New York State TrueAllele® Validation on DNA Mixtures of Known Composition

Mark W. Perlin[1], PhD, MD, PhD, Jennifer Hornyak[1], MS
Jay Caponera[2], MS, Barry Duceman[2], PhD

[1]Cybergenetics, Pittsburgh, PA
[2]Forensic Investigation Center, New York State Police, Albany, NY

October 15, 2013

Corresponding Author:

Mark W. Perlin, PhD, MD, PhD
Chief Scientific Officer
Cybergenetics
160 North Craig Street, Suite 210
Pittsburgh, PA 15213 USA
(412) 683-3004
(412) 683-3005 FAX
perlin@cybgen.com

Table of Contents

Introduction

DNA mixtures contain two or more contributors, and are a common form of biological evidence. Mixtures arise naturally in rape, homicide and property crimes. The interpretation of mixture evidence considers more genotype possibilities than in single source examination, and so poses challenges for manual review.

Short tandem repeat (STR) testing of DNA mixtures produces a quantitative pattern of allelic (and other) peaks at a locus (1). The peak height pattern expresses a sum of contributing genotypes, with each allele pair appearing in rough proportion to its contributor's DNA amount (2). The STR pattern shows natural peak height variation arising from known polymerase chain reaction (PCR) artifacts (e.g., stutter (3), relative amplification), as well as random PCR amplification effects (4). Other variation factors include degraded or inhibited DNA template, volumetric sampling error, baseline noise, and random artifact deviations.

Mathematical models have been developed for PCR (5) and STR (6, 7) experiments. These models can predict DNA mixture data patterns and their statistical variation, often using a hierarchical Bayesian model (8). Computer systems have been developed that can solve these probability equations (9-13). Markov chain Monte Carlo (MCMC) (14) statistical search is used for larger models that have more variables.

Computer interpretation of DNA mixtures has several potential advantages, relative to manual data review:

- sensitivity: extract more identification information from the same data
- specificity: reduce false matches and quantify DNA exclusion
- reproducibility: provide consistent results from independent data analyses

Implementing these computer features could improve laboratory workflow, reduce DNA backlogs, and provide better information to the criminal justice system.

TrueAllele® Casework is the most established computer system for DNA mixture interpretation. Developed by Cybergenetics (Pittsburgh, PA) 15 years ago, TrueAllele has been used in the criminal casework since 2009 (15). Over a hundred TrueAllele reports have been filed in criminal cases, resulting in many guilty pleas and convictions for serious crimes. TrueAllele has been approved for use in casework by the New York State Commission on Forensic Science and its DNA Subcommittee, and has withstood admissibility challenges in three states. TrueAllele has appellate precedent in Pennsylvania (16).

The New York State Police (NYSP) Forensic Investigation Center has conducted extensive TrueAllele validation studies on DNA mixtures. One published NYSP peer-reviewed study (9) of casework evidence items established the greater sensitivity of TrueAllele, relative to manual review. Another NYSP peer-reviewed paper (17) demonstrated TrueAllele's high specificity and reproducibility on a larger casework data set. Other TrueAllele validation studies have been conducted on DNA mixtures of known composition (18, 19).

This NYSP TrueAllele Casework validation study assesses the system's performance on laboratory synthesized DNA mixtures of known composition. The mixtures were constructed from two or three known contributors in known ratios, and

tested using ABI's Identifiler® Plus STR panel. The PCR products were read out on both ABI 3130 and 3500 xl genetic analyzers. The sensitivity, specificity and reproducibility of the TrueAllele genotyping system were assessed using the DNA match statistic (20), which is a standard measure of identification information (21).

Methods

STR data

Three DNA sample groups were comprised of different individuals in two contributor mixtures, each having the integer ratios 1:19, 1:9, 1:5, 1:2, 1:1, 2:1, 5:1, 9:1 and 19:1. Two other sample groups were three contributor mixtures, each constructed from known genotypes in the integral ratios 1:1:1, 1:2:1, 1:5:1, 1:10:1, 1:2:3, 2:2:1 and 3:3:1. The total DNA input for each sample was 1 ng. Every item was amplified twice.

Sorenson Forensics (Salt Lake City, UT) amplified the samples using the ABI Identifiler® Plus STR panel, and analyzed the data on ABI 3130xl and 3500xl sequencers using 5 and 10 second injection times.

The STR electropherogram data were recorded in .fsa files by the genetic analyzer. The New York State Police sent the data to Cybergenetics in October 2012 and March 2013 in batches that were organized by sequencer type. Some of the mixture references were provided as text listings in files.

Genotype inference

The TrueAllele system uses hierarchical Bayesian probability modeling (22) to represent genotypes and data in a DNA mixture problem. TrueAllele considers all the STR data and many explanatory variables when solving DNA mixtures. Using MCMC statistical sampling, the system infers genotypes, mixture weights, and other variables, inferring probability distributions for each one. The computer inference is *thorough*, considering tens of thousands of STR pattern possibilities, and *objective*, working solely from the evidence data without knowing subject references.

TrueAllele has a client-server architecture. All human analyst activities are conducted on a client computer workstation. STR data and interpretation requests are uploaded to a central computer server. This server houses a relational PostgreSQL database, and conducts fully automated genotype interpretation and match operations in parallel on multiple processors. Computed answers are stored on the database. To review computed results, an analyst downloads data, genotypes and other information to their visual workstation.

Match information

In order to quantify the strength of match between evidence and a reference, TrueAllele compares their genotypes, relative to a population. This match comparison is done only after the computer has objectively (i.e., without knowledge of the references) inferred

contributor genotypes for a mixture problem. This objective approach eliminates issues that might arise from subjective examination bias (23, 24).

The match information for each known contributor in a mixture item was calculated as a likelihood ratio (LR), and collated as logarithms. The $\log_{10}$(LR) is a standard measure of information expressed in "ban" units (25). This information can be used to assess forensic match sensitivity, specificity, and reproducibility (17). TrueAllele can present its calculated LR values textually or visually in multiple ways, depending on the explanatory context (26).

Procedure

Design

The study was divided into 24 subgroups comprising two contributor numbers, two injection times, two sequencer types and three cycle numbers (Table 1). This grouping of experiments permitted a more refined statistical analysis.

Processing

The .fsa electropherogram files were processed through the TrueAllele Casework Visual User Interface (VUIer™) Analyze module. The quality checked peak data were then uploaded to a TrueAllele processing database in the Data module.

A trained TrueAllele operator created interpretation requests in the VUIer Request module after downloading DNA data from the database. Each sample was processed assuming the same number of unknown contributors as the actual known number. The mixture requests were processed in duplicate, with burn-in/read-out times of 25,000/25,000, 50,000/50,000, and 100,000/100,000.

Reporting

After TrueAllele processing was completed, the inferred evidence genotypes were compared to known reference genotypes. This comparison was conducted in the VUIer Report module, which calculated log(LR) match values. A contributor genotype's corresponding reference was identified through its maximum match score. A total of 2,448 genotype comparisons (Table 1b) were formed from 47 items (Table 1a).

The reported match statistic was the average value of the replicated computer log(LR) results. For each genotype comparison, the smallest of three FBI ethnic population (African-American, Caucasian and Hispanic) log(LR) values was recorded. The co-ancestry coefficient (theta value) was set to 1%.

Sensitivity, specificity and reproducibility were assessed across the entire set of data (sequencers, injection times, amplifications) for both two contributors (ncon2) and three contributors (ncon3). The log(LR) mean, standard deviation, and within-group standard deviation were calculated for each of the 24 subgroups.

Results

The information content of a genotype inferred from STR mixture data can be quantified through a log(LR) value. This value is obtained by comparing the contributor genotype with a reference genotype (known from the experimental design), relative to a population. Quantifying STR data and genotype information in this way permits the development of an empirical frequency distribution. An information distribution can be examined both visually and statistically. All TrueAllele-inferred genotypes, both major and minor, were included in these information characterizations.

Sensitivity

Sensitivity measures the extent to which a mixture interpretation method correctly includes a true contributor. TrueAllele's information sensitivity was examined separately for two and three contributor mixtures.

The log(LR) match frequency distributions are shown for each mixture contributor number (Figure 1). The log(LR) values were calculated as the average of two independent computer runs. The vast majority of log(LR) values fell well to the right of zero information, indicating high match sensitivity with few false exclusions. As the number of contributors increased from two to three (Figure 1, a & b), the information distribution shifted to the left as it shrank towards zero.

The statistics in Table 2 show that the average log(LR) value was around 14 ban (a hundred trillion) for two contributors. With three contributors, this information average

fell to about 8 ban (a hundred million). The standard deviations for two or three contributors were comparable. Some negative log(LR) values were observed (Figure 1, and Table 2a row "min"), indicating that the genotype was falsely excluded from having contributed to its mixture of known composition.

Table 2b counts the occurrence of false exclusions. With two contributors, there were 9 events out of 1,296 total genotype comparisons, for a false exclusion rate of 0.694%. This error rate was the same with three contributors, where there were 7 false negatives out of 1,008 genotype comparisons.

Specificity

Specificity measures the extent to which a mixture interpretation method correctly excludes a non-contributor. To evaluate specificity, each inferred evidence genotype was compared with a thousand randomly generated genotypes drawn from a population. Three different FBI ethnic populations were used, for a total of 3,000 comparisons per evidence genotype. Comparing a contributor genotype (separated from a mixture by TrueAllele inference) with a random genotype should produce an exclusion, which can be quantified by its negative log(LR) value.

The negative match information distributions for the inferred mixture genotypes are shown (Figure 2). The vast majority of TrueAllele's log(LR) values fell far to the left of zero information, indicating high match specificity that supports true exclusions.

TrueAllele showed considerable exclusionary power. The average log(LR) value was around -24 ban with two contributor mixtures (Table 3a). With three contributors,

the average log(LR) value was around -18 ban. This reduction in the (absolute value of) exclusionary information showed a shrinkage towards zero as the number of contributors increased.

False inclusion rates were determined by counting how many positive log(LR) values occurred in the random match comparisons. With two contributors, there were 18 false positive scores out of 3,888,000 genotype comparisons (across all three ethnic groups) for an error rate of 0.000463%. With three contributors, 79 false inclusions were found out of 3,024,000 comparisons for an error rate of 0.00261%. The false inclusion rate across the entire data set was 0.00140%, or approximately 1 in 70,000.

One false inclusion was seen with a LR value over 100. Out of 6,912,000 total comparisons, this gave a false inclusion rate of 0.0000147% that was well under 1 in a million. No false inclusions were seen when the LR exceeded 1,000.

Reproducibility

Reproducibility measures how precisely the identification information is repeated in independent computer runs on the same mixture data. To assess TrueAllele's reproducibility, comparison was made between the identification information obtained in duplicate concordant computer runs on the same sample.

Figure 3 shows a reproducibility scatterplot for two and three contributors. Each point gives the log(LR) results from a first (x-axis) and second (y-axis) computer run. Since the points resided near the equal information line (i.e., y = x), the computer analyses were seen to be reproducible. The scatterplot width with two contributors

(Figure 3a) was narrower than with three contributors (Figure 3b), indicating greater reproducibility.

Within-group standard deviations were calculated to quantify computer run reproducibility. Calculated over four independent computer runs, the values for each subgroup are listed in Table 4. The within-group standard deviation (σ_w) averaged 1.10 ban across the 24 groups, indicating good reproducibility for the mixture analyses.

The σ_w precision was relatively constant within a subgroup as the MCMC sampling time was increased from 25,000 to 100,000 read out cycles. This relative invariance indicates that 25,000 sampling cycles may be sufficient for these data.

Overall, the 5 second injections showed more information consistency than the 10 second injections. Moreover, the two contributor mixtures had better reproducibility than the three person mixtures when using 5 second injections. Of note, TrueAllele measured smaller σ_w values for the 3500 sequencer than with the 3130, suggesting improved information reproducibility with the newer sequencer model.

Subgroups

The data were divided into 24 subgroups to permit more refined analysis (Table 1). Table 4 shows the sensitivity (mean and standard deviation) and reproducibility (within-group standard deviation) for each subgroup.

Some qualitative information contrasts warrant mention. With two contributors, the identification information averaged about 14 ban; there was less information with

three contributors, averaging around 8 ban. The 3500 sequencer showed better reproducibility (smaller σ_w) than the 3130 model.

Conclusion

DNA mixtures are an abundant source of biological evidence, and can be critical to a criminal investigation or prosecution. Manual review of STR mixtures often understates their probative value, or discards them entirely as "inconclusive" (27). Review consistency between human analysts is not assured. These long-standing issues have been underscored by recent manual interpretation guidelines that propose making even less use of the available evidence (28).

The NYSP Forensic Investigation Center has pioneered the use of automated computer interpretation of DNA samples. They were the first to validate an expert system for reference samples, and publish their findings (29). And they were the first to validate a genotype modeling system for interpreting DNA mixture casework items (9, 17). The forensic DNA statistical community now advocates using these more informative genotype modeling computer methods (30).

This validation study extends the earlier NYSP scientific results on the TrueAllele Casework system to DNA mixtures of known genotype composition. Assessing performance on known mixtures containing two or three contributors, the results quantified TrueAllele's sensitivity, specificity and reproducibility. These results validated TrueAllele's applicability as a useful computational tool that performs reliably on DNA mixtures for forensic casework.

References

1. Gill P, Sparkes R, Pinchin R, Clayton TM, Whitaker JP, Buckleton J. Interpreting simple STR mixtures using allele peak area. Forensic Sci Int. 1998;91:41-53.
2. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sci. 2001;46(6):1372-7.
3. Perlin MW, Lancia G, Ng S-K. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am J Hum Genet. 1995;57(5):1199-210.
4. Gill P, Curran J, Elliot K. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. Nucleic Acids Res. 2005;33(2):632-43.
5. Stolovitzky G, Cecchi G. Efficiency of DNA replication in the polymerase chain reaction. Proc Natl Acad Sci USA. 1996 November 12;93(23):12947–52.
6. Perlin MW. Simple reporting of complex DNA evidence: automated computer interpretation. Promega's Fourteenth International Symposium on Human Identification, 2003; Phoenix, AZ. 2003.
7. Curran J. A MCMC method for resolving two person mixtures. Sci Justice. 2008;48(4):168-77.
8. Gelman A, Carlin JB, Stern HS, Rubin D. Bayesian Data Analysis. Boca Raton, FL: Chapman & Hall/CRC, 1995.
9. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. J Forensic Sci. 2011;56(6):1430-47.
10. Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Identifying contributors of DNA mixtures by means of quantitative information of STR typing. J Comput Biol. 2012;19(7):887-902.
11. Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet. 2013;7(5):516–28.
12. Puch-Solis R, Rodgers L, Mazumder A, Pope S, Evett I, Curran J, Balding D. Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Forensic Sci Int Genet. 2013;7(5):555–63.
13. Cowell R, Graversen T, Lauritzen S, Mortera J. Analysis of DNA mixtures with artefacts. ArXiv. 2013:arXiv:1302.4404 (submitted).
14. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice: Chapman and Hall, 1996.
15. Perlin MW. The Blairsville slaying and the dawn of DNA computing. In: Niapas A, editor. Death Needs Answers: The Cold-Blooded Murder of Dr John Yelenic. New Kensington, PA: Grelin Press; 2013.
16. Commonwealth of Pennsylvania v. Kevin James Foley. Superior Court of Pennsylvania; 2011.
17. Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® Casework validation study. J Forensic Sci. 2013;58(6):DOI 10.1111/556-4029.12223

18. Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS ONE. 2009;4(12):e8327.
19. Ballantyne J, Hanson EK, Perlin MW. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: Combining quantitative data for greater identification information. Sci Justice. 2013;53(2):103-14.
20. Perlin MW. Scientific validation of mixture interpretation methods. Promega's Seventeenth International Symposium on Human Identification, 2006 Oct 10-12; Nashville, TN. 2006.
21. Aitken CG, Taroni F. Statistics and the Evaluation of Evidence for Forensic Scientists. Second ed. Chicester, UK: John Wiley & Sons, 2004.
22. O'Hagan A, Forster J. Bayesian Inference. Second ed. New York: John Wiley & Sons, 2004.
23. Thompson WC. Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation. Law, Probability and Risk. 2009;8(3):257-76.
24. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Science & Justice. 2011;51(4):204-8.
25. Good IJ. Probability and the Weighing of Evidence. London: Griffin, 1950.
26. Perlin MW. Explaining the likelihood ratio in DNA mixture interpretation. Promega's Twenty First International Symposium on Human Identification, 2010; San Antonio, TX. 2010.
27. Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, Morling N, Prinz M, Schneider PM, Weir BS. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Sci Int. 2006;160(2-3):90-101.
28. SWGDAM. Interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories. 2010.
29. Kadash K, Kozlowski BE, Biega LA, Duceman BW. Validation study of the TrueAllele® automated data review system. J Forensic Sci. 2004;49(4):1-8.
30. Kelly H, Bright J-A, Buckleton JS, Curran JM. A comparison of statistical models for the analysis of complex forensic DNA profiles. Science & Justice. 2013;http://dx.doi.org/10.1016/j.scijus.2013.07.003.

Tables

Table 1: Study Design. The 24 subgroups are comprised of two contributor numbers (ncon), two injection times (inj), two sequencers (seq), and three cycle numbers (cycles). Table (a) shows item totals, while (b) gives genotype totals.

(a) Item totals

ncon	*sets*	*weights*	*item totals*
1	3	2	6
2	3	9	27
3	2	7	14
		Overall	47

(b) Genotype totals

ncon	*sets*	*weights*	*seq*	*amp*	*inj*	*cycles*	*totals*
1	3	2	2	2	2	3	144
2	3	9	2	2	2	3	1296
3	2	7	2	2	2	3	1008
						Overall	2448

Table 2: Sensitivity. Statistics were calculated for 2 and 3 contributors, and combined results from the two sequencers. Table (a) shows the number, minimum, mean, median, standard deviation and maximum for 2 and 3 contributors, giving the log(LR) values in ban units. Table (b) shows the number of false exclusions occurring in each log(LR) interval (e.g., "0" indicates the interval [0, 1]).

(a) Summary statistics

ncon	**2**	**3**
N =	1,296	1,008
min	-4.767	-4.757
mean	14.384	8.558
median	16.367	8.453
std dev	4.655	4.224
max	20.049	17.153

(b) False exclusions

ncon	**2**	**3**
-1	3	2
-2	0	2
-3	5	1
-4	0	1
-5	1	1
Total	*9*	*7*

Table 3: Specificity. Statistics were calculated for 2 and 3 contributors across all three FBI ethnic populations. The sequencer results are combined. Table (a) shows the number of comparisons, along with the log(LR) minimum, mean, maximum and standard deviation. Table (b) gives the number of false inclusions, stratified by log(LR) interval (e.g., "0" indicates the interval [0, 1]).

(a) Summary statistics

ncon	**2**			**3**		
ethnicity	BLK	CAU	HIS	BLK	CAU	HIS
N =	1,296,000	1,296,000	1,296,000	1,008,000	1,008,000	1,008,000
min	-30.000	-30.000	-30.000	-30.000	-30.000	-30.000
mean	-24.543	-23.534	-23.997	-19.433	-17.413	-17.691
max	-0.004	2.029	0.831	1.645	1.865	1.756
std	4.840	5.193	5.094	5.697	5.863	5.825

(b) False inclusions

ncon	**2**			**3**		
ethnic	BLK	CAU	HIS	BLK	CAU	HIS
0	0	9	6	16	29	19
1	0	2	0	3	9	3
2	0	1	0	0	0	0
Total	*0*	*12*	*6*	*19*	*38*	*22*

Table 4: Reproducibility. The mean (μ), standard deviation (σ) and within-group standard deviation (σ_w) measure of reproducibility are shown for each of the 24 subgroups. Table (a) gives the log(LR) results for the 2 contributor groups, while table (b) gives results for the 3 contributor groups.

(a) 2 contributors

injection time		5 sec			10 sec		
seq	cycle	μ	σ	σ_w	μ	σ	σ_w
3130	25K	14.61	4.41	0.79	14.40	4.76	1.97
	50K	14.46	4.69	0.90	14.26	5.09	2.08
	100K	14.38	4.81	0.86	14.18	5.11	1.92
3500	25K	14.34	4.55	0.67	14.63	4.23	1.04
	50K	14.20	4.73	0.71	14.55	4.32	1.01
	100K	14.09	4.92	0.73	14.50	4.30	0.75

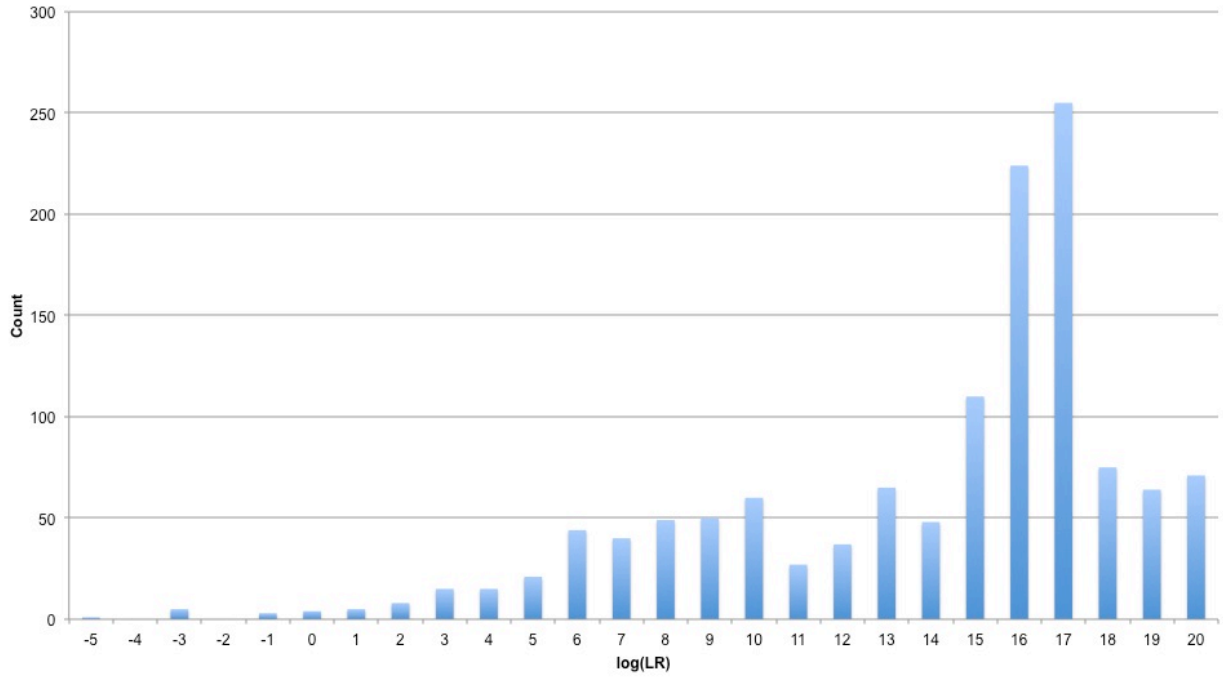
(b) 3 contributors

injection time		5 sec			10 sec		
seq	cycle	μ	σ	σ_w	μ	σ	σ_w
3130	25K	8.95	4.27	1.10	9.06	4.36	1.33
	50K	8.91	4.15	1.13	8.87	4.51	1.56
	100K	8.70	4.11	1.05	8.74	4.27	1.28
3500	25K	8.29	4.29	1.10	8.85	4.19	0.97
	50K	7.87	4.17	0.96	8.53	4.07	0.77
	100K	7.58	4.08	0.90	8.29	4.31	0.92

Figures

Figure 1: Sensitivity. Histograms show the log(LR) genotype match distribution for (a) 2 and (b) 3 contributor mixtures. Results from the two sequencers were combined.

(a) 2 contributors



(b) 3 contributors

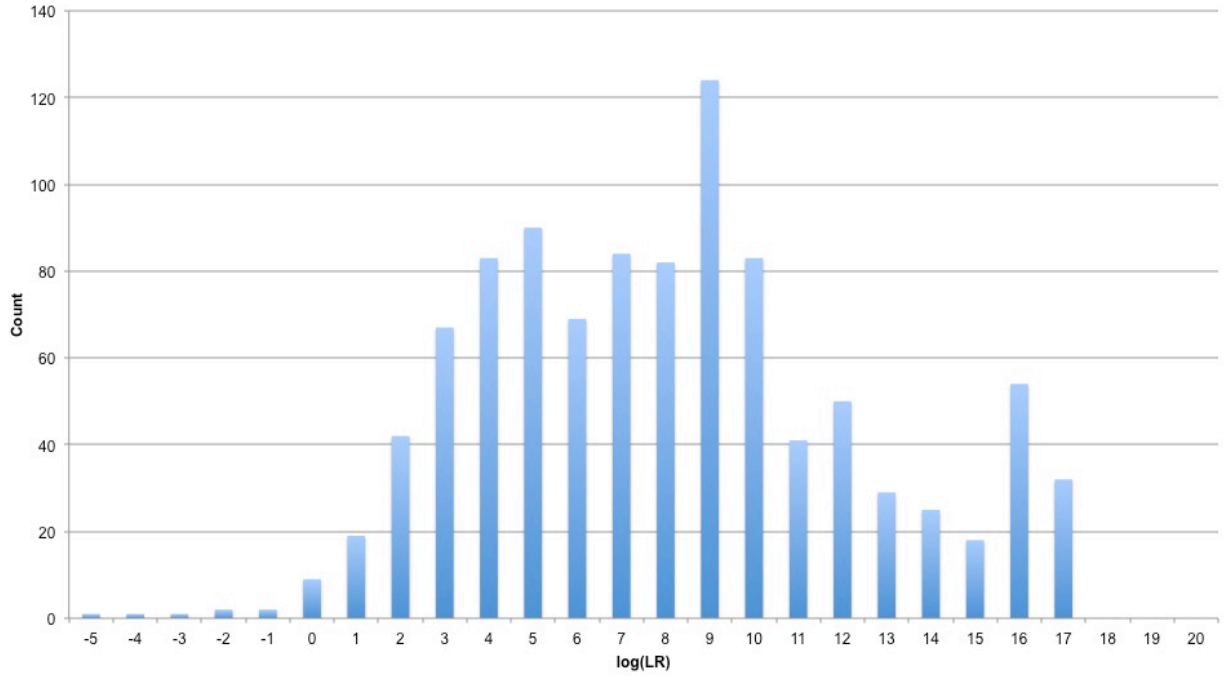
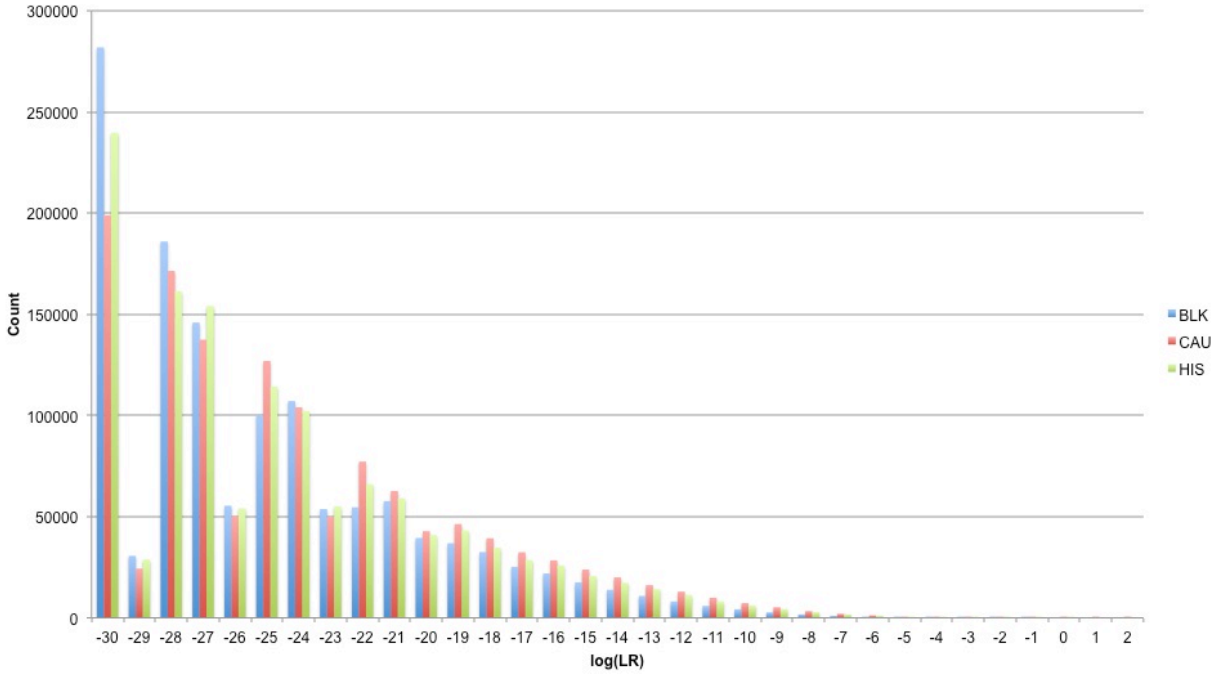


Figure 2: Specificity. Histograms show the log(LR) genotype match distribution for (a) 2 and (b) 3 contributor mixtures, relative to a thousand randomly generated profiles. Each ethnic population is depicted in a different color. Results from the two sequencers were combined.

(a) 2 contributors



(b) 3 contributors

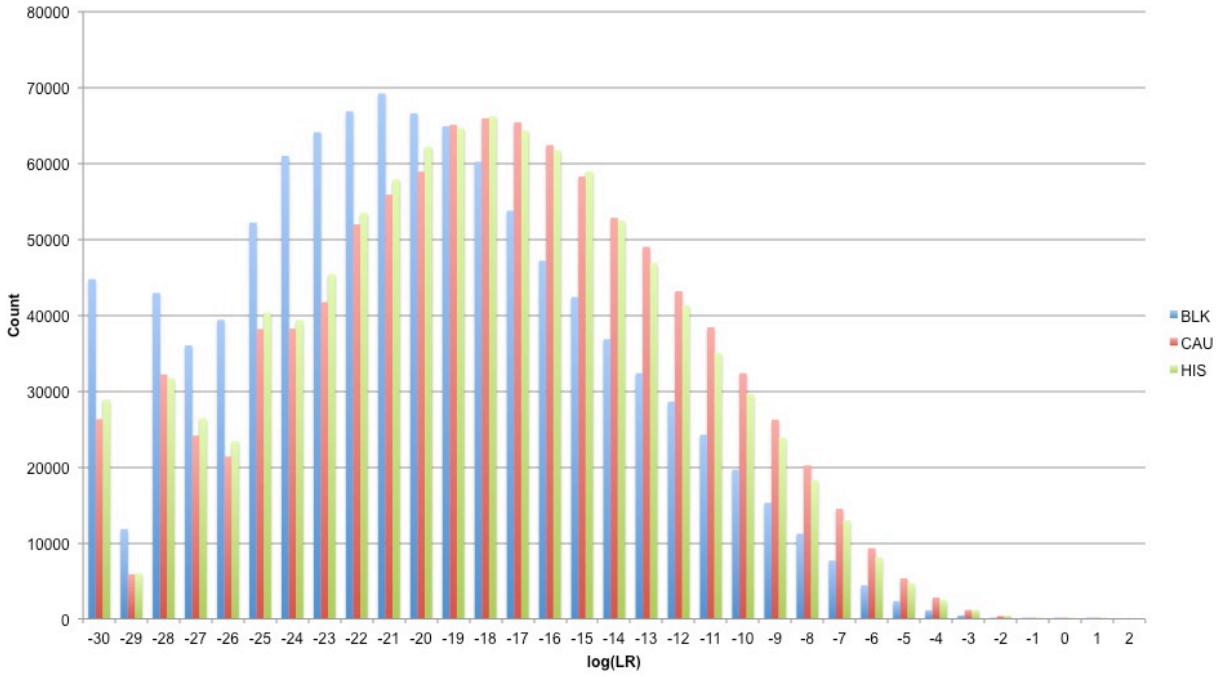
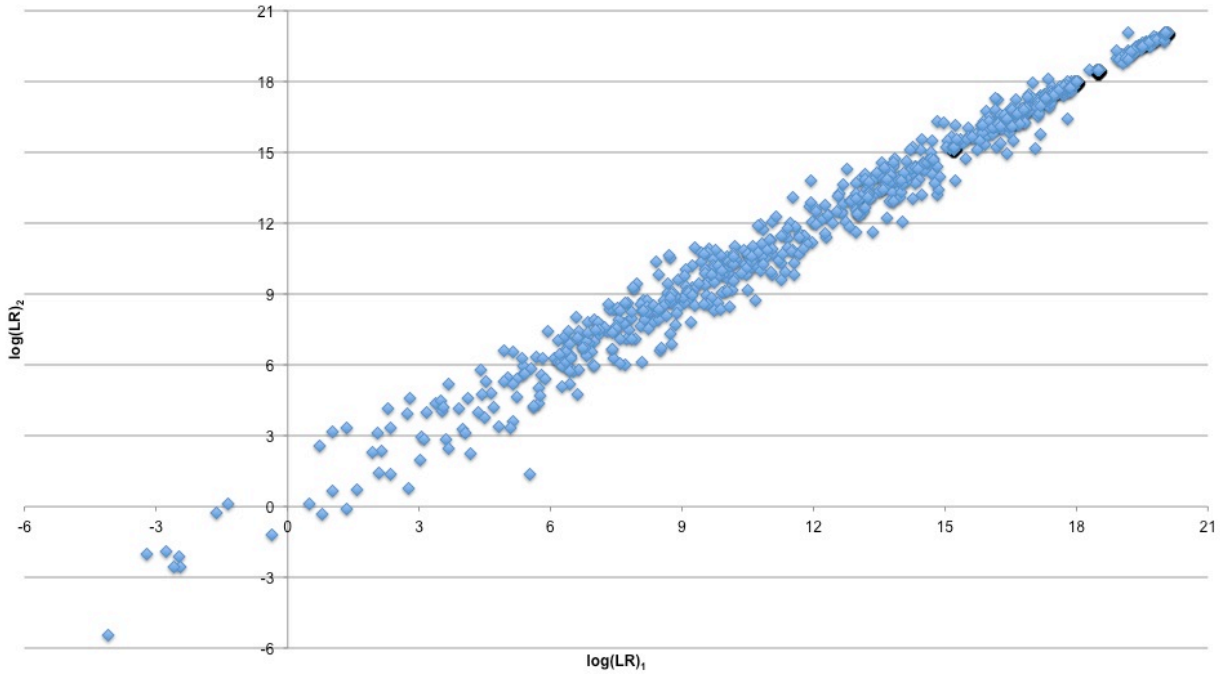


Figure 3: Reproducibility. The scatterplots show log(LR) genotype match values for duplicate computer runs on the same evidence for (a) 2 and (b) 3 contributor mixtures. Each point depicts the two match values on the first (x) and second (y) run. Results from the two sequencers were combined.

(a) 2 contributor mixtures



(b) 3 contributor mixtures