Jan 21, 2020

Targeted gene sequencing using LR-PCR-NGS approach

Sathiya Maran[1]

[1]Monash University, Malaysia Campus

1 Works for me dx.doi.org/10.17504/protocols.io.bbjtiknn

Sathiya Maran

1 Genomic DNA of the recruited samples were extracted from peripheral blood from patients (n=48) using a commercially available kit GeneAll® ExgeneTM Blood SV mini, GeneAll, Korea, at the Human Genome Center, Universiti Sains Malaysia (HGC-USM). For patients who refused to give their blood (n=29), saliva samples were collected instead using PSP® SalivaGene DNA kit, STRATEC. Findings from a high-throughput genotyping study reported that DNA isolates from blood draws and DNA isolates form saliva showed no comparable differences in term of results [20]. The concentration and purity of the extracted DNA was measured using NanoQuant (Tecan, USA).

2 Long-range PCR development and amplification

3 Genomic DNA of the recruited samples were extracted from peripheral blood from patients (n=48) using a commercially available kit GeneAll® ExgeneTM Blood SV mini, GeneAll, Korea, at the Human Genome Center, Universiti Sains Malaysia (HGC-USM). For patients who refused to give their blood (n=29), saliva samples were collected instead using PSP® SalivaGene DNA kit, STRATEC. Findings from a high-throughput genotyping study reported that DNA isolates from blood draws and DNA isolates form saliva showed no comparable differences in term of results [20]. The concentration and purity of the extracted DNA was measured using NanoQuant (Tecan, USA).

4 MYH3 sequence from Ensembl Genome Browser (ID: ENSG00000109063, GRCh38) was used to design the LR-PCR primers. The entire promoter, 5' and 3' MYH3 untranslated and coding regions were amplified in four distinct LR-PCR reactions. LR-PCR was performed using a Max Taq Polymerase (Vivantis, USA), according to the manufacturer's protocol at the HGC-USM.

5 DNA thermal cycling was performed using a SureCycler 8800 (Agilent Technologies, UK); 94°C for 2 min, 94°C 12 sec, annealing temperature for 30 secs, 68°C for 10 mins, 68°C for 7 mins for a total of 35 cycles. The size of the amplified PCR products was determined by gel electrophoresis using SYBR® Green I nucleic acid gel stain (Life Technologies, USA) under ultraviolet light. The amplified products were then purified by using an Illustra Exo-ProStar (GE Healthcare, UK). Subsequently, the LR-PCR fragments were quantified using Qubit® 1.0 Fluorometer (Invitrogen, USA) with Qubit dsDNA BR Assay Kits (Invitrogen, USA) and were pooled together at equal molar ratios.

6 NGS library preparation and sequencing

7 For each sample, 5 µl (1 ng in total) at 0.2 ng/ul of pooled LR-PCR products was used to generate indexed paired-end libraries with Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA) according to the manufacturer's protocol. A fragment length of the libraries was ascertained using the High Sensitivity DNA kit (Agilent Technologies, Santa Clara, CA) on the 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA). Normalized libraries were subjected to a 300-cycle sequencing run with MiSeq Reagent Nano Kit v2 (Illumina, San Diego, CA) on the MiSeq (Illumina, San Diego, CA), were carried out the Genetic Laboratory, faculty of Science, University of Malaya.

8 Data enrichment and variant calling

9 The sequencing data was evaluated with FastQC (version 1.0, Babraham Bioinformatics, UK) for quality control on the raw data. Subsequently, the reads were aligned to a reference genome (hg19) by BWA-MEM (version 1.0) algorithm based on the default settings. The SNPs were filtered by a set of filters. A variant Studio v2.1 software (Illumina, San Diego, CA) was used to identify and annotate exonic and intronic variants and to determine if the variants have been reported in public databases. An integrative Genomics Viewer (IGV version v1.0.0) was used to examine the read counts of the target amplicons and confirm the detected variants. All variants identified by NGS were resequenced with Sanger sequencing.

10 Genotyping in controls cohort

11 The identified variants were also genotyped in the control population either by made-to-order TaqMan SNP Genotyping Assay (ThermoFisher, USA), custom TaqMan SNP Genotyping Assay (ThermoFisher, USA) or Sanger sequencing, based on the manufacturer's protocol.

12 Statistical and bioinformatics analyses

13 The study protocol and reporting were developed according to the STREGA guidelines for case-control studies [21]. The allelic and genotypic frequencies of the patient and control groups were determined using a Fisher's exact test with odds ratio and 95% confidence interval (CI). A x2 P-value of ≤ 0.05 was considered as statistically significant. Haploview version 4.2 (Broad Institute of MIT and Harvard, USA) was used for analysis of linkage disequilibrium (LD) and haplotype block [22]. LD blocks were defined according to the haplotype block definition of Gabriel and colleagues [23]. The method defines pairs to be in strong LD if the one-sided upper 95% confidence bound on D' is >0.98 and the lower bound is above 0.7.

14 In-silico analysis of pathogenic potential of identified non-synonymous variants were predicted using Sorting Intolerant From Tolerant (SIFT) and Polymorphism Phenotyping v2 (PolyPhen-2). SIFT scores which range between 0 and 1 as well as scores below 0.05 suggest that the amino acid change is not tolerated whereas, PolyPhen-2 scores >0.85 are interpreted as "probably damaging" and scores 0.15–0.85 as "possibly damaging". The evolutionary conservation of the identified non-synonymous variants was conducted using a Clustal Omega programme (EMBL-EBI, European Bioinformatics Institute). The splicing effects were predicted by the Human Splicing Finder (HSF) version 3.0 (http://www.umd.be/HSF/) and ESEfinder 3.0 (http://rulai.cshl.edu/).

15 Modelling of the mutant protein structure

16 Prior to the template selection, both wild and mutant types of complete sequences of human myosin MYH3 were aligned against the sequences from Protein Data Bank (PDB), (www.rcsb.org/pdb). Our results showed that protein structures with the PDB ID: 5TBY and 5WJ7 with 72% and 83% sequence identity, respectively are suitable templates for homology modelling. Multiple templates approach homology modelling was adopted in this study. Both the selected templates indicated to belong to the family protein from Homo sapiens myosin 7, which is closely related to the modeled sequences.

17 A total of 20 models for wild and mutant type sequence were built using Modeller 9v20 (www.salilab.org/modeller) and the best models with the lowest discrete optimize potential energy (DOPE) scoring was selected for each wild and mutant types. The structural differences between the mutant against the wild type built model was calculated using a root mean square deviation (RMSD) method.