Visual Microhap: A custom haplotype caller to analyze sequence-based data of microhaplotypes

Ye-Lim Kwon[1,2] · Su Min Joo[1,2] · Kyoung-Jin Shin[1,2]

[1]*Department of Forensic Medicine, Yonsei University College of Medicine, Seoul, Korea*
[2]*Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University, Seoul, Korea*

Introduction

Microhaplotypes (microhaps) are promising forensic marker to be used for multipurpose, such as human identification, kinship testing, and biogeographic ancestry inference. Through introduction of massively parallel sequencing (MPS) technology in forensic genetics, MPS studies on microhaps are actively being reported. Sequence-based microhaps data have been usually analyzed with the Genome Analysis Toolkit (GATK) pipeline or Ion Torrent Suite Software, but there is a desire to straightforwardly analyze MPS data. Therefore, we developed a web-based open-source haplotype caller, ***Visual Microhap*** *(http://forensic.yonsei.ac.kr/VisualMH/index.html)*, to extract SNP-based haplotypes from sequence-based data obtained by STRait Razor 3.0 and applied this tool to analysis of in-house 56 microhaps panel.

Materials and Methods

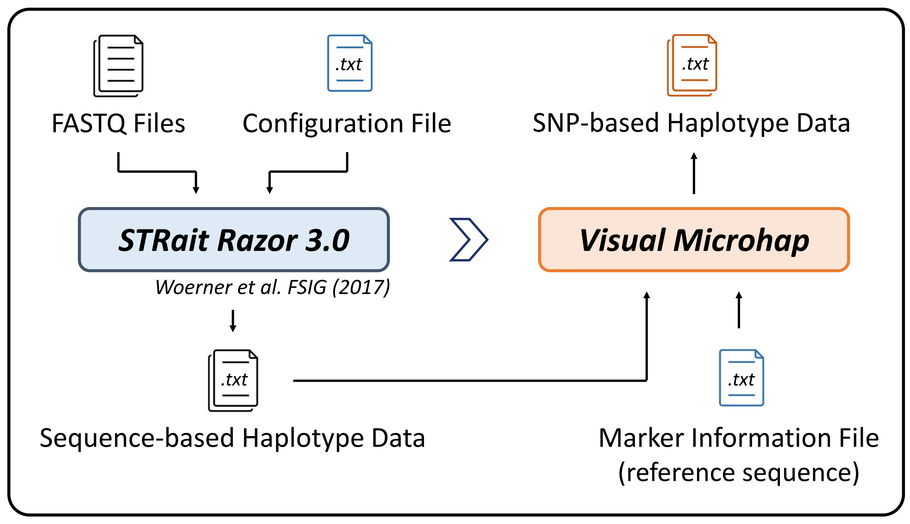
MPS library preparation using in-house 56 microhaps panel

- *Marker*: 56 microhaps (top 51 of A_e and top 25 of I_n) among the 130 microhaps reported in Kidd's study [1].
- *In-house MPS panel*: simultaneously amplified the 56 microhaps through a two-step PCR method.
- MPS library was prepared for 1ng of 2800M using an in-house 56 microhaps panel and sequenced on the MiSeq system (Illumina) after purification and quantification of library.

Microhap MPS data analysis using open-source tools

- Genotype of 56 microhaps was determined through two open-source tools (Fig 1).
  - > 1st step: ***STRait Razor 3.0*** [2] extracted *Sequence-based Haplotype Data* from FASTQ files by matching both the leading and trailing anchor sequences designated by *Configuration File*.
  - > 2nd step: ***Visual Microhap*** detected and annotated variations from the *Sequence-based Haplotype Data* by comparing it with the reference sequence included in the *Marker Information File*.

Fig. 1. Sequence-based microhap data analysis workflow



Results

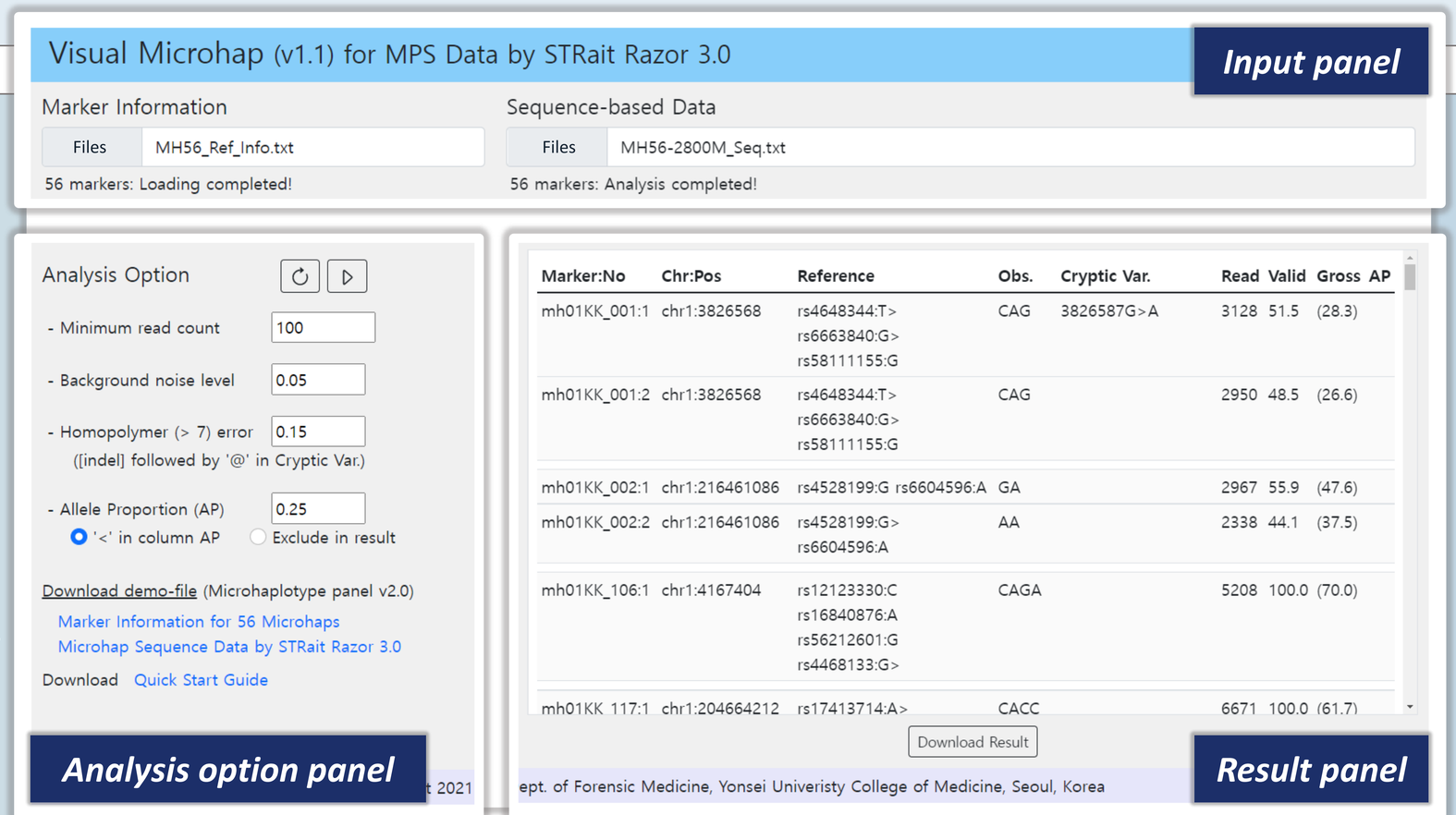
- The following is a screenshot of ***Visual Microhap*** coded in Vanilla JavaScript with Bootstrap v5.0 for graphic user interface. Visual Microhap consists of 3 panels (Input, Analysis option, and Result panels), and the detailed description is as follows.

(1) Marker Information File

A text file composed of the marker's name and chromosome number, position for the target SNPs, the start position of the reference sequence, and the reference sequence corresponding to the result of STRait Razor.

(2) Sequence-based Haplotype Data

The result file of STRait Razor 3.0 containing unique phase-known haplotype sequences and their length and coverage observed for each microhap.



(1) Minimum read count

Filters out reads below the designated number.

→ provide the coverage necessary to obtain reliable reads.

(2) Background noise level

Excludes reads with under the assigned proportion over gross read.

(3) Homopolymer (>7) error

Marks '@' in the column *Cryptic Var.* if a proportion of reads with homopolymer indels is above the threshold.

(4) Allele proportion (AP)

Marks '<' in the column *AP* if a proportion of reads is below the threshold among observed haplotypes in each marker.

→ to verify the read balance among haplotypes

: SNP-based Haplotype Data

- **Chr:Pos**: the position of the first target SNP in each microhap
- **Reference**: rs number of target SNPs and genotype of reference sequence
- **Obs.**: observed genotype of target SNPs
- **Cryptic Var.**: additional variations
- **Valid/Gross**: allele proportion for reads above *'Background noise level'* and *'Minimum read count'*, respectively

Homopolymer indels were observed only in mh11KK-191 among 56 microhaps and cautious analysis is required because they produce reads with various lengths as below.

Marker:No	Chr:Pos	Ref.	Obs.	Cryptic Var.	Read	Valid	Gross	AP
mh11KK-191:1	chr11:100009431	TAAT	TAAC	100009452-[A]@	3862	76.3	(41.7)	
mh11KK-191:2	chr11:100009431	TAAT	TAAC	100009452-[AA]@	1198	23.7	(12.9)	<

Conclusion

- We established an open-source workflow consisting of ***STRait Razor and Visual Microhap*** to analyze sequence-based microhap data.
- ***Visual Microhap***, a custom haplotype caller, extracted SNP-based haplotypes from the sequence-based data of STRait Razor through four analysis options.
- The custom data analysis pipeline could facilitate MPS analysis of microhaps in forensics by producing tailored results for microhaps compared to the conventional GATK method.

Visual Microhap

Acknowledgement

This study was supported by a faculty research grant of Yonsei University College of Medicine (2021), Republic of Korea.

Reference

[1] K.K. Kidd, W.C. Speed, A.J. Pakstis, D.S. Podini, R. Lagacé, J. Chang, S. Wootton, E. Haigh, U. Soundararajan, Evaluating 130 microhaplotypes across a global set of 83 populations, Forensic Sci. Int. Genet. 29 (2017) 29-37.
[2] A.E. Woerner, J.L. King, B. Budowle, Fast STR allele identification with STRait Razor 3.0, Forensic Sci. Int. Genet. 30 (2017) 18-23.