# Statistical Analysis

The frequency of occurrence of evidence sample types will be reported based on the type of analysis that was performed. When DNA sequencing analysis is done on a given comparison, only the DNA sequencing statistical analysis will be reported. Nevertheless, the statistics, when performed, will be included in the case file.

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 evidence sample(s) and reference sample(s) used in the comparison (see previously discussed sequence reporting criteria).

Statistics may also be presented comparing evidentiary samples, in which case the statistical analysis will be limited to the shortest range and most conservative reporting of the sequence in common between the evidence samples.

A. For sequencing data, use the database and the procedure suggested by the FBI.

1. Database

The database used to obtain a frequency estimate is maintained by the FBI (Budowle et al 1999, Monson et al 2002) and is available for download at the following web address: http://www.fbi.gov/hq/lab/fsc/backissu/april2002/miller1.htm.

A copy of the database including the search window is found on mtDNA analysts' computers. The database contains HVI (16024-16365) and HVII (73-340) sequences from a variety of unrelated individuals.

2. Searching Profiles

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).

Click on the mtDNA icon on your screen. The search window will open. Several options are pre-selected as indicated below.

Mode: - search
Database: - forensic

Under options (in edit menu):

| | |
|---|---|
| Listing profiles: | - not checked |
| Length variants: | - consider multiple insertions as one difference |
| Partial profiles: | - not checked |
| Statistics: | - display up to 2 differences |
| Listing haplotypes: | - check to list haplotypes that appear multiple times |
| Date: | - check "all profiles" |
| Heteroplasmic scenarios: | - not checked |
| Helper Apps: | - not specified |

Enter your profile ID.

Enter the searchable basepair range for HVI and click **Add**. Repeat for HVII. If your sample has the standard read length (see above) just double-click on the HVI or HVII icons. Enter all differences from the rCRS.

Click **search**.

Select a temporary directory and name for the results file.

The search result consists of the number of samples with 0-2 mismatches to the searched sample in the combined database and divided into different ethnic groups.

**ATTENTION:**

**When sequence heteroplasmy is present at a given position in the mtDNA sequence, the mtDNA database will be searched with an "N" at that position.**

**Even though mtDNA sequence HVII polycytosine length variants are entered, multiple C-stretch length variants at the same position are considered as one difference during the database searches of concordant sequences containing this region and will not add additional rarity. In addition, the number of "C" residues in samples with HVI length heteroplasmy is not considered for comparison purposes.**

3. Frequency estimate

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

**Raw frequency estimates for the occurrence of a given mtDNA profile in the general population is based on the counting method as follows:**

$$p = x/N \quad \text{(Eq. 1)}$$

Where p is the frequency estimate; x is the number of times a profile has been observed in the population database, and N is the number of profiles in the population database.

A confidence interval must be calculated from the results of the database search in order to correct the counting results for sampling errors according to the following equation

$$p \pm 1.96\,[(p)\,(1\text{-}p)/N]^{1/2} \quad \text{(Eq. 2)}$$

The upper 95% confidence interval value (upper bound = $p + 1.96\,[(p)\,(1\text{-}p)/N]^{1/2}$) is calculated as the maximum frequency of occurrence within each population of the same mtDNA sequence as the searched profile.

The upper bound estimate can be calculated automatically using the Popstats spreadsheet for sequencing statistics found on the Forensic Biology network drive.

Example #1: A mtDNA sequence is observed 3 times in a database containing 2000 sample profiles. The frequency estimate is 3/2000 = 0.0015; the upper bound of the confidence interval is equal to $0.0015 + 1.96[(0.0015)(0.9985)/2000]^{1/2} = 0.0015 + 0.0017 = 0.0032$.

Meaning of example #1: With 95% confidence, the maximum true frequency of the mtDNA profile is 0.0032 or 0.32%, or 1 in 310. In other words, at least 99.68% of the population can be excluded as the source of the evidence.

| STATISTICAL ANALYSIS | | |
|---|---|---|
| DATE EFFECTIVE<br>06-20-2016 | APPROVED BY<br>MITOCHONDRIAL DNA TECHNICAL LEADER | PAGE<br>4 OF 5 |

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

The following equation is used:

$1 - \alpha^{1/N}$ (Eq. 3)

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

Example #2: A mtDNA sequence is observed 0 times in the 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 mitotypes or sequence profiles, the frequency of a mtDNA profile not observed in the database is 0.001 or 0.1%; or 1 in 1000, or, with 95% confidence, 99.9% of the population can be excluded as being the source of the evidence.

**c. Based on the FBI database, the mtDNA population database search software supplies separate results of the frequency estimates for four major populations (African-American, Hispanic, Caucasian, and Asian Origin). It is not the intent of the report to draw any inference as to the population origin of the contributor(s) of the evidence.**

**d. 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 estimate 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.**