Basic Priniciples for Estimating the Rarity of Y-STR Haplotypes Derived from Forensic Evidence

Bruce Budowle, Jianye Ge, and Ranajit Chakraborty

Introduction

The Y-chromosome STR loci are male-specific and thus typing of these genetic markers enable analysis of previously problematic forensic biological samples, such as mixtures comprised of low quantities of male DNA amidst a large background of female DNA. In addition, some paternal lineage questions may be elucidated with Y STR loci. (1-9). The loci that have been validated for forensic applications include DYS19, DYS385 (counted as two loci), DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS456, DYS458, DYS635, DYS448, and Y GATA H4. Their use is promoted by availability of robust and reliable commercial multiplex kits - Powerplex Y (12 locus multiplex kit; Promega Corp., Madison, WI) (10) and Y filer (17 locus multiplex kit; Applied Biosystems, Foster City, CA) (11).

When the evidence and reference sample Y STR profiles are sufficiently similar (i.e., matches or can be a component of a mixed sample) and thus potentially could have originated from the same source, some significance is placed on the observation, typically by estimating the rarity of the evidentiary profile. (12). The interpretation of inclusion or exclusion for Y STR loci profiles is the same as for autosomal STR loci profiles. However, when there is a failure to exclude, the approach for estimating the rarity of a Y DNA profile differs from that applied to autosomal DNA markers, because of the non-recombining features of the Y-STR loci typed. This paper describes the basic aspects of generating reliable statistical inferences of the rarity of Y STR evidentiary profiles. Three concepts are summarized. They are:

1. Use of the counting method and the accompanying upper bound haplotype frequency for estimating a conservative frequency of the observed haploptype

2. Accommodating reasonable possible effects of population substructure

3. Assessing the significance of a mixture

The Weight of Y STR Haplotype Evidence

The forensically important Y STR loci reside in the non-recombining portion of the hemizygous Y chromosome. Thus, the assumption of independence between and among loci does not apply for the Y STR loci. The product rule, even the modified product rule (13), is not appropriate for estimating the rarity of a Y STR haplotype. Essentially, because of linkage, each haplotype is treated as an allele and the total number of possible haplotypes comprises the alleles of a single locus. This is the same for all Y STR systems, no matter how many different loci are used to generate the haplotype profile.

Because the haplotypes are treated as alleles of a single locus, the counting method, which is simple, robust, and conservative, can be applied for assessing the significance of a Y STR haplotype derived from an evidence sample (14,15). The counting method is carried out as follows:

1. The profile of a Y STR haplotype generated from an evidence sample is searched against a reference database(s) of unrelated individuals;

2. The number of times the Y STR haplotype is observed in a database(s) is counted (termed “count”);

3. That count is divided by the number of profiles in the reference database(s);

4. A confidence interval is placed on the proportion of count/total profiles in a database(s).

For the foreseeable future, the size of the available reference databases will be limited such that most haplotypes (certainly the 12-17 loci profiles generated using commercial kits) will not be observed or only observed once in a reference population database. Because of this constraint, an estimate of the frequency of the haplotype in a population is not possible by calculating the proportion of the number of matching profiles divided by the total number of profiles in the database. To address this limitation, a conservative bound is placed on that estimate (count/total) to correct for possible sampling error. The confidence interval allows for a measure of the amount of confidence which can be placed on a value lying between two specified limits (i.e., the interval). One can calculate the upper bound of the confidence interval and this value can be used to convey with a high degree of confidence that the rarity of the evidence Y-haplotype among unrelated individuals in a given population(s) is less than the upper bound of the estimate. The assumption of a Normal distribution may not apply for Y-STR haplotype frequency estimates (16); however, assuming normality will provide a conservative upper bound estimate. Exact methods for constructing confidence intervals are available (17,18).

The counting method has several features that make it desirable for forensic applications. First, it is a conservative approach. Although the counting method treats all loci as physically linked (which they are biologically), not all Y STR loci are in linkage disequilibrium to the same extent. Some pairs of Y-STRs fail to show significant linkage disequilibrium and there are other pairs for which the linkage disequilbrium is incomplete (i.e., less than the maximum possible; 14, 19). Kayser et al. (20), Heyer et al (21), and

Budowle et al (14), have typed confirmed father/son pairs for 12 or more Y-STR loci and have observed mutation rates on the order 10^{-3}. This rate of mutation will somewhat destabilize the effects of linkage, because of the expansion-contraction (of allele size) nature of mutations. Therefore, the assumption of complete linkage disequilibrium contributes to the conservative nature of the counting method. Second, although Y-STR profiles are not as individualizing as are an equal (and even fewer) number of autosomal STR loci, it is still unlikely to draw two unrelated people at random with the same minimal haplotype. Furthermore, in some cases, where no result can be obtained by autosomal STR typing, Y STR typing may be the only viable approach to obtain a result. Thus, even with their combined lower power, the Y STR loci may be the best tool available in certain cases to exclude those individuals falsely associated with the sample. Although Y STR haplotypes are treated as single locus systems, the forensic Y multiplex systems exhibit greater diversity than any single autosomal STR locus. This observation contributes to exceedingly low Fst values used for correcting for the effects of possible substructure (see below). Third, a similar counting approach has been in use for more than 10 years for estimating the rarity of evidentiary mtDNA haplotypes (22). So the counting method is a well established application in the forensic community and in legal proceedings.

Population Substructure

The effects of population structure are expected to be greater for the Y-chromosome than with the forensically important autosomal STR loci. This is due to: 1) the non-recombining Y STR loci are not shuffled during meiosis as are the autosomal

genetic markers; and 2) Y chromosomes have an effective population size that is 1/4 that of autosomal loci. This linkage and smaller effective population size contribute to population specific distributions, which are affected by genetic drift and geographic differentiation. Indeed, population substructure effects have been shown to be more substantial for Y loci compared with that observed for the autosomal STR loci (see e.g., 23 and their cited references). This would suggest that 1) reference population databases may have to be defined more restrictively (most likely by geography and ethnohistory) than is typically done for autosomal STR loci; or 2) current practices of pooling databases based on major population groups can be employed if corrections are employed for effects of substructure when estimating the rarity of the profile (13). It is incumbent upon forensic scientists to assess the effects of population substructure and employ statistical approaches that address those effects.

As recommended by the NRC II Report (13), θ can be used to adjust for possible population substructure. Using the general theory, the unconditional frequency of a haplotype (A_i), which is the count divided by the sample size, can be modified to obtain the conditional probability

$$\Pr(A_i|A_i) = [p_i^2 + \theta\, p_i(1-p_i)]/p_i$$

$$= p_i + \theta(1-p_i) \text{ or}$$

$$= \theta + p_i\,(1 - \theta)$$

Based on this formula, the conditional probability always exceeds the quantity of θ, making θ the minimum bound for estimating the rarity of a Y STR haplotype. So the

probability of observing a particular haplotype in an unrelated individual given that it is seen in another male is highly dependent on the value of θ deemed appropriate for the relevantY-STR haplotype database.

The value of θ (computed by either R_{ST} or F_{ST}) is dependent indirectly on the number of Y STR loci comprising the haplotype. Generally, as more loci are included in the haplotype, most haplotypes in a data set will become differentiated. Therefore, the greater the number of loci within the haplotypes, the smaller should be the value of θ.

R_{ST} has been used to estimate the value of θ for forensic calculations (24). However, the analyses based on R_{ST} are better applied to evolutionary biology for studying the phylogenetic history of Y-chromosomes. The R_{ST} values are based on allele size variance, exploiting the extent of difference between different haplotypes. Such an approach, however, typically does not apply to forensic inferences. Forensic applications assess the evidence in terms of match or non-match. In these terms, the θ values for Y STRs should not be computed on mismatch based approaches (such as AMOVA), but instead by simply treating all haplotypes as different alleles. Thus, F_{ST} (or G_{ST}) is a better estimator for θ relevant for forensic applications, because haplotypes are identified solely by their distinctiveness (i.e., haplotypes are considered simply in terms of identity by state). This generally leads to a more appropriate and much smaller θ value than estimated by R_{ST}.

The F_{ST} values across major populations range from less than 0.002 to less than 0.0008 (depending on the number of loci comprising the haplotype). The population-specific values are smaller, excluding Native Americans (manuscript in preparation). Therefore, with the size of current reference population databases, F_{ST} rarely will

influence the upper bound of the Y STR frequency. Since F_{ST} does not affect the upper bound frequency estimate on any practical level for most population groups (with the current commercially available systems), increasing the size of reference population data sets and including more populations would be more valuable and a better use of resources than adding more loci to the current multiplexes for exploiting the full power of Y STR typing. Dedicating efforts to expanding the battery of Y STR loci will not significantly exploit the power of the assays.

As population comparisons are carried out, significant, but small F_{ST} values at the level of haplotypes for the Y-STRs may and have been observed (24 and manuscript in preparation). This raises the question of the legitimacy of pooling of population samples for forensic calculations, as is done for autosomal STR loci (25). However, as long as the effects of population substructure are addressed adequately, pooling of populations is appropriate. Since most haplotypes are unique in a reference population data set, the conservative counting method already more than compensates for population substructure effects, when pooling geographic samples of populations of broad definitions (e.g., US Caucasian, African American, Hispanic, etc.).

One might attempt to suggest that because of significant differences (albeit small) in population comparisons that no general reference (population specific) database can be used for forensic statistical computations. This position would ignore the fact that regardless of population distinctions, observing an extended Y STR haplotype (12-17 loci) is a rare event in any of the major population groups routinely considered for forensic analyses. For examining the effect of population substructure, haplotype diversity is a very useful indicator. The haplotype diversities are greater than 0.99

(Kayser) and with the extended Y STR haplotypes (with the current commercially available kits) diversities tend to be greater than 0.999 for most populations. Other pragmatic justifications of pooling populations for Y-STR haplotype databases include the unspecified nature of the lineage of paternal (population) ancestry based on self-identification of individuals as member of a specified population.

Mixtures of Y STR Haplotype Profiles

The theory and logic (for use of the exclusion (PE) and/or the likelihood ratio (LR)) for assessing the significance of mixture samples are well-described (26,27 and references within). The PE provides an estimate of the portion of the population that has a genotype composed of at least one allele not observed in the mixed profile. The LR provides the odds ratio of two competing hypotheses, given the mixture evidence. There are two limitations, however, for Y STR mixture analyses that need to be addressed. These are: what frequencies to apply for all possible haplotypes that contribute to the mixture; and 2) user-friendly algorithms of interpretation of Y-STR evidence profiles.

Assume that there 2 alleles observed in a mixed sample at each locus of a 12 locus Y STR profile. Assuming independence across the loci (an obviously incorrect assumption), there can be 2^{12} possible haplotypes (4,096) that could possibly contribute to the mixed profile. Most of these possible profiles will not have been seen in a database and some of these possible haplotypes may not exist in the greater population. So for computing the PE (that is for a random male to be a (part) contributor of a DNA mixture) a minimal haplotype frequency (similar to the logic of a minimum allele frequency for autosomal loci (28)) is not practical. However, one can employ the same principles used

for the counting method described above and count the all contributing possible haplotypes that can be part contributors observed in a reference database(s), divide that by the total number of samples in the database(s), calculate the upper bound of the counted proportion and subtract the upper bound estimate from 1.

The PE evaluation does not use any information about the Y-STR profiles of the known persons tested or the combinations of profiles that can explain the observed evidence. Hence, the total strength of observations is not fully captured by the concept of PE. The likelihood ratio is a more informative approach for assessing the significance of a mixed evidentiary profile.

For addressing the LR only a two person mixture will be considered for simplicity. There are four general scenarios for a two contributor sample:

1. S1 and S2 are the source
2. S1 and an unknown male are the source
3. S2 and unknown male are the source
4. two unknown males are source

where S1 and S2 are identified, DNA-typed suspects. Since scenarios 2 and 3 are the simpler versions of the same calculations as for 1-PE (or the probability of inclusion) described above (with the modification that the unknown male must have alleles that explain all of the alleles in the mixture, taken in combination with the alleles in the profile of S1 or S2), only scenarios 1 and 4 are described below.

For simplicity, only three loci are shown for the LR. Assume two alleles reside at each of the three loci in a mixture. Two male suspects are identified and the combined Y STR haplotypes of the suspects explain all the alleles observed in the evidence (Table 1).

Table 1. Possible 3 locus haplotypes for fictitious mixed sample DYS319 - 13/15; DYS437 - 14/16; DYS389II - 28/30; S1 has the profile DYS319 - 13; DYS437 - 16; DYS389II - 30; S2 has the profile DYS319 - 15; DYS437 - 14; DYS389II – 28.

DYS319	DYS437	DYS389II	
13	14	28	--- haplotype 1
15	14	28	--- haplotype 2
13	14	30	--- haplotype 3
15	14	30	--- haplotype 4
13	16	28	--- haplotype 5
15	16	28	--- haplotype 6
13	16	30	--- haplotype 7
15	16	30	--- haplotype 8

Under the prosecution hypothesis (Hp), the probability is 1 of observing the evidence given the two suspects are the sources of the DNA in the mixed profile. In contrast, under the defense hypothesis (Hd), where two unknown individuals are the source of the evidence, the probability of the evidence is calculated by determining the proportion of the population of pairs of unrelated individuals drawn together at random whose combined profiles are the same as the evidence (note: alternate hypotheses can be generated; only one is described herein for illustration). Essentially, the denominator is the same as the probability of inclusion (1-PE), but with consideration of a prescribed number of contributors. If one considers the haplotypes in Table 1 that could not be excluded as possibly comprising a part of the evidence profile, not all possible pairwise

combinations of the 8 haplotypes could explain all the alleles observed in the mixture. By considering only those pairings the LR (for the 3 loci) is:

$$LR = 1/2[Pr(H_1)Pr(H_8) + Pr(H_2)Pr(H_7) + Pr(H_3)Pr(H_6) + Pr(H_4)Pr(H_5)]$$

The formulation for the LR is correct and has been described by others (see 29) but it is not particularly useful, because no direction is offered for what values (i.e., frequencies to be used for the haplotypes drawn randomly from a population) should be used in the denominator. Again the same principle can be applied as described above for the counting method. All pairwise combinations of profiles in a reference database(s) can be generated and then those pairs that match the evidence can be counted. Then an upper bound of the confidence interval of the proportion of counts over total pairs is calculated.

Conclusion

The Y STR loci currently available through commercial kits are highly informative. For statistical purposes the counting method (and sampling correction) is simple, conservative, and well-established for assessing the rarity of a profile. To correct for possible effects of substructure, gene diversity-based F_{ST} (treating the haplotypes as alleles of a single composite locus) is the more appropriate measure of θ to be used in a conditional probability calculation. Because Y STR haplotype data are treated similarly to that of single locus data and haplotype diversity is high, the F_{ST} values are very small for most populations. Thus, for current size databases, the upper bound estimate of the count proportion is more than adequate to compensate for uncertainty for most populations (excluding for example Native Americans). Lastly, the basic logic for

calculating Y STR mixtures is the same as for autosomal markers, i.e., the application of the PE and/or the LR. Appropriate frequencies can be estimated for the unknowns in the LR by using the same logic as that of the counting method. Software is under development to facilitate mixture calculations (manuscript in preparation).

Acknowledgment

This is publication number 07-11 of the Laboratory Division of the Federal Bureau of Investigation. Names of commercial manufacturers are provided for identification only, and inclusion does not imply endorsement by the Federal Bureau of Investigation.

References

1. Carracedo A, Beckmann A, Bengs A, Brinkmann B, Caglia A, Capelli C, Gill P, Gusmão L, Hagelberg C, Hohoff C, Hoste B, Kihlgren A, Kloosterman A, Myhre Dupuy B, Morling N, O'Donnell G, Parson W, Phillips C, Pouwels M, Scheithauer R, Schmitter H, Schneider PM, Schumm J, Skitsa I, Stradmann-Bellinghausen B, Stuart M, Syndercombe Court D, Vide C. Results of a collaborative study of the EDNAP group regarding the reproducibility and robustness of the Y-chromosome STRs DYS19, DYS389I and II, DYS390 and DYS393 in a PCR pentaplex format. Forensic Sci. Int. 119:28-41, 2001.

2. Gill P, Brenner C, Brinkman B, Budowle B, Carracedo A, Jobling MA, Knijff P de, Kayser M, Krawczak M, Mayr WR, Morling N, Olaisen B, Pascali V, Prinz M, Roewer L, Schneider PM, Sajantila A, Tyler-Smith C. DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. Int. J. Leg. Med. 114:305-309, 2000.

3. Honda K, Roewer L, de Knijff P. Male DNA typing from 25-year-old vaginal swabs using Y chromosomal STR polymorphisms in retrial request case. J. Forensic Sci. 44:868-872, 1999.

4. Jobling MA, Pandya A, Tayler-Smith C. The Y chromosome in forensic and paternity testing. Int. J. Legal Med. 110:118-124, 1997.

5. Kayser M, de Knijff P, Dieltjes P, Krawczak M, Nagy M, Zerjal T, Pandya A, Tyler-Smith C, Roewer L. Applications of microsatellite-based Y-chromosome haplotyping. Electrophoresis 18:1602-1607, 1997.

6. Kayser M, Kruger C, Nagy M, Geserick G, Roewer L. Y-Chromosomal experiences and recommendations; In Olaisen B, Brinkmann B, Lincoln PJ (Eds): Progress in Forensic Genetics 7. Elsevier Science: Amsterdam, The Netherlands; pp 494–496, 1998.

7. Prinz M, Boll K, Baum H, Shaler B. Multiplexing of Y-chromosome specific STRs and performance of mixed samples. Forensic Sci. Int. 85:209-218, 1997.

8. Redd AJ, Clifford SL, Stoneking M. Multiplex DNA typing of short-tandem-repeat loci on the Y-chromosome. Biol. Chem. 378:923-927, 1997.

9. Sinha SK, Budowle B, Arcot SA, Richey SL, Chakraborty R, Jones MD, Wojtkiewicz PW, Schoenbauer DA, Gross AM, Sinha SK, Shewale JG. Development and validation of a multiplexed Y-chromosome STR genotyping system, Y-PLEX™ 6, for forensic casework. J. Forensic Sci. 48:93-101, 2003.

10. Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, Gross AM, Gornall T, Frappier JRH, Eisenberg AJ, Barna C, Aranda XG, Adamowicz MS, Budowle B. Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. Forens. Sci. Int. 148(1):1-14, 2005.

11. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J. Forensic Sci. 51(1):64-75, 2006.

12. Budowle B, Planz JV, Campbell R, Eisenberg AJ.Molecular diagnostic applications in forensic science. In: Molecular Diagnostics, Patrinos G, Ansorge W, eds, Elsevier, Amsterdam, pp. 267-280, 2005.

13. National Research Council: The Evaluation of Forensic DNA Evidence; National Academy Press: Washington, DC, 1996.

14. Budowle B, Adamowicz M, Aranda X, Barna C, Chakraborty R, Eisenberg AJ, Frappier R, Gross AM, Lee HS, Milne S, Prinz M, Saldanha G, Krenke BE. Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America. Forens. Sci. Int. 150(1):1-15, 2005.

15. Budowle B, Sinha SK, Lee HS, Chakraborty R. Utility of Y-chromosome STR haplotypes in forensic applications. Forens. Sci. Rev. 15(2):153-164, 2003.

16. Snedecor GW and Cochran WG. Statistical Methods, 6th ed., Iowa State Univ. Press, Ames, Iowa, p.209, 1967.

17. Ricker WE. The concept of confidence or fiducial limits applied to the Poisson frequency distribution; J. Amer. Stat. Assoc. 32:349-356, 1937.

18. Exact Bionomial and Poisson Confidence Intervals at http://statpages.org/confint.html.

19. Sinha S, Budowle B, Chakraborty R, Paunovic A, Guidry RD, Larsen C, Lal A, Shaffer M, Pineda G, Sinha SK, Schneida E, Nasir H, Shewale JG. Utility of the Y-STR Y-PLEX™ 6 and Y-PLEX™ 5 in forensic casework and 11 Y-STR haplotype database for three major population groups in the United States. J. Forens. Sci. 49:691-700, 2004.

20. Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T, Szibor R, de Knijff P, Stoneking M, Sajantila A. Characterization and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Amer. J. Hum. Genet. 66:1580-1588, 2000.

21. Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. Estimating Y-chromosome specific microsatellite mutation frequencies using deep rooted pedigrees. Hum. Mol. Genet. 6:799-803, 1997.

22. Budowle B, Wilson MR, DiZinno JA, Stauffer C, Fasano MA, Holland MM, Monson KL. Mitochondrial DNA regions HVI and HVII population data. Forens. Sci. Int. 103:23-35, 1999.

23. Lessig R, Willuweit S, Krawczak M, Wu FC, Pu CE, Kim W, Henke L, Henke J, Miranda J, Hidding M, Benecke M, Schmitt C, Magno M, Calacal G, Delfin FC, de Ungria MC, Elias S, Augustin C, Tun Z, Honda K, Kayser M, Gusmao L, Amorim A, Alves , Hou Y, Keyser , Ludes B, Klintschar M, Immel UD, Reichenpfader B, Zaharova B, Roewer L. Asian online Y-STR haplotype reference database, Leg. Med. 5 (Suppl. 1): S160-S163, 2003.

24. Kayser M, Brauer S, Schädlich H, Prinz M, Batzer MA, Zimmerman PA, Boatin BA, Stoneking M. Y-Chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry. Genome Research 13:624-634, 2003.

25. Gusmão L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, Morling N, Prinz M, Roewer L, Tyler-Smith C, Schneider PM. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Int. J. Legal Med.. 120(4):191-200, 2006.

26. Devlin B. Forensic inference from genetic markers. Stat. Meth. Med. Res. 2:241–262, 1992.

27. Weir BS, Triggs CM, Starling L, Stowell LI, Walsh KA, Buckleton J. Interpreting DNA mixtures. J. Forensic Sci. 42(2):213-222, 1997.

28. Budowle B, Monson KL, Chakraborty R. Estimating minimum allele frequencies for DNA profile frequency estimates for PCR-based loci. Int. J. Leg. Med. 108:173-176, 1996.

29. Buckleton BS, Triggs CM, Walsh SJ. Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, p. 329, 2005.