ОТЧЕТ О ПРОДЕЛАННОЙ РАБОТЕ С ИСПОЛЬЗОВАНИЕМ ОБОРУДОВАНИЯ ИВЦ НГУ

Аннотация:

Микробная жизнь в геотермальных водах обнаружена еще в 1903 году, но возможность изучения ее в полной мере появилась только с развитием молекулярно-биологических методов. Благодаря изучению термофильных микроорганизмов были получены термостабильные ферменты, например полимеразы, лигазы, без которых невозможно представить современную молекулярную биологию. Сейчас изучение микробных сообществ геотермальных источников интересно в первую очередь с филогенетической точки зрения, так как все чаще и чаще находят новые микроорганизмы, которые помогают нам понять филогению прокариот, их родство с эукариотическими организмами и эволюцию жизни. В частности, в 2017 году была открыта новая группа Архей, названных Асгард археями, представители которой в настоящее время являются ближайшими прокариотическими родственниками эукариотам. Эти данные были получены при помощи метагеномных методов, которые позволяют получать наиболее полные данные о микробных сообществах, включая возможность описывать некультивируемые микроорганизмы.

Тема работы:

Метагеномный анализ микробных сообществ геотермальных источников острова Кунашир.

Состав коллектива:

- Розанов Алексей Сергеевич, к.б.н., н.с. ИЦИГ СО РАН; руководитель.
- Коржук Антон Владимирович, ФЕН НГУ, 4 курс, кафедра информационной биологии; лаборант ИЦИГ СО РАН; исполнитель.

Научное содержание работы:

1. Постановка задачи:

Осуществить обработку первичных метагеномных данных, сборку метагеномов и их анализ. В результате должны быть получены таксономические составы сообществ и экстрагированы некоторые геномы.

2. Современное состояние проблемы:

Исследование структуры микробных сообществ экстремальных экосистем, в том числе геотермальных, – одна из фундаментальных задач микробиологии. Анализ экстремофильных микробных сообществ позволяет выявлять закономерности организации прокариотических сообществ, не подверженных влиянию современных эукариот. Метагеномный подход позволяет получать наиболее полные данные о микробных сообществах, включая возможность описывать некультивируемые микроорганизмы.

Экстремальные экосистемы могут консервировать в себе виды организмов, предки которых населяли Землю в давние геологические эпохи. Поэтому их исследование может дополнить наши представления о происхождении и эволюции жизни на Земле. Наиболее интересными с точки зрения изучения ранних стадий эволюции считаются археи, населяющие горячие источники, так как они, с одной стороны, часто являются литотрофами, что должно было быть характерным для ранних стадий формирования жизни, с другой стороны, считается, что археи сыграли важную роль в появлении эукариот. Кроме того, адаптация микроорганизмов к экстремальным местообитаниям обуславливает высокую геномную и метаболическую гибкость микробных сообществ в этих экосистемах и делает термофилов и их термостабильные белки перспективными для некоторых промышленных и биотехнологических применений.

Южно-Курильские острова являются одной из самых труднодоступных территорий

России и мира из-за действующего на их территории пограничного режима. Поэтому микробные сообщества геотермальных источников Южных Курил практически не изучались микробиологически и совсем не изучались при помощи метагеномных подходов. Используемый нами подход полногеномного метагеномного секвенирования не применялся ранее в России для исследования природных объектов.

3. Описание работы:

Выделение ДНК проводилось при помощи набора NucleoSpin Soil компании Macherey-Nagel. Разрушение клеток производится физическим и химическим методами: перетиранием образца керамическими шариками и действием лизирующего раствора. В наборе имеется два альтернативных буфера для лизиса (SL1 и SL2) и специальная добавка (Enhancer SX), которую можно комбинировать с обоими буферами.

Количественный контроль выделенной ДНК проводился при помощи электрофореза ДНК в агарозном геле.

Библиотеки для секвенирования были подготовлены в Центре Геномных Исследований ИЦиГ СО РАН и проанализированы с помощью 2100 Bioanalyzer. Секвенирование произведено Центром Генетики и Репродуктивной Медицины "ГЕНЕТИКО" в г. Москве на платформе Illumina NovaSeq 6000 (парные прочтения длиной 100 bp).

Обработка данных секвенирования осуществлена программами FastQC и Trimmomatic. FastQC анализирует файлы формата FASTQ и выводит всю необходимую информацию о ридах, например, количество ридов, их длина, боксплоты распределения качества, содержание GC, сверхпредставленные последовательности (если они есть) и др. Trimmomatic – это быстрый многопоточный инструмент командной строки, использующийся для обрезки и фильтрации данных Illumina, а также для удаления технических последовательностей, которые могут представлять реальную проблему для дальнейшей работы с данными.

Сборка контигов de novo была проведена программой metaSPAdes, использующей алгоритм на основе графов де Брюйна. Информация о качестве сборки получена с помощью metaQUAST. Этот инструмент предоставляет информацию о количестве и длине контигов, значение N50 и др.

Кластеризация контигов была выполнена с помощью MaxBin, программы автоматической кластеризации по двум характеристикам последовательностей: покрытию и частоте тетрануклеотидов. MaxBin, используя модули FragGeneScan и HMMER3, предоставляет для каждого кластера найденные в нем маркерные гены, транслированные в аминокислотные последовательности. Сканирование бинов осуществляется по набору из 107 маркерных генов, присутствующих в единственной копии в геноме и содержащихся у 95% секвенированных бактерий. Часть этих генов универсальны, и встречаются также у архей и эукариот. Параллельно метагеномные данные были визуализированы программой VizBin. Алгоритм программы представляет геномные сигнатуры (нормированные частоты тетрануклеотидов) в виде векторов в 136-мерном (количество уникальных тетрануклеотидов) пространстве, которое затем преобразуется в двумерное пространство методом нелинейного уменьшения размеров BH-SNE (Barnes-Hut Stochastic Neighbor Embedding). Таким образом, VizBin предоставляет диаграмму рассеяния точек (объектов кластеризации - контигов) в виртуальном двумерном пространстве.

Для определения таксономической принадлежности бинов был написан скрипт на языке Python, который для каждого бина брал маркерные белки, найденные в нем, и направлял в BLASTP, а затем записывал результат в файл. Таким образом определялись ближайшие гомологи. Поиск производился по базе данных "nr". Представленность считалась по средним покрытиям бинов.

Филогенетический анализ проведен с использованием BLAST и MEGA6. Для определения таксономии и постройки филогенетических деревьев использовались белковые последовательности маркерных генов, найденные программой MaxBin для каждого бина. Деревья построены с использованием метода максимального правдоподобия.

Экстрагированные геномы проаннотированы сервером RAST. RAST использует GLIMMER3 для идентификации генов-кандидатов. Сначала гены-кандидаты сравниваются с белками из Subsystems при помощи аминокислотных k-меров. Если таким образом были найдены соответствия, то генам-кандидатам присваивается статус белок-кодирующего гена и назначается функциональная роль (≥1). Оставшимся генам-кандидатам присваивается функция и статус белок-кодирующего гена (если они не перекрывают уже определенные гены), используя BLASTP поиск по 30 ближайшим организмам. Последовательности рРНК выявляются с помощью BLASTN, а тРНК с помощью tRNAscan-SE. Полученные геномные данные экпортируются в форматах GenBank, EMBL, GFF3, GTF, Excel и TXT.

4. Полученные результаты:

Определен таксономический состав микробных сообществ ультракислого источника оз. Фауста и Третьяковского источника, о. Кунашир. Выявлено наличие множества потенциально новых прокариотических видов.

В образце донных отложений из оз. Фауста преобладают эукариоты, представленные единственным видом – *Cyanidioschyzon merolae*. Это небольшая одноклеточная булавовидная красная водоросль, обитающая в богатых сульфатами ультракислых горячих источниках. Отличается обладанием самым маленьким геномом из всех фотосинтезирующих эукариот и минималистичным строением: содержит одно ядро, один хлоропласт и одну митохондрию, не содержит вакуолей и клеточной стенки. Геном C. merolae секвенирован в 2004 г., и был первым полным секвенированным геномом водорослей.

В домене бактерий выявлены типы *Actinobacteria, Proteobacteria и Firmicutes*. Среди актинобактерий были обнаружены два класса: *Actinobacteria* и *Acidimicrobii*. Первый представлен родом *Mycobacterium*, виды которого характеризуются высокой устойчивостью к факторам внешней среды (выше только у спорообразующих бактерий), в том числе к кислым условиям и высокой температуре. Второй представлен штаммами, тесно связанными с термофильными ацидофильными железобактериями *Acidimicrobium ferrooxidans* и *Ferrimicrobium acidiphilum*, а также неклассифицированными микроорганизмами. Были обнаружены организмы из рода ацидофильных бактерий *Acidocella* (альфапротеобактерии), из рода фотосинтезирующих бактерий *Thiohalocapsa* (гаммапротеобактерии) и неклассифицированные дельтабактерии. Тип Firmicutes представлен видом, гомологичным *Sulfobacillus Thermosulfidooxidans* – умеренно термофильной ацидофильной бактерии, окисляющей железо и серу.

Среди архей преобладаеют *Euryarchaeota*, представленные организмами, близкими к видам *Aciduliprofundum boonei* и *Aciduliprofundum* sp. MAR08-339, и организмами из класса *Thermoplasmata*, а именно гомологичными экстремальным ацидофилам *Cuniculiplasma divulgatum* и *Thermoplasma acidophilum* и неклассифицируемыми термоплазмами. Также были обнаружены археи из типов *Candidatus Parvarchaeum, Candidatus Marsarchaeota, Thaumarchaeota* и группа организмов, гомологичная типу *Candidatus Micrarchaeota*.

В образце донных отложений Третьяковского источника на уровне доменов абсолютное большинство занимают бактерии – 99,95%. Доминирующим типом является *Proteobacteria,* который представлен классами: *Alphaproteobacteria* (29.18%), *Betaproteobacteria* (21.41%), *Gammaproteobacteria* (10.24%), *Deltaproteobacteria* (4.52%), *Epsilonproteobacteria* (0.98%), *Hydrogenophilali*a (0.43%) и неклассифицированными протеобактериями (33.25%). Класс *Hydrogenophilali*a был определен в 2017 в результате отделения от класса *Betaproteobacteria.* Микроорганизмы из класса *Hydrogenophilali*a являются термофилами, могут использовать молекулярный водород в качестве донора электронов, молекулярный кислород в качестве конечного акцептора электронов, а также некоторые могут использовать нитраты для процессов денитрификации. Тип *Deinococcus-Thermus* представлен родами *Meiothermus* (94.56%) и *Thermus* (5.44%) из порядка *Thermales*, представители которого являются термофильными хемоорганогетеротрофами. Основным фотосинтезирующим организмом в Третьяковском источнике являются цианобактерии, в отличии от озера Фауста. Такая малая

численность цианобактерий (7.04%) объясняется тем, что для образца были взяты донные отложения, а не фрагменты микробного мата. Тип *Chloroflexi* представлен классами *Anaerolineae* (36.92%), *Caldilineae* (25.37%), *Chloroflexia* (4.56%) и неклассифицированными организмами (32.25%), а тип *Bacteroidetes* – классами *Saprospiria* (18.89%), *Cytophagia* (4.58%), *Bacteroidia* (0.60%) и неклассифицированными организмами (75.93%).

Таблица 2. Информация об экстрагированных геномах (M - набор метагеномных данных: R3 – оз. Фауста, R4 – Третьяковский источник. ANI-value – геномная идентичность)

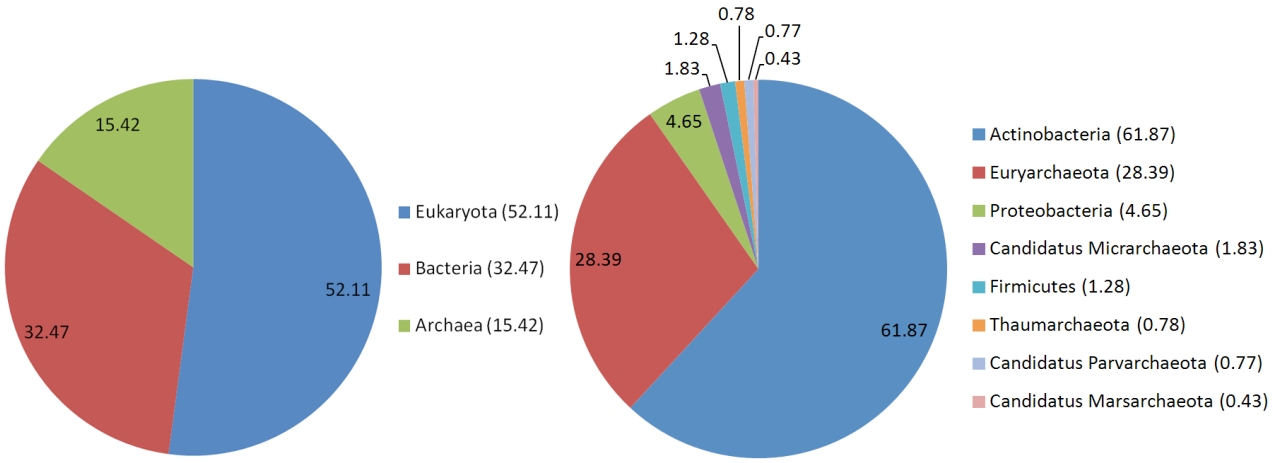
Информация о бине				Информация о геноме ближайшего организма			ANI-value
M	№	Размер, мб	GC%	Вид	Размер, мб	GC%	
R3	19	1.92880	38.5	*Cuniculiplasma divulgatum*	1.90881	37.3	74.00%
	23	1.73336	46.9	*Thermoplasma acidophilum* DSM 1728	1.56491	46.0	79.68%
R4	13	3.60861	63.8	*Dechloromonas agitata* is5	3.62723	62.8	90.18%
	22	3.44674	34.2	*Ignavibacterium album* JCM 16511	3.65900	33.9	88.37%
	74	3.25294	49.4	*Treponema caldarium* DSM 7334	3.23934	45.6	79.55%
	126	4.21820	63.5	*Pannonibacter indicus*	4.17068	63.5	99.88%
	190	1.89239	57.7	*Thermanaerovibrio velox* DSM 12556	1.88084	58.8	79.64%
	197	3.99218	46.4	*Anaerosporomusa subterranea*	3.96819	47.1	77.15%
	201	2.77489	64.8	*Thermoanaerobaculum aquaticum*	2.66093	63.0	78.58%

Полученные данные о гомологии с известными организмами хорошо коррелирую с анализом по гомологии белковых последовательностей.

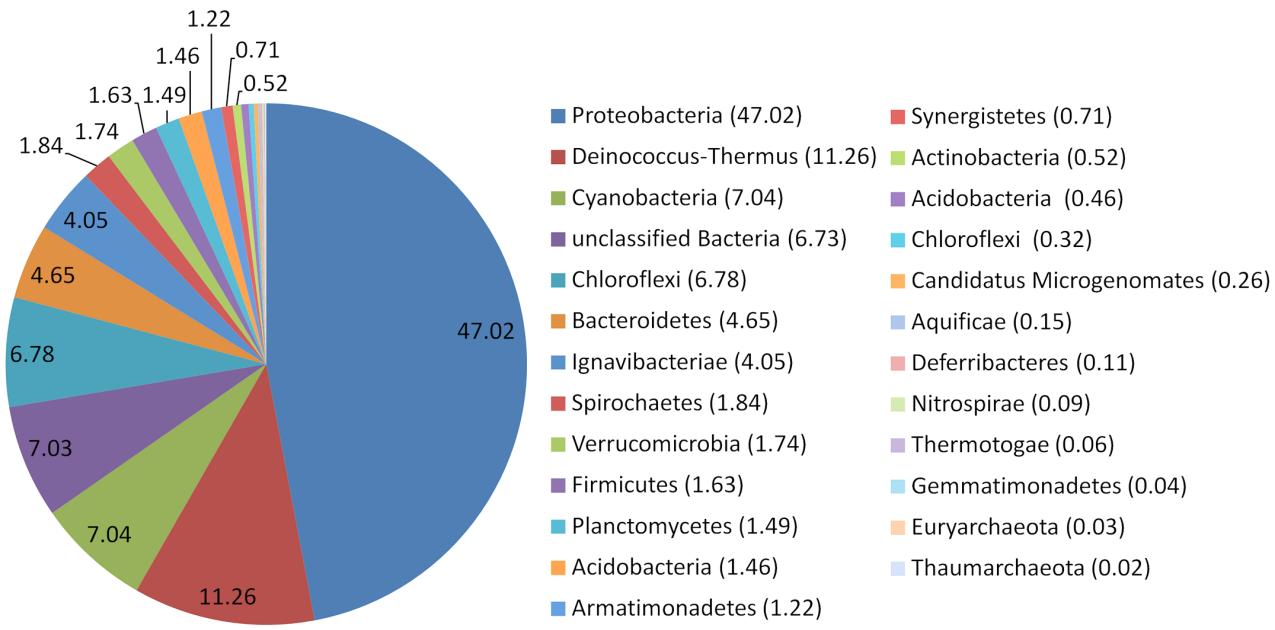
Как видно из данной таблицы, бин 126 из R4 достоверно является видом *Pannonibacter indicus*, тогда как остальные - потенциально новыми видами.

Данные бины были проанализированы при помощи веб-сервиса RAST, в результате чего была получена функциональная аннотация генов и общая информация о геномах.

5. Иллюстрации, визуализация результатов:



Таксономический состав микробного сообщества верхнего слоя донных отложений источника оз. Фауста по метагеномным данным R3 (%). Слева - на уровне доменов. Справа - на уровне типов, без учета представленности эукариот.



Таксономический состав на уровне типов в образце донных отложений Третьяковского источника (%).

6. Эффект от использования кластера в достижении целей работы:

Осуществление сборки метагеномов требует большого количества вычислительных мощностей. В подобных задачах использование многопроцессорных суперкомпьютеров является обязательным условием, без которого невозможно достижение целей работы.

7. Впечатления от работы с ИВЦ НГУ и пожелания:

Впечатления положительные, особенно отмечу отличную техподдержку пользователей. Пожелания: оптимизировать сайт ИВЦ НГУ, а именно улучшить навигацию и меню, а также расширить справочную информацию по работе с кластером (привести больше примеров и т.д.)