IDENTIFICATION OF HIGHLY POLYMORPHIC Y-STRS BASED ON UNDERLYING SEQUENCE VARIATION

Maiko Takahashi[1], Jonathan L. King[1], Bruce Budowle[1,2]
[1]Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center
[2]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University

STRs on the Y chromosome have been widely used in forensic investigations. These markers are useful particularly in analyses of mixtures containing a small quantity of male DNA amidst a high quantity of female DNA, sexual assault cases, and kinship analyses. One limitation of Y-STRs in forensic applications compared with autosomal STRs is a reduced power of discrimination due to a lack of recombination throughout most of the Y-chromosome. The strength of a match of a Y-STR profile from an evidence sample with that of a suspect is evaluated by how frequently the Y-STR haplotype is observed in a database. To increase the strength of Y-STR evidence, two approaches can be considered: 1) increase the size of the database; and 2) improve the discrimination power of Y-STR typing.

Massively parallel sequencing (MPS) allows for detection of genetic variation at the sequence level as opposed to solely relying on length as is achieved by capillary electrophoresis. In the present study, a set of Y-STR markers demonstrating potential intra-allelic sequence variation within the repeat motifs and flanking regions was identified. STR data in the STR Catalog Viewer (http://strcat.teamerlich.org/) and SNP data in the UCSC Genome Browser (http://genome.ucsc.edu/index.html) were searched. In an initial screen, 45 candidate STRs were chosen from 7111 potential Y-STR loci in Phase 1 of the 1000 Genomes Project by visual examination using the STR Catalog Viewer. The following criteria were used: (1) repeat size of 3-6 bp, (2) average heterozygosity >0.6 for the four populations (African, Ad Mixed American, Asian and European) and (3) average heterozygosities of 0.2-0.6 for SNPs including those SNPs which have a minor allele frequency (MAF) of >0.3 within flanking regions up to 200 bases on either side of the STR repeat region from dbSNP build 146 by the UCSC Genome Browser. Also, the 23 commonly used Y-STRs and 15 rapid mutating Y-STRs (RM Y-STRs) reported previously [1] were added to the initial set (a few loci were in common). The total number of STR loci in the initial set is currently 77. Over 900 SNPs were identified in these data sets. These STRs are being screened (using vcftools software) to determine those that would contain increased diversity and be helpful for mixture DNA deconvolution. The initial genotype data are from 1000 Genomes Project.

[1] K.N. Ballantyne et al., Am. J. Hum. Genet. 87 (3) (2010) 341–353