Медицинская генетика. **2024. Том 23. Номер 12**
***Medical genetics* 2024. Vol. 23. Issue 12**

https://doi.org/10.25557/2073-7998.2024.12.16-21

Прогнозирование патогенности миссенс-мутаций в гене TCF4

Государкина С.Н., Савченко Р.Р., Скрябин Н.А.

ФГБНУ Томский национальный исследовательский медицинский центр Российской академии наук,
Научно-исследовательский институт медицинской генетики
634050, г. Томск, ул. Набережная реки Ушайки, д. 10

Введение. Подавляющее большинство обнаруженных на данный момент миссенс-вариантов имеет неизвестное клиническое значение. В связи с этим классификация таких вариантов является актуальной проблемой медицинской генетики, поскольку невозможность установить клиническую значимость варианта затрудняет диагностику наследственных болезней, а также разработку или применение существующих терапевтических стратегий. В данной работе использован новый биоинформатический инструмент AlphaMissense для оценки эффективности классификации вариантов в гене *TCF4*.
Цель: прогнозирование патогенных эффектов всех возможных миссенс-вариантов в гене *TCF4* с помощью инструмента AlphaMissense, основанного на машинном обучении, и оценка способности классификации вариантов данным инструментом с использованием ROC-анализа.
Методы. Для создания и анализа данных, рассматриваемых в работе, были использованы среда разработки Google Colab, язык программирования Python v3.10, библиотеки Biopython для работы с биологическими последовательностями, scikit-learn для проведения ROC-анализа. В качестве референса была использована последовательность гена *TCF4* из геномной сборки версии GRCh38.p14 (транскрипт NM_001083962.2), содержащаяся в базе данных NCBI. Были созданы 1241319 вариантов однонуклеотидных полиморфизмов (SNP), из которых 6906 вариантов находятся в кодирующей последовательности, из них 3747 были определены, как миссенс-варианты. Аннотация полученных данных производилась по базам данных ClinVar и AlphaMissense с использованием инструмента OpenCRAVAT. Из всех обнаруженных миссенс-вариантов оценку AlphaMissense получили 979 варианта, из которых всего 101 вариант был указан в базе данных ClinVar.
Результаты. При сравнении показателей чувствительности (Se), специфичности (Sp), а также графиков ROC-кривых и значений показателей площади под кривой (AUC) явное отличие имеет оценка классификации SNP, как вероятно патогенных (AUC = 0,81, Se = 0,68, Sp = 0,78). Она может быть использована как дополнительный критерий при определении клинической значимости вариантов в диагностике синдрома Питта-Хопкинса. И напротив, классификация вариантов как вероятно доброкачественных или неопределенных не обладает достаточными чувствительностью и специфичностью, а показатели AUC характеризуют их как модели со средним качеством. Таким образом, варианты, вошедшие в эти группы, требуют дополнительной переоценки другими инструментами.
Заключение. Измеренные показатели показывают, что лучше всего инструмент AlphaMissense определяет вероятно патогенные варианты. Однако стоит с сомнением относиться к вариантам, определенным как вероятно доброкачественные или неопределенные и делать проверку с использованием других инструментов. Варианты, полученные в ходе искусственного мутагенеза и оцененные как вероятно патогенные, но не указанные в базах данных, могут быть полезны при определении ранее неизвестных вариантов в гене *TCF4* и помочь в диагностике и разработке терапии ассоциированных заболеваний.

Ключевые слова: миссенс-варианты, биоинформатика, ROC-анализ, TCF4, синдром Питта-Хопкинса.

Для цитирования: Государкина С.Н., Савченко Р.Р., Скрябин Н.А. Прогнозирование патогенности миссенс-мутаций в гене *TCF4*. *Медицинская генетика* 2024; 23(12): 16-21.

Автор для корреспонденции: Государкина С.Н.; **e-mail:** sophia.gosudarkina@medgenetics.ru
Финансирование. Работа выполнена при поддержке гранта РНФ № 23-75-01138.
Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.
Поступила: 03.12.2024

Predicting the pathogenicity of missense mutations in the TCF4 gene

Gosudarkina S.N., Savchenko R.R., Skryabin N.A.

Tomsk National Research Medical Center of the Russian Academy of Sciences, Research Institute of Medical Genetics
10, Naberejnaya Ushaiki, Tomsk, 634050, Russian Federation

Background. The vast majority of currently discovered missense variants have unknown clinical significance. In this regard, the classification of such variants is an urgent problem of medical genetics, since the inability to establish the clinical significance of a variant complicates the diagnosis of inherited diseases, as well as the development or application of existing therapeutic strategies. In this work, a new bioinformatics tool AlphaMissense was used to assess the efficiency of variant classification in the *TCF4* gene.

Aim: prediction of the pathogenic effect of all possible missense variants in the TCF4 gene using the AlphaMissense tool based on machine learning, and evaluation of the ability to classify variants by this tool using ROC analysis.
Methods. The following were used to create and analyse the data discussed in this paper: Google Colab development environment, Python v3.10 programming language, Biopython library for working with biological sequences, scikit-learn library for ROC analysis. The *TCF4* gene sequence contained in the NCBI database was used as a reference. 1241319 single nucleotide polymorphism (SNP) variants were generated, among which 6906 variants are in the coding sequence, of which 3747 were identified as missense variants. Annotation of the obtained data was performed according to ClinVar and AlphaMissense databases using the OpenCRAVAT tool. Of all the detected missense variants, 979 variants were scored by AlphaMissense, of which only 101 variants were reported in the ClinVar database.
Results. When comparing sensitivity (Se), specificity (Sp), ROC curve plots and area under the curve (AUC) values, there is a clear difference in the evaluation of SNP classification as likely pathogenic (AUC = 0.81, Se = 0.68, Sp = 0.78). It can be used as an additional criterion in screening of candidate variants for Pitt-Hopkins syndrome. In contrast, classifying variants as likely benign or ambiguous lacks sensitivity and specificity, and their AUC scores characterise them as models of medium quality. Therefore, the variants included in these groups require further reassessment by other tools.
Conclusions. The measured values make it evident that the AlphaMissense tool is best at identifying likely pathogenic variants. However, variants identified as likely benign or ambiguous should be considered questionable and should be tested with other tools. Variants obtained by artificial mutagenesis and assessed as likely pathogenic but not listed in databases may be useful in identifying previously unknown variants in the *TCF4* gene and help in the diagnosis and development of therapies for associated diseases.

Keywords: missense mutations, bioinformatics, ROC analysis, *TCF4* gene, Pitt-Hopkins syndrome.

For citation: Gosudarkina S.N., Savchenko R.R., Skryabin N.A. Predicting the pathogenicity of missense mutations in the *TCF4* gene. *Medical genetics [Medicinskaya genetika].* 2024; 23(12): 16-21. (In Russian).

Corresponding author: Sophia N. Gosudarkina; **e-mail:** sophia.gosudarkina@medgenetics.ru
Funding. The study was supported by the Russian Science Foundation grant No. 23-75-01138.
Conflict of Interest. The authors declare no conflict of interest.
Accepted: 03.12.2024

Введение

Активное развитие технологий секвенирования генома позволило выявить обширную генетическую изменчивость в популяциях человека. Миссенс-мутации в генах могут приводить к неправильному формированию третичной структуры белка, и, как следствие, к нарушениям функции или стабильности белков. При этом подавляющее большинство обнаруженных миссенс-вариантов имеет неизвестное клиническое значение (2% от 4 млн вариантов в базах данных) [1], что затрудняет диагностику и разработку терапевтических стратегий для лечения болезней, передающихся по наследству. В связи с этим классификация и интерпретация таких вариантов является актуальной проблемой медицинской генетики, которую могут помочь решить подходы машинного обучения, использующие паттерны и закономерности в биологических данных для предсказания патогенности ранее не аннотированных вариантов.

В данной работе использован новый биоинформатический инструмент AlphaMissense [1] для оценки эффективности классификации вариантов в гене *TCF4*. Данный ген кодирует фактор транскрипции TCF4, имеющий критическое значение для развития и функционирования мозга [2]. Патогенные варианты в гене *TCF4* служат причиной развития синдрома Питта-Хопкинса [2]. Более того, ряд вариантов ассоциирован с развитием социально значимых заболеваний, таких как шизофрения, биполярное расстройство, большое депрессивное расстройство и посттравматическое стрессовое расстройство [3-5]. Поиск и классификация вариантов в гене *TCF4* имеют большое значение для понимания патогенеза ассоциированных с данным геном заболеваний.

Таким образом, целью исследования стало прогнозирование патогенных эффектов всех возможных SNP в гене *TCF4* с помощью инструмента AlphaMissense, основанного на машинном обучении, и оценка способности классификации данного инструмента.

Методы

Для создания данных, анализируемых в работе, были использованы среда разработки Google Colab и язык программирования Python версии 3.10 с применением библиотеки для обработки биологических данных Biopython [6]. В качестве референса была использована последовательность гена *TCF4*, содержащаяся в базе данных NCBI [7]. Алгоритм представляет собой инкрементирующий перебор референсной последовательности, сравнение со сло-

варем нуклеотидов и запись трех возможных однонуклеотидных замен в vcf-файл для дальнейшего анализа. Таким образом были созданы 1241319 SNP.

Аннотация полученных данных производилась при помощи инструмента с открытым исходным кодом OpenCRAVAT [8] с использованием модулей ClinVar [9] и AlphaMissense [1].

AlphaMissense – это адаптация алгоритма AlphaFold2 [10], обученная на базе данных о частоте встречаемости вариантов в популяциях людей и приматов для предсказания патогенности миссенс-вариантов. Согласно литературным источникам, AlphaMissense входит в пятерку лучших алгоритмов предсказания тяжести миссенс-мутаций при корреляции с функциональным влиянием. Также при изучении предсказания для генов, связанных с умственной отсталостью (*DDX3X*), онкогенов (*MSH2, PTEN, BRCA1*) и генов, связанных с прогрессирующей потерей слуха (*KCNQ4*), является лучшим алгоритмом для двух из пяти (*PTEN* и *BRCA1*). Кроме того, он превосходит по показателям инструменты, используемые в стандартных протоколах биоинформатической обработки (ClinPred, CADD, PolyPhen-2, PROVEAN) [11].

Из всех созданных вариантов оценку AlphaMissense получили 979, из которых всего 101 был указан в базе данных ClinVar. ROC-анализ для оценки эффективности классификации вариантов в гене *TCF4* алгоритмом AlphaMissense производился с помощью библиотеки для машинного обучения scikit-learn [12]. Во избежание смещения результатов, данные из выборки с известным классом патогенности по ClinVar выбирались в случайном порядке.

Использование ROC-анализа заключается в его применении к задачам бинарной классификации и отнесении объекта (в данном случае оценки патогенности миссенс-варианта) к одному из двух классов, обозначаемых как «истинный» (true, 1) и «ложный» (false, 0). Параметры алгоритма такого классификатора определяются в результате обучения на известных истинных и ложных примерах, затем классификатор тестируется на примерах, не вошедших в тестовую выборку.

ROC-кривая (Receiver Operator Characteristic) – кривая, которая наиболее часто используется для представления результатов бинарной классификации в машинном обучении. Ось X данного графика представляет собой измерения False Positive Rate (ложнопо-

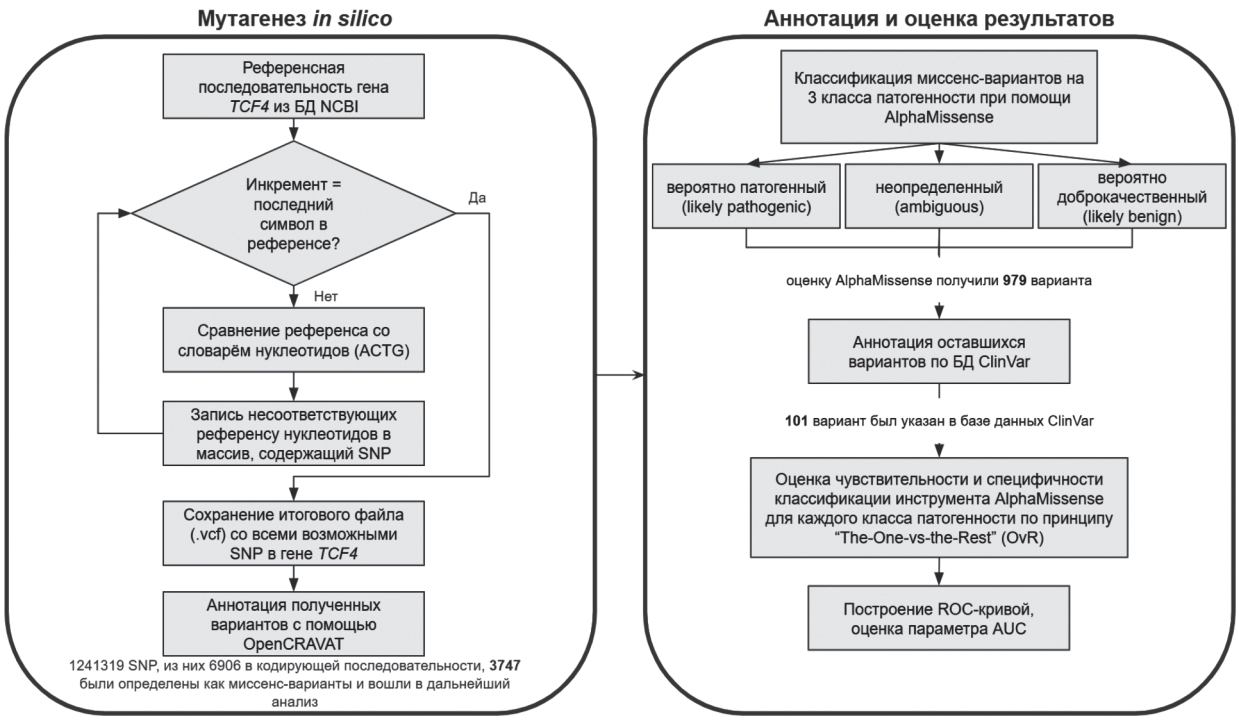

Рис. 1. Дизайн исследования

Fig. 1. Research design

ложительная частота), FPR или (1 − специфичность) определяет долю ошибочно классифицированных отрицательных результатов относительно всех отрицательных результатов. А ось Y — True Positive Rate (истинно положительная частота). TPR также известна как чувствительность, и определяется как доля правильно классифицированных положительных результатов относительно всех положительных результатов [13]. Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода, т.е. обнаруживает положительные случаи. Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода, т.е. обнаруживает отрицательные случаи.

Показатель AUC (Area Under the ROC Curve) − это мера, которая позволяет суммировать производительность модели одним числом, измеряя площадь под кривой ROC. AUC колеблется от 0 до 1, где более высокое значение AUC указывает на более высокую производительность модели. AUC равный 0,5 указывает на отсутствие дискриминационной способности модели, для сравнения приведен на графиках в виде пунктирной диагональной линии (**рис. 2**).

Результаты и обсуждение

Алгоритм AlphaMissense классифицирует данные на три группы: вероятно патогенный (likely pathogenic),

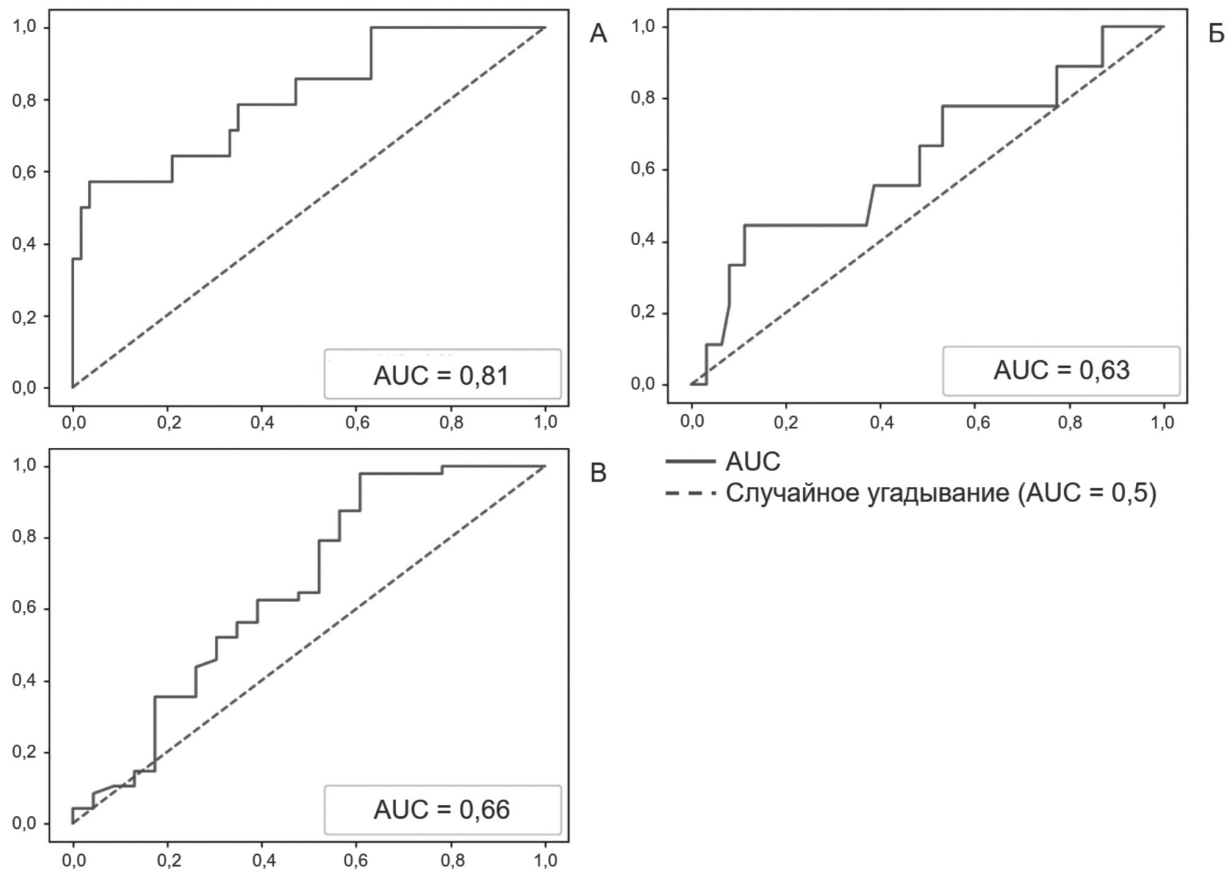

Рис. 2. ROC-кривые оценки классификации SNP инструментом AlphaMissense: ось X – ложноположительная частота (1 – Sp); ось У – чувствительность. А – вероятно патогенные VS неопределенные и вероятно доброкачественные; Б – вероятно доброкачественные VS неопределенные и вероятно патогенные; В – неопределенные VS вероятно доброкачественные и вероятно патогенные.

Fig. 2. ROC curves for SNP classification by the AlphaMissense tool: X-axis – false positive rate (1 – Sp); Y-axis – sensitivity. A – probably pathogenic vs. uncertain and probably benign; B – probably benign vs. uncertain and probably pathogenic; C – uncertain vs. probably benign and probably pathogenic.

неопределенный (ambiguous), вероятно доброкачественный (likely benign). Для минимизации ошибки при многоклассовом сравнении был использован сценарий эксперимента «The-One-vs-the-Rest» (OvR). При таком сценарии один из классов рассматривают как «истинный», а все оставшиеся как «ложный». В качестве модели была выбрана логистическая регрессия, для которой зависимой переменной был выбран класс патогенности по ClinVar (приведен к бинарному виду в соответствии со сценарием эксперимента), а независимой – значение оценки AlphaMissense (вероятностная переменная). Состав данных согласно оценке AlphaMissense: 9 вариантов неопределенной значимости, 61 вариант вероятно доброкачественный, 31 вариант вероятно патогенный. Состав данных согласно классификации ClinVar: 3 варианта доброкачественные, 12 вероятно доброкачественные, 11 вероятно патогенные, 8 патогенные, 67 неявной клинической значимости.

При сравнении показателей чувствительности, специфичности, а также графиков ROC кривых и значений показателей AUC явное отличие имеет оценка классификации SNP, как вероятно патогенных. Значение AUC = 0,81 характеризует модель как очень хорошую; Se = 0,68 и Sp = 0,78 показывают, что 68% патогенных вариантов получат верную классификацию, а 78% SNP без патогенного влияния не войдут в ложноположительный результат. Таким образом, эта классификация может быть использована для приоритизации вариантов во время интерпретации.

Способность инструмента к классификации вариантов как вероятно доброкачественных является недостаточной, чтобы доверять этой оценке. Показатель AUC = 0,63 указывает на то, что данная модель классификации является моделью со средним качеством. При показателях Se = 0,79 и Sp = 0,43 количество ложноположительных результатов будет достаточно высоким. 57% вероятно патогенных или неопределенных вариантов будут охарактеризованы неверно. Такая модель завышает показатель вероятно доброкачественных вариантов.

Способность инструмента к классификации вариантов как неопределенных является самой низкой из представленных моделей. Показатели Se = 0,12 и Sp = 0,97 указывают на то, что данная модель не идентифицирует многие варианты с неопределенной значимостью как таковые. В то же время показатель ложноположительных результатов является самым низким среди всех трех оценок (3%). Показатель AUC = 0,66 также характеризует эту модель как модель со средним качеством. Однако применительно к классификации вариантов это можно интерпретировать как положительное событие, т.к. варианты с неопределенной клинической значимостью из баз данных могут получить свой класс патогенности, что может способствовать в будущем более точной диагностике.

Основываясь на полученных результатах оценки способности AlphaMissense к классификации вариантов, были проанализированы оставшиеся 878 вариантов, не указанные в базе данных ClinVar. Данные нашего исследования для гена *TCF4* имеют схожее распределение долей по сравнению с результатами исследований авторов AlphaMissense [1]. Однако из-за завышения инструментом показателей вероятно доброкачественных вариантов, реальное соотношение классов может быть смещено в сторону вероятно патогенных. Это показывает необходимость увеличения клинических баз данных для лучшего обучения предсказательных инструментов и их дальнейшего использования в практике.

Большая часть вариантов, предсказанных AlphaMissense, приходится на функциональные домены белка TCF4. Таким образом, основная структура спираль-петля-спираль (bHLH) включает 15 предсказанных вариантов. На консервативные активаци-

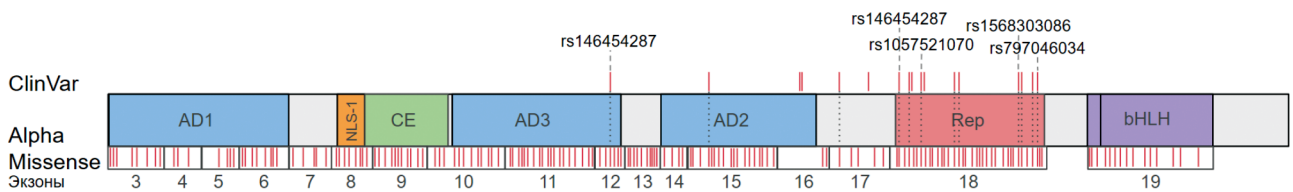

Рис. 3. Схема расположения вероятно патогенных миссенс-вариантов, предсказанных AlphaMissense и зарегистрированных в базе данных ClinVar, относительно доменов белка TCF4.

Fig. 3. Schematic diagram of the arrangement of likely pathogenic missense variants predicted by AlphaMissense and registered in the ClinVar database relative to the TCF4 protein domains.

онные домены (AD1, AD2 и AD3) попадают 23, 22 и 34 варианта соответственно. Данные домены способны модулировать транскрипцию, а также, в зависимости от типа клеток, способны независимо или совместно регулировать экспрессию генов-мишеней [14].

На домен сигнала ядерной локализации (NLS-1) приходится 8 вариантов из предсказанных. Однако, согласно имеющимся исследованиям, сложно сказать, как именно этот домен влияет на регуляцию активности белка или как взаимодействует с NLS-2 и сигналами ядерного экспорта (NES-1 и NES-2), входящими в состав домена bHLH [14].

Внутримолекулярный домен CE, названный «консервативным элементом», включает в себя 15 предсказанных вариантов. Он способен подавлять активность домена AD1. Другой внутримолекулярный «репрессионный домен» Rep, включает в себя 36 вариантов. Согласно литературным источникам, он репрессирует активность как домена AD1, так и AD2. Вероятно, оба внутримолекулярных регуляторных домена действуют путем предотвращения рекрутирования транскрипционных кофакторов [14].

Патогенные SNP, указанные в базе данных ClinVar, в основном попадают на функциональный домен Rep, при этом 8 из 11 вариантов пересекаются с предсказаниями AlphaMissense. Из оставшихся вариантов 3 расположены в домене AD2 (1 вариант пересекается с предсказаниями), 1 пересекающийся с предсказаниями относится к домену AD1. Также примечательно, что 2 варианта попадают в нефункциональную область белка, и 1 из них имеет пересечение с предсказанным вариантом.

Выводы

Инструмент AlphaMissense превосходит по способности к классификации патогенности вариантов инструменты, используемые сейчас в стандартных протоколах биоинформатической обработки данных (ClinPred, CADD, Polyphen2, PROVEAN), и может быть рекомендован для приоритизации вероятно патогенных миссенс-вариантов при интерпретации результатов секвенирования. Измеренные показатели дают понять, что лучше всего инструмент определяет вероятно патогенные варианты, однако стоит с сомнением относиться к вариантам, определенным как вероятно доброкачественные и делать проверку другими инструментами. Варианты, полученные в ходе *in silico* мутагенеза и оцененные как вероятно патогенные, но не указанные в базах данных, могут быть полезны при определении ранее неизвестных вариантов в гене *TCF4* и помочь в диагностике ассоциированных заболеваний.

Литература/References

1. Cheng J., Novati G., Pan J., et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664):eadg7492.
2. Teixeira J.R., Szeto R.A., Carvalho V.M.A. et al. Transcription factor 4 and its association with psychiatric disorders. Translational psychiatry. 2021.;11(1):19.
3. Stefansson H., Ophoff R.A., Steinberg S. et al. Common variants conferring risk of schizophrenia. Nature. 2009;460(7256):744–747.
4. Smoller J.W., Kendler K.K., Craddock N. et al.Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013;381(9875):1371–1379.
5. Wray N.R., Ripke S., Mattheisen M. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature genetics. 2018;50(5):668–681.
6. Cock P.J., Antao T., Chang J.T. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(1422):3.
7. Sayers E.W., Bolton E.E., Brister J.R. et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50(D1):D20-D26.
8. Pagel K.A., Kim R., Moad K. et al. Integrated Informatics Analysis of Cancer-Related Variants. JCO Clin Cancer Inform. 2020;4:310-317.
9. Landrum M.J., Lee J.M., Riley G.R. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(D980):5.
10. Tunyasuvunakool K., Adler J., Wu Z. et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590-596.
11. Ljungdahl A., Kohani S., Page N.F. et al. AlphaMissense is better correlated with functional assays of missense impact than earlier prediction algorithms. bioRxiv [Preprint].2023.
12. Pedregosa F., Varoquaux G., Gramfor A. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.
13. Sonego P., Kocsor A., Pongor S. ROC analysis: applications to the classification of biological sequences and 3D structures. Briefings in Bioinformatics. 2008;9(3):198–209.
14. Teixeira J.R., Szeto R.A., Carvalho V.M.A., Muotri A.R., Papes F. Transcription factor 4 and its association with psychiatric disorders. Transl Psychiatry. 2021;11(1):19.