ISSN 2782-3806
ISSN 2782-3814 (Online)
УДК 575.1

ИЗУЧЕНИЕ ВРОЖДЕННЫХ РИСКОВ ЗАБОЛЕВАНИЙ С ПОМОЩЬЮ МЕТОДОВ ВЫЧИСЛИТЕЛЬНОЙ ГЕНЕТИКИ

Артемов Н. Н. [1, 2, 3, 4, 5]

[1]Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр имени В. А. Алмазова» Министерства здравоохранения Российской Федерации, Санкт-Петербург, Россия
[2]Научный центр мирового уровня «Центр персонализированной медицины» , Санкт-Петербург, Россия
[3]Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики», Санкт-Петербург, Россия
[4]Институт Броада, Кембридж, США
[5]Массачусетская больница общего профиля, Бостон, США

Контактная информация:
Артемов Никита Николаевич,
ФГБУ «НМИЦ им. В. А. Алмазова»
Минздрава России,
ул. Аккуратова, д. 2, Санкт-Петербург,
Россия, 197341.
E-mail: artemov_nn@almazovcentre.ru

Статья поступила в редакцию
03.09.2021 и принята к печати
25.10.2021.

РЕЗЮМЕ

Обширный объем накопленных геномных данных за последние годы сделал возможным переход от ассоциативных исследований, направленных на поиск новых генов риска, к интерпретации индивидуальных рисков и прогнозу.

В данном обзоре рассмотрены ключевые этапы развития вычислительной биологии, а также основные принципы и ограничения методов прогнозирования рисков моногенных и полигенных наследственных заболеваний.

Ключевые слова: GWAS, биоинформатика, врожденные риски, вычислительная биология, генетика, история науки.

Для цитирования: Артемов Н.Н. Изучение врожденных рисков заболеваний с помощью методов вычислительной генетики. Российский журнал персонализированной медицины. 2021;1(1):136-145.

ВВЕДЕНИЕ

В настоящий момент массив накопленных геномных данных начинает становиться достаточным не только для понимания молекулярных причин заболеваний, но и для оценки индивидуальных врожденных рисков. Таким образом, происходит переход к прогнозированию заболеваний и персонализации медицинских вмешательств, особенно в сфере превентивной медицины.

Использование вычислительных методов в биологии открыло возможности для развития целой отрасли исследований — биоинформатики. Разработка методов для эффективной работы с массивами геномных и клинических данных — ключ для извлечения максимальной практической пользы из фундаментальных генетических исследований.

В данном обзоре представлена краткая история развития вычислительной биологии и текущее состояние научной области, посвященной оценке врожденных рисков заболеваний.

РАЗВИТИЕ ВЫЧИСЛИТЕЛЬНОЙ БИОЛОГИИ И БИОИНФОРМАТИКИ

19 сентября 1957 года Френсис Крик обозначил ключевые идеи о функции генов. В частности, сформулировал так называемую «центральную догму» молекулярной биологии, утверждающую, что генетическая информация может передаваться от нуклеиновых кислот к белку, но не в обратном направлении [1, 2].

Несмотря на ключевую роль ДНК в качестве носителя генетической информации, возможность установить последовательность аминокислот в белке появилась значительно раньше, чем методы секвенирования нуклеиновых кислот. К примеру, секвенирование Эдмана, ставшее доступным в 1949 году благодаря автоматизации, в последующее десятилетие позволило получить последовательности более 15 семейств белков [3]. Одно из важных ограничений метода Эдмана — невозможность секвенировать более 50–60 аминокислот за один раунд [4]. Для секвенирования белков, состоящих из большего количества аминокислот, их необходимо было предварительно фрагментировать.

Таким образом, основной проблемой стал не сам процесс секвенирования, а сборка последовательности аминокислот в белке, исходя из сотен небольших полипептидов, секвенированных с помощью метода Эдмана. Данная задача дала основной толчок для появления биоинформатики. Задача сборки последовательности больших белков оказалась чрезвычайно сложной для аналитического решения. Это привело к тому, что в начале 1960-х годов была создана первая биоинформатическая программа, которая позволяла автоматически решать данную задачу [2].

Маргарет Дейхофф (англ. Margaret Dayhoff) и Роберт С. Ледли (англ. Robert S. Ledley), считающиеся прародителями современной вычислительной биологии, в 1962 году создали COMPROTEIN — первую компьютерную программу, записанную на перфокартах, которая решала задачу, актуальную и в настоящее время, *de novo* сборку секвенированной последовательности [5]. Стоит отметить, что внедрение программных решений и существовавшие ограничения производительности вычислительной техники привели к нововведениям, которые мы используем и сейчас, говоря о белковых последовательностях. Так, из-за сложности хранения трехбуквенных названий аминокислот в памяти компьютеров того времени Маргарет Дейхофф предложила современную однобуквенную кодировку аминокислот [6]. Вслед за COMPROTEIN применение вычислительных методов для анализа аминокислотных замен и последовательностей дало значительный толчок к развитию биоинформатики.

К 1970-м годам стало очевидно, что секвенирование отдельных белков является неэффективным, так как требует выделения и очистки каждого белка в отдельности, в то время как секвенирование ДНК позволит определить аминокислотную последовательность всех белков одновременно. Появление секвенирования Максама-Гилберта [7] и Сангера [8] привело к тому, что поиск генов и изучение влияния мутаций в ДНК на возникновение заболеваний стали одним из основных направлений в биологии.

Первые попытки обнаружить положение генов в ДНК и связать мутации с риском заболевания начались задолго до начала проекта «Геном человека». Джеймс Гузелла (англ. James Gusella), Ненси Векслер (англ. Nancy Wexler) и коллеги в 1976–1983 годах, в ходе изучения болезни Хантингтона — редкого моногенного заболевания с аутосомным доминантным наследованием, смогли обнаружить и описать ген *HT*, мутации в котором приводили к возникновению заболевания. Это стало первым примером гена, связанного с генетическим заболеванием. В ходе исследования был сделан анализ сегрегации ДНК-маркеров с фенотипом у большого количества семей с наследственной болезнью Хантингнона [9]. Очевидно, что масштабный анализ генеалогических деревьев — чрезвычайно трудоемкая задача, которая с успехом может решаться алгоритмически, что и привело к разработке целого ряда методов для анализа сцепленного наследо-

вания генов (англ. Linkage analysis) [10] и построения карт сцепленности генов [11].

В случае с полигенными заболеваниями проведение подобного анализа затрудняется необходимостью следить за совместной сегрегацией многих ДНК-маркеров одновременно. Программа GENEHUNTER позволила решить эту задачу и обнаружить сотни ДНК-локусов, ассоциированных с заболеваниями еще до завершения проекта «Геном человека» [12].

С появлением современных методов высокопроизводительного секвенирования, а также генотипирования с помощью микрочипов, удалось накопить большой массив данных для проведения ассоциативных исследований. На текущий момент накопленные результаты таких исследований позволили обнаружить тысячи генов и индивидуальных ДНК-вариантов, связанных с рисками заболеваний [13, 14]. Как следствие, более глубокое понимание молекулярных причин нарушения функционирования организма постепенно приводит к смещению фокуса исследований в сторону использования данных об эффектах ДНК-вариантов на фенотип для оценки индивидуальных предрасположенностей к заболеваниям.

МОНОГЕННЫЕ БОЛЕЗНИ

Для примерно 20 % генов, кодирующих белки в геноме человека, к текущему моменту достоверно установлена связь с одним или несколькими фенотипами, что показывает значительный прогресс, достигнутый за последние 40 лет, но в то же время показывает огромный объем работы, который еще предстоит выполнить.

Моногенные болезни могут быть вызваны ДНК-вариантами различного типа, от однонуклеотидных полиморфизмов до комплексных геномных перестроек [15, 16]. В случае конкретного пациента оценка риска моногенного заболевания заключается в понимании роли и пенетрантности отдельных ДНК-вариантов, обнаруженных в гене, связанном с заболеванием. К примеру, наследственный рак молочной железы часто вызван отдельными ДНК-вариантами в гене *BRCA1*. Задача отделения полиморфизмов с незначимым риском от высокорисковых редких ДНК-вариантов — одна из наиболее актуальных тем для изучения рисков наследственных заболеваний и постановки молекулярных диагнозов [17].

На текущий момент обширное количество вычислительных методов, основанных на различных источниках функциональной информации (консервативности аминокислот, функции белковых доменов и т. д.), используется для численной оценки риска, связанного с конкретным ДНК-вариантом [18–20]. Однако проверка качества таких предсказаний, основанных на вычислительных моделях, до последнего времени оставалась затруднительной.

Финдлей (англ. Findlay) и коллеги в 2018 году с помощью технологии CRISPR создали все возможные однонуклеотидные мутанты в гаплоидной клеточной линии HAP1, чувствительной к нормальному функционированию гена *BRCA1* [21]. Измеряя клеточную виабильность для каждой мутантной клеточной линии, была проведена оценка функциональной важности каждого из ДНК-вариантов. Данный подход позволил с высокой точностью (~98 %) воспроизвести фенотип для известных клинически значимых вариантов в *BRCA1* [22], а также оценить эффективность популярных алгоритмов для функциональной аннотации ДНК-вариантов. Их невысокая чувствительность остается основной проблемой для интерпретации индивидуальных рисков моногенных заболеваний, однако достаточной для приоритизации ДНК-вариантов при поиске новых генов риска.

В 2018 году Каммингс (англ. Cummings) и соавторы смогли показать, что анализ геномного секвенирования для постановки молекулярного диагноза у пациентов с мышечными дистрофиями по-прежнему недостаточен для ~50 % пациентов. Однако одновременный анализ геномного или экзомного секвенирования с РНК-секвенированием помогает идентифицировать генетическую причину заболевания для дополнительных 35 % пациентов [23].

Секвенирование ДНК предоставляет уникальные возможности для постановки молекулярных диагнозов моногенных болезней. Разработка методов для приоритизации и классификации ДНК-вариантов, в частности в некодирующей последовательности ДНК, в будущем позволит в еще большей степени раскрыть ценность генетической информации для медицины моногенных болезней.

ПОЛИГЕННЫЕ БОЛЕЗНИ

В отличие от моногенных болезней, врожденный компонент риска возникновения комплексных (или полигенных) заболеваний в большей степени связан с кооперативным действием целого набора ДНК-вариантов, каждый из которых имеет небольшой эффект. Однако у отдельных пациентов, несмотря на комплексный характер заболевания, риск может быть связан с редкими вариантами, оказывающими большой фенотипический эффект в одном гене [24–26]. Такого рода суммирование рисков от разных групп ДНК-вариантов еще более усложняет оценку индивидуальных рисков.

Информации от одного ДНК-варианта для оценки риска полигенной болезни недостаточно. Вместо интерпретации генотипов отдельных вариантов используется оценка генетической «нагрузки» с помощью величины, включающей все рисковые ДНК-варианты. Существует несколько подходов к комбинированию информации, полученной из нескольких ДНК-локусов. Наиболее распространенным способом является оценка полигенного риска (polygenic risk score — PRS) — взвешенной суммы количества рисковых аллелей, присутствующих у пациента [27]. В отдельных случаях данный критерий оказывается достаточным для идентификации пациентов, имеющих врожденный полигенный риск, сравнимый с наличием высокопенетрантных мутаций, предрасполагающих к моногенным заболеваниям [28].

Изначально PRS применялся в исследованиях «случай-контроль» для детекции генетических отличий между когортами, что позволило подтвердить важность генетических факторов риска для широкого спектра комплексных заболеваний. Это стало особенно важным для исследований генетики психиатрических заболеваний, требующих наличия десятков тысяч участников для достижения достаточной статистической мощности [29].

Использование данного подхода для выявления пациентов с повышенным риском полигенных заболеваний сталкивается с рядом сложностей на пути к реальному применению в клинической практике. Одним из важных ограничений является отсутствие стандартизованного перевода относительного риска, который предоставляет метрика PRS, в абсолютный риск [30].

Также ограничена возможность трансфера между популяциями результатов GWAS, которые используются в качестве референса для построения модели полигенного риска. Методы, позволяющие делать поправку на популяционную структуру, стали появляться совсем недавно [31].

Модели для предсказания риска комплексных заболеваний, включающие комбинацию из клинических и биохимических факторов, а также факторов, связанных с образом жизни, достаточно хорошо показывают себя в реальных условиях [32, 33]. Однако добавление PRS к данным моделям может значительно помочь в более ранней идентификации людей с повышенными рисками, еще задолго до появления клинических симптомов [34].

Таким образом, оценка полигенных рисков находится в процессе трансформации из метода, позволяющего детектировать разницу в генетической «нагрузке» в исследованиях типа «случай-контроль», в метод, который нуждается в стандартизации и более глубоком понимании на пути трансляции в клиническую практику.

ЗАКЛЮЧЕНИЕ

Разработка методов для работы с «большими данными» стала достаточно привычной в экономике и статистике. Применение подобного подхода в медицине и генетике — очень динамично развивающаяся область исследований. Накопление данных о геномном разнообразии, структуризация и объединение клинических данных — ключевые шаги для создания фундаментальной базы персонифицированной медицины.

Конфликт интересов

Авторы заявили об отсутствии потенциального конфликта интересов. / The authors declare no conflict of interest.

СПИСОК ЛИТЕРАТУРЫ

1. Cobb M. 60 years ago, Francis Crick changed the logic of biology. PLoS Biol. 2017; 15 (9): e2003243. DOI: 10.1371/journal.pbio.2003243.
2. Gauthier J, Vincent AT, Charette SJ, et al. A brief history of bioinformatics. Brief Bioinform. 2019; 20(6):1981–1996. DOI: 10.1093/bib/bby063.
3. Hagen JB. The origins of bioinformatics. Nat Rev Genet. 2000;1(3):231–236. DOI: 10.1038/35042090.
4. Edman P, Begg G. A protein sequenator. Eur J Biochem. 1967;1(1):80–91. DOI: 10.1007/978-3-662-25813-2_14.
5. Dayhoff MO, Ledley RS. Comprotein: A computer program to aid primary protein structure determination. AFIPS. 1962;1:262–274. DOI: 10.1145/1461518.1461546.
6. IUPAC-IUB Commission on Biochemical Nomenclature. A one-letter notation for amino acid sequences. Tentative rules. Biochemistry. 1968;7(8):2703–2705. DOI: 10.1021/bi00848a001.
7. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977 Feb;74(2):560–564. DOI: 10.1073/pnas.74.2.560.
8. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–5467. DOI: 10.1073/pnas.74.12.5463.
9. Gusella JF, Wexler NS, Conneally PM, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983;306(5940):234–238. DOI: 10.1038/306234a0.
10. 10. Bryant SP. Software for genetic linkage analysis. In: Mathew CG. (eds) Protocols in human molecular genetics. Methods in Molecular Biology. Springer, Totowa, NJ, 1991;9:403–418. doi: 10.1385/0-89603-205-1:403.

11. Lander ES, Green P Abrahamson J, et al. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics. 1987;1(2):174–181. DOI: 10.1016/0888-7543(87)90010-3.

12. Kruglyak L, Daly MJ, Reeve-Daly MP, et al. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996;58(6):1347–1363.

13. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514-7. DOI: 10.1093/nar/gki033.

14. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901. DOI: 10.1093/nar/gkw1133.

15. Posey JE. Genome sequencing and implications for rare disorders. Orphanet J Rare Dis. 2019;14(1):1–10. DOI: 10.1186/s13023-019-1127-0.

16. Lupski JR. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environ Mol Mutagen. 2015;56(5):419–436. DOI: 10.1002/em.21943.

17. Cooper DN, Krawczak M, Polychronakos C, et al. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013;132(10):1077-130. DOI: 10.1007/s00439-013-1331-2.

18. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–249. DOI: 10.1038/nmeth0410-248.

19. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–3814. DOI: 10.1093/nar/gkg509.

20. Samocha KE, Kosmicki JA, Karczewski KJ, et al. Regional missense constraint improves variant deleteriousness prediction. bioRxiv. 2017;148353. DOI: 10.1101/148353.

21. Findlay GM, Daza RM, Martin B, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562(7726):217–222. DOI: 10.1038/s41586-018-0461-z.

22. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D1067. DOI: 10.1093/nar/gkx1153.

23. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386):eaal5209. DOI: 10.1126/scitranslmed.aal5209.

24. Saint Pierre A, Génin E. How important are rare variants in common disease? Brief Funct Genomics. 2014;13(5):353–361. DOI: 10.1093/bfgp/elu025.

25. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011;43(11):1066–1073. DOI: 10.1038/ng.952.

26. Singh T, Neale BM, Daly MJ. Exome sequenc26. ing identifies rare coding variants in 10 genes which confer substantial risk for schizophrenia. medRxiv. 2020.09.18.20192815. DOI: 10.1101/2020.09.18.20192815.

27. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12(1):44. DOI: 10.1186/s13073-020-00742-5.

28. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219–1224. DOI: 10.1038/s41588-018-0183-z.

29. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia thatoverlaps with bipolar disorder. Nature. 2009;460(7256):748–752. DOI: 10.1038/nature08185.

30. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–590. DOI: 10.1038/s41576-018-0018-x.

31. Martin AR, Kanai M, Kamatani Y, et al. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584–591. DOI: 10.1038/s41588-019-0379-x.

32. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Münster (PROCAM) study. Circulation. 2002;105(3):310–305. DOI: 10.1161/hc0302.102575.

33. Wilson PWF, Meigs JB, Sullivan L, et al. Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study. Arch Intern Med. 2007;167(10):1068–1074. DOI: 10.1001/archinte.167.10.1068.

34. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. Nat Med. 2020;26(4):549–557. DOI: 10.1038/s41591-020-0800-0.

Информация об авторе:

Артемов Никита Николаевич, руководитель НИЛ популяционной генетики НМИЦ им. В. А. Алмазова, доцент-исследователь Университета ИТМО, научный сотрудник Института Броада, инструктор по исследованиям Массачусетской больницы общего профиля.