Notice to U.S. Forensic Laboratories on the status of the U.S. Y-STR Database

Funded by the National Institute of Justice, the U.S. Y-STR Database (http://usystrdatabase.org) has been managed by the National Center for Forensic Science at the University of Central Florida since 2007 (Ballantyne et al. 2006), and its Release 4.2.1 (January 26, 2018) contains 35,658 Y-STR haplotypes generated by thirty U.S. forensic, academic and commercial laboratories. As of November 26, 2018, 217,758 search queries of the database had been performed. To mitigate encumbrances in the administration of resources and to ensure long-term operational stability, the U.S. Y-STR Database haplotypes have been permanently transferred to the Y- Chromosome Haplotype Reference Database (YHRD, http://yhrd.org) for continuance of usage, and the U.S. Y-STR Database will be decommissioned (scheduled for June 30, 2019).

The Scientific Working Group on DNA Analysis Methods *Interpretation Guidelines for Y-Chromosome STR Typing* (approved January 9, 2014) provides guidance on the calculation of profile probabilities and match probabilities subsequent to the determination of a match between two samples subjected to Y-STR typing. In accordance with the *Guidelines*, the U.S. Y-STR Database may be searched to establish the frequency of the Y-STR haplotype in the database and, with application of the population structure parameter theta (θ), to estimate the conditional probability of a random match to the haplotype given that it has already been observed once. The *Guidelines* also support usage of the Y- Chromosome Haplotype Reference Database (YHRD, http://yhrd.org), should a specific population(s) other than those provided in the U.S. Y-STR Database be required.

Originating in 1999, YHRD Release 59 (November 1, 2018) consists of 265,324 nine-locus (“minimal”) haplotypes, 224,657 PowerPlex Y, 205,059 Yfiler, 50,692 PowerPlex Y23, 42,506 Yfiler Plus, and 5,516 29-locus (“maximal”) haplotypes from many worldwide meta-populations including the United States. The U.S. database within YHRD includes the addition of the majority of haplotypes contained in the U.S. Y-STR Database. Table 1 compares the compositions of the U.S. Y-STR Database to the National United States Database contained within YHRD.

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

While not the preferred reporting statistic by that organization, YHRD may be used to estimate haplotype frequencies using the counting method (Holland and Parsons 1999), which entails searching a haplotype against the database and determining the database frequency. The U.S. Y-STR Database reports observed database frequencies using the direct count (x/n, where x is the number of matching haplotypes, and n is the size of the relevant database) accompanied by the 1-sided 95% Exact Confidence Interval (Clopper and Pearson 1934). YHRD reports x and n, but the "Expected" values add 1 to both the numerator and denominator and are calculated as (x+1)/(n+1), accompanied by the 2-sided 95% Exact Confidence Interval. This method generates a frequency estimation and differs from the U.S. Y-STR approach recommended by SWGDAM (2014) as shown in Table 2.

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

In spite of these calculation differences, the administrators of YHRD have kindly modified the National Database search calculations in Release 59 to include those that are compliant with the 2014 SWGDAM *Interpretation Guidelines for Y-Chromosome STR Typing*. Therefore, when searching a haplotype using the "National Database (with Subpopulations, 2014 SWGDAM-compliant)" option in YHRD, a user will be provided with the 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 and two combined theta-corrected match probabilities (with and without the Native American population).

Important differences exist between the theta-corrected match probabilities reported in each of the databases. YHRD limits theta-corrections to profiles with fewer than 23 loci, regardless of which loci are searched and the multiplex selected, while the 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. Although 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 (African American, Asian, Caucasian, Hispanic, and Native American), while YHRD combines all populations (without and, where data exists, with the Native American population) to calculate the Overall theta-corrected match probabilities.[1] Relevant case information regarding the pool of possible alternate contributors may be used as a guide when selecting between YHRD match probabilities that exclude or include the Native American data. If desired, the population-level match probabilities that are not supplied by YHRD can be calculated

[1] The theta values provided in Appendix 1 of the 2014 SWGDAM *Guidelines* were calculated comparing average within-population match proportions to between population match proportions. As such, they address the issue of substructure between the major population groups within the total population, not between subpopulations within the major population groups. While they do not directly address within-population substructure, it was anticipated that they would provide conservative surrogates under the expectation that differences between major population groups would be larger than differences within major populations groups. In that spirit, U.S. Y-STR Database applied these theta estimates to population-specific match probabilities. YHRD instead, applies them only to the combined (Overall) database results for which these theta estimates are directly applicable.

outside of that website using the YHRD search results for each population, Eq. 3 from the 2014 *Guidelines*, and theta values from Appendix 1 of the 2014 *Guidelines*.

To demonstrate the similarity of estimates for the same PP Y23 and Yfiler Plus haplotypes when searched against U.S. Y-STR Database and YHRD, refer to Tables 3 through 6.

A sample from the University of North Texas (UNT) population database of individuals from the state of Texas (Davis et al. 2013) known to match at all 23 PP Y23 loci to a presumed unrelated sample also in the UNT database, was searched in both the U.S. Y-STR and YHRD databases using the full haplotype (PP Y23) and again using subsets of the PP Y23 loci that corresponded to those in the Yfiler and PP Y multiplexes. Both samples are contained in the U.S. Y-STR Database and YHRD; therefore, matches were expected. The haplotype profile probability (expressed as 1/Profile Probability) for the loci found in each dataset (PowerPlex Y, Yfiler, and PowerPlex Y23) is presented in Table 3 (total of all subpopulations) and Table 4 (Caucasian subpopulation only). For the U.S. Y-STR Database, there were 2 matches to 5,305 haplotypes in the total PP Y23 database, giving a profile probability of 1 in 847. Using the YHRD "National Database (with Subpopulations, 2014 SWGDAM-compliant)" search option that contains only the haplotypes from the United States, 2 matches to the 5,717 total PP Y23 haplotypes were observed giving a profile probability of 1 in 908. Finally, including a theta correction for the PP Y23 match including Native Americans reduced the match probability to 1 in 714, given the same observation of 2 in 5,717. The theta corrected estimate without the Native Americans (2 of 4,836 observations) gave a match probability of 756. As shown in Table 4, in the U.S. Y-STR Caucasian database, there were 2 matches to 1,494 PP Y23 haplotypes and in the YHRD Caucasian database, there were also 2 matches to 1,549 PP Y23 haplotypes. The profile probabilities are 238 (U.S. Y-STR database) and 246 (YHRD).

Table 3. Comparison of haplotype profile and match probabilities (expressed as 1/Profile Probability or 1/Match Probability) after searching a PP Y23 haplotype against PowerPlex Y, Yfiler, and PowerPlex Y23 datasets using the U.S. Y-STR Database (4.2.1) and YHRD (R59).

	U.S. Y-STR (95% UCI)		**YHRD (95% UCI)**		**YHRD Match Probability (theta w/ NA)**		**YHRD Match Probability (theta w/o NA)**	
	United States (Total)		**United States (Total)**		**United States (Total)**		**United States (Total)**	
DATASET	**# of Haplotypes**	**1/Profile Prob.**	**# of Haplotypes**	**1/Profile Prob.**	**# of Haplotypes**	**1/Match Prob.**	**# of Haplotypes**	**1/Match Prob.**
PP Y	3 of 32,594	4,545	3 of 35,864	4,626	3 of 35,864	896	3 of 31,495	1,548
Yfiler	3 of 26,007	3,571	3 of 29,277	3,776	3 of 25,696	2,764	3 of 29,277	1,504
PP Y23	2 of 5,305	847	2 of 5,717	908	2 of 5,717	714	2 of 4,836	756

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

Similar to the PP Y23 example above, a sample from the NIST Caucasian population database known to match a non-related sample also in the NIST database at all 27 Yfiler Plus loci was searched in both the U.S. Y-STR and YHRD databases using the full haplotype (Yfiler Plus) and again using subsets of the Yfiler Plus loci that corresponded to those in the Yfiler and PP Y multiplexes. Autosomal STRs and mtDNA data were used to confirm the sources of the samples are not closely related. Both samples are contained in the U.S. Y-STR Database and YHRD; therefore, matches were expected. The haplotype profile probability (expressed as 1/Profile Probability) for the loci found in each dataset (PowerPlex Y, Yfiler, and Yfiler Plus) is presented in Table 5 (total of all subpopulations) and Table 6 (Caucasian subpopulation only). For the U.S. Y-STR Database, there were 3 matches to the 2,092 haplotypes in the total Yfiler Plus database, giving a profile probability of 1 in 270. Using the YHRD “National Database (with Subpopulations, 2014 SWGDAM-compliant)” search option that contains only the haplotypes from the United States, 3 matches to the 2,124 total Yfiler Plus haplotypes were observed giving a profile probability of 1 in 274. As the Yfiler Plus haplotype contains more than 22 loci, a theta corrected match probability was not calculated for the matches to the Yfiler Plus dataset. As shown in Table 6, in the U.S. Y-STR Caucasian database, there were 3 matches to 576 Yfiler Plus haplotypes and in the YHRD Caucasian database, there were also 3 matches to 577 Yfiler Plus haplotypes. The profile probabilities are 75 for both the U.S. Y-STR Database and YHRD.

Table 5. Comparison of haplotype profile and match probabilities (expressed as 1/Profile Probability or 1/Match Probability) after searching a Yfiler Plus haplotype against PowerPlex Y, Yfiler, and Yfiler Plus datasets using the U.S. Y-STR Database (4.2.1) and YHRD (R59).

	U.S. Y-STR (95% UCI)		**YHRD (95% UCI)**		**YHRD Match Probability (theta w/ NA)**		**YHRD Match Probability (theta w/o NA)**	
	United States (Total)		**United States (Total)**		**United States (Total)**		**United States (Total)**	
DATASET	**# of Haplotypes**	**1/Profile Prob.**	**# of Haplotypes**	**1/Profile Prob.**	**# of Haplotypes**	**1/Match Prob.**	**# of Haplotypes**	**1/Match Prob.**
PP Y	7 of 32,594	2,500	8 of 35,864	2,485	8 of 35,864	768	7 of 31,495	1,224
Yfiler	6 of 26,007	2,273	7 of 29,277	2,227	7 of 29,277	1,178	6 of 25,696	1,920
Yfiler Plus	3 of 2,092	270	3 of 2,124	274	n/a	n/a	n/a	n/a

Table 6. Comparison of haplotype profile probabilities (expressed as 1/Profile Probability) after searching a Yfiler Plus haplotype against PowerPlex Y, Yfiler, and Yfiler Plus 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	6 of 9,855	833	7 of 10,889	828
Yfiler	5 of 7,449	714	6 of 8,483	717
Yfiler Plus	3 of 576	75	3 of 577	75

In general, the profile probability estimates between U.S. Y-STR Database and YHRD [National Database (with subpopulations) United States] are all similar and within the same order of magnitude. This is to be expected as many of the same sets of Y-STR data from the U.S. have historically been contributed to both databases by researchers, commercial entities, and practitioners.

The impact of switching to YHRD from the U.S. Y-STR Database should show a minimal impact in resulting statistics. Previous statistics generated using the U.S. Y-STR Database are valid and should not require a recalculation with YHRD. Updates on the status of both databases may be found at the respective websites. A revision to the Scientific Working Group on DNA Analysis Methods *Interpretation Guidelines for Y-Chromosome STR Typing* is forthcoming. A U.S. user's guide on how to use YHRD for haplotype and reduced locus searches can be found at http://usystrdatabase.org, http://yhrd.org, and https://www.swgdam.org/.

References

SWGDAM (2014) Interpretation Guidelines for Y-Chromosome STR Typing by Forensic DNA Laboratories. Available at http://media.wix.com/ugd/4344b0_da25419ba2dd4363bc4e5e8fe7025882.pdf

Ballantyne J, Fatolitis L and Roewer L (2006) Creating and Managing Effective Y-STR Databases. *Profiles in DNA* **9(2)**: 10–13. Available at https://www.promega.com/resources/profiles-in-dna/2006/creating-and-managing-effective-ystr-databases/

Holland MM and Parsons TJ (1999) Mitochondrial DNA Sequence Analysis – Validation and Use for Forensic Casework. *Forensic Science Reviews* **11(1):** 22-50.

Clopper CJ and Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26:** 404–416.

Davis C, Ge J, Sprecher C, Chidambaram A, Thompson J, Ewing M, Fulmer P, Rabbach D, Storts D, Budowle B. (2013) Prototype PowerPlex® Y23 System: A concordance study. *Forensic Sci Int Genet.* **7(1):** 204-208.