Forensic Science International: Genetics 26 (2017) 66–69

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Short communication

SEQ Mapper: A DNA sequence searching tool for massively parallel sequencing data

James Chun-I Lee[a,*], Bill Tseng[a], Liang-Kai Chang[b], Adrian Linacre[c]

[a] Department of Forensic Medicine, College of Medicine, National Taiwan University, No.1 Jen-Ai Road Section 1, Taipei 10051, Taiwan, ROC
[b] Laboratory of Cancer Genomic Medicine, LIHPAO Life Science. CO., Ltd., 8F Med Sci & Tech Bldg, No 201, Sec 2 Shipai Rd, Taipei 11217, Taiwan, ROC
[c] School of Biological Sciences, Flinders University, Adelaide 5001, Australia

ARTICLE INFO

Article history:
Received 30 March 2016
Received in revised form 5 October 2016
Accepted 13 October 2016
Available online 20 October 2016

Keywords:
STR
SNP
Bioinformatics
Massively parallel sequencing
Next generation sequencing
Forensic science

ABSTRACT

The development of massively parallel sequencing (MPS) has increased greatly the scale of DNA sequencing. The analysis of massive data-files from single MPS analysis can be a major challenge if examining the data for potential polymorphic loci. To aid in the analysis of both short tandem repeat (STR) and single nucleotide polymorphisms (SNP), we have designed a new program called SEQ Mapper to search for genetic polymorphisms within a large number of reads generated by MPS. This new program has been designed to perform sequence mapping between reference data and generated reads. As a proof-of-concept, sequences derived from the allelic ladders of five STR loci and data from the amelogenin locus were used as reference data sets. Detecting and recording the polymorphic nature of each STR loci was performed using four levels of search criteria: the entire STR locus spanning the two primers; the STR region plus the two primer sequences; the STR region only; and the two primers only. All the genotypes of 5 STR loci and the amelogenin gene were identified correctly using SEQ Mapper when compared to results obtained from capillary electrophoresis based on 10 test samples in this study. SEQ Mapper is a useful tool to detect STR or SNP alleles generated by MPS in both clinical medicine and forensic genetics.

1. Introduction

The advent of massively parallel sequencing (MPS), or next generation sequencing (NGS), technologies can lead to sequence data from millions of individual DNA strands in a single test and generate whole genome sequences [1] in forensic applications. Short tandem repeats (STRs) and single nucleotide polymorphism (SNPs) can be incorporated as targets for MPS [2–4]. The potential to determine a whole array of STR loci plus forensically informative SNPs from single genomes is enormous but brings with it issues with data handling. Many millions of reads from a single genome sequence can be generated from recently developed platforms but these data require software programs for their subsequent analysis. Examples of recently developed software capable of analysing huge genetic data sets created by MPS include lobSTR [5] and Razor [6].

We report on a new software program called SEQ Mapper that is designed to evaluate the polymorphic content of potential STR loci by performing sequence mapping between reference sequences and reads generated by MPS. SEQ Mapper is a web-based program that detects and records polymorphic STR loci using varying levels of search criteria. These criteria comprise: the entire STR locus spanning the two primers; the STR region plus the two primer sequences; the STR region only; and the two primer sequences only. The reliability of the program can best be demonstrated by comparison to genotype data from a range of different individuals using both standard separation of amplified STR fragments by capillary electrophoresis (CE) and using MPS data from the same individuals using SEQ Mapper.

2. Materials and methods

2.1. Sample preparation and STR genotyping

Ten oral swabs were collected with informed consent and following approval from the Institute Review Board. DNA was extracted from the swabs using a Tissue & Cell Genomic DNA Purification Kit (GeneMark, Taipei, Taiwan) following the

* Corresponding author.
E-mail addresses: jimlee@ntu.edu.tw, jimlee108@gmail.com (J.C.-I. Lee), steven.chang@llsc.com.tw (L.-K. Chang), adrian.linacre@flinders.edu.au (A. Linacre)
.

http://dx.doi.org/10.1016/j.fsigen.2016.10.006
1872-4973/© 2016 Elsevier Ireland Ltd. All rights reserved.

manufacturer's recommended protocol. The resulting quantity of DNA was determined using Nanodrop®. Amplification of STR loci was performed using reagents from an AmpFlSTR® Identifiler® PCR Amplification Kit (Life Technologies, NY, USA), using a thermal cycler (GeneAmp PCR System 2400, Life Technologies), following the manufacturer's protocol. PCR products were separated using POP-4 polymer in ABI PRISM® 3730 Genetic Analyzer (Life Technologies) and analyzed by GeneMapper 3.1 (Life Technologies).

2.2. STR genotyping by MPS

The primer sequences of the 5 STR loci and amelogenin used in this study are shown in the Supplementary data. Multiplex PCR amplifications were performed in a reaction volume of 20 μL, which contained approximately 0.2 ng of genomic DNA, 0.2 mM of each primer, 200 mM dNTP, 0.2 μL of AmpliTaq Gold*R* DNA Polymerase (Life Technologies) and 2 μL of 10 X Reaction Buffer (Life Technologies). PCR amplifications were conducted in a thermal cycler (GeneAmp PCR System 2400, Life Technologies) at 95 °C for 10 min, and for 30 cycles of 94 °C for 30 s and 62 °C for 30 s, with a final extension at 72 °C for 10 min.

Approximately 1 μg of PCR products were used for library preparation. Addition of barcodes, adaptor ligation and library amplification was performed according to the manufacturer's instructions as stipulated in the Ion Xpress Plus Fragment Library Kit (Life Technologies). The size distribution of the DNA fragments was analyzed on the Agilent Bioanalyzer using the High Sensitivity Kit (Agilent, CA, USA). Template preparation, emulsion PCR, and Ion Sphere Particle (ISP) enrichment was performed using an Ion PGM Hi-Q OT2 Kit (Life Technologies) following the manufacturer's instructions. The ISPs were loaded onto a 316 chip and sequenced using an Ion PGM Hi-Q Sequencing Kit (Life Technologies). After a successful sequencing reaction, the raw signal data were converted into the FASTA/Q format using FileExporter v4.4.3.0.

2.3. SEQ mapper

SEQ Mapper is a .NET web-based application (http://forensic.mc.ntu.edu.tw:9000/SEQMapperWeb/Default.aspx) for non-commercial use. The user submits DNA sequence data in the FASTA or FASTQ format and DNA sequences from files of reference loci in comma separated values (CSV) format. A comprehensive record is made for each allele at the specified loci that contains the name of the allele, the complete sequences of the allele, and STR sequences (or variant sequences). Note the STR sequences in an allele should have sufficient flanking bases at both ends to specify each allele. A requirement is that each FASTA or FASTQ read should follow a valid FASTA or FASTQ header. To assist, and be user-friendly, sample data are available for download from the web site.

For the preparation of reference alleles, all known sequences with microvariants such as degenerate bases within a primer, SNPs in flanking region or non-consensus repeat sequences should be included. In order to be compatible with CE-based STR data, the nomenclature of reference alleles should be followed using the same rule. As an example, in the reference allele of sample data of TH01 locus, a single base deletion from allele 10 is designated as allele 9.3. In the case of SNPs, since these variants can be recorded by MPS rather than CE, the repeat number and the polymorphic base are used to designate the allele from the MPS data. As an example, in the reference alleles of D5S818 there are sequence-based variants recorded as Xc and Xt where X represents repeat number and the nucleotide (c and t) followed X represents SNP: rs25768, which is located 13 bp from the 5' end of the repeat region [7].

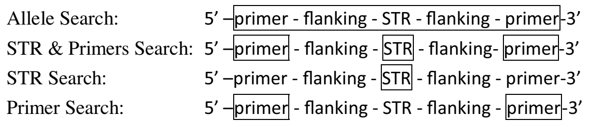
Allele Search: 5' –primer - flanking - STR - flanking - primer-3'
STR & Primers Search: 5' –primer - flanking - STR - flanking- primer-3'
STR Search: 5' –primer - flanking - STR - flanking - primer-3'
Primer Search: 5' –primer - flanking - STR - flanking - primer-3'

Fig. 1. SEQ Mapper Searching Methods showing the four levels of searches performed. These are: the entire locus from the 5' base of the up-stream primer to the 5' end of the downstream primer; the STR repeat motif plus the primer sequences; only the STR repeat motif; and the primer sequences alone.

SEQ Mapper provides an interface allowing the user to search for specific DNA primer sequences by defining the starting position, called the 'Begin Index', and the 'Length' of primers. This allows flexibility in any search by modifying the DNA sequence used, particularly if no matches are recorded in Primer Search Reports. Users will be able to determine the 5' start of the primers based on the reference alleles. In addition to specifying the length of primers, primer sequences can be extracted from the data from any reference alleles as specified by the user.

Once the primer sequences are determined, these will be used in the 'STR & Primers Search' and also the 'Primers Search'. The 'STR & Primers Search' will use sequence data from the STR repeat sequence provided in the reference locus data file. Note that this search assumes that this is the forward strand although the app is capable of computing the reverse sequence for the specified region and then performing search using the reverse sequence. Note that the 'Allele Search' and 'STR Search' will not use any potential primer sequences, rather SEQ Mapper will use allele and STR sequences prepared by user in the reference sequences data files. The reverse of the allele sequence, STR repeat sequences, and primer sequences will be generated automatically and used in the search.

Four levels of search can be performed allowing greatest flexibility. These different levels used by SEQ Mapper are illustrated in Fig. 1.

1. Allele Search: this search is at the highest stringency and requires a full match using the entire reference allele sequence against the specific FASTA/Q read(s). Note that the alleles sequences used are based on the length defined by both the 'Begin Index' of primers or the 5′ and 3′ terminal bases.
2. STR & Primers Search: this is the next level of stringency and requires a match using sequences of the STR repeat plus the defined two primer sequences.
3. STR Search: this is the next level of search and only requires a match based on the STR repeat sequences only.
4. Primers Search: this is the lowest level of search requires a match to both of the defined 5′ and 3′ primer sequences only.

At the end of the search process multiple reports will be generated and saved in the CSV format. These reports comprise the following.

1. A Summary Report includes all the following information: the number of FASTA/Q reads found in the respective Allele Search Report; STR & Primers Search Report, and STR Search Report; and the number of reads obtained in the Primer Search Report. The Summary Report also provides information on the total distribution and length of reads used.
2. Allele Search Report contains the number of FASTA/Q reads that were found to match the entire allele for a specific reference locus.
3. STR & Primers Search Report contains the number of FASTA/Q reads found matching the STR repeat sequence and DNA sequences of the two primers used for each specified loci.