Am. J. Hum. Genet. 59:185–203, 1996

Paleolithic and Neolithic Lineages in the European Mitochondrial Gene Pool

Martin Richards,[1] Helena Côrte-Real,[1] Peter Forster,[3] Vincent Macaulay,[1] Hilde Wilkinson-Herbots,[5] Andrew Demaine,[6] Surinda Papiha,[7] Robert Hedges,[2] Hans-Jürgen Bandelt,[4] and Bryan Sykes[1]

[1]Department of Cellular Science, Institute of Molecular Medicine, and [2]Research Laboratory for Archaeology, University of Oxford, Oxford; [3]Heinrich-Pette-Institut für Experimentelle Virologie und Immunologie and [4]Mathematisches Seminar, University of Hamburg, Hamburg; [5]Department of Statistical Science, University College London, London; [6]Department of Medicine, University of Plymouth, Plymouth; and [7]Division of Human Genetics, University of Newcastle, Newcastle-upon-Tyne

Summary

Phylogenetic and diversity analysis of the mtDNA control region sequence variation of 821 individuals from Europe and the Middle East distinguishes five major lineage groups with different internal diversities and divergence times. Consideration of the diversities and geographic distribution of these groups within Europe and the Middle East leads to the conclusion that ancestors of the great majority of modern, extant lineages entered Europe during the Upper Paleolithic. A further set of lineages arrived from the Middle East much later, and their age and geographic distribution within Europe correlates well with archaeological evidence for two culturally and geographically distinct Neolithic colonization events that are associated with the spread of agriculture. It follows from this interpretation that the major extant lineages throughout Europe predate the Neolithic expansion and that the spread of agriculture was a substantially indigenous development accompanied by only a relatively minor component of contemporary Middle Eastern agriculturalists. There is no evidence of any surviving Neanderthal lineages among modern Europeans.

Introduction

Current Theories on European Population Structure

Genetic interpretations of European prehistory have hitherto been based largely on the analysis of classical, nuclear-encoded protein polymorphisms. Cavalli-Sforza and others (Ammerman and Cavalli-Sforza 1984; Cavalli-Sforza et al. 1993, 1994; Sokal 1991; Sokal et al. 1991) have assimilated extensive data on allele frequencies and, by principal component analysis or spatial correlation analysis, compiled synthetic gene maps that demonstrate geographic clines. The overall topological similarity between the genetic map produced by the first principal component and an archaeological map of radiocarbon dates tracing the spread of farming from the Near East led to the formulation of the demic diffusion model (Ammerman and Cavalli-Sforza 1984). In this model, there is a slow expansion of people from the Neolithic source population into Europe that is driven by population growth resulting from agricultural surpluses and either displacing or absorbing the less numerous Mesolithic hunter-gatherer populations as it proceeds. The demic diffusion model predicts a dramatic effect on the European gene pool, and, though never precisely quantified, an implied consequence is that the major component of the modern European gene pool is derived from Near-Eastern farmers rather than the indigenous Mesolithic population (Bodmer and Cavalli-Sforza 1976).

Its opposing model, cultural diffusion, proposes that, on the contrary, there was minimal intrusion of peoples from the Near East but that some of the local hunter-gatherer groups in Europe entered the Neolithic either independently or as a result of the diffusion of ideas and the trade of crops (Dennell 1983). An intermediate model, pioneer colonization, assumes some role for migrations from Western Asia to Europe but sees this in terms of selective colonization by fairly small groups (Zvelebil 1986; Willis and Bennett 1994; van Andel and Runnels 1995).

Here we examine which, if any, of these models best explains the observed distribution of mtDNA lineages in Europe. Mitochondria are maternally inherited and nonrecombining, and the effectively haploid genome accumulates mutations faster than nuclear DNA. The most variable region of the mitochondrial genome is the 1,122-bp noncoding control region between bp 16024 and 00576 (numbering after Anderson et al. 1981) within which the variation is concentrated in two regions (I and II) (Stoneking et al. 1991). Control region variation,

Received October 18, 1995; accepted for publication April 11, 1996.

Address for correspondence and reprints: Dr. Bryan C. Sykes, Department of Cellular Science, Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, United Kingdom. E-mail: bsykes@molbiol.ox.ac.uk

0002-9297/96/5901-0026$02.00

as well as restriction site polymorphism throughout the mitochondrial genome, has been used extensively to study the evolution of modern populations (e.g., Cann et al. 1987; Horai et al. 1990; Ward et al. 1991; Vigilant et al. 1991; Torroni et al. 1994; Graven et al. 1995). As a method, it differs from allele-frequency–based surveys of nuclear-encoded variants in several respects: (i) the variation is very extensive and is not scrambled by recombination; (ii) the effective population size is roughly one quarter of that for nuclear variants, which enhances the effect of drift; (iii) being maternally inherited, only female lineages are relevant; and (iv) deducing the phylogenetic relationships within and between haplotype clusters is relatively straightforward and allows divergence and expansion times to be estimated.

Subjects and Methods

Sample Collection and Processing

Samples, for which most of the maternal grandmothers were from rural districts, were collected with informed consent from the following populations (fig. 1): northern Germany (107, including northwest Germany [67], northeast Germany [24], and the north Frisian islands, mainly Föhr [16]); Bavaria (49, from Altötting [35] and Bad Tölz [14]); Denmark (33, from Jutland [16] and the Baltic islands [17]); Wales (92, from North Wales [Clwyd and Gwynedd, 75], mid-Wales [Dyfed and Powys, 11], and south Wales [Glamorgan and Gwent, 6]); Cornwall (69); Finland (29, widely distributed over central and southern Finland); the Basque country (47 from Alava and 14 from Vizcaya); northern Spain (30 Spanish-speaking immigrants from Alava); Portugal (30); and Turkey (22). Three further data sets were incorporated into the analysis: Sardinia (69) and the Middle East (42) (Di Rienzo and Wilson 1991), Switzerland (74) (Pult et al. 1994), and "U.K. Caucasians" (100) (Piercy et al. 1993).

DNA was extracted from hair roots or blood spots by use of the Chelex method (Walsh et al. 1991), or from whole blood by standard methods, and amplified and sequenced across the first hypervariable segment of the control region between positions 16090 and 16365, using the Dynal method for the isolation of single-stranded DNA and the Sequenase version 2.0 kit as described by Sykes et al. (1995). mtDNA was amplified using primers con L1 (5′ TCA AAG CTT ACA CCA GTC TTG TAA ACC) and con H2-B (5′ TGA TTT CAC GGA GGA TGG TG, biotinylated at the 5′ end), and the L-strand was sequenced using con L2 (5′ CAC CAT TAG CAC CCA AAG CT) or con L3 (5′ AAG CAG ATT TGG GTA CCA CC); the H strand was sequenced as described by Sykes et al. (1995). When haplotypes contained an intact polycytosine tract between 16184 and 16193, band blurring occurred beyond the tract distal to the sequencing primer because of length heteroplasmy, and it was necessary to sequence toward the tract from both ends (Bendall and Sykes 1995).

Statistical and Phylogenetic Methods

Simple haplotype diversity (h), an estimate of the probability that two individuals picked at random from a population have different haplotypes but which disregards their phylogenetic relationships, was estimated as per the method of Nei (1987). Mean pairwise differences were computed using the program MacPairwise (Macaulay and Micklem 1995). Minimum divergence times were estimated from these differences by using transversional divergence between the human and chimpanzee consensus sequences (Morin et al. 1994) on the assumption of a transition-to-transversion ratio of 30:1 (Ward et al. 1991). Phylogenetic analysis was performed using reduced median networks, which display the principal character relationships present in the data and resolve likely parallel events while retaining character conflicts in the form of reticulations when ambiguity remains (Bandelt et al. 1995). Lineage groups were defined as clusters of haplotypes deriving from a putative common ancestor and separated from other haplotypes by one or more mutations. Reliability of phylogenies was evaluated qualitatively, since no appropriate quantitative evaluation (such as bootstrapping) is yet available for either networks or trees constructed from intraspecific data (see appendix).

Results

Diversity Analysis

Diversity analysis reveals a European population having considerable haplotype diversity, with most of the haplotypes (77%) occurring only once and consequently very few being shared even between adjacent populations. The haplotype diversity values (Nei 1987) show high diversity within all European populations (range 0.93–0.98; average 0.96), whereas mean pairwise difference estimates show relatively low values (range 2.66–4.08; average 3.62) (table 1). Mean pairwise comparisons and derived genetic distance estimates were also comparatively low between most populations, a finding that left the authors of earlier reports less than sanguine about the prospects for mtDNA analysis in Europe (Pult et al. 1994; Bertranpetit et al. 1995). Relatively few interpopulation diversities were statistically significant in a permutation test (Hudson et al. 1992) (table 2), although it is important to note that two populations descended from a common ancestor will not be discriminated by intermatch pairwise comparisons if the branch lengths are the same, even if the variants accumulated are different. (This explains the lack of discrimination between different European populations when us-

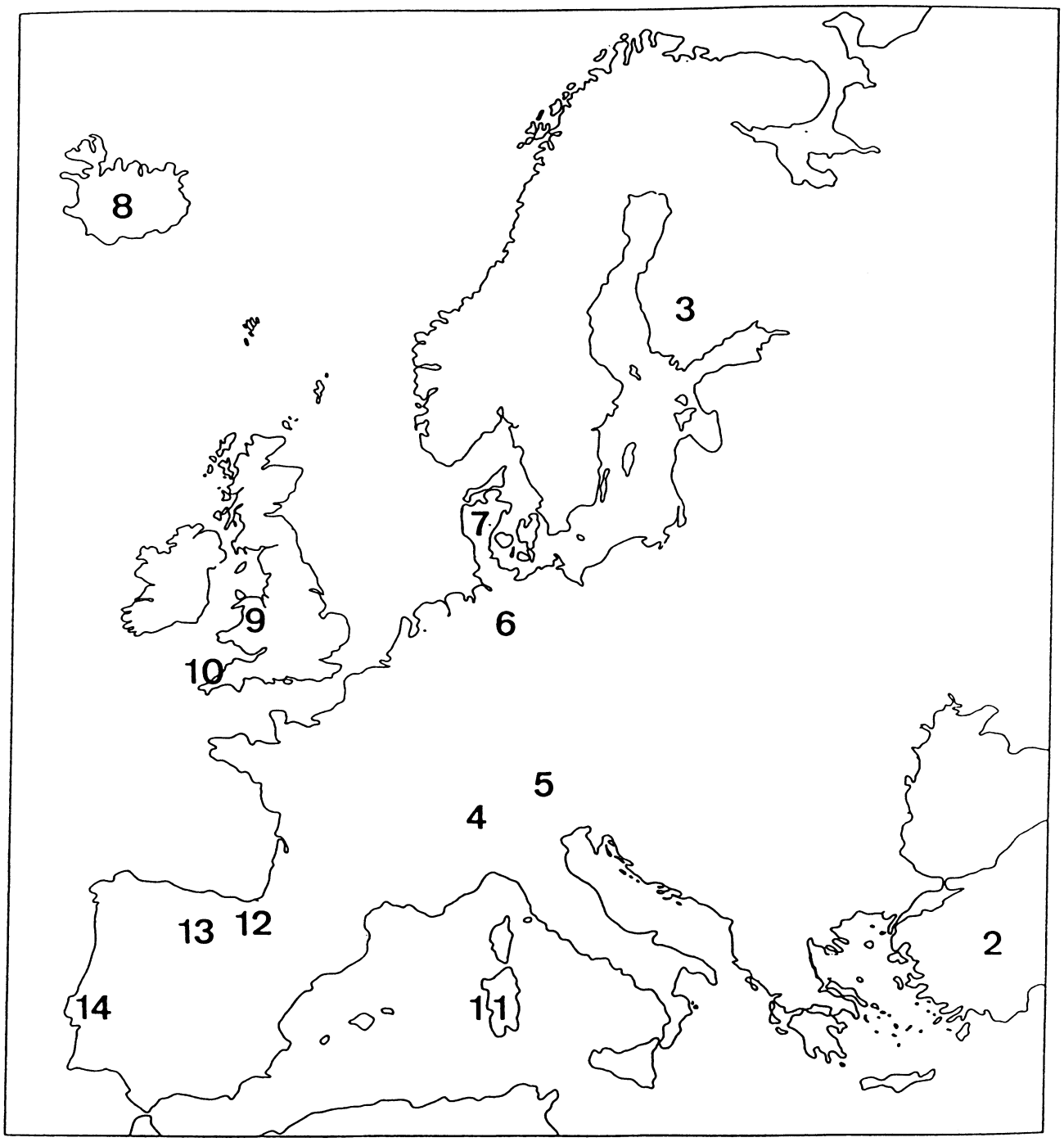


Figure 1 Location of sampling sites. 2 = Western Turkey; 3 = southern and central Finland; 4 = Switzerland (Pult et al. 1994); 5 = Bavaria; 6 = northern Germany; 7 = Denmark; 8 = Iceland; 9 = Wales; 10 = Cornwall; 11 = Sardinia (Di Rienzo and Wilson 1991); 12 = Alava and Vizcaya; 13 = northern Spain; and 14 = Portugal. Sites 1 and 15 (not on the map) refer to the Middle Eastern sample of Di Rienzo and Wilson (1991) and the "U.K. Caucasians" of Piercy et al. (1993), respectively.

ing intermatch comparisons and the permutation test reported by Pult et al. 1994.) There were differences between the Middle East sample and all others in Europe. The only consistently significantly different European population by this test were the Basques, who also had the lowest intrapopulation pairwise diversity. Unimodal distributions of pairwise sequence differences such as is found in all European populations (not shown) have been taken as evidence for demographic expansions (Sherry et al. 1994), although it should be pointed

Table 1

Sample Sizes, Number of Different Haplotypes, and Diversity Estimates for Europe, Turkey, and the Middle East

Location	Sample Size	No. of Haplotypes	Mean Pairwise Difference ± Standard Error[d]	Diversity h ± Standard Error
Middle East[a]	42	37	6.76 ± .48	.993 ± .002
Turkey	22	18	4.09 ± .63	.957 ± .023
Finland	29	21	3.32 ± .46	.956 ± .018
Switzerland[b]	74	42	3.15 ± .27	.960 ± .009
Bavaria	49	35	3.75 ± .35	.981 ± .005
North Germany	107	69	3.87 ± .27	.973 ± .006
Denmark	33	19	3.29 ± .49	.930 ± .021
Iceland	14	12	3.85 ± .74	.962 ± .025
Wales	92	45	3.21 ± .28	.926 ± .014
Cornwall	69	43	3.59 ± .37	.955 ± .012
Sardinia[a]	69	44	4.08 ± .43	.935 ± .017
Basques	61	34	2.66 ± .29	.926 ± .019
Spain	30	26	4.11 ± .40	.984 ± .009
Portugal	30	21	3.82 ± .59	.926 ± .028
All Europe[c]	757	314	3.62 ± <.10	.959 ± <.001

[a] Data of Di Rienzo and Wilson (1991).
[b] Data of Pult et al. (1994).
[c] Includes data of Piercy et al. (1993) and excludes the Middle East and Turkey.
[d] Standard errors of mean pairwise differences estimated from 200 bootstrap replications of the data.

out that without a phylogenetic analysis dating such events would be meaningless (because different clusters may have expanded at different times); furthermore, non-starlike clusters are not amenable to dating in the same way as star phylogenies.

Phylogenetic Analysis

The entire data set is too large and contains too many character conflicts to be displayed in a phylogenetic diagram, as younger haplotypes distort the ancestral structure through recent parallelisms (H. Wilkinson-Herbots, M. Richards, P. Forster, and B. Sykes, unpublished data). Because of this, we focus on the haplotypes represented more than once in Europe since, on the assumption of neutrality in a young expanding population, the more frequent haplotypes tend to be older (Donnelly and Tavaré 1986) and therefore generate a clearer network topology. These are compiled, along with their geographic distribution, in table 3. We have used reduced median networks to explore the phylogenetic relationships between these haplotypes (Bandelt et al. 1995). Missing ancestral haplotypes were reconstructed by this method in the form of empty nodes, which were checked against our worldwide mtDNA database, and most of these could indeed be filled with singly occurring European or Middle Eastern haplotypes. The result is a network (fig. 2) with remarkably few empty nodes, which provides a skeleton with which to examine the general topological features of European mtDNA diversity. Only 93 of the 314 European haplotypes are more than one mutational step away from any haplotype in this network (62 at two, 17 at three, 7 at four, 4 at five, and 3 at six steps away). These fit readily into the groupings we describe below, with the exception of seven sequences that resemble extant African haplotypes (Vigilant et al. 1991; Graven et al. 1995).

Identification of Different Lineage Groups

The most striking and unusual feature of this skeleton network, and also networks drawn for individual European populations, is the starlike phylogeny in which the consensus haplotype is by far the most frequent, at ~20%, in all European populations with the exception of Iceland (for a network analysis of a number of other data sets, see Bandelt et al. 1995). Although most haplotypes in this skeleton network are within two mutational events of the consensus haplotype (which is identical to the Cambridge Reference Sequence (CRS) of Anderson et al. 1981) and the star pattern is characteristic of a population expansion from that root (Castelloe and Templeton 1994), several branches protrude significantly beyond this limit. A number of independent lines of evidence suggest that most of these branches are not, in fact, derived from the European CRS but are the result of a more ancient expansion from a common root and should be grouped separately. The most compelling comes from an examination of the data of 100 "U.K. Caucasians," in which sequences from both regions I

Table 2

Mean Pairwise Comparisons within Populations (on the diagonal) and between Populations (above the diagonal)

	Middle East	Turkey	Finland	Switzerland	Bavaria	Northern Germany	Denmark	Iceland	Wales	Cornwall	Sardinia	Basque	Spain	Portugal
Middle East	**6.76**	5.58*	5.31**	5.21**	5.38**	5.50**	5.24**	5.50*	5.23**	5.31**	5.61**	5.09**	5.59*	5.49**
Turkey	.154	**4.09**	3.70	3.62	3.92	3.99	3.77*	3.92	3.69	3.88	4.10	3.45**	4.08	4.02
Finland	.269	.007	**3.32**	3.23	3.55	3.59	3.41*	3.58	3.33*	3.52*	3.73	3.06*	3.76*	3.67*
Switzerland	.257	.007	.004	**3.15**	3.45	3.51	3.27*	3.49	3.19	3.40	3.65*	2.94*	3.66	3.55*
Bavaria	.122	.005	.011	.004	**3.75**	3.80	3.54	3.77	3.50	3.66	3.92	3.27**	3.93	3.79
Northern Germany	.186	.011	.001	.004	.011	**3.87**	3.64*	3.85	3.58**	3.75	3.98	3.33**	4.01	3.89
Denmark	.212	.087	.111	.053	.024	.059	**3.29**	3.54	3.26	3.46	3.76*	3.02*	3.76	3.62
Iceland	.193	.051	.002	.008	.032	.004	.026	**3.85**	3.48	3.66	3.98	3.22	3.96	3.82
Wales	.247	.039	.065	.015	.023	.043	.012	.052	**3.21**	3.42	3.69**	2.95	3.70	3.55
Cornwall	.135	.037	.059	.029	.015	.020	.018	.055	.017	**3.59**	3.85	3.18**	3.86	3.71
Sardinia	.188	.012	.030	.032	.004	.007	.074	.011	.050	.015	**4.08**	3.42**	4.11	3.96
Basque	.380	.080	.066	.034	.063	.061	.048	.028	.021	.054	.052	**2.66**	3.45**	3.30*
Spain	.152	.015	.050	.035	.004	.017	.063	.015	.043	.005	.014	.067	**4.11**	3.98
Portugal	.201	.063	.099	.070	.004	.045	.066	.016	.031	.007	.012	.061	.019	**3.82**

NOTE.—Significance of the interpopulation diversities tested by the permutation test using the statistic K_s (*$.01 < P \leq .05$; **$P < .01$). Genetic distances (D) between populations are below the diagonal. D was calculated using the formula $D = D_{12} - (D_1 + D_2)/2$, where D_{12} is the pairwise mean between populations 1 and 2 and D_1 (D_2) is the pairwise mean within population 1 (2).

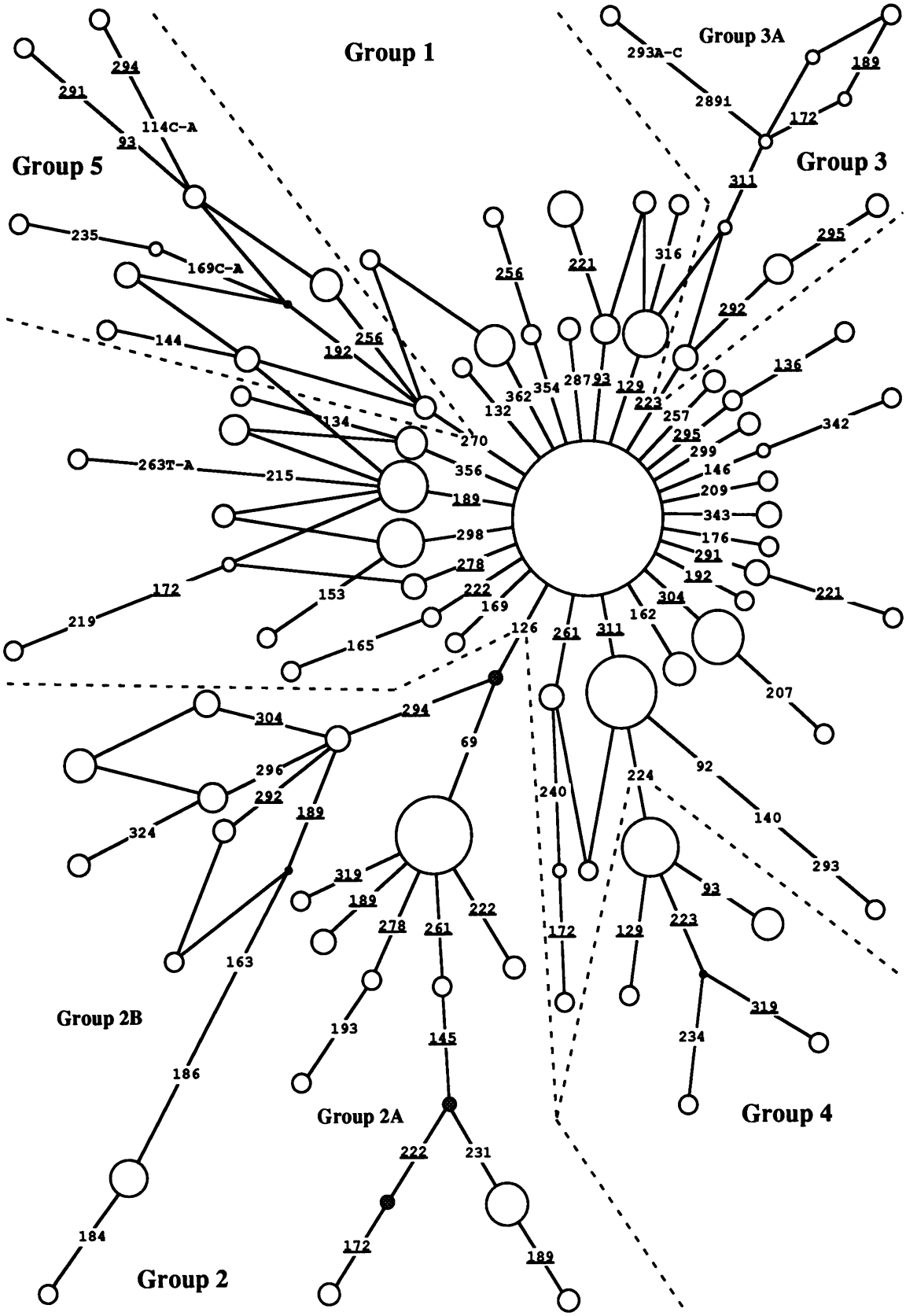
Table 3

Variants in 82 mtDNA Sequence Haplotypes Occurring More Than Once in Europe and Their Geographical Distribution

Group and Haplotype	Variants			Geographical Distribution																
	00011111111111111111111111222222222222222222222	2	222222222333333333																	
	69912233344445666677888990011222233334556667781	8	999999990111244556																	
	92346924604563235926469237959123414506713087	9 i	123456894169423462	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	All	Europe
1:																				
1	CTTCTGACTTTGAGAAACTCCCTCCATAACCCTTCAACCCTCCC	...	CCACCCTATTAGTTACTT	.	5	6	11	5	14	8	.	23	13	17	16	4	8	14	144	139
2	?.C...	...	....	.	.	.	.	.	.	.	.	.	2	1	2	.	.	.	5	5
3	?....A......................................	...	....	.	.	.	.	.	1	2	.	.	1	1	4	.	.	3	12	12
4	?.....G.....................................	...	....	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	2
5	?.............G.............................	...	....	.	.	2	.	.	1	.	1	.	.	.	.	.	.	1	5	5
6	?................T..........................	...	....	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	2	2
7	?..................T........................	...	....	.	.	.	.	.	.	.	.	1	.	.	1	.	.	.	2	2
8	?.....................C.....................	...	....	.	.	1	1	1	8	.	.	.	2	1	.	.	.	2	16	16
9	?......................T....................	...	....	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	2	2
10	?.........................C.................	...	....	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	2	2
11	?.............................T.............	...	....	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	2	2
12	?.....................................T.....	...	....	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	3	3
13	?......................................T....	...	....	.	.	.	2	1	1	.	.	.	.	.	.	.	.	.	4	4
14	?...T.	...	....	.	.	.	.	.	.	.	.	3	.	.	1	.	.	.	4	4
15	?..T	...	....	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	3	3
16	?...	...	T.................	.	.	.	.	.	.	.	.	.	1	.	2	.	.	1	4	4
17	?...	...	....T.............	.	.	.	.	1	.	1	.	.	.	.	.	.	.	.	2	2
18	?...	...	......C...........	.	.	.	3	1	1	1	.	2	1	.	1	1	1	2	14	14
19	?...	...	.......G..........	.	.	.	.	.	.	.	1	.	.	.	.	.	.	1	2	2
20	?...	...	........C.........	.	1	.	2	1	2	.	1	7	1	1	1	.	.	2	19	18
21	?...	...	.........C........	2	.	1	9	3	3	2	.	2	2	3	.	2	.	.	29	27
22	?...	...	..............G...	.	.	.	.	2	2	.	.	.	.	.	.	.	.	.	4	4
23	?...	...	...............T..	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	2	2
24	?...	...	................C.	.	1	.	.	1	.	.	1	.	2	.	.	.	1	1	7	6
25	?...	...	C	.	1	.	.	.	.	1	.	.	2	.	3	.	.	3	10	9
26	?.C..A......................................	...	....	.	.	.	1	.	.	.	.	.	.	1	.	.	.	1	3	3
27	?.C..........................T..............	...	....	.	.	.	2	1	1	.	.	3	1	.	.	.	.	.	8	8
28	?....A......................................	...	..........G.......	.	.	.	.	.	.	1	.	.	.	.	.	.	.	1	2	2
29	?......T....................................	...	................C.	.	.	.	1	1	.	.	.	.	.	.	.	.	.	.	2	2
30	?.......C...................................	...	....T.............	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	2	2
31	?...........G...............................	...	.............C....	.	.	.	1	.	.	.	.	.	.	.	1	.	.	.	2	2
32	?............A..............................	...	......C...........	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	2	2
33	?...............G.............T.............	...	....	.	.	.	.	.	2	.	.	.	.	.	.	.	.	.	2	2
34	?.....................C.....................	...	......C...........	.	.	.	1	.	.	.	.	.	.	1	1	.	.	.	3	3
35	?.....................C.....................	...	................C.	.	.	1	2	.	1	.	.	.	1	.	.	.	.	.	5	5
36	?........................G..................	...	........C.........	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	2	2
37	?............................T..............	...	T.................	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	2	2
38	?....................................T......	...	...............T..	.	.	.	.	.	1	1	.	.	.	.	.	.	.	.	2	2
39	?......................................T....	...	.........C........	.	.	.	.	.	.	.	.	.	.	.	1	.	.	1	2	2
40	?..T..	...	C	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	2
41	?.................C.................G..T....	...	....	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	2
42	?.....................C....G............A...	...	....	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	2	2
43	?C.......C..................................	...	..G......C........	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	2
44	?.................C...C.....G.............T.	...	....	.	.	.	.	.	.	.	.	.	.	.	.	.	2	.	2	2

2A:																		
45	T...C..	.	.	2	2	3	3	3	3	7	6	.	.	1	.	7	37	37
46	T...C.................C..	.	1	.	1	.	.	.	.	.	1	1	.	.	.	.	4	3
47	T...C.........................T..................................	.	.	.	.	2	1	.	.	.	.	.	.	.	.	.	3	3
48	T...C..................................T.........................	.	.	1	1	.	.	.	.	.	.	.	.	.	.	.	2	2
49	T...C.....................................T......................	.	.	.	.	.	.	.	.	.	1	.	.	.	1	.	2	2
50	T...C..A.......	.	.	.	.	.	.	.	.	1	1	.	.	.	.	.	2	2
51	T...C...................T.................T......................	.	1	.	.	.	.	.	1	.	.	1	.	.	.	.	3	2
2A–C:																		
52	T...C......A.....................C.....T.........................	.	.	.	.	2	3	3	.	.	.	.	.	.	.	3	11	11
53	T...C......A..........C..........C.....T.........................	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	3	3
2A–W 54	T...C......A......C...........T........T.........................	.	.	.	.	.	.	.	.	1	.	.	.	.	.	2	3	3
2B:																		
55	....C..T...............	.	.	.	.	1	.	.	.	.	2	1	.	.	.	.	4	4
56	....C..T.T...............	.	.	.	.	.	1	.	.	.	.	2	.	.	.	.	3	3
57	....C..T.T.............	1	.	.	.	.	.	.	.	.	1	.	2	.	.	2	6	5
58	....C..T....C..........	.	.	.	1	.	1	.	.	1	.	.	.	.	.	1	4	4
59	....C.................C........................T.T...............	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	2	2
60	....C..T.T..C..........	.	.	.	1	.	.	2	.	.	.	.	.	.	1	1	5	5
61	....C..T.T......C......	.	.	.	.	.	1	1	.	.	.	.	1	.	.	.	3	3
62	....C..........G.....TC..........................T...............	2	.	.	.	2	2	.	.	1	1	.	.	.	1	2	11	9
63	....C..........G....TTC..........................T...............	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	2	2
3:																		
64	?..............................T.................................	.	.	.	1	.	1	.	.	.	.	.	.	.	1	1	4	4
65	?..............................T...............T.................	.	.	1	.	2	.	.	.	.	.	1	.	1	.	.	5	5
66	?..............................T...............T..T..............	.	1	.	1	.	.	1	.	.	.	.	.	.	.	.	3	2
3A:																		
67	?....A............C...C........T.......................C.........	.	.	.	.	.	.	.	.	.	2	.	.	.	.	.	2	2
68	?....A.........................T.............C..C......C.........	.	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	2
4:																		
69	?...............................C......................C.........	.	.	1	1	1	7	.	.	2	1	1	2	.	2	2	20	20
70	?.C.............................C......................C.........	.	1	.	1	.	1	.	.	2	.	1	.	1	.	.	7	6
71	?....A..........................C......................C.........	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	2	2
72	?..............................TC.T....................C.........	.	.	.	.	.	.	.	.	1	.	.	.	.	.	1	2	2
73	?..............................TC......................C.A.......	.	.	.	2	.	.	.	.	.	.	.	.	.	.	.	2	2
5:																		
74	?..T.......................	.	.	.	.	.	.	.	.	.	.	.	1	.	.	2	3	3
75	?.....................C..................T.......................	.	1	.	1	1	1	.	.	.	.	.	.	.	.	.	4	3
76	?....................................T...T.......................	.	1	.	2	.	2	.	.	1	.	.	.	1	.	.	7	6
77	?.........C...........C..................T.......................	.	.	2	.	.	.	.	.	.	.	.	.	.	.	.	2	2
78	?.....................CT.................T.......................	.	.	1	.	2	1	.	.	.	.	.	.	.	.	.	4	4
79	?......................T.............T...T.......................	.	.	.	1	.	1	1	.	.	.	.	.	.	.	.	3	3
80	?................A.....T...........G.....T.......................	.	.	.	.	.	.	.	.	.	.	3	.	.	.	.	3	3
81	?.C....................T.............T...T....T..................	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	2	2
82	?..A...................T.............T...T.......T...............	.	.	.	.	.	1	.	.	.	1	.	.	.	.	.	2	2
Total		5	14	21	61	34	68	29	9	70	50	44	46	11	18	59	539	520

Variable positions are denoted by the numbering of Anderson et al. (1981) less 16,000. Dots indicate matching with the reference sequence (haplotype 1). 289i indicates a single cytosine insertion in the region 16,290–292. Geographical distribution as in figure 1. Geographical counts are numbers of individuals per haplotype in each location; a dot indicates 0. All = total Europe, Turkey and Middle East.


Group 1
Group 3A
Group 3
Group 5
Group 2B
Group 2A
Group 4
Group 2
294
291
114C-A
93
235
169C-A
144
192
256
134
263T-A
215
219
172
153
165
270
356
189
298
278
222
169
126
132
362
354
287
93
129
223
257
295
299
146
209
343
176
291
192
304
162
207
221
342
136
292
316
221
256
311
2891
293A-C
189
172
295
261
311
240
172
69
304
294
296
292
189
324
163
186
184
319
189
278
261
222
193
145
222
231
172
189
92
224
140
293
93
129
223
319
234

and II of the mitochondrial control region are presented (Piercy et al. 1993). An apparently stable A-G transitional variant at bp 00073 in region II segregates the region I haplotypes roughly in half between the CRS together with most of its close neighbors (A at position 00073) and four longer branches (H. Wilkinson-Herbots, M. Richards, P. Forster, and B. Sykes, unpublished data). These all have G at bp 00073, and are defined in region I by transitions relative to the CRS at 126, 223, 224–311, and 270, respectively (here, and henceforth, omitting the 16000 prefix, for simplicity). Region II sequencing from >50 of our sample confirmed the segregation, and we used this correlation to divide our European data set into the five lineage groups in figure 2.

Group 1 is the CRS and its derivatives, defined by default after subtracting groups 2 to 5. Group 2 is characterized by a transition at 126. Several of the ancestral haplotypes, including the deepest node within this branch, are not found in Europe but in the Middle East (Di Rienzo and Wilson 1991) where the branch as a whole occurs at a much higher frequency (almost 50%). Group 2 splits into two subgroups: one with a transition at 069 and a second with a transition at 294. These two subgroups both occur in the Europeans of the studies by Di Rienzo and Wilson (1991) and Piercy et al. (1993). However, only in the Middle East do we find the ancestral state of both of these branches: a transition at 126 unaccompanied by transitions at either 069 or 294. We consequently sequenced our group 2 individuals to 16069 and added the character to figure 2. We refer to the former branch (069-126) as group 2A, and the latter branch (126-294) as group 2B. Group 3A, characterized by the transitional motif 129-223, is also found in the Middle East and corresponds to haplogroup I of Torroni et al. (1994, p. 769), which they consider to be "one of the most ancient Caucasian-specific lineages." It is part of a much broader group 3 which differs from the CRS by the transition at 223 (and at 00073) and which on the available data appears to encompass most of the extant mtDNA variation in Asia and Africa (e.g., Horai et al. 1991; Vigilant et al. 1991; Graven et al. 1995; Sykes et al. 1995).

The two other lineage groups have not yet been described outside Europe. Members of lineage group 4 share the transitional motif 224-311. The root haplotype is the fourth-most-common in Europe and was recovered from the 5,300-year-old Tyrolean Iceman (Handt et al. 1994). At first sight, it appears to be derived from haplotype 21 (table 3) defined by a transition from the CRS at position 311. However, ancestral haplotypes in a young, expanding population are expected to be not only frequent but to have numerous radiating branches (Castelloe and Templeton 1994). In fact, haplotype 21 has fewer such branches than does its seeming derivative with the transition at position 224. Position 00073 confirms that this lineage does not derive from the European CRS. In analogous fashion, the longest branches in the skeleton network do not converge on the CRS but rather on a haplotype defined by a transition at position 270 (and at 00073). This cluster also includes a central node, defined by transitions at positions 270 and 192, which does not appear to exist any longer in any population sample. We refer to this cluster as group 5.

In conclusion, only the shorter branches displayed in the network are plausible descendants of the European CRS. (Haplotype 5, which has a G at position 00073, is a possible exception; it may either have originated earlier than the 00073 mutation or have undergone reversion, but it is not considered further here.) The five groups converge on a more ancient haplotype with the sequence of the CRS in region I but distinguished at bp 00073 in region II, which is identical to the Middle East consensus, suggesting an eastern origin for European mtDNA, and comparisons with global data suggest that this ancient haplotype itself is likely to be an offshoot of group 3.

The existence of these groups receives independent support from a combined RFLP and sequencing approach on 48 Italian lineages (A. Torroni, K. Huponen, P. Francalacci, M. Petrozzi, L. Morelli, R. Scozzari, D. Obinu, M. L. Savountaus, and D. C. Wallace, unpublished data). Group 1 is equivalent to haplogroup H, group 2A to haplogroup J, group 3A to haplogroup I, and group 4 to haplogroup K of Torroni et al. (1994).

Figure 2 European skeleton network, a reduced median network constructed from the 82 haplotypes occurring at frequency ≥2 Europe, using our data and those of Di Rienzo and Wilson (1991), Piercy et al. (1993), and Pult et al. (1994). Unfilled circles are haplotypes; the circle area is proportional to the haplotype frequency. The smallest circles are singly occurring haplotypes from a global database, which fill nodes in the skeleton. The eight small unfilled circles are European, whereas the three shaded circles represent singly occurring haplotypes from the Middle East (Di Rienzo and Wilson 1991). Nodes remaining empty after this search are represented by black points. The numbers on the branches denote mutations and correspond to base positions, less 16000, in the CRS (Anderson et al. 1981) at which mutation events have occurred. Transversions are defined, but transitions do not require further definition. The single insertion of a cytosine in the region 16290-292 is denoted 289i. Positions at which parallel mutations have been uncovered by network reduction are underlined. Reticulations in the network indicate parallel events, which cannot be resolved by this method. Base notation is given only once in any reticulation but is the same in parallel connections with it. In view of the partial information on position 00073, we have further reduced one cube by favoring one-step connections, thus postulating a parallel event at position 16311. Dotted lines delineate the five lineage groups identified. Only one haplotype (74) may be associated with more than one group (namely, groups 1 and 5) as a result of recurrent mutation (H. Wilkinson-Herbots, M. Richards, P. Forster, and B. Sykes, unpublished data), in this case at position 16270.

In addition, group 2B matches haplogroup T, group 5 approximates haplogroup U, and group 3 as a whole encompasses haplogroups I, W, and X (A. Torroni, K. Huponen, P. Francalacci, M. Petrozzi, L. Morelli, R. Scozzari, D. Obinu, M. L. Savountaus, and D. C. Wallace, unpublished data). It should be noted that a tree-based bootstrap analysis of our data would not resolve these groups clearly and, in particular, would be unable to distinguish the earlier from the later CRS expansion.

Geographic Patterns

Next we examined whether there is any geographic pattern in the frequency distribution of the lineage groups within Europe (table 4). Group 1 is by far the most frequent in all European populations, accounting for nearly two-thirds of the total, and with the highest frequency among the Basques (corroborated by the recent data of Bertranpetit et al. 1995). Group 4 is also widely distributed but at a much lower frequency, averaging 7%. Group 5 lineages are also uncommon but widespread in Europe, at around 7%, but are unusually common in Finland, where they reach 21%, possibly suggesting Saami influence (Sajantila et al. 1995). Like group 4, they are absent in other parts of the world sampled to date. Group 3 lineages are even less frequent in Europe, and 3A lineages are confined within our sample mainly to Britain. They are also found in the Middle East, and most 3A members share transitions at 129-223-311 with Papuan lineages (Sykes et al. 1995).

The most intriguing distribution is found within group 2. While group 2B is widespread and consistent at ~8%, 2A lineages vary widely in frequency within Europe, where the range is from 2% (Basques) to 22% (Cornwall). This group also shows a very interesting geographic substructure. Western Europeans and central/north-central Europeans share some deep nodes (namely, haplotypes 45, 46, and 48), but others are geographically specific. A western European cluster (2A-W) found in Wales and Cornwall comprises haplotype 54 and its derivatives, while a central European cluster (2A-C) found in Denmark, Germany, Switzerland, and Finland comprises haplotype 52 (three steps from haplotype 54) and its derivatives. Both clusters are found among the "U.K. Caucasians" of Piercy et al. (1993). It is striking that, in the Middle East, but not elsewhere in the world, we find the two missing ancestral haplotypes that link the western and central European clusters. Finally, we also find the ancestral haplotype of the entire group 2 branch (with a single transition at position 126) in the Middle East but in none of our European samples or elsewhere. This pattern strongly suggests that group 2A lineages originated in the Middle East and that several different lineages migrated into Europe, splitting into the western and central European clusters but having little impact on the Iberian peninsula, especially the Basque country.

Intragroup Diversities

The pairwise sequence diversity that has accumulated within groups can be used, in conjunction with an appropriate mutation rate, to give an estimate of their minimum divergence times from a common ancestor (table 5). With this treatment, group 3 emerges as the oldest of the European lineage clusters, with a divergence time of 50,500 years, although whether this divergence occurred entirely in Europe is open to question, since this group includes several singly occurring outliers resembling modern African and Asian haplotypes (in the data of, e.g., Horai et al. 1991; Vigilant et al. 1991; Graven et al. 1995; Sykes et al. 1995). However, group 3A, in agreement with Torroni et al. (1994), is among the oldest apparently Caucasian-specific lineage groups in Europe, with a *minimum* age of ~34,000 years. Groups 2B and 5 are similar in age to group 3A, dating to ~35,500 and 36,500 years, respectively. Groups 1 and 4 are both considerably younger, at 23,500 and 17,500 years, respectively, and group 2A also dates to 23,500 years.

These minimum ages would only correlate with arrival or expansion times in Europe if there were little or no diversity within the colonizing groups. To evaluate this, we examined the European and Middle Eastern populations for shared haplotypes and nodes since, in most cases, a haplotype evolves only once, and each match then represents an instance of migration. Among group 1 lineages there was a conspicuous absence of the root (CRS) haplotype in the Middle East sample, although a reduced median network for the Middle East (not shown) reveals that, even though not in the sample, it is present as an intermediate node separating groups 2 and 3. This suggests that, although the source of group 1 had its origins in the Middle East, the bulk of the current diversity arose in Europe and the correlation between age and arrival/expansion time is not unduly distorted. Groups 4 and 5 have not yet been found outside Europe, so the preexisting diversity in the source population cannot properly be estimated, although, if they too had Middle Eastern origins, the ancestral haplotypes are now very rare or extinct, so it is likely that the diversity seen within Europe has, for the most part, developed locally. Group 3 is well represented in the Middle East, and its divergence time in Europe is therefore subject to some uncertainty.

It is striking that group 2 as a whole dates to nearly 60,000 years in the Middle East, where we find the ancestral node of the branch with many derivatives in addition to the 2A and 2B clusters. Group 2B occurs in the Middle East at low frequency, where its diversity is similar to that in Europe (dating to 42,000, as against

Table 4

Geographical Distribution of Lineage Groups

Group	No. of Types	Middle East[a]	Turkey	Finland	Switzerland[b]	Bavaria	Northern Germany	Denmark	Iceland	Wales	Cornwall	Sardinia[a]	Basque	Spain	Portugal	U.K. Caucasian[c]	All Europe
1	183	12 (.29)	12 (.55)	16 (.55)	51 (.69)	27 (.55)	65 (.61)	20 (.61)	8 (.57)	61 (.66)	39 (.57)	41 (.59)	47 (.77)	20 (.67)	18 (.60)	55 (.55)	468 (.62)
2	75	20 (.48)	3 (.14)	4 (.14)	10 (.14)	13 (.27)	19 (.18)	9 (.27)	4 (.29)	18 (.20)	21 (.30)	13 (.19)	5 (.08)	6 (.20)	7 (.23)	20 (.20)	149 (.20)
2A	36	7 (.17)	3 (.14)	4 (.14)	8 (.11)	7 (.14)	10 (.09)	6 (.18)	4 (.29)	14 (.15)	15 (.22)	4 (.06)	1 (.02)	2 (.07)	3 (.10)	12 (.12)	90 (.12)
2B	31	5 (.12)	0 (.00)	0 (.00)	2 (.03)	6 (.12)	9 (.08)	3 (.09)	0 (.00)	4 (.04)	6 (.09)	9 (.13)	4 (.07)	4 (.13)	4 (.13)	8 (.08)	59 (.08)
3	38	10 (.24)	4 (.18)	1 (.03)	3 (.04)	4 (.08)	5 (.05)	1 (.03)	0 (.00)	4 (.04)	4 (.06)	6 (.09)	0 (.00)	2 (.07)	2 (.07)	7 (.07)	39 (.05)
4	24	0 (.00)	1 (.05)	2 (.07)	4 (.05)	1 (.02)	10 (.09)	1 (.03)	1 (.07)	7 (.08)	2 (.03)	4 (.06)	5 (.08)	1 (.03)	3 (.10)	10 (.10)	51 (.07)
5	31	0 (.00)	2 (.09)	6 (.21)	6 (.08)	4 (.08)	8 (.07)	2 (.06)	1 (.07)	2 (.02)	3 (.04)	5 (.07)	4 (.07)	1 (.03)	0 (.00)	8 (.08)	50 (.07)
Total	351	42	22	29	74	49	107	33	14	92	69	69	61	30	30	100	757

NOTE.—Columns under location show number of individuals classified within each group with their respective frequencies in parentheses.

[a] Data of Di Rienzo and Wilson (1991).

[b] Data of Pult et al. (1994).

[c] Data of Piercy et al. (1993).

Table 5

Mean Pairwise Differences (mpd), Standard Errors (SE), and Estimated Minimum Divergence Times (years) for Lineage Groups in Europe and the Middle East

	Europe[a]		Middle East[b]	
Group	mpd ± SE	Divergence Time[c]	mpd ± SE	Divergence Time[c]
1	2.25 ± .10	23,500	5.01 ± .66	52,500
2	3.46 ± .20	36,500	5.47 ± .51	57,500
2A	2.22 ± .23	23,500	2.38 ± .81	25,000
2A–C	.57 ± .20	6,000	...	...
2A–W	1.20 ± .57	12,500	...	...
2B	3.40 ± .21	35,500	4.00 ± 1.87	42,000
3	4.81 ± .49	50,500	6.20 ± 1.22	65,000
3A	3.11 ± .45	34,000	2.00	21,000
4	1.69 ± .22	17,500	...	...
5	3.46 ± .24	36,500	...	...

[a] Data of authors, Di Rienzo and Wilson (1991), Piercy et al. (1993), and Pult et al. (1994).

[b] Data of Di Rienzo and Wilson (1991).

[c] Standard errors of mean pairwise differences estimated from 200 replications of the data.

[d] Using a divergence rate of 1/10,500 years and given to the nearest 500 years.

35,500, years). Group 2A is younger but also has similar diversity in the Middle East and Europe (dating to ~25,000 years). However, since we find both ancestral haplotypes separating the western and central European 2A clusters only in the Middle East, and never in Europe, implying at least two distinct founding lineages, we estimated the ages of the western and central European clusters separately (table 5). This results in minimum age estimates of ~12,500 years in western Europe and only 6,000 years in central/northern Europe.

Discussion

Archaeological Context for mtDNA Haplotypes

The beginning of the Upper Paleolithic 40,000–50,000 years ago marks the first appearance of anatomically and behaviorally modern humans in Europe, prior to which it had been occupied by the Neanderthals for ≥250,000 years (Stringer 1993). Although the estimation of divergence times is by no means uncontroversial, it is extremely unlikely that the majority of contemporary European lineages are derived from Neanderthals, since it would require at least an eightfold overestimate of the mutation rate to accommodate a minimum divergence time of 300,000 years from the mean pairwise difference value for all European haplotypes. Furthermore, no individual lineages in our extensive data set are sufficiently diverged to be realistically attributed to Neanderthal ancestors. We conclude (in agreement with Torroni et al. 1994) that there are no surviving Neanderthal lineages among the sample, supporting the view that Neanderthals became extinct (Mellars 1992) and, though coexisting in Europe with anatomically modern humans, did not interbreed to any significant extent.

The approximate coincidence of divergence times for groups 2B, 3A, and 5 (at ~35,000 years) suggests to us that these most diverse lineage groups were brought into Europe during the early Upper Paleolithic colonization by anatomically modern humans. However, these three lineage groups account for <20% of the modern European sample. The major groups, 1 and 4, which together comprise nearly 70% of European haplotypes, are considerably younger. Both show starlike patterns of expansion from their respective root haplotypes at some time since 25,000 years ago. These expansions may have been linked to climatic improvements following the end of the last Ice Age (Gamble 1986). However, they appear to have been confined to only these two groups, whereas a general improvement in conditions might be expected to benefit all lineage groups extant at the time. An interpretation of the genetic evidence would favor new colonizations and/or selective expansions by the ancestors of groups 1 and 4 after ~25,000 years ago, perhaps accompanying postglacial climatic change.

The most important influence of the Middle East on Europe since the Upper Paleolithic was the spread of agriculture 10,000–6,000 years ago. It has been suggested on the basis of archaeological evidence that there were two Neolithic colonization routes from the Middle East through Europe, one through central to northern Europe and another around the west Mediterranean littoral and possibly up the Atlantic coast of France to Britain (Renfrew 1987). These two routes are linked to the development of the archaeologically defined ceramic cultures, the LBK (Linienbandkeramik) of Central Europe and the Impressed Ware and Cardial Ware complexes of the Mediterranean coastline and Atlantic west. The Neolithic transition reached the Mediterranean earlier than Central Europe, and there was initially minimal Neolithic penetration of the Iberian peninsula (Whittle 1985). These archaeological findings correlate remarkably well with the geographic origin, ages, and distributions of the western and the central European group 2A clusters. The ancestral haplotypes of both groups trace back to Turkey and the Middle East, the western cluster is older than the central cluster, and marker haplotypes for these two groups do not overlap geographically in their respective regions in spite of being very widespread. Different regions of Europe have different frequencies of group 2A: ~12% in north and north-central Europe, 18% in northwestern Europe, 6% in Sardinia, and only ~5% in the Iberian peninsula. The lack of evidence for a clear northwest-southeast gradient may imply that, as archaeologists have suggested, Neolithic colonization

was much more heterogeneous than the demic diffusion model implies (Zvelebil 1986; Willis and Bennett 1994). It should be possible to test these hypotheses by the analysis of ancient DNA extracted from skeletal remains at relevant archaeological sites.

Our interpretation of the mtDNA data is that the majority of modern Europeans are descended from the settlement of Europe by anatomically modern humans during the Upper Paleolithic with expansion of selected groups at around the end of the last Ice Age. We see evidence of later colonization from the Middle East, which appears to coincide with the spread of agriculture. Its distribution is well defined, but the overall demographic influence on modern Europeans is relatively small. Of the three models for the spread of agriculture outlined earlier, our interpretation favors the pioneer colonization model, whereby there was selective penetration by fairly small groups of Middle Eastern agriculturalists of a Europe numerically dominated by the descendants of the original Paleolithic settlements. The ensuing conversion of this population from a hunter-gatherer-fishing economy to one based on agriculture would then have been achieved by technology transfer rather than large-scale population replacement.

Comparison with Other Current Models

Although arguing for a demic component in the transfer of the Neolithic to Europe, our interpretation differs from the conclusions of previous investigators (Ammerman and Cavalli-Sforza 1984; Cavalli-Sforza et al. 1993, 1994; Sokal 1991; Sokal et al. 1991), which have emphasized, though not precisely quantified, the genetic contribution of the Neolithic immigrants. How can this disagreement be accounted for when both models rely on genetic evidence?

Dating branching events is controversial, and a faster rate for the molecular clock (such as that calculated from coalescent theory by Lundstrom et al. 1993) might suggest that the peak of group 1 branching events from the CRS detected in our analysis of mtDNA using pairwise comparisons represent Neolithic rather than Paleolithic incursions.

Our analysis argues against such an interpretation. Every European population shows the same pattern of divergence from the CRS, and the CRS frequency is similar throughout the continent. Indeed, it is highest among the Basques, who, in addition to being the only Western Europeans speaking a non-Indo-European language, have long been recognized as the most genetically unusual population as well as one in which the demographic influence of agriculturalists is thought to have been the least. If, as is thought, the Basques are descended from a pre-Neolithic population, then one would have expected a low frequency of the CRS and its recent derivatives if the latter were to represent Neolithic intrusion. In fact, the reverse is the case within the Basques, both in our data and that of Bertranpetit et al. (1995): they have the highest group 1 frequency in Europe and, because of this, the lowest overall pairwise sequence diversity. It may be that the Basque population appears as a distinctive outlier in most classical genetic analyses not because it is the sole relict of a pre-Neolithic population but rather because of a long period of isolation and genetic drift that has accentuated allele frequency differences at some loci compared with other European populations. It is interesting that the small sample of Finns, who also speak a non-Indo-European language, do not resemble the Basque pattern but are distinguished by an unusually high frequency of group 5, which may be attributable to Saami influence, since group 5 includes a haplotype motif present at high frequency in the Saami (Sajantila et al. 1995), who also speak a Finno-Ugric language.

If the first principal component truly reflects the spread of agriculture our data imply that, in the absence of some form of sexual selection (Bodmer 1993), a small number of incoming lineages had a dramatic effect on the genetic landscape of Europe. However, other interpretations are possible. A recent investigation of alternative models examining the southeast-northwest nuclear-encoded allele frequency cline, which forms the basis of the demic diffusion model, suggests there is little to differentiate between an expansion from the Middle East into a Europe with a low-density hunter-gatherer population that was progressively absorbed and an expansion into an essentially "empty" Europe (Barbujani et al. 1995). Therefore, it seems theoretically possible that the cline was established not by Neolithic immigration but by earlier Upper Paleolithic colonizations that, our analysis suggests, brought with them the ancestors of most modern Europeans.

Overall, this work shows that mtDNA control region haplotype analysis, in conjunction with appropriate phylogenetic treatment, is not only capable of revealing population structure and origins on the widest intercontinental scale but also has the resolving power to disentangle even relatively homogeneous populations such as that of Europe.

Appendix

Phylogenetic Reconstruction

We have used reduced median network construction (Bandelt et al. 1995) as the basis of our phylogenetic analysis, in preference to traditional tree-construction methods, such as maximum parsimony (MP) or neighbor-joining (NJ) in conjunction with bootstrapping. These methods, which were developed for interspecific data, are not appropriate for our intraspecific data, in

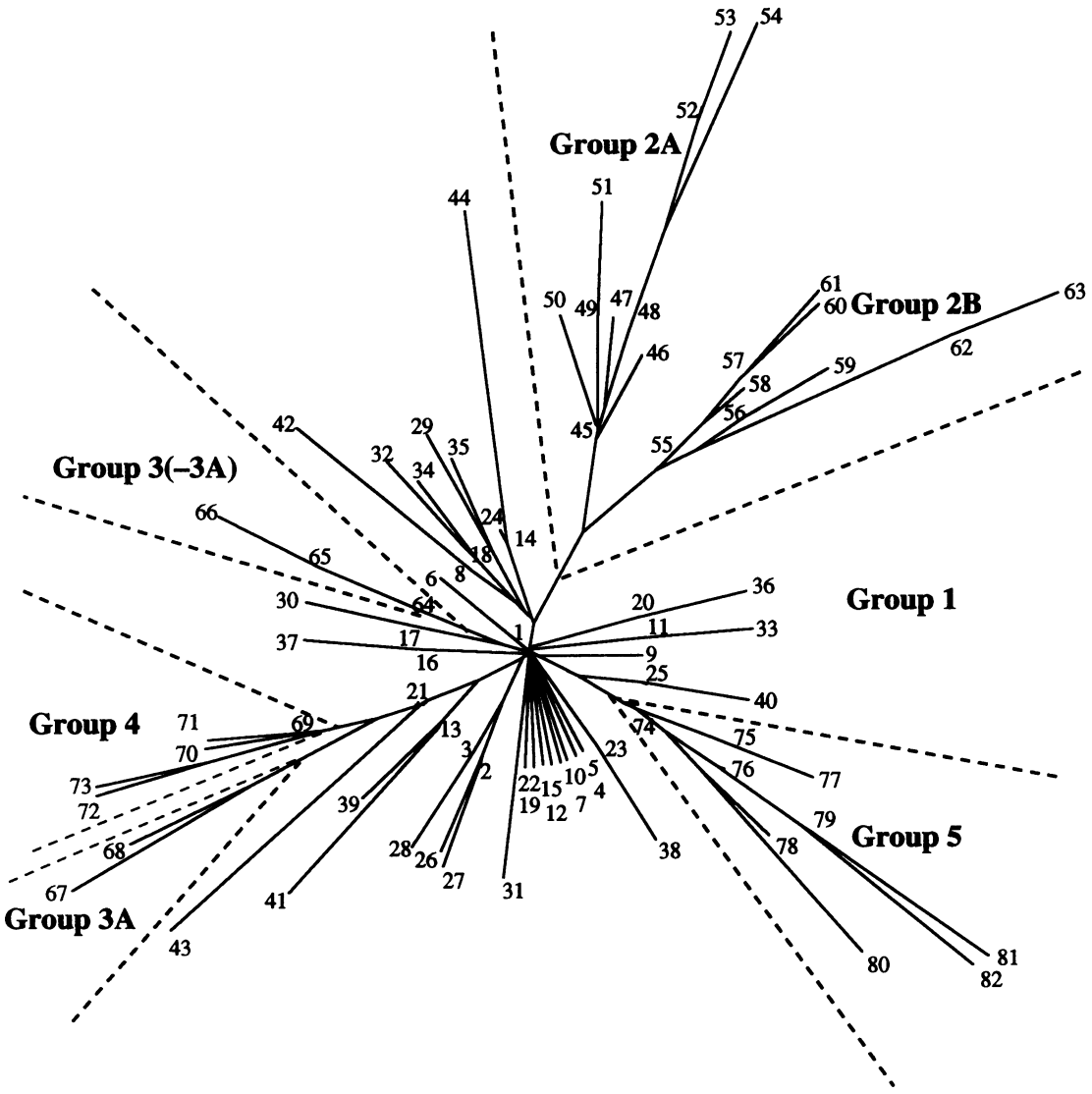


Figure A1 Unrooted NJ tree constructed using the NEIGHBOR program of the PHYLIP package version 3.57e (Felsenstein 1989) from the same data set as figure 2. Haplotype numbers are labeled, and groups 1–5 distinguished.

which the sequences are very closely related. The skeleton network (fig. 2) is almost entirely one-step connected, that is to say, there is only a single mutational step between any two adjacent haplotypes in most cases. However, there are regions of ambiguity resulting from recurrent mutations that need to be identified rather than arbitrarily resolved.

Bootstrapping could in principle be applied to median networks but would be quite inappropriate for testing the reliability of the topology. It is straightforward to demonstrate this simply by considering what happens during the resampling of simple cases. Bootstrapping requires a branch length of three mutations in the absence of homoplasy in order that the probability that the branch be resampled reach 95% (Felsenstein 1985). Branches that comprise a single mutational step, but whose existence is strongly indicated by the presence of haplotypes both with and without the mutation, would receive quite low (only ~63%) support, whereas each link between any four haplotypes forming a cycle in a network would receive 43% support (Bandelt et al. 1995). In the latter case, one of the four possible spanning trees (each containing three of the four possible links) might be favored, on the basis of coalescent theory, by taking into account differences in haplotype frequencies, information that a classical analysis would ignore. Bootstrapping can therefore be misleading, since it ascribes to branches values of supposed robustness that are spuriously low in some instances and high in others: in extreme cases, it even provides quite strong

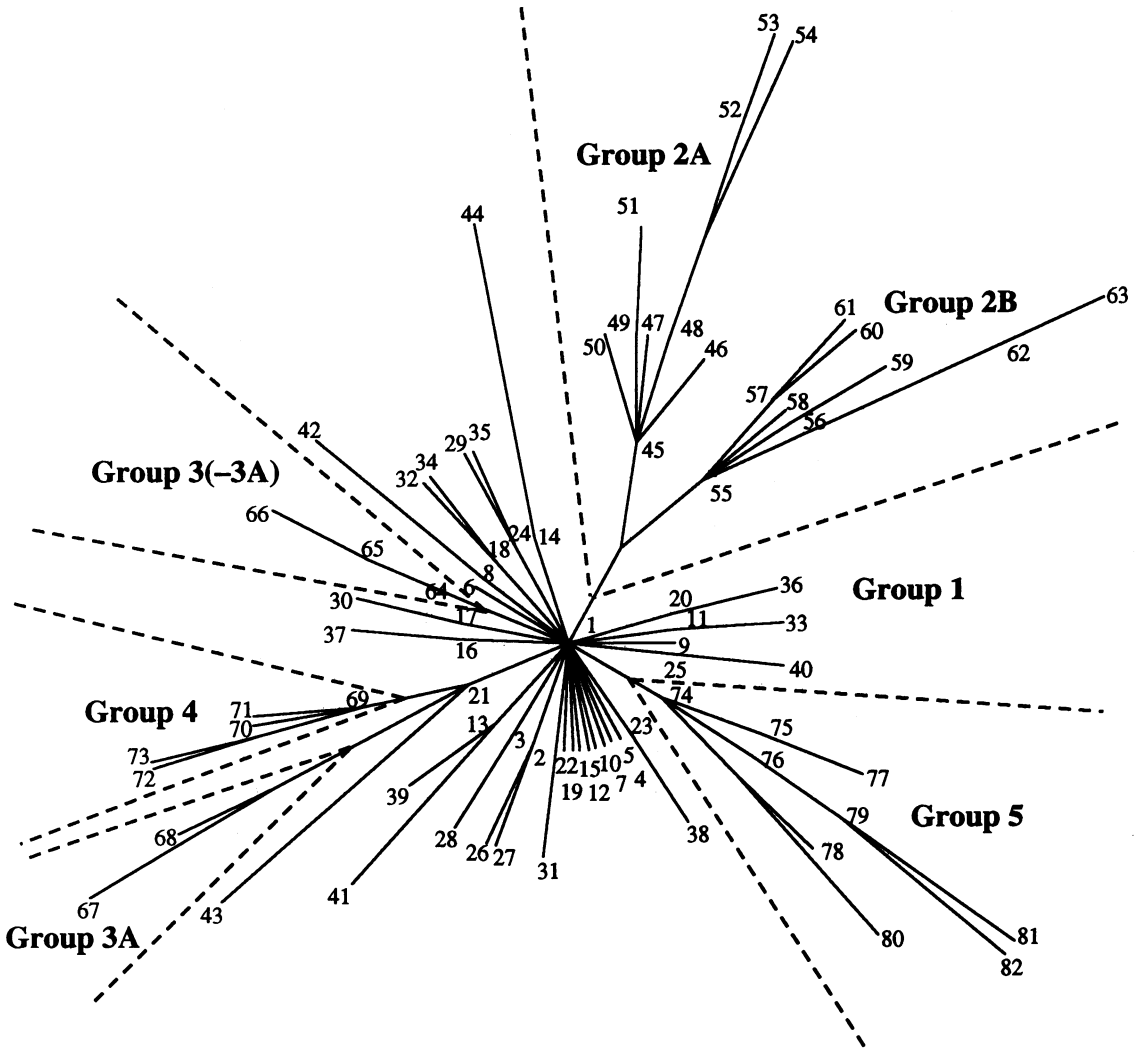


Figure A2 Unrooted NJ tree as in figure A1 but postprocessed by rounding branch lengths to whole numbers. Haplotype numbers are labeled, and groups 1–5 distinguished.

support for branches that are mere artifacts of the tree reconstruction algorithm (Bandelt et al. 1995).

We therefore consider that no existing method usefully quantifies the robustness of branches for population data, whether in tree or network form. However, several lines of qualitative reasoning allow us to be confident of the European skeleton network topology. Because the European population has undergone dramatic expansions, established haplotypes have rarely been lost and almost all of the ancestral nodes can be filled with extant haplotypes. Therefore, there are few parts of the branching structure that are not directly supported by the evidence available. Furthermore, the network is predictive in the following sense: it is constructed using haplotypes occurring more than once, and the network generation algorithm introduces hypothetical intermediate nodes. Each of these nodes identifies a haplotype not present in the data set used for the construction process but potentially existing elsewhere, provided the reconstruction is accurate. A search can be carried out for these inferred haplotypes among a worldwide database of all known sequences, including those occurring only once. Of 22 empty nodes present in the network constructed from the 82 haplotypes occurring more than once, we filled 8 from the European data and a further 3 from the Middle Eastern data, leaving only 11 empty nodes in the final structure.

The topology itself further supports the conclusion that the reconstruction is accurate. As predicted by coalescent theory for a recently expanded population, the most frequent haplotypes are the deepest in a starlike pattern, with rare haplotypes at the tips (Donnelly and Tavaré 1986). Furthermore, the common haplotypes are widely distributed geographically, as one would also ex-

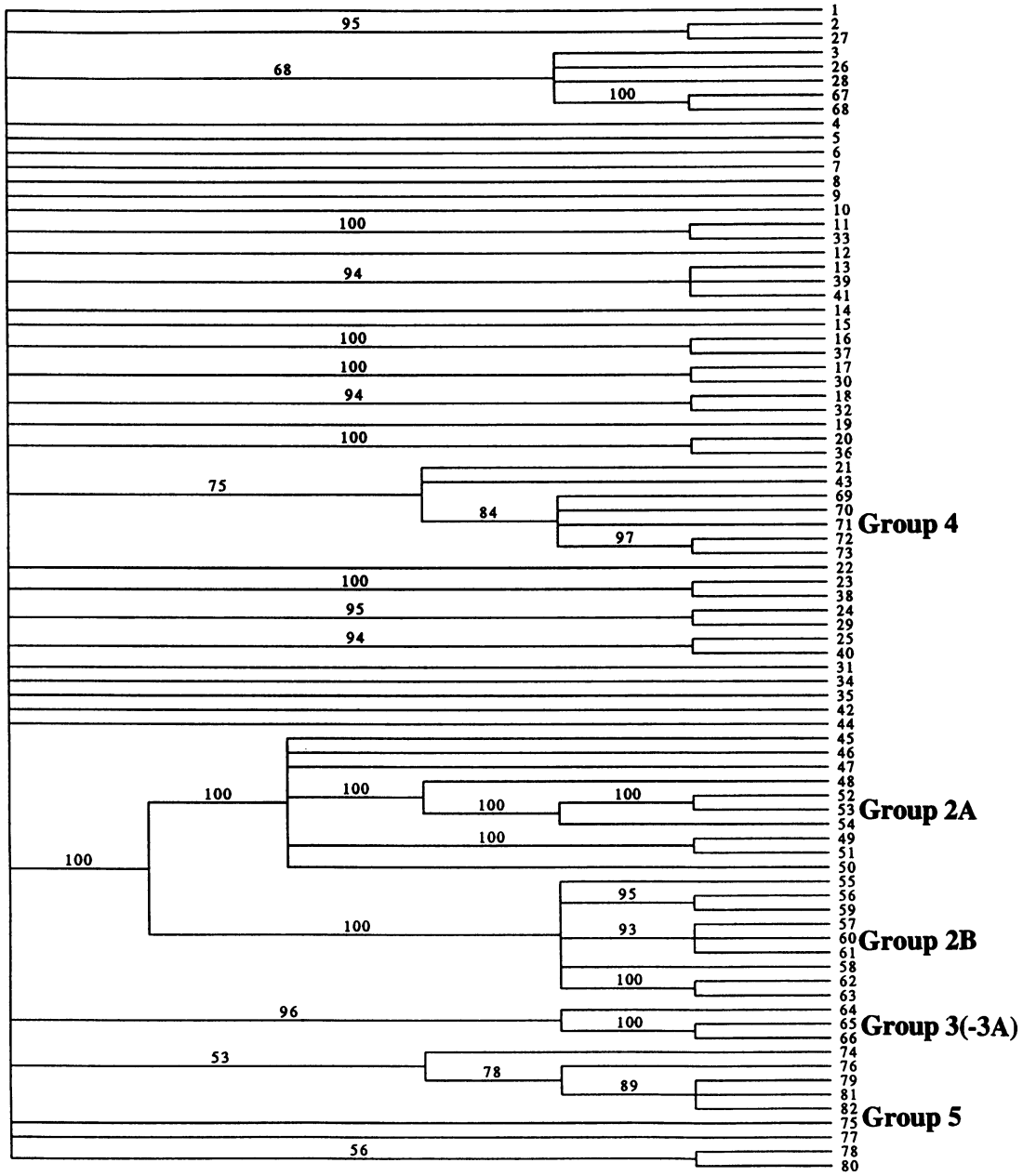


Figure A3 Unrooted majority-rule consensus of 4003 MP trees of length 102 constructed by a 24-h heuristic TBR run of the PAUP 3.1.1 package (Swofford 1993) from the same data set as figure 2. Haplotype numbers and consensus values are labeled, and groups 2–5 marked.

pect of older haplotypes (Templeton 1993). This is the case for groups 1, 2A, and 4, which are the most recently expanded groups on the basis of mean pairwise differences. Groups 2B, 3, and 5 give a different pattern, being less starlike, implying greater age. This is supported by the within-group diversities (which may, however, may be underestimates for less starlike clusters). We would readily admit that the clustering of non-starlike groups is more debatable, but nevertheless plausible ancestral haplotypes can be identified for each of the groups, with additional supporting evidence coming from position 00073 (H. Wilkinson-Herbots, M. Richards, P. Forster, and B. Sykes, unpublished data). It might be appropriate to subdivide these groupings further when more data become available; we note that the divisions between RFLP haplogroups (A. Torroni, K. Huponen, P. Francalacci, M. Petrozzi, L. Morelli, R. Scozzari, D. Obinu, M. L. Savountaus, and D. C. Wallace, unpublished data) match ours, and that their groups are further subdivided in ways that could straightforwardly be matched with control region sequences.

For comparison with the skeleton network, we show an

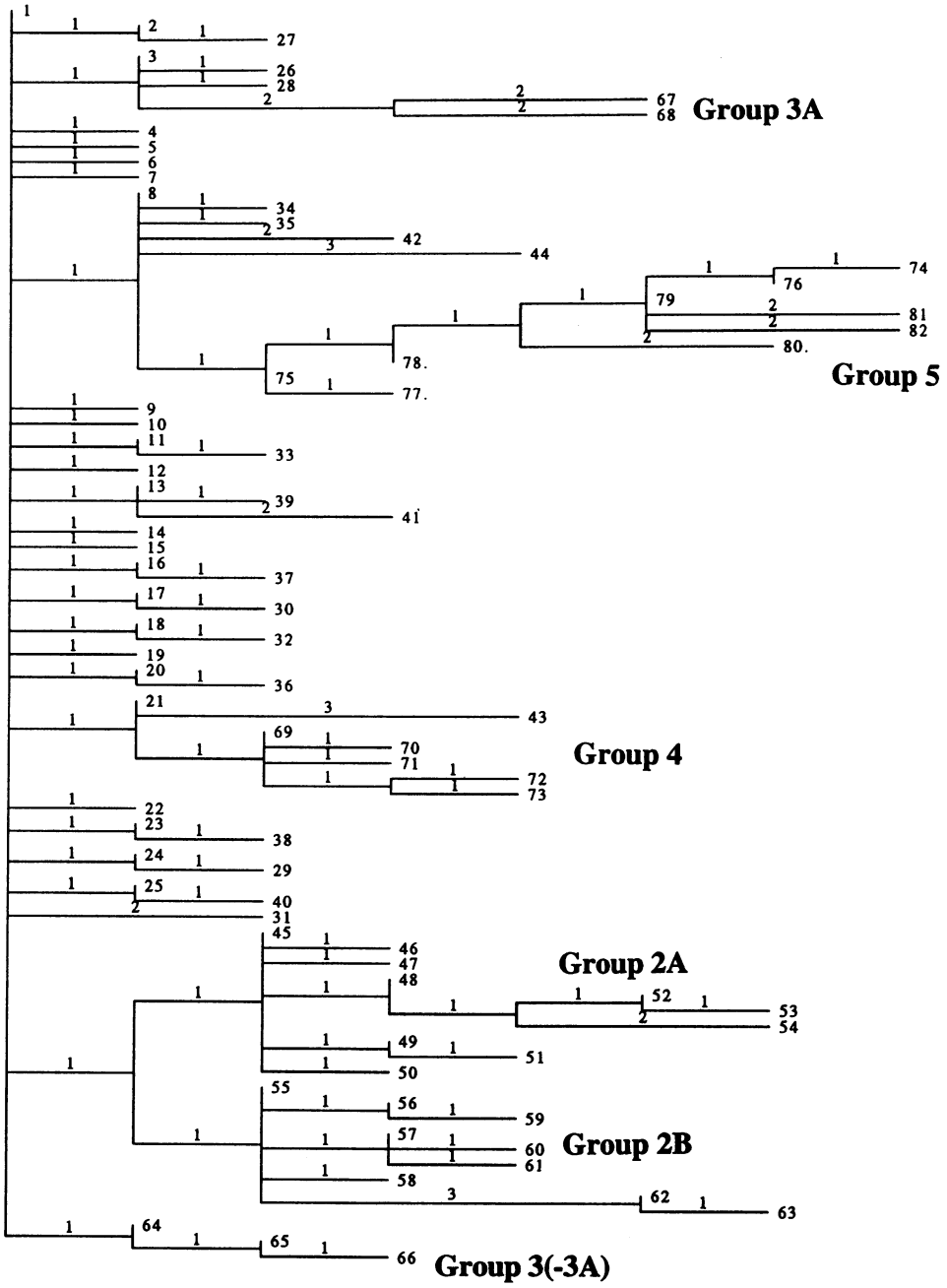


Figure A4 One randomly chosen unrooted MP tree of length 102 from the PAUP run detailed in figure A3, which is contained within the network of figure 2. Haplotype numbers and branch lengths are labeled, and groups 2–5 marked.

unrooted NJ tree constructed from the same data set using PHYLIP (Felsenstein 1989) (fig. A1). While this appears superficially similar to the network, close examination reveals crucial differences. The algorithm does split groups 2A, 2B, 4, and 5 from group 1 cleanly but attaches group 3A and the remaining members of group 3 to group 1 at different points. Group 3 is actually very heterogeneous, since it is likely to be close to the root of the entire phylogeny, and in fact the full reduced median network would allow this pathway, and classification of group 3A relies on the presence of extant intermediates along one of the pathways as well as additional information from position 00073 (see legend to fig. 2).

More seriously, group 1 itself is broken into two clusters in the NJ tree with no fewer than 13 separate nodes representing the CRS, making a total of 12 branches to which no mutations can be assigned (for which we have suggested the expression "ghost links"). There are many

more examples of ghost links throughout the tree and a total of 44 empty bifurcating nodes, in comparison with just 9 in the network (before filling with singletons). These links can be eliminated by postprocessing in the manner of Nerurkar et al. (1993), by rounding branch lengths up or down to whole numbers (fig. A2). Without postprocessing, there are also numerous examples of artifactually early branching caused by the presence of parallel mutations: for example, haplotype 46 branches before the rest of group 2A due to the presence of a parallel mutation at position 189, haplotypes 70 and 71 branch before the rest of group 4 due to the presence of parallel mutations at 93 and 129, respectively, 20 and 36 are pulled toward group 2 by the presence of 304, and 11 and 33 by the presence of 222. The entire artifactual subgroup (8, 42, 14, 44, 18, 32, 34, 24, 29, 35) presumably could be explained in the same way. Again, rounding branch lengths up or down corrects the problem in this example.

The processed NJ tree resembles a putative MP tree 101 steps in length (omitting the insertion); multiple heuristic runs of the package PAUP (Swofford 1993) indeed produce trees of 102 steps. However, problems with MP approaches to mitochondrial data are notorious (e.g., Goldman and Barton 1992). The consensus of 4,003 PAUP trees of length 102 found during one 24-hour heuristic run does identify groups 2A, 2B, and 3A on 100% of occasions and group 4 on 84% of occasions (fig. A3), but this clearly leaves room for many possible branching topologies. One of these, chosen at random, is shown in figure A4: this also resolves group 5, but group 3A has been arbitrarily connected to haplotype 3 rather than to haplotype 64. This also occurs in the consensus tree, whereas in the NJ tree it attaches to haplotype 21, the third possibility. We eliminated the latter possibility from the network by checking for intermediate haplotypes, but the network leaves open the former two possibilities that can be formally resolved only by incorporating additional information from site 00073 in the second hypervariable segment.

The final criticism of any tree is the general point that in many cases the branching pathway cannot be defended (and would not be chosen by MP, which enumerates multiple trees) simply because it is undecidable on the data available, and in such cases to choose one pathway over another arbitrarily is misleading. When no criterion for choice is available, it is preferable that cycles be allowed to remain in the topology so that areas of uncertainty can be isolated and taken into account.

Acknowledgments

We thank Christophe Theopold, Risto Pentinnen, Ann Milward, Debbie Fewster, Mr. and Mrs. David Peterson, Antje Arfsten, and Aysegül Taylan Özkan, for help in sample collection; Kate Smalley and Susannah Tetzner, for technical work; Kingsley Micklem, for help with computation; Antonio Torroni, Paolo Francalacci, Anna Di Rienzo, and Irmgaard Pult, for access to unpublished information; and Rosalind Harding, Paul Pettitt, and Andrew Chamberlain, for critical advice. This work was supported by grants from the Wellcome Trust, Natural Environment Research Council, and the European Union Human Capital and Mobility Network (Biological History of European Populations contract no. ERBCHBGCT920032). The sequence data referred to in this manuscript has been deposited in the Genbank data library under accession numbers U53932–U54467.

References

Ammerman AJ, Cavalli-Sforza LL (1984) The Neolithic Transition and the Genetics of Populations in Europe. Princeton University Press, Princeton

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, et al (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457–465

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Barbujani G, Sokal RR, Oden NL (1995) Indo-European origins: a computer-simulation test of five hypotheses. Am J Phys Anthrop 96:109–132

Bendall KE, Sykes BC (1995) Length heteroplasmy in the first hypervariable segment of the human mtDNA control region. Am J Hum Genet 57:248–256

Bertranpetit J, Sala J, Calafell F, Underhill PA, Moral P, Comas D (1995) Human mitochondrial DNA variation and the origin of Basques. Ann Hum Genet 59:63–81

Bodmer WF (1993) The genetics of Celtic populations. Proc Br Acad 82:37–57

Bodmer WF, Cavalli-Sforza LL (1976) Genetics, evolution, and man. WH Freeman, San Francisco

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Castelloe J, Templeton AR (1994) Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol Phylogen Evol 3:102–113

Cavalli-Sforza LL, Menozzi P, Piazza A (1993) Demic expansions and human evolution. Science 259:639–646

——— (1994) The history and geography of human genes. Princeton University Press, Princeton

Dennell R (1983) European economic prehistory: a new approach. Academic Press, London

Di Rienzo A, Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc Natl Acad Sci USA 88:1597–1601

Donnelly P, Tavaré S (1986) The ages of alleles and a coalescent. Adv Appl Prob 18:1–19

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

——— (1989) PHYLIP: phylogeny inference package version 3.2. Cladistics 5:164–166

Gamble C (1986) The palaeolithic settlement of Europe. Cambridge University Press, Cambridge

Goldman N, Barton NH (1992) Genetics and geography. Nature 357:440–441

Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, Excoffier L (1995) Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. Mol Biol Evol 12:334–345

Handt O, Richards M, Tromsdorff M, Kilger C, Simanainan J, Georgiev O, Bauer K, et al (1994) Molecular genetic analyses of the Tyrolean ice man. Science 264:1775–1778

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46:828–842

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographic sub-structure. Mol Biol Evol 9: 138–151

Lundstrom R, Tavaré S, Ward RH (1993) Estimating substitution rates from molecular data using the coalescent. Proc Natl Acad Sci USA 89:5961–5965

Macaulay VA, Micklem K (1995) MacPairwise 5.0. Locally developed computer program

Mellars PA (1992) Archaeology and the population-dispersal hypothesis of modern human origins in Europe. Phil Trans R Soc Lond B 337:225–234

Morin PA, Moore JJ, Chakraborty R, Jin L, Goodell J, Woodruff DS (1994) Kin selection, social structure, gene flow and the evolution of chimpanzees. Science 265:1193–1201

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Nerurkar VR, Song K-J, Saitou N, Melland RR, Yanagihara R (1993) Interfamilial and interfamilial genomic diversity and molecular phylogeny of human T-cell lymphotrophic virus type 1 from Papua New Guinea and the Solomon islands. Virology 196:506–513

Piercy R, Sullivan KM, Benson N, Gill P (1993) The application of mitochondrial sequence DNA typing to the study of white Caucasian genetic identification. Int J Leg Med 106: 85–90

Pult I, Sajantila J, Simanainan O, Georgiev W, Schaffner W, Pääbo S (1994) Mitochondrial DNA sequences from Switzerland reveal striking homogeneity of European populations. Biol Chem Hoppe-Seyler 375:837–840

Renfrew C (1987) Archaeology and language. Penguin, London

Sajantila A, Lahermo P, Anttinen T, Lukka M, Sistonen P, Savontaus M-L, Aula P, Beckman L, et al (1995) Genes and languages in Europe: an analysis of mitochondrial lineages. Genome Res 5:42–52

Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T, Stoneking M (1994) Mismatch distributions of mtDNA reveal recent human populations expansions. Hum Biol 66: 761–775

Shields GF, Schmeichen AM, Frazier BL, Redd A, Voevoda MI, Reed JK, Ward RH (1993) mtDNA sequences suggest a recent evolutionary divergence for Beringian and North American populations. Am J Hum Genet 53:549–562

Sokal RR (1991) The continental population structure of Europe. Ann Rev Anthropol 20:119–140

Sokal RR, Oden NL, Wilson C (1991) Genetic evidence for the spread of agriculture in Europe by demic diffusion. Nature 351:143–145

Stoneking M, Hedgecock D, Higuchi RG, Vigilant L, Erlich HA (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence specific oligonucleotide probes. Am J Hum Genet 48:370–382

Swofford DL (1993) PAUP: phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign

Stringer C (1993) Secrets of the pit of bones. Nature 362: 501–2

Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet 57:1463–1475

Templeton AR (1993) The "Eve" hypothesis: a genetic critique and reanalysis. Am Anthropol 95:51–72

Torroni A, Lott MT, Cabell MF, Chen Y-S, Lavergne L, Wallace DC (1994) mtDNA and the origin of Caucasians: identification of ancient Caucasian-specific haplogroups, one of which is prone to a recurrent somatic duplication in the D-loop region. Am J Hum Genet 55:760–776

van Andel TH, Runnels CN (1995) The earliest farmers in Europe. Antiquity 69:481–500

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Walsh PS, Metzger DA, Higuchi R (1991) Chelex-100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques 10:506–513

Ward RH, Frazier BL, Dew-Jager K, Pääbo S (1991) Extensive mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci USA 88:8720–8724

Whittle A (1985) Neolithic Europe: a survey. Cambridge University Press, Cambridge

Willis KJ, Bennett KD (1994) The Neolithic transition—fact or fiction? Palaeo-ecological evidence from the Balkans. Holocene 4:326–330

Zvelebil M (1986) Mesolithic prelude and neolithic revolution. In: Zvelebil M (ed) Hunters in transition: mesolithic societies of temperate Eurasia and their transition to farming. Cambridge University Press, Cambridge, pp 5–15