DNA Done Right

Introduction

Deoxyribonucleic acid (DNA) is the information molecule that encodes the instruction set for biological operation. As James Watson and Francis Crick discovered in 1953, the self-replicating DNA polymer writes double helix sentences in a four-letter nucleic acid alphabet. Packaged into 23 chromosome pairs, and spanning three billion letters, this genetic software can direct the workings of a cell, or grow it into a human being.

Reading distinctive passages from the DNA book of life can help identify people and distinguish them from one another. But actual biological evidence is often challenging – paragraphs from multiple people merge (mixtures), DNA ink fades (low amounts) and the paper crumbles (degradation). Evidence that cannot be interpreted goes unused. However, accurate computer modeling can interpret this DNA data to extract scientific truth for criminal justice.

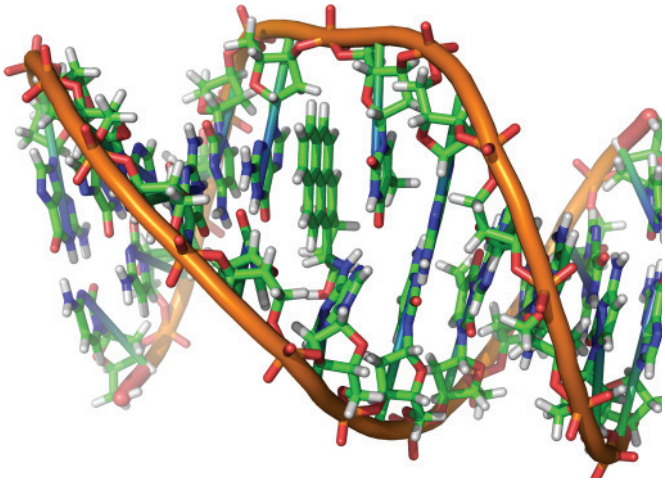
DNA Identification

The DNA prose written in our chromosomes records a "genotype," the genetic text copied from parent to child. The expression of these genes into observable physical traits forms a "phenotype." For example, a blue/brown and brown/brown genotype for eye color will both manifest themselves as a brown-eyed phenotype, following Gregor Mendel's 19th century laws of dominant gene inheritance. These phenotypic rules are followed by the ABO blood group antigens, which are early 20th century markers for establishing paternity and biological identity.

Late 20th century molecular biology rewrote these rules, as scientists began to read directly from the genetic text. The DNA genotype, long hidden within the cell's nuclear membrane, emerged as a new observable phenotype. Professor Alec Jeffreys of Leicester University rolled the first DNA fingerprints using early gene detection methods, and demonstrated their manifold applications to human identification.

Around the same time, Nobel laureate Kary Mullis developed the polymerase chain reaction (PCR). Commandeering nature's own enzymatic machinery, scientists could breed DNA like bacteria, doubling their number every cycle to magnify a gene sequence into millions of molecular copies.

Scattered throughout our genome are hundreds of thousands of genetic locations (loci) that contain short, tandemly repeating (STR) DNA words. Lacking any known function, this DNA text is largely unconstrained by Charles Darwin's natural selection, and so can evolve into a diversity of possible repeat lengths. Each person has just two of these "allele" numbers (one inherited from each parent) at a genetic locus, and so people's STR genotypes generally differ.

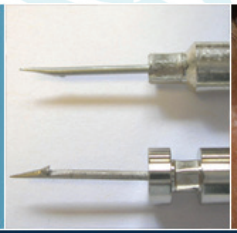


Examining 10 or so different genetic loci multiplies into an astronomical number (billions of billions) of genotype possibilities. Relative to the size of the human population (only billions), the STR genotype provides a virtually unique barcode for identifying people from their DNA.

Forensic Application

The United Kingdom Home Office's former Forensic Science Service (FSS) recognized the power of DNA identification for fighting crime. They began with Jeffrey's DNA fingerprinting methods, and then developed their own STR chemistries. The FSS soon became the world leader in forensic DNA identification. In short time, their scientists had launched the National DNA Database (NDNAD).

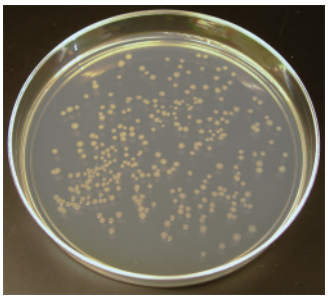
Imagine if the genotypes of all criminals were stored on a database, along with the genotypes from all crime scenes. Then computers could solve cold cases by connecting crime scene to criminal through their common genotypes.




FHC APPROVED EXPERT WITNESS'

The UK built the world's foremost DNA database, eventually housing the genotypes of millions of potential perpetrators. The FSS aggressively processed DNA evidence from property crime, which is often the starter offense for young criminals. The outcome was cost-effective policing based on DNA match, with a reduction in crime and interrupted criminal careers.

As robotic FSS laboratories churned out DNA data, a new problem arose – the interpretation bottleneck. The DNA signals from each suspect's criminal justice (CJ) sample had to be reliably read as a genotype, and then entered onto the NDNAD. Even with two independent readers, and a third person to resolve discrepancies, the CJ error rate was 1 in 2,000 samples.



The FSS had about a hundred people working in Priory House on CJ interpretation in Birmingham. This data analysis factory worked three shifts to manually transform electronic DNA signals into genotypes, and shepherd this information onto the national database. In the late 1990's, the CJ backlog stood at 350,000 samples. This was also the expected annual volume, so simply hiring and training another hundred manual data reviewers was not a scalable solution.

The FSS contacted Cybergenetics in 1998. Our Pittsburgh-based company had developed an expert computer system for automated genotyping of STR data, primarily used for genetic research and diagnosis. We adapted our TrueAllele system for forensic use, deploying it at the FSS on two desktop computers. The FSS used TrueAllele to eliminate their CJ backlog and genotyping errors, and shift staff to casework operations. Their CJ interpretation effort scaled down to a handful of people who worked normal business hours on TrueAllele computers.

Mixture Problem

CJ reference samples have abundant DNA taken from a single person under controlled conditions. These reference items produce relatively pristine STR data that TrueAllele expert system rules can easily handle. But real DNA evidence is usually more complex. Mixtures contain DNA from multiple people, low DNA amounts are harder to PCR copy, and degraded DNA is cut into smaller fragments that cannot be copied at all.

Mixtures introduce uncertainty. Whereas easy single source data has only one genotype solution, complex STR patterns can admit many genotype explanations, often with no best answer. Experts no longer see an obvious match with a huge statistic. There is far more explaining to do, both in time-consuming data interpretation and when testifying in court.

CJ analysts were used to drawing a horizontal line through STR data peaks at a predetermined vertical "threshold" height. Peaks over the threshold would usually correspond to the one or two STR alleles from a person's genotype allele pair. But mixtures of multiple individuals showed more than two allele peaks, and simple rules no longer applied.

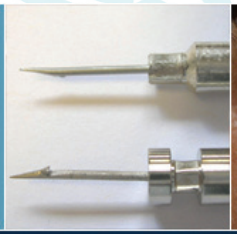
Height now mattered, because small amounts of one contributor gave short peaks, large amounts of another contributor gave tall peaks, and these quantities added up to produce composite mixture patterns. The threshold simplification ignored small peaks, did not discriminate between tall peaks, and lacked a mathematical basis for interpreting quantitative mixture patterns.

Forensic analysts also saw mixtures where an individual person's allele could "drop out" from the data. Either the allele peak was visible (but under the threshold), or it wasn't seen at all (too little DNA present). Allele dropout methods were developed to statistically conjure phantom peaks for alleles not seen in the data. Ignoring the data you have (mixture peaks, heights, and patterns) for the data you want (simpler, but not actually observed) is not the most elegant science. Surely there is a more rigorous way to reason from uncertain evidence.

Genotype Modeling

In the mid-18th century, Scottish philosopher David Hume was unpersuaded that miracles or other empirical evidence could prove the existence of a deity. However, his British contemporary, the Reverend Thomas Bayes, developed a mathematical rule for updating belief in any hypothesis (whether scientific or religious) based on observed evidence. Bayes rule says that after seeing evidence, our belief in a hypothesis changes in direct proportion to how well that hypothesis explains the data.

We need not invoke thresholds, phantoms or deities to solve the DNA mixture problem.




Bayes rule suffices, for it tells us how to update our belief about the genotypes (of each contributor to a mixture) after examining the STR evidence. Before seeing any data, the probability of observing a genotype allele pair is just its prevalence in the population.

When examining data, all possible genotype combinations of the contributors are considered (along with many other variables), assessing how well each combination explains the STR peak height pattern. Better explanations confer higher probability to their constituent genotype values. After seeing data, an evidence genotype will place more probability on allele pairs that better explain the data, and less on those that do not.

This genotype modeling approach requires a computer. The earliest versions (from 1999) of Cybergenetics TrueAllele Casework system could solve DNA mixture problems in seconds. However, more variables are needed for more accurate models that can better explain the STR data and not make mistakes. In 2009, after ten years of development, twenty five software versions, hundreds of reengineered variable models, eighteen thousand World Trade Center victim remains, and a hundred thousand tested casework samples, TrueAllele Casework was ready for use on criminal DNA evidence.

Match Statistic

TrueAllele's Bayesian modeling produces evidence genotypes – one probability distribution for every contributor at each locus – objectively computed solely from the evidence data without any knowledge of a suspect. The genotype summarizes all the identification information contained in the genetic STR data. This summary suffices for making comparisons with candidate contributors to the DNA evidence, and calculating a match statistic.

The DNA match statistic expresses the gain (or loss) in identification information after examining the evidence. Formulated for ABO blood group paternity testing in the 1930's, the "likelihood ratio" (LR) is a two-hypothesis form of Bayes rule. Either a suspect contributed their DNA to the evidence (hypothesis H) or they did not (alternative ~H). As developed for cracking German codes during World War II by British computer pioneer Alan Turing, the LR gives the odds of H after having seen the evidence, relative to before.

UK statistician Dennis Lindley showed in the 1970's how the LR could be used for forensic glass identification. In the late 1990's, FSS and other English-speaking scientists demonstrated the LR's applicability to DNA mixtures.

TrueAllele Casework makes full use of quantitative STR data to objectively infer evidence genotypes. An operator compares this Bayesian-modeled evidence genotype with a reference (e.g., suspect's) genotype, relative to a random population genotype, to immediately calculate a match statistic. Expressing the LR in plain English, a scientist can state that a match between the evidence and suspect is (the LR number) times more probable than coincidence.

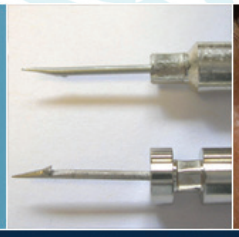
Scientific Reliability

Scientific evidence must be sufficiently reliable for it to be admissible in court of law. Yet laboratory signals such as PCR-amplified DNA have natural variation. Mixtures and small DNA amounts exhibit even more pattern fluctuation. How can solid results be derived from inconstant data?

A first application of Bayes rule lets us infer genotypes solely from the data, representing genetic uncertainty with allele pair probability. A second Bayesian turn with the LR then compares these probabilistic genotypes with reference and population genotypes to calculate the change in identification information. Bayes done twice (drawing on considerable computing power) allows a thorough examination of STR data, and an objective determination of match strength.

A DNA match is expressed in a single statistic, the LR, whose logarithm (powers of 10) is a standard measure of information. A positive log(LR) supports inclusion, a negative value suggests exclusion, while numbers around zero are inconclusive.

The reliability of a scientific process is assessed through validation. With DNA mixtures, we want an interpretation method to be sensitive (include the contributors), specific (exclude non-contributors) and reproducible. Validation studies can measure these axes of DNA mixture information through log(LR) values calculated from genotype comparisons.

Cybergenetics and other groups have conducted many TrueAllele validation studies on DNA mixture interpretation. Two published peer-reviewed studies used DNA samples of known composition, while two other journal papers assessed casework items and compared with manual review. These studies have established that the system is sensitive (match statistics are a million times higher than some threshold methods), specific (false matches are rejected by factors of a billion billion) and reproducible.

Court Appearances

TrueAllele Casework was admitted into evidence in three homicide cases where there were defense challenges, one in the United Kingdom and two in the United States. In a US case, the Pennsylvania Superior and Supreme courts upheld the conviction and TrueAllele's reliability, establishing a statewide precedent. The system was not admitted in a 2010 UK arson retrial where the judge had wanted more validation, although he "did not give a reasoned judgment explaining his decision."

More TrueAllele validation studies were done, with regulatory approval granted by the New York State Commission on Forensic Science. A year and a half later, a voir dire was held in the Northern Ireland Massereene Barracks attack trial. In his ruling, the Honourable Mr. Justice Hart was satisfied that the system could be "regarded as being reliable and accepted" and admitted TrueAllele into evidence.

TrueAllele results have been reported in over a hundred criminal cases, most often for the prosecution. TrueAllele experts have testified in fifteen criminal trials, for offenses including murder, rape, child abduction, child molestation, bank robbery and terror. The system can separate out genotypes from mixtures of relatives. TrueAllele has helped lawyers defend innocent clients. The police use this computer interpretation to sharpen their DNA evidence, whether they need a more informative match statistic for a suspect in a crime, or they want to conduct a more effective DNA database search to solve a cold case.

Genotype Database

The original UK vision was to use the NDNAD to prevent crime through cold case DNA match by retiring criminals early in their careers. But DNA mixtures, and other challenging evidence, have dimmed the success of that mission.

The NDNAD can only store simple single-source genotypes, with allowance for uncertain alleles. This late 20th century database was designed for pristine evidence and reference samples. The NDNAD cannot effectively represent today's DNA mixtures, and so most of that hard-won taxpayer-funded evidence is not used for crime prevention.

A TrueAllele genotype database could help fulfill the original NDNDA goal. The system can resolve all DNA evidence, regardless of complexity or number of contributors, into its constituent genotypes. All genotypes can be uploaded to a national TrueAllele database. The high specificity of TrueAllele's mathematical LR database match greatly reduces false hits. Its high sensitivity finds more cold hits that solve (and ultimately prevent) crime more effectively than existing government technology. A public-private partnership could use all of Britain's DNA evidence in a national genotype database that would better prevent needless victimization.

Conclusion

DNA identification began in the United Kingdom. For over two hundred years, Britain has been innovating the science that backs this forensic gold standard. By using only the most accurate interpretation methods, a nation can keep its wealth of DNA evidence from transmuting into fool's gold.

Cybergenetics pioneering TrueAllele technology is an accepted part of the DNA landscape. The FSS and Cellmark Forensic Services have genotyped millions of CJ samples through TrueAllele computers. The Casework system has interpreted DNA evidence in UK criminal cases, with expert testimony given on reported matches. TrueAllele can help burnish the DNA gold standard, mathematically preserving DNA evidence to find the guilty, free the innocent and make the world a safer place.

by Mark W. Perlin

Dr. Mark Perlin is Chief Scientific and Executive Officer at Cybergenetics. He has twenty years experience developing computer methods for information-rich interpretation of DNA evidence, and providing TrueAllele® products and services to the criminal justice community.