The Science of DNA Search

The Changing Face of DNA:
The Science, Law and Ethics of Familial Searches

Duquesne Univeristy
May, 2011
Pittsburgh, PA

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

Cybergenetics

DNA Investigation

25 years ago: the "blooding"

case-to-case

UK rape/homicide of two 15 year old girls

1983 victim
Lynda Mann

1986 victim
Dawn Ashworth

DNA evidence

DNA evidence

DNA MATCH: serial criminal

DNA Investigation

25 years ago: the "blooding"

case-to-suspect

Confession of 17 year old boy

crime scenes

1986 suspect
Richard Buckland

DNA evidence

DNA reference

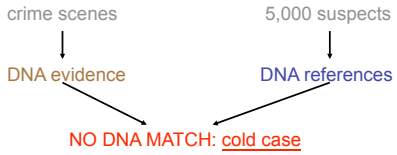
NO DNA MATCH: exonerated

DNA Investigation

25 years ago: the "blooding"

case-to-suspects

Tested 5,000 men in Leicestershire

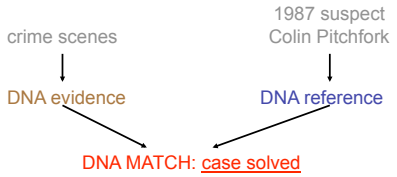


DNA Investigation

25 years ago: the "blooding"

case-to-suspects

Tested one more Leicestershire man



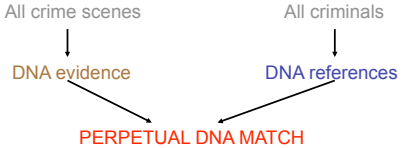
DNA Database

15 years ago: UK NDNAD

cases ⟷ suspects

- violent crime
- property crime

- all arrestees
- computerization



Vision: Preventing Crime

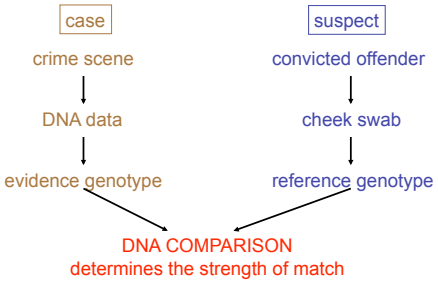
Continual crime scene surveillance by DNA

2004. Ray Wickenheiser: Complete DNA databasing (sexual assault, property crime, convicted offender) would prevent 100,000 stranger rapes

The UK did exactly this, with over 50% DNA database hit rate, stopping criminal careers early on at property crime

Protect the public through DNA databases by stopping criminals before they strike again

How DNA Matching Works



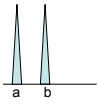
Simple DNA Evidence

evidence

reality

Simple DNA Evidence

evidence —Lab→ STR data

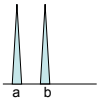


at 1 of 13 CODIS loci

reality observation

Simple DNA Evidence

evidence —Lab→ STR data —Infer→ genotype



a, b

at 1 of 13 CODIS loci

reality observation model

Objective Match Statistic

evidence genotype a, b ⟷ reference genotype a, b

Objective Match Statistic

evidence genotype a, b ⟷ reference genotype a, b

likelihood ratio $LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{100\%}{5\%} = 20$

Objective Match Statistic

evidence genotype a, b ⟷ reference genotype a, b

likelihood ratio $LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{100\%}{5\%} = 20$

information measure $\log(LR) = 1.30$ (since $10^{1.30} = 20$, in powers of 10)

Objective Match Statistic

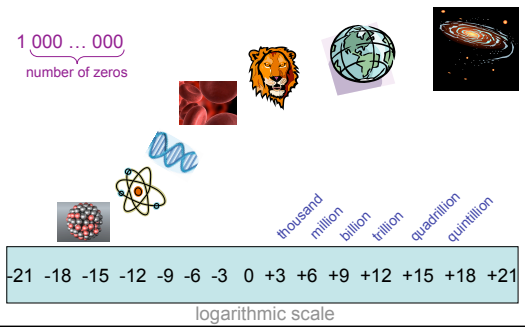
evidence genotype a, b ⟷ reference genotype a, b

likelihood ratio $LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{100\%}{5\%} = 20$

information measure $\log(LR) = 1.30$ (since $10^{1.30} = 20$, in powers of 10)

total information $13 \times \log(LR) = 13 \times 1.30 = 16.9$ 17 zeros is 100 quadrillion

Powers of Ten



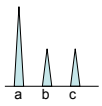
Complex DNA Evidence

evidence

reality

Complex DNA Evidence

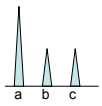
evidence → Lab → STR data



reality observation

Complex DNA Evidence

evidence → (Lab) → STR data → (Infer) → genotype



a, a @ 30%
a, b @ 50%
a, c @ 20%

reality | observation | model

Objective Match Statistic

evidence genotype		reference genotype
a, b @ 50%	⟷	a, b

Objective Match Statistic

evidence genotype		reference genotype
a, b @ 50%	⟷	a, b

likelihood ratio

$$LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{50\%}{5\%} = 10$$

Objective Match Statistic

evidence genotype a, b @ 50% ⟷ reference genotype a, b

likelihood ratio: $LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{50\%}{5\%} = 10$

information measure: log(LR) = 1 (since $10^1 = 10$, in powers of 10)

Objective Match Statistic

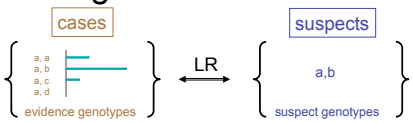
evidence genotype a, b @ 50% ⟷ reference genotype a, b

likelihood ratio: $LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{50\%}{5\%} = 10$

information measure: log(LR) = 1 (since $10^1 = 10$, in powers of 10)

total information: 13 x log(LR) = 13 x 1 = 13 — 13 zeros is 10 trillion

Investigative DNA Database



- *genotype probability* representation
- fully *preserves* DNA identification *information*
- enables *LR calculation* with every match
  - connect crimes to criminals
  - disaster victim identification (WTC)
  - find missing people
  - automatic familial search
  - combat terrorism through DNA

Down A Different Path

15 years ago: United States CODIS
a second generation DNA database

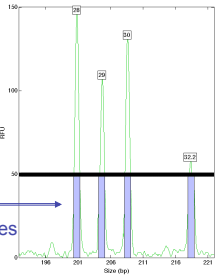
New feature: represents mixture evidence

The FBI used *simplified* mixture interpretation

+ easier for crime labs to do and testify about

- loses considerable identification information

Qualitative Thresholds



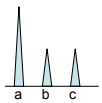
list of included alleles

Over threshold, peaks are treated as identical allele events

Under threshold, alleles do not exist

Simplified Mixture Interpretation

evidence → (Lab) → STR data → (Infer) → genotype



a, a
a, b
a, c
b, b
b, c
c, c

All possibilities given equal likelihood, regardless of how well or poorly they explain the data

reality observation model

Reduced Match Statistic

diffuses probability

evidence genotype a, b @ 15%

reference genotype a, b

likelihood ratio

$$LR = \frac{Pr(match)}{Pr(coincidence)} = \frac{15\%}{5\%} = 3$$

information measure

log(LR) = 0.48 (since $10^{0.48} = 3$, in powers of 10)

total information

13 x log(LR) = 13 x 0.48 = 6.2

6 zeros is one million

The DNA Investigator™ Newsletter, 2009
Same Data, More Information – Murder, Match and DNA

Fingernail: 7% Mixture

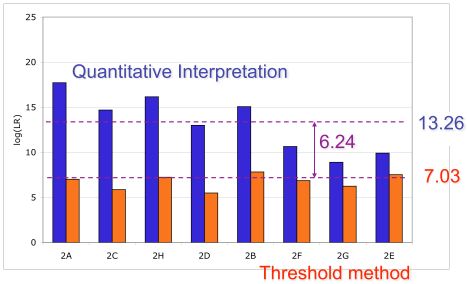
Commonwealth v. Foley

ScoreMethod
13 thousandthreshold
23 millionuse victim
189 billionquantitative

- probability modeling preserves information
- peak threshold discards information

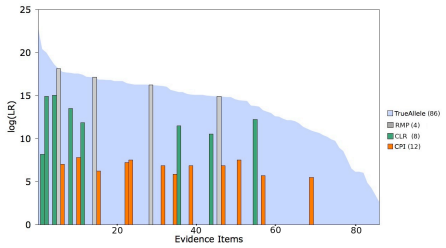
MW Perlin, MM Legler, CE Spencer, JL Smith, WP Allan, JL Belrose, BW Duceman. Validating TrueAllele DNA mixture interpretation. Journal of Forensic Sciences, 2011.

Preserve vs. Discard



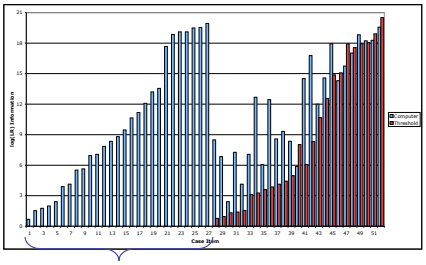
Perlin MW, Duceman BW. Profiles in productivity: greater yield at lower cost with computer DNA interpretation. Twentieth International Symposium on the Forensic Sciences of the Australian and New Zealand Forensic Science Society, Sydney, Australia. 2010.

Preserve vs. Discard



- quantitative interpretation preserves information - every time
- peak threshold discards information - 70% of the time

Preserve vs. Discard



Computer: informative
Threshold: inconclusive

The DNA Investigator™ Newsletter, 2011
DNA Intelligence and Forensic Failure – What you don't know can kill you

CODIS: Second Generation

less informative evidence genotype	simplified genotype representation
allele pairs	allele list
a, a	a, b, c
a, b	
a, c	
b, b	
b, c	
c, c	

- identification information lost
- sacrifices probability accuracy
- probative evidence discarded
- match *trillions* reduced to *millions*

Truth or Consequences

Exchanging best science for practical procedures

"High stringency" search. With high thresholds, lose sensitivity - can't find criminals

"Low stringency" search. Overly broad, so lose specificity - finds hundreds of CODIS matches

For example, with 100,000 convicted offenders, finding 100 database hits means a 1/1000 hit rate

Many dead end leads, and wasted police work

Everyone is a Suspect

Isn't it "fair" to put everyone on the DNA database? Why store only criminal genotypes?

With high CODIS specificity (low false positives), OK

But the CODIS mixture representation reduces *trillion*-fold evidence specificity down to *million*-fold

Each evidence item would falsely implicate *hundreds* of innocent people, soon making all Americans suspects

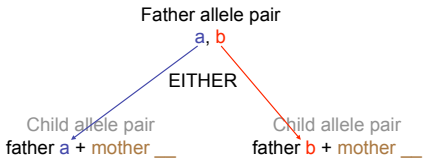
Weak science can have poor consequences

Familial Search: A Good Consequence

Serendipity from low stringency searches

CODIS "*low stringency*": genotypes share at least one allele

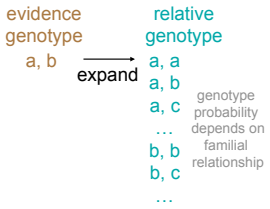
Family Genetics



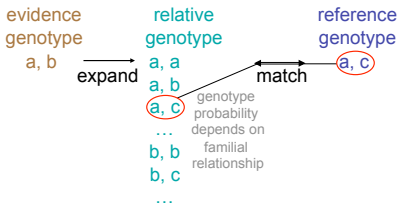
Parent and child share at least one allele
(siblings and other kin also share alleles, but fewer)

So "*low stringency*" match can be used for familial search

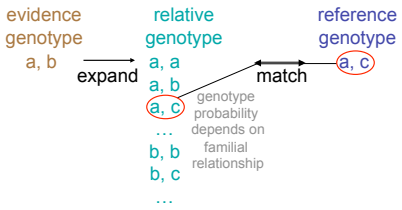
How the Science Works



How the Science Works



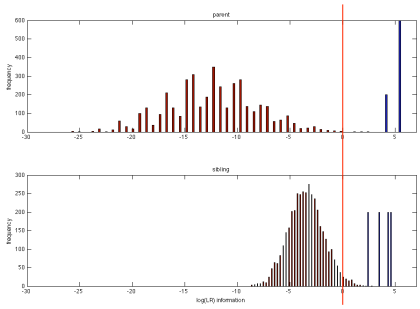
How the Science Works



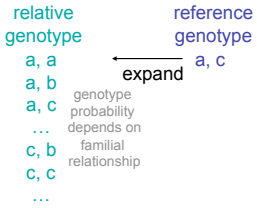
likelihood ratio $LR = \frac{Pr(match)}{Pr(coincidence)}$

information measure $\log(LR)$

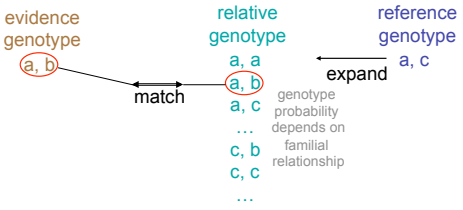
Results: Evidence Mode



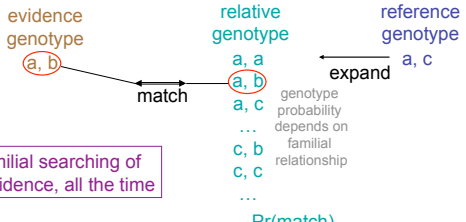
Continuous Mode



Continuous Mode



Continuous Mode



Familial searching of all evidence, all the time

likelihood ratio $LR = \frac{Pr(match)}{Pr(coincidence)}$

information measure $\log(LR)$

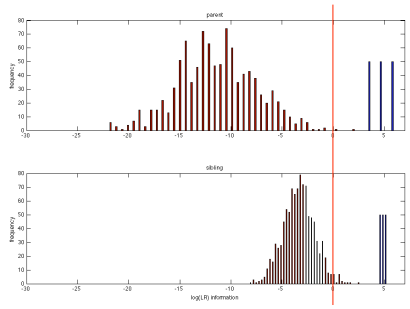
Mixtures in Familial Search

- Most DNA evidence items are *mixtures*
- CODIS representation has weak information, greatly increasing *familial search* false hits

But probabilistic genotypes overcome these obstacles
- informative computer *interpretation*
  (SWGDAM 2010 guidelines, paragraph 3.2.2)
- information preserving DNA *database*
  (ANSI/NIST-ITL standard, fields 18.020 & 18.021)

So the full DNA evidence information is preserved

Results: Minor Contributor



Familial Search Cost/Benefit

most agree that familial search is worthwhile in catching criminals and preventing further crime

- mainly nonmixture evidence with CODIS
- only dozens of hits made in ten years
- cost is $20,000 per search (Y-STR confirm)
- yield is 10% hits per search

Hit cost is $200,000 ($20,000 / 10%)

Preserving Information Cost/Benefit

better use of the DNA data *we already have*
can catch more criminals and prevent more crime

- all evidence types (including mixtures)
- no confirmation testing required (non-familial)
- computer system costs less than one familial hit
- better information doubles evidence yield
- can produce thousands of hits every year
- expands familial search capability to mixtures

More information
More hits
Lower cost

Science & Technology

- past: DNA evidence and DNA databases
  1st generation, single allele pair
- present: familial search and other innovative DNA paths to public safety
  2nd generation, lists of alleles
- future: more informative DNA methods using computers to preserve evidence
  3rd generation, probabilistic genotypes

DNA identification is an information science

Learning More

The science of DNA search

www.cybgen.com/information

- Newsletters
  gentle introduction to ideas
- Courses
  for lawyers and scientists
- Presentations
  handouts, movies, transcripts
- Publications
  abstracts, manuscripts