NEW YORK STATE POLICE CRIME LABORATORY SYSTEM
TRUEALLELE® CASEWORK VALIDATION ADDENDUM

TABLE OF CONTENTS

NEW YORK STATE POLICE

MEMORANDUM

Station: FIC - Headquarters

Date: June 26, 2013

To: **Dr. Barry Duceman, Director of Biological Sciences**
Dr. Russell Gettig, Associate Director of Biological Sciences

From: **Jay Caponera, Forensic Scientist III**

Subject: **Proposal to incorporate in-house Identifiler® Plus data (known samples) in the TrueAllele® Casework probabilistic genotyping software validation**

Objective:

To expand the current TrueAllele Casework validation by incorporating in-house generated Identifiler Plus data from the 3130*xl* sequencing platform.

Scope:

Probabilistic genotyping methodology has been recognized by SWGDAM and approved for use by the DNA Subcommittee of the NYS Commission on Forensic Science. Further, the NYSP TrueAllele Casework probabilistic genotyping software validation includes multiple peer-reviewed publications. However, the lab does not currently have any TrueAllele Casework data for the newly validated Identifiler Plus amplification chemistry. More specifically, the lab should seek to confirm TrueAllele Casework performance on known mixture and low-template samples in accordance with QAS Section 8.2.1.

The following experiments are suggested to provide additional data on the reliability, reproducibility, and robustness of the TrueAllele Casework probabilistic genotyping software:

1. Mixture Study (n = 48)

TrueAllele Casework performance on two separate two person mixtures will be assessed with previously amplified Identifiler Plus data (including male/female mixing ratios of 1:1, 1:2, 1:5, 1:9, and 1:19). All two person mixture requests will be run in duplicate for 50K cycles (25K burn-in and 25K read out). The experiment will then be repeated for two separate three person mixtures (including mixture ratios of 1:1:1, 1:2:1, 1:5:1, 1:10:1, 1:2:3, 2:2:1, and 3:3:1). All three person mixture requests will be run in duplicate for 100K cycles (50K burn-in and 50K read out).

The resulting mixture weights, standard deviations, KL scores, and match statistics for all samples will be recorded for each known donor. Markov Chain Monte Carlo (MCMC) convergence will be evaluated with Gelman-Rubin statistics and by assessing all mixture weight Markov chain histograms. The accuracy of TrueAllele match statistics will be measured by comparing the data with CPI for the same samples. For ease of comparison, all weight of evidence will be expressed in log10 form. Specificity of the software will be evaluated by running approximately 17-18 known staff profiles against all two and three person mixtures, with the expectation that all resulting log (LR) match statistics will be negative for non-donor staff. Reproducibility and precision of the software will be evaluated by quantifying the variation (measured as standard deviation) between replicated match statistics.

2 person mixtures:

M/F 1	**M/F 2**
1:1	1:1
1:2	1:2
1:5	1:5
1:9	1:9
1:19	1:19

Two mixture sets n = 10
Duplicate TA requests n = 20

3 person mixtures:

MIX 1	**MIX 2**
1:1:1	1:1:1
1:2:1	1:2:1
1:5:1	1:5:1
1:10:1	1:10:1
1:2:3	1:2:3
2:2:1	2:2:1
3:3:1	3:3:1

Two mixture sets n = 14
Duplicate TA requests n = 28

2. Sensitivity Study (n = 32)

Single donor TrueAllele requests will be made from Identifiler Plus data derived from two separate serial dilutions (500, 250, 125, 62.5, 31.25, 15.6, 7.8, and 3.9pg). The accuracy of TrueAllele match statistics will be measured by comparing the data with CMP for the same samples. For ease of comparison, all weight of evidence will be expressed in log10 form. Reproducibility and precision of the software will be evaluated by quantifying the variation (measured as standard deviation) between replicated match statistics. Sensitivity will be assessed by quantifying the effect of diminished template samples on match statistics.

Two Serial dilution series:

500, 250, 125, 62.5, 31.25, 15.6, 7.8, 3.9pg

Two dilution series n = 16
Replicate TA requests n = 32

3. MCMC Burn-In/Read Out Study (n =42)

To evaluate the performance of TrueAllele Casework with differing MCMC burn-in/read out parameters, shorter and longer cycle times than those described in *Study 1* above (25/25K and 50/50K) will be tested. Specifically, one single source sample amplified with 250, 62.5, and 15.6pg will be run in duplicate with TrueAllele Casework at 5K/5K, 10K/10K, and 15K/15K. For mixture profiles, a portion of the two person mixture set (M/F 1 at mixing weights of 1:1, 1:2, and 1:5) will be run in duplicate at 75K/75K and 125K/125K. Match statistics, convergence, KL scores, and standard deviations will then be compared to the values obtained in *Study 1*. Additionally, a portion of the three person mixture set (MIX 1 at mixing weights of 1:1:1, 1:2:1, and 1:5:1) will be run in duplicate at 75K/75K and 125K/125K. Match statistics, convergence, KL scores, and standard deviations will also be compared to the values obtained in *Study 1*.

Note:

All Identifiler Plus samples were previously amplified from staff buccal swab extracts as part of the in-house Sorenson validation. All mixture samples had an approximate target DNA input of 1.00ng. All samples were run on a 3130*xl* genetic analyzer with 1.0uL DNA input in 9.0 HIDI/LIZ with 10 second injection times used for all TrueAllele Casework requests. All CPI and CMP statistics used for comparison purposes assume 50 RFU detection and 140 RFU stochastic thresholds. An alpha level of 0.05 will be used for all statistical tests. This work is intended to serve as a companion to the original NYSP FIC Developmental Validation Study (TrueAllele System for Forensic Casework STR DNA Data Interpretation), approved for use on May 2, 2013.

Budget: None needed. All data will be taken from in-house Identifiler Plus validation studies.

Timeline: The experiments outlined above combined with the final write-up and revisions to current protocols are expected to take six to eight weeks for completion.

Jay Caponera
Bioscience Casework
18, July 2013

II. Project Summary

Experiment 1

Two Person Mixture Study

The accuracy of TrueAllele Casework probabilistic genotyping software was assessed by comparing match results from a suite of two person mixtures to the corresponding CPI results. For all mixture sets tested, TrueAllele was able to recover more genetic information than the simpler inclusion method (CPI) and returned higher match statistics (mean of 9.17 log units). TrueAllele match results were also highly specific to the known mixture donors (mean donor log(LR) of 13.7), while always returning negative log(LR) values for non-donor references (mean non-donor log(LR) of -22.96). Known donor and non-donor separation (based in log units) was significant, with a mean log separation of 36.7 for all mixtures tested. The maximum individual non-donor log(LR) did not exceed -6.50. TrueAllele Casework results, including mixture weight inferences, KL statistics, and log(LR) values were also highly reproducible between replicated software runs. Convergence diagnostics for all two person mixtures were acceptable with the exception of one sample requiring a new request due to Markov chain failure from insufficient burn-in time.

Three Person Mixture Study

The two person mixture study described above was extended to an examination of three person mixtures. Similar to the two person mixture data, TrueAllele recovered more genetic information than CPI for all three person mixtures and yielded higher match statistics (mean of 5.31 log units). TrueAllele genotype inference was also highly specific for the three person mixture sets, with a mean donor/non-donor separation of 30.35 log units and a maximum non-donor log(LR) of -3.50. Mixture weight inferences, KL statistics, and log(LR) values were also reproducible between replicate TrueAllele runs. Convergence diagnostics for all three person mixtures were acceptable.

Experiment 2

Sensitivity Study

The accuracy of TrueAllele Casework was assessed by running two sets of serially diluted single source samples (3.9 to 500pg) and comparing the resulting match statistics to conditional match probability (CMP) statistics calculated for identical samples. Results indicate that at DNA input values of 125pg and higher, CMP and TrueAllele match statistics were approximately equal. However, TrueAllele was able to provide

significantly higher match information than CMP for lower template samples in the 15.6 to 62.5pg range. Data below 15.6pg was generally not reproducible or reliable with either method. TrueAllele specificity for known donors was also very high when examining samples with DNA input greater than or equal to 15.6pg. Below this level, TrueAllele failed to discriminate between known and non-donor reference samples, although in no instance was a positive log(LR) returned for a non-donor reference. KL statistics and log(LR) values showed high reproducibility between all replicated results.

Experiment 3

MCMC Burn-In/Read Out Study

An examination of shortened Markov Chain Monte Carlo (MCMC) cycle times revealed no significant difference in KL statistics or log(LR) values between 5K/5K and 25K/25K for single source samples. Longer cycle times (75K/75K and 125K/125K) for both two and three person mixture sets were also assessed, with results again indicating no significant difference in KL statistics or log(LR) values between short and long cycles. The equivalence of short and long cycle times with respect to the data quality documented in this study is concordant with similar research completed by Cybergenetics.

Jay Caponera
Bioscience Casework
16 July, 2013

III. Procedures and Methods

Experiment 1: Mixture Study (n = 48)

TrueAllele Casework performance on two separate two person mixtures was assessed using Identifiler Plus-amplified male/female mixture ratios of 1:1, 1:2, 1:5, 1:9, and 1:19. Original amplifications performed by Sorenson used a target DNA input of 1.0ng for each mixture sample. Two person mixture requests were created and run in duplicate for 50K total cycles (25K burn-in and 25K read out). The process was then repeated for two separate three person mixtures using mixture ratios of 1:1:1, 1:2:1, 1:5:1, 1:10:1, 1:2:3, 2:2:1, and 3:3:1. All three person mixture requests were created and run in duplicate for 100K cycles total (50K burn-in and 50K read out). For all mixture samples, the resulting mixture weights, mixture weight standard deviations, Kullback-Leibler (KL) divergence scores, Gelman-Rubin (GR) statistics for MCMC convergence, and match statistics for all samples were recorded for each known donor (see Note below on these metrics).

The accuracy of TrueAllele Casework match statistics was measured by comparing the data with CPI for all samples, with a peak detection threshold of 50 RFU and a stochastic threshold of 140 RFU specific to Identifiler Plus data used to calculate inclusion statistics. Since the log(LR) is a standard additive measure of information and can be used to compare different DNA methods, all CPI values were converted to log values for comparison to TrueAllele log(LR) values. Information gain between statistical methods was calculated as log(TA) – log(CPI). Specificity of the software was evaluated by running 18 non-donor staff profiles against all two person mixtures, and 17 non-donor staff profiles against all three person mixtures. Reproducibility and precision of the software were evaluated by quantifying the variation (measured as standard deviation) between replicated match statistics and KL scores. Monte Carlo Markov Chain (MCMC) convergence was evaluated with GR statistics and by visually assessing all mixture weight Markov chain histograms and Markov chain plots.

Experiment 2: Sensitivity Study (n = 32)

Single donor TrueAllele Casework requests were made in duplicate from Identifiler Plus data derived from two separate serial dilutions (500, 250, 125, 62.5, 31.25, 15.6, 7.8, and 3.9pg), with all dilutions below 125pg considered low template (LT-DNA) samples. The accuracy of TrueAllele Casework match statistics was measured by comparing the data with CMP. Specificity and sensitivity were assessed by quantifying the effect of diminished template samples on match statistics, measured specifically as mean

separation between known donor and non-donor log(LR) values. Reproducibility and precision of the software were evaluated by quantifying the variation (measured as standard deviation) between replicated log(LR) values and KL scores.

Experiment 3: MCMC Burn-In/Read Out Study (n = 42)

To evaluate the possibility of shortening MCMC cycle times for single source samples, one single source sample amplified with 250, 62.5, and 15.6pg was run in duplicate with TrueAllele Casework at 5K/5K, 10K/10K, and 15K/15K and the results were compared to the 25K/25K results from the identical samples from Experiment 2 listed above. To evaluate longer cycle times for mixture profiles, a portion of the two person mixture set (M/F 1 at mixing weights of 1:1, 1:2, and 1:5) was run in duplicate at 75K/75K and 125K/125K. Log(LR) match statistics, KL scores, and mixture weight standard deviations were then compared to the 25K/25K values obtained in Experiment 1. A portion of the three person mixture set (MIX 1 at mixing weights of 1:1:1, 1:2:1, and 1:5:1) was also run in duplicate at 75K/75K and 125K/125K. Match statistics, KL scores, and mixture weight standard deviations were then compared to the 50K/50K values obtained from Experiment 1.

Note:

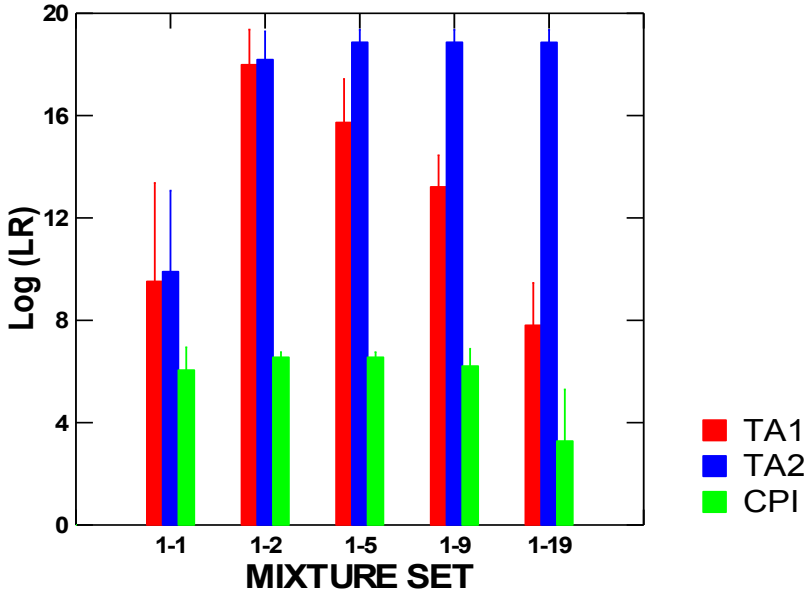
Cybergenetics TrueAllele Casework version 3.3.4764.1 (7-Nov-2012) was used for all validation experiments listed above. All TrueAllele requests used data originally run on a 3130*xl* using 10 second injection times. Gelman-Rubin (GR) convergence statistics and Kullback-Leibler (KL) statistics for profile information content were calculated by TrueAllele (see *Section IX. Glossary of Terms* for definitions), and may be viewed in the Report Module of the software. GR scores under 1.5 typically indicate acceptable mixture weight convergence based on vendor experience (pers. comm.). All statistical tests (ANOVA, non-parametric Kruskal Wallis, and regression analyses) were performed in Systat v. 13.1 with an alpha level of 0.05. CPI and CMP statistics were calculated separately with Popstats. All TrueAllele and CMP statistics were calculated using a theta value of 0.01. All graphs were created with Systat v. 13.1, and all tables were created with Excel.

IV. Experiment 1 Results: Two Person Mixture Study

A. Accuracy

In all two person mixture ratios tested, match statistics for the separately inferred contributors using TrueAllele Casework were greater than those calculated with CPI (Figure IV.A.1.), indicating that probabilistic genotyping more effectively preserves identification information. The mean match statistic was 12.85 for contributor 1 (TA1 below), 16.94 for contributor 2 (TA2), and 5.73 for all donors with CPI. This equates to an overall mean information gain of 9.17 log units (over 1 billion) for probabilistically inferred TrueAllele genotypes over the traditional CPI inclusion method. Examination of pairwise differences with a t-test shows a statistically significant improvement in log (LR) values over CPI ($p < 0.001$). From the data shown below, the information gain with TrueAllele Casework is greater when mixture ratios diverge from 1:1, although even at equal mixture weights the information gain over CPI is still 3.15 log units (over 1,000). The finding here with Identifiler Plus data is in accord with the 2011 JFS paper "Validating TrueAllele Mixture Interpretation" in which a mean information gain of 2.5 log units was reported over CPI for 1:1 mixtures with Profiler Plus/COfiler.

Figure IV.A.1. Comparisons of log(LR) match information as a function of mixture ratio. Mean values from two separate two person mixture sets are shown below; all TrueAllele samples were solved in duplicate with TA1 and TA2 denoting the separately inferred contributors for the two person mixture sets. Error bars represent one standard deviation.



B. Specificity

Data from the two person mixture set 1 examination show a mean separation of 35.54 log units (decillion) between known donor and non-donor reference profiles, and a mean separation of 37.86 log units (undecillion) for mixture set 2 (Figures IV.B.1. and IV.B.2). For all mixtures tested, non-donor log (LR) values were negative with maximum values of -9.26 in mixture set 1 and -6.76 in mixture set 2. Further, all known donor log (LR) values were positive, with individual minimums of 5.87 in mixture set 1 and 6.04 (million) in mixture set 2. The data indicate that TrueAllele Casework is highly specific and can easily discriminate between matching and non-matching reference genotypes.

Figure IV.B.1. Dot plot showing the specificity of TrueAllele Casework by reference sample. Mean values from the two person mixture set 1 are shown below for duplicate TrueAllele requests; reference donors 4 and 18 (far right) were used in the creation of all mixture ratios from this mixture set. Error bars represent one standard deviation; dashed line is set at zero.

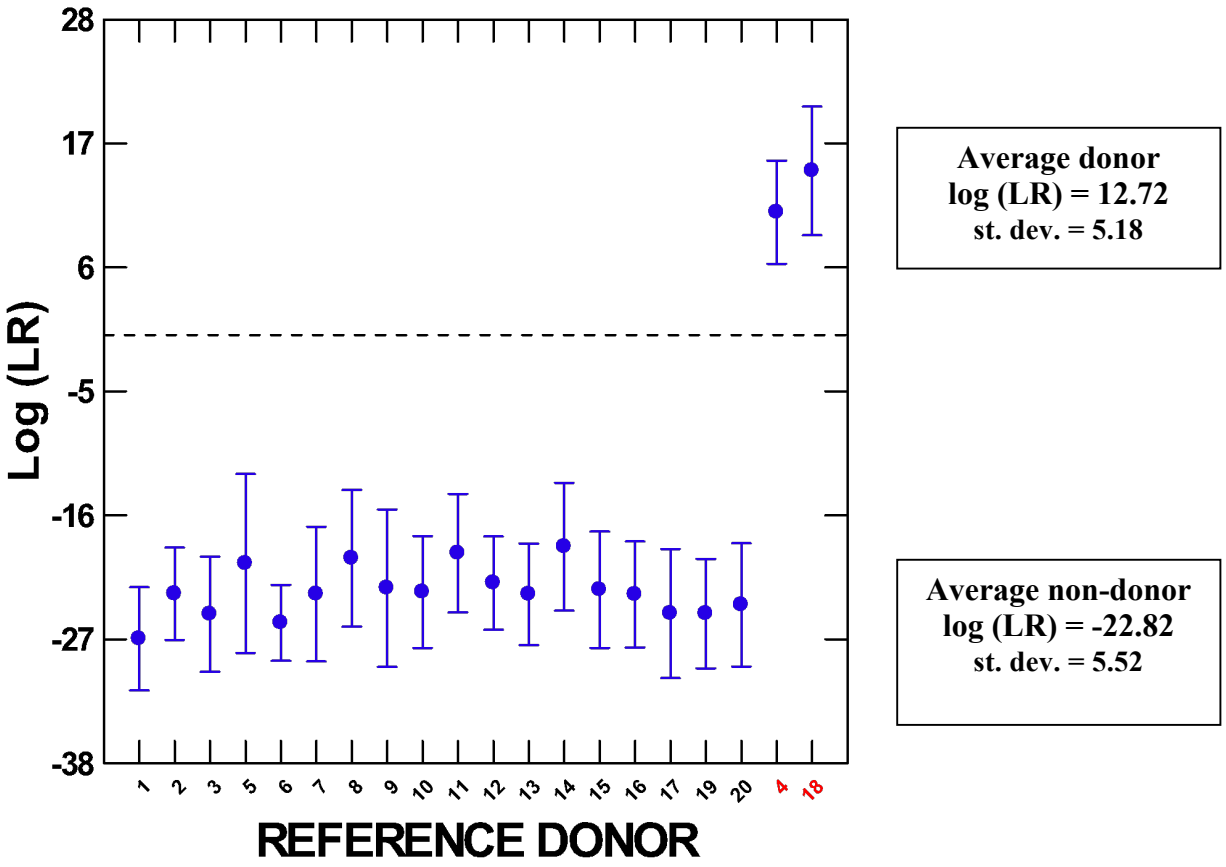
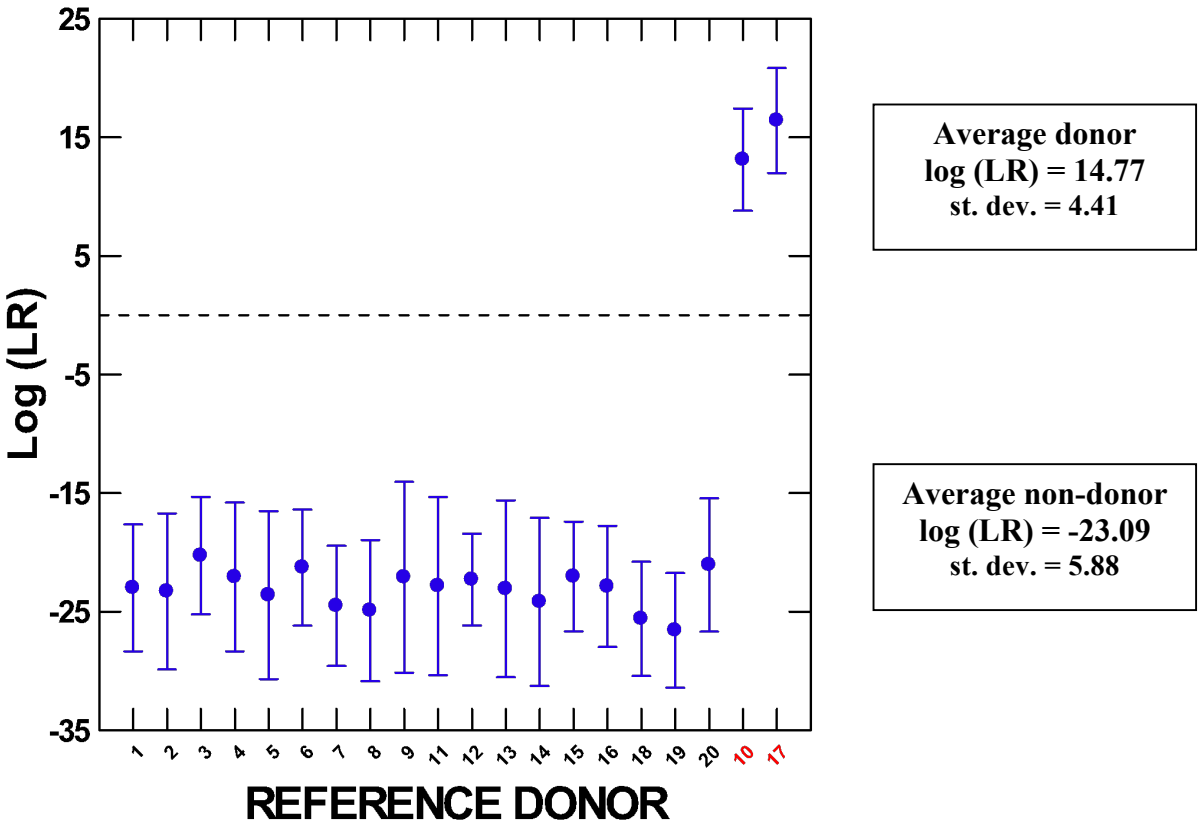


Figure IV.B.2. Dot plot showing the specificity of TrueAllele Casework by reference sample. Mean values from the two person mixture set 2 are shown below; reference donors 10 and 17 (far right) were used in the creation of all mixture ratios from this mixture set. Error bars represent one standard deviation; dashed line is set at zero.



C. Reproducibility – Mixture Weights and Log(LR)

The reproducibility of TrueAllele Casework was assessed by running duplicate identification requests for all two person Identifiler Plus mixtures. The resulting inferred mixture weights for both test mixture sets correlated strongly with the theoretical mixture weights created by Sorenson, with a mean standard deviation of 0.023 (Table IV.C.1 and see two person mixture Match Table). Additionally, log (LR) match statistics for both contributors were highly reproducible, with a maximum standard deviation of 1.720 (Table IV.C.2.). The largest variations were associated with mixture profiles of approximately equal mixture weights and with the minor reference donor. KL statistics (divergence of inferred profile from prior distribution) were also reproducible across all mixture ratios, indicating that TrueAllele consistently infers highly informative genotypes from the mixture samples tested (Figure IV.C.1.).

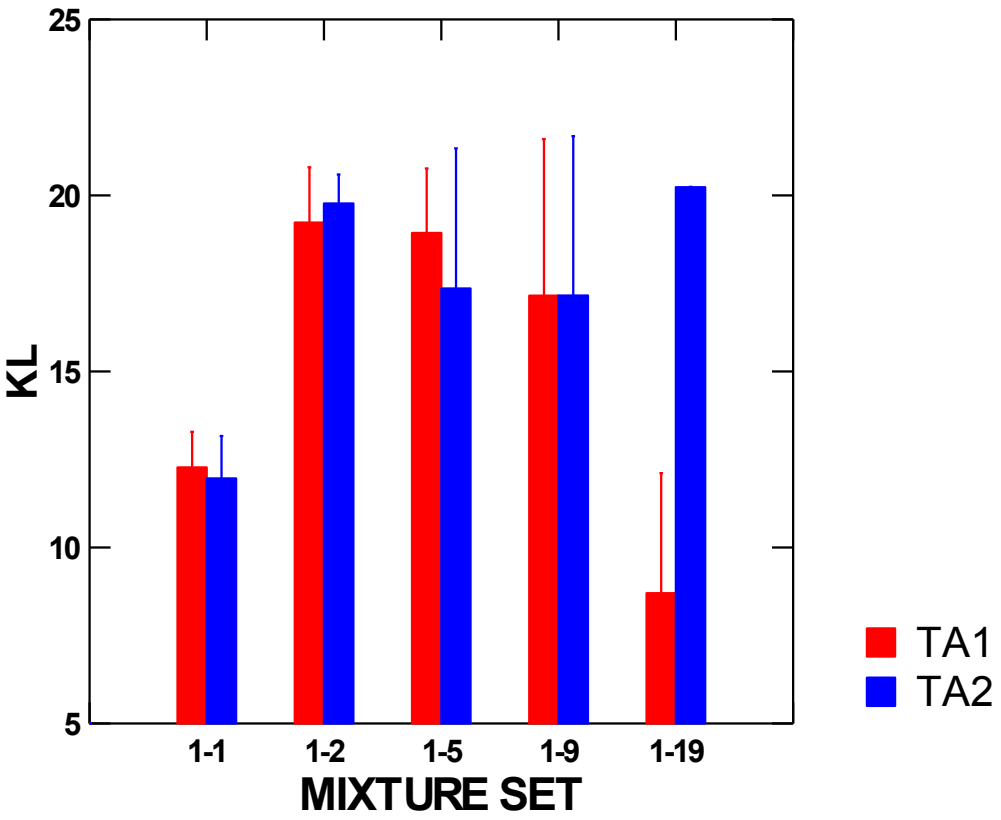
Table IV.C.1. Comparison of theoretical two person mixture weights to those inferred by TrueAllele Casework; mixture weights shown are averaged across all respective mixture categories for mixture sets 1 and 2.

Theoretical Mixture Weight Ratios	TrueAllele Inferred Mixture Weights (as ratios)	TrueAllele Inferred Mixture Weight	Mean Standard Deviation
1:1	1:1.00	.501/.499	0.038
1:2	1:2.30	.310/.690	0.020
1:5	1:6.09	.141/.859	0.021
1:9	1:10.1	.090/.910	0.018
1:19	1:24.6	.039/.961	0.017

Table IV.C.2. Log (LR) match statistics and associated standard deviations listed by contributor across both mixture sets tested.

Theoretical Mixture Weight Ratios	Mixture Set	Log (LR) Donor 1	Standard Deviation Donor 1	Log (LR) Donor 2	Standard Deviation Donor 2
1:1	1	6.742	0.041	7.623	0.001
1:2	1	17.014	0.010	17.418	0.150
1:5	1	14.742	1.481	18.509	0.000
1:9	1	12.485	1.022	18.509	0.000
1:19	1	8.351	1.265	18.509	0.000
1:1	2	12.297	0.444	12.179	0.486
1:2	2	18.710	0.001	19.218	0.000
1:5	2	16.714	0.076	19.218	0.000
1:9	2	13.947	0.064	19.218	0.000
1:19	2	7.253	1.720	19.213	0.004

Figure IV.C.1. Bar graph of KL scores as a function of mixture ratio. Mean values from two separate two person mixture sets are shown below; all TrueAllele samples were solved in duplicate with TA1 and TA2 referring to the separate contributors for the two person mixture sets. Error bars represent one standard deviation.



D. MCMC Convergence

Convergence was assessed with Gelman Rubin statistics (GR hereafter) and by visual inspection of the mixture weight Markov chains and histograms. For all two person mixtures tested, the mean GR was 1.067 with a standard deviation of 0.132 (GR values between 1.0 and 1.2 may indicate acceptable convergence). From the 48 mixture requests created and run, only one (MF1_1-19) exhibited poor Markov chain convergence with a GR statistic of 2.054 for both contributors, and was not used in the preceding analysis (Figure IV.D.1). A GR statistic of this magnitude is indicative of between-chain variance dominating within-chain variance and therefore an ineffective sampling of the target distribution. The inferred mixture weight for this 1:19 sample was .738/.262 (1:2.82), whereas the mixture weights for two additional replicates of this sample with acceptable convergence had a mean inferred weight of .039/.961 (1:24.6). The histogram for the failed sample (below, left) shows a bimodal distribution and excessively narrow mixture weight peaks. The associated Markov chain (below, right) shows a failure to reach a stationary distribution until the read out is approximately 40% complete.

GR statistics are not entirely diagnostic of proper convergence however, as seen in the replicate of sample MF1_1-19 (MF1_1-19A; see Figure IV.D.2). The appearances of both the histogram and the Markov chain have improved dramatically, although the GR statistic is still greater than 1.2 (1.572 for both contributors). However, even with a GR value higher than 1.2, the mixture weights, standard deviations, KL values, and match statistics are all similar to the additional replicate (MF1_1-19_copy) that exhibited a GR of 1.001 for both contributors (Table IV.D.1). This finding underscores the need to evaluate convergence by viewing the mixture weight histograms and Markov chains in addition to GR statistics, since such statistics are not entirely diagnostic (see Figure IV.D.3. for all histograms and Markov chains for each two person mixture set).

Figure IV.D.1. Histogram and Markov Chain from mixture sample MF1_1-19 showing poor convergence due to insufficient burn-in cycle time. Note that the chains eventually reach stationary distributions at approximately 10,000 cycles into the read out phase.

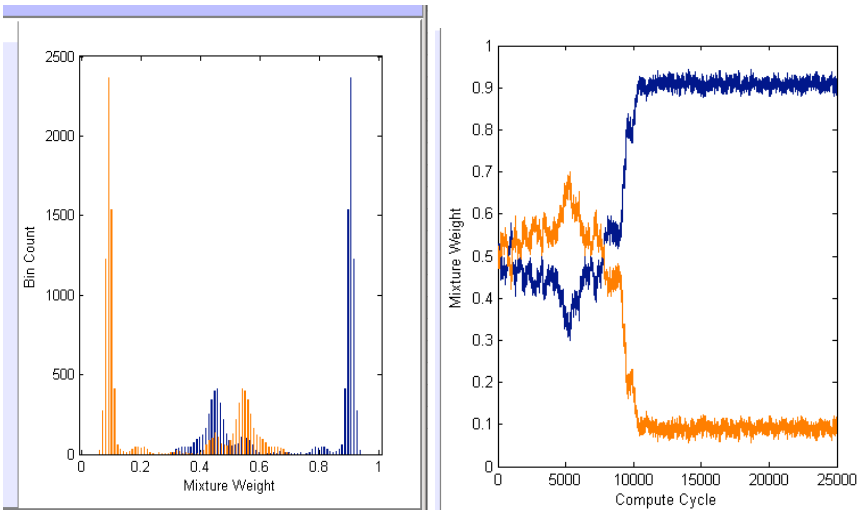


Table IV.D.1. Match table showing results for poorly converged sample MF1_1-19 (request Q41/Q42; see red starred items below) and two replicates of the same sample (requests Q3/Q4 and Q5/Q6). Note the inaccurate mixture weight and elevated standard deviation for Q41/Q42 as compared to the replicated samples and the failure to match known reference sample 4.

Match

File

	Evidence	Contrib...	N Contrib	Weight	Std Dev	KL	18	4
Q3	MF1_1-19A	1	2	0.040	0.022	8.500	-10.356	9.245
Q4	MF1_1-19A	2	2	0.960	0.022	20.234	18.509	-30.000
Q5	MF1_1-19_copy	1	2	0.038	0.008	12.620	-17.563	7.456
Q6	MF1_1-19_copy	2	2	0.962	0.008	20.234	18.509	-30.000
Q41	MF1_1-19	1	2	0.738	0.346	18.198	17.431	-25.014
Q42	MF1_1-19	2	2	0.262	0.346	13.775	8.498	-2.097

Figure IV.D.2. Histogram and associated Markov Chain from replicate of mixture sample MF1_1-19A showing improved convergence, even with a GR statistic greater than 1.2.

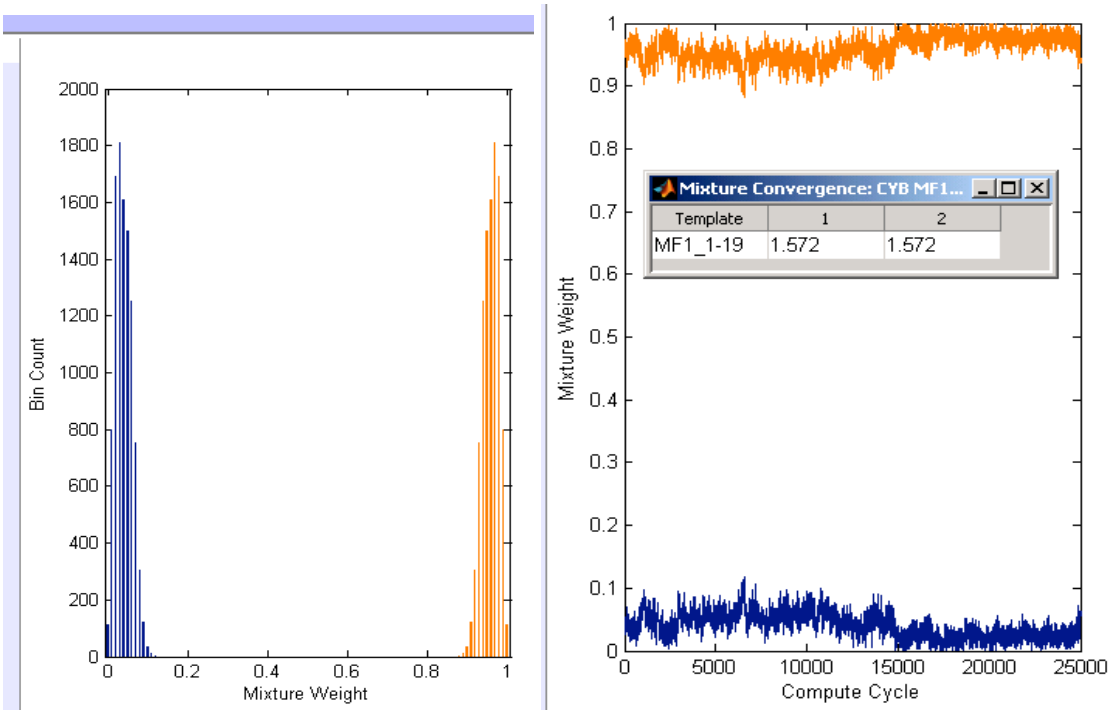
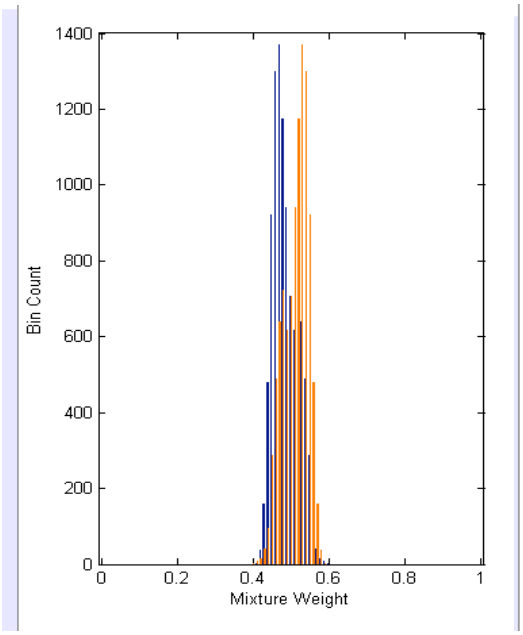
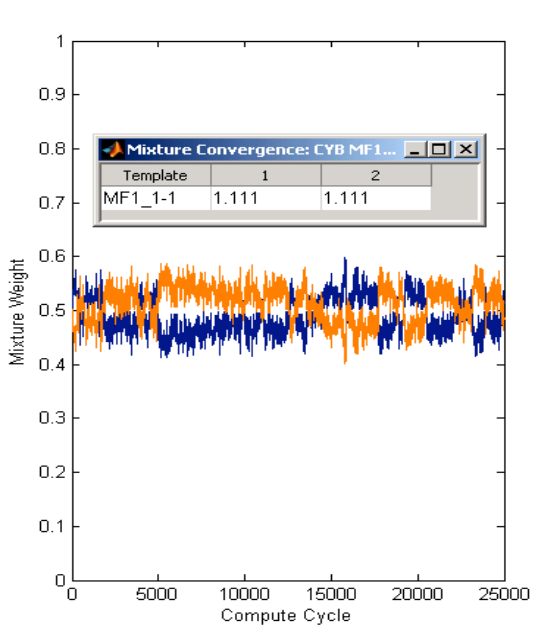
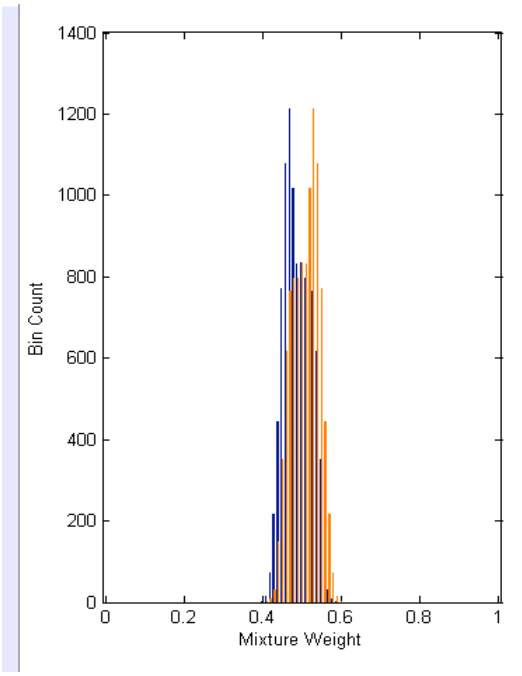


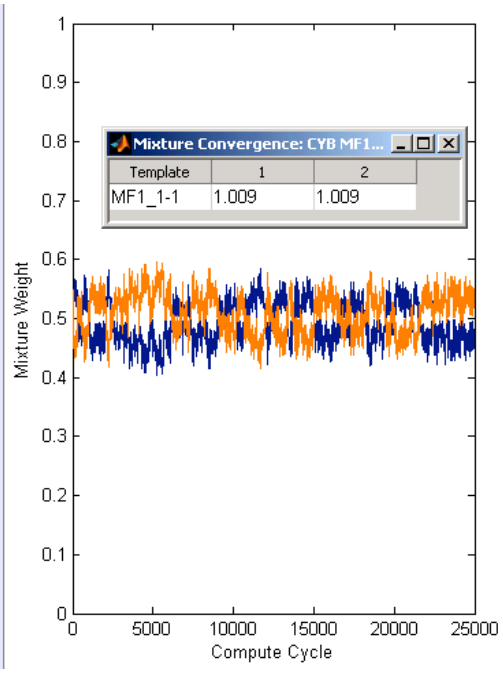
Figure IV.D.3. Histograms and associated Markov Chains from all two person mixture samples; GR convergence scores are indicated in the Markov chains at right.



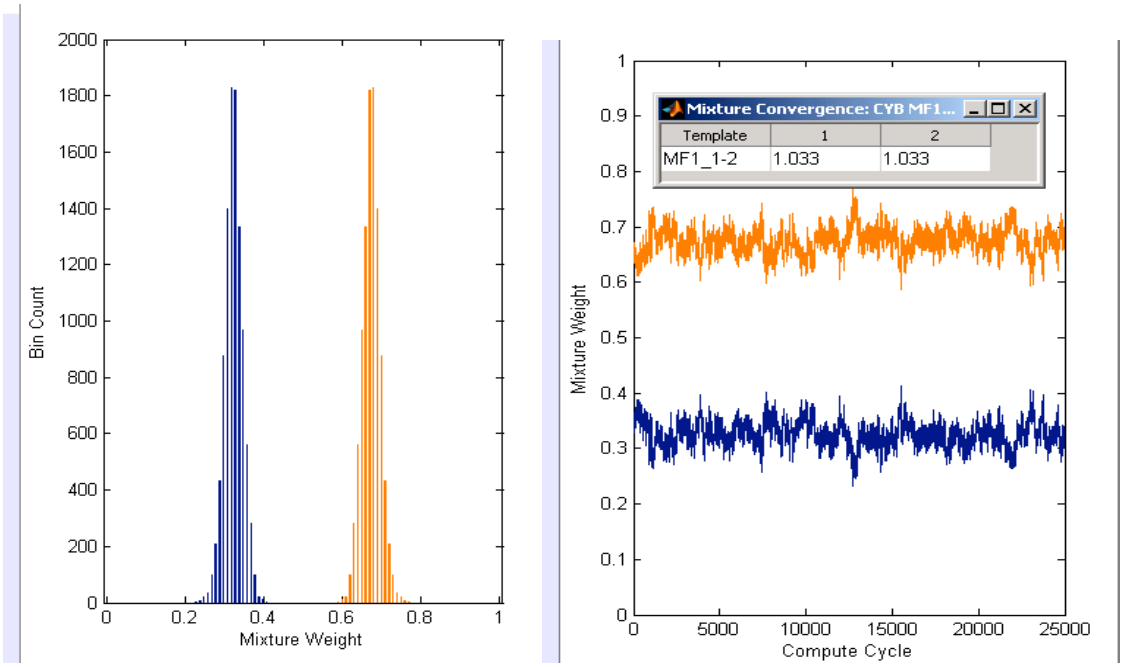


Mix Set 1
1:1

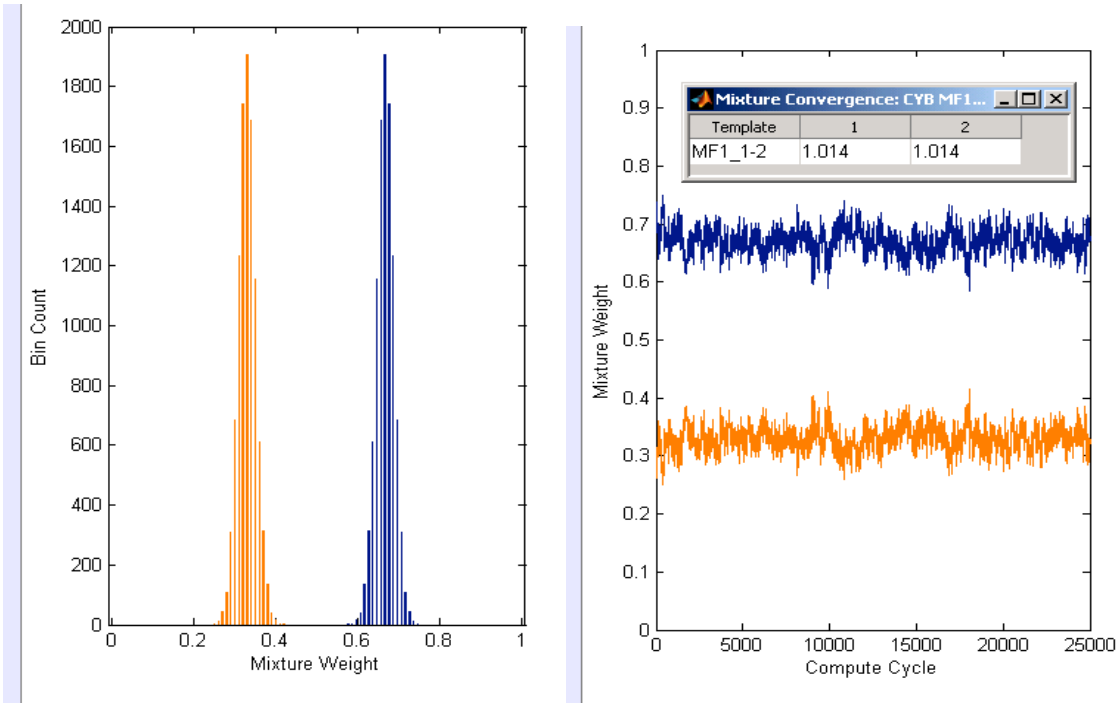




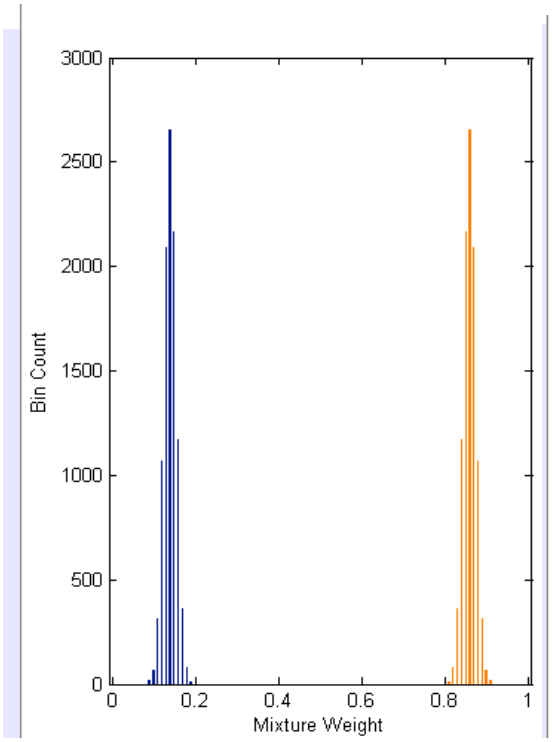
Mix Set 1
1:1
replicate

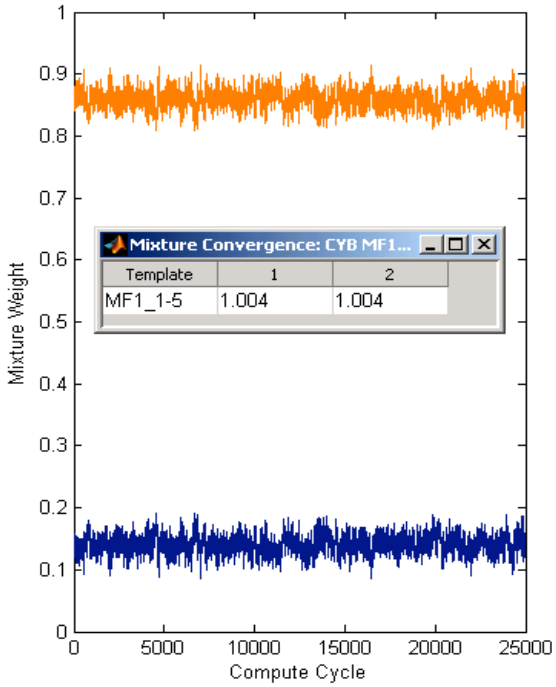


Mix Set 1 1:2

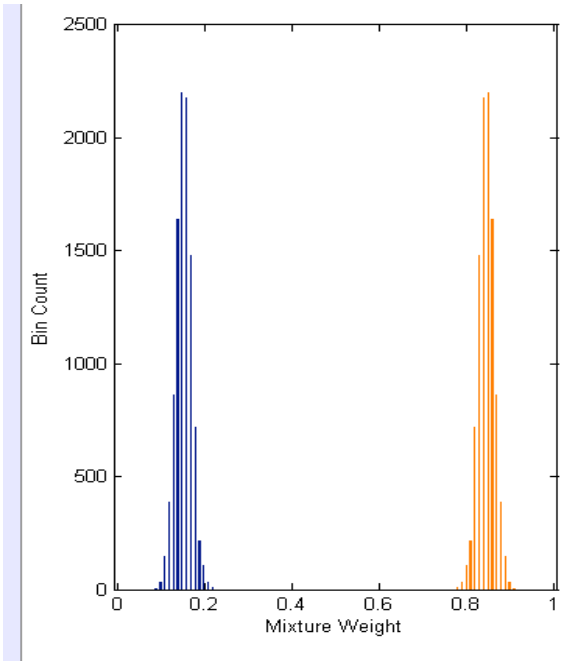


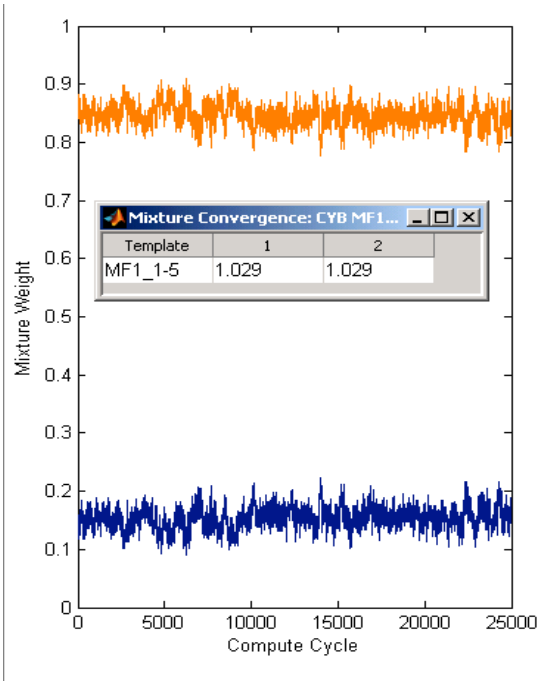
Mix Set 1 1:2 replicate



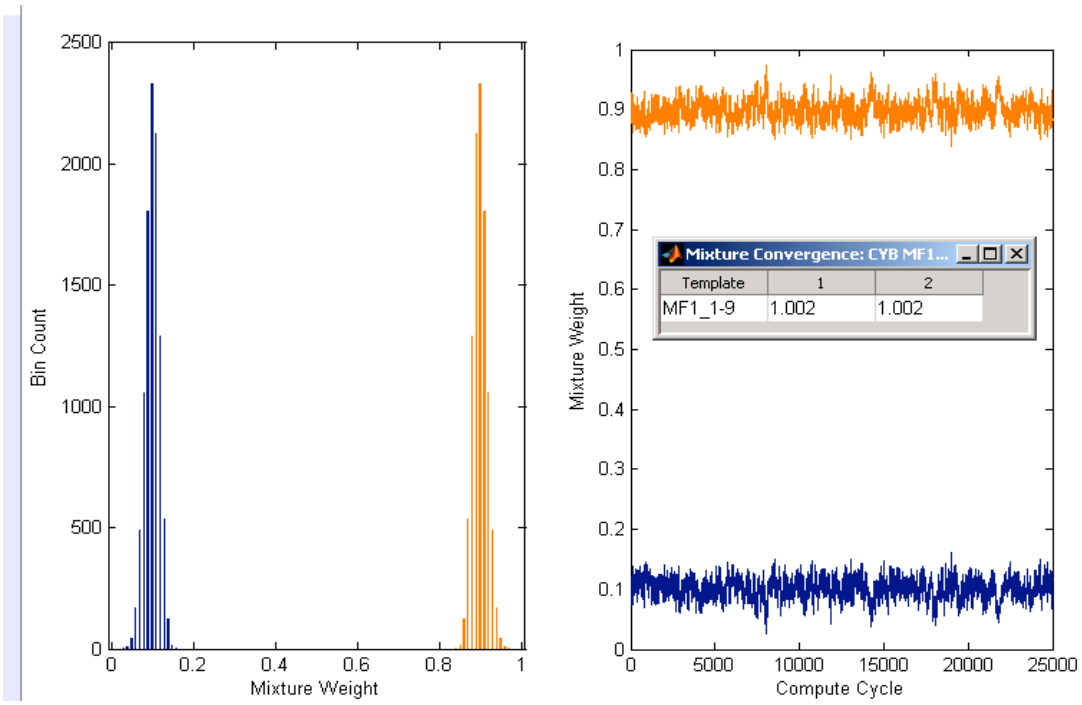


Mix Set 1 1:5

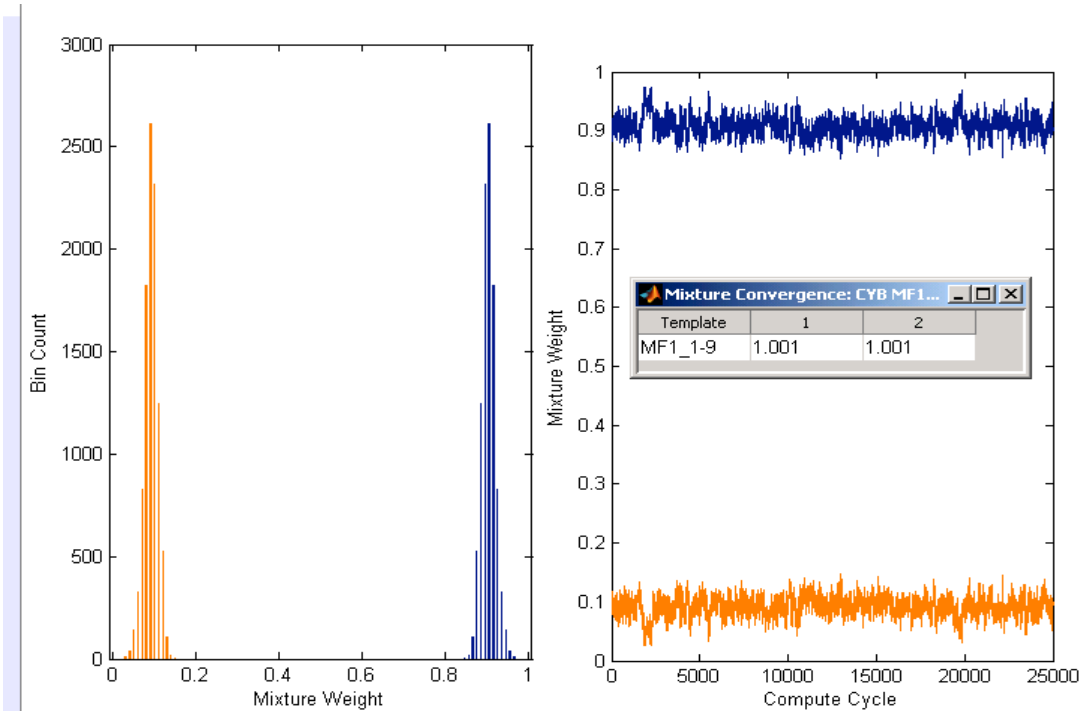




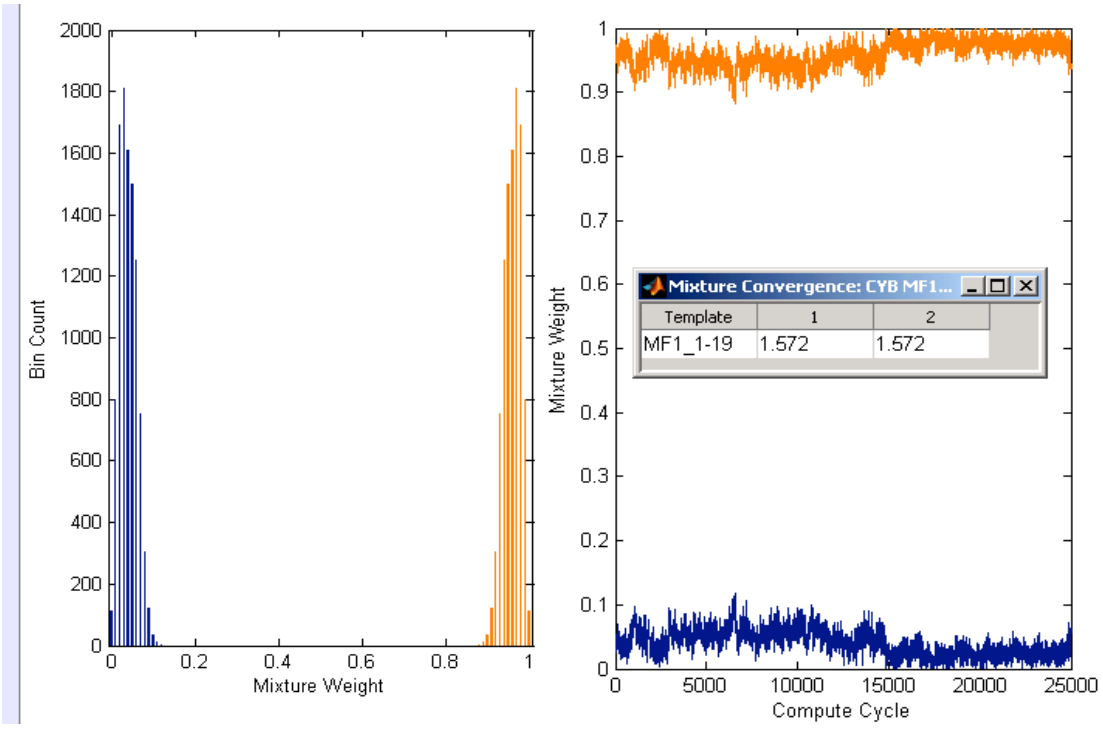
Mix Set 1 1:5 replicate



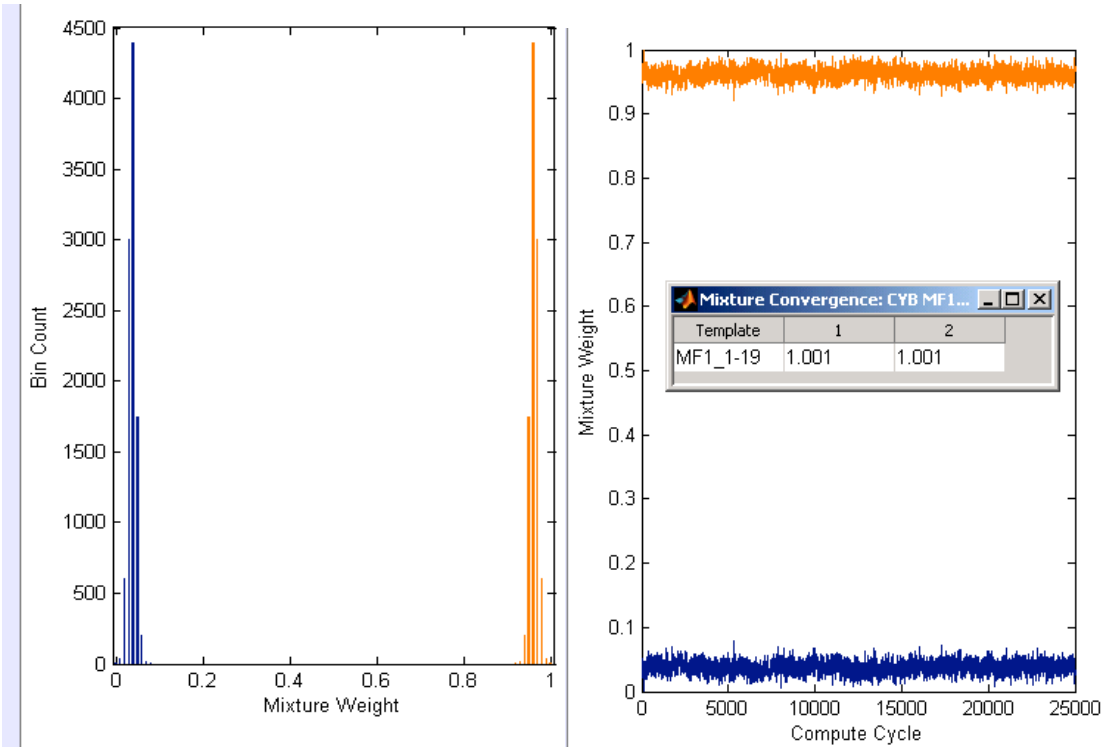
Mix Set 1 1:9



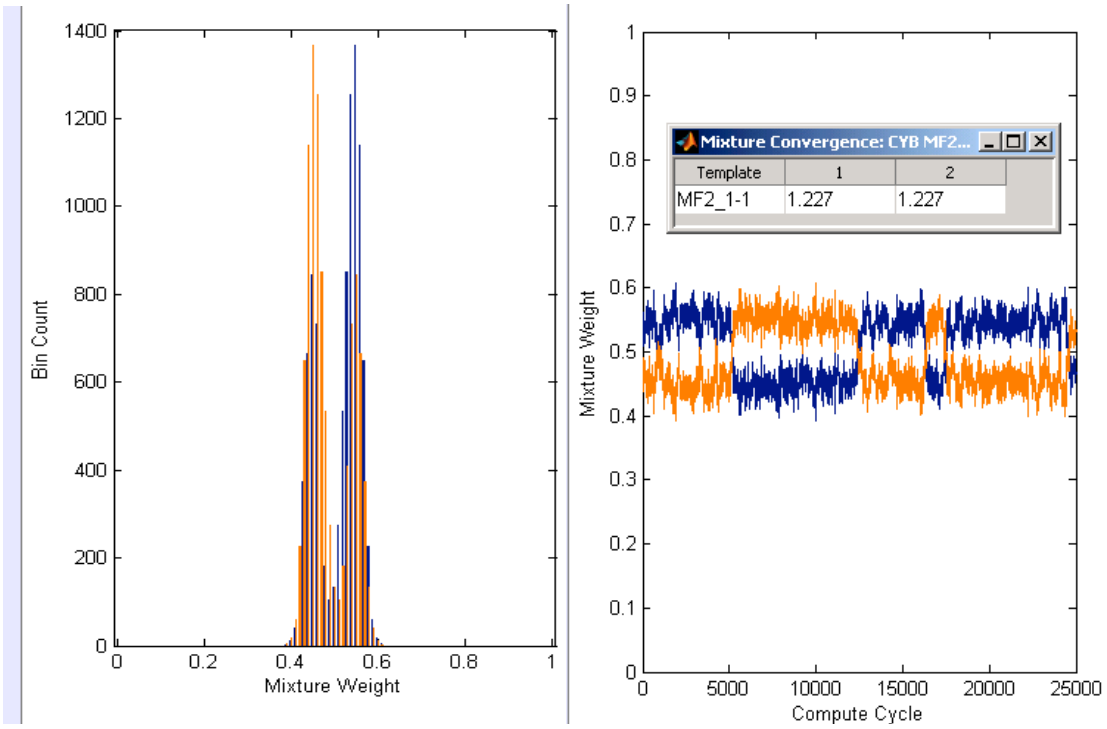
Mix Set 1 1:9 replicate



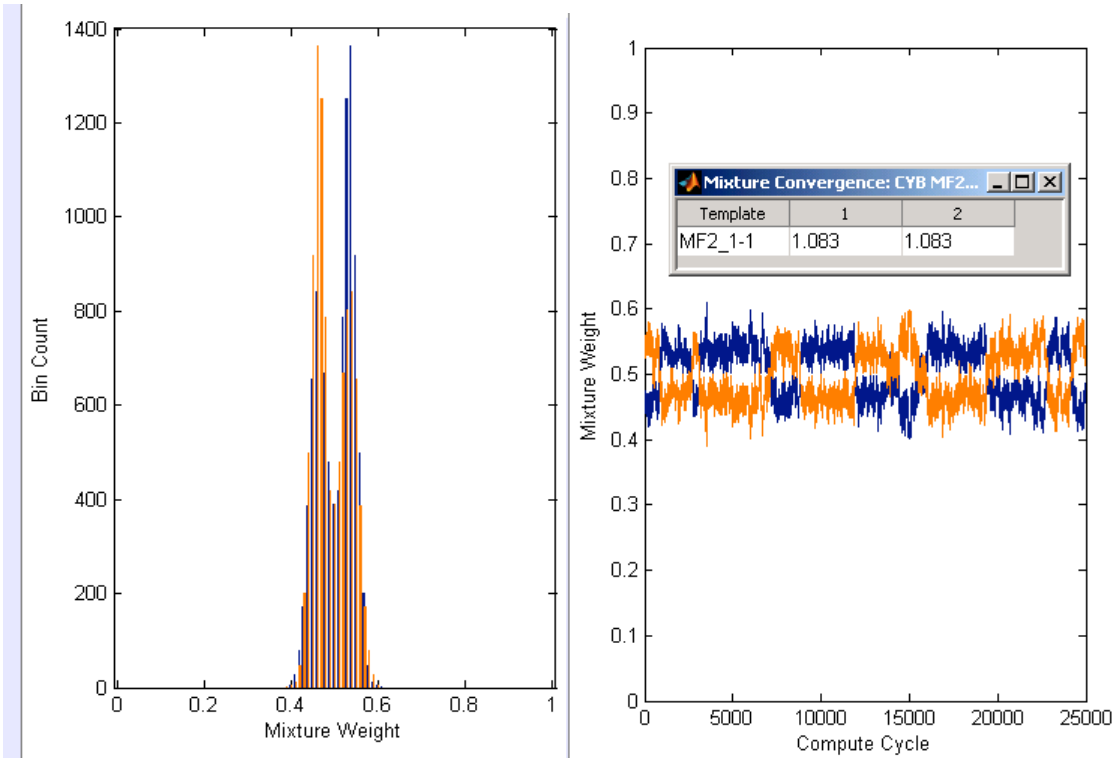
Mix Set 1 1:19



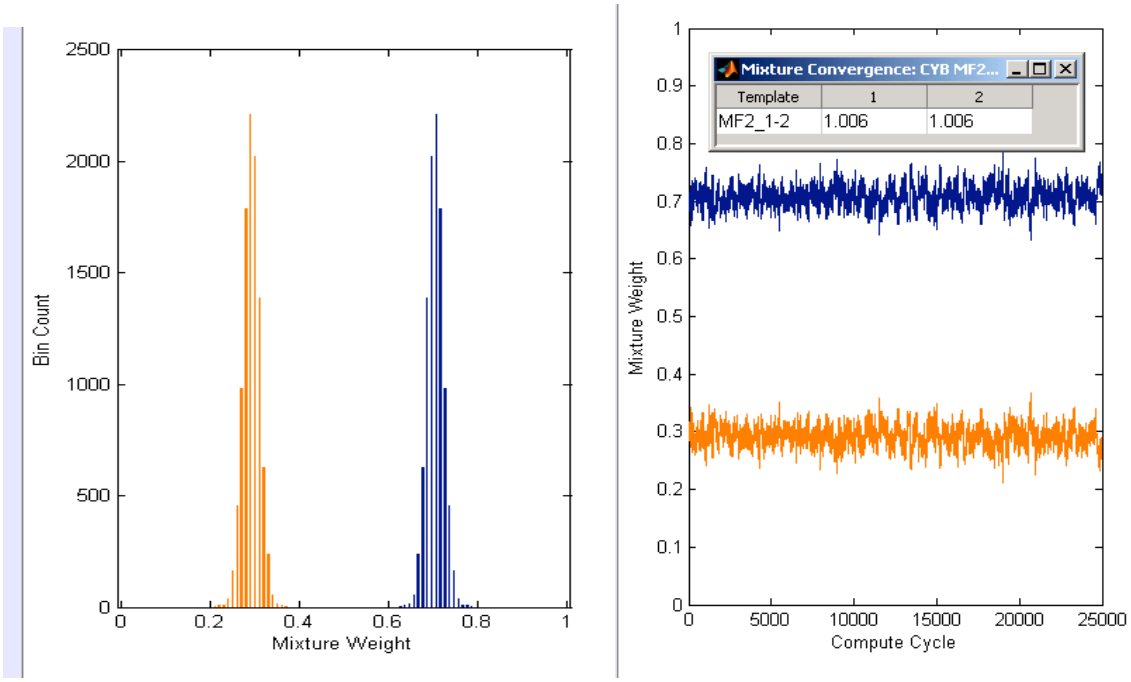
Mix Set 1 1:19 replicate



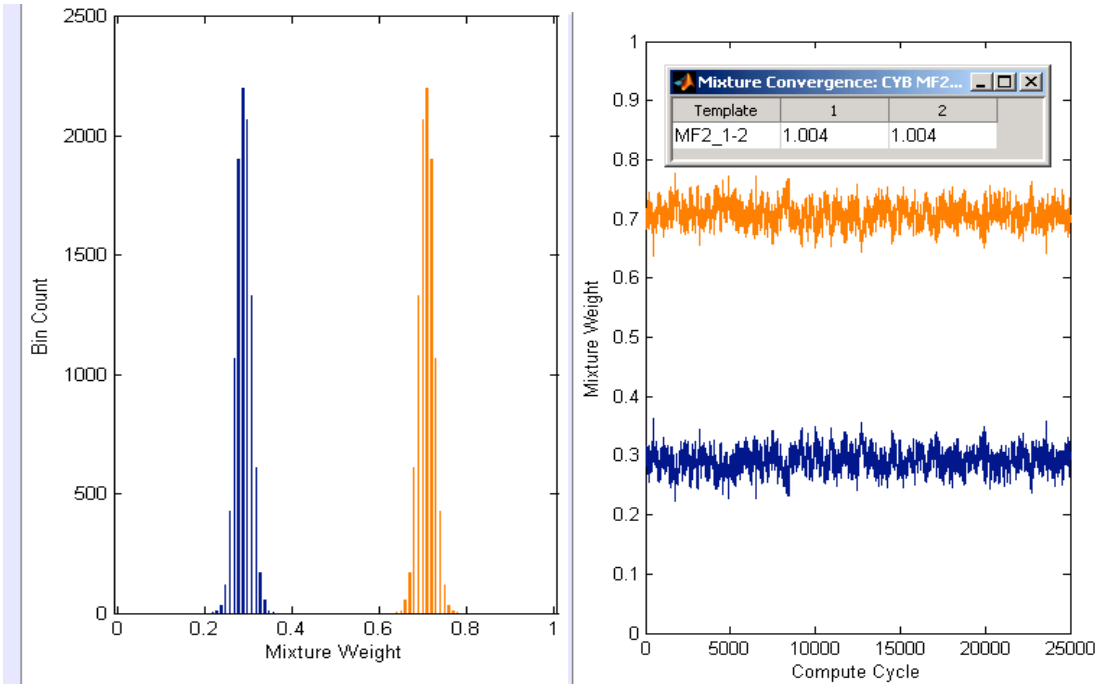
Mix Set 2 1:1



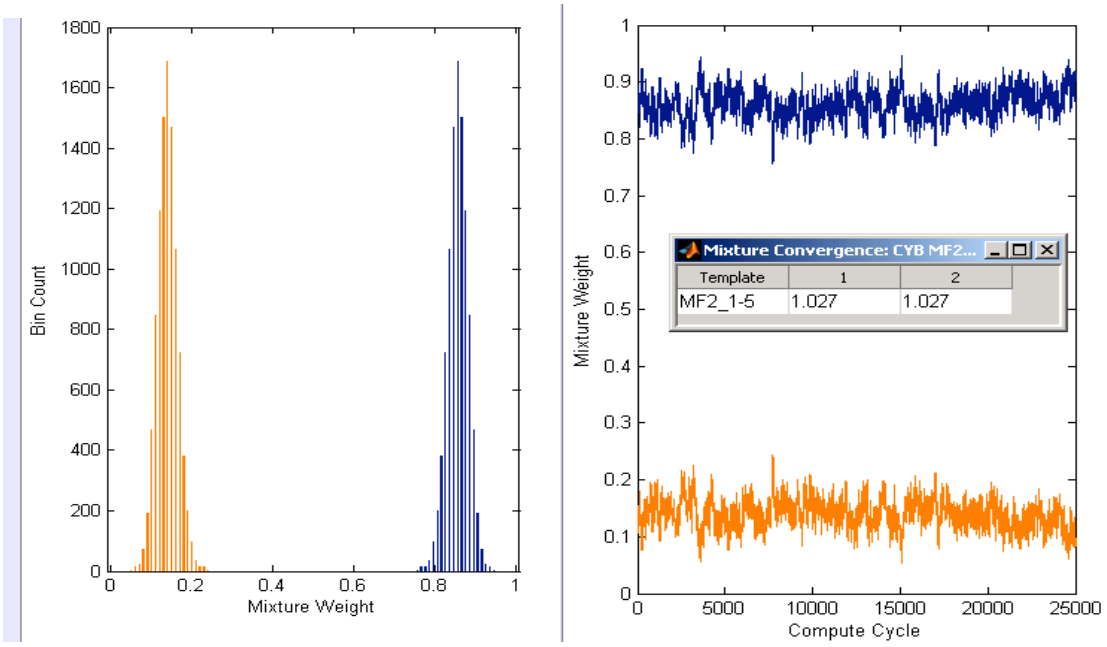
Mix Set 2 1:1 replicate



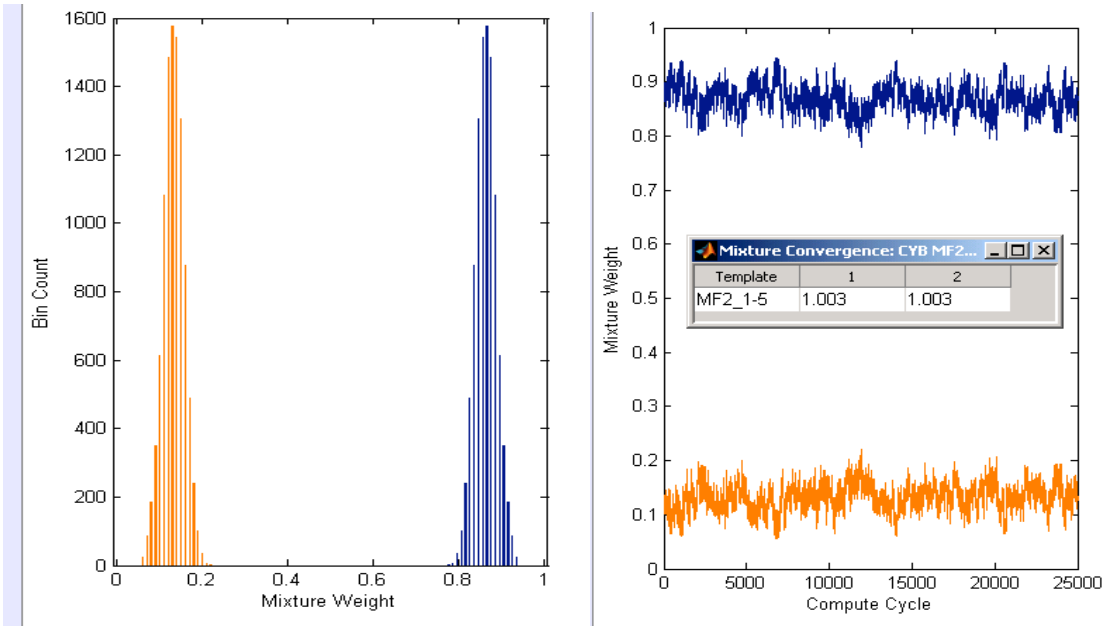
Mix Set 2 1:2



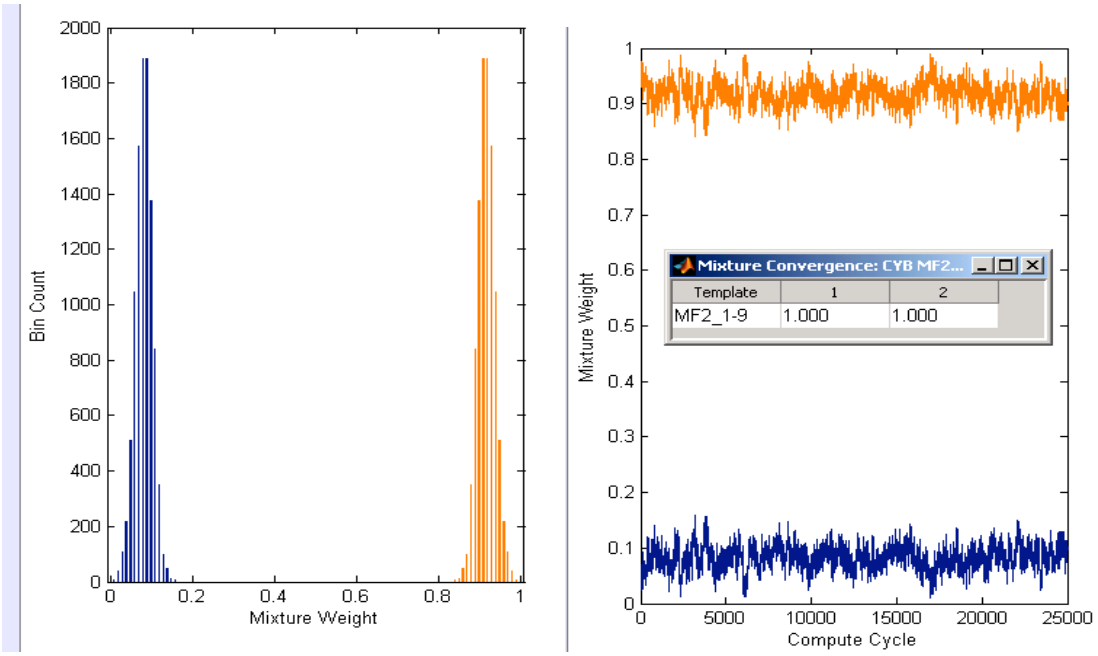
Mix Set 2 1:2 replicate



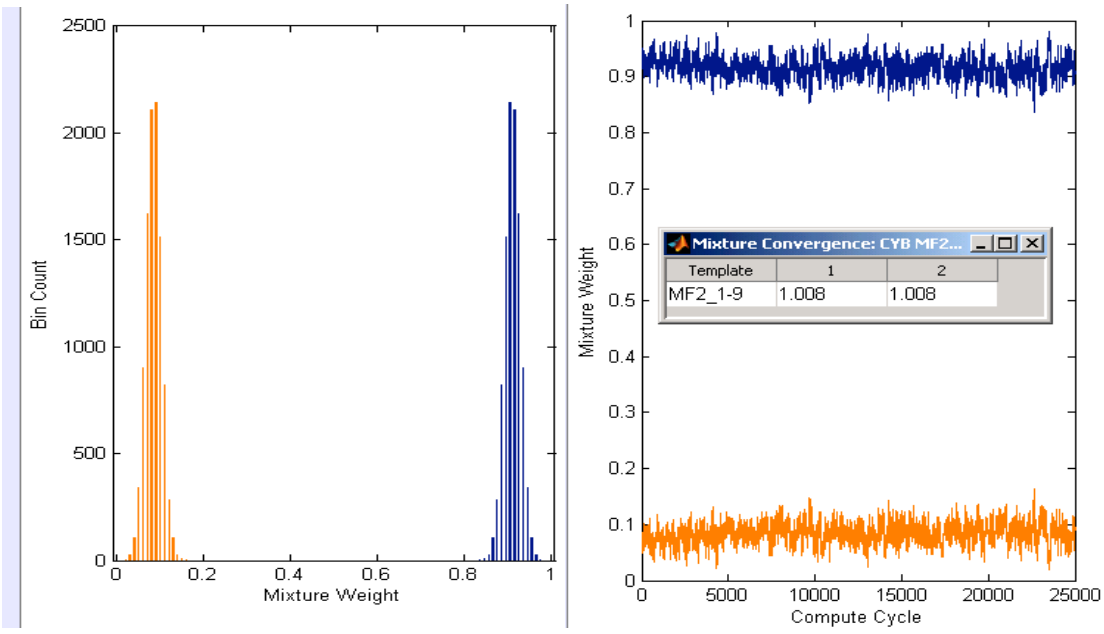
Mix Set 2 1:5



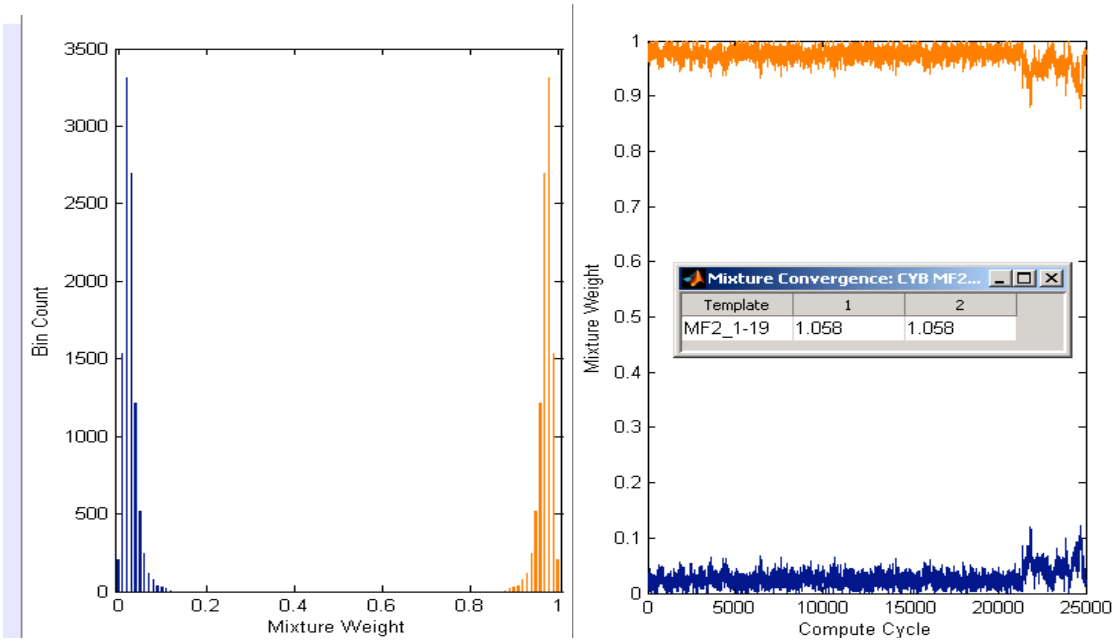
Mix Set 2 1:5 replicate



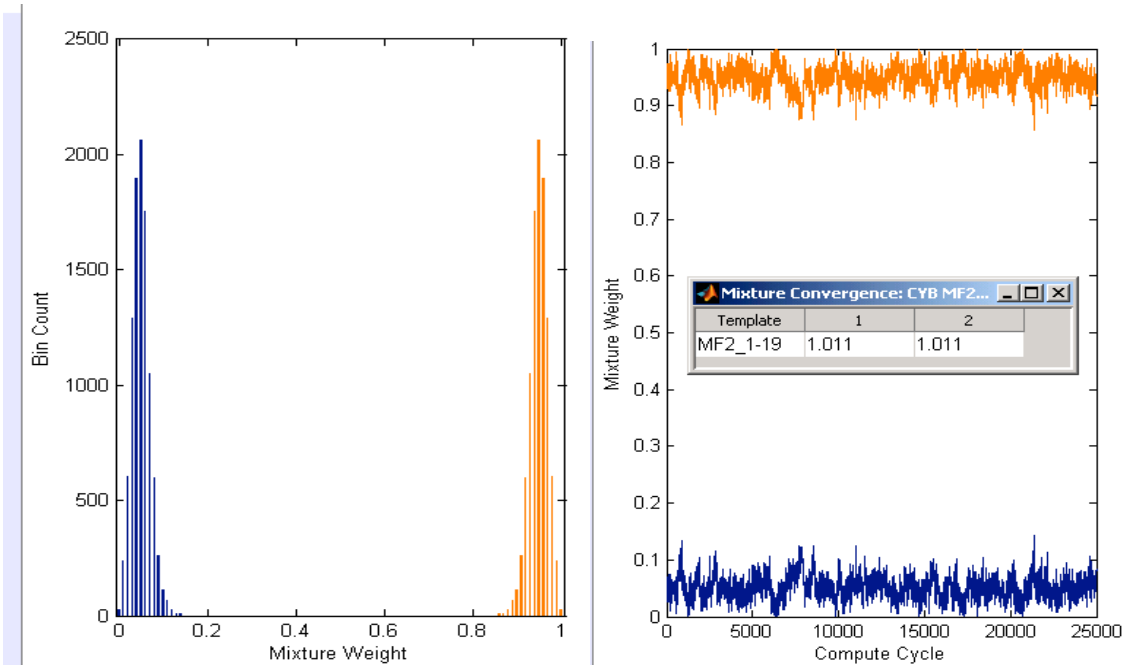
Mix Set 2 1:9



Mix Set 2 1:9 replicate



Mix Set 2 1:19



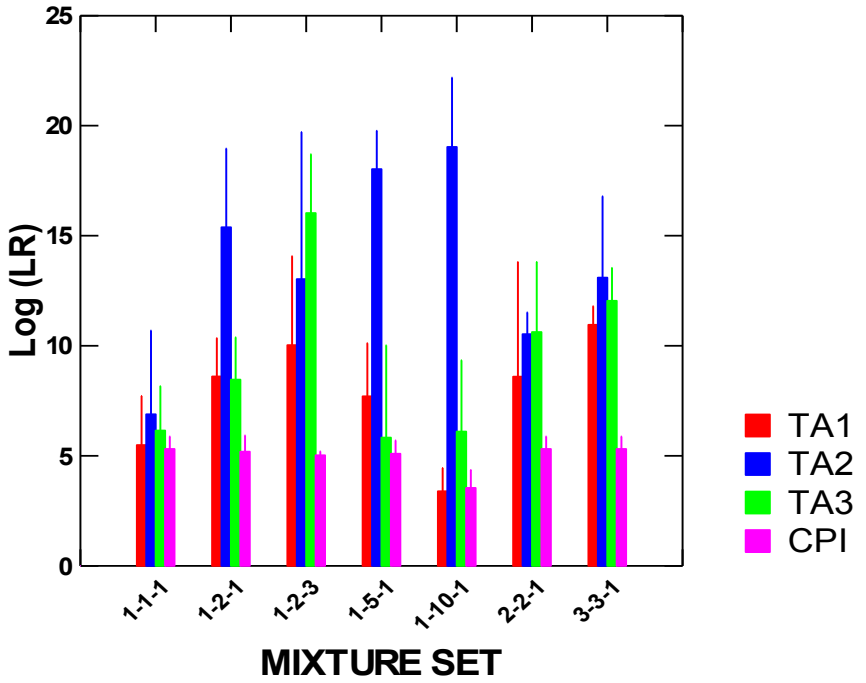
Mix Set 2 1:19 replicate

V. Experiment 1 Results: Three Person Mixture Study

A. Accuracy

Similar to the two person mixture dataset, all three person match statistics for the separately inferred contributors using TrueAllele Casework were greater than those calculated with CPI, (Figure V.A.1.) again indicating that probabilistic genotyping is more effective at preserving identification information. The mean match statistic was 7.82 for contributor 1 (TA1 below), 13.71 for contributor 2 (TA2), 9.32 for contributor 3 (TA3), and 4.97 for all donors with CPI. This equates to an overall mean information gain of 5.31 log units for probabilistically inferred TrueAllele genotypes over the traditional CPI inclusion method. Examination of pairwise differences with a t-test shows a statistically significant improvement in log(LR) values over CPI ($p < 0.001$).

Figure V.A.1. Comparisons of log(LR) match information as a function of mixture ratio. Mean values from two separate three-person mixture sets are shown below; all TrueAllele samples were solved in duplicate with TA1, TA2, and TA3 referring to the separate contributors for the three person mixture sets. Error bars represent one standard deviation.



B. Specificity

Specificity of the TrueAllele Casework software was further assessed with a suite of three person mixtures using known donor and non-donor reference profiles. Data from the three person mixture set 1 examination show a mean separation of 28.17 log units (octillion) between known donor and non-donor reference profiles, and a mean separation of 25.16 log units (septillion) for mixture set 2 (Figures V.B.1. and V.B.2.). For all three person mixtures tested, non-donor log(LR) values were negative with maximum values of -3.50 in mixture set 1 and -4.85 in mixture set 2. Further, all known donor log(LR) values were positive, with minimums of 1.07 in mixture set 1 and 1.88 in mixture set 2 (several instances of cross-matching across known donors were encountered; see Section VIII for Interpretation Guidelines). Even with cross-matching, the minimum separation between known donors and non-donors for both three person mixture sets was still 5.64 log units (over 436,000) (see *Section IX. Glossary of Terms* for cross-matching).

Figure V.B.1. Dot plot showing the specificity of TrueAllele Casework by reference sample. Mean values from the three person mixture set 1 are shown below; reference donors 12, 16, and 19 (far right) were used in the creation of all mixture ratios from this mixture set. Error bars represent one standard deviation; dashed line is set at zero.

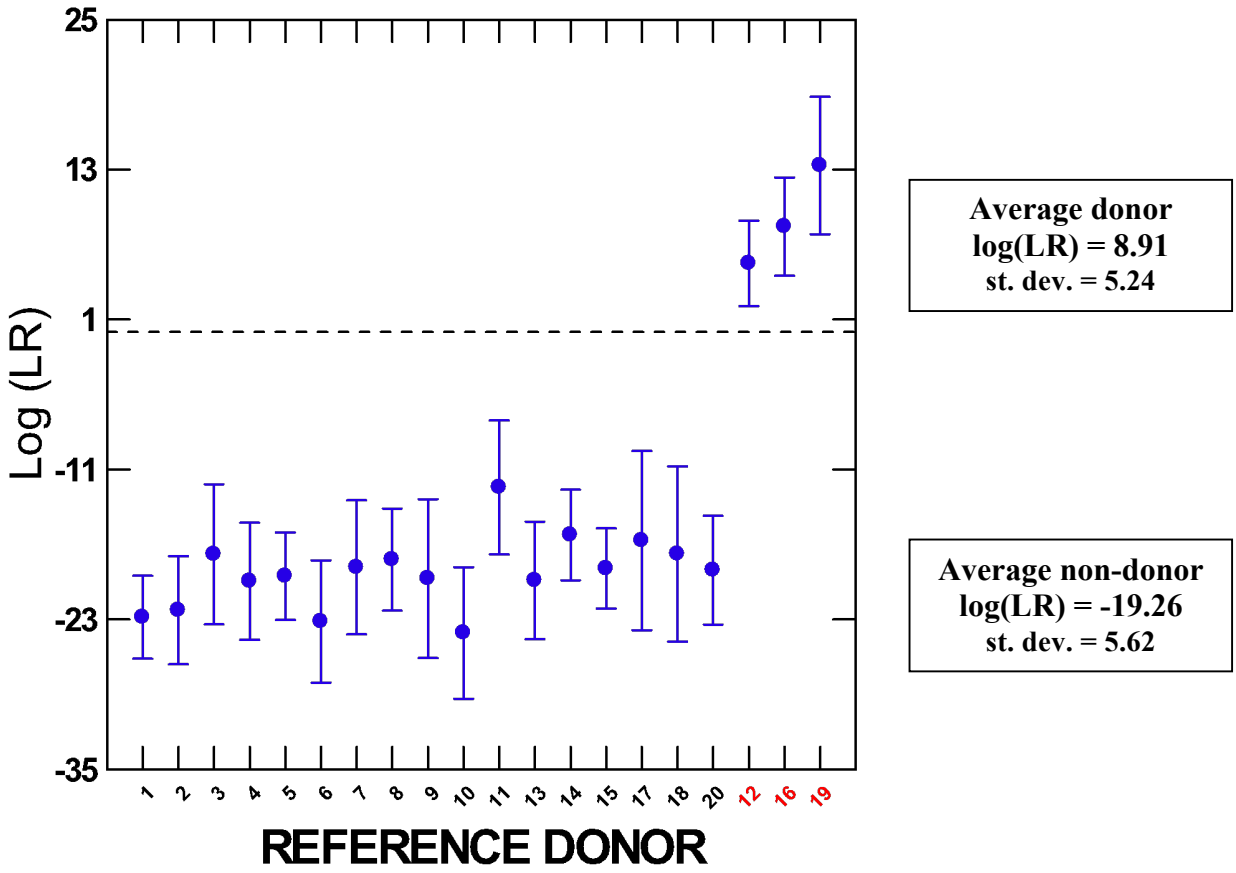
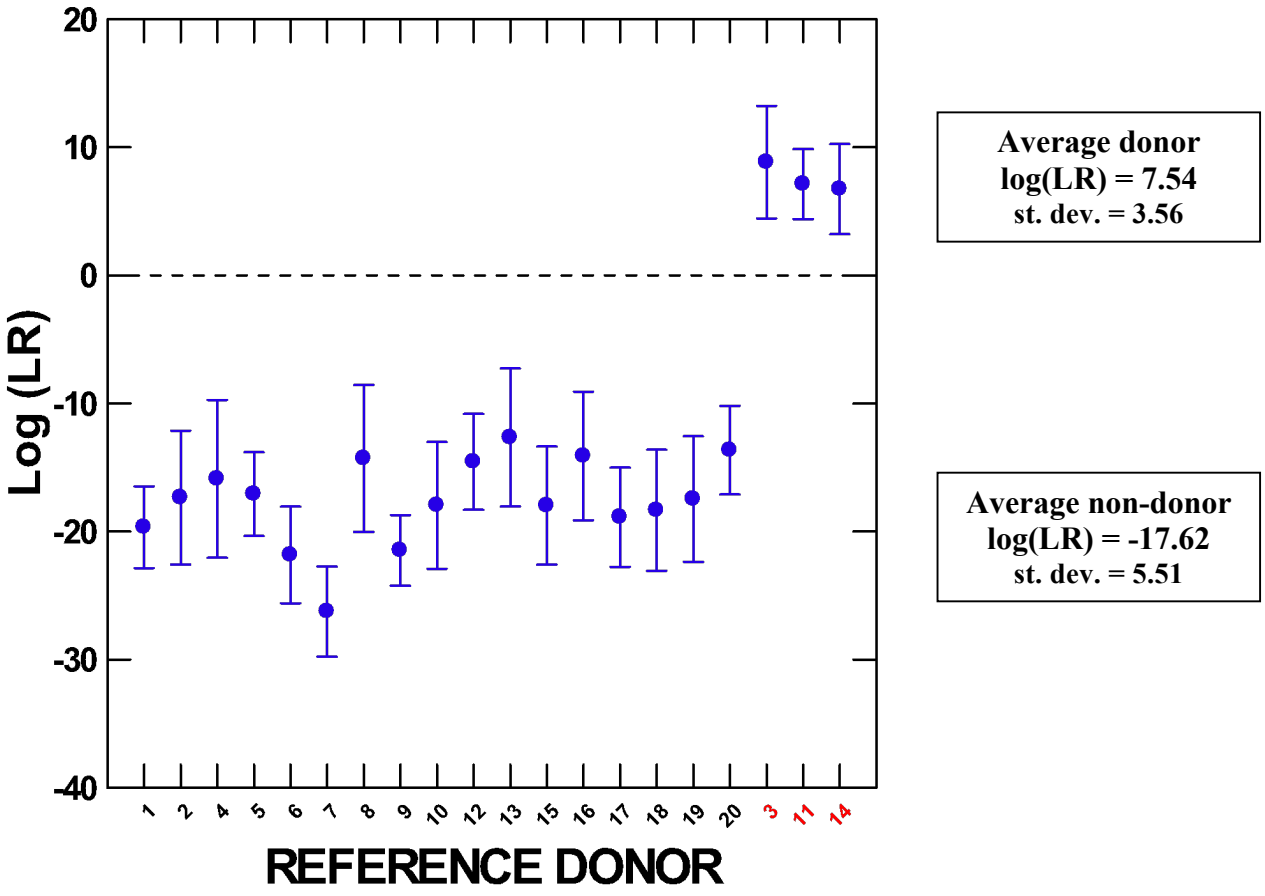


Figure V.B.2. Dot plot showing the specificity of TrueAllele Casework by reference sample. Mean values from the three person mixture set 2 are shown below; reference donors 3, 11, and 14 (far right) were used in the creation of all mixture ratios from this mixture set. Error bars represent one standard deviation; dashed line is set at zero.



C. Reproducibility – Mixture Weights and Log(LR)

Similar to the two person mixture sets, the reproducibility of TrueAllele Casework was assessed by running duplicate identification requests for all three person Identifiler Plus mixtures. Log(LR) match statistics for all contributors were highly reproducible, with mean and maximum standard deviations of 0.585 and 2.707, respectively (Table V.C.1. and see three person Match Table). Values are in accord with the two person mixture sets where the largest variations were associated with mixture profiles of approximately equal mixture weights or with minor reference donors. Additionally, the inferred mixture weights for all mixture sets correlated strongly with the theoretical mixture weights created by Sorenson, with a mean standard deviation of 0.028 (Table V.C.2.). KL statistics were also reproducible across all three person mixture ratios (Figure V.C.1.).

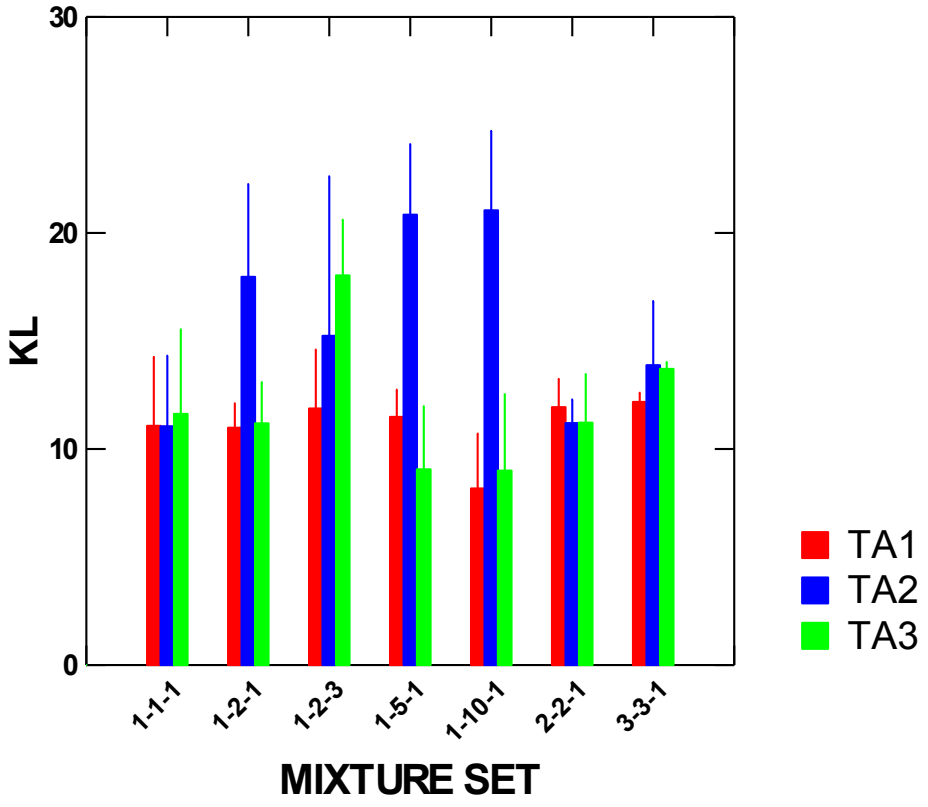
Table V.C.1. Log(LR) match statistics and associated standard deviations listed by contributor across both three person mixture sets tested.

Theoretical Mixture Weight Ratios	Mixture Set	Log(LR) Donor 1	Standard Deviation Donor 1	Log(LR) Donor 2	Standard Deviation Donor 2	Log(LR) Donor 3	Standard Deviation Donor 3
1:1:1	1	6.738	2.022	9.609	0.725	6.663	2.707
1:2:1	1	8.629	0.669	17.974	0.006	10.573	0.329
1:5:1	1	2.374	0.653	19.289	0.064	9.033	2.064
1:10:1	1	3.819	0.148	21.306	0.000	3.842	0.009
1:2:3	1	12.797	1.385	16.879	0.352	18.763	0.250
2:2:1	1	11.075	0.784	12.853	0.858	4.834	0.042
3:3:1	1	13.126	0.085	15.711	1.085	10.435	0.647
1:1:1	1	4.238	0.125	4.162	0.221	5.628	0.301
1:2:1	2	7.159	0.308	12.800	0.025	7.765	0.393
1:5:1	2	8.822	0.680	16.765	0.000	6.371	0.933
1:10:1	2	8.353	1.269	16.756	0.013	2.959	1.260
1:2:3	2	7.253	1.235	8.278	0.688	14.198	0.697
2:2:1	2	12.360	0.284	9.977	0.470	8.390	0.767
3:3:1	2	11.469	0.000	10.964	0.000	10.484	0.000

Table V.C.2. Comparison of theoretical three person mixture weights to those inferred by TrueAllele Casework; mixture weights and standard deviations shown are averaged from duplicate TrueAllele requests within all respective mixture categories for mixture sets 1 and 2.

Theoretical Mixture Weight Ratios	TrueAllele Inferred Mixture Weight	Mean Standard Deviation
1:1:1	.305/.340/.356	0.032
1:2:1	.239/.498/.264	0.025
1:5:1	.162/.719/.119	0.028
1:10:1	.065/.851/.084	0.011
1:2:3	.177/.294/.530	0.050
2:2:1	.332/.411/.258	0.043
3:3:1	.382/.411/.046	0.011

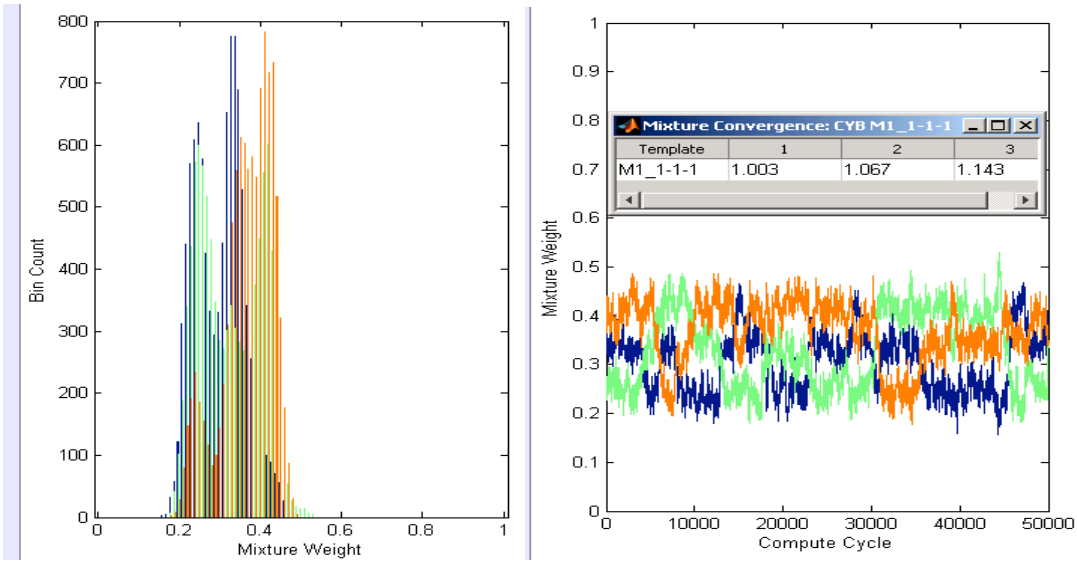
Figure V.C.1. Bar graph of KL statistics as a function of mixture set. Mean values from two separate three person mixture sets are shown below; all TrueAllele samples were solved in duplicate with TA1, TA2, and TA3 referring to the separate contributors for the three person mixture sets. Error bars represent one standard deviation.



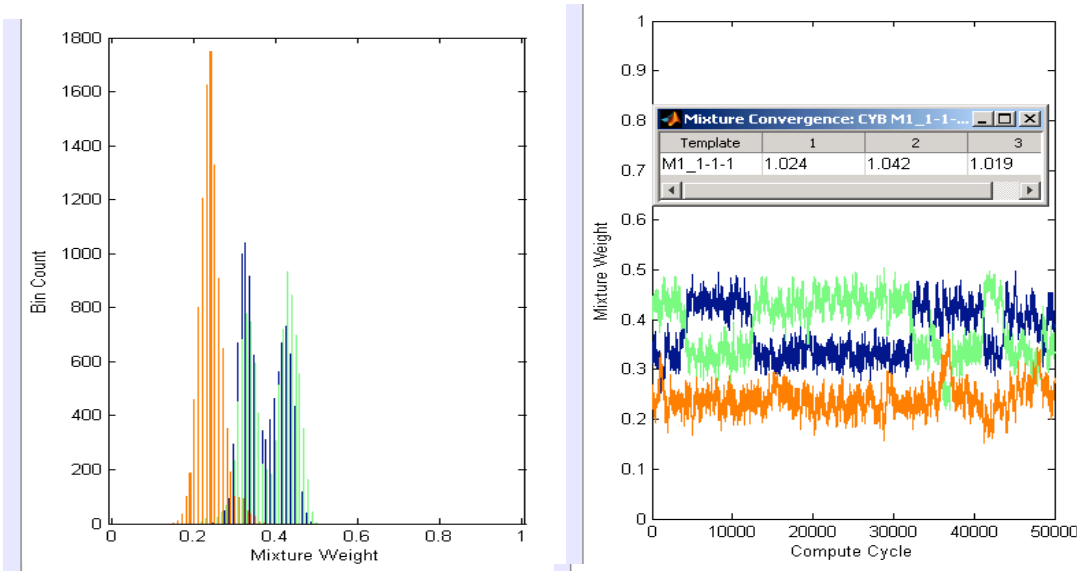
D. MCMC Convergence

For all three person mixtures tested, the mean Gelman Rubin convergence statistic (GR hereafter) was 1.142 with a standard deviation of 0.227 and maximum of 2.181. All three person mixture samples exhibited acceptable convergence based on assessing GR statistics and viewing the mixture weight histograms and Markov chains (Figure V.D.1).

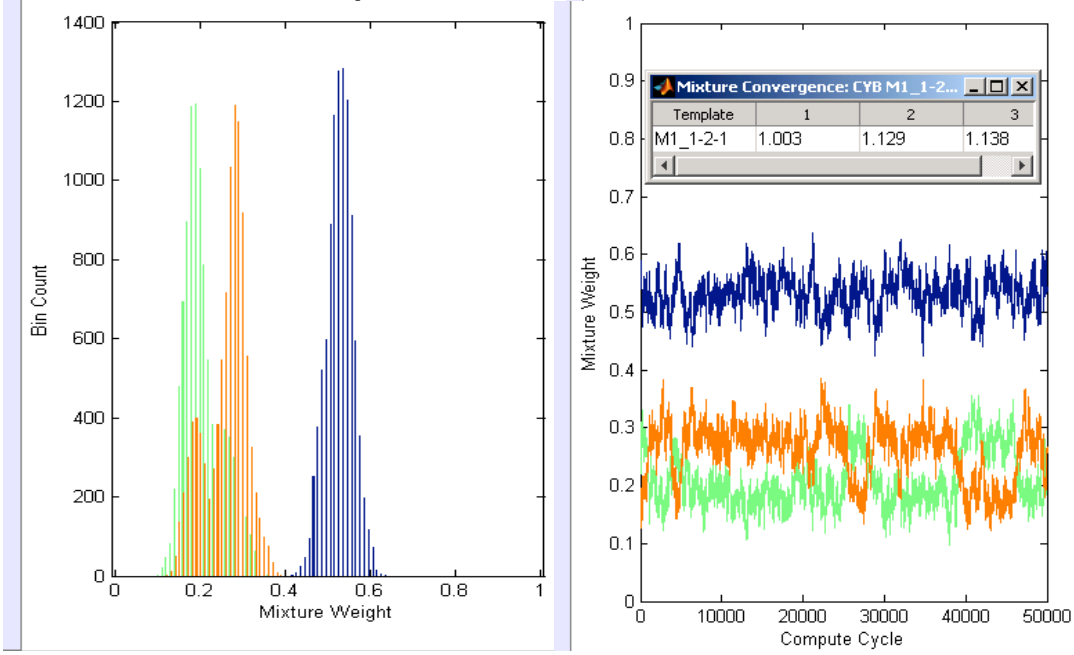
Figure V.D.1. Histograms and associated Markov Chains from all three person mixture samples; GR convergence scores are indicated in the Markov chains at right.



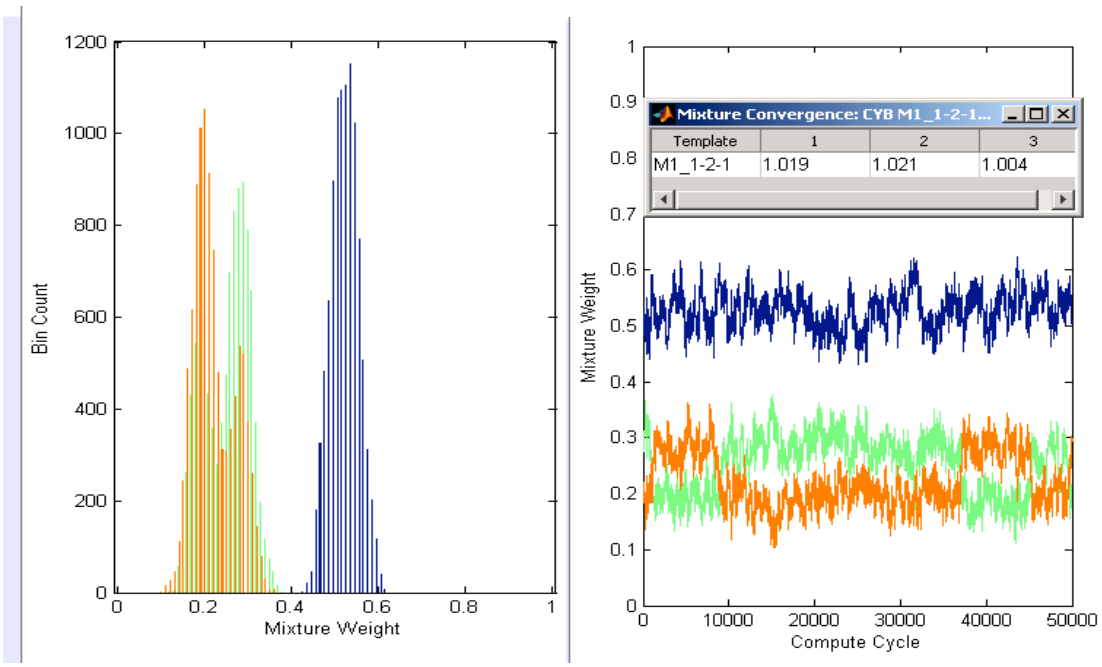
Mix Set 1
1:1:1



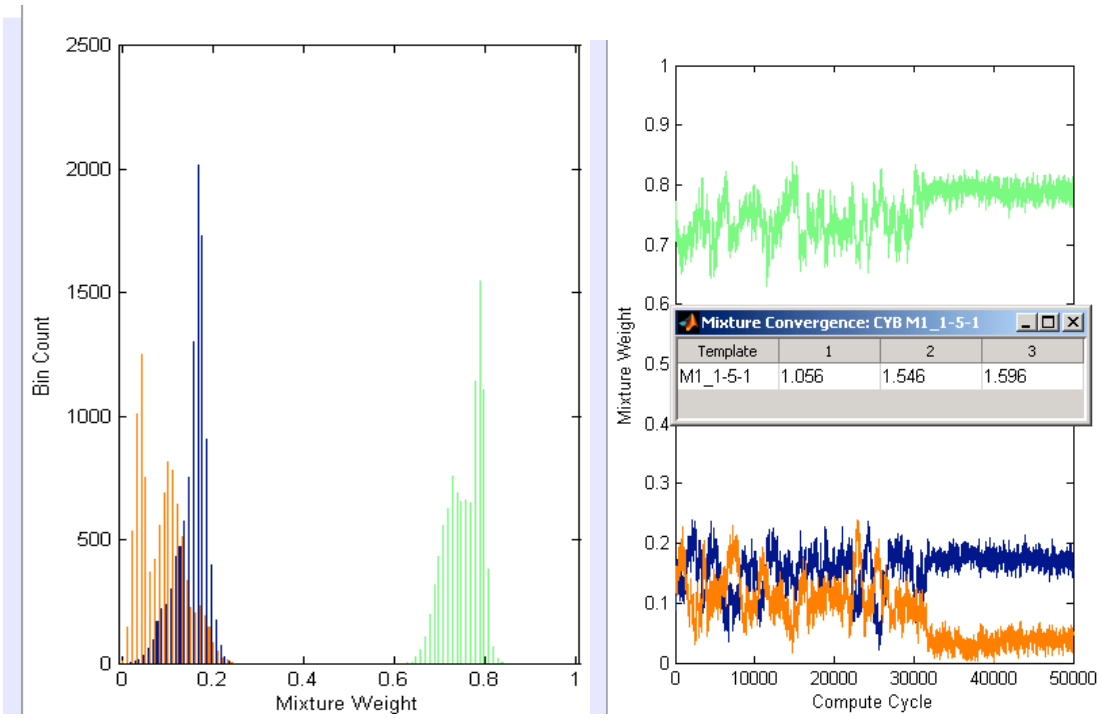
Mix Set 1 1:1:1 replicate



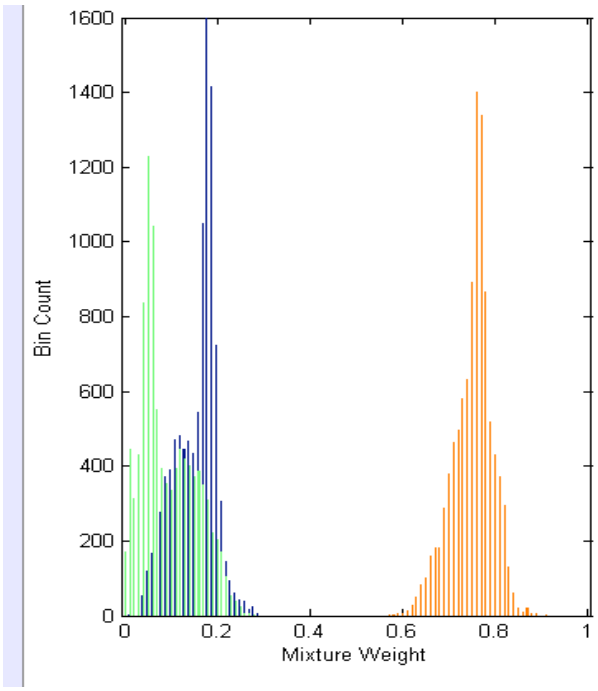
Mix Set 1 1:2:1

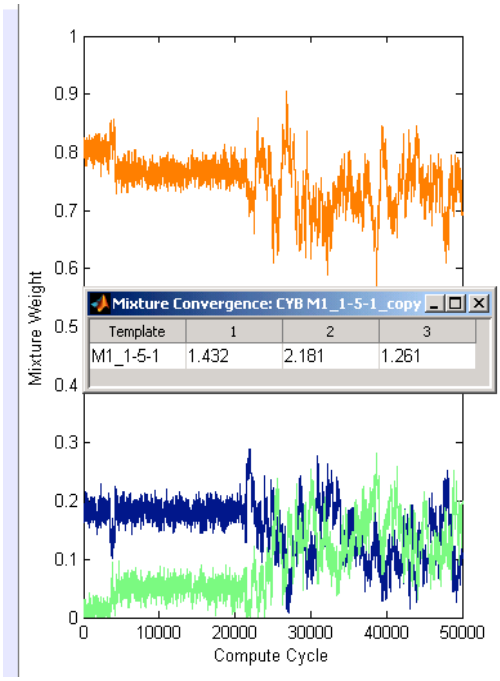


Mix Set 1 1:2:1 replicate

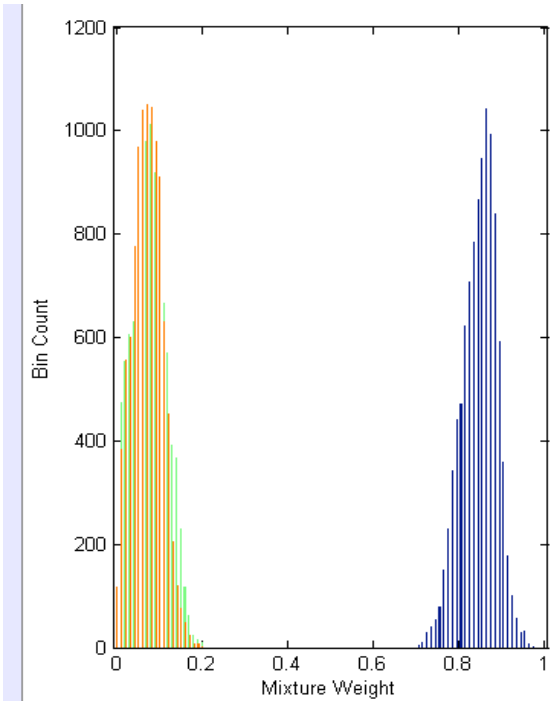


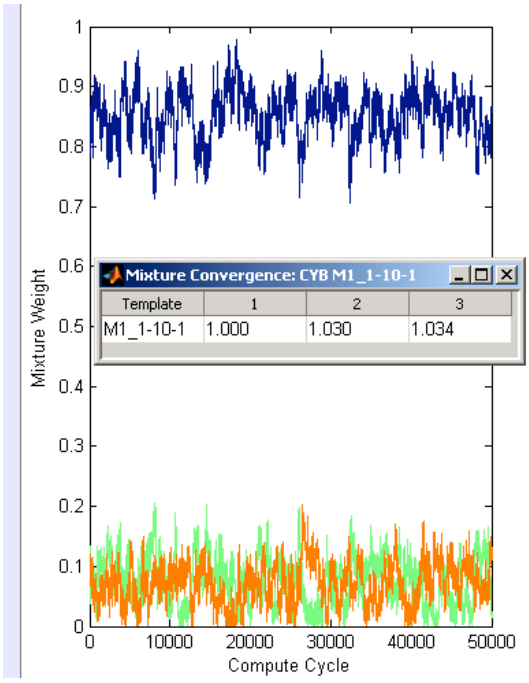
Mix Set 1 1:5:1



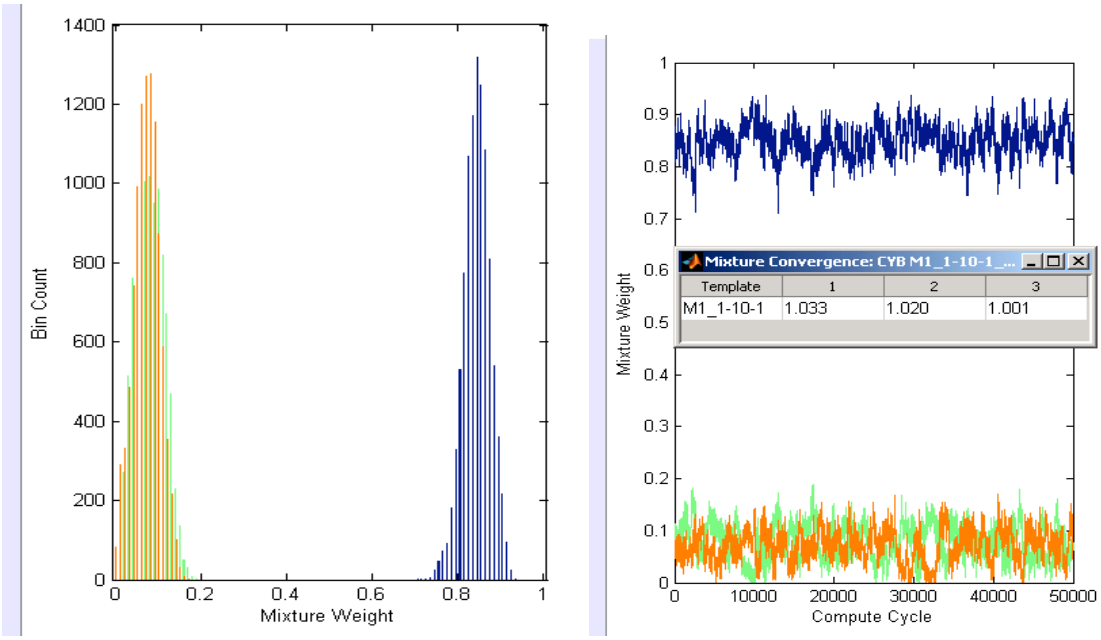


Mix Set 1 1:5:1 replicate

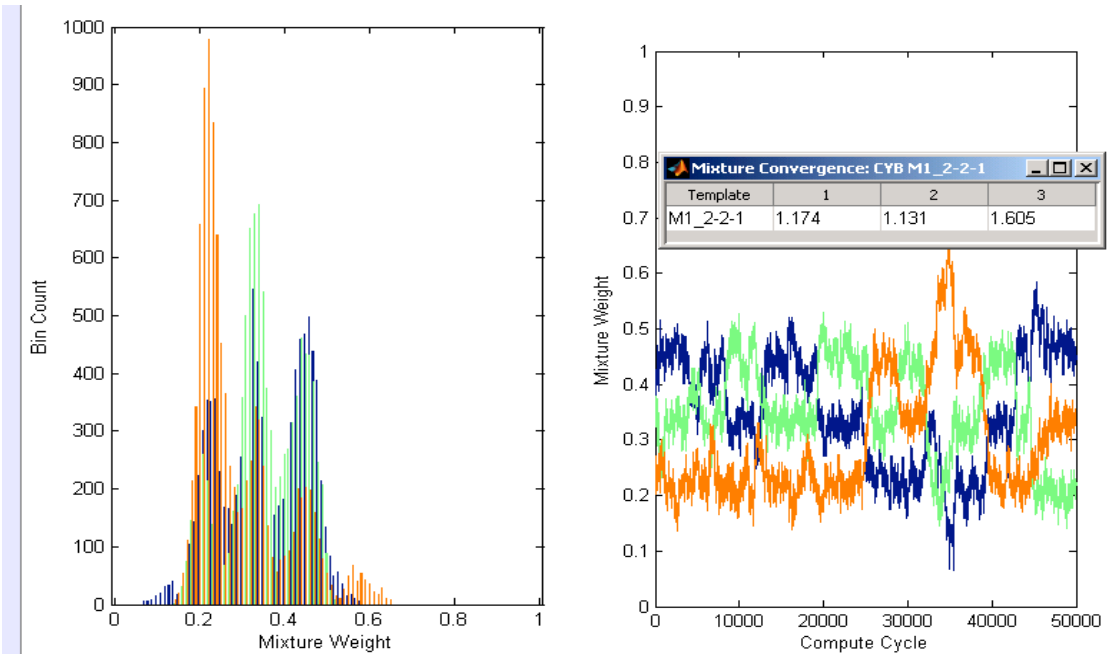




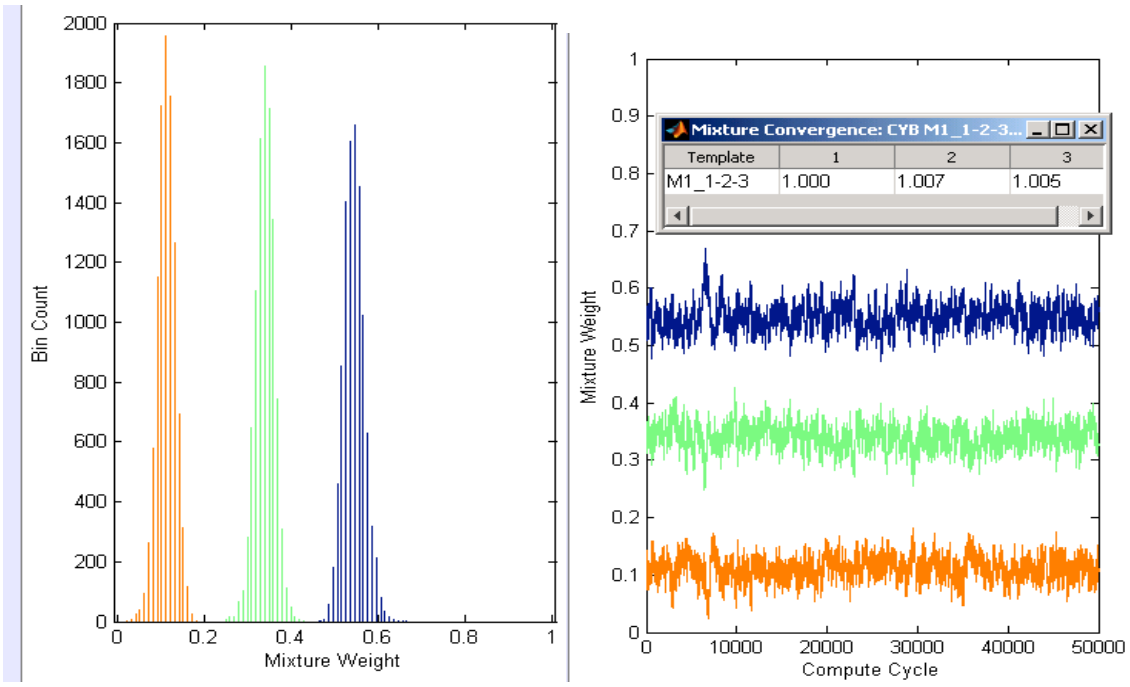
Mix Set 1 1:10:1



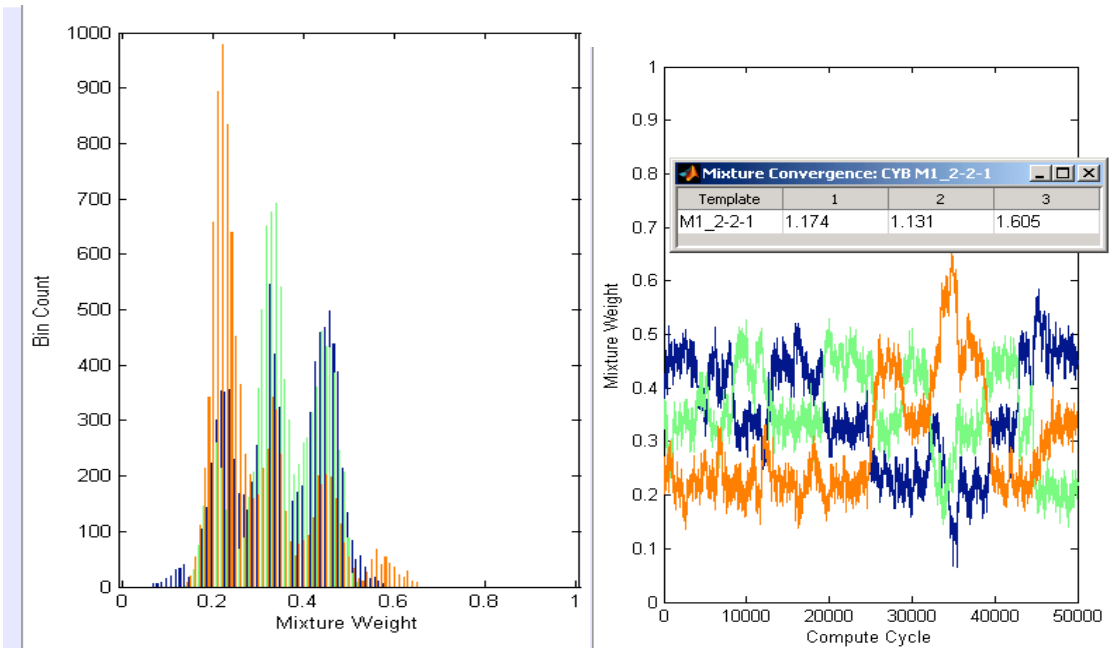
Mix Set 1 1:10:1 replicate



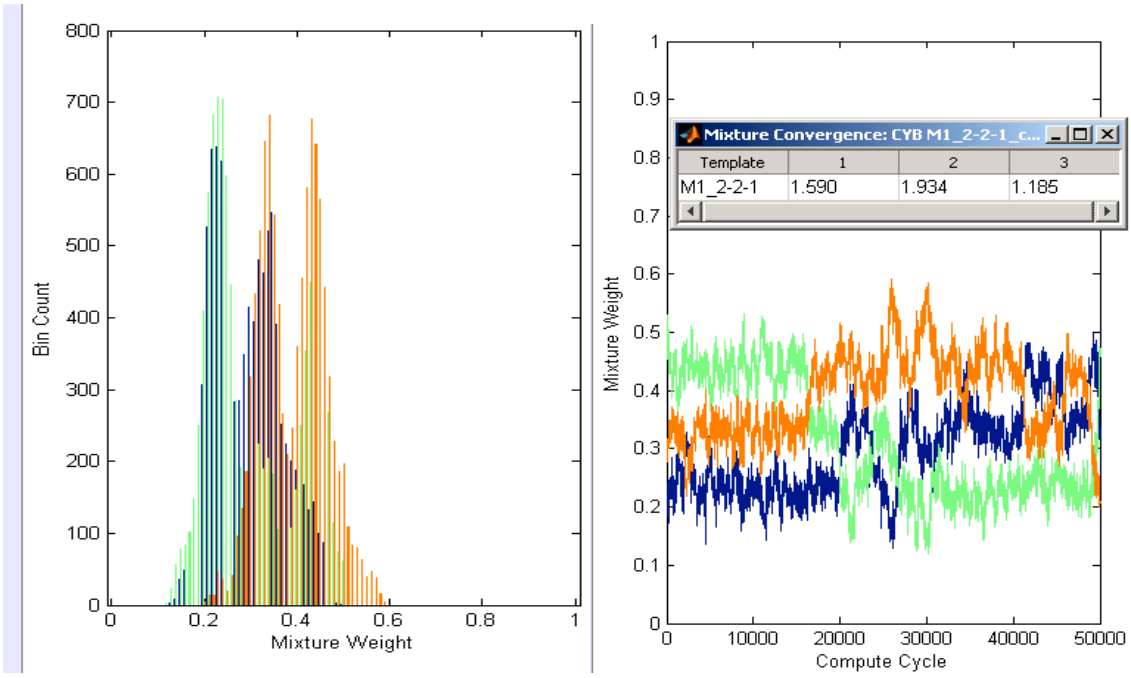
Mix Set 1 1:2:3



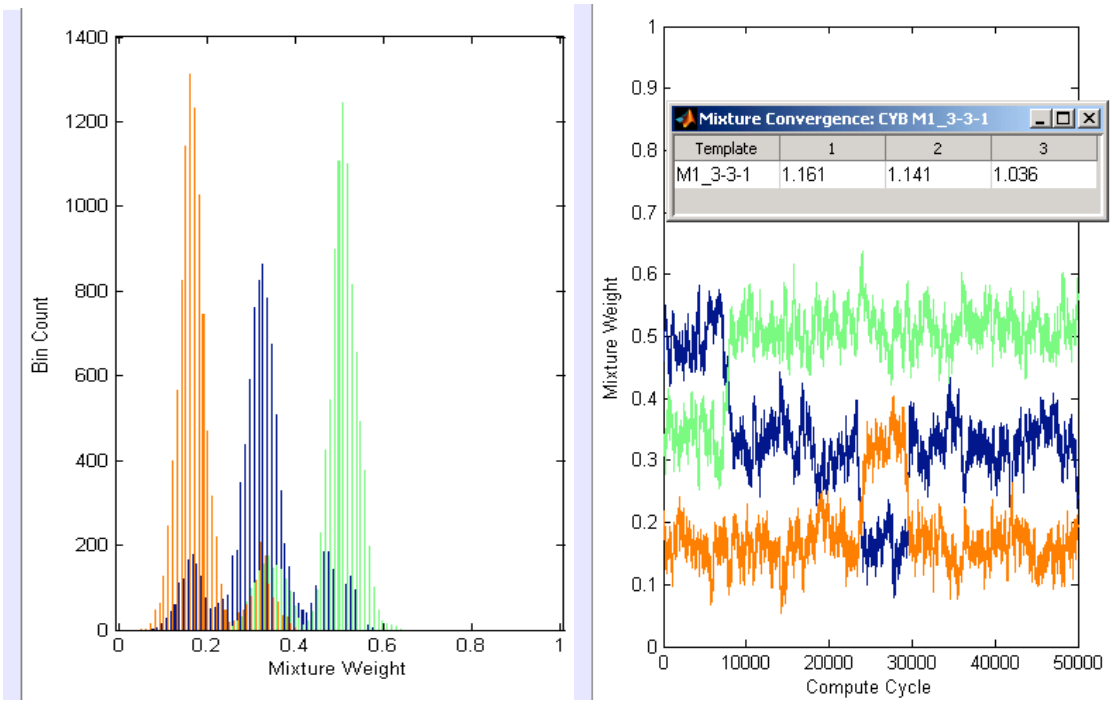
Mix Set 1 1:2:3 replicate



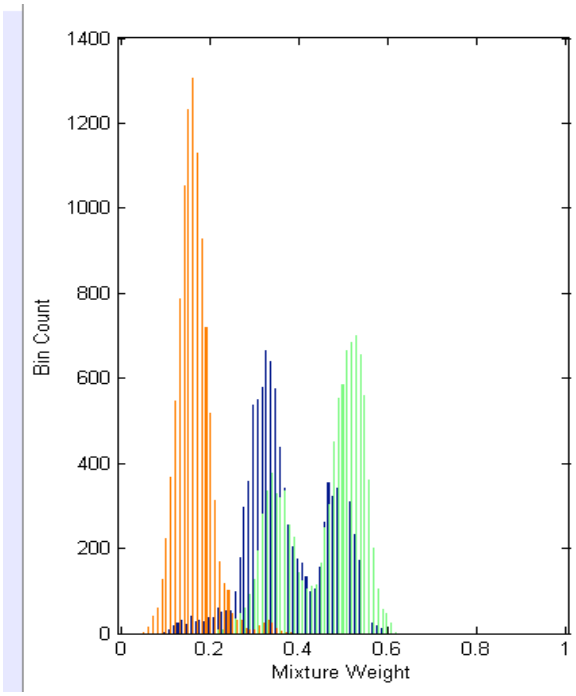
Mix Set 1 2:2:1

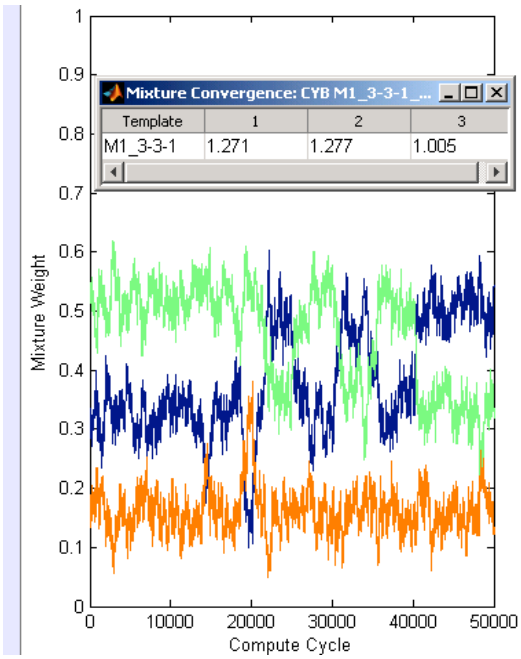


Mix Set 1 2:2:1 replicate

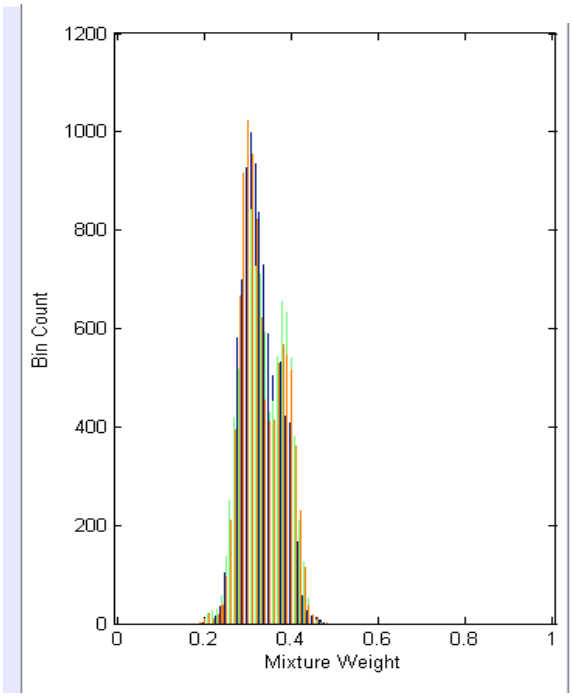


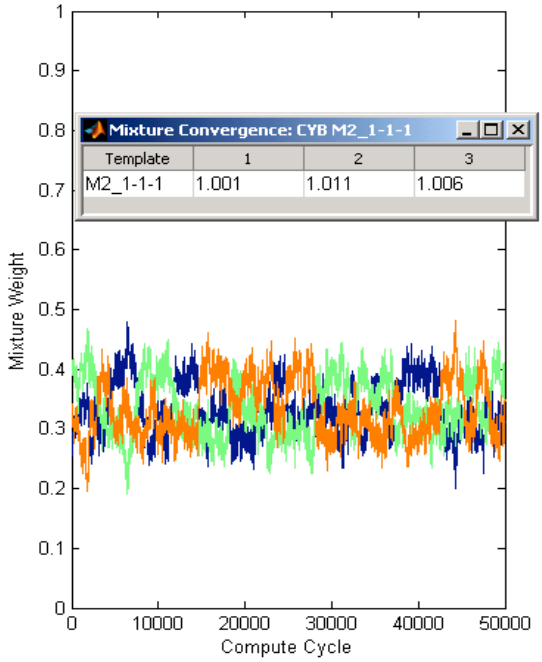
Mix Set 1 3:3:1



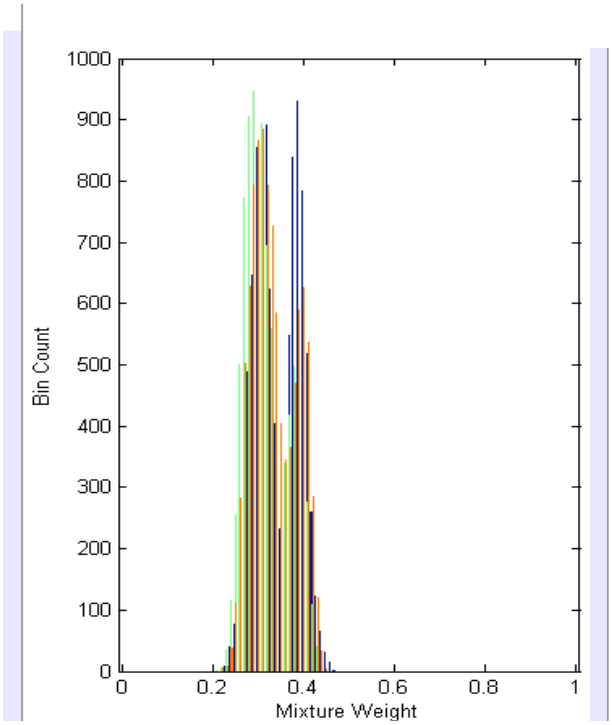


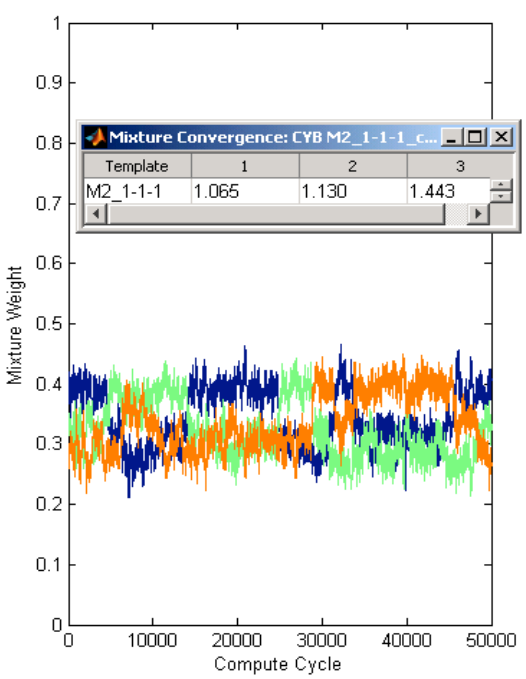
Mix Set 1 3:3:1 replicate



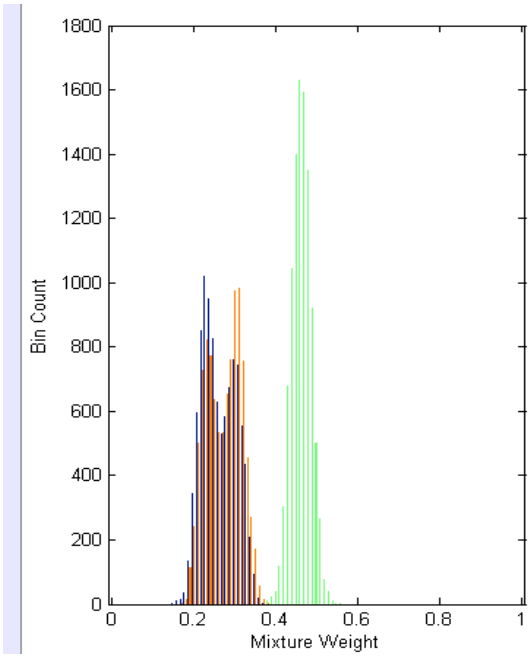


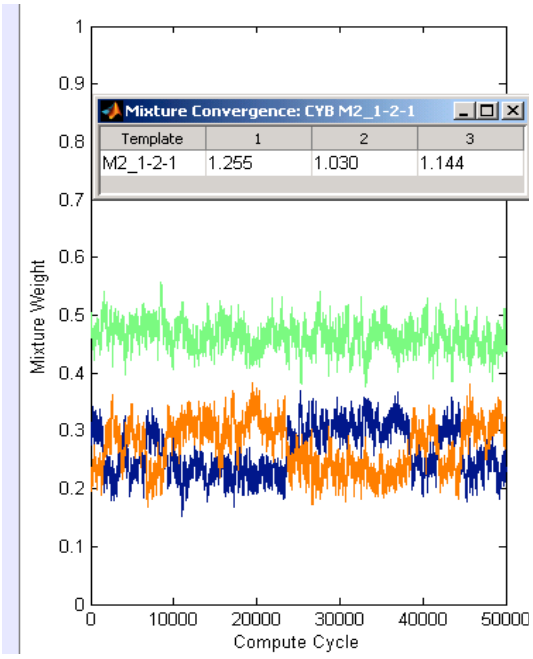
Mix Set 2 1:1:1



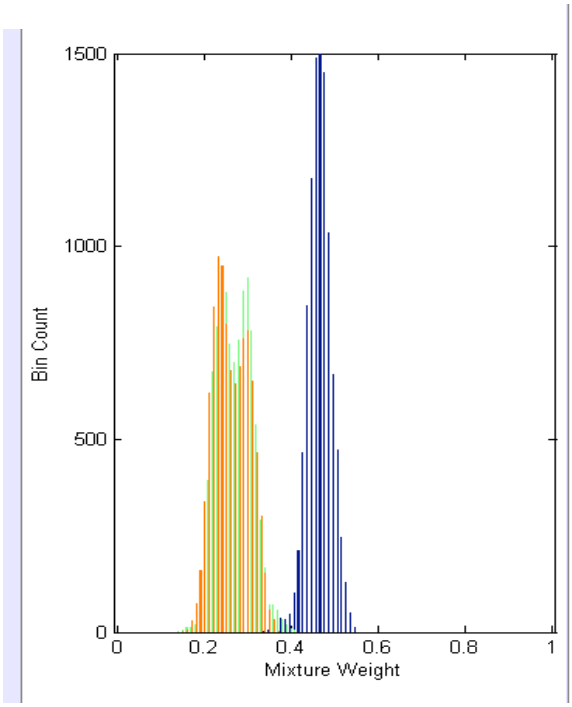


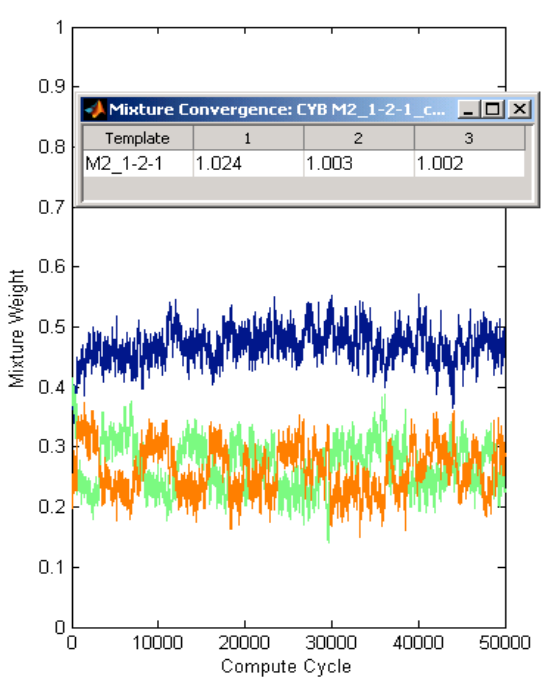
Mix Set 2 1:1:1 replicate



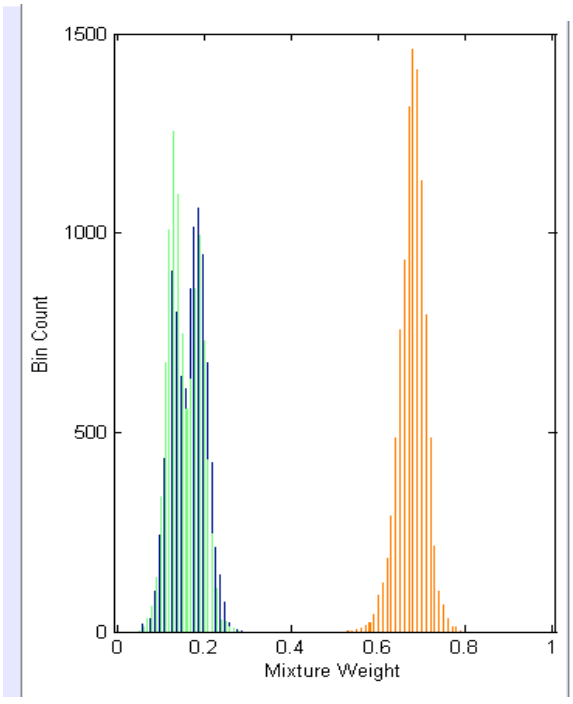


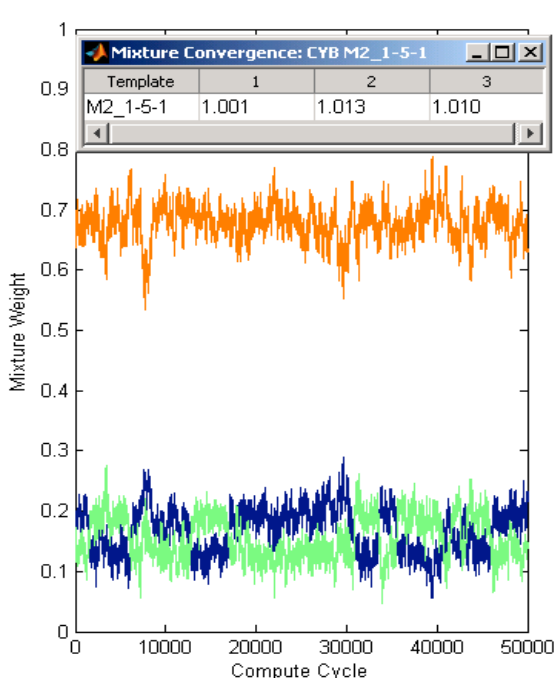
Mix Set 2 1:2:1



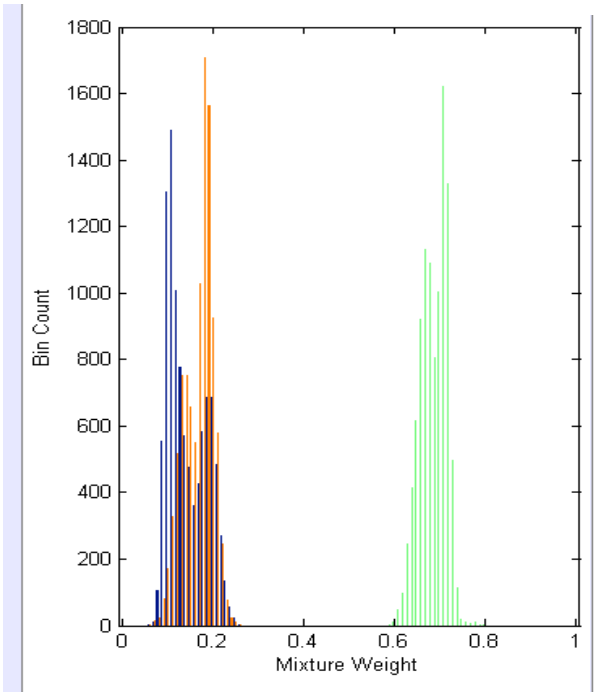


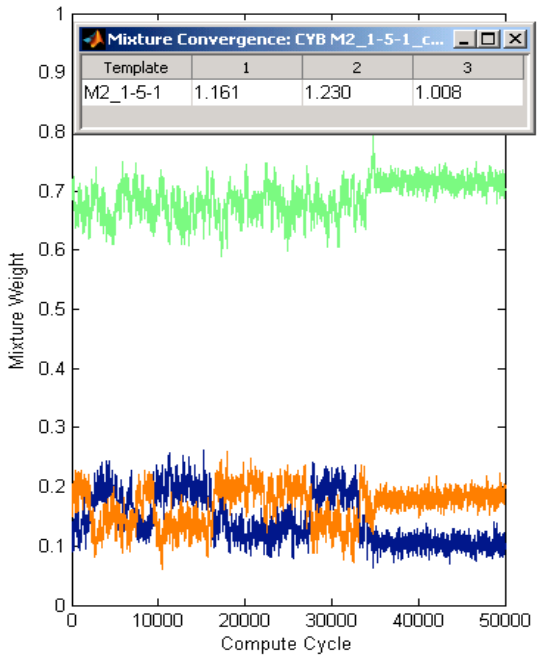
Mix Set 2 1:2:1 replicate



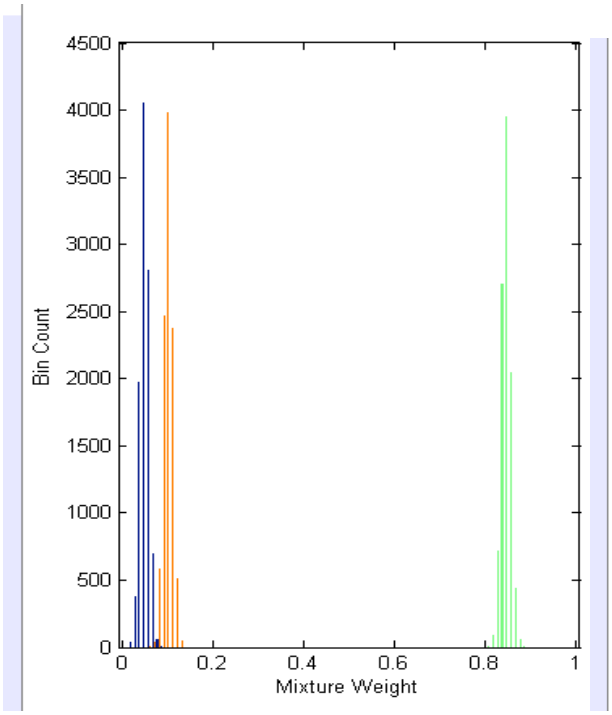


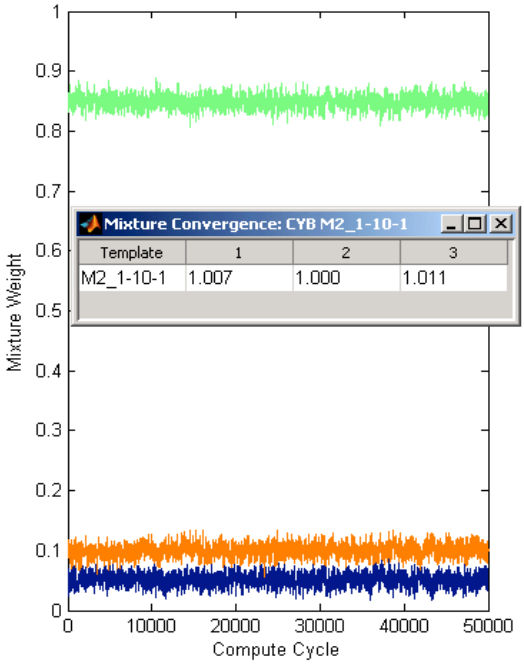
Mix Set 2 1:5:1



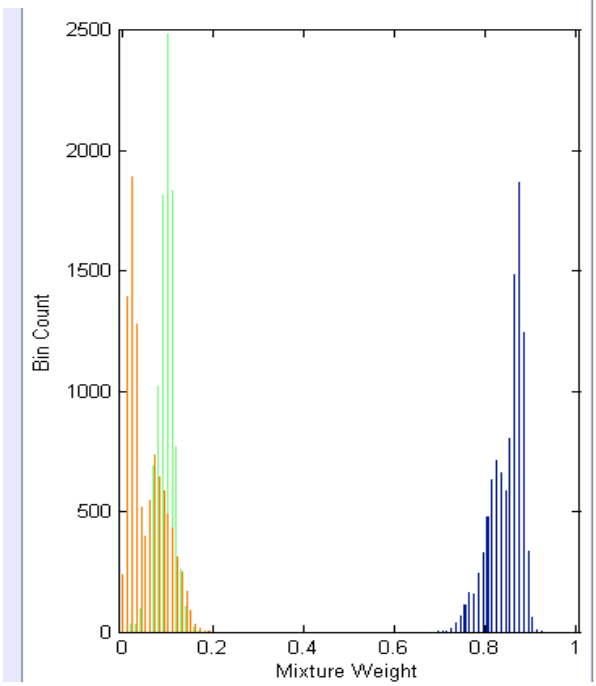


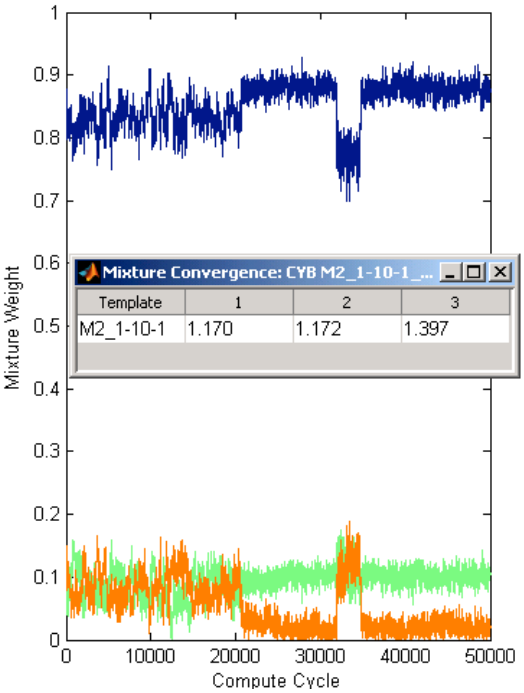
Mix Set 2
1:5:1
replicate



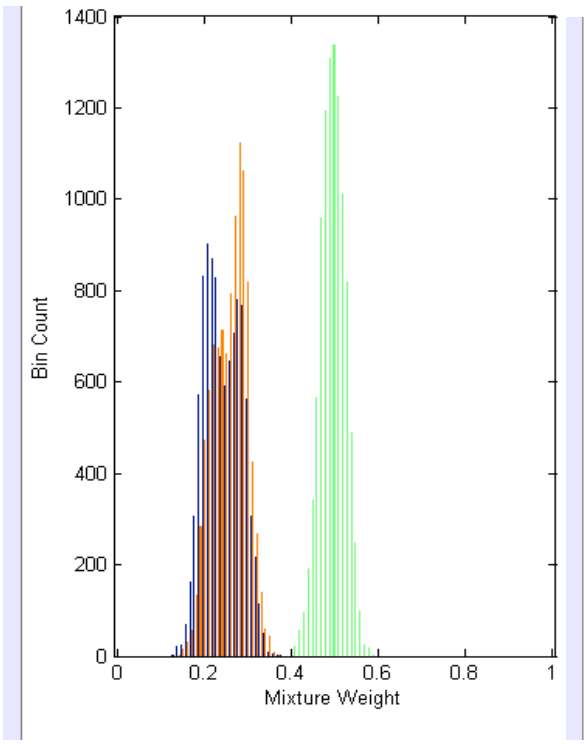


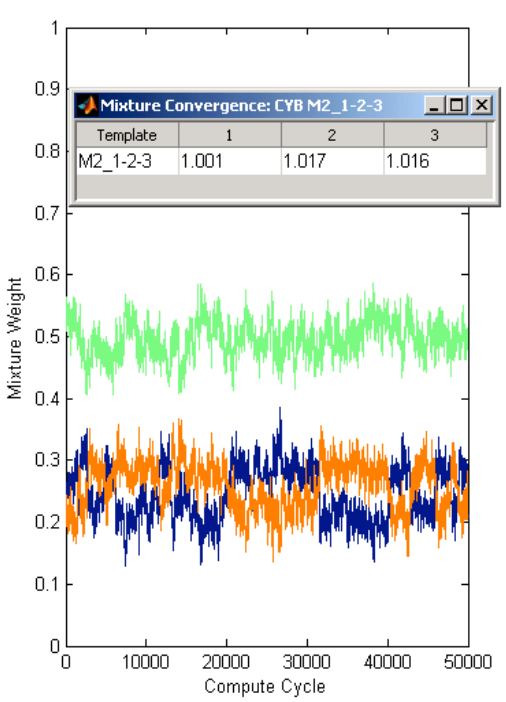
Mix Set 2 1:10:1



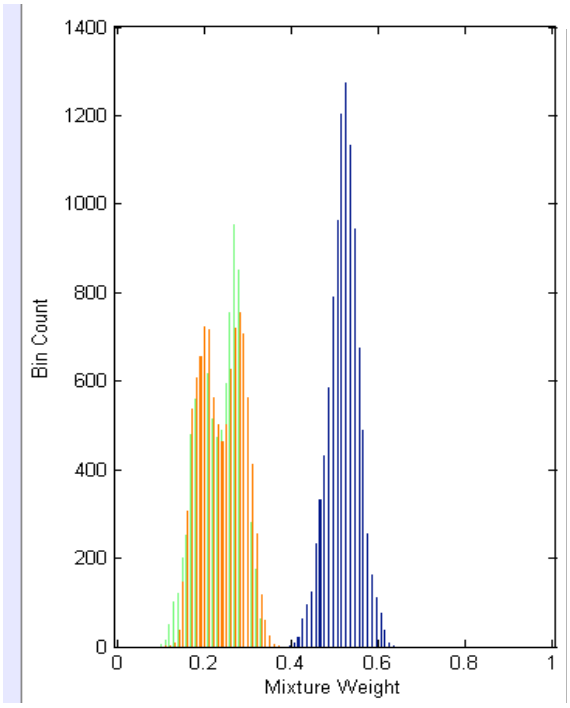


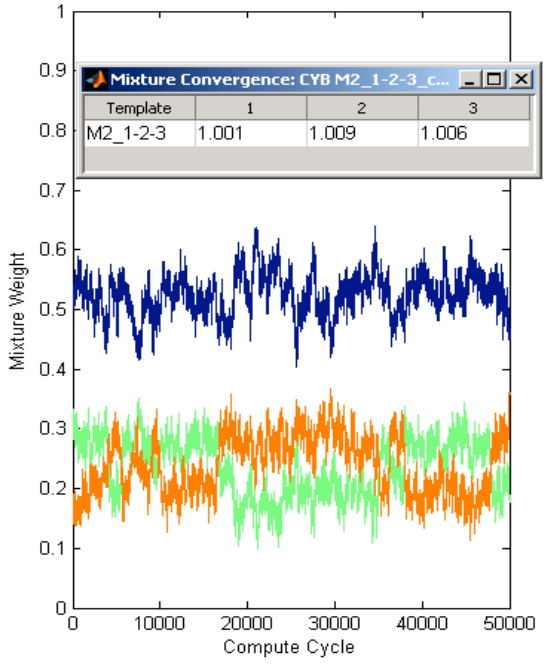
Mix Set 2 1:10:1 replicate



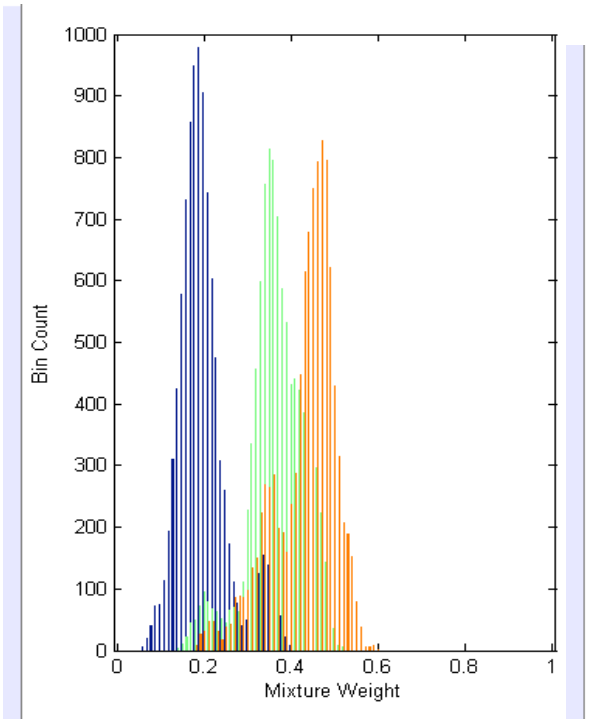


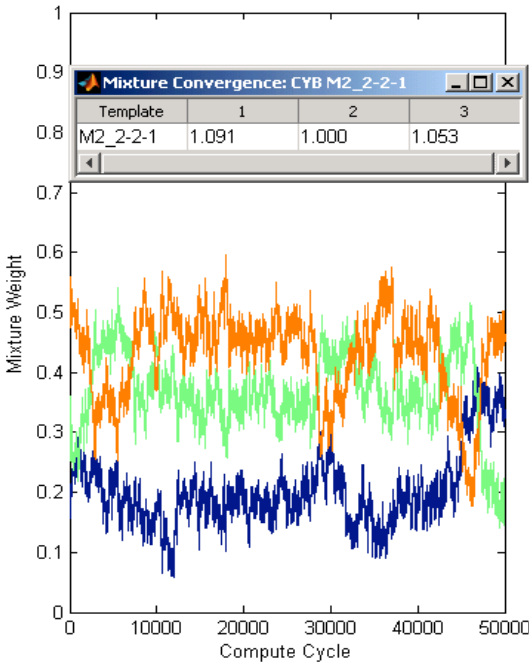
Mix Set 2 1:2:3



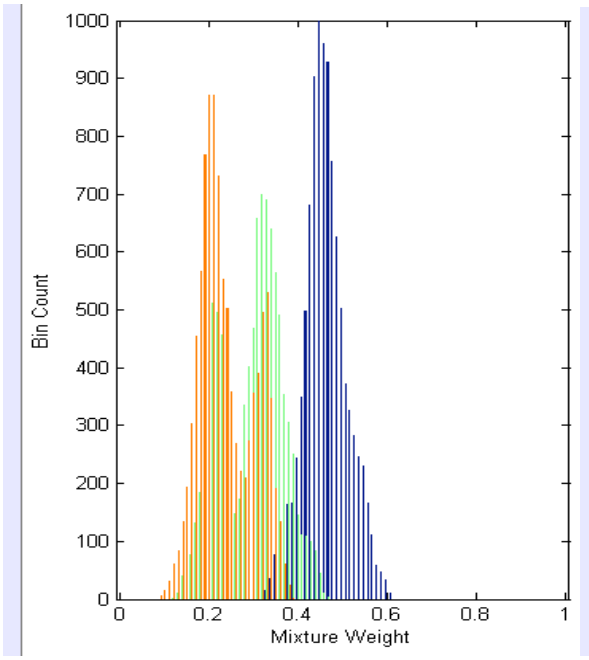


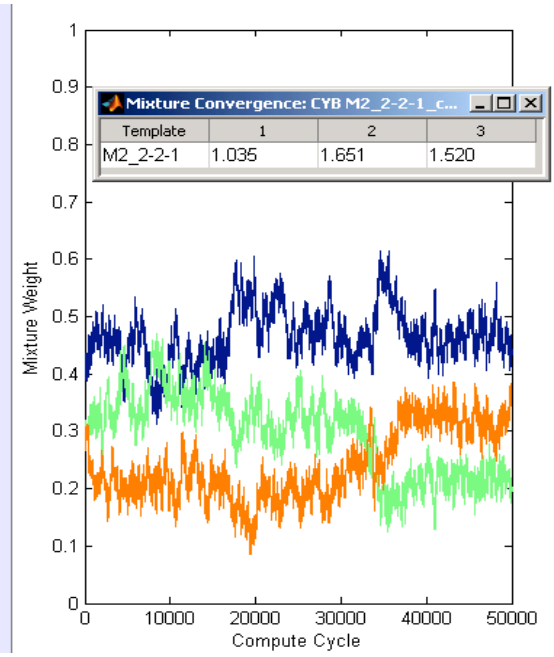
Mix Set 2 1:2:3 replicate



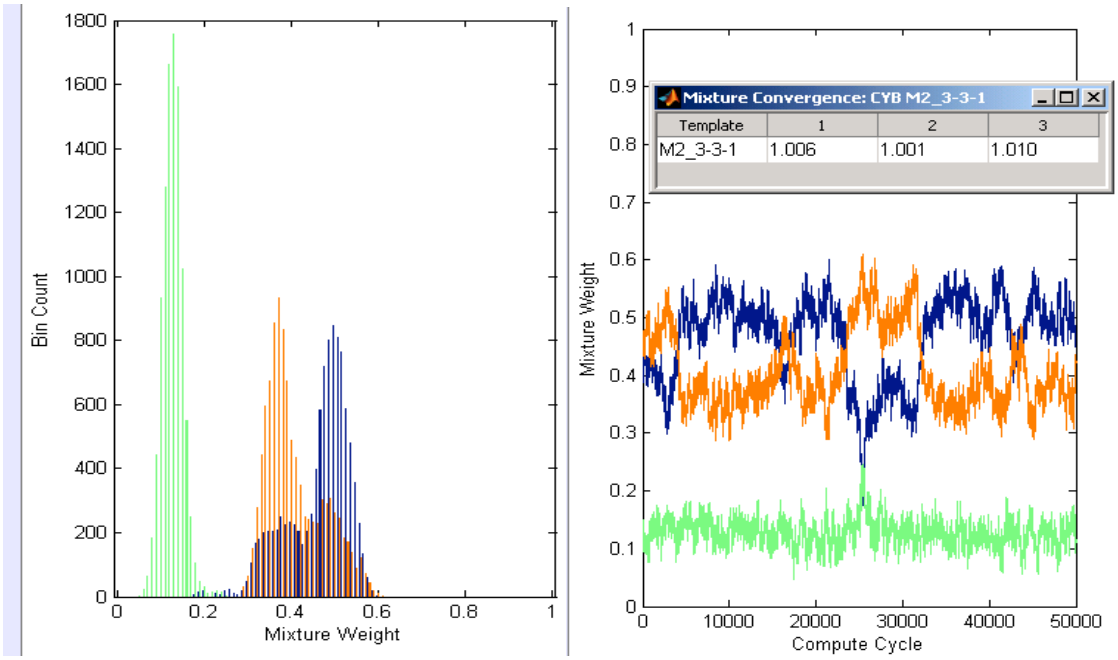


Mix Set 2 2:2:1

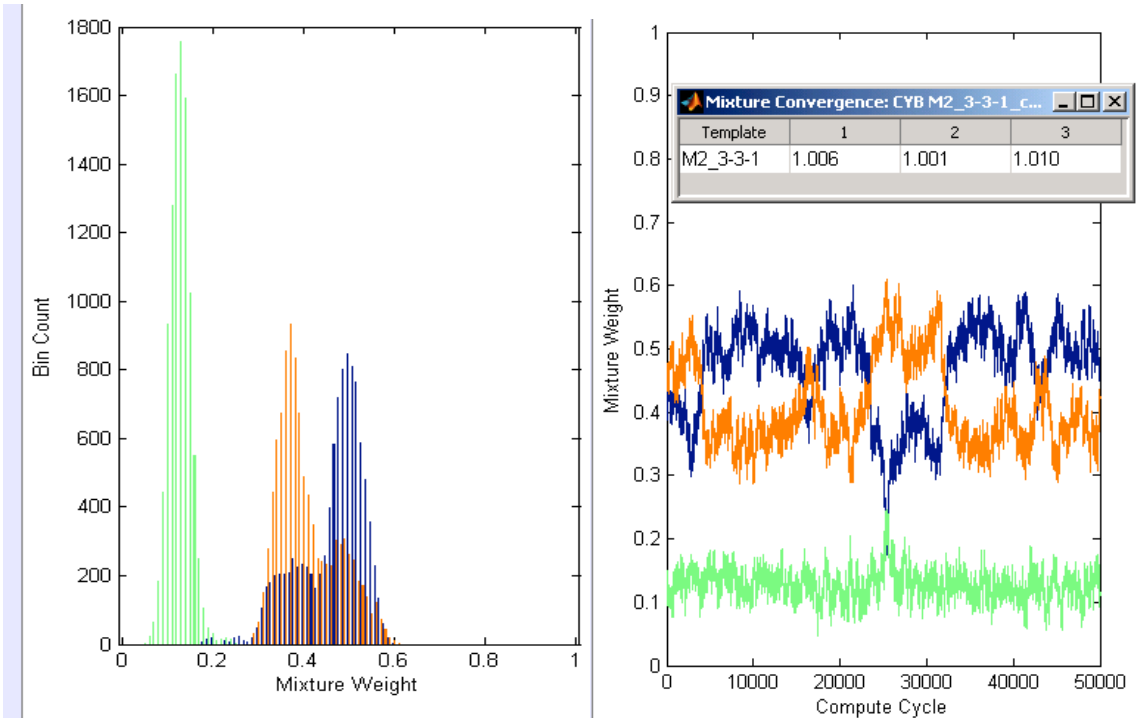




Mix Set 2 2:2:1 replicate



Mix Set 2 3:3:1



Mix Set 2 3:3:1 replicate

Jay Caponera
Bioscience Casework
11 July, 2013

VI. Experiment 2 Results: Sensitivity Study

A. Accuracy (LTDNA Samples)

The accuracy of TrueAllele Casework software in calculating weight of evidence for single source profiles exhibiting stochastic effects is described herein. Statistical comparison between TrueAllele Casework and the conditional match probability method (CMP) did not reveal a significant difference between the two methods when all DNA input categories were tested (Student's t test p = 0.430). However, when comparing data from the 15.6, 31.25, and 62.5pg categories only, TrueAllele Casework generated significantly higher match statistics (p = 0.018), and was able to provide match statistics when CMP could not (Figure VI.A.2). Data for the two lowest dilutions tested (3.9 and 7.8pg) did not yield reliable results, which may indicate a quantification threshold below which probabilistic genotyping requests for single source samples should either not be attempted, or analyzed with caution. Approaching a DNA input of 125pg and above, both methods converge on the same approximate match statistic, indicating that the maximum amount of genetic information has been recovered (Figure VI.A.1.).

Figure VI.A.1. Comparison of mean conditional match probability (CMP) and TrueAllele single source match statistics as a function of DNA input. Data include two amplified sets of serially diluted single source samples pooled by input amount. All samples were solved in duplicate in TrueAllele with 25K/25K burn-in and read out cycles (n = 32). Error bars represent one standard deviation.

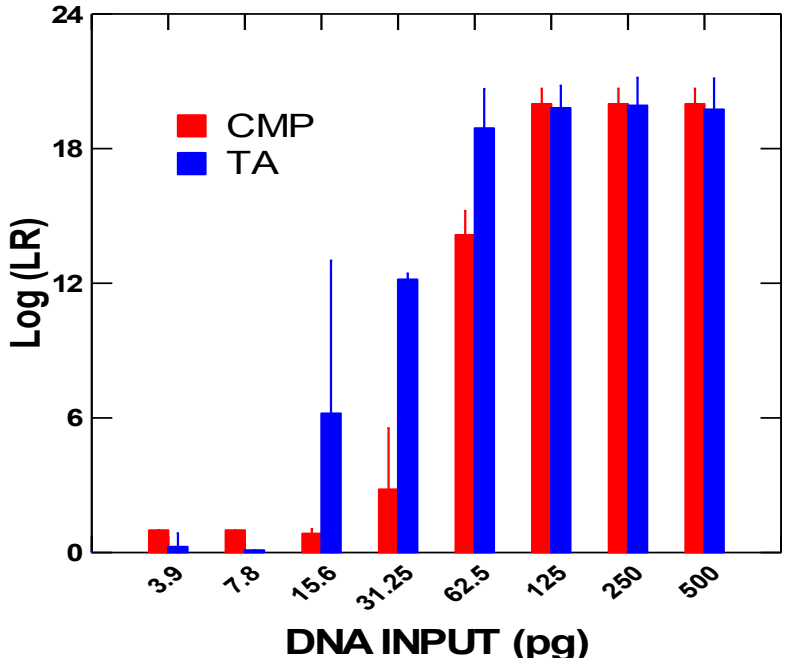
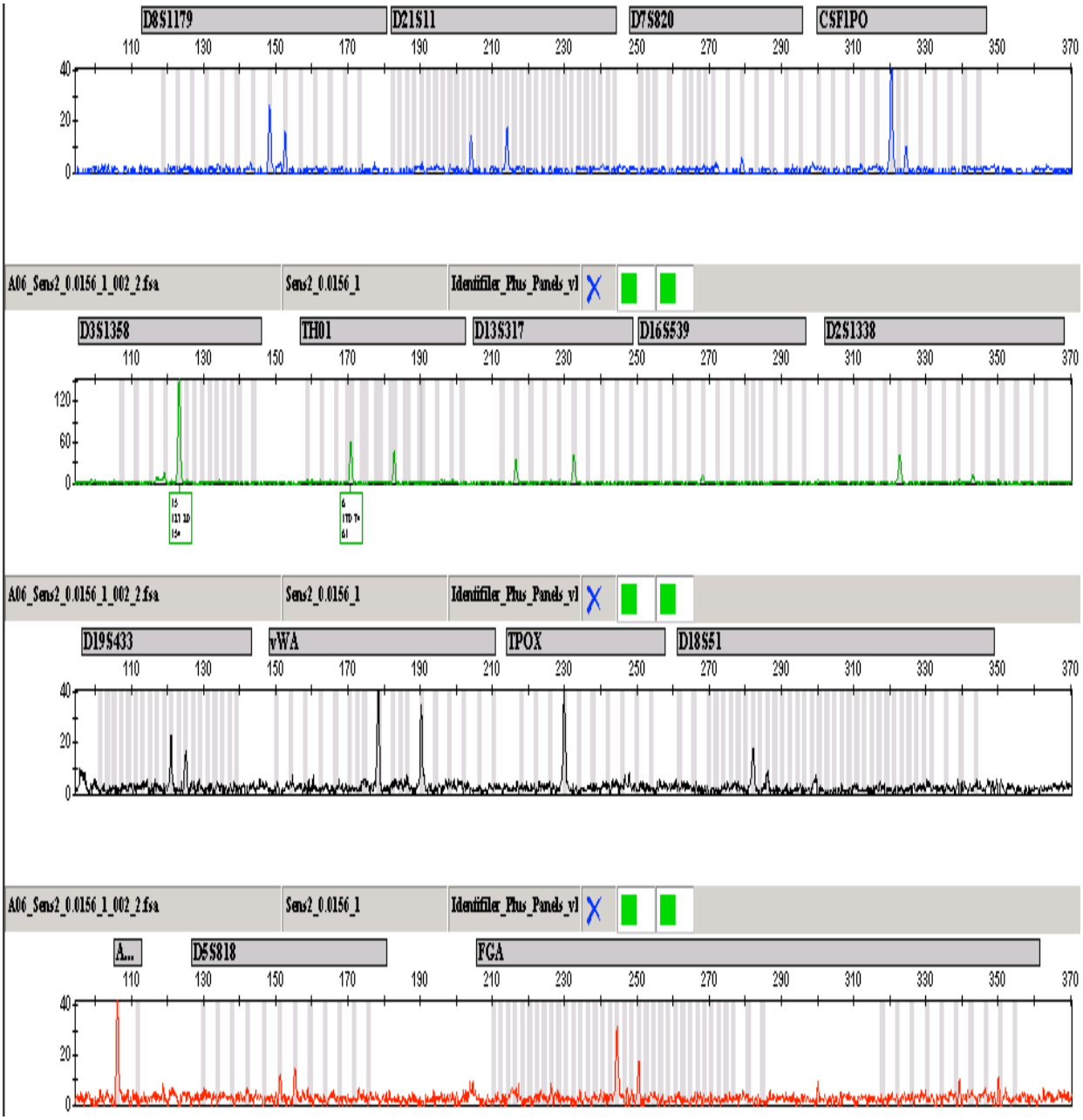


Figure VI.A.2. Electropherogram of 15.6pg Identifiler Plus sample from sensitivity set 2. The mean TrueAllele log(LR) from duplicated requests was 11.13 (billion); CMP was able to provide a match statistic of only 0.699 log units from the single 154 RFU "15" allele above stochastic threshold at D3S1358. In practice, the profile below would be deemed insufficient for comparison purposes in the threshold-based inclusion/exclusion paradigm.



B. Specificity (LTDNA SAMPLES)

Data from the sensitivity set 1 examination show a mean separation of 31.756 log units (decillion) between known donor and non-donor reference profiles, and a mean separation of 30.299 log units (nonillion) for sensitivity set 2. All single source non-donor reference log(LR) values were negative with maximum values of -2.645 in sensitivity set 1 and -3.213 in sensitivity set 2. Similar to the accuracy data from section A above, the LT-DNA sensitivity data indicate that TrueAllele Casework fails to discriminate between known donor and non-donor reference profiles when examining DNA input concentrations at or below 7.8pg (Figure VI.B.1.). In one instance, the non-donor log(LR) exceeded the known donor log(LR) from the 7.8pg sample set 2 data (match statistics from both were each negative). However, separation between known donor and non-donor profiles at or above 15.6pg improved dramatically, with a minimum of 8.586 log units (million) seen for the combined 15.6pg samples (Table VI.B.1).

Figure VI.B.1. Dot plot showing the specificity of TrueAllele Casework as a function of DNA input. Reference samples include one known donor and 19 non-donors from both LTDNA sensitivity sets. Mean values from all replicated single unknown requests are pooled (n = 32). Error bars represent one standard deviation; dashed line is set at zero.

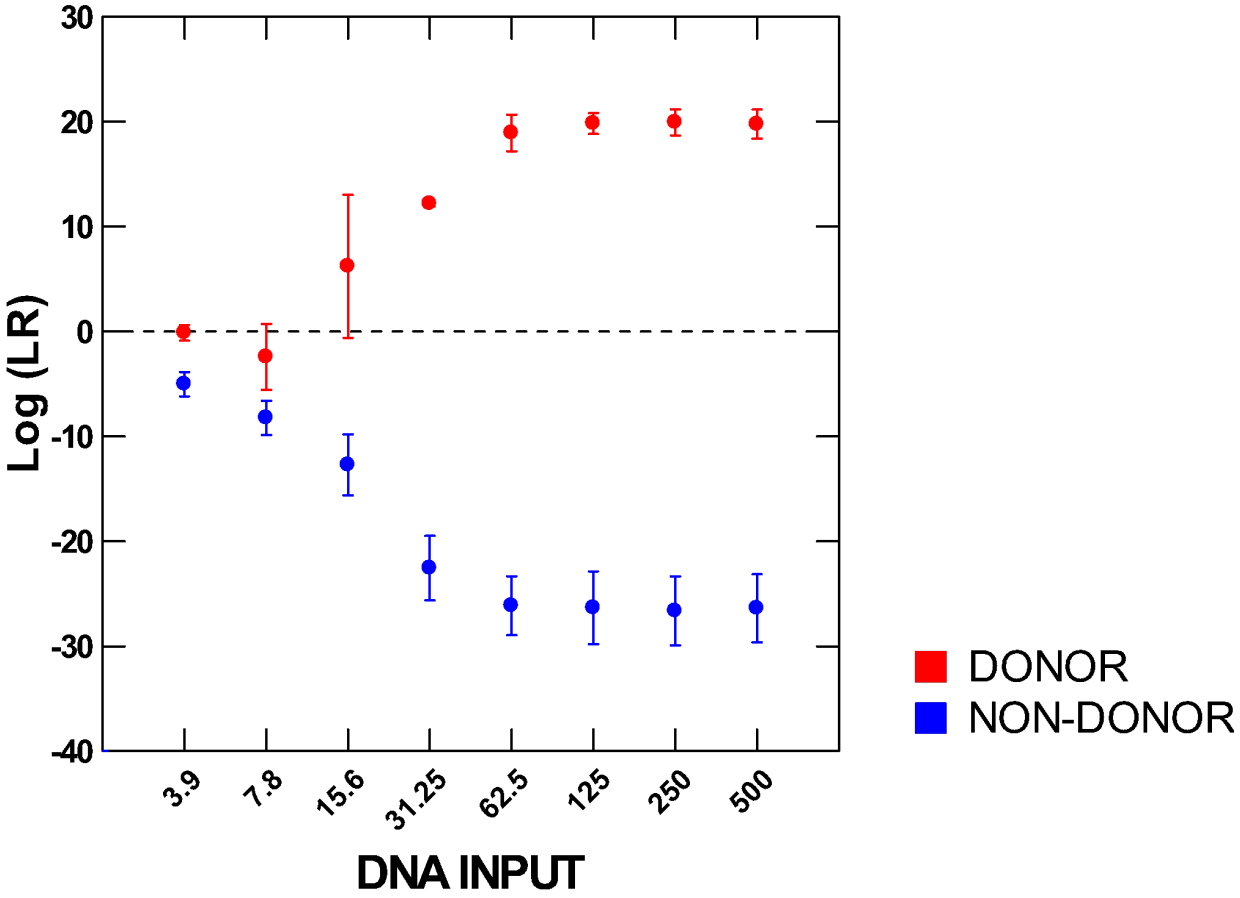


Table VI.B.1. Mean log(LR) values for known donor and non-donor reference samples with separation in log units for both combined sensitivity sets. Values from all replicated single unknown donor requests are pooled (n = 32).

DNA INPUT (pg)	Mean Donor Log(LR)	Mean Non-Donor Log(LR)	Mean Donor/Non-Donor Separation (log units)	Minimum Donor/Non-Donor Separation (log units)
3.9	-0.127	-5.025	4.898	1.649
7.8	-2.430	-8.239	5.809	0.637 *
15.6	6.203	-12.718	18.921	8.586
31.25	12.172	-22.551	34.723	25.104
62.5	18.911	-26.131	45.042	34.785
125	19.813	-26.324	46.137	35.165
250	19.925	-26.637	46.562	34.098
500	19.749	-26.382	46.131	35.176

*** Log(LR) for non-donor exceeded log(LR) for known donor: sample set 2**

C. Reproducibility (LTDNA SAMPLES)

The reproducibility of TrueAllele Casework was further assessed by running duplicate identification requests for all Identifiler Plus sensitivity set profiles. Match statistics were highly reproducible for all samples with a DNA input greater than 15.6pg (Figure VI.C.1 and Table VI.C.1.). The largest standard deviation was associated with the 15.6pg samples, where a mean difference of 9.857 log units was noted between both sensitivity sets. Within-group KL statistics were reproducible across all LTDNA samples, with the previously noted trend of increasing KL values with increasing DNA input (Figure VI.C.2.). Given that match scores at or below 7.8pg were generally not reliable or reproducible, KL scores at or below 5 (corresponding to similar DNA input levels) should be viewed with caution.

Figure VI.C.1. Mean log(LR) match scores as a function of DNA input and replicate amplification. Data include two amplified sets of serially diluted single source samples (n = 32). All samples were solved in duplicate in TrueAllele with 25K/25K burn-in and read out cycles. Error bars represent one standard deviation; dashed line set at zero.

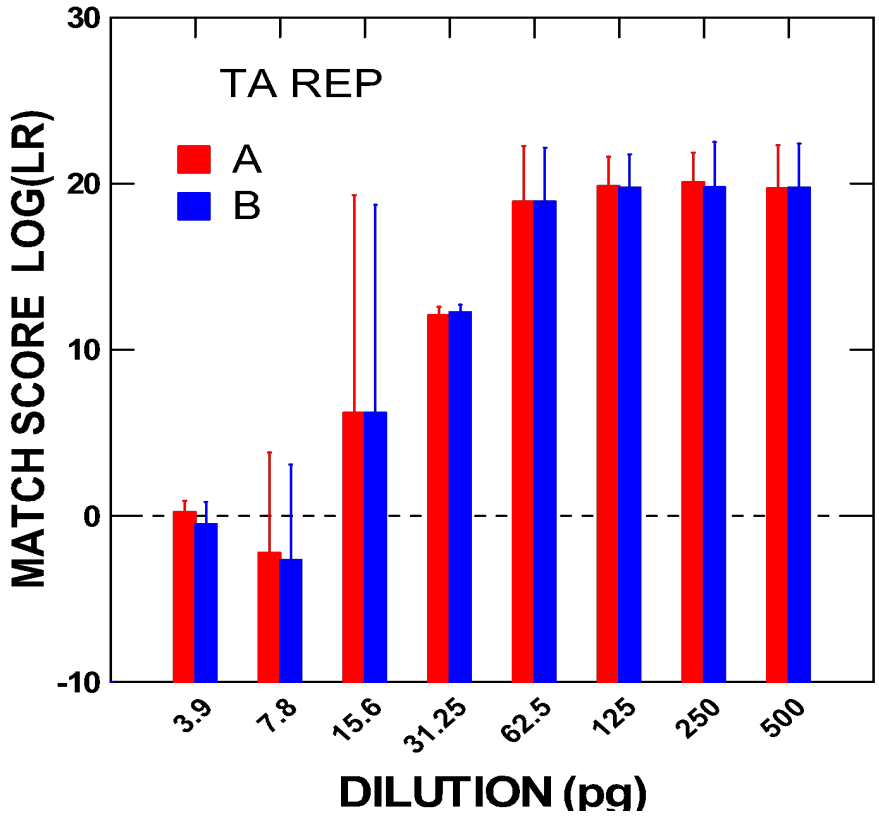


Figure VI.C.2. Mean KL scores as a function of DNA input. Data include TrueAllele Casework requests for two amplified sets of serially diluted single source samples (n = 32). All samples were solved in duplicate with 25K/25K burn-in and read out cycles. Error bars represent one standard deviation.

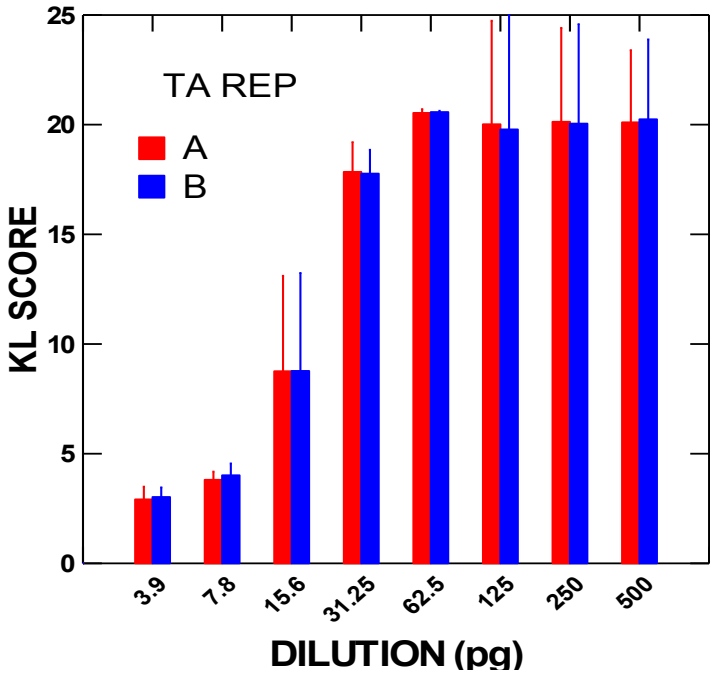


Table VI.C.1. Mean log(LR) match statistics and KL scores with associated standard deviations as a function of DNA input. Values from all replicated single unknown donor requests are pooled (n = 32).

DNA INPUT (pg)	Mean Donor Log(LR)	Standard Deviation	Mean KL score	Standard deviation
3.9	-0.127	0.625	2.971	0.236
7.8	-2.430	2.623	3.912	0.238
15.6	6.203	5.692	8.765	1.959
31.25	12.172	0.225	17.808	0.546
62.5	18.911	1.462	20.552	0.063
125	19.813	0.832	19.899	2.279
250	19.925	1.037	20.089	1.960
500	19.749	1.155	20.173	1.546

Jay Caponera
Bioscience Casework
15 July, 2013

VII. Experiment 3 Results: MCMC Burn-In/Read Out Study

A. Short Cycle Time Comparisons (Single Donor)

Reproducibility of TrueAllele Casework results for single donor requests with varying MCMC cycle times was assessed in this study. Differences in the resultant log(LR) match statistics as a function of MCMC cycle time were not statistically significant (Kruskal-Wallis p = 0.854; n = 24). The effect of varied MCMC cycle time on KL scores was also not statistically significant (p = 0.971). Results indicate that both log(LR) statistics and KL values for single unknown donor samples exhibiting stochastic effects can be run with a minimum of 5K/5K MCMC cycles with no deleterious effects to match statistic reproducibility, magnitude, or the associated information content of the evidence profiles (see Figures VII.A.1 and VII.A.2).

Figure VII.A.1. Comparison of mean log(LR) match statistics as a function of varied MCMC cycle time and DNA input. Data from the 25K requests were taken from sample set 2 of Study 2 (Sensitivity); error bars represent one standard deviation.

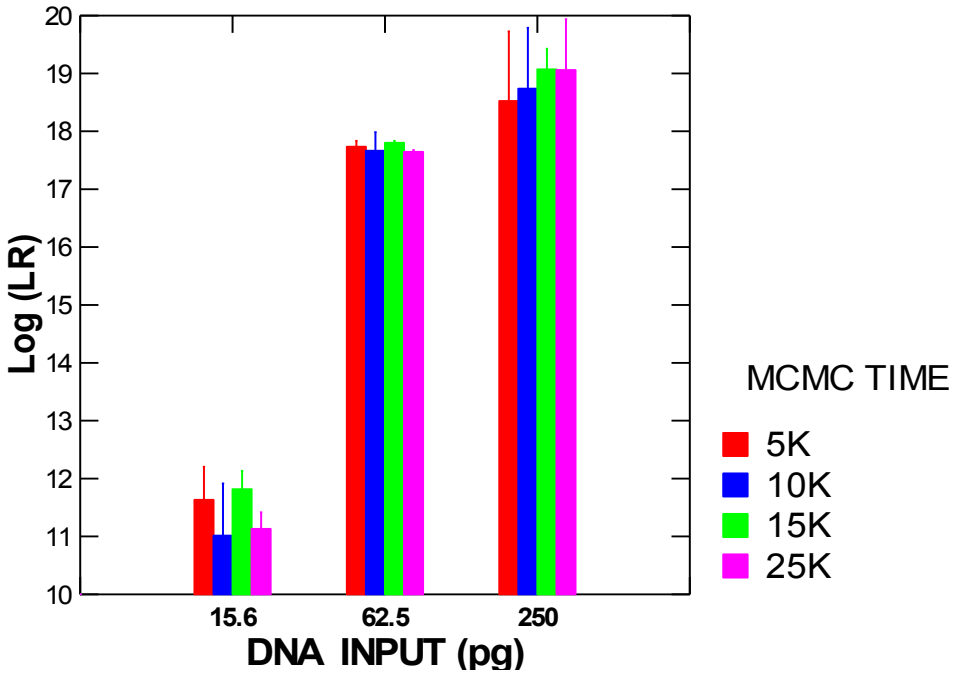
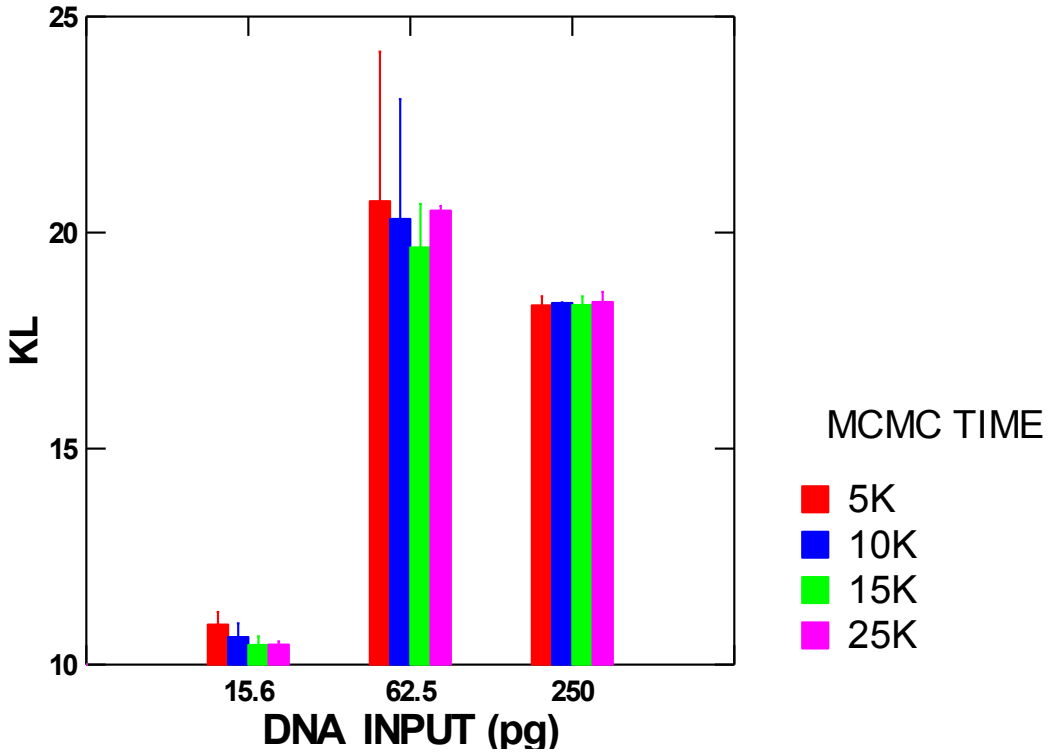


Figure VII.A.2. Comparison of mean KL scores as a function of varied MCMC cycle time and DNA input. Data from the 25K requests were taken from sample set 2 of Study 2 (Sensitivity); error bars represent one standard deviation.



B. Extended Cycle Time Comparisons (Two Person Mixtures)

Similar to the results from the short cycle study above, differences in log(LR) match statistics as a function of the extended MCMC cycle times were not statistically significant (ANOVA p = 0.853; Figure VII.B.1.). The effect of varied MCMC cycle time on KL scores was also not statistically significant (Kruskal Wallis p = 0.975; Figure VII.B.2.). Further, no significant differences in mixture weight standard deviations were detected across the cycle times tested (Kruskal Wallis p = 0.759; Figure VII.B.3.). Results from all metrics described above indicate that shorter MCMC run cycles of 25K/25K perform just as well as longer run times without sacrificing match statistic reproducibility, mixture weight inference, or information content from two person mixture profiles.

Figure VII.B.1. Comparison of log(LR) match statistics resulting from extended cycle times and grouped by mixture set (n = 36). Error bars represent one standard deviation.

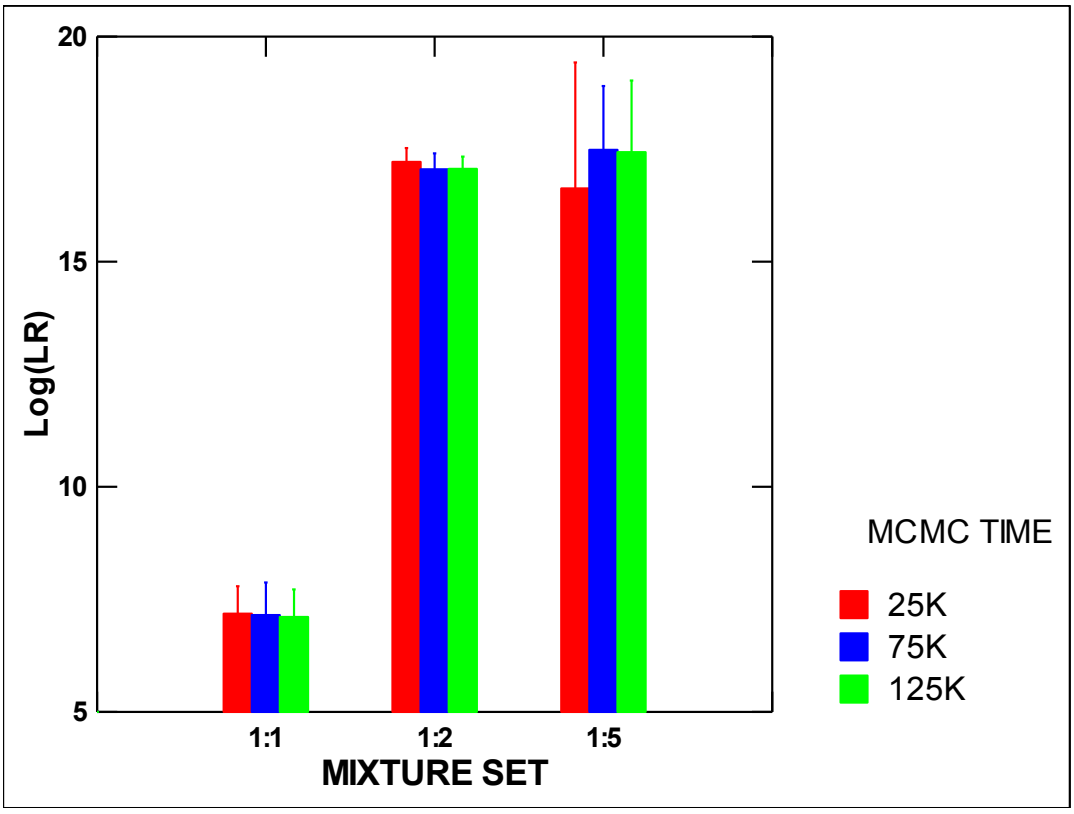


Figure VII.B.2. Comparison of KL scores resulting from extended cycle times and grouped by mixture set (n = 36). Error bars represent one standard deviation.

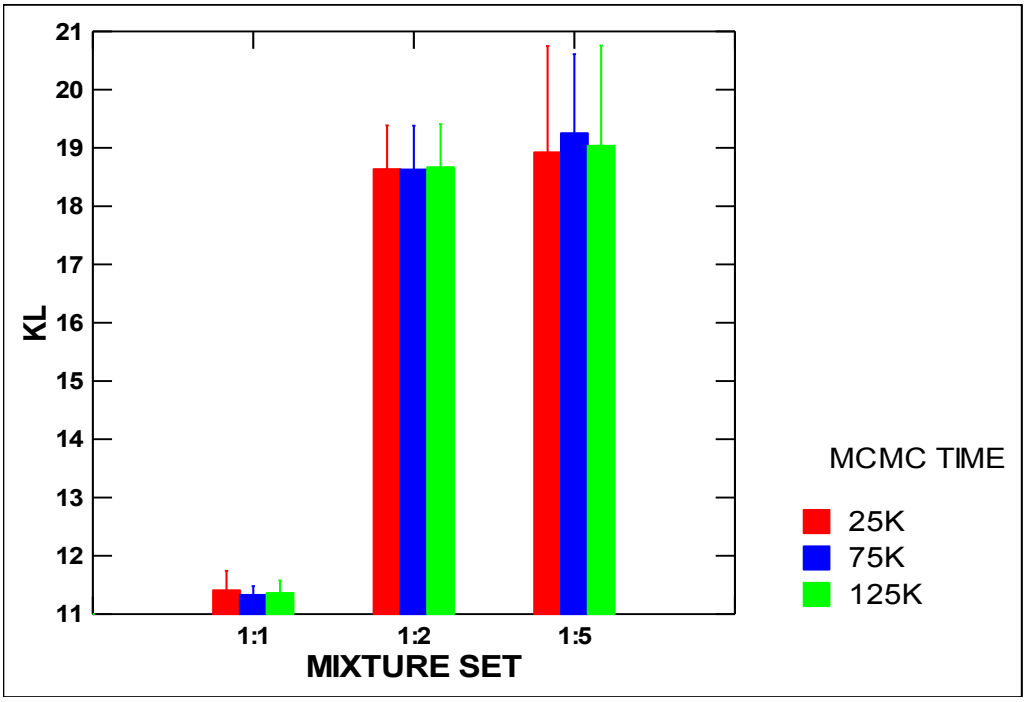
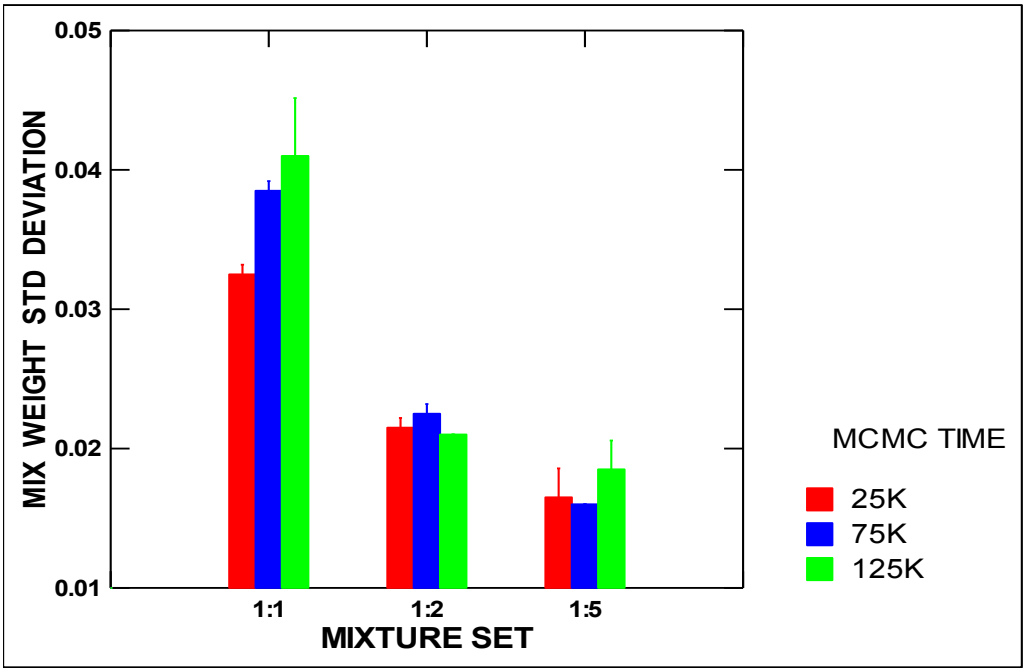


Figure VII.B.3. Comparison of mixture weight standard deviations resulting from extended cycle times and grouped by mixture set (n = 36). Error bars represent one standard deviation.



C. Extended Cycle Time Comparisons (Three Person Mixtures)

The general pattern of long and short MCMC run time equivalence described above with the two person mixture data is extended here with the three person data set. Differences in log(LR) match statistics as a function of the extended MCMC cycle times were not statistically significant (Kruskal Wallis p = 0.698; Figure VII.C.1.). The effect of varied cycle time on KL scores was also not statistically significant (Kruskal Wallis p = 0.681; Figure VII.C.2.). While some variation exists between cycle time mixture weight standard deviations, particularly at the 75K/75K level, differences detected were not significant (Kruskal Wallis p = 0.133; Figure VII.C.3.). Further investigation of pooled three person cycle time observations revealed a slight inverse relationship between mixture weight standard deviations and log(LR) values that was not statistically significant (Regression ANOVA p = 0.325; Figure VII.C.4.). However, regression analysis between mixture weight standard deviation and KL scores showed a steeper inverse relationship that was significant (Regression ANOVA p = 0.002), indicating that mixture weight inference is less sharply defined with samples exhibiting low information content (Figure VII.C.5.), but overall match information is unaffected. Results from all metrics described above indicate that shortened run cycles of 50K/50K for three person mixture samples perform equally well as longer cycle times with minimal effect on data quality.

Figure VII.C.1. Comparison of log(LR) match statistics resulting from extended cycle times and grouped by mixture set (n = 54). Error bars represent one standard deviation.

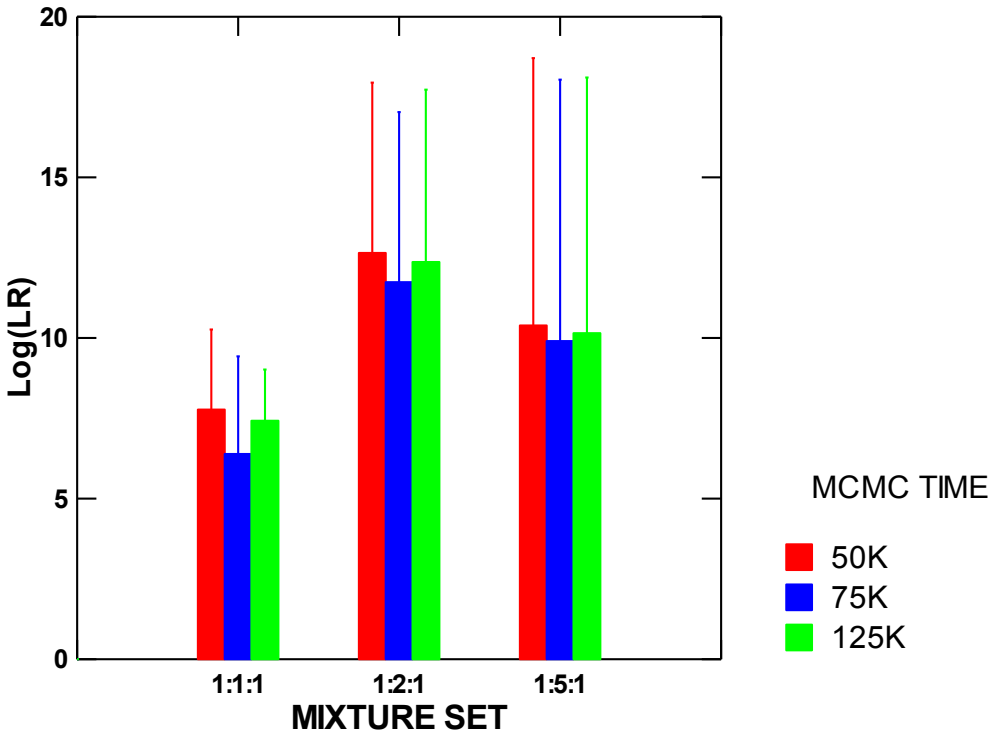


Figure VII.C.2. Comparison of KL scores resulting from extended cycle times and grouped by mixture set (n = 54). Error bars represent one standard deviation.

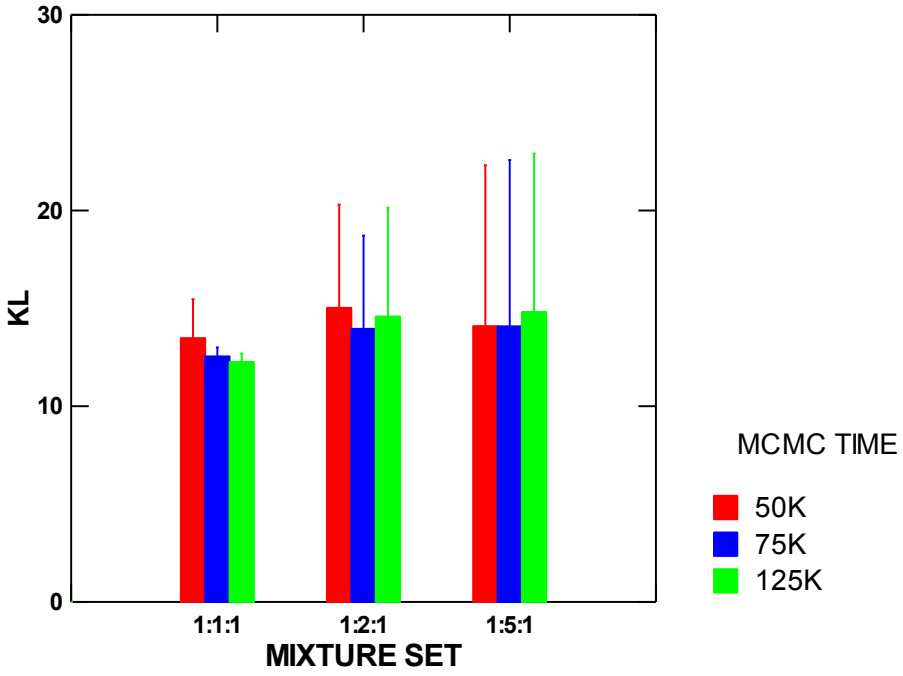


Figure VII.C.3. Comparison of mixture weight standard deviations resulting from extended cycle times and grouped by mixture set (n = 54). Error bars represent one standard deviation.

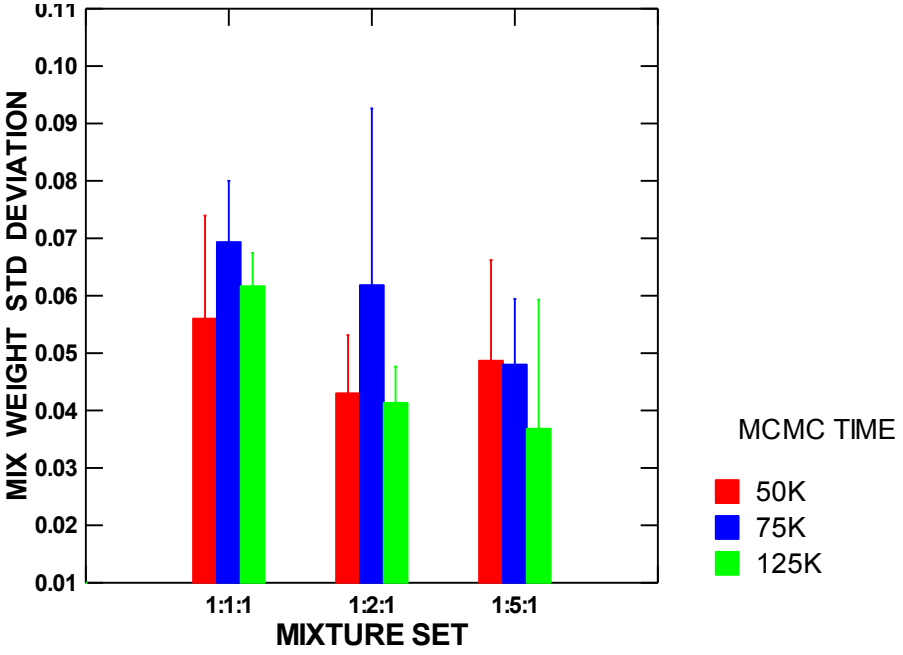


Figure VII.C.4. Scatterplot of three person mixture weight standard deviations against log(LR) values with all MCMC cycle times pooled (n = 54).

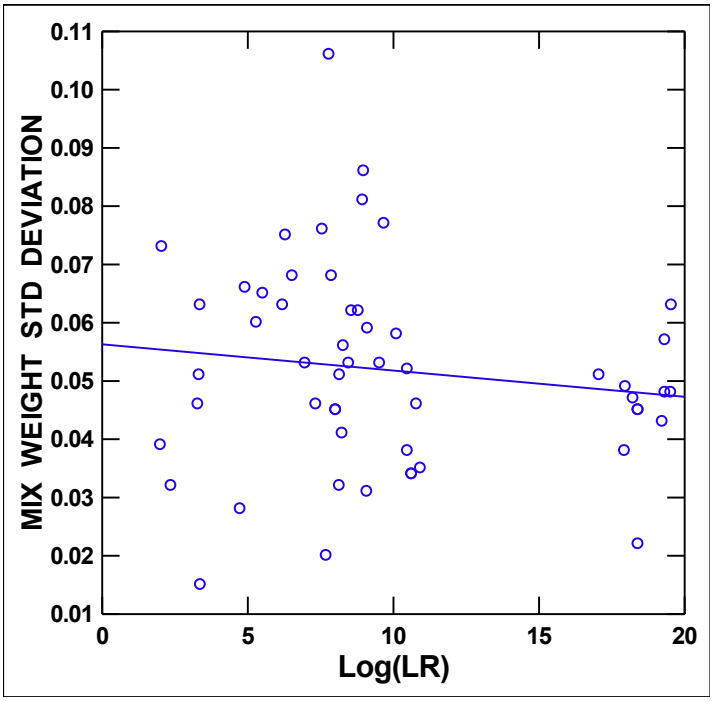
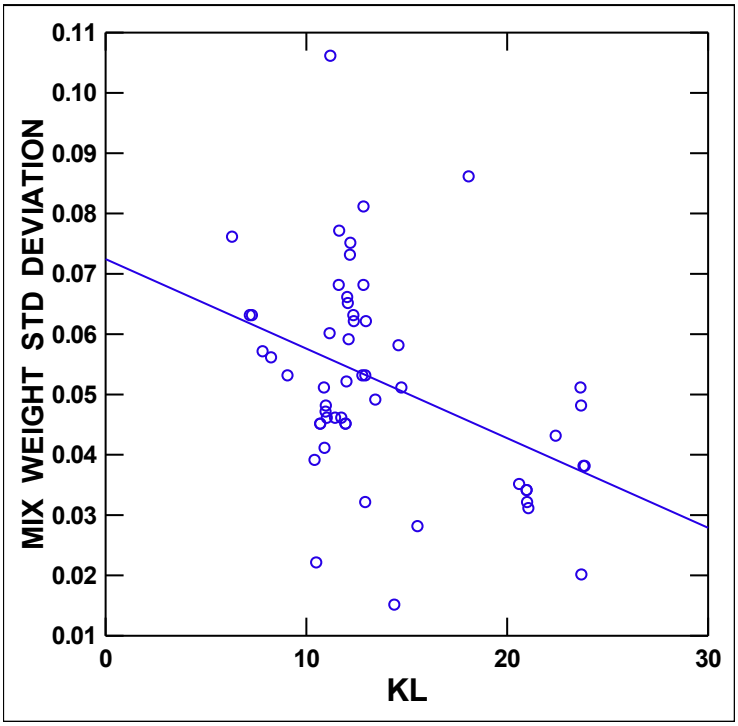


Figure VII.C.5. Scatterplot of three person mixture weight standard deviations against KL score values with all MCMC cycle times pooled (n = 54).



Jay Caponera
Bioscience Casework
19 July, 2013

VIII. Conclusions/ Recommendations

Results from the two and three person mixture component of the validation suggest that TrueAllele Casework probabilistic genotyping software is capable of providing robust and reproducible match statistics when CPI cannot. The overall information gain provided by TrueAllele is typically several orders of magnitude greater than the threshold-based inclusion method. These findings suggest that the software is more sensitive, uses more of the available genetic information, and will ultimately expand the scope of suitable evidentiary samples by providing weight of evidence to profiles that are currently deemed inconclusive.

For all mixture samples tested, TrueAllele Casework demonstrated consistently that log(LR) match statistics were specific to known donors while all non-donors yielded negative log(LR) values. Specificity results from the single source Sensitivity Experiment also support this finding, and further show that TrueAllele offers a significant increase in information gain as compared to CMP for low template samples. However, it is clear from the Sensitivity data that a lower limit (~15.6pg, or the approximate equivalent of five diploid cells) can be placed on the software, below which results may not be reproducible. All genotype inferences at or above 15.6pg were highly reproducible with respect to both information content (KL statistics) and log(LR) values. Based on these data, FIC scientists may confidently conclude that given appropriate DNA input, the resulting match information provided by TrueAllele Casework will be both accurate and reliable for reporting.

Data from the MCMC cycle time study show that run times of as little as 5K/5K for single source samples are comparable to longer run times of 25K/25K, even for low template samples exhibiting stochastic effects. Longer run times of 75K/75K and 125K/125K for the mixture sets also show no significant improvement in either information content or log(LR) match statistics, suggesting that shortened cycle times can be used to increase processor efficiency while still maintaining high data quality. Regardless of the cycle time used for a given mixture, the validation data support a thorough assessment of all convergence diagnostics (Gelman-Rubin statistics and Markov chain histogram and history graphs) before reporting to ensure the integrity of TrueAllele mixture weight inference. Analysts should also note that DNA mixtures exhibiting little separation between contributors may have reduced information content and therefore reduced match statistics.

The validation findings herein support and extend the data from multiple peer-reviewed Journal of Forensic Sciences publications from the New York State Police Forensic Investigation Center in collaboration with Cybergenetics staff, and from the original TrueAllele Casework validation approved May 2, 2013.

IX. Glossary of Terms

Ban: A unit of measure based on $\log_{10}$. For example, 3 ban is the equivalent of 3 $\log_{10}$ units or 1,000.

Burn-In: The initial set of cycles in an MCMC search used to search for the underlying stationary distributions of the variables of interest. Burn-in cycles are not used in computation.

Coancestry Coefficient: The probability of an allele being identical by descent (IBD). Also known as theta, the coancestry coefficient of 1% selected in TrueAllele assumes that we expect no more than 1% of all alleles to be IBD.

Convergence: The point at which it is reasonable to assume test samples are truly representative of the underlying stationary distribution of the Markov chain.

Cross-Matching: A match for multiple inferred contributors to the same reference sample. Cross-matching is more likely to occur when there is little separation between inferred genotypes or when mixture weights are approximately equal.

Gelman-Rubin (GR) statistic: A statistical technique used to monitor the convergence of MCMC output by comparing within-chain and between-chain variances. GR values of approximately 1 are expected when all chains have escaped the influence of their starting points and traversed all of the target distribution (i.e. with-in chain variance dominates between chain variance). The approach emphasizes reducing bias in estimation.

Inferred Genotype: The concentration of probability on a certain allele pair for a given locus.

Kullback-Leibler (KL) statistic: Measures the divergence of an inferred profile from the prior distribution. KL statistics give an indication of the expected log(LR) based on profile information content. More informative profiles will typically exhibit higher KL values.

Likelihood Ratio: A standard measure of information that summarizes the data support for the identification hypothesis in a single number. The LR is also the TrueAllele match statistic used in DNA reporting, and compares the probability of an evidence match to the probability of a match by coincidence.

TrueAllele LR = Probability (Evidence Match) / Probability (Coincidental Match)

Monte Carlo Markov Chain (MCMC): A Bayesian method of integration that samples successively from a target distribution, with each sample depending on the previous one (hence the Markov chain). Monte Carlo integration achieves statistical inference by averaging the Markov chain samples.

Read-Out: The second set of cycles in an MCMC search where posterior probability distributions for all variables are determined.

Standard Deviation: A measure of the variation or dispersion from the average. TrueAllele Casework lists the standard deviations from the mixture weight inferences of each donor. Low values indicate data points are centered closely on the average. Larger values indicate greater uncertainty in mixture weight inference, with data points spread over a larger range.