MITOCHONDRIAL DNA SEQUENCING USING THE PRECISION ID NGS SYSTEM AND CONVERGE ANALYSIS SOFTWARE: A ROBUST AND SENSITIVE TILED AMPLICON ASSAY FOR FORENSIC CASEWORK APPLICATIONS

R. Hasegawa[1], S. Wootton[1], C. Roth[1], CW. Chang[1]; Vijaychander, S. [1], Deng. J [1], R. Lagacé[1]
[1]Thermo Fisher Scientific, South San Francisco, CA 94080

ABSTRACT

In forensic casework, mitochondrial DNA (mtDNA) is useful in the context of recalcitrant samples that fail to produce a standard STR profile. Traditional Sanger sequencing using capillary electrophoresis (CE) compels a limitation of sequencing of the mtDNA genome to the hypervariable region as sequencing of the whole mitochondrial genome (mtGenome) is both time consuming and cost-prohibitive. With the availability of massively parallel sequencing systems (MPS), the mtGenome can easily be prepared and sequenced using a tiled amplicon multiplex of 162 amplicons. Additionally, the forensic mtDNA analysis module developed on Converge™ Software and optimized specifically for the Precision ID Control Region and Whole Genome panels provides streamlined analysis for haplotype and haplogroup designations as well as robust detection of nuclear mitochondrial DNA segments (NUMTs) and point and length heteroplasmies. DNA from samples with known haplotypes were obtained through Coriell and NIST. Libraries were prepared on the Ion Chef using the Precision ID mtDNA Control and Whole Genome Panel and sequenced on the Ion S5. Reads generated on the system were aligned and compared to the rCRS and were evaluated for concordance, amplicon coverage uniformity, presence of artefacts, heteroplasmies, and NUMTs using the mtDNA analysis module on Converge™.

INTRODUCTION

Traditionally, mtDNA sequencing has been limited to the control region due to the cost and time demands of whole genome Sanger sequencing. With MPS, the entire genome can be sequenced in approximately the same amount of time as the control region, and samples can be multiplexed. Additionally, the interrogation of each base at high read depths allows for identification of possible mixtures as well as point and length heteroplasmies at rates above the baseline noise of the system.

MATERIALS AND METHODS

The Precision ID mtDNA Whole Genome Panel was used to amplify the whole mito genome (mtGenome) using a two pool system of 81 primer pairs each with minimal overlap between pools. The Precision ID mtDNA Control Region Panel contained two pools with 7 primer pairs each covering the hypervariable regions I, II, and III (15954– 610). To ensure full genome coverage, both panels were designed with degenerate bases covering SNPs that may affect primer binding.

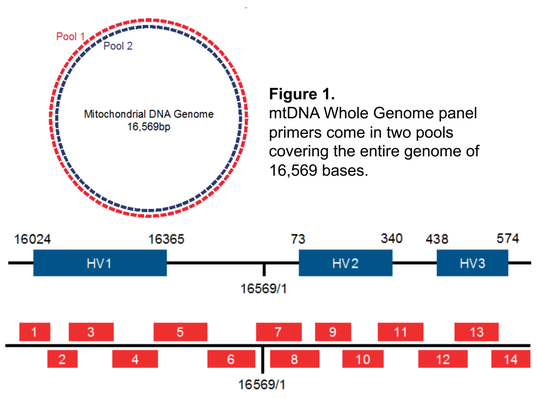


Figure 1. mtDNA Whole Genome panel primers come in two pools covering the entire genome of 16,569 bases.

Figure 2 The Hypervariable Regions of the mtDNA Control Region and insert locations of the Precision ID mtDNA Control Region Panel.

Samples with Sanger sequenced mtGenomes were selected for investigating concordance. DNA from HL-60, CHR, and GM09947A were obtained as part of NIST standard reference material (SRM 2392[1,2]), and GM10742 from the Coriell Institute (Table 1). Additionally, samples from International Genome Sample Resource (IGSR)[3] were obtained, covering a small range of populations (Table 2). Two samples (HG01260 and HG01389) were run at four varying input DNA quantities in quadruplicate. The rest of the IGSR samples were run at 100pg starting DNA input to check coverage uniformity and to test for amplicon dropouts.

Table 1. Samples Used for Concordance with Sanger

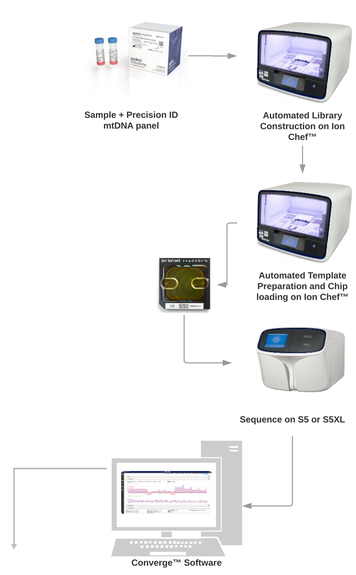
Sample ID	SRM ID	Input DNA (gDNA)
HL-60	SRM2392	100pg
CHR	SRM2392	100pg
GM09947A	SRM2392	100pg
GM10742	SRM2392	100pg

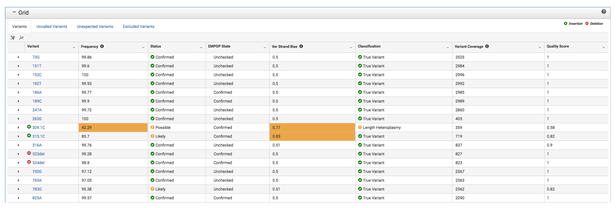
Table 2. Samples Used for Sensitivity and Amplicon Performance

Sample ID	Population Description	Gender	Input DNA quantity (gDNA)	
HG00336	Finnish in Finland	male		IGSR
HG00369	Finnish in Finland	male		IGSR
HG00428	Southern Han Chinese, China	female		IGSR
HG00525	Southern Han Chinese, China	female		IGSR
HG00590	Southern Han Chinese, China	female		IGSR
HG00689	Southern Han Chinese, China	male		IGSR
HG01124	Colombian in Medellin, Colombia	male		IGSR
HG01260	Colombian in Medellin, Colombia	female	100pg, 10pg, 5pg, 2pg	IGSR
HG01378	Colombian in Medellin, Colombia	female		IGSR
HG01389	Colombian in Medellin, Colombia	male	100pg, 10pg, 5pg, 2pg	IGSR
HG01462	Colombian in Medellin, Colombia	female		IGSR
HG01464	Colombian in Medellin, Colombia	male		IGSR

Ampliseq™ libraries were constructed using the manufacturer's protocol for the Precision ID DL8 Kit on the Ion Chef™. Templates were prepared using the Ion S5™ Precision ID Chef and Sequencing Kit and loaded on 510, 520, and 530 chips. Sequence was analyzed using a beta version of the NGS mtDNA module of Converge™ Software using default parameters. Reads were first mapped to nodes in Phylotree and then re-aligned using a context aware custom Smith-Waterman alignment algorithm that integrates phylotree and EMPOP information into the scoring function. Variants were called with reference to the rCRS. Additionally, the closest haplogroup was calculated, and variants were evaluated based on their occurrence in the haplogroup as well as other general metrics including frequency, strand bias, and coverage.

Figure 3. Workflow for mito DNA using Precision ID and Converge™ Software.





Grid view of variants on Converge™. Variant information including variant frequency, status (confirmed, likely, possible, unclear, and unlikely), EMPOP state (whether the variant is expected in the closest haplogroup), strand bias, classification (true variant, PHP, LHP, NUMT, or artefact), coverage, and quality score.

Linear coverage plot on Converge™. Forward (blue) and reverse (red) coverage shown across the entire mtGenome (upper pane is the view of the zoom window). Variants plotted below the coverage diagram are colored by their status (green, yellow), and low coverage regions are marked by red boxes.

RESULTS

The panel showed high uniformity and efficiency for both the whole genome and control region panels (Table 3). Compared to Sanger (Table 4), variant calling with SRM 2392 samples showed little discordance, limited mostly to the 309 position. Sample normalized coverage across all amplicons in the mtGenome (Figure 4) and also sorted coverage (Figure 5) shows amplicon uniformity across the mtGenome.

Table 3. Performance of Panel

Metric	Whole Genome	Control Region
Uniformity	97.46%	88.09%
On Target Percentage	100%	99.98%

Table 4. Concordance with SRM 2392 and SRM 2392-1 haplotypes

Sensitivity	Specificity	Accuracy
98.5%* (n=132)	99.997%** (n=66144)	99.994% (n=66276)

*2 false negatives: 1 missed 309.2 and reported 2841 T-> A in GM10742.
**2 extra variants: 2 different samples with 309DEL at about 50% freq.

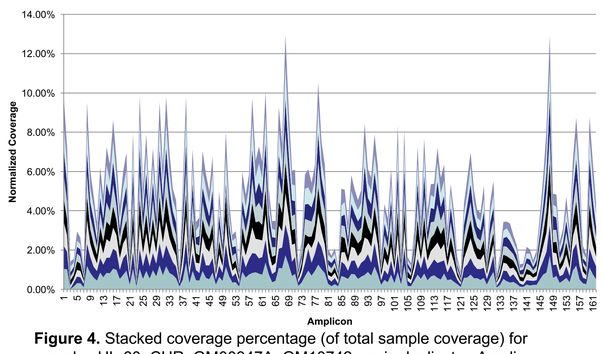


Figure 4. Stacked coverage percentage (of total sample coverage) for samples HL-60, CHR, GM00947A, GM10742 run in duplicate. Amplicon number (up to amplicon 162) are on the x-axis covering the mtGenome.

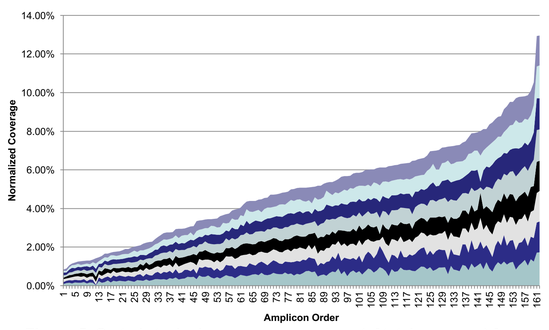


Figure 5. Sorted stacked coverage percentage (of total coverage) for samples HL-60, CHR, GM00947A, GM10742 run in duplicate ordered from lowest coverage to highest.

PHPs were detected at a rate comparable to the published recharacterization of NIST SRM-2392 and SRM-2392-I using MPS[1]. All expected point heteroplasmies were detected except at position 3242 (G/A) in 9947A, which is possibly at lower frequency (it was not detected in another study using an alternative sequencing platform[4]) and 2445 (T/C) in HL-60.

Table 5. PHP detected

Sample	Position	PHP	Frequency	Detected in Sanger
CHR	64	C/T	29.5	Yes
9947A	1393	G/A	18.9%	No
9947A	7861	T/C	85.6%	No
HL-60	5149	T/C	11.1	No
HL-60	12071	T/C	49.8	Yes

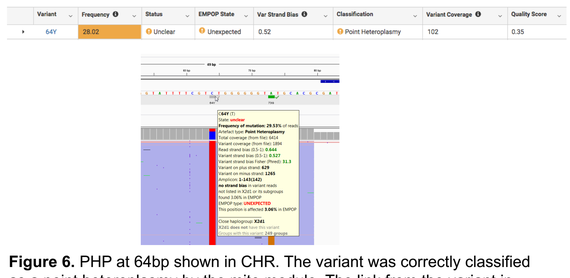


Figure 6. PHP at 64bp shown in CHR. The variant was correctly classified as a point heteroplasmy by the mito module. The link from the variant in the grid view shows the supporting reads in mito IGV.

No complete amplicon dropouts were observed in the runs, including the DNA input titration from 100pg starting gDNA down to 2pg. Coverage across the genome and uniformity (% of amplicons less than 0.2x mean amplicon coverage) remained stable across input amounts (Table 6).

Table 6. Sensitivity of mtDNA panel

Sample	Input (pg)	% Uniformity	Avg Amplicon Coverage	Number of Variants detected
HG1389	100	91.6	2008	90.75 (98.6%)
	10	89.6	2413	91.5 (99.5%)
	5	89.1	2293	91.5 (99.5%)
	2	91.2	2318	90.75 (98.6%)
HG1260	100	93.4	1935	89 (100%)
	10	92	2528	89 (100%)
	5	93.1	2243	89 (100%)
	2	91.1	2430	89 (100%)

*Each input level was sequenced in quadruplicate and the average of the four runs is presented in the table. Input is the gDNA quant for the sample.

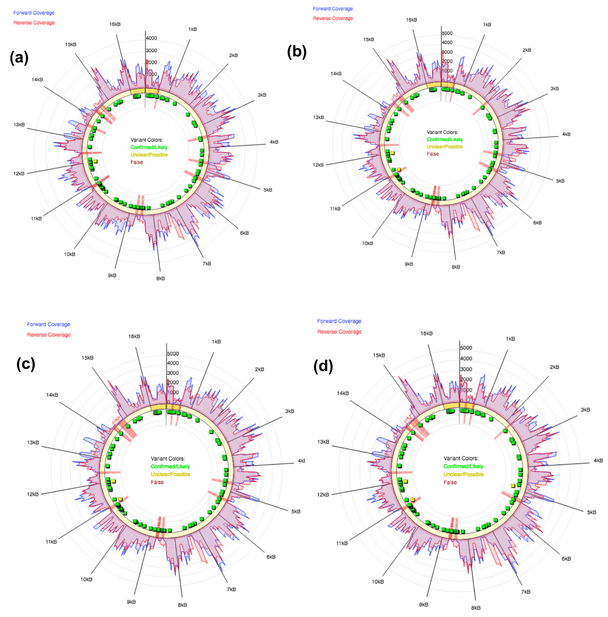


Figure 7. Performance of the panel and system at 100pg (a), 10pg (b), 5pg (c), and 2pg (d) gDNA input.

CONCLUSIONS

Control region and whole genome mito sequencing can be applied in cases where DNA is limited or has been degraded. Sequencing of the panel with the Precision ID mtDNA panel on S5 using the mtDNA analysis module on Converge™ showed high concordance with Sanger, primarily with exception of the 309 position, where Converge™ marked the call as "unclear." Heteroplasmies could be detected at a rate similar to previously published MPS studies, and PHP and LHP thresholds could be lowered as a custom setting to detect lower level heteroplasmies not detected here. Sequencing of samples from 100pg of starting input down to 2pg showed remarkably little difference in coverage uniformity, dropouts, and variant calling performance. In a practical casework laboratory, the Precision ID system can be implemented in routine missing persons / disaster victim identification workflows as a robust mtDNA analysis option in cases were STRs have failed. The benefits of MPS testing relative to standard HV region sequencing using Sanger methods (e.g., increased discrimination with mtGenome sequencing, improved heteroplasmy detection and overall system sensitivity with limited and/or degraded DNA) for mtDNA analysis offer benefits for missing persons identification.

REFERENCES

1. Riman, S., Kiesler, K. M., Borsuk, L. A., & Vallone, P. M. (2017). Forensic Science International: Genetics. *Forensic Science International: Genetics*, *29*, 181–192. http://doi.org/10.1016/j.fsigen.2017.04.005
2. Levin BC, Hancock D, Holland K, Cheng H, Richie K. Human mitochondrial DNA-- amplification and sequencing standard reference materials--SRM 2392 and SRM 2392-I. US Department of Commerce, National Institute of Standards and Technology; 2003.
3. The 1000 Genomes Project Consortium, BGI-Shenzhen, Broad Institute of MIT and Harvard, et al. (2011). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073. http://doi.org/10.1038/nature09534
4. Mikkelsen M, Frank-Hansen R, Hansen AJ, Morling N. Massively parallel pyrosequencing of the mitochondrial genome with the 454 methodology in forensic genetics. Forensic Sci Int Genet. 2014 Sep;12:30-7. doi: 10.1016/j.fsigen.2014.03.014. Epub 2014 Apr 6. PubMed PMID: 24879032.

TRADEMARKS/LICENSING