БИОФИЗИКА, 2023, том 68, № 3, с. 496–500

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 519.876.5

АНАЛИЗ ПОКРЫТИЯ ALU-ПОВТОРОВ ВЫРОВНЕННЫМИ ГЕНОМНЫМИ ПРОЧТЕНИЯМИ

© 2023 г. Г.С. Тамазян*, А.А. Канапин*, А.А. Самсонова*

**Институт трансляционной биомедицины Санкт-Петербургского государственного университета, Университетская наб., 7–9, Санкт-Петербург, 199034, Россия*

E-mail: a.samsonova@spbu.ru

Поступила в редакцию 06.12.2022 г.
После доработки 06.12.2022 г.
Принята к публикации 11.01.2023 г.

Alu-повторы занимают значительную часть генома человека и оказывают большое влияние на процессы, связанные с поддержанием целостности генома. Одним из основных способов изучения вариабельности в геноме, включая Alu-повторы, является геномное секвенирование и последующее выравнивание полученных прочтений на последовательность референсного генома. Ключевой характеристикой выравнивания является глубина покрытия участков референсного генома размещенными на них прочтениями. В данной работе мы предлагаем новый метод анализа покрытия в регионах Alu-повторов и примыкающих к ним геномных регионов прочтениями из набора данных полногеномного секвенирования и исследуем распределение средних величин покрытия в двух указанных разновидностях геномных регионов.

Ключевые слова: секвенирование генома, выравнивание прочтений, геномные повторы.

DOI: 10.31857/S0006302923030109, **EDN:** FRPXRZ

Alu-повторы представляют собой короткие ретротранспозоны, т. е. геномные элементы способные к образованию новых копий посредством механизма «copy–paste». Средняя длина Alu-повтора составляет около 300 пар нуклеотидов (п. н.), поэтому их также относят к типу SINE (short interspersed nuclear elements, короткие рассеянные ядерные элементы). Alu-повторы широко распространены в геномах приматов, в частности в геноме человека они составляют около 11% от общей его длины [1, 2] в количестве около одного миллиона копий, преимущественно расположенных в интронах и межгенных промежутках [3, 4]. Новые вставки Alu-повторов в геноме приводят к разнообразным негативным последствиям для организма хозяина, таким как мутации генов, нарушение структуры регуляторных сайтов и сайтов сплайсинга, а также возникновение новых сайтов неаллельной рекомбинации, приводящей к образованию вариантов копийности (CNV, copy number variants) [5–10]. Данные процессы вносят существенный вклад в нарушение стабильности генома, особенно в случае онкологических заболеваний, когда механизмы подавления активности мобильных элементов нарушены.

Жизненный цикл Alu-повторов включает в себя обратную транскрипцию РНК, считанной с геномных копий. Поскольку сами элементы не содержат кодирующих участков, для этого используется обратная транскриптаза, кодируемая другими типами ретроэлементов, таких как LINE (long interspersed nuclear repeats, длинные рассеянные ядерные повторы). Изучение динамики распространения мобильных элементов в геноме принципиально важно при исследовании таких феноменов, как клональная эволюция опухолевых клеток, в ходе которой они приобретают устойчивость к лекарственным средствам. Ранее нами была разработана математическая модель, описывающая динамику появления новых геномных копий LINE- и SINE-элементов в условиях конкуренции за клеточные ресурсы, такие как АТФ и рибосомы [11]. Тем не менее, наиболее полную информацию о динамике числа копий транспозонов в геноме можно в основном получить из данных полногеномного секвенирования. Существует значительное число алгоритмов и способов поиска вставок мобильных элементов в геном, таких как RepeatMasker [12], SINE_scan [13], Transposome [14] и другие. Эти методы в основном используют статистические подходы,

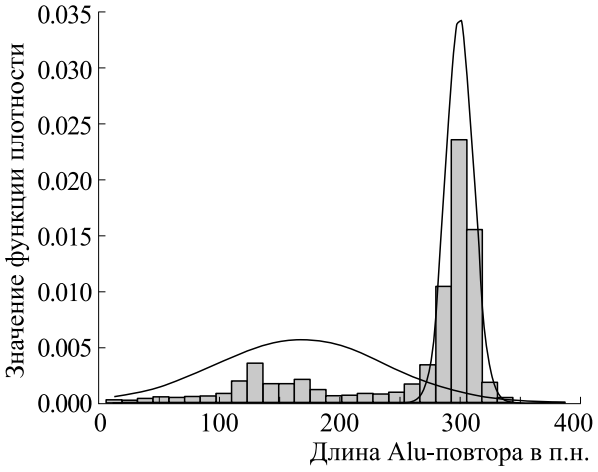

Рис. 1. Гистограмма распределения длин Alu-повторов и оценка распределения длин в виде смеси двух нормальных распределений. Плотности нормальных распределений красного и синего цвета соответствуют частичным и полным Alu-повторам.

анализ спектра нуклеотидных κ-меров, анализ расщепленных фрагментов выровненных на геном прочтений (split reads) и другие.

Одной из ключевых характеристик для вычислительного анализа Alu-повторов является глубина покрытия референсного генома выровненными прочтениями. Особенности реализации алгоритмов выравнивания прочтений в соответствующих программах (например, BWA [15]) могут приводить к системным ошибкам в получаемых выравниваниях в силу наличия большого числа сходных между собой Alu-последовательностей в геноме. В данной работе мы рассматриваем глубину покрытия Alu-повторов и примыкающих к ним регионов выровненными прочтениями для целых повторов и их фрагментов на примере одного образца из проекта «1000 геномов» [16]. Представленный в работе метод может быть распространен как на другие данные секвенирования, так и на геномные регионы иного рода.

МАТЕРИАЛЫ И МЕТОДЫ

Источники данных. Представленный в работе анализ использует следующие наборы данных: выровненные парные прочтения (paired-end reads) для одного образца из проекта «1000 геномов» (идентификатор образца HG00448), полученная программой RepeatMasker аннотация повторов в сборке последовательности генома человека hg19 (загружена с веб-сайта программы https://www.repeatmasker.org) и размеры последовательностей используемой сборки генома человека. Для ускорения вычислений были рассмотрены выравнивания только на собранной последовательности 20-й хромосомы генома человека. Выровненные прочтения были представлены одним файлом в формате BAM [17, 18], аннотированные повторы – в выходном формате программы RepeatMasker.

Выделение геномных повторов. Регионы Alu-повторов в последовательности 20-й хромосомы генома человека были выделены из файла аннотации RepeatMasker с помощью библиотеки pygenomics-ext из состава пакета pygenomics [19]. Анализ распределения длин Alu-повторов для разделения полных и частичных повторов был проведен в среде R [20] с помощью пакета mclust [21].

Обработка выровненных прочтений. Распределение длин выровненных прочтений было проанализировано с помощью Python-скрипта из набора pygenomics-examples, входящего в пакет pygenomics. Длины прочтений оказались лежащими в диапазоне от 43 до 81 п. н. включительно. Средний размер вставки (то есть расстояние между выровненными парными прочтениями) был оценен программой CollectInsertSizeMetrics из пакета Picard [22] и составил 480 п. н. По наибольшей длине прочтения и средней величине размера вставки была вычислена длина примыкающего к повтору региона для последующего анализа покрытия: 81 + 480 + 81 = 642 п. н. Координаты примыкающих к Alu-повторам регионов были определены из координат повторов с использованием программы flank из пакета BEDTools [23]. Усредненные величины покрытия в Alu-повторах и примыкающих регионах были получены из BAM-файла выровненных прочтений применением программы bedcov из пакета SAMtools [17, 24].

Анализ величин покрытия. Alu-повторы были разделены на полные и частичные согласно модели смеси двух нормальных распределений их длины (см. рис. 1). Оценки математического ожидания распределений составили 168 и 300 п. н. соответственно для частичных и полных Alu-повторов; оценки стандартного отклонения – соответственно 70 и 12 п. н.. Мы также исключили из дальнейшего анализа регионы со средним значением покрытия больше 9 прочтений на п. н. Количества рассматриваемых и исключенных регионов приведены в табл. 1, число исключенных из-за высокого покрытия регионов – 0.15% от их общего числа.

Распределения величин покрытия рассмотренных регионов показаны с помощью графиков вида «ящик с усами» (boxplots) на рис. 2 и 3. Нам не удалось обнаружить статистически значимую разницу в среднем покрытии между полными и частичными повторами с примыкающими к ним регионами (см. рис. 2): применение критерия Уэлча для двух выборок (Welch two sample t-test)

Таблица 1. Разбиение рассматриваемых регионов по их типу, полноте соответствующих Alu-повторов и уровню покрытия регионов выровненными прочтениями

Тип повтора	Тип региона	Количество регионов	
		с покрытием < 9	с покрытием > 9
полный	повтор	20667	18
полный	примыкающий	20659	26
частичный	повтор	8402	22
частичный	примыкающий	8404	20

дало значение вероятностного уровня (p-value), равное 0.271. Разница в среднем покрытии между повторами и примыкающими к ним регионами оказалась статистически значимой для обоих видов повторов (см. рис. 3): применение критерия Уэлча дало 95%-е доверительные интервалы для разницы в покрытии от 0.521 до 0.555 прочтений на п. н. для полных Alu-повторов и от 0.344 до 0.409 прочтений на п. н. для частичных Alu-повторов.

К распределениям покрытия повторов и примыкающих регионов был также применен реализованный в R-пакете mclust метод оценивания смеси нормальных распределений. В обоих случаях наиболее подходящими по байесовскому информационному критерию (Bayesian information criterion, BIC) оказались модели смеси двух нормальных распределений: для повторов – с одинаковыми дисперсиями, для примыкающих регионов – с разными. Оценки параметров моделей приводятся в табл. 2. Среди повторов отмечается определенная доля регионов с повышенным средним значением покрытия (компонента 2), причем регионы с более высоким покрытием больше представлены среди неполных повторов (см. табл. 3). Среди примыкающих к повторам регионов подобной значимой разницы в распределении обнаружено не было (см. табл. 4).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Представленные результаты показывают, что распределения глубины покрытия Alu-повторов и примыкающих к ним регионов не показывают системных ошибок, которых можно было бы ожидать из-за особенностей алгоритмов выравнивания прочтений. Распределения глубины покрытия как для повторов, так и для примыкающих регионов можно описывать с помощью простых статистических моделей вида смеси нормальных распределений. Статистически значимую разницу в представленности полных и частичных Alu-повторов в компонентах смеси нормальных распределений и отсутствие такой разницы для примыкающих к повторам регионов можно объяснить тем, что частичные Alu-повторы могут показывать большее разнообразие в своей копийности в геноме. Представленный в рабо-

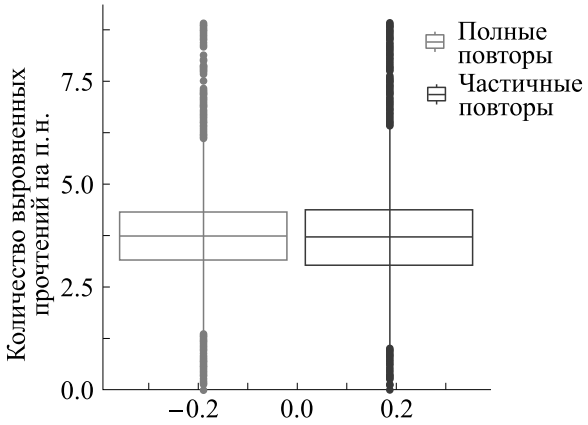

Рис. 2. Распределение средних величин покрытия регионов Alu-повторов вместе с примыкающими к ним регионами в зависимости от типа Alu-повторов: полных или частичных. Распределения показаны в виде «ящиков с усами». Утолщенная линия показывает медианное значение, границы «ящика» – квартили 25% и 75%, «усы» – 95%-й доверительный интервал в предположении о нормальном распределении значений покрытия, точки – выбросы согласно тому же самому предположению.

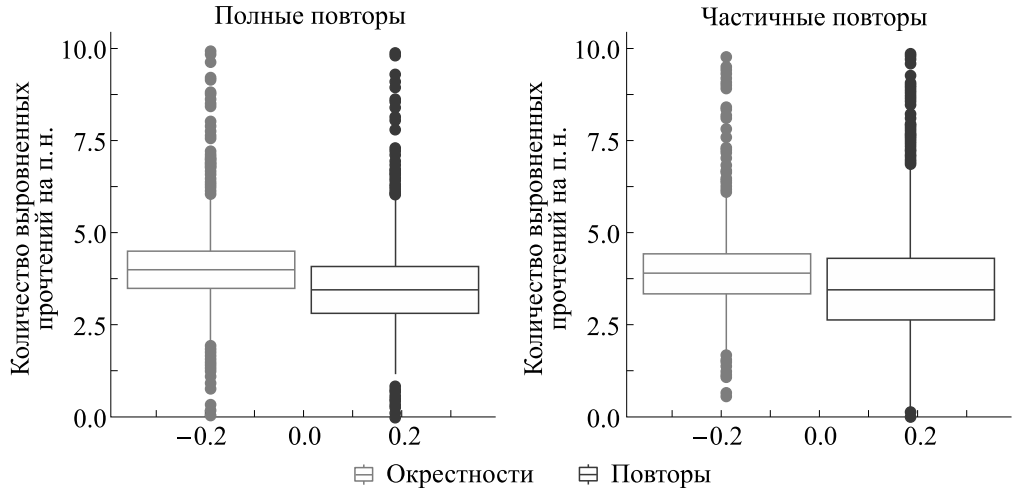

Рис. 3. Распределение средних величин покрытия регионов Alu-повторов и примыкающих к ним регионов (окрестностей) в виде «ящиков с усами».

Таблица 2. Параметры представления распределений средней величины покрытия для Alu-повторов и примыкающих к ним регионов в виде смеси двух нормальных распределений

Параметр	Повторы		Примыкающие регионы	
	Компонента 1	Компонента 2	Компонента 1	Компонента 2
Среднее	3.40	5.87	3.96	4.30
Стандартное отклонение	0.99	0.99	0.74	1.52
Доля регионов	97%	3%	96.5%	3.5%

Примечание. Величины стандартного отклонения и среднего даны в количестве прочтений на п. н. в регионах рассматриваемого рода. Модель для повторов задает равную дисперсию (и, следовательно, стандартные отклонения) для обоих компонент распределения, в отличие от модели для примыкающих регионов.

Таблица 3. Таблица сопряженности для регионов Alu-повторов в соответствии с их полнотой и принадлежности к одной их двух комнонент в ранее построенной модели смеси двух нормальных распределений

Вид повтора	Количество в компоненте 1	Количество в компоненте 2
Полный	20518	149
Частичный	8085	317

Таблица 4. Таблица сопряженности для примыкающих к Alu-повторам регионов в соответствии с полнотой соответствующих регионам повторов и их принадлежности к одной из двух компонент в ранее построенной модели смеси двух нормальных распределений

Вид соответствующего региону повтора	Количество в компоненте 1	Количество в компоненте 2
Полный	20565	94
Частичный	8356	48

Примечание. Применение точного критерия Фишера не приводит к отвержению нулевой гипотезы о равномерности распределения регионов от повторов обоих видов в компонентах смеси распределений.

те алгоритм оценки значений покрытия секвенирования для геномных повторов можно расширить как на большее число секвенированных образцов, так и на геномные регионы другого рода.

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 20-14-00072).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Данная работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. M. A. Batzer and P. L. Deininger, Nat. Rev. Genet., **3** (5), 370 (2002).
2. F. Hormozdiari, M. K. Konkel, J. Prado-Martinez. et al., Proc. Natl. Acad. Sci. USA, **110** (33), 13457 (2013).
3. E. S. Lander, L. M. Linton, B. Birren, et al., Nature, **409** (6822), 860 (2001).
4. J. C. Venter, M. D. Adams, E. W. Myers, et al., Science, **291** (5507), 1304 (2001).
5. F. C. Chen, Y. Z. Chen, and T. J. Chuang, Bioinformatics, **25** (11), 1419 (2009).
6. J. M. Chen, E. Masson, C. Le Marechal, et al., Cytogenet Genome Res, **123** (1–4), 102 (2008).
7. P. Deininger, Genome Biol., **12** (12), 236 (2011).
8. L. M. Payer, J. P. Steranka, W. R. Yang, et al., Proc. Natl. Acad. Sci. USA, **114** (20), E3984 (2017).
9. S. Shen, L. Lin, J. J. Cai, et al., Proc. Natl. Acad. Sci. USA, **108** (7), 2837 (2011).
10. I. Vorechovsky, Hum, Genet., **127** (2), 135 (2010).
11. S. Pavlov, V. V. Gursky, M. Samsonova, et al., Life (Basel), **11** (11), 1209 (2021). DOI: 10.3390/life11111209
12. A. Smit, R. Hubley, and P. Green, *RepeatMasker Open-4.0* (accessed 03/18/2022).
13. H. Mao and H. Wang, Bioinformatics, **33** (5), 743 (2017).
14. S. E. Staton and J. M. Burke, Bioinformatics, **31** (11), 1827 (2015).
15. H. Li and R. Durbin, Bioinformatics, **25** (14), 1754 (2009).
16. S. Fairley, E. Lowy-Gallego, E. Perry, et al., Nucl. Acids Res., **48** (D1), D941 (2020).
17. H. Li, B. Handsaker, A. Wysoker, et al., Bioinformatics, **25** (16), 2078 (2009).
18. J. K. Bonfield, J. Marshall, P. Danecek, et al., Gigascience, **10** (2), giab007 (2021). DOI: 10.1093/gigascience/giab007
19. G. Tamazian, N. Cherkasov, A. Kanapin, et al., in *BGRS/SB-2022* (Novosibirsk, Russia, 2022), pp. 1121-1122.
20. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2022).
21. L. Scrucca, M. Fop, T. B. Murphy, et al., The R Journal, **8** (1), 289 (2016).
22. Broad Institute, *Picard: A set of command line tools for manipulating high-throughput sequencing data* (2022).
23. A. R. Quinlan and I. M. Hall, Bioinformatics, **26** (6), 841 (2010).
24. P. Danecek, J. K. Bonfield, J. Liddle, et al., Gigascience, **10** (2), giab008 (2021). DOI: 10.1093/gigascience/giab008

Analysis of Coverage of Alu Repeats by Aligned Genomic Reads

G.S. Tamazian*, A.A. Kanapin*, and A.A. Samsonova*

**Institute of Translational Biomedicine, St. Petersburg State University, Universitetskaya nab. 7–9, St. Petersburg, 199034 Russia*

Alu repeats occupy a notable part of the human genome and greatly affect processes related to genome integrity maintenance. One of the basic methods for studying variation in a genome, including Alu repeats is genome sequencing followed by mapping the sequenced reads to a reference genome sequence. The key feature of the read alignment is the depth of reference genome region coverage by mapped reads. In this paper, a new method is proposed for analyzing the coverage of Alu repeats and their flanking regions by whole-genome sequencing reads and the distribution of mean coverage in two aforementioned region types is explored.

Keywords: genome sequencing, read alignment, genomic repeats