

**Тема работы:** Метагеномный анализ микробного мата и донных отложений озера Соленого НСО.

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

**Научное содержание работы:**

### 1. Задачи:

1. Провести секвенирование и сборку метагеномов микробных сообществ образцов разных слоев донных отложений и микробного мата озера.
2. Выбрать оптимальный метод таксономической классификации полученных метагеномных данных.
3. Выполнить сравнительный таксономический анализ исследуемых микробных сообществ.
4. Провести сравнительный анализ разнообразия метаболических путей микробных сообществ Соленого озера №48.

### 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 нескольких слоев донных отложений и микробного мата; секвенирование.

Для проверки качества и последующей обработки ридов были использованы программы FastQC, Trimmomatic и cutadapt. Риды были очищены от контаминации путем их картирования на геном человека при помощи bowtie2. Сборка ридов в контиги осуществлялась программой metaSPAdes 3.11.1. Характеристика собранных контигов была выполнена с помощью инструмента metaQUAST. Были отброшены контиги длиной менее 1000 пар нуклеотидов. Разделение на кластеры в зависимости от покрытия контигов и частоты встречаемости тетра-нуклеотидов было выполнено с использованием программы MaxBin 2.2.4.

Для таксономического анализа микробных сообществ отдельных образцов было использовано три независимых метода: 1) метод, реализованный в программе Kraken 2, классифицирующий последовательности (в данной работе – контиги) на основе сходства их k-мерного состава с k-мерным составом эталонной последовательности генома; 2) метод, реализованный в программе phyloFlash, основанный на выравнивании ридов против эталонной базы данных rPHK SSU; 3) метод, основанный на выравнивании контигов/белков каждого кластера против баз данных NCBI nt и NCBI nr.

Была проведена функциональная аннотация метагеномных данных (Prokka и kofam\_scan); получена информация о представленности генов ферментов в образцах. Также при помощи программы MinPath был выявлен минимальный набор метаболических путей, позволяющих охарактеризовать микробные сообщества образцов.

#### 4. Полученные результаты.

Из отобранных образцов соленого озера было проведено выделение ДНК. Парно-концевое секвенирование было выполнено на приборах NovaSeq и NextSeq.

В таблице 1 представлена общая информация о результатах секвенирования, сборки и кластеризации контигов.

	Количество ридов	Длина ридов	Секвенатор	Общее кол-во контигов	Кол-во контигов длиной >1000	Самый длинный контиг	Длина метагенома	N50	Кол-во кластеров
<b>MM</b>	398.852.702	35-100	NovaSeq	1567052	156467	387984	632144908	7195	205
<b>1</b>	398.845.300	35-100	NovaSeq	1706464	201488	476707	752234840	6114	261
<b>2</b>	139.130.832	35-150	NextSeq	8111942	243069	148335	565334284	2525	242
<b>4</b>	23.204.124	35-150	NextSeq	4220403	63969	125468	113099631	1676	83
<b>5</b>	12.648.886	35-150	NextSeq	2027426	36714	70530	62704972	1629	37

Табл. 1: информация о результатах секвенирования, сборки и кластеризации.

Из трех принципиально различных методов таксономического анализа метагеномов наиболее оптимальным оказался метод классификации исходных данных секвенирования (т.е. прочтений) по базе данных 16S рРНК, при котором количество неклассифицированных последовательностей не превышало 6%. По результатам данного анализа таксономического состава (рис. 1) выделяются 2 принципиально различные группы: (1) сообщества микробного мата и двух первых слоев донных осадков и (2) сообщества 4-го и 5-го слоя донных отложений. В первой группе преобладают бактерии типа *Cyanobacteria* и *Proteobacteria*, класса *Gammaproteobacteria*. Во второй группе доминируют бактерии типа *Desulfobacterota* и *Gemmatimonadota*, отсутствуют *Cyanobacteria*, но присутствуют археи.

Менее продуктивным был метод, нацеленный на таксономическую классификацию кластеров контигов по аминокислотным и нуклеотидным последовательностям в их составе (рис. 2), т.к. среди кластеров присутствовало большое количество неклассифицированных. Метод классификации, основанный на сравнении k-мерного состава отдельных контигов с референсной базой данных, показал наименьшую эффективность, поскольку не определил таксономическую принадлежность для более 60% последовательностей в каждом метагеноме.

Анализ выявленного набора метаболических путей в микробных сообществах показал наличие путей фиксации углерода, метаногенеза, сульфатного и нитратного путей дыхания. По тепловой карте на рисунке 3А видно, что представленность генов ферментов в различных образцах разная. При анализе метаболических путей, найденных сразу во всех пяти метагеномах, было обнаружено, что в их составе наиболее различны по представленности (рис. 3Б) гены, ассоциированные со жгутиковой сборкой (mar02040), гены, ответственные за фотосинтез (mar00195), и гены, связанные с образованием биопленки (mar02025). Наиболее вероятно, что такие различия связаны, в первую очередь, с различиями между доминирующими таксономическими группами в метагеномах.

## 5. Иллюстрации, визуализация результатов.

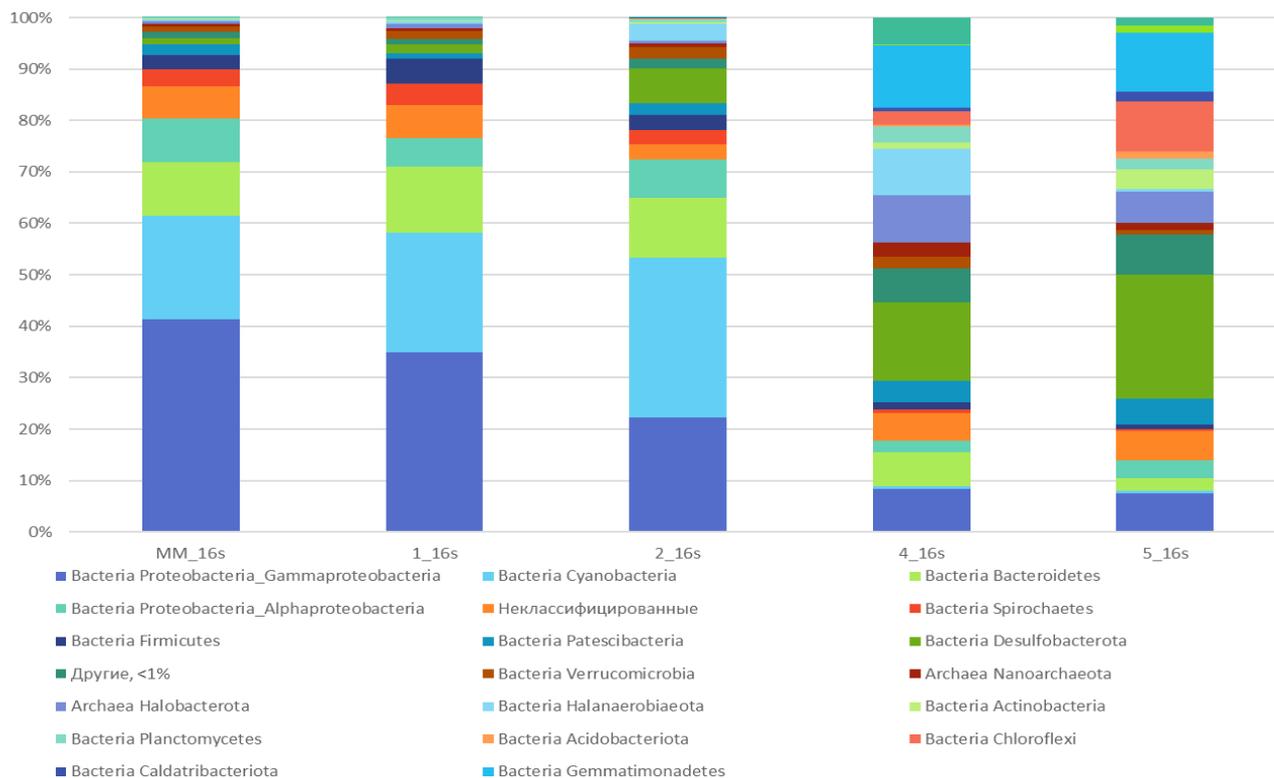


Рис. 1. Нормированная гистограмма с накоплением, отражающая таксономическую представленность бактерий и архей в образцах озера, полученную при помощи phyloFlash. MM – образец микробного мата; 1, 2, 4, 5 – слои донных отложений. Малочисленные микроорганизмы (имеющие долю менее 1% в образце) обозначены как «Другие, <1%».

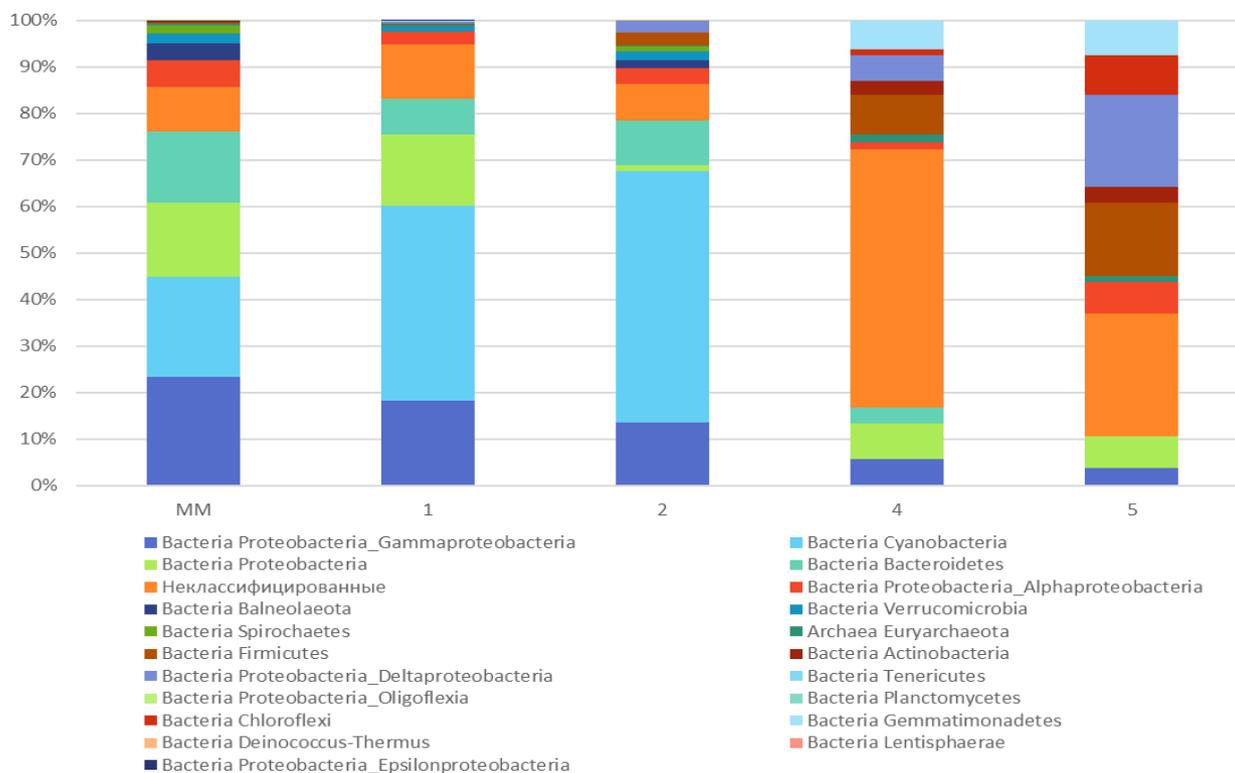
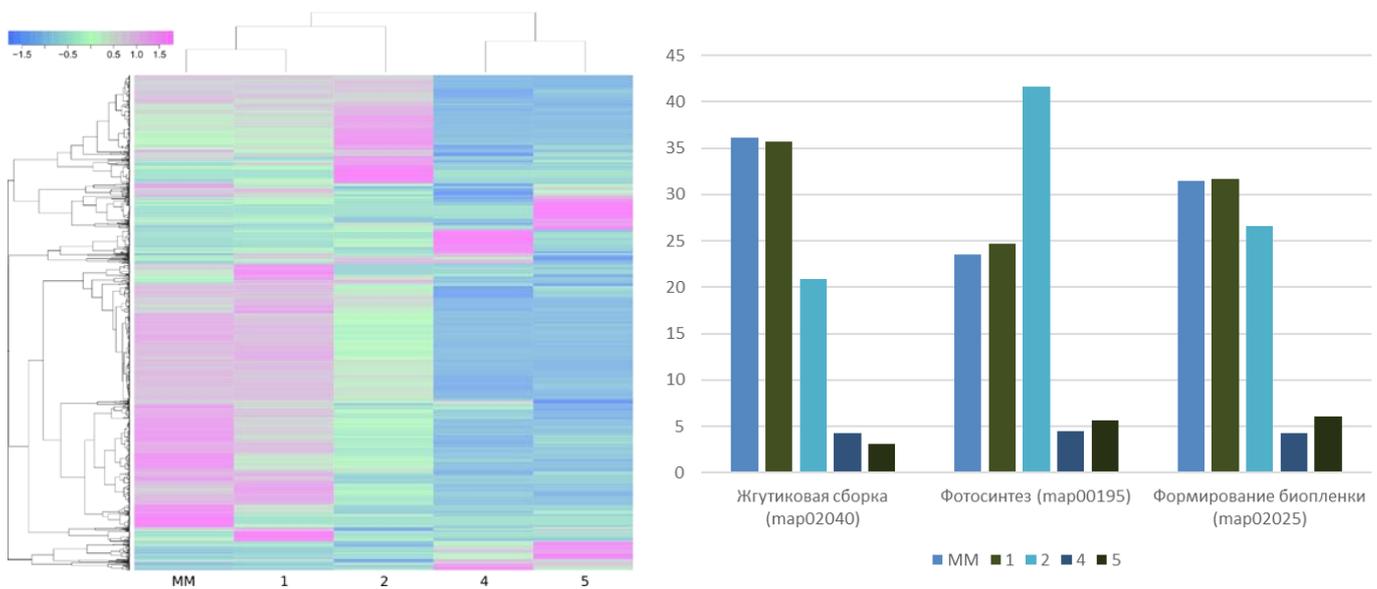


Рис. 2. Нормированная гистограмма с накоплением, отражающая таксономическую представленность микроорганизмов в образцах озера, полученную при помощи анализа результатов кластеризации контигов. MM – образец микробного мата; 1, 2, 4, 5 – слои донных отложений.



А

Б

Рис. 3А: Тепловая карта представленности генов метаболических путей (мар01100). Строки – гены (КО) определенного метаболического пути, столбцы – образцы. Цветовая шкала отражает представленность фермента среди образцов: высокопредставленный – розовый, низкопредставленный – синий; зеленый – промежуточное значение. MM – образец микробного мата; 1, 2, 4, 5 – слои донных отложений.

Рис. 3Б: Диаграмма распределения генов конкретного пути среди метагеномов. MM – образец микробного мата; 1, 2, 4, 5 – слои донных отложений.

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

Объем данных, полученных после секвенирования пяти образцов, был достаточно большим: суммарно около 130 гб. 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, сборка контигов, создание базы данных для программы Kraken2, анализ ридов при помощи phyloFlash.