



III. International Forensic Y-User Workshop **Y chromosome haplotype database(s): state of the art and future developments**

Y-SNPs

Chris Tyler-Smith

Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

The accumulating data on Y-SNPs and their potential uses in forensics and elsewhere will be summarised. Expectations. The Y chromosome is estimated to contain about 25 Mb of euchromatic, non-recombining DNA, and two Y chromosomes chosen at random from the world population are expected to vary at 1 position in approximately 7 kb. There will therefore be about 3,500 SNP differences between two random Y chromosomes. Over 23,000 Y-SNPs were reported in the database 'Ensembl' by March 2002. About 1 SNP mutation is expected per Y chromosome per meiosis. The potential of SNPs for discriminating between Y chromosomes therefore seems to be great. Reality. In practice, these expectations may not easily be fulfilled. It is not known what proportion of the >23,000 database entries represent true Y-SNPs, but the majority appear to represent differences between the X and Y chromosomes, or between sequences that are repeated on the Y. Furthermore, the lack of recombination means that many Y-SNPs will inevitably identify the same lineages. A collaborative analysis of 245 Y markers (mostly SNPs) by the Y Chromosome Consortium has proposed a tree and unified nomenclature for Y haplogroups. These show strong geographical structuring, often making them poor markers for discrimination within a population, but providing some indirect information about geographical origins. Y-SNPs are thus excellent markers for many anthropological studies, but the forensic usefulness of a small panel of Y-SNPs seems limited.



III. International Forensic Y-User Workshop **Y chromosome haplotype database(s): state of the art and future developments**

Interpretation issues in forensic Y-STR casework

Mechthild Prinz

Department Forensic Biology, Office of Chief Medical Examiner, New York City

The emergence of DNA databases for evidence and convicted offender profiles has shifted the emphasis from individual cases, where the evidence in question needs to be compared to known subjects, to a lot of non-suspect casework. Now it is not sufficient to be able to do the comparison, the main priority is the generation of a database eligible profile. For most semen evidence such a profile can only be produced after a differential lysis extraction since unresolved mixed samples create many fortuitous matches in a database.

While in the past, the OCME had been using Y-STR testing to screen for the presence of typable "male" DNA before proceeding with the time consuming differential lysis, this strategy was changed after taking on all non-suspect cases. Now a differential lysis is performed first and Y-STR amplification is only included if no DNA that could not have come from the victim is found, or if it is necessary to determine the number of semen sources in a case.

For saliva evidence the approach is reversed. We could show that especially on vaginal swabs, Y-STR testing has a higher success rate for this type of body fluid. Therefore the first DNA test will be a Y-STR multiplex and only if this test is successful an autosomal amplification is performed.

The presence of allele duplications, female artefacts, and even a negative result seemingly contradicting a high presumptive test value, can create interpretation and reporting challenges. Wording conclusions has to be careful and should avoid absolute statements about the number of semen donors or the absence of male DNA.



III. International Forensic Y-User Workshop Y chromosome haplotype database(s): state of the art and future developments

NIST Y Chromosome Standards and Multiplex Assays

John M. Butler¹, Richard Schoske^{1,2}, Peter M. Vallone¹, Janette W. Redman¹, and Margaret C. Kline¹

¹National Institute of Standards and Technology, Gaithersburg, MD, USA

²American University, Department of Chemistry, Washington, DC, USA

At the National Institute of Standards and Technology (NIST), we have been developing a human Y chromosome Standard Reference Material, SRM 2395, which will soon be available to enable calibration of Y STR results across laboratories worldwide. SRM 2395 includes 5 male DNA samples selected to exhibit a diverse set of alleles across the commonly used Y chromosome short tandem repeat (STR) and single nucleotide polymorphism (SNP) markers. A female DNA sample is also included to serve as a negative control for male-specific DNA tests. The five male samples in SRM 2395 have been sequenced at more than 20 commonly used Y STR markers (in both forward and reverse directions) to confirm allele calls. The Y STR loci typed and sequenced for SRM 2395 include DYS19, DYS385a/b, DYS388, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS426, DYS435, DYS436, DYS437, DYS438, DYS439, DYS447, DYS448, DYS460, DYS464 a/b/c/d, Y-GATA H4, and YCAII a/b.

Typing results from the NIST Y STR 20plex [1], the NIST Y STR 10plex [2], and the ReliaGene Y-PLEX™ 6 and Y-PLEX™ 5 kits (see www.reliagene.com) will be discussed. We have also developed some additional multiplexes that include new Y STR markers discovered by Mike Hammer's group [3]. Casework examples from forensic specimens will be demonstrated with the various Y STR multiplexes.

In addition, we are developing a number of Y SNP assays using multiplex PCR and multiplexed SNP detection from the SNaPshot™ assay on multi-color fluorescence capillary electrophoresis (ABI 310 and ABI 3100 platforms). Several dozen Y SNP markers are under evaluation and being studied in various U.S. population groups. Results from 42 Y SNPs (Marligen Biosciences' Signet™ Y SNP Identification System) run on a liquid array Luminex platform will also be shown.

We are posting recent developments in Y chromosome DNA tests on the STRBase web site: <http://www.cstl.nist.gov/biotech/strbase/>. Y STR fact sheets that describe primer sequences, allele sizes and sequences, and references to population data are available for the loci listed above. A standard reference format for Y STR allele nomenclature is also posted on STRBase.

References

[1] Butler, J.M., *et al.* (2002) *Forensic Sci. Int.* 129: 10-24

[2] Ruitberg, C.M. and Butler, J.M. (2000) New primer sets for Y chromosome and CODIS STR loci. Poster presented at the Eleventh International Symposium on Human Identification, Biloxi, MS.

[3] Redd, A.J., *et al.* (2002) *Forensic Sci. Int.*, *in press*.

Quality control and standardization in Y chromosome typing

Angel Carracedo, Leonor Gusmão

Institute of Legal Medicine. Santiago de Compostela.

If DNA analysis is nowadays accepted in countries all over the world, it is in part due to the progress made in standardization.

Standards are crucial for forensic geneticists. This is due to the fact that only with an agreement about standards is it possible to develop quality control programs and quality assurance programs. In other words, standards are the only way to guarantee the judges, juries and the public that the tests performed and laboratory efficiency are reliable in any specific case. In addition, standards are necessary to allow for second opinions for the exchange of data between labs and for the creation of uniform searching procedures in cross border crime.

Two types of standards need to be addressed: technical and procedural. Technical standards include matters such as the genetic systems to be used (including type, nomenclature and methodology), the statistical methods for evaluating the evidence and the communication of the final report. Procedural standards encompass matters of operation such as laboratory accreditation, laboratory performance, accreditation and licensing of personnel, record keeping and proficiency testing.

There are many national and international standardization groups for forensic DNA typing and some of them have addressed standardization issues of the forensic use of Y chromosome polymorphism. The DNA Commission of the ISFG provides a general framework for the coordination of the main standardization groups and it has recently published recommendations for the forensic use of Y STRs.

Concerning procedural standards proficiency testing is a crucial step for accreditation. Up to now among the main PT programs only GEP-ISFG proficiency testing program has regularly included during the last few years Y chromosome STRs. The results of the GEP-ISFG PT scheme for the Y chromosome show that there are still some problems on Y STR nomenclature that are progressively being solved. This demonstrates the usefulness of proficiency testing as an important instrument for the dissemination of common standards.

The lack of Y STR typing in most of the PT schemes both for paternity testing and criminal casework is a problem for labs with accreditation requirements following ISO17025 or similar guidelines. An open universal PT scheme for the Y chromosome would be an important achievement in the field.

There is still a general problem with the evaluation of the evidence. First because there is a need of standards in some particular aspects. Second because statistics is included as paper challenges in some PT schemes and the results are far from good. However in the case of Y STRs LR calculations have been facilitated by the statistical facility included in the European Y STR database which fulfil ISFG recommendations.



III. International Forensic Y-User Workshop **Y chromosome haplotype database(s): state of the art and future developments**

Do Y-STR databases always provide reliable estimates of haplotype frequencies ?

Peter de Knijff

Forensic Laboratory for DNA Research (FLDO), Department of Human Genetics, Center for Human and Clinical Genetics (CHCG), Leiden University Medical Center (LUMC), PoBox 9503, 2300 RA Leiden, The Netherlands.

Although, at least in Holland, Y-STRs are not frequently used for routine forensic case-work, they can be extremely important in a few cases. Obtaining a reliable Y-STR profile should no longer be a problem. However, what about the interpretation of the obtained results ? Consider the case where a Y-STR haplotype of a crime-stain matches with a suspect. The most obvious approach would be to search the Berlin-based YSTR database – or any other representative Y-STR frequency database - to get an impression of the relative frequency of this haplotype. Obviously, in the report a disclaimer should be included stating that at least all paternal relatives of the suspect will carry an identical profile and can not be excluded as the donor of this stain solely on the basis of DNA-evidence. If the profile of our crime-sample has not been observed before we assume it is a rare haplotype. Under normal circumstances – and by the uninitiated prosecutor - this will be interpreted as a good and strong indication that the crime-sample comes from the suspect.

This interpretation entirely relies on the frequency of the observed haplotype in a Y-STR database. In the case of the Berlin database, this is filled with data supplied by most of us using our random DNA-sample collections. These collections were primarily meant to be used for all our autosomal loci. We have thus assumed that these "random" DNA-collections could be used for the inference of autosomal allele frequencies and allele frequencies of lineage-markers such as Y-STRs. I will argue that this could be a dangerous assumption for a number of reasons.

Most of our DNA collections are ideally suited for obtaining reliable allele frequencies for autosomal loci. However, we should not call them random. In contrast, they are extremely selected for unrelatedness. It is this aspect which renders these Y-STR databases, at least in theory, unsuitable for obtaining reliable Y-STR haplotype frequencies. Theory predicts that common Y-STR haplotypes have underestimated frequencies and that rare haplotypes have overestimated frequencies. I will illustrate this with one recent example.

A simple solution for this problem does not exist. In those cases where Y-STR (or mtDNA) is the only available evidence directly linking a suspect to a criminal offence, a good understanding of local population structure would be ideal. However, in many cases such an understanding is simply not available and not easily to be obtained. It is my personal opinion that we should treat matching Y-STR and mtDNA evidence with great caution, especially in front of the uninitiated legal representatives of the defence and prosecution.

Towards the identification of ethnic and geographic origin by Y-chromosomal genetic markers: Lessons from the South Amerindian communities

**Daniel Corach, Miguel Marino, Gustavo Penacino, Andrea Sala,
Carina Argüelles, Alberto Fenocchio, Veronica Tesa, Lutz Roewer,
Omar Rocabado, Peter Nürnberg, Mohammad Reza Toliat, Nico Ruf,
Michael Krawczak**

Human identification using DNA typing is usually based upon the analysis of autosomal short tandem repeats (STRs). Although highly informative, these markers are however unable to provide sufficient information about the ethnic or geographic origin of the donor of a sample. In cases where no suspect is available, such as in rape, homicide or burglary, or when fragmentary human remains is the only source of genetic material (e.g. in suicide terrorist attacks), it would nevertheless be desirable to determine at least the ethnicity or the geographic origin of the evidentiary material. Uniparentally inherited genetic markers (i.e. Y-chromosomal or mtDNA polymorphisms) may provide the necessary information for such identification. Since most violent crimes are perpetrated by males, gender-specific genetic markers may indeed represent an valuable identification tool.

Mammalian Y chromosomes are characterized by the absence of recombination for most of the chromosome (except PAR1 and PAR2 at the telomers of both arms) and by the presence of a wide range of polymorphic loci. These include alphoid satellite variants, minisatellites (MYS1) and microsatellites (Y-STR), all of which have high mutation rates, as well as so called Unique Event Polymorphisms (UEPs) that are characterized by low mutation rates. The latter are usually biallelic and comprise Single Nucleotide Polymorphisms (Y-SNPs) as well as Insertion/Deletion events (ins/del). UEPs, which also include Y-chromosomal Alu Polymorphisms (YAPs), can provide valuable extra information about sample origin when analyzed in combination with highly variable Y-STRs.

In order to investigate this forensically relevant aspect further, different South-Amerindian communities were investigated with different Y-chromosomal markers. Three Bolivian aboriginal communities from the La Paz region were selected, namely the Lecos (N=22), Mosevenes (N=16) and Moxeños (N=25). All samples were typed by a Y-STR nonaplex (DYS19, DYS385, DYS389I and II, DYS390, DYS391, DYS392 and DYS393). In addition, three communities of the North-Argentinian Guarani-M'bya tribe originating from different localities were investigated using the Y-STR nonaplex, YAPs, SNPs (SRY4064, PN1, PN2 and DYS199), and the mtDNA Region V 9bp ins/del polymorphism. HVRI sequencing of samples carrying the 9 bp deletion allele was carried out in order to confirm the presence of Amerindian mtDNA haplogroup B. The total genetic information was then compared with that of an Argentinian sample randomly selected from different regions of the country (N=50).

A remarkable haplotype diversity was observed within the Bolivian tribes and, although located in geographical proximity, no Y-STR haplotypes were shared by the different communities. In addition, a reduced number of their haplotypes were detected, at low frequency, in the existing YHRDs (Caucasian, North American or Asian). In contrast, the Guarani communities, although more distantly located, showed remarkably similar haplotype distributions. They were characterized by a reduced haplotype diversity and an ethnic-specific combined haplogroup (ESCH), defined as DYS199/T, DYS19/13, DYS392/14, DYS393/11, whose frequency is 43%. The differences detected between tribes may reflect the distinct historical processes that lead to the extant communities.

Is the ethnic origin of male DNA detectable by Y-STR haplotyping and the use of large reference databases ? Examples and discussion.

Lutz Roewer

Institut für Rechtsmedizin, Humboldt-Universität, Hannoversche Str. 6, D-10115 Berlin

The detection of the population or ethnic background of trace DNA can reduce the time required for identifying a (fugitive) perpetrator. It may also help in the identification of mutilated victims of crime and disasters. This should not however lead to a situation where the police reduces their efforts in criminal investigation unduly. Any violation of human rights must of course also be excluded.

Owing to its large size, reduced effective population size and patrilinear transmission, the Y chromosome is probably best suited to analyse the evolution and stratification of populations. *It contains within it a relatively simple record of the human past and represents the most informative haplotypic system in the human genome* as stated in: Genetics (2002) 160(1): 289-303. Different strategies exist to mine the polymorphic sequence archives of population history provided by the Y-chromosome. It has been shown that sets of Y-SNPs, typed in hierarchical order, are probably the best tools to identify the ancestral patrilineages of humankind and to follow their dispersal throughout the globe. Forensically evaluated and established Y-STR-based haplotypes are characterized by a much higher variability, discriminative power and mutation rate than that of SNPs. STR-based haplotypes form clusters of related profiles, emerging from one another through slippage mutagenesis, which evolve on the background of ancient lineages that can be defined nearly unequivocally by SNP-analysis. The advantage of Y-STR as compared to Y-SNP typing is (i) the sensitivity to more recent demographic events and population genetic effects, (ii) the established typing technologies, used by nearly all forensic labs in their routine work (including the chemistry based on multiplex kits with fluorescent primers), and (iii) the availability of large easily accessible databases.

In a given case, it is straightforward to type a trace or tissue at first by autosomal STRs for personal identification via National offender (or missing persons) databases and secondly by Y-STRs for further individualization and a subsequent search in large population databases. Most of these database are standardized and comprise loci that are also included in commercial kits. The online-availability of more than 15.000 standardized haplotypes ("minimal haplotype" comprising 9 loci) from Europe, Asia and the US within the Y-STR haplotype reference databases (YHRD) facilitates these searches. By examples taken from real forensic casework, we show how Y-STR based haplotype distributions differ even among the genetically homogeneous regions of Europe. Clines exist in Europe e.g. between the Mediterranean West and Mediterranean East territories, between the North East and North Central Europe, the South East and the North Central regions (defined according to Gamble, 1986) due to topographic, linguistic, religious, and socio-cultural barriers or disastrous historical events (war, famine, epidemic diseases). Isolates like, for example, the Hungarian and Bulgarian Romani are clearly distinct from the surrounding populations. Even more obvious is the separation between Europeans and East Asians (and less pronounced of course within the Indo-Hittite linguistic supergroup stretching between India and Europe), Africans and Europeans and other populations separated for a long time by genetic isolation.



III. International Forensic Y-User Workshop

Y chromosome haplotype database(s): state of the art and future developments

Recently, prosecutors and courts in Germany have ordered for the first time Y-STR haplotyping “to report on the ethnic affiliation of the donors of traces” left at the scene. These pilot applications show the suitability of the Y-STR approach. Instead of reporting probability values for the affiliation of a haplotype to a population - which is currently impossible - we recommend to give observed counts in clearly - geographically and linguistically- defined reference samples accompanied by a statement about the most probable ancestry of the patrilineage in question. This could be more or less decisive, since due to clinal frequency distributions and different time-depth of lineages a haplotype of a Polish male, for example, could be found to be either of “Eurasian descent”, “European descent”, “Eastern North European descent”, “descent from a Slavic speaking population” or truly “most probably of Polish descent”. In any case, geographically or linguistically separable clusters have to be identified in the databases since single matches are of no significance.



III. International Forensic Y-User Workshop Y chromosome haplotype database(s): state of the art and future developments

An in-depth survey of European Y-chromosomal STR haplotypes

Michael Krawczak

Institut für Medizinische Informatik und Statistik, Brunswiker Str. 10, 24105 Kiel, Germany

An in-depth analysis has been performed of the European data included in the Y-STR database, maintained at the Humboldt-University, Berlin, Germany. In order to avoid practical and/or theoretical problems arising from possible allelic ambiguities, the analysis was confined to those seven STRs that allow unambiguous haplotyping, namely DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393. By the time of analysis, the database contained 2433 different haplotypes for these markers, as observed in 11610 males from 81 different European populations. The most frequent haplotype, 14-13-29-24-11-13-13, was present in 62 populations and, with 627 copies, outnumbered the next most frequent haplotype by more than two-fold. However, the overall haplotype diversity of 99.3% was high and reflects a substantial number of rare haplotypes. Indeed, 2232 haplotypes (91.7%) were observed less than 10 times, and 1403 haplotype (57.7%) were only observed in single copy. Populations showed substantial genetic differentiation, with 1350 haplotypes (55.5%) being unique to a single population, and 2011 haplotypes (82.7%) observed in five or fewer populations. The average pair-wise ϕ_{ST} was equal to 0.0734, ranging from -0.0159 (South-Norway vs. West-Norway) to 0.3863 (Baranya Romani vs. South-Ireland). In general, genetic similarity of Y-STRs among European males is found to be correlated with both geographic distance and linguistic relatedness. For comparison, an Asian sample comprising 14 populations (N=1924) was also included in the analysis. Of the 1101 different haplotypes found in this dataset, 408 were shared with the European sample. The average pair-wise ϕ_{ST} between Asians and Europeans of 0.1318 was only marginally higher than the intra-Asian value of 0.1181. Also, Asians exhibited a higher haplotype diversity (99.8%) and a higher proportion of private haplotypes (880 = 79.9%) than European males.