TrueAllele® Interpretation of DNA Mixture Evidence

**9th International Conference on
Forensic Inference and Statistics
August, 2014
Leiden University, The Netherlands**

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

Cybergenetics

Cybergenetics © 2003-2014

TrueAllele® Casework

ViewStation
User Client

Database
Server

Interpret/Match
Expansion

Visual User Interface
VUIer™ Software

Parallel Processing Computers
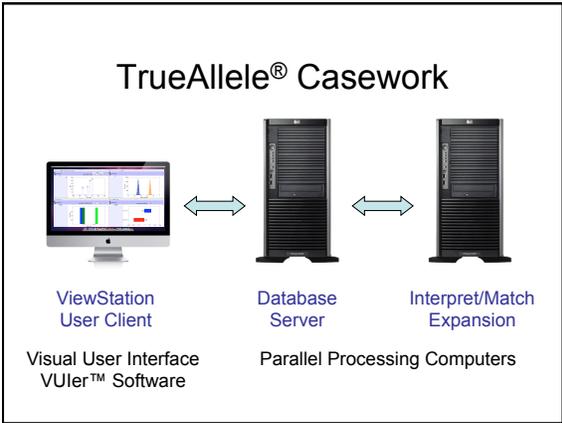
On the Origin of TrueAllele

1993 @ CMU: stutter deconvolution
American Journal of Human Genetics

1999 @ Cybergenetics: mixture deconvolution
Journal of Forensic Sciences

Scope: STR mixtures, degraded, kinship, database

Data: (respect) use everything, add nothing
Objective: never consider suspect reference
General: same for evidentiary & investigative

Variation Under Domestication

Single source DNA → locus → 8 [DNA icon] 12 [DNA icon]

Hypothesis: evidence and suspect share a common contributor

Data analysis: simple threshold

peak height

peak size

Variation Under Nature

victim ⇒ 8 [DNA icon] 12 [DNA icon]

other ⇒ 10 [DNA icon] 13 [DNA icon]

Hypothesis: evidence and suspect share a common contributor

Data analysis: patterns & variation

victim

other

Struggle for Existence

Likelihood ratio (LR) requires genotype probability

$$LR = \frac{O(\,H\,|\,\text{data})}{O(\,H\,)}$$

Bayes theorem + probability + algebra …

$$= \frac{\Sigma x \; P\{d_X | X=x, \ldots\} \; P\{d_Y | Y=x, \ldots\} \; P\{X=x\}}{\Sigma\Sigma x,y \; P\{d_X | X=x, \ldots\} \; P\{d_Y | Y=y, \ldots\} \; P\{X=x, Y=y\}}$$

genotype probability: posterior, likelihood & prior

Natural Selection

Mixture weight variables

Hierarchical Bayesian model induces a set of forces in a high-dimensional parameter space

α β

ω σ_ω

G γ δ

Genotype variables

$l=1,\ldots,n_l$ ω_l Hierarchical mixture weight locus variables

ϕ_l

X^2

• small DNA amounts
• degraded contributions
• K = 1, 2, 3, 4, 5, 6, ... unknown contributors
• joint likelihood function
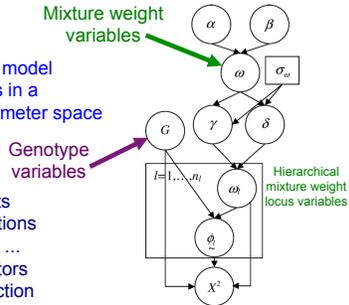
Survival of the Fittest

Markov chain Monte Carlo
Sample from the posterior probability distribution

Next state?
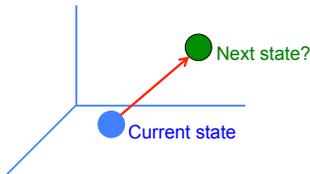
Current state

$$\text{Transition probability} = \frac{P\{\text{Next state}\}}{P\{\text{Current state}\}}$$

Laws of Variation

genotype
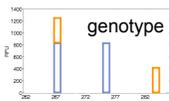
Hierarchy of successive pattern transformations

data
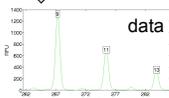
Variance parameters

Hierarchical
(e.g., customized for DNA template or locus)

Differential degradation
Mixture weight
Relative amplification
PCR stutter
PCR peak height
Background noise

Difficulties of the Theory

procedures & rules vs data-driven science

comfort in certainty vs tackling uncertainty

probability and likelihood ratios
can use all the data to
quantify uncertainty

What is the aim of Forensic Science?
comfort vs truth

Miscellaneous Objections

• too complex?
• black box?
• source code?
• insufficient validation?

Validation Studies

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

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):1430-47.

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):103-14.

Perlin MW, Belrose JL, Duceman BW. New York State TrueAllele® Casework validation study. *Journal of Forensic Sciences*. 2013;58(6):1458-66.

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.

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;*in press*.

Sensitive

The extent to which interpretation
identifies the correct person

True DNA mixture inclusions

101 reported genotype matches
82 with DNA statistic over a million

TrueAllele Sensitivity

log(LR) match distribution

11.05 *(5.42)*
113 billion

TrueAllele
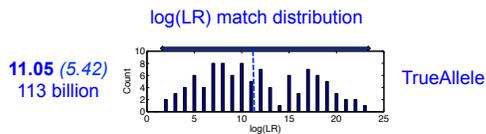
Specific

The extent to which interpretation does
not misidentify the wrong person

True exclusions, without false inclusions

101 matching genotypes x 10,000 random references
x 3 ethnic populations,
for over 1,000,000 nonmatching comparisons

TrueAllele Specificity

log(LR) mismatch distribution

− 19.47   0
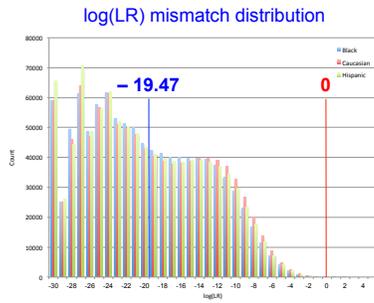
Reproducible

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

MCMC computing has sampling variation

duplicate computer runs
on 101 matching genotypes

measure log(LR) variation

TrueAllele Reproducibility

Concordance in two independent computer runs

standard deviation
(within-group)
0.305
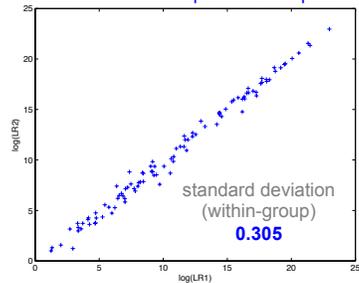
Manual Inclusion Method

Over threshold, peaks become binary allele events

All-or-none allele peaks,
disregard quantitative data

Analytical
threshold

Allele pairs
7, 7
7, 10
7, 12
7, 14
10, 10
10, 12
10, 14
12, 12
12, 14
14, 14

CPI Information

6.83 *(2.22)*
6.68 million

CPI
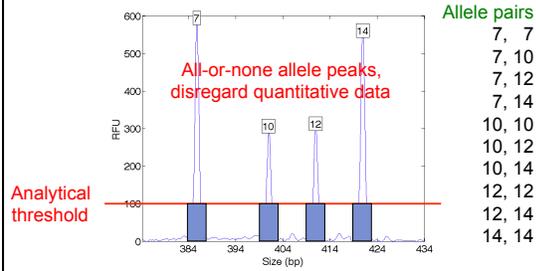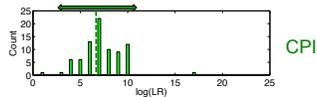
Combined probability of inclusion

Simplify data, easy procedure,
apply simple formula
$PI = (p_1 + p_2 + ... + p_k)^2$

Modified Inclusion Method

Higher threshold for human review

Apply two thresholds,
doubly disregard the data

Stochastic
threshold

in 2010

Analytical
threshold

in 2000
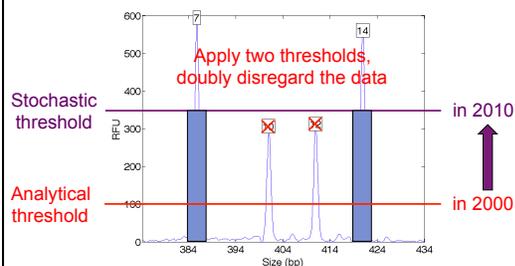
Modified CPI Information
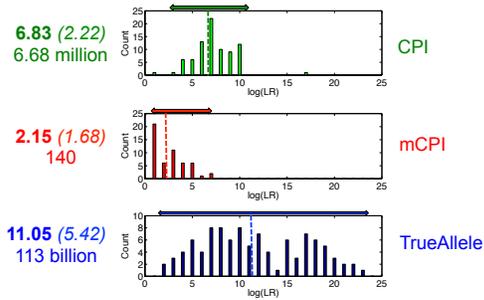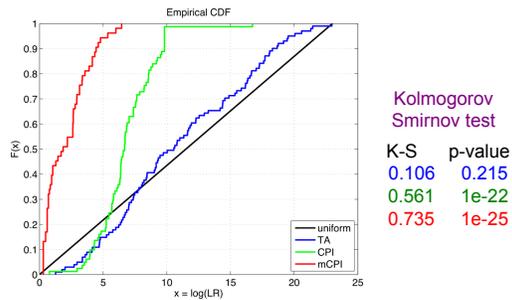
6.83 *(2.22)*
6.68 million

CPI

2.15 *(1.68)*
140

mCPI

Method Comparison

6.83 *(2.22)*
6.68 million

CPI

2.15 *(1.68)*
140

mCPI

11.05 *(5.42)*
113 billion

TrueAllele

Method Accuracy

Empirical CDF

Kolmogorov
Smirnov test

K-S	p-value
0.106	0.215
0.561	1e-22
0.735	1e-25

Slide 1

Invariant Behavior

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

Slide 2
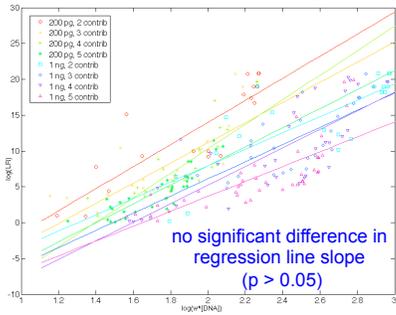
Sufficient Contributors

small negative slope values
statistically different from zero
(p < 0.01)

Slide 3
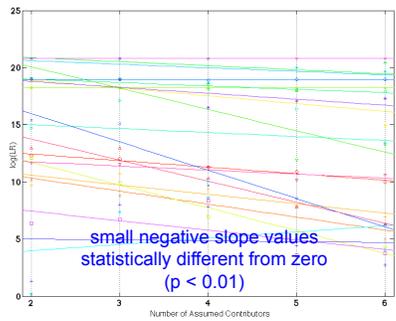
Admissibility Hearings

- California
- Pennsylvania
- Virginia
- United Kingdom
- Australia

Appellate precedent in Pennsylvania

Westchester, NY: Daughter Rape

Quantitative peak heights at locus D16S539
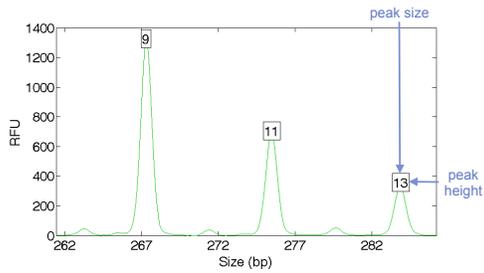
How TrueAllele Thinks

Consider every possible genotype solution
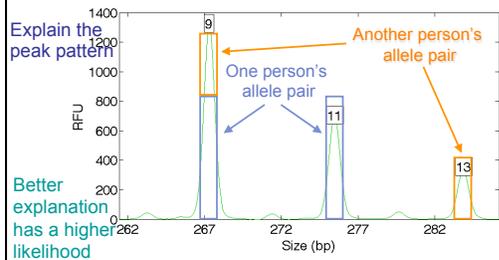
Evidence Genotype

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

DNA Match Information

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

14.5x

Match Information at 15 Loci
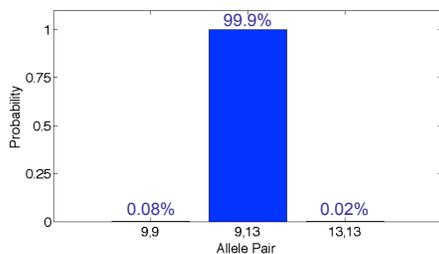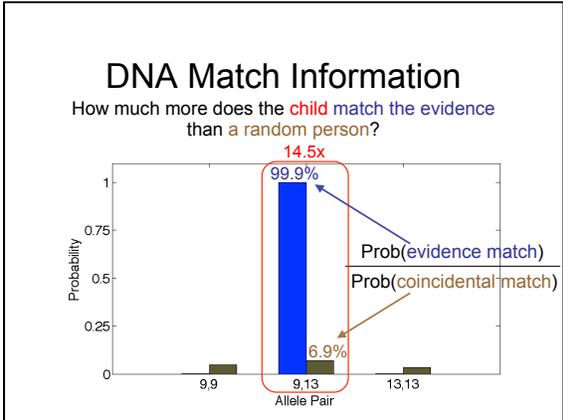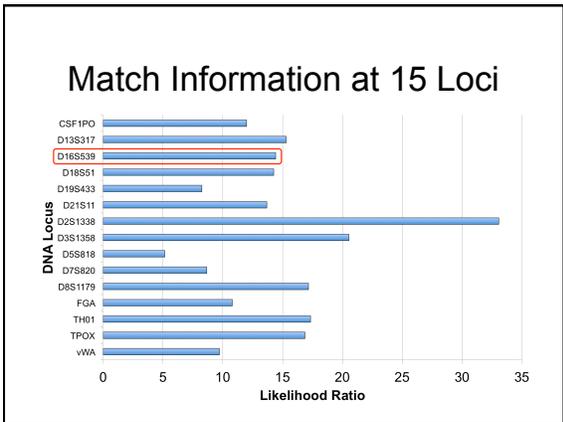
Likelihood Ratio Results

A match between the **blanket** and the **child** is
68 quadrillion times more probable than coincidence.

A match between the **blanket** and the **father** is
33 quadrillion times more probable than coincidence.

Father pleaded guilty to rape and
was sentenced to seven years in prison.
Young daughters spared further torment.

Other Cases

Mixtures of family members
- child rape
- homicide
- 3 person mixture
- 5 person mixture

Over 200 case reports
- Many states & countries

On-line in crime labs
- California
- Virginia
- Middle East

Investigative DNA Database

Infer genotypes, and then match with LR

Evidence   Reference

Database

Match
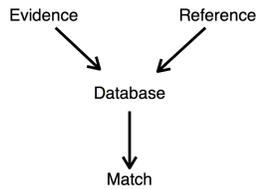
World Trade Center disaster

DNA Mixture Crisis: MIX05

National Institute of Standards and Technology
Two Contributor Mixture Data, Known Victim

Some Differences in Reporting Statistics

LablD	Kits Used	Caucasians	African Americans	Hispanics
			Case1	
90	ProPlus/Cofiler	1.18E+15	2.13E+14	3.09E+15
34	ProPlus/Cofiler	2.40E+11	7.66E+09	9.80E+10
33	ProPlus/Cofiler	2.94E+08	1.12E+08	1.74E+09
6	ProPlus/Cofiler	40,000,000	3,500,000	260,000,000
9	ProPlus/Cofiler	1.14E+07	1.97E+07	1.54E+08
79	ProPlus/Cofiler	930,000	17,900	1,350,000
16	ProPlus/Cofiler	434,600	31,710	399,100

213 trillion (14)

31 thousand (4)

Remember that these labs are interpreting the same MIX05 electropherograms

DNA Mixture Crisis: MIX13

MIX13 Case 5 Outcomes with Suspect C
(whose genotypes were not present in the mixture)

# Labs	Report Conclusions	Reasons given
6	Exclude Suspect C	detailed genotype checks (ID+); TrueAllele negative LR (ID+); assumed major/minor and suspects did not fit (ID+); 3 labs noted Penta E missing allele 15 (PP16HS)
3	Inconclusive with C only (A & B included)	All these labs used PP16HS
21	Inconclusive for A, B, and C	
70	Include & provide CPI statistics	All over the road…

Range of CPI stats for Caucasian population:
FBI allele frequencies: **1 in 9** to **1 in 344,000**

DNA Mixture Crisis: USA

375 cases/year x 4 years = 1,500 cases
320 M in US / 8 M in VA = 40 factor
1,500 cases x 40 factor = 60,000 inconclusive

1,000 cases/year x 4 years = 4,000 cases
320 M in US / 8 M in NY = 40 factor
4,000 cases x 40 factor = 160,000 inconclusive

+ under reporting of DNA match statistics

DNA evidence data in 100,000 cases
Collected, analyzed & paid for – but unused

The Rule of Science & Law

unworkable rules vs validated science

"How wonderful it is that nobody need wait a single moment before starting to improve the world."
ANNE FRANK

Why do we practice Forensic Science?

Learn More About TrueAllele

http://www.cybgen.com/information

- Courses
- Newsletters
- Newsroom
- Presentations
- Publications

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

Cybergenetics

You Tube

perlin@cybgen.com