Phylogeography of mitochondrial DNA in western Europe

M. B. RICHARDS*, V. A. MACAULAY†, H.-J. BANDELT‡ and B. C. SYKES§

* Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DJ, United Kingdom
† Mathematisches Seminar, Universität Hamburg, Bundesstraße 55, D-20146 Hamburg, Germany

(Received 16.1.98. Accepted 27.4.98)

SUMMARY

For most of the past century, prehistorians have had to rely on the fossil and archaeological records in order to reconstruct the past. In the last few decades, this evidence has been substantially supplemented from classical human genetics. More recently, phylogenetic analyses of DNA sequences that incorporate geographical information have provided a high-resolution tool for the investigation of prehistoric demographic events, such as founder effects and population expansions. These events can be dated using a molecular clock when the mutation rate and founder haplotypes are known. We have previously applied such methods to sequence data from the mitochondrial DNA control region, to suggest that most extant mitochondrial sequences in western Europe have a local ancestry in the Early Upper Palaeolithic, with a smaller proportion arriving from the Near East in the Neolithic. Here, we describe a cladistic notation for mitochondrial variation and expand upon our earlier analysis to present a more detailed portrait of the European mitochondrial record.

INTRODUCTION

Europe is a small continent, a Eurasian peninsula with a peripheral relationship in the Palaeolithic to core regions in both Asia and Africa (Gamble, 1986). Nevertheless, classical analyses of gene frequency data have indicated some patterns within the region. In particular, gradients of allele frequencies have been detected by means of principal component and spatial autocorrelation analyses running south-east to north-west, which have been interpreted as indicating a substantial influx of early farming communities from the Near East in the early Neolithic (Menozzi et al. 1978; Sokal et al. 1991). In this respect, however, population genetics has come into conflict with archaeology, since in recent years evidence has accumulated from the archaeological and palaeobotanical records suggesting that the onset of farming in Europe was a heterogeneous and substantially indigenous development (Whittle, 1996).

Mitochondrial DNA (mtDNA) allows us to approach this issue from a fresh perspective. MtDNA, which is maternally inherited, fast-evolving and non-recombining, allows one to take a fundamentally different approach from that of classical population genetics (Wilson et al. 1985). DNA sequence variation, sampled in the present, can be analysed historically by reconstructing a phylogenetic tree – or several alternative trees arranged in a network – that displays the inferred genealogical relationships between individual sequences. The structure of this gene tree contains demographic information which, in conjunction with a calibrated mutation rate for the DNA sequences under study, can be used to estimate a time-scale for events in human prehistory. Moreover, using what has become known as the 'phylogeographic' approach (Avise
et al. 1987), the geographical distribution of the lineages on a tree or network can be used to detect prehistoric movements from one region to another.

Our recent analysis of 757 individuals from various parts of Europe suggested that most of the extant mitochondrial sequences in Europe (probably around 85%) had their origins in the European Upper Palaeolithic, and that the remaining 15% or so arrived more recently in Europe from the Near East (Richards et al. 1996). This analysis would therefore concur with the archaeological evidence that the initial Neolithic immigration from the Near East was relatively minor (Zvelebil, 1989). We concluded that the interpretation of the genetic gradients in Europe in terms of a ‘wave of advance’ of Near-Eastern populations into the continent (Ammerman & Cavalli-Sforza, 1984) may have been too bold, and the contribution of the indigenous Mesolithic peoples substantially underestimated.

Several criticisms can be made of our study, concerning the sampling procedures, phylogenetic analyses, age estimates and conclusions drawn from cluster frequency information. They include: (1) a sampling bias towards the more northerly and westerly extremes of Europe; (2) an incomplete phylogenetic analysis, encompassing only the common sequences, rather than the entire data set; (3) the assumption that cluster frequencies in present-day Europe resemble those thousands of years ago, despite the intervening action of genetic drift; (4) an inadequate founder analysis due to paucity of Near Eastern data.

Here we address the first three of these points using the earlier published sequences and incorporating a further 189 individuals from Iberia, Italy and Bulgaria into the analysis. We present phylogenetic networks for each major European lineage cluster (using a new cladistic notation for the cluster classification scheme) and an analysis of the time depth of each cluster using the age estimator, \( \rho \). This statistic is more powerful than the commonly-used mean of pairwise differences since it takes into account more prior information, by requiring the explicit designation of an ancestral haplotype (Morral et al. 1994; Forster et al. 1996). Finally, we model a population with plausible post-Neolithic demographics to show that the effects of drift are unlikely to have been sufficient to render invalid our conclusion that the low frequency of Near-Eastern-derived lineages extant in Europe reflects relatively low immigration during the Neolithic.

**Subjects and Methods**

**Cladistic notation for mitochondrial clusters**

Nomenclature for clusters of mitochondrial sequences has run into some difficulties, as a number of distinct classification systems have arisen in different publications as a result of different researchers working with imperfect phylogenetic analyses and data sets. We propose here a cladistic notation based on set theory, with the aim of providing a partial codification of a reliable phylogenetic tree, for reference in studies focusing on single populations. The classification also provides motifs, or signature mutations, for clusters. Ideally, such a system would be based on complete mitochondrial sequences from a large world-wide sample. Since complete mtDNA sequences are extremely rare at present and probably will remain so in the near future, the coding and classification system can only refer to the current information provided by the hypervariable segments of the control region, RFLPs, and particular mitochondrial coding-region sequences. The system aims to be sufficiently conservative to avoid confusion, but at the same time flexible enough to adjust to the evolution of the database, with the concomitant greater precision of phylogeny reconstruction, and to the obvious needs of phylogeographic analysis. We depart mainly from the RFLP nomenclature since it is the most well established, and because 14-enzyme RFLP analysis has hitherto provided greater phylogenetic resolution than other studies in most cases.

Following the RFLP nomenclature, therefore, we denote principal clusters by capital letters.
Nesting of clusters is allowed, e.g. $C \subset M$, or even $M \subset L_3$. Subclusters of single-letter-coded clusters may have a non-negative integer suffix. For subsequent levels of the hierarchy, a hierarchical notation may be used if desired, although this may be unnecessary if relatively few subclusters need to be distinguished. Following a hierarchical notation, the next layer of clusters would receive a small letter as a suffix; for all following levels, numbers and small letters alternate. Thus, the code signifies subcluster relationships, such as $Z_7e4a \subset Z_7e4 \subset Z_7c \subset Z_7 \subset Z$.

Every cluster thus codified should constitute a monophyletic clade in the human mtDNA phylogeny. If, in the course of obtaining more information, a cluster thought to be of this kind turns out to be paraphyletic, it should either be enlarged or shrunk. As an example, the new evidence of Hofmann et al. (1997) indicates that the previous RFLP haplogroup U (Torroni et al. 1996) should be enlarged to embrace haplogroup K as well. Further, the African-specific group L1 as defined by Chen et al. (1995) contains the root of the human mtDNA phylogeny (Watson et al. 1997) and therefore should be partitioned into monophyletic clades as soon as sufficient information becomes available.

To economize on letters, and develop the RFLP haplogroup notation, a clade may be referred to using the names of its prominent subclades. $YZ$ will refer to the smallest monophyletic clade including $Y$ and $Z$, provided that the clade does not contain any named subclades that are not already included in $Y$ or $Z$. For example, one could not refer to $HU$, since the smallest clade containing $H$ and $U$ would also include $V$, but $HVU$ might well be a feasible designation. If it turns out that $YZ$ is indeed a feasible notation, then so too is pre-$YZ$, which denotes the largest clade including $YZ$ but excluding every other named clade which is not already part of $Y$ or $Z$.

In order to refer to specific descendants (in a geographical area, say) of a certain sequence type, it may be desirable to accept paraphyletic clusters as temporarily defined groups, although such clusters need to be clearly distinguished from the principal clades. In such a case an exponent suffix, such as $*$, can be used: e.g. if cluster $Z$ has only named subclusters $Z_1$, $Z_2$, and $Z_3$, then $Z^*$ comprises all $Z$ sequences outside $Z_1$, $Z_2$ or $Z_3$, that is: $Z^* = Z - (Z_1 \cup Z_2 \cup Z_3)$. One should bear in mind that these derived cluster notations always have a preliminary status, which may change when more sequence information becomes available.

Details of the 942 sequences used in this study can be found in Richards et al. (1996) and Corte-Real et al. (1996), which incorporate the data of Piercy et al. (1993), Di Rienzo & Wilson (1991), Pult et al. (1994), Bertranpetit et al. (1995), and Francalacci et al. (1996). Four individuals from the data of Pult et al. (1994) were omitted since they shared recent maternal ancestry with other subjects of the same sample (I. Pult, personal communication). In addition, 30 Bulgarians (Calafell et al. 1996) were included. Several Iberian samples with the cluster X motif were tested for position 3592 HpaI status (defining the African-specific RFLP haplogroups L1 and L2: Chen et al. 1995; Watson et al. 1997).

Phylogenetic networks, which reflect ambiguity in the branching structure by incorporating reticulations, were constructed by means of the median algorithm and were reduced (or simplified) using the frequency and compatibility criteria of Bandelt et al. (1995). The program Network (Röhl & Minh, 1997) was used to compute the networks. A reduction rule, the cube rule, additional to those of Bandelt et al. (1995) was applied manually: a character, $a$, incompatible with two other characters, $b$ and $c$, themselves incompatible, was resolved into two independent events if (i) the character states $(abc)$ present were $(000)$, $(001)$, $(010)$, $(011)$, $(100)$, and $(111)$, (ii) the frequency favoured the $a \bar{0}$ haplotypes over those with $a \bar{1}$ (explicitly, $f(000)+f(001)+f(010)>f(100)$ and $f(011)+f(001)+f(010)>f(111)$), and (iii) the added weight of $b$ and $c$ was more than twice the weight...
of a. Triangles were added by hand in cases where more than two bases at a particular position led to ambiguity of the order of mutations. The networks were constructed using sequence data from the first hypervariable segment (HVS I) of the control region (positions 16090–16365: Anderson et al. 1981), using weights based on the list of Hasegawa et al. (1993). Weights were chosen in geometric progression with factor 2 in order to allow resolution of incompatibilities between characters of different weight classes. We divided the list of nucleotide positions into three classes of transition rate, fast (16129, 16189, 16311, 16362), intermediate (16093, 16172, 16209, 16223, 16278, 16293, 16294, 16325) and slow (the remainder of positions between 16090 and 16365), and assigned the classes weights of \(0 + 25\), \(0 + 5\), and \(1 + 0\) respectively. Transversions were weighted \(2 + 0\), except for those adjacent to the polycytosine tract 16184–16193, which were ignored as probable artefacts of length variation (Bendall & Sykes, 1995). Control-region sequence clusters were identified with the RFLP haplogroups of Torroni et al. (1994; 1996), using the data set comparisons of Torroni et al. (1996) and Macaulay et al. (1998).

The time to the most recent common ancestor of each cluster was estimated using \(\rho\), the average transitional distance from the putative founder sequence, and calibrated using a transition rate of 1 in 20180 years (Forster et al. 1996), here rounded to 1 in 20200 years.

The geographic origin of samples was classified by region using the ecological model of Gamble (1986): North-Central (northern Germany and Denmark), Alpine (Bavaria and Switzerland), Mediterranean-West (Iberia, including the north but excluding the Basque country); Mediterranean-Central (Italy and Sardinia). The Basques, who are genetically somewhat distinct from other Europeans (Bertranpetit & Cavalli-Sforza, 1991; Cörte-Real et al. 1996) were represented separately, as was Iceland. We included two small data sets from the North-East (Finns) and South-East (Bulgarians) (Calafell et al. 1996) for comparison but excluded Saami (Sajantila et al. 1995) and Ladins (Stenico et al. 1996) since, atypically for European populations, they appear to have undergone drastic founder effects and drift.

The skeleton network, constructed using only data scored for position 00073 in addition to HVS I, corrects the northern bias of the data set of Richards et al. (1996) by including a similar number of samples from south-west Europe: for Britain, \(n = 100\) (Piercy et al. 1993); for non-Basque Iberia, \(n = 125\) (Cörte-Real et al. 1996); the Basque country is included separately, \(n = 61\) (Cörte-Real et al. 1996), as is Italy, \(n = 49\) (Francalacci et al. 1996) and Bulgaria, \(n = 30\) (Calafell et al. 1996).

We report HVS I sequence haplotypes, between positions 16090 and 16365 (except in the case of cluster J, in which the motif position 16069 was also included), in terms of transitions from the Cambridge Reference Sequence (CRS: Anderson et al. 1981), e.g. ‘16129–16223–16311’ describes an HVS I sequence which differs from the CRS by transitions at positions 16129, 16223, and 16311. Transversions are further indicated in superscript by the nucleotide change, e.g. 16129\(^{\text{A-T}}\), and insertions by the adjacent 5' position plus the inserted nucleotide in superscript, e.g. 16289\(^{+\text{C}}\). No deletions were found.

**Effect of drift on the proportion of Neolithic mtDNA**

In order to test the potential of genetic drift since the beginning of the Neolithic to obscure cluster frequency information, we quantitated the effect of drift under a number of simple demographic scenarios, as follows:

(a) Exponential expansion throughout the Neolithic from an effective female population size of 45000 (based on an initial census size of 250000: McEvedy & Jones, 1978) to one of 360000 at the end of the Neolithic (based on a census size of two million: McEvedy & Jones, 1978), followed by a constant population size until the present day. This corresponds to approximately eight-fold
growth during the Neolithic, or about 2% per generation; 3% is about the maximum expected (Cavalli-Sforza et al. 1994).

(b) Exponential expansion from the Neolithic transition to 1950 from an initial effective female population size of 45000 as in (a) to one of 81 million (census size of about 500 million: McEvedy & Jones, 1978). This amounts to about 2.3% growth per generation.

(c) As (a) but starting with a 10-fold lower initial effective female population size, which rises to 350000 and then remains constant. This amounts to 80-fold growth, i.e. 4.2% per generation.

(d) A constant female effective population size of 45000 subdivided into 10 subpopulations of size 4500. The Neolithic was taken to run from 8300 to 5700 years ago, that is, 104 generations (assuming one generation = 25 years). We estimated the probability distribution of the proportion of mtDNA introduced at the start of the Neolithic,
Fig. 2. Schematic tree for European mitochondrial variation. Clusters of sequences comprising named clades are outlined and labelled. The node marked CRS corresponds to the Cambridge Reference Sequence. The branches are labelled with HVS I motif mutations if present, and otherwise by HVS II motif or coding-region mutations in square brackets. Note that motif positions may occasionally revert within a cluster, particularly in the case of the rapidly-evolving positions in the control region. L3a is a large clade of African origin, based on the root sequence 16223 (Watson et al. 1997), which encompasses virtually all Eurasian mitochondrial variation, as well as several subclusters that have remained restricted to Africa.
Table 1. Principal mitochondrial clusters in western Europeans and their HVS I motifs relative to the CRS

<table>
<thead>
<tr>
<th>Cluster</th>
<th>00073-HVS I motif</th>
<th>Former nomenclature&lt;sup&gt;b,c,d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td></td>
<td>Subset of 00073A group 1</td>
</tr>
<tr>
<td>V</td>
<td>16298</td>
<td>Sub-set of 00073A group 1</td>
</tr>
<tr>
<td>U</td>
<td>00073</td>
<td>00073G group 1, 4, 5, 6</td>
</tr>
<tr>
<td>K</td>
<td>00073-16224–16311</td>
<td>Group 4</td>
</tr>
<tr>
<td>U3</td>
<td>00073-16343</td>
<td>Sub-set of 00073G group 1</td>
</tr>
<tr>
<td>U4</td>
<td>00073-16356</td>
<td>Group 5</td>
</tr>
<tr>
<td>U5</td>
<td>00073-16270</td>
<td>Sub-set of group 5</td>
</tr>
<tr>
<td>U5a</td>
<td>00073–16192–16270</td>
<td>Sub-set of group 5</td>
</tr>
<tr>
<td>U5a1</td>
<td>00073–16192–16256–16270</td>
<td>Sub-set of group 5</td>
</tr>
<tr>
<td>U5a1a</td>
<td>00073–16256–16270</td>
<td>Sub-set of group 5</td>
</tr>
<tr>
<td>U5b</td>
<td>00073–16189–16270</td>
<td>Sub-set of group 5</td>
</tr>
<tr>
<td>U5b1</td>
<td>00073–16144–16189–16270</td>
<td>Sub-set of group 5</td>
</tr>
<tr>
<td>U6</td>
<td>00073–16172–16219</td>
<td>Group 6</td>
</tr>
<tr>
<td>J</td>
<td>00073–16069–16126</td>
<td>Group 2A</td>
</tr>
<tr>
<td>J1</td>
<td>00073–16069–16126–16261</td>
<td>Group 2A-C</td>
</tr>
<tr>
<td>J1a</td>
<td>00073–16069–16126–16143–16231–16261</td>
<td>Sub-set of group 2A</td>
</tr>
<tr>
<td>J2</td>
<td>00073–16069–16126–16193</td>
<td>Group 2A</td>
</tr>
<tr>
<td>T</td>
<td>00073–16126–16294</td>
<td>Group 2B</td>
</tr>
<tr>
<td>T1</td>
<td>00073–16126–16163–16186–16189–16294</td>
<td>Group 2B</td>
</tr>
<tr>
<td>I</td>
<td>00073–16129–16223–16391</td>
<td>Group 3A</td>
</tr>
<tr>
<td>W</td>
<td>00073–16223–16292</td>
<td>Group 3C</td>
</tr>
<tr>
<td>X</td>
<td>00073–16223–16278</td>
<td>Group 3B</td>
</tr>
</tbody>
</table>

<sup>a</sup> Note that occasional reversions within the sequence motif may occur.
<sup>b</sup> Richards <i>et al</i>. 1996.
<sup>c</sup> Corê-Real <i>et al</i>. 1996.
<sup>d</sup> Wilkinson-Herbots <i>et al</i>. 1996.

α<sub>i</sub>, given the (inferred) present-day proportion of these sequences, α<sub>f</sub>, by evaluating the likelihood pr(α<sub>f</sub> | α<sub>i</sub>) [since pr(α<sub>i</sub> | α<sub>f</sub>) £ pr(α<sub>f</sub> | α<sub>i</sub>) when we assign a uniform prior distribution to α<sub>i</sub>]. From this distribution, we evaluated the central 95% credible region of α<sub>i</sub> (Berger, 1985). A computer program was written to determine pr(α<sub|i</sub> | x<sub>i</sub>) for a range of x<sub>i</sub> given each demographic scenario: each scenario was run 10000 times from each starting x<sub>i</sub> and the proportion that achieved x<sub>f</sub> (within a tolerance of ±0.005) counted. The proportion x<sub>f</sub> was taken from the current frequency of cluster J, as described below.

RESULTS

Phylogeographic analyses

Figure 1 shows a skeleton network of 42 sequence haplotypes (from 202 individuals) chosen from the complete data set using the following criteria: (1) they are present more than once in the data set (except for a singleton member of the infrequent cluster W); (2) they have been sequenced (or typed) for position 00073 in the second hypervariable segment of the control region (HVS II). Position 00073 is useful for distinguishing some members of cluster H from cluster U (Torrioni <i>et al</i>. 1996), a distinction which seems to resolve a fairly recent demographic expansion from a more ancient one.

The skeleton provides an overview of the west European mitochondrial phylogenetic structure, and illustrates the clusters. The overall topology reconstructed remains similar to Figure 2 of Richards <i>et al</i>. (1996) despite the incorporation of more south-western data, correcting the earlier north-western bias. Each sequence in the complete data set (with very few exceptions: see below) can be assigned to a cluster with a particular motif, and the resulting clusters can then be examined in more detail for phylogeographic information. The structure—very
Table 2. Sample size, n, and values of ρ (the average mutational distance to an assumed common ancestor) for each haplogroup.

The divergence time, \( t \), is calculated from \( \rho \) using a mutation rate of one transition per 20200 years (Forster et al. 1996), a value close to the widely-used divergence rate of 33% per million years (Ward et al. 1991). Values for \( J \) are presented both without subdivision and with subdivision into the major subclusters \( J^* \) and \( J_{1a} \) (see Fig. 7).

| Group   | H^a | V   | K   | U5  | J   | J^* | J_{1a} | T   | I   | W   | X   |
|---------|-----|-----|-----|-----|-----|-----|-------|-----|-----|-----|-----|-----|
| Ancestor | 173^a | 40  | 64  | 67  | 105 | 67  | 18    | 71  | 15  | 12  | 16  |
|         |   | 16126 | 16126 | 16126 | 16126 | 16126 | 16126 | 16126 | 16126 | 16223–16226 |

| \( \rho \) (transitions) | 1.01 | 0.63 | 0.77 | 2.61 | 1.39 | 0.40 | 0.33 | 2.30 | 1.73 | 0.92 | 1.19 |
| \( \Delta \rho \) (transitions)^b | 0.08 | 0.13 | 0.11 | 0.20 | 0.12 | 0.08 | 0.14 | 0.18 | 0.34 | 0.28 | 0.27 |
| \( t \) (years) | 20500 | 12500 | 15500 | 52500 | 28000 | 8000 | 6500 | 46500 | 35000 | 18500 | 24000 |
| \( \Delta t \) (years)^c | 2500 | 3000 | 3000 | 6500 | 4000 | 2000 | 3000 | 6000 | 8500 | 6500 | 6500 |

^a Data taken from the subset of the data set that has been typed for position 00073 in HVS II.

^b Root-mean-square error estimate for \( \rho \), calculated from \( \sqrt{(\rho/n)} \), i.e. assuming a perfect star phylogeny – when this is inappropriate, this error estimate should be considered a lower-bound on the true error.

^c Root-mean-square error estimate of \( t \) (to nearest 500 years), calculated by combining the estimated error in \( \rho \), \( \Delta \rho \), with the estimated error in the mutation rate of 5% (Forster et al. 1996).
roughly, two star-like phylogenies, one with fairly short-branch derivatives and one with longer branches – suggests two principal expansions, an ancient one based on the 00073G sequence and a more recent one based on 00073A (Wilkinson-Herbots et al. 1996). The overall topology receives support from both RFLP data (Torroni et al. 1996) and further analysis of almost one-third of the coding region (Hofmann et al. 1997), which indicates, however, that RFLP haplogroup K nests within cluster U, which we have therefore enlarged from the RFLP haplogroup U (as described in Methods). This misassignment arose due to the high rate of recurrent mutation in the region 10394–10398.

The overall mitochondrial tree for Europe is represented schematically in Figure 2 and the control-region sequence motifs listed in Table 1. The major clusters are labelled using our new scheme: this notation is intended to supersede the classification into sequence groups of Richards et al. (1996) and Côrte-Real et al. (1996), and those of Calafell et al. (1996) and Comas (1996), as well as modifying the RFLP notation of Torroni et al. (1996 and references therein). The root probably lies at the 16223 node (rather than the CRS, as suggested by Comas et al. 1997), which is ancestral to the clusters derived from both the CRS (designated JTUHV) and from 16223 (IWX). However, the exact root is uncertain; outgroup rooting using African sequences from the ancient RFLP haplogroups L1 and L2 (Chen et al. 1995) would place the root at the 16189–16223–16278 node, but since it seems that the major ‘out-of-Africa’ sequence haplotype was 16223 (Watson et al. 1997), it is likely that positions 16189 and 16278 (both of which have high mutation rates: Hasegawa et al. 1993; Wakeley, 1993) have subsequently reverted. Ages for the major clusters based on the statistic $\rho$ are given in Table 2.

Table 3 shows 16 sequences which have not been assigned to one of these clusters. Of these, seven fall within the African clusters of Watson et al. (1997), and one sequence (present in two individuals in the Mediterranean-West and Mediterranean-Central regions) has been assigned to the Asian haplogroup M (Ballinger et al. 1992; Macaulay et al. 1998). These outliers, as well as the three U6 sequences (see below), therefore represent recent admixture from Africa and Asia, which is then estimated at $\sim 1\%$. The three 16126-based sequences may belong to JT*, a singleton North-West sequence to cluster U, and there is a possible long-branch derivative of cluster I. The remaining 16223-based sequences may be related to clusters I, W and X.
Figure 3 shows clusters H and V. Since cluster H is the largest in Europe, comprising half of the data, it contains too many haplotypes in the present data set to be usefully represented on the page. In addition, there remains an ambiguity amongst sequence data not typed for position 00073 as to whether they are members of cluster H or U; the previous strategy of using ‘group 1’ as a default cluster (Richards et al. 1996) unwittingly included members of the more ancient clusters U*, U3, U4 and U6 as well as H, leading to an over-estimate of the age of the cluster. For this reason, as with the skeleton, for clusters H, V and U (excluding K and U5, which have well-defined HVS I motifs) we took only samples that had been characterized for position 00073. Cluster V (comprising 4% of the data) is included on the same diagram as H since the higher resolution of 14-enzyme RFLP analysis suggests that V and H are sister clusters (Torroni et al. 1998).

This cluster shows a pronounced star-like phylogeny with the centre of the star (the CRS) geographically widespread. Cluster H dates to
about 21000 years; while cluster V is 12500 years old (Table 2). These values are younger than those of Torroni et al. (1996), based on RFLP data, who used different rate estimates and diversity measures; Torroni et al. (1998) find little difference in the RFLP and HVS I ages. The high degree of reticulation shows that a considerable degree of homoplasy has occurred during this time. Nevertheless, most nodes in the network are filled and several subclusters are discernible. Most of these have frequently-occurring root sequence haplotypes, as one would predict for a recently expanded population (Donnelly & Tavare, 1986). These include subclusters based on sequence haplotypes 16129, 16304, 16291, 16311 and 16362, the transitions to which are therefore presumably amongst the oldest mutations in the cluster. All of these root sequences are present in both the Basque country and in Britain (the apparent exception, 16311, is present in the additional British data not typed for 00073). Members of the subclusters are also widely distributed. Since the Last Glacial Maximum occurred about 20000 years ago (Dansgaard et al. 1993), post-glacial expansion from south to north seems to be a plausible explanation for this pattern; the widespread distribution of the subclusters, as well as the cluster as a whole, suggests that the earlier mutations had occurred before the geographical expansion north, unless massive reflux has subsequently taken place, implying that the actual date for the expansion could be slightly more recent, perhaps ~ 15000 years ago (cf. Torroni et al. 1998). Note, however, that one cannot exclude the possibility that some minor parts of those subclusters were generated by independent mutations at the site in question. (We therefore do not recommend naming these putative clades unless they can be supported by additional characters.)

Figure 4 shows cluster K, a subgroup of cluster U, which accounts for nearly 7% of the total data. This shows an early split at position 16093, but otherwise is exceptionally star-like. The widespread geographic distribution and age (16000 years) are also similar to H/V, suggesting a similar explanation in post-glacial expansion. The RFLP-based divergence time for K of 13500–17500 (Torroni et al. 1994) agrees well with our estimate.
Figure 5 shows cluster U5; the remainder of cluster U, with the position of clusters K and U5 indicated, is shown separately (using the restricted data set of 00073-typed samples) in figure 6. The U5 cluster, comprising more than 7% of the data, dates to around 50000 years in Europe. The considerable time depth of U5 is reflected in the ambiguity near the root of the U5 phylogeny. Nevertheless, certain more plausible pathways in the network can be inferred. Cluster U5 as a whole has the motif 16270, U5a has an additional transition at 16192, and U5a1 also has a transition at 16256. A small additional sub-cluster, U5a1a, lacks the transition at 16192, and is therefore defined simply by the motif 16256–16270. Its lack of diversity and geographic restrictedness (mainly to North-West and North-Central Europe) imply a recent origin within the last 20000 years. A transition at a site outside the region of HVS I usually studied, but nevertheless assayed in some data sets, at position 16399, is shared between all members of U5a1a and some members of U5a1, implying that the 16256–16270 haplotype evolved by back-mutation at position 16192 from the 16192–16256–16270 haplotype (rather than via a parallel mutation at position 16256 from the 16270 haplotype). Subclusters U5a and U5a1 date to between approximately 40000 and 30000 years, implying that they evolved and expanded...
Fig. 6. Reduced median network of cluster U, using the restricted data set as in Figures 1 and 3, presented as above, with the CRS indicated with *. Haplogroups K and U5, which are subclusters of U, are indicated with triangles.

well before the Last Glacial Maximum. In addition to its considerable age, U5 has an interesting geographical distribution. U5* is found in both northern and southern Europe, whereas U5a* appears to be mainly restricted to southern Europe, with some diverged individuals present in the North-West, and U5a1* is found mainly in North-West Europe. An apparent exception to this pattern, a subcluster of U5a defined by transitions at 16189–16192–16270, which occurs only in the Alpine and North-Central regions and Finland, has been assigned to U5a* by the reduction procedure on the basis of the weighting scheme but, on the evidence of its geographic distribution, is more likely to belong to U5b. U5b itself, defined by the motif 16189–16270, is widely distributed in western and central Europe, but a derived subcluster, U5b1, with the motif 16144–16189–16270, is restricted to the Saami (where it has been elevated to high frequency) and to neighbouring populations (Sajantila et al. 1995). Figure 6 shows that the remainder of cluster U, including the ancestral sequence, occurs mainly in the south, and possibly also (given the high frequency in Bulgaria) to the east. It cannot be dated, since there is no evidence that it forms a distinct clade, but it comprises a number of distinct subclusters which will need to be analysed separately when more data is available. The largest, U4, dates to more than 25000 years, suggesting an expansion before the Last Glacial Maximum, as for U5. The most distinctive, U6, occurs at high frequency in Berber-speaking North Africans, and is likely to represent historically-attested introgression into Iberia (Córte-Real et al. 1996; Rando et al. 1998), since it is absent from the rest of Europe.
Figure 7 shows cluster J, comprising more than 11% of the total data. In many ways, this group has the most striking phylogeographic distribution (Richards et al. 1996). There appear to have been several distinct ancestors to this cluster in Europe which arrived from the Near East within the last 10000 years. A series of nodes in the network, including some unoccupied by extant sequences in the European data set, are also present in the Near East (marked +) where they have diversified into subclusters to a considerably greater extent than in Europe (Richards et al. 1996). Figure 7 shows, in particular, two empty branching nodes, characterised by transitions at 16069–16126–16145–16261 in one case and 16069–16126–16193 in the other. The first sequence is found only in the Near East, as shown previously (Richards et al. 1996); it is ancestral to a small star-like subcluster, J1a, characterised by a transition at 16231, found mainly in Alpine and North-Central Europe and dating to only about 7000 years. A closely related sequence, 16069–16126–16145–16222–16261, giving rise to subcluster J1b, is also found mainly in the Near East but also, rarely, in southern Iberia (Córté-Reale et al. 1996), and is the ancestor of a small subcluster (with an additional 16172 transition), J1b1, found as yet only in Britain, with a derivative from the ancestor also occurring in Italy. The second empty node, 16069–16126–16193, giving rise to subcluster J2, is extant in Turkey (Calafell et al. 1996) and has derivatives in Turkey, Italy, Sardinia, Iberia and Iceland.

Neither of these minor clusters localised in southern and western Europe includes sufficient sequences for a meaningful age estimate. However, the major founder subcluster, J*, 16069–16126 and its recent derivatives, is present in 67 individuals and can therefore be dated with
somewhat more confidence. This subcluster, which is widely distributed in western and central Europe, but is rare in Iberia, is around 8000 years old in Europe. This further analysis therefore supports the suggestion that these lineages were introduced from the Near East when the Neolithic economy spread into Europe, starting about 10000 years ago, with J2 and J1b spreading with J* along the Mediterranean coastline and into Atlantic Europe, and J1a spreading, again with J*, into central and northern Europe. The overall age of cluster J in Europe, at 28000 years (reflecting its origin in the Near East), is somewhat older than that calculated previously from RFLP data (15000–19000 years or 17000–23000 years: Torroni et al. 1994; Torroni et al. 1996); however, a more recent estimate based solely on Italian samples yields the older divergence time of 35000–46000 years (Torroni et al. 1997). This reflects the uncertainty of dating clusters which show multiple founder haplotypes and a pronounced geographic substructure, since the estimated age for J as a whole depends strongly on the extent to which the various subclusters are present in any given sample. The mismatch diagram for cluster J (not shown), which serves as a summary of the phylogeny, also indicates the ancestral heterogeneity by virtue of its bimodality, in contrast to the diagrams for each of the other major European clusters, which are all smooth and unimodal.

Figure 8 shows cluster T, comprising nearly 8% of the data. This cluster has a common origin in the Near East with cluster J, with which it shares the transition at 16126 and several coding-region variants (Hofmann et al. 1997; Torroni et al. 1997). There are two distinct subclusters, one of which, T*, includes a four-cycle adjacent to the putative root, 16126–16294.
The overall age of the cluster is at least 46500 years. The minor subcluster, T1, is defined by 16126–16163–16186–16189–16294. Treating these subclusters separately would result in an estimate of $\sim$32000 years for T* and 9000 years for T1. The RFLP data previously suggested a much more recent divergence time of 8000–11000 years (Torroni et al. 1996), although, as with J, a more recent estimate based solely on an Italian sample suggests the older age of 22000–29000 years (Torroni et al. 1997); the lower estimates may simply have been an artefact of small sample size. Evidence from the Near East suggests that there may indeed have been several founders for cluster T (Richards et al. 1998).

Figure 9 shows clusters I, W and X, characterised by the ancestral state at position 16223 (which was probably the major ‘out-of-Africa’ sequence: Calafell et al. 1996; Watson et al. 1997). Together, these clusters comprise only 5% of the total data set. Cluster I may be fairly ancient: $\sim$35000 years old in Europe (cf. 26000–34000 years: Torroni et al. 1994), although this age strongly depends on the time of appearance of the mutation at 16311 on the 16129–16223 background. Cluster I is now very rare, occurring at only $\sim$2% in the present data set, and seemingly distributed mainly in the north and west of Europe. Cluster W has a more recent age of 18500 years (although W may, as with T, break down into several founder clusters of more recent age) and appears more diverse in southern than northern Europe, but this may again simply reflect its rarity—only $\sim$1% of the present data. The 16189–16223–16278 signature characteristic of most of cluster X seems to have been at the centre of an expansion, forming a simple star-like phylogeny, with an age of $\sim$24000 years (although the low frequency of these lineages, at again only $\sim$2%, again render any conclusion extremely tentative). Surprisingly, cluster X also occurs at low frequency amongst native North Americans (Ward et al. 1991; Bandelt et al. 1995; Scozzari et al. 1997). This may suggest a common origin with European cluster X in the vicinity of the Near East.

The other sequence types with the 16223–16278 motif, listed in Table 3, are all closely related to modern African lineages (Watson et al. 1997), although the 16278 mutation in X probably occurred independently on the 16223 ‘out-of-Africa’ sequence type, given the lack of X in Africa and the clustering of X with W and I when more coding-region data is used (e.g., Hofmann et al. 1997). Two sequences, from Sardinia and Portugal, are members of RFLP haplogroup L2 (confirmed by testing for the HpaI site at position 3592 characterizing L1 and L2 in Africans: Chen et al. 1995). One, from Iberia, is a one-step derivative of the most frequent and widespread member of L3b. An individual from North Germany, one from Britain and one from Sardinia are members of L1, and the 16209–16223–16311 sequence is a member of an African subcluster of L3a, and indeed is found in a Portuguese subject with Angolan ancestry.

**Effect of drift on the proportion of Neolithic mtDNA**

It is possible that, as a result of genetic drift, the present-day frequency of cluster J no longer reflects the frequency at which it was found in Europe immediately following the proposed introgression into the European population at the beginning of the Neolithic. We used simple demographic scenarios to test whether or not this was likely. With the present-day proportion of mtDNA introduced at the beginning of the Neolithic into the European Mesolithic population taken as 12% (corresponding to cluster J), and an initial effective female population size of 45000, followed by 2% growth throughout the Neolithic, the 95% credible region for the frequency of J at the start of the Neolithic was 9–14%. Allowing the population to continue expanding (at 2% per generation) to the present led to a range of 10–14%. For a 10-fold lower initial population of 4500 expanding 80-fold in the Neolithic, drift was greater and the range broadened to 7–17%. Even for a small constant population of 45000, subdivided into 10 subpopulations, since the start of the Neolithic,
the range was 6–17%. If additional subsets of the modern mtDNA gene pool were to be identified as of Neolithic origin, it is still very unlikely that the Neolithic component was ever the majority; these analyses were repeated assuming twice the currently identified Neolithic component, namely 24%, and none of the credible regions extended above 50%. It appears, therefore, that genetic drift cannot have dramatically transformed the proportion of putative founding Neolithic mtDNAs.

DISCUSSION

In this paper we have (a) reconstructed a western European skeleton network that exhibits less geographical bias than previously and incorporates the important motif position 00073 from HVS II; (b) formulated a new cladistic notation for the mitochondrial diversity in Europe which can be generalized to world-wide variation; (c) constructed complete reduced median networks for each of the major European
clusters, which we have examined for geographic
distribution of sequences and dated using the
statistic \( \rho \); (d) verified under a number of
demographic scenarios that the cluster frequency
information can be considered meaningful and
not merely the product of drift.

The results support earlier conclusions con-
cerning those aspects of European prehistory
which are preserved in the mitochondrial record.
It now seems likely that all of the clusters we
have analysed underwent expansions during
European prehistory, albeit to differing extents.
Furthermore, most of the principal motif control-
region mutations seem to have occurred during
the Early Upper Palaeolithic, 20000–50000 years
ago, or even earlier. There remain traces today,
albeit minor, of the first settlement of the
continent by anatomically modern humans,
40000–50000 years ago. Cluster U5, which is the
only ancient cluster that is specific to Europe
(Richards et al. 1998), may have been the only
cluster to arrive in Europe with the first wave of
settlers. This comprises only 7% of extant
western European lineages. A more precise
analysis of the time of entry of the other clusters
requires a more exact identification of founder
sequences for each cluster by comparison with a
larger sample from the Near East, which is the
subject of our next paper (Richards et al. 1998).
However, it appears likely that clusters I, X,
H, K and probably other subclusters of U were
also present in Europe around the time of the
Last Glacial Maximum. Clusters H, V and K,
which show the most dramatically star-like
phylogenies, seem to have undergone fresh
expansions in the Late Upper Palaeolithic,
probably as post-glacial hunters expanded into
the North European Plain from the southern
refugial zones (cf. Torroni et al. 1998). The whole
of cluster J appears to have been introduced into
Europe within the last 10000 years from the
Near East, but with considerable ancestral
heterogeneity; we have identified at least five
founder sequence haplotypes. A fuller founder
analysis (Richards et al. 1998) indicates that
small subsets of other clusters may have migrated
with J from the Near East during the Neolithic.

A tiny minority of sequences (~1%) from North
African (U6), sub-Saharan African (L1, L2, L3a
and L3b) and Asian (M) clusters appear to have
contributed to the European mitochondrial pool
within the last few thousand years.

The picture that is now emerging can be
summarized as follows. Cluster U5 is likely to
have been introduced, probably from the Near
East, alongside the Aurignacian industry, the
first major European industry to be accompanied
by the remains of anatomically modern humans
(Mellars, 1992). Most of the other clusters were
introduced subsequently during the Early Upper
Palaeolithic, perhaps 20000–25000 years ago,
but populations remained relatively sparse until
the post-glacial warming around 15000 years
ago, when there were dramatic expansions from
south to north, possibly from refugial zones
between the Pyrenees and the Ukraine and
perhaps also from the Mediterranean coastline
(Gamble, 1993). Finally, there was immigration
of early farming communities from the Near
East at the beginning of the Neolithic, from
about 10000 years ago, which eventually stimu-
lated the development of agricultural production
throughout the continent.

We thank Irmgard Pult and Antonio Torroni for
access to unpublished information, Peter Forster and
Antonio Torroni for critical advice, and Nick Barton,
Peter Donnelly, Rosalind Harding, Toomas Kivisild
and Ryk Ward for valuable discussions and suggestions. This
work was supported by the Wellcome Trust and a travel
grant from the DAAD.

REFERENCES

Neolithic transition and the genetics of populations in

ANDERSON, S., BANKIER, A. T., BARRELL, B. G., DE
(1981). Sequence and organization of the human

AVISÉ, J., ARNOLD, J., BALL, R. M., BERMINGHAM, E.,
phylogeography: the molecular bridge between popu-
18, 489–522.

BALLINGER, S. W., SCHURE, T. G., TORRONI, A., GAN, G.


