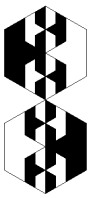
Scientific Validation of Mixture Interpretation Methods

Mark W. Perlin
Cybergenetics, Pittsburgh, PA

December 5, 2006

In the Proceedings of Promega's
Seventeenth International Symposium on Human Identification



Contact information:

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

Abstract

For courtroom admissibility, Federal Rule of Evidence (FRE) 702 mandates the reliability of (a) data, (b) method, and (c) application of method to data. The "reliability" of each component is determined (according to jurisdiction) by the Frye (1923) or Daubert (1993) standard. Whereas Frye entails only general acceptance, Daubert also provides for a testable approach, whose error rate has been determined, and has been communicated by peer review dissemination. These Daubert criteria are typically met by conducting scientific validation studies that establish the reliability of an approach by testing it and determining an error rate.

Forensic STR data have undergone extremely rigorous scientific validation in this country, with validation studies of laboratory processes routinely introduced as courtroom evidence in order to establish admissibility. Similarly, the DNA science of interpreting and matching single source profiles is solidly grounded in the rigor of population genetics. Since there is only one correct designation of a pristine single source profile, concordance studies can compare (the theoretically identical) results of two different examiners.

However, interpretation of mixed DNA samples need not produce unambiguous results. Different laboratories follow different mixture interpretation guidelines. Moreover, different examiners within the same laboratory who are following the same guidelines often infer different STR profiles. Therefore, there is no concordance in current forensic practice on what constitutes a "correct" mixture solution. Thus, it is not possible to conduct a mixture interpretation concordance study in order to validate a mixture interpretation method. But it is essential to have some way of testing the reliability of mixture interpretation methods so that inferred profiles from DNA mixtures can be scientifically validated and admitted as legal evidence.

We have developed a general approach to scientifically validating mixed DNA profiles. Instead of conducting a concordance comparison (which is not possible), our approach determines the amount of information present in the DNA match between an inferred mixture profile and a reference profile. By examining these numerical measures of match information, one can assess the reliability of a mixture interpretation method.

We have tested our validation approach on mixed DNA samples from multiple crime laboratories, each of whom uses their own interpretation methods. We found that different interpretation methods produce different inferred profiles with varying match specificity. However, regardless of match information, once a lab's interpretation method has been scientifically validated, its inferred profiles (and DNA matches that include those profiles) become admissible as reliable evidence in court.

In this paper, we present our novel approach to scientifically validating a lab's guidelines for interpreting DNA mixtures. We describe different mixture interpretation guidelines used in forensic practice, and show how these guidelines can all be validated as reliable methods. And we illustrate how our approach can be used for presenting DNA results in court. By scientifically validating the mixture interpretation method that it uses, a crime lab can go beyond just the admissibility of its validated DNA laboratory data, and also ensure the admissibility of its validated interpretation methods, inferred profiles and DNA matches.

Table of Contents

DNA Mixture Interpretation Admissibility

For scientific evidence to be admissible in court of law, it must be reliable. The Federal Rules of Evidence (FRE) Rule 702 requires that (i) the underlying data, (ii) the method of interpreting the data, and (iii) the application of this method to the data must all be reliable.

The older Frye 1923 ruling (1) defined the reliability of expert evidence as general acceptance by the scientific community. This standard can inadvertently institutionalize junk science, or block the introduction of better science. Therefore the more recent (and widely embraced) Daubert 1993 ruling (2) added to this general acceptance criterion four additional tests to help a judge assess the underlying scientific merit of the proffered evidence. These Daubert prongs are:

(1) Testable. Is the method inherently testable, and has it been tested?
(2) Error rate. Is it possible to determine error rate of the method, and has this error rate been determined?
(3) Peer review. Has the method been disseminated to the relevant scientific community in ways that foster critical review?
(4) Standards. Have standards been established for the use of the method?

While no one test is required, Daubert provides useful criteria for ascertaining scientific reliability.

The science of DNA identification coevolved with the legal Daubert standard (3). The forensic emphasis largely centered on the reliability of DNA laboratory data and the underlying population statistics. However, little attention was paid to the method of interpreting these data. With clean single source reference profiles, reliable data can produce only one correct answer, and so concordance between interpretations is sufficient for demonstrating reliability. Such concordance studies form the basis for validating the interpretation of reference STR profiles (4, 5).

However, the situation is not so clear when interpreting mixed DNA profiles. Different laboratories use different methods of mixture interpretation. Moreover, different people following the same mixture interpretation protocol on the same data often derive different STR profiles. No concordance study is therefore possible since discordant, but valid, inferred mixture profiles cannot be meaningfully compared. Hence there is a need for a general validation approach, which can establish the reliability of STR mixture interpretation in accordance with FRE 702 and the Frye and Daubert requirements.

It is reasonable to question whether DNA mixture evidence is currently admissible in American courts under FRE 702. Certainly the underlying laboratory data can be demonstrated as reliable using established STR validation procedures (6, 7). However, DNA interpretation experts have not validated the reliability of their mixture interpretation methods, nor the reliability of how these methods are applied to their STR data.

DNA mixture evidence currently fails the general acceptance test of both Frye and Daubert, since there are no generally accepted methods for interpreting mixed stains. Additionally, mixture evidence also fails all four scientific Daubert criteria:

(1) Testable. Inferred mixture profiles have not been tested for reliability, since the usual concordance comparisons cannot work.
(2) Error rate. DNA laboratories generally do not determine nor publish their mixture interpretation error rates.
(3) Peer review. Laboratories tend to not share their mixture interpretation guidelines, and many consider their interpretation methods to be confidential.
(4) Standards. Mixture interpretation standards do not exist, since each group uses its own interpretation protocols. Moreover, imposing one group's standards (8) on other practitioners without a rigorous scientific theory would be harmful to both science and the law.

The legal weakness of unvalidated scientific methods has not been lost on the defense bar. Recent articles in the legal defense literature provide recipes for decimating unvalidated methods. For example, in their Champion article "Evaluating and Challenging Forensic Identification Evidence," Tobin and Thompson describe the successful admissibility challenge to the unvalidated Comparative Bullet Lead and Analysis (CBLA) method, and extend it into a general strategy, specifically targeting the potential weaknesses of DNA evidence (9). The authors use four phases of forensic comparison to identify vulnerable targets for legal attack, when considering incompletely validated interpretation methods. As applied to DNA evidence, these phases are:

(1) infer a DNA mixture profile
(2) match the profile with a suspect profile
(3) assess the relative frequency of the profile
(4) draw conclusions

For DNA analysis, the matching methods of Phase 2 are largely agreed upon, the population statistics of Phase 3 has been adequately addressed by the courts, and the conclusions of Phase 4 are largely up to the finders of fact (i.e., the judge or the jury). It is Phase 1, the inferred mixture profile, that is currently without validation support. In order to introduce DNA mixture profiles as evidence without fear of a successful defense admissibility challenge, the DNA expert must provide a scientific validation of their mixture interpretation method to establish its reliability.

In this paper, I present a scientific approach to validating DNA mixture interpretation methods. This approach does not rely on concordance studies, which are not scientifically meaningful in this context. Rather, one can associate with each inferred mixture profile a "match information" value which indicates how strongly the inferred profile matches the true profile. (The true profile can be known in advance in a scientific study, or it can be determined from a legal outcome such as a confirmed match with a guilty verdict.) I show how to use this match information number to carefully examine the accuracy, precision and reproducibility of inferred mixture profiles. By conducting these mixture profile match information calculations on a standard spreadsheet, a DNA laboratory can rapidly and effectively assess the reliability of its mixture interpretation methods. This quantitative assessment is sufficient to scientifically

validate the laboratory's mixture interpretation method (and the application of this method to its STR data), and thereby overcome an admissibility challenge.

Different Mixture Interpretation Methods

I present here an extended example of alternative interpretation methods and resulting profiles on the same data. The example uses a 50:50 mixture data sample that was presented together with a reference sample in a mock sexual assault study, as previously described (10, 11). The mixture sample data (Figure 1A) shows the peaks for this 50:50 mixture. In a focused schematic of the D5S818 locus data (Figure 1B), one can see how the [12, 13] reference heterozygote combined in equal amounts with the [12, 12] unknown genotype to produce a 3:1 peak height ratio. The examiner's task is to accurately infer the [12, 12] unknown genotype at the D5S818 locus, given the reference and mixture data at all loci.

This section compares the results of five different mixture interpretation reviews. These reviews were conducted on STR data derived from the same DNA samples, but using different interpretation methods and reviewers. The conservative lab protocols were followed by two independent reviewers (government scientists A and B), the aggressive protocols were used by two different independent people (private lab scientists A and B), and the objective reviews were conducted by computer using the TrueAllele® Casework System (10). (The participating scientist teams are listed in the Acknowledgements section below.)

The "conservative" government review method for mixture interpretation is designed to avoid overcalling the DNA profile results. In the schematic example of data having an uncertain interpretation (Figure 2), one of the alleles must have designation "b". However, it is uncertain what the other allele is: it could be designated as "a", as "b", or even as something else. Therefore, this conservative protocol calls for reporting out 0, 1 or 2 alleles of which the examiner is entirely certain. The inferred profile here has allele 1 set to "b", and allele 2 able to assume any value. This is indicated by the [b, *] allele pair designation, where * denotes a wild card symbol that can match any allele.

The "aggressive" private lab review method for mixture interpretation strives for more profile specificity by trying to rule out unlikely allele pair combinations. Referring to the uncertain data example (Figure 3), there are two likely allele calls: the heterozygote [a, b] and the homozygote [b, b]. Any other allele pairs would be highly unlikely. This interpretation method reports a list of the likely allele calls.

The "objective" TrueAllele computer review method for mixture interpretation is designed to preserve match information (12, 13). It does this by reporting out a set of allele pairs, each having an associated probability (10). When the data are unambiguous, the method reports out just one allele pair with a probability of 1. With uncertain data, multiple allele pairs can be reported, having probability values which add up to 1. Note that this probability representation does not rank the allele pairs; rather, a contributor profile inferred from the mixture data is a probability distribution described by the allele pairs and their probabilities.

The use of allele probability distributions is common in genetic science . For example (Figure 4A), following Mendelian inheritance in the example genetic pedigree, the parent genotypes of [a, b] and [c, c] would produce a child with a 50% chance of having genotype [a, c] and an equal chance of having genotype [b, c]. Over the last few decades, this statistical representation of genetic uncertainty has been used successfully with computers to find genes, develop diagnostic tests, and understand the underlying molecular mechanisms that can lead to curing disease (14-16).

The data example (Figure 4B) shows how TrueAllele computer interpretation of uncertain data can produce genotypes with probability. If allele calls [a, b] and [b, b] were both likely solutions, then they would both be included in the probability distribution, each having a probability that depended on the data. In this example, there is a hypothetical probability assignment of 0.5 to each of the two possibilities.

Mixture Interpretation and Match Information

The first case review to consider follows the "conservative" interpretation method, as conducted by government Reviewer A. For clarity, let us consider all four phases of forensic DNA comparison.

Phase 1: Infer DNA profile. The reviewer inferred that any genotype was possible, yielding the designation [*, *], where * denotes a wild card symbol that can match any allele. As shown (Figure 5A) by the entirely filled in blue Punnett square (17), all allele pairs are possible.

Phase 2: Match with known. A match comparison can be made against the known genotype [12, 12]. As shown (Figure 5B), this genotype is represented by the small red square at allele coordinate [12, 12]. Looking at this figure, as well as the next (Figure 6A), the blue inferred set of all possible genotypes overlaps the red [12, 12] known genotype, and so there is a match.

Phase 3: Relative frequency. The inferred profile [*, *] includes all possible genotypes. Therefore, the relative frequency of the inferred profile is 100% of the population (Figure 6B), since everyone has some genotype.

Phase 4: Draw conclusions. Since the inferred profile matches everyone, there is no information in this match. Quantitatively, one can express this fact by stating that the match information is equal to zero (since 100% of the population has a frequency of 1, and $log(1) = 0$).

A scientist can calculate match information directly from the population frequencies. The match likelihood ratio (which fully describes the match information) can be roughly defined as the probability of observing a specific match divided by the probability of observing a random

match. For most situations, this is exactly the random match probability – the reciprocal of the relative frequency. The Match Information (MI) is obtained by calculating this likelihood ratio when matching the inferred profile against the true, known profile. To properly measure and add up information, it is useful to work with the logarithm of the likelihood ratio. To summarize then:

$$\begin{aligned}\text{Match Information} &= \log\left(\text{match likelihood ratio}\right)\\ &= \log\left(\frac{\text{Probability (specific match)}}{\text{Probability (random match)}}\right)\end{aligned}$$

In the usual cases seen in current forensic practice, there is a specific match (so that the probability of the observed specific match is certainty, i.e., equal to 1) and the random match probability is the population frequency, so that the Match Information becomes:

$$\begin{aligned}\text{Match Information} &= \log\left(\frac{1}{\text{population frequency}}\right)\\ &= -\log\left(\text{population frequency}\right)\end{aligned}$$

since the logarithm of a reciprocal equals the negative of a logarithm. To summarize, for mixture interpretation validation, most DNA laboratories would calculate match information at a locus from population frequency data using the formula:

$$\text{Match Information} = -\log\left(\text{population frequency}\right)$$

(For clarity in this paper, base 10 logarithms are used throughout. That is, if $y = 10^x$, then $x = \log_{10}(y)$. For example, when y is 1,000,000, i.e., $y = 10^6$, then $x = \log_{10}(10^6) = 6$. For the reader unpracticed in the use of logarithms, it can be helpful to take the log base 10 by counting the number of zeros to the decimal point.)

Different Mixture Interpretation Results

The second government reviewer B, looking "conservatively" at the same STR case data, inferred a different mixture profile at locus D5S818. Reviewer B inferred [12, *], specifying the first allele as 12, and leaving the second allele undesignated and free to match any allele. This inferred genotype is more specific than Reviewer A's [*, *] interpretation (Figure 7), and it only matches those genotypes having at least one allele designated as 12. The increased specificity reduces the relative population frequency of the inferred profile from 100% down to 50%. The result is an increase in match information (computed as the negative logarithm of the population frequency 0.50) from zero up to 0.3.

The first private lab reviewer A, using a different "aggressive" mixture interpretation method, arrived at the same inferred [12, *] genotype as the "conservative" government reviewer B (Figure 7). This replicated result has the same 50% relative frequency of the inferred profile and the same match information of 0.3.

However, the second private lab reviewer B, "aggressively" inferred a more specific result, comprising the two allele pairs [12, 12] and [12, 13]. This increased specificity (Figure 8) fills in even fewer of the Punnett square entries. The relative population frequency of Reviewer B's inferred profile is concomitantly reduced to 20%, with an increased match information of 0.7.

The TrueAllele "objective" computer review inferred only one answer, with probability one: [12, 12]. This is the most specific inferred profile, which exactly matches the known answer (Figure 9). The increased specificity reduces the relative frequency of the inferred profile down to 10%, with an associated increase of match information to 1 (since $-log_{10}(0.1) = +log_{10}(10) = 1$).

To obtain the match information for an entire inferred profile, simply add up the computed match information from every locus in the profile. (This works because of locus independence and the use of logarithms.) Combining the population frequencies in this additive way, one can write:

$$MI_{profile} = \sum_{locus} MI_{locus}$$

which means that the match information of the entire profile is equal to the sum of the match information at each of the loci.

An example of how the individual locus match information values add up to the total profile match information can be seen for this case in the publicly accessible study section of the www.trueallele.net web site (Figure 10).

Scientific Validation of Mixture Interpretation

With five independent reviews for the D5S818 locus in this 50:50 mixture case, the reviewers inferred four different genotype solutions. These differences demonstrate that a concordance study between incommensurable profiles would not be possible. However, the match information statistic introduced above fully captures a profile's information content in ways that can be used for scientific validation. This section shows how accuracy, precision and (most importantly) reproducibility studies can be conducted based on this match information measurement.

I chose accuracy, precision and reproducibility measures because they are familiar to forensic scientists who perform laboratory data validations in accordance with DAB guidelines (6). The reproducibility measure is the most important one for scientific and legal reliability, and its characterization is a key result of this paper. The accuracy measure is also useful in validation studies, since a wholly inaccurate mixture interpretation method would not be considered to be reliable. The precision measure used here was selected from several options, and helps motivate a statistic based on an average of differences.

Importantly, these (and many other) validation measures are based on the match information (MI) value of an inferred profile. The ultimate legal use of an inferred DNA profile is the match information that it produces, so the MI statistic is quite appropriate for assessing the efficacy and reproducibility of forensic STR interpretation methods.

Inferred DNA profiles can be represented at a locus in different ways, including as lists of alleles, as lists of allele pair genotypes, and as probability distributions. These different representations, as well as the different answers within one representation, are often incommensurable – there is no logical way to compare the profiles. This incommensurability makes it impossible to conduct a concordance study based on direct profile comparison, whether between or within laboratories. However, regardless of its representation, the full profile has a match information relative to the true profile. This MI is just an ordinary real number (and not a large, unwieldy profile representation), which can be compared against other MI numbers or used in mathematical formulas. Therefore, these profile MI numbers can be used in place of the actual profiles in a statistical validation study.

The data set used in this section is taken from the two contributor cases previously described (10). Specifically, we analyzed "conservative" interpretation results from the 1 ng DNA mixture samples having 30%, 50%, 70% and 90% unknown contributors for two pairs of individuals. There were two reviews (A and B) performed for each of these eight cases. The match information (MI) statistic for each profile was computed in the TrueAllele Casework system using logarithms of population frequency, as described above.

Accuracy asks the question of whether an answer is correct. Webster's dictionary defines accuracy as the "degree of conformity of a measure to a standard or a true value." Figure 11 shows a bar plot of the MI statistic for each pair of reviewer (A and B) inferred profiles, for all eight cases. It is visually apparent that the MI values lie between 9 (match likelihood ratio exceeding a billion to one) and 15 (match likelihood ratio exceeding a million billion to one). Therefore, on average, the inferred profiles are clearly identifying the correct individual.

This overall accuracy is well expressed through the average of the match information values for the N different case reviews:

$$\overline{MI} = \frac{1}{N} \sum_{i=1}^{N} MI_i$$

which says that the average case match information ($\overline{MI}$) equals the sum of the match information for each case (MI_i), divided by the total number of cases (N). The case MI values taken from the Excel spreadsheet that generated the figure are shown (Table 1). These cases have an average match information of 11.91, or roughly a likelihood ratio of one trillion to one (since $log_{10}(1{,}000{,}000{,}000) = 12$).

Precision asks the question of how precisely an answer is described. Webster's defines precision as "the degree of refinement with which an operation is performed or a measurement stated." The match information statistic describes the number of significant digits that an inferred profile has attained in reaching the full discriminating power of a DNA profile. So in asking, "How close is this profile to the true profile?", the difference between an inferred profile's MI and the true profile's MI provides a precise measurement of how close it is in each

case. Figure 12 shows a bar chart, similar to that in Figure 11, but now adding a bar within each case group that represents the MI of the true profile. The underlying data from this Excel spreadsheet are shown in Table 2, along with the MI differences between the inferred and true profiles for every case.

A reasonable measure of precision for validating a mixture interpretation method is the average MI difference between inferred and true profiles:

$$\overline{\Delta MI} = \frac{1}{N}\sum_{i=1}^{N}\left(MI_i - MI_i^{true}\right)$$

which says that the average case match information difference ($\overline{\Delta MI}$) equals the sum of the match information differences between inferred and true for each case ($MI_i - MI_i^{true}$), divided by the total number of cases (N). These differences are computed and shown (Table 2). For the "conservative" interpretation method in this validation example, the average MI difference is –5.41. Since $log_{10}(1{,}000{,}000) = 6$, the inferred profiles are, on average, within a million to one likelihood ratio of the true answer.

Reproducibility concerns the repeatability of an answer. Consider again the data plotted in Figure 11. Observe that the two reviewers, while not always generating exactly the same profile, are producing match information values that are very close in every case. This repeatability goes to the heart of the scientific validation of mixture interpretation methods, asking, "How repeatable is the solution?" The answer to this question squarely addresses the reliability of the DNA mixture interpretation process.

Scientists and statisticians often measure reproducibility through the variance of a process (18). It is often useful to express the variance through its square root, the standard deviation, which preserves the original units. The standard deviation for mixture interpretation is the average square deviation between the MI of a case and the average MI for that case.

The standard statistical formula for *variance* σ_{MI}^2 within case groups is:

$$\sigma_{MI}^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(MI_{i,j} - \overline{MI_i}\right)^2$$

which says that the average case match information squared deviation (σ_{MI}^2) equals the sum of the match information squared deviations between inferred and true for each case $\left(MI_{i,j} - \overline{MI_i}\right)^2$, divided by the total number of cases (N). Note that the variance here is measuring the variation between each individual review of a case ($MI_{i,j}$) relative to group average for all reviews of that particular case ($\overline{MI}_i$).

(Remarks on notation: The subscript i refers to the i^{th} case, while the subscript pair i,j refers to the j^{th} review of the i^{th} case. The double summation $\sum_{i=1}^{m}\sum_{j=1}^{n_i}$ indicates that the N total reviews can be grouped into multiple reviews of m different cases. The i^{th} case has n_i separate reviews.)

The *standard deviation* σ_{MI} is defined in statistics as the square root of the variance:

$$\sigma_{MI} = \sqrt{\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(MI_{i,j} - \overline{MI_i}\right)^2}$$

The MI standard deviation σ_{MI} is helpful in understanding reproducibility because it is expressed in the same units as match information. (The variance σ_{MI}^2, on the other hand, is scaled in the less intuitive "squared MI" units.)

An Excel spreadsheet where each of the eight paired case reviews is shown in a separate row (Table 3). The columns A and B provide the individual MI scores from Reviewers A and B for each case. The next column, labeled "m," gives the average (or mean) value of the MI values recorded within each case. The last two columns in the spreadsheet compute the square deviation $\left(MI_{i,j} - \overline{MI_i}\right)^2$ between each reviewer's case match information $MI_{i,j}$ and the average match information $\overline{MI}_i$ for that case row.

At the bottom of the table, an entry adds up all the squared deviations to form the sum of squares SS. The estimated variance σ_{MI}^2, or average sum of squares, is calculated as SS/N. Finally, the estimated standard deviation σ_{MI} is computed as the square root of the variance. The resulting MI standard deviation equals 0.14, which is a small amount of MI variation relative to one logarithm unit of human identification. That is, the variation between multiple reviews of the same cases is very small. Hence, this "conservative" mixture interpretation method is quantitatively determined to be highly reproducible, with a known σ_{MI} reliability measured as 0.14 MI units.

Comparison of Validation Results

The validation results for three different mixture interpretation methods are shown in Table 4. The first row, validating the "conservative" method, shows an accuracy of 11.91 match information units, a precision of -5.41 MI units, and a reproducibility of 0.140 MI units. These validation numbers do not in any way constitute a value judgment on the reliability of the interpretation method. However, by having these MI statistics available for presentation in court (particularly the all-important reproducibility determination), a laboratory can demonstrate that it has scientifically measured the reliability of its mixture interpretation method.

The match information comparison bar chart for the "aggressive" mixture interpretation method is shown in Figure 13. Each case cluster shows the MI statistic for reviewer A, reviewer B and the true profile. Visually, as compared with the "conservative" method, there appears to be more match information on average, but with greater variation between analysts' deduced profiles. This observation is borne out in the second row of Table 4, where there is an improvement of 4 MI units for accuracy and precision, but a six-fold increase in MI standard

deviation to 0.833, indicating a less reproducible process. Again, these MI validation statistics are neither good nor bad, but merely provide the scientific assessment of the method's reliability required by the court.

In comparing the "conservative" and "aggressive" mixture interpretation methods, one might ask if there is a fundamental trade-off between accuracy and precision on the one hand, and reproducibility on the other. That is, are more accurate interpretation methods inherently less reproducible? And does greater reproducibility imply lower precision? That might well be the case for the human interpretation of DNA mixture data. However, as demonstrated next, there is no such inherent limitation on statistical computer review.

The "objective" computer mixture interpretation method's match information comparison bar chart is shown in Figure 14. Each case cluster shows MI statistics for four different computer interpretation run profiles, and the true STR profile. (It is far easier to have a computer conduct multiple independent reviews of the same case than it is to ask people to perform this repetitive task.) There appears to be virtually complete match information, and very little variation between the interpretations. This observation is confirmed in the third row of Table 4, where one observes virtually no loss of match information relative to the true profile (accuracy and precision), and a high reproducibility indicated by a standard deviation of less than 0.1 MI units. These data indicate that automated computer mixture interpretation methods are inherently amenable to rapid and complete scientific validation.

Conclusion

The law requires the admissibility of scientific evidence in order for it to be presented in court. This admissibility is based on the reliability of the proffered evidence, including the data, the method and the application of the method to the data (FRE 702). The reliability of these components is demonstrated by conducting validation studies that address jurisdictional requirements (Frye or Daubert).

In forensic DNA practice, mixture interpretation methods have not been validated. This scientific gap is partially due to the absence of any rigorous method for validating mixture interpretation. For example, conventional concordance studies cannot work. However, mixture interpretation must be validated in order to introduce DNA evidence in court without significant risk of successful challenge.

This paper remedies the scientific gap by presenting a new statistical approach for rigorously validating mixture interpretation methods. The approach uses match information (a likelihood statistic based on population frequencies) to assess accuracy and precision. By measuring the standard deviation between different reviewers within the same case, the approach can quantitatively determine the reproducibility of a mixture interpretation method.

The paper showed how to implement the statistical validation approach in a standard Excel spreadsheet (no proprietary software is required). Laboratories already perform multiple

reviews of the same case, and use population statistics to quantitatively assess match information. It takes only a few hours to enter these match information results into a spreadsheet, and calculate the standard deviation to measure reproducibility. This calculation will validate the lab's mixture interpretation method, generating the scientific reliability results needed to refute admissibility challenges.

Acknowledgements

I would like to thank my scientific collaborators, including Drs. Robin Cotton, Barry Duceman and Jay Kadane. The DNA laboratories providing interpretation results for this study were the New York State Police Forensic Identification Center in Albany, NY and Orchid Cellmark in Germantown, MD. The End Point company provided systems support for Cybergenetics Internet-based TrueAllele Casework computer servers, databases and web interfaces used in performing this study.

The New York State Police case reviewers included John Brenner, Russell Gettig, Melissa Lee, Maria Mick, Shannon Morris, Urfan Mukhtar, Michael Portzer, Laura Post and Diana Seaburg. The Orchid Cellmark case reviewers included Christine Baer, Jason Befus, Julie Black, Paula Clifton, Kathryn Colombo, Lisa Grossweiler, Juliet Harris, Jeff Hickey, Jacki Higgins, Christopher Knickerbocker, Jason Kokoszka, Lewis Maddox, Jennifer Reynolds, Leslie Rosier, Ryan Satcher, Alissa Shofkom, Margaret Terrill, Melissa Thompson, Charlotte Word and Cindy Zimmerman. The Cybergenetics staff who processed data and information for this study included Bill Allan, Meredith Clarke, Matthew Legler, Donna Scheuble and Dr. Alex Sinelnikov.

References

1. Frye v. United States. In: Court of Appeals of District of Columbia; 1923.
2. Daubert v. Merrill Dow Pharmaceuticals, Inc. In: Supreme Court; 1993.
3. Faigman DL, Kaye DH, Saks MJ, Sanders J. Science in the Law: Forensic Science Issues: West Group, 2002
4. Kadash K, Kozlowski BE, Biega LA, Duceman BW. Validation study of the TrueAllele® automated data review system. Journal of Forensic Sciences 2004;49(4):1-8.
5. NDIS. Appendix B: Guidelines for submitting requests for approval of an expert system for review of offender samples: Federal Bureau of Investigation; 2005.
6. DNA Advisory Board. Quality assurance standards for forensic DNA testing laboratories and for convicted offender DNA databasing laboratories. Forensic Sci Commun (FBI) 2000;2(3).
7. Butler JM. Debunking some urban legends surrounding validation within the forensic DNA community. Profiles in DNA 2006;9(2):3-6.
8. Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, et al. DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. Forensic Science International 2006;160:90-101.
9. Tobin WA, Thompson WC. Evaluating and challenging forensic identification evidence. The Champion 2006;30(6):12-21.
10. Perlin MW. Simple reporting of complex DNA evidence: automated computer interpretation. In: Promega's Fourteenth International Symposium on Human Identification; 2003; Phoenix, AZ; 2003.
11. Perlin MW. Real-time DNA investigation. In: Promega's Sixteenth International Symposium on Human Identification; 2005; Dallas, TX; 2005.
12. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. Journal of Forensic Sciences 2001;46(6):1372-1377.
13. Perlin MW. Method for DNA mixture analysis. US Patent; 2004. Report No.: #6,807,490.
14. Lange K, Weeks DE, Boehnke M. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genetic Epidemiology 1988;5:471-472.
15. Andrews C, Devlin B, Perlin M, Roeder K. Binning clones by hybridization with complex probes: statistical refinement of an inner product mapping method. Genomics 1997;41(2):141-154.
16. Roeder K, Escobar M, Kadane JB, Balazs I. Measuring heterogeneity in forensic databases using hierarchical Bayes models. Biometrika 1998;85(2):269-287.
17. Butler JM. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. Second ed. New York: Academic Press, 2005
18. Bevington PR. Data Reduction and Error Analysis for the Physical Sciences. New York: McGraw-Hill, 1969

Figures

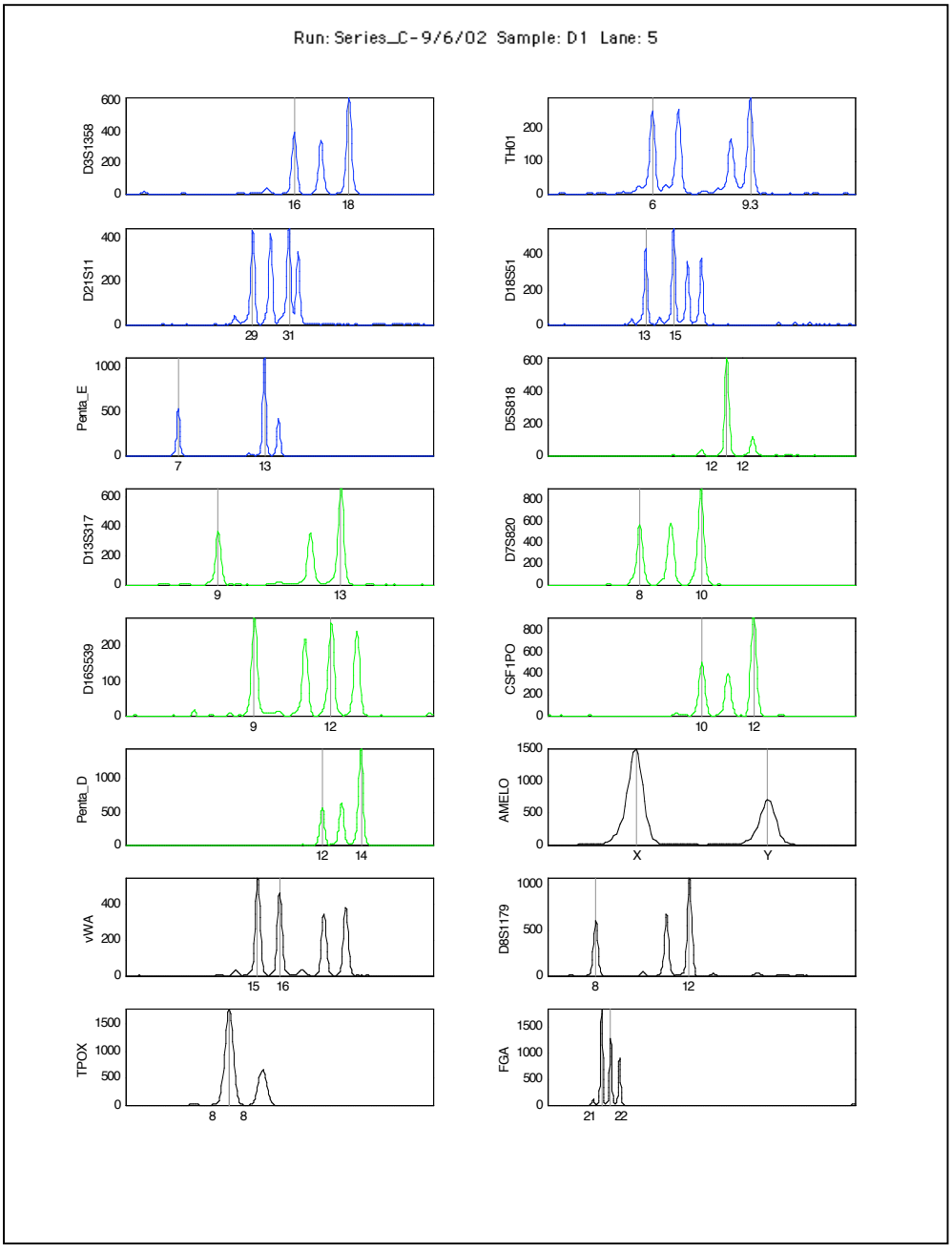


Figure 1A. A 50:50 mixture profile for PowerPlex 16 STR data on an ABI/310 DNA sequencer.

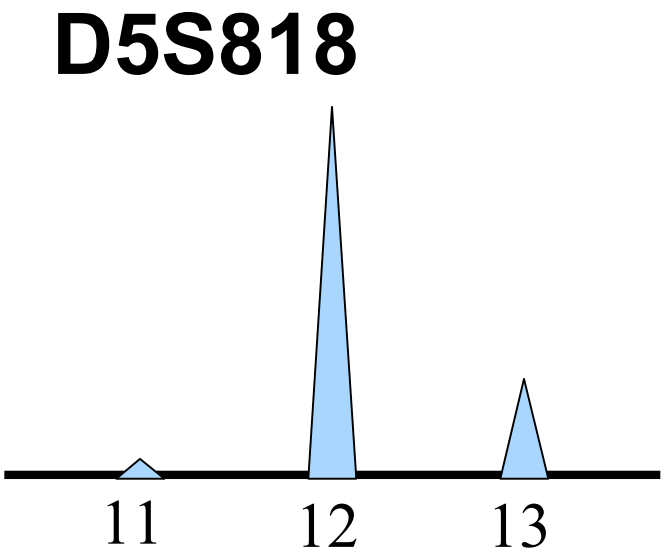


Figure 1B. A zoomed in schematic of the D5S818 locus data showing how a heterozygote genotype combined in equal amounts with a homozygote genotype produces a 3:1 peak height ratio.

Figure 2. The goal of "conservative" government review of uncertain mixture data is to avoid overcalling the results. The inferred locus profile can report out 0,1 or 2 alleles.

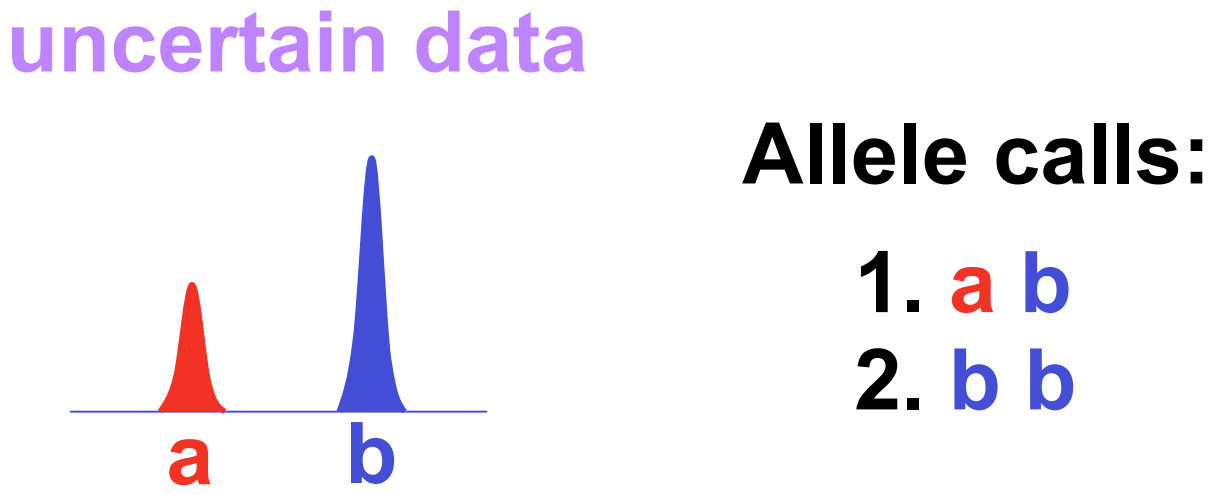
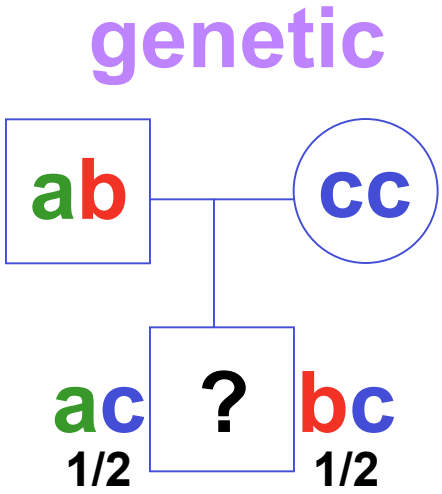


Figure 3. The goal of "aggressive" private lab review of uncertain mixture data is to try ruling out unlikely mixture combinations. The inferred locus profile reports a list of possible allele calls.

(A)



(B)

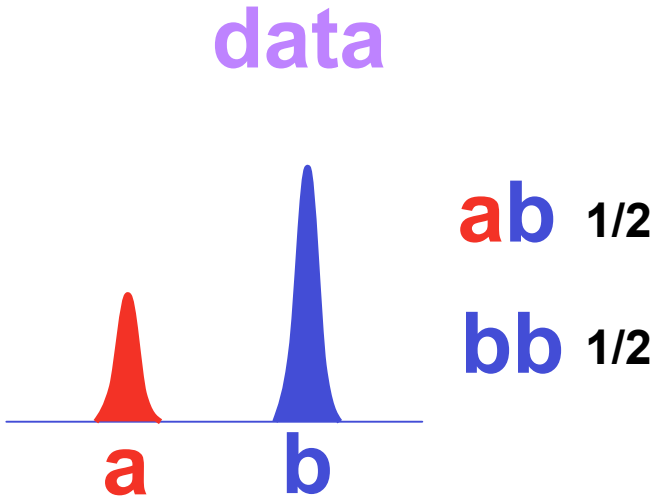


Figure 4. The goal of "objective" computer review is to preserve match information. **(A)** Genetic uncertainty produces STR profiles that have probabilities associated with each possible genotype. **(B)** Data uncertainty can similarly produce an STR profile genotype with a probability distribution of allele calls.

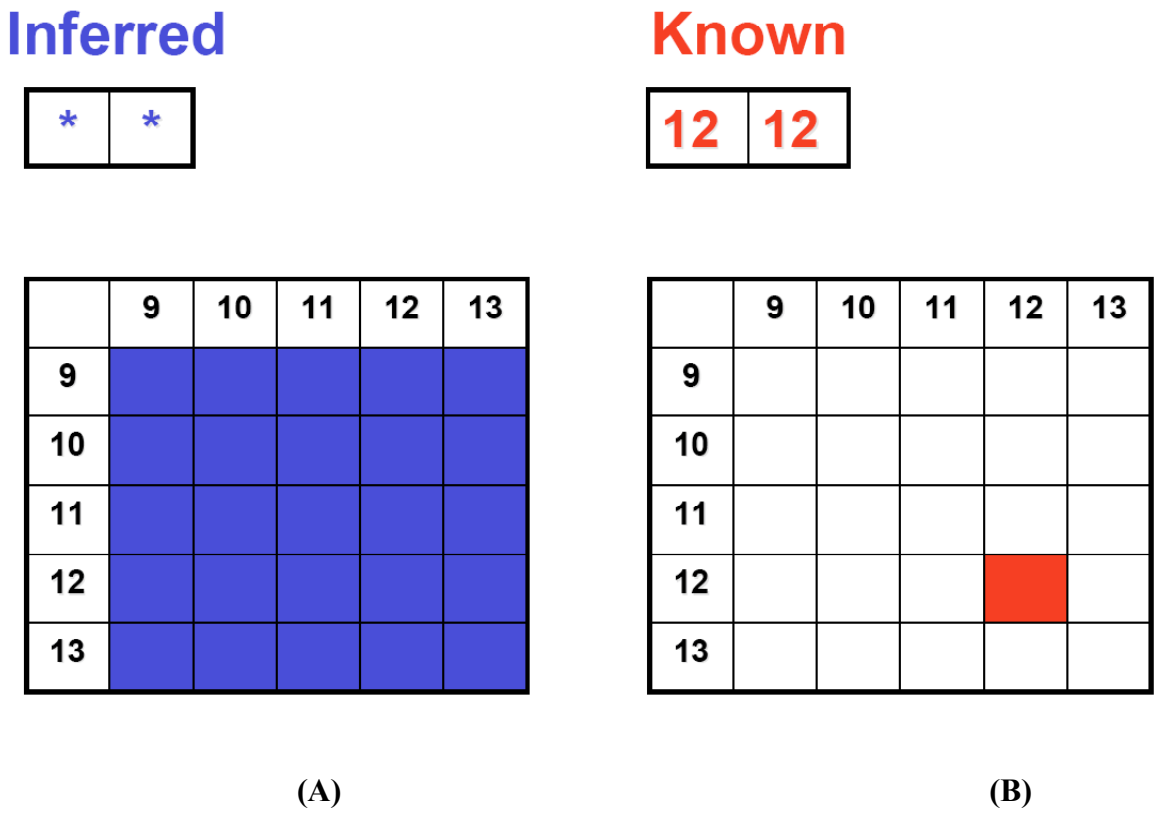


Figure 5. **(A)** Conservative human review in a government laboratory infers all possible allele pairs [*, *], which completely fills the square and represents all possible individuals. **(B)** The known genotype is the allele pair [12, 12].

(A) **(B)**

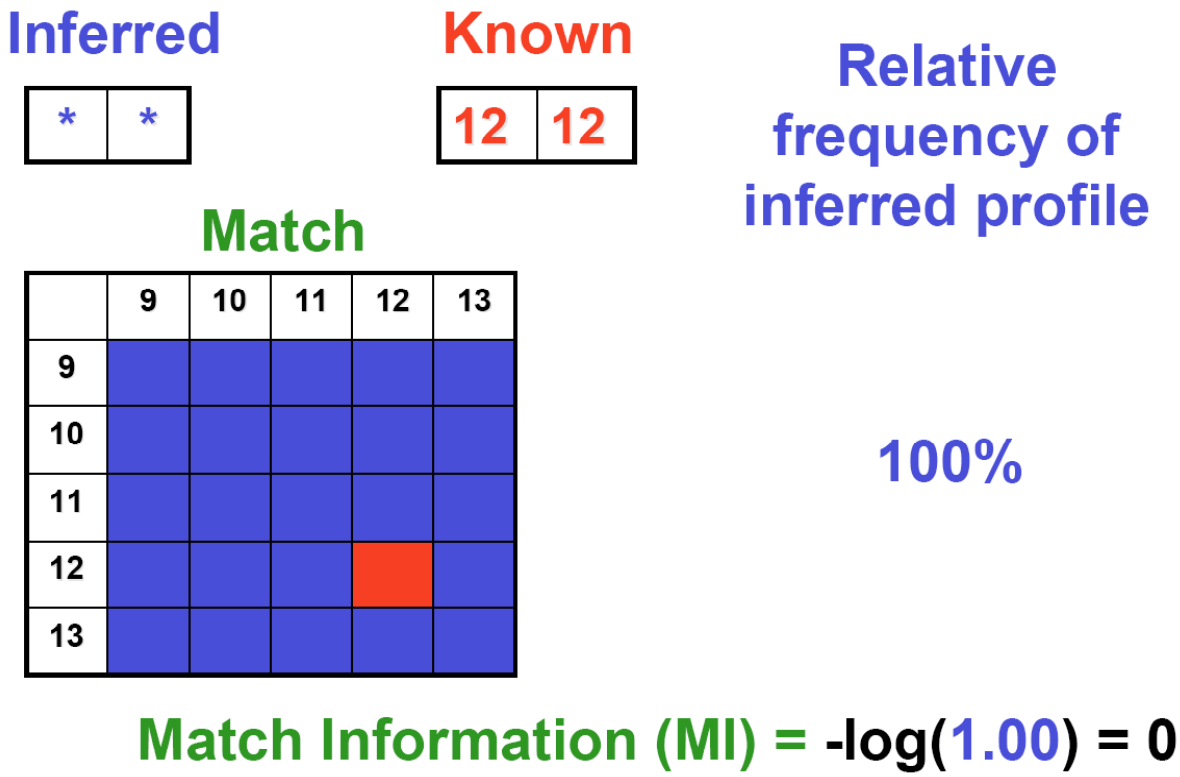


Figure 6. (A) The inferred profile [*, *] includes all possible genotypes, which matches the known genotype [12, 12]. **(B)** The relative frequency of the inferred profile is 100% of the population, which yields no useful match information. I.e., MI = 0.

\

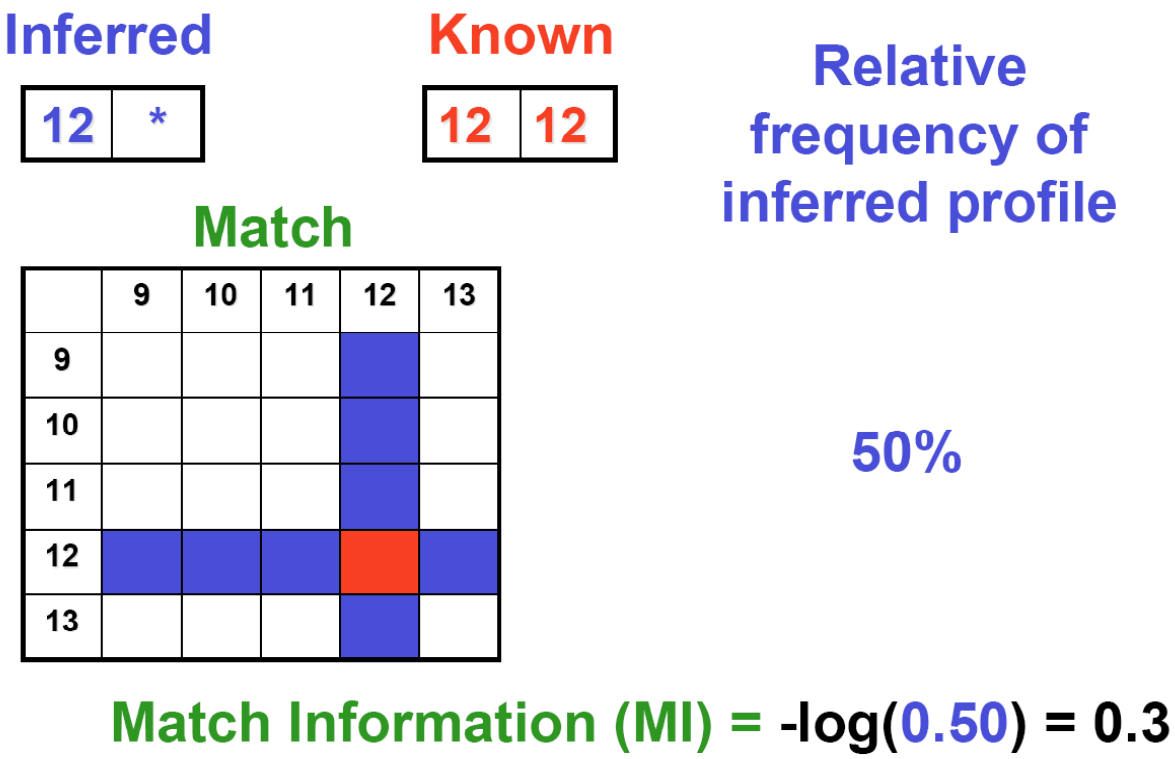


Figure 7. The second government reviewer B, looking "conservatively" at the same STR case data, inferred a mixture profile at locus D5S818 of [12, *]. So did the first private lab reviewer following a more "aggressive" mixture interpretation protocol. This genotype set matches 50% of the population, including the known profile [12, 12]. The information for matching 50% is 0.3.

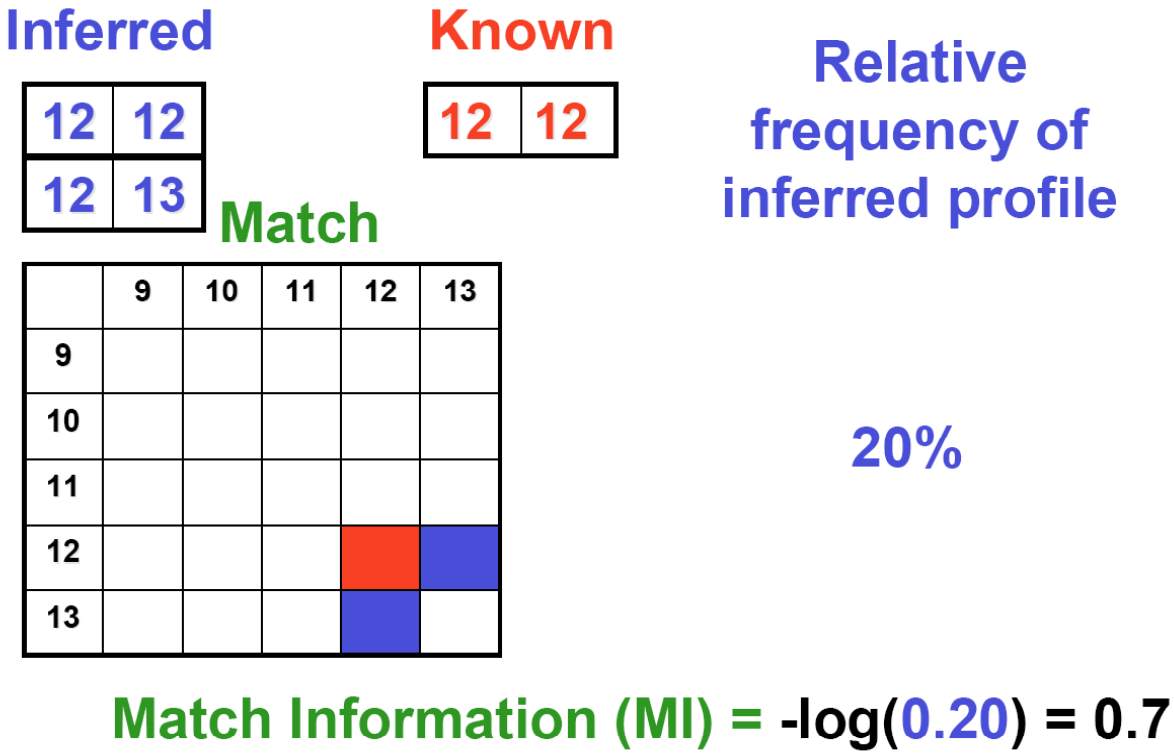


Figure 8. The second private lab reviewer B "aggressively" inferred the two allele pairs [12, 12] and [12, 13]. There are fewer Punnett square entries, hence a lower relative population frequency and increased match information.

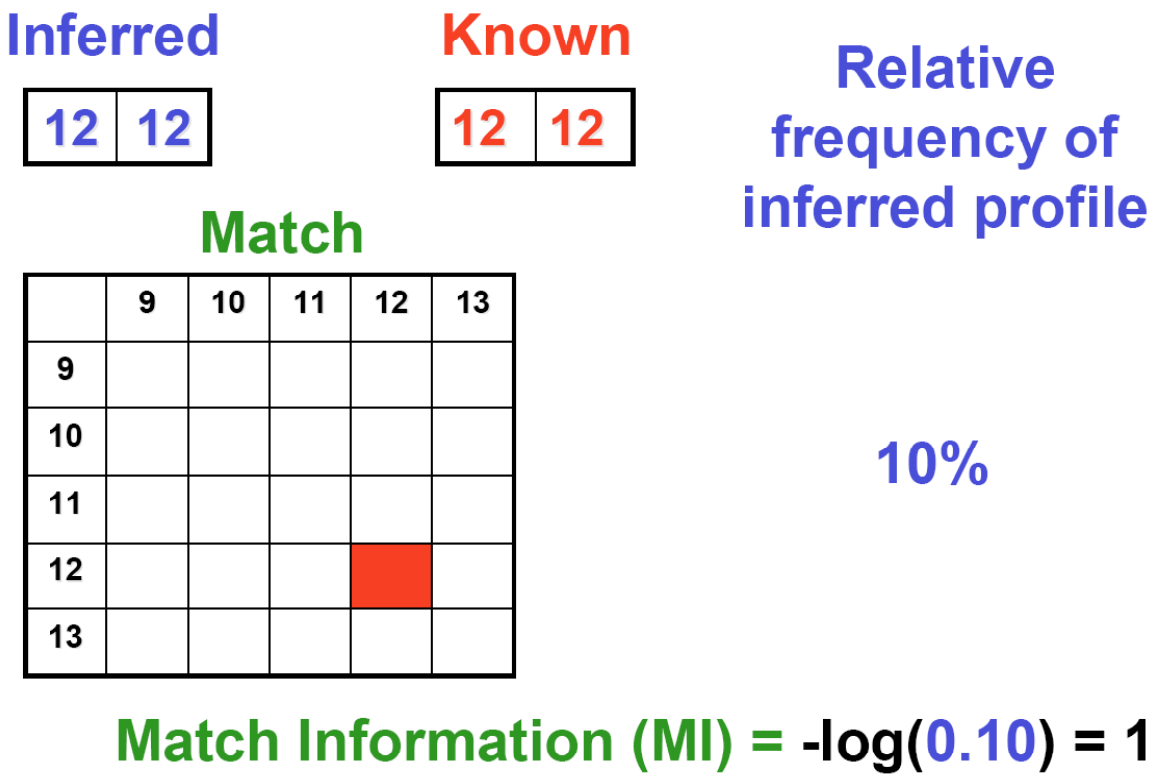


Figure 9. The TrueAllele "objective" computer review inferred only one answer, [12, 12], which exactly matches the known answer. The relative frequency of the correct inferred profile at 10% is the smallest possible, yielding maximal match information.

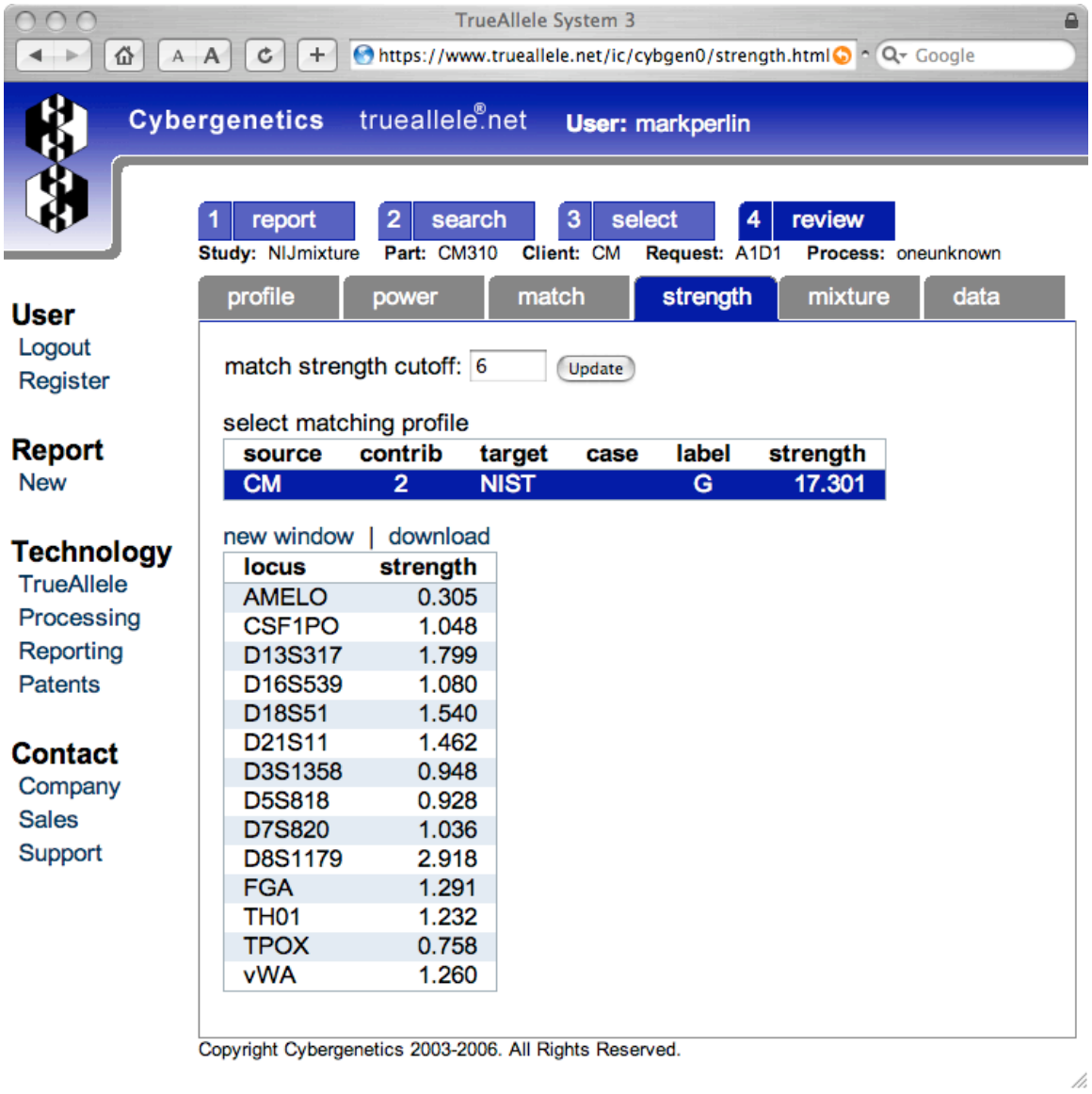


Figure 10. The "match strength" page of the www.trueallele.net web site shows how the individual locus match strengths add to the total match strength for the full profile. Since the inferred profile is being compared here against the true profile, the match strength values provide the match information (at each locus and for the full profile).

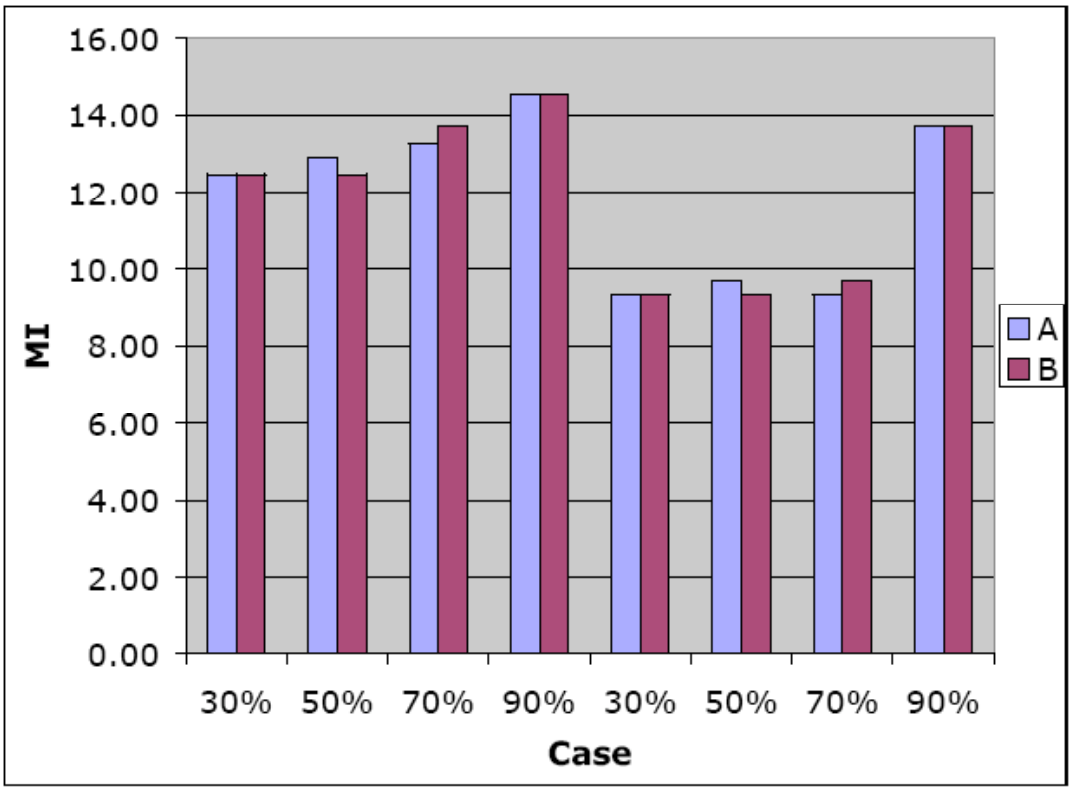


Figure 11. A bar plot of the MI statistic for each pair of "conservative" reviews (A and B) inferred profiles. The eight cases are for two different pairs of contributors, each case having different mixture weights. The y-axis shows the match information statistic in log units.

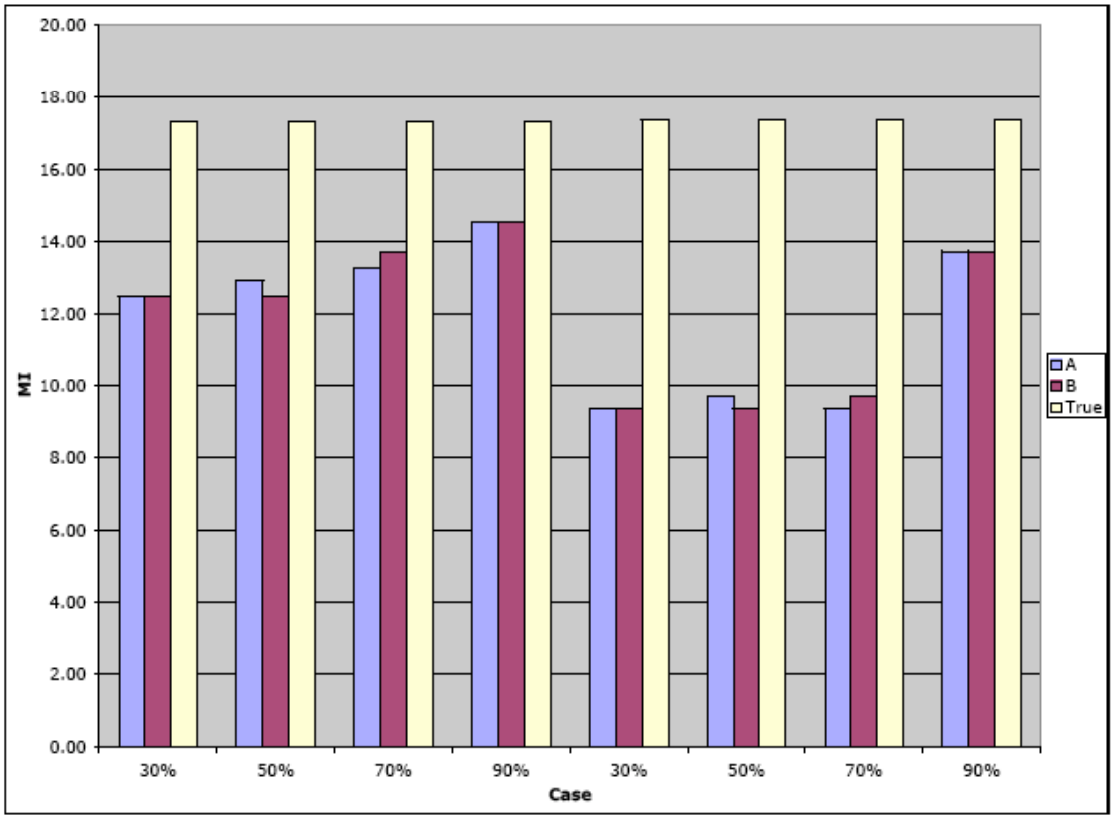


Figure 12. A bar plot of the MI statistic for the "conservative" reviewers (A and B) that now includes a bar showing the maximum attainable match information for the known profile.

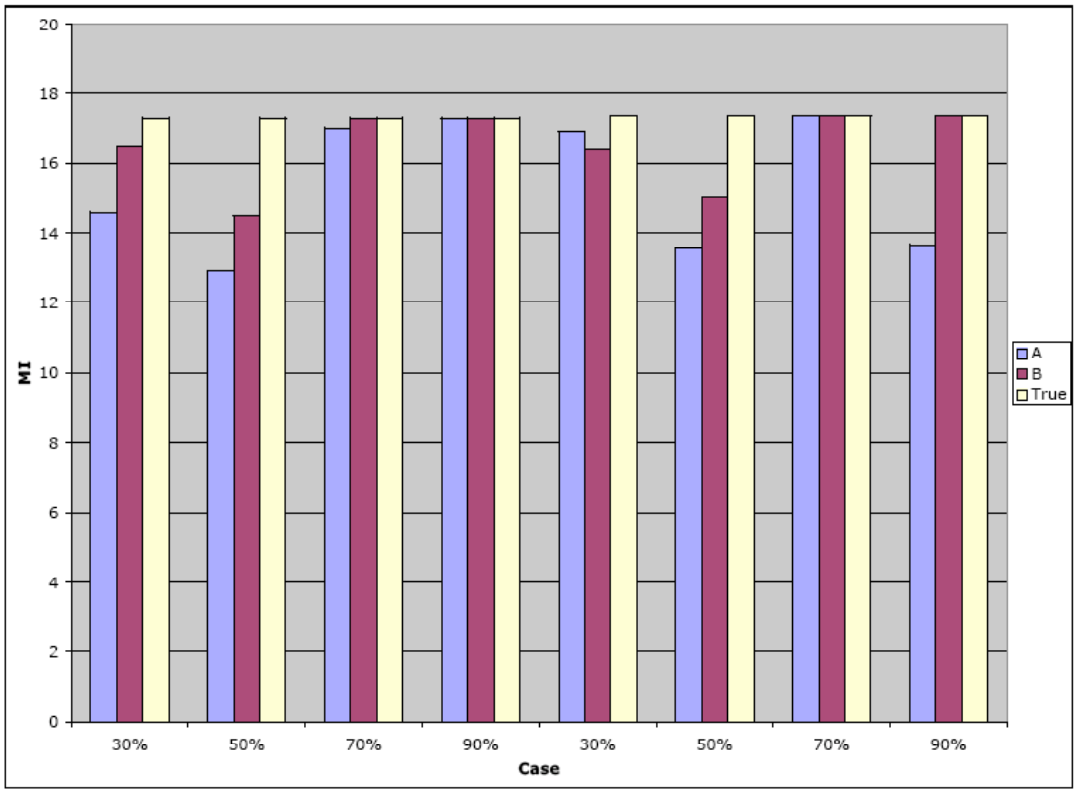


Figure 13. A bar plot of the MI statistic for the "aggressive" reviewers (A and B), including a bar showing the maximum attainable match information for the known profile.

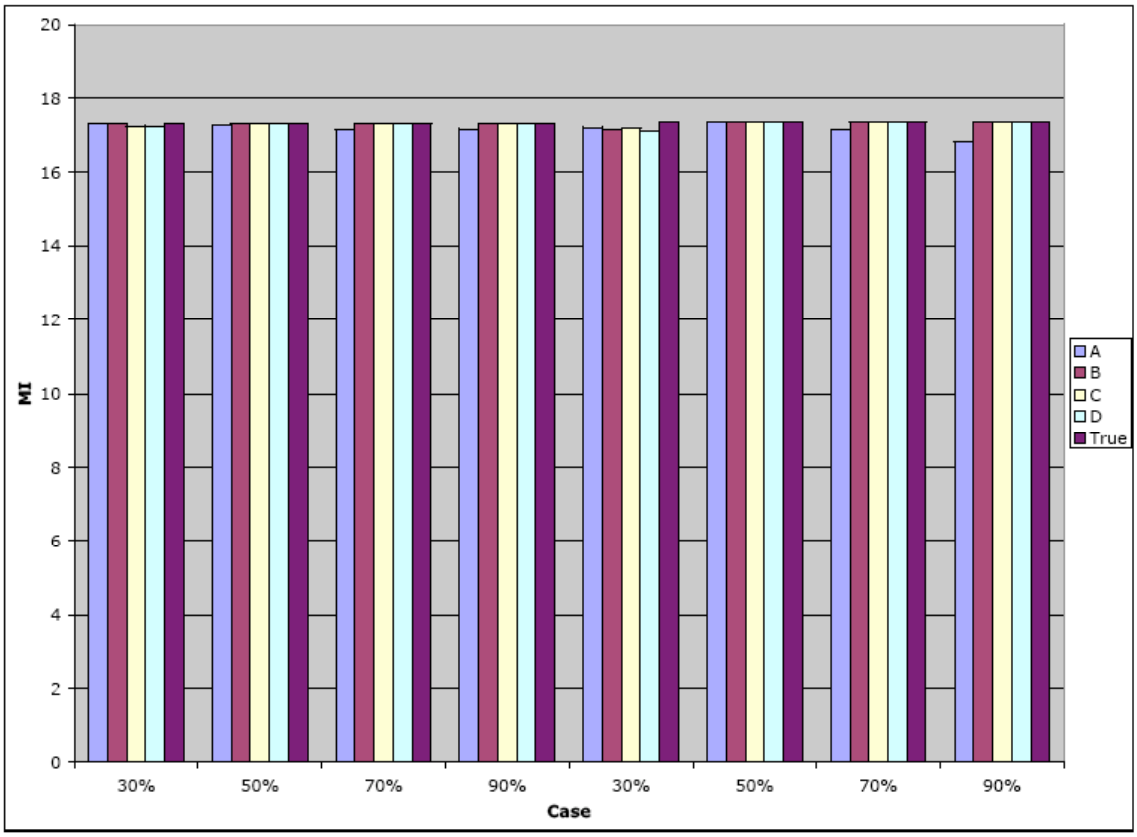


Figure 14. A bar plot of the MI statistic for the four "objective" reviews (A, B, C, D) by the TrueAllele Casework computer, including a bar showing the maximum attainable match information for the known profile.

Tables

case	A	B
30%	12.47	12.47
50%	12.90	12.47
70%	13.27	13.69
90%	14.54	14.54
30%	9.34	9.34
50%	9.71	9.34
70%	9.34	9.71
90%	13.72	13.72
	Average	**11.91**

Table 1. Accuracy: The match information values for each case for the two "conservative" reviewers A and B. The average of these 16 values is shown.

\

case	A	B	True	A-True	B-True
30%	12.47	12.47	17.30	-4.83	-4.83
50%	12.90	12.47	17.30	-4.40	-4.83
70%	13.27	13.69	17.30	-4.03	-3.61
90%	14.54	14.54	17.30	-2.76	-2.76
30%	9.34	9.34	17.35	-8.01	-8.01
50%	9.71	9.34	17.35	-7.64	-8.01
70%	9.34	9.71	17.35	-8.01	-7.64
90%	13.72	13.72	17.35	-3.63	-3.63
			Average difference		**-5.41**

Table 2. Precision: The match information values for the "conservative" reviewers A and B, showing the MI differences between the inferred and true profiles. The spreadsheet calculated the average of these 16 MI differences as –5.41.

case	A	B	m	(A-m)^2	(B-m)^2
30%	12.47	12.47	12.47	0.000	0.000
50%	12.90	12.47	12.68	0.045	0.045
70%	13.27	13.69	13.48	0.045	0.045
90%	14.54	14.54	14.54	0.000	0.000
30%	9.34	9.34	9.34	0.000	0.000
50%	9.71	9.34	9.53	0.034	0.034
70%	9.34	9.71	9.53	0.034	0.034
90%	13.72	13.72	13.72	0.000	0.000

sum of squares	**SS**	**0.314**
average SS	**SS/N**	**0.0197**
standard deviation	**sqrt(SS/N)**	**0.1402**

Table 3. Reproducibility: The Excel spreadsheet for eight paired case reviews. For each case (row), Reviewers A and B inferred profiles having the MI scores shown (columns) for each case. The mean "m" column has the average value of the MI values recorded within each case row. The last two columns show the square deviation between the reviewer's MI and the average MI for that case. Also shown are the SS, average SS and the standard deviation results.

	Accuracy	Precision	Reproducibility
Conservative	11.91	-5.41	0.140
Aggressive	15.93	-1.39	0.833
Objective	17.24	-0.09	0.095

Table 4. Validation results for different mixture interpretation methods using a set of Match Information statistics. The interpretation methods methods (rows) are government "conservative" and private lab "aggressive" human review, and TrueAllele "objective" computer review. The statistics (columns) are accuracy (average match information), precision (average loss of match information), and reproducibility (standard deviation within case).