14 Statistics

Once a match has been identified and an individual included as a possible source of evidentiary material, the significance of that match is estimated to allow investigators, the legal sector, and ultimately a jury of lay persons to place the appropriate emphasis on the conclusion. Although every locus analyzed is evaluated, some loci may provide no information with regard to a particular comparison. Because it is the significance of the match that is important, only the matching loci can be taken into account in estimating significance and only after the analyst has determined that the profiles match. The different methods of calculating significance, when each is to be applied, and the population data from which they are calculated are addressed in this protocol.

If significance estimates are not presented in the report, inclusionary statements may not be provided, with the exception of intimate samples. Intimate samples are samples that originate from within an individual (e.g., vaginal swabs, tampon collected from complainant's body) or from an individual's body (e.g., fingernail scrapings, lip swabs). Garments and other external items, such as panties and feminine pads, are to be treated as non-intimate samples. The laboratory must perform statistical analysis in support of any inclusion, other than for intimate samples, that are determined to be relevant in the context of a case, irrespective of the number of alleles detected and the quantitative value of the statistical analysis.

14.1 Significance Estimation for Forensic Profiles

For all estimates, statistical significance will be expressed as an inverse probability of inclusion and likelihood ratios will not be calculated.

Definitions

1. Source identification/attribution: When the estimate for each calculated population group exceeds 1 in 100 billion probability of a random match, the reporting statement may include identification of the source. This number reflects a 99% confidence limit, assuming a US population of less than 1 billion.
2. Single source significance: The single source significance calculation can be applied if the evidentiary profile compared fits the criteria for either a single source or a major component of a mixed source.
3. Single source: Evidence profiles can be identified as single source if there are no more than two alleles per locus and the alleles at each locus are balanced. For this determination, alleles are considered balanced if the smaller peak height is at least 60% of the larger peak height. The peak height ratio (PHR) is calculated by dividing the peak height of an allele with a lower RFU value by the peak height of an allele with a higher RFU value, then multiplying by 100 to express the PHR as a percentage. Possible contribution of stutter is not subtracted prior to application of this 60% criterion. More than two peaks is possible for a single-source sample in the event of a trisomy. If suspected, the sample should be interpreted with extreme caution.
4. Major component of a mixed source: Some mixed source profiles may be clearly

differentiated into major and minor components. For STR analysis, if the major component fulfills certain criteria, it may be treated the same way that a single source profile is treated for estimating match significance. These criteria are: 1) Interpretable loci show a major component and 2) no minor peak is greater than ~30% (≤30.5%) of the height of the shortest major peak in the same locus for Identifiler amplifications and ~ 35% (≤35.5%) for Identifiler Plus amplifications. Heterozygous alleles considered major must also meet the 60% PHR requirement. Instances of homozygosity of the major donor or the minor donor, along with allelic sharing between contributors, may influence the balance such that the 30% or 35% parameter may be exceeded; the Technical Leader must approve instances where a major contributor is declared, despite a >30% or >35% PHR. If a clear major donor cannot be identified at the majority of the loci examined, the sample will be reported and treated statistically as an indistinguishable mixture. As with the single source evaluation criteria, no prior adjustment for stutter is assumed in the 30% or 35% maximum. It is permissible to determine a major profile at most loci, even if at some loci the major component cannot be determined. In this case, only the loci for which the major component can be determined may be used in the single source significance determination.

Stochastic Threshold

The analytical threshold is the "minimum height requirement at and above which detected peaks can be reliably distinguished from background noise"[1]; peaks above this threshold are generally not considered noise and are either artifacts or true alleles. The analytical threshold for Identifiler amplifications is 100 RFUs; the analytical threshold for Identifiler Plus amplifications is 50 RFUs. The stochastic threshold is the "RFU value above which it is reasonable to assume that, at a given locus, allelic dropout of a sister allele has not occurred[1]". The validation data support a stochastic threshold for Identifiler Plus samples of 200 RFUs, using a 5 or 10 second injection. If the sample is presumed to be single-source, any homozygous alleles must meet or exceed the 200 RFU stochastic threshold for comparison and subsequent statistical analysis. Alternatively, a presumed single-source locus with two alleles may be used for comparison and statistical analysis, should one or both of the alleles not exceed the stochastic threshold. Mixture loci with alleles below the stochastic threshold may not be used for statistical purposes to support an inclusion.

Single source significance calculation

For evidence profiles that meet the criteria for either single source or major component of a mixture, and for Identifiler Plus samples that meet or exceed the stochastic threshold criteria as indicated above, each locus frequency calculation will use NRC II Recommendation formulae:

For homozygotes, $f = p^2 + p(1-p)\theta$ where $\theta = 0.01$ (NRC II formula 4.4a).

For heterozygotes, $f = 2pq$ (NRC II formula 4.1b).

Mixed source significance calculation

Evidence profiles that do not fit the criteria for a single source are considered mixtures of two or more sources of biological material. An Identifiler locus should not be used in the statistical interpretation if there is evidence of possible additional alleles below threshold, as determined by alleles called if the analysis minimum threshold is lowered to 50 RFU. Mixture loci from an Identifiler Plus amplification with alleles below the stochastic threshold may not be used for statistical purposes to support an inclusion.

Mixture calculations may not be performed on mixtures believed to contain DNA from related individuals, as the calculation assumes unrelated individuals. Indistinguishable mixtures of related individuals will be reported as inconclusive, given a statistical analysis is not possible to support the inclusion.

The probability of inclusion (PI) is calculated as (sum of allele frequencies)^2 for each locus. The combined probability of inclusion (CPI) is the product of the individual locus PIs.

CPE (combined probability of exclusion) = 1 - CPI

For the frequency of each locus, $PI = (p_1 + p_2 + p_3 + p_4 + \ldots + p_k)^2$ where p is the estimated frequency of the allele detected for each allele 1 through k, and k is the number of alleles detected at the locus.

For all loci used to identify the match, $CPI = (PI_1 \times PI_2 \times PI_3 \times PI_4 \times \ldots \times PI_k)$, where k is the number of loci.

PopStats is configured to use Theta in the Probability of Exclusion/Inclusion Calculations:

$PE = 1 - [p^2 - \theta(1 - p)p]$, where $p = p_1 + p_2 + \ldots + p_k$

Interpretation of Potential Stutter Peaks in a Mixed Sample[1]

For mixtures in which minor contributors are determined to be present, a peak in stutter position (generally n-4) may be determined to be 1) a stutter peak, 2) an allelic peak, or 3) indistinguishable as being either an allele or stutter peak. This determination is based principally on the height of the peak in the stutter position and its relationship to the stutter percentage expectations established by the laboratory. Generally, when the height of a peak in stutter position exceeds the laboratory's stutter expectation for a given locus, that peak is consistent with being of allelic origin and should be designated as an allele. If a peak is at or below this expectation, it is generally designated as a stutter peak. However, it should also be considered as a possible allelic peak, particularly if the peak height of the potential stutter peak(s) is consistent with (or greater than) the heights observed for any allelic peaks that are conclusively attributed (i.e., peaks in non-stutter positions) to the minor contributor(s).

Filtered minor peaks must be included in the mixture calculation if they are similar to (≥60%) or greater than peaks that are conclusively attributed to the minor component of the mixed sample.

Minimum and null allele frequencies

Following NRC II recommendations, minimum allele frequencies are calculated using 5/2N where N is the number of individuals in the population database, and null allele frequencies are set to 0.

Off-ladder alleles

Off-ladder alleles that have been confirmed by reinjection (in the case of STR analysis) or appearance in more than one sample will be used to determine a match and estimate the significance of that match. The allele frequency will be the calculated minimum allele frequency for that locus and population group.

Software

The latest available and installable version of the FBI's PopStats software will be configured to use the above formulae and used to calculate significance estimates. The Forensic-Single Sample data input option will be used for single source significance calculations; the Forensic Mixture Case data input option and the Mixture Formula will be used for mixed source significance calculations.

In PopStats[2]:

The African American (Blk), U.S. Caucasian (Cau), Hispanic (SWH) databases for the 13 core STR loci were created from data developed by the FBI Forensic Science Research Unit in Quantico, VA. The data are published in the Journal of Forensic Sciences under the title "Population Data on the Thirteen CODIS Core Short Tandem Repeat Loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians", by Bruce Budowle, Ph.D., Tamyra R. Moretti, Ph.D., Anne L. Baumstark, B.S., Debra A. Defenbaugh, B.S., and Kathleen M. Keys, B.S. (J. of Forensic Sci., 1999; 44(6):1277-1286).

The Southeast Hispanic (SEH) data was collected and analyzed as part of the FBI's CODIS STR Study and is published in the Journal of Forensic Science under the title of "CODIS STR Loci Data from 41 Sample Populations" by Budowle, et. al, (J. Forensic Sci. 2001;46(3):453-489). In fact, the FBI\STR*.SEH bin files are essentially the same ones in the FBI\CODIS STR Study\Hispanic\Florida, except for the following minor differences:

-- allele frequency for D21S11 "33.2" is 0.0367 in FBI\STR\D21S11.SEH; and is 0.0366 in FBI\CODIS STR Study\Hispanic\Florida\D21S11.Florida

-- allele "8" for D8S1179 in FBI\STR\D21S11.SEH is lumped in "<9" for D8S1179 in FBI\CODIS STR Study\Hispanic\Florida\D21S11.Florida

The African American (Blk), U.S. Caucasian (Cau), Hispanic (SWH), Chamorro, and Filipino databases for D2S1338 and D19S433 were copied from Table 1 and Table 2 of "Population Data on the STR Loci D2S1338 and D19S433" by Bruce Budowle, Patrick J. Collins, Pero Dimsoski, Constance K. Ganong, Lori K. Hennessy, Craig S. Leibelt, Sulekha Rao-Coticone, Farideh Shadravan, and Dennis J. Reeder; Forensic Science Communications, July 2001, Volume 3, Number 3. The article may be found at the web site http://www.fbi.gov/hq/lab/fsc/backissu/july2001/budowle2.htm and http://www.fbi.gov/hq/lab/fsc/backissu/july2001/budtabs.htm.

14.2 Significance Estimation for Forensic Parentage and Relationship Cases

This section refers to paternity calculations, as the vast majority of cases will involve questions of the paternity of a child. However the same statistical approach and formulas can be used in significance estimations in maternity cases.

Formulas for Paternity Index Calculations

Capital letters refer to the allele(s) present in each individual tested, and small letters refer to the frequency of the allele(s).

M	C	AF	PI
BD	AB	AC	1/2a
BC	AB	AC	1/2a
BC	AB	AB	1/2a
BC	AB	A	1/a
B	AB	AC	1/2a
B	AB	AB	1/2a
B	AB	A	1/a
AB	AB	AC	1/ [2(a+b)]
AB	AB	AB	1/ (a+b)
AB	AB	A	1/ (a+b)

Formulas for Paternity Index Calculations, cont'd

M	C	AF	PI
AB	A	AC	1/2a
AB	A	AB	1/2a
AB	A	A	1/a
A	A	AB	1/2a
A	A	A	1/a
Unknown	AB	AC	1/4a
Unknown	AB	AB	(a+b)/4ab
Unknown	AB	A	1/2a
Unknown	A	AC	1/2a
Unknown	A	A	1/a

The CPI will be calculated by multiplying these individual PI's. The Probability of Paternity will be calculated, using the formula below, which assumes a prior probability of 50%:

Probability of paternity = (CPI/(CPI +1)) X 100

14.3 Significance Estimation for Y-STR Samples[3]

Y-STR loci are located on the nonrecombining part of the Y-chromosome and, therefore, should be considered linked as a single locus. A Y-STR database must consist of haplotype frequencies rather than allele frequencies. A consolidated U.S. Y-STR database (http://usystrdatabase.org) has been established and should be used for population frequency estimation.

At (http://usystrdatabase.org):

- Tabular results display the ancestries selected, number of haplotypes in database having data for the selected loci, number of haplotypes in the database matching the entered haplotype, the frequency, and frequency upper bound (95%)

- "Overall Database Summary" gives statistics statements for the total database and for each ancestry selected

- Blue links under "Number of Haplotypes (with Selected Alleles)" gives a pop-up listing of haplotype searched and matching haplotypes from database

All races searched should be reported statistically, along with the database used for analysis.

14.4 Population Databases

The significance estimate calculations above use empirically determined allele frequencies for each of the represented population groups.

Analysts will routinely calculate and report significance estimates using Caucasian, Black, and Hispanic databases. STR allele frequency data is also available for Japanese, Chinese, and Vietnamese populations for use at the analyst's discretion.

STR allele frequency data for all thirteen NDIS core STR loci are configured for each Popstats installation in the form of data files. The analyst will routinely use the FBI Population database for allele frequency data and calculations. Although Amelogenin is used for declaration of a match, it is not used in calculation of match significance.

References:

1. SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, Approved 1/14/10.
2. CODIS 5.7.4 PopStats
3. Ballantyne, J. (March, 2011). Interpretation of Y STR Mixtures and Statistical Applications. Presentation at the NFSTC DNA Mixture Interpretation Conference, Largo, FL.