HLA genetic diversity and linguistic variation in East Asia

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

CHAPTER 16

HLA GENETIC DIVERSITY AND LINGUISTIC VARIATION IN EAST ASIA

Alicia SANCHEZ-MAZAS, Estella S. POLONI, Guillaume JACQUES and Laurent SAGART

8043 words

INTRODUCTION

Molecular anthropology - the study of human genetic polymorphisms - is now often used to investigate the accuracy of archaeological and/or linguistic hypotheses. One of the classic examples is the use of genetics in an attempt to discriminate between two alternative models for the spread of agriculture in Europe - the demic and the cultural diffusion models - which finally led to a general approval of the former by geneticists, who regard this spread as possibly linked to the expansion of Indo-European languages (Ammerman and Cavalli-Sforza 1984, Renfrew 1992, Barbujani *et al.* 1995, Weng and Sokal 1995, Chikhi *et al.* 2002). More generally, because genetic clines can give evidence for population migrations (Barbujani 2000), the analysis of genetic patterns is particularly interesting for the analysis of early agriculturalist diasporas and their link to the diffusion of human languages (Barbujani and Pilastro 1993, Bellwood 2001). Molecular anthropology can also be useful in estimating the contribution of different gene pools to the make-up of present-day populations, when attempting

to ascertain the origin of specific linguistic families (such as the Austronesian family, see further in this chapter); to test the permeability of linguistic boundaries to gene flow (Dupanloup de Ceuninck *et al.* 2000); or to investigate precise linguistic hypotheses (Excoffier *et al.* 1987; Poloni *et al.*, this volume; this study), although genetics alone cannot be used to discriminate between alternative linguistic models.

The present work aims at bringing genetic evidence to bear on the vexing question of East Asian linguistic relationships. The phylogenetic links between the main language phyla of this region (Sino-Tibetan, Austroasiatic, Tai-Kadai, Austronesian and Altaic) are still deeply controversial (see the introduction to the volume for a review of the main theories). To investigate these relationships from a genetic point of view, we report here on the results of a population genetics analysis of one molecular polymorphism, HLA-DRB1. The DRB1 locus of the major histocompatibility complex (MHC) in humans is a cell surface protein-encoding gene, located on the short arm of chromosome 6 and surrounded by other HLA loci. Its allelic variability is amongst the highest known in the human genome thus far, with 418 DRB alleles detected by DNA oligotyping and sequencing techniques (IMGT/HLA sequence database 2003). Besides this high level of polymorphism, the DRB1 locus also has the advantage of having been extensively tested at the DNA level in human populations for at least 15 years (mostly using the HLA International Workshop typing kits), and abundant population data with high-resolution allelic definition are thus available. In this

study, we analyse this polymorphism to explore a possible congruence between genetic and linguistic relationships in East Asia.

MATERIAL AND METHODS

Populations analysed

We collected population data tested by high-resolution DNA typing for HLA-DRB1 through a thorough review of the literature, adding some samples submitted to the 11th, 12th and 13th HLA workshops and samples obtained through personal communications (Table 16.1). Our aim was to represent all regions of East and Southeast Asia as far as eastern Indonesia. We tried to avoid statistical bias due to low sample sizes, low allelic resolution, or heterogeneous population samples (Sanchez-Mazas 2002). We thus excluded samples with less than 40 individuals, samples with more than 5% "blank" frequency corresponding to undefined alleles, and samples for which only a generic definition of HLA-DRB1 alleles (HLA "broad" specificities) was available. We also excluded all samples the linguistic affiliation of which was unclear or ambiguous. These criteria left us with a final list of 46 linguistically well characterized East Asian populations, defined by a total of 76 DRB1 allele frequencies (Table 16.1). We also included two West Asian populations (Mansi and North Indians) to represent the western edge of the area under study[1]. Overall, these 48 populations are represented by a total of 6,613 individuals.

Table 16.1: Populations considered in this study[a]

#	N	Population	Country	Location	Lat	Long	Language	LF
1	68	Mansi	Russia	Khanty-Mansi	60.2	70.7	Uralic	UY
2	59	Chukchi	Siberia	Several regions	65	185	Chukchi	CK
3	92	Koryak	Siberia	Kamchatka	60	164	Koryak	CK
4	80	Yupik	Siberia	Behring coast	66	185	Eskimo	EA
5	47	Indian	India	North	28.4	77.2	Indo-European	IE
6	53	Nivkhi	Russia	Siberia, Nogliki	51.5	143	Gilyak	GI
7	42	Kazakh	China	Ürümqi	43.4	87.4	Turkic	AL
8	160	Manchu	China	Heilongjiang	45.2	126	Tungus	AL
9	41	Khalkh	Mongolia	Ulaanbaatar	47.5	107	Mongolian	AL
10	201	Khalkh	Mongolia	Kharkhorum	45	100	Mongolian	AL
11	57	Uighur	China	Ürümqi	43.4	87.4	Turkic	AL
12	190	Tuvin	Russia	Kyzyl	51.4	94.3	Turkic	AL
13	44	Tuvin	Russia	Kyzyl	51.4	94.3	Turkic	AL
14	73	Ulchi	Russia	Khabarovsk	54	136	Tungus	AL
15	43	Tofalar	Russia	Nizhneudinsk	54.9	99	Turkic	AL
16	197	Ryukyuan	Japan	Okinawa	26	127	Ryukyuan	AL
17	371	Japanese	Japan	Centre	35.4	139	Japanese	AL
18	916	Japanese	Japan	n.d.[b]	35.4	139	Japanese	AL
19	510	Korean	Korea	Seoul	37.3	127	Korean	AL
20	199	Korean	Korea	Heilongjiang	46	127	Korean	AL
21	91	Chinese	China	Guan County	39.3	116	Sinitic	SI
22	89	Chinese	China	Shanghai	31.1	121	Sinitic	SI
23	59	Chinese	China	Ürümqi	43.4	87.4	Sinitic	SI
24	162	Chinese	China	Xiamen, Fujian	24.3	118	Sinitic	SI
25	1012	Taiwanese	Taiwan	Tainan	23	120	Sinitic	SI

26	190	Taiwanese	Taiwan	n.d.	24	121	Sinitic	SI
27	70	Buyi	China	n.d.	26.2	106	Tai-Kadai	KA
28	140	Thai	Thailand	Bangkok	13.4	100	Tai-Kadai	KA
29	96	Dai Lue	Thailand	North	17	101	Tai-Kadai	KA
30	106	Dai Dam	Thailand	North	17.6	102	Tai-Kadai	KA
31	100	Kinh	Vietnam	Hanoi	21.1	106	Mon-Khmer	AU
32	81	Muong	Vietnam	Hoa Binh	20.5	105	Mon-Khmer	AU
33	40	Indonesian	Indonesia	Molucca	0	128	Mal.-Pol.[d]	AN
34	49	Indonesian	Indonesia	Nusa Tenggara	-9	117	Mal.-Pol.	AN
35	77	Indonesian	Indonesia	Java, Jakarta	-6.1	106	Mal.-Pol.	AN
36	77	Malay	Malaysia	n.d	3.9	101	Mal.-Pol.	AN
37	105	Filipino	Philippines	South Luzon[e]	14.4	121	Mal.-Pol.	AN
38	50	Ivatan	Philippines	Batan islands	20.3	122	Proto-Filipino	AN
39	65	Paiwan	Taiwan	South	23.5	121	Paiwanic	AN
40	51	Paiwan	Taiwan	C. mountains[d]	22.2	121	Paiwanic	AN
41	50	Atayal	Taiwan	C. mountains	24.3	121	Atayalic	AN
42	57	Saisiat	Taiwan	C. mountains	24.5	121	Western Plains	AN
43	88	Bunun	Taiwan	C. mountains	23.2	121	Paiwanic	AN
44	51	Tsou	Taiwan	C. mountains	23.4	120	Tsuic	AN
45	50	Rukai	Taiwan	C. mountains	22.4	120	Tsuic	AN
46	50	Ami	Taiwan	East coast	23.1	121	Sirayan	AN
47	50	Puyuma	Taiwan	East coast	22.4	121	Puyumic	AN
48	64	Yami	Taiwan	Orchid Island	22	121	Proto-Filipino	AN

[a] N: sample size; Lat: latitude; Long: longitude; LF: Linguistic family (UY: Uralic-Yukaghir, IE: Indo-European, AL: Altaic, AN: Austronesian, AU: Austroasiatic, CK: Chukchi-Kamchatkan, KA: Tai-Kadai, EA: Eskimo-Aleut, GI: Gilyak, SI: Sinitic); n.d.: not determined. References: 1: Uinuk-Ool *et al.* 2002; 2-4: Grahovac *et al.* 1998; 5: Rani *et al.* 1998; 6: Lou *et al.* 1998; 7: Mizuki *et al.* 1997; 8: XIIth Workshop data (personal

communication to ASM); 9: Munkhbat *et al.* 1997; 10: Chimge *et al.* 1997; 11: Mizuki *et al.* 1998; 12: Martinez-Laso *et al.* 2001; 13-15: Uinuk-Ool *et al.* 2002; 16: Hatta *et al.* 1999; 17: Saito *et al.* 2000; 18: Hashimoto *et al.* 1994; 19: Park *et al.* 1999; 20: XIIth Workshop data (personal communication to ASM); 21: Gao *et al.* 1991; 22: Wang *et al.* 1993; 23: Mizuki *et al.* 1997; 24: Lee 1997; 25: XIIIth Workshop data (personal communication to ASM); 26: Chu *et al.* 2001; 27: Imanishi *et al.* 1992; 28-30: Chandanayingyong *et al.* 1997; 31: Vu-Trieu *et al.* 1997; 32: XIIIth Workshop data (personal communication to ASM); 33-34: Mack *et al.* 2000; 35: Gao *et al.* 1992; 36: Mack *et al.* 2000; 37: Bugawan *et al.* 1994; 38: Chu *et al.* 2001; 39: XIIth Workshop data (personal communication to ASM); 40-48: Chu *et al.* 2001. [b] not determined. [c] Malayo-Polynesian. [d] Central mountains. [e] Typed in USA.

Linguistic phylogeny

Linguistic phylogenetic trees with absolute differentiation dates were established by one of us (LS) to represent what we consider to be the "least controversial phylogeny" for each of the phyla under consideration: Koreo-Japonic, Altaic (tentatively accepted here on the basis of shared pronominal paradigms), Sino-Tibetan, Tai-Kadai, Austroasiatic, and Austronesian. The trees were established on the basis of the literature, or through consultation with specialists. Because we needed to integrate all the different trees into one so as to obtain separation dates for languages belonging to different families, and in order to avoid the controversial issues of higher sub-grouping between these families, each proto-language was directly linked to a root node, the date of which was arbitrarily set at 50,000 years BP. The overall phylogeny thus obtained for the present analyses

(see the Results section) does not necessarily reflect our own ideas (or anyone else's for that matter), but we believe it integrates largely uncontroversial information concerning the linguistic affiliation of each language, as well as some relatively widely-held views about the internal sub-grouping and times of separation within each family, while remaining neutral on higher sub-grouping.

Statistical methods

Pairwise F_{ST} indexes among populations (a measure of their genetic variation) were computed from their HLA-DRB1 allele frequency distributions and tested for significance by a permutation procedure (Schneider *et al.* 2000). A matrix of coancestry coefficients (Reynolds *et al.* 1983) was used as a genetic distance matrix to plot the populations according to the technique of multidimensional scaling (MDS) (Kruskal 1964, Rohlf 2000). Geographic coordinates were determined for all populations, and were used to compute geographic distances based on the arc length of a sphere and transformed to natural logarithms (Nicolas Ray, personal communication). The date of the most recent common ancestor (or proto-language) of two given languages, as given by the linguistic phylogeny that we constructed, was taken as a "linguistic distance" between the populations speaking those languages (see also the companion paper by Poloni *et al.* in this volume). Correlation coefficients were computed between the genetic, geographic, and linguistic distance matrices and assessed for significance by two-way and three-way Mantel tests (Mantel 1967). In three-way tests, the first two

matrices are adjusted to take into account their possible covariation with a third matrix (Sokal and Rohlf 1994).

Different fixation indices, F_{ST}, F_{CT} and F_{SC}, were estimated to assess the levels of genetic diversity among populations at different hierarchical levels of subdivision (Excoffier 2001). When a single set of populations is considered, F_{ST} represents the overall genetic variation among these populations. When several groups of populations are considered (linguistically defined groups, for example), one may also estimate F_{CT} and F_{SC} to represent the levels of genetic variation *among groups* and *among populations within groups*, respectively. This genetic structure is analysed using an analysis of variance (ANOVA) framework, where the significance of the statistics is assessed by a resampling procedure (Schneider *et al.* 2000). We also estimated the gene diversity within each population (*h*) by its expected heterozygosity (Nei 1987). All resampling and permutation procedures were done with a total of 10,000 runs.

Models of genetic evolution

With the aim of investigating different mechanisms of population differentiation in relation to the history of East Asian linguistic families, F_{ST} and *h* (averaged on all populations considered) were estimated simultaneously within each linguistic family in order to describe the genetic diversity, both *among* and *within* populations, of that family. This led us to consider four distinct patterns of genetic diversity (A to D) corresponding to extreme variations ("high" or "low") of these

two statistics taken together. When one of these patterns is identified in a given linguistic family, one or several modes of genetic evolution can be inferred for that family[2] (Table 16.2).

Table 16.2. Main patterns of population genetic diversity and their possible explanations in terms of genetic evolutionary mechanisms

Patterns	Observed genetic diversity		Inferred evolutionary mechanisms
	Among populations (F_{ST})	Within populations (h)	
A.	high	high	1. Early differentiation of populations, maintenance of genetic diversity among populations by limited gene flow, maintenance of genetic diversity within populations by large population sizes 2. Intensive gene flow from highly diversified external populations
B.	low	high	1. Intensive gene flow among populations after differentiation from a highly diversified population 2. Recent differentiation from a highly diversified population
C.	high	low	Genetic drift and/or founder effects in small-sized and isolated populations
D.	low	low	1. Intensive gene flow among populations after differentiation from a population with reduced diversity

2. Recent differentiation from a population with reduced diversity

RESULTS

HLA genetic diversity in East Asian populations

As with most human MHC loci, HLA-DRB1 genetic profiles are generally highly heterogeneous within human populations, i.e. they are commonly characterized by numerous alleles at low frequencies. This is what we observe for East Asia (Plate VIII), where, at first glance, genetic profiles do not reveal a clear population structure. Nevertheless, a finer examination shows that some alleles reach relatively high frequencies in specific East Asian populations. This is the case for *1402 and *0401 in Siberians, *1201, *07, or *0301 in Altaic, *0405 in the Japanese, *0901 in the Chinese, *1401, *0803, *1101, *0403, or *0404 in different Aboriginal populations from Taiwan, and *1502 and *1202 in Southeast Asians, with extreme frequencies of the latter alleles in most Austronesians (up to 0.507 for *1202 in Java, and up to 0.479 for *1502 in Nusa Tenggara, while *1502 is very rare in Taiwan Aborigines) (Plate VIII). These patterns indicate that some East Asian populations deeply differ genetically from each other, and that a high level of genetic diversity characterizes this continental area. This is confirmed by F_{ST} measures. The overall HLA-DRB1 genetic diversity among the 46 East Asian populations considered in this study (thus excluding the West Asians Mansi

and Indians) is 4.6% (P<0.001). This is much higher than values

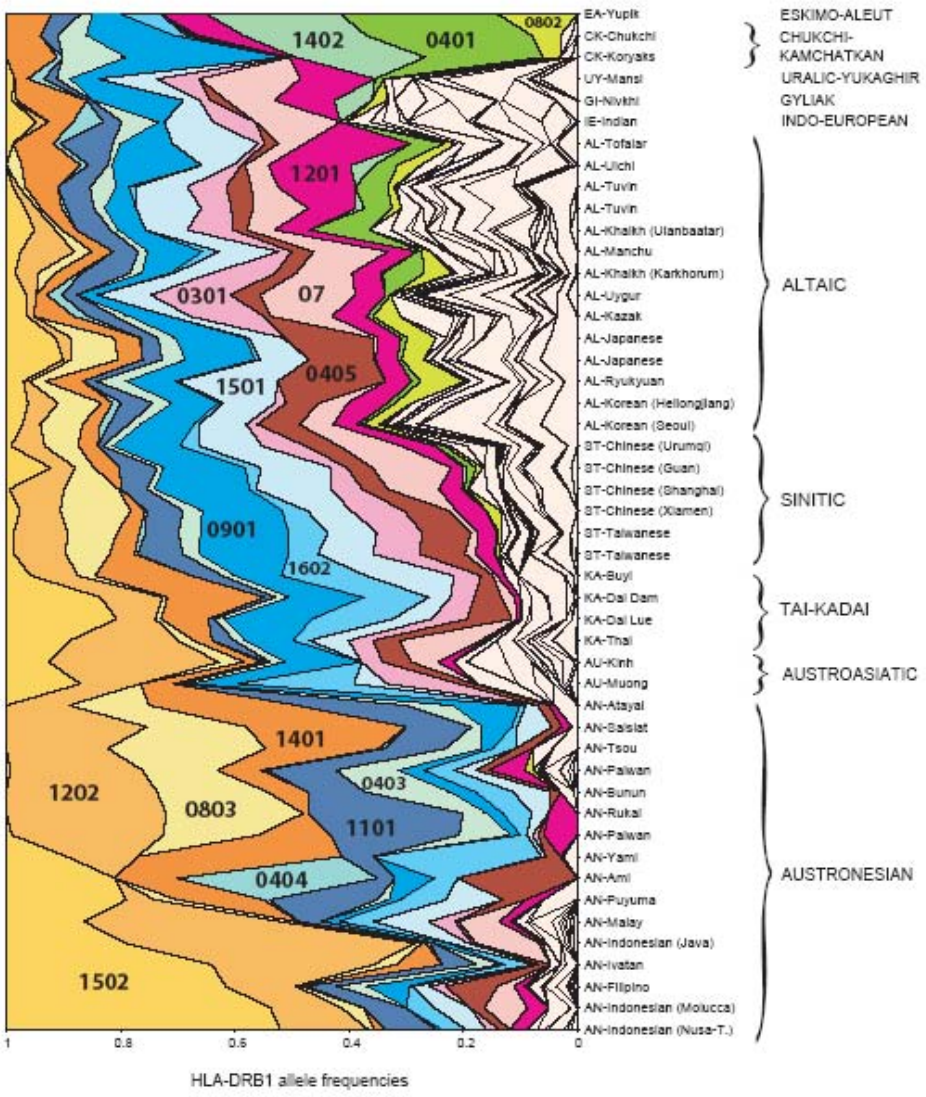


Plate VIII: HLA-DRB1 allele frequencies in 48 Asian populations ordered by linguistic

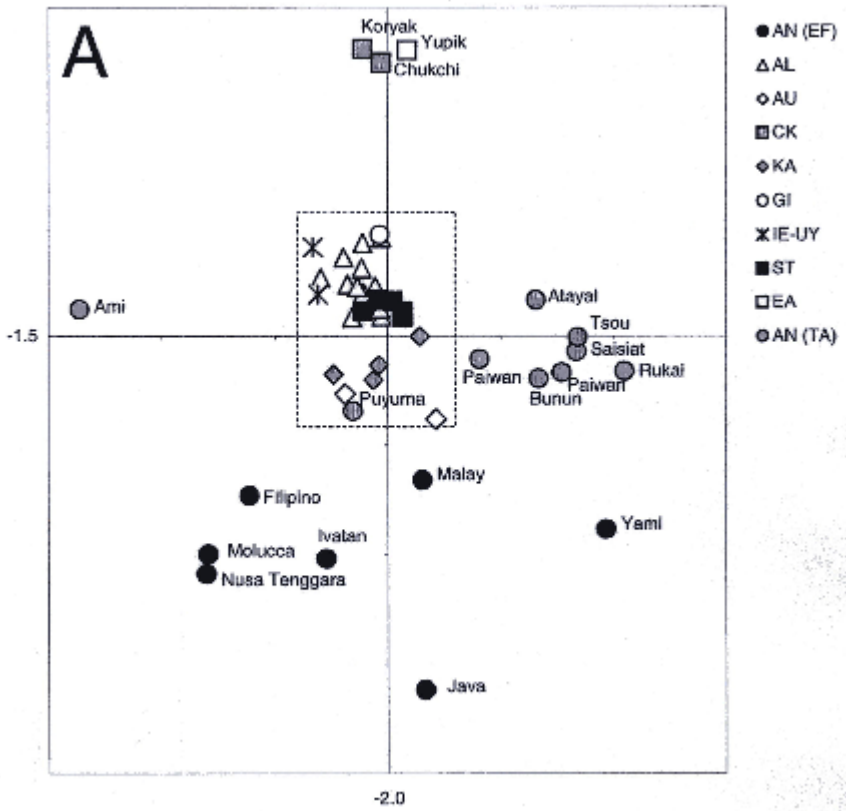
families. The most frequent alleles (frequency > 14% in at least one population) are represented with bright coloured areas. AN: Austronesian; AL: Altaic; AU: Austroasiatic; CK: Chukchi-Kamchatkan; KA: Tai-Kadai; GI: Gilyak; IE: Indo-European; UY: Uralic-Yukaghir; ST: Sino-Tibetan (here only Sinitic); EA: Eskimo-Aleut.

estimated for Europe (1.4-2%), and higher than values found in sub-Saharan Africa (3.4-4%), as already suggested on the basis of DRB1 analyses carried out on more limited sets of populations (Sanchez-Mazas 2001, forthcoming).

A two-dimensional scaling analysis (MDS) of the 46 populations, plus the Mansi and Indians, is presented in Figure 16.1A. An overall correspondence is observed between the genetic pattern and geography: continental East Asian populations (Chinese, Japanese, Koreans, Mongolians, Thai, Vietnamese, West Asians and Nivkhi) plus the Puyuma from Taiwan are tightly clustered in the centre of the MDS (dotted box in Figure 16.1A). The Siberians (Koryak, Chukchi and Yupik) segregate at the top, and the Malaysians, Filipinos, and Indonesians at the bottom. The northwest Asian Mansi (Uralic-Yukaghir speakers) and the Indians (Indo-Europeans) are close to the Uighur, the westernmost East Asian population (a non-significant F_{ST} is even found between Mansi and Uighur, see legend for Figure 16.1). We also note that the Aborigines from the central mountains of Taiwan (Atayal, Saisiat, Bunun, Tsou, Rukai, Paiwan) cluster together on one side of the continental East Asians, while those from the east coast (Ami, Puyuma) and Orchid Island (Yami) are more heterogeneous. This correspondence with

geography is confirmed by a high and significant correlation between genetic and geographic distances among the 48 populations ($r=0.279$, $P<0.001$).

If we now consider the linguistic information in Figure 16.1A, populations belonging to one linguistic group tend to cluster together. However, in many instances linguistic diversity is not paralleled by genetic differentiations. For example, three Siberian populations speaking languages belonging to different linguistic phyla: Yupik (Eskimo-Aleut) and Chukchi and Koryak (both Chukchi-Kamchatkan), are nevertheless genetically similar (non-significant F_{ST}s between Chukchi and the other two). Conversely, the Austronesian-speaking populations are genetically highly heterogeneous, despite their linguistic relatedness.


A
Koryak
Yupik
Chukchi
Ami
Atayal
Tsou
Saisiat
Rukai
Paiwan
Bunun
Paiwan
Puyuma
Malay
Filipino
Ivatan
Yami
Molucca
Nusa Tenggara
Java
-1.5
-2.0
AN (EF)
AL
AU
CK
KA
GI
IE-UY
ST
EA
AN (TA)

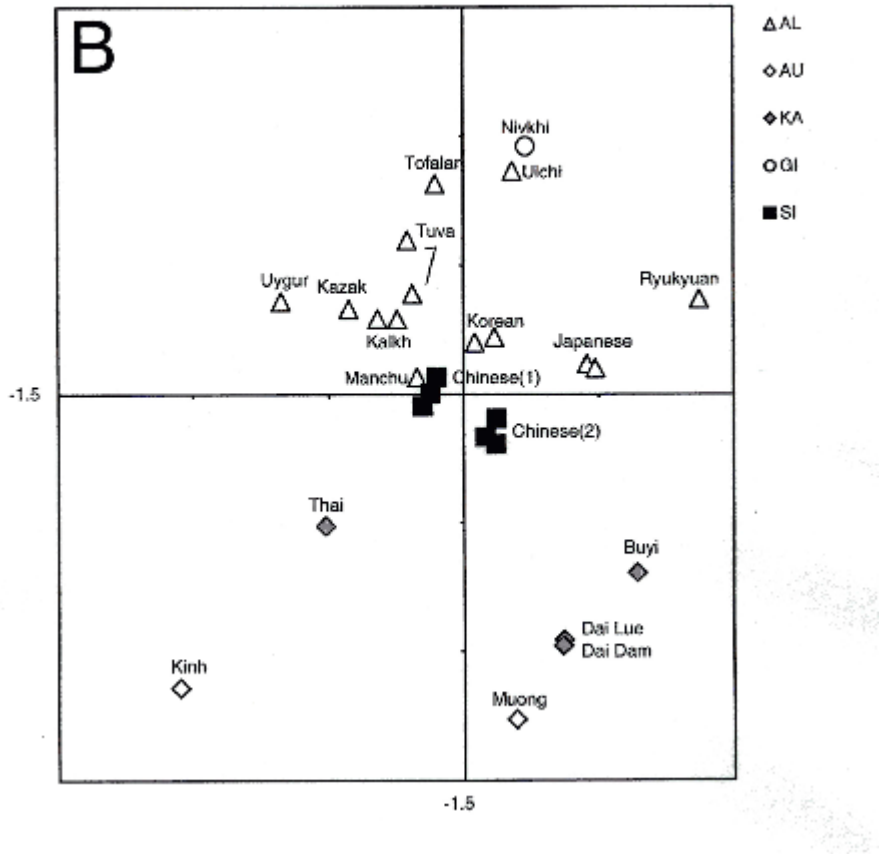


Figure 16.1. Multidimensional analysis (MDS) of Asian populations based on the HLA-DRB1 allelic polymorphism.

A: 48 populations, stress value=0.267; B: 27 populations from the dotted box of Figure 16.1A (excluding 3 IE-UY and AN populations), stress value=0.291.

AL: Altaic; AN (EF): Austronesian (Extra-Formosan); AN (TA): Austronesian (from Taiwan); AU: Austroasiatic; CK: Chukchi-Kamchatkan; EA: Eskimo-Aleut; GI: Gilyak; IE-UY: Indo-European and Uralic-Yukaghir; KA: Tai-Kadai; SI: Sinitic; ST: Sino-Tibetan (here only Sinitic). Chinese (1): Northern Chinese; Chinese (2): Southern Chinese.

Non-significant F_{ST}s (1% level) are found for the following population pairs (see Table 16.1 for population numbers): 1-7, 1-11, 1-13, 2-3, 2-4, 6-14, 7-9, 7-10, 7-11, 8-21, 8-23, 9-10, 9-12, 9-13, 12-13, 13-15, 19-20, 21-22, 21-23, 22-23, 24-26, 25-26, 29-30, 33-34, 33-37, 34-37, 34-38, 39-40, 39-43, 40-42, 40-43, 40-45, 41-42, 41-44, 42-44.

To further investigate the genetic relationships among continental East Asian populations, we performed a second MDS (Figure 16.1B, stress=0.291) of 27 populations among those projected in the centre (dotted box) of Figure 16.1A. A general correspondence with geography is again observed, as Nivkhi and Altaic segregate at the top, Chinese at the centre, and Southeast Asians at the bottom, with only a few exceptions. Also matching relative geographic locations are the significant differentiation of Japanese and Ryukyuans, the close genetic relationship of Ulchi and Nivkhi (both located in northeast Russia close to Sakhalin), and the close genetic relationship of northern Chinese and Manchu. Nevertheless, a few examples contradict those findings: the Vietnamese Kinh and Muong are geographically close but genetically distant. The same is true of the Buyi and Chinese in southern China. When linguistic information is taken into account (symbols in Figure 16.1B), we note that, as in Figure 16.1A, linguistic groups do not overlap substantially: this indicates a relatively fine-grained correspondence between genetic and linguistic relationships.

Correlations between genetic, geographic, and linguistic distances

We attempted to evaluate the contribution of linguistics and/or geography in the genetic structure of East Asian populations. To this end, we statistically compared genetic, geographic and linguistic distance matrices computed on an identical population data set of 40 populations in our data. Linguistic distances were computed from the linguistic phylogeny shown (as explained above) in Figure

16.2. Correlation coefficients and the results of two-way and three-way Mantel tests between the three matrices are presented in Table 16.3.

Table 16.3. Correlation coefficients among genetic (GEN), geographic (GEO), and linguistic (LING) distances in East Asia

	Group size	$r_{GEN,GEO}$ [a,b]	$r_{GEN,LING}$ [a,c]	$r_{GEO,LING}$ [a]
All populations	40	0.131* (0.137*)	0.015 [n.s.] (-0.042 [n.s.])	0.401***
Continental East Asians	25	0.352*** (0.288***)	0.333*** (0.264***)	0.270***
Austronesians only	15	0.340* (0.357*)	0.110 [n.s.] (0.158 [n.s.])	-0.112 [n.s.]

[a] *** : P < 0.001; * : 0.01 < P < 0.05; n.s. : not significant. [b] In parentheses: partial correlation coefficient between genetics and geography controlled for covariation with linguistics. [c] In parentheses: partial correlation coefficient between genetics and linguistics controlled for covariation with geography.

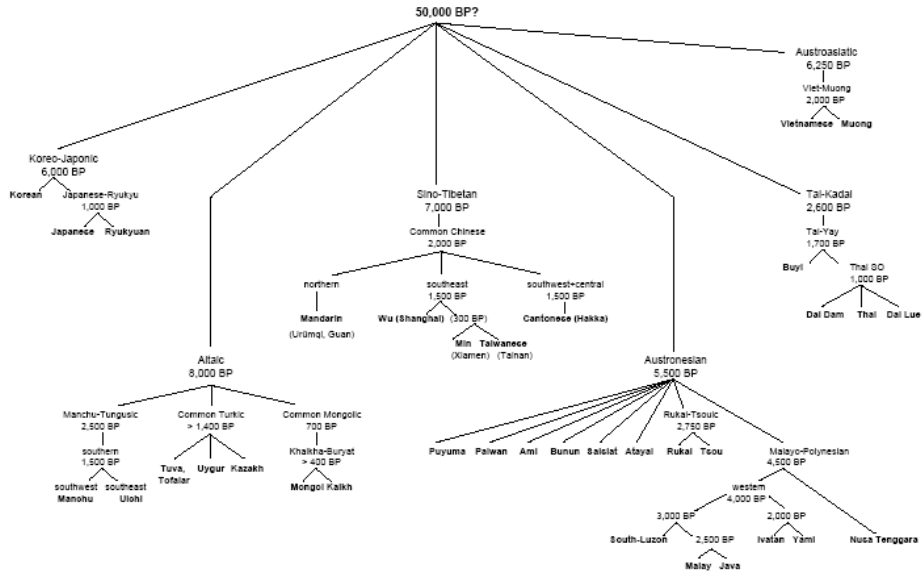


Figure 16.2. The "least controversial "phylogeny of the 40 East Asian languages (bold) used in the correlation analysis (Table 16.3). This tree has been reconstructed on the basis of linguistic and archaeological information (see text). Absolute divergence dates are given in years before present (BP).

The correlations between genetics, on the one hand, and geography or linguistics, on the other hand, are not significant ($r=0.131^{n.s.}$ and $r=-0.022^{n.s.}$, respectively)[3] when we include the 40 populations. Results do not differ substantially when the covariation with the third matrix is taken into account ($r=0.155$*, and $r=-0.085^{n.s}$, respectively). Conversely, a high and very significant correlation coefficient is found between geographic and linguistic distances ($r=0.415$***). We conclude that linguistic families are well differentiated geographically in East Asia, but that this structure does not match the genetic structure.

Different results are obtained when Austronesian populations are considered separately from continental East Asians ("non-Austronesians"). As the Austronesian group was found to be genetically highly heterogeneous (Figure 16.1A), we recomputed correlation coefficients for continental East Asians and Austronesians independently (Table 16.3, lines 2 and 3, respectively). For continental East Asians, a high and significant correlation is found between all pairs of distance matrices (genetics-geography, genetics-linguistics, and geography-linguistics), even when covariation with the third matrix is taken into account (0.225***<r<0.450*** for all coefficients). For Austronesians, the correlation between genetics and geography is also high, although less significant (r=0.340* to 0.361*), but neither genetics nor geography are correlated with linguistics (r=0.146$^{n.s}$ to r=0.194*, for genetics, and r=-0.105 $^{n.s}$, for geography). These results suggest that continental East Asians and Austronesians followed very different modes of evolution, as discussed below.

Genetic diversity within and among linguistic groups

We further conducted ANOVA analyses on the 40 populations considered above, in order to assess the levels of genetic diversity within (Table 16.4) and among (Table 16.5) linguistic families. Table 16.4 indicates that the Austronesian group is the most diverse genetically (high F_{ST} of 9.72%). Among Austronesians, Extra-Formosans are slightly more diversified than Taiwan Aborigines (Formosan) (F_{ST} =8.98% and 7.1%, respectively). At the opposite, the Chinese (Sinitic) are the

most homogeneous group (low F_{ST} of 0.5%), and the Altaic, Tai-Kadai and Austroasiatic groups are intermediate between the Chinese and the Austronesians (F_{ST} =1.7%, 2.39%, and 5.38%, respectively). For each linguistic family, we also estimated the average gene diversity (*h*) within populations. The lowest value is observed in the Austronesians (0.815), and the highest in the Altaic (0.941) and Chinese (0.931). These statistics are plotted in Figure 16.3, together with the average number of detected alleles (*n*) in each linguistic group. A drastic reduction in the number of alleles is observed in Austronesians (mostly Formosan), while this number is above 25 in Sinitic and Altaic (both Koreo-Japonic and Altaic proper). We checked that these results are not due to a sample size effect (not shown).

Thus, from a genetic point of view, the Austronesians represent a highly heterogeneous group of homogeneous populations (pattern C in Table 16.2), while the Chinese, and, to a lesser extent, the Altaic, represent homogeneous groups of heterogeneous populations (pattern B in Table 16.2). The other linguistic groups show intermediate characteristics.

Table 16.4. Amounts of HLA-DRB1 genetic diversity observed among (F_{ST}) and within (h) populations within each East Asian linguistic group

Linguistic group	Group size[a]	F_{ST} (%)[b]	*h* (s.d.)[c]
Altaic	14	1.70***	0.941 (0.016)
- Altaic proper	*9*	*1.27****	*0.946 (0.014)*
- Koreo-Japonic	*5*	*0.88****	*0.932 (0.016)*
Sinitic	6	0.50***	0.931 (0.001)

Tai-Kadai	4	2.39***	0.906 (0.007)
Austroasiatic	2	5.38***	0.891 (0.019)
Austronesian	16	9.72***	0.815 (0.054)
- *Formosan*	*9*	*7.10****	*0.845 (0.042)*
- *Extra-Formosan*	*7*	*8.98****	*0.776 (0.055)*

[a] See Table 16.1 for a list of populations and linguistic groups. [b] ***: $P < 0.001$. [c] The gene diversity has been averaged over the corresponding number of populations; s.d.: standard deviation.

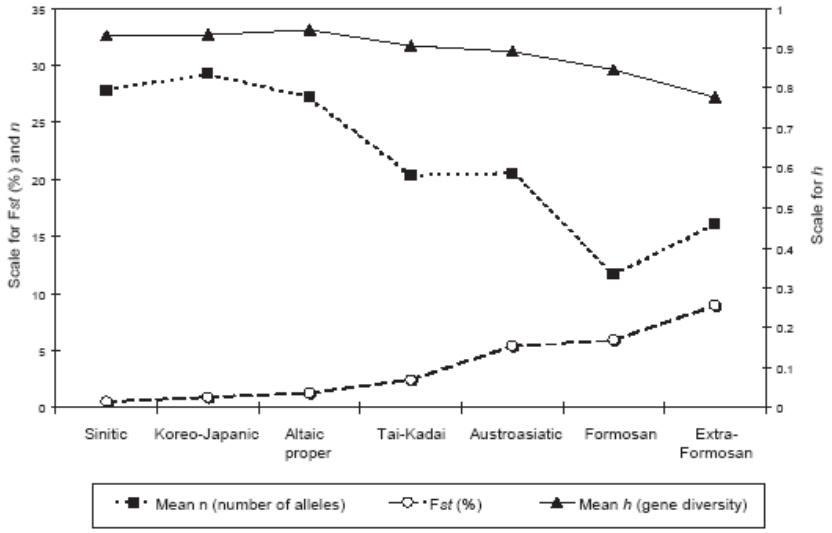


Figure 16.3. Genetic diversity within (*h*) and among (F_{ST}) populations (see Table 16.4), and mean number of HLA-DRB1 alleles detected (*n*) within the main East Asian linguistic groups considered in this study. Group sizes: Sinitic=6, Koreo-Japonic=5, Altaic proper=9, Tai-Kadai=4, Austroasiatic=2, Formosan=9, Extra-Formosan=7.

Finally, linguistic families were compared genetically two by two[4] (Table 16.5). In only one case does the genetic diversity due to a difference *between groups* (F_{CT}) exceed the genetic diversity due to a difference *between populations within groups* (F_{SC}). This case is the pair Altaic proper – Koreo-Japonic (14.5% for F_{CT}, against 1.0% for F_{SC}). At the opposite, Tai-Kadai is not significantly differentiated from Austronesian (F_{CT} =0.5%, not significant), whereas a high level of differentiation is found between populations within these groups (F_{SC} =7.8%). Also, Altaic and Sinitic, and, to a lesser extent, Sinitic and Austronesian, are only weakly differentiated (0.01<P<0.05). In fact, several Altaic and northern Chinese populations are genetically undifferentiated from each other, as indicated by non-significant F_{ST}s (see Figure 16.1, legend). The remaining pairs (Altaic – Tai-Kadai, Altaic – Austronesian, Sinitic – Tai-Kadai, and Taiwan Aborigines (Formosan) – Extra-Formosan) exhibit highly significant differentiations both among groups and among populations within groups.

Table 16.5. Amounts of HLA-DRB1 genetic diversity observed among linguistic groups (F_{CT}), and among populations within linguistic groups (F_{SC}) in East Asia

Linguistic groups[a]			F_{CT} (%)[b]	F_{SC} (%)
Altaic	*versus*	Sinitic	0.6*	1.3***
		Tai-Kadai	1.8***	1.8***
		Austronesian	3.5***	3.7***
Sinitic	*versus*	Tai-Kadai	1.4**	1.0***
		Austronesian	2.4*	4.9***
Tai-Kadai	*versus*	Austronesian	0.5 [n.s.]	7.8**

Altaic proper	*versus*	Koreo-Japonic	14.5***	1.0***
Formosan	*versus*	Extra-Formosan	4.8**	7.3***

[a] Group sizes: Altaic=14, Sinitic=6, Tai-Kadai=4, Austronesian=16, Altaic proper=9, Koreo-Japonic=5, Formosan=9, Extra-Formosan=7. Linguistic groups with less than 4 populations represented have been excluded. [b] ***: P < 0.001; **: 0.001 < P < 0.01; *: 0.01 < P < 0.05; n.s.: not significant.

DISCUSSION

Linguistic hypotheses considered through HLA genetic analyses

This study reveals complex relationships between the processes of genetic and linguistic differentiations in East Asia. At first sight, HLA genetic diversity among populations is geographically structured, as found with most classical systems (Cavalli-Sforza *et al.* 1994; Dugoujon *et al.* forthcoming) as well as with some DNA polymorphisms (Karafet *et al.* 2001). In some instances, however, the observed genetic patterns fit specific linguistic relationships, and may lend some support to one of several competing linguistic hypotheses.

The Altaic family

In mainland Asia, a highly significant differentiation of populations into Turkic-Mongolic-Manchu-Tungusic- (Altaic proper), on the one hand, versus Koreo-Japonic, on the other hand, is observed (Table 16.5). This is consistent with theories that view the two groups as unrelated, or as distantly related within a large macro-Altaic or Eurasiatic phylum. The Altaic-proper group itself does not exhibit a clear genetic subdivision into linguistic families (Turkic, Mongolic,

Manchu-Tungusic). For example, Ulchi (a Tungusic population) are genetically closer to Nivkhi (a Gilyak linguistic isolate) than to Manchu (also Tungusic), and the Mongolian Kalkh are very close to some Turkic populations (Kazakh, Tuva) (Figure 16.1B). Gene flow between neighbouring populations depending on specific environments (like steppe or mountainous areas) as well as language shifts due, for example, to territorial invasions by dominant empires (like shift of Buryat from Turkic to Mongolian, following the Mongol invasion (Pakendorf *et al.* 2003)) probably resulted in intricate relationships between genetic and linguistic patterns. A remarkable result of our study is the very high level of internal genetic diversity (*h*) observed within Altaic (mainly Altaic-proper) populations (Plate VIII, Table 16.4). It indicates that intensive contacts among populations and/or with external groups played a significant role in the evolution of this family. The genetic legacy of multiple migrations or re-colonisations (e.g. Turks, Mongols) in Northeast Asia would thus be reflected in the present heterogeneous Altaic profiles.[5]

The Altaic-Sinitic linguistic border

Compared to Altaic, Chinese populations exhibit more homogeneous HLA-DRB1 genetic profiles (Plate VIII, Table 16.4). Nonetheless, significant differences are found between northern and southern Chinese; northern Chinese are genetically undifferentiated from several Altaic populations (Manchu and Mongolian). Following Hashimoto (1986), some linguists consider that northern Chinese dialects have been "altaicized" due to recurring episodes of domination of

northern China by Altaic-speaking peoples, especially Mongolians and Manchus, followed by shift to Chinese of large numbers of these speakers. Such gene flow into Chinese would then explain the genetic closeness between Altaic and northern Sinitic speakers; the observed genetic differentiation between northern and southern Chinese populations would be a direct consequence of Altaic influence in the north, or of different influences on northern and southern Chinese.

Northern-southern East Asian differentiation

Geneticists are currently debating patterns of genetic differentiation between northern and southern East Asian populations. According to one view, all East Asian populations share a unique origin in mainland Southeast Asia, with a further migration to the north (Su *et al.* 1999, Jin and Su 2000). Revised versions of this theory state that northern populations differ genetically from those located further south due to late genetic contributions from Central Asia (Chu *et al.* 1998, Karafet *et al.* 2001, Jin and Su 2000), but the time and magnitude of these contributions are not clear. A second view, known as the "pincer model", more explicitly invokes two independent migrations into East Asia along a southern and a northern route, with the influence from the Central Asian gene pool being predominant in the north (Ding *et al.* 2000).

In the present study, we find a high correlation between genetic and geographic distances (Table 16.3) and a continuous pattern of genetic differentiation, which roughly follows a north-south geographic axis (Figures 16.1A and 16.1B).

Actually, this pattern suggests isolation by distance (as proposed by Ding *et al.* 2000). It is compatible with both models mentioned above. However, the HLA genetic profiles of northern populations (as discussed above for Altaic) are more diverse (higher *h*) than those of southern populations, which is also true for RH and GM (Sanchez-Mazas 1990, Poloni *et al.*, this volume). This argues against the hypothesis of a unique southern origin whereby northern genetic profiles are a subset of southern ones. Also, Northeast Asians are genetically closer to the Indians and Mansi on the western edge of the region (Figure 16.1A). Such genetic continuity is in keeping with historical relationships established along a northern route.

Proto-East-Asian and Austric hypotheses

At the southern edge of China, populations are in genetic continuity (albeit with large genetic distances) with Tai-Kadai and, to a lesser extent, with Austroasiatic speakers, on the one hand, and Taiwan Aborigines (except the Amis), on the other (Figures 16.1A and 16.1B, and Table 16.5). Such results do not favour any specific linguistic hypothesis linking Sino-Tibetan to Southeast Asian linguistic phyla (e.g. the Sino-Austronesian hypothesis advocated by Sagart in this volume), or the "Greater Austric" hypothesis proposed by Benedict (1942), Ruhlen (1987) and Peiros (1998). On the other hand, Benedict (1942) argues that Tai-Kadai and Austronesian subgroup together as the two branches of the Austro-Tai phylum, and Sagart (2001 and this volume) claims that Tai-Kadai is originally an Austronesian language from Taiwan having been partly relexified by a Southeast

Asian language. Here we show that Tai-Kadai and Austronesian populations do not differ significantly from each other for HLA (Table 16.5), in agreement with both these theories.[6] Moreover, in Sagart's view, the Austronesian language ancestral to Tai-Kadai belongs to a primary branch of Proto-Austronesian to which all extra-Formosan languages and Amis also belong: it is noteworthy, from this point of view, that high frequencies of allele *1502 characterise that set of populations (Tai-Kadai, Amis, and Extra-Formosan Austronesian, in addition to Puyuma, but not central-mountain Taiwanese), as discussed below. Unfortunately, a deeper analysis of Southeast Asian population relationships is not possible due to a lack of representative samples in our data (e.g. only two Austroasiatic, no Hmong-Mien, and no Tibeto-Burman populations are represented). At this point we can simply state that Southeast Asian populations are highly differentiated from each other compared to "continental" groups located further north, in agreement with results recently obtained for mtDNA (Oota *et al.* 2002).

The remarkable HLA diversity of Taiwan Aborigines

The results of a recent HLA analysis of nine aboriginal tribes from Taiwan (Chu *et al.* 2001; Lin *et al.*, this volume), are worth discussing in detail. As shown above (Figure 16.1A), a rough distinction can be drawn between the aboriginal populations from the central mountains (Atayal, Tsou, Saisiat, Rukai, Paiwan and Bunun) and those located on or off the East coast (Amis, Puyuma, and Yami) of the island. In fact, the coastal populations are genetically highly heterogeneous. The Yami live on Orchid (or Lan Yu) Island where they probably settled after

migrating from Bataan Islands (northern Philippines), since their language is closely related to the Batanic languages of the northern Philippines (Chen 1967). This isolation would explain their genetic divergence from other Taiwanese and Extra-Formosan populations. The Puyuma live together with native Chinese in a few villages around the city of T'ai-Tung (Chen 1967), raising the possibility of some degree of gene flow, as recently suggested by mtDNA analyses (Trejaut *et al.* forthcoming); although this may be the case generally for Taiwan Aborigines living in the plains.

The Amis exhibit a highly peculiar HLA-DRB1 genetic profile (Plate VIII) with high frequencies for some uncommon alleles (*0404, *0405) and a very reduced genetic repertoire (only 7 alleles detected, compared to averages of 20.6 for the total set 48 populations and between 11.6 and 29.2 in the different linguistic groups, see Figure 16.3). The uniqueness of the Amis is observed with several other genetic systems: for example, haplotypes R^1 of the Rhesus system, and GM*1,3;5* of the immunoglobulin-associated Gm polymorphism, reach frequencies of 90% and 95%, respectively, in this population (Lin and Broadberry 1998, Sewerin *et al.* 2002, Schanfield *et al.* 2002). Schanfield *et al.* (2002) explain the GM result by a selective effect linked to resistance to malaria in lowland populations; Lin and co-workers suggest a relationship between the Amis and both Papuans from New Guinea and Australian Aborigines, based on a common segregation of these populations in HLA neighbour-joining trees (Lin *et al.* 2000, Lin *et al.,* this volume, Chu *et al.* 2001), and Sewerin *et al.* (2002) notice a genetic

similarity, based on 6 point mutation loci, between the Amis and native Americans.

Our own explanation is that the Amis underwent a founder effect and rapid genetic drift due to isolation. Mabuchi (1954) suggests that a group ancestral to them and to the Ketagalans and Kavalans migrated from the west to the east coast of Taiwan, with an intermediate period of settlement on an undetermined island off the East Coast. This period of isolation could correspond to a bottleneck leading to an impoverishment of their genetic repertoire and accidental genetic convergences with genetically homogeneous populations from other continents. On the other hand, isolation at higher altitudes caused the Aboriginal populations located in the central mountains to evolve independently from coastal tribes, also through rapid genetic drift.

Genetic drift and linguistic variation along the route to the Pacific

Two main alternative models have been proposed to explain the expansion of the Austronesians in the Pacific. These models are popularly known as “the express-train to Polynesia”, proposing a rapid expansion from Taiwan (Bellwood 1978, Diamond 1988), and “the entangled bank” (Terrell 1988), assuming a more complex history of interactions between Polynesians, Melanesians and Southeast Asian populations. The first model has been supported by mtDNA, nuclear DNA, and HLA genetic studies (Melton *et al.* 1995, 1998, Zimdahl *et al.* 1999), but the speed of the Austronesian spread across the Pacific may not have been as fast as

previously assumed: according to the “slow-boat” hypothesis (Kayser *et al.* 2000), various contact phenomena occurred between Polynesian ancestors and Melanesians. Despite recent alternative views (Hurles *et al.* 2002, Jin and Su 2000, Richards *et al.* 1998, Su *et al.* 2000), both scenarios agree that the Austronesian expansion started in Taiwan and/or nearby East Asia about 5-6,000 years ago, and reached Polynesia by migrating southwards and eastwards through the Philippines, Indonesia, and coastal and islands Melanesia.

If we consider this migration theory, the tentative scenario proposed below would account for the observed HLA-DRB1 allelic distribution shown in Plate VIII.

The Proto-Austronesians, originating somewhere in mainland China, were characterized by relatively even HLA allelic distributions (i.e. rather low frequencies for all alleles), as currently observed in continental East Asia.

Migration of these proto-Austronesians to Taiwan was followed by a main differentiation between coastal and central mountain tribes. The central mountain tribes virtually lost allele *1502 and acquired high frequencies of alleles *1202 and *0803 by genetic drift. At the same time, the East coast tribes, here represented by the Amis and the Puyuma, acquired a higher frequency of allele *1502. The Amis diverged to a greater extent due to a bottleneck, loosing allele *1202 and acquiring a high frequency of *0404, very rare elsewhere.

The Extra-Formosans began to differentiate from the rest of the Austronesians on the east coast, where *1502 was frequent, and the frequency of this allele increased rapidly during their migrations southwards to Indonesia, eastwards to the Pacific, and also back to the mainland where they would form the first Tai-Kadai nucleus.

Overall, as also concluded for Amerindians on the basis of HLA studies (Monsalve *et al.* 1999, Sanchez-Mazas forthcoming), Austronesian populations would have experienced rapid differentiations through genetic drift, both within Taiwan, and as soon as they expanded away from this island. Contrary to continental East Asians, the Austronesian genetic pool is indeed characterized by high genetic variation *among*, and low genetic variation *within* populations (pattern C in Table 16.2, Table 16.4 and Figure 16.3), as well as a low number of detected HLA-DRB1 alleles (n=13.6 in average) compared to other linguistic groups (from 20.3 to 29.2). Their allelic repertoire has thus been impoverished during successive founder effects, in agreement with the hypothesis of isolation and migrations in insular environments. It is even more reduced in Island Melanesia, with only 9-14 alleles detected at the DRB1 locus (Hagelberg *et al.* 1999, Zimdahl *et al.* 1999), indicating that the founder effect / genetic drift processes continued during the colonisation of the Pacific. Moreover, the dispersal of small and endogamous[7] population groups across insular Southeast Asia (and the Pacific) may explain why so many different languages are identified within the Austronesian phylum (some 1,262 languages (Grimes and Grimes 2000)[8] –

about one fifth of the total number of human languages). Supposing that genetic and linguistic change are random and independent processes, genetic variation presumably ceased to reflect linguistic relationships in this area. On the other hand, gene flow maintained a certain amount of genetic relatedness among neighbouring populations without necessarily affecting linguistic affinities, as shown by Zimdahl *et al.* (1999) for some populations of the Solomon Islands. This may be the reason why geography still explains 12% of the Austronesian genetic variation, and linguistics less than 2% (according to determination coefficients (Sokal and Rohlf 1994) estimated by the square of the correlation coefficients given in Table 16.3), while geographic and linguistic differentiation explains an equivalent amount of genetic variation in continental East Asia (r^2=12% and 11%, respectively).

Inferring mechanisms of population differentiation within language families

In this study, we have sought to emphasize the role of different evolutionary mechanisms in shaping the patterns of genetic variation within linguistic families. F_{ST} and *h* were taken as two complementary measures of genetic diversity (*among* and *within* populations, respectively) allowing the description of four different patterns from which evolutionary processes might be inferred (Table 16.2). For each linguistic family under study, we reported these two statistics on a graph, together with the average number of detected alleles (*n*) (Figure 16.3). This approach allowed us to identify two main patterns in the HLA-DRB1 data: pattern C (high F_{ST} and low *h*), observed in Austronesians, was explained by genetic drift;

pattern B (low F_{ST} and high *h*) was observed in Sinitic and Altaic, where either intensive gene flow occurred among populations, or each group as a whole underwent a recent differentiation from a highly diversified population. Different mechanisms can also be inferred for the remaining patterns (A and D), although they were not observed in our study. Pattern D (low F_{ST} and low *h*, or "genetic undifferentiation") can be due either to a very recent common origin, or to intensive gene flow between populations. Pattern A (high F_{ST} and high *h*, or "high differentiation / high diversity") can suggest at least two contrasting explanations: a remote ancestry of populations, with large population sizes and reduced gene flow maintaining genetic diversity within and among populations, respectively, or intensive gene flow from external and genetically diverse populations. The latter situation could occur, for example, at different linguistic boundaries of a given family (e.g. the Altaic and Tai-Kadai on the boundaries of Chinese).

The approach described above should prove very useful in investigating the evolution of populations belonging to different linguistic phyla, given that numerous populations are tested in each group. Moreover, an application of this method to several independent genetic systems (such as RH and GM) or the different HLA loci would allow us to distinguish between patterns resulting from the genetic history of populations – if congruent results are obtained for all systems - and those resulting, for example, from selective effects, or even from methodological biases – if discordant results are obtained (Sanchez-Mazas *et al.* 2003).

CONCLUSION

In this study, we used a large set of molecular HLA-DRB1 data (48 populations represented by 6,613 individuals) to investigate the genetic structure of East Asian populations in relation to some currently debated linguistic hypotheses. We looked at the data through several complementary statistical approaches (correlation analysis between genetic, geographic and linguistic distances, F_{ST} significance among populations, analysis of variance across linguistic groups, and multidimensional scaling analysis) not with a view to *prove* or *disprove* linguistic hypotheses, but to explore the compatibility between genetic and linguistic relationships in the continent.

While the HLA polymorphism reveals a complex genetic structure in East Asian, and especially Austronesian, populations, some of our findings confirm, or support, various aspects of linguistic classification. First, although Japanese, Korean and Altaic proper (Mongolic, Manchu-Tungusic and Turkic) are included by some authors into such macrophyla as Altaic and Eurasiatic, few, if any, regard Koreo-Japonic and Altaic proper as linguistically very close. Not surprisingly, we observe a major genetic differentiation between Koreo-Japonic on the one hand and Altaic-proper on the other hand. Second, we find a high degree of genetic proximity between populations on both sides of the Altaic-Sinitic linguistic boundary, paralleling the linguistic evidence for 'altaicization' of northern Chinese (Hashimoto 1986). While very few linguists would argue for a genetic connection

between Chinese and Altaic, the Altaic features in northern Chinese dialects clearly are of the type resulting from imperfect learning of Chinese by Altaic speakers, suggesting that Altaic speakers in northern China have been shifting to Chinese *en masse* in historical times: our genetic observations support Hashimoto's altaicization hypothesis. Third, we find evidence of a genetic continuity between Austronesian, especially extra-Formosan, and Tai-Kadai. This finding is compatible with the hypothesis of an Austronesian origin of Tai-Kadai. Fourth, the results of the present investigation are congruent with a Taiwanese homeland of Austronesian: we propose a tentative historical scenario for the Austronesian expansion, in which Extra-Formosan originated on the east coast of Taiwan. We also drew a parallel between the high level of genetic differentiation among Austronesians and their high number of different languages, both probably resulting independently from the rapid dispersal of small population groups in an island environment. Finally, we have proposed a simple but efficient way of inferring the modes of evolution of different linguistic families through the computation of two statistics. This has also allowed us to contrast the evolutions of continental East Asians and insular peoples. Of course, as in other disciplines, the conclusions reached by genetic studies strongly depend on the quantity and quality of the data and on the methods used. Our present interpretation of the HLA-DRB1 polymorphism in East Asia should therefore be considered tentative.

ACKNOWLEDGEMENTS

This work was supported by the French CNRS (OHLL grant to ESP and LS) and the Swiss FNS (grant 3100-49771.96) to ASM.

BIBLIOGRAPHY

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

Barbujani, G. (2000) 'Geographic patterns: how to identify them and why', *Human Biology*, 72**:**133-53.

Barbujani, G. and Pilastro, A. (1993) 'Genetic evidence on origin and dispersal of human populations speaking languages of the Nostratic macrofamily', *Proceedings of the National Academy of Sciences USA*, 90**:**4670-3.

Barbujani, G., Sokal, R.R. and Oden, N.L. (1995) 'Indo-European origins: a computer-simulation test of five hypotheses', *American Journal of Physical Anthropology*, 96**:**109-32.

Bellwood, P. (1978) *Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania*, Oxford: Oxford University Press.

Bellwood, P. (2001) 'Early agriculturalist population diasporas? Farming, languages, and genes', *Annual Review of Anthropology*, 30**:**181-207.

Benedict, P.K. (1942) ‘Thai, Kadai and Indonesian: a new alignment in Southeastern Asia’ *American Anthropologist, n.s.*, 44, 576-601.

Bugawan, T.L., Chang, J.D., Klitz, W. and Erlich, H.A. (1994) 'PCR/oligonucleotide probe typing of HLA class II alleles in a Filipino population reveals an unusual distribution of HLA haplotypes', *American Journal of Human Genetics*, 54**:**331-40.

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

Chandanayingyong, D., Stephens, H.A., Klaythong, R., Sirikong, M., Udee, S., Longta, P., Chantangpol, R., Bejrachandra, S. and Rungruang, E. (1997) 'HLA-A, -B, -DRB1, -DQA1, and -DQB1 polymorphism in Thais', *Human Immunology*, 53**:**174-82.

Chen, K.-C. (1967) *Taiwan Aborigines, a genetic study of tribal variations*, Cambridge, Massachusetts: Harvard University Press.

Chikhi, L., Nichols, R.A., Barbujani, G. and Beaumont, M.A. (2002) 'Y genetic data support the Neolithic demic diffusion model', *Proceedings of the National Academy of Sciences USA*, 99**:**11008-13.

Chimge, N.O., Tanaka, H., Kashiwase, K., Ayush, D., Tokunaga, K., Saji, H., Akaza, T., Batsuuri, J. and Juji, T. (1997) 'The HLA system in the population of Mongolia', *Tissue Antigens*, 49**:**477-83.

Chu, C.C., Lin, M., Nakajima, F., Lee, H.L., Chang, S.L., Juji, T. and Tokunaga, K. (2001) 'Diversity of HLA among Taiwan's indigenous tribes and the Ivatans in the Philippines', *Tissue Antigens*, 58**:**9-18.

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 Science USA*, 95**:**11763-8.

Diamond, J.M. (1988) 'Express train to Polynesia', *Nature*, 336**:**307-308.

Ding, Y.-C., Wooding, S., Harpending, H.C., Chi, H.-C., Li, H.-P., Fu, Y.-X., Pang, J.-F., Yao, Y.-G., Xiang Yu, J.-G., Moyzis, R. and Zhang, Y.-P. (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) 'The GM haplotype diversity of 82 populations over the world suggests a centrifugal model of human migrations', 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 barriers on population differentiation: application to the Afro-Asiatic/Indo-European case', *European Journal of Human Genetics*, 10**:**750-6.

Excoffier, L. (2001) 'Analysis of population subdivision', in D.J. Balding, M. Bishop and C. Cannings (eds) *Handbook of Statistical Genetics*, Chichester: John Wiley & Sons, LTD.

Excoffier, L., Pellegrini, P., Sanchez-Mazas, A., Simon, C. and Langaney, A. (1987) 'Genetics and history of Sub-Saharan Africa', *Yearbook of Physical Anthropology*, 30**:**151-94.

Gao, X., Zimmet, P. and Serjeantson, S.W. (1992) 'HLA-DR,DQ sequence polymorphisms in Polynesians, Micronesians, and Javanese', *Human Immunology,* 34**:**153-61.

Gao, X.J., Sun, Y.P., An, J.B., Fernandez-Vina, M., Qou, J.N., Lin, L. and Stastny, P. (1991) 'DNA typing for HLA-DR, and -DP alleles in a Chinese population using the polymerase chain reaction (PCR) and oligonucleotide probes', *Tissue Antigens,* 38**:**24-30.

Grahovac, B., Sukernik, R. I., O'hUigin, C., Zaleska-Rutczynska, Z., Blagitko, N., Raldugina, O., Kosutic, T., Satta, Y., Figueroa, F., Takahata, N. and Klein, J. (1998) 'Polymorphism of the HLA class II loci in Siberian populations', *Human Genetics,* 102**:**27-43.

Grimes, B.F. and Grimes, J.E. (eds) (2000) *Ethnologue: Languages of the World,* Dallas: SIL International.

Hagelberg, E., Kayser, M., Nagy, M., Roewer, L., Zimdahl, H., Krawczak, M., Lió, P. and Schiefenhövel, W. (1999) 'Molecular genetic evidence for the human settlement of the Pacific: analysis of mitochondrial DNA, Y chromosome and HLA markers', *Philosophical Transactions of the Royal Society of London B,* 354**:**141-52.

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

Hashimoto, M., Kinoshita, T., Yamasaki, M., Tanaka, H., Imanishi, T., Ihara, H., Ichikawa, Y. and Fukunishi, T. (1994) 'Gene frequencies and haplotypic associations within the HLA region in 916 unrelated Japanese individuals', *Tissue Antigens*, 44**:**166-73.

Hatta, Y., Ohashi, J., Imanishi, T., Kamiyama, H., Iha, M., Simabukuro, T., Ogawa, A., Tanaka, H., Akaza, T., Gojobori, T., Juji, T. and Tokunaga, K. (1999) 'HLA genes and haplotypes in Ryukyuans suggest recent gene flow to the Okinawa Islands', *Human Biology*, 71**:**353-65.

Hurles, M.E., Nicholson, J., Bosch, E., Renfrew, C., Sykes, B.C. and Jobling, M.A. (2002) 'Y chromosomal evidence for the origins of oceanic-speaking peoples', *Genetics*, 160**:**289-303.

Imanishi, T., Akaza, T., Kimura, A., Tokunaga, K. and Gojobori, T. (1992) 'Allele and haplotype frequencies for HLA and complement loci in various ethnic groups', in K. Tsuji, M. Aizawa and T. Sasazuki (eds) *HLA 1991*, Vol. 1, Oxford: Oxford University Press.

IMGT/HLA sequence database (2003) Online. Available HTTP: http://www.ebi.ac.uk/imgt/hla/intro.html (accessed 27 August 2003).

Jin, L. and Su, B. (2000) 'Natives or immigrants: modern human origin in east Asia', *Nature Reviews Genetics*, 1**:**126-33.

Karafet, T., Xu, L., Du, R.F., 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.

Kayser, M., Brauer, S., Weiss, G., Underhill, P.A., Roewer, L., Schiefenhövel, W. and Stoneking, M. (2000) 'Melanesian origin of Polynesian Y chromosomes', *Current Biology,* 10**:**1237-46.

Kruskal, J. B. (1964) 'Nonmetric multidimensional scaling: a numerical method', *Psychometrika,* 29**:**28-42.

Lee, J. (1997) 'Chinese normal', in Terasaki, P. I. and Gjertson, D. W. (eds) *HLA 1997,* Los Angeles: UCLA Tissue Typing Laboratory.

Lin, M. and Broadberry, R.E. (1998) 'Immunohematology in Taiwan', *Transfusion Medicine Reviews,* 12**:**56-72.

Lin, M., Chu, C.C., Lee, H.L., Chang, S.L., Ohashi, J., Tokunaga, K., Akaza, T. and Juji, T. (2000) 'Heterogeneity of Taiwan's indigenous population: possible relation to prehistoric Mongoloid dispersals', *Tissue Antigens,* 55**:**1-9.

Lou, H., Li, H.C., Kuwayama, M., Yashiki, S., Fujiyoshi, T., Suehara, M., Osame, M., Yamashita, M., Hayami, M., Gurtsevich, V., Ballas, M., Imanishi, T. and Sonoda, S. (1998) 'HLA class I and class II of the Nivkhi, an indigenous population carrying HTLV-I in Sakhalin, Far Eastern Russia', *Tissue Antigens,* 52**:**444-51.

Mabuchi, T. (1954) Takasagozoku no ido oyobi bumpu (part 2), *Minzoku Gaku Kenkyu* 18.4.

Mack, S.J., Bugawan, T.L., Stoneking, M., Saha, M., Beck, H.-P. and Erlich, H.A. (2000) 'HLA class I and class II loci in Pacific/Asian populations', in M. Kasahara (ed.) *Major Histocompatibility Complex, Evolution, Structure, and Function,* Tokyo: Springer Verlag.

Mantel, G. (1967) 'The detection of disease clustering and a generalized regression approach', *Cancer Research*, 27**:**209-20.

Martinez-Laso, J., Sartakova, M., Allende, L., Konenkov, V., Moscoso, J., Silvera-Redondo, C., Pacho, A., Trapaga, J., Gomez-Casado, E. and Arnaiz-Villena, A. (2001) 'HLA molecular markers in Tuvinians: a population with both Oriental and Caucasoid characteristics', *Annals of Human Genetics*, 65**:**245-61.

Melton, T., Peterson, R., Redd, A.J., Saha, N., Sofro, A.S.M., Martinson, J. and Stoneking, M. (1995) 'Polynesian genetic affinities with southeast Asian populations as identified by mtDNA analysis', *American Journal of Human Genetics*, 57:403-14.

Melton, T., Clifford, S., Martinson, J., Batzer, M. and Stoneking, M. (1998) 'Genetic evidence for the proto-Austronesian homeland in Asia: mtDNA and nuclear DNA variation in Taiwanese Aboriginal tribes', *American Journal of Human Genetics*, 63:1807-23.

Meyer, D. (2002) 'Studies on selection and recombination in human Major Histocompatibility Complex genes', unpublished thesis, University of California, Berkeley.

Meyer, D. and Thomson, G. (2001) 'How selection shapes variation of the human major histocompatibility complex: a review', *Annals of Human Genetics*, 65**:**1-26.

Mizuki, N., Ohno, S., Ando, H., Sato, T., Imanishi, T., Gojobori, T., Ishihara, M., Goto, K., Ota, M., Geng, Z., Geng, L., Li, G. and Inoko, H. (1998) 'Major histocompatibility complex class II alleles in an Uygur population in the Silk Route of Northwest China', *Tissue Antigens*, 51**:**287-92.

Mizuki, N., Ohno, S., Ando, H., Sato, T., Imanishi, T., Gojobori, T., Ishihara, M., Ota, M., Geng, Z., Geng, L., Li, G., Kimura, M. and Inoko, H. (1997) 'Major histocompatibility complex class II alleles in Kazak and Han populations in the Silk Route of northwestern China', *Tissue Antigens*, 50**:**527-34.

Monsalve, M.V., Helgason, A. and Devine, D.V. (1999) 'Languages, geography and HLA haplotypes in Native American and Asian populations', *Proceedings of the Royal Society of London B*, 266:2209-16.

Munkhbat, B., Sato, T., Hagihara, M., Sato, K., Kimura, A., Munkhtuvshin, N. and Tsuji, K. (1997) 'Molecular analysis of HLA polymorphism in Khoton-Mongolians', *Tissue Antigens*, 50**:**124-34.

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

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.

Pakendorf, B., Wiebe, V., Tarskaia, L.A., Spitsyn, V.A., Soodyall, H., Rodewald, A. and Stoneking, M. (2003) 'Mitochondrial DNA evidence for admixed origins of central Siberian populations', *American Journal of Physical Anthropology*, 120**:**211-24.

Park, M.H., Kim, H.S. and Kang, S.J. (1999) 'HLA-A,-B,-DRB1 allele and haplotype frequencies in 510 Koreans', *Tissue Antigens*, 53**:**386-90.

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

Rani, R., Fernandez-Vina, M.A. and Stastny, P. (1998) 'Associations between HLA class II alleles in a North Indian population', *Tissue Antigens*, 52**:**37-43.

Renfrew, C. (1992) 'Archaeology, genetics and linguistic diversity', *Man* 27**:**445-78.

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

Richards, M., Oppenheimer, S. and Sykes, B. (1998) 'MtDNA suggests Polynesian origins in Eastern Indonesia', *American Journal of Human Genetics*, 63:1234-6.

Rohlf, F.J. (2000) *NTSYSpc: numerical taxonomy and multivariate analysis system*, New York: Exeter Software.

Ruhlen, M. (1987) *A Guide to the World's Languages. V.1, Classification*, London: Edward Arnold.

Sagart, L. (2001) 'Comment: Malayo-Polynesian features in the AN-related vocabulary in Kadai', paper presented at the International Meeting *Perspectives on the Phylogeny of East Asian Languages*, Périgueux, August 2001.

Saito, S., Ota, S., Yamada, E., Inoko, H. and Ota, M. (2000) 'Allele frequencies and haplotypic associations defined by allelic DNA typing at HLA class I and class II loci in the Japanese population', *Tissue Antigens*, 56**:**522-9.

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

— (2001) 'African diversity from the HLA point of view: influence of genetic drift, geography, linguistics, and natural selection', *Human Immunology*, 62**:**937-48.

— (2002) 'HLA data analysis in anthropology: basic theory and practice', paper presented at the 16th European Histocompatibility Conference of the European Federation for Immunogenetics (EFI), Strasbourg, March 2002.

— (forthcoming) 'HLA genetic diversity of the 13th IHWC population data relative to worldwide linguistic families', to appear in J.A. Hansen and B. Dupont (eds) *HLA 2002: Immunobiology of the Human MHC*, Seattle: IHWG Press.

Sanchez-Mazas, A., Poloni, E., Jacques, G. and Sagart, L. (2003) 'Processus de différenciation des locuteurs des grandes familles de langues est-asiatiques: hypothèses de la génétique', paper presented at the annual meeting of the Société d'Anthropologie de Paris, Museum National d'Histoire Naturelle, Paris, January 2003.

Schanfield, M., Ohkura, K., Lin, M., Shyu, R. and Gershowitz, H. (2002) 'Immunoglobulin allotypes among Taiwan Aborigines: evidence of malarial selection could affect studies of population affinity', *Human Biology*, 74**:**363-79.

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.

Sewerin, B., Cuza, F.J., Szmulewicz, M.N., Rowold, D.J., Bertrand-Garcia, R.L. and Herrera, R.J. (2002) 'On the genetic uniqueness of the Ami aborigines of Formosa', *American Journal of Physical Anthropology*, 119**:**240-8.

Sokal, R.R. and Rohlf, F.J. (1994) *Biometry*, New York:W.H. Freeman and Co.

Su, B., Jin, L., Underhill, P., Martinson, J., Saha, N., McGarvey, S.T., Shriver, M.D., Chu, J., Oefner, P., Chakraborty, R. and Deka, R. (2000) 'Polynesian origins: insights from the Y chromosome', *Proceedings of the National Academy of Sciences USA*, 97**:**8225-8.

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.

Terrell, J. (1988) 'History as a family tree, history as an entangled ban: constructing images and interpretations of prehistory in the South Pacific', *Antiquity*, 62:642-57.

Trejaut, J.A., Loo, J.H., Li, Z.Y., Lee, H.L., Chang, H.L., Chu, C.C. and Lin, M. (forthcoming) 'Mitochondrial DNA diversity in nine Taiwan indigenous tribes', to appear in *Genetics*.

Uinuk-Ool, T.S., Takezaki, N., Sukernik, R.I., Nagl, S. and Klein, J. (2002) 'Origin and affinities of indigenous Siberian populations as revealed by HLA class II gene frequencies', *Human Genetics,* 110**:**209-26.

Vu-Trieu, A., Djoulah, S., Tran-Thi, C., Ngyuyen-Tanh, T., Le Monnier de Gouville, I., Hors, J. and Sanchez-Mazas, A. (1997) 'HLA-DR and -DQB1 DNA polymorphisms in a Vietnamese Kinh population from Hanoi', *European Journal of Immunogenetics,* 24**:**345-56.

Wang, F.Q., Semana, G., Fauchet, R. and Genetet, B. (1993) 'HLA-DR and -DQ genotyping by PCR-SSO in Shanghai Chinese', *Tissue Antigens,* 41**:**223-6.

Weng, Z. and Sokal, R.R. (1995) 'Origins of Indo-Europeans and the spread of agriculture in Europe: comparison of lexicostatistical and genetic evidence', *Human Biology,* 67**:**577-94.

Zimdahl, H., Schiefenhövel, W., Kayser, M., Roewer, L. and Nagy, M. (1999) 'Towards understanding the origin and dispersal of Austronesians in the Solomon Sea: HLA class II polymorphism in eight distinct populations of Asia-Oceania', *European Journal of Immunogenetics*, 26:405-16.

[1] Missing from this list are Tibeto-Burman and Hmong-Mien populations; Austro-Asiatic populations are limited to Kinh (=Vietnamese) and Muong, two closely related languages of the Vietic branch of Mon-Khmer. Our three Tai-Kadai populations come from the Thai branch, and we do not have any Kra populations from south China or Li populations from Hainan. For the Austronesian family, we looked at the Western part, especially Taiwan where the family underwent its primary diversification. There is only one Central Malayo-Polynesian population (Nusa Tenggara) and no Oceanic: this is because we have concentrated our attention on the region where a majority of linguists consider the Austronesian homeland to be located.

[2] These models apply to linguistic families considered *a priori* as monophyletic.

[3] The correlation between genetics and geography drops from 0.279 to 0.131 when 40 instead of 48 populations are considered, probably because of the exclusion of three Northeast Siberian populations (Chukchi, Koryak, Yupik), which are both genetically and geographically very distant from all other populations.

[4] Austroasiatic was not considered because it included only two populations.

[5] The evolution of the HLA polymorphism is possibly influenced by selective pressure maintaining a high level of diversity in human populations (Meyer and Thomson 2001, Meyer 2002). We thus checked the possibility of a departure from selective neutrality in all populations considered in this study, and we found that 7 out of 48 population samples (thus 15%) were significantly deviant towards an excess of heterozygotes at the 1% level, namely Manchu ($P=0.004$), Khalkh ($P=0.008$), Tuvin ($P=0.004$), Japanese (two samples, $P=0.001$ and $P=0.002$), and Koreans (two samples, $P=0.0006$ and $P<0.0001$). As all these populations belong to the Altaic family, it is reasonable to suppose that historical events like gene flow, rather than selection, maintained such high levels of genetic diversity.

[6] Note, however, that this test was not applied to Austroasiatic speakers as these are here represented by only two populations.

[7] Endogamous here refers to the fact that the genetic pool of isolated populations is generally more homogeneous than in outbred populations due to a higher kinship between individuals.

[8] http://www.ethnologue.com/ (accessed July 2003)