Clinical Pediatric Endocrinology

Vol.29 / No.3
July 2020
pp 123–126

Short Communication

Predicting the pathogenicity of *NKX2-1* and *IGSF1* variants with in silico bioinformatic tools

Satoshi Narumi[1]

[1]*Department of Molecular Endocrinology, National Research Institute for Child Health and Development, Tokyo 157-8535, Japan*

Key words: congenital hypothyroidism, *NKX2-1*, *IGSF1*, mutation, genetics, in silico

Introduction

Congenital hypothyroidism (CH) is the most common congenital endocrine disorder, affecting 1 in approximately 2,000–3,000 newborns worldwide (1). Systematic genetic studies have revealed that about 20% of CH patients have single-gene mutations (2–4). Each genetic defect differs in inheritance patterns, complications, and clinical courses. For example, thyroid hypoplasia due to *PAX8* and *NKX2-1* mutations exhibit dominant inheritance, whereas thyroid dyshormonogenesis, such as *DUOX2* and *TG* mutations, exhibits recessive inheritance (5). Central hypothyroidism is more common among boys than among girls because *IGSF,* the most frequently mutated gene, is located on the X chromosome (i.e., X-linked inheritance) (6).

Recently, genetic analysis has become a widely used method for making an etiologic diagnosis of CH at the molecular level. As a result, many rare genetic variants are being detected in the clinical setting. For some variants including common mutations, nonsense variants, frameshift variants, and variants in essential splice sites, it is easy for clinicians to interpret their pathogenicity. In contrast, assessing the pathogenicity of missense variants could be difficult if the variants are novel, and the effect of amino acid alteration is vague. To solve this difficulty, a number of in silico bioinformatic tools have been developed. They are based on amino acid sequence homology, machine learning, or integration of the results of multiple bioinformatic tools (7). Tools based on sequence homology use the information of amino acid sequence conservation for prediction, e.g., PolyPhen-2 performs a BLAST search of mutated and non-mutated sequences in the UniRef100 database and calculates the profile score based on multiple alignment results. Tools based on machine learning use various types of variant information, such as allele frequency in the publicly available variant databases and the nature of the affected amino acid, and are trained by supervised machine learning with datasets of disease-causing variants, e.g., Human Gene Mutation Database (HGMD), and common polymorphisms. For example, VEST3 uses 86 quantitative features available through the SNVBox database. Tools that integrate results of multiple bioinformatic tools, also referred to as the ensemble method, are trained similarly to those based on machine learning. For example, CADD incorporates scores of three tools (Grantham, SIFT, and PolyPhen-2) to calculate the CADD score.

The effectiveness of each tool depends on the gene being analyzed due, in part, to the different datasets used for training the tools. Therefore, using the best tool to analyze the gene of interest would contribute to more reliable interpretations of obtained results. In this study, the performance of 13 currently available in silico tools were tested to discriminate disease-causing and non-causing variants in *NKX2-1* and *IGSF1*, two genes associated with CH (8, 9).

Materials and Methods

Selection of genes

Examining the performance of in silico tools requires a wealth of "gold standard" data of both disease-causing and non-disease-causing variants. For disease-causing variants, databases of human genetic disorders, such as HGMD, can be used, while databases for non-disease-causing variants are more problematic.

Received: January 28, 2020 Accepted: April 11, 2020
Corresponding author: Satoshi Narumi, M.D., Ph.D., Department of Molecular Endocrinology, National Research Institute for Child Health and Development, 2-10-1 Okura, Setagaya-ku, Tokyo 157-8535, Japan.
E-mail: narumi-s@ncchd.go.jp

Publicly available population databases, such as the 1000 Genomes Project and the genome aggregation database (gnomAD), can provide lists of non-disease-causing variants. However, the use of these population-derived data requires caution because disease-causing variants in genes with autosomal recessive inheritance can be included in such databases. For example, *TSHR* p.Arg450His, a common genetic mutation in Japanese individuals (2), is registered in gnomAD v2.1.1 (https://gnomad.broadinstitute.org/) with allele frequency 48 in 19,952 East Asians. Considering the possible "contamination" of truly pathogenic mutations in the population databases, only CH-related genes with dominant or X-linked inheritance were evaluated in this study. There were five candidate genes: *PAX8* (dominant), *NKX2-1* (dominant), *IGSF1* (X-linked), *TBL1X* (X-linked), and *IRS4* (X-linked). However, owing to their limited numbers of disease-causing variants (*TBL1X* and *IRS4*) or non-disease-causing variants (*PAX8*), these three genes were omitted. Therefore, *NKX2-1* and *IGSF1* were selected for analysis.

Dataset

Two types of datasets were retrieved from publicly or commercially available variant databases. For disease-causing variants, HGMD Professional Version_2019.4 (Qiagen, Hilden, Germany) was used, and there were 153 *NKX2-1* and 44 *IGSF1* distinct variants in the original dataset, including missense and other variants (e.g., nonsense, frameshift, and splice site variants). Variants other than the 38 *NKX2-1* and 15 *IGSF1* missense variants were excluded as this study aimed to evaluate the pathogenicity of only missense variants. Then the flag "disease" was referred, and five variants labeled as "neuroendocrine cell hyperplasia of infancy", "Hirschsprung disease", "multinodular goiter and papillary thyroid carcinoma", or "pituitary stalk interruption syndrome" were excluded because these phenotypes could be irrelevant to CH. As a result, 35 *NKX2-1* and 14 *IGSF1* disease-causing missense variants remained and were subject for in silico analyses.

For non-disease-causing variants, data were obtained from the gnomAD database v2.1.1. In the original dataset, there were 185 *NKX2-1* and 493 *IGSF1* variants, including missense and other variants. First, 162 *NKX2-1* and 453 *IGSF1* variants with observed allele number ≤ 4 were excluded. In the gnomAD database, each variant has been sequenced in more than 100,000 individuals. Thus, "allele number of five or more" corresponds to > 1/20,000 frequency of the variant carriers. This threshold was set based on the frequency of the most common genetic form of CH, the *DUOX2* defect, with about 1/20,000 frequency (10), considering that the *NKX2-1* defect and the *IGSF1* defect are less frequent than the *DUOX2* defect. For the *IGSF* (X-linked inheritance), male-limited allele count data were referred. After the exclusion of one *NKX2-1* in-frame deletion variant, 22 *NKX2-1* and 40 *IGSF1* non-disease-causing missense variants were selected for analyses.

In silico analyses

A total of 57 *NKX2-1* variants (35 disease-causing; 22 non-disease-causing) and 54 *IGSF1* variants (14 disease-causing; 40 non-disease-causing) were subject to analyses with 13 in silico bioinformatic tools, including CADD (https://cadd.gs.washington.edu/), DANN (https://omictools.com/dann-tool), FATHMM (http://fathmm.biocompute.org.uk/), FATHMM-MKL (http://fathmm.biocompute.org.uk/fathmmMKL.htm), GenoCanyon (https://omictools.com/genocanyon-tool), MetaLR (https://m.ensembl.org/info/genome/variation/prediction/protein_function.html), MetaSVM (https://omictools.com/meta-svm-tool), MutationTaster (http://www.mutationtaster.org/), REVEL (https://sites.google.com/site/revelgenomics/), PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/), PROVEAN (http://provean.jcvi.org/), SIFT (https://sift.bii.a-star.edu.sg/), and VEST3 (https://karchinlab.org/apps/appVest.html). These were classified into three categories based on the principal methods for development: (i) sequence homology (FATHMM, FATHMM-MKL, MutationTaster, PolyPhen-2, PROVEAN, and SIFT), (ii) machine learning (DANN, GenoCanyon, and VEST3), and (iii) ensemble (CADD, MetaLR, MetaSVM, and REVEL).

To quantitatively measure the performance of the 13 tools, Receiver Operating Characteristic (ROC) curve analyses were performed under the hypothesis that the HGMD missense variant dataset included true disease-causing variants, and the gnomAD missense variant dataset included true non-disease-causing variants.

Results

ROC curve analyses of the 13 in silico tools revealed that their performance was not uniform (**Fig. 1**). Area under the curve (AUC) values of the ROC curve varied, ranging from less than 0.6 to more than 0.9. The average values of AUC were 0.79 and 0.81 for *NKX2-1* variants and *IGSF1* variants, respectively (**Fig. 2A**).

When the AUC values were compared between *NKX2-1* and *IGSF1* variants, there was no significant correlation of AUC values between the two (**Fig. 2B**). Four tools (MetaLR, PROVEAN, REVEL, and VEST3) showed AUC values of more than 0.85 for both gene variants. No tool had AUC values less than 0.7 for both gene variants. There was no correlation between the performance (AUC for the two gene variants) and the types of tools (sequence homology, machine learning, or ensemble).

Discussion

In the present study, 13 publicly available in silico tools were tested for the discrimination of disease-causing and non-disease-causing variants in *NKX2-1*

doi: 10.1297/cpe.29.123

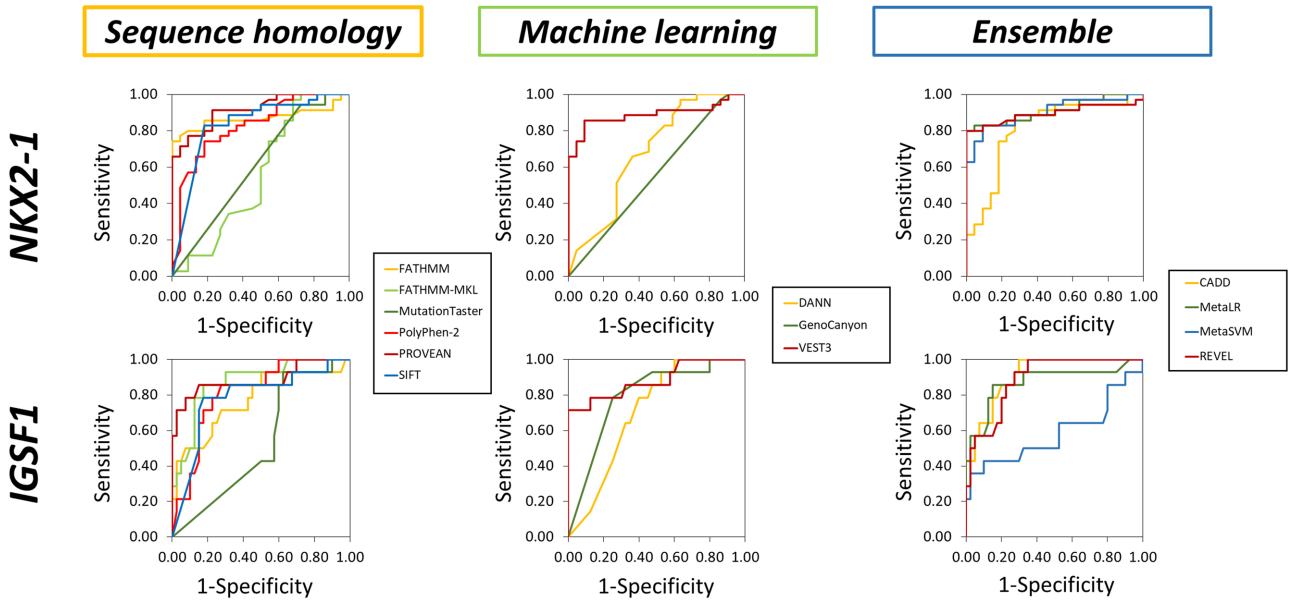

Fig. 1. Receiver Operating Characteristic (ROC) curve analyses of 13 in silico bioinformatic tools. The 13 tools were classified into three groups (sequence homology, machine learning and ensemble) according to the principal method of development. Each ROC curve is shown in the three groups.

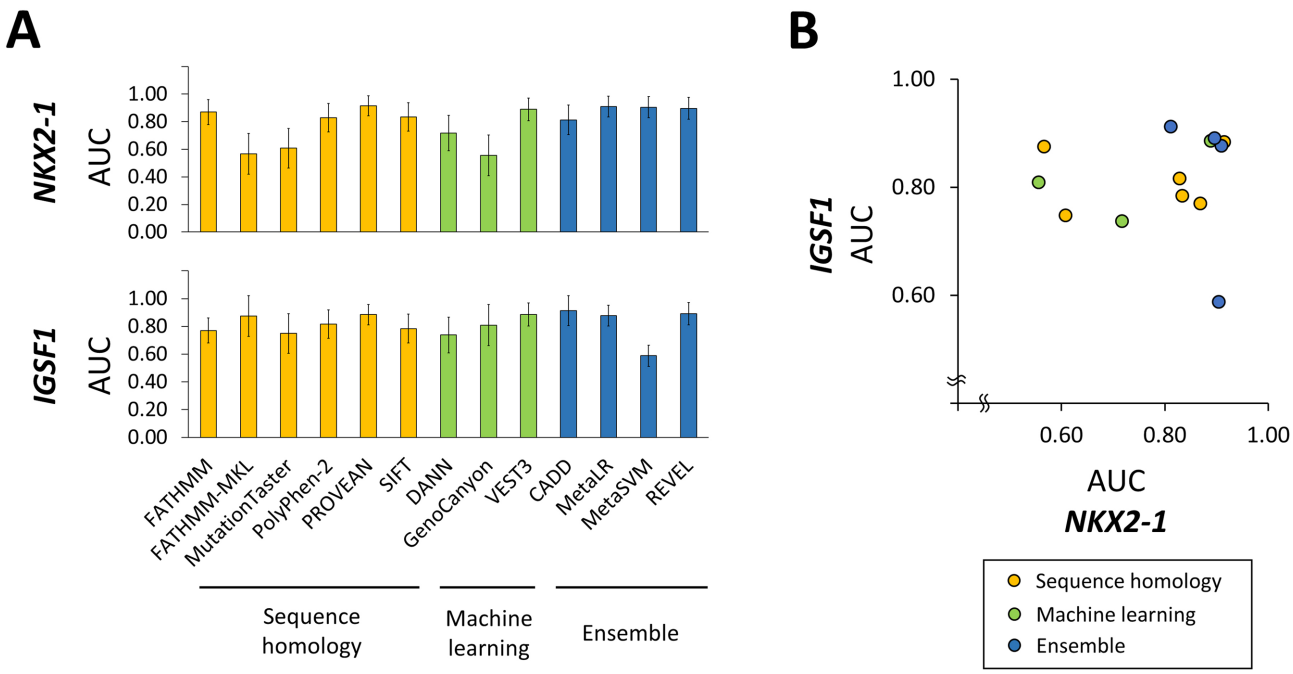

Fig. 2. Comparison of the 13 in silico tools. A) Values of area under the curve (AUC) of the ROC curves for the two genes (*NKX2-1* and *IGSF1*) are shown. Bars indicate 95% confidence intervals. B) A scatter plot showing the relationship between AUC of ROC for *NKX2-1* variants (horizontal axis) and *IGSF1* variants (vertical axis). No significant correlation was observed between AUC values for *NKX2-1* and ones for *IGSF1*.

and *IGSF1*. Interestingly, tool performance did not correlate between *NKX2-1* and *IGSF1* variants. This illustrates the existence of "tool preference" for genes. Comparisons of multiple in silico tools have been attempted with the use of various variant datasets. Mahmood K *et al.* compared the accuracy of eight in silico tools (GERP++, fitCons, SIFT, PolyPhen-2, CADD, Condel, REVEL, and FATHMM) for the prediction of seven deleterious/benign variant datasets and reported that their accuracy was highly variable (7). Excluding the nature of tools themselves, there are two explanations for their performance variability. First, for tools based on machine learning and the ensemble method, their training datasets included a subset of

doi: 10.1297/cpe.29.123

erroneously classified variants because the curation of disease-causing variants requires the consideration of many factors (e.g., frequencies in the patient cohort and general population, mode of inheritance and presumed functional impact), and thus these tools were prone to human error. Second, the variant datasets used to measure performance accuracy could also include erroneously classified variants due to the same reason. At present, there seems to be no rational approach to avoid the contamination of the variant databases. In this study, datasets for disease-causing variants and non-causing variants were carefully selected mainly by restricting genes to ones with autosomal dominant or X-linked inheritance. In silico tool users should be aware that no tool is optimal for all genes. Notably, results of in silico tools will not increase the evidence levels of variants regardless of how many tools are used.

Causative genes for autosomal dominant disorders are known to be depleted for loss of function variants compared with causative genes for autosomal recessive disorders. This situation makes the finding of non-disease-causing variants for genes associated with autosomal dominant genetic disorders feasible. However, finding those for genes associated with autosomal recessive disorders is more complicated because of true disease-causing variants with relatively high allele frequency (e.g., 1 in 200–1,000) can be found among the general population. Creating high-quality datasets of disease-causing and non-causing variants in these autosomal recessive genes is currently virtually impossible without performing functional assays.

In this study, the best tools that predicted the pathogenicity of variants in *NKX2-1* and *IGSF1* were MetaLR (ensemble method), PROVEAN (sequence homology method), REVEL (ensemble method), and VEST3 (machine learning method). Results on the diversity of development methods suggest that the performance of in silico tools might be determined by complex factors, including datasets of protein sequence used, datasets of human genome variants used, combinations of tools in the ensemble method, and computational methods (e.g., logistic regression and machine learning). This finding also indicates the difficulty in predicting which tool works satisfactorily for analyzing the gene of interest.

Conclusion

The performance of 13 currently available in silico bioinformatic tools was tested using real-world variant data, and their performance varied depending on the gene analyzed. For *NKX2-1* and *IGSF1* variants, the MetaLR, PROVEAN, REVEL, and VEST3 tools performed best. Further studies are needed to evaluate the performance of these tools on genes associated with autosomal recessive diseases.

Conflict of Interests: The authors have nothing to disclose.

Acknowledgments

This work was supported by the JSPS KAKENHI grant number 19K22607.

References

1. Ford G, LaFranchi SH. Screening for congenital hypothyroidism: a worldwide view of strategies. Best Pract Res Clin Endocrinol Metab 2014;28: 175–87. [Medline] [CrossRef]
2. Narumi S, Muroya K, Abe Y, Yasui M, Asakura Y, Adachi M, *et al*. TSHR mutations as a cause of congenital hypothyroidism in Japan: a population-based genetic epidemiology study. J Clin Endocrinol Metab 2009;94: 1317–23. [Medline] [CrossRef]
3. Narumi S, Muroya K, Asakura Y, Adachi M, Hasegawa T. Transcription factor mutations and congenital hypothyroidism: systematic genetic screening of a population-based cohort of Japanese patients. J Clin Endocrinol Metab 2010;95: 1981–5. [Medline] [CrossRef]
4. Narumi S, Muroya K, Asakura Y, Aachi M, Hasegawa T. Molecular basis of thyroid dyshormonogenesis: genetic screening in population-based Japanese patients. J Clin Endocrinol Metab 2011;96: E1838–42. [Medline] [CrossRef]
5. Peters C, van Trotsenburg ASP, Schoenmakers N. DIAGNOSIS OF ENDOCRINE DISEASE: Congenital hypothyroidism: update and perspectives. Eur J Endocrinol 2018;179: R297–317. [Medline] [CrossRef]
6. Sugisawa C, Takamizawa T, Abe K, Hasegawa T, Shiga K, Sugawara H, *et al*. Genetics of congenital isolated TSH deficiency: mutation screening of the known causative genes and a literature review. J Clin Endocrinol Metab 2019;104: 6229–37. [Medline] [CrossRef]
7. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, *et al*. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics 2017;11: 10. [Medline] [CrossRef]
8. Devriendt K, Vanhole C, Matthijs G, de Zegher F. Deletion of thyroid transcription factor-1 gene in an infant with neonatal thyroid dysfunction and respiratory failure. N Engl J Med 1998;338: 1317–8. [Medline] [CrossRef]
9. Sun Y, Bak B, Schoenmakers N, van Trotsenburg AS, Oostdijk W, Voshol P, *et al*. Loss-of-function mutations in IGSF1 cause an X-linked syndrome of central hypothyroidism and testicular enlargement. Nat Genet 2012;44: 1375–81. [Medline] [CrossRef]
10. Abe K, Narumi S, Suwanai AS, Adachi M, Muroya K, Asakura Y, *et al*. Association between monoallelic *TSHR* mutations and congenital hypothyroidism: a statistical approach. Eur J Endocrinol 2018;178: 137–44. [Medline] [CrossRef]

doi: 10.1297/cpe.29.123