Workflow for replicable Sanger sequencing of NGS-derived mutations in clinical application

M.Orlova[1], A.Shershebnev[1], D.Borisevich[1,2], I.Stetsenko[1], D.Plakhina[1] , A.Krasnenko[1,3], D.Korostin[1]

1. Genotek, Moscow, Russia, orlova@genotek.ru

2. Bioengineering and Bioinformatics Department, Lomonosov Moscow State University, Moscow, Russia

3. Pirogov Russian National Research Medical University(RNRMU), Moscow, Russia

Introduction

Next-generation sequencing (NGS) is a powerful tool for calling mutations in patients with inherited disorders. However, Sanger validation of findings often required because of medical diagnostics registration laws. Stable workflow for delivering reproducible Sanger sequencing is required to apply NGS at the bedside.

Materials and Methods

Database of primers pairs for 66100 exons of genes mentioned in Agilent FocusedExome panel was prepared using PrimerBlast [1] and Python scripts. 195 mutations from 110 patients were sequenced, using these primers. The oligonucleotide synthesis was performed by Evrogen [2] and Lumiprobe [3] on an ABI 3900 DNA synthesizer following the manufacturer’s protocol. The PCR was performed on Biorad T100 PCR cycler using Phusion DNA polymerase. Resulting PCR products were analyzed by electrophoresis in 1% agarose gel. Sanger sequencing was performed using BigDye v3.1 Sanger sequencing kit and ABI 3500 genetic analyzer following manufacturer’s protocol. Pipeline for chromatogram processing was developed using biopython [4] and SangerSeqR [5]. Algorithm for quality control and mutation calling from chromatograms was created and implemented using Python and R programming languages. Google Cloud Platform was used as underlying infrastructure [6].

Algorithm

Reference sequence with 1200 base pairs around position being validated is extracted. Alleles from ab1-file are called with sangerseqR according to reference sequence. Then alleles are separately aligned to reference. In order to get left-most alignment (it is essential for indels vaildation) the part of this alignment around position subject for validation is extracted and realigned with Biopython.

5' end and 3' end low-quality nucleotides of each allele is replaced by N. Phred score is used as quality metrics. All short regions of low quality and abnormally high or low peaks are also replaced by N. Algorithm carefully treats heterozygotes as they also lead to low peaks with low Phred score. Thus possible dye blobs and other low-quality regions are excluded from consideration. Besides, alleles are tested for polymelase slippage: if a homopolymer is followed by indel it is replaced with N.

Finally, validation decision is made.

Results

Workflow for reproducible Sanger validation was established. First, we created fixed database of 66100 pairs of primers, so every validation of the same mutation will be performed identically in the laboratory. Next, we developed algorithm for detection of polymerase slippage, dye blobs and low-quality nucleotides in chromatogram. Further, we created algorithm for automated calling of position subject for validation. Finally, we implemented automated processing of two and more chromatograms covering the same position in one patient to produce joint decision. We tested our algorithm using library of 195 chromatograms. No false positive validations (validation of not-present mutation) were found.

Here are detailed results of tests:

True positive - 129 chromatograms

True negative - 56 chromatograms

Validation failed (algorithm was unable to provide decision for chromatogram of decent quality) - 7 chromatograms

False positive - 0 chromatograms

False negative - 3 chromatograms

Resulting software will be available at https://validation-166217.appspot.com/
by 31 May 2017.

Discussion

Chromatogram visualizers as Emboss, Chromas, FinchTV require manual analysis by qualified biologist. TraceTuner and sangerseqR can call bases and split chromatogram sequence into two alleles, but they require bioinformatics skills to be used, and are not complete solutions for automated chromatograms processing. We found only one tool which can validate mutation in exact position - MutAid. This software, however, require installation and can be used only by bioinformaticians. Instead, our solution provide simple interface for laboratory. Also our software takes into consideration several chromatograms and provide joint consensus decision on mutation status, not available from other solutions. As a result, our solution decrease Sanger results processing time and improve quality of the processing, which is essential for medical application.

Conclusions

Sanger validation system was developed to reproducibly validate mutations detected by NGS. Algorithm was created and implemented using Python and R. Software was tested using manually curated 195 chromatograms and demonstrated precision and recall of 100% and 97.7% respectively. Software will be available for non-commercial purposes at https://validation.appspot.com/ by 31 May 2017.

1. https://www.ncbi.nlm.nih.gov/tools/primer-blast/
2. http://evrogen.ru
3. https://ru.lumiprobe.com
4. http://biopython.org/
5. http://bioconductor.org/packages/release/bioc/html/sangerseqR.html
6. https://cloud.google.com/