Probabilistic Genotyping
Webcast (Part 1)
May 28, 2014

A Basic Overview of Probabilistic Genotyping

Michael D. Coble
National Institute of Standards and Technology

NIST FORENSIC SCIENCES

NIST and NIJ Disclaimer

<u>Past and Present Funding</u>: Interagency Agreement between the National Institute of Justice and NIST Office of Law Enforcement Standards

Points of view are mine and do not necessarily represent the official position or policies of the US Department of Justice or the National Institute of Standards and Technology.

Certain commercial equipment, instruments, software and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Two Parts to Mixture Interpretation

- Determination of alleles present in the evidence and **deconvolution of mixture components** where possible
  - Many times through comparison to victim and suspect profiles
- **Providing some kind of statistical answer** regarding the weight of the evidence
  - There are multiple approaches and philosophies

Statistical Approaches with Mixtures

See Ladd *et al.* (2001) *Croat Med J.* 42:244-246

“Exclusionary” Approach

Random Man Not Excluded (RMNE)

Combined Prob. of Inclusion (CPI)

Combined Prob. of Exclusion (CPE)

“Inferred Genotype” Approach

Random Match Probability (RMP)

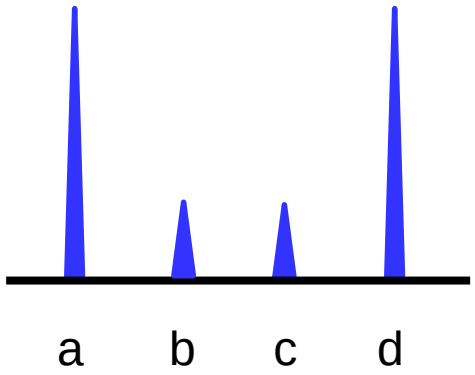
(mRMP)

Likelihood Ratio (LR)

Exclusionary Approach

Statistical Approaches with Mixtures

- **Random Man Not Excluded (CPE/CPI)** - The probability that a random person (unrelated individual) would be included/excluded as a contributor to the observed DNA mixture.

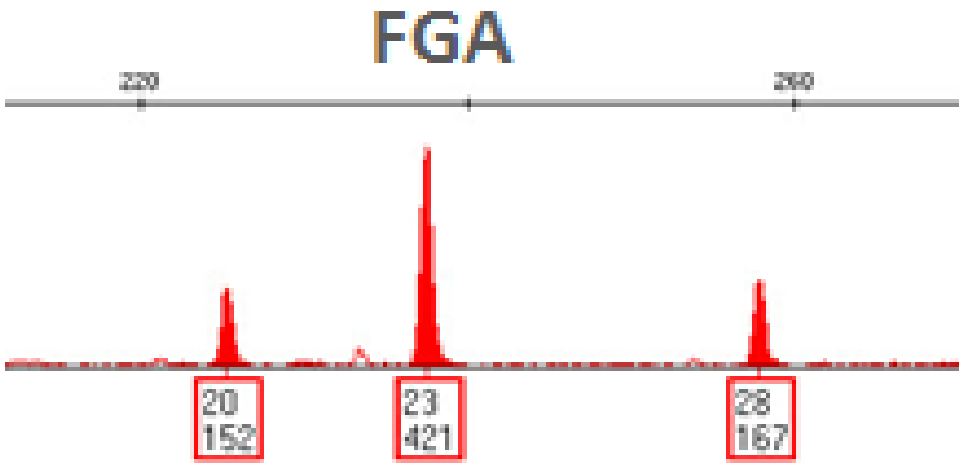


$$CPI = (f(a) + f(b) + f(c) + f(d))^2$$

$$CPI = PI_{M1} \text{ X } PI_{M2} \cdots$$

$$CPE = 1 - CPI$$

RMNE example with FGA

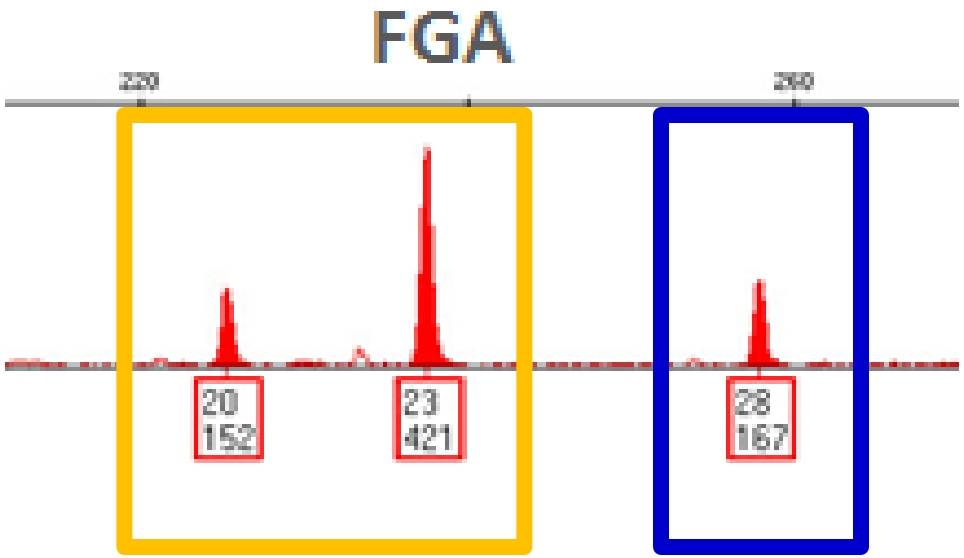


Possible Combinations

20, 28 and 23, 23
20, 23 and 23, 28

Assume ST = 150 RFU

RMNE example with FGA



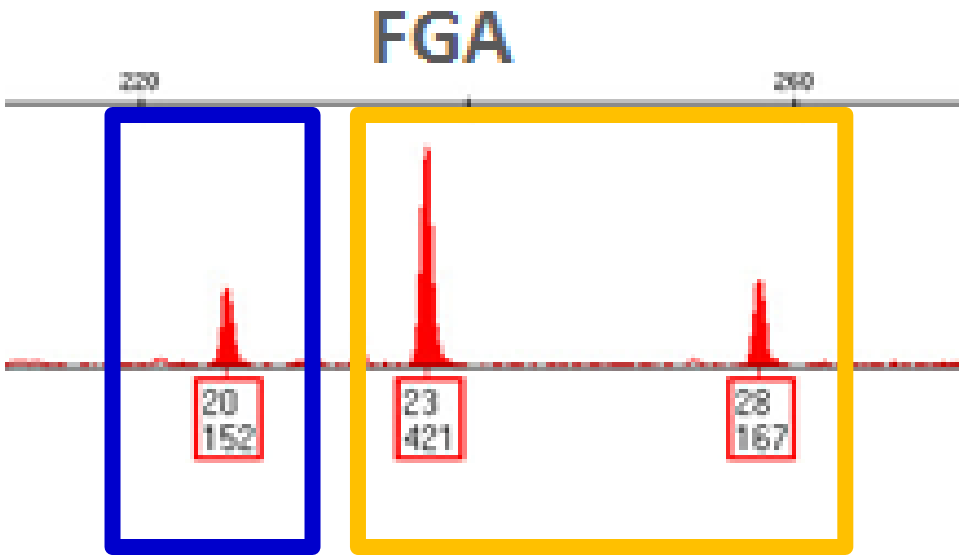
Possible Combinations

20, 28 and 23, 23
20, 23 and 23, 28

20, 23 and 28, 28

Assume ST = 150 RFU

RMNE example with FGA



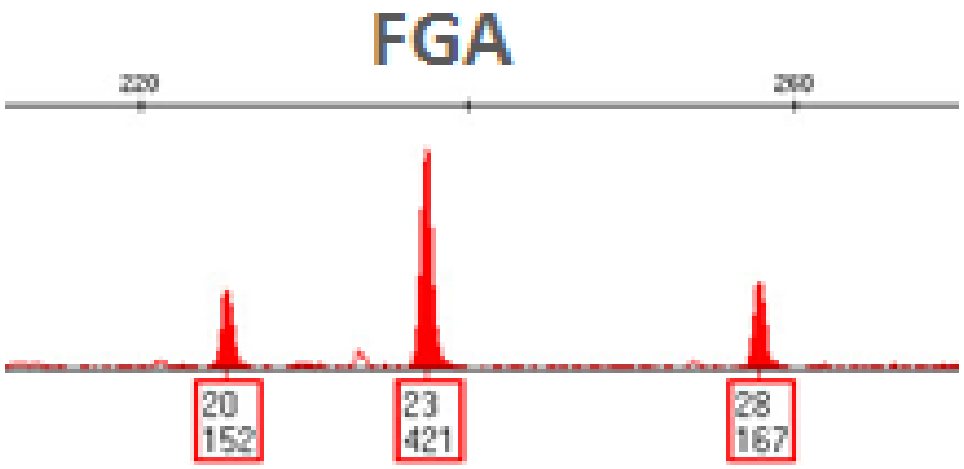
Possible Combinations

20, 28 and 23, 23
20, 23 and 23, 28

20, 23 and 28, 28
20, 20 and 23, 28

Assume ST = 150 RFU

RMNE example with FGA



Possible Combinations

20, 28 and 23, 23
20, 23 and 23, 28

20, 23 and 28, 28
20, 20 and 23, 28

$PI = (p + q + r)^2$

$PI = (f_{20} + f_{23} + f_{28})^2$

$PI = (0.145 + 0.158 + 0.013)^2$

$PI = (0.316)^2$

$PI = 0.099$

$PE = 1 - CPI = 0.901$

Assume ST = 150 RFU

“Advantages and Disadvantages” RMNE

RMNE (CPE/CPI)

Advantages

- Does not require an assumption of the number of contributors to a mixture
- Easier to explain in court
- Deconvolution is not necessary

Disadvantages

- Weaker use of the available information (robs the evidence of its true probative power because this approach does not consider the suspect’s genotype).
- Alleles below ST cannot be used for statistical purpose
- There is a potential to include a non-contributor

Summarized from John Buckleton, *Forensic DNA Evidence Interpretation*, p. 223
Buckleton and Curran (2008) *FSI-G* 343-348.

Notes from Charles Brenner's AAFS 2011 talk

The Mythical "Exclusion" Method for Analyzing DNA Mixtures – Does it Make Any Sense at All?

1. The claim that it requires **no assumption about number of contributors** is mostly wrong.
2. The supposed **ease of understanding** by judge or jury is really an illusion.
3. **Ease of use** is claimed to be an advantage particularly for complicated mixture profiles, those with many peaks of varying heights. The truth is the exact opposite. **The exclusion method is completely invalid for complicated mixtures.**
4. The exclusion method is only **conservative** for guilty suspects.

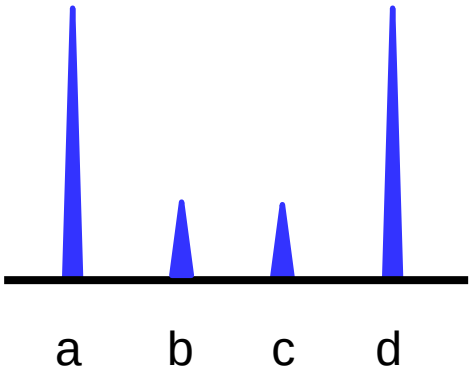
Conclusion: "Certainly no one has laid out an explicit and rigorous chain of reasoning from first principles to support the exclusion method. It is at best guesswork."

Brenner, C.H. (2011). The mythical "exclusion" method for analyzing DNA mixtures – does it make any sense at all? *Proceedings of the American Academy of Forensic Sciences*, Feb 2011, Volume 17, p. 79

modified Random Match Probability

Statistical Approaches with Mixtures

- **Random Match Probability (RMP)** – The major and minor components can be successfully separated into individual profiles. A random match probability is calculated on the evidence as if the component was from a single source sample.



$$RMP_{minor} = 2pq$$

$$= 2 \times f(b) \times f(c)$$

2013 JFS Article

JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, March 2013, Vol. 58, No. 2
doi: 10.1111/1556-4029.12067
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

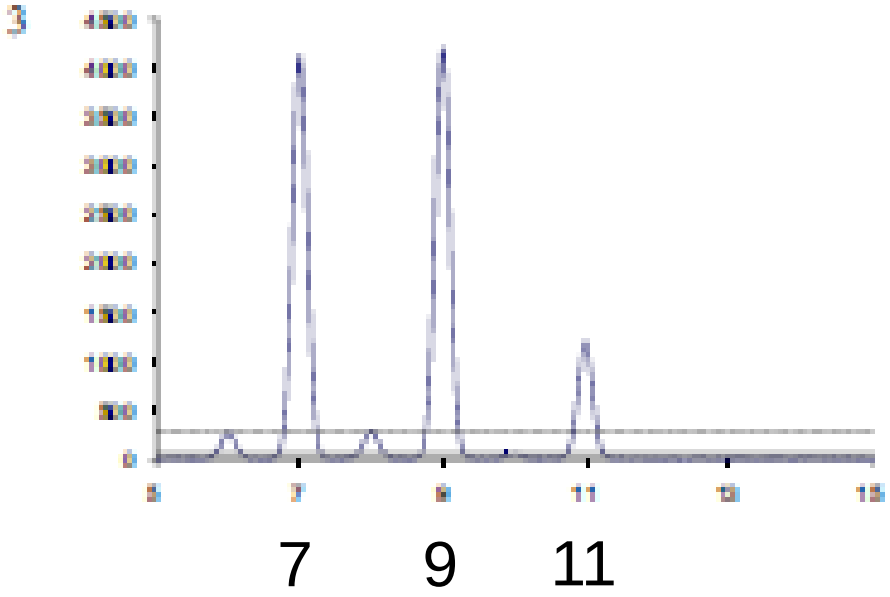
CRIMINALISTICS

Todd Bille,[1] *M.Sc.; Jo-Anne Bright,*[2] *M.Sc.; and John Buckleton,*[2] *Ph.D.*

Application of Random Match Probability Calculations to Mixed STR Profiles

When data is above ST

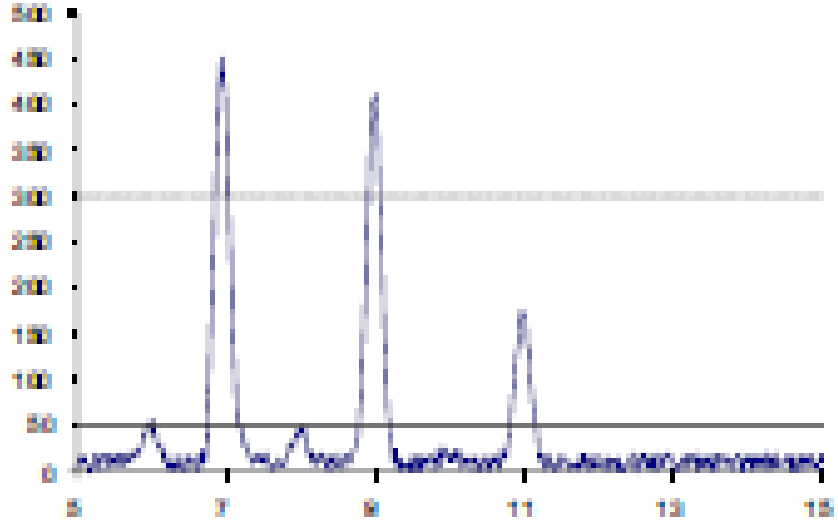
K = 7,9
S = 7,11
U = 7,11
9,11 or
11,11



$$\text{mRMP} = 2f_7 f_{11} + 2f_9 f_{11} + (f_{11})^2$$

$$\text{CPI} = (f_7 + f_9 + f_{11})^2$$

When data is below ST

#	Example EPG	K	S	U
7		7,9	7,11	7,11 or 9,11 or 11,11 or 11,Q

7 9 11

Q = any other allele

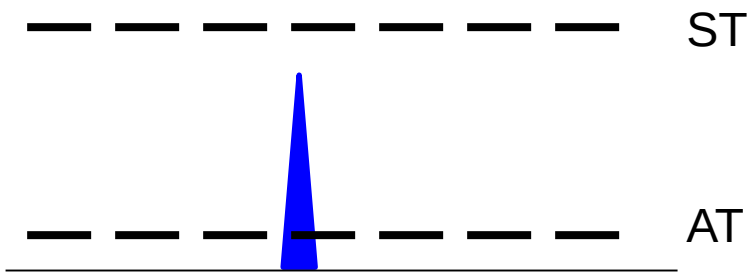
“2p” rule

CPI = n/a

mRMP = 2p

The “2p” Rule

- The “2p” rule can be used to statistically account for zygosity ambiguity – i.e. is this single peak below the stochastic threshold the result of a homozygous genotype or the result of a heterozygous genotype with allele drop-out of the sister allele?



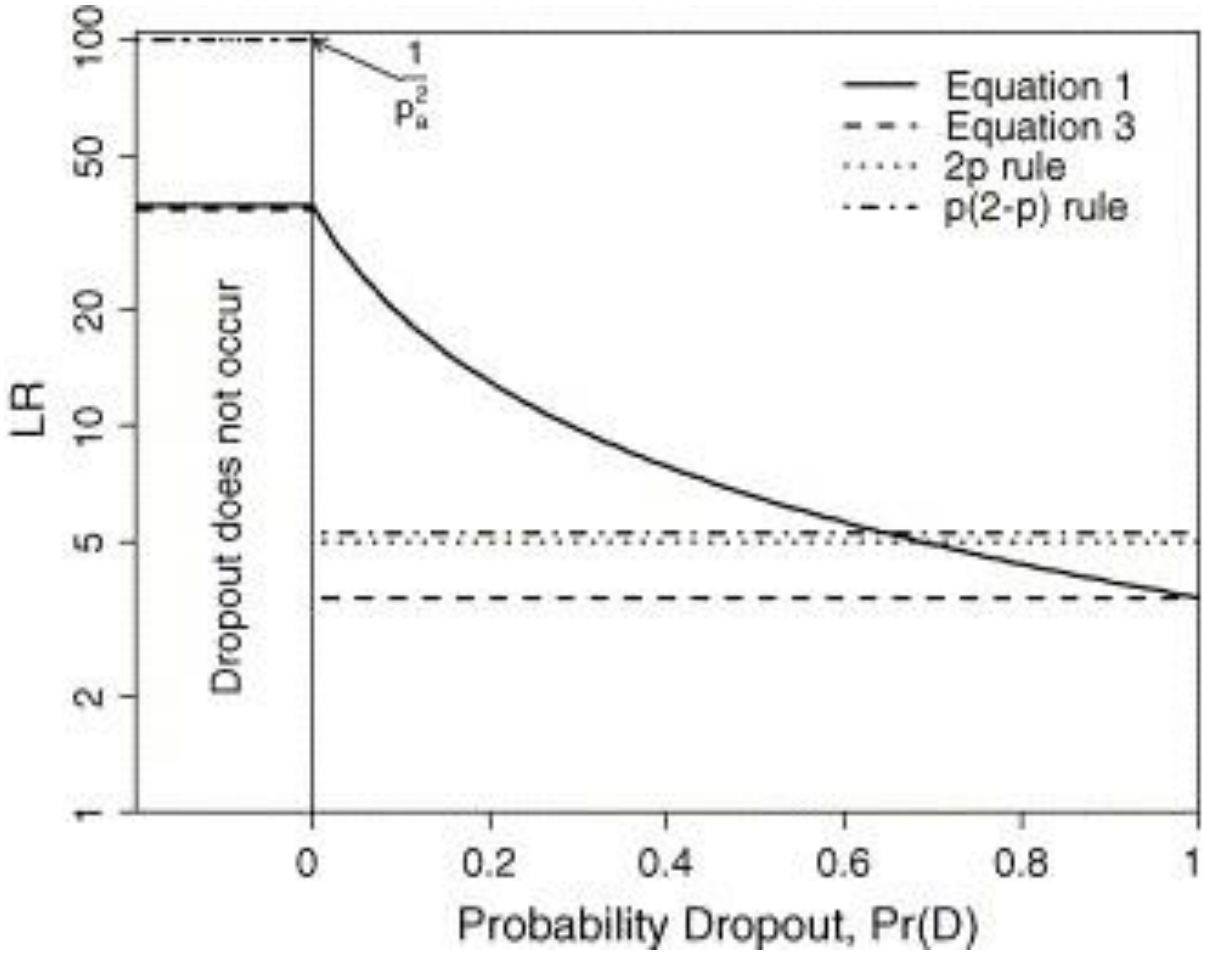
The “2p” Rule

- “This rule arose during the VNTR era. At that time many smaller alleles “ran off the end of the gel” and were not visualised.”

- Buckleton and Triggs (2006)

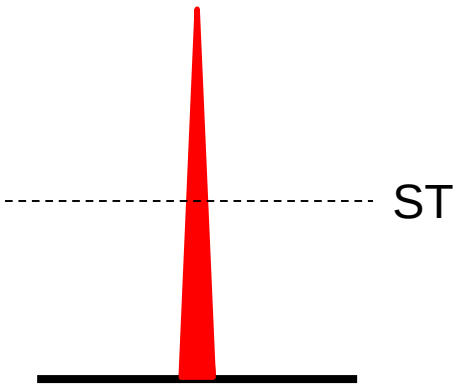
Is the 2*p* rule always conservative?”

The “2p” Rule



Stain = AA

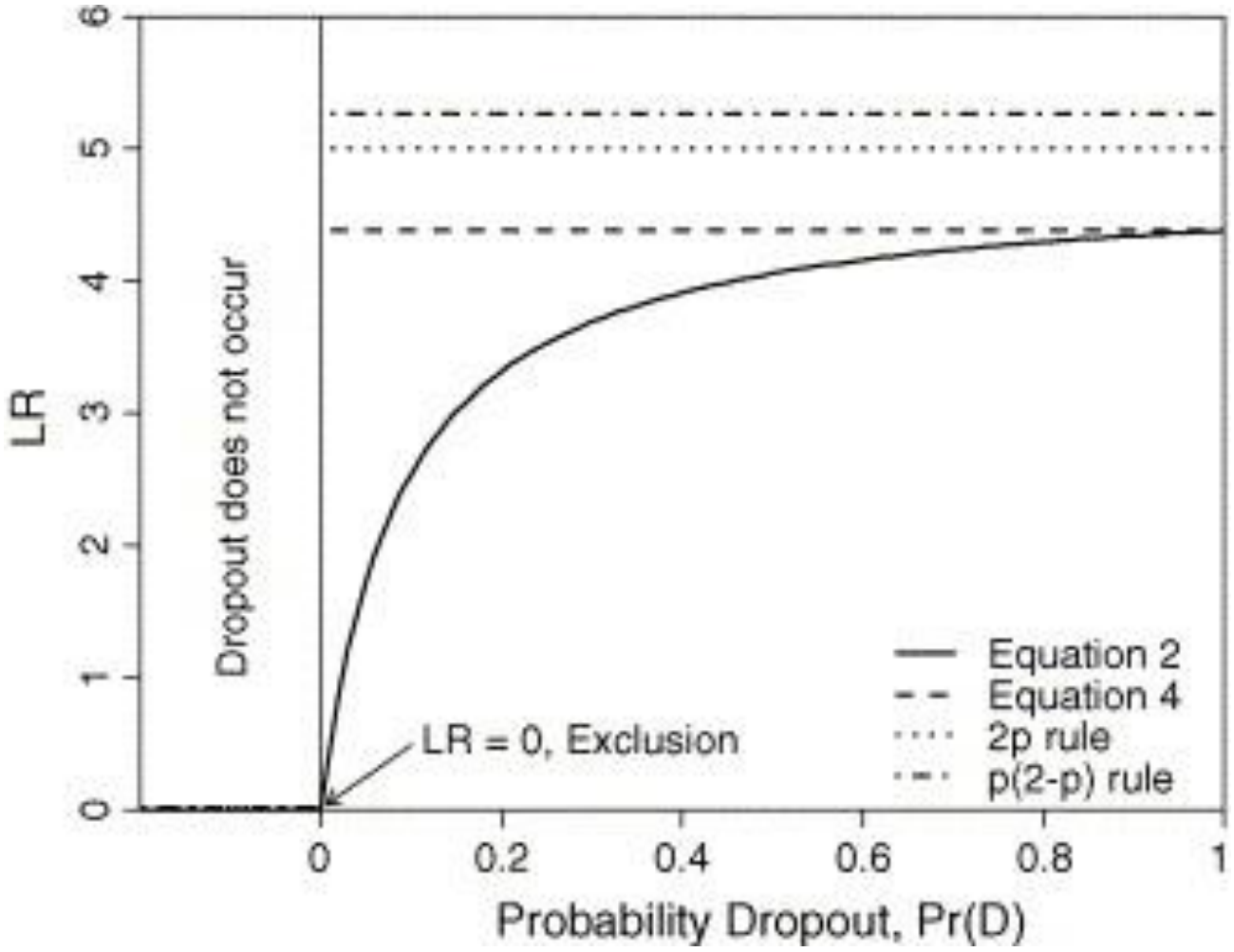
Suspect = AA



LR = 100

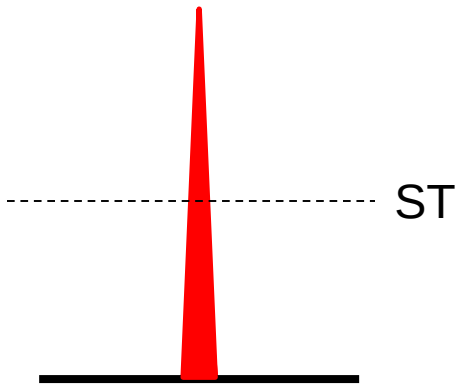
f(a) = 0.10 $1/p^2 = 100$ 1/2p = 5

The “2p” Rule



Stain = AA

Suspect = AB



Exclusion

f(a) = 0.10 1/2p = 5

Likelihood Ratio

Statistical Approaches with Mixtures

- **Likelihood Ratio** - Comparing the probability of observing the mixture data under two (or more) alternative hypotheses

Likelihood Ratios in Forensic DNA Work

- We evaluate the evidence (*E*) relative to alternative pairs of hypotheses

- Usually these hypotheses are formulated as follows:
  - The probability of the evidence if the crime stain originated with the suspect or Pr(*E*|*S*)
  - The probability of the evidence if the crime stain originated from an unknown, unrelated individual or Pr(*E*|*U*)

$$LR = \frac{\Pr(E \mid S)}{\Pr(E \mid U)}$$

The numerator

The denominator

Slide information from Peter Gill

Likelihood Ratio (LR)

- Provides ability to express and evaluate both the prosecution hypothesis, H_p (the suspect is the perpetrator) and the defense hypothesis, H_d (an unknown individual with a matching profile is the perpetrator)

$$LR = \frac{H_p}{H_d}$$

- **The numerator, H_p, is usually 1** – since in theory the prosecution would only prosecute the suspect if they are 100% certain he/she is the perpetrator

- The denominator, **H_d**, is typically the profile frequency in a particular population (based on individual allele frequencies and assuming HWE) – i.e., **the random match probability**

Slide information from Peter Gill

FSI GENETICS

Forensic Science International: Genetics 2 (2008) 343–348

A discussion of the merits of random man not excluded and likelihood ratios

John Buckleton [a,*], James Curran [b]

[a] *ESR, PB 92021, Auckland, New Zealand*

[b] *Department of Statistics, University of Auckland, PB 92019, Auckland, New Zealand*

Received 15 January 2008; received in revised form 29 April 2008; accepted 1 May 2008

We conclude that the two matters that appear to have real force are:
(1) LRs are more difficult to present in court and
(2) the RMNE statistic wastes information that should be utilised.

What kind of mixtures were being seen in the early days of STR testing?

- Torres et al. (2003) published the casework experience in a Spanish laboratory over a four-year time period (Jan 1997 to Dec 2000)
- 2412 samples typed
  - 955 samples from sexual assaults
  - 1408 samples from other offenses
  - 49 samples from human remains identifications
- 163/2412 samples (**6.7% showed a mixed profile**)
- Only 8 samples (**0.3% of total samples**) **were a >2 person mixture!**

Torres, Y., et al. (2003). DNA mixtures in forensic casework: a 4-year retrospective study. *Forensic Science International*, 134, 180-186.

From Torres et al. (2003)

Torres, Y., et al. (2003). DNA mixtures in forensic casework: a 4-year retrospective study. *Forensic Science International*, 134, 180-186.

- "In our own and other authors' experience (Clayton et al. 1998) **two-person mixtures account for the overwhelming majority of mixtures encountered during casework**, but occasionally mixtures of three or more persons are seen with more than four alleles at some loci. Eight of the 163 mixed samples (0.3% of the total typed samples) corresponded to such higher-order profiles."

Clayton, T.M., et al. (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiling. *Forensic Science International*, *91*, 55-70.

Gathered Case Summary Data

During 2007 and early 2008, **Ann Gross** (MN BCA) from the SWGDAM Mixture Interpretation Committee **coordinated the collection of case summary data** from **14 different forensic labs** who collectively reported on **>4500 samples**.

A preliminary summary of this information is divided by crime classifications: sexual assault, major crime (homicide), and high volume (burglary). **Over half of the samples examined were single source and ~75% of all reported mixtures were 2-person.**

This is why the SWGDAM 2010 Interpretation Guidelines focused on 2-person mixtures

Mixture Case Summaries (2007-2008)

Data Set from 14 Different Labs

http://www.cstl.nist.gov/strbase/pub_pres/Promega2008poster.pdf

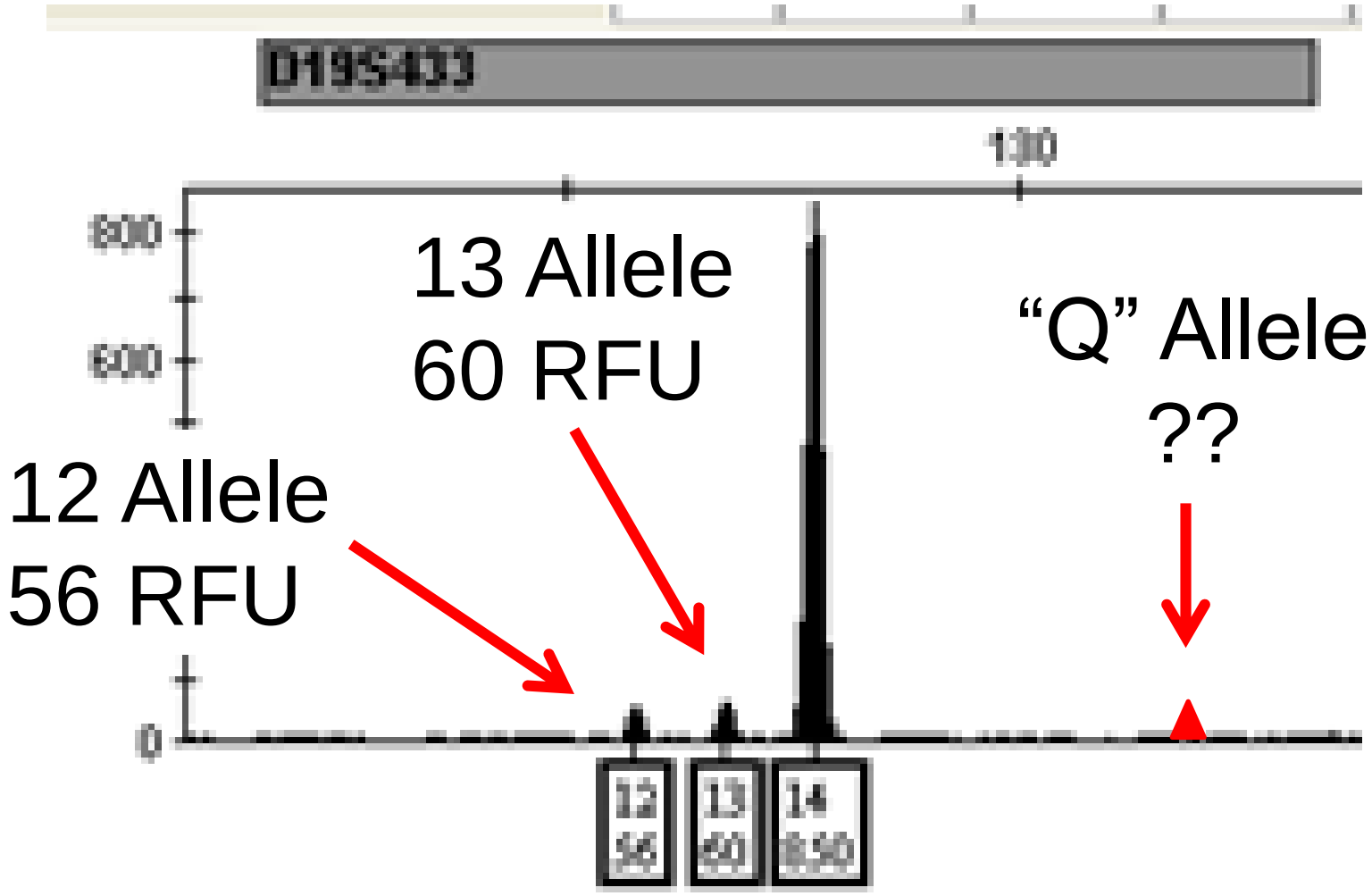
	minimum # of contributors					
Crime Class	**1**	**2**	**3**	**4**	**>4**	**N**
Sexual Assault	884	787	145	11	0	1827
Major Crime	1261	519	182	32	0	1994
High Volume	344	220	140	11	5	720
Total	2489	1526	467	54	5	**4541**
	54.8%	**33.6%**	10.3%	1.2%	0.1%	**mixtures**

Single source

2-person mixtures

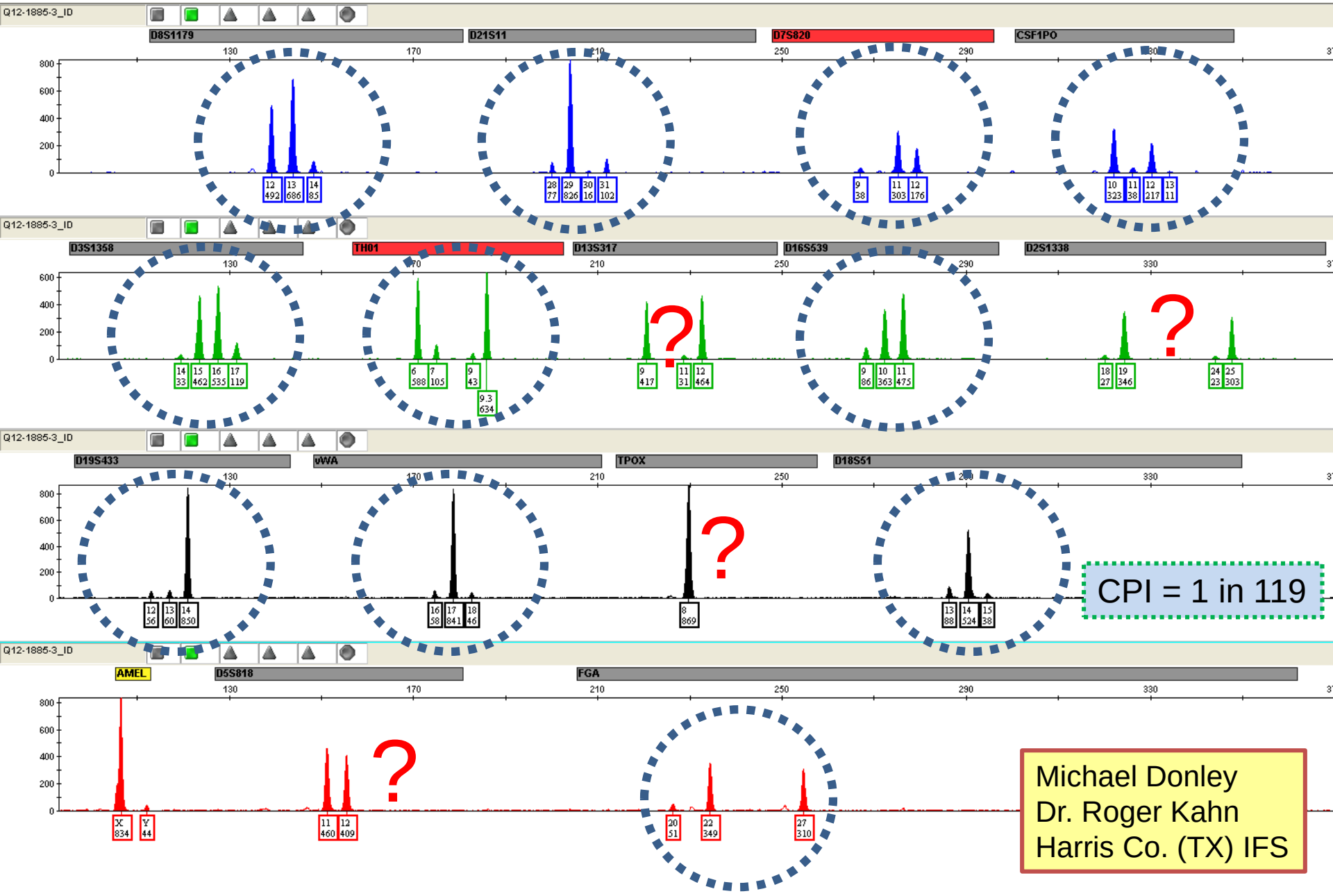
11.6%
>2-person mixtures

Challenging Mixtures



How to handle low level data

- Continue to use RMNE (CPI, CPE)

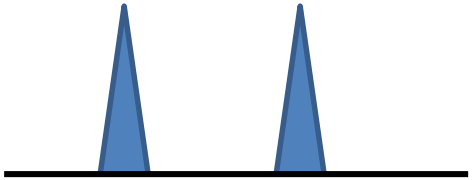


Michael Donley
Dr. Roger Kahn
Harris Co. (TX) IFS

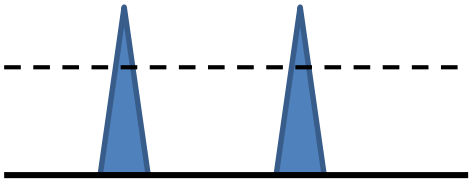
What should we do with data below our Stochastic Threshold?

- Continue to use RMNE (CPI, CPE) (not optimal)
- Use the Binary LR with 2p (not optimal)

Suspect

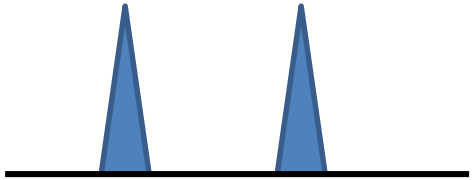


Evidence

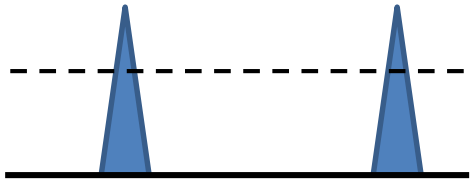


$$LR = \frac{1}{2pq}$$

Suspect

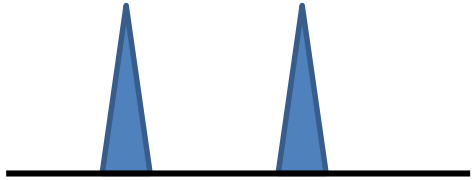


Evidence

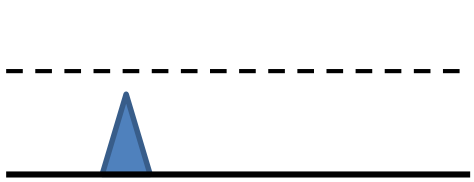


$$LR = \frac{0}{2pq}$$

Suspect



Evidence



$$LR = \frac{?}{2pq}$$

The Binary LR approach

“2p”

Probabilistic Approaches

- “Semi-Continuous” or “Fully Continuous”
- Semi-Continuous – information is determined from the alleles present – peak heights are not considered.
- Fully Continuous – incorporation of biological parameters (PHR [Hb], Mx ratio, Stutter percentage, etc…).

What should we do with discordant data?

- Continue to use RMNE (CPI, CPE) (not optimal)
- Use the Binary LR with 2p (not optimal)
- Semi-continuous methods with a LR (Drop models)

R. v Garside and Bates

- James Garside was accused of hiring Richard Bates to kill his estranged wife, Marilyn Garside.
- Marilyn was visiting her mother when someone knocked on the door. Marilyn answered and was stabbed to death.
- A profile from the crime scene stain gave a low-level DNA profile of the perpetrator.

Summary

Locus	Mrs Garside	Bates	CSP: minor component
D3	16,16	13,16	13
VWA	15,17	16,16	16
D16	11,12	11,12	–
D2	20,20	**19**,22	22
D8	12,13	8,13	8
D21	30,32.2	30,31.2	31.2
D18	14,14	**12,15**	–
D19	12,14	12,15	15
THO1	9.3,9.3	7,7	7
FGA	23,25	21,21	21

Three alleles were not present in the evidence

Court case

- Crown expert dropped the D18 locus (gave a LR = 1) from the statistical results and used “2p” for D2 to give an overall odds for Bates of 1 in 610,000.
- David Balding argued for the defense that dropping loci is not conservative.

Balding and Buckleton (2009)

Forensic Science International: Genetics 4 (2009) 1–10

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

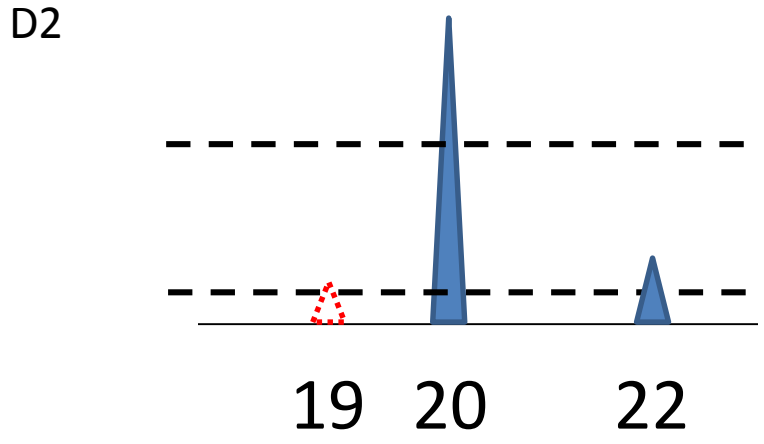
Interpreting low template DNA profiles

David J. Balding [a,*], John Buckleton [b]

[a] Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK
[b] ESR Private Bag 92021, Auckland, New Zealand

Present the "Drop model" for interpreting LT-DNA profiles

Drop Model



V = 20, 20

S = 19, 22

Pr(Drop-out) = 0.05

Pr(Drop-in) = 0.01

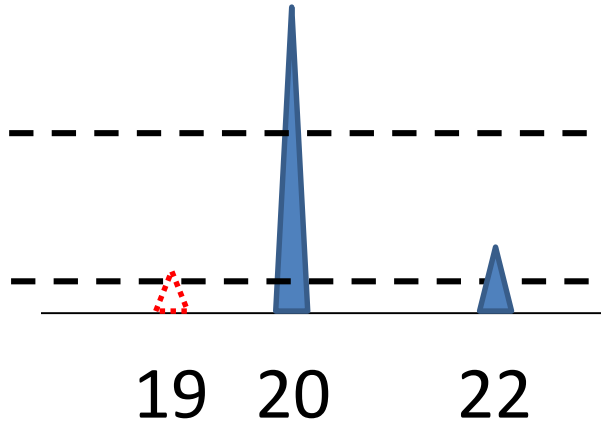
$P(E \mid H_1)$ = Pr(no Drop-out at 22) Pr(Drop-out at 19) Pr(No Drop-in)

= 0.95 0.05 0.99

= 0.047

Drop Model

D2



V = 20, 20

S = 19, 22

Pr(Drop-out) = 0.05

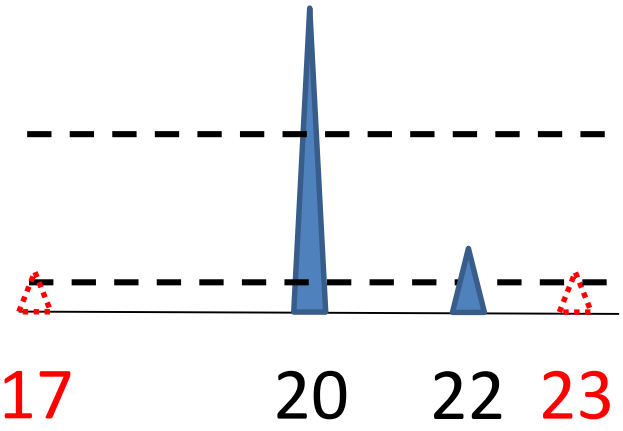
Pr(Drop-in) = 0.01

$$\frac{P(E \mid H_1)}{P(E \mid H_2)} = 0.047$$

The defense can now argue that someone else in the population unrelated to Bates was the true perpetrator!

Drop Model

D2



V = 20, 20

UC = 17, 23

Pr(Drop-out) = 0.05

Pr(Drop-in) = 0.01

$P(E \mid H_2)$

Pr(Drop-out at 17)	Pr(Drop-out at 23)	Pr(Drop-in at 22)
0.05	0.05	0.01

= 0.000025 x $2pq_{17,23}$ (0.027) = 0.000000675

Summary

- Using “2p” for D2 gave a LR = 11. This is non-conservative compared to the probabilistic approach where a Pr(D) was incorporated into the calculation, the LR = 2.8

- The use of a probabilistic approach uses all of the information in the profile.

Some Semi-Continuous Examples

- LR mix (Haned and Gill)
- Balding (likeLTD - R program)
- FST (NYOCME, Mitchell *et al.*)
- Kelly *et al.* (University of Auckland, ESR)
- Lab Retriever (Lohmueller, Rudin and Inman)
- Armed Expert (NicheVision)
- Puch-Solis *et al.* (LiRa and LiRaHT)
- GenoProof Mixture (Qualitype)

Semi-continuous methods

- Use a Pr(DO) and LRs
- Speed of analysis – “relatively fast”

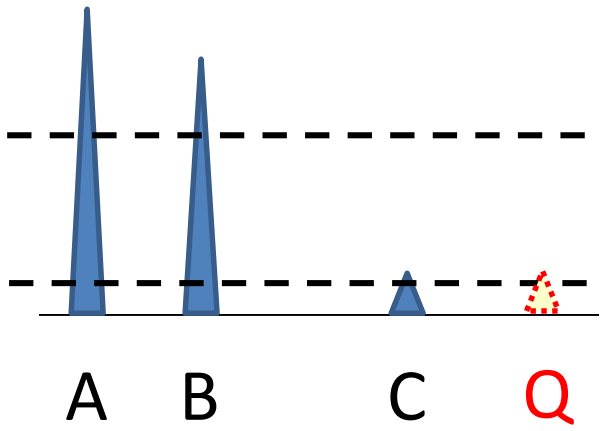
- The methods do not make full use of data - only the alleles present.

What should we do with discordant data?

- Continue to use RMNE (CPI, CPE) (not optimal)
- Use the Binary LR with 2p (not optimal)
- Semi-continuous methods with a LR (Drop models)
- Fully continuous methods with LR

Continuous Models

- Mathematical modeling of “molecular biology” of the profile (mix ratio, PHR (Hb), stutter, etc…) to find optimal genotypes, giving **WEIGHT** to the results.



Distribution of

Probable Genotypes

AC – 40%

BC – 25%

CC – 20%

CQ – 15%

Some Continuous Model Examples

- TrueAllele (Cybergenetics)
- STRmix (ESR [NZ] and Australian collaboration)
- DNA-View Mixture Solution (Charles Brenner)
- DNAmixtures (Graversen 2013a,b) – open source, but requires HUGIN.

Weights may be determined by performing simulations of the data (Markov Chain Monte Carlo - MCMC).

Fully continuous methods

- Use a Pr(DO) and LRs
- Biological modeling of the data parameters
- Speed of analysis – can vary

- Attempts to use all of the data

Summary

- Probabilistic Methods make better use of the data than RMNE or the binary LR with 2p.
- The goal of the software programs should not be to simply “get bigger numbers” but to understand the details of these approaches and not treat the software as a “black box.”
- Semi-continuous approaches will produce a LR that could be replicated by hand if necessary.
- Each approach has it’s own advantages and disadvantages.

Use modern tools for today's mixtures!

Statistical Evaluation of Forensic DNA Profile Evidence

Christopher D. Steele and David J. Balding

UCL Genetics Institute, University College London, London WC1E 6BT, United Kingdom;
email: c.steele.11@ucl.ac.uk, d.balding@ucl.ac.uk

Annual Review of Statistics and Its Application **(2014)**: 361-384

Acknowledgments

National Institute of Justice and NIST OLES

John Butler
Charlotte Word
Robin Cotton
Catherine Grgicak
Bruce Heidebrecht

John Paul Jones

Points of view are mine and do not necessarily represent the official position or policies of the US Department of Justice or the National Institute of Standards and Technology.

Certain commercial equipment, instruments, software and materials are identified in order to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that any of the materials, instruments or equipment identified are necessarily the best available for the purpose.

Contact info:
mcoble@nist.gov
+1-301-975-4330