BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 6 2002
Pages 890–891

MOUSE (Mitochondrial and Other Useful SEquences) a compilation of population genetic markers

Florian Burckhardt

GSF-National Research Center for Environment and Health, Institute of Epidemiology, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

Received on December 7, 2001; revised on February 1, 2002; accepted on February 5, 2002

ABSTRACT

Summary: Mitochondrial and Other Useful SEquences (MOUSE) is an integrated and comprehensive compilation of mtDNA from hypervariable regions I and II and of the low recombining nuclear loci Xq13.3 from about 11 200 humans and great apes, whose geographic and if applicable, linguistic classification is stored with their aligned sequences and publication details. The goal is to provide population geneticists and genetic epidemiologists with a comprehensive and user friendly repository of sequences and population information that is usually dispersed in a variety of other sources.

Availability: http://www.gen-epi.de/mouse

Contact: florian.burckhardt@gsf.de

Supplementary Information: Documentation and detailed information on population subgroups is available on the homepage: http://www.gen-epi.de/mouse

The ever increasing number of sequences used in phylogenetic analysis of primates poses a serious challenge for adequate information management. Individual publications usually focus on one subgroup only, or alternatively, on a limited set of different subgroups. The large number of human and great ape sequences dispersed in many different publications makes literature research a tedious task. Meta-databases of specific sequence subsets are necessary for efficient data handling.

Mitochondrial and Other Useful SEquences (MOUSE) was created to provide the scientific community with a user friendly compilation of published, aligned non-recombining DNA-markers of humans and great apes, including geographic and if applicable, linguistic classifications of the individuals as well as sequence and publication details.

Population genetic studies frequently use DNA from the hypervariable regions I (HV1) and II (HV2) of the mitochondrial D-loop and several databases have been set up to catalog different aspects of the mitochondrial genome (e.g. Burckhardt *et al.*, 1999; Lanave *et al.*, 2000; Attimonelli *et al.*, 2000; Kogelnik *et al.*, 1998). However, mtDNA entails some limitations, such as maternal inheritance only or a high mutation rate. Therefore, non-coding nuclear loci with very low recombination rates, such as the 10kb Xq13.3-sequence of the X-chromosome, are increasingly used to overcome some limitations of mtDNA.

Currently, MOUSE encompasses 11 114 individuals HV1 and/or HV2 sequenced and 130 individuals with Xq13.3. Table 1 summarizes the current mtDNA and Xq13.3 distribution of human individuals per continent as of November 2001. Europeans are the most frequently sequenced population group (40.3%) followed by Asians (18.7%), Africans (13.7%), Australia/Oceanians (11.0%), North Americans (8.6%) and Central and South Americans (7.6%). Table 2 gives the mtDNA and Xq13.3 distribution for great apes. *Pan trogdlodytes* spp. (76.5%) is followed by *Pongo pygmaeus* spp. (12.3%), *Gorilla gorilla* spp. (8.0%) and *Pan paniscus* (7.3%). More details on the 345 different subpopulations are given on the homepage of MOUSE.

HV1 and HV2 lineages are aligned globally and stored with the frequency they occur in the database. Xq13.3 data are stored unaligned. Alignment of primate HV1/2 is based on Anderson *et al.* (1981), but takes into account the various insertions that have accrued over the years (Burckhardt *et al.*, 1999). Global insertions at the following positions had to be made to accommodate human and great ape HV1 and HV2 sequences: 16104.1, 16139.1, 16169.1, 16174.1-2, 16183.1-4, 16227.1, 16259.1, 16296.1, 16366.1, 16386.1 (HV1) and 56.1-2, 174.1, 190.1, 291.1-2, 294.1, 302.1-4, 315.1-2 (HV2).

The compilation incorporates only data previously published in public databases like PubMed. Sequences and data on individuals were obtained systematically. A database search of 'Web of Science' (http://wos.mimas.ac.uk/) and 'Ovid' (http://biomed.niss.ac.uk/ovidweb/ovidweb.cgi) was conducted using the keywords 'd-loop' and 'mtdna', limited to the years 1999–2001.

Table 1. Distribution of individuals for mtDNA and Xq13.3 per continent (humans). note: XQ13.3 was sampled to represent language groups, not continents

Continent	Number of individuals with mtDNA (human)	Number of individuals with Xq13.3 (human)
Africa	1014	23
North America	971	1
Central and South America	857	2
Asia	2005	23
Europe	4513	11
Australia/Oceania	1229	9
Ancient Human	17	0
Total	**10 606**	**69** (+1 n.d.)

Table 2. Distribution of individuals for mtDNA and Xq13.3 per ape species

Species	Number of individuals with mtDNA (apes)	Number of individuals with Xq13.3 (apes)
Pan troglodytes spp.	387	30
Gorilla gorilla spp.	33	11
Pongo pygmaeus spp.	53	14
Pan paniscus	35	5
Total	**508**	**60**

The resulting publications on primate DNA were then screened for sequences with the help of PubMed. It should be noted that it was not always possible to retrieve a publication's set of sequences using names of authors or publication title. In that case, the publication was left out.

MOUSE currently covers 95 publications of D-Loop and Xq13.3 data, an exhaustive list is given on the website.

MOUSE features a rich and intuitive graphical user interface with online help and comes with an extensive documentation and examples. Individuals and sequences can be searched for by using any one or more of the following: sequence type, sequence motif, species, continent, origin, population, language, language phylum, original name of sequence in publication or Genbank, author, journal or year of publication. Primate sequence specific literature can be searched for in great detail.

Query results are displayed and exported in different formats for subsequent analysis. Summary statistics are also available for more complex tasks, e.g. to find out how many subpopulations with how many individuals from the African continent have a certain nucleotide sequence in HV1. Extraction of lineages and a set of basic sequence editing tools are also available. An online version of the database using HTML, PHP and MySQL is currently under development and will be made available for online access or local installation.

The MOUSE-database is updated in regular intervals to include the rapidly growing number of sequences and loci. Please refer to the homepage for detailed and updated statistics. User feedback is encouraged and will be used for future improvements. Notification of new sequences or sequences that have been omitted is greatly appreciated.

ACKNOWLEDGEMENTS

I want to thank R.Külbel for support and all colleagues who kindly provided their sequence data and their feedback. I also would like to thank N.Dostert and R.Perry for help and discussion and S.Meyer and A.v.Haeseler for collaboration on one predecessor of MOUSE.

REFERENCES

Anderson,S., Bankier,A.T., Barell,B.G., de Bruijn,M.H.L., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F. *et al.* (1981) Sequence and Organization of the Human Mitochondrial Genome. *Nature*, **290**, 457–465.

Attimonelli,M., Altamura,N., Benne,R., Brennicke,A., Cooper,J.M., D'Elia,D., de Montalvo,A., de Pinto,B., De Robertis,M., Golik,P. *et al.* (2000) MitBASE: a comprehensive and integrated mitochondrial DNA database. The present status. *Nucleic Acids Res.*, **28**, 148–152.

Burckhardt,F., von Haeseler,A. and Mayer,S. (1999) HvrBase: compilation of mtDNA control region sequences from primates. *Nucleic Acids Res.*, **27**, 138–142.

Kogelnik,A.M., Lott,M.T., Brown,M.D., Navathe,S.B. and Wallace,D.C. (1998) MITOMAP: a human mitochondrial genome database—1998 update. *Nucleic Acids Res.*, **26**, 112–115.

Lanave,C., Liuni,S., Licciulli,F. and Attimonelli,M. (2000) Update of AMmtDB: a database of multi-aligned Metazoa mitochondrial DNA sequences. *Nucleic Acids Res.*, **28**, 153–154.