Past Human Migrations in East Asia

Matching archaeology, linguistics and genetics

Edited by
Alicia Sanchez-Mazas,
Roger Blench,
Malcolm D. Ross, Ilia Peiros and
Marie Lin

Routledge Studies in the Early History of Asia

Past Human Migrations in East Asia

Matching archaeology, linguistics and genetics

Edited by
Alicia Sanchez-Mazas, Roger Blench, Malcolm D. Ross, Ilia Peiros and Marie Lin

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2008
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

Typeset in Times New Roman by
HWA Text and Data Management, London
Printed and bound in Great Britain by
MPG Books Ltd, Bodmin

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN13: 978–0–415–39923–4 (hbk)
ISBN13: 978–0–203–92678–9 (ebk)

ISBN10: 0–415–39923–8 (hbk)
ISBN10: 0–203–92678–1 (ebk)

Methodological issues

Linking genetic, linguistic and archaeological evidence

Roger Blench, Malcolm Ross, and Alicia Sanchez-Mazas

1. The problem: linking linguistics, archaeology and genetics

The concept of linking linguistics, archaeology and genetics in the reconstruction of the past is becoming a commonplace at certain types of academic conference, but the reality is that each discipline largely pursues its own methods and what little interaction there is remains marginal. Generally, despite much talk of interdisciplinary work, many of the questions asked are internal to the individual discipline, are addressed to colleagues and do not concern the larger sphere of understanding the past.

Some of the talk of interdisciplinary work seems to imply that one day there will be a super-discipline investigating the early human past in which the disciplines of archaeology, historical linguistics and genetics will somehow be merged. This is a misunderstanding. The methods and data of these disciplines are distinct and, importantly, they provide independent support for hypotheses about the past. For example, interdisciplinary work on the history of Austronesian speakers has been reasonably successful, especially where archaeology and linguistics are concerned. This success came, however, only when members of the two disciplines stopped piecing together their findings in a multidisciplinary jigsaw. Archaeologists who prior to the 1980s accepted the glottochronological findings of linguists (e.g. Bellwood 1979) found themselves led up the garden path. It was only when scholars in the two disciplines correlated the results of their single-discipline researches that cross-disciplinary work began to make real sense (Spriggs 1989). Such correlation of course entails some understanding of the status of results in the other discipline.

Since the middle of the 20th century two sets of reasoning procedures have been used by historical linguists, and both are used to produce phylogenetic trees. The comparative method dates from the 19th century, and identifies groups of related languages by reconstructing shared innovations (Ross, Chapter 6, this volume): it is inferred that a set of languages forms a subgroup, i.e. shares a common ancestor, if they share innovations. The members of such a subgroup may appear quite dissimilar: this is irrelevant to subgrouping. On the other hand, a set of languages may appear rather similar, and yet not form a subgroup within the family because their similarities are shared retentions from the protolanguage of the entire family. The comparative method establishes groups of related languages, and subgroups within groups, and subsubgroups within subgroups, and so on recursively. It thereby provides a chronological sequence of language splits. It typically relies on the correlation of its results with those of archaeology for absolute dating.

The second set of reasoning procedures consists of lexicostatistics and glottochronology and is due to Morris Swadesh (1952). Here similarity is king. One takes a list of perhaps 200 basic meanings and finds the words representing these in the languages to be compared. It is assumed that basic vocabulary changes at a constant rate, and that the percentage of meanings that are represented by similar words in a pair of languages is a measure of the phylogenetic relationship between them. Glottochronology builds on the constant-rate assumption and calibrates the lexicostatistics-based tree against time-depth (Peiros, Chapter 7, this volume).

Practitioners of both sets of procedures would regard these accounts as oversimplifications, and rightly so. The point, however, is that there is an underlying difference in reasoning between the two approaches, and they may generate quite different results. This fact is not infrequently overlooked by linguists (let alone practitioners of neighbouring disciplines), who attempt to combine what are in essence logically incompatible procedures.

We referred above to 'methods' and 'data' in linguistics. There is an important distinction within 'methods', however, between reasoning procedures and the tools used to apply them. This is another source of confusion. For example, a team led by Russell Gray employs wordlists to generate dated phylogenetic trees of language relationships, and there is a widespread misapprehension that this is an application of lexicostatistics and glottochronology. Its practitioners insist, however, that they are performing a computational simulation of the comparative method, identifying probable shared innovations, and using archaeological datings to calibrate time depth (Greenhill and Gray 2005; Atkinson and Gray 2006). Computation based on a collection of wordlists, typically associated with lexicostatistics, is being used as a tool in the application not of lexicostatistics but of the comparative method.

Geneticists also use two distinct sets of reasoning procedures. These deal with different data types and lead to different kinds of interpretation. Population genetics applies to gene frequencies estimated at the population level from the genetic typing of representative individuals. Population genetics theory allows one to predict the evolution of such frequencies under both neutral and selective models, by taking account of the effect of gene flow due to population migration, of genetic drift during periods of isolation, of demographic expansion or contraction, and also, in the case of selective models, of the different selective forces which may affect genetic evolution (Cavalli-Sforza and Bodmer 1999). By applying specific methods such as multivariate analyses and analyses of genetic variance, the genetic variation observed in a set of presently living populations may then be interpreted backwards in relation to their history, once significant selective effects have been ruled out (Cavalli-Sforza *et al.* 1994).

The second approach is commonly known as phylogeography (e.g. Underhill, Chapter 19, this volume). It is based on the reconstruction of molecular genealogies – phylogenetic trees – of DNA haplotypes identified in a set of individuals from different populations. Here, gene frequencies are only estimated in a second step, when groups of phylogenetically related haplotypes, known as haplogroups, are inferred from the tree. Unlike haplotypes, such haplogroups may reach significant frequencies. Then a relationship is sought between the molecular genealogy of each haplotype and the geographic area(s) where the corresponding haplogroup is at a high incidence; hence the term phylogeography.

There is a major contrast between the ways results in population genetics and phylogeography are interpreted. Molecularly remote haplotypes are sometimes found within a single population while closely related ones are observed in distantly related populations. Therefore, the genealogy of a set of observed haplotypes does not necessarily (and generally does not) reflect the genealogy of the populations represented in the study (Nei 1987). A simple manifestation of this is that different genetic markers may reveal different genealogies even though the genealogy of the populations must be unique. The link between the history of peoples (i.e. groups of individuals) and the history of genes is thus not straightforward. Genes can be transmitted paternally or maternally, or both, and may appear in a given population through migration or recurrent mutation. A consequence is that estimated dates for the common ancestors of DNA haplotypes in a given genealogy dȯ not generally correspond to the actual times of population migrations or differentiations (see section 3.1).

2. Congruence

A key assumption of the trans-disciplinary enterprise, at least with regard to linguistics and archaeology, is that results can be matched. Patterns of language distribution are, in principle, congruent with archaeology. There is some departure from congruence when a community shifts from one language to another, but this incongruence is often greatly overestimated. As Ross (Chapter 6, this volume) notes, a majority of instances of language shift during the Austronesian dispersal were associated with a shift from foraging to agriculture, resulting in an incongruence between genetics and archaeology/linguistics, but not between archaeology and linguistics.

Incongruence between archaeology and linguistics arises, superficially at least, where, for example, an Austronesian language and culture have been sinicized or papuanized. The Tsat of Hainan island speak an Austronesian language, but it has gone through two stages of transformation; first austroasiaticization as a consequence of long residence in Vietnam, and then sinicization through a millennium of bilingualism with Chinese on Hainan (Thurgood and Li 2003). The Takia of Karkar Island speak an Austronesian language with singularly Papuan grammar reflecting past bilingualism in a Trans New Guinea language (Ross 1996). From a macro-perspective, the Takia speak an Austronesian language but culturally resemble Papuan speakers. However, when one looks at the language in detail, there is no real incongruence: the Takia speak a papuanized language and have a papuanized culture.

Regrettably, the possibility of congruence between archaeology and linguistics is rejected by many archaeologists, for whom linguistics is simply a separate discipline and for whom 'the makers of the pots must remain silent'. For them a local incongruence resulting from contact appears to vitiate the broader congruence between the two disciplines. We argue, however, that since both archaeology and linguistics are direct reflections of human activities, they must, in some way, be congruent. One good reason for thinking this is that there is a clear congruence in the present; culture and language *are* clearly linked and divergences can be explained by relatively simple sociolinguistic processes.[1] The single biggest problem in linking various approaches is that within a discipline it is neither fashionable nor popular to frame hypotheses to be tested in terms of the questions asked by another discipline. So archaeologists give almost no time to matching the patterns of the cultures they delineate with historical linguistics and linguists are often uninterested in reconstructing terms and concepts that could illuminate historical hypotheses. Austronesian and Papuan scholarship constitutes an honourable exception to this in the level of cooperation between some historical linguists and archaeologists (Bellwood *et al.* 1995; Pawley *et al.* 2005).

The potential for congruence between genetics and either archaeology or linguistics is much less. The two different genetic approaches outlined above need to be considered separately. Genes are not peoples, and they have a distributional logic quite different from languages and cultures. The diffusion of a given gene does not necessarily reflect the geographic expansion of a given population or population group, nor the diffusion of a given culture or language, as it may simply represent a diffusion through population contact. Hence the tendency of phylogeographers to consider a given haplotype or haplogroup as a marker of the diffusion of a cultural complex is unfounded; like the proposal, for example, to make this link for Upper Palaeolithic cultures such as the Gravettian or Aurignacian (Semino *et al.* 2000). Rather, genes reflect extensive and complex patterns of human interaction with each other and with the environment in one-to-one and one-to-many relationships. A lesson to be learned from population geneticists is that maps of different genetic markers generally reflect geography rather than ethnicity, with gradients of allele frequencies extending over the entire world (Serre and Pääbo 2004). Global genetic discontinuities are often the result of geographic barriers (Barbujani and Belle 2006), while cultural differences may not be an obstacle to intermarriages (Blanc *et al.* 1990); cultural boundaries are more permeable to genetic exchanges.

Where geography can be eliminated as the reason for congruence, a match between genetic maps and linguistics (or archaeology) implies that cultural boundaries may also influence the extent of gene flow among populations. Such a match is sometimes detectable at the world scale (Chen *et al.* 1995; Poloni *et al.* 1997; Belle and Barbujani 2007) and is particularly meaningful when independent genetic markers converge to give similar results. However, on any large land mass, contiguous populations interact in such an intensive and complex fashion as to oblige researchers to analyse congruence in much greater depth, seeking to explain why and how cultural and genetic patterns share common histories (e.g. Karafet *et al.*, Chapter 18, this volume). But a major difficulty is discriminating between different contributing factors – e.g. between geography and linguistics – when a genetic structure corresponds with both of them. Typical examples arise when the distribution of linguistic families itself is geographically structured, as in most of Eurasia (with the exception of some isolated linguistic groups like the Basques, who live in the midst of Indo-European-speaking populations). In Africa, speakers of unrelated language phyla live adjacent to one another, for example, Khoesan and Niger-Congo in the south, Nilo-Saharan, Niger-Congo and Afroasiatic in the north-east. Therefore, while genetic structures cannot be tested against well-differentiated geographic or linguistic structures in East Asia (Sanchez-Mazas *et al.* 2005), an integrated account may prove more successful in Africa (Excoffier *et al.* 1991). In reality, many of the recent successes in genetics have more to do with geographic and demographic than with cultural parameters. For example, geneticists have emphasized the role of fragmented environments, such as islands or other isolated locales, in explaining the genetic heterogeneity observed among related populations throughout large geographic areas, like Oceania (Hagelberg, Chapter 16, this volume; Sanchez-Mazas *et al.* 2005). They have been able to demarcate possible migration routes in the first expansion of modern humans from a single origin, even though there is still much debate on their relative importance and dating (Forster and Matsumara 2005; Mellars 2006). Movements associated with cultural processes, on the other hand, have been much more difficult to isolate genetically.

A related issue often raised by geneticists is that of language diversity. Genetics can often put a quantitative measure on diversity and wonder whether this can be mapped against linguistic diversity. This seems as if it ought to work, but it does not, because languages diversify in different ways. The Australian and Trans New Guinea language areas are well-known for being highly diverse lexically and extremely uniform phonologically. Daic languages are quite uniform lexically but extremely diverse tonally. Khoesan and Nilo-Saharan languages are diverse in almost every conceivable way. Mountain *et al.* (1992) report on measures of diversity within Sinitic, but show that different categories of linguistic feature show different levels of diversity. This is not to say that diversity carries no information at all. The diversity within the Australian and Trans New Guinea regions clearly reflects their long-term settlement, but whether anything more precise can be extracted from this variety is open to question.

3. Dating

3.1. Genetic dating

Another aspect of genetics that is difficult to match to the other disciplines is dating. Absolute dates for population divergence are usually proposed on the basis of molecular phylogenies. To construct a molecular phylogeny implies that a molecular 'clock' measuring a constant speed of genetic divergence is accepted a priori. The constancy of the molecular clock is of course an approximation (Ho and Larson 2006), and is more valuable for greater time depths, except in the case of rapidly evolving genes for which recurrent mutations, or homoplasies, will be too frequent (e.g. mtDNA). Molecular clocks are usually calibrated against absolute dates for the common ancestors of humans and chimpanzees based on the fossil record. But this raises at least two major problems: (1) the lack of well-documented palaeontological evidence for these dates (Stauffer *et al.* 2001); and (2) the arbitrary choice of the 'outgroup', i.e. the ape DNA used to calibrate the tree. This second point is important because apes exhibit a much higher level of intra-specific genetic divergence than humans (Gagneux *et al.* 1999; Kaessmann *et al.* 1999). What is more, genetic dates are usually inferred with such large confidence intervals that they can easily match multiple historical or cultural events and thus satisfy any hypothesis defended *a priori* by the researcher. Then we can only take the lower and upper limits of confidence intervals as the time-frame for the events under study.

There is a further important obstacle to dating historical events through a genetic approach; phylogenies give times for the most tecent common ancestors (TMRCA) of a set of haplotypes or haplogroups, i.e. molecules. As mentioned above, genes are not peoples and there is no reason for the nodes of a phylogenetic tree (the MRCA) to correspond to identifiable events in population history, such as migrations or differentiations. In reality, genetic tree nodes are usually older than population events. For example, phylogenetic trees in genetics usually describe events that happened long before the putative origins of the language groups under discussion. This also explains why genealogies obtained for independent sets of genetic markers, like mtDNA (172,000 years, Ingman *et al.* 2000) and Y chromosome (59,000 years, Underhill *et al.* 2000) do not match (or match only thanks to the huge confidence intervals of their TMRCA): each gene has its own history. It is not surprising, therefore, that acute contradictions between published results sometimes appear. Two chapters in Bellwood and Renfrew (2002) provide a spectacular example of such a contradiction, with Oppenheimer and Richards (Chapter 22) interpreting the so-called 'Polynesian motif' in mtDNA as utterly inconsistent with the Austronesianist archaeology/linguistics consensus, and Hurles (Chapter 23) presenting an opposing view. Indeed the writings of Oppenheimer constitute a broader problem for the credibility of genetic dating for this region since his 'findings' are so completely at odds with any standard archaeological consensus (Oppenheimer 2004; Oppenheimer and Richards 2001a, 2001b).

Coming back to population genetics approaches, one cannot superpose any time scale on genetic maps of population differentiations, for the simple reason that the rate of evolution of gene frequencies is not constant. It depends on demography (rapid in small populations, slow or null in large populations). Here again, different genetic markers (when transmitted independently from one generation to the other through different chromosomes or distant regions on the same chromosome) often provide heterogeneous information on population history. They may tell stories related to different periods. Interestingly, however, plausible scenarios for the peopling history of given geographic areas may now be proposed through simulation approaches (e.g. Currat and Excoffier 2005). Basically, real genetic data are compared to virtual data obtained by simulating alternative evolutionary scenarios based on different parameters chosen a priori. These parameters include time elapsed since a population origin or differentiation, demographic parameters such as population size and migration rate, and intensity of selection for the marker under study. One or several scenarios – and hence, sets of parameters – will finally be favoured according to their greater likelihood in explaining real data. This field of research has developed recently thanks to the increasing power of computer processors. But, of course, while these methods may offer interesting applications, the arbitrary choice of different sets of parameters remains open to discussion.

3.2. Lexicostatistics and glottochronology

Lexicostatistics and glottochronology are held by most linguists to have been largely discredited, but they have undergone a major revival recently. The late Sergei Starostin (Chapter 10, this volume) devised a new series of algorithms that provided dates for the major language phyla studied by his group (including Sino-Tibetan and Altaic) and these dates are assigned to phyla on the Santa Fe website as well as in their publications. Greenberg (1987) also proposed new methods to calculate glottochronological dates. Two volumes of edited papers published by the McDonald Institute (Renfrew *et al.* 2000) consider dating issues at some length, but individual authors reach very contradictory conclusions. Ehret (2000), for example, finds the glottochronology of Bantu in harmony with his own projections of south-east African history, but without more cogent links to the findings of other archaeologists this can only be given limited credence.

To its adherents, there is something very persuasive about glottochronology, as it seems to be a magical shortcut to dates that linguists otherwise cannot provide. However, it relies on the exceedingly shaky assumption that basic vocabulary changes at a constant rate. This is so obviously false that we believe the temptations of glottochronology should be resisted. The new enthusiasts for glottochronology have one feature in common: a disdain for the hard work of trawling the archaeological literature. For example, Starostin *et al.* (2003) reconstruct a number of crop names in proto-Altaic, yet they date the break-up of Altaic some millennia prior to the inception of agriculture in this region.

4. When linguists disagree on classification

A key issue in linguistics that can be very perplexing for outsiders in the East Asian region is the matter of macrophylic schemas. A number of scholars consider that many of the language phyla of East Asia are related to one another. Unfortunately, their maps of these relationships are very diverse. The affiliation of Sino-Tibetan has been a particular problem, with a more 'conventional' view linking it with Miao-Yao or Daic, as well as wider hypotheses that bring in Caucasian or

Austronesian. Similarly, Austronesian, Austroasiatic and Daic are often linked. Indeed, some authors seem to think that all these phyla will ultimately prove to be related. When working on the problem of correlation with other disciplines it is best to retain a minimalist view; namely that while these views *may* reach a consensus among scholars in the future, at present we need to look for the interdisciplinary correlates of agreed groupings.

At least from the perspective of the comparative method, the detection of a language phylum entails a different reasoning procedure from the identification of subgroups within an already known phylum. The procedure for identifying a phylum is outlined by Nichols (2006) and entails what she calls individual-identifying evidence, i.e. formal parallels across languages that could not have arisen by chance. The parallel forms may be paradigms of affixes, lexical morphemes of three or more syllables, or collections of lexical items. There are two problems in applying this procedure in East Asia, the first of which is limited to the mainland South-east Asian region.

Languages of mainland South-east Asia are typically isolating in morphological type and tend to have monosyllabic lexical morphemes of a particular phonological type (which among other things includes tone). This makes the application of Nichols's procedure difficult, and perhaps impossible, as neither the required paradigms of affixes nor lexical morphemes of three or more syllables exist, and one is thrown back on collections of lexical items with regular sound correspondences as the only evidence of phylic relationship.

This leads to a question which is not specific to East Asia: when is such a collection of lexical items large enough to demonstrate a relationship?[2] Sagart (Chapter 5, this volume) defends the Sino-Tibetan/Austronesian macrophylum hypothesis. He has argued for this elsewhere (see his references) on the basis of lexical items with regular sound correspondences. This evidence has been criticized both for insufficiency and for considerable meaning differences among allegedly cognate items. Our purpose here is not to assert a position with regard to Sagart's claim, but to point to the level to which historical linguists do not agree, although in principle Nichols provides pointers to statistical procedures for determining whether such evidence is probative.

This problem means that even some formerly established groupings are now disputed, Altaic being the most salient example. Altaic is taken in its most extensive form to include Turkic, Mongolic, Tungusic and Korean-Japonic, which share a more or less common morphological type: they are agglutinative, verb-final and suffixing, and should offer ample material for individual-identifying evidence. The weak form of Altaic (excluding Japonic) was long held by most scholars following the work of Poppe (1965) and Miller (1996, and Chapter 11, this volume). However, the coherence of Altaic was questioned by Janhunen (1994), for whom the resemblances between its branches are a mosaic of loanwords. More recently, Starostin *et al.* (2003) have published a very large number of Altaic etymologies (over 2,000), and it has yet to be shown that these are all false, but the integrity of Altaic remains controversial, as debate in the pages of the journal *Diachronica* has shown (Georg 2004, 2005; Starostin 2005).

The tiny Korean-Japonic family is also controversial. Korean has most commonly been claimed as part of Altaic (Martin 1991a, b), although it shares many typological features with Japanese. For Janhunen (1994: 10–13) this is most probably the result of intensive interaction over an extended period rather than evidence for a genetic relationship between the languages. For Whitman and Frellesvig (Whitman 1985; Frellesvig 2001; Frellesvig and Whitman forthcoming) it is because Korean and Japonic form a family.

Languages in East Asia have often been classified together on the grounds of common morphological type, be it isolating or agglutinative, but shared morphological type (and a corresponding shared lexical organization) may result from long-standing and at times intense contact (Enfield 2003). It does not necessarily reflect linguistic phylogeny. On the other hand, if recent work by Ostapirat (2005) and Sagart (2004) is headed in the right direction (and we remain agnostic on this topic) Daic is in fact a branch of Austronesian, the phylic origins of which have been obfuscated by contact-induced change in morphological type in the better described Daic languages. Typology and phylic affiliation must be kept absolutely distinct.

The identification of macrophyla is a problem because their postulation entails great time depths. Changes in lexicon and in other language features at time depths over, say, 8,000 years are so great that the search for individual-identifying features (and the more stringent form of lexicostatistics that demands the analysis of sound correspondences) becomes impossible. In this context Nichols (1997) adds the term 'quasi-stock' to the vocabulary of historical linguistics. A quasi-stock is a grouping of well-supported groups into a larger grouping with promising markers of relatedness but with no regular sound correspondences and few clear cognates. Nichols's example of a quasi-stock is Afro-Asiatic.[3] By her rough rubric Trans New Guinea and Sino-Tibetan are also quasi-stocks. They are groupings which lie at the very limit of the reach of the comparative method. Perhaps Reid's (1999, 2005) arguments for an Austronesian/Austroasiatic nexus place it in this category, too: there is cognate bound morphology and apparently a certain amount of cognate basic vocabulary – promising, but in need of further research.

Outside Asia, Joseph Greenberg has used a different set of reasoning procedures, dubbed 'mass comparison', to identify language phyla in three areas, Africa (1963), the languages of the Andaman Islands, New Guinea and Tasmania (1971), and the Americas (1987). Despite claims that mass comparison is an application of the comparative method (e.g. Greenberg and Ruhlen 1992), it is not the linguistic comparative method as understood by most historical linguists. The initial steps of the linguistic comparative method are (1) the diagnosis of a family by individual-identifying evidence; (2) collecting sets of cognate words and affixes; and (3) working out the sound correspondences of the cognate sets (Ross and Durie 1996). Step 3 provides a validation of step 1, establishing the possible existence of a group (and a benchmark for finding the innovations that identify subgroups). It shows that present-day languages belonging to the family are descended through regular sound changes from a common protolanguage. Mass comparison conflates steps 1 and 2: resemblant words and morphemes with similar meanings are assumed to be outcomes of relatedness, but they do not reach the individual-identifying threshold and so relatedness remains undemonstrated (Ringe 1992). Worse, validation through step 3 is omitted.

Ironically, Greenberg's (1963) identification of phyla in Africa has been accepted by Africanist historical linguists, apparently because some of the materials he used could be validated by an application of step 3 (Nichols 1992: 5). This is not quite the triumph it has been represented as, because three of the four phyla he named existed *de facto* in the literature, and much of his work on internal groupings has been extensively revisited (Blench 2006). Greenberg's 1971 application of mass comparison to the 'Indo-Pacific' has drawn little attention and been largely supplanted, and his identification of American phyla has provoked a torrent of criticism of the method.

Geography can often play a role in the classification of languages; thus the proximity of two language phyla leads them to be regarded as related. For example, Daic (Thai) was long held to be related to Chinese partly because of its similar morphology, but also because its most diverse members were geographically embedded in Chinese populations. George van Driem (Chapter 9, this volume) has regularly pointed out that the classic internal structure of Sino-Tibetan (a primary branching between Sinitic and Tibeto-Burman) is not based on linguistic arguments, but rather a perception of the geographic and cultural separateness of China from the Himalayas. Even the recent conspectus of Sino-Tibetan (Thurgood and La Polla 2003) presents no arguments for the primary split of Sinitic, simply assuming it without evidence (Blench, Chapter 4, this volume). It is for linguists to insist that this is an incoherent approach. Linguistic classifications must be entirely based on linguistic arguments. It is noted above that there are good geographic and archaeological arguments for assuming the pre-Austronesians came from the Chinese mainland, but we should be methodologically wary of calling any mainland culture Austronesian without a single fragment of linguistic evidence.

5. Trees, rakes and linkages: internal classification of language phyla

Historical linguists tend to work with 'tree' models, where languages split, usually in binary fashion, and this is evidently convenient when trying to fashion a correspondence with archaeology, as a chronology can be developed. But some linguists are sceptical of these models and it is clear that languages do not always develop in such a convenient fashion. There are at least two major sources of 'inconvenient' patterns.

One is that, at a certain stage in the history of a family, a language may diversify into a dialect network, some of the dialects of which become geographically isolated through their speakers' emigration and develop into distinct subgroups of languages, whilst the stay-at-home dialects continue both to diversify and to interact with each other, acquiring a pattern of overlapping innovations but having no exclusively shared innovation that identify them as a subgroup.

The other source of inconvenient patterns is contact with languages that belong to another phylum or a more or less distantly related part of the same phylum. As we noted above, it is likely that the common pattern of mainland East Asian languages with reduced morphology, complex tones and simplified word structures represents massive convergence between different language phyla. However, the fine-grained description of contact in a particular language or group of languages can tell us a good deal about the culture history of its speakers and help us to correlate linguistic and archaeological findings.

It is difficult to work with non-trees, because they present no sense of chronology, and it is for this reason that a distinction is made in Austronesian linguistics between innovation-defined subgroups like Oceanic (the conventional subgroups of the comparative method) and innovation-linked subgroups or linkages like Western Malayo-Polynesian, the member languages of which are linked by overlapping patterns of innovations.[4] Both can be incorporated into a tree of sorts (see Ross, Chapter 6, this volume).

6. Sampling frames in genetics

An important but little-discussed aspect of the methodology of genetics is the targeting of sample collection. The hard-science aspect of genetics has often blinded journal referees to the highly unscientific nature of the samples that are analysed. Thus we can find 'West or South Africans' compared to 'Caucasians', the latter term being simply a euphemism for 'white race', a concept that is nowadays unacceptable. Even now, many studies depend on 'out of the freezer' materials, often exchanged between laboratories, where samples really collected within a serious anthropological or ethnolinguistic framework are often a minority. Moreover, sample sizes are generally too low to be statistically representative, and this crucial issue – the ABC of population genetics – is almost never addressed. The problem is all the more serious now that DNA typing technologies allow us to define haplotypes at a much more precise level than before, such that the number of detectable haplotypes is always much higher than the number of individuals sampled (which was not the case with studies based on blood groups or proteins). If we are really to solve some of the major problems of correlating genes and language, then what is required is targeted sampling; i.e. collecting samples that are statistically valid and reflect closely the particular groups that are the focus of the study. It is thus unacceptable to make claims about – to take an extreme but common example – 'Africans' (versus 'non-Africans', e.g. Yu *et al.* 2002) when in fact a handful of population samples are supposed to represent the huge diversity of African ethnic and linguistic groups (not to mention the nonsensicality of 'non-Africans'). Ethnolinguistically targeted sample collections, such as the Taiwanese (Sanchez-Mazas *et al.*, Chapter 13, and Trejaut *et al.*, Chapter 14, this volume) or those planned within the framework of the Languages and Genes of the Greater Himalayan Region project headed by George van Driem and co-workers, are presently under way and more coherent results may emerge within a few years.

7. Local factors that may confuse results

7.1. Teleology in archaeology

Linguistics and archaeology are not driven by the spirit of pure enquiry; archaeology in particular is often prone to hijacking by nationalist agendas. This is not a new point, but the development of the nation state in the 20th century has resulted in a bizarre framing of accounts of the past in terms of the boundaries of the present. It encourages archaeological accounts to view the horizons of the past as leading inexorably towards those of the present. Typically, in China, ancient cultures become precursors of the Han state, rather than, perhaps, dead ends.[5] This is persuasive but misleading: most of what we know about Sinitic suggests that the Han expansion is quite recent and therefore almost any older archaeological culture is *not* likely to associated with Sinitic speakers.

7.2. Confusion associated with written texts

The reconstruction of some parts of Sino-Tibetan has been confused by the existence of archaic written texts. Much historical scholarship has gone into the reconstruction of Old Chinese, a language that would consistently account for the system of ancient texts. But there is, and can be, no evidence that such a language was ever spoken, and no necessary link with proto-Sinitic, a language reconstructed from the wide range of modern dialects. Similar problems have arisen by confusing Sanskrit with proto-Indo-Aryan, as Turner (1966) does in his magisterial volumes. Probably if we had a better reconstruction of proto-Sinitic, there would fewer problems about its place within the larger Sino-Tibetan schema.

8. Conclusion

The collection of methodological problems raised in this introductory chapter may seem to give a rather negative impression of the interdisciplinary enterprise. But if this were our thinking we would not have put together this volume. Rather our purpose has been to enter some caveats about simplistic models of congruence and to help each disciplinary specialist to be aware of the pitfalls of reading literature outside their immediate ambit (and sometimes even within it). But demonstrations that researchers are increasingly becoming aware of the need to read around their subject in a geographical and historical frame are found in recent publications: see, for example, Pawley (2002) on the Austronesian dispersal or the interdisciplinary collection edited by Pawley *et al.* (2005) on the Papuan peoples.

Notes

[1] English is the most intensively studied language in the world, and recent explorations of its varieties make it perfectly possible to account for both variation and the congruence or otherwise of the cultures of those who speak it.

2 As Starostin (this volume) notes, regular sound correspondences, preferably in basic vocabulary, are essential if shared inheritance is to be demonstrated and the possibility of borrowing eliminated.

3 This is also an example of the curiously inconsistent way outsiders evaluate evidence for the existence of particular phyla. Compared with Trans New Guinea, Afro-Asiatic has hundreds of proposed etymologies, and some well-established and distinctive phonological and morphological features.

4 There are other ways in which a linkage may come into being, but detailed discussion would require at least a paper to itself.

5 It is interesting to compare these with Stephen J. Gould's strictures on models of evolution that are structured so as they always finish with the evolution of modern humans, rather than being full of byways and forking paths that lead nowhere.

References

Atkinson, Q.D. and Gray, R.D. (2006) 'How old is the Indo-European language family? Progress or more moths to the flame?', in P. Forster and C. Renfrew (eds) *Phylogenetic Methods and the Prehistory of Languages*, pp. 91–110, Cambridge: McDonald Institute for Archaeological Research.

Barbujani, G. and Belle, E.M. (2006) 'Genomic boundaries between human populations', *Human Heredity*, 61: 15–21.

Belle, E.M. and Barbujani, G. (2007) 'Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity', *American Journal of Physical Anthropology*, 133 [Epub 16 May 2007].

Bellwood, P. (1979) *Man's Conquest of the Pacific*, New York, NY: Oxford University Press.

Bellwood, P. and Renfrew, C. (eds) (2002) *Examining the Farming/Language Dispersal Hypothesis*, Cambridge: McDonald Institute of Archaeological Research.

Bellwood, P., Fox, J. and Tryon, D. (eds) (1995) *The Austronesians: Historical and Comparative Perspectives*, Canberra: Department of Anthropology, Research School of Pacific and Asian Studies, Australian National University.

Blanc, M., Sanchez-Mazas, A., Hubert van Blyenburgh, N., Sevin, A., Pison, G. and Langaney, A. (1990) 'Inter-ethnic genetic differentiation: Gm polymorphism in eastern Senegal', *American Journal of Human Genetics*, 46: 383–92.

Blench, R.M. (2006) *Archaeology, Language and the African Past*, Lanham, MD: Altamira Press.

Cavalli-Sforza, L.L. and Bodmer, W.F. (1999) *The Genetics of Human Populations*, Mineola, NY: Dover Publications.

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

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

Currat, M. and Excoffier, L. (2005) 'The effect of the Neolithic expansion on European molecular diversity', *Proceedings: Biological Science*, 272: 679–88.

Ehret, C. (2000) 'Testing the expectations of glottochronology against the correlations of language and archaeology in Africa', in C. Renfrew, A. McMahon and L. Trask (eds) *Time Depth in Historical Linguistics*, 2 vols, Cambridge: McDonald Institute for Archaeological Research.

Enfield, N.J. (2003) *Linguistic Epidemiology: Semantics and Grammar of Language Contact in Mainland Southeast Asia*, London and New York, NY: RoutledgeCurzon.

Excoffier, L., Harding, R.M., Sokal, R.R., Pellegrini, B. and Sanchez-Mazas, A. (1991) 'Spatial differentiation of RH and GM haplotype frequencies in Sub-Saharan Africa and its relation to linguistic affinities', *Human Biology*, 63: 273–307.

Forster, P. and Matsumura, S. (2005) 'Did early humans go north or south?', *Science*, 308: 965–6.

Frellesvig, B. (2001) 'A common Korean and Japanese copula', *Journal of East Asian Linguistics*, 10: 1–35.

Frellesvig, B. and Whitman, J. (forthcoming) 'The Japanese-Korean vowel correspondences', in M. Endo Simon and P. Sells (eds) *Japanese/Korean Linguistics*, Stanford, CA: CSLI.

Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P.A., Boesch, C., Fruth, B., Hohmann, G., Ryder, O.A. and Woodruff, D.S. (1999) 'Mitochondrial sequences show diverse evolutionary histories of African hominoids', *Proceedings of the National Academy of Science USA*, 96: 5077–82.

Georg, S. (2004) Review of Starostin *et al.* (2003), *Diachronica*, 21: 445–50.

Georg, S. (2005) Reply to Starostin (2005), *Diachronica*, 22: 455–7.

Greenberg, J.H. (1963) *The Languages of Africa*, Bloomington, IN: Indiana University Press.

Greenberg, J.H. (1971) 'The Indo-Pacific hypothesis', in T.A. Sebeok (ed.) *Current Trends in Linguistics*, 8: *Linguistics in Oceania*, pp. 807–71, The Hague: Mouton.

Greenberg, J.H. (1987) *Language in the Americas*, Stanford, CA: Stanford University Press.

Greenberg, J.H. and Ruhlen, M. (1992) 'Linguistic origins of native Americans', *Scientific American* (Nov.): 60–5.

Greenhill, S.J. and Gray, R.D. (2005) 'Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees, and Austronesian languages', in R. Mace, C.J. Holden and S. Shennan (eds) *The Evolution of Cultural Diversity: Phylogenetic Approaches*, pp. 31–52, London: UCL Press.

Ho, S.Y. and Larson, G. (2006) 'Molecular clocks: When times are a-changin'', *Trends in Genetics*, 22: 79–83.

Hurles, M. (2002) 'Can the hypothesis of language/agriculture co-dispersal be tested with archaeogenetics?', in P. Bellwood and C. Renfrew (eds) *Examining the Farming/Language Dispersal Hypothesis*, pp. 299–309, Cambridge: McDonald Institute of Archaeological Research, ch. 23.

Ingman M., Kaessmann, H., Paabo, S. and Gyllensten, U. (2000) 'Mitochondrial genome variation and the origin of modern humans', *Nature*, 408: 708–13.

Janhunen, J. (1994) 'Additional notes on Japanese and Altaic', *Journal de la Société Finno-Ougrienne*, 85: 236–40, 256–60.

Kaessmann, H., Wiebe, V. and Paabo, S. (1999) 'Extensive nuclear DNA sequence diversity among chimpanzees', *Science*, 286: 1159–62.

Martin, S.E. (1991a) 'Morphological clues to the relationships of Japanese and Korean', in P. Baldi (ed.) *Patterns of Change, Change of Patterns: Linguistic Change and Reconstruction Methodology*, pp. 235–62, Berlin: Mouton de Gruyter.

Martin, S.E. (1991b) 'Recent research on the relationships of Japanese and Korean', in S. Lamb (ed.) *Sprung from Some Common Source*, pp. 269–92, Stanford, CA: Stanford University Press.

Mellars P. (2006) 'Going east: New genetic and archaeological perspectives on the modern human colonization of Eurasia', *Science*, 313: 796–800.

Miller, R.A. (1996) *Languages and History: Japanese, Korean and Altaic*, Bangkok: White Orchid Press.

Mountain, J.K., Wang, W.S.-Y., Du, R., Yuan, Y. and Cavalli-Sforza, L.L. (1992) 'Congruence of genetic and linguistic evolution in China', *Journal of Chinese Linguistics*, 20: 315–30.

Nei, M. (1987) *Molecular Evolutionary Genetics*, New York, NY: Columbia University Press.

Nichols, J. (1992) *Linguistic Diversity in Space and Time*, Chicago, IL: University of Chicago Press.

Nichols, J. (1997) 'Modeling ancient population structures and movement in linguistics', *Annual Review of Anthropology*, 26: 359–84.

Nichols, J. (2006) 'The comparative method as heuristic', in M. Durie and M. Ross (eds) *The Comparative Method Reviewed: Irregularity and Regularity in Linguistic Change*, pp. 39–71. New York: Oxford University Press.

Oppenheimer, S.J. (2004) 'The "Express Train from Taiwan to Polynesia"; on the congruence of proxy lines of evidence', *World Archaeology*, 36: 591–600.

Oppenheimer, S.J. and Richards, M. (2001a) 'Slow boat to Melanesia?', *Nature*, 410: 166–7.

Oppenheimer, S.J. and Richards, M. (2001b) 'Fast trains, slow boats, and the ancestry of the Polynesian islanders', *Science Progress*, 84: 157–81.

Oppenheimer, S.J. and Richards, M. (2002) 'Polynesians: Devolved Taiwanese rice farmers or Wallacean maritime traders with fishing, foraging and horticultural skills?', in P. Bellwood and C. Renfrew (eds) *Examining the Farming/Language Dispersal Hypothesis*, pp. 287–97, Cambridge: McDonald Institute of Archaeological Research.

Ostapirat, W. (2005) 'Kra-Dai and Austronesian: Notes on phonological correspondences and vocabulary distribution', in L. Sagart, R. Blench and A. Sanchez-Mazas (eds) *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, London: RoutledgeCurzon.

Pawley, A.K. (2002) 'The Austronesian dispersal: Languages, technologies and people', in P. Bellwood and C. Renfrew (eds) *Examining the Farming/Language Dispersal Hypothesis*, Cambridge: McDonald Institute of Archaeological Research.

Pawley, A.K., Attenborough, R., Golson, J. and Hide, R. (eds) (2005) *Papuan Pasts: Cultural, Linguistic and Biological Histories of Papuan-speaking Peoples*, PL 572, Canberra: ANU.

Poloni, E.S., Semino, O., Passarino, G., Santachiara-Benerecetti, A.S., Dupanloup, I., Langaney, A. and Excoffier, L. (1997) 'Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics', *American Journal of Human Genetics*, 61: 1015–35.

Poppe, N.N. (1965) *Introduction to Altaic linguistics*, Wiesbaden: Otto Harrassowitz.

Reid, L.A. (1999) 'New linguistic evidence for the Austric hypothesis', in E. Zeitoun and P.J. Li (eds) *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, pp. 1–30, Taipei: Institute of Linguistics (Preparatory Office), Academia Sinica.

Reid, LA. (2005) 'The current status of Austric: A review and evaluation of the lexical and morphosyntactic evidence', in L. Sagart, R. Blench and A. Sanchez-Mazas (eds) *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, pp. 132–60, London: RoutledgeCurzon.

Renfrew, C., McMahon, A. and Trask, L. (eds) (2000) *Time Depth in Historical Linguistics*, 2 vols, Cambridge: McDonald Institute for Archaeological Research.

Ringe, D.A. (1992) 'On calculating the factor of chance in language comparison', *Transactions of the American Philosophical Society*, 82: 1–110.

Ross, M. (1996) 'Contact-induced change and the comparative method: cases from Papua New Guinea', in M. Durie and M. Ross (eds) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, pp. 180–217, New York, NY: Oxford University Press.

Ross, M. and Durie, M. (1996) 'Introduction', in M. Durie and M. Ross (eds) *The Comparative Method Reviewed: Regularity and Irregularity in Language Change*, pp. 3–38, New York, NY: Oxford University Press.

Sagart, L. (2004) 'The higher phylogeny of Austronesian and the position of Tai-Kadai', *Oceanic Linguistics*, 43: 411–44.

Sanchez-Mazas A., Poloni, E.S., Jacques, G. and Sagart, L. (2005) 'HLA genetic diversity and linguistic variation in East Asia', in L. Sagart, R. Blench and A. Sanchez-Mazas (eds) *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, pp. 273–96, London and New York, NY: RoutledgeCurzon.

Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Arbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P., Kouvatsi, A., Limborska, S., Marcikiae, M., Mika, A., Mika, B., Primorac, D., Santachiara-Benerecetti, A.S., Cavalli-Sforza, L.L. and Underhill, P.A. (2000) 'The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective', *Science*, 290: 1155–9.

Serre, D. and Pääbo, S. (2004) 'Evidence for gradients of human genetic diversity within and among continents', *Genome Research*, 14: 1679–85.

Spriggs, M. (1989) 'The dating of the Island Southeast Asian Neolithic: an attempt at chronometric hygiene and linguistic correlation', *Antiquity*, 63: 587–613.

Starostin, S. (2005) 'Response to Stefan Georg's review of the Etymological Dictionary of the Altaic Languages', *Diachronica*, 22: 451–4.

Starostin, S.A., Dybo, A. and Mudrak, O. (2003) *Etymological Dictionary of the Altaic Languages*, 3 vols, Leiden: Brill.

Stauffer, R.L., Walker, A., Ryder, O.A., Lyons-Weiler, M. and Hedges, S.B. (2001) 'Human and ape molecular clocks and constraints on paleontological hypotheses', *Journal of Heredity*, 92: 469–74.

Swadesh, M. (1952) 'Lexicostatistic dating of prehistoric ethnic contacts', *Proceedings of the American Philosophical Society*, 96: 453–62.

Thurgood, G. and LaPolla, R.J. (eds) (2003) *The Sino-Tibetan Languages*. London and New York, NY: Routledge.

Thurgood, G. and Li, F. (2003) 'Contact induced variation and syntactic change in the Tsat of Hainan', in D. Bradley, R. LaPolla, B. Michailovsky and G. Thurgood (eds) *Language Variation: Papers on Variation and Change in the Sinosphere and in the Indosphere in honour of James A. Matisoff*, Canberra: Pacific Linguistics.

Turner, R.L. (1966) *A Comparative Dictionary of the Indo-Aryan Languages*, London: Oxford University Press.

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J.R., Mehdi, S.Q., Seielstad, M.T., Wells, R.S., Piazza, A., Davis, R.W., Feldman, M.W., Cavalli-Sforza, L.L. and Oefner, P.J. (2000) 'Y chromosome sequence variation and the history of human populations', *Nature Genetics*, 26: 358–61.

Whitman, J.B. (1985) 'The phonological basis for the comparison of Japanese and Korean', Ph.D. dissertation, Harvard University.

Yu, N., Fu, Y.X. and Li, W.H. (2002) 'DNA polymorphism in a worldwide sample of human X chromosomes', *Molecular Biology and Evolution*, 19: 2131–41.