Forensic Science International: Genetics 32 (2018) 88–93

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

Short communication

Massive parallel sequencing of mitochondrial DNA genomes from mother-child pairs using the ion torrent personal genome machine (PGM)

Ke Ma[a,*], Xueying Zhao[a], Hui Li[a], Yu Cao[b,c], Wei Li[d], Jian Ouyang[d], Lu Xie[d], Wenbin Liu[a,*]

[a] Shanghai Key Laboratory of Crime Scene Evidence, Shanghai Research Institute of Criminal Science and Technology, Shanghai 200083, China
[b] Key Laboratory of Forensic Evidence and Science Technology, Ministry of Public Security, Institute of Forensic Science, Shanghai Public Security Bureau, Shanghai 200083, China
[c] State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, Institute of Genetics, School of Life Sciences, Fudan University, Shanghai 200438, China
[d] Shanghai Center for Bioinformation Technology, Shanghai Academy of Science and Technology, Shanghai 201203, China

ARTICLE INFO

Keywords:
mtDNA
Population genetics
Heteroplasmy
Massively parallel sequencing (MPS)

ABSTRACT

Mitochondrial genome analysis is a potent tool in forensic practice and in the understanding of human phylogeny in the maternal lineage. With the development of molecular biology and bioinformatics techniques, high-throughput sequencing has enabled mtDNA analysis during whole genome sequencing, which provides more comprehensive information and raises the power of discrimination. In this study, peripheral blood samples were taken from 194 mother-offspring pairs and sequenced by Ion Torrent Personal Genome Machine and obtained high-coverage mitochondrial sequencing data, demonstrating the mutation levels at each position in the mitochondrial DNA (mtDNA) between maternally related pairs. A total of 14,332 variants were observed at 891 nucleotide positions from 388 individuals, and the result shows that all maternally related pairs shared the same detailed homoplasmic SNPs and haplotypes. With appropriate criteria for avoiding false positives due to sequencing errors and contamination by nuclear mitochondrial pseudogenes, we identified 33 heteroplasmies at a frequency of ≥10% at 32 sites in 30 pairs. The maternally related pairs had the same heteroplasmic sites but with different allele frequencies. The dataset is available through EMPOP under accession number EMP00684 and will serve as an mtDNA reference database in forensic casework in Eastern China.

1. Introduction

The mitochondrial genome harbors 37 genes in a circular molecule of approximately 16.6 kb in size [1]. Mitochondrial DNA (mtDNA) possesses three mainly unique characteristics, including a uniparental, non-recombining mode of inheritance and its high mutation rate compared to that of the nuclear genome [2,3]. MtDNA is widely used as a marker in molecular evolution, population genetics, and forensics [4–8]. In forensic science, it is hard to obtain full short tandem repeat (STR) profiles from specimens like highly decomposed remains, bones, or hair shafts, and in such cases, mtDNA typing can provide an alternative clue to investigators by comparison with known maternal relatives [9–12]. Due to its maternal mode of inheritance and lack of recombination, mtDNA is particularly distinctive and informative in kinship analyses.

In recent years, massively parallel sequencing (MPS) technologies have matured in step with cost reductions, which provides a powerful tool for studying mtDNA [13–16]. MPS technology can increase both sample throughput and overall process efficiency, thereby holding great potential for efforts to expand mtDNA typing beyond current capabilities. MPS technology is currently being used in mtDNA validation, large mtGenome reference data establishment, and the investigation of heteroplasmy occurrence [17–19]. Heteroplasmy, the presence of more than one mtDNA variant in a cell or a tissue, occurs with appreciable frequency in the general population and must be considered in the interpretation of forensic evidence [20]. Individuals may possess mtDNA molecules that differ in their length (length heteroplasmy; LHP) or at single nucleotide positions (point heteroplasmy; PHP) [21]. In the past 10 years, MPS technologies have quickly developed and replaced more traditional methods of DNA sequencing and typing. The high-coverage generated by high-throughput sequencing methods provides a powerful tool for the study of mtDNA heteroplasmy [22,23]. For heteroplasmy variant detection, MPS technology can detect heteroplasmic variants at mutation levels as low as 1% [24]. However, it is hard to distinguish if such low-level heteroplasmy sites are due to true mutations or false positive mutations because of amplification errors, sequencing errors,

* Corresponding authors.
E-mail addresses: nmgdxmark@126.com (K. Ma), wbliu1981@163.com (W. Liu).

https://doi.org/10.1016/j.fsigen.2017.11.001
Received 24 July 2017; Received in revised form 26 September 2017; Accepted 5 November 2017
Available online 06 November 2017

contamination by nuclear mitochondrial pseudogenes (NUMTs) and true mtDNA sequencing due to the baseline error rates of high-throughput sequencing. Sanger-type sequencing is not sensitive enough to confirm the very low heteroplasmy levels [24,25].

To conduct a population study of heteroplasmy frequency and its transmission between maternal pairs, we analyzed full-length mtDNA in 194 mother-children pairs using the Ion Torrent PGM platform. High sequencing depth was used to uncover very low-level heteroplasmy sites and improve our understanding of mtDNA heteroplasmy and mutation, leading to further refinement of its treatment and utility in forensic mtDNA casework. With those data, we addressed i) mtDNA differences between mothers and children in variants levels and heteroplasmy levels, ii) how common heteroplasmy is in a Chinese population, and iii) how heteroplasmy frequency changes between generations.

2. Materials and methods

2.1. Sample collection and DNA extraction

Permission for the study was granted by the Ethical Committee of Fudan University. Peripheral blood samples were taken from194 mother-offspring pairs by venipuncture with written informed consent. All of the samples were collected from Han Chinese in the Shanghai region. Details of each family and their haplogroup information are described in Table 1. Genomic DNA was extracted using a BioRobot EZ1 Advanced XL and EZ1 DNA Investigator Kit (QIAGEN,Hilden, Germany) according to manufacturer's protocols. The quantity of recovered DNA was determined using a Qubit® dsDNA BR Assay Kit on a Qubit 2.0 Fluorometer (Thermo Fisher, Foster City, CA, USA).

2.2. Mitochondrial enrichment and sequencing

Genomic DNA was amplified using a KOD FX Neo PCR Kit (TOYOBO, Japan) with four primer sets; details on this method are described in our previous publication [26]. This protocol specifically amplifies the entire mitochondrial genome from genomic DNA using overlapping primers to eliminate the bias that may be introduced from the PCR method. The enriched mtDNA was barcoded and sequenced using the Ion Torrent PGM high-throughput sequencing platform (Thermo Fisher, Foster City, CA,USA) as described by Jiang et al. [17].

2.3. Sequencing data analysis

The mtGenome sequencing data were analyzed with the Ion Torrent Software Suite (v 5.0.1) using the plug-in variant caller (v 5.0). The output of the variant caller was presented as a list of base differences relative to the revised Cambridge Reference Sequence (GenBank: NC_012920, rCRS) (Andrews et al., 1999). Mitochondrial haplogroups were assigned to haplotypes for each individual using Mito Tool, a web server based on PhyloTree Build 17. Variants not previously observed in the database or variants expected but not observed were validated by visualizing BAM files using Integrative Genomics Viewer (IGV). For the purpose of this study, the following criteria were used for variant calling: a heteroplasmy threshold of 0.1, and the coverage threshold of 40 × . It meant that positions would only be interpreted if there was a minimum of 40 × coverage and point heteroplasmy at which would be displayed ≥4 × [9,23]. Length variation of the poly C-tracts in regions 303–316 and 16184–16193 was not considered in this study because most of them were heteroplasmic for multiple poly C-tracts (> 10 Cs) and could not be accurately sequenced by the Ion Torrent PGM [14].

2.4. Sanger-type sequencing

For the heteroplasmic mutation sites, Sanger-type sequencing was employed to confirm PGM sequence data. Primers for amplification and sequencing are listed in Table S4. Sequencing was performed on a 3130 × l Genetic Analyzer (Thermo Fisher, Foster City, CA, USA) with a BigDye® Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher, Foster City, CA, USA). Raw data were analyzed using Vector NTI Advance11 Software v11 (Thermo Fisher, Foster City, CA, USA).

3. Results and discussion

3.1. Sequencing overview

A total of 4.6 billion reads were generated from the 194mother-children samples, and 88% of the reads could be mapped to the rCRS. Though the average coverage provides some measure of data quality, the reliability of any given base call is dependent on read coverage at that particular position. Sufficient data were obtained to reliably determine the mtGenome sequence, and the coverage was similar among all samples. The mean sequencing depth per sample was 647 × (± 235) (mean ± SEM). The average coverage of both forward and reverse strands at each nucleotide position is displayed in Fig. 1.The observed pattern of coverage was consistent with what has been described in previous studies [17,26], in which coverage was reduced around the poly-C stretches 303–316 and 16184–16193, and low coverage around 3552–3575 and 8605–8625. Positions with low coverage might be attributed to homopolymeric stretches. The strand balance percentage (lower coverage/higher coverage) of all samples at all positions was calculated and is displayed in Fig. 2. In all 388 samples, 90% of all positions had a strand balance percentage > 40%, which was similar with the previous studies by Seo et al. [14] and zhou et al. [17]. From Fig. 1, some regions with low coverage and imbalance were observed. Those regions were mainly located around the poly stretches which were hard to be sequenced. After reviewing the BAM files of the samples in IGV, it was found that poly-C regions were more difficult to be sequenced than other poly stretches [13,14], which would explain why those regions were with high imbalance. Compared with poly-C region in the forward strand sequencing, poly-G region was much easier to be sequenced for the reverse strand sequencing. Potential methods to overcome the problem of bias and coverage variation would improve both quantity and quality of sequencing data in mtGenome research.

3.2. Sequence variants and haplogroups in maternal pairs

MPS reads were represented using a graphical viewer to understand MPS data features and interpret the discrepancy of variant calling, as well as to discriminate sequencing noise [9]. In this study, the mtGenome sequencing data were analyzed with the Ion Torrent Software Suite (v 5.0.1) using the plug-in variant caller (v 5.0) and manually confirmed using IGV. The output of the variant caller was presented as a list of base differences relative to the revised Cambridge Reference Sequence (rCRS, NC_012920.1).In total, 14,332 variants were observed at 891 nucleotide positions from 388 Eastern Han Chinese (Table S1), where 13,536(94.44%) were substitutions and 796 (5.56%) were indels. Among the substitutions, 13,114 transitions and 422 transversions were observed (31.07:1 ratio of transitions to transversions). A "heatmap" was generated for variants observed from 388 Chinese Han and constructed by plotting mtGenome position versus the number of variants observed at each variants position (Fig. 3, Table S2). Variants 750G, 4769G, 7028T, 8860G, 11719A, 14766G and 15326G were observed in all samples, and variants 73G, 263G, 1438G, and 2706G were observed in at least from 360 samples. There were 1874(13.08%) variants within HVS-I (Table S1), which ranked as the highest polymorphism density (Fig. 3), and 3836 of the variants (26.77%) resided in the Control region (CR). Thus, 10,496 of the variants (73.23%) were located outside of the CR of the mtGenome, which demonstrates the potential value of the coding region to improve the discrimination power of mtDNA typing in forensic testing. The haplogroup