SEQ MAPPER: A DNA SEQUENCE SEARCHING TOOL FOR MASSIVE PARALLEL SEQUENCING DATA

James Chun-I Lee[1], Bill Tseng[1], Liang-Kai Chang[2], Adrian Linacre[3]

[1]Department of Forensic Medicine, College of Medicine, National Taiwan University, No.1 Jen-Ai Road Section 1, Taipei 10051, Taiwan, ROC.
[2]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
[3]School of Biological Sciences, Flinders University, Adelaide 5001, Australia.

The development of massive parallel sequencing (MPS) has increased greatly the scale of DNA sequencing. The massive data-files from one 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, we have designed a new program called SEQ Mapper to search for genetic polymorphisms within a large number of reads generated by MPS. The new program has been designed to perform sequence mapping between reference data and generated reads. As 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. Three types of STR-related loci were used for sequence mapping including: the STR repeat region only; the STR region plus the two primer sequences; and the entire STR locus spanning the two primers, flanking DNA and repeat region. The program identified STR alleles from complex DNA data obtained from MPS using these three parameters with lowest stringency using the STR regions only, and the highest stringency when using the entire STR locus (including primers and flanking DNA sequences). Polymorphic variation was observed in the flanking sequences by comparison to the entire STR loci. The genotypes of these loci were correctly identified using SEQ Mapper when compared to results obtained from capillary electrophoresis based on 10 test samples. Novel and recorded micro-variants can be identified using this new program. SEQ Mapper accepts FASTA and FASTQ format of reads. Searching parameters including 5' and 3' primer sequences and flanking DNA sequences can be defined by users. SEQ Mapper has been found to be a valuable tool to detect STR or SNP alleles generated by MPS in both clinical medicine and forensic genetics.