DOCUMENTING HIGHLY INFORMATIVE MICROHAPLOTYPE LOCI FOR IMPLEMENTATION USING MPS

Kenneth Kidd[1], Willian Speed[1], Daniele Podini[2], Sharon Wootton[3], Joseph Chang[3], Robert Lagace[3]

[1] Department of Genetics, Yale University
[2] Department of Forensic Sciences, The George Washington University
[3] Thermo Fisher Scientific

In 2013 we proposed microhaplotypes as a new type of forensic marker designed for study by massively parallel sequencing (MPS). MPS yields phase-known haplotypes of multiple SNPs that define three or more alleles within a read length of up to 300 bp. Data on 31 statistically phased microhaplotype loci evaluated on 54 populations was published in 2014. In 2015 we posted on our web site the definitions and statistics for 129 loci evaluated on 55 populations. We have now defined 132 microhaplotype loci (362 SNPs) evaluated on 83 populations (5100 individuals) using TaqMan assays and statistical phasing.

Microhaplotype loci can be used for individual identification with very low random match probabilities. Moreover, microhaplotype loci have highly significant advantages over forensic short tandem repeat polymorphisms (STRPs): improved ability to resolve mixtures (no stutter bands), improved ability to identify more distant relationships (lower mutation rates), and better inference of biogeographic ancestry (some loci have large global variation). To evaluate loci for mixture detection we have used the concept of effective number of alleles, Ae. To evaluate loci for ancestry inference we have used Rosenberg's Informativeness, In. Given the current importance in forensics of mixture deconvolution, we have concentrated on identifying loci with high Ae but some of the loci are also useful for ancestry inference at resolution of seven biogeographic regions. One advantage of Ae is that it is also a rough measure of the ability to infer relationships because of the multiple alleles.

Currently, over 70% of the loci evaluated on the 83 populations have a global average Ae >2.0, making them more heterozygous, on average, than is possible for a simple diallelic locus, i.e., a SNP or Indel. More significant is that 28 of the loci have a global average Ae >3 making them very useful for mixture deconvolution. Given the distribution of the Ae values for those 28 loci, (two have an Ae >5) the probability of *not* detecting a mixture of two or more random individuals using these 28 loci is less than 10e-7. While these are global average values, we are including many populations that are small and more isolated; even those have considerable heterozygosity for most of these loci.

The question remaining is translation of our results into practice through implementation of these loci into multiplex kits for MPS. That has been done for several of the loci and mixtures are detectable with sufficient sensitivity to detect low levels of the minor component. That work will be discussed.