Genetics inMedicine | LETTER TO THE EDITOR

Standard enrichment methods for targeted next-generation sequencing in high-repeat genomic regions

To the Editor: In addition to the increasing use of whole-exome and whole-genome sequencing for diagnosis of symptomatic recessive genetic diseases, researchers are developing targeted disease and gene panels using next-generation sequencing (NGS) for potential screening applications in adults and for newborn screening.[1] Because significant actions could be taken based on the results of such screens, validation of the methods is necessary. To test the accuracy in a repeat genomic region of a published targeted NGS screening test for 448 childhood recessive diseases, we sequenced three samples reported to be positive by the authors for mutations in a newborn screening disease gene. This complex published method generated raw data from Agilent SureSelect hybrid capture enrichment of sheared DNA that was filtered using a variety of approaches. This method specified a minimum number of uniquely aligned reads of quality score >20; it specified that >14% of reads were needed to call a variant; and it restricted mutation calls to those associated with severe disease phenotypes in databases such as the Human Gene Mutation Database (HGMD). Using this highly curated approach, the authors reported mutations found in 104 Coriell samples, including two congenital adrenal hyperplasia (CAH) mutations in *CYP21A2* in three samples clinically affected with diseases other than CAH. HGMD mutation CM071683, Ala392Thr (1174G>A), was reported in samples NA01881 and NA03365, and mutation CM920233, Pro454Ser (1360C>T), was reported in sample NA00059. The second mutation, CM920233, is a recognized cause of mild, nonclassical CAH,[2] but the former, CM071683, was reported in 2006 as a novel mutation by Robins and coworkers, who were studying a Caucasian child presenting with premature adrenarche at 6 years of age and advanced bone age of 8.5 years.[3]

CYP21A2 is near a highly homologous pseudogene in the high-repeat human leukocyte antigen class III genomic region on chromosome 6. Due to the high similarity between the pseudogene and the functional gene (>98%), standard hybrid probe sets and primers created using standard design software are not always specific to the functional gene. We have sequenced *CYP21A2* in the three samples reported to carry the two *CYP21A2* mutations by (i) traditional Sanger sequencing of a long-range *CYP21A2* amplicon specific to the functional gene using the ABI 3730, (ii) Fluidigm PCR enrichment of genomic DNA using standard design approaches for primer design with NGS using the Roche 454, and (iii) shearing of the *CYP21A2* long-range amplicon specific to the functional gene followed by NGS using the Roche 454. Methods (i) and (iii) employ a long-range amplicon template that is the result of specific amplification of the functional gene for sequencing. This long-range amplicon was generated using primers that make use of multiple base differences between the functional and pseudogenes upstream of the start codons (forward primer) and sequence that is deleted in TNXA downstream from the pseudogene for the reverse primer.

The CM920233 mutation reported in sample NA00059 was confirmed by all three methods. However, the CM071683 mutation reported in samples NA01881 and NA03365 was not found in the functional gene by Sanger sequencing or by NGS of the sheared functional gene–specific long-range amplicon (methods (i) and (iii)). Conversely, Sanger sequencing of a long-range amplicon specific for the pseudogene for samples NA01881 and NA03365 did detect this mutation in the pseudogene. This G>A change was seen in both samples with the second sequencing method, using primers designed by a standard software approach (Primer3; Whitehead Institute for Biomedical Research, Cambridge, MA) with genomic DNA.

The hybrid capture probe used for capture of the region containing the CM071683 mutation in the publication[1] matches the reference sequences (UCSC Human Genome Browser, GRCh37/hg19 assembly) for both the functional gene *(CYP21A2)* and pseudogene *(CYP21AP)*, and therefore this nonspecific probe fails to separate *CYP21A2* from its pseudogene. The forward and reverse PCR primers used to amplify the CM071683 mutation in the second NGS sequencing approach using genomic DNA also match the reference sequences (UCSC Human Genome Browser, GRCh37/hg19 assembly) for both the functional gene and the pseudogene, and likewise fail to separate *CYP21A2* from its pseudogene. When these data are analyzed, the sequence reads of the pseudogene incorrectly align with the reference sequence of the functional gene.

The original report of the CM071683 mutation[3] did not specify the primer sequences that were used. We requested the primer sequences for this mutation from the authors, but they could not be released due to their institutional policy.

Standard approaches for NGS enrichment will not always work in high-repeat genomic regions and will not reliably separate functional genes from pseudogenes. This problem is common to all NGS methods that depend on libraries of DNA fragments that are not long enough to capture sufficient differences between highly homologous functional genes and pseudogenes. To capture reliable differences between the functional *CYP21A2* and its pseudogene, our long-range amplicons are >5,000 bp in length, and sequence

LETTER TO THE EDITOR

spans of this length cannot be directly sequenced by current, commonly used NGS instruments. Those developing targeted NGS tests for genes, gene panels, and the exome, as well as whole-genome sequencing applications, need to be aware of the large number of pseudogenes in the human genome and the likelihood of including genes that have pseudogenes. Other approaches are needed for enrichment of these genes, such as sequencing a more specific amplicon template isolated within the region. Flagging these genes and regions in reference sequences, hybrid capture arrays, primer design software, and data analysis software would help those designing tests and analyzing data to be aware of potential problems.

DISCLOSURE

The authors declare no conflict of interest.

Patricia W. Mueller, PhD[1], Justine Lyons, PhD[1], Gregory Kerr, BS[1], Chad P. Haase, BS[2] and R. Benjamin Isett, MS[3]

[1]Molecular Risk Assessment Laboratory, Newborn Screening and Molecular Biology Branch, Centers for Disease Control and Prevention, Atlanta, Georgia, USA; [2]Department of Human Genetics, Emory GRA Genomics Core, Emory University, Atlanta, Georgia, USA; [3]Cancer Genomics Shared Resource, Winship Cancer Institute, Emory University School of Medicine, Atlanta, Georgia, USA. Correspondence: Patricia W. Mueller (pwm2@cdc.gov)

REFERENCES

1. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011;3:65ra4.
2. Nimkarn S, New MI. 21-Hydroxylase-deficient congenital adrenal hyperplasia. Gene Reviews NCBI Bookshelf ID: NBK1171, PMID: 20301350 (Last revision: 2010).
3. Robins T, Bellanne-Chantelot C, Barbaro M, Cabrol S, Wedell A, Lajic S. Characterization of novel missense mutations in CYP21 causing congenital adrenal hyperplasia. *J Mol Med* 2007;85:247–255.

doi:10.1038/gim.2013.119