DNA Identification Information: Science and Statistics

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

June, 2010

Cybergenetics

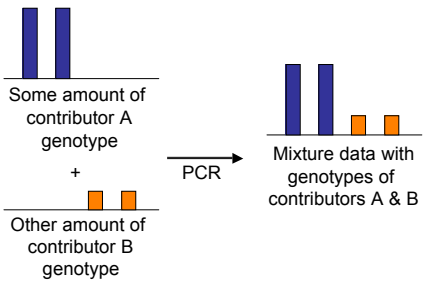
Objective DNA Identification

Given uncertain STR data d

(1) Infer questioned genotype Q

(2) Match with suspect genotype S
relative to random genotype R
to form likelihood ratio LR

DNA Mixture Data



Data Model: Likelihood

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

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

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

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

Prior Probability

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

$$\mathbf{w} \sim Dir(\mathbf{1})$$

$$m_l \sim N_+\left(5000, 5000^2\right)$$

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

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

$$\psi^{-2} \sim Gam(1/2, 1/200)$$

Genotype Inference

$$\Pr\{Q = x \mid d_{l,1}, d_{l,2}, ..., d_{l,z}, ...\} \propto \Pr\{Q = x\} \cdot \prod_{i=1}^{I} \Pr\{d_{l,j} \mid Q = x, ...\}$$

$$\Pr\{W = w \mid d_1, d_2 \ldots, d_J, \ldots\} \propto \Pr\{W = w\} \cdot \prod_{j=1}^{J} \Pr\{d_j \mid W = w, \ldots\}$$

$$\Pr\left\{\sigma^2 = s^2 \middle| d_1, d_2, ..., d_j, ...\right\} \propto \Pr\left\{\sigma^2 = s^2\right\} \cdot \prod_{j=1}^{J} \Pr\left\{d_j \middle| \sigma^2 = s^2, ...\right\}$$

$$\Pr\left\{\tau^2 = t^2 \middle| d_1, d_2, ..., d_j, ...\right\} \propto \Pr\left\{\tau^2 = t^2\right\} \cdot \prod_{j=1}^{J} \Pr\left\{d_j \middle| \tau^2 = t^2, ...\right\}$$

Information Gain (LR)

identification hypothesis:
the suspect contributed to the evidence

information gain (likelihood ratio) = $\frac{\text{Odds(hypothesis | data)}}{\text{Odds(hypothesis)}}$

after ↑ data
before

Additive information units: log(LR)
Order of magnitude, powers of ten

(apply Bayes theorem)

$$\frac{O(H|d_Q,d_R,d_S)}{O(H)} = \frac{\Pr\{H|d_Q,d_R,d_S\}/\Pr\{\overline{H}|d_Q,d_R,d_S\}}{\Pr\{H\}/\Pr\{\overline{H}\}}$$
$$= \frac{\Pr\{H|d_Q,d_R,d_S\}/\Pr\{H\}}{\Pr\{\overline{H}|d_Q,d_R,d_S\}/\Pr\{\overline{H}\}}$$
$$= \frac{\Pr\{d_Q|H,d_R,d_S\}/\Pr\{d_Q\}}{\Pr\{d_Q|\overline{H},d_R,d_S\}/\Pr\{d_Q\}}$$
$$= \frac{\Pr\{d_Q|H,d_R,d_S\}}{\Pr\{d_Q|\overline{H},d_R,d_S\}}$$

Likelihood Ratio (LR)

$$LR = \frac{\Pr\{d_Q|H,d_R,d_S\}}{\Pr\{d_Q|\overline{H},d_R,d_S\}}$$

(extend conversation)

$$\frac{\Pr\{d_Q|H,d_R,d_S\}}{\Pr\{d_Q|\overline{H},d_R,d_S\}} = \frac{\sum_{x\in G}\Pr\{d_Q|H,d_R,d_S,Q=x\}\cdot\Pr\{Q=x|H,d_R,d_S\}}{\sum_{x\in G}\Pr\{d_Q|\overline{H},d_R,d_S,Q=x\}\cdot\Pr\{Q=x|\overline{H},d_R,d_S\}}$$
$$= \frac{\sum_{x\in G}\Pr\{d_Q|d_R,d_S,Q=x\}\cdot\Pr\{S=x|d_R,d_S\}}{\sum_{x\in G}\Pr\{d_Q|d_R,d_S,Q=x\}\cdot\Pr\{R=x|d_R,d_S\}}$$
$$= \frac{\sum_{x\in G}\Pr\{d_Q|Q=x\}\cdot\Pr\{S=x|d_S\}}{\sum_{x\in G}\Pr\{d_Q|Q=x\}\cdot\Pr\{R=x|d_R\}}$$
$$= \frac{\sum_{x\in G}\lambda_Q(x)\cdot s(x)}{\sum_{x\in G}\lambda_Q(x)\cdot r(x)}$$

Genotype-Weighted Likelihood

$$LR = \frac{\sum_{x\in G}\lambda_Q(x)\cdot s(x)}{\sum_{x\in G}\lambda_Q(x)\cdot r(x)}$$

(apply Bayes theorem)

$$LR = \frac{\sum_{x\in G}\lambda_Q(x)\cdot s(x)}{\sum_{x\in G}\lambda_Q(x)\cdot r(x)}$$
$$= \sum_{x\in G}\frac{\lambda_Q(x)}{\sum_{y\in G}\lambda_Q(y)\cdot\pi_Q(y)}\cdot s(x)$$
$$= \sum_{x\in G}\frac{q(x)}{\pi_Q(x)}\cdot s(x)$$
$$= \sum_{x\in G}\frac{q(x)\cdot s(x)}{r(x)}$$

Genotype Probability Gain

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

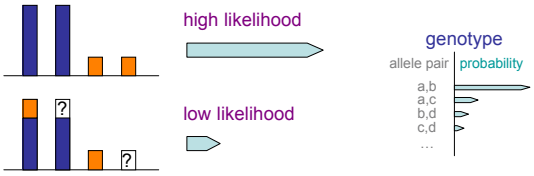
Summary

- Other modeling variables (stutter, …)
- LR can also consider coancestry
- Infer genotypes, then match them
- Objective: inference never sees suspect
- Goal: preserve identification information
  - + Bayesian modeling for genotype
  - + Bayesian information gain (LR)

Quantitative Mixture Interpretation

Step 1: infer genotype

- consider every possible allele pair
- compare pattern with DNA data
- Rule: *better fit's more likely it*



Information Gain (LR)

Step 2: match genotypes

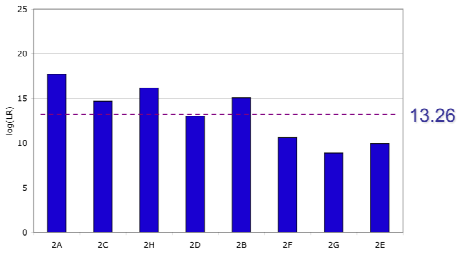
At the suspect's genotype allele pair,
what is the locus *information gain*?

$$\text{information gain (likelihood ratio)} = \frac{\text{Prob(allele pair} \mid \text{data)}}{\text{Prob(allele pair)}}$$

after ↑ data
before (population)

Computer objectivity:
(Step 1) infer evidence genotype from data
(Step 2) compare genotype with suspect

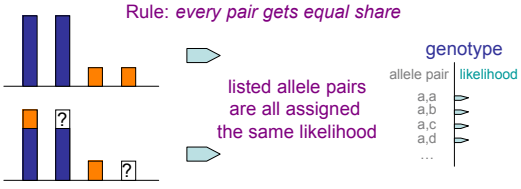
Efficacy (2 unknown)



Qualitative Manual Review

Step 1: infer genotype

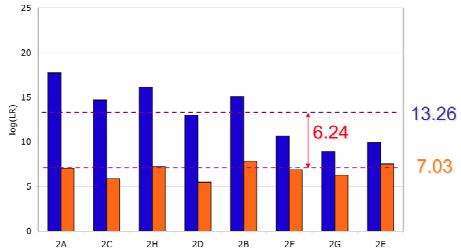
Rule: *every pair gets equal share*



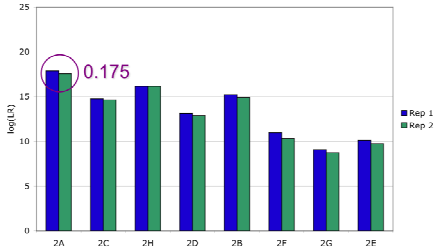
Step 2: match genotype

lower probability means lower information gain (LR)

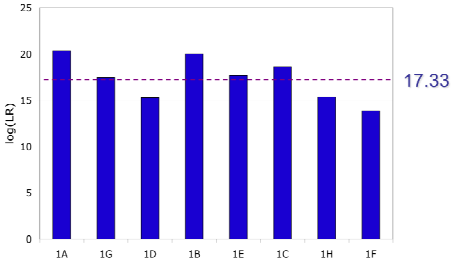
Improvement



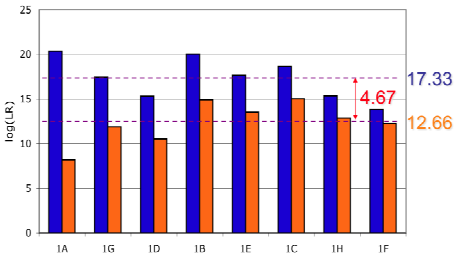
Reproducibility



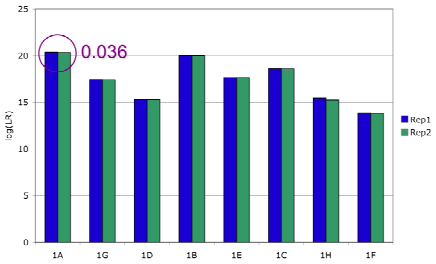
Efficacy (1 unknown)



Improvement



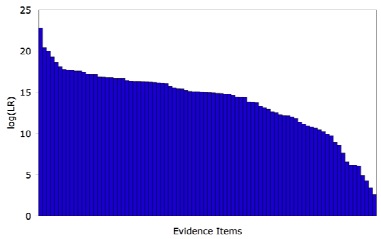
Reproducibility



Comparison

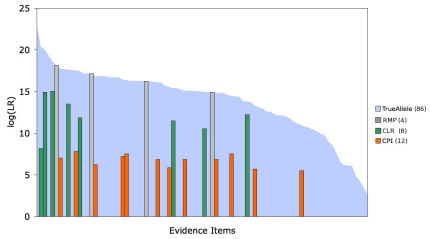
interpretation method	two unknown (without victim)	one unknown (with victim)
quantitative computer	13.26 (0.175) (ten trillion)	17.33 (0.036) (hundred quadrillion)
qualitative human	7.03 (ten million)	12.66 (fifty trillion)
improvement	6.24 (one million)	4.67 (fifty thousand)

TrueAllele Information: 86 Match Stats (100%)



Preserves all the identification information

Human Review Information: 24 Match Stats (28%)



Preserves 20% of the identification information

Summary

- information gain (LR) is a universal DNA metric
- *efficacy*: computer extracts useful information
- *improvement*: computer mixture interpretation is more informative than human review
  with victim 50,000x - without victim 1,000,000x
- *reproducibility*: tenths of a log(LR) unit
- *objectivity*: "parallel unmasking", infer then match
- *productivity*: lab gives statistic for 1 of 3 items
- *utility*: science, investigation and evidence

Commonwealth vs. Foley

Apr 2006: Blairsville Dentist John Yelenic murdered

Nov 2007: Trooper Kevin Foley charged with crime

Feb 2008: Defense questions 13,000 DNA match score

DNA Evidence

- DNA from under victim's fingernails (Q83)
- two contributors to DNA mixture
- 93.3% victim & 6.7% unknown
- 1,000 pg DNA in 25 ul
- STR analysis with ProfilerPlus®, Cofiler®
- know victim contributor genotype (K53)
- TrueAllele® computer interpretation
  (using genotype addition method)
  infer unknown contributor genotype
- only after having inferred unknown,
  compare with suspect genotype (K2)

Three DNA Match Statistics

Score	Method
13 thousand	inclusion
23 million	subtraction
189 billion	addition

- Why are there different match results?
- How do mixture interpretation methods differ?
- What should we present in court?

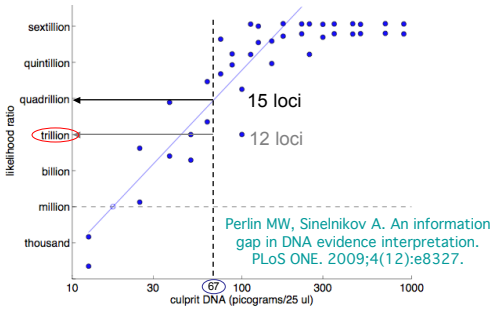
Different Interpretation Methods

Data Used	inclusion	subtraction	addition
victim profile	NO	YES	YES
quantitative data	NO	NO	YES

Frye: General Acceptance in the Relevant Community

- Quantitative STR Peak Information
- Genotype Probability Distributions
- Computer Interpretation of STR Data
- Statistical Modeling and Computation
- Likelihood Ratio Literature
- Mixture Interpretation Admissibility
- Computer Systems for Quantitative DNA Mixture Deconvolution
- TrueAllele Casework Publications

Expected Result



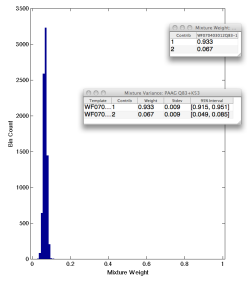
Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS ONE. 2009;4(12):e8327.

Expert Testimony

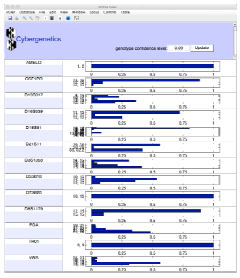
Dr. Perlin explained to the jury why these apparently different results were expected by DNA science. "The less informative methods ignored some of the data," said Dr. Perlin, "while the TrueAllele computation considered all of the available DNA data."

"A scientist may look at the same slide using the naked eye, a magnifying glass, or a microscope," analogized Dr. Perlin. "A computer that considers all the data is a more powerful DNA microscope."

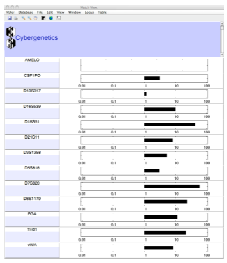
Mixture Weight



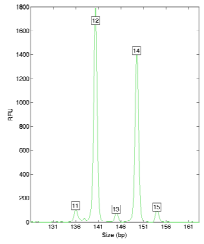
Inferred Genotype



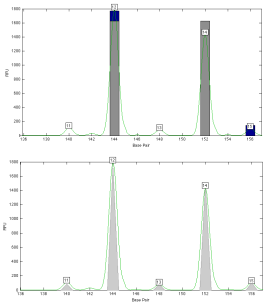
log(LR) Match Information



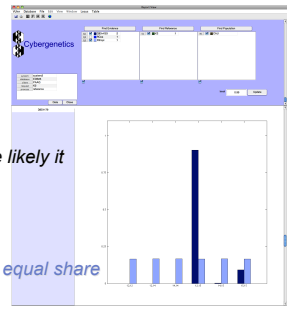
Locus D8S1179 Data



Explain D8S1179 Genotype



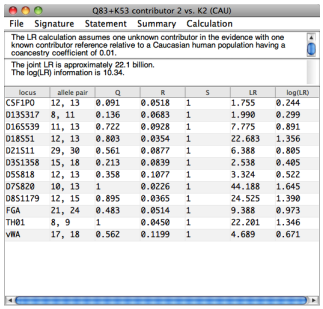
Likelihood Comparison



better fit's more likely it

every pair gets equal share

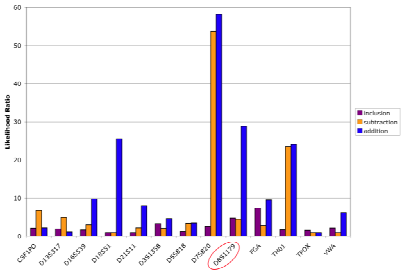
Generate Report



Locus information gain is genotype probability ratio: LR = after/before

Joint information is the sum of the locus information

More Data In, More Information Out



13 thousand (4)
23 million (7)
189 billion (11)

Case Observations

- objective review never saw suspect
- easy to testify about in court
- understandable to judge and jury
- have precedent: admitted, testified
- preserve match information in data
- rapid response to attorney
- multiple match scores presented
  all information to the triers of fact –
  nothing was withheld from the jury
  this should be standard practice


Indiana Gazette.com
Indiana, Pennsylvania
IN PRINT DAILY. ONLINE ALWAYS.

HOME SUBSCRIBERS MARKETPLACE NEWS OBITUARIES SPORTS BUSINESS MULTIMEDIA O
FYI JOURNALS

ARCHIVES > NEWS

One Verdict

Print | E-mail | Comment (No comments posted.) | Rate | Text Size

Jury convicts trooper of dentist slaying

Published: Thursday, March 19, 2009 12:46 AM EDT

An Indiana County Court jury this evening convicted state trooper Kevin Foley of first degree murder in the April 13, 2006, slashing death of Blairsville dentist John Yelenic.

"John Yelenic provided the most eloquent and poignant evidence in this case," said the prosecutor, senior deputy attorney general Anthony Krastek. "He managed to reach out and scratch his assailant," capturing the murderer's DNA under his fingernails.

The DNA Investigator Newsletter. Same Data, More Information - Murder, Match and DNA, Cybergenetics, 2009.
www.cybgen.com/information/newsletters/CybgenNews1.pdf