Nucleic Acids Research Advance Access published November 28, 2016

Nucleic Acids Research, 2016 **1**
doi: 10.1093/nar/gkw1066

HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor

Rosanna Clima[1,2,†], Roberto Preste[1,†], Claudia Calabrese[3], Maria Angela Diroma[1], Mariangela Santorsola[1], Gaetano Scioscia[4], Domenico Simone[5], Lishuang Shen[6], Giuseppe Gasparre[2,‡] and Marcella Attimonelli[1,*,‡]

[1]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari, 70126 Bari, Italy, [2]Department of Medical and Surgical Sciences-DIMEC, Medical Genetics Unit, University of Bologna, 40126 Bologna, Italy, [3]European Bioinformatics Institute EMBL Outstation - Hinxton, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK, [4]IBM Italia S.p.A., GBS BAO Advanced Analytics Services and MBLab Bari, Italy, [5]Department of Biology and Environmental Science, Centre for Ecology and Evolution in Microbial model Systems (EEMiS), Linnaeus University, Barlastgatan 11, Kalmar, Sweden and [6]Center for Personalized Medicine, Children's Hospital Los Angeles, Los Angeles, California, CA 90027, USA

Received September 15, 2016; Revised October 20, 2016; Editorial Decision October 21, 2016; Accepted November 04, 2016

ABSTRACT

The HmtDB resource hosts a database of human mitochondrial genome sequences from individuals with healthy and disease phenotypes. The database is intended to support both population geneticists as well as clinicians undertaking the task to assess the pathogenicity of specific mtDNA mutations. The wide application of next-generation sequencing (NGS) has provided an enormous volume of high-resolution data at a low price, increasing the availability of human mitochondrial sequencing data, which called for a cogent and significant expansion of HmtDB data content that has more than tripled in the current release. We here describe additional novel features, including: (i) a complete, user-friendly restyling of the web interface, (ii) links to the command-line standalone and web versions of the MToolBox package, an up-to-date tool to reconstruct and analyze human mitochondrial DNA from NGS data and (iii) the implementation of the Reconstructed Sapiens Reference Sequence (RSRS) as mitochondrial reference sequence. The overall update renders HmtDB an even more handy and useful resource as it enables a more rapid data access, processing and analysis. HmtDB is accessible at http://www.hmtdb.uniba.it/.

INTRODUCTION

The investigation of human mitochondrial DNA (mtDNA) variation and genetics over the past decades has provided unique and often startling insights into human evolution, degenerative diseases and aging (1,2). Hence, great effort is undertaken to fully sequence human mtDNA genomes of healthy individuals from different populations around the world, and of subjects showing phenotypes definitely or potentially related to mitochondrial dysfunctions (3).

In the last years the availability of next generation sequencing (NGS) technologies has offered the unprecedented opportunity to understand mitochondrial variations, allowing the quantification of heteroplasmy (4,5) (the coexistence of different mtDNA genotypes within the same cell, tissue or individual) by providing good quality and high coverage mtDNA sequences. Indeed, since mtDNA mutations, in most cases, exert their phenotypic effect only above a certain mutation load threshold (6,7), a wide range of low heteroplasmic variants is identified even in healthy individuals (8,9).

The widespread application of NGS by clinicians to highly heterogeneous disorders, such as mitochondrial diseases, is limited by substantial technical and bioinformatics challenges involving the storage and correct identification and prioritization of the extensive number of sequence variants that occur in each individual, compared to a reference sequence. This effort has engaged over a hundred international mitochondrial clinicians, researchers and bioinformaticians in implementing the Mitochondrial Disease Se-

*To whom correspondence should be addressed. Tel: +39 0805443308; Email marcella.attimonelli@uniba.it
†These authors equally contributed to this work as the first authors.
‡Co-last authors.

quence Data Resource (MSeqDR) consortium, established in June 2012 with the aim to collect, integrate, organize and critically analyze mitochondrial sequence data (10). In this frame the Human Mitochondrial Database (HmtDB, http://www.hmtdb.uniba.it/), developed in 2005, is actively and largely contributing to populate the MSeqDR portal with both data and hosted tools, such as the MToolBox package (11) with its recently implemented multi-parametric workflow for the prioritization of mtDNA variants of clinical interest (12). At its birth, the database stored 1255 human mitochondrial sequences (13). Driven by the advent of NGS, HmtDB now provides sequence records for thousands of genomes obtained from both healthy and pathologic samples, and offers a direct link to high-performance pipelines devoted to the analysis of human mtDNA, which always requires variant annotation, based on the alignment of a subject sequence with a human mtDNA reference.

Since the beginning of the sequencing era, in the early 1980s, the first human mitochondrial genome, namely the Cambridge Reference Sequence (CRS) (14) subsequently revised and thereby renamed rCRS (15), has been used to compare newly sequenced mtDNA genomes. In 2012 a virtual sequence, the Reconstructed Sapiens Reference Sequence (RSRS) was proposed to represent the inferred root haplotype of human phylogeny (16), differently from rCRS, which is a recent genome of one woman of European descent and hence a haplogroup pattern descriptor. The usage of RSRS required revisions of current mtDNA databases along with the tools for haplogroup classification, therefore, HmtDB has implemented the RSRS for variant annotation and haplogroup prediction of each human mtDNA genome stored in the database. However, since the rCRS remains largely used especially in the clinical practice, both RSRS and rCRS are used for variant annotation and prioritization tools (12) implemented within the MToolBox package (11), now directly accessible through HmtDB.

HmtDB resource description

Web interface. The HmtDB web interface has been redesigned and made more lightweight, in order to provide users with shorter page loading times and a more enjoyable experience. The whole website is built upon the W3.CSS framework (http://www.w3schools.com/w3css/default.asp) that allows a safe and consistent page rendering on every platform and web browser, including mobile devices. Several other graphics improvements allow users to easily locate, access and visualize data in a more modern fashion with respect to the previous versions of the website (13,17).

Figure 1 shows a screenshot of the HmtDB Menu page, with the *Query* link to access the database, the *Download* link to download the latest versions of mitochondrial genome alignments and site variability data, as well as the link to the novel tools that allow human mtDNA classification described in details in the *MToolBox and prioritization workflow* paragraph below.

Features of the query page. The number of genomes and data made available through the HmtDB platform has dramatically increased since its first publication, now counting more than 30 000 human genomes and 10 000 variant sites. Such an amount of data has required the development of a more efficient way of consulting the database. For this reason, the Query function of HmtDB has undergone a deep restructuring, mostly involving fields that used to synchronously access data from the HmtDB database when entering the Query page, and that in the previous version of the database determined a longer loading times. The *HmtDB Genome Identifier* and *GenBank Accession Number* fields now offer an open text area instead of a dropdown menu with all the HmtDB Genome Identifiers and GenBank Accession Numbers (AN); furthermore, the *GenBank Accession Number* field has now been renamed to *Reference DB Source Identifier*, referring to the genome identifier reported in the database where the genome comes from. The *Subjects' Geographical Origin* field requires users to first select a specific continent. Only countries for which an entry in HmtDB is available, belonging to that continent, will then be loaded in the *Country* field. However, users may also choose to perform a query for a single continent, without selecting any specific country, so that genomes coming from all the countries belonging to that continent will be returned. The former *Haplogroups* field is now split in two different dropdown menus, using the same approach as above: after choosing a specific macro-haplogroup (e.g. A, B, etc.), all its related haplogroups will be loaded (e.g. A1, A1b, etc.) in the second menu; in this case, however, users must select both the macro-haplogroup and the specific haplogroup.

A new query function is also implemented in the *SNP Position* field, in addition to the standard single-position query, allowing users to search for genomes containing SNPs located in a range of positions (e.g. 1120–2780 will return all genomes containing a SNP in at least one of the positions in the specified range), as well as to search for genomes containing SNPs in one or more positions of a list (e.g. 245, 2145, 11 789 will return the list of genomes that contain variants in position 245 OR in position 2145 OR in position 11 789).

Once the query is submitted, a list of genomes consistent with the selected parameters will be shown, and after selecting one or more specific genomes users may either download related sequences or the alignment against RSRS, or view the related Genome Cards, which report the whole set of annotations as in Figure 2. The Genome Card view has also been improved, particularly regarding the haplogroup prediction results, as further described in the *Haplogroup prediction* section below.

Database content

Eleven years ago, the HmtDB project started with the aim to store and analyze human mitochondrial genomes (13). At that time, there were 1255 human mitochondrial sequences publicly available, and only 695 of them were complete genomes, while the remaining were sequences of the coding region, i.e. the entire genome except the regulatory D-loop. In the last 10 years, the number of sequenced genomes has continued to increase exponentially (Figure 3). Currently there are 32 922 mitochondrial genomes publicly available in HmtDB, including 1427 mitochondrial genomes reconstructed as off-target sequences from exome data (9,18) generated by the 1000 Genomes Project

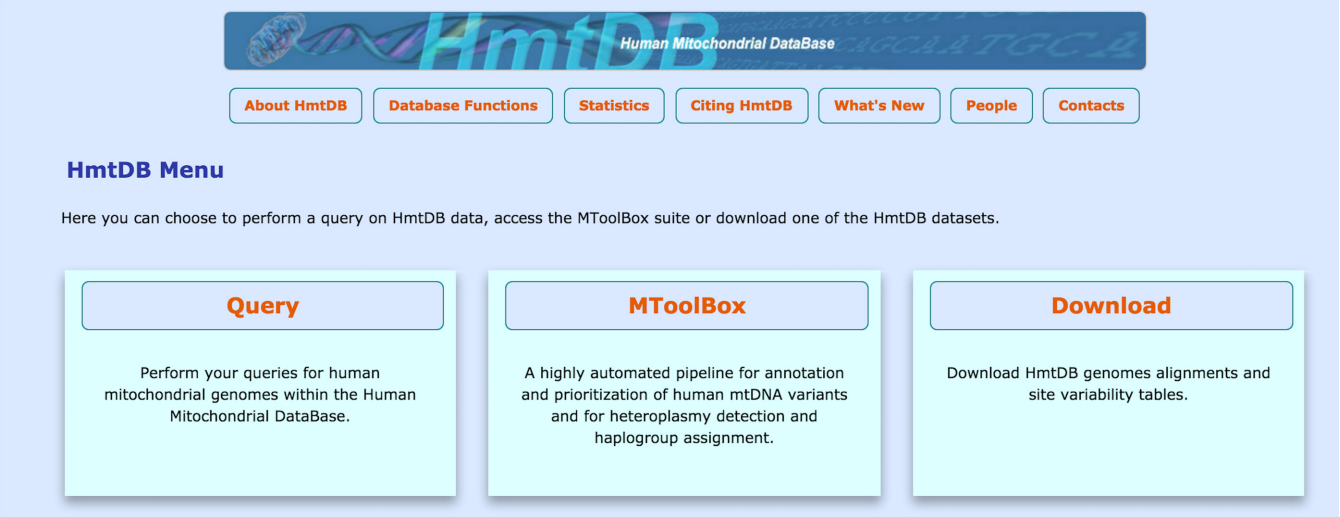


Figure 1. HmtDB Menu page. This page provides access to three sections: **(i)** a Query page to retrieve data from HmtDB; **(ii)** MToolBox, in which links to the command-line stand-alone and/or web versions of the MToolBox pipeline are provided and **(iii)** Download, that allows to download the latest versions of mitochondrial genome alignments and site variability data.

(19,20). The sequences are organized in two data sets representing individuals with healthy and disease phenotype, including 29 274 and 3648 samples, respectively. Samples reported as 'healthy' mainly derive from population studies or control in clinical studies, while 'pathologic' genomes come from individuals affected by mitochondrial diseases or other stated clinical conditions. It must be underlined here that the 'healthy' feature merely refers to what is reported in the study that produced that specific sequence, meaning that subjects referred to as controls in clinical studies are reported as healthy on the basis of not having *that* specific pathology, but not of not having *any* disease. Similarly, subjects included in population studies are not necessarily free of disease, if this has not been taken into account in the study, and are therefore reported as 'healthy' in the absence of more specific details. Both sets are grouped in continent-specific subsets (AF: Africa, AM: America, AS: Asia, EU: Europe, OC: Oceania, XX: Undefined Continent). The number of items of each data set is available through the *Statistics* window, where the world map is shown, offering a straightforward view of the number of genomes and variants annotated in the database for each continent, followed by detailed information about genome type (Normal/Patient Complete/Only Coding Region) (Figure 4).

Thanks to its constant and continuous updating, HmtDB is nowadays considered the most reliable source of consultation for mtDNA variants, along with MITOMAP (21). It is worth mentioning, however, that a relevant number of papers still cite mtDB (22) as the reference database when assessing the novelty, the polymorphic nature or even the inferred pathogenicity of a mtDNA variant. Since mtDB is a dead resource, whose last updating procedure according to the database site itself traces back to 2007, included data are based on 1865 genomes only (against the nearly 33 000 of HmtDB), whereas only about 3800 variants are annotated against the 10 000 of HmtDB, rendering variability far from reliable. With the aim not to mislead researchers in the interpretation of novel mtDNA sequences, MITOMAP has recently deleted the link to mtDB and the international MSeqDR consortium now recognizes that the database is of relatively low use when assessing variability parameters.

Improvements in the updating procedure

The HmtDB updating procedure described in our earlier article (14) was further improved by an auto-update protocol. The previous updating protocol used for retrieving and storing genomes and references in HmtDB required manual supervision of the results from several scripts, to ensure that all the data were correctly included in the database. However, the constantly evolving rapidity in genome sequencing and publication of related data on the web rendered this approach a cumbersome one. We circumvented this drawback by creating an automated protocol capable of periodically connecting to the *International Nucleotide Sequence Database Collaboration* (INSDC) databases and looking for mitochondrial genomes added subsequently to the latest HmtDB update. Such genomes are then downloaded in both GenBank and FASTA format, and the information contained in these entries is extracted and used to populate several HmtDB database tables. The multi-alignment of new genomes and the RSRS sequences to the data set of genomes already annotated in HmtDB is then performed by MAFFT software (version 7.304) by selecting the –auto option. Multi-alignment softwares may generate ambiguous alignments around homopolymeric regions where short insertions/deletions are frequently observed. Thus, in order to avoid the random insertion of gaps in the automatically generated multi-alignment, a single human-supervised step is hence required to fine-tune the MAFFT output within homopolymer-rich regions. The well-aligned genomes are then added to the previously stored multi-alignment and further processed through an automated protocol in order to estimate site specific variability. Each new genome is annotated with (i) information extracted from the GenBank

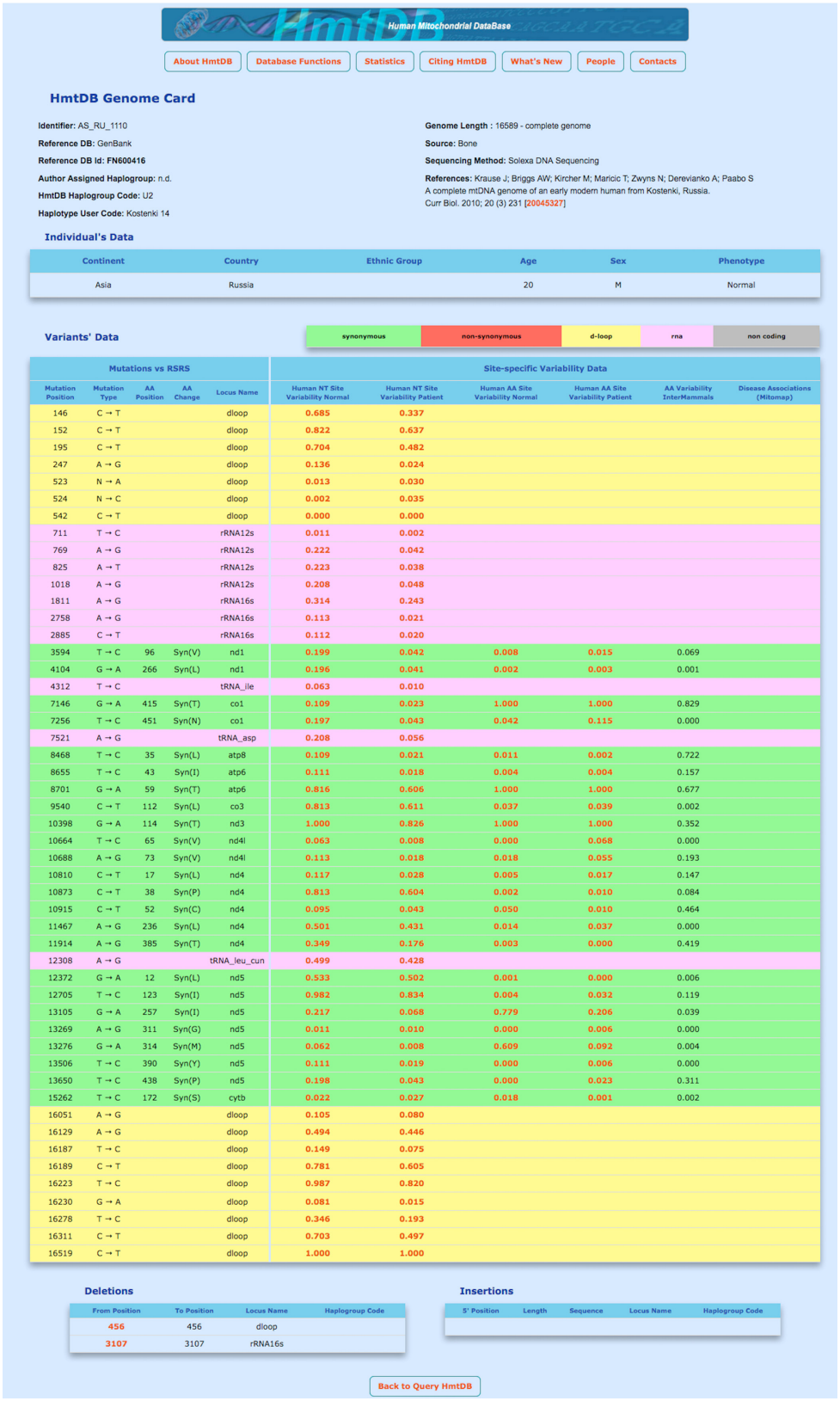


Figure 2. Genome card. The genome card shown is associated to the genome AS_RU_1110, a genome from a Russian sample. The top of the genome card shows the following information: identifier of the genome, reference database, accession number, haplogroup assigned by the author, haplogroup predicted by MToolBox, haplotype user code, genome length, source tissue, sequencing method and references. In the section 'Individual's data' information related to the subject are described: continent, country, ethnic group, age, sex and phenotype. In the table 'Variants Data' all the mutations identified in the genome are reported, annotated with variability data.

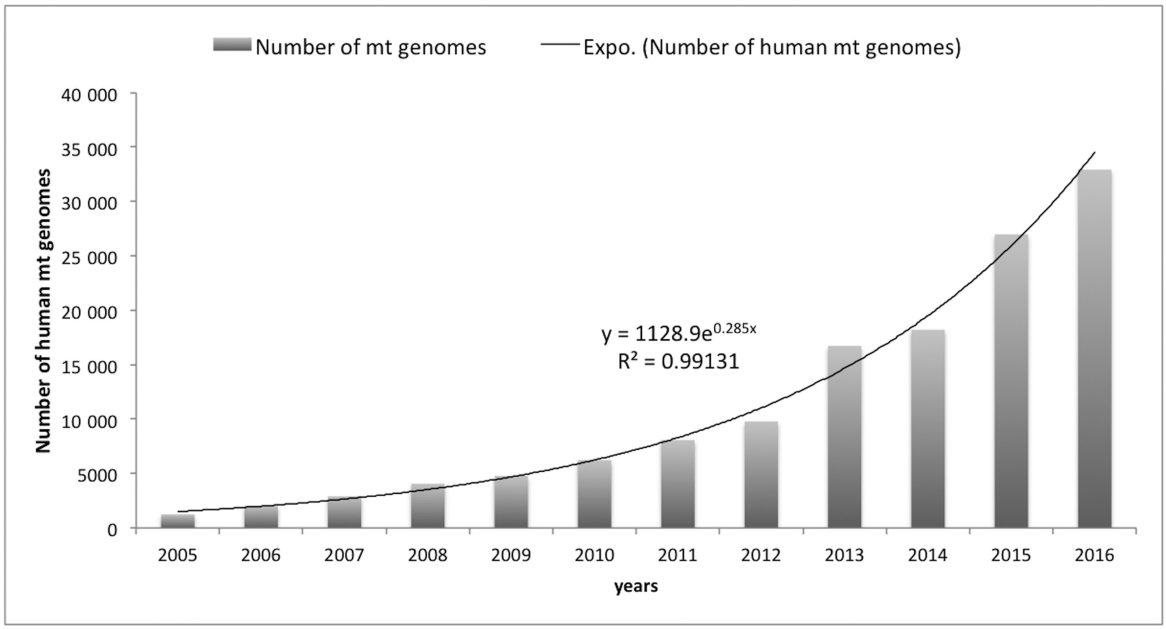


Figure 3. HmtDB database growth trend. An exponential growth of the number of genomes deposited in the database can be observed through the years, mirroring advances in sequencing technologies.

entry and/or from the related publication; (ii) variant sites detected through the comparison with RSRS; (iii) the best predicted haplogroup according to the Phylotree classification (23), as described in the next paragraph. Data gathered with this procedure are then made available through the *Genome Card*.

Haplogroup prediction

Human mtDNAs may be grouped in clusters of evolutionarily closely-related haplotypes, namely the haplogroups, defined by the pattern of genetic markers occurring in the entire mtDNA and reflecting the migration of human populations over the different continents (24,25). Human mitochondrial phylogeny is described through the haplogroup classification, whose complexity clinicians may or may not be familiar with, and that nonetheless is of support to clinical studies, where genetic association between the variant pattern and the haplogroup may be traced (26–29). Upon sequencing and reconstruction of a novel mtDNA, the first step in the analysis is the identification of all positions that are different from the reference sequence. Next, with the aim to identify those positions that determine the correct genetic background of the investigated genome, a haplogroup assignment procedure is required. The pipeline available within the MToolBox package fulfils this need (11). It is based on the Phylotree haplogroup classification (23), and care is taken to always implement the latest Phylotree build as soon as it is published; currently the most recent build17 is implemented.

In each *Genome Card* the best predicted haplogroup, obtained by MToolBox, is reported as 'HmtDB Haplogroup Code'. In addition, if available, the authors/sequencing-centre assigned haplogroup is reported as 'Author assigned haplogroup'.

HmtDB data downloads

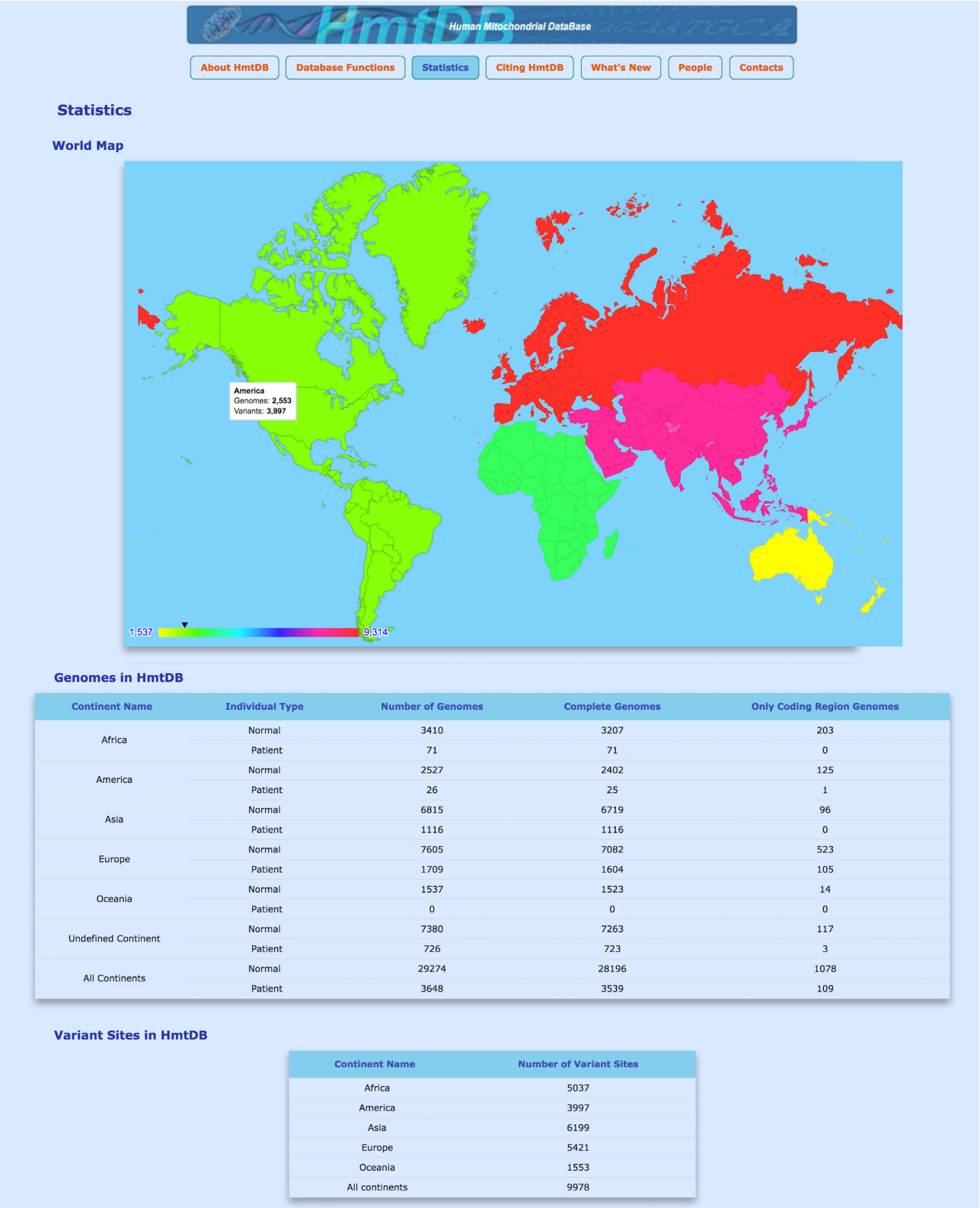
Multi-alignments and variability data regarding the entire data set of 'healthy' or 'pathologic' samples, as well as of continent specific data sets, may be downloaded through the HmtDB *Download* section. Moreover, after a *Query* session, data regarding only the selected genomes, fasta sequences and their multi-alignment may be obtained. In this respect, it is worth mentioning that if the selected sequences are a mix of 'healthy' and 'pathologic' samples, or a subset of a specific continent, a multi-alignment with sites containing only gaps along the entire set may be obtained. Such gaps are insertions belonging to other sequences not selected within the subset. This, however, does not represent a problem for any further analysis where the downloaded multi-alignment is used as input.

Tools for annotation of mtDNA variants

The estimation of the pathogenicity of novel variants is now performed thanks to the availability of links to more powerful resources such as MToolBox and the HmtDB tracks implemented in MSeqDR.

MToolBox and the mtDNA variants prioritization workflow

The processing of mitochondrial data from NGS requires the development of appropriate tools, which take into account the peculiarities of the mitochondrial genomes, including heteroplasmy. To this aim, users can directly gain access to the command-line stand-alone (https://github.com/mitoNGS/MToolBox) (8) and/or web versions (https://mseqdr.org/mtoolbox.php) of the MToolBox pipeline (7,8), through either the *Database Functions* section or the *MToolBox* section in the HmtDB Menu page. It includes



Continent Name	Individual Type	Number of Genomes	Complete Genomes	Only Coding Region Genomes
Africa	Normal	3410	3207	203
	Patient	71	71	0
America	Normal	2527	2402	125
	Patient	26	25	1
Asia	Normal	6815	6719	96
	Patient	1116	1116	0
Europe	Normal	7605	7082	523
	Patient	1709	1604	105
Oceania	Normal	1537	1523	14
	Patient	0	0	0
Undefined Continent	Normal	7380	7263	117
	Patient	726	723	3
All Continents	Normal	29274	28196	1078
	Patient	3648	3539	109

Continent Name	Number of Variant Sites
Africa	5037
America	3997
Asia	6199
Europe	5421
Oceania	1553
All continents	9978

Figure 4. HmtDB statistics. For each data set and sub-data set, the number and the type of HmtDB stored genomes and the number of variant sites are summarized.

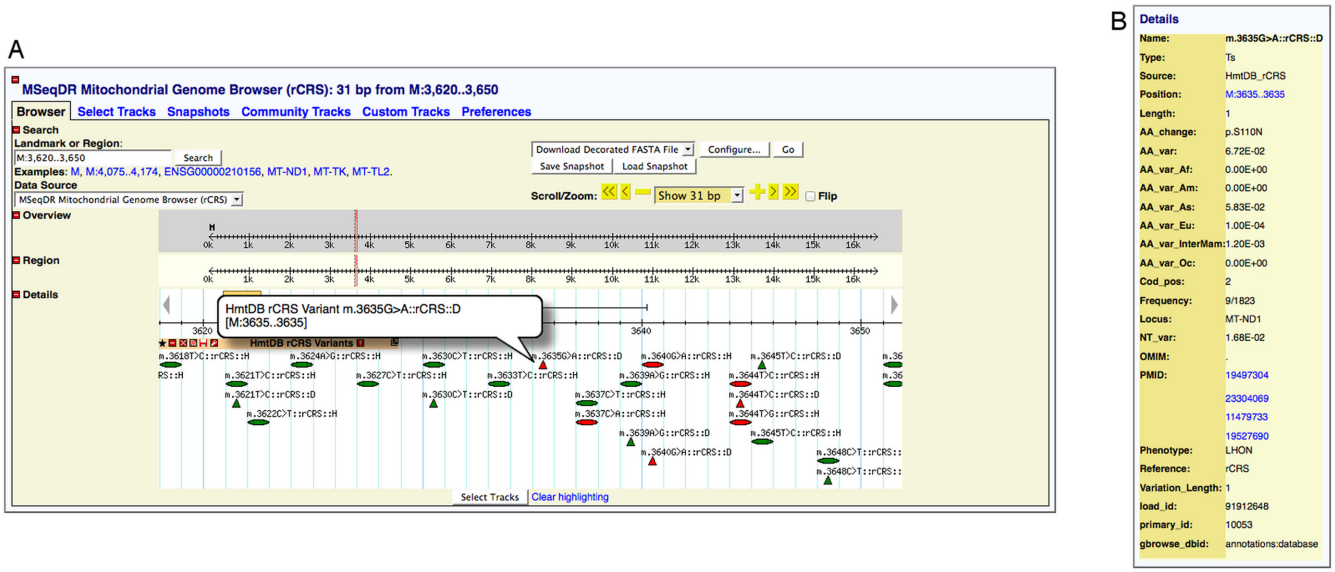


Figure 5. HmtDB tracks in MSeqDR GBrowse. (**A**) The HmtDB track visualized in the GBrowse at MseqDR website, after selecting 'MSeqDR Mitochondrial Genome Browser (rCRS)' as data source. A specific genomic region of interest can be searched showing all the variant sites annotated with respect to one of the two possible reference sequences (in this case, rCRS) in healthy (circles) and patient (triangles) genomes deposited in HmtDB database. Different colors are used to discriminate synonymous (green) from non-synonymous (purple) variants, and variants in protein coding sequences from those in other genomic regions (pink for tRNAs and rRNAs, yellow for D-Loop, here not shown). (**B**) Detailed information can be shown by clicking on a specific variant site. Nucleotide and amino acid variability scores are calculated on complete genomes in HmtDB database. Variability values are reported only for variants in protein coding genes together with the information regarding the amino acid change. All the fields are widely explained in the related README file at the website.

several steps as read mapping and NumtS filtering, post-processing mapping, genome assembly, haplogroup prediction and variant annotation and provides a VCF file with allele-specific heteroplasmy as an output. This pipeline also supports results interpretation, through a prioritization process intended to easily target the mtDNA variants that most likely affect the gene or protein function and recognizing a reliable pool of candidate variants for further investigations (12). Specifically, a phylogeny-based prioritization workflow that compares each variant against the related Macro Haplogroup Consensus Sequences (9) and exploits an extended set of annotation resources. These include variability values, calculated on mitochondrial nucleotide multi-aligned sequences from all the healthy individuals hosted in HmtDB, and *disease scores*, defined as a weighted mean of probabilities that a mutation can be pathogenic estimated by several predictors (12), implemented in the MToolBox pipeline to prioritize the nucleotide variations mostly relevant to the disease.

HmtDB tracks on MseqDR GBrowse

MSeqDR (http://mseqdr.org/) is a central data resource for mitochondrial diseases, supported by the United Mitochondrial Disease Foundation and the National Institutes of Health. Its website is an innovative international genomic data resource for mitochondrial disease community that facilitates the coherent compilation, organization, annotation and analysis of sequence data from both nuclear and mitochondrial genomes (10,30). The MSeqDR website provides access to a wide range of web-based bioinformatics tools, including the GBrowse instance that enables integrated visualization and analysis of heterogeneous variation and other genomic data contained in optimized annotation tracks.

HmtDB data have now been successfully converted to GBrowse tracks, a comprehensive collection of all the variants annotated in complete genomes deposited in the database (17), made available to the general public. Two tracks are generated using the two mitochondrial reference sequences, rCRS (15) and RSRS (16), respectively. The tracks generation is completely automated and its updating goes hand in hand with the update of the HmtDB database. Once HmtDB tracks are selected in the GBrowse, together with other tracks there implemented reporting mtDNA site specific information, a specific genomic region of interest can be searched showing all the variant sites annotated in healthy and patient data sets (Figure 5A). Currently, users are allowed to display an essential annotation of each variation, including the reference used, the phenotype ('healthy' or 'pathologic'), the locus, the amino acid change in protein coding genes, reference SNP (rs) identifiers reported in dbSNP database (31), site-specific nucleotide and amino acid variability scores, calculated for complete healthy genomes stored in HmtDB by applying SiteVar (32) and MitVarProt (32) algorithms, respectively (Figure 5B). Both nucleotide and amino acid variability scores range between 0 and 1: the score is 0 if all the genomes have the same allele in a specific site position, whereas it is > 0 if more alleles are observed in the same site in the whole data set. The higher the number and the frequency of different alleles annotated for a specific site, the higher the score calculated for that position. In this fashion, HmtDB tracks data specifically contribute, in a comparative approach with data reported in the other tracks, to the recognition of private variants or mutations

with a potential pathogenic role, representing a significant tool to prioritize variants and diagnose primary mitochondrial diseases. The whole tracks may also be downloaded in GFF3 (General Feature Format), FASTA or GenBank format.

CONCLUSIONS

A great effort has been made to improve the HmtDB database content and its web interface, and to make it consistent with current standards. The enriched information and search facilities within the current HmtDB portal with a direct link to other mitochondria-focused databases provide today one of the most comprehensive and informative resource on human mtDNA together with other resources such as MITOMAP and Phylotree, whose data are implemented as tracks in the MSeqDR GBROWSE.

ACKNOWLEDGEMENTS

The authors thank the MSeqDR Consortium, particularly dr Marni Falk (The Children's Hospital of Philadelphia) and Dr Xiaowu Gai (Center for Personalized Medicine, Children's Hospital Los Angeles). The HmtDB portal is hosted on the IT resources made available by ReCaS, a project financed by the MIUR (Italian Ministry for Education, University and Research) within the 'PON Ricerca e Competitività 2007–2013-Azione I-Interventi di rafforzamento strutturale' PONa3_00052, Avviso 254/Ric.

FUNDING

FP7 ITN-People Marie Curie Action grant [MEET – Mitochondrial European Educational Training; GA #317433 to G.G.]; The Italian Association for Cancer Research (AIRC) [#IG14242]; Italian Ministry of Health [GR-2013-02356666]; Worldwide Cancer Research UK grant DHOMOS. Funding for open access charge: FP7 ITN-People Marie Curie Action grant [MEET – Mitochondrial European Educational Training; GA #317433 to G.G.]; The Italian Association for Cancer Research (AIRC) [#IG14242]; Italian Ministry of Health [GR-2013-02356666]; Worldwide Cancer Research UK grant DHOMOS.
Conflict of interest statement. None declared.

REFERENCES

1. Wallace,D.C. (2010) Mitochondrial DNA mutations in disease and aging. *Environ. Mol. Mutagen.*, **51**, 440–450.
2. Stewart,J.B. and Chinnery,P.F. (2015) The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.*, **16**, 530–542.
3. Ye,K., Lu,J., Ma,F., Keinan,A. and Gu,Z. (2014) Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 10654–10659.
4. Tang,S. and Huang,T. (2010) Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. *Biotechniques*, **48**, 287–296.
5. Zaragoza,M.V., Fass,J., Diegoli,M., Lin,D. and Arbustini,E. (2010) Mitochondrial DNA variant discovery and evaluation in human Cardiomyopathies through next-generation sequencing. *PloS One*, **5**, e12295.
6. Rossignol,R., Faustin,B., Rocher,C., Malgat,M., Mazat,J.-P. and Letellier,T. (2003) Mitochondrial threshold effects. *Biochem. J.*, **370**, 751–762.
7. Gasparre,G., Kurelac,I., Capristo,M., Iommarini,L., Ghelli,A., Ceccarelli,C., Nicoletti,G., Nanni,P., De Giovanni,C., Scotlandi,K. *et al.* (2011) A mutation threshold distinguishes the antitumorigenic effects of the mitochondrial gene MTND1, an oncojanus function. *Cancer Res.*, **71**, 6220–6229.
8. Payne,B.A.I., Wilson,I.J., Yu-Wai-Man,P., Coxhead,J., Deehan,D., Horvath,R., Taylor,R.W., Samuels,D.C., Santibanez-Koref,M. and Chinnery,P.F. (2013) Universal heteroplasmy of human mitochondrial DNA. *Hum. Mol. Genet.*, **22**, 384–390.
9. Diroma,M.A., Calabrese,C., Simone,D., Santorsola,M., Calabrese,F.M., Gasparre,G. and Attimonelli,M. (2014) Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. *BMC Genomics*, **15**(Suppl. 3), S2.
10. Falk,M.J., Shen,L., Gonzalez,M., Leipzig,J., Lott,M.T., Stassen,A.P.M., Diroma,M.A., Navarro-Gomez,D., Yeske,P., Bai,R. *et al.* (2015) Mitochondrial Disease Sequence Data Resource (MSeqDR): a global grass-roots consortium to facilitate deposition, curation, annotation, and integrated analysis of genomic data for the mitochondrial disease clinical and research communities. *Mol. Genet. Metab.*, **114**, 388–396.
11. Calabrese,C., Simone,D., Diroma,M.A., Santorsola,M., Guttà,C., Gasparre,G., Picardi,E., Pesole,G. and Attimonelli,M. (2014) MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, **30**, 3115–3117.
12. Santorsola,M., Calabrese,C., Girolimetti,G., Diroma,M.A., Gasparre,G. and Attimonelli,M. (2016) A multi-parametric workflow for the prioritization of mitochondrial DNA variants of clinical interest. *Hum. Genet.*, **135**, 121–136.
13. Attimonelli,M., Accetturo,M., Santamaria,M., Lascaro,D., Scioscia,G., Pappadà,G., Russo,L., Zanchetta,L. and Tommaseo-Ponzetta,M. (2005) HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. *BMC Bioinformatics*, **6**(Suppl. 4), S4.
14. 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.
15. Andrews,R.M., Kubacka,I., Chinnery,P.F., Lightowlers,R.N., Turnbull,D.M. and Howell,N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147.
16. Behar,D.M., van Oven,M., Rosset,S., Metspalu,M., Loogväli,E.-L., Silva,N.M., Kivisild,T., Torroni,A. and Villems,R. (2012) A 'Copernican' reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.*, **90**, 675–684.
17. Rubino,F., Piredda,R., Calabrese,F.M., Simone,D., Lang,M., Calabrese,C., Petruzzella,V., Tommaseo-Ponzetta,M., Gasparre,G. and Attimonelli,M. (2012) HmtDB, a genomic resource for mitochondrion-based human variability studies. *Nucleic Acids Res.*, **40**, D1150–D1159.
18. Picardi,E. and Pesole,G. (2012) Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat. Methods*, **9**, 523–524.
19. Consortium, T. 1000 G.P. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
20. Consortium, T. 1000 G.P. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
21. Lott,M.T., Derbeneva,O., Chalkia,D., Sarmady,M., Procaccio,V. and Wallace,D.C. (2013) mtDNA variation and analysis using mitomap and mitomaster. *Curr. Protoc. Bioinformatics*, **44**, doi:10.1002/0471250953.bi0123s44.
22. Ingman,M. and Gyllensten,U. (2006) mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res.*, **34**, D749–D751.
23. van Oven,M. and Kayser,M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, **30**, E386–E394.
24. Torroni,A., Schurr,T.G., Cabell,M.F., Brown,M.D., NeelMeel,J.V., Larsen,M., Smith,D.G., Vullo,C.M., Wallace,D.C., Forster,P., Richards,M. and Bandelt,H.J. (1993) Mitochondrial footprints of human expansions in AfricaAsian affinities and continental radiation

of the four foundingNative American mtDNAs. *Am. J. Hum. Genet.*, **53**, 563–590.
25. Balter,M. (2011) Was North Africa the launch pad for modern human migrations? *Science*, **331**, 20–23.
26. Ghelli,A., Porcelli,A.M., Zanna,C., Vidoni,S., Mattioli,S., Barbieri,A., Iommarini,L., Pala,M., Achilli,A., Torroni,A. *et al.* (2009) The background of mitochondrial DNA Haplogroup J increases the sensitivity of Leber's hereditary optic neuropathy cells to 2,5-Hexanedione toxicity. *PLoS One*, **4**, e7922.
27. Khan,N.A., Govindaraj,P., Jyothi,V., Meena,A.K. and Thangaraj,K. (2013) Co-occurrence of m.1555A>G and m.11778G>A mitochondrial DNA mutations in two Indian families with strikingly different clinical penetrance of Leber hereditary optic neuropathy. *Mol. Vis.*, **19**, 1282–1289.
28. Peng,G.-H., Zheng,B.-J., Fang,F., Wu,Y., Liang,L.-Z., Zheng,J., Nan,B.-Y., Yu,X., Tang,X.-W., Zhu,Y. *et al.* (2013) Mitochondrial 12S rRNA A1555G mutation associated with nonsyndromic hearing loss in twenty-five han Chinese pedigrees. *Yi Chuan*, **35**, 62–72.
29. Zhang,J., Zhao,F., Fu,Q., Liang,M., Tong,Y., Liu,X., Lin,B., Mi,H., Zhang,M., Wei,Q.-P. *et al.* (2013) Mitochondrial haplotypes may modulate the phenotypic manifestation of the LHON-associated m.14484T>C (MT-ND6) mutation in Chinese families. *Mitochondrion*, **13**, 772–781.
30. Shen,L., Diroma,M.A., Gonzalez,M., Navarro-Gomez,D., Leipzig,J., Lott,M.T., van Oven,M., Wallace,D.C., Muraresku,C.C., Zolkipli-Cunningham,Z. *et al.* (2016) MSeqDR: A centralized knowledge repository and bioinformatics web resource to facilitate genomic investigations in mitochondrial disease. *Hum. Mutat.*, **37**, 540–548.
31. Sherry,S.T., Ward,M.-H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
32. Pesole,G. and 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.