Title: Investigation of length heteroplasmy in mitochondrial DNA control region by massively parallel sequencing

Authors: Chun-Yen Lin, Li-Chin Tsai, Hsing-Mei Hsieh, Chia-Hung Huang, Yu-Jen Yu, Bill Tseng, Adrian Linacre, James Chun-I Lee

PII: S1872-4973(17)30145-X
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2017.07.003
Reference: FSIGEN 1744

To appear in: *Forensic Science International: Genetics*

Received date: 20-2-2017
Revised date: 3-7-2017
Accepted date: 7-7-2017

Investigation of length heteroplasmy in mitochondrial DNA control region by massively parallel sequencing

Chun-Yen Lin[1], Li-Chin Tsai[2], Hsing-Mei Hsieh[2], Chia-Hung Huang[1], Yu-Jen Yu[3], Bill Tseng[3], Adrian Linacre[4], James Chun-I Lee[3*]

[1] Chun-Yen Lin, Institute of Forensic Medicine, Ministry of Justice, No. 123 Min'an Street, Zhonghe Dist., New Taipei City 235, Taiwan ROC.

[2] Li-Chin Tsai, Hsing-Mei Hsieh, Department of Forensic Science, Central Police University, 56 Shu-Jen Road, Kwei-San, Taoyuan 33304, Taiwan ROC.

[3] James Chun-I Lee, Yu-Jen Yu, Bill Tseng, Department of Forensic Medicine, College of Medicine, National Taiwan University, No.1 Jen-Ai Road Section 1, Taipei 10051, Taiwan ROC.

[4] Adrian Linacre, School of Biological Sciences, Flinders University, Adelaide 5001, Australia.

[*] **Corresponding author**: James Chun-I Lee, Ph.D.

Department of Forensic Medicine, College of Medicine, National Taiwan University, No.1 Jen-Ai Road Section 1, Taipei 10051, Taiwan ROC.

Tel: 886-2-23123456 ext 65493, **Fax:** 886-2-23218438

Email: jimlee@ntu.edu.tw

Highlights

- Accurate decoding of mitochondrial hypervariable regions can be problematic due to variable numbers of cytosines in poly C stretches
- Comparison of data is provided generated by Sanger sequencing and massively parallel sequencing for HVI, HVII and HVII
- Varying designations of bases occurred depending on the software used
- Applying the program SEQ Mapper to MPS data gave the most reliable results and reported the greatest number of variants

Abstract

Accurate sequencing of the control region of the mitochondrial genome is notoriously difficult due to the presence of polycytosine bases, termed C-tracts. The precise number of bases that constitute a C-tract and the bases beyond the poly cytosines may not be accurately defined when analyzing Sanger sequencing data separated by capillary electrophoresis. Massively parallel sequencing has the potential to resolve such poor definition and provides the opportunity to discover variants due to length heteroplasmy. In this study, the control region of mitochondrial genomes from 20 samples was sequenced using both standard Sanger methods with separation by capillary electrophoresis and also using massively parallel DNA sequencing technology. After comparison of the two sets of generated sequence, with the exception of the C-tracts where length heteroplasmy was observed, all sequences were concordant. Sequences of three segments 16184-16193, 303-315 and 568-573 with C-tracts in HVI, II and III can be clearly defined from the massively parallel sequencing data using the program SEQ Mapper. Multiple sequence variants were observed in the length of C-tracts longer than 7 bases. Our report illustrates the accurate designation of all the length variants leading to heteroplasmy in the control region of the mitochondrial genome that can be determined by SEQ Mapper based on data generated by massively parallel DNA sequencing.

Keywords: mitochondrial DNA; control region; length heteroplasmy; C-tract; massively parallel sequencing; SEQ Mapper; forensic science

1. Introduction

Sequencing of the human mitochondrial control region has been a niche application of forensic science since the earliest reports [1]. It is particularly used in instances where there is insufficient DNA for nuclear testing. Well characterised length polymorphisms occur at nucleotide positions (np) 16184-16193 of the hypervariable region I (HVI), at 303-315 of HVII, and 456-463, 514-523 and 568-573 of HVIII [2-5]. With the exception of a length polymorphism at np 514-523, which is due to a CA repeat polymorphism, all of the other sequence variants are polycytosine stretches (C-tracts). The C-tracts are classified based on whether the C-tracts are interrupted with a thymine or whether they are uninterrupted and occur as a string of cytosines. If a C-tract is not interrupted, or the number of cytosines is typically more than 7 bases, a phase shift pattern and heteroplasmy may be recorded in the resulting elctropherogram [6]. This phase shift results in uncertainty in the number of cytosines and also makes the subsequent DNA sequence difficult to interpret.

Length heteroplasmy within the mtDNA control region has been studied due to the relationship with a variety of diseases [7,8]. Additional other analytical strategies for this determination of the length of poly-C tracts have been reported [9-15]. Despite the need for accurate base-calling, this can still be highly problematic. Although many variants with different length can be detected by massively parallel sequencing (MPS), the proportions of every variant may still be ambiguous [16]. More recently, our latest study [17] has designed a program called SEQ Mapper to detect alleles in MPS data, provided that a reference allele sets has been established. This program provides an opportunity to evaluate MPS data for all the variants at the C-tracts. In this study we demonstrate the different sequence variants and their proportions of poly-C length observed after standard Sanger sequencing followed by separation by capillary electrophoresis (CE) compared to MPS analysis and the application of the program SEQ Mapper.

2. Materials and methods

2.1 Sample preparation and DNA extraction

Twenty buccal swabs were collected from volunteers with informed consent after the project had been approved by the Institutional Review Board of National Taiwan University Hospital. DNA was extracted using the Tissue & Cell Genomic DNA Purification Kit (GeneMark, Taipei, Taiwan) following the manufacture’s protocol and quantified using the NanoDrop 2000 spectrophotometer (Thermo Fisher, DE, USA).

2.2 Control region sequencing by CE

PCR amplification of the whole mtDNA control region was performed using the primer pair L15969 (5’-GGACAAATCAGAGAAAAAGTC-3’) and H638 (5’-ACCAAACCTATTTGTTTATGG-3’). Amplification was performed in 20 µL which contained approximately 1 ng of DNA, 0.5 µM of each primer, Phire Hot Start DNA Polymerase (Thermo Fisher) and its PCR buffer according to the manufacturer’s recommendation. PCR amplification was conducted in the Applied Biosystems 2720 Thermal Cycler (Thermo Fisher) at 98 ^{o}C for 5 minutes, then 35 cycles of 98 ^{o}C for 5 seconds, 57 ^{o}C for 5 seconds and 72 ^{o}C for 15 seconds, with a final extension at 72 ^{o}C for 2 minutes. PCR products were sequenced using the commercial BigDye^{TM} Terminator Kit (Thermo Fisher). There are 4 sequencing primers, primers L15969, H638, L16488 (5’-CTGTATCCGACATCTGGTTCC-3’) and H29 (5’-GTGGTTAATAGGGTGATAGACC-3’). The cycle sequencing products were separated and detected using an Applied Biosystems 3730 DNA Analyzer (Thermo Fisher). Sequences were interpreted by the software Sequence Analysis (Thermo Fisher).

2.3 Mitochondrial DNA Genome sequencing by MPS

Mitochondrial genome (mtGenome) amplification was performed in two separate reactions using the TaKaRa LA PCR Kit (TaKaRa Bio, Shiga, Japan). Primers were described previously by Gunnarsdo´ttir et al. [18]. The overlapping two amplicons are approximately 8.3 and 8.6 kb, respectively. Libraries were prepared using the Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) according to the manufacturer’s protocol. DNA quality was measured using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). DNA quantity was determined using 7500 Real-Time PCR System (Thermo Fisher) with SYBR Green. Unique indices were added to DNA fragments of each sample. Each library was normalized for sequencing to 4 nM in the same run on the MiSeq^{TM} (Illumina). Sequencing reactions were carried out using the MiSeq Reagent Nano v2500-Cycle Kit (Illumina) according to the manufacturer’s protocol.

2.4 MPS data analysis

MPS data were analyzed using the BaseSpace App Guide of the mtDNA Variant Processor (MVP) v1.0 (Illumina) with a threshold of Q30, analysis threshold of 5 %, an interpretation threshold of 5 % and a minimum read count of 5 to search for variants in the mtGenome. Sequences were compared to the revised Cambridge Reference Sequence (rCRS) [19]. For C-tract regions, SEQ Mapper [17] was used to search sequence variants and coverages. Reference alleles used in SEQ Mapper, listed in Table 1, had an extra conserved 1 to 5 bases at both ends. The sequence variants of C-tracts were edited based on the laboratory database of mtDNA control region in the Taiwan Han population (350 individuals). Variants in MPS data were recorded with coverage threshold of 5 % and a minimum read count of 5 detected by SEQ Mapper.

3. Results

The complete mitochondrial genomes of the 20 samples were successfully sequenced by MPS and compared against the rCRS. Haplotypes based on the whole genome data are listed in Table S1 (Supplementary Material). Sequences of the entire control region were also obtained using Sanger sequencing separated by CE. Comparison of DNA sequence data from both Sanger sequencing and MPS revealed that with the exception of the C-tracts, where length heteroplasmy was observed, all 20 sequences were concordant; this included all the examples of point heteroplasmy. An example is sample 1 where a point heteroplasmy at nucleotide position 146 is observed in the electropherogram, shown in Fig. 1, and also detected by MPS, listed in Table S1.

It was found that for all 20 samples there was an interruption by a thymine in C-tracts at np 460 within the region np 456-463 of HVII and no length heteroplasmy was observed. This region was not therefore the focus of further study. Three examples of length heteroplasmy were found in C-tracts in HVI at np 16184-16193, HVII at np 303-315, and HVIII at np 568-573. These three C-tracts where length heteroplasmy was observed were recorded in electropherograms from the CE data although the precise length of C-tracts could not be determined. Reads of MPS interpreted by MPS-MVP were aligned and compared against the relevant position of the rCRS. As illustrations of length heteroplasmy, examples of the variant call format (VCF) data generated by MPS-MVP are provided in Table S1 (Supplementary Material). It remained, however, the case that the length of C-tracts were problematic to determine easily. Sample 1 exhibits some prime examples at nucleotide position 309, 315, 573 and 16181-16183, listed in Table S2.

Analyses of the same MPS data by the program SEQ Mapper where examples of length heteroplasmy was observed are shown in Table 2 and where no such heteroplasmy occurred in Table 3. Results of sequence variants of 20 samples in polyCI (np 16184-16193), II (np 303-315) and III (np 568-573) interpreted using CE and SEQ Mapper are shown in Table 4. The coverage and percentage of sequence variants of 20 samples in polyCI to III analyzed by SEQ Mapper are shown in Table 5 to 7 respectively.

4. Discussion

The VCF data (Table S2) obtained from MVP is capable of identifying nucleotide variants such as point heteroplasmy. Length heteroplasmy can affect the ability to accurately determine sequence variants due to the presence of multiple of sequence and length variants. Any such accurate determination of the variants can be problematic if based only on the VCF data. The application of the program SEQ mapper was found to resolve the sequence variants of C-tracts in reads of MPS in this study.

4.1 PolyCI

Sequences of C-tracts based on CE data were clearly and accurately recorded when interrupted by thymine in polyCI for 12 samples (No. 2, 5-7, 10-13, 15, 18-20). Most of these samples detected contained a major and many trace variants as detected by SEQ Mapper (Table 3 and 5). Since the coverages of all trace variants were below the interpretation threshold, and as such they were not recorded, with only major variants being recorded in the data of SEQ Mapper.

In the remaining 8 samples, which were not interrupted by thymine in the C-tract, a frame shift pattern was observed after the C-tracts. An example is sample 1 (Fig. 2) where there are 10 clear homo-signals for a C from np 16183 followed by 3 hetero-signals of a C and A and/or T at the 11-13th cytosine of this region. Three length variants (recorded as $C_{10,}$ C_{11} and C_{12} in Table 4) resulted in a frame shift in the sequences after the C-tract. Based on the VCF data of MPS-MVP, no point and length heteroplasmies were recorded (np 16183-16193), as shown in Table S2. There are, however, 12 sequence variants from C_6 to C_{17} detected by MPS-SEQ and are shown in Table 2. The frequency of occurrence of all the variants created a bell shape from C_6:1 increasing to C_{11}:525 and decreasing to C_{17}:2. This resulted in many sequence variants not being observed by CE and MPS-MVP as they were below this threshold, but were detected by SEQ Mapper. Only sequence variants C_{10-13} with both coverage and frequency above interpretation thresholds were recorded in this study (Table 5). SEQ Mapper provided details on the length and sequence variants within C-tracts and their coverage.

3.2 PolyCII

Although there are C-tracts interrupted by thymine in this study, many samples still showed length heteroplasmy before the thymine base insert (np 310) based on CE, MPS-MVP and SEQ Mapper data. For example, in sample 1 there is an interrupting thymine at np 310. There are 4 hetero-signals of C and T in this region (Fig. 3). These data suggest that at least 4 sequence variants (C_8, C_9, C_{10} and C_{11}) existed before np 310 in this sample with C_9 as the major variant. The variants in CE are $C_{8-11}TC_6$ (Table 4). The VCF data of MPS-MVP recorded three instances of point heteroplasmy, 309.1c, 309.2c and 309.3c, as shown in Table S2. SEQ Mapper recorded the sequence variants as $C_{8-10}TC_6$ as there are variants C_8TC_6, C_9TC_6 and $C_{10}TC_6$ all with coverage and frequency above interpretation thresholds respectively (Table 4 and 6).

In sample 3, close inspection shows that there are still trace cytosine and thymine signals at np 310 and 311 in Fig. 4, although no length heteroplasmy was recorded by CE. The VCF data of MPS-MVP also recorded no instances of length heteroplasmy (Table S2) although SEQ Mapper recorded the sequence variants as $C_{7-8}TC_6$ (Table 4 and 6). While trace indication of length heteroplasmy can be evident by both CE and MPS-MVP, accurate recording of the number of length and sequence variants can still be problematic except when applying SEQ Mapper.

3.3 PolyCIII

At the polyCIII region, 14 of the 20 samples exhibited a 6 base C-tract and no length heteroplasmy as detected by CE, MPS-MVP and SEQ Mapper. The 6 exceptions were samples 1, 4, 11, 13, 14 and 16 where the C-tract was longer than 6 nucleotides. For example in sample 1, there are 9 homo-signals of C and 4 hetero-signals of A and C; these are from the 10th to 13th cytosines as viewed from the reverse direction np 573 (Fig. 5). In this sample a number of sequence variants with longer than 9 homo-polymer cytosines were observed and recorded as C_{9-13} by CE, shown in Table 4. VCF data records a C-tract from np 568 to 573 with 4 inserted nucleotides (573.1c, 573.2c, 573.3c and 573.4c) (Table S2) underscoring the complexity of accurately recording the length and number of sequence variants. These data contrast with SEQ Mapper where 5 sequence variants were detected. Since the total coverage of Three of these (C_{9-11}) had interpretation thresholds of coverage and frequency both above 5 and 5% respectively (see Table 4 and 7).

Accurate recording of the 6 bases of C in this region was performed using CE, but this was not possible for longer than 7 Cs. Both SEQ Mapper and MPS-MVPs could detect instances of length heteroplasmy but only SEQ Mapper provided the length, number and coverage of sequence variants.

It may be that some variants present at trace levels as detected by CE may not be presented in the MPS data and therefore will not be interpreted by MVP and SEQ Mapper, shown in Table 4. The average coverage for samples that exhibited length heteroplasmy was 139, compared to 1,524 for samples that did not exhibit any length heteroplasmy. The relatively low coverage of variants in polyCIII is most likely the reason for this discordance.

5. Conclusions

From the 20 samples analyzed, examples of length heteroplasmy caused by polycytosine stretches were observed in the mtDNA control region found in HVI, HVII and HVIII. The presence of these variants can be visually identified in sequence electropherograms generated by CE. It is, however, problematic to accurately determine the length, number and distribution of sequence variants. Although sequence data generated by MPS does not produce such an electropherogram, it has the potential to answer these questions. The MVP and SEQ Mapper software were used to detect and record the total data in this study. Evidence of point heteroplasmy can be easily obtained using MVP even if length heteroplasmy is less straightforward to interpret. SEQ Mapper applied to MPS data was able to record accurately the length, number and coverage of sequence variants from these mtDNA samples. The length and number of sequence variants can be evaluated by the number of homo-signal at C-tracts and extended hetero-signals from the CE electropherogram after comparison SEQ Mapper results. SEQ Mapper provided valuable information in the determination of the length, number and coverage of sequence variants of C-tracts in the control region of mtDNA.

Acknowledgements

We thank the Ministry of Science and Technology for funding (104-1301-IFM(8)-08) and the Department of Medical Research at the National Taiwan University Hospital (Taipei, Taiwan) for the assistance of capillary electrophoresis.

References

1. J.M. Butler, B.C. Levin. Forensic applications of mitochondrial DNA. Trends Biotechnol. 16(4)(1998)158-162.
2. J.A. Irwin, J.L Saunier, H. Niederstätter, K.M. Strouss, K.A. Sturk, T.M. Diegoli, A. Brandstätter, W. Parson, T.J. Parsons, Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples, J. Mol. Evol. 68(2009)516-27.
3. W. Parson, T.J. Parsons, R. Scheithauer, M.M. Holland, Population data for 101 Austrian Caucasian mitochondrial DNA d-loop sequences: application of mtDNA sequence analysis to a forensic case, Int. J. Legal Med. 111(1998)124-32.
4. S. Lutz, H.J. Weisser, J. Heizmann, S. Pollak, Location and frequency of polymorphic positions in the mtDNA control region of individuals from Germany, Int. J. Legal Med. 111(1998)67-77.
5. A. Bodenteich, L.G. Mitchell, M.H. Polymeropoulos, C.R. Merril, Dinucleotide repeat in the human mitochondrial D-loop, Hum. Mol. Genet. 1(1992)140.
6. N. Howell, C.B. Smejkal, Persistent heteroplasmy of a mutation in the human mtDNA control region: hyper mutation as an apparent consequence of simple-repeat expansion/contraction, Am. J. Hum. Genet. 66(2000)1589-1598.
7. J. Poulton, M.S. Brown, A. Cooper, D.R. Marchington, D.I. Phillips, A common mitochondrial DNA variant is associated with insulin resistance in adult life, Diabetologia. 41(1998)54-58.

8. P.F. Chinnery, H.R. Elliott, S. Patel, C. Lambert, S.M. Keers, S.E. Durham, M.I. McCarthy, G.A. Hitman, A.T. Hattersley, M. Walker, Role of the mitochondrial DNA 16184–16193 poly-C tract in type 2 diabetes, Lancet. 366(2005)1650-1651.
9. M.G. Shin, B.C. Levin, H.J. Kim, H.R. Kim, I.K. Lee, D. Cho, S.J. Kee, J.H. Shin, S.P. Suh, D.W. Ryang, Profiling of length heteroplasmies in the human mitochondrial DNA control regions from blood cells in the Korean population. Electrophoresis. 27(2006)1331-1340.
10. C. Berger, P. Hatzer-Grubwieser, C. Hohoff, W. Parson, Evaluating sequence-derived mtDNA length heteroplasmy by amplicon size analysis, Forensic Sci. Int. Genet. 5(2011)142-145.
11. S. Lutz-Bonengel, T. Sänger, S. Pollak, R. Szibor, Different methods to determine length heteroplasmy within the mitochondrial control region, Int. J. Legal Med. 118(2004)274-281.
12. L. Forster, P. Forster, S.M. Gurney, M. Spencer, C. Huang, A. Röhl, B. Brinkmann, Evaluating length heteroplasmy in the human mitochondrial DNA control region, Int. J. Legal Med. 124(2010)133-142.
13. S.B. Seo, B.S. Jang, A. Zhang, J.A. Yi, H.Y. Kim, S.H. Yoo, Y.S. Lee, S.D. Lee, Alterations of length heteroplasmy in mitochondrial DNA under various amplification conditions, J. Forensic Sci. 55(2010)719-722.
14. J.C. Lee, L.C. Tsai, Y.J. Yu, C.Y. Lin, A. Linacre, H.M. Hsieh, Investigation into length heteroplasmy in the mitochondrial DNA control region after treatment with bisulfite, J. Formosa Med. Asso. 115(2016)284-287.
15. W. Parson, L. Gusmão, D.R. Hares, J.A. Irwin, W.R. Mayr, N. Morling, E. Pokorak, M. Prinz, A. Salas, P.M. Schneider, T.J. Parsons, DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. Forensic Sci Int Genet. 13(2014)134-142.
16. C. Davis, D. Peters, D. Warshauer, J. King, B. Budowle, Sequencing the hypervariable regions of human mitochondrial DNA using massively parallel sequencing: Enhanced data acquisition for DNA samples encountered in forensic testing, Leg. Med. 17(2015)123-7.
17. J.C. Lee, B. Tseng, L.K Chang, A. Linacre, SEQ Mapper: A DNA sequence searching tool for massively parallel sequencing data, Forensic Sci. Int. Genet. 26(2017)66-69.

18. E.D. Gunnarsdóttir, M. Li, M. Bauchet, K. Finstermeier, M. Stoneking, High-throughput sequencing of complete human mtDNA genomes from the Philippines, Genome Res. 21(2011)1–11.

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

Figure legends

Fig. 1. Sequence electropherogram of polyCI in the control region of sample 1. Arrows indicate point heteroplasmy.

Fig. 2. Sequence electropherogram of polyCI in the control region of sample 1. Arrows indicate hetero-signals.

Fig. 3. Sequence electropherogram of polyCII in the control region of sample 1. Arrows indicate trace hetero-signals of C and T.

Fig. 4. Sequence electropherogram of polyCII in the control region of sample 3. Arrows indicate hetero-signals of C and T.

Fig. 5. Sequence electropherogram of polyCIII in the control region of sample 1. Arrows indicate hetero-signals.

Fig.1

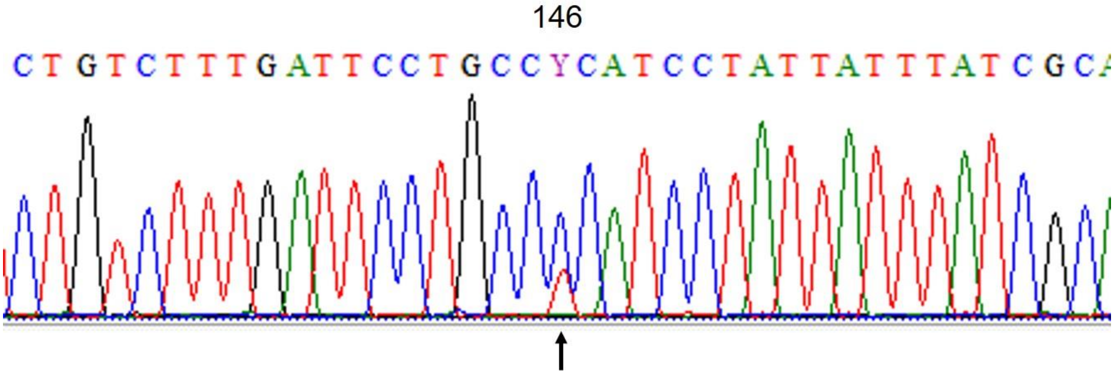

Fig.2

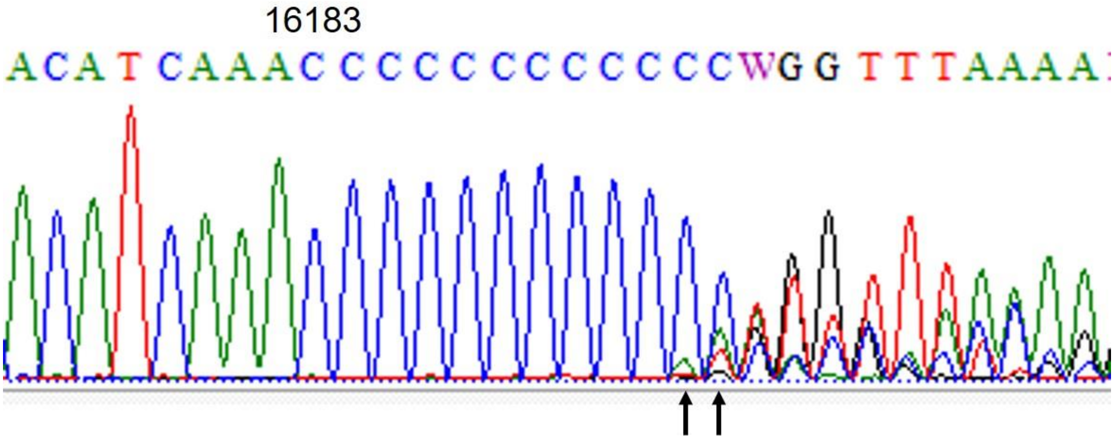

Fig.3

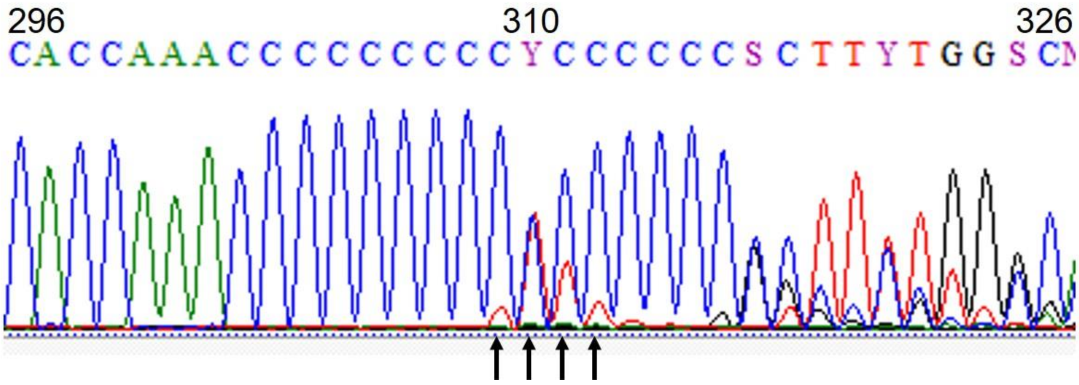

Fig.4

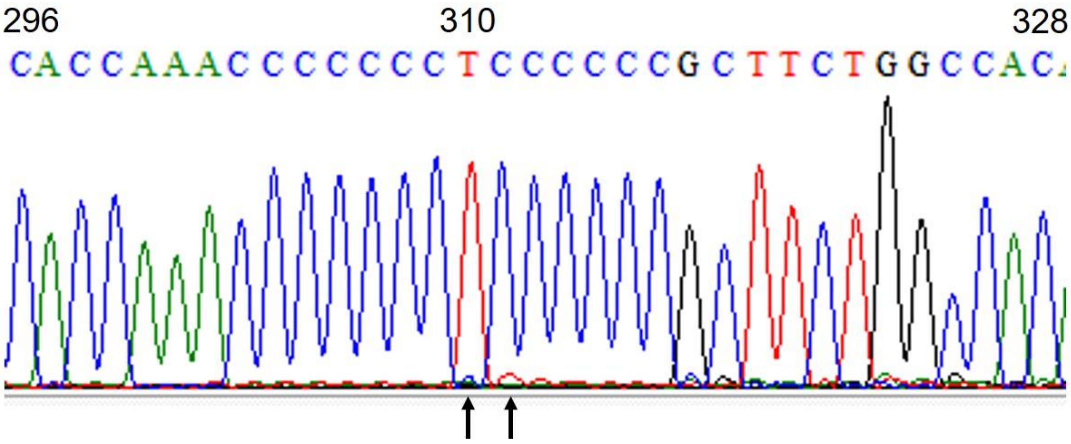
296
310
328
CACCAAACCCCCCCTCCCCCCGCTTCTGGCCACA

Fig.5

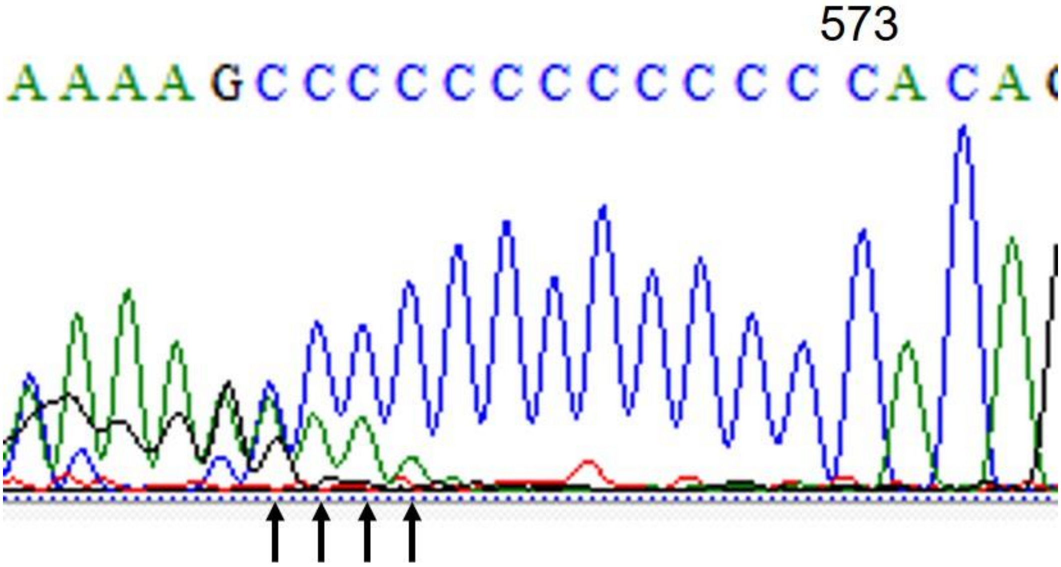
573
AAAAGCCCCCCCCCCCCACAC

Table 1 Sequence types of polyCI, II and III in the mtDNA control region defined in this study.

Types	Sequences	Types	Sequences
polyCI_C$_2$TC$_2$TC$_4$	CCTCCTCCCC	polyCII_C$_4$TC$_6$	CCCCTCCCCCC
polyCI_C$_3$TC$_6$	CCCTCCCCCC	polyCII_C$_7$TC$_5$*	CCCCCCCTCCCCC
polyCI_C$_3$TCTC$_4$	CCCTCTCCCC	polyCII_C$_7$TC$_6$	CCCCCCCTCCCCCC
polyCI_C$_4$T$_2$C$_4$	CCCCTTCCCC	polyCII_C$_8$TC$_5$	CCCCCCCCTCCCCC
polyCI_C$_5$T$_2$C$_3$	CCCCCTTCCC	polyCII_C$_8$TC$_6$	CCCCCCCCTCCCCCC
polyCI_C$_5$TC$_2$TC	CCCCCTCCTC	polyCII_C$_8$TC$_6$A	CCCCCCCCTCCCCCCA
polyCI_C$_5$TC$_3$T	CCCCCTCCCT	polyCII_C$_9$TC$_5$	CCCCCCCCCTCCCCC
polyCI_C$_5$TC$_4$*	CCCCCTCCCC	polyCII_C$_9$TC$_6$	CCCCCCCCCTCCCCCC
polyCI_C$_5$TCTC$_2$	CCCCCTCTCC	polyCII_C$_{10}$TC$_5$	CCCCCCCCCCTCCCCC
polyCI_CTC$_3$TC$_4$	CTCCCTCCCC	polyCII_C$_{10}$TC$_6$	CCCCCCCCCCTCCCCCC
polyCI_TC$_4$TC$_4$	TCCCCTCCCC	polyCII_C$_8$-C$_{15}$	C$_8$-C$_{15}$†
polyCI_C$_6$-C$_{17}$	C$_6$-C$_{17}$#	polyCIII_C$_6$*-C$_{12}$	C$_6$-C$_{12}$§

*Sequence of rCRS.

#6 cytosines were the shortest number and Cs and 17 the longest number observed.

†8 cytosines were the shortest number of Cs and 15 the longest.

§6 cytosines were the shortest number of Cs and 12 the longest.

Table 2 Searching results generated by SEQ Mapper from the MPS sequence data for sample 1.

Sequence type	Is Reverse*	Count
polyCI_C_6	TRUE	1
polyCI_C_7	FALSE	4
polyCI_C_8	TRUE	3
polyCI_C_8	FALSE	7
polyCI_C_9	TRUE	35
polyCI_C_9	FALSE	23
polyCI_C_{10}	TRUE	153
polyCI_C_{10}	FALSE	119
polyCI_C_{11}	TRUE	349
polyCI_C_{11}	FALSE	176
polyCI_C_{12}	TRUE	344
polyCI_C_{12}	FALSE	102
polyCI_C_{13}	TRUE	165
polyCI_C_{13}	FALSE	6
polyCI_C_{14}	TRUE	65
polyCI_C_{15}	TRUE	19
polyCI_C_{16}	TRUE	4
polyCI_C_{17}	TRUE	2
polyCII_C_8TC_6	TRUE	29
polyCII_C_8TC_6A	TRUE	1
polyCII_C_9TC_6	TRUE	55
polyCII_$C_{10}TC_6$	TRUE	32
polyCII_$C_{11}TC_6$	TRUE	3
polyCII_C_{13}	FALSE	2
polyCII_C_{14}	FALSE	5
polyCII_C_{15}	FALSE	3
polyCII_C_{15}	TRUE	3
polyCIII_C_8	FALSE	1
polyCIII_C_9	TRUE	2
polyCIII_C_9	FALSE	6
polyCIII_C_{10}	TRUE	1
polyCIII_C_{10}	FALSE	12
polyCIII_C_{11}	FALSE	5
polyCIII_C_{12}	FALSE	1

*SEQ Mapper searches both the forward and reverse sequences.

Table 3 Searching results generated by SEQ Mapper from the MPS sequence data for sample 2.

Sequence type	Is Reverse*	Count
polyCI_C_5TC_3T	TRUE	4200
polyCI_C_5TC_3T	FALSE	4516
polyCI_C_5TC_4	TRUE	9
polyCI_C_5TC_4	FALSE	8
polyCII_C_7TC_5	TRUE	1
polyCII_C_7TC_6	TRUE	60
polyCII_C_8TC_5	FALSE	1
polyCII_C_8TC_5	TRUE	5
polyCII_C_8TC_6	TRUE	1212
polyCII_C_8TC_6A	FALSE	1
polyCII_C_8TC_6A	TRUE	1
polyCII_C_9TC_6	TRUE	49
polyCII_$C_{10}TC_6$	TRUE	5
polyCII_C_9	TRUE	1
polyCII_C_{10}	TRUE	1
polyCII_C_{13}	TRUE	3
polyCII_C_{15}	TRUE	8
polyCIII_C_6	TRUE	368
polyCIII_C_6	FALSE	1270
polyCIII_C_7	TRUE	1

*SEQ Mapper searches both the forward and reverse sequences

Table 4 Sequence variants of 20 samples based on the polyCI, II and III regions of the mtDNA control region obtained using Sanger sequencing separated by CE and MPS-SEQ Mapper.

Samples	PolyCI		PolyCII		PolyCIII	
	CE	MPS SEQ*	CE	MPS SEQ*	CE	MPS SEQ*
1	C_{10-12}	C_{10-13}	$C_{8-11}TC_6$	$C_{8-10}TC_6$	C_{9-13}	C_{9-11}
2	C_5TC_3T	C_5TC_3T	C_8TC_6	C_8TC_6	C_6	C_6
3	C_{11-13}	C_{9-13}	C_7TC_6	$C_{7-8}TC_6$	C_6	C_6
4	C_{12-14}	C_{10-14}	C_4TC_6	C_4TC_6	C_{8-12}	C_{9-11}
5	C_5TC_4	C_5TC_4	$C_{7-9}TC_6$	$C_{7-9}TC_6$	C_6	C_6
6	C_5TC_4	C_5TC_4	$C_{8-9}TC_6$	$C_{8-9}TC_6$	C_6	C_6
7	C_5TC_4	C_5TC_4	$C_{7-8}TC_6$	$C_{7-8}TC_6$	C_6	C_6
8	C_{9-11}	C_{9-11}	$C_{8-9}TC_6$	$C_{8-9}TC_6$	C_6	C_6
9	C_{10-12}	C_{9-12}	$C_{7-9}TC_6$	$C_{7-9}TC_6$	C_6	C_6
10	C_5TC_4	C_5TC_4	$C_{8-9}TC_6$	$C_{8-9}TC_6$	C_6	C_6
11	CTC_3TC_4	CTC_3TC_4	C_7TC_6	C_7TC_6	C_{8-11}	C_{8-9}
12	C_5TC_4	C_5TC_4	$C_{9-10}TC_6$	$C_{8-10}TC_6$	C_6	C_6
13	C_5TC_4	C_5TC_4	$C_{7-9}TC_6$	$C_{7-9}TC_6$	C_{9-12}	C_{8-10}
14	C_{10-12}	C_{9-12}	$C_{8-9}TC_6$	$C_{8-9}TC_6$	C_{9-13}	C_{9-11}
15	C_5TC_4	C_5TC_4	$C_{8-9}TC_6$	$C_{7-9}TC_6$	C_6	C_6
16	C_{10-12}	C_{9-12}	$C_{8-9}TC_6$	$C_{8-9}TC_6$	C_{9-12}	C_{8-11}
17	C_{11-13}	C_{10-13}	$C_{8-10}TC_6$	$C_{8-9}TC_6$	C_6	C_6
18	TC_4TC_4	TC_4TC_4	$C_{7-8}TC_6$	$C_{7-8}TC_6$	C_6	C_6
19	C_5TC_4	C_5TC_4	C_7TC_6	C_7TC_6	C_6	C_6
20	C_5TC_4	C_5TC_4	$C_{8-9}TC_6$	$C_{8-9}TC_6$	C_6	C_6

*Only sequence variants with coverage and frequency both above 5 and 5% respectively were recorded.

Table 5 Coverage counts and percentage of sequence variants of 20 samples based on polyCI in the mtDNA control region obtained using SEQ Mapper.

Samples Variants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
polyCI_$\mathrm{C_2TC_2TC_4}$						3 (0.04)	5 (0.10)	1 (0.11)		3 (0.05)		9 (0.21)	1 (0.03)		13 (0.28)				18 (0.10)	8 (0.18)
polyCI_$\mathrm{C_3TC_6}$				1 (0.08)				1 (0.11)	1 (0.07)							2 (0.04)				
polyCI_$\mathrm{C_3TCTC_4}$						3 (0.04)	4 (0.08)			4 (0.07)			3 (0.1)		2 (0.04)				10 (0.05)	1 (0.02)
polyCI_$\mathrm{C_4T_2C_4}$						3 (0.04)	8 (0.16)			5 (0.09)		5 (0.12)	2 (0.07)		4 (0.09)			1 (0.00)	10 (0.05)	
polyCI_$\mathrm{C_5T_2C_3}$						2 (0.03)	1 (0.02)			2 (0.04)		2 (0.05)							5 (0.03)	
polyCI_$\mathrm{C_5TC_2TC}$						5 (0.07)	5 (0.10)			3 (0.05)		6 (0.14)			9 (0.20)				7 (0.04)	1 (0.02)
polyCI_$\mathrm{C_5TC_3T}$		8716 (99.81)			3 (0.06)	4 (0.06)	1 (0.02)			4 (0.07)		8 (0.19)			2 (0.04)				11 (0.06)	3 (0.07)
polyCI_$\mathrm{C_5TC_4}$		17 (0.19)		1 (0.08)	5240 (99.56)	6733 (99.20)	4970 (99.38)	4 (0.44)	5 (0.37)	5538 (99.14)	47 (0.51)	4171 (98.91)	3025 (99.25)		4569 (99.07)	1 (0.02)		44 (0.18)	18755 (99.20)	4304 (99.35)
polyCI_$\mathrm{C_5TCTC_2}$						3 (0.04)	1 (0.02)			1 (0.02)		2 (0.05)	6 (0.2)						9 (0.05)	1 (0.02)
polyCI_$\mathrm{CTC_3TC_4}$					8 (0.15)	6 (0.09)				6 (0.11)	9217 (99.48)	6 (0.14)	1 (0.03)		3 (0.07)				13 (0.07)	2 (0.05)
polyCI_$\mathrm{TC_4TC_4}$					4 (0.08)	6 (0.09)				4 (0.07)		2 (0.05)	1 (0.03)		1 (0.02)	2 (0.04)		24424 (99.81)	7 (0.04)	1 (0.02)
polyCI_$\mathrm{C_6}$	1 (0.06)		6 (0.27)	3 (0.24)				3 (0.33)	6 (0.45)					7 (0.42)		7 (0.16)	11 (0.30)			1 (0.02)
polyCI_$\mathrm{C_7}$	4 (0.25)		10 (0.45)	1 (0.08)	1 (0.02)	1 (0.01)		6 (0.66)	3 (0.22)	1 (0.02)				8 (0.49)		2 (0.04)	4 (0.11)			
polyCI_$\mathrm{C_8}$	10 (0.63)		19 (0.86)	3 (0.24)		1 (0.01)		22 (2.44)	6 (0.45)					17 (1.03)		66 (1.46)	15 (0.41)		1 (0.01)	
polyCI_$\mathrm{C_9}$	58 (3.68)		129 (5.82)	19 (1.50)	1 (0.02)			252 (27.91)	100 (7.45)	3 (0.05)		2 (0.05)	1 (0.03)	135 (8.19)	2 (0.04)	593 (13.14)	175 (4.82)	1 (0.00)	11 (0.06)	1 (0.02)
polyCI_$\mathrm{C_{10}}$	272 (17.25)		472 (21.28)	113 (8.90)	5 (0.01)	14 (0.21)	6 (0.12)	461 (51.05)	391 (29.11)	11 (0.20)	1 (0.01)	4 (0.09)	5 (0.16)	473 (28.68)	7 (0.15)	1579 (34.98)	788 (21.73)	1 (0.00)	48 (0.25)	9 (0.21)
polyCI_$\mathrm{C_{11}}$	525 (33.29)		781 (35.21)	257 (20.24)	1 (0.02)	3 (0.04)		111 (12.29)	508 (37.83)	1 (0.02)			3 (0.1)	599 (36.33)		1319 (29.22)	1212 (33.42)		2 (0.01)	
polyCI_$\mathrm{C_{12}}$	446 (28.28)		545 (24.57)	322 (25.35?)				36 (3.99)	246 (18.32)					319 (19.35)		743 (16.46)	878 (24.21)			
polyCI_$\mathrm{C_{13}}$	171 (10.84)		161 (7.26)	306 (24.09)				2 (0.22)	64 (4.77)					67 (4.06)		158 (3.50)	335 (9.24)			

polyCI_C$_{14}$	65 (4.12)		79 (3.56)	198 (15.59)				3 (0.33)	12 (0.89)					21 (1.27)		39 (0.86)	179 (4.94)			
polyCI_C$_{15}$	19 (1.2)		15 (0.68)	38 (2.99)				1 (0.11)	1 (0.07)					3 (0.18)		3 (0.07)	24 (0.66)			
polyCI_C$_{16}$	4 (0.25)		1 (0.05)	6 (0.47)													4 (0.11)			
polyCI_C$_{17}$	2 (0.13)			2 (0.16)													2 (0.06)			
Sum	1577	8733	2218	127	5263	6787	5001	903	1343	5586	9265	4217	3048	1649	4612	4514	3627	24471	18907	4332

* Within bracket indicates coverage percentage; gray block indicates coverage and frequency both above 5 and 5% respectively.

Table 6 Coverage counts and percentage of sequence variants based on data from the polyCII in the mtDNA control region from 20 samples obtained using SEQ Mapper.

Samples Variants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
polyCII_C$_{4}$TC$_{6}$				2857 (99.06)			1 (0.09)													
polyCII_C$_{7}$TC$_{5}$		1 (0.07)	4 (0.61)					1 (0.13)			3 (0.39)				1 (0.18)			5 (0.17)	13 (0.74)	
polyCII_C$_{7}$TC$_{6}$		60 (4.45)	581 (89.11)		62 (5.52)	32 (3.14)	146 (13.57)	32 (4.18)	30 (5.25)	22 (3.75)	736 (94.97)		48 (11.85)	20 (3.10)	39 (7.13)	75 (4.03)	9 (1.38)	480 (16.46)	1653 (94.13)	26 (3.82)
polyCII_C$_{8}$TC$_{5}$		6 (0.45)	1 (0.15)		2 (0.18)	2 (0.20)	4 (0.37)		1 (0.18)	1 (0.17)				1 (0.16)		11 (0.59)		6 (0.21)		
polyCII_C$_{8}$TC$_{6}$	29 (21.8)	1212 (89.91)	45 (6.90)		936 (83.35)	909 (89.21)	882 (81.97)	648 (84.60)	494 (86.51)	527 (89.93)	32 (4.13)	42 (26.75)	324 (80.00)	532 (82.48)	473 (86.47)	1439 (77.28)	101 (15.54)	2364 (81.07)	76 (4.33)	606 (89.12)
polyCII_C$_{8}$TC$_{6}$ A		2 (0.15)				1 (0.10)			1 (0.18)	1 (0.17)			1 (0.25)	1 (0.16)				2 (0.07)		
polyCII_C$_{9}$TC$_{5}$	1 (0.75)											1 (0.64)				1 (0.05)	2 (0.31)			2 (0.29)
polyCII_C$_{9}$TC$_{6}$	55 (41.35)	49 (3.64)	5 (0.77)		111 (9.88)	62 (6.08)	35 (3.25)	68 (8.88)	31 (5.43)	30 (5.12)		92 (58.6)	28 (6.91)	74 (11.47)	31 (5.67)	262 (14.07)	295 (45.38)	45 (1.54)	5 (0.28)	40 (5.88)
polyCII_C$_{10}$TC$_{5}$																	1 (0.15)			
polyCII_C$_{10}$TC$_{6}$	32 (24.06)	5 (0.37)	1 (0.15)		7 (0.62)	8 (0.79)		8 (1.04)		3 (0.51)		21 (13.38)	1 (0.25)	6 (0.93)	2 (0.37)	46 (2.47)	210 (32.31)	1 (0.03)		3 (0.44)
polyCII_C$_{11}$TC$_{5}$																				
polyCII_C$_{11}$TC$_{6}$	3 (2.26)											1 (0.64)				5 (0.27)	17 (2.62)			
polyCII_C$_{8}$					1 (0.09)															
polyCII_C$_{9}$		1 (0.07)	1 (0.15)																	
polyCII_C$_{10}$		1 (0.07)		1 (0.03)												1 (0.05)	2 (0.31)		1 (0.06)	
polyCII_C$_{11}$				13 (0.45)					2 (0.35)							4 (0.21)	1 (0.15)	1 (0.03)		
polyCII_C$_{12}$			2 (0.31)	2 (0.07)		1 (0.10)	1 (0.09)	1 (0.13)						3 (0.47)		4 (0.21)	1 (0.15)	2 (0.07)	2 (0.11)	1 (0.15)
polyCII_C$_{13}$	2 (1.5)	3 (0.22)	1 (0.15)	2 (0.07)		1 (0.10)	1 (0.09)	5 (0.65)	2 (0.35)						1 (0.18)	7 (0.38)	3 (0.46)	2 (0.07)	2 (0.11)	
polyCII_C$_{14}$	5		6	2	1	1	3		7		3		2	5		4	5	1	3	

	(3.76)		(0.92)	(0.07)	(0.09)	(0.10)	(0.28)		(1.23)		(0.39)		(0.49)	(0.78)		(0.21)	(0.77)	(0.03)	(0.17)	
polyCII_C$_{15}$	6	8	5	7	3	2	3	3	3	2	1		1	3		3	3	7	1	2
	(4.51)	(0.59)	(0.77)	(0.24)	(0.27)	(0.20)	(0.28)	(0.39)	(0.53)	(0.34)	(0.13)		(0.25)	(0.47)		(0.16)	(0.46)	(0.24)	(0.06)	(0.29)
Sum	133	1348	652	2884	1123	1019	1076	766	571	586	775	157	405	645	547	1862	650	2916	1756	680)

* Within bracket indicates coverage percentage; gray block indicates coverage and frequency both above 5 and 5% respectively.

Table 7 Coverage counts and percentage of sequence variants based on data from the polyCIII in the mtDNA control region from 20 samples obtained using SEQ Mapper.

Samples Variants	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
polyCIII_C$_{6}$		1638 (99.94)	815 (100.00)		1574 (100.00)	1306 (99.92)	1863 (99.95)	1369 (100.00)	1042 (100.00)	895 (99.89)		684 (99.42)			912 (99.78)		2282 (99.83)	3341 (100.00)	2501 (99.88)	1099 (99.82)
polyCIII_C$_{7}$		1 (0.06)				1 (0.08)	1 (0.05)			1 (0.11)	14 (3.88)	4 (0.58)	1 (1.45)		2 (0.22)	1 (0.34)	3 (0.13)		3 (0.12)	2 (0.18)
polyCIII_C$_{8}$	1 (3.57)			2 (5.56)							254 (70.36)		10 (14.49)	1 (1.96)		26 (8.93)	1 (0.04)			
polyCIII_C$_{9}$	8 (28.57)			5 (13.89)							76 (21.05)		30 (43.48)	21 (41.18)		135 (46.39)				
polyCIII_C$_{10}$	13 (46.43)			19 (52.78)							13 (3.60)		25 (36.23)	19 (37.25)		99 (34.02)				
polyCIII_C$_{11}$	5 (17.86)			8 (22.22)							4 (1.11)		2 (2.90)	9 (17.65)		27 (9.28)				
polyCIII_C$_{12}$	1 (3.57)			2 (5.56)									1 (1.45)	1 (1.96)		3 (1.03)				
Sum	28	1639	815	36	1574	1307	1864	1369	1042	896	361	688	69	51	914	291	2286	3341	2504	1101

* Within bracket indicates coverage percentage; gray block indicates coverage and frequency both above 5 and 5% respectively.