# Geographical, Linguistic, and Cultural Influences on Genetic Diversity: Y-Chromosomal Distribution in Northern European Populations

*Tatiana Zerjal,* Lars Beckman,† Gunhild Beckman,†‡ Aavo-Valdur Mikelsaar,§ Astrida Krumina,‖ Vaidutis Kučinskas,¶ Matthew E. Hurles,# and Chris Tyler-Smith**

*Department of Biochemistry, University of Oxford, Oxford, England; †Department of Medical Genetics, Umeå University, Umeå, Sweden; ‡Gotland University College, Visby, Gotland, Sweden; §Institute of General and Molecular Pathology, University of Tartu, Estonia; ‖Department of Medical Biology and Genetics, Medical Academy of Latvia, Riga, Latvia; ¶Center of Human Genetics, University of Vilnius, Vilnius, Lithuania; #McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, England

We analyzed 10 Y-chromosomal binary markers in 363 males from 8 populations in Northern Europe and 5 Y microsatellites in 346 of these individuals. These populations can be grouped according to cultural, linguistic, or geographical criteria, and the groupings are different in each case. We can therefore ask which criterion best corresponds to the distribution of genetic variation. In an AMOVA analysis using the binary markers, 13% of the Y variation was found between populations, indicating a high level of differentiation within this small area. No significant difference was seen between the traditionally nomadic Saami and the neighboring, historically farming, populations. When the populations were divided into Uralic speakers and Indo-European speakers, 8% of the variation was found between groups, but when they were divided according to geographical location, 14% of the variation was between groups. Geographical factors have thus been the most important in limiting gene flow between these populations, but linguistic differences have also been important in the east.

## Introduction

Human genetic variation is not distributed evenly across the globe (Cavalli-Sforza, Menozzi, and Piazza 1994), so as medicine becomes increasingly based on genetics, it will become more and more important to understand the distribution of these genetic differences. Geographical patterns of variation are seen on many scales. Continental differences exist, such as the higher level of heterozygosity in Africa, explained by a longer time of occupation and/or a larger effective population size (Jorde, Bamshad, and Rogers 1998), but there are also local differences which form gradual clines or more abrupt changes, called genetic boundaries (Barbujani 2000). Although selection is responsible for some differences in gene frequencies, most DNA variation is thought to be neutral, or nearly so. Differences between populations arise largely through random genetic drift when they are separated by factors such as distance, geographical barriers, or culture, and may or may not be maintained when distinct populations come into contact. There is debate about the relative importance of these factors. Europe provides an excellent area in which to investigate the mechanisms responsible because its archaeology, history, culture, linguistics, and genetics are relatively well known.

The northern part of Europe was an inhospitable area during the last glacial maximum (around 20,000 B.C.), and there is little evidence for human occupation during this period. Archaeological sites dating back to 10,000–8,000 B.C. are known from the southern part of the Scandinavian Peninsula, but much of the interior was probably only colonized after 6,000 B.C. (Larsson 1996). The Scandinavian Peninsula is separated from Finland and the Baltic countries by the Baltic Sea, which is potentially a substantial geographical barrier and is likely to have affected human migration in prehistoric times. Later, however, intense commercial and cultural exchanges across the Baltic Sea are well documented, particularly during the Viking Age, and the Swedish island of Gotland, situated in the middle of the Baltic Sea at the ''crossroads'' of trading pathways (Tønneson 1994), seems to have been particularly exposed to influence from the east. The extent to which these contacts have led to genetic exchanges is an issue that so far has received relatively little attention.

Europe shows a remarkable linguistic homogeneity, with most of the populations speaking languages belonging to the Indo-European family. In northern Europe, Indo-European speakers include the Swedes and Norwegians (on the western side of the Baltic Sea), whose languages belong to the Germanic subfamily, and the Latvians and Lithuanians (on the eastern side of the Baltic), whose languages belong to the Baltic subfamily. In contrast, the Saami, Finns, and Estonians speak languages belonging to the Uralic family, otherwise spoken across a broad region of northern Asia (Hajdú 1976).

Finally, this area provides an example of cultural and livelihood differences that have separated the Saami from their neighboring populations for thousands of years. The Saami were nomadic herders and hunters of reindeer, while the other populations were, after ~3,000 B.C., traditionally farmers (Cavalli-Sforza, Menozzi, and Piazza 1994). Only in the last half century have the Saami undergone significant acculturation.

Thus, in northern Europe, strong geographical, linguistic, and cultural barriers can all be identified, but they divide the populations into different groups. It is therefore a particularly informative region: we can ask which grouping most closely corresponds to the genetic

Key words: Y chromosome, binary marker, microsatellite, genetic boundary.

Address for correspondence and reprints: Chris Tyler-Smith, CRC Chromosome Molecular Biology Group, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England. E-mail: chris@bioch.ox.ac.uk.

*Mol. Biol. Evol.* 18(6):1077–1087. 2001

grouping and thus evaluate the relative importance of these factors in determining the current distribution of genetic variation.

Previous work has provided some information on the genetics of these populations but has left several questions unanswered. While most of the present populations are assumed to have a similar origin to other Europeans (Carpelan 1998), the origin of the Saami has been the subject of controversy. One hypothesis suggests that they were originally Mongoloids who moved from Western Siberia (Cavalli-Sforza, Menozzi, and Piazza 1994), while an alternative hypothesis suggests, on the basis of archaeological findings, a possible homeland in the Onega and Ladoga Lakes region (today in the Karelia Republic) (Eriksson 1988; Carpelan 1998). Whatever their origin, the Saami are thought to have gradually retreated northward after ~2,000 B.C. as a result of the Neolithic expansion of European farmers (Eriksson 1988; Carpelan 1998), such that they are now confined to the extreme north of Scandinavia and Fennoscandia.

Genetic studies using autosomal and mtDNA markers usually find that the Saami represent an outlying population in Europe (Cavalli-Sforza, Menozzi, and Piazza 1994; Sajantila et al. 1995; Lahermo et al. 1996; Simoni et al. 2000), and the same conclusion was reached when a small number of Y-chromosomal markers were used (Sajantila et al. 1996). In contrast, other northern European populations (including the Finns, the Estonians, and the Swedes) were not significantly different from the rest of Europe when analyzed using mtDNA markers (Sajantila et al. 1995; Simoni et al. 2000), although the Finns show unusual frequencies at many autosomal loci, usually attributed to a bottleneck (de la Chapelle 1993). A few genetic investigations are available for the southern Baltic populations, such as the Estonians, the Latvians, and the Lithuanians. Although they generally resemble other European populations, recent studies using autosomal markers have shown some distinct features (Beckman et al. 1998; Sistonen et al. 1999).

Most of the Y chromosome does not recombine, so haplotypes are transmitted unchanged from father to son apart from rare mutations and can be used in the interpretation of population history. The effective population size for this chromosome is four times as small as that of any autosome, making it particularly susceptible to drift and thus a very sensitive genetic tool with which to investigate not only population movements, but also the elements that can obstruct them. A wide range of markers are now available, including stable binary polymorphisms and more variable microsatellites (Jobling and Tyler-Smith 1995; Ayub et al. 2000; Underhill et al. 2000). Binary markers have the advantage that they usually provide unambiguous identification of a set of chromosomes that share a common ancestor, but they suffer from ascertainment bias. They can usefully be combined with microsatellites, which are variable in all populations. Using a Y-chromosomal base substitution (''Tat'') that seems to have arisen in Asia, Zerjal et al. (1997) proposed a significant Asian contribution to the paternal gene pool of the Finns, the Saami, and the Estonians but less of a contribution to the Norwegians; subsequent work has identified the same marker in the Latvians (Lahermo et al. 1999). Thus, the Y chromosome is an effective genetic tool for revealing the patterns of variation in this area.

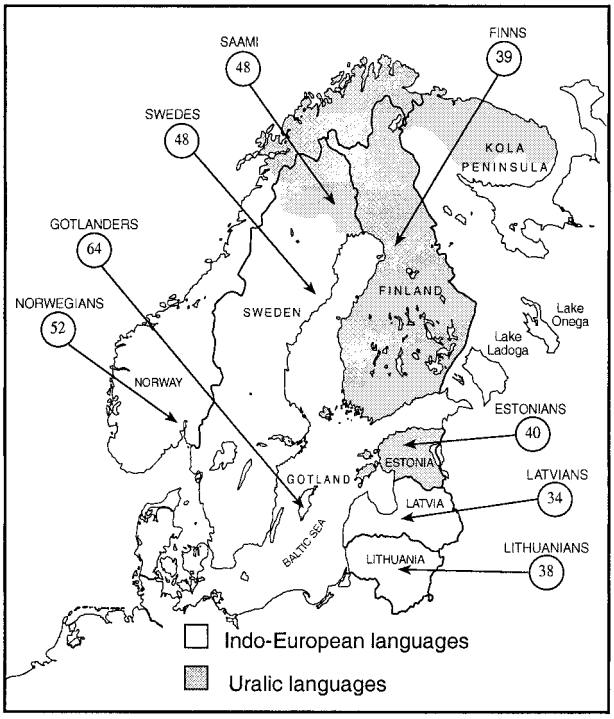


FIG. 1.—Populations sampled and language family distributions.

In the present study, we therefore used a set of Y-chromosomal markers with different mutation rates to investigate the relative importance of geographical, linguistic, and cultural factors in determining the distribution of genetic variability in closely located populations.

## Materials and Methods

### Samples

This study used a set of 363 males from eight populations from Scandinavia and the Baltic region (fig. 1). Four sets of samples were from the eastern side of the Baltic: Finns and Estonians (both Uralic speakers) and Lithuanians and Latvians (Indo-European speakers, Baltic subfamily). Three sets of samples were from the western side of the Baltic: Swedes, Gotlanders, and Norwegians (all Indo-European speakers, Germanic subfamily). The samples from the island of Gotland were of Swedish nationality but were considered separately in order to investigate their origin and degree of admixture with the neighboring populations. The Saami samples were from the Swedish territory of Norrbotten, and they were from Uralic speakers. Although they now live to the west of the Gulf of Bothnia, they are thought to have come from Karelia and should for many purposes be considered an eastern population.

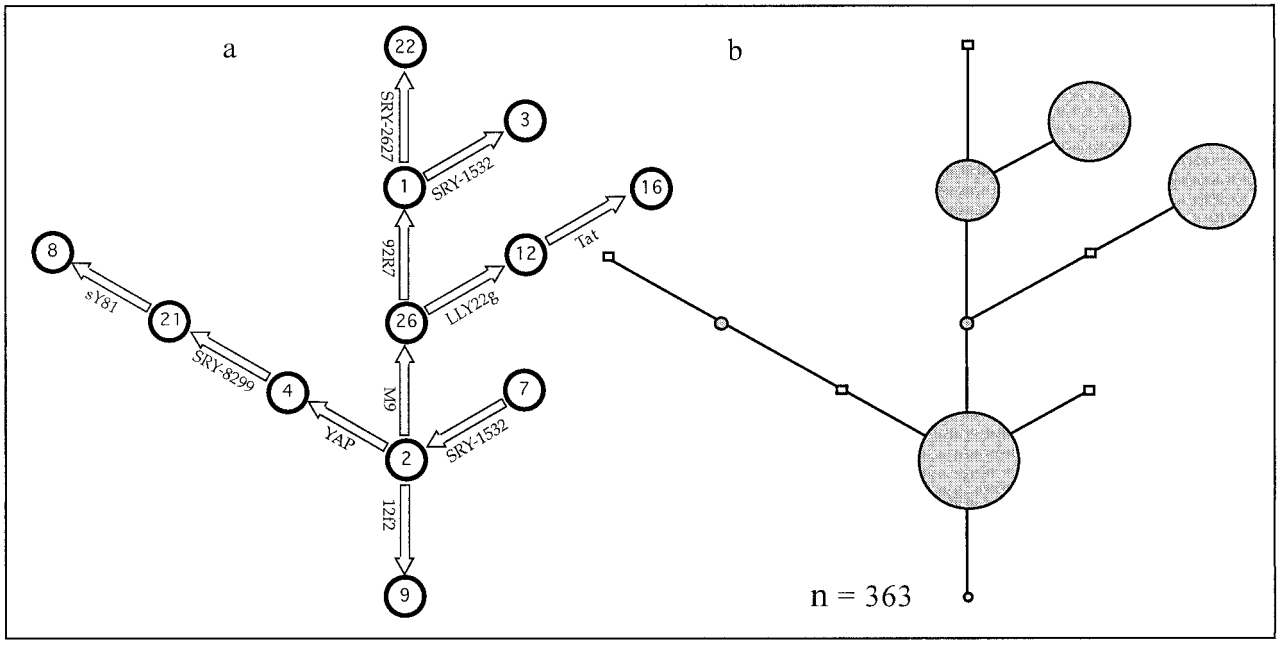


FIG. 2.—*a*, Haplogroups (in circles) defined by binary Y markers (arrows). *b*, Haplogroup frequencies in the 363 individuals. Frequency is proportional to circle area; squares = haplogroup not found.

## Binary Polymorphism Typing

Ten binary markers known to identify polymorphisms within Europe were used in this study. Nine appear to originate from unique mutational events: 12f2, YAP, *SRY*-8299, sY81, M9, LLY22g/*Hin*dIII, Tat, 92R7, and *SRY*-2627 (references below). One, *SRY*-1532, has undergone recurrent mutation. Ten markers with these mutational characteristics (nine unique, one recurrent) allow 12 haplotypes to be defined. These haplotypes often contain large numbers of related chromosomes and are designated ''haplogroups'' to distinguish them from the haplotypes defined by microsatellites alone or combinations of binary and microsatellite markers. Haplogroup relationships and nomenclature have been established in previous work (Jobling, Pandya, and Tyler-Smith 1997; Jobling and Tyler-Smith 2000), where a tree was constructed using the principle of parsimony assuming the minimum number of mutational events. The relevant portion of this tree is shown in figure 2*a*. Although two of the markers (sY81 and *SRY*-2627) did not detect variation in the samples analyzed, the results are reported in order to facilitate comparisons with other studies.

All the samples were typed with each biallelic marker. Although the allelic states of some polymorphisms have previously been found to be associated with one another, this redundancy provides an internal check on the reliability of the typing (Jobling, Pandya, and Tyler-Smith 1997; Pandya 1998). Typing conditions were similar for all of the markers, changing only the cycling program. The reaction volume was 25 μl and contained 20–50 ng of DNA, 1 μM of each primer, 100–200 μM dNTPs, and 1.25 U *Taq* polymerase in the PCR buffer described previously (Zerjal et al. 1997; Hurles et al. 1998). All PCR reactions were carried out in an MJR PTC-200 thermal cycler. For restriction fragment length polymorphism analysis, 1–3 U of the appropriate restriction enzyme in 10 μl of 1 × digestion buffer was added directly to 13 μl of PCR reaction and incubated at the appropriate temperature overnight. Digests were analyzed by electrophoresis on agarose gels (2%–4% NuSieve : Seakem, 2:1) containing ethidium bromide in 0.5 × Tris-acetate EDTA (TAE) buffer.

The polymorphisms *SRY*-1532, YAP, *SRY*-8299, sY81, 12f2, M9, 92R7, Tat, and *SRY*-2627 were typed according to procedures based on those described previously (Seielstad et al. 1994; Hammer and Horai 1995; Whitfield, Sulston, and Goodfellow 1995; Underhill et al. 1997; Zerjal et al. 1997; Hurles et al. 1999; Santos et al. 1999; Blanco et al. 2000).

## Microsatellite Typing

Samples were typed with five microsatellite markers: *DYS390*, *DYS391*, *DYS19*, *DYS392*, and *DYS393*. All have tetranucleotide repeat units, except for *DYS392* which has a trinucleotide repeat unit. The five microsatellites were amplified in a single multiplex PCR reaction (Thomas, Bradman, and Flinn 1999) and run on a denaturing 4.25% polyacrylamide gel in 1 × Tris-borate EDTA (TBE) on an ABI 377 DNA sequencer. Gels were analyzed by ABI PRISM GeneScan Analysis 2.0.2 to produce sample files, which were then imported into Genotyper 1.1 (both from PE Applied Biosystems).

## Data Analysis

Haplogroup frequencies were calculated for all populations, and the chromosomes in each haplogroup were analyzed with a set of five microsatellites. With these data, a median-joining network (Bandelt, Forster,

**Table 1**
**Haplogroup Frequencies and Diversities**

| Population (*n*) | Haplogroups | | | | | | | Haplogroup Diversity |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 16 | 9 | 21 | 26 | |
| Norwegians (52) | 15 (29%) | 17 (33%) | 16 (31%) | 2 (4%) | 1 (2%) | 1 (2%) | 0 | 0.722 (±0.014) |
| Gotlanders (64) | 11 (17%) | 38 (59%) | 10 (16%) | 4 (6%) | 0 | 0 | 1 (2%) | 0.603 (±0.07) |
| Swedes (48) | 11 (23%) | 23 (48%) | 9 (19%) | 4 (8%) | 0 | 1 (2%) | 0 | 0.688 (±0.06) |
| Saami (48) | 3 (6%) | 15 (31%) | 10 (21%) | 20 (42%) | 0 | 0 | 0 | 0.690 (±0.04) |
| Finns (39) | 0 | 11 (28%) | 3 (8%) | 25 (64%) | 0 | 0 | 0 | 0.518 (±0.09) |
| Estonians (40) | 2 (5%) | 14 (35%) | 10 (25%) | 13 (32.5%) | 0 | 0 | 1 (2.5%) | 0.720 (±0.04) |
| Latvians (34) | 5 (15%) | 4 (12%) | 14 (41%) | 11 (32%) | 0 | 0 | 0 | 0.700 (±0.06) |
| Lithuanians (38) | 2 (5%) | 5 (13%) | 13 (34%) | 18 (47%) | 0 | 0 | 0 | 0.660 (±0.06) |

and Rohl 1999) for each haplogroup was constructed using the program Network, version 1.8.

The distribution of Y-chromosomal diversity was measured as within- and between-population variation calculated using analysis of molecular variance (AMOVA) (Excoffier, Smouse, and Quattro 1992) using the Arlequin computing package, version 1.1. With the same program, pairwise genetic distances between populations were calculated either from the haplogroup frequencies or from the microsatellite haplotypes. In each case, a distance matrix was created by counting the number of mutational steps separating each pair of haplogroups or haplotypes. The $F_{ST}$ analogs calculated in this way are referred to as $\Phi_{ST}$ values. Haplotype frequencies were calculated for all populations. Except for the haplogroup 16 AMOVA results, where only haplogroup 16 haplotypes were taken into account, the microsatellite analyses did not use information about the haplogroups. Diversities and their standard errors were calculated according to Nei (1987). Haplogroup frequencies and pairwise genetic distances from microsatellite haplotypes were represented in two-dimensional space with multidimensional scaling (MDS) in the SPSS, version 7.0, software package.

Genetic boundaries and their significance assessments were calculated using the ORINOCO program (Hurles 1999). This program uses methods similar to those described earlier (Barbujani, Oden, and Sokal 1989) to identify the geographical regions in a chosen landscape where abrupt genetic change occurs (known formally as ''genetic boundaries'') and subsequently applies a permutation test to assist in evaluating their significance. It involved the following steps: (1) A frequency distribution surface was calculated for each haplogroup throughout the landscape by interpolation from the irregular sampling sites to a regular 100 × 100 grid. (2) These surfaces were differentiated to determine the rate of change with respect to distance at each grid point. (3) The differentials were summed to produce the raw output (e.g., as illustrated in fig. 6*a*), which is thus a representation of the overall rate of change of Y haplogroup frequency with distance extant at the time of sampling. In order to assess whether regions of rapid change are likely to be significant, a simple threshold filter was first applied to retain only the top 5% of values, and then a permutation test was used: (4) Population samples with the same size and geographical location as the real data were constructed by randomly sampling from the pooled data, and steps 1–3 were repeated. Step 4 was carried out 1,000 times. The real data were then compared with these simulations, and grid points at which the rate of change in the real data was greater than that in 95% of the simulations were retained. Thus, insignificant boundaries that appeared because of small sample size were excluded.

## Results

Ten binary Y markers were used to classify 363 samples from eight northern European populations into haplogroups (table 1). This set of markers potentially identifies 12 haplogroups (fig. 2*a*); however, in the entire region, only four were observed at high frequencies (haplogroups 1–3 and 16; fig. 2*b*). Three haplogroups were present at low frequencies (haplogroups 21, 26, and 9), while five haplogroups (haplogroups 12, 7, 4, 8, and 22) were not seen at all.

Haplogroup frequencies were examined in individual populations. Haplogroups 2 and 3 did not show large differences in frequency between the populations, but haplogroups 1 and 16 showed substantial differences within this area (fig. 3). Haplogroup 1 was strongly represented on the western side of the Baltic Sea but was quite rare among the Saami and the populations on the eastern side (fig. 3*a*). This clinal distribution fits its known pattern of distribution, with a high incidence in western European populations (English, Irish, Basques, Catalans) and a lower frequency in the east (Pandya 1998; Hill, Jobling, and Bradley 2000). Haplogroup 16 showed the opposite pattern: high frequency among the populations on the eastern side of the Baltic Sea, accounting for more than 60% of the Finnish, 42% of the Saami, and 47% of the Lithuanian chromosomes, and low frequency on the western side, falling to 4% among the Norwegians (fig. 3*c*). Again, this fits its known worldwide distribution (Zerjal et al. 1997; Karafet et al. 1999), where it is well represented in Northern Asia but absent from most of Europe. Haplogroup 3 accounts for a high proportion of Asian and European chromosomes (Zerjal et al. 1999) and shows considerable local variation, especially in Asia, where it has not been detected on the eastern edge of the continent. In Europe, its highest frequencies are among central-eastern populations, and it is rare in the southeast and southwest regions of the continent. In Scandinavia and in the Baltic regions, it is well represented among all the populations, with

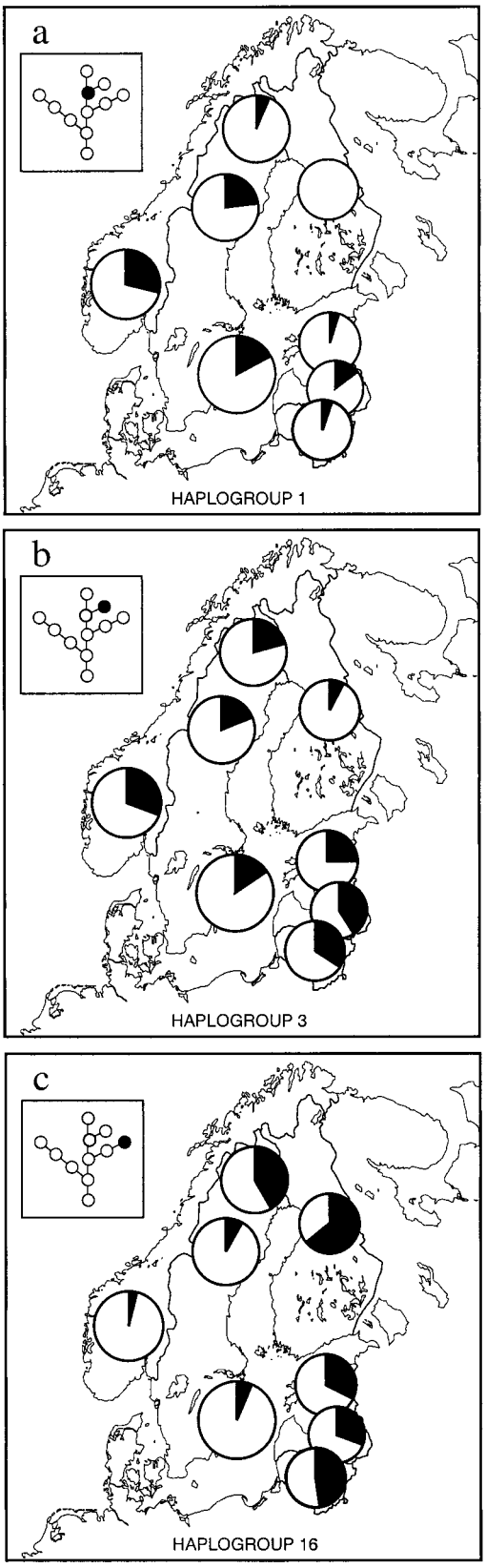


FIG. 3.—Geographical distribution of selected haplogroups. Circle area is proportional to sample size. *a,* Haplogroup 1. *b,* Haplogroup 3. *c,* Haplogroup 16.

the highest frequencies among the Latvians (41%) and the Norwegians (31%) and a lower frequency in Finland (8%) (fig. 3*b*). Haplogroup 2 contains a heterogeneous set of chromosomes that are not necessarily closely related. It is common in much of Asia and in Europe but, without further subdivision, its distribution is not very meaningful and is not shown. Haplogroups 9, 21, and 26 are rare in this region. The first two are common in northern Africa and southern Europe (Pandya 1998; Bosch et al. 1999) while haplogroup 26 is present, but rare, in much of Europe.

Y chromosomes were also typed using five microsatellites. Full data were available from 346 samples, among which 92 different five-locus microsatellite haplotypes were found, or 97 combination binary plus microsatellite haplotypes (summary table available as supplementary information). Average haplotype diversity was 0.95 for all of the populations, but in the Finns it was slightly lower ($0.83 \pm 0.04$), as was the haplogroup diversity (0.68 compared with $0.52 \pm 0.09$ in the Finns; table 1). A previous study (de Knijff et al. 1997) found a higher haplotype diversity of 0.99 using seven Y microsatellites in four European populations, as might be expected with more microsatellites.

To investigate the phylogenetic relationships between microsatellite haplotypes within each haplogroup, we constructed median-joining networks using the program Network, version 1.8. Haplogroups 1 (not shown) and 3 (fig. 4*a*) showed a high degree of haplotype sharing among populations from both sides of the Baltic Sea and from both linguistic families, with no evident haplotype clustering on the basis of geography or linguistics. In haplogroup 3, the four common haplotypes were shared by most of the populations; some of the minor haplotypes were population-specific but were scattered randomly around the main core (fig. 4*a*). A different result was found in the network analysis of haplogroup 16. In this case, a concentration of Latvian and Lithuanian haplotypes was present on the left-hand side of the network, while Finns, Saami and Estonians were mostly on the right-hand side (fig. 4*b*). The two groups of haplotypes were distinguished by a difference at *DYS19.* Swedes, Gotlanders, and Norwegians were also on the right-hand side of the network, perhaps indicating northeastern ancestors for the western haplogroup 16 chromosomes.

Genetic relationships between the populations were examined using the binary and microsatellite data in separate analyses. Haplogroup frequencies were used to produce a two-dimensional plot using multidimensional scaling (fig. 5*a*). The strongest division was between eastern and western populations, separated by the Baltic Sea. This can readily be understood from the distributions of haplogroups 1 and 16, shown in figure 3. The Saami and the Estonians are the closest populations, while the Finns are separate and the Latvians and Lithuanians are close together but somewhat separated from the rest.

Microsatellite haplotype frequency data were used to calculate pairwise genetic distances between populations using AMOVA (table 2) and then displayed in

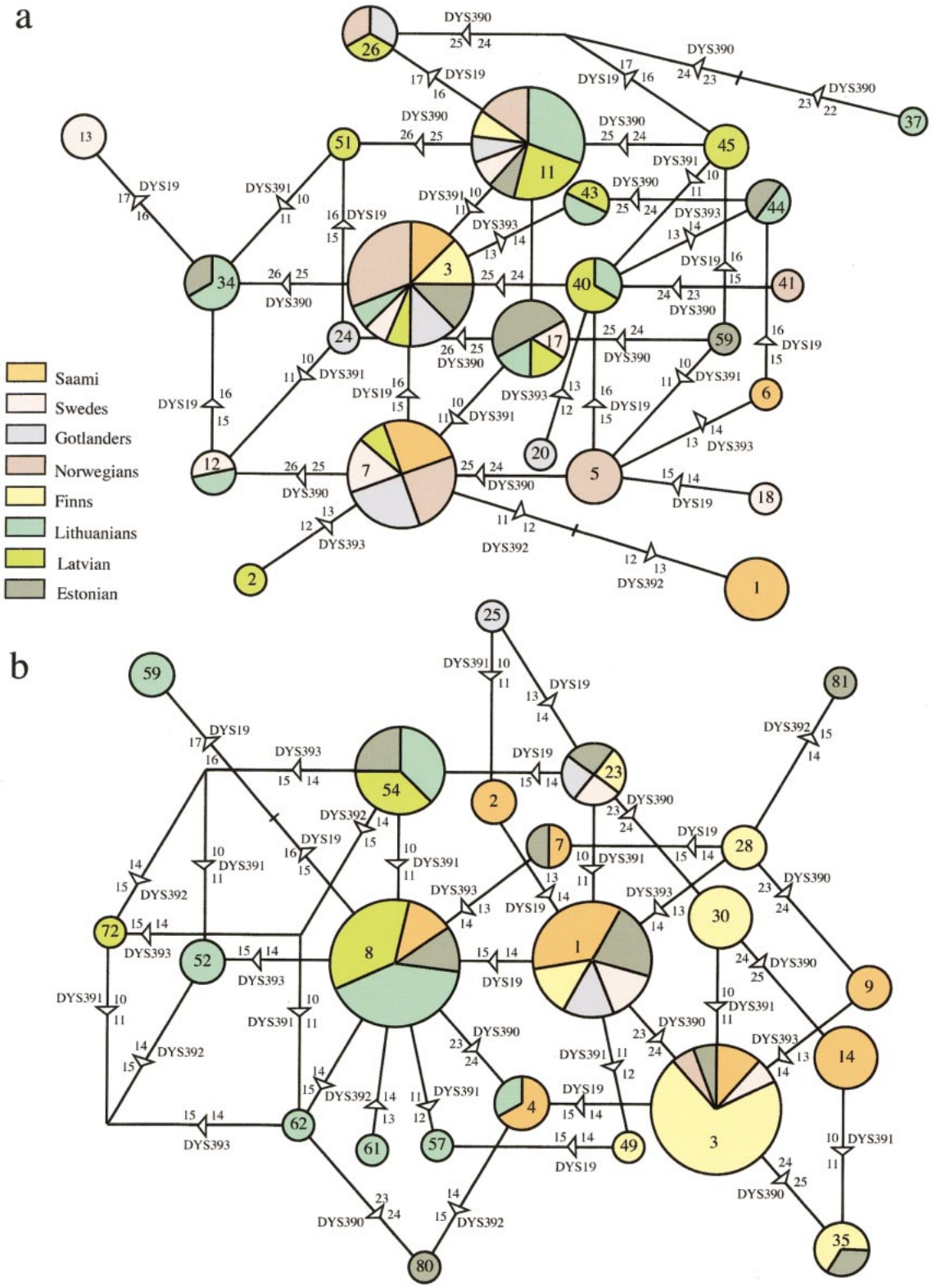


FIG. 4.—Median-joining networks of microsatellite haplotypes. Each circle represents a haplotype, and the circle area is proportional to the number of chromosomes. Each color indicates a population, and the microsatellite mutational steps are shown on the lines linking haplotypes. *a*, Haplogroup 3. *b*, Haplogroup 16.

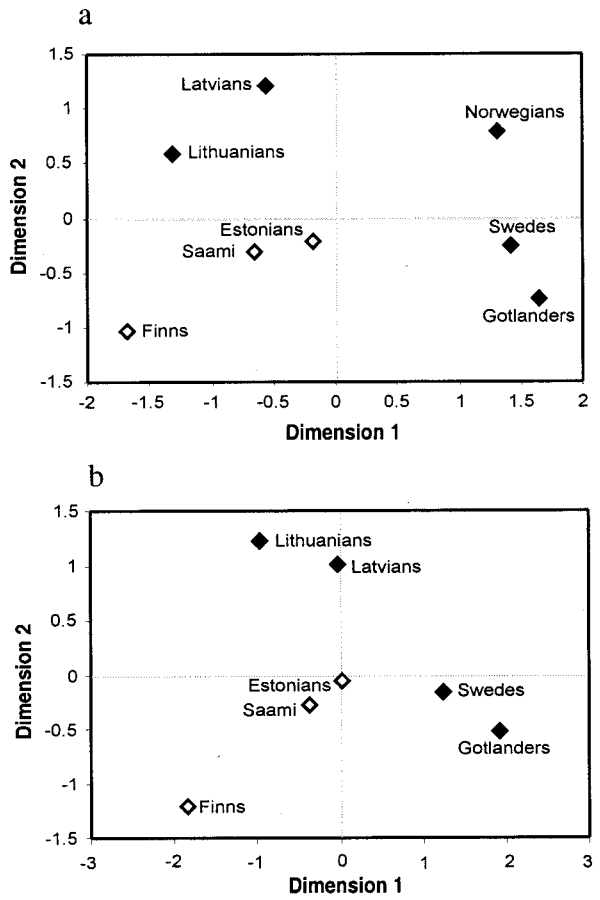


FIG. 5.—Genetic relationships between populations represented by multidimensional scaling. Open symbols, Uralic speakers; closed symbols, Indo-European speakers. Relationships are based on (*a*) haplogroup frequencies and (*b*) pairwise $\Phi_{ST}$ values derived from microsatellite haplotype frequencies and molecular distances between haplotypes.

two dimensions, again using multidimensional scaling (fig. 5*b*). Despite using independent markers, the plots are quite similar. In figure 5*b*, the major division is between the Latvians and Lithuanians and the rest of the populations, and there is a secondary division between the eastern and western populations. As in figure 5*a*, the Saami and the Estonians lie very close together, and on both plots Swedes and Gotlanders are clustered together, illustrating the genetic similarity existing between them.

The genetic structure of these data was analyzed in more detail using AMOVA (Excoffier, Smouse, and Quattro 1992), which allows the percentage of genetic variation to be calculated in hierarchical levels: within populations, among populations, and among groups of populations (grouped according to geography, linguistic family, or linguistic subfamily). When haplogroup frequencies were analyzed without grouping of the populations, the highest fraction of variability was due to within-population differences, as expected (87%; $P < 0.0001$), but a substantial minor fraction, 13% ($P < 0.0001$), was due to differences among populations (table 3). When microsatellite haplotype frequencies were used, the latter fraction was 9%.

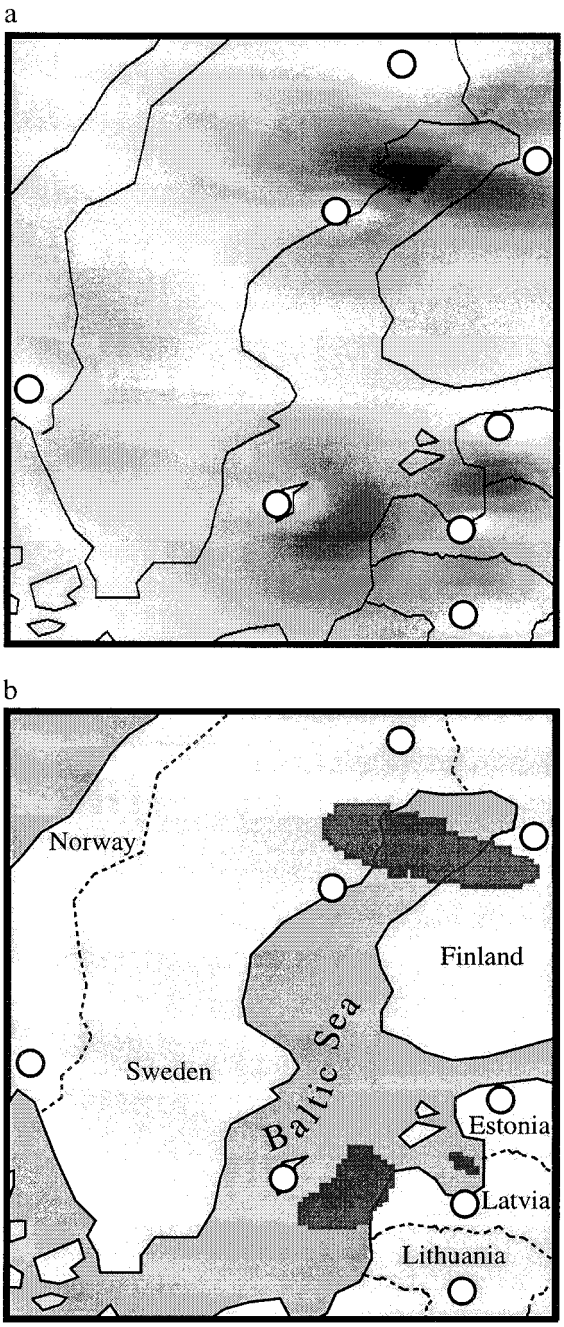


FIG. 6.—Visualization of genetic boundaries. *a*, Raw data. The coastline is shown in black and the locations of the populations are shown as white circles. Boundaries are coded according to their intensity, with higher boundaries shown as darker gray. *b*, Boundaries remaining after application of the 5% threshold and 95% permutation filters. Land is shown in light gray, the sea in intermediate gray, and the boundaries in dark gray.

When a hierarchical approach was taken, populations were first divided according to geography. Populations on the western side of the Baltic (Norwegians, Swedes, and Gotlanders) were grouped together and compared with populations on the northern and eastern sides (Saami, Finns, Estonians, Latvians, and Lithuanians). The second grouping was on the basis of language family: Indo-European speakers (Norwegians,

**Table 2**
**$\Phi_{ST}$ Values Calculated from Microsatellite Haplotypes and Their Significance**

| | Gotlanders | Swedes | Saami | Finns | Estonians | Latvians | Lithuanians | Norwegians |
|---|---|---|---|---|---|---|---|---|
| Gotlanders | | — | + | + | + | + | + | + |
| Swedes | −0.009 | | + | + | + | + | + | — |
| Saami | 0.105 | 0.08 | | + | — | + | + | + |
| Finns | 0.221 | 0.196 | 0.042 | | + | + | + | + |
| Estonians | 0.055 | 0.033 | 0.006 | 0.093 | | — | — | + |
| Latvians | 0.137 | 0.109 | 0.063 | 0.168 | 0.012 | | — | + |
| Lithuanians | 0.197 | 0.161 | 0.066 | 0.158 | 0.037 | −0.011 | | + |
| Norwegians | 0.038 | 0.013 | 0.081 | 0.215 | 0.04 | 0.077 | 0.124 | |

NOTE.—Significance of $\Phi_{ST}$ $P$ values is shown above the diagonal. Population pairwise $\Phi_{ST}$ values are shown below the diagonal.

Swedes, Gotlanders, Lithuanians, and Latvians) compared with Uralic speakers (Saami, Finns, and Estonians). Finally, the third grouping was also based on language but divided the Indo-Europeans speakers according to their language subfamilies: Germanic speakers (Norwegians, Swedes, and Gotlanders), Baltic speakers (Latvians and Lithuanians), and Uralic speakers.

A significant amount of differentiation between groups was present when geography and linguistic subfamily were taken into account (14% in both cases for haplogroup frequency variation; 8% and 10%, respectively, when microsatellite haplotype variation was used). Nonsignificant results (table 3) were obtained instead when the language family grouping was taken into account, indicating that the differentiation that is detectable in this case is not entirely ascribable to the linguistic division. It is interesting to observe that the AMOVA results from the biallelic marker data always indicate a higher degree of group differentiation than that obtained with microsatellite data, perhaps because of homoplasy among the microsatellite haplotypes.

Populations consist of individuals subjected to common demographic processes that affect the different genetic lineages together and not independently, but they are not necessarily stable over time. The present composition does not always reflect the past composition. Since haplogroup 16 seems from the network analysis (fig. 4) to show a different population distribution pattern, distinct from the rest of the haplogroups found in these populations, we carried out an independent AMOVA analysis for the haplogroup 16 chromosomes. A larger amount of population differentiation was seen in this haplogroup, apportioning 28% ($P < 0.001$) of the genetic variation to among-populations differences (table 4). When geography was taken into consideration, the amount of variation among groups fell to 0.5% ($P = 0.38$). This suggests that the small number of haplogroup 16 haplotypes in the western countries are quite similar to the Saami/Finnish/Estonian ones, perhaps because of gene flow. When language subfamily was taken in to consideration, the amount of differentiation among groups increased even further, assigning 31% ($P = 0.003$) of the variation among groups and 5% of the variation to populations within the groups. This result shows a strong differentiation among groups accompanied by a striking similarity among populations within the same group. Nonsignificant values were obtained instead when populations were grouped simply into Uralic speakers and Indo-European speakers. Thus, for this haplogroup, the major differentiation is between (Latvians + Lithuanians) and (Saami + Estonians + Finns), as illustrated by the network (fig. 4*b*).

To locate the zones of sharpest genetic change within this area, we used the Orinoco program on the haplogroup data for the eight populations (fig. 6). In figure 6*a*, the strongest boundaries divide western populations from eastern ones, indicating that the geographical boundary has played a major role in determining genetic frequencies in this part of Europe. Nevertheless, a zone of sharp genetic change is also present between Estonians and Latvians and between Swedes and the Saami, supporting the idea that other factors also contribute to genetic boundaries. When the top 5% threshold filter was combined with the 95% permutation filter, a reduced boundary was still present between the Es-

**Table 3**
**Apportionment of Genetic Variation Using Two Different Sets of Markers**

| Grouping | Source of Variation | Biallelic Markers (%) | Microsatellites (%) |
|---|---|---|---|
| No grouping | Among populations | 13 ($P < 0.0001$) | 9 ($P < 0.0001$) |
| | Within populations | 87 ($P < 0.0001$) | 91 ($P < 0.0001$) |
| Geography | Among groups | 14 ($P = 0.02$) | 8 ($P = 0.02$) |
| | Amoung populations within groups | 4 ($P < 0.0001$) | 4 ($P < 0.0001$) |
| | Within populations | 81 ($P < 0.0001$) | 87 ($P < 0.0001$) |
| Language family | Among groups | 8 ($P = 0.08$) | 4 ($P = 0.09$) |
| | Amoung populations within groups | 9 ($P < 0.0001$) | 7 ($P < 0.0001$) |
| | Within populations | 84 ($P < 0.0001$) | 89 ($P < 0.0001$) |
| Language subfamily | Among groups | 14 ($P = 0.003$) | 10 ($P = 0.003$) |
| | Among populations within groups | 3 ($P < 0.0001$) | 2 ($P < 0.0001$) |
| | Within populations | 84 ($P < 0.0001$) | 88 ($P < 0.0001$) |

**Table 4**
**Apportionment of Genetic Variation in Haplogroup 16 Chromosomes**

| Grouping | Source of Variation | Percentage of Variation |
|---|---|---|
| No grouping | Among populations | 28 ($P < 0.0001$) |
| | Within populations | 72 ($P < 0.0001$) |
| Geography | Among groups | 0.5 ($P = 0.38$) |
| | Among poulations within groups | 28 ($P < 0.0001$) |
| | Within populations | 72 ($P < 0.0001$) |
| Language family | Among groups | 20 ($P = 0.15$) |
| | Among populations within groups | 14 ($P < 0.0001$) |
| | Within populations | 66 ($P < 0.0001$) |
| Language subfamily | Among groups | 31 ($P = 0.003$) |
| | Among populations within groups | 5 ($P < 0.0001$) |
| | Within populations | 64 ($P < 0.0001$) |

tonians and the Baltic populations, although the major boundary was between the western populations and the eastern/northern ones.

## Discussion

The present study provides an example of the power of a genealogical approach to Y chromosome analysis based on a hierarchical use of different markers to investigate human genetic variation and the elements that can affect it. The subdivision of the Y chromosomes into distinct lineages (haplogroups), defined by unique-event binary polymorphisms, was followed by investigation of the diversity within haplogroups using more variable loci, microsatellites. This provided information about the genetic background and relationship among populations, which can be considered in the light of their known history. Each haplogroup represents a unique lineage that has originated from a single man sometime in the past and somewhere in the world. The spread of each haplogroup is assumed to be unaffected by selection and to be the result of male migrations. The influences of factors like genetic founder effects, gene flow, and genetic drift, as well as geographical, linguistic, and cultural barriers, can be investigated. The overall distributions of Y haplogroups within Europe have now been documented (Rosser et al. 2000; Semino et al. 2000).

The origins of two populations, the Saami and the Gotlanders, were of particular interest, as were the more general questions about the relative importance of different factors in determining genetic boundaries between populations.

### Origin of the Saami

A striking finding was that the Saami Y chromosomes, characterized using either binary markers or microsatellites, were very similar to those of the Estonians (fig. 5) and distinct from those of their immediate neighbors, the Swedes and the Finns. One explanation would be that these chromosomes represent the ancestral pool for the northeastern Uralic-speaking populations, perhaps from the Ladoga and Onega Lake region, and that the Finns now differ because of their unique demographic history: a limited number of founders, isolation even in historical times, and a population bottleneck followed by rapid population growth. The $\Phi_{ST}$ values obtained from the microsatellite data were significantly different between the Finns and all of the other populations investigated and this is shown in figure 5, where the Finns always occupy an outlying location. The Y data, however, contrast with findings from other regions of the genome (Cavalli-Sforza, Menozzi, and Piazza 1994). In mtDNA, there is a distinct ''Saami motif'' that distinguishes a major proportion of Saami mtDNA from all other European lineages (Sajantila et al. 1995; Lahermo et al. 1996). Thus, there could be different genetic histories for males and females, due to different migration patterns or gene flow. For example, if there were high levels of male-mediated gene flow into the Saami, the original Y lineages, but not all of the original mtDNA lineages, could have been replaced.

### Origin of the Gotlanders

In contrast, the origin of the Gotland Y chromosomes is clear: using either binary or microsatellite markers, the Gotlanders and the Swedes form the most closely related pair of populations (fig. 5). They share 14 microsatellite haplotypes among 31 and 27, respectively, and the $\Phi_{ST}$ value between them is not significantly different from 0. Thus, the Gotlanders' Y chromosomes have a predominantly western origin. This conclusion again contrasts with the findings from other genomic regions. Although no information is available on Gotland mtDNA, some autosomal markers show evidence of gene flow from the east. The blood group gene $LW^b$ has been found at frequencies of about 6% in the Latvians and the Lithuanians, 4% in the Estonians, and 2.9% in the Finns, but it is present at a very low frequency in mainland Swedes (0.3%) and elsewhere in Western Europe (0%–0.1%) and is apparently absent from Asian and African populations (Sistonen et al. 1999). Thus, it can be considered a ''Baltic tribal marker.'' In Gotland, it is found at a frequency of 1.0%, perhaps reflecting female, rather than male, migration.

### Genetic Boundaries

Genetic differences can be correlated with cultural, linguistic, and geographical differences, but it is usually difficult to disentangle the individual effects of these factors because culture, language, and geography are all correlated with one another. Distinguishing individual

effects requires an informative genetic system and populations in which culture, language, and geography are not entirely correlated. The Y chromosome in northern Europe provides this.

AMOVA analysis of autosomal markers in worldwide populations has typically found that about 15% of the variation is between populations or groups of populations (Barbujani et al. 1997), while a similar analysis of Y-chromosomal markers found that the equivalent figure was 41% (Santos et al. 1999), illustrating the greater differentiation of Y markers. In a smaller geographical area, the populations are expected to be more similar, so 10%–13% (table 3) represents substantial genetic differentiation.

The one population with a distinct lifestyle, the Saami, does not show a distinct set of Y chromosomes. Cultural differences may not have led to genetic differentiation, or differences that had accumulated in the past may have been erased by subsequent gene flow. However, significant genetic differences between populations grouped according to language or geography are seen. The percentage of variation between groups is larger when grouping is according to geographical location (table 4), suggesting that geography has had the more important influence. This conclusion is supported by the genetic boundary of the Baltic Sea being stronger than that of the Estonian/Latvian border (fig. 6) and a consideration of the key populations where the geographical and linguistic boundaries do not correspond: the Latvians and the Lithuanians. Haplogroup frequencies, particularly for haplogroups 1 and 16 (fig. 3), resemble those in other eastern populations, not in other Indo-European speakers.

This finding raises a new question: what is the origin of the discrepancy? If haplogroup frequencies alone were considered, possible explanations would include replacement of an earlier Uralic language in these populations by the present Indo-European languages, or flow of Y chromosomes but not language from the north, or another, unsampled, region with the characteristic high frequency of haplogroup 16 and low frequency of haplogroup 1. There is no evidence for language replacement, and the differentiation of the haplogroup 16 chromosomes between the Baltic countries and the rest (fig. 4*b* and table 4) shows that the Latvian and Lithuanian haplogroup 16 chromosomes have not originated from recent gene flow. When haplogroup 16 chromosomes were omitted from the analysis, in order to understand to what extent the other lineages were distinct, a similar clustering effect for Latvians and Lithuanians was observed in an MDS analysis (data not shown). This result supports the idea that the genetic history of Y chromosomes within these two populations is distinct from that of the Uralic speakers. This conclusion does not challenge the earlier suggestion that haplogroup 16 chromosomes arose in Asia, but suggests that there were two distinct early migrations of haplogroup 16 chromosomes into Europe.

In summary, our interpretation of the Y-chromosomal data is that the major genetic difference in this area is geographical, distinguishing populations living on the western side of the Baltic from those on the eastern side. However, a significant difference was also detectable between Finno-Ugric speakers and Baltic speakers on the eastern side, where the Latvians showed greater genetic similarity to the Lithuanians than to the Estonians, demonstrating that linguistic differences can have a lesser, but still important, influence on the distribution of genetic diversity.

## Acknowledgments

We thank the donors of the DNA samples for making this work possible, and Vincent Macaulay, Guido Barbujani, and Mark Jobling for helpful explanations and comments. T.Z. was supported by the Wellcome Trust, and C.T.-S. was supported by the CRC.

## LITERATURE CITED

AYUB, Q., A. MOHYUDDIN, R. QAMAR, K. MAZHAR, T. ZERJAL, S. Q. MEHDI, and C. TYLER-SMITH. 2000. Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. Nucleic Acids Res **28**:e8.

BANDELT, H. J., P. FORSTER, and A. ROHL. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16**:37–48.

BARBUJANI, G. 2000. Geographic patterns: how to identify them and why. Hum. Biol. **72**:133–153.

BARBUJANI, G., A. MAGAGNI, E. MINCH, and L. L. CAVALLI-SFORZA. 1997. An apportionment of human DNA diversity. Proc. Natl. Acad. Sci. USA **94**:4516–4519.

BARBUJANI, G., N. L. ODEN, and R. R. SOKAL. 1989. Detecting regions of abrupt change in maps of biological variables. Syst. Zool. **38**:376–389.

BECKMAN, L., C. SIKSTROM, A. V. MIKELSAAR, A. KRUMINA, D. AMBRASIENE, V. KUCINSKAS, and G. BECKMAN. 1998. Transferrin variants as markers of migrations and admixture between populations in the Baltic Sea region. Hum. Hered. **48**:185–191.

BLANCO, P., M. SHLUMUKOVA, C. A. SARGENT, M. A. JOBLING, N. AFFARA, and M. E. HURLES. 2000. Divergent outcomes of intra-chromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. J. Med. Genet. **37**:752–758.

BOSCH, E., F. CALAFELL, F. R. SANTOS, A. PÉREZ-LEZAUN, D. COMAS, N. BENCHEMSI, C. TYLER-SMITH, and J. BERTRANPETIT. 1999. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am. J. Hum. Genet. **65**:1623–1638.

CARPELAN, C. 1998. Die Herkunft der Finnen und Saamen im Licht der Archäologie (20 000–500 v. Chr). Jahrb. Finn. Dtsch. Literaturbezieh. Helsinki **30**:31–39.

CAVALLI-SFORZA, L. L., P. MENOZZI, and A. PIAZZA. 1994. The history and geography of human genes. Princeton University Press, Princeton, N.J.

DE KNIJFF, P., M. KAYSER, A. CAGLIÁ et al. (25 co-authors). 1997. Chromosome Y microsatellites: population genetic and evolutionary aspects. Int. J. Legal Med. **110**:134–149.

DE LA CHAPELLE, A. 1993. Disease gene mapping in isolated human populations: the example of Finland. J. Med. Genet. **30**:857–865.

ERIKSSON, A. W. 1988. Anthropology and health of the Lapps. Coll. Anthropol. **12**:197–235.

EXCOFFIER, L., P. E. SMOUSE, and J. M. QUATTRO. 1992. Analysis of molecular variance inferred from metric distances

among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131**:479–491.

HAJDÚ, P. 1976. Ancient cultures of the Uralian peoples. Corvina Press, Budapest, Hungary.

HAMMER, M. F., and S. HORAI. 1995. Y chromosomal DNA variation and the peopling of Japan. Am. J. Hum. Genet. **56**:951–962.

HILL, E. W., M. A. JOBLING, and D. G. BRADLEY. 2000. Y-chromosome variation and Irish origins. Nature **404**:351–352.

HURLES, M. E. 1999. Mutation and variability of the human Y chromosome. Ph.D. thesis, Department of Genetics, University of Leicester, Leicester, England.

HURLES, M. E., C. IRVEN, J. NICHOLSON, P. G. TAYLOR, F. R. SANTOS, J. LOUGHLIN, M. A. JOBLING, and B. C. SYKES. 1998. European Y-chromosomal lineages in Polynesians: a contrast to the population structure revealed by mtDNA. Am. J. Hum. Genet. **63**:1793–1806.

HURLES, M. E., R. VEITIA, E. ARROYO et al. (18 co-authors). 1999. Recent male-mediated gene flow over a linguistic barrier in Iberia, suggested by analysis of a Y-chromosomal DNA polymorphism. Am. J. Hum. Genet. **65**:1437–1448.

JOBLING, M. A., A. PANDYA, and C. TYLER-SMITH. 1997. The Y chromosome in forensic analysis and paternity testing. Int. J. Legal Med. **110**:118–124.

JOBLING, M. A., and C. TYLER-SMITH. 1995. Fathers and sons: the Y chromosome and human evolution. Trends. Genet. **11**: 449–456.

———. 2000. New uses for new haplotypes: the human Y chromosomes, disease and selection. Trends. Genet. **16**: 356–362.

JORDE, L. B., M. BAMSHAD, and A. R. ROGERS. 1998. Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays **20**:126–136.

KARAFET, T. M., S. L. ZEGURA, O. POSUKH et al. (14 co-authors). 1999. Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. Am. J. Hum. Genet. **64**: 817–831.

LAHERMO, P., A. SAJANTILA, P. SISTONEN, M. LUKKA, P. AULA, L. PELTONEN, and M. L. SAVONTAUS. 1996. The genetic relationship between the Finns and the Finnish Saami (Lapps): analysis of nuclear DNA and mtDNA. Am. J. Hum. Genet. **58**:1309–1322.

LAHERMO, P., M. L. SAVONTAUS, P. SISTONEN, J. BERES, P. DE KNIJFF, P. AULA, and A. SAJANTILA. 1999. Y chromosomal polymorphisms reveal founding lineages in the Finns and the Saami. Eur. J. Hum. Genet. **7**:447–458.

LARSSON, L. 1996. The earliest settlement of Scandinavia and its relationship with neighbouring areas. Almquist and Wiksell International, Stockholm, Sweden.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

PANDYA, A. 1998. Human Y-chromosomal DNA variation. D.Phil. thesis, Department of Biochemistry, University of Oxford, Oxford, England.

ROSSER, Z. H., T. ZERJAL, M. E. HURLES et al. (63 co-authors). 2000. Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet. **67**:1526–1543.

SAJANTILA, A., P. LAHERMO, T. ANTTINEN et al. (13 co-authors). 1995. Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res. **5**:42–52.

SAJANTILA, A., A. H. SALEM, P. SAVOLAINEN, K. BAUER, C. GIERIG, and S. PAABO. 1996. Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. USA **93**:12035–12039.

SANTOS, F. R., A. PANDYA, C. TYLER-SMITH, S. D. PENA, M. SCHANFIELD, W. R. LEONARD, L. OSIPOVA, M. H. CRAWFORD, and R. J. MITCHELL. 1999. The central Siberian origin for native American Y chromosomes. Am. J. Hum. Genet. **64**:619–628.

SEIELSTAD, M. T., J. M. HEBERT, A. A. LIN, P. A. UNDERHILL, M. IBRAHIM, D. VOLLRATH, and L. L. CAVALLI-SFORZA. 1994. Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum. Mol. Genet. **3**:2159–1261.

SEMINO, O., G. PASSARINO, P. J. OEFNER et al. (17 co-authors). 2000. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science **290**:1155–1159.

SIMONI, L., F. CALAFELL, D. PETTENER, J. BERTRANPETIT, and G. BARBUJANI. 2000. Reconstruction of prehistory on the basis of genetic data. Am. J. Hum. Genet. **66**:1177–1179.

SISTONEN, P., K. VIRTARANTA-KNOWLES, R. DENISOVA, V. KUCINSKAS, D. AMBRASIENE, and L. BECKMAN. 1999. The $LW^b$ blood group as a marker of prehistoric Baltic migrations and admixture. Hum. Hered. **49**:154–158.

THOMAS, M. G., N. BRADMAN, and H. M. FLINN. 1999. High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum. Genet. **105**:577–581.

TØNNESON, K. 1994. Norden og Baltikum, Rapport 1, Det 22. Nordiske historikermöte, Oslo, Norway.

UNDERHILL, P. A., L. JIN, A. A. LIN, S. Q. MEHDI, T. JENKINS, D. VOLLRATH, R. W. DAVIS, L. L. CAVALLI-SFORZA, and P. J. OEFNER. 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. **7**:996–1005.

UNDERHILL, P. A., P. SHEN, A. A. LIN et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. Nat. Genet. **26**:358–361.

WHITFIELD, L. S., J. E. SULSTON, and P. N. GOODFELLOW. 1995. Sequence variation of the human Y chromosome. Nature **378**:379–380.

ZERJAL, T., B. DASHNYAM, A. PANDYA et al. (18 co-authors). 1997. Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am. J. Hum. Genet. **60**:1174–1183.

ZERJAL, T., A. PANDYA, F. R. SANTOS et al. (17 co-authors). 1999. The use of Y-chromosomal DNA variation to investigate population history: recent male spread in Asia and Europe. Pp. 91–101 *in* S. S. PAPIHA, R. DEKA, and R. CHAKRABORTY, eds. Genomic diversity: applications in human population genetics. Kluwer Academic/Plenum Publishers, New York.

PEKKA PAMILO, reviewing editor

Accepted February 15, 2001