8-2007

CSI Revisited: The Science of Forensic DNA Analysis

Michael L. Raymer
Wright State University - Main Campus, michael.raymer@wright.edu

CSI Revisited

The Science of Forensic DNA Analysis

Michael L. Raymer, Ph.D.

forensic bioinformatics

Growth in the importance of DNA

- Roughly 900,000 felony *convictions* per year in the U.S.
- DNA profiles are generated primarily for sexual offenses, murder, and assault
  - Often the key source of physical evidence
- The F.B.I. has established the CODIS database, with over 2 million DNA profiles
  - Allows "cold hit" searches for unresolved cases

DNA evidence misconceptions

- Everyone's DNA profile is unique
- DNA testing is always an objective and scientific process
- DNA testing is infallible
- DNA evidence is carefully evaluated by both the prosecution and the defense

Science and Art

The science of DNA testing is sound

but

not all DNA testing is done scientifically

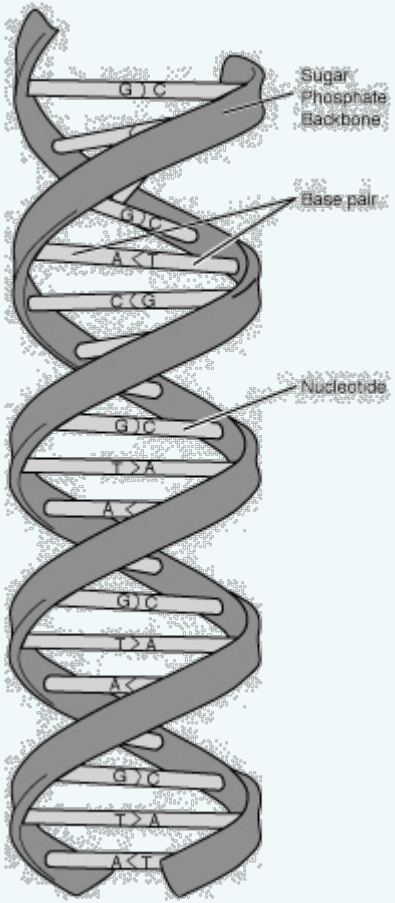
Background: DNA

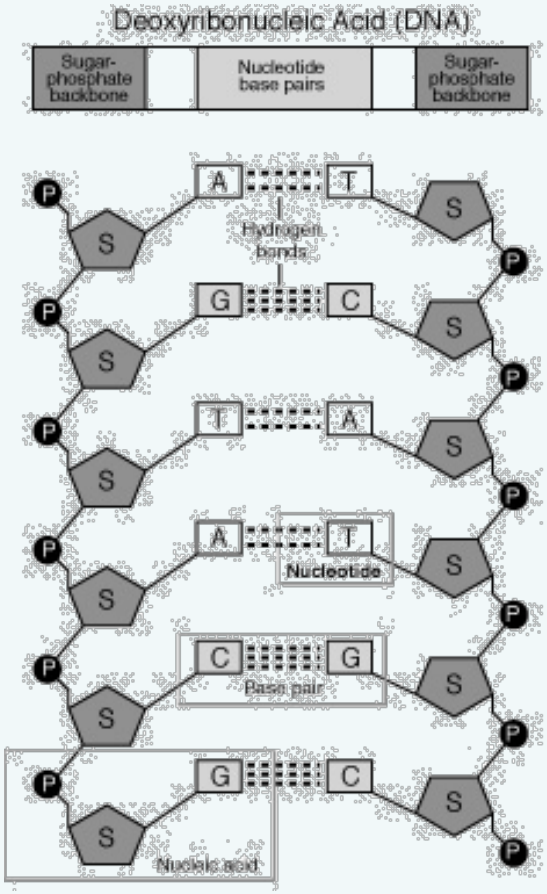
- DNA is found in each human cell

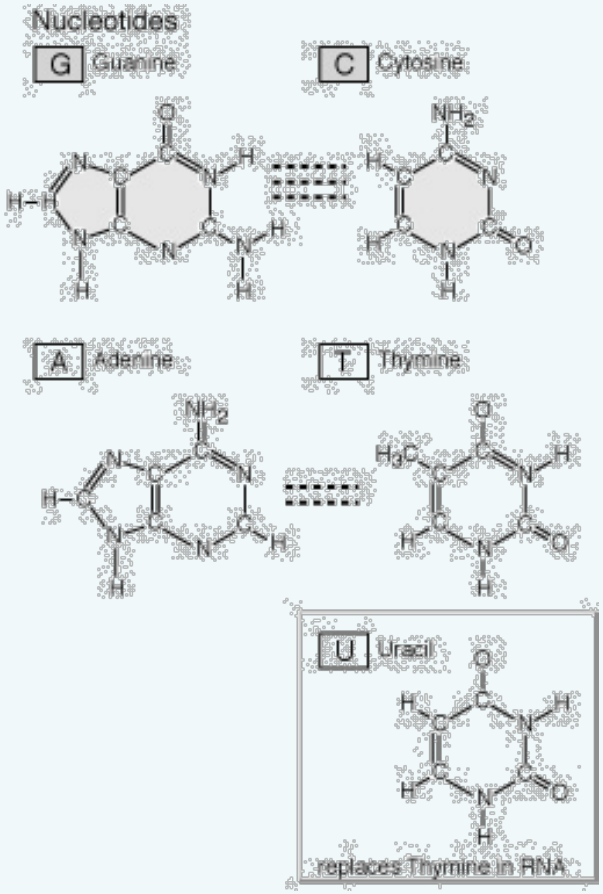
Type of Sample	Amount of DNA
Blood	30,000 ng/mL
1cm^2 stain	200 ng
1mm^2 stain	2 ng
Semen	250,000 ng/mL
postcoital vaginal swab	0 – 3,000 ng
Hair	
plucked	1 – 750 ng/hair
shed	1 – 12 ng/hair
Saliva	5,000 ng/mL
Urine	1 – 20 ng/mL

Background: DNA structure

- DNA is a polymer of nucleotides
  - Four building blocks: A, C, G, T







Background: DNA information content

- Most DNA (as much as 90%) is non-coding, or "junk" DNA
- More than 99% of the DNA is identical between any two humans
  - Regions of difference: "polymorphic"
- Changes to DNA are *random*, and usually ***bad***

Non-coding DNA exhibits higher polymorphism

STRs

- Short Tandem Repeat = STR
- Describes a type of DNA polymorphism in which:
  - a DNA sequence repeats
  - over and over again
  - and has a short (usually 4 base pair) repeat unit
- A length polymorphism – alleles differ in their length

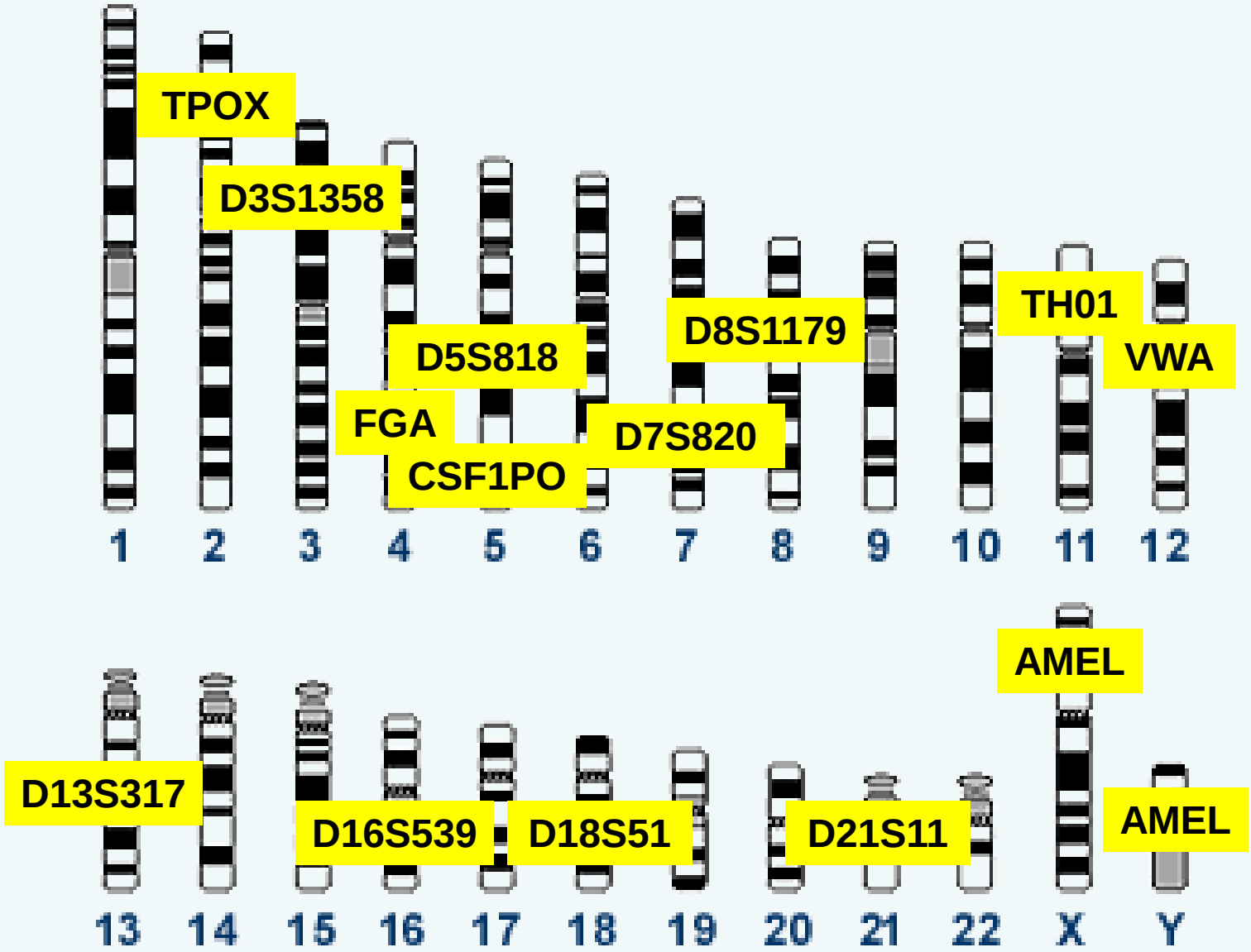
3 repeats: **AATG AATG AATG**

4 repeats: **AATG AATG AATG AATG**

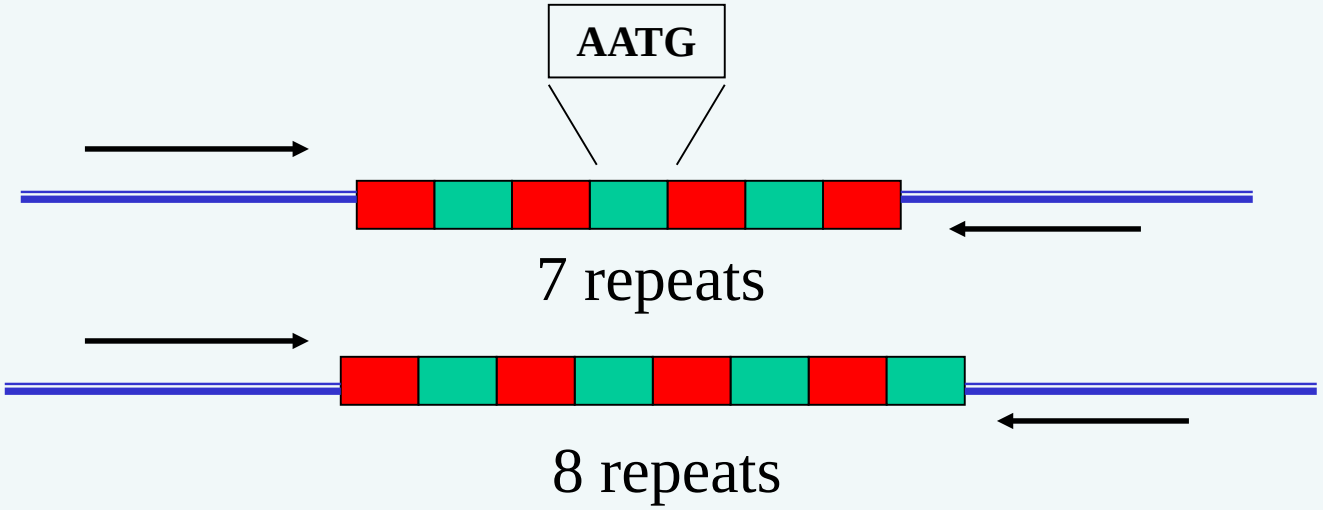
5 repeats: **AATG AATG AATG AATG AATG**

6 repeats: **AATG AATG AATG AATG AATG AATG**

13 CODIS core STR loci



Short Tandem Repeats (STRs)



the repeat region is variable between samples while the flanking regions where PCR primers bind are constant

Homozygote = both alleles are the same length

Heterozygote = alleles differ and can be resolved from one another

Extract and Purify DNA

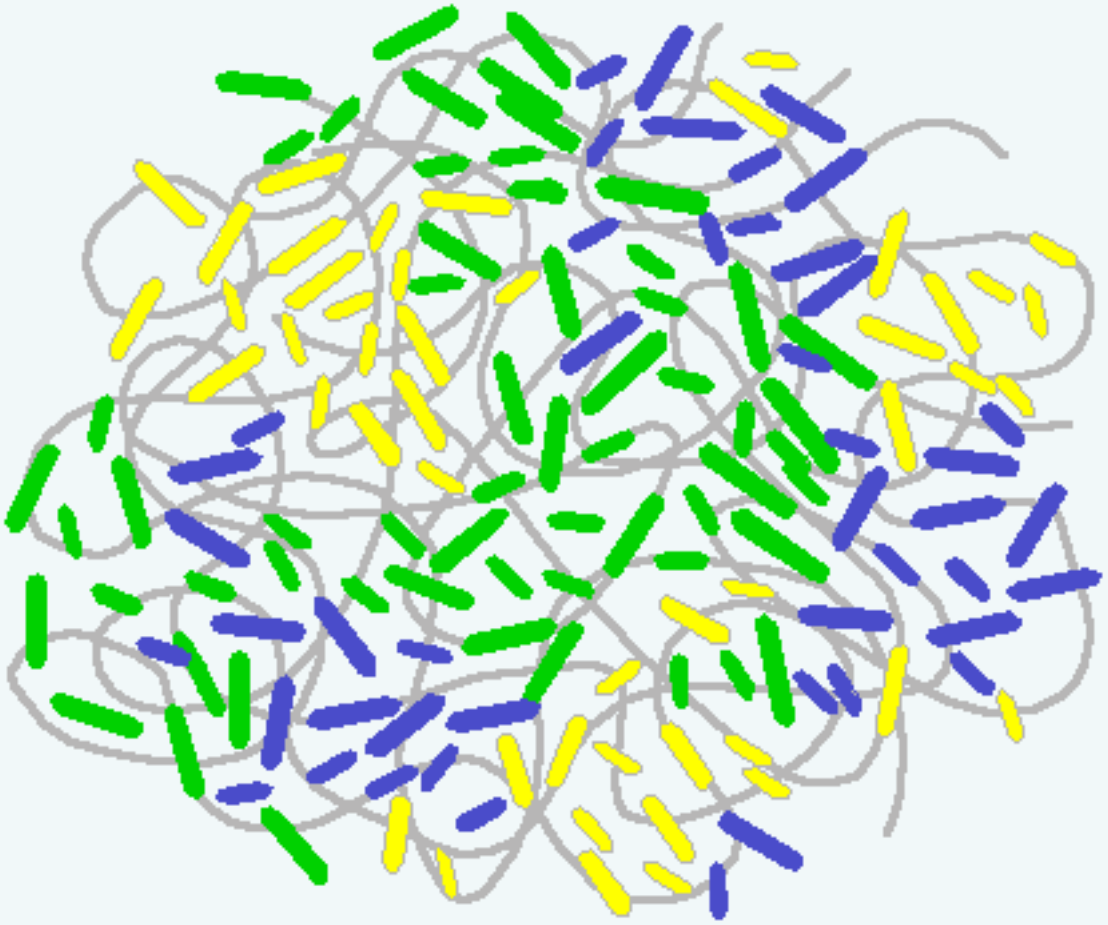
- Add primers and other reagents

PCR Amplification

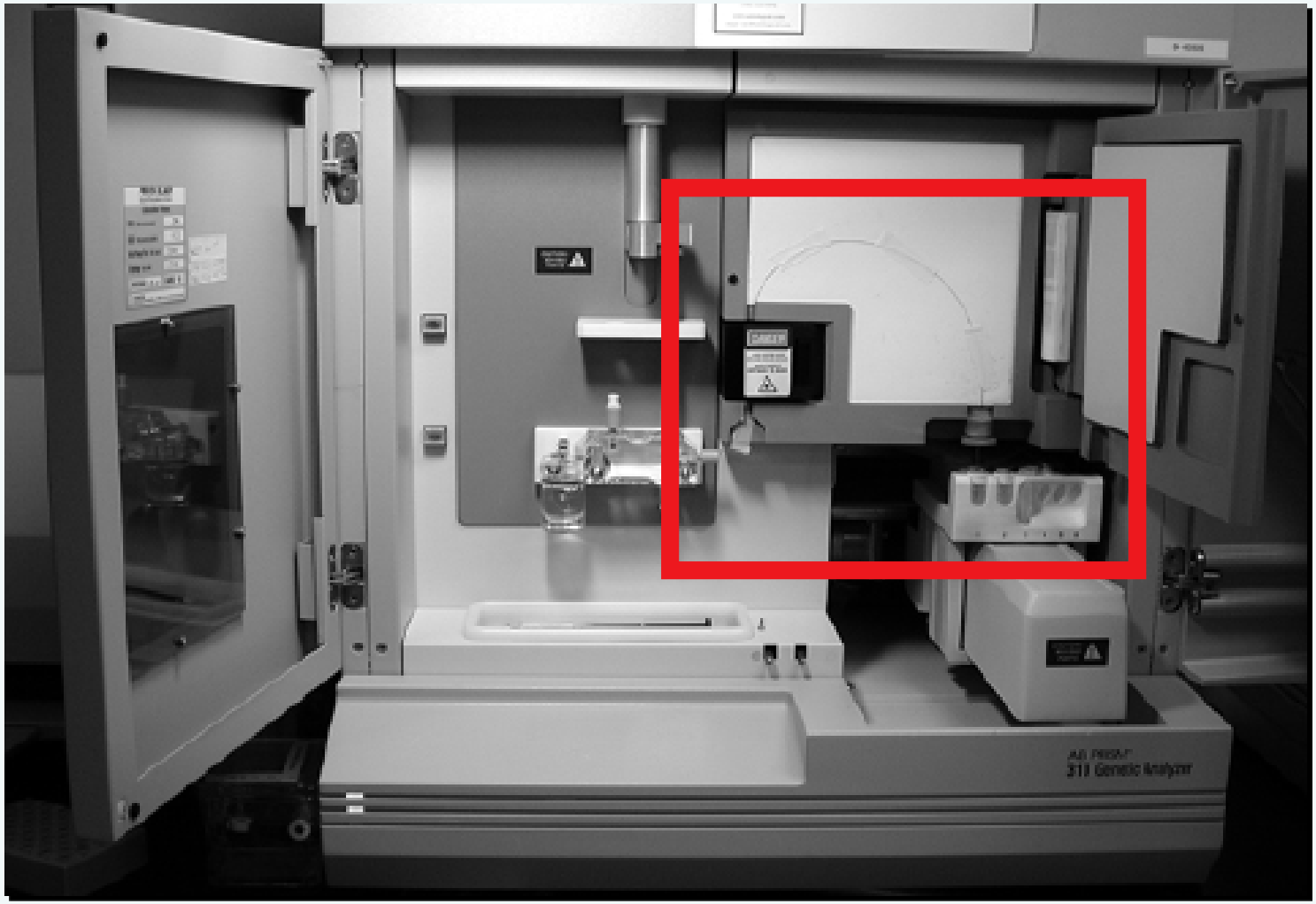
- DNA regions flanked by primers are amplified

Groups of amplified STR products are labeled with different colored dyes (blue, green, yellow)

Profiler Plus: After Amplification

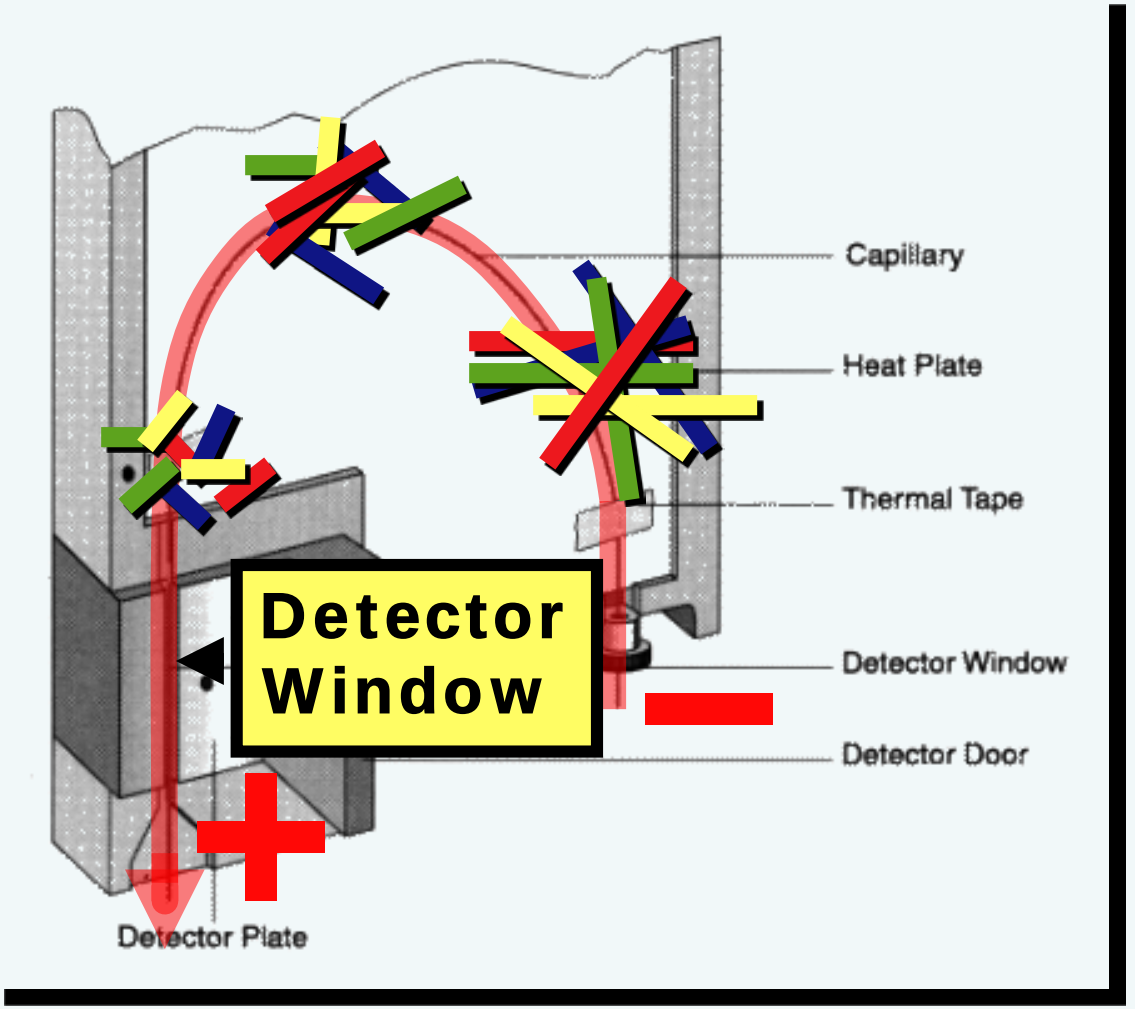


The ABI 310 Genetic Analyzer:

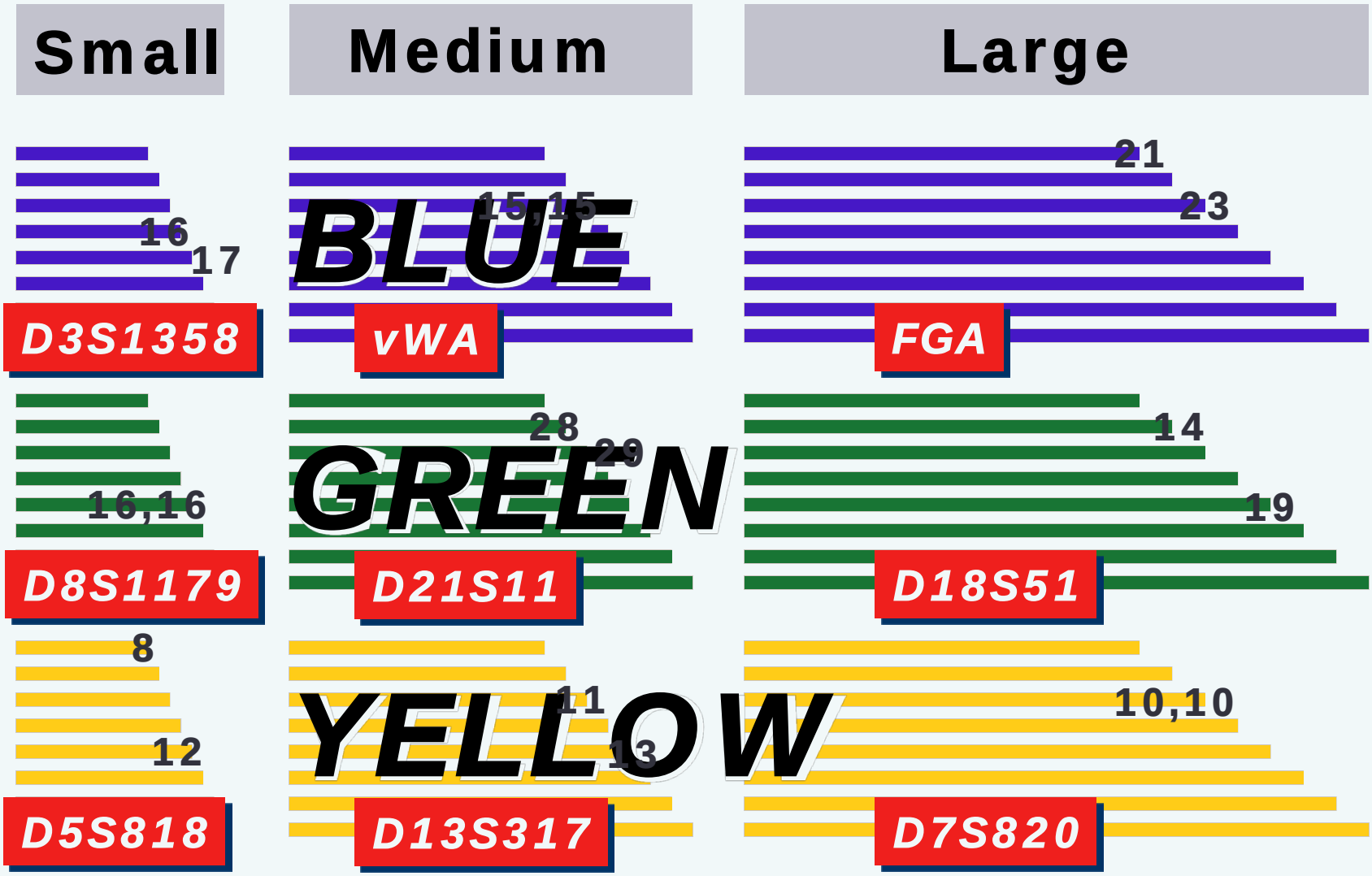


Capillary Electrophoresis

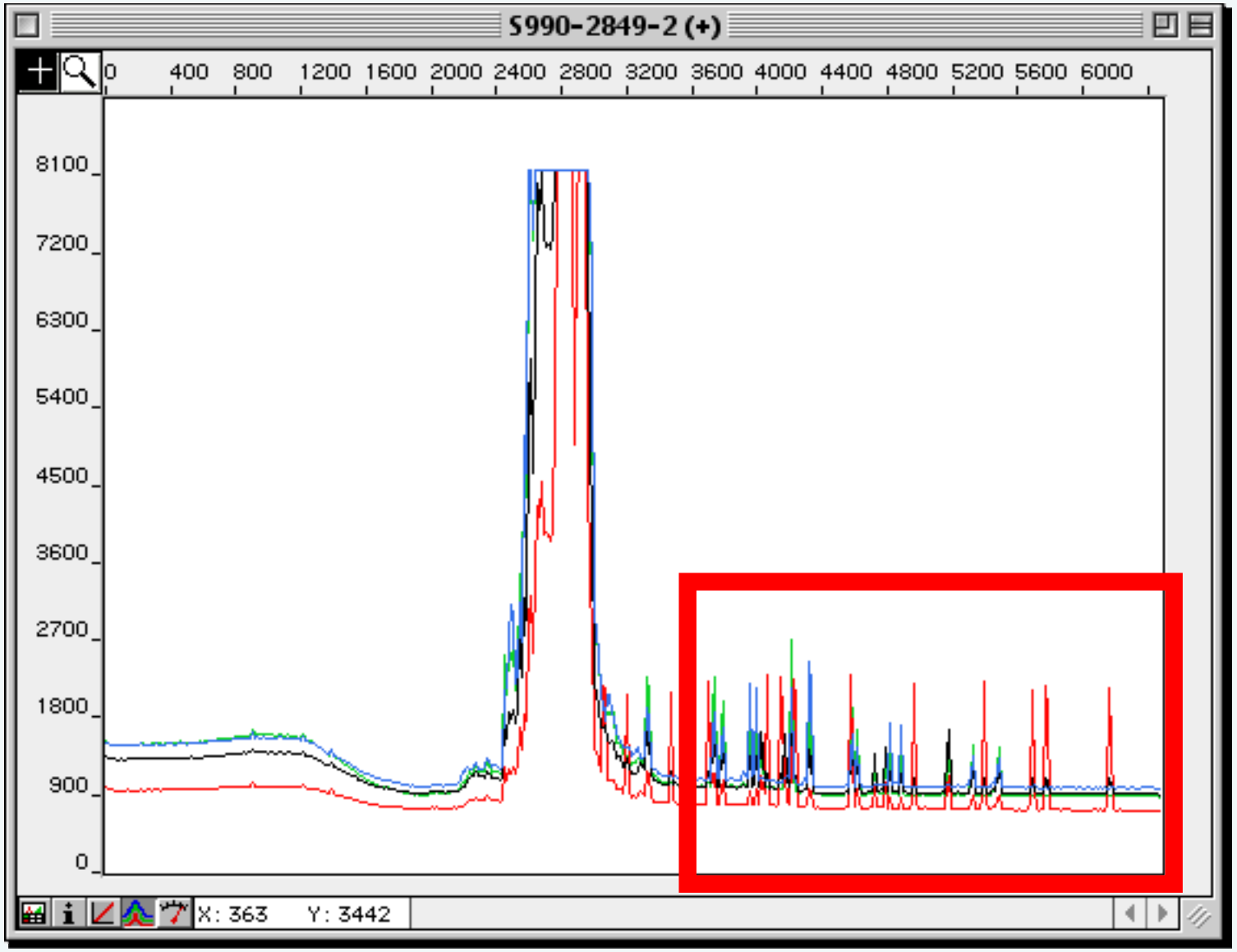
- **Amplified STR DNA injected onto column**
- **Electric current applied**
- **DNA pulled towards the positive electrode**
- **DNA separated out by size:**
  - **Large STRs travel slower**
  - **Small STRs travel faster**
- **Color of STR detected and recorded as it passes the detector**



‘Nested’ STR alleles: Profiler Plus



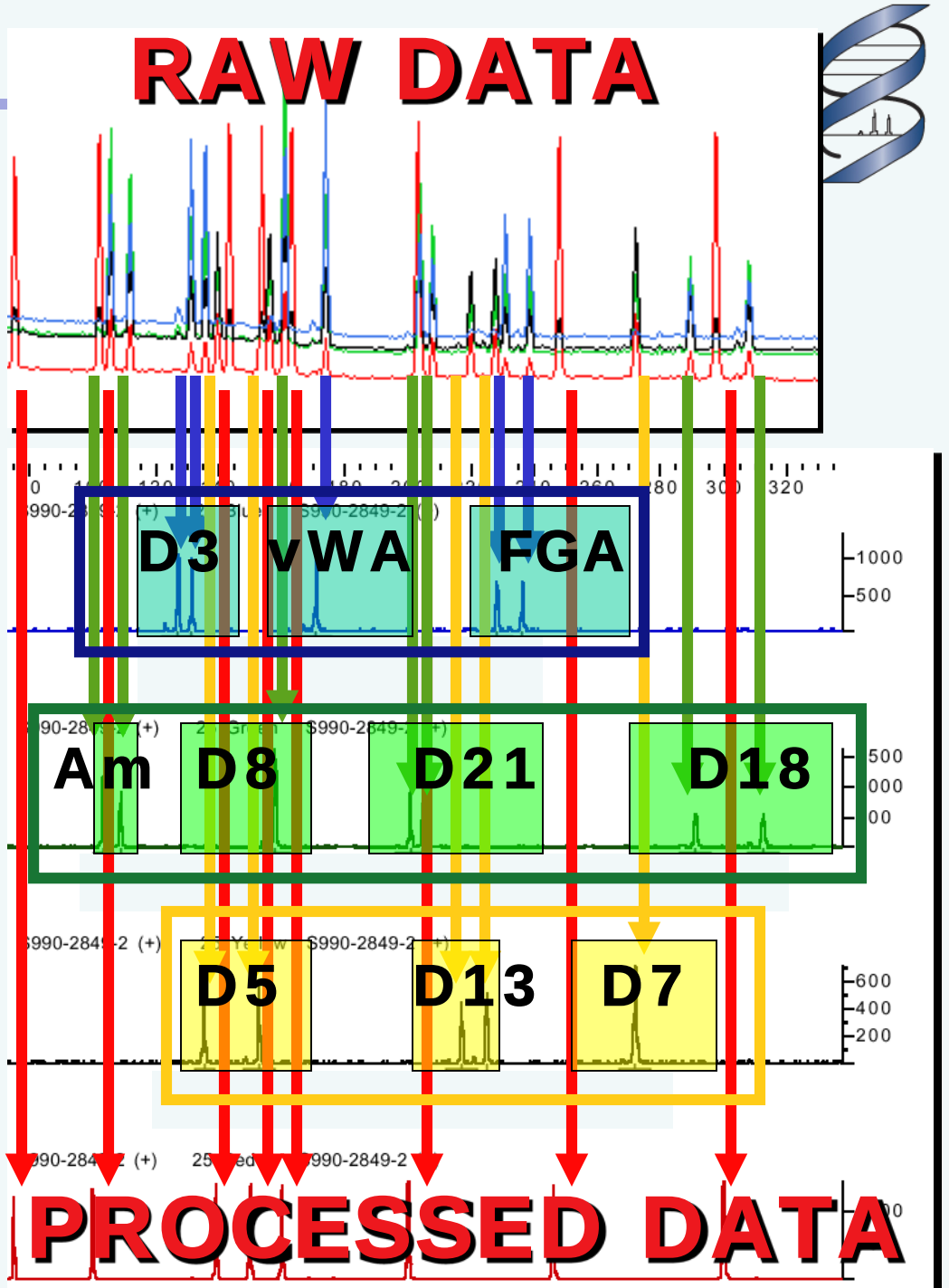
Profiler Plus: Raw data



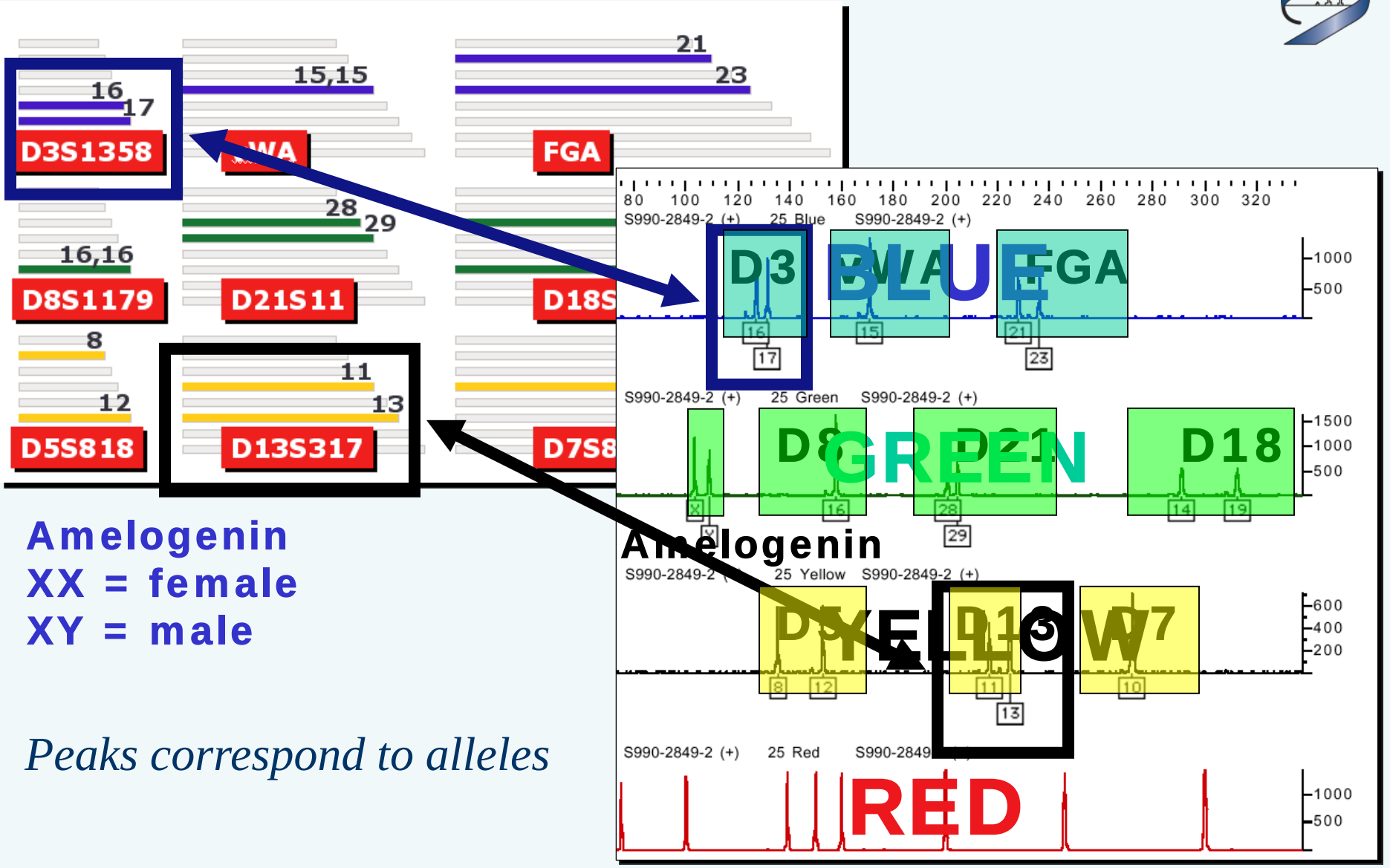
- **GENESCAN divides the raw data into a separate electropherogram for each color:**
  - **Blue**
  - **Green**
  - **Yellow**
  - **Red**

  - **D3: 16, 17**
  - **vWA: 15, 15**
  - **FGA: 21,23**
  - **Amelogenin: X, Y**
  - **D8: 16, 16**
  - **D21: 28, 29**
  - **D18: 14, 19**

  - **D5: 8, 12**
  - **D13: 11, 13**
  - **D7: 10 10**



Reading an electropherogram



Amelogenin
XX = female
XY = male

Peaks correspond to alleles

Statistical estimates: product rule

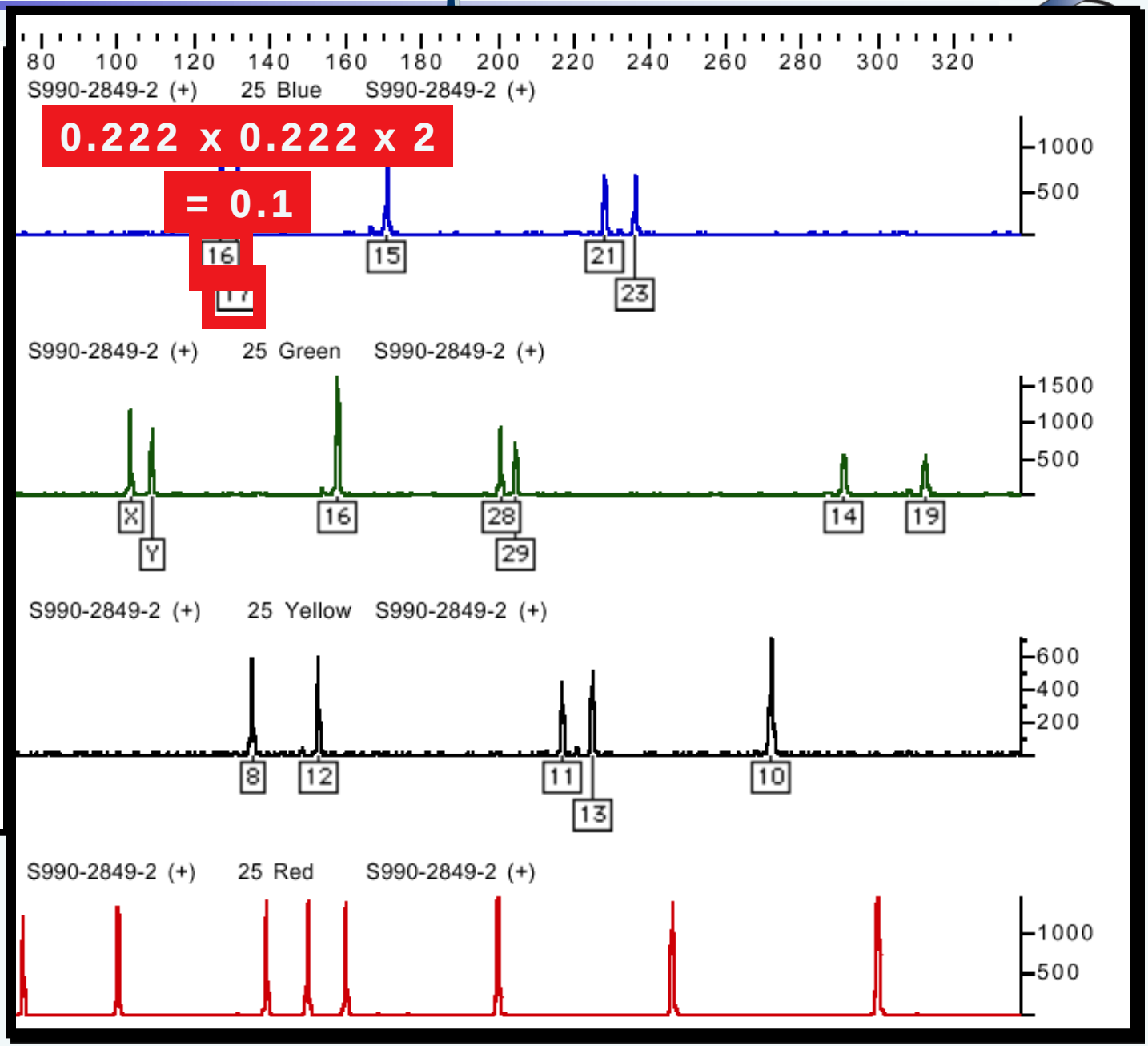
Allele Frequencies

Locus D3S1358
Race Caucasian
(N = 203)

Allele	Frequency
12	0.012
13	0.012
14	0.140
[redacted]	[redacted]
19	0.012

Locus vWA
Race Caucasian
(N = 196)

Allele	Frequency
11	0.012
12	0.012
13	0.012
14	0.102
15	0.082



The product rule

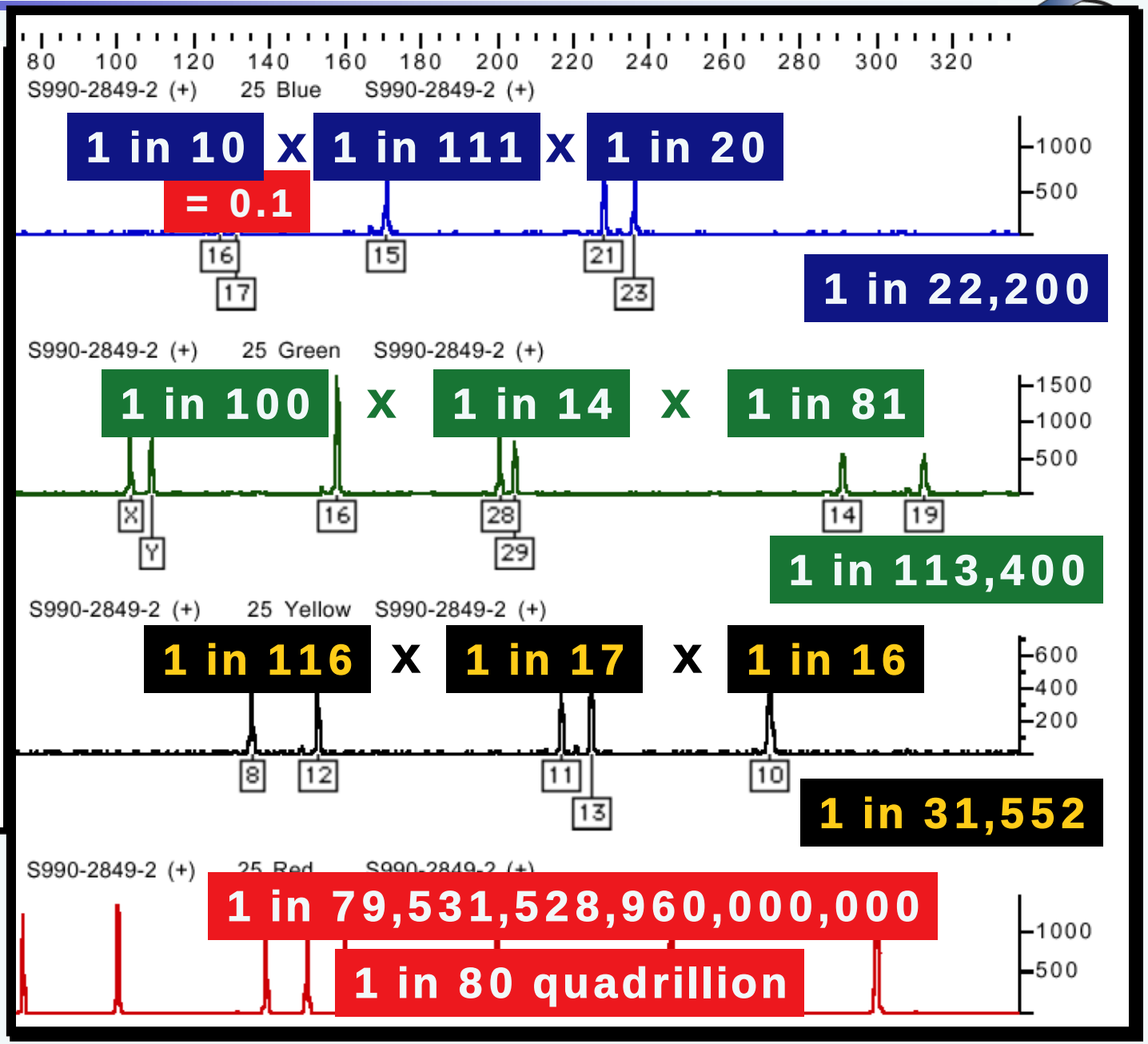
Allele Frequencies

Locus D3S1358
Race Caucasian
(N = 203)

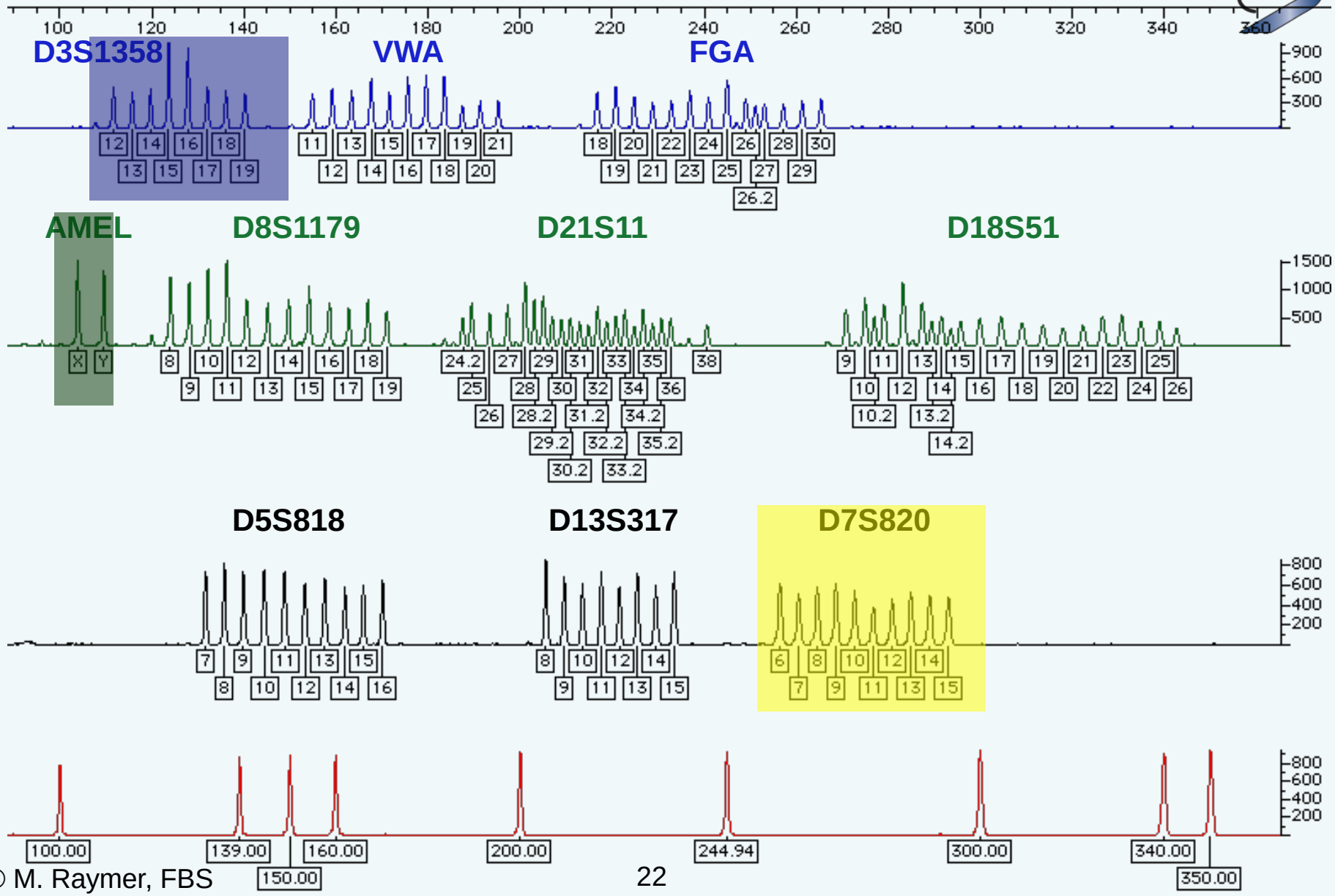
Allele	Frequency
12	0.012
13	0.012
14	0.140
15	0.246
16	0.222
17	0.222
18	0.163
19	0.012

Locus vWA
Race Caucasian
(N = 196)

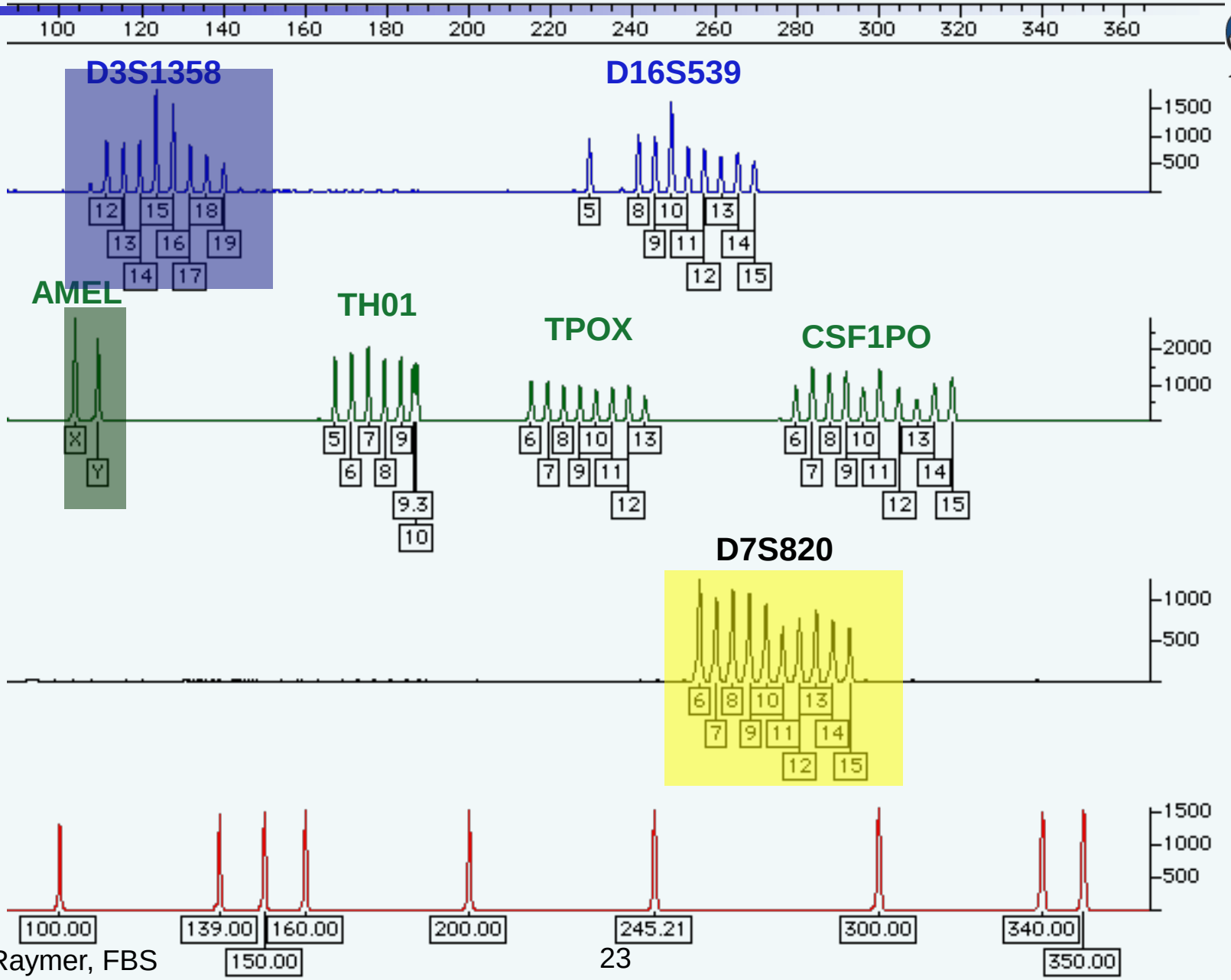
Allele	Frequency
11	0.012
12	0.012
13	0.012
14	0.102
15	0.082



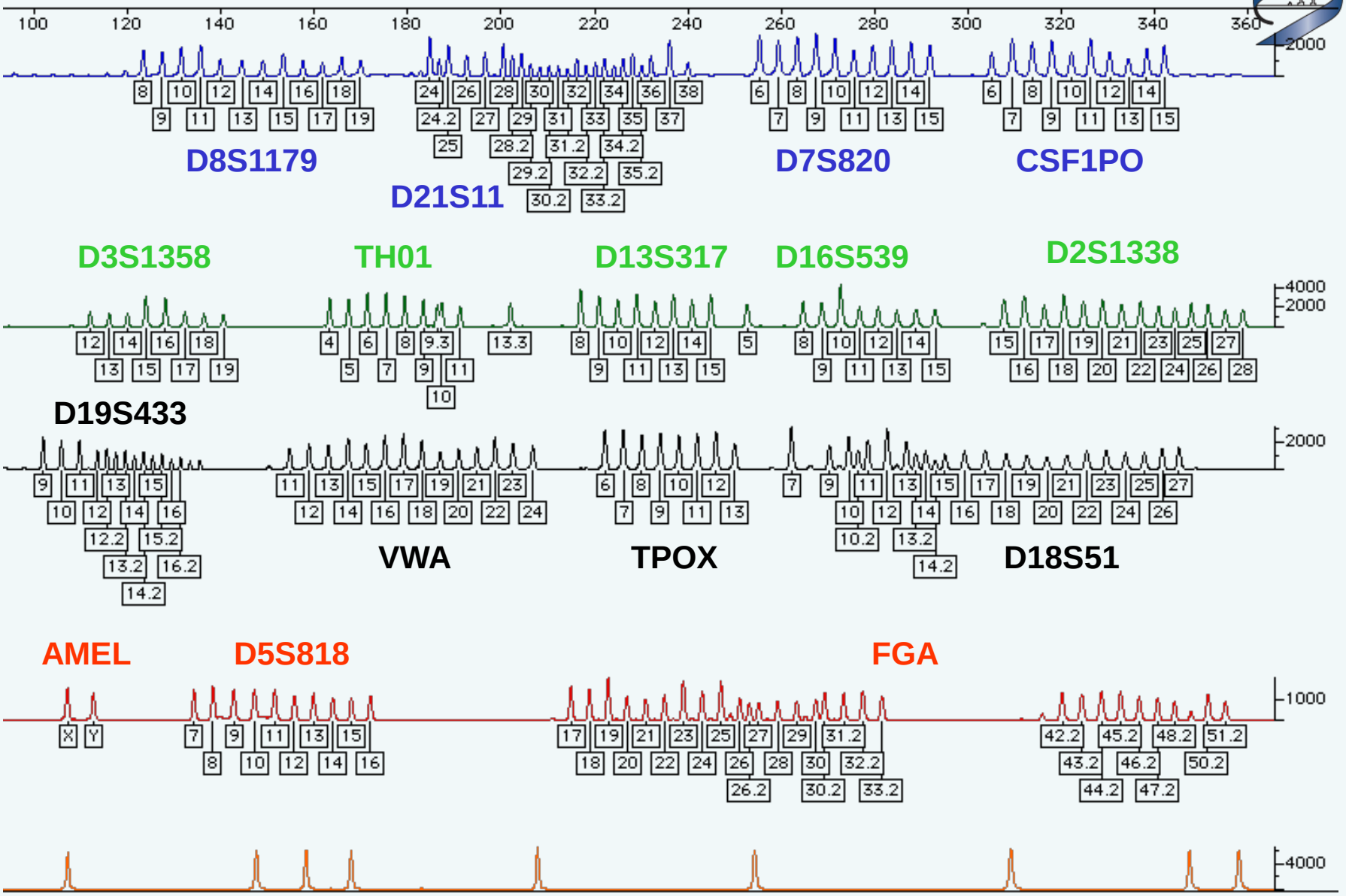
Profiler Plus



Cofiler



Identifiler



Components of a DNA report

- The samples tested
  - Evidence samples (crime scene)
  - Reference samples (defendant, suspect)
- The lab doing the testing
- The test used:
  - Profiler Plus, Cofiler, Identifiler, mtDNA
- The analyst who did the testing
- Results and conclusions:
  - Table of alleles
  - Narrative conclusions

Table of alleles

TABLE OF RESULTS

ITEM	DESCRIPTION	D3S1358	vWA	FGA	AMEL	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820
1	Reference From [redacted]	15,18	18,20	26,28	X,X	10,13	30,31	12,17	12,12	11,12	10,12
2	Reference From [redacted]	15,16	15,16	19,26	X,Y	12,13	31,31	16,21	11,12	11,12	10>11
3	Neck Swab	15,16 (18)	15,16 (18,20)	19,26 (28)	X,Y	12,13 (10)	31,31 (30)	16,21 (12,17)	11,12	11,12	10,11 (12)
4	Chest Swab	15,16 (17<18)	15,16 (18,20)	19,26 (28)	X>Y	12,13 (10)	31,31 (30)	16,21 (12,17)	11,12	11,12	10,11
	Extraction Blanks	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Key: NA = No activity. < = Less than.
() = Weak results for types in parenthesis. X,X = Female DNA.
> = Greater than. X,Y = Male DNA.

- Some labs include more information than others
- Usually includes information about mixed samples
- May also include:
  - Indication of low level results
  - Indication of results not reported
  - Relative amounts of different alleles (in mixed samples)
- No standard format

Narrative conclusions

CONCLUSIONS

1. The neck and chest swabs (items 3 and 4) have an elevated level of amylase 1 present in the extracts. These results strongly indicate saliva on the swabs.

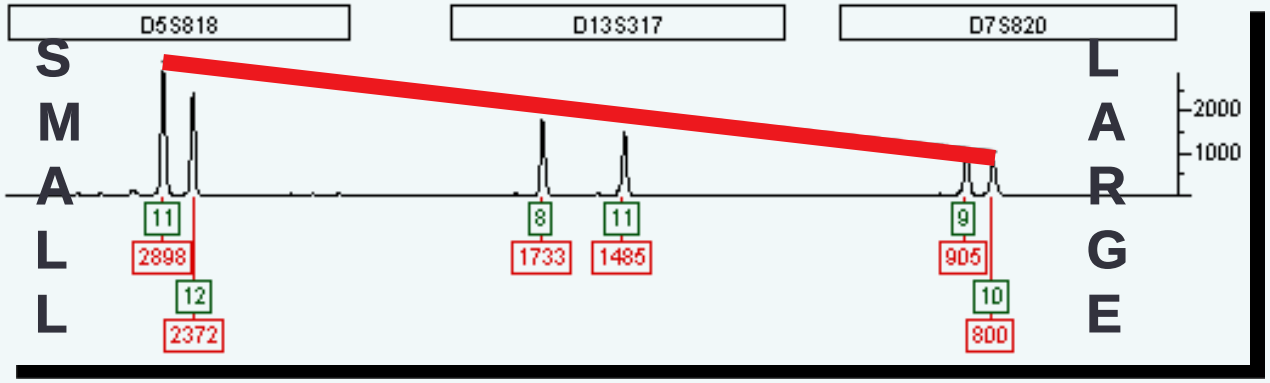
2. The genetic marker results in the DNA extracted from the neck and chest swabs (items 3 and 4) are a mixture of at least two persons. The results indicate a major (or stronger donor) and a secondary (or weaker donor). [redacted] is, in my opinion, the major DNA donor on items 3 and 4. Due to the presence of weak typing results at some loci, it is possible that minor components of the mixture have dropped out in the larger loci. As a result, [redacted] cannot be excluded as a contributor to the secondary DNA profile obtained from the neck and chest swabs (items 3 and 4). In addition, a weak amount of D3S1358 type 17 was detected on item 4 which could not have originated from [redacted] or [redacted]. It is unclear as to whether this allele is artifactual in origin or from another donor.

- Indicates which samples match
- Includes a statistical estimate
- Identifies samples as mixed
- May include an 'identity statement' i.e., samples are from the same source to a scientific degree of certainty (FBI)
- May allude to problems (e.g. interpretative ambiguity, contamination)

Sources of ambiguity in STR interpretation

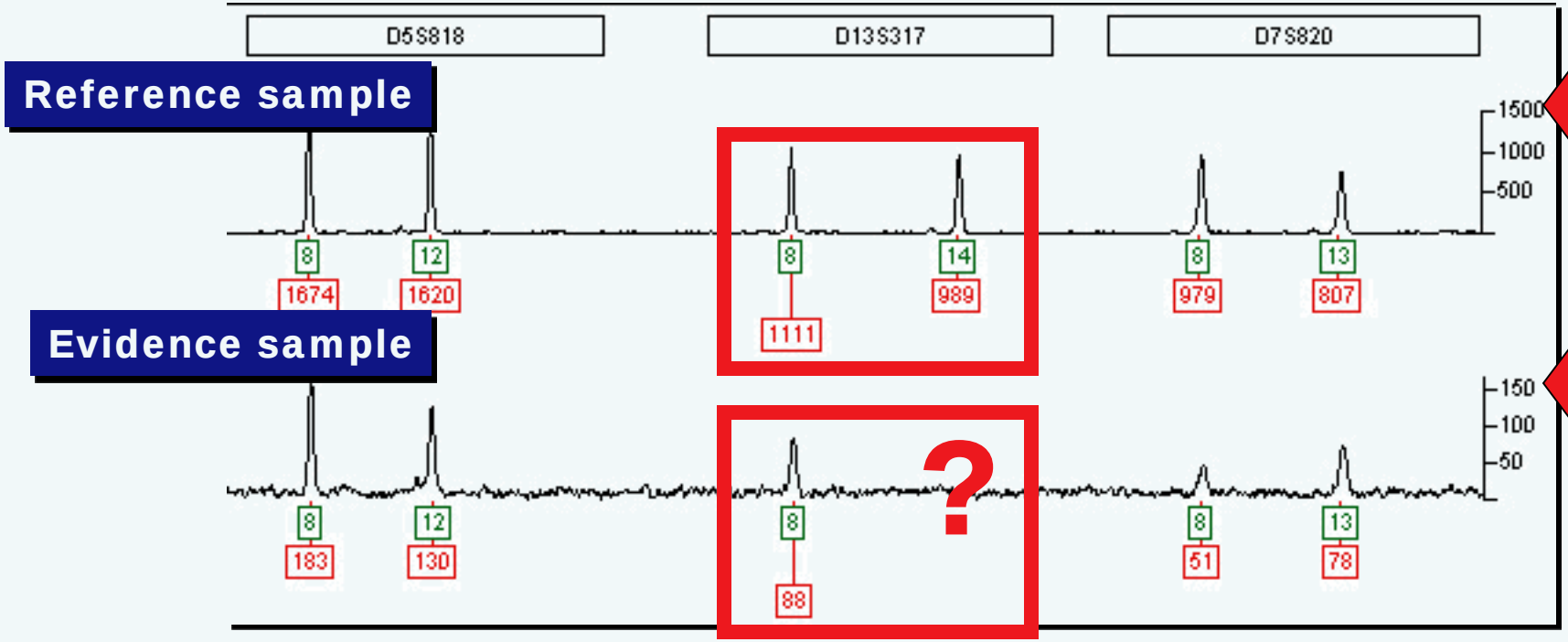
- Degradation
- Allelic dropout
- False peaks
- Mixtures
- Accounting for relatives
- Threshold issues -- marginal samples

Degradation



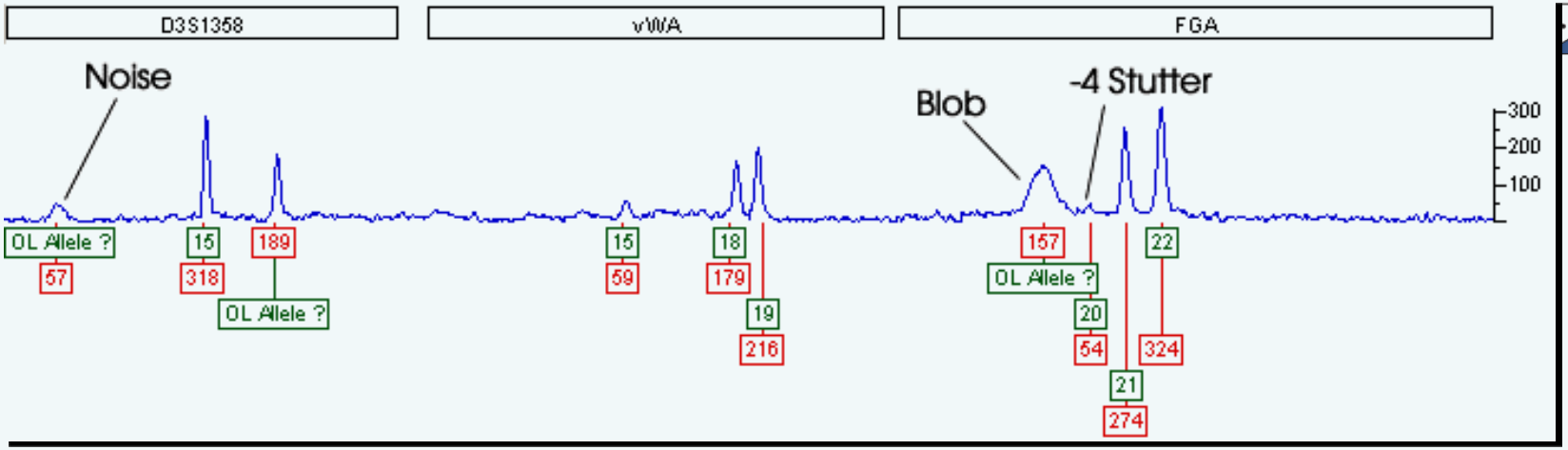
- When biological samples are exposed to adverse environmental conditions, they can become degraded
  - Warm, moist, sunlight, time
- Degradation breaks the DNA at random
- Larger amplified regions are affected first
- Classic ‘ski-slope’ electropherogram
- Peaks on the right lower than peaks on the left

Allelic Dropout



- Peaks in evidence samples all very low
  - Mostly below 150 rfu
- Peaks in reference sample much higher
  - All well above 800 rfu
- At D13S817:
  - Reference sample: 8, 14
  - Evidence sample: 8, 8
- 14 allele has dropped out -- or has it?
- Tend to see with ‘marginal samples’

False peaks & machine problems



- False peaks:
  - Contamination
  - Dye blob
  - Electrical spikes
  - Pull-up
- Machine problems:
  - Noise
  - Baseline instability
  - Injection failures

Summary Sheet

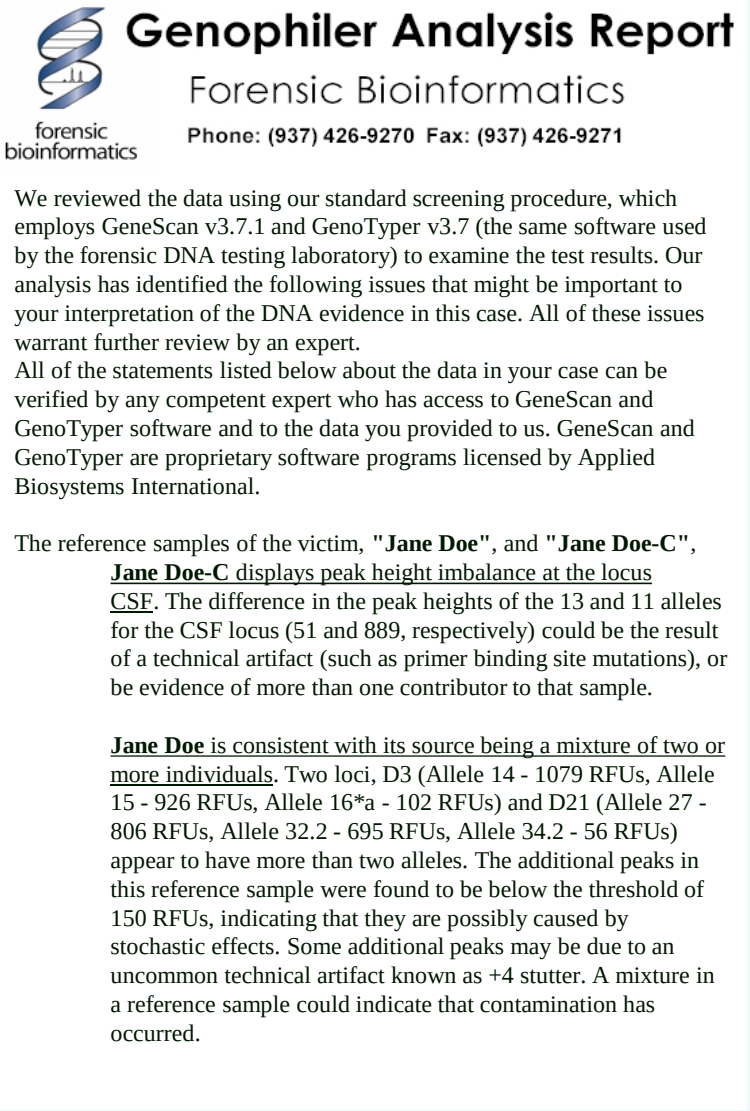
Sample Name	GT Results	D3	vWA	FGA	D16
Jane Doe Victim PP	GT Graph	14 (1079) 15 (926) 16*[a] (102)	16 (755) 19 (664)	21 (1284)	_
Jane Doe-C	GT Graph	14 (858) 15 (794)		_	13 (1034)
John Doe Defendant PP	GT Graph	13 (159) 14*[b] (2820) 19 (58)		22 (881) 24 (765)	_
John Doe-C Defendant CO	GT Graph	14 (2168)	_	_	9 (805) 12 (675)
Dress - Blood Stain Evidence1 PP	GT Graph	14 (851) 15 (759)	16 (7		_
Dress - Blood Stain-C Evidence1 CO	GT Graph	14 (422) 15 (380)	_		13 (716)
Knife - Blood Stain Evidence2 PP	GT Graph	14 (1161) 15 (1084)	15 (6		_
Knife - Blood Stain-C Evidence2 CO	GT Graph	14 (1003) 15 (895)	_		13 (1219)
Towel Evidence3 PP	GT Graph	14 (1032) 15 (1089)	16 (822) 19 (690)	21 (1071)	_
Towel-C Evidence3 CO	GT Graph	14 (669) 15 (640)	_	_	13 (774)
Positive Control Run 1 POSITIVE PP run1	GT Graph	14 (1399) 15 (1282)	16 (104) 17 (1358) 18 (1084)	23 (1143) 24 (911)	_
Positive Control Run 1-C POSITIVE CO run1	GT Graph	14 (1032) 15 (858)	_	_	11 (1076) 12 (889)
Positive Control Run 2 pos_cont_co run2	GT Graph	14 (989) 15 (871)	_	_	11 (804) 12 (855)
Positive Control Run 3 pos_cont_pp run3	GT Graph	14 (1262) 15 (1129)	16 (94) 17 (1253) 18 (1243)	23 (901) 24 (890)	_
Reagent Blank Run 1 BLANK run1	GT Graph	_	_	_	_
Reagent Blank Run 2 blank run2	GT Graph	_	_	_	_

The * indicates that this peak may be involved in pullup…

Analysis Report

A locus by locus description of issues that may warrant further review by an expert, including:

- Peak height imbalance
- Presence of a mixture
- Possible degradation
- Possible pullup
- Inconsistent results from multiple runs
- Problems with control runs and reagent blanks



forensic bioinformatics

Genophiler Analysis Report

Forensic Bioinformatics

Phone: (937) 426-9270 Fax: (937) 426-9271

We reviewed the data using our standard screening procedure, which employs GeneScan v3.7.1 and GenoTyper v3.7 (the same software used by the forensic DNA testing laboratory) to examine the test results. Our analysis has identified the following issues that might be important to your interpretation of the DNA evidence in this case. All of these issues warrant further review by an expert.
All of the statements listed below about the data in your case can be verified by any competent expert who has access to GeneScan and GenoTyper software and to the data you provided to us. GeneScan and GenoTyper are proprietary software programs licensed by Applied Biosystems International.

The reference samples of the victim, "**Jane Doe**", and "**Jane Doe-C**",

Jane Doe-C displays peak height imbalance at the locus CSF. The difference in the peak heights of the 13 and 11 alleles for the CSF locus (51 and 889, respectively) could be the result of a technical artifact (such as primer binding site mutations), or be evidence of more than one contributor to that sample.

Jane Doe is consistent with its source being a mixture of two or more individuals. Two loci, D3 (Allele 14 - 1079 RFUs, Allele 15 - 926 RFUs, Allele 16*a - 102 RFUs) and D21 (Allele 27 - 806 RFUs, Allele 32.2 - 695 RFUs, Allele 34.2 - 56 RFUs) appear to have more than two alleles. The additional peaks in this reference sample were found to be below the threshold of 150 RFUs, indicating that they are possibly caused by stochastic effects. Some additional peaks may be due to an uncommon technical artifact known as +4 stutter. A mixture in a reference sample could indicate that contamination has occurred.

What can be done to make DNA testing more objective?

- Distinguish between signal and noise
  - Deducing the number of contributors to mixtures
  - Accounting for relatives

Where do peak height thresholds come from (originally)?

- Applied Biosystems validation study of 1998
- Wallin et al., 1998, "TWGDAM validation of the AmpFlSTR blue PCR Amplification kit for forensic casework analysis." *JFS* 43:854-870.

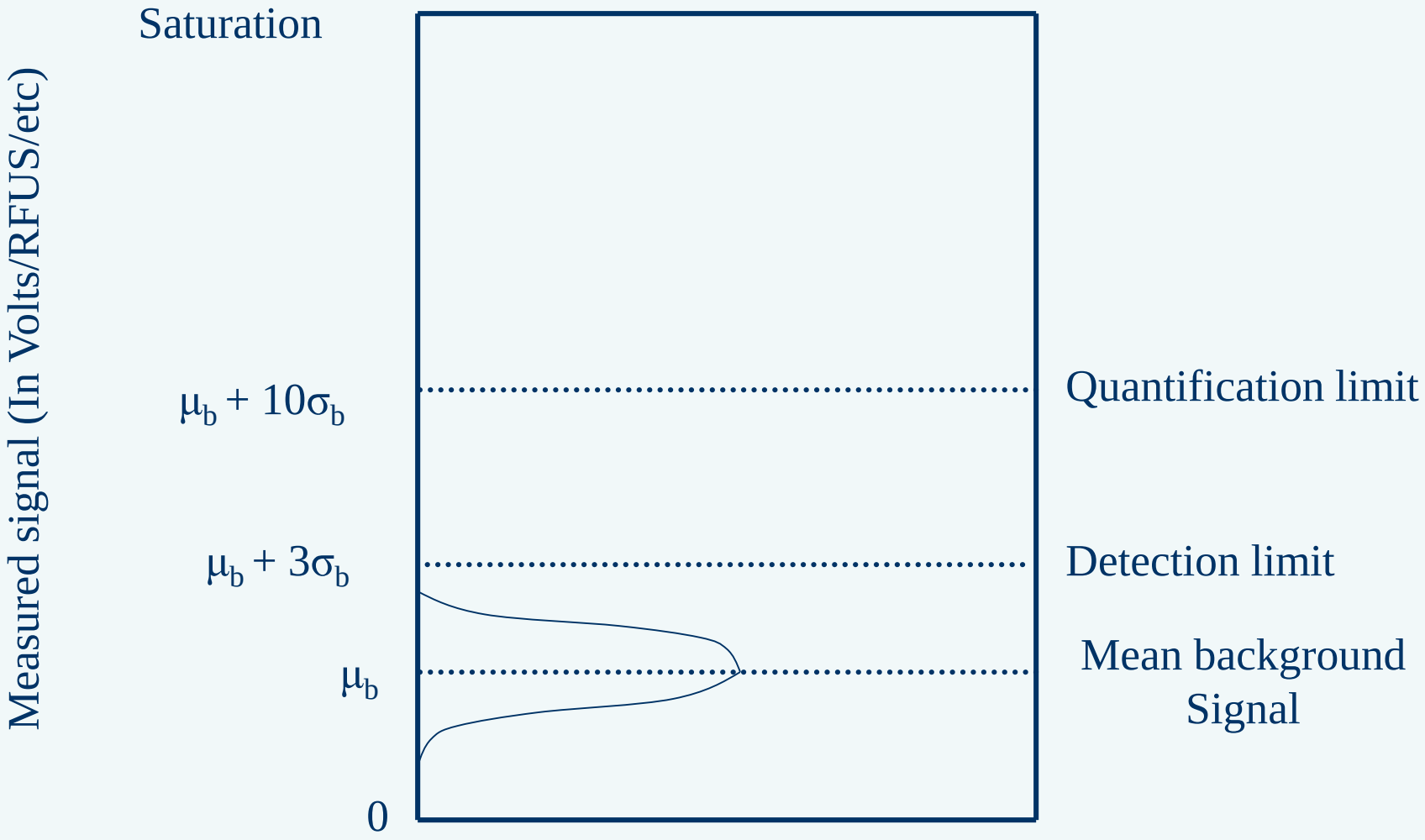
Where do peak height thresholds come from (originally)?

PCR products were examined on both the 377 DNA Sequencer and the 310 Genetic Analyzer. The results of 0.25 to 1.0 ng were clearly typable with peak heights of approximately 150 RFU and greater (data not shown). At 0.125 ng and less, the peak heights in both samples were not significantly above the background (< 150 RFU) or were undetectable. At 0.0313 ng specifically, peaks were extremely low or undetectable, and thus, DNA quantities as low as approximately 35 pg did not produce a typable result. Based on these results, we employed a peak height threshold of 150 RFU, below which peaks were interpreted with caution. Laboratories should determine a minimum peak height threshold for their instruments using high quality, single source genomic DNA samples which provides them with the desired sensitivity while not allowing for detection of low copy DNA. This is particularly important as the overall sensitivity of the assay may vary between laboratories.

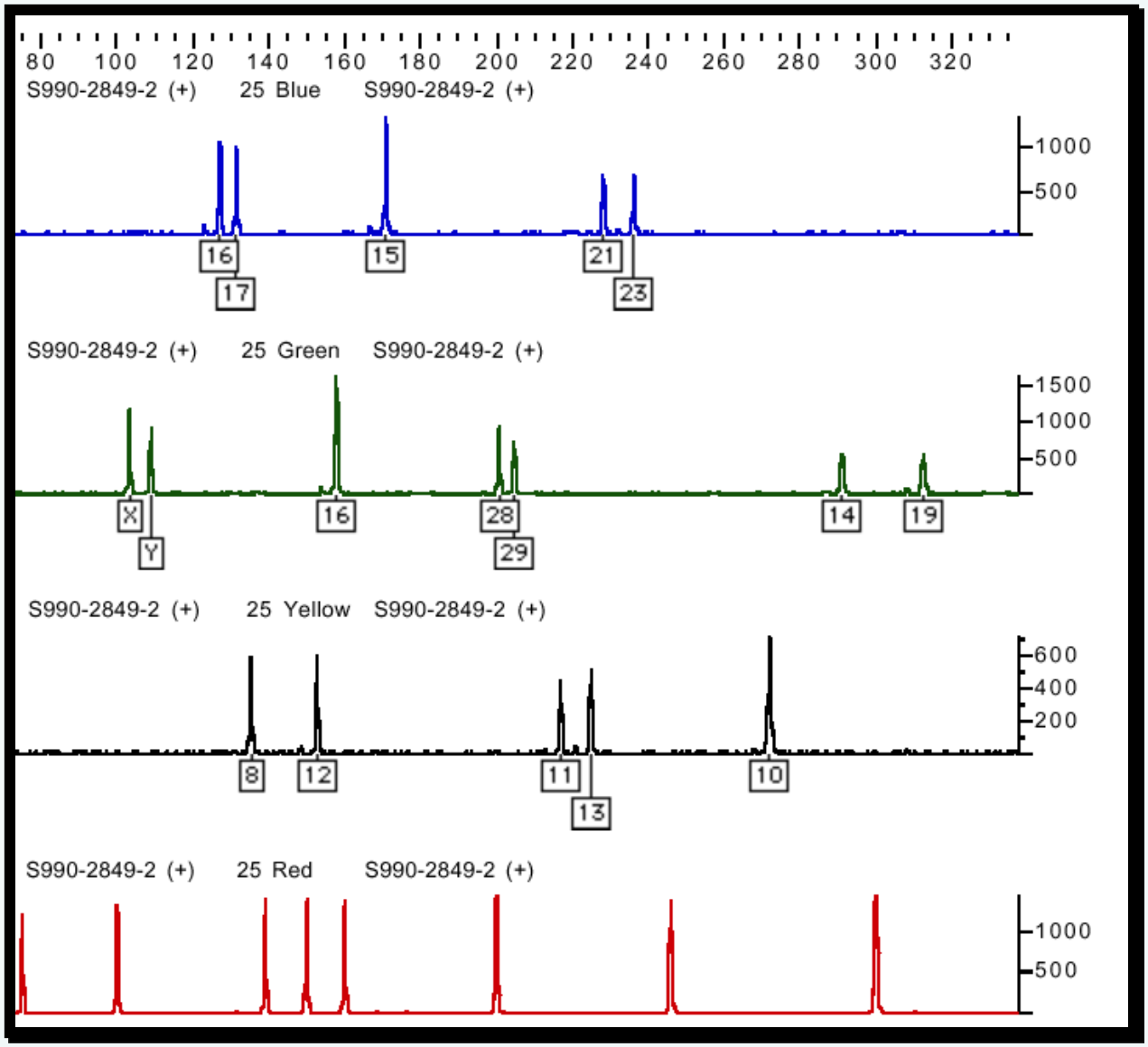
Where do peak height thresholds come from?

- "Conservative" thresholds established during validation studies
- Eliminate noise (even at the cost of eliminating signal)
- Can arbitrarily remove legitimate signal
- Contributions to noise vary over time (e.g. polymer and capillary age/condition)
  - Analytical chemists use LOD and LOQ

Signal Measure



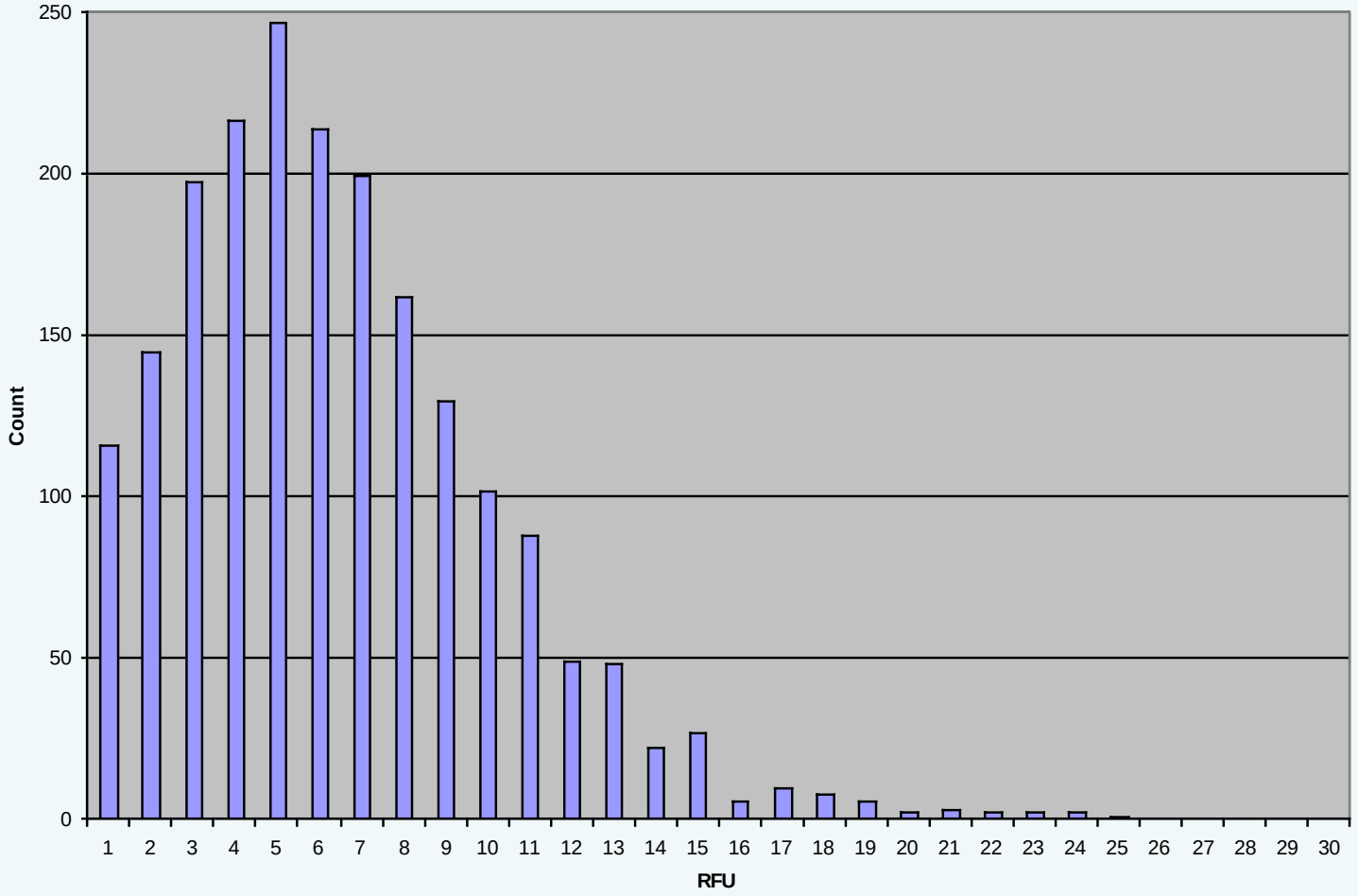
Opportunities to measure baseline



Control samples

- Negative controls: 5,932 data collection points (DCPs) per run (σ = 131 DCPs)
- Reagent blanks: 5,946 DCPs per run (σ = 87 DCPs)
- Positive controls: 2,415 DCP per run (σ = 198 DCPs)
- DCP regions corresponding to size standards and 9947A peaks (plus and minus 55 DCPs to account for stutter in positive controls) were masked in all colors

RFU levels at all non-masked data collection points



Variation in baseline noise levels

Positive Control		μ_b	σ_b	$\mu_b + 3\sigma_b$	$\mu_b + 10\sigma_b$
	Maximum	6.7	6.9	27.4	75.7
	Average	5.0	3.7	16.1	42.0
	Minimum	3.7	2.4	10.9	27.7

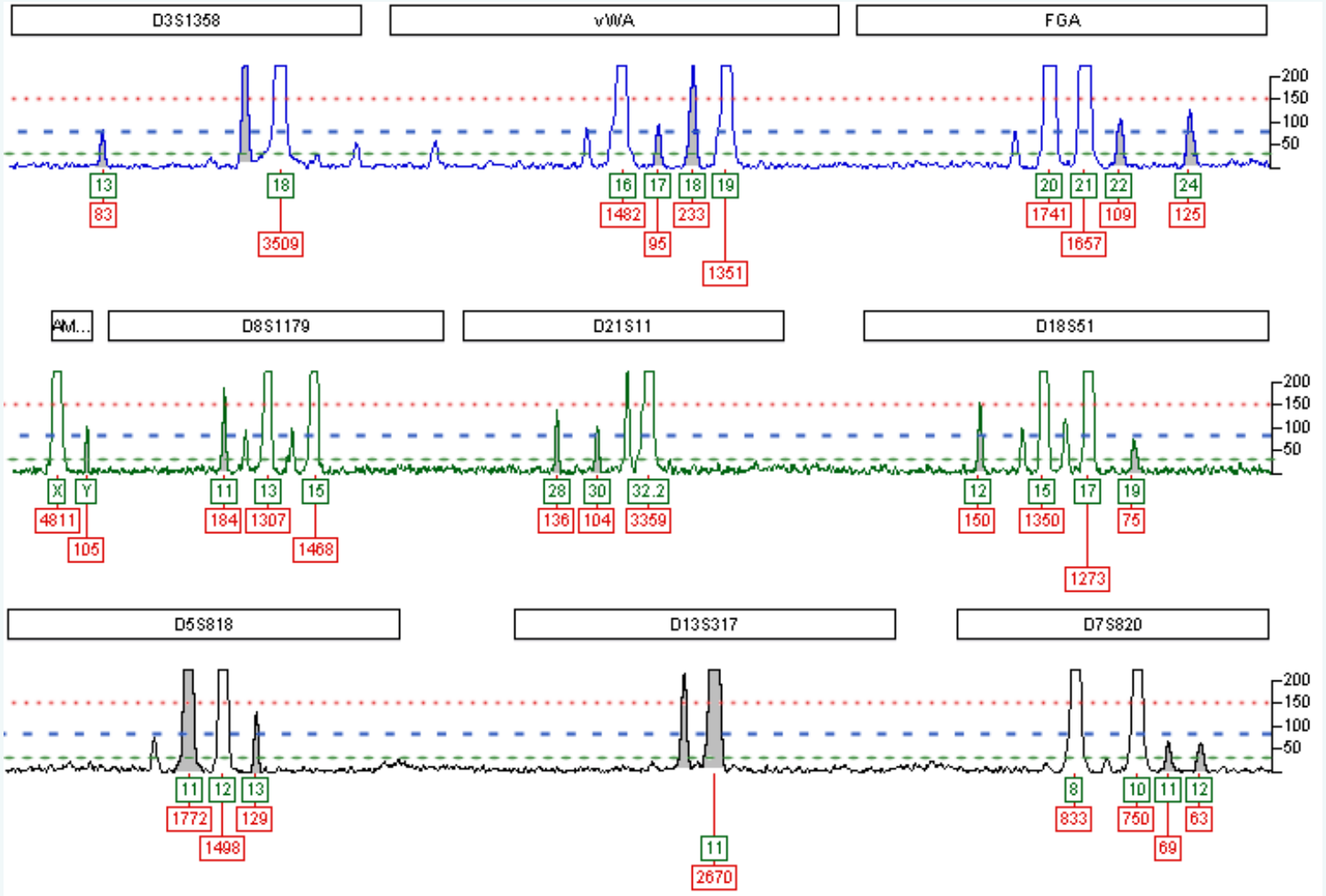
Negative Control		μ_b	σ_b	$\mu_b + 3\sigma_b$	$\mu_b + 10\sigma_b$
	Maximum	13.4	13.2	53.0	145.4
	Average	5.4	3.9	17.1	44.4
	Minimum	4.0	2.6	11.8	30.0

Reagent Blank		μ_b	σ_b	$\mu_b + 3\sigma_b$	$\mu_b + 10\sigma_b$
	Maximum	6.5	11.0	39.5	116.5
	Average	5.3	4.0	17.3	45.3
	Minimum	4.0	2.6	11.8	30.0

All three controls averaged		μ_b	σ_b	$\mu_b + 3\sigma_b$	$\mu_b + 10\sigma_b$
	Maximum	7.1	7.3	29.0	80.1
	Average	5.2	3.9	16.9	44.2
	Minimum	3.9	2.5	11.4	28.9

Average (μ_b) and standard deviation (σ_b) values with corresponding LODs and LOQs from positive, negative and reagent blank controls in 50 different runs. BatchExtract: ftp://ftp.ncbi.nlm.nih.gov/pub/forensics/

Lines in the sand: a 2-person mix?



Two reference samples in a 1:10 ratio (male:female). Three different thresholds are shown: 150 RFU (red); LOQ at 77 RFU (blue); and LOD at 29 RFU (green).

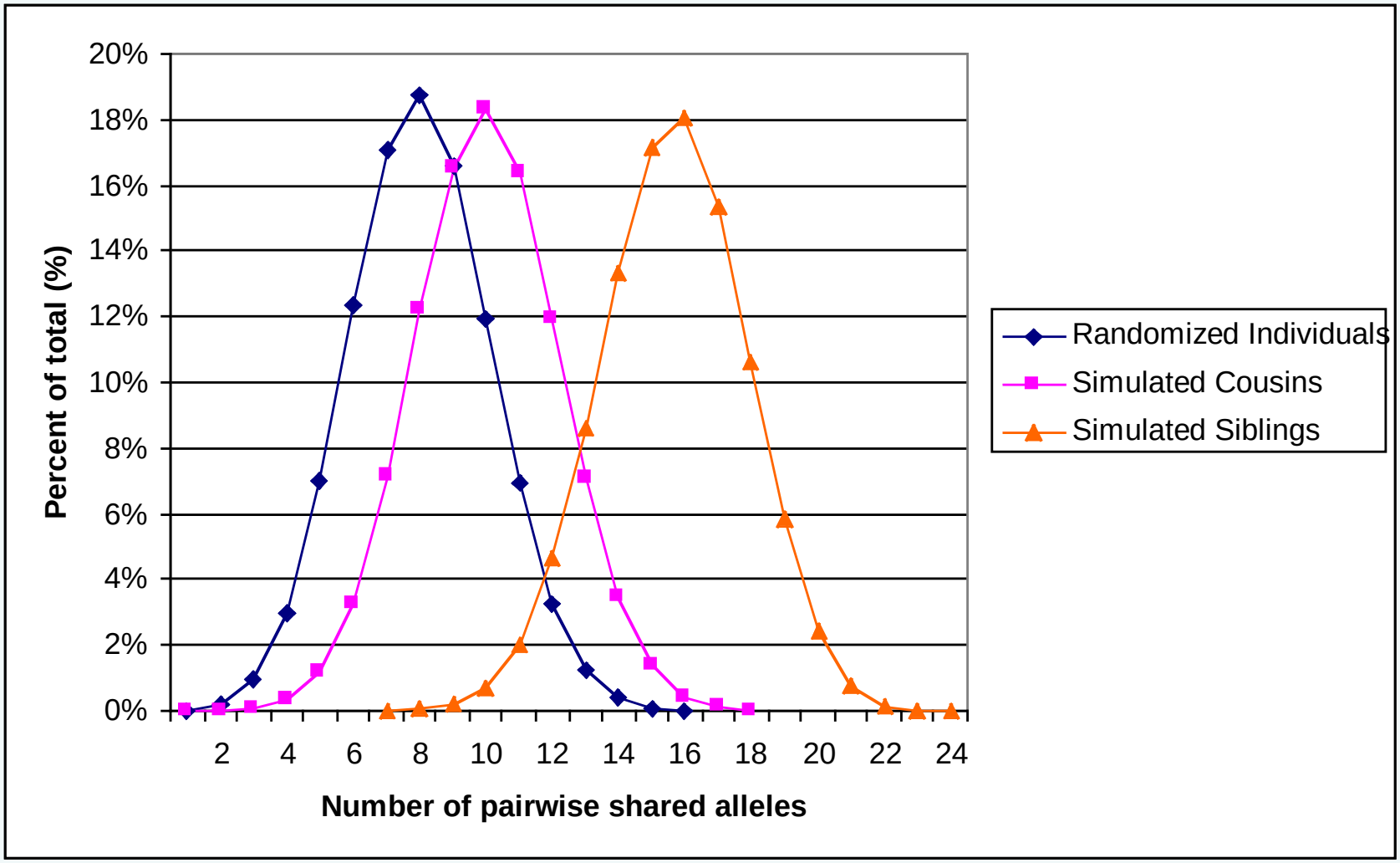
Familial searching

- Database search yields a close but imperfect DNA match
- Can suggest a relative is the true perpetrator
- Great Britain performs them routinely
- Reluctance to perform them in US since 1992 NRC report
- Current CODIS software cannot perform effective searches

Three approaches to familial searches

- Search for rare alleles (inefficient)
- Count matching alleles (arbitrary)
- Likelihood ratios with kinship analyses

Pair-wise similarity distributions



Is the true DNA match a relative or a random individual?

- Given a closely matching profile, who is more likely to match, a relative or a randomly chosen, unrelated individual?

- Use a $$LR = \frac{P(E \mid relative)}{P(E \mid random)}$$

Is the true DNA match a relative or a random individual?

- What is the likelihood that a relative of a single initial suspect would match the evidence sample perfectly?
- What is the likelihood that a single randomly chosen, unrelated individual would match the evidence sample perfectly?

$$LR = \frac{P(E \mid relative)}{P(E \mid random)}$$

Probabilities of siblings matching at 0, 1 or 2 alleles

$$P(E \mid sib) = \begin{cases} \dfrac{P_a \cdot P_b \cdot HF}{4}, & if \quad shared = 0 \\ \dfrac{P_b + P_a \cdot P_b \cdot HF}{4}, & if \quad shared = 1 \\ \dfrac{1 + P_a + P_b + P_a \cdot P_b \cdot HF}{4}, & if \quad shared = 2 \end{cases}$$

HF = 1 for homozygous loci and 2 for heterozygous loci; P_a is the frequency of the allele shared by the evidence sample and the individual in a database.

Probabilities of parent/child matching at 0, 1 or 2 alleles

$$P(E \mid parent/child) = \begin{cases} 0, & if \quad shared = 0 \\ \frac{P_b}{2}, & if \quad shared = 1 \\ \frac{P_a + P_b}{2}, & if \quad shared = 2 \end{cases}$$

HF = 1 for homozygous loci and 2 for heterozygous loci; P_a is the frequency of the allele shared by the evidence sample and the individual in a database.

Other familial relationships

Cousins:

$$P(E \mid cousins) = \begin{cases} \dfrac{6 \cdot P_a \cdot P_b \cdot HF}{8}, & if \quad shared = 0 \\ \dfrac{P_b + 6 \cdot P_a \cdot P_b \cdot HF}{8}, & if \quad shared = 1 \\ \dfrac{P_a + P_b + 6 \cdot P_a \cdot P_b \cdot HF}{8}, & if \quad shared = 2 \end{cases}$$

$$P(E \mid GG / AUNN / HS) = \begin{cases} \dfrac{2 \cdot P_a \cdot P_b \cdot HF}{4}, & if \quad shared = 0 \\ \dfrac{P_b + 2 \cdot P_a \cdot P_b \cdot HF}{4}, & if \quad shared = 1 \\ \dfrac{P_a + P_b + 2 \cdot P_a \cdot P_b \cdot HF}{4}, & if \quad shared = 2 \end{cases}$$

HF = 1 for homozygous loci and 2 for heterozygous loci; P_a is the frequency of the allele shared by the evidence sample and the individual in a database.

Familial search experiment

- Randomly pick related pair or unrelated pair from a synthetic database
- Choose one profile to be evidence and one profile to be initial suspect
- Test hypothesis:
  - H_0: A relative is the source of the evidence
  - H_A: An unrelated person is the source of the evidence

Paoletti, D., Doom, T., Raymer, M. and Krane, D. 2006. Assessing the implications for close relatives in the event of similar but non-matching DNA profiles. *Jurimetrics*, 46:161-175.

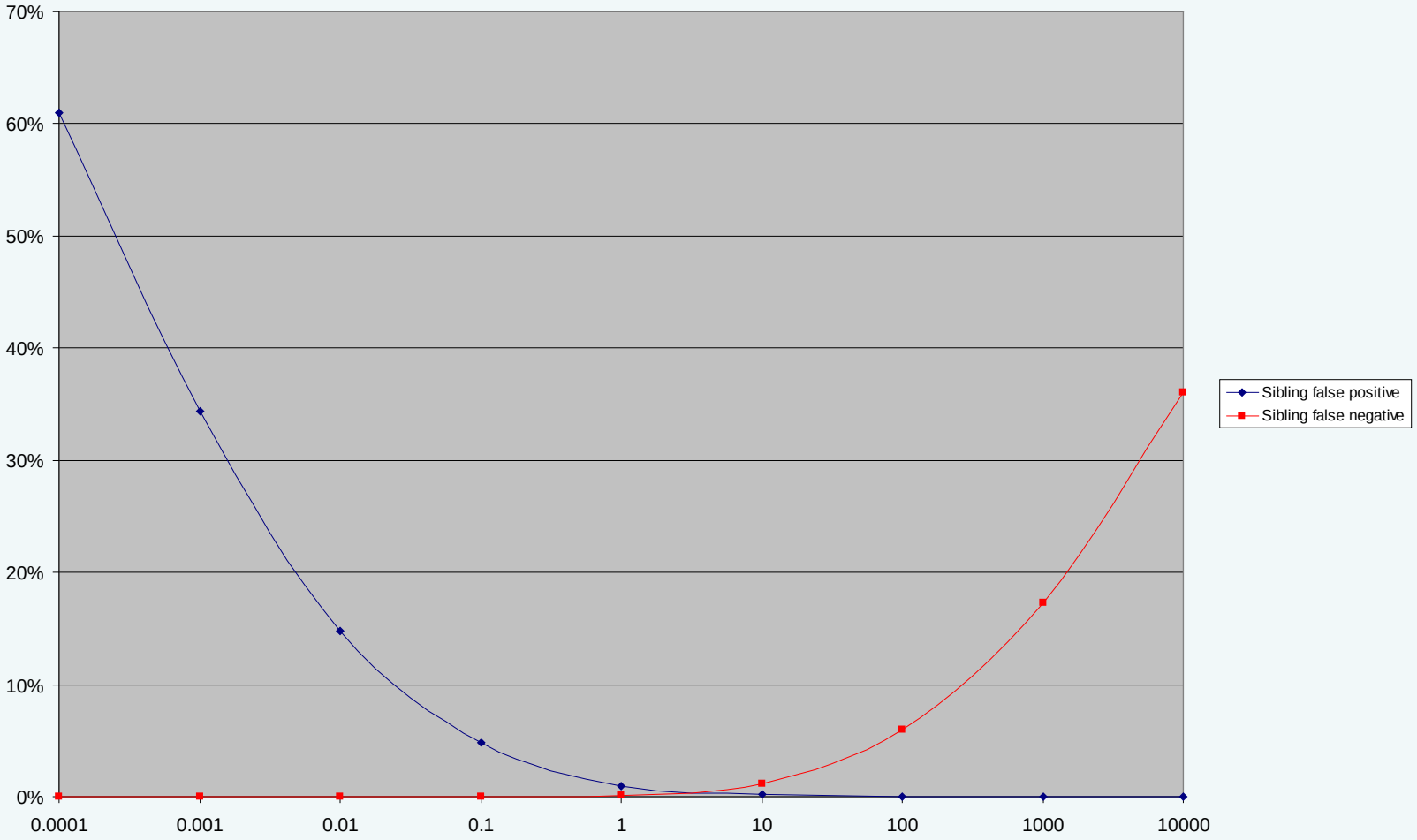
Hypothesis testing: LR threshold of 1 with prior odds of 1

		True state	
		Evidence from Unrelated individual	Evidence from sibling
Decision	Evidence from unrelated individual	~ 98% [Correct decision]	~4% [Type II error; false negative]
	Evidence from sibling	~ 2% [Type I error; false positive]	~ 96% [Correct decision]

Two types of errors

- False positives (Type I): an initial suspect's family is investigated even though an unrelated individual *is* the actual source of the evidence sample.
- False negatives (Type II): an initial suspect's family is *not* be investigated even though a relative really is the source of the evidence sample.
- A wide net (low LR threshold) catches more criminals but comes at the cost of more fruitless investigations.

Type I and II errors with prior odds of 1



Is the true DNA match a relative or a random individual?

- What is the likelihood that a close relative of a single initial suspect would match the evidence sample perfectly?
- What is the likelihood that a single randomly chosen, unrelated individual would match the evidence sample perfectly?

$$LR = \frac{P(E \mid relative)}{P(E \mid random)}$$

Is the true DNA match a relative or a random individual?

- What is the likelihood that the source of the evidence sample was a relative of an initial suspect?

$$P(sib \mid E) = \frac{P(E \mid sib) \cdot P(sib)}{P(E \mid sib) \cdot P(sib) + P(E \mid random) \cdot P(random)}$$

$$P(sib) = \frac{s}{popsize}$$

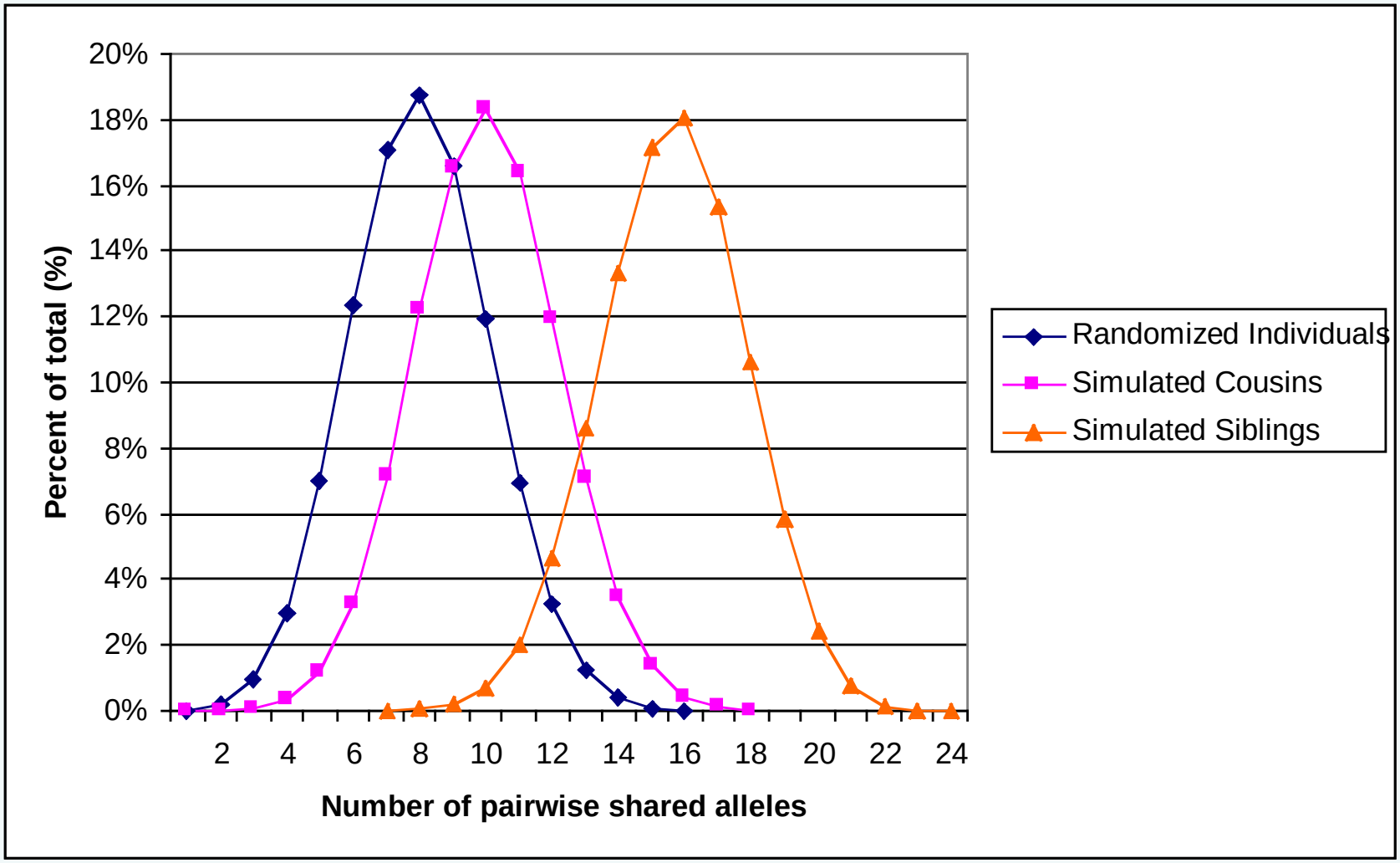
$$P(random) = \frac{popsize - s}{popsize}$$

Is the true DNA match a relative or a random individual?

- This more difficult question is ultimately governed by two considerations:
  - What is the size of the alternative suspect pool?
  - What is an acceptable rate of false positives?

$$LR = \frac{P(E \mid sib)}{P(E \mid random)}$$

Pair-wise similarity distributions



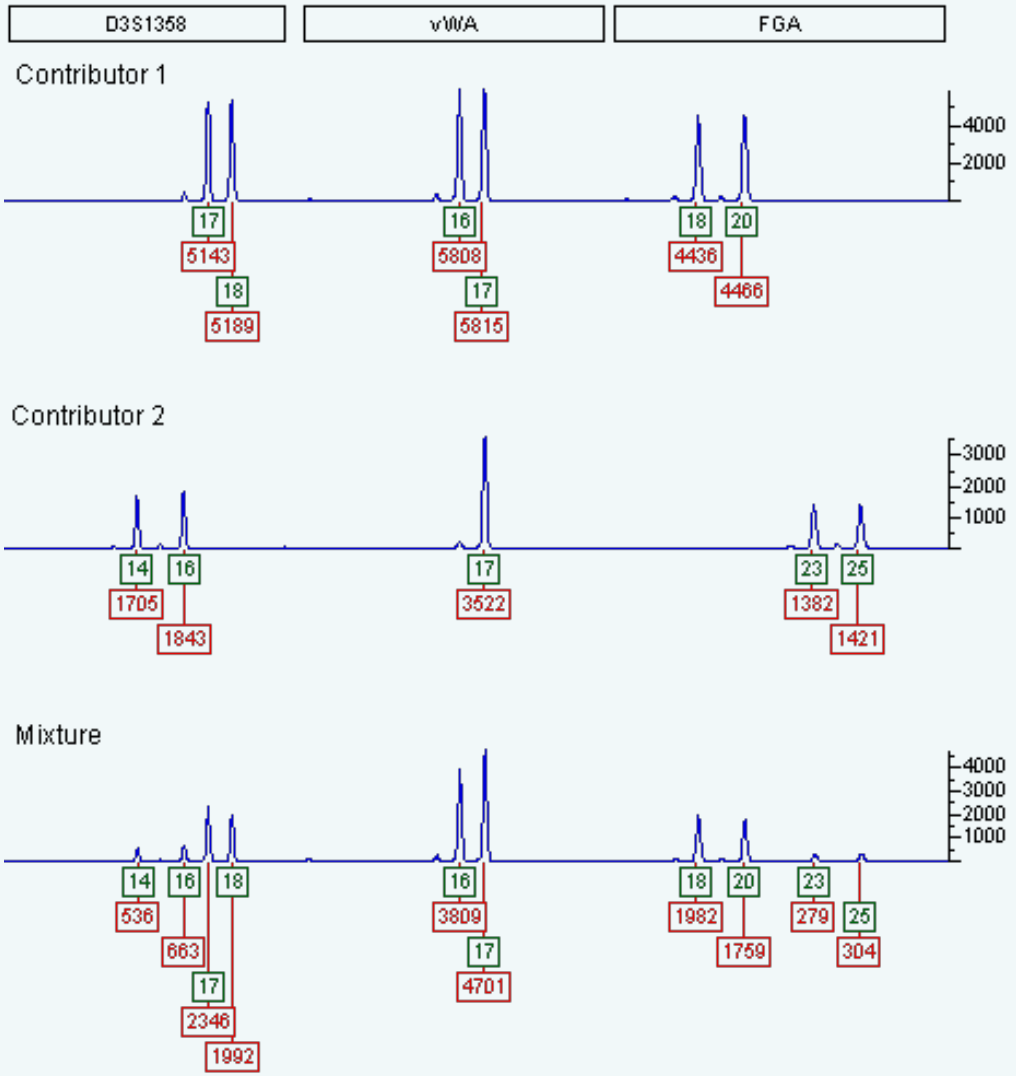
How well does an LR approach perform relative to alternatives?

- Low-stringency CODIS search identifies all 10,000 parent-child pairs (but only 1,183 sibling pairs and less than 3% of all other relationships and a high false positive rate)
- Moderate and high-stringency CODIS searches failed to identify any pairs for any relationship
- An allele count-threshold (set at 20 out of 30 alleles) identifies 4,233 siblings and 1,882 parent-child pairs (but fewer than 70 of any other relationship and with no false positives)

How well does an LR approach perform relative to alternatives?

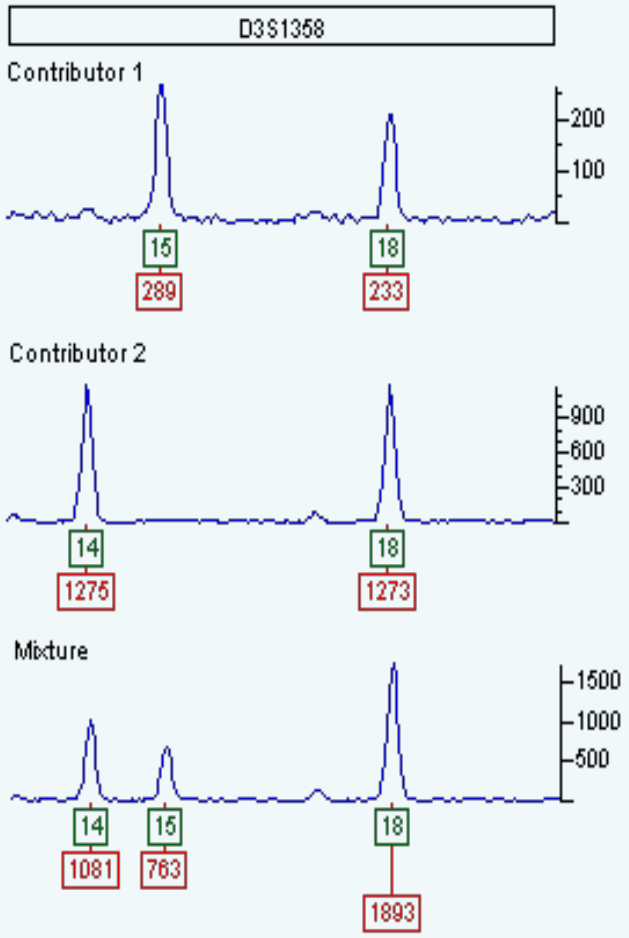
- LR set at 1 identifies > 99% of both sibling and parent-child pairs (with false positive rates of 0.01% and 0.1%, respectively)
- LR set at 10,000 identifies 64% of siblings and 56% of parent-child pairs (with no false positives)
- Use of non-cognate allele frequencies results in an increase in false positives and a decrease in true positives (that are largely offset by either a ceiling or consensus approach)

Introduction to Mixtures



- Mixtures can exhibit up to two peaks *per contributor* at any given locus
- Mixtures can exhibit as few as 1 peak at any given locus (regardless of the number of contributors)

Introduction to Mixtures



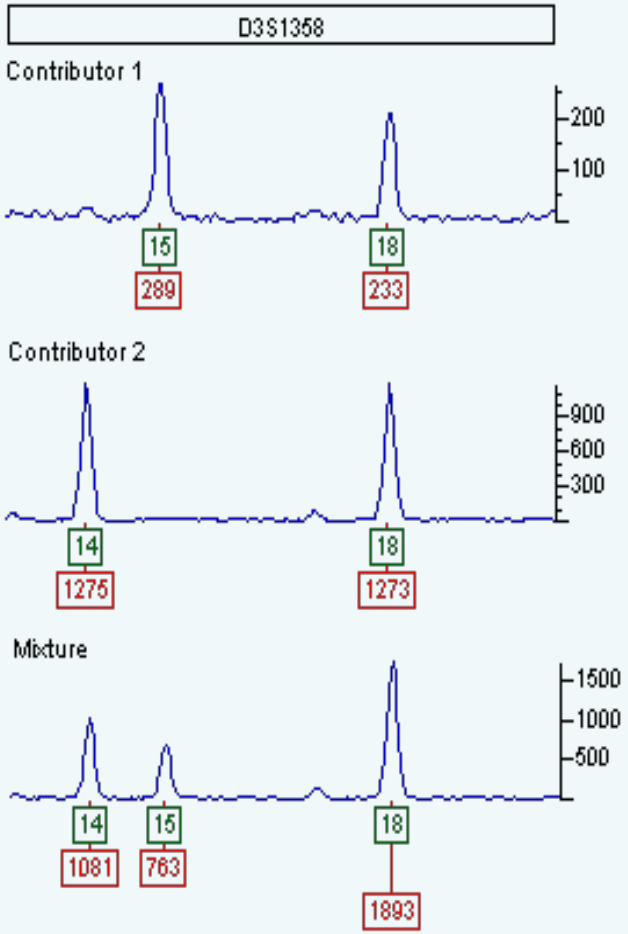
- Determining if two genotypes *could* be contributors is relatively easy

Possible contributors to a mixture:

	D3 locus genotype
Individual #1:	15, 18
Individual #2:	14, 18
Mixture:	14, 15, 18

- But beware – the opposite is *not* true

Introduction to Mixtures



- Determining what genotypes *created* the mixture is non-trivial

	D3 locus genotype
Mixture:	14, 15, 18

Option #1

Individual A: 15, 18

Individual B: 14, 18

Option #2

Individual B: 14, 18

Individual C: 15, 15

Option #3

Individual #D: 14, 15

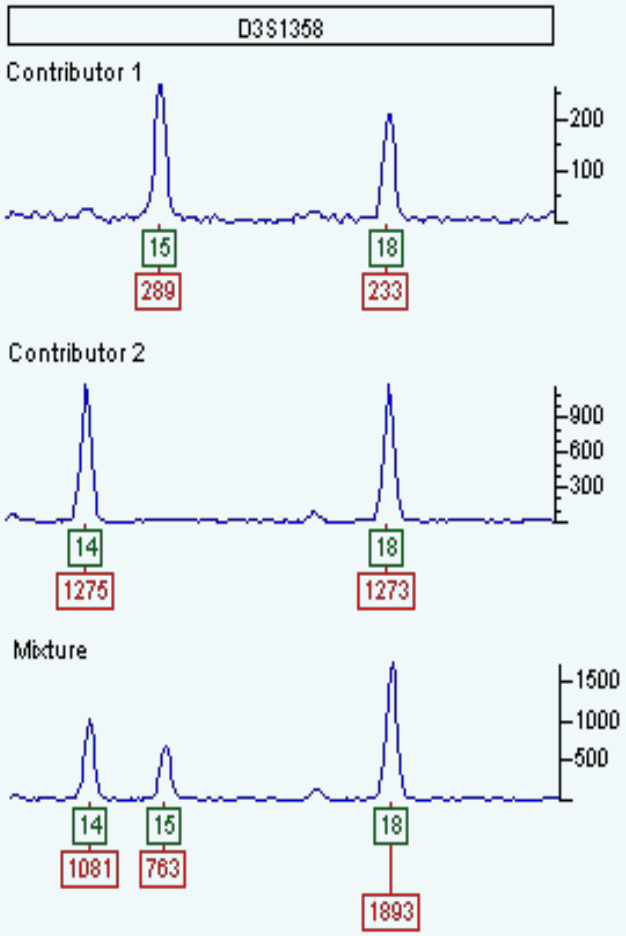
Individual #E: 18, 18

Option #4

Individual #A: 15, 18

Individual #F: 14, 14

Introduction to Mixtures



- Even determining the number of contributors is non-trivial

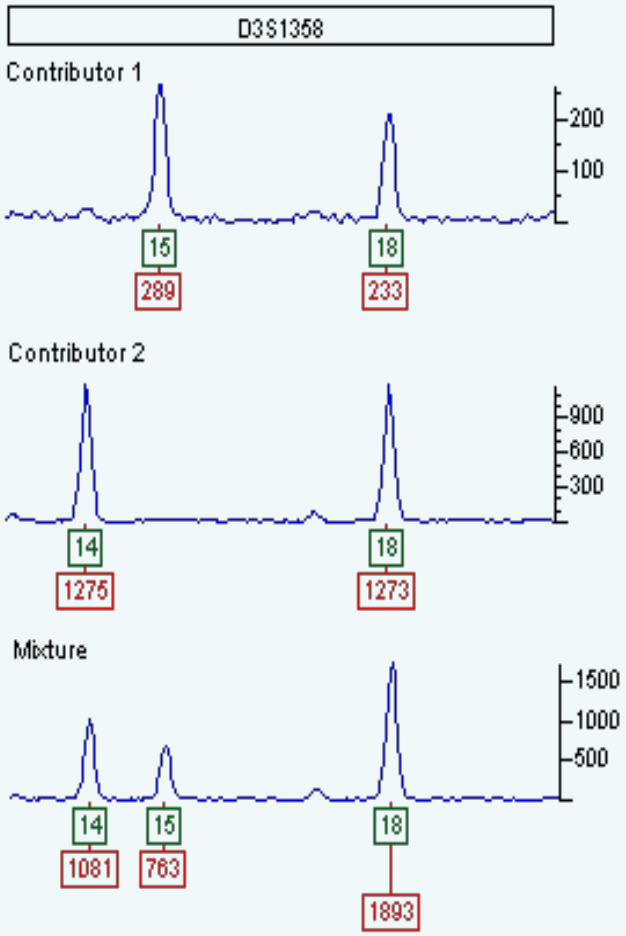
	D3 locus genotype
Mixture:	14, 15, 18

Another Option

Individual C:	15, 15
Individual D:	14, 15
Individual E:	18, 18

- There is no "hard" mathematical upper bound to the number of contributors possible

Introduction to Mixtures



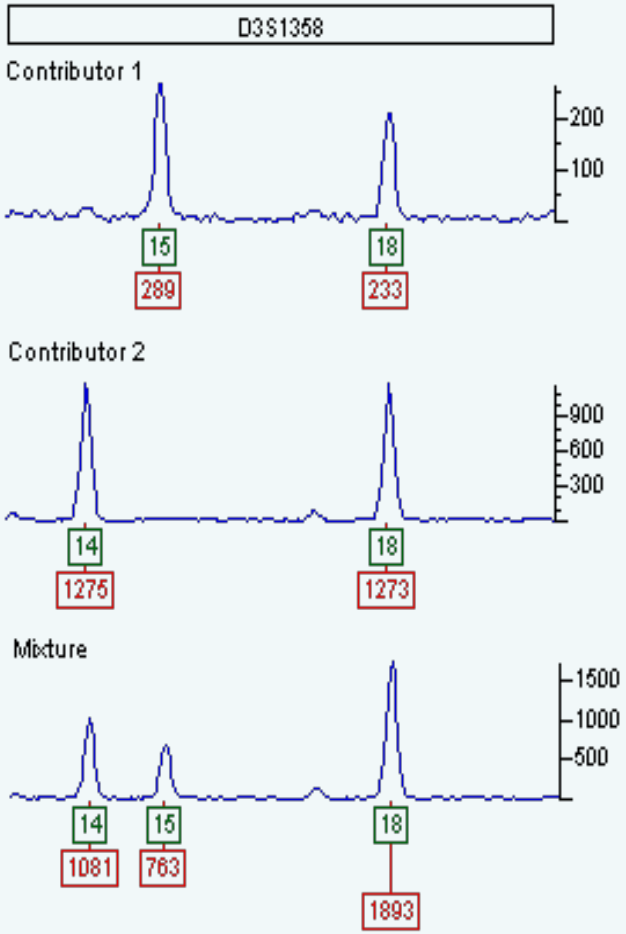
- Usually the victim's genotype is known, but this does not always make the defendant's genotype clear

	D3 locus genotype
Mixture:	14, 15, 18
Victim:	14, 15

Possible genotypes for a single perpetrator:

Individual C:	14, 18
Individual D:	15, 18
Individual E:	18, 18
Individual F:	14, 14 ?

Introduction to Mixtures

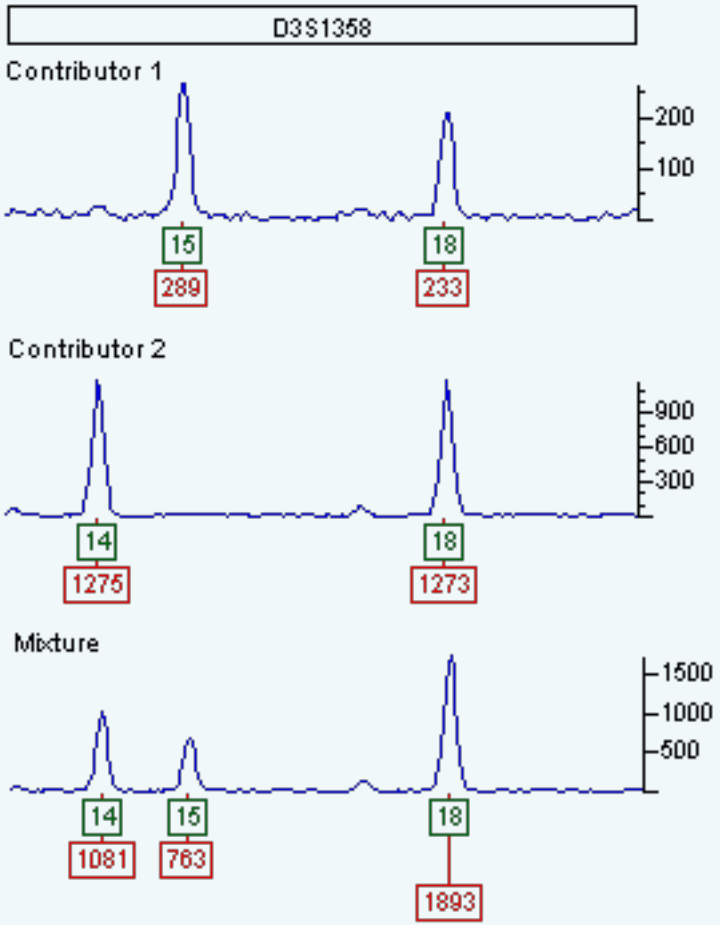


- The large number of potential genotypes consistent with the mixture allows for a *VERY* wide net to be cast
  - This greatly increases the likelihood of accusing an innocent suspect, particularly in database trawls
  - This is generally *not* reflected in the statistics reported by the DNA testing laboratory
  - Case History: Sutton

Making sense of mixtures

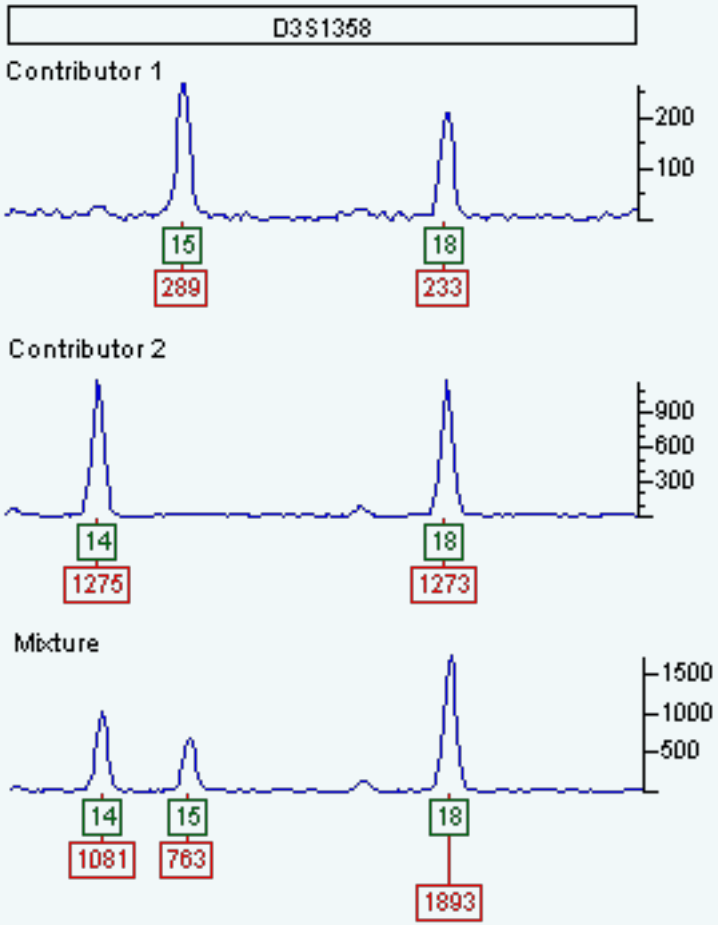
- There are two major open research areas:
  - Determining the most likely number of contributors
  - Determining the genotypes of each contributor
- Factors that can aid in deconvolution
  - Mixture ratios
  - Peak height additivity
- Factors that can greatly complicate deconvolution results
  - Allowing alleles to be discarded as artifacts ("analyst's discretion")

Mixture ratios



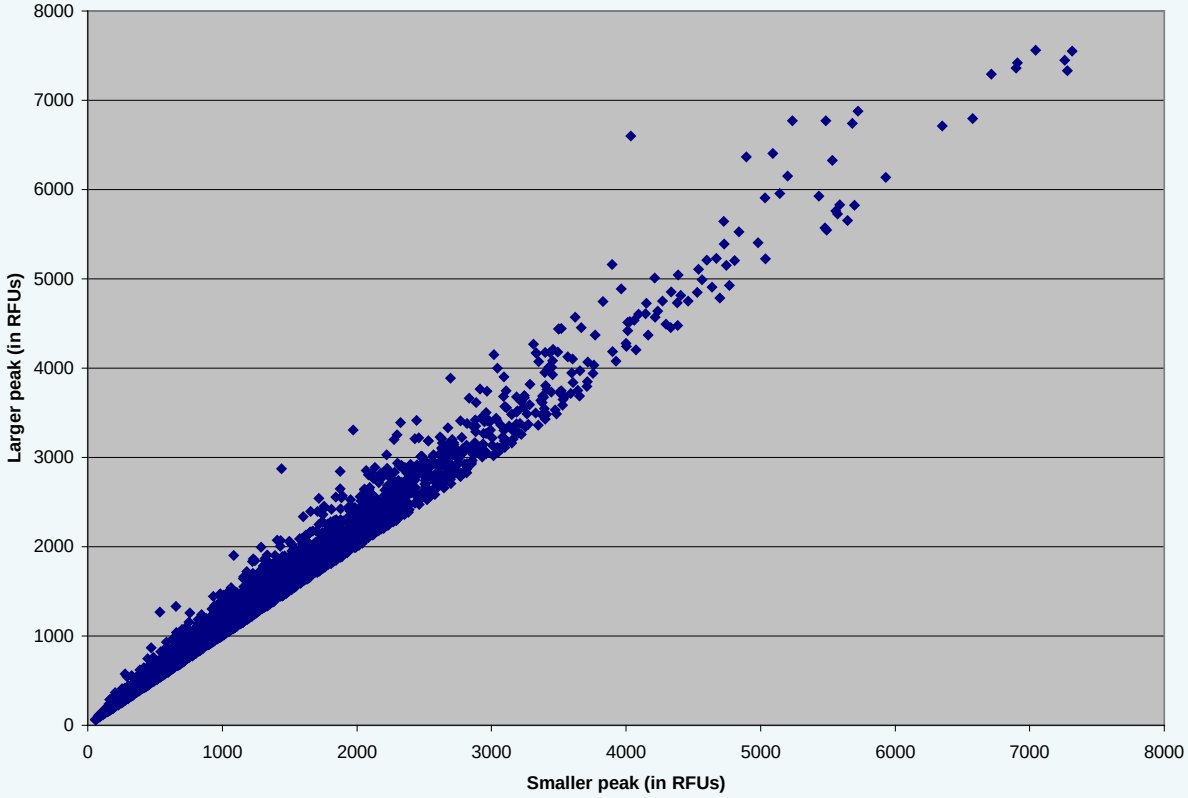
- Different individuals may contribute different "amounts" of DNA to the mixture. This difference should be reflected (relatively uniformly) throughout the entire sample.

Peak height additivity



- Assume one individual contributes an amount of DNA that measured at *n* RFUs
- Assume a second individual contributes an amount of DNA that measures at *m* RFUs
- In a two person mixture, any allele which they share should measure at roughly *n* + *m* RFUs

Evidence of additivity



Relationship between the smaller and larger peaks in heterozygous loci of reference samples.

Making sense of mixtures

- There are two major research areas:
  - **Determining the most likely number of contributors**
  - Determining the genotypes of each contributor

- How can we determine the mostly likely number of contributors?
  - We (Paoletti *et al*.) create mixtures from an existing database in order to determine how often the *actual* number of contributors differs from the *perceived* number of contributors.
  - The Minnesota BCA database uses twelve (12) loci

Minnesota BCA database

BCA ID#	D3S1358	vWA	FGA	TH01	TPOX	CSF1PO	D5S818	D13S317	D7S820	D8S1179	D21S1
PB0005	17,18	16,16	21,24	6,8	10,11	11,12	12,14	11,12	8,10	13,14	29,29
PH0070	15,17	16,17	21,25	7,7	10,11	11,12	11,12	11,12	8,10	13,14	29.2,
PH0138	17,17	14,16	24,25	7,8	11,11	10,11	11,11	10,10	10,11	14,14	29,30
Mixture1	15,17,18	14,16,17	21,24,25	6,7,8	10,11	10,11,12	11,12,14	10,11,12	8,10,11	13,14	29,29
PB0155	16,17	16,16	24,24	8,9.3	10,11	11,12	11,13	12,12	10,11	12,15	29,29
PH0014	17,17	17,18	19,22	6,9.3	11,12	12,12	11,11	9,9	11,11	13,15	28,29
PN0166	15,16	17,17	19,22	9.3,9.3	11,11	12,13	11,11	12,13	9,10	12,12	30,30
Mixture2	15,16,17	16,17,18	19,22,24	6,8,9.3	10,11,12	11,12,13	11,13	9,12,13	9,10,11	12,13,15	28,29
PB0022	15,16	15,16	22,23	7,7	9,11	11,12	11,12	14,14	11,12	14,15	32.2,
PB0078	15,17	15,15	23,24	7,8	10,10	11,12	11,12	13,13	10,10	13,13	28,28
PH0146	17,17	16,16	24,24	8,9.3	9,11	10,12	12,12	8,8	10,12	13,14	28,32
Mixture3	15,16,17	15,16	22,23,24	7,8,9.3	9,10,11	10,11,12	11,12	8,13,14	10,11,12	13,14,15	28,32
PB0024	17,18	16,18	22,24	7,8	6,9	10,11	11,11	9,12	8,10	15,15	29,29
PB0067	17,18	16,19	22,24	7,8	11,11	10,10	12,13	11,12	8,8	12,13	29,30
PB0111	15,18	16,16	23,24	8,9.3	6,9	10,11	11,12	11,12	10,12	12,15	30,31
Mixture4	15,17,18	16,18,19	22,23,24	7,8,9.3	6,9,11	10,11	11,12,13	9,11,12	8,10,12	12,13,15	29,30
PB0024	17,18	16,18	22,24	7,8	6,9	10,11	11,11	9,12	8,10	15,15	29,29
PB0075	16,18	16,16	22,24	9.3,9.3	8,8	7,10	8,11	11,11	8,8	14,14	29,32
PC0090	16,17	14,18	22,25	7,8	8,8	10,11	12,12	11,11	8,12	12,15	29,30
Mixture5	16,17,18	14,16,18	22,24,25	7,8,9.3	6,8,9	7,10,11	8,11,12	9,11,12	8,10,12	12,14,15	29,30
PB0030	14,16	15,15	22,22	7,7	8,9	11,11	11,13	12,13	10,11	14,16	28,29
PH0055	16,16	16,18	24,24	7,9	8,11	11,12	11,12	12,14	8,11	13,14	28,29
PN0108	15,16	18,18	22,23	9.3,9.3	11,11	11,11	11,11	12,14	8,8	14,16	29,30
Mixture6	14,15,16	15,16,18	22,23,24	7,9,9.3	8,9,11	11,12	11,12,13	12,13,14	8,10,11	13,14,16	28,29

All 3-way MN BCA mixtures

- There are 45,139,896 possible different 3-person mixtures of the 648 individuals in the MN BCA database

Maximum # of alleles observed	# of occurrences	As Percent
2	0	0.00%
3	310	0.00%
4	2,498,139	5.53%
5	29,938,777	66.32%
6	12,702,670	28.14%

- 6% of three contributors mixtures “look like” two contributors

All 3-way MN BCA mixtures

- What if "analyst's discretion" is invoked exactly once (at the "worst" locus)

Maximum # of alleles observed	# of occurrences	As Percent
1, 2	0	0.00%
	0	0.00%
3	310	0.00%
	8,151	0.02%
4	2,498,139	5.53%
	11,526,219	25.53%
5	29,938,777	66.32%
	32,078,976	71.01%
6	12,702,670	28.14%
	1,526,550	3.38%

- 26% of three contributors mixtures "look like" two contributors

All 4-way MN BCA mixtures

Maximum # of alleles observed	# of occurrences	As Percent
1, 2, 3	0	0.00%
	6	0.00%
4	42,923	0.07%
	731,947	1.25%
5	9,365,770	15.03%
	30,471,965	52.18%
6	34,067,153	58.32%
	25,872,024	44.29%
7	13,719,403	23.49%
	1,328,883	2.28%
8	1,214,261	2.08%
	4,695	0.01%

- 73% of four contributors mixtures “look like” three contributors

All 4-way MN BCA mixtures

Maximum # of alleles observed	# of occurrences	As Percent
1, 2, 3	0	0.00%
	6	0.00%
4	42,923	0.07%
	731,947	1.25%
5	9,365,770	15.03%
	30,471,965	52.18%
6	34,067,153	58.32%
	25,872,024	44.29%
7	13,719,403	23.49%
	1,328,883	2.28%
8	1,214,261	2.08%
	4,695	0.01%

- 96% of four contributors mixtures “look like” three contributors when one locus can be dropped from consideration

Removing possible relationships

Individual	vWA	
	Original	Redistributed
1	18,19	15,18
2	18,18	18,18
. . .	. . .	. . .
648	14,15	14,19

- Redistribute alleles at each locus randomly
- New database of “synthetic” unrelated individuals with the same allele frequencies

3-way mixtures with all 12 loci

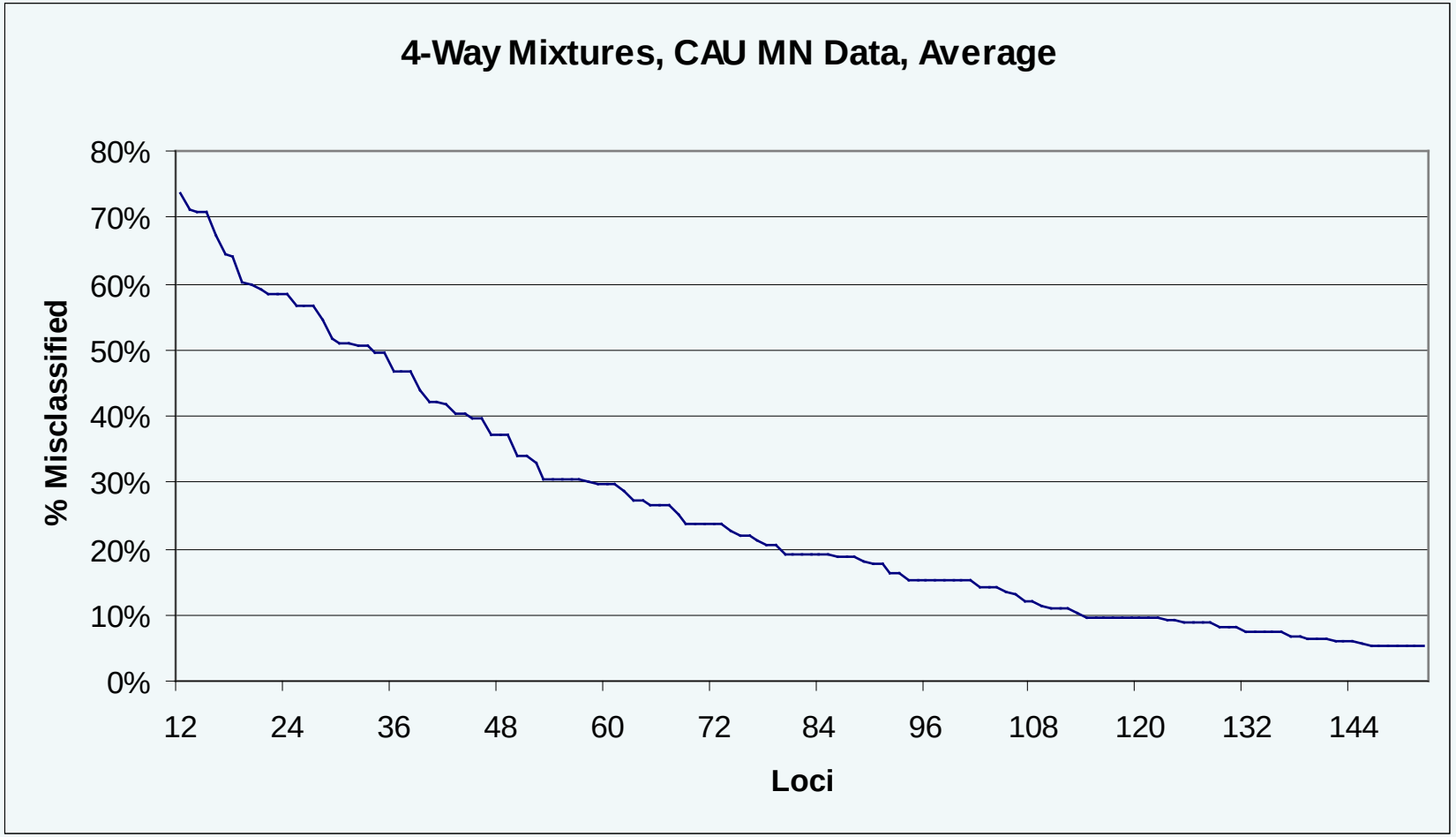
Maximum # of alleles observed in a 3-person mixture	# of occurrences	Percent of mixtures
2	0	0.00%
3	310	0.00%
4	2,498,139	5.53%
5	29,938,777	66.32%
6	12,702,670	28.14%

MN BCA Original Data

Maximum # of alleles observed in a 3-person mixture	# of occurrences	Percent of mixtures
2	0.0	0.00%
3	139.4	0.00%
4	2,233,740.8	4.95%
5	29,829,482.0	66.08%
6	13,076,533.8	28.97%

Synethtic “Unrelated” Data

How many loci until 4-way mixture doesn't look like a 3-way mixture?



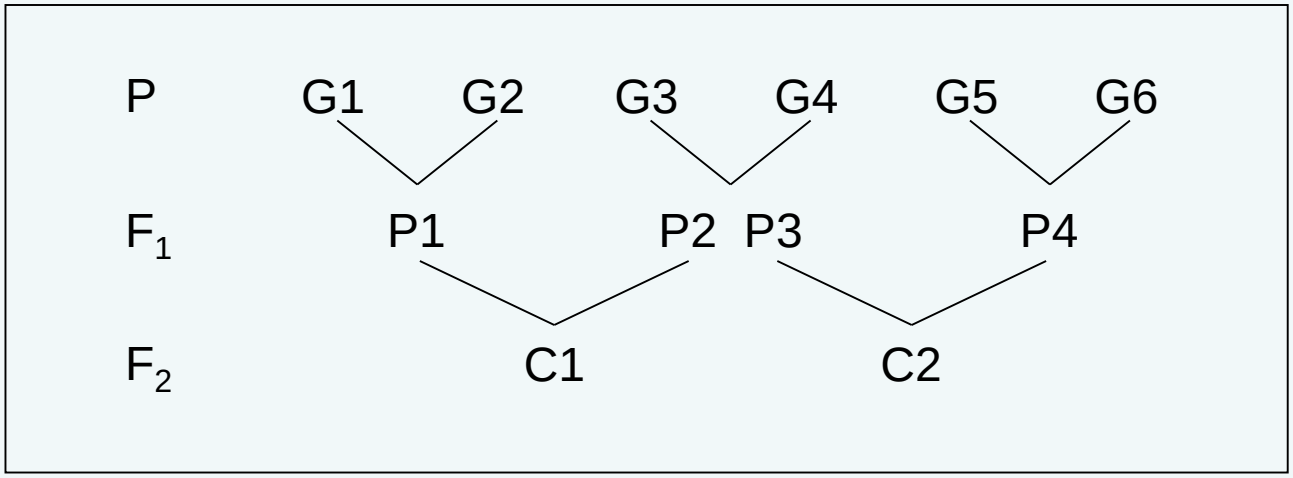
- Redistribute alleles across all individuals (by locus) and add to database

What if contributors *are* related?

- Clearly, determining the number of contributors to a DNA mixture is difficult when the contributors are unrelated

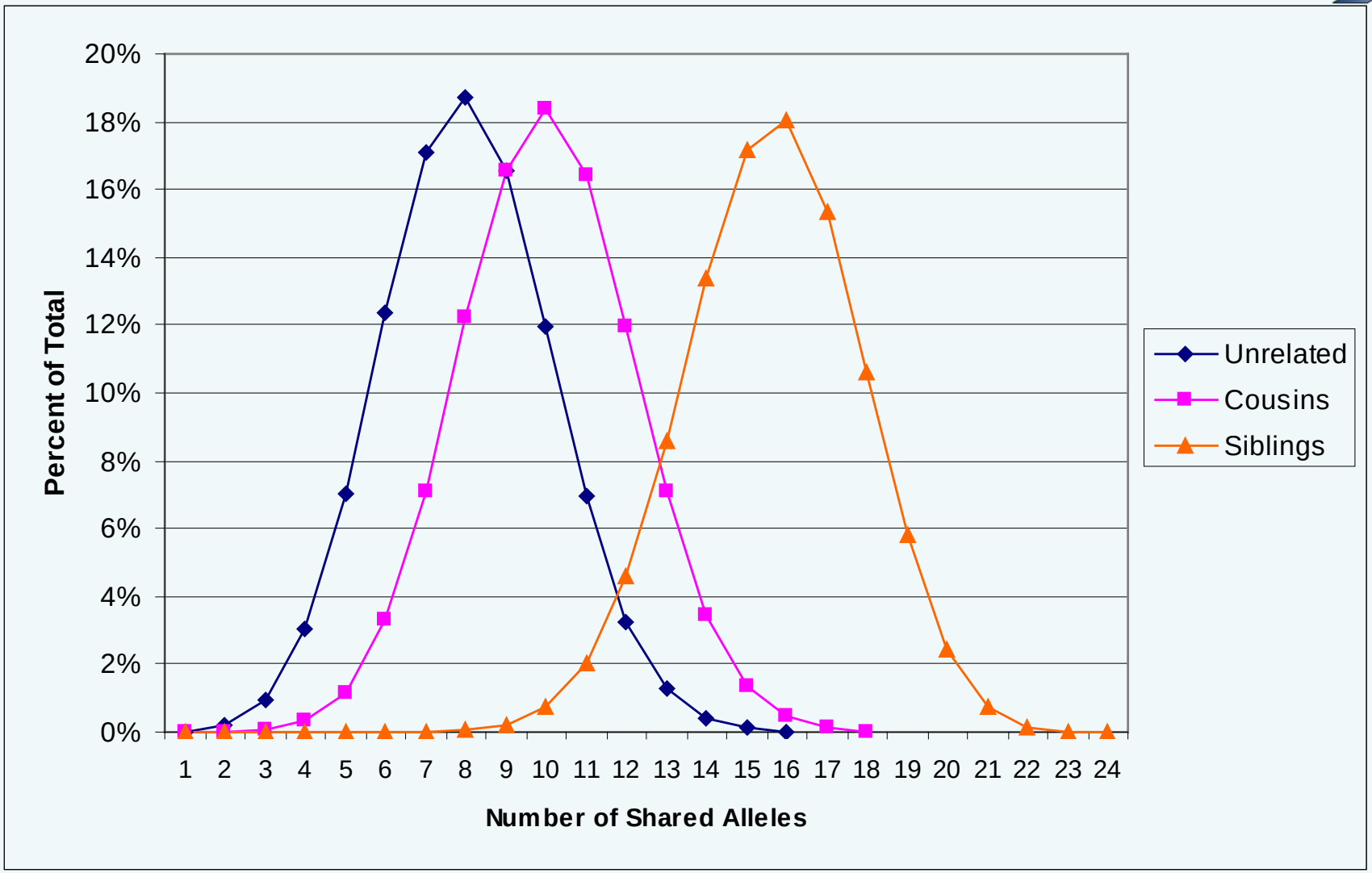
- How much harder does it become when they *are* related?

Virtual families

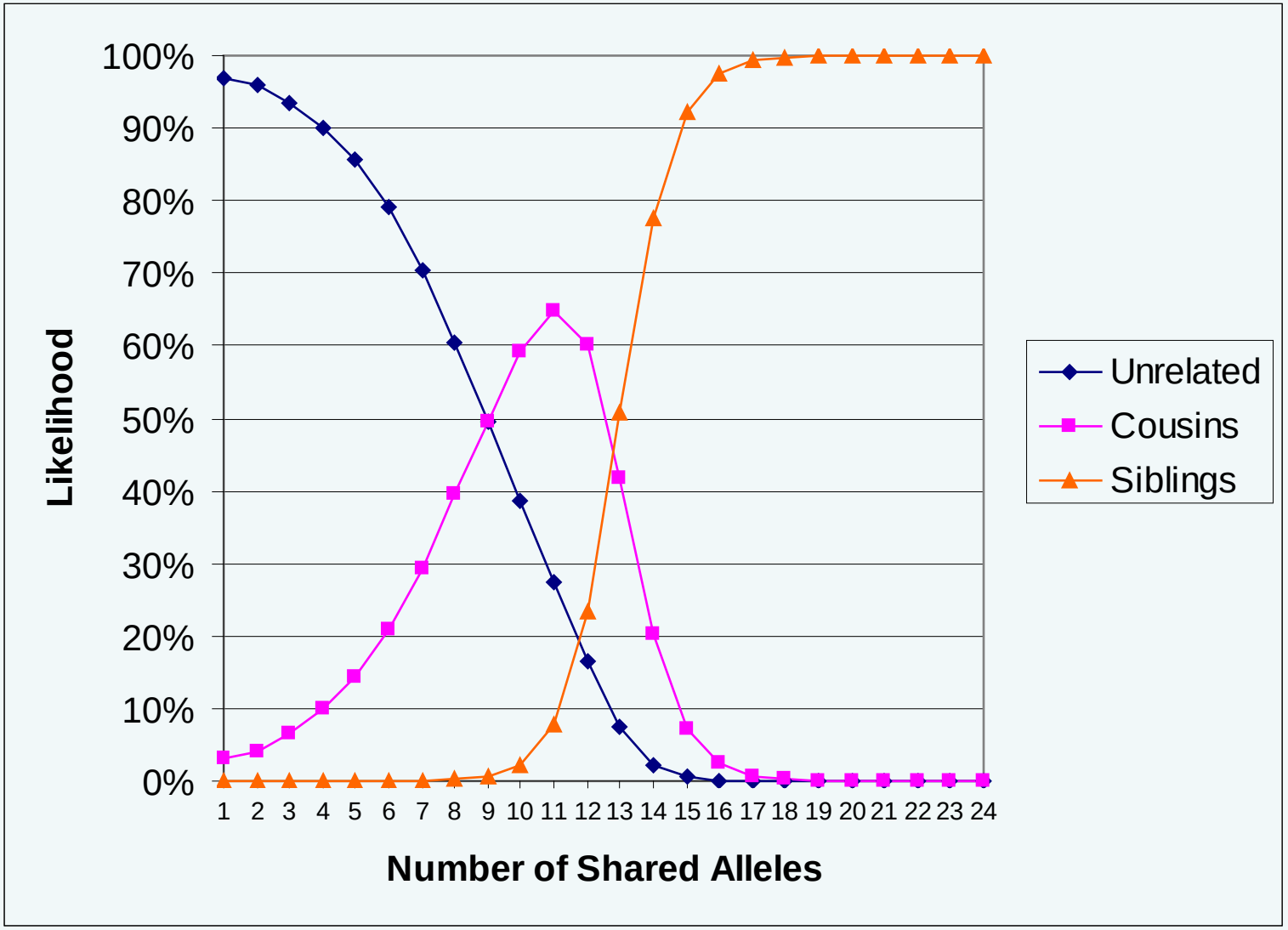


- Parents randomly chosen from unrelated (randomized) database
- Random mating
- Creates databases of grandparents, parents, and grandchildren

Distributions of shared alleles



Likelihoods of shared alleles

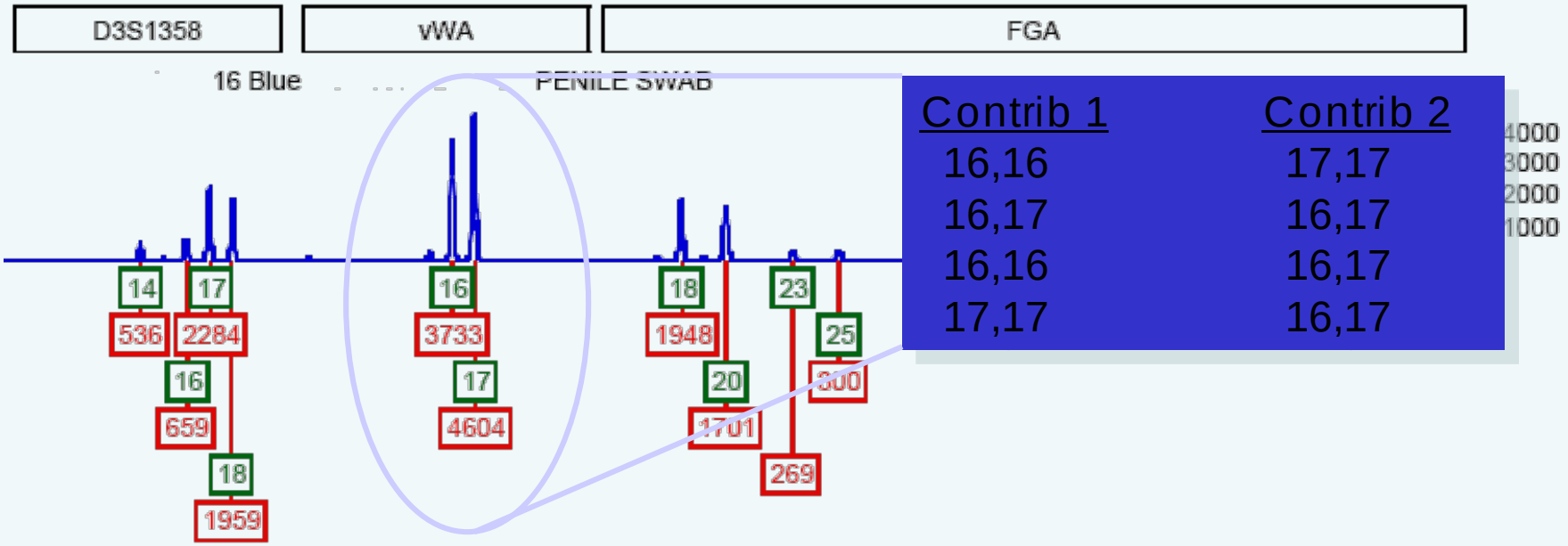


Analysis of Allele Sharing

- Clearly, it is difficult to definitively assign the number of contributors to a mixture
- This difficulty must be fairly reported in random probability match statistics in order for such statistics to remain objective
- Analyst discretion should be invoked cautiously, and always carefully double-checked for error
- Likelihoods allow a analysts to infer the possible relationship between two individuals

Mixture Deconvolution

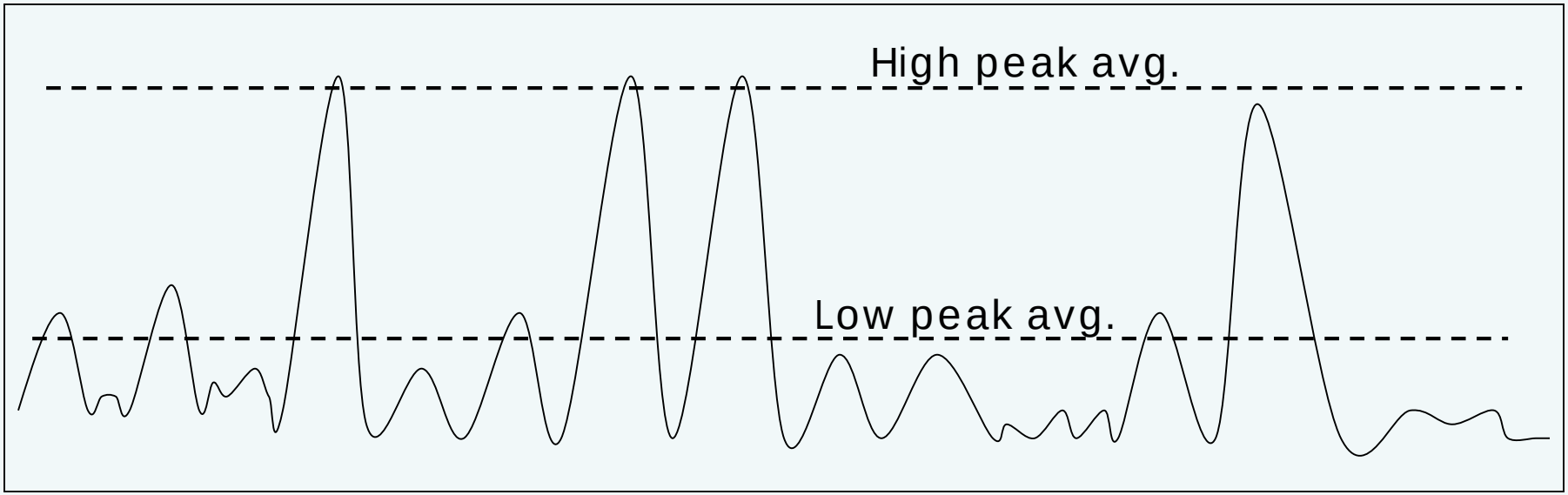
- Even when the number of contributors is known (or assumed), separating mixtures into their components can be difficult



Contrib 1	Contrib 2
16,16	17,17
16,17	16,17
16,16	16,17
17,17	16,17

Current Methods

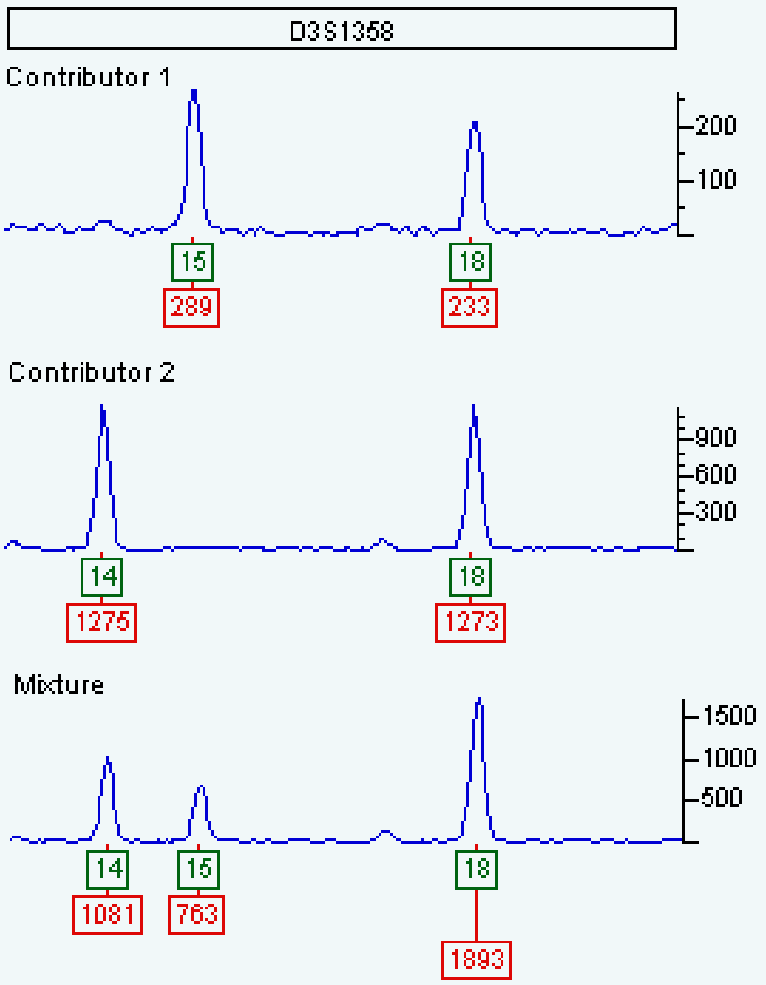
- Most methods start by inferring the mixture ratio:



Simple example: All loci heterozygous, two contributors

Minimal Basic Assumptions

- A primary assumption of all methods is **peak additivity**
- Most labs assume peaks from the same source will vary by ≤ 30%



Objectives

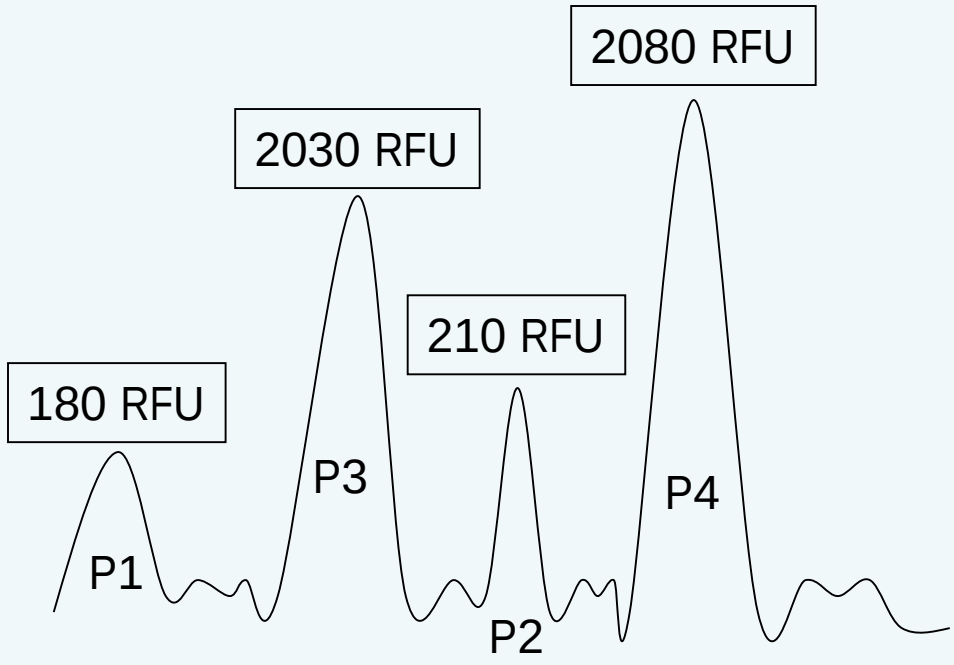
- Start with **simple assumptions**:
  - Additivity with constant variance: c
  - Peaks below a minimum threshold (often 50 or 150 RFU) are not observable
  - Peaks above the saturation threshold (often 4000 RFU) are not measurable
- Obtain **provably correct** deconvolution where possible
- Identify when this is not possible

Method

- Assume the number of contributors
- Enumerate all possible mixture contributor combinations
- Determine which pairs of profiles contain peaks **in balance**

Peak Balance

- Example: assume two contributors, four peaks:
  - For this locus, and $c = 1.3$, the combination (P1,P3) is out of balance because:



Peaks are numbered by height

$$180 \times c < 2030$$

Example: Mixture of four peaks

Contributor 1	Contributor 2	Mixture Condition 1	Mixture Condition 2
P4 P3	P2 P1	$P4 \leq cP3$	$P2 \leq cP1$
P4 P2	P3 P1	$P4 \leq cP2$	$P3 \leq cP1$
P4 P1	P3 P2	$P4 \leq cP1$	$P3 \leq cP2$

- P4 >= P3 >= P2 >= P1 >= Min. Threshold

Sweet Spot

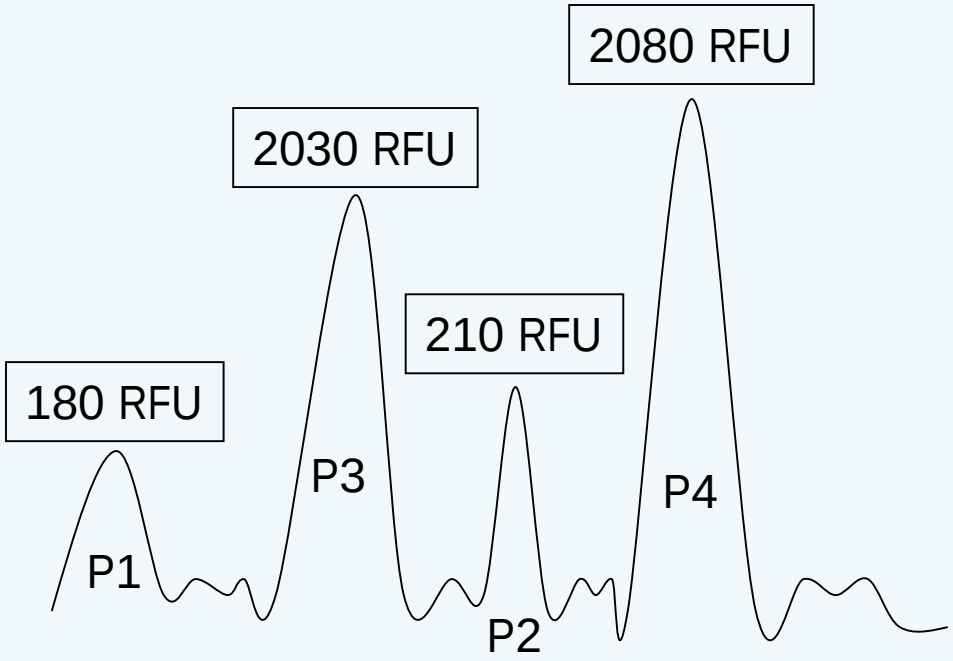
Contributor 1	Contributor 2	Mixture Condition 1	Mixture Condition 2
P4 P3	P2 P1	$P4 \leq cP3$	$P2 \leq cP1$
P4 P2	P3 P1	$P4 \leq cP2$	$P3 \leq cP1$
P4 P1	P3 P2	$P4 \leq cP1$	$P3 \leq cP2$

- If only one row is satisfied, then the genotypes can be **unambiguously** and **provably** determined

Example: In the sweet spot

Contributor 1	Contributor 2	Mixture Condition 1	Mixture Condition 2
P4 P3	P2 P1	$P4 \le cP3$	$P2 \le cP1$
~~P4 P2~~	~~P3 P1~~	$P4 \le cP2$	$P3 \le cP1$
~~P4 P1~~	~~P3 P2~~	$P4 \le cP1$	$P3 \le cP2$

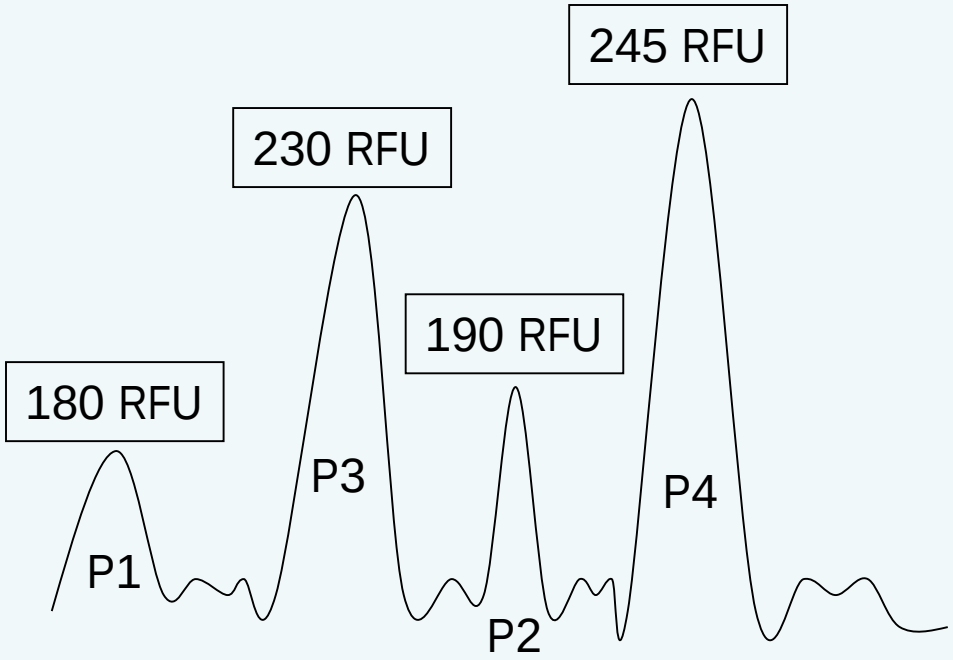
- $P4 > cP2$
  so we can't have (P4,P2)
- $P4 > cP1$
  so we can't have (P4, P1)



Example: Ambiguous Locus

Contributor 1	Contributor 2	Mixture Condition 1	Mixture Condition 2
P4 P3	P2 P1	$P4 \le cP3$	$P2 \le cP1$
P4 P2	P3 P1	$P4 \le cP2$	$P3 \le cP1$
~~P4 P1~~	~~P3 P2~~	$P4 \le cP1$	$P3 \le cP2$

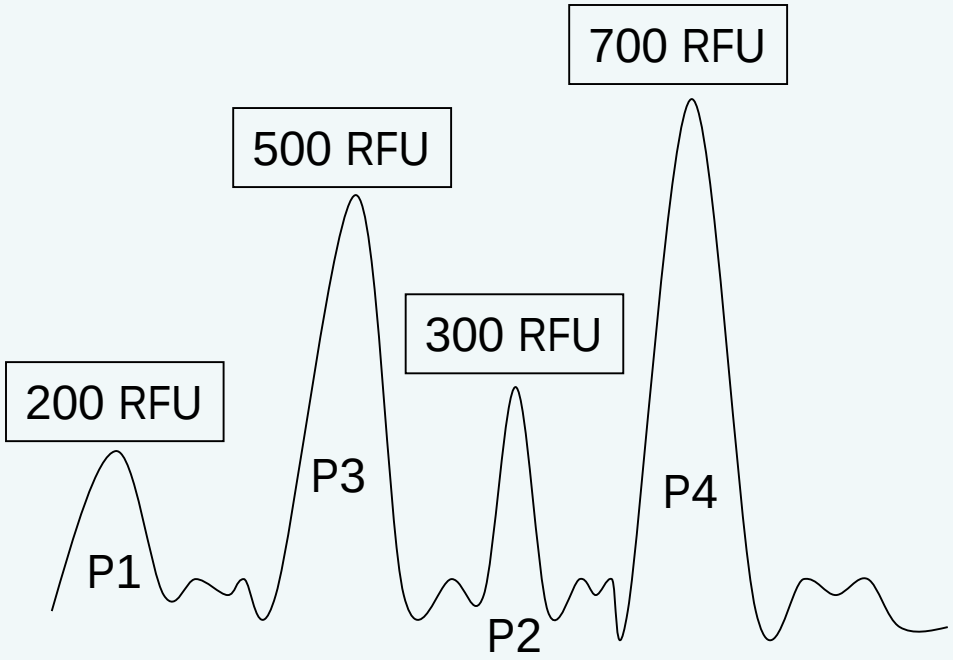
- P2 is within *c* of both P1 and P4, so we can have
  - (P1,P3) (P2,P4), or
  - (P1,P2) (P3,P4)
- P4 cannot pair with P1



Example: No row satisfied

Contributor 1	Contributor 2	Mixture Condition 1	Mixture Condition 2
~~P4 P3~~	~~P2 P1~~	P4 ≤ *c*P3	P2 ≤ *c*P1
~~P4 P2~~	~~P3 P1~~	P4 ≤ *c*P2	P3 ≤ *c*P1
~~P4 P1~~	~~P3 P2~~	P4 ≤ *c*P1	P3 ≤ *c*P2

- P4 (for example) cannot pair with any other peak
- One of our assumptions (*c* or the number of contributors) is incorrect



Three Peaks

Contributor 1	Contributor 2	Mixture Condition 1	Mixture Condition 2
P3 P3	P2 P1	None (homozygote)	$P2 \leq c \times P1$
P3 P2	P3 P1	$P3 \leq c \times (P2+P1)$	$P3 \geq (1/c) \times (P2+P1)$
P3 P2	P2 P1	$P2 \leq c \times (P3+P1)$	$P2 \geq (1/c) \times (P3+P1)$
P3 P2	P1 Pmpht	$P3 \leq c \times P2$	$P1 \leq c \times Pmpht$
P3 P2	P1 P1	$P3 \leq c \times P2$	None
P3 P1	P2 Pmpht	$P3 \leq c \times P1$	$P2 \leq c \times Pmpht$
P3 P1	P2 P2	$P3 \leq c \times P1$	None
P3 P1	P2 P1	$P1 \leq c \times (P3+P2)$	$P1 \geq (1/c) \times (P3+P2)$
P3 Pmpht	P2 P1	$P3 \leq c \times Pmpht$	$P2 \leq c \times P1$

Advantages of the method

- If you accept the simple assumptions, the resulting mixture interpretations directly follow
- Interprets mixtures on a locus by locus basis
- Does not interpret ambiguous loci

Future work

- Mixture ratio can be inferred only from unambiguous loci, and then applied to perform an more aggressive interpretation of the ambiguous loci when desired
- Confidence values can be applied to the more aggressively interpreted possitions

Acknowledgements

- Research Students
  - David Paoletti (analysis of allele sharing)
  - Jason Gilder (data collection, additivity study, mixture deconvolution)
- Faculty
  - Travis Doom
  - Dan Krane
  - Michael Raymer