NEST Update

John M. Butler
for the NIST Human Identity Project Team

NEST Meeting at AAFS
Denver, CO
February 17, 2009

Topics to Discuss

- NEST vs. NIST efforts
- Expert system software and tools examined at NIST
- Data sets available
  - Single source
  - Mixtures
- Excel tools developed by Dave Duewer
- Presentations made

NEST vs. NIST Efforts

- **NEST** **evaluates data interpretation software systems and provides training workshops** at Marshall University to provide forensic scientists with exposure to various expert systems
- **NIST** Human Identity Project Team **has examined several software programs in a research context** and given some presentations describing our experiences with several data sets; Excel tools have also been developed to aid data review (concordance studies)

Expert Systems and Data Analysis Tools Examined at NIST

Single Source Samples

- FSS-i3 (i-STRess)
- GeneMapper*ID* v3.2
- OSIRIS

Mixture Samples

- FSS-i3 (i-STReam)
- Web-LSD
- DNA_DataAnalysis
- GeneMapper*ID-X* v1.1

Data Sets Available

Single Source Samples

- Identifiler: >1100
- PowerPlex 16: ~650
- Profiler Plus: ~150
- MiniFiler: >1100
- Yfiler: >1400
- NIST 23/26plex: >1100

Mixture Samples

- MIX05 (2-person)
  - Identifiler, PP16, ProPlus, COfiler, SGM Plus
- 2-person & 3-person
  - Identifiler
- NEST (2-person)
  - Identifiler, PP16, etc.
- ATF (2-person)
  - Identifiler

U.S. Population Samples: 663 males (C, AA, H)
Father-Son Samples: ~800 males (C, AA, H, A)

Mixture Data Sets Examined

(kit, #contributors, input DNA, contributor ratios)

Data from Collaborators

- **ATF** (Identifiler)
  - 2-person, 1ng DNA, 1:40 to 40:1, pristine:pristine and pristine:degraded samples)
- **NEST** (Identifiler, PP16)
  - 2-person, 0.25-1.5ng DNA, 1:30 to 30:1)

Data generated in-house

- NIST MSS3 (Identifiler, PP16)
  - 2-person, 1-4ng DNA, 3:1 to 10:1)
- NIST MSS3 (Identifiler, PP16)
  - 3-person, 3ng DNA, 4:2:1)
- NIST MIX05 (ProPlus)
  - 2-person, 1ng DNA, 1:1 to 1:8)
- NIST additional 1 (Identifiler)
  - 2-person, 1ng DNA, 1:3, 3:1, 1:5, 5:1)
- NIST additional 2 (Identifiler)
  - 3-person, 1ng DNA, 5:2:1)

Sources: ATF data supplied by Todd Bille, NEST data by Amy Christian and Rhonda Roby

Initially purchased in January 2006

Cost to the End User

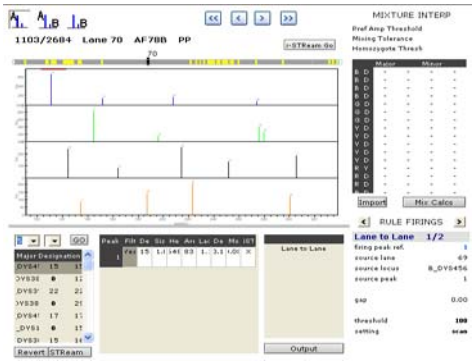
- **Software**
  - Originally v4.0.1 (upgraded to v4.1.3)
  - single copy, single computer **$20,000**
- **Maintenance agreement**
  - **$4,000** per year (20% of total software cost per year, max $15,000)
  - Software upgrades and patches are included
- **Training**
  - **$2,000** if at Promega (plus your travel expenses)
  - $12,000 for up to 5 people if performed in your lab
- Requires GeneMapper ID or GeneScan/Genotyper software to already be in place in your lab

Minimum starting cost of $26,000

STR Data Examined with FSS-i^3

- Identifiler
- PowerPlex 16
- Yfiler
- MiniFiler
- Some in-house assays
- Some mixtures

Spikogram view of Y-filer data



From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Spikogram view of MiniFiler data

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Allele Concordance Studies at NIST with Single Source Samples

- Manual calls
  - with GeneScan/Genotyper v3.7
  - with GeneMapper ID v3.2
- Automated calls with GM/FSS-i^3
- Comparison of output with Excel spreadsheets written by Dave Duewer (NIST)

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Single Source Samples Examined with i-STRess

- We have previously examined **262 Identifiler** samples with v4.0.1 and **656 PowerPlex16** samples with v4.1.3. Excellent concordance was found and the results can be found in past presentations:
  http://www.cstl.nist.gov/biotech/strbase/pub_pres/NIST_FSSi3_Mar2006.pdf
  http://www.cstl.nist.gov/biotech/strbase/pub_pres/PromegaTechTour_NIST_FSSi3.pdf
  http://www.cstl.nist.gov/biotech/strbase/pub_pres/Promega2006_FSSi3.pdf
- In this presentation I will present results from **982 MiniFiler** samples run with v4.1.3.

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Data Comparison Between Methods

- Dave Duewer (NIST Analytical Chemistry Division) has written several computer programs to convert and compare FSS-i³ data that utilize Excel macros
  - DNA_FSSi3_Convert.xls (converts data format)
  - STR_MatchSamples.xls (compares samples)
- These programs are currently available to the community
  - **http://www.cstl.nist.gov/biotech/strbase/software.htm**

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

DNA_FSSi3_Convert.xls

First five columns in FSS-i³ output are converted to be like GeneMapper ID allele designation table

Batch ID	Sample ID	Locus ID	Major Designation 1	Major Designation 2
MiniFiler RAW AA	GT37019	D13S317	11	12
MiniFiler RAW AA	GT37019	D7S820	10	11
MiniFiler RAW AA	GT37019	AMEL	X	Y
MiniFiler RAW AA	GT37019	D2S1338	16	
MiniFiler RAW AA	GT37019	D21S11	28	
MiniFiler RAW AA	GT37019	D16S539	11	
MiniFiler RAW AA	GT37019	D18S51	17	
MiniFiler RAW AA	GT37019	CSF1PO	11	12
MiniFiler RAW AA	GT37019	FGA	23	24

Each row is an individual locus

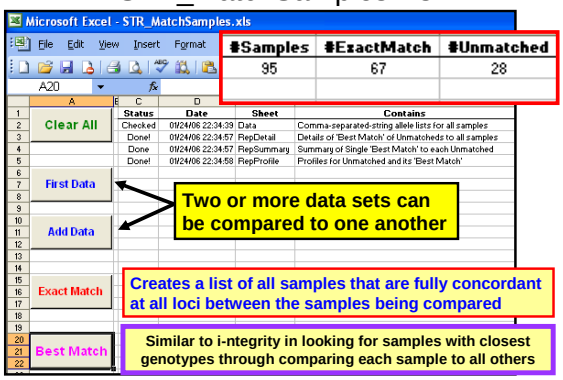
Data Transformation

SampleCode	D13S317	D13S317	D7S820	D7S820	AMEL	AMEL	D2S1338	D2S1338	D21S11	D21S11	D16S539	D16S539	D18S51	D18S51	CSF1PO	CSF1PO	FGA	FGA
GT37019	11	12	10	11	X	Y	16	17	28	30	11	11	17	19				
GT37020	11	13	9	11	X	Y	19	26	28	28	11	12	16	16				
GT37026	12	12	9	10	X	Y	17	21	28	28	11	12	12	17				
GT37027	12	13	8	10	X	Y	17	20	29	31	9	12	15	17				
GT37032	11	12	8	8	X	Y	20	22	29	31	9	11	17	19				

Each row is an individual sample

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

STR_MatchSamples.xls



From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Exact Matches (Full Concordance) Observed with STR_MatchSamples.xls Program

Sample Type	Sample Description	Penta_D	Penta_E	TH01	TPOX	vWA
Unmatched	GT37019:all data	2.2,11	5,13	6,7	8,9	17
Unmatched	GT37019:all PP16 GM samples	2.2,11	5,13	6,7	8,9	14,17
ExactMatch	BC11352:all data, BC11352:all PP16 GM samples	10,11	7,12	6,9.3	8	14,17
ExactMatch	GA05070:all data, GA05070:all PP16 GM samples	13,14	7,17	7,9	8,12	14,19
ExactMatch	GA05071:all data, GA05071:all PP16 GM samples	10,11	11,12	7,9.3	8,11	16,17
ExactMatch	GC03394:all data, GC03394:all PP16 GM samples	10,11	12,15	6,7	8	17,18

- Unmatched sample type flags discordant calls
- ExactMatch sample type indicates full concordance between FSS-i³ and GeneMapper ID samples

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Concordance Evaluation

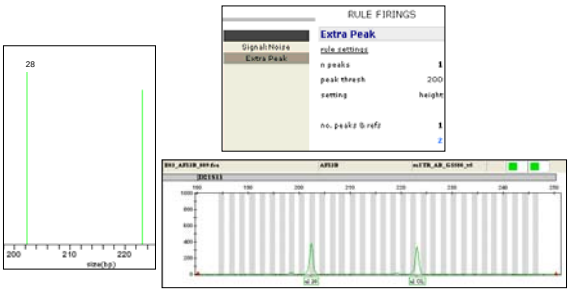
- MiniFiler collected on ABI 3130xl; 982 samples processed in GeneMapper ID and FSS-i³
- Typed manually with GeneMapper ID
- Same data processed through GeneMapper ID/FSS-i³
- When rules were fired, profiles were reviewed
- Results from **982 samples** compared with STR_MatchSamples.xls
- **Examination of mismatches to determine which rules were fired and if user would be able to make correct calls following editing: All calls were concordant after review**

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Example 1- Microvariant

Unedited FSS-i^3 Data:

D21S11 Allele 33.1



No virtual bin for this microvariant was entered into software when MiniFiler was created therefore this allele was not being called in FSSi^3

In Summary

- FSS-i^3 has the capability to create new multiplex kits (Y-filer, PowerPlex Y, MiniFiler and custom assays)
- Dave Duewer software programs are currently available on STRBase: http://www.cstl.nist.gov/biotech/strbase/software.htm
- A total of **2162** profiles have been analyzed using the FSS-i^3 software with concordance checks performed. Full concordance has been achieved after careful review.
- In general, FSS-i^3 i-STReam is conservative in its mixture deconvolution; however, **26** out of **4080** allele calls were called incorrectly (**0.64%**).

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

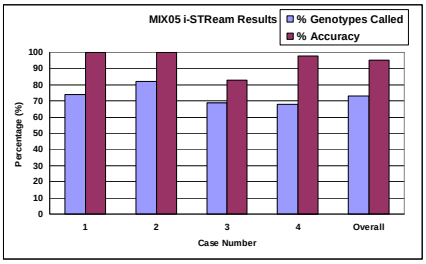
i-STReam Results

Case 1
3 female : 1 male

Case 2
1 female : 3 male

Case 3
1 female : 1 male

Case 4
7 female : 1 male



i-STReam
Stand-alone version

Case #	Genotypes Called	Accuracy
1	74	100
2	82	100
3	69	83
4	68	98
Overall	73	95

From Becky Hill presentation for Promega Summer Technology Tour – August 2007

Mixtures

Mixture Analysis at NIST

Amy Decker | John Butler | Angie Dolph (summer 2007) | Michelle Burns (summer 2008+)

- AAFS Feb 2008 Poster
  - Angie Dolph: "The DNA Mixture Conundrum: Sample Variation and Its Effects on Mixture Deconvolution Tools"
- Promega Oct 2008 Poster
  - Amy Decker: "Determining Contributor Profiles from DNA Mixtures of Varying Ratios"

Some Recent Mixture Workshops

60 DNA analysts from 16 labs (all MD labs, DE, NMS)

25 DNA analysts from AFDIL

Mixture Interpretation and Other Topics January 27-28, 2009
50 DNA analysts from Harris County, Texas (HPD, MEO, TX DPS)

Mixture Analysis Efforts at NIST

- Interlaboratory Studies: MSS1,2,3 and MIX05
  - **Future ones planned** when software tools and guidelines are available
- Software testing (see posters from AAFS 2008 and Promega 2008)
  - DNA_DataAnalysis (USACIL) – user's manual written
  - FSS-i3 (Promega)
  - Web-LSD (UTenn)
  - GeneMapper *ID-X* v1.1 (ABI)
  - GenoProof Mixture 1.0 (Qualitype)

  Some conversations with Mark Perlin regarding TrueAllele 3 software
  Some work coordinated with NEST Project (Marshall University)
- Work with SWGDAM Mixture Committee
  - Case summaries
- Training workshops and discussion groups
  - AAFS Feb 2008, MD Apr 2008, FDLE May 2008, CE Users Dec 2008, AFDIL Jan 2009, Harris County, TX Jan 2009

NIST Interlaboratory Mixture Studies

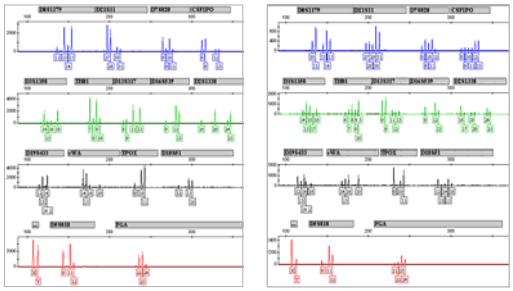
http://www.cstl.nist.gov/biotech/strbase/interlab.htm

- **Provide a big-picture view of the community** (not graded proficiency tests)
  - offers laboratories an opportunity to directly compare themselves to others in an anonymous fashion
- Some lessons learned:
  - instrument sensitivities can vary significantly
  - amount of input DNA plays important role in ability to detect minor component(s)
  - **protocols and approaches are often different between forensic DNA labs**
- Studies Conducted
  - **Mixed Stain Study #1 (MSS1)** – Apr-Nov 1997
    (6 single-source, 4 two-source, 1 three-source stains)
  - **Mixed Stain Study #2 (MSS2)** – Jan-May 1999
    (4 single-source, 1 two-source, 1 three-source stains)
  - **Mixed Stain Study #3 (MSS3)** - Dec 2000-Oct 2001
    (1 single-source, 5 two-source, 1 three source DNA extracts)
  - **Mixture Interpretation Study (MIX05)** – Jan-Aug 2005
    (4 two-person mixture "case" data with victim profiles supplied – *data only*)

Creating Known Mixtures for Testing Software Tools

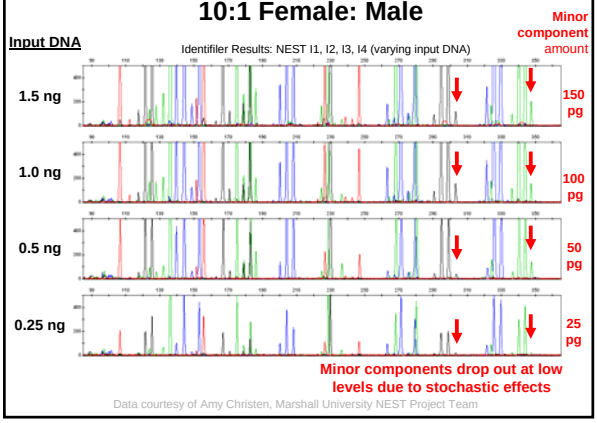
NIST 2-person mixture
(Identifiler data, 1ng DNA, **1:5**)

NIST 3-person mixture
(Identifiler data, 1ng DNA, **5:2:1**)



Mixtures were created for research purposes and are synthetic mixtures of extracted DNA created in a controlled environment without PCR inhibitors or an unknown amount of degraded DNA as may be found in forensic casework.

10:1 Female: Male



Data courtesy of Amy Christen, Marshall University NEST Project Team

Mixture Case Summaries

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 **4780 samples**. A preliminary summary of this information is shown below 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.**

	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
Single source	54.8%	33.6%	10.3%	1.2%	0.1%	mixtures

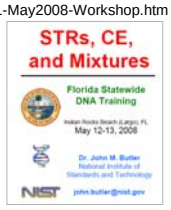
http://www.cstl.nist.gov/biotech/strbase/pub_pres/Promega2008poster.pdf

CFS Toronto Case Summary Data

Case type	N = 276	1	# contributors 2	3	4	>4
Sexual Assault	N = 152	42%	52%	7%	1%	--
High Volume	N = 56	69%	16%	16%	--	--
Major Crime	N = 68	59%	34%	7%	--	--
		Single source	Mixtures			

Mixture Questionnaire Summaries

…strbase\training\FL-May2008-Workshop.htm



42 participants from 13 different labs

28 responses

>80 analysts from 10 different FL labs

48 responses

76 responses representing >20 different laboratories

20 Questions Asked

Interpretation Guidelines
11 questions including:
What would you like to see in national guidelines on how to perform DNA mixture interpretation and statistical analysis?

Validation and Training
4 questions including:
What kind of training materials would be beneficial to help your laboratory more effectively solve mixtures?

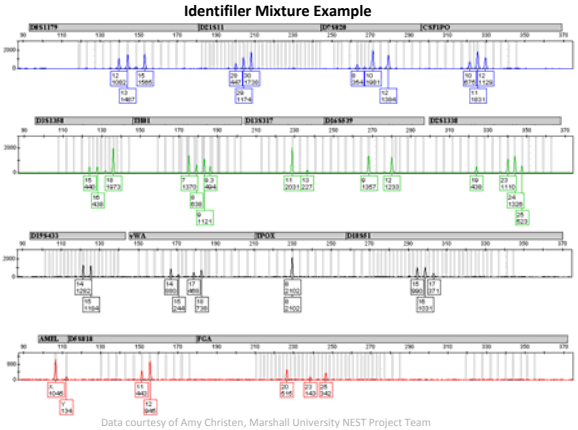
Other Topics
5 questions including:
What are the biggest obstacles you face in your lab in terms of mixture interpretation?

CE User's Group (December 5, 2008)

- Bruce Heidebrecht organized
- Held at Maryland State Police Forensic Lab
- Presentations & discussion on 4 mixture cases
- ~60 people attended from 16 labs
- Bruce has developed several helpful tools for mixtures…

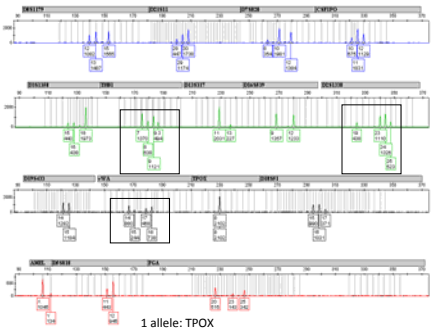
Amy Decker reviewed some of these mixture cases as worked examples

Identifiler Mixture Example



Data courtesy of Amy Christen, Marshall University NEST Project Team

Profile Overview



Evaluation Notes:

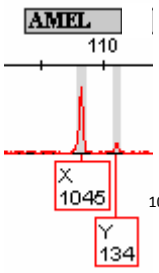
1. Loci seen with 1,2,3,&4 alleles (**a mixture** with at least 2 contributors)
2. Imbalance at amelogenin (female & male mixture with female as major)
3. Decent overall signal with D8 in ~1500 RFU (out of stochastic range)
4. Large MW loci have decent signal with D18 in ~1000 RFU range (degradation unlikely)
5. Ratio of **major to minor around 3:1** (from amelogenin X/Y ratios)

1045/134 = 7.8
~3 female (X,X):
1 male (X,Y)

1 allele: TPOX
2 alleles: D19, D5, D13, D16
3 alleles: D8, D21, D7, CSF, D3, D18, FGA
4 alleles: TH01, D2, VWA

Amelogenin Ratio

In many cases, amelogenin provides a helpful guide to assessing the mixture ratio



Female/Male ratio = X:X / X:Y

X/3 = 1045/3 = 348

348/134 = 2.6 (closest to 3 parts female to 1 part male)

1045/134 = 7.80

Chart of Expected Ratios

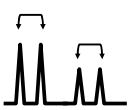
F:M	Chr ratio
1:1	3X:1Y
2:1	5X:1Y
3:1	7X:1Y
4:1	9X:1Y

1045/134 = 7.8
~3 female (X,X):
1 male (X,Y)

Potential problems with X or Y amplicon deletions

Locus-by-Locus Breakdown…

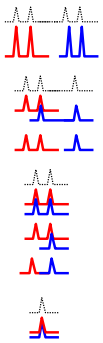
- Start with 4 allele loci…
  - Assume two person mixture
  - With non-overlapping heterozygotes
  - Pair peaks with similar peak heights



Possible but not as likely depending on ratios

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

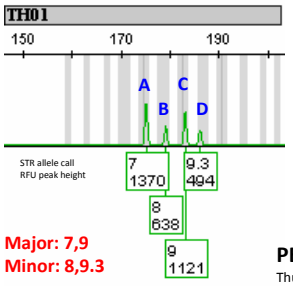
TH01

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	0.19380	0.09643
9	0.11424	0.15116	0.15000
9.3	0.36755	0.10465	0.24643
10	0.00828	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

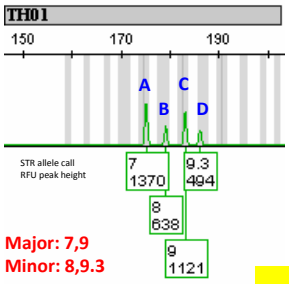
PE = 1 – PI = 1 – 0.572 = 0.428

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 = 494/648 = 0.774 **minor**

D/C = 494/1121 = 0.441

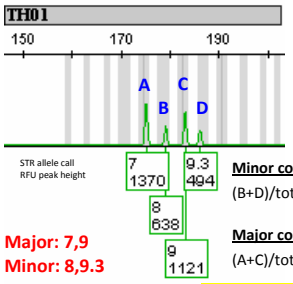
Major: 7,9
Minor: 8,9.3

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**

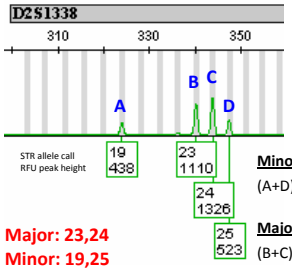
Major: 7,9
Minor: 8,9.3

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)

4 Allele Locus: D2S1338

Mix Ratio



Total of all peak heights
= 438 + 1110 + 1326 + 523
= 3397 RFUs

Minor component:
(A+D)/total = (438+523)/3397 = **0.283**

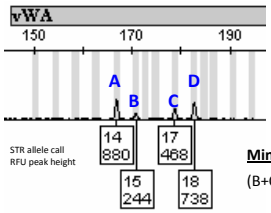
Major component:
(B+C)/total = (1110+1326)/3397 = **0.717**

Major: 23,24
Minor: 19,25

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

4 Allele Locus: vWA

Mix Ratio



Total of all peak heights
= 880 + 244 + 468 + 738
= 2330 RFUs

Minor component:
(B+C)/total = (244+468)/2330 = **0.306**

Major component:
(A+D)/total = (880+738)/2330 = **0.694**

Major: 14,18
Minor: 15,17

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

NIST Human Identity Project Team

…Bringing traceability and technology to the scales of justice…


John Butler
Group Leader


Amy Decker


Becky Hill


Margaret Kline


Jan Redman


Pete Vallone


Dave Duewer
(data analysis)


Angie Dolph
(summer 2007)


Michelle Burns
(summer 2008+)

Current Collaborators
Mike Coble (AFDIL) **Tom Hall** (IBIS)
Bruce McCord (FIU) **Danielle Podini** (GWU)
Tom Reid (DDC) **Lisa Forman-Neall** (NCBI)
Manfred Kayser (The Netherlands)

Funding from the **National Institute of Justice (NIJ)**
through NIST Office of Law Enforcement Standards

Our team publications and presentations are available at:
http://www.cstl.nist.gov/biotech/strbase/NISTpub.htm

Thank you for your attention…

Funding from the **National Institute of Justice (NIJ)**
through NIST Office of Law Enforcement Standards

http://www.cstl.nist.gov/biotech/strbase

Questions?

john.butler@nist.gov
301-975-4049


Margaret Kline


Pete Vallone


Jan Redman


Amy Decker


Becky Hill


Dave Duewer

Summer Interns
Angie Dolph ('07)
Angela Gorman ('07)
Michelle Burns ('08)

Collaborators
Mike Coble (AFDIL)
Bruce McCord (FIU)
Tom Reid (DDC)