ISHI25 Poster #103: Microhaps: a powerful new class of forensic markers

Kenneth K. Kidd[a], Andrew J. Pakstis[a], William C. Speed[a], Robert Lagacé[b], Joseph Chang[b], Sharon Wootton[b], Eva Haigh[a], Judith R. Kidd[a]

(a) Department of Genetics, Yale University School of Medicine, New Haven, CT 06520 USA

(b) Human Identification Group, Thermo Fisher Scientific, 180 Oyster Point Blvd. South San Francisco, CA 94080 USA

ABSTRACT

Forensic uses of DNA include individual identification, inference of ancestry, inference of phenotype, and, in mass disaster and missing person cases, linking unknowns to a family. We are proposing that haplotypes of SNPs can become the markers of choice for all such tasks. Microhaplotypes (two or more SNPs within a span of <200bp) are a powerful new type of multiallelic forensic marker that can be genotyped by NGS (the massively parallel "Now" Generation Sequencing) since 200bp is within the span of current desktop sequencing platforms. In addition, because they are multiallelic and co-dominant when typed by sequencing, microhaplotypes can detect and deconvolute mixtures in a sample, a very common problem encountered in crime scene samples.

Population genetics globally of over 62 microhaplotype loci selected to define multiple alleles has now been fully documented in over 2500 individuals from at least 54 different populations. All have multiple alleles (haplotypes) and almost all have global average heterozygosities >0.5 (many over 0.6) and over 74% of individual heterozygosities are greater than the 0.5 maximum possible for any single SNP. Ongoing research on 19 selected microhaplotypes comprised of 4 SNPs is yielding nearly 97% of heterozygosities >0.5 and a global mean of over 0.7. These high heterozygosities make microhaplotypes a highly efficient type of forensic marker for typing by sequencing. For individual identification large numbers of microhaplotype loci can be multiplexed at affordable costs allowing high statistical power, much greater than even the expanded set of CODIS STR polymorphisms. The microhaplotypes we have fully characterized not only give low match probabilities but also allow ancestry inference. We have easily identified in HapMap and 1000 Genomes an additional several dozen such loci now being studied on our population samples. The phase-known data from a forensic sample will detect mixtures for these multialleleic loci allowing deconvolution not possible with single SNPs. Thus, microhaplotype loci constitute a statistically powerful new type of genetic marker ready for forensic applications using existing sequencing methods.

The value of marker typing by sequencing

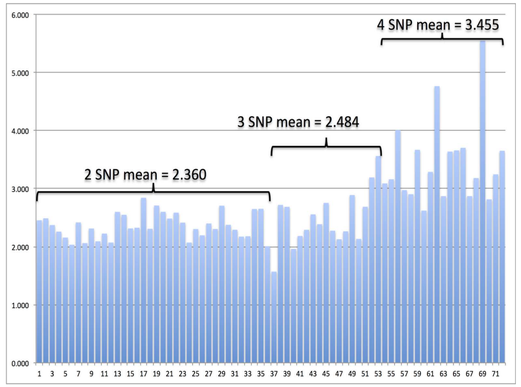
Marker Type	Method		
	electrophoresis	chip typing	sequencing
STRP	yes	no	yes (smaller)
Indel	yes	limited	yes
SNP	limited	yes	yes
Microhap	no	no	yes
Cryptic variation	no	no	yes

As the table shows, sequencing is the only method that allows all types of markers to be typed. As such, it is the logical single method for future use in forensic laboratories.

Cryptic variation around Microhap048 (see figure)

rs number (build 138)	Microhap role	Position (nt) chr14 GRCh37/hg19	clustered allele frequency
rs149195448		74,250,553	0.006
rs12717560	SNP 1	74,250,557	0.331
rs76446474		74,250,562	0.005
rs374425620		74,250,591	n/a
rs191001036		74,250,647	0.001
rs113480934		74,250,694	n/a
rs12878166	SNP 2	74,250,715	0.377
rs12879393		74,250,730	0.286

Effective number of alleles for 72 microhaps



Five most heterozygous 3-SNP microhaps

Locus	Avg. Het	The SNPs
MicroHap061	0.687	rs763040; rs5764924; rs763041
MicroHap049	0.654	rs9937467; rs17670098; rs17670111
MicroHap033	0.637	rs10815466; rs9408671; rs17431629
MicroHap017	0.632	rs4699748; rs2584461; rs1442492
MicroHap058	0.628	rs6122890; rs6095836; rs6012881

Five most heterozygous 4-SNP microhaps ("microtetrads")

Locus	Avg.Het	The SNPs
MicroTet315	0.790	rs8126597; rs6517970; rs8131148; rs6517971
MicroTet180	0.750	rs12802112; rs28631755; rs7112918; rs4752777
MicroTet017	0.730	rs17413714; rs2772234; rs1610401; rs1610400
MicroTet223	0.727	rs1192204; rs1192205; rs3825483; rs3825481
MicroTet324	0.725	rs6518223; rs2838868; rs7279250; rs8133697

MIXTURE DETECTION

A mixture is qualitatively detected by observing three or more alleles at a locus in the analysis of a sample. This is no different from mixture detection using STRPs. Mixture detection requires a locus with three or more alleles; such qualitative data are impossible with a diallelic locus. As more heterozygous loci with more alleles are used, the probability of identifying a mixture rapidly approaches certainty assuming the sensitivity is great enough. With high numbers of sequence reads, the sensitivity will be quite high.

Cumulative probability of a mixture having three or more alleles at two or more loci

Effective # alleles	Number of Loci Studied			
	2	3	4	5
3	0.69131	0.82849	0.90471	0.94706
4	0.88184	0.95938	0.98604	0.99520
5	0.94618	0.98751	0.99710	0.99933

PROGRESS

As the number of characterized microhaps has increased from 31 (Kidd et al., 2014) to 72 (analyses underway), the match probabilities have continued to become smaller, the ancestry resolution has increased, the ability to identify close relatives has increased, and the ability of microhaplotypes to qualitatively identify mixtures in a DNA sample has been demonstrated. A further advantage of using sequencing is detection of cryptic variation.

INDIVIDUAL IDENTIFICATION

The random match probabilities vary considerably among populations with the existing markers in contrast to our earlier IISNP panel. Nonetheless, the values ranged from 10^{-13} to 10^{-20} for our first 31 microhaplotypes (Kidd et al., 2014) most of which were only 2-SNP loci. However, as the number of loci and their average heterozygosity have increased, we are able to select markers that will show less variation globally. Microhap061 illustrates one with less global variation in frequencies.

ANCESTRY INFERENCE

The original 31 microhaps defined from five global clusters to six global clusters. With the new markers the global pattern is now approaching eight clusters. An important aspect of the ancestry inference results is that these markers were selected for high global heterozygosity, not for high global variation. This automatically biases against markers that are good for ancestry inference. However, as can be seen for the microhaps illustrated, several contain considerable variation in haplotype frequencies.

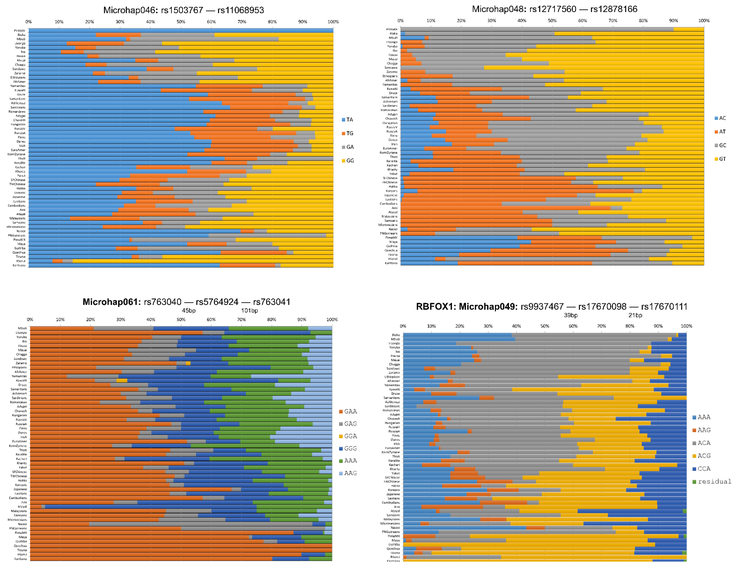
PHENOTYPE INFERENCE

As yet there are no characterized microhaplotypes for phenotype informative SNPs.

LINEAGE INFERENCE

As we have identified more complex microhaplotypes with three to five effective alleles, we are also identifying many relatively infrequent alleles. Collectively these will greatly improve the ability to identify close relatives.

Examples of Microhap Population Frequencies



SNP Database Resources

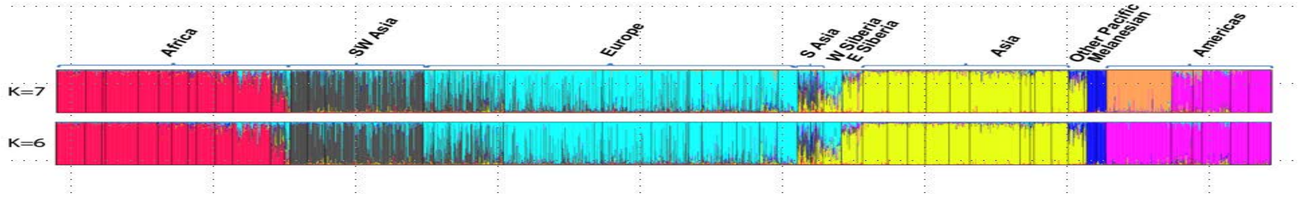
ALFRED: the ALlele FREquency Database http://alfred.med.yale.edu

ALFRED is making available allele frequencies for forensic applications of many SNP panels. Currently ALFRED has over 37 million allele frequency tables involving several hundred populations.

FROG-kb: Forensic Resource/Reference on Genetics--knowledge base http://frog.med.yale.edu

FROG-kb has data for 11 different forensic SNP panels allowing calculations of probabilities for individual SNP profiles based on the populations for which allele frequencies are available

STRUCTURE ANALYSES OF 62 MICROHAPS



REFERENCES

K.K. Kidd, J.R. Kidd, W.C.Speed, R.Fang, M.R. Furtado, F.C. Hyland, A.J. Pakstis, 2012. Expanding data and resources for forensic use of SNPs in individual identification. *Forensic Science International:Genetics* 6:646-652

K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagace, J. Chang, S. Wootton, N. Ihuegbu, 2013. Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series* 4:e123-e124.

K.K. Kidd, A.J. Pakstis, Wi.C. Speed, R.Lagace, J. Chang, S. Wootton, E. Haigh, J.R. Kidd, 2014. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. *Forensic Science International: Genetics* 12:215-224.

Pakstis A.J., W.C. Speed, R. Fang, F.C.L. Hyland, M.R. Furtado, J.R. Kidd, K.K. Kidd. 2010. SNPs for a universal identification panel. *Human Genetics* 127:315-324.

ACKNOWLEDGMENTS

This work was funded primarily by NIJ Grants 2013-DN-BX-K023, 2010-DN-BX-K225, and 2007-DN-BX-K197 to KKK awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice. Assembly of the population resource was funded by several NIH grants over many years. We thank the many collaborating researchers who helped assemble the samples from diverse populations and the many graduate students and postdocs who collected SNP data on these samples. Special thanks are due to the many hundreds of individuals who volunteered to give blood samples for studies of gene frequency variation. ALFRED is supported by NSF grant BCS1444279.