МЕТАГЕНОМНЫЙ АНАЛИЗ И ЕГО ПРИМЕНЕНИЕ ДЛЯ СРАВНЕНИЯ СООБЩЕСТВ В ВОДНЫХ ЭКОСИСТЕМАХ

Букин Ю.С. [1,2]

[1] ФГБУН Лимнологический институт СО РАН 664033 г. Иркутск, ул, Улан-Баторская, д. 3. E-mail: bukinyura@mail.ru.

[2] Иркутский национальный исследовательский технический университет 664074, г. Иркутск, ул. Лермонтова 83. E-mail: bukinyura@mail.ru.

Аннотация

Метагеномный анализ одно из наиболее быстро развивающихся направлений современной молекулярной генетики и биоинформатики, направленное на изучения сообществ организмов с применением технологии высокопроизводительного секвенирования нуклеотидных последовательностей (NGS). Метагеномом называют смесь геномов нескольких организмов, извлеченных из некоторого образца биологического материала. Обычно метагеномные исследования проводиться для изучения состава сообществ микроорганизмов и сравнения таких сообществ друг с другом. Методы метагеномнгого анализа подразделяются на две большие групп. Первая группа методов основана на анализе апликона метагенома. Для амплификации и секвенирования выбирается вариабельный участок генома, присутствующий у вех организмов рассматриваемой группы. Втора ягруппа методов основана на расшифровки и анализе случайных фрагментов метагенома, так называемый метагеномный анализ методом дробовика. Даная группа методов, кроме установления таксономичнеского разнообразия сообщества, позволяет установит функциональную активность генов в составе организмов сообщества.

Annotation

Metagenomic analysis of one of the most rapidly developing areas of modern molecular genetics and bioinformatics. Metagenomic study the communities of organisms using the technology of high-throughput sequencing of DNA (Next-Gen Sequencing NGS). Metagenome is a mixture of the genomes of several organisms extracted from a sample of biological material. Usually metagenomic conducted to study the taxonomic composition of microbial communities and comparing these communities to each other. Methods of metagenomic analysis are divided into two large groups. The first group of methods is based on the analysis of amplicon of metagenome. For the amplification and sequencing of the selected variable region of the genome of the most organisms of this group. Second groups of method is based on the sequencing and analysis of random fragments metagenome, the so-called metagenomic analysis by a shotgun. This group of methods, in addition to establishing taxonomic diversity of community, allows you to set the functional activity of genes in the composition of the community of organisms.

Ключевые слова: метагеном, высокопроизводительное секвенирование (NGS), сообщества, биологические базы данных, сборка геномов, контиг, аннотация геномов.

Keywords: metagenom, Next-Gen Sequencing (NGS), communities, biological databases, genome assembly, contig, genome annotation.

Ведение

Метагеномный анализ – раздел молекулярной генетики, в котором изучаются образцы ДНК и РНК выделенные из среды, содержащей биологические продукты смеси организмов (Riesenfeld C. S., et al 2004; Edwards R. A., Rohwer F., 2005). Метагеномный анализ нашел

широкое применение при изучении бактериальных и вирусных сообществ. Преимуществом данного вида исследований является возможность определить в окружающей среде не культивируемые традиционными способами виды и штаммы микробов и вирусов (Tringe S. G., Rubin E. M., 2005).

Толчком к быстрому развитию метагеномного анализа послужили два фактора. Первый фактор это развитие методов высокопроизводительного секвенирования (NGS) и второй фактор это быстрое пополнение баз данных информацией о расшифрованных последовательностях ДНК и геномов организмов разных видов (Ребриков Д.В., и др. 20014). Сочетания этих факторов позволяют получать одновременно большие объемы данных (расшифрованных нуклеотидных последовательностей) и идентифицировать видовой состав и функциональную активность генов в смеси метагеномного материала.

Традиционно метагеномные исследования развиваются по двум направлениям:

1) таксономический анализ сообщества на основе высокопроизводительного секвенирования ампликона метагенома (Petrosino J. F. et al. 2009). Для амплификации выбираются определенные маркеры (последовательности ДНК определенных генов организмов) которые традиционно используются для проведения молекулярно-филогенетических исследования и идентификации видов организмов. Для идентификации видового состава сообществ микроорганизмов традиционно используется секвенирования ампликона гена, кодирующего 16S рибосомальную РНК. Базы данных расшифрованных нуклеотидных последовательностей, например, такие как Genbank (http://www.ncbi.nlm.nih.gov/), содержат информации приблизительно о 10000 16s рибосомальных РНК идентифицированных видов микроорганизмов. Данное обстоятельство позволяет путем сравнения информации, полученной при сиквенировании амипликона 16s рибосомальных РНК с информацией из баз данных идентифицировать часть видового состава микроорганизмов образца биологического материала. В перспективе возможно использование данного метода для идентификации видов в сообществе одноклеточных и многоклеточных эукориотических организмов. Особо полезен метагеномный анализ будет для исследования видового состава таких групп как одноклеточные зеленые водоросли, простеющие животные и грибы, имеющие в своем составе трудно культивируемые в лабораторных условиях виды. Для подобных целей может быть применен такой универсальный филогенетический маркер как ген, кодирующий 18s рибосомальную РНК эукариот. Базы данных содержат огромные массивы информации о расшифрованных последовательностях гена, кодирующего 18s рибосомальную РНК видов, идентифицированных по морфологическим и другим признакам. Однако скорость накопления замен в 18s рибосомальной РНК достаточно мала, что во многих случаях снижает разрешающую способность этого маркера. Зачастую при использовании 18s рибосомальной РНК можно различить только разные рода или даже семейства организмов. На внутривидовом уровне перспективным является использование митохондриального маркера – гена, кодирующего первую субъединицу белка цитохром с оксидазы (COI) для животных, в том числе и одноклеточных. Для растений в перспективе виды можно будет различить, используя хлоропластный генетический маркер, кодирующий болящую субъединицу белка рибулозобисфосфаткарбоксилаза (rbcl). Количество расшифрованных последовательностей маркеров COI и rbcl от идентифицированных видов организмов постоянно увеличивается. Данные маркеры становиться стандартом для баркода, используемого для идентификации видов.

2) таксономический и функциональный анализ сообщества на основе информации о расшифрованных случайных последовательностях из метагенома (Eisen J. A. et al. 2007). По другому подобное высокопроизводительное сиквенирование называется исследование метагенома методом дробовика. Из образца биологически материала выделяется общая ДНК,

которая затем фрагментируется и используется для создания библиотеки случайных фрагментов для NGS. Для обработки метагеномов, полученных методом дробовика создан ряд специальных баз данных, расшифрованных фрагментов ДНК и полных геномов прокориотических организмов позволяющих идентифицировать видовой состав анализируемого образца и функциональную активность сообщества микроорганизмов. Функциональная активность сообщества определяется генами, кодирующими белки, участвующие в различных метаболических циклах в нутрии организмов и между организмами. Анализируя ферментативный состав метагенома, можно например выяснить, какие реакции являются источником первичного органического вещества в сообществах хемотрофнх микроорганизмов.

Метагенгомный анализ используется и для идентификации состава вирусного сообщества в биологическом образце (Edwards R. A., Rohwer F., 2005; Mardis E. R.б 2008; Thurber R. V. et al. 2009). Так как генетическое разнообразие вирусов чрезвычайно велико, то не существует определенного маркерного гена, с помощью которого удалось бы идентифицировать видовой состав вирусного сообщества. Поэтому для исследования вирусного состава биологического образца используется метагеномный анализ методом дробовика. Расшифрованные случайные последовательности ДНК и РНК вирусов сравниваются с базой данных полных вирусных геномов, которая на данный момент содержит информацию о 5000 видов вирусов.

Кроме получения первичных данных о расшифрованных паследовательностях ДНК метагенома биологического образца трудоемкой задачей является биоинформационная обработка результатов секвенирования. Для извлечения информации из массивов расшифрованных нуклеотидных последовательностей необходимо использовать широкий спектр программного обеспечения (Ребриков Д.В., и др. 20014; Mendoza M. L. Z., et al 2015). Нескорые этапы анализа требуют большого объема вычислений с привлечением суперкомпьторов и параллельных вычислительных технологий. Кроме того высокая квалификация необходима для специалиста биоинформатика, обрабатывающего данные метагеномного секвенирования. Кроме общебиологических знаний подобный специалист должен разбираться в методах математической статистики, обладать навыками программирования на скриптовых языках R, Phyton, или Perl.

Широкое применение методов метагеномных исследований позволило получить новые данные о функционировании сообществ микроорганизмов водных экосистемах (Парфенова В. В., и др. 2013; Гладких А.С., и др. 1014; Colatriano D. et al. 2015; Ininbergs K. et al. 2015). Метагеномный анализ стал одним из инструментов мониторинга состояния водных и других экосистем. Проводя периодические исследования состава микробного сообщества в водоеме можно выяснить, идут ли в данном водоеме процессы, направленные на изменение текущего состояния экосистемы, происходит ухудшение или улучшение качества воды в водоеме.

Одним из перспективных направления метагеномных исследования является анализ бактериального и вирусного сообщества человека (Марданов А. В. и др. 2013; Kim Y., et al 2015; Tilg H., Adolph T. E., 2105; Santiago-Rodriguez T. M. et al. 2015; Jagathrakshakan S. N. et al. 2015). В настоящее момент проводятся масштабные исследования в рамках проекта метагеном человека. Исследования в этой области помогут выявить механизмы формирования многих заболеваний, связанных с изменение микрофлоры в организме человека.

Получение метагеномных данных

Как уже описывалось в ведении существует два направления метагеномного анализа: первое направление это исследование таксономического состава сообщества с применение

секвенирования ампликона (ПЦР продукта) полученного для стандартного генетического маркера (например 16s рибосомальная РНК прокариот или 18s рибосомальная РНК эукариот) (Petrosino J. F. et al. 2009) и второе, это анализ таксоносимического состава и функциональной анктивности сообщества с применением секвенирования случайных фрагментов нуклеиновых кислот метагеномного образца (метод дробовика) (Eisen J. A. et al. 2007). Обе технологии в плане реализации отличаются друг от друга. Рассмотрим этапы реализации их отдельно.

а) Сиквенирование определенных стандартных генетических маркеров (сиквенирование ампликонов)

Определение видового состава сообщества организмов необходимо начать с выбора молекулярного маркера для идентификации таксономической идентификации организмов. При исследовании бактериальных сообществ, стандартом стало использование фрагмента гена кодирующего 16s рибосомальную РНК. В большинстве случаях этот маркер способен идентифицировать в метагеноме не только прокориотичнескую ДНК но 16s рибосомальную РНК, входящую в состав митохондрий и хлоропластов эукариот. Схема получения первичных метагеномных данных в виде расшифрованных последовательностей ДНК представлена на рисунке 1.

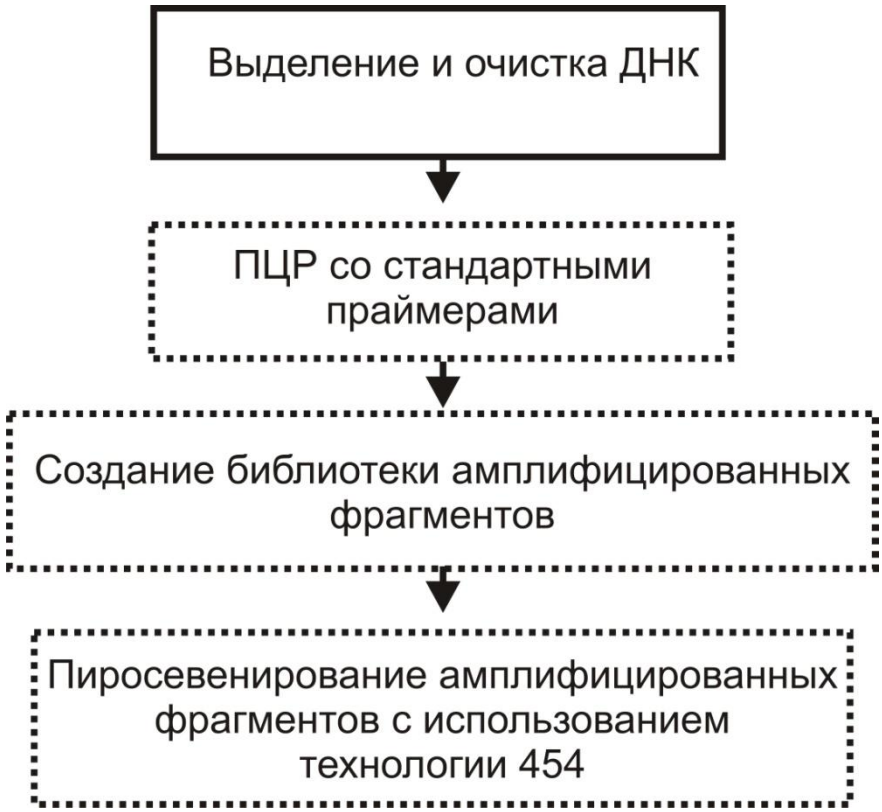

Рис. 1. Этапы получения первичных расшифрованных последовательностей ДНК при метагеномном анализе ампликонов стандартных генетических маркеров. Этапы работы, выполняемые по стандартным протоколам, выделены пунктирными квадратами.

Первой и самой сложной задач получения первичных метагенгомных данных является выделение и очистка ДНК из биологического образца. Проблема выделения ДНК заключается в том, что образец может быть сильно гетерогенных (неоднородным), например почва, придонный или поверхностный слой воды, мазки со слизистых оболочек человека. Для каждого типа образцов необходимо выдрать такую методику выделения ДНК, чтобы не происходило искажение образцов по соотношению видового состава микроорганизмов [].

Сравнительный анализ метогеномов возможен только при использовании одной технологии выделения и очистки ДНК. Иногда перед выделением ДНК производят фильтрацию водной взвеси образца, для того, чтобы изолировать микроорганизмы определенного размера, обогатить образец какими либо видами, интересными для анализа.

ПЦР реакция амплификации выборного для анализа метагенома фрагмента РНК производиться по стандартной методике. Если используется маркер 16s рибосомальная РНК для анализа сообщества микроорганизмов, то используется определенный достаточно консервативный участок этого маркера (Chun J. et al. 2010).

Первоначально, при метагеномном анализе бактериальных сообществ, применялась технология секвенирования по Сенгеру. При реализации этой методики ПЦР продукт клонировался в плазмидные вектора (Escherichia coli). После чего каждый клон секвенировался с подошью метода Сенгера (Rogers Y. H., Venter J. C. 2005). Использование этого метода позволяло получить последовательности высокого качества с достаточной длинной. Однако метод этот являеться достаточно трудоемким и не позволяет получить достаточное количество последовательностей для выделения редких видов и определение долевых соотношений микроорганизмов различных таксономических групп в образце.

Применение технология NGS позволило расшифровать достаточно большие выборки расшифрованных последовательностей ДНК для получения статистически сходящегося по соотношениям видов и других таксономических групп результатов. Для расшифровки достаточно длинных фрагментов применяется технология пиросеквенирования «454 Life Sciences» (Shendure J. et al. 2004), которая позволяет получить последовательности длинной до 400 пар нуклеотидов. В некоторых случаях применяется технология сиквенирования на платформах Illumina (Quail M. A. et al. 2008) и Ion Torrent (Quail M. A. et al. 2012), при этом прочтения производиться с обоих концов последовательности из ампликона.

б) Секвенирование метагенома методом дробовика

Технология получения случайных фрагментов метагенома образца отличается от технологии секвенирования ампликона. Этапы получения метагеномных данных методом дробовика представлены на рисунке 2.

Выделение и очистка ДНК производиться по той же технологии и встречает той же затруднения что и выделения ДНЛ для секвенирования ампликона. В некоторых случаях. Для анализа состава РНК содержаших вирусов в образце или исследования транскриптома образца биологического материал необходимо выделить РНК. В некоторых случаях после выделения ДНК или РНК из образца производят концентрирование нуклеиновых кислот с применением специального оборудования. Чаше всего концентрирование применеться при исследовании метагенома вирусных сообществ в виду низкой концентрации ДНК и РНК в исходном образце.

Если из образца выделялась РНК, то для дальнейших манипуляций ее переводят в ДНК реакцией обратной транскрипции.

Увеличение концентрации ДНК в образце может быть достигнуто применением технологии полногеномной амплификации ДНК (WGA – whole genome amplification) (Lasken R. S., Egholm M. 2003; Huang L. et al. 2015). Следует заметить, что полногеномная амплификация ДНК может вносить искажение в долевой состав видов в исследуемом сообществе организмов. Особо это заметно при амплификации ДНК вирусных сообществ. Геномы вирусов состоящих из одной цепочки ДНК амплифицируються эффективней геномов вирусов. состоящих из двухцепочесной ДНК. Если для увеличения концентрации ДНК примялось полногеномная амплификация, то вносимые ей искажения необходимо учитывать при обработкt результатов методами биоинформатики.

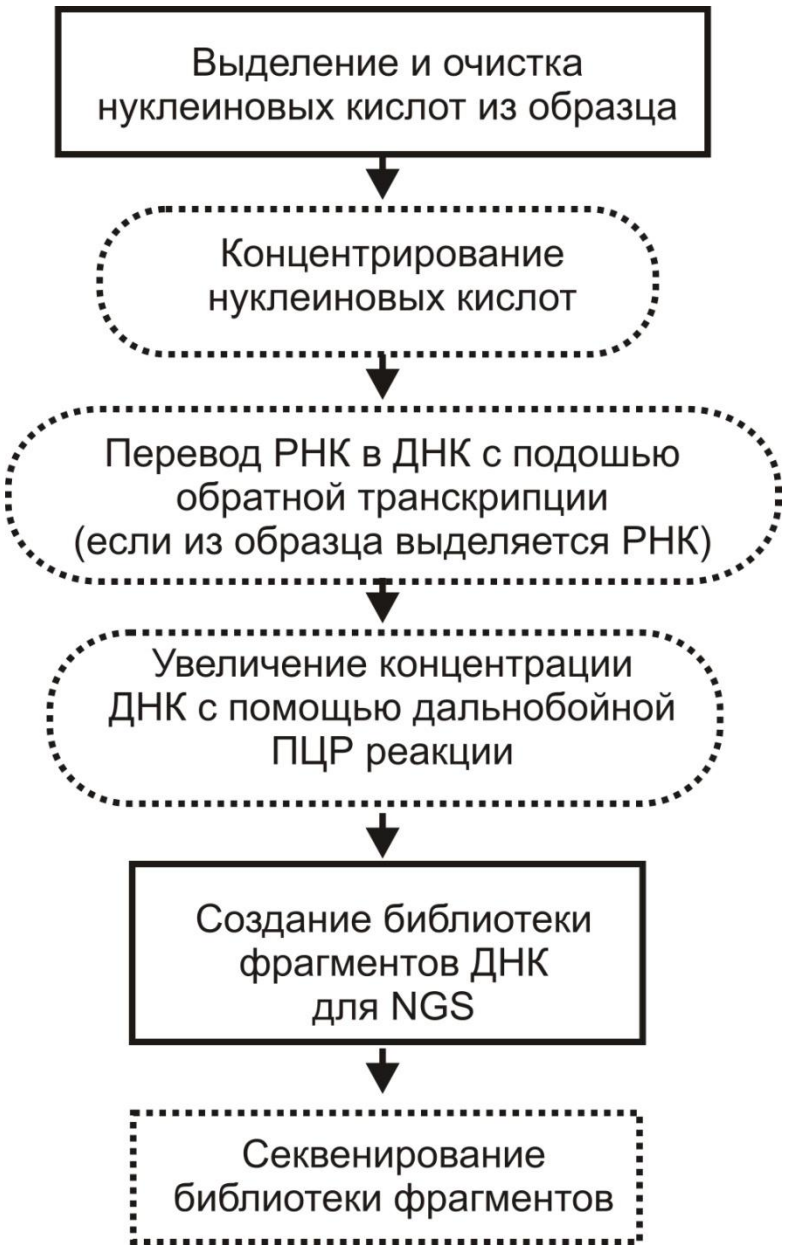

Рис. 2. Этапы получения первичных расшифрованных последовательностей ДНК при метагеномном анализе случайных фрагментов нуклеиновых кислот (секвенирование метагенома методом дробовика). Этапы работы, выполняемые по стандартным протоколам, выделены пунктирными квадратами. Не обязательно присутствующие этапы работы, выполняемые по стандартным протоколам, обведены пунктирными овалами.

Перед приготовлением библиотеки метагеномную ДНК необходимо раздробить на короткие фрагменты с помощью физических либо ферментативных способов. Из полученного набора коротких фрагментов ДНК отбирают фракцию заданной длинны для приготовления библиотеки фрагментов для NGS.

Существуют две технологии создание фрагментов ДНК библиотеки для NGS. Обычные библиотеки и инвертирование библиотеки. Обычные библиотеки позволяю проводить чтение фрагмента ДНК с одной стороны или с обеих сторон. При прочтении последовательности библиотеки с обеих сторон в случае если прочтения не пересеклись, то остаток не прочтенной длины не известен. При использовании инвертированной библиотеки в случае прочтения фрагмента библиотеки с обеих сторон исследователю известна длинна фрагмента, заключенного между прочитанными концами. Эту информацию используют

биоинформационные программы, собирающие прочитанные фрагменты в контиги. В любом случае прочтение последовательности из библиотеки дает исследователю информацию о том, что данная пара прочтений принадлежит одному организму. Подробнее с информацией о типах библиотек фрагментов и особенностях их подготовки описана в книге (Ребриков Д.В. и др. 2014).

Предварительная обработка и оценка качества прочтения метагеномных данных

Большинство исследователей заказывают получение первичных расшифрованных последовательностей ДНК метагенома в центрах коллективного пользования, располагающих дорогостоящими приборами для высокопроизводительного секвенирования (NGS). Такие центры, как правило, осуществляют создание библиотеки фрагментов ампликона или случайных фрагментов метагенома и секвенирования этих библиотек высокопроизводительными методами. На выходе, исследователь, заказавший метагеномное секвенирования, получает в электронном виде набор расшифрованных последовательностей ДНК. Расшифрованные последовательности представляют собой файлы, хранящие информацию о последовательностях букв в расшифрованных фрагментах и о качестве или мере доверия к нахождению именно этой буквы в данной позиции кодона. Наиболее распространенный формат файлов для хранения информации о последовательностях и о качестве расшифрованных букв является формат fastq (Cock P. J. A. et al. 2010). Большинство приборов NGS, в месте, с их программным обеспечением, могут выдавать первичные данные в этом формате. Ниже приведен пример записи информации об одной последовательности в формате fastq:

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC6

в этой записи знак @ является идентификатором строки, на которой расположено имя паследовательности, следующей строкой записан нуклеотидный состав последовательности, знак + разделяет строки с нуклеотидным составом последовательности и оцененным качеством букв в последовательности. Строка с информацией о качестве прочтения букв в последовательности содержит такое же количество символов, что и нуклеотидная последовательность. Каждая буква строки качества кодирует качество прочтения соответствующей буквы в последовательности. Символы строки качества кодируют цифры от 0 до 40. Цифра 0 соответствует наихудшему качеству прочтения, цифра 40 наилучшему качеству прочтения. Пороговом значением качества считается 20, при таком качестве прочтении соответствующую букву можно использовать в SNP анализе. Сушествует два типа кодировки качества Sanger/Illumina 1.9 и Illumina 1.5, эту информацию необходимо узнать для использования ее при редактировании последовательностей на качество прочтения.

Анализ качества прочтения, большого массива данных последовательностей формата fastq можно осуществить с помощью программы FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Данная программа выдает распределение качества букв в последовательностях, в различных позициях их длинны. График подобного распределения можно увидеть на рисунке 3. Кроме информации о распределении качества прочтения в различных позициях по длине последовательностей программа FastQC рисует график распределения длин прочитанных фрагметов в наборе данных, и сообщанет информацию о том, имеются ли в составе последовательностей в начале в конце или в средней части адаптеры – короткие последовательности пришитые к основным фрагментам метагенома при подготовки библиотеки для прочтения ДНК.

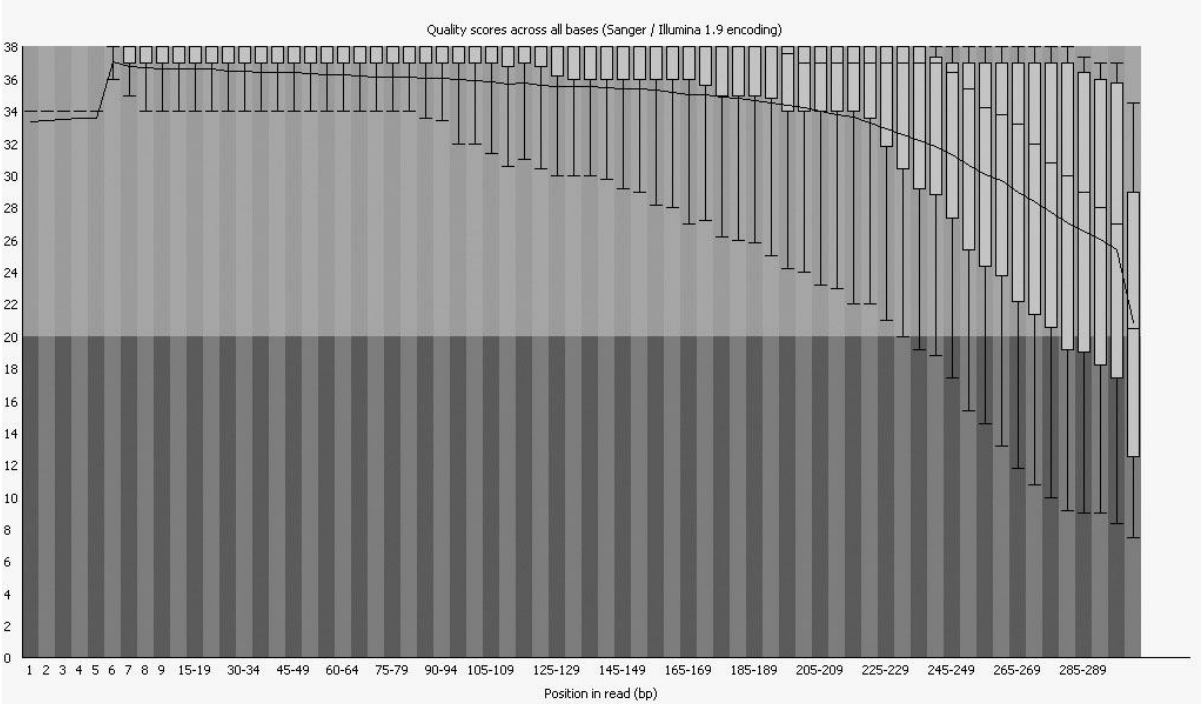

Рис. 3. График распределения качества прочтения последовательностей выданный программой FastQC на основе файла формата fastq, содержащего результаты секвенирования метагенома. Распределеня качества букв отображено в виде боксплотов. По оси икс отображен номер буквы в последовательности, по оси игрек качество прочтения.

Для примера проанализируем распределения качества прочтения на рисунке 3. Мы видим, что начина с позиции 229 в некотором количестве последовательностей качество почтения букв уменьшилось ниже критического предела в 20. Ближе к концу прочтения, начиная с позиции 269, количество таких последовательностей увеличивается до 50%. Общее качество прочтения уменьшаться ближе к концу прочтения.

Для метагеномного анализа на основе секвенирования ампликона качество прочтения является критическим параметром. Большинство методов биоинформационного анализа используют SNP. Поэтому качество прочтения каждой буквы в сравниваемых последовательностях должно быть больше 20 единиц. Для анализа метагенома с применением случайных последовательностей (метод дробовика) качество почтения букв является мене критических, особенно при установлении функциональной активности сообщества. В случае проведения анализа метагенома методом дробовика качество прочтении, по крайней мере 95% букв, из всего массива даны должно быть больше 20 единиц.

Для редактирования качества прочтения ДНК путем исключения части последовательностей или последовательностей целиком используется программа Trimmomatic (Bolger A. M., et al 2014). Даная программа позволяет по заданному шаблону находить участки последовательностей с низким качеством почтения и удалять последовательности или части последовательностей. Когда речь идет об удалении части последовательностей, то иметься в виду начало или конец последовательности. Именно в этих участках чаще всего содержаться буквы с плохим качеством прочтения. В принципе используя программу Trimmomatic можно исключить все начальные и конечные участки последовательностей или последовательности, которые в срединной части, содержат буквы с плохим качеством. Редактируя последовательности с помощью программы Trimmomatic нужно соблюдать определенный компромисс, чтобы не удалить из набора чрезмерно

большое количество последовательностей и критически не уменьшить длину прочтения, что приведет к падению статистической сходимости результатов метагеномного анализа.

На рисунке 4 приведен пример обработки программой Trimmomatic метагенома, анализируемого методом дробовика. Рисунок 4 представляет собой распределения качества нуклеотидов в последовательностях прочтения, после редактирования образца, с распределением, изображенном на рисунке 3. В результате редактирования среднее 97.5% нуклеотидов в совокупности последовательностей имеют качество больше 20 единиц. До редактирования эта цифра составляла 86%. Количество последовательностей в наборе уменьшилось на 30%, что является значительной величиной. Для того, чтобы осуществить подобное редактирования для работы программы Trimmomatic были заданы следующие опции: 1) удалять с конца последовательности хвост, протяженностью 10 нуклеотидов если он содержит хотя бы две буквы с качеством прочтения ниже 20 единиц, 2) сканировать последовательно всю последовательность рамкой в 20 нуклеотидов и удалять последовательность, которая содержит в этой рамке хотя бы два нуклеотида с качеством прочтения ниже 20 единиц 3) удалять все последовательности длиной меньше чем 150 нуклеотидов.

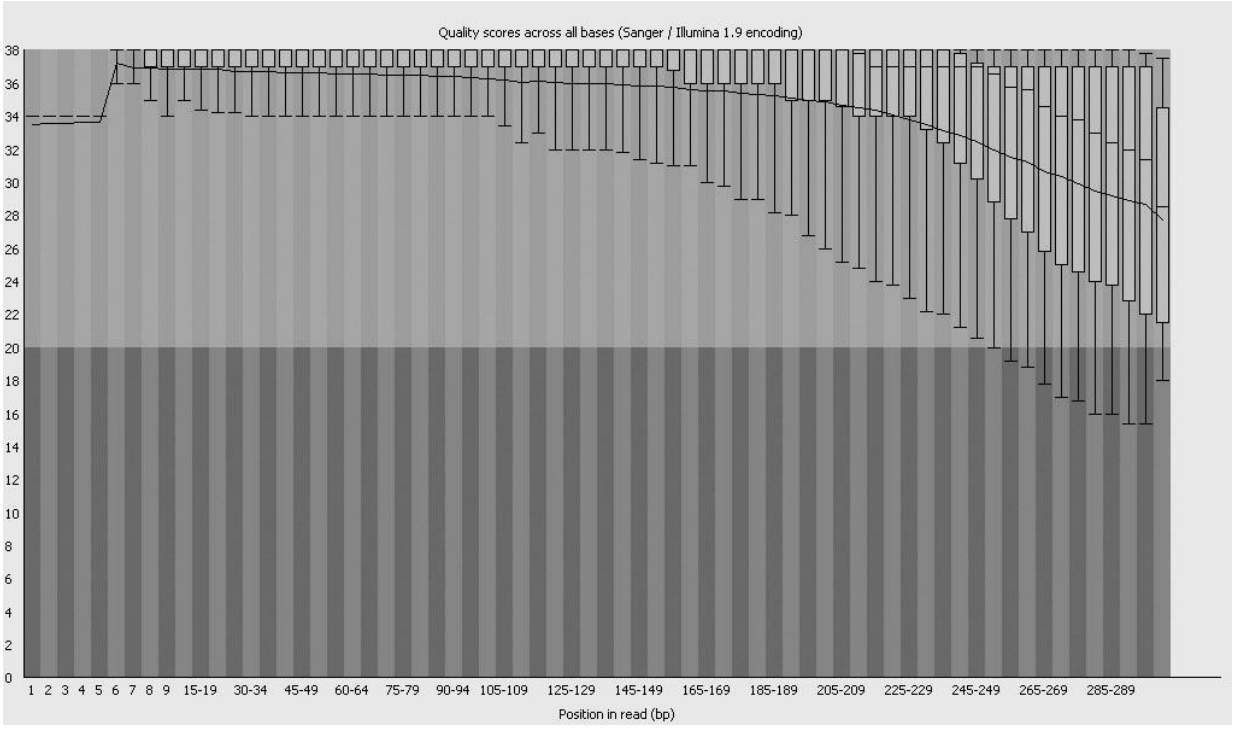

Рис. 4. График распределения качества прочтения последовательностей выданный программой FastQC на основе файла формата fastq, содержащего результаты секвенирования метагенома, после редактировании программой Trimmomatic. Распределеня качества букв отображено в виде боксплотов. По оси икс отображен номер буквы в последовательности, по оси игрек качество прочтения.

При планировании эксперимента по анализу метагенома необходимо учесть тот факт, что часть полученных данных должна быть отброшена из-за низкого качества прочтения. При метагеномном анализе ампликона потери при отбрасывании последовательностей недостаточной длинны и качества могут достигать 50%. Следовательно, необходимо секвенировать последовательности с некоторым запасом чтобы достигнуть сходящегося результата при анализе долей представленности организмов различных таксономических групп в метагеноме.

Обработка метагеномных данных, полученных на основе секвенирования стандартных генетических маркеров (сиквенирование ампликонов)

В настоящее время достаточно подробно отработана технология анализа сообществ микроорганизмов на основе ампликонов участков гена, кодирующего 16s рибосомальную РНК (Ребриков Д.В. и др. 2014) (идентифицируют состав прокариот в сообществе) и на основе участков генов 18s рибосомальной РНК (идентифицируют состав эукариот в сообществе). Для проведения такого анализа разработан широкий спектр коммерческого и не коммерческого программного обеспечения, позволяющего осуществить все этапы сравнительного анализа биологического разнообразия метагеномного образца. К подобному программному обеспечению относятся MOTHUR (Schloss P. D. et al. 2009), QIIME (Caporaso J. G. et al. 2010). Одной из самых популярных программ из перечисленного списка является MOTHUR. Данная программа является кросплатформенной (иметься версия для различных операционных систем (Windows, Linux, MAC OS) и обеспечивает встроенную поддержку многоядерных процессоров и многопроцессорных компьютеров, что существенно ускоряет вычислительный процесс. Иметься ряд онлайн сервисов для проведения метагеномного анализа ампликона на предмет установления биологического разнообразия. Самым популярным онлайн сервисом является MG-RAST (Meyer F. et al. 2008). При использовании онлайн сервисов пользователь загружает свои данные в виде расшифрованных последовательностей ДНК через веб интерфейс на сервер, задает опции для анализа и через некоторое время через веб интерфейс получает результат. Преимуществом данного подхода является то, что интернет сервисы для расчетов используют высокопроизводительные суперкомпьютеры, ускоряющие вычисления. Недостаток веб технологий проявляется при обращении большого количества пользователь за короткий промежуток времени к интернет ресурсу, что создает очереди на ожидания к началу расчетов.

Кроме программного обеспечения для идентификации последовательностей метагенома на основе участков гена, кодирующего 16S или 18S рибосомальную РНК, понадобиться референсная база данных последовательностей 16s или 18S рибосомальной РНК микроорганизмов с идентифицированным таксономическим статусом. Эти базы данных используются программами для сравнения с расшифрованных последовательностей ДНК ампликона метагенома, с целью таксономической идентификации микроорганизмов в исследуемом образце. Коллективами ученых разработаны и поддерживаются несколько баз данных полноразмерных последовательностей 16S и 18S рибосомальных РНК: SILVA (Quast C. et al. 2012), RDB (Cole J. R. et al. 2013), NCBI (Pruitt K. D., et al 2005). Кроме непосредственного сопоставления последовательностей базы данных хранят информацию о выравнивании последовательностей. Выравнивание последовательностей в базах данных осуществляется с использованием алгоритмов, учитывающих вторичную структуру 16S и 18S рибосомальной РНК, и консервативные домены этих маркеров. Выровненные последовательности в базе данных используются как шаблоны для выравнивания последовательностей из ампликона метагенома исследуемого образца.

К сожаления для исследования ампликонов метагенома сообществ на основе эукориотических генетических маркеров таких как COI или rbcl пока не разработано стандартных методов и подходов. Для метагенмных данных такого типа можно использовать базу данных Genbank (NCBI) и приложения BLAST для поиска гомологий последовательностей метагенома с последовательностями из базы данных. Имеются базы данных расшифрованных последовательностей этих маркера для некоторых групп многоклеточных эукариотических организмов. В основном такие базы данных используемых для баркодирования (Blaxter M. et al. 2005). Эти базы данных совместно с приложением BLAST можно использовать для идентификации организмов метагенома. Кроме того при

работе с ампликоном метагенома эукариот можно использовать набор методов, основанных на анализе разнообразия операционных таксономических единиц []

Процесс исследования биологического разнообразия на основе ампликона 16S рибосомальных РНК метагенома состоит из нескольких стадий: 1) удаление химерных последовательностей, 2) определение таксономического состава метагенома, 3) сравнительный анализ состава сообществ. Рассмотрим по порядку все три стадии анализа.

а) удаление химерных последовательностей

В процессе ПЦР реакции при амплификации фрагмента гена 16S рибосомальной РНК микроорганизмов в ампликон могут формироваться химерные последовательности ДНК. Части таких последовательностей содержать участи ДНК от одного вида организмов, а части последовательности участки ДНК другого вида организмов. При анализе ампликона такие химерные последовательности будут вносить некоторое искажение в состав разнообразия организмов анализируемой пробы. Большинство из химерных последовательностей не будут идентифицироваться по принадлежности к определенной таксономической группе в базе данных. Химерные последовательности необходимо удалить из исходного набора данных перед таксономической классификацией ампликона.

Для удаления химерных последовательностей необходимо выбрать референсную базу данных 16S рибосомальных РНК. Эта же база, в дальнейшем, будет использована для таксономической идентификации последовательностей метагенома. Из всех последовательностей выбранной референсной базы данных необходимо выделить участок 16S рибосомальной РНК, который использовался для амплификации ДНК анализируемого метагенома. Полученный таким образом массив последовательностей из части 16S рибосомальных РНК послужит новой референсной базой данных для обработки и идентификации видов в анализируемом метагеноме. Назовем эту базы данных укореченой референсной базой данных для анализа метагенома.

Следующим шагом является выравнивание с помощью функций используемых программ набора расшифрованных последовательностей ампликона метагенома путем сопоставления с последовательностями из укороченной базы данных. Выровненный набор последовательностей ампликона метагенома используется на всех стадиях дальнейшего анализа.

Большинство программ анализа таксаномического разнообразия метагенома на основе амипликона 16S рибосомальной РНК содержан набор функция для удаления химерных паследовательностей. Удаления происходит на основе сравнения участков последовательностей ампликона метагенома с последовательностями из укороченной базы данных. Если один из участков последовательности из ампликона близок к одной последовательности из базы данных, а другой участок близок другой последовательности из базы данных, то такая последовательность удаляется из анализируемого набора как химерная.

а)определение таксономического состава метагенома.

Для анализа таксономического состава сообщества на основе расшифрованных последовательностей ампликона метагенома 16S рибосомальной РНК используются два независимых подхода: 1) идентификация таксономического состава на основе сравнения с референсной базой данных 16S рибосомальной РНК, 2) определение операционных таксономических единиц (OUT) в составе метагенома на основе генетических дистанций между расшифрованными последовательностями. Первый подход имеет преимущество в том, что мы можем сравнить видовой состав сообщества и доли различных видов и других таксономических категорий микроорганизмов в рамка разных исследований, выполняемых по различным проектам. Недостатком метода является то, что многие редкие виды микроорганизмов могут попасть в разряд не идентифицированных, что приведет к потере

части информации о биоразнообразии сообщества. Подход на основе OTU позволяет учесть весь спектр организмов в метагегоме, но не позволяет непосредственно сравнивать результаты разных исследований. Если OUT были идентифицированы в рамках различных сессий по обработки данных, то сопоставить принадлежность расшифрованных последовательностей ДНК одной сессии обработки данных номерам OUT другой сессии обработки данных не представляется возможным. Чаше всего применяется комплексный подход с таксономической классификации по базе данных и классификации виде OUT.

Перед началом таксономической идентификации видов необходимо помнить следующие цифры: для микроорганизмов в пределах вида последовательности 16S рибосомальной РНК различаются мене, чем на 3% замен (несовпадающих нуклеотидов), в пределах рода между различными видами генетические дистанции лежат в пределах >%3 и ≤ 6% замен, для различных родов в нутрии семейства характерны дистанции >6% и ≤ 10%. Если исследуемые последовательности ДНК расшифрованы, с качеством каждого нуклеотида больше чем 20 единиц, то мы можем напрямую использовать набор данных для дальнейшей обработки. Если в наборе данных менее 1.5% букв имеют качество прочтении ниже 20 единиц, то разрешающая способность анализа уменьшиться до рода, виды микроорганизмов в этой ситуации разделить не возможно. Если от 1.5% до 3% процентов букв в последовательностях имеют качество прочтения менее 20 единиц, то разрешающая способность анализа сократиться до семейства. В любом случае если в наборе сравниваемых последовательностей имеются буквы с плохим качеством (менее 20 единиц), то необходимо произвести прекластеризацию. Процедура преклатеризации заключается в том, что в наборе те последовательности, которые разделены дистанцией менее чем выбранная величина, считаются одинаковыми, для дальнейшего анализа в наборе оставляется одна, из таких последовательности, информация о наличии остальных последовательностей идентичной искомой сохраниться отдельно, и используется в остальных видах анализа. Порог генетической дистанции, для того чтобы считать последовательности одинаковыми, выбирается исходя их характеристик качества прочтения и равен удвоенному доли букв с качеством прочтении менее 20 единиц.

После прекластерезации можно приступить к таксономической классификации по укороченной базе данных. Сопоставляя последовательности из базы данных с последовательностями из ампликона метагенома можно идентифицировать их принадлежность виду, роду, семейств и др. таксономическим категориям. Информация сохранеться в специальных текстовых файлах, которые используются в дальнейшем анализе. Ниже приведен пример файла, сохраненного программой MOTHUR для нескольких последовательностей 16S рибосомальной РНК ампликона.

```
6N_97702        Bacteria(99);Actinobacteria(47);Actinobacteria(47);Micrococcales(13);Microbacteriaceae(13);Cryocola(13);
6N_113188       Bacteria(94);Proteobacteria(64);Gammaproteobacteria(54);Chromatiales(47);Chromatiaceae(47);Thiorhodovibrio(47);
6N_112368       Bacteria(96);Firmicutes(19);Bacilli(14);Bacillales(14);Thermoactinomycetaceae(10);Thermoactinomyces(9);
```

В приведенном примере произведена идентификация последовательностей с точностью до рода. С лева записано имя последовательности и далее идет строка с установленным таксономическим статусом организма, обладающего этой последовательностью. При анализе таксономического статуса оценивается вероятность, того, что данная последовательность действительно соответствует установленной систематической категории. В вышеприведенной записи вероятность указана в скобках в процентах.

Как уже упоминалось выше, на уровне различных таксономических категорий для маркера 16S рибосомальной РНК установлены соответствующие пороги различия. Обозначим количество не совпадающих нуклеотид в последовательностях как x – генетическая дистанция. Соответственно при идентификации OUT применяются следующие

пороги различия (генетические дистанции) между последовательностями: виды ($x \leq 3\%$), роды ($3 < x \leq 6\%$), семейства ($6 < x \leq 10\%$), порядки ($10 < x \leq 15\%$), классы ($15 < x \leq 20\%$), филум ($20 < x \leq 25\%$) (Petrosino J. F. et al. 2009). Для того чтобы использовать данные пороги различия между последовательностями ДНК рассчитывается матрица генетических дистанций. При хорошем качестве прочтения и длине последовательностей порядка 400 пар основании в качестве меры генетической дистанции можно учитывать только количество совпадающих нуклеотидов при попарном сравнении выровненных последовательностей. При меньшей длине прочтения можно учитывать и инделы (пробелы или гепы) возникающие в процессе выравнивания. На основе матрицы генетических дистанций производиться реконструкция филогенетического древа. Обычно для такой реконструкции используется дистанционный метод объединения ближайших соседей (NJ) (Saitou N., Nei M. 1987) и его модификации. При анализе древа на нем можно выделить ряд кластеров. Если для выделения кластеров использовать порог с дистанцией $x \leq 3\%$, то получаться кластеры, объединяющие последовательности на уровне видов. Выделенные кластеры можно пронумеровать и обозначить OTU1, OTU2, OTU3, и т. д. Определенные таким образом таксономические единицы можно считать условными видами, количество условно соответствует количеству видов в метагеноме. Количество последовательностей в каждой из выделенных OUT будет условно соответствовать количеству особей вида в наборе данных. Таким же образом можно выделить кластеры с в пороге дистанций $3 < x \leq 6\%$, в этом случае кластеры будут соответствовать OUT на уровне рода. Кластеры так же можно будет пронумеровать и посчитать количество последовательностей – организмов в данном кластере и количество OUT рангом ниже – количество видов в роде. Таким же образом можно выделить OTU на любо таксономическом уровне. На рисунке 5 приведена общая схема определения таксономического состава метагенома.

Рис. 5. Схема определения таксономического состава сообщества на основе расшифрованных последовательностей ДНК ампликона метагенома.

После того, как были выделены необходимые для дальнейшего анализа уровни OTU исходя из набора последовательностей ампликона, можно сопоставить таксономические идентификаторы последовательностей, полученные на основе базы данных со списком последовательностей, входящих в состав OUT. Таким образом, можно идентифицировать таксономический статус некоторых OTU. Обычно при проведении такого анализа удается идентифицировать OUT для достаточно распространенных видов родов и семейств микроорганизмов. Чем выше таксономический статус исследуемых OTU, тем большую долю их удается идентифицировать на основе базы данных. В конечном итоге в исследуемом наборе последовательностей и выделенных на их основе OTU сохраняются не идентифицированные таксоны редко встречающихся видов микроорганизмов.

Точно по такому же алгоритму осуществляется анализ эукариотической составляющей сообщества микроорганизмов на основе ампликона 18S рибосомальной РНК. Единственная проблема здесь заключается в том, что эукариот не существует единой системы порогов различий на разных таксономических уровнях. Для разных групп видов характерны разные пороги различия на уровнях видов, родов, семейств и т. д.

б) Сравнительный анализ состава сообществ с применением метагеномного анализа на основе стандартных генетических маркеров

Обычно при проведении метагеномного исследования отбираются пробы из нескольких точек сбора или от нескольких исследуемых объектов. Для того чтобы провести анализ таких данных необходима следующая информация: 1) полный набор расшифрованных последовательностей ДНК сохраняемый в одном файле, 2) информация, сохраняемая обычно в отдельном файле в формате, требуемом программой по обработки метагеномных данных, содержащая соответствие между именем последовательности и принадлежности последовательности определенной точки сбора образцов, или объекту сбора материала. Данная информация является основой для реализации всех методов сравнительного анализа сообществ с помощью метагеномных исследований. Таким образом, получается, что мы имеем дело с набором последовательностей ДНК, разделенных на группы, такие, что каждая группа характеризует свое сообщество микроорганизмов.

Первый этап анализа расшифрованных последовательностей метагеномных ДНК состоит в определение уровней сходимости представленности долей OTU видов и других таксономических групп в метагеноме. Обычно для этого используют кривые насыщения образца количеством OTU в зависимости от количества расшифрованных последовательностей в образце. Анализ сходимости долевого состава сообщества необходимо проводить для каждой из исследуемых групп последовательностей. На рисунке 6 приведен пример кривой сходимости состава бактериального сообщества, оцененного по количеству OUT кластерного анализа расшифрованных последовательностей фрагмента гена 16S рибосомальной РНК. Представлена сходимость количества OTU на уровне видов (x=3%), родов (x=6%), семейств (x=10%) и порядков (x=15%).

Анализируя рисунок 6, мы видим, что исследуемое сообщество микроорганизмов представлено группой из 240 расшифрованных последовательностей 16S рибосомальной РНК. По мере вовлечения в анализ новых последовательностей от 0 до 240 растет количество OTU на исследуемых таксономических уровнях. Кривые для количества OTU на уровнях порядков и семейств выходят на насыщение при количестве анализируемых последовательностей в 200 штук. Кривые для количества OTU на уровнях вида и рода на насыщение не выходят. Это означает, что для проведения сравнительного анализа данного сообщества микроорганизмов с другим сообществом на уровне видов и родов потребует

увеличение размера выборки расшифрованных последовательностей ДНК. Сравнительный анализ на уровне семейств и порядков данного сообщества с другими сообществами даст адекватный результат на уже имеющимся количестве последовательностей.

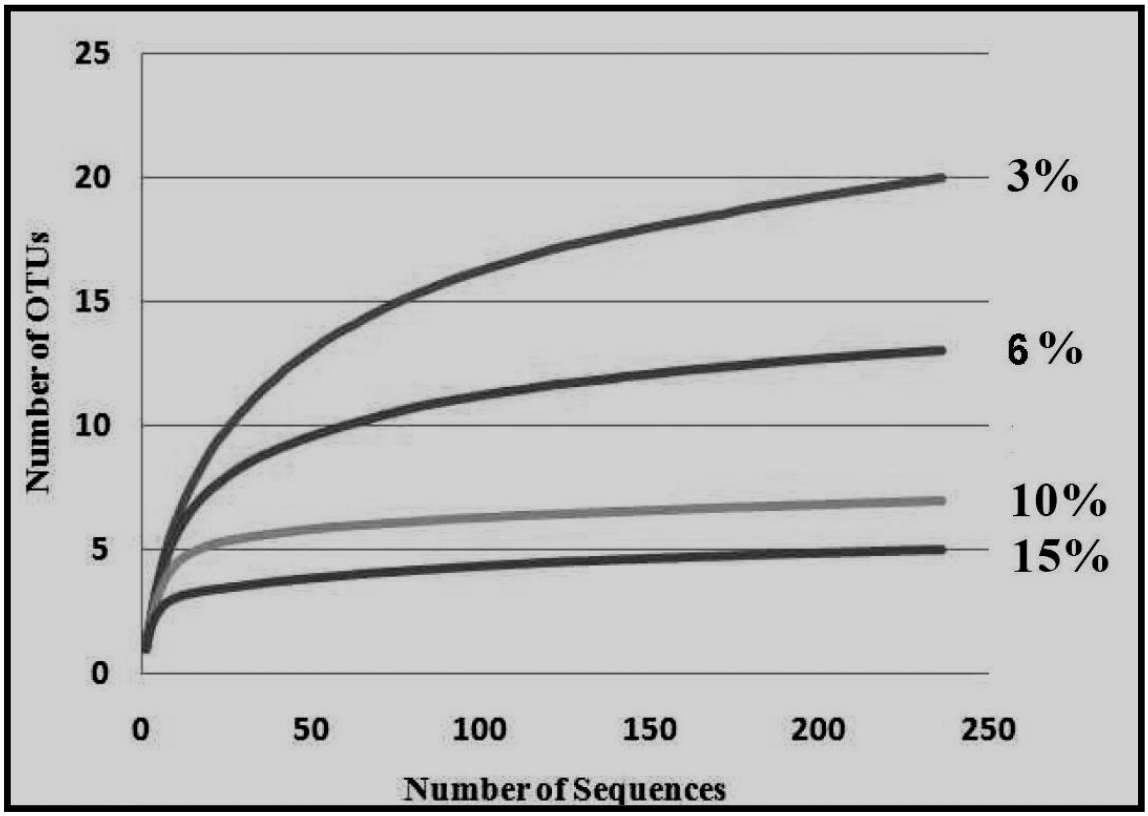

Рис. 6. Пример кривых насыщения при анализе количества OTU в сообществе микроорганизмов исследуемом на основе расшифрованных последовательностей ампликона 16S рибосомальной РНК. Представлены OTU на уровнях видов (*x*=3%), родов (x=6%), семейств (x=10%) и порядков (x=15%).

Отдельно следует заметить, что результат статистической сходимости данных, представленный на рисунке 6, не отражает общего состояния сходимости результатов метагеномного анализа на различных уровнях OTU. При анализе других сообществ, сходимость на уровнях семейств и порядков, может достигаться только при наличии нескольких десятков тысяч последовательностей, характеризующих сообщество или точку сбора материала. Чем богаче исследуемое сообщество по составу микроорганизмов, тем больше последовательностей потребуется для получения статистически сходящегося результата. Сообщество, охарактеризованное на рисунке 6, является достаточно бедным по таксономическому составу. Предварительный анализ состава на основе литературных и других данных может помочь в планировании эксперимента по исследованию метагенома на основе расшифровки ампликона. Используя предварительную информацию можно определить изначальное количество расшифровываемых последовательностей ДНК для анализа.

Если было установлено, что в процессе секвенирования расшифровано достаточно последовательностей ДНК для получения статистически достоверного результата, то можно переходить к следующие части метагеномного анализа, это оценка представленности различных таксономических категорий в исследуемом образце в абсолютных цифрах и в долях. Большинство средств для метагеномного анализа, такое как MOTHUR, QIIME, MG-RAST выдают информации об абсолютном количестве классифицированных и не

классифицированных OTU в текстовом виде. Текстовые файлы при этом оформлены таким образом, что информация в них содержится в виде столбцов с определенным типом разделителя. Чаще всего разделителем служит знак табуляции. Такое устройство файла позволяет без особых проблем импортировать его содержимое в одну из программ по созданию и обработке электронных таблиц (Microsoft Excel или OpenOffice Calc). Используя электронные таблицы можно оценить доли представленности различных видов, родов, семейств, и других таксономических категорий микроорганизмов среди последовательностей метагенома в виде столбчатых или круговых диаграмм. Используя визуализированные данные, можно стравить сообщества микроорганизмов друг с другом. Примеры столбчатой и круговой диаграммы приведены на рисунках 7 и 8 соответственно.

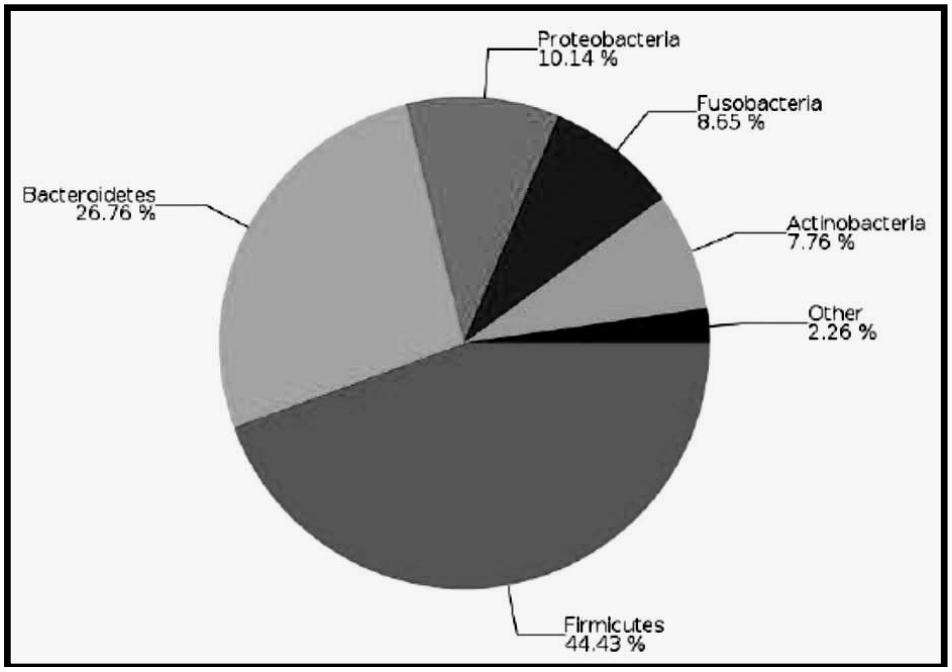

Рис. 7. Круговая диаграмма распределения бактерий в метагеноме расшифрованных последовательностей 16S рибосомальной РНК (показано распределения OTU на уровне филумов).

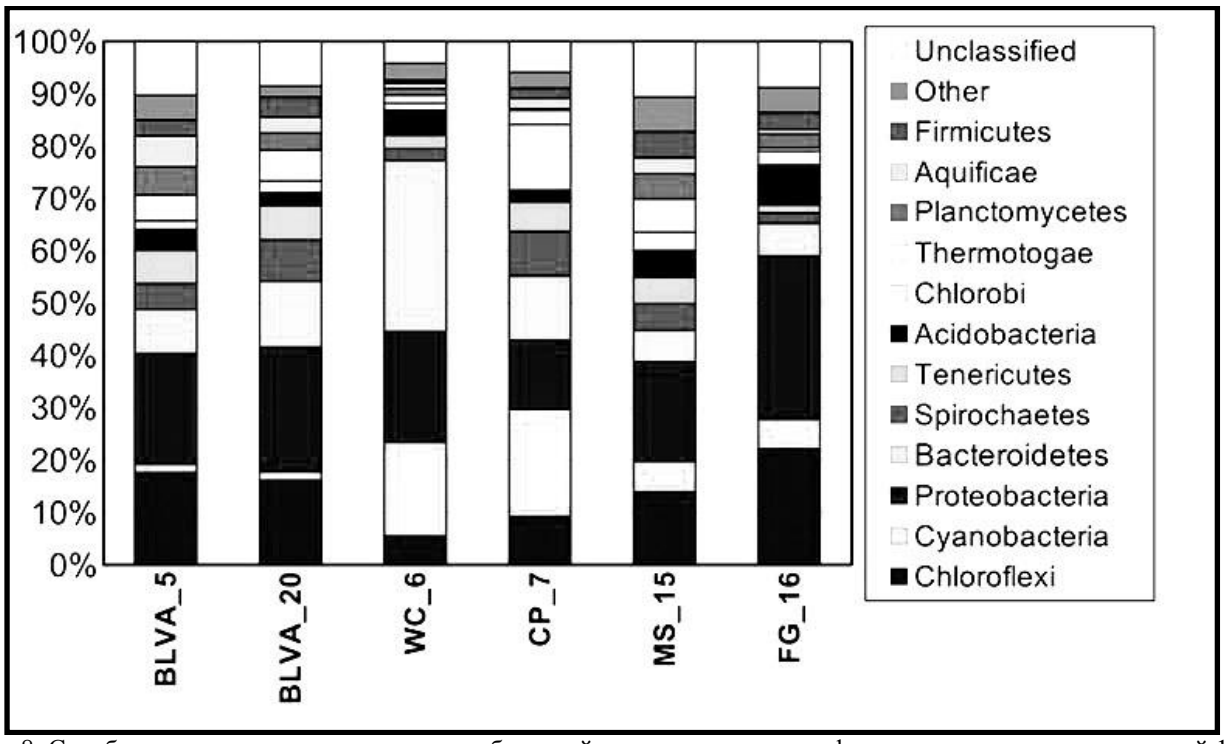

Рис. 8. Столбчатая диаграмма распределения бактерий в метагеноме расшифрованных последовательностей 16S рибосомальной РНК (показано распределения OTU на уровне филумов). Разные столбцы по разными названиями характеризуют состав метагенома в разных точках сбора образцов.

Кроме электронных таблиц для визуализации данных метагеномного анализа можно использовать средства языка программирования R (https://www.r-project.org/). Язык программирования R специально разрабатывался как средство для обработки и графической визуализации статистических данных. Результаты метагенгомного анализа как раз и являются одним их видов статистической информации. Язык R содержит достаточно обширный набор функция для предварительного импорта текстовой информации и набор библиотек и функция для графической визуализации статистических данных стандартными методами в виде различного рода столбчатых и круговых диаграмм. Одним из наиболее популярных пакетов в R для графической визуализации метагеномных данных стал ggplot2 (Ginestet C. 2011).

Круговые и столбчатые диаграммы позволяют на визуальном уровне сравнить доли различных таксономических групп микроорганизмов в метагеномах отобранных из разных сообществ (точек сбора, образцов биологического материала и т. д.). Наблюдаемые различия между долями представленности таксономических групп микроорганизмов в сравниваемых сообществах могут носит случайный характер и могут быть закономерными. Случайность. наблюдаемых различий может быть обусловлена тем, что анализируемая выборка расшифрованных последовательностей ДНК является частью генералной совокупности последовательностей всех организмов сообщества. Для того, чтобы определить, носят ли различия между сообществами случайный или закономерный характер применяется статистический метод AMOVA (Analysis of Molecular Variance) (Excoffier L., et al 1992). Метод позволяет вовлечь в анализ несколько сравниваемых сообществ. Для нескольких или попарно сравниваемых сообщества тестируется нулевая гипотеза H0 – достоверных различи в долях встречаемости последовательностей, характеризующих сравниваемы сообщества, нет. И альтернативная гипотеза H1 – различия во встречаемости разного тапа последовательностей в сообществах существуют. Как правило нулевая гипотеза отвергаеться.

если ее вероятность (P value) меньше чем 0.05. Метод AMOVA основан на анализе отношения генетических дистанций, характеризующих различия между организмами в нутрии сообществ и между сообществами. При сравнении нескольких сообществ одновременно, можно ответить только на качественный вопрос, различия между сравниваемыми сообществами существуют или нет. Если сравнивать два сообщества, то отношение средних внутригрупповых и межгрупповых дистанция может дать дополнительную информацию о составах сообщества. Если отношение дистанций >1 либо <1 гипотеза H0 принимается, то различие в составе сообществ нет. Если отношение средних внутригрупповых и межгрупповых дистанция <1 и гипотеза H0 отвергается, значит, видовые составы сообществ различаются, чем меньше значение отношение дистанций, тем больше различается видовой состав сообществ. Если отношение средних внутригрупповых и межгрупповых дистанция >1 и гипотеза H0 отвергается, то это означает, что одно из сообществ является подмножеством другого сообщества. Сообщество, с меньшей средней внутригрупповой генетической дистанцией является подмножеством сообщества с большей внутригрупповой генетической дистанцией.

Графическим средством визуализации метода AMOVA для сравнения сообществ микроорганизмов по составу их метагенома является метод многомерного шкалирования (multidimensional scaling) (Толстова Ю.Н. 2006). Метод основан на преобразовании положения в многомерном пространстве исследуемых объектов характеризуемых определенным многомерным вектором в обозримое двух или трехмерное пространство, так, чтобы расстояние между объектами в многомерном пространстве как можно более точно соответствовала расстоянию в обозримом пространстве. В нашем случае объект – это организм от которого была взята расшифрованная последовательность ДНК, вектор – это последовательность ДНК характеризующая организм. Размерность исходного пространства – количество букв и гепов в расшифрованной последовательности после выравнивания. Для того чтобы осуществить многомерное шкалирование, необходимо выбрать меру расстояния между многомерными объектами и рассчитать матрицу дистанция, на основе выбранной меры расстояния. При многомерном шкалировании в метагенгомном анализе в качестве меры расстояния используется доля несовпадающих букв между последовательностями. Для расчета многомерных преобразований применяется матрица дистанций, уже рассчитанная для кластеризации OTU и реализации метода AMOVA. Чаше всего при обработки метагеномных данных применяться двухмерное из многомерного в двухмерное обозримое пространно. Пример графика положения объектов (организмов) при двухмерном преобразовании методом многомерного шкалирования метагеномных данных представлен на рисунке 9.

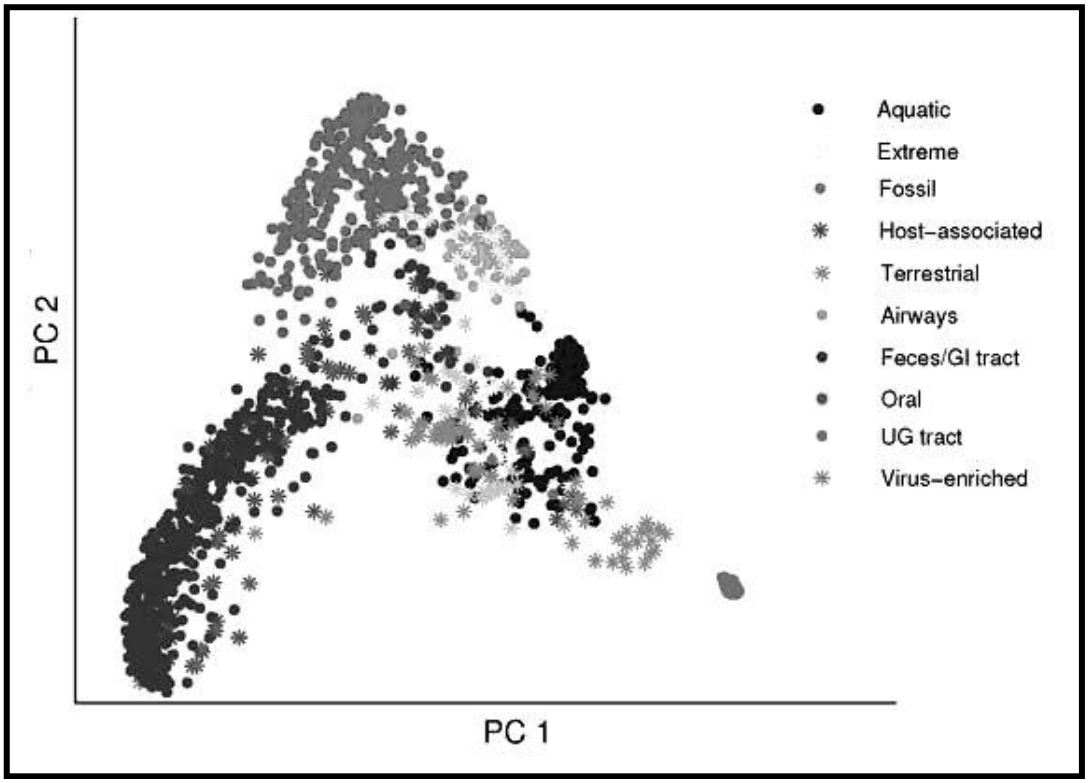

Рис. 9. Распределение положения объектов (микроорганизмов) в двухмерном пространстве на основе преобразования методом много мерного шкалирования инвормации, заключенной в векторах – паследовательностях расшифрованных фрагментов генов 16S рибосомальной РНК метагеномов различных сообществ микроорганизмов.

На рисунке 9 принадлежность организма сего расшифрованной последовательностью ДНК определенному сообществу обозначена различными геометрическими фигурами отличающихся цветов. Организмы сообществ образуют облака точе с определенной локализацией в пространстве. Чем ближе точки между собой в пространстве, тем более похожи их расшифрованные последовательности ДНК. Чем больше степень перекрытия облаков, тем более похожи сообщества по составу организмов. Если одно облако находиться в нутрии другого облака, то это демонстрирует ситуации при попарном сравнении сообществ с помощью AMOVA теста, когда отношение средних внутригрупповых и межгрупповых дистанция >1 и гипотеза H0 отвергается, и сообщество, с меньшей средней внутригрупповой генетической дистанцией является подмножеством сообщества с большей внутригрупповой генетической дистанцией.

Существуют два метода анализа степени похожести сообществ по составу OTU. Оба метода дают качественную оценку по наличию одинаковых и разных OTU в сопоставляемых сообществах. Первый метод это кластерный анализ представленности OTU в пробах и второй метод, это графическая визуализация наличия одинаковых OTU помощью диаграмм венна (Venn diagram).

Для кластерного анализа и для построения диаграмм венна необходимы следующие данные: номера OTU на заданном пороге кластерного расстояния между таксонами в сравниваемых сообществах (образцах, пробах), информация о наличии OTU заданного номера в конкретной пробе. Стоит заметить, что для анализа необходимо выбрать тот уровень кластерного расстояния, на котором наблюдается сходимость по количеству OTU данного номера во всех сравниваемых образцах.

Для построения дендрограммы сходства между сообществами по наличию или отсутствию OTU на заданном пороге кластерного расстояния рассчитывают матрицы дистанция между сообществами на основе следующей меры: считают суммарное количество OTU представленных в одном сообществе и отсутствующих в другом сообществе. Делят полученную величину на общее количество OTU, характеризующие все сообщества, сравниваемые в анализе. Полученную таким образом матрицу попарных дистанций используют для кластерного анализа методом UPGMA (невзвешенного попарного среднего) или NJ (объединения ближайших соседей) (Мандель И. Д. 1988). Рисунок 10 приводит пример кластерной дендрограммы, сравнивающей метагеномные образцы по наличию общих OTU.

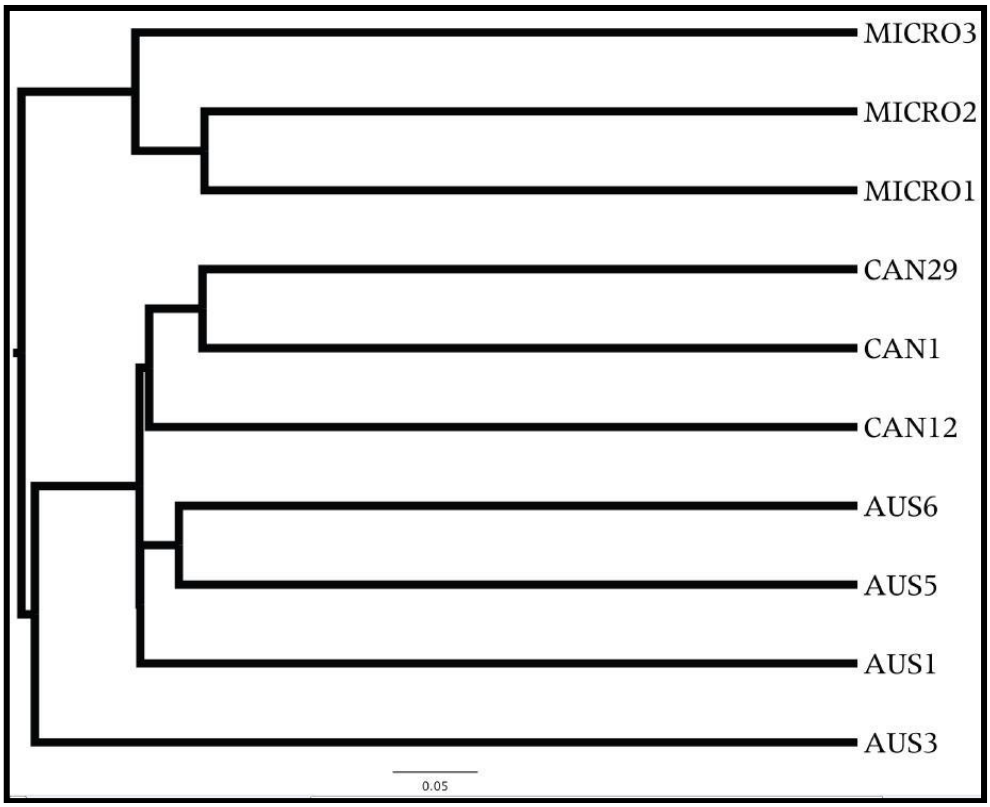

Рис. 10. Кластерная дендрограмма, построенная методом UPGMA, характеризующая различия меду составами сравниваемых сообществ по наличию одинаковых OTU на выбранном уровне кластерного расстояния. Длинна ветвей на древе соответствует доли несовпадающих OTU в сравниваемых сообществах.

Диаграммы венна (Кузичев А. С. 1968) строятся на основе того – же принципа и с использованием тех же данных, что и кластерный анализ по сходству OTU. Для каждого сообщества определяется количество OTU, характеризующего его состав. Для каждой пары сообществ определяется количество общих OTU, для каждой тройки сообществ определяется количество общих OTU и так уваливается ранг сравнения, пока не будет рассчитано количество общих OTU для всех сравниваемых сообществ. После того как искомая информация получена, результат визуализируется в виде геометрических фигур (обычно эллипсов или окружностей) количество которых равно количеству сравниваемых сообществ на пересечении геометрических фигур отмечают количество общих OTU. Рисунок 11 отображает пример диаграммы венна для сравнения четырех сообществ по содержанию OTU на уровне кластерного расстояния 0.05 – уровень рода.

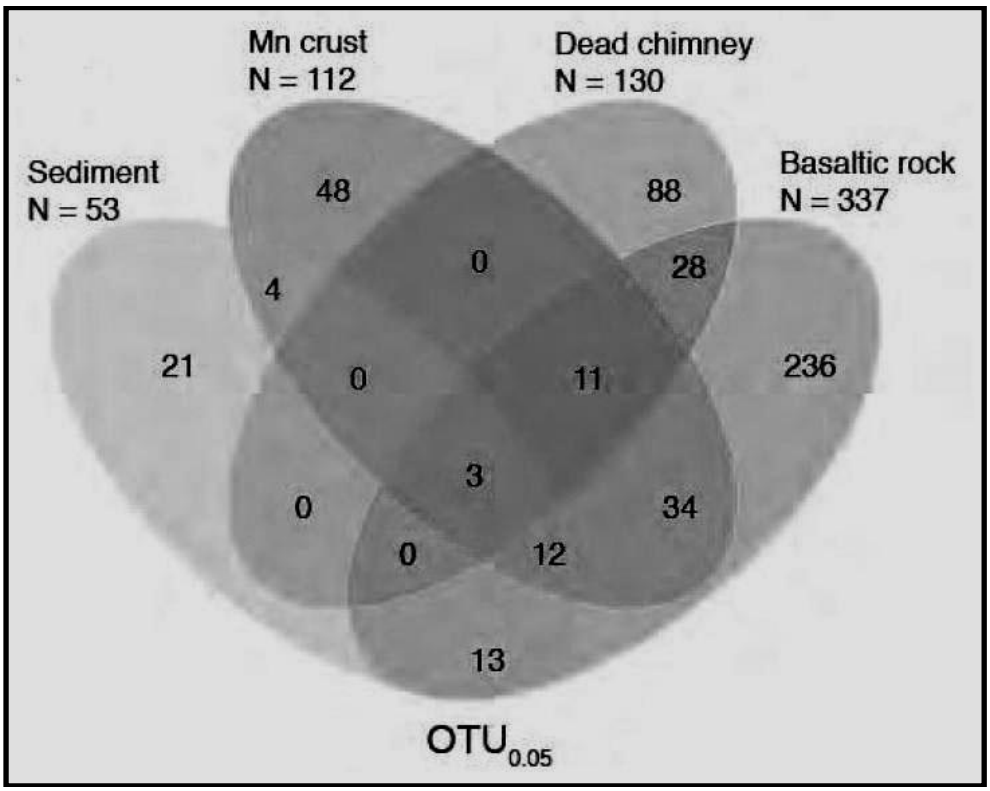

Рис. 11. Диаграмма венна сравнения 4 сообществ по содержанию OTU на уровне кластерного расстояния 0.05 – уровень рода. Отмечены выбранные для исследования названия сообществ и общее количество OTU на уровнях рода в них.

Следующим одним из наиболее информативных методов сравнения сообществ на основе результатов метагеномного анализа является метод главных компонент (Дубров А. М. 1978). Метод главных компонент относится к одному из методов уменьшения размерности массива данных. Исходной информацией для реализации анализа этим методом является таблица, содержащая информацию о наличии и количестве OTU на заданном уровне кластеризации в сравниваемых сообществах. Пример исходных данных для реализации метода главных компонент:

	P1	P2	P3	P4	P5	P6
OTU1	123	56	78	89	45	21
OTU2	156	15	38	83	46	0
OTU3	134	16	75	85	49	12
OTU4	67	186	67	78	34	19
OTU5	13	17	98	43	78	34
OTU6	23	19	7	23	90	63
OTUN ……						
OTU100	345	0	120	18	12	22

В таблице приведены названия сообществ (точек сбора, образцов биологического материала) в виде названий столбцов в таблице (P1, P2, P3 и т. д.) и названия OTU в качестве идентификаторов строк в таблице. На пересечении столбцов и строк содержится информация о количестве расшифрованных последовательностей ДНК соответствующего OTU в исследуемом сообществе. Количество OTU в таблице для реализации метода главных компонент должно превышать количество анализируемых сообществ.

Изначально, посмотрев на таблицу данных, мы видим, что каждое сообществ характеризуется числовым вектором с размерностью, равным количеству анализируемых

OTU. Человеческое восприятие способно адекватно обозреть визуализацию на графиках в двух или трехмерном пространстве. Для того чтобы уменьшит размерность пространства в методе главных компонент применяется серия сложных математических преобразований []. В результате для охарактеризовывания массива данных получают набор новых переменных, называемых главными компонентами. Каждое сообщество, после преобразование исходного массива данных получает новые числовые характеристики – главные компоненты. Каждая главная компонента характеризуется долей вариабельности от исходного массива данных, охватываемой рассматриваемой главной компонентой. Исходный массив компонент номеруется по вкладу в охват вариабельности исходного массива данных. Первая главная компонента охватывает наибольшую вариабельность, вторая главная компонента охватывает меньшую вариабельность массива данных, третья еще меньшую вариабельность массива данных и т. д. Для адекватной сравнительной оценки состав сообществ необходимо использовать такое количество первых компонент которое, в сумме, охватывает 95% и более вариабельности состава сообществ в исходном массиве данных. Обычно, при анализе методом главных компонент первых трех компонент достаточно для охвата 95% вариабельности исходного массива данных. Используя числовые значения главных компонент, рисуется координатная плоскость, с осями выбранных для анализа компонент, на которой отмечаются точки, соответствующие положению сообществ. На рисунке 12 приведен пример координатной плоскости с расположенными на нем точками, характеризующими положение сравниваемых сообществ, в пространстве первых двух главных компонент.

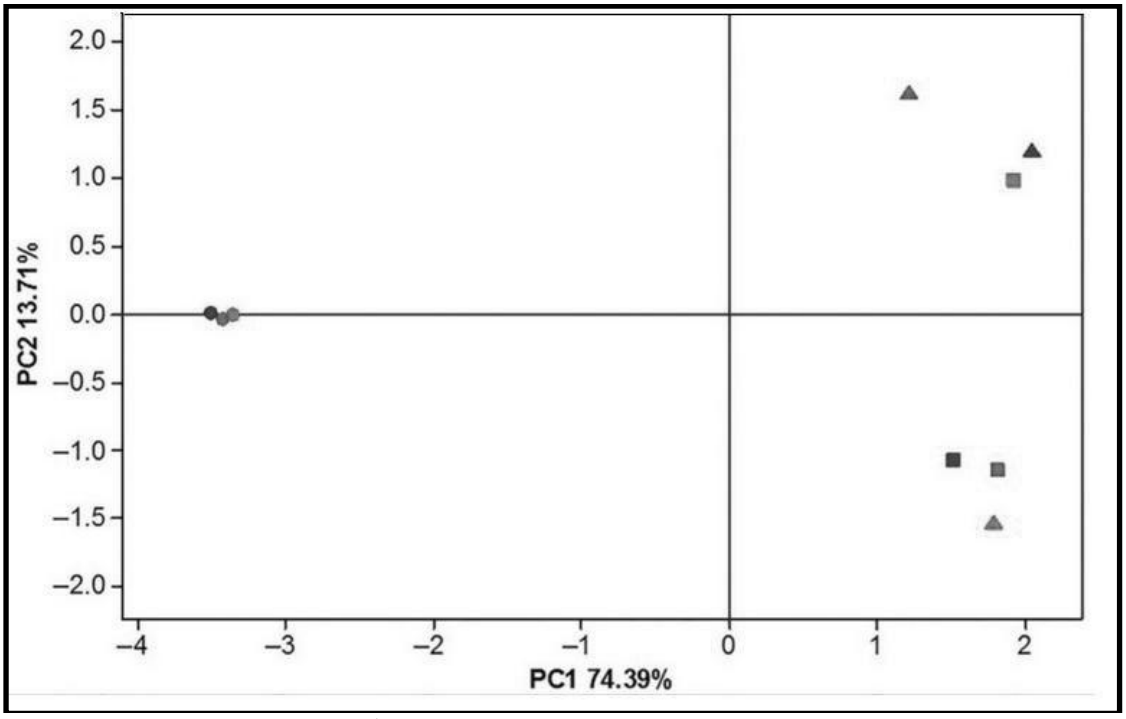

Рис. 12. Положение сравниваемых сообществ, в плоскости первых двух главных компонент на основе анализа метагеномных данных. Геометрическими фигурами разных оттенков отмечены разные анализируемые сообщества.

На рисунке 12 мы видим в плоскости двух главных компонент расположение сообществ микроорганизмов. Первая главная компонента охватывает 74.39% вариабельности массива данных, втора главная компонента охватывает 13.71% вариабельности массива данных, что в сумме составляет 88.1% процента вариабельности. Близость точек на

плоскости указывает близость состава сообществ по OTU. Чем ближе точки на плоскости друг к другу, тем более похожи сообщества по представленности OTU и по количеству расшифрованных последовательностей в соответствующих OTU. Так как суммарный охват вариабельности первыми двух компонентами меньше чем 95% целесообразно для анализа привлечь третью главную компоненту.

Кроме вышеуказанной информации метод главных компонент дает набор возможностей для анализа вклада разных OTU в формирования разнообразия главных компонент. Подробно с возможностями метода главных компонент можно ознакомиться в работах (Андерсон Т. 1963; Дубров А. М. 1978).

Для числового сопоставления степени сходства между сравниваемыми сообществами микроорганизмов на основе расшифрованных последовательностей ампликона метагенома все большее значение приобретает, достаточно новы метод, называемый в англоязычном сообществе исследователей «UniFrac» (Martin A. P. 2002; Lozupone C., Knight R. 2005). Метод основан на анализе филогенетического древа сообщества и рассчитывает степень перемешанности клад микроорганизмов на филогенетическом древе. Мера перемешанности рассчитывается так, что ее значение равно 1 при абсолютно независимых по составу сообществах образующих независимые клады на филогенетическом древе. И мера перемешанности равна 0 при различий в составе сравниваемых сообществ нет. Для филогенетического анализа обычно используется дистантный методы NJ или UPGMA. Матрица дистанций для филогенетического метода рассчитывается как доля несовпадающих нуклеотидов в выровненных последовательностях. Метод позволяет сравнить несколько сообществ между собой, построив для них общее филогенетическое древо. На рисунке 13 приведен пример анализа степени сходства между тремя сообществами. взятый из работы [].

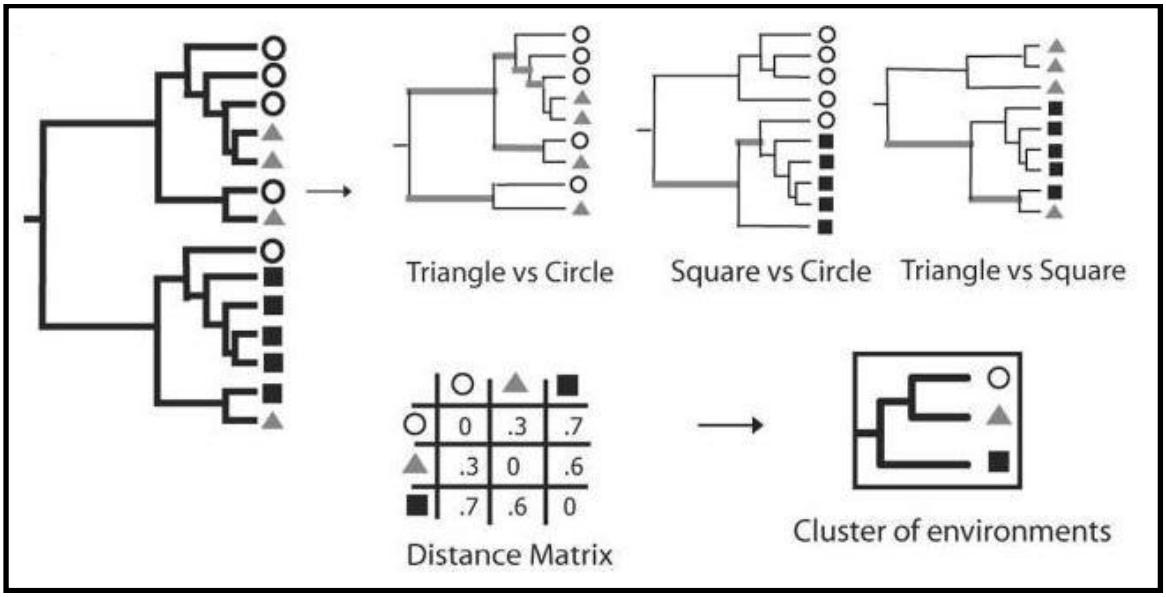

Рис. 13. Схема сравнительного анализа трех сообществ микроорганизмов с помощью метода «UniFrac», основанного на анализе филогенетического древа сообществ, реконструированного на основе расшифрованных последовательностей ДНК ампликона метагенома. Показано филогенетическое древо, схема анализа сходства – различия между сообществами, матрица значений мер различий между сообществами и кластерная дендрограмма степени сходства сообществ.

Кроме анализа степени сходства - различия между сообществами метод «UniFrac» позволяет протестировать нулевую гипотезу H0 – различий по составу организмов в сравниваемых сообществах нет. Альтернативная гипотеза H1 – различия в составе сообществ существуют и они достоверны. Для расчетов вероятности принятия нулевой гипотеза (P

value) применяется метод имитационного моделирования (метод монтекарло), описанный в работе []. Таким образом метод «UniFrac» позволяет не только рассчитать сперени различия – сходства в составах сравниваемых сообществ но и определить с помощью статистического теса степень доверия к полученным результатам. В какой-то мере, метод «UniFrac» повторяет уже рассматривавшейся ранее метод AMOVA.

Если на каком то уровне кластерного расстояния в сравниваемых сообществах выделены OTU, то для каждого из выделенных OTU в сравниваемых сообществах можно определить репрезентативную последовательность. Репрезентативная последовательность, это такая последовательность для которой среднее расстояние до всех других членов кластера, образующего OTU является минимальным. Для одних и тех же OTU в разных сообществах, анализируемых в исследовании, репрезентативная последовательность может оказаться разной. Во многих случаях репрезентативную последовательность удается найти в списке последовательностей, с определенным таксономическим статусом. Для разных сообществ, для одного и того же OTU репрезентативная последовательность может оказаться разной. Это свидетельствует пользу того, что сравниваемые сообщества отличаются друг от друга по соотношению таксонов в одних и тех же OTU. Для выявления репрезентативных последовательностей обычно рассматриваются OTU на уровне кластерного расстояния x=0.25 филумы. Для каждого выделенного OTU на уровне филума в каждом сообществе определяется репрезентативная последовательность. Репрезентативные последовательности ищется в списке с последовательностей с определенным таксономическим статусом. Затем, проводят сравнительны анализ OTU на уровне филумов, чем больше различается таксономический статус репрезентативной последовательности в пределах одного OTU в разных сообществах, тем больше различий в таксономическом составе рассматриваемого OTU наблюдается в сообществах.

Для сравнительного анализа и охарактеризования сообществ микроорганизмов, на основе расшифрованных последовательности ДНК ампликона метагенома, применяются и ряд других методов, используемых в классических биологических исследованиях. Часто для описания общества микроорганизмов применяется коэффициент биологического разнообразия Шеннона (Лебедева Н. В. и др. 2002). Некоторые программы для метагеномного анализа содержат процедуры для расчета этого коэффициента. Кроме того коэффициент Шеннона легко рассчитывается на основе результатов метагеномного анализа по разнообразию OTU или разнообразию таксономированных последовательностей, импортированных в электронные таблицы или среду R для обработки статистических данных. Коэффициент Шеннона связан с вероятностью того, что две случайно взятые из набора данных для сообщества микроорганизмов последовательности ДНК будут принадлежать разным таксонам одного ранга или разным OTU одного кластерного расстояния. Чем больше значения коэффициента Шеннона тем больше искомая вероятность. Таким образом, можно сказать, что если коэффициент Шеннона, рассчитанный по OTU одного кластерного расстояния или разнообразию таксонов одного ранга первого сообщества больше чем у второго, то биоразнообразие в первом сообществе больше чем во втором.

Обработка метагеномных данных, полученных методом дробовика

Как уже отмечалось ранее, метагеномный анализ путем секвенирования случайных нуклеотидных последовательностей применяется для решения нескольких задач. После предварительной обработки расшифрованных последовательностей ДНК метагенома производиться биоинформационный анализ, состоящий из нескольких этапов (см. рис. 14). Первая этап это идентификация таксономического состава организмов метагеномного образца и вторая этап это анализ функциональной активности генов в составе метагенома. Задача определения функциональной активности генов позволяет установить основные

метаболические пути синтеза и утилизации органического вещества сообществом. Заключительным этапом идет сравнение информации, полученной в ходе таксономической идентификации и анализа функциональной активности. Нам данном этапе исследователи пытаются определить, соответствует ли информация о таксономическом разнообразии сообщества информации о его функциональной активности.

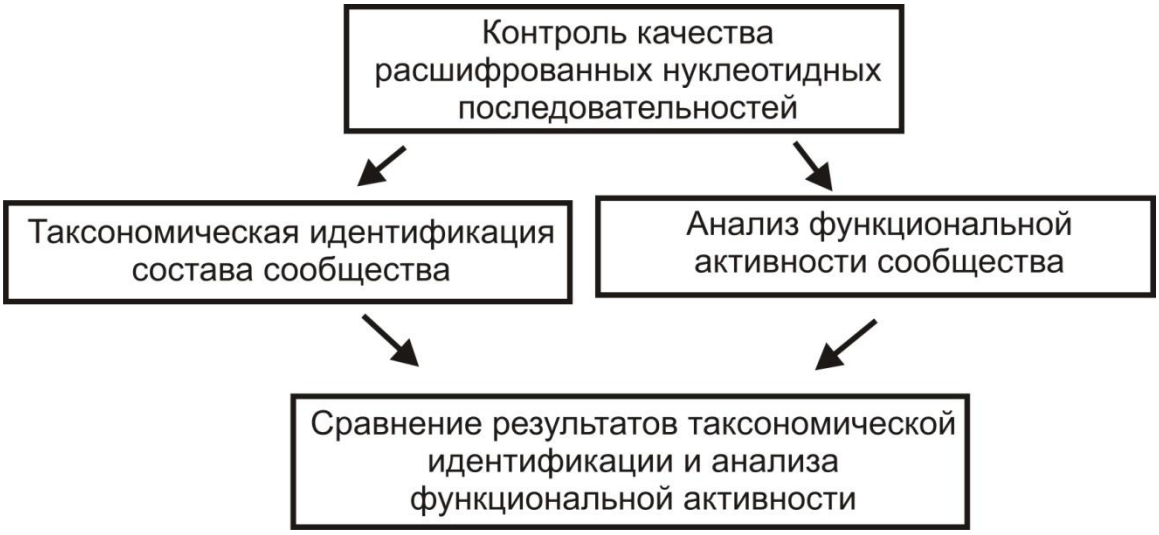

Рис. 14. Этапы анализа нуклеотидных последовательностей, полученных в ходе расшифровки случайных последовательностей метагенома.

Метагеномный анализ методом дробовика является единственным методом исследования вирусных сообществ. Далее мы рассмотрим поп порядку все этапы биоинформационного анализа расшифрованных нуклеотидных последовательностей метагенома, анализируемого методом дробовика. По ходу будут даны дополнительные комментарии по поводу анализа вирусных сообществ.

а) Таксономическая классификация сообщества

Таксономическая идентификация состава сообщества микроорганимов прокариот или эукариот может осуществляться двумя путями (см. рис .15). Первый путь состоит в выделении из набора расшифрованных последовательностей метагенома, фрагменты, кодирующие традиционные филогенетические маркеры 16S или 18S рибосомальную РНК. В итоге из общей массы расшифрованных последовательностей ДНК будет получена часть массива данных, которая может быть использована для таксономической идентификации состава сообщества с применением методов анализа ампликона метагенома. Весь путь проведения подобного анализа и спектр его методов описан в предыдущем разделе. Возникает естественный вопрос, с помощью какой методики из общего массива данных расшифрованных случайных последовательностей можно выделить подмассив фрагментов, кодирующих 16S или 18S рибосомальную РНК. Для этого можно использовать программы BLASTN из пакета программ BLAST (Altschul S. F. et al. 1990) и опорную базу данных, на основе которой затем буде производиться идентификация последовательностей. Опорная база данных используется для поиска с помощью приложения BLASTN фрагментов ДНК из общего массива данных гомологичных последовательностям базы данных. Выделенные из общего массива данных последовательности объединяются в отдельный файл (набор последовательностей), который выравниваться по опорной базе данных и используется для идентификации таксономического статуса организмов.

Обычно для секвенирования случайных последовательностей ДНК метагенома используют методы, прочитывающие достаточно короткие фрагменты ДНК (NGS

секвенирования пот технологии SOLiD (Hedges D. J. et al. 2011)) 100-300 пар нуклеотидов. В отличии от пиросеквенирования, выдающего фрагменты длинной до 400 пар оснований, разрешающая способность 100-300 пар основания позволяет определить расшифрованные последовательности ДНК максимум до уровня семейства. Об этом необходимо помнить, при определении таксономического состава сообщества по базам данных филогенетических маркеров 16S или 18S рибосомальных РНК при использовании секвенирования метагенома методом дробовика.

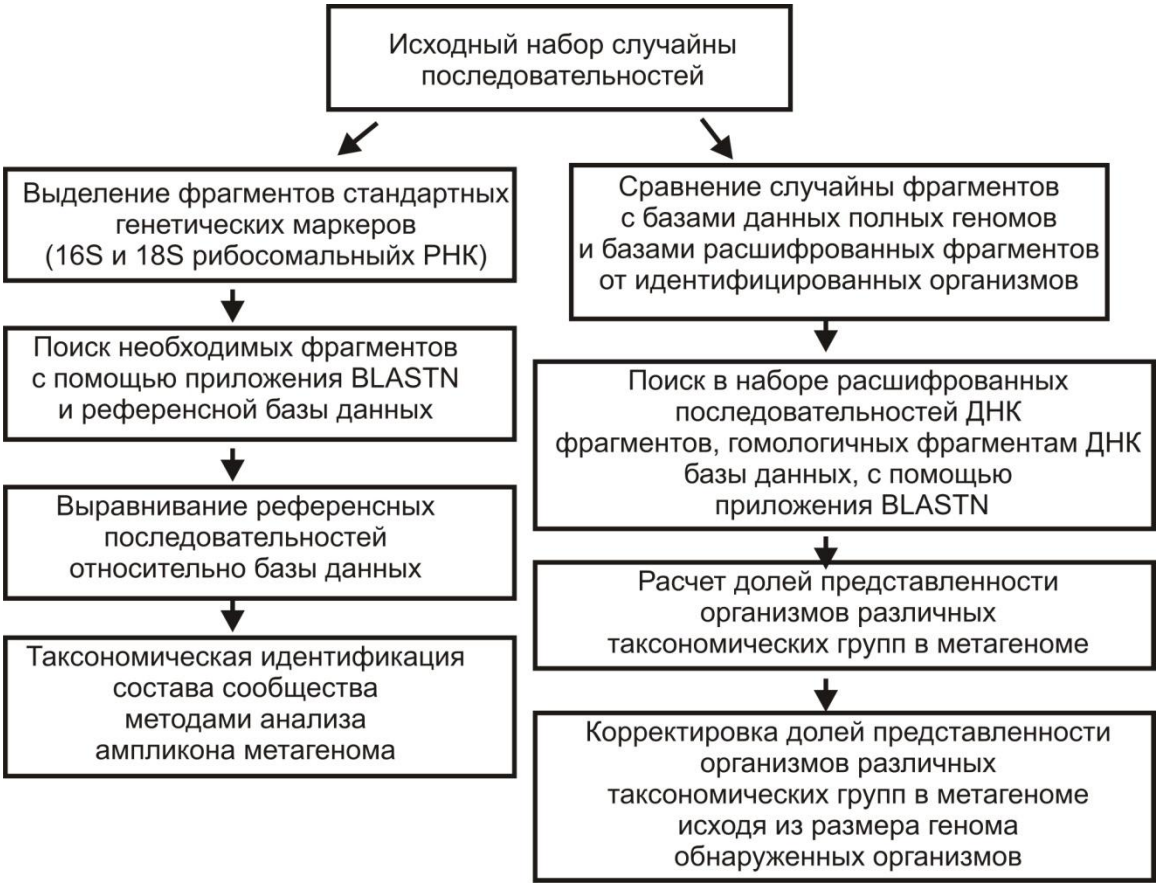

Рис. 15. Схема тасономической идентификации состава сообщества метагеномного образца, исследуемого методом секвенирования случайных фрагментов (метод дробовика).

Второй путь идентификации таксономической идентификации состава сообщества заключается в использовании информации из баз данных полных геномов прокариот, эукариот и вирусов (Pruitt K. D., et al 2005). Идентификацию состава сообщества можно провести самостоятельно, воспользовавшись приложением BLASTN. Для этого необходимо выбрать референсную базу данных расшифрованных полных геномов и сопоставить расшифрованные последовательности метагенома с выбранной базой данных. Если таким поиском с помощью BLASTN обнаружена гомология между последовательностью метагенома и последовательность из базы данных, то можно говорить о родственной связи сравниваемых организмов. Далее необходимо оценит генетическую дистанция между последовательностью метагенома и часть полного генома, с которым было найдено сходство. Генетическую дистанцию можно определить на основе информации, выдаваемой программой BLASTN. Приложение BLASTN определяет степень гомологии между последовательностями в процентах, соответственно генетическая дистанция x в долях замен будет определяться как 100 минус степень гомологии, в процентах, деленная на 100. Для

прокариот, можно выделить коридоры генетических дистанций для различий на соответствующем таксономическом уровне также как и для 16S рибосомальноф РНК. В пределах вида нуклеотидные последовательности различаются мене, чем на 3% замен (несовпадающих нуклеотидов), в пределах рода между различными видами генетические дистанции лежат в пределах >%3 и ≤ 6% замен, для различных родов в нутрии семейства характерны дистанции >6% и ≤ 10%. Соответственно, ели между расшифрованной последовательность метагенома и участком последовательности из базы данных дистанция $x<3\%$ то последовательность из метагенома принадлежит тому же виду, что и последовательность из базы данных. Если дистанция лежит в пределах $3\%\leq x\leq 6\%$, то сопоставляемые последовательности принадлежат одному роду микроорганизмов. Таким образом, делая подобные рассуждения можно определить таксономический статус с определенным разрешением до уровня вида, рода, семейства и т. д. для части последовательностей в прочитанном метагеноме. Обычно, таким способом удается идентифицировать небольшую часть последовательностей (≈10%).

Далее необходимо рассчитать на заданном таксономическом уровне доли представленности последовательностей. Эти доли необходимо скорректировать, с учетом длинны геномов микроорганизмов. Длинна геномов разных микроорганизмов может отличаться в разы. Естественно, что представленность среди последовательностей метагенома, организма, с большей длинной генома будет большей. Для коррекции долей за основу берут самую короткую длину генома из представленных микроорганизмов. Эту длину берут за единицу и называют референсной длинной. Затем делят длины геномов остальных микрооганизмов на референсную длину и получат цифру, показывающую во сколько раз необходимо уменьшить представленность рассматриваемого микроорганизма в наборе данных.

Для эукариотических организмов и вирусов подобная технология определения таксономического состава сообщества также применяется. Однако ситуация затрудняется отсутствием четких коридоров генетических дистанция для разных таксономических категорий.

Осуществить процесс таксономической идентификации состава сообщества по вышеописанному алгоритму можно с помощью языков программирования Perl (https://www.perl.com) и Phyton (https://www.python.org/) с использование средств библиотек BioPerl (Stajich J. E. et al. 2002) и BioPhyton (Cock P. J. A. et al. 2009). Однако в этом случае потребуются знание в области программирование. Обычно подобные навыки позволят биоинформатикам совершать широкие маневры в области обработки и извлечению информации из метагеномных данных. Альтернативной использованию скриптовых языков программирования являться использование готовых программных продуктов, таких как (Wood D. E., et al 2014). Многие из разработчиков этих программных средств вкладывают усилия и в формирование специальных баз данных расшифрованных последовательностей ДНК, входящих в состав программных пакетов. Эти базы данных содержат не только полные расшифрованные геномы прокариот, эукариот и вирусов, но и разнообразные последовательности различных генов и локусов от идентифицированных видов организмов. Такая база данных расширяет возможности по точной идентификации таксономического разнообразия сообществ.

Для обработки данных, по метагеномному секвенирования вирусных сообществ также может применяться база данных полных расшифрованных вирусных геномов и приложение BLASTN. Технология определения состава сообщества аналогичная той, которая описана в предыдущих абзацах. Особенность функционирования вирусов состоит в том, что для многих видов вирусов существует стадия, когда их генетическая информация встроена в состав генома хозяина. в некоторых случаях это можно установит факт встраивания генетической

информации вируса в геном хозяина. Если при NGS секвенирования были произведены парные прочтения (парные риды) – расшифровка протяженного участка ДНК с двух концов при этом не произошло перекрытия прочтения, так что оба прочтения остаются разделенными протяженным не расшифрованным фрагментом ДНК. При идентификации таких последовательностей по базам данных, один рид может оказаться гомологичен участку вирусного генома а парный ему рид участку генома хозяина. Это и является признаком того, что вирусный геном встроен в состав генома хозяина.

Для обработки вирусных метагеномов используют и специальное программное обеспечение и онлайн сервисы, например, такие как MetaVir (Roux S. et al. 2011), VIROME (Wommack K. E. et al. 2012), VMGAP (Lorenzi H. A. et al. 2011). Данные инструменты позволят не только провести таксономическую классификацию сообществ, но провести анализ функциональной активности вирусов в метагеноме.

б) Функциональная характеристика сообщества

Поиск различных категорий генов и анализ функциональной активности сообществ на основе случайных расшифрованных последовательностей ДНК состоит из нескольких этапов. Последовательность реализации анализа такого вида представлена на рисунке 16.

Рис. 16. Схема проведения анализа функциональной активности сообщества на основе данных по секвенирования случайных последовательностей метагенома сообщества.

Первым этапом анализа функциональной активности сообщества на основе исследования его метагеномя является сборка расшифрованных последовательностей ДНК в контиги. Контигом в этом случае называют набор перекрывающихся сегментов ДНК,

которые в совокупности представляют собой консенсусную область ДНК. При сборке последовательности в контиги получат достаточно протяженнее участки ДНК, содержащие одну или несколько белок кодирующих последовательностей (генов) ил их части. Для сборки последовательностей в контиги применяется программное обеспечение, такое как Velvet (Zerbino D. R., Birney E. 2008), Abyss (Simpson J. T. et al. 2009). Сушествет и специально адаптированное к сборке в контиги метагеномных данных программное обеспечение: Meta-IDBA (Peng Y. et al. 2011), MetaVelvet (Namiki T. et al. 2012). Визуализировать, собранный контиг, можно с помощь программы CBrowse (Li P. et al. 2012).

Следующий этап анализа связан с поиском в собранных контигах открытых рамок считывания и предсказания участков кодирующих белки. По другому данная часть исследования носит название аннотация метагенома. Способы аннотации и применяемые средства зависят от средней длинны собранных контигов. Если средняя длинна контига 30 тысяч пар нуклеотидов, то для поиска белок кодирующих участков можно применить те же программ, что используются для аннотации полных геномов: IMG (Markowitz V. M. et al. 2012). Если набор собранных контигов обладает меньшей средней длинной, то необходимо воспользоваться специальными средствами для аннотации метагеномов: MG-RAST (Meyer F. et al. 2008) MetaGeneMark (Zhu W., et al 2010), Metagene (Noguchi H., et al. 2006). После того, как в контигах найдены части последовательностей, кодирующих белки, эти участки можно транслировать в аминокислотные последовательности и попытаться определить функциональный класс. Фукционалный класс генов определяется теме же программами для аннотации контигов. Существует несколько функциональных классов белков: ферменты, рецепторы, транспортеры, сигнальные белки, структурнее белки. Каждый из итих классов белков в свою очередь подразделяется еще на подклассы.

Выделенные из метагенома аминокислотные последовательности белков и кодирующие их нуклеотидные последовательности используются для точного функционального ахарактеризовывания генов в метагеноме. Для этого используется сравнение с помощью программы BLASTP (Altschul S. F. et al. 1990) аминокислотных последовательностей белков с уже известными по функциональной активности белками из базы данных. Существует несколько баз данных белков с известной функциональной активностью. К таким базам данных относятся: KEGG (Kanehisa M., Goto S. 2000), COG (Tatusov R. L. et al. 2000), The SEED (Overbeek R. et al. 2014), UniProt (Apweiler R. et al. 2004). Белкам в этих базах данных присвоены некоторые категории, соответствующие определенному уровню иерархии. Часть подобных баз данных содержат встроенные средства для поиска гомологии в последовательностях на основе BLAST алгоритма с выдачей систематизированного результата, по обнаруженному совпадению последовательностей из контигов метагенома с последовательностями из базы данных. При сопоставлении аминокислотных последовательностей с помощью BLAST алгоритма выделяют следующие коридоры гомологии, определяющие функциональное сходство белков: 30-35% сходства – сравниваемые белки отделились в прошлом от одного предка, 40-50% сходства – сравниваемые белки имеют общий план трехмерного строения и т. д. Если аминокислотные последовательности двух белков совпадают на 95%, после выравнивания, то можно говорить о том, что они выполняют одну и ту же функцию в организме, например, катализируют одну и ту же ферментативную реакцию. Следует отметить, что современные методы аннотации и определения функциональной активности белков в метагеноме позволяю охарактеризовать только 20-50% последовательностей в метагеномных данных.

После того как функциональна активность части генов в контигах определена, можно приступить к изучению представленности последовательностей метагенома, гомологичных исследуемым генам контига. Англизируя представленность последовательность метагенома, отвечающих за различные функции можно определить основные пути получения

трансформации и утилизации органического вещества в сообществе микроорганизмов и пути преобразования энергии. Представленность последовательностей в метагеноме. гомологичных какому либо гену контига можно определить с помощью приложения BLASTN. На рисунке 17 приведен пример исследование динамики изменения долей генов в метагеноме, участвующих в метаболизме углеводов участника проекта МАРС-500. Рисунок взят из работы (Марданов А. В. и др. 2013).

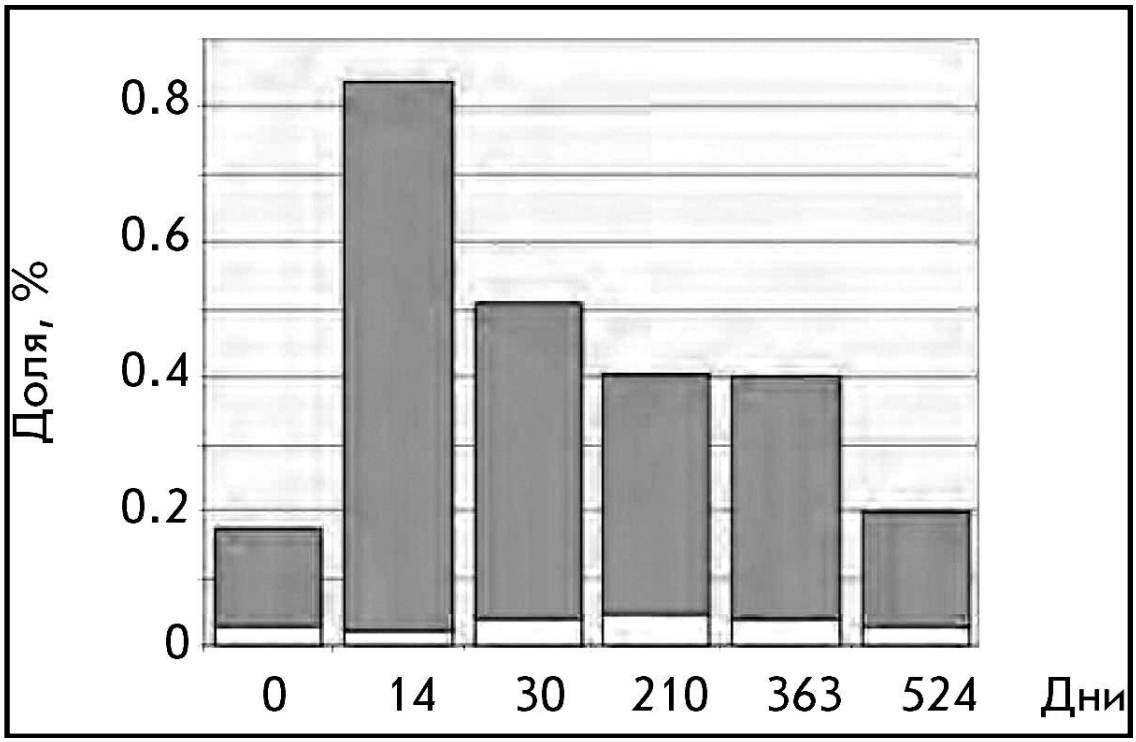

Рис. 17. Динамика изменения доли бактериальных генов в кишечнике человека, участвующих в метаболизме углеводов в ходе выполнения проекта «МАР-500». Белым указаны доли генов, отнесенные к Bacteroidetes, серым – к Firmicutes

Для оценки доле представленности определенных генов в метагеноме можно использовать комплексные специализированные средства для метагеномного анализа, такие как MG-RAST. Данный интернет сервис например использовалось для получения информации, отображенных на диаграмме рисунка 17.

В данном обзоре приведен только краткий список возможностей и средств метагеномного анализа. Метагеномный анализ является наиболее быстро развивающимся инструментом функционального исследования сообществ живых организмов. Кроме широкого применения в исследовании сообществ микроорганизмов, метагеномный анализ займет свою нишу и при исследовании других сообществ.