A paradigm shift in DNA interpretation

John Buckleton

Specialist Science Solutions
Manaaki Tangata Taiao Hoki
protecting people and their environment through science
I sincerely acknowledge conversations with

- Jo-Anne Bright  XX
- Duncan Taylor  XY
- Steven Myers  XY
- Michael Coble  XY
- Ian Evett  XY
Variability in interpretation

- DNA science has been criticised for producing different interpretations of the same profile
- Part of the diversity is subjectivity but
- Part is systemic
- Different laboratories use different methods for interpretation
- Yet we nearly all use strongly similar typing technology
- Why?
"The method I invented is the exactly correct mix of complexity and information usage"
I claim a special right to critique

CPI
Cumulative probability of inclusion

RMP
Random match probability
Todd Bille, Jo-Anne Bright

LR binary
selection of genotypes
Peter Gill, Jonathan Whittaker, Tim Clayton

Drop model
Peter Gill, Jonathan Whittaker, David Balding

Continuous models
Duncan Taylor, Jo-Anne Bright
CPI, RMP, Drop model, Continuous model

Drop model
Continuous model

Continuous model
If we were starting new how would we choose.

**CPI**
Cumulative probability of inclusion

**RMP**
Random match probability

**LR binary**
Selection of genotypes

**Drop model**

**Continuous models**

Simple answer: the one that gets it right?
Getting it right

• Use ground truth known samples
• Eg mix person A and person B
• What is the right answer if we test the hypotheses
• $H_1: A + B$
• $H_2: A + \text{unknown}$
• Should be between 1 and $1/\Pr(B)$
• But that is a pretty wide range
• If the PCR is unusual $<1$ is even the “right” answer
Getting it right?

How old are they now?

How old were their parents when they died?

What is the probability that this person will live to 75+?

Do they have any health risks?

What does their doctor say?

The person dies at 76.

Was I “right”?

The more relevant information

Used properly

The better

22%
We cannot decide from this one event

Was 22% right? Was it wrong?

How can an answer be neither right nor wrong?
Is the answer right?

It is the best answer that I can produce?

John Buckleton ESR
But is it right?

I cannot tell if it is right or wrong but it makes the best use of the available information

John Buckleton ESR
Q. Can you answer the question?

A. Yes. Can we make that the last time you yell at me.
Q. Well if you'd answered the question then I wouldn't need to repeat it.

A. OK
We cannot decide from this one event

But we might be able to score methods from a lot of events with known outcomes, Known ground truth. There are scoring methods
If we were starting new how would we choose

<table>
<thead>
<tr>
<th>CPI</th>
<th>RMP</th>
<th>LR binary</th>
<th>Drop model</th>
<th>Continuous models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulative probability of inclusion</td>
<td>Random match probability</td>
<td>selection of genotypes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Which one makes best use of the available information?
CPI

- CPI (cumulative probability of inclusion)
- The probability that a person would be included (not excluded) usually on straight allele presence
Person could be 10,10 10,11 10,12 10,13 11,11 11,12 11,13 12,12 12,13 13,13

\[ CPI = f_{10}^2 + f_{11}^2 + f_{12}^2 + f_{13}^2 + 2f_{10}f_{11} + 2f_{10}f_{12} + 2f_{10}f_{13} + 2f_{11}f_{12} + 2f_{11}f_{13} + 2f_{12}f_{13} \]
\[ = \left( f_{10} + f_{11} + f_{12} + f_{13} \right)^2 \]

use the reasonable inference of

Does not assume a number of contributors, this is seen as a good thing.

Is it good for Mr 10,10?
If we assume two people then one of them could be

10,10  10,11
10,12  10,13
11,11  11,12
11,13  12,12
12,13  13,13

\[
RMP = f_{10}^2 + f_{11}^2 + f_{12}^2 + f_{13}^2 + 2f_{10}f_{11} + 2f_{10}f_{12} + 2f_{10}f_{13} \\
+ 2f_{11}f_{12} + 2f_{11}f_{13} + 2f_{12}f_{13}
\]

Does use the reasonable inference of a number of contributors.
Does use the reasonable inference of a number of contributors.

\( LR \): We now need two hypotheses

Some people think this is bad

\( H_1 \): POI = 10,11 + U

\( H_2 \): 2U

If we assume two people they must be

10,11 and 12,13 or
10,12 and 11,13 or
10,13 and 12,13 or
11,12 and 10,13 or
11,13 and 10,12 or
12,13 and 10,11

\[
LR = \frac{Pr(E \mid H_1)}{Pr(E \mid H_2)}
\]

\[
LR = \frac{2f_{12}f_{13}}{24f_{10}f_{11}f_{12}f_{13}} = \frac{1}{12f_{10}f_{11}}
\]
If we assume two people they must be

10,11 and 12,13 or
10,12 and 11,13 or
10,13 and 12,13 or
11,12 and 10,13 or
11,13 and 10,12 or
12,13 and 10,11

\[ LR = \frac{2f_{12}f_{13}}{24f_{10}f_{11}f_{12}f_{13}} = \frac{1}{12f_{10}f_{11}} \]
Add information

V = 12,13 high vaginal swab, no consensual partners

$H_1$: POI = 10,11 + V

$H_2$: U + V

$$\text{CPI} = (f_{10} + f_{11} + f_{12} + f_{13})^2$$

$$\text{RMP} = 2f_{10}f_{11} + 2f_{10}f_{12} + 2f_{10}f_{13} + 2f_{11}f_{12} + 2f_{11}f_{13} + 2f_{12}f_{13}$$

$$\text{LR} = \frac{1}{12f_{10}f_{11}}$$
Principle

- Adding relevant information improves the power of our statistics
- On average
- Higher $LR$ when $H_1$ true, lower when $H_2$ true
- Benefits the innocent, bad for the guilty
Let’s ask the automobile association?

Is the mountain pass open?
We could ring the gas station on the other side and see if people are coming over.

Nah I don’t like information it might bias me?
Nah might bias.
Best if we just drive blind?
\[ \text{CPI} = (f_{10} + f_{11} + f_{12} + f_{13})^2 \]

\[ \text{RMP} = 2f_{10}f_{11} + 2f_{12}f_{13} \]

What about Mr 11,12?

\[ \text{LR} = \frac{2f_{12}f_{13}}{2f_{10}f_{11}2f_{12}f_{13}} = \frac{1}{2f_{10}f_{11}} \]
LR binary
selection of genotypes

RMP
Random match probability

CPI
Cumulative probability of inclusion

Drop model

Where does this one go?

Information
The drop model

Take the profile
Throw away much of the information
Then start the interpretation
So why did we even develop it?
This graphic is only true for a good profile with no drop out possible.

**LR binary**
Selection of genotypes

**CPI**
Cumulative probability of inclusion

**RMP**
Random match probability

**Drop model**

Where does this one go?
Non-concordance
All non-concordances are problematic but some more so than others. POI = 13,15

Strong evidence

DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods Forensic Science International: Genetics, Volume 6, Issue 6, December 2012, Pages 679-688 P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, B.S. Weir
So why did we even develop it?
Drop model

We can probably extend the drop model a lot further by incorporating aspects of height information.
Identifiler 28 cycles
NGM SEElect 29 cycles
SGMPlus 34 cycles

![](image)

Log(Hb) vs APH plot.
Experiments with a composite approach might catch a lot of the information content

Luigi Armogida
USACIL

PHr works well
Drop model wastes info

PHr unreliable
Drop allows interpretation
Degradation slopes differ hence drop-out probabilities are profile and locus specific.
Locus specific amplification example
Locus effects are not steady over time, they may be batch or even profile specific

Modelling one drop-out probability per profile misses these effects
Modelling a degradation slope gets some but not all
D8S1179

V = 13,17  POI=14,15

\[ CPI = \left( f_{12} + f_{13} + f_{14} + f_{15} + f_{16} + f_{17} \right)^2 \]

\[ RMP = 2 f_{14} f_{15} \]

\[ LR_B = \frac{1}{2 f_{14} f_{15}} \]

\[ LR_C = \frac{1}{2 f_{14} f_{15}} \]
D7S820  V = 9,9  POI=11,11

\[
CPI = \left( f_8 + f_9 + f_{11} \right)^2 = 0.09
\]

\[
RMP = f_{11}^2 + 2f_{11}f_Q = 0.19
\]

\[
LR_B = \frac{1}{f_{11}^2 + 2f_{11}f_Q} = 5.26
\]

\[
LR_C = \frac{0.427 \times 1}{0.218 \times 2f_8f_{11} + 0.191 \times 2f_9f_{11} + 0.427 \times f_{11}^2 + 0.165 \times 2f_{11}f_Q}
\]

\[= 8.20\]
Have we gone too far? Will anyone follow?
A paradigm shift in DNA interpretation

- LR binary
- RMP
- CPI
- Drop model
- Continuous