HLA genetic diversity and linguistic variation in East Asia

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

To cite this version:


HAL Id: halshs-00104753
https://halshs.archives-ouvertes.fr/halshs-00104753v2
Submitted on 10 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
CHAPTER 16

HLA GENETIC DIVERSITY AND LINGUISTIC VARIATION IN EAST ASIA

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

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. Overall, these 48 populations are represented by a total of 6,613 individuals.
<table>
<thead>
<tr>
<th>#</th>
<th>N</th>
<th>Population</th>
<th>Country</th>
<th>Location</th>
<th>Lat</th>
<th>Long</th>
<th>Language</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68</td>
<td>Mansi</td>
<td>Russia</td>
<td>Khanty-Mansi</td>
<td>60.2</td>
<td>70.7</td>
<td>Uralic</td>
<td>UY</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>Chukchi</td>
<td>Siberia</td>
<td>Several regions</td>
<td>65</td>
<td>185</td>
<td>Chukchi</td>
<td>CK</td>
</tr>
<tr>
<td>3</td>
<td>92</td>
<td>Koryak</td>
<td>Siberia</td>
<td>Kamchatka</td>
<td>60</td>
<td>164</td>
<td>Koryak</td>
<td>CK</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>Yupik</td>
<td>Siberia</td>
<td>Bering coast</td>
<td>66</td>
<td>185</td>
<td>Eskimo</td>
<td>EA</td>
</tr>
<tr>
<td>5</td>
<td>47</td>
<td>Indian</td>
<td>India</td>
<td>North</td>
<td>28.4</td>
<td>77.2</td>
<td>Indo-European</td>
<td>IE</td>
</tr>
<tr>
<td>6</td>
<td>53</td>
<td>Nivkh</td>
<td>Russia</td>
<td>Siberia, Nogliki</td>
<td>51.5</td>
<td>143</td>
<td>Gilyak</td>
<td>GI</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>Kazakh</td>
<td>China</td>
<td>Ürümqi</td>
<td>43.4</td>
<td>87.4</td>
<td>Turkic</td>
<td>AL</td>
</tr>
<tr>
<td>8</td>
<td>160</td>
<td>Manchu</td>
<td>China</td>
<td>Heilongjiang</td>
<td>45.2</td>
<td>126</td>
<td>Tungus</td>
<td>AL</td>
</tr>
<tr>
<td>9</td>
<td>41</td>
<td>Khalkh</td>
<td>Mongolia</td>
<td>Ulaanbaatar</td>
<td>47.5</td>
<td>107</td>
<td>Mongolian</td>
<td>AL</td>
</tr>
<tr>
<td>10</td>
<td>201</td>
<td>Khalkh</td>
<td>Mongolia</td>
<td>Kharkhorum</td>
<td>45</td>
<td>100</td>
<td>Mongolian</td>
<td>AL</td>
</tr>
<tr>
<td>11</td>
<td>57</td>
<td>Uighur</td>
<td>China</td>
<td>Ürümqi</td>
<td>43.4</td>
<td>87.4</td>
<td>Turkic</td>
<td>AL</td>
</tr>
<tr>
<td>12</td>
<td>190</td>
<td>Tuvin</td>
<td>Russia</td>
<td>Kyzyl</td>
<td>51.4</td>
<td>94.3</td>
<td>Turkic</td>
<td>AL</td>
</tr>
<tr>
<td>13</td>
<td>44</td>
<td>Tuvin</td>
<td>Russia</td>
<td>Kyzyl</td>
<td>51.4</td>
<td>94.3</td>
<td>Turkic</td>
<td>AL</td>
</tr>
<tr>
<td>14</td>
<td>73</td>
<td>Ulchi</td>
<td>Russia</td>
<td>Khabarovsk</td>
<td>54</td>
<td>136</td>
<td>Tungus</td>
<td>AL</td>
</tr>
<tr>
<td>15</td>
<td>43</td>
<td>Tofalar</td>
<td>Russia</td>
<td>Nizhneudinsk</td>
<td>54.9</td>
<td>99</td>
<td>Turkic</td>
<td>AL</td>
</tr>
<tr>
<td>16</td>
<td>197</td>
<td>Ryukyuan</td>
<td>Japan</td>
<td>Okinawa</td>
<td>26</td>
<td>127</td>
<td>Ryukyuan</td>
<td>AL</td>
</tr>
<tr>
<td>17</td>
<td>371</td>
<td>Japanese</td>
<td>Japan</td>
<td>Centre</td>
<td>35.4</td>
<td>139</td>
<td>Japanese</td>
<td>AL</td>
</tr>
<tr>
<td>18</td>
<td>916</td>
<td>Japanese</td>
<td>Japan</td>
<td>n.d.(^b)</td>
<td>35.4</td>
<td>139</td>
<td>Japanese</td>
<td>AL</td>
</tr>
<tr>
<td>19</td>
<td>510</td>
<td>Korean</td>
<td>Korea</td>
<td>Seoul</td>
<td>37.3</td>
<td>127</td>
<td>Korean</td>
<td>AL</td>
</tr>
<tr>
<td>20</td>
<td>199</td>
<td>Korean</td>
<td>Korea</td>
<td>Heilongjiang</td>
<td>46</td>
<td>127</td>
<td>Korean</td>
<td>AL</td>
</tr>
<tr>
<td>21</td>
<td>91</td>
<td>Chinese</td>
<td>China</td>
<td>Guan County</td>
<td>39.3</td>
<td>116</td>
<td>Sinitic</td>
<td>SI</td>
</tr>
<tr>
<td>22</td>
<td>89</td>
<td>Chinese</td>
<td>China</td>
<td>Shanghai</td>
<td>31.1</td>
<td>121</td>
<td>Sinitic</td>
<td>SI</td>
</tr>
<tr>
<td>23</td>
<td>59</td>
<td>Chinese</td>
<td>China</td>
<td>Ürümqi</td>
<td>43.4</td>
<td>87.4</td>
<td>Sinitic</td>
<td>SI</td>
</tr>
<tr>
<td>24</td>
<td>162</td>
<td>Chinese</td>
<td>China</td>
<td>Xiamen, Fujian</td>
<td>24.3</td>
<td>118</td>
<td>Sinitic</td>
<td>SI</td>
</tr>
<tr>
<td>25</td>
<td>1012</td>
<td>Taiwanese</td>
<td>Taiwan</td>
<td>Tainan</td>
<td>23</td>
<td>120</td>
<td>Sinitic</td>
<td>SI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Country</td>
<td>Region</td>
<td>N: Sample Size</td>
<td>Lat: Latitude</td>
<td>Long: Longitude</td>
<td>LF: Linguistic Family</td>
<td>Notes</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---------</td>
<td>---------</td>
<td>----------------</td>
<td>---------------</td>
<td>-------------------</td>
<td>----------------------</td>
<td>-------</td>
</tr>
<tr>
<td>26</td>
<td>190</td>
<td>Taiwanese</td>
<td>Taiwan</td>
<td>n.d.</td>
<td>24</td>
<td>121</td>
<td>Sinitic</td>
<td>SI</td>
</tr>
<tr>
<td>27</td>
<td>70</td>
<td>Buyi</td>
<td>China</td>
<td>n.d.</td>
<td>26.2</td>
<td>106</td>
<td>Tai-Kadai</td>
<td>KA</td>
</tr>
<tr>
<td>28</td>
<td>140</td>
<td>Thai</td>
<td>Thailand</td>
<td>Bangkok</td>
<td>13.4</td>
<td>100</td>
<td>Tai-Kadai</td>
<td>KA</td>
</tr>
<tr>
<td>29</td>
<td>96</td>
<td>Dai Lue</td>
<td>Thailand</td>
<td>North</td>
<td>17</td>
<td>101</td>
<td>Tai-Kadai</td>
<td>KA</td>
</tr>
<tr>
<td>30</td>
<td>106</td>
<td>Dai Dam</td>
<td>Thailand</td>
<td>North</td>
<td>17.6</td>
<td>102</td>
<td>Tai-Kadai</td>
<td>KA</td>
</tr>
<tr>
<td>31</td>
<td>100</td>
<td>Kinh</td>
<td>Vietnam</td>
<td>Hanoi</td>
<td>21.1</td>
<td>106</td>
<td>Mon-Khmer</td>
<td>AU</td>
</tr>
<tr>
<td>32</td>
<td>81</td>
<td>Muong</td>
<td>Vietnam</td>
<td>Hoa Binh</td>
<td>20.5</td>
<td>105</td>
<td>Mon-Khmer</td>
<td>AU</td>
</tr>
<tr>
<td>33</td>
<td>40</td>
<td>Indonesian</td>
<td>Indonesia</td>
<td>Molucca</td>
<td>0</td>
<td>128</td>
<td>Mal.-Pol.</td>
<td>AN</td>
</tr>
<tr>
<td>34</td>
<td>49</td>
<td>Indonesian</td>
<td>Indonesia</td>
<td>Nusa Tenggara</td>
<td>-9</td>
<td>117</td>
<td>Mal.-Pol.</td>
<td>AN</td>
</tr>
<tr>
<td>35</td>
<td>77</td>
<td>Indonesian</td>
<td>Indonesia</td>
<td>Java, Jakarta</td>
<td>-6.1</td>
<td>106</td>
<td>Mal.-Pol.</td>
<td>AN</td>
</tr>
<tr>
<td>36</td>
<td>77</td>
<td>Malay</td>
<td>Malaysia</td>
<td>n.d</td>
<td>3.9</td>
<td>101</td>
<td>Mal.-Pol.</td>
<td>AN</td>
</tr>
<tr>
<td>37</td>
<td>105</td>
<td>Filipino</td>
<td>Philippines</td>
<td>South Luzon</td>
<td>14.4</td>
<td>121</td>
<td>Mal.-Pol.</td>
<td>AN</td>
</tr>
<tr>
<td>38</td>
<td>50</td>
<td>Ivatan</td>
<td>Philippines</td>
<td>Batan islands</td>
<td>20.3</td>
<td>122</td>
<td>Proto-Filipino</td>
<td>AN</td>
</tr>
<tr>
<td>39</td>
<td>65</td>
<td>Paiwan</td>
<td>Taiwan</td>
<td>South</td>
<td>23.5</td>
<td>121</td>
<td>Paiwanic</td>
<td>AN</td>
</tr>
<tr>
<td>40</td>
<td>51</td>
<td>Paiwan</td>
<td>Taiwan</td>
<td>C. mountains</td>
<td>22.2</td>
<td>121</td>
<td>Paiwanic</td>
<td>AN</td>
</tr>
<tr>
<td>41</td>
<td>50</td>
<td>Atayal</td>
<td>Taiwan</td>
<td>C. mountains</td>
<td>24.3</td>
<td>121</td>
<td>Atayalic</td>
<td>AN</td>
</tr>
<tr>
<td>42</td>
<td>57</td>
<td>Saisiat</td>
<td>Taiwan</td>
<td>C. mountains</td>
<td>24.5</td>
<td>121</td>
<td>Western Plains</td>
<td>AN</td>
</tr>
<tr>
<td>43</td>
<td>88</td>
<td>Bunun</td>
<td>Taiwan</td>
<td>C. mountains</td>
<td>23.2</td>
<td>121</td>
<td>Paiwanic</td>
<td>AN</td>
</tr>
<tr>
<td>44</td>
<td>51</td>
<td>Tsou</td>
<td>Taiwan</td>
<td>C. mountains</td>
<td>23.4</td>
<td>120</td>
<td>Tsuic</td>
<td>AN</td>
</tr>
<tr>
<td>45</td>
<td>50</td>
<td>Rukai</td>
<td>Taiwan</td>
<td>C. mountains</td>
<td>22.4</td>
<td>120</td>
<td>Tsuic</td>
<td>AN</td>
</tr>
<tr>
<td>46</td>
<td>50</td>
<td>Ami</td>
<td>Taiwan</td>
<td>East coast</td>
<td>23.1</td>
<td>121</td>
<td>Sirayan</td>
<td>AN</td>
</tr>
<tr>
<td>47</td>
<td>50</td>
<td>Puyuma</td>
<td>Taiwan</td>
<td>East coast</td>
<td>22.4</td>
<td>121</td>
<td>Puyumic</td>
<td>AN</td>
</tr>
<tr>
<td>48</td>
<td>64</td>
<td>Yami</td>
<td>Taiwan</td>
<td>Orchid Island</td>
<td>22</td>
<td>121</td>
<td>Proto-Filipino</td>
<td>AN</td>
</tr>
</tbody>
</table>


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*

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Observed genetic diversity</th>
<th>Inferred evolutionary mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Among populations ($F_{ST}$)</td>
<td>Within populations ($h$)</td>
</tr>
<tr>
<td>A.</td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C.</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>D.</td>
<td>low</td>
<td>low</td>
</tr>
</tbody>
</table>
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.
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.


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**

<table>
<thead>
<tr>
<th>Group</th>
<th>size</th>
<th>( r_{GEN,GEO}^{ab} )</th>
<th>( r_{GEN,LING}^{ac} )</th>
<th>( r_{GEO,LING}^{a} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>All populations</td>
<td>40</td>
<td>0.131* (0.137*)</td>
<td>0.015 n.s. (-0.042 n.s.)</td>
<td>0.401***</td>
</tr>
<tr>
<td>Continental East Asians</td>
<td>25</td>
<td>0.352*** (0.288***)</td>
<td>0.333*** (0.264***))</td>
<td>0.270***</td>
</tr>
<tr>
<td>Austronesians only</td>
<td>15</td>
<td>0.340* (0.357*)</td>
<td>0.110 n.s. (0.158 n.s.)</td>
<td>-0.112 n.s.</td>
</tr>
</tbody>
</table>

\( a \)** : \( P < 0.001 \); \( * \) : \( 0.01 < P < 0.05 \); 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) 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

<table>
<thead>
<tr>
<th>Linguistic group</th>
<th>Group size</th>
<th>$F_{ST}$ (%)</th>
<th>$h$ (s.d.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altaic</td>
<td>14</td>
<td>1.70***</td>
<td>0.941 (0.016)</td>
</tr>
<tr>
<td>- Altaic proper</td>
<td>9</td>
<td>1.27***</td>
<td>0.946 (0.014)</td>
</tr>
<tr>
<td>- Koreo-Japonic</td>
<td>5</td>
<td>0.88***</td>
<td>0.932 (0.016)</td>
</tr>
<tr>
<td>Sinitic</td>
<td>6</td>
<td>0.50***</td>
<td>0.931 (0.001)</td>
</tr>
<tr>
<td>Linguistic Group</td>
<td>Populations</td>
<td>Gene Diversity</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------</td>
<td>----------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Tai-Kadai</td>
<td>4</td>
<td>2.39***</td>
<td>0.906 (0.007)</td>
</tr>
<tr>
<td>Austroasiatic</td>
<td>2</td>
<td>5.38***</td>
<td>0.891 (0.019)</td>
</tr>
<tr>
<td>Austronesian</td>
<td>16</td>
<td>9.72***</td>
<td>0.815 (0.054)</td>
</tr>
<tr>
<td></td>
<td>- Formosan</td>
<td>9</td>
<td>7.10***</td>
</tr>
<tr>
<td></td>
<td>- Extra-Formosan</td>
<td>7</td>
<td>8.98***</td>
</tr>
</tbody>
</table>

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 (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

<table>
<thead>
<tr>
<th>Linguistic groups</th>
<th>(F_{CT}) (%)</th>
<th>(F_{SC}) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altaic (v) Sinitic</td>
<td>0.6*</td>
<td>1.3***</td>
</tr>
<tr>
<td>Tai-Kadai</td>
<td>1.8***</td>
<td>1.8***</td>
</tr>
<tr>
<td>Austronesian</td>
<td>3.5***</td>
<td>3.7***</td>
</tr>
<tr>
<td>Sinitic (v) Tai-Kadai</td>
<td>1.4**</td>
<td>1.0***</td>
</tr>
<tr>
<td>Austronesian</td>
<td>2.4*</td>
<td>4.9***</td>
</tr>
<tr>
<td>Tai-Kadai (v) Austronesian</td>
<td>0.5 n.s.</td>
<td>7.8**</td>
</tr>
</tbody>
</table>
Altaic proper \textit{versus} Koreo-Japonic \hspace{1cm} 14.5*** \hspace{1cm} 1.0***
Formosan \textit{versus} Extra-Formosan \hspace{1cm} 4.8** \hspace{1cm} 7.3***

\textsuperscript{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. \textsuperscript{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 \textit{et al.} 1994; Dugoujon \textit{et al.} forthcoming) as well as with some DNA polymorphisms (Karafet \textit{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 Eurasiatian 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. 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 R¹ 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 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)).
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 macrophyyla 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 \textit{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


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 Austroasiatic was not considered because it included only two populations. 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.
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.