Forensic Science International: Genetics Supplement Series 8 (2022) 79–81

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

journal homepage: www.elsevier.com/locate/fsigss

Introduction of the python script MHinNGS for analysis of microhaplotypes

Carina G. Jønck, Claus Børsting *

Section of Forensic Genetics, Department of Forensic Medicine, University of Copenhagen, Denmark

ARTICLE INFO

Keywords:
Microhaplotype
SNP
Indel
Next generation sequencing
Data analysis
MHinNGS

ABSTRACT

MHinNGS is a Python application developed for analysis of microhaplotypes (MHs) in single-end sequencing data. MHinNGS analyses reads in standard formats and store each sequence into bins, one bin for each MH as defined by the two flanking sequences. MHinNGS requires a reference genome and a configuration file with information about each locus. Four mandatory and 15 optional criteria defined in the configuration file allow detailed locus-specific analyses of the MH loci. The program 1) removes noise, 2) identify and name alleles, 3) test the genotypes, and 4) test unique sequences not identified as noise or alleles. MHinNGS produces a result file, where every unique sequence that passed the noise filter is presented with MH allele, read depth, warning flags based on the genotyping criteria, sequence, heterozygote balance, and MH name. Furthermore, variation in other parts of the fragment that is not defined as SNPs in the MH, linked variants, or rare SNPs are listed in a separate column of the result file.

1. Introduction

Microhaplotypes (MHs) consist of two or more polymorphic loci (typically SNPs or small indels) within a short stretch of DNA (typically 2–300 nucleotides) [1]. The relative short distances between the variants allow for efficient PCR amplification and sequencing of the entire amplicon, which makes PCR-NGS assays targeting MH loci highly sensitive and potentially interesting for forensic genetic applications [2,3].

MHs have three important advantages compared to the standard STR loci used in forensic genetics: 1) Amplification of MHs do not generate stutter artefacts, that complicates data analysis of mixture samples [4]. 2) The mutation rates of MHs are 4–6 orders of magnitude lower than the mutation rates of STRs [1], which is particularly important for relationship testing. 3) The amplicon lengths of the different MH alleles are the same. This prevents NGS read count variation due to differently sized alleles, which is observed for most STRs [5–7] and may be a problem in the analysis of highly degraded samples.

2. Materials and methods

MHinNGS is a freely available python script (https://hub.docker.com/r/bioinformatician/mhinngs) developed for analysis of MHs in single-end sequencing data. MHinNGS is built upon the program STRinNGS v2.0 [8], that is used for analysis of STR sequences, and they have many similar features. MHinNGS needs three input files: 1) One file or folder containing the reads (FASTQ, FASTA, BAM, SAM, or CRAM format), 2) A reference genome in FASTA format, and 3) A configuration file containing information about each locus. The configuration file has five mandatory elements and 15 optional criteria (Table 1).

MHinNGS output consist of three files: 1) A log file containing various information about the run such as program version, input files, and parameter settings. 2) A result file in csv format with filtered data and all comments, and 3) A file named raw_results in csv format that contains all data including noise sequences, but without allele name, comments, and heterozygote balance.

* Correspondence to: Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Frederik V's Vej 11, DK-2100 Copenhagen, Denmark.
E-mail address: claus.boersting@sund.ku.dk (C. Børsting).

https://doi.org/10.1016/j.fsigss.2022.09.029
Received 12 September 2022; Accepted 28 September 2022
Available online 29 September 2022

Table 1
Criteria and settings.

Settings	Example	Possible flag(s)
Mandatory information		
locus name	mh11KK-180[a]	
chromosome	chr11	
start	1669536	
stop	1669769	
mh_info	rs12802112:1669561:GAT*AGA:AG rs12360952:1669657:GG*GAC:TC rs28631755:1669681:GGG*TTT:AC rs4752778:1669720:ATG*TAA:CT rs7112918:1669739:CCA*GAA:CT rs4752777:1669754:TGA*GCA:CG	"Unexpected microhaplotype"
Optional criteria		
ignore_pos	1669594.1 G,1669676	
rare_snp	rs117851656:1669542:T	"Rare SNP"
linked_allele	1669540:ACCTTG:A	"Linked allele not linked"
Default criteria		
flank_up_length	-15	
flank_down_length	-15	
mism_up	1	
mism_down	1	
noise_filter (>=)	0.01	
min_reads (>=)	100	"Too few reads" / "Locus Dropout"
max_num_unique	4	"Many unique sequences"
min_frac_genotype	0.8	"Three alleles" / "More than three alleles"
hetero_balance	0.25;0.75	"Heterozygote imbalance"
max_reads_unique_not_called	0.1	"Sequence with many reads not called"
min_unex	15	"Unexpected sequence detected"
slide	2	

[a] Two additional SNPs (rs12360952 and rs4752778) were included in the original MH defined by Kidd and co-workers [9].

3. Results and conclusions

In short, MHinNGS collects and stores sequences in bins, one bin for each MH, according to the two flanking sequences ('flank_up_length' and 'flank_down_length' in Table 1). Next, the program removes noise, identifies and names alleles, tests the genotype, and tests unique sequences (Fig. 1), that were not identified as either noise or alleles. In addition to the criteria defined in STRinNGS [8], four criteria have been added to the MHinNGS configuration file: 'mh_info', 'slide', 'linked_allele' and 'rare_snp' (Table 1). Each variant (SNP or indel) of the MH is defined in the configuration file ('mh_info' in Table 1) with rs number (if known), genome position, surrounding nucleotides, and possible alleles. The variant is identified by searching for the surrounding nucleotides to the variant position. The surrounding nucleotides must be an exact match. If a match is not found, the program will slide one nucleotide to the left or right, and try again, until the surrounding nucleotides match or the slide maximum ('slide' in Table 1) is reached. MHinNGS also searches for additional variants between the start and stop position (Table 1). If a variant is identified, the position and base call is indicated in the result file (Supplementary Tables 1 and 2), but it is not included in the MH name.

In the configuration file, it is possible to ignore specific positions

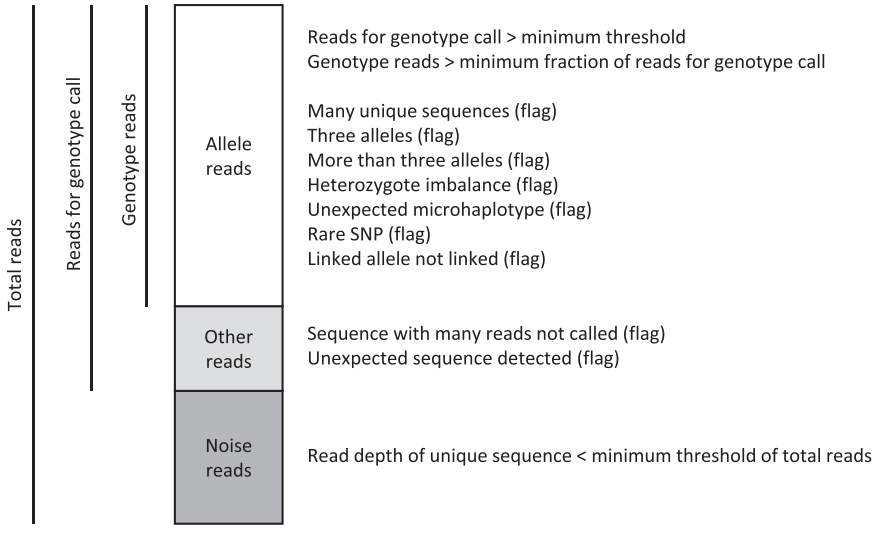

Fig. 1. Genotype calling with MHinNGS. There are three groups of reads for a locus as indicated on the left. 'Total reads' are all reads identified via the upstream and downstream flank. The 'Reads for genotype call' are all the reads that are left after noise reads have been removed. The 'Genotype reads' are the reads that make up the genotype. Thresholds and possible flags (Table 1) for each group of reads are indicated on the right.

('ignore_pos' in Table 1) with frequent errors, that generate multiple unique sequences (example in Supplementary Table 1). Furthermore, it is possible to define rare variants ('rare_snp' in Table 1), that are not part of the MH, with rs number, genome position, and alternative allele. If the alternative allele is detected, a warning flag is raised ("Rare SNP") in the comment column of the result file. However, the SNP is not included in the MH name.

Linked alleles may be defined in the configuration file (linked_allele in Table 1) with genome position, the MH allele that the allele is linked to, and the variant allele. If the allele is detected and the MH allele is identical to the expected, linked MH allele, the position and base call of the SNP allele is not shown in the results file. If another haplotype is detected, the flag "Linked allele not linked" is shown in the comment column of the result file.

In conclusion, MHinNGS is a freely available MH analysis software that provide the user with maximum flexibility and complete control of the analysis process.

Conflict of interest

None.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigss.2022.09.029.

References

[1] K.K. Kidd, A.J. Pakstis, W.C. Speed, R. Lagacé, J. Chang, S. Wootton, E. Haigh, J. R. Kidd, Current sequencing technology makes microhaplotypes a powerful new type of genetic marker for forensics, Forensic Sci. Int. Genet. 12 (2014) 215–224.
[2] F. Oldoni, K.K. Kidd, D. Podini, Microhaplotypes in forensic genetics, Forensic Sci. Int. Genet. 38 (2019) 54–69.
[3] Børsting, Morling, Next generation sequencing and its applications in forensic genetics, Forensic Sci. Int. Genet. 18 (2015) 78–89.
[4] P. Gill, H. Haned, O. Bleka, O. Hansson, G. Dørum, T. Egeland, Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches-twenty years of research and development, Forensic Sci. Int. Genet. 18 (2015) 100–117.
[5] S.L. Fordyce, H.S. Mogensen, C. Børsting, R.E. Lagacé, C.W. Chang, N. Rajagopalan, N. Morling, Second-generation sequencing of forensic STRs using the Ion Torrent™ HID STR 10-plex and the Ion PGM™, Forensic Sci. Int. Genet. 14 (2015) 132–140.
[6] C. Hussing, C. Huber, R. Bytyci, H.S. Mogensen, N. Morling, C. Børsting, Sequencing of 231 forensic genetic markers using the MiSeq FGx™ forensic genomics system - an evaluation of the assay and software, Forensic Sci. Res. 3 (2018) 111–123.
[7] H. Simayijiang, N. Morling, C. Børsting, Sequencing of human identification markers in an Uyghur population using the MiSeq FGxTM Forensic Genomics System, Forensic Sci. Res. 7 (2022) 154–162.
[8] C.G. Jønck, X. Qian, H. Simayijiang, C. Børsting, STRinNGS v2.0: improved tool for analysis and reporting of STR sequencing data, Forensic Sci. Int. Genet. 48 (2020), 102331.
[9] 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.