DNA Mixture Analysis:

Principles and Practice of Mixture Interpretation and Statistical Analysis Using the SWGDAM STR Interpretation Guidelines

---

Different Approaches to Statistical Analysis of Mixtures

Todd Bille

ATFE DNA Laboratory

AAFS 2011 Workshop #17

Chicago, IL

February 22, 2011

Disclaimers

- Todd Bille is NOT here on behalf of SWGDAM or the ATF, and his comments do NOT necessarily represent the views of SWGDAM or the ATF
- My experience is with CPI for mixtures and RMP for single source profiles
- I'm a biochemist…

Thresholds

- Analytical Threshold:
  - The Analytical Threshold is the relative fluorescent unit (rfu) value that, when exceeded by peaks that conform to the parameters defining a peak, allows those peaks to be considered "real" products of amplification.
  - Typically 3 times the standard deviation of the baseline noise.
  - Can be estimated as 2 times the maximum peak to trough signal.

- Stochastic Threshold:
  - Due to the inherent nature of PCR amplification of low levels of DNA, the results may contain dramatic peak height imbalance and allele drop-out. The stochastic threshold is the rfu value that, when exceeded by a single allelic peak, the DNA analyst can be confident that the sister peak of a heterozygous pair would be detected (i.e. would be above the Analytical Threshold). The Stochastic Threshold listed below refers to a single source sample. The use of the Stochastic Threshold must be modified for the interpretation of mixed DNA profiles due to the possible additive effects of allele sharing.

Thresholds

- Stochastic Threshold: how can the stochastic threshold be determined?
  - The laboratory can analyze replicate dilution series samples and determine at what point drop-out of the sister allele is not observed for a heterozygous pair.

OR

  - The laboratory can do what we did…

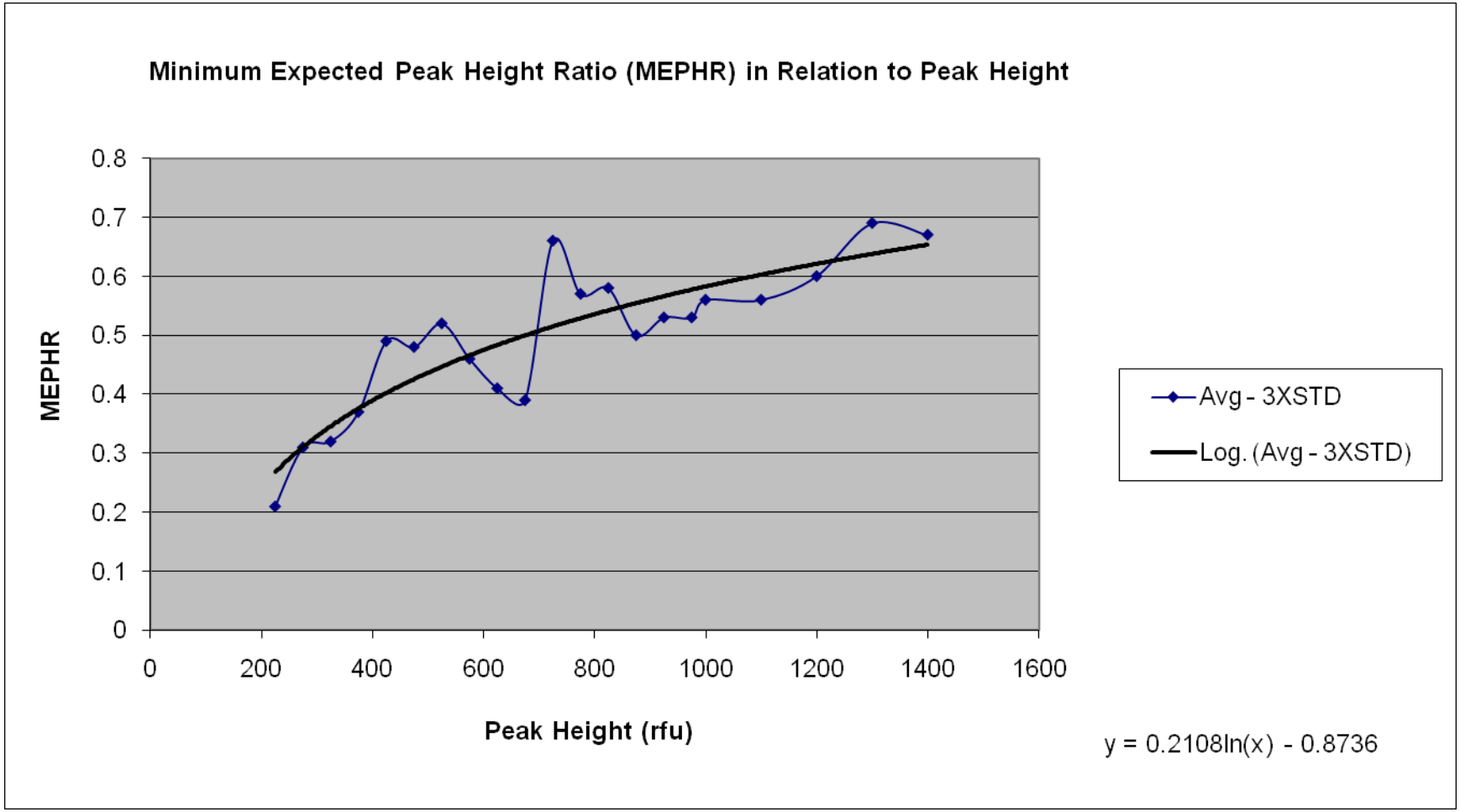
Thresholds

- We examined 1,600 heterozygous loci with rfu values ranging from >2,000 down to < 50.
- We ordered the heterozygous pairs from greatest rfu to the least based on the tallest peak in the pair.
- We then created overlapping ranges within the data and calculated the average peak height balance and the standard deviation.

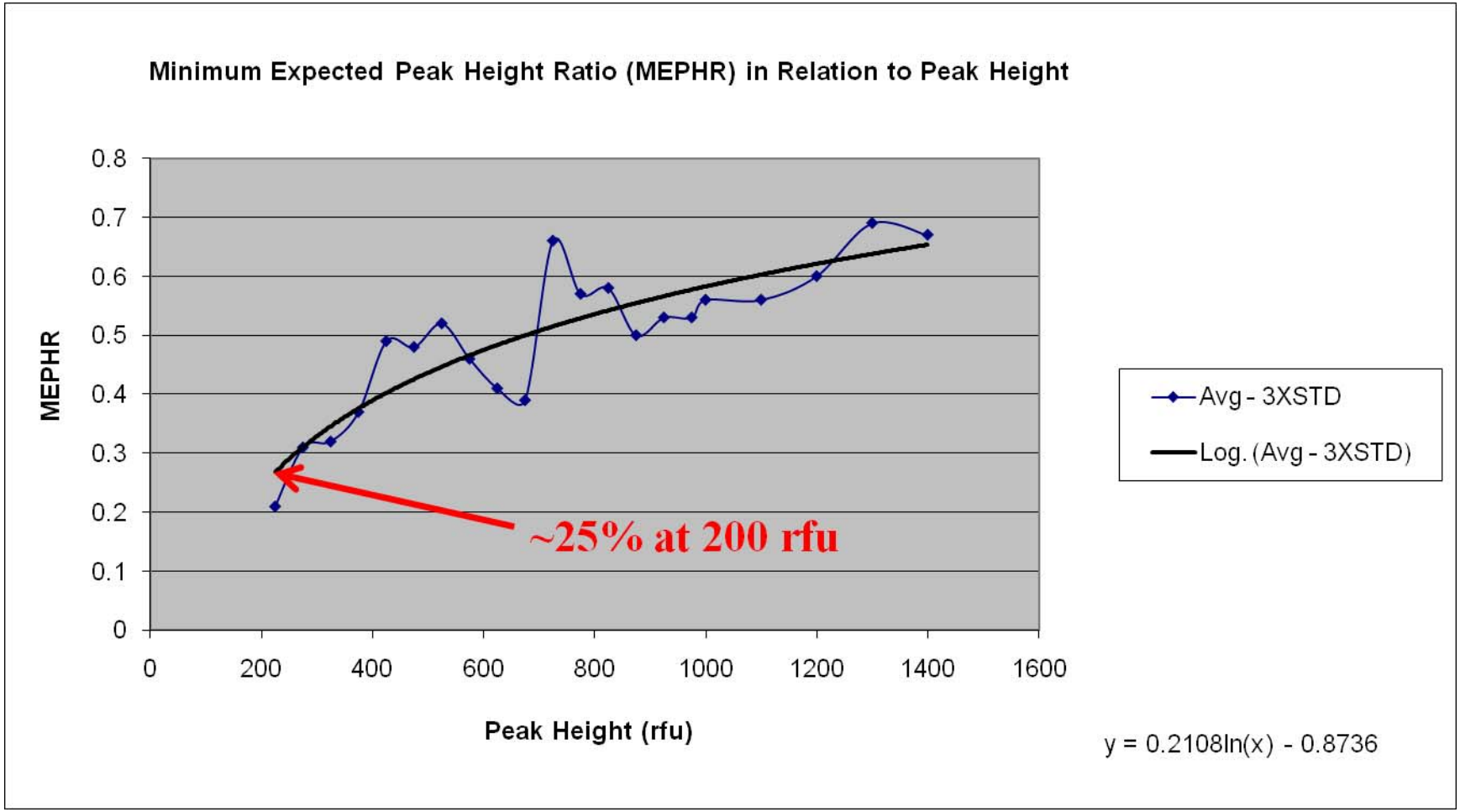
Thresholds

- From this data we create a curve using the average peak height balance – 3 X Std Dev
  - This provides a way to extrapolate the minimum expected peak height balance associated with a single peak's rfu.
  - Since we used 3 X Std Dev, we would expect 99.7% of the events observed should fall within this range. Therefore, 3 in 1,000 would be expected to fall outside this range (or 1 in 333).

Thresholds


Minimum Expected Peak Height Ratio (MEPHR) in Relation to Peak Height
MEPHR
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400
1600
Peak Height (rfu)
Avg - 3XSTD
Log. (Avg - 3XSTD)
y = 0.2108ln(x) - 0.8736

Thresholds


Minimum Expected Peak Height Ratio (MEPHR) in Relation to Peak Height
MEPHR
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
200
400
600
800
1000
1200
1400
1600
Peak Height (rfu)
~25% at 200 rfu
Avg - 3XSTD
Log. (Avg - 3XSTD)
y = 0.2108ln(x) - 0.8736

Thresholds

- Therefore, if a single peak was detected below 200 rfu, the sister peak may fall below the Analytical Threshold of 50 rfu.
- Instead of having a single minimum expected peak height balance, we use this graph.
- The average peak height balance ranged from 90% to 71% . The standard deviation ranged from 8 to 26.

Thresholds

- If a change to a method is made that increases the sensitivity of the amplification or detection, the thresholds must be re-evaluated.
  - For example: we found in our laboratory if we doubled the injection time, the Stochastic Threshold doubled, as well.
  - Other "enhanced detection methods" examples include:
    - Post-amp de-salting
    - Increased cycle #
    - Increased product loaded on capillary

Why Do Stats?

- A way to assess the weight of the following statements:
  - "The DNA profile obtained from the swab of the knife is consistent with the known DNA profile of John Doe."
  - "John Doe cannot be excluded as a contributor to the mixed DNA profile obtained from the swabbings of the grips of the firearm."
- To be non-prejudicial
- 4.1 The laboratory ***must*** perform statistical analysis in support of any inclusion that is determined to be relevant in the context of a case, irrespective of the number of alleles detected and the quantitative value of the statistical analysis.

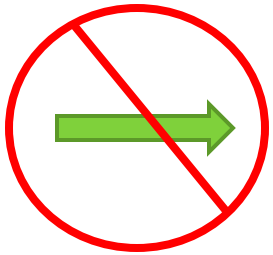
Why Do Stats?

- The statistical analysis differentiates between a DNA match between a known and an evidentiary partial profile with three loci detected and an evidentiary full profile.

- The statistical analysis differentiates between a mixture where six members of the jury could be potential contributors and one where only one person in two hundred million would be potential contributors.

Why Do Stats?

- Reporting that someone is not excluded ***is*** the same as stating that they are included.
- Bottom line, if you can't put number with it, you shouldn't report it as inclusion.
- "You can't make chicken salad out of chicken $#@t."


What Approach??

- CPE/CPI (RMNE)
- RMP
- LR

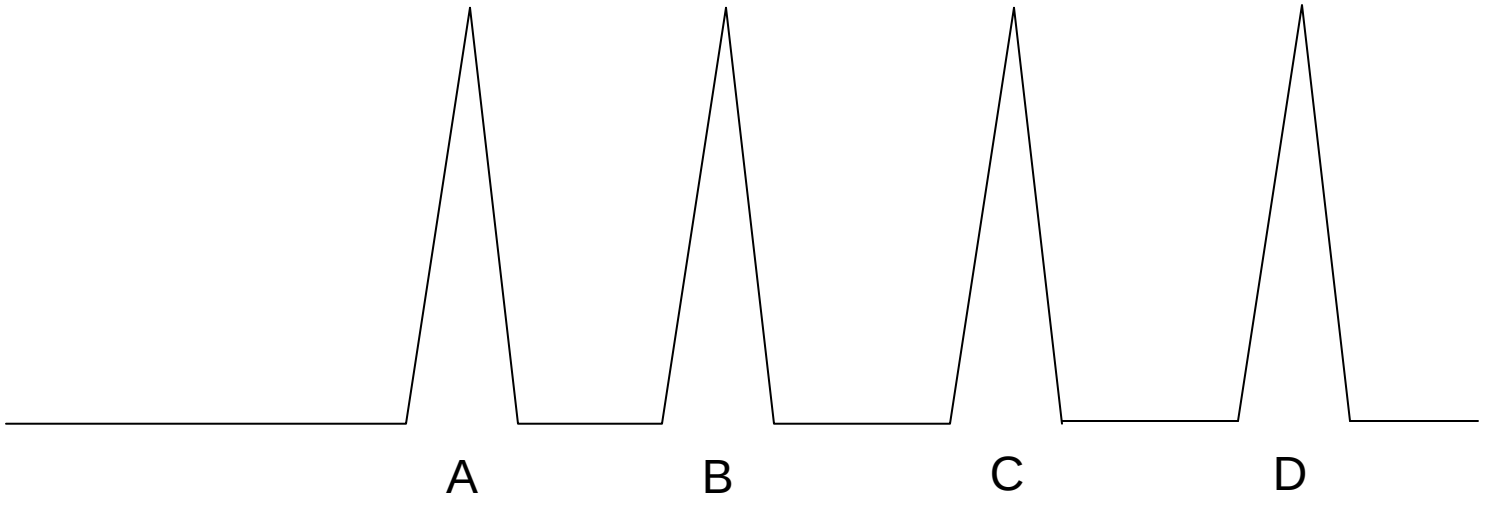
CPE/CPI (RMNE)

- Advantages
  - No assumptions
  - Simple to calculate
  - Simple to explain
  - Independent of reference profiles (statistical calculation could be done prior to comparison to references)
- Disadvantages
  - Doesn't make the best use of available data

What Approach??

- 4.2 For calculating the CPE or RMP, any DNA typing results used for statistical analysis must be derived from evidentiary items and not known samples.

CPE/CPI (RMNE)



Answers the question: what percentage of the population could be a possible contributor to the observed profile

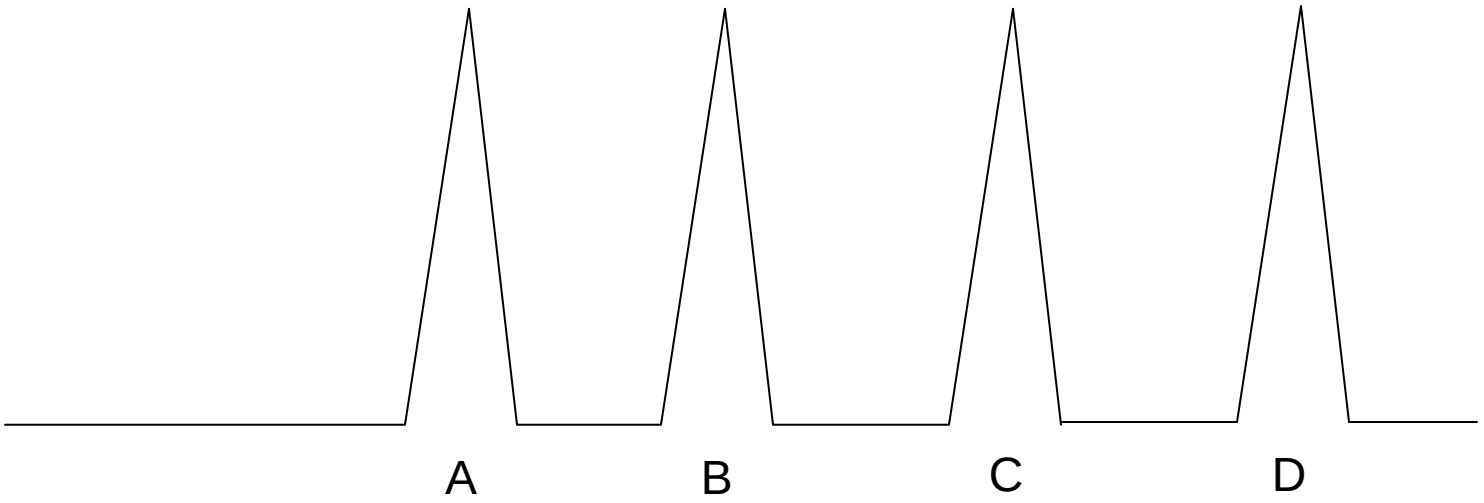
CPI = AA + AB + AC + AD + BB + BC + BD + CC + CD + DD
CPI = $(A + B + C + D)^2$

Therefore, AE, BF, CG, etc. would be excluded

RMP

- Advantages
  - Makes better use of the data
  - Simple to calculate
  - Simple to explain
  - Independent of reference profiles (statistical calculation could be done prior to comparison to references)
- Disadvantages
  - Assumption must be made

RMP



Answers the question: Assuming a specific number of contributors, what percentage of the population could be a possible contributor to the profile. Assuming a two person mixture, the RMP calculation is equal to the following:

$$RMP = (A + B + C + D)^2 - A^2 - B^2 - C^2 - D^2$$

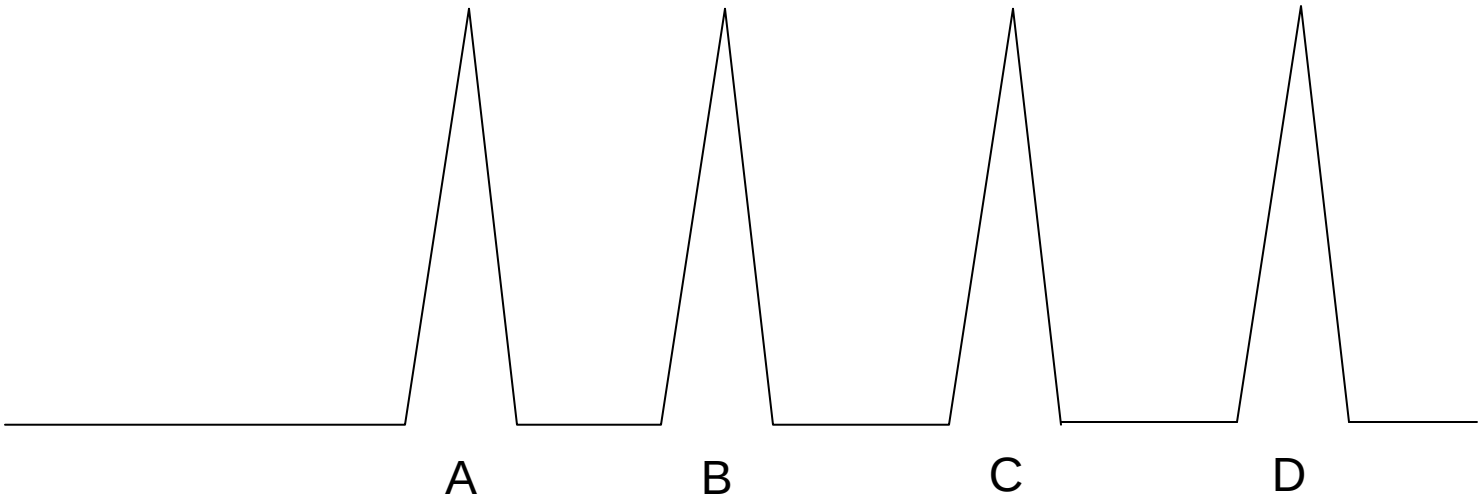
This is the equivalent of the sum of the genotypic frequencies for any combination of heterozygous genotype

LR

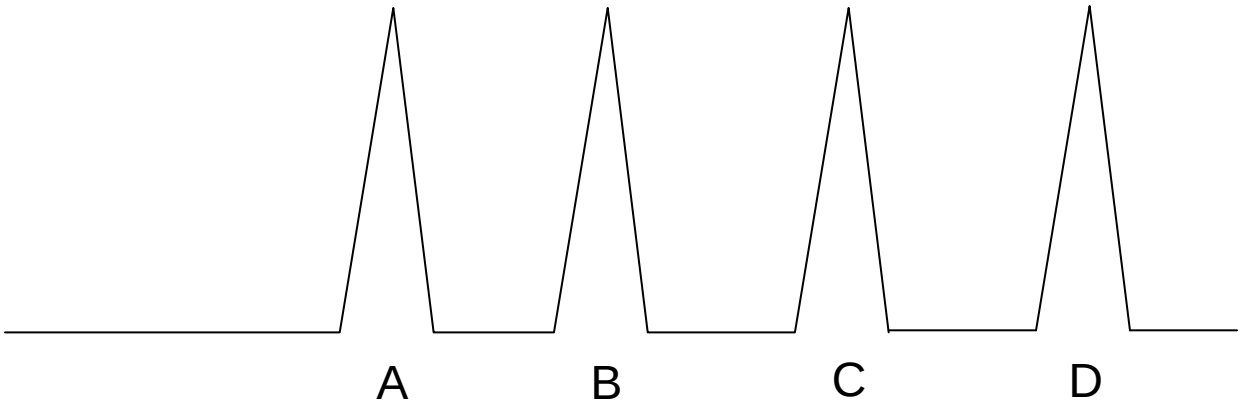
- Advantages
  - Makes even better use of the data
  - Dependent on reference profile (calculation done after comparison to references)
- Disadvantages
  - Assumptions must be made
  - More complex to calculate
  - More difficult to explain

LR

- Likelihood ratio is the comparison of two differing hypotheses: the "prosecutor's" (numerator) and the "defense's" (denominator) hypotheses



LR



Assuming a two person mixture with an unknown second contributor and a suspect profile of AB:

Numerator: Given this evidence, what is the probability of obtaining this result under the assumption that the suspect is a contributor = 2CD

Denominator: Given this evidence, what is the probability of obtaining this result from two random individuals = 24ABCD

LR

- Where the hell did 24ABCD come from?
  - AB / CD = 2AB x 2CD = 4ABCD
  - Six combos of this = 24ABCD
  - Simplified LR = 2CD / 24ABCD = 1 / 12AB

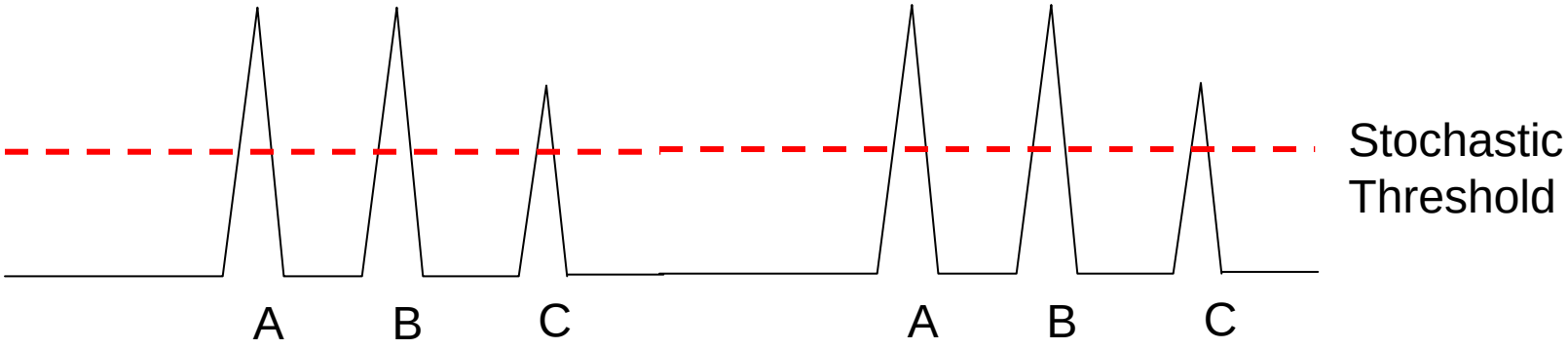
CPI / RMP

- 4.2. For calculating the **CPE or RMP**, any DNA typing results used for statistical analysis must be derived from evidentiary items and not known samples.

CPI

- 4.6.3. When using **CPE/CPI** (with no assumptions of number of contributors) to calculate the probability that a randomly selected person would be excluded/included as a contributor to the mixture, ***loci with alleles below the stochastic threshold may not be used for statistical purposes to support an inclusion***. In these instances, the potential for allelic dropout raises the possibility of contributors having genotypes not encompassed by the interpreted alleles.

CPI / RMP



Locus 1: What genotypes are possible contributors to this mixture.

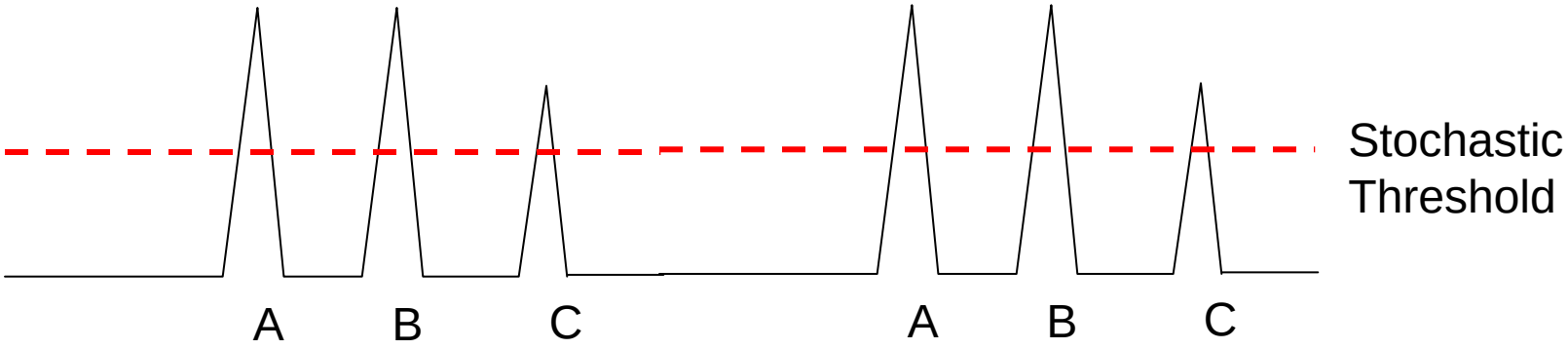
AA	AB	AC
BB	BC	CC

Locus 2: What genotypes are possible contributors to this mixture.

AA	AB	AC
BB	BC	CC

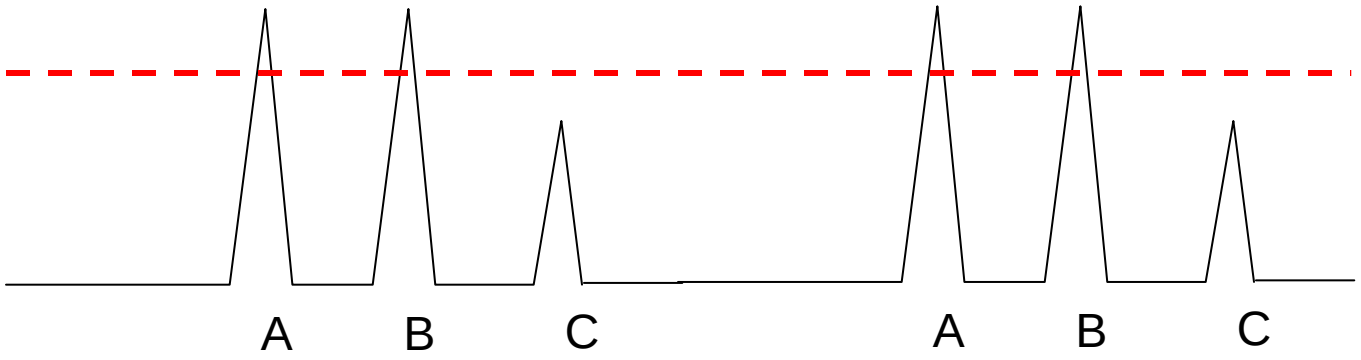
Suspect: AB / BC – Included, CPI and RMP stat equivalent

CPI / RMP



Suspect: AB / **CD** – Excluded

CPI / RMP



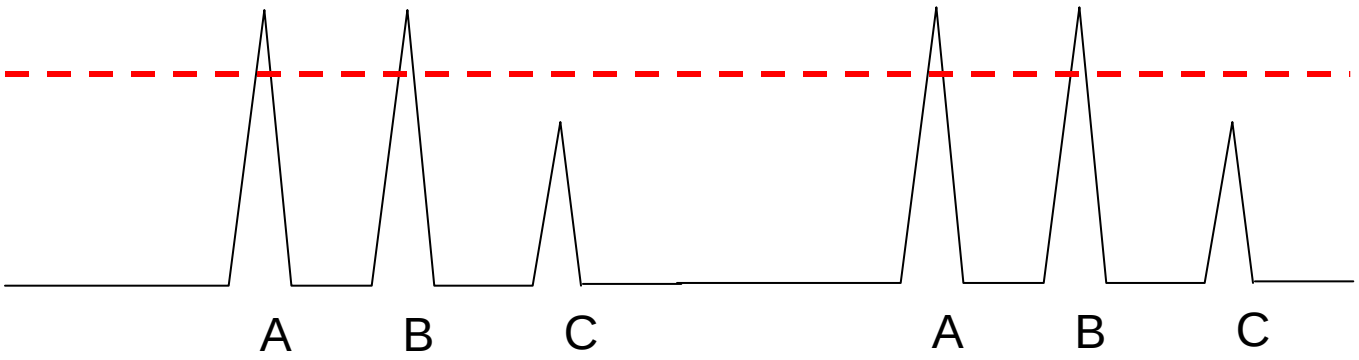
Locus 1: What genotypes are possible contributors to this mixture.

AA AB AC
BB BC CC CD CE CF etc.

Locus 2: What genotypes are possible contributors to this mixture.

AA AB AC
BB BC CC CD CE CF etc.

CPI / RMP

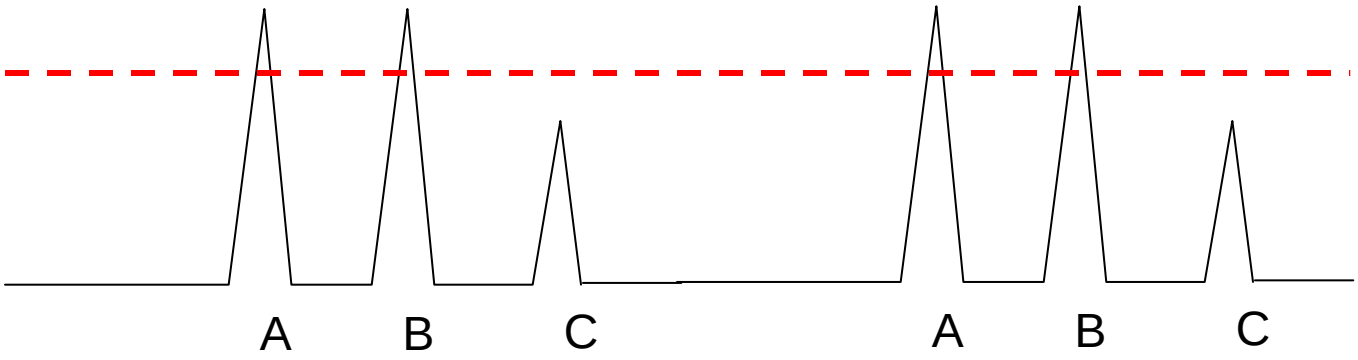


Suspect: AB / BC – Included, neither locus can be used for CPI

RMP – either not use both loci or incorporate 2p for the C allele with both loci.

This decision in both instances is made prior to comparison to the reference sample

CPI / RMP



Suspect: AB / CD

The statistical calculation must be based on the evidence profile, not the reference profile.

Not proper to use Locus 1 for statistics and then not use Locus 2 for statistics

"2p Rule"

- "2p Rule" can be used to statistically account for zygosity ambiguity – i.e. is this single peak below my stochastic threshold that I'm seeing the result of a homozygous genotype or the result of a heterozygous genotype with allele drop-out of the sister allele?

"2p Rule"

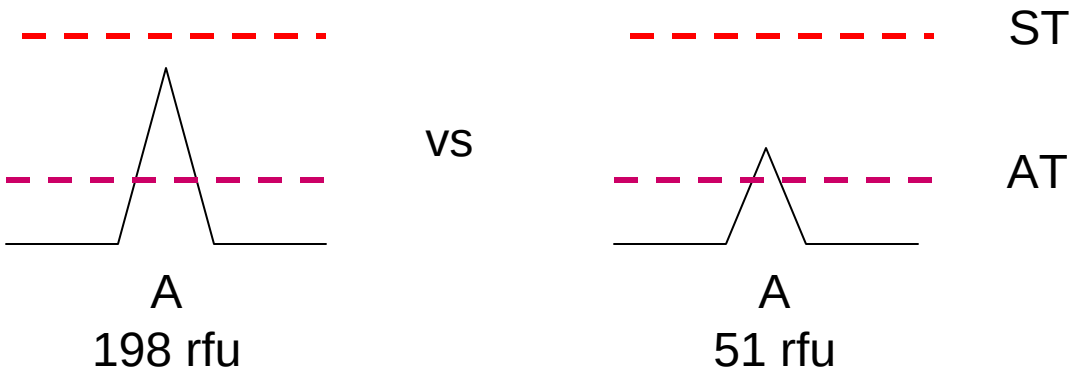
- Where does it come from?
  - Two ways of thinking of it
    - Typical heterozygote: 2pq. In this case, the frequency of the q allele is unknown, use "1" instead. Therefore, the formula is 2p.
    - Another way of looking at it is defining what it accounts for. In a 5 allele system (alleles A – E), where A was detected below the stochastic threshold, any genotype associated with the A allele would not be excluded, i.e. AA, AB, AC, AD, AE.
      - Probability = $A^2 + 2AB + 2AC + 2AD + 2AE$
      - Probability = $A^2 + 2A(B + C + D + E)$
      - Probability = $A^2 + 2A(1 - A)$
      - Probability = $A^2 + 2A - 2A^2$
      - Probability = $2A - A^2$

"2p Rule"

- The "2p Rule" cannot be used with CPI
- 4.6.3 When using CPE/CPI (with no assumptions of number of contributors) to calculate the probability that a randomly selected person would be excluded/included as a contributor to the mixture, ***loci with alleles below the stochastic threshold may not be used for statistical purposes to support an inclusion***. In these instances, the potential for allelic dropout raises the possibility of contributors having genotypes not encompassed by the interpreted alleles.

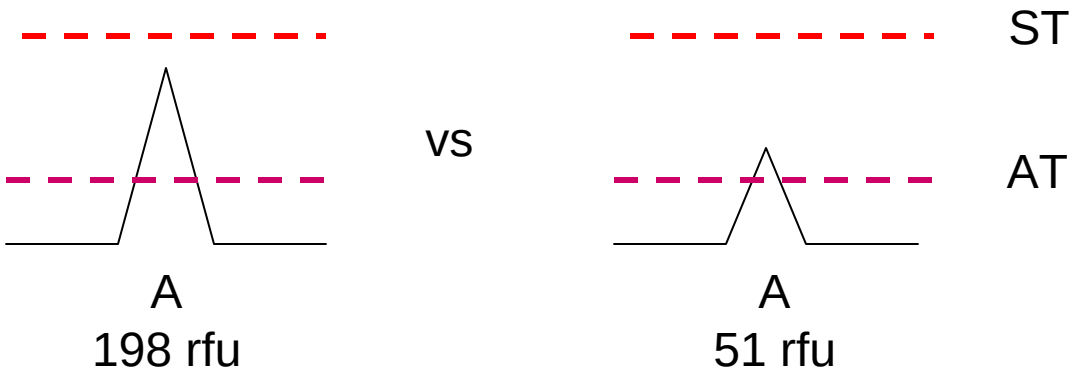
Probability of Drop-out Pr(D)

- The probability that drop-out occurred associated with peaks below the stochastic threshold is not equal across the range of peak heights.



Probability of Drop-out Pr(D)

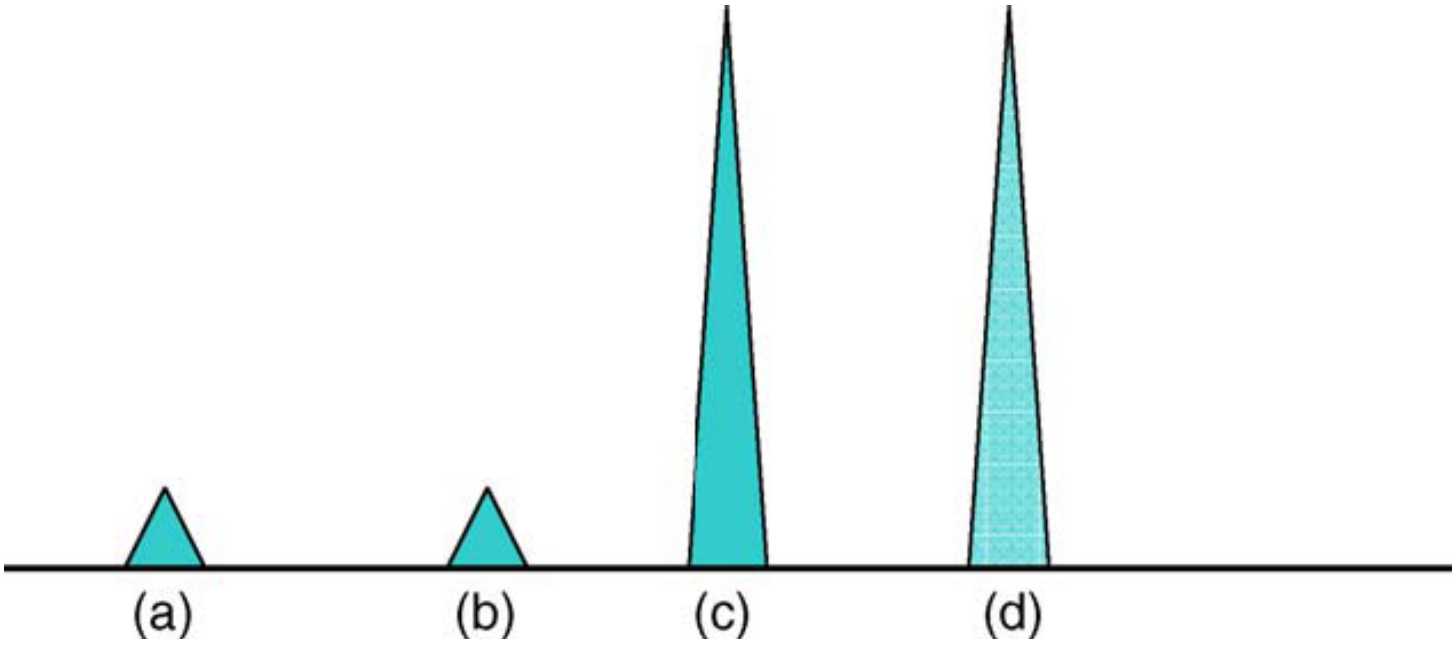
- Therefore, it is not always conservative to use the *2p rule* or drop the locus for statistical purposes.
- If the suspect is an A,B at the locus below, the Pr(D) approaches zero as the allelic peak nears the Stochastic Threshold.



ISFG 2006

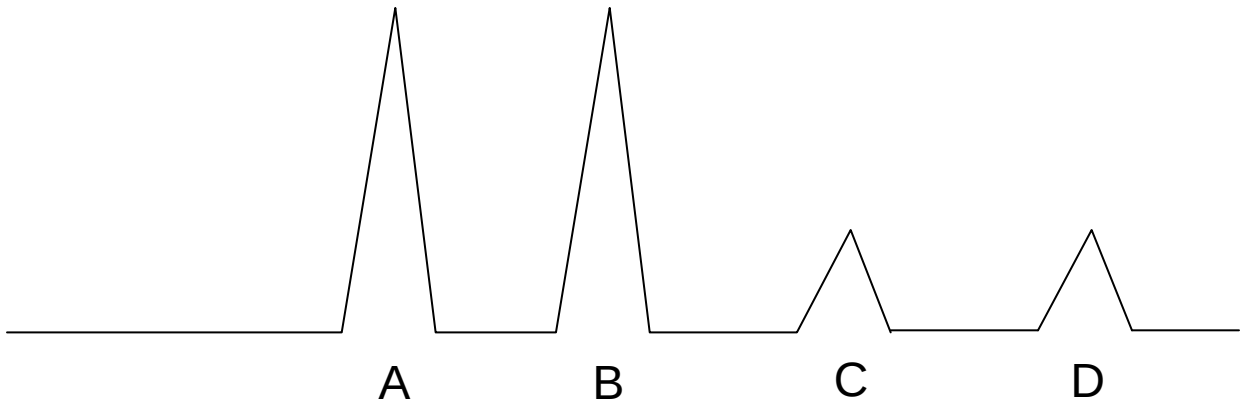
- *"Stutters (from a major contributor) may be the same height/peak area as the minor contributor to the mixture. This means (Fig. 4) that those bands in stutter positions may be allele only, allele plus stutter, or stutter only. In Fig. 4, bands a, b are minor alleles that are very similar in height/area. Band b is in a stutter position and we must assume that it could be from an unknown contributor under Hd. Consequently, if we condition on the number of contributors = 2, then the possible minor contributor genotypes are aa, ac, ad (where b is a stutter), or ab (where b is an allele either with or without a stutter)."*

Figure 4: ISFG 2006

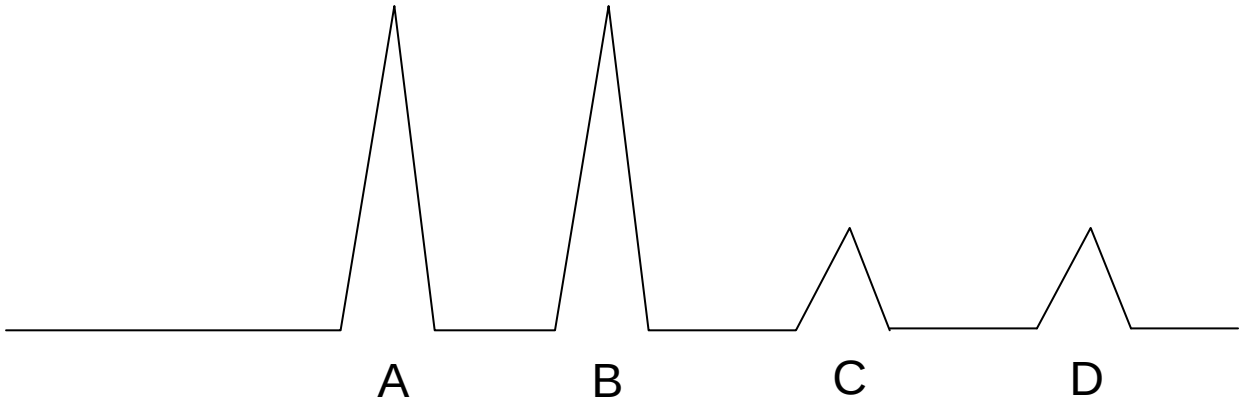


Modified vs Restricted

- ***Modified***: as in Modified RMP, an assumption as to the number of contributors has been made
- ***Restricted***: peak height information is taken into account during the statistical analysis



Restricted vs Unrestricted



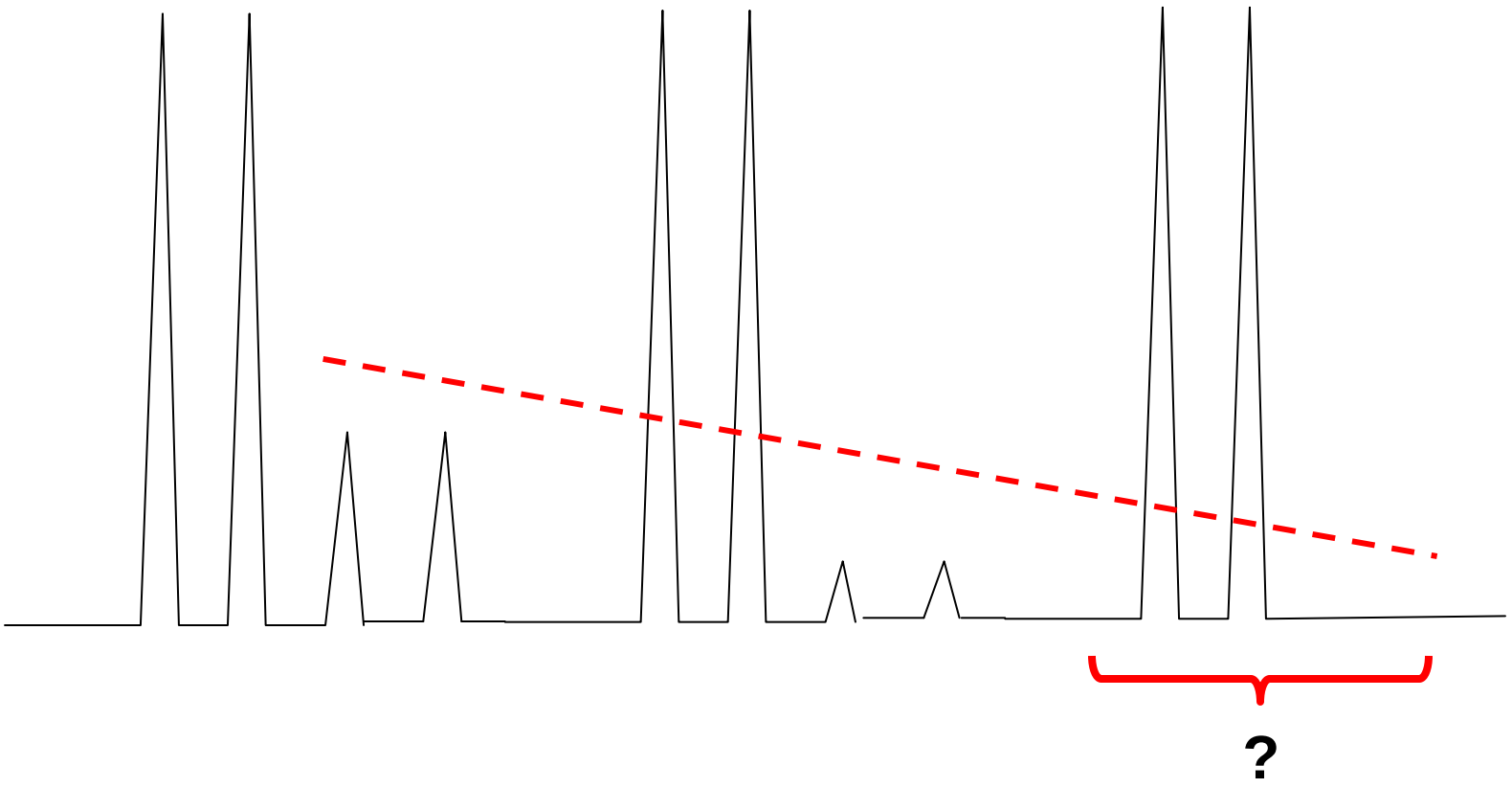
Unrestricted RMP: All heterozygous genotypes considered possible

Restricted RMP: AC, BC, AD, BD genotypes not considered as possible contributors to the mixture

What if the minor component may not be detected at a locus?

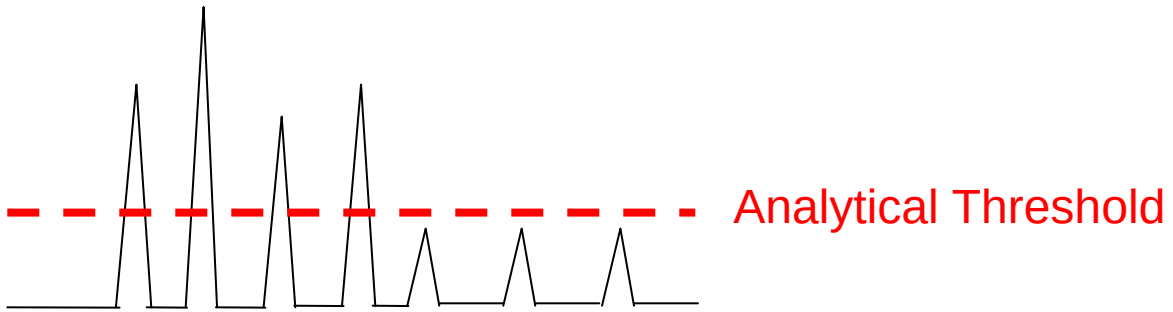
- 5.1 Whenever the statistical analysis at a locus is meant to represent all possible contributors to a mixture, if there is a reasonable possibility that locus dropout could have led to the loss of an entire genotype, then a statistical calculation should not be performed for that locus.

What if the minor component may not be detected at a locus?



Data Below Analytical Threshold

- Should the analyst use data below the analytical threshold?
  - Determination of number of contributors?
  - Only 4 or less alleles detected above the analytical threshold, but all at between 50 rfu and 250 rfu.
  - Multiple peaks are observed below the analytical threshold with the correct peak morphology and fall within allelic bins. Including these peaks, the number of potential alleles at several loci > 4.
  - Still feel comfortable making the assumption the profile is the result of a two person mixture?

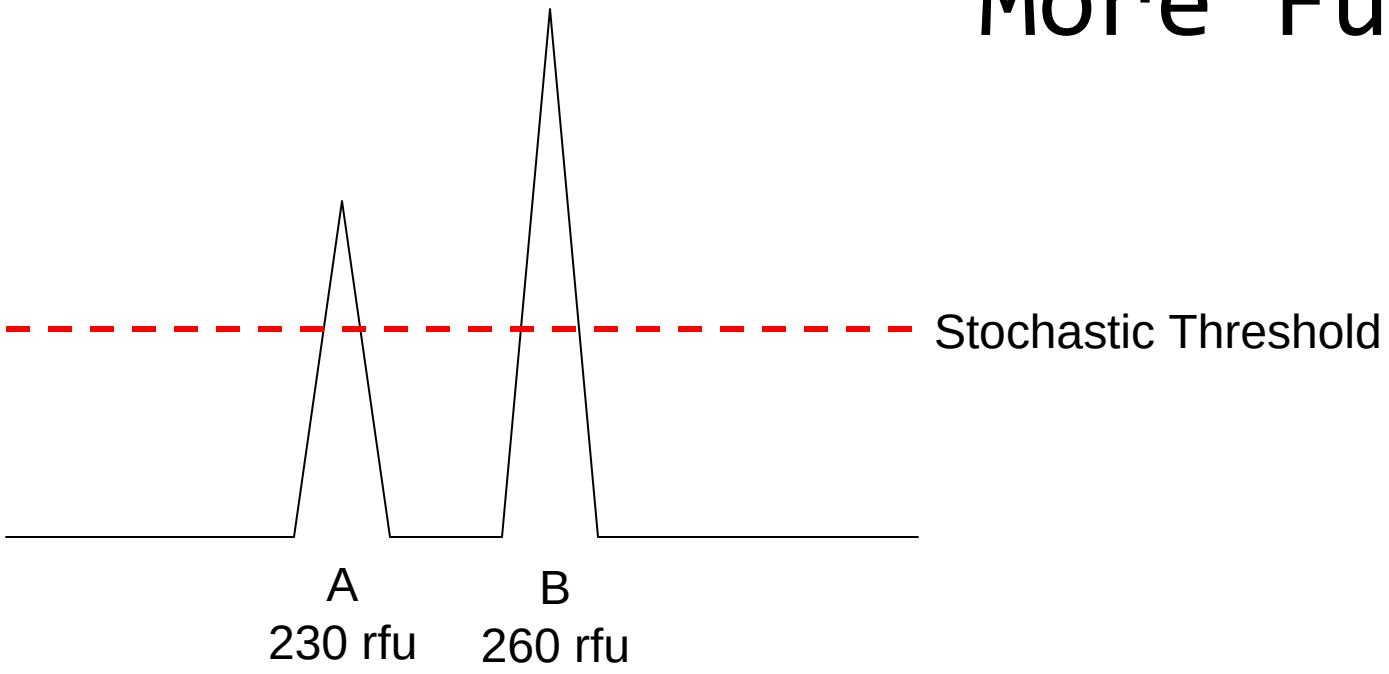


Major Contributor?

- German Stain Commission:

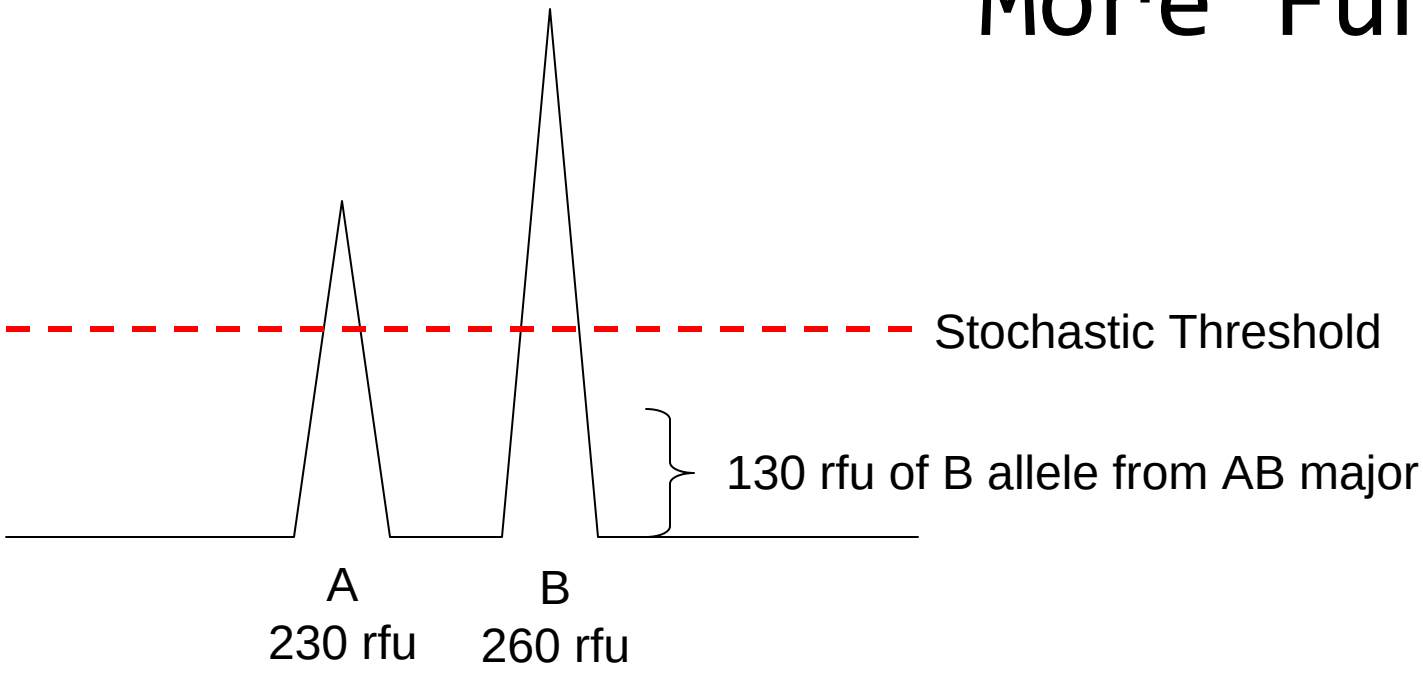
"Classification of mixed stains
Type A has no obvious major contributor with no evidence of stochastic effects.2 **Type B** ***has clearly distinguishable major and minor DNA components; consistent peak height ratios of approximately 4:1 (major to minor component) across all heterozygous systems, and no evidence of stochastic effects.*** **Type C** has mixtures with no major component(s) and evidence of stochastic effects."

More Fun…



Two person mixture, approximately a 3:1 ratio of components.
Both alleles above stochastic threshold, therefore no possible drop-out?

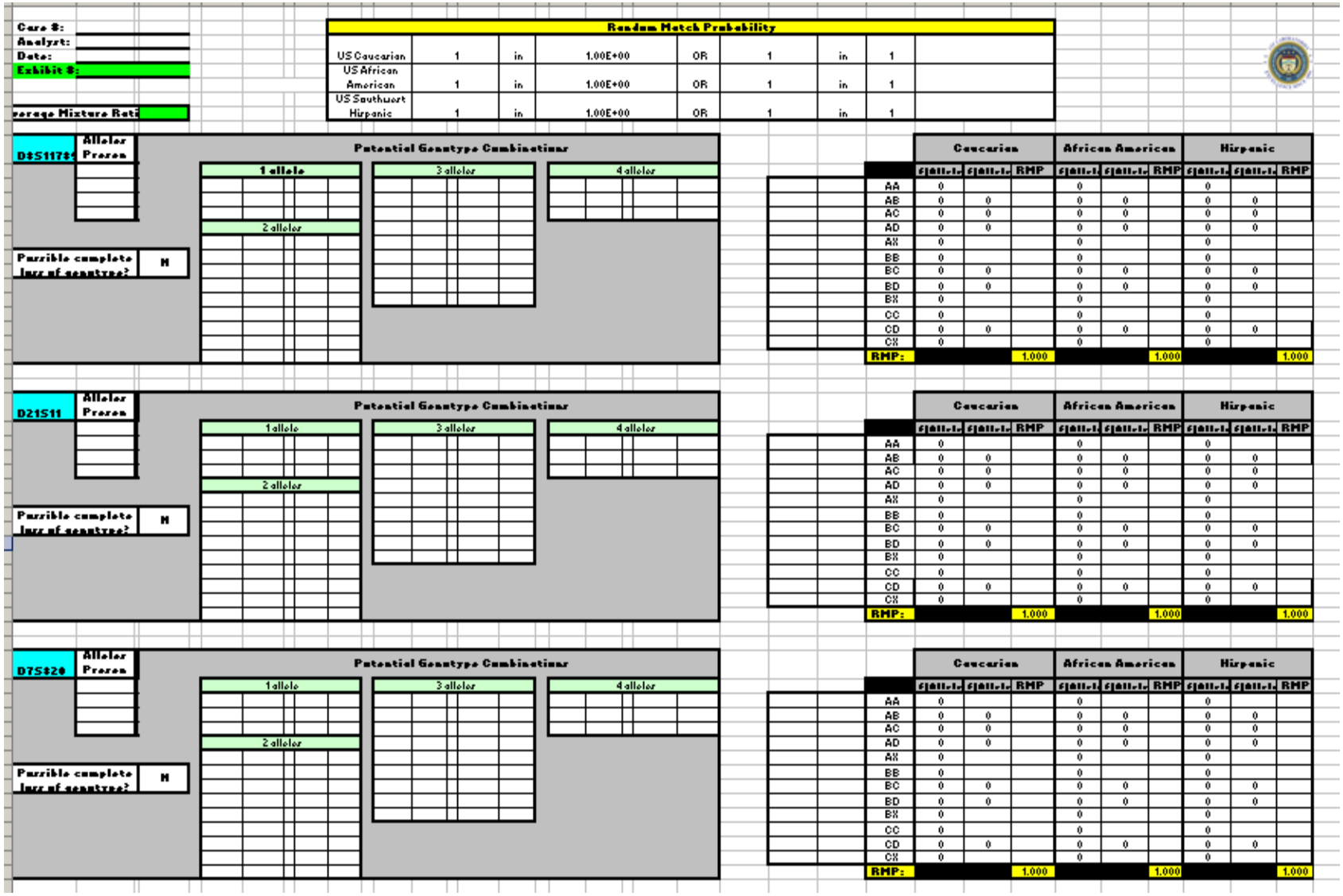
More Fun…



The stochastic threshold was based on a single source sample. The potential for allele sharing must be considered when examining a mixture.

If 130 rfu of the B allele is from the AB major, then 130 rfu of the B allele peak would remain. This is below the stochastic threshold of 200 rfu, but still falls within the 3:1 mixture ratio.

Documentation



Case #:
Analyst:
Date:
Exhibit #:

Random Match Probability

US Caucasian	1	in	1.00E+00	OR	1	in	1
US African American	1	in	1.00E+00	OR	1	in	1
US Southwest Hispanic	1	in	1.00E+00	OR	1	in	1

Allelic Pairs

Potential Genotype Combinations: 1 allele, 2 alleles, 3 alleles, 4 alleles

Possible complete loci of genotype? H

	Caucasian			African American			Hispanic		
	[illegible]	[illegible]	RMP	[illegible]	[illegible]	RMP	[illegible]	[illegible]	RMP
AA	0			0			0		
AB	0	0		0	0		0	0	
AC	0	0		0	0		0	0	
AD	0	0		0	0		0	0	
AX	0			0			0		
BB	0			0			0		
BC	0	0		0	0		0	0	
BD	0	0		0	0		0	0	
BX	0			0			0		
CC	0			0			0		
CD	0	0		0	0		0	0	
CX	0			0			0		
RMP:			1.000			1.000			1.000

D21S11

Allelic Pairs

Potential Genotype Combinations: 1 allele, 2 alleles, 3 alleles, 4 alleles

Possible complete loci of genotype? H

	Caucasian			African American			Hispanic		
	[illegible]	[illegible]	RMP	[illegible]	[illegible]	RMP	[illegible]	[illegible]	RMP
AA	0			0			0		
AB	0	0		0	0		0	0	
AC	0	0		0	0		0	0	
AD	0	0		0	0		0	0	
AX	0			0			0		
BB	0			0			0		
BC	0	0		0	0		0	0	
BD	0	0		0	0		0	0	
BX	0			0			0		
CC	0			0			0		
CD	0	0		0	0		0	0	
CX	0			0			0		
RMP:			1.000			1.000			1.000

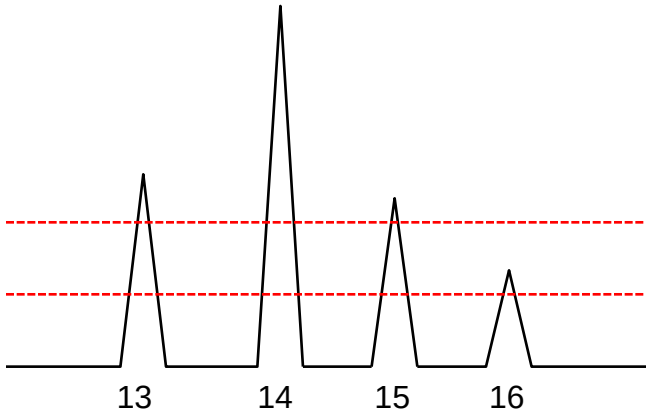
D7S820

Allelic Pairs

Potential Genotype Combinations: 1 allele, 2 alleles, 3 alleles, 4 alleles

Possible complete loci of genotype? H

	Caucasian			African American			Hispanic		
	[illegible]	[illegible]	RMP	[illegible]	[illegible]	RMP	[illegible]	[illegible]	RMP
AA	0			0			0		
AB	0	0		0	0		0	0	
AC	0	0		0	0		0	0	
AD	0	0		0	0		0	0	
AX	0			0			0		
BB	0			0			0		
BC	0	0		0	0		0	0	
BD	0	0		0	0		0	0	
BX	0			0			0		
CC	0			0			0		
CD	0	0		0	0		0	0	
CX	0			0			0		
RMP:			1.000			1.000			1.000



D8S11789	Alleles Present
	13
	14
	15
	16

Possible complete loss of genotype?	N

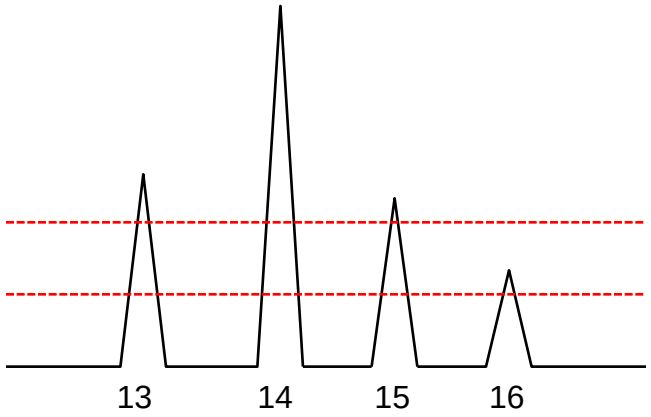
Potential Genotype Combinations

1 allele				

2 alleles				

3 alleles				

4 alleles				
13	14	y	15	16
13	15	n	14	16
13	16	y	14	15



			Caucasian			African American			Hispanic		
			f(Allele 1)	f(Allele 2)	RMP	f(Allele 1)	f(Allele 2)	RMP	f(Allele 1)	f(Allele 2)	RMP
		AA									
13	14	AB	0.339	0.202	0.137	0.222	0.333	0.148	0.325	0.246	0.160
		AC									
13	16	AD	0.339	0.013	0.009	0.222	0.044	0.020	0.325	0.025	0.016
		AX									
		BB									
14	15	BC	0.202	0.110	0.044	0.333	0.214	0.143	0.246	0.116	0.057
		BD									
		BX									
		CC									
15	16	CD	0.110	0.013	0.003	0.214	0.044	0.019	0.116	0.025	0.006
		CX									
		RMP:			0.192			0.329			0.239