Forensic Biology Section

Genetic Analysis

1. Scope

1.1. This procedure provides a means of applying statistical significance to a **match** between two or more STR profiles, the **relatedness** of two individuals being compared for maternity/paternity, the **inclusion** of a person as a potential donor to a STR mixture, or the **corrected match probability** of a Y-STR profile. Which calculations to choose depends on the type of testing being performed, the genetic information available, and the scenario being examined.

1.2. Random match probabilities are used for matches between known and questioned samples having single source DNA profiles, clear major and/or minor donors in mixed DNA profiles, or deduced profiles from intimate body swabs.

1.3. Combined Probability of Inclusion of all possible donor combinations are used when the donors to a mixture of DNA profiles cannot be conclusively separated.

1.4. Likelihood Ratios are used when calculating potential Parent-Offspring relationships.

1.5. Corrected match probabilities are used when the same Y-STR profile is obtained from known and questioned samples.

1.6. Statistics for STR profiles are primarily reported for the Caucasian and African American groups in the U.S., although statistics are available for other groups. Statistics for Y-STR profiles are typically reported for the overall United States groups, although statistics are also available for specific groups.

2. Quality Assurance

2.1. When reporting the probability of a profile, report only three (3) significant figures and do not round up. Digits beyond three significant figures should be truncated (i.e. “1 in 23,456,789” is reported as “1 in 23.4 million”). For figures less than 100,000, the full number may be written out with zeros for any integers beyond the three significant figures (i.e. “1 in 23,456” may be reported as “1 in 23,400” or “1 in 23.4 thousand”).

2.2. Actual testing of a large number of samples was used to identify the frequency of the individual alleles, referred to its observed or “empirical” frequency. The formula 5/2N is used to calculate the minimum allele frequency for autosomal alleles, and 1/N +1 is used to calculate the minimum allele frequency for Y-chromosome haplotypes (N = number of profiles in the population database). If an empirical allele frequency is less than the minimum allele frequency, or if an allele has not been observed in a population database, the minimum allele frequency is used.

2.3. For inclusions or matches between autosomal profiles, the probability for a person’s race (if known) may be reported. If the race is not known, the Caucasian and African American results can be reported.

2.4. For inclusions between Y-chromosome profiles, the overall United States probabilities will be reported. The probability for a person’s race (if known) may also be reported.

2.5. Theta (θ) is used to correct for population subdivision. A Theta value of 0.01 is used to calculate homozygous loci probabilities in autosomal profiles. The Theta value used to calculate match probabilities in Y-chromosome profiles changes with the number of loci in the haplotype and whether or not Native Americans are included in the calculation.

3. Random Match Probability (RMP) for Single-Source Profiles

3.1. RMP is used to denote the strength of a match between a known reference sample and an unknown evidence sample at the overlapping loci.

3.2. RMP is used when single-source DNA profiles are obtained, clear major or minor donors can be interpreted in mixed DNA profiles, or individual DNA profiles can be deduced from mixed DNA profiles on intimate body swabs.

3.3. If alleles at a locus cannot be clearly attributable to a single donor, that locus will not be included in the random match probability calculations.

3.4. The RMP is the inverse of the profile's frequency and expresses the probability that a randomly selected person in a given population would have the same DNA profile.

3.5. Calculating Random Match Probability:

3.5.1. Open PopStats program (a standalone application on desktop computers or as part of the Analyst Workbench on the CODIS Server or Workstation).

3.5.2. The default calculation will be "Forensic Single Sample". Type in the Lab Case # and Item # in the "Specimen ID" box, include any relevant "Comments", and then enter the alleles under "Allele 1" and "Allele 2" for each locus to be included in the RMP calculation.

3.5.3. Click the "Calculate" button. The "Summary of RMP" results window will display the estimated frequency (f) for each locus in each population, as well as the total estimated frequency for the profile in each population. The "Summary of Inverse RMP" displays the estimated probability (1/f).

3.5.4. Click the Print button, choose the "Relatedness Report" format, and print the report. It is optional to print just the Caucasian and African American statistics.

Sample report:

UNCLASSIFIED//FOR OFFICIAL USE ONLY/LAW ENFORCEMENT SENSITIVE

Popstats DNA Relatedness Profile Conditional Match Statistics

Database: \\10.51.24.76\CODIS\Popstats\POPDATA\FBI\Modified Expanded FBI
Probability Formula: NRC '96 with selected alternative Homozygous Calculation: 2p
Theta1: 0.01
Agency ID: MEMSP0000
Specimen ID: 2800M_POS_CONTROL
Comment:

Population Group: Caucasian

Relationship: Unrelated

Locus	Allele 1	Allele 2	Frequency	1/Frequency	Code
D3S1358	17	18	6.8759E-02	15	
D1S1656	12	13	1.1761E-02	85	
D2S441	10	14	9.4831E-02	11	
D10S1248	13	15	1.1667E-01	9	
D13S317	9	11	5.0964E-02	20	
Penta E	7	14	1.9544E-02	51	
D16S539	9	13	3.3987E-02	29	
D18S51	16	18	1.9053E-02	52	
D2S1338	22	25	5.7321E-03	174	
CSF1PO	12		1.0893E-01	9	
Penta D	12	13	8.7484E-02	11	
TH01	6	9.3	1.3715E-01	7	
vWA	16	19	3.3562E-02	30	
D21S11	29	31.2	3.6682E-02	27	
D7S820	8	11	6.6340E-02	15	
D5S818	12		1.2583E-01	8	
TPOX	11		6.6925E-02	15	
D8S1179	14	15	4.3709E-02	23	
D12S391	18	23	2.6109E-02	38	
D19S433	13	14	1.9523E-01	5	
FGA	20	23	4.5492E-02	22	
D22S1045	16		1.0253E-01	10	

Composite frequency = 3.0900E-29
1 out of 32,360,000,000,000,000,000,000,000,000

3.6. Formulas for RMP:

3.6.1. Homozygous Loci: if an individual has one allele at a particular locus, the genotype frequency is calculated as $\mathbf{p^2+p(1-p)\,\theta}$, where "p" is the frequency of the allele and "θ" is a measure of population subdivision (θ = 0.01 per NRC-2).

Example: Locus 1 has a genotype of 12, 12. Frequency of allele 12 = 0.130

Frequency = (freq. allele 12)^2 + (freq. allele 12)(1- freq. allele 12)(**θ**) = $(0.130)^2+(0.130)(1-0.130)(0.01)$ = **0.0180**

Probability = 1/frequency = 1/0.0180= **1 in 55.5**

3.6.2. Heterozygous Loci: if an individual has two alleles at a particular locus, the genotype frequency is calculated as **2pq**, where "p" and "q" are the frequencies of the two alleles.

Example: Locus 2 has a genotype of 14, 16. Frequency of allele 14 = 0.145; frequency of allele 16 = 0.214.

Frequency = 2 (freq. allele 14) x (freq. allele 16) = 2 (0.145 x 0.214) = **0.0620**

Probability = 1/frequency = 1/0.0620= **1 in 16.1**

3.6.3. Product Rule: the probability of a profile by multiplying the RMP of all the loci.

Example: Profile probability = (1/55.5) x (1/16.1) = **1 in 893.**

3.7. RMP and Identity Threshold

3.7.1. Identity Threshold is the point at which a DNA profile is considered unique within the context of a case because the RMP of a match between an evidence sample and a known reference is sufficiently rare that the individual can be identified as the source of that DNA profile with a high degree of statistical confidence. The only exceptions are identical twins (which would have the same DNA profile) and close relatives (which would be more likely to have matching profiles than unrelated individuals).

3.7.2. "Uniqueness within the context of a case" is calculated for a population of size (**N**) and confidence interval (α). The formula to calculate the identity threshold is: $1 - (1- \alpha)^{1/N}$, where **N** = population of 360 million and α= 0.01 (for a confidence interval of 99%).

3.7.3. Expressing the RMP as the inverse of the identity threshold's frequency allows the number to be reported as "1 in…" instead of expressing it as a decimal. For example, reporting an RMP of "1 in 1 million" instead of "0.000001", both of which denote the same probability (1/1,000,000 = 0.000001).

3.7.4. The identity threshold = $1 - (1 - \text{alpha})^{(1/N)}$= 0.0000000000279399 = $2.79399\text{x}10^{-11}$ with a 99% confidence interval and a U.S. population of 360 million or less. The inverse of this is $1/2.79399\text{x}10^{-11}$ = approximately 36 billion, expressed as "**1 in 36 billion".**

N = United States population =	360,000,000 (estimated U.S. population through 2030)
1/N = inverse of U.S. population =	1/360 million = 0.00000000278 = $2.78\text{x}10^{-9}$ = **2.78e-9**
α = 99% confidence level =	**0.01**
Identify threshold = 1 – (1 – alpha)$^{(1/N)}$ = 1-(1 – 0.01)$^{(2.78e-9)}$ = 1 – (0.99)$^{(2.78e-9)}$= 1 – 0.99999999997 = 0.0000000000279399 = $2.79399\text{x}10^{-11}$ = **2.79399e-11**	
Inverse of id threshold = 1 / 2.79399e-11 = 35,791,108,772 rounded off to "**1 in 36 billion"**	

3.7.5. Therefore, with a 99% confidence interval and a population is 360 million or less, the 'identity threshold' for the United States is **1 in 36 billion**.

3.7.6. When a random match probability is *greater than or equal to the identity threshold*, the statistical probability is reported as "less than 1 in 36 billion" and the 'identity statement' is reported in the conclusion.

3.7.7. When a random match probability is *less than the identity threshold*, the actual value of the statistical probability is reported and the phrase 'included as a potential donor' is reported in the conclusion.

4. Combined Probability of Inclusion (CPI) for Mixed Profiles

4.1. The CPI is used to denote the strength of a 'match' or inclusion of known references as potential contributors to a mixed DNA profile. CPI is used when a mixture cannot be differentiated into major/minor contributors and individual contributors can't be deduced from intimate body swabs. The CPI represents the number of people from a particular population that would be included as potential donors to a DNA mixture. It is not the probability of a specific individual being a contributor to a mixture. CPI applies to anyone in a particular population being included to a given mixture and is independent of comparison to any specific known references.

4.2. A CPI of 1 in 64,000 means that approximately 1 in 64,000 people in the population could be included as a potential donor to a mixture of DNA profiles. Conversely, it means approximately 63,999 out of 64,000 people (or 99.998%) would be excluded (also known as the Combined Probability of Exclusion).

4.3. The Mixture module in "PopStats", developed by the FBI, can calculate the CPI statistics for mixtures. The program calculates the frequency for all possible genotypes at each locus in the mixture. These genotype frequencies are added together to determine the probability of inclusion (PI) for each locus. The PI for all of the loci are multiplied together to get the final CPI. Homozygous genotypes use theta ($\theta = 0.01$ per NRC-2).

4.4. Calculating CPI:

4.4.1. Open PopStats and click on "Forensic Mixture". Type in the Lab Case # and Item # in the "Specimen ID" box, include any relevant "Comments".

4.4.2. Enter the alleles in the loci to be included in the CPI calculation.

4.4.3. Click the 'Calculate' button. The "Summary of Inclusion/Exclusion" results window will open, displaying the estimated probability of inclusion for each locus, as well as the Profile Inclusion Probability and the Inverse (1/f) for the mixture.

4.4.4. Click on the Print button, choose the "Probability of Inclusion (short)" format, and print the report.

4.4.5. The Probability of Inclusion will be calculated for Caucasians, African Americans, Southeast Hispanics, and Southwest Hispanics using the FBI Expanded allele frequency tables.

Sample report from CPI:

Probability of Inclusion (Short)

Database: \\10.51.24.76\CODIS\Popstats\POPDATA\FBI\Modified Expanded FBI
Theta1: 0.01

	Mixture	*H1*	*H2*
Agency ID:	MEMSP0000	MEMSP0000	MEMSP0000
Specimen ID:	TEST MIXTURE (Keyboard)	Keyboard	Keyboard
Comment:	Mixture of 2800M and 007.		

Allele Frequency

Locus	Allele	Caucasian	AfricanAmerican	SoutheastHispanic	SouthwestHispanic
Amelogenin	X	N/A	N/A	N/A	N/A
Amelogenin	Y	N/A	N/A	N/A	N/A
D3S1358	15	2.4750E-01	3.1770E-01	3.4030E-01	4.2580E-01
D3S1358	16	2.3270E-01	3.1030E-01	2.3380E-01	2.6560E-01
D3S1358	17	2.1040E-01	2.0350E-01	1.8060E-01	1.2680E-01
D3S1358	18	1.6340E-01	5.6200E-02	1.4640E-01	8.3700E-02
D18S51	15	1.3370E-01	1.6790E-01	1.5780E-01	1.4110E-01
D18S51	16	1.0400E-01	1.8270E-01	1.5210E-01	1.1480E-01
D18S51	18	9.1600E-02	1.2270E-01	5.3200E-02	5.0200E-02

Caucasian Profile Probability of Inclusion:	**1.711E-05**	**= 1 in 58,450**
AfricanAmerican Profile Probability of Inclusion:	**1.759E-05**	**= 1 in 56,850**
SoutheastHispanic Profile Probability of Inclusion:	**5.042E-05**	**= 1 in 19,830**

5. Likelihood Ratios (LR) for Parent-Offspring

5.1. LR is used to denote the strength of DNA allele-sharing between two people that may be related as parent-offspring (PO).

5.2. If two individuals share at least one allele at each locus (or all but one locus, to allow for spontaneous mutation), there are two possible hypotheses to consider: (1) the unknown individual and the known reference are related, or (2) the unknown individual is a randomly selected person unrelated to the known reference that just happens to share an allele at each locus by chance. A LR of 1,000 means the probability that the two individuals are related is 1,000 times greater than they are unrelated.

5.3. The LR of a Parent-Offspring match is = $\frac{\text{Probability (maternity or paternity match)}}{\text{Probability (non-maternity or -paternity)}}$

5.4. The LR for all of the loci are multiplied together to determine the final LR value (product rule).

5.5. Calculating Likelihood Ratio:

5.5.1. Open PopStats program.

5.5.2. Click on "**Kinship**". Type in the Lab Case # and Item # in the "Specimen ID" box and include any relevant "Comments".

5.5.3. Enter the alleles for each individual, unchecking loci that won't be used in the LR calculation. Deselect the locus if there is a one-locus mismatch due to mutation.

5.5.4. In the "relationship" tests box (lower right corner), deselect everything except "**PO**".

5.5.5. Click the 'Calculate' button. The "Kinship Index by Pop Group" results will open.

5.5.6. Click on the Print button, choose the "**PopStats Kinship Locus Report**" format, and print the report.

5.5.7. In the report, the kinship probability, the unrelated probability, and the likelihood ration of the two probabilities for the entire profile (and each locus) will be calculated.

Popstats Kinship Locus Report

Database: \\10.10.144.90\CODIS\Popstats\POPDATA\FBI\Modified Expanded FBI

	Reference	*Evidence*
Lab ID:	MEMSP0000	MEMSP0000
Specimen ID:	dsm (Keyboard)	alleged father (Keyboard)
Comment:		

Population Group: SoutheastHispanic

Kinship	Kinship Conditional Probability	Unrelated Conditional Probability	Likelihood Ratio
Parent-Offspring (PO)	9.582E-15	1.215E-18	7,884

Locus	Likelihood Ratio
D3S1358	1.4693
vWA	1.3284
FGA	5.9773
Amelogenin	Inconclusive
D8S1179	2.3912
D21S11	2.1204
D18S51	1.6437
D5S818	6.1200
D13S317	1.7768
D7S820	2.5046
D16S539	1.5654
TH01	1.0607
TPOX	0.98873
CSF1PO	1.8136

5.6. Formulas for Likelihood Ratio:

5.6.1. If the parent and child are both heterozygous and share only one allele:

LR = 1 / (4 x frequency of shared allele)

Example: An unknown individual has a genotype of 9, 10. The alleged parent has a genotype of 7, 9. The frequency of allele 9 (the shared allele) = 0.090.

LR = 1 / (4 x 0.090) = 1 / (0.36) = 2.77

The statistical probability would be "it is 2.77 times more likely that the remains are the offspring of (alleged mother's name) rather than a random person in the population."

5.6.2. If the parent and child are both heterozygous and share both alleles:

LR = [1 / (4 x frequency of 1st allele)] + [1 / (4 x frequency of 2nd allele)]

Example: An unknown individual has a genotype of 12, 14. The alleged parent has a genotype of 12, 14. The frequency of the shared alleles: 12 = 0.145; 14 = 0.214.

LR = [1 / (4 x 0.145)] + [1 / (4 x 0.214)] = [1 / 0.58] + [1 / 0.856] = [1.72] + [1.16] = 2.88

The statistical probability would be "it is 2.88 times more likely that the remains are the offspring of (alleged mother's name) rather than a random person in the population."

5.6.3. If the parent is homozygous and the child is heterozygous (or vice versa):

LR = 1 / (2 x frequency of shared allele)

Example: An unknown individual has a genotype of 17, 17. The alleged mother has a genotype of 16, 17. The frequency of allele 17 (shared allele) = 0.150.

LR = 1 / (2 x 0.150) = 1 / (0.3) = 3.33

The statistical probability would be "it is 3.33 times more likely that the remains are the offspring of (alleged mother's name) rather than the offspring of a random woman in the population."

5.6.4. If the parent and child are both homozygous (thus sharing the one allele):

LR = 1 / (frequency of shared allele)

Example: An unknown individual has a genotype of 11, 11. The alleged mother has a genotype of 11, 11. The frequency of allele 11 (shared allele) = 0.434.

LR = 1 / (0.434) = 2.30

The statistical probability would be "it is 2.30 times more likely that the remains are the offspring of (alleged mother's name) rather than a random person in the population."

6. Haplotype Frequency for Y-Chromosome Profiles

6.1. A Y-STR profile is a haplotype, which is essentially a one-locus DNA profile, so the product rule does not apply. The counting method estimates the haplotype frequency by counting the number of times a profile appears in a database of Y-STR profiles. The database used is called the Y-chromosome Haplotype Reference Database (YHRD) at yhrd.org/search.

6.2. Only single-source Y-STR profiles can be searched at this time (including single-source deduced and major-contributor profiles) because the method for statistical probabilities of Y-STR mixtures has not been validated yet.

6.3. Manually entering a Y-STR profile:

6.3.1. Go to the YHRD "search the database" page (**yhrd.org/search**).

6.3.2. Click on the "manually enter the haplotype" button and choose "**Y23**" for Dataset and "**PowerPlex Y23**" for Kit.

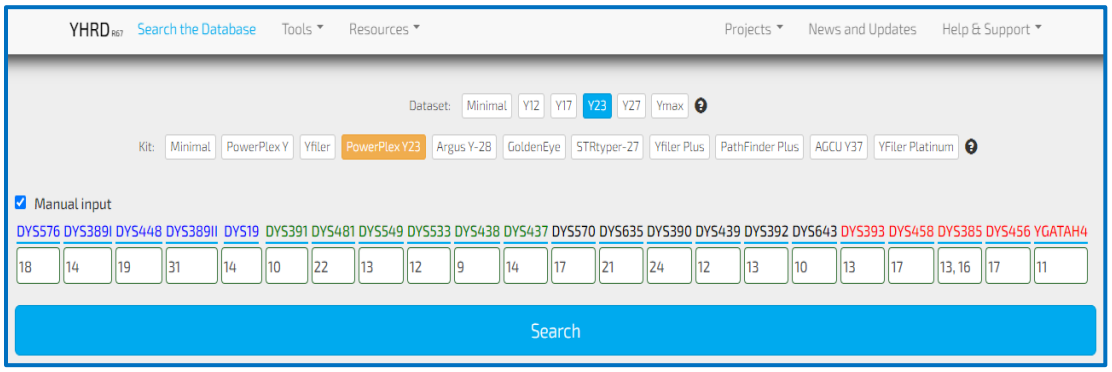
6.3.3. Click in the box beneath a locus and enter the allele(s) for each locus (Tab will move the cursor to the right; Shift-Tab will move the cursor to the left).

6.3.4. If a locus has two alleles (e.g. "13,15" at DYS385), those alleles **must** be separated by a comma. If DYS385 only has one allele (and it is NOT due to drop-out), that allele **must** be entered twice (e.g. "10,10").

6.3.5. If the alleles are not entered in a recognizable way, the box will turn RED, and it will not search the profile until the allele(s) are changed.

6.3.6. Null alleles (loci that do not amplify) are difficult to differentiate from dropout and should be left blank (do not enter any alleles at those loci).

6.3.7. Example of manual data entry screen:



6.4. Importing a Y-STR profile exported from GeneMapper ID-X:

6.4.1. If profiles exported from GMID-X are imported into YHRD, it will display each sample's name next to the profile and print that sample name on the statistical report.

6.4.2. In GMID-X, go to **File → Export Combined Table** and save the .txt file in a folder on the desktop. Even though GMID-X will export all of the samples in the project, YHRD will allow you to choose specific samples for stats in a later step.

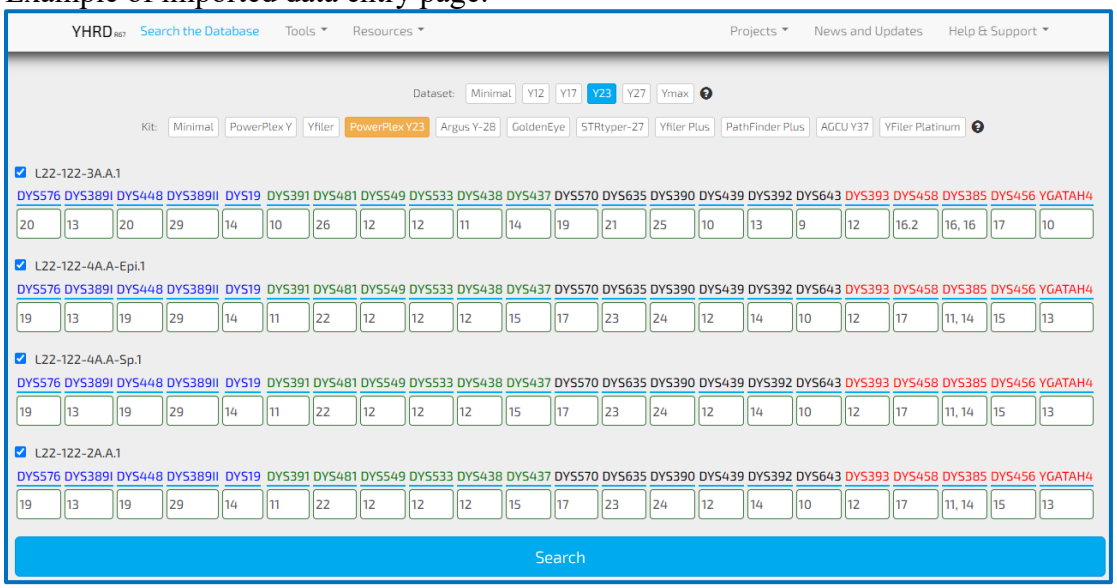
6.4.3. Go to YHRD "search the database" page (**yhrd.org/search**) and click on **Tools→ Data File Validator**, click "choose your file" and select the .txt file exported from GMID-X. Select "Search" in the 'select the type of validation' dropdown list, and then click "Go". The tool points out and corrects any problems (if possible). Click "download

your file as Excel sheet", open the Excel sheet that appears at the bottom of the browser, click 'enable editing', and save it to the folder on the desktop.

6.4.4. Click on "search the database" at the top of the YHRD site, click the 'Search using your… export-file" button, then choose the validated Excel spreadsheet saved earlier.

6.4.5. YHRD will **automatically** set the Dataset (**Y23)** and Kit (**PowerPlex Y23**), display the sample name next to each profile, and place a checkmark next to the first sample.

6.4.6. Example of imported data entry page:



6.4.7. Check boxes next to samples that needs statistics calculated and click "Search" button.

6.4.8. Clicking the back-arrow will return the program to the data entry page (keeping the alleles entered thus far) in case any samples need to be added, removed, or edited.

6.5. Printing Reports:

6.5.1. Choosing the "National database with subpopulations" option applies the formulas from the 2014 SWGDAM Y-STR Interpretation Guideline to the YHRD database and reports the probabilities for Y-STR profiles in the United States.

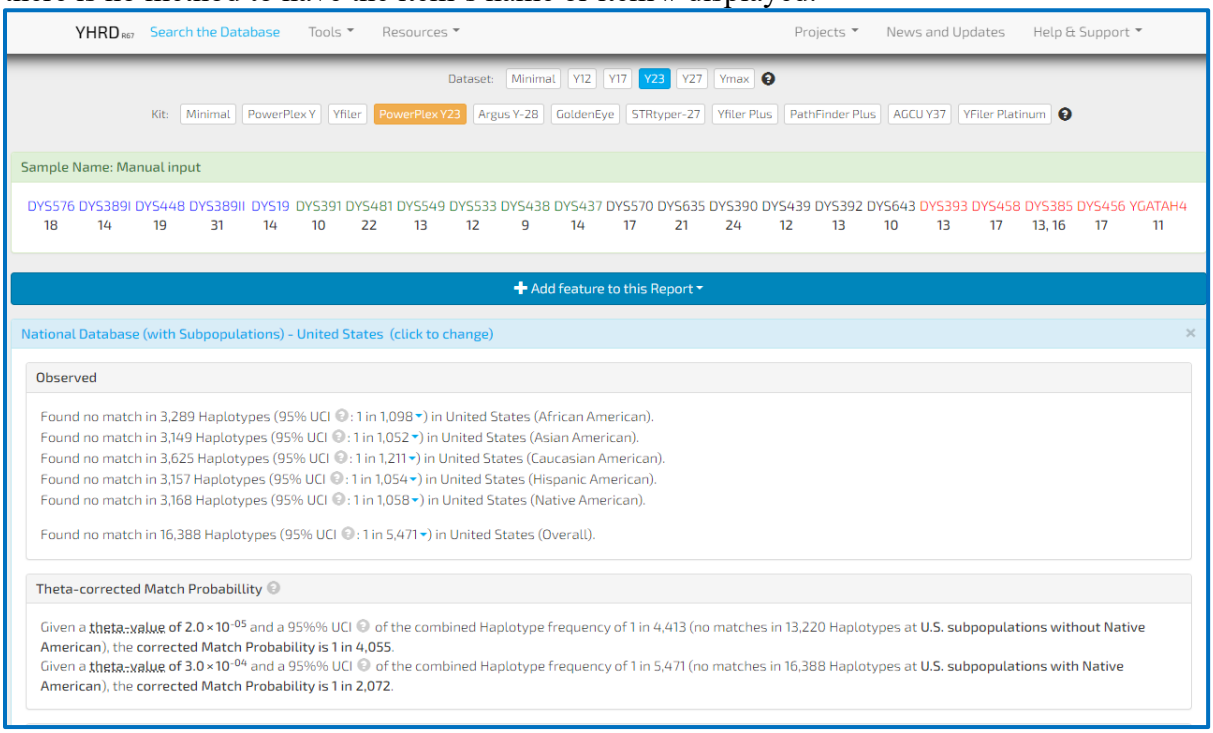
6.5.2. The first section of the report gives the haplotype frequency (= # times observed in database / # of profiles searched against) and the profile probability (= 95% upper confidence limit of the haplotype frequency) for the subsets of the database (African American, Asian American, Caucasian American, Hispanic American, and Native American). It also lists an overall haplotype frequency and overall profile probability that includes all of these subsets.

6.5.3. The second section of the report gives the theta-corrected match probability (= theta-corrected 95% upper confidence limit of the haplotype frequency) with- and without-Native Americans.

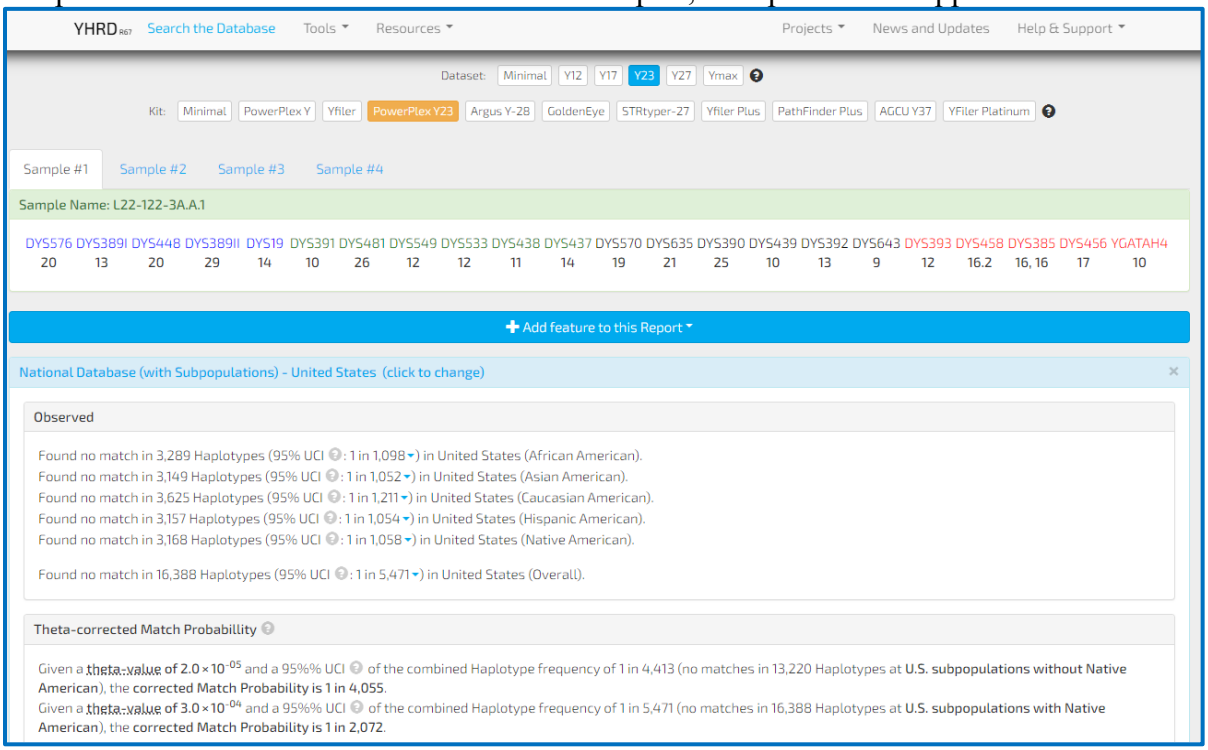
6.5.4. Print the report using the web browser's print function. The 'scale' can be changed to approximately 90% to make the entire report (including footers) fit on one page.

6.5.5. If data was entered manually, write the Case # and Item # at the top of each page.

6.6. Example of manual data entry report. Only one sample is entered and calculated at a time, and there is no method to have the item's name or item # displayed.

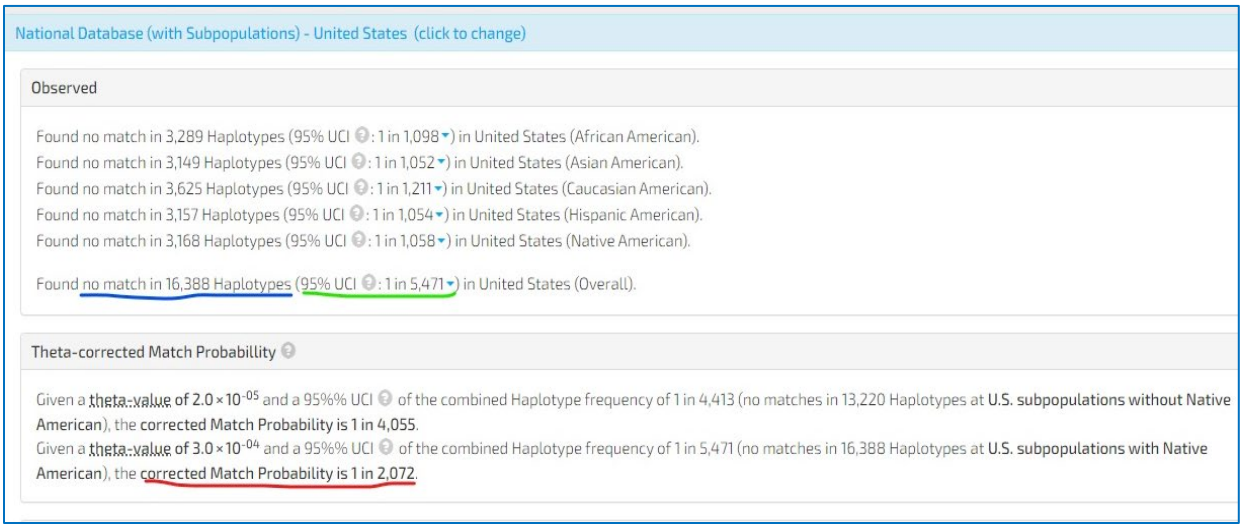


6.7. Example of imported data entry report. There are separate TABS for up to five different samples. If there are stats for more than five samples, a drop-down list appears instead.



6.8. Y-STR statistical probability results are reported as three components:

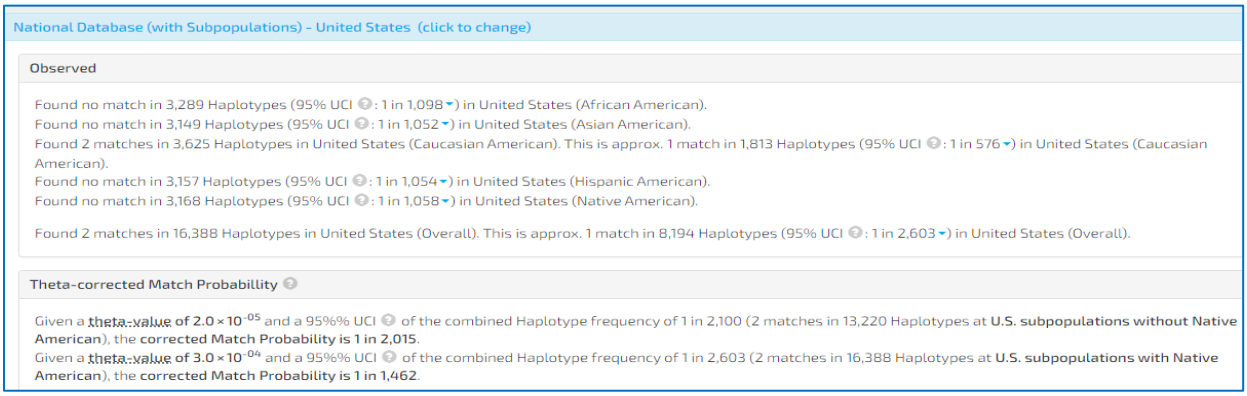
- The profile frequency of being observed in X number individuals, specifying the subpopulation or "overall" for all of the subpopulations.
- The profile probability at the upper 95% confidence limit, specifying the subpopulation or "overall" for all of the subpopulations.
- The (theta-corrected) match probability, specifying "overall" (i.e. including the Native American subset). The YHRD website cannot calculate corrected match probabilities for any specific subpopulations.



The above YHRD result would be reported as:

The Y-STR DNA profile obtained from Item * is consistent with the Y-STR DNA profile of Item *.

- This Y-STR profile was not observed in any United States subpopulations in the YHRD.org database, with a profile probability of 1 in 5,471 individuals (95 % upper confidence limit). This profile has an overall match probability of 1 in 2,072 for all U.S. subpopulations.



The above YHRD result would be reported as:

The Y-STR DNA profile obtained from Item * is consistent with the Y-STR DNA profile of Item *.

- This Y-STR profile was observed in 2 of 16,388 United States Caucasians/Asians/Hispanics/African-Americans/Native-Americans in the YHRD.org database, with a profile probability of 1 in 2,603 individuals (95 % upper confidence limit). This profile has an overall match probability of 1 in 1,462 for all U.S. subpopulations.

7. Glossary

7.1. Autosomal loci = the alleles in a DNA profile from locations NOT on the X- or Y-chromosomes (i.e. all loci except Amelogenin and any Y-STR loci).

7.2. Y-chromosome haplotype = all of the loci in a Y-STR profile. Due to the lack of recombination, the entire Y-chromosome haplotype is treated as a single locus. There are no "allele frequencies", only "haplotype frequencies" as observed in a database of Y-STR profiles.

7.3. Complete DNA profile = when a single-source profile (or deduced profile) has comparable data for all of the loci tested (different than a "complete profile" in the CODIS database).

7.4. Partial DNA profile = when a profile is missing or has inconclusive data at any of the loci tested (different than a "partial profile" in the CODIS database)

7.5. Mixture of DNA profiles = when a profile appears to have DNA from two or more contributors. It is sometimes possible to deduce a single-source profile from a major/minor mixture or from a mixture on an intimate body swab.