JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, July 2014, Vol. 59, No. 4
doi: 10.1111/1556-4029.12441
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

CRIMINALISTICS

Brittania J. Bintz,[1] *M.S.; Groves B. Dixon,*[1] *M.S.; and Mark R. Wilson,*[1] *Ph.D.*

Simultaneous Detection of Human Mitochondrial DNA and Nuclear-Inserted Mitochondrial-origin Sequences (NumtS) using Forensic mtDNA Amplification Strategies and Pyrosequencing Technology*,†

ABSTRACT: Next-generation sequencing technologies enable the identification of minor mitochondrial DNA variants with higher sensitivity than Sanger methods, allowing for enhanced identification of minor variants. In this study, mixtures of human mtDNA control region amplicons were subjected to pyrosequencing to determine the detection threshold of the Roche GS Junior® instrument (Roche Applied Science, Indianapolis, IN). In addition to expected variants, a set of reproducible variants was consistently found in reads from one particular amplicon. A BLASTn search of the variant sequence revealed identity to a segment of a 611-bp nuclear insertion of the mitochondrial control region (NumtS) spanning the primer-binding sites of this amplicon (*Nature* 1995;378:489). Primers (*Hum Genet* 2012;131:757; *Hum Biol* 1996;68:847) flanking the insertion were used to confirm the presence or absence of the NumtS in buccal DNA extracts from twenty donors. These results further our understanding of human mtDNA variation and are expected to have a positive impact on the interpretation of mtDNA profiles using deep-sequencing methods in casework.

KEYWORDS: forensic science, mitochondrial DNA, next-generation sequencing, pyrosequencing, minor variant, pseudogene

The possibilities offered by next-generation sequencing (NGS) technologies are revolutionizing biotechnological laboratories. Over the past five to 7 years, large-scale sequencing has been realized by the development of several NGS platforms and chemistries. These technologies provide an unprecedented tool for numerous biological applications (1,2). Although each chemistry and accompanying instrument varies, the output from an NGS run can exceed several gigabases of sequence data. These technologies are increasingly used for various nucleic acid sequencing-related applications. However, several potential artifacts, including read errors (base calling errors and small insertions/deletions), poor quality reads, and primer or adaptor contamination can occur in the NGS data, which can impact the downstream sequence processing/analysis (3,4). For forensic applications, validation of NGS requires a thorough understanding of these potential sources of interpretational error, within a contextual grasp of the potential sources of error in casework applications and the potential for the actual mistyping of a sample or a misinterpretation of a profile comparison.

DNA extracted from challenging sample matrices including bones and hair can be degraded and/or contain very little DNA. Mitochondrial DNA (mtDNA) analysis is often utilized on these kinds of samples (5–8). However, in a forensic setting, human mtDNA analysis is currently limited in both breadth (the amount of sequence data obtained) and depth (the ability to detect minor variants arising from mutations but present at very low levels).

Due to the presence of relatively fast-changing sites (hot spots) within the mtDNA genome, subtle sequence variants are often observed between cells or tissues within an individual, a phenomena referred to as heteroplasmy (9,10). Rather than being viewed as an anomaly, heteroplasmy is actually a principle of mitochondrial DNA genetics. If it is deemed desirable to use such a locus for forensic purposes, our conception of what constitutes a match, that is, a failure to exclude, should be widened to consider the possibility of observing heteroplasmy in casework. Accordingly, interpretational guidelines should be developed that are cognizant of these facts.

Two distinct dimensions of information available from mtDNA are the number of nucleotide positions that are included in the analysis (breadth), and the rigor with which each position is interrogated (depth). That is, additional information can be obtained by increasing the proportion of the genome that is examined as well as closely interrogating each base pair to determine whether or not minor variants arising from mutations are

[1] Forensic Science Program, Department of Chemistry and Physics, Western Carolina University, 111 Memorial Drive, NSB #231, Cullowhee, NC 28723.

*Supported by the National Institute of Justice through grant award number 2010-DN-BX-K171.

†Presented at the 23rd International Symposium of Human Identification, October 15–18, 2012, in Nashville, TN; at the 2012 Green Mountain DNA Conference, August 1–3, 2012, in Burlington, VT; and at the 65th Annual Meeting of the American Academy of Forensic Sciences, February 18–23, 2013, in Washington, D.C.

Received 27 Aug. 2012; and in revised form 10 May 2013; accepted 1 June 2013.

present at very low levels. Using emerging technologies, an extension of the breadth of sequence data obtained can easily extend to the entirety of the human mtDNA genome. This advance essentially expands the amount of exclusionary power of mtDNA analysis and hence renders this locus of greater utility in forensic casework analyses.

With respect to information in the complementary dimension (depth), NGS technologies are also well suited. In effect, using these methods, each molecule generated during the amplification phase is independently sequenced and reported as a separate, individual item of information called a read. This means that if the original template is mixed, each component of the mixture will each be independently analyzed and presented to the analyst. For most applications, the presence of minor variants among such a vast amount of data is not an issue. However, in some cases, including sequencing a mixture of different templates, which is common in metagenomics applications, the pattern and distribution of the variants becomes much more important and is in fact the purpose of the analysis (11).

Sequencing depth and accuracy arising from NGS applications have very important advantages that are specific to mtDNA analysis and other loci with widely disparate mutation rates. The minor component may be detected if the minor is at least 10% or more of the mixture (8). Using the current technology, the inability to detect the minor components of mixtures below this threshold has lead forensic analysts to interpret one base-pair differences between samples as inconclusive. A method that can reach below this threshold and capture the presence of low abundance components of mixtures could greatly assist in the forensic interpretation of mtDNA sequencing results, potentially revealing common low-level mixtures in both questioned and reference samples. Because of the ability of NGS to now detect low-level mixtures at greatly reduced levels, it is imperative that forensic laboratories begin to implement such improvements and potentially revise the current interpretational approach to casework comparisons.

Studies employing newly emerging DNA sequencing technologies have been designed to interrogate targets down to the single molecule level. Such studies have shown tissue differences (heteroplasmy) within individuals that have lead to suggestions for a reevaluation of current interpretational approaches to forensic mtDNA comparisons (12). To more fully understand this issue, we have performed mixture studies on human DNA templates and have shown that it is possible to detect minor variants at the 1% level using these new chemistries and instruments (data not shown). During these mixture experiments using the Roche GS Junior (Roche Applied Science, Indianapolis, IN) and the accompanying pyrosequencing chemistry, we encountered persistent low-level variants of high quality in a particular amplicon that were present in both strands at a consistent level. A search of the NCBI nucleotide database revealed that this sequence has been previously identified as a nuclear insertion of a mitochondrial DNA fragment (NumtS) (13–15). We performed a series of experiments to confirm that these variants are in fact mostly likely nuclear-inserted elements. We discuss the interpretational ramifications of these findings in the context of forensic science.

Materials and Methods

Forensically relevant samples including buccal swabs, whole blood on Whatman® FTA® (GE Healthcare, Buckinghamshire, U.K.) cards and hairs were collected from twenty donors according to the Human Subjects Institutional Review Board policies implemented at Western Carolina University and following informed consent. Reference sequence data for the mtDNA control region was obtained for all donors using Sanger methods. Pairs of donors exhibiting the highest amount of sequence variation within the control region were chosen for mixture studies using pyrosequencing to elucidate the minor variant limit of detection of the Roche GS Junior instrument (Roche Applied Science).

Obtaining Donor Reference Sequence Data

DNA from bloodstain card punches (1.2 mm) was purified using FTA® Purification Reagent (GE Healthcare) following manufacturers' protocol. The mtDNA hypervariable region was amplified in 4 distinct amplification reactions as follows: purified template DNA on 1.2 mm FTA punches in a reaction mixture containing 5 U of AmpliTaq Gold® DNA polymerase (Applied Biosystems®, Foster City, CA), 1× GeneAmp® PCR Buffer (Applied Biosystems®), 160 ng/μL BSA (Thermo Fisher Scientific, Rockford, IL), 200 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation, Madison, WI), 600 nM forward primer, and 600 nM reverse primer. Primer sequences are shown in Table 1. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems®) with an initial 11 min hold at 95°C, followed by 32 cycles comprised of a 15 s denaturation at 95°C, a 30 s annealing step at 56°C, and a 45 s extension at 72°C with a final hold at 4°C. Following amplification, unincorporated dNTPs and primers were enzymatically removed from each sample with ExoSAP-IT® (Affymetrix, Santa Clara, CA). Resulting amplification products were analyzed using the Agilent 2100 Bioanalyzer with the DNA 1000 kit (Agilent Technologies, Inc., Waldbronn, Germany) and were normalized to 1 ng/uL using TE^{-4} buffer (Teknova, Hollister, CA). Samples (5.0 ng) were cycle-sequenced using the BigDye® Terminator v1.1 (Applied Biosystems®) kit according to manufacturers' instructions, and fragments were separated on a 3130xl Genetic Analyzer (Applied Biosystems®) and analyzed with Sequencher® 5.0 software (Gene Codes Corporation, Ann Arbor, MI).

Sample Preparation and Pyrosequencing—Mixture Study

To determine the minor variant limit of detection of the Roche GS Junior pyrosequencing instrument (Roche Applied Science), a mixture study was designed in which mixtures were prepared from pairs of donors with maximum variability between their known mtDNA control region sequences. The mtDNA control region was amplified in four independent PCRs from whole

TABLE 1—*Mitochondrial DNA control region primer sequences. Primer IDs show strand represented (light vs. heavy) and rCRS position of the 3′ base of the primer sequence.*

HV Region	Primer ID	Primer Sequence 5′–3′
HV1a	A1 (L15997)	CAC CAT TAG CAC CCA AAG CT
	B2 (H16237)	GGC TTT GGA GTT GCA GTT GAT
HV1b	A2 (L16259)	TAC TTG ACC ACC TGT AGT AC
	B1 (H16391)	GAG GAT GGT GGT CAA GGG AC
HV2a	C1 (L 048)	CTC ACG GGA GCT CTC CAT GC
	D2 (H 285)	GGG GTT TGG TGG AAA TTT TTT G
HV2b	C2 (L 177)	TTA TTT ATC GCA CCT ACG TTC AAT
	D1 (H 409)	CTG TTA AAA GTG CAT ACC GCC

blood samples stored on Whatman® FTA® classic cards (GE Healthcare) from four donors (donors 001-CF30, 003-CM54, 005-CF40, and 015-AM30). To reduce the occurrence of polymerase-induced base misincorporations, the Roche FastStart High fidelity PCR system (Roche Applied Science) was used for DNA amplification as follows: purified template DNA on 1.2 mm FTA punches in a reaction mixture containing 1.25 U of Roche FastStart High Fidelity enzyme blend, 1× FastStart High Fidelity reaction buffer with 1.8 mm $MgCl_2$, 200 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation), 400 nM forward primer, and 400 nM reverse primer. Forward and reverse fusion primers (Fig. 1) were designed in which 454 specific adaptor sequences, and multiplex identifiers were included immediately 5′ of an mtDNA template-specific primer sequence (Table 1), enabling NGS sample preparation using PCR amplification. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems®) with an initial 2 min hold at 95°C, followed by 32 cycles comprised of a 30 s denaturation at 95°C, a 30 s annealing step at 60°C, and a 30 s extension at 72°C with a final 7 min extension at 72°C and a long-term 4°C hold. Resulting PCR products were purified with Agencourt® AMPure® XP beads (Beckman Coulter, Indianapolis, IN) for removal of unincorporated primers, and dNTP's. Purified amplicons were quantified in quintuplicate using the Agilent 2100 Bioanalyzer, and concentrations were averaged. The amplified products were normalized to 1 ng/μL and mixed in defined ratios of 10, 5, 2, 1, and 0.5%. Single-donor samples and prepared mixtures were then pooled at equimolar concentrations for multiplexed sequencing on the Roche GS Junior instrument Roche (Roche Applied Science).

Individual single-stranded template molecules from the pooled library were clonally amplified on the surface of paramagnetic DNA capture beads using the Roche GS Junior Titanium Lib-A kit (Roche Applied Science) for emulsion PCR (emPCR). This kit contains two sets of beads, each coated with oligonucleotides complementary to adapters corresponding to either the sense or antisense PCR products to allow for bidirectional sequencing reads. Enrichment for beads with clonally amplified sequencing template was performed, and resulting beads were deposited into independent microwells of a PicoTiter™ Plate (PTP) device. Pyrosequencing and raw image collection were carried out for 200 cycles. Pooled libraries were deep-sequenced across three independent pyrosequencing runs with a target of 5000× coverage per sample to capture minor variants at 1% or lower. Roche estimates 70,000 total reads per amplicon sequencing run on the GS Junior instrument (Roche Applied Science). As a result, sequencing of 14 pooled libraries gives rise to an estimated depth of coverage of 5000× per library, with 1% minor variant coverage of 50×. Previous studies in our laboratory have shown that it is possible to detect a 1% minor variant at and below this depth unambiguously (data not shown).

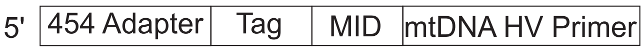

FIG. 1—*Roche 454 Fusion Primer schematic. Adaptor sequences are required for hybridization of amplicons to complementary adaptors on the surface of DNA capture beads where pyrosequencing chemistry takes place. Additionally, the adaptors serve as primer-binding sites for downstream pyrosequencing reactions. Multiplex identifiers (MIDs) are short sequences located upstream of the template-specific primer and allow the user to identify the provenance of each sequence read, while also enabling multiple libraries to be sequenced in a single run. Fourteen primer pairs with unique MIDs were used in this study for template amplification. Pairs of donors were used to create mixtures, and each pair was assigned identical MIDs to allow for postrun mixture-dependent sample pairs.*

Library Preparation—Tissue Comparison Study

DNA from three forensically relevant tissue types including hair, blood, and buccal cells from donor 001-CF30 was deep-sequenced to determine whether minor sequence differences exist between different tissue types originating from the same individual. DNA was extracted from five hair shafts with no root tissue from different regions of the scalp of donor 001-CF30 using the Qiagen DNA Investigator Kit (Qiagen, Hilden, Germany) and from buccal swabs using the Qiagen DNA mini kit (Qiagen, Germany) with no modifications to the vendor recommended protocols. Extracts were quantified using a human mtDNA-specific 5′-nuclease real-time PCR assay (16). DNA was amplified using the Roche FastStart PCR System (Roche Diagnostics, Indianapolis, IN) and fusion primers for 454 library preparation as described above, with 10 μL of template added per reaction from buccal and hair extracts. Additionally, DNA was purified from whole blood samples stored on Whatman® FTA® classic cards and amplified directly using the Roche FastStart PCR System. DNA extracts were assigned unique multiplex identifier (MID's) for postrun sample parsing by tissue type. Resulting PCR products were purified using Agencourt® AMPure® XP beads, were quantified using an Agilent 2100 Bioanalyzer and the DNA 1000 kit (Agilent, Waldbronn, Germany), and were normalized to a concentration of 1 ng/μL using TE^{-4} pH 8.0 (Teknova). Normalized samples were then pooled and clonally amplified using emPCR as described in the section titled Sample Preparation and Pyrosequencing—Mixture Study. A 200-cycle sequencing run was performed.

Roche GS Junior 454 Data Analysis

Roche Amplicon Variant Analyzer (AVA) software v2.7 was used for identification and quantification of minor variants using default analysis parameters. This software is capable of parsing reads according to the MID detected, aligning reads against a specified reference sequence, and generating a list of putative variants and their occurrence frequency within each library. All libraries were aligned against the rCRS (17). Putative variants were called, where $\geq$20 reads differed at a given position from the reference sequence, and were quantified as a proportion of reads from a given library sharing the same MID. In addition to standard quality filters applied to the data set by the AVA software, minor variants were only further considered "real" if they appeared bidirectionally within the library.

PCR Confirmation of NumtS

DNA was extracted from buccal swabs from twenty donors using the QIAamp® DNA mini kit (Qiagen, Valencia, CA). Extracted nuclear DNA was quantified with real-time PCR using the Quantifiler™ Human DNA Quantification Kit (Applied Biosystems) and 7500 real-time PCR system (Applied Biosystems) according to the manufacturer's instructions. Extracts were normalized to a concentration of 1 ng/μL with TE^{-4} buffer, pH 8.0 (Teknova). Two control samples commonly encountered in forensic laboratories were also included in the sample set (9947A and HL60). A nuclear-specific primer set (Integrated DNA Technologies, Coralville, IA) designed by Thomas et al. (15) to flank the region containing the NumtS insertion was used

TABLE 2—*Primer sequences for NumtS-specific amplification (from Thomas et al.) (15). Amplicon size without NumtS insertion = 155 bases; amplicon size with NumtS insertion = 711 bases.*

Primer Sequence	Nucleotide position on chromosome 11
F 5′-AGTCTTGCTTATTACAATGATGG-3′	49,840,047–49,840,069
R 5′-ACAAAGTCCAGGTTTCTAACAG-3′	49,840,174–49,840,195

to confirm the presence of the NumtS in all twenty donors. Samples were amplified using the Roche FastStart High Fidelity PCR System (Roche Diagnostics) (Table 2). Initially, input quantities ranging from 5.8 to 14.2 ng were amplified. The resulting yields of PCR products were low in most cases (0.04–0.95 ng/μL) and it was thought that amplification failure of the NumtS insertion-negative peak might be occurring. As a result, extracted DNA samples were then re-amplified with increased input amounts of DNA as follows: 10 μL of purified buccal extract (ranging from 24.0–147.0 ng input) in a reaction mixture containing 2.5 U of Roche FastStart High Fidelity enzyme blend, 1× FastStart High Fidelity reaction buffer, 3.0 mm $MgCl_2$ (Applied Biosystems), 4% DMSO, 400 μM each dATP, dTTP, dCTP, dGTP from PCR grade nucleotide mix (Promega Corporation), 400 nM forward primer, and 400 nM reverse primer. Primer sequences are shown in Table 2. Reaction mixtures were amplified on a GeneAmp® PCR System 9700 (Applied Biosystems) with an initial 2 min hold at 95°C, followed by 32 cycles comprised of a 30 s denaturation at 95°C, a 30 s annealing step at 60°C, and a 30 s extension at 72°C with a final 7 min extension at 72°C and a long-term 4°C hold. Resulting amplification products were analyzed using the Agilent 2100 Bioanalyzer and DNA 1000 kit (Agilent Technologies). Individuals heterozygous for the NumtS insertion were defined as those who presented two peaks with sizes of *c.* 155 bp and 711 bp. Homozygous individuals included those who showed a single peak at *c.* 155 bp (no NumtS on either chromosome) or 711 bp (NumtS insertion on both chromosomes).

Sanger Sequencing of NumtS

Initially, HV1b amplification primers were used to sequence the NumtS insertion amplicons. PCR products (10 ng) from all twenty donors were sequenced in forward and reverse directions using the BigDye® Terminator v1.1 Cycle Sequencing kit (Applied Biosystems) according to manufacturers' instructions. Fragments were separated on a 3130xl Genetic Analyzer (Applied Biosystems) and analyzed with Sequencher® 5.0 software (Gene Codes Corporation). To show that nuclear DNA was being sequenced exclusively, amplified samples were then Sanger sequenced as described above using the nuclear-specific primers originally used to amplify the NumtS insertion region (Table 2). At the point in the sequence where the insertion is expected, the resulting sequence data from heterozygous individuals were out of phase, resulting from the co-amplification and cosequencing of inserted and noninserted alleles. To confirm the mixed nature of these templates, PCR products were run on a 1.2% agarose gel at 100 V for 20 min and the resulting bands were excised from the gel to isolate NumtS-negative and NumtS-positive products for further sequence analysis. DNA purification was accomplished with the Qiaquick® Gel Extraction kit (Qiagen, Germany) according to manufacturers' instructions. Purified amplicons were quantified using the Agilent 2100 Bioanalyzer and the DNA 1000 kit with internal sizing standards. Quantified amplicons were Sanger sequenced independently as described above with nuclear DNA-specific primers.

Results

Variant Classification

During the studies described herein, we encountered four types of variants, as follows: (i) donor-dependent expected variants (polymorphisms), (ii) potential heteroplasmic sites, (iii) variants associated with a previously characterized (13) nuclear-inserted sequence of mitochondrial origin (NumtS), and (iv) background noise. We classified a variant as heteroplasmic if it was observed across different NGS runs from DNA prepared from the same donor, and forward and reverse reads were balanced. This classification was further strengthened if the position of the variant was a previously characterized mutational hot spot. NumtS-associated variants were easily classified as such because they were found only in nuclear-rich sample types, matched the reported sequence of previously characterized insertions, were reproducible from run-to-run, and were observed as a clustered set within particular reads. Variants classified as background noise appeared to be sporadic and had widely variable frequencies between forward and reverse reads. Further, these variants were not reproducible and are often associated with homopolymeric stretches of 2 or more consecutive identical bases and hence could easily be filtered from the data set by applying minimal quality filtering during analysis.

Mixture Study—HV1a Data

Mixtures of HV1a amplicons were prepared in defined ratios to determine the limit of detection of minor sequence variants using the Roche GS Junior 454 pyrosequencing instrument (Roche Applied Science) as described above. Single-donor amplicons or prepared mixtures were bioinformatically parsed using a sample-dependent MID sequence and showed an average depth of coverage of 1950 reads. The frequencies of variants reported by the Roche AVA software for single donors and prepared mixtures are shown in Table 3. In addition to the expected variants arising from the minor contributor, several variants with frequencies below 1.5% were detected. All HV1a low-level variants (≤ a frequency of 1.5%), except those associated with the minor contributor, appeared in either a forward or reverse read (but not both) and as a result were excluded from the data set by applying a bidirectional read variant confirmation quality filter, and thus were considered to be background noise. This quality filter is built into the AVA software and can easily be applied to any data set. Table 4 shows the variants that remained after quality filtering.

Mixture Study—HV1b Data

Roche AVA software successfully parsed all HV1b samples based on the MID sequence and allowed for detection of all expected variants with an average read depth of *c.* 8158 per sample. In addition to the expected variants arising from the mixture, a subset of unexpected variants was detected with an average combined frequency of 0.99%. In contrast to the HV1a data set, bidirectional filtering did not remove these variants from the HV1b data set, as shown in Table 5. This subset of variants was reproduced with an average combined frequency of

TABLE 3—*Roche GS Junior 454 pyrosequencing mixture experiment: HV1a variants detected using Roche Amplicon Variant Analyzer software. Variants highlighted in gray represent expected variants originally obtained for each sample donor using Sanger sequencing. Mixtures were prepared in defined ratios by combining quantified HV1a amplification products from donors 001-CF30 and 005-CF40. Data are presented as the frequency of each variant detected per multiplex identifier (MID).*

Variant Position	Donor 001	Donor 005	0.5% 001	1% 001	2% 001	5% 001	95% 001	98% 001	99% 001	99.5% 001
16069: C/T	99.95	0	0.09	1.1	1.85	3.89	95.04	98.3	99.43	99.35
16093: T/C	97.97	0	0.09	1.15	1.8	3.95	93.78	96.78	97.86	97.63
16124: T/G	0	0.69	0.67	0.62	0.31	0	0.9	0.6	0.61	1.28
16125: G/A	0	0.69	0.71	0.62	0.31	0	0.99	0.6	0.65	1.31
16126: T/C	99.09	0	0.26	1.1	1.8	3.96	94.1	97.06	98.08	97.62
16129: G/A	0	97.51	98.28	97.53	96.92	94.59	4.49	1.56	0.28	0.7
16130: G/T	0	0	0	0	0	0.45	0.51	0.24	0.33	0
16131: T/G	0	0	0	0	0	0.45	0.51	0.24	0.33	0
16132–16138: DEL (4)	0	0	0	0	0	0.15	0.56	0.24	0.37	0.52
16132–16139: DEL (5)	0	0	0	0	0	0.15	0.56	0.24	0.37	0
16134–16135: CA/–	0	0.69	0.62	0.57	0.31	0.15	1.45	0.84	0.98	1.8
16137–16139: AAA/—	0	0.69	0.62	0.57	0.31	0.15	1.45	0.84	0.98	1.28
16139: A/G	0	0	0	0	0	0	0	0	0	0.7
16236: C/T	0	0	0	1.52	0	0	0	0	0	0
Total Coverage	2094	1163	2272	2117	1953	1988	2350	1679	2153	1726

TABLE 4—*HV1a expected variants remaining after bidirectional read confirmation. Variants highlighted in gray represent expected variants originally obtained for sample donors using Sanger sequencing. Only expected variants remain after applying a bidirectional read quality filter.*

Variant Position	Donor 001	Donor 005	0.5% 001	1% 001	2% 001	5% 001	95% 001	98% 001	99% 001	99.5% 001
16069: C/T	99.95	0	0.09	1.1	1.85	3.89	95.04	98.3	99.43	99.35
16093: T/C	97.97	0	0.09	1.15	1.8	3.95	93.78	96.78	97.86	97.63
16126: T/C	99.09	0	0.26	1.1	1.8	3.96	94.1	97.06	98.08	97.62
16129: G/A	0	97.51	98.28	97.53	96.92	94.59	4.49	1.56	0.28	0.7
Total Coverage	2094	1163	2272	2117	1953	1988	2350	1679	2153	1726

0.72% in an additional NGS run with an average depth of coverage of 1875 per library (data not shown). Additionally, with very few exceptions, the frequencies of unexpected variants were identical within each mixture in both forward and reverse reads. Further, none of the unexpected variants were detected at positions spanning the region shared between the overlapping HV1a and HV1b amplicons, and all negative controls lacked any analyzable sequence. Use of the AVA Global Alignment Consensus viewing option showed that all unexpected HV1b variants were consistently detected as a group within a small subset of the same reads, arguing against their characterization as background noise.

A Blastn search (18) of the rCRS HV1b sequence with unexpected variants against the NCBI nucleotide database revealed that the detected minor unexpected DNA sequence is identical to a previously reported nuclear insertion of mitochondrial DNA found on the short arm of chromosome 11 (NCBI accession #HE613849.1) (14). This insert contains the HV1b primer-binding sites, consistent with our finding of a co-amplification event (see Fig. 2).

Tissue Comparison Study—HV1a Data

The AVA software successfully parsed the read library into sublibraries based on the MID sequence and allowed for detection of all expected variants across all tissue types. No unexpected variants were detected in whole-blood-derived DNA samples, and a single unexpected variant was detected at rCRS position 16199 (frequency of 1.32%, standard deviation of 0.62 between frequencies of forward and reverse reads) in the buccal sample from donor 001-CF30. This position is not a known mutational hot spot. Several unexpected variants were also detected sporadically across five hair samples from donor 001-CF30 that were not removed from the data set with quality filtering. The average frequency of filtered, unexpected variants in hair samples was 1.53%, with an average standard deviation of 0.25 between forward and reverse reads. This low average standard deviation value confirms that there is little difference in the forward and reverse strand frequencies of these variants, suggesting that they warranted further study. It is possible that these variants are low-level intra-individual differences not previously characterized using Sanger methods.

Varying levels of heteroplasmy were detected across all tissue types at position 16093, a known mutational hot spot (19). Sanger reference data from whole blood samples for donor 001-CF30 show a C at this site with no evidence of heteroplasmy (data not shown). Pyrosequencing data show a range of variant frequencies from <25% to >99% for the reported transition, with similar frequencies across all tissue-specific forward and reverse strand reads. This is evidenced by an average standard deviation of forward and reverse strand frequencies of 0.46 at position 16093. It is possible that these variants reveal differences in the level of heteroplasmy between tissues and between samples of the same tissue originating from heteroplasmic mixtures within a single individual. Further studies are ongoing to elucidate the properties of these variants from other sources of variation.

A proportion of the variants appear as short indels, which cluster around homopolymeric stretches of 3 or more identical nucleotides. As expected, homopolymeric-associated variants tended to have large frequency disparities between forward and reverse reads, with some detected 100% in one direction and 0% in the other direction. This class of variants is expected due to the reported inability of pyrosequencing chemistry to sequence through homopolymeric regions accurately (3,4,20). Fortunately, many of these variants can easily be filtered from the data set by

TABLE 5—*Roche GS Junior 454 pyrosequencing mixture experiment: HV1b variants detected using Amplicon Variant Analyzer software. Mixtures were prepared in defined ratios by combining quantified HV1b amplification products from donors 003-CM54 and 015-AM30. Data are shown as the frequency of each variant detected per multiplex identifier (MID). Highlighted variants are those expected based on the Sanger reference data. Little difference is seen after applying the bidirectional read confirmation quality filter.*

Variant Position	Donor 003	Donor 015	0.5% 003	1% 003	2% 003	5% 003	95% 003	98% 003	99% 003	99.5% 003
16189: T/A	0.87	1.47	1.20	1.11	0.92	1.22	0.57	0.74	0.43	0.74
16193: C/T	97.91	0	0.92	1.14	2.18	6.74	94.40	96.34	95.87	97.47
16195: T/C	99.07	0	0.93	1.23	2.17	6.88	95.36	97.56	98.54	98.74
16218: C/T	1.03	1.48	1.41	1.21	1.07	1.18	0.44	0.70	0.51	0.63
16221: C/T	98.50	0	0.93	1.15	2.15	6.82	94.89	96.79	97.86	98.00
16223:T/C	1.54	99.72	98.93	98.69	97.91	95.55	8.02	3.24	2.43	2.36
16224:C/T	1.50	99.69	98.88	98.67	97.92	95.49	7.96	3.21	2.43	2.40
16230: A/G	1.05	1.54	1.49	1.29	1.13	1.34	0.47	0.76	0.55	0.73
16242: C/A	98.74	0	0.92	1.14	2.15	6.82	95.17	97.32	98.32	98.46
16249: T/C	1.08	1.64	1.57	1.26	1.13	1.40	0.52	0.72	0.47	0.73
16259:C/A	1.02	1.59	1.52	1.24	1.11	1.39	0.46	0.70	0.49	0.69
16263: T/C	1.00	1.64	1.53	1.31	1.14	1.44	0.49	0.72	0.49	0.71
16264: C/T	1.02	1.60	1.65	1.28	1.11	1.48	0.49	0.73	0.58	0.70
16270: T/C	1.68	99.90	99.40	99.01	98.22	95.84	8.25	3.27	2.52	2.57
16274: G/A	0	98.80	98.50	98.44	97.36	95.28	6.88	2.36	1.31	1.01
16278: C/T	1.05	1.62	1.68	1.34	1.23	1.44	0.50	0.78	0.55	0.73
16284: A/G	1.03	1.59	1.53	1.32	1.15	1.31	0.58	0.76	0.53	0.73
16288: T/C	0.54	1.06	1.01	0.83	0.63	0.89	0.33	0.43	0.06	0.54
16290: C/T	0.52	0.92	1.08	0.84	0.68	0.95	0.40	0.41	0.06	0.49
16293: A/C	0.60	1.12	1.27	0.85	0.77	1.00	0.31	0.46	0.06	0.51
16301: C/T	1.09	1.60	1.58	1.27	1.21	1.33	0.46	0.76	0.53	0.68
16311: T/C	1.03	1.64	1.63	1.39	1.24	1.44	0.51	0.84	0.66	0.80
16319: A/G	1.03	1.62	1.41	1.34	1.24	1.38	0.53	0.82	0.49	0.78
16352: C/T	95.95	1.64	2.35	2.33	3.27	7.86	91.24	93.57	97.72	95.13
16355: C/T	0.99	1.65	1.67	1.50	1.35	1.48	0.51	0.76	0.51	0.69
16356: T/C	1.03	1.67	1.54	1.37	1.30	1.43	0.50	0.76	0.60	0.73
16357: T/C	0.12	98.83	98.65	98.50	97.48	95.25	7.00	2.30	1.42	1.10
16368: T/C	1.08	1.56	1.53	1.38	1.26	1.38	0.53	0.95	0.53	0.83
16390: G/A	1.02	1.56	1.72	1.25	1.26	1.48	0.55	0.80	0.62	0.67
Total Coverage	6693	6609	9206	9064	8439	8343	8431	8337	4869	11,589

```
TTTTCTTTTGTTGATTGAGCAGCATTCCATTGTGGGAAAATACCAAATGCATGGAGAGCTCC
CGTGAGTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGGAACG
TGTGGGCTATTTAGACTTTATGGCCCTGAAGTAGGAACCAGATGTTGGATACAGTTCACTTT
AGCTACCCCCAAGTGTTATGGGCCCGGAGCGAGGAAAGTAGCACTCTTGTGCGGGATATTG
ATTTCACGGAGGATGGTGGCCAAGGGACTCCTATCTGAGGGGGGTCATCCGTGGGGACGAG
AGAGGATTTGACTGTAATGTGCTATGTACGGTAAATGGCTTTATGTGCTATGTACTATTAAG
GGGGGATGGGTCTGTTGATATTCTAGTGGGTAGGGGTTGGCTTTGGGGTTGCAGTTGATGTG
TGACAGTTGAGGGTTAATTGCTGTACTTGCTTGTAAGCATGGGGTGGGGGTTTGATGTGGA
TTGGGTTTTTATGTACTACAGGTGGTCAAGTATTTATGGTACTGTACAATATTCATGGTGGCT
GGCAGTAATGTACGAAATACTATGGATTGTTTATTCACTCTTCTGTTAGAAACC
```

FIG. 2—*Human mtDNA control region NumtS sequence reported by Thomas et al. (15). Highlighted regions at the 5′ and 3′ ends are nuclear specific and appear in both NumtS-positive and NumtS-negative fragments amplified with nuclear-specific primer sets (see also Fig. 5). The blue region represents the HV1b sequence amplified using human mtDNA-specific primers.*

applying the bidirectional read quality filter. Based on these criteria, these variants were classified as background noise.

Other uncharacterized variants were also detected with frequencies ≤0.26%. These variants potentially arise as a result of PCR-induced base misincorporations, 454 sequencing chemistry artifacts, or random sampling effects. A default minimum read percentage of 0.26% can be set within the AVA software to filter any variants with frequencies below this threshold. Further analysis is currently being conducted in our laboratory to elucidate the cause(s) of these variants and the appropriate use of such filters.

Tissue Comparison Study—HV1b Data

The same set of NumtS variants detected in HV1b mixture data was also detected in the HV1b amplicons derived from nuclear DNA-rich blood and buccal extracts in the tissue comparison study, but not in hair shaft extracts. All NumtS variants were detected consistently in whole-blood-derived DNA extracts with an average frequency of 0.66% and an average standard deviation of 0.23 between forward and reverse strand reads. However, NumtS variants were only detected in buccal tissue samples after applying modified AVA analysis parameters where the minimum read percentage was changed from the default value of 0.25% to 0%. It is likely that this discrepancy is due to the amount of template DNA originally amplified (5–20 ng of input DNA from whole blood on FTA® cards vs. 1 ng of input DNA from buccal extracts). Supporting this idea is the observation that high input nuclear DNA template concentrations were required for NumtS amplification and subsequent Sanger sequencing (24–147 ng of input template DNA). The presence of the NumtS-associated variants was confirmed in whole blood

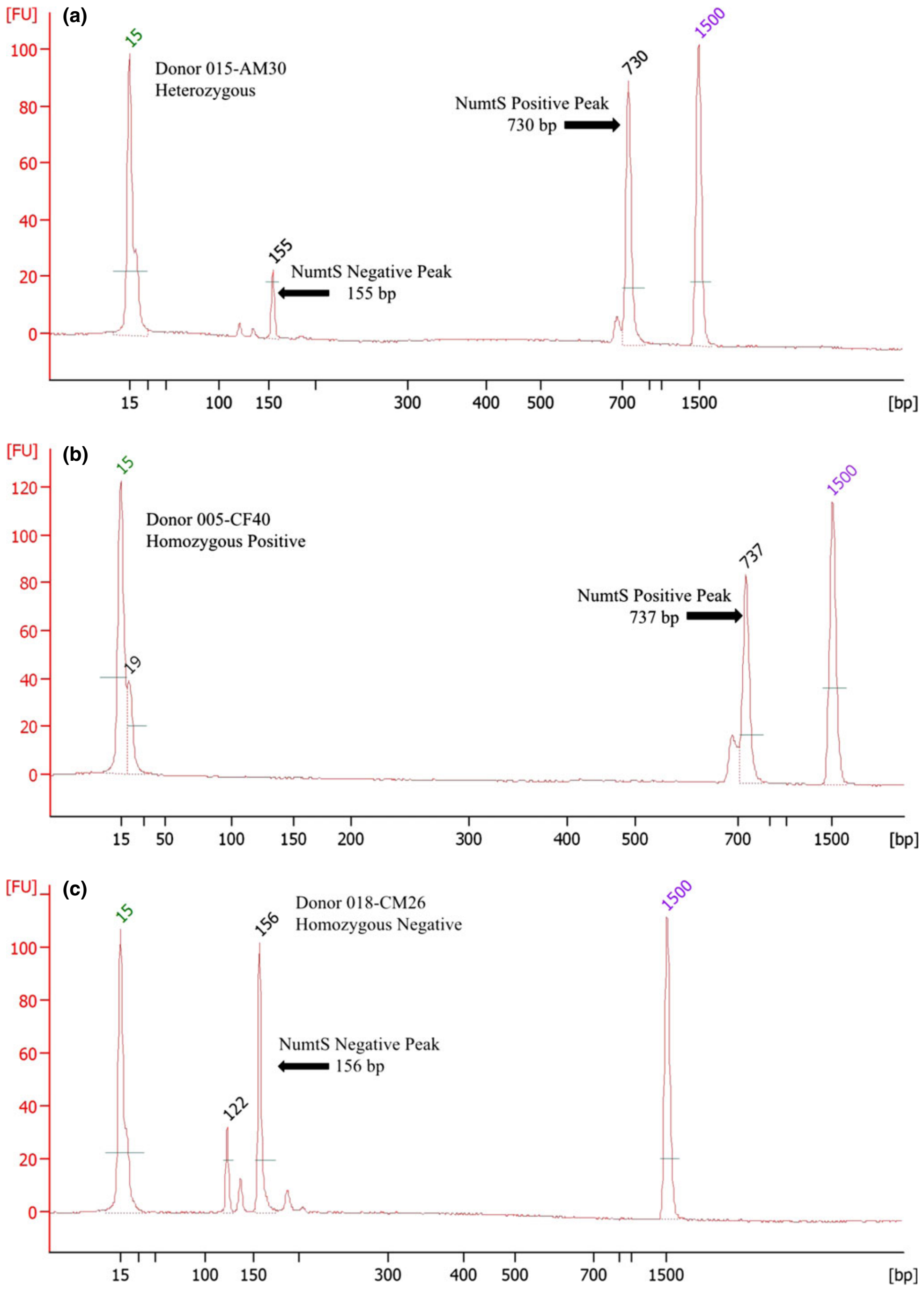

FIG. 3—*Agilent 2100 Bioanalyzer data for NumtS amplification from donors 005-CF40, 015-AM30, and 018-CM26, respectively. Each Bioanalyzer trace shows an upper and lower molecular weight marker (15 and 1500 bp). (a) shows the amplification products from donor 015-AM30, a heterozygous individual for the NumtS insertion. In this trace, a peak is detected at 155 bp (NumtS-negative allele) and a peak at 737 bp (NumtS-positive allele). (b) shows the amplification product from a homozygous-positive donor 005-CF40, a single peak at 737 bp. (c) shows the amplification results from a homozygous-negative donor 018-CM26, resulting in an amplification product at 156 bp. The Agilent Bioanalyzer reports approximate sizes of amplicons in bp.*

and buccal tissues by viewing the data in the Global Alignment Consensus tab of the AVA software to verify that they were read-clustered.

Four SNP variants were detected at frequencies between 1% and 2% from hair samples at positions consistent with the NumtS insertion, even after employing the lower stringency AVA analysis parameters. However, not all of the NumtS variants appear. This is expected due to the overall low amount of nuclear DNA in hair shaft extracts, and hence, the appearance of some of these variants detected from hair samples may be due to stochastic sampling effects from nuclear-encoded templates present in some hair shaft extracts.

In addition to the expected variants, several unexpected variants were sporadically detected within hair shaft-derived DNA samples in the HV1b data set. Many of these sporadic variants were filtered from the data set by applying the bidirectional read quality filter or by increasing the minimum read percentage to 0.26% with the expectation that a majority of the NumtS-associated variants would also be removed from the buccal tissue data set.

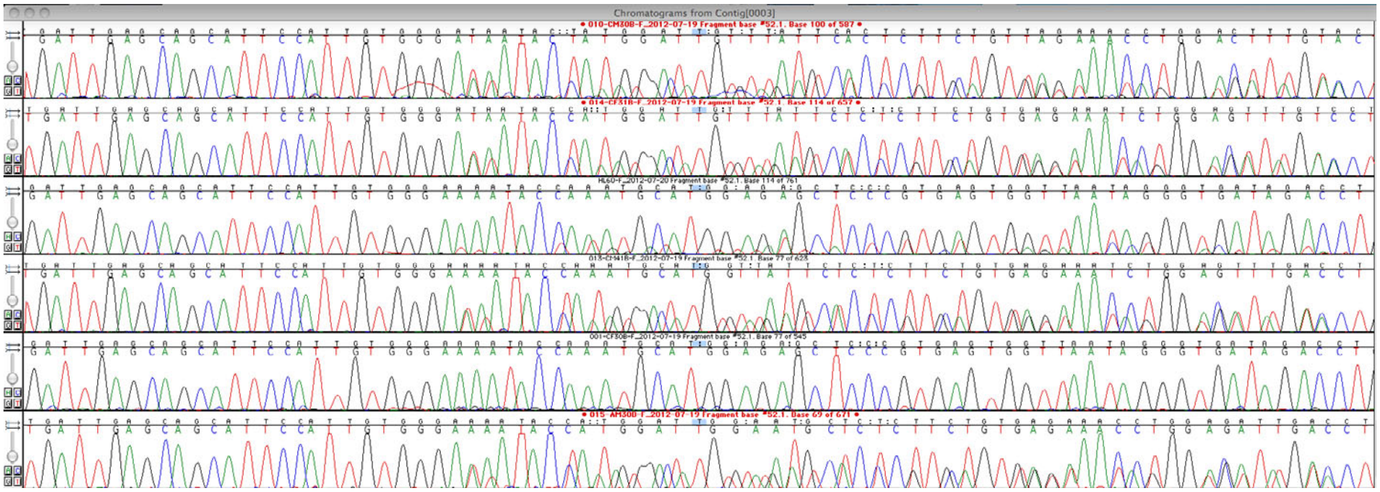

FIG. 4—*Electropherograms showing Sanger sequence data for NumtS heterozygous donors. Resulting sequences were aligned against the full NumtS insertion. Data are out of phase in dissimilar areas between NumtS-positive and NumtS-negative sequences. Sample 001-CF30 (fifth sequence from the top) is homozygous positive for the insertion, with no out-of-phase data seen.*

TTGTATGTATTAGTTTTTTTCTTTTGTTGATTGAGCAGCATTCCATTGTGGGATTATACTATGG
ATTGTTTATTCACTCTTCTGTTAGAAACCTGGACTTTGTA

FIG. 5—*NumtS-negative sequence.*

NumtS-Specific Amplification and Sanger Sequencing

Of the twenty-two donor and control samples amplified, Agilent 2100 Bioanalyzer amplicon analysis revealed that thirteen donors were heterozygous for the NumtS insertion, eight were homozygous positive, and one was homozygous negative. Bioanalyzer traces illustrating this data are found in Fig. 3. Those donors whose sample extracts were used for the deep-sequencing mixture analysis and tissue comparison studies (001-CM30, 003-CM54, and 015-AM30B) all possessed the NumtS insertion on at least one chromosome. This evidence further supports the conclusion that the set of variants obtained with HV1b deep-sequencing is due to the expected co-amplification of the NumtS nuclear insertion and mtDNA using this particular primer set, previously reported to be part of the inserted NumtS sequence (13–15).

The sequence variants observed using the mtDNA HV1b primer set are those from the reported sequence of the NumtS insertion, as well as the expected variants from HV1b mtDNA amplicons. Conversely, no mtDNA-specific polymorphisms were detected in any of the donor sequences amplified with the nuclear-specific primer set, confirming the hypothesis that the source of these variants is the NumtS insertion rather than mtDNA.

Sanger sequence data obtained using the mtDNA HV1b-specific primers align to the rCRS for all samples and controls except in the case of individual 018–26 M. This individual is homozygous negative for the NumtS insertion, and thus, no sequence data aligning to the rCRS are expected to result. The sequence data for homozygous-positive individuals obtained using NumtS-specific primers (Fig. 2) are identical to the NumtS sequence reported by Lang et al. (14). Individuals heterozygous for the insertion show out-of-phase sequence data after regions of sequence similarity between insertion-positive and insertion-negative amplicons (Fig. 4). This pattern was observed for all thirteen heterozygous donors. Sequence data for homozygous individuals did not exhibit regions in the electropherogram that appeared out of phase. Alignment of the 5′ and 3′ ends of both fragments in the NumtS-negative and NumtS-positive samples reveals that these regions are of nuclear origin and not part of a mitochondrial DNA insertion as reported previously (Figs. 2 and 5) (14).

Discussion

Targeted deep-sequencing of the human mtDNA hypervariable region allows for an increase in the depth of mtDNA analysis and hence enhances the ability to detect minor variants present in the sequence. The goal of forensic validation studies is to fully understand all the potential sources of variation, including those that arise from true genetic differences in the sample and those that are artifacts from the technique or chemistry, so that a proper interpretation can be made in a forensic case. *This technical note represents preliminary research findings related to the evaluation of NGS technologies for use in forensic crime laboratories. Additional research is required to identify sources of platform-specific background noise and to evaluate the sensitivity and data quality associated with these instruments. These issues are complex, because they are dependent not only on platform-specific sequencing chemistry, but also on bioinformatic processing of the data.* We have shown that minor SNP variants can be reproducibly and accurately detected at a level of 1% or lower in multiple deep-sequencing runs using the Roche GS Junior 454 pyrosequencing instrument (Roche Applied Science). Research is ongoing in our laboratory to identify a dynamic threshold window at which minor variants can be reliably detected above background noise. In addition to expected minor variants, a subset of read-clustered, unexpected variants was also detected in HV1b reads originating from nuclear DNA-rich blood and buccal cell samples. These data show that these variants are amplified products from a nuclear pseudogene, an ancient nuclear insertion of mitochondrial DNA that contains HV1b control region primer-binding sites. These NumtS sequences, to our knowledge, have never been co-detected in mitochondrial sample preparations using Sanger sequencing, further supporting the assertion that the sensitivity of low-level variant detection reported herein is a vast improvement over current methodologies. Additionally, as expected by the relative amounts of nuclear and mitochondrial DNA found in different tissues, the data show slight differences in the amounts at which the unexpected variants are detected in whole blood, buccal, and hair samples. Because amplification of the NumtS requires a relatively high amount of input template, this difference is likely due mostly to the tissue-type-dependent amount of the NumtS template present in the sample. From a forensic perspective, the presence or absence of the NumtS sequence in a sample should accordingly vary with a host of factors, including the amount of sample DNA amplified and the individual's genotype with respect to the NumtS itself and hence not be expected *a priori* to be present in each analysis. Consequently, until fuller characterization of the genetic variation within each NumtS is elucidated, we recommend that no direct comparisons be made of the NumtS insertion itself.

As expected, deep-sequencing of HV1b from hair shaft samples does not give rise to the read-clustered, unexpected variants observed in blood and buccal tissue. We attribute this observation to the known scarcity of nuclear DNA present in hair shaft DNA extracts. However, additional inconsistent, unexpected variants were detected in all control region amplicons from hair samples that were not also found in blood and buccal tissue from the same individual. Some of these low-level variants may represent intra-individual variation. Variants may in many cases arise as a result of 454 pyrosequencing-specific chemistry and may be omitted from the data set with increased stringency quality filter settings. Additional research is needed to elucidate the cause(s) of these low-level variants.

Heteroplasmy is readily detected using NGS, because individual template molecules are interrogated independently of one another. This is observed at position 16093 in HV1a data from the tissue comparison study. As a result, crime laboratories wishing to implement next-generation sequencing into their mtDNA analysis workflow should reevaluate interpretational criteria accordingly. It should be noted that all of the single-donor blood, buccal, and hair shaft samples amplified and sequenced in this study would be interpreted as an inclusion using the mtDNA interpretation guidelines currently in place in forensic laboratories. All of these samples share the same predominant DNA sequence. Hence, we anticipate no significant interpretational complications due to the co-detection of NumtS sequences in forensic casework. Further, mtDNA evidentiary samples often lack or contain very low amounts or highly degraded nuclear DNA and hence are not expected to reveal any NumtS-specific variants, as we have shown with hair shaft samples. Regardless, the NumtS insertion sequence appears to be highly conserved in most cases, and hence, the ability to detect the NumtS insertions can be included in further validation studies of NGS involving mtDNA. Additionally, if warranted in a particular context and sample type, we have shown that confirmation of NumtS co-detection can be accomplished with nuclear DNA-specific amplification of a suspected NumtS insertion. Future studies will include analysis of additional forensically relevant sample types including bone materials, and blood and buccal samples exposed to environmental insults.

It has been reported that much of the mtDNA genome has been inserted as small fragments into various positions within the nuclear genome, creating a mitochondrial pseudogenome (21). Based on these findings, it is probable that expansion of mtDNA analysis beyond the control region using NGS will result in further co-detection of additional NumtS-specific variants, particularly from nuclear DNA-rich reference samples. Several publications have reported incorrect characterization of co-amplified NumtS as heteroplasmy (21). Careful characterization of NumtS that are co-detected with whole mtDNA genome data will result in better identification and resolution of true mtDNA heteroplasmic mixtures and hence provide a further conceptual foundation to continue to correctly interpret mtDNA comparisons in the future.

Acknowledgments

The authors wish to thank Erin S. Burnside, Brandon C. Smith, and Hilde Stawski for their contributions to this work as well as two anonymous reviewers from the *JFS*.

References

1. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet 2010;11(1):31–46.
2. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24(3):133–41.
3. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 2007;8(7):R143.
4. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, Martin J-F. Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 2011;12(1):245.
5. Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, et al. Mitochondrial DNA regions HVI and HVII population data. Forensic Sci Int 1999;103(1):23–35.
6. Budowle B, Allard MW, Wilson MR, Chakraborty R. Forensics and mitochondrial DNA: applications, debates, and foundations. Annu Rev Genomics Hum Genet 2003;4:119–41.
7. Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, et al. DNA commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 2000;110 (2):79–85.
8. Wilson MR, DiZinno JA, Polanskey D, Replogle J, Budowle B. Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med 1995;108(2):68–74.
9. Wilson MR, Polanskey D, Replogle J, DiZinno JA, Budowle B. A family exhibiting heteroplasmy in the human mitochondrial DNA control region reveals both somatic mosaicism and pronounced segregation of mitotypes. Hum Genet 1997;100(2):167–71.
10. Wilson MR, DiZinno JA, Polanskey D, Budowle B, editors. Assessing heteroplasmy in the control region of human mitochondrial DNA. Proceedings of the Meeting of the American Academy of Forensic Sciences; 1998 Feb 9–14; San Francisco, CA: Colorado Springs, CO: McCormick-Armstrong, 1998.
11. Wooley JC, Ye Y. Metagenomics: facts and artifacts, and computational challenges. J Comput Sci Technol 2009;25(1):71–81.
12. Salas A, Lareu MV, Carracedo A. Heteroplasmy in mtDNA and the weight of evidence in forensic mtDNA analysis: a case report. Int J Legal Med 2001;114(3):186–90.
13. Zischler H, Geisert H, von Haeseler A, Pääbo S. A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. Nature 1995;378(6556):489–92.
14. Lang M, Sazzini M, Calabrese FM, Simone D, Boattini A, Romeo G, et al. Polymorphic NumtS trace human population relationships. Hum Genet 2012;131(5):757–71.
15. Thomas R, Zischler H, Pääbo S, Stoneking M. Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. Hum Biol 1996;68(6):847–54.
16. Kavlick MF, Lawrence HS, Merritt RT, Fisher C, Isenberg A, Robertson JM, et al. Quantification of human mitochondrial DNA using synthesized DNA standards. J Forensic Sci 2011;56(6):1457–63.

17. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. Nature 1981;290(5806):457–65.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10.
19. Stoneking M. Hypervariable sites in the mtDNA control region are mutational hotspots. Am J Hum Genet 2000;67(4):1029–32.
20. Balzer S, Malde K, Jonassen I. Systematic exploration of error sources in pyrosequencing flowgram data. Bioinformatics 2011;27(13):i304–9.
21. Parr RL, Maki J, Reguly B, Dakubo GD, Aguirre A, Wittock R, et al. The pseudomitochondiral genome influences mistakes in heteroplasmy interpretation. BMC Genomics 2006;7:185.

Additional information and reprint requests:
Brittania J. Bintz, M.S.
Forensic Research Scientist
Western Carolina University
111 Memorial Drive, NSB #231
Cullowhee, NC 28723
U.S.A.
E-mail: bbintz@wcu.edu