

Mixture Deconvolution

Florida Statewide Training Meeting
 Indian Rocks Beach, FL
 May 12-13, 2008



Dr. John M. Butler
 National Institute of Standards and Technology
john.butler@nist.gov



Outline

- Points for Consideration
 - DNA quantity and quality
- Deconvolution steps by Clayton *et al.* (1998)
- Worked Example – using NEST data
- Software programs introduced

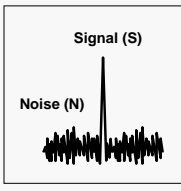
Final version available at
http://www.cstl.nist.gov/biotech/strbase/training/AAFS2008_MixtureWorkshop.htm

Points for Consideration

- Peak height vs peak area
- Thresholds – analytical vs stochastic levels
- Other lab-specific values:
 - Heterozygote peak height balance
 - Locus-specific stutter percentage
- DNA quantity and quality
 - problems with low-level or degraded DNA

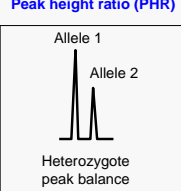
What is a true peak (allele)?

Peak detection threshold



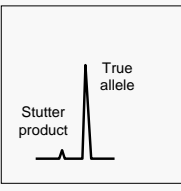
Signal > 3x sd of noise

Peak height ratio (PHR)



**PHR consistent with single source
Typically above 60%**

Stutter percentage



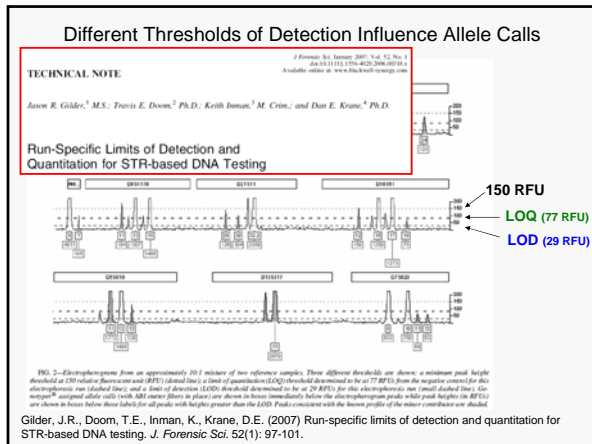
Stutter location below 15%

Validation Studies

- Information from validation studies should be used to set laboratory-specific
 - Stutter %
 - Peak Height Ratios
 - Minimum Peak Heights (detection thresholds)
 - Relative balance across loci
- These values are all dependent on amount of input DNA
 - If low-level DNA is amplified, stutter % may be higher and peak height ratios may be lower

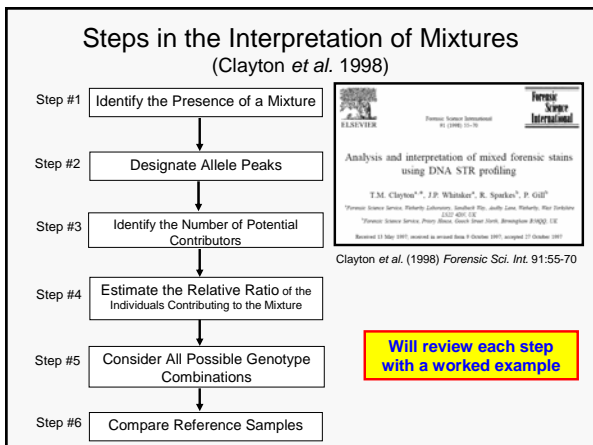
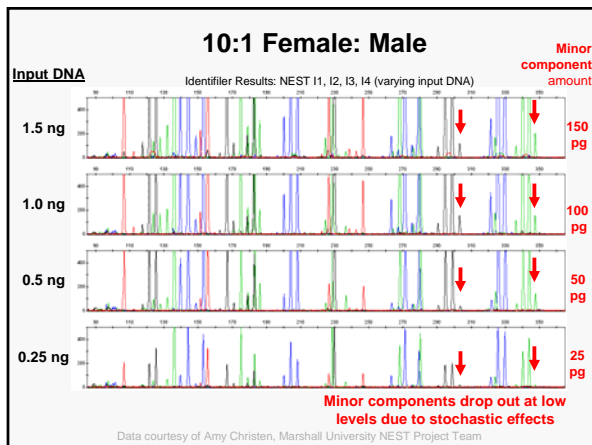
Thresholds

- Validation studies should be performed in each laboratory
- Some labs have set two thresholds:
 - Analytical thresholds – what is a peak? (50 RFU)
 - Stochastic thresholds – what is reliable PCR data? (150 RFU)



The Scientific Reasoning behind the Concept of an Analytical Threshold (limit of detection)

- This is fundamentally an issue of reliability
- For a peak intensity three times the standard deviation of the noise there is a limited chance that such a signal is the result of a random fluctuation
- This is because 99.7 percent of all noise signals fall below this value (from the definition of a Gaussian curve)
- Below this point the very real possibility exists that what you think is a peak is simply a statistical fluctuation in the baseline noise.



Step #1: Is a Mixture Present in an Evidentiary Sample?

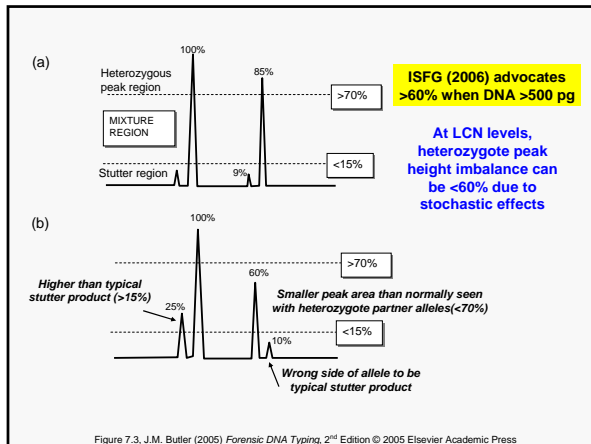
- Examine the **number of peaks present** in a locus
 - More than 2 peaks at a locus (except for tri-allelic patterns at perhaps one of the loci examined)
- Examine **relative peak heights**
 - Heterozygote peak imbalance <60%
 - Peak at stutter position >15%
- Consider all loci tested

Is a DNA Profile Consistent with Being a Mixture?

From J.M. Butler (2005) *Forensic DNA Typing, 2nd Edition*, pp. 156-157

If the answer to any one of the following three questions is yes, then the DNA profile may very well have resulted from a mixed sample:

- Do any of the loci show more than two peaks in the expected allele size range?
- Is there a severe peak height imbalance between heterozygous alleles at a locus?
- Does the stutter product appear abnormally high (e.g., >15-20%)?



Step #2: Designate Allele Peaks

- Use regular data interpretation rules to decipher between true alleles and artifacts
- Use stutter filters to eliminate stutter products from consideration (although stutter may hide some of minor component alleles at some loci)
- Consider heterozygote peak heights that are highly imbalanced (<60%) as possibly coming from two different contributors

Step #3: Identifying the Potential Number of Contributors

- **Important for some statistical calculations**
- Typically if 2, 3, or 4 alleles then 2 contributors
- If 5 or 6 alleles per locus then 3 contributors
- If >6 alleles in a single locus, then >4 contributors
- **JFS Nov 2005 paper by Forensic Bioinformatics on number of possible contributors**
 - Relies on maximum allele count alone
 - Does not take into account peak height information

Forensic Bioinformatics Article

http://www.bioforensics.com/articles/empirical_mixtures.pdf
J. Forensic Sci., Nov. 2005, Vol. 50, No. 6
 Paper ID JFS2004435
 Available online at: www.aafm.org

David R. Pasletti¹ M.S., Travis E. Doorn^{1,2} Ph.D., Carissa M. Krane³ Ph.D.,
 Michael L. Raymer^{1,2} Ph.D., and Dan E. Krane⁴ Ph.D.

Empirical Analysis of the STR Profiles Resulting from Conceptual Mixtures

Using 959 complete 13-locus STR profiles from FBI dataset
 146,536,159 possible combinations with 3-person mixtures
3.39% (4,967,034 combinations) would only show a maximum of four alleles (i.e., appear based on maximum allele count alone to be a 2-person mixture)

TABLE 2—Count and percent of three-person mixtures in which a particular number of unique alleles was the maximum observed across all loci, both for the original and randomized individuals*.

Unique Alleles	Count	Percent (%)
2	0	0.00%
3	78	0.00%
4	4,967,034	3.39%
5	93,037,010	63.49%
6	48,532,057	33.12%

Recent Article by Buckleton et al.

Available online at www.sciencedirect.com
 ScienceDirect
 Forensic Science International: Genetics 1 (2007) 20–28
 www.elsevier.com/locate/bscig

Towards understanding the effect of uncertainty in the number of contributors to DNA stains

John S. Buckleton^a, James M. Curran^{b,c}, Peter Gill^c

^aThe Institute of Environmental Science and Research Ltd., Private Bag 92021, Auckland, New Zealand
^bDepartment of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand
^cThe Forensic Science Service, Hirst Centre, Solihull Parkway, Birmingham Business Park, Solihull B37 7YX, UK

Received 31 May 2006; received in revised form 12 September 2006; accepted 13 September 2006

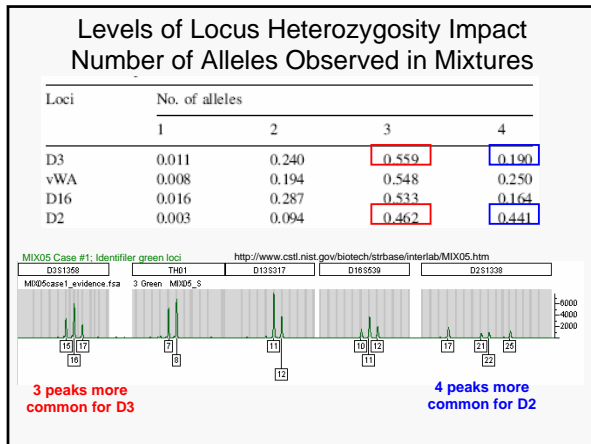
Abstract
 DNA evidence recovered from a scene or collected in relation to a case is generally declared as a mixture when more than two alleles are observed at several loci. However, in principle, all DNA profiles may be considered to be potentially mixtures, even those that show not more than two alleles at any locus. When using a likelihood ratio approach to the interpretation of mixed DNA profiles it is necessary to postulate the number of potential contributors. However, this number is never known with certainty. The possibility of a, say three-person mixture, presenting four or fewer peaks at each locus of the CODIS set was explored by Pasletti et al. (D.R. Pasletti, T.E. Doorn, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, *J. Forensic Sci.* 50 (2005) 1361–1366). In this work we extend this analysis to consider the profile plus and SCM plus mixtures. We begin the assessment of the risk associated with current practice in the calculation of LRs. We open the discussion of possible ways to remove this ambiguity.
 © 2006 Elsevier Ireland Ltd. All rights reserved.

Two-Person Mixtures for Simulated Profiles: Probability by Locus of A Particular Number of Alleles Being Observed

Table 1
 The probability of observing a given number of alleles in a two-person mixtures for simulated profiles at the SGM™ loci

Loci	No. of alleles			
	1	2	3	4
D3	0.011	0.240	0.559	0.190
vWA	0.008	0.194	0.548	0.250
D16	0.016	0.287	0.533	0.164
D2	0.003	0.094	0.462	0.441
D8	0.011	0.194	0.521	0.274
D21	0.007	0.147	0.505	0.341
D18	0.003	0.095	0.472	0.430
D19	0.020	0.261	0.516	0.203
THO	0.016	0.271	0.547	0.166
FGA	0.003	0.116	0.500	0.381

Buckleton et al. (2007) Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *FSI Genetics* 1:20-28



Three-Person Mixtures for Simulated Profiles: Probability by Locus of A Particular Number of Alleles Being Observed

Table 2
 The probability of observing a given number of alleles in a three-person mixtures for simulated profiles at the SGM+™ loci

Loci	No. of alleles showing					
	1	2	3	4	5	6
D3	0.000	0.053	0.366	0.463	0.115	0.002
vWA	0.000	0.037	0.285	0.468	0.194	0.016
D16	0.001	0.086	0.397	0.411	0.100	0.005
D2	0.000	0.008	0.104	0.385	0.393	0.110
D8	0.001	0.041	0.258	0.436	0.236	0.029
D21	0.000	0.023	0.192	0.428	0.302	0.055
D18	0.000	0.007	0.109	0.392	0.396	0.096
D19	0.003	0.078	0.352	0.401	0.152	0.014
THO	0.001	0.074	0.395	0.439	0.088	0.002
FGA	0.000	0.012	0.144	0.424	0.346	0.074

Buckleton et al. (2007) Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *FSI Genetics* 1:20-28

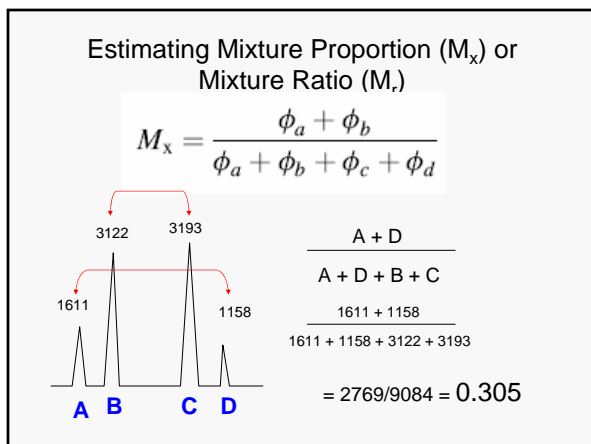
Number of Alleles Observed with Simulated Four-Person Mixtures

- The simulation of four person mixtures suggests that 0.014% of four person mixtures would show four or fewer alleles and that 66% would show six or fewer alleles for the SGM Plus loci.
- The results for the Profiler Plus loci were 0.6% and 75%.
- The equivalent values for the CODIS set from Paoletti et al. were 0.02% showing four or fewer and 76.35% showing six or fewer.

Buckleton et al. (2007) Towards understanding the effect of uncertainty in the number of contributors to DNA stains. *FSI Genetics* 1:20-28

Step #4: Estimation of Relative Ratios for Major and Minor Components to a Mixture

- Mixture studies with known samples have shown that the mixture ratio between loci is fairly well preserved during PCR amplification
- Thus it is generally thought that the peak heights (areas) of alleles present in an electropherogram can be related back to the initial component concentrations
- Start with loci possessing 4 alleles...



Step #5: Consider All Possible Genotype Combinations

Table 3
 Pairwise combinations of two, three and four alleles:

Four alleles (a,b,c,d)	Three alleles (a,b,c)		Two alleles (a,b)	
a,b	a,a	b,c	a,a	a,b
a,c	b,d	b,b	a,c	a,b
a,d	b,c	c,c	a,b	a,a
c,d	a,b	a,c	a,b	b,b
b,d	a,c	b,c	a,c	a,b
b,c	a,d	a,b	b,c	b,b
		b,e	a,a	b,b
		a,e	b,b	
		a,b	c,e	
		a,c	a,b	
		a,e	b,c	
		b,c	a,b	

Key: bold entries represent reciprocal combinations.

Clayton et al. *Forensic Sci. Int.* 1998; 91:55-70

Considering Genotype Combinations

AC
BD
AB
CD
BC
AD

Depends on PHR

Peak Height Ratios (PHR)
Minimum Peak Height (mPH)
Proportion (p) or mixture proportion (M_x)

Step #6: Compare Reference Samples

- If there is a suspect, a laboratory must ultimately decide to include or exclude him...
- **If no suspect is available for comparison, does your laboratory still work the case?** (Isn't this a primary purpose of the national DNA database?)
- Victim samples can be helpful to eliminate their allele contributions to intimate evidentiary samples and thus help deduce the perpetrator

Worked Example

NIJ Expert Systems Testbed (NEST) Project

http://www.promega.com/profiles/1002/ProfilesInDNA_1002_13.pdf
Profiles in DNA (September 2007) 10(2): 13-15

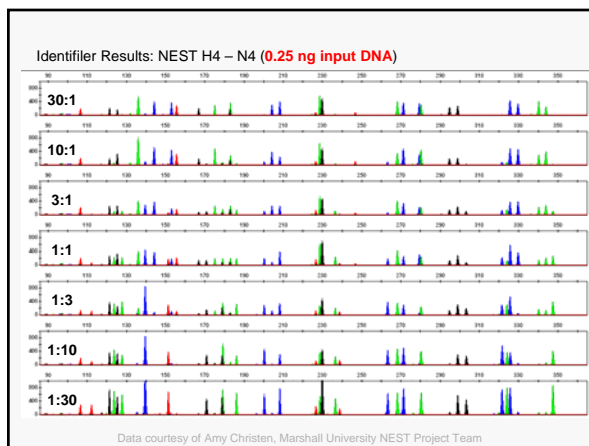
EXPERT SYSTEMS

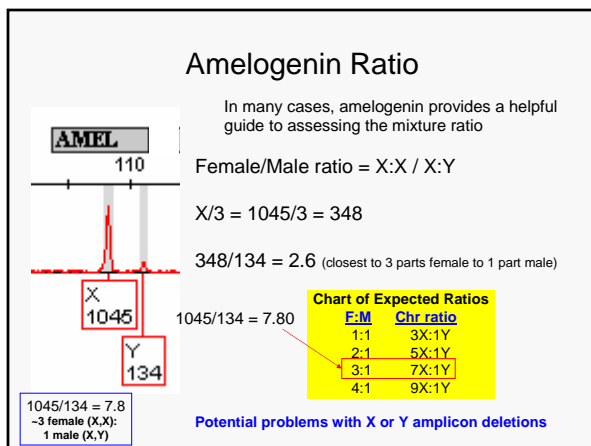
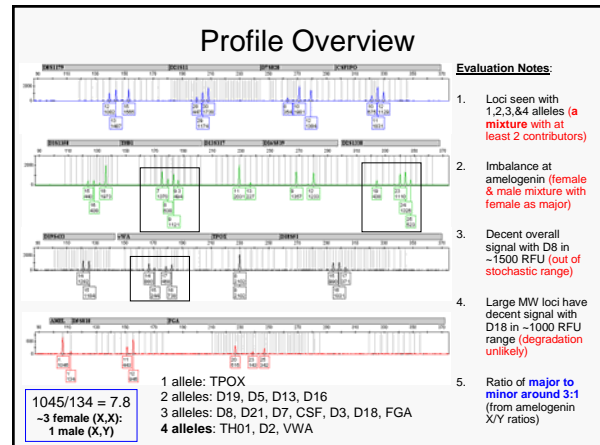
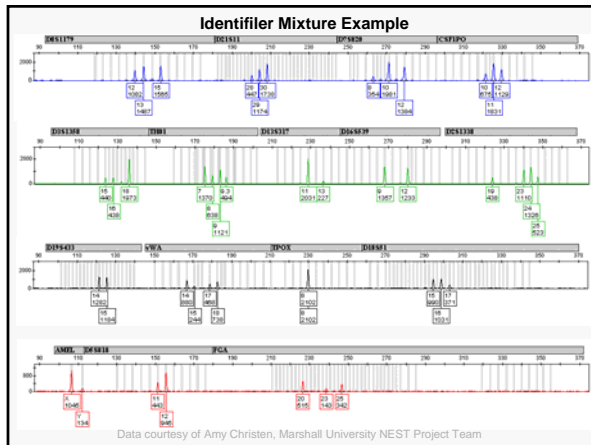
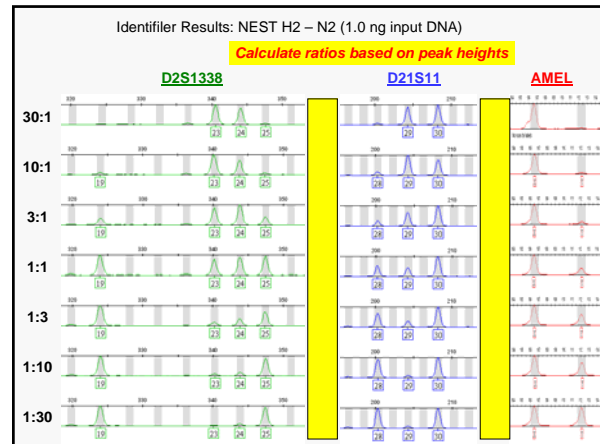
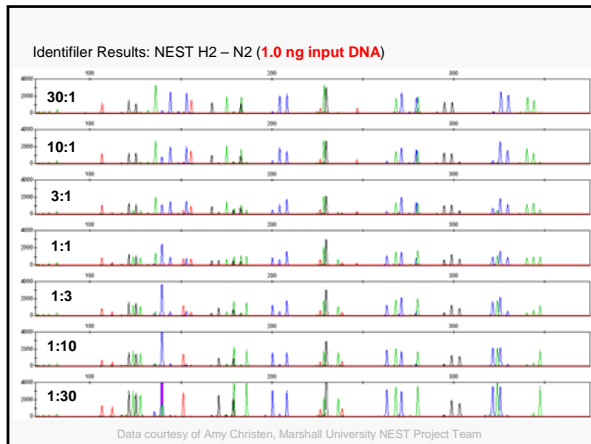
Validating Expert Systems: Examples with the FSS-i3™ Expert Systems Software
By Rhonda K. Roby* and Amy D. Christen†
*Technical Consultant, National Institute of Justice
†Research Analyst, Marshall University

NEST Project Mixture Sample Set

- NIJ Expert Systems Testbed (NEST) Project
 - Marshall University with Rhonda Roby (NIJ consultant)
- Phase II Mixture Sample Analysis
 - **Amy Christen** (Marshall University) produced a dataset while interning at Forensic Science Service in Summer 2006
 - Data to be used for evaluating "expert systems"
- Mixtures tested (280 total samples)
 - **2 different female/male sample combinations:** A:X and B:Y
 - **4 input DNA amounts:** 1.5 ng, 1.0 ng, 0.5 ng, 0.25 ng
 - **5 kits:** Identifiler, ProfilerPlus, COfiler, PowerPlex 16, SGM Plus
 - **7 mixture ratios:** 30:1, 10:1, 3:1, 1:1, 1:3, 1:10, 1:30

I will focus on a subset of this data... e.g., B:Y, 1.0 ng, Identifiler, 3:1





Anomalous Amelogenin Alleles

<http://www.cstl.nist.gov/biotech/strbase/Amelogenin.htm>

- Males possessing only a single X amelogenin amplicon (Y null)** - a male DNA sample will falsely look like a female DNA sample:
 - Santos et al. (1998) reported a rare deletion of the amelogenin gene on the Y-chromosome
 - Y-STR typing can be performed to verify that other portions of the Y-chromosome are present
- Males possessing only a single Y amelogenin amplicon (X null)**:
 - Shewale et al. (2000) observed loss of the X chromosome amplicon in three out of almost 7,000 males examined
 - while this phenomenon should not result in a gender misclassification (as the Y null situation might), its occurrence can impact the expected X and Y amplicon ratios in a mixture (see NIST MIX05 interlab study, case #3)

Running reference samples from suspect and/or victim may help discover potential amelogenin anomalies

Locus-by-Locus Breakdown...

- Start with 4 allele loci...
 - Assume two person mixture
 - With non-overlapping heterozygotes
 - Pair peaks with similar peak heights

Possible Genotype Combinations

See Butler, J.M. (2005) *Forensic DNA Typing, 2nd Edition*, pp. 156-157

- Four Peaks (4 allele loci)**
 - heterozygote + heterozygote, no overlapping alleles (genotypes are unique)
- Three Peaks (3 allele loci)**
 - heterozygote + heterozygote, one overlapping allele
 - heterozygote + homozygote, no overlapping alleles (genotypes are unique)
- Two Peaks (2 allele loci)**
 - heterozygote + heterozygote, two overlapping alleles (genotypes are identical)
 - heterozygote + homozygote, one overlapping allele
 - homozygote + homozygote, no overlapping alleles (genotypes are unique)
- Single Peak (1 allele loci)**
 - homozygote + homozygote, overlapping allele (genotypes are identical)

MUST ALSO CONSIDER STUTTER POSITION

Population Database Used for STR Allele Frequencies

- U.S. population data contained in J.M. Butler (2005) *Forensic DNA Typing, 2nd Edition*, Appendix II (pp. 577-583)
- Published in Butler et al. (2003) *J. Forensic Sci.* 48(4): 908-911
- Available at <http://www.cstl.nist.gov/biotech/strbase/NISTpop.htm>
- Will focus on Caucasians for simplicity

TH01			
Allele	Caucasian N = 302	African-American N = 258	Hispanic N = 140
5	0.00166*	0.00388*	
6	0.23179	0.12403	0.21429
7	0.19040	0.42054	0.27857
8	0.08444	0.19390	0.09643
9	0.11424	0.15116	0.15000
9.3	0.36755	0.10485	0.24843
10	0.00825	0.00194*	0.01429*
11	0.00166*		

Remember that different population databases will have different allele frequencies because they are based on different samples

4 Allele Locus: TH01

Stats

Allele	Frequency
7	0.190
8	0.084
9	0.114
9.3	0.368

$$PI = (P_A + P_B + P_C + P_D)^2$$

$$= (0.190 + 0.084 + 0.114 + 0.368)^2$$

$$= (0.756)^2$$

$$= 0.572$$

Major: 7,9
Minor: 8,9.3

$$PE = 1 - PI = 1 - 0.572 = 0.428$$

Thus ~43% of Caucasian population can be excluded from contributing to this mixture (primarily because allele 6 is missing)

Four Peaks (4 allele loci)
• heterozygote + heterozygote, no overlapping alleles (genotypes are unique)

4 Allele Locus: TH01

PHRs

Consider all possible combinations:

B/A = 638/1370 = 0.466

B/C = 638/1121 = 0.569

C/A = 1121/1370 = 0.818 **major**

D/B = 494/638 = 0.774 **minor**

D/C = 494/1121 = 0.441

Major: 7,9
Minor: 8,9.3

All other combinations <0.60 (60% heterozygote Peak Height Ratio)

Four Peaks (4 allele loci)
• heterozygote + heterozygote, no overlapping alleles (genotypes are unique)

4 Allele Locus: TH01

Mix Ratio

Total of all peak heights
= 1370 + 638 + 1121 + 494
= 3623 RFUs

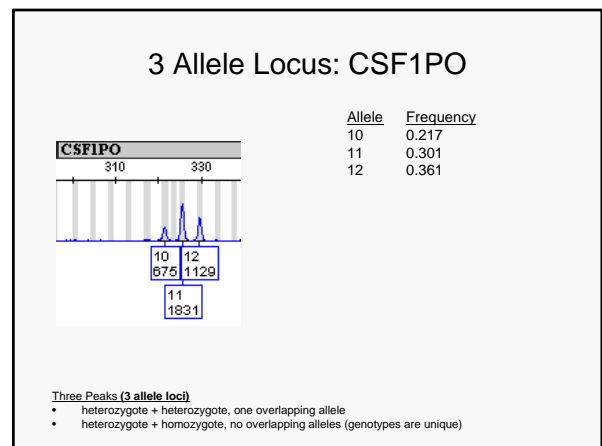
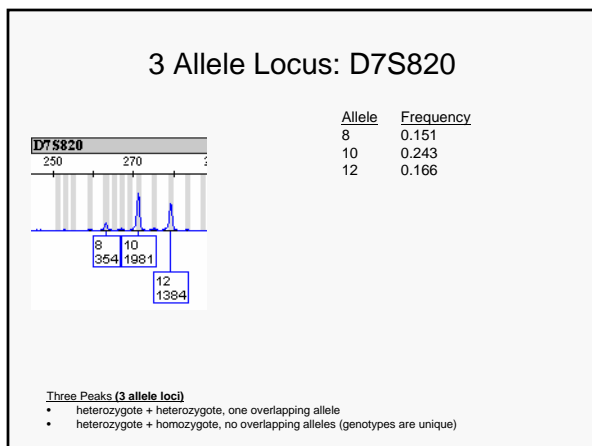
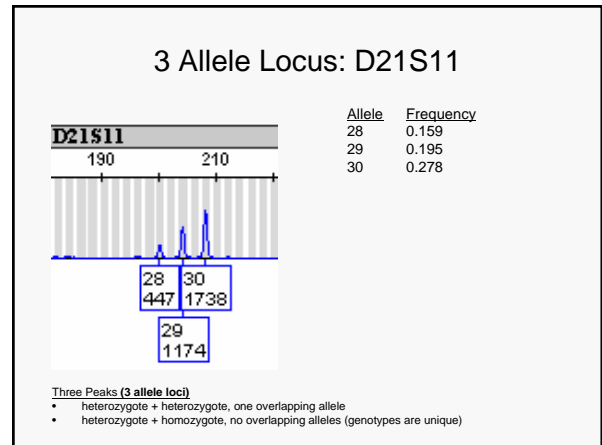
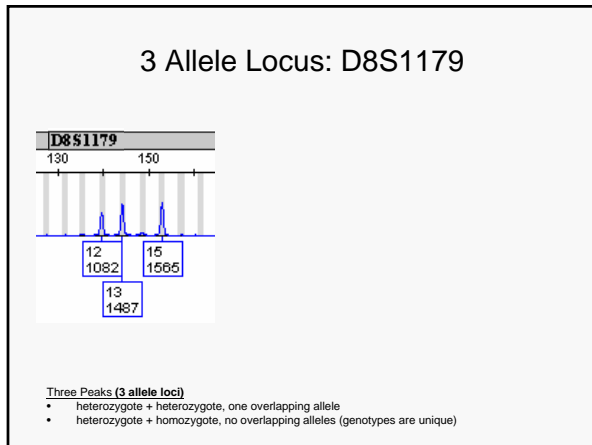
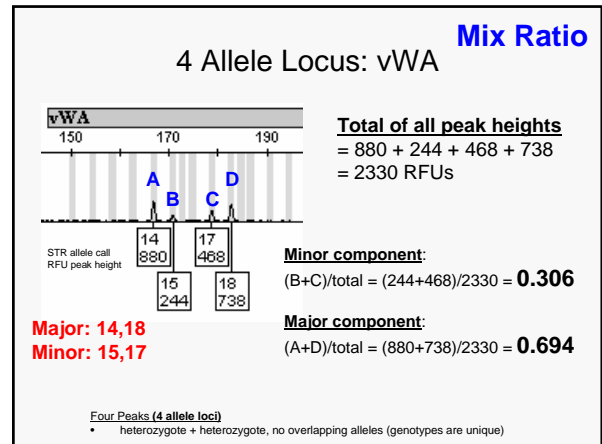
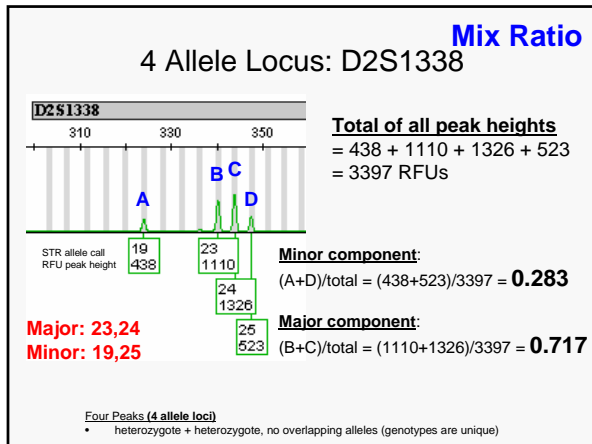
Minor component:
(B+D)/total = (638+494)/3623 = 0.312

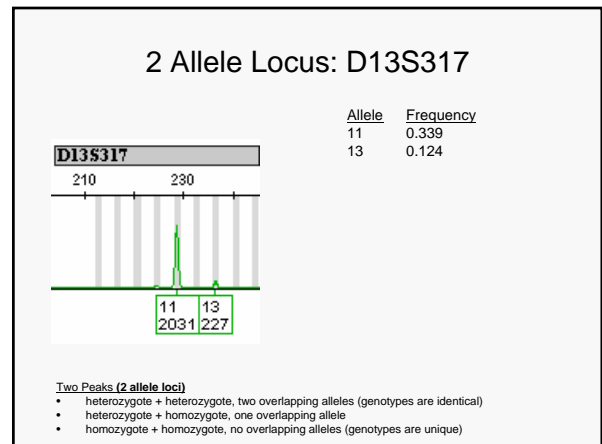
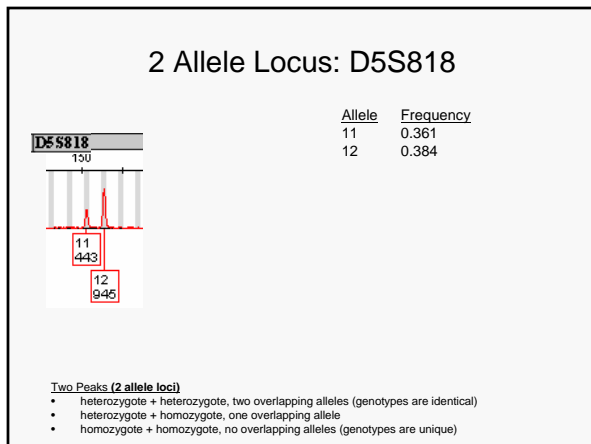
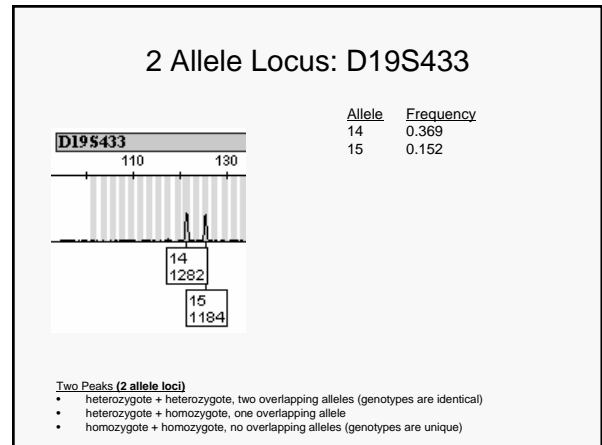
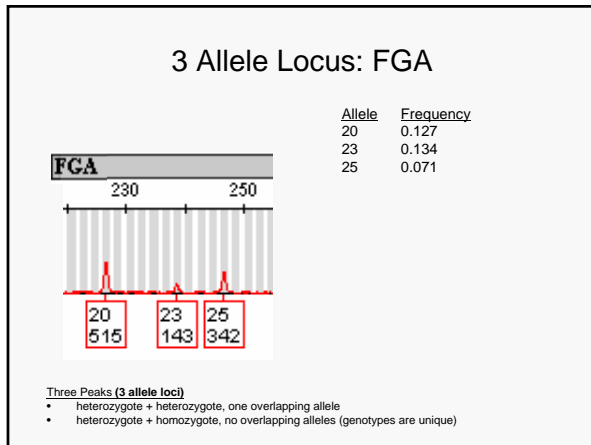
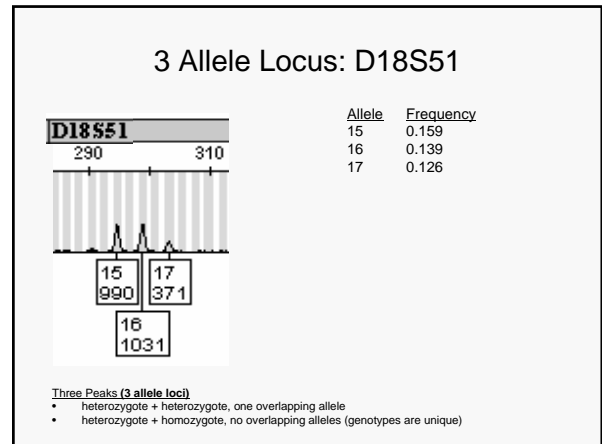
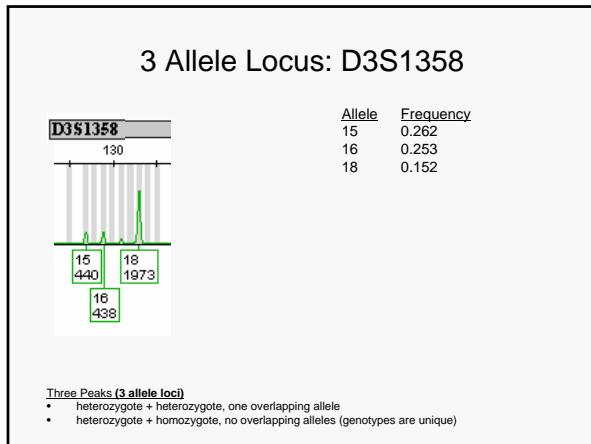
Major component:
(A+C)/total = (1370+1121)/3623 = 0.688

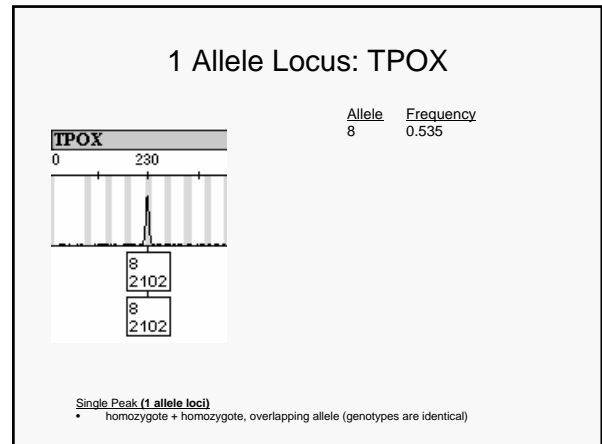
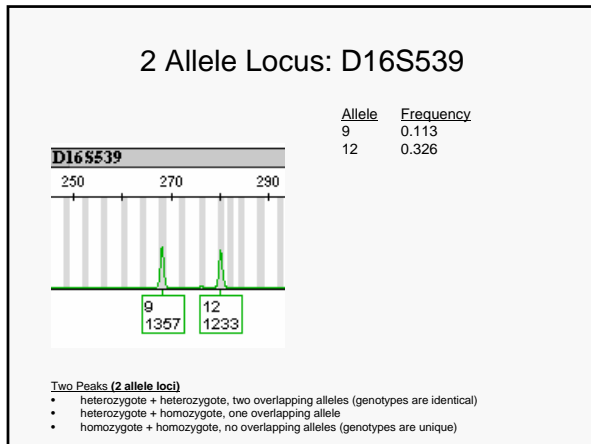
Major: 7,9
Minor: 8,9.3

Close to the ~3:1 predicted by amelogenin X/Y allele ratio – thus major component = female

Four Peaks (4 allele loci)
• heterozygote + heterozygote, no overlapping alleles (genotypes are unique)







Profiles Used In Mixture Samples

	Victim	Suspect
D8S1179	13,15	12,12
D21S11	29,30	28,30
D7S820	10,12	8,10
CSF1PO	11,12	10,11
D3S1358	18,18	15,16
TH01	7,9	8,9,3
D13S317	11,11	11,13
D16S539	9,12	9,12
D2S1338	23,24	19,25
D19S433	14,15	14,15
vWA	14,18	15,17
TPOX	8,8	8,8
D18S51	15,16	16,17
AMEL	X,X	X,Y
D5S818	12,12	11,11
FGA	20,25	20,23

Software Programs (Expert Systems) for Mixture Deconvolution

These programs do not supply stats (only attempt to deduce mixture components)

- Linear Mixture Analysis (LMA)** **U.S. Patent 6,807,490**
 - Part of **TrueAllele** system developed by Mark Perlin (Cybergenetics)
 - Perlin, M.W. and Szabady, B. (2001) Linear mixture analysis: a mathematical approach to resolving mixed DNA samples. *J. Forensic Sci.* 46(6): 1372-1378
- Least Squares Deconvolution (LSD)**
 - Available for use at <https://lsd.lit.net/>
 - Wang, T., Xue, N., Birdwell, J.D. (2006) Least-square deconvolution: a framework for interpreting short tandem repeat mixtures. *J. Forensic Sci.* 51(6):1284-1297.
- PENDULUM**
 - Part of **FSS i-3 software suite (i-STREAM)**
 - Bill, M., Gill, P., Curran, J., Clayton, T., Pinchin, R., Healy, M., and Buckleton, J. (2005) PENDULUM—a guideline-based approach to the interpretation of STR mixtures. *Forensic Sci. Int.* 148(2-3): 181-189

USACIL program developed by Tom Overson called **DNA_DataAnalysis**

Forensic Sci. Int. 2005;148(2-3): 181-189

Available online at www.sciencedirect.com

PENDULUM—a guideline-based approach to the interpretation of STR mixtures

Martin Bill^{a,*}, Peter Gill^b, James Curran^b, Tim Clayton^c, Richard Pinchin^a, Marcus Healy^a, John Buckleton^d

^aThe Forensic Science Service, Trident Court, Solihull Parkway, Birmingham Business Park, Solihull B3773N, UK
^bDepartment of Statistics, University of Waikato, Private Bag 2105, Hamilton, New Zealand
^cThe Forensic Science Service, Southwick Way, Southwick, West Yorkshire, LS227DN, UK
^dCSI, Private Bag 92021, Auckland, New Zealand

Received 7 January 2004; received in revised form 28 May 2004; accepted 1 June 2004

J Forensic Sci. 2006; 51(6):1284-1297

Available for use over internet at <https://lsd.lit.net/>

Tsewei Wang,¹ Ph.D.; Ning Xue,¹ M.Sc.; and J. Douglas Birdwell,² Ph.D.

Least-Square Deconvolution: A Framework for Interpreting Short Tandem Repeat Mixtures^a

- ### Acknowledgments
- Amy Christen (Marshall University NEST Project Team)
 - Angie Dolph (NIST intern/Marshall University)
 - Tim Kalafut (USACIL)