HUMAN MUTATION 27(9), 965–974, 2006

METHODS

Human mtDNA Site-Specific Variability Values Can Act as Haplogroup Markers

Matteo Accetturo,[1] Monica Santamaria,[1] Daniela Lascaro,[1] Francesco Rubino,[1] Alessandro Achilli,[2] Antonio Torroni,[2] Mila Tommaseo-Ponzetta,[3] and Marcella Attimonelli[1]*

[1]Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Bari, Bari, Italy; [2]Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia, Italy; [3]Dipartimento di Zoologia, Università degli Studi di Bari, Bari, Italy

Communicated by Pui-Yan Kwok

Sequencing of entire human mtDNA genomes has become rapid and efficient, leading to the production of a great number of complete mtDNA sequences from a wide range of human populations. We introduce here a new statistical approach for classifying mtDNA nucleotide sites, simply by comparing the mean simple deviation (MSD) of their specific variability values estimated on continent-specific dataset sequences, without the need for any reference sequence. Excellent correspondence was observed between sites with the highest MSD values and those marking known mtDNA haplogroups. This in turn supports the classification of 81 sites (23 in Africa, eight in Asia, eight in Europe, 34 in Oceania, and eight in America) as novel markers of 47 mtDNA haplogroups not yet identified by phylogeographic studies. Not only does this approach allow refinement of mtDNA phylogeny, an essential requirement also for mitochondrial disease studies, but may greatly facilitate the discrimination of candidate disease-causing mutations from haplogroup-specific polymorphisms in mtDNA sequences of patients affected by mitochondrial disorders. Hum Mutat 27(9), 965–974, 2006.

KEY WORDS: mtDNA; site-specific variability; haplogroup markers; mtDNA mutations

INTRODUCTION

Due to the peculiar features of mitochondrial DNA (mtDNA) (maternal inheritance, absence of recombination, role in cellular energy production, and lack of an efficient repair system), analysis of mtDNA sequence variations has proven to be a powerful tool for investigating human origins and dispersals [Macaulay et al., 2005; Thangaraj et al., 2005; Forster and Matsumura, 2005] and identifying single mutations (or combinations of mutations) that either cause human diseases or play an important role in their expression [Howell et al., 2005; Wallace, 2005].

As regards evolutionary studies, the earliest mtDNA work began by digesting entire mtDNA with a number of restriction enzymes [Denaro et al., 1981; Johnson et al., 1983; Cann et al., 1987], sequencing the hypervariable segments (HVS-I and HVS-II) [Vigilant et al., 1991; Richards et al., 1998] of the D-loop, the main regulatory region of metazoan mitochondrial DNA, or using a combination of the two approaches [Torroni et al., 1993a,b, 1996; Macaulay et al., 1999; Richards et al., 2000; Quintana-Murci et al., 2004]. In the last few years, the advent of more advanced molecular techniques has allowed rapid and efficient sequencing of the entire human mitochondrial genome, leading to the production of a great number of mtDNA sequences from a wide range of human populations [Ingman et al., 2000; Finnilä et al., 2001; Herrnstadt et al., 2002; Ingman and Gyllensten, 2003; Kong et al., 2003; Achilli et al., 2004, 2005; Palanichamy et al., 2004; Tanaka et al., 2004; Friedlaender et al., 2005; Kivisild et al., 2006; Macaulay et al., 2005; Merriwether et al., 2005; Thangaraj et al., 2005; Trejaut et al., 2005].

This massive amount of sequence data provides the opportunity of analyzing sequence variations of human mtDNA with new approaches, with results which may be considered significant and reliable only now that the "raw" data are sufficiently plentiful. These analyses contribute to a more accurate definition of mtDNA haplogroups, defined by a unique set of variations acquired from the same ancient common female ancestor. Here, we introduce a novel statistical approach, based on site-specific variability estimates, i.e., a measure of how variable a site is, in the multialigned set of considered sequences. The method allows identification of nucleotide sites, which are mtDNA haplogroup

Received 11 January 2006; accepted revised manuscript 14 April 2006.

*Correspondence to: Prof. Marcella Attimonelli, Dipartimento di Biochimica e Biologia Molecolare, Università degli Studi di Bari, Via E. Orabona 4, 70126 Bari, Italy. E-mail: m.attimonelli@biologia.uniba.it

Grant sponsor: Ministero Italiano dell'Università e Ricerca, Fondo Investimenti Ricerca di Base 2001, "Bioinformatica per la Genomica e la Proteomica"; Grant sponsor: Progetti Ricerca Interesse Nazionale 2003; Grant sponsor: Progetti Ricerca Interesse Nazionale 2005; Grant sponsor: Progetti Ricerca Interesse Nazionale Fondazione Cariplo; Grant sponsor: Progetti Ricerca Interesse Nazionale ESF (P.O.P. 2000–2006); Grant sponsor: Fondazione Cariplo; Grant sponsor: ESF (European Science Foundation) P.O.P. Piano Operativo Puglia 2006.

Matteo Accetturo, Monica Santamaria, and Daniela Lascaro contributed equally to this work.

DOI 10.1002/humu.20365

Published online 24 July 2006 in Wiley InterScience (www.interscience.wiley.com).

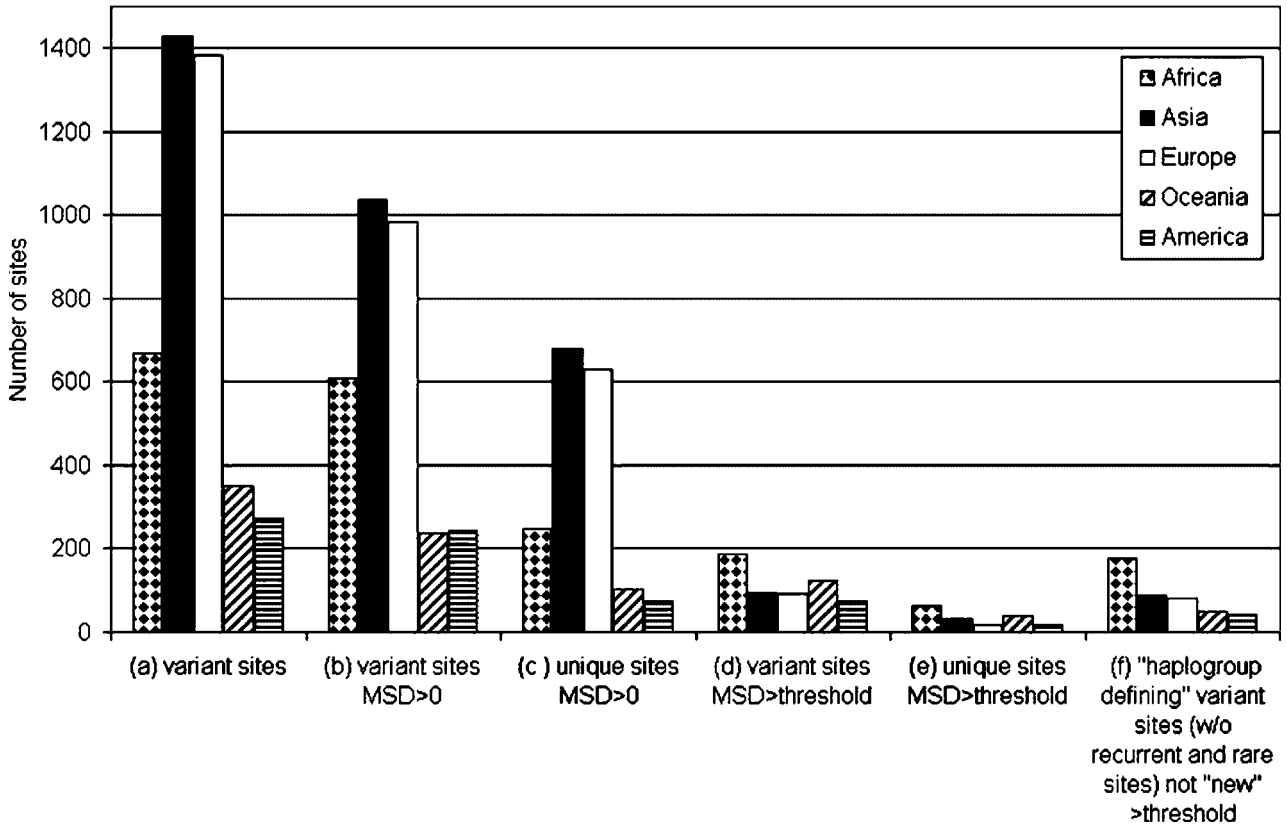

FIGURE 1. **a:** Number of continent-specific variant sites (site-specific variability value >0) in coding part of mitochondrial genome. **b:** Number of variant sites with MSD values greater than zero. **c:** Number of unique continent-specific sites with MSD values greater than zero. **d:** Number of variant sites with MSD values greater than threshold. **e:** Number of unique continent-specific sites with MSD values greater than threshold. **f:** Number of "haplogroup defining" variant sites (with/without those with recurrent or rare mutations) not "new" with values greater than threshold.

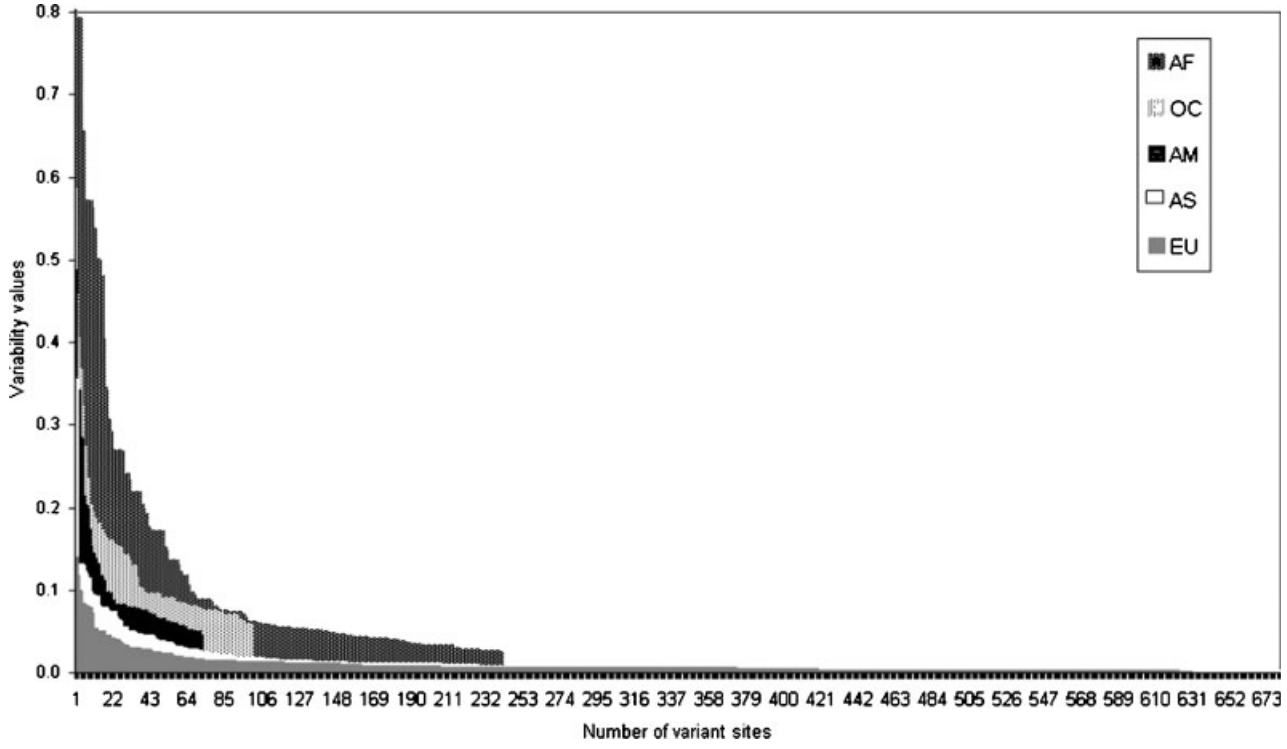

FIGURE 2. **Site-specific variability trends in various continents obtained by application of Site_Var approach.**

TABLE 1. Mean Simple Deviation (MSD) of mtDNA Site Variability Values in Africa With Respect to Other Continents

MSD	Nucleotide position	Confirmed haplogroup(s)[a]	MSD	Nucleotide position	Confirmed haplogroup(s)[a]	MSD	Nucleotide position	Confirmed haplogroup(s)[a]
1.000	825	L0-L1-L5	0.267	6587	L3e1	0.157	921	L3d1-L3d3
0.805	7521	(L w/o L3-L4)	0.267	15942	New	0.156	11899	L1c1
0.802	7256	(L w/o L3-L4)	0.264	14152	L3e1	0.156	5442	L0
0.792	769	(L w/o L3)	0.261	10086	L3b	0.156	13886	L3d
0.792	3594	(L w/o L3-L4)	0.252	13101	L3e3	0.154	7424	L3d
0.792	13650	(L w/o L3-L4)	0.247	5773	L3b	0.154	4454	L1c1a
0.791	4104	(L w/o L3-L4)	0.239	6071	L1c	0.153	750	L3e3
0.791	1018	(L w/o L3)	0.239	9072	L1c	0.151	8618	L3d
0.728	13105	L0-L1-L3b-L3d-New	0.239	14911	L1c	0.151	14034	New
0.683	2352	L1b-L3e-U6b1	0.238	10373	New	0.150	10920	L0k
0.655	9221	L2	0.238	5951	L1c	0.147	15217	L2b-L2c
0.655	10115	L2	0.237	12810	L1c	0.146	3843	L1c1
0.647	8206	L2	0.233	10586	L1c	0.146	9540	L/M
0.632	11944	L2a-L2b-L2c	0.232	5581	New	0.145	10873	L/M
0.597	2416	L2	0.232	3693	L1b-L2d	0.143	1048	L0
0.577	15784	L2a1	0.221	5046	L1b	0.139	1442	L2b-L2c
0.570	2789	L0f-L2a	0.219	10321	L1c1-L1c2	0.139	3200	L2c
0.570	7771	L2a	0.218	5036	L1b	0.138	8701	L/M
0.570	12693	L2a1	0.218	5393	New	0.136	13958	L2c
0.570	13803	L2a	0.218	5655	L1b	0.135	5366	Rare mutation
0.569	7175	L2a	0.218	8248	L1b	0.135	5603	L0a-L0f
0.569	7274	L2a	0.218	14203	L1b	0.135	6875	Rare mutation
0.562	14566	L2a	0.217	1738	L1b	0.135	8428	L0a
0.553	13590	L2-L0k	0.216	3308	L1b	0.135	15136	L0a-L0f
0.542	10819	L3e-L3i	0.213	6827	L1b	0.134	14182	L0k
0.540	14212	L3e	0.211	15115	L1b	0.134	12720	L0a-L0d
0.536	13914	L3b	0.210	9449	L3b	0.133	11641	L0a-L0f
0.513	11914	L0-L2a	0.210	7867	L1b	0.132	5656	U5b1-New
0.500	8655	L0-L1-L5	0.209	14179	New	0.131	5237	L0a2b
0.500	13506	L0-L1-L5	0.205	15670	New	0.131	12630	L3w-New
0.499	2758	L0-L1	0.203	14769	L1b-L3f1	0.130	1692	Rare mutation
0.499	2885	L0-L1	0.202	14905	L3e2	0.128	2000	L3e3
0.498	7146	L0-L1	0.202	6548	L1b	0.128	15431	L0a-L0f
0.498	8468	L0-L1	0.202	6989	L1b	0.127	10667	L3e3
0.494	6221	L1c3-L3b	0.195	8650	New	0.126	7768	U5b
0.483	10688	L0-L1-L5	0.190	15110	L2b-L2c	0.126	6524	L3e3
0.479	14000	L1c	0.190	8566	L0a	0.123	3438	L0d1
0.472	10810	L0-L1-L5	0.188	11800	New	0.122	6680	L3d1-M1
0.404	13880	L1b	0.186	7805	U6a	0.121	8087	L1c1a
0.399	13789	L1	0.184	10915	L0	0.119	15099	Rare mutation
0.397	7389	L1	0.184	12519	L1b	0.118	14180	Recurrent
0.395	14178	L1	0.182	9554	L3e3-New	0.117	6150	New
0.382	7055	L1	0.179	12236	L2b-L2c	0.116	15391	Rare mutation
0.381	15301	L0-L1-L5	0.175	5331	L2b	0.115	5471	Recurrent
0.377	3666	L1	0.172	2245	L0a-L0f	0.115	7385	U5b1b
0.375	14560	L1	0.172	2768	L1b1	0.113	14088	L1c1a
0.358	15244	New	0.172	15229	New	0.111	5231	L0a
0.345	5285	New	0.170	4312	L0	0.111	10927	U5b1b
0.342	15629	New	0.170	4586	L0a-L0k-L0f	0.106	9098	Rare mutation
0.340	3516	Lo	0.170	9042	L0	0.106	9755	L0
0.339	3918	L2a1a	0.170	9347	L0	0.105	14571	Rare mutation
0.319	3796	L1c1	0.170	9818	L0a-L0k-L0f	0.104	3197	U5
0.316	12308	U(U6)	0.170	10664	L0	0.104	13617	U5
0.315	11467	U(U6)	0.170	13276	L0	0.104	9477	U5
0.311	3348	U6	0.167	6185	L0	0.104	6164	Rare mutation
0.305	7624	L2b-L2c	0.167	11002	New	0.103	12618	U5b1b
0.291	3495	New	0.166	5147	L0a2-L3d	0.103	13980	New
0.281	12372	U(U6)	0.165	10589	L0	0.103	6152	Rare mutation
0.270	2332	L2b-L2c	0.162	710	L1b	0.101	12930	New
0.268	3450	L3b	0.162	11176	L0a	0.100	9438	U6b
0.268	13485	L1c	0.159	14148	L1c1	0.100	11204	Rare mutation
0.268	15311	L3b	0.157	4655	L3e3	0.100	11257	New
0.268	15824	L3b	0.157	14284	L3d			

[a]The "New" nucleotide position is a marker of a new predicted haplogroup (see Table 6). The "Rare mutation" was observed only in one or two haplotypes in HmtDB, without defining any specific haplogroup. The "Recurrent" mutation at that nucleotide position is paraphyletic.

markers, and their discrimination from disease-causing mutations, thus greatly facilitating identification of the latter in patients affected by mitochondrial disorders. Indeed, all disease-causing mutations are either rare or recurrent mutations (ones which are less rare and are thus found in mtDNAs belonging to different haplogroups). Instead, haplogroup-specific polymorphisms are

TABLE 2. Mean Simple Deviation (MSD) of mtDNA Site Variability Values in Asia With Respect to Other Continents

MSD	Nucleotide position	Confirmed haplogroup(s)[a]	MSD	Nucleotide position	Confirmed haplogroup(s)[a]
0.757	5178	D	0.083	6962	F1
0.552	9824	M7-D4b2	0.083	8563	A1
0.375	4883	D	0.081	2766	D4d1
0.356	1382	D4b2	0.081	5351	M7b
0.344	8414	D4	0.081	9950	B2-B5-M11
0.327	14668	D4-Z2	0.080	12406	F1
0.273	13928	R9	0.080	10345	M7b2
0.259	10400	M	0.079	5601	G2
0.259	14783	M	0.079	5301	D5
0.233	8020	D4b-F4b-M7b2	0.079	1107	D5
0.216	6455	M7	0.078	13563	G2
0.213	15043	M-I	0.076	7853	M7b
0.171	14569	New-G-B4b1b	0.076	15524	D4b2a
0.166	10310	F-B4c1a	0.075	11017	M7a1a
0.151	5108	B4c2	0.075	15518	New
0.140	8701	L/M-D4g@	0.075	4343	New
0.135	3970	R9	0.075	14200	G2a
0.132	5417	N9	0.074	9180	D5a-D5b
0.131	3206	D4a	0.074	7600	G2a
0.131	14979	D4a	0.073	11969	B4f-C4-M11
0.131	6392	F	0.073	12358	N9a-New
0.130	4386	M7a-N9a1	0.073	11084	M7a1a
0.130	4071	M7b-M7c	0.072	15346	B4c
0.124	4833	G	0.068	14944	A1a1a
0.122	10873	L/M	0.066	9575	G2a
0.120	9540	L/M	0.064	1119	B4c
0.119	9296	D4b2b	0.064	15874	D4e2
0.116	10410	New	0.064	10397	D5
0.114	2626	M7a	0.062	8594	R5
0.114	4958	M7a	0.056	10104	D4b2a
0.110	2772	M7a	0.055	15851	B5b
0.105	8473	D4a	0.055	15662	B5b
0.101	11536	A1	0.054	15508	B5b
0.100	12771	M7a	0.054	827	B4b-B4d-New
0.098	11215	D4e-F1a1a1	0.054	15301	L/M
0.097	709	Recurrent	0.053	10801	A1a1
0.096	10609	F1	0.053	10754	R5
0.095	12405	M7b	0.053	14544	R5
0.095	7684	M7b	0.052	9377	G2a
0.095	13104	D4g-U1	0.052	6026	C4-U1a
0.093	12882	F1	0.051	8684	M8a
0.093	8584	M8-B5-R30	0.051	15954	U1-F1b1a
0.092	14605	New	0.051	13759	F1a-F1c
0.089	4048	M7b	0.051	12811	M7b
0.088	4164	M7b	0.050	15223	B5b
0.087	11647	A1a	0.050	13635	New
0.087	8964	D4b2	0.050	8829	B5b

[a]The "New" nucleotide position is a marker of a new predicted haplogroup (see Table 6). The "Recurrent" mutation at that nucleotide position is paraphyletic.

generally quite common in at least one geographic area/ethnic group, have a deep location in the phylogenetic tree, and have been subjected to selective pressure for tens of thousands of years. Therefore, haplogroup markers cannot be disease-causing mutations, at most, they may play a secondary role in disease expression [Carelli et al., 2006].

The main idea underlying our hypothesis is that mtDNA sites with high variability values found only in a particular geographic area may be considered good haplogroup markers of that area. Variability values satisfying these features are defined here as "discriminating variability values." A major requirement of this approach is that, due to the structure of the algorithm used to estimate variability values, the sample must be sufficiently large for statistically significant results to be obtained, and sufficiently heterogeneous to be sufficiently representative of the populations living in the geographic area of interest. This spurred for the research group to perform a random simulation sampling procedure, starting from the original sequence datasets, in order to assess to what extent the method is dependent on sample characteristics.

MATERIALS AND METHODS

The method is based on the Site_Var algorithm [Pesole and Saccone, 2001], extensively modified in collaboration with the authors in order to adapt it to human mtDNA data and to obtain more precise site-specific variability values. The original version of the algorithm starts from N nucleotide multialigned sequences and for each i-th site in each j-th pair of sequences estimates the δ_{ij} score, which is 1 or 0, depending on the presence or absence of the substitution event in the site, divided by the *j*-th genetic distance calculated according to the stationary Markov model

TABLE 3. Mean Simple Deviation (MSD) of mtDNA Site Variability Values in Europe With Respect to Other Continents

MSD	Nucleotide position	Confirmed haplogroup(s)[a]	MSD	Nucleotide position	Confirmed haplogroup(s)[a]
0.937	15452	J-T	0.118	15833	H5a1
0.892	2706	H	0.118	7864	W1
0.887	7028	H	0.114	1888	T
0.798	11719	pre-HV	0.099	6365	New
0.768	14766	HV	0.099	14793	U5a
0.520	3010	H1-J1	0.092	930	T2b
0.468	4216	J-T	0.090	4769	H2
0.467	11251	J-T	0.087	1719	N1-X2-New
0.461	14798	K-J1c	0.087	15924	I-New
0.337	12308	U	0.086	5004	H4
0.334	11467	U	0.084	14582	H4
0.314	15904	pre*V2-V	0.084	4024	H4
0.303	12372	U	0.084	4793	H7
0.301	4580	V	0.084	3992	H4
0.290	12612	J	0.082	9150	New
0.283	6776	H3	0.081	8869	V1
0.251	709	Recurrent	0.080	14365	H4
0.250	1811	U2-U3-U4-U9-U7-U8	0.080	5495	W1a-New
0.249	9698	U8	0.078	7768	U5b
0.243	9055	U8b-K	0.077	14470	H10-X
0.242	8697	T	0.074	13780	I
0.239	14167	U8b-K	0.073	12669	W1a
0.236	15928	T	0.073	4639	V
0.236	15884	W	0.070	5046	N2
0.236	10463	T	0.069	9899	T1a
0.235	4917	T	0.068	9716	K2
0.234	3480	K	0.065	11377	J2a
0.213	11299	K	0.065	4561	K2a
0.211	13368	T	0.064	8271	New
0.211	10550	K	0.061	12501	N1
0.185	12633	T1	0.061	5263	V1a
0.183	1189	K1	0.060	15218	U5a1-HV1
0.174	11812	T2	0.060	10034	I
0.173	14905	T	0.058	8269	J1b-H4a
0.164	14233	T2	0.057	8705	X2c
0.159	3197	U5	0.056	10044	New
0.158	8251	W-I	0.055	13966	X
0.155	13708	J-X2b	0.055	5656	U5b1
0.153	13617	U5	0.054	10394	New
0.153	9477	U5	0.054	9380	H6a1
0.146	4529	I	0.053	3915	H6a
0.140	4336	H5a	0.050	9066	H1f
0.138	3505	W	0.050	7309	H1f
0.137	11674	N2	0.050	4452	H1f
0.135	11947	W	0.050	15223	B5b
0.128	8994	W	0.050	13635	New
0.121	1243	W	0.050	8829	B5b

[a]The "New" nucleotide position is a marker of a new predicted haplogroup (see Table 6). The "Recurrent" mutation at that nucleotide position is paraphyletic.

[Lanave et al., 1984; Saccone et al., 1990], also known as the general time reversible (GTR model, PAUP package) [Swofford, 2002]. The variability values are then obtained by summing these ratios along N(N−1)/2 pairwise sequences in the multialignment

$$\nu_i = \sum_{j=1}^{N(N-1)/2} \frac{\delta_{ij}}{K_j} \quad (1)$$

In the revised version of the algorithm, the δ_{ij} score is 1 for a transition, 2 for a transversion, and 0 if the site is unchanged. Moreover, in sites where an insertion or deletion is present in some of the multialigned genomes (gapped sites), to variability υ_i is added a score of $2/K_{j\ (mean)}$, where $K_{j\ (mean)}$ is the mean distance along all the N(N−1)/2 pairs of sequences.

The genetic distance is now estimated through the Kimura model [Kimura, 1980], which is more suitable than GTR for intraspecies analyses and in cases in which different weights are assigned to transitions and transversions.

Last, site-specific variability values are also normalized (Eq. 2), relative to the number of all possible pair-wise comparisons, and divided by the maximum ν_i value, ν_{max}, thus producing relative variability γ_i (Eq. 3), with values ranging between 0 and 1;

$$\nu_{i(norm)} = \nu_i/[N(N-1)/2] \quad (2)$$

$$\gamma_i = \nu_{i(norm)}/\nu_{i(max)}. \quad (3)$$

The starting point of our approach is thus a sample of multialigned mtDNA sequences grouped according to their continental origin, on which site-specific variability is estimated.

Data resulting from the application of Site_Var are then automatically processed by introducing the mean simple deviation (MSD) parameter in order to quantify the concept of "discrimi-

TABLE 4. Mean Simple Deviation (MSD) of mtDNA Site Variability Values in Oceania With Respect to Other Continents

MSD	Nucleotide position	Confirmed haplogroup(s)[a]	MSD	Nucleotide position	Confirmed haplogroup(s)[a]
0.967	6719	B4a1a	0.168	3203	P2
0.949	5465	B4a	0.167	15300	New
0.927	9123	B4a	0.166	15852	New
0.919	12239	B4a1a	0.166	12879	New
0.918	15746	B4a1a	0.166	1438	P2
0.883	10238	B4a	0.165	4122	P2
0.862	15607	P	0.164	15885	New
0.832	14022	B4a1a1	0.162	8577	Rare mutation
0.584	8404	S	0.161	13479	Rare mutation
0.489	12705	N/M/L	0.160	14070	Rare mutation
0.472	5843	Q	0.159	5086	Rare mutation
0.466	4117	Q	0.159	5483	Rare mutation
0.464	13500	Q	0.159	6083	Rare mutation
0.462	6366	P1a	0.159	5563	New
0.457	10118	P1	0.157	10700	New
0.456	6077	P1	0.156	7681	Q1a
0.453	12940	Q	0.155	9103	Rare mutation
0.438	8790	Q	0.153	2263	Rare mutation
0.404	15937	P3	0.152	10786	New
0.380	5460	Q1-Q2-New	0.152	15664	New
0.366	14025	Q1	0.150	5302	New
0.322	8964	Q1	0.150	9938	New
0.320	6167	New	0.149	591	Rare mutation
0.319	2380	New	0.147	4025	Rare mutation
0.274	6905	New	0.145	13135	Rare mutation
0.274	13641	Recurrent	0.144	8269	New
0.263	3438	New	0.144	5105	Rare mutation
0.245	593	Rare mutation	0.143	15172	Q3
0.234	9140	New	0.142	11151	New
0.234	14502	New	0.141	10914	P4
0.234	4733	New	0.141	11288	P4
0.233	12346	New	0.140	5492	New
0.230	6755	New	0.140	8152	Rare mutation
0.229	13145	New	0.140	10933	Rare mutation
0.208	9812	B4a2	0.140	15204	New
0.207	8842	Rare mutation	0.137	15663	New
0.202	3645	P3	0.137	15317	New
0.202	15748	P3	0.137	10192	Rare mutation
0.202	13269	Rare mutation	0.134	12750	Rare mutation
0.196	4823	B4a2a	0.133	5177	M27a-Q3b
0.196	14338	P3	0.131	2768	Q3
0.192	5330	Rare mutation	0.130	14385	Rare mutation
0.191	6734	P3	0.129	8525	Rare mutation
0.189	14384	Rare mutation	0.129	14449	Rare mutation
0.189	12519	New	0.129	14954	Rare mutation
0.188	3351	Rare mutation	0.128	14290	Rare mutation
0.185	4335	Q3	0.127	13681	Q3b
0.183	15924	New	0.125	11963	Q3b
0.183	6620	B4a2	0.123	4769	Rare mutation
0.180	11016	P4-New	0.123	5894	Rare mutation
0.179	3394	Rare mutation	0.121	15043	M
0.179	6878	Rare mutation	0.121	4023	Rare mutation
0.179	14890	P2	0.115	1692	M27c
0.179	15443	Rare mutation	0.114	11992	Rare mutation
0.178	8859	P2	0.113	12366	Rare mutation
0.178	1375	New	0.112	9254	Q3a
0.174	3882	P2	0.110	13651	New
0.171	8572	New	0.107	6131	Rare mutation
0.170	13927	Rare mutation	0.103	4892	Rare mutation
0.170	3699	New	0.100	5090	Rare mutation
0.169	10400	M	0.100	9866	Rare mutation
0.169	14783	M			

[a]The "New" nucleotide position is a marker of a new predicted haplogroup (see Table 6). The "Rare mutation" was observed only in one or two haplotypes in HmtDB, without defining any specific haplogroup. The "Recurrent" mutation at that nucleotide position is paraphyletic.

nating variability values" of the starting hypothesis. MSD is defined as:

$$MSD_{i,k} = \sum_{j=1}^{5} (\gamma_i^k - \gamma_i^j)/4 \text{ for } j \neq k, \quad (4)$$

where $MSD_{i,k}$ is the discriminating value of the i-th site in continent k, γ_i^k indicates the variability value of the i-th site in continent k, and γ_i^j is the variability value of the other four continents (indicated as j, for $j \neq k$) in the same nucleotide position.

The MSD values—estimated for each site in each continent—are a measure of the degree of difference in variability values

TABLE 5. Mean Simple Deviation (MSD) of mtDNA Site Variability Values in America With Respect to Other Continents

MSD	Nucleotide position	Confirmed haplogroup(s)[a]	MSD	Nucleotide position	Confirmed haplogroup(s)[a]
0.959	3552	C	0.234	15301	L/M
0.926	7196	M8 (C)	0.228	4977	B2
0.926	15487	M8 (C)	0.228	6473	B2
0.723	5178	D (D1-D2)	0.226	11177	B2
0.721	663	A (A2)	0.226	3547	B2
0.721	1736	A (A2)	0.225	6260	Rare mutation
0.721	4248	A (A2)	0.214	6491	Rare mutation
0.721	8794	A (A2)	0.206	9950	B2
0.700	12007	A2	0.202	7697	New
0.693	4824	A (A2)	0.199	12978	Rare mutation
0.691	8027	A2	0.176	5054	Rare mutation
0.540	9545	C	0.176	13590	B4b
0.487	2092	D1	0.174	4970	Rare mutation
0.480	14318	C	0.173	6216	New
0.475	13263	C	0.173	13855	Rare mutation
0.462	4715	M8 (C)	0.168	11314	Rare mutation
0.439	8584	M8 (C)	0.168	6308	New
0.421	10400	M	0.167	6413	B4b1a
0.421	14783	M	0.163	961	Rare mutation
0.374	15043	M	0.161	6023	B4b1a
0.371	8414	D4 (D1-D2)	0.154	12317	Rare mutation
0.368	1888	New	0.148	15670	Rare mutation
0.363	4883	D (D1-D2)	0.144	12642	Rare mutation
0.353	14668	D4 (D1-D2)	0.144	3010	D4
0.349	7724	Rare mutation	0.139	11147	C1b
0.343	12468	New	0.139	9591	Rare mutation
0.341	4820	B4b	0.138	3316	D4e1
0.340	15535	B4b	0.134	9449	Rare mutation
0.335	827	B4b	0.131	14463	Rare mutation
0.308	14364	New	0.126	6261	Rare mutation
0.290	15930	New	0.124	12811	Rare mutation
0.285	11593	Rare mutation	0.118	11884	Rare mutation
0.283	7112	New	0.116	4315	Rare mutation
0.282	9540	L/M	0.111	15499	Rare mutation
0.282	10873	L/M	0.109	15439	Rare mutation
0.276	11914	C	0.106	10007	Rare mutation
0.274	8701	L/M	0.103	10398	M
0.234	15301	L/M	0.100	5964	Rare mutation

[a]The "New" nucleotide position is a marker of a new predicted haplogroup (see Table 6). A "Rare mutation" means that the mutation was observed only in one or two haplotypes in HmtDB, without defining any specific haplogroup.

necessary to identify a site as a defined continent associated site. Data with MSD values higher than a certain threshold are most probably haplogroup markers.

Here, analysis was performed on the coding region of 1,694 mtDNAs of different ethnic/geographical origin, all belonging to healthy subjects: 104 from Africa, 496 from Asia, 977 from Europe, 54 from Oceania, and 63 Native Americans. These sequences and their variability data are available in the HmtDB human mitochondrial genomic resource (www.hmdb.uniba.it) [Attimonelli et al., 2005]. To test if the inclusion of the D-loop in the analyzed sequences could affect site-specific variability value estimation in the coding part of the mitochondrial genome, we also applied our method to the 1,134 sequences for which both coding and D-loop regions were available. We did not observe any drastic change in the variability values of the coding region: it was simply generally lowered, because of the influence of the higher variability of the D-loop sites (data not shown).

Simulated data were generated randomly by selecting 100 different datasets for each continent [Attimonelli et al., 2005] on which site-specific nucleotide variability values and their mean and standard deviations, were estimated (data not shown, available in HmtDB by clicking on site variability values available in the genome card or downloading variability value tables through the HmtDB downloading function).

RESULTS AND DISCUSSION

We analyzed a total of 15,447 sites in the coding region, but only a fraction of these were informative, for two major reasons: first, only about 10% were variant (with a site-specific variability value greater than zero); second, MSD values lower than a certain value were not able to show already known continent-specific nucleotide positions as the corresponding site specific variability values were very similar in the various continents. This last observation highlighted the need to address the issue of threshold choice—i.e., the choice of the minimum MSD value to be considered significant enough to identify a potential continent-specific mtDNA mutation—and guided our choice in order to adapt it to the peculiar variability features of each geographic area. As shown in Figure 1, the number of sites with variability greater than zero is not constant in the five continents, as Asia and Europe are the most variable continents when the number of variant sites is considered (see below). However, as the great majority of Asian and European variant sites have very low variability values (see variability trend in Fig. 2), often being very similar to these of other continents, they are not sufficiently informative for our goal. If unique sites—those with variability values greater than zero only in one particular continent—are examined (histogram in Fig. 1), the situation improves slightly, in the sense that a certain number

TABLE 6. Novel Predicted Haplogroups on Each Continent

Haplogroup code	Nucleotide position(s)
Africa	
L0a2a1	9554
L1b1a	5393
L1b1a1	13980
L1c1a1	11257, 12930, 14034
L1c2a	6150
L2a1a	5285, 15244, 15629
L2a1d	5581, 15229
L2a1e	3495, 12630
L3b1	10373
L3b1a	11002
L3b1a1	11800
L3e1c	15670, 15942
L3e1c1	8650
L3e2a	13105
L5a1a	5656
U6a1	14179
Asia	
U7a	14569
D4a1	10410
D4b2b1	14605
D4b2b2	12358
D4g1	4343, 15518
G1a1a	827
R5a	13635
Europe	
H1a1	6365
H1a2	8271
H1c1	9150
H4a1	10044
H7a	1719
H16	10394
U5a1a1	5495, 15924
Oceania	
P1a1	3699, 8269, 12346, 12879
P2a	1375, 8572
P3a	13651
N23	6755, 9140
N23a	5460, 5563, 10700, 15300, 15852, 15885, 15924
S2	2380, 3438, 6167
S2a	14502
S3	5302, 5492, 9938, 10786, 11151, 15204, 15317, 15663, 15664
B4a1a2	4733, 12519
B4a1a1a	6905
M42a	11016, 13145
America	
A2a	6308
A2b	7112
A2c	12468, 14364
B4b1a2	6216
C1a	7697
C1c	1888, 15930

of sites certainly not involved in continent-specific site variability can be ignored, but there are still many of sites whose variability is not sufficiently "discriminating." Overall, when deciding the MSD threshold, it is necessary to take into account the general variability trend of the continent in question, thus avoiding over- or underestimation of the number of continent-specific nucleotide positions. For this purpose, we decided to fix the threshold value by taking into account the general variability trend observed in each continent. The Asian case is explicative. Variability values in Asia are generally low (although numerous) and only a very sites have a high MSD value, so that, if a given prefixed threshold is chosen, only a few will exceed it. This would obviously lead to underestimation of the number of continent-specific sites, missing the large majority of them. As a guide for fixing the threshold in the case of the Asian-specific variability trend we used the published haplogroup classification (and thus the number of already defined haplogroup-specific sites). On this basis, a threshold value of 0.05 was chosen for Asian and European data and 0.1 for the other continents.

The results are shown in five continent-specific tables (Tables 1–5), built using MSDgen script. This script was developed in our laboratory and, after estimating MSD values, allows sorting of nucleotide sites by their MSD value, and thus by their capacity to discriminate one geographic area from the rest of the world. After application of the MSDgen script, the resulting sites are searched in the haplogroup classification of HmtDB; if the search produces positive results, the site is assigned to the corresponding most suitable haplogroup.

The continent tables show that those sites with the highest MSD values generally match the already known continent-specific mtDNA haplogroup defining sites [Finnilä et al., 2001; Herrnstadt et al., 2002; Ingman and Gyllensten, 2003; Kong et al., 2003; Achilli et al., 2004, 2005; Palanichamy et al., 2004; Friedlaender et al., 2005; Kivisild et al., 2004, 2006; Macaulay et al., 2005; Merriwether et al., 2005; Salas et al., 2004; Thangaraj et al., 2005; Trejaut et al., 2005]. This finding is supported by the observation that "real" and "simulated" variability values are generally comparable, indicating that the quality of the dataset is good enough to obtain statistically significant results. Consequently, we may assume that the frequency of substitutions and the final result have not been influenced. For instance, Table 2 shows the distribution of haplogroups in Africa, obtained by the application of MSDgen script. Of the 188 sites classified in the analysis (stopping at an MSD value of 0.1), 165 mark known African-specific haplogroups, and 23 were defined as "novel" because they have not yet been reported in the literature as haplogroup-defining sites. Evaluation of the five continents revealed a total of 81 variant nucleotide positions (23 in Africa, eight in Asia, eight in Europe, 34 in Oceania, and eight in America) with a high MSD value, despite the fact that they were not reported in the literature as haplogroup-defining sites. Therefore, they represent very good candidate markers of novel mtDNA haplogroups or subhaplogroups (Table 6). It is worth pointing out that all the new haplogroup markers detected in this study had not been previously revealed by a classic phylogeographic approach [Finnilä et al., 2001; Herrnstadt et al., 2002; Ingman and Gyllensten, 2003; Kong et al., 2003; Achilli et al., 2004, 2005; Palanichamy et al., 2004; Friedlaender et al., 2005; Kivisild et al., 2004, 2006; Macaulay et al., 2005; Merriwether et al., 2005; Salas et al., 2004; Thangaraj et al., 2005; Trejaut et al., 2005], although the analyzed sequence datasets were the same. In fact, the possibility of analyzing such a great number of genomes all together is one of the intrinsic advantages of our approach (and thus of this study), revealing information which would otherwise be hidden in the data. Obviously, the new predicted sites were less abundant in those continents where mtDNA phylogeny has recently been studied in greater detail, such as Asia [Kong et al., 2003; Tanaka et al., 2004; Palanichamy et al., 2004] and Europe [Herrnstadt et al., 2002; Achilli et al., 2004, 2005; Palanichamy et al., 2004]. In addition, taking into account the shared patterns of mutations observed in the available complete sequences of these new 81 sites and the already defined phylogeny reported in the literature, we propose that these sites define a total of 47 mtDNA haplogroups (16 in Africa, seven in Asia, seven in Europe, 11 in Oceania, and six in

America) which have not been previously identified by phylogeographic studies (Table 6).

In conclusion, this study provides a new method for discriminating geographic-specific patterns of mtDNA mutations by employing nucleotide continent-specific variability values and without having to examine any reference sequence. We found an extremely good fit between our results and the mtDNA haplogroup classification currently reported in the literature. Our algorithm was able to identify totally about half all the known haplogroups, although the current literature classification is the result of years of research on a great variety of samples. This supports the reliability of prediction method in recognizing geographically defined patterns of mutations, and validates the identification of a large number of new variant sites which are most probably markers of mtDNA haplogroups not yet identified by phylogeographic studies. This information not only allows refinement of mtDNA phylogeny, an essential requirement for mitochondrial disease studies [Bandelt et al., 2005; Salas et al., 2005], but also facilitates discrimination of candidate disease-causing mutations from haplogroup-specific polymorphisms in mtDNAs from patients affected by mitochondrial diseases.

ACKNOWLEDGMENTS

We thank Dr. David Horner (Dipartmento di Scienze Biomolecolari e Biotecnologie, Università di Milano) for his advice and support in the production of the new version of the Site_Var algorithm. This work was supported by the Ministero Italiano dell'Università e Ricerca: Fondo Investimenti Ricerca di Base 2001, "Bioinformatica per la Genomica e la Proteomica"; Progetti Ricerca Interesse Nazionale 2003 (to M.T.-P.) and 2005 (to A.T.), and Fondazione Cariplo (to A.T.).

REFERENCES

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogväli EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A. 2004. The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910–918.

Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, Magri C, Scozzari R, Babudri N, Santachiara-Benerecetti AS, Bandelt HJ, Semino O, Torroni A. 2005. Saami and Berbers—an unexpected mitochondrial DNA link. Am J Hum Genet 76: 883–886.

Attimonelli M, Accetturo M, Santamaria M, Lascaro D, Scioscia G, Pappadà G, Russo L, Zanchetta L, Tommaseo-Ponzetta M. 2005. HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. BMC Bioinformatics 6:S4.

Bandelt HJ, Achilli A, Kong QP, Salas A, Lutz-Bonengel S, Sun C, Zhang YP, Torroni A, Yao YG. 2005. Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. Biochem Biophys Res Commun 333:122–130.

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

Carelli V, Achilli A, Valentino ML, Rengo C, Semino O, Pala M, Olivieri A, Mattiazzi M, Pallotti F, Carrara F, Zeviani M, Leuzzi V, Carducci C, Valle G, Simionati B, Mendieta L, Salomao S, Belfort R, Sadun AA, Torroni A. 2006. Haplogroup effects and recombination of mitochondrial DNA: novel clues from the analysis of Leber hereditary optic neuropathy pedigrees. Am J Hum Genet 78:564–574.

Denaro M, Blanc H, Johnson MJ, Chen KH, Wilmsen E, Cavalli-Sforza LL, Wallace DC. 1981. Ethnic variation in Hpa 1 endonuclease cleavage patterns of human mitochondrial DNA. Proc Natl Acad Sci USA 78:5768–5772.

Finnilä S, Lehtonen MS, Majamaa K. 2001. Phylogenetic network for European mtDNA. Am J Hum Genet 68:1475–1484.

Forster P, Matsumura S. 2005. Evolution. Did early humans go north or south? Science 308:965–966.

Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, Deka R, Yanagihara R, Merriwether DA. 2005. Expanding Southwest Pacific mitochondrial haplogroups P and Q. Mol Biol Evol 22: 1506–1517.

Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, Ghosh SS, Olefsky JM, Beal MF, Davis RE, Howell N. 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152–1171 [Erratum in Am J Hum Genet 2002;71: 448–449].

Howell N, Elson JL, Chinnery PF, Turnbull DM. 2005. mtDNA mutations and common neurodegenerative disorders. Trends Genet 11:583–586.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408:708–713 [Erratum in Nature 2001;410:611].

Ingman M, Gyllensten U. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res 13:1600–1606.

Johnson MJ, Wallace DC, Ferris SD, Rattazzi MC, Cavalli-Sforza LL. 1983. Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. J Mol Evol 19:255–271.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120.

Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R. 2004. Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 75:752–770.

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis KK, Passarino G, Underhill PA, Scharfe C, Torroni A, Scozzari R, Modiano D, Coppa A, de Knjiff P, Feldman MW, Cavalli Sforza LL, Oefner PJ. 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172:373–387.

Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP. 2003. Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 73:671–676 [Erratum in Am J Hum Genet 2004;75:157].

Lanave C, Preparata G, Saccone C, Serio G. 1984. A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A. 1999. The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232–249.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D,

Bandelt HJ, Oppenheimer S, Torroni A, Richards M. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–1036.

Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, Koki G, Friedlaender JS. 2005. Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proc Natl Acad Sci USA 102:13034–13039.

Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP. 2004. Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75:966–978.

Pesole G, Saccone C. 2001. A novel method for estimating substitution rate variation among sites in a large dataset of homologous DNA sequences. Genetics 157:859–865.

Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Qasim Mehdi S, Torroni A, McElreavey K. 2004. Where West meets East: the complex mtDNA landscape of the Southwest and Central Asian corridor. Am J Hum Genet 74: 827–845.

Richards M, Oppenheimer S, Sykes B. 1998. MtDNA suggests Polynesian origins in Eastern Indonesia. Am J Hum Genet 63: 1234–1236.

Richards M, Macaulay V, Hickey E Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ. 2000. Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251–1276.

Saccone C, Lanave C, Pesole G, Preparata G. 1990. Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol 183:570–583.

Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74:454–465.

Salas A, Yao YG, Macaulay V, Vega A, Carracedo A, Bandelt HJ. 2005. A critical reassessment of the role of mitochondria in tumorigenesis. PLoS Med 2:296.

Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (and other methods) 4.0 beta, Available at: www.sinauer.com/detail.php?id = 8060. Last accessed: 21 May 2006.

Tanaka M, Cabrera VM, Gonzalez AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, Shinoda K, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H. 2004. Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14:1832–1850.

Thangaraj K, Chaubey G, Kivisild T, Reddy AG, Singh VK, Rasalkar AA, Singh L. 2005. Reconstructing the origin of Andaman Islanders. Science 308:996.

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

Torroni A, Sukernik RI, Schurr TG, Starikorskaya YB, Cabell MF, Crawford MH, Comuzzie AG, Wallace DC. 1993b. MtDNA variation of Aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet 53:591–608.

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

Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu CJ, Li ZY, Lin M. 2005. Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biol 3:247.

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

Wallace DC. 2005. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. Annu Rev Genet 39:359–407.