Scientific Working Group on DNA Analysis Methods

SWGDAM

January 17, 2012 – Fredericksburg, VA

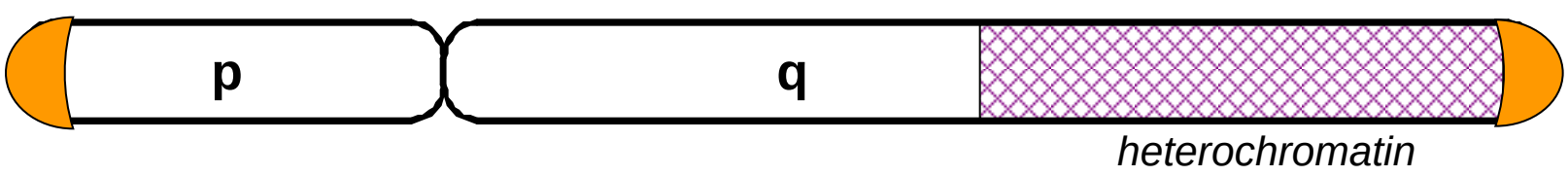
Issues with Y-STR Profile Frequency Estimation

Jack Ballantyne

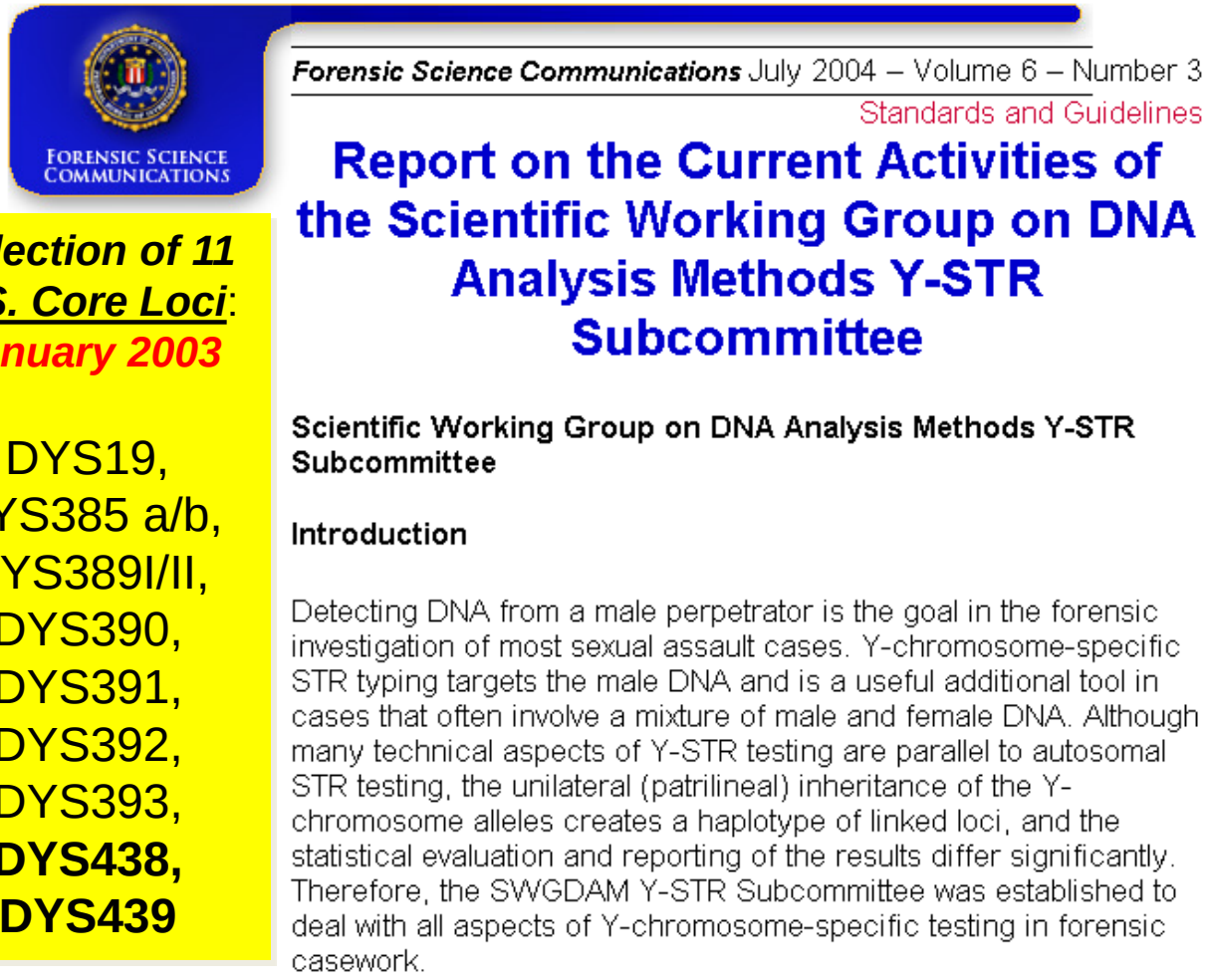
University of Central Florida

John M. Butler

NIST Applied Genetics Group



SWGDAM Y-STR Committee



Selection of 11 U.S. Core Loci*: *January 2003

DYS19,
DYS385 a/b,
DYS389I/II,
DYS390,
DYS391,
DYS392,
DYS393,
DYS438,
DYS439

These 11 loci were part of the Y-PLEX 6 and Y-PLEX 5 kits available at the time from Reliagene and encompassed the 9 loci in the European minimal haplotype (established in 1998) plus DYS438 and DYS439

Committee Members

Not all were present for all meetings

July 2002 – Jan 2008

Jack Ballantyne (UCF) – chair
Mecki Prinz (NYC) – co-chair
John Butler (NIST)
Ann Gross (MN)
Bruce Budowle (FBI)
Jill Smerick (FBI)
Sam Baechtel (FBI)
John Hartmann (Orange Co., CA)
Jonathan Newman (CFS)
Phil Kinsey (OR→MT)
Gary Sims (CA DOJ)
Demris Lee (AFDIL)
Carl Ladd (CT)
Charles Barna (MI)
Debbie Figarelli (Phoenix PD)

http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/july2004/standards/2004_03_standards03.htm

Some Background on the Previous Y-STR Committee

- SWGDAM had a functional Y-chromosome committee from **July 2002 to January 2008**
  - *Many of the committee members came from the prior validation committee and later became part of the mixture committee*
  - Not much happened from July 2005 to Jan 2008 waiting for a decision on subpopulation correction and USYSTR database
  - Mixture committee started in January 2007 and ran in conjunction with the Y-STR committee for three meetings (Jan 2007, July 2007, Jan 2008)

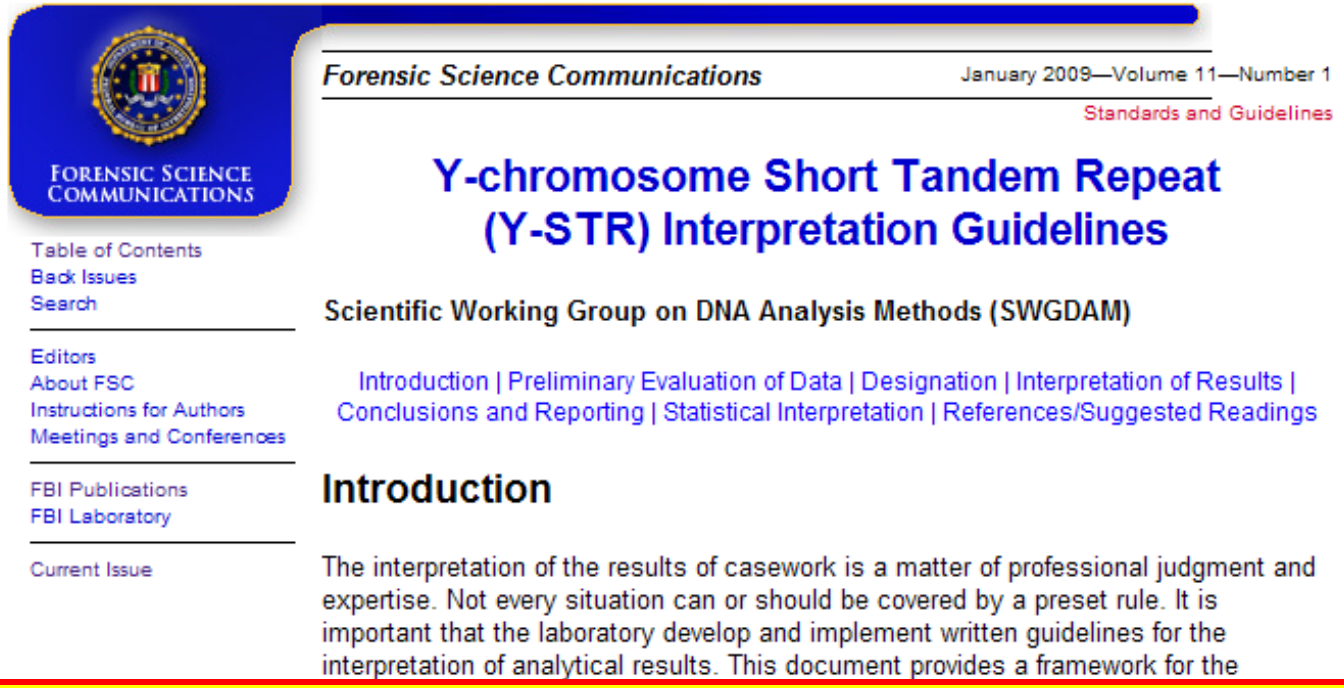
- Two primary accomplishments:
  1. Selection of **core Y-STR loci** (January 2003)
  2. SWGDAM approval (July 2008) and publication (January 2009) of **Y-STR interpretation guidelines**

What has happened in the past decade…

- **Selection of core Y-STR loci** (SWGDAM Jan 2003)
- "Full" Y-chromosome sequence became available in June 2003; over 400 Y-STR loci identified (only ~20 in 2000)
- **Commercial Y-STR kits released**
  - ~~Y-PLEX 6,5,12 (2001-03)~~, **PowerPlex Y** (9/03), **Yfiler** (12/04)
- Many population studies performed and online databases generated with thousands of Y-STR haplotypes
- Forensic casework demonstrations showing value of Y-STR testing along with court acceptance
- Some renewed interest in Y-STRs to aid familial searching

Current (2009) SWGDAM Y-STR Interpretation Guidelines

- **Approved July 15, 2008 by SWGDAM**
- Published in *Forensic Sci. Comm.* Jan 2009 issue



FORENSIC SCIENCE COMMUNICATIONS

Forensic Science Communications January 2009—Volume 11—Number 1

Standards and Guidelines

Y-chromosome Short Tandem Repeat (Y-STR) Interpretation Guidelines

Table of Contents
Back Issues
Search

Editors
About FSC
Instructions for Authors
Meetings and Conferences

FBI Publications
FBI Laboratory

Current Issue

Scientific Working Group on DNA Analysis Methods (SWGDAM)

Introduction | Preliminary Evaluation of Data | Designation | Interpretation of Results | Conclusions and Reporting | Statistical Interpretation | References/Suggested Readings

Introduction

The interpretation of the results of casework is a matter of professional judgment and expertise. Not every situation can or should be covered by a preset rule. It is important that the laboratory develop and implement written guidelines for the interpretation of analytical results. This document provides a framework for the

Modeled largely on the 2000 SWGAM Interpretation Guidelines with Section 5 discussing statistical interpretation

http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/jan2009/standards/2009_01_standards01.htm/

Presentation Outline

- **Elements of Haplotype Frequency Estimates**
  - Differences between Y-STRs and mtDNA
  - Y-STR loci and kits available
  - Databases: YHRD and USYSTR
  - Approaches to profile frequency estimation
    - Counting method
    - 95% confidence interval (normal & Clopper-Pearson)
    - [Bayesian predictor used in YHRD]
    - [Brenner rare haplotype estimation]

- **Current SWGDAM Guidelines**
  - Section 5 point-by-point
    - Produced language to adjust to Clopper-Pearson
  - Points for discussion
    - Population substructure correction
    - Mixtures
    - Combining Y-STR data with autosomal STR information

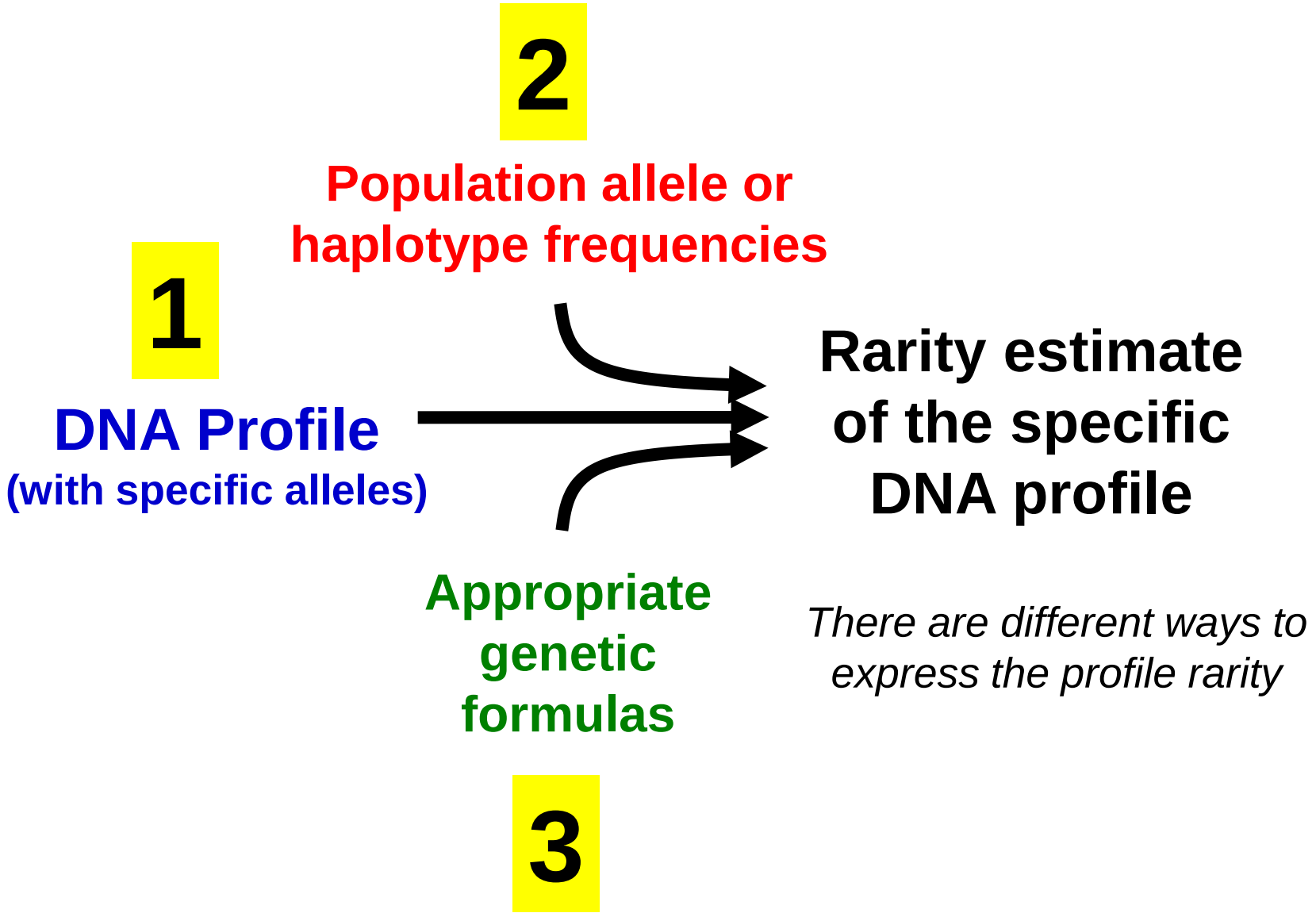
Y-STRs vs. mtDNA

- **Y-STRs are easier to analyze**
  - STR typing at 12 or 17 loci in a single multiplex PCR compared to sequence analysis across at least 610 nucleotides (and multiple strands) and often multiple amplifications with difficult samples
  - Fewer labs are doing mtDNA analysis

- **Y-STRs have larger population databases**
  - Samples are easier to analyze; more labs are doing Y-STR analysis
  - YHRD ~100,000 samples; EMPOP ~16,000 samples

- **Y-STRs offer finer resolution**
  - Effectively more “alleles” (haplotypes)
  - Leads to better separation of unrelated samples (and possibly related ones) due to a higher mutation rate with Y-STR loci

Elements Going into the Calculation of a Rarity Estimate for a DNA Sample



There are different ways to express the profile rarity

Example with an Autosomal STR Profile

Table 11.3 Random match probability for a 13-locus STR profile using the U.S. Caucasian allele frequencies found in Table 11.1.

	Allele 1	Allele 2	Allele 1 Frequency (*p*)	Allele 2 Frequency (*q*)	Formula	Expected Genotype Frequency
D13S317	11	14	0.33940	0.04801	2pq	0.0326
TH01	6	6	0.23179		p^2	0.0537
D18S51	14	16	0.13742	0.13907	2pq	0.0382
D21S11	28	30	0.15894	0.27815	2pq	0.0884
D3S1358	16	17	0.25331	0.21523	2pq	0.1090
D5S818	12	13	0.38411	0.14073	2pq	0.1081
D7S820	9	9	0.17715		p^2	0.0314
D8S1179	12	14	0.18543	0.16556	2pq	0.0614
CSF1PO	10	10	0.21689		p^2	0.0470
FGA	21	22	0.18543	0.21854	2pq	0.0810
D16S539	9	11	0.11258	0.32119	2pq	0.0723
TPOX	8	8	0.53477		p^2	0.2860
VWA	17	18	0.28146	0.20033	2pq	0.1128
AMEL	X	Y				
Product rule						1.20×10^{-15}
Combined frequency						1 in 8.37×10^{14}

Combined STR Profile Frequency (unrelated, Caucasian) = 1 in 837 trillion

Impact of Various Assumptions

With different population group

African Amer.
1 in 17 quadrillion
…

With subpopulation correction (NRC II 4.10)

θ = 0.03
1 in 33 trillion
…

With relatives as a possibility

Full sibling
1 in 248,000
…

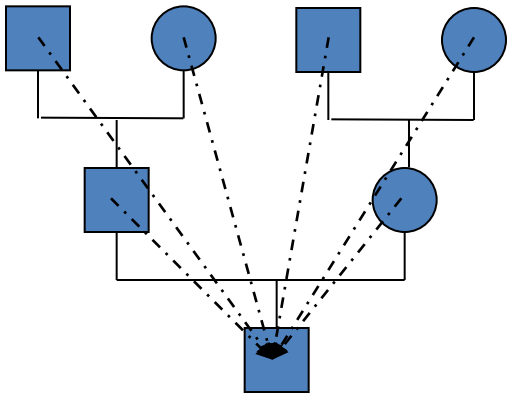
3

John M. Butler (2010) *Fundamentals of Forensic DNA Typing*, Table 11.3

Different Inheritance Patterns

Autosomal Markers

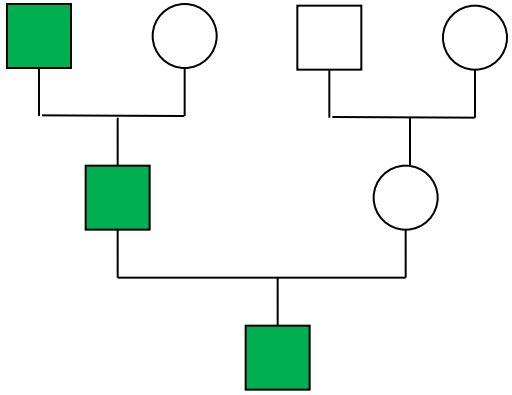
13 CODIS STR Loci



22 pairs of autosomes
(passed on in part,
from all ancestors)

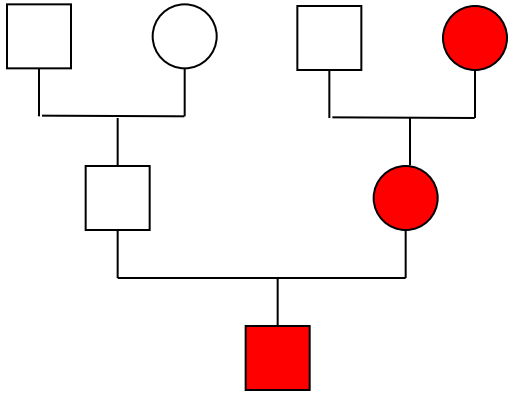
Lineage Markers

12 or 17 Y-STRs



Y-Chromosome
(passed on complete,
but only by sons)

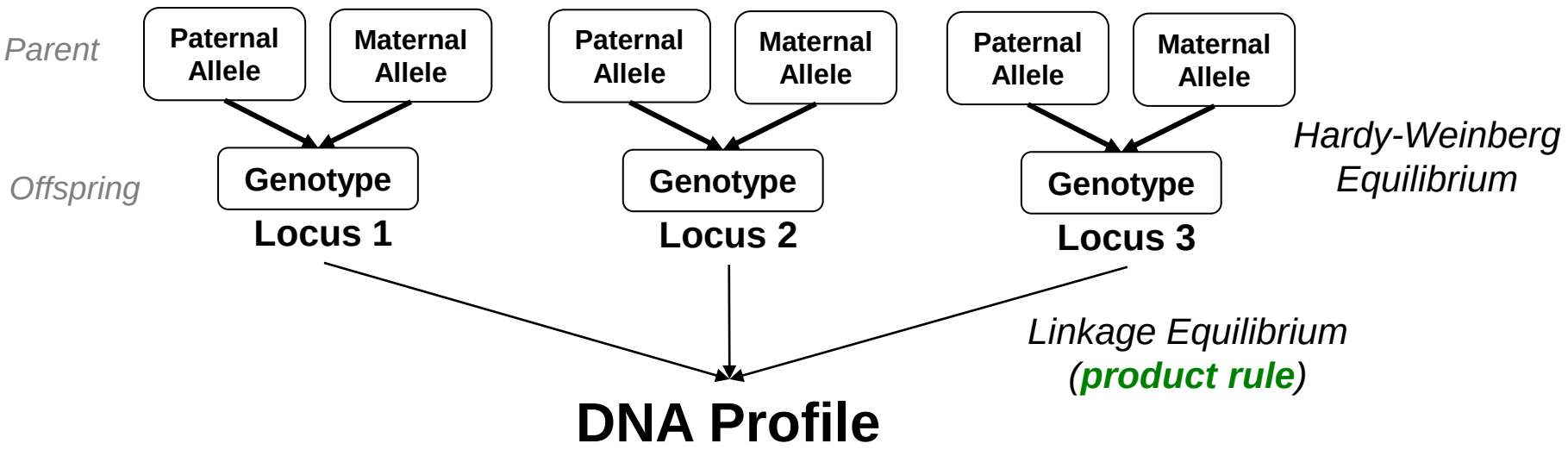
mtDNA control region



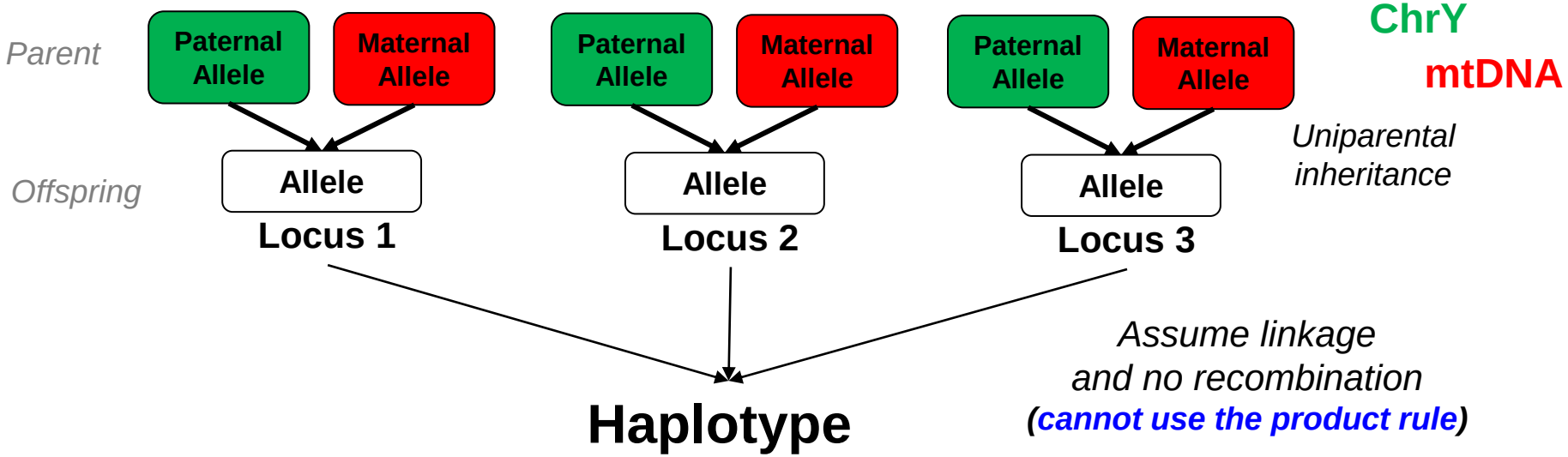
Mitochondrial
(passed on complete,
but only by daughters)

Differences between Autosomal and Lineage Markers

Autosomal Markers



Lineage Markers

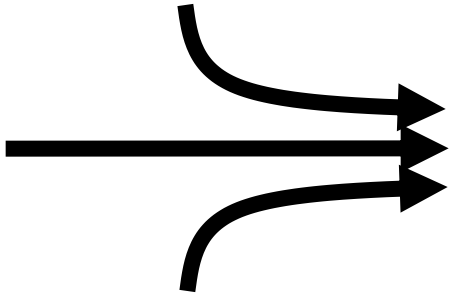


Generating a Y-STR Profile

Population allele or haplotype frequencies

1

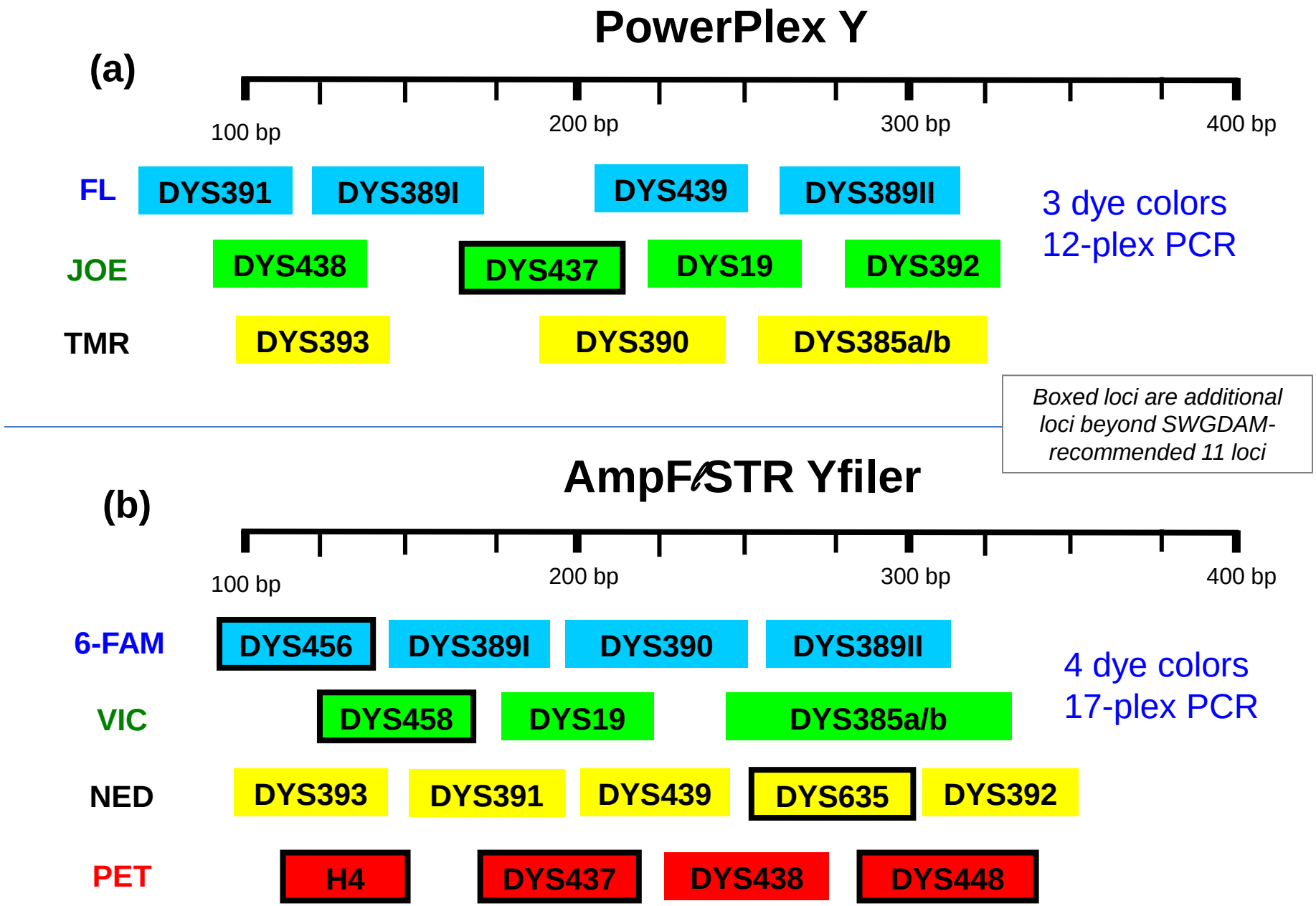
DNA Profile
(with specific alleles)



Rarity estimate of the specific DNA profile

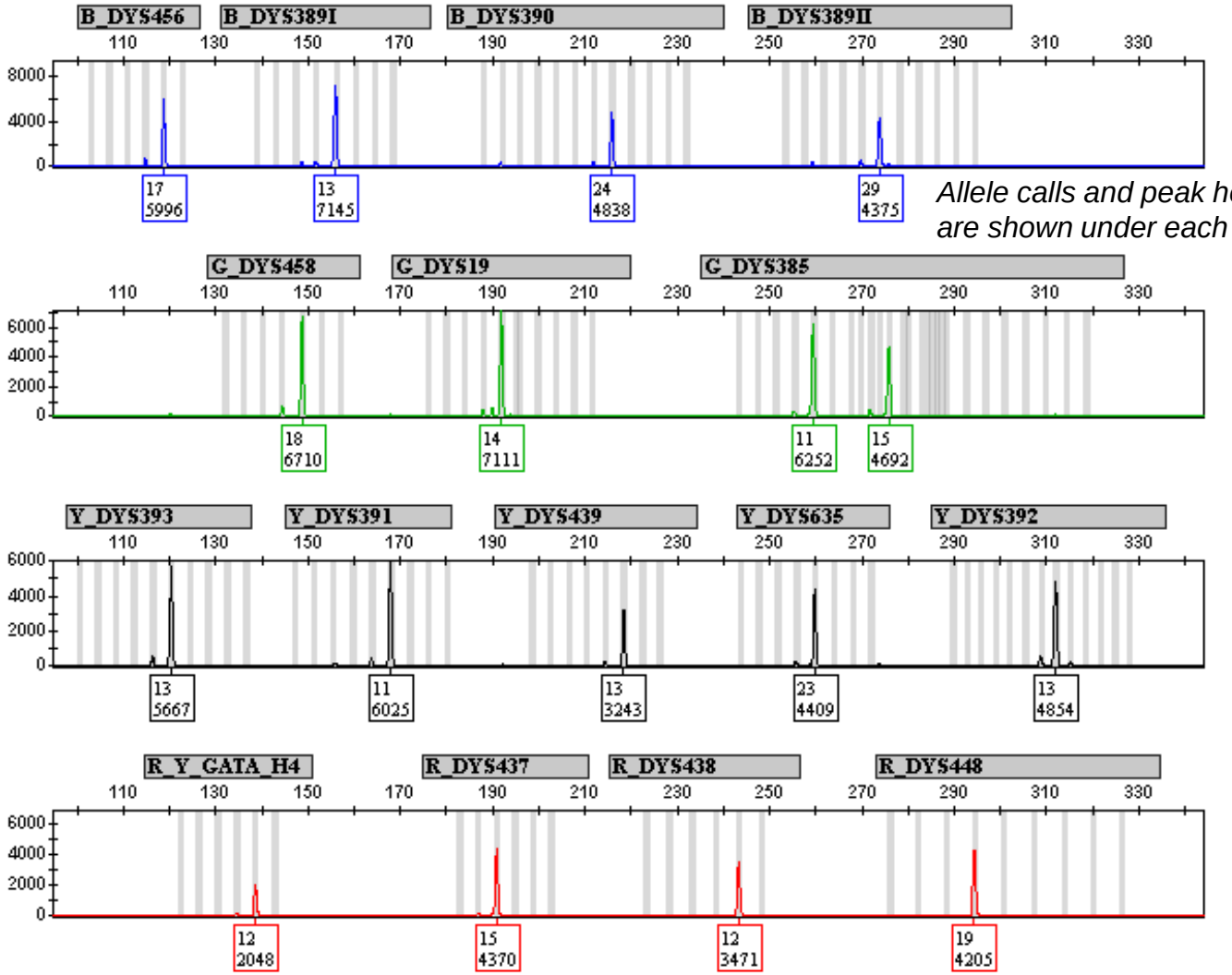
Appropriate genetic formulas

Current Commercial Y-STR Kits (Loci, Dye Colors, Size Ranges)



Yfiler Result (17 Y-STRs)

from a Single-Source Male of European Ancestry



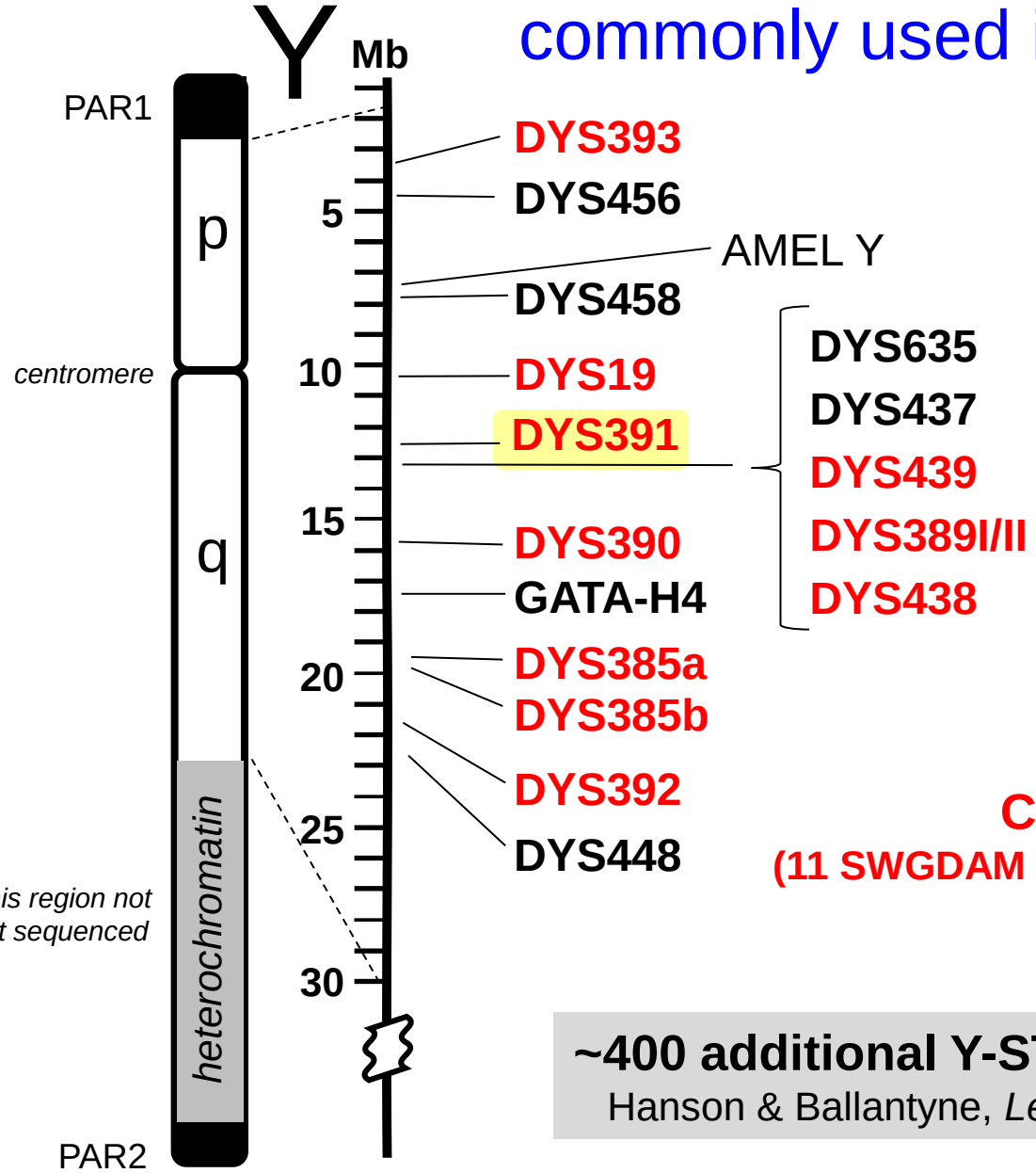
Allele calls and peak heights are shown under each peak

Haplotype

DYS19	389I	389II	390	391	392	393	385 a/b	438	439	437	448	456	458	635	H4
14	13	29	24	11	13	13	11-15	12	13	15	19	17	18	23	12

J.M. Butler (2010) *Fundamentals of Forensic DNA Typing*, Figure 16.4

Relative positions of 17 Y-STR loci
commonly used in ChrY testing



DYS391 has been proposed for inclusion as a future CODIS core locus
(D.R. Hares, *FSI Genetics*, 2012, 6, e52-e54)

Core U.S. loci
(11 SWGDAM recommended in Jan 2003)

~400 additional Y-STRs currently known
Hanson & Ballantyne, *Legal Med* 2006;8(2):110-20

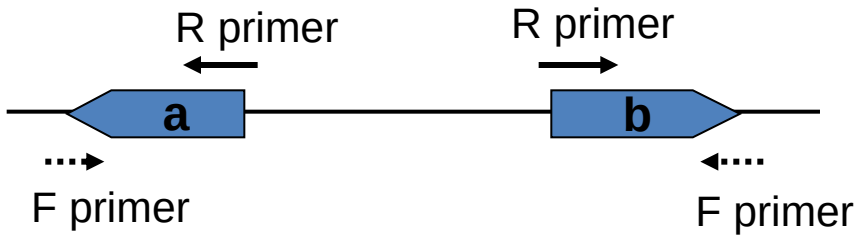
PAR = pseudo-autosomal region (recombines with X-chromosome)

J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, Figure 13.6

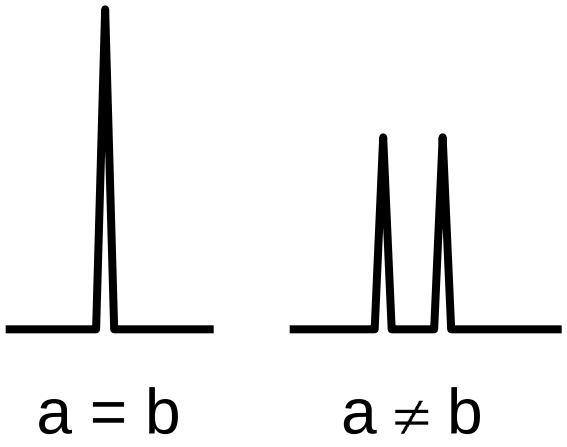
Single Primer Sets Produce Multiple PCR Products

(a) DYS385 a/b

Multi-Copy (Duplicated) Marker

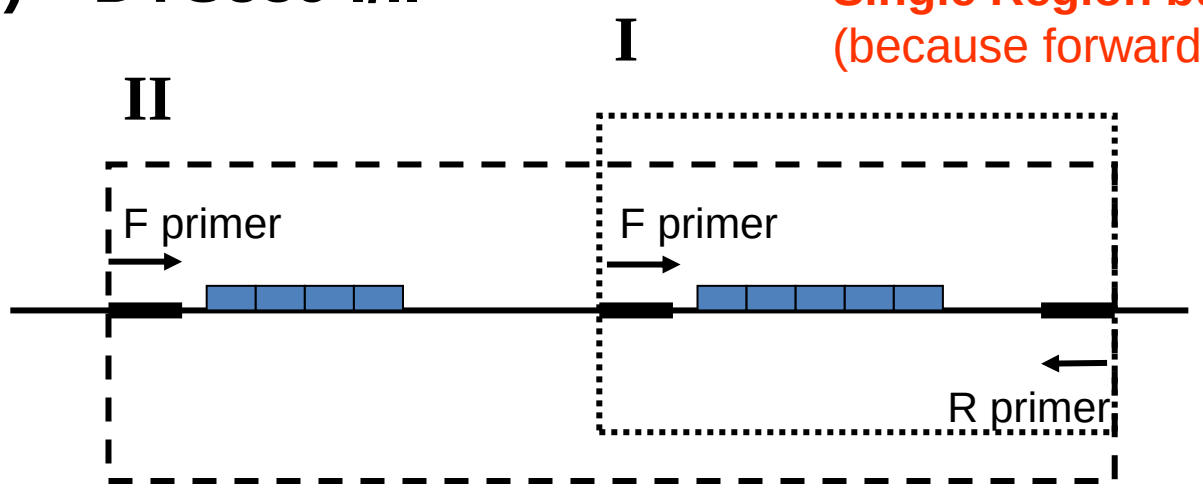


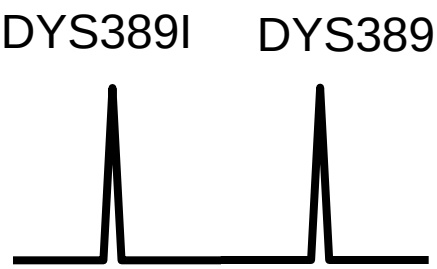
Duplicated regions are 40,775 bp apart and facing away from each other



(b) DYS389 I/II

Single Region but Two PCR Products
(because forward primers bind twice)





17 PCR products
15 primer sets

Characteristics of the 17 Commonly Used Y-STR Loci

STR Marker	Position (Mb)	Repeat Motif	Allele Range	Mutation Rate*
DYS393	3.19	AGAT	8-17	0.10 %
DYS456	4.33	AGAT	13-18	0.42 %
DYS458	7.93	GAAA	14-20	0.64 %
DYS19	10.13	TAGA	10-19	0.23 %
DYS391	12.61	TCTA	6-14	0.26 %
DYS635	12.89	TSTA	17-27	0.35 %
DYS437	12.98	TCTR	13-17	0.12 %
DYS439	13.03	AGAT	8-15	0.52 %
DYS389 I/II	13.12	TCTR	9-17 / 24-34	0.25 % / 0.36 %
DYS438	13.38	TTTTC	6-14	0.03 %
DYS390	15.78	TCTR	17-28	0.21 %
GATA-H4	17.25	TAGA	8-13	0.24 %
DYS385 a/b	19.26	GAAA	7-28	0.21 %
DYS392	21.04	TAT	6-20	0.04 %
DYS448	22.78	AGAGAT	17-24	0.16 %

*Mutation rates are from as many as 15000 meioses described in a YHRD summary of 23 publications in Jan 2011 (see (http://www.yhrd.org/Research/Loci/)

J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, Table 13.2

Recent Developments with Y-STR Typing

- **Promega Corporation** announced at their Oct 2011 ISHI meeting that they were working on a **Y-STR 23plex** which will enable further resolution of Y-STR haplotypes
  - Hopefully a kit will be released in 2012 but population databases will need to be developed with the new extended haplotypes

- Manfred Kayser's group has developed a set of **rapidly mutating (RM) Y-STR loci** that have the capability to resolve fathers and sons in many instances
  - An international collaboration is currently on-going to study these RM Y-STRs in more detail (14 RM Y-STRs in 3 multiplexes)

The Meaning of a Y-Chromosome Match…

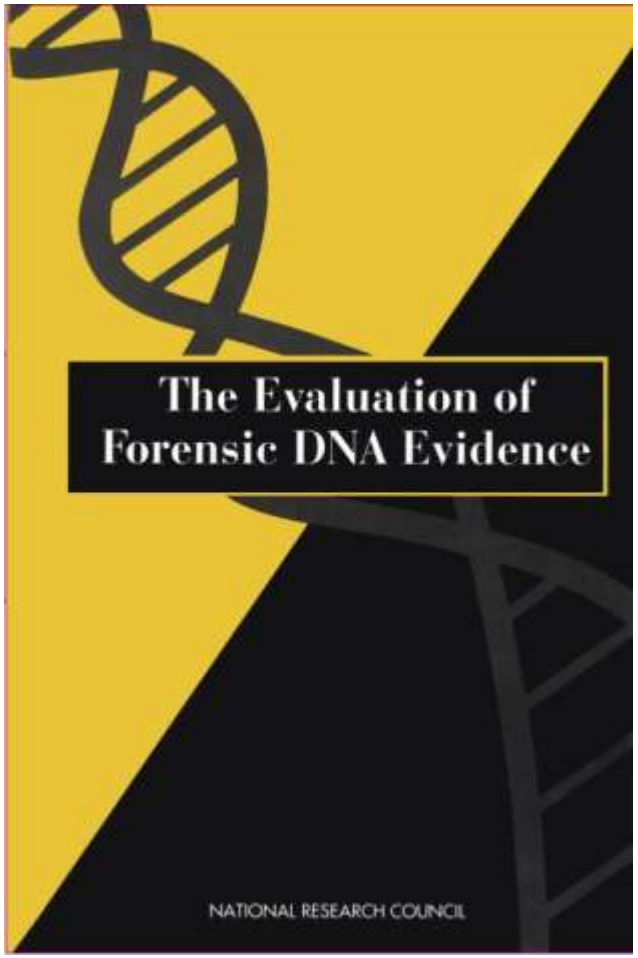
Conservative statement for a match report:

The Y-STR profile of the crime sample matches the Y-STR profile of the suspect (at *xxx* number of loci examined). Therefore, ***we cannot exclude the suspect*** as being the donor of the crime sample. In addition, we cannot exclude all patrilineal related male relatives and an unknown number of unrelated males as being the donor of the crime sample.

de Knijff, P. (2003). Son, give up your gun: presenting Y-STR results in court. *Profiles in DNA, 6(2),* 3-5. Available at http://www.promega.com/resources/articles/profiles-in-dna/2003/son-give-up-your-gun-presenting-ystr-results-in-court/

From Peter de Knijff's Oct 2004 presentation "Presenting Y-chromosome DNA evidence in court" at the International Symposium on Human Identification: http://www.promega.com/~/media/files/resources/conference%20proceedings/ishi%2015/oral%20presentations/deknijff.ashx?la=en

Inclusions (Matches) Require Statistics



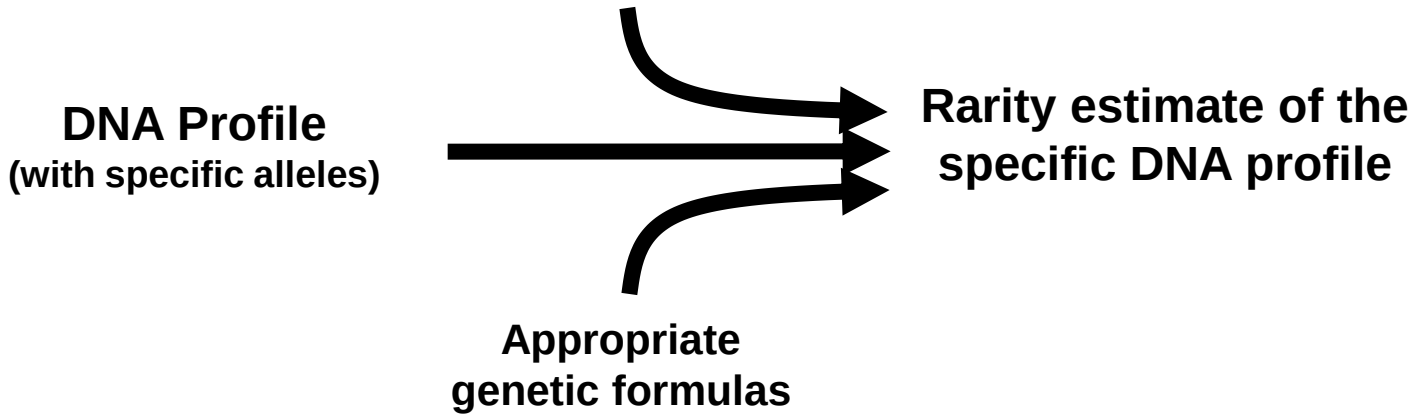
- It would not be scientifically justifiable to speak of a match as proof of identity in the absence of underlying data that permit **some reasonable estimate of how rare the matching characteristics actually are**.

-- NRC II, p. 192

Y-STR Population Haplotype Frequencies

2

Population allele or haplotype frequencies



On-line Y-STR Population Databases

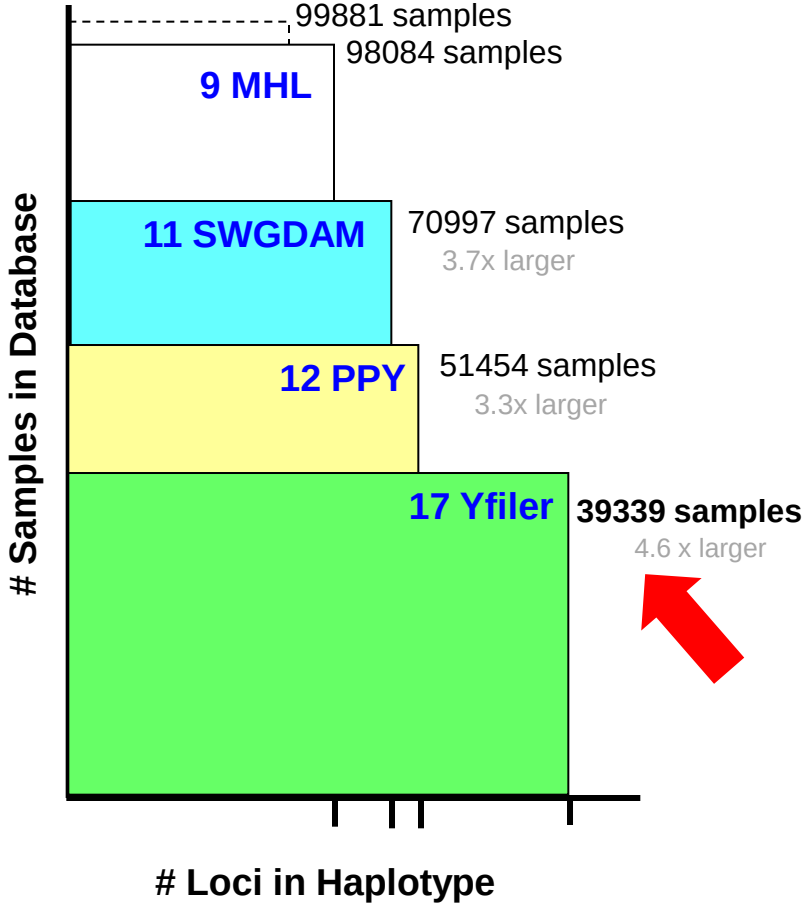
YHRD

Launched Feb 2000

http://www.yhrd.org

Release 38 Dec 30, 2011

750 Populations (109 countries)



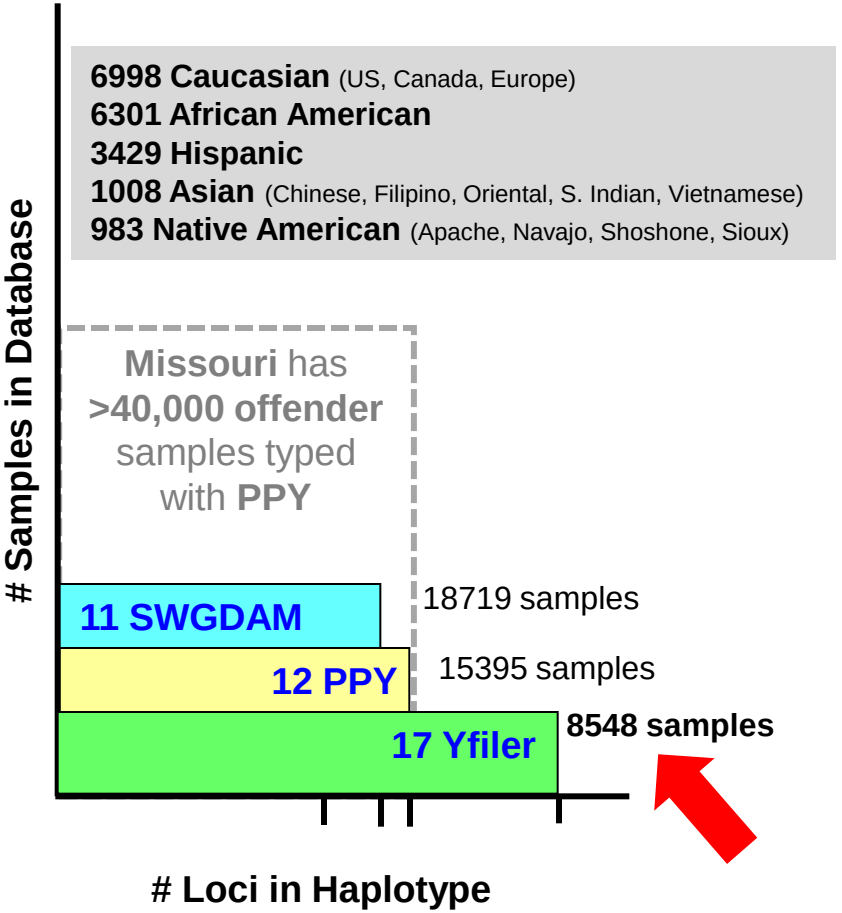
US YSTR

Launched Dec 2007

http://www.usystrdatabase.org

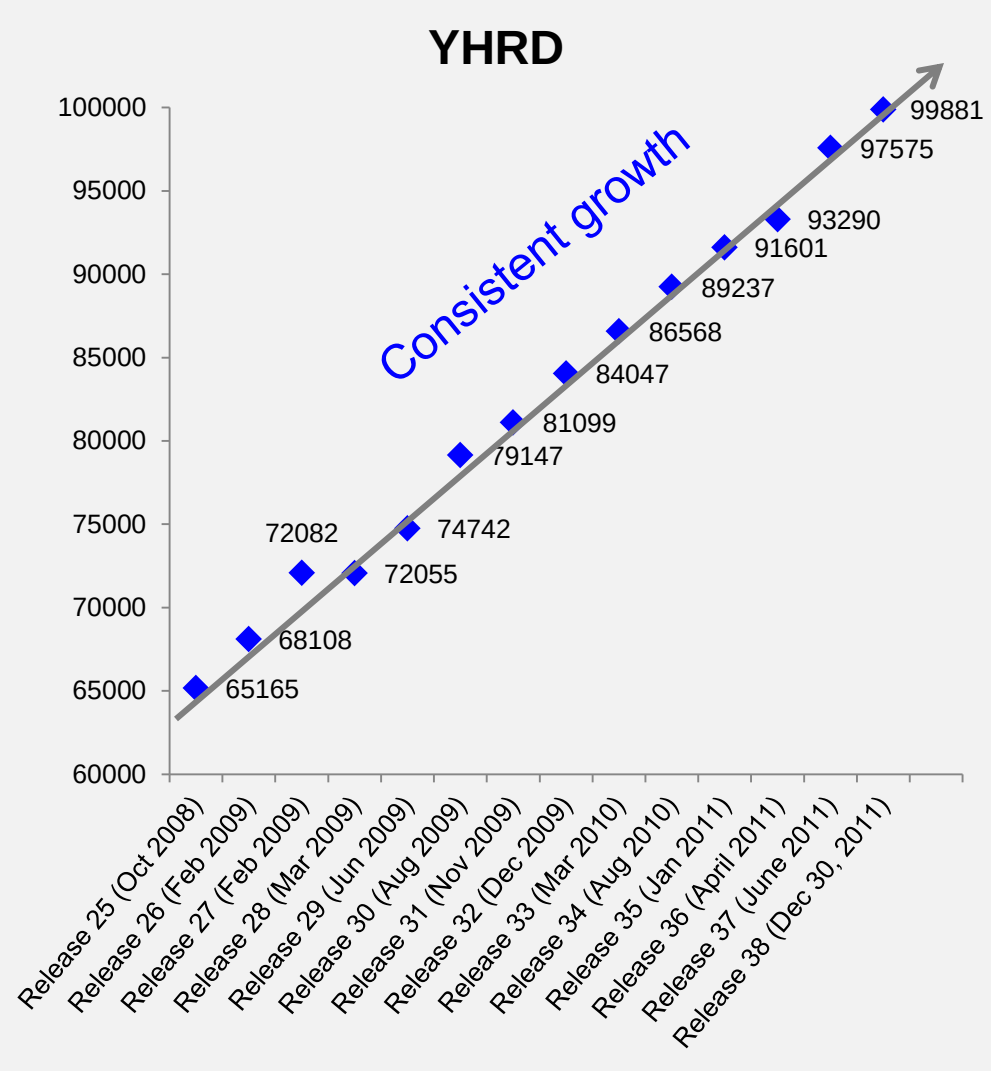
Release 2.6 Jan 3, 2012

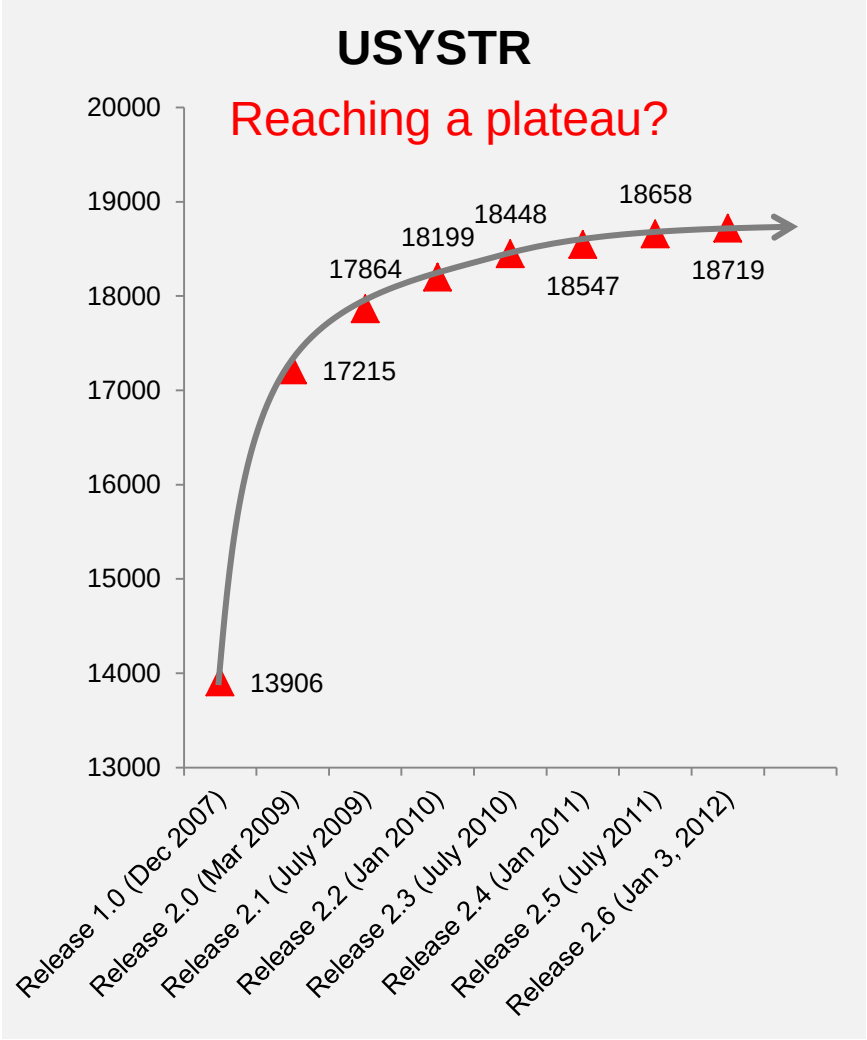
Focus is on U.S. samples



Updated from J.M. Butler (2012) *Advanced Topics in Forensic DNA Typing: Methodology*, Figure 13.10

Y-STR Haplotype Database Growth

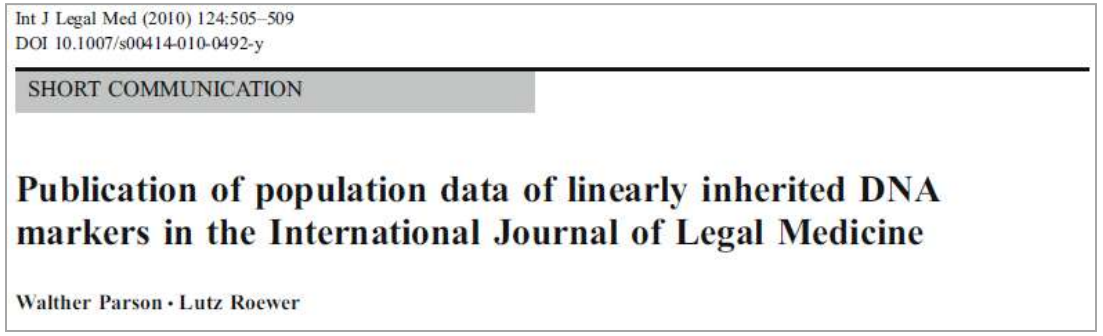




Detailed YHRD data not available on their website below Release 25

Population Data Publications Describing Handling of Y-STR and mtDNA Haplotype Information

Carracedo, A., Butler, J.M., Gusmao, L., Parson, W., Roewer, L., Schneider, P.M. (2010) Editorial: Publication of population data for forensic purposes. ***Forensic Sci. Int. Genet.*** 4: 145-147



Parson, W., Roewer, L. (2010) Publication of population data of linearly inherited DNA markers in the International Journal of Legal Medicine. ***Int. J. Legal Med.*** 124: 505-509

- The leading forensic journals *require* Y-STR and mtDNA population data to be reviewed by and submitted to YHRD and EMPOP

US YSTR Contributions

Contributor to US YSTR	# Samples	% of Database
Applied Biosystems (includes UNTHSC, NIST samples, …)	6,159	33%
Promega	3,800	20%
ReliaGene	3,037	16%
University of Arizona	2,462	13%
NCFS (University of Central Florida)	2,440	13%
Illinois State Police	398	2.1%
Santa Clara Co. CA Crime Lab	143	0.6%
Marshall University	113	0.6%
Washington State Patrol Crime Lab	40	0.2%
San Diego Sheriff's Regional Crime Lab	39	0.2%
CA DOJ	32	0.2%
Orange County CA Coroner	30	0.2%
Richland County Sheriff's Dept.	7	0.04%
Release 2.6 (Jan 3, 2012)	**18,719**	**8548** 17-locus profiles

US YSTR Database Search Results (with 17 loci)

Release: 2.6 | **Last Updated:** 01/03/2012

Select Alleles | Input Haplotype(s) From Your File | Mixture Analysis Tools

Common Markers

DYS19	DYS385	DYS389I	DYS389II
14	11,15	13	29
DYS390	**DYS391**	**DYS392**	**DYS393**
24	11	13	13
DYS437	**DYS438**	**DYS439**	**DYS448**
15	12	13	19
DYS456	**DYS458**	**DYS635 (YGATAC4)**	**YGATAH4**
17	18	23	12

Search By Ancestry

All
African American
Asian
Caucasian

0 matches in 8548 Yfiler profiles

Search Reset

Results: Show Details Hide Details

Ancestry	# of Haplotypes	Number of Haplotypes (with Selected Alleles)	Frequency	Frequency Upper Bound (95%)
African American	2817	0	0.000000	0.001062
Asian	603	0	0.000000	0.004955
Caucasian	3299	0	0.000000	0.000907
Hispanic	1711	0	0.000000	0.001749
Native American	118	0	0.000000	0.025068
Total	8548	0	0	0.000350

Overall Database Summary:

The selected haplotype is found in **0** of **8548** total individuals within the database with a frequency of **0**. Applying the 95% upper confidence interval results in a frequency of **0.000350**, which is equivalent to approximately 1 in every **2857** individuals.

When there is no match in the haplotype database…

Current SWGDAM mtDNA (2003) and Y-STR (2009) Guidelines

In cases where the profile has not been observed in a database, the upper bound on the confidence interval is

$$1-\alpha^{1/N}$$

where α is the confidence coefficient (0.05 for a 95% confidence interval) and N is the number of individuals in the database.

USYSTR: 0 matches in 8548 Yfiler profiles

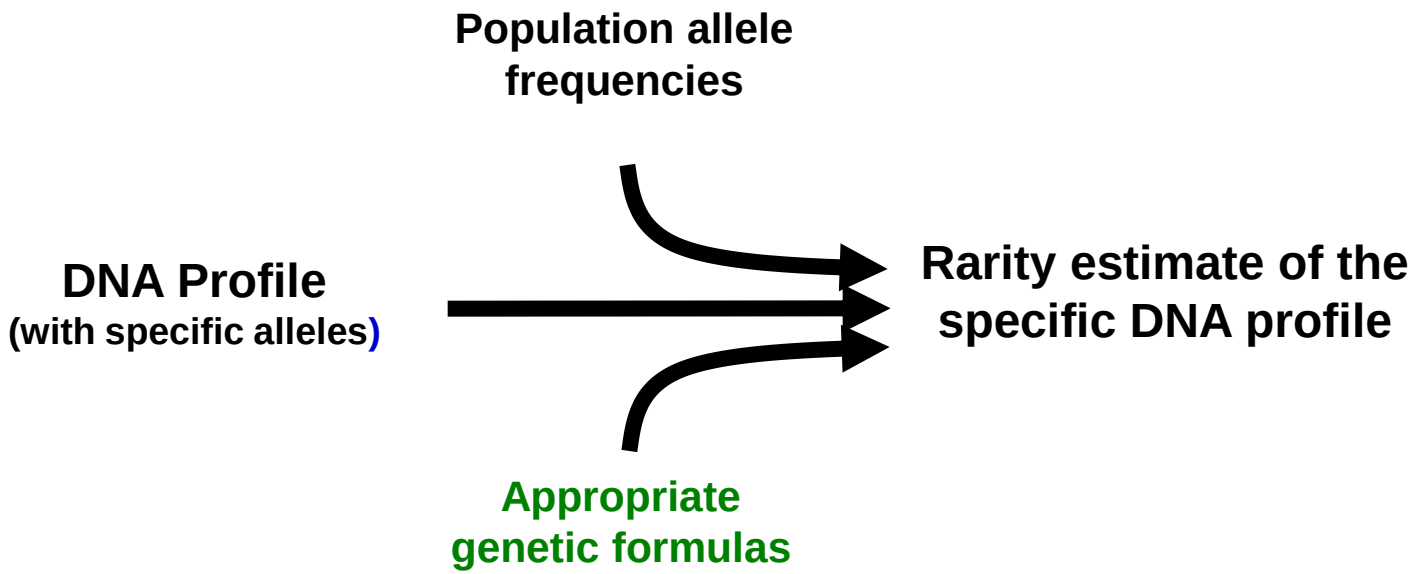
$1-\alpha^{1/N} = 1-(0.05)^{[1/8548]} = 0.000350$

= **0.035% (1 in 2857)**

A simplified calculation is 3/N.

In this example: 3/8548 = 0.000351 = 0.035% (1 in 2849)

Applying Genetic Models and Formulas



3

New Lineage Marker Interpretation Information

Forensic Science International: Genetics 5 (2011) 78–83

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

The interpretation of lineage markers in forensic DNA testing

J.S. Buckleton [a], M. Krawczak [b], B.S. Weir [c,*]

[a] ESR Ltd, Private Bag 92021, Auckland, New Zealand
[b] Institute of Medical Informatics and Statistics, Christian-Albrechts University, 24105 Kiel, Germany
[c] Department of Biostatistics, University of Washington, Box 357232, Seattle, WA 98195-7232, USA

This article reviews and discusses a number of highly relevant topics:

- **Normal vs. binomial (Clopper-Pearson) sampling distributions**
- **Theta corrections**
- **Handling rare haplotypes (Charles Brenner approach)**
- **Combination of lineage and autosomal markers**

Different Approaches/Models for Presenting Haplotype Frequency Estimates

1) **Direct Count** (frequency in population database)

$$p = \frac{x}{n}$$

2) **Confidence Interval for Sampling Correction**
   - Holland & Parsons (1999) *Forensic Sci Rev*

$$p \pm 1.96\sqrt{\frac{(p)(1-p)}{n}}$$

3) David Balding “pseudo-count” Estimate
   - Balding (2005) *Weight-of-evidence for Forensic DNA Profiles*, p. 99

$$\hat{p} = \frac{x+2}{n+2}$$

4) Theta Adjustment for Subpopulation Correction
   - Buckleton et al. (2005) *Forensic DNA Evidence Interpretation*, Chapter 9
   - Balding (2005) *Weight-of-evidence for Forensic DNA Profiles*, p. 100
   - Budowle et al. (2007) *Proc. ISHI*, (2009) *Legal Med*, (2009) *JFS*
   - Buckleton et al. (2011) *FSI Genetics*; Cockerton et al. (2012) *FSI Genetics*

$$f = \theta + (1-\theta)\,p$$

Note if $\theta = 0$, then $f = p$
If $p < \theta$, then θ bounds f

5) YHRD approach
   - Roewer et al. (2000) *Forensic Sci Int* 114: 31-43
   - Willuweit et al. (2011) *FSI Genetics* 5: 84-90

6) Brenner model for rare haplotypes
   - Brenner (2010) *FSI Genetics* 4: 281-291

$$LR = n/(1-\kappa)$$

Different Confidence Intervals

Normal approximation

Sometimes called H-P method for Holland/Parsons who introduced it to mtDNA in a 1999 review article

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

Current SWGDAM mtDNA (2003) and Y-STR (2009) Guidelines

Y-STR only includes the (+) portion of the equation

$$p \pm 1.96\sqrt{\frac{(p)(1-p)}{n}}$$

95% confidence interval with 2-tail

Wilson

$$\frac{\hat{p} + \frac{1}{2n}z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{1}{n}z^2_{1-\alpha/2}}$$

Used with EMPOP

Clopper-Pearson (exact method)

$$\left\{\, \theta \;\middle|\; P\left[\mathrm{Bin}\left(n;\theta\right) \le X\right] \ge \alpha/2 \right\} \bigcap \left\{\, \theta \;\middle|\; P\left[\mathrm{Bin}\left(n;\theta\right) \ge X\right] \ge \alpha/2 \right\}$$

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

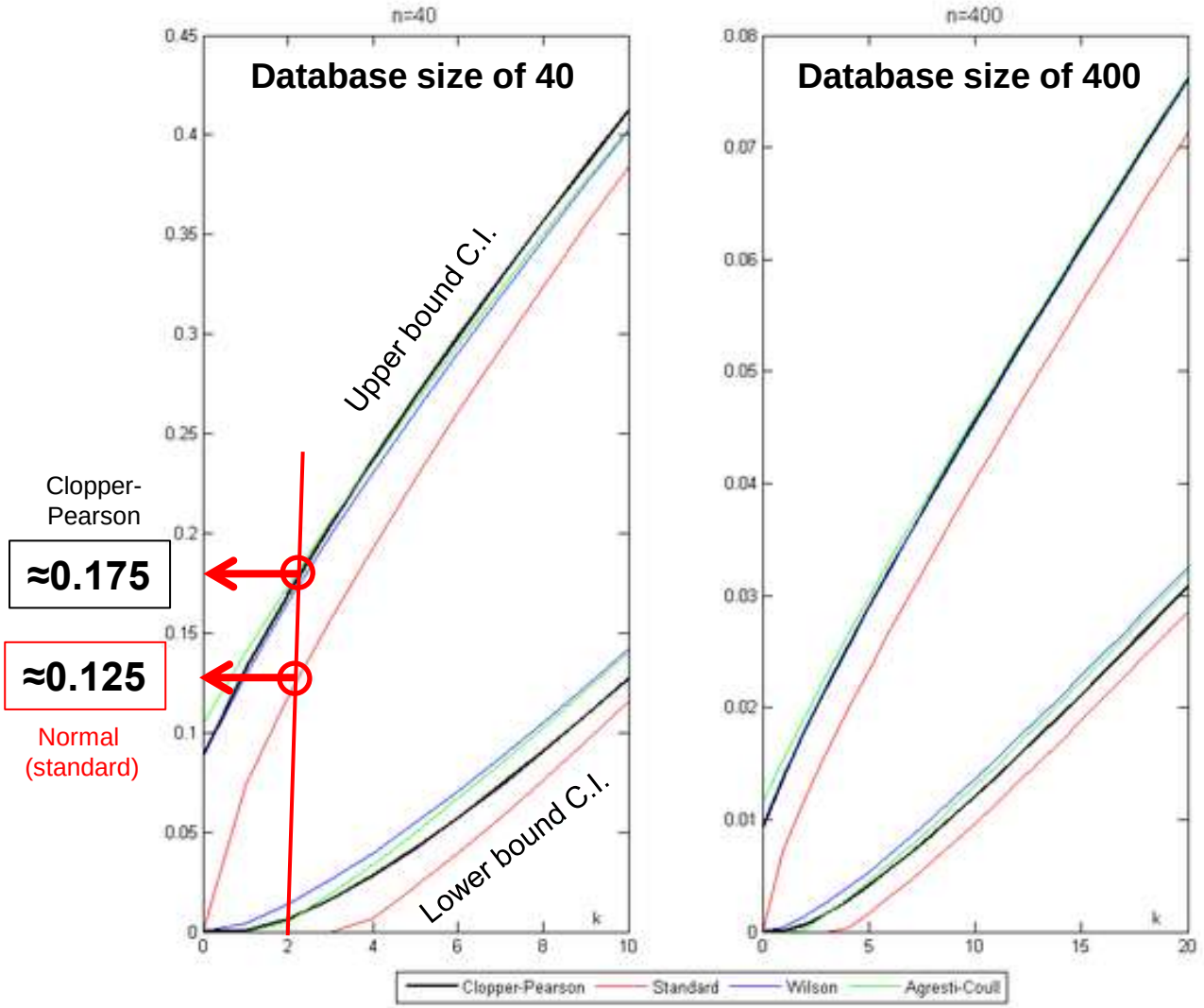
Now used with USYSTR (1-tail)

Recommended in recent review article by Buckleton et al. (2011)

At Jan 2011 SWGDAM meeting, new language was written to incorporate a Clopper-Pearson approach

Comparison of Clopper-Pearson to Normal (Standard) Confidence Intervals (C.I.)

Higher value is more conservative (favors the defendant)



Note that the Wilson upper bound interval is very close to the Clopper-Pearson

With two matches in a database of 40

Exact vs. Normal Confidence Intervals

n	x	P (=x/n)	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

HP: Holland, M.M., & Parsons, T.J. (1999). Mitochondrial DNA sequence analysis - validation and use for forensic casework. *Forensic Science Review, 11,* 21-50.

CP: Clopper, C.J., & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26,* 404-413.

US Y-STR Database versus YHRD

US: Advantages

- Relevance: US population data mainly
- Direct community involvement
- Customer service (e.g., ad hoc searches)
- SWGDAM responsiveness
- Accepted in US Courts (2 Frye hearings)

US: Disadvantages

- Cost of maintenance
- Smaller database size
- Limited number of ancestral populations
- Difficulty in obtaining samples/data from US community
  - Low rate of growth

US Y-STR Database versus YHRD

Y-HRD: advantages

- No cost
- Larger database (world wide)-Too Big to Fail!
- More ancestral populations
- Population genetic parameters well characterized
- Greater rate of growth
- Curated from afar (Europe)

Y-HRD: disadvantages

- Limited customer service
- Not US specific
  - Not yet accepted in US Courts?
- Lack of SWGDAM responsiveness
  - Have their ‘own way of doing things’
- Greater rate of growth
- Curated from afar (Europe)

Standardization is Critical for Success and Data Sharing

Needs	How/When Accomplished
Core Y-STR loci	SWGDAM Y-STR Committee selected 11-loci in January 2003
Consistent allele nomenclature	NIST SRM 2395 (2003); kit allelic ladders; ISFG (2006) and NIST (2008) publications
Commercially available Y-STR kits	Early ReliaGene kits (2001-2003); **PowerPlex Y** (2003) and **Yfiler** (2004)
Accessible, searchable population databases for haplotype frequency estimations	**YHRD** (70,997 11-locus haplotypes from 750 worldwide populations) **US YSTR** (18,719 11-locus haplotypes from primarily U.S. population groups)
Interpretation guidelines	SWGDAM Y-STR Interpretation Guidelines published in January 2009 (*will likely be revised soon*)

Predictions for **the Future** of Y-STR Analysis

- Continued use with casework (with excess female DNA)

- Improved frequency estimates with growing Y-STR databases
  - YHRD now at **70,997 11-locus profiles** (39,339 Yfiler)
  - USYSTR has **18,719 11-locus profiles** (8,548 Yfiler)

- Use with familial searching to eliminate false positives
  - Myers, S.P. et al. (2011) *FSI Genetics* 5(5): 493-500 – describes CA DOJ familial searching

- **New Y-STR kits with additional loci**
  - At the ISHI meeting, Promega announced a Y-STR 23plex was being developed
  - Will take time though to grow large population databases that cover all of the new loci

- Use of fast mutating loci to help resolve paternal lineages (e.g., to separate brothers or father/son haplotypes)
  - Ballantyne, K.N. et al. (2010) *Am J Hum Genet* 87(3): 341-353
  - Ballantyne, K.N. et al. (2012) *FSI Genetics* (*in press*)

- **In some cases, being able to put a lineage name to an unknown Y-STR profiles using on-line genetic genealogy information**

Results of a Genetic Genealogy Search
with an "unknown" profile using (14 of 17) Yfiler loci

Compare	User ID	Pedigree	Last Name	Origin	Haplogroup	Tested With	Markers Compared	Genetic Distance
	KB56Q	Show	Smith	Slievenisky, Down, Northern Ireland	Unknown	Family Tree DNA	14	0
	XU3XE	Show	Butler	Ireland	Unknown	Family Tree DNA	14	0
	VAP7E		Butler	Ireland	R1b (tested)	Family Tree DNA	14	0
	74VV9	Show	Butler	Ireland	Unknown	Family Tree DNA	14	0
	T65UT		Butler	Ireland	Unknown	Family Tree DNA	14	0
	5BJX4		Butler	Ireland	Unknown	Family Tree DNA	14	0
	CYFNX		Butler	Unknown	Unknown	Family Tree DNA	14	0
	2B587		Butler	Unknown	Unknown	Family Tree DNA	14	0
	JSRJW		Butler	Unknown	Unknown	Family Tree DNA	14	0
	QWQG7		Butler	Unknown	Unknown	Family Tree DNA	14	0
	F9W7H		Butler	Unknown	Unknown	Family Tree DNA	14	0
	UXBFW		Butler	Unknown	Unknown	Family Tree DNA	14	0
	SFUSJ		Butler	Unknown	Unknown	Family Tree DNA	14	0
	4ZF4Z		Butler	Unknown	Unknown	Family Tree DNA	14	0
	P66AH		Harris	Unknown	R1b1a2*	Family Tree DNA	14	0
	W27DJ		Butler	Mississippi, USA	Unknown	Family Tree DNA	14	0
	VBVX9		Butler	South Carolina, USA	Unknown	Family Tree DNA	14	0
	FKNWZ		Butler	Mississippi, USA	Unknown	Family Tree DNA	14	0
	2NZ68		Butler	Quitman, Texas, USA	R1b*	Other - Anchestry by DNA	14	0
	ZFN67		Willhite (Adopted)	Tennessee, USA	Unknown	Family Tree DNA	14	0

17 of 20 full matches are "Butlers"

Other 3 are Butlers but didn't know it…
(adoption or other happenings in the gene pool of the past!)

www.Ysearch.org
Search conducted Jan 5, 2012

104,015 Records
80,143 Different Haplotypes
74,907 Surnames

Currently larger than YHRD – but serves a different purpose

YHRD Search Results (with 17 loci)

DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385
14	13	29	24	11	13	13	11,15

DYS438	DYS439	DYS437	DYS448	DYS456	DYS458	DYS635	YGATAH4
12	13	15	19	17	18	23	12

National database | Metapopulations | SNP

Whole database

Search Reset

Matches grouped by Metapopulations | Matches grouped by Continents | Matches grouped by Haplogroups | Frequency surveying estimates

- **All Metapopulation:** Found 0 of 39339 matching haplotypes [f=0 (95% CI: 0 – 9.377×10^{-5})] in 0 of 263 populations.
  - **Eurasian Metapopulation:** Found 0 of 15455 matching haplotypes [f=0 (95% CI: 0 – 2.387×10^{-4})] in 0 of 113 populations.
  - **East Asian Metapopulation:** Found 0 of 12522 matching haplotypes [f=0 (95% CI: 0 – 2.945×10^{-4})] in 0 of 63 populations.
  - **Australian Aboriginal Metapopulation:** Found 0 of 766 matching haplotypes [f=0 (95% CI: 0 – 4.804×10^{-3})] in 0 of 1 populations.
  - **African Metapopulation:** Found 0 of 1533 matching haplotypes [f=0 (95% CI: 0 – 2.403×10^{-3})] in 0 of 10 populations.
  - **Native American Metapopulation:** Found 0 of 384 matching haplotypes [f=0 (95% CI: 0 – 9.56×10^{-3})] in 0 of 9 populations.
  - **Eskimo Aleut Metapopulation:** Found 0 of 301 matching haplotypes [f=0 (95% CI: 0 – 1.218×10^{-2})] in 0 of 2 populations.
  - **Afro-Asiatic Metapopulation:** Found 0 of 1636 matching haplotypes [f=0 (95% CI: 0 – 2.252×10^{-3})] in 0 of 20 populations.
  - **Admixed Metapopulation:** Found 0 of 6742 matching haplotypes [f=0 (95% CI: 0 – 5.47×10^{-4})] in 0 of 45 populations.

0 matches found in 39,339 Yfiler profiles searched from 263 populations worldwide

Geographical projection

With 95% confidence interval

≈3/n = 3/39,339 = **1 in 13,113 ≈ 1 in 13,000**

How Many Butler Y-chromosomes Are Out There?

Katherine **Butler**

PhD student at George Washington University; Former TL of Bode Technology Group; Former VA DFS scientist

John **Butler**

NIST

<u>Some interesting points</u>:

1. **Katherine's father possesses an identical Yfiler 17-locus profile to John**
2. The first known Butler in John's lineage came to Virginia in the early 1700s – Katherine's family has been in Virginia since about the same time
3. Based on review of what they know from their family histories, they **cannot be closer than sixth or seventh cousins** (their 5th great-grandfathers differ)
4. **Potentially thousands of male Butlers have this same Yfiler haplotype** – or one very similar due to mutation at individual Y-STR loci
5. **A YHRD search that results in a value of <u>0 out of 39,339</u> Yfiler profiles does not reflect the true haplotype frequency in the world** (and especially Virginia)

Primary Steps Involved:
1 - Generate profile (Y or mtDNA)
2 - Query population database
3 - Report frequency estimate (with adjustment?)

Summary of Issues

Want good quality data going into database

population studies

Real-World Population Variation

Forensic Sample

1

HV1/HV2 control region mtGenome?

PP Y
Yfiler

mtDNA sequence (or difference from reference)

Y-STR Profile

2

Online **Population Database**

EMPOP
YHRD
US YSTR

Can θ correction help with this sampling issue?

Assume pop data representative of real-world

3 Counting method

Sampling confidence interval (normal or C-P; 1-tail or 2-tail)

?

What the court wants to know…

Haplotype Frequency Estimate
(of lineage <u>not</u> the individual)

UCF Acknowledgments

Funding from …

UCF Past and Present Team Members:
Erin Hanson, Lyn Fatolitis, Paulina Berdos

USYSTR Collaborators and Contributors:
ReliaGene, Promega , Applied Biosystems, University of Arizona

Illinois State Police, Orange County CA Coroner, Santa Clara County CA, DOJ Sacramento, Marshall University, WSP Vancouver, Richland County Sheriff SC, San Diego County Sheriff

NIST Acknowledgments

Funding from interagency agreements 2008-IJ-R-029 (and previously 1999 & 2003) between the National Institute of Justice and the NIST Office of Law Enforcement Standards

NIST Past and Present Team Members:

Amy Decker, Richard Schoske, Christian Ruitberg, Jill Appleby, Mike Coble, Becky Hill, Margaret Kline, Peter Vallone, Dave Duewer

Collaborators:

Mike Hammer, Alan Redd, Tom Reid,
ISFG DNA Commission, SWGDAM Y-STR Committee

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

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