty)$, an estimator is a Bayes estimator if it belongs to the random set:

$$\operatorname*{argmin}_{x\in\Sigma^*} \mathbb{E}^{\pi}\,\mathrm{Loss}(x,X) = \operatorname*{argmin}_{x\in\Sigma^*} \sum_{y\in\Sigma^*} \mathrm{Loss}(x,y)\pi(y).$$

Bayes estimators are not only optimal within the Bayesian decision framework, but also satisfy frequentist optimality criteria such as admissibility [11]. In our case, the loss we used is the Levenshtein [12] distance, denoted $\mathrm{Loss}(x,y) = \mathrm{Lev}(x,y)$ (we discuss this choice in more detail in Section 2).

Since we do not have access to π, but rather to an approximation based on S samples, the objective function we use for reconstruction rewrites as follows:

$$\begin{aligned}\operatorname*{argmin}_{x\in\Sigma^*} \sum_{y\in\Sigma^*} \mathrm{Lev}(x,y)\pi(y) &\approx \operatorname*{argmin}_{x\in\Sigma^*} \frac{1}{S}\sum_{s=1}^{S} \mathrm{Lev}(x,X_s)\\ &= \operatorname*{argmin}_{x\in\Sigma^*} \sum_{s=1}^{S} \mathrm{Lev}(x,X_s).\end{aligned}$$

The raw samples contain both derivations and strings for all protolanguages, whereas we are only interested in reconstructing words in a single protolanguage. This is addressed by marginalization, which is done in sampling representations by simply discarding the irrelevant information. Hence, the random variables X_s in the above equation can be viewed as being string-valued random variables.

Note that the optimum is not changed if we restrict the minimization to be taken on $x \in \Sigma^*$ such that $m \leq |x| \leq M$ where $m = \min_s |X_s|, M = \max_s |X_s|$. However, even with this simplification, optimization is intractable. As an approximation, we considered only strings built by at most k contiguous substrings taken from the word forms in $X_1, X_2, \ldots, X_S$. If $k = 1$, then it is equivalent to taking the min over $\{X_s : 1 \leq s \leq S\}$. At the other end of the spectrum, if $k = S$, it is exact. This scheme is exponential in k, but since words are relatively short, we found that $k = 2$ often finds the same solution as higher values of k.

1.5 Finding Cognate Groups

Our cognate model finds cuts to the phylogeny in order to determine which words are cognate with one another. However, this approach cannot straightforwardly handle instances where the evolution of the words did not follow strictly treelike behavior. This limitation applies to polymorphisms—where multiple words for a given meaning are available in a language—and also for borrowing—where the words did not evolve according to the phylogeny.

However, we can modify the inference procedure to capture these kinds of behaviors using a *post hoc* agglomerative "merge" procedure. Specifically, we can run our procedure to find an initial set of cognate groups, and then merge those cognate groups that produce an increase in model score. That is, we create several initial small subtrees containing some cognates, and then stitch them together into one or more larger trees. Thus, non-treelike behaviors like borrowing will be represented as multiple trees that "overlap." For instance, if two languages each have two words for one meaning (say A and B), then the initial stage might find that the two A's are cognate, leaving the two B's as singleton cognate groups. However, merging these two words into a single group will likely produce a gain in likelihood, and so we can merge them. Note this procedure is unlikely to work well for long distance borrowings, as the sound changes involved in long distance borrowing are likely to be very different from those according to the phylogeny. Nevertheless, we found this procedure to be effective in practice.

2 Experiments

In this section, we give more details on the results in the main paper, namely those concerning validation using reconstruction error rate, and those measuring the effect of the tree topology and the number of languages. We also include additional comparisons to other reconstruction methods, as well as cognate inference results. In Section 2.1, we analyze in isolation the effects of varying the set of features, the number of observed languages, the topology, and the number of iterations of EM. In Section 2.2 we compare performance to an oracle and to two other systems.

Evaluation of all methods was done by computing the Levenshtein distance [12] (uniform-cost edit distance) between the reconstruction produced by each method and the reconstruction produced by linguists. The Levenshtein distance is the minimum number of substitutions, insertions, or deletions of a phoneme required to transform one word to another. While the Levenshtein distance misses important aspects of phonology (all phoneme substitutions are not equal, for instance), it is parameter-free and still correlates to a large extent with linguistic quality of reconstruction. It is also superior to held-out log-likelihood, which fails to penalize errors in the modeling assumptions, and to measuring the percentage of perfect reconstructions, which ignores the degree of correctness of each reconstructed word. We averaged this distance across reconstructed words to report a single number for each method. The statistical significance of all performance differences are assessed using a paired t-test with significance level of 0.05.

2.1 Evaluating system performance

We used the Austronesian Basic Vocabulary Database (ABVD) [13] as the basis for a series of experiments used to evaluate the performance of our system and the factors relevant to its success. The database, downloaded from `http://language.psy.auckland.ac.nz/austronesian/` on August 7, 2010, includes partial cognacy judgments and IPA transcriptions,[2] as well as a several reconstructed protolanguages.

In our main experiments, we used the tree topology induced by the Ethnologue classification [14] in order to facilitate interpretability of the results (i.e. so that we can give well-known names to clades in the tree in our analyses). Our method does not require specifying branch lengths since our unsupervised learning procedure provides a more flexible way of estimating the amount of change between points of the phylogenetic tree.

The first claim we verified experimentally is that having more observed languages aids reconstruction of protolanguages. In the results in this section, we used the subset of the languages under Proto-Oceanic (POc) to speed-up the computations. To test this hypothesis we added observed modern languages in increasing order of distance d_c to the target reconstruction of POc so that the languages that are most useful for POc reconstruction are added first. This prevents the effects of adding a close language after several distant ones being confused with an improvement produced by increasing the number of languages.

The results are reported in Figure 1(c) of the main paper. They confirm that large-scale inference is desirable for automatic protolanguage reconstruction: reconstruction improved statistically significantly with each increase except from 32 to 64 languages, where the average edit distance improvement was 0.05.

We then conducted a number of experiments intended to identify the contribution made by different factors it incorporates. We found that all of the following ablations significantly hurt reconstruction: using a flat tree (in which all languages are equidistant from the reconstructed root and from each other) instead of the consensus tree, dropping the markedness features, dropping the faithfulness features, and disabling sharing across branches. The results of these experiments are shown in Table S.1.

For comparison, we also included in the same table the performance of a semi-supervised system trained by K-fold validation. The system was run $K = 5$ times, with $1 - K^{-1}$ of the POc words given to the system as observations in the graphical model for each run. It is semi-supervised in the sense that target reconstructions for many internal nodes are not available in the dataset, so they are still not filled.[3]

2.2 Comparisons against other methods

The first competing method, PRAGUE was introduced in [15]. In this method, the word forms in a given protolanguage are reconstructed using a Viterbi multi-alignment between a small number of its descendant languages. The alignment is computed using hand-set parameters. Deterministic rules characterizing changes between pairs of observed languages are extracted from the alignment when their frequency is higher than a threshold, and a proto-phoneme inventory is built using linguistically motivated rules and parsimony. A reconstruction of each observed word is first proposed independently for each language. If at least two reconstructions agree, a majority vote is taken, otherwise no reconstruction is proposed. This approach has several limitations. First, it is not tractable for larger trees, since the time complexity of their multi-alignment algorithm grows exponentially in the number of languages. Second, deterministic rules, while elegant in theory, are not robust to noise: even in experiments with only four daughter languages, a large fraction of the words could not be reconstructed.

Since PRAGUE does not scale to large datasets, we also built a second, more tractable baseline. This new baseline system, CENTROID, computes the centroid of the observed word forms in Levenshtein distance. Let

[2] While most word forms in ABVD are encoded using IPA, there are a few exceptions, for example language family specific conventions such as the usage of the okina symbol (') for glottal stops (ʔ) in some Polynesian languages or ad hoc conventions such as using the bigram 'ng' for the agma symbol (ŋ). We have preprocessed as many of these exceptions as possible, and probabilist methods are generally robust to reasonable amounts of encoding glitches.

[3] We also tried a fully supervised system where a flat topology is used so that all of these latent internal nodes are avoided; but it did not perform as well—this is consistent with the -Topology experiment of Table S.1.

$\mathrm{Lev}(x, y)$ denote the Levenshtein distance between word forms x and y. Ideally, we would like the baseline system to return:

$$\operatorname*{argmin}_{x \in \Sigma^*} \sum_{y \in O} \mathrm{Lev}(x, y),$$

where $O = \{y_1, \ldots, y_{|O|}\}$ is the set of observed word forms. This objective function is motivated by Bayesian decision theory [11], and shares similarity to the more sophisticated Bayes estimator described in Section 2. However it replaces the samples obtained by MCMC sampling by the set of observed words. Similarly to the algorithm of Section 2, we also restrict the minimization to be taken on $x \in \Sigma(O)^*$ such that $m \le |x| \le M$ and to strings built by at most k contiguous substrings taken from the word forms in O, where $m = \min_i |y_i|, M = \max_i |y_i|$ and $\Sigma(O)$ is the set of characters occurring in O. Again, we found that $k = 2$ often finds the same solution as higher values of k. The difference was in all the cases not statistically significant, so we report the approximation $k = 2$ in what follows.

We also compared against an oracle, denoted ORACLE, which returns

$$\operatorname*{argmin}_{y \in O} \mathrm{Lev}(y, x^*),$$

where x^* is the target reconstruction. We will denote it by ORACLE. This is superior to picking a single closest language to be used for all word forms, but it is possible for systems to perform better than the oracle since it has to return one of the observed word forms. Of course, this scheme is only available to assess system performance on held-out data: It cannot make new predictions.

We performed the comparison against a previous system proposed in [15] on the same dataset and experimental conditions as used in [15]. The PMJ dataset was compiled by [16], who also reconstructed the corresponding protolanguage. Since PRAGUE is not guaranteed to return a reconstruction for each cognate set, only 55 word forms could be directly compared to our system. We restricted comparison to this subset of the data. This favors PRAGUE since the system only proposes a reconstruction when it is certain. Still, our system outperformed PRAGUE, with an average distance of 1.60 compared to 2.02 for PRAGUE. The difference is marginally significant, $p = 0.06$, partly due to the small number of word forms involved.

To get a more extensive comparison, we considered the hybrid system that returns PRAGUE's reconstruction when possible and otherwise back off to the Sundanese (Snd.) modern form, then Madurese (Mad.), Malay (Mal.) and finally Javanic (Jv.) (the optimal back-off order). In this case, we obtained an edit distance of 1.86 using our system against 2.33 for PRAGUE, a statistically significant difference.

We also compared against ORACLE and CENTROID in a large-scale setting. Specifically, we compare to the experimental setup on 64 modern languages used to reconstruct POc described before. Encouragingly, while the system's average distance (1.49) does not attain that of the ORACLE (1.13), we significantly outperform the CENTROID baseline (1.79).

2.3 Cognate recovery

To test the effectiveness of our cognate recovery system, we ran our system on all of the Oceanic languages in the ABVD, which comprises roughly half of the Austronesian languages. We then evaluated the pairwise precision, recall, F1, and purity scores, defined as follows. Let $G = \{G_1, G_2, \ldots, G_g\}$ denote the known partitions of the forms into cognates, and let $F = \{F_1, F_2, \ldots, F_f\}$ denote the inferred partitions. Let $\mathrm{pairs}(F)$ denote the set of unordered pairs of indices in the same partition in F: $\mathrm{pairs}(F) = \{\{i, j\} : \exists k \text{ s.t. } i, j \in F_k, i \neq j\}$, and similarly for pairs(G). The metrics are defined as follows:

$$\begin{aligned}
\text{precision} &= \frac{|\text{pairs}(G) \cap \text{pairs}(F)|}{|\text{pairs}(F)|},\\
\text{recall} &= \frac{|\text{pairs}(G) \cap \text{pairs}(F)|}{|\text{pairs}(G)|},\\
\text{F1} &= 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},\\
\text{purity} &= \frac{1}{N}\sum_f \max_g |G_g \cap F_f|.
\end{aligned}$$

Using these metrics, we found that our system achieved a precision of 84.4, recall of 62.1, F1 of 71.5, and cluster purity of 91.8. Thus, over 9 out of 10 words are correctly grouped, and our system errs on the side of under-grouping words rather than clustering words that are not cognates. Since the null hypothesis in historical linguistics is to deem words to be unrelated unless proven otherwise, a slight under-grouping is the desired behavior.

Since we are ultimately interested in reconstruction, we then compared our reconstruction system's ability to reconstruct words given these automatically determined cognates. Specifically, we took every cognate group found by our system (run on the Oceanic subclade) with at least two words in it. That is, we excluded words that our system found to be isolates. Then, we automatically reconstructed the Proto-Oceanic ancestor of those words using our system (using the auxiliary variables z described in Section 1.2.1).

For evaluation, we then looked at the average edit distance from our reconstructions to the known reconstructions described in the previous sections. This time, however, we average *per modern word* rather than *per cognate group*, to provide a fairer comparison. (Results were not substantially different averaging per cognate group.)

Using known cognates from the ABVD, there was an average reconstruction error of 2.19, versus 2.47 for the automatically reconstructed cognates, or an increase in error rate of 12.8%. The fraction of words with each Levenshtein distance for these reconstructions is shown in Figure S.1. While the plots are similar, the automatic cognates exhibit a slightly longer tail. Thus, even with automatic cognates, the reconstruction system can reconstruct words faithfully in many cases, only failing in a few instances.

2.4 Computation of functional loads

To measure functional load quantitatively, we used the same estimator as the one used in [17]. This definition is based on associating a *context vector* $\boldsymbol{c}_{x,\ell}$ to each phoneme and language. For a given language with $N(\ell)$ phoneme tokens, these context vectors are defined as follows: first, fix an enumeration order of all the contexts found in the corpus, where the context is defined as a pair of phonemes, one at the left and one at the right of a position. Element i in this enumeration will correspond to component i of the context vectors. Then, the value of component i in context vector $\boldsymbol{c}_{x,\ell}$ is set to be the number of time phoneme x occurred in context i and language l. Finally, King's definition of functional load $\text{FL}_\ell(x, y)$ is the dot product of the two induced context vectors:

$$\text{FL}_\ell(x, y) = \frac{1}{N(\ell)^2}\langle \boldsymbol{c}_{x,\ell}, \boldsymbol{c}_{y,\ell}\rangle = \frac{1}{N(\ell)^2}\sum_i \boldsymbol{c}_{x,\ell}(i) \times \boldsymbol{c}_{y,\ell}(i),$$

where the denominator is simply a normalization that insures $\text{FL}_\ell(x, y) \leq 1$. Note that if x and y are in complementary distribution in language ℓ, then the two vectors $\boldsymbol{c}_{x,\ell}$ and $\boldsymbol{c}_{y,\ell}$ are orthogonal. The functional load is indeed zero in this case.

In Figure 3 of the main paper, we show heat maps where the color encodes the log of the number of sound changes that fall into a given 2-dimensional bin. Each sound change $x > y$ is encoded as pair of numbers in the unit interval, $(\hat{l}, \hat{m})$, where $\hat{l}$ is an estimate of the functional load of the pair and $\hat{m}$ is the posterior fraction of the instances of the phoneme x that undergo a change to y. We now describe how $\hat{l}, \hat{m}$ were estimated. The posterior fraction $\hat{m}$ for the merger $x > y$ between languages $\mathrm{pa}(\ell) \to \ell$ is easily computed from the same expected sufficient statistics used for parameter estimation:

$$\hat{m}_\ell(x > y) = \frac{\sum_{p\in\Sigma} N(S, x, p, \ell, y)}{\sum_{p'\in\Sigma}\sum_{y'\in\Sigma} N(S, x, p', \ell, y')}.$$

The estimate of the functional load requires additional statistics, i.e. the expected context vectors $\hat{\mathbf{c}}_{x,\ell}$ and expected phoneme token counts $\hat{N}(\ell)$, but these can be readily extracted from the output of the MCMC sampler. The estimate is then:

$$\hat{l}_\ell(x, y) = \frac{1}{\hat{N}(\ell)^2}\langle \hat{\mathbf{c}}_{x,\ell}, \hat{\mathbf{c}}_{y,\ell}\rangle.$$

Finally, the set of points used to construct the heat map is:

$$\left\{\left(\hat{l}_{\mathrm{pa}(\ell)}(x, y), \hat{m}_\ell(x > y)\right) : \ell \in L - \{\text{root}\}, x \in \Sigma, y \in \Sigma, x \neq y\right\}.$$

3 Reconstruction lists

In Table S.2–S.4, we show the lists of consensus reconstructions produced by our system (the ‘Automatic’ column). For comparison, we also include a baseline (randomly picking one modern word), and the edit distances (when it is greater than zero). In Proto-Oceanic, since two manual reconstructions are available, we include the distances of the automatic reconstruction to both manual reconstructions (‘P-A’ and ‘B-A’) and the distance between the two manual reconstructions (‘B-P’).

4 Sound changes

In Figures S.2–S.5, we zoom and rotate each quadrant of the tree shown in Figure 2 of the main paper. For information on the most frequent change of each branch, refer to the row in Table S.5 corresponding to the code in parenthesis attached to each branch. By most frequent, we mean the change with the highest expected count in the last EM iteration, collapsing contexts for simplicity. In cases where no change is observed with an expected count of more than 1.0, we skip the corresponding entry—this can be caused for example by languages where cognacy information was too sparse in the dataset. The functional load, normalized to be in the interval $[0, 1]$ is also shown for reference.

5 Frequent errors

In Figure S.6, we analyze the frequent discrepancy between the PAn reconstruction from our system with those from [18]. The most frequent problematic substitutions, insertions, and deletions are shown. These frequencies were obtained by aligning, after running the system, the reconstructions with the references. The aligner is a pair-HMM trained via EM, using only phoneme identity and gap information [19]. The frequencies were extracted from the posterior alignments.

References and Notes

[1] Hastie T, Tibshiranim R, Friedman J (2009) *The Elements of Statistical Learning* (Springer).

[2] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39:1–38.

[3] Liu DC, Nocedal J, Dong C (1989) On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45:503–528.

[4] Lunter GA, Miklós I, Song YS, Hein J (2003) An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.* 10:869–889.

[5] Tierney L (1994) Markov chains for exploring posterior distributions. *The Annals of Statistics* 22:1701–1728.

[6] Holmes I, Bruno WJ (2001) Evolutionary HMM: a Bayesian approach to multiple alignment. *Bioinformatics* 17:803–820.

[7] Bouchard-Côté A, Jordan MI, Klein D (2009) Efficient inference in phylogenetic InDel trees. *Advances in Neural Information Processing Systems* 21:177–184.

[8] Caffo B, Jank W, Jones G (2005) Ascent-based Monte Carlo EM. *Journal of the Royal Statistical Society - Series B* 67:235–252.

[9] Judea P (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* (Morgan Kaufmann).

[10] Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.

[11] Robert CP (2001) *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (Springer).

[12] Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10.

[13] Greenhill S, Blust R, Gray R (2008) The Austronesian basic vocabulary database: From bioinformatics to lexomics. *Evolutionary Bioinformatics* 4:271–283.

[14] Lewis MP, ed (2009) *Ethnologue: Languages of the World, Sixteenth edition.* (SIL International).

[15] Oakes M (2000) Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics* 7:233–244.

[16] Nothofer B (1975) *The reconstruction of Proto-Malayo-Javanic* (M. Nijhoff).

[17] King R (1967) Functional load and sound change. *Language* 43:831–852.

[18] Blust R (1999) Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. *Inst. Linguist. Acad. Sinica* 1:31–94.

[19] Berg-Kirkpatrick T, Bouchard-Côté A, DeNero J, Klein D (2010) Painless unsupervised learning with features. *Proceedings of the North American Conference on Computational Linguistics* pp 582–590.

Condition	Edit dist.
Unsupervised full system	1.87
-FAITHFULNESS	2.02
-MARKEDNESS	2.18
-Sharing	1.99
-Topology	2.06
Semi-supervised system	1.75

Table S.1: Effects of ablation of various aspects of our unsupervised system on mean edit distance to POc. -Sharing corresponds to the restriction to the subset of the features in OPERATION, FAITHFULNESS and MARKEDNESS that are branch-specific, -Topology corresponds to using a flat topology where the only edges in the tree connect modern languages to POc. The semi-supervised system is described in the text. All differences (compared to the unsupervised full system) are statistically significant.

Table S.2. Proto-Austronesian reconstructions

Gloss	Reference	Baseline	Distance	Automatic	Distance
tohold(1)	*ɡemɡem	*hiɡem	3	*ɡemɡem	
smoke(1)	*qebel	*qɨvɨl	3	*qebel	
toscratch(1)	*ka ɾaw	*kaɡaw	1	*ka ɾaw	
and(2)	*mah	*ma	1	*ma	1
leg/foot(1)	*qaqay	*ai	4	*qaqay	
shoulder(1)	*qaba ɾa	*vala	4	*qaba ɾa	
woman/female(1)	*bahi	*babinay	4	*vavaian	5
left(1)	*kawi ɾi	*kayli	3	*kawi ɾi	
day(1)	*qalejaw	*andew	5	*qalejaw	
mother(1)	*tina	*qina	1	*tina	
tosuck(1)	*sepsep	*sipsip	2	*sepsep	
small(2)	*kedi	*kedhi	1	*kedi	
night(1)	*be ɾŋi	*beyi	2	*be ɾŋi	
tosqueeze(1)	*pe ɾeq	*perah	3	*pereq	1
tohit(1)	*palu	*mipalo	3	*palu	
rat(1)	*labaw	*lapo	3	*kulavaw	3
you(1)	*ikamu	*kamo	2	*kamu	1
toplant(1)	*mula	*hɨmula	2	*mula	
toswell(1)	*ba ɾeq	*ba ɾeq		*ba ɾeq	
tosee(1)	*kita	*kita		*kita	
one(1)	*isa	*sa	1	*isa	
tosleep(1)	*tudu ɾ	*maturuɡ	4	*tudu ɾ	
dog(1)	*wasu	*kahu	2	*vatu	2
topound,beat(20)	*tutuh	*nutu	2	*tutu	1
stone(1)	*batu	*batu		*batu	
green(1)	*mataq	*mataq		*mataq	
father(1)	*tama	*qama	1	*tama	
this(1)	*ini	*eni	1	*ani	1
tooth(1)	*nipen	*lipon	2	*nipen	
tochoose(1)	*piliq	*piliʔ	1	*piliq	
star(1)	*bituqen	*bitun	2	*bituqen	
tobuy(23)	*baliw	*taiw	2	*taiw	2
tovomit(1)	*utaq	*mutjaq	2	*utaq	
towork(1)	*qumah	*quma	1	*quma	1
wide(1)	*malawas	*ɣawiɡ	5	*malabe ɾ	3
tocut,hack(1)	*ta ɾaq	*ta ɾaq		*ta ɾaq	
tofear(1)	*matakut	*takuʔ	3	*matakut	
tolive,bealive(1)	*maqudip	*maʔuŋi	3	*maqudip	
thunder(3)	*de ɾuŋ	*zuŋ	3	*deruŋ	1
tofly(2)	*layap	*layap		*layap	
toshoot(1)	*panaq	*fanak	2	*panaq	
name(1)	*ŋajan	*nɡaza	4	*ŋajan	
tobuy(1)	*beli	*poli	2	*beli	
and(1)	*ka	*kae	1	*ka	
when?(1)	*ijan	*piranɡ	3	*pijan	1
todig(1)	*kalih	*kali	1	*kali	1
ash(1)	*qabu	*abu	1	*qabu	
big(1)	*ma ɾaya	*ma ɾaya		*ma ɾaya	
tocut,hack(3)	*tektek	*tutek	2	*tektek	
road/path(1)	*zalan	*dalan	1	*zalan	
tostand(1)	*di ɾi	*diri	1	*di ɾi	
sand(1)	*qenay	*one	4	*qenay	

continued on next page

continued from previous page

Gloss	Reference	Baseline	Distance	Automatic	Distance
meat/flesh(31)	*isi	*ici	1	*isi	
what?(2)	*nanu	*anu	1	*anu	1
toswell(26)	* ɾibawa	*malifawa	4	*abeh	5
bird(2)	*qayam	*ayam	1	*qayam	
hand(1)	*lima	*lime	1	*lima	
in,inside(1)	*idalem	*dalale	4	*idalem	
if(2)	*nu	*no	1	*nu	
toknow,beknowledgeable(2)	*bajaq	*mafanaʔ	5	*mafanaʔ	5
at(20)	*di	*ri	1	*di	
wind(2)	*bali	*feli	2	*beliu	2
new(1)	*mabaqe ɾu	*baʔru	5	*vaquan	6
blood(1)	*da ɾaq	*da ɾaq		*da ɾaq	
breast(1)	*susu	*soso	2	*susu	
i(1)	*iaku	*ako	2	*iaku	
salt(1)	*qasi ɾa	*sie	4	*qasi ɾa	
toflow(1)	*qalu ɾ	*ilir	4	*qalu ɾ	
five(1)	*lima	*lima		*lima	
at(1)	*i	*i		*i	
other(1)	*duma	*duma		*duma	
leaf(2)	*bi ɾaq	*bela	3	*bela	3
all(1)	*amin	*kemon	3	*amin	
rotten(1)	*mabu ɾaq	*mavuk	4	*mabu ɾuk	2
tocook(1)	*tanek	*tanək	1	*tanek	
head(1)	*qulu	*ʔuɣuh	3	*qulu	
mouth(2)	*ŋusu	*ŋutu	1	*ŋuju	1
house(1)	* ɾumaq	*uma	2	* ɾumaq	
if(1)	*ka	*ke	1	*ka	
neck(1)	*liqe ɾ	*liqɨg	2	*liqe ɾ	
needle(1)	*za ɾum	*dagɨm	3	*za ɾum	
he/she(1)	*siia	*ia	2	*siia	
fruit(1)	*buaq	*buaq		*buaq	
back(1)	*likud	*likudeʔ	2	*likud	
tochew(2)	*qelqel	*qmelqel	1	*qmelqel	1
salt(2)	*timus	*timus		*timus	
we(2)	*kami	*sikami	2	*kami	
long(1)	*inaduq	*nandu	3	*anaduq	1
we(1)	*ikita	*itam	3	*kita	1
three(1)	*telu	*tɨlu	1	*telu	
lake(1)	*danaw	*ranu	3	*danaw	
toeat(1)	*kaen	*kman	2	*kman	2
no,not(3)	*ini	*ini		*ini	
where?(1)	*inu	*sadinno	5	*ainu	1
how?(1)	*kuja	*gagua	4	*kua	1
tothink(34)	*nemnem	*nɨmnɨm	2	*kinemnem	2
who?(2)	*siima	*cima	2	*tima	2
tobite(1)	*ka ɾat	*kagat	1	*ka ɾat	
tail(1)	*iku ɾ	*ikog	2	*iku ɾ	

Table S.3. Proto-Oceanic reconstructions

	Reconstructions			Pairwise distances		
Gloss	Blust (B)	Pawley (P)	Automatic (A)	B-P	P-A	B-A
fish(1)	*ikan	*ikan	*ikan			
five(1)	*lima	*lima	*lima			
what?(1)	*sapa	*saa	*sava	1	1	1
meat/flesh(1)	*pisiko	*pisako	*kiko	1	3	3
star(1)	*pituqun	*pituqon	*vetuqu	1	4	3
fog(1)	*kaput	*kaput	*kabu		2	2
toscratch(44)	*karu	*kadru	*kadru	1		1
shoulder(1)	*pa ɾa	*qapa ɾa	*vara	2	4	2
where?(3)	*pai	*pea	*vea	2	1	3
toclimb(2)	*sake	*sake	*cake		1	1
toeat(1)	*kani	*kani	*kani			
two(1)	*rua	*rua	*rua			
dry(11)	*maca	*masa	*mamasa	1	2	3
narrow(1)	*kopit	*kopit	*kapi		2	2
todig(1)	*keli	*keli	*keli			
bone(2)	*su ɾi	*su ɾi	*sui		1	1
stone(1)	*patu	*patu	*patu			
left(1)	*mawi ɾi	*mawi ɾi	*mawii		1	1
they(1)	*ira	*ira	*sira		1	1
toliedown(1)	*qinop	*qeno	*eno	2	1	3
tohide(1)	*puni	*puni	*vuni		1	1
rope(1)	*tali	*tali	*tali			
smoke(2)	*qasu	*qasu	*qasu			
when?(1)	*ŋaican	*ŋaijan	*ŋisa	1	3	3
we(2)	*kamami	*kami	*kami	2		2
this(1)	*ne	*ani	*eni	2	1	2
egg(1)	*qatolu ɾ	*katolu ɾ	*tolu	1	3	3
stick/wood(1)	*kayu	*kayu	*kai		2	2
tosit(16)	*nopo	*nopo	*nofo		1	1
toshoot(1)	*panaq	*pana	*pana	1		1
liver(1)	*qate	*qate	*qate			
needle(1)	*sa ɾum	*sa ɾum	*sau		2	2
feather(1)	*pulu	*pulu	*vulu		1	1
topound,beat(2)	*tutuk	*tuki	*tutuk	3	3	
near(9)	*tata	*tata	*tata			
heavy(1)	*mamat	*mapat	*mamava	1	3	2
year(1)	*taqun	*taqun	*taqu		1	1
old(1)	*matuqa	*matuqa	*matuqa			
fire(1)	*api	*api	*avi		1	1
tochoose(1)	*piliq	*piliq	*vili		2	2
rain(1)	*qusan	*qusan	*usa		2	2
togrow(1)	*tubuq	*tubuq	*tubu		1	1
tosee(1)	*kita	*kita	*kita			
tohear(1)	*roŋo ɾ	*roŋo ɾ	*roŋo		1	1
tochew(1)	*mamaq	*mamaq	*mama		1	1
louse(1)	*kutu	*kutu	*kutu			
wind(1)	*aŋin	*mataŋi	*mataŋi	4		4
bad,evil(1)	*saqat	*saqat	*saqa		1	1
hand(1)	*lima	*lima	*lima			
toflow(1)	*tape	*tape	*tave		1	1
night(1)	*boŋi	*boŋi	*boŋi			

continued on next page

continued from previous page

	Reconstructions			Pairwise distances		
Gloss	Blust (B)	Pawley (P)	Automatic (A)	B-P	P-A	B-A
day(5)	*qaco	*qaco	*qaso		1	1
tospit(14)	*qanusi	*qanusi	*aŋusu		3	3
person/humanbeing(1)	*taumataq	*tamwata	*tamata	3	1	2
tovomit(1)	*mumutaq	*mumuta	*muta	1	2	3
name(1)	*ŋajan	*qajan	*qasa	1	2	3
snake(12)	*mwata	*mwata	*mwata			
man/male(1)	*mwa ɾuqane	*taumwaqane	*mwane	5	5	4
tobreathe(1)	*manawa	*manawa	*manawa			
far(1)	*sauq	*sauq	*sau		1	1
tobuy(1)	*poli	*poli	*voli		1	1
tovomit(8)	*luaq	*luaq	*lua		1	1
tocook(9)	*tunu	*tunu	*tunu			
thick(3)	*matolu	*matolu	*matolu			
leg/foot(1)	*waqe	*waqe	*waqe			
tobite(1)	*ka ɾat	*ka ɾati	*karat	1	2	1
leaf(1)	*raun	*rau	*dau	1	1	2
sky(1)	*laŋit	*laŋit	*laŋi		1	1
todrink(1)	*inum	*inum	*inum			
tostand(2)	*tuqur	*taqur	*tuqu	1	2	1
i(1)	*au	*au	*yau		1	1
warm(1)	*mapanas	*mapanas	*mavana		2	2
moon(1)	*pulan	*pulan	*vula		2	2
how?(1)	*kua	*kuya	*kua	1	1	
three(1)	*tolu	*tolu	*tolu			
toplant(2)	*tanum	*tanom	*tanəm	1	1	1
mosquito(1)	*namuk	*namuk	*namu		1	1
bird(1)	*manuk	*manuk	*manu		1	1
four(1)	*pani	*pat	*vati	2	2	2
water(2)	*wai ɾ	*wai ɾ	*wai		1	1
one(1)	*sakai	*tasa	*sa	3	2	3
skin(1)	*kulit	*kulit	*kulit			
toyawn(1)	*mawap	*mawap	*mawa		1	1
he/she(1)	*ia	*ia	*ia			
nose(1)	*isuŋ	*ijuŋ	*isu	1	2	1
thatch/roof(1)	*qatop	*qatop	*qato		1	1
towalk(2)	*pano	*pano	*vano		1	1
flower(1)	*puŋa	*puŋa	*vuŋa		1	1
dust(1)	*qapuk	*qapuk	*avu		3	3
neck(18)	* ɾuqa	* ɾuqa	*ua		2	2
eye(1)	*mata	*mata	*mata			
father(1)	*tama	*tamana	*tama	2	2	
tofear(1)	*matakut	*matakut	*matakut			
root(2)	*waka ɾa	*waka ɾ	*waka	1	1	2
tostab,pierce(8)	*soka	*soka	*soka			
breast(1)	*susu	*susu	*susu			
tolive,bealive(1)	*maqurip	*maqurip	*maquri		1	1
head(1)	*qulu	*qulu	*qulu			
thou(1)	*ko	*iko	*kou	1	2	1
fruit(1)	*puaq	*puaq	*vua		2	2

Table S.4. Proto-Polynesian reconstructions

Gloss	Reference	Baseline	Distance	Automatic	Distance
rope(9)	*taula	*taura	1	*taula	
short(11)	*pukupuku	*puʔupuʔu	2	*pukupuku	
strikewithfist	*moto	*moko	1	*moto	
coral	*puŋa	*puŋa		*puŋa	
cook	*tafu	*tahu	1	*tafu	
painful,sick(1)	*masaki	*mahaki	1	*masaki	
downwards	*hifo	*ifo	1	*hifo	
dive	*ruku	*uku	1	*ruku	
toface	*haŋa	*haŋa		*haŋa	
grasp	*kapo	*ʔapo	1	*kapo	
bay	*faŋa	*hana	2	*faŋa	
tail	*siku	*siʔu	1	*siku	
toblow(6)	*pupusi	*pupuhi	1	*pupusi	
channel	*awa	*ava	1	*awa	
pandanus	*fara	*fala	1	*fara	
urinate	*mimi	*mimi		*mimi	
navel	*pito	*pito		*pito	
gall	*ʔahu	*au	2	*ʔahu	
wave	*ŋalu	*ŋalu		*ŋalu	
sleep	*mohe	*moe	1	*mohe	
warm(1)	*mafanafana	*mafanafana		*mafanafana	
beak	*ŋutu	*ŋutu		*ŋutu	
tooth	*nifo	*niho	1	*nifo	
dance	*saka	*haka	1	*saka	
leg	*waʔe	*wae	1	*waʔe	
drown	*lemo	*lemo		*lemo	
water	*wai	*vai	1	*wai	
nose	*isu	*isu		*isu	
taro	*talo	*kalo	1	*talo	
tohide(1)	*funi	*huna	2	*suna	2
upwards	*hake	*aʔe	2	*hake	
overripe	*peʔe	*pee	1	*peʔe	
tosit(16)	*nofo	*noho	1	*nofo	
red	*kula	*ʔula	1	*kula	
tochew(1)	*mama	*mama		*mama	
three(1)	*tolu	*tolu		*tolu	
tosleep(10)	*mohe	*moe	1	*mohe	
nine	*hiwa	*iwa	1	*hiwa	
dawn	*ata	*ata		*ata	
feather(1)	*fulu	*fulu		*fulu	
canoe	*waka	*vaʔa	2	*waka	
left(11)	*sema	*hema	1	*sema	
ashes	*refu	*lehu	2	*refu	
dew	*sau	*sau		*sau	
small	*riki	*riki		*riki	
house	*fale	*hale	1	*fale	
voice	*leʔo	*leo	1	*leʔo	
octopus	*feke	*heʔe	2	*feke	
toclimb(2)	*kake	*aʔe	2	*ake	1
sea	*tahi	*kai	2	*tahi	
day	*ʔaho	*ao	2	*ʔaho	
branch	*maŋa	*maŋa		*maŋa	

continued on next page

continued from previous page

Gloss	Reference	Baseline	Distance	Automatic	Distance
lovepity	*ʔaloʔofa	*aroha	5	*ʔaloʔofa	
flyrun	*lele	*rere	2	*lele	
thick(3)	*matolutolu	*matolutolu		*matolutoru	1
tostrippeel	*hisi	*hihi	1	*hisi	
sitdwell	*nofo	*noho	1	*nofo	
black	*kele	*ʔele	1	*kele	
torch	*rama	*ama	1	*rama	

Table S.5. Sound changes

Code		Parent			Child	Functional load	Number of occurrences
(1)	v	(ProtoOcean)	>	p	(AdmiraltyIslands)	0.09884	5.67800
(2)	a	(Northern Dumagat)	>	ʌ	(Agta)	5.132e-05	110.26986
(3)	q	(Bisayan)	>	ʔ	(AklanonBis)	0.03289	46.31015
(4)	a	(Schouten)	>	ə	(Ali)	7.158e-05	5.45255
(5)	e	(UlatInai)	>	a	(Alune)	1.00000	1.31832
(6)	e	(SeramStraits)	>	o	(Amahai)	0.05802	6.91108
(7)	a	(SouthwestNewBritain)	>	e	(Amara)	0.19090	15.52212
(8)	a	(CentralWestern)	>	e	(AmbaiYapen)	0.21003	5.74372
(9)	i	(SeramStraits)	>	a	(Ambon)	0.27139	3.04082
(10)	a	(EastVanuatu)	>	e	(AmbrymSout)	0.09085	14.42495
(11)	s	(BimaSumba)	>	h	(Anakalang)	0.03743	10.01598
(12)	a	(Vanuatu)	>	e	(AnejomAnei)	0.08559	14.43646
(13)	f	(Futunic)	>	p	(Anuta)	0.09282	19.96156
(14)	n	(Wetar)	>	ŋ	(Aputai)	0.04788	7.10435
(15)	t	(WestSanto)	>	r	(ArakiSouth)	0.21996	20.80080
(16)	*Not enough data available for reliable sound change estimates*						
(17)	d	(NorthPapuanMainlandDEntrecasteaux)	>	t	(AreTaupota)	0.06738	2.92590
(18)	ɔ	(Southern Malaita)	>	o	(AreareMaas)	0.01151	1.99913
(19)	ɔ	(Southern Malaita)	>	o	(AreareWaia)	0.01151	1.99971
(20)	a	(Bibling)	>	e	(Aria)	0.29446	4.99706
(21)	*Not enough data available for reliable sound change estimates*						
(22)	ɔ	(SanCristobal)	>	o	(ArosiOneib)	0.00562	1.99971
(23)	*Not enough data available for reliable sound change estimates*						
(24)	b	(ProtoCentr)	>	p	(Aru)	0.13711	9.35892
(25)	a	(RajaAmpat)	>	ɛ	(As)	0.05156	4.74243
(26)	o	(Utupua)	>	u	(Asumboa)	0.07562	3.70741
(27)	a	(Central Manobo)	>	o	(AtaTigwa)	0.00291	7.29063
(28)	s	(Formosan)	>	h	(Atayalic)	0.04098	4.31225
(29)	l	(West NuclearTimor)	>	n	(Atoni)	0.23175	7.02772
(30)	t	(Ibanagic)	>	q	(AttaPamplo)	0.47297	7.87084
(31)	a	(Suauic)	>	e	(Auhelawa)	0.18692	7.98362
(32)	ŋ	(MalekulaCentral)	>	n	(Avava)	0.10747	15.49566
(33)	a	(ProtoCentr)	>	e	(Babar)	0.04891	5.61192
(34)	ɔ	(Choiseul)	>	o	(Babatana)	0.01953	8.94227
(35)	ɔ	(Choiseul)	>	o	(BabatanaAv)	0.01953	1.99952
(36)	ɔ	(Choiseul)	>	o	(BabatanaKa)	0.01953	2.99952
(37)	*Not enough data available for reliable sound change estimates*						
(38)	ɔ	(Choiseul)	>	o	(BabatanaTu)	0.01953	2.99961
(39)	v	(Ivatan)	>	b	(Babuyan)	0.01574	16.97554
(40)	a	(BorneoCoastBajaw)	>	e	(Bajo)	0.04774	21.59109
(41)	a	(NuclearCordilleran)	>	ʌ	(Balangaw)	0.00387	37.89097
(42)	ŋ	(BaliSasak)	>	n	(Bali)	0.19166	13.91349
(43)	e	(ProtoMalay)	>	ə	(BaliSasak)	6.064e-04	26.22332
(44)	h	(BimaSumba)	>	s	(Baliledo)	0.03743	7.64100
(45)	p	(East CentralMaluku)	>	f	(BandaGeser)	0.05559	3.52479
(46)	a	(Sulawesi)	>	o	(BanggaiWdi)	0.06991	8.61615
(47)	a	(Palawano)	>	o	(Banggi)	6.735e-04	7.69053
(48)	e	(LocalMalay)	>	a	(BanjareseM)	0.10667	36.24790
(49)	a	(SouthNewIrelandNorthwestSolomonic)	>	o	(Banoni)	0.16106	5.69268
(50)	r	(Sangiric)	>	d	(Bantik)	0.01606	6.99267
(51)	b	(Sulawesi)	>	w	(Baree)	0.05030	10.71565
(52)	q	(ProtoMalay)	>	ʔ	(Barito)	0.04087	15.54090
(53)	e	(Madak)	>	o	(Barok)	0.04576	1.91240
(54)	u	(Northern EastFormosan)	>	o	(Basai)	0.00130	5.88492
(55)	u	(BashiicCentralLuzonNorthernMindoro)	>	o	(Bashiic)	0.02423	63.00884
(56)	r	(NorthernPhilippine)	>	y	(BashiicCentralLuzonNorthernMindoro)	0.01665	3.54936
(57)	ŋ	(Palawano)	>	n	(BatakPalaw)	0.04752	4.94491
(58)	ɔ	(SanCristobal)	>	o	(BauroBaroo)	0.00562	1.99981
(59)	*Not enough data available for reliable sound change estimates*						
(60)	*Not enough data available for reliable sound change estimates*						
(61)	p	(Vitiaz)	>	f	(Bel)	0.01771	1.06548
(62)	u	(BerawanLowerBaram)	>	o	(Belait)	0.03125	12.34195
(63)	f	(Futunic)	>	h	(Bellona)	0.01009	15.97130
(64)	t	(Pangasinic)	>	s	(Benguet)	0.09759	1.35377

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(65)	u	(BerawanLowerBaram)	>	o	(BerawanLon)	0.03125	13.55252
(66)	k	(NorthSarawakan)	>	ʔ	(BerawanLowerBaram)	0.15797	10.19426
(67)	e	(SouthwestNewBritain)	>	i	(Bibling)	0.03719	2.29657
(68)	ɔ	(RajaAmpat)	>	o	(BigaMisool)	0.03843	5.00642
(69)	k	(CentralPhilippine)	>	c	(BikolNagaC)	9.303e-05	12.87548
(70)	a	(Blaan)	>	ɔ	(BilaanKoro)	0.01132	23.88863
(71)	ŋ	(Blaan)	>	n	(BilaanSara)	0.08320	18.08416
(72)	*Not enough data available for reliable sound change estimates*						
(73)	u	(Northern NuclearBel)	>	o	(Bilibil)	0.03687	2.95970
(74)	a	(ProtoMalay)	>	ɨ	(Bilic)	0.00777	19.43024
(75)	ɔ	(SouthNewIrelandNorthwestSolomonic)	>	o	(Bilur)	0.00242	1.97890
(76)	t	(BimaSumba)	>	d	(Bima)	0.08028	9.97797
(77)	b	(ProtoCentr)	>	w	(BimaSumba)	0.08312	11.67021
(78)	a	(NorthSarawakan)	>	e	(Bintulu)	0.07029	9.74568
(79)	ɨ	(Manobo)	>	u	(Binukid)	0.03862	2.90829
(80)	ɨ	(CentralPhilippine)	>	u	(Bisayan)	0.01651	18.53014
(81)	ɨ	(Bilic)	>	a	(Blaan)	0.08598	17.06499
(82)	*Not enough data available for reliable sound change estimates*						
(83)	*Not enough data available for reliable sound change estimates*						
(84)	t	(GorontaloMongondow)	>	s	(BolaangMon)	0.03159	4.79938
(85)	o	(TukangbesiBonerate)	>	ɔ	(Bonerate)	0.02521	18.76480
(86)	u	(Seram)	>	i	(Bonfia)	0.14937	11.37310
(87)	u	(Bontok)	>	o	(BontocGuin)	0.04335	58.88829
(88)	a	(BontokKankanay)	>	ɨ	(Bontok)	0.10466	3.41602
(89)	q	(Bontok)	>	ʔ	(BontokGuin)	8.294e-04	49.64130
(90)	u	(NuclearCordilleran)	>	o	(BontokKankanay)	0.03806	9.39059
(91)	u	(SuluBorneo)	>	o	(BorneoCoastBajaw)	0.05043	3.53556
(92)	v	(GelaGuadalcanal)	>	p	(Bughotu)	0.09530	1.99015
(93)	a	(Bugis)	>	ə	(BugineseSo)	0.00654	17.32980
(94)	ŋ	(SouthSulawesi)	>	k	(Bugis)	0.10991	2.98167
(95)	s	(Bughotu)	>	h	(Bugotu)	0.03395	8.55895
(96)	e	(KayanMurik)	>	a	(Bukat)	0.06056	4.02644
(97)	a	(SouthHalmahera)	>	e	(Buli)	0.05583	3.50307
(98)	o	(Vanikoro)	>	ɔ	(Buma)	0.13019	23.27248
(99)	a	(Sabahan)	>	o	(BunduDusun)	0.05640	1.58162
(100)	u	(Formosan)	>	o	(Bunun)	2.254e-04	11.98634
(101)	u	(CentralMaluku)	>	o	(BuruNamrol)	0.01570	19.92112
(102)	o	(South Bisayan)	>	u	(ButuanTausug)	0.03947	3.20776
(103)	u	(ButuanTausug)	>	o	(Butuanon)	0.03983	30.32254
(104)	r	(NorthPapuanMainlandDEntrecasteaux)	>	l	(Bwaidoga)	0.10631	8.92246
(105)	a	(NewCaledonian)	>	ɛ	(Canala)	0.05974	4.64394
(106)	ŋ	(ProtoChuuk)	>	n	(Carolinian)	0.04042	23.82083
(107)	u	(Bisayan)	>	o	(Cebuano)	0.03656	8.68282
(108)	l	(SouthHalmaheraWestNewGuinea)	>	r	(CenderawasihBay)	0.11907	11.14761
(109)	u	(EastFormosan)	>	o	(CentralAmi)	4.190e-04	12.82795
(110)	u	(SouthCentralCordilleran)	>	o	(CentralCordilleran)	0.02941	3.43974
(111)	e	(ProtoMalay)	>	ə	(CentralEastern)	6.064e-04	32.49321
(112)	j	(ProtoOcean)	>	s	(CentralEasternOceanic)	0.00410	2.00058
(113)	e	(BashiicCentralLuzonNorthernMindoro)	>	i	(CentralLuzon)	0.01063	2.14101
(114)	ɾ	(ProtoCentr)	>	r	(CentralMaluku)	0.02692	12.80932
(115)	u	(MaselaSouthBabar)	>	ɛ	(CentralMas)	0.04503	4.21108
(116)	ŋ	(RemoteOceanic)	>	ɡ	(CentralPacific)	0.00869	11.59225
(117)	k	(Peripheral PapuanTip)	>	ɡ	(CentralPapuan)	0.11515	5.29705
(118)	u	(MesoPhilippine)	>	o	(CentralPhilippine)	0.01374	38.98219
(119)	x	(NortheastVanuatuBanksIslands)	>	k	(CentralVanuatu)	0.02455	3.46857
(120)	i	(CenderawasihBay)	>	u	(CentralWestern)	0.12684	1.99006
(121)	u	(Bisayan)	>	o	(Central Bisayan)	0.03656	7.64073
(122)	l	(East Nuclear Polynesian)	>	r	(Central East Nuclear Polynesian)	0.02709	2.35911
(123)	a	(Manobo)	>	ɨ	(Central Manobo)	0.09378	18.50241
(124)	a	(SantaIsabel)	>	u	(Central SantaIsabel)	0.27278	1.22175
(125)	t	(Chamic)	>	k	(ChamChru)	0.34269	7.73602
(126)	y	(Malayic)	>	i	(Chamic)	0.00237	1.73963
(127)	b	(ProtoMalay)	>	p	(Chamorro)	0.15451	10.03113
(128)	ɣ	(East SantaIsabel)	>	ɡ	(ChekeHolo)	0.03879	10.10942
(129)	o	(SouthNewIrelandNorthwestSolomonic)	>	ɔ	(Choiseul)	0.00242	8.56855
(130)	a	(ChamChru)	>	ə	(Chru)	0.02139	31.97181
(131)	a	(ProtoChuuk)	>	e	(Chuukese)	0.11108	36.12356
(132)	a	(ProtoChuuk)	>	e	(ChuukeseAK)	0.11108	59.55135
(133)	ə	(Atayalic)	>	a	(CiuliAtaya)	0.02870	6.00504
(134)	e	(North Babar)	>	o	(Dai)	0.02409	5.70633
(135)	n	(North Babar)	>	l	(DaweraDawe)	0.16784	23.27929
(136)	ŋ	(LandDayak)	>	n	(DayakBakat)	0.26785	7.13031
(137)	ŋ	(South West Barito)	>	n	(DayakNgaju)	0.16918	29.67016
(138)	a	(NorthSarawakan)	>	e	(Dayic)	0.07029	5.15706
(139)	a	(LoyaltyIslands)	>	e	(Dehu)	0.20861	4.46889
(140)	ɡ	(Bwaidoga)	>	y	(Diodio)	0.08008	12.32717
(141)	l	(NorthPapuanMainlandDEntrecasteaux)	>	r	(Dobuan)	0.10631	12.30865
(142)	p	(Southern Malaita)	>	b	(Dorio)	8.555e-04	1.99973
(143)	l	(Nuclear WestCentralPapuan)	>	r	(Doura)	0.13489	5.63596
(144)	u	(Northern Dumagat)	>	o	(DumagatCas)	0.01544	18.63984
(145)	s	(SouthwestMaluku)	>	h	(EastDamar)	0.02479	6.75499
(146)	a	(CentralPacific)	>	o	(EastFijianPolynesian)	0.23158	1.54042

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(147)	c	(Formosan)	>	t	(EastFormosan)	0.04224	6.87583
(148)	b	(MaselaSouthBabar)	>	v	(EastMasela)	0.00417	5.01447
(149)	i	(BimaSumba)	>	u	(EastSumban)	0.17438	28.36485
(150)	b	(NortheastVanuatuBanksIslands)	>	v	(EastVanuatu)	0.16115	2.78838
(151)	e	(Barito)	>	i	(East Barito)	0.02495	4.21938
(152)	p	(CentralMaluku)	>	f	(East CentralMaluku)	0.09093	7.47879
(153)	u	(Manus)	>	o	(East Manus)	0.01231	4.11061
(154)	ɡ	(NewGeorgia)	>	h	(East NewGeorgia)	0.02220	5.62094
(155)	a	(NuclearTimor)	>	e	(East NuclearTimor)	0.14170	3.51835
(156)	l	(Nuclear Polynesian)	>	r	(East Nuclear Polynesian)	0.04761	79.14382
(157)	ɔ	(SantaIsabel)	>	o	(East SantaIsabel)	0.02383	4.65059
(158)	ə	(CentralEastern)	>	o	(EasternMalayoPolynesian)	0.00303	18.76994
(159)	a	(RemoteOceanic)	>	e	(EasternOuterIslands)	0.14981	13.43554
(160)	a	(AdmiraltyIslands)	>	e	(Eastern AdmiraltyIslands)	0.16811	7.58136
(161)	a	(BandaGeser)	>	o	(ElatKeiBes)	0.05446	15.50349
(162)	r	(SamoicOutlier)	>	l	(Ellicean)	0.03331	5.08687
(163)	a	(Futunic)	>	e	(Emae)	0.20786	3.97652
(164)	e	(SouthwestBabar)	>	ɛ	(Emplawas)	0.18331	11.05576
(165)	*Not enough data available for reliable sound change estimates*						
(166)	i	(BimaSumba)	>	e	(EndeLio)	0.04706	5.22061
(167)	a	(Sumatra)	>	ə	(Enggano)	0.00149	1.61296
(168)	*Not enough data available for reliable sound change estimates*						
(169)	*Not enough data available for reliable sound change estimates*						
(170)	f	(Wetar)	>	h	(Erai)	0.03346	7.93007
(171)	a	(Vanuatu)	>	e	(Erromanga)	0.08559	16.54778
(172)	ɔ	(SanCristobal)	>	o	(Fagani)	0.00562	1.99961
(173)	ɔ	(SanCristobal)	>	o	(FaganiAguf)	0.00562	1.99952
(174)	ɔ	(SanCristobal)	>	o	(FaganiRihu)	0.00562	1.99690
(175)	*Not enough data available for reliable sound change estimates*						
(176)	u	(WesternPlains)	>	o	(Favorlang)	0.01234	20.09559
(177)	s	(EastFijianPolynesian)	>	c	(FijianBau)	0.04333	9.93879
(178)	f	(Timor)	>	w	(FloresLembata)	0.01025	7.38989
(179)	l	(ProtoAustr)	>	r	(Formosan)	0.01978	16.25315
(180)	r	(Futunic)	>	l	(FutunaAniw)	0.05835	2.94909
(181)	r	(Futunic)	>	l	(FutunaEast)	0.05835	48.57376
(182)	l	(SamoicOutlier)	>	r	(Futunic)	0.03331	74.90293
(183)	l	(WestCentralPapuan)	>	r	(Gabadi)	0.13055	5.24097
(184)	u	(Ibanagic)	>	o	(Gaddang)	0.01362	18.75014
(185)	e	(Are)	>	i	(Gapapaiwa)	0.06186	4.23744
(186)	a	(BimaSumba)	>	o	(GauraNggau)	0.13543	29.26141
(187)	a	(ProtoMalay)	>	ə	(Gayo)	0.00259	24.49089
(188)	r	(Northern NuclearBel)	>	z	(Gedaged)	0.02021	7.64916
(189)	c	(GelaGuadalcanal)	>	s	(Gela)	0.01050	2.08324
(190)	k	(SoutheastSolomonic)	>	ɡ	(GelaGuadalcanal)	0.01749	19.76693
(191)	b	(GeserGorom)	>	w	(Geser)	0.06536	3.91631
(192)	f	(BandaGeser)	>	w	(GeserGorom)	0.05946	4.12917
(193)	*Not enough data available for reliable sound change estimates*						
(194)	ɡ	(Guadalcanal)	>	ɣ	(Ghari)	4.995e-04	9.42542
(195)	*Not enough data available for reliable sound change estimates*						
(196)	h	(Guadalcanal)	>	ɣ	(GhariNggae)	5.670e-05	2.00000
(197)	ɔ	(Guadalcanal)	>	o	(GhariNgger)	0.00601	2.99863
(198)	*Not enough data available for reliable sound change estimates*						
(199)	*Not enough data available for reliable sound change estimates*						
(200)	a	(SouthHalmahera)	>	o	(Giman)	0.11966	11.83686
(201)	a	(GorontaloMongondow)	>	o	(Gorontalic)	0.12679	8.78982
(202)	k	(Gorontalic)	>	ʔ	(GorontaloH)	0.00973	18.65349
(203)	a	(Sulawesi)	>	o	(GorontaloMongondow)	0.06991	26.41042
(204)	s	(GelaGuadalcanal)	>	c	(Guadalcanal)	0.01050	3.50693
(205)	*Not enough data available for reliable sound change estimates*						
(206)	*Not enough data available for reliable sound change estimates*						
(207)	ŋ	(NehanNorthBougainville)	>	n	(Haku)	0.10208	6.26107
(208)	ɨ	(MesoPhilippine)	>	u	(Hanunoo)	0.02599	24.61094
(209)	t	(Marquesic)	>	k	(Hawaiian)	0.31942	56.48048
(210)	u	(Peripheral Central Bisayan)	>	o	(Hiligaynon)	0.02028	1.95625
(211)	r	(Ambon)	>	l	(HituAmbon)	0.03275	16.21990
(212)	ɡ	(West NewGeorgia)	>	ɣ	(Hoava)	0.00152	7.69893
(213)	a	(NorthNewGuinea)	>	e	(HuonGulf)	0.13866	4.84173
(214)	a	(LoyaltyIslands)	>	e	(Iaai)	0.20861	8.25767
(215)	u	(Malayic)	>	o	(Iban)	0.00621	12.93856
(216)	k	(NorthernCordilleran)	>	q	(Ibanagic)	0.40890	11.09862
(217)	ŋ	(Sabahan)	>	n	(Idaan)	0.13834	5.84917
(218)	e	(Futunic)	>	a	(IfiraMeleM)	0.20786	4.95522
(219)	ɨ	(NuclearCordilleran)	>	o	(Ifugao)	0.01254	19.55025
(220)	k	(Ifugao)	>	q	(IfugaoAmga)	0.34057	4.76075
(221)	k	(Ifugao)	>	q	(IfugaoBata)	0.34057	17.99754
(222)	ɨ	(Kallahan)	>	o	(IfugaoBayn)	0.01101	17.75466
(223)	n	(Wetar)	>	ŋ	(Iliun)	0.04788	10.14036
(224)	u	(NorthernLuzon)	>	o	(Ilokano)	0.02660	29.66945
(225)	a	(Peripheral Central Bisayan)	>	u	(Ilonggo)	0.12642	4.07172
(226)	a	(SouthernCordilleran)	>	ɨ	(Ilongot)	0.07296	10.35510
(227)	ŋ	(Ilongot)	>	n	(IlongotKak)	0.06631	9.13131
(228)	a	(Ivatan)	>	e	(Imorod)	0.08734	6.03794

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(229)	n	(SouthwestBabar)	>	m	(Imroing)	0.23281	10.35889
(230)	u	(SamaBajaw)	>	o	(Inabaknon)	0.02973	19.36631
(231)	a	(LocalMalay)	>	e	(Indonesian)	0.10667	3.03924
(232)	u	(Benguet)	>	o	(Inibaloi)	0.03849	64.30170
(233)	y	(MaranaoIranon)	>	i	(Iranun)	0.00959	4.31410
(234)	a	(Ivatan)	>	e	(Iraralay)	0.08734	11.08683
(235)	l	(Ivatan)	>	d	(Isamorong)	0.01340	14.95905
(236)	t	(Ibanagic)	>	s	(IsnegDibag)	0.05600	2.91874
(237)	h	(Ivatan)	>	x	(Itbayat)	0.00445	31.26009
(238)	o	(Ivatan)	>	u	(Itbayaten)	5.820e-04	67.10087
(239)	u	(KalingaItneg)	>	o	(ItnegBinon)	0.02352	60.52923
(240)	l	(Ivatan)	>	d	(Ivasay)	0.01340	14.96335
(241)	g	(Bashiic)	>	y	(Ivatan)	0.02574	5.02873
(242)	e	(Ivatan)	>	ɨ	(IvatanBasc)	0.00723	32.77267
(243)	q	(ProtoMalay)	>	h	(Javanese)	0.07129	15.44980
(244)	a	(Northern NewCaledonian)	>	e	(Jawe)	0.13215	8.47484
(245)	e	(West Barito)	>	o	(Kadorih)	7.673e-05	3.80522
(246)	ɔ	(SanCristobal)	>	o	(Kahua)	0.00562	1.99971
(247)	ɔ	(SanCristobal)	>	o	(KahuaMami)	0.00562	1.99952
(248)	l	(Gorontalic)	>	r	(Kaidipang)	0.01053	21.87631
(249)	k	(KairiruManam)	>	q	(Kairiru)	0.00897	10.55125
(250)	u	(Schouten)	>	i	(KairiruManam)	0.17315	1.37804
(251)	n	(Ilongot)	>	ŋ	(KakidugenI)	0.06631	9.38577
(252)	r	(Mansakan)	>	l	(Kalagan)	0.04611	3.61546
(253)	i	(MesoPhilippine)	>	ɨ	(Kalamian)	0.02195	3.10459
(254)	ɨ	(KalingaItneg)	>	o	(KalingaGui)	0.01076	22.89422
(255)	a	(CentralCordilleran)	>	i	(KalingaItneg)	0.13957	2.25237
(256)	s	(Benguet)	>	h	(Kallahan)	0.02667	15.50239
(257)	s	(Kallahan)	>	h	(KallahanKa)	0.00566	1.92678
(258)	ɨ	(Kallahan)	>	e	(KallahanKe)	0.00403	38.49811
(259)	n	(BimaSumba)	>	ŋ	(Kambera)	0.00600	25.28750
(260)	b	(Tsouic)	>	v	(Kanakanabu)	0.01715	4.07361
(261)	v	(PatpatarTolai)	>	w	(Kandas)	0.03857	3.91962
(262)	u	(BontokKankanay)	>	o	(KankanayNo)	0.04380	49.14539
(263)	n	(CentralLuzon)	>	ŋ	(Kapampanga)	0.01456	6.07915
(264)	t	(Ellicean)	>	d	(Kapingamar)	3.974e-05	35.94977
(265)	v	(LavongaiNalik)	>	f	(KaraWest)	0.02487	4.21517
(266)	a	(SouthHalmahera)	>	e	(KasiraIrah)	0.05583	5.86174
(267)	e	(South West Barito)	>	ɛ	(Katingan)	0.00426	27.03064
(268)	ɪ	(Pasismanua)	>	i	(KaulongAuV)	0.01894	6.17672
(269)	a	(Northern EastFormosan)	>	i	(Kavalan)	0.08596	4.99960
(270)	b	(ProtoMalay)	>	v	(KayanMurik)	1.449e-07	6.64852
(271)	a	(KayanMurik)	>	e	(KayanUmaJu)	0.06056	5.93656
(272)	a	(SarmiJayapuraBay)	>	e	(KayupulauK)	0.32087	3.93441
(273)	a	(FloresLembata)	>	e	(Kedang)	0.10748	14.89342
(274)	b	(KeiTanimbar)	>	β	(KeiTanimba)	4.198e-05	4.13306
(275)	a	(SoutheastMaluku)	>	e	(KeiTanimbar)	0.17157	1.74318
(276)	a	(Dayic)	>	e	(KelabitBar)	0.07691	10.95832
(277)	e	(East NuclearTimor)	>	ɛ	(Kemak)	2.255e-04	5.94197
(278)	i	(NorthSarawakan)	>	e	(KenyahLong)	0.03423	6.39511
(279)	a	(LocalMalay)	>	o	(Kerinci)	0.01029	37.28182
(280)	a	(KilivilaLouisiades)	>	e	(Kilivila)	0.14778	5.83519
(281)	o	(Peripheral PapuanTip)	>	a	(KilivilaLouisiades)	0.17286	2.23572
(282)	o	(Central SantaIsabel)	>	ɔ	(KilokakaYs)	0.02707	5.60509
(283)	ŋ	(MicronesianProper)	>	n	(Kiribati)	0.04469	10.23912
(284)	ʔ	(Manam)	>	k	(Kis)	0.07162	3.84299
(285)	t	(KisarRoma)	>	k	(Kisar)	0.05336	24.56857
(286)	f	(SouthwestMaluku)	>	w	(KisarRoma)	0.03321	7.75331
(287)	a	(BimaSumba)	>	o	(Kodi)	0.13543	22.04728
(288)	l	(ProtoCentr)	>	r	(KoiwaiIria)	0.06793	11.95360
(289)	ɔ	(Central SantaIsabel)	>	o	(Kokota)	0.02707	31.07342
(290)	e	(Pesisir)	>	o	(Komering)	0.00325	8.58447
(291)	a	(Blaan)	>	ɔ	(KoronadalB)	0.01132	30.81323
(292)	s	(NgeroVitiaz)	>	r	(Kove)	0.06561	7.13972
(293)	u	(PatpatarTolai)	>	a	(Kuanua)	0.16803	1.98361
(294)	ɔ	(West NewGeorgia)	>	o	(Kubokota)	0.00573	2.00145
(295)	v	(Nuclear WestCentralPapuan)	>	b	(Kuni)	0.11533	6.21112
(296)	*Not enough data available for reliable sound change estimates*						
(297)	a	(MicronesianProper)	>	e	(Kusaie)	0.12251	14.33579
(298)	ɔ	(Northern Malaita)	>	o	(Kwai)	0.00348	1.99952
(299)	a	(Northern Malaita)	>	o	(Kwaio)	0.18875	8.31055
(300)	l	(Tanna)	>	r	(Kwamera)	0.05929	25.71039
(301)	ŋ	(Northern Malaita)	>	n	(KwaraaeSol)	0.03728	9.80996
(302)	*Not enough data available for reliable sound change estimates*						
(303)	e	(MelanauKajang)	>	ə	(Lahanan)	0.00231	14.95868
(304)	i	(Willaumez)	>	e	(Lakalai)	0.11589	1.41234
(305)	r	(Nuclear WestCentralPapuan)	>	l	(Lala)	0.13489	4.70309
(306)	u	(FloresLembata)	>	o	(LamaholotI)	0.02895	10.53359
(307)	*Not enough data available for reliable sound change estimates*						
(308)	s	(BimaSumba)	>	h	(Lamboya)	0.03743	9.36744
(309)	o	(Bibling)	>	u	(LamogaiMul)	0.12203	5.67124
(310)	a	(Pesisir)	>	ə	(Lampung)	0.00929	12.80317

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(311)	e	(ProtoMalay)	>	a	(LandDayak)	0.04499	11.18206
(312)	*Not enough data available for reliable sound change estimates*						
(313)	k	(Northern Malaita)	>	ɡ	(Lau)	0.07399	10.37078
(314)	*Not enough data available for reliable sound change estimates*						
(315)	*Not enough data available for reliable sound change estimates*						
(316)	r	(NewIreland)	>	l	(LavongaiNalik)	0.13377	4.12434
(317)	a	(East Manus)	>	e	(Leipon)	0.13133	15.77294
(318)	a	(Tanna)	>	e	(Lenakel)	0.05393	9.10010
(319)	*Not enough data available for reliable sound change estimates*						
(320)	h	(Gela)	>	ɣ	(LengoGhaim)	5.895e-05	1.98182
(321)	*Not enough data available for reliable sound change estimates*						
(322)	f	(SouthwestMaluku)	>	w	(Letinese)	0.03321	8.50360
(323)	n	(West Manus)	>	ŋ	(Levei)	0.18931	43.18724
(324)	a	(LavongaiNalik)	>	e	(LihirSungl)	0.07781	10.60106
(325)	i	(West Manus)	>	e	(Likum)	0.08687	4.08584
(326)	e	(EndeLio)	>	ə	(LioFloresT)	0.00290	10.17585
(327)	a	(Malayan)	>	e	(LocalMalay)	0.09530	18.43396
(328)	*Not enough data available for reliable sound change estimates*						
(329)	r	(Manus)	>	ʔ	(Loniu)	0.00619	8.93647
(330)	a	(SoutheastIslands)	>	e	(Lou)	0.09902	12.46015
(331)	i	(LocalMalay)	>	e	(LowMalay)	0.03770	2.99404
(332)	a	(RemoteOceanic)	>	e	(LoyaltyIslands)	0.14981	15.79651
(333)	t	(Ellicean)	>	k	(Luangiua)	0.47404	39.73227
(334)	e	(MonoUruava)	>	a	(LungaLunga)	0.13896	3.85273
(335)	ɔ	(West NewGeorgia)	>	o	(Lungga)	0.00573	2.00155
(336)	ɔ	(West NewGeorgia)	>	o	(Luqa)	0.00573	2.00145
(337)	b	(East Barito)	>	w	(Maanyan)	0.07537	6.13215
(338)	a	(NewIreland)	>	e	(Madak)	0.16211	8.90472
(339)	k	(Madak)	>	ɡ	(MadakLamas)	0.05218	9.51948
(340)	l	(LavongaiNalik)	>	r	(Madara)	0.10006	10.95189
(341)	u	(ProtoMalay)	>	o	(Madurese)	0.01585	39.16274
(342)	l	(CentralPapuan)	>	r	(MagoriSout)	0.12605	3.95178
(343)	a	(Nuclear PapuanTip)	>	i	(Maisin)	0.22933	2.17692
(344)	n	(SouthSulawesi)	>	ŋ	(Makassar)	0.18370	10.70075
(345)	k	(MalaitaSanCristobal)	>	ʔ	(Malaita)	0.09152	5.02310
(346)	v	(SoutheastSolomonic)	>	f	(MalaitaSanCristobal)	0.06368	25.62607
(347)	*Not enough data available for reliable sound change estimates*						
(348)	ŋ	(LocalMalay)	>	n	(MalayBahas)	0.17169	34.31949
(349)	p	(Malayic)	>	m	(Malayan)	0.17398	3.33605
(350)	q	(ProtoMalay)	>	h	(Malayic)	0.07129	31.25704
(351)	a	(Vanuatu)	>	e	(MalekulaCentral)	0.08559	25.93669
(352)	a	(NortheastVanuatuBanksIslands)	>	e	(MalekulaCoastal)	0.08702	31.59187
(353)	a	(Vitiaz)	>	o	(Maleu)	0.11223	11.90975
(354)	e	(Bugis)	>	a	(Maloh)	0.07175	4.70755
(355)	u	(CentralPhilippine)	>	o	(Mamanwa)	0.01883	38.86565
(356)	u	(East NuclearTimor)	>	a	(Mambai)	0.34369	4.99285
(357)	e	(BimaSumba)	>	a	(Mamboru)	0.13262	7.46411
(358)	k	(KairiruManam)	>	ʔ	(Manam)	0.01548	9.58872
(359)	a	(Marquesic)	>	e	(Mangareva)	0.26245	3.72483
(360)	s	(BimaSumba)	>	c	(Manggarai)	2.420e-05	8.94154
(361)	r	(Tahitic)	>	l	(Manihiki)	9.361e-04	13.67126
(362)	ŋ	(SouthernPhilippine)	>	n	(Manobo)	0.07976	13.17836
(363)	ɨ	(AtaTigwa)	>	o	(ManoboAtad)	0.03723	66.54667
(364)	ɨ	(AtaTigwa)	>	o	(ManoboAtau)	0.03723	66.57068
(365)	ɨ	(Central Manobo)	>	a	(ManoboDiba)	0.12386	3.76411
(366)	ɡ	(West Central Manobo)	>	h	(ManoboIlia)	0.03983	13.28241
(367)	a	(South Manobo)	>	ɨ	(ManoboKala)	0.12673	7.79055
(368)	ɨ	(South Manobo)	>	ʌ	(ManoboSara)	2.271e-04	58.20252
(369)	o	(AtaTigwa)	>	ɨ	(ManoboTigw)	0.03723	11.93633
(370)	a	(West Central Manobo)	>	ɨ	(ManoboWest)	0.13743	4.54113
(371)	l	(Mansakan)	>	r	(Mansaka)	0.04611	18.44830
(372)	o	(CentralPhilippine)	>	u	(Mansakan)	0.01883	21.46241
(373)	a	(Eastern AdmiraltyIslands)	>	e	(Manus)	0.13652	4.75105
(374)	v	(Tahitic)	>	w	(Maori)	0.00255	12.29480
(375)	l	(BorneoCoastBajaw)	>	w	(Mapun)	0.03973	7.82813
(376)	u	(MaranaoIranon)	>	o	(Maranao)	0.00377	54.21909
(377)	ɨ	(SouthernPhilippine)	>	e	(MaranaoIranon)	0.00422	7.67944
(378)	*Not enough data available for reliable sound change estimates*						
(379)	*Not enough data available for reliable sound change estimates*						
(380)	*Not enough data available for reliable sound change estimates*						
(381)	*Not enough data available for reliable sound change estimates*						
(382)	s	(East NewGeorgia)	>	c	(Marovo)	0.00144	4.87889
(383)	a	(Marquesic)	>	e	(Marquesan)	0.26245	10.48977
(384)	f	(Central East Nuclear Polynesian)	>	h	(Marquesic)	0.05235	2.06222
(385)	a	(MicronesianProper)	>	e	(Marshalles)	0.12251	42.07442
(386)	i	(South Babar)	>	ɛ	(MaselaSouthBabar)	0.04148	3.02753
(387)	e	(Northern NuclearBel)	>	i	(Matukar)	0.04103	3.86380
(388)	t	(Willaumez)	>	f	(Maututu)	0.00143	5.97137
(389)	*Not enough data available for reliable sound change estimates*						
(390)	*Not enough data available for reliable sound change estimates*						
(391)	*Not enough data available for reliable sound change estimates*						
(392)	*Not enough data available for reliable sound change estimates*						

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(393)	*Not enough data available for reliable sound change estimates*						
(394)	e	(Northern NuclearBel)	>	i	(Megiar)	0.04103	3.96950
(395)	b	(Nuclear WestCentralPapuan)	>	p	(Mekeo)	0.01057	10.53926
(396)	e	(Northwest)	>	a	(MelanauKajang)	0.05269	4.10012
(397)	a	(MelanauKajang)	>	e	(MelanauMuk)	0.05542	8.30760
(398)	ŋ	(LocalMalay)	>	n	(Melayu)	0.17169	15.57376
(399)	e	(LocalMalay)	>	a	(MelayuBrun)	0.10667	57.81798
(400)	*Not enough data available for reliable sound change estimates*						
(401)	r	(Vitiaz)	>	l	(Mengen)	0.09236	7.76084
(402)	a	(WestSanto)	>	e	(Merei)	0.12570	4.40581
(403)	u	(East Barito)	>	o	(MerinaMala)	0.00538	45.42570
(404)	p	(WesternOceanic)	>	b	(MesoMelanesian)	0.06671	3.25436
(405)	e	(ProtoMalay)	>	ɨ	(MesoPhilippine)	0.00192	27.69856
(406)	s	(ProtoMicro)	>	t	(MicronesianProper)	0.14436	10.32542
(407)	e	(Malayan)	>	a	(Minangkaba)	0.09530	36.65853
(408)	b	(RajaAmpat)	>	p	(Minyaifuin)	0.08190	5.44290
(409)	r	(KilivilaLouisiades)	>	l	(Misima)	0.09569	2.85853
(410)	u	(Malayic)	>	o	(Moken)	0.00621	18.67524
(411)	a	(Ponapeic)	>	ɔ	(Mokilese)	0.00863	20.21510
(412)	k	(Bwaidoga)	>	ʔ	(Molima)	0.02226	6.39306
(413)	r	(MonoUruava)	>	l	(Mono)	0.18018	7.90990
(414)	*Not enough data available for reliable sound change estimates*						
(415)	*Not enough data available for reliable sound change estimates*						
(416)	ŋ	(SouthNewIrelandNorthwestSolomonic)	>	n	(MonoUruava)	0.07198	4.50934
(417)	t	(CenderawasihBay)	>	ʔ	(Mor)	0.00191	7.90698
(418)	a	(Sulawesi)	>	o	(Mori)	0.06991	15.10951
(419)	d	(ProtoChuuk)	>	t	(Mortlockes)	0.10821	17.95820
(420)	v	(EastVanuatu)	>	w	(Mota)	0.04447	4.78867
(421)	v	(SinagoroKeapara)	>	h	(Motu)	0.01870	13.39891
(422)	r	(Bibling)	>	x	(Mouk)	3.176e-05	16.89672
(423)	u	(Sulawesi)	>	o	(MunaButon)	0.04575	3.73293
(424)	ŋ	(Western Munic)	>	n	(MunaKatobu)	2.731e-04	6.02727
(425)	d	(UlatInai)	>	r	(MurnatenAl)	0.00181	5.95393
(426)	r	(ProtoOcean)	>	l	(Mussau)	0.16171	3.95147
(427)	a	(EastVanuatu)	>	ɛ	(Mwotlap)	0.01111	27.83527
(428)	r	(EndeLio)	>	z	(Nage)	0.02129	1.96583
(429)	i	(MalekulaCoastal)	>	e	(Nahavaq)	0.05885	9.63004
(430)	n	(Willaumez)	>	l	(NakanaiBil)	0.00412	1.02690
(431)	a	(LavongaiNalik)	>	ə	(Nalik)	0.00211	12.96772
(432)	u	(CentralVanuatu)	>	i	(Namakir)	0.18127	21.52549
(433)	a	(MalekulaCentral)	>	e	(Naman)	0.13582	33.25250
(434)	b	(MalekulaCoastal)	>	p	(Nati)	0.01049	19.54611
(435)	r	(SoutheastIslands)	>	l	(Nauna)	0.10938	5.60587
(436)	a	(ProtoMicro)	>	e	(Nauru)	0.13661	5.67814
(437)	*Not enough data available for reliable sound change estimates*						
(438)	s	(NehanNorthBougainville)	>	h	(Nehan)	0.05608	13.16549
(439)	ŋ	(Nehan)	>	n	(NehanHape)	0.11803	18.93949
(440)	a	(SouthNewIrelandNorthwestSolomonic)	>	o	(NehanNorthBougainville)	0.16106	4.34745
(441)	e	(Northern NewCaledonian)	>	a	(Nelemwa)	0.13215	10.43158
(442)	o	(Utupua)	>	œ	(Nembao)	0.01621	4.98208
(443)	a	(LoyaltyIslands)	>	e	(Nengone)	0.20861	5.30862
(444)	m	(Vanuatu)	>	n	(Nese)	0.35703	18.91500
(445)	a	(MalekulaCentral)	>	e	(Neveei)	0.13582	34.11813
(446)	e	(LoyaltyIslands)	>	a	(NewCaledonian)	0.20861	1.71479
(447)	k	(SouthNewIrelandNorthwestSolomonic)	>	g	(NewGeorgia)	0.08381	5.00348
(448)	e	(MesoMelanesian)	>	i	(NewIreland)	0.06281	3.31247
(449)	w	(BimaSumba)	>	v	(Ngadha)	4.053e-05	14.09575
(450)	i	(Aru)	>	e	(NgaiborSAr)	0.02299	6.77784
(451)	f	(NorthNewGuinea)	>	w	(NgeroVitiaz)	0.01667	3.53470
(452)	*Not enough data available for reliable sound change estimates*						
(453)	s	(Gela)	>	h	(Nggela)	0.05008	17.07009
(454)	b	(CentralVanuatu)	>	p	(Nguna)	0.06113	6.26520
(455)	p	(Sumatra)	>	f	(Nias)	0.00205	8.90278
(456)	f	(NilaSerua)	>	h	(Nila)	0.01125	4.39065
(457)	t	(TeunNilaSerua)	>	l	(NilaSerua)	0.13648	1.88492
(458)	u	(Nehan)	>	w	(Nissan)	0.02937	8.95963
(459)	a	(Tongic)	>	e	(Niue)	0.18204	4.31952
(460)	l	(North Babar)	>	n	(NorthBabar)	0.16784	6.84401
(461)	ɾ	(ProtoCentr)	>	r	(NorthBomberai)	0.02692	10.12083
(462)	v	(WesternOceanic)	>	p	(NorthNewGuinea)	0.12924	8.71798
(463)	a	(Nuclear PapuanTip)	>	e	(NorthPapuanMainlandDEntrecasteaux)	0.27192	3.13375
(464)	a	(Northwest)	>	e	(NorthSarawakan)	0.05269	5.44048
(465)	e	(Babar)	>	ɛ	(North Babar)	0.03978	1.27682
(466)	b	(Sulawesi)	>	w	(North Minahasan)	0.05030	3.03447
(467)	u	(Vanuatu)	>	o	(NortheastVanuatuBanksIslands)	0.06513	4.83269
(468)	r	(NorthernLuzon)	>	g	(NorthernCordilleran)	0.01688	4.17784
(469)	m	(NorthernPhilippine)	>	n	(NorthernLuzon)	0.15955	5.87812
(470)	ŋ	(ProtoMalay)	>	n	(NorthernPhilippine)	0.10751	18.96525
(471)	o	(NorthernCordilleran)	>	u	(Northern Dumagat)	0.01648	1.25949
(472)	l	(EastFormosan)	>	n	(Northern EastFormosan)	0.14981	5.83985
(473)	p	(Malaita)	>	b	(Northern Malaita)	0.00727	4.99257
(474)	a	(NewCaledonian)	>	o	(Northern NewCaledonian)	0.17212	3.00312

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(475)	b	(Bel)	>	p	(Northern NuclearBel)	0.08564	2.04429
(476)	ŋ	(Sangiric)	>	n	(Northern Sangiric)	0.10504	18.92173
(477)	q	(ProtoMalay)	>	ʔ	(Northwest)	0.04087	30.80231
(478)	l	(CentralCordilleran)	>	k	(NuclearCordilleran)	0.14865	1.01516
(479)	b	(Timor)	>	f	(NuclearTimor)	0.08459	2.80832
(480)	l	(PapuanTip)	>	n	(Nuclear PapuanTip)	0.10099	4.22343
(481)	ŋ	(Polynesian)	>	n	(Nuclear Polynesian)	0.01742	5.34206
(482)	r	(WestCentralPapuan)	>	l	(Nuclear WestCentralPapuan)	0.13055	1.52585
(483)	t	(Ellicean)	>	d	(Nukuoro)	3.974e-05	48.84748
(484)	f	(HuonGulf)	>	w	(NumbamiSib)	0.03605	5.99986
(485)	t	(CenderawasihBay)	>	k	(Numfor)	0.18043	10.67064
(486)	f	(Seram)	>	h	(Nunusaku)	0.02050	9.64666
(487)	a	(LocalMalay)	>	ə	(Ogan)	0.00113	22.90283
(488)	ŋ	(Javanese)	>	n	(OldJavanes)	0.12968	22.81888
(489)	t	(CentralVanuatu)	>	r	(Orkon)	0.18595	20.59925
(490)	ɔ	(Southern Malaita)	>	o	(Oroha)	0.01151	1.99952
(491)	r	(EastVanuatu)	>	l	(PaameseSou)	0.17094	18.23342
(492)	s	(Formosan)	>	t	(Paiwan)	0.10317	11.06529
(493)	a	(ProtoMalay)	>	e	(Palauan)	0.04499	11.70163
(494)	ɨ	(Palawano)	>	ə	(PalawanBat)	4.146e-04	36.88428
(495)	ɨ	(MesoPhilippine)	>	u	(Palawano)	0.02599	5.66511
(496)	*Not enough data available for reliable sound change estimates*						
(497)	u	(Pangasinic)	>	o	(Pangasinan)	0.03899	25.70031
(498)	*Not enough data available for reliable sound change estimates*						
(499)	a	(WesternOceanic)	>	o	(PapuanTip)	0.15356	3.56136
(500)	ŋ	(SouthwestNewBritain)	>	n	(Pasismanua)	0.09977	3.11319
(501)	v	(PatpatarTolai)	>	h	(Patpatar)	0.00182	8.75307
(502)	a	(SouthNewIrelandNorthwestSolomonic)	>	e	(PatpatarTolai)	0.15474	6.20403
(503)	e	(SeramStraits)	>	i	(Paulohi)	0.18832	3.98409
(504)	u	(Formosan)	>	o	(Pazeh)	2.254e-04	6.81642
(505)	f	(Tahitic)	>	h	(Penrhyn)	0.02693	5.16875
(506)	n	(Wetar)	>	ŋ	(Perai)	0.04788	4.13934
(507)	u	(Central Bisayan)	>	o	(Peripheral Central Bisayan)	0.02827	1.93754
(508)	e	(PapuanTip)	>	a	(Peripheral PapuanTip)	0.19334	3.54732
(509)	q	(ProtoMalay)	>	h	(Pesisir)	0.07129	14.80076
(510)	v	(EastVanuatu)	>	f	(PeteraraMa)	0.01242	17.18743
(511)	a	(ChamChru)	>	ɨ	(PhanRangCh)	0.00636	12.19166
(512)	*Not enough data available for reliable sound change estimates*						
(513)	v	(EastFijianPolynesian)	>	f	(Polynesian)	0.05871	17.52595
(514)	a	(Ponapeic)	>	e	(Ponapean)	0.16942	10.52294
(515)	a	(PonapeicTrukic)	>	e	(Ponapeic)	0.13099	19.78210
(516)	t	(MicronesianProper)	>	d	(PonapeicTrukic)	0.04291	17.55398
(517)	e	(BimaSumba)	>	a	(Pondok)	0.13262	6.53158
(518)	β	(TukangbesiBonerate)	>	ɸ	(Popalia)	0.00989	10.98695
(519)	e	(CentralEastern)	>	ə	(ProtoCentr)	0.00248	3.97165
(520)	o	(PonapeicTrukic)	>	a	(ProtoChuuk)	0.09341	2.83945
(521)	ŋ	(ProtoAustr)	>	n	(ProtoMalay)	0.13485	12.99231
(522)	v	(RemoteOceanic)	>	f	(ProtoMicro)	0.04980	13.02087
(523)	a	(EasternMalayoPolynesian)	>	o	(ProtoOcean)	0.09981	7.99270
(524)	f	(SamoicOutlier)	>	w	(Pukapuka)	0.00204	8.90305
(525)	a	(NorthBomberai)	>	e	(PulauArgun)	0.13318	12.93777
(526)	u	(ProtoChuuk)	>	ʊ	(PuloAnna)	8.595e-06	28.37331
(527)	d	(ProtoChuuk)	>	t	(PuloAnnan)	0.10821	26.92081
(528)	d	(ProtoChuuk)	>	t	(Puluwatese)	0.10821	32.33646
(529)	u	(KayanMurik)	>	o	(PunanKelai)	0.00695	11.47865
(530)	b	(Formosan)	>	v	(Puyuma)	0.02080	5.97573
(531)	s	(EastVanuatu)	>	h	(Raga)	0.03550	19.03806
(532)	r	(CenderawasihBay)	>	l	(RajaAmpat)	0.10632	10.89494
(533)	f	(East Nuclear Polynesian)	>	h	(RapanuiEas)	0.05254	6.51440
(534)	a	(Tahitic)	>	e	(Rarotongan)	0.23947	2.99725
(535)	a	(ProtoMalay)	>	ə	(RejangReja)	0.00259	33.94479
(536)	i	(CentralEasternOceanic)	>	a	(RemoteOceanic)	0.30671	1.87121
(537)	r	(Futunic)	>	ɡ	(Rennellese)	0.01560	46.78448
(538)	ɔ	(Choiseul)	>	o	(Ririo)	0.01953	8.10758
(539)	r	(NorthNewGuinea)	>	z	(Riwo)	0.01094	1.23968
(540)	r	(KisarRoma)	>	ɾ	(Roma)	0.00324	9.86839
(541)	k	(Nuclear WestCentralPapuan)	>	h	(Roro)	0.04267	7.79458
(542)	f	(West NuclearTimor)	>	b	(RotiTerman)	0.05510	4.89359
(543)	t	(WestFijianRotuman)	>	f	(Rotuman)	0.07237	18.85436
(544)	ɔ	(West NewGeorgia)	>	o	(Roviana)	0.00573	6.86288
(545)	u	(Formosan)	>	o	(Rukai)	2.254e-04	9.26163
(546)	k	(Tahitic)	>	ʔ	(Rurutuan)	0.00737	41.17228
(547)	u	(Palawano)	>	o	(SWPalawano)	7.014e-04	10.93716
(548)	f	(Southern Malaita)	>	h	(Saa)	0.00833	23.36029
(549)	*Not enough data available for reliable sound change estimates*						
(550)	ɔ	(Southern Malaita)	>	o	(SaaSaaVill)	0.01151	2.00000
(551)	ɔ	(Southern Malaita)	>	o	(SaaUkiNiMa)	0.01151	1.99961
(552)	*Not enough data available for reliable sound change estimates*						
(553)	u	(Tsouic)	>	o	(Saaroa)	0.00593	29.89765
(554)	a	(Northwest)	>	o	(Sabahan)	0.01254	7.65235
(555)	d	(ProtoChuuk)	>	t	(SaipanCaro)	0.10821	30.24687
(556)	u	(Formosan)	>	o	(Saisiat)	2.254e-04	32.49754

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(557)	a	(Vanuatu)	>	ɛ	(SakaoPortO)	1.415e-05	7.37912
(558)	v	(Suauic)	>	h	(Saliba)	0.01548	5.59439
(559)	e	(ProtoMalay)	>	a	(SamaBajaw)	0.04499	11.37100
(560)	a	(SuluBorneo)	>	ɨ	(SamalSiasi)	0.03121	4.55207
(561)	u	(CentralLuzon)	>	o	(SambalBoto)	0.03570	51.44942
(562)	k	(SamoicOutlier)	>	ʔ	(Samoan)	7.913e-05	40.57180
(563)	r	(Nuclear Polynesian)	>	l	(SamoicOutlier)	0.04761	2.52276
(564)	*Not enough data available for reliable sound change estimates*						
(565)	h	(Northern Sangiric)	>	r	(SangilSara)	0.01859	10.57112
(566)	q	(Northern Sangiric)	>	ʔ	(Sangir)	0.17663	33.66747
(567)	ʔ	(Northern Sangiric)	>	q	(SangirTabu)	0.17663	5.11841
(568)	a	(Sulawesi)	>	e	(Sangiric)	0.06360	9.41251
(569)	l	(SanCristobal)	>	r	(SantaAna)	0.20803	20.89020
(570)	ɔ	(SanCristobal)	>	o	(SantaCatal)	0.00562	1.99971
(571)	s	(SouthNewIrelandNorthwestSolomonic)	>	h	(SantaIsabel)	0.01828	6.35726
(572)	l	(NehanNorthBougainville)	>	n	(SaposaTinputz)	0.13005	8.23827
(573)	a	(Blaan)	>	u	(SaranganiB)	0.10239	1.65720
(574)	o	(SarmiJayapuraBay)	>	a	(Sarmi)	0.30507	3.74591
(575)	l	(NorthNewGuinea)	>	r	(SarmiJayapuraBay)	0.09033	9.81639
(576)	a	(BaliSasak)	>	ə	(Sasak)	0.07763	7.22450
(577)	a	(ProtoChuuk)	>	e	(Satawalese)	0.11108	19.48892
(578)	a	(BimaSumba)	>	e	(Savu)	0.13262	10.66529
(579)	n	(NorthNewGuinea)	>	ŋ	(Schouten)	0.12461	3.55253
(580)	a	(Atayalic)	>	u	(Sediq)	0.15623	4.94540
(581)	f	(Western AdmiraltyIslands)	>	h	(Seimat)	0.03838	5.52135
(582)	u	(NorthBomberai)	>	i	(Sekar)	0.07251	14.57543
(583)	b	(SoutheastMaluku)	>	h	(Selaru)	0.00154	5.27609
(584)	*Not enough data available for reliable sound change estimates*						
(585)	ɛ	(Pasismanua)	>	e	(Sengseng)	0.00635	6.51904
(586)	r	(East CentralMaluku)	>	l	(Seram)	0.17834	9.26050
(587)	l	(Nunusaku)	>	r	(SeramStraits)	0.08036	17.65607
(588)	e	(MaselaSouthBabar)	>	ɛ	(Serili)	0.03391	27.97054
(589)	f	(NilaSerua)	>	w	(Serua)	0.09929	7.71214
(590)	ŋ	(PatpatarTolai)	>	n	(Siar)	0.11079	7.91352
(591)	n	(FloresLembata)	>	ŋ	(Sika)	0.12345	10.33270
(592)	f	(Ellicean)	>	h	(Sikaiana)	0.01500	18.73830
(593)	ɔ	(West NewGeorgia)	>	o	(Simbo)	0.00573	3.68466
(594)	a	(CentralPapuan)	>	o	(SinagoroKeapara)	0.25340	1.00322
(595)	a	(LandDayak)	>	o	(Singhi)	0.01654	17.32791
(596)	u	(EastFormosan)	>	o	(Siraya)	4.190e-04	8.28489
(597)	ɔ	(Choiseul)	>	o	(Sisingga)	0.01953	13.70360
(598)	*Not enough data available for reliable sound change estimates*						
(599)	r	(BimaSumba)	>	z	(Soa)	0.00401	6.96601
(600)	a	(Sarmi)	>	e	(Sobei)	0.23073	10.06460
(601)	k	(CentralMaluku)	>	ʔ	(Soboyo)	0.00549	7.46123
(602)	a	(NehanNorthBougainville)	>	e	(Solos)	0.10574	4.14147
(603)	*Not enough data available for reliable sound change estimates*						
(604)	l	(Ambon)	>	r	(SouAmanaTe)	0.03275	3.58352
(605)	r	(NorthernLuzon)	>	l	(SouthCentralCordilleran)	0.03231	7.74272
(606)	n	(MaselaSouthBabar)	>	l	(SouthEastB)	0.25859	6.79647
(607)	a	(CentralVanuatu)	>	e	(SouthEfate)	0.06242	9.08441
(608)	v	(SouthHalmaheraWestNewGuinea)	>	p	(SouthHalmahera)	0.04413	6.86021
(609)	u	(EasternMalayoPolynesian)	>	i	(SouthHalmaheraWestNewGuinea)	0.15907	2.91215
(610)	i	(NewIreland)	>	u	(SouthNewIrelandNorthwestSolomonic)	0.17058	3.74428
(611)	u	(Sulawesi)	>	o	(SouthSulawesi)	0.04575	8.90368
(612)	t	(Babar)	>	k	(South Babar)	0.08233	5.96986
(613)	l	(Bisayan)	>	y	(South Bisayan)	0.06356	5.93967
(614)	a	(Manobo)	>	ɨ	(South Manobo)	0.09378	14.76928
(615)	y	(West Barito)	>	i	(South West Barito)	0.00602	4.16231
(616)	o	(Eastern AdmiraltyIslands)	>	u	(SoutheastIslands)	0.01939	2.52762
(617)	ɾ	(ProtoCentr)	>	r	(SoutheastMaluku)	0.02692	16.51716
(618)	r	(CentralEasternOceanic)	>	l	(SoutheastSolomonic)	0.18698	6.93482
(619)	t	(SouthCentralCordilleran)	>	b	(SouthernCordilleran)	0.17112	1.97141
(620)	e	(ProtoMalay)	>	ɨ	(SouthernPhilippine)	0.00192	17.69275
(621)	l	(Malaita)	>	n	(Southern Malaita)	0.09412	3.02851
(622)	o	(South Babar)	>	u	(SouthwestBabar)	0.05720	5.18676
(623)	b	(Timor)	>	w	(SouthwestMaluku)	0.05085	6.39151
(624)	i	(Vitiaz)	>	u	(SouthwestNewBritain)	0.23570	3.84871
(625)	u	(Atayalic)	>	o	(SquliqAtay)	0.00681	13.92923
(626)	k	(Suauic)	>	ʔ	(Suau)	0.02509	20.06806
(627)	r	(Nuclear PapuanTip)	>	l	(Suauic)	0.04206	7.42557
(628)	ɨ	(Subanun)	>	o	(SubanonSio)	0.03138	42.58704
(629)	a	(SouthernPhilippine)	>	ɨ	(Subanun)	0.05979	13.92604
(630)	ɡ	(Subanun)	>	d	(SubanunSin)	0.15626	6.97087
(631)	e	(ProtoMalay)	>	a	(Sulawesi)	0.04499	11.63780
(632)	u	(SamaBajaw)	>	o	(SuluBorneo)	0.02973	3.89781
(633)	e	(ProtoMalay)	>	o	(Sumatra)	0.00512	23.81901
(634)	ŋ	(ProtoMalay)	>	n	(Sunda)	0.10751	21.82474
(635)	u	(South Bisayan)	>	o	(Surigaonon)	0.03947	24.45484
(636)	a	(Erromanga)	>	e	(SyeErroman)	0.11628	12.15031
(637)	t	(SouthSulawesi)	>	ʔ	(TaeSToraja)	0.06897	3.10286
(638)	ŋ	(Tboli)	>	n	(Tagabili)	0.10214	23.38929

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(639)	u	(CentralPhilippine)	>	o	(TagalogAnt)	0.01883	10.16369
(640)	k	(Kalamian)	>	q	(TagbanwaAb)	0.42879	5.41452
(641)	q	(Kalamian)	>	k	(TagbanwaKa)	0.42879	17.73591
(642)	u	(Tahitic)	>	o	(Tahiti)	0.19238	6.98869
(643)	e	(Tahitic)	>	a	(TahitianMo)	0.23947	3.68949
(644)	a	(Tahitic)	>	e	(Tahitianth)	0.23947	2.71372
(645)	f	(Central East Nuclear Polynesian)	>	h	(Tahitic)	0.05235	6.11830
(646)	v	(SaposaTinputz)	>	f	(Taiof)	0.04497	6.79178
(647)	l	(Ellicean)	>	r	(Takuu)	0.01827	30.88407
(648)	*Not enough data available for reliable sound change estimates*						
(649)	*Not enough data available for reliable sound change estimates*						
(650)	*Not enough data available for reliable sound change estimates*						
(651)	*Not enough data available for reliable sound change estimates*						
(652)	h	(Guadalcanal)	>	ɣ	(TalisePole)	5.670e-05	1.97576
(653)	f	(Wetar)	>	h	(Talur)	0.03346	6.74401
(654)	e	(Vanikoro)	>	a	(Tanema)	0.28248	10.81075
(655)	a	(PatpatarTolai)	>	e	(Tanga)	0.10990	12.52910
(656)	e	(Utupua)	>	u	(Tanimbili)	0.10880	6.71919
(657)	a	(Vanuatu)	>	ə	(Tanna)	0.01056	13.80984
(658)	r	(Tanna)	>	l	(TannaSouth)	0.05929	26.23479
(659)	a	(Vanuatu)	>	e	(Tape)	0.08559	36.62961
(660)	*Not enough data available for reliable sound change estimates*						
(661)	f	(Sarmi)	>	p	(Tarpia)	0.09285	4.76042
(662)	o	(ButuanTausug)	>	u	(TausugJolo)	0.03983	38.80244
(663)	*Not enough data available for reliable sound change estimates*						
(664)	a	(Bilic)	>	ɔ	(Tboli)	0.01846	18.79298
(665)	ɔ	(Tboli)	>	o	(TboliTagab)	0.02446	2.78683
(666)	ɔ	(Vanikoro)	>	o	(Teanu)	0.13019	24.94671
(667)	l	(SouthwestBabar)	>	n	(TelaMasbua)	0.17076	14.35748
(668)	s	(SaposaTinputz)	>	h	(Teop)	0.06456	11.64051
(669)	ŋ	(East NuclearTimor)	>	n	(TetunTerik)	0.02632	3.88184
(670)	t	(TeunNilaSerua)	>	ʔ	(Teun)	0.01477	8.79107
(671)	e	(SouthwestMaluku)	>	ɛ	(TeunNilaSerua)	0.00128	6.36734
(672)	b	(WesternPlains)	>	f	(Thao)	0.01723	9.66201
(673)	u	(LavongaiNalik)	>	a	(Tiang)	0.23417	7.41175
(674)	a	(LavongaiNalik)	>	o	(Tigak)	0.10194	2.85222
(675)	r	(Futunic)	>	l	(Tikopia)	0.05835	3.75175
(676)	ɾ	(ProtoCentr)	>	r	(Timor)	0.02692	17.70656
(677)	e	(Dayic)	>	o	(TimugonMur)	0.00467	20.83722
(678)	s	(Northern Malaita)	>	ɵ	(Toambaita)	0.01236	5.00367
(679)	k	(Sumatra)	>	h	(TobaBatak)	0.01209	10.18636
(680)	s	(SamoicOutlier)	>	h	(Tokelau)	0.02620	9.71242
(681)	ɡ	(Guadalcanal)	>	h	(Tolo)	0.04461	9.43634
(682)	a	(Tongic)	>	o	(Tongan)	0.28401	6.49634
(683)	s	(Polynesian)	>	h	(Tongic)	0.04938	24.85227
(684)	l	(North Minahasan)	>	d	(Tonsea)	0.05604	2.89546
(685)	b	(North Minahasan)	>	w	(Tontemboan)	0.12446	9.61430
(686)	n	(MonoUruava)	>	l	(Torau)	0.17242	4.80739
(687)	a	(Tsouic)	>	o	(Tsou)	0.00357	26.52983
(688)	d	(Formosan)	>	c	(Tsouic)	0.02777	7.43830
(689)	n	(Tahitic)	>	ŋ	(Tuamotu)	0.00257	17.01166
(690)	a	(Wetar)	>	i	(Tugun)	0.34504	3.34154
(691)	a	(MunaButon)	>	o	(TukangbesiBonerate)	0.19256	6.52913
(692)	a	(LavongaiNalik)	>	e	(TungagTung)	0.07781	8.41577
(693)	u	(Barito)	>	o	(Tunjung)	0.00125	6.26453
(694)	s	(Ellicean)	>	h	(Tuvalu)	0.01029	12.42513
(695)	p	(Are)	>	f	(Ubir)	0.00167	1.99937
(696)	*Not enough data available for reliable sound change estimates*						
(697)	p	(Aru)	>	f	(UjirNAru)	0.01141	10.59153
(698)	h	(Nunusaku)	>	b	(UlatInai)	0.07007	6.57134
(699)	o	(Erromanga)	>	e	(Ura)	0.05761	9.45115
(700)	l	(MonoUruava)	>	r	(Uruava)	0.18018	8.71608
(701)	a	(EasternOuterIslands)	>	o	(Utupua)	0.19429	6.30089
(702)	s	(SamoicOutlier)	>	h	(UveaEast)	0.02620	28.16233
(703)	r	(Futunic)	>	l	(UveaWest)	0.05835	27.69390
(704)	r	(Futunic)	>	l	(VaeakauTau)	0.05835	41.42212
(705)	u	(Choiseul)	>	ə	(Vaghua)	3.654e-05	19.13794
(706)	*Not enough data available for reliable sound change estimates*						
(707)	a	(EasternOuterIslands)	>	e	(Vanikoro)	0.33088	4.43031
(708)	o	(Vanikoro)	>	e	(Vano)	0.10677	4.13392
(709)	o	(RemoteOceanic)	>	u	(Vanuatu)	0.11711	14.34763
(710)	o	(Choiseul)	>	ɔ	(Varisi)	0.01953	6.43868
(711)	*Not enough data available for reliable sound change estimates*						
(712)	r	(SinagoroKeapara)	>	l	(Vilirupu)	0.17037	8.15153
(713)	a	(NgeroVitiaz)	>	e	(Vitiaz)	0.13267	4.47748
(714)	s	(MesoMelanesian)	>	d	(Vitu)	0.02619	6.27238
(715)	t	(HuonGulf)	>	r	(Wampar)	0.06174	5.81284
(716)	s	(BimaSumba)	>	h	(Wanukaka)	0.03743	14.17816
(717)	*Not enough data available for reliable sound change estimates*						
(718)	v	(CenderawasihBay)	>	w	(Waropen)	0.08865	4.74200
(719)	r	(GeserGorom)	>	l	(Watubela)	0.17521	13.68799
(720)	l	(AreTaupota)	>	r	(Wedau)	0.04316	3.56623

continued on next page

continued from previous page

Code		Parent			Child	Functional load	Number of occurrences
(721)	i	(BimaSumba)	>	e	(WejewaTana)	0.04706	11.31057
(722)	u	(Seram)	>	v	(Werinama)	3.739e-05	7.78342
(723)	t	(CentralPapuan)	>	k	(WestCentralPapuan)	0.13745	14.10036
(724)	a	(ProtoCentr)	>	o	(WestDamar)	0.02891	9.89554
(725)	f	(CentralPacific)	>	h	(WestFijianRotuman)	0.00402	4.95883
(726)	k	(NortheastVanuatuBanksIslands)	>	h	(WestSanto)	0.01801	2.59859
(727)	a	(Barito)	>	e	(West Barito)	0.06510	2.67923
(728)	a	(Central Manobo)	>	ɨ	(West Central Manobo)	0.12386	21.47034
(729)	a	(Manus)	>	e	(West Manus)	0.16506	7.10396
(730)	ɔ	(NewGeorgia)	>	o	(West NewGeorgia)	0.00631	2.83443
(731)	r	(NuclearTimor)	>	l	(West NuclearTimor)	0.14621	3.15074
(732)	*Not enough data available for reliable sound change estimates*						
(733)	ɨ	(West Central Manobo)	>	e	(WesternBuk)	4.647e-04	76.71920
(734)	s	(WestFijianRotuman)	>	c	(WesternFij)	0.00926	4.99871
(735)	*Not enough data available for reliable sound change estimates*						
(736)	a	(ProtoOcean)	>	e	(WesternOceanic)	0.12139	2.58873
(737)	d	(Formosan)	>	s	(WesternPlains)	0.04913	4.48311
(738)	s	(AdmiraltyIslands)	>	h	(Western AdmiraltyIslands)	0.01688	4.34207
(739)	a	(MunaButon)	>	o	(Western Munic)	0.19256	11.48592
(740)	t	(SouthwestMaluku)	>	k	(Wetar)	0.09657	3.25493
(741)	r	(MesoMelanesian)	>	l	(Willaumez)	0.14060	7.57978
(742)	b	(CentralWestern)	>	v	(WindesiWan)	0.16193	3.04357
(743)	ʔ	(Manam)	>	k	(Wogeo)	0.07162	8.86144
(744)	k	(ProtoChuuk)	>	g	(Woleai)	0.00179	35.44518
(745)	a	(ProtoChuuk)	>	e	(Woleaian)	0.11108	60.39379
(746)	a	(Sulawesi)	>	o	(Wolio)	0.06991	8.71005
(747)	w	(Western Munic)	>	v	(Wuna)	0.00508	2.73532
(748)	t	(Western AdmiraltyIslands)	>	ʔ	(Wuvulu)	0.00203	11.89389
(749)	a	(HuonGulf)	>	e	(Yabem)	0.13370	6.51629
(750)	a	(SamaBajaw)	>	e	(Yakan)	0.04299	27.73964
(751)	a	(KeiTanimbar)	>	e	(Yamdena)	0.18822	17.50706
(752)	o	(Bashiic)	>	u	(Yami)	0.00368	5.79971
(753)	a	(ProtoOcean)	>	i	(Yapese)	0.27352	4.64937
(754)	a	(Javanese)	>	e	(Yogya)	0.08208	4.09979
(755)	o	(West SantaIsabel)	>	ɔ	(ZabanaKia)	0.02351	13.69682

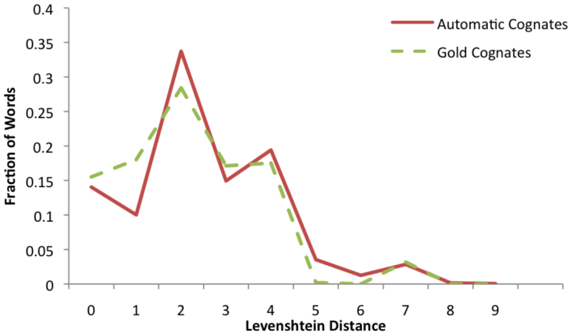

Figure S.1: Percentage of words with varying levels of Levenshtein distance. Known Cognates (gold) were hand-annotated by linguists, while Automatic Cognates were found by our system.

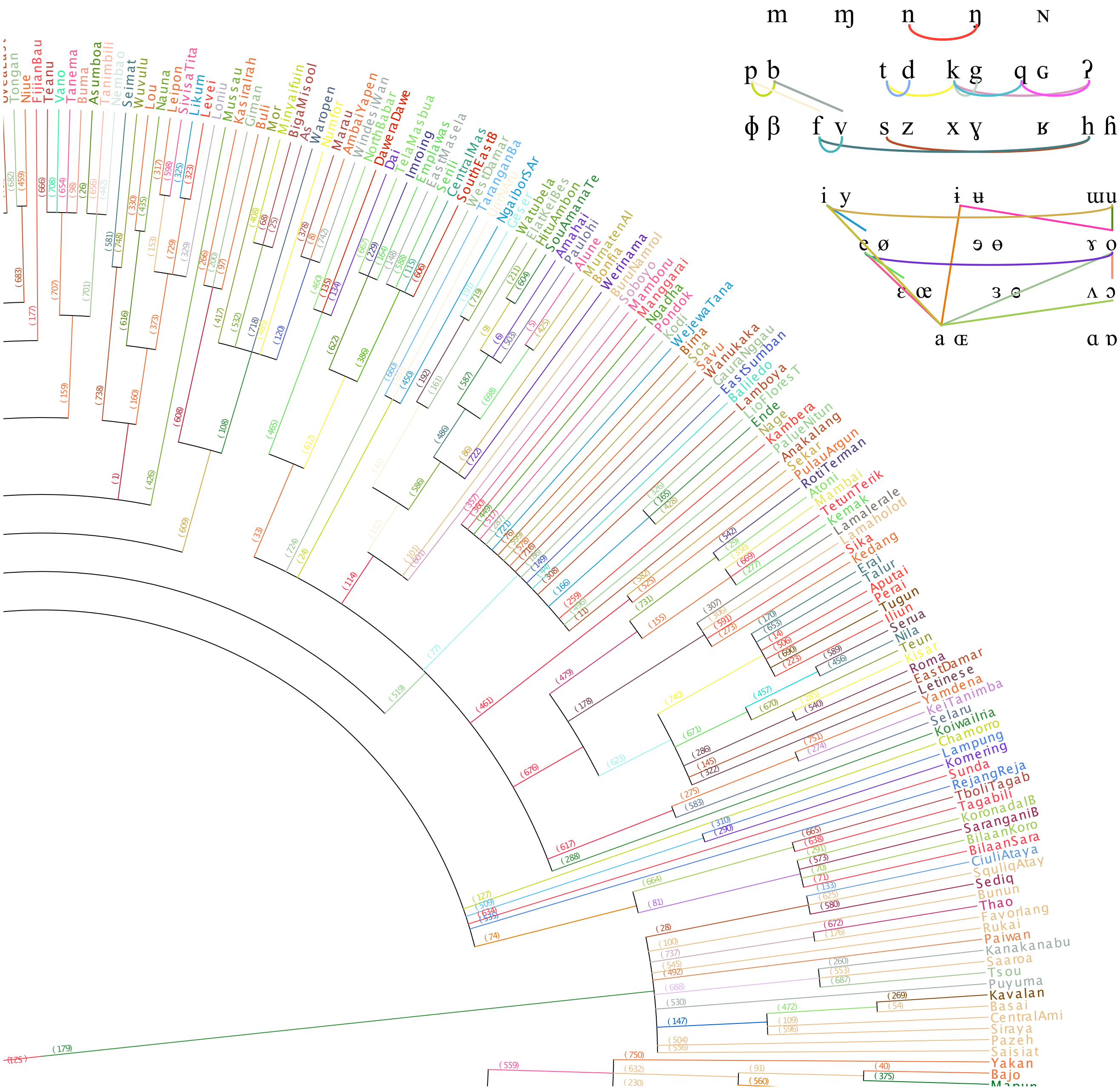

Figure S.2: Branch-specific, most frequent estimated changes. See Table S.5 for more information, cross-referenced with the code in parenthesis attached to each branch.

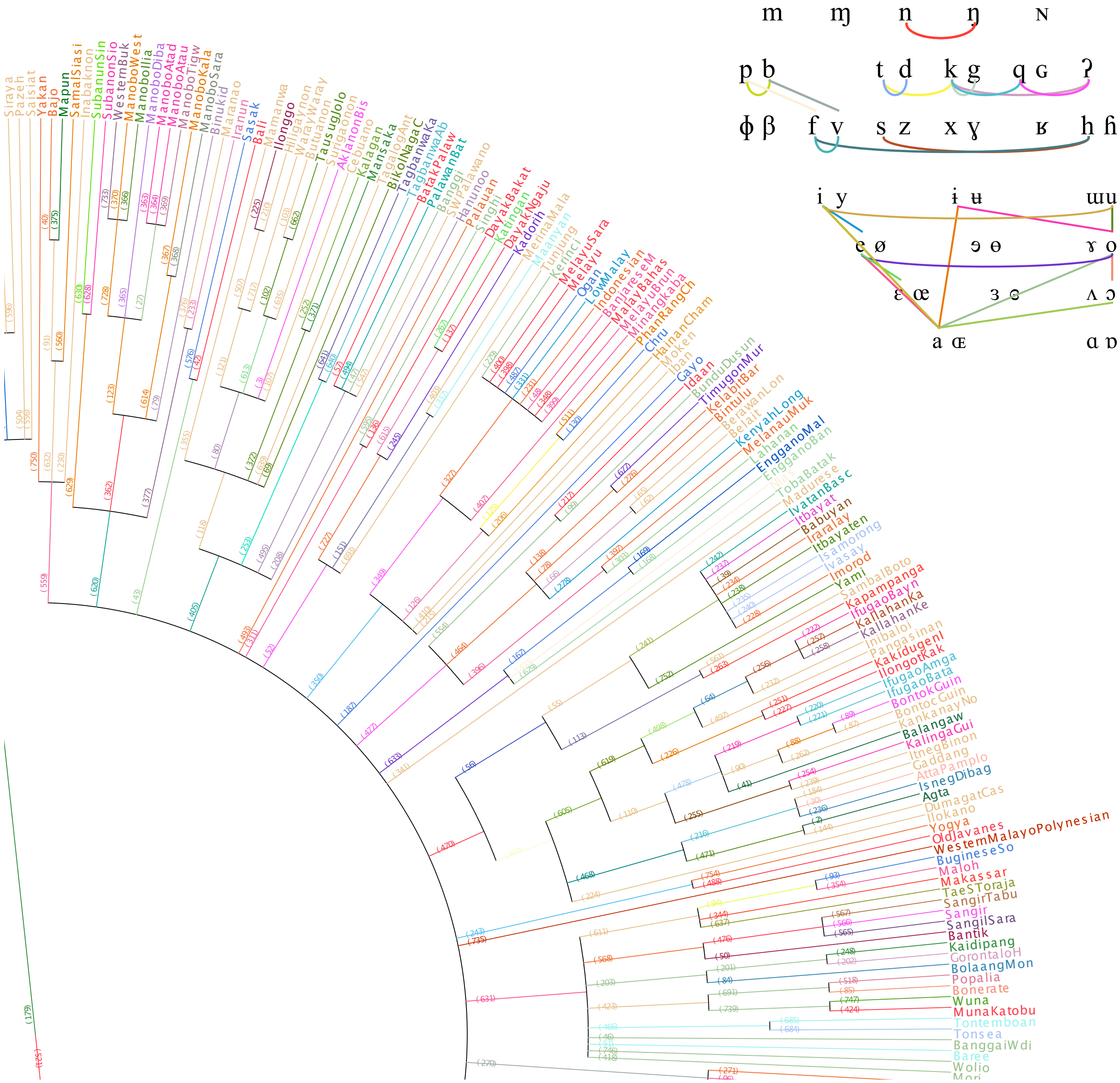

Figure S.3: Branch-specific, most frequent estimated changes. See Table S.5 for more information, cross-referenced with the code in parenthesis attached to each branch.

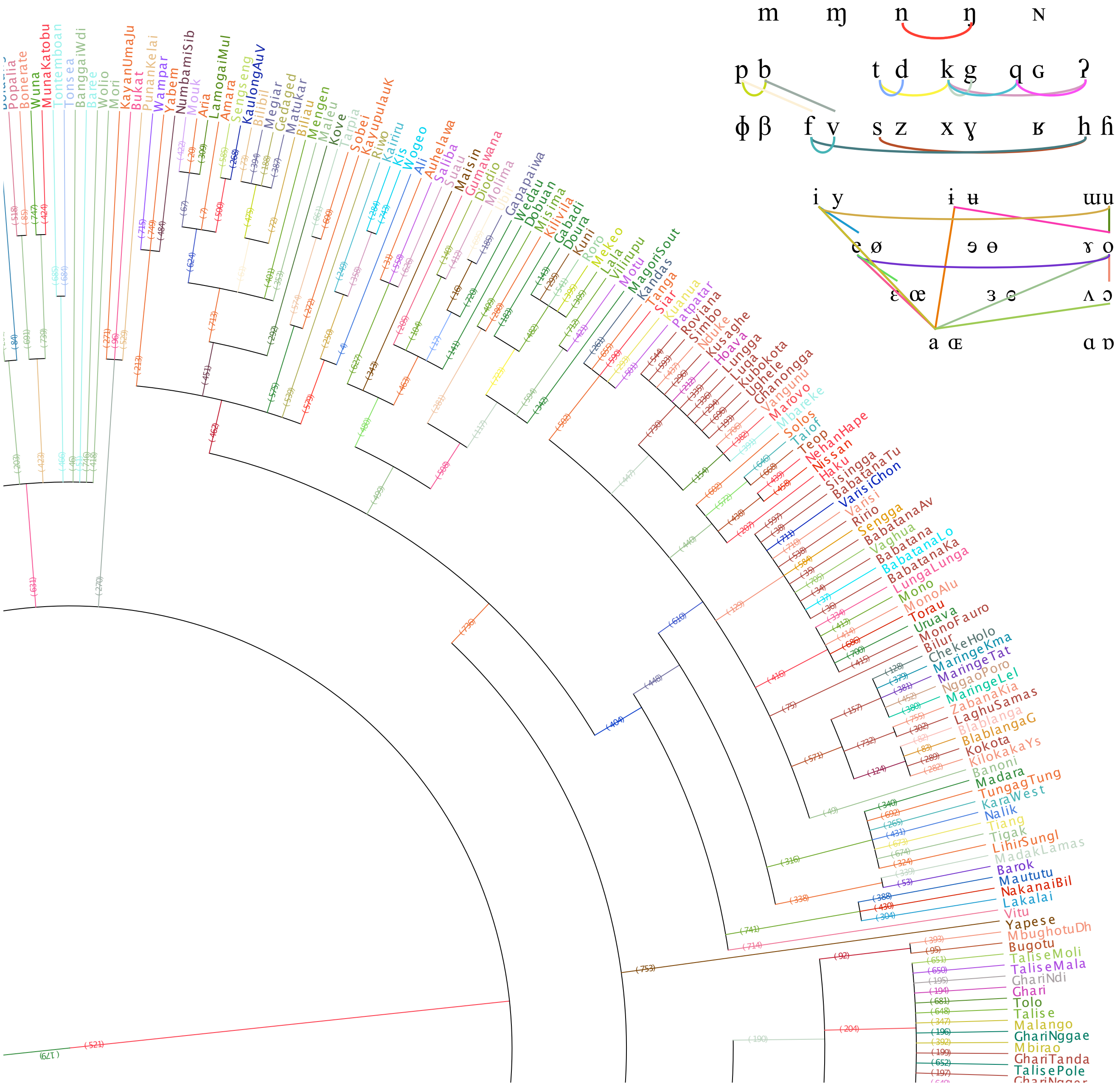

Figure S.4: Branch-specific, most frequent estimated changes. See Table S.5 for more information, cross-referenced with the code in parenthesis attached to each branch.

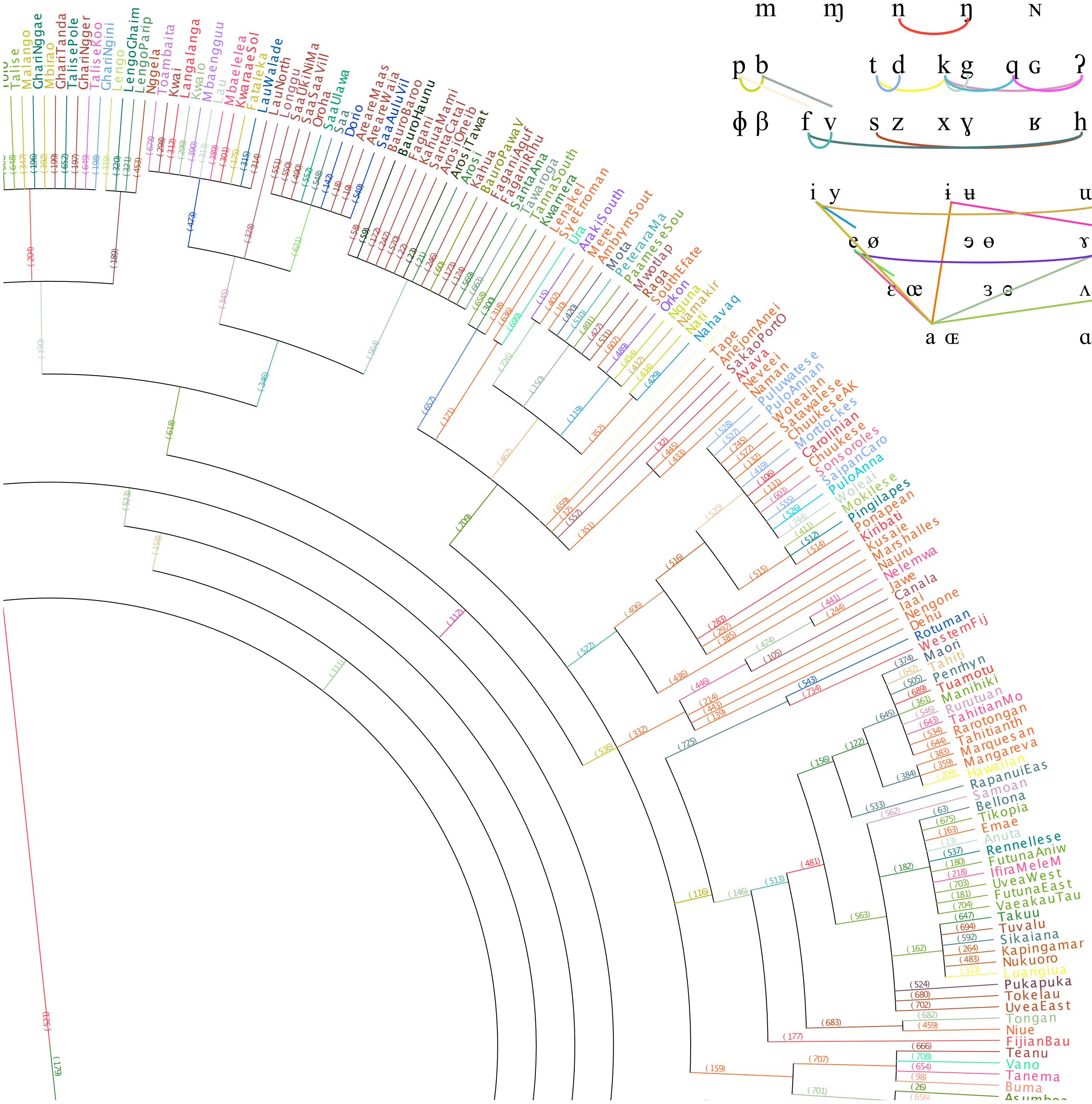

Figure S.5: Branch-specific, most frequent estimated changes. See Table S.5 for more information, cross-referenced with the code in parenthesis attached to each branch.

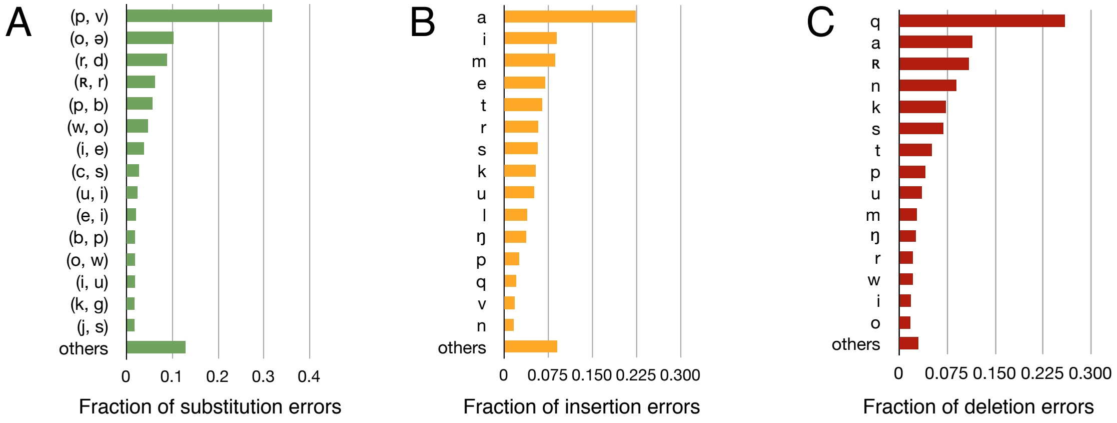

Figure S.6: Most common substitution errors, insertion errors, and deletion errors in the PAn reconstructions produced by our system. In (A), the first phoneme in each pair (x, y) represents the reference phoneme, followed by the incorrectly hypothesized one. In (B), each phoneme corresponds to a phoneme present in the automatic reconstruction but not in the reference, and vice-versa in (C).