126–129 Nucleic Acids Research, 1998, Vol. 26, No. 1

Compilation of human mtDNA control region sequences

Oliva Handt+, Sonja Meyer and Arndt von Haeseler*

Zoologisches Institut, Universität München, Luisenstraße 14, D-80333 München, Germany

Received September 29, 1997; Revised and Accepted October 24, 1997

ABSTRACT

This paper describes the organisation of a database for human mitochondrial control-region sequences. The data are divided into three ASCII files that contain aligned sequences from the hypervariable region I (HVRI), from the hypervariable region II (HVRII), and the available information about the individuals, from whom the sequences stem. The current collection comprises 4079 HVRI and 969 HVRII sequences. From 728 individuals sequences of both HVRI and HVRII are available. For easy access, the collection is made available to the scientific community via World Wide Web at URL http://www.zi.biologie.uni-muenchen.de/~meyers/mtdna.html

INTRODUCTION

The history of human populations is studied for a wealth of different genetic systems (1–4). Because the mitochondrial genome is maternally inherited and accumulates substitutions at a higher rate than the nuclear genome, it is well suited to analyse the population history of humans based on simple models of population history. Especially the hypervariable regions HVRI and HVRII (5) of the control region have been studied extensively (cf. 6 and references therein). Since 1981 the amount of available HVRI and HVRII data has increased exponentially (Fig. 1). We have collected and aligned a large number of control-region sequences. This paper describes the organisation of the database.

COMPILATION OF SEQUENCES

Sequences were collected from publications (7–40) or were retrieved from GenBank (41) and stored as plain ASCII files. Sequences from GenBank were compared to the sequences in the corresponding publications. If discrepancies occurred the sequences were stored as given in the paper. If only sequence positions deviating from the reference sequence (7) were published these deviations were added to the reference sequence and the resulting sequence was stored. When the publication did not clearly state the start and end of a sequence, the first, respectively the last variable sites were used as limitation. Unfortunately, it was not always evident how often each lineage was found or to which population it belonged when individuals of more than one population were studied. If this could not be unraveled the data were not added to the collection.

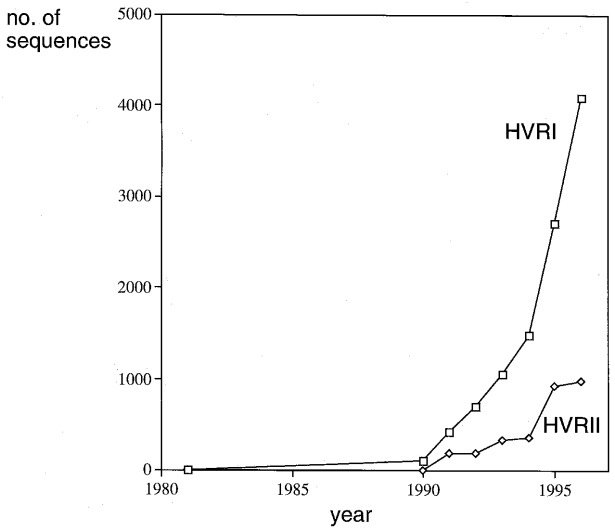

Figure 1. Accumulation of HVRI and HVRII sequences during the last 15 years.

Sequences were manually aligned. For the HVRI region we aligned positions 16001–16408 and for the HVRII region positions 1–408 were aligned (7). If sequences were longer than this alignment, they were truncated to the corresponding sites, if they were shorter, question marks were introduced to achieve the length required by the alignment. All non-determined nucleotides within a sequence are also represented by question marks. A dash (–) indicates an insertion or deletion of a nucleotide.

ORGANIZATION OF COLLECTED DATA

The data are divided into an information file (info12.txt) and two sequence files (alld1.txt, alld2.txt). To reduce the amount of storage the sequence files contain only '(database)-lineages', which differ in at least one position from the remaining entries of the collection. The file info12.txt contains available information

* To whom correspondence should be addressed. Tel: +49 89 5902 327; Fax: +49 89 5902 474; Email: arndt@zi.biologie.uni-muenchen.de

+Present address: Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, 72 King William Road, North Adelaide SA 5006, Australia

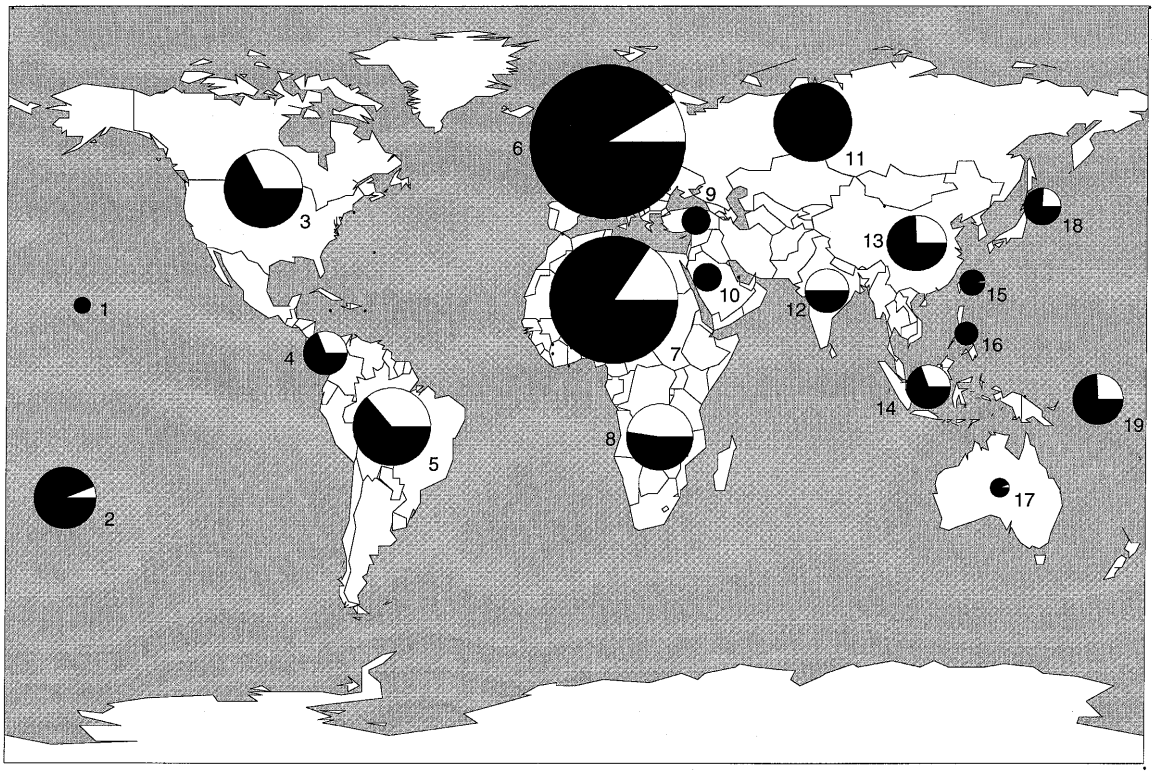

Figure 2. Worldwide distribution of mitochondrial hypervariable region I and II sequences. The size of the circles reflects the proportion of individuals from the corresponding region in the collection. The frequencies of the hypervariable region I are shown as solid portions of the pie charts. The approximate origin of the individuals are: 1, Hawaii; 2, Polynesia; 3, North America; 4, Panama; 5, South America; 6, Europe; 7, North Africa; 8, South Africa; 9, Turkey; 10, Jordania; 11, Russia; 12, India; 13, China; 14, Indonesia & Malaysia; 15, Taiwan; 16, Phillipines; 17, Australia; 18, Japan; 19, Micronesia & Melanesia.

about the individuals. Currently the following categories are defined:

- I: <number>. This number specifies the HVRI lineage found in the individual. The corresponding sequence in alld1.txt has the same number. A zero indicates that HVRI was not sequenced for that individual.
- II: <number>. This number refers to the corresponding HVRII lineage in alld2.txt.
- Continent the individual stems from. The following abbreviations are used: AFRI, Africa; AMER, Americas; ASIA, Asia; A/OC, Australia & Oceania; EURO, Europe.
- N: specifies the name of the sequence in the original publication or the GenBank accession number.
- R: gives the original reference.
- O: shows the country of origin.
- P: gives the population the individual belongs to.
- L: gives the language and the language phylum of the individual.
- +9bpdel/–9bpdel indicates the presence or absence of the 9 bp deletion (42).

The file alld1.txt contains the alignment of HVRI lineages. Each lineage in the file is indexed by a number. If an individual from info12.txt has the same number, the corresponding sequence was found in that individual. The file alld2.txt is organised as alld1.txt. It comprises the alignment of the HVRII.

Program

A C-program, that should run on most computers, allows the retrieval of all individual sequences that match a user defined keyword in the information file. The search results are stored in four files: kw-info contains the information about the individuals that match the keyword. In kw-I and kw-II the HVRI and HVRII sequences of the individuals are given and the file kw-I-II contains the sequences of the individuals where both variable regions have been sequenced.

DESCRIPTION OF THE COMPILATION

The current collection comprises 4079 HVRI, 969 HVRII, and 728 human sequences where HVRI and HVRII are known. This amounts to 2298 and 580 (database)-lineages for HVRI and HVRII, repectively. 539 lineages are found among individuals where both HVRI and HVRII have been determined. These numbers also include some unpublished sequences [K.Bauer, H.Geisert, M.Krings, M.Laan, A.Salem, A.Sajantila and S.Pääbo (1997), manuscript in preparation], that will be made available as soon as they are in press.

Table 1. Number of sequenced individuals and lineages for the five continents

Region		Africa	Europe	Americas	Asia	Australia & Oceania
	sequences	878	1330	617	914	340
HVRI	lineages	557	802	255	616	143
	sequences	321	204	166	220	58
HVRII	lineages	183	108	72	175	52
HVRI &	sequences	247	114	166	143	58
HVRII	lineages	139	108	92	143	54

Table 2. Collection of sequences according to language phyla

Language phylum		HVRI	HVRII	HVRI + II
1	Khoisan	43	57	43
2	Niger-Kordofanian	318	249	197
3	Nilo-Saharan	208	5	0
4	Afro-Asiatic	366	0	0
5	Caucasian	0	0	0
6	Indo-Hitite	540	100	100
7	Uralic-Yukaghir	574	0	0
8	Altaic	14	0	0
9	Chuchi-Kamchatkan	0	0	0
10	Eskimo-Aleut	0	0	0
11	Elamo-Dravidian	0	0	0
12	Sino-Tibetian	0	0	0
13	Austric	25	25	25
14	Indo-Pacific	20	20	20
15	Australian	0	0	0
16	Na-Dene	49	0	0
17	Amerind	388	165	165
18	Language Isolates	47	0	0
19–21	Miscellaneous			

Classification and numbering follows Ruhlen (43). Miscellaneous represents, unclassified (19), pidgins and creoles (20) and invented (21). For 1657 individuals no information about the language could be obtained.

Geographical sampling

Table 1 shows the number of sequences and lineages for each continent. An overview of the world wide sampling is displayed in Figure 2. Obviously, some regions of the world are sampled well whereas sampling is still poor in other regions. Except for India and South Africa, where the number of HVRI and HVRII sequences is balanced, we note a strong preponderance for the former. For some regions only HVRI sequences are available.

Language sampling

Table 2 shows the number of sequences according to language phyla. Sequences are available for 12 of the 18 language phyla, classified according to Ruhlen (43). Unfortunately, for 1657 individuals the publications did not specifiy the linguistic affiliation of the sequences.

Alignment

The alignment of the HVRI sequences is 419 bp long and starts at position 16001 according to the human reference sequence (7). Gaps of varying length were introduced at positions 16104.1, 16169.1, 16174.1, 16183.1–16183.4, 16227.1, 16259.1, 16366.1, 16386.1. Especially, the region from position 16183 to 16193 shows a high degree of length variants (19,31). Among the 419 positions are 275 variable sites. 188 sites carry two different nucleotides (164 sites with transitions and 24 with transversions), 66 with three nucleotides and 21 sites show all four nucleotides.

The HVRII sequence alignment, which starts at position one comprises 418 bp with gaps at positions 56.1, 65.1, 190.1, 294.1, 302.1–302.4 and 310.1–310.2. Only 105 of 418 positions show different nucleotides. Two nucleotides are found in 89 of these positions (77 transitions and 12 transversions). The rest are 15 positions with three different nucleotides and one position that shows all four nucleotides.

QUALITY AND COMPLETENESS OF THE DATA AND FUTURE DIRECTIONS

Our data have been largely compiled from published sequences. Although we have taken great pains to minimise mistakes, there may still be sequences in our collection that contain errors or where some annotations are not correct. To ensure a high quality of the data, we are grateful if bugs or obscurities are pointed out to us.

We solicit everybody to furnish new sequences via electronic mail together with the relevant information. We would also be grateful to receive already published sequences which are missing in our collection.

Besides regular updates of the collection of human control-region sequences we are planning to add DNA sequences from the hypervariable region of the mitochondrial control region from chimpanzees. There are currently 377 sequences published (44–46).

While we have collected only the control region sequences from humans there are other databases like MITOMAP (47) that collect information about the variabilitiy of the entire human mitochondrial genome.

AVAILABILITY

The collection is available on request (meyers@zi.biologie.uni-muenchen.de or arndt@zi.biologie.uni-muenchen.de) It can also be retrieved free of charge over the internet from http://www.zi.biologie.uni-muenchen.de/~meyers/mtdna.html . We also distribute a simple program that allows retrieval of sequences according to specific keywords. The program is written in standard C and should run on most computers equipped with a C-compiler. It can also be obtained from the internet address given above.

ACKNOWLEDGEMENTS

We are grateful to all colleagues who provided their sequence data as a computer file and gave additional information when needed. We want to express our special thanks to Matthias Krings, Martin

Richards, Antti Sajantila, and Svante Pääbo. Financial support from the DFG is gratefully acknowledged.

REFERENCES

1 Nei,M. and Roychoudhury,A.K. (1988) *Human Polymorphic Genes: World Distribution*. Oxford University Press, New York.
2 Bowcock,A.M., Ruiz-Linares,A., Tomfohrde,J., Minch,E., Kidd,J.R and Cavalli-Sforza,L.L. (1994) *Nature* **368**, 455–457.
3 Cavalli-Sforza,L.L., Menozzi,P. and Piazza,A. (1994) *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.
4 Deka,R., Jin,L., Shriver,M.D., Yu,L.M., DeCroo,S., Hundrieser,J., Bunker,C.H., Ferrell,R.E. and Chakraborty,R. (1995) *Am. J. Hum. Genet.* **56**, 461–474.
5 Vigilant,L., Stoneking,M., Harpending,H., Hawkes,K. and Wilson,A.C. (1991) *Science* **253**, 1503–1507.
6 von Haeseler,A., Sajantila,A. and Pääbo,S. (1996) *Nature Genet.* **14**, 135–140.
7 Anderson,S., Bankier,A.T., Barell,B.G., deBruijn,M.H.L., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreier,P.H., Smith,A.J.H., Staden,R. and Young,I.G. (1981) *Nature* **290**, 457–465.
8 Batista,O., Kolman,C.J. and Bermingham,E. (1995) *Hum. Mol. Genet.* **4**, 921–929.
9 Bertranpetit,J., Sala,J., Calafell,F., Underhill,P.A., Moral,P. and Comas,D. (1995) *Ann. Hum. Genet.* **59**, 63–81.
10 Betty,D.J., Chin-Atkins,A.N., Croft,L., Scraml,M. and Easteal,S. (1995) *Am. J. Hum. Genet.* **58**, 428–433.
11 Comas,D., Calafell,F., Mateu,E. and Bertranpetit,J. (1996) *Mol. Biol. Evol.* **13**, 1076–1077.
12 Corte-Real,H.B., Macaulay,V.A., Richards,M.B., Hariti,G., Issad,M.S., Cambon-Thomsen,A., Papiha,S., Bertranpetit,J. and Sykes,B.C. (1996) *Ann. Hum. Genet.* **60**, 331–350.
13 DiRienzo,A. and Wilson,A.C. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1597–1601.
14 Easton,R., Merriwether,A., Crews,E. and Ferrell,R. (1996) *Am. J. Hum. Genet.* **59**, 213–225.
15 Francalacci,P., Bertranpetit,J., Calafell,F. and Underhill,P.A. (1996) *Am. J. Phys. Anthrop.* **100**, 443–460.
16 Ginther,C., Corach,D., Penacino,G., Rey,J.A., Hutz,M.H., Carnese,F.R., Anderson,A., Just,J., Salzano,F.M. and King,M.C. (1993) *EXS.* **67**, 211–219.
17 Graven,L., Passarino,G., Semino,O., Bourset,P., Santachiara-Benerecetti,S., Langaney,A. and Excoffier,L. (1995) *Mol. Biol. Evol.* **12**, 334–345.
18 Handt,O., Richards,M., Trommsdorff,M., Kilger,C., Simanainen,J., Georgiev,O., Bauer,K., Stone,A., Hedges,R., Schaffner,W., Utermann,G., Sykes,B. and Pääbo, S. (1994) *Science* **264**, 1775–1778.
19 Horai,S. and Hayasaka,K. (1990) *Am. J. Hum. Genet.* **46**, 828–842.
20 Horai,S., Kondo,R., Nakagawa-Hattori,Y., Hayashi,S., Sonada,S. and Tajima,K. (1993) *Mol. Biol. Evol.* **10**, 23–47.
21 Jorde,L.B., Bamshad,M.J., Watkins,W.S., Zenger,R., Fraley,A.E., Krakowiak,P.A., Carpenter,K.D., Soodyall,H., Jenkins,T. and Rogers,A.R. (1995) *Am. J. Hum. Genet.* **57**, 523–538.
22 Kolman,C.J., Bermingham,E., Cooke,R., Ward,R.H., Arias,T.D. and Guionneau-Sinclair,F. (1995) *Genetics* **140**, 275–283.
23 Kolman,C.J., Sambuughin,N. and Bermingham,E. (1996) *Genetics* **112**, 1321–1334.
24 Lum,J.M., Rickards,O., Ching,C. and Cann,R.L. (1994) *Hum. Biol.* **4**, 567–590.
25 Mountain,J.L., Hebert,J.M., Bhattacharyya,S., Underhill,P.A., Ottolenghi,C., Gadgil,M. and Cavalli-Sforza,L.L. (1995) *Am. J. Hum. Genet.* **56**, 979–992.
26 Piercy,R., Sullivan,K.M., Benson,N. and Gill,P. (1994) *Int. J. Legal. Med.* **106**, 85–90.
27 Pinto,F., Gonzales,A., Hernandez,M., Larruga,J. and Cabrera,V. (1996) *Ann. Hum. Genet.* **60**, 321–330.
28 Pult,I., Sajantila,A., Simanainen,J., Georgiev,O., Schaffner,W. and Pääbo, S. (1994) *Biol. Chem. Hoppe-Seyler* **375**, 837–840.
29 Redd,A.J., Takezaki,N., Sherry,S.T., McGarvey,S.T., Sofro,A.S.M. and Stoneking,M. (1995) *Mol. Biol. Evol.* **12**, 604–615.
30 Sajantila,A., Lahermo,P., Anttinen,T., Lukka,M., Sistonen,P., Savontaus,M.-L., Aula,P., Beckman,L., Tranebjaerg,L., Gedde-Dahl,T., Issel-Tarver,L., DiRienzo,A. and Pääbo, S. (1995) *Gen. Res.* **5**, 42–52.
31 Santos,M., Ward,R.H. and Barrantes,R. (1994) *Hum. Biol.* **6**, 963–977.
32 Santos,S., Ribeiro-Dos-Santos,A., Meyer,D. and Zago,M. (1996) *Ann. Hum. Genet.* **60**, 305–319.
33 Stenico,M., Nigro,L., Bertorelle,G., Calafell,F., Capitanio,M., Corrain,C. and Barbujani,G. (1996) *Am. J. Hum. Genet.* **59**, 1363–1375.
34 Sykes,B.C., Leiboff,A., Low-Beer,J., Tetzner,S. and Richards,M. (1995) *Am. J. Hum. Genet.* **57**, 1463–1475.
35 Torroni,A., Schurr,T.G., Cabell,M.F., Brown,M.D., Neel,J.V., Larsen,M., Smith,D.G., Vullo,C.M. and Wallace,D.C. (1993) *Am. J. Hum. Genet.* **53**, 563–590.
36 Torroni,A., Sukernik,R.I., Schurr,T.G., Starikovskaya,Y.B., Cabell,M.F., Crawford,M.H., Comuzzie,A.G. and Wallace,D.C. (1993) *Am. J. Hum. Genet.* **53**, 591–608.
37 Vigilant,L. (1990) Control region sequences from African populations and the evolution of human mitochondrial DNA. PhD thesis, University of California, Berkeley, CA.
38 Ward,R.H., Frazier,B.L., Dew-Jager,K. and Pääbo,S. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8720–8724.
39 Ward,R.H., Redd,A., Valencia,D., Frazier,B. and Pääbo, S. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 10663–10667.
40 Watson,E., Bauer,K., Aman,R., Weiss,G., von Haeseler,A. and Pääbo,S. (1996) *Am. J. Hum. Genet.* **59**, 437–444.
41 Benson,D.A., Boguski,M.S., Lipman,D.L. and Ostell,J. (1997) *Nucleic Acids Res.* **25**, 1–6. [See also this issue, *Nucleic Acids Res.* (1998) **26**, 1–7.].
42 Wrischnik,L.A., Higuchi,R.G., Stoneking,M., Erlich,H.A., Arnheim,N. and Wilson,A.C. (1987) *Nucleic Acids Res.* **15**, 529–542.
43 Ruhlen,M. (1991) *A Guide to the World's Languages, Volume 1: Classification*. Edward Arnold, A Division of Hodder & Stoughton, London, Melbourne, Auckland.
44 Morin,P.A., Moore,J.J., Jin,L., Chakraborty,R., Goodall,J. and Woodruff,D.S. (1994) *Science* **265**, 1193–1201.
45 Wise,C.A., Sraml,M., Rubinsztein,D.C. and Easteal,S. (1997) *Mol. Biol. Evol.* **14**, 707–716.
46 Goldberg,T.L. and Ruovolo,M. (1997) *Nucleic Acids Res.* **25**, 1–6.
47 Kogelnik,A.M., Lott,M.T., Brown,M.D., Navathe,S.B. and Wallace,D.C. (1997) *Nucleic Acids Res.* **25**, 196–199. [See also this issue, *Nucleic Acids Res.* (1998) **26**, 112–115.]