ОПРЕДЕЛЕНИЕ САЙТОВ ОДНОНУКЛЕОТИДНЫХ ПОЛИМОРФИЗМОВ С ПОМОЩЬЮ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ ПО ДАННЫМ ГЕНОМНОГО СЕКВЕНИРОВАНИЯ

Е. В. Смолякова

Белорусский государственный университет, г. Минск;
smolyakova580@gmail.com;
науч. рук. – Н. Н. Яцков, канд. физ.-мат. наук, доц.

В работе реализован и исследован алгоритм определения сайтов однонуклеотидных полиморфизмов с помощью биномиального распределения по данным геномного секвенирования.

Ключевые слова: однонуклеотидный полиморфизм; секвенирование; биномиальное распределение.

ВВЕДЕНИЕ

В живых организмах состав геномов варьируется. Различия называются генетическим полиморфизмом или вариациями генома. Включают в себя однонуклеотидные полиморфизмы (SNP), вариации одиночных нуклеотидов (SNV), полиморфизм с небольшими вставками, вариации числа копий (CNV) и структурные вариации (SV). Генетические полиморфизмы влияют на фенотипы и заболевания [1].

SNP являются одним из наиболее распространенных типов генетических вариаций в геноме человека. SNP в генах, которые регулируют репарацию несоответствия ДНК, клеточные циклы, метаболизм и иммунитет, связаны с генетической предрасположенностью к раку [2]. Знание генов, участвующих в развитии рака, в сочетании с возможностью секвенирования генов и бионформатического анализа, является мощным инструментом для скрининга пациентов с риском и помощи в генетическом консультировании [3]. Поэтому нахождение сайтов полиморфизмов является одной из важной задач биомедицины.

Существует много методов определения сайтов SNP, такие как: точный тест Фишера, критерий хи-квадрат, тест биномиального отношения правдоподобия [1]. Они достаточно просты, однако вычислительно затратны и трудно применимы при анализе экспериментальных данных с высоким уровнем шума.

Цель работы – реализовать и исследовать метод определения сайтов SNP с помощью биномиального распределения, и предложить способы автоматического подбора параметров метода.

ВЕРОЯТНОСТЬ ОПРЕДЕЛЕНИЯ САЙТА ОДНОНУКЛЕОТИДНОГО ПОЛИМОРФИЗМА С ПОМОЩЬЮ БИНОМИАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Представим набор оснований, покрывающий конкретный сайт следующим образом: $\mathcal{D} = \{b_1, \dots, b_n\}$, где $n -$ количество ридов на одном сайте и $b_i -$ нуклеотид, соответствует i -ому риду. Допущение заключается в том, что все не референсные варианты в $\mathcal{D}$ генерируется ошибками секвенирования. Пусть случайная величина X будет количеством вариантов среди n оснований. Обозначим $Pr_n(X = k)$ как вероятность наблюдения k вариантов в $\mathcal{D}$.

Предположим, что в $\mathcal{D}$ есть K не референсные вариантов. Будем считать, что сайт является SNP если $Pr(X \geq K) = \sum_{k \geq K} Pr_n(X = k)$ меньше определенного пользователем порога.

Считая ошибки секвенирования n оснований независимыми и вероятность ошибки p известной, тогда X будет подчиняться биномиальному распределению. Вероятность наблюдения k вариантов в $\mathcal{D}$ можно записать следующим образом: $Pr_n(X = k) = C_n^k p^k (1-p)^{n-k}$.

ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ

Таблица 1

Экспериментальные данные

	Название последовательности	Расположение	Референсное значение	A	C	G	T	SNP
0	chr20	41029600	G	0	0	51	0	
1	chr20	41029601	G	0	0	51	0	
2	chr20	41029602	G	0	0	51	0	

Экспериментальные данные представляют собой массив размером 20000 строк на 8 столбцов и содержат информацию о местоположении нуклеотида в последовательности, типе референсного нуклеотида, количестве ридов при секвенировании оснований A, C, G и T, маркере SNP.

Сайт SNP выделяется среди остальных. На рисунке 1 нуклеотид под индексом 2447 является сайтом SNP.

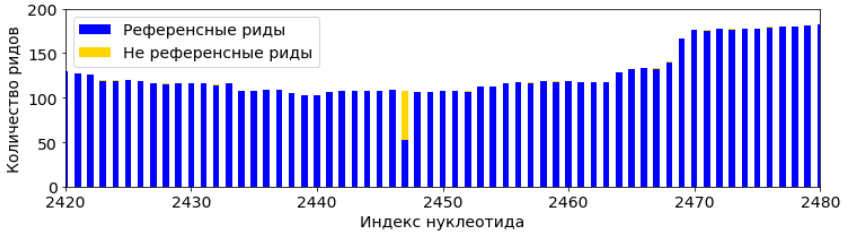

Рис. 1. Количество ридов при прочтении нуклеотидных последовательностей

РЕЗУЛЬТАТЫ

Для того чтобы сайт являлся SNP должно выполняться следующее неравенство:

$$\sum_{k=K}^{n} Pr_n(X = k) < T,$$

где n – общее количество ридов, K – количество не референсных ридов, p – вероятность ошибки секвенирования, T – пороговое значение.

По представленным экспериментальным данным возможно определить общее количество ридов и количество не референсных ридов, но нужно определить оставшиеся два параметра: вероятность ошибки секвенирования и пороговое значение.

Пороговое значение должно зависеть от количества ридов. Один из вариантов определения порогового значение является следующим: $T = 10^{-A}$, где A – это среднее количество ридов по всем экспериментальным данным. В нашем случае $A = 78$.

Чтобы определить является ли сайт SNP, нужно пройтись по данным два раза: первый раз для определения среднего количества ридов и второй раз уже непосредственно для определения сайтов SNP.

Ограничения: объем данных большой, двойное прохождение требует много времени и ресурсов; трудности при потоковой обработке данных.

Ограничения могут быть устранены следующим путём: выбрав интервал данных, определим для него среднее количество ридов и определим сайты SNP.

После прохода по массиву экспериментальных данных были получены результаты поиска сайтов SNP, представленные на рисунке 2.

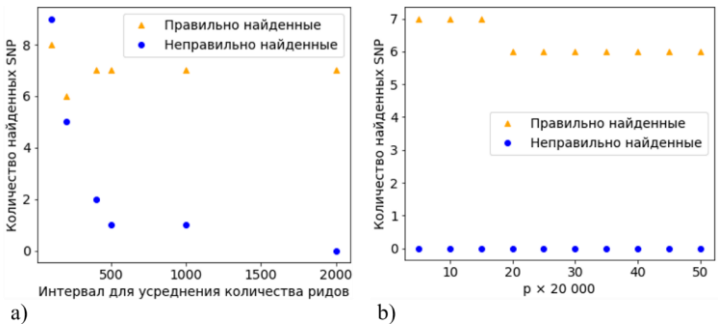

Рис. 2. Количество правильно и неправильно найденных сайтов SNP от количества нуклеотидов в интервале для усреднения количества ридов (a) и от ошибки секвенирования (b)

Как видно на рисунке 2 (a), безошибочно находятся семь из девяти сайтов SNP, когда количество сайтов для усреднения количества ридов

2000 и более. Для построения графика на рисунке 2 (b) выбран интервал усреднения количества ридов равный 2000 сайтов. С увеличением значения ошибки уменьшается количество найденных сайтов SNP.

Следует сделать вывод, что лучший результат, которого удалось достичь: семь правильно определенных сайтов SNP из девяти. Алгоритм не находит сайты SNP на местах с индексами: 14250 и 18493 (таблица 2). Ошибка связана с различием количества ридов между разными сайтами, а также с различием в отношении не референсных ридов к общему количеству ридов.

Таблица 2

Однонуклеотидные полиморфизмы в данных

	Название последовательности	Расположение	Референсное значение	A	C	G	T	SNP
2447	chr20	41032047	A	52	56	0	0	+
3985	chr20	41033585	G	0	28	25	0	+
4602	chr20	41034202	G	92	0	79	0	+
6425	chr20	41036025	T	0	65	0	103	+
12309	chr20	41041909	T	0	0	32	22	+
14250	chr20	41043850	G	10	0	19	0	+
14378	chr20	41043978	T	23	0	0	19	+
18199	chr20	41047799	C	0	22	0	24	+
18493	chr20	41048093	A	33	0	0	10	+

ЗАКЛЮЧЕНИЕ

В данной работе реализован и исследован метод определения сайтов SNP с помощью биномиального распределения, найдена зависимость количества правильно и неправильно найденных сайтов SNP от параметров метода: порогового значения и ошибки секвенирования.

Библиографические ссылки

1. *Wing-Kin Sung* Algorithms for next-generation sequencing / *Wing-Kin Sung* // Chapman & Hall/CRC Computational Biology Series – 2017 – P. 175-185
2. *Na Deng* Single nucleotide polymorphisms and cancer susceptibility / *Na Deng, Heng Zhou, Hua Fan, Yuan Yuan* // Oncotarget – 2017 Nov 7
3. *Melanie Kappelmann-Fenzl* Next Generation Sequencing and Data Analysis / *Melanie Kappelmann-Fenzl [et al.]* // Springer Learning Materials in Biosciences – 2021 – P. 17-36