Identifying human remains using TrueAllele® technology

Mark W. Perlin, PhD, MD, PhD
Cybergenetics, Pittsburgh, PA

29 August 2007

Book chapter appears in:

Perlin, M. W. (2007). Identifying human remains using TrueAllele® technology. Forensic Investigation and Management of Mass Disasters. M. I. Okoye and C. H. Wecht. Tucson, AZ, Lawyers & Judges Publishing Co: 31-38.

Corresponding author contact information:
Dr. Mark W. Perlin
Cybergenetics
160 North Craig Street, Suite 210
Pittsburgh, PA 15213 USA
412.683.3004
412.683.3005 FAX
perlin@cybgen.com

Introduction

The DNA laboratory transforms biological specimens into STR sequencer data. To accomplish this task automatically, the laboratory applies a series of DNA analysis instruments. These instruments include a DNA extraction robot, a DNA quantitation device, a PCR DNA amplifier, and a fluorescent DNA sequencer. But the resulting DNA data must then be further transformed into human identification information: genetic profiles and matches. This information can be obtained automatically using a DNA interpretation instrument – the TrueAllele® Scientific Calculator (TASC™). The TASC instrument is a small supercomputer specifically designed for transforming DNA sequencer data into genetic profiles and matches.

TrueAllele System

The TrueAllele Casework system [1] consists of an analysis workstation, a TASC supercomputer, and visual DNA review software for forensic analysts. The analysis workstation reads the laboratory's DNA sequencer data, and performs a number of automated quality checks. A human operator then visually checks the data in the context of the computer's quality information. Questionable data is routed back to the laboratory for reanalysis, while acceptable data is further quantitated and uploaded to the TASC database. The entire human interaction takes about one minute for each plate of 96 DNA samples and controls.

The TASC database holds the DNA data and interpretation requests for specific case questions. Once all the data has been uploaded for an interpretation request, the request is ready for processing by one of the many parallel TASC interpretation computers. (For example, the TASC-16 model runs 16 parallel interpretation processes.) An interpretation process statistically examines all the data and their uncertainty in the context of a mathematical model of the STR process. After inferring genetic profiles from the data, the interpretation computer uploads its results to the database. Another TASC computer then automatically matches the inferred profiles on the database against one another, according to the lab's specific application requirements. For example, with mass disasters, victim remains profiles are compared against personal effects profiles.

Forensic scientists use the TrueAllele visual user interface client software to review the inferred genetic profiles and matches in the context of their original DNA data. Information is displayed both visually and textually, and the presented visualizations dynamically adapt to the user's focus of interest. From the TrueAllele interface, the user can annotate the data and initiate additional TrueAllele interpretation.

TrueAllele Science

The TrueAllele casework science broadly encompasses three areas: data quality, genetic profiles, and genetic matches. Data quality is quantified by statistical modeling that determines the relative confidence in each data component. In mass disasters, this

is important, since the DNA data can be highly uncertain, even when data quality is high. For example, damaged victim remains may contain little DNA and produce peaks that are below the validated human detection threshold. However, these data can be highly informative when analyzed by a TrueAllele computer.

To preserve the genetic identity information that is present in uncertain STR data, forensic scientists often generate lists of genotype possibilities, and give them an equal weighting. However, the data may strongly suggest that some genotype candidates are more probable than others. Therefore, TrueAllele Casework represents the genetic profile of a contributor at a locus as a probability distribution of genotype possibilities, assigning more probable genotypes a higher probability.

Quantitative DNA match information is obtained by comparing the probability of the genetic profile matching a specific target, relative to the probability of that profile matching a random person in the population. With straightforward single source DNA profiles, the specific match probability (numerator) is 1, and the random match probability (denominator) is the frequency of the genotype in the population. TrueAllele matching generalizes this match ratio concept in order to compare genetic profiles that are described using probability distributions. This mathematical refinement was specifically designed for handling uncertain DNA data. Therefore, it is not surprising that TrueAllele match can preserve more human identity information than the current DNA match statistics, since these older methods originated from comparisons based on unambiguous pristine DNA data.

TrueAllele Technology

The TrueAllele casework technology is highly useful in the challenging DNA interpretation problems found with mass disasters. Victim remains are often damaged by heat and moisture, which degrade the DNA molecules and produce low signals. Personal effects (e.g., toothbrushes) often contain a mixture of DNA contributors, such as the missing person and a spouse. A missing person's family references can be used to combine relatives' profiles to infer a genetic profile. In all these situations, the DNA data are highly uncertain and genetic profiles with refined probability representation are needed to preserve identification information.

TrueAllele Workflow

The sexual assault DNA identification problem is quite similar to identification issues in a mass disaster [2]. In a mass disaster, unknown genetic profiles from victim remains are matched against the missing person reference profiles from personal effects and kinship DNA. Similarly, in a sexual assault case, the unknown assailant's genetic profile is inferred from uncertain DNA data that contains a mixture of the victim and the assailant. This unknown inferred profile is then matched against reference profiles, such as a particular suspect or a database of known criminal convicted offenders.

The DNA process workflow for sexual assault or mass disaster begins with a selection of relevant biological specimens. The specimens go through the DNA laboratory process, are entered into a Laboratory Information Management System (LIMS), and ultimately produce DNA sequencer data. The TrueAllele Analysis workstation conducts a quality control check, followed by a rapid human review; this check provides laboratory feedback to the LIMS database, and uploads quality checked DNA peaks to the TrueAllele database on the TASC instrument. A case's interpretation request specifies which DNA mixture evidence and victim reference to use. A TrueAllele interpretation process examines the data indicated by this request, inferring the unknown genetic profile and uploading this profile to the TrueAllele database. TrueAllele can then match the inferred profile against a suspect database.

A forensic scientist thus has the original DNA data, the inferred genetic profiles, and the DNA matches as an organizing foundation for their review. This computer-inferred information is available to the examiner from the TrueAllele database throughout their own visual review. The scientist can specify interesting data subsets for the TrueAllele scientific calculator to solve with new DNA interpretation and matching.

The property crime (burglary, car theft, etc.) DNA identification process is also similar to the mass disaster task [2]. Here, the unknown profiles come from DNA left by the criminal at the crime scene, and the reference profiles are largely derived from a suspect database of known criminal offenders. Property crime DNA data is often uncertain due to the low levels of biological material left behind by the perpetrator.

Since burglars are serial offenders who often progress to more violent crimes (e.g., sexual assault), it is important to match their unknown crime scene genetic profiles to a criminal database so that they can be identified and apprehended by the police before they commit more crimes.

In both mass disasters and property crime, much of the crime scene DNA evidence produces genetic profiles that do not match any reference profile. It is therefore useful to modify the genetic analysis and interpretation workflow to maximize the utility of the human scientist's time. This can be done by screening the DNA information so that a forensic scientist only needs to look at STR data and genetic profile after a DNA match has already been found. In the TrueAllele Casework system, this workflow is implemented by having the computer automatically generate interpretation requests, and match computer-inferred genetic profiles to the reference DNA database. The first human look at the DNA data (plus genetic profiles and match results) can then occur for just the DNA evidence that has a known match.

So Much Data, So Little Time

With the advent of automated laboratory instrumentation, the DNA lab has witnessed an exponential increase in its data generation capability. Robotic and computer systems for DNA extraction, amplification, detection, and database organization have also improved the reproducibility of the STR experiment. However, most DNA data review is still done manually without the aid of scientific calculation for inferring and matching

genetic profiles. So DNA data review only scales linearly: to double the capacity, a lab must double the number of people. In order to meet the DNA data challenge of a mass disaster, the data review must scale exponentially along with the vast amounts of DNA data.

The goal of every forensic scientist and their laboratory director is to do the best job possible in making a human identification with DNA evidence. This entails (1) looking at all the good data, (2) ignoring the bad data, (3) considering every possibility, and (4) obtaining the most match information for human identification. The number of interpretation possibilities and interactions between data grows much faster than the amount of data. This increase in interpretation difficulty is even greater with the highly uncertain data often found in a mass disaster.

TASC Instrument - A Scientific Calculator

For rapid DNA identification in a large-scale mass disaster, another DNA instrument is required after the DNA sequencer. This "scientific calculator" instrument is needed to interpret STR data and generate genetic profiles by doing the best job possible. The TASC instrument performs the necessary functions of this scientific calculator. First, it interprets DNA evidence and generates genetic profiles by considering every possibility and preserving match information using probability representations. Second, it matches genetic profiles (e.g., between victim remains and personal effects or kinship profiles) using genotype probabilities in order to retain match information. Third, the

supercomputer instrument maintains a database for storing and retrieving DNA information.

The TASC engine interprets DNA data through repeated application of the scientific method. Specifically, at each decision it forms a new hypothesis about one of its statistical modeling parameters. These parameters include genotypes, mixing weights, PCR artifacts, DNA amounts, background noise, data uncertainty, and parameter uncertainty. Combining these parameters, the computer generates a hypothetical data pattern, which it compares against the STR data in order to form a probability value. The statistical decision process prefers higher probabilities to lower probabilities, and is therefore able to compute the probability distribution of each parameter. The end result is that the TrueAllele interpretion assigns higher probabilities to those genotypes which best fit the observed DNA data.

Solving a DNA mixture problem simply entails considering additional parameters. These parameters include the number of contributors, the genetic profile at each locus of a new contributor, and the mixing weight between the contributors. The hypothesized data patterns are more complex, because they include alleles for more than one contributor. However, the basic mechanism (of comparing hypothesized patterns against the actual data in order to form probabilities and make decisions) is the same.

In the process of converging to an answer, the computer probabilistically searches the values of a parameter. This random variation captures the statistical variation of that

parameter, and can be represented by a histogram. The center value of the histogram estimates the mean value of the unknown parameter, while the width of the histogram describes its standard deviation. In this way, the statistical TrueAllele engine determines both the values of genetic parameters (e.g. genotypes, mixing weights) and the scientific confidence in those values based on the data uncertainty.

Forensic Review

The TrueAllele database contains the genetic profiles of each interpreted mass disaster sample, and the appropriate matches between them. These results can be viewed by accessing the database via a local computer network to their own lab's in-house TASC instrument, or by Internet to a remote TASC server. Either way, a user connects to a TASC database by logging on to their secure trueallele.net account. A web browser client provides complete access to the genetic profiles, matches and data for each interpreted request. The browser window presents in table form the genotypes and probabilities for each locus of each contributor, along with discriminating power and match information statistics. All of the original data peaks are also viewable as tables.

A separate TrueAllele database client for viewing results is provided by a Windows-based visual user interface. After logging in to their TASC database, and selecting a case to review, the user sees all of their original DNA sequencer data signals in an easily navigable graphical interface. With one or two mouse clicks, the user can zoom in on the sequencer lanes and genetic loci of greatest interest. Another window shows

the genetic profiles (including genotypes and probabilities) in an intuitive visual form. Focusing on information in one window will automatically zoom in on comparable information in the other windows. The result is a user experience entirely determined by the immediate genetic information interests of the visual user.

The visual user interface also provides the capability for the forensic scientists to conduct "what if...?" scenarios. The user can submit requests for interpretation and matching that include additional samples. The forensic scientist has complete control over which samples should be used in reinterpretations of the evidence, and can ask questions like "Is there too much data?" or "Should low-quality data be discarded?" The forensic user can combine or separate low signal data in order to obtain the most statistically reliable answer to their forensic DNA identification questions.

Mass Disaster DNA Identification

The DNA mass disaster process [2] begins by obtaining three sets of biological specimens: victim remains from the mass disaster site, personal effects from the missing people, and family references from their immediate relatives and spouse. These biological specimens are transformed by the DNA laboratory into DNA signals, where each signal peak has a size (in base pair units) and an amount (in relative frequency units).

True Allele interpretation requests are organized by biological specimen, so that all of the DNA sequencer lanes developed from one biological specimen will be interpreted together. This grouping together of all related data for a specimen provides far greater statistical power than analyzing each sequencer lane separately, and having a person try to subjectively reconcile multiple genotypes across the multiple experiments. With joint data interpretation, a single hypothesized genotype pattern must match well at all of the experiments for that specimen at a locus. Consistent use of mathematical probability functions that represent these pattern comparisons and their statistical uncertainty can be rigorously performed by a computer.

For each biological specimen, all the DNA peak data are examined together in order to infer a genetic profile. When there is no ambiguity about the genotype at a locus of the contributor to the DNA sample, the genetic profile at that locus has only one feasible genotype. However, in mass disasters, damaged victim remains will typically yield uncertain DNA data, and so the genetic profile at such a locus will contain multiple genotypes, each with an associated probability derived from the data. The inferred genetic profiles are organized into groups, depending on their role in the DNA identification process (e.g., victim remains, missing people).

As each genetic profile is inferred, it is compared against the previously determined profiles in other profile groups. For example, a victim remains profile will be compared against all the personal effects profiles and all of the kinship derived profiles. Similarly, when a new personal effects genetic profile is inferred, it is compared against all of the

victim remains profiles. In this way, all of the victim remains DNA information is matched against the missing person DNA information, as new genetic profiles become available. The use of multiple coordinating computers to infer and match genetic profiles, and communicate via a central database, enables such real-time dynamic updating of DNA match information.

When a forensic scientist examines a candidate DNA match between a victim remain and a missing person, the TrueAllele casework system provides all the relevant information for assessing the validity of that candidate match. The inferred genetic profile of the victim remain is shown, along with its discriminating power. The same genetic information, in both tabular and visual forms, is similarly provided for the missing person. The match information is represented numerically and graphically, along with the individual locus hits and misses. Ready access to the original DNA sequencer data signals also helps the scientist during forensic examination. Since the TrueAllele scientific calculator is merely an aid to human decision-making, all final decisions are made by forensic scientists and their supervisors.

World Trade Center TrueAllele System

The TASC supercomputer that we are using to analyze the World Trade Center data has a central multiprocessor TrueAllele database computer, 20 DNA interpretation processors, and a DNA match computer. In its current configuration, optimized for extensive analysis of damaged victim remains STR data, the system can automatically

infer and match a new genetic profile every two minutes. Doubling the number of interpretation processors would halve this instrument turnaround time.

We are currently working with New York City's Office of the Chief Medical Examiner to help identify more World Trade Center victim remains. They anticipate that the use of the TrueAllele technology on the World Trade Center effort will yield additional results. We are currently well along in the process of providing DNA decision-support to their forensic scientists using the TrueAllele Casework System.

Conclusion

The TrueAllele Casework system is being used to identify human remains. The heart of its speed and accuracy is the TASC supercomputer, a multiprocessor instrument specifically designed for inferring and matching DNA profiles. The primary goal of the TrueAllele Casework system is to preserve DNA evidence. When inferring genetic profile and match information, the TrueAllele system considers every possibility, has high automated capacity, and produces rapid results, relative to conventional human review. Applying dedicated DNA statistical support to mass disasters can accelerate and improve human decision-making in identifying human remains.

References

[1] Perlin MW. Simple reporting of complex DNA evidence: automated computer interpretation. In: *Promega's Fourteenth International Symposium on Human Identification*; Phoenix, AZ; 2003.

[2] Perlin MW. Real-time DNA investigation. In: *Promega's Sixteenth International Symposium on Human Identification*; Dallas, TX; 2005.