Integration of genomic variation and phenotypic data using HmtPhenome

Preste R.[1], Attimonelli M.[1]

1. Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Bari 70126, Italy

Abstract

A full understanding of relationships between variants, genes, phenotypes and diseases is often overlooked when investigating mitochondrial functionality in both healthy and pathological situations. Gaining a comprehensive overview of this network can indeed offer interesting insights, and guide researchers and clinicians towards a full-spectrum knowledge of the mitochondrial system.

Given the current lack of tools addressing this need, we have developed HmtPhenome (https://www.hmtphenome.uniba.it), a new web resource that aims at providing a visual network of connections among variants, genes, phenotypes and diseases having any level of involvement in the mitochondrial functionality. Data are collected from several third party resources and aggregated on the fly, allowing users to clearly identify interesting relations between the involved entities. Tabular data with additional hyperlinks are also included in the output returned by HmtPhenome, so that users can extend their analysis with further information from external resources.

Introduction

The relationship between mitochondria and disease has gained a lot of interest by clinicians and researchers during recent years, due to the involvement of the mitochondrion in several biological processes, the most important being aerobic ATP production. Defective mitochondria can be, directly or indirectly, associated to the onset and progression of neurodegenerative diseases[1–3], diabetes[4], cancer[5,6], metabolic syndromes[7] and other pathologies, with a broad spectrum of phenotypic traits and outcomes.

Moreover, human cells present a high number of mitochondria, each one of which can contain thousands of copies of mtDNA with a variable ratio of wild-type/mutated genomes, a condition defined as mitochondrial heteroplasmy. Aberrant mitochondrial phenotypes can only arise when the number of mutated genomes is sufficient to take over the wild type ones and exert their detrimental effects. This can explain the pathologic heterogeneity usually observed in mitochondrial diseases: the same disease can show slightly different phenotypes among different individuals, based on the heteroplasmic fraction of the given causative mutation(s). Recent clinical literature has shown several examples of pathologies characterised by different levels of severity or phenotypic effects due to variable mitochondrial heteroplasmic fraction[8–10].
Furthermore, mitochondrial dysfunction can also occur when mutations are located in nuclear-encoded genes whose functionality affects that of mitochondria. More than 1000 genes exist that are coded by the nuclear genome but are involved in mitochondrial physiology, and most of them are implicated in energy production processes at various levels. A comprehensive list of such genes is maintained by Mitocarta[11], which provides an index of 1158 genes involved in mitochondrial activities.

Many online resources offer data about mitochondrial variations and their involvement in pathological conditions, such as Mitomap[12], LOVD[13], HmtDB[14], HmtVar[15], Mitobreak[16], MitImpact[17]. However, while these softwares extensively analyse mitochondrial mutations at different levels, from large genomic rearrangements to SNPs, they usually only report diseases related to a specific variation, lacking to properly highlight potential relationships between variations and their phenotypic effects. These information can indeed be of much help for researchers and clinicians to obtain a much broader view of the studied subject.
An additional source of difficulty that is frequently encountered when investigating phenotypes and diseases is represented by the different ways in which the same condition can be potentially referred to. Phenotype and disease names can differ over distinct resources, and the actual cut-off line between these two entities is often blurred, confusing phenotypic effects with disease causing them, and vice-versa. This issue was effectively addressed by ontology services, which offer standardised vocabularies that unambiguously identify entities and relationships among them[18]. Terms are arranged in a hierarchical manner, with terms referring to much broader concepts located as top nodes, and more specific elements can be found traversing this tree-like structure until reaching the final leaves; each node usually contains other additional information about the described element, with links pointing to further external resources. The Human Phenotype Ontology (HPO)[19] and the Disease

Ontology (DO)[20] are probably the most commonly used examples of such schema as related to phenotypes and diseases, respectively, but many other similar services exist, like the Mammalian Phenotype Ontology (MPO)[21], PhenomeNET[22], Medical Subject Headings (MeSH)[23] and the Unified Medical Language System (UMLS)[24].

In order to better understand the pathological mechanisms of mitochondrial syndromes and diseases, it is then fundamental to build a comprehensive network of diseases and related phenotypes directly or indirectly associated to mitochondrial mutations; along with them, nuclear and mitochondrial genes with some level of involvement in mitochondrial functionality should be considered as well.
For this reason we developed HmtPhenome (https://www.hmtphenome.uniba.it), a system that integrates information about variants, genes, phenotypes and diseases associated to mitochondrial functions, and allows to perform queries that can start from any one of these entry points, retrieving data related to the other three entities and building a full-fledged information network.

Materials and methods

The aim of HmtPhenome is to provide information about four different biological entities and the relationship among them, namely variants, genes, phenotypes and diseases with any involvement in mitochondrial functionality. In order to do this, HmtPhenome retrieves the relevant data from several online resources, interconnects these information when possible and returns a network that highlights relations of interest for researchers and clinicians.

Due to the high number of resources involved and the great amount of data considered in each query, a classic data retrieval system using a local database would have not been suitable for these purposes; instead, all the needed information is collected from external resources on the fly after a user query, and only a small amount of data is actually stored in a local database. The local database is mostly used by the web interface to populate query menus and provide autocompletion suggestions based on user input, thanks to the Awesomplete Javascript framework[25]. This set of locally-stored information contains:

- the list of 1158 mitochondrial- and nuclear-encoded genes involved in mitochondrial functionality, collected from Mitocarta and used in the Gene query section; this list is further extended with the 22 mitochondrial tRNA genes, totalling 1180 genes;
- a list of diseases collected from OMIM[26] and Orphanet[27], used to provide suggestions in the Disease query section; this basic list is further enriched using data coming from DisGeNET[28], to integrate additional information about variants and genes associated with each disease;
- a list of phenotypes, collected from HPO and used to provide suggestions in the Phenotype query section; additional data are also collected from HPO with details about the association of these phenotypes with specific diseases and genes.

All the other information related to variants, genes, phenotypes and diseases and their mutual relationships are gathered from third-party resources, exploiting their APIs, such as HPO, the Experimental Factor Ontology (EFO)[29], Ensembl REST services[30] and BioMart[31].

Having to deal with a large amount of data flowing in from several external services, there is the risk that the system may face long-running tasks when waiting for their response, and thus users could experience delays while using HmtPhenome. This possibility was taken into account and addressed using two complementary strategies: first of all, HmtPhenome was built using the Quart Python framework[32], which is specifically suited to handle all the requests to external services separately and in parallel, instead of queueing them one after the other, reducing waiting times sensibly. In addition, in case these third-party resources take too long or fail to provide a valid response, a fallback set of basic information retrieved from the local database is returned, so that a minimum amount of data is always available in the query results.

The final data, collected from both the local database and external resources, are aggregated to create a dictionary-like structure, where the keys are represented by the above-mentioned four different biological entities, i.e. disease, genes, variants and phenotypes, and the value associated to each key is a list of all the available information in that context. The key referring to the query starting point will obviously contain a single element, the query subject, with the only exception being when the query starts from a variant position, in which case multiple variants can be found on the same genomic position. This dictionary structure is rendered as both a visual network and a tabular form, thanks to the vis.js[33] and DataTables[34] Javascript frameworks, respectively. In addition, a JSON-formatted version of the same data is also provided, for users that may want to perform further downstream analyses.

Results and discussion

HmtPhenome offers a powerful Query page, where users can select which biological entity they want to focus on, and retrieve all the available information about it.
Queries on HmtPhenome are organised on different levels, and can start from one of these entry points:

- variant position;
- gene;
- phenotype;
- disease.

In general, when a query is launched on HmtPhenome, a basic set of information about the chosen biological entity is first retrieved from the local database, depending on the specific entry point; then, API requests are performed to the relevant external services, in order to collect as many additional data as possible.
All these information are finally integrated to build the final network shown on the Results page, where every different piece of information available is connected to another according

to their biological relationships; variants, genes, phenotypes and diseases are shown in different colours to facilitate the visual recognition of interesting relations and patterns in this network. The same data are also shown as a table listing all the different variants, genes, phenotypes and diseases found, with links pointing to external resources with more information about each of them. A JSON-formatted text with the same data is also available for further manipulation.

For queries starting from a variant position, users will first have to select a specific chromosome to restrict the search; after that, users can type a specific position to investigate on that chromosome. After the query is launched, variants found are connected to the gene they belong to (if applicable), to the diseases they are involved in and to the phenotypes they cause or the phenotypes that may cause the disease, when these information are available (Fig. 1).

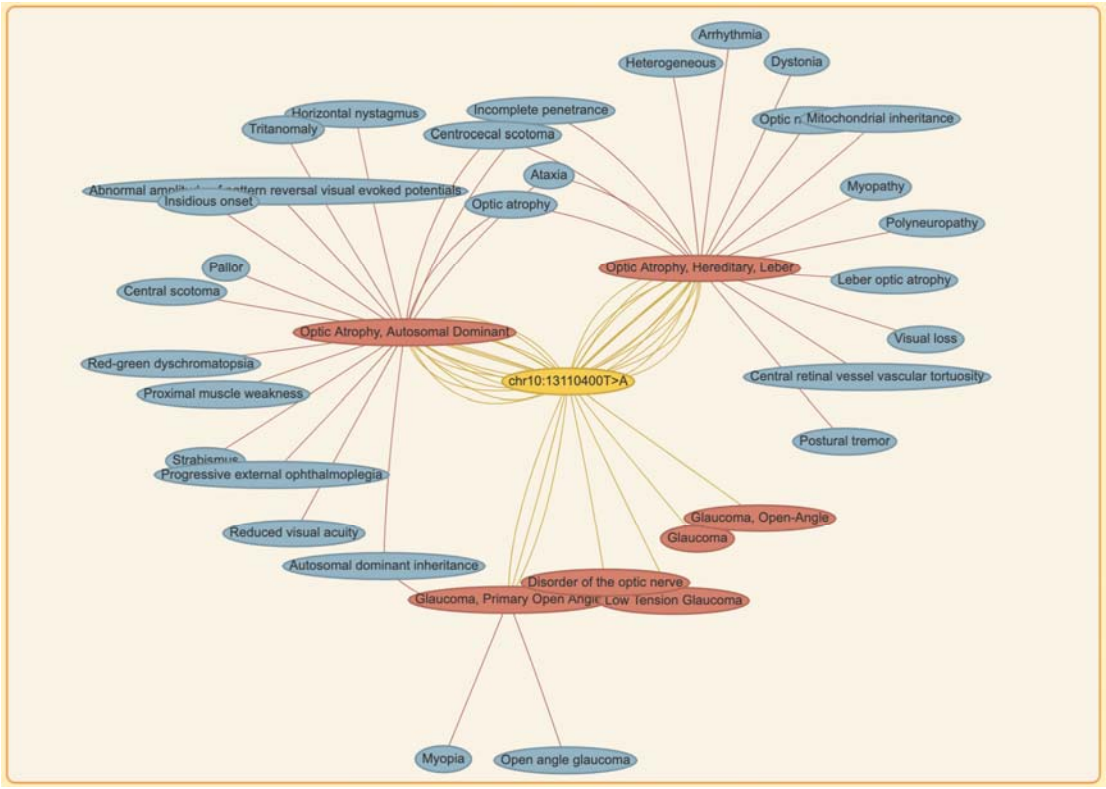

Fig. 1. Network returned after a query by variant position (chr10:13110400T>A).
Yellow nodes: variants, green nodes: genes (not shown here), red nodes: diseases, blue nodes: phenotypes.

If users want to investigate a specific gene, they can insert the desired Ensembl Gene ID or start typing the gene name in the related text search area, and a few suggestions from the list of Mitocarta genes will be shown to facilitate the search. Query results report variants associated with the chosen gene, diseases and phenotypes related to these variants as well as

those related to the queried gene, when further information about involved variants is not available (Fig. 2).

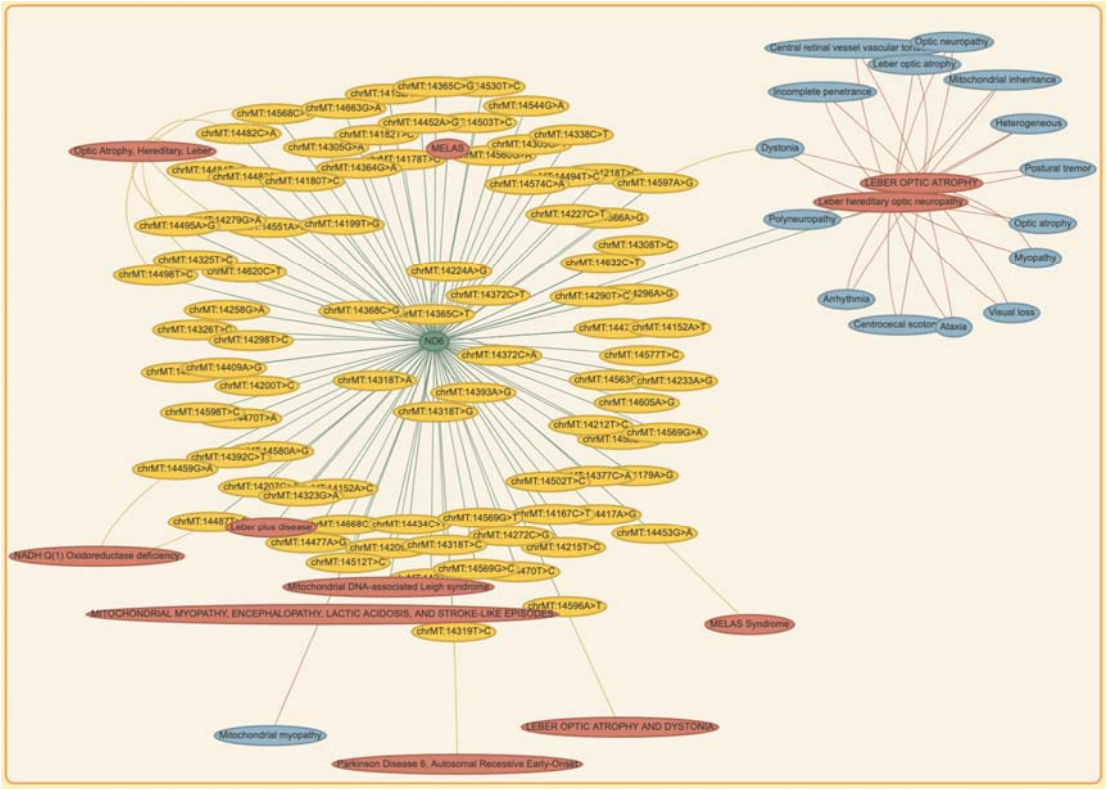

Fig. 2. Network returned after a query by gene name (ND6, Ensembl Gene ID ENSG00000198695). Yellow nodes: variants, green nodes: genes, red nodes: diseases, blue nodes: phenotypes.

Queries involving diseases also present an open text search, which can accept either Orphanet or OMIM identifiers. As an alternative, users can also start typing the desired common disease name, and a set of related suggestions will show up to guide the user; when a suggested element is selected, the associated identifier will be typed in automatically. Disease queries return information about genes and variants known to be involved in the given disease onset or progression, as well as phenotypes caused by or somewhat associated to it (Fig. 3). The query disease name is also converted to its proper ontology term, according to the UMLS service.

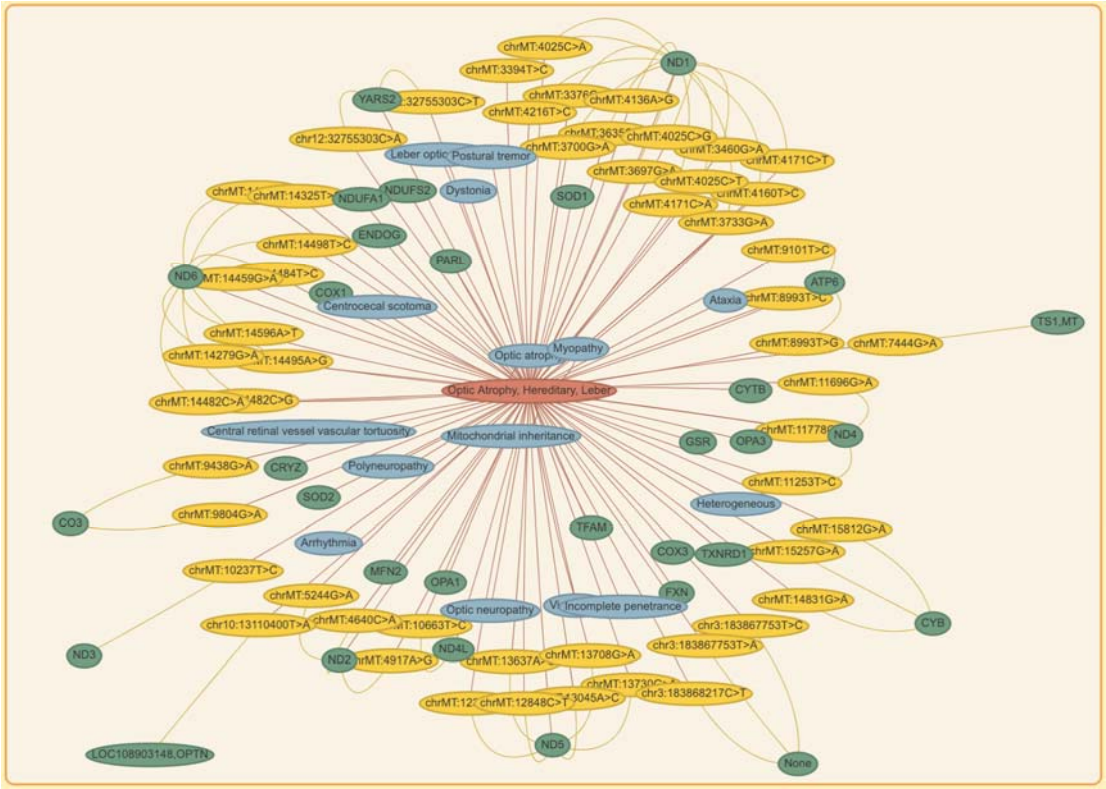

Fig. 3. Network returned after a query by disease name (Leber Optic Atrophy, OMIM:535000). Yellow nodes: variants, green nodes: genes, red nodes: diseases, blue nodes: phenotypes.

The query field for phenotypes also features an open text search with autocompletion, but in this case it accepts HPO identifiers related to phenotype names. Query results contain diseases characterised by the given phenotype, together with genes and variants that are related with it (Fig. 4).

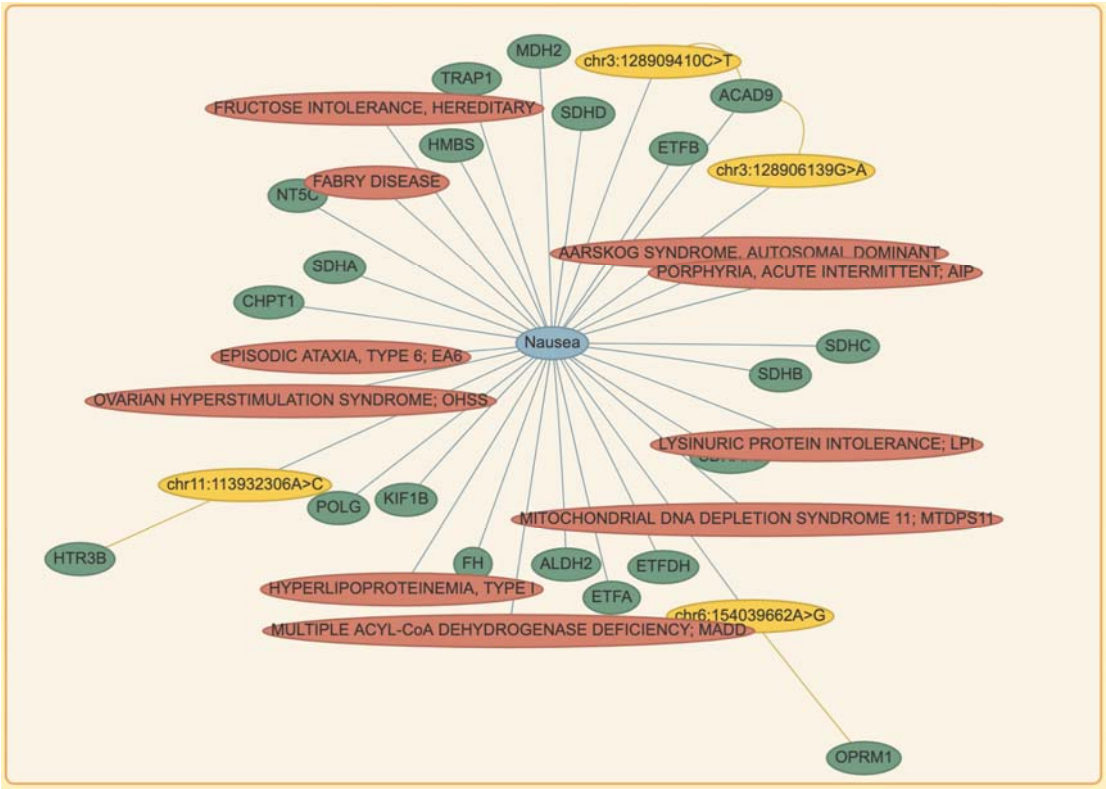

Fig. 4. Network returned after a query by phenotype name (Nausea, HP:0002018).
Yellow nodes: variants, green nodes: genes, red nodes: diseases, blue nodes: phenotypes.

Conclusions

HmtPhenome (https://www.hmtphenome.uniba.it) offers a comprehensive view of relationships between variants, genes, phenotypes and diseases with a particular involvement in mitochondrial functionality. These data can be extremely useful for researchers and clinicians looking to integrate information coming from different resources and focusing on separate aspects of physiological and pathological mitochondria.

Given the high number of resources involved and the impressive amount of data generated and manipulated by HmtPhenome, a great effort was made to ensure this system would work seamlessly and efficiently, through the usage of state-of-the-art programming frameworks and a set of default fallback data. The most useful online resources with data about variants, gene, diseases and phenotypes are scanned to gather information about these biological entities, and these information are then thoroughly aggregated and returned to the user both in graphical and textual form. This allows to identify and further investigate existent biological relationships at a glance, as well as finding potential new players involved in mitochondrial functionality to some extent.

Although the current implementation of HmtPhenome queries a limited number of third-party resources, the returned set of information is extensive and serves efficiently its purpose of providing a wide overview of the relationships existing among variants, genes, diseases and phenotypes, particularly as regarding mitochondrial functionality. More external resources will be integrated into this system with the upcoming updates, with particular attention to mitochondria-focused ones, in order to increase the amount of data processed and augment the added value of the information integration performed by HmtPhenome.

References

1. Perez Ortiz, J. M. & Swerdlow, R. H. Mitochondrial dysfunction in Alzheimer's disease: Role in pathogenesis and novel therapeutic opportunities. *Br. J. Pharmacol.* (2019). doi:10.1111/bph.14585
2. Juarez-Flores, D. L., Gonzalez-Casacuberta, I. & Garrabou, G. Mitohormesis and autophagic balance in Parkinson disease. *Aging* (2019). doi:10.18632/aging.101779
3. Rango, M. & Bresolin, N. Brain Mitochondria, Aging, and Parkinson's Disease. *Genes* **9**, (2018).
4. Williams, M. & Caino, M. C. Mitochondrial Dynamics in Type 2 Diabetes and Cancer. *Front. Endocrinol.* **9**, 211 (2018).
5. Emmings, E. *et al.* Targeting Mitochondria for Treatment of Chemoresistant Ovarian Cancer. *Int. J. Mol. Sci.* **20**, (2019).
6. Németh, K. *et al.* Next-generation sequencing identifies novel mitochondrial variants in pituitary adenomas. *J. Endocrinol. Invest.* (2019). doi:10.1007/s40618-019-1005-6
7. Jha, S. K., Jha, N. K., Kumar, D., Ambasta, R. K. & Kumar, P. Linking mitochondrial dysfunction, metabolic syndrome and stress signaling in Neurodegeneration. *Biochim. Biophys. Acta Mol. Basis Dis*. **1863**, 1132–1146 (2017).
8. Stefano, G. B., Bjenning, C., Wang, F., Wang, N. & Kream, R. M. Mitochondrial Heteroplasmy. *Adv. Exp. Med. Biol*. **982**, 577–594 (2017).
9. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).
10. Wallace, D. C. & Chalkia, D. Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. *Cold Spring Harb. Perspect. Biol.* **5**, a021220 (2013).
11. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* **44**, D1251–D1257 (2016).

12. Lott, M. T. *et al.* mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinforma.* **44**, 1.23.1-26 (2013).

13. Fokkema, I. F. A. C. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* **32**, 557–563 (2011).

14. Clima, R. *et al.* HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. *Nucleic Acids Res.* **45**, D698–D706 (2017).

15. Preste, R., Vitale, O., Clima, R., Gasparre, G. & Attimonelli, M. HmtVar: a new resource for human mitochondrial variations and pathogenicity data. *Nucleic Acids Res.* **47**, D1202–D1210 (2019).

16. Damas, J., Carneiro, J., Amorim, A. & Pereira, F. MitoBreak: the mitochondrial DNA breakpoints database. *Nucleic Acids Res.* **42**, D1261-1268 (2014).

17. Castellana, S., Rónai, J. & Mazza, T. MitImpact: an exhaustive collection of pre-computed pathogenicity predictions of human mitochondrial non-synonymous variants. *Hum. Mutat.* **36**, E2413-2422 (2015).

18. Gkoutos, G. V., Schofield, P. N. & Hoehndorf, R. The anatomy of phenotype ontologies: principles, properties and applications. *Brief. Bioinform.* **19**, 1008–1021 (2018).

19. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).

20. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).

21. Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399 (2009).

22. Rodríguez-García, M. Á., Gkoutos, G. V., Schofield, P. N. & Hoehndorf, R. Integrating phenotype ontologies with PhenomeNET. *J. Biomed. Semant.* **8**, 58 (2017).

23. MeSH Browser. Available at: https://meshb.nlm.nih.gov/search. (Accessed: 15th February 2019)

24. Unified Medical Language System (UMLS). Available at: https://www.nlm.nih.gov/research/umls/. (Accessed: 15th February 2019)

25. Awesomplete: Ultra lightweight, highly customizable, simple autocomplete, by Lea Verou. Available at: https://leaverou.github.io/awesomplete/. (Accessed: 21st February 2019)

26. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. **43**, D789-798 (2015).

27. RESERVED, I. U.--A. R. Orphanet. Available at: http://www.orpha.net/consor/www/cgi-bin/index.php?lng=EN. (Accessed: 21st February 2019)

28. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).

29. Malone, J. *et al.* Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics* **26**, 1112–1118 (2010).

30. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

31. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database J. Biol. Databases Curation* **2011**, bar030 (2011).

32. Quart documentation — Quart 0.8.1 documentation. Available at: https://pgjones.gitlab.io/quart/. (Accessed: 21st February 2019)

33. vis.js - A dynamic, browser based visualization library. Available at: http://visjs.org/#.

(Accessed: 21st February 2019)

34. DataTables | Table plug-in for jQuery. Available at: https://datatables.net/. (Accessed: 21st February 2019)