CHAPTER 3

TESTING POPULATION DISPERSAL HYPOTHESES: PACIFIC SETTLEMENT, PHYLOGENETIC TREES AND AUSTRONESIAN LANGUAGES

Simon J Greenhill and Russell D Gray

The poets made all the words, and therefore language is the archives of history.
Ralph Waldo Emerson, Essays: 'The Poet'

Dispersals have been commonplace throughout the history of genus *homo* (Templeton 2002). However, it is only recently that scenarios about human population expansions have begun to be studied again after a long period of marginalisation (Anthony 1990; Burmeister 2000 and associated commentaries). Some authors, such as Diamond and Bellwood (2003), have argued that dispersals, especially those linked to the development of agriculture, are the 'most important process in Holocene human history' (p 597). The increased emphasis on the importance of migrations has been followed by the proliferation of dispersal scenarios. Diamond and Bellwood (2003) present evidence for a number of agriculture-driven dispersals: the spread of Bantu languages from Nigeria and Cameroon through subequatorial Africa, the colonisation of the West Indies by the ancestors of the Taino people, three expansions in prehistoric China involving speakers of Austro-Asiatic, Tai ('Daic') and Sino-Tibetian languages, the spread from northern Mesoamerica to the southwest of the United States associated with the Uto-Aztecan languages, the expansion of Oto-Manguean, Mixe-Zoquean and Mayan languages through Central America, a shift from the highlands of New Guinea by the speakers of Trans-New-Guinea languages, a major population expansion from Korea into China, the spread of Dravidian languages into southern India, a major influx of Afro-Asiatic languages and their speakers into Egypt and North Africa, the settlement of Europe by speakers of Indo-European languages, and finally, the Austronesian expansion.

Unfortunately, many expansion scenarios are little more than plausible narratives. A common feature of these narratives is the assertion that a particular line of evidence (archaeological, linguistic or genetic) is 'consistent with' the scenario. 'Consistent with' covers a multitude of sins. Rigorous tests require a measure of exactly how well the data matches the proposed scenario. They also require an explicit evaluation of alternative hypotheses. Perhaps the data are equally 'consistent with' many alternative hypotheses. Given the interest in hypotheses about human dispersal scenarios, a framework for the rigorous evaluation of these hypotheses is clearly desirable. Here we describe our attempts to apply a phylogenetic framework to linguistic data in an effort to test one of these scenarios – the Austronesian expansion.

PEOPLING THE PACIFIC

The prehistory of the Pacific region has long intrigued scholars. The brief outline that follows is based on Kirch's (2000) excellent overview. European speculation on Polynesian origins began with Fornander (1878) who proposed that Polynesians arose from the same parent group as the Vedic branch of the Arian 'race'. He suggested that they had spread through India before the Vedic Arians and intermixed with the Dravidian 'races' (see Kirch 2000). Other scholars were less explicit, but more lyrical – 'these far-extended Oceanic languages, sprung from the abysm of prehistoric time, were manifestly and admittedly of one stock or origin. What then was that origin?' (MacDonald 1907: vi–vii). Later, Smith (1921) also proposed an Indian origin for the Polynesians, based on oral traditions. Worldwide interest peaked in 1947 when Thor Heyerdahl embarked on a trip from Peru to Raroia in the Tuamotus in the balsa raft Kon-Tiki. This voyage, undertaken in an attempt to prove Heyerdahl's conviction that the Polynesians originated in South America, was not entirely successful (the raft had to be towed in the initial stages). Suggs (1960) noted that Heyerdahl's South American origin theory was so flawed as to be 'equivalent to saying that America was discovered in the last days of the Roman Empire by King Henry the Eighth, who brought the Ford Falcon to the benighted aborigines' (Suggs 1960: 224; see also Kirch 2000). Despite Heyerdahls' sea-faring exploits, converging evidence from linguistics, archaeology and genetics leaves no doubt that the origin of the Austronesians was Island Southeast Asia and not South America (Bellwood 1991; Pawley and Ross 1993; Bellwood *et al* 1995; Melton *et al* 1998; Spriggs 1999; Kirch 2000; Hurles *et al* 2003). The fact that Polynesian populations have the South American sweet potato (*kumara*) is a testament to the voyaging skills of the Polynesians. That is, the Polynesians travelled to South America and returned with cultural items and a food crop, whilst the South Americans kept their feet dry (Green 1985; Irwin 1992 (see new list); Finney 1996).

Current debates about the settlement of the Pacific continue to be equally flamboyant and fierce. The settlement of the Pacific provides an ideal model for testing dispersal hypotheses (Hurles 2003; Hurles *et al* 2003). There are relatively clear, testable scenarios about the colonisation of the Pacific. There is also a large amount of archaeological, genetic and linguistic data available to test these hypotheses and, as the proposed population expansions have occurred relatively recently, linguistic data are likely to show stronger signal than for dispersals that occurred at greater time depths.

Current evidence suggests that humans colonised the Pacific first around 56000 years BP (Roberts *et al* 2001), when stone-age hunter gatherers travelled from Island Southeast Asia through New Guinea to Australia and areas of Island Melanesia, eventually reaching the Bismarck archipelago by 39,500 BP (Leavesley *et al* 2002). More recently, around 5500 to 6000 BP (Bellwood 1991), another expansion occurred, spreading throughout the Pacific, rapidly reaching as far north as Hawaii, as far east as Easter Island (Rapanui) and as far south as New Zealand. This Austronesian expansion has been explained by a number of settlement models, all with convenient media-friendly labels. The first scenario, dubbed the 'Express Train' by Diamond (1988), proposes that this expansion was

made by a population with distinct languages, genotypes and cultural innovations, who arose from or near Taiwan. These Austronesians moved relatively rapidly through Island Southeast Asia, the Philippines, Oceania, Polynesia, and then onto New Zealand and Hawaii.

The Express Train scenario as proposed by Diamond proposes an explicit migration sequence:

1 Taiwan (5500 BP)

2 The Philippines by 5000 BP

3 Through Borneo (4500 BP) and into Sumatra (4000 BP), Java (4000 BP) and Madagascar (1500 BP)

4 Into Indonesia

5, 6 Along the coast of Papua New Guinea (3600 BP)

7 Through the Solomons (3600 BP)

8 Into Near Oceania around 3200 BP

9 Into Remote Oceania and Western Polynesia by 3200 BP

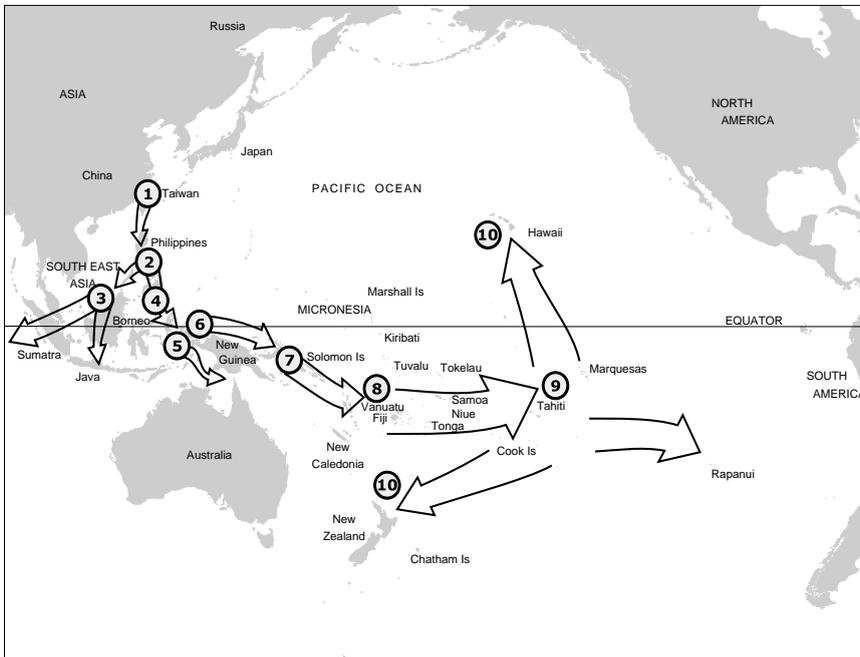10 Once in Polynesia, the Express Train moved north to Hawaii by 1500 BP and south to New Zealand by about 1000 BP.

Figure 3.1 Map of the Pacific showing the Express Train model of Pacific settlement (adapted from Diamond 1988, 1997).

The Express Train model makes a number of specific and testable predictions (adapted from Hurles *et al* 2003):

1   Austronesian languages originated in Taiwan.

2   The spread of Austronesian in the Pacific is closely associated with Lapita culture that contained innovations in agriculture, horticulture, fishing, pottery, weaving and long-distance sailing.

3   The Austronesians are genetically distinct from the indigenous non-Austronesian speaking populations.

4   The Austronesians spread relatively rapidly throughout Island Southeast Asia and then east throughout the Pacific.

5   In the initial Austronesian expansion there was little genetic mixing between the Austronesians and the pre-existing populations (but admixture has occurred since this expansion within Near Oceania).

6   There were no substantial pauses in the expansion from Taiwan to Polynesia.

A phylogenetic tree of languages that supports the Express Train scenario should have a typology that is a good fit to the geographic sequence shown in Figure 3.1. It should also have short internode distances (or branch lengths) as a result of the speed of the Austronesian expansion.

In contrast, the 'Entangled Bank' scenario (aka the Bismarck Archipelago Indigenous Inhabitants Scenario, BAIIS; see Green 2003) proposes that the Austronesians arose somewhere in Melanesia, as a result of 'interlocking, expanding, sometimes contracting and everchanging set of social, political, and economic subfields' (Terrell 1988: 647) between populations which 'kept more or less in touch with one another ever since the first arrival of people at least 45,000 years ago' (Terrell *et al* 2001: 107). While this hypothesis is obviously a theoretical possibility, it lacks specific predictions. As Lum has recently remarked, proponents of this hypothesis risk being 'vague to the point of uselessness' (Lum 2001: 116). At its most extreme this model would predict a maximally interconnected network. However, our best guess is that proponents of this model propose a partially interconnected network, where promiscuous prehistoric mingling has washed out any colonisation signal. So from this viewpoint, if we force linguistic data into a tree structure, we should expect to see no geographically ordered sequence, strong conflicting signals (ie a very low consistency index) and weak branch support values.

Numerous intermediate hypotheses exist on the continuum of possibilities between the Express Train and Entangled Bank scenarios (see Green 2003). The 'Slow Train' hypothesis (Hurles *et al* 2002) retains the same sequence as the Express Train model, but argues for a much higher level of mixing at the stops along the sequence. This model would predict the same nested tree-like structure as the Express Train model, but as a result of the multiple contacts between the Austronesian and the existing non-Austronesian speakers, there should be higher levels of conflicting signal (lower consistency index than the pure Express Train

model). The branches should be longer in these areas of contact as language contact generally increases the rate of linguistic evolution (Nichols 1997).

The next intermediate model is the 'Slow Boat' model (Oppenheimer and Richards 2001a; 2001b). According to this model the Austronesians had a much deeper history in Island Southeast Asia than is proposed by the Express Train model. This model draws on evidence from sources such as Irwin (1992) who have noted the presence of a safe 'voyaging corridor' extending from eastern Indonesia through to the Bismarck archipelago and Solomon Islands. Archaeological data suggest that as far back as 20000 BP, the inhabitants of Near Oceania were capable of 100–300 km inter-island voyages (Green 2003). Under this model the Austronesian expansion is seen as arising from these existing cultures, before beginning its spread east. Proponents of this view focus on an eastern Indonesian (around Sulawesi or Maluku) origin. Consequently, the Austronesian origin can only be traced as far back as Island Southeast Asia, not China or Taiwan as claimed by the Express Train scenario. A language tree supporting the slow boat model should show two main branches, one branch leading north and west containing languages spoken around the Philippines and Borneo, the other leading east into the Micronesian and Polynesian languages. The greatest language diversity should be around Wallacea (Eastern Indonesia) due to the greater length of time spent there.

The final, and most complex, intermediate hypothesis is a combination of the previous three models. This model lacks a catchy acronym but has been characterised best by Green (2003) as a Voyaging Corridor Triple I model (VC Triple I hereafter). In Green's model there is voyaging back and forth in the area between eastern Indonesia and the Bismarcks and the Solomons from 6000 to 3500 BP. The Lapita cultural complex is assembled through process of Intrusion, Integration and Innovation. The sequence of migration of the Austroneisan languages is similar to the Express Train, but is interspersed with a number of pauses. The first 'pulse' of migration occurred around 4,000 years ago from Taiwan to the Philippines, and was followed by an 800 year pause. The next pulse at around 3200 BP coincides with the development of the Lapita culture. The third pulse of colonisation occurred at 2000 BP, in marginal Western Polynesia, and was followed by the final two pulses around 1,000 and 800 years ago. These final pulses took place in Eastern and Southern Polynesia.

The pulse/pause aspects of this model make very specific predictions about the sequence, timing and location of population movements (from Green 2003 and Green, pers comm):

1   Taiwan (5500 BP)

2   The Philippines (around 4000 BP)

3, 4 Through Borneo, Sumatra, and Java

5, 6 Along the coast of Papua New Guinea

7, 8 Near Oceania (eg the Bismarck and Solomon Islands) around 3500–3300 BP

9   Into Remote Oceania and Western Polynesia by 3200 BP

10   Finally north to Hawaii by 1300 BP and south to New Zealand by about 800 BP

Dialect chains are an additional aspect of Green's model of Pacific settlement (eg Green 1999). Dialect chains are clusters of languages where closely situated languages are mutually intelligible, but further apart languages become increasingly unintelligible. Green's model proposes the existence of several large-scale dialect chains at avroious time points (eg across Remote Oceania, between Fiji and Western Polynesia, and between northern and southern Proto Polynesia (see Pawley 1975; Pawley and Green 1984; Green 1985; Pawley 1996; Kirch and Green 2001). As these dialect chains broke up, a number of non-hierarchical subgroups were formed.

In biology the equivalent of a dialect chain (a *metaspecies*) is known to cause problems for the construction of bifurcating trees because of the conflicting signals generated in the partial breakup, contact and separation that occurs along the dialect chain (Hoelzer and Melnick 1994, see Figure 3.2 below). The dialect chain
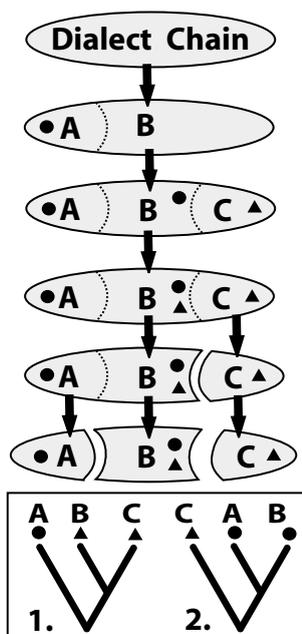

Figure 3.2   Schematic diagram showing the problems dialect chains cause for the construction of bifurcating trees. The dialects A, B and C are initially all mutually intelligible (note the permeable boundaries between the dialects). Innovations evolve in these dialects (● ▲) and diffuse through the network. However, if a dialect splits off from the network (eg the split between C and the other two languages), and this diffusion is only partially complete, then conflicting character histories can result. The ● character supports topology 1 whereas the ▲ character supports topology 2. So, under the Dialect Chain/Network Breaking model, areas where dialect chains were present should be poorly resolved in a phylogenetic analysis, and are better represented by a network diagram rather than a tree.

model predicts that in areas where pauses occurred, the tree topology should show a clear branching pattern with longer branches. In contrast, areas with dialect chains should have very weakly supported branches and substantial conflicting signal. The tree would therefore show a flat 'rake-like' topology where these dialect chains existed. There are thus rapid, slow and reticulate phases of population history under the VC Triple I model. Most linguists and archaeologists advocate some variant of this scenario. (For some recent genetic work that supports this model, see Matisoo-Smith and Robins 2004.)

We should stress at this point that both the 'Express Train' and 'Entangled Bank' models are improbable given current archaeological and genetic evidence, and that the prehistory of the Pacific is probably best characterised by one of the intermediate models.

LANGUAGE AND PACIFIC SETTLEMENT

The idea that languages can aid the study of ancient history has a long pedigree. For example, Schleicher (1865: 80) proposed that the '… observation and classification of languages also gives us the basis to conclude even more exact views of the prehistory of our race'. Some readers might assume that this view has been rendered obsolete by the advent of modern genetic techniques. Numerous genetic studies have indeed been conducted to test hypotheses about Pacific settlement. However, their results have often been contradictory. Studies of the maternally inherited mitochondrial DNA are generally interpreted as supporting an Express Train scenario (Melton *et al* 1998). In contrast, studies of paternally inherited Y chromosome haplotypes are claimed to support weak Entangled Bank or Slow Boat scenarios (Hurles *et al* 2002). The task of making accurate inferences about our past is a demanding one that requires the integration of genetic, linguistic and archaeological data (Hurles *et al* 2003). So exactly how well does the linguistic data fit the various models of Pacific settlement outlined above?

Two groups of languages are present in the Pacific: The Austronesia language family, and a diverse assemblage of languages known as Papuan or Non-Austronesian languages. Austronesian is a well-defined family of languages that is the world's largest and the most widely distributed (Lynch 1998; Bellwood *et al* 1995). These languages are spoken throughout the Pacific and out into the Indian Ocean, covering an area extending over half the world. The ~1,200 Austronesian languages are classified into 10 subfamilies, nine of which are spoken only by indigenous Taiwanese (Formosans). In contrast, languages of the 10th subfamily, Malayo-Polynesian, are spoken from Madagascar (47° east) to Easter Island (109° west), and their relative similarity suggests that they share a recent common origin. Moreover, non-Austronesian languages in Oceania are extremely diverse and are consequently expected to be much older. Although these languages are often lumped together in a heterogeneous catch-all Papuan 'group', they might eventually be classified into at least 12 major family groupings (Kirch 2000). These linguistic relationships can be interpreted as predicting two major genetic groups in the Pacific – the older Papuan lineages, and a more recent Austronesian group. However, language replacement (where one language replaces another) is a

well-known process in language evolution. Thus, we would not always expect a one-to-one relationship between languages and genetic markers.

Linguistic research on Austronesian languages has examined both the deep relationships (eg Blust 1977, 1978, 1981, 1984, 1999; Pawley and Green 1984) and the lower level subgroups (eg Pawley 1967, 1972; Ross 1988; Zorc 1986; Marck 2000). There is considerable consensus on many of the higher order groupings (eg Western Malayo-Polynesian and Oceanic) and the lower level subgroups (eg Eastern Polynesian). The sequence of these groups is 'consistent with' (*sic*) the Express Train, VC Triple I and Slow Train models of Pacific settlement (see Figure 3.3). However, as we mentioned at the beginning of this chapter, rigorous tests of migration scenarios require a measure of exactly how well the data matches the proposed scenario and an explicit evaluation of alternative hypotheses. The computational phylogenetic methods developed by biologists enable evolutionary hypotheses to be tested in a quantitative manner (Harvey and Pagel 1991; Pagel 1999a; Huelsenbeck *et al* 2002). In the sections below we will outline our efforts to apply these phylogenetic methods to hypotheses about the Austronesian expansion.

Testing the Express Train: an earlier attempt

Despite numerous parallels between the processes of biological and linguistic evolution (eg Schleicher 1865; Darwin 1871; Hoenigswald and Wiener 1987; Kirch
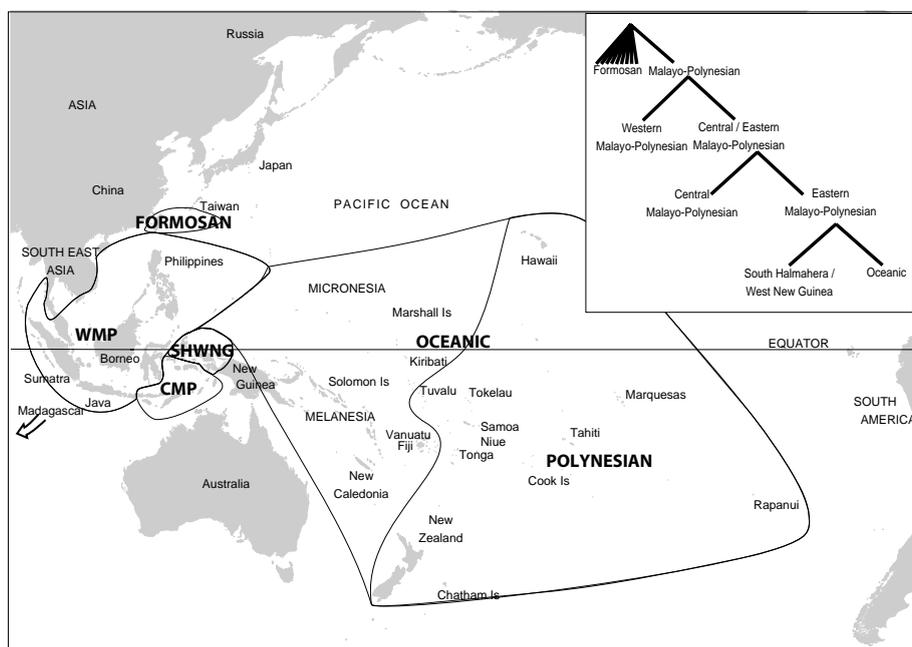

Figure 3.3   Map of the Pacific Ocean showing the distribution of Austronesian language subgroups and a tree (inset) depicting the generally accepted view of their relationships.

and Green 1987; Mace and Pagel 1994), historical linguists have not typically utilised the quantitative phylogenetic methods that have revolutionised evolutionary biology in the last 20 years (eg Felsenstein 2004). So, although linguists routinely use the 'comparative method' to construct language family trees from discrete lexical, morphological and phonological data, they do not use an explicit optimality criterion to select the best tree, nor do they typically use an efficient computer algorithm to search for the best tree. This is odd given that the task of finding the best tree is inherently a combinatorial optimisation problem of considerable computational difficulty (Felsenstein 1978; Graham and Foulds 1982; Warnow 1997).

Our first preliminary attempt to test the Express Train scenario of Pacific settlement in a quantitative manner used parsimony methods derived from evolutionary biology to construct an optimal language tree (Jordan 1999; Gray and Jordan 2000a, 2000b). A simplified example of the steps we followed is shown in Figures 3.4 and 3.5. In the first step of this process the lexical data were obtained

Figures 3.4 and 3.5 show the process of testing a migration scenario, in this case the Express Train model. This sample data set is obviously very small and provided for illustrative purposes only. (A) First the lexical data is grouped into cognate sets using systematic sound correspondences. Each semantic slot (e.. 'Eye', 'To Drink', 'Fire') may contain one or more cognate sets. These cognate sets are marked on the columns with different geometric symbols. (Note that for convenience we have organised the data into semantic slots. This is often not the case with linguistic data, and was not the case for the ACD.) The semantic slot for 'Right' contains two sets. The first set (denoted by squares) has words in the languages of Paiwan, Itbayaten, Bare'e, and Manggarai. The second set contains words that are related in Fijian, Tongan and Maori (denoted by a star). Where cognacy is unknown, it is denoted by a question mark. Note the presence of polymorphisms in 'To Drink' (Fijian), and 'To Choose' (Manggarai).

The second step (B) is to translate this lexical data into a binary matrix. This is simply done by coding all items in a cognate set as 1, and all other items as 0. Where there are two or more cognate sets per semantic slot, the item is duplicated with each cognate set being entered into its own column. So, the first column in the slot 'Bone' corresponds to the cognate sets marked with stars in part A, the second corresponds to the set marked by squares, and the third is that denoted by circles.

The next (underrated) step involves expressing the dispersal hypothesis as precisely as possible. The Express Train model can be formalised as an ordered 10-stage geographical character state tree reflecting the sequence of expansion from Taiwan through the Pacific (C). If this scenario is correct, then it should map onto our soon-to-be-built language trees very well. The optimal fit of the Express Train hypothesis (9 character steps) is shown in D.

The final step is to generate some trees from our binary data matrix in B. The tree shown in E was found using a parsimony search in PAUP* (Swofford 2002). Comparing this tree to the optimal tree (D) shows a number of departures: specifically Buli is shunted down the tree and grouped with Javanese, whilst Itbayaten and Bare'e are located further up the tree. Mapping the Express Train geographical character (C) on to the parsimony tree (E) gives a scenario fit of 11. The maximum likelihood tree fits slightly better (10 steps).

40   The Evolution of Cultural Diversity: A Phylogenetic Approach

(A) Table of Lexical Items

	Bone		To Drink		Fire		Right	
Paiwan	tsuqela⁺	★	t-em-ekel		sapuy	★	ka-navaɬ	■
Itbayaten	tuqgaŋ	★	quminun	★	hapuy	★	wanan	■
Javanese	baluŋ	?	ŋombe		geni		teŋen	
Bare'e	wuku	■	inu	★	apu	★	kana	■
Manggarai	toko	■	inuŋ	★	api	★	wanaŋ	■
Buli	loŋ		dom		yap	★	wela	
Motu	turia	●	inu-a	★	lahi	★	idiba	
Fijian	sui-na	●	gunu-va, unu	★	bukawaqa		i matau	★
Tongan	hui	●	inu	★	afi	★	mata?u	★
Maori	iwi	●	inu	★	ahi	★	matau	★

	Eye		Blood		To Choose	
Paiwan	matsa	★	djamuq		p-n-iliq	★
Itbayaten	mataq	★	rayaq	■	mamiliq	★
Javanese	mata	★	getih		milih	★
Bare'e	mata	★	daa	■	pili	★
Manggarai	mata	★	dara	■	lir, pilé	★
Buli	mta	★	laflaf		fisan	
Motu	mata	★	rara	■	hidi-a	?
Fijian	mata-na	★	drā	■	digi-a	
Tongan	mata	★	toto	★	fili	★
Maori	mata	★	toto	★	whiri	★

Figure 3.4

A) Express Train Model
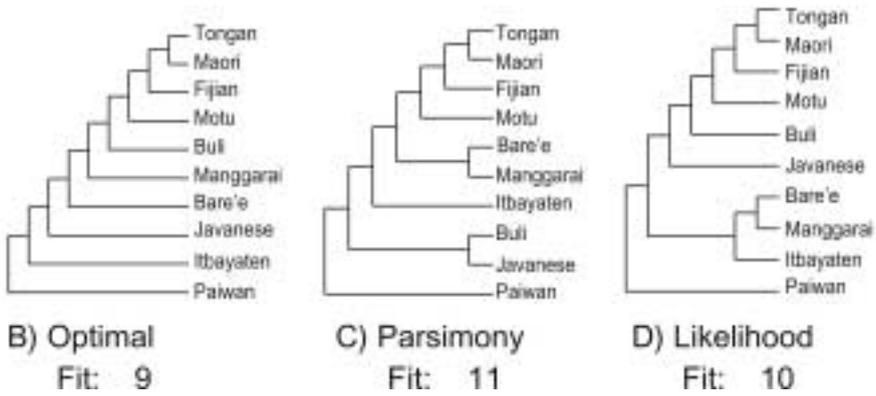
B) Optimal
Fit: 9

C) Parsimony
Fit: 11

D) Likelihood
Fit: 10

Figure 3.5

by converting information from Blust's Austronesian Comparative Dictionary (ACD, Blust pers comm) into a matrix suitable for phylogenetic analysis. The ACD is about 25% complete and is comprised of 5,185 cognate sets from over 200 languages. Linguists infer that words are 'cognate' (related by descent) by if they show recurrent sound correspondences and have a similar meaning (see Figure 3.4 for examples). We converted this lexical information into a binary matrix of 5,185 cognate sets and 77 languages (languages with little data were removed from the analysis). It could be argued that we should have grouped the data into semantic slots, as linguists sometimes do when they construct wordlists of basic vocabulary. Instead of binary characters we would then construct multi-state characters with the different cognate sets within each semantic slot having a separate character state. While this has the advantage that the characters are then independent, we did not use multi-state coding because of the conceptual and practical problems it creates (see Atkinson and Gray, in press for a discussion of this issue). Semantic slots are not 'objects' of linguistic evolution (cf Evans, Ringe and Warnow, in press). Rather, they are vaguely defined artificial linguistic constructs with mercurial boundaries. Cognate sets, on the other hand, constitute discrete, relatively unambiguous heritable evolutionary units with a birth and death. It is also much easier to develop tractable maximum likelihood for binary charcters (Pagel 2000a; Gray and Atkinson 2003; Atkinson and Gray, in press). We thus used binary coding where the presence of a cognate was coded as a one in the matrix and an absence as zero. This data was analysed using the programme PAUP* (Swofford 2002) to search for the most parsimonious tree. The Express Train scenario of Pacific settlement was then formalised as an ordered geographical character state tree, with each language assigned to a 'station' according to its place in the sequence (Figure 3.5C). This ordered geographical character was then mapped onto the most parsimonious tree to measure the fit between language tree and the Express Train model. The most parsimonious tree fitted the model quite closely (18 steps with a minimum possible fit of 9). A null distribution was constructed by randomly assigning the geographical character states on the character state tree to produce 200 random geographical sequences. These random sequences were then mapped onto the language tree. This null distribution had a range of values from 95 to 122. We concluded that these results supported the *sequence* of movements proposed by the Express Train scenario of Pacific settlement and was incompatible with the Entangled Bank model. We would now note that this sequence is also compatible with the VC Triple I and Slow Train models of Pacific settlement (albeit with a different rooting for the later).

Evaluation of Gray and Jordan's test

(a) Phylogenetic uncertainty

The number of potential trees is vast (Felsenstein 1978). For 10 taxa there are 34 million rooted trees, for 77 taxa the number of trees contains about 130 zeros. The problem of finding the optimal tree is known to be 'NP-Complete' (Graham and Foulds 1982). This means that a search for a tree is not guaranteed to return the

optimal tree in polynomial time. Although Gray and Jordan ran a large number of randomly seeded heuristic searches they still sampled a small number of trees from the space of all possible trees. This means that they may have missed shorter trees. Indeed, three more parsimonious trees with minor differences in topology have been found post-publication (R Gray and D Saul, pers comm). These new trees fit the Express Train scenario slightly less well than the original tree (the ordered geographical character has a length of 19 rather than 18 when it is mapped on to these trees). However, the existence of a slightly more parsimonious tree is not the main problem facing Gray and Jordan's analysis. Phylogenetic methods attempt to estimate the true tree (Felsenstein 2004). All estimates have uncertainty associated with them. A more significant problem for the Gray and Jordan analysis is that the use of one tree does not take into account the uncertainty in the estimation of this tree.

The technique of bootstrapping (Diaconis and Efron 1983) was introduced to phylogenetics by Felsenstein (1985a), as a means of estimating the uncertainty associated with tree estimates. Bootstrapping randomly samples from the original data set with replacement, generating numerous bootstrap samples, or 'pseudo-replicates', from which the statistic of interest is calculated. This process mimics repeated sampling from the population (Diaconis and Efron 1983). Greenhill (2002) applied this bootstrap technique to the ACG data. He conducted 1,000 bootstrap replicates, which generated 1,255 most parsimonious trees. The tree majority rule consensus topology (Figure 3.7) largely conforms to that expected from previous linguistic work (Grimes *et al* 1995). However, many of the branches of the tree are not strongly supported. It is important to remember that the Express Train scenario does actually predict some areas of weak support, namely those areas where rapid population dispersal or large amounts of contact occurred. In the former case the lack of signal will lead to low bootstrap values, while in the latter low bootstrap values will be produced by conflicting signals.

Rather than map the Express Train scenario on to a single tree, Greenhill (2002) mapped it onto the entire bootstrap sample of trees. The mean fit of the scenario on the bootstrap trees was slightly greater than that obtained by Gray and Jordan, but still an order of magnitude less than their null model. Greenhill therefore concluded that, despite the slightly poorer fit, these bootstrap results still provided strong support for the *sequence* of Pacific settlement proposed by the Express Train model (of course this sequence is also compatible with the VC Triple I and Slow Train models as well).

(b) Better models and Bayesian inference

In the last decade or so, parsimony methods have been largely superseded by maximum likelihood techniques for constructing evolutionary trees. These likelihood techniques (Felsenstein 1981; Yang 1996a, 1996b) calculate the likelihood of a tree and associated branch lengths. This likelihood score is proportional to the probability of observing the data given the model of evolution (Pagel 1999b; Steel and Penny 2000). Likelihood methods are capable of utilising
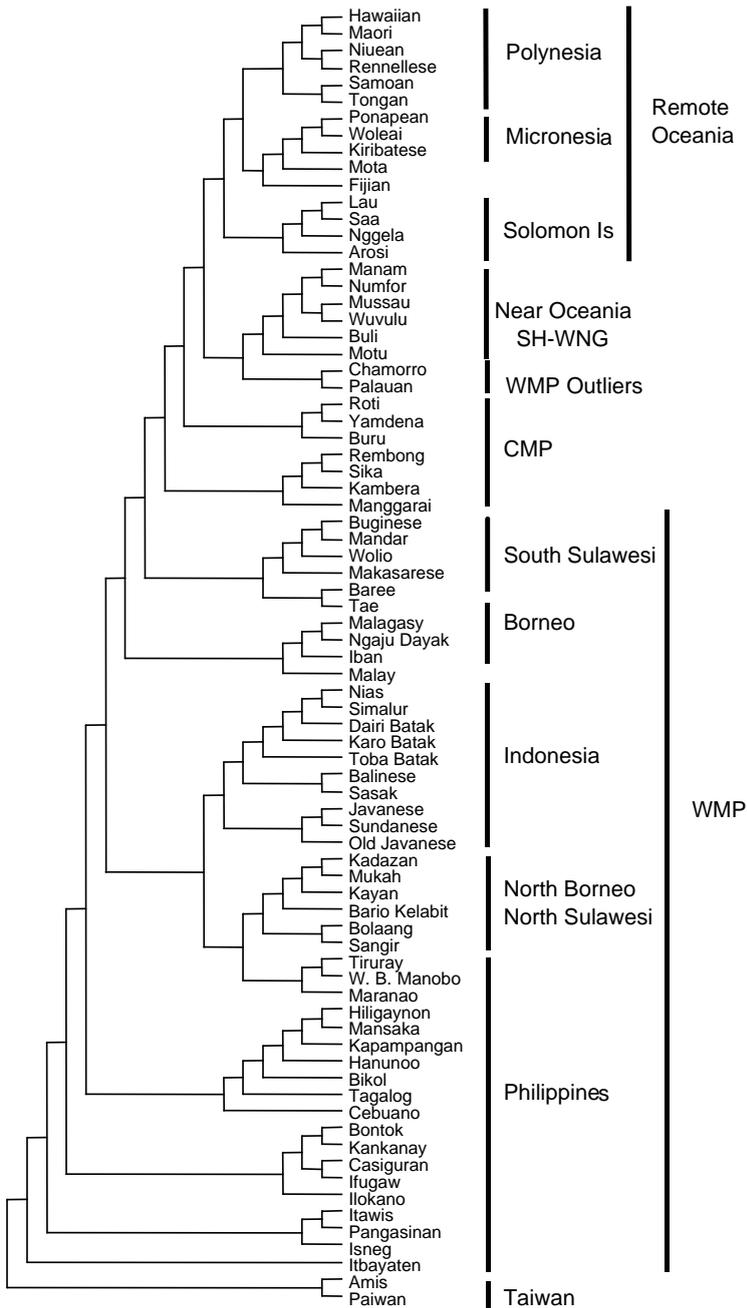
Figure 3.6 Gray and Jordan (2000a) tree, with some language subgroups marked. WMP = Western Malayo Polynesian, CMP = Central Malayo Polynesian, SH-WNG = South Halmahera, Western New Guinea. The tree is rooted with the Taiwanese languages.
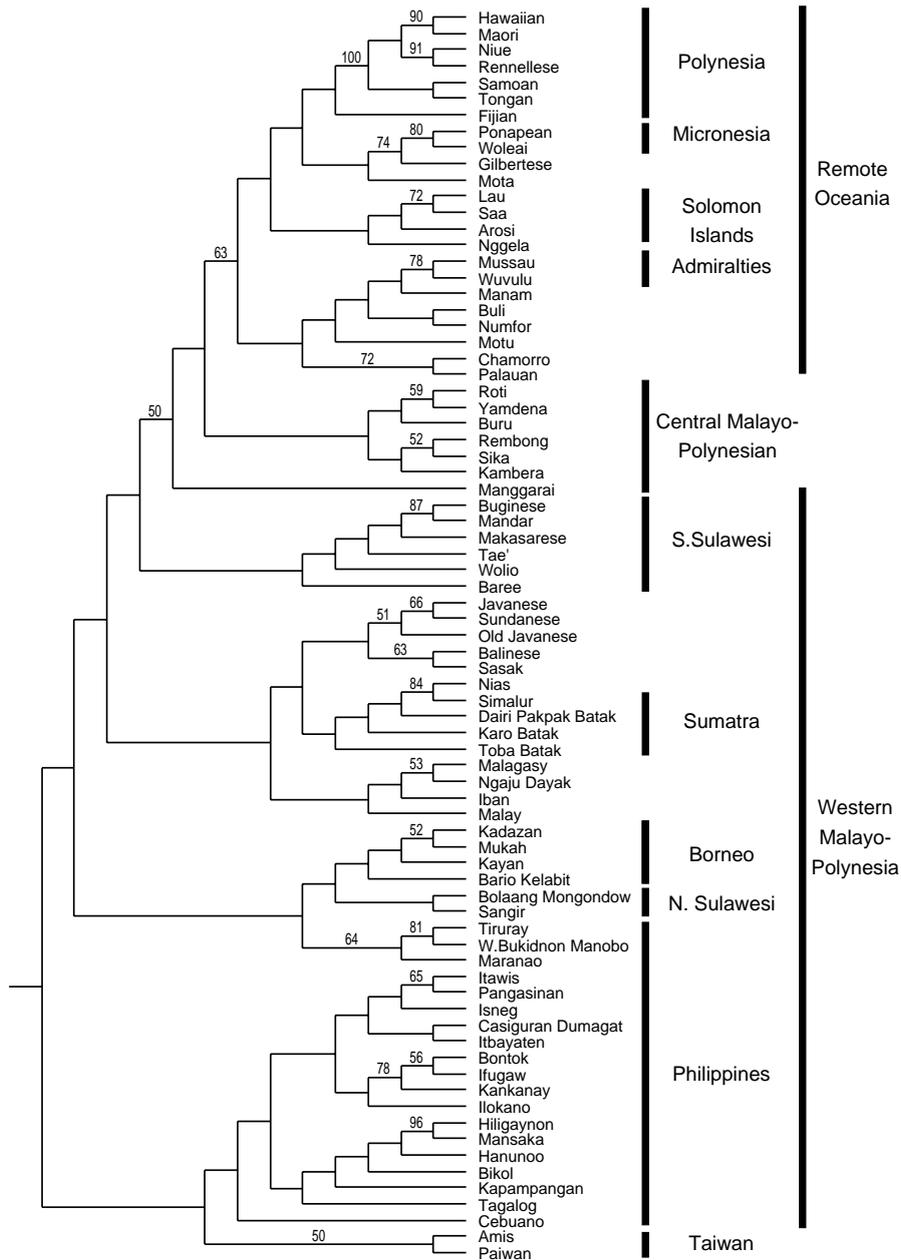
Figure 3.7 Majority rule consensus tree of 1,000 bootstrap replicates under a 5:1 cognate loss/gain ratio. Bootstrap support values greater than 50% are shown on the tree. Some geographic and linguistic subgroups are labelled. Note that the 5:1 coding roots the tree with a clade that includes the two Taiwanese and most Philippines languages. This might be interpreted as cause for excitement among proponents of the Slow Boat model. However, analyses we have conducted reveal that the absence of straight Taiwanese root for the tree reflects the unevenness of the sampling of cognate sets. The dataset contained only a relatively small number of cognates for the Taiwanese languages (210 for Amis, 317 for Paiwan).

more information in the data than the parsimony approach. Additionally, likelihood methods allow inferences to be based on more accurate models of the evolution process than that implicit in the use parsimony methods. These more realistic models thus lead to more accurate inferences (Felsenstein 2004). For example, allowing different sites to evolve at different rates using a gamma distribution (Yang 1993, 1994) can improve tree estimation (Swofford *et al* 1996; Posada and Crandall 2001; Felsenstein 2004).

One major drawback of the likelihood approach is that any analysis with greater than about 20 taxa becomes so time-consuming as to be impractical. Fortunately, the recent development of Bayesian Markov Chain Monte Carlo (MCMC) methods have provided a way around this problem. Bayesian methods combine one's prior beliefs about a hypothesis (prior probability) with the likelihood of that tree given the data and specific model of evolution (Huelsenbeck *et al* 2001; Shoemaker *et al* 1999). The posterior probability distribution of trees can then be used to simultaneously estimate the tree topology and the uncertainty in this estimate. This posterior probability distribution can be approximated using MCMC methods that perform a random walk in tree space, preferentially visiting areas of high probability (Lewis 2001).

Greenhill (2002) conducted a preliminary MCMC run on the ACD data using a two state 'restriction site' model in Mr Bayes (2.01), that allowed unequal character state frequencies. Rate variation across sites was estimated using a gamma distribution (Yang 1993, 1994). All other parameters were estimated using flat priors (ie equal probabilities were placed over the parameters' range) (Huelsenbeck *et al* 2002). This MCMC analysis ran for 2,500,000 generations, sampling one tree every 1,000 generations for a total of 2,500 trees. The first 1,000 trees were then discarded as 'burn-in', (Huelsenbeck *et al* 2001). A majority rule consensus tree was constructed from the remaining 1,500 trees and is shown in Figure 3.7.

Predictably perhaps, given the more realistic likelihood model used, the topology of the MCMC trees was more congruent with previous linguistic work than the parsimony trees. The MCMC consensus tree also shows far better resolution. In contrast to the bootstrap tree the majority of the tree is well supported. All of the major geographic and linguistic groupings are positioned appropriately, with only a few recalcitrant languages like Motu, Chamorro and Palau, and the languages from Borneo and Brunei situated oddly. Motu groups fall within the remote Oceanic languages, instead of the near Oceanic languages; this incongruent position is possibly the result of Motu's use as a trading language and *lingua franca* by missionaries in the area (Tryon 1995). We would therefore expect Motu to have a large amount of borrowing and language contact-induced change. Chamorro and Palau are spoken in the Philippines, but are inserted in between the South Halmahera/Western New Guinea languages and those of Central Malayo-Polynesia. This misplacement also appears to be the result of large-scale language borrowing (Gray and Jordan 2000a). Finally, the Borneo and Brunei-based languages, Bario Kelabit, Kadazan, Kayan and Mukah, are placed basally next to the Formosan languages. These four languages are spoken primarily by foraging populations. These languages may have retained more

'archaic' word forms than the other Western Malayo-Polynesian languages, that have undergone a period of linguistic levelling. These retained archaic forms could therefore produce the basal placement of these Western Malayo-Polynesian languages. Despite these minor incongruities, the MCMC trees are clearly more congruent with the standard Austronesian subgroupings than the parsimony trees. The Express Train scenario geographical character also fits the 1,500 MCMC trees slightly better than the bootstrap trees (Figure 3.8).

(c) Null models and alternative hypotheses

Demolishing straw men is always a tempting strategy. In biology it is not uncommon for researchers to reject some lame 'null' model and then happily conclude that their favourite alternative hypothesis is supported. Gray and Jordan (2000a) fell into this trap. Recall that the original null model used by Gray and Jordan randomly assigned the geographical character states across the tree (subject to the constraints of the same ordering and frequencies of these states). This distribution is a null model for the lack of any geographic signal rather than for the Express Train *sequence*. The critical part of the Express Train model is the sequence of settlement events, and even if all of our language subgroups were monophyletic (comprised of a common ancestor and all its descendant languages), good model fits could be obtained without the deeper relationships fitting the predicted sequence. To test this possibility we calculated the fit that would result if the languages at each of the geographic 'stations' formed monophyletic groups but were visited in a random order. As the Express Train model had 10 character states or 'stations' (Figure 3.5C), we generated all possible unrooted trees with 10 taxa (N = 2,027,025). The Express Train character was then mapped for each of these trees to produce a null distribution for the settlement sequence. In Gray and Jordan's optimal tree some of the main language subgroups were not monophyletic. This increased the number of steps required by the Express Train scenario by eight additional steps. We therefore increased the lengths of the 10 taxa tree model fits by eight character steps to facilitate a fair comparison with this more appropriate null model. The results of this analysis are shown in Figure 3.8. In contrast to the initial null distribution the fit of this more appropriate null model is only slightly worse than the MCMC trees. This suggested that most of the signal in the ACD data is for the language subgroups rather than the sequence they are linked in. The support for the Express Train *sequence* is therefore weaker than Gray and Jordan concluded. (However, this does not mean that the results support the Entangled Bank hypothesis. The data do contain clear phylogenetic signal. It is this signal that correctly reproduces most of the subgroups.) The question that remains is whether the lack of resolution of the deeper branches (see the low posterior probabilities in Figure 3.9) reflects a lack of signal, or the presence of conflicting signals (as would be predicted by the VC Triple I and Slow Train models).

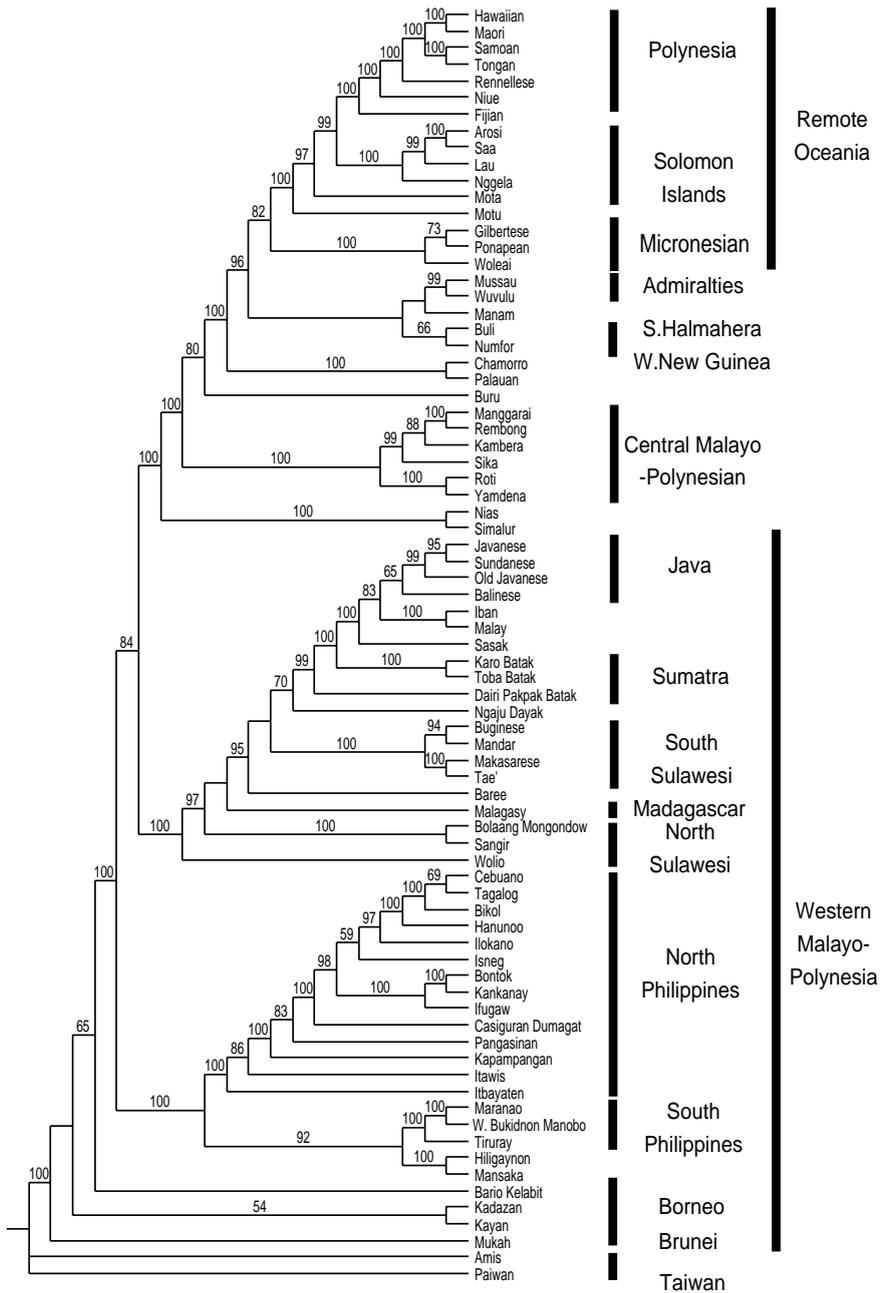48   The Evolution of Cultural Diversity: A Phylogenetic Approach

Figure 3.8   Majority rule consensus tree of the 1,500 post-burn-in MCMC trees. Posterior probability values greater than 50% are shown on the branches. Some geographic and linguistic subgroups are labelled. The tree is rooted with the Taiwanese languages.
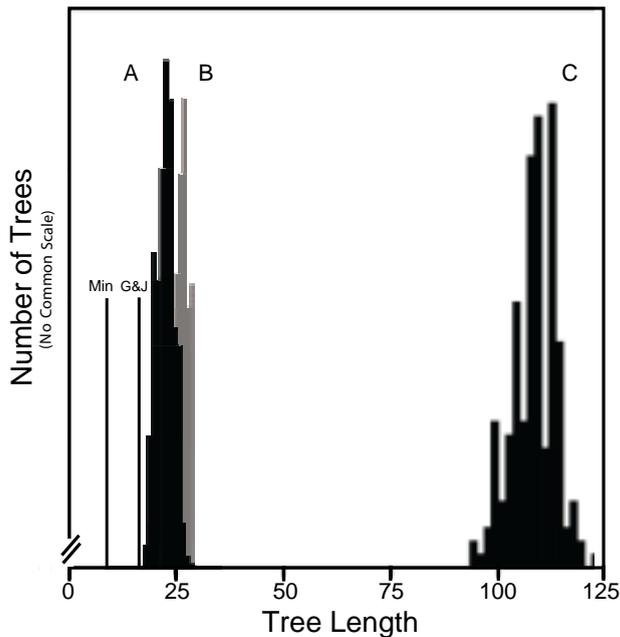
Figure 3.9 An evaluation of the strength of support for the Express Train sequence of Pacific settlement. The fit obtained by Gray and Jordan (2000a) is compared to that obtained when the Express Train character is mapped on the MCMC sample of trees (A). (B) shows the improved null distribution discussed above and (C) show the null distribution used by Gray and Jordan.

(d) How tree-like are Austronesian language relationships?

Gray and Jordan (2000a) report a consistency index for their parsimony analysis of the ACD data of 0.25. This statistic measures the fit of the data on the tree, and has a maximum value of 1. The relatively low value suggests that either borrowing has produced substantial amounts of conflicting signal or there have been a large number of cognate losses. Both of these possibilities are reasonable. Recent borrowing is common in Austronesian languages. For example, Indonesian languages have borrowed locally between Javanese, Sundanese, Minangkabau, and Balinese, as well as from more exotic languages like Sanskrit, Arabic, Persian, Chinese, Portuguese, Spanish, Dutch, French, and English (Moeliono 1994). The frequent 'founder events' that occurred when relatively small populations initially settled vacant islands in the Pacific could also have produced lots of cognate losses (Green 1966; Kirch and Green 1987). Phylogenetic programmes will generate a tree regardless of how well the data is explained by a tree model (Bateman *et al* 1990; Moore 1994b; Terrell 1988; Terrell *et al* 2001). To assess whether the low posterior probabilities for the basal parts of the ACD tree (Figure 3.8) are produced by strong conflicting signal or lack of signal, a method is required that does not force the data into a tree. NeighbourNet analysis provides such a method

(Bryant and Moulton 2004). Bryant *et al* (Chapter 5, this volume) give a detailed account of this approach. In brief, NeighbourNet is an agglomerative method of network construction that can recover known reticulations in linguistic data. If the data is tree-like, NeighbourNet produces a tree. Where there is conflicting signal in the data, NeighbourNet produces a box-like section of the graph. A NeighborNet analysis of the ACD data generates a network which can be seen in Figure 3.10. The network displays many of the major Austroneisan subgroupings. This suggests that the Entangled Bank model is not an adequate explanation of the data. However, the analysis also shows considerable amounts of conflicting signal (eg within the Philippines and South Sulawesi languages) and is poorly resolved in many areas. The lack of resolution is particularly pronounced for the deeper relationships where the graph resembles a star phylogeny. This suggests that the lack of strong support for the Express Train *sequence* of Pacific settlement reflects a lack of signal in the ACD dataset rather than strong conflicting signal.
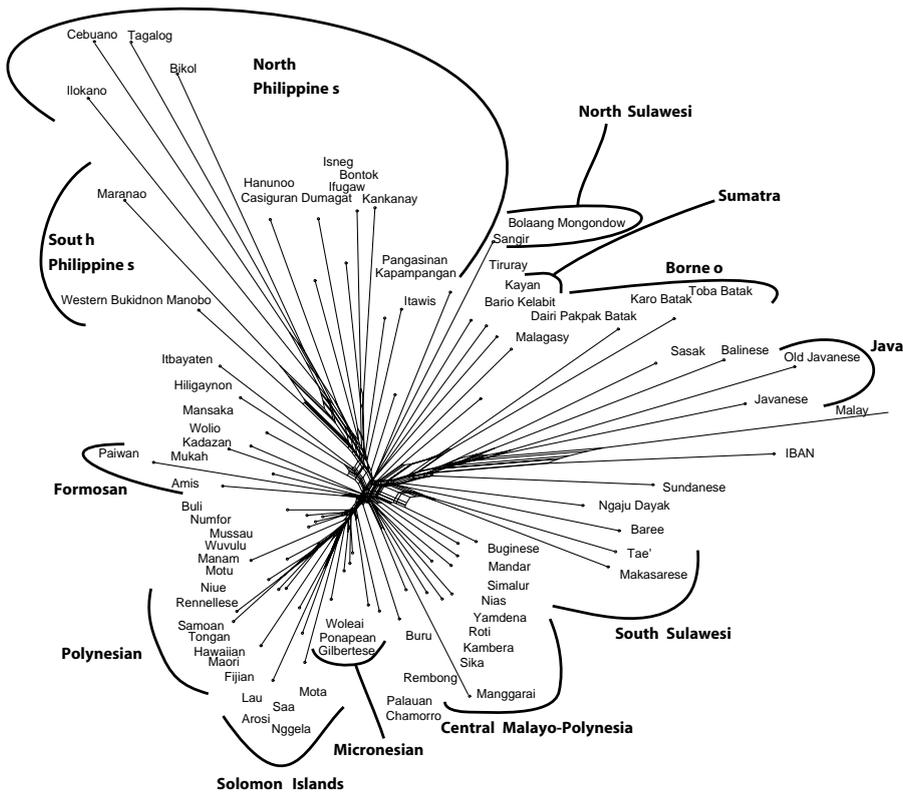

Figure 3.10   A splits graph showing the results of a NeighbourNet analysis of the 77 Austronesian languages coded from the ACD and used in Gray and Jordan's (2000a) parsimony analysis. Some subgroups are marked. Box-like sections of the graph show conflicting signals due to borrowing.

(e) Data problems

Bad workers reportedly blame their tools. Bad phylogeneticists might be accused of blaming their data. However, the major problem faced by the analyses discussed above really is the lack of sufficient, evenly sampled data. Although the ACD is a large database with over 5,000 cognate sets, it is only about 25% complete (we should point out that the database was not compiled with phylogenetic analyses in mind). In an effort to provide a balanced sampling of Austronesian subgroups, Gray and Jordan were forced to include nine languages with fewer than 150 cognates. Their dataset therefore included languages with very little information, such as Mussau, which has 38 cognates, Wuvulu with 46 cognates, and Manam with 59 cognates. Phylogenetic methods are known to have problems if the sequences are short (eg Nei *et al* 1998) or, as in this case, the amount of information for each language is small. In addition to lack of information for specific languages, there is uneven sampling of the language subgroups. For example, there are only two of the 20 Formosan (Taiwainese) languages in the dataset (Amis and Paiwan), and about eight of the 500 or so Oceanic languages. In total, the dataset contains only 77 of the more than 1,200 Austronesian languages. These deficiencies are crucial because accurate branch length estimates would enable us to discriminate between the Express Train and Slow Train scenarios. The unequal amounts of information for each language make accurate branch length estimates impossible. Similarly, unevenness in the sampling of the different language subgroups makes inferences about the settlement sequence less powerful than would otherwise be the case.

Conclusions

The results discussed above can be summarised in just a few words: one scenario down, four to go. Our results do not support the 'Entangled Bank' model of Pacific settlement. There is phylogenetic signal in the lexical data. This signal enables many of the major Austronesian subgroups to be correctly inferred (see Figures 3.7 and 3.9). There is also some signal reflecting the sequence of movements predicted by the Express Train (and the VC Triple I and Slow Train scenarios). However, when this signal is evaluated with an appropriate null model it appears relatively weak (see Figure 3.8). The weakness of the signal is a consequence of both limitations in the ACD database, and the presence of some borrowing (see Figure 3.10). With the present data we are unable to distinguish between the Express Train, the VC Triple I, the Slow Train, and the Slow Boat scenarios. To be able to discriminate between these scenarios we need:

1 Better estimates of the deeper relationships, including an evaluation of alternative rootings of the tree. This would provide a more powerful test of both the Express Train sequence and the Slow Boat model.

2 Accurate estimates of the amount of borrowing (preferably with an indication of whether this borrowing is recent or ancient). This would help us to discriminate between the Express Train, VC Triple I and Slow Train models.

3   Accurate branch length estimates. Rate variation is likely to be pronounced between Austronesian languages (Blust 2000) but this problem is not insurmountable. Gray and Atkinson (2003) have recently used a combination of Bayesian phylogenetic inference and rate smoothing methods to infer the divergence time for the Indo-European languages. Such techniques do not assume a constant 'glottoclock' or rate of lexical change, and as such have enormous potential to facilitate a closer synthesis of linguistic, archaeological and genetic data. Given accurate branch lengths and several uncontroversial calibration points, we should be able to distinguish between the differing time-depth predictions of the Express Train, the Slow Boat and the VC Triple I models. The VC Triple I model with its five dispersals pulses and four pauses makes especially bold testable time-depth predictions.

We have the computational methods required to achieve these aims, but we need to improve two aspects of our data. Recall that the ACD is only about a quarter complete, and that we had to remove a number of languages due to a lack of data. This led to patchy data sampling of both cognates and languages. The uneven sampling of cognates means that the branch lengths are extremely variable and poorly estimated. There is also uneven sampling of the language subgroups. To be able to test between the settlement scenarios, we need to have a more representative sampling of the languages. We are currently working on solving both of these problems. We are collaborating with Robert Blust (University of Hawaii) to produce a much larger, evenly sampled dataset. This dataset currently contains about 280 Austronesian languages and uses a modified version of the Swadesh 200 wordlist. The Swadesh wordlist includes items of basic vocabulary such as basic nouns (eg, body parts, colours, animals), simple verbs and kinship terms. Retention rates for items on the Swadesh list are claimed to be higher, and borrowing rates lower (Embleton 1986), although this does appear to vary within Swadesh items (McMahon and McMahon 2003). This data should therefore enable us to obtain better estimates of the deeper relationships between the Austronesian subgroups. It should also enable us to measure branch lengths and the extent of borrowing much more accurately. Most crucially, it will enable us to estimate divergence times without assuming constant rates of lexical change. This is critical because it will enable us to discriminate between the more complex hypotheses about the settlement of the Pacific. The challenge of making rigorous inferences about human population dispersals is not easy, but it is fascinating, engaging, and increasingly tractable.

Acknowledgements