Kern County Resolves the DNA Mixture Crisis

Mark W. Perlin and Kevin W.P. Miller

Introduction

DNA mixtures containing DNA from two or more people comprise most biological evidence samples seen in United States crime labs. Mixtures can greatly complicate data interpretation since DNA analysts must account for data patterns that include many peaks of varying height. There are a vast number of weighted genotype combinations, with multiple ways to explain the data. This complexity makes it hard for analysts to readily differentiate probable from improbable genotypes.

The difficulty in manually interpreting complex DNA mixtures has consequences. Hundreds of thousands of evidentiary items have been collected and processed into DNA identification data that could implicate or exonerate—but these items have not been reported conclusively. A failure to fully use data from evidence is a failure of science to promote justice. This ongoing mixture crisis permits needless victimization by those whom DNA should have identified.

The forensic community is discussing genotype probability modeling as a way to interpret DNA mixtures. The Kern Regional Crime Laboratory (KRCL) was an early adopter of probabilistic genotyping for mixture interpretation. KRCL's adoption of these computer methods enables Kern County to use complex mixture evidence in routine casework, and easily report their match results.

Getting Started

KRCL is located in Bakersfield, California, just two hours north of Los Angeles. KRCL, which operates as the Forensic Science Division of the Kern County District Attorney's Office (KCDA), is only one of three crime laboratories in California to be administered by a District Attorney. The Laboratory's new Director, who arrived in October of 2011, has spent the past two years updating the DNA Analysis Unit with new kit chemistries, robotic instrumentation, a new LIMS, and software for the interpretation of DNA mixtures.

This article chronicles the partnership between KRCL and Cybergenetics, a bioinformatics company, in order to realize KRCL's goal—creating meaningful workflows that enhance both the quantity and quality of genetic information obtainable from a variety of (often challenging) biological evidentiary samples.

Initial Assessment

In late October of 2011, KRCL contacted the company about purchasing their genotype probability modeling system. KRCL initially assessed the software by sending electronic data from three mixtures obtained from three of its challenging cases. These data were processed and presented in a webinar customized for the laboratory. The system found match statistics between:

- A beer can left at a homicide scene (20% contributor) and a suspect of a billion
- An oral swab from a suspect (three person mixture) in a male-on-male sexual assault and the victim of a hundred thousand
- A glove from a casino robbery (at least four contributors) and three suspects of a hundred thousand, hundred million, and sextillion, respectively

KRCL purchased the system through a National Institute of Justice grant in May of 2012.

Planning the Deployment

At the May 2012 meeting of the California Association of Criminalists (CAC), both parties met to plan the system validation and subsequent deployment into casework. A rollout document mapped the system hardware and software requirements, user training plan, mixture validation study, and timelines for documenting how the system would function in casework.

Science and the Law

Cases

During the validation process, KCDA began sending criminal cases to the bioinformatics company for analysis. This gave KRCL a first-hand look at the system's operation and, equally importantly, its reception and use by prosecutors—the laboratory's primary consumer of forensic data.

The first case ever prosecuted in California using probabilistic genotyping was that of West Side Crips street gang members Charles Lawton and Dupree Langston. In 2011, Lawton and Langston committed eight armed robberies of jewelry and check advance stores in several cities. During their five week crime spree, Lawton and Langston robbed employees and brutalized their victims—some of whom were beaten or forced to disrobe at gunpoint.

In February of 2012, KRCL developed STR data from ten clothing and touch DNA items that showed mainly low-level three or four person mixtures. The majority of data were not interpreted because they fell below traditional threshold cut-offs.

The company analyzed the mixture data using their supercomputer, with 60 interpret computer processes solving for the genotypes in under a week. Once the company received reference data for the five victims and five suspects, 9 DNA matches were found. These likelihood ratios (LRs) were reported in time for a pretrial hearing in June.

Validation

In June of 2012, KRCL expanded on twenty previous system validation studies. Their new study examined up to five unknown contributors, with a random mixture design using both high and low DNA quantities that simulated casework observations. The supercomputer completed its interpretation of 120 random mixture samples over several weeks. All analyses were repeated manually by KRCL's most experienced analysts, and the results were compared.

The study found genotype probability modeling to be sensitive, specific, and reproducible. Assuming more contributors than the number actually present usually did not materially affect the LR. Moreover, the system's interpretation behavior was relatively invariant, regardless of the number of contributors or DNA quantity (Figure 1). These results were presented at the 2014 meeting of the American Academy of Forensic Sciences.[1]

Admissibility

The gang case went to trial in January of 2013. The defense requested an admissibility hearing to assess the reliability of genotype probability modeling mixture interpretation. The prosecutor presented validation studies, peer-reviewed papers, related scientific articles, regulatory approvals, forensic applications, and judicial admissibility opinions. The bioinformatics company provided dozens of these documents.

KRCL DNA analysts observed the admissibility hearing for continuing education credit. After qualifying a scientist from the company as an expert witness, the prosecutor introduced the exhibits. Kern's study showed that the match statistic in this case was not unexpected. A scatterplot of DNA (x-axis) versus the number of zeros in the DNA match statistic (y-axis) displayed how (on average) the statistic increases with mixture proportion (Figure 2). The LR in this case had about nine zeros for the 42% evidence mixture (green circle), which falls within the scatterplot.

The science showed it was extremely unlikely to obtain a log (LR) of nine by chance. Comparing LRs of mixture contributor genotypes with non-contributor references, the exclusionary range of log (LR) values was –30 to 0 (Figure 3, red bars). The defendant's nine statistic (green circle) was far from these non-contributor values. After cross-examination, the judge admitted

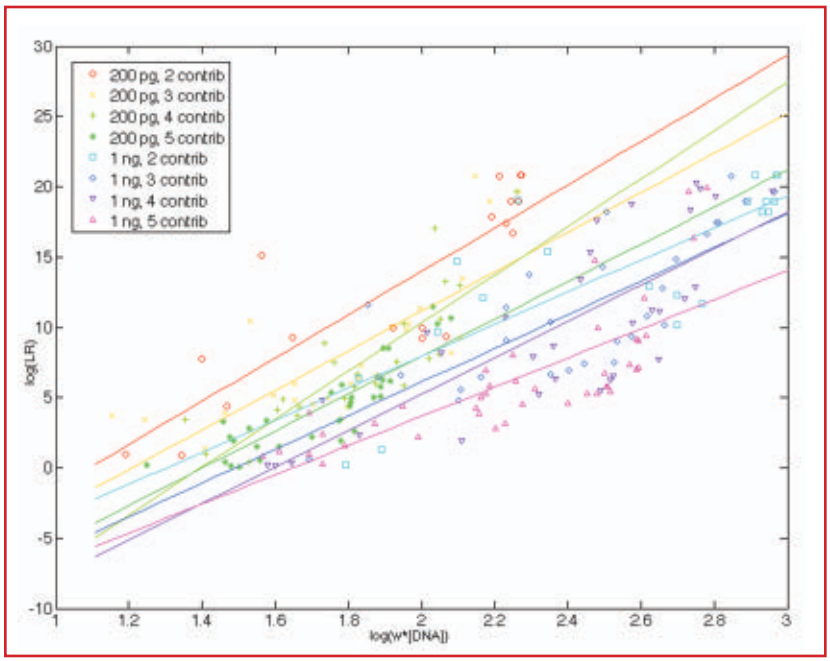

Figure 1: Invariant information response. Whether there are 2, 3, 4, or 5 contributors, or high (1 ng) or low (200 pg) DNA template, in all eight situations the regression slope is the same in a log-log plot of DNA quantity (x-axis) vs. match statistic (y-axis).

genotype probability modeling into evidence.

Trial

Over a hundred witnesses were called, including the police who collected the evidence, KRCL analysts who generated data, and the bioinformatics expert. Security footage showed a robber leaping over a front counter with his hand touching the countertop.[2] DNA swabs from the counter produced a low-level mixture containing at least three contributors.

A PowerPoint presentation outlined the probability modeling evidence, explaining genotypes, STR data (Figure 4), and its interpretation: first objectively infer genotypes, then compare them to references. The computer generates millions of possible peak height patterns. Those that better explain the observed data (Figure 5) confer higher probability to the contributor genotypes, while poorer explanations give lower probability. The result is a genotype (probability distribution) at every locus for each contributor.

The LR was explained in a bar chart showing a contributor's genotype at locus D8S1179 both before (brown) and after (blue) examining the STR data (Figure 6). At the defendant's 14,15 genotype (red) the ratio of posterior probability (blue) to prior probability (brown) is eight. Thus the LR at this locus showed the suspect matches the front counter with eight times more probability than a coincidental match.[3]

Genotype probability modeling found that a match between the front counter and Dupree Langston was 553 million times more probable than a coincidental match to an unrelated individual. Cross exam lasted an hour. In February, the jury found Langston guilty of multiple counts of robbery (with firearm and gang enhancements).[2] In April of 2013, Langston was sentenced to 73 years in prison.

Analyzing Mixtures at Kern

Cases

The software yielded 11 probative matches and 5 exclusions in KCRL's first 20 cases. All cases had samples that would have previously been reported as inconclusive. Kern scientists presented representative cases in an April webinar this year:[4]

- A sexual assault with dozens of challenging evidence items (e.g., touch DNA, low-level mixtures) and over ten references. Without this system, only one sample produced a match statistic. However, the use of probabilistic genotyping allowed five additional reported matches to the offender across multiple cases.
- A soda can left at a homicide had a three-person DNA mixture. Probability modeling placed the can owner in the mixture with a LR of a 100 trillion. Accounting for differential degradation of the contributors, the system gave a LR of 300 thousand between a 20% minor component and the shooter.
- An ax handle and blade showed mixtures having 3 to 4 people. Manual review gave a 1 in 8 statistic. Genotype probability modeling of the same data gave a LR of 2.4 million for a 9% contributor.

Database

KRCL is reanalyzing all past cases using genotype probability modeling, and looking for DNA matches between them. Their new system provides a built-in DNA database matching capability that automatically compares evidence genotypes with other evidence or reference genotypes. The system uploads all DNA mixtures (as separated probabilistic genotypes) to this database for investigative comparison, unlike CODIS which disallows most mixture uploads. The genotype-matching DNA database calculates LR statistics that quantify the strength of match, which CODIS cannot do.

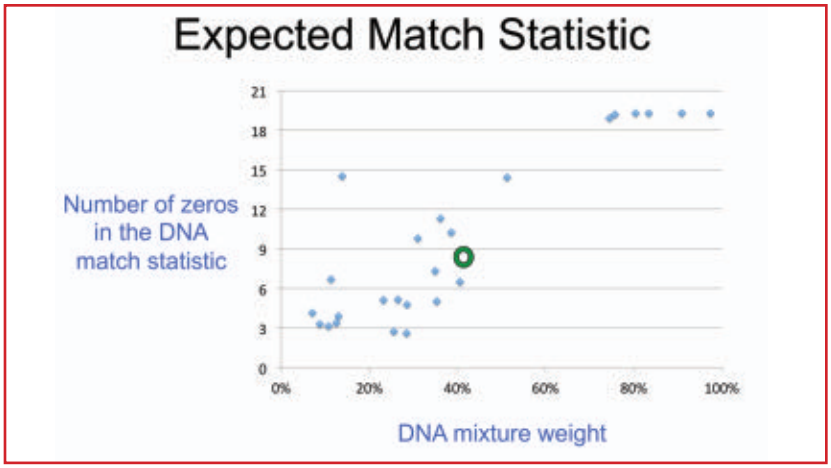

Figure 2: Match statistic predicted. Validation study scatterplot (blue diamonds) shows the mixture weight (x-axis) and match statistic (y-axis) for low-template three-person mixtures. The contributor's statistic (green circle) is consistent with these data.

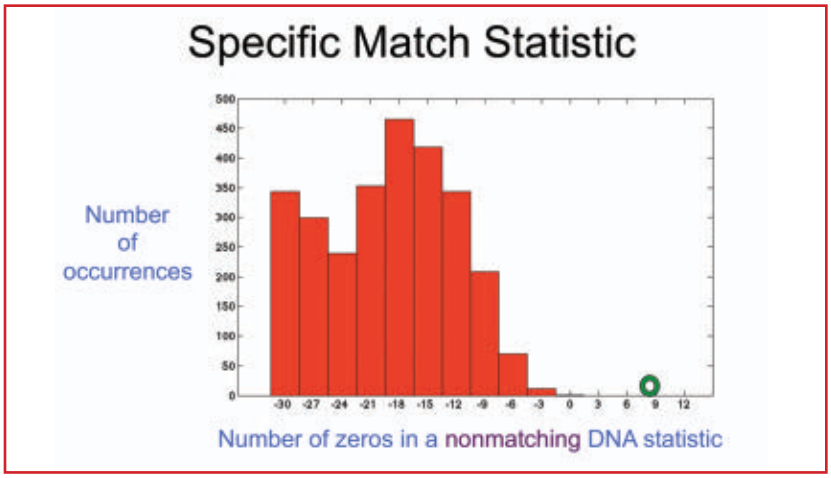

Figure 3: Non-contributor match distribution. A log (LR) histogram (red bars) for 2,500 mixture comparisons with non-contributors is centered left of zero. The match to the defendant has a value around nine (green circle) that lies far to the right of this distribution.

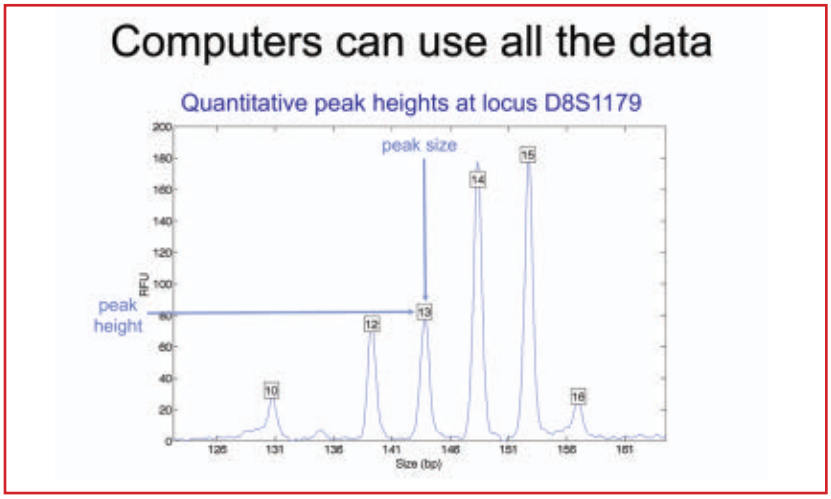

Figure 4: STR mixture data. The DNA mixture at the D8 locus shows six allelic data peaks that arise from at least three individuals. The peak heights in the pattern suggest pairs of peaks that may correspond to contributor genotype allele pair values.

Audit

The first FBI-QAS assessment of the KRCL DNA program after the changes were implemented occurred in April 2014. No deficiencies were found in Kern's rapid transformation to 21st century workflows.

Impact

With the use of the new system, no DNA is left behind.[5] KCDA now expects all DNA evidence data to be interpreted. Exclusions help exonerate the innocent, while inclusions produce accurate DNA match statistics. At KRCL, previously impossible DNA analysis has become routine.

Conclusion

Most DNA mixture interpretation methods ignore STR data and lose identification information. A 2005 National Institute of Standards and Technology (NIST) mixture study showed that thresholds give highly inconsistent match statistics, ranging over ten orders of magnitude (MIX05). In 2013, NIST further showed that adding a "stochastic" threshold does not improve the situation (MIX13).

Forensic scientists want to give accurate answers based on their data. They see the consequences when mixture data are visible, but impotent methods impose silence. On hundreds of thousands of informative DNA mixture items, innocent people cannot be exonerated, criminals cannot be implicated, and prosecutors and defenders alike are incorrectly told of "inconclusive" results that sever truth from justice.

KRCL has shown that the DNA mixture problem is illusory and easily resolved. Highly sophisticated, extensively validated statistical computing solves the problem. A KRCL scientist uploads data, asks some questions, and produces results with fast and informative

workflows. After investing in validation and training, Kern scientists now enjoy accurate DNA mixture match statistics computed automatically, and are freed up to work on other forensic tasks.

The Kern success highlights outmoded national policies. SWGDAM's threshold-based mixture guidelines discard evidentiary information, which leads to inaccuracies. CODIS adds ever more loci in trying to overcome threshold-based limitations, but does not address the fundamental issue that all data should be used. Kern's database reaches into the past to solve cold cases while it prepares for the future by readying reliable mixture evidence for court.

Other crime labs are now turning to KRCL to learn from their experience. KRCL showed how scientists use DNA data to get the best results, and best serve society. Laboratories can follow in Kern's footsteps and avoid many pitfalls in bringing on board an accurate genotyping solution. Working within accepted industry standards, KRCL deployed their system in under two years. The result is better justice in a safer Kern County.

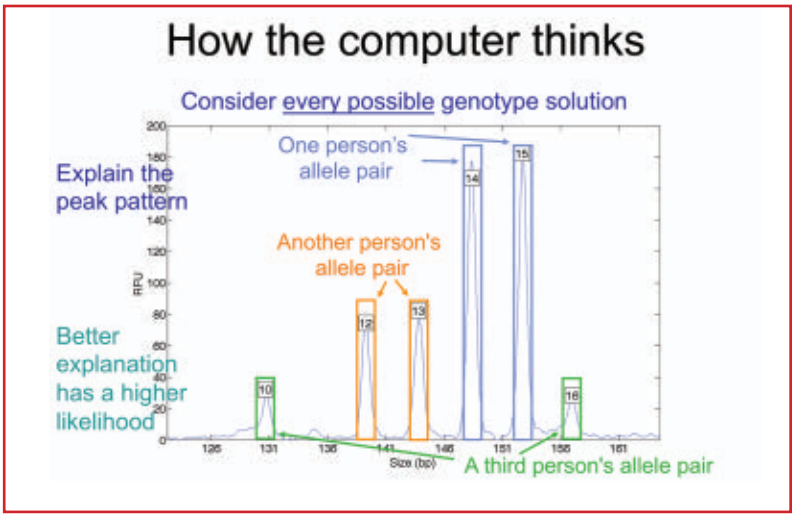

Figure 5: How the computer thinks. The computer proposes patterns as a combination of three genotypes (colored bars). The genotype probability software compares the bar heights with the data's peak heights. A better fit gives higher probability to a pattern's genotypes.

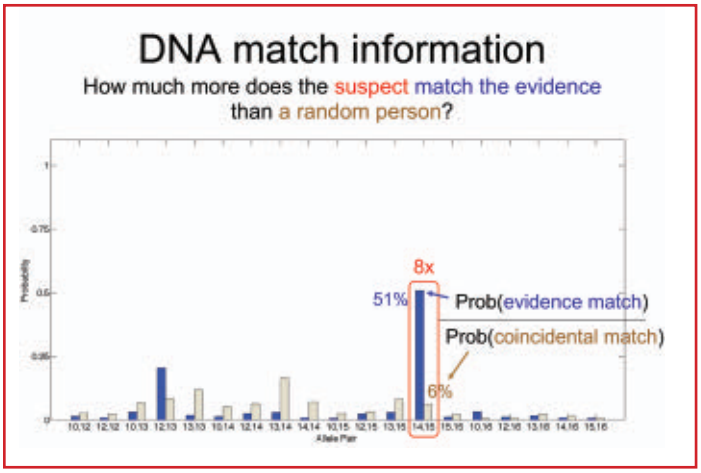

Figure 6: Genotype probability distributions. The bar chart shows the prior (brown) and posterior (blue) probabilities at a locus. Only genotype values (x-axis) with appreciable posterior probability (y-axis) are shown.

View references at www.forensicmag.com/articles/2014/08/kern-county-resolves-dna-mixture-crisis.

Dr. Mark Perlin *is Chief Scientific and Executive Officer for 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. perlin@cybgen.com; www.cybgen.com*
Dr. Kevin Miller *is the Laboratory Director and Acting DNA Technical Lead Criminalist at the Kern Regional Crime Laboratory.*