1

Probabilistic genotyping of evidentiary DNA typing results: TrueAllele® technology

Federal Bureau of Investigation
& National Institute of Justice
May, 2019

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

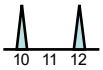
Cybergenetics

2

Genotyping

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

Simple DNA from one person



10, 12 @ 100%

Compare

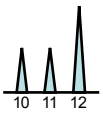
Known genotype
10, 12

3

Probability

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

Complex DNA from two people



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

Compare

Known genotype
11, 12

4

Bayes theorem

Evidence updates belief

Pr{belief|evidence} ∝ Pr{evidence|belief} × Pr{belief}

Posterior belief	**Likelihood**	**Prior belief**
after seeing evidence	how well belief explains the evidence	*before* seeing evidence

5

Probabilistic genotyping

STR data updates genotype probability

Pr{genotype|data} ∝ Pr{data|genotype} × Pr{genotype}

Posterior genotype	**Likelihood**	**Prior genotype**
after seeing data	how well genotype explains the data	*before* seeing data

6

Likelihood ratio

Measures identification information in DNA evidence

Posterior probability

$$LR(geno) = \frac{q(geno)}{p(geno)}$$

Prior probability

Perlin, M.W. Inclusion probability is a likelihood ratio: Implications for DNA mixtures. *Promega's Twenty First International Symposium on Human Identification*. 2010 7

Probability of inclusion

CPI was the first "probabilistic genotyping" method

Pr{genotype|data} ∝ Pr{data|genotype} × Pr{genotype}

Posterior genotype	**Likelihood**	**Prior genotype**
raises selected allele frequencies	all-or-none based on peak threshold	population frequency

Makes little use of the data
Equally weights genotypes
Largely **uninformative**

$$LR_{CPI}(geno) = 1/CPI(geno)$$

8

Genotyping issues

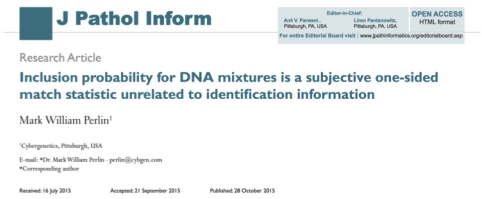
CPI "ProbGen" had Problems

- peak thresholds
- drop out & drop in
- data variance
  (*aka* stochastic effects)
- stutter artifacts
- allele sharing
- data & parameter choices
- subjectivity & bias
- "inconclusive" results
- inaccurate match statistics
- lack of validation

9

CPI failure

Can't work, won't work, will never work



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

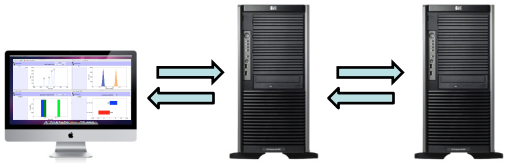
10

Lessons learned

- use all the data
  - all peaks, all loci
  - no data choices
- deliver accurate results
  - prove reliability
  - have a valid basis
- fully automate the process
  - no people, no bias
  - no parameter choices

11

TrueAllele® solution



ViewStation
User Client

Visual User Interface
VUIer™ Software

Database
Server

Interpretation
Servers

Parallel Processing Computers
10 to 100 simultaneous questions
2 hours for each question

12

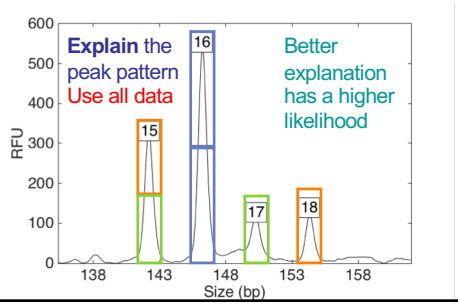
TrueAllele infers genotypes

- separates DNA mixtures (10 unknowns)
- handles very low-level DNA amounts
- no thresholds or data choices
- no setting or parameter choices
- solves maternity, paternity & kinship
- finds missing person genotypes
- combines multiple STR tests
- uses all the data
- delivers accurate results
- fully automates the process

13

How the computer thinks

Consider every possible genotype solution



14

TrueAllele matches genotypes

- calculates likelihood ratios
- adjusts for co-ancestry
- reports LR error
- visually displays results
- compares evidence-to-evidence
- has a built-in genotype database
- automates familial search
- uses all the data
- delivers accurate results
- fully automates the process

15

Standard likelihood ratio

Posterior *q* over prior *p*

Information is the change in probability

Posterior Q $q(r)$

Prior P $p(r)$

Evaluated at a known reference *r* genotype

Perlin, M.W. Explaining the likelihood ratio in DNA mixture interpretation. *Promega's Twenty First International Symposium on Human Identification*. 2010 16

Evidence-to-evidence LR

Posterior Q Posterior R

$$LR = \sum_{x \in G} \frac{q(x)r(x)}{p(x)}$$

Prior P

Sum over **unknown** genotypes *x*

17

TrueAllele history

1995. Published stutter deconvolution
1997. Published Bayes MCMC mapping
1999. Implemented mixture deconvolution
2000. Filed first U.S. mixture patent
2001. Published linear mixture analysis
2004. First U.S. mixture patent issued
2009. Published first validation paper
2009. First trial & admissibility challenge
2012. First appellate precedent set
2014. U.S. & European patents issued
2015. Published seventh validation paper
2016. First TrueAllele exoneration
2017. Third U.S. mixture patent issued
2019. Case reports issued in 43 states

18

2005. World Trade Center

18,000
victim remains

match

2,700
missing people

19

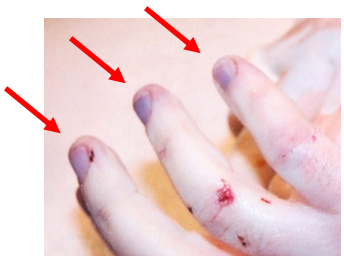
2009. Pennsylvania vs. Foley

World's first "probabilistic genotyping" trial

November 2007: Kevin Foley charged with murder

20

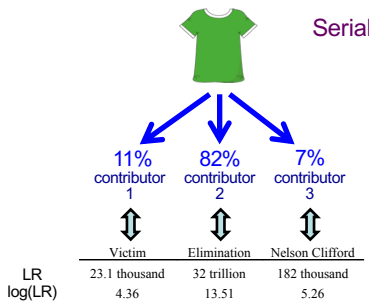
Fingernail DNA Evidence



93.3% victim + 6.7% DNA component

21

2014. Maryland v. Clifford



22

2014. United States v. Tsarnaev

Boston Marathon Bombing

Five person mixture

30% Dzhokhar

30% Tamerlan

15%

15%

10%

Separate sibling genotypes

23

2016. Indiana v. Pinkins

1989 – 5 men raped an Indiana woman
Darryl Pinkins and 2 others misidentified
1991 – wrongfully convicted, 65 year sentence

2001 – DNA mixture evidence
2 contributors found, not the accused
but 5 were needed, post-conviction relief denied

24

TrueAllele match table

References

		V	P	G	J	S	H	JJ	JS
	J	-15.39	-18.00	-15.39	11.07	1.21	4.15	-6.19	-7.66
	S	-15.39	-15.17	-15.39	1.21	10.22	-2.39	-5.81	-8.09
	H	-15.39	-18.00	-15.39	4.15	-2.39	10.31	-3.52	-5.80
	JJ	-2.78	-8.50	-8.43	-6.19	-5.81	-3.52	7.05	-1.49
	JS	-10.47	-5.79	-12.60	-7.66	-8.09	-5.80	-1.49	8.06

25

TrueAllele Pinkins findings

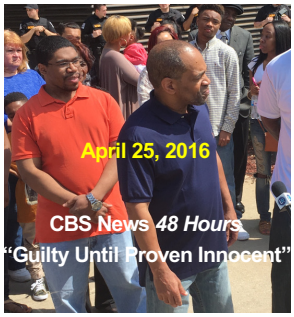
1. compared *evidence with evidence*
2. calculated *exclusionary match statistics*
3. revealed 5% *minor mixture contributor*
4. *jointly analyzed* DNA mixture data
5. showed three perpetrators were *brothers*

found 5 unidentified genotypes,
defendants not linked to the crime

Pinkins & Glenn exonerated

26

Pinkins released after 24 years



27

TrueAllele validation papers

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.

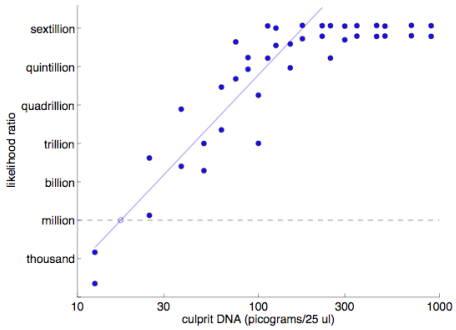
Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*. 2011;56(6):1430-1447.

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.

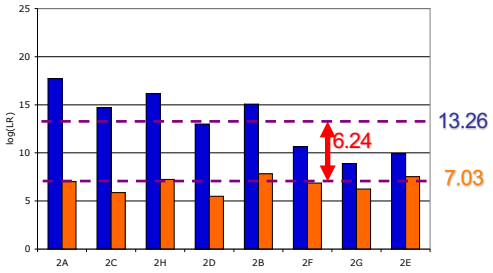
28

Accurate & predictable



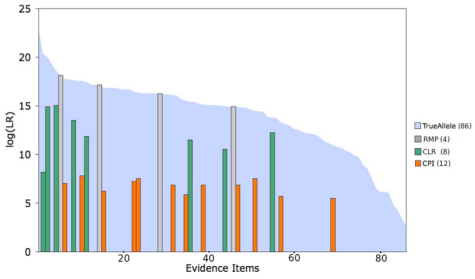
29

More information



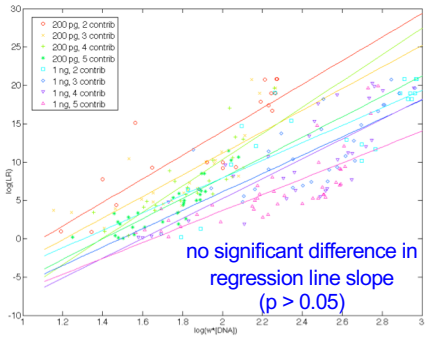
30

Less “inconclusive”



Any number of contributors or DNA amount

31

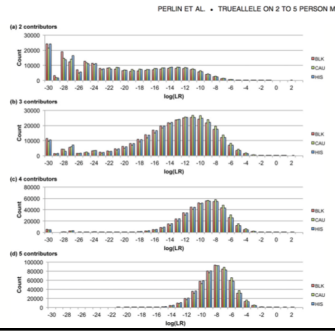


no significant difference in regression line slope ($p > 0.05$)

Perlin, M.W. Scientific validation of mixture interpretation methods. *Promega's Seventeenth International Symposium on Human Identification*. 2006

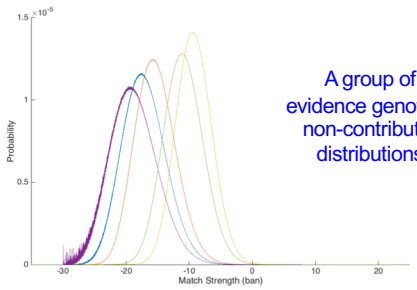
32

LR specificity



33

Validation averages genotypes

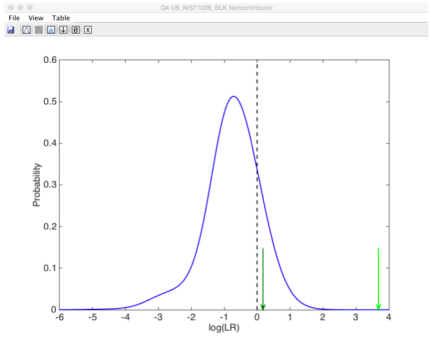


A group of evidence genotype non-contributor distributions

34

Perlin, M.W. Efficient construction of match strength distributions for uncertain multi-locus genotypes. *Heliyon*, 4(10):e00824, 2018.

One genotype: LR error rate



35

Verbal equivalents? (hide LR)

LR for (H_1) and 1/LR for (H_2)	Qualitative Equivalent
1	Uninformative
2 to <100	Limited Support
100 to <10,000	Moderate Support
10,000 to <1,000,000	Strong Support
≥1,000,000	Very Strong Support

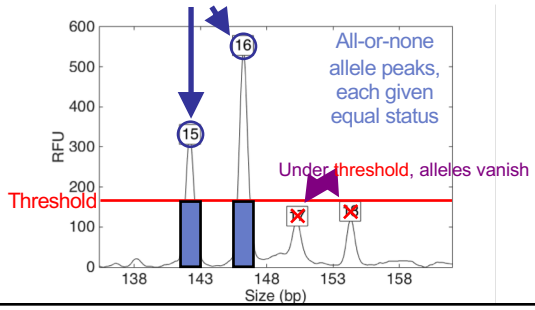
- hides the real match strength information
- not what a DNA expert actually believes
- misleads the jury about "million" (Koehler)

Just report LR error, along with the LR, when the match strength is under a million

36

Thresholds? (discard data)

Over threshold, peaks are labeled as allele events



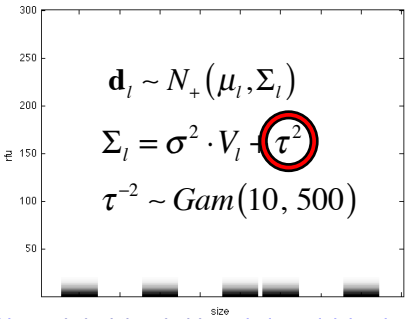
Perlin, M.W. and Sinelnikov, A. An information gap in DNA evidence interpretation. *PLoS ONE*, 4(12):e8327, 2009. 37

Background noise

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

$$\Sigma_l = \sigma^2 \cdot V_l + \tau^2$$

$$\tau^{-2} \sim Gam(10, 500)$$



No *analytical threshold* needed; model the data

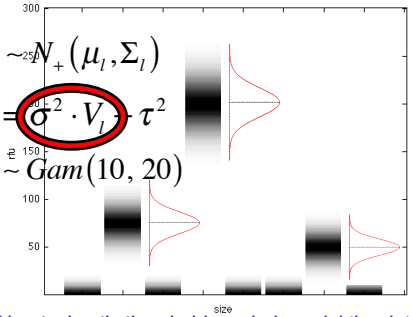
Perlin, M.W., Legler, M.M., Spencer, C.E., Smith, J.L., Allan, W.P., Belrose, J.L., and Duceman, B.W. Validating TrueAllele® DNA mixture interpretation. *Journal of Forensic Sciences*, 56(6):1430-47, 2011. 38

PCR variation

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

$$\Sigma_l = \sigma^2 \cdot V_l + \tau^2$$

$$\sigma^{-2} \sim Gam(10, 20)$$

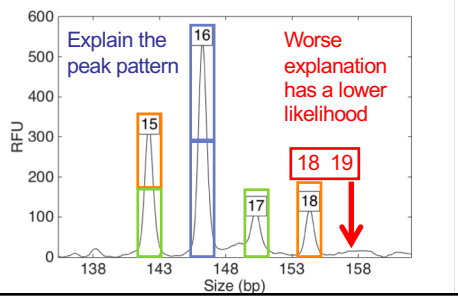


No *stochastic threshold* needed; model the data

39

Dropout? (concoct data)

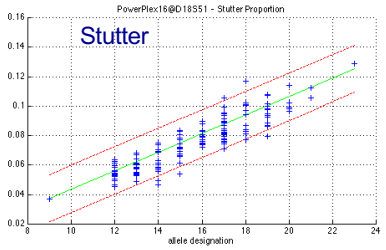
No allele data, lower genotype probability, lower match statistic



Perlin, M.W. Expert systems for automated STR analysis, *Scientific Working Group on DNA Analysis Methods (SWGDAM)*, Quantico, VA, 14 January 2003. 40

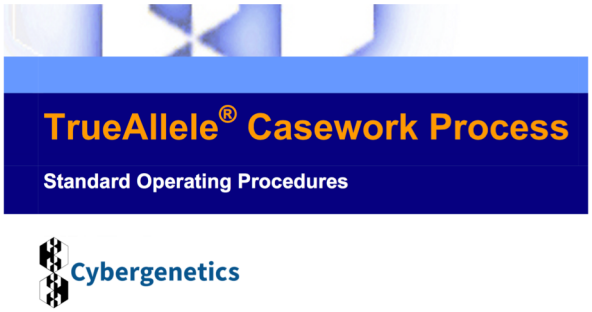
Calibration? (old data)

Fully Bayesian model learns parameters from data
Eliminates calibration bias and limitations



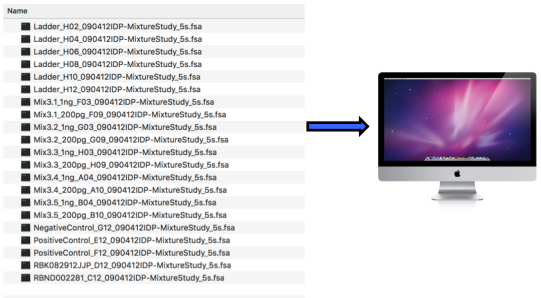
41

Guided Tour



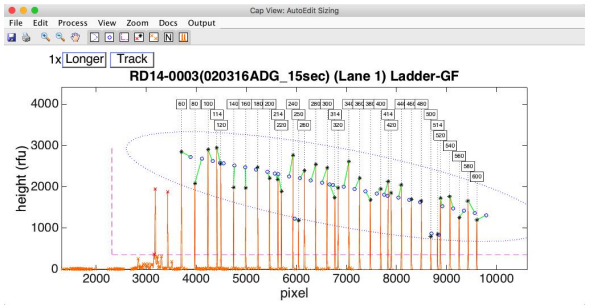
42

Enter DNA data files



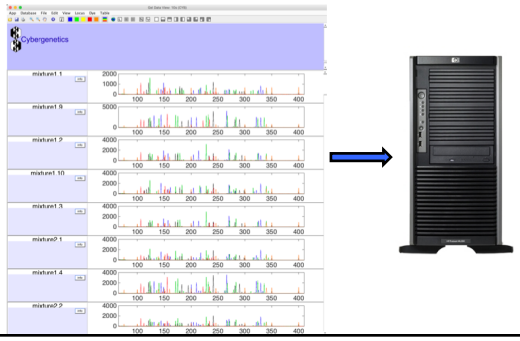
43

Automated size tracking



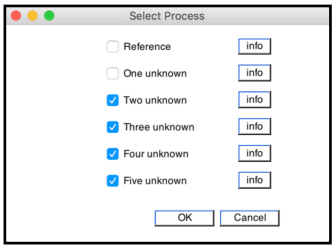
44

Upload DNA data



45

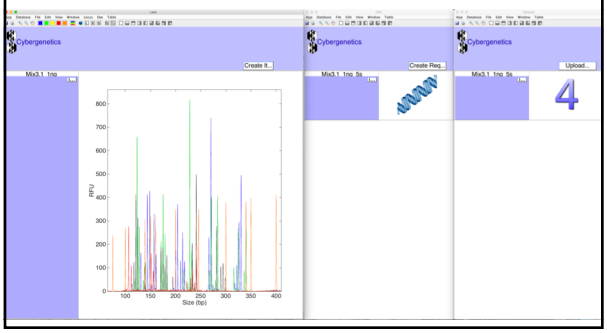
Automatic request upload



No need to choose the “right” number of contributors

46

Custom request upload



47

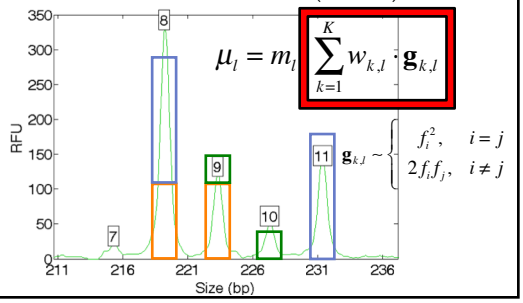
Perlin, M.W. and Szabady, B. Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *Journal of Forensic Sciences*, 46(6), 1372-77, 2001.

Unmixing genotypes

$$\mathbf{d}_l \sim N_+\left(\mu_l, \Sigma_l\right)$$

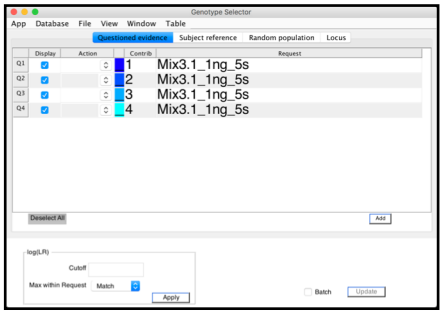
$$\mu_l = m_l \sum_{k=1}^{K} w_{k,l} \cdot \mathbf{g}_{k,l}$$

$$\mathbf{g}_{k,l} \sim \begin{cases} f_i^2, & i = j \\ 2 f_i f_j, & i \neq j \end{cases}$$



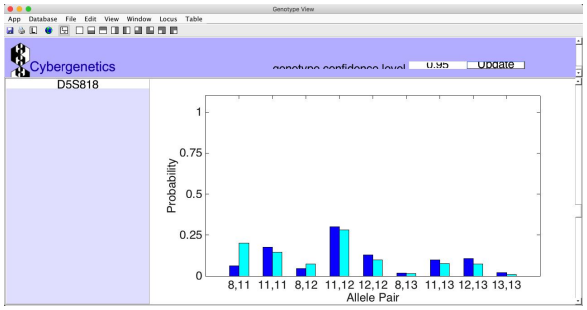
48

Genotype selector



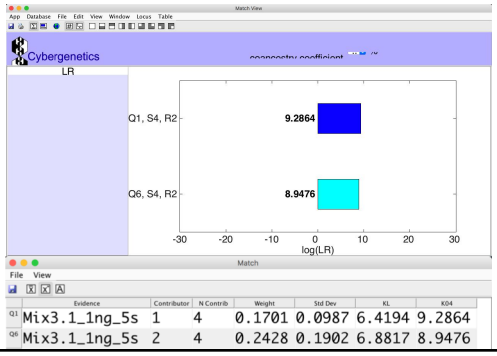
49

Genotype window



50

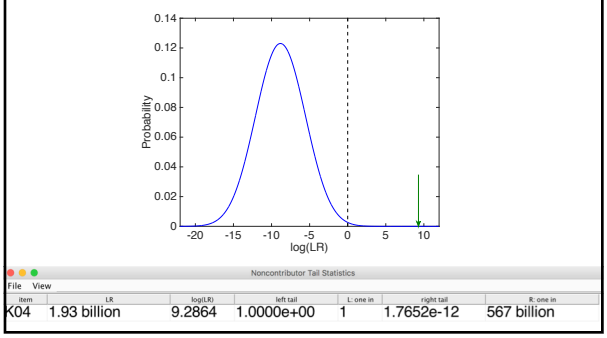
Match window

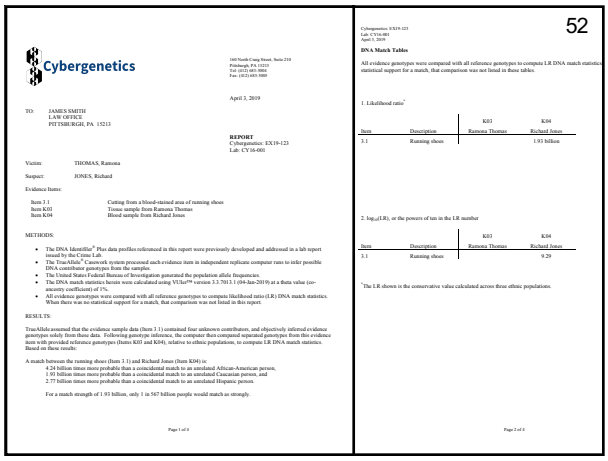


51

LR error rate

Probability of misleading evidence





Cybergenetics

165 North Craig Street, Suite 210
Pittsburgh, PA 15213
Tel: (412) 683-3004
Fax: (412) 683-3005

April 3, 2019

TO: JAMES SMITH
LAW OFFICE
PITTSBURGH, PA 15213

REPORT
Cybergenetics: EX19-123
Lab: CY16-001

Victim: THOMAS, Ramona

Suspect: JONES, Richard

Evidence Items:

Item 3.1	Cutting from a blood-stained area of running shoes
Item K03	Tissue sample from Ramona Thomas
Item K04	Blood sample from Richard Jones

METHODS:

- The DNA Identifiler® Plus data profiles referenced in this report were previously developed and addressed in a lab report issued by the Crime Lab.
- The TrueAllele® Casework system processed each evidence item in independent replicate computer runs to infer possible DNA contributor genotypes from the samples.
- The United States Federal Bureau of Investigation generated the population allele frequencies.
- The DNA match statistics herein were calculated using VUlan™ version 3.2.701.1 (04-Jan-2019) at a theta value (co-ancestry coefficient) of 1%.
- All evidence genotypes were compared with all reference genotypes to compute likelihood ratio (LR) DNA match statistics. When there was no statistical support for a match, that comparison was not listed in this report.

RESULTS:

TrueAllele assumed that the evidence sample data (Item 3.1) contained four unknown contributors, and objectively inferred evidence genotypes solely from those data. Following genotype inference, the computer then compared separated genotypes from this evidence item with provided reference genotypes (Items K03 and K04), relative to ethnic populations, to compute LR DNA match statistics. Based on these results:

A match between the running shoes (Item 3.1) and Richard Jones (Item K04) is:

- 4.24 billion times more probable than a coincidental match to an unrelated African-American person,
- 1.93 billion times more probable than a coincidental match to an unrelated Caucasian person, and
- 2.77 billion times more probable than a coincidental match to an unrelated Hispanic person.

For a match strength of 1.93 billion, only 1 in 567 billion people would match as strongly.

Page 1 of 4

Cybergenetics: EX19-123
Lab: CY16-001
April 3, 2019

DNA Match Tables

All evidence genotypes were compared with all reference genotypes to compute LR DNA match statistics. When there was no statistical support for a match, that comparison was not listed in these tables.

1. Likelihood ratio*

Item	Description	K03 Ramona Thomas	K04 Richard Jones
3.1	Running shoes		1.93 billion

2. $\log_{10}$(LR), or the powers of ten in the LR number

Item	Description	K03 Ramona Thomas	K04 Richard Jones
3.1	Running shoes		9.29

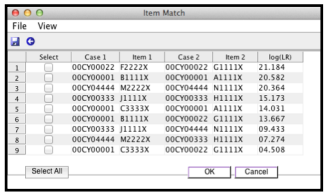
*The LR shown is the conservative value calculated across three ethnic populations.

Page 2 of 4

52

53

TrueAllele genotype database



All genotype uploaded
All genotypes compared

54

Perlin, M.W. Investigative DNA databases that preserve identification information. *American Academy of Forensic Sciences 64th Annual Meeting*, 2012.

CODIS allele database

Type of DNA database	Sensitivity	Specificity
Probabilistic genotype	LR average is about a quadrillion	False positive rate < 0.01%
Allele list (moderate stringency)	Upload fails about 1/3 of the time	Hits 5%-25% of the DNA database

Perlin, M.W. Explaining the likelihood ratio in DNA mixture interpretation. *Promega's Twenty First International Symposium on Human Identification*. 2010 55

LR proof sketch

$$LR(geno) = \frac{q(geno)}{p(geno)}$$

$$= \frac{Pr\{geno|data\}}{Pr\{geno\}}$$

Uses Bayes theorem

$$= \frac{Pr\{data|geno\}}{Pr\{data\}}$$

$$= \frac{Pr\{data|H\}}{Pr\{data|\sim H\}}$$

56

LR hypotheses? (bad math)

$$LR(geno) = \frac{Pr\{data|H\}}{Pr\{data|\sim H\}}$$

With H_p = H, and H_d = ~H, we have:

$$LR(geno) = \frac{Pr\{data|H_p\}}{Pr\{data|H_d\}}$$

But if not, the LR expression may be meaningless.

57

TrueAllele transparency

Cybergenetics standard disclosure

- validation reports and articles
- validation testing data
- error rate methods and results
- standards compliance documents
- standard operating procedures
- peer-reviewed scientific papers
- mathematical method descriptions
- general acceptance information
- Daubert/Frye admissibility decisions
- tutorials for lawyers and scientists
- software user manuals
- software installer for case review
- case data for reviewing results
- cloud access for independent testing
- source code access for defense experts

58

Throughput

At 2 hours for simple mixtures,
a basic 12-core TrueAllele system
solves 100 mixtures per day.

And an advanced 100-core system
solves 1,000 mixtures per day.

More complex mixtures take more time,
but without limits on the number of contributors
or the amount of DNA.

59

Open the past

We can overcome the past failures
of DNA mixture interpretation.

There have been 100,000's of mixtures
wrongly reported as "inconclusive"
or given inaccurate match statistics.

TrueAllele automation can open the past.
It can use all the data, without people,
to accurately reprocess old cases.

To free the innocent, and find the guilty.

60

More TrueAllele information

http://www.cybgen.com/information

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

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

Cybergenetics

perlin@cybgen.com