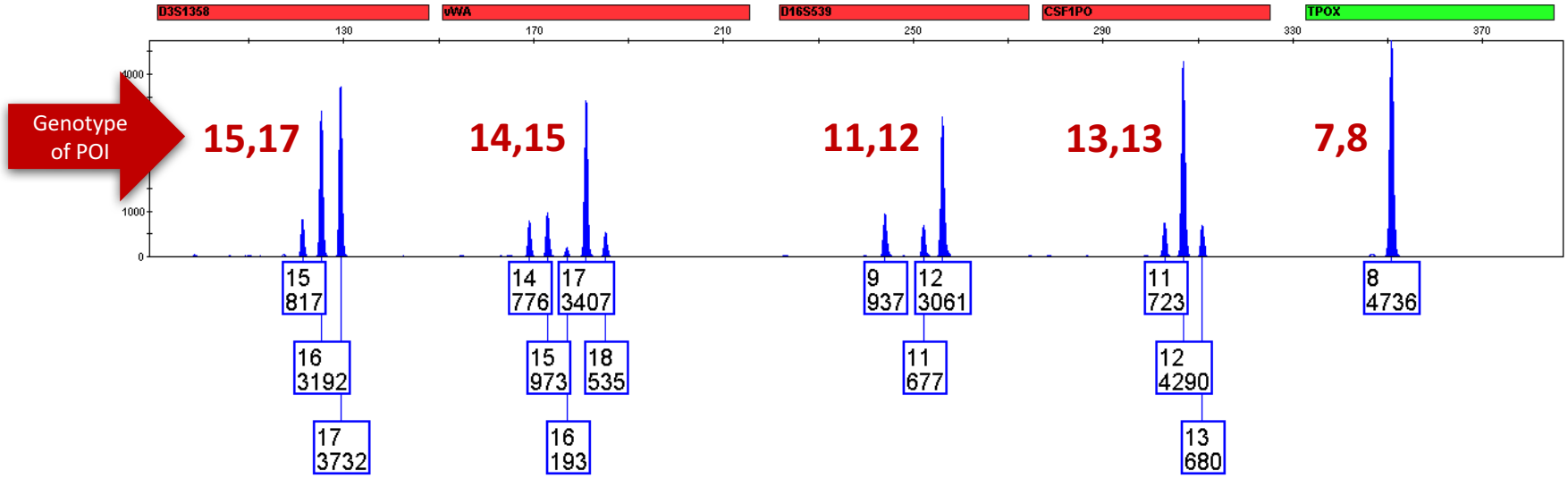
Genotyping Errors in the FBI STR Allele Frequency Database Used for Estimating Match Probabilities in Forensic Investigations

Tamyra Moretti, Lilliana Moreno and Anthony Onorato
FBI Laboratory

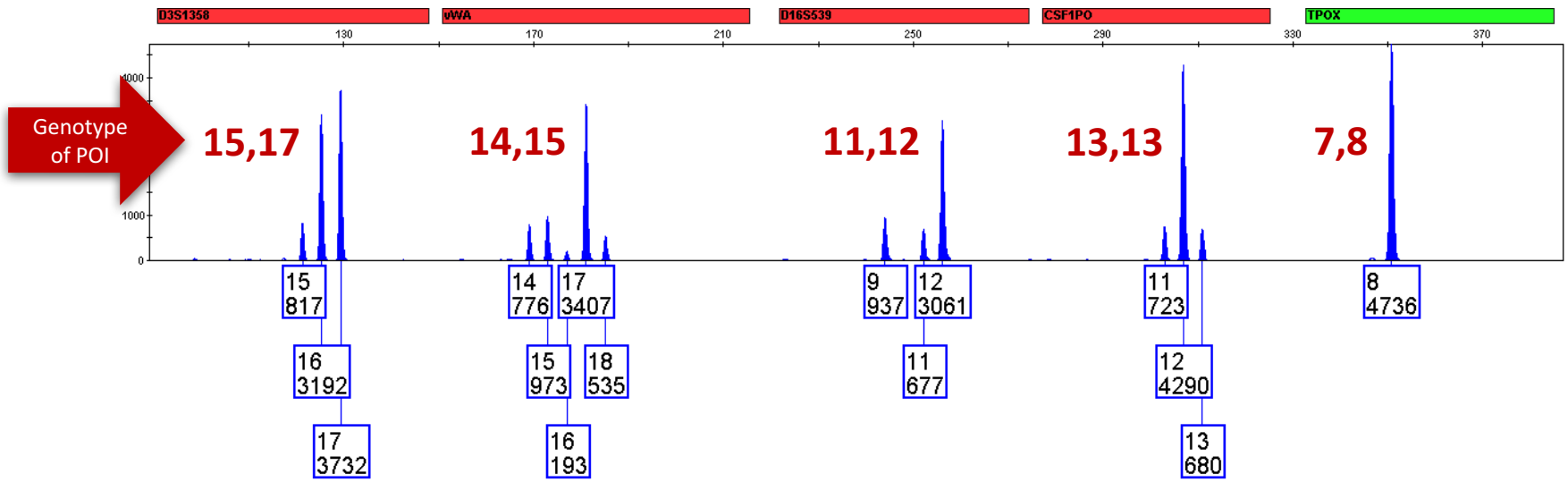
2017 International Symposium on
Forensic Science Error Management
Gaithersburg, MD

When a Person of Interest cannot be excluded as a potential contributor of the DNA obtained from evidence…



- A profile probability is calculated to estimate the statistical weight of the evidence
- The calculation incorporates frequency estimates for the observed alleles

When a Person of Interest cannot be excluded as a potential contributor of the DNA obtained from evidence…



- Such allele frequencies have been obtained from various population samples, permitting profile probabilities to be calculated for different population groups since the reference population of the true contributor to the evidence is unknown

FBI Population Groups

- Major populations
  - African American, Caucasian, Southeast Hispanic, Southwest Hispanic
- Additional populations
  - Native American – Apache, Navajo, Minnesota Native American
  - Caribbean Islands – Bahamian, Jamaican, Trinidadian
  - Guam – Filipino, Chamorro

- Allele frequencies from these populations have been used since 1999 by the FBI and other laboratories for calculating match statistics in criminal investigations and other human identity testing applications

Autosomal STR Amplification Kits Tested by the FBI Laboratory

Early Progenitor Kits	*Core CODIS 13 Kits*	*Single-Amp Kits*	*Expanded CODIS Loci Kits*
GenePrint® FFFL, CTT, CTTv, GammaSTR® *AmpFlSTR®* Blue, Green I, Green II, Yellow	*GenePrint®* PowerPlex™ 1.1, 1.2, 2.1 *AmpFlSTR®* Profiler™, Profiler™ Plus, COfiler™	*GenePrint®* PowerPlex™ 16, 16HS *AmpFlSTR®* Identifiler™, Identifiler™ Plus	PowerPlex™ Fusion Globalfiler™

Since the development in the late 1990s of the original STR typing systems for the 13 core CODIS STR loci, new test kits that expand the number of loci to 24-27 are now commercially available & required of NDIS laboratories as of January 2017 for typing the CODIS 20 Core STR loci.

1998 – 2001

Budowle *et al*. (1999a) J Forensic Sci 44(6):1277-86

Budowle *et al*. (2001a) Forensic Sci Comm 3(3)

Budowle *et al*. (2001a) J Forensic Sci 46(3):453-89

Budowle *et al*. (2000) Legal Med 2:26-30

Population Sample (data source)	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820	CSF1PO	TPOX	TH01	D16S539	D10S1248	D22S1045	D2S441	D15S1656	D12S391	D2S1338	D19S433	DYS391	SE33	Penta D	Penta E
Caucasian (FBI)																								
African American (FBI)																								
Southwest Hispanic (FBI)																								
Southeast Hispanic (FDLE, PBSO, Metro-Dade/Miami Childrens Hospital)																								
Bahamian (FBI)																								
Jamaican (FBI)																								
Trinidadian (FBI)																								
Apache (Arizona DPS)																								
Navajo (Arizona DPS)																								
Minnesota Native American (Minnesota BCA)																								
Filipino (FBI & Applied Biosystems)																								
Chamorro (FBI & Applied Biosystems)																								

Budowle & Moretti (1999b) Forensic Sci Comm 1(2) | Budowle *et al*. (2001b) Forensic Sci Comm 3(3)

Electronic genotype data are available in the cited online references

Population Sample (data source)	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820	CSF1PO	TPOX	TH01	D16S539	D10S1248	D22S1045	D2S441	D15S1656	D12S391	D2S1338	D19S433	DYS391	SE33	Penta D	Penta E
Caucasian (FBI)																								
African American (FBI)																								
Southwest Hispanic (FBI)																								
Southeast Hispanic (FDLE, PBSO, Metro-Dade/Miami Childrens Hospital)																								
Bahamian (FBI)																								
Jamaican (FBI)																								
Trinidadian (FBI)																								
Apache (Arizona DPS)																								
Navajo (Arizona DPS)																								
Minnesota Native American (Minnesota BCA)																								
Filipino (FBI & Applied Biosystems)																								
Chamorro (FBI & Applied Biosystems)																								

2014 – 2015

• Moretti *et al*. (2016) Forensic Science International: Genetics 25:175-181

Population Sample (data source)	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820	CSF1PO	TPOX	TH01	D16S539	D10S1248	D22S1045	D2S441	D15S1656	D12S391	D2S1338	D19S433	DYS391	SE33	Penta D	Penta E
Caucasian (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
African American (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Southwest Hispanic (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Southeast Hispanic (FDLE, PBSO, Metro-Dade/Miami Childrens Hospital)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Bahamian (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Jamaican (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Trinidadian (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Apache (Arizona DPS)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Navajo (Arizona DPS)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Minnesota Native American (Minnesota BCA)																								
Filipino (FBI & Applied Biosystems)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Chamorro (FBI & Applied Biosystems)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Electronic genotypes from the1999/2001 & 2014 typing were assessed for concordance

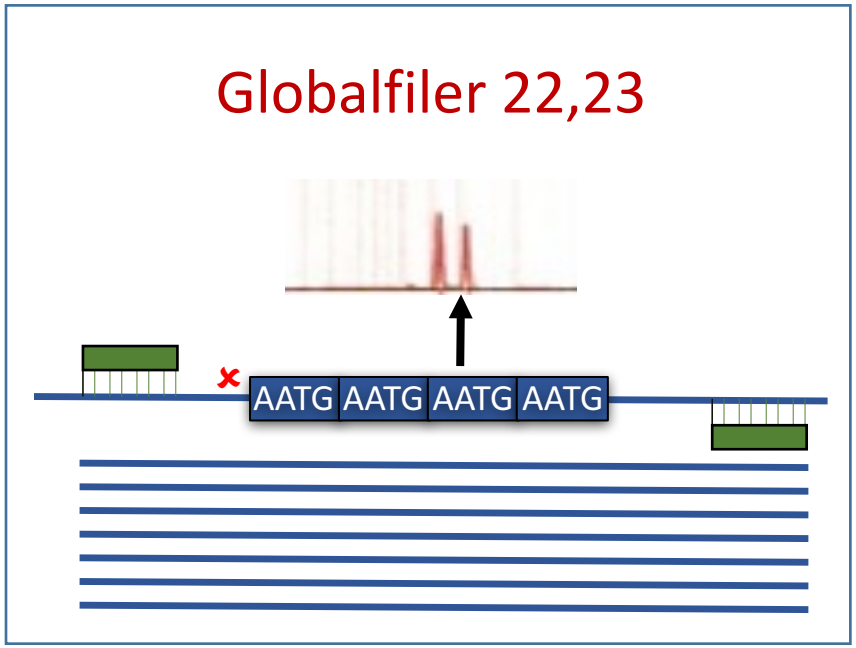
1175 samples
~30,000 alleles

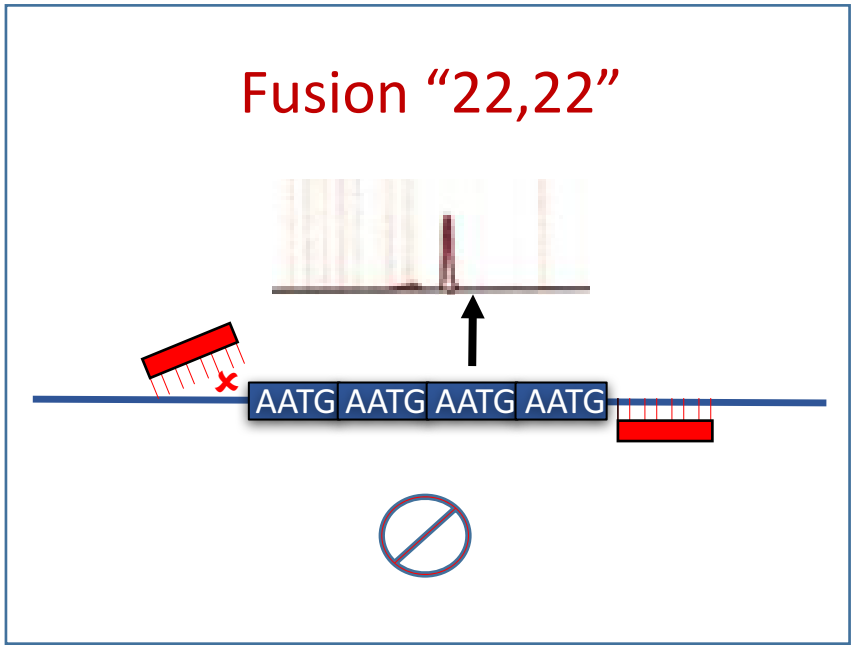
Pop sou	D3S1358	vWA	FGA	D8S1179	D21S11	D18S51	D5S818	D13S317	D7S820	CSF1PO	TPOX	TH01	D16S539	D10S1248	D22S1045	D2S441	D15S1656	D12S391	D2S1338	D19S433	DYS391	SE33	Penta D	Penta E
Caucasian (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
African American (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Southwest Hispanic (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Sou (FDL Child	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Bahamian (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Jamaican (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Trinidadian (FBI)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Apa (Ariz	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Nav (Ariz	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Min (Min																								
Filipino (FBI & Applied Biosystems)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Chamorro (FBI & Applied Biosystems)	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

We expected to identify a few rare, normal genetic variants

When two kits have different primers for amplifying a given locus, an allele may fail to amplify detectably in one kit due to a variant in the DNA sequence that impedes typing of the allele

Same sample typed with different kits, exhibiting drop-out of allele #23 with one kit





Null alleles due to primer binding site variants

Kit	Locus	Undetected allele	Population	Kit(s) showing allele
GlobalFiler	D12S391	21	Southeast Hispanic	Fusion
GlobalFiler	D12S391	23	Southwest Hispanic	Fusion
Fusion	D13S317	8	Southeast Hispanic	GlobalFiler
GlobalFiler	D13S317	13	Filipino	Fusion
Fusion	D16S539	9	Southeast Hispanic	GlobalFiler
PowerPlex 1.1	D16S539	10	Jamaican	GlobalFiler, Fusion
Fusion	D16S539	11	Filipino	GlobalFiler
PowerPlex 1.1	D16S539	12	Jamaican	GlobalFiler, Fusion
GlobalFiler	D1S1656	15	African American	Fusion
Fusion	D22S1045	14	Southwest Hispanic	GlobalFiler
Profiler Plus	FGA	22	African American	GlobalFiler, Fusion
Identifiler Plus	vWA	18	Southeast Hispanic	GlobalFiler, Fusion

However…

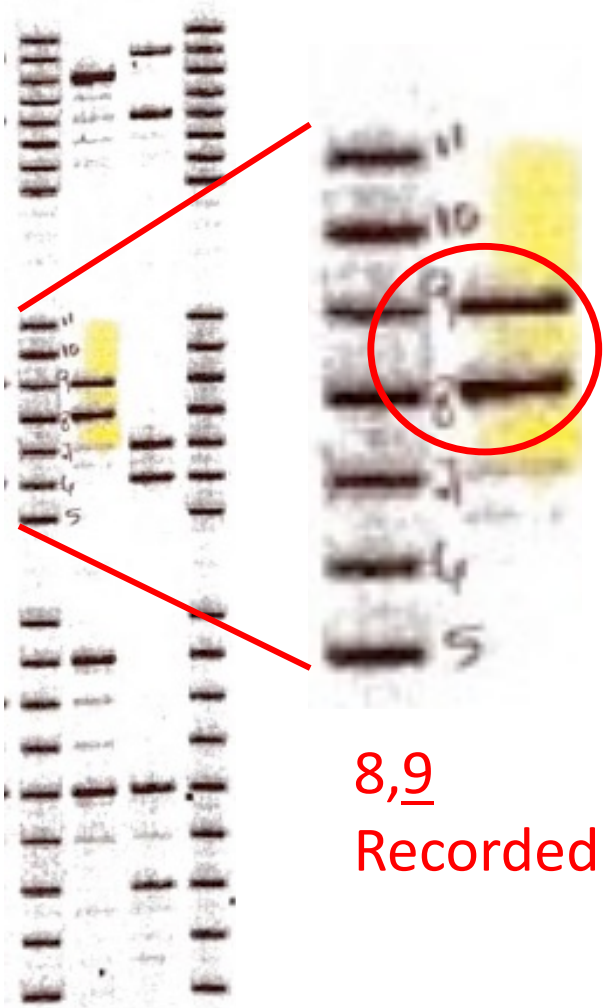
Comparison of the original and new data for the same samples revealed other genotyping discrepancies that were determined to be errors in the original population dataset

There are two general categories of genotyping errors:

- Clerical errors
  - Due to manual data recording and data manipulation
- Errors due to technological limitations
  - Inherent to the STR typing systems and data analysis software available in the 1990s

Examples of Clerical Errors

Manual data analysis with transcription error



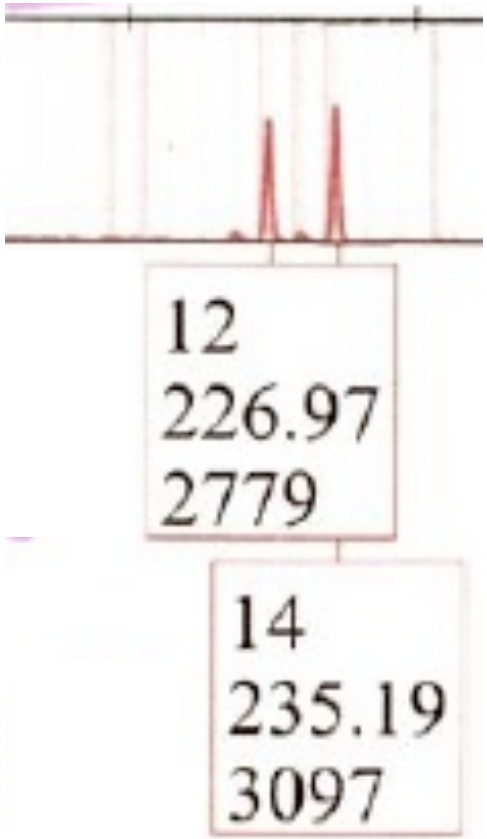
8,9

Recorded as 8,10

1CH 505 FL	Penta E	D18S51	D21S11	TH01	D3S1358
C180	15-20	12-19	31.2-32.2	9-9.3	15-16
C181	7 12	15-17	27-32.2	7-9	16-16
C182	7-12	14-18	30-32.2	6-8	15-17
C183	11-14	13-16	30-31.2	7-9.3	15-18
C184	5-16	13-15	28-30	9.3-9.3	14-18
C185	10-13	14-15	28-30	7-7	14-18
C186	11 17	14-15	28-30	6-9.3	14-15
C187	13-14	15-17	27-32	7-9	15-18
C188	10-17	14-15	29-31	6-9.3	14-17
C189	10-17	14-15	29-31	6-9.3	14-17

Data were recorded manually and manually tallied or hand-typed into spreadsheets for population genetic analyses

Software-assisted analysis with transcription error



12,14

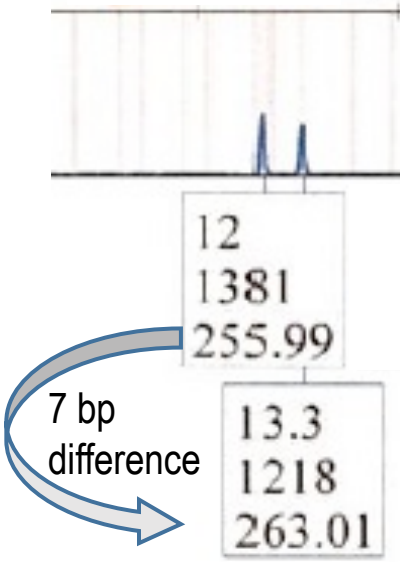
Recorded as 8,12

Examples of Technological Limitations

Then…12,14

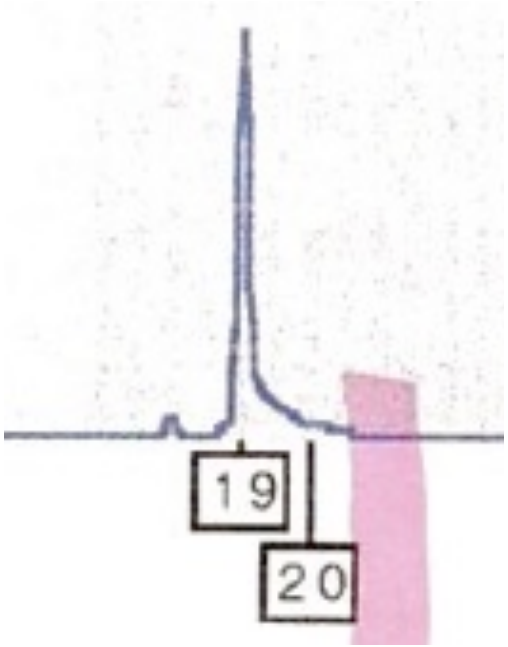
Now…12,13.3

Difficulty in distinguishing a microvariant allele using an early separation & detection method (polyacrylamide gel electrophoresis)

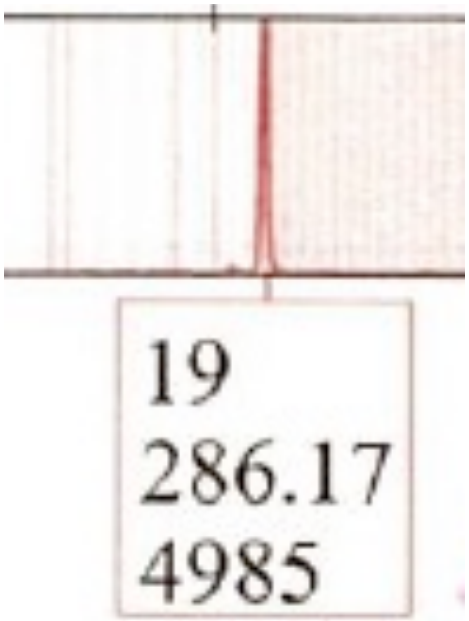


Labeling of a ‘shoulder’ from an early electrophoresis technology, not properly edited

Then…19,20

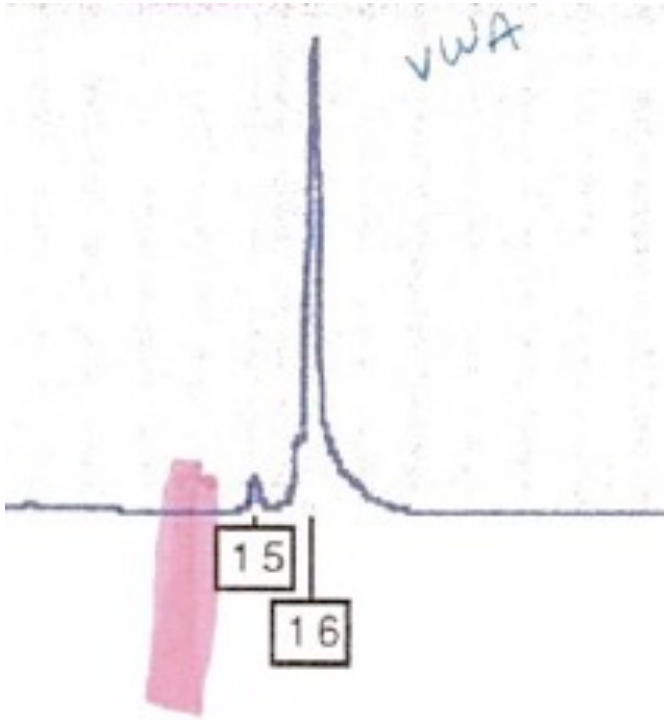


Now…19,19

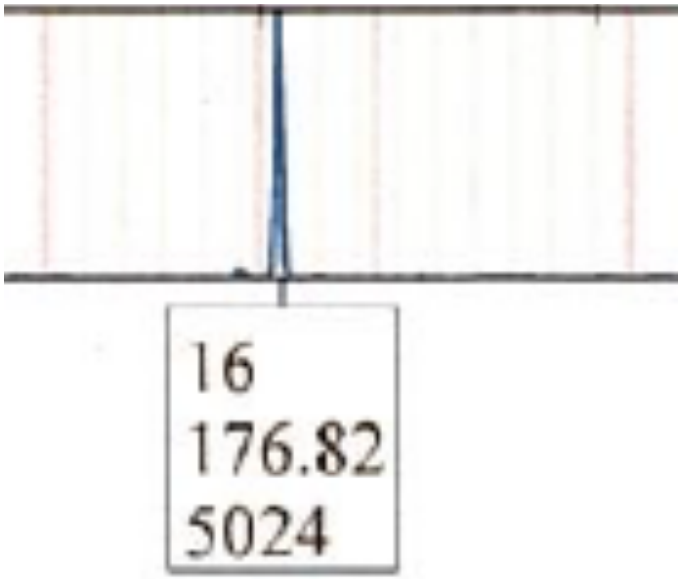


Labeling of a common STR artifact (stutter) in the absence of data filters in an early version of the analysis software, not properly edited

Then. 15,16

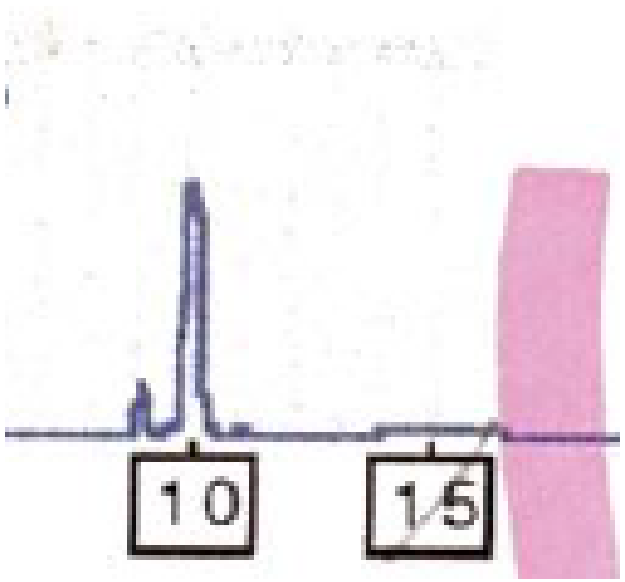


Now. 16,16

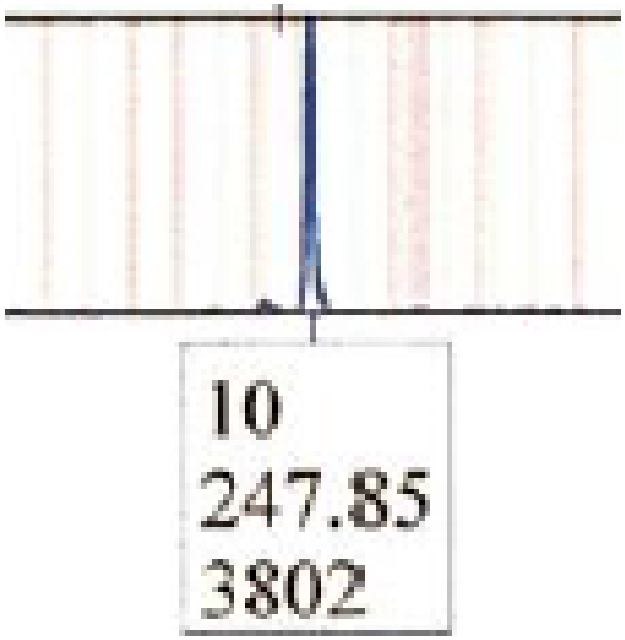


Labeling of elevated baseline using an early detection system, not properly edited

Then…10,15



Now…10,10



Some errors impact *ALLELE* count

Effect on allele frequencies of an error that affects *allele count*

Example: 9,10 called instead of 9,11

Allele	Allele Count	Allele Frequency (Allele Count/2*N*)
6	0	0.000000
7	0	0.000000
8	221	0.547030
9	50	0.123762
10	~~15~~ ⬇ 14	~~0.037129~~ 0.034653
11	~~103~~ ⬆ 104	~~0.254950~~ 0.257426
12	15	0.037129
13	0	0.000000
Total	**404** ✓	**1**

Some errors impact *SAMPLE* count, as well as allele count

Effect on allele frequencies of an error that affects *sample count*

Example: duplicate profile removed (one locus with 8,11 shown, with the number of profiles, *N*, changed from 202 to 201)

Allele	Allele Count	Allele Frequency (Allele Count/2*N*)
6	0	0.000000
7	0	0.000000
8	~~221~~ ⬇ 220	~~0.547030~~ 0.547264
9	50	~~0.123762~~ 0.124378
10	15	~~0.037129~~ 0.037313
11	~~103~~ ⬇ 102	~~0.254950~~ 0.253731
12	15	~~0.037129~~ 0.037313
13	0	0.000000
Total	**~~404~~ ⬇ 402**	**1**

Errors including both clerical and technical:

Over 30,000 alleles in the originally published FBI datasets, the average change in allele frequency due to error is only 0.002 (1999/2001) (range 0.000012 to 0.018181)

Allele Count Errors

Occurred in:

- 28 SAMPLES
  out of 1175
- 47 ALLELE FREQUENCIES
  impacted
  out of ~30,000 typed

Sample Count Errors

Occurred in:

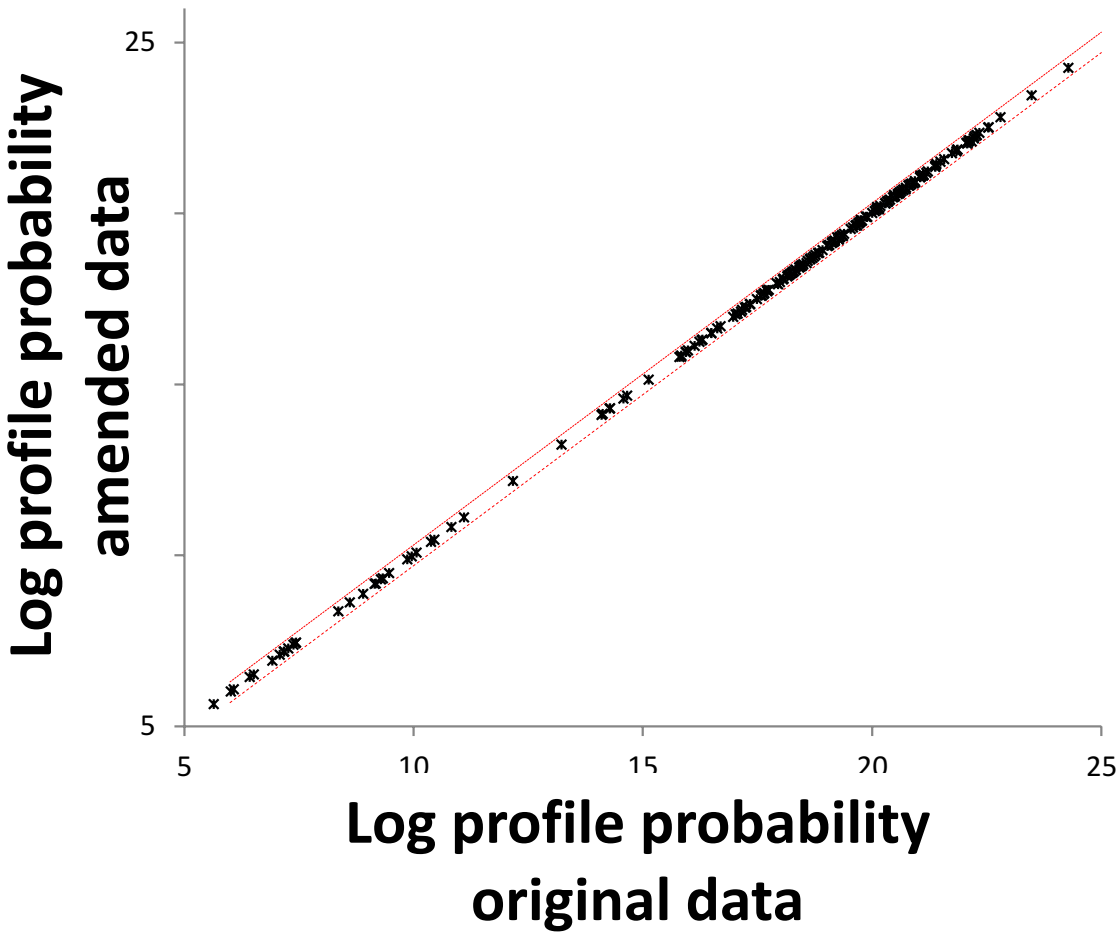
- 6 SAMPLES
  out of 1175
- 208 additional ALLELE FREQUENCIES impacted
  out of ~30,000 typed

Of the 1239 different allele frequencies at 15 loci across 8 populations, 255 frequencies for the alleles noted above required correction.

The FBI Laboratory partnered with Drs. Bruce Budowle (UNTHSC) and John Buckleton (ESR) to perform an assessment of the impact of these errors on profile probability estimates

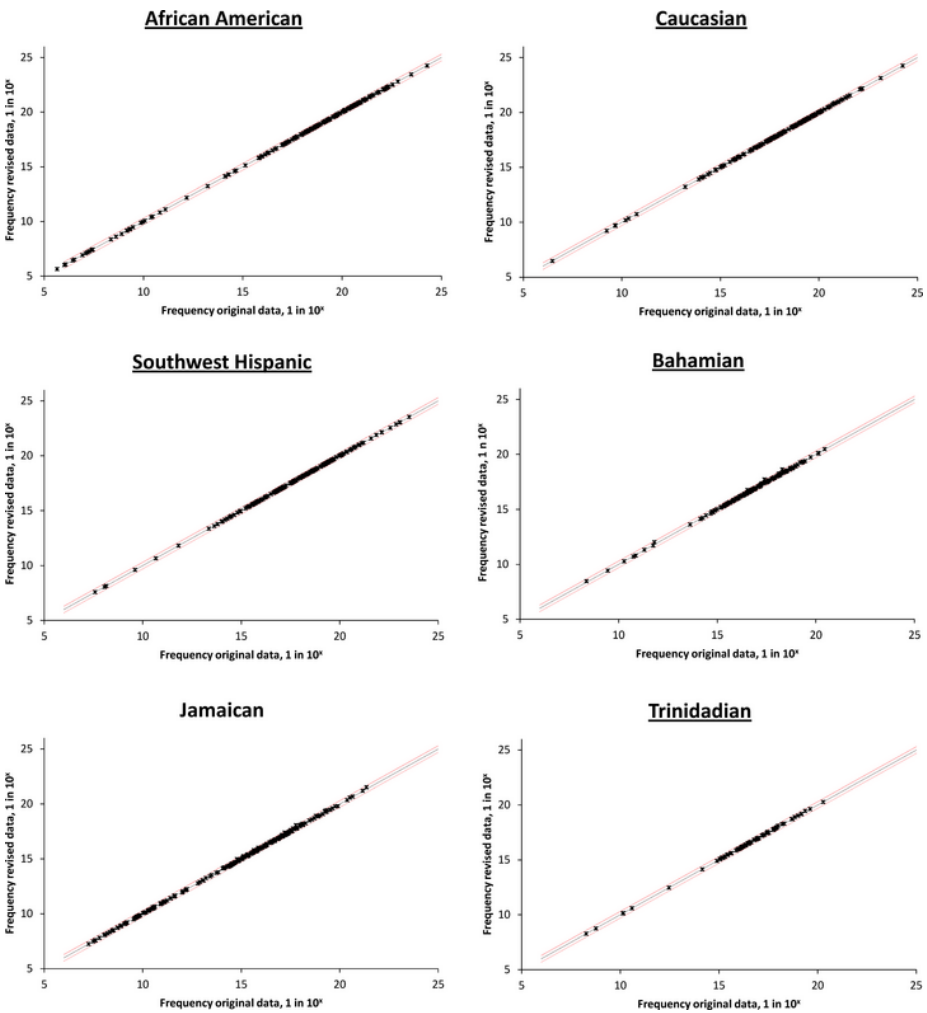
- Only relevant if the evidentiary profile has one or more alleles for which the allele frequency has been corrected.
- If multiple affected alleles occur in a profile, the effect of a correction that makes the allele more-rare could essentially be cancelled out if another allele has a more-common frequency change.

The difference in profile probabilities calculated using the original and updated frequencies is nominal



African American

These *population genetic studies* support the expectation that minor changes in allele frequencies such as these have little effect on statistical calculations performed in forensic or other human identity testing applications



Journal of Forensic Sciences
Volume 60, Issue 4, pages 1114-1116, 3 JUN 2015 DOI: 10.1111/1556-4029.12806
http://onlinelibrary.wiley.com/doi/10.1111/1556-4029.12806/full#jfo12806-fig-0001

Calculated differences in profile probabilities comparing the original and updated frequencies

	Blk	Cau	SW Hisp	Bahamas	Jamaica	Trinidad
15 loci comb.	**1.32**	**1.13**	**1.14**	**1.40**	**1.30**	**1.30**
CSF1PO		1.01		1.03		
D13S317	1.14	1.02		1.03		
D16S539		1.01	1.03	1.03		1.07
D18S51	1.01			1.03	1.18	1.14
D19S433	1.14					
D21S11		1.05		1.03		
D2S1338						
D3S1358		1.01		1.01		
D5S818			1.02	1.04		
D7S820		1.01		1.03		
D8S1179				1.03	1.07	1.07
FGA			1.06	1.02	1.03	
TH01		1.01		1.03		
TPOX		1.01		1.03		
vWA			1.03	1.04		

The magnitude of the impact is <u>no greater</u> for partial profiles

- Of particular interest is the scenario whereby a *more common* estimate is generated using the amended data as compared to the original data

	Full Profiles	Partial Profiles
Worst case scenario: Greatest "more common" difference	1.40-fold	1.18-fold
Dataset	Bahamian	Jamaican (D18S51)

- It is intuitively obvious, as well as demonstrated in this assessment, that such a relatively small number of errors of small magnitude would have little impact on statistical calculations

JOURNAL OF FORENSIC SCIENCES

ERRATUM

J Forensic Sci, July 2015, Vol. 6 0, No. 4
doi: 10.1111/ 1556-4 029.12806
Available online at: o nlinelibrar y.wiley.com

Reference: Budowle B, Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM. Population data on the thirteen CODIS core short tandem repeat loci in African Americans, US Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J Forensic Sci 1999;44(6):1277–86.

Since the development in the late 1990s of the original short tandem repeat (STR) typing systems that included the 13 CODIS

The published allele frequencies (1,2) have been used in the past to generate profile probabilities for autosomal STR typing results using FBI PopStats software. Empirical testing suggests that any discrepancy between profile probabilities calculated using the original and corrected data is expected to be less than a factor of two in a full profile. The actual minimum ratio that we could obtain for a constructed profile with three loci, in the

- Accuracy of the data and rapid dissemination of information was paramount
- To mitigate any potential misunderstanding or exaggeration of the extent, magnitude and impact of the errors:
  - We published an Erratum to the original JFS publication within a month
  - We published an Authors' Response to a Commentary on the Erratum, addressing incorrect assertions
  - We disseminated an information bulletin to NDIS Labs, providing an FBI POC
    - FBI DNA Support Unit responded in real time to nearly a hundred inquires

The Forensic Technology Center of Excellence Presents

ASCLD: FBI Allele Frequency Amendments – Technical Discussion

APPROXIMATLEY 75 MINUTES

Register Now »

June 30, 2015 and August 25, 2015 Webinars

Course Description

Crime Labor[illegible]s have e[illegible]ion of public trust over time through providing reliable quality results. Keeping public trust [illegible]es being self-correcting [illegible] an issue is discovere[illegible] the Federal Bureau [illegible] quality managers, DNA technical leaders and forensic scientists to provide timely assessment, correction, notification, reporting and documentation, enabling Crime Labs t[illegible]ntinue to earn and keep public trust [illegible] have handled the [illegible]

Course [illegible]

- The FBI Laboratory communicated with accrediting bodies and the Consortium of Forensic Science Organizations to discuss supporting them in disseminating information beyond analysts, to lab management and other stakeholders, including attorneys
  - We participated in Webinars focused on technical and non-technical audiences
  - We presented at a meeting of Scientific Working Group on DNA Analysis Methods and the Technical Leaders' Summit at the CODIS Conference
- Several labs reported their own findings, confirming the FBI's impact assessment

Presenters


Tamyra Moretti, Ph.D.


Dr. Bruce Budowle


John Buckleton, Ph.D

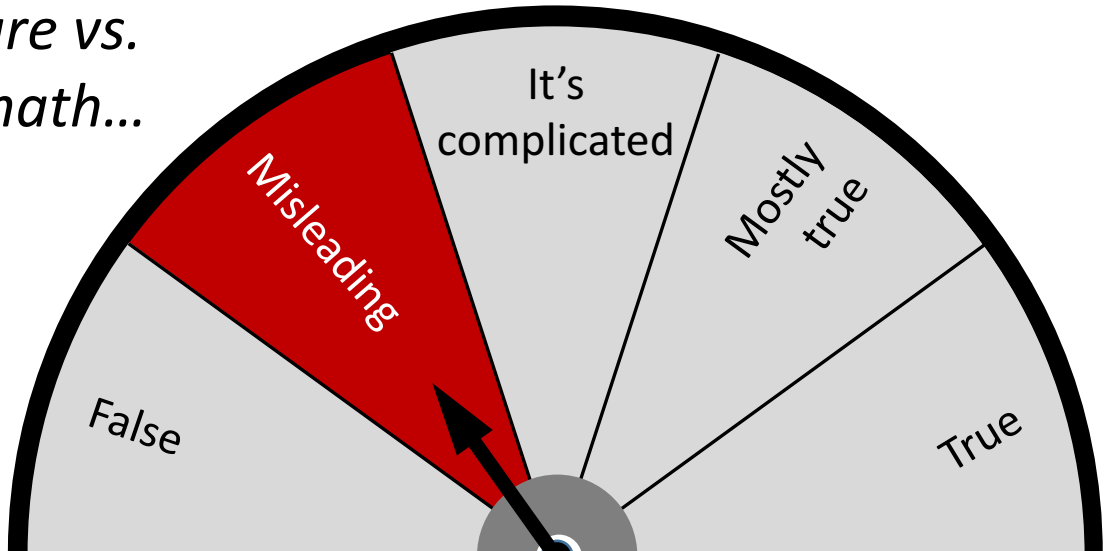

Cyndi Hall


Alyson Saadi, MNS

"While juries might well reach the same decision if errors mean that an individual has a 1 in 1 billion chance of matching a crime-scene sample instead of 1 in 10 billion, for example, that might not be so if errors were to halve, say, assertions that the person had a 1 in 180 chance of matching, as [Daniel] Krane said came up in a case that he testified in this month."

http://www.dispatch.com/content/stories/national_world/2015/05/30/fbi-concedes-errors-in-dna-stats-since-1999.html

Conjecture vs.
Do the math…



The difference in profile probabilities calculated using the original and amended allele frequencies is

less than two-fold

- Sampling variation: We know that individual within-race profile frequencies calculated in different databases can differ by up to10-fold in either direction.
  - Budowle et al. 1993a,1993b
  - FBI Worldwide Compendium
  - NRC II (p. 149-156)

- Greater differences were seen when comparing profile probabilities calculated with the FBI & NIST population databases than with the FBI-original and FBI-amended population databases

Calculations for limited, partial profiles (single source or mixture)

- May be evident in the two truncated digits

  1 in 180 as per original frequencies
  1 in 150 with corrected frequencies (1.18-fold more common)

- According to the NRCII factor of ten expectation, 1 in 180 ≈ 1 in 18 to 1 in 1800

- The corrected frequency, 1 in 150, is well within this range of expectation

For any given DNA typing result, such small differences in probability estimates are simple to independently confirm with the published allele frequencies (FSIG & fbi.gov) & are well within expectations supported by NRC II

# loci/ profile	FBI Original	FBI Amended	NIST	10-fold expectation
15	24 quintillion	25 quintillion	15 quintillion	2.4 quintillion to 240 quintillion
4	59 billion	58 billion	12 billion	5.9 billion to 590 billion
2	7200	7300	9200	720 to 72,000
1	690	690	660	69 to 6900
1	10	11	14	10 to 100

Worldwide survey of STR population data:

250 papers, 446 populations, 24 loci, nearly 500,000 profiles!

Buckleton, Curran, Goudet, Taylor, Thiery, Weir (2016) Population-specific FST values for forensic STR markers: A worldwide survey. Forensic Sci Int Genet 23:91-100

- Errors within the published databases were apparent in "significant number"
- Mostly typographical errors, also miscalls and swapped loci
- Evident in published summary data when:
  - Allele frequencies for a given locus did not add sufficiently close to one
  - Allele frequencies multiplied by 2*N* were not sufficiently close to integers (e.g., back calculating allele counts)

Confirmation of genotypes among multiple typing systems speaks to the quality of the FBI population databases

- Every sample in the dataset now assembled has been typed in three or four multiplexes, often in duplicate.
- Following review and verification, the typing results were authenticated further by concordance among multiplexes.
- This effort of retyping and assessment provided the best assurance of detecting all genotyping errors in the original data sets.
- These data have thus been scrutinized to a level beyond most population studies used for DNA typing statistics.
- The data processing has been undertaken independently by the FBI and the Institute of Environmental Sciences and Research.
- The original, amended and expanded population data are published in peer-reviewed scientific journals.

In Summary…

- Particularly given the methods used more than 10 years ago, it was expected and accepted that some typing errors would occur
- Observations of error in various population databases demonstrate that a small number of genotyping errors is normative and, in fact, an anticipated element of population databasing in which state-of-the-art methodologies are used at any given time
- We support providing the updated frequencies tables and informing the community, and recognize that based on empirical studies, errors of the magnitude found in the 1999/2001 FBI population databases are expected to have at most a nominal impact on match statistics

Acknowledgments

- John Buckleton, ESR
- Bruce Budowle, UNTHSC
- Steven Myers, CA-DOJ
- FBI DNA Support Unit: Anthony Onorato, Jocelyn Carlson, Amber Carr, Jerrilyn Conway, Rosana Hizon, Jodi Irwin, Lilliana Moreno, Michelle Pignone, Jill Smerick & Leah Willis
- FBI DNA Casework Unit: Jade Gray