**Тема работы:** Метагеномный анализ микробных сообществ соленого озера №48 НСО.

**Состав коллектива:** Шипова Александра Андреевна, студент НГУ; Розанов Алексей Сергеевич, к.б.н., н.с., лаборатория Молекулярных биотехнологий ИЦиГ.

**Научное содержание работы:**

1. Задачи:

Выделение ДНК из образцов микробных сообществ и метагеномное секвенирование;

Сборка скэффолдов и биннинг;

Филогенетический анализ микробного состава сообществ.

2. Современное состояние проблемы.

Микроорганизмы, обитающие в соленых озерах, обладают высокой устойчивостью к осмотически активным веществам и часто к экстремально низким или высоким значениям pH. Поэтому экосистемы соленых озер интересны как для получения фундаментальных знаний о филогении микроорганизмов и эволюции микробных сообществ, так и для поиска белков биотехнологического назначения.

Микробные сообщества экстремальных соленых экосистем изучаются исследователями по всему миру. При помощи метагеномных методов описано много соленых озер. Но биоразнообразие соленых озер России, в том числе, соленых озер НСО, остается малоисследованным, что делает данную работу актуальной.

3. Подробное описание работы, включая используемые алгоритмы.

Было выполнено выделение ДНК из образцов верхнего слоя донных отложений и цианобактериальной биомассы; секвенирование.

Для проверки качества и последующей обработки ридов были использованы программы FastQC и Trimmomatic. Сборка ридов в скэффолды осуществлялась программой metaSPAdes 3.11.1. Характеристика собранных скэффолдов была выполнена с помощью инструмента metaQUAST. На Python был написан скрипт, при помощи которого были отброшены скэффолды длиной менее 1000 пар нуклеотидов. Для аннотации генов и определения филогенетического разнообразия скэффолды были загружены на онлайн ресурс MG-RAST.

Разделение на кластеры в зависимости от покрытия скэффолдов и частоты встречаемости тетрануклеотидов было выполнено с использованием программы MaxBin 2.2.4. Помимо файла с метагеномом на вход программе подавался файл с информацией о покрытии каждого скэффолда. Для сравнения параллельно была выполнена визуализация кластеризации скэффолдов при помощи программы VizBin.

Для филогенетического анализа кластеров на Python был написан скрипт, суть работы которого заключается в следующем. Из файла, содержащего белковые последовательности маркерных генов, найденных для определенного кластера программой MaxBin, брались первые 3 последовательности. Затем программа обращалась к белковой базе данных blastp, по которой происходил поиск гомологичных последовательностей. Данная процедура повторялась для белковых последовательностей каждого кластера для обоих образцов. Если для отдельного кластера гомологичные к каждому из трех белков последовательности принадлежали к одному бактериальному/археальному/эукариотическому типу с идентичностью >60% и e-value<0, то считалось, что кластер принадлежит этому типу.

Также из белковых последовательностей, найденных программой MaxBin, были выбраны два наиболее встречающихся среди кластеров белка отдельно для двух образцов. Для каждого из этих белков была сформирована выборка в формате fasta, содержащая номер кластера и белковую последовательность, найденную в нем. Для удобства выполнения этих действий были написаны скрипты на Python.

Поскольку наиболее интересными микроорганизмами для исследования в данной работе являются цианобактерии, они исследовались более подробно. После определения таксономии каждого кластера среди белковых последовательностей, относящихся к цианобактериям, был выбран наиболее встречающийся (среди двух образцов) белок. Была сформирована выборка, содержащая: последовательности этого белка из обоих образцов; наиболее к ним близкие последовательности из базы данных NCBI; последовательности этого белка, принадлежащие разным порядкам цианобактерий из базы данных NCBI; для построения филогенетического дерева в качестве аутгруппы была выбрана последовательность этого белка, относящаяся к фирмикутам.

Для выравнивания белковых последовательностей и построения филогенетических деревьев была использована программа MEGA 6.06. Сначала последовательности были выровнены при помощи метода MUSCLE. Затем для каждой выборки для построения деревьев методом максимального правдоподобия была найдена подходящая модель эволюции аминокислотных последовательностей с наименьшим значением информационного критерия BIC. Для проверки устойчивости дерева было построено 100 бутстрепов.

4. Полученные результаты.

Из отобранных образцов соленого озера было проведено выделение ДНК. Секвенирование было выполнено в ЦГРМ Генетико на приборе NovaSeq. В результате секвенирования был получен массив сырых парных ридов, длиной 100 пар оснований, в количестве примерно 400 миллионов шт на образец.

После анализа сборки метагеномов была получена информация, представленная в таблице 1. Для образца цианобактериальной биомассы было получено всего 1,5 млн скэффолдов; общая длина метагенома составила 994 млн нуклеотидов. Для образца донных отложений был получен 1,7 млн скэффолдов; общая длина метагенома составила 1 млрд нуклеотидов.

| | Образец цианобактериального мата | Образец донных отложений |
|---|---|---|
| Кол-во скэффолдов (>= 0 bp) | 1 567 052 | 1 706 464 |
| Кол-во скэффолдов (>= 1000 bp) | 156 467 | 201 488 |
| Самый длинный скэффолд | 387 984 | 476 707 |
| Общая длина (>= 0 bp) | 994 597 198 | 1 191 582 506 |
| Общая длина (>= 1000 bp) | 632 144 908 | 752 234 840 |
| GC (%) | 54.02 | 53.88 |

Табл. 1: информация о сборке.

При помощи онлайн сервера MG-RAST было получено филогенетическое разнообразие сообществ на разных таксономических уровнях. На рисунке 1 приведена представленность разных доменов. Наше внимание привлекло наличие большого числа последовательностей, отнесенных к эукариотам. Для образца цианобактериального сообщества это были насекомые, для образца верхнего слоя донных отложений – растения. Несмотря на то, что попадание насекомых и растений в пробы было возможным, такая значительная доля эукариот вызвала сомнения в правильности полученных данных.

Поэтому при помощи программы MaxBin была проведена кластеризация скэффолдов. По заявлению разработчиков кластер представляет собой геномную последовательность одного вида микроорганизма. Для образца цианобактериального мата было получено 158 кластеров, для

образца донных отложений – 187. Программа также находит гены, кодирующие 107 белков, которые встречаются у 95% бактерий. Таким образом, для каждого кластера были получены белковые последовательности. Каждый кластер был филогенетически охарактеризован по трем белкам при помощи поиска наиболее схожих последовательностей.

Результаты филогенетического анализа были обобщены и представлены в виде столбчатой диаграммы (рисунок 2). В обоих образцах преобладали Bacteroidetes, Cyanobacteria и Proteobacteria. Представленность цианобактерий оказалась высока в обоих образцах, что, скорее всего, связано с тем, что мы исследовали свежеосевший мат, в котором не в полной мере произошло изменение филогенетического состава. Содержание архей в образцах оказалось низким, что, скорее всего, связано с низкой соленостью озера в 2017 году. Археи были представлены единственным типом Euryarchaeota.

Для цианобактериальных кластеров было построено филогенетическое дерево. Наиболее многочисленными оказались цианобактерии порядка *Chroococcales*, близкий родственник из рода *Euhalothece*. Также к этому порядку были отнесены значительно реже встречающиеся в образце цианобактерии, филогенетически ближе всего располагающиеся к роду *Halothece*. Вторые по численности цианобактерии были отнесены к порядку *Oscillatoriales*, близко расположенные к *Phormidium*. Очень редко встречающиеся цианобактерии были отнесены к порядку *Synechococcales*, их белки гомологичны с *Leptolyngbya*. Большинство обнаруженных нами видов цианобактерий имеют близких родственников, но в то же время исследуемые последовательности достаточно удалены от ближайших известных гомологов, что позволяет предположить, что они являются новыми видами, кроме кластеров, принадлежащих к порядку *Oscillatoriales*.

Для каждого образца также была получена функциональная аннотация генов. В образце микробного сообщества донных отложений оказалась более высокой доля генов, ответственных за разрушение карбогидратов, что, возможно, вызвано более гетеротрофным типом метаболизма микроорганизмов; и доля генов, относящихся к метаболизму серы и азота, что, скорее всего, связано с переходом на анаэробный тип дыхания.

5. Иллюстрации, визуализация результатов.

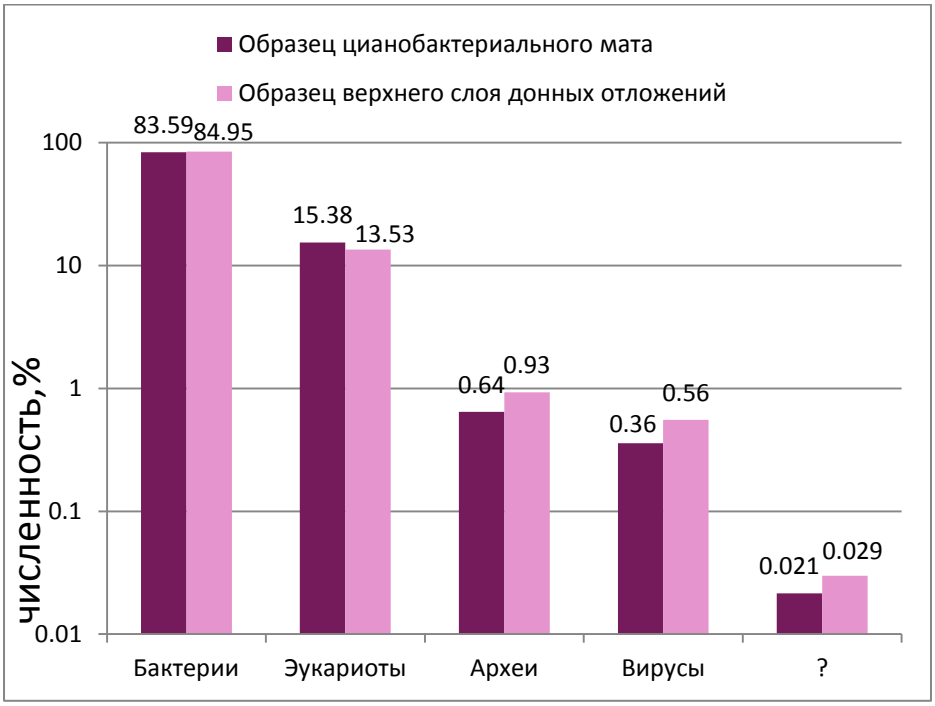


Рис 1. Филогенетическое разнообразие микробных сообществ на уровне домена, полученное при помощи сервера MG-RAST.

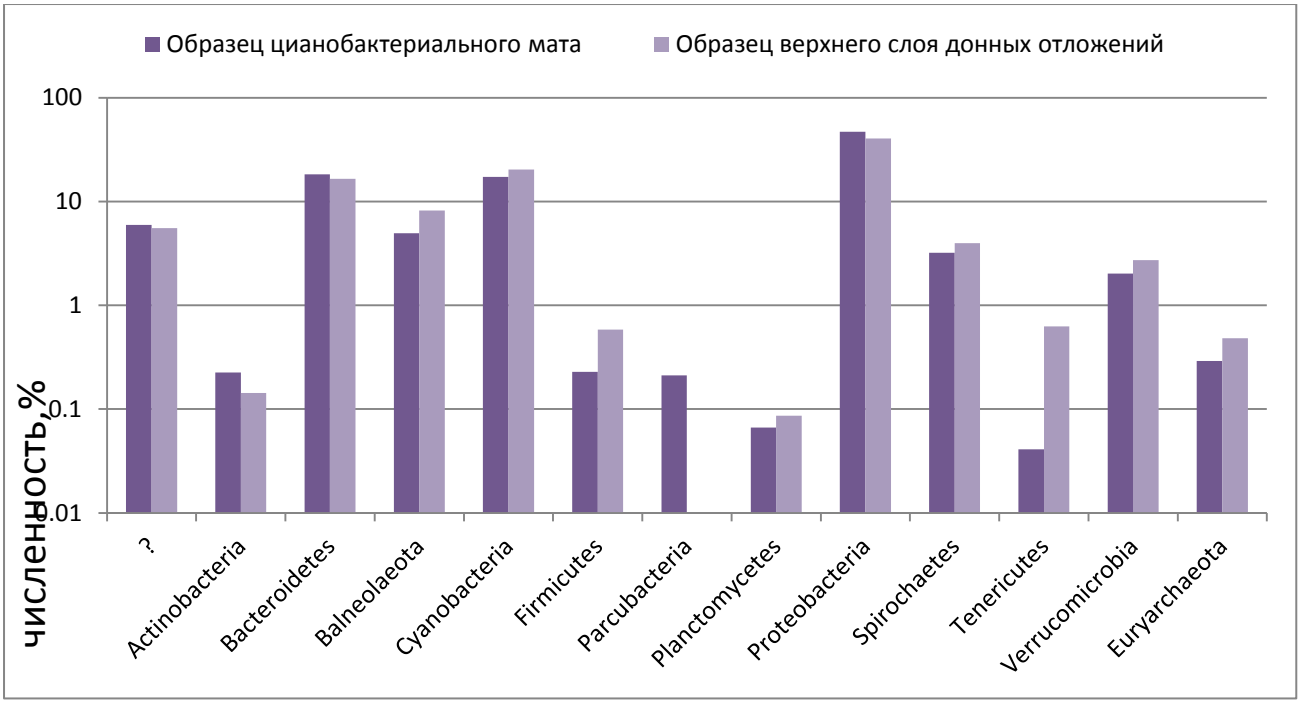


Рис 2. Филогенетическое разнообразие микробных сообществ на уровне типа, полученное при анализе данных кластеризации.

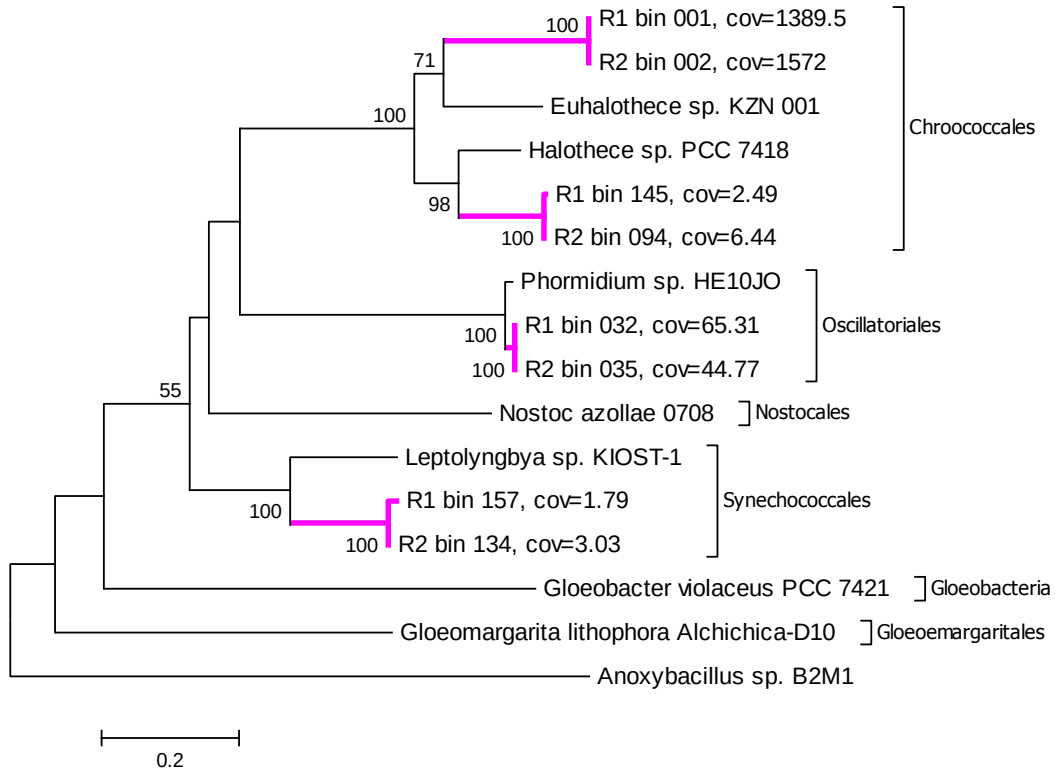


Рис 3. Филогенетическое дерево цианобактерий (по белку NAD-зависимая ДНК-лигаза). Кластеры, полученные при анализе образцов цианобактериального мата и верхнего слоя донных отложений, отмечены как «R1» и «R2» соответственно. Покрытие бина обозначено «cov». Отображены устойчивые узлы (>50).

**Эффект от использования кластера в достижении целей работы:**

Объем данных, полученных после секвенирования, был достаточно большим: по 40 гб на каждый образец. Обработка этих данных, в соответствии с поставленными задачами, не могла быть выполнена на ПК: не хватало памяти и мощности ПК. Поэтому часть биоинформатической работы выполнялась на кластере НГУ: предобработка ридов и сборка скэффолдов.