Letters to the Editor

Am. J. Hum. Genet. 66:1173–1177, 2000

mtDNA Haplogroups and Frequency Patterns in Europe

To the Editor:

Recently, an article by Simoni et al. (2000), who used (i) SAAP analysis to analyze the population frequencies of mtDNA haplogroups and (ii) AIDA analysis to examine both the frequency and the sequence similarity of truncated mtDNA sequences, appeared in this *Journal*. The main outcome of their study was that "the overall patterns of mtDNA diversity appear to be poorly significant in Europe." The raw data comprised 2,619 hypervariable segment I (HVS-I) sequences (denoted as "HVR-I" [hypervariable region I] sequences by Simoni et al. [2000]) that were obtained from 36 regions or populations of Europe, the Near East, and the Caucasus and that were collected from both the literature and unpublished sources. Simoni et al. ostensibly grouped the HVS-I sequences according to haplogroup motifs proposed elsewhere (Richards et al. 1998), and they reported the resulting frequencies for each region/population in table 3 in their study. We have checked the input data displayed in table 3 and have found serious technical errors affecting numerous entries. More critically, the mtDNA categories that they report correspond neither to their own criteria nor to the haplogroup definitions established in the literature (to which they refer). Furthermore, their decision to truncate HVS-I information (and to disregard RFLP information) renders these data inadequate to differentiate even African and East Asian sequences from European sequences in many cases.

Inspection of table 3 in the study by Simoni et al. (2000) reveals that (i) the data in the "Galicia" and "Spain: Central" rows have been, in part, crossed-over, (ii) the data in the "Belgium," "Alps," and "Turkey" rows have been computed with the use of sample sizes smaller than those reported in table 1 in the same study, (iii) the haplogroup "J" column has been totally randomized, and (iv) the "Other" column is complementary to the last four "superhaplogroup" columns but not to the first 11 haplogroup columns. As for item (iii), almost all positive entries in the haplogroup "J" column have been either displaced or calculated with the use of sample sizes corresponding to nearby rows. Hence, most entries in this column diverge widely from the real haplogroup J frequencies (see the last column of table 1 in the present study).

As an example of their haplogroup assignment, Simoni et al. (2000) specifically referred to the motif 16069T–16126C for haplogroup J, but they overlooked the fact that this criterion cannot formally be applied to the sequences in the study by Richards et al. (1996), since these were reported only between 16090 and 16365. This might explain some of the many "0" entries in the haplogroup "J" column of table 3 in the Simoni et al. study (see table 1 in the present study). Simoni et al. should have either adopted the haplogroup J frequencies reported by Richards et al. (1996), excluded these population samples from their study, or trimmed all data to the shortest common segment. In the latter case, by employing the motif 16126C–16294C, one could take the default cluster JT-T (comprising all JT sequences that are not T) as a crude default criterion for haplogroup J (see table 1 in the present study).

The discrepancies in haplogroup frequencies are by no means restricted to haplogroup J. Table 2 in the present study shows the marked contrast between published haplogroup frequencies and those assumed by Simoni et al. (2000) for the well-characterized Tuscan, Druze, and Adygei samples (which were typed for RFLPs as well as for HVS-I sequences by Torroni et al. [1996] and Macaulay et al. [1999]). The large differences in frequency for haplogroup H, the most-common European haplogroup, are due to the premise of Simoni et al. (2000) that haplogroup "H contains all sequences . . . that show none of the 22 substitutions considered in this study." This extreme simplification results, on the one hand, in the dumping of large numbers of haplogroup H mtDNAs mainly into the default category "Other" and, on the other hand, in the inclusion of several non-H sequences within their haplogroup H category. For instance, by their criterion, 10/20 haplogroup H mtDNAs from the Tuscan sample (Torroni et al. 1996) would no longer be scored as "H," whereas the U sequence 16051G–16309G–16318C would be scored as "H." In consequence, the haplogroup H category described by Simoni et al. (2000) is bound to be highly polyphyletic

Table 1

Haplogroup J Frequencies According to Simoni et al. (2000), a Crude Default Criterion, and Inference in the Present Study

Population/Region[a]	Sample Size[b]	Haplogroup J Frequency (%) According to		
		Simoni et al. (2000)	Crude Default Criterion (16126C–16294C)[c]	Inference in the Present Study[d]
Austria	117	.0	12.8	11.1
Cornwall	69	.0	21.7	21.7
United Kingdom mainland	100	1.1	12.0	12.0
Wales	92	.0	15.2	15.2
Bulgaria	30	6.7	10.0	10.0
Adygei	50	.0	4.0	4.0
Denmark	33 (32)	.0	18.2	18.2
Estonia	28	.0	7.1	7.1
Finland	79	4.4	8.9	8.9
North Germany	107 (108)	17.8	9.3	8.4
South Germany	249	.0	10.0	8.8
Iceland	53	1.4	17.0	17.0
Druze	45	.0	11.1	6.7
Tuscany	49	.0	18.4	14.3
Karelia	83	.0	6.0	3.6
Near East	42	.0	35.7	19.0
Norway	30	1.9	.0	.0
Portugal	54	.0	7.4	5.6
Saami[e]	312 (240)	1.9	3.8	.0
Basques	106	.0	2.8	2.8
Catalunya	15	.0	6.7	6.7
Mixed Spain[f]	74	.4	9.5	8.1
Galicia	92	.0	9.8	8.7
Sweden	32	2.7	9.4	9.4
Switzerland	70 (72)	.0	11.4	11.4
Turkey	96 (95)	.0	16.7	15.6
Volga-Finnic	34	3.2	17.6	17.6

[a] Population samples from published data tables and data banks cited by Simoni et al. (2000).

[b] Samples sizes are taken from the original sources, except for the Swiss sample, where four close maternal relatives were excluded (Richards et al. 1998, p 243). The sample sizes actually employed by Simoni et al. (2000) are given in parentheses whenever there is a discrepancy.

[c] The mechanical application of this criterion captures a number of non-J sequences with motif 16126C–16362C, especially in Near Eastern populations.

[d] The inference was made on the basis of (i) the motif 16069T–16126C and in conjunction with partial screening of 16069 (Richards et al. 1996), (ii) incorporating of HVS-II (J motif 00295T; Torroni et al. 1996) and RFLP information whenever available, and (iii) appreciating recurrent mutations at 16126.

[e] According to the original sources, an unspecified number of the 312 Saami may be related to each other.

[f] Denoted as "Central Spain" by Simoni et al. (2000).

in the mtDNA genealogy and does not reflect the spatial patterns of haplogroup H.

At this point, it is important to clarify what haplogroup classification entails. An mtDNA haplogroup, when properly defined, is a monophyletic clade of the mtDNA genealogy. Originally, high-resolution RFLP analysis (employing 14 enzymes) had been used for identification of clades by signature sites (Torroni et al. 1992, 1993, 1994*a,* 1994*b,* 1996; Chen et al. 1995), and current haplogroup nomenclature originated in that context. In retrospect, this approach is indeed quite reliable, although recurrent changes at a few sites, such as 10394 *Dde*I, may occasionally cause problems. Potential ambiguities can largely be resolved by incorporation of information from other segments of mtDNA sequences or specific positions of the coding regions (Torroni et al. 1997; Brown et al. 1998; Starikovskaya et al. 1998; Macaulay et al. 1999; Quintana-Murci et al. 1999; Schurr et al. 1999). For instance, haplogroup K is now understood to be a clade (as are U1–U6) within haplogroup U. HVS-I data in combination with partial RFLPs can sometimes serve as a satisfactory substitute for a full RFLP analysis (Rando et al 1998, 2000; Kivisild et al. 1999*a,* 1999*b*).

Table 2

Haplogroup Frequencies, According to Simoni et al. (2000) vs. the Original Studies, in Tuscan, Druze, and Adygei Populations

Population and Study	Sample Size	Haplogroup Frequency (%)											
		H	I	J	K	T	U3	U4	U5	V	W	X	U[a]
Tuscan:													
Simoni et al. (2000)	49	22.4	4.1	0	6.1	6.1	0	4.1	6.1	0	2.0	4.1	16.3
Francalacci et al. (1996), Torroni et al. (1996)[b]	48	41.7	4.2	14.6	6.3	10.4	0	2.1	4.2	0	2.1	8.3	16.7
Adygei:													
Simoni et al. (2000)	50	22.0	6.0	0	2.0	14.0	14.0	4.0	8.0	0	0	0	28.0
Macaulay et al. (1999)	50	30.0	0	4.0	2.0	14.0	14.0	2.0	8.0	0	2.0	0	34.0
Druze:													
Simoni et al. (2000)	45	24.4	4.4	0	15.6	4.4	0	0	0	0	0	17.8	15.6
Macaulay et al. (1999)	45	13.3	2.2	6.7	15.6	4.4	0	0	0	0	0	26.7	26.7

[a] Simoni et al. (2000) had labeled this category as "KU, " which is a misnomer since U encompasses not only U3–U5 but also K and other clusters (Richards et al. 1998; Macaulay et al. 1999).

[b] Of the 49 Tuscans reported in Francalacci et al. (1996), 48 were RFLP analyzed by Torroni et al. (1996).

Unfortunately, HVS-I data alone, which have been produced en masse, often do not contain sufficient information for confident assignment of haplogroup affiliation. The truncation of the HVS-I data to only 13–22 variant positions, as performed by Simoni et al. (2000), yields even poorer results. For example, the motif 16223T–16278T, which was used by Simoni et al. to identify haplogroup X, would transfer most African L1/L2 sequences (Watson et al. 1997; Rando et al. 1998) into the then artefactual category "X." For Europe, this is relevant insofar as a few L1/L2 sequences are present in Iberia (Rocha et al. 1999), and there even resides an African L1c sequence with the motif 16223T–16278T in the British data (Piercy et al. 1993). In addition, as was previously pointed out (Torroni et al. 1996; Macaulay et al. 1999), one has to be prepared for recurrent mutations in the HVS-I motifs (compare also figs. 4, 5, 8, and 9 of the study by Richards et al. [1998]). For instance, the frequency discrepancy (17.8% vs. 26.7%) for haplogroup X in the Druze sample (see table 2 in the present study) is due to the fact that Simoni et al. did not include four haplogroup X mtDNAs that have mutated to 16223C. Another of the many possible examples of misclassification caused by the use of truncated motifs is illustrated by 16129A–16223T, the motif used by Simoni et al. for classification of haplogroup I mtDNAs. Use of this truncated motif has led them to classify both the Asian haplogroup C mtDNAs (16129A–16223T–16298C–16327T) of the Adygei (6.0%) and the East African haplogroup M1 mtDNA (16129A–16189C–16223T–16249C–16311C–16359C) of the Druze (2.2%) as members of haplogroup I (see table 2 in the present study).

The issue of haplogroups only affects the SAAP analysis. However, there are also serious difficulties with the AIDA analysis. Ideally, AIDA should be applied to full DNA-sequence data, but Simoni et al. (2000) included only 22/241 variant positions. One cannot expect that such a truncated data set would show much evidence of geographic patterns within Europe. Most of the haplogroup diagnostic variants in western Eurasian mtDNA are very ancient, and they probably evolved in the Near East and subsequently spread to Europe (Torroni et al. 1998; Macaulay et al. 1999); at any event, they occur throughout western Eurasia. The more recent "rare substitutions," which have evolved since the earlier dispersals and which Simoni et al. (2000) discarded as "statistical noise," are precisely those that are most likely to show regional distributions. The exclusion of such mutations severely restricts the capacity to identify phylogeographic units and, thus, is bound to have seriously reduced the power of the approach to detect autocorrelation.

Even when haplogroup assignment is done with care, failure to detect significant clines in haplogroup frequencies does not prove the absence of any spatial structure in the mtDNA pool. Such structure would rather be manifest at a phylogenetically finer scale (defined on the basis of more-recent mutations). In any case, one would not expect that meaningful patterns of mtDNA diversity could emerge from analyses based on categories with no demonstrable phylogenetic support.

Antonio Torroni,[1,2] Martin Richards,[3] Vincent Macaulay,[4] Peter Forster,[5] Richard Villems,[6] Søren Nørby,[7] Marja-Liisa Savontaus,[8] Kirsi Huoponen,[8] Rosaria Scozzari,[2] and Hans-Jürgen Bandelt[9]

[1]*Istituto di Chimica Biologica, Università di Urbino, Urbino, Italy;* [2]*Dipartimento di Genetica e Biologia Molecolare, Università "La Sapienza," Rome;* [3]*Galton Laboratory, Department of Biology, University*

College London, London; [4]Department of Statistics, University of Oxford, Oxford; [5]McDonald Institute for Archaeological Research, Cambridge, United Kingdom; [6]Estonian Biocentre, Tartu, Estonia; [7]Laboratory of Biological Anthropology, Institute of Forensic Medicine, University of Copenhagen, Copenhagen; [8]Department of Medical Genetics, University of Turku, Turku, Finland; and [9]Mathematisches Seminar, Universität Hamburg, Hamburg

References

Brown MD, Hosseini SH, Torroni A, Bandelt H-J, Allen JC, Schurr TG, Scozzari R, et al (1998) mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? Am J Hum Genet 63:1852–1861

Chen YS, Torroni A, Excoffier L, Santachiara-Benerecetti AS, Wallace DC (1995) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am J Hum Genet 57:133–149

Francalacci P, Bertranpetit J, Calafell F, Underhill PA (1996) Sequence diversity of the control region of mitochondrial DNA in Tuscany and its implications for the peopling of Europe. Am J Phys Anthropol 100:443–460

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, Laos S, et al (1999*a*) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. Curr Biol 9:1331–1334

Kivisild T, Kaldma K, Metspalu M, Parik J, Papiha S, Villems R (1999*b*) The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World. In: Papiha S, Deka R, Chakraborty R (eds) Genomic diversity: applications in human population genetics. Plenum, New York, pp 135–152

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, et al (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232–249

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

Quintana-Murci L, Semino O, Bandelt H-J, Passarino G, McElreavey K, Santachiara-Benerecetti AS (1999) Genetic evidence for an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. Nat Genet 23:437–441

Rando JC, Cabrera VM, Larruga JM, Hernández M, Gonzáles AM, Pinto F, Bandelt H-J (2000) Phylogeographic patterns of mtDNA reflecting the colonisation of the Canary Islands. Ann Hum Genet 63:413–428

Rando JC, Pinto F, Gonzáles AM, Hernández M, Larruga JM, Cabrera VM, Bandelt H-J (1998) Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. Ann Hum Genet 62:531–550

Richards M, Côrte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, et al (1996) Paleolithic and neolithic lineages in the European mitochondrial gene pool. Am J Hum Genet 59:185–203

Richards MB, Macaulay VA, Bandelt H-J, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62:241–260

Rocha H, Flores C, Campos Y, Arenas J, Vilarinho L, Santorelli FM, Torroni A (1999) About the "pathological" role of the mtDNA T3308C mutation.... Am J Hum Genet 65: 1457–1459

Schurr TG, Sukernik RI, Starikovskaya YB, Wallace DC (1999) Mitochondrial DNA variation in Koryaks and Itel'men: population replacement in the Okhotsk Sea–Bering Sea region during the Neolithic. Am J Phys Anthropol 108:1–39

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000) Geographic patterns of mtDNA diversity in Europe. Am J Hum Genet 66:262–278

Starikovskaya YB, Sukernik RI, Schurr TG, Kogelnik AM, Wallace DC (1998) mtDNA diversity in Chukchi and Siberian Eskimos: implications for the genetic history of ancient Beringia and the peopling of the New World. Am J Hum Genet 63:1473–1491

Torroni, A, Bandelt H-J, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, et al (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. Am J Hum Genet 62: 1137–1152

Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, et al (1996) Classification of European mtDNAs from an analysis of three European populations. Genetics 144:1835–1850

Torroni A, Lott MT, Cabell MF, Chen YS, Lavergne L, Wallace DC (1994*a*) 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

Torroni A, Miller JA, Moore LG, Zamudio S, Zhuang J, Droma T, Wallace DC (1994*b*) Mitochondrial DNA analysis in Tibetmplications for the origin of the Tibetan population and its adaptation to high altitude. Am J Phys Anthropol 93:189–199

Torroni A, Petrozzi M, D'Urbano L, Sellitto D, Zeviani M, Carrara F, Carducci C, et al (1997) Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. Am J Hum Genet 60: 1107–1121

Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, Larsen M, Smith DG, et al (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53:563–590

Torroni A, Schurr TG, Yang C-C, Szathmary EJE, Williams RC, Schanfield MS, Troup GA, et al (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 130:153–162

Watson E, Forster P, Richards M, Bandelt H-J (1997) Mitochondrial footprints of human expansions in Africa. Am J Hum Genet 61: 691–704

Address for correspondence and reprints: Dr. Guido Barbujani, Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, I-44100 Ferrara, Italia. E-mail: bjg@dns.unife.it

0002-9297/2000/6603-0040$02.00

Am. J. Hum. Genet. 66:1177–1179, 2000

Reconstruction of Prehistory on the Basis of Genetic Data

To the Editor:

In their letter, Torroni et al. (2000) express a radical disagreement with the assumptions, methods, and conclusions of Simoni et al.'s (2000) article. We think that their many criticisms can be reduced to four points:

1. Haplogroups have been incorrectly defined, and therefore the spatial autocorrelation analysis (SAAP) of their frequencies is flawed;
2. Aside from these errors, the frequencies of haplogroup J and of superhaplogroup JT do not match previous reports;
3. Only 22 polymorphic sites have been considered, and therefore the results of AIDA are flawed;
4. Meaningful patterns of mtDNA diversity can only be identified by the analysis of the distributions of recent mutations.

Point 2 is correct. In the article by Simoni et al. (2000), the column with the frequencies of haplogroup J is wrong, and the frequencies of several haplogroups in Galicia and Spain have been put in each other's places. We apologize to the readers for these errors. However, the correct data (see the erratum published in this issue of the *Journal*) were used in all the analyses, including SAAP, and therefore the autocorrelation results in table 5 in the article by Simoni et al. (2000) are correct. Before we consider the other points, it is important to exactly define the subject of this discussion.

The general question being asked in our study and in similar studies is: What combination of evolutionary factors is most likely to account for the current levels and patterns of genetic diversity? To answer this question, one has to study as many loci as possible and has to study them by using the same statistical methods, so that the results will be comparable. The methods of SAAP and AIDA are especially suitable, because they have long been used to summarize both protein (Sokal and Menozzi 1982; Sokal et al. 1989; O'Rourke et al. 1992; Epperson and Li 1996; Crawford et al. 1997) and DNA (Barbujani et al. 1995; Chikhi et al. 1998; Casalotti et al. 1999; Krings et al. 1999; Rickards et al. 1999) diversity.

Point 1: haplogroup definition and SAAP.—AIDA can be directly applied to any set of DNA data, whereas SAAP processes frequencies and therefore requires prior definition of the entities whose frequencies will be analyzed. AIDA found very little spatial structuring of mtDNA. To confirm this result, we reanalyzed the data by using SAAP, and hence we had to identify evolutionarily meaningful clusters of hypervariable region 1 (HVR-1) haplotypes, or haplogroups.

The categories that we used for that purpose—and that Torroni et al. (2000) question—were proposed by Richards et al. (1998) in a paper cosigned by two other authors of Torroni et al.'s letter. The classification of mitochondrial haplotypes is no easy task; there is an unresolved uncertainty about the best way to cluster and interpret mitochondrial data. Analysis, at the nucleotide level, of the whole mitochondrial genome will be a suitable approach only in the not-so-near future. Indeed, as we stated in the "Database" section of the article by Simoni et al. (2000), only three European samples have been typed at the RFLP level. There is no current alternative to the study of HVR-1 sequences, if one wants to understand whether mitochondrial variation shows any structuring in Europe. On the basis of 22 polymorphic sites, Richards et al. (1998) identified what they consider to be monophyletic clades in the HVR-1 phylogeny, and we chose to use those sites to define haplogroups. Of course, the frequencies of the haplogroups defined in this way do not perfectly overlap with those which are based on RFLPs (table 1 in Torroni et al. 2000). We have since discovered that they even differ between table 2 and figure 2 of Richards et al. (1998), because site 16189 is mentioned as being part of the "X motif" only in the former, and we trusted the latter. That is not our fault.

In quantitative terms, a nonparametric discriminant analysis that we ran on worldwide data shows that 15.3% of suitable mitochondrial data are assigned to different haplogroups that are based on RFLPs or on HVR-1 sequences, with variable levels of disagreement for the different haplogroups—for example, 7.1% for haplogroup J (for more details, see Simoni 2000); that 7.1% of uncertainty accounts, for example, for most of the persisting differences between the haplogroup J frequencies that we considered and the frequencies presented in Torroni et al.'s (2000) table 2.

Answering the criticisms raised by Torroni et al. (2000), which we do not feel are justified, would entail reclassification of just a few sequences, <20 for haplogroups X and U4, of a total of >800 distinct sequences. After they have been reallocated, the SAAP coefficients do not change, up to the second decimal place. Table 1