GENES, LANGUAGE, AND ETHNIC GROUPS: RECONSTRUCTING ORANG ASLI PREHISTORY

Alan G. Fix

Department of Anthropology, University of California, Riverside, CA 92521, USA

ABSTRACT

Darwin's phrase, "descent with modification," identifies the two core historical processes in evolution: characteristics are transmitted from generation to generation but not faithfully. The natural selection of genes is a major "modifier," but the frequency of traits may also change due to mutation, genetic drift, and gene flow. Analogous but not identical processes also shape the histories of languages and cultures. Current distributions of genetic traits, then, may reflect the interaction of both processes, descent and modification. Although called "markers," alleles do not necessarily label populations through time. Since multiple mechanisms cause allele frequency distributions, as well as linguistic and cultural traits, several hypotheses must always be considered to explain any observed bio-cultural-linguistic pattern. Recent genetic studies of global human variation have emphasized "descent," constructing tree diagrams purporting to show the history of human populations. These models are based on a branching process of genetic divergence, with population splits followed by isolation and gene drift providing the evolutionary dynamic. On the other hand, numerous studies of the population structures of small-scale human groups have found that networks of mobile individuals (foragers) or interlinked village clusters (farmers) are common. These studies imply that reticulate, anastomosing population histories provide a better model for human genetic history than do phylogenetic tree diagrams. Recent reconstructions of the linguistic and cultural prehistory of the Malaysian Peninsula along with contemporary surveys of biological variation can be used to assess a branching evolutionary model of fissioning and population replacement versus a mosaic model of long-term population interaction and gene flow.

INTRODUCTION

Given the current interest in reconstructing the history of human populations from genetic "markers," and especially the use of such genetic reconstructions by prehistorians (eg, Bellwood 1996a), I thought it would be useful in this paper to make clear some basic assumptions of historical genetics, as well as some pitfalls when these assumptions are not met. Good consumers of the products of another discipline need to be aware of its potential shortcomings.

Initially, I will provide some examples of how the history of genes may not tell us much about the history of populations. Then I will briefly summarize recent reconstructions of the population prehistory of the Malaysian Orang Asli based on linguistics and socio-cultural data and attempt to compare this inferred history with that suggested by the genetic data.

Evolution is an historical process. Darwin described it as "descent with modification," descent being the inheritance of characteristics from generation to generation, and modification being the change in the frequency of characteristics produced by evolutionary forces (primarily natural selection for Darwin but now considered to include mutation, genetic drift, and gene flow). The canonical representation of evolution is a tree diagram, or dendrogram, the branches of which are species or lineages diverging from a common ancestor. Populations of the ancestral species become geographically isolated, accumulate genetic differences, and ultimately become reproductively isolated as separate evolutionary units. Once such a split occurs, differences continue to accumulate over time without the possibility of genetic exchange between units to reduce them. The history of speciation, then, is a history of divergence through time.

Genes, like species, are discrete non-recombining units and their evolution can be described by a branching process. All present-day alleles (alternative forms of an ancestral gene at a locus) can in theory be traced back to a common ancestral condition. Each allele arose by mutation from a preceding ancestral state. Most such new variants are lost by chance, but some increase in frequency in populations and persist for long periods. New mutations may continue to occur, leading to new branches of the genic tree of descent.

Both species and genes share the fundamental property of discrete individuality; they cannot merge with other such units to form hybrids or mixtures. Their histories are correctly shown by tree diagrams with non-recombining branches. In contrast, infra-species biological taxa are (by definition) capable of interbreeding; that is, they are *not* closed reproductive communities and may fuse with other such groups repeatedly. It is possible for such divisions of a species to form by fission and diverge biologically in isolation, but the possibility for reamalgamation always exists. More importantly for historical study, intermixture may have occurred in the past so that simple divergence no longer accounts for genetic differences. Thus the basic assumption of dendrogram analysis that evolution occurs through successive splits is only a *hypothesis* when populations rather than species or genes are the taxonomic units.

Further complication arises from the fact that genetic differentiation or convergence among sub-specific populations may come about through several processes other than simply temporal accumulation. Ignoring this fact, Cavalli-Sforza and colleagues (1990) argue that history is a sequence of fissions followed by differentiation at a rate roughly proportional to time. As they acknowledge, later episodes of gene flow may "blur" the picture but "not obscure it entirely" (1990:18). However, they simply disregard the several other processes that may affect allele frequencies. Most importantly, Darwin's discovery, natural selection, may have profound consequences, causing allele frequency changes in *either* direction: toward greater similarity between population or greater divergence, depending on the environments.

THE HISTORY OF GENES

The human genome is a palimpsest of histories. As is often pointed out, we share some 98% of our functional genes with chimpanzees. Thus human hemoglobin A, the most common allele, is identical to chimpanzee hemoglobin along the entire length of the molecule, and the corresponding gene is also identical base-by-base for some 438 DNA bases. Cytochrome c, a molecule of great physiological importance, is nearly identical between humans and blue-green algae, testifying to the critical importance of its structure for its function. Clearly these genes tell us nothing about recent human history but rather are the conserved products of adaptive evolution.

A further example of the retention of functional molecules across species boundaries comes from the major histocompatibility complex (MHC). The MHC is crucial in immunological resistance and variants are found in most vertebrates. Since each new allele is "descended" from a previous allele by mutation or recombination, it is possible to reconstruct the genealogy of current alleles and present this information in a tree diagram.

In humans, the MHC is labeled the "Human Leucocyte Antigen" (HLA) system. A tree diagram representing HLA genes of humans and the corresponding genes in chimpanzees (eg, Parham *et al.* 1995, Figure 1) shows that several human HLA-A alleles cluster with chimpanzee alleles; that is, some human HLA-A alleles are more similar to those found in chimpanzees than they are to other human alleles. Such similarity does not mean that some humans or human populations are historically "closer" to chimpanzee populations. Rather, this situation demonstrates the functional retention of variability across species boundaries, a trans-specific polymorphism. The original mutations responsible for the present-day variation must have occurred prior to the split between ancestral chimpanzees and humans and has been maintained by natural selection. Again, the point is that the history of genetic loci is not equivalent to the history of populations and may tell us nothing useful about recent human history.

Similar arguments apply to all genetic reconstructions. Thus the recent massive compendium of Cavalli-Sforza and colleagues (1994), *The History and Geography of Human Genes*, is aptly titled since it does not necessarily tell us about the history and geography of human populations. Even currently popular studies of mitochondrial (mt) and nuclear DNA do not offer a panacea for historical reconstruction. Some of these problems will be examined below in the trees of Southeast Asian mtDNA. Before evaluating the genetic evidence bearing on Malaysian prehistory, however, the major hypotheses to explain present day diversity on the Peninsula will be set out.

MALAYSIAN PALEOSOCIOLOGY

The indigenous peoples of the Malaysian Peninsula (Orang Asli) are traditionally divided into three main groups: Semang (or "Negritos"), Senoi, and Melayu Asli ("proto"- or "aboriginal" Malay). Table 1 provides a capsule summary of the major differences (excluding the biological) among these groups.

The traditional explanation for the presence of three distinct cultural patterns coexisting on the Peninsula was migrational. That is, each group originated elsewhere and

Table 1: Malaysian Orang Asli "traditions".[1]

Tradition	Language	Technology/economy	Societal Pattern
1. Semang ("Negritos"}	Northern Aslian	nomadic foragers	exogamy, mobile conjugal families, extensive networks
2. Senoi	Central Aslian	sedentary swiddeners	nodal kindreds, fission-fusion
3. Melayu Asli ("Aboriginal Malays)	Southern Aslian	sedentary farmers, collectors-for-trade	endogamy

[1] based on Benjamin (1986)

immigrated at different times, each subsequent wave displacing some of the previous inhabitants (see Carey 1976 for a summary of this literature). The foraging Semang (together with other "Oceanic Negritos" such as the Andaman Islanders and various Philippine groups like the Aeta and Agta) represent descendants of a previously widely distributed population now mostly replaced by later migrants. The second wave included the swidden farming Senoi peoples, followed later by the Melayu Asli.

As an alternative to the wave theory of successive colonization of the Peninsula, Benjamin (1976, 1980, 1985, 1986) has proposed an *in situ* model for the origin and diversification of Orang Asli cultures. The three traditions differentiated from a common linguistic and cultural matrix within the last few thousand years as the northern Aslian groups continued (and became more committed to) foraging, central groups adopted swidden farming and became increasingly sedentized, and southern groups became involved in trade relations with outsider groups (probably Austronesian-speaking peoples of the islands).

Clearly, these polar hypotheses do not exhaust the possibilities for explaining present-day cultural diversity. Bellwood (1993), for instance, has proposed one such intermediate position, maintaining that both regional continuity and successive flows of people into the Peninsula have occurred. He sees the population increases that follow from adopting agriculture as the cause for the spread of farmers to new regions (Bellwood 1996b). This argument extends the demic diffusion model of Cavalli-Sforza and colleagues (1993), originally applied to the spread of the European Neolithic through population expansion from the Near East.

But how might these different scenarios for Orang Asli prehistory be tested? Each is consistent with the broad outlines of the situation; each depends on different mechanisms to produce current cultural, linguistic and genetic similarities and differences among the three groups. The most direct test (that is, one based on evidence from the past rather than inference from current distributions) would be archaeological: Are there clear indications of population replacements and/or new influxes of peoples arriving in the Peninsula bearing diagnostic cultural markers of the traditions? My impression from the literature is that such evidence is not unequivocal (but I will not presume to argue either position in this forum!). Similarly, skeletal remains from archaeological sites with diagnostic morphological and/or metrical traits might provide direct evidence of migration. Again, such data as exist seem unable to resolve the issue (but see Bulbeck 1981 and his paper this volume).

Turning to inference from present-day distributions, linguistics has been used to substantiate both the *in situ* and the demic diffusion models. Benjamin's (1976) reconstruction of Malaysian cultural history was based on a comparative analysis of Aslian languages. The pattern of diversification of northern Aslian (mostly Semang) suggested a split between their ancestors and those of central Aslian speakers (Senoi) some 5000 years ago. This analysis links Semang and Senoi ancestry. Bellwood (1993), on the other hand, interprets the linguistic evidence to support population movement. He sees ancestral Senoi, speaking Austroasiatic languages, arriving from the north into the Peninsula beginning around 4000 BP as a result of expanding populations of rice farmers. Already resident ancestral Semang, speaking some now unknown languages, through a process of interaction with the migrant farmers, adopted their language. Meanwhile, similar pressures were driving expanding populations of Austronesian speakers, who somewhat later entered the Peninsula from the south as the ancestors of the Melayu Asli. Since the southern Melayu Asli now speak dialects of Malay, the Benjamin model requires a language switch on the part of these peoples as they assimilated to their maritime trading partners from the islands. Much more could be said about both these models' congruence with linguistic patterns, but it should be obvious that a definitive test has not been achieved.

How might genetics contribute to the solution of this problem? As our previous discussion suggests, the task of inferring history from current genetic patterns will not be easy (see also Fix 1979). Most importantly, the three traditions are not *species*; that is, not self-contained

evolutionary units. Their evolution is not simply a process of continuous divergence after a definitive split. Just as words may be borrowed between languages, so may genes flow between populations.

Despite the theoretical inappropriateness of dendrograms for describing relationships among infra-specific taxa, this method seems to be the principal tool of many geneticists (see especially Cavalli-Sforza *et al.* 1994). As Cavalli-Sforza *et al.* (1994:38) state, "The study of trees is the major technique for understanding the complex relationships between different *populations*. It offers a simple graphic aid for visualizing those relationships and a path to infer the possible evolutionary history behind them" (emphasis added). The problem is, of course, that by using the tool, the underlying assumption of fission and divergence is accepted. Often this assumption will be wrong.

Cavalli-Sforza *et al.* (1994) present such a tree based on some 31 "classic genetic markers" (blood groups, enzymes, etc.) from a series of Southeast Asian populations including the Semai, an Orang Asli (Senoi) population. I won't attempt to evaluate all the branches in this dendrogram but will concentrate on inferred relationships of the Semai.

Initially I might note that the population units of analysis in the tree are quite heterogeneous in size and geographical extent. For example, the Semai number perhaps 20,000 people; the "South Chinese," one of the other "populations" being clustered, are considerably more numerous. Leaving aside questions of sampling strategies (are the allele frequencies used in this analysis actually representative of the population unit labels?), the hierarchy of clusters surrounding the Semai do not suggest an obvious historical scenario. In the tree, Semai are closest to the Zhuang, a Tai-speaking "minority" population (some 13 million strong) of South China. Resisting any temptation to infer historical connection or derivation at this point, I should point out that the Semai are in a different major cluster (defined by the first two "fissions" separating these populations) from the Malays with whom they are now geographically contiguous and are also widely separated from their close linguistic cousins, the Khmer (mainly Cambodians). I might also note that Filipinos are an "outlier" group to the cluster that includes the Semai.

If gene trees accurately portray the history of populations, trees based on different genes should still show the same relationships; ie, all gene loci should be affected in identical ways by the historical processes of isolation and divergence. In fact, another tree (Saha *et al.* 1995) based on polymorphic allele frequences but clustering a different set of populations than those in the Cavalli-Sforza *et al.* (1994) study, is not identical to the Cavalli tree at all. This tree clusters Semai closely with Khmer and shows this pair being split rather distantly from the other populations. Since hemoglobin E, a non-neutral allele, is high in both these groups and not the others, natural selection may be strongly influencing this "history."

A major problem for genetic studies that cluster many populations is the number of alleles that can be compared; in theory, the greater the number of alleles, the greater the statistical power of the analysis. Thus Cavalli-Sforza *et al.* (1994:238) do not place great confidence in their Southeast Asian tree due to the few loci available for study. However, results from Saha *et al.* (1995) suggest that this is not the only problem. When Saha and colleagues increased the number of alleles in their study to 53, the Javanese population clustered closer to the Semai, not an intuitively obvious result.

If classical genetic loci do not provide a clear picture of Orang Asli affinities and history, what of the molecular data? The much more rapid rate of mutation in the mitochondrial genome and its uniquely maternal pattern of inheritance insuring that no recombination occurs can provide a different perspective on genetic relationships.

Although most attention has been focused on the origins of modern *Homo sapiens*, a number of studies have examined mtDNA variation in Asia and the Pacific (see paper by Baer this volume). Thus Melton and colleagues (1995) studied the distribution of a 9 base-pair deletion in the mitochondrial genome in populations ranging from Pakistan to Taiwan, the Philippines, and eastern Indonesia and including a sample of 30 Semai from Malaysia. The Semai were broadly similar in the frequency of this mutation to several insular Southeast Asian groups. It is worth noting (along with the paper by Baer in this volume) that a previous study (Ballinger *et al.* 1992) found a very different frequency of this "marker" (only 3% — one person in a sample of 32 Orang Asli, perhaps a dozen of whom were Semai, versus 37% in the Melton *et al.* study), pointing up the problem of representativeness of the data.

The tree of populations based on this deletion and three other substitutions in the control region of mtDNA (Melton *et al.* 1995) shows the Semai as closest to Filipinos (while in the dendrogram of Cavalli-Sforza *et al.* 1994, based on allele frequencies, Filipinos were a distant outlier to the group containing the Semai). But, more importantly, *all* the branches of this mtDNA tree are very deep, extending back nearly to the "common ancestor." Rather than a tree, this is a stump sprout. Indeed, it suggests that all the "splits" are roughly contemporaneous and ancient. Similar "star clusters" typify much of the human mtDNA data (Rogers and Jorde 1995), a pattern that has been attributed to an ancient bottleneck followed by population increase. Rogers and Jorde (1995) argue that this ancient demographic expansion of the human population has obscured more recent population history.

Finally, a genealogy of mtDNA haplotypes based on 191 polymorphic sites using restriction enzymes has been presented by Ballinger *et al.* (1992). This is not a tree of *populations* but rather one of haplotypes, that is, individual mtDNA types descended by mutation from a common ancestral type. The Orang Asli are represented in this tree by 33 individuals of various ethnolinguistic designations. Probably most are Semai or Temiar Senoi but 2 persons were perhaps Semang (Jehai?) and one or two were Melayu Asli. The key observation, however, is that Orang Asli haplotypes are found in almost every branch of this genealogy. That is, some Orang Asli are close to Sabah indigenes, others to Koreans, Vietnamese, Malays, etc. In other words, this molecular tree tells us almost nothing about the history of the Orang Asli other than that they are similar to other Asians. To go further, it may represent molecular variation that "predates the present geographic subdivision" of the region (Ballinger *et al.* 1992). Alternatively, this pattern might simply reflect gene flow among populations across the entire region over the last several thousand years.

These different trees are obviously telling us about different things. Thus the histories of genes or haplotypes are not necessarily those of the current ethnolinguistically defined populations whose histories we wish to reconstruct. The branching nodes (or coalescents) for some HLA genes are about 60 million years ago, not very informative for Orang Asli history. The coalescent for the 9bp deletion discussed above has been inferred to be about 58 thousand years ago (Redd *et al.* 1995). Clearly these systems will play different roles in historical analysis. Going beyond the history of *genes* to the history of *populations* requires careful attention to the units of analysis and the nature of the biological variation.

As the title of Baer's paper in this volume states, the genetic data are a patchwork. Allele frequencies are determined by many causal factors in addition to descent from a common ancestor. As she pointed out, many human alleles are strongly selected in malarious environments. Such alleles are well represented in the Orang Asli, most of whom inhabited malaria-ridden environments and continue to do so (Baer, manuscript). The alleles include Hemoglobin (Hb) E, G6PD deficiency, and Southeast Asian Ovalocytosis (SAO). To treat these adaptive alleles as neutral "markers" of population affinity is a mistake.

HEMOGLOBIN E, OVALOCYTOSIS AND HISTORY IN SOUTHEAST ASIA

Both of these adaptive alleles protect against malaria. Hemoglobin E is associated with mainland SE Asia, ovalocytosis with Island SE Asia.

Table 2 compares populations at several levels of a hierarchy from settlements (SA, RU, BU) of Semai, to regions (Perak, Pahang) of Semai, to ethnolinguistic groups (Semai, Temiar), and so forth. There is considerable variation among settlements even for these adaptive alleles. There are also clear historical patterns linking Semai with mainland Southeast Asia (Hb E) but also with Papua New Guinea (ovalocytosis), probably through intermarriage with Melayu Asli (traders with connections to sea-faring Malays) and the islanders of Indonesia (Fix 1995).

To represent the Semai (or the Orang Asli) as a node on a tree might provide a phenetic measure of net genetic differences from other such units. However, the actual evolutionary history of the Orang Asli is not one of binary fission and isolation. We could ask: Are the Semai on a branch with the Khmer or a separate branch with Sulawesi and the coastal New Guineans? Actually, they are connected with all these populations via a long history of widely

Table 2: Hemoglobin E and ovalocytosis frequencies.

		Ovalocytosis			Hb E
Population	Size	N	% Positive	N	q_E
Senoi	39000				
Temiar	13000	?	6.7	80	0.319
Semai	19500				
Perak	12000	242	6.6	332	0.255
Pahang	7500	545	21.3	520	0.215
SA	272	196	25.2	198	0.168
RU	107	81	28.4	75	0.346
BU	107	79	5.1	80	0.250
Melayu Asli	28000				
Temuan	10500	315	35.2	406	0.015
Jakun	11500	?	19.0	116	0.017
Malays	8×10^6				
Selangor		?	<0.3	536	0.015
Negri Sembilan		629	13.2	629	0.026
Indonesians	190×10^6				
Sulawesi	14×10^6	?	40.0-50.0		
Bali	3×10^6			219	0.018
Minangkabau		83	7.2	235	0.011
New Guinea	4×10^6				
Kar Kar Island		334	13.8		

ramifying gene flow. Indeed, this is the general conclusion from a "world-wide analysis of genetic and linguistic relationships of human populations" by Chen and colleagues (1995:610), "In view of the continuum of variation and the numerous interconnections resulting from gene flow, distances may be a far better representation than trees."

The take-home message, then, is a cautionary tale: History may not be read simply from gene trees.

REFERENCES

Baer, A. Rainforest malaria, mosquitos, and people. Unpublished manuscript.

Ballinger, S.W., T.G. Schurr, A. Torroni, Y.Y. Gan, J.A. Hodge, K. Hassan, K.-H. Chen and D.C. Wallace 1992. Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient Mongoloid migrations. *Genetics* 130:139-52.

Bellwood, P. 1993. Cultural and biological differentiation in Peninsular Malaysia: The last 10,000 years. *Asian Perspectives* 32:37-60.

Bellwood, P. 1996a. Early agriculture and the dispersal of the southern Mongoloids. In T. Akazawa and E.J.E. Szathmary (eds), *Prehistoric Mongoloid Dispersals*, pp. 287-301. Oxford: Oxford University Press.

Bellwood, P. 1996b. The origins and spread of agriculture in the Indo-Pacific region: Gradualism and diffusion or revolution and colonization? In D.R. Harris (ed.), *The Origins and Spread of Agriculture and Pastoralism in Eurasia,* pp. 465-98. Washington: Smithsonian Institution Press.

Benjamin, G. 1976. Austroasiatic subgroupings and prehistory in the Malay Peninsula. In P. Jenner, L.C. Thompson and S. Starosta (eds), *Austroasiatic Studies,* pp. 37-128. Honolulu: University of Hawaii Press.

Benjamin, G. 1980. Semang, Senoi, Malay: Culture-history, kinship and consciousness in the Malay Peninsula. Unpublished manuscript.

Benjamin, G. 1985. In the long term: Three themes in Malayan cultural ecology. In K. Hutterer and T. Rambo (eds), *Cultural Values and Tropical Ecology in Southeast Asia,* pp. 219-78. Ann Arbor: Center for South and Southeast Asian Studies.

Benjamin, G. 1986. Between isthmus and islands: Reflections on Malayan palaeo-sociology. Singapore: University of Singapore Sociology Department, *Working Paper No. 71.*

Bulbeck, F.D. 1981. Continuities in Southeast Asian Evolution Since the Late Pleistocene. MA Thesis: Australian National University, Canberra.

Carey, I. 1976. *Orang Asli: The Aboriginal Tribes of Peninsular Malaysia.* Kuala Lumpur: Oxford University Press.

Cavalli-Sforza, L.L., P. Menozzi and A. Piazza 1993. Demic expansions and human evolution. *Science* 259:639-46.

Cavalli-Sforza, L.L., P. Menozzi and A. Piazza 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Cavalli-Sforza, L.L., A. Piazza, P. Menozzi and J. Mountain 1990. Comment on Bateman et al., "Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language". *Current Anthropology* 31:16-9.

Chen, J., R.R. Sokal and M. Ruhlen 1995. Worldwide analysis of genetic and linguistic relationships of human populations. *Human Biology* 67:595-612.

Fix, A.G. 1979. Anthropological genetics of small populations. *Annual Review of Anthropology* 8:207-30.

Fix, A.G. 1995. Malayan paleosociology: Implications for patterns of genetic variation among the Orang Asli. *American Anthropologist* 97:313-23.

Melton, T., R. Peterson, A.J. Redd, N. Saha, A.S.M. Sofro, J. Martinson and M. Stoneking 1995. Polynesian genetic affinities with southeast Asian populations as identified by mtDNA analysis. *American Journal of Human Genetics* 57:403-14.

Parham, P., E.J. Adams and K.L. Arnett 1995. The origins of HLA-A,B,C polymorphism. *Immunological Reviews* 143:141-80.

Redd, A., N. Takezaki, S. Sherry, S. McGarvey, A.S.M. Sofro and M. Stoneking 1995. Evolutionary history of the CoII/tRNA (Lys) intergenic 9-bp deletion in human mitochondrial DNAs from the Pacific. *Molecular Biology and Evolution* 12:604-15.

Rogers, A.R. and L.B. Jorde 1995. Genetic evidence on modern human origins. *Human Biology* 67:1-36.

Saha, N., J.W. Mak, J.S.H. Tay, Y. Liu, J.A.M.A. Tan, P.S. Low and M. Singh 1995. Population genetic studies among the Orang Asli (Semai Senoi) of Malaysia: Malayan Aborigines. *Human Biology* 67:37-57.