Coretention of Words and Genes on the Island of Sumba, Eastern Indonesia

Abstract

Numerous studies have shown strong associations between languages and genes among human populations. These continental scale genetic and linguistic patterns must arise from processes originating in the communities where people live. We examine the patterns of linguistic and genetic variaion on the eastern Indonesian island of Sumba. Approximately 3500 years ago, this island was the cite of contact between aboriginal foraging societies and seafaring Austronesian farmers. We find the the proportion of words and of genes that can be traced to Austronesian origins varies across the island with a positive correlation between the percentage of Y chromosome lineages derived from Austronesian ancestors and retained cognates of the Proto-Austronesian language.

1 Introduction

Human populations and the languages that they speak change over time. The movement of people and the innovations they make in their language are difficult to observe and quantify over short periods and impossible to witness over long periods. Consequently, researchers have been forced to undertake indirect approaches to infer associations between human languages and human genes. Most of the well-known studies focus their questions on the movement of people on continental scales. These studies led Diamond and Bellwood to suggest that many of the correlations that we presently see in languages and genes result from the movement of prehistoric farmers from the places in which these agricultural techniques first arose. The simplest form of their hypothesis is that genes and languages evolve in parallel after migrating farmers encounter and replace indigenous hunter-gatherer populations. Examples of these source homelands can be found in Africa, the Near East and Europe, Asia and the Americas.

2 The History and the People of Sumba

Both genetic and archeological evidence place the first migration of anatomically modern humans (AMH) in Southeast Asia and Oceana between 40,000 and 45,000 years ago. Further archeological evidence place the transition on Sumba from the original hunter-gatherer technology to the neolithic technology between 4,000 and 3,500 years ago. At that time, a small number of farmers speaking an Austronesian language likely came into contact with resident foragers speaking presumably a Papuan language.

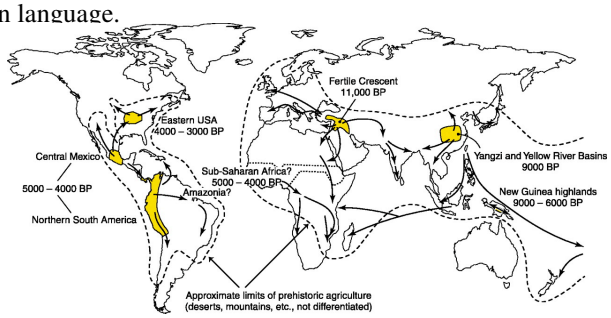

Figure 1: Archaeological map of agricultural homelands and spreads of Neolithic/Formative cultures, with approximate radiocarbon dates, from Diamond and Bellwood

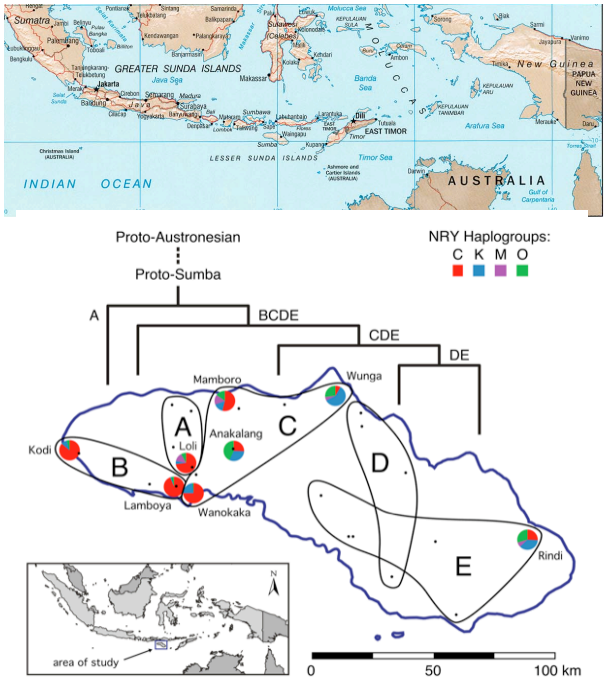

Figure 3: Phylogenetic and geographic distribution of languages and Y chromsomes

Sumba is a remote island approximately 220 km $\times$ 75 km in the Lesser Sunda Islands. This isolation explains some of the conservative cultural aspects of Sumban life. For example, the majority of residents of this island live in traditional farming villages, adhere to a patrilocal social structure, and practice ancestral tribal or pagan religions. Residents rarely have contact with anyone outside their own village and population density, particularly in the dry eastern half of the island, is low. One telling consequence of the infrequency of contact is that despite the smallness of the island the variety of languages spoken is quite large.

3 Language Variation

Our linguistic data consists consists of twenty-nine 200-word Swadesh lists from sites well distributed throughout the island. Swadesh lists are build from words ascribed to meanings that are basic to everyday life. Using the methodology of comparative linguistics, some words from different language lists can be traced to a common ancestral word. Such a pair of words is called cognates. This investigation also discovers sound correspondences, language innovations and loan words. These techniques have been used previously to construct a Proto-Austronesian (PAn) language. On average, 35% of the words from a given Swadesh lists can be determined as direct descendents of PAn. In addition, a phylogenetic tree built from these 29 word lists branches to form five major language subgroups and leads us to the conclusion that the present day Sunbanese all speak a language derived from a single common ancestral Proto-Sunbanese.

4 Genetic Variation

In order to investigate the paternal histories of the Sumbanese villages, we obtained genetic samples from 352 men inhabiting eight villages. Our genetic information is derived from the Y-chromosome.

The Y Chromosome Consortium (YCC) collaborates to study genetic variation on the human Y chromosome. Based on single nucleotide polymorphisms (SNP), the YCC built a worldwide genealogical tree of paternal ancestry. The major branches of the tree, the so-called Y chromosome lineages, are given labels using the letters from **A** to **R**. The tips of these branches are based on the polymorphism that define the Y chromosome groups or *haplogroups*.

Figure 4: Collecting language data in the village of Wunga

	r	P-value
genetics/geography	0.011	0.518
genetics/language	0.358	0.023
geography/language	0.673	< 0.001

Table I: Mantel test results for correlation r.

In our data, we found 17 haplogroups belongs to the **C**, **K**, **M**, and **O** lineages. The first three are associated with the first expansion of AMHs. Men belonging to these haplogroups are found nearly exclusively in eastern Indonesia, Papua New Guinea, and Melanesia. Haplogroup O appears to be associated with the expansion of Austronesian societies from southeast Asia to Indonesia and Oceana. Overall, only 16% of the Sumbanese Y chromosomes belong to this haplogroup. As can be seen in Figure 3, the proportion of O haplogroup individuals in our sample varies from community to community. Overall, the proportion decreases as one moves from east to west across Sumba.

5 Associations among Linguistic, Genetic, and Geographic Distances

We organize our data for the 8 genetically sample villages into three symmetric matrices. The first is the linguistic distance determined from the language tree. The second is the genetic distances between villages determined by the SNP data. The third is the actual geographic distance between villages. We perform a Mantel test, a statistical test of the correlation between two matrices, and conclude that a statistically significant positive correlation exists between linguistic and genetic distances ($r = 0.358$, P-value $= 0.023$). To verify that this correlation emerged during the time of the Austronesian expansion, we estimated the divergence time between Rindi and Kodi, the geographically most distant communities in out sample. The upper limit of 4,875 years for the 95% confidence interval for this divergence time is consistent with this population expansion.

In addition, we found strong correlations between linguistic and geographic distances, but no correlation between genetic and geographic distance. This latter result is not surprising. The original settles have resided on Sumba for so long that geography are no longer an indicator of haplogroup. In addition, because only one-sixth of the individuals in our sample are members of the O haplogroup, we have little statistical power to detect any genetic/geographic correlation due to the Austronesian expansion.

Focussing on the Austronesian component of genetic/linguistic variation, we do find using a parametric bootstrap approach a significant positive correlation ($r = 0.627$, P-value = 0.047) between the percentage of O haplogroup individuals and Proto-Austronesian cognates. This correlation, unprecedented at such small spatial scales, provides additional evidence for the coevolution of languages and genes on Sumba.

Settlement history has a major role in the evolution of language on Sumba. To investigate this, we computed the correlations between the percentage of retained PAn cognates and geographic distances assuming that a given village was the site of the source for Austronesian settlers. Of the 29 populations, only 3 showed strongly significant correlations, Wunga ($r = -0.503$, P-value = 0.006), Rambangaru ($r = -0.501$, P-value = 0.006) and Kanatanu ($r = -0.507$, P-value = 0.006). These three villages are located on Sumba's central northern coast. This result matches well with Sumban oral history which suggests an origin near the village of Wunga.

6 Models of Language Change

The variation of language and genetics on the island of Sumba is not easily explained by models of language evolution formulated on the basis of larger-scale patterns of language variation. For example, simple models of language replacement without gene flow (e.g., elite dominance) or complete replacement of genes and languages are not appropriate given the evidence for genetic admixture between Austronesian farmers and indigenous Papuan populations. Indeed, the prevalence of indigenous Y chromosomes on Sumba is consistent with a pattern of demic diffusion whereby the incremental spread of farmers from their point of entry on the island was accompanied by frequent intermarriage with resident hunter-gatherers. Moreover indigenous languages were unlikely to have been fully replaced during the inital expansion of Austronesian on Sumba as we observe a high proportion of words (65%) that cannot be traced to proto-Austronesian, and loan words shared between different language groups that may have been absorbed from a now extinct indigenous source. Evidence for the latter hypothesis comes from the presence of non-Austronesian words (in particular, culturally significant words such as husband, animal, dog, and sea) in groups A and B (which do not form a subgroup) and their absence in subgroups C, D, and E to the east. Given the phylogenetic relationships in Figure 3, this pattern is more easily explained by loans of these vocabulary items from a common non-Austronesian source rather than by losses of ancestral Austronesian words in ProtoSumban and later recovery in groups C, D, and E.

To account for these patterns of linguistic and genetic variation we propose an alternative model of language evolution appropriate for the spatial scale of Sumba (Figure 7). In this model, intermarriage between expanding farmers and resident hunter-gatherers leads to progressively lower frequencies of haplogroup O Y chromosomes at increasing distances from the source population. Climate and population density data suggest that eastern Sumba remained sparsely populated during this expansion, and that new agricultural communities were relatively isolated. This contrasts with wetter western Sumba where expanding farmers likely came into contact with a larger indigenous population speaking non-Austronesian languages. As new farming villages proliferated in the populous west, the proportion of settlers of Austronesian descent would decrease, while the opportunities for linguistic contact would increase. Over time, these community-level processes gave rise to differential rates of language divergence/lexical borrowing, and the association between languages and genes on Sumba.

This scenario suggests a novel mechanism for language change: rather than elite dominance – where a few individuals of an invading culture impose their language on a resident population – the extent of retention of PAn items is governed by the proportion of men in the population with Austronesian paternal ancestry. This co-dominant model

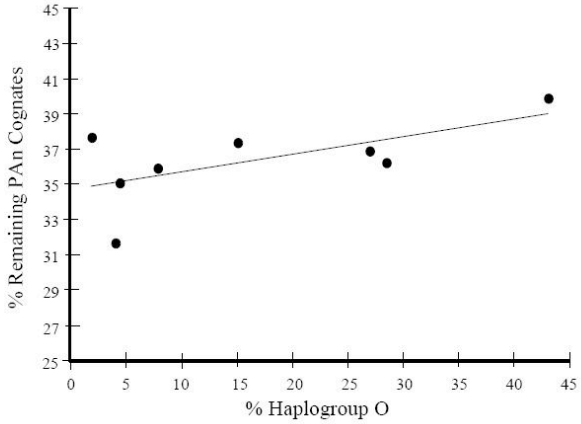

Figure 5: Scatterplot of PAN cognates versus the percentage of sample from haplogroup O

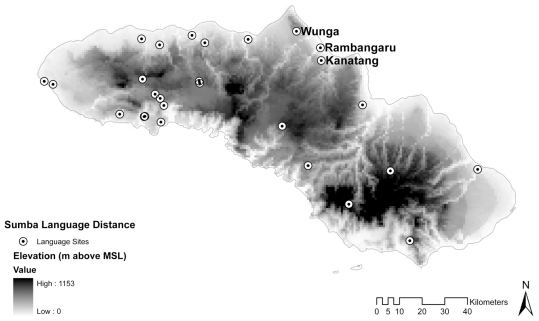

Figure 6: Geographic location of language lists with names of villages having strong language distance correlation to all other villages

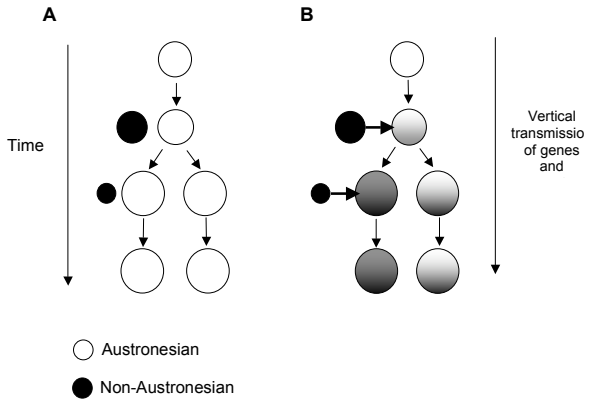

Figure 7: Geographic location of language lists with names of villages having strong language distance correlation to all other villages

Figure 8: A village in Sumba

also differs from the basic hypothesis of Diamond and Bellwood in that a linguistic-genetic association evolves despite ongoing processes of demic diffusion and language shift. Whether the processes integrated in this model can explain patterns observed at continental scales remains an open question; however, a link can be postulated because large-scale patterns are contingent on processes occurring at local scales. This may be particularly true in the many cases of languages and genes spread by the recent dispersal of farmers. More local scale studies in contact zones with variable degrees of interaction among groups speaking different languages (e.g., Bantu and Khoisan in southern Africa and Indo-Iranian and Tibeto-Burman in south Asia, etc.) would be particulary helpful for determining the generality of the model presented here. By incorporating lists of culturally appropriate words reflecting functional differences between farming and indigenous populations, future studies also may reveal more about the social dynamics favouring the retention of borrowed words. For now, the evidence provided here strongly suggests that language change can potentially vary in magnitude and character depending on factors inherent in the individual contact situation, and that genetic analysis is a powerful tool that can be used to help formulate hypotheses of incipient language speciation.

References

[1] Ammerman, A. J. and Cavalli-Sforza, L. L. (1984) *Neolithic Transition and the Genetics of Populations in Europe*, Princeton University Press, Princeton.

[2] Bahasa, P. (2002) *Kosakata Dasar Swadesh di Kabupaten Belu, Ngada, Sumba Barat, Sumba Timur, dan Timor Tengah Utara* Departemen Pendidikan Nasional, Rawamangun, Jakarta.

[3] Bellwood, P. (1997) *Prehistory of the Indo-Malaysian Archipelago* University of Hawaii Press, Honolulu.

[4] Diamond, J. and Bellwood, P. (2003) Farmers and their language: the first expansions, *Science* **300**, 597-603.

[5] Labov, W. (1994) *Principles of Linguistic Change: Internal Factors*, Blackwell Publishers, Oxford.

[6] Renfrew, C. (1987) *Archaeology and Language* Jonathan Cape, London.

[7] Tryon, D. T. ed. (1995) *The Austronesian Languages* Mouton de Gruyter, Berlin.

STATISTICAL PROCEDURES

All of the statistical procedures here are based on extensions of the ideas of correlation and covariance.

7 Mantel Test

The Mantel test computes a correlation between two positive symmetric $n \times n$ matrices. The ij-entry in matrix is meant to give a distance between to sites i and j. In this example, we have $n = 8$ villages and three measures of distance - geographic, linguistic, and genetic.

- geographic - distance in kilometers
- linguistic -
- genetic -

Distance matrices have a lot of constraints. For example, if we change one entry in the matrix, by altering distance from one village to the next, then we must also change lots of other entries in the matrix to compensate.

The null hypothesis is that the observed relationship between the two distance matrices could have been obtained by any random arrangement in space of the observations.

The distance matrices M and N are square and the calculations for the test are carried out on the entries above the diagonal. The computation yields a Z statistic:

$$Z = \sum_{1=1}^{n} \sum_{j=i+1}^{n} M_{ij} N_{ij}.$$

To see if this is bigger than what we might find by chance, we perform a permutation on one of the matrices. For for, example, if we were to keep the labels on the villages geographic distances in M and rearrange the distances in the N matrix with a permutation π, then we have a new Z-statistic:

$$Z^{\pi} = \sum_{1=1}^{n} \sum_{j=i+1}^{n} M_{ij} N_{\pi(i)\pi(j)}.$$

If the null hypothesis holds, then Z^{π} is equally likely to be bigger or smaller than Z. If a correlation between geographic distances does exist, then Z is likely to be bigger than Z^{π}.

The significance of the test statistic can be assessed by one of two methods. For small matrices, we can compute Z^{π} for all $n!$ permutations π. For larger matrices, we will randomly choose many permutations π. With either method, we rank their values. The P-value of the test is fraction of the Z^{π} that are greater than Z.

8 Parametric Bootstrap

The bootstrap analysis is based on a null assumption of no correlation between the fraction of O-haplotype individuals and retained Swadesh list proto-Austronesian cognates among the 8 villages. In our sample, the observed correlation was $r = 0.627$.

The alternative is that these two quantities are positively correlated. The standard test is based on a sufficiently large samples of people and words so that normal approximations are valid. With the small sample of Austronesian individuals in some villages, a bootstrap analysis is a more reliable procedure.

In addition, the following parametric bootstrap procedure also takes into account the size of the genetic sample in each village and the size of the sample of neutral words as exemplified by the Swadesh word list.

For village i, let

- S_i be the number of individual in the genetic sample

- p_i be the fraction that are typed as O haplogroup.
- W_i be the number of words in the Swadesh word list, W_i = 200 words
- q_i be the fraction of words that are PAn cognates

In addition, let p be the fraction of individuals throughout all of Sumba typed as O haplogroup and let q be the fraction of words throughout all of Sumba that are PAn cognates. Under the null hypothesis, each of the villages has a fraction p belonging to haplogroup O and a fraction q of neutrally evolving words that are PAn cognates. Consequently, for village i, draw from a binomial distribution $Bin(S_i, p)$ and divide by S_i. This gives a bootstrap sample for the fraction of individuals of haplotype O. In addition, for village i, draw from a binomial distribution $Bin(200, q)$ and divide by 200. This gives a bootstrap sample for the fraction of PAn cognates.

Compute the correlation using the 8 bootstrapped pairs.

$$\text{corr}(S, W) = \frac{1}{n-1} \sum_{i=1}^{n} \frac{S_i - \bar{S}}{s_S} \frac{W_i - \bar{W}}{s_W}.$$

Pepeat this procedure many times. The bootstrapped P-value for the test of no correlation is simply the fraction of bootstrap correlations are greater than the observed value of 0.627.

9 Correlation

For each village, we now look for the correlation in the genetic distance and the linguistic distance. Let

- D_{ij} be the geographic distance from village i to village j
- L_{ij} be the linguistic distance from village i to village j

Thus, we compute

$$\text{corr}(D_{\cdot j}, L_{\cdot j}) = \frac{1}{n-2} \sum_{i \neq j} \frac{D_{i,j} - \bar{D}_{\cdot j}}{s_{D_{\cdot j}}} \frac{L_{i,j} - \bar{L}_{\cdot j}}{s_{L_{\cdot j}}}.$$

Here, we expect the correlations from a source village to be negative - the farther one village is from the source village, the more dissimilar are their languages.

In order to avoid an excess of false positives, we will need to employ the *Bonferroni correction*. This is a multiple-comparison correction used when several statistical tests are being performed simultaneously. In this case,

$$\begin{aligned} &P\{\text{some test rejects the null hypothesis} | \text{the null hypothesis holds}\} \\ &\leq \sum_{j=1}^{n} P\{\text{test } j \text{ rejects the null hypothesis} | \text{the null hypothesis holds}\} \end{aligned}$$

The Bonferroni correction states that if we set the level of each of the tests at α/n, then the overall level of the test is at most α. In this case, we set $\alpha = 0.05$ and $n = 29$.