D700–D704 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj030

HvrBase++: a phylogenetic database for primate species

Jochen Kohl, Ingo Paulsen, Thomas Laubach, Achim Radtke and Arndt von Haeseler[1,2,3,4,*]

Heinrich-Heine-University Duesseldorf, Universitaetsstrasse 1, 40225 Duesseldorf, Germany, [1]Max F. Perutz Laboratories, University of Vienna, Center for Integrative Bioinformatics Vienna, Dr.-Bohr-Gasse 9/6, A-1030 Vienna, Austria, [2]University of Vienna, [3]Medical University of Vienna and [4]University of Veterinary Medicine Vienna, Vienna, Austria

Received August 15, 2005; Revised and Accepted September 22, 2005

ABSTRACT

HvrBase++ is the improved and extended version of HvrBase. Extensions are made by adding more population-based sequence samples from all primates including humans. The current collection comprises 13 873 hypervariable region I (HVRI) sequences and 4940 hypervariable region II (HVRII) sequences. In addition, we included 1376 complete mitochondrial genomes, 205 sequences from X-chromosomal loci and 202 sequences from autosomal chromosomes 1, 8, 11 and 16. In order to reduce the introduction of erroneous data into HvrBase++, we have developed a procedure that monitors GenBank for new versions of the current data in HvrBase++ and automatically updates the collection if necessary. For the stored sequences, supplementary information such as geographic origin, population affiliation and language of the sequence donor can be retrieved. HvrBase++ is Oracle based and easily accessible by a web interface (http://www.hvrbase.org). As a new key feature, HvrBase++ provides an interactive graphical tool to easily access data from dynamically created geographical maps.

INTRODUCTION

HvrBase was originally started as compilation of hypervariable region I (HVRI) and hypervariable region II (HVRII) mitochondrial sequences (1,2). These regions are situated in the non-coding mitochondrial control region and play an important role in population genetics (3,4). With some exceptions, mitochondrial DNA (mtDNA) follows a maternal clonal inheritance pattern without recombination (5–7). Therefore, population genetics analyses allow studying the population history of maternal inherited mitochondrial genomes. Furthermore, mtDNA variation correlates with the geographic origin of the population, and has been linked to a wide range of degenerative diseases, preferentially affecting the central nervous system, heart, muscle, renal and endocrine systems, and is generally used in forensic comparisons (8–10).

HvrBase++ focuses on aspects of population genetics and collects meta information like ethnic groups and spoken languages for each individual. For this reason, sequences have only been included if a minimum of meta information was available. Moreover, meta information is linked to a geographical information system (GIS), which allows intuitive searches supported by geographical maps to obtain additional information about countries. This interactive map searching feature and the presence of meta information predestine HvrBase++ as a database for population genetics analysis.

Among other databases, like MITOMAP (11) as a general resource for mtDNA-related data and the 'mtDNA Population Database' (12) for forensic studies, HvrBase++ contributes to the wide area of mtDNA analysis.

Wherever in the course of a phylogenetic analysis mitochondrial data are used which at best reflect matrilineal history, a closer look at nuclear DNA (nDNA) is indispensable to answer questions concerning phylogenetic history in their entirety. When drawing a comparison between evolutionary pathways of the pyruvate dehydrogenase E1α (PDHA1) subunit and mtDNA, J. Hey showed that 'variation at nuclear genes and mtDNA are not both consistent with a common demographic history' (13).

While both hypervariable regions of mtDNA are commonly used for phylogenetic studies, no equivalent sequence markers exist when dealing with nuclear DNA. With HvrBase++, we introduce a set of ready-to-use nDNA sequence markers. Genes that code for the human immune defence, and non-coding regions around microsatellite DNA markers, are promising candidates for nDNA sequence markers owing to their mutation rate (3).

*To whom correspondence should be addressed. Tel: +43 1 4277 24007; Email: arndt.von.haeseler@univie.ac.at

Moreover, the detection of nuclear mitochondrial-like sequences called 'numts' in the last decade has shed doubts whether mtDNA data have been classified correctly (14–16). This is caused by a maximum of 94% similarity between mtDNA and numts fragments.

To meet these concerns, researchers have begun to incorporate nuclear markers in their studies. It is a matter of fact that HvrBase++ now carries nuclear markers as well.

Compilation of sequences

In HvrBase++ the term 'sequence' represents a piece of DNA from one individual. A 'lineage' in contrast means a piece of DNA from possibly different individuals, which share the same nucleotide sequence. Meta information about the sampled individuals was collected from publications (supplemented information). If different sequence sources were available, they have been chosen in the following order: (i) public databases like GenBank (17), (ii) supplemental data from publications, (iii) data manually extracted from publications and (iv) data requested from authors.

After collecting and extracting sequences and meta information for a gene or region, a global nucleotide alignment was created. For the HVRI and HVRII regions, HvrBase++ carries a manual alignment (2) and an alignment generated with MAFFT (18). Automatically calculated alignments can be obtained from HvrBase++ for complete mitochondrial genomes and nuclear sequences, respectively. A procedure checks sequence alterations in GenBank and updates the data in the next release. It is worth noting that every single update step is logged in our database system and can be traced via the HvrBase++ web interface.

Sequences enter HvrBase++ if meta information can be retrieved from the corresponding publication. Meta information must be attributable to each sequence in the paper and consider: (i) geographic origin, (ii) population, (iii) spoken language and (iv) bibliographic information. Owing to those filtering criteria, not all data from publications and most of the forensic data, for example the comprehensive forensic dataset from 'mtDNA Population Database' (12), could not be integrated into HvrBase++.

Since there is no unique way to gain the above named meta information either from publications or from sequence files, it is difficult to build a fully automated tool that identifies meta information that is located in different resources.

Synonyms and context-dependent meanings of a word may pose a challenge as well. Where it is easy for humans to associate certain information, it is a hard task for computers. Seeing that, HvrBase++ categorizes ambiguous data to facilitate a broad range of complex search patterns. Bibliographic information, like authors, publication date, journal and PubMed publication identifier are standardized. Each country is assigned to just one continent, e.g. in HvrBase++ Turkey is assigned to Asia, the Canary Islands belong to the sovereign territory of Spain.

All 258 language entries in HvrBase++ have been adapted to comply with the SIL (Summer Institute of Linguistics) and ISO/DIS 639-2 language code standards respectively from Ethnologue vol. 14 (19). In order to avoid information loss and to compensate the incompleteness of any of the standards, it was necessary to integrate both language codes (Table 1). The following example clearly shows the hassle of associating a mother tongue of an individual deduced from a publication with the SIL or ISO language codes.

Table 1. Assignment of language names to the SIL and ISO/DIS 639-2 codes in HvrBase++ for the mitochondrial dataset

SIL	ISO	No. of individuals	Language family or population
Yes	Yes	7248	English
Yes	No	41	Mandenka (population from Senegal, 'Mandinka' in SIL)
No	Yes	1951	Bantu (Africa's largest language family)
No	No	454	Mbenzele (population from Central African Republic)
		4611	Language information missing or not assignable
Total		14 305	

This year, the SIL and ISO/DIS 639-2 codes have converged. We will account for them in the next major release.

Table 2. Number of sequence categories in humans, great apes and Neanderthalers across all sequence types in HvrBase++

	Number of Humans	Great Apes	Neanderthalers
HVRI	13 350	520	3
HVRII	4925	13	2
Mitochondrial genomes	1376	0	0
Nuclear sequences	386	21	0
Total	20 037	554	5

It is known that a certain tongue belongs to the Niger-Kordofanian language family. Niger-Kordofanian is a collective language code only used in the ISO standard whose languages can be found throughout Southern and Central Africa as well as in Sub-Saharan Western Africa. Since that language family does not have a SIL code, a more in-depth knowledge about the very tongue (e.g. language name and habitation of a tribe) would be essential to find a suitable SIL code.

Technical organization

HvrBase++ is managed in an Oracle 10g relational database system. Sequence data and accompanying information are extracted and stored in HvrBase++ via Perl programs that use object-oriented modules from the BioPerl-Project (20) and the Perl DBI module.

The web client is based on the Apache web server technology. For the geographical interface, a map server (MapServer version 4.6 from the University of Minnesota) is integrated into the web client using geographical maps from publicly available resources.

UMN MapServer provides the core functionality of a GIS system for an intuitive data access from dynamically created geographical maps.

Description of the compilation

The HvrBase++ database now comprises not only HVRI and HVRII sequences but also mitochondrial genomes and nuclear sequences from several chromosomal loci. Not surprisingly, human sequences are overrepresented with a total amount of 20 037 sequences (Table 2). Table 3 displays an excerpt

Table 3. Human HVRI datasets over six continents

Continent	Lineages	Human samples	Number of Countries	References	Populations	Languages	SIL	ISO
Europe	2033	4358	17	39	25	31	20	16
Africa	1046	1680	25	22	47	47	22	25
North America	824	1581	7	19	34	9	9	8
South America	267	473	7	10	11	19	7	7
Asia	2867	4778	23	49	102	67	31	47
Australia/Oceania	224	473	10	10	12	28	9	16
World	7036	13 343	89	103	220	194	81	118

Note that the last row does not depict the arithmetic sum in columns 2, 5–9 as some relevant subsets overlap across continents.

Table 4. Location, length and amount of the 407 nuclear sequences in HvrBase++

Amount	Gene	Gene function	Chromosome	Length in bp
8	pdh1	Pyruvate dehydrogenase E1-α subunit gene, partial seq.	X	1769
41	Factor ix	Factor IX gene, intron 4	X	3740
42	rrm2p4	Ribonucleotide reductase M2 pseudogene 4, partial seq.	X	2392
42	tnfsf5	Tumor necrosis factor ligand superfamily 5 gene, partial seq.	X	5239
1	amelx	Amelogenin X chromosome gene, complete seq.	X	5323
71	xq13.3	Xq13.3 non-coding region	X	10 178
56	mc1r promoter	Melanocortin 1 receptor gene, promoter	16	6599
1	mc1r	Melanocortin 1 receptor gene	16	953
8	lpl	Lipoprotein lipase gene, partial seq.	8	542–1636
61	ch1	Membrane protein CH1 gene, partial seq.	1	9626
59	β-globin	β-globin gene, complete seq.	11	3008
17	β-globin repl. init. reg.	β-globin gene, repl. ori. init. reg. and partial seq.	11	1312

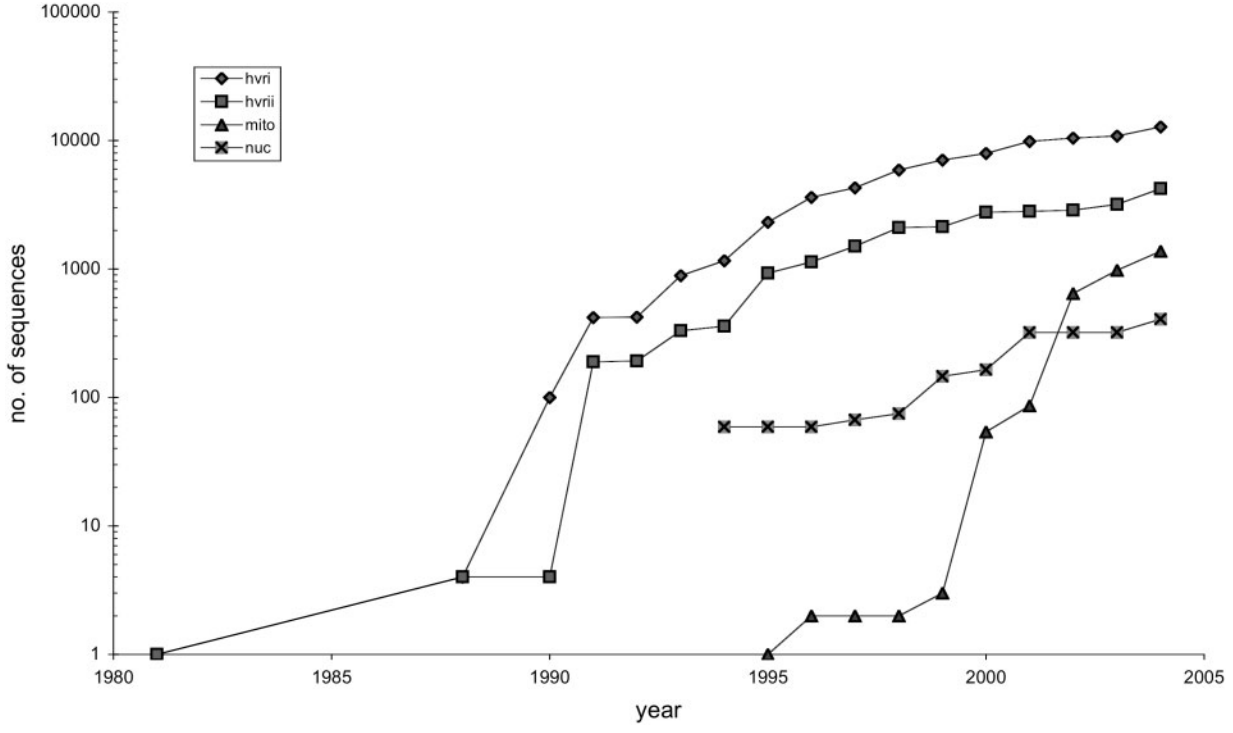
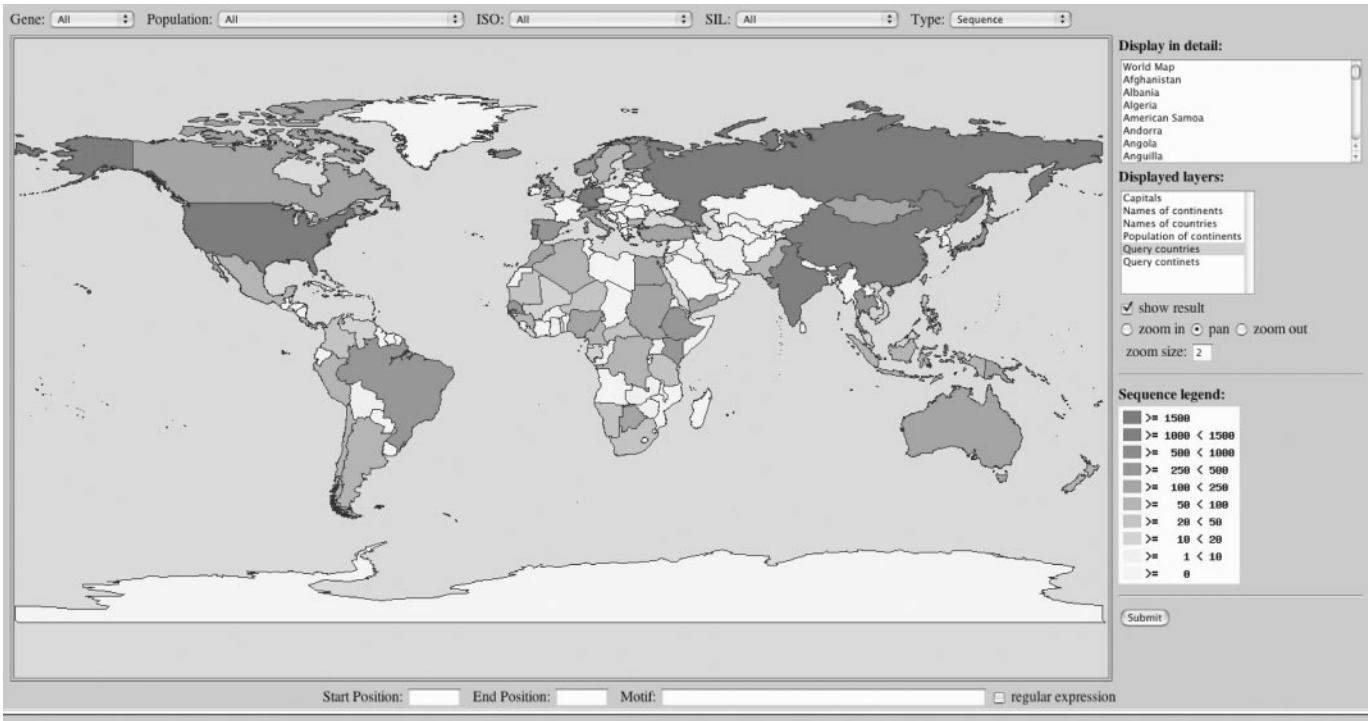


Figure 1. Accumulation of HVRI, HVRII, mitochondrial genomes and nuclear sequences over the last 25 years.

from the human HVRI dataset gathered from 103 publications which encompasses sequences from 89 countries and 220 ethnic groups.

Table 4 describes the 10 loci for the 407 nuclear sequences. The amount of nuclear markers in HvrBase++ is currently not very high because the compilation is at an early stage. We feel confident that it will get more and more important to sequence and analyze nuclear genes for studies in population genetics due to possibly contradicting histories of nuclear genes and mtDNA (3,21). Figure 1 shows the sequence increase of



357 Sequences for Brazil

Genes	Number of				
	Sequences	Populations	Language sources		
			Publications	ISO-639-2	SIL vol.14
tnfsf5	4	2	0	0	2
rrm2p4	4	2	0	0	2
hvri	299	8	5	3	4
hvrii	50	1	1	0	0

Figure 2. Geographical map interface in HvrBase++. The upper frame contains elements for searching sequences, the search results are displayed in the map and at the bottom. A countries' colour represents the number of sequences for a given gene. The main table shows the results of all available genes for a selected country. Additional information for each gene is displayed in separate tables (data not shown). Sequences are accessible by selecting them from the table.

HVRI/II, mitochondrial genomes and nuclear sequences for all available publications within the past 25 years. Thus, it can be assumed that this upward trend will continue.

User interface

The new geographic map search interface is the centre of the web interface, which provides an intuitive search method and presents the results clearly structured. On the other hand, the well-tried form-based search function from HvrBase is recommended for more systematic searches. Supported sequence output formats are Phylip, GenBank, XML and simple text files. The form-based and map searches in combination make it possible to find any kind of sequence available from a sampled individual.

Figure 2 shows the geographic map functionality in HvrBase++. It is possible to search for all genes in countries and continents. A more sophisticated search can be obtained by specifying populations and language codes. Sequence patterns can be detected within genes for a whole sequence or a given range. Moreover, regular expressions allow for complex motif searches. Each country (or continent) is pictured in the world map and colour-coded, depending on the number of sequences from the respective country. More detailed information is displayed at the bottom of the world map after choosing a country from the map.

Quality and completeness of the data and future directions

Although HvrBase++ represents a large compilation of HVRI and HVRII sequences, completeness cannot be claimed. The collection of mitochondrial genomes and nuclear genes will be extended, and gaps will have to be closed in future releases.

Therefore, we solicit everybody to furnish new sequences and respective information by electronic mail. We would also be grateful to receive already published sequences that are missing in our collection.

This database gives easy access to freely available sequences without altering them in any way. That means

we have not checked the data for typos or any other kind of sequence errors that might have occurred between their acquisition and their publication (22–24). Our intention is not to fix putative errors in other publications and finally to hold in our hand another dataset. This could cause confusion by the use of sequences in comparative analyses from two different sources.

We recommend our colleagues to control their datasets carefully and to follow the instructions proposed by Bandelt *et al*. (25) to detect suspicious sequence positions.

New sequence versions in GenBank are investigated automatically and continuously included into HvrBase++. Since there is no common way to update sequences from non-public databases, we have done this manually. As we aim at a high quality of data, we will welcome any cues regarding programming bugs, misinterpretations or other discrepancies.

ACKNOWLEDGEMENTS

We thank all colleagues who have provided their sequence data in computer readable format and have given us additional information when needed. Funding to pay the Open Access publication charges for this article was provided by the Deutsche Forchungsgemeinschaft (DFG).

Conflict of interest statement. None declared.

REFERENCES

1. Handt,O., Meyer,S. and von Haeseler,A. (1998) Compilation of human mtDNA control region sequences. *Nucleic Acids Res*., **26**, 126–129.
2. Burckhardt,F., von Haeseler,A. and Meyer,S. (1999) HvrBase: compilation of mtDNA control region sequences from primates. *Nucleic Acids Res*., **27**, 138–142.
3. Zhang,D.X. and Hewitt,G.M. (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Mol. Ecol*., **12**, 563–584.
4. Avise,J.C. (1998) The history and purview of phylogeography: a personal reflection. *Mol. Ecol*., **7**, 371–379.
5. Kondo,R., Satta,Y., Matsuura,E.T, Ishiwa,H., Takahata,N. and Chigusa,S.I. (1990) Incomplete maternal transmission of mitochondrial DNA in *Drosophila*. *Genetics*, **126**, 657–663.
6. Gyllensten,U., Wharton,D., Josefsson,A. and Wilson,A.C. (1991) Paternal inheritance of mitochondrial DNA in mice. *Nature*, **352**, 255–257.
7. Skibinski,D.O.F., Gallagher,C. and Beynon,C.M. (1994) Mitochondrial DNA inheritance. *Nature*, **368**, 817–818.
8. Coskun,P.E., Beal,M.F. and Wallace,D.C. (2004) Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. *Proc. Natl Acad. Sci. USA*, **29**, 10726–10731.
9. Sukernik,R.I., Derbebeva,O.A., Starikovskaya,E.B., Volodko,N.V., Mikhailovskaya,I.E., Buychov,I.Yu., Lott,M., Brown,M. and Wallace,D. (2002) The mitochondrial genome and humans mitochondrial diseases. *Russ. J. Genet*., **38**, 161–170.
10. Budowle,B., Allard,M.W., Wilson,M.R. and Chakraborty,R. (2003) Forensics and mitochondrial DNA: applications, debates, and foundations. *Annu. Rev. Genomics Hum. Genet*., **4**, 119–141.
11. Brandon,M.C., Lott,M.T., Nguyen,K.C., Spolim,S., Navanthe,S.B., Baldi,P. and Wallace,D.C. (2004) MITOMAP: a humans mitochondrial genome database—2004 update. *Nucleic Acids Res*., **33**, D611–D613.
12. Monson,K.L., Miler,K.W.P., Wilson,M.R., DiZinno,J.A. and Budowle,B. (2002) The mtDNA population database: an integrated software and database resource for forensic comparison. *Forensic Sci. Commun*., **4**. Available at http://www.fbi.gov/hqlab/fsc/backissu/april2002/miller1.htm.
13. Hey,J. (1997) Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol*., **14**, 166–172.
14. Mishmar,D., Ruiz-Pesini,E., Brandon,M. and Wallace,D.C. (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Hum. Mutat*., **23**, 125–133.
15. Bensasson,D., Zhang,D.X., Hartl,D.L. and Hewitt,G.M. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol*., **16**, 314–321.
16. Thalman,O., Hebler,J., Poinar,H.N., Pääbo,S. and Vigilant,L. (2004) Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. *Mol. Ecol*., **13**, 321–335.
17. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res*., **32**, D23–D26.
18. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*., **33**, 511–518.
19. Grimes,B.F. (2000) In Grimes,B.F. (ed.), Ethnologue: Volume 1 Languages of the World, (14th Edn) ISBN 1-55671-103-4.
20. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I., Lapp,H. *et al*. (2002) The Bioperl Toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1161–1168.
21. Petit,R.J., Duminil,J., Fineschi,S., Hampe,A., Salvini,D. and Vendramin,G.G. (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. *Mol. Ecol*., **14**, 689–701.
22. Bandelt,H.-J., Quintana-Murci,L.L., Salas,A. and Macaulay,V. (2002) The fingerprint of phantom mutations in mitochondrial DNA data. *Am. J. Hum. Genet*., **71**, 1150–1160.
23. Herrnstadt,C., Preston,G. and Howell,N. (2003) Errors, phantom and otherwise, in humans mtDNA sequences. *Am. J. Hum. Genet*., **72**, 1585–1586.
24. Forster,P. (2003) To err is human. *Ann. Hum. Genet*., **67**, 2–4.
25. Bandelt,H.-J., Lahermo,P., Richards,M. and Macaulay,V. (2001) Detecting errors in mtDNA data by phylogenetic analysis. *Int. J. Legal Med*., **115**, 64–69.