## 9.1 PURPOSE:

9.1.1 To provide a general guideline for the analysis and interpretation of mtDNA profiles. However, it is not an exhaustive list of all possible casework scenarios. Scientist's training, experience and judgment is considered when conducting mtDNA analysis. Individual case results may require specific interpretational strategies where directed by the TL.

## 9.2 RESPONSIBILITY:

9.2.1 Forensic Science Examiners from the CT DESPP Division of Scientific Services who have been trained in the discipline of Mitochondrial DNA Analysis according to the Mitochondrial DNA Section Training Manual.

## 9.3 Data Analysis

9.3.1 After sample processing is completed, the sequence data is analyzed. Both strands (one from the forward primer and one from the reverse primer) must be used to confirm sequence edits whenever possible. If only one strand can be sequenced, two independent amplifications of that strand can be used to confirm the sequence edits. Single-stranded sequence in samples or RB/NCs can be used, but can only be edited toward the most conservative edit call (N, Y, or R) in the appropriate context.

9.3.2 The steps outlined below describe the general steps involved in analyzing sequencing runs. The user's manual for Sequence Analysis™ contains details on using the software that may be used by the analyst.

9.3.3 Launch Sequencing Analysis version 5.2 by double-clicking on the sample file within the run folder or from the Applied Biosystems menu on the Start menu.

9.3.4 Click on File and select 'Add Samples'. Open the desired run folder and choose the appropriate sample files to add to the Sample Manager window. Click 'Add Selected Samples' and 'OK'.

9.3.5 To view the electropherograms, double-click on the sample file name, or check the 'Show' box.

9.3.6 For each sample, click on the 5th button (raw data button) to display raw data. Signal intensity may be reviewed here. Note the beginning and ending scan numbers by moving the cursor to the beginning and end of the data peaks. (The scan numbers should be similar for all

capillaries within a run.)

9.3.7 In Sample Manager enter the beginning scan number in the 'Peak 1' column for all the sample files (this value will automatically fill in the 'Start' column as well). This assures that the first base off the primer is included.

9.3.8 Also in Sample Manager, enter the ending scan number in the 'Stop' column for all the sample files. This will
decrease the number of N's at the end of the run and focus the signal intensity calculation on the appropriate section.

Note: A single Peak 1 and Start value, as well as a single Stop value can be used for all files, as long as these numbers encompass all sequence data in each file.

9.3.9 Ensure the 'BC' (basecalling) column is checked and the 'P' (print) and 'PP' (post processing) columns are not checked.

9.3.10 Click on the Start button.

9.3.11 When analysis is complete, save the samples and close the program.

## 9.4 Identification of Polymorphic Sites Using Sequencher

9.4.1 Launch Gene Codes Sequencher™ software (version 4.1.4Fb19 for Windows).

9.4.2 Open New Project under the File pull-down menu.

9.4.3 Highlight the data files you wish to import and drag them into the project window. Data can also be imported into the project window by clicking on File, scrolling down to Import and selecting Sequences. Find the data you want to import, select it and click 'open'. Once all the data is imported into the Sequencher project, open each sequence individually and delete unanalyzable sequence from the ends. Note which samples are unanalyzable (typically this includes RBs and NCs).

9.4.4 Highlight analyzable forward and reverse sequences for assembly and click on "Assemble Automatically". For analyzable RBs and NCs, the default assembly parameters (85 and 20) may be lowered to 75 and 15 in order to get the sequences to assemble.

9.4.5 View the contig produced. Make sure the direction of the sequence files is appropriate. If the sequence files are not in the appropriate direction, go to View Menu and select "Reverse Comp".

9.4.6 Click on the Bases box to observe the nucleotide sequence, highlight bases in the consensus sequence and click on "Show Chromatograms" to observe the electropherograms.

9.4.7 If the sequence is not of the requisite quality, (for example: high background, poor peak resolution, or indicative of a mixture) the sequence may not be used. The sample may be re-extracted, re-amplified, recycle sequenced, or re-injected as appropriate.

9.4.8 Edit the sequences using information from all strands in the electropherogram window. Each base in the sequence must be manually reviewed for correct identification using the electropherogram windows. It is possible that ambiguous bases will remain after the sequence is fully edited. These ambiguities can be designated according to International Union of Pure and Applied Chemistry (IUPAC) nomenclature.

9.4.9 When the editing is completed, import the revised Cambridge Reference Sequence (rCRS) into the project. Highlight rCRS and the CONTIG(s) relevant to the sample (for example: HV1 and HV2) in the project window and click on "Assemble Automatically". A new contig will be created which will consist of the sample's consensus sequence and the rCRS. The contig name should include the case #, item#, and sample designation.

9.4.10 Differences from rCRS can be viewed by either of the following methods: highlight rCRS, go to Sequence →Compare To →Consensus. Alternatively, highlight the contig, go to Contig → Difference Review. Click on the 'export as text' button that appears in the contig difference review window to save it as a text file . Print out the Sequencher project window in its expanded form. Note the sequence ranges (double stranded data) of both the HL60 and the sample. The shortest of these two ranges shall serve as the range for both.

9.4.11 Open the text file and add the following information: laboratory number, submission/item number, analyst's initials, date, and the sequence ranges. Other information may need to be added according to the sample. Print the file. Use the text files to compare and interpret differences among samples.

## 9.5 Sequence Nomenclature

9.5.1 Polymorphisms are noted on the text file printouts. They are noted by stating the nucleotide position followed by the code for the polymorphic base (for example, 263 G). Nucleotide base positions can be designated according to the standard nomenclature set by the IUPAC.

| CODE | DESIGNATION |
|---|---|
| A | Adenosine |
| B | C, G, or T |
| C | Cytosine |
| D | A, G, or T |
| G | Guanosine |
| H | A, C, or T |
| K | G or T (Keto) |
| M | A or C (aMino) |
| N | aNy base |
| R | A or G (puRine) |
| S | G or C |
| T | Thymine |
| V | A, C, or G |
| W | A or T |
| Y | C or T (pYrimidine) |

| OTHER | DESIGNATION |
|---|---|
| - | Deletion |
| X.1(.2,...) | Insertion after position X |



9.5.2 Profiles are characterized so that the least number of differences from rCRS are present. Insertions are listed to the right of a particular nucleotide position to minimize the number of differences from the rCRS. Insertions are notated by first noting the site immediately 5' to the insertion followed by a point and a 1 (for the first insertion), a 2 (if there is a second insertion), and so on, and then by the nucleotide that is inserted (for example, 315.1 C).

If there is more than one way to maintain the same number of differences with respect to the rCRS, differences should be prioritized as follows: (i) substitutions: transitions are favored over transversions (ii) insertions/deletions (indels).

9.5.3 Length heteroplasmy in HV1 most commonly arises when there is a substitution of a C for a T at position 16189. The reference type in HV1 is $C_5TC_4$. HV1 length heteroplasmy will not be recorded in casework samples. Rather, sequences will be truncated to fit the $C_5TC_4$ format.

9.5.4 Length variants in HV2 are commonly observed in the number of C residues preceding a T residue at position 310. The reference type in HV2 is $C_7TC_5$. A 309.1 C insertion in HV2 will be annotated $C_8TC_5$, and the insertions 309.1 C, 315.1 C will be annotated as $C_8TC_6$. It is possible to determine the dominant length variant in this region, as well as some or all of the minor length variants. The sequence can be notated to reflect these variants. Where a

sequence does not have a T at 310 and length heteroplasmy is present, it is notated 310 C and no insertions or deletions relative to rCRS at 303-315 are noted. If a sequence does not have a T at 310 and length heteroplasmy is not present, then the guidance described above can be used.

9.5.5 Additional sequence nomenclature guidance can be found in the SWGDAM Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories (dated 07/18/2013) and the SWGDAM Nomenclature Examples Document (dated 08/07/2013).

## 9.6 Evaluation of Control Sequences

9.6.1 The positive control that is run with the sample should type correctly in order to report the sequence for that sample.

9.6.2 After RB and NC sequence data are analyzed, a determination is made as to whether any analyzable sequence is present. If there is not, proceed to "Comparing Sequences" below. If there is, the sequence obtained from that sample must first be compared to any sequence obtained from the RB and NC associated with that sample before comparing it to another sample in the case. The RB and NC shall have at least one base pair difference from the relevant sample in each region (HV1A, HV1B, HV2A, HV2B, or HV1, HV2) or that region will not be further used. Refer to mtDNA SOP-01 section 1.8.

## 9.7 Mixtures

9.7.1 A mixture is defined as 3 or more sites of point heteroplasmy. Mixture samples will not be interpreted for comparison purposes. A notation will be made within the case workbook and an evaluation of the controls will be performed. An analysis packet (Difference Files for the sample, positive control, and any analyzable RB's/NC's as well as the Sequencher project) will not be printed out for mixture samples, nor will second analysis by another qualified mtDNA analyst occur. If possible, the sample will be re-extracted (for both HV1 and HV2) in an effort to obtain an interpretable profile (see DNA SOP-1 for Evidence Consumption policy). No further analysis is conducted if the sample remains a mixture or re-extraction cannot occur.

## 9.8 Comparing Sequences

9.8.1 When comparing sequences obtained from samples, only the regions with a common range will be evaluated. For example, if a partial sequence (16024-16365 and 73-284) is obtained for the evidentiary sample and a full sequence is obtained for the known sample, the comparison will be conducted on positions 16024-16365 and 73-284. In addition,

sequence before and after the defined HV1 and HV2 regions will also be used for comparison purposes, provided this range is common to both samples.

## 9.9 Comparing Sequences with Length Heteroplasmy

9.9.1 Length heteroplasmy in HV1 most commonly arises when there is a substitution of a C for a T at position 16189. The reference type in HV1 is $C_5TC_4$. HV1 length heteroplasmy will not be recorded in casework samples. Rather, sequences will be truncated to fit the $C_5TC_4$ format.

9.9.2 The number of C nucleotides exhibited in samples with HV2 length heteroplasmy is highly variable and care must be exercised when making comparisons. In order for sequence concordance to be declared, a common length variant must be observed in the samples being compared. For instance, if a questioned sample has predominately 7 C residues preceding the T at position 310, but shows the presence of 8 Cs as well, it is concordant with samples containing any mixture of length variants of 7 or 8 Cs.

9.9.3 Differences in sequences due to the absence of a common length variant are exclusionary in combination with a base difference outside of the 303 – 310 region. For instance, if a questioned sample shows only the presence of 7 Cs and a known sample shows 9 Cs, no common length variant is found between the two samples and a difference at position 309 is noted. This difference should be considered in the context of the entire comparison and the appropriate conclusion reported.

9.9.4 Where a sequence does not have a T at 310 and length heteroplasmy is present, it is notated 310 C and no insertions or deletions relative to rCRS at 303-315 are noted. If a sequence does not have a T at 310 and length heteroplasmy is not present, then the guidance described in section 9.5.4 can be used.

## 9.10 Interpretation of Sequence Comparisons

9.10.1 The following interpretations are available for sequence comparisons:

9.10.1.1 Exclusion:
If the samples differ at two or more positions, they can be excluded as coming from the same source.

9.10.1.2 Inconclusive:
If the samples differ at a single position, with no evidence of heteroplasmy (including non-overlap of length variants), the comparison should be reported as inconclusive. Because of the possibility of undetected heteroplasmy, additional

samples may be analyzed when two sequences differ by a single base. These samples may include blood, buccal swabs, and hair. Hair fragments from a known hair standard may be combined and processed as a single known sample.

9.10.1.3 Cannot Exclude:
If samples demonstrate sequence concordance, they cannot be excluded as coming from the same source. Sequence concordance is defined as having a common base, or HV2 length variant, at each position. For example, if one sample has evidence of a C and a T at a given position, and the other has a C, they share the C in common at that position and are concordant. Similarly, overlap of length variants is considered sequence concordance. However, if one sample has evidence of a C and a T at a given position, and the other has a G, these sequences are not concordant.

## 9.11 Sequence Confirmation

9.11.1 A second qualified analyst must confirm all non-mixture sequences regardless of interpretation or destination. Confirmation involves independently assembling sequences for the NC, RB, positive control, and evidentiary samples from the analysis data. Confirmed sequence range is defined as the shortest length of sequence obtained by two analysts.

## 9.12 Databases

9.12.1 The CODIS 7.0 database, containing 10,629 individuals, is used to determined match probability estimates. The database represents a broad geographic sampling from across the United States and covers the entire control region (16024-576). All profiles reflect the Scientific Working Group on DNA Analysis Methods (SWGDAM) mtDNA Interpretation Guidelines document dated 07/18/2013, which includes standards for sequence nomenclature. The database is maintained by the FBI Laboratory in collaboration with SWGDAM and is updated periodically.

## 9.13 Searching Profiles

9.13.1 The CODIS 7.0 database, containing 10,629 individuals, is used to search casework profiles in the forensic database. The range of the profile to be searched consists of the shortest range of sequence agreed upon by the analysts reviewing the data. Ambiguous base positions in both database samples and case work profiles are searched as an N (meaning the search includes all four bases at that position) and do not have any exclusionary value when compared with other sequences. Deletions are searched as a (-). Those database samples whose range(s) are fully included within the casework profile range are used in the database search result. Presently, this includes the regions 16024-16365 (HV1) and 73-340 (HV2).

The database search result provides the number of database profiles that match the casework profile, as well as the number of profiles in the database that differ by up to five positions.

9.13.2 The number of C residues in samples with HV1 length heteroplasmy around positions 16183-16194 (see section 9.8) are not considered for comparison purposes.

9.13.3 All sequence polymorphisms in the sample searched are entered; however, all length variants at nucleotide positions 16193 and 309 are ignored in database searches of relevant concordant sequences. Hence, length variability in this region will not add any additional rarity to a database profile search.

## 9.14 Upper Bound Frequency Estimate

9.14.1 An upper bound 95% confidence interval can be calculated from the results of a database search in order to estimate the population frequency of a profile. The upper bound estimate is dependent on the size of the database, and these estimates may change as the database size changes.

9.14.2 The following formulae may be used to calculate the upper bound estimate:

9.14.2.1 In cases where the profile has been observed in a database:

$$p + 1.96\,[(p)\,(1-p)/N]^{1/2}$$

$p = x/N$, where x is the number of times a profile has been observed in a population and N is the number of profiles in that population.

9.14.2.2 In cases where the profile has not been observed in a database:

$$1-\alpha^{1/N}$$

$\alpha$ is the confidence coefficient (0.05 for a 95% confidence interval), and N is the number of individuals in the population.

9.14.3 The results of these upper bound frequency estimate calculations for the three major population groups in the United States (African-American, Caucasian, and Hispanic) are included in the report. The calculations for other groups, whose total number in the database exceeds 100 people, may be included in the report where appropriate.

9.14.4 See mtDNA WI-11 and WI-12 for further instructions on analysis.

## 9.15 References

Scientific Working Group on DNA Analysis Methods (SWGDAM). Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic Testing Laboratories (2013) [Online]. Available: http://www.swgdam.org/SWGDAM%20mtDNA_Interpretation_Guidelines_APPROVED_073013.pdf

Scientific Working Group on DNA Analysis Methods (SWGDAM). Mitochondrial DNA Nomenclature Examples Document (2013) [Online]. Available: http://swgdam.org/SWGDAM_Mitochondrial_DNA_Nomenclature_Examples_FINAL-080713.pdf