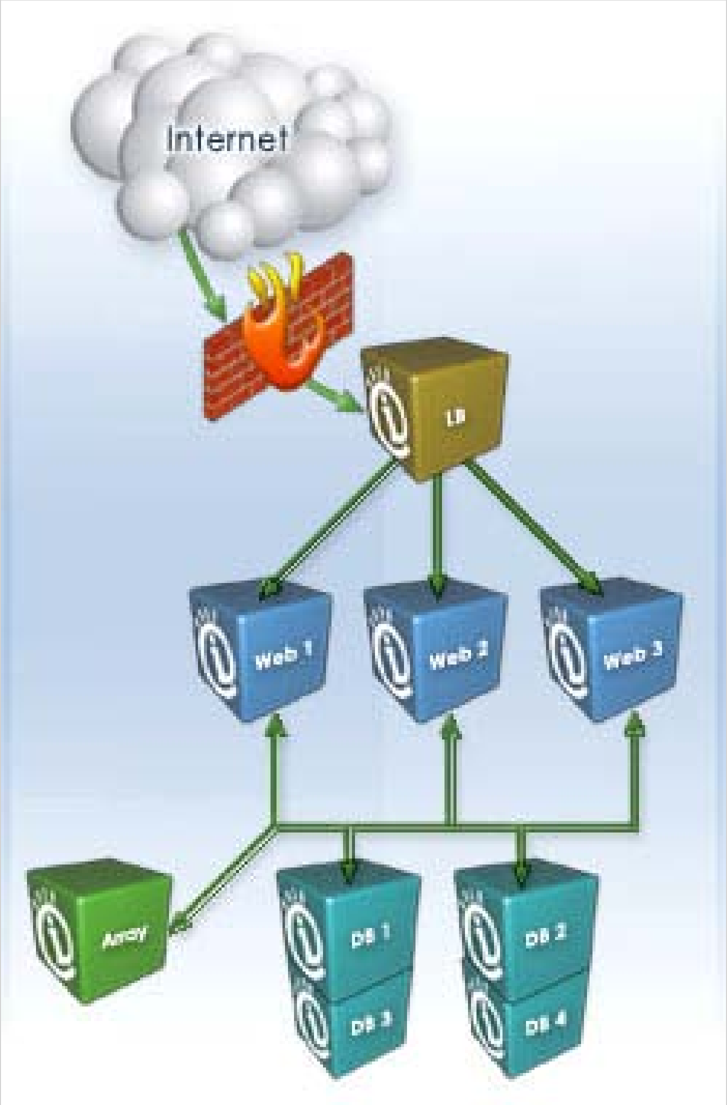
NIJ
National
Institute
of Justice

DNA Mixture Interpretation Workshop | ***Jack Ballantyne***

Interpretation of Y STR Mixtures and Statistical Applications

FREE
DNA MIXTURE INTERPRETATION
WORKSHOP

NFSTC | LARGO, FL | MARCH 15 -17 | 2011



1. ***Y chromosome biology***
2. ***Y-STRs in casework (incl. statistics)***
3. ***Use of the National YSTR Database***
4. ***Y-STR mixture analysis***

NIJ
National Institute of Justice

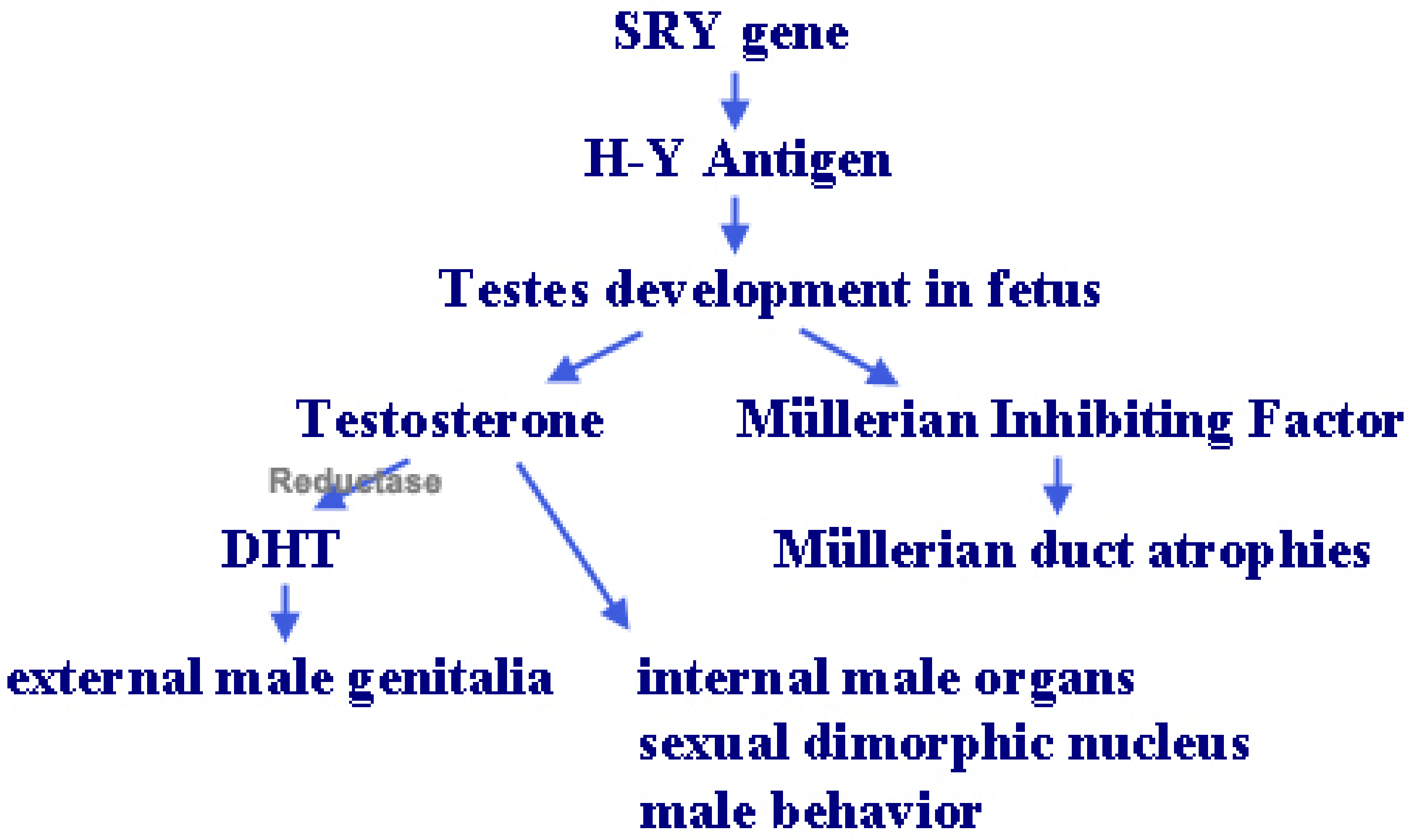
Y-Chromosome

Biology

Function of Y Chromosome?

- Aneuploidies for the X and Y
  - 47, XXY (Klinefelter synd.) → males
    - 48, XXXY; 49, XXXXY; 50, XXXXXY → males
  - 45, XO (Turner synd.) → females
  - 47, XXX (triple-X karyotype)→ ‘normal’ female
  - 47, XYY karyotype →’ normal’ male
- Sex Reversed Humans
  - XY → female (Y minus TDF)
  - XX → male (X plus TDF)

Sex Determination in Mammals



SRY gene

H-Y Antigen

Testes development in fetus

Testosterone

Müllerian Inhibiting Factor

Reductase

DHT

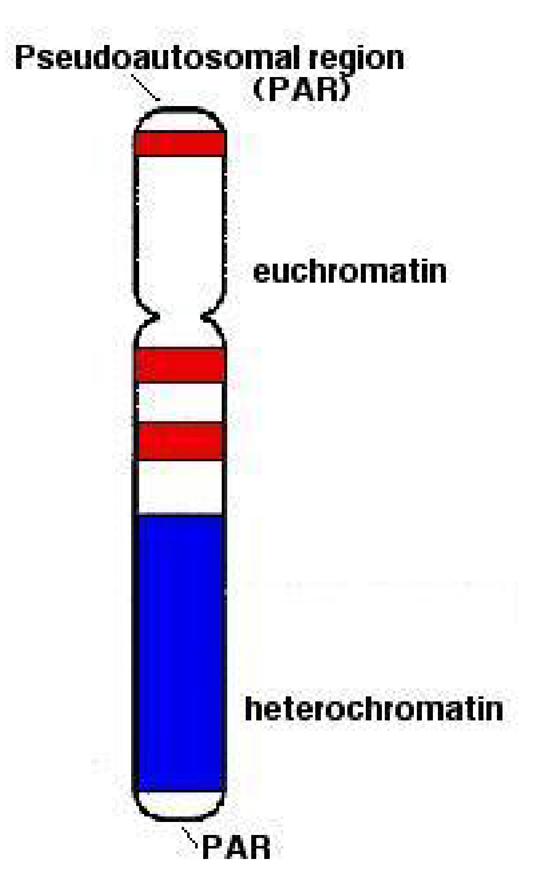
Müllerian duct atrophies

external male genitalia

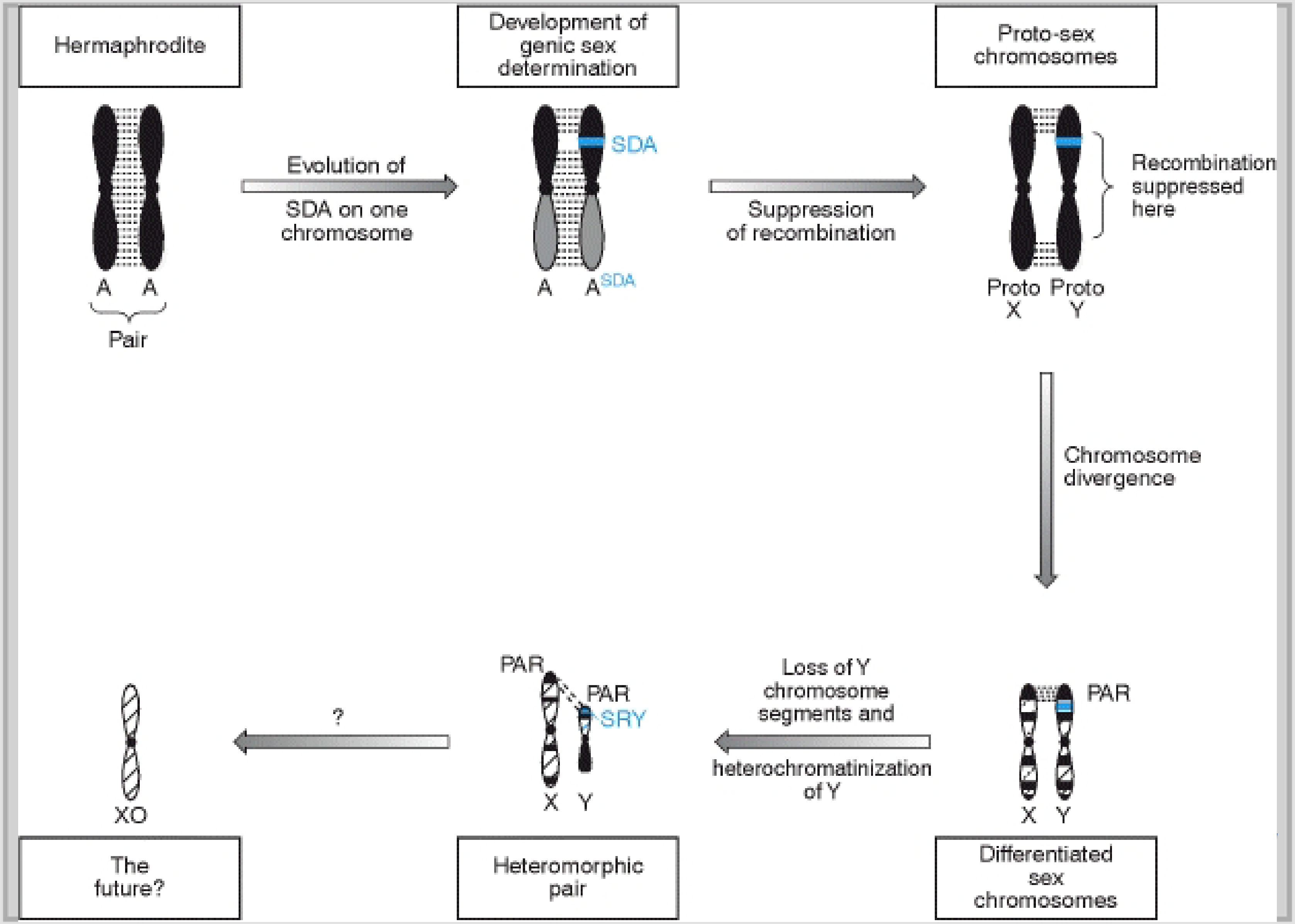
internal male organs
sexual dimorphic nucleus
male behavior

Female system is
always the default

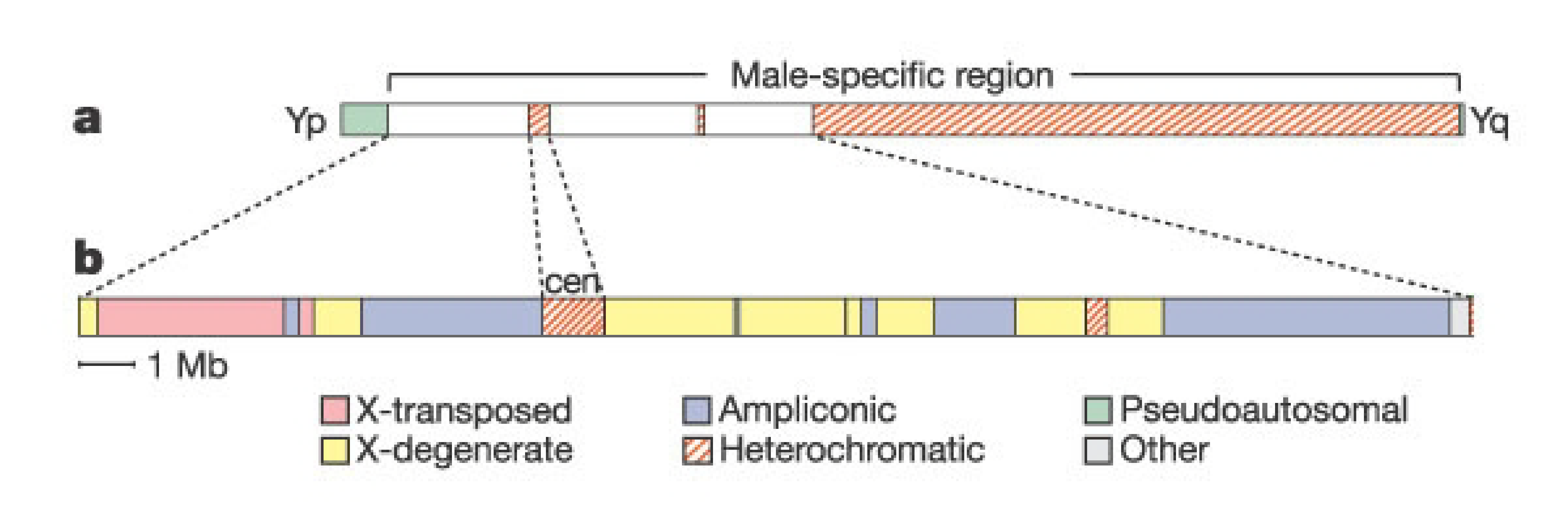
Classic View of Y-Chromosome

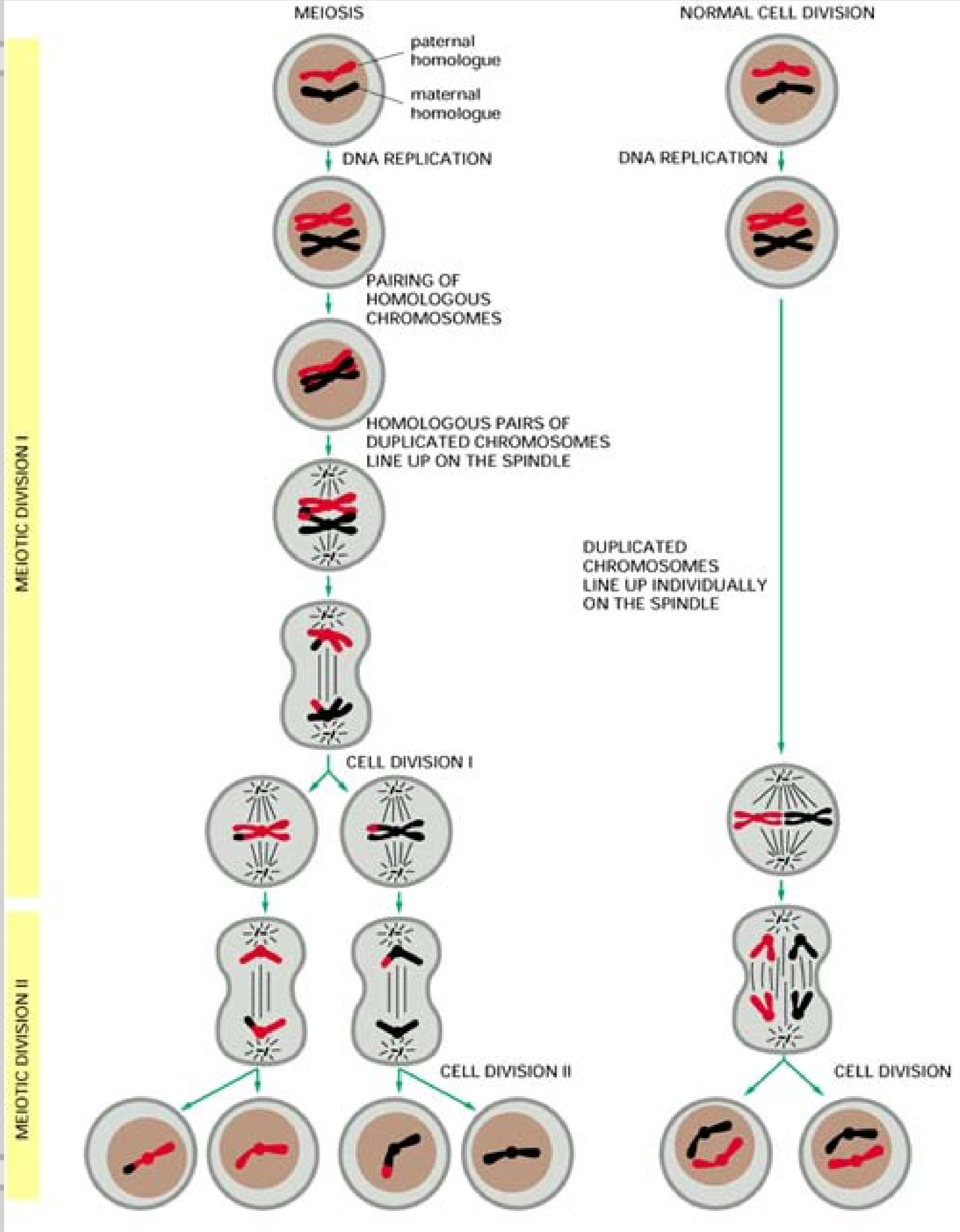


- TDF master gene
- patrilineal inheritance
- no recombination in NRY
- recombination in PAR
- junk-rich, gene poor


Hermaphrodite
Development of genic sex determination
Proto-sex chromosomes
Evolution of
SDA on one chromosome
SDA
A A
Pair
A A^SDA
Suppression of recombination
Recombination suppressed here
Proto X
Proto Y
Chromosome divergence
PAR
X Y
Differentiated sex chromosomes
Loss of Y chromosome segments and
heterochromatinization of Y
PAR
PAR
SRY
X Y
Heteromorphic pair
?
XO
The future?

Y Chromosome NRY is a Mosaic of Discrete Sequence Classes





MEIOSIS

NORMAL CELL DIVISION

paternal homologue

maternal homologue

DNA REPLICATION

DNA REPLICATION

PAIRING OF HOMOLOGOUS CHROMOSOMES

HOMOLOGOUS PAIRS OF DUPLICATED CHROMOSOMES LINE UP ON THE SPINDLE

DUPLICATED CHROMOSOMES LINE UP INDIVIDUALLY ON THE SPINDLE

CELL DIVISION I

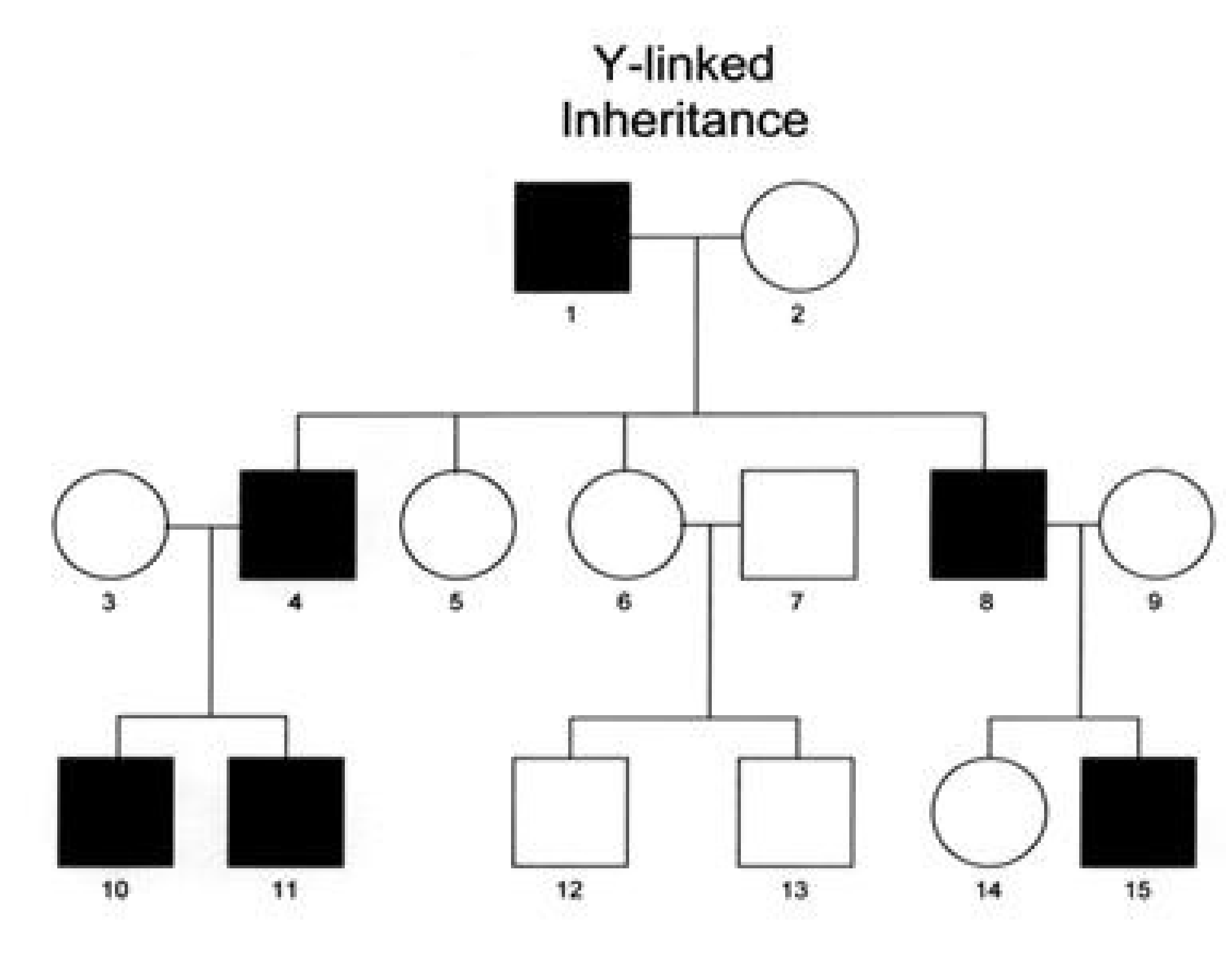
CELL DIVISION II

CELL DIVISION

MEIOTIC DIVISION I

MEIOTIC DIVISION II

NIJ National Institute of Justice

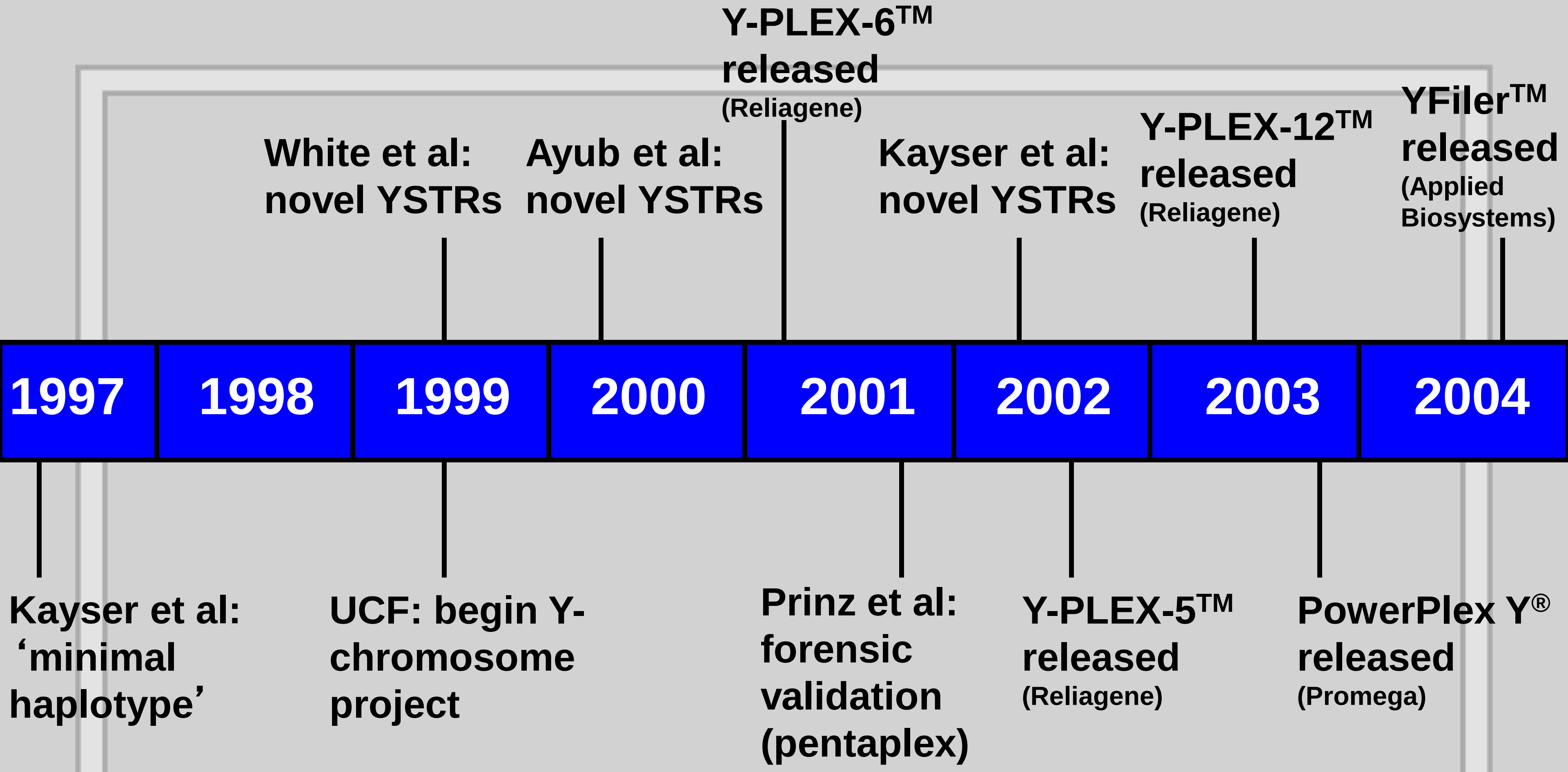

Y-linked
Inheritance
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Reasons Y?

- **Males (Criminal Victimization in United States, BJS 2001)**
  - **80% of all violent crime**
  - **95% of all sex offenses**
- **When trying to determine the genetic profile of the male donor in a male/female DNA admixture (when F/M > 20, often >1000) and autosomal STR analysis fails (is not informative) or not possible**
  - **sexual assault cases (saliva/saliva; saliva/vaginal secretions; extended interval post coital samples)**
    - **pre-mature lysis into non-sperm fraction**
  - **aspermia/oligospermia**
  - **normal degradation/loss over time**

Reasons Y? (cont’d)

- No need for differential extraction
- Determination of number of semen donors
- Missing persons (MP)
  - criminal paternity/disaster victim ID
  - haplotype of MP determined by typing male relative
    - son, brother, father, uncle, nephew
- Additional statistical discrimination
  - mixture/relative cases
- Y-SNPs: useful for ethnogeographic ancestry prediction
- Familial Searching
  - Reduce number of potential male relatives obtained by low stringency match of sample profile to offendor database



1997

Kayser et al: 'minimal haplotype'

1998

1999

White et al: novel YSTRs

UCF: begin Y-chromosome project

2000

Ayub et al: novel YSTRs

2001

Y-PLEX-6™ released (Reliagene)

Prinz et al: forensic validation (pentaplex)

2002

Kayser et al: novel YSTRs

Y-PLEX-5™ released (Reliagene)

2003

Y-PLEX-12™ released (Reliagene)

PowerPlex Y® released (Promega)

2004

YFiler™ released (Applied Biosystems)

Y-STR Development Timeline

"Minimal Haplotype" Loci

(Kayser et. al. 1997):

- DYS 19 DYS 389I DYS 389II DYS 390,

  DYS 391 DYS 392 DYS 393 DYS 385 I/II

  - First used in Europe
  - Formed the basis for multiplex development in the U.S.

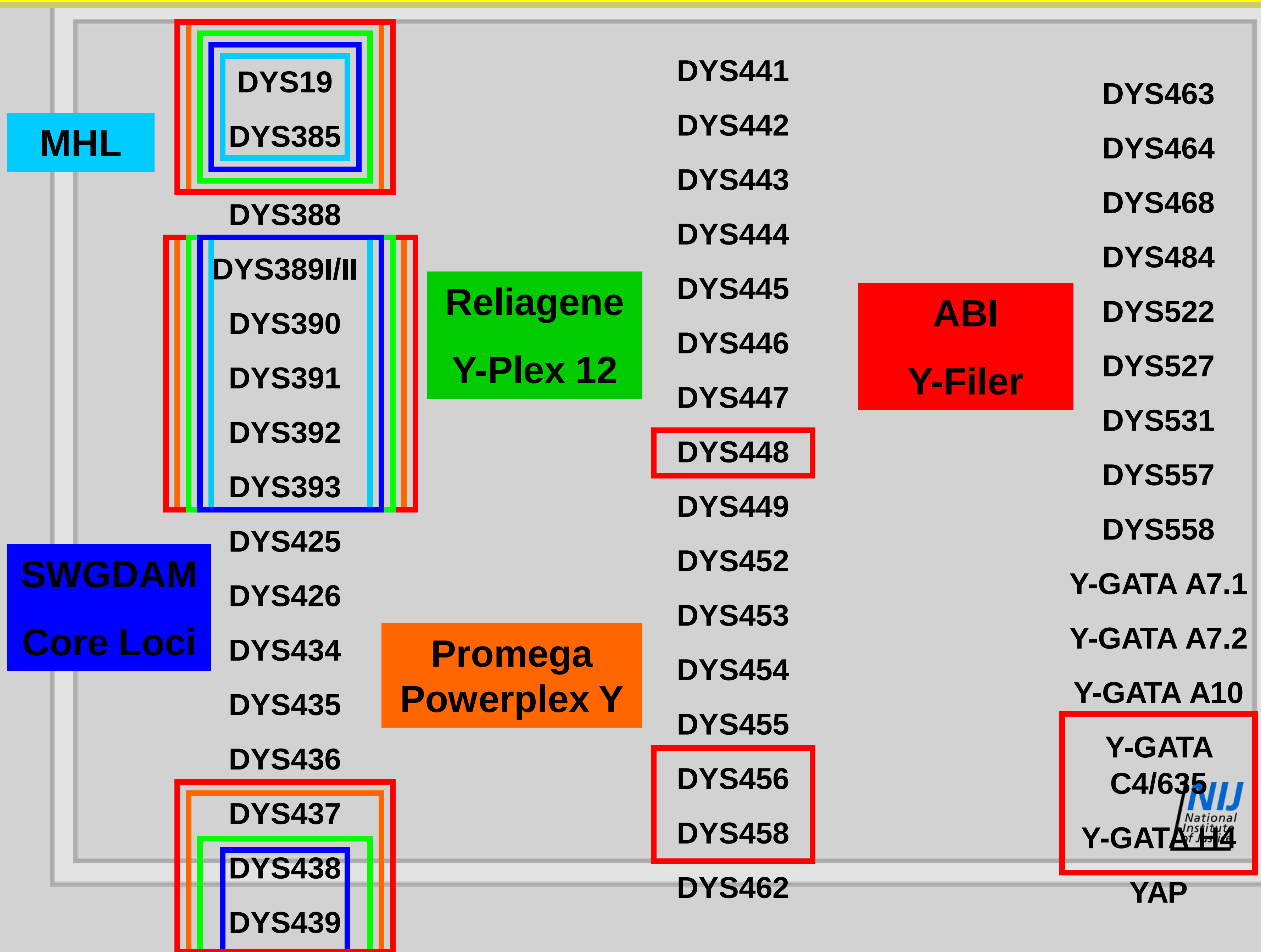
SWGDAM Core Loci

- **Recommended in 2003**
- **DYS19 DYS389I DYS389II DYS390, DYS391 DYS392 DYS393 DYS385 I/II**
- **Also: DYS438, DYS439**

Commercially Available Y-STR Multiplex Kits

- Reliagene: Y-PLEX-5, Y-PLEX-6, Y-PLEX-12
  - No longer available
- ABI: Yfiler^{TM}
  - 17 loci in a single reaction
  - Includes SWGDAM core loci
- Promega: PowerPlex Y^{TM}
  - 12 loci in a single reaction

Y-STR Markers in Commercial kits



MHL

Reliagene Y-Plex 12

SWGDAM Core Loci

Promega Powerplex Y

ABI Y-Filer

DYS19
DYS385
DYS388
DYS389I/II
DYS390
DYS391
DYS392
DYS393
DYS425
DYS426
DYS434
DYS435
DYS436
DYS437
DYS438
DYS439

DYS441
DYS442
DYS443
DYS444
DYS445
DYS446
DYS447
DYS448
DYS449
DYS452
DYS453
DYS454
DYS455
DYS456
DYS458
DYS462

DYS463
DYS464
DYS468
DYS484
DYS522
DYS527
DYS531
DYS557
DYS558
Y-GATA A7.1
Y-GATA A7.2
Y-GATA A10
Y-GATA C4/635
Y-GATA H4
YAP

Statistics

“A detailed understanding of the influence of all factors on the evolution of profile proportions requires a lifetime of study, and more”

Balding 2005

'Some people believe football is a matter of life and death. I'm very disappointed with that attitude. I can assure you it is much, much more important than that. ‘

Bill Shankly, 1960s

Basic Y-STR Interpretation Guidelines

- Similar general issues to autosomal STRs
  - Thresholds for detection and interpretation
  - Probability of a match (STATISTICS!)
  - Mixtures – what constitutes a mixture
  - Stutter
  - Validation studies in concert with guidelines

Probability of a Match

Issues

- estimating the rarity of a Y DNA profile is performed differently than for autosomal DNA markers
- because of linkage, each haplotype is treated as an allele and the total number of possible haplotypes comprise the alleles of a single locus
  - Composite multi-locus profile is treated as a single locus or haplotype
- no evidence for recombination across the majority of the Y-chromosome
- cannot employ the product rule to estimate the rarity of the Y types in a profile

Counting Method

- The counting method is very simple
- A Y-haplotype (evidence sample) is compared to a reference database(s) of unrelated individuals
- The number of times the Y-haplotype is observed in a database
  - The size of a database can be and is often limited
  - With databases (e.g., *n* = 100 to 1000), many possible haplotypes will not be observed and there will be sampling error
- A confidence interval can be placed on the observation
  - Can convey with a high degree of confidence that the rarity of the evidence Y-haplotype among unrelated individuals in a given population(s) is less than the upper bound of the estimate

For Y haplotype observed, count the number of times the profile is observed (x) in a database of N individuals

$$p = x/N$$

$$CI = p \pm 1.96\sqrt{p(1-p)/N}$$

$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} = 0.05$$

If the haplotype has not been observed in the database, then:

The upper (95%) bound of the CI is

$$1-(\mathbf{0.05})^{1/N}$$

Or the ‘rule of 3’ = 3/N

Population Substructure Issues

- Correction for population structure may be necessary although depends upon no. of loci typed

- Effective population size ¼ of autosomal loci

- Substructure effects less in US than ancestral populations

- Use when reference database considered not representative

Substructuring Formulae?

$$f\ (\text{haplotype}) = p_i + \theta(1-p_i)$$

or equivalent

$$f\ (\text{haplotype}) = \theta + p_i(1-\theta)$$

In either case θ *plays a significant role in determining f(haplotype)*

Y-STR Database Goals

- **To compile and consolidate Y-STR data from all available ‘legitimate’ sources**
- **To create a Y-STR Consortium comprised of stake holders and data contributors from the forensic community**
- **Expand data**
  - **Type additional samples using core loci**
- **Provide custodial and managerial responsibility**
- **Develop quality indicators for data inclusion and submission**
  - **‘Proficiency testing’ for labs who wish to contribute data**
  - **Screen data and remove duplicate & related samples**
- **Ensure allele-call consistency among different primer sets**
- **Provide accessibility and statistical data to the forensic community via the Internet**

Y-STR Consortium Members

- NCFS
  - Jack Ballantyne
  - Lyn Fatolitis
- Applied Biosystems
  - Lisa Calandro
- FBI
  - Bruce Budowle
- University of Arizona
  - Mike Hammer
- NIST
  - John Butler
- MN Dept of Public Health
  - Ann Marie Gross
- NYC OCME
  - Mecki Prinz
- University of North Texas
  - Arthur Eisenberg
- Promega
  - Curtis Knox
- ReliaGene
  - Sudhir Sinha
- Orchid Cellmark
  - Cassie Johnson
- NIJ
  - John Paul Jones

The Y-STR Consortium was formed at the 2006 AAFS Meeting in Seattle, WA to assist in sample consolidation and the design and development of the database.

Data Consolidation

- “Fast-track” consolidation of data from only AB, NCFS, Promega, ReliaGene, and U of AZ
- Removed obvious identical samples
- Turned data over to Bruce Budowle of the FBI for calculation of F_{st} (theta)
- Further investigated unresolved matches to remove duplicate and related samples
- Consolidated data, validated the database, and made Release 1.0 available to the forensic community via the Internet on January 3, 2008
- Continue to expand the database by additional sampling and typing
  - State and local crime labs
  - In-house

Consolidated Database: Release 1.0 (N = 13,906)

Agency	Beginning Total	Samples Removed	Ending Total
NCFS	1401	42	1359
ReliaGene	3406	369	3037
Promega	4004	204	3800
AB	3502	254	3248
AZ	2486	24	2462
Total	14,799	893	13,906

Y-STR Interpretation Guidelines

Scientific Working Group on DNA Analysis Methods (SWGDAM)
FSC January 2009

5. Statistical Interpretation

5.1. Y-STR loci are located on the nonrecombining part of the Y-chromosome and, therefore, should be considered linked as a single locus. A Y-STR database must consist of haplotype frequencies rather than allele frequencies. The source of the population database(s) used should be documented. Relevant population(s) for which the frequency will be estimated should be identified. A consolidated U.S. Y-STR database (http://usystrdatabase.org) has been established and should be used for population frequency estimation. A number of other Y-STR haplotype frequency databases exist online. (See available listing on the NIST [National Institute of Standards and Technology] STRBase Web site at http://www.cstl.nist.gov/biotech/strbase/y_strs.htm.)

Release 2.0

- Was made available on March 1, 2009
- Comprised of 17,216 haplotypes
  - An additional 3,310 haplotypes were uploaded
    - 1062 African American
    - 115 Asian (Southern Indian)
    - 1062 Caucasian
    - 1071Hispanic
  - Applied Biosystems donated 2,912 17-locus haplotypes
    - 950 African American, 957 Caucasian, and 1,005 Hispanic
  - Illinois State Police donated 283 11-locus haplotypes
    - 112 African American, 105 Caucasian, and 66 Hispanic
    - 115 12-locus haplotypes (Asian/Southern Indian)

Release 2.1

- Was made available on July 1, 2009
- Comprised of 17,864 haplotypes
  - An additional 649 haplotypes were uploaded
    - 442 Caucasian
    - 200 African American
    - 7 Asian
  - NCFS typed 619 of these samples using Yfiler (17 loci)
  - The Orange County California Coroner’s Office donated 30 12-locus haplotypes

Release 2.2

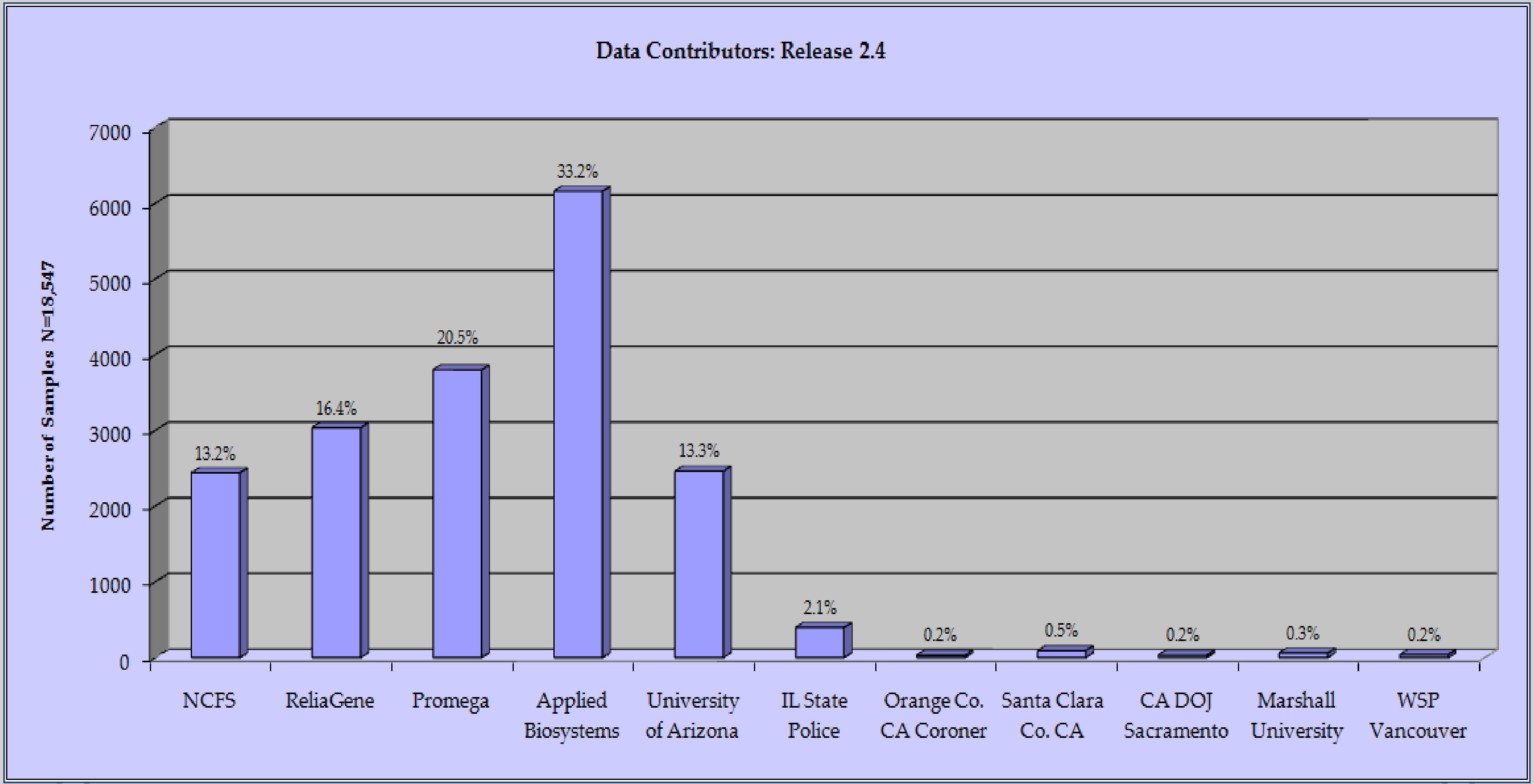
- Was made available on January 24, 2010
- Comprised of 18,199 haplotypes
  - An additional 335 Yfiler haplotypes were uploaded
    - 212 Caucasian
    - 102 African American
    - 13 Hispanic
    - 8 Asian
  - NCFS typed 277of these samples using Yfiler
  - The Santa Clara County Crime Laboratory donated 58 Yfiler samples

Release 2.3

- Was made available on July 31, 2010
- Comprised of 18,448 haplotypes
    - An additional 249 Yfiler haplotypes were uploaded
        - 70 Caucasian
        - 119 African American
        - 54 Hispanic
        - 6 Asian
    - NCFS typed 185 of these samples
    - The Santa Clara County Crime Laboratory donated 32 samples
    - The California Department of Justice Sacramento Crime Laboratory donated 32 samples

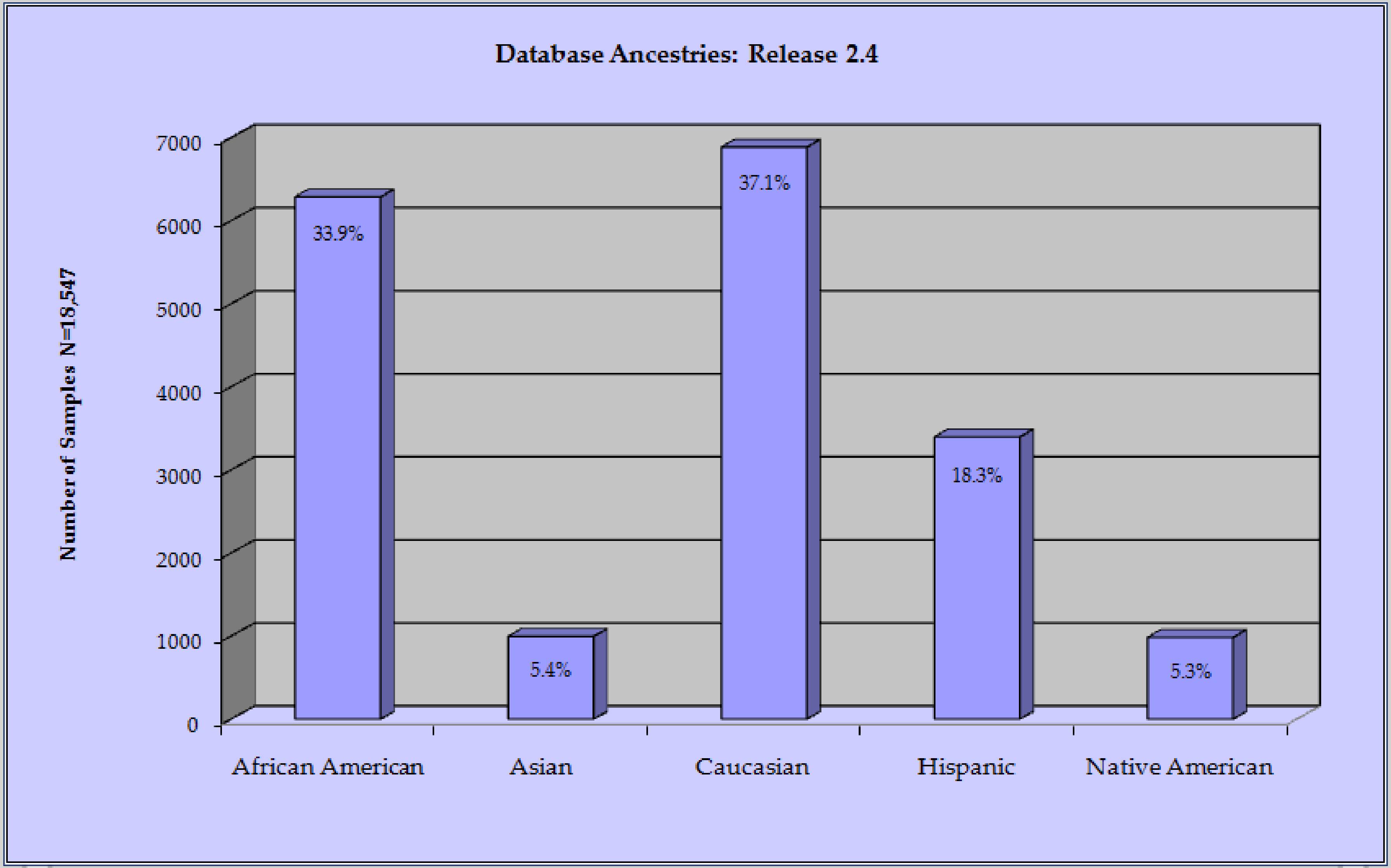
Release 2.4 – Current Version

- Was made available on January 2, 2011
- Comprised of 18,547 haplotypes
  - An additional 99 Yfiler haplotypes were uploaded
    - 52 Caucasian
    - 7 African American
    - 40 Asian
  - The Marshall University Forensic Science Center donated 59 samples
  - The Washington State Police Crime Laboratory in Vancouver donated 40 samples



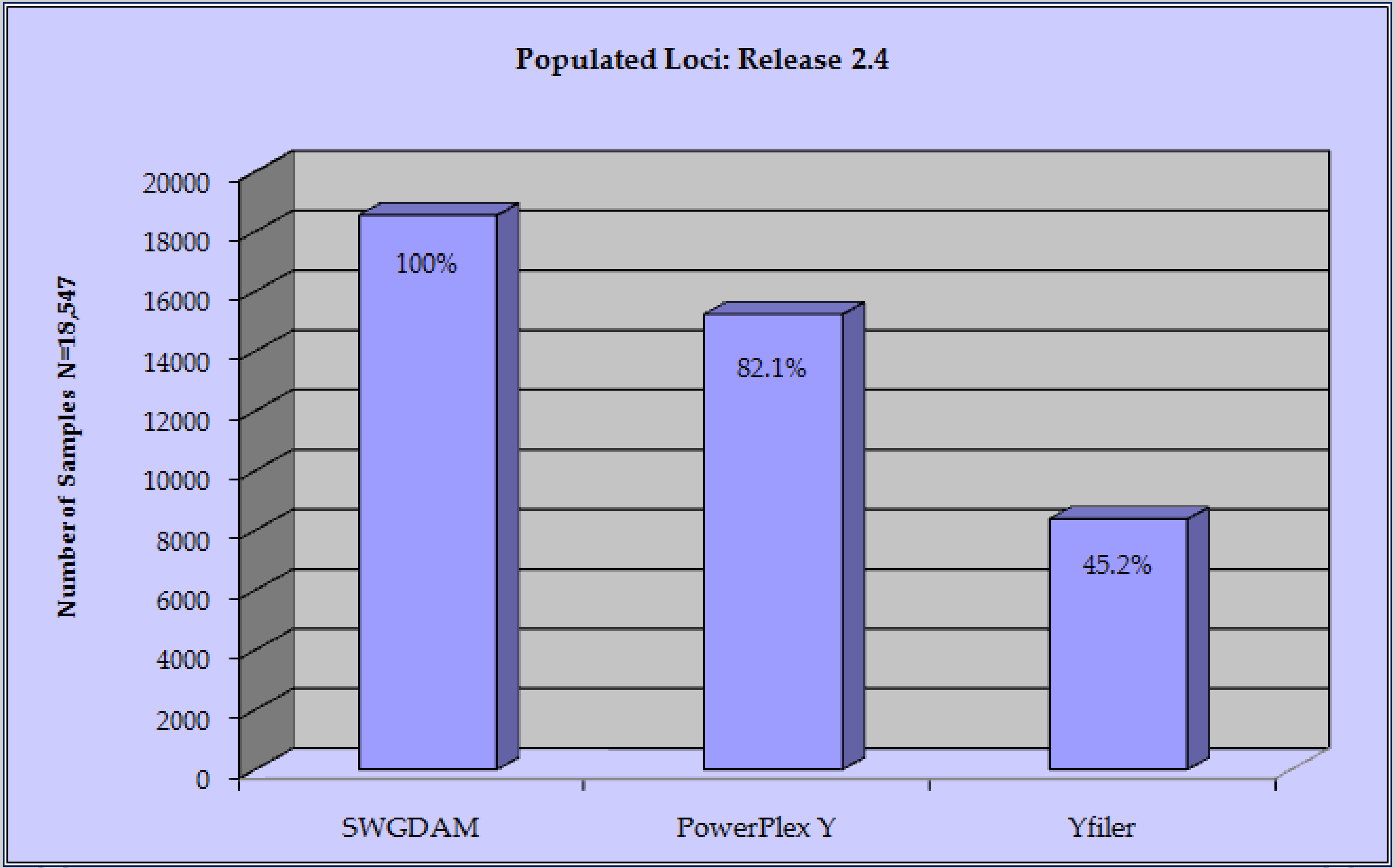
- NCFS contributed 2,440 samples
- ReliaGene contributed 3,037 samples
- Promega contributed 3,800 samples
- Applied Biosystems contributed 6,159 samples
- University of AZ contributed 2,462 samples
- Illinois State Police contributed 398 samples

- Orange County CA Coroner contributed 30 samples
- Santa Clara CA Crime Lab contributed 90 samples
- The CA DOJ in Sacramento contributed 32 samples
- Marshall University contributed 59 samples
- Washington State Police in Vancouver contributed 40 samples

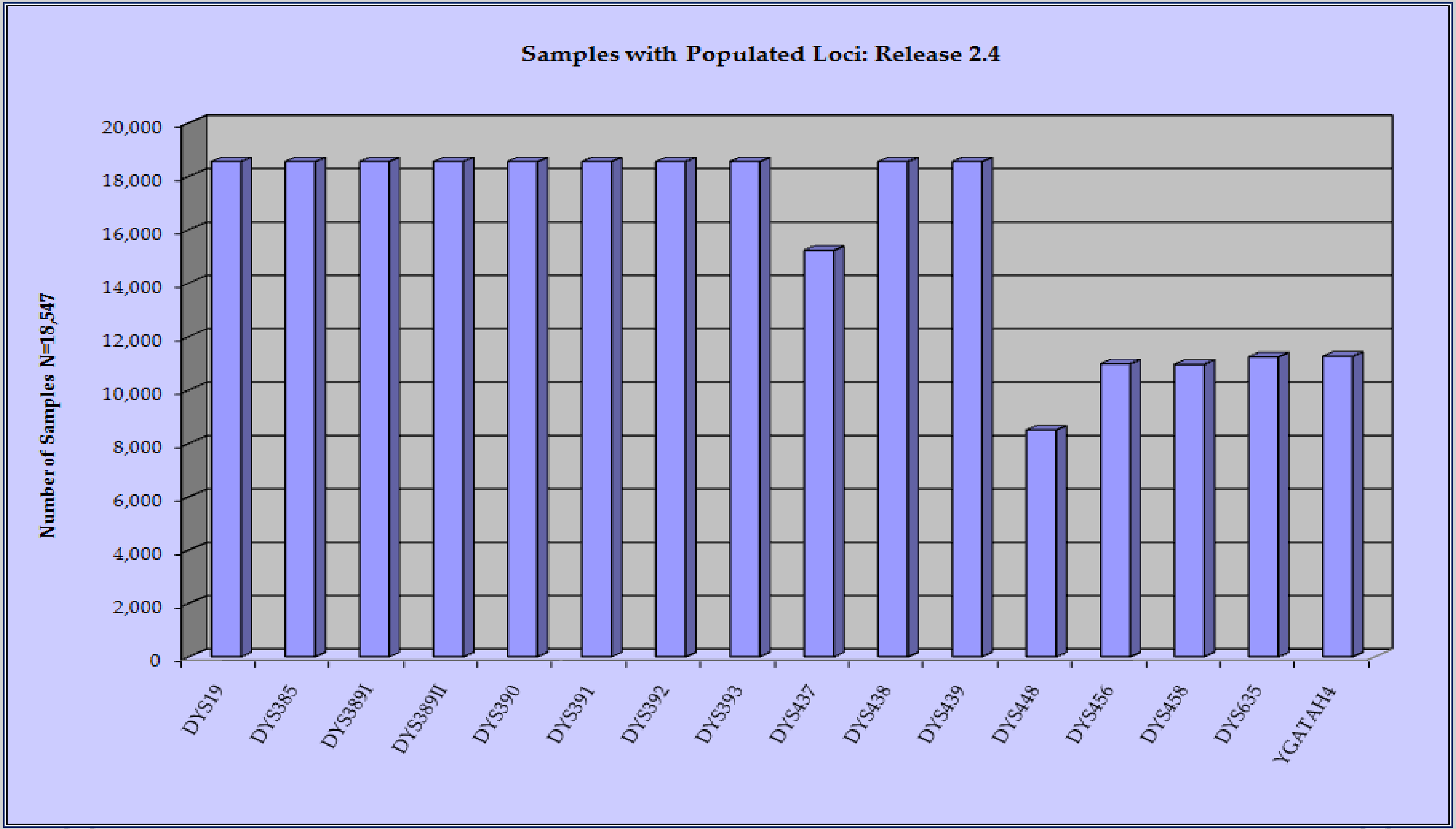


- 6286 African American (All, Undefined, or Select by State)
- 996 Asian (All, Asian, Chinese, Filipino, Oriental, Southern Indian, Vietnamese)
- 6885 Caucasian (All, US, Canada, Europe, Undefined)
- 3397 Hispanic (All, Undefined, or Select by State)
- 983 Native American (All, Apache, Navajo, Shoshone, Sioux)

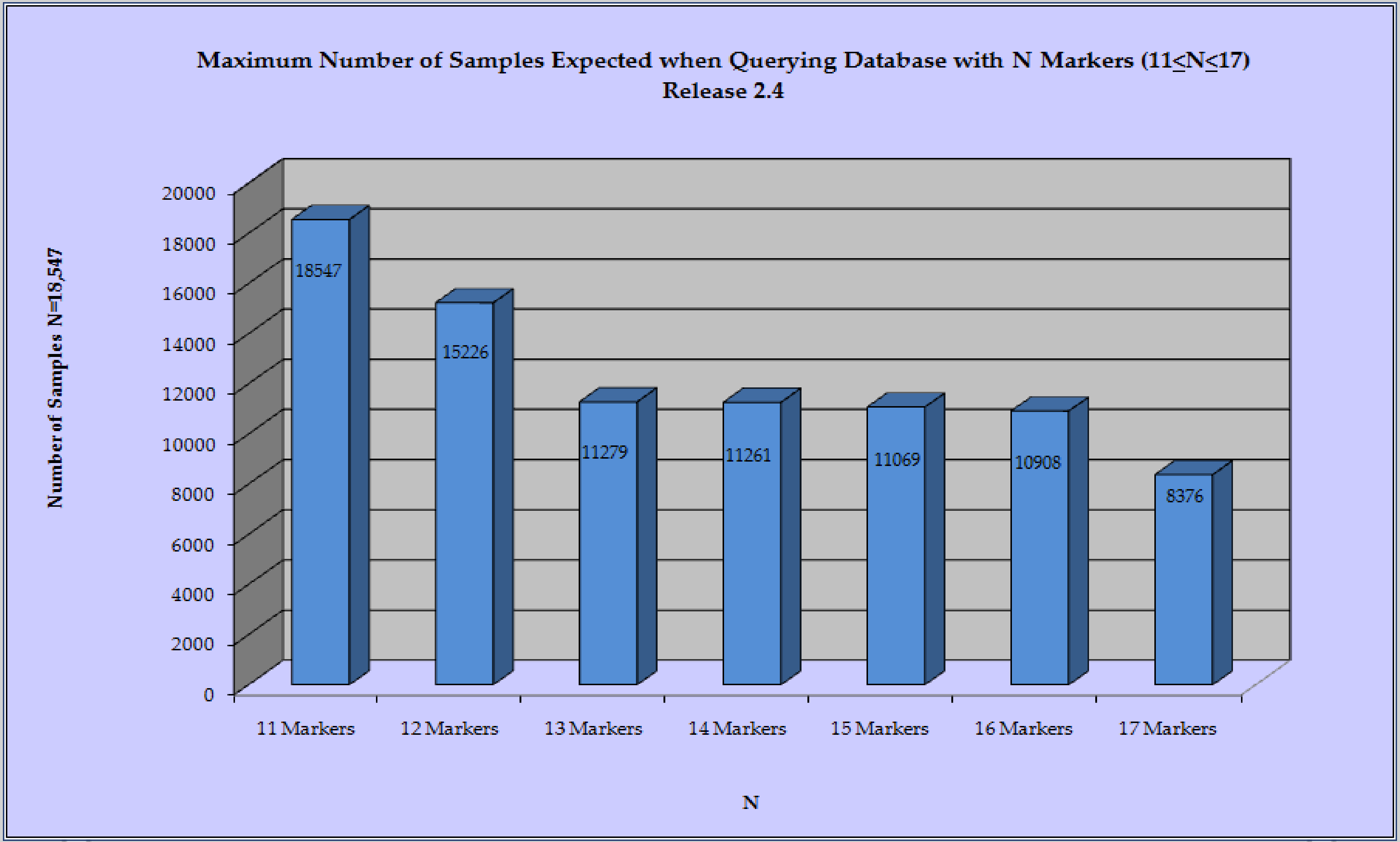
NIJ National Institute of Justice



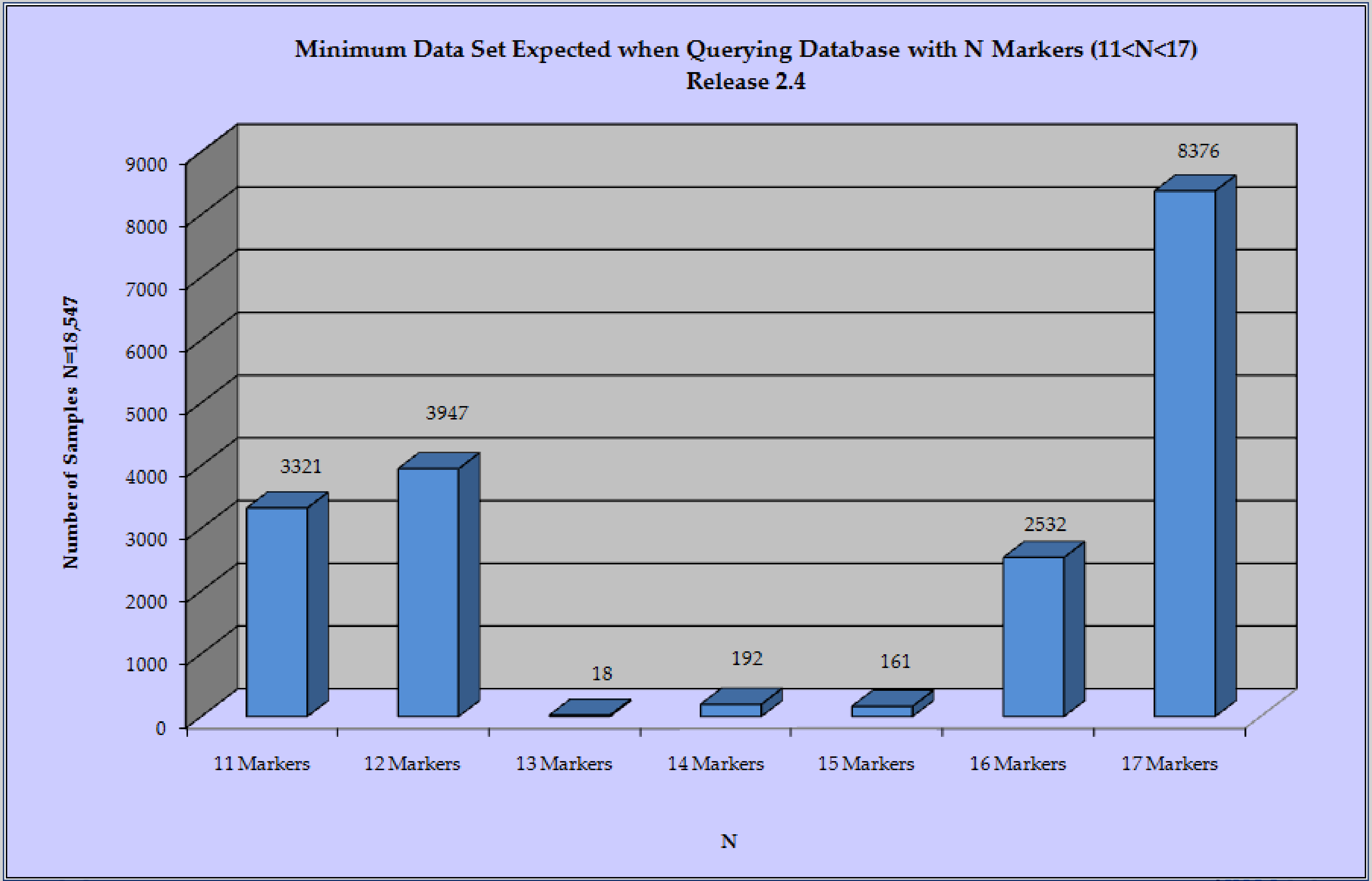
- Release 2.4 contains 18,547 samples with a complete 11-marker SWGDAM core haplotype
- 15,223 samples have a complete 12-marker PowerPlex Y haplotype
- 8,376 samples have a complete 17-marker Yfiler haplotype



- Sample data supplied by NCFS and the University of Arizona were generated using Y-STR typing systems developed at these institutions rather than the more-commonly used commercial kits. This has resulted in a variation in the number of populated markers within the database.

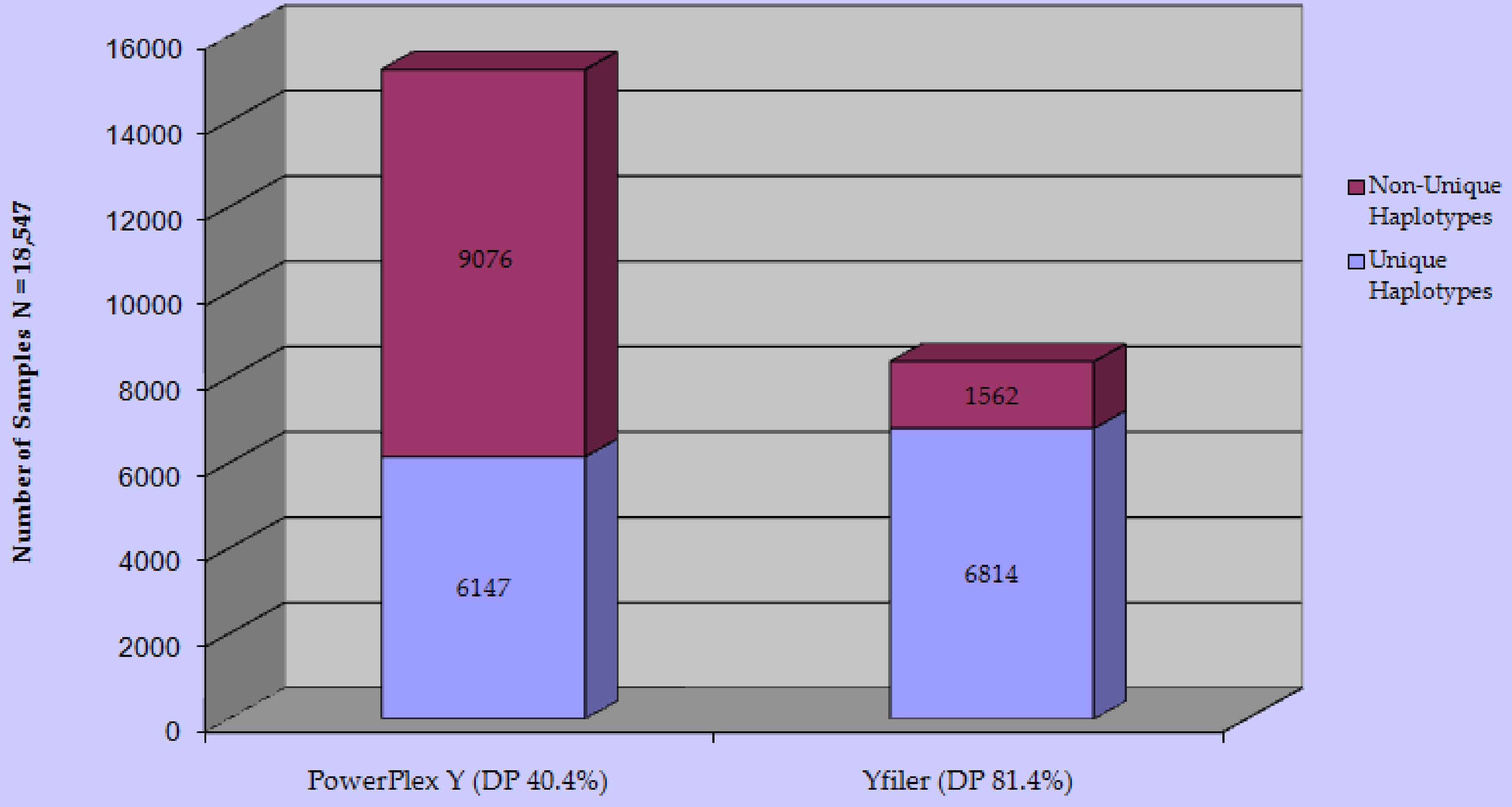


- The database is designed to query only those samples that possess data at the particular markers chosen by the user, resulting in a data set that varies depending on which markers are selected. This graph illustrates the maximum data set expected when querying the database.



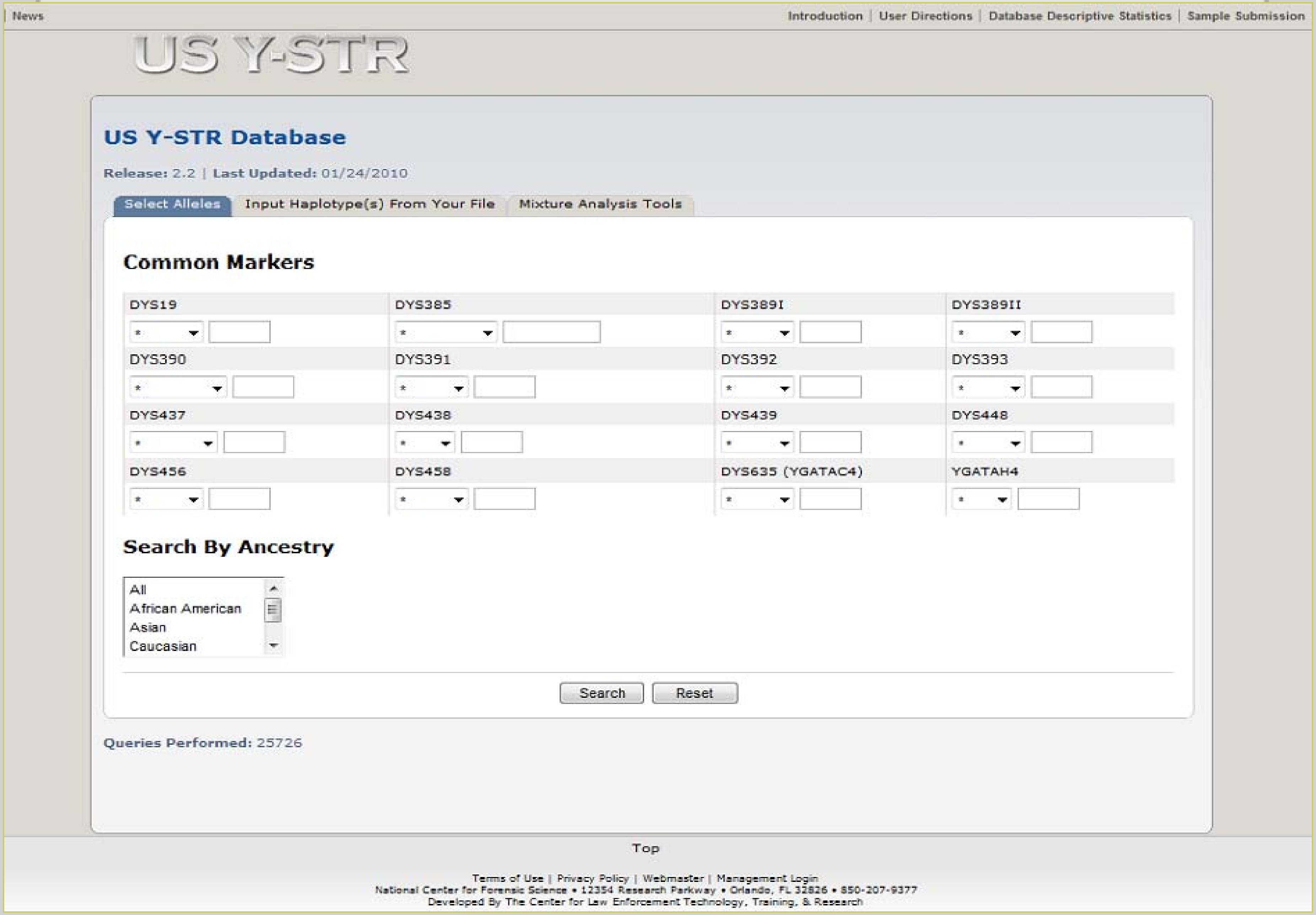
- The data set for 12-16 markers will vary depending on which particular markers are chosen. This graph illustrates the minimum data set expected when querying the database.

Discrimination Potential (DP): Release 2.4



- Release 2.4 contains 15,223 complete PowerPlex Y (12-locus) haplotypes. Of these, 6147 haplotypes are unique (i.e., seen only once in the database) while 9076 haplotypes are seen more than once, giving a DP of 40.4%
- The Database contains 8376 complete Y-filer (17-locus) haplotypes. Of these, 6814 haplotypes are unique while 1562 haplotypes are seen more than once, giving a DP of 81.6%.

Database Home Page: www.usystrdatabase.org



News | Introduction | User Directions | Database Descriptive Statistics | Sample Submission

US Y-STR

US Y-STR Database

Release: 2.2 | **Last Updated:** 01/24/2010

Common Markers

DYS19	DYS385	DYS389I	DYS389II
*	*	*	*
DYS390	DYS391	DYS392	DYS393
*	*	*	*
DYS437	DYS438	DYS439	DYS448
*	*	*	*
DYS456	DYS458	DYS635 (YGATAC4)	YGATAH4
*	*	*	*

Search By Ancestry

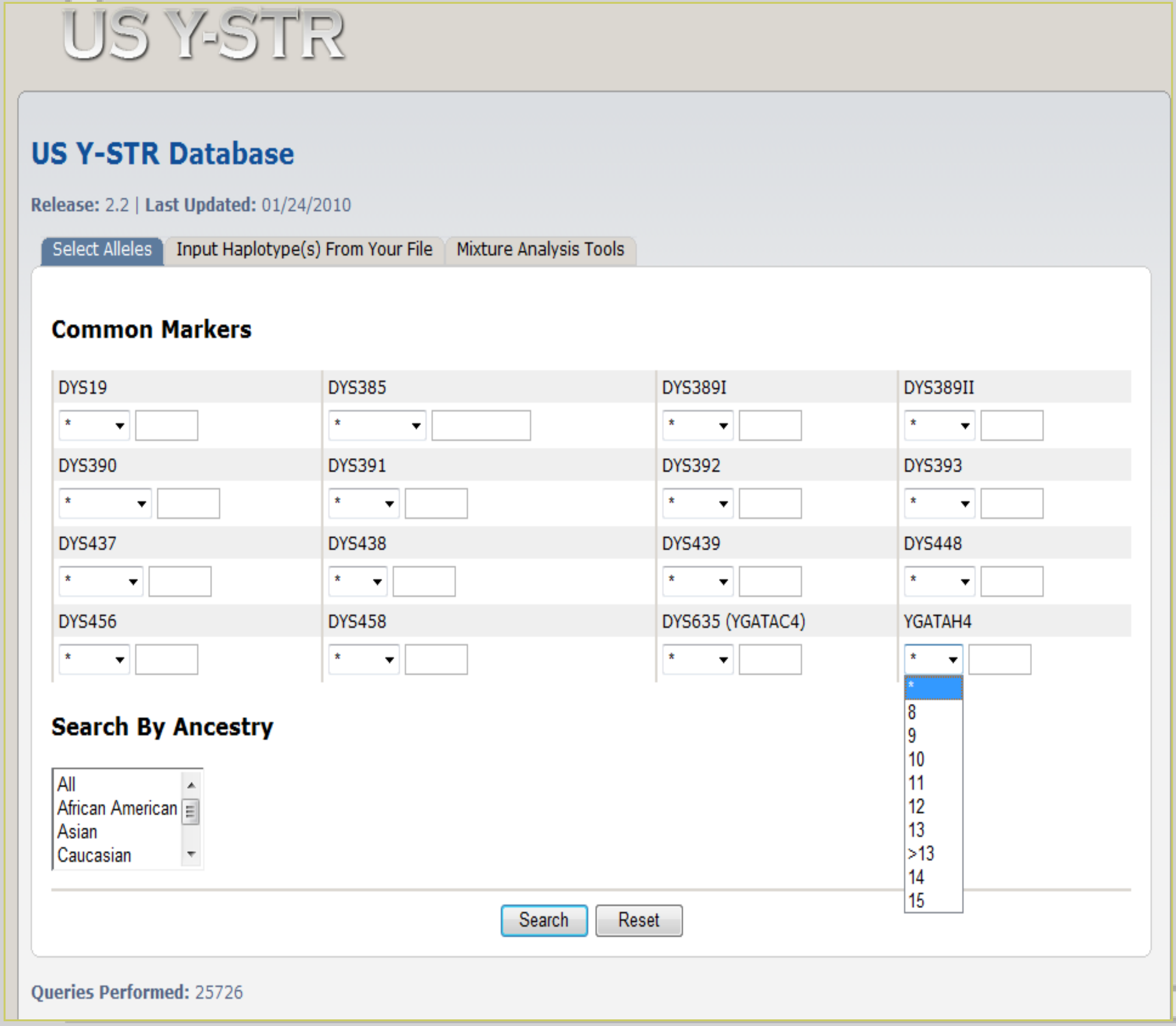
All
African American
Asian
Caucasian

Queries Performed: 25726

Top

Terms of Use | Privacy Policy | Webmaster | Management Login
National Center for Forensic Science • 12354 Research Parkway • Orlando, FL 32826 • 850-207-9377
Developed By The Center for Law Enforcement Technology, Training, & Research

Searching the Database



US Y-STR

US Y-STR Database

Release: 2.2 | **Last Updated:** 01/24/2010

Common Markers

DYS19	DYS385	DYS389I	DYS389II
*	*	*	*
DYS390	DYS391	DYS392	DYS393
*	*	*	*
DYS437	DYS438	DYS439	DYS448
*	*	*	*
DYS456	DYS458	DYS635 (YGATAC4)	YGATAH4
*	*	*	*

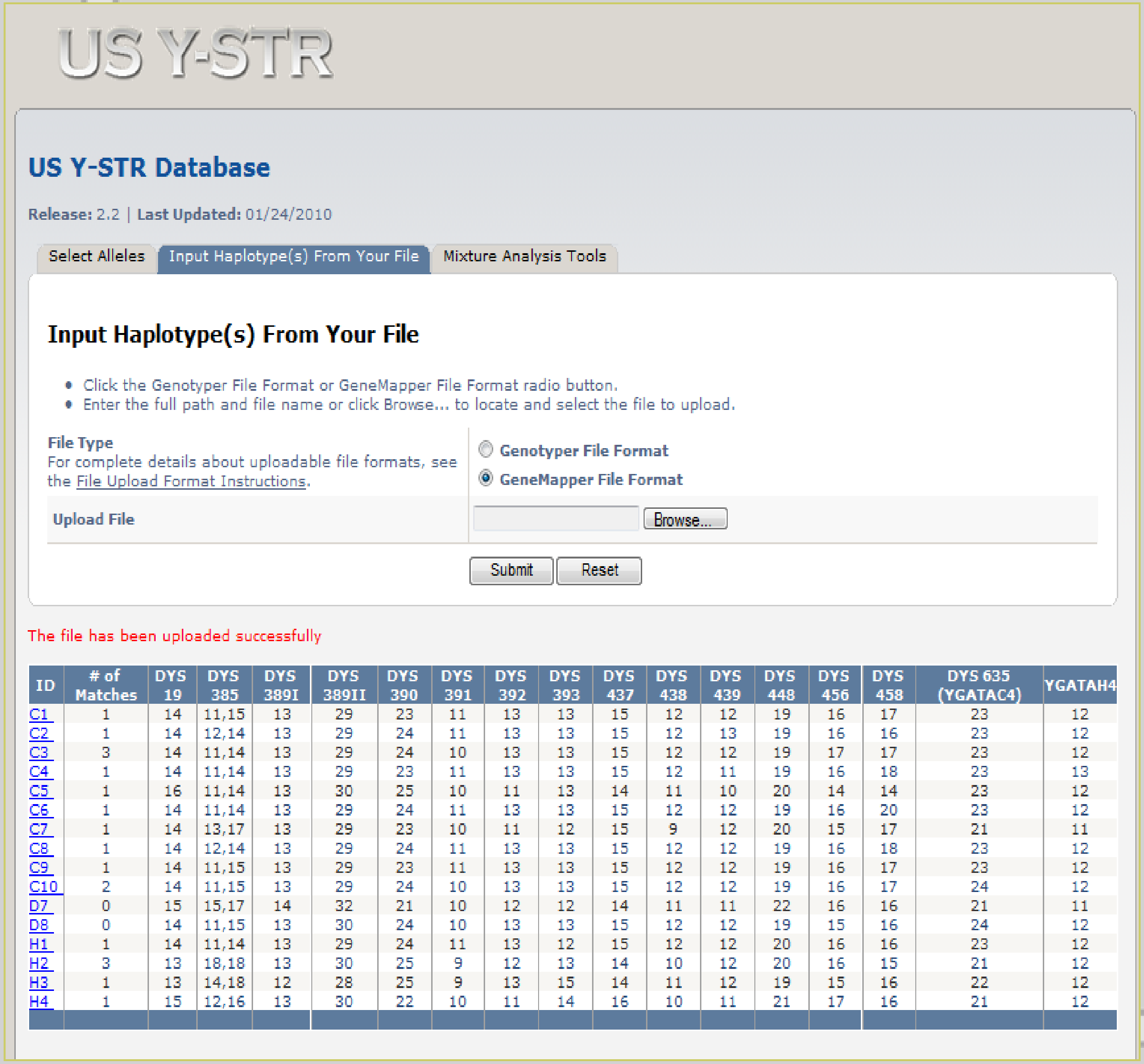
Search By Ancestry

Queries Performed: 25726

- The haplotype of interest is entered by selecting alleles from the drop-down menus, by manually entering alleles into the text boxes, or uploading haplotypes from text files
- Select "All" to query all samples or select ancestry of interest
- Multiple selections can be made using the Ctrl key
- Click "Search" to query the database and review results

NIJ National Institute of Justice

Input Y-STR Data Directly from File



US Y-STR

US Y-STR Database

Release: 2.2 | **Last Updated:** 01/24/2010

Input Haplotype(s) From Your File

- Click the Genotyper File Format or GeneMapper File Format radio button.
- Enter the full path and file name or click Browse... to locate and select the file to upload.

File Type
For complete details about uploadable file formats, see the File Upload Format Instructions.

○ Genotyper File Format
● GeneMapper File Format

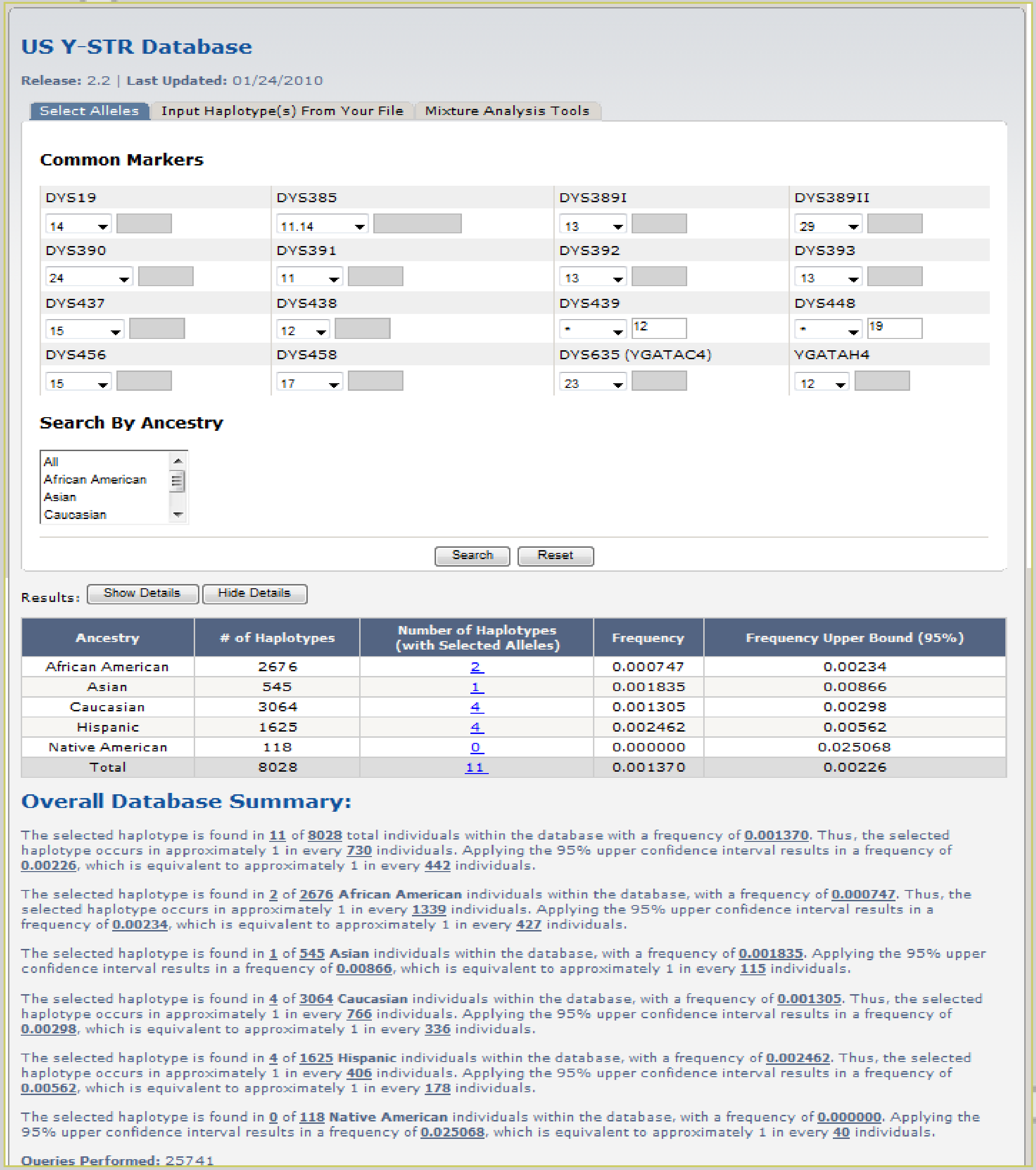
Upload File

The file has been uploaded successfully

ID	# of Matches	DYS 19	DYS 385	DYS 389I	DYS 389II	DYS 390	DYS 391	DYS 392	DYS 393	DYS 437	DYS 438	DYS 439	DYS 448	DYS 456	DYS 458	DYS 635 (YGATAC4)	YGATAH4
C1	1	14	11,15	13	29	23	11	13	13	15	12	12	19	16	17	23	12
C2	1	14	12,14	13	29	24	11	13	13	15	12	13	19	16	16	23	12
C3	3	14	11,14	13	29	24	10	13	13	15	12	12	19	17	17	23	12
C4	1	14	11,14	13	29	23	11	13	13	15	12	11	19	16	18	23	13
C5	1	16	11,14	13	30	25	10	11	13	14	11	10	20	14	14	23	12
C6	1	14	11,14	13	29	24	11	13	13	15	12	12	19	16	20	23	12
C7	1	14	13,17	13	29	23	10	11	12	15	9	12	20	15	17	21	11
C8	1	14	12,14	13	29	24	11	13	13	15	12	12	19	16	18	23	12
C9	1	14	11,15	13	29	23	11	13	13	15	12	12	19	16	17	23	12
C10	2	14	11,15	13	29	24	10	13	13	15	12	12	19	16	17	24	12
D7	0	15	15,17	14	32	21	10	12	12	14	11	11	22	16	16	21	11
D8	0	14	11,15	13	30	24	10	13	13	15	12	12	19	15	16	24	12
H1	1	14	11,14	13	29	24	11	13	12	15	12	12	20	16	16	23	12
H2	3	13	18,18	13	30	25	9	12	13	14	10	12	20	16	15	21	12
H3	1	13	14,18	12	28	25	9	13	15	14	11	12	19	15	16	22	12
H4	1	15	12,16	13	30	22	10	11	14	16	10	11	21	17	16	21	12

- Users can uploaded samples directly from Genotyper and GeneMapper text files to perform multiple, simultaneously searches of the database
- Results returned show the uploaded haplotypes and the number of matches found in the database
- Clicking on the “ID” automatically populates the search fields on the homepage to display the frequency information for that haplotype

Search Results



US Y-STR Database

Release: 2.2 | **Last Updated:** 01/24/2010

Common Markers

DYS19	DYS385	DYS389I	DYS389II
14	11.14	13	29
DYS390	DYS391	DYS392	DYS393
24	11	13	13
DYS437	DYS438	DYS439	DYS448
15	12	- 12	- 19
DYS456	DYS458	DYS635 (YGATAC4)	YGATAH4
15	17	23	12

Search By Ancestry

All
African American
Asian
Caucasian

Results:

Ancestry	# of Haplotypes	Number of Haplotypes (with Selected Alleles)	Frequency	Frequency Upper Bound (95%)
African American	2676	2	0.000747	0.00234
Asian	545	1	0.001835	0.00866
Caucasian	3064	4	0.001305	0.00298
Hispanic	1625	4	0.002462	0.00562
Native American	118	0	0.000000	0.025068
Total	8028	11	0.001370	0.00226

Overall Database Summary:

The selected haplotype is found in **11** of **8028** total individuals within the database with a frequency of **0.001370**. Thus, the selected haplotype occurs in approximately 1 in every **730** individuals. Applying the 95% upper confidence interval results in a frequency of **0.00226**, which is equivalent to approximately 1 in every **442** individuals.

The selected haplotype is found in **2** of **2676** **African American** individuals within the database, with a frequency of **0.000747**. Thus, the selected haplotype occurs in approximately 1 in every **1339** individuals. Applying the 95% upper confidence interval results in a frequency of **0.00234**, which is equivalent to approximately 1 in every **427** individuals.

The selected haplotype is found in **1** of **545** **Asian** individuals within the database, with a frequency of **0.001835**. Applying the 95% upper confidence interval results in a frequency of **0.00866**, which is equivalent to approximately 1 in every **115** individuals.

The selected haplotype is found in **4** of **3064** **Caucasian** individuals within the database, with a frequency of **0.001305**. Thus, the selected haplotype occurs in approximately 1 in every **766** individuals. Applying the 95% upper confidence interval results in a frequency of **0.00298**, which is equivalent to approximately 1 in every **336** individuals.

The selected haplotype is found in **4** of **1625** **Hispanic** individuals within the database, with a frequency of **0.002462**. Thus, the selected haplotype occurs in approximately 1 in every **406** individuals. Applying the 95% upper confidence interval results in a frequency of **0.00562**, which is equivalent to approximately 1 in every **178** individuals.

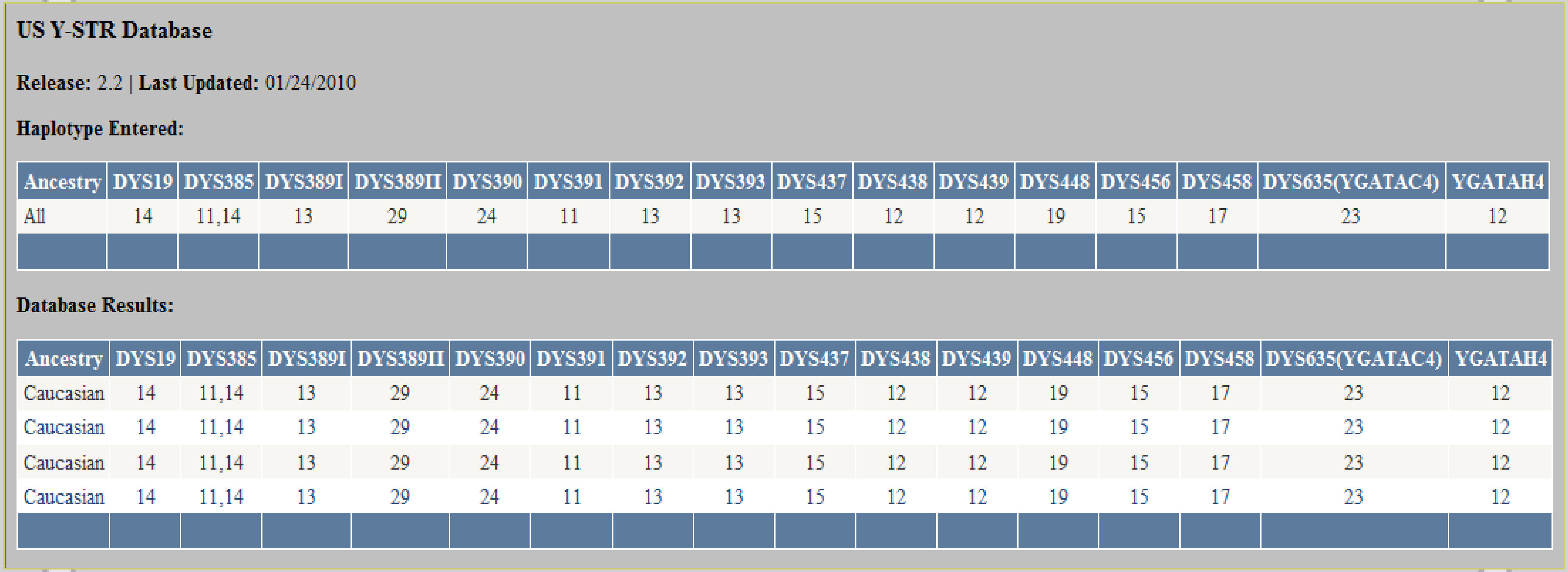
The selected haplotype is found in **0** of **118** **Native American** individuals within the database, with a frequency of **0.000000**. Applying the 95% upper confidence interval results in a frequency of **0.025068**, which is equivalent to approximately 1 in every **40** individuals.

Queries Performed: 25741

- Tabular results display the ancestries selected, number of haplotypes in database having data for the selected loci, number of haplotypes in the database matching the entered haplotype, the frequency, and frequency upper bound (95%)
- "Overall Database Summary" gives statistics statements for the total database and for each ancestry selected
- Blue links under "Number of Haplotypes (with Selected Alleles)" gives a pop-up listing of haplotype searched and matching haplotypes from database

NIJ National Institute of Justice

Matching Haplotypes



US Y-STR Database

Release: 2.2 | **Last Updated:** 01/24/2010

Haplotype Entered:

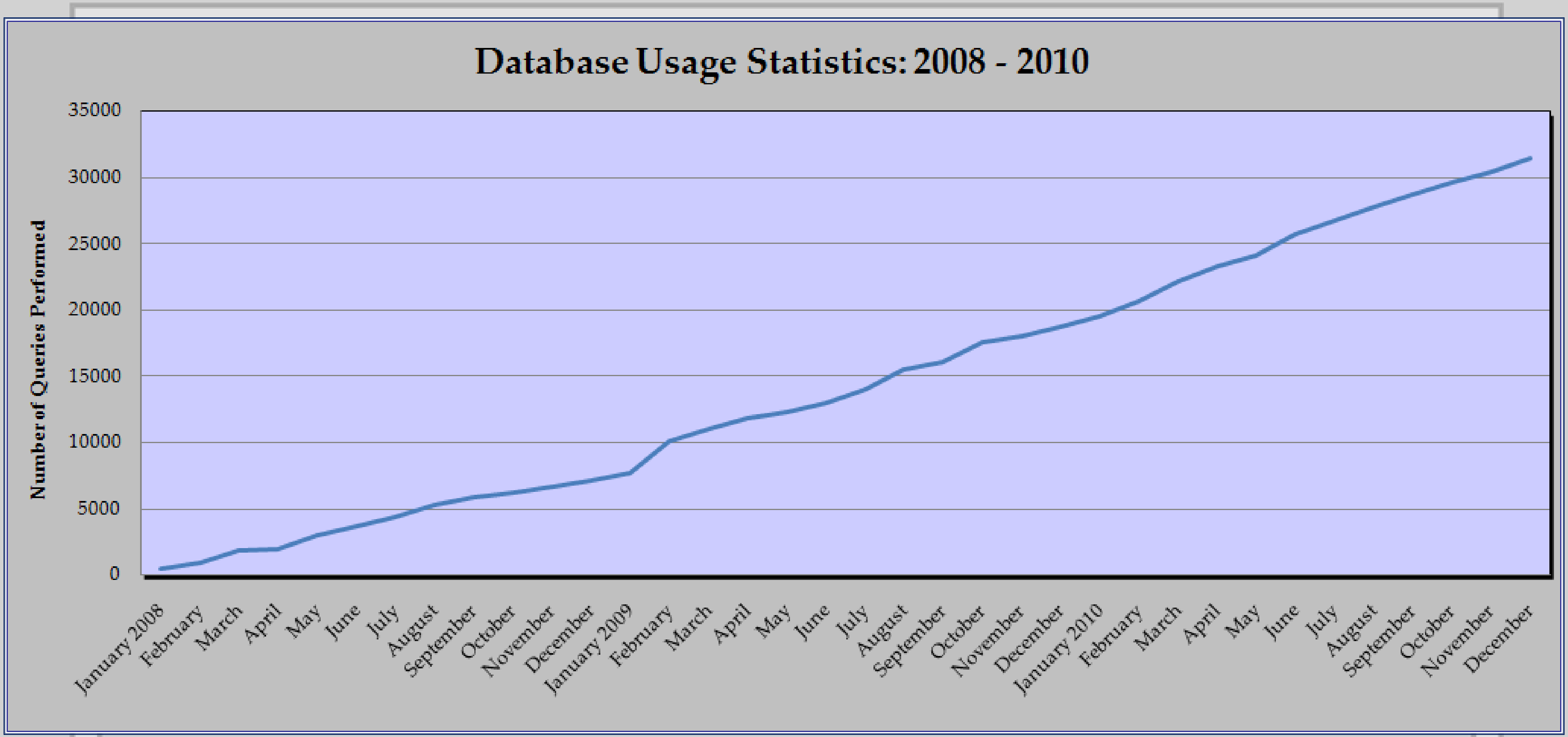
Ancestry	DYS19	DYS385	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635(YGATAC4)	YGATAH4
All	14	11,14	13	29	24	11	13	13	15	12	12	19	15	17	23	12

Database Results:

Ancestry	DYS19	DYS385	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS448	DYS456	DYS458	DYS635(YGATAC4)	YGATAH4
Caucasian	14	11,14	13	29	24	11	13	13	15	12	12	19	15	17	23	12
Caucasian	14	11,14	13	29	24	11	13	13	15	12	12	19	15	17	23	12
Caucasian	14	11,14	13	29	24	11	13	13	15	12	12	19	15	17	23	12
Caucasian	14	11,14	13	29	24	11	13	13	15	12	12	19	15	17	23	12

- Pop-up screen shows haplotype entered, ancestry searched, database results of matching haplotypes, and ancestry in which the matches were found

NIJ
National Institute of Justice



- **Since the release of the US Y-STR Database in January 2008, over 31,000 database search queries have been performed, an average of over 800 per month.**

Publicizing the Database

- An announcement of our intentions to create a consolidated Y-STR database was published in Promega's September 2006 issue of *Profiles in DNA*
- An announcement of the database's availability and its web location was published in Applied Biosystems' February 2008 issue of *Forensic News*
- An article formally announcing the database, including a description of its development and composition, a solicitation for additional samples, and SWGDAM's recommendations for use was published in Promega's March 2008 issue of *Profiles in DNA*
- NCFS designed a brochure specific to the US Y-STR Database that solicits for additional data and/or samples for our own processing, and provides contact information

NIJ National Institute of Justice

Creating and Managing Effective Y-STR Databases

By Jack Ballantyne[1,2], Lyn Fatolitis[2] and Lutz Roewer[3]
[1]University of Central Florida, Department of Chemistry, Orlando, Florida, U.S.A.
[2]National Center for Forensic Science, Orlando, Florida, U.S.A. [3]Institute of Legal Medicine, Charité—University Medicine Berlin, Germany

EDITOR'S NOTE: *This article describes the creation of a new national Y-STR database in the U.S. by the National Center for Forensic Science and the management of an established, worldwide Y-STR database by the Institute of Legal Medicine, Charité—University Medicine Berlin.*

COMPILATION AND MANAGEMENT OF A COMPREHENSIVE U.S. Y-STR REFERENCE DATABASE

By Jack Ballantyne and Lyn Fatolitis

When the DNA profile of a known suspect or victim matches the DNA profile from crime scene evidence, the individual is "included" as a potential source of that evidence. In the U.S., the strength of the match is most often expressed as a statistic that describes the estimated frequency of occurrence of the DNA profile in unrelated individuals within various population groups. Due to the lack of recombination along most of the length of the Y chromosome, Y-STR loci are not statistically independent of one another (unlike standard autosomal STR markers) and are co-inherited as extended haplotypes of linked markers. Therefore, multiplication of single-locus allele frequencies to obtain estimated Y-STR haplotype frequencies is not appropriate. An estimation of the frequency of occurrence of a particular Y-STR haplotype necessitates the use of a counting method, which, with the limited sizes of databases available, produces an estimate that depends entirely upon the size of the database used. Thus, large databases of multi-locus Y-STR haplotypes need to be generated to maximize the probty of Y-STR evidence.

Estimates of the frequency of a particular Y-STR haplotype depend upon the size of the database used. Thus, large databases of multi-locus Y-STR haplotypes need to be generated to maximize the probity of Y-STR evidence.

A large comprehensive European-based Y-STR database is maintained by the Institute of Legal Medicine, Charité—University Medicine Berlin (www.yhrd.org). However, although a subset of this database comprises the SWGDAM core loci, it is less useful for frequency estimates from haplotypes that have been generated using the two most popular commercial kits in the U.S., namely the PowerPlex® Y System[a,b] and AmpFlSTR® Yfiler™ PCR Amplification Kit. There are presently four online searchable Y-STR haplotype databases based in the United States and intended for forensic use. Three are maintained by commercial vendors: Reliagene, Inc., Promega Corporation and Applied Biosystems, Inc. The fourth is maintained by the University of Arizona. The National Center for Forensic Science (NCFS) also maintains a Y-STR database that will soon be available online. These databases differ in the number of Y-STR markers and individuals represented (Table 1), although all possess the SWGDAM core loci. However, these databases are somewhat limited in the number of individuals and loci profiled, which sometimes limits their operational usefulness. For example, the biggest U.S.-based database comprises haplotypes from 4,623 individuals. By combining data from these U.S. databases, a much larger Y-STR database of approximately 16,000 individuals can be created (Table 1), resulting in a significant increase in the probative value of Y-STR evidence. Also, merging the NCFS and University of Arizona databases will increase the number of samples with extended Y-STR loci haplotypes, which may be of assistance to those interested in developing the next generation of Y-STR multiplex systems.

Table 1. Current U.S.-Based Y-STR Databases.

Agency	URL	Number of Markers	Number of Samples
National Center for Forensic Science	To be determined	76	1,396
University of Arizona	http://amadeus.biosci.arizona.edu/~kcaldero/str.php	38	2,518
Applied Biosystems	www.appliedbiosystems.com/yfilerdatabase/	17	3,561
Promega Corporation	www.promega.com/techserv/tools/pplexy/	12	4,004
Reliagene	www.reliagene.com/index.asp?menu_id=rd&content_id=y_frq	11	4,623
Potential Size of National Y-STR Database			16,102

Establishing a national database that incorporates data from a multitude of sources requires the implementation of a number of quality indicator metrics. Quality assurance procedures must be developed to govern the suitability and quality of data from diverse sources. For example, it may be necessary for donors of data to establish analytical prowess by testing externally provided proficiency samples. Since each commercial kit or academic multiplex system uses different primer sets, it will also be essential to ensure that allele calls are equivalent regardless of the multiplex system employed. Importantly, merged data must be purged of duplicate samples that have been submitted by the same donor to multiple databases.

To effectively manage the data, a Y-STR Database Consortium comprised of database stakeholders from commercial companies, academia, the FBI and U.S. crime laboratories was formed at the February 2006 AAFS meeting in Seattle (Table 2). It was agreed that NCFS, a program of the National Institute of Justice (NIJ) hosted by the University of Central Florida, would maintain and manage the consolidated Y-STR database on behalf of stakeholders. The National Institute of Justice is funding this effort. As a group, we are working to collate existing Y-STR data from various commercial and academic sources and have enlisted the aid of geographically diverse crime laboratories to furnish additional samples.

In addition to the immediate goal of expanding the number of individuals in each population group, another key component of the strategy is to allow continuous updating of sample haplotype data using the same samples. This ensures that, as new Y-STR markers are developed, the same samples would be re-typed and a new extended haplotype would be developed. Thus, any laboratory needing haplotype data for any combination of Y-STR markers would be served. The National Y-STR Database will be made available to the U.S. forensic DNA community via the internet, along with tools for obtaining Y-STR haplotype frequencies with confidence intervals needed for calculating matching or paternity probabilities. Other features will include the ability to include or exclude sampled populations, a report-style printout of the results, and flexibility to choose up to 76 Y-STR markers with the potential to search additional Y-STR markers as they become validated and employed by the forensic community.

We expect that the consolidated online database will be accessible to users in the Fall of 2007.

Table 2. Y-STR Database Consortium Members.

Applied Biosystems
Lisa M. Calandro, M.P.H.

Federal Bureau of Investigation
Eric Pokorak, Forensic Examiner
Bruce Budowle, Ph.D.

Minnesota Department of Public Safety
Ann Marie Gross, M.S., F-ABC

National Center for Forensic Science
Jack Ballantyne, Ph.D.
Lyn Fatolitis, Database Manager

National Institute of Justice
John Paul Jones, II, Program Manager

National Institute of Standards and Technology
John M. Butler, Ph.D.

New York City Office of the Chief Medical Examiner
Mecki Prinz, Ph.D.

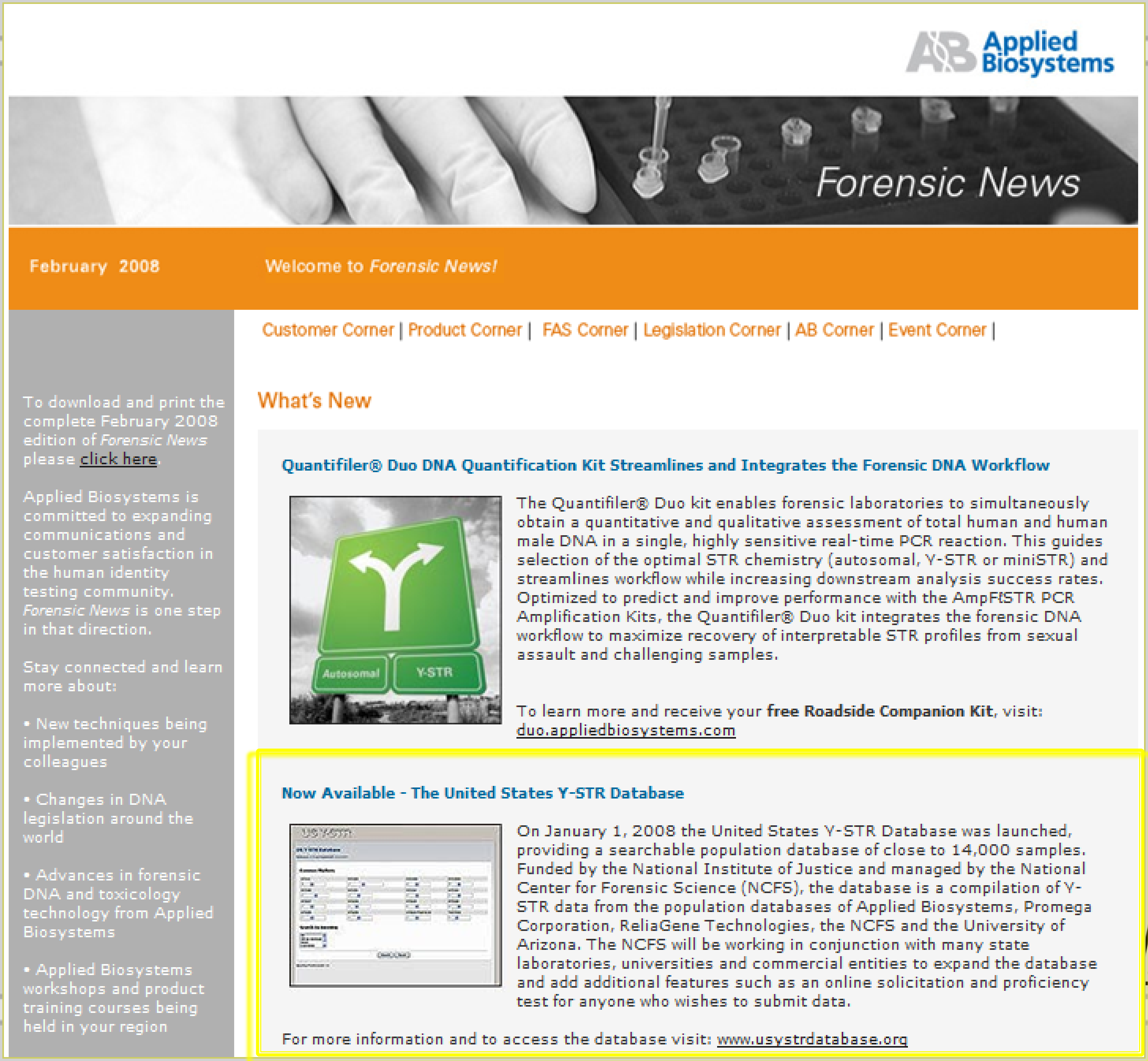
Orchid Cellmark
Cassie Johnson, M.S.

Promega
Curtis D. Knox, B.S., Product Manager

Reliagene
Sudhir Sinha, Ph.D.

University of Arizona
Mike Hammer, Ph.D.

University of North Texas
Arthur Eisenberg, Ph.D.



February 2008

Welcome to *Forensic News!*

Customer Corner | Product Corner | FAS Corner | Legislation Corner | AB Corner | Event Corner |

To download and print the complete February 2008 edition of *Forensic News* please click here.

Applied Biosystems is committed to expanding communications and customer satisfaction in the human identity testing community. *Forensic News* is one step in that direction.

Stay connected and learn more about:

- New techniques being implemented by your colleagues
- Changes in DNA legislation around the world
- Advances in forensic DNA and toxicology technology from Applied Biosystems
- Applied Biosystems workshops and product training courses being held in your region

What's New

Quantifiler® Duo DNA Quantification Kit Streamlines and Integrates the Forensic DNA Workflow

The Quantifiler® Duo kit enables forensic laboratories to simultaneously obtain a quantitative and qualitative assessment of total human and human male DNA in a single, highly sensitive real-time PCR reaction. This guides selection of the optimal STR chemistry (autosomal, Y-STR or miniSTR) and streamlines workflow while increasing downstream analysis success rates. Optimized to predict and improve performance with the AmpFℓSTR PCR Amplification Kits, the Quantifiler® Duo kit integrates the forensic DNA workflow to maximize recovery of interpretable STR profiles from sexual assault and challenging samples.

To learn more and receive your **free Roadside Companion Kit**, visit: duo.appliedbiosystems.com

Now Available - The United States Y-STR Database

On January 1, 2008 the United States Y-STR Database was launched, providing a searchable population database of close to 14,000 samples. Funded by the National Institute of Justice and managed by the National Center for Forensic Science (NCFS), the database is a compilation of Y-STR data from the population databases of Applied Biosystems, Promega Corporation, ReliaGene Technologies, the NCFS and the University of Arizona. The NCFS will be working in conjunction with many state laboratories, universities and commercial entities to expand the database and add additional features such as an online solicitation and proficiency test for anyone who wishes to submit data.

For more information and to access the database visit: www.usystrdatabase.org

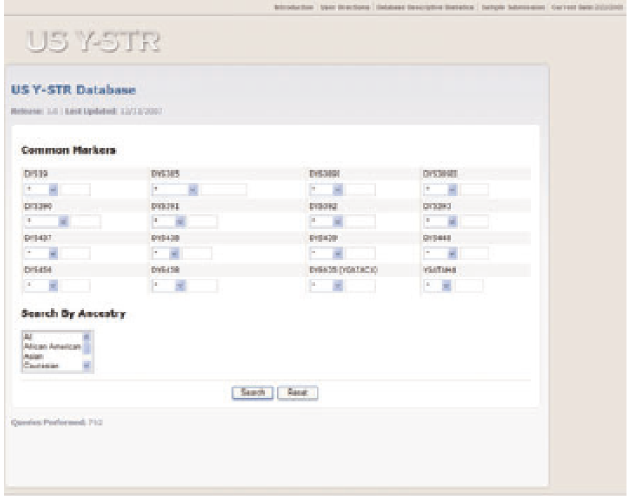
NIJ National Institute of Justice

The US Y-STR Database

By Lyn Fatolitis and Jack Ballantyne
National Center for Forensic Science
University of Central Florida, Orlando, Florida, USA

The National Center for Forensic Science (NCFS), a program of the National Institute of Justice hosted by the University of Central Florida, in conjunction with the Y-STR Consortium created at the American Academy of Forensic Science meeting in 2006 has created a large comprehensive Y-STR reference database of more than 13,000 haplotypes, which is now available online at: **www.usystrdatabase.org** (Figure 1). The US Y-STR Database, a searchable listing of 11- to 17-locus Y-STR haplotypes, was developed by combining data from NCFS with online databases maintained by the University of Arizona, Applied Biosystems, Inc., ReliaGene, Inc., and Promega Corporation (Figure 2).

The database provides tools to obtain Y-STR haplotype frequencies needed to calculate matching or paternity probabilities with confidence intervals. Other features include the ability to simultaneously upload multiple haplotypes for searches directly from Genotyper® and GeneMapper® text files, the ability to include or exclude sampled populations, and a report-style printout of the results. Samples are divided into five forensically relevant ancestries: African-American, Asian, Caucasian, Hispanic and Native American.



Introduction | User Directions | Database Descriptive Statistics | Sample Submission | Current Date [illegible]

US Y-STR

US Y-STR Database

Release: 1.0 | Last Updated: 12/13/2007

Common Markers

DYS19 DYS385 DYS389I DYS389II
DYS390 DYS391 DYS392 DYS393
DYS437 DYS438 DYS439 DYS448
DYS456 DYS458 DYS635 (YGATAC4) YGATAH4

Search By Ancestry

All
African American
Asian
Caucasian

Search | Reset

Queries Performed: 712

Figure 1. The US Y-STR Database interface.

The US Y-STR Database is a comprehensive and searchable Y-STR reference database containing more than 13,000 11- to 17-locus Y-STR haplotypes divided into five forensically relevant ancestries.

The goal of the database is to expand continuously the number of individuals (N) for each ancestral group and geographical location. NCFS is currently creating quality assurance procedures to govern the suitability of data solicited from diverse sources for inclusion in the database, including a proficiency testing procedure for labs who wish to contribute Y-STR haplotypes in the future, ensuring that they can correctly genotype samples. Information about submitting Y-STR data to expand the database will soon be accessible from the database web site.

It is important to note that a number of individual samples were shared among the contributing data sets. All duplicate samples were removed to ensure that each sample in the consolidated database is from a unique individual. Any population group that did not contain at least 50 samples was also removed. These data reconciliation and reorganization steps have resulted in the consolidated US Y-STR Database having slightly different sample numbers than those found in the curated databases currently maintained by the individual contributing institutions. This population database is intended for use in estimating haplotype population frequencies for forensic casework purposes. All donors are anonymous, and original electropherograms do not exist in a curated fashion. All submitting entities are solely responsible for their data. In the event that details of a certain population sample are requested via the judicial process, the request will be redirected to the collaborating scientists and their institutions.

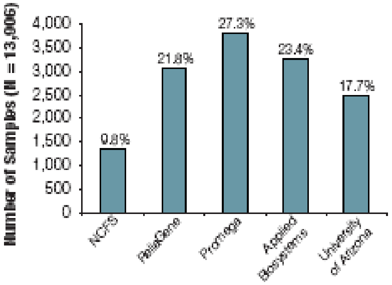


Figure 2. Data contributors to the US Y-STR Database, Release 1.0.

SWGDAM is currently considering recommending the use of the consolidated database for population frequency estimation in casework. In the reporting of matches, haplotype searches of the population database should be conducted using all loci for which results were obtained from the evidentiary sample. In cases where less information is obtained from the known sample, only those loci for which results were obtained from both the known and evidentiary sample should be used in the population database search.

Our Faculty

Carrie Whitcomb, MSFS
Director, National Center for Forensic Science

Jack Ballantyne, PhD
Associate Director of Research, National Center for Forensic Science;
Assistant Director for Biological Evidence

Michael Sigman, PhD
Assistant Director for Physical Evidence

Philip Craiger, PhD
Assistant Director for Digital Evidence

Richard Blair, PhD
Assistant Professor, Chemistry and Forensic Science

NCFS Partners

Bureau of Alcohol, Tobacco, Firearms & Explosives
Federal Bureau of Investigation
I.D.E.A.L. Technology, Inc
National Institute of Justice
NEWTEC Services Group, Inc.
University of Central Florida
US Secret Service
Xiotech

Affiliated Organizations

Digital Forensics Certification Board
Orange County Sheriff's Office
Scientific Working Group on Imaging Technology
Scientific Working Group on Digital Evidence
Seminole County Sheriff's Office
South Carolina Law Enforcement
Technical Working Group on Fire and Explosives

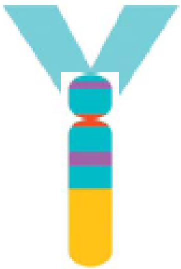
How to Contact Us

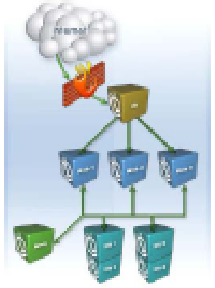
National Center for Forensic Science
University of Central Florida
P.O. Box 162367
Orlando, FL 32816-2367

Phone: (407) 823-6469 Fax: (407) 823-3162
http://www.ncfs.org
natlctr@mail.ucf.edu

March 2010

National Center for Forensic Science





US Y-STR Database

www.usystrdatabase.org

NIJ

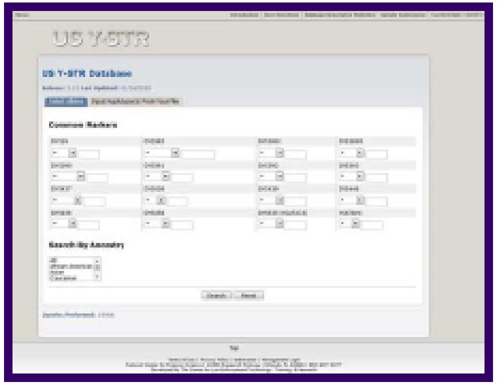
The US Y-STR Database

The National Center for Forensic Science (NCFS) in conjunction with the Y-STR Consortium has created a large comprehensive Y-STR reference database that can be used to obtain the Y-STR haplotype frequencies needed to calculate matching or paternity probabilities with confidence intervals for forensic casework purposes.

The Y-STR Consortium, formed at the 2006 American Academy of Forensic Sciences meeting in Seattle, Washington to effectively manage the data and to aid in the developmental design of the Database, was comprised of database stakeholders from commercial companies, academia, the FBI and US crime laboratories.

The US Y-STR Database, a searchable listing of 11- to 17-locus Y-STR haplotypes, was developed by combining data from NCFS with the online databases maintained by the University of Arizona, Applied Biosystems, Inc., ReliaGene, Inc., and Promega Corporation. By combining data from these US databases, a much larger Y-STR database was created, resulting in a significant increase in the probative value of Y-STR evidence.

Release 1.0 of the US Y-STR Database was made available online to the forensic community on January 3, 2008 at www.usystrdatabase.org and www.usystrdatabase.com and was initially comprised of 13,906 Y-STR haplotypes.

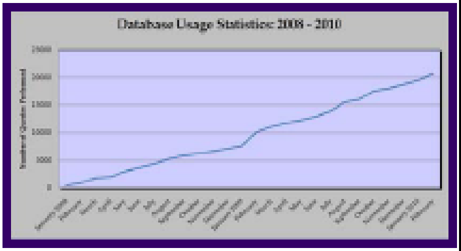


US Y-STR Database Homepage/ User Interface

The user interface provides additional database information via the links at the top of the page, allows users to enter haplotypes for database queries by either selecting alleles from the drop-down menus, manually entering alleles in the text boxes, or automatically uploading multiple haplotypes directly from their genotyping software applications. It also provides the ability to include or exclude sampled populations and a report-style printout of the results. Samples are divided into five forensically-relevant ancestries: African-American, Asian, Caucasian, Hispanic, and Native American.

The objectives for the US Y-STR Database are to expand continuously the number of individuals (N) for each ancestral group and geographical location and to incorporate improvements in design and functionality based on recommendations from users in the field.

The current version of the Database, Release 2.2, is comprised of 18,199 haplotypes. SWGDAM has recommended that the US Y-STR Database be used for calculating Y-STR population frequencies. Since its release, there have been over 20,000 search queries performed, an average of nearly 800 per month.



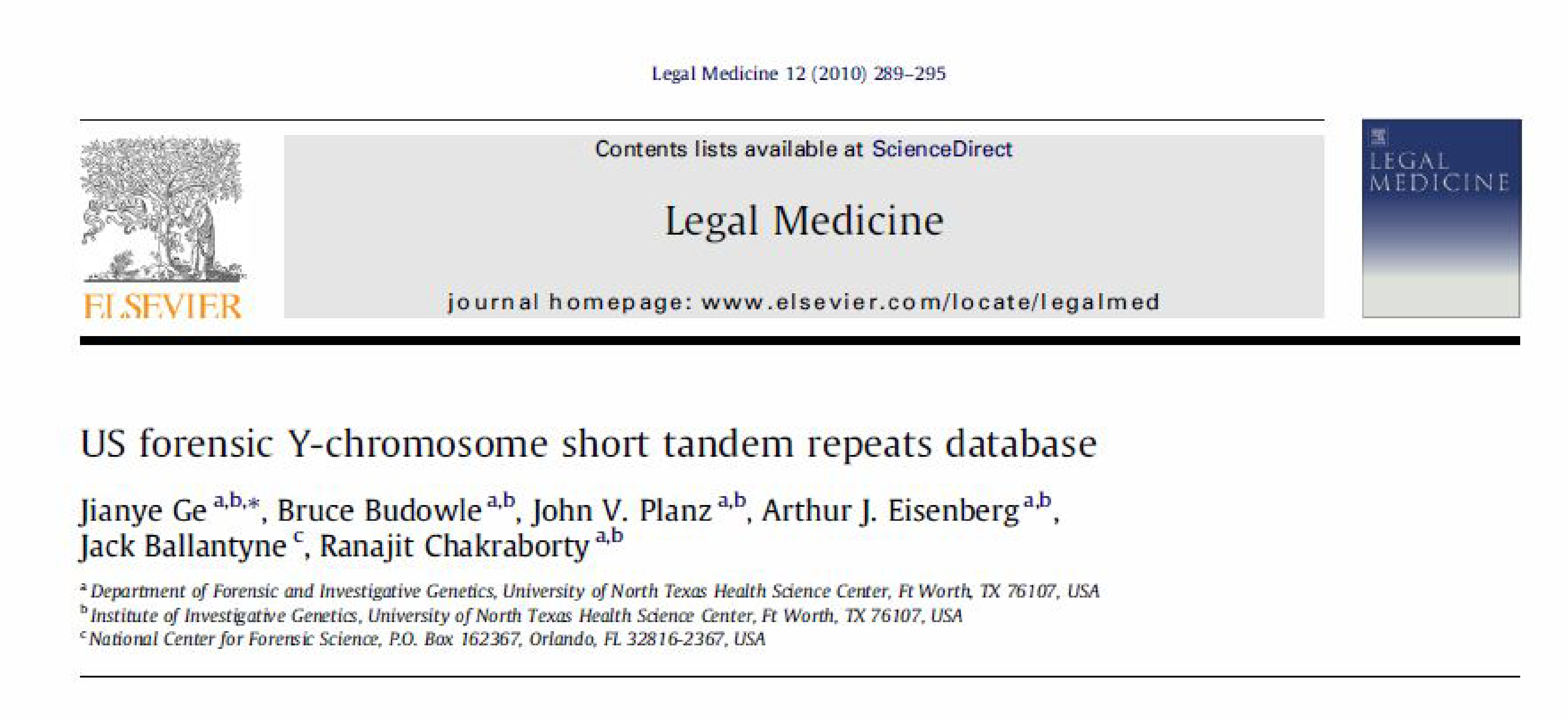
Database Usage Statistics

NCFS has developed a sample submission protocol that invites all forensic laboratories and institutions to contribute data to the US Y-STR Database by first successfully completing a five-sample quality control exercise and then typing samples from any specific population group with any commercial kit. **If you are interested in donating samples or Y-STR data, please contact us or visit the Database website for more information.**

How to Contact Us

National Center for Forensic Science
University of Central Florida
P.O. Box 162367
Orlando, FL 32816-2367

Phone: 407-823-4041 Fax: 407-823-4042
lfatolit@mail.ucf.edu
www.usystrdatabase.org



Legal Medicine 12 (2010) 289–295

Contents lists available at ScienceDirect

Legal Medicine

journal homepage: www.elsevier.com/locate/legalmed

US forensic Y-chromosome short tandem repeats database

Jianye Ge [a,b,*], Bruce Budowle [a,b], John V. Planz [a,b], Arthur J. Eisenberg [a,b],
Jack Ballantyne [c], Ranajit Chakraborty [a,b]

[a] Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Ft Worth, TX 76107, USA
[b] Institute of Investigative Genetics, University of North Texas Health Science Center, Ft Worth, TX 76107, USA
[c] National Center for Forensic Science, P.O. Box 162367, Orlando, FL 32816-2367, USA

Database Expansion / Data Solicitation

- **Created Sample Submission section on database website**
  - **Created quality control competency testing procedure using liquid blood samples donated by UNT for this purpose**
  - **Certificate of Participation is issued to qualifying laboratories**
  - **Sample submission template and information available on website**
- **Solicitation for data was posted on our behalf by Lynne Burley of Santa Clara County Crime Lab on the Yahoo group, *forens-DNA,* a technical discussion group of forensic DNA technology**
- **Updated U.S. Department of Justice and NCFS websites to solicit for samples and / or data**
- **Routinely make appeals for samples and data at all meetings, presentations, workshops, etc.**
- **Hired Research Technician to obtain and process samples in-house for inclusion into the database**

Certificate No.: 00010

DuPage County Forensic Science Center
501 North County Farm Road
Wheaton, IL 60187
USA

The National Center for Forensic Science
University of Central Florida
12354 Research Parkway, Suite 225
Orlando, FL 32816-2367
USA

US Y-STR Database

Certificate of Participation

DuPage County
Forensic Science Center

has participated in the Y-STR Haplotyping
Quality Assurance Exercise

The alleles at all loci tested have been typed correctly
according to the published nomenclature and the
ISFG guidelines for Y-STR Analysis
(Int J Legal Med 114 (2001) 305-309)

Granted: March 30, 2010

Jack Ballantyne, Ph.D.
Associate Director (Research)

Lyn Fatolitis
US Y-STR Database Manager

QC Participants

- To date, ten laboratories have requested and completed the QC exercise, allowing submission of their haplotype data for inclusion into the database
  - IL State Police Crime Laboratory
  - Jan Bashinski DNA Crime Laboratory, CA Department of Justice
  - Orange County CA Sheriff – Coroner, Forensic Science Services
  - State of Connecticut Forensic Science Services Laboratory
  - County of Santa Clara CA Crime Laboratory, Office of the DA
  - CA Department of Justice, Sacramento Crime Laboratory
  - WA State Patrol Crime Laboratory – Vancouver
  - Marshall University Forensic Science Center
  - AZ Department of Public Safety Central Regional Crime Laboratory
  - DuPage County IL Forensic Science Center

Users' Feedback – Database Improvements

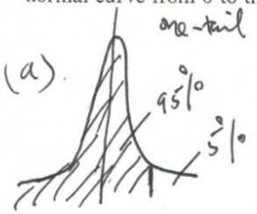
- Several changes and improvements have been made based upon recommendations and suggestions from users in the field
  - Followed suggestions from the Santa Clara County DA Crime Laboratory and the Centre of Forensic Sciences in Ontario to alter some of the verbiage in the displayed results
  - Added a "News" section to the database homepage to allow for announcements and updates to keep users informed
  - Created and validated an automatic haplotype upload interface, allowing users to simultaneously upload multiple haplotypes directly from Genotyper® and GeneMapper® text files for database searches, modeled after Applied Biosystems' Yfiler Database interface
  - Adjusted the > (greater than) and < (less than) queries of the database. Rather than returning just exact matches, the query now returns all alleles greater than or less than the entry and calculates these haplotypes into the statistic statements.

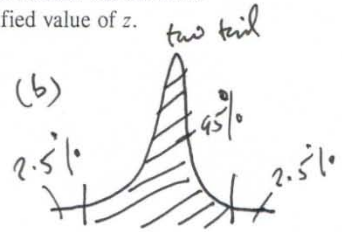
...continued

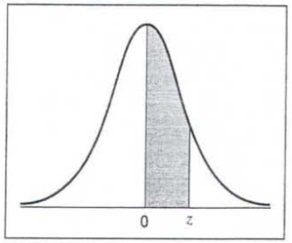
- Followed suggestions from the New Jersey State Police to add links to database publications, to add verbiage clarifying the difference between the US Y-STR Database and CODIS, to provide a single PDF download of all database information contained on website, and to update the “Database Descriptive Statistics” PDF to include additional information on the discrimination potential
- Followed suggestion from DNA Labs International to provide descriptive statistics for all prior releases for court purposes
- AFDIL alerted us to an automatic file upload error when uploading multiple samples from GeneMapper ID-X, discovered table exported by GMID-X is different than previous versions, worked with IT to fix the issue
- Followed suggestion from Bruce Weir and John Buckleton and corrected the formula used to calculate confidence intervals when an entered haplotype is observed in the database
- Received requests from numerous agencies to add mixture interpretation / deconvolution tools, recently added two to the website

Areas of a Standard Normal Distribution

The table entries represent the area under the standard normal curve from 0 to the specified value of *z*.

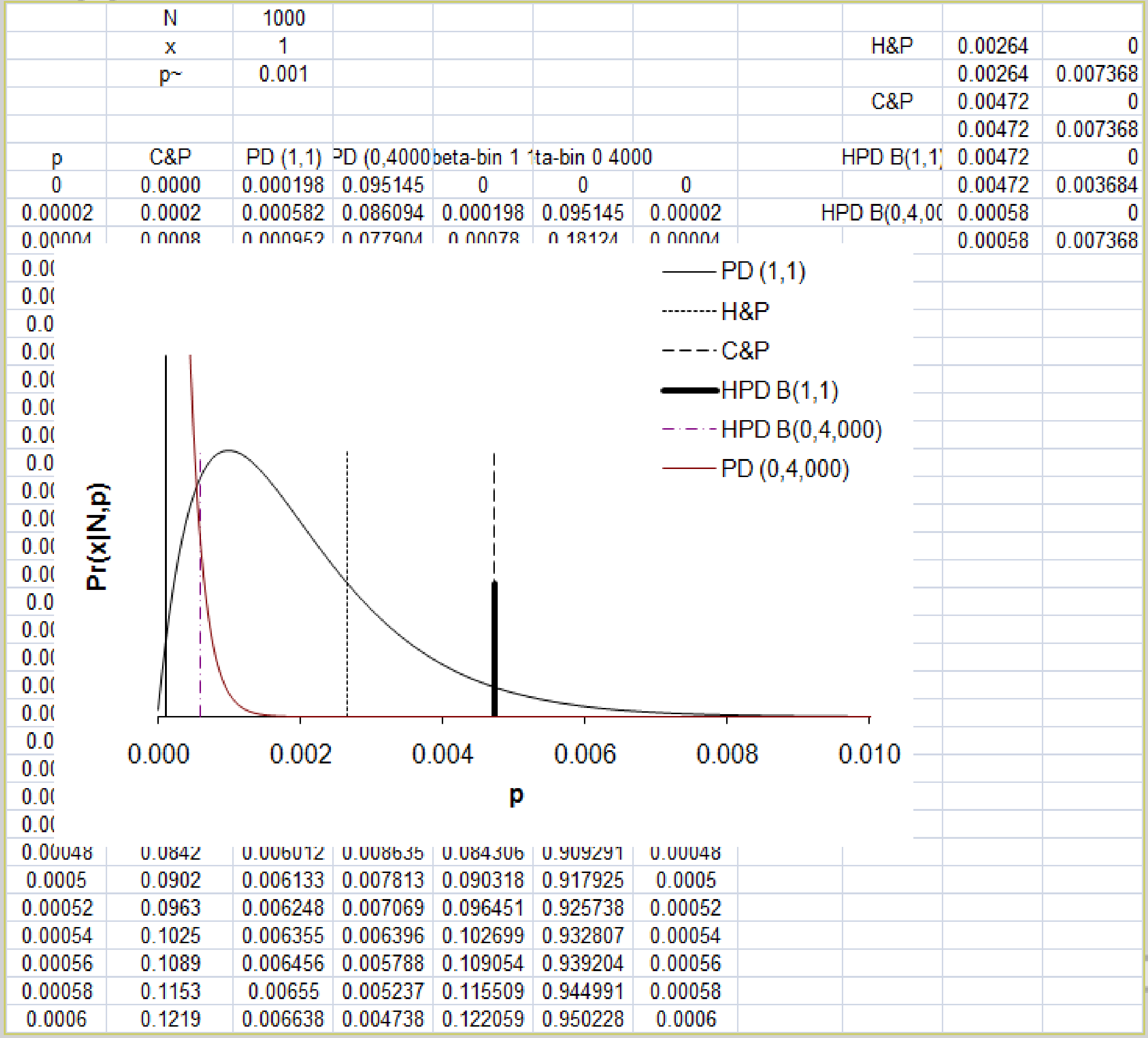






z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									

Clopper-Pearson Exact Confidence Interval



	N	1000								
	x	1						H&P	0.00264	0
	p~	0.001							0.00264	0.007368
								C&P	0.00472	0
									0.00472	0.007368
p	C&P	PD (1,1)	⁾D (0,4000	beta-bin 1 1	ta-bin 0 4000			HPD B(1,1	0.00472	0
0	0.0000	0.000198	0.095145	0	0	0			0.00472	0.003684
0.00002	0.0002	0.000582	0.086094	0.000198	0.095145	0.00002		HPD B(0,4,0(	0.00058	0
0.00004	0.0008	0.000952	0.077904	0.00078	0.18124	0.00004			0.00058	0.007368
0.00048	0.0842	0.006012	0.008635	0.084306	0.909291	0.00048				
0.0005	0.0902	0.006133	0.007813	0.090318	0.917925	0.0005				
0.00052	0.0963	0.006248	0.007069	0.096451	0.925738	0.00052				
0.00054	0.1025	0.006355	0.006396	0.102699	0.932807	0.00054				
0.00056	0.1089	0.006456	0.005788	0.109054	0.939204	0.00056				
0.00058	0.1153	0.00655	0.005237	0.115509	0.944991	0.00058				
0.0006	0.1219	0.006638	0.004738	0.122059	0.950228	0.0006				

- CI formula was changed to Clopper-Pearson ‘exact’ CI formula
- Change came about as a result of discussions with Bruce Weir and John Buckleton who provided us with an excel spreadsheet to perform the more correct calculations
- Spreadsheet was incorporated into the database on March 26, 2010

Clopper-Pearson Exact Confidence Interval

Clopper, C.J. and E.S. Pearson, *The use of confidence or fiducial intervals illustrated in the case of the binomial*. Biometrika (1934). **26**: p. 64-69.

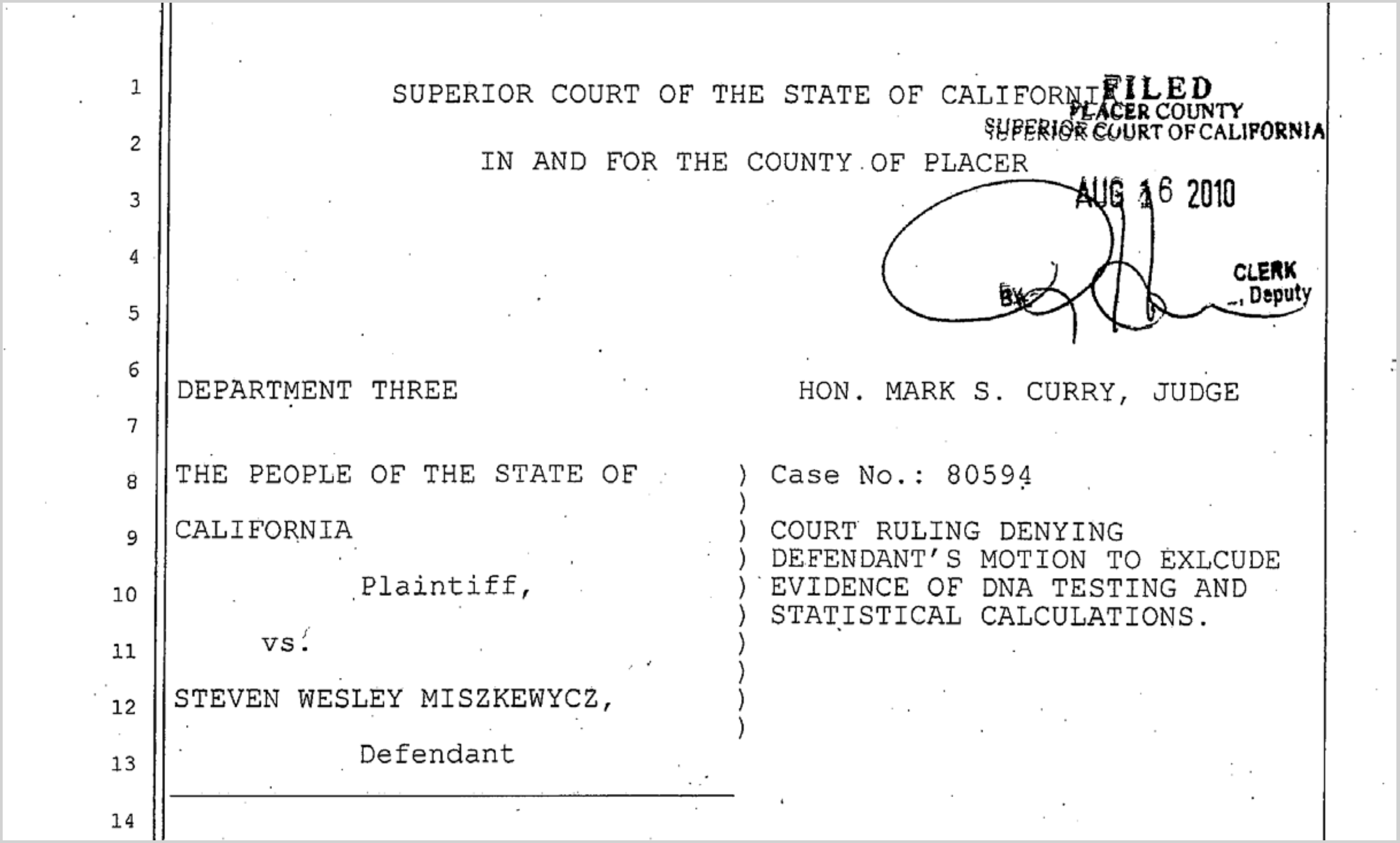
$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1 - p_0)^{n-k} = 0.05$$

The formula is the cumulative binomial distribution for all values from 0 matches to k matches given a database size of N and a frequency of p. Since N and k are fixed after a search, the goal is to determine the p at which 95% of the observations are expected to be more than k, and 5% of the observations (the 0.05 in the formula) are expected to be between 0 and k. This program increases p by small increments until this balance point (~95% of the possible comparisons expected to give you >k matches, and ~5% expected to give you k or fewer matches) has been reached. By finding what amounts to a left-hand 1-sided 95% confidence interval (i.e., the *lower* limit) for the distribution of possible matches given the frequency p (all as it relates to the k and N observed from the search), this then also provides an 95% *upper* limit for p. Beyond that point, it is considered too unlikely that a haplotype with a more common frequency would give so few matches.

Exact vs. Normal Confidence Intervals

n	X	P	HP (1-tail)	HP (2-tail)	CP
100	1	0.01	0.026	0.029	0.047
	2	0.02	0.043	0.047	0.062
	10	0.10	0.149	0.159	0.164
1,000	1	0.001	0.0026	0.0029	0.0047
	2	0.002	0.0043	0.0048	0.0063
	10	0.010	0.0152	0.0162	0.0169
10,000	1	0.0001	0.0003	0.0003	0.0005
	2	0.0002	0.0004	0.0005	0.0006
	10	0.0010	0.0015	0.0016	0.0017

Frye Hearings on the National Y-STR Database



SUPERIOR COURT OF THE STATE OF CALIFORNI

IN AND FOR THE COUNTY OF PLACER

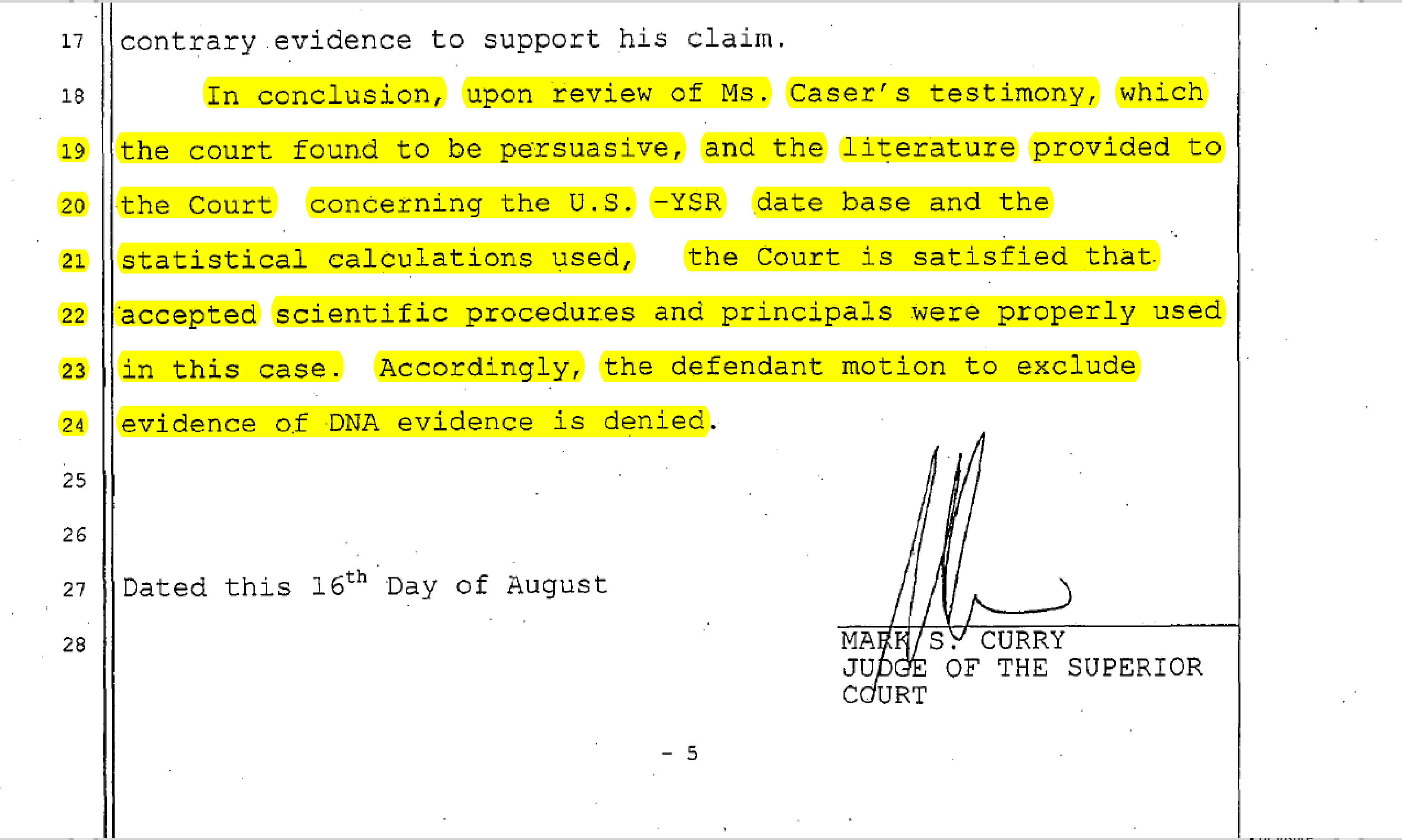
FILED
PLACER COUNTY
SUPERIOR COURT OF CALIFORNIA
AUG 16 2010
CLERK
By ___, Deputy

DEPARTMENT THREE HON. MARK S. CURRY, JUDGE

THE PEOPLE OF THE STATE OF CALIFORNIA, Plaintiff, vs. STEVEN WESLEY MISZKEWYCZ, Defendant	Case No.: 80594 COURT RULING DENYING DEFENDANT'S MOTION TO EXLCUDE EVIDENCE OF DNA TESTING AND STATISTICAL CALCULATIONS.

"Specifically, he contends that the statistical analysis applied in this case is faulty because of an unreliable or unknown data base used to formulate the statistics"

NIJ National Institute of Justice



contrary evidence to support his claim.

In conclusion, upon review of Ms. Caser's testimony, which the court found to be persuasive, and the literature provided to the Court concerning the U.S. -YSR date base and the statistical calculations used, the Court is satisfied that accepted scientific procedures and principals were properly used in this case. Accordingly, the defendant motion to exclude evidence of DNA evidence is denied.

Dated this 16th Day of August

MARK S. CURRY
JUDGE OF THE SUPERIOR
COURT

Frye Hearings on the National Y-STR Database

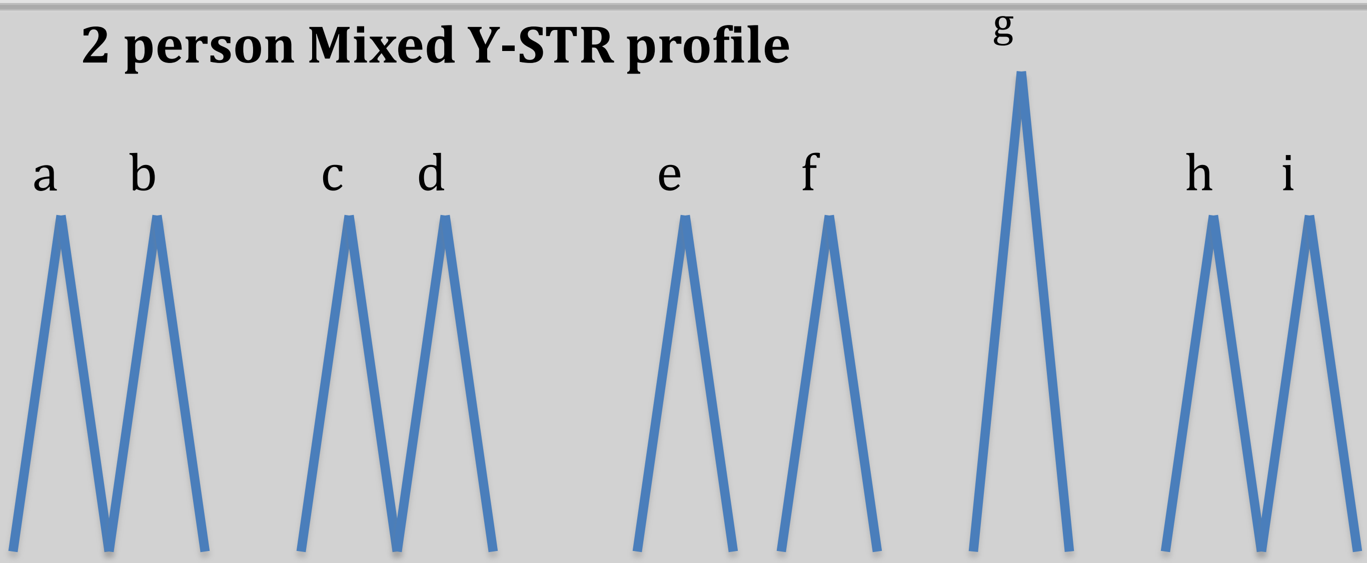
- State of Kansas v Gonzalez
- Judge Pokorny, Seventh Judicial District for the District of Kansas, Douglas County, KS
- Challenge that over a period of six months the frequency of the evidence/suspect haplotype changed from 1/2717 to 1 in 1786
- 4 October 2010: found that Y-STR database is fit for purpose and motion to deny/exclude Y-STR haplotype evidence denied

Future Goals of the US Y-STR Database

- To continue to solicit data and / or samples from forensic laboratories in an effort to expand continuously the number of individuals (N) for each ancestral group and geographical location, plan updates approximately every 6 months if samples are available
- To continuously incorporate the suggestions and recommendations received from users to improve the design and functionality of the database to better serve the needs of the forensic community
- Funding/Location??

NIJ
National Institute of Justice

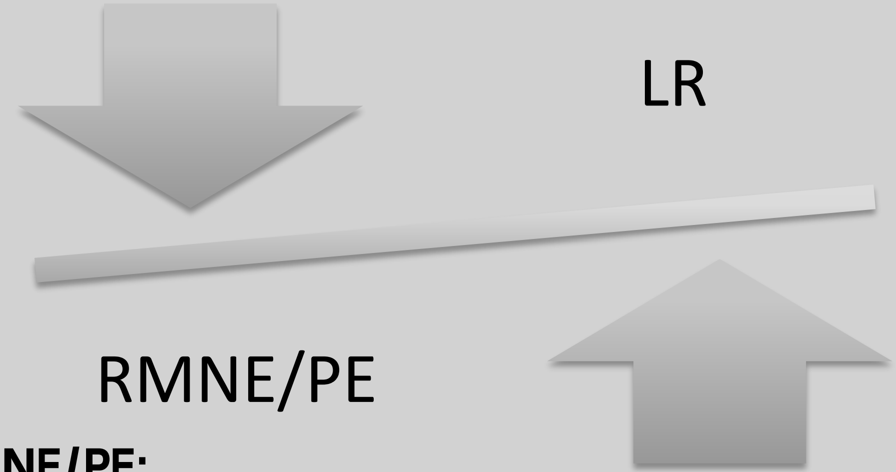
2 person Mixed Y-STR profile



No. of possible haplotypes = 2^n, where n = no of (non 385) loci exhibiting two alleles (as opposed to one) = 2^4 = 16

No of haplotypes = 2^n x (k(k + 1)/2) where k = no of 385 peaks a database search (N = 5000) reveals that 9 of the 16 haplotypes have been observed at least once:

acegh	**acegi**	**acfgh**	**acfgi**	**adegh**	**adegi**	**adfgh**	**adfgi**
1	**1**	**0**	**0**	**0**	**0**	**1**	**1**
bcegh	**bcegi**	**bcfgh**	**bcfgi**	**bdegh**	**bdegi**	**bdfgh**	**bdfgi**
0	**2**	**1**	**1**	**0**	**0**	**1**	**1**



LR

RMNE/PE

V = acegh
S = bdfgi = 0.0034

<u>RMNE/PE</u>:

1. ***<u>Using only haplotypes OBSERVED in the database</u>***
   Count how many times each of the possible haplotypes are <u>observed</u> in the database, sum them, determine an overall frequency and that constitutes the PI/RMNE.
   Thus 10/5000 = 1/500 = 0.002 which, with the binomial 95% CI, is = 0.00338 (1/296)
   Then 1-PI = PE = 1-0.0038 = 0.9962 (ie 99.6 % can be excluded)
2. ***<u>Using all haplotypes whether or not they have been observed in the database</u>***
   However, what about the 7 haplotypes that could be components of the mixture but whose presence has not been accounted for?

- **The 7 haplotypes could each occur with a frequency of 3/N**
  - **Thus 7 x 3/N = 21/N = 21/5000 = 0.0042**
- **The PI to take into account all possible contributors to the mixture is 0.00338 + 0.0042 = 0.00758 (1/132)**
  - **PE = 1- 0.00758 = 0.99242 = 99.2%**
- **PE = 99.6% (observed possible haplotypes) versus 99.2% (all possible haplotypes)**

$$LR = \frac{\Pr\,(E|Hp)}{\Pr(E|Hd)}$$

Hp = Prosecution hypothesis
= Mixture comprises (1 known + 1 unknown) OR (2 known) individuals

In this case assume = victim + suspect DNA comprises the mixture, thus Pr (E|Hp) = 1

Hd = Defense hypothesis
= the mixture comprises DNA from 2 random unrelated males

haplotype	alleles	count	Pr (H_i) (with binomial sampling correction)
H_1 = victim	**acegh**	1	0.0034
H_2	**acegi**	1	0.0034
H_3	**acfgh**	0	0.0006
H_4	**acfgi**	0	0.0006
H_5	**adegh**	0	0.0006
H_6	**adegi**	0	0.0006
H_7	**adfgh**	1	0.0034
H_8	**adfgi**	1	0.0034
H_9	**bcegh**	0	0.0006
H_{10}	**bcegi**	2	0.0068
H_{11}	**bcfgh**	1	0.0034
H_{12}	**bcfgi**	1	0.0034
H_{13}	**bdegh**	0	0.0006
H_{14}	**bdegi**	0	0.0006
H_{15}	**bdfgh**	1	0.0034
H_{16} = suspect	**bdfgi**	1	0.0034

There are sixteen combinations of haplotypes that can produce the evidence haplotype (e.g. $H_1 + H_{16}$)

NIJ
National Institute of Justice

Combinations of haplotypes that could comprise the 2 person mixture

- **(1 + 16) = (16 + 1) Pr = 0.0034 x 0.0034 = 0.00001156**
- **(2 + 15) = (15 + 2) Pr = 0.0034 x 0.0034 = 0.00001156**
- **(3 + 14) = (14 + 3) Pr = 0.0006 x 0.0006 = 0.00000036**
- **(4 + 13) = (13 + 4) Pr = 0.0006 x 0.0006 = 0.00000036**
- **(5 + 12) = (12 + 5) Pr = 0.0006 x 0.0034 = 0.00000204**
- **(6 + 11) = (11 + 6) Pr = 0.0006 x 0.0034 = 0.00000204**
- **(7 + 10) = (10 + 7) Pr = 0.0034 x 0.0068 = 0.00002312**
- **(8 + 9) = (9 + 8) Pr = 0.0034 x 0.0006 = 0.00000204**

$\sum$ = 0.00005308

LR Calculation:

$$LR = \frac{1}{2\left\{\begin{array}{c}\Pr(H1)\Pr(H16)+\Pr(H2)\Pr\,(H15)+\Pr(H3)\Pr(H14)+\Pr(H4)\Pr(H13)+\\ \Pr(H5)\Pr(H12)+\Pr(H6)\Pr(H11)+\Pr(H7)\Pr(H10)+\Pr(H8)\Pr(H9)\end{array}\right\}}$$

= 1/(2 x 0.00005308)
= 1/0.00010616
= 9419

Thus the DNA profiling results were 9419 times more likely if the mixture comprised DNA from the victim and the suspect than if it came from two random unrelated individuals

-is this true? (random individuals chosen from the database or might be expected to be present in the database?)

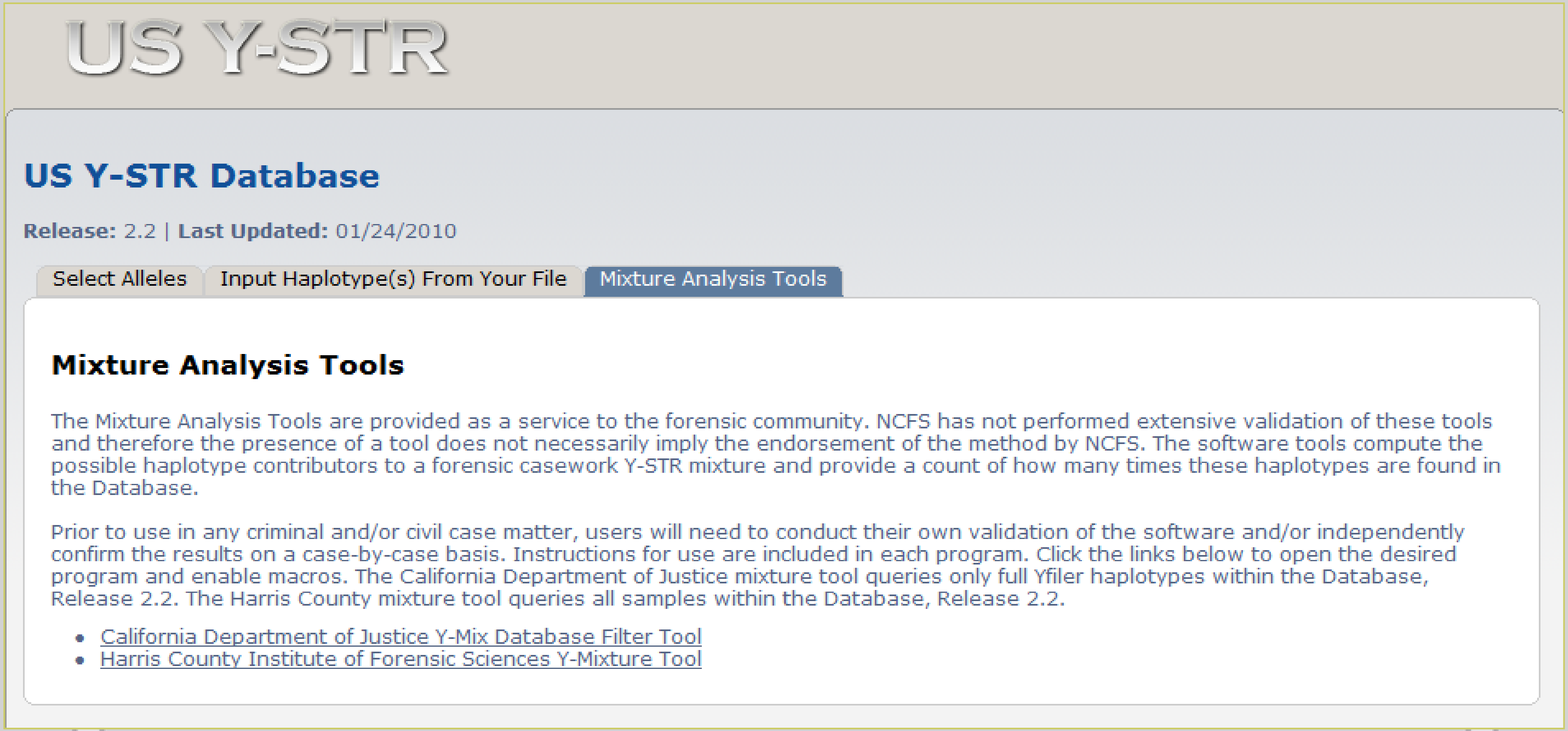
RMNE versus LR and Unresolved Questions

- RMNE = 1 in 296 or 1 in 132 (taking into account unobserved haplotypes) males are included as potential donors to the mixture
- LR = DNA results are 9400 times more likely if the suspect is admixed with the victim than if DNA from two random males' DNA is present (cf LR of 1/0.0034 = 294 for the suspects haplotype if single source)
- Population substructure correction added instead of or in addition to, binomial sampling correction?
- Denominator of LR
  - Instead of haplotype frequencies some suggest use frequency of pair wise haplotypes from database that can explain the mixture versus all other pairs of haplotypes that are possible
    - (8)/(1/2 x 5000)(4999) = 8/12497500 = 0.00000064 (without binomial correction)
    - LR = 1/0.00000064 ≈ 1,562,000

NIJ National Institute of Justice

Mixture Interpretation / Deconvolution

- In early March 2008, NCFS entered into collaboration with the Harris County Medical Examiner's Office (MEO) in Houston, Texas in an effort to create a query for Y-STR mixture interpretation / deconvolution
- NCFS obtained approval from the Y-STR Consortium members to send Y-STR data to Harris County MEO
- Their scientists created the mixture tool using the supplied data and the tool was supplied to NCFS for inclusion into the database
- NCFS also received a mixture tool from the California Department of Justice, tool was validated and added to the database
- We are currently working to add one additional tool to the database supplied by the Illinois State Police Crime Lab



US Y-STR Database

Release: 2.2 | **Last Updated:** 01/24/2010

Mixture Analysis Tools

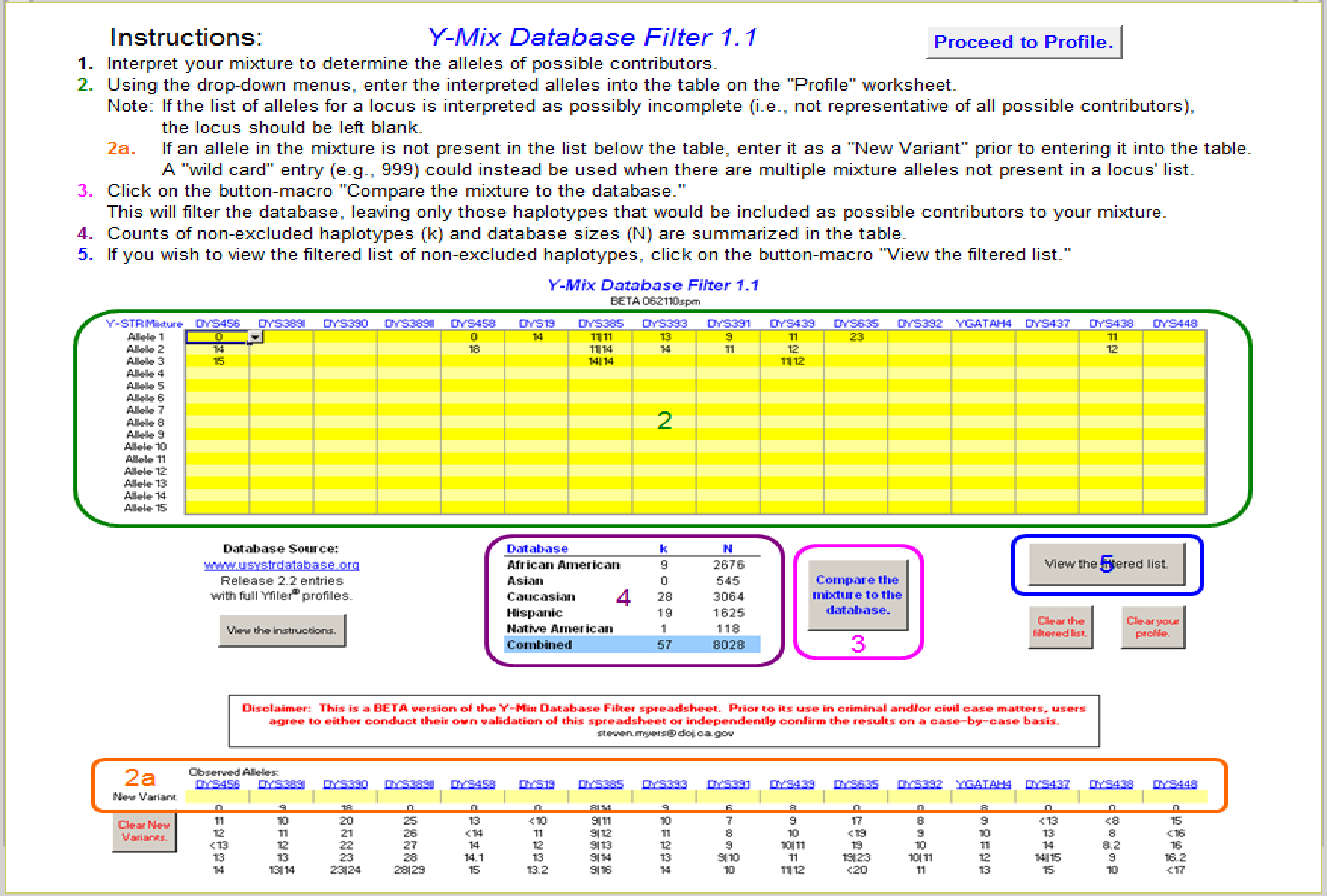
The Mixture Analysis Tools are provided as a service to the forensic community. NCFS has not performed extensive validation of these tools and therefore the presence of a tool does not necessarily imply the endorsement of the method by NCFS. The software tools compute the possible haplotype contributors to a forensic casework Y-STR mixture and provide a count of how many times these haplotypes are found in the Database.

Prior to use in any criminal and/or civil case matter, users will need to conduct their own validation of the software and/or independently confirm the results on a case-by-case basis. Instructions for use are included in each program. Click the links below to open the desired program and enable macros. The California Department of Justice mixture tool queries only full Yfiler haplotypes within the Database, Release 2.2. The Harris County mixture tool queries all samples within the Database, Release 2.2.

- California Department of Justice Y-Mix Database Filter Tool
- Harris County Institute of Forensic Sciences Y-Mixture Tool

- **Y-STR Mixture Tools were added to the database website on June 20, 2010**
- **Users can select the tool that best suits their needs and the tool opens in a new window.**

CA DOJ Y-STR Mixture Tool



Instructions: *Y-Mix Database Filter 1.1*

Proceed to Profile.

1. Interpret your mixture to determine the alleles of possible contributors.
2. Using the drop-down menus, enter the interpreted alleles into the table on the "Profile" worksheet.
   Note: If the list of alleles for a locus is interpreted as possibly incomplete (i.e., not representative of all possible contributors), the locus should be left blank.
   2a. If an allele in the mixture is not present in the list below the table, enter it as a "New Variant" prior to entering it into the table. A "wild card" entry (e.g., 999) could instead be used when there are multiple mixture alleles not present in a locus' list.
3. Click on the button-macro "Compare the mixture to the database."
   This will filter the database, leaving only those haplotypes that would be included as possible contributors to your mixture.
4. Counts of non-excluded haplotypes (k) and database sizes (N) are summarized in the table.
5. If you wish to view the filtered list of non-excluded haplotypes, click on the button-macro "View the filtered list."

Y-Mix Database Filter 1.1
BETA 062110spm

Y-STR Mixture	DYS456	DYS389I	DYS390	DYS389II	DYS458	DYS19	DYS385	DYS393	DYS391	DYS439	DYS635	DYS392	YGATAH4	DYS437	DYS438	DYS448
Allele 1	0				0	14	11|11	13	9	11	23				11	
Allele 2	14				18		11|14	14	11	12					12	
Allele 3	15						14|14			11|12						
Allele 4																
Allele 5																
Allele 6																
Allele 7																
Allele 8																
Allele 9																
Allele 10																
Allele 11																
Allele 12																
Allele 13																
Allele 14																
Allele 15																

Database Source:
www.usystrdatabase.org
Release 2.2 entries
with full Yfiler® profiles.

View the instructions.

Database	k	N
African American	9	2676
Asian	0	545
Caucasian	28	3064
Hispanic	19	1625
Native American	1	118
Combined	57	8028

Compare the mixture to the database.

View the filtered list.

Clear the filtered list.

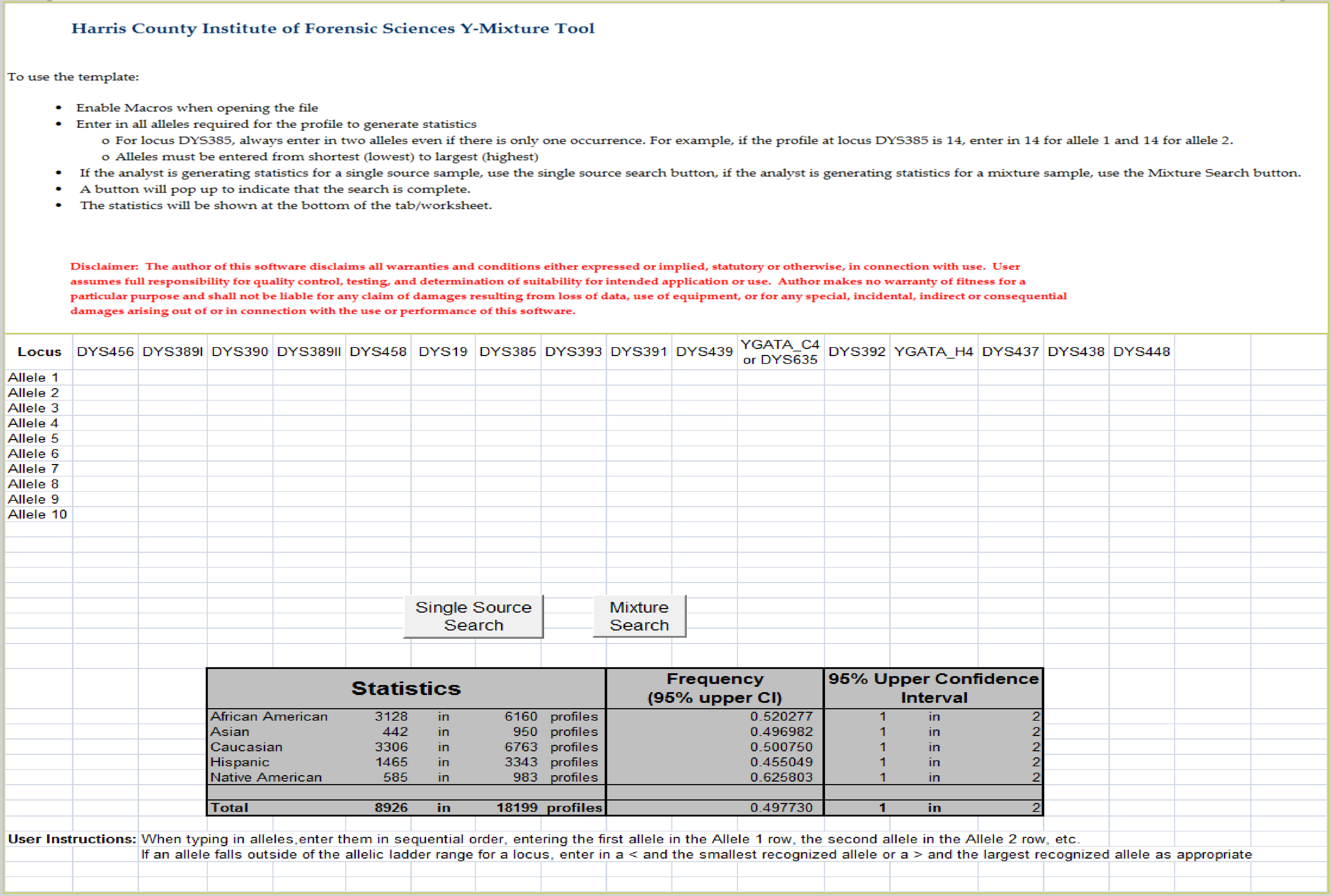
Clear your profile.

Disclaimer: This is a BETA version of the Y-Mix Database Filter spreadsheet. Prior to its use in criminal and/or civil case matters, users agree to either conduct their own validation of this spreadsheet or independently confirm the results on a case-by-case basis.
steven.myers@doj.ca.gov

Observed Alleles:

New Variant	DYS456	DYS389I	DYS390	DYS389II	DYS458	DYS19	DYS385	DYS393	DYS391	DYS439	DYS635	DYS392	YGATAH4	DYS437	DYS438	DYS448
	0	9	18	0	0	0	8|14	9	6	8	0	0	8	0	0	0
Clear New Variants.	11	10	20	25	13	<10	9|11	10	7	9	17	8	9	<13	<8	15
	12	11	21	26	<14	11	9|12	11	8	10	<19	9	10	13	8	<16
	<13	12	22	27	14	12	9|13	12	9	10|11	19	10	11	14	8.2	16
	13	13	23	28	14.1	13	9|14	13	9|10	11	19|23	10|11	12	14|15	9	16.2
	14	13|14	23|24	28|29	15	13.2	9|16	14	10	11|12	<20	11	13	15	10	<17

Harris County MEO Y-STR Mixture Tool



Harris County Institute of Forensic Sciences Y-Mixture Tool

To use the template:

- Enable Macros when opening the file
- Enter in all alleles required for the profile to generate statistics
  - For locus DYS385, always enter in two alleles even if there is only one occurrence. For example, if the profile at locus DYS385 is 14, enter in 14 for allele 1 and 14 for allele 2.
  - Alleles must be entered from shortest (lowest) to largest (highest)
- If the analyst is generating statistics for a single source sample, use the single source search button, if the analyst is generating statistics for a mixture sample, use the Mixture Search button.
- A button will pop up to indicate that the search is complete.
- The statistics will be shown at the bottom of the tab/worksheet.

Disclaimer: The author of this software disclaims all warranties and conditions either expressed or implied, statutory or otherwise, in connection with use. User assumes full responsibility for quality control, testing, and determination of suitability for intended application or use. Author makes no warranty of fitness for a particular purpose and shall not be liable for any claim of damages resulting from loss of data, use of equipment, or for any special, incidental, indirect or consequential damages arising out of or in connection with the use or performance of this software.

Locus	DYS456	DYS389I	DYS390	DYS389II	DYS458	DYS19	DYS385	DYS393	DYS391	DYS439	YGATA_C4 or DYS635	DYS392	YGATA_H4	DYS437	DYS438	DYS448
Allele 1																
Allele 2																
Allele 3																
Allele 4																
Allele 5																
Allele 6																
Allele 7																
Allele 8																
Allele 9																
Allele 10																

Statistics					**Frequency (95% upper CI)**	**95% Upper Confidence Interval**		
African American	3128	in	6160	profiles	0.520277	1	in	2
Asian	442	in	950	profiles	0.496982	1	in	2
Caucasian	3306	in	6763	profiles	0.500750	1	in	2
Hispanic	1465	in	3343	profiles	0.455049	1	in	2
Native American	585	in	983	profiles	0.625803	1	in	2
Total	**8926**	**in**	**18199**	**profiles**	0.497730	**1**	**in**	2

User Instructions: When typing in alleles,enter them in sequential order, entering the first allele in the Allele 1 row, the second allele in the Allele 2 row, etc.
If an allele falls outside of the allelic ladder range for a locus, enter in a < and the smallest recognized allele or a > and the largest recognized allele as appropriate

“The test of all knowledge is experiment. Experiment is the sole judge of scientific truth.”

Richard Feynman

Acknowledgements: Lyn Fatolitis
Mirianette Gayoso
Erin Hanson
NIJ

NIJ National Institute of Justice

Contact Information

Jack Ballantyne

University of Central Florida, Orlando, FL

jballant@mail.ucf.edu