Y-STR DATABASES

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

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.

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 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 AmpF/STR® 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

Y-STR DATABASES

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

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.

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.

Y-STR DATABASES

THE WEB-BASED Y-CHROMOSOME HAPLOTYPE REFERENCE DATABASE FOR FORENSIC USE

By Lutz Roewer

A long-term international project initiated by the Forensic Y-Chromosome Research and User Group to facilitate forensic exploitation of human Y-chromosomal STR markers has resulted in a large database of haplotypes collected worldwide. This repository, known as the Y-STR haplotype reference database (YHRD, available at: **www.yhrd.org)**, supports the decision-making process of forensic analysts, regardless of their national affiliation. Nine forensically evaluated Y-STR loci with high levels of variability in all world populations were chosen as the basis of this database and to define the minimal haplotype (minHt). These markers are part of popular commercial multiplex kits, such as AmpF*l*STR® Yfiler™, PowerPlex® Y, Mentype® Argus Y-MH QS and *genRES*® DYSplex-1 and -2. The extended haplotype (extHt), already determined for a third of the samples in the database, includes the markers DYS438 and DYS439 and is endorsed by SWGDAM.

Currently, the 19th release of YHRD includes 41,965 7-locus haplotypes from 357 populations (blue loci in Figure 1). 40,108 of these (95%) are typed for the 9-locus minHt, including DYS385a/b (320 populations), and 14,837 (35%) in 98 populations for the 11-locus extHt (purple loci in Figure 1). The YHRD was established through the joint efforts and continuing work of a steadily growing network of laboratories (currently about 120 labs in 45 countries). Online searches for complete or partial Y-STR haplotypes from evidentiary or nonprobative material can be performed on a noncommercial basis to yield observed haplotype counts as well as extrapolated population frequency estimates (Figure 2).

To ensure accurate Y-STR genotyping and use of an approved standard nomenclature, proficiency tests are an obligate requirement for all participating laboratories prior to data submission. The quality control test (QC) involves blind haplotyping of five DNA samples for the extHt. Once a contributing laboratory has passed the QC, haplotype profiles are submitted electronically to the central YHRD server at the Institute of Legal Medicine, Charité—University Medicine Berlin. Data files are transmitted in a standardised format to avoid extra work (and introduction of errors). Haplotype profiles go through a number of plausibility checks prior to data entry. Each record comprises the respective allele designations, geographic coordinates, population and metapopulation affiliation (e.g., a sample from the recruitment unit "Paris, France" is assigned to three population and search levels: 1. Eurasian; 1.1 European; 1.1.1 Western European). Also included is a unique proband Y-chromosome identifier that is used internally to identify samples for possible future haplotype expansion by analysis of additional STRs.

Two approaches are currently available to evaluate the probability of a coincidental match between two Y-STR haplotypes: the counting method and the haplotype surveying method. The latter is a Bayesian approach that attempts to extract more information from the structure of Y-STR haplotype databases than does the counting method. The surveying method yields a β-type posterior distribution, the mean of which is a robust estimator

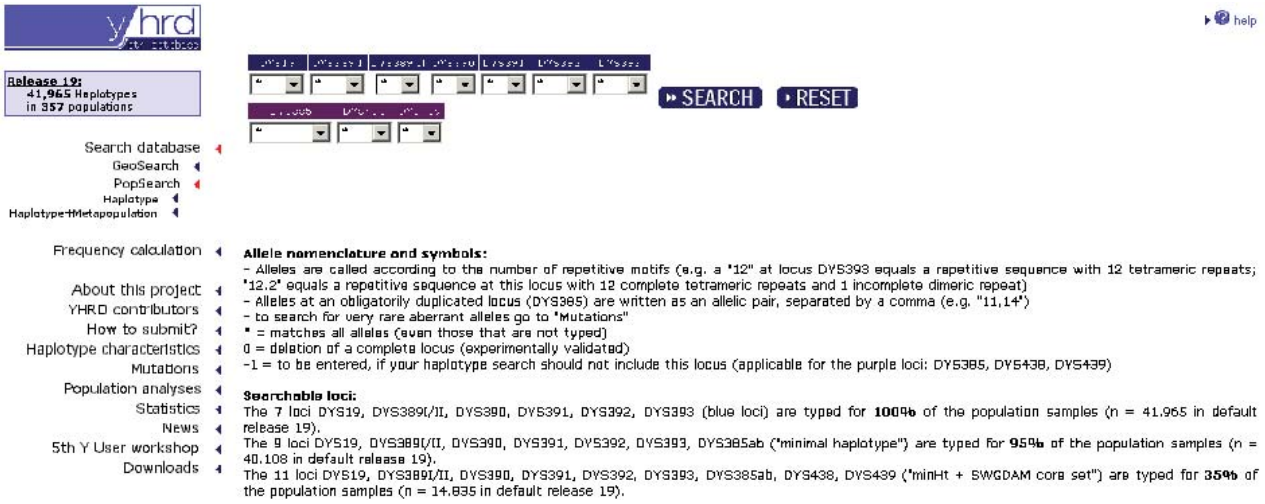

Figure 1. PopSearch menu of the Y-STR Haplotype Reference Database.

Y-STR DATABASES

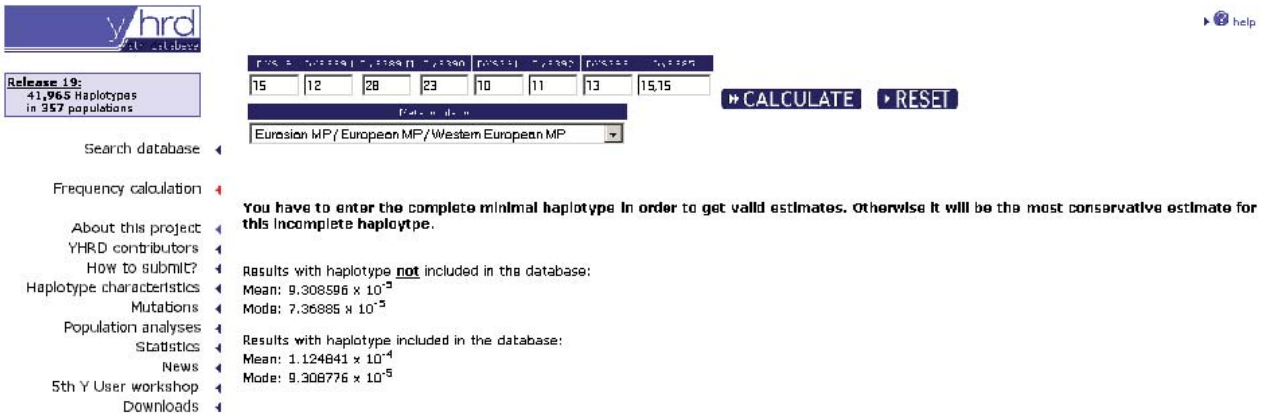

Figure 2. Extrapolation of the frequency of the author's rare haplotype using the Western European metapopulation as the reference. (Haplotype count in the whole YHRD, release 19: 2/40,108; count in the Western European metapopulation (MP): 2/13,237).

of haplotype frequency (Figure 2). This posterior distribution is obtained by extrapolating from the structure and frequency of all other haplotypes in the relevant subdatabase (e.g., Western Europeans). The method makes use of the fact that, in the same population, the frequency of a haplotype positively correlates with the combined frequency of all "neighbouring" haplotypes that differ by only a small number of mutational steps. Since it is not included in the estimation process, the haplotype in question does not have to be present in the database. The match probabilities obtained with the counting or surveying method are in good agreement for frequent haplotypes (>0.1% in a reference population). For rare haplotypes, the discrepancies are slightly more pronounced since the database size represents a lower limit of the prior frequency estimates that can be obtained via the counting method. The frequency calculation program is already implemented in the YHRD for three well sampled European metapopulations (Figure 2), whereas the easily defensible counting method is still recommended for all other metapopulations until sufficiently large databases are available.

The sensitivity of Y-chromosome-specific genetic markers to population differentiation necessitates appropriate sampling strategies. Populations such as Europeans, regarded as sufficiently homogeneous for the purpose of autosomal STR typing, show a marked geographic differentiation of Y-chromosome haplotypes. Frequencies vary strongly between Western, Eastern and Southeastern Europe and many embedded isolates, reflecting historic and prehistoric demographic events. Intense efforts by the forensic genetics community are currently underway to increase sample sizes for populations shown to be genetically homogeneous entities. The progress of these efforts can be monitored; the YHRD (with continuous updates of haplotype and population counts) allows online searches via the "PopSearch" menu (Figure 1). For this search algorithm, the haplotypes in the database are classified according to widely accepted, linguistically and genetically defined metapopulations (Figure 3). Thus, frequencies retrieved by counting or extrapolation can be reported for the whole database as well as for the relevant metapopulation (Figure 2).

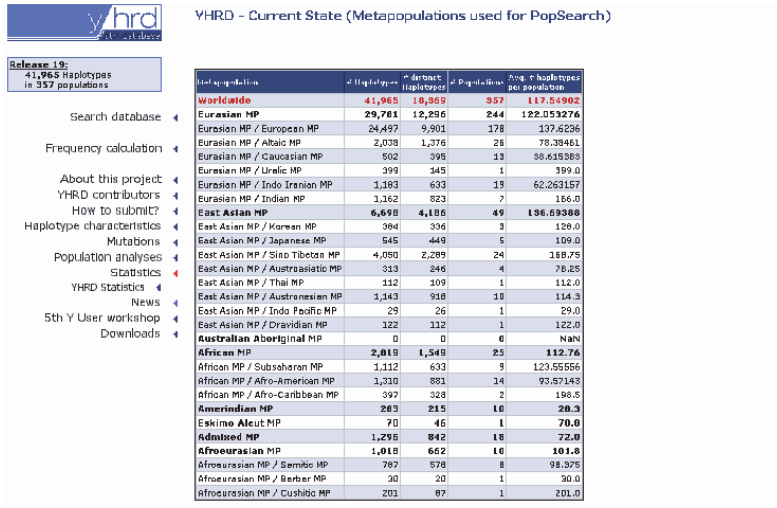

Figure 3. Metapopulation structure and number of samples per metapopulation in the YHRD (release 19, August 1, 2006). Note that population names as "Caucasians", often misused for Europeans, are replaced by "Eurasian MP/European MP", correctly returning the designation "Caucasian" to the language group from the Caucasus region.