MSeqDR mvTool: a Mitochondrial DNA Web and API Resource for Comprehensive Variant Annotation, Universal Nomenclature Collation, and Reference Genome Conversion

Lishuang Shen[1], **Marcella Attimonelli**[2], **Renkui Bai**[3], **Marie T. Lott**[4], **Douglas C. Wallace**[4,5], **Marni J. Falk**[5,6], and **Xiaowu Gai**[1,7,*]

[1]Center for Personalized Medicine, Department of Pathology & Laboratory Medicine, Children's Hospital Los Angeles, Los Angeles, California, USA

[2]Department of Biosciences, Biotechnologies and Biopharmaceutics, University of Bari, Bari, Italy

[3]GeneDx, Gaithersburg, MD, USA

[4]Center for Mitochondrial and Epigenomic Medicine, Department of Pathology, The Children's Hospital of Philadelphia, Philadelphia, USA

[5]University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA

[6]Division of Human Genetics, Department of Pediatrics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA

[7]Keck School of Medicine, University of Southern California, California, USA

Abstract

Accurate mitochondrial DNA (mtDNA) variant annotation is essential for the clinical diagnosis of diverse human diseases. Substantial challenges to this process include the inconsistency in mtDNA nomenclatures, the existence of multiple reference genomes, and a lack of reference population frequency data. Clinicians need a simple bioinformatics tool that is user-friendly, and bioinformaticians need a powerful informatics resource for programmatic usage. Here, we report the development and functionality of the MSeqDR mtDNA Variant Tool set (mvTool), a one-stop mtDNA variant annotation and analysis web service. mvTool is built upon the MSeqDR infrastructure (https://mseqdr.org), with contributions of expert curated data from MITOMAP (http://www.mitomap.org) and HmtDB (http://www.hmtdb.uniba.it/hmdb). mvTool supports all mtDNA nomenclatures, converts variants to standard rCRS- and HGVS-based nomenclatures, and annotates novel mtDNA variants. Besides generic annotations from dbNSFP and Variant Effect Predictor (VEP), mvTool provides allele frequencies in more than 47,000 germline mito-genomes, and disease and pathogenicity classifications from MSeqDR, Mitomap, HmtDB and ClinVar (Landrum *et al.*, 2013). mvTools also provides mtDNA somatic variants annotations. 'mvTool API' is implemented for programmatic access using inputs in VCF, HGVS, or classical mtDNA

*Corresponding Author: Xiaowu Gai, PhD, 4650 Sunset Blvd., # 173, Los Angles, LA 90027, Phone: 323-644-8506, xgai@chla.usc.edu.

variant nomenclatures. The results are reported as hyperlinked html tables, JSON, Excel and VCF formats. MSeqDR mvTool is freely accessible at https://mseqdr.org/mvtool.php.

1 Introduction

The human mitochondrial DNA (mtDNA) genome is small (16.6 kilobases, Kb), circular, and maternally inherited through the female germline. The lack of recombination between mtDNA genomes leads to distinct mitochondrial lineages that track evolutionary populations, called 'haplogroups'. Each cell has tens to hundreds of mtDNA genomes and high mtDNA mutation rate, which together can result in non-identical mtDNA genome populations co-existing in the same cell or tissue in a state of 'heteroplasmy'. In addition, the mtDNA genome is prokaryotic in origin and utilizes mitochondrial-specific genetic codons. These unique characteristics of mtDNA are not addressed by most nuclear-centric bioinformatics analytic tools, yet are key factors that influence the analysis and interpretation of mtDNA variant pathogenicity.

Historic negligence and limited coordination within the fields present additional obstacles for studying mtDNA variants and their associations with various human diseases. Firstly, comparing and cross-referencing mtDNA variants between studies is challenging due to multiple standards, and non-standard ways to name mtDNA variants in contrast to nuclear DNA variants. Therefore, each mtDNA variant must be converted, normalized, and/or harmonized before being able to work with contemporary bioinformatics tools and resources designed for nuclear DNA variant nomenclature. A second major barrier to accurate mtDNA variant analysis is the multiple mitochondrial reference genomes that have been used in the literature, various genomic resources and technologies including design and manifest files, commercial analysis software, and custom sequence analysis pipelines. The first mtDNA reference genome was the Cambridge Reference Sequence ("CRS") [Anderson *et al.*, 1981], derived from Sanger sequencing of the mtDNA genome of a Caucasian individual who belonged to the H haplogroup; this sequence represents one end branch of the phylogenetic tree but has been shown to contain sequencing errors. The Revised Cambridge Reference Sequence ("rCRS", NCBI Reference Sequence Acc#: NC_012920.1, 16,569 bp in length) was introduced to correct those errors [Andrews *et al.*, 1999] and added one artificial spacer at position 3,107 to preserve the historical CRS position numbering. This rCRS is the most commonly used reference genome for human mtDNA research. However, earlier releases of the GRCh37 human reference genome and, as a result the UCSC hg19 genome release, used the full mtDNA genome sequence of an African individual as the mtDNA reference genome YRI (NCBI Reference Sequence Acc#: NC_001807.4, Yoruban, L3e haplogroup, 16,571 bp in length), which has different position numbering due to the genome length difference and over 40 differences from rCRS. Although the error was fixed in later releases of GRCh37, as well as the newest GRCh38 human genome build, it has been used by major genotyping companies including Affymetrix and Illumina in various genomics platforms (e.g., Affymetrix Genome-Wide Human SNP Array 6.0; Illumina 550 v.1, 550 v.3, 610 v.1). The popularity of these genotyping platforms has led to discordant nucleotide numbers and alleles in variant calling from those based on the rCRS sequence. The same problem has occurred with many academic pipelines that rely upon UCSC hg19 reference genome.

Adding to the complexity has been the proposition and use of yet another mtDNA reference genome, revised sapient reference sequence (RSRS), which was an inferred mtDNA sequence falling between haplogroups L0 and L1′2′3′4′5′6 evolutionarily intended to represent the root mtDNA genome sequence for better rooting of phylogenetic trees [Behar *et al.*, 2012]. RSRS is 16,569 bp in length, including three spacers (positions 523, 524 and 3107) to preserve the historical CRS position numbering. Overall, the literature is filled with mtDNA variants called based on different reference sequences that have led to widespread confusion and mistakes. These problems need to be resolved to avoid further error propagation. A third major barrier to mtDNA genome analysis is the lack of comprehensive and up-to-date mtDNA pathogenic variant databases. Most mtDNA web resources are either limited in data scope or not updated for years. Database exceptions include MitoMap [Lott *et al.*, 2013; http://www.mitomap.org], HmtDB [Clima *et al.*, 2016; http://www.hmtdb.uniba.it/hmdb/], and MSeqDR [Falk *et al.*, 2015; Shen *et al.*, 2016; https://mseqdr.org]. Finally, a fourth barrier to mtDNA analysis is the paucity of mtDNA variants reported in large sequencing projects, established bioinformatics pipelines, and resources. This results in a significant lack of mtDNA variant population frequency data despite the ease to obtain and analyze mtDNA sequence data in large sequencing projects, including the absence of mtDNA data from most widely used reference database such as ExAC and its expanded successor gnoMAD [Lek *et al.*, 2016], 1KG [1000 Genomes Project Consortium, 2013] and EVS [http://evs.gs.washington.edu/EVS]. Thus, no reliable reference population frequency data exist for annotating mtDNA genome variants except for Mitomap, Hmtdb and MSeqDR.

These limitations present significant challenges to non-bioinformaticians, including clinicians, trying to annotate mtDNA variants of interest and interpret their likely biological and clinical significance. Furthermore, while a free and easy-to-use web–based bioinformatics tool is much desired to facilitate these analyses for clinical purposes, bioinformaticians would benefit from access to convenient, powerful web services for batch processing a large number of mtDNA variants in an automated way from the command line and programmatically. Indeed, whole-exome sequencing (WES) or whole-genome sequencing (WGS) data usually contain sufficient reads from mtDNA genome to allow for mtDNA variant analysis in association with various human diseases almost as by-products but rarely utilized and could be fruitful to mine by bioinformaticians across a wide breadth of clinical and research projects.

To overcome these challenges of mitochondrial DNA genomics, we established the versatile mvTool web service to exploit the expertise and resources of the international MSeqDR consortium, including MitoMap and HmtDB, and timely sharing of expert-curated data by MSeqDR members. The mvTool converter supports all classical and current mtDNA variant nomenclature, while the mvTool annotator module pools population frequency data about mtDNA variants from multiple sources and provide over 100 types of *in silico* annotations and predictions.

2. Methods

2.1. mvTool Chromosome and Variant Converter Module

The main mvTool feature for non-bioinformaticians is its user-friendly support of accepting flexible variant input formats. As discussed above, many mtDNA variant nomenclatures remain in use. mvTool can support all input formats, whether separately or combined (Table 1). Mitochondrial chromosome names currently used include "M", "MT", "ChrM", "ChrMT", and "NC_01290.1". mvTool recognizes and automatically converts each to "M". Variants may be entered in free-text style, saving user's effort to prepare variant input formats and allowing mvTool to clean up and split complex input into individual variant entries. mvTool iteratively works through each entry, converting each variant to formats supported by internal and external annotation tools. mtDNA variant nomenclatures supported by mvTool include: 1) Classical I: 8527, 8993G, 5787_5789d, 1494.1T, 7472.XA. This has been used in literature extensively, with only position provided for transitions and only the alternate allele shown for transversions; 2) Classical II: T8993G. Both reference and variant bases are provided; 3) VCF-style input: this is a tab-delimited format requiring at least the first 5 columns that is widely used in bioinformatics variant analysis tools [Danecek *et al.*, 2011]; 4) HGVS: NC_012920.1:m.8993T>G. This is the recommended format and has become widely used in recent literature; 5) Ensembl: MT:g.8993T>G. This is a commonly accepted input format for annotation tools such as VEP [McLaren *et al.*, 2016]; 6. Mutalyzer: NC_012920.1:g.8993T>G; and 7) Other non-standard format: 8527A>G, which is commonly used but non-standard. Input variants for mvTool converter can be entered in a mix of various formats for convenience, with one format used for each variant entry. mvTool conducts variant naming conversions to obtain standard genomic coordinate based on HGVS nomenclature using a custom text conversion tool, and web service APIs from Mutalyzer and Ensembl VEP. mvTool also converts YRI-based (African Yoruba, GenBank Acc#: AF347015) positions and naming to standard rCRS-based (NC_012920.1) positions and naming per HGVS format.

2.2. mvTool Variant Annotation Module

Once mtDNA variant formats are converted and standardized, they can be annotated in the mvTool annotation module. The mvTool annotation module shares much of the backend data with the MSeqDR VariantOneStop and HBCR annotation tools, which are also useful for annotation of nuclear DNA variants [Falk *et al.*, 2015; Shen *et al.*, 2016]. mvTool annotation module individually iterates through the list of input variants of potentially mixed input formats. For a variant that has not been annotated previously by any MSeqDR user, mvTool conducts new prediction by calling Ensembl VEP RESTful API, and stores its genomic annotations in an internal database that mvTool will search first. For previously annotated variants, mvTool simply extracts and returns these existing annotations, but with updated population and pathogenicity data from the MSeqDR Consortium members [Falk *et al.*, 2015; Shen *et al.*, 2016]. For mtDNA-specific variant function, mvTool integrates population frequencies from multiple major resources including Mitomap [Lott *et al.*, 2013; http://www.mitomap.org], HmtDB - the Human Mitochondrial Database [Clima *et al.*, 2016; Rubino *et al.*, 2012; http://www.hmtdb.uniba.it/hmdb/], GeneDx (personal communication), the 1000 Genomes Project data [1000 Genomes Project Consortium, 2013; Diroma et al.,

2014] and the MSeqDR community exome dataset M1 (1,534 exomes) [Falk *et al.*, 2015; Shen *et al.*, 2016]. These sources provide variant allele frequencies from 37,000, 35,942, 6,391, 1,700, and 1,534 full-length mtDNA genomes, respectively. The total combined population size of these sources exceeds 47,000 unique subjects, making it the largest of any mitochondrial genomic databases. In addition, when the input includes all mtDNA variants of a given sample, accurate mtDNA haplogroup assignment can be obtained by mvTool, which calls Phy-Mer [Navarro-Gomez *et al.*, 2014; https://mseqdr.org/phymer.php] that is also hosted by MSeqDR.

2.3. mvTool Implementation

mvTool is implemented as a web service hosted on MSeqDR, the Mitochondrial Disease Sequence Resource [https://mseqdr.org]. The mvTool back-end database is built using MySQL 5.7 with JSON support. The front end web interface is written in PHP (www.php.net), JQuery (www.jquery.com), HTML, and JavaScript libraries. Custom Python, bash and perl scripting are also used at the back-end.

3. Results

3.1. Running mvTool

mvTool functions in two modes, "web" and "API". In web mode, the user can paste a list of mtDNA variants into the web form and have annotations returned as either an HTML table or as a downloadable Excel file. The input form is pre-populated with example variants in mixed formats to enable quick start, and includes help documentation describing the technical details shown when the tool is first opened. Upon submission, each job is sent to a 24-CPU dedicated server with 32 GB of RAM and 4 TB of HDD. With this set-up, most analyses of one mtDNA genome (usually comprised of less than 100 variants) take under 1 minute to complete.

In the API mode, users can use the UNIX curl command to remotely upload a file in VCF, HGVS, or classical mtDNA variant formats to retrieve annotation back as either JSON or annotated VCF files. There are three ways to access MSeqDR mvTool API, using syntax similar to the following example commands:

1. VCF input, new VCF returned with MSeqDR annotations appended to INFO column:

    ```
    curl -s -X POST https://mseqdr.org/mtannotapi.php?format=cpmvcf --
    data-binary @demo00001.MT.vcf -o demo00001.MT.annot.vcf
    ```

2. VCF input, JSON return, with full annotation details:

    ```
    curl -s -X POST https://mseqdr.org/mtannotapi.php?format=vcf --
    data-binary @demo00001.MT.vcf -o demo00001.MT.vcf.json
    ```

3. HGVS or classical mtDNA variant input formats, JSON return, with full annotations:

```
curl -s -X POST https://mseqdr.org/mtannotapi.php?format=hgvs --
data-binary @mvtool_hgvs.txt -o mvtool_hgvs.txt.json
```

Data transfer is performed securely through an SSL certificate-backed web server. The download links are randomly regenerated and do not reveal any user and input information. The mvTool API is currently being used by Children's Hospital Los Angeles Clinical Exome Sequencing pipeline and the North American Mitochondrial Disease Consortium (NAMDC, personal communications).

3.2. Output

Variant annotations provided by mvTool consist of six categories: 1) MSeqDR community population allele frequency, 2) diseases and phenotypes, 3) general annotations, 4) dbNSFP [Liu *et al.*, 2013] and CADD scores [Kircher *et al.*, 2014], 5) HmtDB pathogenicity predictions, and 6) somatic variant data from the International Cancer Genome Consortium (ICGC) [Zhang *et al.*, 2011] and the Catalogue of Somatic Mutations in Cancer (COSMIC) [Forbes *et al.*, 2016]. In category 1, the population allele frequency of each variant is provided, based on resources and populations listed in Table 2. In category 2, reported associations with diseases and phenotypes of each variant is provided, based on data obtained from MSeqDR locus specific database (LSDB), ClinVar, HmtDB, and literature curated by MitoMap. In category 3, genomic annotations are provided by VEP for each variant, namely the affected gene, transcript and codon all defined in HGVS nomenclature, as well as SIFT and PolyPhen predictions. In category 4, more comprehensive genomic annotations are provided based on dbNSFP, including the CADD scores but limited to the coding region for protein variants. In category 5, the predicted pathogenicity of each possible coding or non-coding variant, whether prior reported or not, is provided by HmtDB [Clima *et al.*, 2016]. In category 6, if this is a somatic mtDNA variant that has been reported previously in either COSMIC or ICGC, the corresponding COSMIC or ICGC mutation ID is returned and hyperlinked to the COSMIC or ICGC webpage. When a job is complete, mtDNA variant annotation results are returned as multiple HTML tables, as well as downloadable Excel files. The HTML table entries are hyperlinked by URL links to MSeqDR, as well as to external databases when applicable. All results can be batch downloaded as a single combined Excel file, or as a JSON format file, by clicking on the corresponding links (randomly named for anonymity) on the page.

For API calls, annotation results are returned as either an augmented new VCF file with MSeqDR annotations appended to the INFO column, or as full detail nested JSON format where data is divided into several sections. JSON data can be programmatically parsed and used, or visualized as vertical and horizontal nested HTML tables using the MSeqDR json2table report tool (https://mseqdr.org/json2table.php), which is publicly accessible.

4. Conclusion

We have developed a powerful web service for standardized mtDNA variant annotations that are comprehensive and incorporate user-friendly input and output format options. The mvTool API generates VCF and JSON output formats to facilitate the ready integration of mvTool into other sequence analysis pipelines. mvTool harnesses actively updated mtDNA annotations from highly-recognized and widely-used mitochondrial genome curation and analysis resources, namely MitoMap, HmtDB and MSeqDR. These resources collectively offer the largest collection of reference mtDNA genomes and variant data sets. mvTool complements the MSeqDR Phy-mer (Navarro-Gomez *et al.*, 2014) and MToolBox (Calabrese *et al.*, 2014) resources that work on sequence data in FASTQ, BAM and classical FASTA formats. Together, these resources provide a full dynamic range of mtDNA genomic data analyses and processing capability in the unified MSeqDR website (https://mseqdr.org/). We welcome suggestions from users on this mvTool and the MSeqDR resource in general (https://mseqdr.org/feedback.php).

References

Anderson S, et al. Sequence and organization of the human mitochondrial genome. Nature. 1981; 290(5806):457–465. [PubMed: 7219534]

Andrews RM, et al. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nature Genetics. 1999; 23(2):147. [PubMed: 10508508]

Behar D, et al. A "Copernican" Reassessment of the Human Mitochondrial DNA Tree from Its Root. The American Journal of Human Genetics. 2012; 90(5):936.

Calabrese C, et al. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics. 2014; 30(21):3115–3117. [PubMed: 25028726]

Clima R, et al. HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor. Nucleic Acids Research. 2016; 45(D1):D698–D706. [PubMed: 27899581]

Danecek P, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–2158. [PubMed: 21653522]

Diroma M, et al. Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. BMC Genomics. 2014; 15(Suppl 3):S2.

Falk M, et al. Molecular Genetics and Metabolism. 2015; 114(3):388–396. [PubMed: 25542617]

Forbes S, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research. 2016; 45(D1):D777–D783. [PubMed: 27899578]

Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014; 46(3):310–315. [PubMed: 24487276]

Landrum M, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research. 2013; 42(D1):D980–D985. [PubMed: 24234437]

Lek M, et al. Analysis of protein-coding genetic variation in 60, 706 humans. Nature. 2016; 536(7616):285–291. [PubMed: 27535533]

Liu X, et al. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Human Mutation. 2013; 34(9):E2393–E2402. [PubMed: 23843252]

Lott M, et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. Current Protocols in Bioinformatics. 2013:1.23.1–1.23.26.

McLaren W, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016; 17(1)

Navarro-Gomez D, et al. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. Bioinformatics. 2014; 31(8):1310–1312. [PubMed: 25505086]

Shen L, et al. MSeqDR: A Centralized Knowledge Repository and Bioinformatics Web Resource to Facilitate Genomic Investigations in Mitochondrial Disease. Human Mutation. 2016; 37(6):540–548. [PubMed: 26919060]

van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human Mutation. 2009; 30(2):E386–E394. [PubMed: 18853457]

Zhang J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database. 2011; 2011:bar026–bar026. [PubMed: 21930502]

Table 1

Mitochondrial DNA variant nomenclature.

mtDNA Variant Format	Information and Example
1. Classical I	Alleles are omitted for transition and only variant alleles are included in transversions. Insertions are denoted as ".1" or ".X" for single or multiple insertions. Widely used in literature, PhyloTree, Haplogrep. Example: 8527, 8993G, 8993d, 5787_5789d, 1494.1T, 7472.XA
2. Classical II	T8993G
3. VCF-style input	Tab-delimited, with at least the first 5 columns in vcf format.
4. HGVS	This is the recommended mtDNA variant format and is widely used in literature. This is also a format used by NCBI and ClinVar. Reference version is required, but may still be omitted in literature. Example: NC_012920.1:m.8993T>G, m.8993T>G, m.8993_8994delTG, m.1494_1495insT
5. Ensembl	Example: MT:g.8993T>G
6. Mutalyzer	Example: NC_012920.1:g.8993T>G
7. Other non-standard formats	Example: 8527A>G, ChrM:g.8993T>G, ChrMT:g.8993T>G

Table 2

mvTool Mitochondrial DNA Variant Major Data Source

Data Source	Data nature	Data content	Reference and URL
MitoMap	Mitochondrial DNA Variants and expert curated disease associated variants	12,400 genomic variants from 37,528 mito-genomes, 6,016 *in silico* prediction of tRNA pathogenicityscore, 605 Disease-associated variants	Lott *et al.*, 2013; http://www.mitomap.org
HmtDB	Mitochondrial DNA Variants and in silico predicted pathogenicity information	Total number of genomes: 35,942 (October 2017): 10,093 mito-genomic variants from 32,240/30,860 normal (including 1,427 from 1K Genome project) and 3,702/3,558 patient all/complete mito-genomes. *In silico* prediction of any potential mtDNA non-synonymous and tRNA variants.	Clima *et al.*, 2016; http://www.hmtdb.uniba.it/hmdb/
MSeqDR Consortium	Mitochondrial DNA Variants from non-patients and patient exomes	2,164 genomic variants from 1,534 mito-genomes	Falk *et al.*, 2015; Shen *et al.*, 2016; https://mseqdr.org
GeneDx	Mitochondrial DNA Variants from non-patients	5,946 genomic variants from 6,391 mito-genomes	http://genedx.com
Enembl and dbSNP	Mitochondrial DNA Variants	About 1,300 genomic variants without population allele information	www.ensembl.org
PhyloTree	Haplogroup-defining mitochondrial DNA variants	4,560 genomic variants without population allele information (version 16)	van Oven M, Kayser M. 2009, http://www.phylotree.org/