AN EXPERT SYSTEM FOR SCORING DNA DATABASE PROFILES

Mark W. Perlin, Ph.D., MD
Cybergenetics, Pittsburgh, PA

Abstract
Forensic DNA databases are becoming an increasingly valuable law enforcement tool for convicting repeat offenders and exonerating the innocent. However, constructing such databases is quite laborious. After generating STR profiles in the lab, people expend even greater effort visually reviewing the data before it enters the database. All artifacts must be detected, and no error can be tolerated. With millions of forensic samples to be analyzed, this bottleneck has become a formidable task.

We have developed software analysis methods that can automate this data review and potentially eliminate 90% of the work. Our fully automated TrueAllele™ system inputs raw fluorescent DNA sequencer (gel or capillary) files, processes the gel image (separating colors, tracking and sizing lanes), and analyzes the STR experiments (quantitating and sizing peaks, comparing with ladder peaks, calling alleles). For each allele call, TrueAllele™ assigns a quality score and applies artifact detection rules. These quality checks enable a user to focus on just the 10% of suspect data, thereby eliminating most of the review effort.

TrueAllele™ models every step of the STR data generation process. By computing hundreds of variables for each genotype, the system can compare the observed data against expected behavior. Large deviations between expected and observed enable the software to flag potentially problematic data for human review.

TrueAllele™ automatically extracts the information it needs from the raw data (e.g., ABI/377 collection files). After automated image processing, the software provides a user quality assurance review for assessing data runs (lane tracking, gel quality, control lanes, etc.). Automated data processing continues with peak quantitation, allele designation, and quality scoring. In the allele-based quality assurance review, the user focuses on those designations which TrueAllele™ has flagged as having specific problems. After reviewing (and possibly editing) this small subset of data, TrueAllele™ then generates files for submission to the DNA database (e.g., CODIS).

TrueAllele™ operates independently of DNA sequencer manufacturer or technology, and runs on all major computer platforms (Macintosh, Windows, and UNIX). The program can analyze any panel of STR loci, allelic ladders, or internal size standards. Evaluation software is available from “www.cybgen.com”.

The TrueAllele™ expert system is designed to provide an automated computer-based “second scorer” for STR database profiles. The British Forensic Science Service (FSS) has selected TrueAllele™ automated scoring for scaling up the UK National DNA Database. We anticipate that TrueAllele™ will have a role in significantly reducing the human review of forensic STR

data, assessing the quality of STR data and providing laboratory feedback, and fully automating the forensic STR data scoring process.

Introduction

Convicted offender DNA databases are being developed by governments for forensic applications (1, 2). Much like fingerprints, these databases permit the police to link crime scene evidence to a likely candidate offender. Such "cold hits" can identify perpetrators who have no other apparent connection to the crime scene. In this way, rapists and other violent criminals can be revealed and convicted, even when their earlier profiled offenses were less violent (3, 4).

DNA databases are effective because of the long-term stability of the deoxyribonucleic acid ("DNA") molecule, nature's way of storing and transmitting genetic information (5). Like computer information, genes are encoded as linear text in an alphabet, formed from four DNA letters. Most cells in the human body contain two complete copies (maternal and paternal) of the human genome in DNA form. Sampling from the three billion letter human genome, forensic scientists typically examine about a dozen sites ("loci"), examining a few hundred DNA letters at each locus.

For human identity, scientists use short tandem repeat ("STR") loci (6). Each STR locus exhibits variation in DNA molecule length. One person will inherit two specific lengths from their parents, which is likely to be different from the pair of lengths of another person. These likelihoods multiply. Therefore, when (for example) ten loci are used, it is extremely improbable that the 20 numbers (i.e., 10 length pairs) from one individual will identically match the 20 numbers of an unrelated individual. This uniqueness serves as a "fingerprint" of genetic identity.

To form a DNA profile, scientists generate and analyze STR data. For DNA databasing, such data is derived from a blood (or other) sample taken from the convicted offender. Since no error can be tolerated, the quality assurance ("Q/A") must be extremely high. This leads to a bottleneck in the review and scoring of STR data. After first elaborating on the labor-intensive data scoring problem, this paper will describe our computer-based expert system solution. Before concluding, we will touch on some new approaches to automated DNA analysis in casework applications.

The Data Scoring Problem

In nature, STR locus of an individual has two "alleles," each corresponding to a true DNA fragment length. During laboratory data generation, the forensic scientist conducts experiments to transform these unknown DNA lengths into observable data. Later on, during data analysis, the scientist must somehow reliably transform these data back into the actual alleles.

The data transformation process is shown in Figure 1 (left-hand column). In summary:

1. Perform polymerase chain reaction ("PCR") amplification on the DNA sample to transform the STR lengths into PCR products. PCR makes millions of DNA copies of each fragment allele.

2. Size separate the amplified PCR products on a DNA sequencer to form electrophoretic bands. The locations of these bands are related to their size.
3. Detect the bands to acquire data as pixels in an electronic data image. This detection generally measures fluorescently labeled DNA fragments.

This laboratory process is not perfect. As shown in Figure 1 (right-hand column), data artifacts can be introduced at every step of the data generation process. There are dozens of potential artifacts. Some include:

- The PCR process can introduce PCR stutter or preferential amplification (also called relative amplification, or heterozygote imbalance), or amplify low-level contaminating DNA material.
- The size separation introduces peak spread, and crosstalk from neighboring lanes can produce unwanted peaks. There are no guarantees that the STR allele molecules will migrate reproducibly relative to internal size standards.
- The data acquisition may exhibit shifts in the baseline, color bleedthrough from other dyes, and other size distortions.

The human data analysis reverses the laboratory processes. Starting from the observed data pixels, the scientist tries to infer the lengths of the STR alleles (Figure 2, left-hand side). In summary:

1. First the operator recovers DNA signals from the observed data pixels in order to identify electrophoretic bands (i.e., peaks).
2. Then, these bands are sized and quantitated to estimate the relative amount of DNA present in each PCR product fragment.
3. From these quantitative measurements, the scientist designates the alleles of the PCR products, thereby inferring the lengths of the true alleles.

Each analysis step serves to invert its corresponding data generation step. The potential data imperfections (or, "artifacts") are well known. Therefore, in each step the analyst scrutinizes the data, checking for these artifacts. Human data analysts perform these quality assurance steps when reviewing the data and editing the results (Figure 2, right-hand side).

Such high-quality STR data scoring can be quite labor intensive (Figure 3). Following current quality assurance guidelines, two skilled analysts independently score the data. After that, the discrepancies are reviewed in order to reach concordance. This review may be done together with more experienced supervisory personnel. This process assures the quality of the data prior to submission to a DNA database.

It is expensive to have so many people involved in the data analysis process. One can develop a spreadsheet to compute these costs (Figure 4); the Excel file shown is downloadable from "www.cybgen.com". The right side of the spreadsheet determines the number of people needed, based on the required throughput and other assumptions:

- There are roughly fifteen million felony arrests in America every year. To upload this one year of DNA profiles onto the US database over a five year period would require processing three million samples each year. Using the current CODIS loci, sixteen STR experiments

must be performed and analyzed. This entails scoring about 48,000,000 genotypes each year.

- Suppose that one person can completely review one STR locus experiment every minute (i.e., about 500 calls/person). With double scoring, over 750 people would be needed. Note that it is not easy to recruit, train, or retain such highly skilled forensic analysts.

The left side of the spreadsheet looks at the labor costs:

- Assume a low estimate of $25,000 as an analyst's annual salary.
- Accounting for overhead, training, supervision, benefits, equipment, and other personnel costs, this figure is increased by 100% to 200%. With the assumptions shown, the actual annual cost becomes $64,000.
- Multiplying the actual personnel cost times the number of personnel required, we arrive at an annual cost of about $50,000,000 for scoring the data.
- Adding up over the five year project, the total cost of scoring **one year** of US felony arrests is **one quarter of a billion dollars**.

Thus, one STR locus experiment costs about a dollar for data generation, and an equal dollar for quality assurance in the data scoring. Using about fifteen loci per DNA profile, that $30 cost, added to another $15 in administrative costs, gives the current commercial rate of $50 per sample for DNA database processing.

It would be desirable to reduce this cost, speed up the process, obtain more accurate results, and achieve a completely objective, impartial data review. This can be done by building a computer-based expert system that replicates the quality assurance procedures of a human STR review expert (Figure 5). Once validated, such an expert system could replace the tedium of the human scorer with a tireless computer program. We have developed an intelligent automated system for this task, as described in the next section.

An Expert System Solution

TrueAllele™ is a flexible, automated DNA analysis technology (Figure 6). The TrueAllele™ computer program performs exquisitely accurate sizing and quantitation of DNA fragment data. It does this by mathematically modeling every step of the data generation process, and then performing the analysis via a succession of computational inversion operations (7). At each step, the models know what to expect from the data. Therefore, automated comparison of the expected results with the observed data provides an explicit quality assurance mechanism for assessing results and pinpointing potential data artifacts. Although TrueAllele™ is used for diverse genetic analyses, this paper focuses solely on automated STR analysis for human identity.

TrueAllele™ can read data in from any gel or capillary DNA sequencing instrument (Figure 6). It performs all necessary image and signal processing on the data for quantitative analysis. Further downstream, the program designates the alleles, and assigns quality measures to every designation. In forensic analysis, the software applies dozens of rules, each of which checks for a particular quality assurance problem. TrueAllele™ can run on any standard high-end computer, including the Macintosh®, Windows®, and UNIX® platforms.

The result of all this fully automated analysis is that the computer determines which data are good, and which are not. A human analyst can then focus on just the 10% of problematic data. There is no need to waste time on the remaining 90% – these good data have already passed dozens of highly quantitative quality assurance tests. By focusing the user on just the small fraction of problematic data, TrueAllele™ can reduce the human scoring effort by an order of magnitude.

TrueAllele™ processing is conducted in four phases:

1. Input. A set of run data is read from its native sequencer file format, and automatically prepared for TrueAllele™ processing.
2. Run processing & Q/A. Each gel or capillary run is automatically processed (color separation, tracking, sizing, etc.). The user can then review the results in order to accept, reject, or edit a run.
3. Allelic processing & Q/A. Each experiment (one individual at one locus) is automatically analyzed (peak quantitation, allele designation, rule application, etc.). The problematic allele calls are presented to the human analyst for more careful review.
4. Output. The final high-quality allele designations are automatically formatted for export to a DNA database.

We describe each phase in turn.

1. Input

A site can set their preferences for the AutoSetup initialization phase (Figure 7). These preferences include a DataDisk template that records STR panel properties, DNA sequencer instrument, size standards, user preferences, and other run-independent information.

To analyze data, a set of runs (e.g., about 10 gel or 96-capillary runs) is placed in a designated input folder. TrueAllele™ automatically extracts all necessary STR data, calibrations, and sample information from the run files, and creates a DataDisk. This DataDisk is suitable for TrueAllele™ processing on any computer platform.

2. Run processing and Q/A

TrueAllele™ automatically processes gel and capillary runs in six steps (Figure 8):

- Acquire data. TrueAllele™ reads in the data from the native file format. These data are in "pixel" coordinates: the scan lines of the actual data acquisition.
- Process signal. Basic signal processing is done, such as baseline removal or any necessary smoothing.
- Separate colors. To analyze the multiplexed STR data, spectral color separation is required. TrueAllele™ can compute (in under a second) the separation matrix directly from the observed data. This feature lets the program customize color separation to the actual data in the capillary or on the gel.
- Remove primers. The primer peaks are stripped off the signal.
- Track sizes. TrueAllele™ spends most of its time at this stage carefully analyzing the internal lane size standards. The goal is to ensure the best possible automated lane tracking (gel data) or alignment of the expected sizes to the observed data peaks (capillary data).

- Extract profiles. Using the tracked size standards, TrueAllele™ transforms the signals from the initial pixel coordinates into a "size" (in bp) coordinate system. This size coordinate system is better suited for comparisons between lanes. The resulting set of color separated, sized profiles – one profile for each lane or capillary – is independent of sequencer type, and is suitable for downstream allelic processing.

After TrueAllele's™ automated run processing, the user can perform a visual quality assurance check. Figure 9 shows the size standard color plane of a TrueAllele™-analyzed Applied Biosystems ABI/377 gel (from the UK Forensic Science Service (FSS)) in the ImageView interface. This Q/A review user interface shows the data in the context of the program's analysis (here, the automated lane/size tracking grid). The user can optionally view TrueAllele's™ rule firings on the run quality. Typical gel run rules check the negative and positive control lanes, the ladder lanes, and look for particular data artifacts.

Other visual user interfaces are also available to the reviewer during the gel/capillary run quality assurance check. Ideally, the user should spend no more than a minute per run in their review. However, some of these less frequently used interfaces can help when exploring particular data artifacts.

- It is sometimes helpful to view the STR data planes. Figure 10 is an ImageView of STR data and allelic ladders from a TrueAllele™-analyzed Hitachi/FMBIO II gel (from Dr. Cecelia Crouse's lab in the Palm Beach County Sheriff's Office (PBSO) in Florida).
- With capillary data, one may need to review the relative alignment of the sized, color separated, one-dimensional (1D) capillary traces. Figure 11 is a view of STR data and ladders from a TrueAllele™-processed ABI/3700 capillary run (from the UK FSS).
- To check cross-capillary ladder sizing consistency, an overlay view of the ladder traces can be helpful. Figure 12 shows an allelic ladder overlay (blue) and size standards (red) from a TrueAllele™-processed Amersham/Molecular Dynamics MegaBACE™ 96-capillary sequencer run (from the FSS).
- One dimensional signal views can show possible bleedthrough problems from other dyes, lanes, or loci. Figures 8 and 13 show such 1D views from TrueAllele™-processed SpectruMedix SCE/9600 96-capillary data.

3. Allelic Processing and Q/A

Once the run Q/A has been completed, TrueAllele™ then starts its overnight allelic processing of the DataDisk containing multiple gel/capillary runs. The automated allelic analysis is done in several steps (Figure 13):

- Derive allelic ladder. The program carefully matches the expected allele sizes to the observed allelic ladder data peaks. This matching can transform each allele's fractional size (based on internal lane molecular weight standards) into its true integer-valued DNA fragment length.
- Transform coordinates. Using the derived allelic ladder, the STR data is transformed from artificial size coordinates into (the more natural) length coordinates. In length coordinates, the fractional part denotes experimental deviation from the true allele length.

- Quantitate trace. TrueAllele™ then performs a computationally intensive least-squared fit on every STR trace to accurately estimate the relative DNA concentration in each peak. This process is described more fully in the next paragraph.
- Call alleles. Applying multiple algorithms, TrueAllele™ uses the STR peak quantitations and sizes to designate the alleles.

When estimating the relative concentration of a DNA band, most analysis software simply records the observed peak height or the area immediately under the curve. For truly quantitative genetic analysis, though, this simple approach is inadequate. The best estimate models the data curve as a sum of model peak functions, each having their own location, height and widths. This approach accounts for band overlap, and other electrophoretic signal distortions. Minimizing the deviation between the model and the data provides reliably accurate estimates of the actual DNA amount for each DNA fragment (8). By expending this computational effort, TrueAllele™ can determine reproducible DNA quantities. Accurate numerical estimates are essential for robust downstream comparisons.

After calling the alleles, TrueAllele™ prepares for the user's quality assurance data review. To do this, for each STR locus experiment, TrueAllele™:
- Applies several dozen rules to check for possible data artifacts.
- Computes a quality measure on the genotype (0 is bad, 1 is great).

Following these determinations, TrueAllele™ can sort the experiments by quality. Low quality results (a rule firing, or a low quality measure score) will be ranked for earlier review. This prioritization focuses the user on the problematic data calls.

Once the computer has called the alleles, TrueAllele™ can present the processed STR results to the user in an AlleleView interface (Figure 14). The purpose of this interface is to minimize the time spent visually inspecting data. While there are many data views available from this interface (e.g., by clicking on any subwindow), each has a specialized role for a particular situation. (Spending time looking at all possible visualizations would be highly inefficient.) Key AlleleView interface elements (Figure 14) are kept visible:
- A navigator for selecting which data to view (gel, sample, locus, etc.).
- An editor for making changes (e.g., via popup menus) to designations.
- Several visualizations for checking the data signal, its quantitation, and the genotype call. To make size shifts immediately evident, the allelic ladder is displayed in the background.
- A list of fired rules. If a rule fires, then the STR experiment has failed some user-defined criterion and should be examined. The fired rules also set the context for further inspection of the data.

In AlleleView, the user reviews all interesting experiments at one locus (across all the gels) before moving on to the next locus. This ordering establishes a user context for carefully examining one locus, keeping the locus-specific patterns of the allelic ladder and other issues fixed in mind. Moreover, since all scorable gel/capillary runs in the DataDisk have passed quality assurance, one rarely needs to view the run again.

Interestingly, one could review **all** the data (not just the computer-identified problematic calls) with very little extra effort. The reviewer has tremendous prior knowledge about the "good"

experiments: each has passed 20+ quantitative rules, without triggering a low quality score. Thus, by rapidly skimming through pages of profiles (say, ten per page in the LaneView interface), and looking only for weird patterns, the AlleleView analyst can review all 90% of good data even faster than the bad subset.

4. Output

When the Q/A has been completed, TrueAllele™ can then automatically export the data for database upload. Different databases have different requirements. TrueAllele™ supports the CODIS standard, including the new XML Common Message Format (Figure 15).

The TrueAllele™ Technology

The TrueAllele™ software is written in the MATLAB programming language (The MathWorks, Natick, MA), and we develop it in a reasonably automated way (Figure 16). Based on the program design plan, and feedback from testing, the software modules are extended and updated. Building the software entails code compilation, file assembly, electronic packaging, and electronic distribution across all supported computer platforms (Macintosh, Windows, and UNIX). Done manually, this task would ccupy several error-ridden person-months. Therefore, we automated the build process. Over several hours, four computers coordinate their efforts to compile, assemble, package, and e-distribute the software for all the computer platforms. This automation frees up programmer time for more constructive software development, and permits us to generate new (and reliable) versions over very short time periods.

The Cybergenetics web site (www.cybgen.com) provides considerable support for the TrueAllele™ technology and software (Figure 17). Considering each frame:

- Company. These web pages describe Cybergenetics.
- Technology. This section provides the downloadable "LaborCost" spreadsheet, and lists recent publications. It also enumerates our patent claims on TrueAllele™. The technology is protected by broad and deep claims covering all aspects of automated STR analysis, including automated scoring, lane tracking, PCR artifact removal, quality measures, forensic applications, and high-throughput (e.g., 96-capillary) analysis.
- Software. Software updates are downloadable from this section. Visitors can also download PDF user documentation, and request additional information.
- Support. On these pages, users can report bugs, provide feedback, and obtain other useful information.
- News. Recent software and press releases are reproduced in these pages.
- Contact. This page gives instructions for contacting Cybergenetics.

TrueAllele™ is currently in use by the FSS in the UK. The UK National DNA Database will soon exceed one million DNA profiles, and is expected to be augmented with several million additional profiles. The model (Figure 18) is to virtually eliminate the human STR review. Using diverse DNA sequencers (both gel and capillary), the TrueAllele™ system will process the generated data, designating alleles and assessing data quality. For extra assurance, the computer will check its results against the FSS's own STRess expert system (9). In the near-term, a human forensic reviewer will check the 10% of computer-identified problematic calls.

Longer-term, human intervention will be eliminated entirely. The automatically processed STR calls will then be uploaded to the UK National DNA Database.

We recently began a Justice Department funded project to validate TrueAllele™ on diverse American STR data. The project will automatically rescore 30,000 previously analyzed CODIS samples, and assess the quality of TrueAllele's™ results.

- Florida. In collaboration with David Coffmann's group at the Florida Department of Law Enforcement (FDLE) in Tallahassee, we will reanalyze 15,000 samples generated on ABI/310 and ABI/3700 capillary DNA sequencers using the Applied Biosystems ProfilerPlus and Cofiler STR panels.
- Virginia. In collaboration with Dr. Paul Ferrara's group at the Division of Forensic Science in Richmond, we will reanalyze 15,000 samples generated on an Hitachi FMBIO2 gel system using the Promega PowerPlex® 1.1 and 2.1 STR panels.

The goal is to demonstrate that TrueAllele™ is a platform-independent solution for automating the scoring of STR data in forensic DNA databases.

Complex DNA Analysis

The future of DNA databases lies in their application to solving crimes in the casework laboratory. Toward this end, we have developed automated methods for resolving data artifacts. These include:

- PCR stutter. The PCR stutter of an STR locus is reproducible (10). Therefore, one can calibrate and mathematically remove the artifact by stutter deconvolution (11). The deconvolved signal is far easier to interpret than is the original data.
- Relative amplification. Similarly, one can model and account for preferential amplification (i.e., heterozygote imbalance) (7).
- Band overlap quantitation. When quantitating DNA peaks, the overlap of neighboring bands can be modeled for more precise quantitation (8). We have been developing newer mathematical results that may prove far faster than our current search-based implementations.
- DNA mixtures. We have developed new methods for resolving DNA mixtures. These are discussed in the next paragraph.

We solve these problems by mathematically modeling the natural data generation process. Then, mathematical inversion lets us apply our computer programs to recover the true alleles from the artifact-distorted data.

In casework, mixed DNA samples can complicate the analysis, and confuse a jury. When performing database matches from DNA mixtures, even simple two-person mixtures can suggest 3 or 4 alleles at many loci, leading to thousands of spurious matches. It would be useful to have an entirely objective approach to automatically resolving DNA mixtures, thereby finding the unique DNA profile. We have started developing such techniques.

A representative mixture analysis is shown (Figure 19). The data were generated in our laboratory using the ten locus SGM Plus panel on a sample we prepared containing 30% of individual A (say, the victim), and 70% of individual B (say, the perpetrator). The problem was to determine the profile of perpetrator B entirely from the given data.

1. The first row shows the quantitative mixture STR profile A+B. This represents a mixture of two individuals – a known victim A, and an unknown perpetrator B. At each locus, the bar chart displays the relative quantitative data for the relevant alleles.
2. The second row shows the STR profile of the victim A.
3. The third row shows the computed STR profile for B, the unknown perpetrator. This was computed automatically in under 0.1 sec from the quantitative data for mixture A+B, and the known profile A. A's mixture fraction was estimated to be 29.6%.

These results, and more recent extensions, suggest that automated mixture analysis is a feasible technology.

Conclusion

Currently, virtually all steps in the STR profiling pipeline are automated.

1. robots prepare DNA samples;
2. thermocyclers enable robust and routine PCR;
3. automated DNA sequencers separate DNA fragments by size;
4. armies of people agonize over the laborious high-quality scoring of STR data; and
5. computers instantly link criminals to crimes via DNA profiles.

The TrueAllele™ expert system addresses step 4.

TrueAllele™ automates the last remaining manual step in forensic DNA databasing – scoring the data. This technological advance enables full automation of the entire STR process. We expect that TrueAllele™ will help reduce the time, cost, and complexity of building national DNA databases of convicted offender profiles.

Acknowledgements

Meredith Clarke and Michael Breen of Cybergenetics are key members of the TrueAllele™ development team. The Forensic Science Service supplied much of the multi-platform Applied Biosystems and Amersham data used here. Dr. Cecelia Crouse of the PBSO provided the Hitachi data. I would particularly like to thank the Promega Corporation for inviting us to present our work at the Eleventh International Symposium on Human Identification.

References

1. Gill P, Urquhart A, Millican ES, Oldroyd NJ, Watson S, Sparkes R, et al. Criminal intelligence databases and interpretation of STRs. Advances in Forensic Haemogenetics 1996;6:235-42.

2. McEwen JE. Forensic DNA data banking by state crime laboratories. Am. J. Hum. Genet. 1995;56:1487-92.

3. Blumstein A, Cohen J, Das S, Moitra SD. Specialization and seriousness during adult criminal careers. J. of Quantitative Criminology 1988;4(4):303-45.

4. McCue C, Smith GL, Diehl RL, Dabbs DF, McDonough JJ, Ferrara PB. Criminal histories of sex offenders identified through DNA "cold hits". submitted 2000.

5. Watson JD, Gilman M, Witkowski J, Zoller M. Recombinant DNA. second ed. New York, New York: W.H. Freeman and Company; 1992.

6. Weber J, May P. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. 1989;44:388-96.

7. Ng S-K. Automating computational molecular genetics: solving the microsatellite genotyping problem [Doctoral dissertation]: Carnegie Mellon University; 1998.

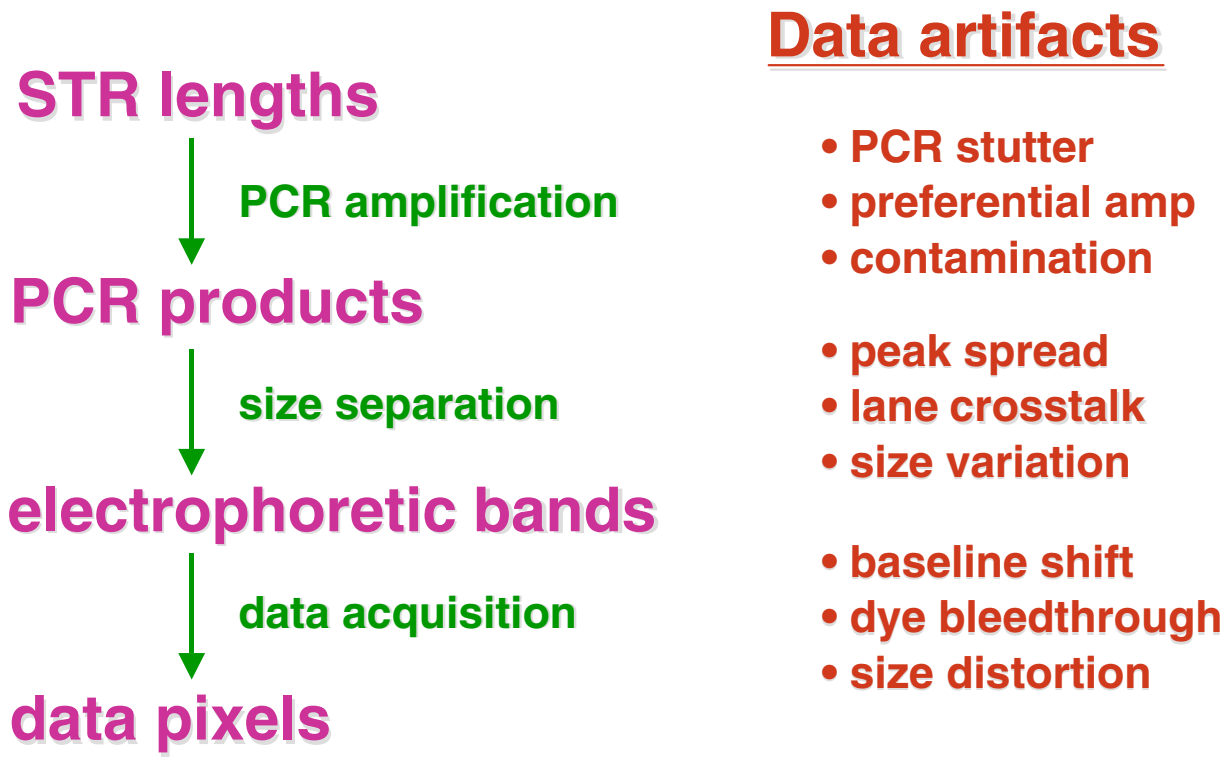
8. Richards DR, Perlin MW. Quantitative analysis of gel electrophoresis data for automated genotyping applications (Abstract). Amer. J. Hum. Genet. 1995;57(4 Supplement):A26.

9. Werrett D, Pinchin R, Hale R. Problem solving: DNA data acquisition and analysis. Profiles in DNA 1998;2(1).

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):777-87.

11. Perlin MW, Lancia G, Ng S-K. Toward fully automated genotyping: genotyping microsatellite markers by deconvolution. Am. J. Hum. Genet. 1995;57(5):1199-210.

Figure 1. Generate STR profiles. The data generation process successively transforms the STR lengths found in nature into the observed data pixels (left-hand side). In the process, data artifacts are introduced (right-hand side).



°

Figure 2. Analyze STR profiles. The data analysis process inverts every step of the data generation process (left-hand side). At each step, human data editors check for known artifacts to assure quality results (right-hand side).

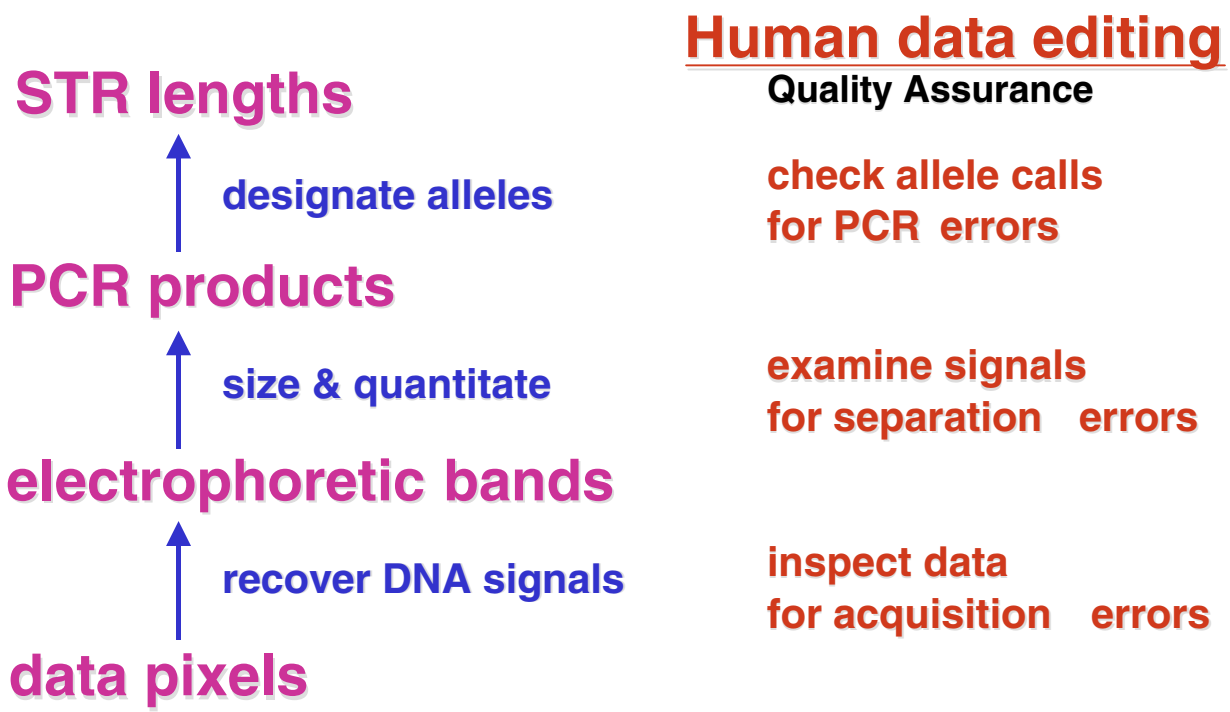
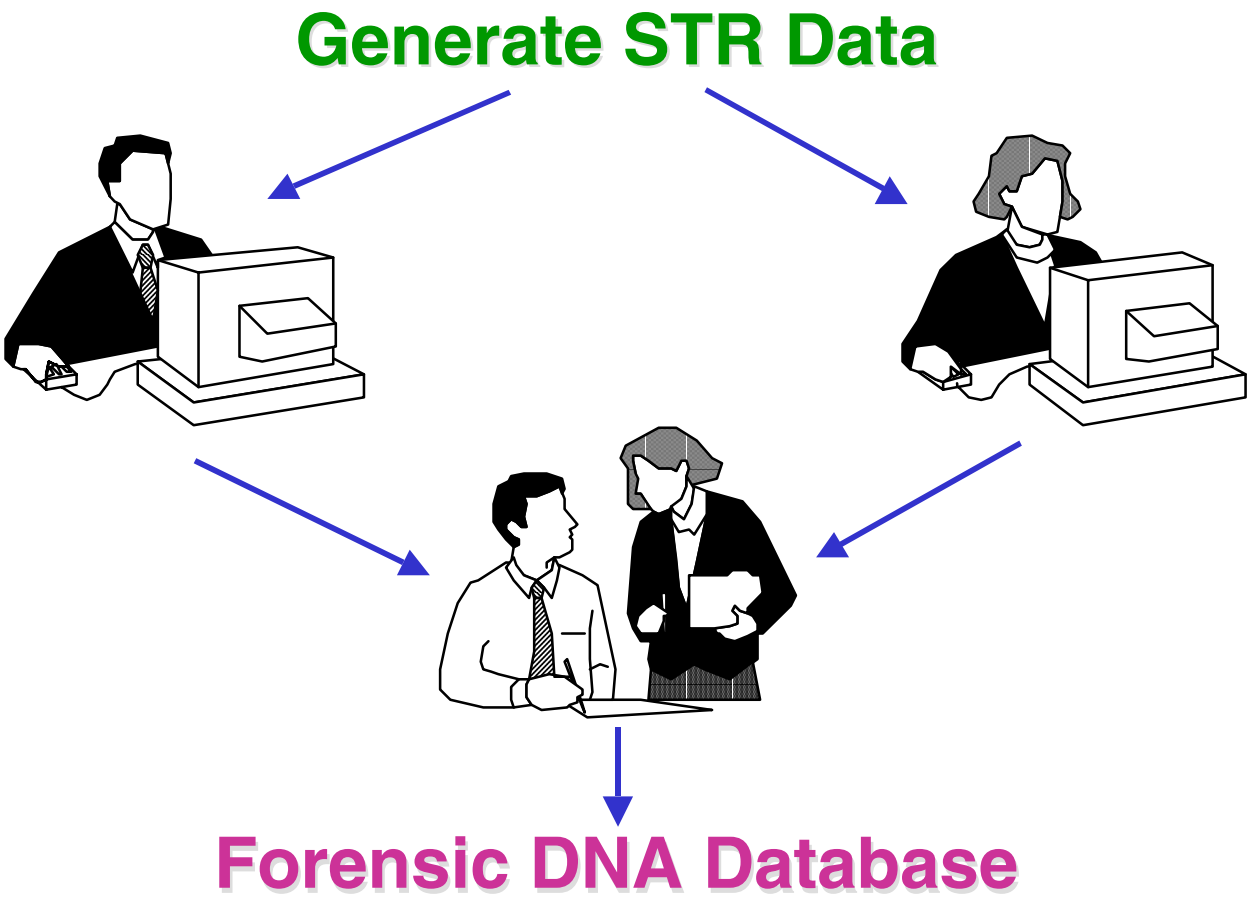
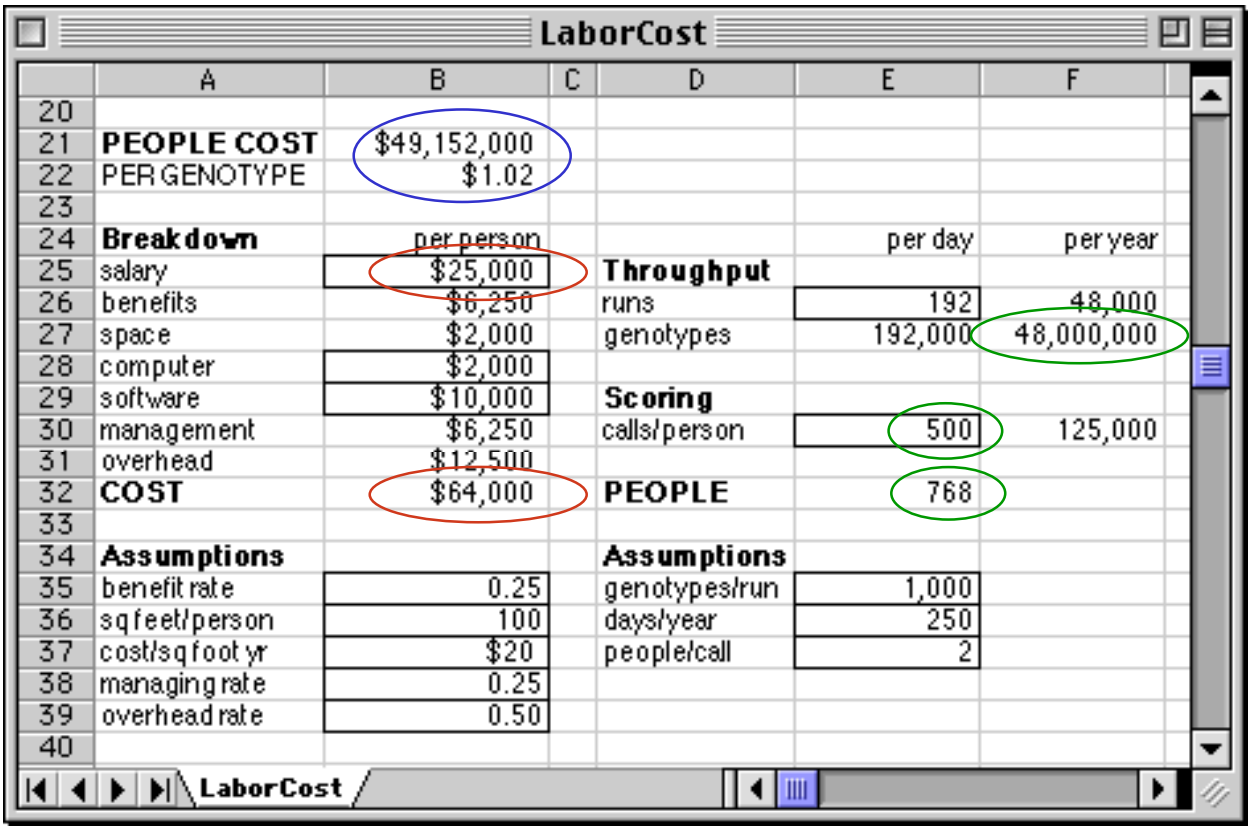


Figure 3. High-quality STR data scoring. For quality assurance, two people typically score the STR data (with additional review) prior to uploading the results onto a forensic DNA database.



°

Figure 4. Data analysis: labor cost. One can compute the actual fully-overheaded labor cost of scoring STR data, based on throughput and other assumptions. High-quality data analysis costs are comparable with the data generation costs, typically exceeding $1 per genotype locus. This spreadsheet is downloadable from "www.cybgen.com".



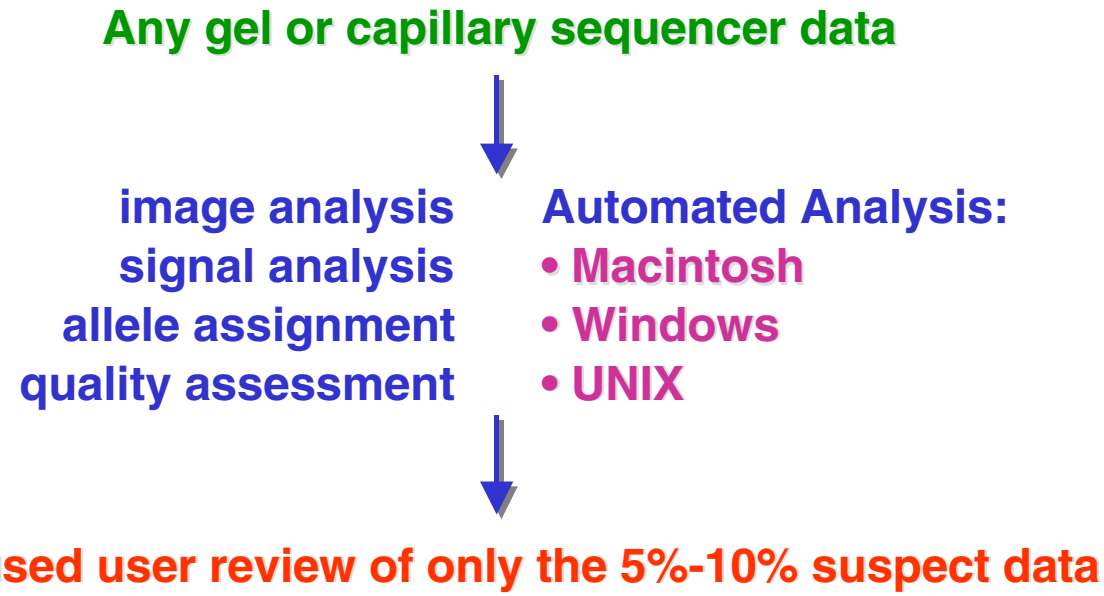
	A	B	C	D	E	F
20						
21	**PEOPLE COST**	$49,152,000				
22	PER GENOTYPE	$1.02				
23						
24	**Breakdown**	per person			per day	per year
25	salary	$25,000		**Throughput**		
26	benefits	$6,250		runs	192	48,000
27	space	$2,000		genotypes	192,000	48,000,000
28	computer	$2,000				
29	software	$10,000		**Scoring**		
30	management	$6,250		calls/person	500	125,000
31	overhead	$12,500				
32	**COST**	$64,000		**PEOPLE**	768	
33						
34	**Assumptions**			**Assumptions**		
35	benefit rate	0.25		genotypes/run	1,000	
36	sq feet/person	100		days/year	250	
37	cost/sq foot yr	$20		people/call	2	
38	managing rate	0.25				
39	overhead rate	0.50				
40						

°

Figure 5. Computer data scoring. One solution to the labor bottleneck is using an expert system computer program (right-hand side). Such software can be faster, more accurate, and more objective than a labor-based approach.

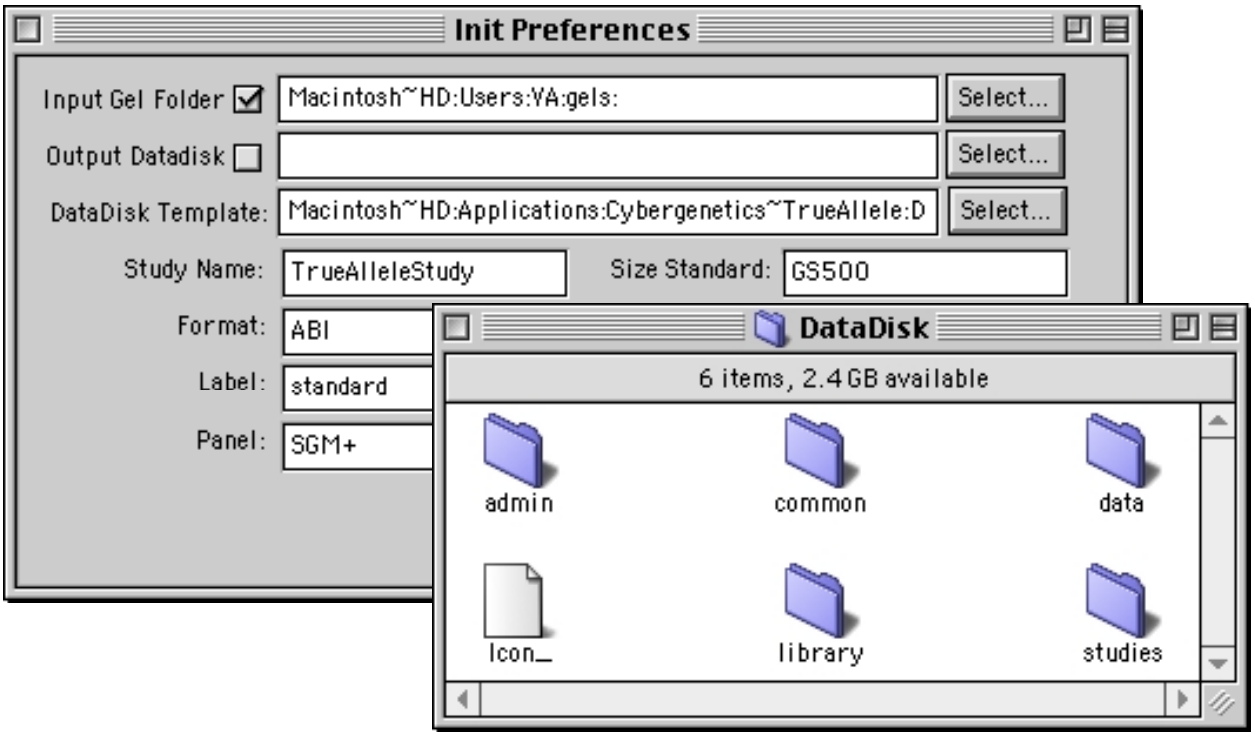
°

Figure 6. TrueAllele™ automation. TrueAllele™ is an expert system technology that automates the STR scoring and quality assurance process. It provides a flexible automated analysis for many DNA fragment sizing and quantitation applications, including forensic databasing. The process is independent of DNA sequencer, STR panel, or computer system.



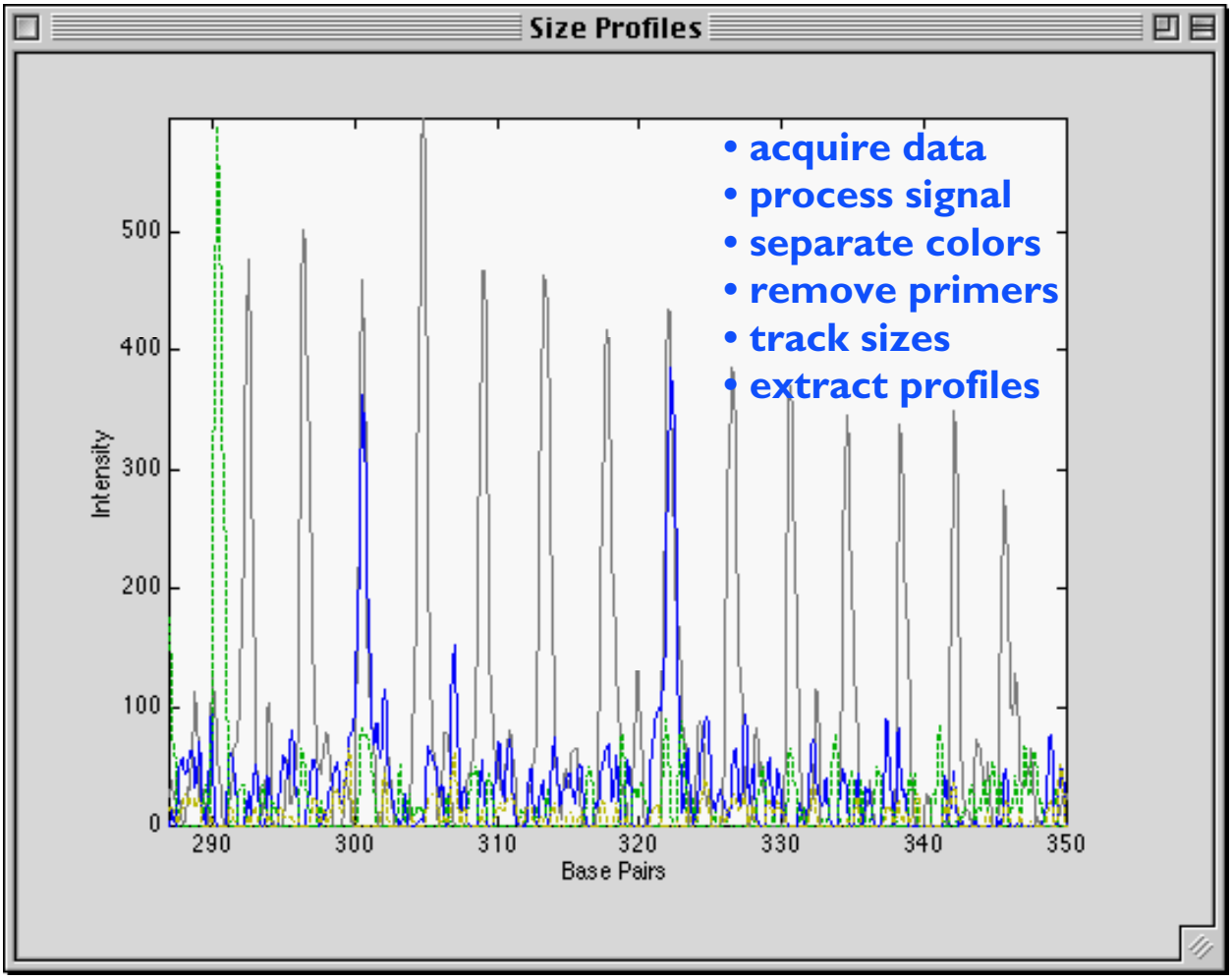
°

Figure 7. (1) Input. The transformation of raw sequencer data into TrueAllele™-processable form is automated by the program's AutoSetup module.



。

Figure 8. (2) Run processing & Q/A. The steps performed for initial processing of gel or capillary data are shown. Also shown are the results of performing this processing on data from a SpectruMedix SCE/9600 high-throughput capillary instrument.

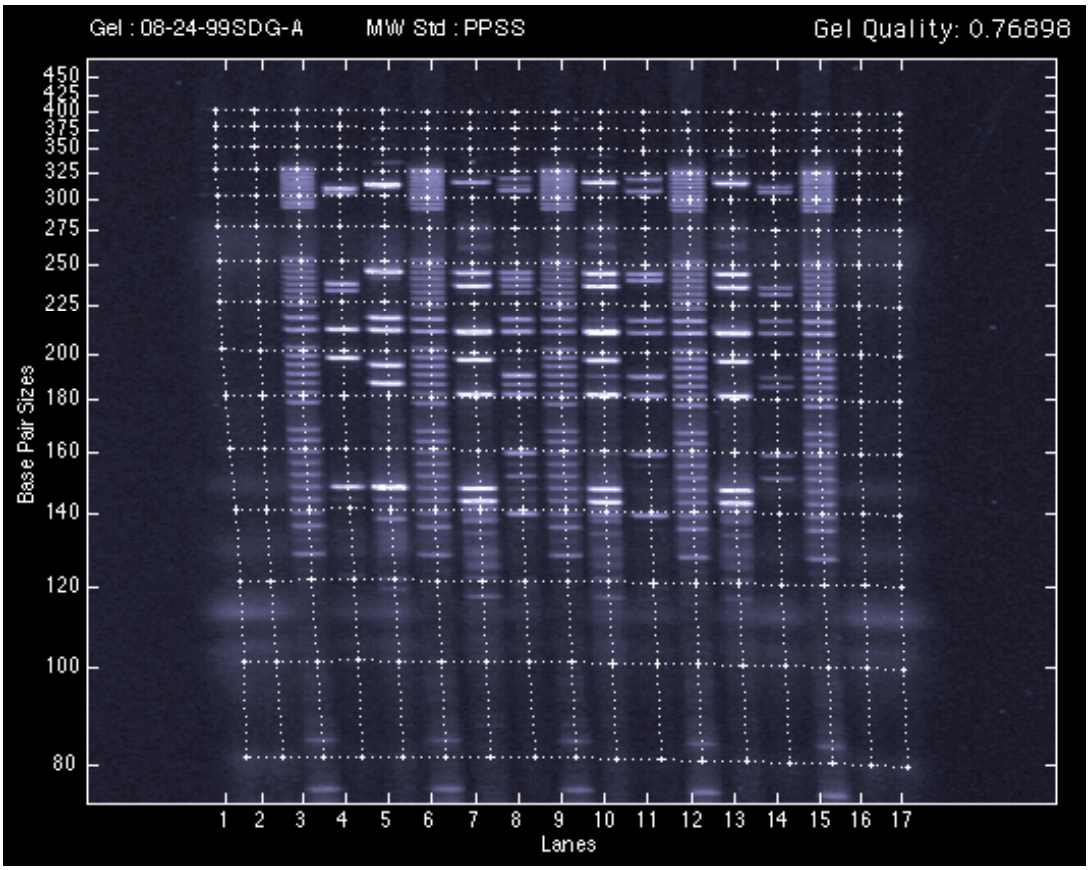


°

Figure 9. ImageView user interface. For visual quality assurance, ImageView is applied after TrueAllele's™ run analysis has completed. Starting from rule-based quality checks, the user can assess the run data quality, and decide whether to accept, reject, or edit the results. Shown is the ROX size standard plane of an ABI/377 gel from the FSS.

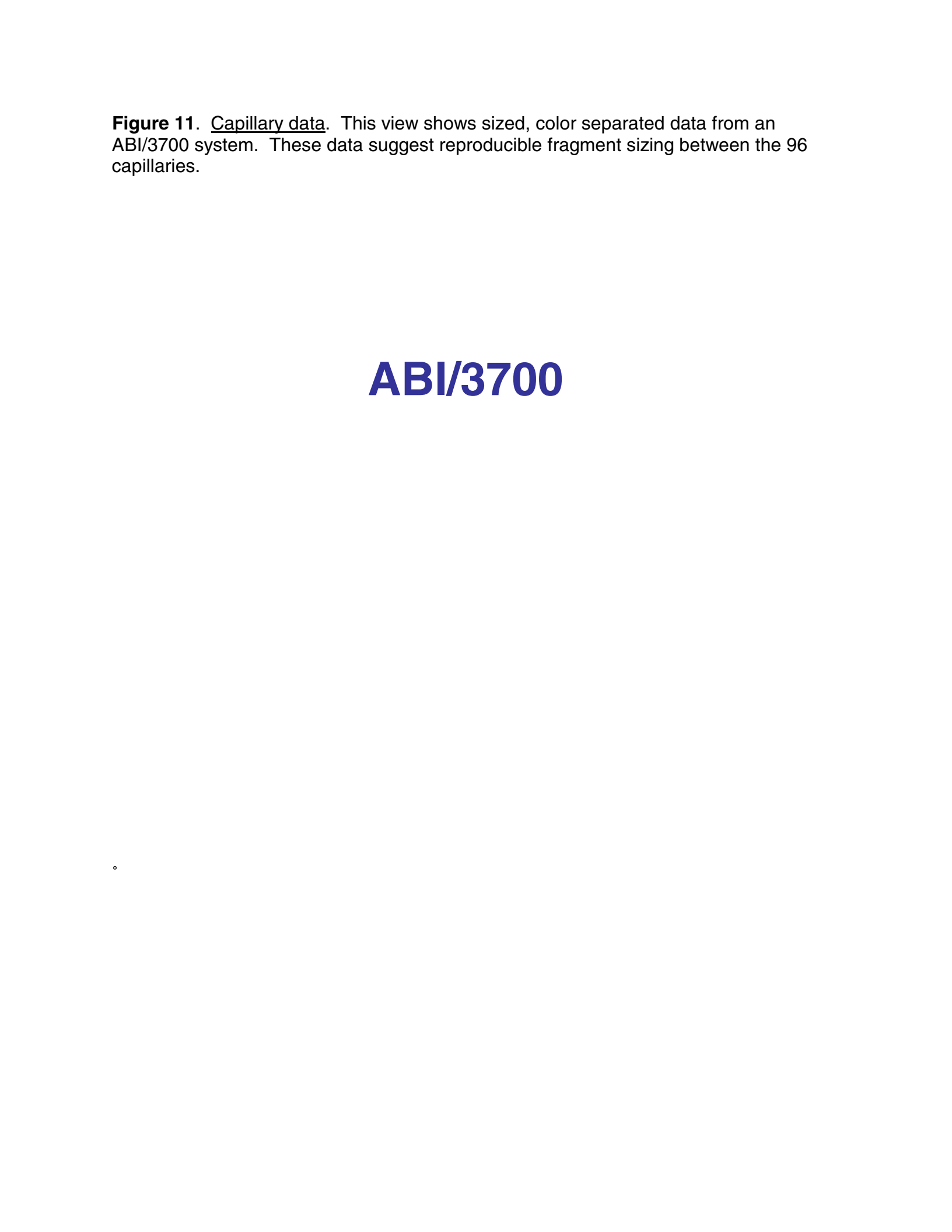
°

Figure 10. ImageView STR data. TrueAllele's™ automatic lane tracking size standard grid is shown, superimposed on STR data and ladders. The data are from an Hitachi FMBIO II gel provided by Dr. Cecelia Crouse.



°

Figure 11. Capillary data. This view shows sized, color separated data from an ABI/3700 system. These data suggest reproducible fragment sizing between the 96 capillaries.



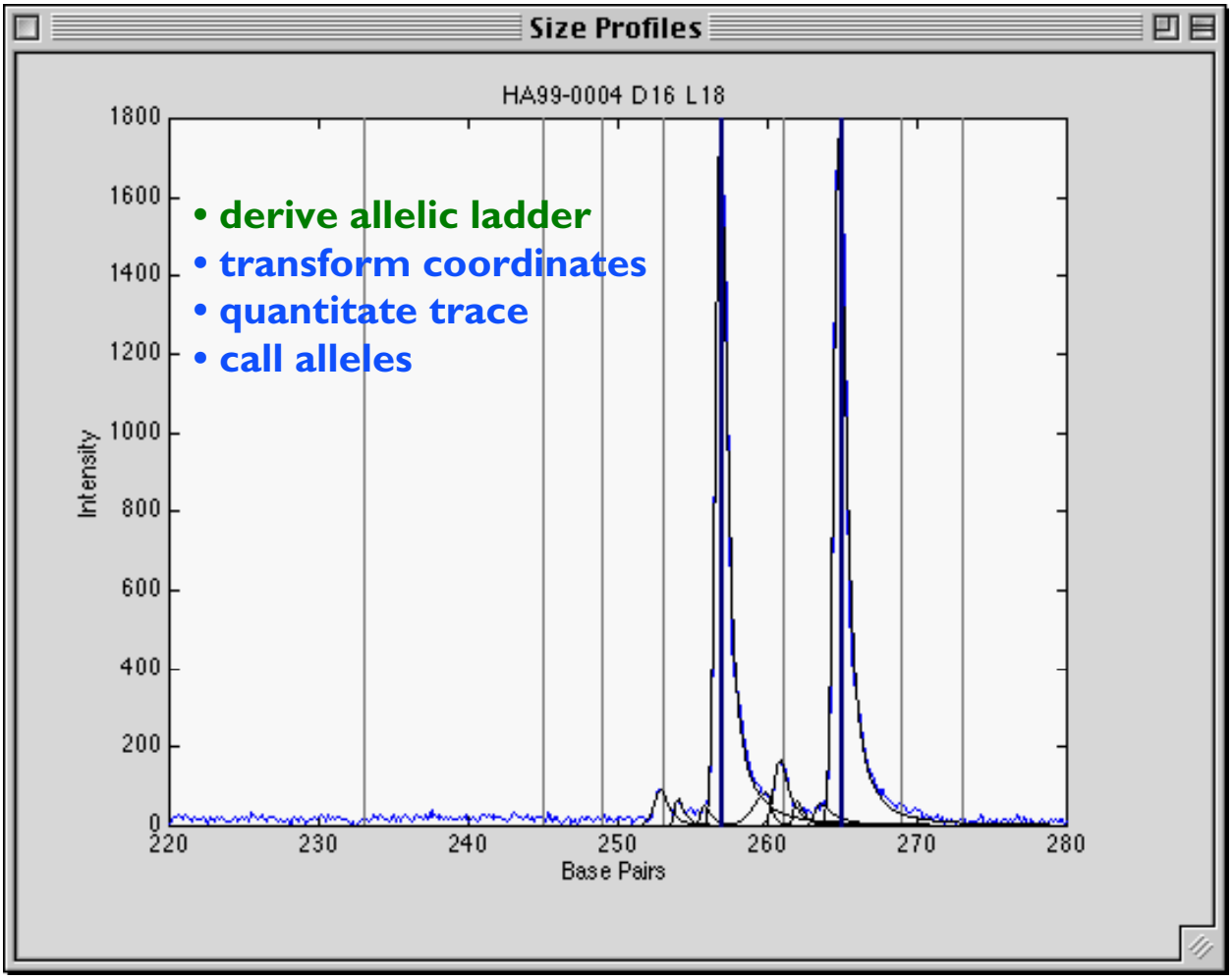
°

Figure 12. Signal overlays. This LadderView shows the overlay of six allelic ladders (blue curves), and demonstrates the sizing reproducibility of the Amersham MegaBACE 96-capillary sequencing instrument.

MegaBACE

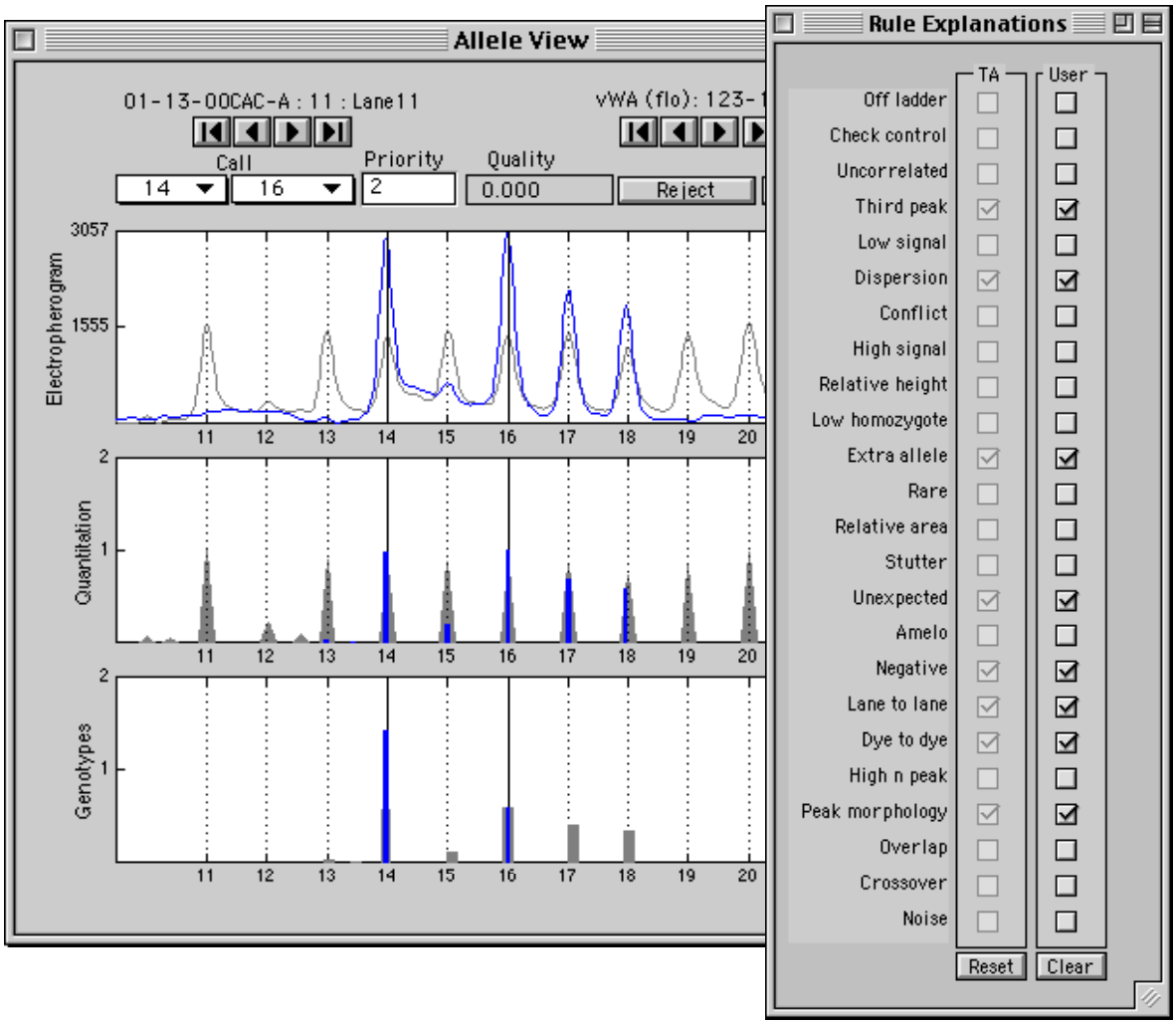
°

Figure 13. (3) Allelic processing & Q/A. The four steps of TrueAllele's™ allelic processing are shown, together with the results of such processing. The data are from a SpectruMedix SCE/9600 96-capillary sequencing instrument.



。

Figure 14. AlleleView. The AlleleView navigator window is the primary data review mechanism in TrueAllele™. It is typically viewed together with the rule firings window, which sets the context for examining problematic data.



。

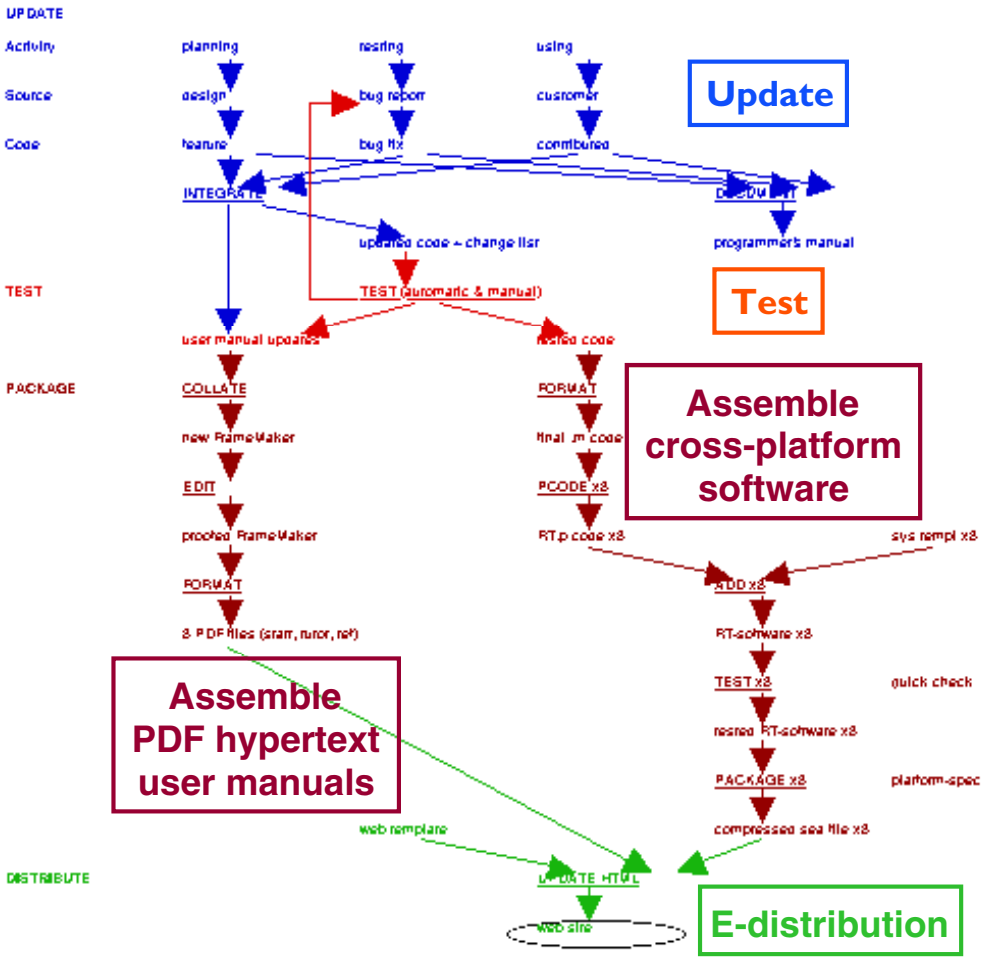
Figure 15. (4) Output. After TrueAllele™ processing, a file is generated for uploading to DNA databases. Shown is one output format: CODIS CMF 2.0 results, with XML text (left-hand side) and web browser display (right-hand side).

```
<PCRREADING>
       <PCRLANE>
              <PCRLANENUMBER>1</PCRLANENUMBER>
              <LOCUS>
                     <LOCUSNAME>D16</LOCUSNAME>
                     <PCRALLELE>
                            <PCRVALUE>  11</PCRVALUE>
                            <PCRVALUE>  11</PCRVALUE>
                     </PCRALLELE>
              </LOCUS>
              <LOCUS>
                     <LOCUSNAME>D2</LOCUSNAME>
                     <PCRALLELE>
                            <PCRVALUE>  17</PCRVALUE>
                            <PCRVALUE>  25</PCRVALUE>
                     </PCRALLELE>
              </LOCUS>
              <LOCUS>
                     <LOCUSNAME>D3</LOCUSNAME>
                     <PCRALLELE>
                            <PCRVALUE>  15</PCRVALUE>
                            <PCRVALUE>  15</PCRVALUE>
                     </PCRALLELE>
              </LOCUS>
```

```
1
  D16
    11
    11
  D2
    17
    25
  D3
    15
    15
  VW
    14
    17
  AMELO
    X
    X
  D18
    14
    16
  D21
    28
    31.2
  D8
    10
    12
  D19
    13
    13.2
  FG
    21
    22
  TH
    6
    8
```

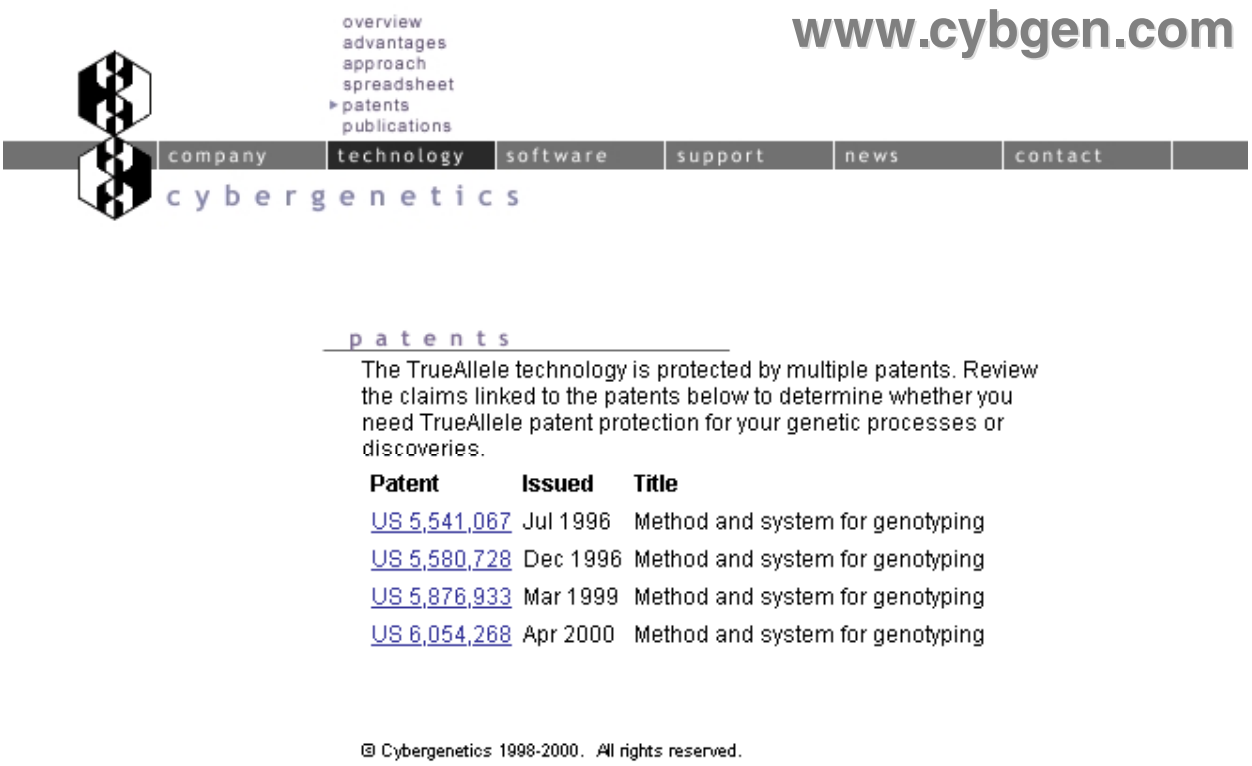
。

Figure 16. Automated build process. The TrueAllele™ software construction is highly automated, as outlined in the flow diagram.



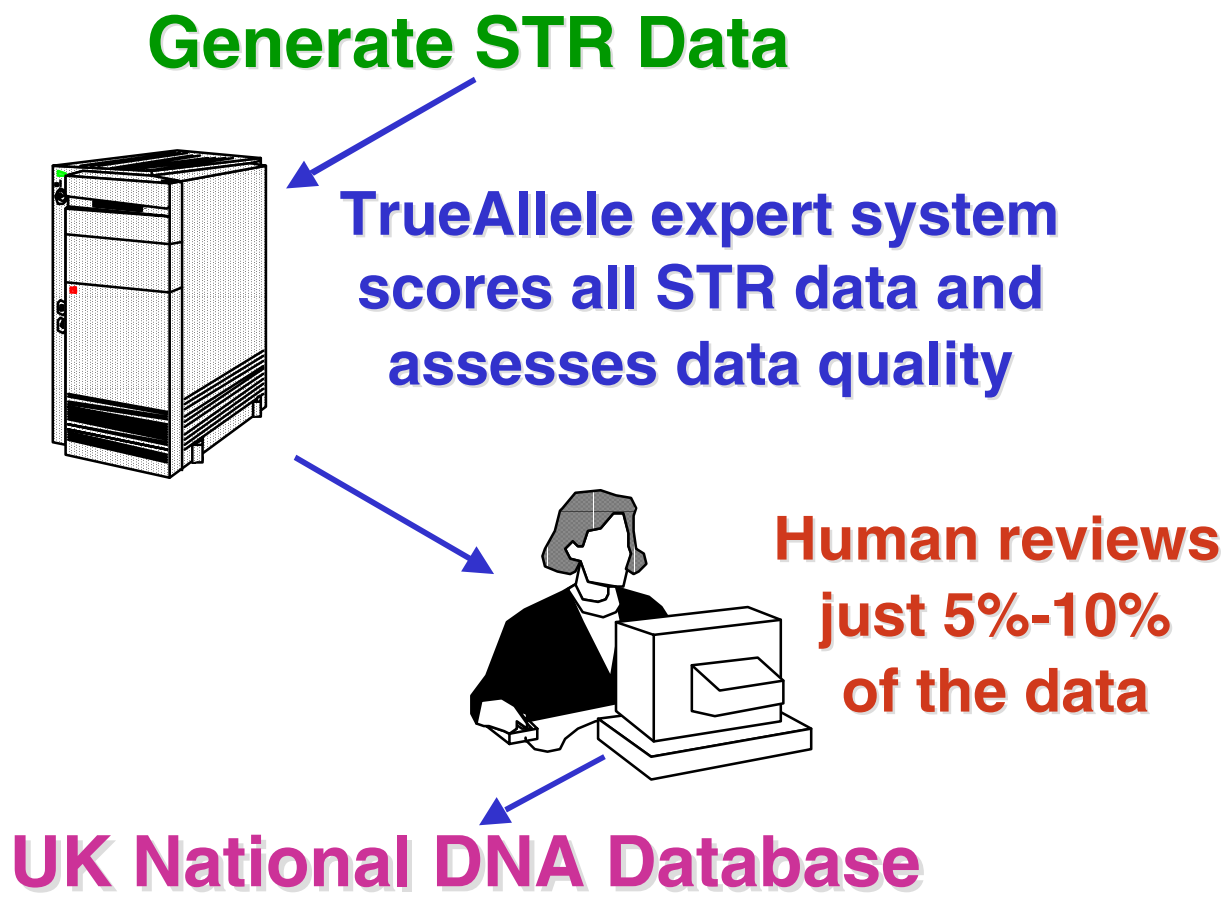
°

Figure 17. Web site. Considerable support for the TrueAllele™ technology is provided via the "www.cybgen.com" site.



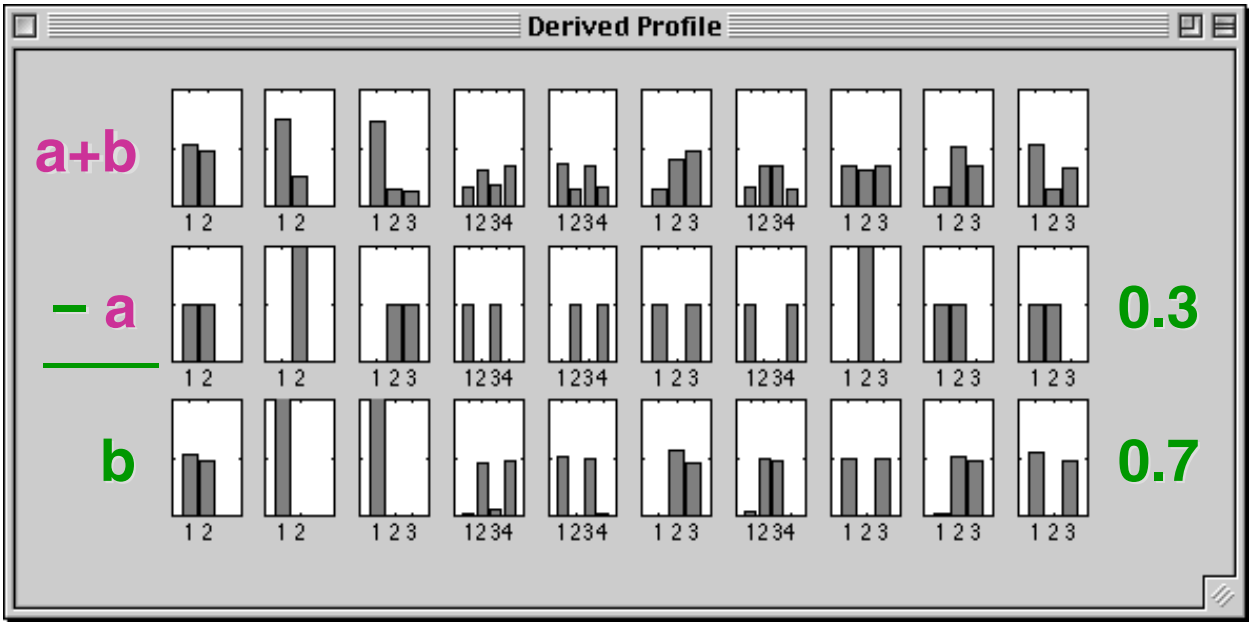
°

Figure 18. Automated scoring: FSS/UK. The TrueAllele™ process for the FSS in the UK largely eliminates the human review of STR data.



°

Figure 19. Mixture deconvolution. Shown are data and analysis results for automatically resolving mixed DNA samples. Each row shows the quantitative data at all ten loci for the samples indicated. The method determines the third row 's genotype (unknown b), and the mixture weight, using only the data in first two rows (mixture a+b, reference a).



°