МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2012, том 46, № 2, с. 361–366

БИОИНФОРМАТИКА

УДК 575.113

ПРЕДВАРИТЕЛЬНАЯ ВЕРСИЯ ГЕНОМА МАКРОНУКЛЕУСА ИНФУЗОРИИ *Euplotes crassus*

© 2012 г. Д. В. Виноградов[1]*, О. В. Цой[1, 2], А. В. Заика[1, 2], А. В. Лобанов[3], А. А. Туранов[3], В. Н. Гладышев[3], М. С. Гельфанд[1, 2]

[1]*Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, 127994*

[2]*Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова, Москва, 119991*

[3]*Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115, USA*

Поступила в редакцию 06.06.2011 г.
Принята к печати 28.09.2011 г.

Проведен базовый биоинформатический анализ предварительной версии макронуклеарного генома и транскриптома инфузории *Euplotes crassus*. Более четверти генов *E. crassus* содержат несколько экзонов. Среди интронов обнаружена большая фракция "крошечных" интронов длиной 20–30 н. Определено 63 случая альтернативного сплайсинга, а также найдено 14 интронов с нестандарными сайтами сплайсинга. Около 2000 предполагаемых генов *E. crassus* не имеют каких-либо гомологов в других инфузориях, но в большинстве из них наблюдается значительное сходство с бактериальными генами, что может быть результатом загрязнения при приготовлении образцов из инфузорий. Сравнение геномов *E. crassus* и других инфузорий свидетельствует об экспансии одних и тех же групп генов, отвечающих за жизнедеятельность свободноживущих гетеротрофных организмов.

***Ключевые слова*: анализ генома, инфузория, геном, биоинформатика.**

DRAFT MACRONUCLEAR GENOME OF A CILIATE *Euplotes crassus*, by *D. V. Vinogradov*[1]*, *O. V. Tsoy*[1, 2], *A. V. Zaika*[1, 2], *A. V. Lobanov*[3], *A. A. Turanov*[3], *V. N. Gladyshev*[3], *M. S. Gelfand*[1, 2] ([1]Kharkevich Institute for Information Transmission Problems, Russian Academy of Science, Moscow, 127994 Russia, *e-mail: dimavin@bioinf.fbb.msu.ru; [2]Department of Bioengineering and Bionformatics, Moscow State University, Moscow, 119991 Russia, *e-mail: bioeng@genebee.msu.ru; [3]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, 02115 USA). Basic bioinformatical analysis of the draft *Euplotes crassus* macronuclear genome and transcriptome suggests that more than a quarter of *E. crassus* genes contain several exons. A large fraction of all introns is formed by "tiny" introns having length 20-30 bp. Analysis of the transcriptome revealed 63 possible cases of alternative splicing, and also 14 introns with non-standard splicing sites. About 2000 hypothetical genes do not have homologs in other ciliates, and since most of them have the closest homologs in bacterial genomes, they are likely an artifact of the sample preparation. Comparison of the *E. crassus* genome to the genomes of other ciliates showed an expansion of the same gene families, responsible for the free-living heterotrophic lifestyle.

***Keywords*: genome analysis, ciliate, bioinformatics.**

Геном инфузории *Euplotes crassus* (класс *Spirotrichea*) обладает рядом особенностей, которые отличают его от геномов остальных инфузорий и определяют интерес к его изучению.

Хотя инфузории используются как модельные организмы для изучения таких процессов, как РНК-катализ, образование теломер, ацетилирование гистонов, их геном изучен мало: в базах данных есть полногеномные последовательности только двух инфузорий – *Tetrahymena thermophila* [1] и *Paramecium tetraurelia* [2]. Ряд особенностей генома инфузорий затрудняет секвенирование и анализ, и, в первую очередь, вследствие ядерного диморфизма (макронуклеус и микронуклеус). Микронуклеус в вегетативном состоянии неактивен, но участвует в передаче генетического материала между поколениями. Макронуклеус активен во время вегетативного роста инфузории. У

Принятые сокращения: EST (Expressed Sequence Tags) – фрагменты экспрессированных последовательностей; ЭИК – экзон-интронные контиги; н. – нуклеотид.

*Эл. почта: dimavin@bioinf.fbb.msu.ru

Основные характеристики результатов секвенирования и сборки генома и последовательностей EST макронуклеса *Euplotes crassus*

	Число	Длина			
		минимальная	средняя	N50	максимальная
Контиги	70328	43	937.75	595	15440
EST	18742	200	533.56	512	3165

большинства из таксонов он характеризуется высокой степенью полиплоидии. При половом процессе он разрушается и заменяется новым, который образуется из микронуклеуса путем перестроек ДНК.

Все представители рода *Euplotes* обладают еще одной отличительной особенностью – их генетический код отклоняется от универсального. Кодон UGA не является стоп-кодоном, а отвечает за включение цистеина [3] или селеноцистеина [4]. Кроме того, у *Euplotes* механизмы трансляции белков отличаются от универсальных: для трансляции некоторых генов требуется программируемый сдвиг рамки считывания [5]; детали этого процесса неизвестны. Предложен гипотетический механизм для сдвигов типа +1 [5] в тех случаях, когда транслируемая последовательность заканчивается слабыми стоп-кодонами, за которыми располагается сдвигающая последовательность.

Впервые подобный механизм описан при изучении гена, кодирующего рекомбиназу в составе транспозона Tec2 в *Euplotes* [6]. Это далеко не единственный пример, и, возможно, программируемый сдвиг рамки считывания достаточно часто наблюдается как механизм трансляции у *Euplotes*. Ранее при анализе неполного генома Клобутчер (Klobutcher) [5] предположил, что более 5% генов этого организма обладают такой особенностью.

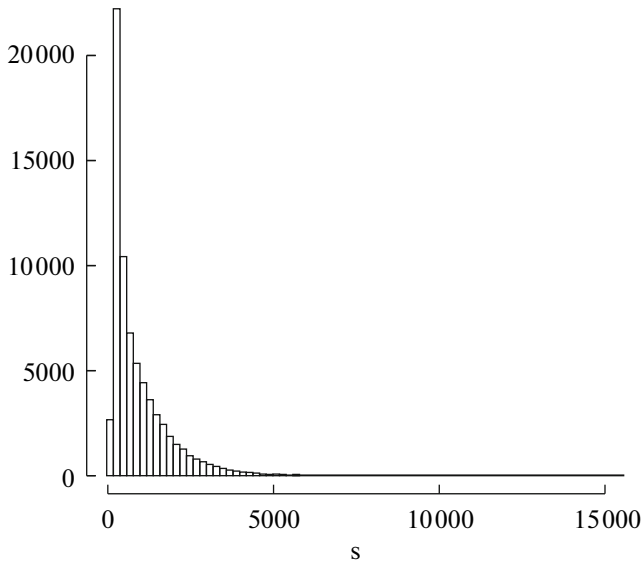

Рис. 1 Распределение длин контигов *E. crassus*. По оси *X* – длина контига (н.), по оси *Y* – число контигов указанной длины.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Библиотеку кДНК штамма *E. crassus* CT5 получили из клеток, находящихся в логарифмической фазе роста; для питания клеток использовали культуру одноклеточных водорослей *Dunaliella salina*. Для создания библиотеки использовали вектор pDNR-LIB, а ее конструирование осуществляли по протоколу SMART компании Clontech (“Mountainview”, США). Рекомбинатные клоны помещали в лунки, содержащие 1.2 мл среды с добавлением канамицина (30 мкг/мл). После культивирования в течение 18 ч при температуре 37°C плазмидную ДНК выделяли при помощи лизиса с использованием оборудования фирмы “Beckman Biomek FX”. Полученные клоны секвенировали по Сэнгеру с использованием реагентов BigDye Terminator (“Applied Biosystems”, США). Пробы для секвенирования осаждали изопропанолом и суспендировали в 7 мкл карбамальдегида. Последовательности определяли на капиллярном секвенаторе Applied Biosystems 3730XL 96. Из 8400 клонов получили примерно 15000 последовательностей EST, которые являются частью двух тысяч транскриптов.

Макронуклеарный геном *E. crassus* секвенировали на оборудовании компании 454 Life Science (www.454.com). Общее число последовательностей составило 2550648 (средняя длина 236 н.). Для создания геномной последовательности использовали программный пакет PCAP [7]. Контиги собирали в соответствии с инструкцией, прилагаемой к пакету. Базовые статистические характеристики используемой сборки приведены в табл. 1 и рис. 1.

Полученные EST картировали на контиги методом сплайсированного выравнивания, реализованного в программе Pro-EST [8]. Выравнивания с низким уровнем сходства (<95% вдоль всего выравнивания или <90% хотя бы в одном экзоне) отбрасывали. Выравнивания, содержащие интроны с нестандартными сайтами, обрабатывали вручную. Минимальную длину интрона ограничивали

16 н., в соответствии с данными о длине интронов в других представителях Ciliates [9, 10].

Неполные экзон-интронные структуры, полученные в результате выравнивания EST и контигов, объединяли в *экзон-интронные контиги* (ЭИК). При этом использовали следующую итеративную процедуру. Две структуры ЭИК объединяли, если все сайты сплайсинга (минимум два) в их пересечении в точности совпадали. В частности, в перекрывающемся участке не допускали удлинения экзонов за счет альтернативных сайтов, кассетные экзоны, присутствующие в одном ЭИК, но отсутствующие в другом, а также удержанные интроны.

Теломеры на концах контигов искали с помощью шаблона c{3-6}a{3-6}c{3-6}a{3-6}. Хотя классической последовательностью теломер является c{4}a{4}c{4}a{4} [15], технология пиросеквенирования часто допускает ошибки именно при определении длины гомополимерного участка [16], поэтому в таких участках и допускали вариации.

Поиск гомологов в базе RefSeq [11] проводили при помощи программы blastx из пакета BLAST [12].

Используемые для сравнения полные геномы *Paramecium tetraurelia* и *Tetrahymena thermophila* взяты из баз данных ParameciumDB [9] и Tetrahymena Genome Database [10] соответственно.

Паралогичные семейства выделяли в два этапа. На первом находили группы ортологичных генов, которые составили ядро паралогичного семейства. Для автоматического определения ортологов использовали алгоритм двусторонних лучших гомологов: ортологами считаются гены, которые в обе стороны являются друг для друга лучшими гомологами (BeT, best hit [13]). Все попарные сравнения проводили с использованием программы blastx из пакета Blast [12] при величине e-value $< 10^{-5}$. Из-за многочисленных дупликаций в геномах инфузорий не исключено наличие коортологов и, как следствие, несколько типов ортологичных соотношений: один ген или имеет одного ортолога в другом организме, или много ортологов, или целая группа генов может соответствовать группе генов.

На следующем этапе строили паралогичные семейства – дополнение ядра ортологичных генов паралогами. Паралогичные гены определяли по следующему критерию: сходство между ними должно было быть больше, чем сходство с наиболее близким ортологом из другого организма.

При анализе функционального разнообразия паралогичных семейств их разделяли на несколько групп, соответствующих метаболизму аминокислот, компонентам цитоскелета, энергетическому метаболизму, внутриклеточному транспорту, синтезу липидов, транспортерам, протеолитическим ферментам, белкам сигнальных путей (киназы и фосфатазы), комплексам транскрипции и трансляции, а также белкам с неизвестной функцией.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Статистический анализ предварительного генома E. crassus

В результате сборки генома получено 70328 контигов (из них 70312 уникальных) и 18742 последовательностей EST (из них 15170 уникальных).

После завершения процедуры объединения в большинстве контигов (4244) содержался один ЭИК, в 246 – два ЭИК, и в 5 контигах – три. Случаи нескольких ЭИК на контиг распределились так: 197 пар ЭИК не пересекаются друг с другом, один – целиком лежит в интроне другого ЭИК, и еще 63 пары, вероятно, представляют собой примеры альтернативного сплайсинга. В первом случае не вполне ясно, являются ли эти ЭИК двумя различными генами или (что более вероятно) артефактами, вызванными недостаточным покрытием EST.

Около четверти ЭИК содержат более одного экзона. Поскольку имеющиеся EST покрывают лишь небольшую часть генома, доля многоэкзонных генов, скорее всего, превышает 25%.

Гистограмма длин экзонов (рис. 2) выходит на плато в районе 30–200 н., далее следует пик на участке 200–250 н., а затем наблюдается плавное снижение. Максимальная длина внутреннего экзона – 2211 н., минимальная – 7 н.

Гистограмма длин интронов (рис. 3) имеет резкий пик в районе “крошечных” (tiny) интронов (23–30 н.) и затем убывает почти монотонно. Длина больше 300 н. встречается крайне редко, а около 40% интронов короче 30 н. Тем не менее, максимальная длина интрона – 2821 н., а минимальная – 21 н. (более короткие интроны встречаются в выдаче программы, но их отбрасывали как артефактные после индивидуального анализа). Стоит отметить, что в области коротких длин распределение длин интронов выглядит бимодальным, причем второй (меньший) пик приходится на значения длин 37–45 н. Это наблюдение можно объяснить тем, что имеются два механизма сплайсинга, и, как следствие, две группы интронов; граница длины между этими группами – 33–35 н.

Примерно 69662 контига *E. crassus* имеют значимых гомологов в других организмах. Из них около 2000 не имеют каких-либо гомологов и в других инфузориях. Таксономическое распределение последних представлено на рис. 4. Заметно, что большинство из этих генов имеют наиболее сходные гомологи в бактериальных геномах, причем во многих случаях – с очень высоким уровнем сходства (>75% одинаковых аминокислотных остатков). У подавляющего большинства таких контигов (~1400) отсутствуют теломеры (см. ни-

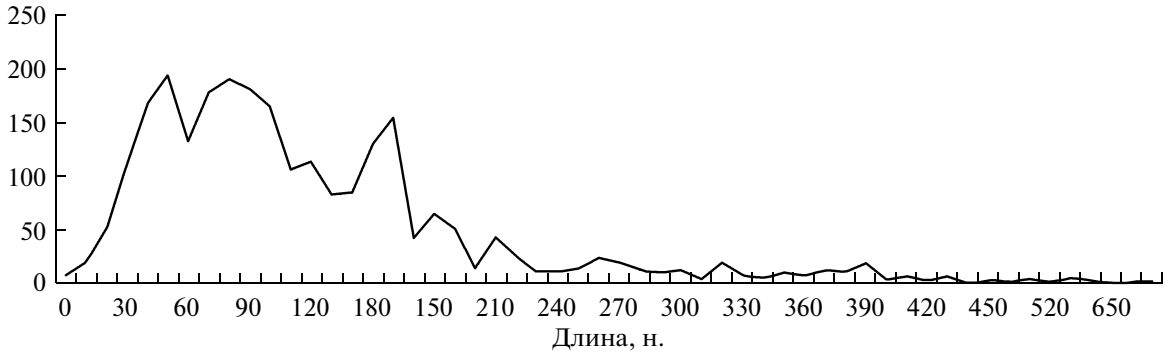

Рис. 2. Распределение длин экзонов *E. crassus*. По оси *X* – длина экзона (н.), по оси *Y* – число экзонов указанной длины.

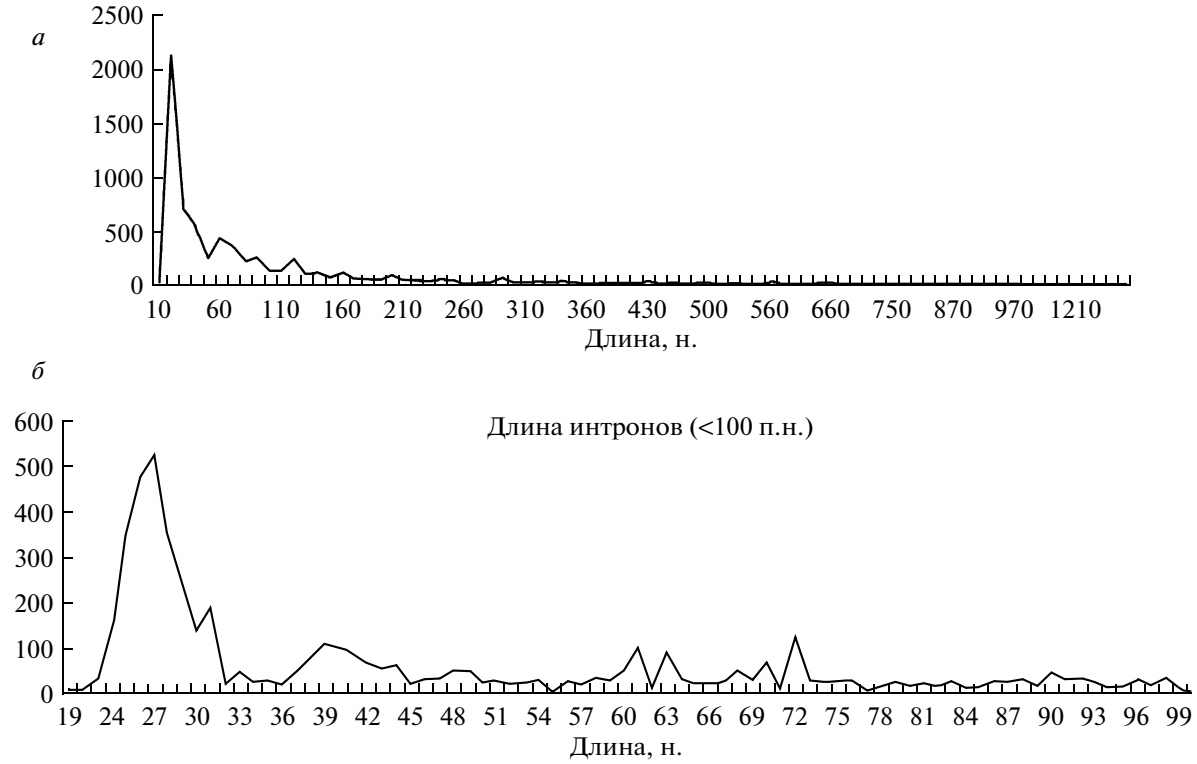

Рис. 3. Распределение длин интронов *E. crassus*. По оси *X* – длина интрона (н.), по оси *Y* – число интронов указанной длины. *а* – Все интроны. *б* – Интроны не длиннее 100 н.

же). Таким образом, наиболее вероятно, что появление этих контигов – следствие бактериального загрязнения исходного материала.

Альтернативный сплайсинг

Найдено 63 примера альтернативного сплайсинга, среди них 55 относятся к типу "удержанный интрон", т.е. экзон одного ЭИК полностью или частично покрывает интрон другого ЭИК. В 48 случаях один из ЭИК представляет собой единственный экзон, и, строго говоря, нельзя гарантировать, что соответствующие EST надо относить к одному и тому же типу альтернативного сплайсинга. Более того, невозможно по формальным критериям отличить случай удержанного интрона (аналогичный случаям, наблюдаемым среди частично перекрывающихся ЭИК) от случая удлинения экзона из-за альтернативного сайта. Однако в семи случаях перекрытие было полным, что дает основания для отнесения таких вариантов к разряду удержанных интронов или к случаям неполного сплайсинга. Альтернативные сайты обнаружены в восьми парах ЭИК: один альтернативный донор, два альтернативных ак-

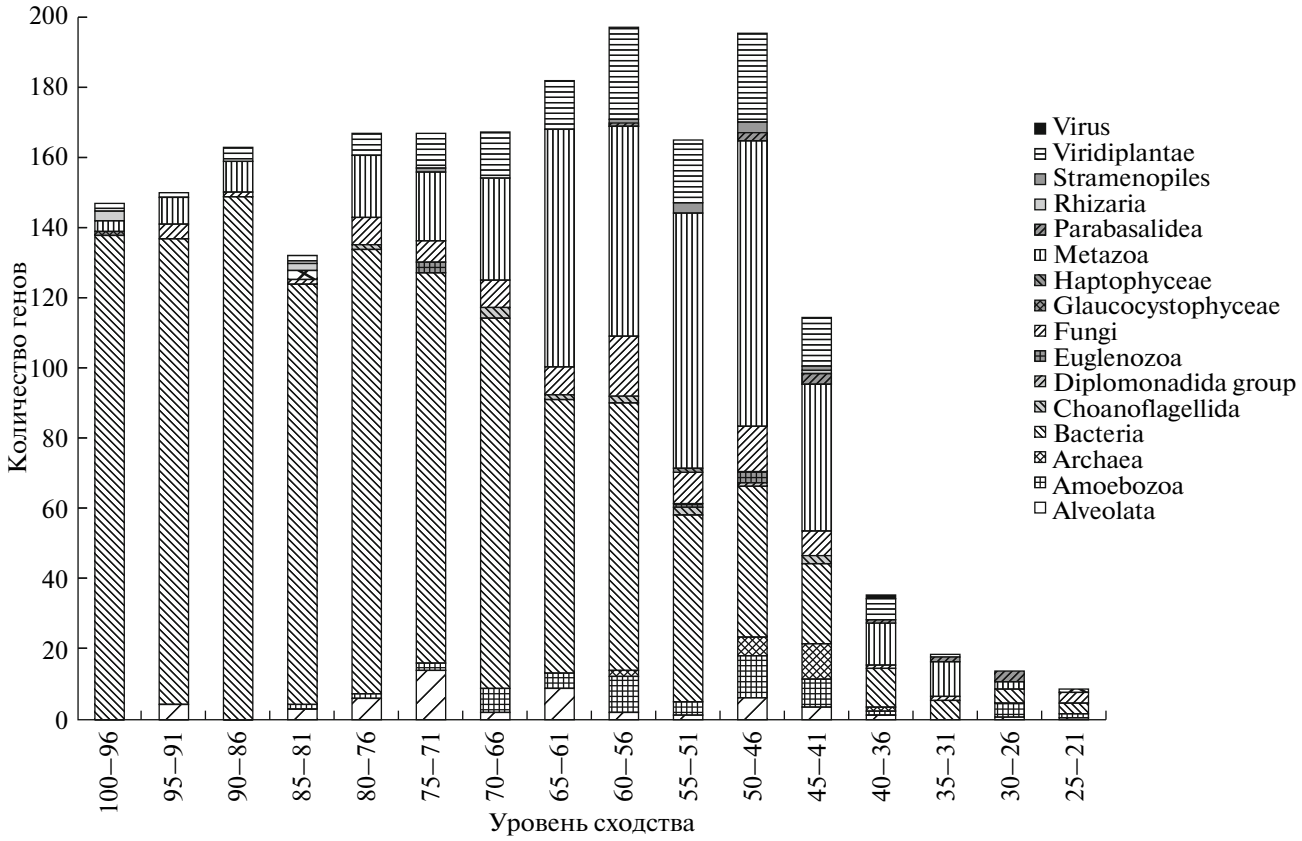

Рис. 4. Таксономическое распределение ближайших гомологов контигов *E. crassus*, не относящихся к инфузориям. По оси *X* – уровень сходства по идентичным аминокислотным остаткам (в %), по оси *Y* – число генов с таким уровнем сходства.

цептора и пять случаев, в которых варьировали оба сайта.

Нестандартные сайты сплайсинга

Все выравнивания, содержащие нестандартные сайты сплайсинга, проверяли вручную. После отбрасывания ненадежных сайтов и псевдоинтронов (вероятно, возникших из-за повторов, вызванных артефактами клонирования или ошибками при секвенировании) осталось 14 потенциальных интронов с нестандартными сайтами. Среди них: четыре интрона GC-AG, три – GT-TG, три AT-AG интрона, три GT-AT и один интрон GT-AC. Вывод о функциональности семи нестандартных акцепторных сайтов поддерживается тем, что перед всеми ними находится остаток T. Это согласуется с тем, что доля TAG среди интронов с каноническими сайтами составляет 93.63% (остальные – это в основном CAG, 4.73%)

"Одна хромосома – один ген"

При первичном анализе контигов было отмечено, что практически в каждом из них содержится всего один ген. При детальном рассмотрении 70328 контигов обнаружили всего 36 (0.05%) случаев, когда один контиг выравнивается более чем с одним белком из базы данных, с *e*-value < 0.001. Из этих 36 только один можно выровнять также с двумя различными EST.

При этом 12819 (18%) контигов ограничены теломерами с обеих сторон, т.е. с большой вероятностью соответствуют одной хромосоме макронуклеуса. Еще в 26773 (38%) контигах теломеры обнаруживаются с одного конца последовательности.

Таким образом, видимо, большинство хромосом *E. crassus* содержат один ген.

Паралогичные семейства

Вследствие многочисленных дупликаций генов, геном *E. crassus* содержит семейства паралогичных генов. Мы проанализировали функциональное разнообразие 1861 паралогичного семейства, общего для трех инфузорий. Эти семейства содержат 3781 ген *E. crassus*. Среди них оказалось больше всего генов, кодирующих белки сигнальных путей, а именно, киназы и фосфатазы. На их

долю приходится 13% от числа генов *E. crassus* в этих семействах, в то время как доля каждой из других функциональных групп была меньше 8%. Кроме того, самые обширные паралогичные семейства содержат гены, кодирующие белки, ответственные за жизнедеятельность свободноживущих гетеротрофных организмов: упомянутые выше гены сигнальных путей, а также компоненты цитоскелета (кинезины и динеины), транспортеры и протеолитические ферменты (катепсины).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Изучение предварительной версии генома *E. crassus* выявило ряд особенностей, не характерных для ранее изученных *T. thermophila* и *P. tetraurelia*: он содержит большое число крошечных интронов, а большинство хромосом содержит, видимо, один ген.

Анализ паралогичных семейств подтвердил независимую экспансию одних и тех же групп генов в различных инфузориях, связанную, очевидно, со сходным образом жизни. Таким образом, для инфузорий в целом необходимы такие же гены, какие, как было показано ранее, присутствуют в относительно большом количестве в каждой отдельной инфузории [1, 2]. Это согласуется с данными анализа генома *T. thermophila* [1].

Длина интронов у различных видов инфузорий может отличаться довольно сильно: если в *P. tetraurelia* большинство интронов имеет длину 18–35 н. ("крошечные"), а в *T. thermophila* – 53–978 н., то в *E. crassus* встречаются оба типа (длина варьирует от 21 до 2821 н.). Сплайсинг в целом достаточно развит: встречаются как нетривиальные альтернативные изоформы, так и неканонические сайты сплайсинга.

В последнее время становятся доступны неполные геномы других представителей этого таксона [14]. Это позволит расширить область исследования и провести детальный функциональный анализ генов, специфичных как для инфузорий в целом, так и для каждой инфузории в отдельности.

Авторы выражают благодарность А. А. Миронову за ценное обсуждение и комментарии.

Работа получила финансовую поддержку Государственных контрактов 14.740.11.0003, 2.740.11.0101 и 14.740.11.0738, Российского фонда фундаментальных исследований (09-04-92745) и Российской академии наук (программа "Молекулярная и клеточная биология").

СПИСОК ЛИТЕРАТУРЫ

1. Eisen J.A., Coyne R.S., Wu M., Wu D., Thiagarajan M., et al. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **4**, e286.
2. Aury J.M., Jaillon O., Duret L., Noel B., Jubin C., et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*. **444**, 171–178.
3. Meyer F., Schmidt H. J., Plumper E., Hasilik A., Mersmann G., Meyers H. E., Engstrom A., Heckmann K. 1991. UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. *Proc. Natl. Acad. Sci.* **88**, 3758–3761.
4. Turanov A.A., Lobanov A.V., Fomenko D.E., Morrison H.G., Sogin M.L., Klobutcher L.A., Hatfield D.L., Gladyshev V.N. 2009. Genetic code supports targeted insertion of two amino acids by one codon. *Science*. **323**, 259–261.
5. Klobutcher L.A. 2005. Sequencing of random *Euplotes crassus* macronuclear genes supports a high frequency of +1 translational frameshifting. *Eukaryot. Cell*. **4**, 2098–2105.
6. Krikau M.F., Jahn C.L. 1991. Tec2, a second transposon-like element demonstrating developmentally programmed excision in *Euplotes crassus*. *Mol. Cell Biol.* **11**, 4751–4759.
7. Huang X., Yang S.P. 2005. Generating a genome assembly with PCAP. *Curr. Protoc. Bioinformatics*. Oct; Chapter 11:Unit11.3.
8. Mironov A.A., Fickett J.W., Gelfand M.S. 1999. Frequent alternative splicing of human genes. *Genome Res.* **9**, 1288–1293.
9. Arnaiz O., Cain S., Cohen J., Sperling L. 2007. Paramecium DB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucl. Acids Res.* **35**, D439–D444.
10. Stover N.A., Krieger C.J., Binkley G., Dong Q., Fisk D.G., Nash R., Sethuraman A., Weng S., Cherry J.M. 2005 Tetrahymena Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. *Nucl. Acids Res.* **34**, D500–D503.
11. McEntyre J., Ostell J. 2002. *The NCBI handbook [Internet]*. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information.
12. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
13. Tatusov R.L., Koonin E.V., Lipman D.J. 1997. A genomic perspective on protein families. *Science*. **278**, 631–637.
14. Doak T.G., Cavalcanti A.R., Stover N.A., Dunn D.M., Weiss R., Herrick G., Landweber L.F. 2003. Sequencing *Oxytricha trifallax* macronuclear genome: a pilot project. *Trends Genet.* **19**, 603–607.
15. Jonsson F., Lipps H.J. 2000. The biology of telomeres in hypotrichous ciliates. *Madame Curie Bioscience Database [Internet]*. NCBI Bookshelf ID: NBK6195
16. Margulies M., Egholm M., Altman W.E., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. **437**, 376–380.