mtDNA Variation and Analysis Using Mitomap and Mitomaster

Marie T. Lott,[1] Jeremy N. Leipzig,[2] Olga Derbeneva,[1] H. Michael Xie,[2] Dimitra Chalkia,[1] Mahdi Sarmady,[2] Vincent Procaccio,[3] and Douglas C. Wallace[1,4]

[1]Center for Mitochondrial and Epigenomic Medicine, Children's Hospital of Philadelphia Research Institute, Philadelphia, Pennsylvania
[2]Center for Biomedical Informatics, Children's Hospital of Philadelphia Research Institute, Philadelphia, Pennsylvania
[3]Biochemistry and Genetics Department, Angers Hospital; UMR CNRS 6214/INSERM U1083, Angers, France
[4]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

ABSTRACT

The Mitomap database of human mitochondrial DNA (mtDNA) information has been an important compilation of mtDNA variation for researchers, clinicians, and genetic counselors for the past 25 years. The Mitomap protocol shows how users may look up human mitochondrial gene loci, search for public mitochondrial sequences, and browse or search for reported general population nucleotide variants as well as those reported in clinical disease. Within Mitomap is the powerful sequence analysis tool for human mitochondrial DNA, Mitomaster. The Mitomaster protocol gives step-by-step instructions showing how to submit sequences to identify nucleotide variants relative to the rCRS, determine the haplogroup, and view species conservation. User-supplied sequences, GenBank identifiers, and single-nucleotide variants may be analyzed. *Curr. Protoc. Bioinform.* 44:1.23.1-1.23.26. © 2013 by John Wiley & Sons, Inc.

Keywords: biological database • information retrieval • human mitochondrial DNA • haplogroups • species conservation • GenBank sequences • single nucleotide variants

INTRODUCTION

The Mitomap database of human mitochondrial DNA (mtDNA) information has been an important resource for information about the human mitochondrial DNA (mtDNA) for researchers, clinicians, and genetic counselors for the past 25 years. Essential information about the mitochondrial reference sequence is provided, along with an extensive compilation of mtDNA variants. The Mitomap curators search research literature for published reports of mitochondrial DNA variants and index those variants in the database. Those variants that are reported as having possible association with disease are noted. A new addition to Mitomap is the inclusion of data from full-length human mtDNA sequences in GenBank.

The Mitomap protocol section (Basic Protocol 1) shows how users may look up human mitochondrial gene loci, search for public mitochondrial sequences, and browse or search for reported general population nucleotide variants as well as those reported in clinical disease. Within Mitomap is the powerful sequence-analysis tool for human mitochondrial DNA, Mitomaster.

The Mitomaster protocol section (Basic Protocol 2) gives step-by-step instructions showing how to submit sequences to identify nucleotide variants relative to the rCRS, determine the haplogroup, and view species conservation. User-supplied sequences, GenBank sequences, and single-nucleotide variants may be analyzed.

Current Protocols in Bioinformatics 1.23.1-1.23.26, December 2013
Published online December 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi0123s44

BASIC PROTOCOL 1

EXPLORING mtDNA VARIANTS WITH MITOMAP

The Mitomap database can be accessed at *http://www.mitomap.org* (Fig. 1.23.1).

Mitomap consists of three main sections: (1) background information about the human mitochondrial DNA; (2) an annotated listing of mtDNA variants, both general population and patient; and (3) The Mitomaster analysis tool (Basic Protocol 2).

Necessary Resources

Hardware

Internet connection

Software

An up-to-date Web browser such as Mozilla Firefox, Google Chrome, Apple Safari, or Microsoft Internet Explorer (version 9 or higher)

Mitomap's background information about human mitochondrial DNA

1. Access the Mitomap database at *http://www.mitomap.org*. Several pages of important background information are available, shown in Figure 1.23.2.
2. Click on Link1, The Annotated Human Mitochondrial DNA Sequence (Fig. 1.23.3), to find the information about the revised Cambridge reference sequence (Andrews et al., 1999) and the sequence itself.

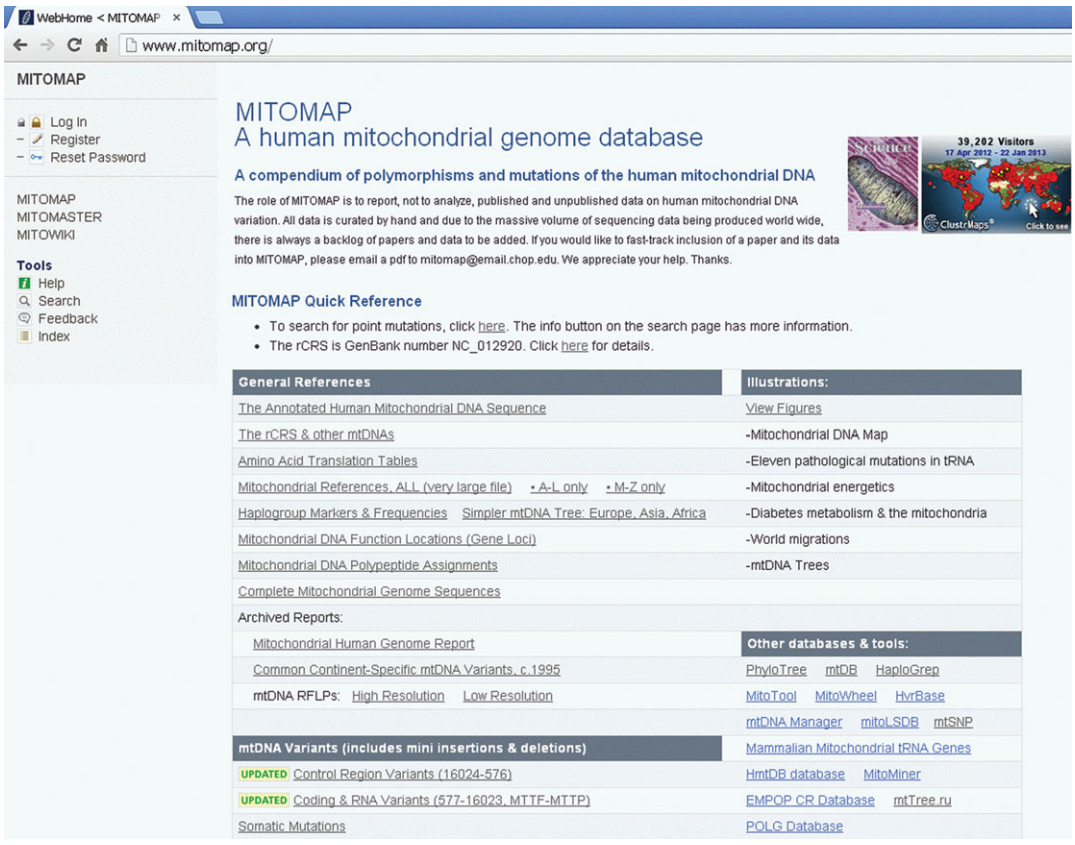


Figure 1.23.1 The home page of *http://www.mitomap.org*.

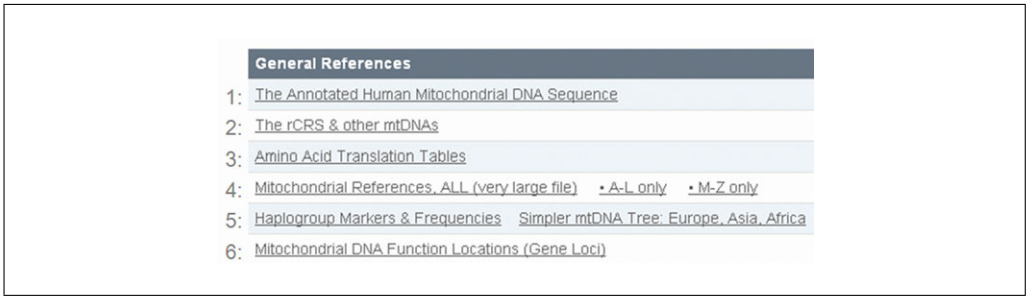


Figure 1.23.2 Essential mtDNA Background Information.

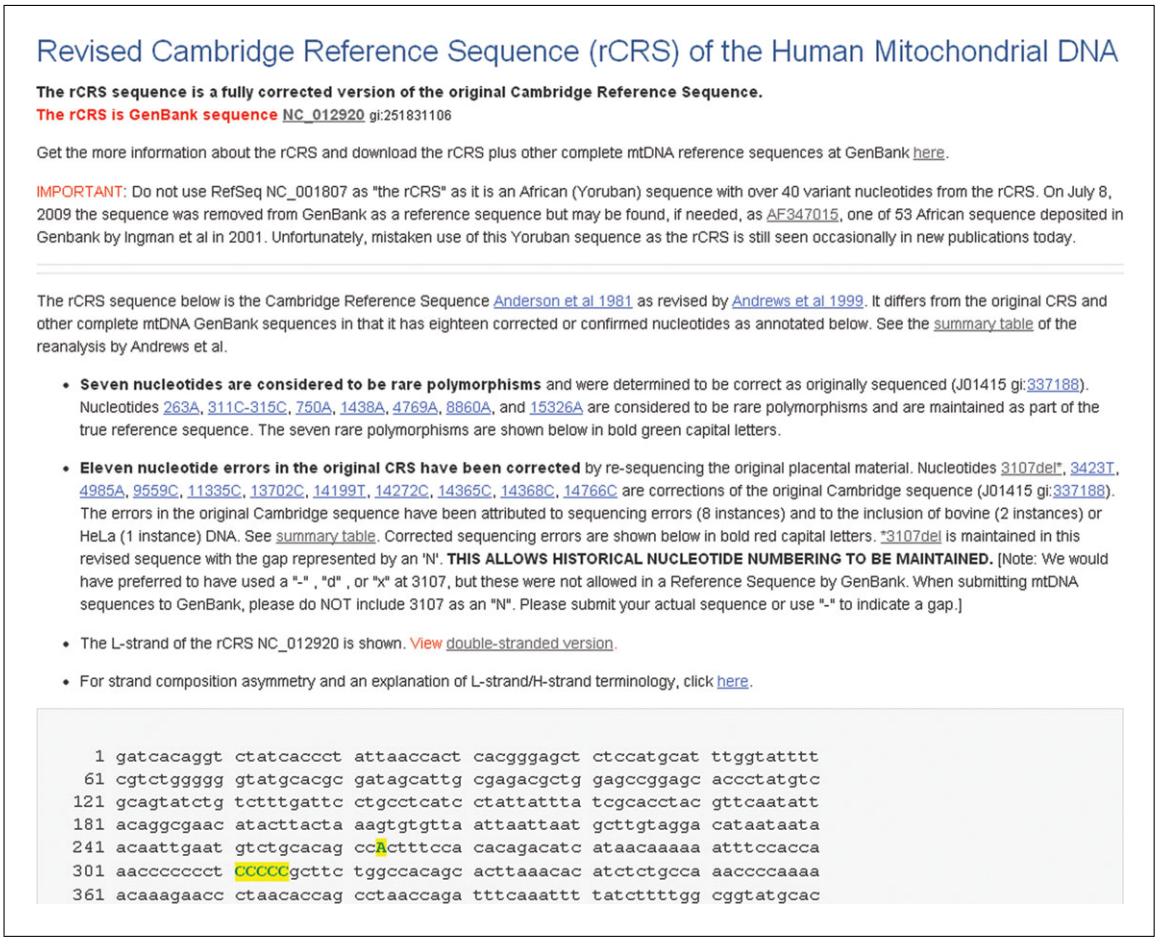
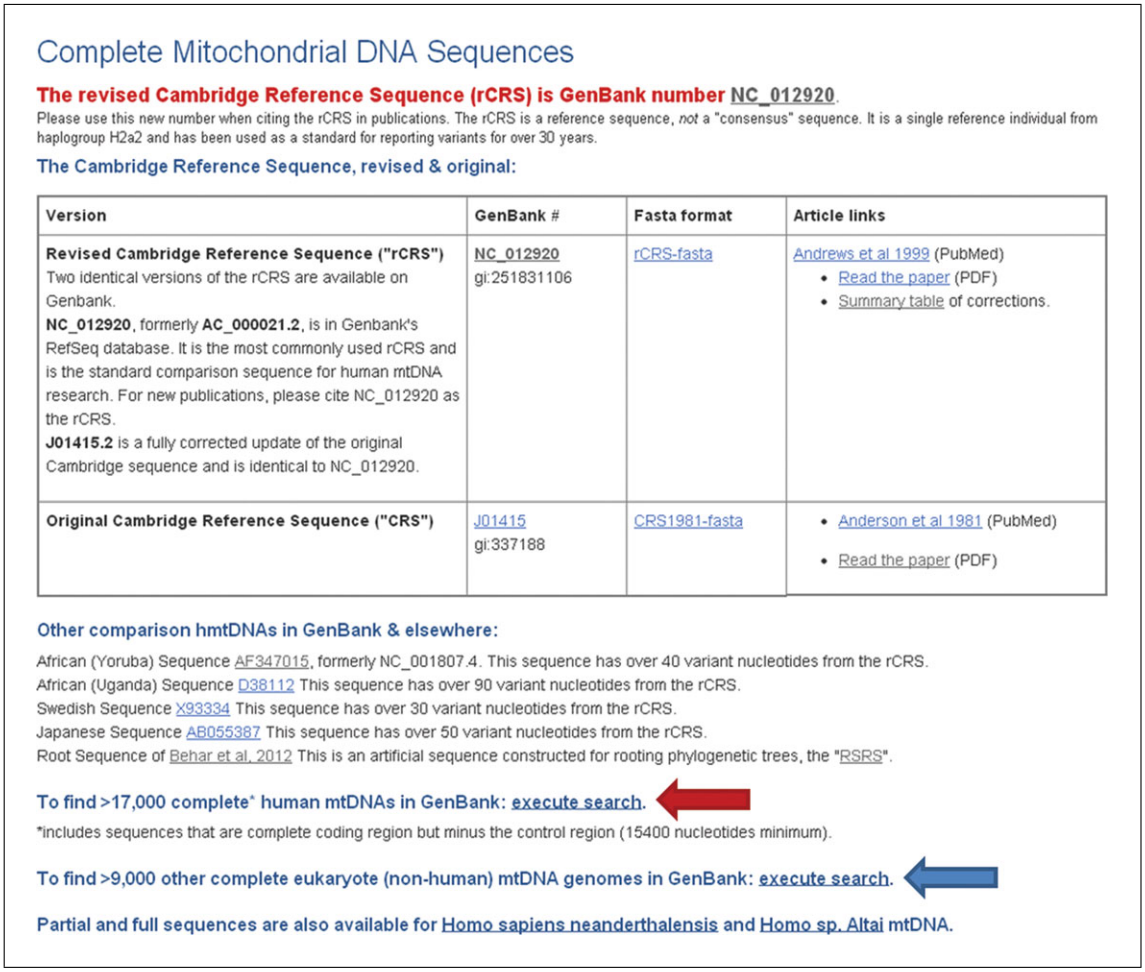


Figure 1.23.3 The rCRS.

This is critical information because, for many years, researchers mistakenly used either the original but outdated version of the reference sequence (Anderson et al., 1981) or an African Yoruban sequence that was listed as the primary reference sequence in GenBank for several years (AF347015; formerly RefSeq NC_001807.4). Use of the Yoruban sequence on DNA sequencing chips has often resulted in confusion and, occasionally, misinterpretation of data.

3. Click on Link 2, The rCRS & Other mtDNAs (Fig. 1.23.4), to go to a companion page where you will find the version history of the rCRS and links to other representative mtDNA sequences from different continental populations, as well as search tools to retrieve full length mtDNA sequences from GenBank.

4. Click on the "search for complete human mtDNAs in GenBank" link (Fig. 1.23.4, red arrow) to retrieve sequences with a minimum length of 15400 bp and a maximum of 16600 bp.



Complete Mitochondrial DNA Sequences

The revised Cambridge Reference Sequence (rCRS) is GenBank number NC_012920.

Please use this new number when citing the rCRS in publications. The rCRS is a reference sequence, *not* a "consensus" sequence. It is a single reference individual from haplogroup H2a2 and has been used as a standard for reporting variants for over 30 years.

The Cambridge Reference Sequence, revised & original:

Version	GenBank #	Fasta format	Article links
Revised Cambridge Reference Sequence ("rCRS") Two identical versions of the rCRS are available on Genbank. **NC_012920**, formerly **AC_000021.2**, is in Genbank's RefSeq database. It is the most commonly used rCRS and is the standard comparison sequence for human mtDNA research. For new publications, please cite NC_012920 as the rCRS. **J01415.2** is a fully corrected update of the original Cambridge sequence and is identical to NC_012920.	NC_012920 gi:251831106	rCRS-fasta	Andrews et al 1999 (PubMed) • Read the paper (PDF) • Summary table of corrections.
Original Cambridge Reference Sequence ("CRS")	J01415 gi:337188	CRS1981-fasta	• Anderson et al 1981 (PubMed) • Read the paper (PDF)

Other comparison hmtDNAs in GenBank & elsewhere:

African (Yoruba) Sequence AF347015, formerly NC_001807.4. This sequence has over 40 variant nucleotides from the rCRS.
African (Uganda) Sequence D38112 This sequence has over 90 variant nucleotides from the rCRS.
Swedish Sequence X93334 This sequence has over 30 variant nucleotides from the rCRS.
Japanese Sequence AB055387 This sequence has over 50 variant nucleotides from the rCRS.
Root Sequence of Behar et al. 2012 This is an artificial sequence constructed for rooting phylogenetic trees, the "RSRS".

To find >17,000 complete* human mtDNAs in GenBank: execute search.

*includes sequences that are complete coding region but minus the control region (15400 nucleotides minimum).

To find >9,000 other complete eukaryote (non-human) mtDNA genomes in GenBank: execute search.

Partial and full sequences are also available for Homo sapiens neanderthalensis and Homo sp. Altai mtDNA.

Figure 1.23.4 Complete Mitochondrial DNA Sequences.

5. Click on the link "complete eukaryote (non-human) genomes in GenBank" (Fig. 1.23.4, blue arrow) to retrieve mtDNA sequences of other species.

   In February 2013, these searches returned 17851 full-length human mitochondrial DNA sequences and 9608 complete non-human mitochondrial DNA sequences.

6. Click on Link 3, Amino Acid Translation Table (Fig. 1.23.5), to take you to the human mitochondrial genetic code, with important notes as to the differences between it and the nuclear genetic code.

7. Click on Link 4, Mitochondrial References, to browse a library of the ~5000 publications indexed by the Mitomap curators, with links to PubMed.

8. Select Link 5, Haplogroup Markers and Frequencies, to reach background haplotyping information (Figs. 1.23.6 and 1.23.7).

   Helpful maps of world migrations and haplogroup relationships are located nearby (Figs. 1.23.8 and 1.23.9).

9. Click on Mitochondrial DNA Function Locations (Link 6) to pull up delineated positions of gene loci (Fig. 1.23.10).

   Mitomap's classic map of loci with selected pathological DNA variants is available in the adjacent section of illustrations (Fig. 1.23.11).

The Human Mitochondrial Genetic Code

Phe F	UUU UUC	**Thr** T	ACU ACC ACA ACG	**Asp** D	GAU GAC
Leu (1) L (UUA/G)	UUA UUG	**Ala** A	GCU GCC GCA GCG	**Glu** E	GAA GAG
Leu (2) L (CUN)	CUU CUC CUA CUG	**Tyr** Y	UAU UAC	**Cys** C	UGU UGC
Ile I	AUU* AUC	**Ter**	UAA UAG	**Trp** W	UGA UGG
Met M	AUA AUG	**His** H	CAU CAC	**Arg** R	CGU CGC CGA CGG
Val V	GUU GUC GUA GUG	**Gln** Q	CAA CAG	**Ser (2)** S (AGU/C)	AGU AGC
Ser (1) S (UCN)	UCU UCC UCA UCG	**Asn** N	AAU AAC	See note below*	AGA AGG
Pro P	CCU CCC CCA CCG	**Lys** K	AAA AAG	**Gly** G	GGU GGC GGA GGG

Figure 1.23.5 The human mitochondrial genetic code.

Estimated Worldwide Haplotype Frequencies (%)

Compiled for Mitomap by O. Derbeneva 2009
Please note: These numbers are for illustrative purposes, not for solid calculations.
They are simple means from published frequencies and not always have all haplogroups been typed.
Frequencies ≥ 20% are bolded; frequencies ≥10% are highlighted.

REGION	A	B	C	D	E	F	G	H	I	J	K	L	M	N	R	S	T	U	V	W	X	Y	Z	n.d.
Africa								17	0	5	2	**46**	3	1	0		4	15	3	0	1			3
Middle East	1	<0.5	1	1		0	0	**22**	2	13	5	6	2	4	1	0	9	15	1	2	3	0	0	13
South East Asia		13																						87
Australia and Oceania		**23**										1	7	**48**	7									14
West Europe				1		0		**41**	2	9	5	1	1	1	0		8	18	7	2	2		0	3
East Europe	1	<0.5	2	2		1	1	**35**	2	8	4		2	1	1		11	**22**	3	1	1	0	1	-
Caucasus	1	<0.5	4	4		1		**23**	2	7	6	0	1	3	3		10	**22**	1	2	4		0	7
Central Asia	7	5	12	15		5	5	15	1	3	1	0	6	2	1		6	10	0	2	0	1	2	3
East Asia	7	16	5	**26**	0	11	4	1		1			15	3	2		0	<0.5		0		1	2	7
Siberia	8	3	**30**	19	0	4	5	4	0	2	0		3	3	1	1	2	6	0	0	0	4	1	5
North America	**43**	**23**	18	10	1																3			2
South and Central Americas	**28**	**27**	**21**	19	1							2								0				2

Figure 1.23.6 Haplogroup frequency estimates.

Haplogroup Markers

www.mitomap.org

Nucleotide positions for RFLPs represent the first base of the recognition sequence in the top strand, not the actual cut site.
"-" = enzyme site absent, "+" = enzyme site present
All diagnostic markers are indexed to the rCRS, NC_012920.

Haplogroup	Diagnostic RFLP site*	Diagnostic SNP The rCRS nucleotide is on the left of the SNP position, the diagnostic value is on the right.	HVS1 motif All sites are transtions except where indicated.
L0, L1, L2	+3592h	C3594T	no specific sites
L3	-3592h	C3594C	no specific sites
M	+10397a (+10394c)	C10400T + A10398G	no specific sites
N	-10397a and +10871z	C10400C + A10398A and T10873T	no specific sites
M subgroups:			
C	+13262a (-13259o)	A13263G	16223 16327 16298
D	-5176a	C5178A	16223 16362
E	-13619x	C13626T	16362 16390
G	+4830n (+4831f)	A4833G	16223 16278 16362
Q	no RFLP	A5843G	16129 16241 16311
Z	no RFLP	T9090C	16185 16223 16260 16298
N/R subgroups:			
R	+12704j	C12705C	16223C (=CRS)
B	8281-8289 nps 9bp del	8280:8290 =A[delCCCCCTCTA]G	16183C 16189 16217

Figure 1.23.7 Diagnostic RFLP and SNPs for major haplogroups.

Human mitochondrial DNA variants

10. An extensive database of published mitochondrial DNA variants is available. These can be searched (Fig. 1.23.12A, *http://www.mitomap.org/bin/view/Main/SearchAllele*) or browsed (Fig. 1.23.13). To search for variants:

    a. Locate the bulleted "to search for point mutations, click here" link in the Mitomap Quick Reference section near the top of the Mitomap home page (Fig. 1.23.12A). Click "here" to open the Allele Search box (Fig. 1.23.12B).
    b. Enter one of the following:
        i. A single nucleotide position in the "Start" box.
        ii. A range of up to 100 nucleotides by also entering a position number in the "End" box.

11. Click the "Search" button. Results will be returned as a listing of reported variants and their references as seen in Figure 1.23.12C.

Population variants

12. There are two links for Control Region Variants and Coding/RNA Variants. Information is listed for nucleotide position, rCRS nucleotide, variant nucleotide, and, for coding-region variants, amino-acid change and codon position.

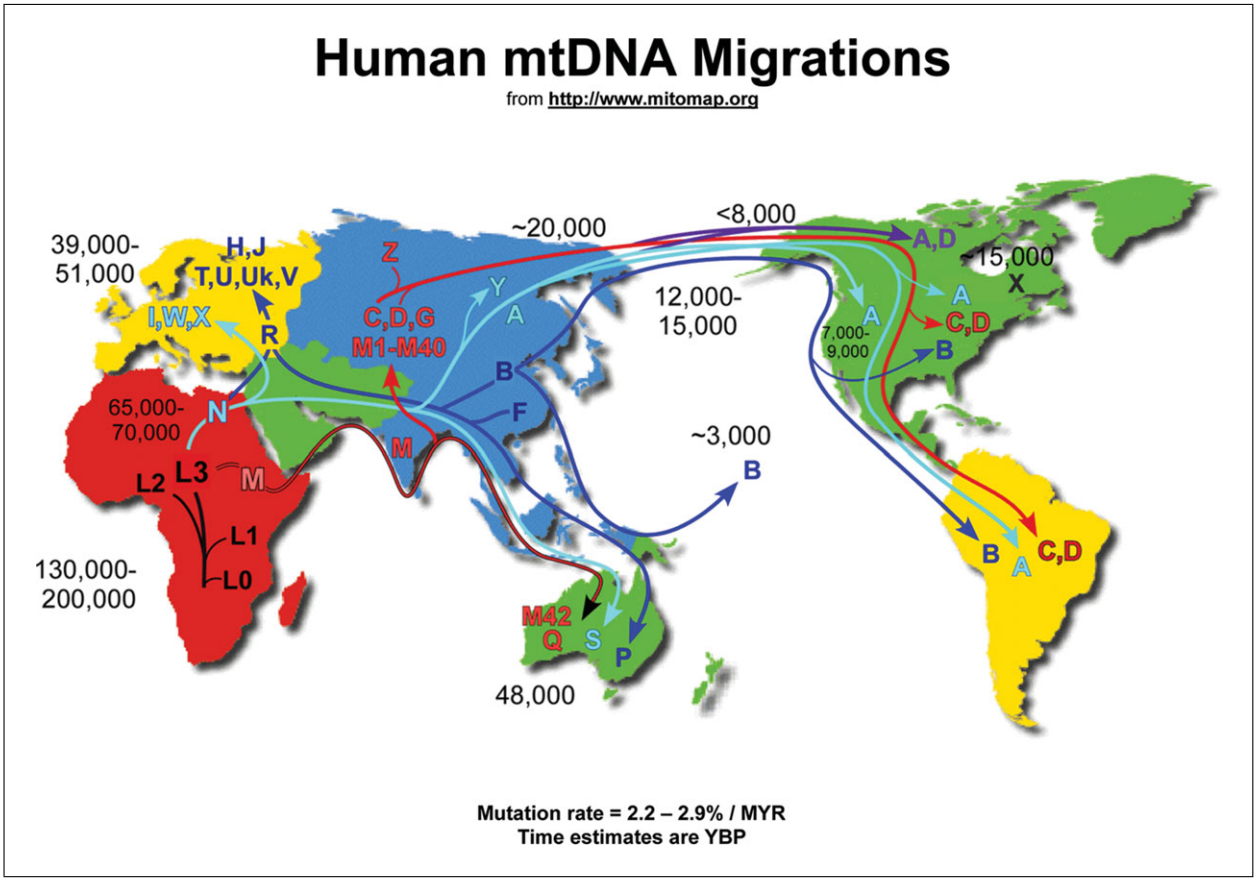


Figure 1.23.8 mtDNA haplogroup migration map.

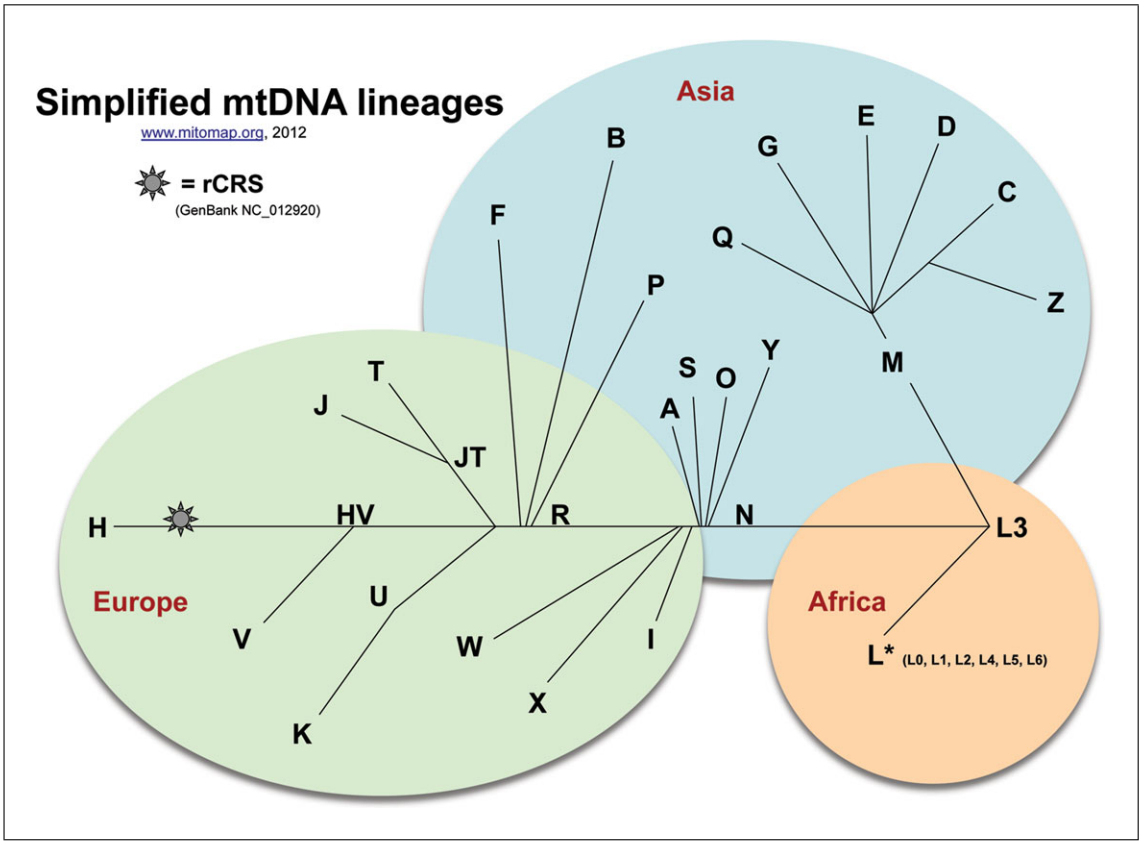
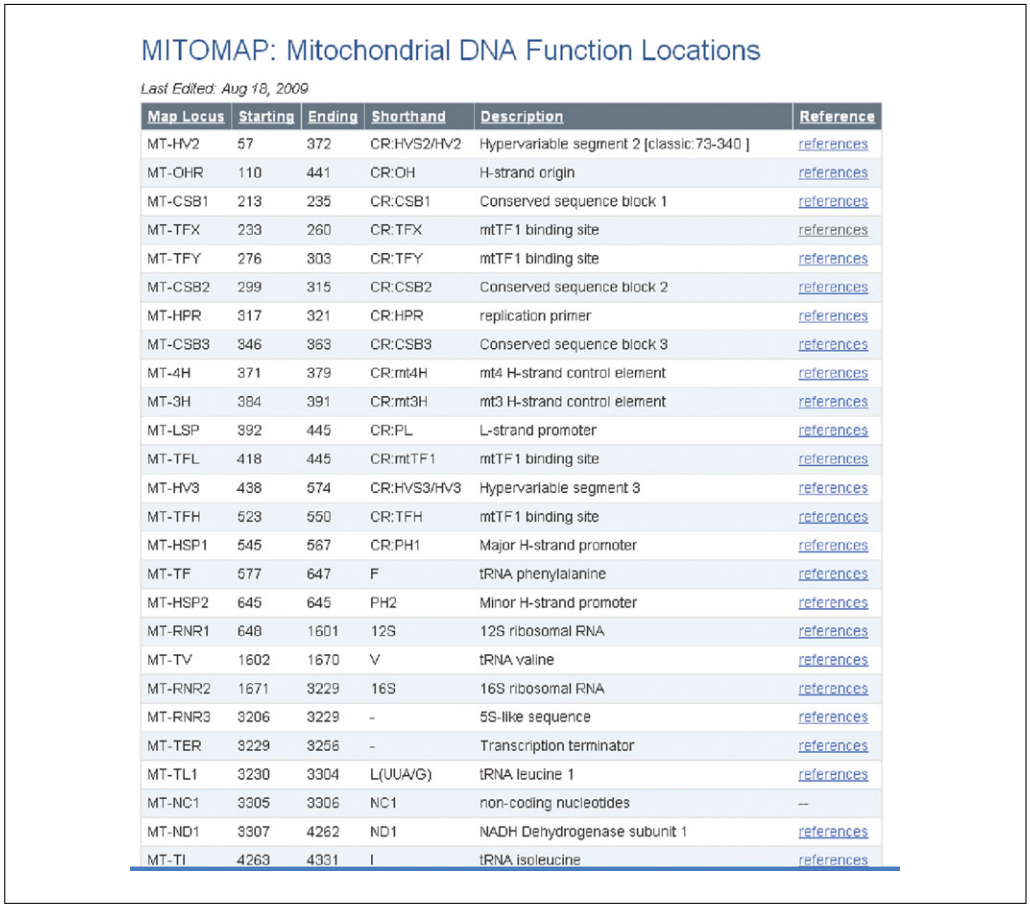


Figure 1.23.9 Simplified mitochondrial haplogroup relationships.



MITOMAP: Mitochondrial DNA Function Locations

Last Edited: Aug 18, 2009

Map Locus	Starting	Ending	Shorthand	Description	Reference
MT-HV2	57	372	CR:HVS2/HV2	Hypervariable segment 2 [classic:73-340]	references
MT-OHR	110	441	CR:OH	H-strand origin	references
MT-CSB1	213	235	CR:CSB1	Conserved sequence block 1	references
MT-TFX	233	260	CR:TFX	mtTF1 binding site	references
MT-TFY	276	303	CR:TFY	mtTF1 binding site	references
MT-CSB2	299	315	CR:CSB2	Conserved sequence block 2	references
MT-HPR	317	321	CR:HPR	replication primer	references
MT-CSB3	346	363	CR:CSB3	Conserved sequence block 3	references
MT-4H	371	379	CR:mt4H	mt4 H-strand control element	references
MT-3H	384	391	CR:mt3H	mt3 H-strand control element	references
MT-LSP	392	445	CR:PL	L-strand promoter	references
MT-TFL	418	445	CR:mtTF1	mtTF1 binding site	references
MT-HV3	438	574	CR:HVS3/HV3	Hypervariable segment 3	references
MT-TFH	523	550	CR:TFH	mtTF1 binding site	references
MT-HSP1	545	567	CR:PH1	Major H-strand promoter	references
MT-TF	577	647	F	tRNA phenylalanine	references
MT-HSP2	645	645	PH2	Minor H-strand promoter	references
MT-RNR1	648	1601	12S	12S ribosomal RNA	references
MT-TV	1602	1670	V	tRNA valine	references
MT-RNR2	1671	3229	16S	16S ribosomal RNA	references
MT-RNR3	3206	3229	-	5S-like sequence	references
MT-TER	3229	3256	-	Transcription terminator	references
MT-TL1	3230	3304	L(UUA/G)	tRNA leucine 1	references
MT-NC1	3305	3306	NC1	non-coding nucleotides	--
MT-ND1	3307	4262	ND1	NADH Dehydrogenase subunit 1	references
MT-TI	4263	4331	I	tRNA isoleucine	references

Figure 1.23.10 Mitochondrial functional locations.

a. Click on the link Control Region Variants (16024-576) to find variants located between tRNA Proline and tRNA Phenylalanine.
b. Click on the link Coding & RNA Variants (577-16023, MTTF-MTTP) to find variants located in the region including the beginning of tRNA Phenylalanine through the end of tRNA Proline.

New for 2013 is the frequency of each variant in a large set of over 18,000 human mitochondrial DNA sequences from GenBank. These sequences have a minimum length of 15.4 kb and are extracted from GenBank on a quarterly basis (Figs. 1.23.14 and 1.23.15).

13. Click on the GB set frequency for a given variant to retrieve a listing of sequences that contain the variant of interest (Fig. 1.23.16) and the relevant PubMed reference. In addition, Mitomap's companion analysis tool Mitomaster (detailed in Basic Protocol 2) predicts haplogroup for each sequence, calculates the number of different haplogroups seen carrying this variant, and displays the total number of each haplogroup found in the GenBank set of sequence. Clicking on each ID and haplogroup yields more information (Figs. 1.23.16, 1.23.17, 1.23.18, 1.23.19, and 1.23.20). Haplotyping is based on Phylotree (van Oven and Kayser, 2009) and is generated by Mitomaster using the Haplogrep engine (Kloss-Brandstatter et al., 2011).

Patient variants

14. To locate information about reported mtDNA variation in patients, click on one of the links listed in the section mtDNA Mutations with Reports of Disease-Associations, shown in Figure 1.23.13.

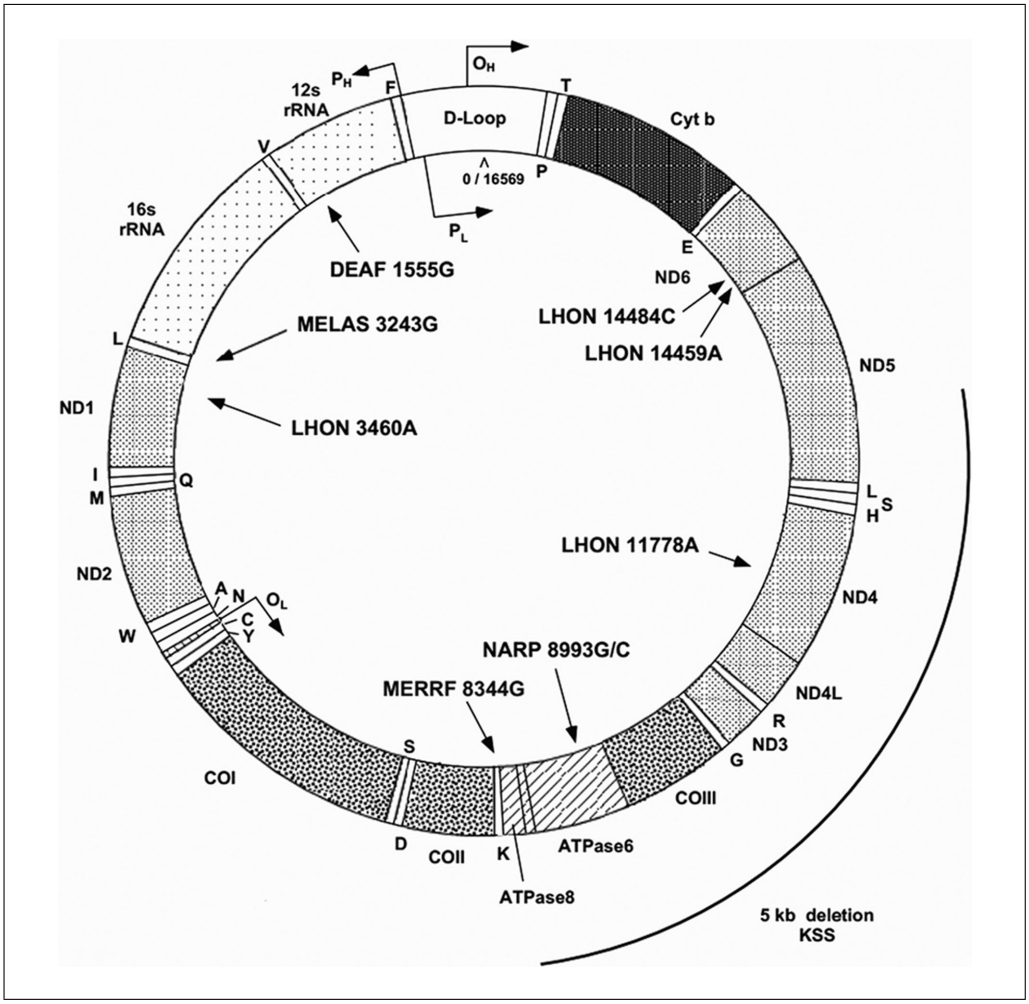


Figure 1.23.11 Gene map of the human mitochondrial DNA with representative disease variants shown.

Variants reported as having possible association with disease are grouped into two categories: those found in the Coding or Control Region (Fig. 1.23.21) and those found in the mitochondrial rRNA or tRNA loci (Fig. 1.23.22). In addition to the same information presented in the population variant tables, reports of heteroplasmy are indicated, as well as a general category of pathogenicity. A pathogenic status of Reported indicates that a publication has considered the mutation as possibly pathologic. A status of "Cfrm" (confirmed) indicates that several independent laboratories have published strong evidence of the pathogenicity of a specific mutation. These mutations are generally accepted by the mitochondrial research community as being pathogenic.

ANALYZING mtDNA VARIANTS WITH MITOMASTER

BASIC PROTOCOL 2

Mitochondrial sequence analysis typically begins with specialized mitochondrial SNP genotyping microarrays using Sanger-based capillary sequencing or various forms high-throughput sequencing. Mitomaster is designed to accommodate output from each of three strategies. Small batches of sequence can be "copy-and-pasted" into a text field or uploaded from FASTA-formatted files. Query by GenBank ID is available for individual or multiple record numbers. Single nucleotide variants, identified by microarray, can be submitted using the SNV Query tab.

Necessary Resources

Hardware

Internet connection

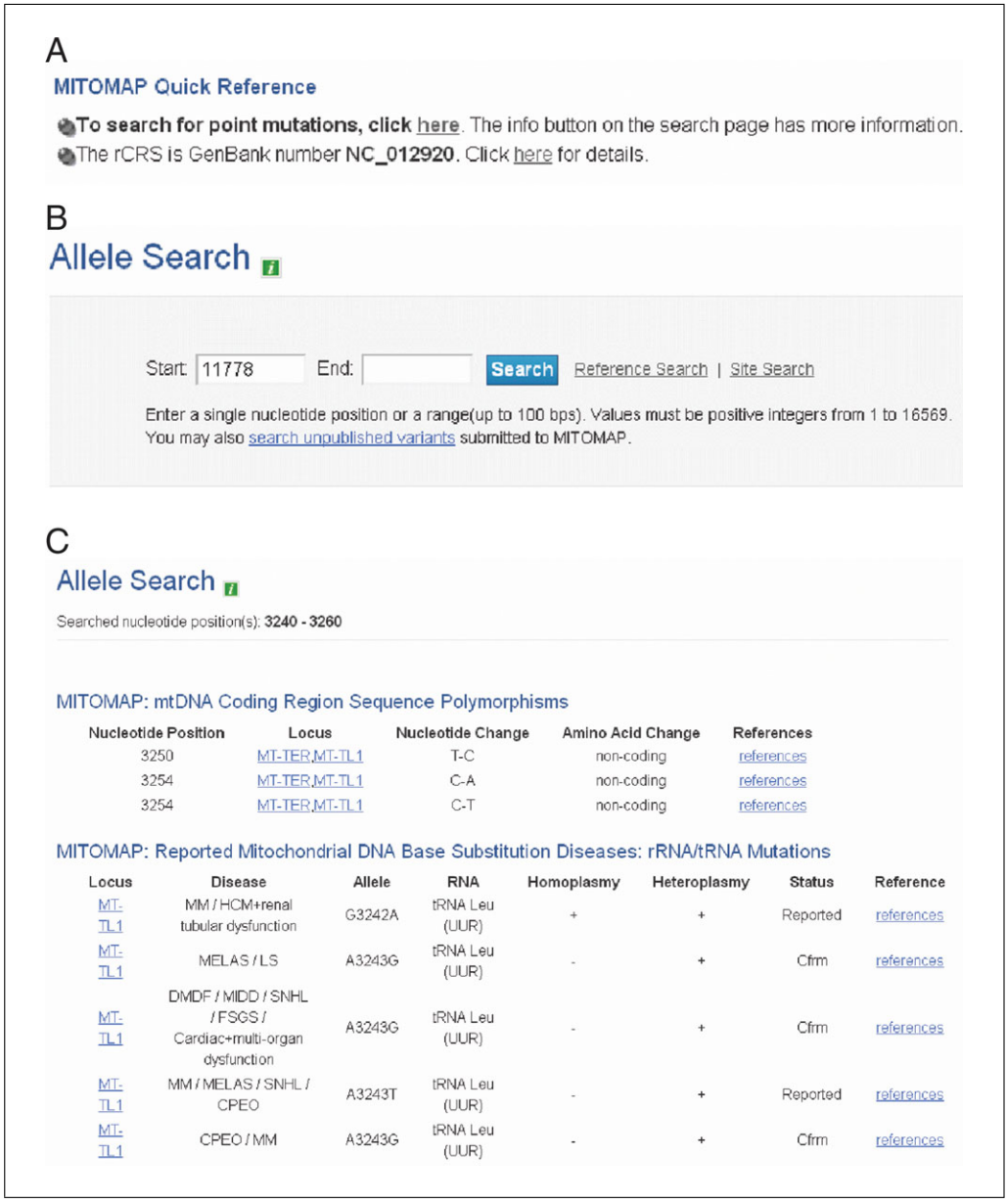


Figure 1.23.12 (**A**) Search box for specific variant(s). (**B**) Search box for specific variant(s). In this example, a single position is queried. (**C**) Results for variant search. In this example, a range was queried. This screenshot shows only a portion of the results returned.

Software

An up-to-date Web browser; Mitomaster is optimized for Mozilla Firefox, Google Chrome, or Apple Safari (if Microsoft Internet Explorer is used, it must be version 9 or higher)

Files

Your data of interest, either as a FASTA file (such as sequence.fasta, downloaded in this protocol; see *APPENDIX 1B* for FASTA format) or as a text file from which you can copy and paste

Sequence query walkthrough

1. Navigate to *http://www.ncbi.nlm.nih.gov/nuccore/EU915478*.

   This sequence, a full length mitochondrial sequence, can demonstrate some of the features (Fig. 1.23.23):

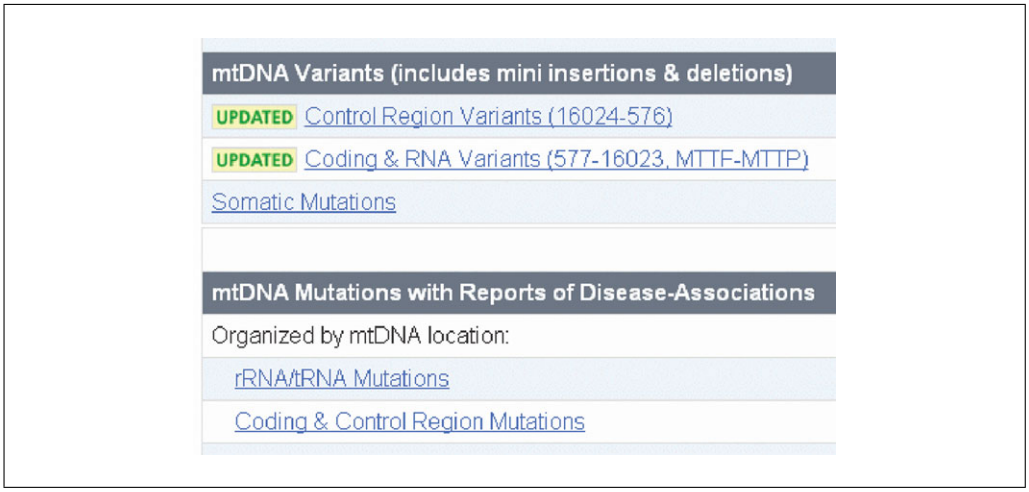


Figure 1.23.13 Links to Mitomap Variants. Variants are organized into two categories—general variants (top section) and those with reports of possible disease-associations (lower section).

2. Click on the Send pop-up dialog link in the upper right of the main section of the sequence record. Select File as a destination and FASTA as a format (Fig. 1.23.24).

   The file will be saved to your local drive as `sequence.fasta` *in your default download folder. Make a note of this folder location. For details of the FASTA format, see APPENDIX 1B and http://www.bioinformatics.nl/tools/crab_fasta.html.*

3. Navigate to *http://mitomaster.mitomap.org/*. This is also accessible as a frame within the Mitomap Web site at *http://mitomap.org/MITOMASTER*. You may also use the Mitomaster link on the left menu of the Mitomap home page.

   Please note that currently the only browsers supported are Chrome, Mozilla, Firefox, and Internet Explorer 9 or higher.

   By default the Sequence tab in the top menu is selected (Fig. 1.23.25).

4. Skip Step 1 in Figure 1.23.25 for now. We will discuss this feature later in the walk-through.

   By default, 45 species are selected for comparison.

5. Click Select File in Step 2, Option 1 in Figure 1.23.25 and navigate to the `sequence.fasta` file that you downloaded earlier (see step 2 in this protocol).

6. Click Submit. The Alignment Summary page (Fig. 1.23.26) is then shown after the sequence is processed:

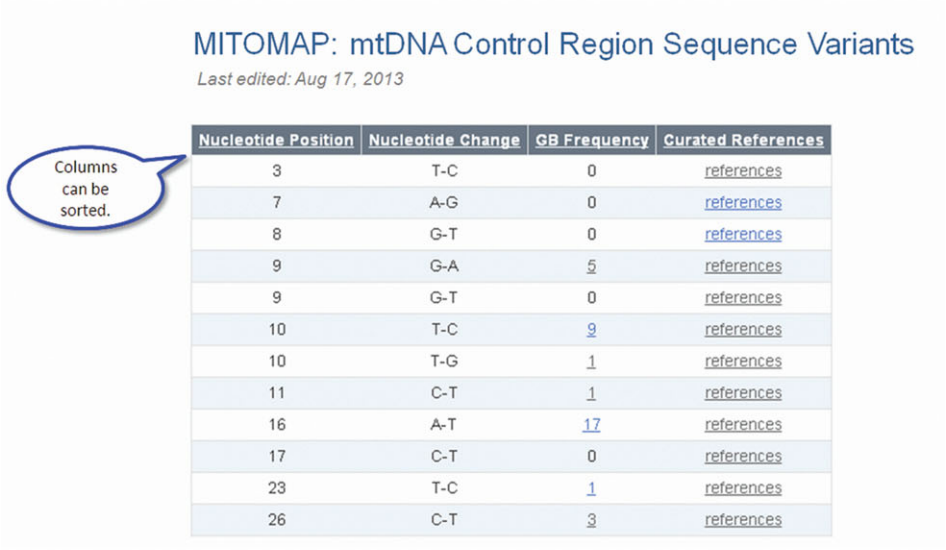
   There are two portions to the Alignment Summary screen:

   rCRS track view: The rCRS track view (Fig. 1.23.27) shows the coverage of the query sequence with respect to the rCRS reference sequence. The example sequence EU915478, indicated by a bar at the top of the screen, displays complete coverage. This is expected, as it is a full-length sequence. Tracks below the query sequence correspond to the locations of protein-coding genes, ribosomal RNAs, and transfer RNAs.

   Sequence alignment: For each query sequence analyzed, you will find (1) the predicted haplogroup as calculated by HaploGrep, (2) the total number of variants relative to the rCRS, and (3) a condensed list of the variants observed. In the example given (Fig. 1.23.28, circled in blue), the predicted haplogroup is J1b and the total number of variants detected is 41, with a summary listed in the right-most column.

7. Next, click on the sequence name to bring up the Alignment Details page (Fig. 1.23.29). In the example above, you will click on "gi|194441041|gb|EU915478.1" (Fig. 1.23.28, blue arrow). The sequence name is taken from the first line of the

A



Nucleotide Position	Nucleotide Change	GB Frequency	Curated References
3	T-C	0	references
7	A-G	0	references
8	G-T	0	references
9	G-A	5	references
9	G-T	0	references
10	T-C	9	references
10	T-G	1	references
11	C-T	1	references
16	A-T	17	references
17	C-T	0	references
23	T-C	1	references
26	C-T	3	references

B

MITOMAP: mtDNA Coding Region & RNA Sequence Variants

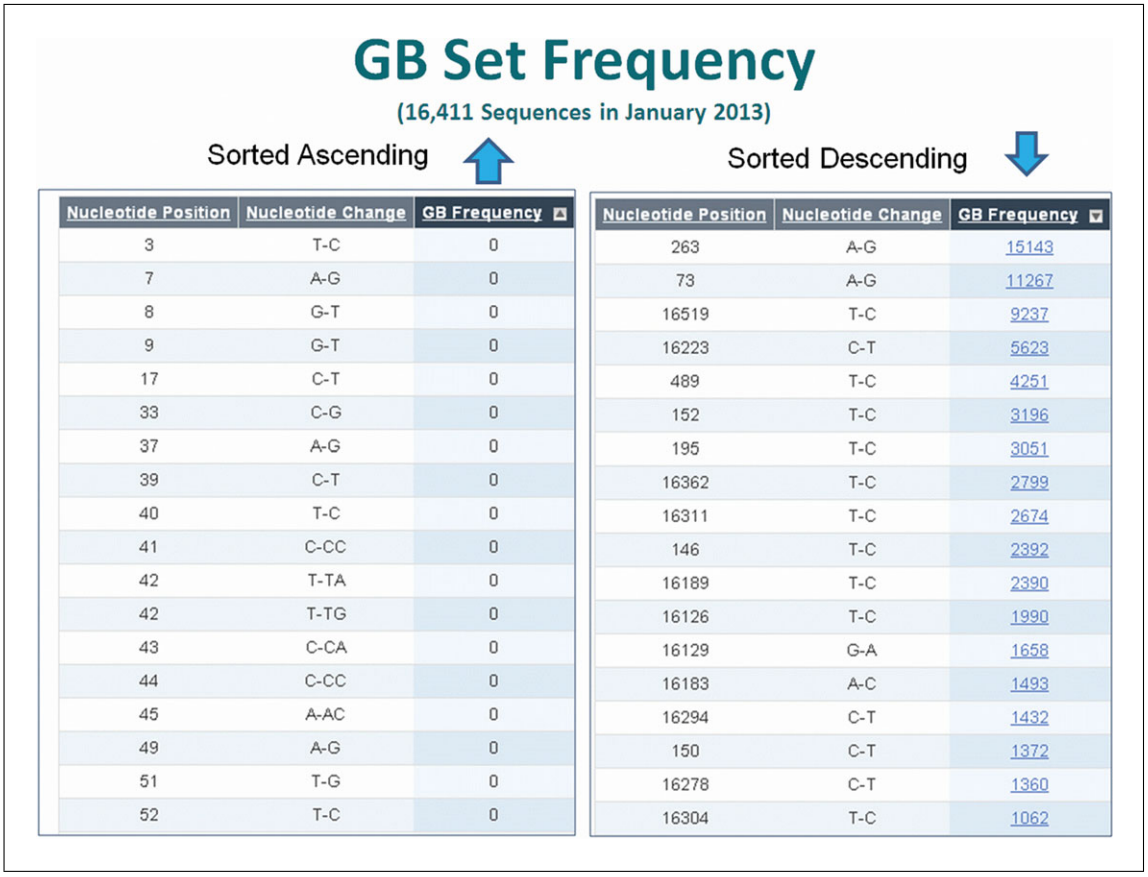
Last Edited: Feb 19, 2013

Locus	Nucleotide Position	Nucleotide Change	Codon number	Codon Position	Amino Acid Change	GB frequency	Curated References
MT-ND3	10373	G-A	105	3	syn	153	references
MT-ND3	10376	A-G	106	3	syn	9	references
MT-ND3	10379	A-G	107	3	syn	2	references
MT-ND3	10385	A-C	109	3	K-N	0	references
MT-ND3	10387	G-C	110	2	G-A	0	references
MT-ND3	10388	T-C	110	3	syn	0	references
MT-ND3	10389	T-C	111	1	syn	15	references
MT-ND3	10391	A-G	111	3	syn	1	references
MT-ND3	10394	C-T	112	3	syn	32	references
MT-ND3	10397	A-G	113	3	syn	119	references
MT-ND3	10398	A-G	114	1	T-A	7081	references

Figure 1.23.14 (**A**) Control Region Variants. (**B**) Outside of the Control Region: Coding and RNA Variants.

`sequence.fasta` file. After clicking on the sequence name, the Alignment Details page will open (Fig. 1.23.29).

The Alignment Detail page (Fig. 1.23.29) shows the rCRS reference position, the query sequence position, the reference base, the query base with respect to the L-strand of the rCRS (NC_012920), the mutation type (substitution type or indel type), the locus or loci intersected, and the predicted transcript effect, if any.



GB Set Frequency

(16,411 Sequences in January 2013)

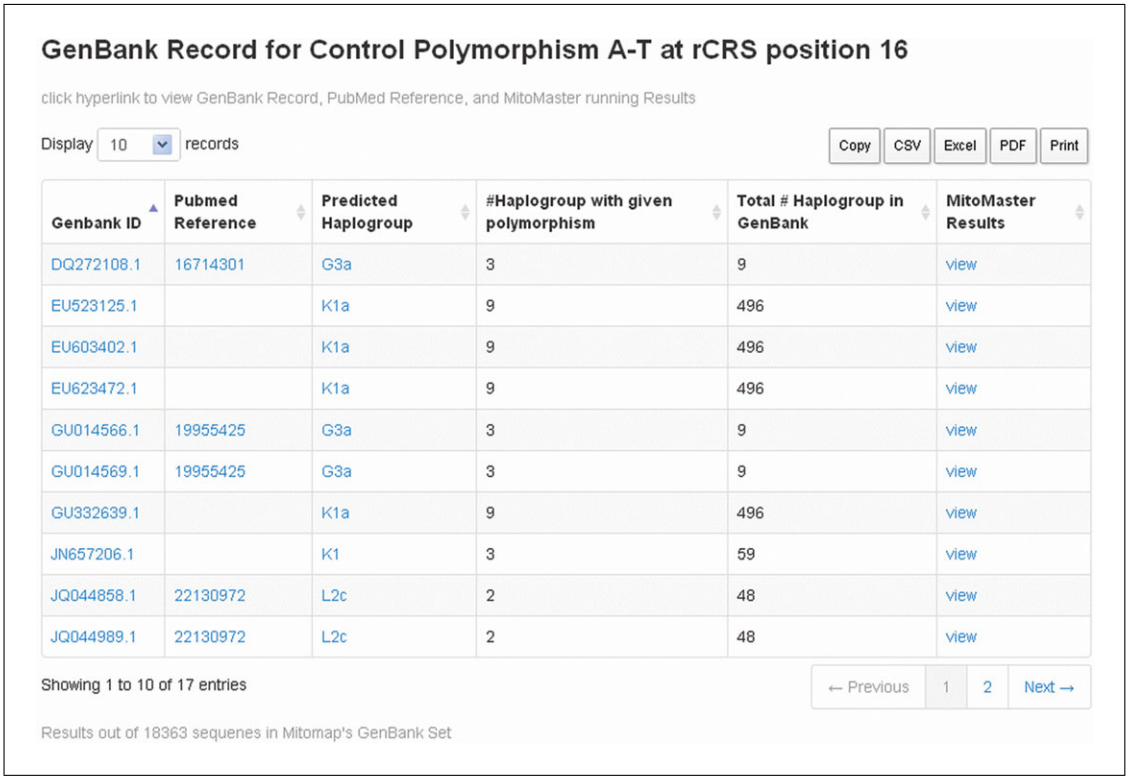
Sorted Ascending

Nucleotide Position	Nucleotide Change	GB Frequency
3	T-C	0
7	A-G	0
8	G-T	0
9	G-T	0
17	C-T	0
33	C-G	0
37	A-G	0
39	C-T	0
40	T-C	0
41	C-CC	0
42	T-TA	0
42	T-TG	0
43	C-CA	0
44	C-CC	0
45	A-AC	0
49	A-G	0
51	T-G	0
52	T-C	0

Sorted Descending

Nucleotide Position	Nucleotide Change	GB Frequency
263	A-G	15143
73	A-G	11267
16519	T-C	9237
16223	C-T	5623
489	T-C	4251
152	T-C	3196
195	T-C	3051
16362	T-C	2799
16311	T-C	2674
146	T-C	2392
16189	T-C	2390
16126	T-C	1990
16129	G-A	1658
16183	A-C	1493
16294	C-T	1432
150	C-T	1372
16278	C-T	1360
16304	T-C	1062

Figure 1.23.15 Variant information can be sorted.



GenBank Record for Control Polymorphism A-T at rCRS position 16

click hyperlink to view GenBank Record, PubMed Reference, and MitoMaster running Results

Display 10 records

Copy CSV Excel PDF Print

Genbank ID	Pubmed Reference	Predicted Haplogroup	#Haplogroup with given polymorphism	Total # Haplogroup in GenBank	MitoMaster Results
DQ272108.1	16714301	G3a	3	9	view
EU523125.1		K1a	9	496	view
EU603402.1		K1a	9	496	view
EU623472.1		K1a	9	496	view
GU014566.1	19955425	G3a	3	9	view
GU014569.1	19955425	G3a	3	9	view
GU332639.1		K1a	9	496	view
JN657206.1		K1	3	59	view
JQ044858.1	22130972	L2c	2	48	view
JQ044989.1	22130972	L2c	2	48	view

Showing 1 to 10 of 17 entries

← Previous 1 2 Next →

Results out of 18363 sequenes in Mitomap's GenBank Set

Figure 1.23.16 Details of GB set frequency.

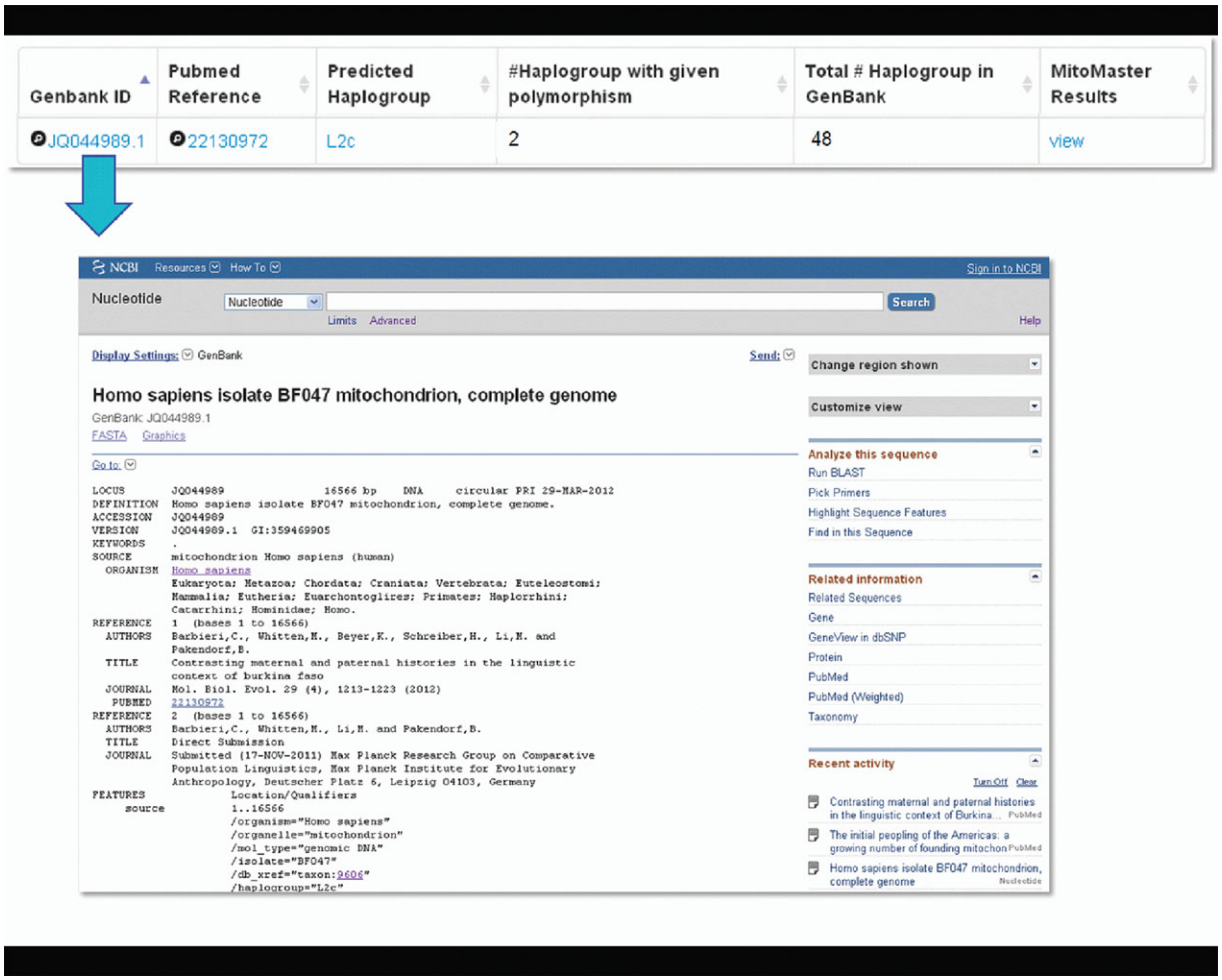


Figure 1.23.17 GenBank ID links to sequence.

8. The amount of information displayed can be adjusted with the drop-down menu at the top, "Display 10 records."

   The default value is 10 records per page, but more records (25, 50, 100, or "All") can be selected.

9. Each column can be sorted by clicking the column headers. To explore the data columns, first click on the header Patient Report twice to sort descending (Fig. 1.23.30).

   In Figure 1.23.30, the Patient Report column shows variants in this sequence that have been published in the literature as possibly having some disease association in patients with Leber Hereditary Optic Neuropathy (LHON), cyclic vomiting, and AD/PD. Do not assume that a listing of a published patient report is a confirmation of a variant's disease association. In this screenshot, 14484C is a well known LHON mutation; however, 3010A is a relatively common variant, found in 17% of all sequences in the mined GenBank set (currently numbering over 18,000) and in 98% of haplogroup J1b sequences. Please read more in the Commentary section about interpreting disease reports.

10. For coding mutations, a Conservation value is calculated based on the user's initial selection of species. By default, 45 species are used.

    a. To explore this, select the first Conservation value (51.11%, circled in red dashes in the above figure).

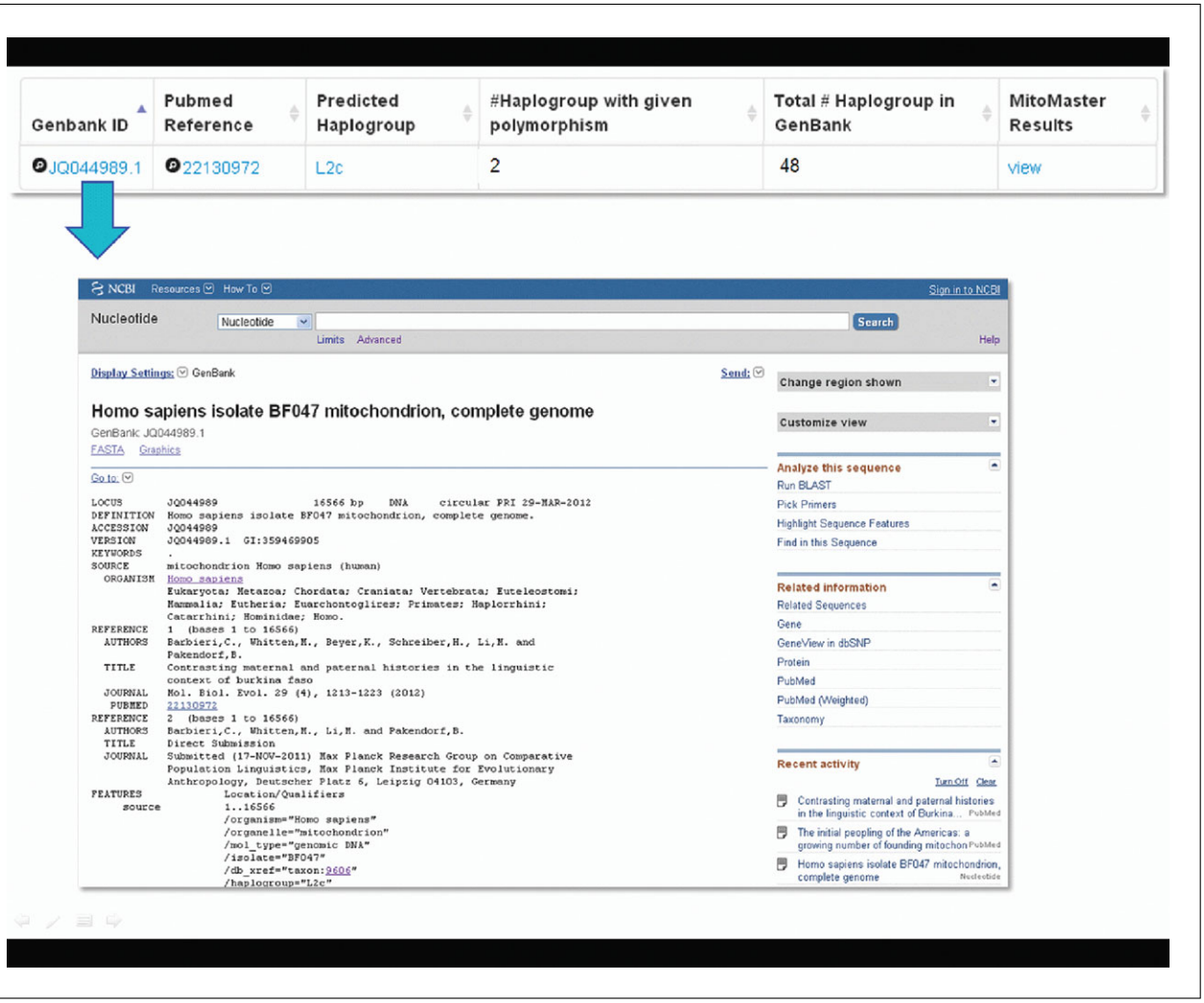


Figure 1.23.18 Pub Med ID links to publication.

b. Open this blue "51.11%" link in a new tab or window by right-clicking if using a PC or control-clicking if using a Mac. This will open the Conservation Grid (Fig. 1.23.31A). Keep this window or tab open.

11. Also open the Species link found in the top menu bar (Fig. 1.23.30, circled in red dots) in a new tab. Keep this window or tab open as well.

12. First, let us examine the Conservation Grid (Fig. 1.23.31A). Amino acid conservation is shown for genic mutations. Note that the query sequence has an A residue while rCRS has a T residue. Finally, 23 of the 45 species (51.11%) have the T residue at that position—note this is the fraction that matches the rCRS value, not the query residue.

13. Now examine the Species Selection tab or window (Fig. 1.23.31B). Species are subcategorized into checkbox groups. Users select which species are used in the conservation index calculations.

14. As an example, click on both Primates and Artiodactyla to open the drop-down view of the contents of these two categories. You many move your mouse to hover over any genus/species name to view the common name. Click each category name again to close the drop downs.

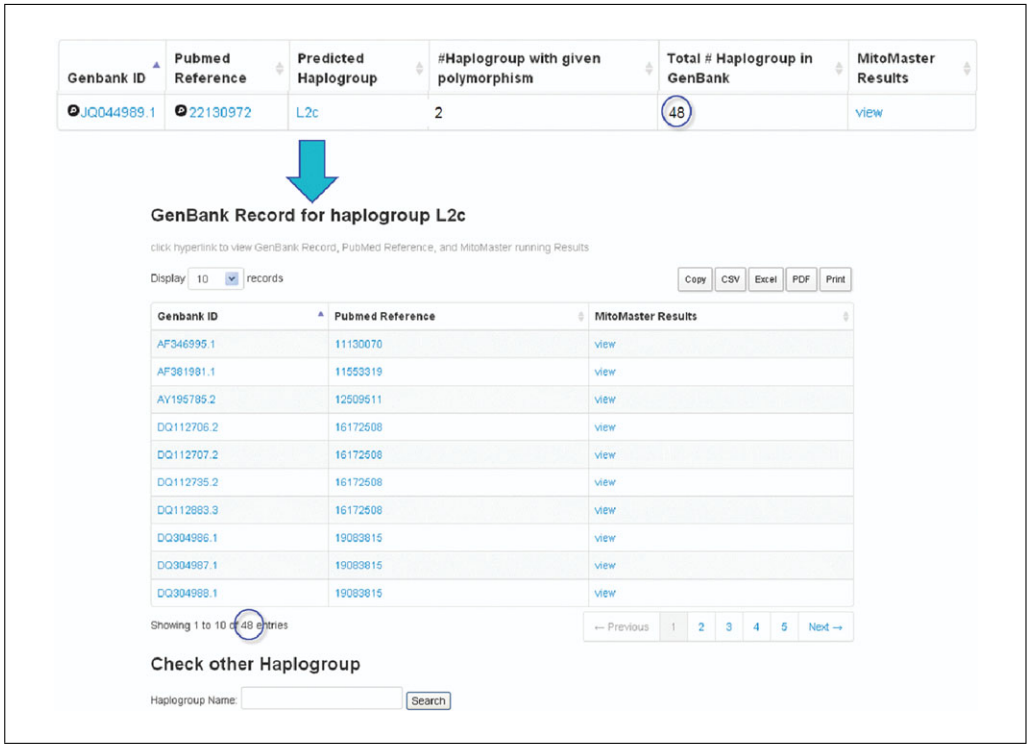


Figure 1.23.19 Predicted Haplogroup links to total listing of sequences with that haplogroup in the GenBank sequence set.

15. Another popular selection choice is to use the Select All, Clear All, or Primates Only options at the top of the page.

16. Click on Primates Only—the change in selected species is instant. Subsequent queries will use this subset of 12 primate species only.

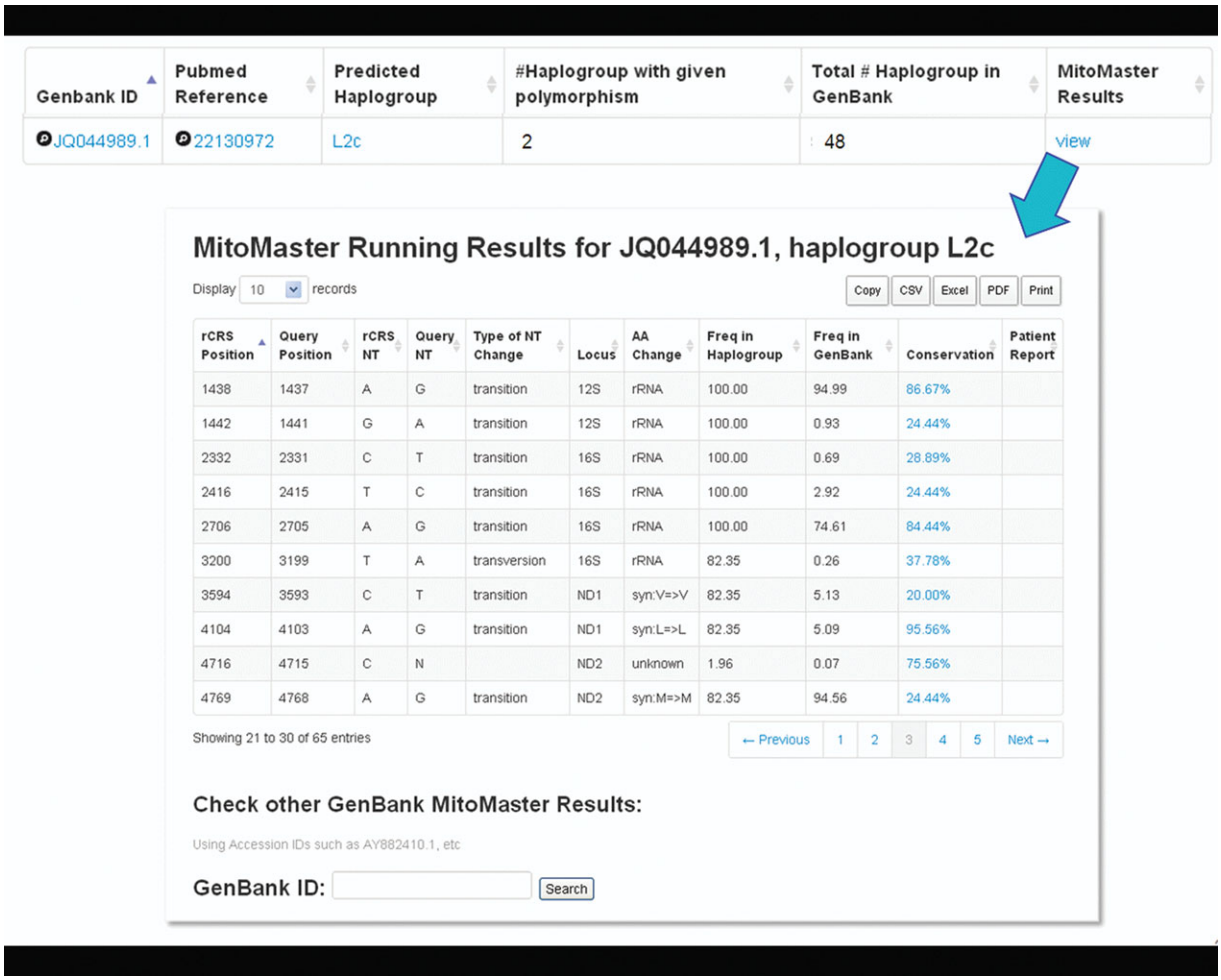
GenBank queries

17. To aid basic research, Mitomaster offers one-step analysis of any human mitochondrial sequences stored in GenBank—104,705 sequences at the time of this publication (17,869 sequences 15.40 kb to 16.66 kb in length plus another 86,836 sequences less than 15.4 kb). One or more GenBank sequence identifiers can be used as query parameters; Mitomaster fetches these sequences using the NCBI Web service and returns reports. Acceptable inputs are a single GenBank accession number or GI sequence identifier (e.g., `EF060316` or `93116889`), a comma-separated list of identifiers with or without intervening spaces (e.g., `EF060316, 302376313, DQ112752` or `EF060316, 302376313, DQ112752`), or a range of sequence numbers (`AF346963-AF346968`). Data reports are produced as described earlier in walk-through of the FASTA sequence analysis. (Fig. 1.23.32)

SNV submission

18. Single Nucleotide Variants can be analyzed alone, without the need for a complete sequence. Data reports are produced as described earlier in walk-through of the FASTA sequence analysis.

    a. Single Nucleotide Variants can be formatted the following ways and may either be pasted into the “copy/paste” window or uploaded as a text file (Fig. 1.23.33).
    b. A four-column tab-delimited format consisting of sample name, rCRS mitochondrial position, reference base, and query base (sample, pos, ref, and var; Fig. 1.23.34).



Genbank ID	Pubmed Reference	Predicted Haplogroup	#Haplogroup with given polymorphism	Total # Haplogroup in GenBank	MitoMaster Results
JQ044989.1	22130972	L2c	2	48	view

MitoMaster Running Results for JQ044989.1, haplogroup L2c

Display 10 records

Copy CSV Excel PDF Print

rCRS Position	Query Position	rCRS NT	Query NT	Type of NT Change	Locus	AA Change	Freq in Haplogroup	Freq in GenBank	Conservation	Patient Report
1438	1437	A	G	transition	12S	rRNA	100.00	94.99	86.67%	
1442	1441	G	A	transition	12S	rRNA	100.00	0.93	24.44%	
2332	2331	C	T	transition	16S	rRNA	100.00	0.69	28.89%	
2416	2415	T	C	transition	16S	rRNA	100.00	2.92	24.44%	
2706	2705	A	G	transition	16S	rRNA	100.00	74.61	84.44%	
3200	3199	T	A	transversion	16S	rRNA	82.35	0.26	37.78%	
3594	3593	C	T	transition	ND1	syn:V=>V	82.35	5.13	20.00%	
4104	4103	A	G	transition	ND1	syn:L=>L	82.35	5.09	95.56%	
4716	4715	C	N		ND2	unknown	1.96	0.07	75.56%	
4769	4768	A	G	transition	ND2	syn:M=>M	82.35	94.56	24.44%	

Showing 21 to 30 of 65 entries

← Previous 1 2 3 4 5 Next →

Check other GenBank MitoMaster Results:

Using Accession IDs such as AY882410.1, etc

GenBank ID: Search

Figure 1.23.20 Mitomaster analysis report on each sequence.

MITOMAP: Reported Mitochondrial DNA Base Substitution Diseases: Coding and Control Region Point Mutations

Last Edited: Feb 18, 2013

Locus	Disease	Allele	Nucleotide Position	Nucleotide Change	Amino Acid Change	Homo-plasmy	Hetero-plasmy	Status	References
MT-CO1	MM & Rhabdomyolysis	G6708A	6708	G-A	G-Ter	-	+	Reported	references
MT-CO1	Acquired Idiopathic Sideroblastic Anemia	T6721C	6721	T-C	M-T	-	+	Reported	references
MT-CO1	Acquired Idiopathic Sideroblastic Anemia	T6742C	6742	T-C	I-T	-	+	Reported	references
MT-CO1	Multisystem Disorder	G6930A	6930	G-A	G-Ter	-	+	Reported	references
MT-CO1	Mild EXIT and MR	G6955A	6955	G-A	G-D	+	+	Reported	references
MT-CO1	MELAS-like syndrome	G7023A	7023	G-A	V-M	-	+	Reported	references
MT-CO1	Prostate Cancer	G7041A	7041	G-A	V-I	+	-	Reported	references
MT-CO1	Prostate Cancer	T7080C	7080	T-C	F-L	+	-	Reported	references
MT-CO1	Prostate Cancer	A7083G	7083	A-G	I-V	+	-	Reported	references
MT-CO1	Prostate Cancer	A7158G	7158	A-G	I-V	+	-	Reported	references
MT-CO1	Prostate Cancer	A7305C	7305	A-C	M-L	+	-	Reported	references
MT-CO1	DEAF	A7443G	7443	A-G	Ter-G	+	-	Reported	references
MT-CO1	LHON/SNHL/DEAF	G7444A	7444	G-A	Ter-K	+	-	Reported	references
MT-CO1	DEAF	A7445C	7445	A-C	Ter-S	+	-	Reported	references
MT-CO1	SNHL	A7445G	7445	A-G	Ter-Ter	+	+	Cfrm	references
MT-CO2	Mitochondrial Encephalomyopathy	T7587C	7587	T-C	M-T	-	+	Reported	references
MT-CO2	Possible LHON helper variant	G7598A	7598	G-A	A-T	-	+	Reported	references

Figure 1.23.21 Reported Coding and Control Region mutations found in patients.

MITOMAP: Reported Mitochondrial DNA Base Substitution Diseases: rRNA/tRNA mutations

Last Edited: Nov 25, 2012

Locus	Disease	Allele	RNA	Homo-plasmy	Hetero-Plasmy	Status	References
MT-TF	Mitochondrial myopathy	T582C	tRNA Phe	-	+	Reported	references
MT-TF	MELAS / MM & EXIT	G583A	tRNA Phe	-	+	Cfrm	references
MT-TF	Extrapyramidal disorder with akinesia-rigidity, psychosis and SNHL	G586A	tRNA Phe	-	+	Reported	references
MT-TF	Axial myopathy with encephalopathy	C602T	tRNA Phe	-	+	Reported	references
MT-TF	Myoglobinuria	A606G	tRNA Phe	+	+	Unclear	references
MT-TF	Tubulo-interstitial nephritis	A608G	tRNA Phe	+	-	Reported	references
MT-TF	MERRF	G611A	tRNA Phe	-	+	Reported	references
MT-TF	Maternally inherited epilepsy	T616C/G	tRNA Phe	+	+	Reported	references
MT-TF	MM	T618C	tRNA Phe	-	+	Reported	references
MT-TF	Ptosis CPEO MM & EXIT	T618G	tRNA Phe	-	+	Reported	references
MT-TF	EXIT & Deafness	G622A	tRNA Phe	-	+	Reported	references
MT-TF	SNHL & Epilepsy	G625A	tRNA Phe	-	+	Reported	references
MT-TF	DEAF	A636G	tRNA Phe	+	-	Reported	references

Figure 1.23.22 Reported rRNA and tRNA mutations found in patients.

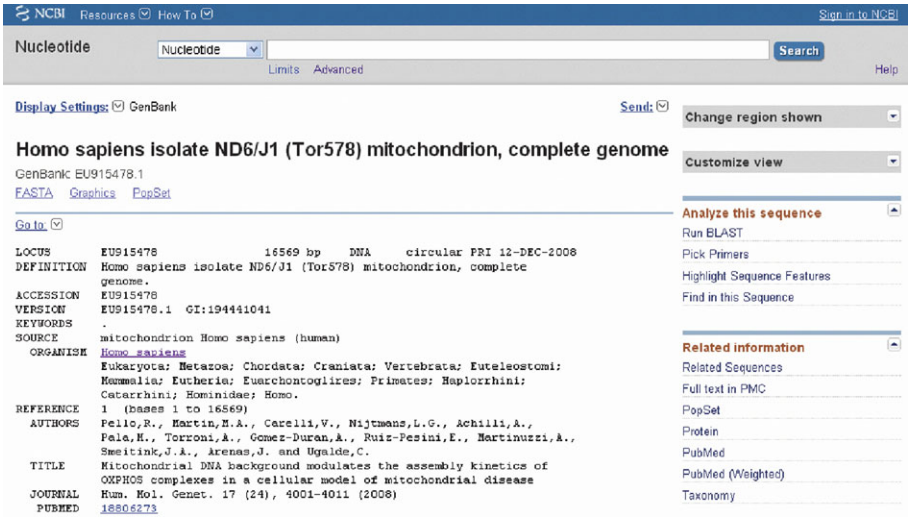


Figure 1.23.23 Sample Sequence EU915478 at GenBank.

c. A more compact non-delimited reference-position-query format is also permitted, with or without sample names (Fig. 1.23.35).
d. Indels can be represented using the adjacent-base notation (C573CC) or with the use of decimal positions (573.1C), explicitly (573insC), or using dash or colon symbols (:573C) as shown in Figure 1.23.36.

GUIDELINES FOR UNDERSTANDING RESULTS

With the recent explosion of sequence data, Mitomap has found it necessary to augment the hand-curated portion of the database with mined sequence data from GenBank. The new GenBank frequency data are derived from sequences with size equal to or larger than the complete coding region. These sequences have been pre-loaded into Mitomaster and represent almost all haplogroups known to date. As of August 2013, this data set contained over 18000 sequences. The size of the sequence set is expected to increase with quarterly scans of GenBank and possibly other public sequence repositories. Keep

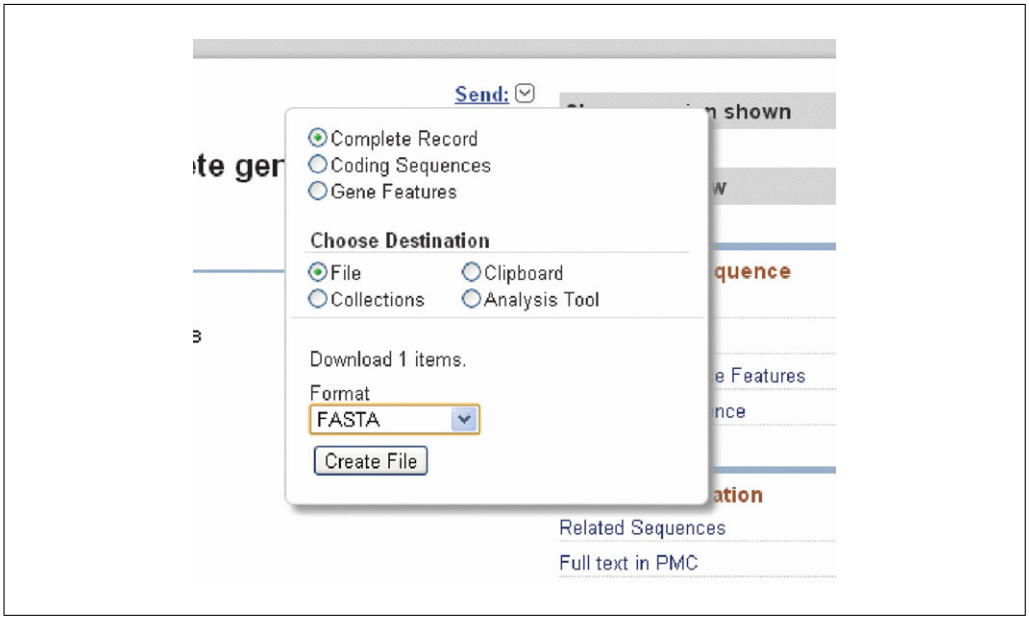


Figure 1.23.24 Send FASTA dialog box.

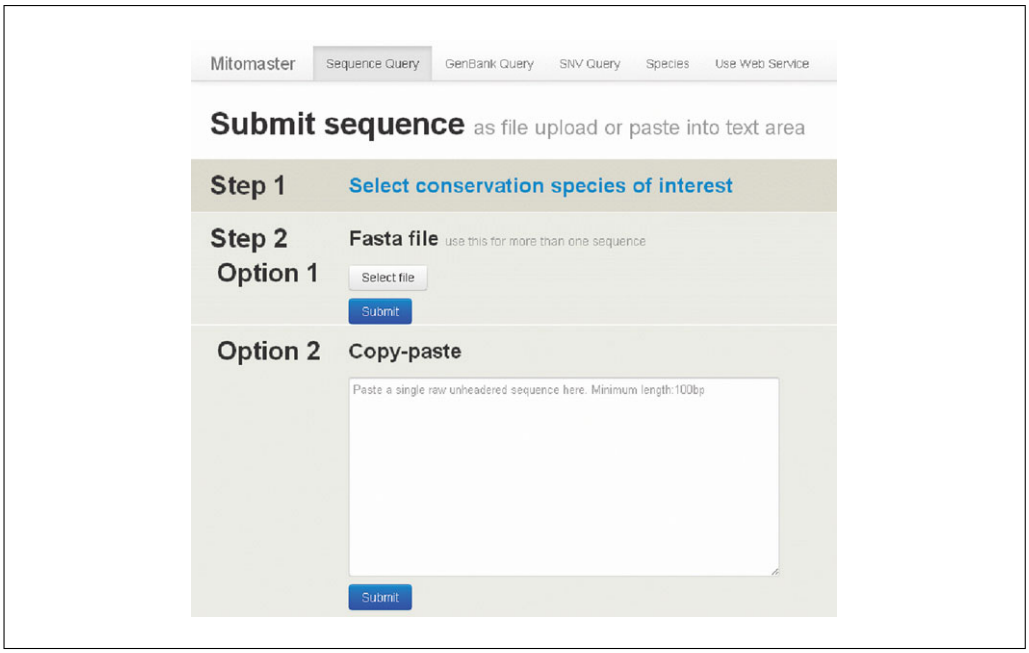


Figure 1.23.25 Submit Sequence screen.

in mind that pre-loaded sequences from GenBank have not been individually reviewed by Mitomap.

GenBank sequences may not be of equal quality (Yao et al., 2009). Published results may have sequencing errors, partial data, and analysis mistakes that, even if corrected, might still not be downloadable as a corrected sequence directly from GenBank and might only be found as published erratum to the corresponding publications.

When considering the frequency of any particular variant in GenBank, one must understand that the numbers are not the true worldwide frequency but are only a reflection of the actual sequences currently collected. Many populations in the world are underrepresented in sequence repositories due to remote location and economics. Some sequences in public databases are occasionally duplicated with different record numbers

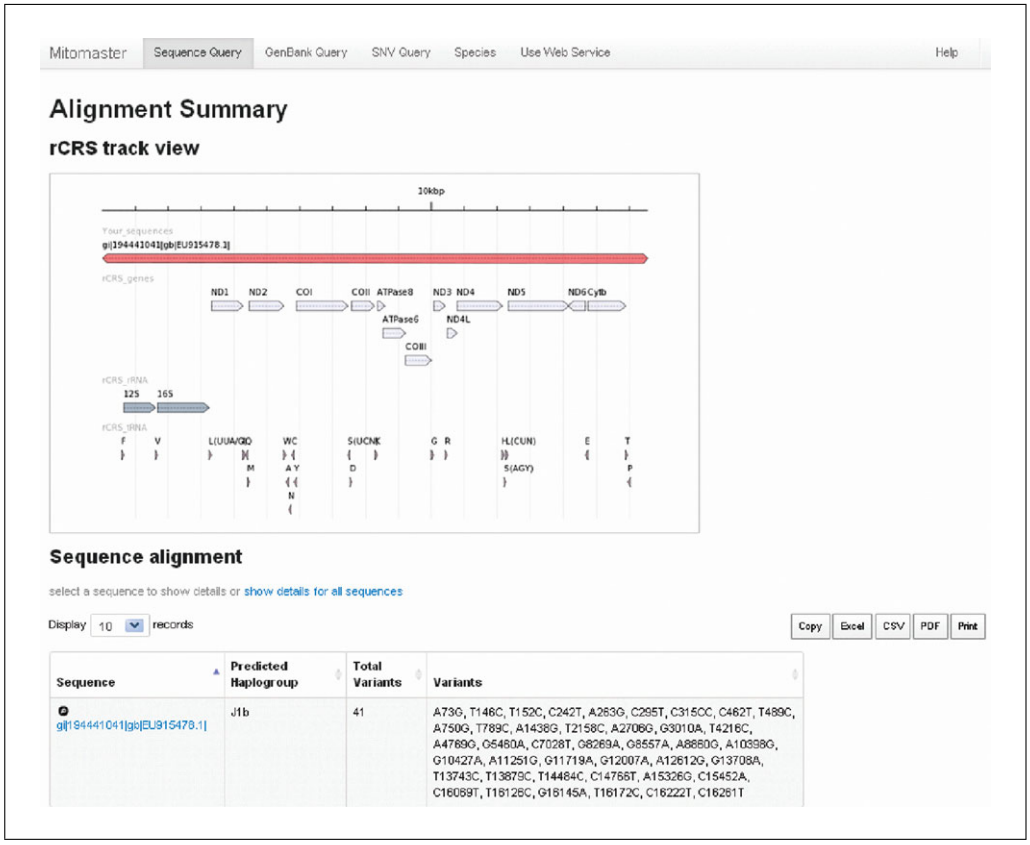


Figure 1.23.26 Alignment Summary.

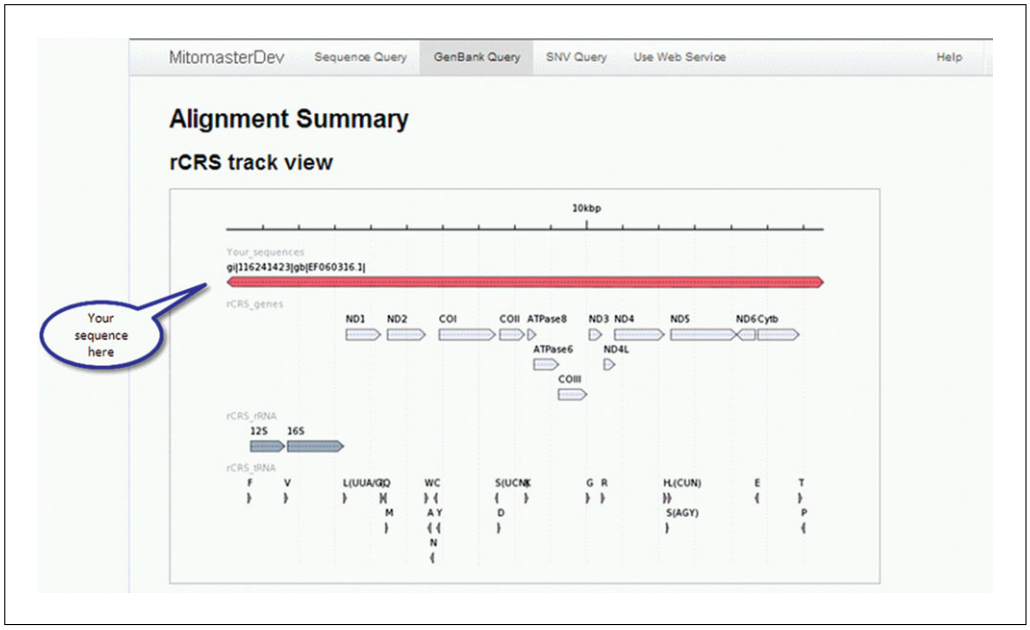


Figure 1.23.27 rCRS Track View.

and therefore do not represent unique individuals. Also, multiple sequences might be from closely related family members with identical mitochondrial DNA.

More and more sequences from patients with known or suspected mitochondrial disease are now being banked in public databases. It is important to keep in mind that any human mitochondrial sequence has the potential to contain variants relevant to past, current, or future disease. This is especially true in the case of diseases that develop later in life or

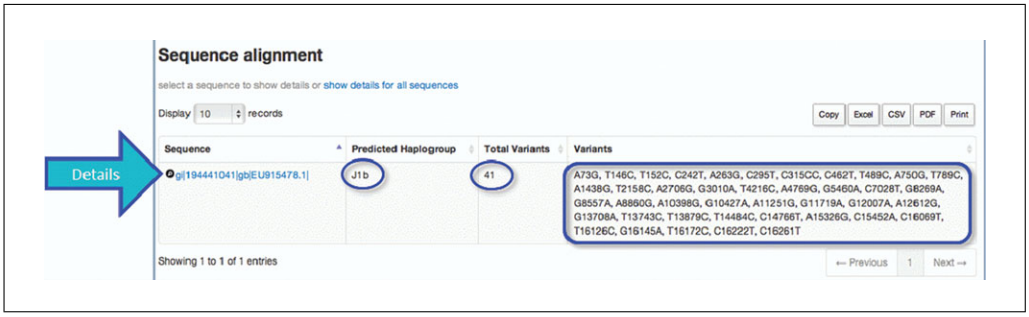
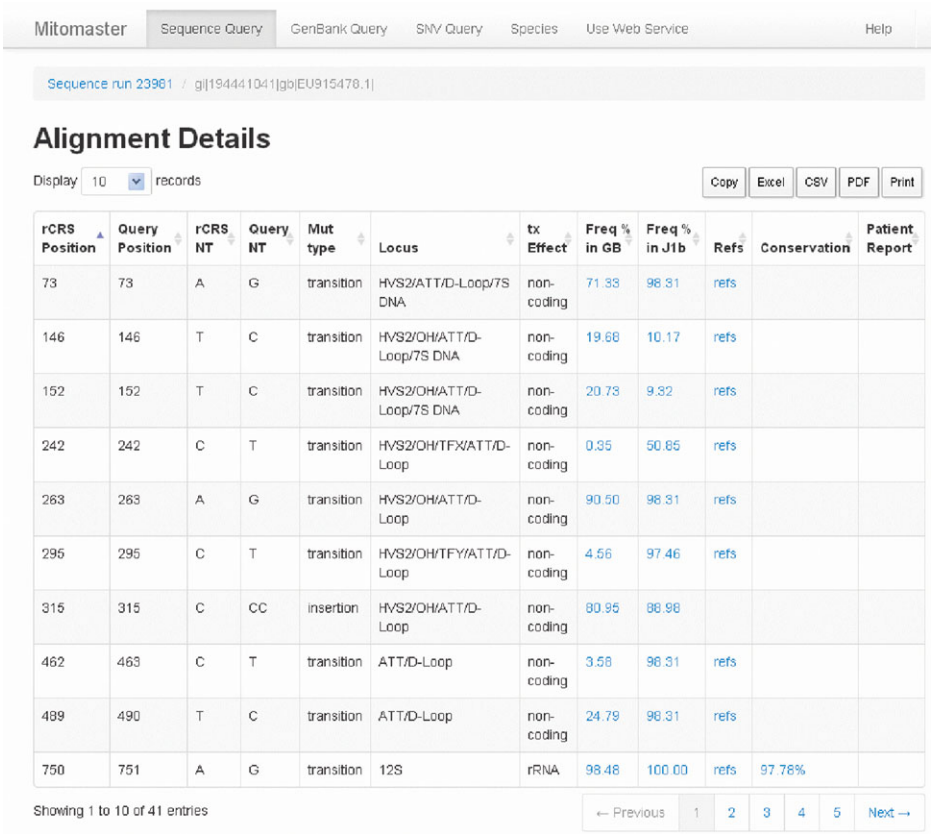


Figure 1.23.28 Sequence Alignment, areas of interest circled.



rCRS Position	Query Position	rCRS NT	Query NT	Mut type	Locus	tx Effect	Freq % in GB	Freq % in J1b	Refs	Conservation	Patient Report
73	73	A	G	transition	HVS2/ATT/D-Loop/7S DNA	non-coding	71.33	98.31	refs		
146	146	T	C	transition	HVS2/OH/ATT/D-Loop/7S DNA	non-coding	19.68	10.17	refs		
152	152	T	C	transition	HVS2/OH/ATT/D-Loop/7S DNA	non-coding	20.73	9.32	refs		
242	242	C	T	transition	HVS2/OH/TFX/ATT/D-Loop	non-coding	0.35	50.65	refs		
263	263	A	G	transition	HVS2/OH/ATT/D-Loop	non-coding	90.50	98.31	refs		
295	295	C	T	transition	HVS2/OH/TFY/ATT/D-Loop	non-coding	4.56	97.46	refs		
315	315	C	CC	insertion	HVS2/OH/ATT/D-Loop	non-coding	80.95	88.98			
462	463	C	T	transition	ATT/D-Loop	non-coding	3.58	96.31	refs		
489	490	T	C	transition	ATT/D-Loop	non-coding	24.79	98.31	refs		
750	751	A	G	transition	12S	rRNA	98.48	100.00	refs	97.78%	

Figure 1.23.29 Alignment Details.

after particular environment stressors, and thus might not be evident in an individual at the time of sequencing.

COMMENTARY

Background Information

Mitomap had its beginnings as a 1994 report to the Human Genome Committee, appearing in print form in 1995 (Wallace et al., 1995). This was the first attempt to record and document all published human mitochondrial DNA variation in the general population as well as summarize continent-specific and disease-specific variants. It also served as a unified reference for human mitochondrial gene loci. Mitomap was first launched as an online mtDNA database in 1996, and has been an essential tool for genetic researchers, counselors, and clinicians ever since. With its well documented and hand-curated dataset, Mitomap allows users to browse or search for reported mtDNA variants from both the general and clinical disease populations.

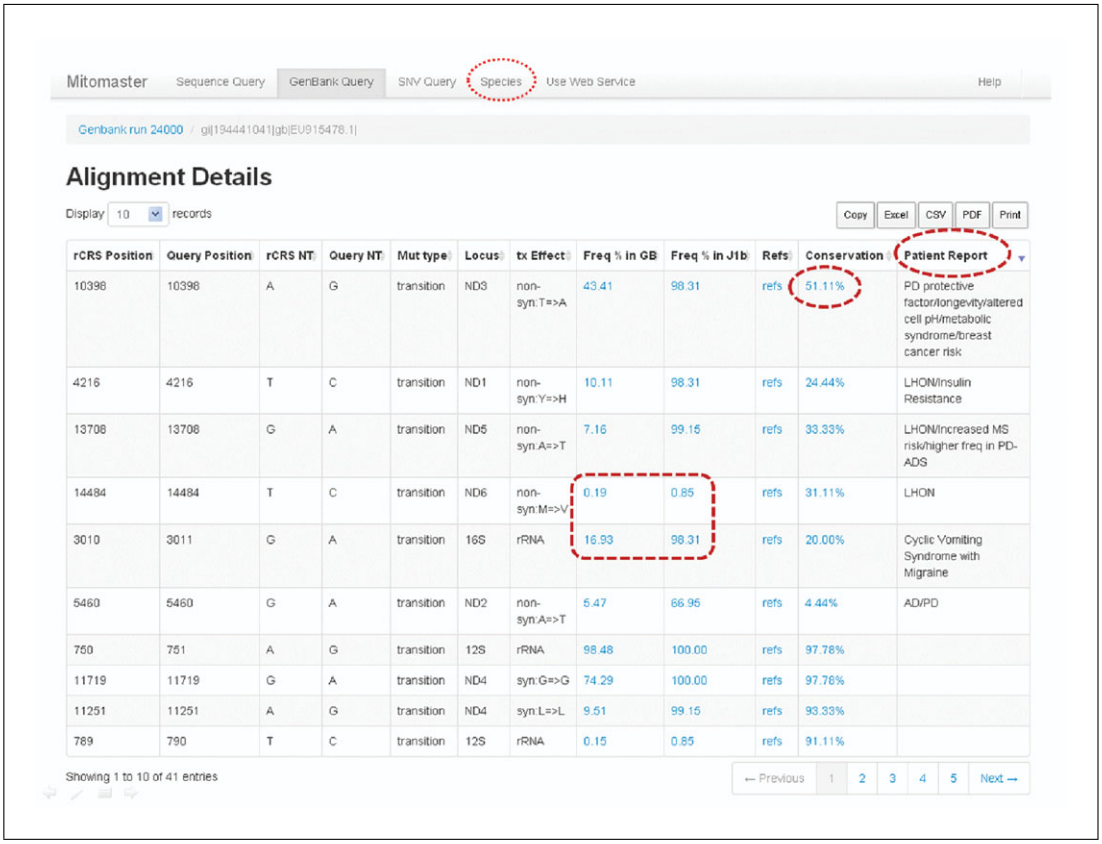


Figure 1.23.30 Alignment Details with areas of interest circled.

Critical Parameters

Alignment and positional caveats

Mitomaster uses a standard pair-wise BLAST local alignment to determine substitutions and mismatches. A local alignment will only include subsequences to achieve optimal alignment given a set of parameters, and so areas of extreme dissimilarity will be unaligned and no results will be given. This can result in truncated alignments or alignments composed of discontinuous high-scoring pairs. The genome viewer on the summary page shows the portion of the rCRS reference covered by the query sequence.

Calculation of a mutation's effect is done on an individual basis, isolated from any potential interactions of other mutation events within the query sequence. For instance, an insertion followed closely by a deletion within a protein-coding gene would result in both being reported as independent frameshift mutations.

Also, be aware that insertions and deletions in regions of nucleotide polytracts (for example, "ACCCCCT") are often reported using different conventions. Some publications list the insertion/deletion point at the beginning of nucleotide run, some at the end. Further complications arise in repetitive polytract regions (for example, "CCCCCTCTACCCCCTCTAG"). This will often result in inaccurate sequence frequency numbers when Mitomaster scans published indel reports for matches with GenBank sequences.

Large data sets

Large data sets may take a long time to process in Mitomaster, depending upon the number of FASTA files submitted and the server load at the time of submission. To allow high-throughput sequence analyses using programmatic scripts, a Web service utility is provided. This service is for use by programmers and database support personnel. Data may be submitted in large file batches (e.g., 100 or more sequences) to the Mitomaster Web service via script clients to programmatically submit queries using a simple POST mechanism. Sample client scripts in Perl and Python are provided on the Mitomaster site. Input is by FASTA file, and output is in a tab-delimited text format with similar content as delivered via the interactive Web application.

Pathogenicity cautions

Mitochondrial DNA is highly variable within and between populations. New variants will continue to be discovered as more and more sequencing studies are performed.

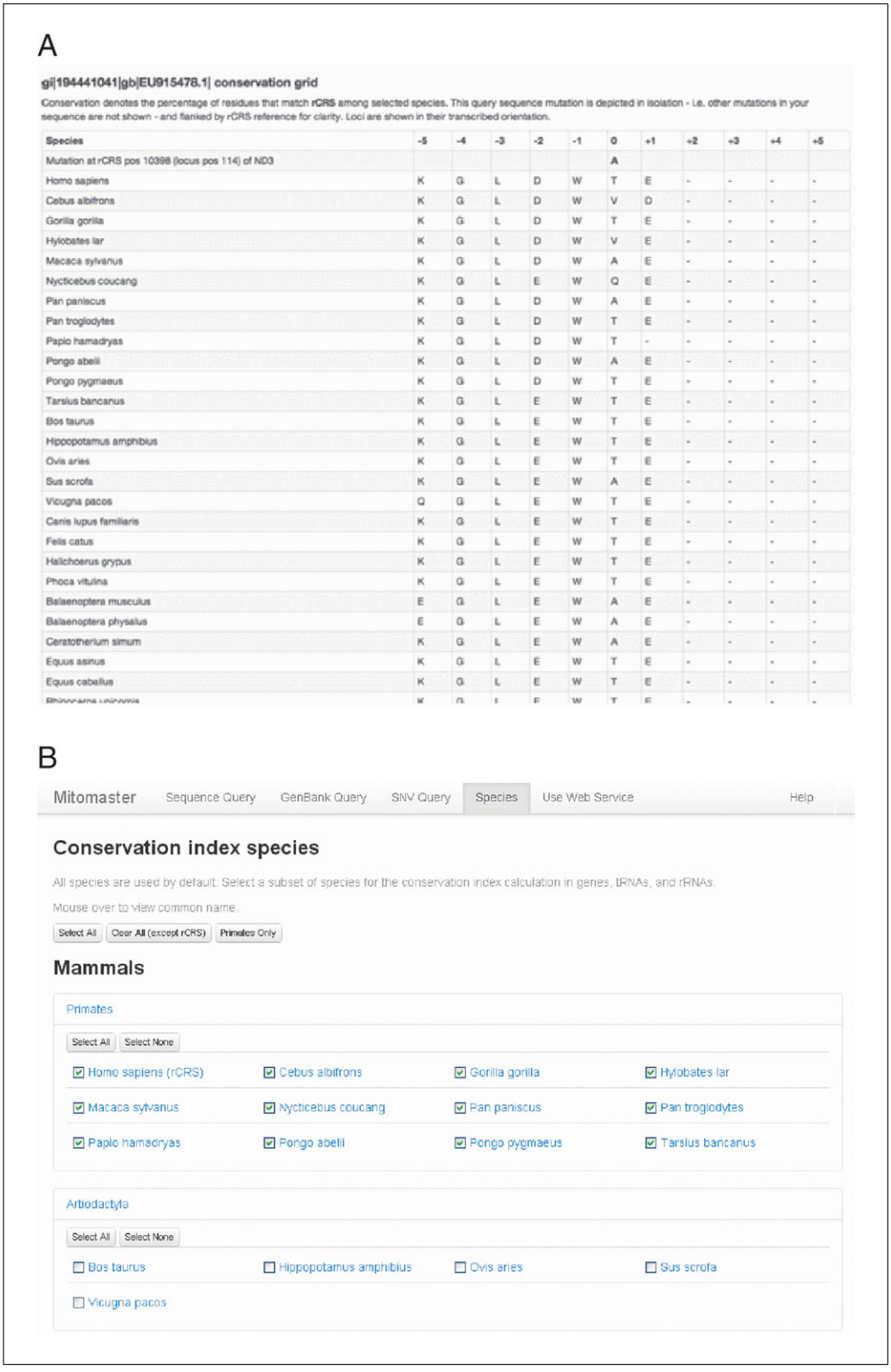


Figure 1.23.31 (**A**) Conservation Grid. (**B**) Conservation Grid of Species.

It is possible that some sporadic mutations as well as known haplogroup-defining or polymorphic variants might be involved in a disease, but to make any conclusions concerning pathogenicity, more evidence and data analyses are required. Caution is advised when using the listing of mtDNA variants found in patient groups. A status of Confirmed ("Cfrm") in Mitomap is not an assignment of pathogenicity but is a general consensus of what is reported in published literature. Researchers and clinicians are advised that

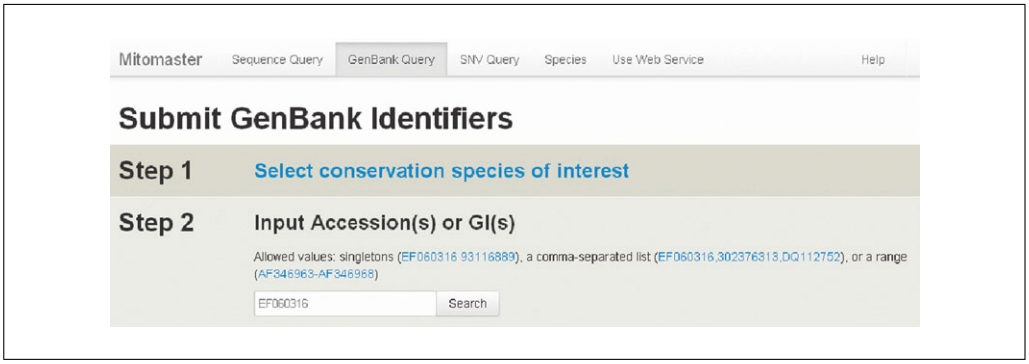


Figure 1.23.32 Submitting GenBank Identifiers.

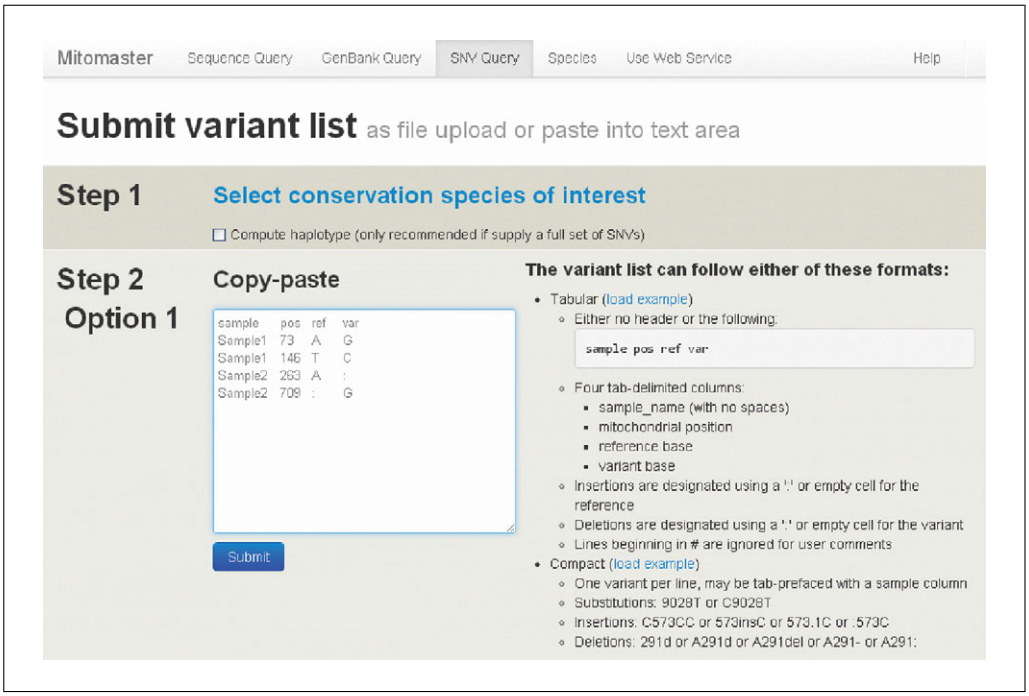


Figure 1.23.33 Single Nucleotide Variant Submission.

With header row				Without header row			
Sample	pos	ref	var	Sample1	73	A	G
Sample1	73	A	G	Sample1	146	T	C
Sample1	146	T	C	Sample2	263	A	:
Sample2	263	A	:	Sample2	709	:	G
Sample2	709	:	G				

Figure 1.23.34 SNV submission format option 1: a four-column tab-delimited format consisting of sample name, rCRS mitochondrial position, reference base, and query base (sample, pos, ref and var).

additional data, searches of other databases, and/or analyses are usually required to confirm the pathological significance of some of these mutations. Such due diligence will also reduce the number of false reports of "novel" and "pathogenic" mtDNA variants in the literature.

Haplogroups

Haplogroup definition, while in some cases possible on a truncated sequencing data set, is best done using a complete mtDNA sequence. Caution is advised against the popular "allelic" analysis approach where a researcher is only looking at a single nucleotide variant

With sample names		Without sample names	Without rCRS value
Sample1	A73G	A73G	73G
Sample2	T146C	T146C	146C

Figure 1.23.35 SNV submission format option 2: a more compact non-delimited reference-position-query format is also permitted, with or without sample names.

With sample names		Without sample names	Explicit /decimal format	-/: format
Sample123	A263d	A263d	263delA	A263-
Sample456	C573CCC	573CC	573.1insC	:573C

Figure 1.23.36 SNV submission format option 3: indels can be represented using the adjacent-base notation (C573CC) or with the use of decimal positions (573.1C), explicitly (573insC), or using dash or colon symbols (:573C).

in relation to a disease phenotype. This approach is not directly applicable to mtDNA analysis. There are sets of different variants that are linked together and define mtDNA highly hierarchical phylogeny and haplogroup assignments. Some variants are polymorphic and could be found in many different mitochondrial lineages on different haplogroup backgrounds and some could be haplogroup defining and rarely found anywhere else, while there are variants that are haplogroup-defining and polymorphic at the same time. The latter variants could be pathogenic on a different (non-defining) haplogroup background or in a different environment, but be nonpathogenic in their defining haplogroup. Thus, they might occasionally be found in Mitomap as being listed in both the polymorphism table and the disease table. Ethnic, geographical, and historic factors can also come into play when attempting to correlate a haplogroup with a medical condition.

Acknowledgments

This work was supported by NIH grants NS21325, NS070298, AG24373, and DK73691. Also, the Simons Foundation grant 205844 was awarded to D.C.W.

Literature Cited

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R., and Young, I.G. 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290:457-465.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* 23:147.

Kloss-Brandstatter, A., Pacher, D., Schonherr, S., Weissensteiner, H., Binna, R., Specht, G., and Kronenberg, F. 2011. HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* 32:25-32.

van Oven, M. and Kayser, M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat. (Online)* 30:E386-E394.

Wallace, D.C., Lott, M.T., Brown, M.D., Huoponen, K., and Torroni, A. 1995. Report of the committee on human mitochondrial DNA. *In* Human Gene Mapping 1994, a Compendium (A.J. Cuticchia ed.) pp. 910-954. The Johns Hopkins University Press, Baltimore, Md.

Yao, Y.G., Salas, A., Logan, I., and Bandelt, H.J. 2009. mtDNA data mining in GenBank needs surveying. *Am. J. Hum. Genet.* 85:929-933.

Key References

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D. and Birney, E. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12:1611-1618. *Supporting publication for Bioperl.*

Steinbiss, S., Gremme, G., Scharfer, C., Mader, M., and Kurtz, S. 2009. AnnotationSketch: A genome annotation drawing library. *Bioinformatics* 25:533-534. *Supporting publication for AnnotationSketch*

Yao et al., 2009. See above.
Important caveats about use of data mined from public sequences.

Zaragoza, M.V., Brandon, M.C., Diegoli, M., Arbustini, E. and Wallace, D.C. 2011. Mitochondrial cardiomyopathies: How to identify candidate pathogenic mutations by mitochondrial DNA sequencing, MITOMASTER and phylogeny. *Eur. J. Hum. Genet.* 19:200-207.
Using MITOMASTER to investigate pathogenicity.

Internet Resources

http://www.mitomap.org
Mitomap.

http://mitomaster.mitomap.org
Mitomaster.

http://www.phylotree.org
Phylotree.

http://haplogrep.uibk.ac.at/
Haplogrep.

http://www.ncbi.nlm.nih.gov
NCBI.

http://blast.ncbi.nlm.nih.gov/
BLAST.

http://www.ncbi.nlm.nih.gov/nuccore/NC_012920.1
The rCRS: Homo sapiens mitochondrion, complete genome.

http://www.bioinformatics.nl/tools/crab_fasta.html
FASTA format.