J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Pittsburgh, PA, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Research Article

Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information

Mark William Perlin[1]

[1]Cybergenetics, Pittsburgh, USA

E-mail: *Dr. Mark William Perlin - perlin@cybgen.com
*Corresponding author

Received: 16 July 2015 Accepted: 21 September 2015 Published: 28 October 2015

Abstract

Background: DNA mixtures of two or more people are a common type of forensic crime scene evidence. A match statistic that connects the evidence to a criminal defendant is usually needed for court. Jurors rely on this strength of match to help decide guilt or innocence. However, the reliability of unsophisticated match statistics for DNA mixtures has been questioned. **Materials and Methods:** The most prevalent match statistic for DNA mixtures is the combined probability of inclusion (CPI), used by crime labs for over 15 years. When testing 13 short tandem repeat (STR) genetic loci, the CPI^{-1} value is typically around a million, regardless of DNA mixture composition. However, actual identification information, as measured by a likelihood ratio (LR), spans a much broader range. This study examined probability of inclusion (PI) mixture statistics for 517 locus experiments drawn from 16 reported cases and compared them with LR locus information calculated independently on the same data. The log(PI^{-1}) values were examined and compared with corresponding log(LR) values. **Results:** The LR and CPI methods were compared in case examples of false inclusion, false exclusion, a homicide, and criminal justice outcomes. Statistical analysis of crime laboratory STR data shows that inclusion match statistics exhibit a truncated normal distribution having zero center, with little correlation to actual identification information. By the law of large numbers (LLN), CPI^{-1} increases with the number of tested genetic loci, regardless of DNA mixture composition or match information. These statistical findings explain why CPI is relatively constant, with implications for DNA policy, criminal justice, cost of crime, and crime prevention. **Conclusions:** Forensic crime laboratories have generated CPI statistics on hundreds of thousands of DNA mixture evidence items. However, this commonly used match statistic behaves like a random generator of inclusionary values, following the LLN rather than measuring identification information. A quantitative CPI number adds little meaningful information beyond the analyst's initial qualitative assessment that a person's DNA is included in a mixture. Statistical methods for reporting on DNA mixture evidence should be scientifically validated before they are relied upon by criminal justice.

Key words: DNA mixture interpretation, forensic science, identification information, inclusion probability, likelihood ratio

Access this article online

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.168525

Quick Response Code:

This article may be cited as:
Perlin MW. Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information. J Pathol Inform 2015;6:59.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2015/6/1/59/168525

BACKGROUND

"Among existing forensic methods, only nuclear DNA analysis has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between an evidentiary sample and a specific individual or source… However,… there may be problems… with how the DNA was… interpreted, such as when there are mixed samples…"[1]

National Academy of Sciences, "Strengthening Forensic Science" (2009), page 100.

DNA mixtures arise when more than one person contributes their DNA to a biological sample. Greater sample diversity and instrument sensitivity have increased the volume of mixture evidence in the crime laboratory. Whereas single source DNA can uniquely identify an individual, mixtures give a statistical association between evidence and person. Courts rely on these match statistics to establish the probative value of DNA mixture evidence.

A likelihood ratio (LR) quantifies the evidential impact of data on a hypothesis. The base 10 logarithm of the LR is a standard measure of information, expressed in "ban" units. In forensic biology, a usual hypothesis is that some particular person contributed their DNA to a mixture sample. The data are derived from short tandem repeat (STR) experiments performed on the mixture DNA molecules. All DNA match statistics used in forensic identification are LRs, at least formally.

One way to calculate the LR for a DNA mixture is to separate the STR data into the genotypes of each contributor. Since a person's genotype at a locus is a pair of alleles, and there may be uncertainty in the separation, a contributor's genotype is a discrete probability distribution over allele pair possibilities. When using a mathematical model that faithfully accounts for the quantitative data, sources of variation, and known artifacts, an inferred genotype becomes a useful statistical summary of a contributor's genetic identity. Comparison of this separated evidence genotype with that of a known subject, relative to a population, yields an LR for that person having contributed their DNA to the mixture. The $\log_{10}$(LR) of this number quantifies the identification information.

Previous validation studies[2,3] on computer-based mixture separation methods show a broad distribution of log(LR) match information over an ensemble of randomly chosen DNA casework samples, as shown in Figures 1a and 2a (blue). This happens because log(LR) information is roughly proportional to DNA contributor amounts, and

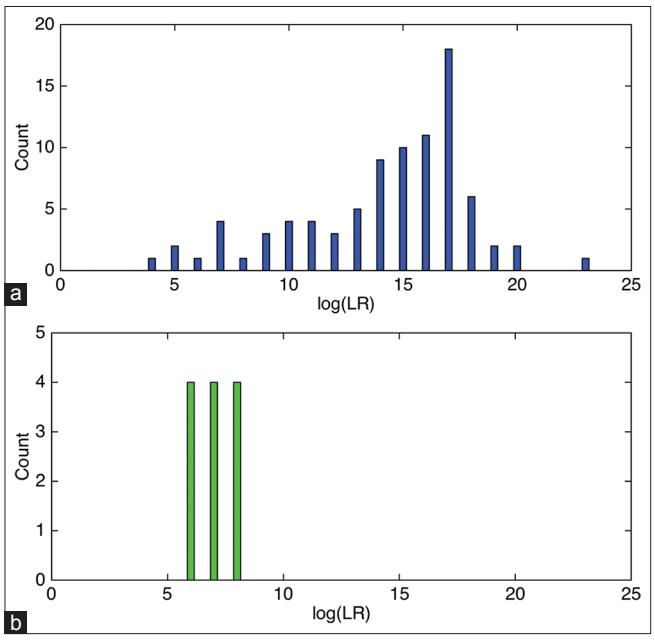

Figure 1: New York validation study histograms compare TrueAllele and combined probability of inclusion (CPI) on the same mixture data. A New York validation study recorded match information for casework DNA mixture samples that were tested using the Federal Bureau of Investigation 13 core short tandem repeat loci. Two log(LR) histograms are shown. (a) Computer-based mixture separation methods showed a broad log(LR) frequency distribution over an ensemble of 87 match comparisons (blue). (b) Human review of 12 mixtures gave a log(CPI^{-1}) frequency distribution that was narrowly centered around six, corresponding to a CPI^{-1} value of around a million (green)

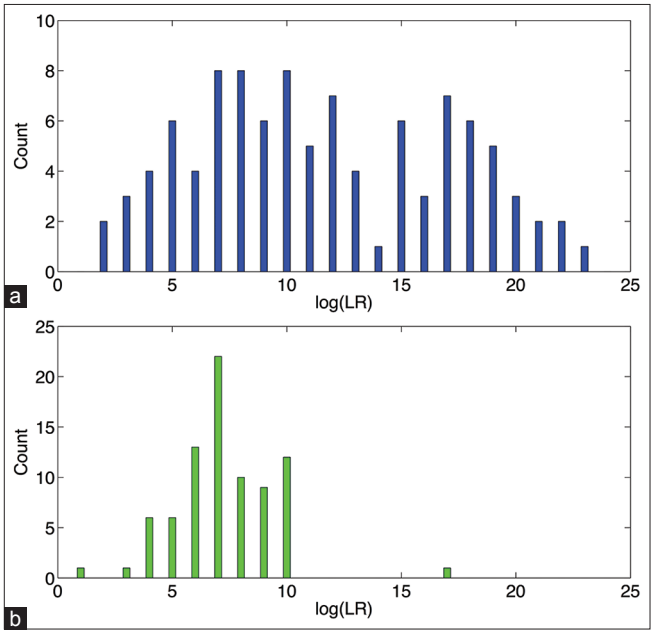

Figure 2: Virginia validation study histograms compare TrueAllele and combined probability of inclusion (CPI) on the same mixture data. A Virginia validation study recorded match information for casework DNA mixture samples that were tested using 15 short tandem repeat loci. Two log(LR) histograms are shown. (a) Computer-based mixture separation methods showed a broad log (LR) frequency distribution over an ensemble of 101 match comparisons (blue). (b) Human review of 81 genotype comparisons gave a log(CPI^{-1}) frequency distribution that was largely centered around seven, corresponding to a CPI^{-1} value of around 10 million (green). (Reproduced with permission from Figure 7 of Perlin *et al.*, PLoS ONE 2014;9(3): E92837)

so random DNA quantities give similarly random log(LR) values.[4,5] Therefore, with accurate and detailed computer modeling of DNA mixture evidence, one expects to see a uniform distribution of identification information, ranging from none ($\log_{10}(1) = 0$) to single source levels ($\log_{10}(10^{24}) = 24$) when using 13 STR loci.

Current forensic practice generally does not use sophisticated quantitative modeling to analyze DNA mixtures. An STR data signal [Figure 3] is comprised of quantitative peaks that correspond to allele sizes. The peak heights range from 10 to 10,000 relative fluorescent units (RFU). Applying an RFU threshold simplifies STR data into all-or-nothing "allele" events; peaks over the threshold are in the allele set, while those under are out.

At a locus, if the subject's one or two alleles are included in the mixture's allele set, then the subject is said to be "included" in the mixture. A probability of inclusion (PI) is then calculated at the locus by adding together the population frequencies of the alleles over threshold and squaring their sum. The PI reciprocal PI^{-1} is formally an LR, so $\log_{10}(PI^{-1})$ measures this method's information content.

Multiplying the PI values of included loci produces the standard combined probability of inclusion (CPI) match statistic for the mixture. CPI has been the dominant DNA mixture statistic in the United States for 15 years,[6] used on hundreds of thousands of evidence items. Yet CPI has an interesting property – it usually returns the same number, regardless of the item analyzed. With the Federal Bureau of Investigation (FBI's) 13 core loci, CPI^{-1} is around a million;[2] with 15 loci (13 core, plus 2 more), the statistic is around 10 million,[3] as shown in Figures 1b and 2b (green).

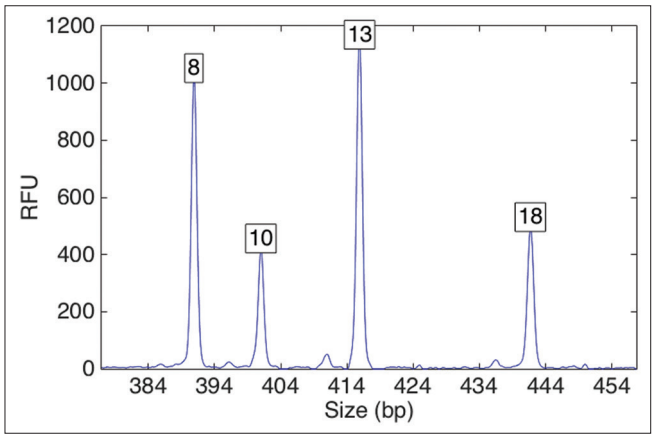

Figure 3: Short tandem repeat (STR) locus data with PI giving statistical false inclusion. An STR signal at the Penta E locus comprised of data peaks. The x value is allele size (bp), y is allele quantity (RFU), and the allele designation is shown at the top of a peak (box). Four allele peaks are shown – a tall 8.13 pair at 1000 RFU, and a short 10.18 pair at 400 RFU. These two allele pairs correspond to the two actual contributor genotypes in this mixture, with each pair having allele peaks of comparable height. Inclusion mixture analysis does not account for peak height, and so the PI match statistic can be overstated, falsely suggesting an inclusion

If sophisticated analysis shows a full spectrum of log(LR) information between 0 and 24, why is the $\log(CPI^{-1})$ value centered around six? This paper examines the disparity between relatively constant CPI^{-1} logarithm values and the broad range of true log(LR) information values. The explanatory hypothesis is that CPI is essentially unrelated to identification information. The approach here is to assess this hypothesis by fractionating the CPI statistics of many mixture samples into their individual locus PI values and studying their empirical distribution.

The paper describes materials and methods for analyzing STR mixture data. It then compares the LR and CPI methods in case examples of false inclusion, false exclusion, a homicide, and criminal justice outcomes. Statistical analysis of crime laboratory STR data shows that inclusion match statistics exhibit a truncated normal distribution having zero center, with little correlation to actual identification information. These statistical findings explain why CPI is relatively constant, with implications for DNA policy, criminal justice, cost of crime, and crime prevention. Statistical methods for reporting on DNA mixture evidence should be scientifically validated before they are relied upon by criminal justice.

MATERIALS AND METHODS

Separating Genotypes for Identification Information

Genotype separation

A hierarchical Bayesian probability model can be used to model DNA mixture data.[4,7] Matrix algebra linearly combines genotype allele pairs to form an STR peak height pattern vector as the mean of a multivariate distribution.[8] Variance parameters describe peak event and artifact variation, computed from a sample's locus experiments;[9] a peak's variation scales with its height.[10] Markov chain Monte Carlo (MCMC)[11] solution of the hierarchical model[12] separates the genotypes into allele pairs for each contributor at a locus. A contributor's genotype probability is a marginal distribution of the joint MCMC posterior distribution.

Likelihood ratio

The LR numerically summarizes the information change from one probability measure to another.[13] With $h(x)$ the prior probability of a genotype allele pair x at a locus based on population data, and $f(x)$ the MCMC posterior genotype probability after having seen STR data, comparison with subject genotype $g(x)$ gives the LR

$$LR = \sum_{x \in G} \frac{f(x)g(x)}{h(x)}$$

where the sum is taken over the discrete set G of genotype allele values. When genotype $g(x)$ corresponds to a reference subject, the function places all of its

probability mass at the subject's one allele pair value s, and the LR reduces to

$$LR = \frac{f(s)}{h(s)}$$

a posterior-to-prior genotype probability ratio.[14] Using Bayes theorem,[15] other LR forms can be written.[16] The $\log_{10}$(LR) is a standard dimensionless measure of information.[17]

TrueAllele software

TrueAllele® Casework (Cybergenetics, Pittsburgh, PA) is a computer program that mathematically separates STR mixture data into contributor genotypes, and then compares such genotypes to calculate LRs. The two-step process is an objective procedure, since the initial mixture separation step does not have the subject's genotype available. Sufficiently long MCMC sampling ensures a thorough statistical assessment of the genotype (and other parameter) possibilities. Numerous validation studies show that the program produces accurate log(LR) information values.[2-5,16,18,19]

Applying Thresholds for Inclusion Probability

Threshold application

A simple way to interpret STR mixtures is to heuristically classify some data peaks as confident "allele" events, and then develop a match statistic from these putative alleles. A DNA laboratory sets a "threshold" as an RFU value above which they are comfortable calling a data peak an allele that represents signal, rather than noise. At a sample's STR locus, the allele designations having peaks over threshold are collected into an allele set. When the one or two alleles of a subject are all included in the mixture sample's allele set, the subject is said to be "included" at the locus. The comparison is made between a genotype and (features of) the data, rather than between two genotypes. There is potential for subjective bias[20] because subject information (e.g., for a defendant) is used in making this "inclusion" determination.[21]

Inclusion probability

Once an inclusion has been declared, an inclusion probability can be calculated for the locus. The locus PI is the sum of the population frequencies of all allele pairs included in the mixture allele set. This probability can be divided into a sum of homozygous genotype possibilities, plus a sum of heterozygous possibilities

$$PI = \sum_{i \in I} p_i^2 + 2 \sum_{\substack{i,j \in I \\ i \neq j}} p_i p_j$$

where I is the set of included alleles and p_i is the allele frequency of allele i. Algebraically rearranging terms gives the more familiar squared sum of allele frequencies

$$PI = \left(\sum_{i \in I} p_i \right)^2$$

so that

$$PI^{-1} = 1 \Big/ \left(\sum_{i \in I} p_i \right)^2$$

The PI^{-1} is a qualitative LR based on an all-or-nothing likelihood function for a set of included alleles.[22] Therefore, log(PI^{-1}) inclusion statistics can be meaningfully compared with log(LR) information values.

Population statistics software

PopStats population statistics software (FBI, Quantico, VA) can calculate a sample's CPI statistic based on allele inclusion sets. The user also supplies locus allele frequency databases for human populations of interest. PopStats then gives the PI values for the inclusion sets at each locus, relative to a population. Multiplying these PI values together yields the combined CPI match statistic.

Mixture Data and Match Statistics

Mixture data

An accredited American crime laboratory analyzed 31 mixture items from 16 criminal cases. These items were amplified using a PowerPlex® 16 STR kit (Promega, Madison, WI), and size separated on an AB 3130® genetic analyzer (Life Technologies, Foster City, CA). The genetic analyzer's electropherogram (EPG) signal data were recorded for each item in a fragment size analysis (.fsa) electronic file.

Likelihood ratio statistics

TrueAllele Casework analyzed the DNA mixture .fsa files to separate genotypes and calculate LR match statistics (VUIer™ version 2014a). The MCMC process sampled at least 50,000 burn in and read out cycles. All items were minimally run in duplicate, and the average concordant log(LR) value was recorded. Locus LR values were collated for all loci. A co-ancestry coefficient of 1% was applied, using a generalization of the θ correction formula of NRC 4.10.[23]

Locus inclusion

The crime laboratory determined that 41 individuals were included in the 31 mixtures. An individual could be included in an item at up to 15 STR loci. A stochastic threshold of 150 RFU was used to identify inclusionary loci that had sufficient peak heights for statistical reporting.[24] A total of 517 statistically usable locus inclusions were identified.

Probability of inclusion statistics

The crime laboratory used PopStats to calculate a PI value at every included locus. A co-ancestry coefficient of 1% was used, following the θ correction formula of NRC 4.4.[23] Co-ancestry accounts for relatedness within a human population, and conservatively reduces a DNA match statistic.

Population databases

African-American, Caucasian, and Hispanic databases developed by the Pennsylvania State Police were used

to calculate allele frequencies. Multiplying these allele frequencies together produces the prior probability $h(x)$ of a random person, providing a genotype rarity for the match statistic denominator. The smallest match statistic across the three populations was recorded.

CASE RESULTS

Using the case locus data, examples are given of how inclusion analysis can falsely include (hence wrongly implicate) or falsely exclude (i.e., incorrectly exonerate) someone from a mixture. A homicide case is presented that shows considerable disparity between accurate LR analysis and threshold-based inclusion. Finally, the impact of reliable mixture analysis on criminal justice is demonstrated by examining the outcomes of 72 cases from a method comparison study.

False Inclusion

The inclusion method applies a threshold to simplify a quantitative STR peak pattern into a list of putative "allele" events. Imposing a threshold discards considerable information. Peak heights are not used. Also lost are the patterns of taller and shorter peaks that can help separate genotypes and assess data variation. Ignoring data can affect the PI, distorting the match statistic's reporting of inclusion.

Figure 3 shows four allele peaks at the Penta E locus – a tall 8.13 pair at 1000 RFU and a short 10.18 pair at 400 RFU. These two allele pairs correspond to the actual contributor genotypes in this mixture, with each pair of peaks having comparable height. However, the inclusion method ignores peak height; each of the 10 possible pairings of the four alleles (8, 10, 13, and 18) is given equal likelihood. In this case, an individual with a 10.13 genotype was falsely included at the locus. Inclusion ignores the peak pattern (two tall, two short), and so cannot recognize that the 10.13 genotype is an unlikely pairing of dissimilar (short and tall) allele peak heights. The overstated PI for this noninclusion is a high 9.0866 ($10^{0.9584}$).

False Exclusion

Figure 4 shows the STR data pattern from locus D3S1358. There are six peaks over threshold, a tall pair (15 and 19) at 1200 RFU, and four shorter ones (14, 16, 17, and 18) at 200 RFU. The obvious 15.19 peak pair corresponds to a person with a 15.19 genotype who contributed their DNA to this mixture. However, inclusion with thresholds ignores peak height and so confers equal status to all six alleles; equal likelihood is then assigned to the twenty-one ($n(n + 1)/2$, $n = 6$) allele pair possibilities. The inclusion method does not identify the major contributor as a highly informative 15.19 genotype. Therefore, the low PI^{-1} of 1.4484 ($10^{0.1609}$) greatly understates the true probative value of this locus data. An inability to separate genotypes leads to a statistical false exclusion.

Yelenic Homicide Case

In April of 2006, Blairsville dentist John Yelenic was repeatedly slashed on his face and throat in his Pennsylvania home.[25] Dr. Yelenic was partially decapitated after his head was thrust through a side pane window by his front door. He died after exsanguinating onto his living room floor. The decedent's upper extremities showed multiple defensive wounds [Figure 5, black arrows].

The main suspect was Pennsylvania State Trooper Kevin Foley, who at the time resided with Yelenic's estranged wife, Michele. Found on the living room coffee table were the unsigned Yelenic divorce papers, splattered with the victim's blood.

Trace DNA under the decedent's fingernails [Figure 5, blue arrows] revealed a mixture containing 93.3% of the dentist and 6.7% of an unknown person. The FBI tested the fingernail DNA using Profiler Plus® and COfiler® STR kits (Life Technologies, Foster City, CA). The FBI determined that trooper Foley's DNA was included at 11 of 13 core STR loci having sufficient peak heights and reported a CPI statistic of 13 thousand. The locus breakdown of the PI statistics is shown [Table 1, log(1/PI)].

Cybergenetics independently analyzed the same STR data, using TrueAllele to separate out the minor genotype of the 6.7% unknown contributor. Comparison of this minor genotype with Foley's known reference gave the LR match statistic of 189 billion.[4] The locus breakdown of this LR statistic is shown [Table 1, log(LR)]. Because the TPOX data had EPG spike artifacts, that locus was not used.

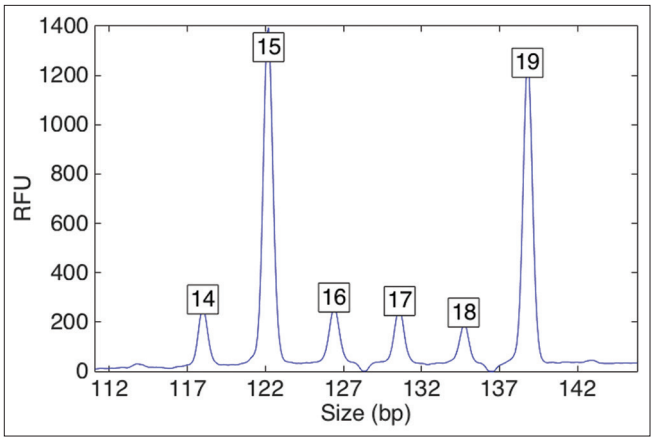

Figure 4: Short tandem repeat (STR) locus data with PI giving statistical false exclusion. The data show an STR data pattern at locus D3S1358. The x value is allele size (bp), y is allele quantity (RFU), and the allele designation is shown at the top of a peak (box). There are six peaks over threshold, a tall pair (15 and 19) at 1200 RFU and four shorter ones (14, 16, 17, and 18) at 200 RFU. The tall 15.19 peak pair corresponds to a genotype present in this mixture clearly differentiated from the four shorter peaks. The inclusion method does not separate mixture peaks into genotypes, and so a PI match statistic can be greatly understated, falsely suggesting exclusion

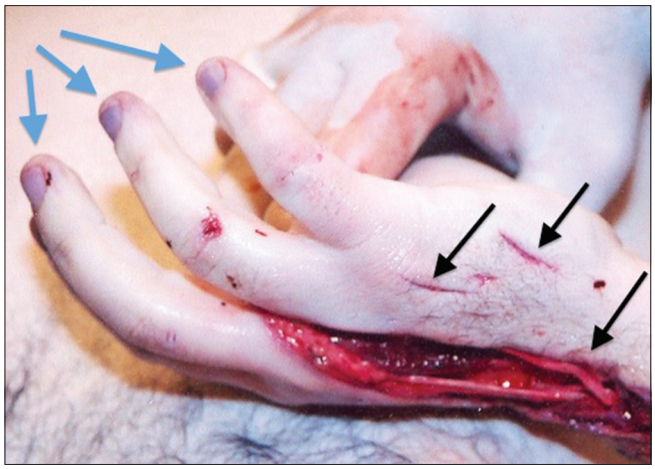

Figure 5: Decedent's hand with defensive wounds and fingernail DNA. An autopsy photograph of the decedent's right hand shows multiple defensive wounds (black arrows). Trace DNA under the fingernails (blue arrows) contained a mixture of two people, primarily the decedent, but also a small amount of an unknown person

Table 1: Locus breakdown of TrueAllele and CPI statistics from Yelenic case shows little correlation

Locus	log (LR)	log (1/PI)	Difference
CSF1PO	0.329	0.314	0.015
D13S317	0.060	0.277	−0.216
D16S539	0.989	0.222	0.768
D18S51	1.408		
D21S11	0.903		
D3S1358	0.671	0.520	0.150
D5S818	0.541	0.082	0.459
D7S820	1.765	0.411	1.354
D8S1179	1.460	0.681	0.779
FGA	0.980	0.870	0.110
TH01	1.382	0.229	1.154
TPOX		0.196	
vWA	0.791	0.327	0.464
Total	11.279	4.127	5.037
Average	0.940	0.375	0.504

Match statistics are shown for the Yelenic fingernail evidence relative to the Foley reference genotype. The first column lists the STR loci examined, the second gives the computer's log(LR) value, the third has the human-scored logarithmic PI value, while the last column gives the numerical difference log(LR) – log(PI^{-1}). The "total" match statistic sums over all analyzed loci, and the "average" statistic per locus is shown. PI: Probability of inclusion, LR: Likelihood ratio, CPI: Combined probability of inclusion

The TrueAllele LR was larger than the PI number at 9 of the 10 common loci that were analyzed using both methods [Table 1, Difference]. The average log(LR) information was 0.940 ban per locus (ban/loc) for TrueAllele, and 0.375 ban/loc for PI, showing an average inclusion method information loss of 0.504 ban/loc. The small coefficient of determination (r^2 = 0.097) indicates little correlation between TrueAllele's identification information and the heuristic inclusion statistic.

At Foley's trial, the FBI testified about their CPI results for Yelenic's fingernail DNA, and (following an admissibility hearing) Cybergenetics presented TrueAllele statistics on the same data. Foley was convicted of first-degree murder and is serving a life sentence. Appellate courts denied Foley's appeal, establishing a legal precedent for TrueAllele computer DNA interpretation in Pennsylvania.[26]

Criminal Justice Outcomes

The Virginia Department of Forensic Science had Cybergenetics report on 72 criminal cases where CPI analysis of DNA mixture items was inconclusive or uninformative.[3] TrueAllele reanalysis of the same data provided match statistics for 101 of the 111 requested genotype comparisons (91%). The computer's match statistics reintroduced DNA mixtures into these cases, facilitating evidence-based criminal justice outcomes.

The crime laboratory's original log(CPI^{-1}) average was 6.825 ban (6.68 million match statistic). However, stochastic thresholds rendered half of the items inconclusive, and statistically removed two thirds of the loci from the remainder to give an uninformative average of 2.145 ban (140). Using all items and loci, the computer's log(LR) averaged 11.054 ban (113 billion).

The 72 cases spanned a full range of violent offenses, including 18 prosecuted murders, 12 robberies, and 6 rapes or sexual assaults [Table 2, Reports]. There were 20 cases involving weapons, reflecting a considerable number of touch DNA firearm cases. The DNA evidence items were all mixtures, most having three contributors and some with four.

Defendants were found guilty in 59 cases [Table 2, Convictions], a 78.7% conviction rate (out of

Table 2: Criminal justice outcomes for Virginia cases where TrueAllele was used instead of CPI

Offense	Reports	Convictions	Guilty pleas	Life	Years
Homicide	18	16	11	4	32.1
Rape/sexual assault	6	6	2	1	40.1
Robbery	12	11	7		30.5
Weapons	20	17	12		11.2
Drugs	4	4	4		5.9
Other	12	5	1	1	10.2
Total	72	59	37	6	19.6

Criminal justice outcomes are shown for 72 reported Virginia cases that were analyzed by TrueAllele (three cases had two defendants, for a total of 75 defendants). For each criminal offense (rows), the number of reports, convictions, and guilty pleas are listed in separate columns. The last two columns list the number of life sentences, along with the average number of years in prison for nonlife sentences. CPI: Combined probability of inclusion

75 defendants). TrueAllele trial testimony was given in 10 cases. A guilty plea was entered in 37 of the 59 convictions (62.7%), thereby avoiding a trial. Indeed, most of the 72 cases with TrueAllele reports led to a guilty plea (51.4%). Relative to having a trial, a guilty plea is a cost-effective strategy for reducing crime.[27]

Of 18 homicides, 2 defendants were found "not guilty by insanity". The other 16 were convicted, with 4 sentenced to life in prison and 12 receiving an average of 32 years in prison [Table 2, Life and Years]. With rape or sexual assault, all defendants were convicted, with one life sentence, and average prison time of 40 years. There were 11 convictions for robbery (91.7%) and 17 for weapons violations (85%).

STATISTICAL RESULTS

This section provides statistical support for the hypothesis that inclusion probability for DNA mixtures is a subjective, one-sided match statistic unrelated to identification information. It also shows why more STR loci generally yield a higher CPI statistic (regardless of information), and why the CPI value has tended to be around a million.

One-sided Match Statistic

A probability is a number (inclusively) between zero and one.[28] Since an "inclusion" event must include the subject's allele pair, the PI is bounded below by that person's genotype frequency in the population, which is a positive number. The locus PI can attain unity only when all possible alleles are observed. That surfeit of STR alleles would indicate too many contributors for a crime lab's protocols to permit CPI analysis. Therefore, a PI at a locus is strictly greater than zero and less than one.

It follows that the reciprocal PI^{-1} value is a finite number greater than one. Hence, $\log(PI^{-1})$ is a positive number; exclusionary loci are given no weight. The $\log(PI^{-1})$ distribution in the data set is shown in a histogram [Figure 6]. The minimum locus value is 0.0023 ban and the maximum is 2.0644 ban. Clearly, $\log(PI^{-1})$ is a one-sided match statistic that only assumes positive real values.

Truncated Normal Distribution

The Figure 6 histogram looks like the right half of a unimodal distribution. One can symmetrize this distribution by augmenting the $\log(PI^{-1})$ set with corresponding negative values. The symmetrical histogram in Figure 7 is shaped like a normal distribution. Normality is tested here in two different ways.

A normal probability plot of the symmetrized $\log(PI^{-1})$ data shows excellent agreement with a straight line [Figure 8]. The Lilliefors test[29] accepts the null hypothesis that the data come from a distribution in the normal family ($p = 0.3825$), confirming statistically that the symmetrized data are normally distributed. Therefore, the $\log(PI^{-1})$ data follow a truncated normal distribution.

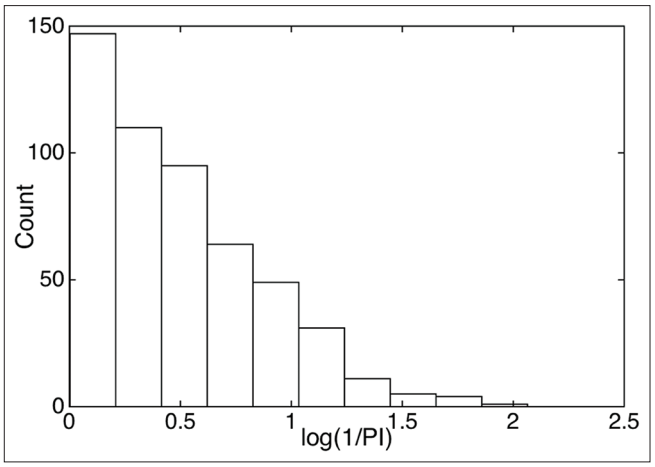

Figure 6: Inclusion has a positive-valued truncated normal distribution centered at zero. A histogram of 517 $\log(PI^{-1})$ locus values that shows a positive valued distribution. The values form the right half of a $N_+(0, 0.6220)$ truncated normal distribution that is centered at zero and has a positive variance

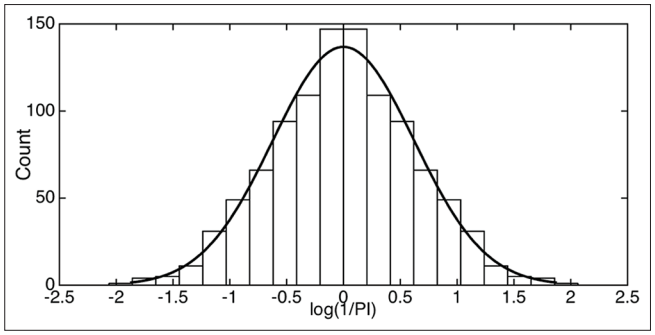

Figure 7: Symmetrized inclusion distribution formed by adding negative values. Augmenting the positive locus $\log(PI^{-1})$ values with their corresponding negative values forms a symmetrical histogram shaped like a normal distribution. A fitted normal curve is superimposed

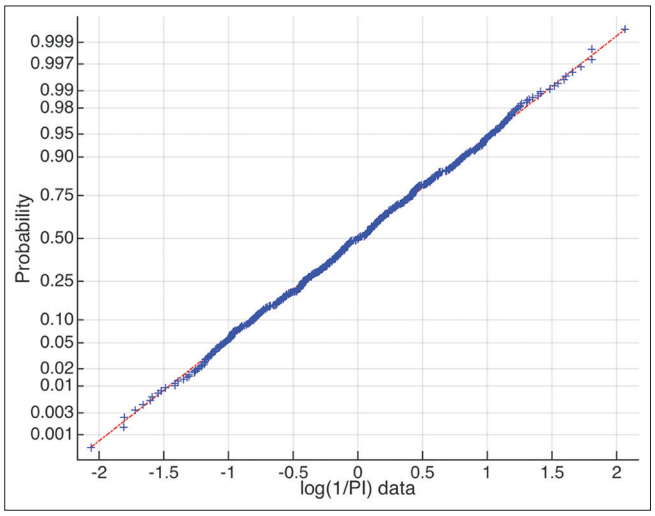

Figure 8: Normal probability plot shows symmetrized inclusion follows normal distribution. A normal probability plot of the symmetrized locus $\log(PI^{-1})$ data. The points (blue plus) are situated along a straight line (red dash) which indicates the data are normally distributed

Positive Tail Centered at Zero

The parameters of a nonnegative truncated normal distribution are its center μ and the spread σ. Maximum likelihood fitting of the log(PI^{-1}) data to a truncated normal on the nonnegative real numbers gives the estimates μ = 0 ban/loc and σ = 0.6220. Thus, the log(PI^{-1}) values form the right half of a normal distribution, centered at zero with a positive variance [Figure 6].

Uncorrelated with Identification Information

For each reported locus, the log(PI^{-1}) inclusion statistic can be compared with reliable log(LR) identification information computed from the same data. A scatterplot comparison is shown in Figure 9. The x-axis gives real-valued log(LR) information, whereas the y-axis shows the reported positive log(PI^{-1}) statistic.

The correlation coefficient of this joint data is $r = 0.4360$. The low r^2 value of 0.1901 demonstrates little statistical correlation between the log(PI^{-1}) inclusion statistic and log(LR) identification information.

Inclusion Distribution Has a Positive Mean

The locus log(PI^{-1}) values are distributed as a truncated normal [Figure 6]. The positive values in this distribution have an empirical mean $\bar{x} = 0.4937$ ban/loc, with a standard deviation of $\bar{s} = 0.3787$. Raising this logarithmic mean to a power of 10, the expected PI^{-1} value at an included locus is 3.1167 ($10^{0.4937}$).

Law of Large Numbers

The law of large numbers (LLN) holds that a sum of L identically distributed random variables, each with mean $\bar{x}$, will have an expected value of $L \cdot \bar{x}$[15] Therefore, as the number loci L increases, so too will the average combined CPI value. On average, more loci will yield a higher CPI statistic, regardless of the sample's information content.

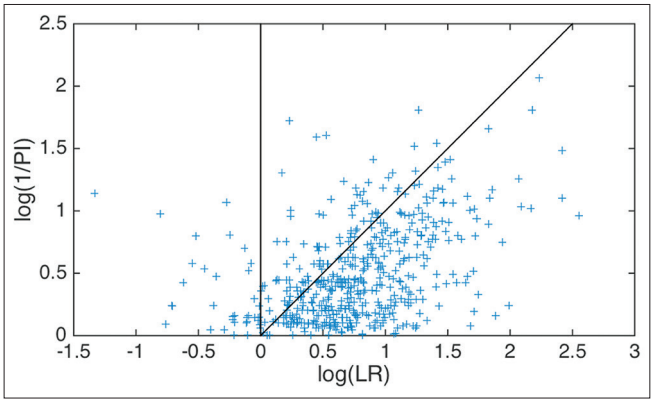

Figure 9: Scatterplot shows little correlation between TrueAllele and inclusion locus values. A scatterplot compares log(PI^{-1}) inclusion statistics (*y*-axis) with log(LR) identification information (x-axis) calculated on the same locus data. At each data point (blue plus), x gives the log(LR) information, while y is the (positive valued) log(PI^{-1}) statistic. Visually, there is little correlation between PI inclusion and likelihood ratio information, since the points are widely dispersed and do not reside along the y = x (black) line

In the New York (NY) study,[2] the 12 log(CPI^{-1}) values averaged 6.581 ban. There were L = 13 STR loci used. Dividing log(CPI^{-1}) by 13 loci gives a NY locus log(PI^{-1}) average $\bar{x}_{NY}$ of 0.5062 ban/loc. The Virginia (VA) study[3] had 81 log(CPI^{-1}) values averaging 6.825 ban. With L = 15 STR loci per sample, the average VA locus log(PI^{-1}) $\bar{x}_{VA}$ is 0.455 ban/loc.

Why CPI is Always a Million

The average log(PI^{-1}) statistic was $\bar{x} = 0.4937$ ban/loc in the study. Different studies on other CPI data show similar locus log(PI^{-1}) averages around a half, with $\bar{x}_{NY} = 0.506$ ban/loc and $\bar{x}_{VA} = 0.455$ ban/loc. By LLN, the expected log(CPI^{-1}) statistic is $L \cdot \bar{x}$, where L is the number of STR loci that a lab reported for the mixture.

With the NY study, $L \cdot \bar{x}_{NY}$ = (13 loci) × (0.5062 ban/loc), or 6.581 ban. Raising 10 to the power of this logarithmic value gives $10^{6.581}$, or 3.81 million, the expected CPI^{-1} statistic from NY. For the VA study, $L \cdot \bar{x}_{VA}$ = (15 loci) × (0.455 ban/loc), which equals 6.825 ban. Exponentiating this logarithmic value as $10^{6.825}$ gives an average VA CPI^{-1} statistic of 6.68 million.

DISCUSSION

The statistical PI locus results demonstrated little correlation between CPI and DNA identification information. This CPI deficiency has an impact on forensic science, criminal justice, and DNA policy. The Virginia case outcomes, where TrueAllele was used to resolve DNA mixture evidence when CPI proved to be uninformative, have implications for the cost and prevention of crime.

CPI has Zero Information

The log(PI^{-1}) locus match statistics follow a $N_+(0, 0.6220)$ truncated normal distribution [Figure 6]. Sampling from this distribution produces positive numbers from a right tail centered at zero. The sampling is censored since the inclusion method cannot produce an exclusionary PI statistic.

Perhaps PI provides no actual identification information since its truncated distribution is centered at 0 ban. An observed CPI statistic may merely be a (logarithmic) sum of positive numbers, randomly sampled from the truncated distribution's positive right tail.

Uncorrelated with Identification Information

The scatterplot visually indicates little correlation between PI inclusion and LR information [Figure 9]. Positive x values with zero y values suggest false exclusions, whereas negative x values with positive y values indicate false inclusions. Higher points along the y-axis show inclusionary PI statistics that have no actual identification information. A strong correlation would place most of the points along a line of slope one in the first quadrant, but this arrangement is not seen.

For a DNA match statistic to meaningfully identify people, it should correlate with identification information. CPI does not possess such correlation, so it may not be suitable for measuring identification strength.

DNA Policy and Number of Loci

The log(PI^{-1}) average is positive. By LLN, testing more loci will give a larger CPI^{-1} match number. However, CPI is largely uncorrelated with identification information, so testing more loci may not actually add useful information.

The FBI is requiring crime laboratories to use larger STR kits containing seven additional loci.[30] This policy is expensive since kits with more loci cost more money. Moreover, there is the cost of transition and kit revalidation. More loci impose a heavier data burden that taxes analyst productivity through increased DNA laboratory analysis, data artifacts, data review, case reporting, and trial testimony effort. Since there is little information gained from CPI interpretation of mixture data, the additional loci may not be worth the extra cost.

A "Match Estimator" module in the PopStats software uses inclusion probability to determine whether a DNA mixture may be uploaded to the FBI's Combined DNA Index System (CODIS) DNA database.[31] This filtering policy blocks the CODIS upload of most mixture evidence, impeding criminal investigation. Testing additional STR loci can increase Match Estimator values, allowing more mixtures to be uploaded to CODIS. However, this increase would come from extra tests, and not from extra identification information.

Criminal Justice and Cost of Crime

Eliciting more identification information from DNA mixture evidence has an impact on criminal justice. DNA evidence without a match statistic is usually not admissible; the absence of DNA can hinder the prosecution of violent crimes. By providing a meaningful statistic where CPI cannot, a computer can bring DNA back into a case.

Societal impact can be quantified by examining the total cost of crime, both tangible and intangible.[32] Tangible costs include loss of productivity and property. Intangible costs include medical expenses, lost earnings, and psychological damage. Total crime cost is estimated to be $8,982,907 for a homicide and $240,776 for rape or sexual assault.[32]

In the Virginia study, TrueAllele restored DNA evidence in 20 violent crimes where the defendant was convicted. The total cost of crime can be estimated for these cases. Combining 14 homicides × $8,982,907/homicide, with 6 rape or sexual assaults × $240,776/rape, gives a total crime cost of $127,205,354.

Crime Prevention Through Incarceration

Criminals often reoffend. After release from prison, a violent offender has a 33% recidivism rate of being arrested for another violent crime within 5 years.[33] Therefore, 15 years of violent offender incarceration roughly translates into preventing 1 violent crime. This rate is conservative since just 72% of homicide offenses (and only 21% of rapes) result in an arrest.[34]

In Virginia's computer reexamination of 72 mixture cases, the average prison term (excluding life sentences) for the 20 convicted murderers and rapists was 35 years [Table 2]. Thus, the total incarceration time for these violent offenders is 700 years, assuming the criminals serve out their full sentences. Multiplying this total time by the "1 reoffending violent crime every 15 years" rate conservatively estimates that TrueAllele's reanalysis of the 72 Virginia cases will help prevent about 46.7 violent crimes.

CONCLUSIONS

Forensic DNA laboratories apply thresholds to their STR data, simplifying the quantitative peak pattern into a qualitative list of "allele" peaks. When inclusionary criteria are met for a mixture and a subject, a PI is calculated for the included loci; these numbers are multiplied together to calculate a CPI match statistic. With a full complement of 13 core STR loci, many labs will report a CPI^{-1} of about a million, often with little variation around this average. A DNA statistic at this level can be persuasive to juries.[35]

The results reported in this paper explain why CPI^{-1} is often around a million. The positive log(PI^{-1}) values were distributed as a truncated normal centered at zero, with a standard deviation of 0.6220. The inclusion method censors the left half of the true distribution, as only positive inclusionary values can be reported. The positive-valued log(PI^{-1}) distribution had mean 0.4937 ban/loc. By LLN, randomly sampling log(PI^{-1}) values for 13 STR loci gives an expected total statistic of $10^{13 \times 0.4937}$ or 6.42 million. With 15 loci, the $10^{15 \times 0.4937}$ expected total is 25.44 million.

CPI did not correlate well with identification information ($r^2 = 0.19$), leaving 81% of the variance unexplained. This lack of correlation between the CPI statistic and LR information has been previously observed.[3] Courts generally require a match statistic for DNA evidence, and CPI can supply such a number. Testing more STR loci usually yields a bigger CPI number, as explained by LLN. Since log(PI^{-1}) is largely uncorrelated with DNA information, a bigger CPI statistic is just a bigger number, not more identification information. While a larger LR conveys greater probative weight for DNA identification, a larger CPI does not.

DNA mixtures are a highly prevalent type of biological evidence, with hundreds of thousands of items having been tested in criminal cases. How these mixtures are interpreted affects criminal justice outcomes and

public safety. For the last 15 years, CPI has been the predominant match statistic reported for mixtures. The inclusion method subjectively compares evidence data features with a reference genotype, relying on CPI to provide objective statistical support. However, if CPI does not accurately convey identification information, its patina of science lends no meaningful support.

The National Academy of Sciences report on "Strengthening Forensic Science" provides a way forward.[1] Valid forensic results must be based on solid science: not just the biological data, but also the statistical evaluation of that data. Scientific relevance is demonstrated through empirical validation studies that establish accuracy and reliability. Rigorous studies have not been conducted for CPI and other statistical DNA mixture methods, yet courts need such validation assessments. The match statistics used to report on DNA mixture evidence should be subject to the same scientific scrutiny as unproven methods in other forensic disciplines.

Acknowledgments
The author would like to thank Michael Gorin, Jennifer Hornyak, Matthew Legler, and two anonymous reviewers for helpful comments that improved the manuscript. William Allan gathered the criminal justice outcome data for the 72 reported Virginia cases.

Financial Support and Sponsorship
Nil.

Conflict of Interest
The author is an officer, employee, and shareholder in Cybergenetics, the company that develops TrueAllele Casework technology. He holds related patents in the United States (5,541,067, 5,580,728, 5,876,933, 6,054,268, 6,750,011, 6,807,490, and 8,898,021) and Europe (1,229,135).

REFERENCES

1. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: National Academies Press; 2009.
2. Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® casework validation study. J Forensic Sci 2013;58:1458-66.
3. Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. TrueAllele® Casework on Virginia DNA mixture evidence: Computer and manual interpretation in 72 reported criminal cases. PLoS ONE 2014;9:e92837.
4. Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS ONE 2009;4:e8327.
5. Perlin MW, Hornyak JM, Sugimoto G, Miller KW. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. J Forensic Sci 2015;60:857-68.
6. Scientific Working Group on DNA Analysis Methods (SWGDAM). Short Tandem Repeat (STR) interpretation guidelines. Forensic Science Communications; 2000;2.
7. Curran JM. A MCMC method for resolving two person mixtures. Sci Justice 2008;48:168-77.
8. Perlin MW, Szabady B. Linear mixture analysis: A mathematical approach to resolving mixed DNA samples. J Forensic Sci 2001;46:1372-7.
9. Gelfand AE, Smith AF. Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 1990;85:398-409.
10. Stolovitzky G, Cecchi G. Efficiency of DNA replication in the polymerase chain reaction. Proc Natl Acad Sci U S A 1996;93:12947-52.
11. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21:1087-92.
12. O'Hagan A, Forster J. Bayesian Inference. 2nd ed. New York: John Wiley & Sons; 2004.
13. Billingsley P. Probability and Measure. 3rd ed. New York: John Wiley & Sons; 1995.
14. Essen-Möller E. The evidential value of similarity as proof of paternity, fundamental principles. Transactions of the Anthropological Society in Vienna 1938;68:9-53.
15. Feller W. An Introduction to Probability Theory and Its Applications. 3rd ed. New York: John Wiley & Sons; 1968.
16. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. J Forensic Sci 2011;56:1430-47.
17. Good IJ. Probability and the Weighing of Evidence. London: Griffin; 1950.
18. 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;53:103-14.
19. Greenspoon SA, Schiermeier-Wood L, Jenkins BC. Establishing the limits of TrueAllele® Casework: A validation study. J Forensic Sci 2015;60:1263-76.
20. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Sci Justice 2011;51:204-8.
21. Curran JM, Buckleton J. Inclusion probabilities and dropout. J Forensic Sci 2010;55:1171-3.
22. Perlin MW. Inclusion Probability is a Likelihood Ratio: Implications for DNA Mixtures (poster #85). Promega's Twenty First International Symposium on Human Identification; San Antonio, TX; 2010.
23. National Research Council. Evaluation of Forensic DNA Evidence: Update on Evaluating DNA Evidence. Washington, DC: National Academies Press; 1996.
24. Scientific Working Group on DNA Analysis Methods (SWGDAM). Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. Available from: http://www.fbi.gov/about-us/lab/codis/swgdam-interpretation-guidelines2010. [Last accessed on 2015 Sep 29].
25. Smith C. Dying for Love: The True Story of a Millionaire Dentist, his Unfaithful Wife, and the Affair that Ended in Murder. New York: St. Martin's True Crime; 2011.
26. 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.
27. Noam EM. Jury Trial Vs. Guilty Plea: A Prosecutor's Cost-benefit Comparison. New York: Columbia University; 1980;318A.
28. Lindley DV. Understanding Uncertainty. Hoboken, NJ: John Wiley & Sons; 2006.
29. Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc 1967;62:399-402.
30. Hares DR. Expanding the CODIS core loci in the United States. Forensic Sci Int Genet 2012;6:e52-4.
31. In: Department of Forensic Science, editor. Biology Program Manager. CODIS Operating Policies and Procedures Manual. Richmond, VA: Commonwealth of Virginia; 2015. p. 1-39.
32. McCollister KE, French MT, Fang H. The cost of crime to society: New crime-specific estimates for policy and program evaluation. Drug Alcohol Depend 2010;108:98-109.
33. Durose MR, Cooper AD, Snyder HN. In: Department of Justice, editor. Recidivism of prisoners released in 30 States in 2005: patterns from 2005 to 2010. Washington, DC: Office of Justice Programs, Bureau of Justice Statistics; 2014. p. 1-31.
34. Federal Bureau of Investigation. Crime in the United States. In: Division CJIS, editor. Uniform Crime Reports. Washington, DC: United States Department of Justice; 2013.
35. Koehler JJ. When are people persuaded by DNA match statistics? Law Hum Behav 2001;25:493-513.