

YONSEI UNIVERSITY COLLEGE OF MEDICINE

Abstract

Mitochondrial DNA (mtDNA) has become a potent tool in population-, medical-, and forensic genetics investigation. However, the analysis and comparison of mtDNA sequences may not be easy to do for researchers who are not familiar with mtDNA nomenclature conventions. Therefore, we developed a web program, mtDNAprofiler, which enables users to analyze and compare mtDNA sequences whether they are indicated as FASTA format or as a difference to revised Cambridge Reference Sequence (rCRS). The mtDNAprofiler consists of two systems; mtDNA nomenclature system by alignment with the rCRS, and mtSNP conversion and comparison system. mtDNA nomenclature system sequentially performs range determination and mtSNP calculation for input sequences followed alignment with the rCRS. In mtSNP conversion and comparison system, conversion mode allows mtDNA data indicated as a difference to the rCRS to be converted to FASTA sequence so that all the difference-coded mtSNP data followed various alignment rules may be easily used for further analyses. Comparison mode provides comparison results between two mtSNP data gathered independently or generated by different alignment rules, thereby ensuring the validity of the input mtSNP nomenclature. Therefore, the mtDNAprofiler will provide useful tools to characterize and analyze mtDNA sequences for researcher dealing with mtDNA. mtDNAprofiler is available at http://mtprofiler.yonsei.ac.kr.

Background

Forensic mitochondrial DNA (mtDNA) variations are described and reported as a list of difference to revised Cambridge Reference Sequence (rCRS), namely difference coded single nucleotide polymorphism (mtSNP), which is obtained by alignment and comparison of sample sequence against rCRS. However, the alignment of mtDNA sequences varies mostly in the vicinity of homopolymeric or dinucleotide-repeat regions that are prone to generate insertions or deletions according to their length. Furthermore, the position of insertion or deletion can varies on alignment result, since there are many different alignment rules according to the program or algorithm.

As a first attempt to standardize the description of length variations in mtDNA, Wilson et al. proposed an alignment and nomenclature protocol (Wilson Rules). Recently, Budowle et al. developed a automated program using a hierarchical set of rules, called the Mitotyper Rules to accomplish absolute consistency and stability in mtDNA nomenclature.

We present here a web-based and publicly available program, mtDNAprofiler, which is composed of mtDNA nomenclature tool and mtSNP conversion and comparison tool, providing users with easy nomenclature and typing from mtDNA sequence and consistent management of mtSNP data, respectively.

Results

1. mtDNA nomenclature system

1-1. Alignment protocol



Alignment in mtDNAprofiler basically uses parsimonious Smith-Waterman algorithm and follows Mitotyper Rules. Its protocol is composed of two main rules as following: least number of differences and indels (insertions and deletions) detection and rearrangement. Second rule is further divided into three sub-rules, which selects one rule according to the region such as AC repeat, non-HV2 C-stretch, or HV2 C-stretch regions.

1-2. Screen shot

A	mtDNApro	fíler	mtDNA nome	nclatur	e tool					
	Please provide nucleotide sequence in FASTA format. >AB241275_samp1e_YM89 TIGGGTACCACCCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCA									
	↑ Start page FASTA format	of mtDN.	A nomenc	lature	tool. Inp	out data	© Coding re	cleotide	e seque	nce is u
B				- 7 u	Results	laorrogion	obuilight		Download	button
	 I. AB241275_SAM a. Range : 16045 ~ 16 b. Alignment with rCRS x <l< th=""><th>IPLE_YM89 569, 1 ~ 543 ← View new wind CRS IIGGGTACC ← III</th><th>- Range dow ← Butto 016060 ACCCAAGTATTGAC</th><th>on to viev 16070. ICACCCAT Alignm</th><th>v alignment in </th><th>new windc 16090 STATTICGIAC</th><th>DW 16100 ATTACTGCCA ATTACTGCCA</th><th>c. Mini-m</th><th>of mtSNP Origin (H) Mini-map</th><th></th></l<>	IPLE_YM89 569, 1 ~ 543 ← View new wind CRS IIGGGTACC ← III	- Range dow ← Butto 016060 ACCCAAGTATTGAC	on to viev 16070. ICACCCAT Alignm	v alignment in 	new windc 16090 STATTICGIAC	DW 16100 ATTACTGCCA ATTACTGCCA	c. Mini-m	of mtSNP Origin (H) Mini-map	
	d. mtSNP profile	Control region	np 438 - 576	mtS	NP profile		Coding regio	on		
	16223 16287 16319 16362 16399	73 263 309.1C 315.1C 431	489 523d 524d							
e. Sequence TTGGGTACCACCCAGGTATTGACTCACCCATCAACAACCGCTATGTATTTCGTACATTACTCCCACCACCATGAATATTGTACGGTACCATAAATATTACATAAAAACCCAATCCAACTACAAACCCCCTTCCCCCAGGTTACAA TACATAAAAACCCAATCCAACTAAAACCCCCTTCCCCCAGGTTACAA ACTAGGATACCAACAAACTTACCACCCTTAACAGTACATAGTACAT BCACATTACAGTCAAACTTACCCACCCTTAACAGTACATAGTACAT						ctigaccacci agccaccccic catggaigacc e of sar	Terrag Taccc Tocccc mple sec			
	alignment re- alignment. In And downloa sample seque	sult with r addition, d functior ence (Fig 1	CRS, mini alignment for the r D).	-map with ntSNP	on circul rCRS cou data allo	ar mtD Id be c w users	NA, and lisplayed s to stor	d mtŠl d in a re thei	NP data new wi r data o	ndow (F calculate
С	rCRS AB241275_SAMPLE	4 CCCCCAA CCCCCAA	404 CTAACACATT	50		CCCATACT	ACTAATCTO	CATCAAT. S	490. ACAA ACAA	
	rCRS AB241275_SAMPLE I: Insertion, D	5 cccccccc cccccccc : Deletion, between so	005: CCATCCTACC CCATCCTACC V: Transv ample seq	LO CAGCACA Prsion, UENCE	s: Transi	TGCTAAC	CCCATACCO CCCATACCO Heterop W wind	olasmy OW.		
	٨		B C	D	E	E	G	Ц	T	1
D	1 Samples	SNP	lists	U	E	F	0	П	1	,
	2 AB241275_SAMPL	E_YM89 1622	16287	16319	16362	16399	73	263	309.1C	315.1C
	• • •				_					

↑ Downloaded mtSNP list in a "CSV" file format, which is compatible with MS Excel.

mtDNAprofiler: a web-based mitochondrial DNA sequence analysis tool In Seok Yang, Hwan Young Lee, Woo Ick Yang, and Kyoung-Jin Shin Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, Korea



used in



quence, ated by Fig. 1C). ed from

2. mtSNP conversion and comparison system

2-1. Protocol



Comparison mode

mtSNP conversion and comparison tool performs two functions, conversion of mtSNP data to FASTA sequence (Fig. 2A-B) and comparison between two mtSNP data gathered independently or generated by different alignment rules (Fig. 2C-E). Both also carry out mtSNP re-calculation for the converted FASTA sequence, which enables comparison between original input and recalculated mtSNP for the validity checkup of original input mtSNP data (third tab menu in Fig 2B and 2E).

Major difference between conversion and comparison modes is only in the number of input mtSNP data sets. While single data set can be used only in conversion mode (Fig 2A), two data sets with the same number of samples can be tested in comparison mode (Fig 2C).

2-2. Screen shot

2A	mtDNAprofiler mtSNP conversion and comparison tool	
	Mode: [®] Conversion [©] Comparison ► [Conversion from mtSNP profile to FASTA sequence] ← Mod	le selection
	Control region : I All (16024-576) HV1 (16024-16365) HV2 (73-340) HV3 (438-574)	Banga coloction
	O User define range :	
	ex) 1-576, 7500-8000, 14500-15000, 16500-200	
	mtSNP data :	
	PK-070 16111A 16223 16519 73 228 234 249d 263 315.1C 345 489 569 PK-042 16148 16223 16271 16399 16519 73 146 152 200 263 315.1C 489 ← mtSNP data inpu	t

↑ Start page of conversion mode of mtSNP coversion and comparison tool. After mode selection, users may provide mtSNP data and its range. Input format of mtSNP data is a sample name followed by mtSNP list separated by tab or space.

D	Select All Inverse Reset	Download FASTA seq				
D	Results	↑ Download button of converted sequence				
	✓ 1. PK-070					
	Alignment with rCRS Converted FASTA sequences Validity checkup of mtSNP ← Three	e tab menus				
	a. Range : 16024 ~ 16569 b. Included mtSNPs : 16111A 16223 16519 c. Alignment with rCRS View new window					
	rcrss 1603016040160501606016070160801609016100 rcrss ITCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCAACAACCGCCTATGTATTTCGTACATTACTG 16024-16569 ITCTTTCATGGGGAAGCAGATTTC Alignment with rCRS in range 1 CGCTATGTATTTCGTACATTACTG					
	a. Range : 1 ~ 576 b. Included mtSNPs : 73 228 234 249d 263 315.1C 345 489 569 c. Alignment with rCRS View new window					
	rcrs 110					
	d. Excluded mtSNPs : None ← Excluded mtSNPs					

↑ Result page of the conversion mode has three tab menus, alignment with rCRS, converted sequences, and validity checkup of mtSNP. In alignment tab, range, included mtSNPs, alignment result with rCRS in each ranges, and excluded mtSNPs are represented.

2C	Mode: O Conversion O Comparison F [Compari	ison between two mtSNP data sets] \leftarrow Mode select	ion
	mtSNP data set 1: PK-070 16111A 16223 16519 73 228 234 249d 263 315.1C 345 489 569 969 9K-042 16148 16223 16271 16399 16519 73 146 152 200 263 315.1C 489 9K-001 16223 16271 16343 16399 16519 73 146 152 263 315.1C 489 9K-036 16223 16271 16399 16519 73 146 152 263 315.1C 489 9K-036 16223 16271 16399 16519 73 146 152 263 315.1C 489 489 46519 73 146 152	mtSNP data set 2: PK-070 161111A 16223 16519 73 228 234 249d 263 315.1C 345 489 569 PK-042 16148 16223 16271 16399 16519 73 146 152 200 263 315.1C 489	← mtSNP data input

↑ Start page of comparison mode of mtSNP conversion and comparison tool. Two mtSNP data sets are usually independently obtained from duplicate experiments using the same samples. Input format is the same as used in the case of conversion mode. Number and order of samples in two data set should be the same each other.

Comparison mode allows users to compare converted FASTA sequences as well as mtSNPs of two data sets, and to easily confirm their difference between both data sets and validity for the original input mtSNPs in each data set (Fig 2A-E).

20	Sample 1.			Direct comparison			
20	a. Comparison between two mtSNP data sets - mtSNPs between two samples (PK-070 and PK-070) are not equal.			7 result between two original mtSNP data			
	PK_070 16111A 16223 16519 73 228 234 249d 263 314.1C 345 489 569 PK_070 16111A 16223 16519 73 228 234 249d 263 315.1C 345 489 569 9K_070 16111A 16223 16519 73 228 234 249d 263 315.1C 345 489 569						
	Magenta : consensus mtSNPs, Black : different mtSNPs						
	↑ Result of checkup wh	direct comparison be ether both data is eo	etween two mtSN qual or not.	IP data in comparison m	ode, which enables		
	b. Comparison of sequences converted from two mtSNP data sets						
2E	Alignment with rCRS Converted FASTA sequences Validity checkup of mtSNP						
	mtSNP data set 1	mtSNP data set 2 Consensus m	tSNPs ← Three sub-	ab menus			
	1. PK-070 a. Range : 16106 ~ 16569 b. Included mtSNPs : 16111A 16223 16519 c. Alignment with rCRS						
	rcrss 1611016120161301614016150161601617016180 16106-16569 GCCAGCACCATGAATATTGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAACCCAATCCACATCAAAAACCCCC V. GCCAGACACCATGAATATTGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAAACCCAATCCACATCAAAAACCCCC						
	a. Range :	1~574					
	b. Included mtSNPs	8 : 73 228 234 249d 263 314	4.1C 345 489 569				
		1					
	1-574	GATCACAGGTCTATCACCCTATTAACC	ACTCACGGGAGCTCTCCATGCAT				
	1-074		ACICACOCONSCICICCATOCAT	•			
	d. Excluded mtSNP	s : None					
	Alignment with rCRS	Converted FASTA sequences	Validity checkup of mtSNP				
	Converted sequences from two mtSNP data sets are the same sequence. Sequence comparison result between two data						
a. From mtSNP data set 1 : 16106-16569 GCCAGACACCATGAATATTGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAAACCCAAT(
	1-574 GAT	CACAGGTCTATCACCCTATTAAC	CACTCACGGGAGCTCTCC	ATGCATTTGGTATTTTCGTCTGGG			
	b. From mtSNP data set 2 :						
16106-16569 GCCAGACACCATGAATATTGTACGGTACCATAAATACTTGACCACCTGTAGTACATAAAAAACCCAAT(
	Alignment with rCRS	Converted FASTA sequences	Validity checkup of mtSNP				
	mtSNP data set 1	- At least, one mtSNP is not correct	t	Different position between original and re-calculated mtSNP			
	Original	PK-070 16111A 16223 16519 7	3 228 234 249d 263 314.	C 345 489 569			
	Re-calculated	PK-070 16111A 16223 16519 7	3 228 234 249d 263 315.	LC 345 489 569			
	mtSNP data set 2	- All mtSNPs are correct.					
	Original	PK-070 16111A 16223 16519 7	3 228 234 249d 263 315.	LC 345 489 569			
	Re-calculated	PK-070 16111A 16223 16519 7	3 228 234 249d 263 315.	LC 345 489 569			

1 Result page of three tab menus in comparison mode, which provides results of alignment with rCRS, comparison between two converted FASTA sequences, and comparison between original input and re-calculated mtSNP data. In particular, last result enables users to ensure validity of mtSNP data.

Conclusion

1. The description and analysis of mtDNA sequence variations compared to rCRS can be easily carried out using mtDNA nomenclature tool of mtDNAprofiler.

2. mtSNP conversion and comparison tool perform comparison and management of the data gathered by two independent experiments.

3. Therefore, mtDNAprofiler will provide useful tools to characterize and analyze mtDNA sequences for researcher dealing with mtDNA.

References

- 1. Bär W. *et al.* Int J Legal Med. 2000, 113(4):193-6.
- 2. Wilson MR. et al. Forensic Sci Int. 2002, 129(1):35-42.
- 3. Budowle B. et al. J Forensic Sci. 2010, in press.