**HmtDB tracks**

The HmtDB tracks are a comprehensive collection of all the variants annotated in the Human mitochondrial database HmtDB [1] that contains 32,922 human mitochondrial genomes from healthy people and patients derived from GenBank or submitted by end users. Five continent-specific subsets are identified within the two main datasets. The data contained in the tracks refer only to complete genomes.

**Methods**

An accurate protocol is applied to the healthy and the pathologic datasets as well as their continent-specific subsets aiming to characterize the variability status of each variant site.

A revised version of***SiteVar*** algorithm [2] performed on multialigned genomes is used to calculate site-specific variability score for each position in the multialignment. Site-specific amino acid variability is also estimated for protein encoding genes applying a suitable version of ***MitVarProt*** algorithm [3].

Both nucleotide and amino acid variability scores range between 0 and 1. The score is:

* 0 if all the genomes report the same allele in the considered site
* > 0 if more alleles are observed in the same site.

The higher the number and the frequency of different alleles annotated for a specific site, the higher the score calculated for that position.

HmtDB tracks are generated with respect to both mitochondrial reference sequences, rCRS (revised Cambridge Reference Sequence) and RSRS (Reconstructed Sapiens Reference sequence), differing in 52 positions [4]. Manual curation is done to ensure accuracy.

**Release III: some numbers**

**HmtDB rCRS tracks**: 21525 nucleotide variants within 28196 healthy genomes: 15070 SNPs (10941 transitions, 4050 transvertions, 79 transitions/ transvertions (due to the presence of nucleotide ambiguity)), 640 indels (347 deletions, 293 insertions).

5815 nucleotide variants within 3539 diseased genomes: 5505 SNPs (4405 transitions, 1079 transvertions, 21 transitions/ transvertions (due to the presence of nucleotide ambiguity)), 310 indels (172 deletions, 138 insertions).

**HmtDB RSRS tracks**: 16370 nucleotide variants within 28196 healthy genomes: 15090 SNPs (10956 transitions, 4055 transvertions, 79 transitions/ transvertions (due to the presence of nucleotide ambiguity)), 1280 indels (694 deletions, 586 insertions).

6125 nucleotide variants within 3539 diseased genomes: 5505 SNPs (4405 transitions, 1079 transvertions, 21 transitions/ transvertions (due to the presence of nucleotide ambiguity)), 620 indels (344 deletions, 276 insertions)

GenBank is the reference database for the majority of healthy and patients’ genomes, with the exception of 1427 mitochondrial genomes reconstructed as off-target sequences from exome data generated by the 1000 Genomes Project and 263 diseased genomes directly deposited in HmtDB.

**MSeqDR GBrowse display conventions**

***Name***

Variant name, followed by indication of the reference sequence used (rCRS or RSRS) and the dataset where the variant was annotated (H=healthy, D=disease).

***Variant\_Type***

Event type: Ts=transition, Tv=transversion, ins=insertion, del=deletion.

***Source***

Name of the track (HmtDB\_rCRS or HmtDB\_RSRS).

***Position***

1-based start and end coordinates of the variant in mitochondrial genome or 5’-3’ flanking positions for insertions.

***Score***

Same value of “NT\_variability” field.

***Reference***

Mitochondrial genome reference sequence (rCRS or RSRS).

***Phenotype***

Healthy or Disease. Pathological phenotype is specified for variants annotated exclusively in the patients’dataset.

***Length***

Number of altered bases in the mitochondrial genome.

***Locus***

Mitochondrial gene or region.

***NT\_var***

SiteVar variability value calculated on complete healthy or diseased genomes.

***Cod\_Pos\****

Nucleotide position within the codon.

***AA\_change\****

Type of amino acid change (synonymous or non-synonymous mismatch, STOP-loss, STOP-gain, frameshift or in frame indel).

***AA\_var\****

MitVarProt amino acid variability value calculated on complete healthy genomes.

***AA\_var\_InterMam\****

MitVarProt amino acid variability value for 200 mammals annotated in GenBank.

***AA\_var\_Af\****

MitVarProt amino acid variability value calculated on complete healthy genomes of African origin.

***AA\_var\_Am\****

MitVarProt amino acid variability value calculated on complete healthy genomes of American origin.

***AA\_var\_As\****

MitVarProt amino acid variability value calculated on complete healthy genomes of Asian origin.

***AA\_var\_Eu\****

MitVarProt amino acid variability value calculated on complete healthy genomes of European origin.

***AA\_var\_Oc\****

MitVarProt amino acid variability value calculated on complete healthy genomes of Oceania origin.

***AN***

Link to accession number of each genome in GenBank harboring the variant. PA\_\* codes correspond to diseased genomes submitted only in HmtDB.

***PMID***

Link to PubMed IDs of relative publications.

***ClinVar***

ClinVarID of the mutation

***dbSNP***

dbSNP ID of the mutation.

***OMIM***

OMIM ID of the mutation.

***gbrowse\_dbid***

annotations:database.

***AN***

Link to accession number of each genome in GenBank harboring the variant. PA\_\* codes correspond to diseased genomes submitted only in HmtDB.

***PMID***

Link to PubMed IDs of relative publications.

\*only for variants in protein encoding genes

**Credits**

These data were provided by Maria Angela Diroma (Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Italy), Rosanna Clima (Department of Medical and Surgical Sciences, University of Bologna, Italy) and Marcella Attimonelli (Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Italy).

**References**

[1] 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-calculated human variability studies. Nucleic Acids Res, 40, D1150-1159. [www.hmtdb.uniba.it](http://www.hmtdb.uniba.it)

[2] 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.

[3] Horner, D.S. and Pesole, G. (2003) The estimation of relative site variability among aligned homologous protein sequences. Bioinformatics, 19, 600-606.

[4] Behar, D.M., van Oven, M., Rosset, S., Metspalu, M., Loogvali, 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.