# Analyzing Y-STR mixtures and calculating inclusion statistics

Rick W. Staub, Ph.D. and Cassie Johnson, M.S.
Orchid Cellmark Inc.
Dallas, TX

## ABSTRACT

Forensic evidence samples analyzed with Y-STR multiplexes frequently produce profiles that are mixtures of DNA derived from more than one individual. Sometimes these mixed profiles can be separated into major and minor components and haplotype counts can be produced for the individual components using Y-STR haplotype databases. Often, however, the mixed profiles cannot be separated into individual components and it is necessary to perform a count of all possible haplotypes for the mixed profile. In our laboratory, we have developed a method for analyzing these mixed profiles and accessing various databases to perform these counts. We have encountered a good number of forensic scientists who have difficulty accessing Y-STR haplotype population databases to derive counts of possible matching profiles. Thus, we will demonstrate the techniques we use in our laboratory by employing specific exemplary mixture profiles to explain the logic used. In the process of performing this demonstration, we will show how to directly access on-line data to obtain counts of all possible Y-STR haplotypes that can be found in a mixture profile, as well as how to develop a simple spreadsheet to accomplish the same end result. We will conclude by examining the conservative nature of statistical assessments made in this manner.

## INTRODUCTION

Y-STRs provide forensic scientists with a valuable tool to analyze samples which contain a mixture of large amounts of female DNA and very small amounts of male DNA. Utilization of autosomal STRs in these types of samples may result in successful amplification of only the female profile because it out-competes the male template in the amplification reaction. Because Y-STRs are only found in male genomes, their use allows the forensic scientist to obtain only the male profile, even when there are massive amounts of female DNA. In forensic evidence samples, it is not uncommon to encounter mixtures of cellular material from more than one male. Since single-source Y-STR samples typically possess only one allele per locus (unless duplications occur at a locus (1) [see Figure 1]), if a sample displays more than one allele at several Y-STR loci it is assumed to consist of a mixture of DNA from two or more males.

The standard autosomal STRs used in forensic casework have been shown to be genetically independent with respect to their distribution in populations (Budowle, Chakraborty, etc.). Autosomal STR databases of approximately 100 – 200 individuals suffice to determine allele frequencies and assess independence. Furthermore, the product rule can be used to arrive at a match probability and, when using the thirteen CODIS loci, random match probabilities are frequently obtained from single-source profiles that are low enough to enable forensic scientists to statistically determine, with a reasonable degree of scientific certainty, the identity of the DNA sample through comparison to known standards.(2)

The situation for Y-STR profiles is considerably different. Since all Y-STRs show linkage to the Y chromosome, the product rule cannot legitimately be used to determine the frequency of a Y-STR profile (haplotype). Instead, large databases of haplotypes must be maintained (typically by race or ethnic group) and these databases are searched for haplotypes that "match" the haplotype of interest. This is called the counting method. Obviously, the counting method fails to provide the capability of source attribution for Y-STRs. Their inheritance pattern alone would prohibit that since all male offspring of a father will share his Y-STR haplotype, barring mutation. Thus, Y-STRs typically pale in comparison to autosomal STRs statistically because it is difficult to build databases that are of ample size to reasonably estimate population frequencies for Y-STR haplotypes. Several on-line databases are available for use in performing haplotype frequency assessments, however. Those sites are listed below:

- Applied Biosystems' Yfiler database: www.appliedbiosystems.com/yfilerdatabase

- Promega's PowerPlex Y database: www.promega.com/techserv/tools/pplexy

- U.S. Y-STR database:  www.usystrdatabase.com

- Y-HRD database: www.yhrd.org

- Reliagene's Y-Plex database: http://www.reliagene.com/index.asp?menu_id=rd&content_id=y_frq

The software available for accessing these databases works well for assessing frequencies of single source profiles, but producing counts of included haplotypes for mixture profiles can be much more complex. We will describe three different sorts of mixtures that we have typically encountered in our casework and explain how we assess statistics in each case. These mixtures were all analyzed through amplification by AmpFℓSTR® Yfiler™ PCR Amplification Kits produced by Applied Biosystems. (3)

## YFILER MIXTURE WITH A MAJOR AND MINOR COMPONENT

An electropherogram of a mixture of DNA from two males (developed from the epithelial cell fraction of a boxer short cutting in a male-on-male sexual assault) is shown in Figure 2. One of the components of the mixture is present in significantly higher quantity than the other; however the peak ratios at each locus where they differ are highly disparate, varying from 2.8:1 at DYS437 to 16.6:1 at DYS456.  At any rate, it is still an easy matter to de-convolute the mixture into major and minor components. The two haplotypes can then be entered into the Yfiler Haplotype Database (www.appliedbiosystems.com/yfilerdatabase). When this is done, the complete haplotype obtained for the major component (DYS456=14, DYS389I=12, DYS390=23, DYS389II=30, DYS458=16, DYS19=14, DYS385a/b=14,14, DYS393=13, DYS391=10, DYS439=12, DYS635=22, DYS392=11, YGATAH4=11, DYS437=16, DYS438=10, DYS448=20) is unique in the Applied Biosystems database of 3,561 males.  The partial haplotype for loci at which the minor component actually possesses a different allele than the major (DYS456=17, DYS389I=14, DYS390=21, DYS389II=32, DYS458=17, DYS19=15, DYS385a/b=15,16, DYS393=14, DYS635=21, DYS437=14, DYS438=11) is also unique in the Applied Biosystems haplotype database.  Obviously, it would be reasonable to assume that the major and minor contributors share an allele where there is only one peak, but since the minor contributor may also possess a null allele at any of those loci or may have an allele that simply dropped out (failed to amplify), it is more conservative to omit those loci when determining the minor haplotype (in this case, a partial haplotype).

## YFILER MIXTURES WITH COMPONENTS THAT CANNOT BE DE-CONVOLUTED

When a mixture consists of two or more contributors that cannot be de-convoluted we treat the mixture statistics somewhat analogously to mixtures in autosomal STRs that cannot be de-convoluted into separate components.  In these cases, we compute an inclusion probability for all possible STR profiles in the mixture. However, determining all haplotypes that could be included in a Y-STR mixture profile is a bit more difficult than computing a statistic for autosomal STRs by formula. For this type of Y-STR mixture profile, one must be able to perform a count of all possible profiles in the Y-STR database which may have contributed to the mixture.  In order to do this with website-based databases, one must either compute all possible haplotypes and enter them individually into the website (a very time-consuming process if there are many loci with multiple peaks) or have the capability sorting all haplotypes in the database by locus for the necessary alleles.

For mixtures of this type, a technique that sometimes works to enable use of website data directly is to enter all the single-peak loci (assumedly sites at which contributors share alleles). An example of this is demonstrated by the mixture profile in Figure 3 (a Yfiler profile that is a mixture of at least two males produced from one side of a cutting from the finger of a latex glove found at the scene of a homicide). Any haplotype that is part of that mixed profile must have DYS389I=14, DYS458=18, DYS635=24, YGATAH4=11, and DYS437=15.  When these allele are entered into the Applied Biosystems Yfiler

Haplotype website database of 3,561 profiles no matching haplotypes are found. This obviates the necessity of entering all possible haplotypes into the website, since it has now been determined that there will be no profiles that match. Of course, one would expect that this will not happen very often, since most partial haplotypes that are shared by two or more individuals would be expected to be found in the database.

## YFILER MIXTURES FOUND IN SAMPLES WITH LOW LEVELS OF MALE DNA

Often, Y-STR mixed profiles are obtained in forensic casework from samples that have low levels of male DNA. In these samples, it is not uncommon to obtain complex mixtures of two or more males yielding partial profiles with some peaks below the laboratory's established threshold value. Figure 4 is an example of this type of mixed profile. It is a profile obtained from a cutting of sweatpants found at the scene of a murder. In our laboratory, to determine all haplotypes that would be included in a mixture of two or more males, we omit loci that clearly show below-threshold alleles. When this method is used for determining all haplotypes that would be included in the mixture, only those with DYS456=15, DYS391=10 or 11, YGATAH4=11 or 13, and DYS448=19 would be included in the count. When the previously explained technique of entering single-peak loci is used, the only loci that are usable are DYS456 and DYS448. When DYS456=15, and DYS448=19 are entered into the Applied Biosystems Yfiler Haplotype website database, 556 haplotypes are returned. If one clicks on this number and waits, all 556 haplotypes are downloaded in spreadsheet format and can be copied into a spreadsheet and sorted for haplotypes that have DYS391=10 or 11 and YGATAH4=11 or 13. However, the Applied Biosystems Yfiler Haplotype Database does contain 16 null alleles (DYS456=5, DYS389II=1, DYS458=1, DYS635=1, DYS392=2, and DYS448= 6) and 67 instances of duplicated alleles (DYS456=1, DYS389I=3, DYS390=2, DYS389II=7, DYS458=2, DYS19=10, DYS393=2, DYS391=4, DYS439=6, DYS635=3, DYS392=4, DYS437=2, DYS448=21) at non-duplicate loci (i.e., excluding DYS385a/b) among its 3,561 haplotypes. Since selecting the single alleles for DYS456 and DYS448 does not select haplotypes that have a null allele at those loci, the final count after sorting for DYS391=10 or 11 and YGATAH4=11 or 13 will be short any haplotypes that have any possible combination of those alleles, as well as a null allele at DYS456 or DYS448.

It is clear that the best way to avoid missing null alleles or duplicate alleles in database counts is to have access to the entire database of haplotypes with accompanying racial/ethnic data. Of the databases mentioned, the only one that presently provides the user with the capability of accessing all haplotype data to extract into a sortable spreadsheet to determine inclusion counts for mixtures is the Applied Biosystems Yfiler Haplotype Database. The entire database can be accessed by leaving asterisks in all data entry cells and mouse-clicking on "Search". Then, in the column labeled "# **Haplotypes** (with Selected Alleles)" counts of all 3,561 haplotypes arranged by racial group will appear in blue, underscored fonts that are accessible by the click of a mouse. When the total of 3561 is selected, a servlet downloads the entire database in a format that can be transferred to an Excel spreadsheet.

When we downloaded the entire database to an Excel spreadsheet and sorted it for the requisite alleles in the mixture shown in Figure 4 (DYS456=15, DYS391=10 or 11, YGATAH4=11 or 13, and DYS448=19), we obtained different database counts in Caucasians and African Americans from the results obtained using the previously described method in which single alleles are entered first and then the resultant 556 haplotypes are downloaded and sorted. The first method produces 65 Caucasians with matching haplotypes while the second produces 66 Caucasians (one with a null allele at DYS448). Additionally, the first method produces 49 African Americans with matching haplotypes while the second method produces 52 African Americans (three with a null allele at DYS456). Thus, entering data from only single peak markers is not recommended as a method for performing inclusion counts. It may not always be accurate due to the presence of null alleles and duplicate alleles in some haplotypes of the Applied Biosystems Yfiler Database.

### DEVELOPING MIXTURE CALCULATION SPREADSHEETS

In our laboratory at Orchid Cellmark, we have developed a spreadsheet that consists of two sorting worksheets with the entire Applied Biosystems Yfiler Haplotype Database.  These are sorted by two different analysts and the third worksheet compares the first two worksheets and compiles the data for entry into final case reports.  Analysts are trained to always include null alleles at all loci and to include duplicate profiles if both alleles are components of the mixture.  This produces the most conservative inclusion count for the mixture.

We are presently in the process of developing an automated spreadsheet that will use macros to perform the sorting operations in order to eliminate the necessity of utilizing two analysts to perform manual sorts of the Yfiler Haplotype Database.  An Excel spreadsheet of this type that has been prepared with the U.S. Y-STR database is reported in a poster presented at this meeting. (4)

## REFERENCES

1.    Johnson C.L., Warren J.H., Giles R.C., Staub R.W. Validation and uses of a Y-chromosome STR 10-plex for forensic and paternity laboratories.   J Forensic Sci 2003; 48(6):1260-1268.

2.    Budowle B., Chakraborty R., Carmody G., Monson K.L. Source attribution of a forensic DNA profile. Forensic Science Communications July 2000; Vol 2 No 3: 1-6 and 1-2. http://www.fbi.gov/hq/lab/fsc/backissu/july2000/source.htm.

3.    Mulero J.J., Chang C.W., Calandro L.M., Green R.L., Li Y., Johnson C.L., Hennessy L.K. Development and validation of the AmpFlSTR® Yfiler™ amplification kit: A male specific single amplification 17 Y-STR multiplex system. J Forensic Sci 2006; 51(1):64-75.

4.    MacMillan K., Gefrides, L., Klein C., Fatolitis l., Ballantyne J., Kahn R. A Y-STR mixture frequency estimator. Poster presentation at Nineteenth International Symposium on Human Identification; 2008 Oct 13-16; Hollywood (CA).
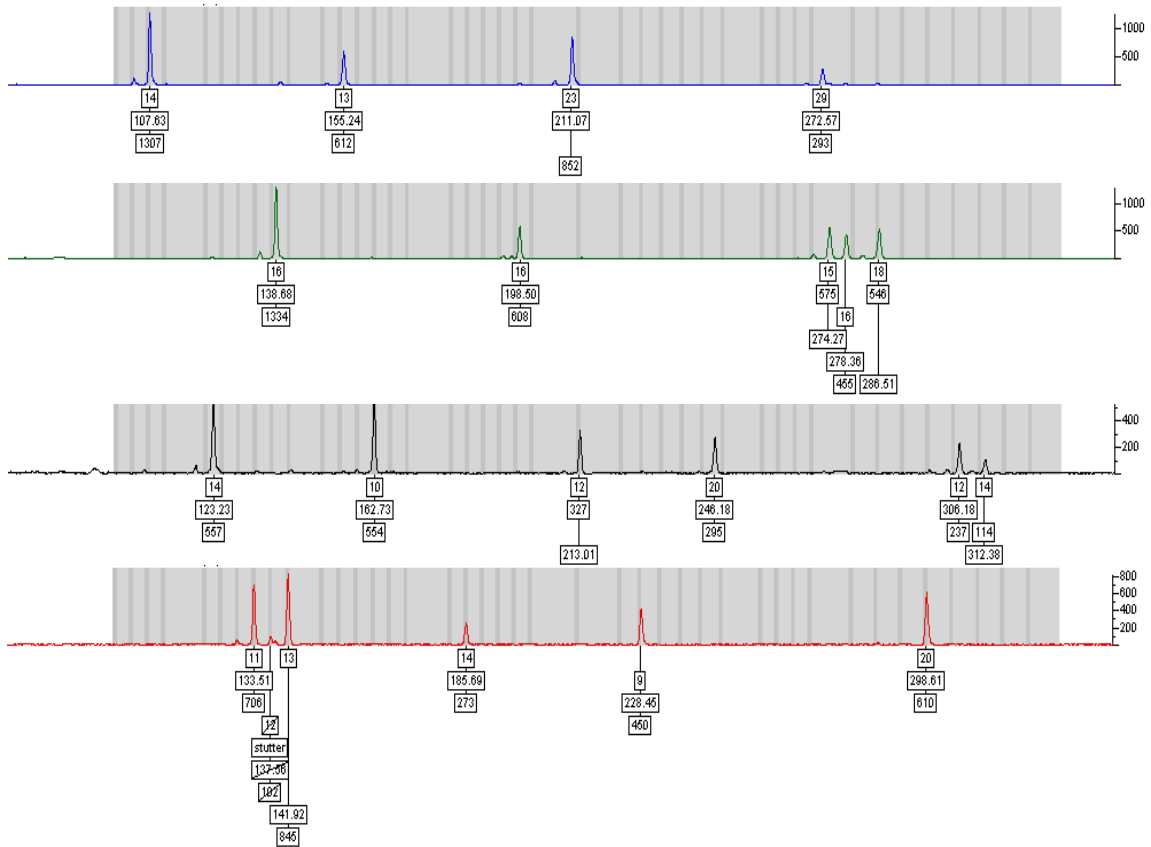
Figure 1.  Yfiler haplotype profile of a reference sample from a male with duplications at DYS385a/b, DYS392, and YGATAH4.  The profile is notable in that it mimics a mixture profile from a forensic evidence sample.
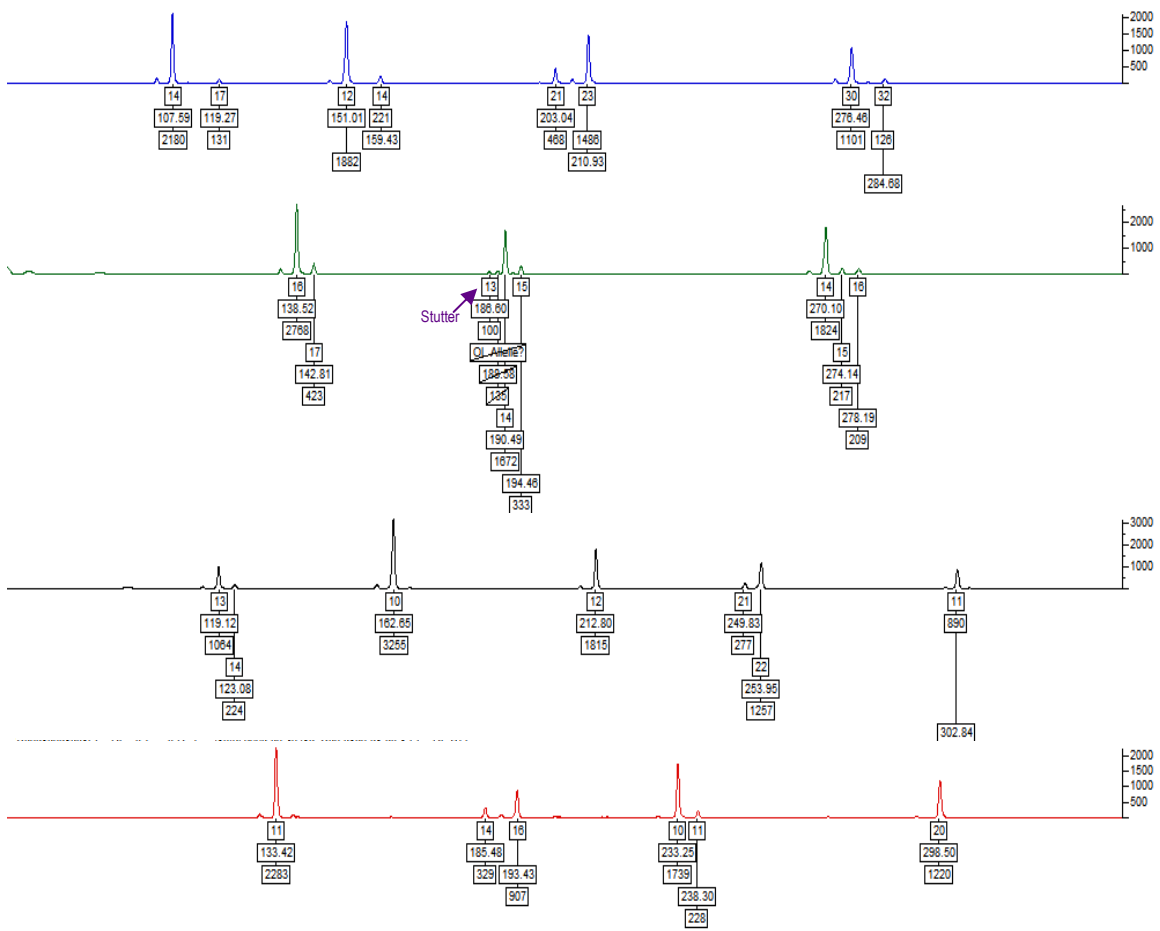
Figure 2. Yfiler profile of a mixture of two males in an epithelial cell fraction of a cutting from some boxer shorts in a male-on-male sexual assault. The mixed profile shows a major and a minor profile which were consistent with the profiles of the victim and suspected assailant, respectively.
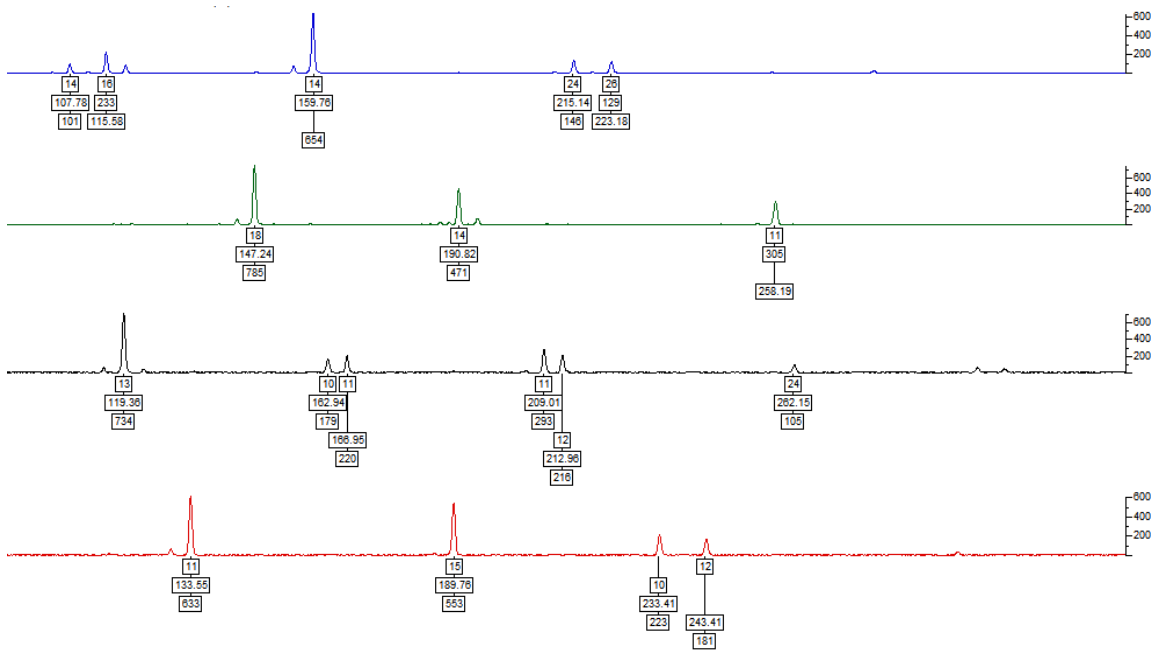
Figure 3. Yfiler profile of non-deconvolutable mixture of at least two males found on one side of a cutting from the finger of a latex glove found at the scene of a homicide. In this case the suspect could not be excluded and a count of all included haplotypes in the database was computed.
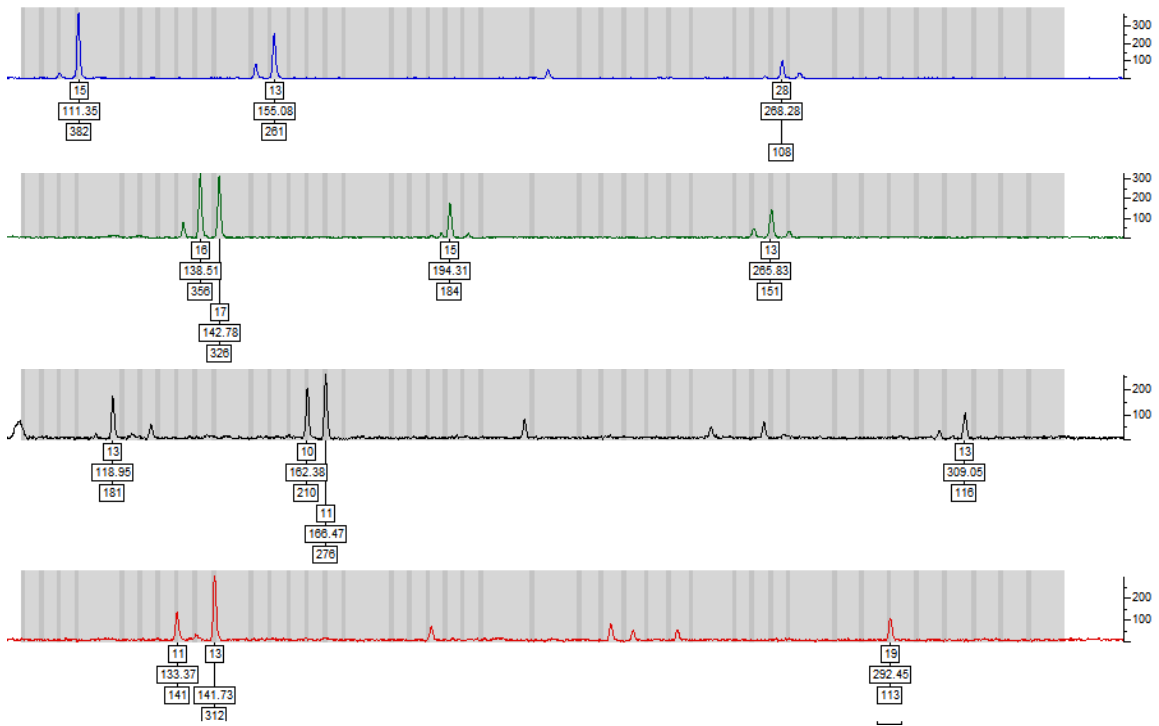
Figure 4. Yfiler profile of the epithelial cell fraction of a stain from some sweatpants found at the scene of a homicide. This mixture is exemplary of those often found in forensic evidence samples with reduced quantities of Y chromosome template DNA. In this case, the suspect could not be excluded as a possible contributor to the mixture. Following Orchid Cellmark protocols, the only loci that could be used to determine haplotype counts are DYS456=15, DYS391=10 or 11, YGATAH4=11 or 13, and DYS448=19.