Зверков О.А.[1], Рубанов Л.И.[2] Селиверстов А.В.[3]

Институт проблем передачи информации им. А.А. Харкевича РАН, г. Москва
[1] н.с., zverkov@iitp.ru
[2] в.н.с., rubanov@iitp.ru
[3] в.н.с., slvstv@iitp.ru

Поиск ультраконсервативных элементов у простейших типа Apicomplexa

КЛЮЧЕВЫЕ СЛОВА

Ультраконсервативный элемент, простейшие, Apicomplexa, кластер.

АННОТАЦИЯ

В статье обсуждается кластеризация последовательностей и представлены результаты поиска на её основе ультраконсервативных элементов у простейших типа Apicomplexa.

В 2004 г. в геномах позвоночных были открыты ультраконсервативные элементы, сначала как точные повторы, затем как повторы, которые могут незначительно отличаться [1]. Первоначально был выделен 481 такой участок с сохранением синтении. Функциональная роль таких элементов до сих пор не определена, но их очень высокая консервативность, сложность и специфическое расположение в геноме указывают на вероятную роль в регуляции экспрессии генов эукариот. Ультраконсервативные элементы рассматриваются как новые источники филогенетической информации в геномах: они многочисленны, консервативны и соотнесены у далёких видов, они позиционно не пересекаются с большинством генных семейств с большим числом паралогов, редко содержат вставки мобильных элементов. Подобные характеристики указывают на их наследование от общего предка, жёсткий характер стабилизирующего отбора и функциональную значимость геномного контекста. С увеличением размера генома и сложности его организации, от дрожжей к позвоночным, увеличивается и доля ультраконсервативных элементов, приходящихся на некодирующую (возможно, регуляторную) часть генома. Например, найдено более 5000 уникальных ультраконсервативных элементов у птиц и рептилий [2], более 2000 – у мух и позвоночных [3]. Ультраконсервативные последовательности описаны у дрозофилид и у позвоночных животных с близким уровнем филогенетического расхождения видов, измеренного по кодирующим белки областям. Как длины, так и числа ультраконсервативных элементов оказались больше у позвоночных. Однако у простейших такие элементы до

сих пор не исследовались.

Причины консервативности остаются неизвестными. Возможно, часть из них является результатом горизонтального переноса от симбионтов или паразитирующих видов с широким распространением. Примером возбудителя протозойных инфекций, имеющего всесветное распространение служит *Toxoplasma gondii*, который относится к наиболее важным для медицины и ветеринарии представителям типа Apicomplexa. Церебральный токсоплазмоз, как оппортунистический паразитоз, занимает третье место в структуре летальных исходов при ВИЧ-инфекции.

Описание ультраконсервативных участков позволяет получать новые сведения о распространении возбудителей инфекций и путях заражения. В свою очередь это позволит целенаправленно проводить противоэпидемиологические мероприятия.

С другой стороны, такие участки могут служить маркёрами (зондами) для определения филогенетического положения малоизученных видов беспозвоночных животных. Это особенно актуально для видов, составляющих зоопланктон и в значительной степени определяющих продуктивность водных экосистем.

За последнее время значительно увеличилось число секвенированных геномов простейших, включая представителей типа Apicomplexa.

Геномные данные брались из базы данных Eukaryotic Pathogen Database Resources (http://eupathdb.org/). Нами выполнен широкомасштабный поиск ультраконсервативных элементов длиной от 90 п.н. и выше. Для этого из совокупности консервативных участков ДНК для анализа отобраны участки, полностью совпадающие у *Toxoplasma gondii* ME49 и *Neospora caninum*.

Найдено 30 таких участков, но подавляющее большинство из них перекрывает кодирующие области. Соответствующие гены были определены парным выравниванием ДНК против ДНК [4]. Наиболее длинные фрагменты (до 697 п.н.) расположены целиком в кодирующей области гена 28S рРНК на хромосоме IX или в гомологичных генах на хромосомах IX и Ia. Последние, возможно, представляют собой результаты недавних дупликаций генов рРНК.

Менее длинные консервативные участки перекрывают кодирующие области генов TGME49_245620 (рибосомный белок RPS27A), TGME49_301250 (гипотетический белок), TGME49_237130 (цитохром b), TGME49_237120 (гипотетический белок), TGME49_295710 (гипотетический белок), TGME49_242340 (рибосомный белок RPS29), TGME49_210690 (рибосомный белок RPS6), TGME49_254915 (гипотетический белок), TGME49_296010 (фосфатидалинозитол-3-4-киназа), TGME49_296000 (гипотетический белок), TGME49_266785 (белок, содержащий мотив zinc finger типа CCCH), TGME49_286090 (фактор инициации трансляции SUI1), TGME49_248480 (рибосомный белок RPS9) и TGME49_231140 (рибосомный

белок RPS25).

Пять ультраконсервативных участков из *Toxoplasma gondii* ME49 с длинами 166, 158, 143, 112 и 90 п.н. не перекрывают известные гены и являются кандидатами на роль ультраконсервативных некодирующих элементов. Координаты этих участков TGME49_chrV:2898396..2898561, TGME49_chrV:2898399..2898556, TGME49_chrVIIb:3224157..3224299, TGME49_chrIII:999998..1000109 и TGME49_chrVIIa:371446..371535.

Первый из них наиболее консервативен среди видов типа Apicomplexa. Точное совпадение наблюдается у штаммов *T. gondii* FOU, p89, VAND, TgCATBr9, RUB, GAB2-2007-GAL-DOM2, CtCo5, GT1, ME49, VEG, а также у видов *N. caninum* Liverpool и *Hammondia hammondi* H.H.34. Этот элемент, но уже с тремя заменами нуклеотидов представлен у *Sarcocystis neurona* SN3. Близкие последовательности, но с бóльшим числом замен, найдены у кокцидий *Eimeria maxima* Weybridge и *E. praecox* Houghton и у пироплазмид: *Babesia microti* RI, *B. bovis* T2Bo, *Theileria equi* WA, *Th. orientalis* Shintoku, *Th. parva* Muguga, *Th. annulata* Ankara. Этот элемент с ещё бóльшими отличиями (по сравнению с *T. gondii*) наблюдаются у *Cryptosporidium hominis* TU502, *C. parvum* Iowa II и *C. muris* RN66. Последовательности этого элемента с одной делецией выравниваются у *Plasmodium gallinaceum* 8A, *Pl. vivax* Sal-1, *Pl. cynomolgi* B, *Pl. knowlesi* H, *Pl. falciparum* 3D7, *Pl. falciparum* IT, *Pl. chabaudi* chabaudi, *Pl. berghei* ANKA, *Pl. yoelii* yoelii 17XNL и *Pl. yoelii* yoelii YM. У *Plasmodium* spp. эти последовательности хорошо выравниваются друг с другом.

Также найдены последовательности этого ультраконсервативного элемента с одной делецией у различных видов эймерий: *Eimeria tenella* strain Houghton, *E. maxima* Weybridge, *E. brunetti* Houghton, *E. necatrix* Houghton, *E. falciformis* Bayer Haberkorn 1970, *E. mitis* Houghton, *E. acervulina* Houghton. Пример выравнивания показан на рис. 1.

```
T: 4      aattaaacttacctggcaggcgctcggggtggctcacgcatcaccctgtcgtagttcgga 63
          ||||||||||||||||||||||||||||||||||   ||||||||||||||||| ||||
E: 24397  aattaaacttacctggcaggcgctcggggtggctcgtccatcaccctgtcgtagtccgga 24338

T: 64     gcagggcactgcactctgctg-ctgtgatgagctatgggctcccaatcgggggtgccaac 122
          |||| |||||||||||| |||| |||| ||| ||||||||| ||| |||||||||||
E: 24337  gcagagcactgcactcagctgtctgtaatgggctatgggcctccatatgggggtgccaac 24278

T: 123    tgcagaatttctggtagcggcaggttgcgttcgcgc 158
          ||||||||||||| |||||||| ||||||||||||
E: 24277  tgcagaatttctgatagcggcatgttgcgttcgcgc 24242
```

Рис.1. Выравнивание ультраконсервативных элементов из Toxoplasma gondii ME49 и Eimeria acervulina Houghton (локус HG670352). Длина выравнивания 156 из 166 поданных на вход программы blast. Совпадает 89% нуклеотидов (140/156), делеция состоит из одного нуклеотида

Другие найденные ультраконсервативные элементы также имеют близкие по последовательности участки в большинстве видов типа Apicomplexa. Количество нуклеотидных замен не более 11%.

Для анализа найденных консервативных элементов был проведён

поиск соответствий по базе данных Rfam [5]. Три из рассмотренных пяти консервативных элементов в *T.gondii* ME49 (два перекрывающихся элемента на хромосоме V и ещё один на хромосоме VIIb) гомологичны РНК U1, участвующей в сплайсинге и регулирующей разнообразие изоформ [6–7]. Четвёртый элемент (на хромосоме III) гомологичен РНК U6, также участвующей в сплайсинге [8–11]. Самый короткий элемент, который расположен на хромосоме VIIa, не имеет значительного сходства с известными образцами из базы Rfam. Результаты поиска в базе Rfam приведены в таблице.

Длина	Позиции в геноме *T.gondii* ME49	Rfam	Bits score
166	TGME49_chrV:2898396..2898561	U1	89.8
158	TGME49_chrV:2898399..2898556	U1	80.9
143	TGME49_chrVIIb:3224157..3224299	U1	69.6
112	TGME49_chrIII:999998..1000109	U6	102.4
90	TGME49_chrVIIa:371446..371535	N/A	

Отметим, что пять перечисленных ультраконсервативных элементов специфичны для типа Apicomplexa и не обнаружены у других простейших. Поэтому они могут служить филогенетическими маркёрами для установления принадлежности видов таксономическим группам.

В биоинформатике часто рассматривают взвешенные многодольные графы, у которых вес каждого ребра отражает сходство последовательностей, приписанных его концам. Доли соответствуют видам. Ультраконсервативным элементам соответствуют такие *m*-плотные подграфы взвешенного многодольного графа, которые содержат рёбра большого веса. Последние условия предполагают заданными некоторые пороги. Такие подграфы называют *кластерами*.

На этой основе проведён поиск ультраконсервативных элементов для списка из 10 хорошо собранных геномов споровиков: *Toxoplasma gondii* ME49, *Neospora caninum* Liverpool, *Sarcocystis neurona* SN3, *Eimeria tenella* Houghton, *Cryptosporidium parvum* Iowa II, *Plasmodium falciparum* 3D7, *Babesia bovis* T2Bo, *Theileria parva* Muguga, *Th. annulata* Ankara и *Gregarina niphandrodes*. Из них учитывались только контиги длиной 500 п.н. и более.

При попарном выравнивании участков длиной не менее 150 п.н. допускалась величина штрафа не более 20 (при штрафе за несовпадение 1 и постоянном штрафе за делецию 5.1). Практически во всех подходящих парах встречалось не более 3 делеций. Кроме того, участки отбирались по сложности (коэффициент сжатия не более 2.5). Всего подходящих пар, т.е. рёбер предварительного графа, нашлось 23887.

Далее проводилась склейка вершин из одного вида, участвующих в различных ребрах, исходя из величины перекрытия соответствующих участков генома (например, 120).

Если требовать, чтобы каждая вершина кластера в графе была смежна с 6 или более долями (когда параметр алгоритма поиска кластеров m=7), то ответом является пустое множество (нет ни одного кластера).

Если же ослабить это требование, то найдутся кластеры с представителями всех 10 долей. Наиболее интересен вариант *m*=3, в котором найдено 3 кластера с 10 видами, по одному с 9 и 8 видами, 6 кластеров с 7 видами, 3 кластера с 6 видами, 5 кластеров с 5 видами, 1 кластер с 4 видами и 7 кластеров с 3 видами. Однако среди соответствующих последовательностей встречаются участки большой субъединицы рибосомной РНК.

Дальнейшее изучение ультраконсервативных элементов предполагает одновременное изучение последовательностей фланкирующих их локусов сегментов некодирующей ДНК в геномах. Анализ полногеномных данных позволяет изучать редкие явления или сочетания явлений, которые проливают свет на различные аспекты межгенных взаимодействий. Идея анализа в следующем: синергический эпистаз между вредными мутациями в пределах одного генотипа должен приводить к изменениям в распределении числа таких мутаций на генотип. Конкретно, поскольку синергический эпистаз на приспособленность приводит к избирательному действию отбора против особей, несущих большое число вредных мутаций, он должен снижать дисперсию числа вредных мутаций на генотип по сравнению с ожидаемой в отсутствие эпистаза. Другими словами, при эпистазе особи, несущие вредные мутации сразу во многих локусах, должны быть более редки в популяции, чем ожидается.

Это теоретическое предсказание можно проверить, используя данные по генетической изменчивости в природных популяциях. Поскольку для подсчета числа вредных аллелей в генотипе требуются полные генотипы, а для обнаружения сигнала эпистаза – анализ больших выборок генотипов, данные, необходимые для подобного исследования, начали появляться только недавно. А именно нужно рассматривать различные классы полиморфизмов, против которых ожидается действие отбора разной силы: от максимального (например, нонсенс-мутации, радикальные миссенс-мутации в консервативных аминокислотных сайтах, мутации в ультраконсервативных некодирующих позициях) до минимального (синонимические мутации). Отклонение распределения числа вредных аллелей на геном от ожидаемого будет означать действие синергического эпистаза против вредных мутаций на уровне всего генома.

Помимо глобального, в геноме также может действовать локальный эпистаз между мутациями в пределах одного функционального элемента (гена, ультраконсервативного некодирующего элемента, сайта связывания фактора транскрипции и т.п.). Свойства локального и глобального эпистаза, по-видимому, радикально различаются: локальный эпистаз определяется необходимостью действия функционального элемента как целого.

Взаимодействия между вредными аллелями в пределах функционального элемента могут быть как синергическими, так и антагонистическими. Антагонистический (положительный) эпистаз можно ожидать для мутаций большого эффекта: например, в ситуации, когда одна

замена в ключевой позиции сайта посадки транскрипционного фактора может полностью «выключать» функцию такого сайта, так что дальнейшее накопление мутаций в этом сайте будет лишь очень слабо вредным или даже нейтральным. С другой стороны, слабовредные мутации малого эффекта могут взаимодействовать синергически, так что приобретение одной мутации делает более вредными мутации, накапливающиеся после неё. На отдельных функциональных элементах наблюдался как синергический, так и антагонистический эпистаз; однако понимание вклада того и другого фактора в изменчивость можно получить только полногеномным анализом.

Работа выполнена при частичной финансовой поддержке РФФИ (проект 13-04-40196-Н).

Литература

1. Bejerano G.., Pheasant M., Makunin I., Stephen S., Kent W.J., Mattick J.S., Haussler D. Ultraconserved elements in the human genome // Science. – 2004. – V. 304 (5675). – P. 1321–13254.
2. Faircloth B.C., McCormack J.E., Crawford N.G.., Harvey M.G.., Brumfield R.T., Glenn T.C. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales // Systematic biology. – 2012. – V. 61(5). – P. 717–726.
3. Makunin I.V., Shloma V.V., Stephen S.J., Pheasant M., Belyakin S.N. Comparison of Ultra-Conserved Elements in Drosophilids and Vertebrates // PloS one – 2013. – V. 8(12), e82362.
4. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. – 1997. – V.25. – P.3389–3402.
5. Burge S.W., Daub J., Eberhardt R., Tate J., Barquist L., Nawrocki E.P., Eddy S.R., Gardner P.P., Bateman A. Rfam 11.0: 10 years of RNA families // Nucleic Acids Research. – 2012. – doi: 10.1093/nar/gks1005.
6. Zwieb C. The uRNA database // Nucleic Acids Research. – 1997. – V. 25 (1). – P. 102–103.
7. Berg M.G., Singh L.N., Younis I., Liu Q., Pinto A.M., Kaida D., Zhang Z., Cho S., Sherrill-Mix S., Wan L., Dreyfuss G. U1 snRNP determines mRNA length and regulates isoform expression // Cell. – 2012. – V. 150 (1) – P. 53–64.
8. Brow D.A., Guthrie C. Spliceosomal RNA U6 is remarkably conserved from yeast to mammals // Nature. – 1988. – V. 334 (6179). – P. 213–218.
9. Marz M., Kirsten T., Stadler P.F. Evolution of spliceosomal snRNA genes in metazoan animals // J. Mol. Evol. – 2008. – V. 67 (6). – P. 594–607.
10. Butcher S.E., Brow D.A. Towards understanding the catalytic core structure of the spliceosome // Biochem. Soc. Trans. – 2005. – V. 33 (Pt 3). – P. 447–449.
11. Karaduman R., Dube P., Stark H., Fabrizio P., Kastner B., Lührmann R. Structure of yeast U6 snRNPs: arrangement of Prp24p and the LSm complex as revealed by electron microscopy // RNA. – 2008. – V. 14 (12). – P. 2528–2537.