Transition from U.S. Y-STR Database to YHRD

Dixie Peters, MS
UNTHSC Center for Human Identification
National Technical Leader Summit
Monday, December 3, 2018

Y-STR Databases

- Provides frequency estimates of a Y-STR haplotype
- Estimates can be divided by population (AA, Asian, Caucasian, Hispanic, Native American, etc.)
- Y-STR databases are dependent on contributions of sample data
  - Research/Academic, Forensic Laboratories
  - Database quality parameters
  - Sample info: kit and donor geographic and ethnic ancestry information
- As new Y-STR kits are employed, contributions of new sample data are needed

2014 SWGDAM Y-STR Interp Guidelines

- Section 10: Statistical Analysis of DNA Typing Results, pg. 10

A consolidated U.S. Y-STR database (http://usystrdatabase.org) has been established and may be used. Should a specific population(s) other than those provided in the U.S. Y-STR Database be required, the Y-Chromosome Haplotype Reference Database (YHRD) may be used (http://yhrd.org). A number of other Y-STR haplotype databases exist online [listed at http://www.cstl.nist.gov/biotech/strbase/y_strs.htm.]

U.S. Y-STR Database

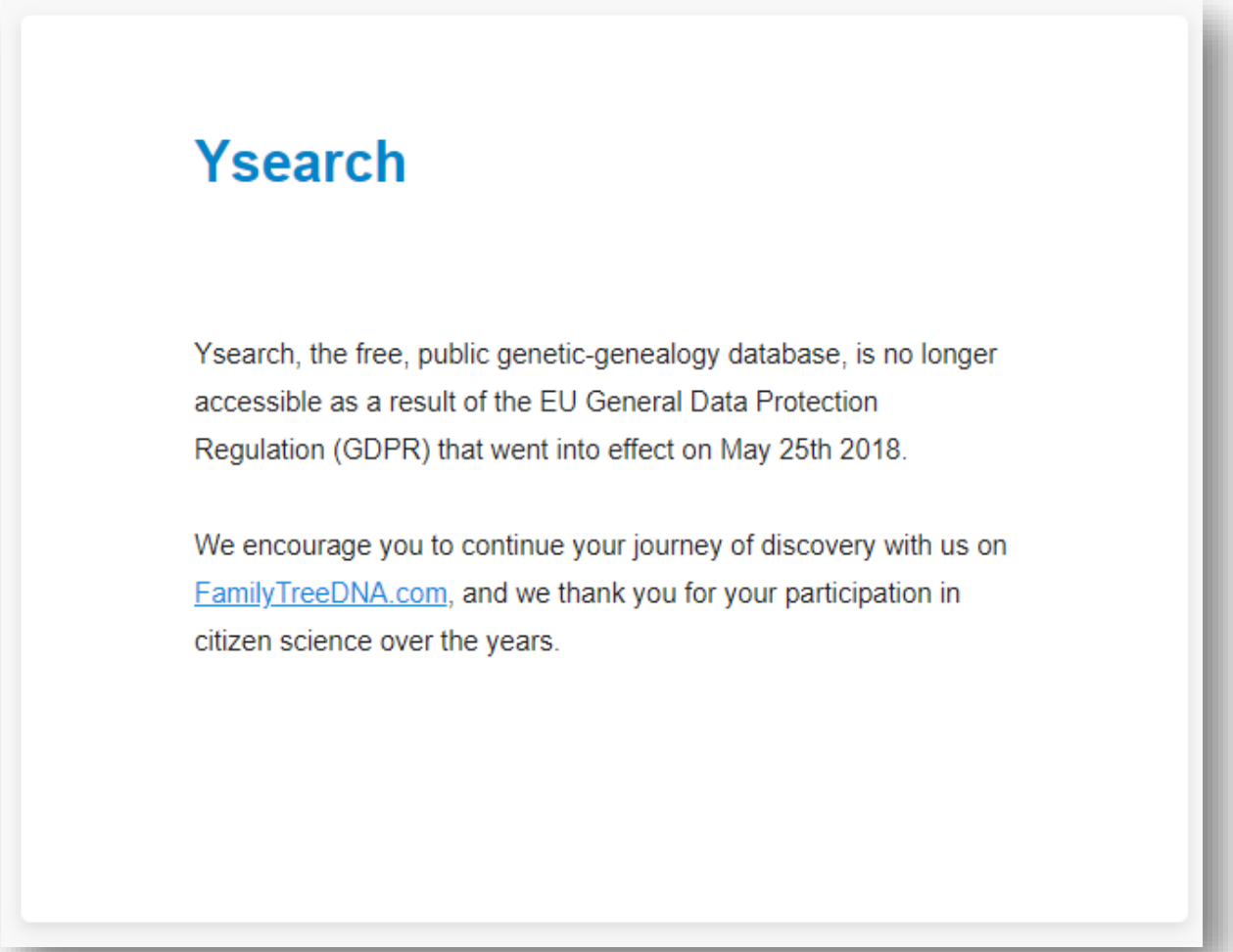
- Online since 2008 (https://www.usystrdatabase.org)
- Most recent update January 26, 2018 (Release 4.2.1)
- Initially funded by the National Institute of Justice
- Managed by the National Center for Forensic Science (NCFS) in conjunction with the University of Central Florida
  - Jack Ballantyne
  - Lyn Fatolitis
- Currently 35,658 haplotypes

YHRD

- Online since August 1, 1999 (https://yhrd.org)
- Most recent update November 1, 2018 (R59)
- Endorsed by ISFG; partial funding by LT and Promega
- Managed by Institute of Legal Medicine and Forensic Sciences, Charité - Universitätsmedizin Berlin
  - Lutz Roewer
  - Sascha Willuweit
- Currently houses 265,324 haplotypes (40,923 US samples)

Other Y-STR Databases

- *Genetic Genealogy:* FamilyTree DNA Y Search
- http://www.ysearch.org/



Ysearch

Ysearch, the free, public genetic-genealogy database, is no longer accessible as a result of the EU General Data Protection Regulation (GDPR) that went into effect on May 25th 2018.

We encourage you to continue your journey of discovery with us on FamilyTreeDNA.com, and we thank you for your participation in citizen science over the years.

Other Y-STR Databases

- *Genetic Genealogy:* Sorenson Molecular Genealogy Foundation
- http://www.smgf.org/

Sorenson Molecular Genealogy Foundation

We regret to inform you the site you have accessed is no longer available.

Sorenson Molecular Genealogy Foundation (SMGF) was founded in 2000 with the philanthropic goal of helping connect mankind. It was the organization's goal through the sharing of genetic data, to show how the similarities we possess are greater than our differences. The site was created in the spirit of openness and it is in that spirit AncestryDNA purchased the DNA assets from SMGF to further its mission and support the intentions on which it was founded. Unfortunately, it has come to our attention the site has been used for purposes other than that which it was intended, forcing us to cease operations of the site.

We understand the site has been a helpful resource for genealogists and plan to advance the original vision of Mr. Sorenson by continuing to develop tools like ethnicity estimates, matching, DNA Circles, and New Ancestor Discoveries, which are connecting mankind. There are no plans to destroy the DNA that was contributed, but have no plans to make the service available in the future.

Ancestry is committed to helping people understand their family's unique story and through AncestryDNA, make new discoveries about their family's past and cultural roots. Like the original founders of SMGF, Ancestry also beleves one can have a better understanding of who we are and where we come from. Through our continued work on family history and DNA, we will encourage the same mission of SMGF in hopes of making the world a smaller, more relatable place.

If you have any concerns, complaints, or questions about the SMGF study, or wish to withdraw consent to participate in the study, please contact AncestryDNA at:

Ancestry.com DNA, LLC
Member Services
1300 W Traverse Parkway
Lehi, Utah 84043
memberservices@ancestrydna.com
1-800-262-3787 or fax to 801-705-7001

Other Y-STR Databases

- Applied Biosystems Yfiler Haplotype Database
- http://www6.appliedbiosystems.com/yfilerdatabase/

ThermoFisher SCIENTIFIC
The world leader in serving science

Thermo Fisher Scientific
Life Sciences Solutions Group
180 Oyster Point Blvd.
South San Francisco, CA, 95080
T (650) 244-1000
www.thermofisher.com

November 7, 2018

RE: Discontinuation of the Applied Biosystems Yfiler™ Haplotype Database

Dear Valued Customer,

This letter is to inform you that effective December 9th, 2018, Thermo Fisher Scientific will decommission the Applied Biosystems Yfiler Haplotype Database (http://www6.appliedbiosystems.com/yfilerdatabase/).

The decision to no longer host the haplotype database aligns with the Interpretation Guidelines for Y-Chromosome STR Typing publication by the Scientific Working Group on DNA Analysis Methods in 2014. This publication recommends the usage of the U.S. Y-STR database (http://usystrdatabase.org) and the YChromosome Haplotype Reference Database (http://yhrd.org). Additionally, this action supports the 2014 announcement that additional haplotypes would not be added to the Applied Biosystems Haplotype Database upon release of the Yfiler™ Plus PCR Amplification Kit instead referring users to the two database resources listed above.

Thank you for your continued support of Thermo Fisher Scientific products. If you have any questions or comments, please do not hesitate to contact your local Human Identification representative.

Sincerely,

Human Identification Team
Life Sciences Solutions
Thermo Fisher Scientific

Why transition?

- U.S. Y-STR Database
  - Initially funded by NIJ, not permanent
  - Limited resources (admin staff, equipment, IT support)
  - Will be decommissioned June 30, 2019
- YHRD
  - Outside the U.S., most labs rely on YHRD
  - No reliance on grant funding
  - Worldwide database

Transition Process

1. Data transfer
2. Modifications to YHRD for 2014 SWGDAM compliance
3. Creation of notification and user’s guide

Data Transfer

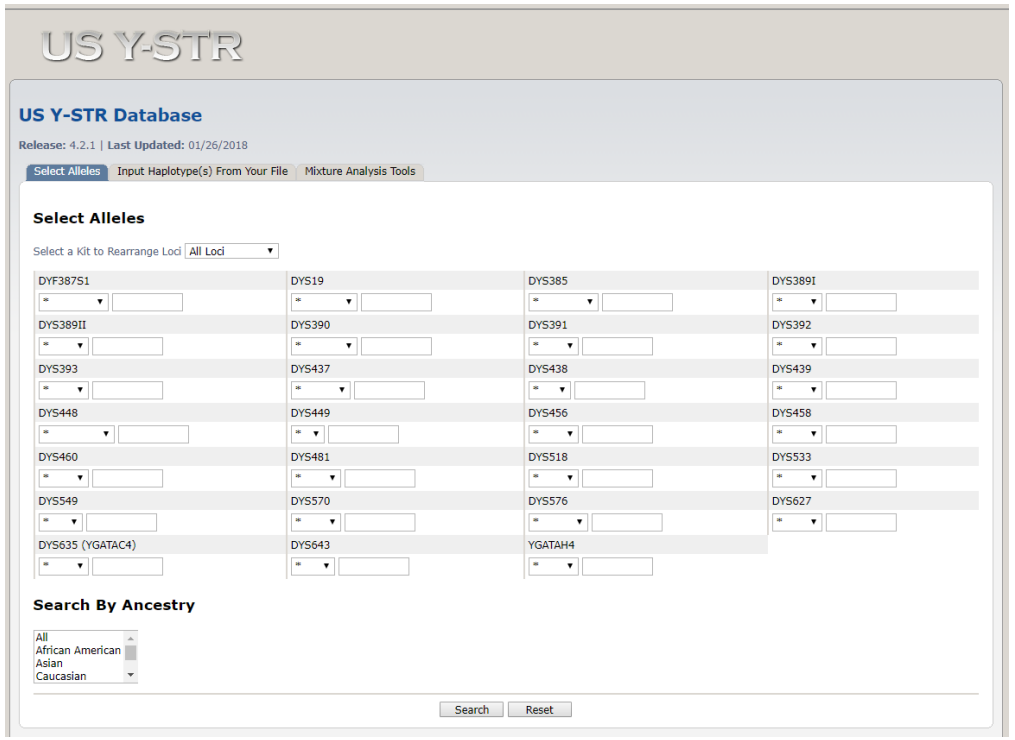
- Disassembled U.S. Y-STR database into single-origin files and provided required information:
  - Actual haplotype data (population sample)
  - Designation or Name of the population sample
    - State, United States, [Ethnicity]
- Created a list of Points of Contact: original submitter of haplotype data and maintains control of the data
- Created a list of local ethnicities to be visible at YHRD under National Database as sub-datasets

Modifications

- Alignment with 2014 SWGDAM Y-STR Interp Guidelines required calculation modifications
- YHRD provided outlines and test site of a “SWGDAM compliant” view of YHRD to reflect methods available at U.S. Y-STR Database
- U.S. Y-STR Database administrators and SWGDAM provided feedback until approved

Issues

- POC information
- 493 Rejected samples
  - Contained ‘non-designated alleles’ (< or >)
  - 277 were corrected
  - 216 population samples not transferred
  - YHRD requires confirmation of non-standard alleles
- Population sample duplicates
- Accommodating statistics used by U.S.



U.S. Y-STR + YHRD = BFF

Available as of November 1, 2018 (R59)

No additional updates to U.S. Y-STR Database

Notification and User’s Guide

- Released last week
- Notification will be posted on U.S. Y-STR Database and SWGDAM websites
- User’s Guide will be posted on U.S. Y-STR Database, SWGDAM, and YHRD websites

Notification

- Includes background information on U.S. Y-STR and YHRD
- Compares the U.S. database compositions
- Compares calculation differences
- Compares search results of PP Y23 and Yfiler Plus haplotypes

Table 1. Comparison of U.S. database compositions.

	U.S. Y-STR Database	National United States in YHRD
Minimal	35,658 (11 loci)	40,923 (9 loci)
PP Y	32,594	35,864
Yfiler	26,007	29,277
PP Y23	5,305	5,717
Yfiler Plus	2,094	2,124
Maximal	575	575

Table 2. Calculation differences between U.S. Y-STR Database and YHRD.

	U.S. Y-STR Database	YHRD
Theta Value used	Yes	No
Count Method	x/n	(x+1)/(n+1)
Confidence Interval	1-sided	2-sided

- YHRD Release 59 includes calculations that are compliant with the 2014 SWGDAM *Interpretation Guidelines for Y-Chromosome STR Typing*
- Use the "National Database (with Subpopulations, 2014 SWGDAM-compliant)" option in YHRD
- YHRD provides:
  - Observed haplotype frequencies (x/n) for each subpopulation (African American, Asian, Caucasian, Hispanic, Native American, and overall) including a 1-sided 95% upper confidence interval
  - Two combined theta-corrected match probabilities (with and without the Native American population)

Theta-corrected Match Probabilities

- Both databases use the theta values described in Appendix 1 of the 2014 SWGDAM *Interpretation Guidelines for Y-Chromosome STR Typing*
- U.S. Y-STR Database separates the theta-corrected match probabilities by major population group while YHRD combines all populations (without and, where data exists, with the Native American population) to calculate the Overall theta-corrected match probabilities

Theta-corrected Match Probabilities

- YHRD limits theta-corrections to profiles with fewer than 23 loci, regardless of which loci or multiplex are searched
  - If needed, population-level match probabilities can be calculated outside of YHRD using Eq. 3 and theta values from Appendix 1 from the 2014 *Guidelines*
- U.S. Y-STR Database will apply theta to all searches of any number of loci as long as the Yfiler Plus kit locus order is not selected for profile entry

Search Results Comparison

- Four tables included in notification
- A PP Y23 and a Yfiler Plus haplotype were searched
- Profile probability differences are minimal

Table 4. Comparison of haplotype profile probabilities (expressed as 1/Profile Probability) after searching a PP Y23 haplotype against PowerPlex Y, Yfiler, and PowerPlex Y23 Caucasian datasets using the U.S. Y-STR Database (4.2.1) and YHRD (R59).

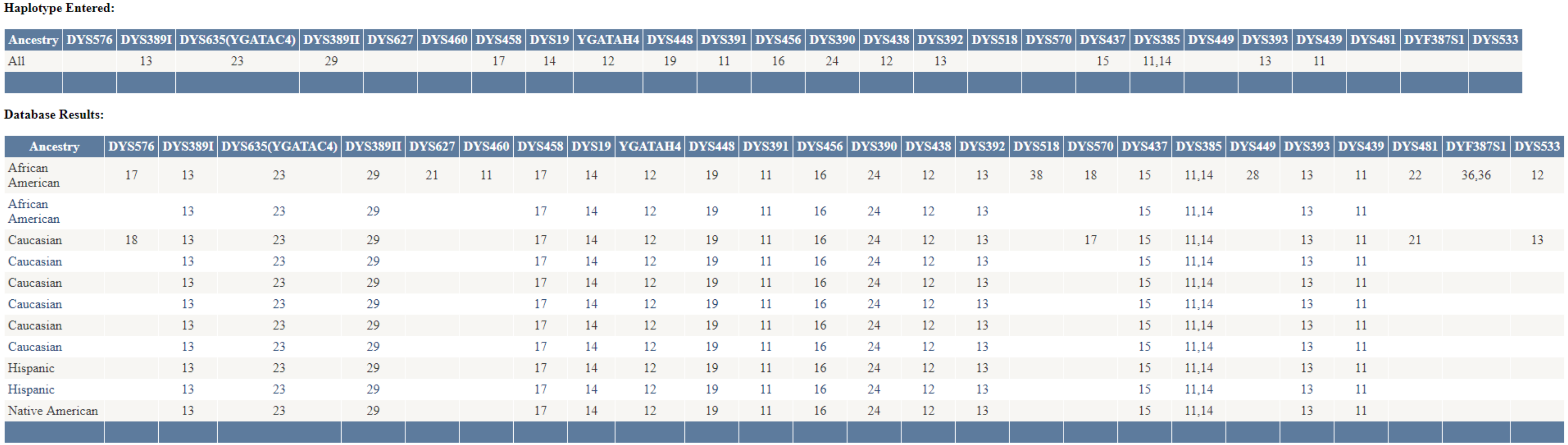
	U.S. Y-STR (95% UCI)		YHRD (95% UCI)	
	United States (Caucasian)		United States (Caucasian)	
DATASET	**# of Haplotypes**	**1/Profile Prob.**	**# of Haplotypes**	**1/Profile Prob.**
PP Y	3 of 9,855	1,282	3 of 10,889	1,405
Yfiler	3 of 7,449	962	3 of 8,483	1,094
PP Y23	2 of 1,494	238	2 of 1,549	246

User’s Guide

- 2014 SWGDAM Compliant YHRD User’s Guide
- Overview
  - Section 1: How to search a Y-STR Haplotype
  - Section 2: How to search a Y-STR Haplotype with reduced loci
  - Section 3: How to check the current state of the database
  - Section 4: How to check release history
  - Section 5: How to obtain locus information

Reduced Locus Searches

- 2014 SWGDAM Interp Guidelines, Section 10.1 describes use of reduced locus searches
- Unlike U.S. Y-STR Database, YHRD does not provide haplotypes of matching population samples for exclusion evaluation



Haplotype Entered:

Ancestry	DYS576	DYS389I	DYS635(YGATAC4)	DYS389II	DYS627	DYS460	DYS458	DYS19	YGATAH4	DYS448	DYS391	DYS456	DYS390	DYS438	DYS392	DYS518	DYS570	DYS437	DYS385	DYS449	DYS393	DYS439	DYS481	DYF387S1	DYS533
All		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			

Database Results:

Ancestry	DYS576	DYS389I	DYS635(YGATAC4)	DYS389II	DYS627	DYS460	DYS458	DYS19	YGATAH4	DYS448	DYS391	DYS456	DYS390	DYS438	DYS392	DYS518	DYS570	DYS437	DYS385	DYS449	DYS393	DYS439	DYS481	DYF387S1	DYS533
African American	17	13	23	29	21	11	17	14	12	19	11	16	24	12	13	38	18	15	11,14	28	13	11	22	36,36	12
African American		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Caucasian	18	13	23	29			17	14	12	19	11	16	24	12	13		17	15	11,14		13	11	21		13
Caucasian		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Caucasian		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Caucasian		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Caucasian		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Caucasian		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Hispanic		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Hispanic		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			
Native American		13	23	29			17	14	12	19	11	16	24	12	13			15	11,14		13	11			

How to search a Y-STR Haplotype with reduced loci

YHRD Reduced Locus Search Instructions

Underlined text represents an example for illustration purposes. In the example spreadsheet accompanying these instructions, the corresponding cells are as follows: x_1 = G4, n_1 = H4; x_2 = G8, n_2 = H8 and H11, x_3 = G9, x_4 = G10, x_5 = G11.

Note: this procedure requires calculations outside of the YHRD website.

1. Search the full multiplex haplotype (PP Y23) in YHRD against the Worldwide and US National Database. Record x_1, n_1, 1 in (x/n), and 1 in (95% UCI) values.
2. Within the YHRD report results page, select the next smallest multiplex (Yfiler) to search the haplotype with a reduced number of loci. Record x_2 and n_2.
3. Next, reselect the full multiplex haplotype within the YHRD report results page to search the reduced number of loci against the dataset for the larger multiplex (Yfiler loci against PP Y23 dataset). Record the number of matches (x_3).
4. Subtract the number of matches (x_3) obtained from the third search (PP Y23 dataset with Yfiler loci) from the number of matches (x_2) obtained from the second search (Yfiler). The value obtained (x_4) represents the number of haplotypes from the smaller dataset (Yfiler) that are excluded from the larger dataset (PP Y23) and are representative of that multiplex only (Just Yfiler). Note: haplotypes are excluded because they were tested with the larger multiplex and did not match at the additional loci.
5. To obtain the total number of matches from both datasets, add the number of matches (x_1) observed from the original full multiplex haplotype (PP Y23) to the number of matches (x_4) calculated in the previous step (Just Yfiler). Record this value (x_5). Use the n_2 from the larger dataset (from search in step 2) to calculate the frequency estimate (1 in x/n) and the profile probability (1 in 95% UCI).
6. The process from step 2 through step 5 can be repeated for all existing smaller multiplexes (PPY and Minimal). Combine the number of matches observed with each dataset (Total PPY = PPY23 + Just Yfiler + Just PPY; Total Minimal = PPY23 + Just Yfiler + Just PPY + Just Minimal) to calculate new frequency estimates and profile probabilities.
7. Determine the most and least discriminatory profile probabilities and report per your laboratory's guidelines.

These instructions guide the user through the entire process. The following slides will illustrate these steps with an example. Note: these instructions refer to items on a spreadsheet that is shown in the next slide.

How to search a Y-STR Haplotype with reduced loci

A spreadsheet is recommended to record values obtained from reduced locus searches. Note: cells in this spreadsheet are referenced in the previous slide.

	A	B	C	D	E	F	G	H	I	J
1			Worldwide				US (SWGDAM Total)			
2			PPY23 or YFP loci				PPY23 or YFP loci			
3			x	n	1 in (x/n)	1 in (95% UCI)	x	n	1 in (x/n)	1 in (95% UCI)
4		PPY23 or YFP profiles	2	50692	25,346	8,052	2	5717	2,859	908
5										
6			Yfiler loci				Yfiler loci			
7			x	n	1 in (x/n)	1 in (95% UCI)	x	n	1 in (x/n)	1 in (95% UCI)
8		Yfiler profiles	26	205059			12	29277		
9		PPY23 or YFP profiles	10				5			
10		Just Yfiler	16				7			
11		Total Yfiler	18	205059	11,392	7,683	9	29277	3,253	1,864
			PPY loci				PPY loci			
			x	n	1 in (x/n)	1 in (95% UCI)	x	n	1 in (x/n)	1 in (95% UCI)
		PPY profiles	732	224657			233	35864		
		Yfiler profiles	646				189			
		Just PPY	86				44			
		Total PPY	104	224657	2,160	1,835	53	35864	677	538
			Minimal loci				Minimal loci			
21			x	n	1 in (x/n)	1 in (95% UCI)	x	n	1 in (x/n)	1 in (95% UCI)
22		Minimal profiles	1906	265324			576	40923		
23		PPY profiles	1569				503			
24		Just Minimal	337				73			
25		Total Minimal	441	265324	602	556	126	40923	325	280
26										
27										
28		Notes:		= a cell in which you'll enter search results						
29			Total Yfiler = Just Yfiler + (PPY23 or YFP)							
30			Total PPY = Just PPY + Total Yfiler = Just PPY + Just Yfiler + (PPY23 or YFP)							
31			Total Minimal = Just Minimal + Total PPY = Just Minimal + Just PPY + Just Yfiler + (PPY23 or YFP)							
32			Per YHRD Search Instructions, G4 = X_1; H4 = n_1; G8 = X_2; H8 & H11 = n_2; G9 = X_3; G10 = X_4; G11 = X_5							

What’s next?

- Laboratories:
  - Update SOPs as appropriate
  - Searches performed in U.S. Y-STR are still valid and do not need to be searched again in YHRD
  - Searches performed using YHRD R58 may need to be repeated in YHRD R59
  - Perform any necessary internal performance checks
  - Submit new data to YHRD
- SWGDAM:
  - Updating Y-STR Interp Guidelines

Performance Check

- 32 Yfiler haplotypes
  - 22 with no matches/observations in U.S. Y-STR Database
  - 10 with matches/observations in U.S. Y-STR Database
- Compare to matches observed in YHRD
- Compare profile probabilities

No matches in U.S. Y-STR Database

- All 22 Yfiler haplotypes that did not match to any U.S. Y-STR Database sample, also did not match to any YHRD sample
- Obtained the same profile probabilities for all 22 haplotype searches in the respective databases
- Differences in profile probabilities are due to differences in database sizes

Profile Probabilities for 0 matches											
African Amer		Asian		Caucasian		Hispanic		Native Amer		TOTAL	
US Y-STR	YHRD	US Y-STR	YHRD	US Y-STR	YHRD	US Y-STR	YHRD	US Y-STR	YHRD	US Y-STR	YHRD
1 in...	1 in...	1 in...	1 in...	1 in...	1 in...	1 in...	1 in...	1 in...	1 in...	1 in...	1 in...
2083	2377	1335	1363	2488	2832	1592	2007	1190	1196	8696	9773

Matches in U.S. Y-STR Database

Sample	OBSERVATIONS & PROFILE PROBABILITIES (1 in...)																							
	African Amer				Asian				Caucasian				Hispanic				Native Amer				TOTAL			
	US Y-STR (N = 6234)		**YHRD** (N = 7118)		**US Y-STR** (N = 3998)		**YHRD** (N = 4083)		**US Y-STR** (N = 7449)		**YHRD** (N = 8483)		**US Y-STR** (N = 4765)		**YHRD** (N = 6012)		**US Y-STR** (N = 3561)		**YHRD** (N= 3581)		**US Y-STR** (N= 26007)		**YHRD** (N = 29277)	
	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...
1001	2	1000	2	1131	0	1335	0	1363	6	633	6	717	2	758	2	955	1	758	1	755	11	1429	11	1608
5001	0	2083	0	2377	0	1335	0	1363	1	1613	1	1789	0	1592	0	2007	0	1190	0	1196	1	5556	1	6172
5003	1	1316	1	1501	0	1335	0	1363	5	714	5	807	0	1592	0	2007	0	1190	0	1196	6	2273	6	2472
6002	0	2083	0	2377	0	1335	0	1363	2	1190	2	1348	0	1592	0	2007	0	1190	0	1196	2	4167	2	4651
6004	1	1316	1	1501	0	1335	0	1363	0	2488	0	2832	0	1592	0	2007	0	1190	0	1196	1	5556	1	6172
6011	2	1000	2	1131	0	1335	0	1363	1	1613	2	1348	0	1592	0	2007	0	1190	0	1196	3	3571	4	3199
6015	0	2083	0	2377	0	1335	0	1363	1	1613	0	2832	0	1592	0	2007	0	1190	0	1196	1	5556	0	9773
6017	0	2083	0	2377	0	1335	0	1363	1	1613	1	1789	2	758	3	776	0	1190	0	1196	3	3571	4	3199
6020	0	2083	0	2377	0	1335	0	1363	1	1613	1	1789	0	1592	0	2007	0	1190	0	1196	1	5556	1	6172
6024	0	2083	0	2377	0	1335	0	1363	5	714	8	588	0	1592	0	2007	1	758	1	755	6	2273	9	1864

- 10 Yfiler haplotypes: 1-11 matches in U.S. Y-STR Database
- 6 of 10 haplotypes: same frequencies in YHRD
- 3 of 10 haplotypes: larger frequencies in YHRD
- 1 of 10 haplotypes: no observations/matches in YHRD (DYS439 issue)
- Generally, can expect the same or similar frequencies in both databases
- Similar profile probabilities

Maximal Profile (29 loci)

- Both databases contain same number of maximal profiles (575)
- Searched 1 maximal profile
- Both databases generated same frequencies and profile probabilities

Sample	OBSERVATIONS & PROFILE PROBABILITIES (1 in...)																			
	African Amer				Asian				Caucasian				Hispanic				TOTAL			
	US Y-STR (N = 242)		**YHRD** (N = 242)		**US Y-STR** (N = 2)		**YHRD** (N = 2)		**US Y-STR** (N = 229)		**YHRD** (N = 229)		**US Y-STR** (N = 102)		**YHRD** (N = 102)		**US Y-STR** (N= 575)		**YHRD** (N = 575)	
	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...	Freq	1 in...
Maximal	0	81	0	81	0	1	0	1	1	49	1	49	0	35	0	35	1	122	1	122