УДК 577.212.2

# Поиск сильно дивергировавших SINE повторов в геноме риса

© **2020** Ю.М. СУВОРОВА[1*], А.М. КАМИОНСКАЯ[1], Е.В. КОРОТКОВ[1, 2]

[1] *Федеральное государственное учреждение «Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук», Москва, 119071*

[2] *Национальный исследовательский ядерный университет «МИФИ», Москва, 115409*

**e-mail:* suvorovay@gmail.com

Поступила в редакцию 23.03.2020 г.
После доработки 17.04 2020 г.
Принята к публикации 24.07.2020 г.

Представлен новый метод поиска размытых копий SINE повторов в геномных последовательностях, основанный на учете корреляции соседних символов, как при построении позиционно весовой матрицы, так и при проведении сканирования, что позволяет увеличить алфавит и соответственно разрешающую способность метода. При поиске с его помощью SINE повторов в геноме риса найдены новые копии SINE повторов, не включенные в стандартную аннотацию. Проведено тестирование метода, сравнение с программой RepeatMasker и оценка ложных позитивов.

*Ключевые слова:* SINE повторы, ретротранспозоны, методы выравнивания, подобие последовательностей

**doi:** 10.21519/0234-2758-2020-36-4-15-20

Мобильные элементы (транспозоны) — это участки ДНК, способные перемещаться по геному и создавать свои копии. Известно, что транспозоны занимают большую часть геномов эукариот, а в случае растений до 90% от общей длины генома. Долгие годы эти участки генома считались «мусорными», однако, в последнее время показано, что они могут выполнять полезные функции, включая, регуляторные функции [1] и участвовать в адаптации организма к окружающей среде [2], что делает важным их дальнейший поиск и изучение. Транспозоны подразделяются на классы в зависимости от способа перемещения по геному. Транспозоны первого класса при перемещении используют РНК-копии (ретротранспозоны) и метод «копирования и вставки»; второго класса — перемещаются непосредственно при помощи ДНК, их метод — «вырезание и вставка». Элементы первого класса подразделяются на содержащие и не содержащий длинные концевые повторы (LTR — Long Terminal Repeats и не-LTR элементы). Среди не-LTR повторов выделяют семейства LINE (Long Interspersed Nuclear Elements) и SINE (Short Interspersed Nuclear Elements). Повторы семейства SINE — это неавтономные мобильные элементы, т. е. они не кодируют собственные белки, а для перемещения используют белки, кодируемые элементами LINE [3]. В геномах млекопитающих SINE повторы широко представлены (в основном семействами Alu и MIR [4, 5]) и достаточно хорошо изучены. Вместе с тем показано, что повторы, обнаруженные стандартными методами их поиска, это далеко не все копии, содержащиеся в этих геномах [6]. В геномах растений большую часть занимают LTR повторы, а SINE элементы распространены не так широко.

Обычно структура SINE элемента включает в себя тРНК часть, тело (происхождение которого до конца не ясно) и А-богатый хвост [7]. Длина этих элементов варьируется от 100 до 600 нуклеотидов. Известно, что копии SINE элементов после встраивания довольно быстро дивергируют (накапливают мутации) [8], что предотвращает дальнейшую активность транспозона и защищает клетку от его бесконтрольного копирования. Также распространены неполные копии этих повторов. Большое количество замен и других

*Список сокращений*: ПВМ — позиционно-весовая матрица; LTR — Long Terminal Repeats; LINE — Long Interspersed Nuclear Elements; SINE — Short Interspersed Nuclear Elements; МПСДП — метод поиска сильно дивергировавших повторов.

мутаций, происходящих после встраивания новой копии в геном, затрудняют их поиск и изучение современными методами биоинформатики.

Большинство биоинформатических методов разработанных для поиска копий SINE элементов в геномах (как и других типов транспозонов) можно условно разделить на два класса. К первому классу относятся структурные методы, использующие для поиска и классификации транспозонов определенные структурные свойства их последовательностей (такие как тРНК часть, А-богатый хвост и другие). Это так называемые *de-novo* методы, которые могут применяться к новым геномам для поиска в них различных типов повторов. К структурным методам, разработанным для поиска SINE повторов, относятся SINE-Finder [9] и его последователь — программа SINE-Scan [10]. Обычно, структурные методы находят только полные и хорошо сохранившиеся копии повторов. Методы второго класса основаны на сходстве последовательностей — выравнивании. Для поиска новых копий повторов этим методам требуется начальная библиотека последовательностей, собранная обычно методами первого класса. RepeatMasker, метод основанный на выравнивании, является наиболее используемым на сегодняшний день методом поиска всех типов повторов, включая SINE [11]. RepeatMasker использует собственную библиотеку последовательностей повторов Repbase [12] или библиотеку пользователя для идентификации новых копий этих последовательностей в исследуемом геноме. Однако, сильно размытые копии повторов вызывают трудности у методов, основанных на выравниваниях [6].

Поэтому мы разработали новый метод поиска сильно дивергировавших повторов (МПСДП). Особенностью метода является то, что при сравнении последовательностей одновременно учитывается как подобие между последовательностями, так и корреляция пар нуклеотидов внутри последовательностей. МПСДП позволяет находить статистически значимое подобие при большом проценте замен нуклеотидов в случае присутствия вставок и делеций. Ранее аналогичный метод был использован для поиска аминокислотных повторов [13] и сдвигов рамки считывания в белок-кодирующих последовательностях [14]. В данной работе проведено исследование четырех семейств SINE элементов в геноме риса с целью определения их более древних, сильно дивергировавших копий.

## УСЛОВИЯ ЭКСПЕРИМЕНТА

### Описание метода

МПСДП основан на одновременном учете подобия между двумя сравниваемыми последовательностями, и корреляции внутри них соседних нуклеотидов. Ранее нами было показано, что даже при значительной дивергенции сравниваемых последовательностей относительно друг друга (до уровня 3,0 случайных замен на одно основание), когда построить традиционное выравнивание уже не удается, метод, основанный на учете корреляций соседних символов, позволяет построить статистически значимое выравнивание. Алгоритм работы МПСДП включает в себя три основные части. В первой рассчитывается множественное выравнивание для всех известных копий исследуемого повтора и определяется оптимальная позиционно весовая матрица (ПВМ). ПВМ учитывает корреляции соседних нуклеотидов внутри созданного множественного выравнивания, поэтому ее размерность $16 \times L$, где $L$ — длина повтора. Оптимальная ПВМ подбирается для каждого рассматриваемого повтора с использованием генетического алгоритма [13].

На втором этапе происходит нормализация ПВМ и подбор параметров. На третьем с помощью полученной ПВМ производится поиск повторов в последовательности генома с использованием модифицированного метода динамического программирования [15].

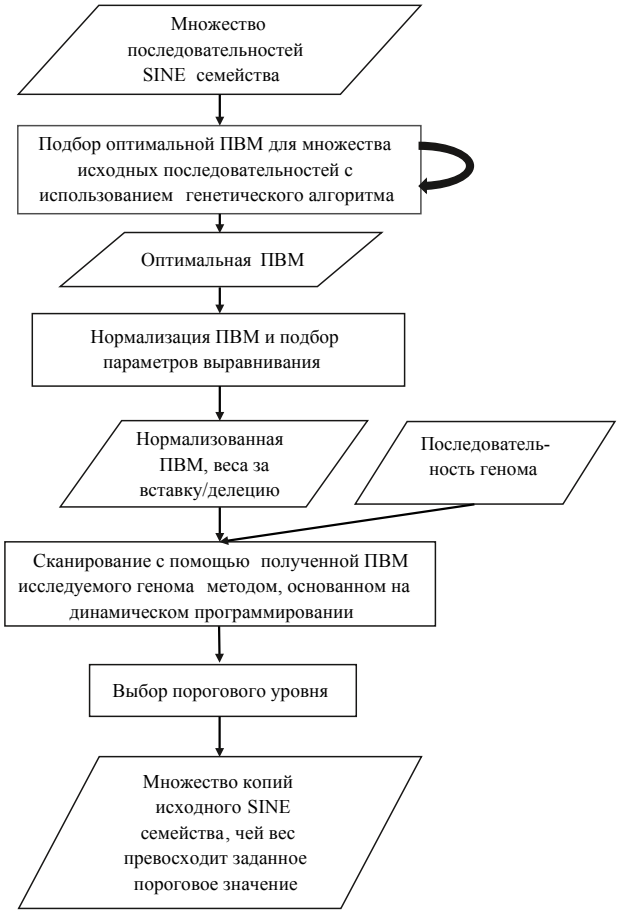


**Рис. 1.** Схема работы МПСДП.

**Fig. 1.** Scheme of the MPSDP algorithm.

Таблица 1

**Результаты работы программ на тестовых последовательностях. В каждом тесте в последовательность хромосомы было сделано 300 вставок последовательности OsSN1 с советующим уровнем случайных замен.**

**The results of the programs on the test sequences. In each test, 300 insertions of the OsSN1 sequence with a corresponding level of random substitutions, were made into the chromosome.**

| | Число случайных замен на одно основание | | | |
|---|---|---|---|---|
| | **0,25** | **0,5** | **0,75** | **1,0** |
| RepeatMasker, первая серия тестов | 300 | 300 | 209 | 32 |
| RepeatMasker, вторая серия тестов | 299 | 266 | 87 | 11 |
| МПСДП, первая серия тестов | 300 | 300 | 300 | 298 |
| МПСДП, вторая серия тестов | 300 | 300 | 288 | 133 |

После сканирования происходит отбор найденных подобий на основании выбранного порогового значения, которое определяется заданным уровнем ложных позитивов. Учет корреляций на этапе построения матрицы и на этапе поиска новых копий позволяют находить сильно дивергировавшие копии исследуемых последовательностей. На рис. 1 представлена схема работы метода, а сам он подробно описан в работе [14].

## Тестирование

Для того, чтобы оценить работу МПСДП и сравнить его с программой RepeatMasker мы провели ряд тестов, имитирующих встраивание копий SINE повторов в геном. В качестве исходной последовательности SINE повтора была использована консенсусная последовательность семейства OsSN1 из базы данных SineBase. Длина последовательности — 293 нуклеотида. Было проведено две серии тестов.

В первой серии использовалась консенсусная последовательность целиком. В первом тесте этой серии последовательность OsSN1 размывалась путем введения случайных замен до уровня 0,25 случайных замен на одну позицию. Всего было создано 300 размытых копий с этим уровнем.

Далее в каждой из этих последовательностей было сделано от двух до пяти случайных вставок или делеций. Полученные копии были встроены в последовательность одной из хромосом генома риса. Используемая последовательность хромосомы была предварительно перемешана случайным образом для того, чтобы убрать следы повторов, которые уже в ней присутствовали. Аналогичным образом были созданы искусственные хромосомы, содержащие копии повтора OsSN1 с уровнем размытия 0,5, 0,75 и 1,0 случайных замен на позицию. Было сгенерировано по 300 последовательностей с заданным уровнем размытия, в каждой были сделаны случайные вставки и делеции, и полученные последовательности были случайным образом встроены в перемешанную хромосому риса.

Во второй серии тестов в качестве исходной последовательности была использована только часть последовательности OsSN1 — первые 150 нуклеотидов. Далее аналогично первой серии исходная последовательность размывалась до уровня 0,25, 0,5, 0,75 и 1,0 случайных замен на одну позицию с добавление случайных вставок и делеций. В каждом тесте было также создано по 300 копий с заданным уровнем размытия, которые были встроены в последовательность перемешанной хромосомы риса.

Искусственные хромосомы со встроенными размытыми повторами были обработаны программами МПСДП и RepeatMasker. В качестве библиотеки в обеих программах во всех тестах была использована полная последовательность OsSN1. В результате сканирования искусственных хромосом из первой серии МПСДП обнаружил все встроенные копии. В то время как RepeatMasker на тестовой хромосоме с уровнем замен в повторах 0,75 нашел 70% встроенных последовательностей и только 11% при уровне замен 1,0 на одно основание. Результаты представлены в табл. 1. Во второй серии тестов, в которых был использован неполный повтор, программа RepeatMasker обнаружила 100%, 89%, 29% и 4% размытых копий соответственно. В то время как, процент копий, обнаруженных МПСДП, составил 100%, 100%, 96% и 44% соответственно (табл. 1).

Результаты проведенных тестов показывают, что метод МПСДП способен находить более древние копии повторов, по сравнению с RepeatMasker, как в случае полных, так и не полных вставок повторов. Также следует упомянуть, что МПСДП определяет более верно границы самих повторов, в то время как RepeatMasker склонен находить более короткие подобия, чем повтор, который был встроен в тестовых последовательностях.

Таблица 2

**Статистика сканирования 12 хромосом риса.**

**The results of the analysis of 12 rice chromosomes.**

| | Всего найдено, МПСДП | Случайных, МПСДП | С учетом пересечения, МПСДП | Найдено, RepeatMasker |
|---|---|---|---|---|
| **OsSN1** | 3077 | 44 | 939 | 999 |
| **OsSN2** | 3351 | 36 | 2770 | 3223 |
| **OsSN3** | 1426 | 36 | 1 081 | 1416 |
| **p-SINE1** | 1019 | 30 | 1018 | 1015 |
| **Всего** | **8873** | **146** | **5808** | **6653** |

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Консенсусные последовательности четырех семейств SINE элементов, относящихся к геному риса, были загружены из базы данных SINEBase [8]. Для каждого семейства разработанным методом была построена ПВМ и подобраны оптимальные параметры. С помощью полученных ПВМ были просканированы все 12 хромосом генома риса (*Oryza sativa* Japonica). Для определения процента ложных позитивов были просканированы 12 хромосом риса, перемешанных случайным образом. В результате сканирования было найдено 8873 копии SINE элементов четырех исследуемых семейств. На перемешанных хромосомах найдено 146 случаев, превышающий выбранный порог. Статистика по каждому из четырех семейств представлена во втором и третьем столбцах табл. 2.

Консенсусные последовательности четырех исследуемых SINE семейств имеют большую степень подобия, и поэтому результаты сканирования генома могут пересекаться между семействами. Табл. 3 демонстрирует уровень пересечения результатов сканирования между четырьмя SINE семействами в результатах работы нашей программы.

Таблица 3

**Пересечение результатов работы программы МПСДП между исследованными семействами SINE.**

**Intersection of the results of the program between the studied SINE families.**

| | OsSN1 | OsSN2 | OsSN3 | p-SINE1 |
|---|---|---|---|---|
| **OsSN1** | 3077 | 2268 | 804 | 2 |
| **OsSN2** | 2268 | 3351 | 206 | 0 |
| **OsSN3** | 804 | 206 | 1426 | 0 |
| **p-SINE1** | 2 | 0 | 0 | 1019 |

Из результатов работы МПСДП мы исключили избыточность, связанную с подобием последовательностей исходных SINE элементов. В случае пересечения копий повторов, соответствующих разным консенсусам в геноме, только один повтор с максимальным весом был помещен в финальную выборку, пересечение не допускается, но копии могут следовать непосредственно друг за другом. В результате осталось 5808 копий повторов (четвертая колонка табл. 2).

Консенсусные последовательности тех же четырех SINE семейств были переданы программе RepeatMasker в качестве библиотеки. Мы использовали пониженное значение порогового уровня для RepeatMasker по сравнению с настройками по умолчанию (-cutoff 160) для того, чтобы результаты работы были сравнимы с МПСДП по числу найденных ложных позитивов. В результате сканирования генома риса RepeatMasker обнаружил 6653 копии этих повторов (пятая колонка табл. 2). При выбранном пороговом уровне RepeatMasker также нашел 172 копии повторов при использовании случайно перемешанных хромосом.

При сравнении результатов работы двух программ, мы обнаружили 1204 подобия, найденных RepeatMasker и отсутствующих в результатах МПСДП, и 545 случаев, найденных МПСДП, но отсутствующих в результатах RepeatMasker. Стоит отметить, что большинство последовательностей, которые были найдены только программой RepeatMasker отличаются короткой длиной (83% из них короче 100 нуклеотидов). Распределение длин подобий уникальных для каждой программы приведено на графике (рис. 2).

Можно заключить, что RepeatMasker позволяет находить короткие фрагменты повторов с высоким уровнем подобия, в то время как МПСДП способен находить более полные и сильно дивергировавшие копии повторов. Также следует заметить, что (как показали тесты) программа RepeatMasker склонна обрезать копии повтора. Для получения полной картины распространения

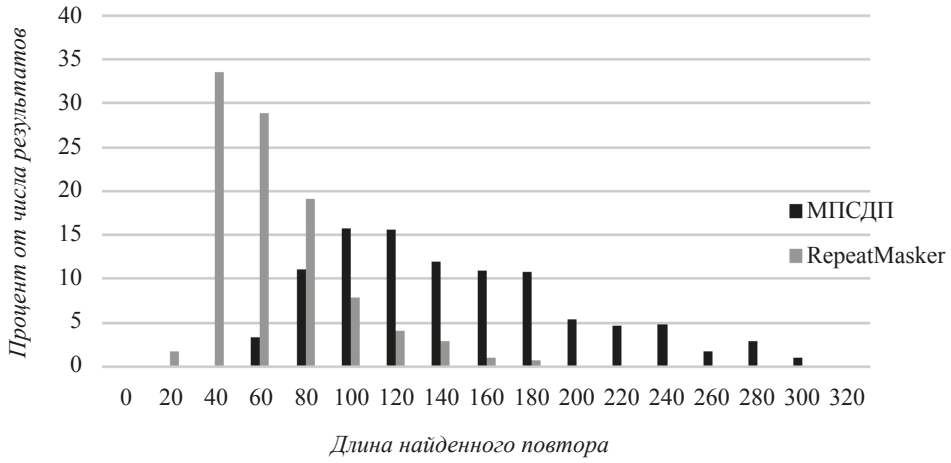


**Рис. 2.** Распределения длин найденных копий SINE элементов уникальных для результатов программы RepatMasker и программы МПСДП.

**Fig. 2.** The distribution of length of the found copies of SINE elements unique to the results of the RepatMasker program and the MPDSD programs.

SINE повторов в геноме можно рекомендовать использовать совместно программы МПСДП и RepeatMasker.

Разработанный метод является универсальным и подходит для поиска сильно дивергировавших копий повторов в различных геномах. МПСДП может быть использован для поиска в геноме не только SINE повторов, но и других транспозонов, при наличии набора консенсусных последовательностей, на основании которых можно построить ПВМ, которая будет использоваться для дальнейшего сканирования. Программа МПСДП доступна по запросу для некоммерческого использования.

ФИНАНСИРОВАНИЕ

Работа выполнена при частичной поддержке РФФИ (грант № 20-016-00057 А).

## ЛИТЕРАТУРА

1. Elbarbary R.A., Lucas B.A., Maquat L.E. Retrotransposons as regulators of gene expression. *Science*, 2016, 351(6274). doi: 10.1126/science.aac7247.
2. Casacuberta E., González J. The impact of transposable elements in environmental adaptation., *Mol. Ecol*., 2013, 22(6), 1503–17. doi: 10.1111/mec.12170.
3. Kumar A., Bennetzen J.L. Plant Retrotransposons. *Annu. Rev. Genet.,* 1999, 33, 479–532, doi:10.1146/annurev.genet.33.1.479.
4. Korotkov E. V, Korotkova M.A., Rudenko V.M. MIR-family of repeats common for vertebrate genomes. *Mol. Biol.*, 2000, 34(4), 553–559. doi: 10.1007/BF02759556
5. J S Tulko, E V Korotkov, D A Phoenix, MIRs are present in coding regions of human genes. *DNA. Seq.,* 1997, 8(1–2), 31–38. doi: 10.3109/10425179709020882.
6. de Koning A.P.J., Gu W., Castoe T.A., et.al. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet*., 2011, 7(12), doi: 10.1371/journal.pgen.1002384.
7. Kramerov D.A., Vassetzky N.S. Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, 2011, 107(6), 487–495. doi: 10.1038/hdy.2011.43
8. Vassetzky N.S., Kramerov D.A. SINEBase: A database and tool for SINE analysis. *Nucleic. Acids. Res*., 2013 41(Database issue), D83–9. doi: 10.1093/nar/gks1263.
9. Wenke T., Döbel T., Sörensen T.R., et.al. Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. *Plant Cell*., 2011, 23(9), 3117–3128. doi: 10.1105/tpc.111.088682.
10. Mao H., Wang H. SINE-scan: An efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics*, 2017, 33(5), 743–745. doi: 10.1093/bioinformatics/btw718.
11. Smit A., Hubley R., Green P. RepeatMasker Open-3.0, 1996 http://www.repeatmasker.org/.
12. Bao W., Kojima K.K., Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*., 2015, 6:11. doi: 10.1186/s13100-015-0041-9.
13. Pugacheva V., Korotkov A., Korotkov E. Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming. *Stat. Appl. Genet. Mol. Biol.,* 2016, 15(5), 381–400. doi: 10.1515/sagmb-2015-0079.

14. Suvorova Y.M., Korotkova M.A., Skryabin K.G., Korotkov E.V. Search for potential reading frameshifts in cds from Arabidopsis thaliana and other genomes. *DNA Res*. 2019, 26(2), 157–170. doi: 10.1093/dnares/dsy046.

15. Laskin A.A., Korotkov E. V, Chaley M.B., Kudryashov N.A. The Locally Optimal Method of Cyclic Alignment to Reveal Latent Periodicities in Genetic Texts: the NAD-binding Protein Sites. *Mol. Biol.*, 37, 561–570 (2003).

# Search for Highly Divergent SINE repeats in the Rice Genome

Yu. M. SUVOROVA[1*], A.M. KAMIONSKAYA[1], and E.V. KOROTKOV[1, 2]

[1] *Fundamentals of Biotechnology, Federal Research Centre, Russian Academy of Sciences., Moscow, 119071, Russia*

[2] *National Research Nuclear University MEPhI, Moscow, 115409, Russia*

*e-mail:*suvorovay@gmail.com

Received March 23, 2020
Revised April 17, 2020
Accepted July 24, 2020

**Abstract–**In this article, we present a new method for searching for highly divergent copies of SINE elements in a genome. The method is based on the correlation of neighboring symbols both in constructing a positional weight matrix for a sequence of interest and in the genome scanning procedure. This makes possible to increase the alphabet size and, accordingly, the resolution capacity of the method. Using it, we found new copies of SINE repeats in the rice genome that have not been annotated before. The method was tested and compared with the RepeatMasker program; false positives were evaluated.

*Key words:* SINE elements, retrotransposons, alignment-based methods, sequence similarity

**Funding–**The work was partly supported by the RFBR Grant no. 20-016-00057.

**doi:** 10.21519/0234-2758-2020-36-4-15-20