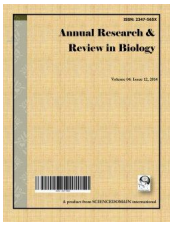

Annual Research & Review in Biology

25(4): 1-5, 2018; Article no.ARRB.40312
ISSN: 2347-565X, NLM ID: 101632869

Pivotal Role of High Sensitivity Variant Calls and Confirmation Methods for Next-Generation Sequencing Findings: A Case Report

Mario Gorenjak[1*], Gregor Jezernik[1], Katja Repnik[1,2], Natasa Marcun Varda[3] and Uros Potocnik[1,2]

[1]Centre for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, Taborska 8, 2000 Maribor, Slovenia.
[2]Laboratory of Biochemistry, Molecular Biology and Genomics, Faculty of Chemistry and Chemical Engineering, University of Maribor, Smetanova Ulica 17, 2000 Maribor, Slovenia.
[3]Division of Paediatrics, University Medical Centre Maribor, Ljubljanska Ulica 5, 2000 Maribor, Slovenia.

Authors' contributions

This work was carried out in collaboration between all authors. Author MG designed the bioinformatics NGS pipeline, performed the analysis and wrote the first draft of the manuscript. Authors GJ and KR performed the DNA library preparations and sequencing runs. Author NMV provided the case and biological material. Author UP helped with the study design and manuscript writing. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/ARRB/2018/40312
Editor(s):
(1) Moacir Marocolo Júnior, Professor, Physiology, Institute of Biological Sciences, Federal University of Juiz de for a, Brazil.
(2) George Perry, Dean and Professor of Biology, University of Texas at San Antonio, USA.
Reviewers:
(1) Joel K. Weltman, Brown University, USA.
(2) Linnea Baudhuin, Mayo Clinic, USA.
Complete Peer review History: http://www.sciencedomain.org/review-history/23948

Case Report

Received 15th January 2018
Accepted 22nd March 2018
Published 2nd April 2018

ABSTRACT

Aims: The aim of this case report is to demonstrate the importance of prioritization of sensitivity over specificity, coupled with additional confirmation by using standard methods. Next-generation sequencing has revolutionized genetic research as it has allowed sequencing of human genomes within days. Generated raw sequencing data are manipulated using bioinformatic approaches for variant detection. Variant discovery should be performed on appropriately pre-processed data with the aforementioned prioritization of sensitivity over specificity.

**Corresponding author: E-mail: mario.gorenjak@um.si;*

Presentation of the Case: Here, we report a case of a low quality variant call, emitted due to prioritization of sensitivity over specificity. This call was found to be a causative variant for the patient's phenotype. DNA extracted from peripheral venous blood of a young female with encephalopathy was sequenced on a MiSeq apparatus. The obtained and analyzed call set emitted a low quality heterozygous insertion with a high probability of a false negative call. Annotation revealed a known pathogenic insertion rs758946412 with a frameshift consequence flagged with "Early infantile epileptic encephalopathy type 9" in ClinVar. The emitted insertion was validated and confirmed by using SANGER sequencing and RFLP.
Conclusion: In the presented case, the variant could have easily been missed without the prioritization of sensitivity over specificity. Furthermore, the presented case also demonstrates the importance of additional methods for confirmation of NGS calls that do not meet the thresholds.

Keywords: Next-generation sequencing; DNA sequencing; bioinformatics; variant-calling.

1. INTRODUCTION

Rapid development of Next-generation sequencing (NGS) technologies has revolutionized genetic research and we are now able to sequence the human genome in a matter of days. On the other hand, due to the rapid development of new technologies, technical and data management challenges are emerging [1]. One sequencing run generates millions of small fragmented DNA sequences (reads), which need further analysis. With the use of bioinformatic analyses the generated raw data is further manipulated in terms of mapping the multiple individual fragments to the reference genome sequence, thus providing the depth of fragment coverage and variant detection [2]. To detect the variations, variant-calling workflows are used. A good variant-calling workflow involves methods of data preparation, which compensate for various errors, such as amplification bias, machine errors, software errors, and mapping artefacts [3]. To construct a robust variant-calling workflow, publicly available program packages BWA and GATK are used [4-6]. For variant discovery, appropriate pre-processing of data with prioritization of sensitivity over specificity is paramount in order to include as many potential variants as possible, without missing any. Subsequently, filters compensating for lower specificity can be applied [3].

2. PRESENTATION OF CASE

Here, we report a case of a low quality variant call, emitted due to employing prioritization of sensitivity over specificity, and found to be a causative variant for the patient's phenotype. Moreover, we also demonstrate the importance of additional confirmation of such variant calls, since low quality calls have a high probability of false positive emission.

DNA was extracted by TRI-reagent (Sigma-Aldrich, Munich, Germany) from the peripheral venous blood of a 5-year-old female with epileptic encephalopathy who presented with global mental retardation and a history of seizures. A written informed consent was obtained from both parents. We generated sequencing libraries for paired-end read (2×150 bp) exome sequencing on a MiSeq apparatus (Illumina) using TruSight One Sequencing Kit (Illumina, San Diego, California). The obtained reports of generated raw sequencing data showed decreased median read length of 54 bp and a decreased mean region coverage depth of 20.3× with 64.6% of reads at target coverage at 10× and 40.9% at 20×. However, the read quality at Q30 was 96.3%, indicating accurate sequencing of the sample. Generated raw data files were additionally checked by using FastQC v0.11.5. Alignment was carried out with the BWA v0.7.12-r1039 program package and reads were mapped to the UCSC hg19 reference genome. PCR duplicates were marked by using PicardTools v2.3.0. Further, the reads were locally realigned around indels by using GATK v3.3 and paired end mate information was subsequently fixed by using PicardTools. Base quality scores were recalibrated and variants were joint called with other exome samples by using HaplotypeCaller algorithm in the GATK package. The minimum phred-scaled emission and calling confidence threshold were set to 5 and 30, respectively. Variant quality scores were recalibrated and normalized by using GATK. Furthermore, the obtained vcf files were annotated by using ANNOVAR v2017-06-01.

With regard to the clinical presentation, the obtained call set was carefully inspected for variants in the genes correlated with epileptic encephalopathy and specifically in the *PCDH19* gene region, which is known to play a role in

early infantile epileptic encephalopathy restricted to females [7]. The inspected single nucleotide variants (SNV) were either synonymous or flagged tolerated/benign according to the Sift and PolyPhen-2 algorithms, which are tools for prediction of amino acid substitution effects and can be used for SNV evaluation [8,9]. In the *PCDH19* gene region, two good synonymous benign SNV's with "passed filter" status were called (rs1953337 at the coordinate ChrX:99661969 and rs41300169 at the coordinate ChrX:99663194) (Fig. 1). However, a heterozygous insertion was also called in the *PCDH19* target gene region, which was flagged as a low quality call with low read depth of 12 and hence could be a possible false positive variant.

3. DISCUSSION

The retained variant was "caught" in the final call set due to low calling emission threshold and higher sensitivity; it was also included in the annotation step. Annotation revealed that the called variant is a known insertion of the C base, with ID rs758946412 (Fig. 1) at the coordinate ChrX:99662504 with the frameshift consequence p.Tyr366LeufsTer10, and is flagged in ClinVar as pathogenic with disease name "Early infantile epileptic encephalopathy type 9" (EIEE9).

EIEE9 is an X-linked disease limited to females, with high but incomplete penetrance [10]. A study of two unrelated families with affected female family members also suggests a somatic germline mosaicism [11].

Although the detected insertion is more than compliant with the patient's phenotype, validation is necessary in cases like this due to low read depth and high probability of false positive emission. Insufficiently validated tests tend to present a threat to the patient and should be unacceptable in a clinical diagnostic setting [1] and even in research oriented tests. In order to stringently validate the emitted call, primers for SANGER sequencing [12] and Restriction Fragment Length Polymorphism analysis (RFLP) were subsequently designed (Table 1). Both analyses confirmed the heterozygous insertion rs758946412 (Figs. 2 and 3).

The present case report clearly elucidates the meaning of prioritization of sensitivity over specificity and supports the idea of lowering the emission thresholds. Appropriate filters, such as VariantRecalibrator in the GATK package during the variant quality score recalibration step can still be used and applied on high-sensitivity call sets in order to achieve the desired balance between sensitivity and specificity [3].

Table 1. Primers and restriction enzyme

Gene	Sequence accession number	Primers 5' to 3' (SANGER+RFLP)	Restriction enzyme
PCDH19	NC_000023.11	TGGACGTGCAGGCTAAGGACT TGGAGACATAGGTGAAGACAGGCAT	BseL I

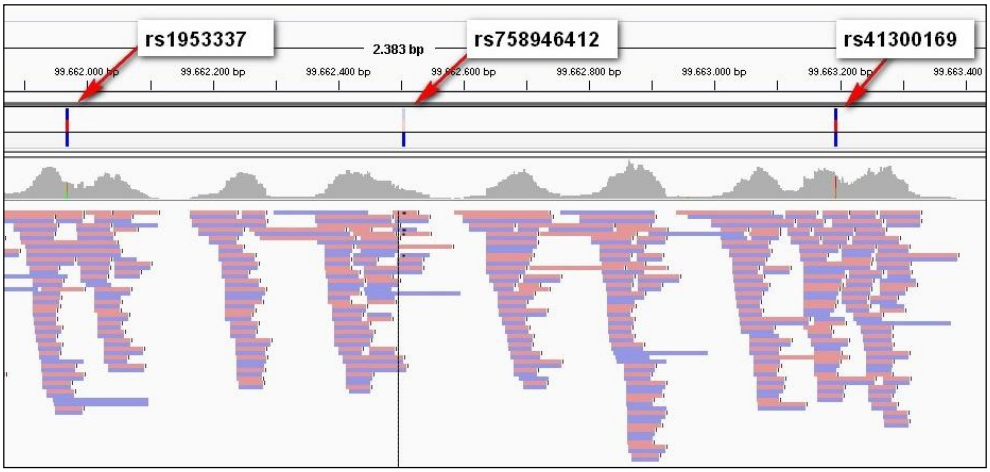

Fig. 1. Graphical view of the patient's binary alignment file
Arrows indicate the called SNV's and the insertion in the PCDH19 gene region

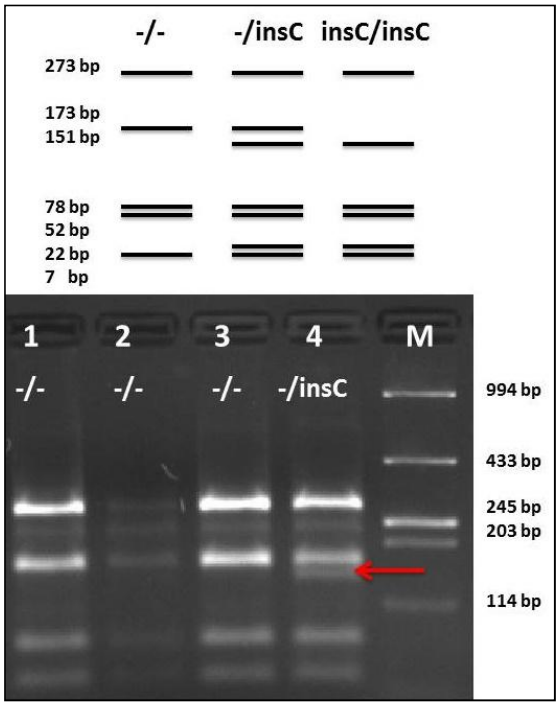

Fig. 2. RFLP design and agarose gel electrophoresis results

Arrow indicates the 22-bp shorter fragment of the allele with insertion. Lane 1: Female reference sample; Lane 2: Male reference sample; Lane 3: Female reference sample; Lane 4: Sample of the presented case; Lane M: DNA marker

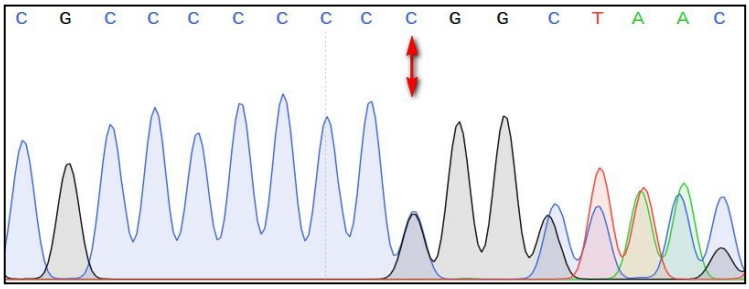

Fig. 3. SANGER chromatogram data

Arrow indicates the insertion of the C base and the beginning of the frameshift

4. CONCLUSION

The variant could easily have been missed in the presented case if the prioritization of sensitivity over specificity had not been used. Furthermore, the presented case also demonstrates the importance of additional methods for confirmation of NGS calls that do not meet the thresholds.

CONSENT

All authors declare that a written and informed consent was obtained from the patient's parents.

ACKNOWLEDGEMENT

The authors thank dr. Boris Gole and dr. Helena Sabina Celesnik from Centre for Human Molecular Genetics and Pharmacogenomics, Faculty of Medicine, University of Maribor, for critical evaluation of the manuscript. This work was supported by Slovenian Research Agency: Pharmacology and Pharmacogenomics (P3-0067) and Analysis and development of rare diseases field in Slovenia (V3-1505).

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Matthijs G, Souche E, Alders M, Corveleyn A, Eck S, Feenstra I, et al. Guidelines for diagnostic next-generation sequencing. Eur J Hum Genet. 2016;24(1):2-5.
2. Behjati S, Tarpey PS. What is next generation sequencing? Arch Dis Child Educ Pract Ed. 2013;98(6):236-8.
3. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11 0 1-33.
4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5): 491-8.
5. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.
6. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303.
7. Depienne C, LeGuern E. PCDH19-related infantile epileptic encephalopathy: An unusual X-linked inheritance disorder. Hum Mutat. 2012;33(4):627-34.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
9. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81.
10. Ryan SG, Chance PF, Zou CH, Spinner NB, Golden JA, Smietana S. Epilepsy and mental retardation limited to females: An X-linked dominant disorder with male sparing. Nat Genet. 1997;17(1):92-5.
11. Dibbens LM, Kneen R, Bayly MA, Heron SE, Arsov T, Damiano JA, et al. Recurrence risk of epilepsy and mental retardation in females due to parental mosaicism of PCDH19 mutations. Neurology. 2011;76(17):1514-9.
12. Depienne C, Bouteiller D, Keren B, Cheuret E, Poirier K, Trouillard O, et al. Sporadic infantile epileptic encephalopathy caused by mutations in PCDH19 resembles Dravet syndrome but mainly affects females. PLoS Genet. 2009;5(2): e1000381.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history/23948