30th CONGRESS OF THE INTERNATIONAL SOCIETY FOR FORENSIC GENETICS
SANTIAGO DE COMPOSTELA
9-13 SEPTEMBER 2024

POSTER NR.
442

Interpreting Microhaplotype Data

Brian Young[1,2,3], Joop Theelen[4], Daniele Podini[5]
[1]NicheVision Forensics, [2]Florida International University, [3]Syracuse University,
[4]NimaGen, [5]George Washington University

BACKGROUND

Microhaplotypes are polymorphic markers covering two or more single nucleotide polymorphisms (SNPs) selected for their lack of historical recombination (1). The SNPs selected for inclusion are in close enough proximity such that they can be covered by single sequencer reads (e.g., 2x300 sequencing). The individual SNP states are automatically phased by sequencing (2). Microhaplotype alleles are variant sequences defined by their distinct set of SNP state combinations. Microhaplotype alleles have the advantage of reduced stutter under PCR amplification due to a lack of short tandem repeat (STR) loci. However, microhaplotype assays are not completely free of artifacts. Stutter artifacts can be generated when microhaplotypes cover homopolymer stretches, and PCR misincorporation or sequencing errors can occur. To make microhaplotypes useable in forensic practice, some of the same challenges faced by sequence-based STR alleles must be addressed. This includes: 1) useful nomenclatures, 2) software for easy analysis, 3) microhaplotype allele frequency databases, 4) multiplex panels validated for identity or ancestry, and 5) commercially available PCR kits. Here, we describe advances toward addressing some of those challenges.

METHODS

Samples were prepared for sequencing using the IDseek® OmniHAP™ 29plex kit (NimaGen) and sequenced on a MiSeq instrument using a 2x250 protocol. Partially-overlapping mate-pair reads were joined using FLASh (3), and joined reads were analyzed using MixtureAce software (NicheVision). All samples were collected under informed consent. Here we utilize the following terms: a) **CANONICAL SNPs**: the set of SNPs identified as defining the microhaplotype, b) **VARIANTS**: non-reference positions that include non-reference canonical variants, and variants which may or may not be polymorphic, c) **DISTINCT SEQUENCE TYPE (DST)**: the nucleotide sequence of an examined region to include both allelic and artifactual sequences, d) **EXAMINED REGION**: the genomic extent covered by the defined haplotype, e) **DEFINED HAPLOTYPE**: the genomic region between the outermost canonical SNPs inclusive.

FINDINGS

Nomenclature

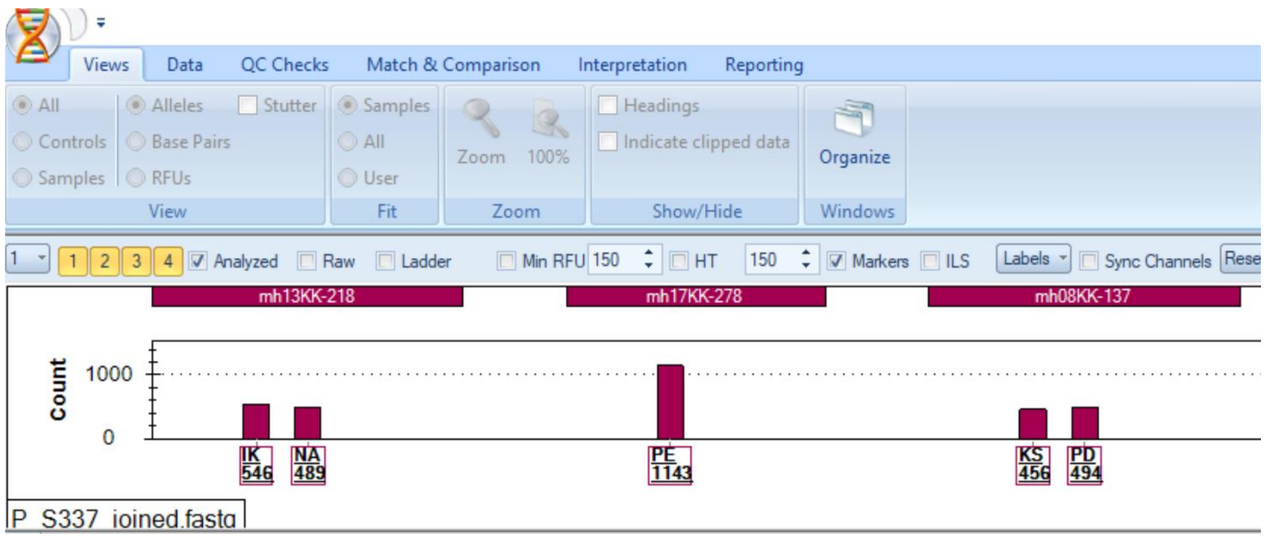
The canonical SNPs selected to define the microhaplotype alleles contain only part of the variability inherent to amplified segments. Haplotypes defined by the canonical SNPs can be used for profile matching, databasing, and allele frequency calculations. However, SNPs occurring at other positions within reads may be valuable for investigative purposes or exclusions. Multiple nomenclatures may be necessary in forensic casework to describe observed DSTs (**Figure 1**).

Long Microhaplotypes and/or Short Sequencing

Kit primer placement may cause PCR amplicons to be longer than the sequencing cycles. Read joining allows recovery of microhaplotype alleles from amplicons too long to be completely covered by both the R1 and R2 reads. Construction of correct haplotypes was validated by separate alignment of R1 and R2 mate pairs (data not shown).

Software

Microhaplotypes can be analyzed like sequence-based STRs, which are SNP-STR haplotypes. Using the SID short designator, and STR-like displays, profiles can easily be visualized and evaluated. **Figure 1** displays three loci from a single source sample. Mixed samples can be displayed in a similar manner.



MixtureAce Sample Report

Marker	DST	Allele	Count	Canonical SNPs	Variants
mh02KK-134v2	ZQ	ZQ	485	ACTCACCA	rs3101043 T>C |rs3111398 C>T |rs72623112 G>A
mh02KK-134v2	PN	PN	485	ACTCACCG	rs3101043 T>C |rs3111398 C>T
mh02KK-014v2	HP	HP	697	GCTAGCCGTCAAG	rs4270334 T>C |rs4332915 G>A
mh20KK-306	SY	SY	587	CGGCTGT	rs11697918 A>G |rs533194 A>G |rs1014897 C>T
mh20KK-306	FT	FT	543	CAACCGT	

Figure 1. (Top) Sample analysis report showing multiple nomenclatures for the same sequences. (Bottom) Bar chart display of microhaplotype alleles and their read count intensities. Alleles are identified by SID labels (4).

Allele Frequencies

To find suitable PCR primer sites, kit manufacturers may choose to include only a subset of the SNPs that officially define the haplotype (**Table 1**). If this occurs, matching between kit-derived alleles and alleles in the relevant allele frequency database will be restricted to the overlapping region. Frequencies for kit-derived microhaplotype alleles can still be derived from the subset region by the method previously described for sequence-based STRs (5).

Table 1. Illustration of deriving microhaplotype allele frequencies at the mh02KK-014v2 locus using the NimaGen 29-plex which excludes three canonical SNP positions. Frequencies used are hypothetical.

Microhaplotype Allele		Freq.	Counts
Canonical SNPs	SID		
GTTAGTCGTCAGGAGT	PG	0.500	50
GTTAGTCGTCAGGGGT	OA	0.200	20
GCTAGCCGTCGGGAGT	MN	0.050	5
GTTAGCCATCAGGAGT	CK	0.100	10
GTTAGCCATCAGGAGC	RO	0.150	15
	Sum	**1.000**	**100**

Microhaplotype Allele		Freq.	Counts
Canonical SNPs	SID		
GTTAGTCGTCAGG---	XW	0.700	50
GTTAGTCGTCAGG---		--	20
GCTAGCCGTCGGG---	VV	0.050	5
GTTAGCCATCAGG---	ZT	0.250	10
GTTAGCCATCAGG---		--	15
	Sum	**1.000**	**100**

CONCLUSIONS

Lessons learned in sequence-based STR markers can be applied to microhaplotype markers to successfully address nomenclature, software and allele frequencies.

ACKNOWLEDGEMENT
Support was provided by NicheVision (MixtureAce software) and NimaGen (OmniHap kit).

CONTACT brian@nichevision.com

REFERENCES

(1) Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum Genet. 2002;71(5):1227-34.
(2) Kidd KK, Pakstis AJ, Speed WC, Lagace R, Chang J, Wootton S, et al. Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics. Forensic Sci Int Genet. 2014;12:215-24.
(3) Magoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957-63.
(4) Young B, Faris T, Armogida L. A nomenclature for sequence-based forensic DNA analysis. Forensic Sci Int Genet. 2019;42:14-20.
(5) Young B, Marciano M, Crenshaw K, Duncan G, Armogida L, McCord B. Match statistics for sequence-based alleles in profiles from forensic PCR-mps kits. Electrophoresis. 2021;42(6):756-65.