DNA: Statistical Guidelines

Frequency calculations for STR analysis

When a probative association between an evidence profile and a reference profile is made, a frequency estimate is calculated to give weight to the association. Frequency estimates are calculated for at least three major population groups, generally Caucasian, African American, and Hispanic.

Additional population/ethnic groups known to be relevant to the case for which data is available may also be calculated, if deemed appropriate or if requested.

Reference data for allele frequencies

AmpFℓSTR® Identifiler Plus frequency estimates will use empirical values tabulated from data in the following references:

- Budowle, B., et al., "Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians," *Journal of Forensic Sciences*, 1999, 44(6) pp. 1277-1286.
- Budowle, B., "Genotype Profiles for Five Population Groups at the Short Tandem Repeat Loci D2S1338 and D19S433," *Forensic Science Communications*, 2001, 3(3).
- Moretti, T., et al., "Erratum," *Journal of Forensic Sciences*, 2015, 60(4) pp. 1114-1116.

Statistical calculations

Based on the interpretation of the profile, the analyst may apply one of the following statistical calculations:

- Random Match Probability (RMP)
  - The RMP estimates the probability that a profile from a random person in the population is consistent with the profile from the evidence sample.
- Combined Probability of Inclusion (CPI)
  - The CPI calculation estimates the frequency that a randomly selected person would be included as a possible contributor to an observed mixture.

Continued on next page

DNA: Statistical Guidelines, Continued

Random match probability

Depending on the mixture type (refer *Steps for Profile Interpretation- Step 4*), either a "restricted" or an "unrestricted" random match probability may be applied.

- The "restricted" RMP is conditioned on the number of contributors and with consideration of quantitative peak height information and inference of contributor mixture ratios. A "restricted" approach will limit the genotypic combinations of possible contributors.
- The "unrestricted" RMP is also conditioned on the number of contributors, but is performed without consideration of quantitative peak height information or inference of contributor mixture ratios.

Both use the following formulas:

$2pq$	Heterozygote genotype frequency
$p^2 + p(1-p)\theta$	Homozygote genotype frequency
$2p - p^2$	Obligate allele with dropout *('2p')*

where p = the frequency of allele p
q = the frequency of allele q
θ = homozygote correction factor (see *Correction factor for homozygotes*, below)

The appropriate calculation to estimate the frequency of all genotypes that include an obligate allele (with a frequency of p) is *'2p.'* The laboratory will use the expanded *'2p'* formula, $2p - p^2$. If the '$2p$' formula is used more than once at a locus, the frequency of one of the duplicated heterozygote genotypes will be removed.

- When the interpretation for a locus or a profile is inconclusive, that locus or profile will not be used for statistical analysis.
- When the interpretation at a locus includes only one genotype, the appropriate formula above is used to calculate the genotype frequency.
- When the interpretation at a locus includes more than one genotype, the RMP is the sum of the individual frequencies for the genotypes included following mixture interpretation. Adding the frequencies of each genotype provides a frequency of A genotype or B genotype.

Block continued on next page

DNA: Statistical Guidelines, Continued

Random match probability (continued)

The frequencies calculated for all of the individual loci are then multiplied together, using the product rule, to give the estimated probability of the profile as a whole. If a number greater than one is generated by adding the individual frequencies together, round the number to 1.0.

For examples, refer to *DNA: Statistical Calculations.*

Combined Probability of Inclusion (CPI)

CPI is applied to mixture profiles where the contributions of individual donors cannot be resolved. There are no assumptions about the number of contributors when using CPI. Loci with alleles below the stochastic threshold may not be used for statistical purposes.

The probability of inclusion (PI) calculation provides an estimate, at a locus, of the portion of the population that has a genotype that is represented in the mixed profile, and therefore would be included as a possible contributor to the mixed profile.

• Example: If evidence includes three alleles (A_1, A_2, A_3) at a locus, then:

$$PI = (a_1 + a_2 + a_3)^2$$

$$PI = (a_1\,a_1 + a_2\,a_2 + a_3\,a_3 + 2a_1\,a_2 + 2a_1\,a_3 + 2a_2\,a_3)$$

In this example, the only genotypes that would be included in the PI calculation and as possible contributors to the mixture would be:

- A_1A_1
- A_2A_2
- A_3A_3
- A_1A_2
- A_1A_3
- A_2A_3

PI at each locus is first determined and then PIs from all of the included loci are multiplied together, using the product rule, to give the combined probability of inclusion (CPI).

A locus that contains a single allele will not be included in a PI calculation, if allele dropout is suspected. If the analyst concludes that there is no allele dropout, and all contributors are represented by the single allele (all homozygous), then the allele can be included in a PI calculation.

Continued on next page

DNA: Statistical Guidelines, Continued

Conservative calculations

The following concepts were implemented to ensure that the frequency estimates are conservatively calculated:

- Correction factor for homozygotes
  - To account for non-random mating, θ is applied to the calculation for a homozygote genotype. Empirical studies have shown that a conservative value for θ is 0.01. A θ value of 0.03 is applied when calculating a homozygote genotype for an isolated sub-population such as a Native American population.
- Five-event minimum allele frequency
  - A five-event minimum allele frequency is used for rare alleles. For each individual allele, an observed allele count less than five is raised to five. This modified allele count is converted to a frequency and used for all subsequent genotype calculations.

Conclusions

The interpretation and comparison of evidence profiles to reference profiles will lead to the conclusions the analyst makes. Refer to *DNA: Profile Interpretation* for additional information.

NOTE: A statistical calculation less than 1 in 2 for any of the three should not be used to make an inclusion. The profile can, however, be used to exclude an individual.

Haplotype statistics for Y-STR analysis

A consolidated United States Y-STR population database (www.usystrdatabase.org) consisting of anonymous Y-STR profiles from various population/ethnic groups has been established and should be used for reporting the significance of a Y-STR inclusion. The website also has the statistical formulas that are used to calculate frequency estimates.

The search of the database provides the number of times a specific haplotype is observed in the database. The basis for the haplotype frequency estimation is the counting method. The application of upper bound of a confidence interval corrects for sampling variation uncertainty. Typically, upper bound frequency estimates for African Americans, Caucasians, and Hispanics will be used for reporting purposes. Examples of frequency estimate calculations can be found in *DNA: Statistical Calculations.*

Continued on next page

DNA: Statistical Guidelines, Continued

Generating paternity statistics

The CODIS *Popstats* software will be used to generate the Combined Paternity Index (CPI) and Probability of Paternity (PP) (see sections that follow) for the three major population groups reported:

- Caucasian,
- African American, and
- Hispanic
  - For Hispanics, the data for the Southwestern Hispanic population group is reported.
  - Data for Southeastern Hispanic, Chamorro, and Filipino population groups are automatically generated, but this data will not be reported.

Popstats compares the child and alleged parent profiles at each locus and automatically computes the following values:

- *Parentage Index*
- *Probability of Exclusion*
- *Probability of Parentage*

This laboratory will only be entering the mother as the 'known parent' and the alleged father as the 'alleged parent.' Therefore, the *Parentage Index* generated will be the reported *Paternity Index* and the *Probability of Parentage* will be the reported *Probability of Paternity*. *Probability of Exclusion* will not be reported.

Continued on next page

DNA: Statistical Guidelines, Continued

Popstats calculations

In order to perform parentage calculations using *Popstats*, three profiles must be entered:

- mother
- child, and
- suspected father

The following procedure is used to obtain results.

Step	Action
1	Log on to CODIS and open the *Analyst Workbench* program. Select *Popstats.*
2	Select Parentage Calculation.
3	Enter the profiles of *Biological parent* (mother) *Child/product of Conception* *Alleged parent* (suspected father)
4	Click on *Calculate* and the window will display the *Parentage Statistics*.

Paternity index

The **Paternity Index (PI)** is a likelihood ratio based on two conditional probabilities:

$$PI_{locus} = \frac{P \text{ (that the alleged father passed an allele to the child)}}{P \text{ (a randomly selected man passed an allele to the child)}}$$

The *Paternity Index* reflects how many more times likely it is to observe a particular set of alleles under the hypothesis that the alleged father is the biological father compared to the hypothesis that a randomly selected man is the biological father. The *Paternity Index* is based on the assumption that the randomly selected man has a similar ethnic background to the alleged father.

Formula tables for the *Paternity Index* are listed in the Parentage *Formula Table* in the *Popstats* software. The exact formula for the *Paternity Index* depends on the obligate paternal allele and the homozygosity of the alleged father. Obligate paternal alleles are alleles that the biological father is required to have based on the relationship between the mother and the child.

Continued on next page

DNA: Statistical Guidelines, Continued

Combined paternity index

The **Combined Paternity Index (CPI)** for a DNA profile is calculated using the Product Rule by multiplying the individual PIs for each locus tested.

Probability of paternity

The **Probability of Paternity (PP)** is based upon Bayes Theorem. The probability of paternity tests the hypothesis that the alleged father is the biological father by incorporating a prior probability that the alleged father is the true biological father.

$$PP = \frac{\text{CPI (prior probability)}}{\text{CPI (prior probability)} + [1-(\text{prior probability})]}$$

The laboratory uses a prior probability set to a neutral value of 0.5 which simplifies the formula to:

$$PP = \frac{CPI}{(1+CPI)}$$

For example, a probability of paternity of 99% reflects a 99% probability that the hypothesis that the alleged father is the biological father is correct and a 1% probability that this hypothesis is incorrect.

Continued on next page

DNA: Statistical Guidelines, Continued

Paternal mutation rates and mean power of exclusion

In cases where there is a paternal mutation, *Popstats* requires a mutation rate and mean power of exclusion to be entered for that locus. The current mutations rates and mean powers of exclusion for each population group can be found in the following table:

Locus	Paternal Mutation Rate	Mean Power of Exclusion		
		African American	Caucasian	Hispanic
D8S1179	0.002031	0.5747954	0.6122379	0.6027166
D21S11	0.001709	0.7233475	0.7082609	0.6412956
D7S820	0.001348	0.5757591	0.61646	0.5622929
CSF1PO	0.002021	0.5777266	0.4923128	0.4524626
D3S1358	0.001691	0.5433942	0.5888644	0.4918526
TH01	0.00007	0.5111277	0.5719272	0.5316202
D13S317	0.001743	0.4655392	0.568405	0.6544299
D16S539	0.001127	0.6025692	0.5579725	0.5609566
D2S1338	0.001526	0.7808132	0.7376168	0.6755826
D19S433	0.000745	0.6923103	0.5729251	0.6757108
VWA	0.003258	0.6239527	0.6252292	0.562939
TPOX	0.00013	0.554485	0.3769857	0.3609418
D18S51	0.00253	0.7442273	0.7483559	0.7463515
D5S818	0.001742	0.5050401	0.4260790	0.4937583
FGA	0.003713	0.7279926	0.7170999	0.7515565