УДК 57:51-76;57.02:001.57;57:007:001.891.57;616-006

Яцков Н. Н., Смолякова Е. В., Скакун В. В., Гринев В. В.

ОБНАРУЖЕНИЕ САЙТОВ ОДНОНУКЛЕОТИДНОГО ГЕНЕТИЧЕСКОГО ПОЛИМОРФИЗМА НА ОСНОВЕ ЭНТРОПИИ

Белорусский государственный университет, Минск, Республика Беларусь

В работе предложен метод обнаружения однонуклеотидных полиморфизмов на основе вычисления энтропии сайтов геномной ДНК. Эффективность разработанного алгоритма подтверждена в ходе анализа экспериментальных данных геномного секвенирования.

Введение. Генетический полиморфизм влияет на фенотип человека и других живых организмов 6. Однонуклеотидные полиморфизмы (SNP, от англ. single nucleotide polymorphism) являются одним из наиболее распространенных типов генетических вариаций в геноме человека. Знание генов, участвующих в развитии рака, в сочетании с возможностью секвенирования генов и биоинформатического анализа, является важным инструментом скрининга пациентов с риском и помощи в генетическом консультировании 2. Среди существующих способов определения сайтов SNP следует отметить методы подсчета числа покрытий (ридов), точного теста Фишера, критерия хи-квадрат, биномиального отношения правдоподобия 6. Методы достаточно универсальны и просты для программной реализации, однако вычислительно затратные и трудно применимы при анализе экспериментальных данных с высоким уровнем шума (небольшой глубиной секвенирования), что часто наблюдается в данных геномного секвенирования, полученных с помощью технологий PacBio и Oxford Nanopore 3. Критерии на основе энтропии широко используются в ходе анализа данных геномного секвенирования, в основном для решения задач ассоциаций областей генома и выявления мутаций при заболеваниях 4. Однако для прямой идентификации сайтов SNP по данным отдельного эксперимента секвенирования генома метод энтропии не используется. Критерии на основе энтропии могут иметь преимущества в точности и эффективности при анализе экспериментальных данных с невысоким числом покрытий и более вычислительно эффективны в сравнении с существующими статистическими методами.

В работе предложен метод обнаружения однонуклеотидных полиморфизмов на основе вычисления энтропии сайтов геномной ДНК. Для оценки эффективности предложенного метода проведено его сравнение с точным тестом Фишера и тестом биномиального отношения правдоподобия на примерах опубликованных экспериментальных данных геномного секвенирования.

Определения SNP на основе вычисления энтропии нуклеотидного сайта. Введем меру энтропии сайта

$$E = \sum f_r^i \ln f_r^i , \tag{1}$$

где f_r^i – нормированная к общему числу покрытий частота нуклеотида r в позиции рассматриваемого сайта.

В случае, если сигналы источников искажений, вносящие неопределённости в оценку числа нуклеотидных оснований в экспериментах геномного секвенирования, распределены нормально, то для описания вероятностного распределения величины энтропии можно предположить распределение хи-квадрат 5. Для расчетной статистики распределения хи-квадрат воспользуемся выражением $2 \cdot n \cdot E$ 6, где n – число покрытий сайта. Для оценки значимости необходимо задать число степеней свободы v равным 4, числу параметров, представленных количеством нуклеотидных оснований, изменяющихся при вычислении энтропии. Для расчётной статистики вычислим p-величину

$$p = \int_{2nE}^{+\infty} \frac{(1/2)^{v/2}}{\Gamma(v/2)} u^{v/2-1} e^{-u/2} du , \tag{2}$$

где Γ – обозначение гамма-функции. p-Величина представляет собой вероятность того, что случайная величина, имеющая распределение хи-квадрат, примет значение не меньше по абсолютной величине, чем наблюдаемое $2 \cdot n \cdot E$. При формировании гипотез $H_0 : E = 0$ и $H_1: E > 0$, если $p > \alpha/2$, где α – уровень значимости критерия, то гипотеза H_0 принимается. Например, взяв α = 0,05 сайт будет идентифицирован как SNP если p-величина для расчётной статистики $2 \cdot n \cdot E$ меньше α.

Приведенный метод не учитывает ситуацию, когда максимальное число покрытий приходится на нереференсный нуклеотид. В работе предлагается способ решения данной задачи, суть которого состоит в награждении (числом покрытий) нуклеотида референсного генома и штрафе (числом покрытий) для нереференсного нуклеотида в половины от числа покрытий нереференсного нуклеотида, так чтобы общее число покрытий оставалось неизменным. Последнее предполагает допустимость применения распределения хи-квадрат при оценке значимости энтропии сайта.

Для программной реализации алгоритма необходимо задать пороговое значение для числа прочтений T_n и уровень значимости критерия α. Отфильтровываются сайты, число покрытий n которых ниже порогового значения T_n. Алгоритм идентификации сайта включает вычисления величины энтропии E и p-величины. Если $p < \alpha/2$, то сайт определяется как SNP.

Экспериментальные данные для сравнительного анализа алгоритмов идентификации сайтов SNP. В работе рассмотрены экспериментальные данные, полученные консорциумом GIAB 7. Выбор данных GIAB обусловлен тем, что на сегодняшний день это наиболее надежные бенчмарк-данные для решения задач, связанных с изучением геномного полиморфизма у человека (от разработки новых инструментальных методов «мокрой» биологии до сравнения алгоритмов обнаружения полиморфных сайтов).

Результаты. Выполнен анализ экспериментальных наборов данных на основе разработанного энтропийного теста (ЭТ), модифицированного точного теста Фишера (МТТФ) из R-пакета *Rsubread* 8 и теста биномиального отношения правдоподобия (ТБОП) из R-пакета *VariantTools* 9. Эффективность алгоритмов оценена с помощью мер точности *Precision* (*P*), чувствительности *Recall* (*R*) и счета F1 *score F1* (F_1), характеризующих свойства алгоритмов не включать ложноположительные события (*P*, неверно классифицированные сайты как SNP), истинно положительные события (*R*, верно классифицированные как SNP) и их комбинированного вклада (F_1) 10. Результаты идентификации сайтов SNP для 5 наборов по 20 000 сайтов, считанных с позиций номер 3, 9, 15, 21, и 27×10^6 в хромосомах 10 и 22, представлены в таблице 1.

Таблица 1

Точность алгоритмов ЭТ, МТТФ и ТБОП

i [1]	**Хромосома 10**									**Хромосома 22**								
	ЭТ			**МТТФ**			**ТБОП**			**ЭТ**			**МТТФ**			**ТБОП**		
	R	***P***	***F*$_1$**	***R***	***P***	***F*$_1$**	***R***	***P***	***F*$_1$**	***R***	***P***	***F*$_1$**	***R***	***P***	***F*$_1$**	***R***	***P***	***F*$_1$**
3	100	100	100	95.0	100	97.4	100	20.2	33.6	75.0	9.7	17.2	50.0	6.7	11.8	75.0	2.3	4.5
9	56.6	8.3	14.5	79.2	4.7	8.9	79.2	4.7	8.9	97.3	97.3	97.3	89.2	100	94.3	100	37.0	54.0
15	100	88.9	94.1	96.9	96.9	96.9	100	25.6	40.8	100	91.7	95.7	87.9	93.5	90.6	100	25.8	41.0
21	100	94.1	97.0	96.9	96.9	96.9	100	37.2	54.2	100	70.8	82.9	94.1	88.9	91.4	100	10.3	18.7
27	100	93.9	96.9	96.8	96.8	96.8	100	37.3	54.3	100	80.0	88.9	100	95.2	97.5	100	28.2	44.0
с.з. [2]	91.3	77.0	**80.5**	93.0	79.1	**79.4**	95.8	25.0	**38.4**	94.5	69.9	**76.4**	84.2	76.9	**77.1**	95.0	20.7	**32.4**

[1] ***i***, $\times 10^6$ – номер первой позиции набора из 20 000 сайтов в хромосомах 10 и 22;

[2] с.з. – среднее значение

ЭТ не ниже в оценке меры F_1, чем МТТФ и значительно превосходит ТБОП при анализе данных консорциума GIAB. Однако для МТТФ характерно снижение точности идентификации сайтов SNP при повышении порога числа покрытий T_n 11. В работе дополнительно выполнен анализ данных с высоким экспериментальным шумом (малым числом покрытий сайтов), результаты которого свидетельствуют о превосходстве ЭТ. Следует отметить, что в ЭТ не требуется ручной подбор пороговых значений параметров, он является автоматизированным, а также предоставляет *p*-величины статистического критерия, что неприменимо в сравниваемых методах.

Заключение. В работе предложен новый метод определения однонуклеотидного полиморфизма на основе вычисления энтропии сайта, позволяющий повысить эффективность определения однонуклеотидного полиморфизма в молекулах ДНК по данным геномного секвенирования. Работоспособность алгоритма подтверждена в ходе сравнительного анализа с наиболее эффективными из существующих алгоритмов на примерах экспериментальных данных. Наилучшие результаты получены для метода на основе энтропии; менее значимые – для модифицированного критерия Фишера; наихудшие – для теста биномиального отношения правдоподобия. Разработанный метод имеет преимущества в точности при анализе экспериментальных данных с невысоким числом покрытий и предоставляет *p*-величины для автоматического или интуитивно понятного ручного подбора пороговых значений статистических параметров при определении значимости сайтов SNP.

Список литературы

1. Sung, W.-K. Algorithms for next-generation sequencing / W.-K. Sung. – 1st ed. – Chapman & Hall/CRC. – 2017. – 350 p.

2. Kappelmann-Fenzl, M. Next Generation Sequencing and Data Analysis / ed. M. Kappelmann-Fenzl. – Cham : Springer. – 2021. – 218 p.

3. Masoudi-Nejad, A. Next Generation Sequencing and Sequence Assembly. Methodologies and Algorithms / A. Masoudi-Nejad, Z. Narimani, N. Hosseinkhan. – New York : Springer. – 2013. – 86 p.

4. Optimal Design of Low-Density SNP Arrays for Genomic Prediction: Algorithm and Applications / X. L. Wu [et al] // PLoS One. – 2016. – Vol. 11, iss. 9 : e0161719.

5. An Entropy-Based Approach for Testing Genetic Epistasis Underlying Complex Diseases / G. Kang [et al] // J. Theor. Biol. – 2008. – Vol. 250, iss. 2. – P. 362-74.

6. de Andrade, M. Entropy Based Genetic Association Tests and Gene-Gene Interaction Tests / M. de Andrade, X. Wang // Stat. Appl. Genet. Mol. Biol. – 2011. Vol. 10, iss. 1 : 38.

7. Zook, J. M. An Open Resource for Accurately Benchmarking Small Variant and Reference Calls / J. M. Zook [et al] // Nature Biotechnology. – 2019. – Vol. 37, iss. 5. – P. 561-566.

8. Liao, Y. The R Package Rsubread is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads / Y. Liao, G. K. Smyth, W. Shi // Nucleic Acids Research. – 2019. – Vol. 47 : e47.

9. VariantTools: Tools for Exploratory Analysis of Variant Calls. [Electronic resource]. – URL: https://www.bioconductor.org/packages/release/bioc/html/VariantTools.html (date of access: 10.03.2023).

10. Murphy, K. P. Probabilistic Machine Learning / K. P. Murphy. – London : The MIT Press, 2022. – 854 p.

11. Сравнительный анализ алгоритмов обнаружения сайтов однонуклеотидных вариаций / Я. В. Шинкевич [и др.] // Информационные системы и технологии = Information Systems and Technologies [Электронный ресурс] : материалы междунар. науч. конгресса по информатике. Ч. 2, Респ. Беларусь, Минск, 27–28 окт. 2022 г. / Белорус. гос. ун-т ; редкол.: С. В. Абламейко (гл. ред.) [и др.]. – Минск : БГУ, 2022. С. 61–66.