JIG Journal of Interdisciplinary Genomics

https://doi.org/10.22742/JIG.2022.4.2.19
Journal of Interdisciplinary Genomics 2022;4(2):19-23
eISSN 2671-6771

Pseudogenes: Nuances and Nuisances in Molecular Diagnostics

Seung Hwan Oh

Department of Laboratory Medicine, Pusan National University Yangsan Hospital, Yangsan, Korea

Pseudogenes are genomic regions that contain gene-like sequences that have a high similarity to the known genes but are non-functional. They are categorized into processed, unprocessed, and unitary pseudogenes. Unprocessed pseudogenes generated by duplications can be problematic in sequencing approaches in molecular diagnostics. We discuss the risk of misdiagnosis when investigating genes with pseudogenes of high homology, and describe a method for identifying these small and annoying differences between parent genes and pseudogenes, including parent gene-specific assay design.

Key words: Pseudogene, Parent gene, Retrotransposition, Duplication

REVIEW ARTICLE

Received: October 4, 2022
Accepted: October 13, 2022

Correspondence to: Seung Hwan Oh
Department of Laboratory Medicine, Pusan National University Yangsan Hospital,
20 Geumo-ro, Mulgeum-eup, Yangsan 50612, Korea
Tel: +82-55-360-1870
Fax: +82-55-360-1880
E-mail: paracelsus@pusan.ac.kr

ORCID
https://orcid.org/0000-0002-1946-9939

INTRODUCTION

Pseudogenes (Ψ) are genomic regions that contain gene-like sequences that have high similarity to the known genes (parent genes, i.e., their functional paralogs) but are non-functional. Therefore, pseudogenes have been considered as 'junk DNA' [1-6]. However, recent discoveries unveil this junk DNA's biological meanings, such as a regulatory role in cancer and developmental processes, and a genetic repository role to store and expand genetic information [7-14]. Explorations and annotations of pseudogenes can give an insight into their genetic evolution, diversification, and functionalization [15-17].

Pseudogenes can be introduced into our genome through several mechanisms such as spontaneous mutation that results in 'unitary pseudogene' [18], duplication that constitutes 'unprocessed pseudogene' [19], and, retrotransposition of processed mRNAs that produces 'processed pseudogene' [20]. Among them, unprocessed pseudogene maintaining its exonic and intronic structure could be problematic in molecular diagnostics because the genetic regions of pseudogenes and parent genes can be co-amplified, co-enriched, and co-aligned to the regions of interest. The mutation rate of pseudogenes is higher than that of the parent genes because they are not often under selection pressure. Therefore, mutations of pseudogenes can misguide the diagnostic approaches to the genetic disorders. Excluding out these false positive mutations with tiny differences ('nuances') in their flanking regions is challenging, and, sometimes, missed for geneticists, and, unfortunately, can be reported as a pathogenic variation to clinicians ('nuisances').

Here, we discuss the risk of misdiagnosis when investigating genes with pseudogene counterparts of high homology, and we describe the method of identifying these small and annoying differences between parent genes and pseudogenes, including parent gene-specific assay design.

GENOMIC CONTEXT AND PSEUDOGENES

According to the current GENCODE release (version 41, 01. 2022.) that aims to annotate all genetic features at genome-wide level, human genome contains total 61,852 genes including 19,370 protein-coding genes (31%), 19,095 long non-coding RNA genes, 7,566 small non-coding RNA genes, and 14,736 pseudogenes (24%) [21]. Among pseudogenes, processed pseudogenes (10,662, 72%) comprise over two-thirds.

Pseudogenes could be identified by several characteristics such as the absence of introns, truncating mutations, and the absence of transcription. Although most pseudogenes are non-functioning, increasing numbers of pseudogenes have been demonstrated to play some biological roles.

Processed pseudogenes rarely have regulatory elements, thus lack transcriptional activities. *PGK2* (chr6:49,785,660-49,787,285) is a processed pseudogene of *phosphoglycerate kinase 1* (*PGK1*, chrX:77,910,739-78,129,295) via retrotransposition, but, it is expressed specifically in testis, encoding a functional phosphoglycerate kinase that catalyzes the reversible conversion of 1,3-bisphophoglycerate to 3-phophoglycerate, during spermatogenesis [22]. Some processed pseudogenes have intact open reading frames to be able to encode proteins, and play a biological role [23-26].

Unprocessed pseudogenes that arise via duplication preserve regulatory elements and intronic structures, but they are usually non-functional by disruption like truncation mutations. *NOTCH2* gene (chr1:119,911,553,120,100,779) has 4 identical paralogs (*NOTCH2NLA*, *NOTCH2NLB*, *NOTCH2NLC*, and *NOTCH2NLR*) via segmental duplications. *NOTCH2NLR* is likely to be non-functional, but other 3 *NOTCH2NL*s are expressed throughout corticogenesis [27]. Interestingly, *NOTCH2NLC* gene (chr1:149,390,621-149,471,833) is expressed significantly with age [28]. Recently, a heterozygous trinucleotide repeat expansion in the 5'-untranslated regions of the NOTCH2NLC gene was reported to be a causative mutation of neuronal intranuclear inclusion disease (NIID) [29, 30].

MISLEADING AND CHALLENGING ASPECT OF SMALL DIFFERENCE IN DIAGNOSTICS

DNA sequencing by Sanger method or recent next-generation sequencing (NGS) technology is a crucial step for genetic diagnosis. This strategy can be hampered when there are homology regions with target regions of interest, mostly those are pseudogenes. Mutation data can be erroneous, especially when we analyze target genes with highly homologous unprocessed pseudogene counterparts because segmental duplications can be indistinguishable from their parent region if a laboratory is using short-read sequencing regardless of clinical or research settings.

Autosomal dominant polycystic kidney disease (ADPKD) is a common hereditary kidney disease caused by mutations in *PKD1* and *PKD2*. *PKD1* has 6 homologous pseudogenes (*PKD1P1-P6*) located between 13 and 16 Mb distant to *PKD1* via intrachromosomal duplication. Besides high GC contents and absence of hot spots in *PKD1* gene. high sequence homology (98-99%) of pseudogenes complicate the molecular diagnosis of ADPKD [31].

Gaucher disease is an autosomal recessive lysosomal storage disorder caused by mutations of the *GBA* gene, encoding the lysosomal enzyme acid β-glucosidase. Heterozygous *GBA* mutations are the main genetic risk factor for late-onset Parkinson disease. *GBA* also has a highly homologous (96-98%) pseudogene (*GBAP1*) located approximately 16 kb distant to *GBA* [32]. Thus, detecting a pathogenic variant in *GBA* is challenging with short-read sequencing because of alignment issues [33].

Filamin myopathy is a neuromuscular disorder caused by mutations of *FLNC* gene, encoding filamin C, muscle-specific filamin isoform cross-linking actin. *FLNC* has a highly homologous pseudogene (*pseFLNC*) located 53.6 kb distant to *FLNC*, which is 98% homologous to exons 46, 47 and 48 of the parent gene [34]. Genetic studies for *FLNC* had been erroneous for years until Odgerel et al. [35] discerned the misidentification of *FLNC* mutations and suggested an optimized strategy [36].

A recent study revealed recurrent GNAQ mutation encoding T96S in natural-killer/T cell lymphoma using NGS technologies [37]. However, mutations in this study was revealed to be misaligned to *GNAQ* instead of *GNAQP1*, a processed pseudogene of *GNAQ* [38]. When analyzing somatic mutations using fragmented DNAs, even processed pseudogenes can be problematic.

In the era of NGS technologies, pseudogenes can be still troublesome because short-read sequencing is more susceptible to mis-aligning of homologous sequences than conventional Sanger sequencing. If sequence reads of 150–250 bp containing a pseudogene-derived mutation are mapped to the parent gene, it will generate a false positive result. If sequence reads containing the parent gene-derived variant are mapped to the pseudogene, it will produce a false negative result.

IDENTIFYING SMALL DIFFERENCE BETWEEN PARENT GENE AND PSEUDOGENE

When we devise a diagnostic assay to find a sequence variation in genetic regions of interest with highly homologous pseudogenes, we should confirm that the only desired genetic regions are amplified/enriched, aligned, and analyzed. The first step to do this is to identify small differences in sequence between parent gene and pseudogenes. The BLAST-Like Alignment Tool (BLAT) could be a good starting point to identify small differences [39]. BLAT is typically used to search similar sequence within the same species. Therefore, it is convenient for the identification of discrepant sites between two highly homologous genetic regions.

For example, Fig. 1 demonstrates how to identify the differences between the parent gene (*FLNC*) and pseudogene (*pseFLNC*) in the region of interest. We can annotate the discrepant sequence between parent gene and pseudogene upon the patient's sequencing reads to discriminate whether a significant variation is from parent gene or pseudogene.

NEED FOR ORTHOGONAL METHODS: PARENT GENE-SPECIFIC ASSAY

When we cannot discriminate the significant variations from pseudogene-specific alterations in sequencing analysis, we need to find the other orthogonal methods, i.e., parent gene-specific assay. Mostly, we can adopt allele-specific amplification and subsequent Sanger sequencing that can achieve longer sequencing results than NGS. However, sometimes, we need additional separation or enrichment methods, such as cloning. We should manually design all primers and long-range PCR when confirming variants in regions with high homology and devising a new diagnostic assay for an already known gene with highly homologous pseudogenes. Fig. 2 shows the example of parent gene (*CEL*)-specific assay design and its application in Sanger sequencing. Using the parent gene-specific primer, we can exclude out pseudogene-derived sequence variations.

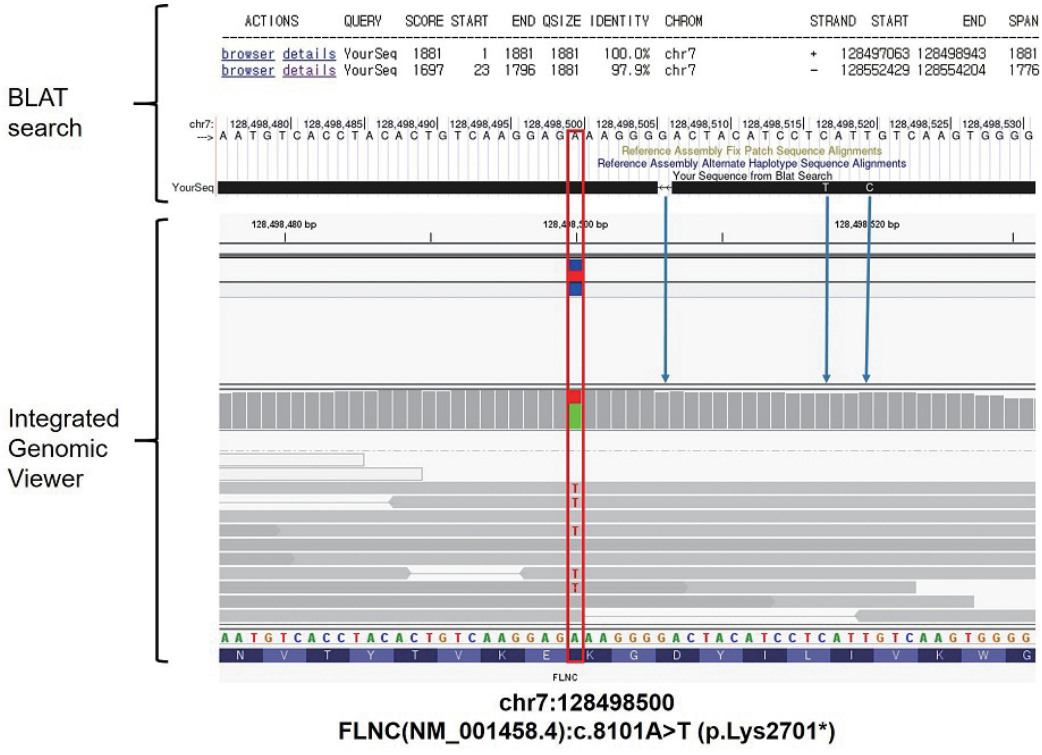

Fig. 1. Identifying differences in sequence between the parent gene and pseudogene using the BLAT search tool and genomic viewers from sequencing results. Through the BLAT search, similarity (97.9%) between *FLNC* exon 46–48 regions (chr7:128,497,063-128,498,943) and *pseFLNC* (chr7:128,552,429-128,554,204) was calculated. Patient's sequence reads are displayed in the Integrated Genomic Viewer (IGV). The red rectangle indicates a pathogenic variation of patient. Blue arrows show the discrepant sequences between *FLNC* and *pseFLNC*.

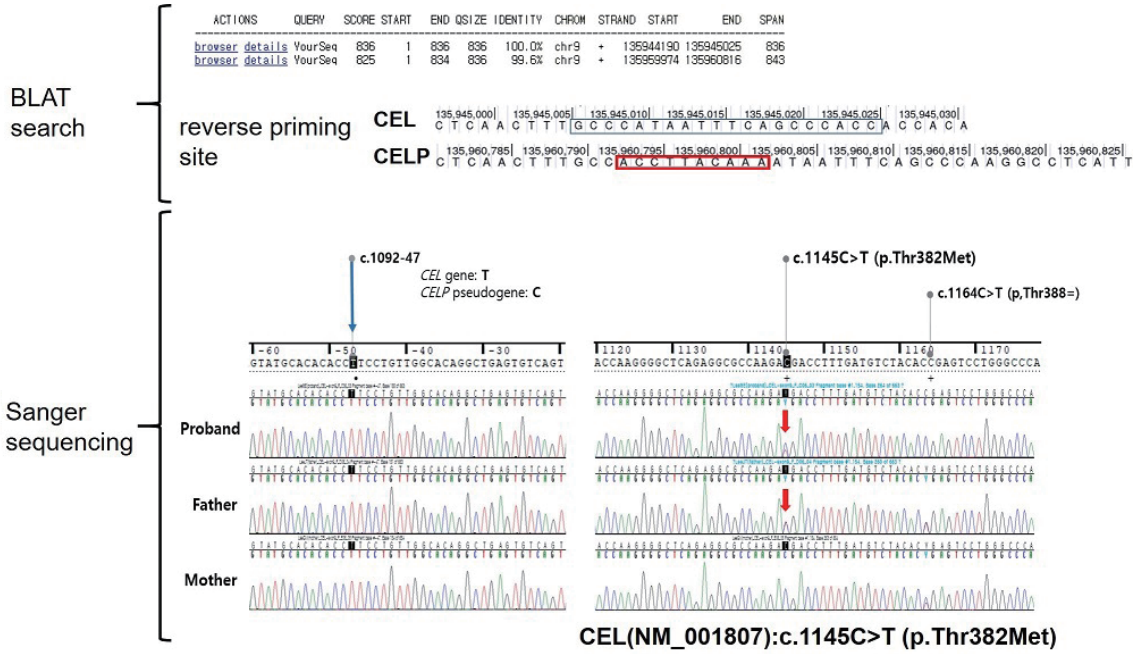

Fig. 2. Design of the parent gene-specific primer and its amplification. In the BLAT search, reverse primer binds parent gene (*CEL*)-specific region (blue rectangle). The red rectangle indicates discrepant sequences of pseudogene counterpart (*CELP*) with this reverse primer. In the Sanger reaction, the amplified product revealed parent gene-specific (c.1092-47T) results. Therefore, c.1145C>T variant comes from the parent gene.

CONCLUSION

Even in the era of NGS, specific attention and considerations to discriminate pseudogenes from parent genes are needed. Investigators and clinicians should be aware of the possibility of false positive and/or false-negative results due to highly homologous pseudogenes. Laboratorians and researchers should be prepared for identifying small differences between parent gene- and pseudogene-derived sequences, and designing the parent gene-specific assays.

ACKNOWLEDGEMENTS

None.

FINANCIAL DISCLOSURE STATEMENT

This review did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICTS OF INTEREST

None.

REFERENCES

1. Balakirev ES, Ayala FJ. Pseudogenes: are they "junk" or functional DNA?. Annu Rev Genet 2003;37:123-51.
2. Zhang Z, Harrison P, Gerstein M. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. Genome Res 2002;12:1466-82.
3. Harrison PM, Gerstein M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 2002;318:1155-74.
4. Balasubramanian S, Harrison P, Hegyi H, Bertone P, Luscombe N, Echols N, et al. SNPs on human chromosomes 21 and 22 - analysis in terms of protein features and pseudogenes. Pharmacogenomics 2002;3:393-402.
5. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, Zhang Z, et al. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. Nucleic Acids Res 2002;30:2515-23.
6. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, et al. Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. Genome Res 2002;12:272-80.
7. Muro EM, Mah N, Andrade-Navarro MA. Functional evidence of post-transcriptional regulation by pseudogenes. Biochimie 2011;93:1916-21.
8. Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DR. Pseudogenes: pseudo-functional or key regulators in health and

disease? RNA 2011;17:792-8.
9. Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell 2011;146:353-8.
10. Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. Characterization of human pseudogene-derived non-coding RNAs for functional potential. PLoS One 2014;9:e93972.
11. Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature 2008;453:534-8.
12. Chen B, Wang C, Zhang J, Zhou Y, Hu W, Guo T. New insights into long noncoding RNAs and pseudogenes in prognosis of renal cell carcinoma. Cancer Cell Int 2018;18:157.
13. Hu X, Yang L, Mo YY. Role of Pseudogenes in Tumorigenesis. Cancers (Basel) 2018;10(8):256.
14. Jiang T, Guo J, Hu Z, Zhao M, Gu Z, Miao S. Identification of Potential Prostate Cancer-Related Pseudogenes Based on Competitive Endogenous RNA Network Hypothesis. Med Sci Monit 2018;24:4213-39.
15. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 2012;22:1760-74.
16. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, et al. The GENCODE pseudogene resource. Genome Biol 2012;13:R51.
17. Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet 2020;21:191-201.
18. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 2010;11:R26.
19. Mighell AJ, Smith NR, Robinson PA, Markham AF. Vertebrate pseudogenes. FEBS Lett 2000;468:109-14.
20. Ohshima K, Hattori M, Yada T, Gojobori T, Sakaki Y, Okada N. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol 2003;4:R74.
21. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 2019;47:D766-73.
22. Vemuganti SA, de Villena FP, O'Brien DA. Frequent and recent retrotransposition of orthologous genes plays a role in the evolution of sperm glycolytic enzymes. BMC Genomics 2010;11:285.
23. Sayah DM, Sokolskaja E, Berthoux L, Luban J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature 2004;430:569-73.
24. Burki F, Kaessmann H. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet 2004;36:1061-3.
25. Hayashi H, Arao T, Togashi Y, Kato H, Fujita Y, De Velasco MA, et al. The OCT4 pseudogene POU5F1B is amplified and promotes an aggressive phenotype in gastric cancer. Oncogene 2015;34:199-208.
26. Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci U S A 2006;103:3220-5.
27. Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, et al. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 2018;173:1356-69.e13.
28. Tian Y, Wang JL, Huang W, Zeng S, Jiao B, Liu Z, et al. Expansion of Human-Specific GGC Repeat in Neuronal Intranuclear Inclusion Disease-Related Disorders. Am J Hum Genet 2019;105:166-76.
29. Okubo M, Doi H, Fukai R, Fujita A, Mitsuhashi S, Hashiguchi S, et al. GGC Repeat Expansion of NOTCH2NLC in Adult Patients with Leukoencephalopathy. Ann Neurol 2019;86:962-8.
30. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat Genet 2019;51:1215-21.
31. Liu W, Chen M, Wei J, He W, Li Z, Sun X, et al. Modification of PCR conditions and design of exon-specific primers for the efficient molecular diagnosis of PKD1 mutations. Kidney Blood Press Res 2014;39:536-45.
32. Sorge J, Gross E, West C, Beutler E. High level transcription of the glucocerebrosidase pseudogene in normal subjects and patients with Gaucher disease. J Clin Invest 1990;86:1137-41.
33. Zampieri S, Cattarossi S, Bembi B, Dardis A. GBA Analysis in Next-Generation Era: Pitfalls, Challenges, and Possible Solutions. J Mol Diagn 2017;19:733-41.
34. van der Ven PF, Odgerel Z, Furst DO, Goldfarb LG, Kono S, Miyajima H. Dominant-negative effects of a novel mutation in the filamin myopathy. Neurology 2010;75:2137-8.
35. Odgerel Z, van der Ven PF, Furst DO, Goldfarb LG. DNA sequencing errors in molecular diagnostics of filamin myopathy. Clin Chem Lab Med 2010;48:1409-14.
36. Kono S, Nishio T, Takahashi Y, Goto-Inoue N, Kinoshita M, Zaima N, et al. Dominant-negative effects of a novel mutation in the filamin myopathy. Neurology 2010;75:547-54.
37. Li Z, Zhang X, Xue W, Zhang Y, Li C, Song Y, et al. Recurrent GNAQ mutation encoding T96S in natural killer/T cell lymphoma. Nat Commun 2019;10:4209.
38. Lim JQ, Lim ST, Ong CK. Misaligned sequencing reads from the GNAQ-pseudogene locus may yield GNAQ artefact variants. Nat Commun 2022;13:458.
39. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res 2002;12:656-64.