*Nucleic Acids Research, 2006, Vol. 34, Database issue* ***D749–D751***
*doi:10.1093/nar/gkj010*

# mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences

**Max Ingman[1,2,*] and Ulf Gyllensten[2]**

[1]Centre for Integrative Genomics, University of Lausanne, Switzerland and [2]Department of Genetics and Pathology, Rudbeck Laboratory, University of Uppsala, Uppsala, Sweden

Received August 9, 2005; Revised and Accepted September 12, 2005

## ABSTRACT

**The mitochondrial genome, contained in the subcellular mitochondrial network, encodes a small number of peptides pivotal for cellular energy production. Mitochondrial genes are highly polymorphic and cataloguing existing variation is of interest for medical scientists involved in the identification of mutations causing mitochondrial dysfunction, as well as for population genetics studies. Human Mitochondrial Genome Database (mtDB) (http://www.genpat.uu.se/mtDB) has provided a comprehensive database of complete human mitochondrial genomes since early 2000. At this time, owing to an increase in the number of published complete human mitochondrial genome sequences, it became necessary to provide a web-based database of human whole genome and complete coding region sequences. As of August 2005 this database contains 2104 sequences (1544 complete genome and 560 coding region) available to download or search for specific polymorphisms. Of special interest to medical researchers and population geneticists evaluating specific positions is a complete list of (currently 3311) mitochondrial polymorphisms among these sequences. Recent expansions in the capabilities of mtDB include a haplotype search function and the ability to identify and download sequences carrying particular variants.**

## INTRODUCTION

The mitochondrial genome supplies parts of the protein machinery that are necessary for oxidative phosphorylation (OXPHOS), by utilizing a series of five multiple-subunit enzymes located within the mitochondrial inner membrane. The complex constituents are encoded by both nuclear and mitochondrial genes. A genetic defect could therefore be due to mutations in genes of either system. Since new mutations are introduced more frequently to the mitochondrial genome, a higher proportion of mitochondrial dysfunction is due to mitochondrial DNA (mtDNA) mutations. A number of human diseases have been shown to be caused by mitochondrial mutations, such as Leber's hereditary optic neuropathy (LHON) (1) and neurogenetic muscle weakness, ataxia, and retinitis pigmentosa (NARP) (2). In the evaluation of a possible functional effect of a mitochondrial variant found in a group of patients, reliable population frequency data for the variant under study is needed. The Human Mitochondrial Genome Database (mtDB) provides such a compilation of available genome sequences information for this purpose.

The mtDNA of most metazoan species (including humans) is predominantly maternally inherited (3). This clonal inheritance coupled with a substitution rate that in vertebrates is typically 5 to 10 times that of nuclear DNA (4) has made mitochondria an attractive source of DNA polymorphism data for population genetics studies in a wide range of species. The lack of recombination among maternal and paternal mitochondrial genomes allows the tracing of a direct genetic line where all polymorphism is due to mutation and the high substitution rate makes it possible to study variation between closely related individuals (i.e. within species). mtDNA sequences have been the main tool in a large number of studies of human evolution. The Human Mitochondrial Genome Database (mtDB) is a repository for these sequences and will provide scientists with access to a common resource for future studies in this field.

Since 2000, with the publication of the first comprehensive study on complete human mitochondrial genome sequences (5), the amount of data available from mitochondrial genomes has been growing rapidly. However, polymorphism information from these data is becoming more time consuming to produce. The mtDB provides a unique resource to both medical and human population genetic researchers. Here,

*To whom correspondence should be addressed. Tel: +41 21 692 3962; Fax: +46 18 471 4931; Email: max.ingman@unil.ch published mitochondrial genome sequences are collected from GenBank and other sources (not all sequences are submitted to GenBank) and made available for download. In addition, extensive polymorphism information from the complete dataset is easily accessible.

## Database content

The mtDB database contains three principal types of content for researchers:

(i) Download of all mitochondrial sequences either as individuals or population sets. The sequences are grouped into 10 major geographic regions based on the population origin of the donor (Table 1). In cases where the geographic origin of the donor is different from their supposed historical background, the sequences are listed under the heading that best fits their donors' ancestry. For example, African American, European American and Asian American sequences are not listed under North America but under the headings Africa, Europe and Asia, respectively. Large sets from the same population are available as batches of individual files. All sequences are cross-referenced to their original publications and to GenBank accession numbers, where available. There are currently 2104 mitochondrial sequences at mtDB.

(ii) A list of all variable positions [numbered as in Cambridge Reference Sequence, CRS (6)] among complete, or near complete, mitochondrial sequences (Figure 1). Currently, 3311 polymorphic sites are identified and characterized in tabulated form. This table comprises a separate line for each variable site with a count of how many sequences contain each particular nucleotide variant at that site, the genic location of that site, the codon number and position and details of amino acid changes. An interested researcher can click on the number of a particular variant to obtain a list of all sequences that contain that particular variant. These sequences can then be downloaded from the list. All insertions relative to CRS have been removed.

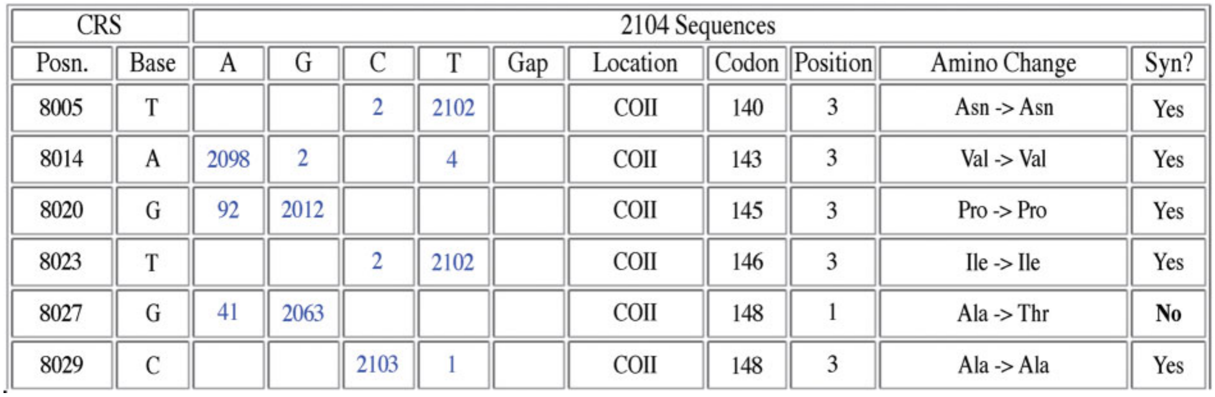
(iii) A search function for mitochondrial haplotypes. This goes a step beyond the list of variable positions in that sequences carrying specific haplotypes can be retrieved by entering the position and nucleotide for up to 10 loci. Only sequences that match all these criteria will be returned. Again, these sequences can then be downloaded from the database.

Some population genetics researchers use predefined haplotypes (haplogroups) purported to designate specific mitochondrial lineages. As a compliment to our search function, this page has a link to a haplogroup tree where clicking the individual haplogroup letters will return a list of all sequences that belong to that particular group.

**Table 1.** Summary table of the number of sequences from each of the 10 geographic regions

| Population | Complete | Coding region |
|---|---|---|
| Africa | | |
| American | | 56 |
| Other | 58 | |
| America (North) | 6 | |
| America (South) | 3 | |
| Asia | | |
| Japanese | 672 | |
| Chinese | 48 | |
| American | | 69 |
| Other | 70 | |
| Australia | | |
| Aborigine | 20 | |
| Europe | 535 | |
| American | 241 | 435 |
| Finnish | 192 | |
| Italian | 62 | |
| Other | 40 | |
| Melanesia | | |
| New Guinea | 21 | |
| Other | 2 | |
| Middle East | 7 | |
| Polynesia | 6 | |
| South Asia | | |
| Indian | 75 | |
| Other | 21 | |
| Total: | 1544 | 560 |

Population groups for which a large number of sequences are available are indicated separately. Coding region sequences extend from np 577 to 16023, relative to CRS (6).



| CRS | | 2104 Sequences | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Posn. | Base | A | G | C | T | Gap | Location | Codon | Position | Amino Change | Syn? |
| 8005 | T | | | 2 | 2102 | | COII | 140 | 3 | Asn -> Asn | Yes |
| 8014 | A | 2098 | 2 | | 4 | | COII | 143 | 3 | Val -> Val | Yes |
| 8020 | G | 92 | 2012 | | | | COII | 145 | 3 | Pro -> Pro | Yes |
| 8023 | T | | | 2 | 2102 | | COII | 146 | 3 | Ile -> Ile | Yes |
| 8027 | G | 41 | 2063 | | | | COII | 148 | 1 | Ala -> Thr | **No** |
| 8029 | C | | | 2103 | 1 | | COII | 148 | 3 | Ala -> Ala | Yes |

**Figure 1.** Truncated table of polymorphic sites. Each row of the table shows nucleotide position [relative to Anderson (Cambridge Reference Sequence, CRS) (6)], CRS nucleotide state at that position, the number of database sequences with A, G, C, T or gap, and the functional region that the site is in. If the functional region is a protein coding gene, also listed is the codon number, the codon position, the amino acid state in CRS and for the variant, whether the change is synonymous or not. Clicking the number of sequences with a particular nucleotide state will retrieve a list of all sequences that carry that particular variant.

### Database interface

To facilitate easy updating of mtDB, all data pages are produced dynamically by PHP scripts. PHP is an easy-to-use scripting language that integrates well with HTML. Data is parsed on the server machine and an HTML output is sent to the client. This is independent of the client's operating system, browser and installed options. The only exception to this, is the polymorphic sites, nucleotide variants and amino acid states list which is produced by a separate script and the HTML output saved to avoid long processing time for individual requests. The core database is a text file of aligned sequences. New sequences can be simply pasted to this list and are then included in searches.

## CONCLUSIONS

mtDB is the only comprehensive online source for the data contained within it. This includes the sequences themselves as many have not been deposited in a publicly available database such as GenBank. The list of mitochondrial polymorphisms continually grows with the addition of new sequences and is an important resource for phylogenetic and medical studies. The ability to search for multiple-variant haplotypes adds further detail to the latent data. We are committed to the maintenance of this database and hope that it will be a useful resource for researchers for years to come.

## ACKNOWLEDGEMENTS

This research has been supported by the Swedish National Research Council (UG). Funding to pay the Open Access publication charges for this article was provided by Swedish National Research Council (UG).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Wallace,D.C., Singh,G., Lott,M.T., Hodge,J.A., Schurr,T.G., Lezza,A.M., Elsas,L.J.,II and Nikoskelainen,E.K. (1988) Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science*, **242**, 1427–1430.
2. Holt,I.J., Harding,A.E., Petty,R.K. and Morgan-Hughes,J.A. (1990) A new mitochondrial disease associated with mitochondrial DNA heteroplasmy. *Am. J. Hum. Genet.*, **46**, 428–433.
3. Giles,R.E., Blanc,H., Cann,H.M. and Wallace,D.C. (1980) Maternal inheritance of human mitochondrial DNA. *Proc. Natl Acad. Sci. USA*, **77**, 6715–6719.
4. Brown,W.M., George,M.,Jr and Wilson,A.C. (1979) Rapid evolution of animal mitochondrial DNA. *Proc. Natl Acad. Sci. USA*, **76**, 1967–1971.
5. Ingman,M., Kaessmann,H., Paabo,S. and Gyllensten,U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature*, **408**, 708–713.
6. Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H., 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.