J Forensic Sci, November 2006, Vol. 51, No. 6
doi:10.1111/j.1556-4029.2006.00268.x
Available online at: www.blackwell-synergy.com

Tsewei Wang,[1] *Ph.D.; Ning Xue,*[1] *M.Sc.; and J. Douglas Birdwell,*[2] *Ph.D.*

Least-Square Deconvolution: A Framework for Interpreting Short Tandem Repeat Mixtures*

ABSTRACT: Interpreting mixture short tandem repeat DNA data is often a laborious process, involving trying different genotype combinations mixed at assumed DNA mass proportions, and assessing whether the resultant is supported well by the relative peak-height information of the mixture sample. If a clear pattern of major–minor alleles is apparent, it is feasible to identify the major alleles of each locus and form a composite genotype profile for the major contributor. When alleles are shared between the two contributors, and/or heterozygous peak imbalance is present, it becomes complex and difficult to deduce the profile of the minor contributor. The manual trial and error procedures performed by an analyst in the attempt to resolve mixture samples have been formalized in the least-square deconvolution (LSD) framework reported here for two-person mixtures, with the allele peak height (or area) information as its only input. LSD operates on the peak-data information of each locus separately, independent of all other loci, and finds the best-fit DNA mass proportions and calculates error residual for each possible genotype combination. The LSD mathematical result for all loci is then to be reviewed by a DNA analyst, who will apply a set of heuristic interpretation guidelines in an attempt to form a composite DNA profile for each of the two contributors. Both simulated and forensic peak-height data were used to support this approach. A set of heuristic guidelines is to be used in forming a composite profile for each of the mixture contributors in analyzing the mathematical results of LSD. The heuristic rules involve the checking of consistency of the best-fit mass proportion ratios for the top-ranked genotype combination case among all four- and three-allele loci, and involve assessing the degree of fit of the top-ranked case relative to the fit of the second-ranked case. A different set of guidelines is used in reviewing and analyzing the LSD mathematical results for two-allele loci. Resolution of two-allele loci is performed with less confidence than for four- and three-allele loci. This paper gives a detailed description of the theory of the LSD methodology, discusses its limitations, and the heuristic guidelines in analyzing the LSD mathematical results. A 13-loci sample case study is included. The use of the interpretation guidelines in forming composite profiles for each of the two contributors is illustrated. Application of LSD in this case produced correct resolutions at all loci. Information on obtaining access to the LSD software is also given in the paper.

KEYWORDS: forensic science, STR/DNA allele peak data, mixture DNA, least-square fit, forensic DNA, mixture resolution

For a variety of crimes, DNA analysis of the body fluids or other trace evidence left behind by either the perpetrator, or the perpetrator mixed with that of the victim or both, turns out to be the most effective in providing a clue to the solution of the crime. Often, resolution of a mixture sample is complicated by a lack of a clear major–minor peak pattern, allele degradation at selected loci, imbalanced heterozygote peak amplification, or variation of DNA mass proportions in the mixture from locus to locus. Earlier effort in mixture resolution was based on calculating the likelihood ratios of the various genotype combinations of alleles at each locus and drawing conclusions based on the comparisons of various likelihood ratios (1–3). Quantitative peak information, such as peak height or area, is often available from DNA sequencing instrumentation that is used to read and analyze a DNA electropherogram. Such data have triggered the development of heuristic methods as well as computer software in an attempt to resolve short tandem repeat (STR) mixtures. The group at Forensic Science Services (FSS) and their colleagues has published a series of articles reporting on interpretation of STR mixtures based on quantitative allele peak data guided by a series of logical steps (4–7). Clayton et al. (7) advocated a series of six logical steps for a DNA analyst to follow in interpreting mixture STR profiles. These steps are well laid out and are based on observations and experiences obtained from laboratory experiments and casework. Recently, three computer software programs for STR mixture resolution have been reported. Perlin and Szabady (8) presented the linear mixture analysis method (LMA) showing the application of LMA to several forensic mixture scenarios, depending on whether the genotype of one, both, or either contributor to the mixture is known. Mortera et al. (9) presented a probabilistic expert system in conjunction with quantitative peak data to resolve DNA mixture. More recently, FSS reported on the availability of PENDULUM (marketed as *i*-Stream, part of the i^3 package by the FSS of U.K.)—a guideline-based approach to the interpretation of STR mixtures, to the forensic community (10), which is based on the steps advocated in Clayton et al. (7) in interpreting mixture profiles. Availability of computer software to carry out systematic mathematical analysis using quantitative peak data is expected to reduce greatly the time required for mixture analysis, improve the consistency among mixture interpretations, and to yield less conservative mixture resolution results.

We report here an interpretation framework guided by the least-square analysis results of the quantitative peak data of either peak area or peak height. In this paper, we use the term "peak height" to refer to the quantitative peak data. Peak area can also be used with this approach. Consistency in using either height or area in interpreting a mixture sample is recommended. This framework is referred to as the "Least-square Deconvolution (LSD)." Reference (11) contains an earlier presentation of this work. LSD differs from the other quantitative interpretation approaches reported earlier, such as the LMA and PENDULUM in the mathematical

[1]Department of Chemical Engineering and Laboratory for Information Technologies, The University of Tennessee, Knoxville, TN 37996-2200.

[2]Department of Electrical and Computer Engineering & Laboratory for Information Technologies, The University of Tennessee, Knoxville, TN 37996-2100.

*Results have been presented at the 13th International Symposium on Human Identification, Oct. 7–10, 2002, Phoenix, AZ, under the title of 'Least-square deconvolution (LSD): a new way of resolving STR/DNA mixture samples.

Received 5 Sept. 2005; and in revised form 6 June 2006; accepted 25 June 2006; published 8 Nov. 2006.

formulation of the least-square problems, and in the heuristic guidelines used in forming the most compatible genotype for the individual contributors of the mixture.

In this paper, the LSD mathematical theory, methodology, as well as the heuristic guidelines for interpreting the raw LSD mathematical results are first presented, followed by applying LSD to a four-, three-, and two-peak locus, respectively. A 13-loci sample is then used to illustrate the steps of LSD and the application of the guidelines in interpreting the raw LSD mathematical results. This is then followed by a brief discussion of the main difference between LSD and the other known quantitative methods. In the context of this paper, the term "peak height" is used to represent the peak data. In practice, either peak height or area can be used with the LSD framework. The key is in its consistent use when applying the LSD framework. Additional examples of LSD applications to other forensic samples are available in the document entitled "LSD Interpretation Guidelines" at https://lsd.lit.net/helpfiles.

LSD Software Availability

Organizations that wish to find out about accessing, via the web, our LSD software may check in https://lsd.lit.net for information (note the "s" in "https").

Requirement for Using LSD

Before LSD can be applied, it is important that DNA analysts already have made proper allele calls to exclude artifact peaks, such as stutter and pull-up peaks. At the present time, LSD does not contain an artificial intelligence element to make decisions on proper allele calls. It acts on allele peak data fed to it for each locus and returns the best-fit mass proportions for each possible genotype combination. It is also important that the quantitative allele peak data information exhibit no peak saturation (resulting from overloading of DNA samples onto the DNA sequencing instrument). Otherwise, the proportionality between peak height and mass proportion is compromised, and the resultant least-square fit would not reflect the appropriate estimation of the true underlying mass proportions. No known contributor profile needs to be given to LSD for processing.

An Overview of the LSD Algorithm

The common practice among forensic DNA analysts faced with a mixture sample of two contributors is to first analyze four-peak loci to see whether a clear major–minor separation of peaks is evident. If it is, the relative peak-height ratio can be used to arrive at approximate DNA mass proportions of the two DNA amount present in the mixture DNA template before amplification (7). Using this approximate mass proportion, the three- and two-peak loci are then examined in an attempt to separate out the genotypes of the two contributors. During the examination of the three- and two-peak loci, if the analyst is not using a formal quantitative procedure, then using the approximate mass proportions derived from four-allele loci, the analyst would go through a series of trial-and-error attempts to fit each of the possible genotype combinations for the two contributors to the given relative peak heights observed for these loci. The mathematical steps in LSD formalize these trial-and-error procedures at each locus, by using the quantitative peak data to find the most compatible genotype combination and the corresponding best-fit mass proportions. It also gives measures that can be used to assess the quality of fits.

LSD analyzes and processes peak data of each locus separately and independently, allowing a different best-fit mass proportion vector to be developed through least-square optimization for each locus. For each locus, all possible genotype combinations are first enumerated and represented by a gene matrix. For each possible gene matrix, the mass proportion vector that best fits the given peak data is calculated using the normalized relative peak-height vector through the least-square optimization method. This step is then followed by the calculation of the corresponding error residual for each possible combination. The top-ranked combination with the smallest residual is regarded to be the genotype combination best supported by the observed peak data. Analysis proceeds to the next locus using exactly the same set of procedures as those for the first locus. After all loci have been processed in this manner, the mathematical results are reviewed by a DNA analyst, who would apply a set of heuristic LSD interpretation guidelines in the analysis of the top-ranked combination cases at all loci for an assessment of the quality and confidence of resolution at each locus, and for putting together a composite genotype profile for each contributor. In particular, the LSD mathematical results for all two-allele loci must be analyzed by an analyst to select subjectively the most compatible genotype combination, following the LSD interpretation guidelines. At some loci, due to compromised peak data, more than one genotype combination resolution is to be made. If no known contributor profile to the mixture sample is available, the developed composite profiles are then to be examined by an analyst to identify and determine the number of loci at which the LSD-suggested resolution results are judged to be sufficiently reliable, and the identity and number of loci at which multiple resolution possibilities exist. One can follow-up with searches in a DNA database, such as Combined DNA Index System (CODIS), for each of the possible resolved composite profiles to see whether a match is found. One can also compare LSD-derived composite profiles against the DNA profiles of suspects in custody for matches.

If a known contributor profile is available, such as that of the victim in a rape kit mixture sample, it is to be brought in post-LSD to check, locus by locus, the LSD-suggested resolution results. For those loci in which one of the contributors' genotypes consistently agrees with that of the known contributor, confidence is increased that the suggested resolution is correct. For loci with more than one possible resolution, the known contributor profile may help in the elimination of the incompatible ones. Details of the LSD formulation and solution steps will now be described.

The LSD Method

LSD Mathematical Formulation

Problem Formulation—Two assumptions underlying the LSD approach are that (1) the multiple alleles within a locus are co-amplified to roughly the same degree during the polymerase chain reaction step; and (2) the allele peaks add when the two contributors to the mixture have a common peak present at a locus (12,13). Thus, ideally, the allele peak height should be proportional to how much DNA of each allele is present in the original mixture, and when an allele is shared between two contributors, the peak heights add to give the combined height. Therefore, knowing the relative peak heights of the alleles present at a locus, for a given postulated possible genotype combination, the most compatible mass proportion vector can be arrived at through least-square's calculations.

Throughout this paper, the amount of DNA of each component present in the mixture is referred to as the "DNA mass" for that component, which does not refer to the molecular weight of the DNA component under consideration. The terms "mass proportion" and "mass proportion ratio" then refer to the amount of DNA of one contributor in proportion to that of the other (unnormalized and normalized, respectively). For instance, if the calculated relative mass proportions are 1.5 and 3.0 for the two contributors, then it means that relatively speaking one contributor has 1.5 parts of DNA and the other contributor has 3.0 parts of DNA in the mixture. If the proportions are normalized against the smaller proportion (dividing the two proportions against the smaller one), then the "mass proportion ratio" would be 1:2 (1.5:3.0 → 1:2).

The first step in LSD is, for each locus, to list all possible genotype combinations of the alleles for the two contributors, and to represent each combination by a gene matrix, composed of "0," "1," or "2" as its elements representing the number of copies of the corresponding allele at that locus. Tables 1–3 list all possible genotype combination cases for a four-, three-, and a two-allele locus, along with the associated gene matrix. The columns of the gene matrix represent the allele presence pattern of the two contributors. For instance, at a three-allele locus, if person 1 has alleles, "A," and "B," and person 2 is homozygous in "C," then the corresponding gene matrix is

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 2 \end{bmatrix}$$

where the first column of A indicates that person 1 has one copy of each of alleles A and B, and none of C, whereas person 2 (column 2) has two copies of C only. Let the x vector designate the mass proportion of the two contributors' DNA in the mixture. When mixed in these proportions, the resulting peak height for this example should have a relative ratio of $x_1{:}x_1{:}2x_2$, as shown below

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_1 \\ 2x_2 \end{bmatrix} \propto \begin{bmatrix} peak1 \\ peak2 \\ peak3 \end{bmatrix} \quad (1)$$

or in symbolic representation

$$Ax = b \quad (2)$$

where A denotes the gene matrix, $\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 2 \end{bmatrix}$ for instance, x the mass proportion vector, and b the relative allele peak-height vector. At each locus, depending on the measured allele peak-height data, and the underlying relative mass proportion, only one of the possible genotype combinations, when combined at the optimum mass proportion, should yield a peak-height vector that comes close to the given peak-height measurement vector.

TABLE 1—*Three possible genotype combination cases for a four-allele locus.**

Case	Genotype Combination: Person 1	Genotype Combination: Person 2	Matrix (A)	Pseudoinverse of the Matrix (A^+)
1	A, B	C, D	$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$
2	A, C	B, D	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$
3	A, D	B, C	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$

*A B C D denote the names of the four alleles. Matrix A denotes the genotype combination matrix, where a 1 indicates the presence, and 0 indicates the absence of the allele it represents. A^+ denotes the pseudoinverse of A, and is used to calculate directly the least-squares solution of the best-fit DNA mass proportion coefficients.

TABLE 2—*Six possible genotype combination cases for a three-allele locus.**

Case	Genotype Combination: Person 1	Genotype Combination: Person 2	Matrix (A)	Pseudoinverse of the Matrix (A^+)
1	A, A	B, C	$\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}$
2	B, B	A, C	$\begin{bmatrix} 0 & 1 \\ 2 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$
3	C, C	A, B	$\begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 2 & 0 \end{bmatrix}$	$\frac{1}{2}\begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$
4	A, B	B, C	$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{3}\begin{bmatrix} 2 & 1 & -1 \\ -1 & 1 & 2 \end{bmatrix}$
5	A, B	A, C	$\begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\frac{1}{3}\begin{bmatrix} 1 & 2 & -1 \\ 1 & -1 & 2 \end{bmatrix}$
6	A, C	B, C	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$	$\frac{1}{3}\begin{bmatrix} 2 & -1 & 1 \\ -1 & 2 & 1 \end{bmatrix}$

*A B C denote the names of the three alleles. Matrix A denotes the genotype combination matrix, where a 1 indicates the presence, and 0 indicates the absence of the allele it represents. A^+ denotes the pseudoinverse of A, and is used to calculate directly the least-squares solution of the best-fit DNA mass proportion coefficients.

TABLE 3—*Four possible genotype combination cases for a two-allele locus.**

Case	Genotype Combination		Matrix (A)	Pseudoinverse of the Matrix (A^+)
	Person 1	Person 2		
1	A, A	B, B	$\begin{bmatrix}2 & 0\\0 & 2\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$
2	A, B	A, A	$\begin{bmatrix}1 & 2\\1 & 0\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}0 & 2\\1 & -1\end{bmatrix}$
3	A, B	B, B	$\begin{bmatrix}1 & 0\\1 & 2\end{bmatrix}$	$\frac{1}{2}\begin{bmatrix}2 & 0\\-1 & 1\end{bmatrix}$
4	A, B	A, B	$\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$	$\frac{1}{4}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$

*A B denotes the names of the two alleles. Matrix A denotes the genotype combination matrix, where a 1 indicates the presence, and 0 indicates the absence of the allele it represents. A^+ denotes the pseudoinverse of A, and is used to calculate directly the least-squares solution of the best-fit DNA mass proportion coefficients.

LSD Mathematical Solution—The least-square problem to be solved is the following. At each locus, the relative peak-height vector is known and the gene matrix for each possible genotype combinations is also known, corresponding to knowing the matrix, *A,* and the vector, *b*, in Eq. (2). The best-fit mass proportion vector, *x*, can be computed directly by using the pseudoinverse of the *A* matrix (14–16), and thus bypass any iterative search for the optimum *x*. The pseudoinverse approach is shown in the following equation:

$$x_{\mathrm{ls}} = A^{+}b \tag{3}$$

where x_{ls} denotes the least-square solution for *x*, and A^+ denotes the pseudoinverse of the *A* matrix. The pseudoinverse of a matrix always exists for any type of matrix *A* of any dimension regardless of whether their columns or rows are all independent of each other. When the columns (rows) of *A* are all independent, the pseudoinverse is the left (the right) inverse of *A*. In this case, the product of A^+A (AA^+) gives an identity matrix, of size equal to the minimum of (m, n) where (m, n) denote the number of rows and columns of *A*, respectively. The most reliable way of calculating the pseudoinverse of a matrix is by using the singular value decomposition (SVD) of *A* (14–16). However, for our application purposes, the relevant *A* matrices all have small dimensions, with the largest dimension being 4×2 (four alleles and two people), and the pseudoinverse of these matrices can be computed using matrix computation software such as Matlab® (The Mathworks, Natick, MA) and stored in a look-up table of the LSD software for use. They can also be hard coded into any software implementing the LSD algorithm. For reference, pseudoinverse matrices of the various gene matrices are also included in Tables 1–3.

Calculation of the Mass Proportion Ratio—The x_{ls} vector from the least-square solution contains the two relative mass proportion values that best fit the given allele peak height for the genotype combination matrix, *A*, under consideration. The mass proportion ratio is calculated by dividing the larger-valued mass proportion element by the smaller-valued mass proportion element.

Assessment of the Least-Square Fit—Calculation of the Error Residuals—In order to assess how well the best-fit genotype combination fits the observed allele peak data, the residual representing the associated fitting error between the predicted *b* peak vector, calculated from $A \times x_{\mathrm{ls}}$ and the given *b* peak-height vector, is calculated, The error residual is calculated as the square of the Euclidian length of the error vector, and can be treated as a measure of the lack of fit of the measured allele peak data by the proposed genotype combination case mixed at the best-fit mass proportions. The smaller the error, the better the fit.

Similarly, for each of the other possible gene matrices for this locus, the associated error residual is to be computed from its respective least-square solution. After the least-square fitting error residuals are calculated for all the possible genotype combinations for this locus, the error residuals are sorted from the smallest to the largest. The corresponding genotype combination cases are thus ranked according to their associated fitting error residuals, from the smallest to the largest. The genotype combination case with the smallest fitting error residual is the one best supported by the given peak data. The farther apart the fitting errors of the top two-ranked cases, the more confident it is in concluding that the top-ranked genotype combination is the most supported genotype for the contributors conditioned on the observed peak-height data for that locus. Notice that this is not the same as saying that the individuals with the genotypes as indicated by the top-ranked genotype combination are the most probable contributors to the mixture at this locus. The residual should in no way be directly connected to probability considerations. It merely gives an indication of how supported the indicated genotype combination is, given that the peaks are observed to have the measured relative peak heights. It is not to be interpreted as a measure of the probability that the corresponding genotype combination is the correct one.

Work now proceeds to the next locus. Exactly the same series of steps are performed for peak data at locus 2 until all loci have been processed. The next major step is to analyze the LSD mathematical results for all loci as a whole and attempt to form a composite genotype resolution profile for each of the two contributors. This is done by using a set of interpretation guidelines, followed by comparing LSD suggested results with that of a known contributor profile if one is available.

Computer Implementation of the LSD Mathematics

Software has been developed to implement the mathematical steps of LSD. The computational modules of the LSD software are implemented using the C++ language. Access to LSD is through the Internet, via secure connections (https://lsd.lit.net; a user account is required). The software is executed as back-end processes to a secure Apache web server (http://www.apache.org/), and uses a MySQL (http://www.mysql.com/) database engine to store profile data and results. The software runs on two mirrored Linux servers, which are protected by a firewall from Internet threats, and all traffic to and from the servers, except for the web pages that provide an introduction and overview of LSD, is encrypted. User authentication is by password, and security mechanisms are used to ensure that each user's Internet traffic originates from a known organization. LSD results shown in this paper are obtained from the version of LSD software that is currently described at the www.lit.net website and accessed via https://lsd.lit.net (a user account is required).

Overview of LSD Mathematical Result Interpretation Guidelines

Two heuristic criteria are used in forming the most compatible LSD-guided composite profiles for the two contributors. One is consistent mass proportion ratio for the chosen genotype combination for all loci, and the second is the relative small error residual for the chosen combination. The top-ranked combination at each locus, by default, is nominally the most compatible genotype combination, given the observed relative allele peak height. However, the second-ranked combination (sometimes the third ranked also) cannot be ruled out if it fits comparably and has a mass proportion ratio consistent with those calculated for the other loci. In this case, two combinations for this locus need to be retained as supported and one may be eliminated only after comparing with the known genotype if one is available. These two criteria apply to all four- and three-allele loci.

When interpreting LSD mathematical results for two-allele loci, a different set of heuristic guidelines is to be used. Because three of the four two-allele gene matrices (see Table 3) are square and have independent rows (and columns), there are two independent equations with two unknowns. An exact mass proportion vector solution exists, resulting in no fitting error. However, this does not mean that one of these three genotype combinations is necessarily the correct resolution. The fourth genotype combination is for the two contributors to be both heterozygous with alleles AB and AB. The associated gene matrix is square but only one of the two rows is independent, indicating that there are two unknowns but only one independent equation to bind them. The least-square solution for this particular matrix would always be [1 1]' mathematically, implying equal mass proportion in the mixture, which may not be the case at all. This is a limitation of the mathematics of least-square solution for square matrices when the matrix does not have fully independent rows. A clue to which genotype combination is most applicable can be obtained by examining the relative peak heights of the two given peaks. If both contributors are heterozygous with alleles [AB], then no matter what the mass proportions are, the peak height for A and B should always be comparable, because there is equal amount present for A and B DNA in the mix. A guideline for judging whether peak heights are comparable for heterozygous alleles may be the 60% rule (12,17). Owing to different allele size, degree of possible degradation, and primer binding site mutation, two heterozygous alleles are usually not amplified equally (7,13,17,18). Clayton et al. (7) and Bill et al. (10), suggested 60–167% (inverse of 60%) to be the acceptable region. This heterozygote balance threshold can be chosen by the user based on the user's laboratory experience for the polymerase chain reaction amplification step. The LSD interpretation guidelines adopt this range from (10) as a fuzzy boundary range to assess whether the two peak heights are considered comparable. If they are, then the genotype combination of [AB] and [AB] cannot be ruled out based on the given allele peak data. If the two peak heights are not comparable, then one of the other genotype combinations with a mass proportion ratio consistent with that for other loci should be chosen to be the most compatible one.

The LSD Framework

The main steps of the LSD framework are presented below. Note that steps 4–6 and 11 describe the calculation of error and residual values and the role they play in the ranking of the possible genotype combinations at each locus. As a reminder, please note that the term "DNA mass proportion" refers to the estimated relative amount of DNA present in the mixture after amplification from one contributor with respect to the amount of DNA from the other contributor. The values for the two mass proportions come from the least-square solution (Eq. [3]).

Summary Steps of LSD

The steps of the LSD framework are summarized as follows:

1. Take the first locus. Normalize (divide) the allele peak heights to the smallest peak height, and put the normalized heights in a column vector.
2. Take each possible genotype combination for this locus, and compute the best-fit mass proportion vector using Eq. (3).
3. Calculate the fitting error vector for each possible combination.
4. Calculate the error residual (sum of the squares of the entries of the error vector from step 3) for each possible combination.
5. Rank the possible genotype combinations according to their residuals, from the smallest to the largest.
6. Calculate the ratio of the error residual of each possible combination case to the residual of the top-ranked combination case.
7. Calculate the mass proportion ratio of each genotype combination case by dividing the larger mass proportion element by the smaller mass proportion element of the mass proportion vector from step 2.
8. Take the next locus and apply steps 1–7. Continue until all loci have been processed.
9. A DNA analyst applies a set of heuristic interpretation guidelines, briefly described in the remaining steps, to the LSD mathematical results in an attempt to form a composite resolution profile for each contributor.

The heuristic interpretation guidelines are as follows:

10. Review the LSD rankings of the possible genotype combinations at all four- and three-allele loci first to assess the consistency in the mass proportion ratio of the top-ranked genotype combination case, and to assess the degree of fit by the second-ranked combination case compared with the top-ranked case. By default, the top-ranked genotype combination case is the most supported one for that locus, conditioned on the given allele peak data. Flag those loci with inconsistent mass proportion ratio for further examination. For each locus, retain the second-ranked combination case as a supported resolution candidate also, if its residual is close to that of the top-ranked case. If by chance, the third or higher ranked combination case also has comparable residual as that of the top ranked, then they are also retained.
11. Review the LSD results of two-allele loci next. For each locus, calculate the ratio of the two original peak heights. If it falls in the range of 0.6–1.67, then combination case 4 is a candidate for resolution supported by the peak data. Then, examine the best-fit mass proportion ratio for the first three genotype combination cases of each locus to decide, subjectively, whether one of these is "close" to those for four- and three-allele loci. If one is, then it is also a supported resolution candidate for the two contributors of this locus conditioned on the peak data.

12. Compile the analysis from steps 10 and 11 and form a composite individual profile for each of the two contributors. Flag those loci for which resolution is not confident.
13. If a known contributor profile is available, bring it in to compare against the profiles formed in step 12. Flag those loci where results do not agree with the known contributor's profile. Delete these loci from the final resolution results.

Examples of applying the LSD framework to four-, three-, and two-allele loci (the mathematical results and their interpretations) are given in Appendix A. Note especially the guidelines in interpreting the LSD mathematical results of two-allele loci.

Heuristic Guidelines for Forming Tentative Composite Profiles for the Two Contributors

The entire LSD interpretation framework encompasses not only the mathematical methodology to compute the least-square solution for each possible genotype combination at each locus but also the follow-up interpretation of the LSD computational results, using the guidelines described here.

Reviewing LSD Mathematical Results for all Four- and Three-Allele Loci First—The LSD mathematical results for all four- and three-allele loci are examined first to analyze the top-ranked genotype combinations and the associated information. The mass proportion ratios for the top-ranked combination for these loci are examined for consistency. The measure for consistency is to be judged subjectively by the analyst. Inconsistent mass proportion ratio across loci usually arise from unequal DNA amplification due to a variety of reasons such as a marked difference in the size of the two alleles (7), mutation in the primer binding site (7), allele degradation at some loci (18), and additive effect of the stutter (7,17). Clayton et al. (7) suggest in step 4 of mixture profile analysis the estimation of the mixture proportion first from four-allele loci peak data information. A manually estimated mixture proportion based on the relative peak heights of these loci should correspond closely to the LSD-calculated mixture proportion ratio for the top-ranked combination case of the corresponding loci. The latter is arrived at by the least-square fit using the peak data information. For any locus that has a much higher or lower mass proportion ratio for the top-ranked combination than that of the majority of the others in this group, one needs to review the associated peak data for that outlier locus to see whether a severe peak imbalance is evident between the major and minor alleles, much more severe than those observed at other loci. If a severe imbalance is evident, then the analyst should examine the peak data associated with this locus to assess if an called allele peak is a false call, such as a stutter peak that should have been deleted from the final allele calls for this locus but was retained by mistake, or the low peak can be from a degraded allele, or as a result of a primer binding site mutation for that allele, both of which will result in reduced DNA amplification (7,17). It is also known that larger alleles have a lower amplification efficiency and tend to stutter more (7,17).

At each locus, the confidence in the resolution given by the top-ranked combination can be assessed by comparing the relative error residuals of the top two-ranked combinations. The further apart they are, the more confident it is that the top-ranked combination is the best supported resolution, given the observed allele peak data. If the error for the second-ranked case is comparable to that of the top-ranked case, AND the mass proportion ratio for the two cases are also comparable, then the top two cases have to be retained for consideration at this point. The notion of comparableness is a subjective determination, at the discretion of the analyst. Bill et al. (10) report that a variation in the estimated mixture proportion M_x from the overall profile mean across all loci for hundreds of possible profiles with close overall fitting residuals can be as high as $\pm$ 0.35. They define M_x as the ratio of the DNA mass of one contributor to that of the total. (This is different from the definition of the mass proportion ratio used in LSD, which refers to the ratio of the larger DNA mass proportion to that of the smaller DNA mass proportion in the mixture. M_x and mass proportion ratio can be calculated from each other.) A change in M_x of this magnitude corresponds to a change in mass proportion ratio of several fold. Table 4 shows an example in which the top two genotype combination cases appear to fit comparably well (1.0 and 2.8 error ratios) and the best-fit mass proportion ratios are also comparable: 1:2.7 and 1:1.6. When the known contributor profile (if available) is brought in for comparison later, one candidate combination can usually be eliminated because neither genotype in the combination would correspond to that of the known genotype at this locus. However, it is recommended that if the top two combinations have comparable fitting error residuals, the combination with a mass proportion ratio more consistent with those at other loci is preferred. Table 5 shows another example where the mass proportion ratios for the top two combination cases are comparable, but the two-error residuals differ by a thousand fold, indicating that the second case is not at all supported by the given measured peak data.

TABLE 4—*LSD results for a three-allele locus showing that the top two-ranked cases are both supported by the given peak area data.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
CSF1PO—three alleles {alleles 10 11 12} {peak area 482 697 617}						
	1	10, 10	11, 12	1.4e − 02	1.0	1.0:2.7
	2	11, 11	10, 12	3.9e − 02	2.8	1.0:1.6
	3	12, 12	10, 11	9.9e − 02	7.2	1.0:1.9
	4	10, 11	11, 12	0.23	17	1.0:1.4
	5	10, 12	11, 12	0.45	33	1.0:1.7
	6	10, 12	10, 11	0.99	72	1.0:1.2

*Mass ratio refers to the mass proportion ratio. Note that the fitting error refers to the fitting residual (sum of squares of the elements of the error vector); the error ratio refers to the ratio of the fitting error of each genotype combination case to the fitting error of the top-ranked combination case; the mass ratio refers to the ratio of the two mass proportions that are given by the least-square solution.

TABLE 5—*LSD results for a second three-allele locus.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
TH01—three alleles {alleles 5 6 8} {peak area 944 935 633}						
	1	8, 8	5, 6	1.0e − 04	1.0	1.0:3.0
	2	5, 5	6, 8	0.11	1.1e+03	1.0:1.7
	3	6, 6	5, 8	0.12	1.2e+03	1.0:1.7
	4	5, 8	5, 6	0.32	3.2e+03	1.0:1.7
	5	6, 8	5, 6	0.34	3.4e+03	1.0:1.7
	6	6, 8	5, 8	1.29	1.3e+04	1.0:1.0

*Mass ratio refers to the mass proportion ratio. Even though the mass ratios for the two top-ranked combination cases are comparable, the top-ranked case is much better supported by the given peak area data, . Note that the fitting error refers to the fitting residual (sum of squares of the elements of the error vector); the error ratio refers to the ratio of the fitting error of each genotype combination case to the fitting error of the top-ranked combination case; the mass ratio refers to the ratio of the two mass proportions that are given by the least-square solution.

TABLE 6—*LSD results for the first of four two-allele loci all from the same mixture sample as that in Tables 6–9.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D18S51—two alleles {alleles 14 15} {peak area 209 276}						
	1	14, 14	15, 15	0.0e+00		1.0:1.3
	2	14, 14	14, 15	0.0e+00		1.0: − 8.2
	3	15, 15	14, 15	0.0e+00		1.0:6.2
	4	14, 15	14, 15	5.1e − 02		1.0:1.0

*Mass ratio refers to the mass proportion ratio. The allele peak areas of the two alleles are considered comparable (within the 0.6–1.67 range). As a result, case 4 is a compatible combination. Note that the first three combination cases have 0 fitting error associated with them, due to the fact that the gene matrix for each of these has two independent rows with two unknowns; thus, an exact solution exists. The mass proportion ratio for the second case is negative, thus rendering this case infeasible.

TABLE 8—*LSD results for the third of four two-allele loci all from the same mixture sample as that in Tables 6–9.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
VWA—two alleles {alleles 16 17} {peak area 1749 362}						
	1	17, 17	16, 16	0.0e+00		1.0:4.8
	2	16, 17	16, 16	0.0e+00		1.0:1.9
	3	17, 17	16, 17	0.0e+00		1.0: − 2.5
	4	16, 17	16, 17	7.34		1.0:1.0

*Mass ratio refers to the mass proportion ratio. The allele peak areas of the two alleles are not comparable at all (well outside the 0.6–1.67 range); thus, case 4 can be eliminated confidently from further consideration. Note that the first three combination cases have 0 fitting error associated with them, due to the fact that the gene matrix for each of these has two independent rows with two unknowns; thus an exact solution exists. The mass proportion ratio for the third case is negative, thus rendering this case infeasible.

The next step in the guideline for reviewing is to collect the mass proportion ratio of the top-ranked genotype combination for these loci and decide what mass proportion ratio range is considered to be consistent across them. This range will be used to evaluate the LSD mathematical results for two-allele loci in the subjective selection of the most compatible genotype combination. The guidelines for which are described in the next section,

Reviewing the LSD Mathematical Results for Two-Allele Loci—Tables 6–9 show the LSD mathematical results for four two-allele loci from the same mixture sample. They will be used to illustrate the heuristic guidelines for interpreting two-allele LSD mathematical results. Note that for these four loci, the peak area information is used in LSD. The use of peak area and peak height is interchangeable in using LSD, as long as the use is consistent within a mixture sample.

It is observed from the LSD mathematical results for these four loci that in each of the first three loci, among all four genotype combination cases of each locus, no two mass proportion ratios can be considered to be close to each other. However, in the fourth locus (D5S818), the best-fit mass proportion ratios for cases 1 and 2 are very close: 1:2.1 versus 1:1.9. This implies that if one combination is chosen to be a candidate, then the other is also to be chosen as a candidate. Examining the peak areas of the two alleles of each of the four loci, it is evident that only the first locus (D18S51) has allele areas close to each other: 209 and 276, or 1:1.32, within the 0.6–1.67 range. This implies that for this locus, the fourth genotype combination case of {[A B], [A B]} should be considered as a candidate. If the peak areas of the two alleles are not comparable, then from the first three combinations, the one with the best-fit mass proportion ratio most consistent with those from other loci is to be chosen. If none exists, then the resolution at this locus is deemed inconclusive, and this locus is to be removed from consideration in forming the overall profile.

For reference, the mass proportion ratios from the top-ranked genotype combination case from the other loci in this particular mixture average to 1:2.0 (mass proportion ratios are 1.3, 1.6, 1.8, 1.8, 1.8, 1.9, 2.3, and 3.6). With this information, the most compatible genotype combination case from the LSD mathematical results for each of these four two-allele loci can now be selected.

- Locus D18S51: In addition to genotype combination case 4 being a supported one (due to comparable peak areas) based on the peak data, case 1 cannot be ruled out with confidence at this point because the corresponding mass proportion ratio of 1:1.3 is close to the overall average of 1:2. At this point, cases 1 and 4 are the potential contenders.
- Locus CSF1PO: The allele peak area ratio of [450: 673] or [1:1.5] is near the boundary of the 0.6–1.67 range for acceptable heterozygous peak balance. Hence, case 4 may or may not be applicable. The mass proportion ratio for case 1, 1:1.5, is consistent with the average of 1:2 from the other loci. Case 3 with a mass proportion ratio of 1:4, is still close enough to 1:2 to be considered as borderline compatible at this point also. Therefore, case 1 is considered to be the most supported, with cases 3 and 4 as less supported, conditioned on the given peak data.
- Locus VWA: The allele peak areas of [1749, 362] are well outside the 0.6–1.67 range. case 4 is deleted from consider-

TABLE 7—*LSD results for the second of four two-allele loci all from the same mixture sample as that in Tables 6–9.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
CSF1PO—two alleles {alleles 7 11} {peak area 450 673}						
	1	7, 7	11, 11	0.0e+00		1.0:1.5
	2	7, 7	7, 11	0.0e+00		1.0: − 6.0
	3	11, 11	7, 11	0.0e+00		1.0:4.0
	4	7, 11	7, 11	0.12		1.0:1.0

*Mass ratio refers to the mass proportion ratio. The allele peak areas of the two alleles are considered comparable (within the 0.6–1.67 range). As a result, case 4 is a compatible combination. Note that the first three combination cases have 0 fitting error associated with them, due to the fact that the gene matrix for each of these has two independent rows with two unknowns; thus, an exact solution exists. The mass proportion ratio for the second case is negative, thus rendering this case infeasible.

TABLE 9—*LSD results for the last of four two-allele loci from the same mixture sample as that in Tables 6–9.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D5S818—two alleles {alleles 12 13} {peak area 1552 749}						
	1	13,13	12,12	0.0e+00		1.0:2.1
	2	12,12	12,13	0.0e+00		1.0:1.9
	3	13,13	12,13	0.0e+00		1.0: − 3.9
	4	12,13	12,13	0.57		1.0:1.0

*Mass ratio refers to the mass proportion ratio. The allele peak areas of the two alleles are not comparable (outsdie the 0.6–1.67 range); thus, case 4 can be eliminated confidently from further consideration. Note that the first three combination cases have 0 fitting error associated with them, due to the fact that the gene matrix for each of these has two independent rows with two unknowns; thus, an exact solution exists. The mass proportion ratio for the third case is negative, thus rendering this case infeasible.

ation. Only case 2, with genotype combinations of {[16, 17], [16, 16]} and a mass proportion ratio of 1:1.9, is consistent with the average of 1:2 from the other loci. Therefore, case 2 is the preferred choice at this time.

- Locus D5S818: The allele peak areas of [1552:749] or [1:2.07] are outside the 1.67 bound. Therefore, case 4 is removed from further consideration. Case 1 of {[13, 13], [12, 12]} and case 2 of {[12, 12], [12, 13]} with the fitted mass proportion ratios of 1:2.1 and 1:1.9, respectively, are both candidates to be the compatible genotype combination choice. Both will be retained for consideration. Note that for this locus, in both cases 1 and 2, one of the two people's genotypes is [12, 12], but it is person 2 in case 1 and person 1 in case 2 that is assigned this genotype. When the known contributor profile is brought in (if one is available), one needs to first determine whether it is person 1 or 2 of LSD mathematical results to which the known person corresponds. Then, using the known genotype, the appropriate case may be selected from the possible choices for the locus. The following section gives the analysis of the LSD mathematical results for a 13-loci mixture sample.

Results and Discussion

A 13-Loci Example

Table 10 shows the allele peak data of a mixture sample with 13 loci (sample kindly provided by R. Wickenheiser of the Acadiana Crime Lab, New Iberia, LA), and Tables 11–14 show LSD mathematical results for this sample. There are four four-allele loci, seven three-allele loci, and two two-allele loci. LSD mathematical results for the 11 four- and three-alelle loci will be analyzed first. The set of the best-fit mass proportion ratio for the top-ranked genotype combination for the 11 loci is very consistent: {2.3, 3.0, 2.4, 2.6, 2.2, 2.4, 2.7, 1.8, 2.1, 2.4, 2.0} with an average of 2.4. This implies that if these genotype combinations are indeed the correct resolutions, then the mass proportion was preserved well across these loci during amplification. Next, the confidence regarding which genotype combinations are the most compatible ones is examined. Examining how much worse the second-ranked case fits the given peak data compared with that of the top-ranked case shows that the error ratios for the second-ranked cases for the 11 loci are {3.2, 1126, 4.8, 30, 7.7, 13465, 2.8, 14, 4.3, 3.7, 13}, respectively. The error ratios of {3.2, 2.8, and 3.7} for D3S1358, CSF1PO, and TPOX may be too low for the resolution rendered by the top-ranked case to be fully confident. However, the mass proportion ratio of the second-ranked case for D3S1358 is 1:4.8, slightly too high compared with the average of 2.4 for the 11 loci. Therefore, only the top-ranked case for this locus is favored at this point. For CSF1PO, the mass proportion ratio for the second-ranked case is 1:1.6, just slightly too low compared with the majority of the top-ranked cases for the 11 loci, which are mostly above 1:2.0. For TPOX, the top-ranked case is also favored, because the mass proportion ratio for the second-ranked case is 1:5.0, too high to be regarded as consistent with those at the other 11 loci.

For the remaining two loci with two alleles in each locus (D5S818, D13S317) genotype combination Case 4 of [AB] and [AB] is not considered supported because the two-allele peak areas are not within the acceptable 0.6–1.67 range of heterozygous peak balance. For D5S818, the second listed case of {[12, 13], [12,12]} is favored because the associated mass proportion ratio of 1:2.3 is the most consistent to those for the 11 four- and three-allele loci examined previously. Finally, for D13S317, both the first- and second-listed cases have to be retained for consideration at this point because they have comparable mass proportion ratios: 1:1.9 and 1:2.2. Table 15 shows the final resolution based on LSD mathematical results and applying the interpretation guidelines. Two choices for D13S317 are listed to reflect the uncertainty. Assume that the victim's profile is available (from Table 10) and is now brought in to check the LSD-suggested resolution and to pick the appropriate resolution at locus D13S317. Comparison shows that person 1 of LSD result corresponds to that of the known contributor profile, and that at locus D13S317, the second-listed genotype combination case (with mass proportion ratio of 1:2.2) contains the known contributor profile. The other contributor of the mixture (the offender) also happens to be known for this example, and results show that the LSD-suggested resolutions are all correct at the remaining 12 loci. In a realistic application, a suspect's DNA profile can be compared with that of the other contributor from LSD (other than the known reference). If they match, one can draw the conclusion that given the victim's profile, and if the prosecutor's hypothesis is correct that the suspect is the

TABLE 10—*Data for the 13-loci mixture sample. LSD results are in Tables 11–14.*[*]

Locus	Alleles in the Mixture	Allele Peak Area	True Genotype Combination	
			Victim	Offender
D3S1358	15	1989	15	15
	16	739	16	
	18	1550		18
VWA	15	1318		15
	16	621	16	
	18	793	18	
	19	1200		19
FGA	21	2414	21	21
	22	1461		22
	23	687	23	
D8S1179	12	1431		12
	13	603	13	
	14	560	14	
	16	986		16
D21S11	28	1410		28
	30	1199	30	
	32.2	1506		32.2
D18S51	12	471	12	
	13	386	13	
	17	1181		17
	18	1029		18
D5S818	12	2561	12	12
	13	463	13	
D13S317	11	1607	11	11
	12	834		12
D7S820	8	723		8
	10	1203	10	10
	11	289	11	
D16S539	11	1262		11
	12	515	12	
	13	1253		13
	14	514	14	
THO1	5	944		5
	6	935		6
	8	633	8	
TPOX	8	1257	8	8
	10	984		10
	11	447	11	
CSF1PO	10	482	10	
	11	697		11
	12	617		12

[*]The true genotypes for both contributors are known and are shown in the Table (data kindly provided by R. Wickenheiser of Acadiana Crime Lab, New Iberia, LA).

TABLE 11—*Partial LSD mathematical result for the 13-loci mixture example.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D3S1358—three alleles {alleles 15 16 18} {peak area 1989 739 1550}						
	1	15, 16	15, 18	5.5e − 02	1.0	1.0:2.3
	2	16, 16	15, 18	0.18	3.2	1.0:4.8
	3	15, 15	16, 18	0.60	11	1.0:1.2
	4	16, 18	15, 18	0.85	15	1.0:4.6
	5	18, 18	15, 16	1.43	26	1.0:1.8
	6	16, 18	15, 16	4.79	87	1.0:1.7
VWA—four alleles {alleles 15 16 18 19} {peak area 1318 621 793 1200}						
	1	16, 18	15, 19	5.6e − 02	1.0	1.0:1.8
	2	16, 19	15, 18	0.79	14	1.0:1.2
	3	15, 16	18, 19	0.84	15	1.0:1.0
FGA—three alleles {alleles 21 22 23} {peak area 2414 1461 687}						
	1	21, 23	21, 22	5.0e − 02	1.0	1.0:2.0
	2	22, 23	21, 21	0.63	13	1.0:1.1
	3	23, 23	21, 22	0.96	19	1.0:5.6
	4	22, 23	21, 22	1.90	38	1.0:13.3
	5	22, 22	21, 23	3.16	63	1.0:2.1
	6	22, 23	21, 23	7.18	144	1.0:3.4

*Tables 11–14 contain the complete results. Mass ratio refers to the mass proportion ratio.

TABLE 13—*Partial LSD mathematical result for the 13-loci mixture example.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D13S317—two alleles {alleles 11 12} {peak area 1607 834}						
	1	12, 12	11, 11	0.0e+00		1.0:1.9
	2	11, 11	11, 12	0.0e+00		1.0:2.2
	3	12, 12	11, 12	0.0e+00		1.0: − 4.2
	4	11, 12	11, 12	0.43		1.0:1.0
D7S820—three alleles {alleles 8 10 11} {peak area 723 1203 289}						
	1	10, 11	8, 10	0.15	1.0	1.0:2.2
	2	8, 11	10, 10	1.13	7.7	1.0:1.2
	3	11, 11	8, 10	1.38	9.5	1.0:6.7
	4	8, 11	8, 10	2.36	16	1.0:29.0
	5	8, 8	10, 11	5.00	34	1.0:2.1
	6	8, 11	10, 11	10.70	73	1.0:3.7
D16S539—four alleles {alleles 11 12 13 14} {peak area 1262 515 1253 514}						
	1	12, 14	11, 13	1.6e-04	1.0	1.0:2.4
	2	13, 14	11, 12	2.09	1.3e+04	1.0:1.0
	3	12, 13	11, 14	2.09	1.3e+04	1.0:1.0

*Tables 11–14 contain the complete results. Mass ratio refers to the mass proportion ratio.

offender, then their DNA, when mixed at the calculated best-fit mass proportions, would best support the quantitative allele peak data of the given mixture sample.

This particular mixture represents a fairly clear-cut mixture for LSD to resolve with very little ambiguity (some uncertainty at the D13S317 locus). Most experienced DNA analysts would probably arrive at a comparable resolution without the help of LSD. The benefit in applying LSD is envisioned to be twofold: (1) it adds objectivity to the deconvolution of a mixture sample, one based on mathematics, and (2) it can act as a peer reviewer in the interpretation of mixture sample. It especially is expected to aid in situations where, at several loci more than one genotype combination choices appear to be compatible. LSD would systematically fit each combination in turn and gives an error for the fit, thus allowing the analyst to assess the degree of fit, and to pick the more compatible genotype combination case for resolution. When a known profile is available and is found to support LSD-suggested resolution at some or all loci, then the confidence in the resolution at those loci is increased.

The robustness of resolution by the LSD approach has been studied by the authors using thousands of simulated mixture data sets that exhibit slightly different mixture proportion ratios between mixture profile samples, and with varying degrees of heterozygous peak imbalance across loci. Some genotype combination patterns (for three-allele loci) are much more sensitive to peak imbalance than others. The results from these studies are the subject of a new manuscript currently under preparation.

TABLE 12—*Partial LSD mathematical result for the 13-loci mixture example.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D8S1179—four alleles {alleles 12 13 14 16} {peak area 1431 603 560 986}						
	1	13, 14	12, 16	0.32	1.0	1.0:2.1
	2	14, 16	12, 13	1.38	4.3	1.0:1.3
	3	13, 16	12, 14	1.44	4.5	1.0:1.3
D21S11—three alleles {alleles 28 30 32.2} {peak area 1410 1199 1506}						
	1	30, 30	28, 32.2	3.2e − 03	1.0	1.0:2.4
	2	32.2, 32.2	28, 30	1.5e − 02	4.8	1.0:1.7
	3	28, 28	30, 32.2	3.3e − 02	10	1.0:1.9
	4	30, 32.2	28, 32.2	0.28	88	1.0:1.3
	5	28, 30	28, 32.2	0.39	121	1.0:1.4
	6	28, 30	30, 32.2	0.68	213	1.0:1.1
D18S51—four alleles {alleles 12 13 17 18} {peak area 471 386 1181 1029}						
	1	12, 13	17, 18	0.10	1.0	1.0:2.6
	2	13, 18	12, 17	3.08	30	1.0:1.2
	3	12, 18	13, 17	3.17	31	1.0:1.0
D5S818—two alleles {alleles 12 13} {peak area 2561 463}						
	1	13, 13	12, 12	0.0e+00		1.0:5.5
	2	12, 13	12, 12	0.0e+00		1.0:2.3
	3	13, 13	12, 13	0.0e+00		1.0: − 2.4
	4	12, 13	12, 13	10.27		1.0:1.0

*Tables 11–14 contain the complete results. Mass ratio refers to the mass proportion ratio.

TABLE 14—*Partial LSD mathematical result for the 13-loci mixture example.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
TH01—three alleles {alleles 5 6 8} {peak area 944 935 633}						
	1	8, 8	5, 6	1.0e − 04	1.0	1.0:3.0
	2	5, 5	6, 8	0.11	1.1e+03	1.0:1.7
	3	6, 6	5, 8	0.12	1.2e+03	1.0:1.7
	4	5, 8	5, 6	0.32	3.2e+03	1.0:1.7
	5	6, 8	5, 6	0.34	3.4e+03	1.0:1.7
	6	6, 8	5, 8	1.29	1.3e+04	1.0:1.0
TPOX—three alleles {alleles 8 10 11} {peak area 1257 984 447}						
	1	8, 11	8, 10	5.1e − 02	1.0	1.0:2.4
	2	11, 11	8, 10	0.19	3.7	1.0:5.0
	3	8, 8	10, 11	0.72	14	1.0:1.1
	4	10, 11	8, 10	0.86	17	1.0:4.9
	5	10, 10	8, 11	1.64	33	1.0:1.7
	6	10, 11	8, 11	5.37	106	1.0:1.7
CSF1PO—three alleles {alleles 10 11 12} {peak area 482 697 617}						
	1	10, 10	11, 12	1.4e − 02	1.0	1.0:2.7
	2	11, 11	10, 12	3.9e − 02	2.8	1.0:1.6
	3	12, 12	10, 11	9.9e − 02	7.2	1.0:1.9
	4	10, 11	11, 12	0.23	17	1.0:1.4
	5	10, 12	11, 12	0.45	33	1.0:1.7
	6	10, 12	10, 11	0.99	72	1.0:1.2

*Tables 11–14 contain the complete results. Mass ratio refers to the mass proportion ratio.

TABLE 15—*LSD-guided resolution of the 13-loci mixture sample.**

Locus	Alleles in the Mixture	LSD Suggested Result			
		Victim	Suspect	Mass Ratio Calculated	Remarks
D3S1358	15	15	15		
	16	16		1:2.3	Correct
	18		18		
VWA	15		15		
	16	16			
	18	18		1:1.8	Correct
	19		19		
FGA	21	21	21		
	22		22	1:2.0	Correct
	23	23			
D8S1179	12		12		
	13	13			
	14	14		1:2.1	Correct
	16		16		
D21S11	28		28		
	30	30		1:2.4	Correct
	32.2		32.2		
D18S51	12	12			
	13	13		1:2.6	Correct
	17		17		
	18		18		
D5S818	12	12	12		
	13	13		1:2.3	Correct
D13S317		11	11		
	11		12	1:2.2	Correct
	12		11		
		12		1:1.9	
D7S820	8		8		
	10	10	10	1:2.2	Correct
	11	11			
D16S539	11		11		
	12	12		1:2.4	Correct
	13		13		
	14	14			
THO1	5		5		
	6		6	1:3.0	Correct
	8	8			
TPOX	8	8	8		
	10		10	1:2.4	Correct
	11	11			
CSF1PO	10	10			
	11		11	1:2.7	Correct
	12		12		

*The resolutions at all 13 loci except for D13S317 are made with confidence. Resolutions at all four- and three-allele loci are the top-ranked genotype combination case, and the resolutions for the two-allele loci are arrived at using the LSD interpretation guidelines where at D13S317, two genotype combination cases are supported by the given peak data.

Difference Between LSD and Other Quantitative Approaches

Both the Pendulum and LMA approaches (8,10) also use the least-square approach to find the optimum DNA proportions in the mixture. The mathematical system in which the least-square solution is sought is posed differently from that of LSD. The main difference between the mathematical formulation of LSD is that in LSD, peak data from a locus are processed independently from all other loci. A set of least-square problems is posed for every locus, using the allele peak data for that locus only. The optimum mass proportion vector is then calculated for each possible genotype combination at that locus. This allows for a different optimum mass proportion vector to be developed for each locus independently from all other loci. As a result, a different optimum mass proportion vector usually results for the top-ranked combination case from the least-square fit for each locus. In sharp contrast, in the formulation of and in the selection of the most supported genotype combination at all loci in Perlin and Szabady (8) and in the first phase of (10), peak data from all loci are concatenated together into a long vector, and one least-square fit to the combined string of peak data is carried out for each possible profile genotype comprised of all loci. This means that a common mixture proportion estimate across all loci, $\hat{M}_x$ (or referred to as weights in Perlin and Szabady (8)) is simultaneously imposed for all loci, and through either iterative search (10) or direct computation (8) the $\hat{M}_x$ that gives the smallest residual across all loci simultaneously is chosen to be the best-fit estimated $\hat{M}_x$ for all loci for a particular possible profile. (M_x in Perlin and Szabady (8) and Bill et al. (10) plays the same role as the mass proportion ratio in LSD) The formulation of a common $\hat{M}_x$ for all loci does not appear to yield the flexibility of accommodating for variations in mixture proportion across loci, which is known to exist (7,12,13). In contrast, LSD calculates a different best-fit estimated mass proportion for each locus separately. Thus, LSD yields a set of estimated best-fit mass proportion values (one for each locus), $\{\hat{M}_x^i\}_{i=1}^n$ where i denotes locus i, and n denotes the number of loci. The approach of (10) accommodates for variation in the true but unknown M_x by including a second phase, in which an updated mixture proportion for each locus is separately calculated based on the relative peak-height data, and the associated genotype combination is passed only if it lies within $\pm$ 0.35 (or user-defined) of the PENDULUM average (derived from the first phase) for the top 500 hits (12). In LMA, an explicit allowance for variation of DNA weights across loci is not provided. However, it is mentioned in Perlin and Szabady (8) that "the relative residual (as measured by "*dev*" in LMA) in the overall residual contributed by individual locus can be examined to see whether any locus has an unusually large residual associated with it, indicating a poor fitting result. It is suggested then that this locus be removed from the profile.

A comparison study of resolution of several well-publicized mixture samples by the method of (8) (applied to the case where one of the contributors is known and used in the computation for the profile of the unknown) and LSD showed somewhat different resolution results, with the LSD approach yielding correct resolution for all samples studied. Comparison results are documented in Wang et al. (11).

Integration of LSD into the Workflow of Forensic Laboratories

The main objective of LSD is to provide a set of resolved profiles for each of two contributors to a mixture sample that are well supported by the given allele peak data. It provides the mathematical results of the best-fit mass proportions for the two contributors to the mixture through a systematic fitting of each possible genotype combination at a locus, through the least-square fitting of the allele peak data to the genotype combination under consideration. It ranks the possible combinations at each locus according to their corresponding residuals. The DNA analysts can take the ranked results and apply the heuristic interpretation guidelines outlined in this paper to assign a degree of confidence to the LSD suggested resolution. Used in this way, LSD acts as a filter to suggest the best-supported genotype combination for each locus based on the associated peak data. This is expected to save the analysts from more laborious manual interpretation of the mixture data from the raw peak data. We suggest that when a sample can be assumed to be a two-contributor mixture, and artifact peaks have been identified and removed from the final allele calls, LSD be applied first to the peak data to yield the ranked result for the possible genotype combinations at each available

locus. Applied in this way, LSD will deliver consistent and systematic fitting results for the DNA analyst to follow up with LSD interpretation guidelines. It is emphasized at this point that a manual review of the LSD mathematical results is required. Successful automation of these two-step approaches to mixture interpretation will require wide field testing of the LSD framework with forensic data to identify its robustness and limitations. Nevertheless, some degree of human review of the final resolution results will always be necessary to assure interpretation integrity.

A second use for LSD is in providing a set of candidate profiles for a database search on systems such as Combined DNA Database System (CODIS). Each candidate profile is to be formed from those loci that are confidently resolved by LSD, in conjunction with all possible combinations of those less-confidently resolved loci in each of which more than one genotype combination is supported by the allele peak data. Unless the profile peak data have been severely compromised across many loci by varying degrees leading to widely inconsistent LSD mathematical results, the total number of candidate profiles resulting from such a combination is expected to be small: 32, for instance, if two candidate genotype combinations exist at each of five loci out of 13 core loci ($2^5 = 32$), which is a relatively small number out of the millions of possible profiles that can be formed from all possible combinations of all the loci alleles if no prefiltering of possible profiles is carried out first.

Summary and Conclusion

In this paper, a least-square-based interpretation framework is presented for interpreting two-contributor STR mixture samples using the quantitative allele peak data information. Applying the least-square principles, a best-fit mass proportion ratio is calculated for all possible genotype combination cases at each locus, independent of all other loci. Based on the ranked best-fit mass proportion ratios and the relative error residual ratios, a composite profile for each contributor can be developed from the LSD mathematical results using the heuristic interpretation guidelines explained in this paper. Results from studies using simulated data as well as from forensic case data show that LSD consistently gives correct resolution of profiles, provided no severe allele degradation or peak-height saturation (from overloading of DNA samples) exists among the given peak data of a locus. Some limitations do exist imposed by the inherent mathematical properties associated with nonuniqueness of solutions for two-allele loci. Using the suggested LSD interpretation guidelines may circumvent some of these limitations. The LSD framework for mixtures with mixture proportions close to 1:1 yields less certain results for four- and three-allele loci, due to possible heterozygous peak imbalance (four-allele loci), which would mask the slight difference in the mass proportion ratio, and due to degeneracy (for three-allele loci), in which two combinations would have the same pattern of relative peak heights. The use of a known contributor profile, when available, aids in the final determination of resolution when ambiguities exist. An expert system applying the interpretation guidelines documented here to interpret the LSD mathematical results is currently being developed by the authors. A follow up paper documenting the results of a sensitivity study of mixture resolution to the relative allele peak imbalance is also in preparation.

Acknowledgments

Encouragement and support from Dr. Barry Brown of FBI are greatly appreciated. Appreciation is extended to members of the LSD Joint Application Design Group, which includes Karen Ambrozy, Bruce Budowle, Kermit Channell, Kevin Miller, John Planz, Walther Parsons, Robyn Ragsdale, Ted Staples, and Ray Wickenheiser. Forensic data, valuable comments, and field testing of LSD from and by Ray Wickenheiser of Acadiana Crime Lab, New Iberia, LA, are especially appreciated.

Work supported by the FBI under contract J-FBI-98-083.

References

1. Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KAJ, Buckleton JS. Interpreting DNA mixtures. J Forensic Sci 1997;42(2):213–22.
2. Evett IW, Weir BS. Interpreting DNA evidence. Sunderland, MA: Sinauer Associates Inc., 1998.
3. Curran JM, Triggs CM, Buckleton JS, Weir BS. Interpreting DNA mixtures in structured populations. J Forensic Sci 1999;44(5):987–95.
4. Gill PD, Sparkes RL, Buckleton JS. Interpretation of simple mixtures when artifacts such as stutters are present—with special reference to multiplex STRs used by the Forensic Science Service. Forensic Sci Int 1998;95:213–24.
5. Evett IW, Gill PD, Lambert JA. Taking account of peak areas when interpreting mixed DNA profiles. J Forensic Sci 1998;43(1):62–9.
6. Gill PD, Sparkes RL, Pinchin R, Clayton TM, Whitaker JP, Buckleton JS. Interpreting simple STR mixtures using allelic peak areas. Forensic Sci Int 1998;91(1):41–53.
7. Clayton TM, Whitaker JP, Sparkes RL, Gill P. Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Sci Int 1998;91(1):55–70.
8. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. J Forensic Sci 2001;46(6):1372–8.
9. Mortera J, Dawid AP, Lauritzen SL. Probabilistic expert system for DNA mixture profiling. Theor Popul Biol 2003;63:191–205.
10. Bill M, Gill PD, Curran JM, Clayton TM, Pinchin R, Healy M, et al. PENDULUM—a guideline-based approach to the interpretation of STR mixtures. Forensic Sci Int 2005;148:181–9.
11. Wang T, Xue N, Wickenheiser R. Least-square deconvolution (LSD): a new way of resolving STR/DNA mixture samples. Proceedings of the Thirteenth International Symposium on Human Identification; 2002 Oct 7–10; Phoenix, AZ. Madison, WI: Sponsored by the Promega Corporation, 2002. Available at http://www.promega.com/geneticidproc/ussymp13proc/contents/wang.pdf.
12. Buckleton JS, Triggs CM, Walsh KAJ, editors. Forensic DNA evidence interpretation. Boca Raton: CRC Press, 2005.
13. Butler JM. Forensic DNA typing: biology, technology, and genetics of STR markers. 2nd ed. Burlington: Elsevier Academic Press, 2005.
14. Strang G. Linear algebra and its applications. 4th ed. San Belmont: Thomson Brooks/Cole, 2006.
15. Noble B, Daniel J. Applied linear algebra. 3rd ed. Englewood Cliffs, NJ: Prentice Hall Inc, 1987.
16. Stewart GW. Introduction to matrix computations. Orlando: Academic Press Inc., 1973.
17. Gill PD, Sprkes RL, Kimpton CP. Development of guidelines to designate alleles using an STR multiplex system. Forensic Sci Int 1997;89:185–97.
18. Whitaker JP, Clayton TM, Urquhart AJ, Millican ES, Downes TJ, Gill PD, et al. STR typing of bodies from the scene of a mass disaster: high success rate and characteristic amplification patterns in highly degraded samples. Bio Techniques 1995;18:670–7.

APPENDIX A

LSD Mathematical Steps: Examples

This section presents examples of applying the LSD mathematical steps to the peak data of a locus with four, three, and two alleles, respectively.

LSD Steps for a Four-Allele Locus—Let the four alleles at a locus be designated as {12, 13, 17, 18} with peak heights shown below

$$allele\_peak\_vector = \begin{bmatrix} 471 \\ 386 \\ 1181 \\ 1029 \end{bmatrix}$$

From inspection, it is clear that the first two alleles come from one contributor, the minor contributor, and the last two come from the second contributor, the major contributor. LSD analysis should yield the same conclusion with a high degree of confidence. First, the peak heights are normalized such that the normalized smallest peak is "1". The normalized peak vector now becomes

$$b = normalized\_allele\_peak\_vector = \begin{bmatrix} \frac{471}{386} \\ \frac{386}{386} \\ \frac{1181}{386} \\ \frac{1029}{386} \end{bmatrix} = \begin{bmatrix} 1.22 \\ 1.0 \\ 3.06 \\ 2.67 \end{bmatrix}$$

For four alleles, there are three possible allele combinations for the two contributors (see Table 1). Case 3 genotype combination is selected for processing first. The corresponding gene matrix for this case is shown below

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

For this case, the assumed allele combination pattern is [12 18] and [13 17], respectively. We are to find the particular combination weights for the two columns of A such that when combined with these weights would yield a fitted peak vector that is "closest" to the given peak-height vector b. The optimum weights give the best-fit mass proportions. The least-square solution is given directly by Eq. (3) shown previously. Using the appropriate pseudoinverse matrix from Table 1, the computed x_{ls} is thus

$$x_{ls} = A^{+} \bullet b = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}\begin{bmatrix} 1.22 \\ 1.0 \\ 3.06 \\ 2.67 \end{bmatrix} = \begin{bmatrix} 1.95 \\ 2.03 \end{bmatrix}$$

The corresponding mass proportion ratio is therefore 1.95:2.03, or 1:1.04. The predicted, or fitted allele peak vector, b_f is computed as follows:

$$b_f = A \bullet x_{ls} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1.95 \\ 2.03 \end{bmatrix} = \begin{bmatrix} 1.95 \\ 2.03 \\ 2.03 \\ 1.95 \end{bmatrix}$$

The error vector between b_f, and the measured b vector is computed as follows:

$$e = error\_vector = b - b_f = \begin{bmatrix} 1.22 \\ 1.0 \\ 3.06 \\ 2.67 \end{bmatrix} - \begin{bmatrix} 1.95 \\ 2.03 \\ 2.03 \\ 1.95 \end{bmatrix} = \begin{bmatrix} -0.73 \\ -1.03 \\ 1.03 \\ 0.72 \end{bmatrix}$$

The residual, or the corresponding square of the length of the error vector, is given as the sum of squares of the entries of the e vector, or

$$residuals = (-0.73)^2 + (-1.03)^2 + (1.03)^2 + (0.72)^2 = 3.17$$

The least-square solution, x_{ls}, calculated this way guarantees that the error, or residual, between the given peak-height vector, b, and the fitted vector, b_f, will be the smallest of all possible x's for the given genotype matrix, A. That is to say, no mass proportion vector, x, when premultiplied by the corresponding A matrix, will yield a vector, b_f, that is "closer" to the given vector, b than x_{ls} would; therefore, the computed x_{ls} vector constitutes the best-fit mass proportion vector for this particular genotype combination being tried. The magnitude of the residual alone is not indicative of how well this genotype combination fits the given allele peak data at the optimum mass proportion just calculated. It is in light of the residuals of all possible genotype combinations that the relative degree of the compatibility of each genotype combination is assessed. This point will be explained further later in the guidelines for interpreting LSD mathematical results.

Next, the best-fit mass proportion vector and the residual for another genotype combination for this locus is calculated, that of case 1. The genotype matrix, A for this case is shown below

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

The least-square solution is given by

$$x_{ls} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1.22 \\ 1.0 \\ 3.06 \\ 2.67 \end{bmatrix} = \begin{bmatrix} 1.11 \\ 2.87 \end{bmatrix}.$$

The corresponding mass proportion ratio is 1.11:2.87, or 1:2.6. The error vector is given by

$$e = error\_vector = b - b_f = \begin{bmatrix} 1.22 \\ 1.0 \\ 3.06 \\ 2.67 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1.11 \\ 2.87 \end{bmatrix}$$

$$= \begin{bmatrix} 0.11 \\ -0.11 \\ 0.19 \\ -0.20 \end{bmatrix}$$

TABLE 16—*Example of the LSD mathematical result for a four-allele locus.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D3S1358—four alleles {alleles 12 13 17 18} {peak area 471 386 1181 1029}						
	1	12, 13	17, 18	0.10	1.0	1.0:2.6
	2	13, 18	12, 17	3.08	30	1.0:1.2
	3	12, 18	13, 17	3.17	31	1.0:1.0

*The genotype combination cases are ranked according to the associated fitting error. Note that the fitting error refers to the fitting residual (sum of squares of the elements of the error vector); the error ratio refers to the ratio of the fitting error of each genotype combination case to the fitting error of the top-ranked combination case; the mass ratio refers to the ratio of the two mass proportions given by the least-square solution.

The residual, as calculated by the sum of squares of the elements of the error vector above is equal to 0.10, compared with 3.17 for the previous genotype combination. The relative error residual of the two fits is: 0.1:3.17, or a ratio of 1:32, indicating that the manner in which genotype combination case 3 explains the observed peak data with a fitting error residual of 3.17, is 32 times worse than the manner of explanation provided by genotype combination case 1.

The fitting residual for genotype combination case 2 of Table 1 is found to be 3.08. Therefore, of the three cases, genotype combination case 1 is considered the most supported one, conditioned on the given allele peak data, because it has the smallest fitting residual (0.1 vs. 3.08, and 3.17). Table 16 shows the ranked LSD mathematical results of the three cases for this locus of four alleles just processed.

LSD Steps for a Three-Allele Locus—Six genotype combination cases are possible for a three-allele locus, three of which involve one shared allele. The remaining three cases involve a mixture of one homozygous and one heterozygous contributor. Table 2 displays all six cases, along with the associated gene matrix and their pseudoinverses.

Exactly the same steps as those for the four-allele locus are carried out for the three-allele locus, resulting in a ranked list of the six cases according to their associated fitting residuals. Table 17 shows an example of the LSD mathematical output of a three-allele locus, in which the measured allele peak heights are {723, 1203, 289}. The top-ranked combination involves a shared allele between the two contributors. In general, when the two-contributor genotypes at a locus share an allele, a balanced peak-height ratio among the respective heterozygous alleles of each contributor would require that the composite peak-height ratio lie between

TABLE 17—*Example of the LSD mathematical result for a three-allele locus.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D7S820—three alleles {alleles 8 10 11} {peak area 723 1203 289}						
	1	10, 11	8, 10	0.15	1.0	1.0:2.2
	2	8, 11	10, 10	1.13	7.7	1.0:1.2
	3	11, 11	8, 10	1.38	9.5	1.0:6.7
	4	8,11	8, 10	2.36	16	1.0:29.0
	5	8, 8	10, 11	5.00	34	1.0:2.1
	6	8, 11	10, 11	10.70	73	1.0:3.7

*The true profiles in the mixture sample are both heterozygous and correspond to the top-ranked combination. Note that the fitting error refers to the fitting residual (sum of squares of the elements of the error vector); the error ratio refers to the ratio of the fitting error of each genotype combination case to the fitting error of the top-ranked combination case; the mass ratio refers to the ratio of the two mass proportions that are given by the least-square solution.

TABLE 18—*Example of the LSD mathematical result for a second three-allele locus.**

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
TH01—three alleles {alleles 5 6 8} {peak area 944 935 633}						
	1	8, 8	5, 6	1.0e − 04	1.0	1.0:3.0
	2	5, 5	6, 8	0.11	1.1e+03	1.0:1.7
	3	6, 6	5, 8	0.12	1.2e+03	1.0:1.7
	4	5, 8	5, 6	0.32	3.2e+03	1.0:1.7
	5	6, 8	5, 6	0.34	3.4e+03	1.0:1.7
	6	6, 8	5, 8	1.29	1.3e+04	1.0:1.0

*The true profiles in the mixture sample consist of one heterozygous and one homozygous and correspond to the top-ranked combination. Note that the fitting error refers to the fitting residual (sum of squares of the elements of the error vector); the error ratio refers to the ratio of the fitting error of each genotype combination case to the fitting error of the top-ranked combination case; the mass ratio refers to the ratio of the two mass proportions that are given by the least-square solution.

0.6 and 1.66, as shown below

$$0.6 \le \frac{\varphi_{\text{shared allele}}}{\varphi_{\text{unshared1}} + \varphi_{\text{unshared2}}} \le 1.66$$

where φ denotes the peak height. In this case, the calculated composite peak-height ratio is 1.18, well within the limit. Further, LSD mathematical results show that the second-ranked genotype combination has a fitting residual about eight times of that of the top-ranked one, indicating a markedly less-supported genotype combination. Moreover, the second-ranked case assigns alleles [8 11] to the same person, whereas the other person is homozygous in allele 10. This assignment is not supported well by the very imbalanced peak heights of 723 and 289 for alleles [8 11] if they are assigned together to the same person. Table 18 shows the LSD mathematical results of a second three-allele locus in which it is known that one of the contributors is homozygous and the other is heterozygous at this locus. The measured allele peaks are {944, 935, 633}. Note that the peaks of the first two alleles are comparable with each other (with a peak ratio of 1.01:1, well within the 0.6–1.67 range), and no peak is close to the sum of the other two peaks. Therefore, a shared allele is not supported by the peak data. It is evident, for this example, that the top-ranked case is clearly the most supported resolution given the allele peak data as indicated, because the next ranked case has a fitting residual of more than 1000 times larger than that of the top-ranked case. The large separation of the two-error residuals indicates that one can be very confident that the top ranked genotype combination gives the best-supported resolution, given the peak data. In addition, the heterozygous peak balance ratio for alleles 5 and 6 is almost 1:1, well within the 0.6–1.67 range indicating balanced peaks.

LSD Steps for Two-Allele Loci: Special Consideration and Interpretation—Table 3 shows that four genotype combinations are possible for a two-allele locus, involving one homozygous–homozygous, two heterozygous–homozygous, and one heterozygous–heterozygous mix, respectively.

The same mathematical procedures as those shown for a four-allele locus are used to process the allele peak data of a two-allele locus. Table 19 gives an example. The two allele peak heights for this locus, {2561, 463}, are not comparable with each other (not within the 0.6–1.67 range). Therefore, resolution given by the fourth combination is not supported based on the given peak-height data. The third case is automatically excluded from further consideration because it has a negative best-fit mass proportion ratio, which is not realistic. Without taking into consideration

TABLE 19—*LSD results for a two-allele locus.*[*]

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D5S818—two alleles {alleles 12 13} {peak area 2561 463}						
	1	13, 13	12, 12	0.0e+00		1.0:5.5
	2	12, 13	12, 12	0.0e+00		1.0:2.3
	3	13, 13	12, 13	0.0e+00		1.0: − 2.4
	4	12, 13	12, 13	10.27		1.0:1.0

[*]The allele peak areas of the two alleles are not comparable (outside the 0.6–1.67 range), implying that case 4 is not a well-supported candidate for the correct resolution. Note that the first three combination cases have 0 fitting error associated with them, due to the fact that the gene matrix for each of these has two independent rows with two unknowns; thus, an exact solution exists. The mass proportion ratio for the third case is negative, thus rendering this case infeasible.

TABLE 20—*LSD results for a second two-allele locus.*[*]

Select	Case	Person 1	Person 2	Fitting Error	Error Ratio	Mass Ratio
D13S317—two alleles {alleles 12 13} {peak area 2105 2591}						
	1	12, 12	13, 13	0.0e+00		1.0:1.2
	2	12, 12	12, 13	0.0e+00		1.0: − 10.7
	3	13, 13	12, 13	0.0e+00		1.0:8.7
	4	12, 13	12, 13	2.7e − 02		1.0:1.0

[*]The allele peak areas of the two alleles are comparable with each other, implying that case 4 is well supported by the given peak area data. Note that the first three combination cases have 0 fitting error associated with them, due to the fact that the gene matrix for each of these has two independent rows with two unknowns; thus an exact solution exists. The mass proportion ratio for the second case is negative, thus rendering this case infeasible.

what the LSD mathematical results are at the other loci, the first two cases, for now, are retained to be supported resolutions conditioned on the given peak data. Table 20 gives another example. In this case, the two-allele peak heights are {2105, 2591}, at a ratio of 0.84:1, and can be considered to be comparable with each other (within the 0.6–1.67 range). As a result, the fourth combination cannot be ruled out for consideration as the compatible resolution. As a matter of fact, it is a strong candidate to be the most supported one at this point given the observed peak data information. The second case is invalid due to its negative mass proportion ratio.

It is emphasized that review by a DNA analyst of the raw LSD mathematical results is required in order to put together the composite profile for each contributor. The analysis of a 13-loci mixture sample given earlier in the paper illustrates the heuristic guidelines the analyst would use in the review of the LSD mathematical results in forming a composite profile for each of the two contributors.

Additional information and reprint requests:
Tsewei Wang, Ph.D.
Department of Chemical Engineering and Laboratory for Information Technologies
431 The University of Tennessee
Dougherty Hall
Knoxville, TN 37996-2200
E-mail: wang@lit.net