

Statistical Cut-Off Study

AmpFISTR® Identifiler® Plus

Report

Appendix A – RMP Study Worksheets

Appendix B – CPI Study Worksheets

Appendix C – Mixture Study Worksheets

Appendix D – RMP Worksheet Validation Worksheets

Appendix E – CPI Worksheet Validation Worksheets

Appendix F – AmpFISTR® Identifiler® Plus User's Guide, Population Data

This study has been technically reviewed and approved for use by:



Jennifer Zeffer, Technical Leader

2/1/15

Date

This study was conducted and written by:



Jessica Skillman, Forensic Scientist III

12/1/14

Date

Introduction

The Forensic Biology Unit is evaluating the use of either a locus specific or numerical statistical cutoff for single source and mixture samples. This value will be used to determine whether a DNA profile from a probative sample contains enough information to provide a useful and/or meaningful result in relation to its frequency within the general population. Profiles which do not meet an appropriate value will not be considered interpretable and will be reported as inconclusive. This cutoff will allow the laboratory to confidently provide weight to both evidence profiles which match an individual and evidence profiles in which an individual cannot be excluded. This study includes the following sections: Random Match Probability Study, Combined Probability of Inclusion Study, Mixture Calculation Study, Random Match Probability Worksheet Validation and Combined Probability of Inclusion Worksheet Validation.

I. Random Match Probability Study

a. Objective

Use the most common alleles to determine the frequency of a one-locus profile through a 15-locus profile. Use this data to determine whether a profile should be considered inconclusive at a specific number of loci.

b. Materials and Methods

Random match probability was calculated using the two most common alleles at each locus for each population based on the Identifiler® Plus User's Manual. The results of each locus were then subtracted one at a time in the following order: D2S1338, D18S51, D16S539, CSF1PO, FGA, TPOX, D13S317, D7S820, D5S818, vWA, TH01, D21S11, D19S433, D3S1358, D8S1179. This order is based on the Identifiler® Plus locus order. The blue dye channel is typically the most intense, followed by the green, yellow and finally red. In addition, samples typically follow a ski slope effect from left to right with the largest loci dropping out first.

Random match probability was then calculated using the single most common allele at each locus for each population. This was done to demonstrate the expected frequency if loci contain single alleles below stochastic threshold. The results of each locus were then subtracted one at a time in the same order to demonstrate possible dropout.

And finally, random match probability was calculated using a combination of the two most common alleles at each locus and the single most common allele at each locus. This was done to demonstrate the possibility of partial dropout at any or all loci.

c. Data Analysis

Random Match Probabilities were calculated using the Forensic Biology Unit's Identifiler® Plus Statistics Calculation Worksheet (previously validated Excel spreadsheet). Each calculation worksheet was printed and the values manually typed into an Excel spreadsheet. Values which are below the population of the District of Columbia were highlighted in red.

d. Results

Number of loci with one most common allele (2p)	Number of loci with two most common alleles	African American Frequency	US Caucasian Frequency	US Hispanic Frequency
0	1	7.142270456	7.201477167	5.793103033
0	2	40.9420186	62.32877959	27.69644203
0	3	326.6329173	315.7953436	272.3499408
0	4	3678.214947	3056.757048	2190.487618
0	5	20699.29912	24131.18499	14313.33514
0	6	187908.1046	210909.2775	96792.75513
0	7	1012336.747	762218.514	421867.6429
0	8	7387049.288	7756828.163	2672978.838
0	9	32603764.59	42255884.27	26637516.02
0	10	210546303.2	152871833	98457750.67
0	11	2787284717	2673069571	1861475365
0	12	17166188792	12805907689	8298415620
0	13	1.43E+11	66293563063	46651154653
0	14	2.02E+12	1.42E+12	9.08E+11
0	15	5.09E+13	2.81E+13	9.41E+12
1	0	1.660577881	1.537515375	1.518372305
2	0	2.719583821	3.031378894	1.939668248
3	0	4.558471038	4.444837088	3.057484629
4	0	9.92266225	8.815622944	5.215770435
5	0	11.57566758	15.15753601	7.756945917
6	0	21.63676183	30.93374696	11.30420565
7	0	29.71266387	39.39600988	14.44073282
8	0	43.12433073	72.36592558	25.07942483
9	0	46.6511583	117.4771519	54.28446933
10	0	64.56014157	110.2037072	54.65613102
11	0	149.652623	325.8536583	158.5154612
12	0	256.8702764	496.5767423	199.8429919
13	0	407.6012002	777.1154026	331.1948822
14	0	1046.741654	2318.363373	1000.588768
15	0	3698.733761	6685.015493	2198.129983
1	14	7.13E+12	4.10E+12	1.99E+12
2	13	1.30E+12	5.70E+11	3.10E+11

3	12	2.47E+11	1.72E+11	91276636754
4	11	68889030325	54839069745	25813032874
5	10	12062469210	9273286694	3959719243
6	9	2584987152	2404563240	1078627522
7	8	633580350.9	716559845.4	234277897.7
8	7	126019131.8	129339052.1	64215571.77
9	6	32122276.69	45579081.96	18821584.08
10	5	6613976.771	10642715.64	4056051.37
11	4	1371077.385	2317983.466	923157.2919
12	3	265029.2585	474954.8122	195801.6616
13	2	55682.6472	137451.9226	31387.00636
14	1	15908.53231	31311.54798	8386.608098

e. Conclusions

Based on the results listed above, a single source profile with two alleles at 8 loci or more will produce a statistic greater than the population of the District of Columbia. If partial results are obtained at any loci, the number of loci required to produce a statistic higher than the DC population will increase.

This data was produced using the most common alleles within each of the populations. It can be expected for casework statistics of single source profiles to be significantly higher. As an example, for the Calculated Probability of Inclusion Study listed next in this evaluation, twenty database profiles were randomly selected and random match probabilities calculated. The most common statistic for a full profile was listed as 2.79×10^{16} . In the above listed study which used the most common alleles, the highest statistic was listed as 5.09×10^{13} . This data indicates that using a specific number of loci as the statistical cutoff for determining inconclusive results and basing it on the most common alleles at each location tested may be an overly conservative approach.

II. Combined Probability of Inclusion (CPI) Study

a. Objective

Use the most common alleles to determine the combined probability of inclusion (CPI) of a one-locus mixture profile through a 15-locus mixture profile with two alleles, three alleles or four alleles at each location. Use this data to determine whether a mixture profile should be considered inconclusive at a specific number of loci.

b. Materials and Methods

CPI was calculated using the two, three and four most common alleles at each locus for each population based on the Identifiler® Plus User's Manual. The results of each locus were then subtracted one at a time in the following order: D2S1338, D18S51, D16S539, CSF1PO, FGA, TPOX, D13S317, D7S820, D5S818, vWA, TH01, D21S11, D19S433, D3S1358, D8S1179. This order is based on the Identifiler® Plus locus order. The blue dye channel is typically the most intense, followed by the green, yellow and finally red. In addition, samples typically follow a ski slope effect from left to right with the largest loci dropping out first.

Data was not calculated for mixtures with more than four alleles. Based on the results obtained, this data would not be needed.

c. Data Analysis

Combined Probabilities of Inclusion were calculated using the Forensic Biology Unit's Identifiler® Plus Mixture Statistics Calculation Worksheet (previously validated Excel spreadsheet). Each calculation worksheet was printed and the values manually typed into an Excel spreadsheet. Values which are below the population of the District of Columbia were highlighted in red.

d. Results

Number of loci with two most common alleles	African American Frequency	US Caucasian Frequency	US Hispanic Frequency
15	71400000	51800000	15100000
14	58300000	54200000	29800000
13	8460000	5240000	3140000
12	2190000	2050000	1130000
11	722000	864000	527000
10	112000	101000	57400
9	37700	64000	34200
8	19000	23800	6990
7	5690	4960	2240
6	2210	2770	1060
5	502	645	334
4	205	169	108
3	37.4	36	27.9
2	9.8	14.4	6.45
1	3.46	3.4	2.83

Number of loci with three most common alleles	African American Frequency	US Caucasian Frequency	US Hispanic Frequency
15	98700	100000	53000
14	17200	21400	17900
13	5130	4270	3960

12	2490	2680	2270
11	1520	2130	1890
10	450	522	430
9	281	433	331
8	210	232	145
7	114	92.8	85.1
6	79.8	75.7	56.4
5	36.6	37.8	31.4
4	21.3	19.5	18.7
3	7.64	7.69	7.61
2	2.94	4.82	2.89
1	1.83	2.14	1.95

Number of loci with four most common alleles	African American Frequency	US Caucasian Frequency	US Hispanic Frequency
15	1180	880	702
14	325	295	339
13	144	100	127
12	103	80.7	96.5
11	75.3	75.2	92.2
10	34.2	30.2	35.2
9	26.6	27.7	31.8
8	22	19.7	18.5
7	16.7	12.5	14.5
6	13.7	11.5	11.2
5	9.32	7.74	8.05
4	7.2	5.98	6.7
3	3.32	3.13	3.57
2	1.59	2.36	1.77
1	1.31	1.64	1.47

e. Conclusions

Based on the results listed above, a mixture profile with two alleles at 12 loci or more will produce a statistic greater than the population of the District of Columbia. However, even full profiles with three or four alleles at each locus will not be expected to produce a CPI higher than the DC population.

This data was produced using the most common alleles within each of the populations. It can be expected for casework statistics of mixture profiles to be higher. However, due to the additive effect of the CPI calculation, the statistics in this study are, in general, much lower than those generated from single source profiles using RMP. Setting a cutoff at such a high number of loci, would not be practical because nearly all mixtures would be considered inconclusive. As with the previous study, using a specific number

of loci as the statistical cutoff for determining inconclusive results and basing it on the most common alleles at each location tested may be an overly conservative approach. Because of the results of this study, a mixture study will be completed using real profiles to determine whether a different approach to determining a cutoff may be appropriate.

III. Mixture Study

a. Objective

Use four profiles in different 2-person mixture combinations to determine the combined probability of inclusion (CPI) of a one-locus mixture profile through a 15-locus mixture profile. Use this data and the data from the previous CPI study to determine whether a mixture profile should be considered inconclusive at a specific number of loci.

b. Materials and Methods

Twenty random profiles from the Quality Assurance database were selected and a Random Match Probability (RMP) calculated for each. The four profiles with the lowest RMP regardless of population group (Samples 1-4) were then selected and combined into six different mixtures (Mixture 1,2, Mixture 1,3, Mixture 1,4, Mixture 2,3, Mixture 2,4 and Mixture 3,4). CPI was calculated for each population based on the Identifiler® Plus User's Manual. The results of each locus were then subtracted one at a time in the following order: D2S1338, D18S51, D16S539, CSF1PO, FGA, TPOX, D13S317, D7S820, D5S818, vWA, TH01, D21S11, D19S433, D3S1358, D8S1179. This order is based on the Identifiler® Plus locus order. The blue dye channel is typically the most intense, followed by the green, yellow and finally red. In addition, samples typically follow a ski slope effect from left to right with the largest loci dropping out first.

c. Data Analysis

Combined Probabilities of Inclusion were calculated using the Forensic Biology Unit's Identifiler® Plus Mixture Statistics Calculation Worksheet (previously validated Excel spreadsheet). Each calculation worksheet was printed and the values manually typed into an Excel spreadsheet. Values which are below the population of the District of Columbia were highlighted in red.

d. Results

	Number of loci	African American	US Caucasian	US Hispanic
Mixture 1, 2	1	3.18	2.44	2.17
Mixture 1, 2	2	9.74	6.69	5.21
Mixture 1, 2	3	28.6	9.64	12
Mixture 1, 2	4	209	45.4	46.3

Mixture 1, 2	5	4.82E+02	8.82E+01	7.79E+01
Mixture 1, 2	6	2.12E+03	3.79E+02	2.47E+02
Mixture 1, 2	7	3.01E+03	4.65E+02	3.72E+02
Mixture 1, 2	8	6.44E+03	1.16E+03	9.45E+02
Mixture 1, 2	9	1.28E+04	3.13E+03	4.80E+03
Mixture 1, 2	10	2.85E+04	6.48E+03	1.26E+04
Mixture 1, 2	11	1.15E+05	2.65E+04	6.42E+04
Mixture 1, 2	12	457000	62800	138000
Mixture 1, 2	13	1780000	161000	383000
Mixture 1, 2	14	4.71E+06	500000	1.12E+06
Mixture 1, 2	15	2.39E+07	2.69E+06	4.88E+06
Mixture 1, 3	1	3.46	3.4	2.83
Mixture 1, 3	2	11.4	10.2	10.4
Mixture 1, 3	3	39.7	22.3	37.8
Mixture 1, 3	4	196	78.5	114
Mixture 1, 3	5	1560	180	343
Mixture 1, 3	6	21300	3190	2860
Mixture 1, 3	7	54900	5700	6040
Mixture 1, 3	8	87400	9350	8780
Mixture 1, 3	9	173000	25200	44700
Mixture 1, 3	10	829000	75200	150000
Mixture 1, 3	11	4.29E+06	3.49E+05	1.10E+06
Mixture 1, 3	12	7.05E+06	4.39E+05	1.31E+06
Mixture 1, 3	13	1.39E+07	7.60E+05	2.57E+06
Mixture 1, 3	14	5.15E+07	2.34E+06	7.93E+06
Mixture 1, 3	15	3.97E+08	1.06E+07	3.23E+07
Mixture 1, 4	1	3.06	3.1	3.16
Mixture 1, 4	2	4.18	4.47	3.8
Mixture 1, 4	3	12.9	9.06	10.1
Mixture 1, 4	4	94.2	42.7	39.1
Mixture 1, 4	5	217	82.9	65.9
Mixture 1, 4	6	1850	759	381
Mixture 1, 4	7	4520	1220	697
Mixture 1, 4	8	9650	3040	1770
Mixture 1, 4	9	17600	5680	6060
Mixture 1, 4	10	38900	8080	9040
Mixture 1, 4	11	201000	37500	66100
Mixture 1, 4	12	331000	47200	79200
Mixture 1, 4	13	1.29E+06	1.21E+05	2.20E+05
Mixture 1, 4	14	3.40E+06	4.16E+05	7.78E+05
Mixture 1, 4	15	2.01E+07	1.54E+06	2.41E+06
Mixture 2, 3	1	3.18	2.44	2.17
Mixture 2, 3	2	5.1	5.49	3.2
Mixture 2, 3	3	16.3	8.76	8.44
Mixture 2, 3	4	80.5	30.9	25.4
Mixture 2, 3	5	182	89.1	50.6
Mixture 2, 3	6	798	383	160

Mixture 2, 3	7	1140	470	242
Mixture 2, 3	8	3220	1310	824
Mixture 2, 3	9	6390	3520	4190
Mixture 2, 3	10	19200	8280	12800
Mixture 2, 3	11	110000	35700	81200
Mixture 2, 3	12	1.81E+05	4.49E+04	9.72E+04
Mixture 2, 3	13	1.14E+06	2.31E+05	5.71E+05
Mixture 2, 3	14	1.20E+07	1.16E+06	3.95E+06
Mixture 2, 3	15	1.77E+08	2.02E+07	9.00E+07
Mixture 2,4	1	2.27	3.25	2.97
Mixture 2,4	2	3.65	7.31	4.4
Mixture 2,4	3	41.8	26.1	14.9
Mixture 2,4	4	306	123	57.5
Mixture 2,4	5	704	239	96.7
Mixture 2,4	6	8890	2040	1160
Mixture 2,4	7	36800	5820	3540
Mixture 2,4	8	156000	28400	30700
Mixture 2,4	9	636000	151000	320000
Mixture 2,4	10	1.02E+06	1.82E+05	4.49E+05
Mixture 2,4	11	5.80E+06	7.86E+05	2.85E+06
Mixture 2,4	12	9.55E+06	9.89E+05	3.41E+06
Mixture 2,4	13	3.72E+07	2.53E+06	9.47E+06
Mixture 2,4	14	3.92E+08	1.27E+07	6.55E+07
Mixture 2,4	15	4.01E+09	1.50E+08	8.25E+08
Mixture 3, 4	1	1.31	1.64	1.47
Mixture 3, 4	2	2.11	3.68	2.18
Mixture 3, 4	3	7.08	8.41	6.71
Mixture 3, 4	4	35	29.7	20.2
Mixture 3, 4	5	57.3	38.4	24.3
Mixture 3, 4	6	488	351	141
Mixture 3, 4	7	1190	563	257
Mixture 3, 4	8	3390	1560	876
Mixture 3, 4	9	6170	2920	3000
Mixture 3, 4	10	18400	4620	5040
Mixture 3, 4	11	219000	47400	80100
Mixture 3, 4	12	361000	59600	96000
Mixture 3, 4	13	7.10E+05	1.03E+05	1.88E+05
Mixture 3, 4	14	7.48E+06	5.16E+05	1.30E+06
Mixture 3, 4	15	1.43E+08	4.80E+06	1.47E+07

e. Conclusions

Based on the results listed above, a mixture profile with results at 10 or more loci may produce a statistic greater than the population of the District of Columbia. However, a full profile would be required to guarantee a statistic higher than the DC population.

While this study produced higher statistics than the previous study, it still indicates that a high number of loci will be required to make conclusions regarding mixture profiles.

IV. RMP Worksheet Validation

a. Objective

Determine if the additional tab in the Identifiler® Plus Statistics Calculation Worksheet accurately indicates whether statistics “will” or “will not” be provided and the statistical values are listed correctly. Samples which do not produce statistics greater than the population of the District of Columbia (650,000) will not be calculated.

b. Materials and Methods

Nine profiles from the RMP study were calculated using the new tab in the Identifiler® Plus Statistics Calculation Worksheet. Each profile was printed and compared to confirm correct values were obtained.

c. Data Analysis

A visual comparison was completed to confirm correct values were calculated and the “will” and “will not” blank is appropriately populated.

d. Results

See Appendix.

e. Conclusions

All calculated values are accurate and “will” or “will not” displayed appropriately.

V. CPI Worksheet Validation

a. Objective

Determine if the additional tab in the Identifiler® Plus Mixture Statistics Calculation Worksheet accurately indicates whether statistics “will” or “will not” be provided and the statistical values are listed correctly. Samples which do not produce statistics greater than the population of the District of Columbia (650,000) will not be calculated.

b. Materials and Methods

Nine profiles from the Mixture Study were calculated using the new tab in the Identifiler® Plus Mixture Statistics Calculation Worksheet. Each profile was printed and compared to confirm correct values were obtained.

c. Data Analysis

A visual comparison was completed to confirm correct values were calculated and the “will” and “will not” blank is appropriately populated.

d. Results

See Appendix.

e. Conclusions

All calculated values are accurate and “will” or “will not” displayed appropriately.

Conclusion

In this evaluation, two methods of determining an appropriate single source and mixture statistical cutoff were used. One method used the most common alleles to establish a theoretical number of loci which would be required in order to guarantee a statistic greater than the population of the District of Columbia (650,000). The other validated the use of an Excel spreadsheet which would indicate if a profile contains enough information to produce a statistic greater than the DC population.

Based on the results obtained from the first method, single source samples would require results from more than half the loci tested and mixture samples would require results at all the loci tested. This theoretical approach may be overly conservative since real profiles generally do not contain such a high number of common alleles. Of twenty randomly selected database profiles, the most common statistic was still 1,000 times higher than the highest theoretical profile statistic which came from the most common alleles at every location.

The second method of allowing the Excel spreadsheet to indicate whether a profile “will” or “will not” have the statistics calculated was validated as accurate. These spreadsheets would allow a profile with less common alleles to be accurately concluded, however it would still produce the same theoretical results as the previous studies.

While both approaches were evaluated in order to set a statistical cutoff, CODIS (Combined DNA Index System) eligibility of the profiles should also be considered before a final decision is made. Results at ten loci are required for NDIS (National DNA Index System) eligibility and seven loci are required for SDIS (State DNA Index System) eligibility AND the rarity of the profile must exceed one in the size of the database itself. It must therefore be determined what type of conclusions may be drawn, if any, on an ineligible profile at the state and/or national level.