Human mitochondrial variant annotation with HmtNote

Preste R.[1], Clima R.[2], Attimonelli M.[1]

1. Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Bari 70126, Italy
2. Department of Medical and Surgical Sciences - DIMEC, Medical Genetics Unit, University of Bologna, Bologna 40126, Italy

Abstract

HmtNote is a Python package to annotate human mitochondrial variants from VCF files.

Variants are annotated using a wide range of information, which are grouped into basic, cross-reference, variability and prediction subsets so that users can either select specific annotations of interest or use them altogether.

Annotations are performed using data from HmtVar, a recently published database of human mitochondrial variations, which collects information from several online resources as well as offering in-house pathogenicity predictions.

HmtNote also allows users to download a local annotation database, that can be used to annotate variants offline, without having to rely on an internet connection.

HmtNote is a free and open source package, and can be downloaded and installed from PyPI (https://pypi.org/project/hmtnote) or GitHub (https://github.com/robertopreste/HmtNote).

Introduction

Recent Next Generation Sequencing (NGS) techniques allow researchers to collect an impressive amount of genomic information, particularly regarding genome variability. This is especially true for mitochondrial genomic research, which has seen a rapid increase in interest during the last few years, mostly due to the historically overlooked central role of the mitochondrion in many biological processes and pathological situations[1]. The presence of specific mitochondrial variants can offer useful insights on different levels, from population and evolution studies to pathogenicity assessment, and the functional analysis of human mitochondrial variations is a very florid research topic[2].

Mitochondrial variants annotation, however, can be tricky to perform correctly, given the issue of mitochondrial heteroplasmy, a situation where a variable ratio of wild-type/variant mitochondrial genomes can be present in the same mitochondria, and the fact that some of the variants may represent benign variation only if related to the sample haplogroup; this may lead to uncertainty about which variants should be investigated further, and what effect they may exert in the living organism. Currently, a few tools for mtDNA variant annotation are available: MSeqDR mvTool[3], SG-ADVISER mtDNA[4], mtDNA-Server[5], Mitomaster[6], MitoTool[7], MitImpact[8]; some of these resources, however, are not timely updated, and some others use a relatively narrow training set of mitochondrial genomes, so their annotations can be less reliable. A recently published resource, HmtVar[9], offers variability and pathogenicity data about variants collected from over 45000 human mitochondrial genomes, thus rendering this information very useful to annotate variants found in a mitochondrial sample with a high confidence.

Here we propose a new tool for human mitochondrial variants annotation, HmtNote (https://github.com/robertopreste/HmtNote), which exploits the great amount of data available on HmtVar to enrich a given VCF file with functional annotations.

Materials and methods

HmtNote aims at providing a simple yet efficient tool to annotate a list of human mitochondrial variants. It is built on Python 3 and functions as both an importable Python module and a command line interface (CLI) tool, so that it can easily be integrated into any NGS data analysis pipeline.

HmtNote operates on an input VCF file with human mitochondrial variants, which can be produced from any mitochondrial variant calling software; variants should be called with respect to the rCRS[10], since HmtVar relies on this mitochondrial reference sequence. Every variant is looked for on HmtVar, exploiting its API, in order to retrieve all the available data. Annotations are distinguished into four annotation groups, which are detailed in Table 1, depending on the type of information they provide:

- basic data about a variant (locus, aminoacid change, pathogenicity, disease score, HmtVar ID);
- cross-reference information (Clinvar[11] ID, dbSNP[12] ID, OMIM[13] ID, diseases associated to the variant according to Mitomap[6]);

- variability and allele frequency data (nucleotide and aminoacid variability in healthy and diseased individuals, allele frequency in healthy and diseased individuals from cumulative and continent-specific datasets);
- pathogenicity predictions from external resources (MutPred[14], Panther[15], PhD SNP[16], SNPs & GO[17], Polyphen2[18]).

Users can choose to use all of these groups of annotations, or select one or more specific annotation if preferred. HmtNote is able to efficiently annotate SNPs, insertions and deletions, as long as they are available on HmtVar.

HmtNote is best suited to work online, since it fetches annotations on-the-fly from HmtVar; nonetheless, it is also possible to use it even when no internet connection is available, thanks to its “offline mode”: users can download the required annotation database on their machine so that it can subsequently be used offline. The downloaded annotation database contains the very same information provided with online annotation, and details can be found in Table 1.

An extensive documentation about HmtNote and its usage and features can be found at http://hmtnote.readthedocs.io/.

Results and discussion

In order to ensure that all annotations are correct and to provide users with some use cases, HmtNote was tested using a VCF file containing mitochondrial variants found in the 1KGenomes[19] dataset. Annotations were performed using full annotation as well as specific basic, cross-reference, variability and prediction annotations options, both in online and offline mode to ensure that results were identical.

Out of the total 4242 variants reported in the VCF file, 3806 were annotated by HmtNote (more than 89% of the total number of variants). The detailed protocol used to download, process and annotate the VCF file is available in the related project on Open Science Framework[20], together with the original and annotated VCF files.

Annotation	Description	Annotation group
Locus	Locus to which the variant belongs	Basic
AaChange	Aminoacidic change determined	Basic
Pathogenicity	Pathogenicity predicted by HmtVar	Basic
DiseaseScore	Disease score calculated by HmtVar	Basic

HmtVar	HmtVar ID of the variant	Basic
Clinvar	Clinvar ID of the variant	Cross-reference
dbSNP	dbSNP ID of the variant	Cross-reference
OMIM	OMIM ID of the variant	Cross-reference
MitomapAssociatedDiseases	Diseases associated to the variant according to Mitomap	Cross-reference
MitomapSomaticMutations	Diseases associated to the variant according to Mitomap Somatic Mutations	Cross-reference
NtVarH/ NtVarP	Nucleotide variability of the position in healthy/patient individuals	Variability
AaVarH/ AaVarP	Aminoacid variability of the position in healthy/patient individuals	Variability
AlleleFreqH/ AlleleFreqP	Allele frequency of the variant in healthy/patient individuals overall	Variability
AlleleFreqH_AF/ AlleleFreqP_AF	Allele frequency of the variant in healthy/patient individuals from Africa	Variability
AlleleFreqH_AM/ AlleleFreqP_AM	Allele frequency of the variant in healthy/patient individuals from America	Variability
AlleleFreqH_AS/ AlleleFreqP_AS	Allele frequency of the variant in healthy/patient individuals from Asia	Variability
AlleleFreqH_EU/ AlleleFreqP_EU	Allele frequency of the variant in healthy/patient individuals from Europe	Variability
AlleleFreqH_OC/ AlleleFreqP_OC	Allele frequency of the variant in healthy/patient individuals from Oceania	Variability
MutPred_Prediction/ MutPred_Probability	Pathogenicity prediction offered by MutPred and relative confidence	Predictions
Panther_Prediction/ Panther_Probability	Pathogenicity prediction offered by Panther and relative confidence	Predictions
PhDSNP_Prediction/ PhDSNP_Probability	Pathogenicity prediction offered by PhD SNP and relative confidence	Predictions
SNPsGO_Prediction/ SNPsGO_Probability	Pathogenicity prediction offered by SNPs & GO and relative confidence	Predictions
Polyphen2HumDiv_Prediction/ Polyphen2HumDiv_Probability	Pathogenicity prediction offered by Polyphen2 HumDiv and relative confidence	Predictions
Polyphen2HumVar_Prediction/ Polyphen2HumVar_Probability	Pathogenicity prediction offered by Polyphen2 HumVar and relative confidence	Predictions

Table 1. Annotations provided by HmtNote.

Conclusions

HmtNote is a simple and efficient tool to annotate human mitochondrial variants from VCF files. It exploits data hosted on HmtVar to perform its annotations using always the latest information available; in addition, it also allows to perform offline annotation if needed. Annotations can be limited to a specific type of information or span all the available data to provide a full enrichment of the annotated VCF file.

HmtNote employs an easy-to-use interface, which makes it appetible for both experienced bioinformaticians as well as scientists who are less familiar with programming, rendering the functional annotation of human mitochondrial variants a trivial task.

HmtNote is a free and open source package, and can be downloaded and installed from PyPI (https://pypi.org/project/hmtnote) or GitHub (https://github.com/robertopreste/HmtNote).

References

1. Gorman, G. S. *et al.* Mitochondrial diseases. *Nat. Rev. Dis. Primer* **2**, 16080 (2016).
2. Bris, C. *et al.* Bioinformatics Tools and Databases to Assess the Pathogenicity of Mitochondrial DNA Variants in the Field of Next Generation Sequencing. *Front. Genet.* **9**, (2018).
3. Shen, L. *et al.* MSeqDR mvTool: A mitochondrial DNA Web and API resource for comprehensive variant annotation, universal nomenclature collation, and reference genome conversion. *Hum. Mutat.* **39**, 806–810 (2018).
4. Rueda, M. & Torkamani, A. SG-ADVISER mtDNA: a web server for mitochondrial DNA annotation with data from 200 samples of a healthy aging cohort. *BMC Bioinformatics* **18**, 373 (2017).
5. Weissensteiner, H. *et al.* mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res.* **44**, W64–W69 (2016).
6. Lott, M. T. *et al.* mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinforma.* **44**, 1.23.1-26 (2013).
7. Fan, L. & Yao, Y.-G. MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. *Mitochondrion* **11**, 351–356 (2011).
8. 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).
9. 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).
10. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference

sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).

11. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862-868 (2016).

12. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

13. 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).

14. MutPred2: inferring the molecular and phenotypic impact of amino acid variants | bioRxiv. Available at: https://www.biorxiv.org/content/10.1101/134981v1.full. (Accessed: 10th April 2019)

15. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377-386 (2013).

16. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinforma. Oxf. Engl.* **22**, 2729–2734 (2006).

17. Capriotti, E. *et al.* WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics* **14 Suppl 3**, S6 (2013).

18. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).

19. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

20. Foster, E. D. & Deardorff, A. Open Science Framework (OSF). *J. Med. Libr. Assoc. JMLA* **105**, 203–206 (2017).