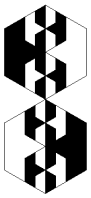
Simple Reporting of Complex DNA Evidence: Automated Computer Interpretation

Mark W. Perlin
Cybergenetics, Pittsburgh, PA

November 18, 2003

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



Contact information:

Mark W. Perlin, PhD, MD, PhD
Chief Executive Officer
Cybergenetics
160 North Craig Street
Suite 210
Pittsburgh, PA 15213
USA
(412) 683-3004
(423) 683-3005 FAX
perlin@cybgen.com

Abstract

Complex DNA evidence (e.g., mixed, degraded, or small amounts of DNA) can confound straightforward approaches to interpretation and reporting. Reporting simplicity here has four components: (1) rapid turnaround time to the police or prosecutors who need DNA answers, (2) high information content that provides a useful discriminating power (DP), (3) understandable results that can be presented to the layperson, and (4) admissibility in court that reflects the underlying reliability of the scientific methods employed.

We have developed a fully automated expert computer system that interprets complex DNA evidence based on mathematical models of the STR process, and a hyper-modern statistical assessment of certainty. After taking several minutes to consider thousands of numerical variables, the computer can present the results of its rigorous deliberations as ordinary probabilities that are understandable to juries. This paper describes the collaborative study design, data generation, analytic interpretation, and scientific results obtained in validating our expert system.

The validation study was designed collaboratively with crime labs in Florida, Maryland, and Virginia. The "mock rape kit" approach analyzed a set of two contributor "sperm fraction" DNA mixtures (10%, 30%, 50%, 70% and 90% proportions; at 1.0, 0.5, 0.25 and 0.125 ng dilutions), along with their "victim" reference samples. NIST prepared the mixed DNA stock solutions in known proportions for two individual pairs, and sent these materials to the ten participating forensic laboratories. Each lab followed detailed study protocols, generated PCR data for 56 templates, and sent their original sequencer data to Cybergenetics for automated computer interpretation. In addition, each lab also provided about 100 single-source samples that were used for calibrating their PCR process and artifacts (stutter, preferential amplification, peak variation, etc.). All STR chemistries and DNA sequencers in current forensic use were represented.

Cybergenetics ran their sequencer-independent TrueAllele® program on the original lab data to generate a database of quality checked, quantitated peaks (under five minutes of human time per gel). The interpretation expert system was then applied to these data (no human time). The calibration produced graphs of stutter and pref amp for each marker. The automated mixture analysis for each lab's 40 unknown suspect cases (mixture and reference data, but no suspect data) yielded an estimate of the mixing proportion, the genotype confidence set at each locus, and quantitative bar graphs that permitted rapid visual comparison between the observed data and the best genotype model. These computer results were reviewed by each laboratory.

The DNA templates and the computer interpretation were the same throughout the study, so we could compare the amount of information present in each laboratory's data. Our key information measure was discriminating power – the probability that a randomly selected person from the population matches the inferred unknown suspect profile. Unlike conventional measures (e.g., CPE, likelihood ratios), our computed DP is an understandable probability result that reflects the computer's detailed consideration of all feasible genotypes, and captures all relevant information. We (1) found that DP is an accurate measure of laboratory data quality, (2) established a quantitative relationship between mixture proportion and template dilution relative to information content, and (3) showed how the computer could combine multiple mixture samples at low DNA concentration without any reference sample to derive very high unknown suspect DP.

The legal admissibility of our scientific approach was established by the reliability demonstrated in our collaborative validation study. All points of Daubert were addressed: testability, error rate, peer review, and general acceptance. These admissibility issues are detailed in this paper. We have validated an expert computer approach to the interpretation of complex DNA evidence that is objective, unbiased, reproducible, reliable and admissible. Moreover, the system generates simple and understandable results that are useful in court.

Table of Contents

Introduction

Justice helps protect society by exonerating the innocent, and apprehending and convicting the guilty. One task of the criminal justice system is to identify the people who were present at a crime scene. In the past decade, DNA identification has become an objective way to obtain such scientific evidence (1). This section describes each of the concepts introduced in the title of this report.

Simple reporting of pristine, single source DNA data is understandable, and readily achieved. The data in Figure 1 show such a clean DNA profile. At each locus, there are exactly 1 or 2 major peaks, and each peak directly corresponds to an allele. Interpretation here is easy – at each locus, look at the peaks, and report the genotype as the pair of alleles that is present.

With complex DNA evidence, however, the underlying interpretation may not be immediately apparent. Figure 2 shows a mixed DNA profile comprised of two contributors. At a locus with three peaks, it is unclear which alleles come from which contributors. There are many such data issues that are common in DNA casework: low peak height, varying peak size, PCR stutter artifact, preferential amplification, mixed DNA stains, low amounts of DNA, degraded DNA, interpreting many samples together, handling multiple underlying contributors, determining the number of contributors, background noise, random peak variation, and low copy number signals. Thus, peaks no longer directly correspond to alleles. Moreover, as society's expectations increasingly demand DNA for every routine case, such fuzzy data will become the future of DNA evidence.

Fortunately, automated computer interpretation can help satisfy these expectations. For every data artifact, scientists can model the phenomenon with mathematics, and use statistics to assess the uncertainty. Although these models can become quite complex, entailing thousands of interacting equations, they are readily solved by computers. The result is that even the most complex DNA evidence can be reported as simply as clean single source data. For each contributor, at each locus, the computer simply reports the genotypes. With unambiguous data, this genotype will be unique; highly ambiguous data may lead to multiple genotype possibilities.

Background

The last decade has witnessed many important revolutions. We focus here on just four.

- Science. In molecular biology, the amplification power of polymerase chain reaction (PCR) (2), and its application to short tandem repeat (STR) polymorphisms (3), have revolutionized scientists' ability to study trace amounts of DNA. And, the complementary automated fluorescent DNA sequencer (4) (which sequenced the human genome) provided a painless way to quantitatively characterize STR fragments.

- Technology. For the last four decades, the ubiquitous computer has exponentially increased each year in power and compactness (5). Whether used in cell phones, automobiles, DNA sequencers, or the Internet, computer intelligence pervades and facilitates our daily productivity.
- Law. The 1993 Daubert decision revolutionized the role of science in the courtroom (6). Instead of relying on accepted practice, judges are now the gatekeepers of scientific evidence, applying objective criteria to assess scientific reliability.
- Statistics. In the past decade, using new mathematics and modern computers, statisticians have moved beyond the null hypothesis, and now apply powerful methods to determine computational truth (7). These new techniques have been used in hundreds of applications (8).

Cybergenetics TrueAllele® expert system automates the interpretation of STR data. An expert system is a computer program that replicates (or transcends) human expertise (9); this is what the TrueAllele system does for allele calling. The TrueAllele project started over ten years ago, with the concept that the PCR stutter artifact in dinucleotide repeat markers was reproducible, and could be mathematically removed by computer. The resulting STR analysis, interpretation and reporting technology led to a series of papers (10, 11), patents (12) and genetic applications (13). Five years ago, the TrueAllele System 2 software was adapted for forensic databasing applications (14-16). Over the last four years, we have been developing TrueAllele System 3, an intelligent system for forensic casework (17). This past year, we designed, built and tested version 15, a System 3 program specifically created for solving complex casework that contains lower quality data.

Validation

Scientific evidence is admissible in court only if it is reliable. Rule 702 of the United States Federal Rules of Evidence states that all of these components must be reliable: the data, the method used, and the application of the method to the data. The older 1923 Frye standard assessed this reliability based on general acceptance. The more modern 1993 Supreme Court Daubert ruling provided three additional prongs for determining the reliability of each component (data, method, application):

(1) Testable. The component should be capable of being tested, and have actually been tested.
(2) Error Rate. The component should have an error rate, which has actually been determined.
(3) Peer Review. The component (and its reliability) should be disseminated in the relevant scientific community.

The usual approach to demonstrating these admissibility criteria is to establish reliability through a validation study.

Our validation study was designed to assess interpretation efficacy for two contributor mixture cases. The experimental design has three axes: varying mixture ratios, serial DNA dilutions, and different contributor pairs. As shown in Table 1, the mixture weights

were 10%, 30%, 50%, 70% and 90%. The DNA amounts were 1 nanogram, 0.5 ng, 0.25 ng and 0.125 ng, standardized to a 25 microliter volume. DNA from four different individuals was used to create two distinct sets of two contributor mixtures.

Premixed DNA templates were prepared by the National Institute of Standards and Technology (NIST). There were 14 stock solutions, 7 for each mixture set, which included the two individuals and the five mixed DNA samples. NIST sent these stock solutions to each of the 10 participating DNA laboratories, located in Florida, Maryland, New York, Ohio, Pennsylvania, Virginia and the United Kingdom. Using detailed laboratory protocols that we had prepared for this study, each lab diluted their stock solutions into the 56 PCR templates of the experimental design. Each laboratory then followed its usual casework protocols to generate electronic DNA sequencer files. A diverse set of STR panels (Promega PowerPlex 1, 2, 16; ABI ProfilerPlus, Cofiler, SGMplus, Identifiler) and DNA sequencers (Hitachi FMBio; ABI 377, 310, 3100, 3700) were used in the STR data generation. Each lab forwarded its completed DNA sequencer files to Cybergenetics for further processing in the study.

Technology

The TrueAllele® Technology automates the three cognitive tasks that follow STR data generation: analysis, interpretation and reporting. Starting from the original sequencer data file, the TrueAllele Analysis program automatically transforms the file into a quality checked peak height database. This is shown in the upper half of Figure 3. The computer independently performs all steps of the analysis process, including image and signal processing, background subtraction, dye color separation, lane tracking, peak detection, peak sizing, ladder derivation and comparison, coordinate transformation, peak quantitation, artifact detection, and quality assurance of the sequencer run and its controls (e.g., positive, negative, sizing and allelic ladders). The human task is then reduced to simply checking the computer's automated quality checking, which takes about 3 minutes of operator time per sequencer run.

In our earlier TrueAllele System 2 for DNA databanking, after the peak analysis there is additionally the interpretation of the single source data (16). The computer uses over 20 rules, applied to each genotype, to identify potentially problematic allele calls. The user then reviews just these problematic genotypes (typically, about 10% of the total), and decides whether to except, reject or edit the allele calls. This computer-based, streamlined STR review process has led to considerable efficiencies in DNA databank construction, particularly in the United Kingdom.

The UK Forensic Science Service (FSS) adds 350,000 STR profiles every year to the UK National DNA Database, using Cybergenetics' TrueAllele expert databank system to automate their data review. The FSS identified several key TrueAllele process improvements relative to their previous semi-automated manual process. These improvements included (a) reduced turn-around time from one week to an eight hour shift, (b) adaptive capacity that could meet or exceed the throughput requirements of

30,000 samples per month, using two desktop computers, (c) reduced manpower requirements from 75 individuals to 2 people per shift, with elimination of associated space, computers and software, (d) a high quality process, leading to greater confidence in the accuracy of the data review results and automated troubleshooting, and (e) standardization of the databank review process, making it reproducible and objective, with automated quality assurance and audit trails.

For casework interpretation, the TrueAllele Interpretation program downloads relevant data from the quality-checked quantitative peak database, automatically interprets the data, and then uploads its results to the reporting database. This is shown in the lower half of Figure 3. When interpreting the data, the TrueAllele program applies the mathematical and statistical models of data behavior described above, and infers the genotypes of each of contributor, along with other useful formation, such as mixture weights and discriminating power. This process takes no human time, since the interpretation is done entirely by computer.

Example

To understand how the TrueAllele interpretation system works, it is useful to step through its problem solving on a straightforward no suspect, mock sexual assault example. The STR data used in this example were amplified in the Cybergenetics laboratory using 1 ng NIST DNA templates with a Promega PowerPlex 16 STR panel, and then detected on an ABI 310 automated fluorescent DNA sequencer. In this mock case, there are two contributors: (A) the victim, and (G) the unknown perpetrator. Two mock evidence data samples are interpreted together:

- A1, the victim control sample, and
- C1, a DNA mixture comprising a 70% contribution from victim A, and a 30% contribution from the unknown perpetrator G.

For objectivity, the computer only evaluates the case evidence, and does not consider any suspect profiles (e.g., perpetrator G) when it performs its interpretation. Suspect matching can be done after the computer has completed all of its processing.

The victim control A1 data are shown in Figure 1 – a clean, single source, easily interpreted DNA profile, where each peak corresponds directly to an underlying allele. The data peaks for mixture C1, shown in Figure 2, have more complex patterns. In some of the four peak patterns (e.g., TH01), one can visually pick out the two minor peaks that probably belong to the unknown minor contributor. However, in the two and three peak patterns, it is more difficult to derive the minor contributor genotype possibilities.

The TrueAllele interpretation computer solves this problem using two inputs: a quantitative peak database for the peaks shown in the two data figures, and an interpretation request that tells the computer where to find the two DNA sequencer lanes used in this case. Using SQL database queries to look into the PostgreSQL relational database (18) that underlies TrueAllele System 3:

- Table 2 shows some of the peak information recorded for a few peak records. Each peak table record corresponds to one data peak, and describes its chain of custody (e.g., laboratory, DNA sequencer, STR panel), peak size information (e.g., pixel, designation), and peak quantitation information (e.g., height, area). All peaks with heights greater than zero are recorded in the database, since statistically reliable interpretation requires all peaks to be considered.
- Table 3 shows the request table that connects each specimen in this case with its DNA sequencer capillary or lane location. This information is provided by a forensic scientist, or by a laboratory information management system (LIMS), in order to specify the scientific interpretation question. The request can include specimens from one case, from many cases, or from one focused part of a case.

The actual processing of this case was demonstrated live during the conference presentation on an Apple G4 PowerBook laptop computer. The computer downloaded the peak and request information from the database, set up the problem, interpreted the data by considering thousands of variables (including all feasible genotypes and mixture weights, each peak and its variation, PCR stutter, preferential amplification, and background noise), and then uploaded its results back to the database. The entire process took 29 seconds.

The TrueAllele report for this case includes results of interest to both scientists and laypeople. The results were queried from the underlying TrueAllele relational database:

- Table 4 shows the mixture weights, which can be used as evidence of contact between biological materials. The victim sample A1 corresponds to the first contributor, since the victim's DNA has a weight of 1 for the first contributor, and a weight of 0 for the second contributor. The probative specimen C1 is a mixture of the two contributors, with a weighting that combines about 70% of the first victim contributor, and 30% of the unknown second contributor. These computer inferred weights accurately correspond to the experimentally created proportions in DNA templates A1 and C1.
- Table 5 shows the genetic identity of the unknown second contributor. (The computer also infers the first contributor victim's genetic identity, which is straightforward since it has available reference sample A1.) In this mock case, even though the unknown is a 30% minor contributor, the TrueAllele computer infers a unique (and correct) genetic identity at the 99% confidence level at each STR locus. This mathematical conclusion is possible because the laboratory data contain sufficient information to support a high level of statistical precision.

These results satisfy the reporting simplicity goals stated in the Abstract.

(1) Time. The total interpretation and reporting time for this minor contributor mock sexual assault case was less than one minute.

(2) Information. As detailed in the next section, the unique inferred genetic profile has full discriminating power for identifying (hence apprehending and convicting) the unknown suspect, and excluding innocent men.

(3) Understandable. The results for the triers of fact were presented as the genotype at each locus of the unknown contributor. (Had there been greater data ambiguity, a

locus might have reported multiple genotypes.) This mixture result is as easy to understand as the reporting of clean, single source DNA profiles.

(4) Admissible. These results are reproducible and scientifically reliable, as reported here and in forthcoming scientific publications. In particular, one can test a computer interpretation system, and determine its error rate, as mandated by Rule 702 under Daubert.

There are 1440 minutes in each day; there are many laptop computers in this world. And yet there is a backlog of hundreds of thousands of no suspect sexual assault cases and convicted offender profiles (19). Clearly, with the TrueAllele technology, the analysis, interpretation and reporting components of this case backlog can be fully addressed within several months. It may be advantageous for society to identify perpetrators, and prevent them from committing further violent crimes, using fast computer-based interpretation methods .

Information

Discriminating power is a good measure of DNA information. Although each person has a unique genetic profile, crime scene evidence can have varying degrees of clarity. Think of a photograph: a sharp picture of someone's face may be uniquely identifying, but a blurry image taken from a distance on a rainy night might match millions of individuals. A similar range of identification power is found with DNA evidence, where a unique genotype may have a discriminating power of one in quadrillion, whereas complete ambiguity yields no discriminating power at all.

Consider the illustrative three peak example in Figure 4. A typical "conservative" interpretation might report six possible genotypes at locus, offering relatively little discriminating power. A less conservative "cautious" group might consider data at other loci or specimens, and eliminate some possibilities in order to increase discriminating power. Alternatively, a mathematically powered, statistically enabled computer solution might scientifically derive the "exact" solution, thereby realizing the greatest discriminating power. Sharper discriminating power provides greater scientific truth to the criminal justice system.

As part of our ongoing TrueAllele concordance studies, we are comparing computer and human discriminating powers on the same cases. In a no suspect case, where no match is performed, discriminating power can provide an objective comparison measure. Table 6 compares the genetic profile results in the example case for the TrueAllele computer interpretation and a double reviewed forensic laboratory human interpretation. The computer accurately designates all 26 of the 26 possible alleles, with population frequencies generating a discriminating power of 3.6×10^{16} (rarer than one in quadrillion). The more conservative human review designates only 18 of the 26 alleles, yielding a discriminating power of 6.1×10^{11} (commoner than one in a trillion). The almost five order of magnitude computer improvement in derived information becomes increasingly important as DNA case data become more ambiguous.

Application

There are many ways to deploy the TrueAllele interpretation technology into a laboratory process. The approach taken depends entirely on the laboratory's workflow, speed and accuracy requirements, and re-engineering objectives. Representative scenarios include using the TrueAllele automation:

- to initially screen the data and organize the case, prior to human review;
- as a second scorer that reviews the case data after an initial human review;
- for post-conviction DNA testing (e.g., in an innocence project) as a way to screen DNA case evidence in-depth, prior to involving a defense expert;
- as a DNA assistant for the forensic scientist that can perform the detailed review of DNA evidence;
- for cleaning up inferred DNA case profiles before uploading them to a forensic crime database, by reducing the set of possible genotypes to just those that are scientifically feasible;
- to reduce from many to one (or none) the number of scientifically feasible DNA suspect profiles obtained from an ambiguous convicted offender database search – this can reduce considerably the police effort of tracking down all these suspects;
- to review the DNA data in large-scale mass disasters;
- for solving serial crime by combining DNA evidence from multiple cases;
- to detect terrorist activity by inferring genetic profiles from low-level DNA evidence; and
- for quality assurance, and in the real-time troubleshooting of automated DNA production lines.

The forensic scientist integrates diverse types of scientific evidence into a unified, coherent presentation. An advanced TrueAllele computer that can reliably interpret DNA data can free the scientist from a task better done by high-powered calculators. Computer interpretation would give the forensic scientist time to work more closely with police, prosecutors and the courts – time to synthesize and communicate information in ways that only people can. By offloading some of the computational DNA burden to the computer, there would be more human time for productive, creative scientific thought.

Conclusion

This paper introduced an automated computer system for DNA casework interpretation. The system is:

- objective, inferring genetic identity results in complex mixture cases without ever seeing a suspect DNA profile;
- unbiased, using only quantitative DNA peak information on the experiments of interest;
- reproducible, consistently generating the same results given the same data;

- reliable, producing scientifically sound answers that are useful to the criminal justice system;
- admissible, based on scientific studies that test reliability and determine error rates; and
- available, as the TrueAllele System 3 software, version 15.

The system uses its mathematical power to generate genetic identity reports that are easy to understand, and simple to explain to nonspecialists. The results are analogous to the jury-tested reporting of single source DNA samples. Our computational approach differs from current mixture reporting methods, whose inherent complexity may confuse the issues or mislead the jury.

We are currently conducting a number of concordance and validation studies with both government and private DNA forensic laboratories. Our concordance studies on no-suspect sexual assault cases use discriminating power as a measure of useful information. When perpetrators have been identified, then match results can be used to assess interpretation accuracy. In our ongoing studies, we expect to analyze up to one thousand cases in the coming year; we will report on these results in the scientific literature.

The goal of this report was to show how simple reporting of complex DNA evidence could be achieved through automated computer interpretation. As described above, we have established this result. Moreover, we presented an actual problem solving case example that illustrated our simplicity criteria of turnaround time, genetic information, layperson understandability, and legal admissibility. As intelligent DNA interpretations systems come into general use, we expect society to benefit from the more rapid and accurate application of DNA evidence in our criminal justice system.

Acknowledgements

I would like to thank our collaborating forensic DNA laboratories in Florida, Maryland, New York, Ohio, Pennsylvania, Virginia, and the United Kingdom. Dr. Margaret Klein at NIST prepared and distributed the mixed DNA templates for this study. Dr. Jay Kadane at Carnegie Mellon University collaborated in the statistical development. Jeff Ban, Dr. Robin Cotton and Dr. Cecilia Crouse were instrumental in designing the study. At Cybergenetics, the software developers, quality process team, and DNA laboratory all contributed significantly to the study.

This research was supported in part under Award number 2001-IJ-CX-K003 from the Office of Justice Programs, National Institute of Justice, Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice.

References

1. Evett IW, Weir BS. Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists. Sunderland, MA: Sinauer Assoc, 1998
2. Mullis KB, Faloona FA, Scharf SJ, Saiki RK, Horn GT, Erlich HA. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol. 1986;51:263-273.
3. Weber J, May P. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. 1989;44:388-396.
4. Strauss E, Kobori J, Siu G, Hood L. Specific-primer-directed DNA sequencing. Anal Biochem 1986;154(1):353-60.
5. Moore GE. Cramming more components onto integrated circuits. Electronics 1965;38(8):114-117.
6. Sanders J, Kaye DH, Faigman DL, Saks MJ. Modern Scientific Evidence. Second ed. Eagan, MN: Thomson/West, 2002
7. Gelman A, Carlin J, Stem H, Rubin D. Bayesian Data Analysis. Boca Raton, FL: Chapman Hall/CRC, 1995
8. Gilks WR, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. Chapman and Hall, 1996
9. Brownston L, Farrell R, Kant E, Martin N. Programming Expert Systems in OPS5 - An Introduction to Rule-Based Programming. Reading, MA: Addison-Wesley, 1985
10. Perlin MW, Burks MB, Hoop RC, Hoffman EP. Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. Am. J. Hum. Genet. 1994;55(4):777-787.
11. Perlin MW, Lancia G, Ng S-K. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am. J. Hum. Genet. 1995;57(5):1199-1210.
12. Perlin MW. Method and system for genotyping. US Patents; 1996-2000. Report No.: #5,541,067, #5,580,728, #5,876,933 and #6,054,268.
13. Pálsson B, Pálsson F, Perlin M, Gubjartsson H, Stefánsson K, Gulcher J. Using quality measures to facilitate allele calling in high-throughput genotyping. Genome Research 1999;9(10):1002-1012.
14. Perlin MW. Computer automation of STR scoring for forensic databases. In: First International Conference on Forensic Human Identification in The Millennium; 1999 Oct 25-27; London, UK: The Forensic Science Service; 1999.
15. Perlin MW. An expert system for scoring DNA database profiles. In: Promega's Eleventh International Symposium on Human Identification; 2000; Biloxi, MS; 2000.
16. Perlin MW, Coffman D, Crouse CA, Konotop F, Ban JD. Automated STR data analysis: validation studies. In: Promega's Twelfth International Symposium on Human Identification; 2001; Biloxi, MS; 2001.
17. Perlin MW, Szabady B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. Journal of Forensic Sciences 2001;46(6):1372-1377.
18. Momjian B. PostgreSQL: Introduction and Concepts. Boston, MA: Addison Wesley, 2001
19. Hart SV. DNA Initiatives. Washington, DC; 2002 May 14.

Figures

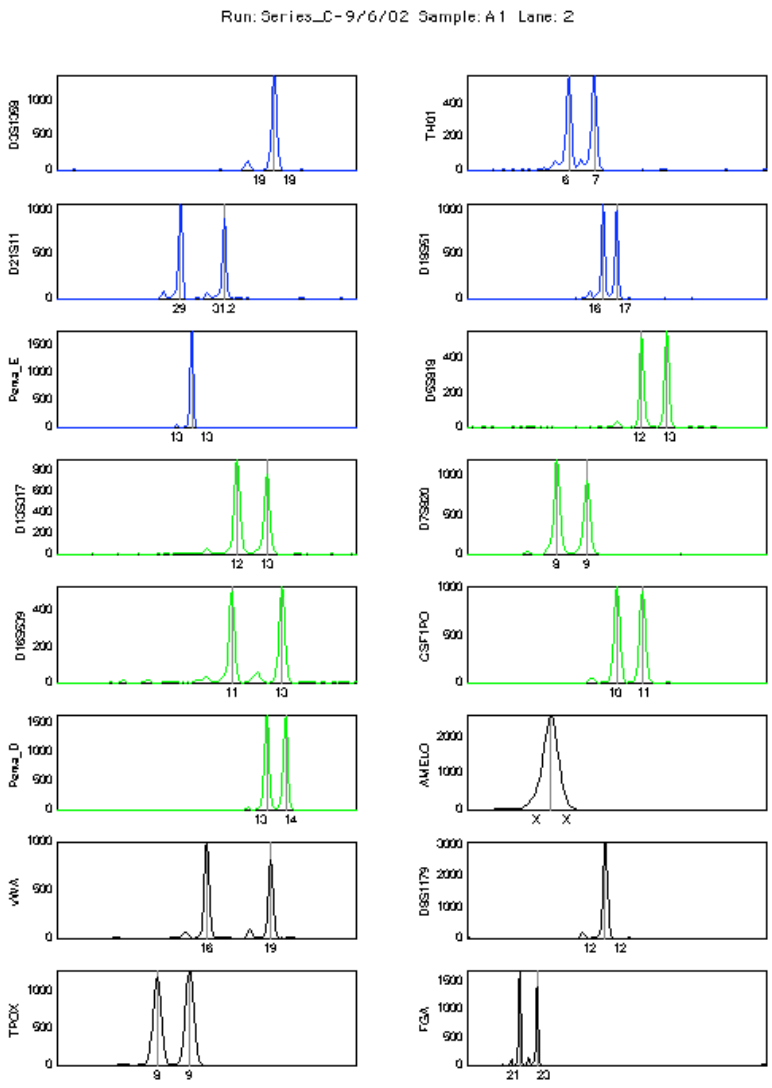


Figure 1. The pristine STR data from single source victim control sample A1.

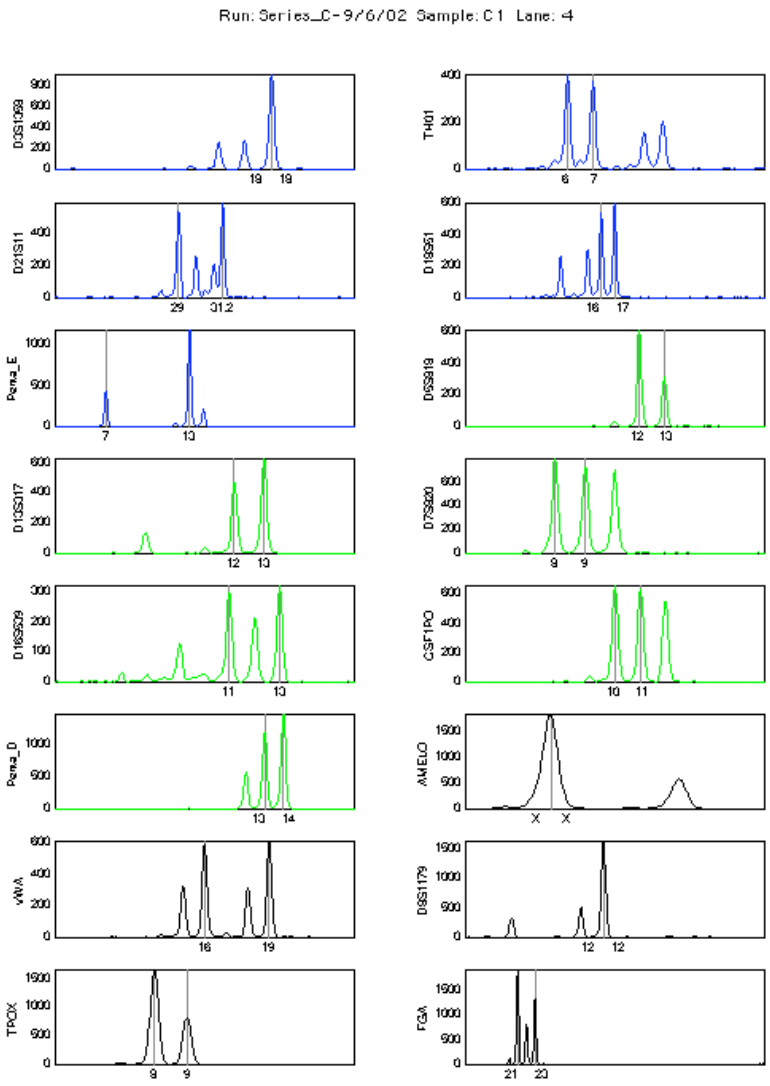


Figure 2. The STR data from the mixed DNA specimen C1. There are two contributors, weighted in a 70:30 mixture ratio.

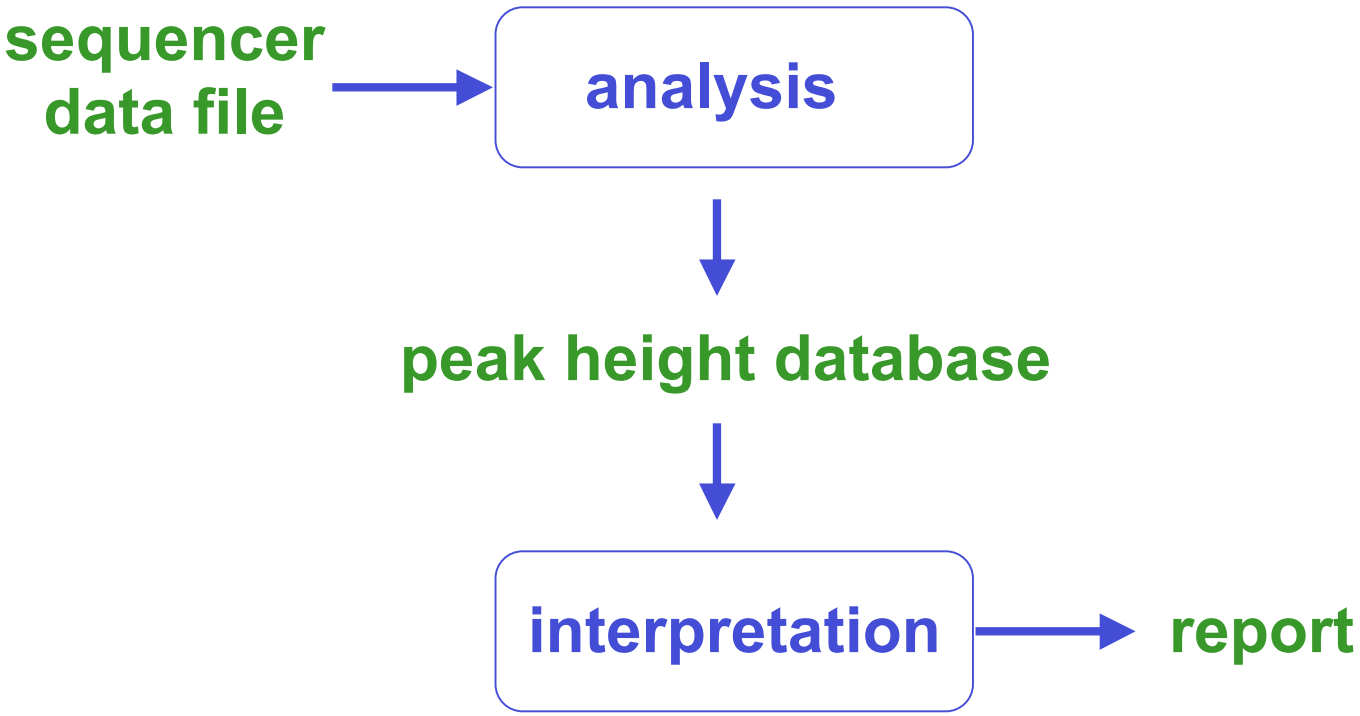
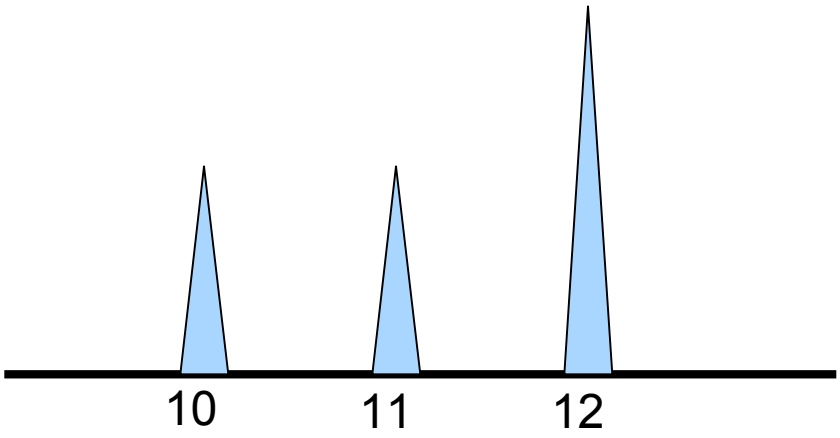


Figure 3. The TrueAllele® process flow for automated analysis and interpretation.



Conservative		**Cautious**		**Exact**	
10	10	10	12	10	12
10	11	11	12		
10	12	12	12		
11	11				
11	12				
12	12				

Figure 4. A simplified three peak mixture example. Different degrees of interpretation lead to different levels of discriminating power.

Tables

Set 1	ng amplified			
Mixture ratio	1	0.5	0.25	0.125
10:0	A1	A2	A3	A4
9:1	B1	B2	B3	B4
7:3	C1	C2	C3	C4
5:5	D1	D2	D3	D4
3:7	E1	E2	E3	E4
1:9	F1	F2	F3	F4
0:10	G1	G2	G3	G4

Set 2	ng amplified			
Mixture ratio	1	0.5	0.25	0.125
10:0	H1	H2	H3	H4
9:1	I1	I2	I3	I4
7:3	J1	J2	J3	J4
5:5	K1	K2	K3	K4
3:7	L1	L2	L3	L4
1:9	M1	M2	M3	M4
0:10	N1	N2	N3	N4

Table 1. The TrueAllele® validation study design. The three validation dimensions are mixture ratio, serial dilution and contributor pair.

lab	seq	gel	lane	sample	panel	locus	pixel	desig	height	area
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5348	7.2	6.48	2.09
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5363	8.0	10.30	3.79
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5391	9.0	38.84	20.38
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5407	9.2	19.23	9.18
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5420	10.0	659.11	361.34
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5449	11.0	645.33	353.79
CYB	ABI310	Series_C	4	C1	PowerPlex16	CSF1PO	5478	12.0	542.75	306.56

Table 2. Some data in the TrueAllele® analyzed 'peak' database table. Each record shows source, sizing and quantitation information for one peak.

lab	name	specimen	cutting	prep	pcr	seq	gel	lane
CYB	A1C1	A1	1	standard	standard	ABI310	Series_C	2
CYB	A1C1	C1	1	standard	standard	ABI310	Series_C	4

Table 3. Two specimen lanes in the user's 'request' database table. These records specify the data used in the mixture case example.

		template	
		A1	C1
contrib	1	1.00	0.68
	2	0.00	0.32

Table 4. The results of querying the 'weight' database table. Each PCR template *column* shows the ratio of underlying contributors in the specimen. Conversely, a contributor *row* shows how much that contributor is contained within each specimen.

locus	allele1	allele2	probability
CSF1PO	12.0	12.0	1.00
D13S317	9.0	13.0	1.00
D16S539	9.0	12.0	1.00
D18S51	13.0	15.0	1.00
D21S11	30.0	31.0	1.00
D3S1358	16.0	17.0	1.00
D5S818	12.0	12.0	0.99
D7S820	10.0	10.0	1.00
D8S1179	8.0	11.0	1.00
FGA	21.0	22.0	1.00
Penta_D	12.0	14.0	1.00
Penta_E	7.0	14.0	1.00
TH01	9.0	9.3	1.00
TPOX	8.0	8.0	1.00
vWA	15.0	18.0	1.00

Table 5. The results of querying the 'genotype' database table after the TrueAllele® system interprets the example case data. The genotypes (pairs of designated alleles) and probabilities are shown for each locus. With ambiguous results, a locus would display more than one row. In this case, at the 99% level, the genotype results are unique, so there is only one row for each locus.

	TrueAllele® Interpretation		**Human Interpretation**	
locus	**allele 1**	**allele 2**	**allele 1**	**allele 2**
CSF1PO	12	12	12	
D13S317	9	13	9	
D16S539	9	12	9	12
D18S51	13	15	13	15
D21S11	30	31	30	31
D3S1358	16	17	16	17
D5S818	12	12		
D7S820	10	10	10	
D8S1179	8	11	8	11
FGA	21	22		22
TH01	9	9.3	9	9.3
TPOX	8	8		
vWA	15	18	15	18

Table 6. A comparison of computer and (doubly reviewed) human reporting on the example case. The computer designates all 26 alleles, while the conservative human review designates only 18 of the alleles. While both answers are correct, the conservative approach loses much information, as measured in discriminating power.