RIBOGROVE – БАЗА ДАННЫХ ПОЛНОРАЗМЕРНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ГЕНОВ 16S РРНК ПРОКАРИОТ

М. А. Сиколенко[1,2], Л. Н. Валентович[2,3]

[1]Кафедра системного анализа и компьютерного моделирования РФиКТ, Белорусский Государственный Университет, Минск, Беларусь
[2]Лаборатория "Центр аналитических и генно-инженерных исследований", Институт микробиологии НАН Беларуси, Минск, Беларусь
[3]Кафедра молекулярной биологии, биологический факультет, Белорусский Государственный Университет, Минск, Беларусь
E-mail: sikolenko@bio.bsu.by

Прокариотические гены 16S рРНК являются удобным и часто использующимся филогенетическим маркером. Существующие базы данных последовательностей генов 16S рРНК в основном содержат последовательности ПЦР-ампликонов, которые часто неполны, а иногда содержат артефакты ПЦР. Чтобы дополнить и расширить существующие общедоступные ресурсы, посвящённые разнообразию последовательностей 16S рРНК, была создана RiboGrove – база данных полноразмерных последовательностей генов 16S рРНК, извлечённых из полностью собранных геномов прокариот. Анализ полученных последовательностей позволил, например, оценить внутригеномную изменчивость генов 16S рРНК, а также определить таксоны, чьи гены 16S рРНК не содержат последовательностей анти-Шайн-Дальгарно.

Ключевые слова: *SSU rRNA, молекулярная таксономическая классификация микроорганизмов, Шайн-Дальгарно.*

Последовательности генов 16S рРНК часто используются для определения видового состава сообществ микроорганизмов и для установления филогенетических связей между отдельными группами прокариот [1]. Для этого используются общедоступные базы данных последовательностей генов 16S рРНК, такие как, например, Silva [2] и RDP [3]. Существующие базы данных естественным образом стремятся максимизировать количество последовательностей, содержащихся в них, и поэтому значительную долю данных в них составляют последовательности, полученные в результате секвенирования ампликонов, полученных путём амплификации ДНК, извлечённой из природных образцов. Вследствие этого последовательности генов, хранящиеся в таких базах данных, часто неполны, а также могут содержать артефакты, сформированные ПЦР [4].

Для преодоления вышеописанных ограничений была создана RiboGrove – общедоступная база данных полноразмерных последовательностей генов 16S рРНК прокариот: бактерий и архей. Последовательности генов были извлечены из полностью собранных геномов отдельных организмов, депонированных в базу данных RefSeq – популярную курируемую базу данных геномных последовательностей.

RiboGrove доступна по адресу http://mbio.bas-net.by/cager/en/ribogrove. На момент подготовки данного текста актуальной является версия RiboGrove 4.210 – четвёртый по счёту выпуск RiboGrove, основанный на данных RefSeq 210 – и именно она описывается в данном тексте. Объём данных, содержащихся в RiboGrove, представлен в Таблице 1.

Таблица 1

Размер базы данных RiboGrove

	Бактерии	**Археи**
Количество последовательностей генов	128 152	699
Количество уникальных последовательностей генов	36 349	504
Количество видов	6 743	321
Количество геномов	24 505	410

RiboGrove содержит последовательности только тех генов, которые входят в состав полностью собранных геномов, следовательно, выборка организмов, представленных в RiboGrove, не репрезентативна в отношении всего разнообразия прокариот. Напротив, в RiboGrove сверхпредставлены часто секвенируемые геномы. Так, например, один лишь вид *Escherichia coli* представлен 1 912-ю геномами, что составляет 7,7% всех геномов-источников последовательностей для RiboGrove.

Присутствие только полноразмерных последовательностей генов 16S рРНК в RiboGrove позволило оценить размеры генов и статистически их описать (см Таблицу 2).

Таблица 2

Размеры генов 16S рРНК, содержащихся в RiboGrove

	Бактерии	**Археи**
Минимальный размер, п.н.	1 448,00	1 439,00
Медианный размер, п.н.*	1 532,00	1 475,50
Средний размер, п.н.*	1 528,38	1 500,67
Среднеквадр. отклонение, п.н.*	26,04	149,64
Максимальный размер, п.н.	2 438,00	3 604,00

* – данные метрики были рассчитаны с предварительной нормализацией – вычислением медианного внутривидового размера генов – для уменьшения искажающего эффекта нерепрезентативности RiboGrove.

Данные RiboGrove показывают, что количество копий гена 16S рРНК в геномах прокариот может изменяться в довольно широких пределах: до 37 копий на геном включительно (такое число копий зафиксировано в геноме бактерий *Tumebacillus avium* AR23208). У архей же, при меньшем количестве геномов в выборке, и размах числа копий скромнее: до 5-ти копий включительно.

Иные способы численного описания нуклеотидных последовательностей также могут приводить к неожиданным результатам. Так, значения ГЦ- и АТ-перекосов [5] последовательностей генов 16S рРНК, содержащихся в RiboGrove, будучи нанесёнными на двумерный график (см. рисунок), демонстрируют сильную отрицательную корреляцию между данными величинами у архей и актинобактерий, но не у иных прокариот. Коэффициенты корреляции (Пирсона) для между ГЦ- и АТ-перекосами актинобактерий и архей равны 0,91 и 0,76, соответственно. Такой эффект уже был продемонстрирован ранее для архей [5]; актинобактериальная же корреляция ранее не была отмечена в публикациях, и она же более статистически значима. Однако несмотря на явно значимую корреляцию, ни в коем случае не очевидно, какая причинно-следственная связь может стоять за таким наблюдением, если таковая вообще существует.

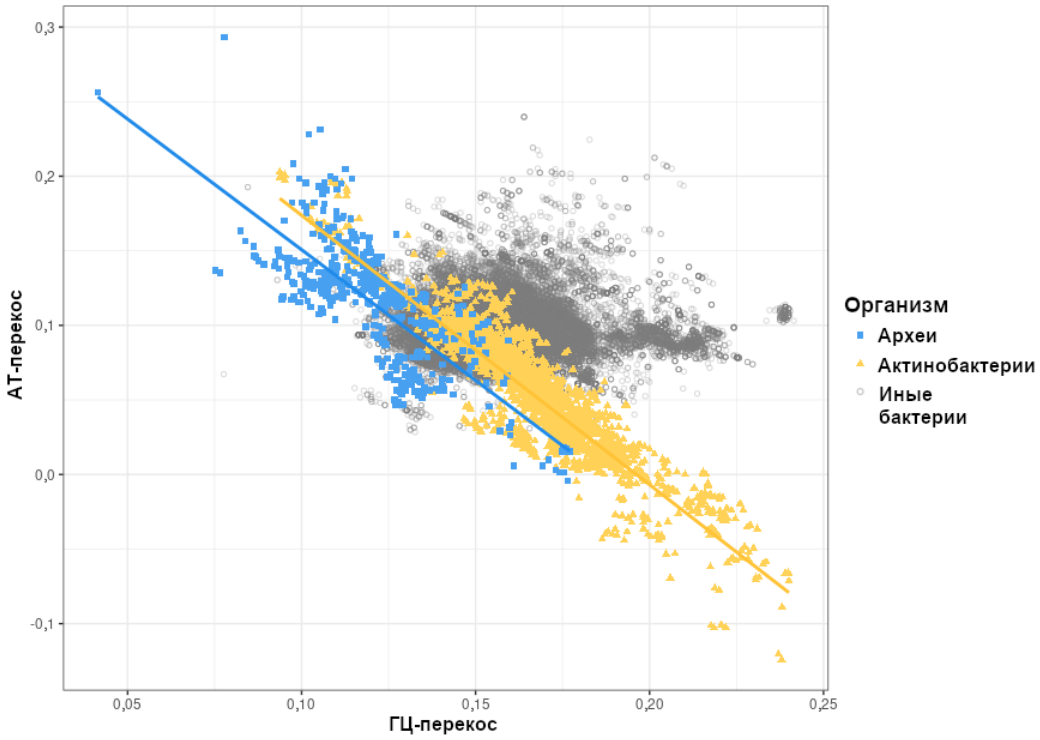

Рис. ГЦ- и АТ-перекосы последовательностей
генов 16S рРНК бактерий и архей

Известная геномная принадлежность каждого гена позволяет оценить внутригеномную вариабельность генов 16S рРНК. Она, будучи высокой, может вносить искажения в результаты оценки разнообразия сообществ микроорганизмов. Например, геном архей *Halomicrobium* sp. ZPS1 содержит два гена 16S рРНК, имеющих процент идентичности 90,70%, что при общепринятом значении в 95% для разделения прокариот по отдельным родам [1] напоминает, что данный порог – лишь договорённость, не всегда отражающая реальное родство.

Наличие в RiboGrove только полноразмерных генов позволяет изучать участки 16S рРНК, расположенные вне регионов, которые могут быть амплифицированы с помощью универсальных пар праймеров. Таковой является консервативная последовательность CCTCCT (т. н. последовательность анти-Шайн-Дальгарно), расположенная у самого 3′-конца транскрипта, и которая, как считается, участвует в процессе инициации трансляции у многих групп бактерий [6]. Анализ последовательностей, содержащихся в RiboGrove, показал, что высокая консервативность этой последовательности не распространяется на отдельные семейства отдела *Bacteroidetes*: 231 геном из 905-ти геномов бактерий этого порядка имел иные последовательности на месте CCTCCT в своих генах 16S рРНК.

Таким образом, созданная база данных RiboGrove может являться эталоном при формировании баз данных для таксономического анализа нуклеотидных последовательностей, делает возможным более точные подсчёты пропорций количества прокариотических клеток в метагеномных образцах и позволяет отслеживать значимые для трансляции участки рРНК у отдельных таксономических групп прокариот. Статья, описывающая RiboGrove и содержащая более подробную информацию, принята в печать журналом Research in Microbiology 23.02.2022 [4].

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Tindall B., Rosselló-Móra R., Busse H-J., et al. Notes on the characterization of prokaryote strains for taxonomic purposes // International Journal of Systematic and Evolutionary Microbiology. 2009. Vol. 60. № 1. P. 249-266. DOI: 10.1099/ijs.0.016949-0
2. Quast C., Pruesse E., Yilmaz P. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools // Nucleic Acids Research. 2013. T. 41, № D1. P. D590-D596. DOI: 10.1093/nar/gks1219
3. Cole J., Wang Q., Fish J. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis // Nucleic Acids Res. 2014. Vol. 42, № Database issue. P. D633-642. DOI: 10.1093/nar/gkt1244
4. Sikolenko M. A., Valentovich L. N. RiboGrove: a database of full-length prokaryotic 16S rRNA genes derived from completely assembled genomes // Res Microbiol. 2022. DOI: 10.1016/j.resmic.2022.103936
5. Guy L., Roten C-A. Genometric analyses of the organization of circular chromosomes: a universal pressure determines the direction of ribosomal RNA genes transcription relative to chromosome replication // Gene. 2004. Vol. 340, № 1. P. 45-52. DOI: 10.1016/j.gene.2004.06.056
6. Amin M., Yurovsky A., Chen Y., et al. Re-annotation of 12,495 prokaryotic 16S rRNA 3’ ends and analysis of Shine-Dalgarno and anti-Shine-Dalgarno sequences // PLOS ONE. 2018. Vol. 13, № 8. P. e0202767. DOI: 10.1371/journal.pone.0202767