Designing And Implementing A Novel Database For Human Mitogenomes

Hrant Hovhannisyan
Institute of Molecular Biology, NAS, Yerevan, Armenia
grant.hovhannisyan@gmail.com

Abo Yesayan
The Institute for Informatics and Automation Problems, NAS, Yerevan, Armenia
abo.yesayan@gmail.com

Levon Yepiskoposyan
Institute of Molecular Biology, NAS, Yerevan, Armenia
lepiskop@yahoo.com

ABSTRACT

In consequence of intensive human population genetics and forensic studies that take place nowadays, human mitochondrial genome data are continuously accumulating, while the effective methods of this massive information storing and management lag behind. For instance, publicly available databases of human mitochondrial DNA (mtDNA) are either not updated regularly or lack functional tools for appropriate data parsing. In this work, we introduce the mtMart – a novel manually curated database for complete human mitochondrial genomes that significantly facilitate effective managing of large-scale genomic data.

Keywords

human mitochondrial genome, forensic science, population genetics, database, mtMart

1. INTRODUCTION

Today, the advances of novel DNA sequencing methods made the mitochondrial genome a versatile tool for phylogenetics, population genetics, forensic science and other disciplines [1, 2]. Despite the human mitochondrial DNA is a tiny molecule ca. 16,570 b.p. [3], continuously accumulating large-scale mtDNA data make problematic the handling, analyzing and comparing the mtDNA gene pools of different human populations. Recently, several attempts were made to design and create publicly available human mitochondrial DNA databases, however, some of them are no longer updated and maintained, while the rest do not provide convenient functionality for effective data management and further analysis. Thus, the aim of our project was to develop new functionally rich, user-friendly and regularly updated database for increasing the effectiveness of management of human mitogenomic data.

2. RELATED WORK

Several major databases were created for human mitochondrial DNA data storage in past decade. For instance, HvrBase++ [4] and mtDB [5] databases were launched in 2006, but have not been updated since 2007, while the number of new mtDNA partial and complete sequences has increased significantly since then. Besides of being outdated, these databases did not have appropriate functional characteristics for managing mtDNA data. For example, lack of the data on human mtDNA haplogroup and sorting/grouping functions were restricting the usage of HvrBase++ and mtDB. Additionally, today when next-generation technologies allow massive sequencing of entire mitochondrial genomes on population scale, the data on hyper-variable region (HVR) 1 and 2 of mtDNA, that were available in these databases, do not fulfill the demands of nowadays population genetics.

On the other hand, database hmtDB, has numerous options for complex data searching – it has a powerful querying system, where user can search data according to mutated positions, haplogoups, geographic regions, tissue, sex, etc; mtDNA haplogorup assignment tool, convenient downloading function and besides containing the data on HVR 1 and 2 hmtDB is designed also for storing the data on complete mitochondrial genomes. However, it does not contain sequences obtained after 2013, which also does not allow researchers to handle all the available mtDNA data.

Phylotree [6], being the reference database which defines the mtDNA haplogroup nomenclature, is updated regularly every 1-2 years, and today the last built (built 16, 19 February 2014) of Phylotree contains 20666 complete published and unpublished human mitochondrial genomes. Nevertheless, the only functional feature of Phylotree is the ability to download the row data of sequences in *.fasta* format according to the published papers, where the data were first described. Hence, the above described features of Phylotree make it powerful for finding the source of row data, but not for data parsing and manipulating.

Mitomap [7] is another regularly updated human mitochondrial genome database, which not only stores mtDNA data, but also contains useful information about mitochondrial DNA and mitochondria in general, i.e., information on mitochondrial polypeptides, references, useful tools, figures, maps, statistics, etc. However, the main disadvantage of Mitomap is also the lack of convenient functionality for data searching and downloading in different formats, which restricts the efficacy of data handling by Mitomap.

Moreover, so far there is no accurately curated human mtDNA database that provides precise information about high-resolution mtDNA haplogroup, ethnic group the subject belongs to and the geographic location of the population, while this information plays a crucial role in human population genetic studies.

Here, we introduce the mtMart (by analogy with BioMart search engine of Ensemble genome browser [8]) – a new database for human complete mitochondrial DNA data, which is designed to fill the gaps of the above described databases and implement our own ideas in order to significantly facilitate the effective treatment of large-scale human mitochondrial genome information.

3. IMPLEMENTATION AND WORKFLOW

The database is designed with MySQL open-source relational database management system. Its functional characteristics are implemented in PHP and JavaScript (jQuery) programming languages.

mtMart retrieves the information on human mitochondrial genomes (however, in principle it can be used for retrieving any query) directly from the National Center of Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/), which provides the API (Application Programming Interface), using the arbitrary set or range of accession numbers of interest. Using in-home PHP script to process the INSDSeq XML file, which is generated from GenBank (http://www.ncbi.nlm.nih.gov/genbank) and contains all the information about the query, mtMart stores the obtained data in internal memory, storing it in columns by the accession number, complete mitochondrial genome in FASTA format, size of the molecule in base pairs and the reference, where the molecule was first described.

One of the main features of mtMart is the possibility to semi-automatically add the information about mtDNA haplogroup, mutation data compared to revised Cambridge Reference Sequence, ethnic group and geographic region of the population to any record of the database. For assigning the haplogroup and defining mutated positions, mtMart is synchronized with Haplofind [9] – fast and reliable web application for high-throughput human mitochondrial genome haplogroup assignment, which is based on the most recent Phylotree built. Besides haplogroup assignation, Haplofind also defines the mitochondrial DNA mutations that were reported to be associated with the particular disorders, and this data is also added to mtMart. The data on population and geographic region are appended to mtMart manually, and we believe this approach ensures precision and high quality of the data stored in the database.

In order to avoid some undesired manipulations of users, the functions of addition and removal of the data are restricted for all users, except the Administrator.

Another important feature of mtMart is a possibility to search, sort and download data in a very customizable way. One can sort it either by haplogroup (at any resolution of phylogenetic tree), population and geographic region or combine all these options in order to obtain a necessary result. After obtaining it, the information can be retrieved in several output formats. For complete sequences and mutation data the mtMART outputs the information into FASTA and FASTA-like formats, respectively, since some widely used software for mitochondrial DNA data analysis, such as MITOTOOL [10] use FASTA-like format composed of mutated sites only, instead of DNA sequences. On the other hand, for downloading the data from sortable columns (haplogroup, population, region) and accession number the database allows to output the result in commonly used text formats (tab, comma, colon, etc.-separated files) with the arbitrary order of columns. Moreover, for haplogroup and population or geographic data we have implemented the algorithm, allowing to automatically generate the *.arp* Arlequin [11] input file with relative haplogroup frequency values.

4. USE CASE

For illustrative purpose we have performed the study on the Iranian and Armenian mitogenomic diversity. Two types of analyses were done using mtMART. First, by our database we have selected the data according to the "region", i.e. Armenia and Iran. Then using the function of automatic Arlequin input file generation, we have created the *.arp* file that includes the frequency of haplogroups for each of the populations. Notable, that maMART allows to automatically calculate these frequencies at any resolution of mtDNA phylogenetic tree instead of manual unification of haplogroup depth of different samples. The process of *.arp* file generation by mtMART takes seconds, while usually to create this type of input file manually for multiple populations might take hours. Consequently, it is used for calculation of different population genetics parameters using Arlequin software. In this case, we have calculated Fst genetic distance between the studied populations based on the lowest haplogroup resolution and performed Fisher's exact test of population differentiation, and as it was expected Iranian and Armenian populations on the low resolution haplogroup level do not differ from each other significantly ($p>0.05$). Secondly, we have automatically downloaded the above mentioned data in *.fasta* format using mtMART and processed it using DNASP5 [12] software by calculating genetic diversity index *h*. The obtained results show that the Iranian mitochondrial gene pool has higher value of genetic diversity than the Armenian ones (0.99 and 0.98, respectively).

In this demonstrative study we have shown that using mtMART and its tools one can significantly shorten the time

of mitochondrial population genomic analysis and make it more convenient to perform.

5. CONCLUSION

In our project, we have developed a new user-friendly database for complete human mitochondrial genomes mtMART which might significantly support large-scale human mitogenomic studies. mtMART is built taking into account the conveniences of previously designed mtDNA databases and filling the gaps of the last. Our database will be continuously developing, allowing researchers to use new functionally convenient features for fast and effective data management.

Availability

mtMART is publicly available at http://genebank.hol.es/search.php?search_val=all, but still under development.

REFERENCES

[1] Bandelt, H. J., Richards, M., & Macaulay, V. (2006). *Human mitochondrial DNA and the evolution of Homo sapiens* (Vol. 18). Springer.

[2] Ca1valli-Sforza, L. Luca, and Marcus W. Feldman. "The application of molecular genetic approaches to the study of human evolution." *Nature Genetics* 33 (2003): 266-275.

[3] Anderson, Sharon, Alan T. Bankier, Bart G. Barrell, M. H. L. De Bruijn, Alan R. Coulson, Jacques Drouin, I. C. Eperon et al. "Sequence and organization of the human mitochondrial genome." (1981): 457-465.

[4] Kohl, Jochen, Ingo Paulsen, Thomas Laubach, Achim Radtke, and Arndt von Haeseler. "HvrBase++: a phylogenetic database for primate species." *Nucleic acids research* 34, no. suppl 1 (2006): D700-D704.

[5] Ingman, Max, and Ulf Gyllensten. "mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences." *Nucleic acids research* 34, no. suppl 1 (2006): D749-D751.

[6] Van Oven, Mannis, and Manfred Kayser. "Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation." *Human mutation* 30, no. 2 (2009): E386-E394.

[7] Ruiz-Pesini, Eduardo, Marie T. Lott, Vincent Procaccio, Jason C. Poole, Marty C. Brandon, Dan Mishmar, Christina Yi, James Kreuziger, Pierre Baldi, and Douglas C. Wallace. "An enhanced MITOMAP with a global mtDNA mutational phylogeny." *Nucleic acids research* 35, no. suppl 1 (2007): D823-D828.

[8] Flicek, Paul, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva et al. "Ensembl 2014." *Nucleic acids research* (2013): gkt1196.

[9] Vianello, Dario, Federica Sevini, Gastone Castellani, Laura Lomartire, Miriam Capri, and Claudio Franceschi. "HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment." *Human mutation* 34, no. 9 (2013): 1189-1194.

[10] Fan, Long, and Yong-Gang Yao. "MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations." *Mitochondrion* 11, no. 2 (2011): 351-356.

[11] Excoffier, Laurent, and Heidi EL Lischer. "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows." *Molecular ecology resources* 10, no. 3 (2010): 564-567.

[12] Librado, P., & Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*(11), 1451-1452.