Comparing linguistic and genetic relationships among east asian populations: a study of the Rh and GM polymorphisms

Estella Poloni, Alicia Sanchez-Mazas, Guillaume Jacques, Laurent Sagart

CHAPTER 15

COMPARING LINGUISTIC AND GENETIC RELATIONSHIPS AMONG EAST ASIAN POPULATIONS: A STUDY OF THE RH AND GM POLYMORPHISMS

E. S. Poloni, A. Sanchez-Mazas, G. Jacques, L. Sagart

6591 words

INTRODUCTION

According to palaeoanthropological and archaeological records, East Asia is probably one of the earliest regions settled by our species, *Homo sapiens sapiens*, after Africa and the Middle East (Lahr and Foley 1994, 1998). Research in this region of the world should thus provide important clues about the history of our species. Moreover, documenting the genetic diversity of East Asian populations is a crucial step in understanding the settlement history of such regions as Japan, insular Southeast Asia and Oceania, as well as the American continent.

Continental East Asian populations have recently attracted the attention of molecular anthropologists, as attested by the numerous studies on variation of molecular markers in these populations published during the last four or five years (e.g. Chu *et al.* 1998; Su *et al.* 1999; Ding *et al.* 2000; Karafet *et al.* 2001; Ke *et al.* 2001; Oota *et al.* 2002; Yao *et al.* 2002a). These studies have provided contradictory results and lead to discrepancies in the interpretation of the genetic history of East Asian populations. There may be several reasons that explain this,

including differential or restricted sampling of populations, but the most important is that each independent component of our genome has its own specific evolutionary history. For instance, gender-specific polymorphisms, such as those studied on the mitochondrial genome and the Y chromosome, have revealed the impact of differential migratory behaviour of men and women on the genetic structure of populations (Poloni *et al.* 1997, Seielstadt *et al.* 1998, Oota *et al.* 2001a). Thus, several polymorphic systems must be analysed if one aims at drawing more conclusive inferences about the genetic history of populations in East Asia.

Continental East Asia is also home to much cultural diversity, as attested among other traits by the number of distinct language families that coexist there. However, the relationships between this linguistic diversity and the genetic variability of East Asian populations are only starting to be investigated (Su *et al.* 2000). This study analyses the genetic structure of East Asian populations with an emphasis on the linguistic classification of these populations, i.e. the classification of their languages into the great East Asian language families. It is part of an ongoing project to analyse multiple genetic systems. As a contribution to the investigation of the evolutionary information held by each specific component of the genome, we present here the results of the analysis of two serological markers, the Rhesus (RH) and GM polymorphisms, which have been extensively tested in East Asian populations. A companion paper in this volume (Sanchez-Mazas *et al.*) investigates the genetic variation of HLA molecular alleles in East Asia.

The results based on the variability of the RH and GM systems indicate that both linguistic classification and geographic proximity explain a significant proportion of the genetic affinities observed among East Asian populations. At present, we interpret these results by suggesting the existence of a commonality in the history of genetic differentiation and linguistic diversification of East Asian populations and language families, with occurrences of strong genetic contacts across linguistic borders.

MATERIALS

The choice of the RH and GM genetic systems, two classical markers[1], is motivated by the fact that numerous samples drawn from populations of distinct geographic locations in East Asia have been tested over the years, providing a large body of data. The RH system consists of specific antigens expressed on the surface of the red cell and encoded in a set of genes on human chromosome 1. The GM system consists of antigens (allotypes) encoded in a set of genes on chromosome 14 and expressed on specific immunoglobulins (IgG class) circulating in the serum. The RH system comprises eight genetic variants ('haplotypes'), with variable frequencies among human populations; the GM system is more polymorphic in that it comprises more haplotypes, nine of which represent the vast majority of the human polymorphism (Steinberg and Cook 1981; Sanchez-Mazas 1990; Dugoujon *et al.* forthcoming).

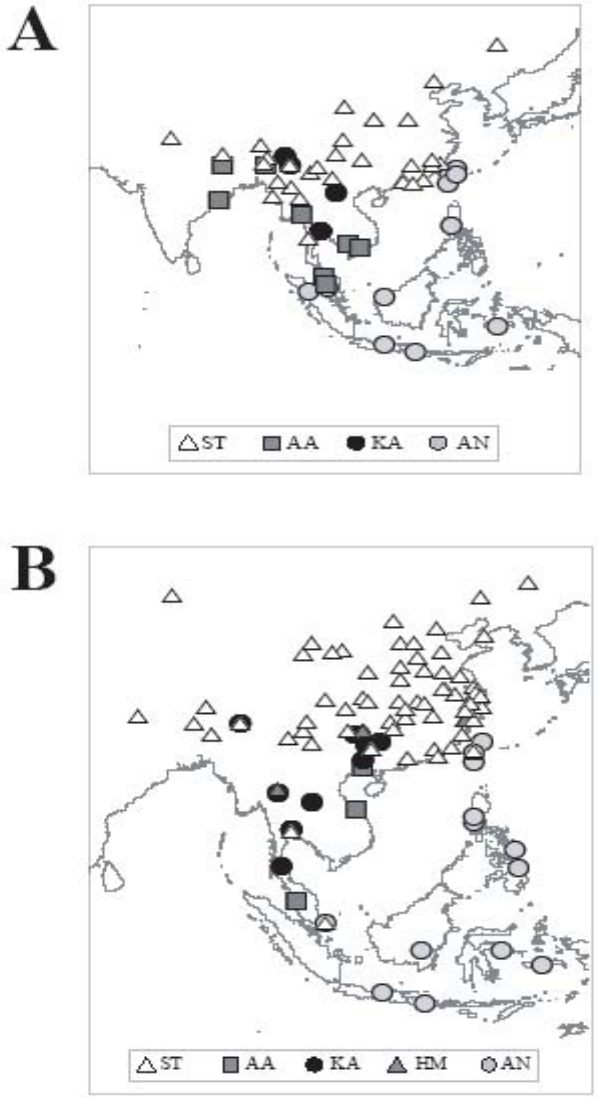

Figure 15.1

A: Geographic location of 61 population samples tested for RH polymorphism. Samples symbols correspond to linguistic families (ST: Sino-Tibetan; AA: Austroasiatic; KA: Tai-Kadai; NCA: North Caucasian; AA: Austronesian). B: Geographic location of 102 population samples tested for GM polymorphism. Samples symbols correspond to linguistic families (ST: Sino-Tibetan; AA: Austroasiatic; KA: Tai-Kadai; HM: Hmong-Mien; AA: Austronesian).

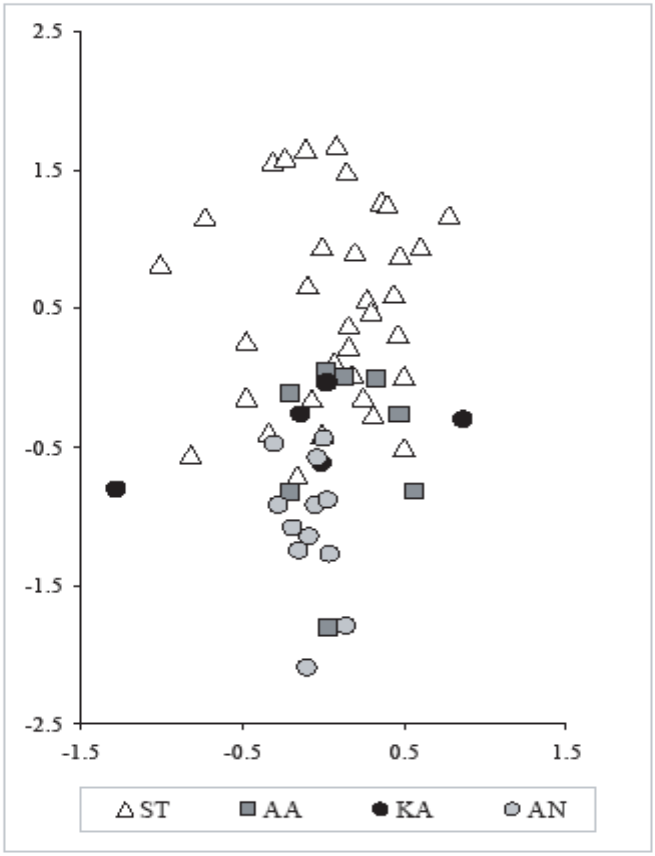

FIGURE 15.2

MDS of Reynolds et al. (1983) genetic distances among 61 population samples computed on RH frequency distributions. The goodness-of-fit of the 2-dimensional projection to the original configuration is fair (stress value = 0.160). Samples symbols as in Figure 15.1A.

The selection of samples was based on linguistic criteria (Table 15.1). We focused on populations whose languages belong to the Sino-Tibetan family and its southern neighbours from mainland and insular South-East Asia approximately down to Kalimantan: the Austroasiatic, Tai-Kadai, Hmong-Mien (only for GM, not available for RH) and Austronesian families (Figures 15.1A and 15.1B). Thus, we did not consider populations north of Sino-Tibetan, e.g. Altaic, Japanese and Korean. Overall, the analyses of the RH and GM genetic systems rely upon 10,972 and 15,437 individuals respectively (Table 15.1). All the genetic data used are included in the *GeneVa* databank (maintained by ASM in Geneva) and have been checked for reliability of gene frequencies.

Table 15.1. Representation of the linguistic families by numbers of population samples (and numbers of individuals) in the analyses.

	RH	GM
Austronesian	12 (2,222)	14 (3,515)
Austroasiatic	9 (1,165)	4 (944)
Tai-Kadai	6 (1,004)	11 (1,548)
Hmong-Mien	- -	3 (345)

Sino-Tibetan	34 (6,581)	70 (9,555)
Total	61 (10,972)	102 (15,437)

Statistical analyses were performed using Arlequin ver. 2.0 (Schneider *et al.* 2000) and NTSYSpc ver. 2.1 (Rohlf 1998) software; great-circle distances between geographic localities were computed by means of a local program (N. Ray, personal communication). For the sake of clarity, the analyses are described in the relevant results sections.

RESULTS

Genetic landscapes of the RH and GM polymorphisms in East Asia

In East Asia, the RH genetic landscape is mainly characterized by a high frequency (>50%) of haplotype R^1 in all populations, concomitant with substantial frequencies of haplotype R^2 and, to a lesser extent, of haplotype R^0 (Plate Va). Actually, the frequency of R^1 increases and that of R^2 decreases as one moves from the north to the south of the continent. The pattern of frequency distributions for the GM system is more diversified, in that more haplotypes are observed at polymorphic frequencies in the populations, especially in North-East Asia (plate Vb). In this region, four variants are present at substantial frequencies, i.e. GM∗1,3;5*, GM∗1,17;21, GM∗1,2,17;21, and GM∗1,17;10,11,13,15,16. When one moves from north to south, the populations become less diversified because of an increase in frequency of haplotype GM∗1,3;5*, concomitant with a decrease

in frequency of the other three common variants. Thus, both genetic systems display a pattern of continuity in variation of the frequency distributions along a north-to-south axis, with no abrupt changes.

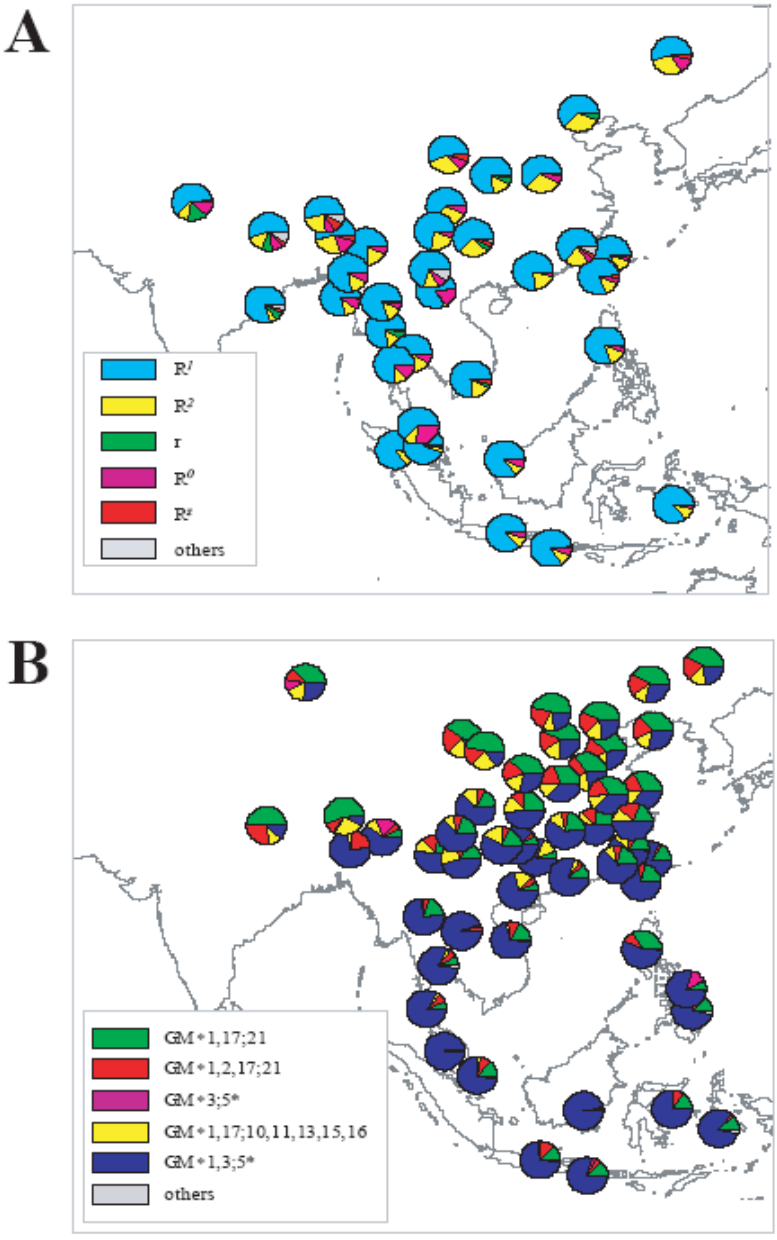

PLATE V (a) Rh frequency distributions (due to sampling density, only 37 samples are represented). (b) GM frequency distributions (due to dampling density 51 samples are represented).

Patterns of genetic affinities among populations

Genetic distances between population pairs were calculated as Reynolds *et al.* (1983) coancestry coefficients based on pairwise F_{ST} statistics estimated from the haplotype frequencies in the samples. The F_{ST} index expresses the proportion of the total genetic variability that is attributable to differences between two populations (the remainder being explained by differences among individuals within the populations). Multivariate analyses of these genetic distances were performed in order to study the patterns of genetic relationships among populations inferred from each genetic system. We used non-metric multidimensional-scaling (MDS) to obtain a graphic projection of the populations on a two-dimensional space in which the distances between the projected points bear a monotone relationship to the original genetic distances between the populations.

FIGURE 15.2

In the resulting MDS on RH data, no clear clustering of the samples is evidenced: the populations tend to group together according to their linguistic affiliation but without any discontinuity between groups (Figure 15.2). Indeed, substantial overlapping of these linguistically defined groups is readily observable, especially for Austroasiatic and Tai-Kadai. A similar pattern of genetic affinities among populations is observed for the GM system (Figure 15.3), with even higher

overlapping among the southern groups (i.e. Austroasiatic, Tai-Kadai, Hmong-Mien and Austronesian). The relationships among Sino-Tibetan populations are further analysed below.

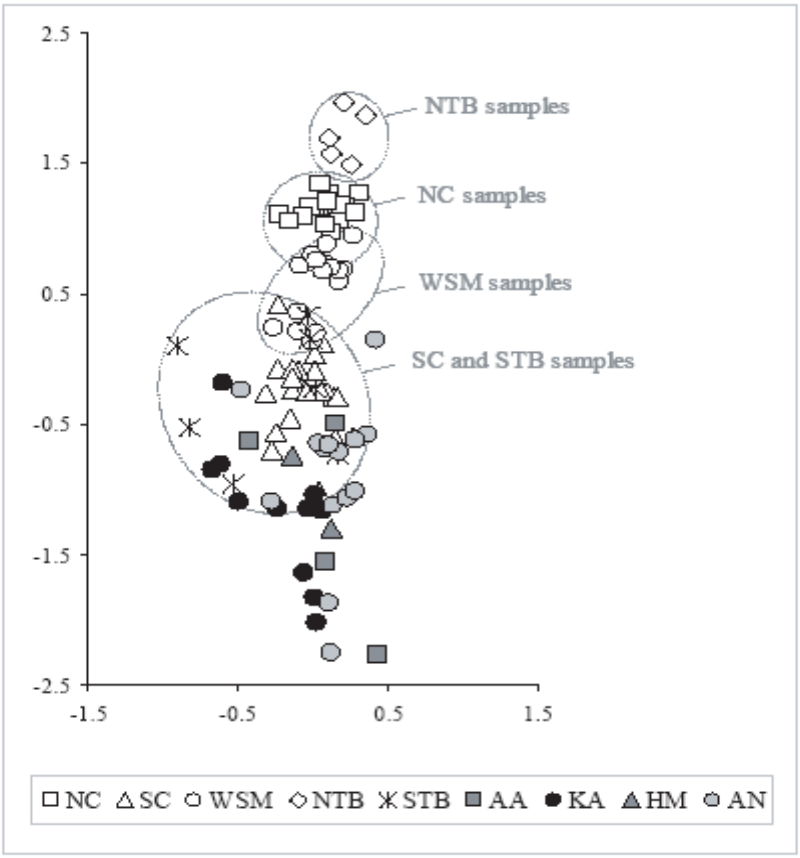
2.5
1.5
0.5
-0.5
-1.5
-2.5
-1.5
-0.5
0.5
1.5
NTB samples
NC samples
WSM samples
SC and STB samples
NC SC WSM NTB STB AA KA HM AN

FIGURE 15.3

MDS of Reynolds et al. (1983) genetic distances among 102 population samples computed on GM frequency distributions. The goodness-of-fit of the 2-dimensional projection to the original configuration is good (stress value =0.085). Samples symbols as in Figure 15.2A, except that Sino-Tibetan (ST) samples are further subdivided into: NC: Northern Chinese (all Mandarin but Southeastern); SC: Southern Chinese (Min, Xiang, Gan, Hakka, Min and Yue); WSM: Wu and Southeastern Mandarin; NTB: Northern Tibeto-Burman; STB: Southern Tibeto-Burman (see text).

Levels of population genetic structure

The level of genetic differentiation in a set of populations, referred to as the level of population genetic structure, can also be estimated from an F_{ST} statistic. In this case, this statistic expresses the proportion of the total genetic variability attributable to differences between all the populations (the remainder being explained by differences between individuals within these populations). The observed levels of genetic differentiation between populations are significant both for the RH and GM systems (Table 15.2). The structure is stronger for GM, with ~14% of the total genetic variance being explained by differences between populations, versus ~4% for the RH system.

Table 15.2. Proportion of the total genetic variation that is due to differences between populations.

	Percent of total genetic variance explained by differences:	
	between populations	among individuals within populations
RH	3.9*	96.1*
GM	14.3*	85.7*

Significance level: * $P < 0.005$.

The level of population structure within each of the linguistic groups represented in the data is also significant for both systems (Table 15.3). In all groups, these levels are always higher for GM than for RH, but for both systems the highest F_{ST} values are observed in the Austroasiatic group, indicating a substantial level of genetic differentiation among Austroasiatic populations.

Table 15.3. Levels of genetic structure among populations within linguistic groups and mean expected heterozygosity in linguistic groups.

	RH			GM		
	Group size[a]	F_{ST}[b]	*h* (s.d.)[c]	Group size[a]	F_{ST}[b]	*h* (s.d.)[c]
Austronesian	12	0.7*	0.26 (.07)	14	4.3*	0.37 (.15)
Austroasiatic	9	3.5*	0.36 (.10)	4	13.6*	0.26 (.23)
Tai-Kadai	6	1.7*	0.36 (.05)	11	4.9*	0.28 (.11)
Hmong-mien				3	2.7*	0.43 (.21)
Sino-Tibetan	34	2.9*	0.47 (.09)	70	8.1*	0.63 (.10)

[a] Number of populations per linguistic group (see Table 1); [b] Expressed as the percent of total genetic variation due to differences among populations of the linguistic group; [c] Gene diversity

(standard deviation) averaged over populations in the linguistic group. Significance level: * $P <$ 0.005.

In Table 15.3, these levels of genetic structure in the linguistically-defined groups are contrasted with a measure of the degree of genetic variability among individuals within the populations, i.e. gene diversity (*h*) averaged over the populations in each linguistic group. For both systems, we observe more intra-population variability in the Sino-Tibetan group than in the other East Asian groups (although the large standard deviations associated with these measures indicate that the differences between the groups are not substantial).

In summary, Sino-Tibetan populations display comparatively high levels of both genetic differentiation and internal diversity. At the opposite, Austronesian and Tai-Kadai populations are both less differentiated and more homogeneous. Austroasiatic populations also display a rather low level of internal diversity, but they are substantially differentiated. By contrast, Hmong-Mien populations are found to be quite heterogeneous and only slightly differentiated, but this group has to be regarded with caution as it is only represented by three samples, one of which (She) was drawn from an almost completely sinicized population (i.e. Hakka speakers).

Genetic and linguistic affinities among populations

A two-level hierarchical ANOVA was used to further investigate whether the genetic structure inferred from both polymorphisms can be related to linguistic

classification (Table 15.4). The populations are first assigned to distinct groups, and the analysis performs a partition of the total genetic variability into three components, i.e. one due to differences between groups of populations, another due to differences between populations within groups, and a third due to differences among individuals within the populations. The groups are defined as Austronesian, Austroasiatic, Tai-Kadai and Sino-Tibetan, plus Hmong-Mien for GM.

Table 15.4. Proportion of the total genetic variation that is due to differences between linguistic groups, and between populations within linguistic groups.

		Percent of total genetic variance explained by differences:		
	N[a]	between groups	among populations within groups	among individuals within populations
RH	4[b]	2.2*	2.6*	95.2*
GM	5[c]	12.1*	6.8*	81.2*

[a] N : number of linguistic groups. [b] The four groups are: Austronesian, Austroasiatic, Tai-Kadai and Sino-Tibetan. [c] The five groups are: Austronesian, Austroasiatic, Tai-Kadai, Hmong-mien and Sino-Tibetan. Significance level: * $P < 0.005$.

The results for the GM system do indeed suggest a correspondence between the genetic structure of the populations and linguistic groupings. We observe almost twice as much genetic variability between linguistic groups (~12%) as between

populations within the linguistic groups (~7%). This correspondence does not apply to the RH system, as the observed level of genetic variability between linguistic groups is comparable to that within those groups (both < 3%).

A high level of genetic structure can arise from just a few diverging populations. To determine which linguistically-defined population groups are differentiated from others we performed two-level hierarchical ANOVAs on pairs of groups (Table 15.5). The analyses of RH data indicate that almost all groups are significantly differentiated, but levels of divergence are rather low. Indeed, in most cases, differentiation levels observed between the linguistically-defined groups are lower than those among populations within these groups, with the notable exception of the significant divergence between Austronesian and Sino-Tibetan.

Table 15.5. Proportion of the total genetic variation[a] that is due to differences between linguistic groups compared two by two. Above diagonal: RH system, below diagonal: GM system

	Austronesian	Austroasiatic	Tai-Kadai	Hmong-Mien	Sino-Tibetan
Austronesian		0.9***	1.0***	- -	4.0***
Austroasiatic	n.s.		n.s.	- -	1.1**
Tai-Kadai	1.3***	n.s.		- -	0.9*
Hmong-Mien	n.s.	n.s.	n.s.		- -

Sino-Tibetan	11.5***	13.9***	15.2***	9.6***

[a] This proportion is underlined when it is superior to the proportion of genetic variance explained by differences between populations within linguistic groups. Significance level: n.s. not significant at the 5% level, * $0.05 > P > 0.01$, ** $0.01 > P > 0.005$, *** $P < 0.005$.

In contrast, with the GM data, the Sino-Tibetan group is highly and significantly differentiated from all other groups (with values of the "between groups" component > 9%), whereas the latter are mostly undifferentiated between them. Among Southeast Asian groups, intra-group divergence levels are always higher than inter-group levels. Thus, by this analysis, the high level of genetic structure of the GM system (~12%, Table 15.4) is mainly attributable to the differentiation of Sino-Tibetan from all other linguistically-defined groups.

However, this result is challenged by the MDS analysis on GM data (Figure 15.3), which does not reveal a clear clustering of Sino-Tibetan populations. Rather, the MDS suggests some degree of genetic structure within the Sino-Tibetan group itself. Indeed, as highlighted in Figure 15.3, and further supported by two-level hierarchical ANOVA analyses (Table 15.6), the Sino-Tibetan group can be subdivided into four partially overlapping groups: a northern Tibeto-Burman group (i.e. Tibetans and Bhutanese), a northern Chinese group (i.e. Hui and Han samples composed of speakers of Jin and all Mandarin dialects except for Southeastern Mandarin), a Han group of Southeastern Mandarin and Wu speakers, and finally a southern group which comprises both Han speakers of southern

Chinese languages (i.e. Xiang, Gan, Hakka, Min and Yue) and southern Tibeto-Burmans (i.e. Kachari, Sonowal, Lahu, Mikir, Tujia and Yi). This latter group displays close genetic affinities with populations from the Southeast Asian language families.

Table 15.6. Proportion of the total genetic variation of the Gm system that is due to differences within (above diagonal) and between (below diagonal) Sino-Tibetan groups[a] compared two by two.

	NTB	NC	WSM	SC	STB
NTB		0.5*	1.1*	1.1*	3.0*
NC	2.5*		0.7*	0.8*	1.2*
WSM	8.6*	2.2*		1.3*	2.1*
SC	26.5*	14.2*	5.9*		2.4*
STB	24.0*	12.5*	4.8*	n.s.	

[a] See legend to Fig. 15.4 for codes to Sino-Tibetan groups; and see text for the composition of these groups. Significance level: n.s. not significant at the 5% level, * $P < 0.005$.

Correlation between linguistic and genetic distances

Another approach in the study of the relationship between genetics and linguistics is to test for a possible correlation between the degree of genetic similarity (or dissimilarity) between populations and the degree of linguistic similarity (or dissimilarity) between the languages they speak. Genetic dissimilarity between populations, or genetic distance, is a classical measure in population genetics, and several statistics have been developed to quantify it. Here, as for the MDS

analyses, genetic distances were computed as coancestry coefficients based on populations pairwise F_{ST}s (Reynolds *et al.* 1983).

We then used phylogenetic classification to infer measures of evolutionary distance between languages. However, the phylogeny of East Asian languages is disputed, especially with respect to higher-order relationships between language families. Different classification schemes are currently being proposed (see the introduction to this volume). In view of this, we have used three different hypotheses for East Asian languages, which we have called, respectively, hypotheses 1, 2 and 3 (Figure 15.4). Hypothesis 1 (Figure 15.4A) is based on a conjecture by Sagart (1994), according to which all the language families of East Asia, south of Altaic, developed from the language of the first domesticators of rice, ca. 10,000 years BP. In the version used here there are three branches: a northern branch consisting of Sino-Tibetan plus Austronesian including Tai-Kadai (see Sagart's contributions to this volume) and two southern branches, i.e. Hmong-Mien and Austroasiatic. For a similar conjecture, with a different internal subgrouping, see Starosta (this volume). Hypothesis 2 (Figure 15.4B) is represented in such works as Ruhlen (1987) and Peiros (1998) which envision an 'Austric' macro-phylum ('Greater Austric' in the introduction to this volume) and a distinct Sino-Tibetan family, intrusive in East Asia, with genetic connections to north Caucasian and Yenisseian, following Starostin’s Sino-Caucasian theory (Starostin 1984 [1991]). Hypothesis 3 (Figure 15.4C) states that no phylogenetic relationships exist between the main language families of East Asia.

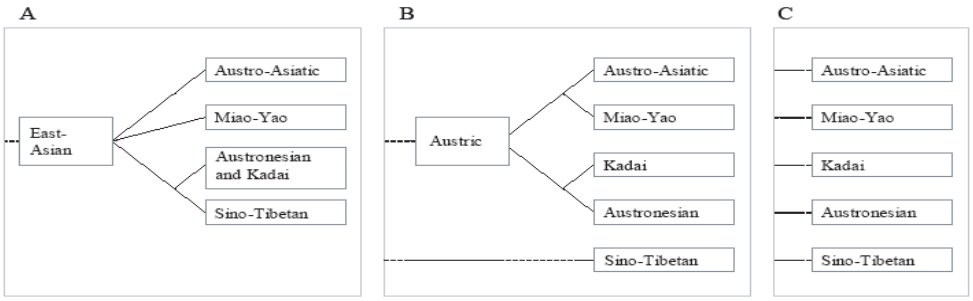

FIGURE 15.4

Three hypotheses on the phylogenetic relationships among languages considered in this study. In A, hypothesis 1 postulates the existence of an East-Asian linguistic macro-family that comprises the Austro-Asiatic, Hmong-Mien, Kadai, Austronesian and Sino-Tibetan families (Sagart 1994). Nodes' ages at the level of accepted language families were defined by LS on the basis of estimates by specialists: W. Ostapirat (personal communication 2001) for Tai-Kadai, G. Diffloth (personal communication 2001) for Austroasiatic, and LS own views, especially for Chinese, Austronesian and Hmong-Mien. Datings of higher-order nodes correspond to archaeological events that LS associates with the upper part of the phylogeny: Proto-East-Asian with the domesticaton of rice, proto-Sino-Austronesian (Sino-Tibetan, Austronesian and Tai-Kadai) with the domestication of millet, and proto-Hmong-Mien with the appearance of iron metallurgy. In B, hypothesis 2 postulates the existence of an Austric macro-family, which relates Austroasiatic, Hmong-Mien, Tai-Kadai and Austronesian. The dates in this phylogeny follow Starostin (1984 [1987], for the root) and Peiros (1998) and are based on glottochonology. Because the Chinese and Austronesian clades are not dealt with in

Peiros (1998), the Chinese and Austronesian internal classifications and datings used in hypothesis 1 were applied to hypothesis 2. The same strategy was applied when detailed statements to construct the classifications dominating specific populations samples included in this study could not be found in Peiros (1998), i.e. central Mon-Khmer, Tai proper and Lolo-Burmese. In C, hypothesis 3 (upon a suggestion raised by R. Blench during the Périgueux workshop) postulates that all the linguistic families considered are unrelated. Here we used, alternatively, the classification and dating schemes of hypotheses 1 (except for Kadai which is treated as a separate family, not as a branch of Austronesian) and 2. Finally, in A, B and C, nodes for which ages were not directly available were assigned dates through equidistant interpolation.

For each of these hypotheses, the linguistic distance between any two languages was equated with the postulated age of the most recent node (i.e. common ancestor) in which they coalesce. When the hypothesis under consideration supposes no genetic relatedness between two languages, the linguistic distance separating them was equated with an arbitrarily high age, to which we refer as the "maximum linguistic distance" (MLX). A description of the dating of ages of nodes in the three hypotheses is given in the legend to Figure 15.4. Here we stress the fact that these three hypotheses differ mainly in that part of the phylogeny nearest to the root (i.e. in the primary branches); lower levels in the phylogenies are less controversial.

Once a matrix of linguistic distances between all pairs of languages was obtained, it was compared to the matrix of genetic distances between all pairs of populations speaking those languages, in order to test the significance of the

resulting correlation coefficient (r). Repeated computations of r were run with values of MLX increased from 15,000 to 50,000 years BP, to account for the effect of the value assigned to the MLX on the correlation coefficient. All of the three linguistic hypotheses lead to significant correlation coefficients ($P < 0.001$) in East Asia, with values increasing with the value of MLX: respectively, from r=0.19 to r=0.31 for the RH system, and r=0.38 to r=0.45 for the GM system.

However, populations that are linguistically related tend to occupy geographically adjacent areas. If genetic and linguistic distances are correlated, then this correlation could be due to the fact that these distances are correlated through geography. Indeed, genetic distances are significantly correlated with geographic distances in East Asia: r=0.24 ($P < 0.001$) for RH and r=0.35 ($P < 0.001$) for GM. To address this fact, we computed partial correlation coefficients between genetic and linguistic distances controlled for geography, i.e. residual correlation coefficients between genetic and linguistic distances once the correlation of both distances with geographic distance has been accounted for (Figure 15.5).

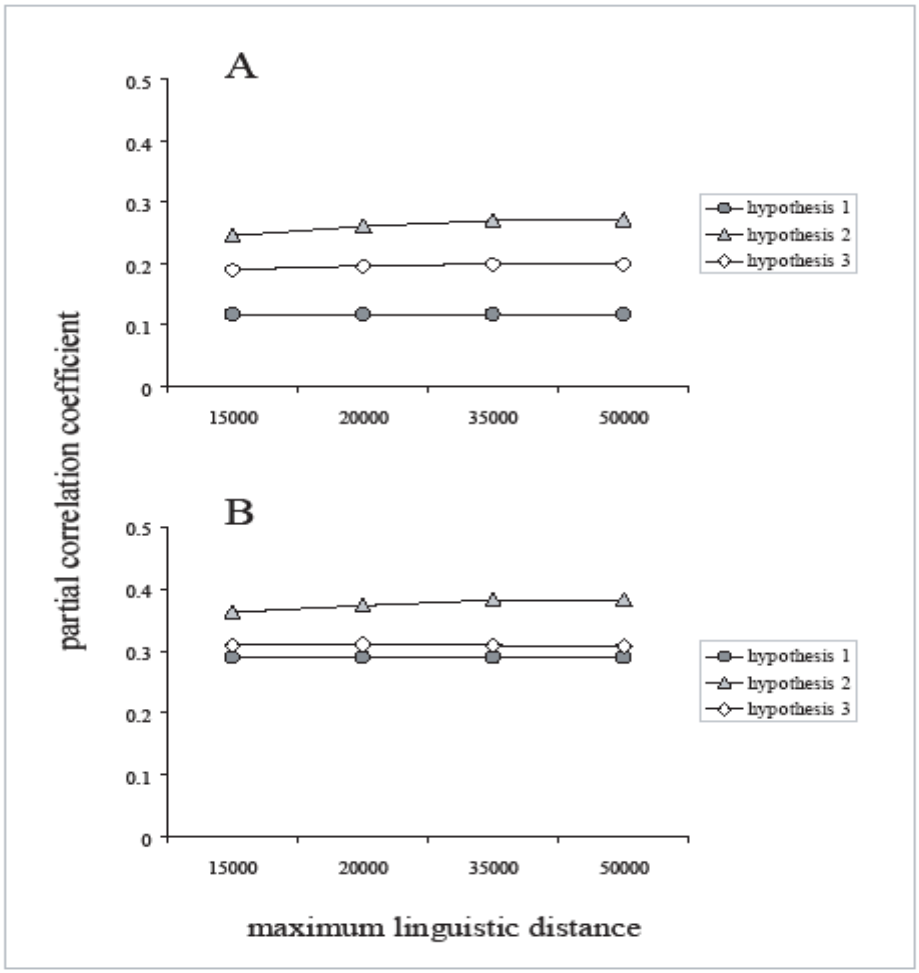
A
B
0.5
0.4
0.3
0.2
0.1
0
15000
20000
35000
50000
hypothesis 1
hypothesis 2
hypothesis 3
partial correlation coefficient
maximum linguistic distance

FIGURE 15.5

Partial correlation coefficients of genetic with linguistic distances, controlled for geography. A: RH data. B: GM data. The three hypotheses of language classification are those of Figure 15.4. To test for the effect of the maximum linguistic distance on the partial correlation with genetic distances, the value assigned to it was varied from 15,000 to 50,000 years. For hypothesis 3, the correlation coefficients reported are those inferred by using the intra-family classification and dating scheme of hypothesis 1 (see legend to Figure 15.4). These coefficients differ from those inferred by using the intra-family classification and dating scheme of hypothesis 2 only at the third decimal (results not shown). All coefficients are statistically significant ($P< 0.001$).

For both the RH and GM systems we observe, firstly, that all three linguistic hypotheses lead to a significant positive partial correlation coefficient between genetic and linguistic distances controlled for geography. Thus, part of the genetic variability among populations observed for both systems is related to the linguistic variability of the languages spoken by these populations. Secondly, hypothesis 2 leads to a slightly higher partial correlation coefficient (from r=0.28 to r=0.30 for RH data and from r=0.36 to r=0.38 for GM data) than both hypotheses 1 and 3. This is because hypothesis 2 postulates that Sino-Tibetan is unrelated to the other linguistic families, in agreement with the observation, both for RH and GM, of a significant genetic differentiation of the Sino-Tibetan group from the other Southeast Asian groups (Table 15.5). However, we cannot assume that the rather small differences in r observed between the three hypotheses are significant because a statistical tool to test for such an assumption is not yet available.

DISCUSSION

Patterns of genetic diversity among East Asian populations

Both the RH and the GM polymorphisms reveal a significant level of genetic structure in East Asia (Table 15.2). This level is quite low for the RH system, in agreement with the fact that only a few haplotypes are inferred from serology (Sanchez-Mazas 1990), and one of those (R^1) dominates the genetic makeup of East Asian populations (Plate Va). Conversely, although the polymorphism of the GM system is also tested here by serology, its variation is more informative[2] and it reveals substantial genetic differentiation among the populations in East Asia (Plate Vb). For comparison, the level of GM genetic differentiation observed in this continent (14.32%) is very similar to that observed among Sub-Saharan African populations (14.96%, based on 51 population samples). This genetic structure seems generally related to the linguistic classification of the languages spoken by East Asian populations since we observe a significant level of variance in genetic diversity among populations from distinct linguistic families (Table 15.4). This seems also to be the case with HLA diversity in continental East Asia (Sanchez-Mazas *et al.*, this volume), and a similar observation is reported in a recent study of Y-chromosome specific biallelic markers (Karafet *et al.* 2001).

At first sight, these results could appear to be compatible with the hypothesis that the populations of a language family share a common genetic origin. By this assumption, distinct models of evolution can be inferred from the comparison of

levels of divergence among populations (F_{ST}) from a linguistically-defined group to levels of diversity within these populations (h), as shown in Table 16.2 of Sanchez-Mazas *et al.* (this volume). According to RH and GM, this comparison suggests roughly three distinct types of evolution in East Asia (Table 15.3). Firstly, the relatively high levels of both internal diversity and inter-population divergence observed among Sino-Tibetans can be explained either by assuming an ancient divergence of Sino-Tibetan populations from a common ancestor, or by substantial incoming gene flow from differentiated sources into Sino-Tibetans, for instance, from populations located north and south of their geographic extension. Secondly, probably because of small population sizes and/or relative geographic isolation, strong genetic drift would characterize the evolution of Austroasiatic populations from a relatively ancient common origin. This would explain both the relatively low internal diversity and high level of inter-population divergence observed among Austroasiatics. Finally, because the Austronesian group is characterized by both low internal diversity and low inter-population divergence, it suggests a recent origin of Austronesians from a rather homogeneous common ancestral population, maybe following a demographic bottleneck. Such an evolution can also be assumed for Tai-Kadai populations. Interestingly, Tai-Kadai is considered as a daughter-group of Austronesian under the hypothesis of an East-Asian linguistic phylum (Figure 15.4A). For the Austronesians, however, the patterns of intra- and inter-population variation (i.e. h and F_{ST}) inferred from HLA (Sanchez-Mazas *et al.*, this volume) are quite different of those inferred from RH

and GM, suggesting either that RH and GM are not as informative as HLA, or that other evolutionary factors, such as selection, could be playing a role here.

More generally, these evolutionary interpretations are challenged by the observation that the genetic relationships among East Asian populations do not show a clear clustering pattern of linguistically-defined groups. For both genetic systems, populations tend to display genetic similarities according to their linguistic relatedness, but also according to geographic proximity. Indeed, we observe non-significant levels of differentiation between several linguistic groups considered two by two (Table 15.5), substantial overlapping of linguistically defined groups in the MDS analyses (Figure 15.2 and 15.3), and similar correlation coefficients of genetic distances with linguistic and geographic distances. In fact, according to both RH and GM polymorphisms, the populations differentiate progressively, along one major axis, from Sino-Tibetan samples down to their Southeastern neighbours, i.e. Hmong-Mien, Tai-Kadai, Austroasiatic and Austronesian, all these latter groups sharing close genetic affinities. This pattern of differentiation very roughly corresponds to a longitudinal axis, and it parallels the frequency increase of haplotypes R^1 (Plate Va) and GM∗1,3;5* (Plate Vb) from the north towards the south of the continent.

Actually, the ANOVA analyses of differentiation among linguistic groups (Table 15.5) indicate a significant genetic differentiation, for GM, between Sino-Tibetan and the other East-Asian groups (Austroasiatic, Tai-Kadai, Hmong-Mien and

Austronesian), but this strong divergence is mainly due to the northernmost Sino-Tibetan populations (northern Han, Tibetans) (Figure 15.3). Sino-Tibetan populations from the south (i.e. Southern Han, Tibeto-Burmans from India, Burma, Thailand, etc.) present genetic similarities with the populations from the Southeast Asian groups, i.e. Hmong-Mien, Tai-Kadai, Austroasiatic and Austronesian, these latter groups being virtually undifferentiated (Table 15.5). However, the differentiation between northern and southern Sino-Tibetan populations is not clear-cut: frequency distributions of Han populations from the Wu and Southeastern Mandarin speaking areas (central-eastern China) are intermediate between those observed in northern and southern populations (Figure 15.3 and Table 15.6). For RH, a fine-scale analysis of Sino-Tibetan populations was not possible with the available sampling. The only notable differentiation observed with this system is between Sino-Tibetans and Austronesians (Table 15.5), but again these two groups include the most differentiated populations in the MDS analyses (Figure 15.2), i.e. the northernmost Sino-Tibetan populations (northern Han, Tibetans) on one hand, and several Austronesian populations of the Malayo-Polynesian daughter-group on the other hand.

Origin(s) of East Asian populations

At present, two alternative hypotheses have been advanced for the origin of continental East Asian populations. One hypothesis postulates that northern East Asians derive from southern populations (e.g. Chu *et al.* 1998; Su *et al.* 1999), whereas the other hypothesis, dubbed the "pincer model" by Ding *et al.* (2000)

suggests two migration routes from the west into East Asia with subsequent contact (e.g. Underhill *et al.* 2001).

The idea of two major contributions to the peopling of the Asian continent (the pincer model) has been put forward to explain a marked genetic differentiation between the north and the south of the continent observed in some studies (Sanchez-Mazas 1990, Cavalli-Sforza *et al.* 1994). However, in agreement with recent analyses of molecular markers (Ding *et al.* 2000; Oota *et al.* 2002; Yao *et al.* 2002), there is little evidence, in our present analyses, for such a clear separation; rather we find a gradual pattern of differentiation. We are aware that our study suffers from the fact that Altaic populations are not represented. However, earlier analyses of GM in East Asia (Sanchez-Mazas 1990) have evidenced continental continuity in changes of frequency distributions, further extending to the north from northern Sino-Tibetans towards such populations as Mongolians, Japanese, Koreans and Siberians (e.g. Buriats, Nentzi, Yakuts). The patterns of genetic relationships among populations observed with mitochondrial DNA (Yao *et al.* 2002b) and Y-specific (Su *et al.* 1999, 2000; Karafet *et al.* 2001) polymorphisms further support a continuous differentiation of Altaic populations from their Sino-Tibetan neighbours, although the sampling of populations in these studies is more restricted.

Could this general pattern of continuity be simply explained in terms of a process of isolation-by-distance, as suggested by Ding *et al.* (2000)? Since genetic and

geographic distances are correlated, this hypothesis cannot be ruled out, but it might be too simplistic, because genetic and linguistic affinities among populations are also correlated (Figure 15.5). Although our results indicate that the linguistic groups do not correspond to separate genetic clusters of populations, this lack of clustering could be due to substantial levels of gene flow between populations across linguistic borders. Such gene flow would both diminish genetic divergence between linguistic groups and raise it between populations within linguistic groups. From our results we can thus hypothesize that gene flow between linguistically distinct sources has substantially contributed to the genetic makeup of East Asian populations, especially for Sino-Tibetan populations.

Advocates of the hypothesis of a southern origin of northern populations claim, among other genetic evidence, that higher numbers of Y-specific haplotypes are observed in the south of the continent (Su *et al.* 1999). However, caution must be exercised when reasoning on the presence or absence of alleles or haplotypes in samples, as the probability of missing a rare allele or haplotype increases very rapidly as sample size decreases[3] (Sanchez-Mazas 2002). The Y-specific study of Karafet *et al.* (2001) reported on higher gene diversity (*h*) in southern populations than in northern ones (although the difference is rather small). To some extent, we observe the opposite pattern with RH and GM, in that Sino-Tibetan populations are more heterogeneous than Austroasiatic, Tai-Kadai, Hmong-Mien or Austronesian populations (Table 15.3). However, RH or GM haplotypes are determined by serology, and any given serological haplotype might include

several distinct molecular variants (Dard *et al.* 1996).[4] Nevertheless, Karafet *et al.* (2001) also reported on higher molecular diversity among Y-specific haplotypes in the north than in the south, suggesting there were more genetic contributions from distinct sources in northern than in southern East Asia, as in our analyses.

Thus, neither our results nor other studies on molecular markers can discriminate, at present, between the two competing hypotheses on the origin of continental East Asian populations. Actually recent hypotheses tend to reconcile both models into a framework that includes substantial gene flow between populations differentiating in East Asia at various times, as well as an important genetic input from central Asia into northern East Asia (Ding *et al.* 2000; Su *et al.* 2000; Karafet *et al.* 2001; Wells *et al.* 2001). Our results are compatible with the hypothesis that Austroasiatic, Tai-Kadai, Hmong-Mien, and Austronesian populations share a common origin. These groups of populations may have differentiated by settling into geographically distinct areas, eventually coming into secondary contact and thus favouring genetic and cultural exchange. Given the present extension of these linguistic families, it is tempting to assume that these differentiation processes took place in southern East Asia, but we have no evidence to link these groups to the first settlers of the continent. Insights into this matter may be gained in the future by analyses of ancient DNA (Oota *et al.* 1999, 2001b; Wang *et al.* 2000).

In turn, at least two scenarios can be envisioned for the origins of Sino-Tibetan populations. Either Sino-Tibetans differentiated from the same common source as the other East Asian groups, a common source that should be linked to the hypothesis of a proto-East-Asian linguistic phylum (Figure 15.4A). In favour of this hypothesis, we observe little differentiation of southern Sino-Tibetans (either Han or Tibeto-Burmans) from other Southeast Asian groups, and in particular from most Austronesians (Figures 15.2 and 15.3). Some Sino-Tibetan populations might thus have differentiated through a northwards expansion, where they would have eventually experienced strong genetic inflow from distinct northern, possibly Altaic, groups. Northern Mandarin has indeed been deeply influenced by Altaic languages (Hashimoto 1986). Alternatively, the Sino-Tibetans have an independent origin. In this case, a scenario that could fit the genetic data would assume a southwards expansion of Sino-Tibetans, where they would have assimilated already settled populations, while imposing their language(s). Under this scenario, substantial gene flow between "intrusive" Sino-Tibetans and already settled Southeastern groups must be invoked to account for the observation of no sharp genetic changes between north and south.

The correlation analyses between linguistic and genetic distances carried out in this study argue in favour of hypothesis 2 (Figure 15.5), i.e. for a common origin of the populations of the Southeast Asian linguistic families (Austroasiatic, Tai-Kadai, Hmong-Mien and Austronesian) and a separate origin of Sino-Tibetans (Figure 15.4B). The case for this hypothesis is not strong since it leads to

correlation coefficients not much higher than those obtained for hypotheses 1 and 3. Moreover, even with the hypothesis that Sino-Tibetan populations do share a common origin with the Southeastern groups, hypothesis 2 could still perform better than the others if the divergence of the Sino-Tibetan group was accentuated by substantial genetic input from other, differentiated, sources (e.g. from Altaic populations).

CONCLUSION

In this study, our purpose is not to confirm or invalidate a linguistic hypothesis of genetic relationships among languages with genetic data. Indeed, there is no *a priori* reason why genetic data could do this. There are several ways by which populations that share a common linguistic and genetic origin might diverge from one another, either genetically, or linguistically, or both. For instance, if linguistically related populations are submitted to strong genetic drift, because population sizes are small, then they can diverge genetically quite rapidly but may retain a strong linguistic relatedness. Alternatively, a population can acquire a new (even unrelated) language, for instance through a process of domination by an elite (see for instance Renfrew 1989), without diverging genetically from their former linguistic relatives.

However, when considering a large set of populations, as was done here, we observe that genetic and linguistic distances are correlated to some extent. We have shown that, among the East Asian groups considered in this study, genetic

distances among populations generally increase with the linguistic distances among their languages, although with some variation. Correlation between genetic similarity and linguistic relatedness has also been described for other regions of the world, and other genetic systems (e.g. Sokal *et al.* 1992). It suggests that there is a relationship between the process of language diversification and that of genetic differentiation of the populations, i.e. that both processes have occurred through a common cause. If this hypothesis is correct, it implies that the origin of the genetic structure we observe today is to be linked to the origin of language families. In other words, since linguists assume that the ages of East Asian linguistic families are 10,000 years or less, then at least part of the genetic structure of today's populations might originate within that period. Of course the genes (i.e. the genetic variants that we observe) might be much older, but the genetic pools (the frequency distributions observed in the populations) can be much more recent. Indeed, the fact that the vocabulary of domestic crops reconstructs in the proto-languages of several of the East Asian linguistic families considered in this study (Blench, this volume; Sagart, forthcoming) strongly suggests that the genetic profiles of East Asian populations have been deeply influenced by the demographic (and territorial) expansion that is concomitant with the transition to food-producing economies. Such expansions would both slow down population differentiation through genetic drift and induce conditions to cultural and genetic exchange. If genetic exchange through secondary contact between populations has been the rule rather than the exception in the history of East Asia, then we need to use appropriate statistical tools, such as spatial

autocorrelation analyses (Sokal and Oden 1978) and analyses of the impact of linguistic boundaries on genetic structure (Dupanloup de Ceuninck *et al.* 2000) to discriminate between specific cases of populations differentiating from a common source and cases of convergence through secondary contact.

ACKNOWLEDGEMENTS

This research was supported by the French CNRS OHLL (*Origine de l'Homme, du Langage et des Langues*) action to ESP and LS.

BIBLIOGRAPHY

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

Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T., Yang, Z. Q., Lin, K. Q., Li, P., Wu, M., Geng, Z. C., Tan, C. C., Du, R. F. and Jin, L. (1998) 'Genetic relationship of populations in China', *Proceedings of the National Academy of Sciences USA*, 95: 11763-8.

Dard, P., Sanchez-Mazas, A., Dugoujon, J.-M., De Lange, G., Langaney, A., Lefranc, M.-P. and Lefranc, G. (1996) 'DNA analysis of the immunoglobulin IGHG loci in a Mandenka population from eastern Senegal: correlation with Gm haplotypes and hypotheses for the evolution of the Ig CH region', *Human Genetics*, 98: 36-47.

Ding, Y.-C., Wooding, S., Harpending, H. C., Chi, H.-C., Li, H.-P., Fu, Y.-X., Pang, J.-F., Yao, Y.-G., Yu, J.-G., Moyzis, R. and Zhang, Y. (2000) 'Population structure and history in East Asia', *Proceedings of the National Academy of Sciences USA*, 97: 14003-6.

Dugoujon, J.-M., Hazout, S., Loirat, F., Mourrieras, B., Crouau-Roy, B. and Sanchez-Mazas, A. (forthcoming) 'GM haplotype diversity of 82 populations over the World', to appear in *American Journal of Physical Anthropology*.

Dupanloup de Ceuninck, I., Schneider, S., Langaney, A. and Excoffier, L. (2000) 'Inferring the impact of linguistic boundaries on population differentiation: application to the Afro-Asiatic-Indo-European case', *European Journal of Human Genetics*, 8: 750-6.

Hashimoto, M. J. (1986) 'The Altaicization of Northern Chinese', in J. McCoy J and T. Light (eds.) *Contributions to Sino-Tibetan Studies*, Leiden: E. J. Brill.

Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R. S., Redd, A. J., Zegura, S. L. and Hammer, M. F. (2001) 'Paternal population history of East Asia: sources, patterns, and microevolutionary processes', *American Journal of Human Genetics*, 69: 615-28.

Ke, Y., Su, B., Song, X., Lu, D., Chen, L., Li, H., Qi, C., Marzuki, S., Deka, R., Underhill, P., Xiao, C., Shriver, M., Lell, J., Wallace, D., Wells, R. S., Seielstad, M., Oefner, P., Zhu, D., Jin, J., Huang, W., Chakraborty, R., Chen, Z. and Jin, L. (2001) 'African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes', *Science*, 292: 1151-3.

Lahr, M. M. and Foley, R. (1994) 'Multiple dispersals and modern human origins', *Evolutionary Anthropology*, 3: 48-60.

Lahr, M. M. and Foley, R. (1998) 'Towards a theory of modern human origins: geography, demography and diversity in recent human evolution'. *Yearbook of Physical Anthropology*, 41: 137-76.

Oota, H., Saitou, N., Matsushita, T. and Ueda S. (1999) 'Molecular genetic analysis of remains of a 2,000-year-old human population in China and its relevance for the origin of the modern Japanese population', *American Journal of Human Genetics*, 64: 250-8.

Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T. and Stoneking, M. (2001a) 'Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence', *Nature Genetics*, 29: 20-1.

Oota, H., Kurosaki, K., Pookajorn, S., Ishida, T. and Ueda, S. (2001b) 'Genetic study of the Paleolithic and Neolithic Southeast Asians', *Human Biology*, 73: 225-31.

Oota, H., Kitano, T., Jin, F., Yuasa, I., Wang, L., Ueda, S., Saitou, N. and Stoneking, M. (2002) 'Extreme mtDNA homogeneity in continental Asian populations', *American Journal of Physical Anthropology*, 118: 146-53.

Peiros, I. (1998) *Comparative Linguistics in Southeast Asia*, Canberra: Pacific Linguistics.

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.

Renfrew, C. (1989) 'The origins of Indo-European languages', *Scientific American*, 261: 82-90.

Reynolds, J., Weir, B. S. and Cockerham, C. C. (1983) 'Estimation of the coancestry coefficient: basis for a short-term genetic distance', *Genetics*, 105: 767-79.

Rohlf, F. J. (1998) *NTSYSpc: numerical taxonomy and multivariate analysis system*, New York: Exeter Software.

Ruhlen, M. (1987) *A Guide to the World's Languages*, London: Edward Arnold.

Sagart, L. (1994) 'Proto-Austronesian and Old Chinese: evidence for Sino-Austronesian', *Oceanic Linguistics*, 33: 271-308.

— (forthcoming) 'The vocabulary of cereal cultivation and the phylogeny of East Asian languages', to appear in *Bulletin of the Indo-Pacific Prehistory Association.*

Sanchez-Mazas, A. (1990) *Polymorphisme des systèmes immunologiques Rhésus, GM et HLA et histoire du peuplement humain*, unpublished thesis, Geneva: University of Geneva.

— (2002) 'HLA data analysis in anthropology: basic theory and practice', in *·16th European Histocompatibility Conference*, European Federation for Immunogenetics (EFI), Strasbourg, France, pp. 68-83.

Schneider, S., Roessli, D. and Excoffier, L. (2000) *Arlequin ver 2.000: a software for population genetics data analysis*, Geneva: Genetics and Biometry Laboratory, University of Geneva.

Seielstad, M. T., Minch, E. and Cavalli-Sforza, L. L. (1998) 'Genetic evidence for a higher female migration rate in humans', *Nature Genetics*, 20: 278-80.

Sokal, R. R. and Oden, N. L. (1978) 'Spatial autocorrelation in biology', *Biological Journal of the Linnean Society*, 60: 73-93.

Sokal, R. R., Oden, N. L. and Thomson, B. A. (1992) 'Origins of the Indo-Europeans: genetic evidence', *Proceedings of the National Academy of Sciences USA*, 89: 7669-73.

Starostin, S. (1984 [1991]) 'On the Hypothesis of a genetic connection between the Sino-Tibetan languages and the Yeniseian and North-Caucasian languages', translation and introduction by William H. Baxter III, in V. Shevoroshkin (ed.) *Dene-Sino-Caucasian.* Bochum: Brockmeyer.

Steinberg, A. G. and Cook, C. E. (1981) *The distribution of the human immunoglobulin allotypes*, Oxford: Oxford University Press.

Su, B., Xiao, J., Underhill, P., Deka, R., Zhang, W., Akey, J., Huang, W., Shen, D., Lu, D., Luo, J., Chu, J., Tan, J., Shen, P., Davis, R., Cavalli-Sforza, L., Chakraborty, R., Xiong, M., Du, R., Oefner, P., Chen, Z. and Jin, L. (1999) 'Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last Ice Age', *American Journal of Human Genetics*, 65: 1718-24.

Su, B., Xiao, C., Deka, R., Seielstad, M. T., Kangwanpong, D., Xiao, J., Lu, D., Underhill, P., Cavalli-Sforza, L., Chakraborty, R. and Jin, L. (2000) 'Y chromosome haplotypes reveal prehistorical migrations to the Himalayas', *Human Genetics*, 107: 582-90.

Underhill, P. A., Passarino, G., Lin, A. A., Shen, P., Mirazon Lahr, M., Foley, R. A., Oefner, P. J. and Cavalli-Sforza, L. L. (2001) 'The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations', *Annals of Human Genetics*, 65: 43-62.

Wang, L., Oota, H., Saitou, N., Jin, F., Matsushita, T. and Ueda, S. (2000) 'Genetic structure of a 2,500-year-old human population in China and its spatiotemporal changes', *Molecular Biology and Evolution*, 17: 1396-400.

Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., Jin, L., Su, B., Pitchappan, R., Shanmugalakshmi, S., Balakrishnan, K., Read, M., Pearson, N. M., Zerjal, T., Webster, M. T., Zholoshvili, I., Jamarjashvili, E., Gambarov, S., Nikbin, B., Dostiev, A., Aknazarov, O., Zalloua, P., Tsoy, I., Kitaev, M., Mirrakhimov, M., Chariev, A. and Bodmer, W. F. (2001) 'The Eurasian heartland: a continental perspective on Y-chromosome diversity', *Proceedings of the National Academy of Sciences USA*, 98: 10244-9.

Yao, Y.-G., Kong, Q.-P., Bandelt, H.-J., Kivisild, T. and Zhang, Y.-P. (2002a) 'Phylogeographic differentiation of mitochondrial DNA in Han Chinese', *American Journal of Human Genetics*, 70: 635-51.

Yao, Y.-G., Nie, L., Harpending, H., Fu, Y.-X., Yuan, Z.-G. and Zhang, Y.-P. (2002b) 'Genetic relationship of Chinese ethnic populations revealed by mtDNA sequence diversity', *American Journal of Physical Anthropology*, 118: 63-76.

[1] "Classical markers" refers here to genetic systems that reveal variation between individuals at the level of the gene product (i.e. the protein), not at the level of the gene itself, as is the case for "DNA markers".

[2] A worldwide analysis of GM variation reveals one of the strongest levels of population genetic structure observed so far for an autosomal marker (39.14%), and this structure globally corresponds to continental groupings of populations (Dugoujon *et al.* forthcoming).

[3] For instance, Y-chromosome mutation M95, considered southern-specific by Su *et al.* (1999), has also been observed in some northern samples (Sino-Tibetan and Altaic) (Su *et al.* 2000; Karafet *et al.* 2001; Wells *et al.* 2001).

[4] Actually caution should also be exercised with Y-chromosome haplotypes defined by biallelic markers, because the former could also include further sub-variants.