Statistical Analysis for Massively Parallel Sequencing		
Status: Published		Document ID: 51242
DATE EFFECTIVE 09/08/2023	APPROVED BY mtDNA Technical Leader	PAGE 1 OF 4

Statistical Analysis for Massively Parallel Sequencing

1 Overview

1.1 DNA statistical analysis is done on a given comparison when the two samples cannot be excluded as originating from the same person or from having a maternal relationship. These statistics will be reported, and the generated paperwork will be included in the case file.

1.2 The extent of the sequence data that will be used for the database search and statistical analysis will be limited to the shortest range and most conservative reporting of the sequence in common between the two samples used in the comparison (see previously discussed sequence reporting criteria).

1.3 Statistics may be presented for sequence comparisons made for body identification cases. Statistics may also be presented when comparing evidentiary samples. In each instance the statistical analysis will be limited to the shortest range and most conservative reporting (e.g., when sequence heteroplasmy is present in one of the samples) of the sequence in common between the samples being compared.

2 Database

2.1 The database used to obtain a frequency estimate is maintained by the FBI and is available to CODIS users. This data set contains Whole Control Region (WCR) data for over 10,000 mitochondrial DNA sequences and was created in collaboration between the Armed Forces DNA Identification Laboratory (AFDIL), European DNA Profiling group (EDNAP), and the FBI Laboratory.

2.2 A copy of the database including the search window is found on dedicated CODIS computers.

3 Searching Profiles

3.1 The base pair range of the profile to be searched is limited to the shortest range of reported sequence in common for both compared samples (see previously discussed reporting criteria).

3.2 **Default Settings.** Under the **Tools** menu select **Options** and then **Popstats/mtDNA** to view the mtDNA default options. Check to make sure that the default settings are the following:

3.2.1 Population groups to search window located on the top right: all groups are checked

3.2.2 Confidence Interval (95%) Method: Clopper-Pearson One-Tailed is checked

3.2.3 Maximum number of sequence differences for match: 3

3.2.4 Maximum number of sequence differences for display: 0

3.2.5 Minimum number of overlapping base pairs for search: 90

3.2.6 Minimum pop. group size for upper bound frequency estimate: 150

3.2.7 List the match pairs: optional

3.2.8 Length Heteroplasmy window located on the right: 309, 573, 16193

3.2.9 Length heteroplasmy option: Ignore insertions at the specified length heteroplasmy sites is selected

3.3 **Profile Entry.** The mtDNA profile can be either retrieved from the database or entered manually. To retrieve a profile, in the main window on the left side select the **Popstats** tab. Enter the specimen ID number and then click on **Retrieve** and proceed with step **3.3.6**(see below). To enter a profile, click on the **mtDNA** tab, located below the specimen details field and continue as follows:

Note: Do not press Save at any point during profile entry. If the profile is saved, it will then need to be expunged from CODIS.

3.3.1 Enter the specimen identifier (e.g., case # - item description) in the **Specimen ID** field at the top of the window.

3.3.2 All of the other fields in the top portion of the Popstats Forensic Single Sample window can be left blank.

3.3.3 To begin entering a profile at the HVI, HVII, and/or HVIII regions, first click on the "+" icon located at the bottom of the window. This will insert the first data entry row in the mtDNA section.

3.3.4 To enter specific region(s) at HVI, HVII and/or HVIII, manually enter the start and end positions of the region(s) being searched. Alternatively, right click on the "I" icon to the left of the data entry row to select a region from the list.

3.3.5 Proceed to enter each difference from the rCRS by clicking on the "+" icon to add a row followed by the position number in one column and then the base change in the next column.

3.3.5.1 Locations exhibiting dropout should be entered as N.Repeat the sequence position and base entries on the same row for quality control purposes (the initial position number and base change entries will be blacked-out immediately after entry). Move from one column to the next by pressing the **Enter** key on the keyboard.

3.3.6 Click the **Calculate** button for the program to perform the necessary mtDNA popstats calculation.

3.3.7 Print the result.

3.4 The search result consists of the number of samples with 0, 1, or >1 mismatches to the searched sample in the combined database and divided into different ethnic groups. It will also calculate the upper bound frequency estimate based on the number of matches for each ethnic group.

3.5 Scan the resulting paperwork for attachment to the LIMS case file and attach this paperwork into the appropriate case file.

ATTENTION:

When sequence heteroplasmy is present at a given position in the mtDNA sequence, choose the appropriate IUPAC code during the data entry to be used for the search. (Please see document "Interpretation Guidelines for Mitochondrial DNA Massively Parallel Sequencing (mitoMPS)", section 5.1.1.1 for a table of current IUPAC codes.)

Even though mtDNA sequence of polycytosine length variants for regions 302-310 (HVII), and 568-573 (HVIII) are entered, multiple C-stretch length variants at these positions are ignored during the database searches of concordant sequences containing this region and therefore will not add any additional rarity.

The number of "C" nucleotides in samples with HVI length heteroplasmy is not considered for comparison purposes.

4 Frequency estimate

4.1 Frequency estimate when the mtDNA sequence is observed at least once in database.

Frequency estimates for the occurrence of a given mtDNA profile in the general population are determined using the Clopper-Pearson formula:

$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \alpha$$

where n = the number of profiles in the population database, x = the number of observations of the mtDNA profile in the database, k = 0, 1, 2, 3 … x observations, p = the 95% confidence limit for the proportion of a population with the mtDNA profile of interest, and □ = 0.05 for a 95% confidence limit.

The upper bound frequency estimate is calculated using the CODIS mtDNA Popstats program.

Example #1: A mtDNA sequence is observed 2 times in the US Caucasian database containing 2,609 sample profiles. The upper bound frequency estimate using the Clopper-Pearson formula (as calculated by the CODIS program) = 0.24%.

Meaning of example #1: With 95% confidence, the maximum true frequency of the mtDNA profile in the African American population is 0.24%, or 1 in 417 individuals (this is rounded down to 1 in 410 as mentioned below in 5d).

4.2 Frequency estimate when the mtDNA sequence is not observed in the database.

In cases where the mtDNA profile is not observed in the database (e.g. x = 0), the Clopper-Pearson formula simplifies to the following equation:

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

where α is the confidence coefficient (use 0.05 for a 95% confidence interval), and n is the number of profiles in the population database.

Example #2: A mtDNA sequence is observed 0 times in a database containing 2000 sample profiles. The frequency estimate is $1\text{-}0.05^{1/2000} = 1\text{-}0.999 = 0.001$.

Meaning of example #2: For a database size of 2000 sequence profiles, the frequency of a mtDNA profile not observed in the database is 0.1% or 1 in 1000.

4.3 Based on the current CODIS mtDNA population database, the search software supplies separate results of the frequency estimates for major population groups. The results for the African-American, US Hispanic, US Caucasian, and US Asian population groups will be reported. It is not the intent of the report to draw any inference as to the population origin of the contributor(s) of the evidence.

4.4 Reports will present the upper bound 95% confidence interval estimate for each population group, and express this as a percentage and a frequency, e.g., an upper bound 95% confidence interval estimate of 0.5% (1 in 200). Frequency estimates will be rounded down to nearest 10 or single whole number. The intent of the report is to present a conservative range of estimates of the strength of the mitochondrial DNA comparison.