Human factors in automated forensic DNA examination

Expert Working Group on Human Factors in Forensic DNA Interpretation: Probabilistic Genotype Software

National Institute of Standards and Technology
National Institute of Justice
January 2021

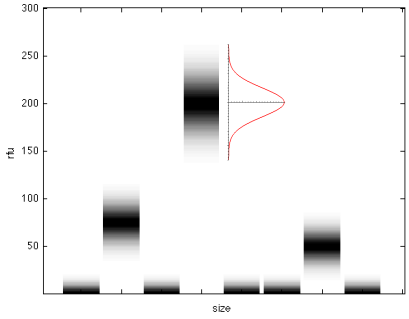
Mark W Perlin, PhD, MD, PhD
Cybergenetics
Pittsburgh, PA

Cybergenetics

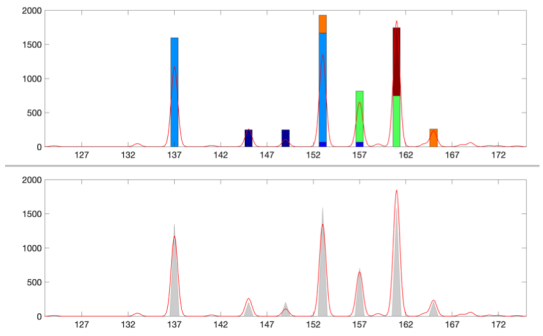
Working group questions

- How DNA complexity affects understanding and communication
- Explaining probabilistic genotyping results to a jury
- Validating a probabilistic genotyping system
- Selecting DNA samples in a validation study
- When do computers outperform a human analyst
- How the LR depends on reference database assumptions
- Any 'manufacturer's warnings' for TrueAllele crime labs
- Automatic genotype comparison with contamination databases

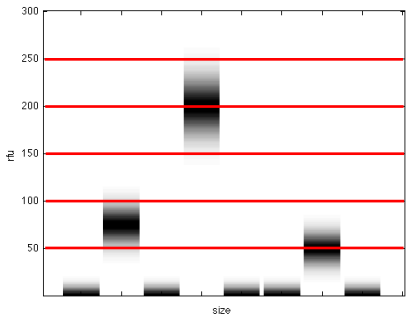
PCR introduces randomness



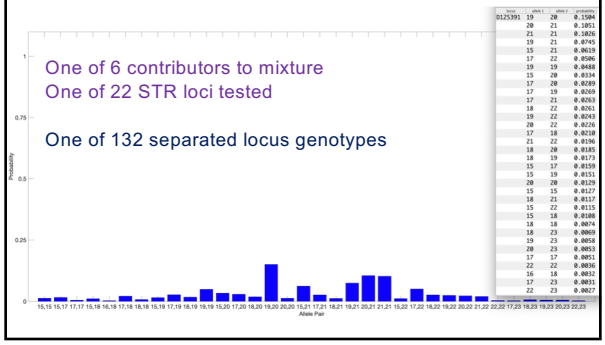
DNA mixtures – many answers



Thresholds introduce error



Find all genotypes, with probability



Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*. 2011;56(6):1430-1447.

Bayesian modeling

contributor components sum to one, so that $\sum_{k=1}^{K} w_{k,l} = 1$. The total DNA quantity at locus l is given by mass parameter m_l. A quantitative linear model of data pattern $\mathbf{d}_l$ at locus l has an expected vector value μ_l given by the weighted genotype sum

$$\mu_l = m_l \cdot \sum_{k=1}^{K} w_{k,l} \cdot \mathbf{g}_{k,l} \quad (8)$$

Additional model variables can include PCR stutter, relative amplification, DNA degradation, and dye separation (57).

A hierarchical model of mixture weight at every locus provides a better fit to the data (6). We therefore draw each individual locus weight $\mathbf{w}_l$ as a hierarchical prior from a common DNA template mixture weight $\mathbf{w}$ using a truncated (simplex) multivariate normal distribution as

$$\mathbf{w}_l \sim N_{[0,1]^K,\,\mathbf{1}}(\mathbf{w}, \psi^2 \cdot I) \quad (9)$$

The mixture weight covariance is an identity matrix scaled by a mixture variance ψ^2.

We write the peak data covariance matrix Σ_l as

$$\Sigma_l = \sigma^2 \cdot V_l + \tau^2 \quad (10)$$

where σ^2 is amplification dispersion, τ^2 is detection variation, and V_l is a diagonal matrix $diag(\mathbf{d}_l)$ of peak heights. We linearly model the data vector $\mathbf{d}_l$ using a truncated ($\geq \mathbf{0}$) multivariate normal distribution N_+ of the mean vector μ_l and covariance matrix Σ_l (18) as

$$\mathbf{d}_l \sim N_+(\mu_l, \Sigma_l) \quad (11)$$

Other square deviation data models can be used (47,58), as well as nonnormal distributions (19).

To infer the pmf $\varphi(x)$ of genotype $\mathbf{g}_{k,l}$, we form the joint probability distribution of Eq. 1 over all the relevant random variables (25). The likelihood function elements $\Pr\{d_{i,l} | \mathbf{g}_{k,l} = x_i, \ldots\}$ are given by Eq. 11. The prior probability assignments are given in Eqs (9) and (12).

$$\begin{aligned} \mathbf{g}_{k,l} &\sim \begin{cases} f_i^2, & i = j \\ 2f_i f_j, & i \neq j \end{cases} \\ \mathbf{w} &\sim Dir(\mathbf{1}) \\ m_l &\sim N_+(5000, 5000^2) \\ \sigma^{-2} &\sim Gam(10, 20) \\ \tau^{-2} &\sim Gam(10, 500) \\ \psi^{-2} &\sim Gam(1/2, 1/200) \end{cases} \quad (12) \end{aligned}$$

The genotype prior probability $\Pr\{\mathbf{g}_{k,l} = x\}$ at allele pair $x = [i\ j]$ is a product of population allele frequencies $\{f_i\}$. The template mixture weight $\mathbf{w}$ is assigned a uniform prior probability over the K contributor simplex. The locus mass m_l prior is a (nonnegative) truncated normal distribution on feasible total peak rfu values. The data variation parameters σ^2 and τ^2 have inverse gamma prior probability distributions, as does the mixture variance ψ^2.

Two-step objective approach

1. Without looking at the suspect's genotype, find:

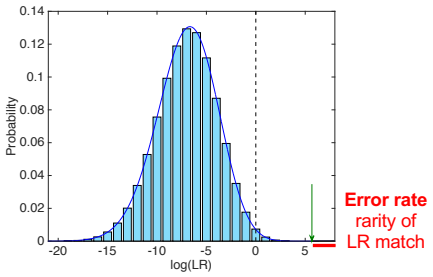
   Genotype probability q, for all allele pairs,
   from analyzing the evidence data
   for each contributor at every locus.

2. Afterwards, plug in the suspect's genotype s:

$$\text{LR(s)} = \frac{q(s)}{p(s)}$$

Genotype match distributions

Frequency distribution of noncontributor genotypes



Heliyon

ELSEVIER

Received:
9 July 2018
Revised:
18 September 2018
Accepted:
24 September 2018

Cite as:
Mark W. Perlin. Efficient construction of match strength distributions for uncertain multi-locus genotypes. Heliyon 4 (2018) e00824. doi: 10.1016/j.heliyon.2018.e00824

Efficient construction of match strength distributions for uncertain multi-locus genotypes

Mark W. Perlin*

Cybergenetics, Pittsburgh, PA, USA

*Corresponding author.
E-mail address: perlin@cybgen.com (M.W. Perlin).

Question 1

How DNA complexity affects understanding and communication

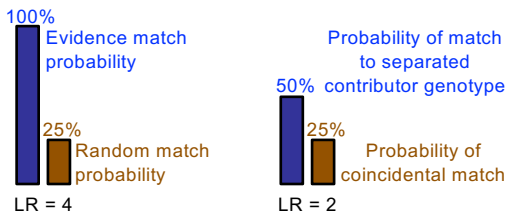
The more complex the analysis gets, the harder it is to communicate.

Are we reaching a point where we cannot understand how it works?

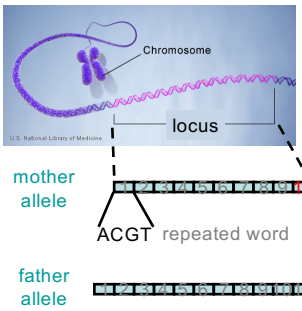
Do we just need to know how often we get it right?

Easy with separated genotypes

Complexity doesn't affect explanation.
Show results for one contributor's genotype.
Then explain just like RMP.



DNA genotype



A genetic locus has two DNA sentences, one from each parent.

An allele is the number of repeated words.

A genotype at a locus is a pair of alleles.

10, 12

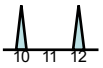
Many alleles allow for many many allele pairs. A person's genotype is relatively unique.

DNA evidence interpretation

Evidence item → Lab → Evidence data → Infer → Evidence genotype

10, 12 @ 100%

DNA from one person



Compare

Known genotype

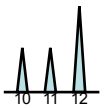
10, 12

DNA mixture interpretation

Evidence item → Lab → Evidence data → Infer → Evidence genotype

10, 11 @ 20%
11, 11 @ 30%
11, 12 @ 50%

DNA from two people



Compare

Known genotype

11, 12

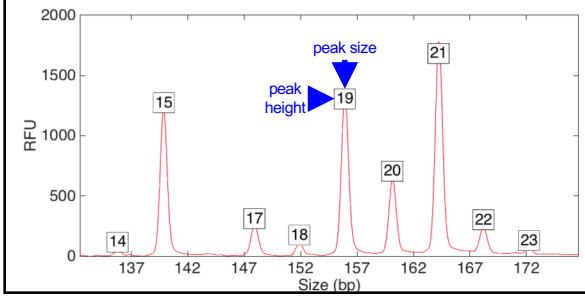
Question 2

Explaining probabilistic genotyping results to a jury

What is the most effective way to explain PGS outputs to a jury?

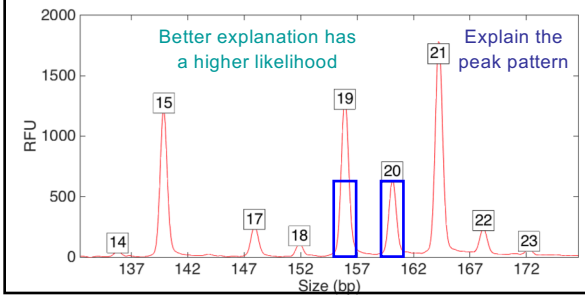
Computers can use all the data

Quantitative peak heights at locus D12S391



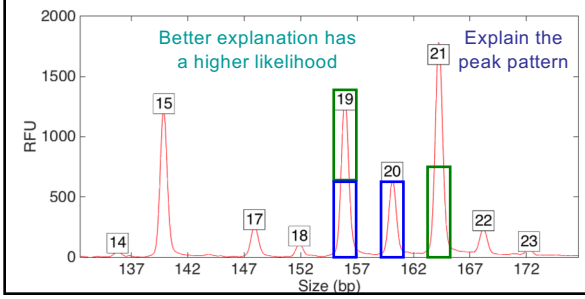
How the computer thinks

Consider every possible genotype solution



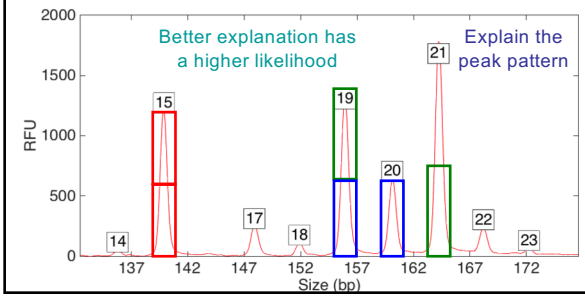
How the computer thinks

Consider every possible genotype solution



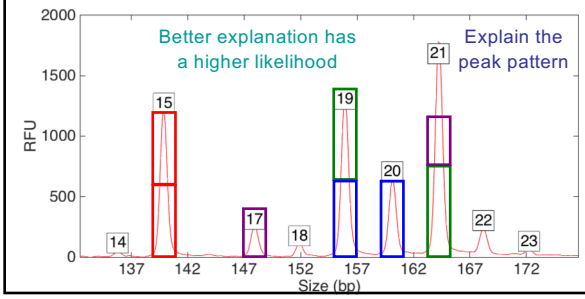
How the computer thinks

Consider every possible genotype solution



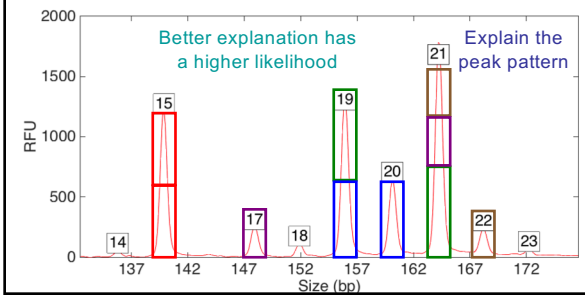
How the computer thinks

Consider every possible genotype solution



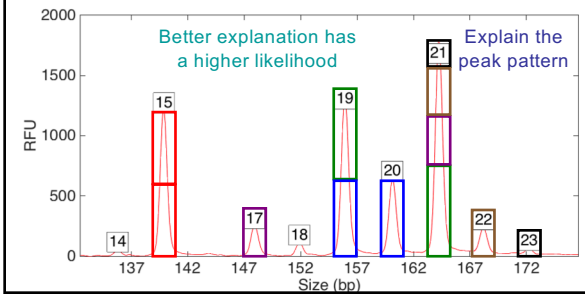
How the computer thinks

Consider every possible genotype solution



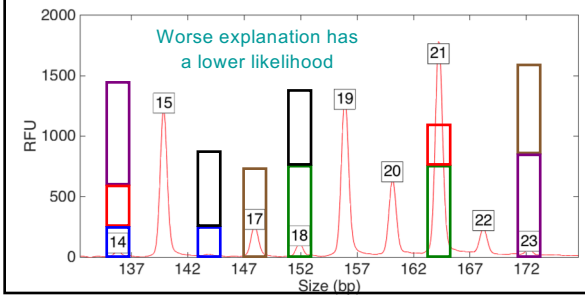
How the computer thinks

Consider every possible genotype solution



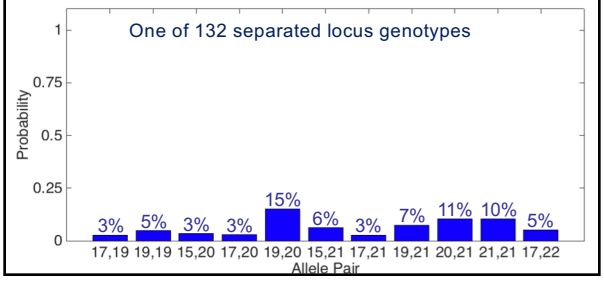
How the computer thinks

Consider every possible genotype solution



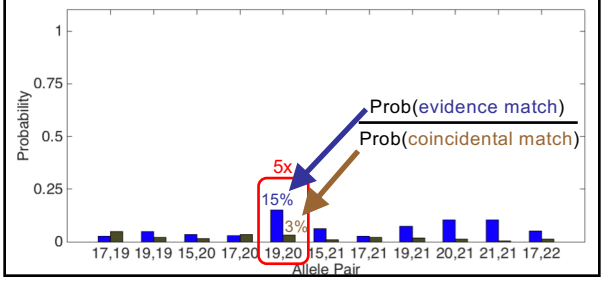
Evidence genotype

Objective genotype determined solely from the DNA data.
Never sees a comparison reference.

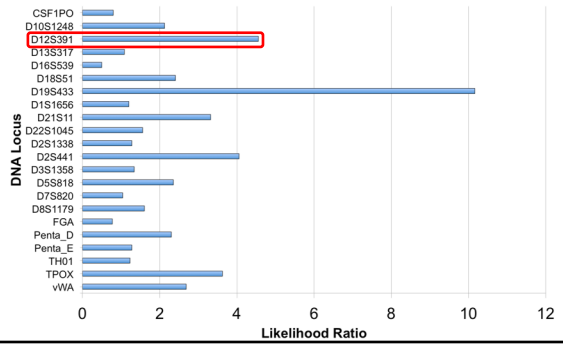


DNA match information

How much more does the suspect match the evidence
than a random person?



Match information at 22 loci



Is the suspect in the evidence?

A match between the khaki pants
and the defendant is:

470 thousand times more probable than
a coincidental match to an unrelated African-American person

7.64 million times more probable than
a coincidental match to an unrelated Caucasian person

5.12 million times more probable than
a coincidental match to an unrelated Hispanic person

Match statistics

Item	Description	007 Victim	009 Defendant
008	khaki pants	204 trillion	470 thousand

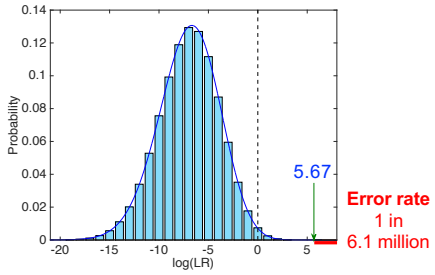
LR

Match statistics

Item	Description	007 Victim	009 Defendant
008	khaki pants	14.31	5.67

log(LR)

Non-contributor analysis



Question 3

Validating a probabilistic genotyping system

What is involved in a validation of a Probabilistic Genotyping System?

Validation metrics

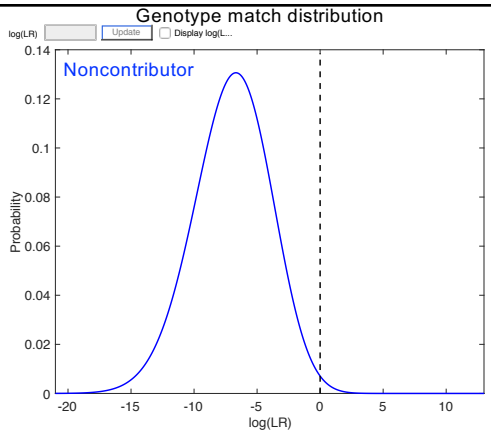
A. Specificity
B. Sensitivity
C. Reproducibility
D. Accuracy
E. Predictability

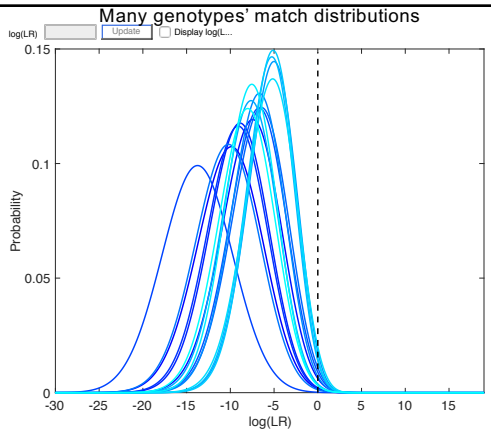
A. Specificity

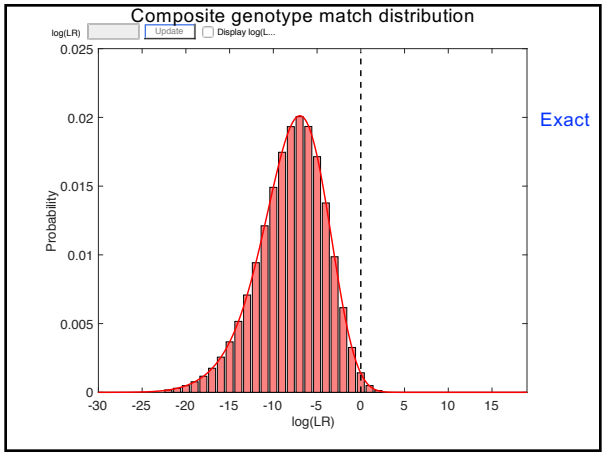
The extent to which interpretation does not misidentify the wrong person

True exclusions, without false inclusions

Average of the individual genotype specificity distributions



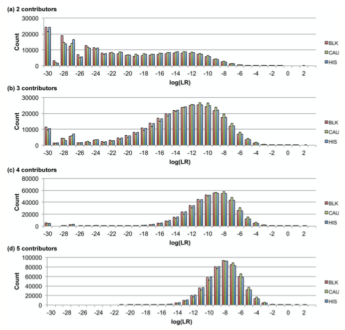




Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*. 2015;60(4):857-868.

TrueAllele specificity

Sampled



Exact vs. sampled

Exact

- all – 10^{24} genotypes
- accurate
- exact probability function
- convolution – fast

Sampled

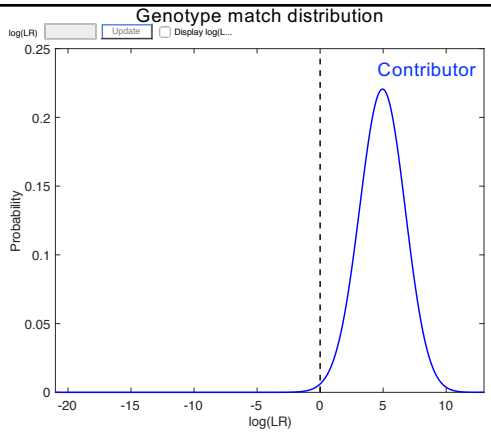
- some – 10^{4} genotypes
- approximate
- sample using random profiles
- Monte Carlo – slow

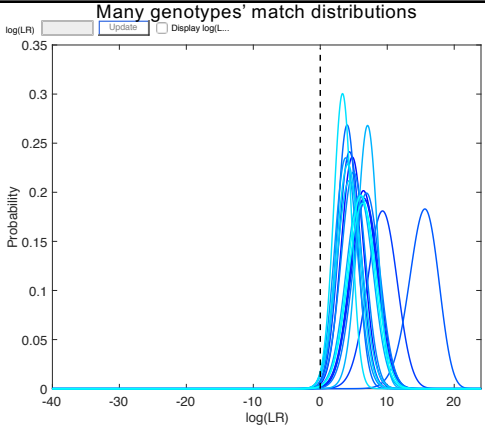
B. Sensitivity

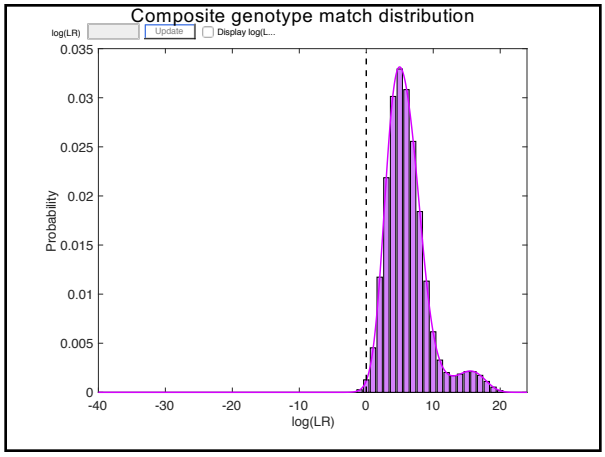
The extent to which interpretation identifies the correct person

True DNA mixture inclusions

Average of the individual genotype specificity distributions

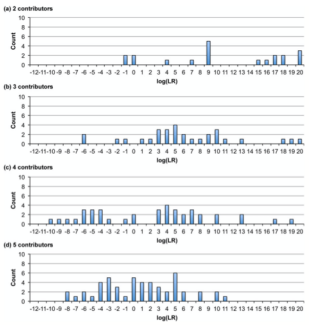






Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*. 2015;60(4):857-868.

TrueAllele sensitivity



C. Reproducibility

The extent to which interpretation gives the same answer to the same question

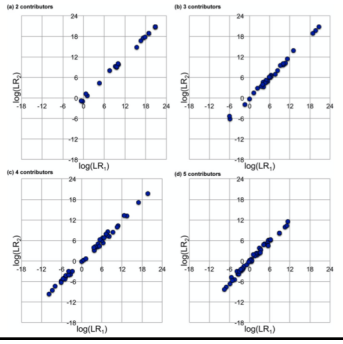
MCMC computing has sampling variation

duplicate computer runs on the matching genotypes

measure log(LR) variation

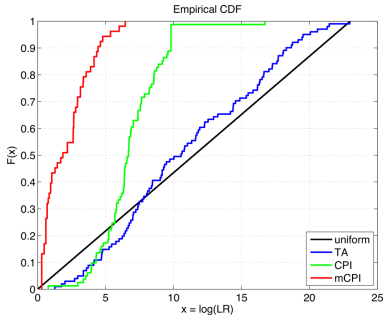
Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*. 2015;60(4):857-868.

TrueAllele reproducibility



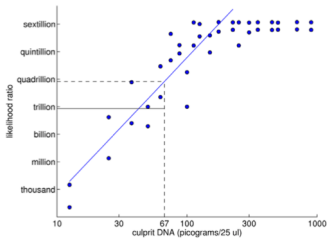
Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. TrueAllele® Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. *PLOS ONE*. 2014;(9)3:e92837.

D. TrueAllele accuracy



Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS ONE. 2009;4(12):e8327.

E. TrueAllele predictability



Question 4

Selecting DNA samples in a validation study

How do you select samples for validation?

Are the samples pristine or degraded?

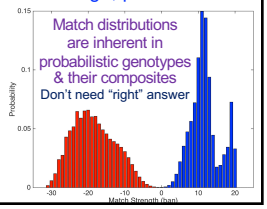
Will different sample types produce different results?

Representative of casework

Typically, uniform aliquots – unrealistic, unrepresentative

Realistic mixtures
- laboratory synthesized – randomized design, pristine
- casework items – most realistic

Both kinds of DNA samples are scientifically workable and informative



Question 5

When do computers outperform a human analyst

What are the circumstances in which PGS outperforms an analyst?

Is there a way to make these criteria more subjective?

JOURNAL OF FORENSIC SCIENCES

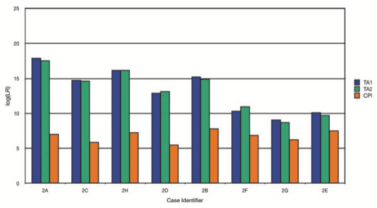
J Forensic Sci, November 2011, Vol. 56, No. 6
doi: 10.1111/j.1556-4029.2011.01859.x
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Mark W. Perlin,[1] M.D., Ph.D.; Matthew M. Legler,[1] B.S.; Cara E. Spencer,[1] M.S.; Jessica L. Smith,[1] M.S.; William P. Allan,[1] M.S.; Jamie L. Belrose,[2] M.S.; and Barry W. Duceman,[3] Ph.D.

Validating TrueAllele® DNA Mixture Interpretation*,†



JOURNAL OF FORENSIC SCIENCES

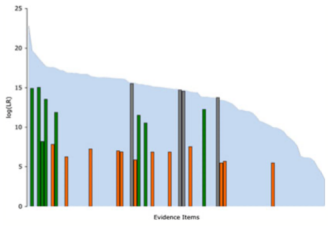
J Forensic Sci, November 2013, Vol. 58, No. 6
doi: 10.1111/1556-4029.12223
Available online at: onlinelibrary.wiley.com

PAPER

CRIMINALISTICS

Mark W. Perlin,[1] M.D., Ph.D.; Jamie L. Belrose,[2] M.S.; and Barry W. Duceman,[3] Ph.D.

New York State TrueAllele® Casework Validation Study*



Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. TrueAllele® Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. *PLOS ONE*. 2014;(9)3:e92837.

Higher human error

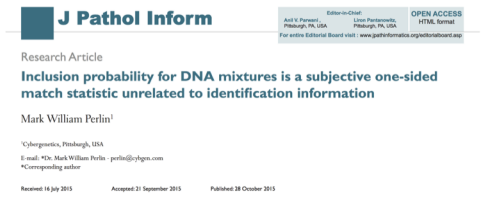
TrueAllele specificity (million samples)
From noncontributor distribution, for LR > 100:
Error rate = 1 in 1,000,000 (**0.0001**)%

CPI – analytical threshold
5 false positives in 81 comparisons
Error rate = 5 in 81 (**6%**)

mCPI – stochastic threshold
17 inconclusive results
1 false positive in 53 comparisons
Error rate = 1 in 53 (**2%**)

Human review should not be done

CPI just counts subjective number of reported loci



J Pathol Inform

Research Article

Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information

Mark William Perlin[1]

[1]Cybergenetics, Pittsburgh, USA

E-mail: *Dr. Mark William Perlin - perlin@cybgen.com
*Corresponding author

Received: 16 July 2015 Accepted: 21 September 2015 Published: 28 October 2015

TrueAllele Casework Workflow

ViewStation
User Client

Database
Server

Interpret/Match
Expansion

Visual User Interface
VUIer™ Software

Parallel Processing Computers

Automated operation

Human involvement
- optional
- no key decisions

Number of contributors
- not needed
- sufficient number
- based on EPG data

Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences.* 2015;60(4):857-868.

Bauer DW, Butt N, Hornyak JM, Perlin MW. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *Journal of Forensic Sciences.* 2020;65(2):380-398.

Question 6

How the LR depends on reference database assumptions

How much does the LR depend on the assumptions that go into the reference database?

TrueAllele calibration-free

TrueAllele has a full Bayesian model
Does not use laboratory "calibration"
Variables derived from evidence data

Baseline variation – no analytic thresholds
PCR variation – no stochastic thresholds
Stutter parameters – no calibration; learn from data

Question 7

Any 'manufacturer's warnings' for TrueAllele crime labs

What 'manufacturer's warnings' do you presently have for labs who purchase licenses to your software?

No ‘warnings’

None needed.
Labs learn methods and test system.
Then validate to their comfort level.

Question 8

Automatic genotype comparison with contamination databases

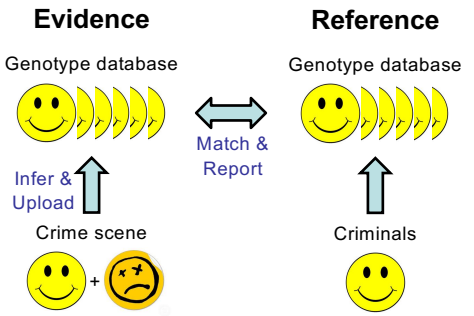
It seems computationally easy now for labs to run PGS against contamination databases, and
to get LRs for the profiles in those databases.
Should labs routinely do these LR database checks on casework,
in order to ferret out low level contamination, or
to give fact-finders more context for low suspect LRs?

Automated TrueAllele database

Human-free genotype matching

Yes – TrueAllele labs auto-check contamination
Automated pre-review of cases (seed & harvest)
Automated DNA investigation (property crime)
Automated post-review of past cases (open the past)

Upload and compare all DNA



Recommendations for forensic genotyping practice

Use TrueAllele computers to fully automate DNA interpretation

Computers solve 100-dimensional problems that people can't

Eliminate people from the interpretation process

waste, cost, time, error, labor, limits, bias, impact, …

Have people interact with society, explain results

Scientific literature and technical knowledge

- Peer-reviewed journal articles
- Internal validation studies
- Academic thesis papers
- Manufacturer method reports
- Data-rich white papers
- Patent specifications
- On-line talks & tutorials

Empirical testing

Peer-reviewed validation studies

Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. *PLoS ONE*. 2009;4(12):e8327.

Ballantyne J, Hanson EK, Perlin MW. DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: Combining quantitative data for greater identification information. *Science & Justice*. 2013;53(2):103-114.

Perlin MW, Hornyak J, Sugimoto G, Miller K. TrueAllele® genotype identification on DNA mixtures containing up to five unknown contributors. *Journal of Forensic Sciences*. 2015;60(4):857-868.

Greenspoon SA, Schiermeier-Wood L, Jenkins BC. Establishing the limits of TrueAllele® Casework: a validation study. *Journal of Forensic Sciences*. 2015;60(5):1263-1276.

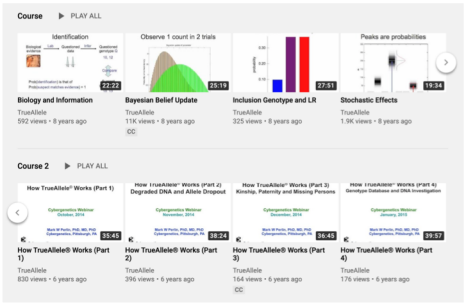
Bauer DW, Butt N, Hornyak JM, Perlin MW. Validating TrueAllele® interpretation of DNA mixtures containing up to ten unknown contributors. *Journal of Forensic Sciences*. 2020;65(2):380-398.

Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*. 2011;56(6):1430-1447.

Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® Casework validation study. *Journal of Forensic Sciences*. 2013;58(6):1458-1466.

Perlin MW, Dormer K, Hornyak J, Schiermeier-Wood L, Greenspoon S. TrueAllele® Casework on Virginia DNA mixture evidence: computer and manual interpretation in 72 reported criminal cases. *PLOS ONE*. 2014;(9)3:e92837.

TrueAllele lectures



More information

http://www.cybgen.com/information

- Courses
- Newsletters
- Newsroom
- Presentations
- Publications
- Webinars

http://www.youtube.com/user/TrueAllele
TrueAllele YouTube channel

Cybergenetics

YouTube

perlin@cybgen.com