mtDNAprofiler: a web-based mitochondrial DNA sequence analysis tool

YONSEI UNIVERSITY COLLEGE OF MEDICINE

In Seok Yang, Hwan Young Lee, Woo Ick Yang, and Kyoung-Jin Shin

Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, Korea

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.

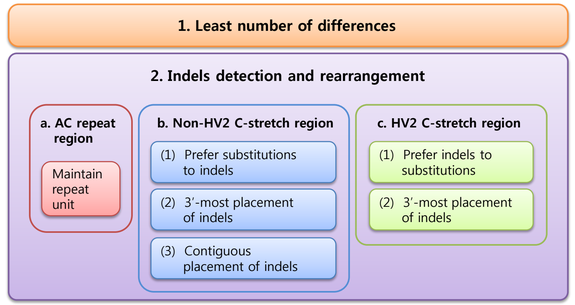
As a first attempt to standardize the description of length variations in mtDNA, Wilson *et al.* proposed an alignment and nomenclature protocol (Wilson Rules). Recently, Budowle *et al.* developed a automated program using a hierarchical set of rules, called the Mitotyper Rules to accomplish absolute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

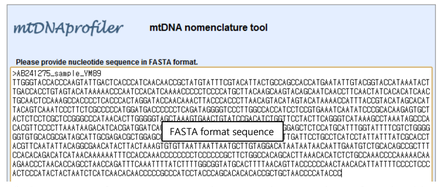
1-1. Alignment protocol



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.

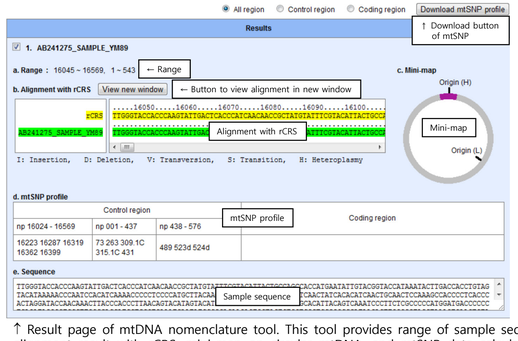
1-2. Screen shot

1A



↑ Start page of mtDNA nomenclature tool. Input data of nucleotide sequence is used in FASTA format.

1B



↑ Result page of mtDNA nomenclature tool. This tool provides range of sample sequence, alignment result with rCRS, mini-map on circular mtDNA, and mtSNP data calculated by alignment. In addition, alignment with rCRS could be displayed in a new window (Fig. 1C). And download function for the mtSNP data allow users to store their data calculated from sample sequence (Fig 1D).

1C

```
                .....440.......450.......460.......470.......480.......490.
rCRS            CCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAATACAA
                ..............................................................
AB241275_SAMPLE_CCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAATCTCATCAACACAA

                .....500.......510.......520.......530.......540.
rCRS            CCCCCGCCATCCTACCCAGCACACACC----CGCTGCTAACCCCATACCC
                ..........................D2.....................
AB241275_SAMPLE_CCCCCGCCATCCTACCCAGCACACACC----CGCTGCTAACCCCATACCC

I: Insertion, D: Deletion, V: Transversion, S: Transition, H: Heteroplasmy
```

↑ Alignment between sample sequence and rCRS in a new window.

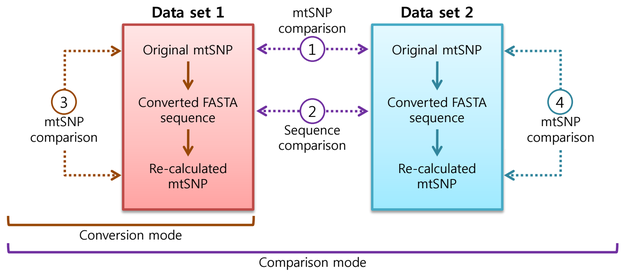
1D

	A	B	C	D	E	F	G	H	I	J
1	Samples	SNP lists								
2	AB241275_SAMPLE_YM89	16223	16287	16319	16362	16399	73	263	309.1C	315.1C

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

2. mtSNP conversion and comparison system

2-1. Protocol

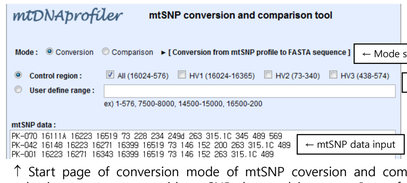


mtSNP conversion and comparison tool performs two functions, conversion of mtSNP data to FASTA sequence (Fig. 2A-B) and comparison between two mtSNP data gathered independently or generated by different alignment rules (Fig. 2C-E). Both also carry out mtSNP re-calculation for the converted FASTA sequence,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

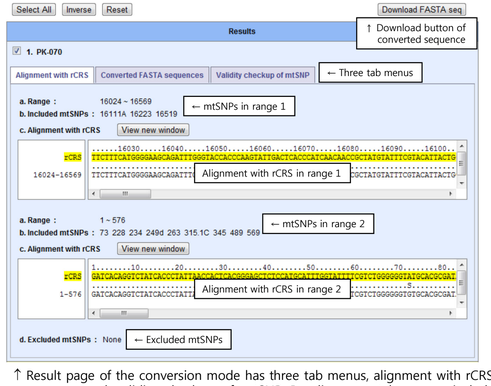
2-2. Screen shot

2A



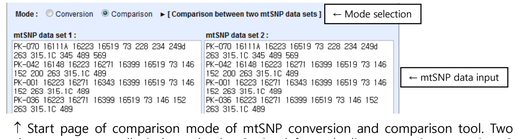
↑ Start page of conversion mode of mtSNP coversion and comparison tool. After mode selection, users may provide mtSNP data and its range. Input format of mtSNP data is a sample name followed by mtSNP list separated by tab or space.

2B



↑ 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.

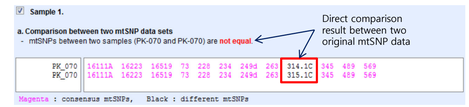
2C



↑ 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.

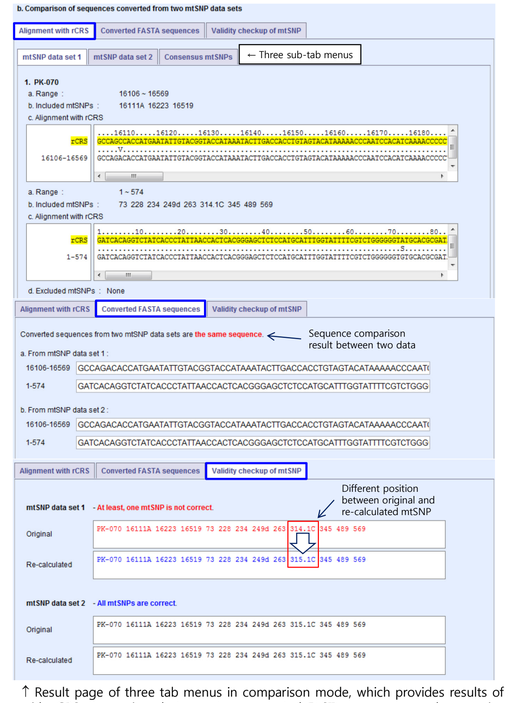
Comparison mode allows users to compare converted FASTA sequences as well as mtSNPs of two data sets, and to easily confirm their difference between both data sets and validity for the original input mtSNPs in each data set (Fig 2A-E).

2D



↑ Result of direct comparison between two mtSNP data in comparison mode, which enables checkup whether both data is equal or not.

2E



↑ 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.