КАЗАНСКИЙ (ПРИВОЛЖСКИЙ) ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
Институт фундаментальной медицины и биологии

М.И. МАРКЕЛОВА, Е.А. БУЛЫГИНА, М.Н. СИНЯГИНА, А.М. СЕНИНА, Т.В. ГРИГОРЬЕВА

АНАЛИЗ ДАННЫХ СЕКВЕНИРОВАНИЯ АМПЛИКОНОВ ГЕНА 16S РРНК С ПОМОЩЬЮ WEB-ПЛАТФОРМЫ MICROBIOMEANALYST

Учебное пособие

КАЗАНЬ
2024

УДК 579.25
ББК 28.0
М23

Печатается по рекомендации учебно-методической комиссии Института фундаментальной медицины и биологии КФУ (протокол № 3 от 16.10.2024 г.)

Рецензенты:
кандидат биологических наук, доцент **Козлова О.С.**
кандидат ветеринарных наук, ассистент **Задорина И.И.**

Маркелова М.И., Булыгина Е.А., Синягина М.Н., Сенина А.М., Григорьева Т.В.
М23 Анализ данных секвенирования ампликонов гена 16S рРНК с помощью web-платформы MicrobiomeAnalyst: учебное пособие / М.И. Маркелова, Е.А. Булыгина, М.Н. Синягина, А.М. Сенина, Т.В. Григорьева// Казанский федеральный университет, 2024. – 95 с.

В учебном пособии приведена пошаговая схема анализа данных секвенирования ампликонов гена 16S рРНК с использованием web-платформы MicrobiomeAnalyst от сырых прочтений до статистического анализа и поиска биомаркеров. Рекомендовано для изучения дисциплины: «Б1.В.ДВ.03.02. Метагеномика» биологических специальностей, а также при подготовке курсовой работы по специальности, научно-исследовательской работы и выпускной квалификационной работы медицинских и биологических направлений.

УДК 579.25
ББК 28.0

© Маркелова М.И., Булыгина Е.А., Синягина М.Н., Сенина А.М., Григорьева Т.В.
© КФУ, 2024

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....5

1. Типы метагеномных исследований 7

1.1. Шотган-метагеномика 7

1.2. Секвенирование ампликонов целевых генов 11

1.2.1. Ген 16S рРНК.....12

1.2.2. Основные этапы метагеномного анализа на основе ампликонов гена 16S рРНК 14

1.2.3. Базы данных референсных последовательностей гена 16S рРНК и популярные программы для анализа данных 21

ОТ СЫРЫХ ПРОЧТЕНИЙ К ASV.....23

2.1. Загрузка данных учебного проекта на персональный компьютер.....23

2.2. Загрузка ридов на web-платформу MicrobiomeAnalyst 23

2.3. Подготовка данных к анализу 27

2.4. Выбор параметров анализа.....29

2.5. Анализ данных 31

2.6. Результаты анализа.....33

СТАТИСТИЧЕСКАЯ ОБРАБОТКА И ВИЗУАЛИЗАЦИЯ ДАННЫХ 42

3.1. Загрузка файлов для визуализации и статистического анализа 42

3.2. Фильтрация и нормализация загруженных данных 45

3.3. Результаты.....47

3.3.1. Классическая визуализация таксономического состава.....47

3.3.2. Разнообразие микробного сообщества 56

3.3.3. Кластеризация и сети корреляций.....65

3.3.4. Сравнение и классификация 73

3.3.5. Сохранение результатов 89

ЗАКЛЮЧЕНИЕ 90

ЛИТЕРАТУРА 91

ВВЕДЕНИЕ

Метагеномом, по аналогии с геномом – совокупностью всей ДНК клетки одного организма, называют совокупность геномов всех организмов, присутствующих в образце окружающей среды (от греч. «мета» – «над, сверх», охватывающий все). Раздел науки геномики, изучающий метагеномы, называют метагеномикой. Совокупность всех генов сообщества называют микробиомом, а совокупность всех микроорганизмов сообщества – микробиотой. Метагеномный подход позволяет исследовать исходное разнообразие микроорганизмов среды, минуя этапы культивирования и выделения отдельных клеток. Данная методика стала возможной благодаря развитию технологии высокопроизводительного секвенирования нуклеиновых кислот (секвенирование следующего поколения – Next-Generation Sequencing, NGS), разработке биоинформатических алгоритмов для анализа получаемых данных и росту вычислительных мощностей.

Для чего может быть полезно определение «суммарной» генетической последовательности микробного сообщества? Среди задач, решаемых с помощью этого подхода, можно выделить следующие:

- идентификация новых видов микроорганизмов;
- поиск генетических систем – детерминант антибиотикоустойчивости, вирулентности, патогенности, и отслеживание их распространения в микробном сообществе;
- мониторинг динамики видового состава микробного сообщества и изучение его ответа на внешние факторы;
- обнаружение патогенных микроорганизмов в клинической практике, определение видов-маркеров патологических состояний, и многое другое.

Метагеномный подход имеет ряд преимуществ перед классическими микробиологическими методами идентификации микроорганизмов, основанными на изолировании бактерий на селективной среде, получении штамма и его культивировании. Во-первых, появляется возможность для описания видов, плохо поддающихся

культивированию, в то время как такие организмы составляют бóльшую часть природных экосистем (от 99%) [Amann *et al.*, 1995]. Во-вторых, сохраняется исходное численное соотношение между микроорганизмами в образце. В-третьих, применение технологии шотган-метагеномного секвенирования позволяет раскрыть не только таксономический, но и функциональный состав сообщества с высоким разрешением, вплоть до количественной представленности отдельных генов.

Ниши, которые чаще всего исследуются с помощью методов метагеномики:

- Кишечник человека (пристеночная микробиота – биопсия стенки кишечника, просветная микробиота – исследование фекалий);
- Кожа, урогенитальный тракт, ротовая и носовая полости человека и т.п.;
- Почвы (разного типа, на разной глубине, с разным антропогенным загрязнением);
- Водоемы (озера, реки, моря);
- Микробиота сельскохозяйственных животных (рубец крупного рогатого скота, кишечник птиц);
- Микробиота культурных растений (поверхность листьев и корней, корневое окружение);
- Микробиота диких животных и дикорастущих растений для оценки биоразнообразия и поиска потенциальных пробиотиков.

В настоящее время существует два основных подхода для определения состава сообщества с помощью технологии NGS: секвенирование тотальной ДНК, выделенной из образца (шотган-секвенирование), и секвенирование ампликонов целевых генов (рис. 1). Термин «метагеномика» был предложен в 1998 г. для обозначения исследований, в которых проводится анализ коллективных *геномов*, полученных из образцов окружающей среды [Handelsman *et al.*, 1998], и в этой трактовке данному термину соответствует подход с использованием шотган-секвенирования – определение последовательности *тотальной*

ДНК из природного образца. Однако в дальнейшем термин «метагеномика» стали применять в более широком смысле – как набор методов для получения генетической информации о микробном сообществе, минуя этап культивирования. В этом смысле метагеномными можно называть также исследования, основанные на амплификации целевых генов.

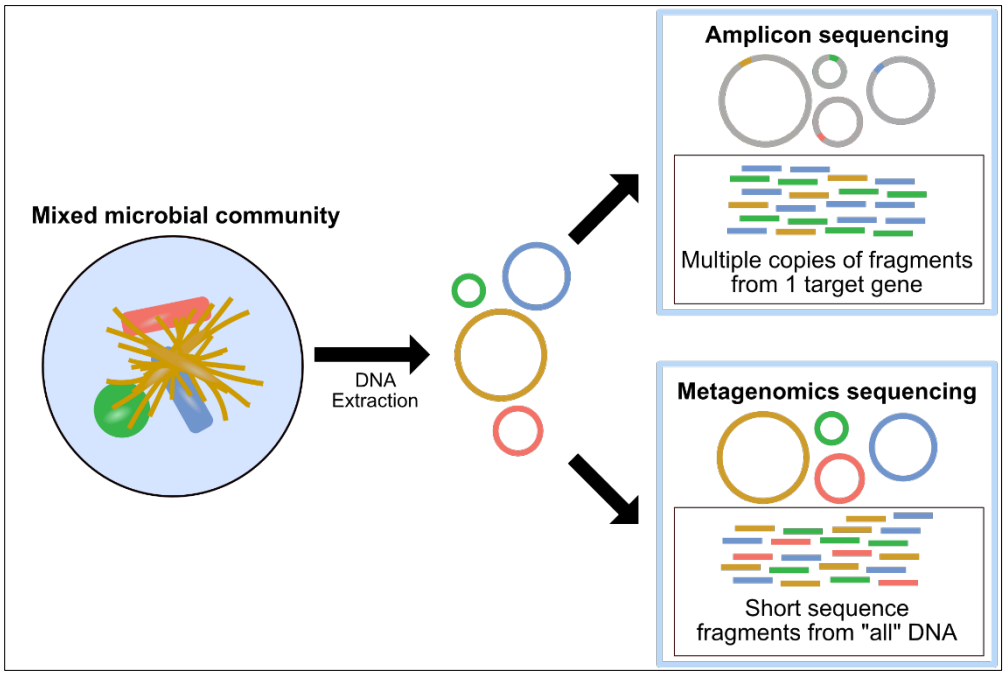

Рис. 1. Основные подходы метагеномики (https://astrobiomike.github.io/misc/amplicon_and_metagen)

Стоит подчеркнуть, что метагеномный анализ позволяет определить структуру не только бактериальной части микробного сообщества, но других его составляющих – архей, грибов, микроскопических эукариот, вирусов.

1. Типы метагеномных исследований

1.1. Шотган-метагеномика

Шотган-секвенирование (от англ. “shotgun” – дробовик) – метод прочтения последовательности нуклеиновых кислот, при котором исходная цепь ДНК разбивается на небольшие случайные фрагменты, которые затем секвенируются параллельно, после чего исходную

последовательность можно восстановить по перекрытию этих фрагментов. Метод используется для определения последовательности полных геномов организмов, а также хорошо подходит для функционального профилирования, так как позволяет прочитать все гены всех микроорганизмов, находящихся в образце.

Этапы анализа шотган-метагеномов:

1. Выделение тотальной ДНК сообщества;
2. Фрагментация ДНК на короткие фрагменты (200-600 п.о.) с помощью ферментов или механически (например, ультразвуком);
3. Затупление концов, пришивание адаптеров (технических последовательностей) и уникальных последовательностей (баркодов), и отбор фрагментов определенной длины, подходящей к имеющемуся секвенатору;
4. Смешивание в эквимолярном соотношении баркодированных библиотек;
5. Высокопроизводительное секвенирование полученных фрагментов (~30 млн. ридов, что в ~1000 раз больше, чем для секвенирования ампликонов).
6. Биоинформатический анализ.

Преимуществами данного метода являются:

- Возможность одновременно определять бактерии, археи, грибы и микроскопические эукариоты в одном образце без использования различных праймеров для амплификации целевого гена.
- Отсутствие этапа проведения полимеразной цепной реакции (ПЦР) с праймерами для целевого гена позволяет избежать смещения таксономического состава.
- Более точное определение таксономического состава сообщества благодаря присутствию в массивах сразу всех маркерных генов.
- Возможность функционального профилирования для определения метаболического потенциала сообщества.

- Возможность полных сборок геномов микроорганизмов, доминирующих в исследуемом сообществе.

Недостатками метода шотган-секвенирования являются:

- Дороговизна из-за трудоемкости метода, необходимости секвенирования с большей глубиной прочтения (на более дорогих высокопроизводительных секвенаторах);
- Требует значительных вычислительных мощностей;
- Требует бóльшего количества ДНК в исследуемом образце;
- Относительно небольшое количество накопленных данных для сравнения и обсуждения с литературой.

Биоинформатический анализ данных шотган-секвенирования микробиома состоит из нескольких принципиальных блоков – таксономическое профилирование, функциональное профилирование и сборка геномов из метагеномных прочтений.

Таксономическое профилирование производится путем картирования коротких прочтений на референсную базу маркерных генов (например, MetaPhlAn4, который имеет базу последовательностей маркерных генов для более чем 1 000 000 геномов 22 000 видов микроорганизмов).

Функциональное профилирование производится также путем картирования коротких прочтений, но на референсную базу генов с известной функцией. Далее эти гены могут быть объединены в метаболические пути, что позволяет оценить как представленность пути (сколько копий генов, составляющих этот путь присутствует в сообществе), так и его полноту (все ли ферменты метаболического пути присутствуют в сообществе). Это позволяет определить метаболический потенциал сообщества – какие источники углерода, азота, фосфора и т.д. могут использоваться, какие метаболиты продуцироваться, к каким антибиотикам устойчиво сообщество и др.

Метагеномная сборка позволяет обнаруживать и собирать геномы даже некультивируемых организмов. Этот анализ ни в каком виде не может быть произведен на данных секвенирования маркерных

генов. На первом этапе производится непосредственно сборка контигов и скаффолдов – объединенных в более длинные фрагменты ридов. Метагеномная сборка имеет несколько особенностей в отличие от классической сборки единичного генома:

- Глубина прочтения разных геномов в одном образце сильно варьирует из-за различий в представленности видов;
- Разные виды имеют схожие участки ДНК – консервативные гены;
- Каждый вид может быть представлен несколькими штаммами.

Следующим этапом метагеномной сборки является биннинг – группировка контигов в отдельные геномы на основе следующих принципов (рис. 2):

1. Оценки покрытия каждого контига (если у двух контигов одинаковое покрытие, то велика вероятность, что они из одного генома).
2. Частота встречаемости кодонов или тетрануклеотидов (разные виды бактерий с разной частотой используют кодоны для кодирования одной и той же аминокислоты – избыточность генетического кода).
3. Однокопийные маркерные гены не могут присутствовать в нескольких копиях в одном геноме (если один и тот же однокопийный ген присутствует в двух контигах, то эти контиги принадлежат разным геномам).

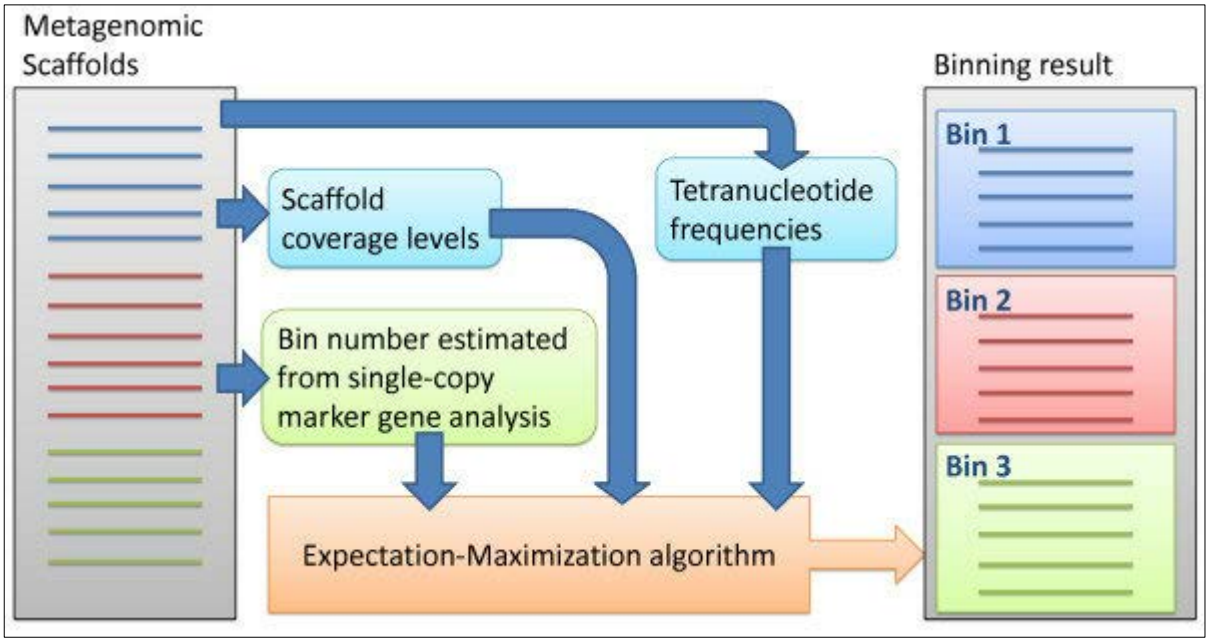

Рис. 2. Принципиальная схема биннинга [Wu *et al.*, 2014]

Далее проводится оценка полноты и контаминации собранных геномов. Полнота сборки оценивается по наличию всех маркерных генов, а контаминация – по их избыточности. Качественной принято считать сборку генома с полнотой более 80%, и с контаминацией менее 5%.

1.2. Секвенирование ампликонов целевых генов

Для первичной оценки таксономического состава микробного сообщества наиболее распространенным подходом является ампликоновое секвенирование. В литературе также можно встретить другие названия метода – метагенетика, метабаркодинг, метатаксономика, структурная метагеномика. Данный подход основан на амплификации и последующем прочтении нуклеотидной последовательности одного гена – филогенетического маркера. Маркерный ген должен соответствовать следующим критериям:

- Быть консервативным – иметь низкую скорость накопления мутаций;
- Быть однокопийным (желательно);
- Должен иметь участки, подходящие для подбора селективных праймеров, подходящих для большинства таксонов.

Наиболее популярные маркерные гены:

- 16S рРНК – самый популярный, используется для исследования бактериальных и архейных сообществ;
- 18S рРНК – используется для исследования сообществ микроскопических эукариот;
- ITS (internally transcribed spacer, внутренний транскрибируемый спейсер) – используется для исследования грибных сообществ;
- Гены хлоропластов (Rubisco) – для исследования фитопланктона;
- Гены митохондрий (цитохром оксидаза) – для исследования зоопланктона.

1.2.1. Ген 16S рРНК

Наиболее популярным маркерным геном является ген 16S рРНК. Его выбор обусловлен несколькими критериями: он присутствует в любом прокариотическом геноме; высококонсервативен, с одной стороны, и содержит гипервариабельные регионы, с другой, что делает его идеальным инструментом для таксономической идентификации – геном каждой бактерии отличается своей уникальной последовательностью вариабельных участков гена, и в то же время, используя универсальные праймеры, подобранные к консервативой части гена, эти маркерные последовательности легко амплифицировать из всех бактериальных геномов, содержащихся в образце.

Ген 16S рРНК имеет длину около 1500 п.о., и состоит из 10 консервативных и 9 вариабельных регионов (рис. 3).

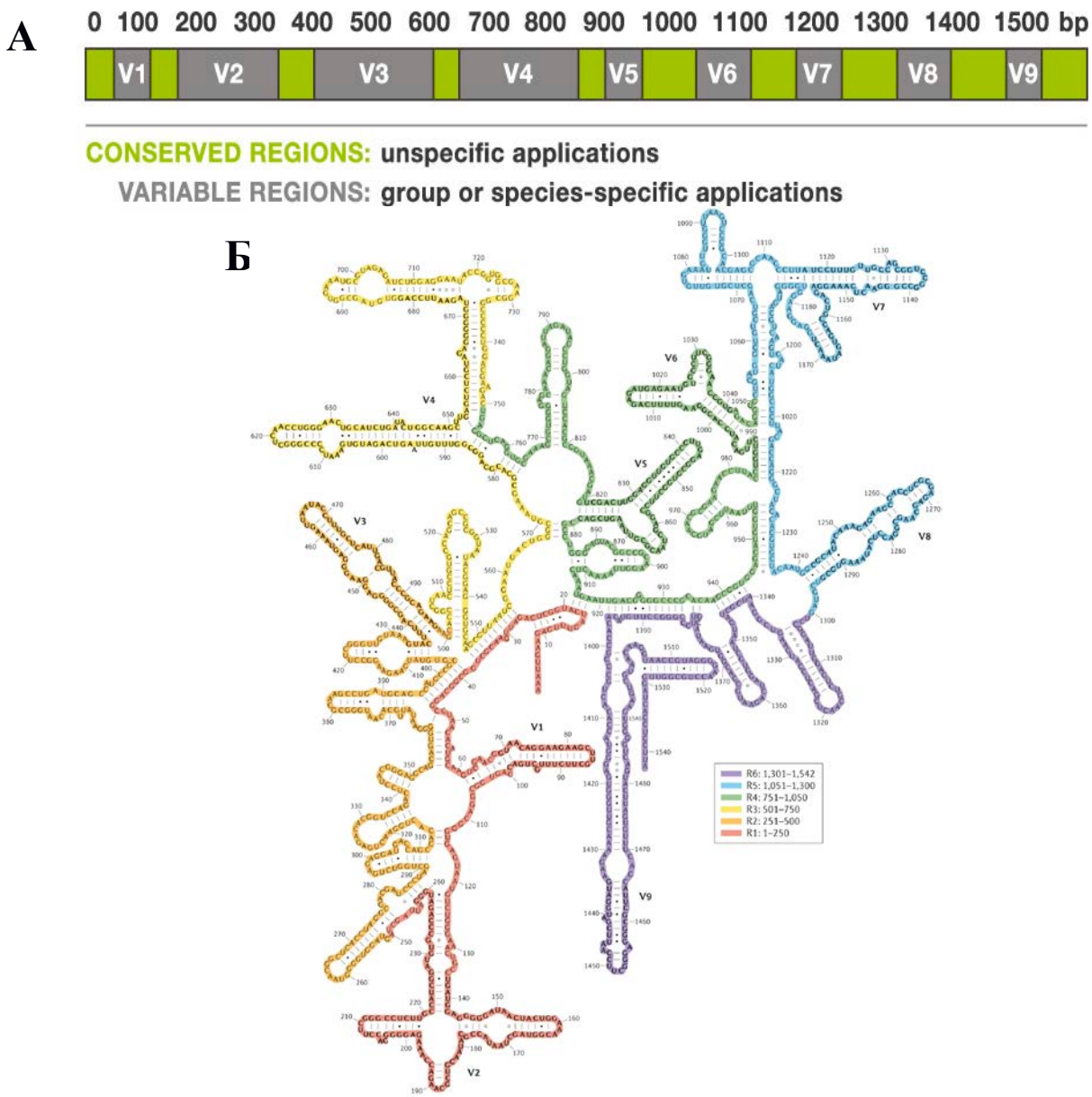

Рис. 3. Первичная (А) и вторичная (Б) структура 16S рРНК [Yarza *et al.*, 2014]

Длина и строение гена 16S рРНК позволяют секвенировать ампликоны различных вариабельных участков в зависимости от располагаемых платформ для секвенирования (рис. 4). Т.к. наиболее популярной платформой NGS является Illumina, то, соответственно, чаще всего для метагеномных исследований амплифицируются регионы V4 или V3-V4, но универсального набора участков не существует. По данным исследований, для изучения кишечной микробиоты лучше всего подходят регионы V4, V6 и V7 [Aloisio *et al.*, 2016], для микробиоты ротовой полости – V1-V3 и V7-V9 [Kumar *et al.*, 2011], для метагенома

почвы – V1-V3 [Soriano-Lerma *et al.*, 2020], арктических вод – V4, V5 [Fadeev *et al.*, 2021] и т.д. Но наиболее информативным методом секвенирования ампликонов гена 16S рРНК на сегодняшний момент является секвенирование всех 9 вариабельных участков (V1-V9) с использованием платформ секвенирования третьего поколения (Oxford Nanopore Technologies, Pacific Bioscience), с помощью которых можно прочитать длинные фрагменты ДНК.

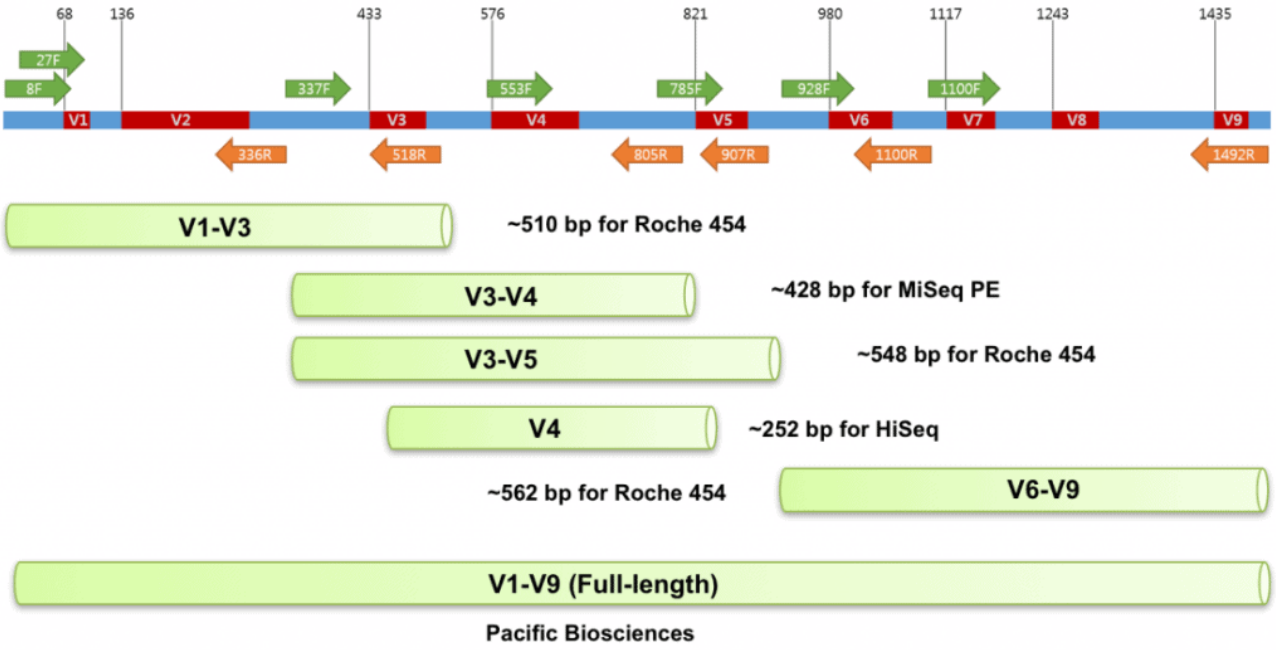

Рис. 4. Вариабельные регионы гена 16S рРНК, использующиеся для секвенирования на различных платформах (https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/)

1.2.2. Основные этапы метагеномного анализа на основе ампликонов гена 16S рРНК

Метагеномное исследование на основе ампликонов гена 16S рРНК в общем случае включает в себя следующие основные этапы:

1. Выделение ДНК из образца, оценка концентрации ДНК методом флуорометрии;
2. Амплификация целевого фрагмента (участка маркерного гена), контроль качества ПЦР-продукта методом электрофореза в агарозном геле, очистка ПЦР-продукта;

3. Индексирование (баркодирование) ПЦР-продуктов методом амплификации: «пришивание» адаптеров и уникальных последовательностей (баркодов или индексов) для отдельных образцов, если планируется секвенировать сразу несколько образцов;

4. Очистка и контроль качества полученных библиотек: оценка концентрации ДНК методом флуорометрии и оценка длины библиотек методом капиллярного электрофореза;

5. Смешивание в эквимолярном соотношении баркодированных библиотек;

6. Секвенирование библиотек, получение нуклеотидных последовательностей – ридов;

7. Биоинформатический анализ данных секвенирования, интерпретация результатов.

Биоинформатический анализ данных секвенирования ампликонов состоит из следующих основных этапов:

1. Контроль качества прочтений

Данный этап анализа актуален для анализа любых данных секвенирования, для шотган-секвенирования микробиома в том числе. Метагеномное NGS-секвенирование дает на выходе множество коротких нуклеотидных прочтений (до 300 п.о.). Эти прочтения хранятся в текстовых файлах формата FASTQ, по паре или по одному файлу на образец, в зависимости от режима секвенирования – парноконцевое (paired-end) или одноконцевое (single-end). Запись прочтений в формате FASTQ характеризуется тем, что, помимо собственно нуклеотидных последовательностей, несет в себе показатели качества каждого прочитанного нуклеотида, которое может быть оценено с помощью специализированных программ, таких как FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc), FastX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) и др. Следуя биоинформатическому принципу “Garbage in, garbage out” («мусор на входе – мусор на выходе», при некачественных исходных данных будут получены некачественные результаты), «сырые» файлы необходимо избавить от ридов низкого качества:

- содержащих остатки служебных последовательностей – индексов, адаптеров и т. п.;
- имеющих неопределенные нуклеотиды (обозначенные как N);
- слишком коротких;
- имеющих нуклеотиды с неудовлетворительным параметром качества (см. п. 2.4.).

Для этих целей можно воспользоваться программой для тримминга сырых ридов, например, cutadapt [Martin *et al.*, 2011] или Trimmomatic [Bolger *et al.*, 2014].

Также следует обратить внимание на общее число прочтений на образец. Стандартного порогового значения не существует, так как оно зависит от цели метагеномного исследования и типа образца, но необходимо понимать, что чем больше ридов получено для образца, тем более полной будет картина профилирования и тем более малопредставленные организмы удастся обнаружить. В целом, для ампликонового секвенирования число высококачественных ридов в образце от 10 000 считается приемлемым.

2. Объединение парных ридов в единый фрагмент

Как упоминалось выше, наиболее популярной платформой NGS для метагеномных исследований является Illumina MiSeq, и чаще всего амплифицируются вариабельные участки V3-V4 гена 16S рРНК в режиме парноконцевого чтения 2*250 п.о. или 2*300 п.о. Длина данного участка в среднем составляет около 460 п.о., таким образом при секвенировании имеется небольшое перекрытие – участок в середине фрагмента секвенируется дважды, что позволяет объединить парные риды в единый фрагмент (рис. 5). После обрезки по качеству риды могут стать короче, поэтому важно учесть, что около 20-30 п.о. необходимы для объединения ридов.

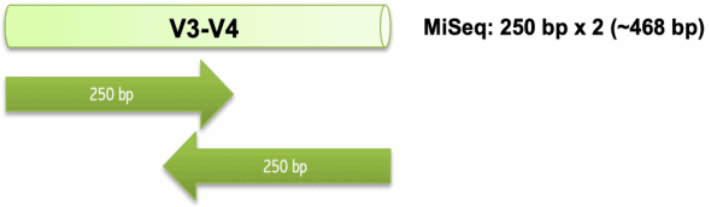

Рис. 5. Перекрывание прямого и обратного рида при секвенировании региона V3-V4 гена 16S рРНК на платформе Illumina MiSeq (https://help.ezbiocloud.net/16s-rrna-and-16s-rrna-gene/)

3. Поиск и фильтрация химерных фрагментов

На следующем шаге необходимо удалить химерные последовательности – артефактные последовательности, которые ошибочно создаются в процессе ПЦР-амплификации из двух или более цепей-шаблонов вместо одной родительской цепи (рис. 6). Таким последовательностям будет невозможно найти однозначное соответствие среди сиквенсов базы данных, поэтому их следует отфильтровать. Химерные фрагменты обнаруживаются путем выравнивания объединенных фрагментов на референсную базу данных и удаляются из анализа. Другой подход – фрагменты выравниваются друг с другом и аномальные химерные фрагменты исключаются из анализа.

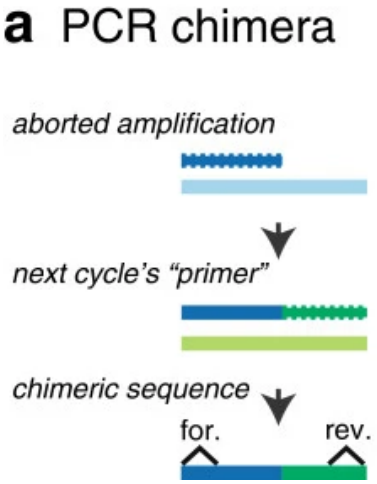

Рис. 6. Механизм образования химерных прочтений [Fichot *et al.*, 2013]

4. Выявление OTU или ASV и определение их таксономической принадлежности

Данный шаг является основным этапом биоинформатического анализа данных ампликонового секвенирования. Фрагменты ДНК сравниваются друг с другом и общие кластеры образуют OTU (operational taxonomic units, операционные таксономические единицы) или ASV (amplicon sequence variants, варианты последовательности ампликона).

OTU – это кластер схожих последовательностей. Традиционно за порог сходства принято значение 97% как эвристическая оценка сходства последовательностей гена 16S рРНК внутри одного бактериального вида.

Существует три основных пути идентификации таксонов в образце:

- Основанный на закрытом референсе (closed-reference). Выравнивание каждого сиквенса образца с референсной базой производятся напрямую, минуя этап кластеризации OTU. Сиквенсы, не похожие на референсные, отбрасываются. Данный подход используется крайне редко, применим для образцов из хорошо изученных сред, для которых создан специализированный референсный каталог последовательностей микроорганизмов – например, для микробиома кишечника.
- Основанный на открытом референсе (open-reference). Гибридный подход, при котором последовательности сначала выравниваются на референс, а затем невыровненные сиквенсы группируются в OTU с последующей таксономической классификацией. Подход отличается высокой скоростью за счет первого этапа и более полной микробной идентификацией за счет второго.
- Основанный на поиске de novo. Последовательности объединяются в кластеры OTU, представленность OTU принимается равной числу последовательностей в кластере. Далее с помощью инструментов классификации, наиболее распространенным среди

которых является RDP Classifier [Wang *et al.*, 2007], кластерам присваиваются таксоны. Преимущество подхода в отсутствии необходимости в референсной базе.

Современные подходы к анализу данных отходят от использования OTU, т.к. появились алгоритмы, которые позволяют увеличить разрешающую способность метода путем выявления ASV. Такой подход позволяет оценить вероятность случайной ошибочной вставки нуклеотида при постановке ПЦР или ошибочного прочтения при секвенировании, и исправить данные ошибки, кластеризуя фрагменты в фактические последовательности – ASV, а не в консенсус из облака схожих на 97% фрагментов (OTU). Это позволяет анализировать последовательности фрагментов, которые действительно присутствуют в образце, что может быть полезно при сравнении результатов разных исследований. Поиск ASV основан на методе de novo, т.е. фрагменты сначала сравниваются друг с другом, а затем им присваивается таксономическая принадлежность на основе баз данных. Также данный подход позволяет различить многие виды бактерий, ген 16S рРНК которых схож более чем на 97%, например, использование ASV позволяет отделить *Neisseria gonorrhoeae* от остальных видов *Neisseria*.

5. Анализ альфа-разнообразия

Альфа-разнообразие – показатель, применимый ко всему сообществу исследуемого образца, характеризует богатство и разнообразие сообщества. Для оценки альфа-разнообразия разработано множество метрик, среди которых популярны следующие:

- Chao1 – оценка скрытого видового богатства, опирающаяся при расчете на экстраполяцию редких OTU или ASV;
- Индекс Шеннона – оценка богатства и равномерности представленности видов;
- Индекс Симпсона также учитывает как число видов, так и их равномерность, но менее чувствителен к редким видам, чем индекс Шеннона;

- Филогенетический индекс (Phylogenetic Distance, PD) – доля филогенетического дерева жизни, покрываемого сообществом. Чем дальше удалены друг от друга обнаруженные таксоны на дереве, тем выше индекс.

6. Анализ бета-разнообразия

Бета-разнообразие – показатель, применимый к нескольким сообществам сразу, характеризует меру различия сообществ. Обычно бета-разнообразие рассчитывается как расстояние между исследуемыми образцами сообществ с помощью метрик Unifrac, расстояния Брея-Кёртиса, Жаккарда. Расстояние между образцами варьирует от 0 до 1, где 0 в общем смысле означает полностью идентичные сообщества по наличию OTU или ASV и их представленности, а 1 – полностью различающиеся сообщества. Так, одной из распространенных метрик является UniFrac (от англ. unique fraction – уникальная доля). Для пары образцов строится общее филогенетическое дерево их сообществ и определяется доля таксонов (ветвей) на дереве, присутствующих в обоих образцах. Если общих ветвей нет, расстояние UniFrac максимально и равно 1. Если образцы не отличаются по микробному составу, UniFrac равен 0. Половина общих микроорганизмов даст значение UniFrac 0,5, и т.д. При расчете невзвешенного расстояния (unweighted UniFrac) учитывается только наличие и отсутствие таксонов (качественные отличия), при расчете взвешенного (weighted UniFrac) – их количественная представленность. Данные бета-разнообразия принято визуализировать с помощью алгоритмов понижения размерности NMDS или PCoA (см. п. 2.6.).

7. Статистический анализ

Последующий анализ зависит от целей метагеномного исследования и может включать поиск ассоциаций между найденными видами и характеристиками образца, поиск значимых отличий в микробном составе между сообществами, оценку метаболического потенциала на основании найденных видов и т.д.

1.2.3. Базы данных референсных последовательностей гена 16S рРНК и популярные программы для анализа данных

Важным этапом анализа метагеномных данных является выравнивание прочтений на референсную базу данных, содержащую известные последовательности маркерного гена различных таксономических групп. Среди наиболее популярных баз, содержащих последовательности гена 16S рРНК – Greengenes [DeSantis *et al.*, 2006; McDonald *et al.*, 2024] и SILVA [Quast *et al.*, 2013].

Greengenes – это курируемая база полных последовательностей гена 16S рРНК. Долгое время ее последней версией являлась 13.8 от 2013 г., но в 2022 г. произошло глобальное обновление до версии Greengenes2, которая включает в себя около 330 тыс. полных последовательностей гена 16S рРНК. Каталог SILVA включает в себя полные последовательности не только гена 16S, но и 18S, 23S/28S для анализа эукариот, число последовательностей превышает 500 тыс. Версии данной базы обновляются на регулярной основе, последней актуальной версией является SILVA v.138.2 от 2024 г.

Наиболее популярными программами для анализа данных ампликонового секвенирования являются QIIME (1 и 2 версия) [Caporaso *et al.*, 2010], Mothur [Schloss *et al.*, 2009], Kraken2 [Lu, Salzberg, 2020]. На сегодняшний момент одной из самых быстрых и аккуратных программ является QIIME2, которая реализует в себе алгоритм DADA2 [Callahan *et al.*, 2016] для выявления и подсчета ASV. Все перечисленные программы требуют значительных вычислительных мощностей, не имеют графического интерфейса, что делает анализ данных трудоемким занятием для биологов, не обладающих биоинформатическими навыками.

Альтернативным вариантом анализа данных является использование web-платформ. Так, EzBioCloud (https://www.ezbiocloud.net) – один из наиболее популярных ресурсов, позволяющих проанализировать данные на удаленном сервере, не используя собственные вычислительные ресурсы и не применяя

навыков работы в командной строке. Преимуществом является одноэтапная загрузка данных, автоматический контроль качества, поиск OTU и визуализация результатов высокого качества. Недостатками является невозможность одновременной загрузки более чем 100 образцов, а также отсутствие функции выгрузки результатов анализа для всех образцов в одну таблицу. Кроме того, ресурс не располагает возможностью проведения статистического анализа для выявления отличий между исследуемыми группами сравнения.

Практически все недостатки, которыми обладает EzBioCloud, отсутствуют на web-платформе MicrobiomeAnalyst (https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/home.xhtml). Это ресурс, позволяющий анализировать как данные ампликонового секвенирования и шотган-секвенирования, так и производить интеграцию с метаболомными данными. Кроме того, MicrobiomeAnalyst позволяет проводить статистический анализ для сравнения микробиомов образцов из разных групп сравнения, учитывая имеющиеся метаданные, а также производить корреляционный анализ, кластеризацию, классификацию образцов для выявления биомаркеров с применением методов машинного обучения. Также данная web-платформа позволяет визуализировать полученные результаты, создавая графики, пригодные для публикации в научных журналах. Web-платформа MicrobiomeAnalyst имеет гибкие фильтры на каждом этапе, позволяющие проводить анализ в четком соответствии с видением исследователя.

Недостатками web-ресурса MicrobiomeAnalyst является невозможность проанализировать более чем 100 образцов ампликоновых метагеномов, и не имеет возможности полного анализа данных шотган-секвенирования от сырых прочтений.

Настоящее учебное пособие содержит в себе описание полного анализа данных секвенирования ампликонов V3-V4 региона гена 16S рРНК с от сырых прочтений через выявление ASV к статистическому анализу и поиску биомаркеров с помощью машинного обучения с использованием web-платформы MicrobiomeAnalyst.

ОТ СЫРЫХ ПРОЧТЕНИЙ К ASV

2.1. Загрузка данных учебного проекта на персональный компьютер

Учебный проект представляет собой данные секвенирования образцов микробиоты кишечника 3 людей с воспалительными заболеваниями кишечника (ВЗК, inflammatory bowel disease, IBD) и 3 здоровых добровольцев (группа контроля, CTRL). Было произведено секвенирование ампликонов V3-V4 региона гена 16S рРНК на приборе Illumina Miseq в режиме парноконцевого прочтения (paired-end) 2*300 п.о.

Сырые прочтения доступны для скачивания по ссылке https://disk.yandex.ru/d/-01fg1aRroRW1Q. Необходимый объем на диске– около 300 Мб.

Извлеките файлы из архива. В папке находятся 12 файлов в формате fastq.gz. Формат fastq служит для записи прочтений, полученных после секвенирования. Подробнее про данный формат можно прочитать на сайте производителя наиболее распространенных секвенаторов Illumina (https://knowledge.illumina.com/software/general/software-general-reference_material-list/000002211). Формат gz – gzip-архив, который служит для сжатия файлов. Обратите внимание на имена файлов. Суффикс R1 в названии указывает, что это прямой рид (forward read), а суффикс R2 – обратный (reverse read). Таким образом, наш учебный проект составляют 6 биологических образцов (по 2 файла fastq.gz на каждый (R1+R2)).

В папке после разархивирования находится также файл metadata.txt, в котором отражена схема эксперимента – каждому файлу с ридами соответствует группа сравнения, к которой он относится (метаданные).

2.2. Загрузка ридов на web-платформу MicrobiomeAnalyst

Откройте в браузере (Opera или Firefox) сайт MicrobiomeAnalyst (https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/home.xhtml). На

главной странице сайта представлены обновления и кратко охарактеризованы программные модули. Кликните на кнопку “Click here to start” (рис. 7).

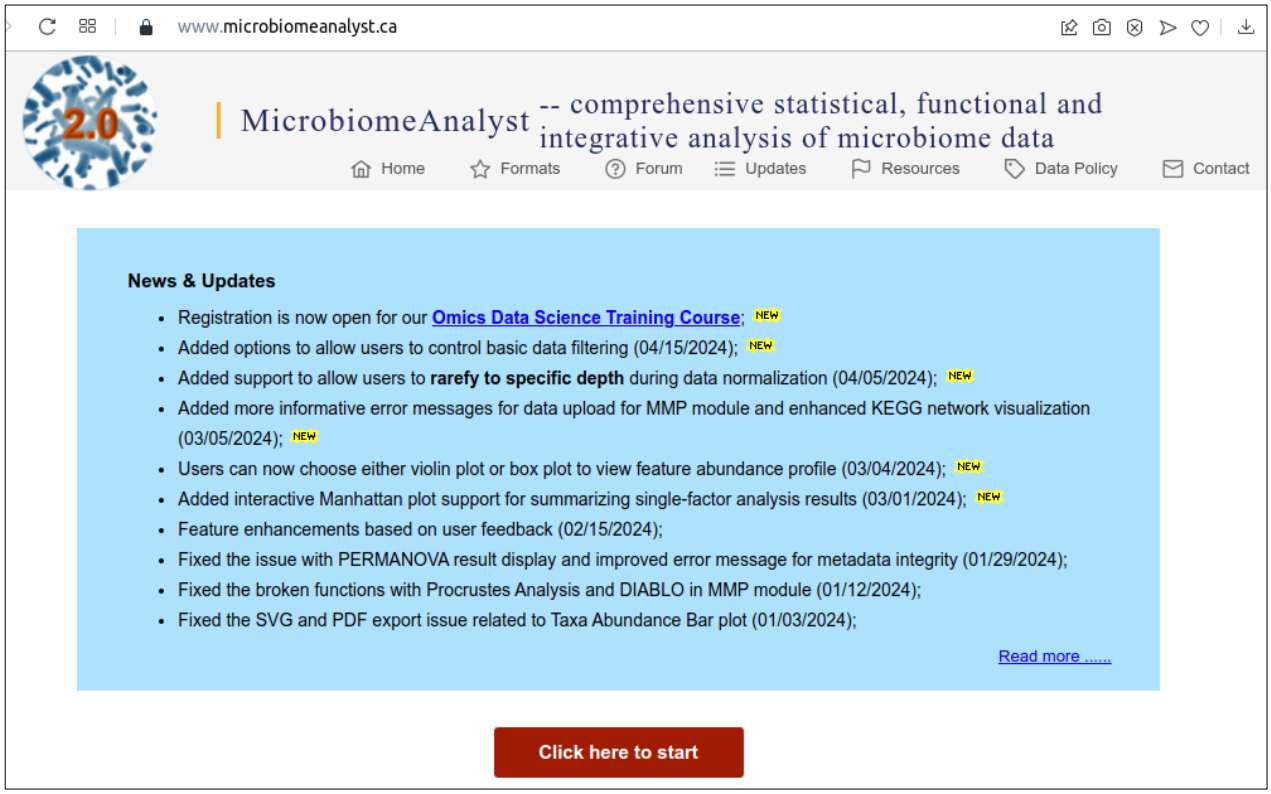

Рис. 7. Главная страница web-платформы MicrobiomeAnalyst

На вновь открывшейся странице выберете модуль Raw Data Processing (рис. 8), который необходим для первичной обработки сырых прочтений.

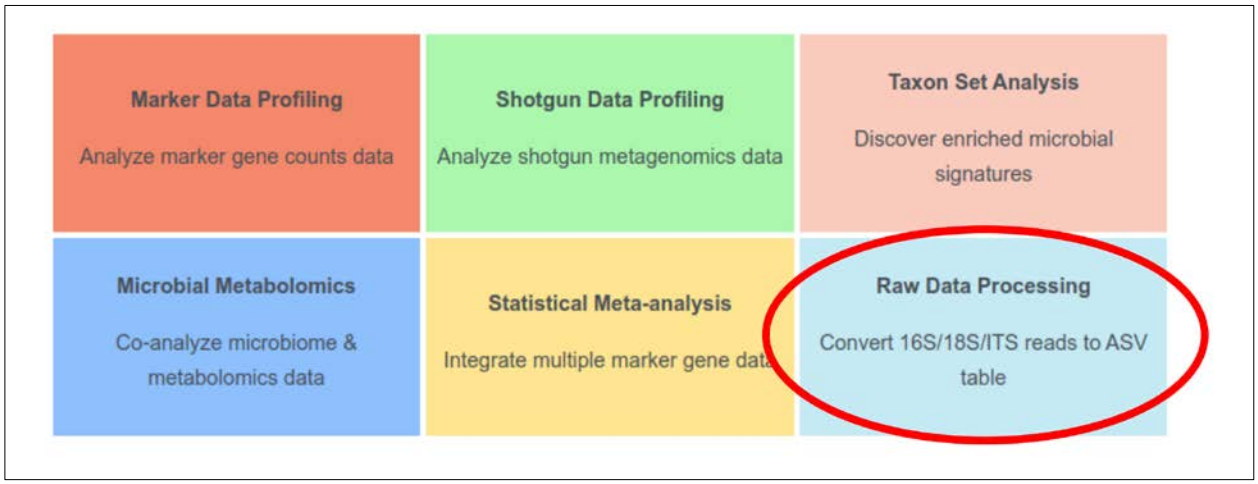

Рис. 8. Страница выбора модуля “Raw Data Processing” web-платформы MicrobiomeAnalyst

Загрузка ридов на web-платформу MicrobiomeAnalyst.

Данный модуль MicrobiomeAnalyst реализует анализ данных секвенирования ампликонов маркерных генов для оценки таксономического состава сообщества с помощью программного пакета DADA2. DADA2 позволяет выполнить профилирование сырых прочтений в таблицы, содержащие данные о представленности ASV в каждом образце. Подобный анализ ампликонов может быть реализован как непосредственно с помощью DADA2 в среде языка программирования R, а также с помощью программного пакета QIIME2 и web-платформы MicrobiomeAnalyst.

MicrobiomeAnalyst требует загрузки файлов прочтений, а также метаданных с описанием эксперимента (Рис. 9). С помощью модуля можно проанализировать данные секвенирования ампликонов генов 16S рРНК, 18S рРНК и ITS, как в режиме парноконцевого (paired-end), так и одноконцевого прочтения (single-end). Имеются ограничения как по размеру загружаемого файла (не больше 100 Мб на 1 файл), так и по количеству файлов (не более 100). Обратите внимание, что парные риды должны называться определенным образом – *_R1.fastq/*_R2.fastq.

Для загрузки файлов нажмите кнопку “Select” и выберете все 12 скачанных файлов прочтений и файл metadata.txt из папки на персональном компьютере (рис. 9).

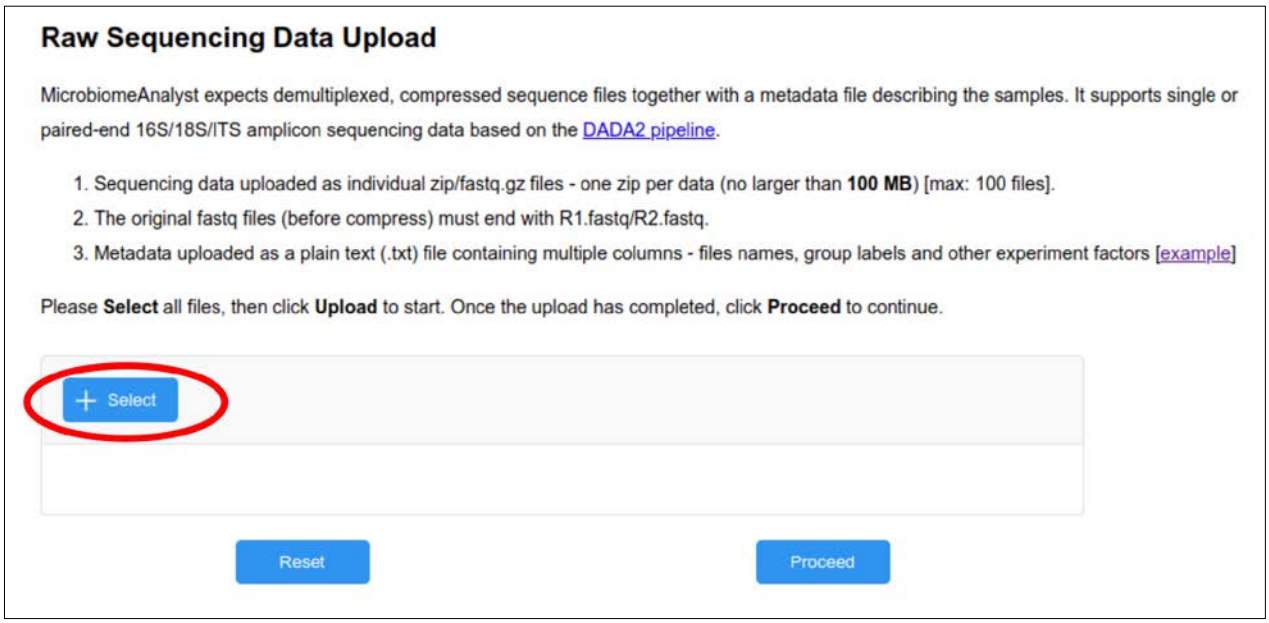

Рис. 9. Страница загрузки файлов сырых прочтений web-платформы MicrobiomeAnalyst

Далее нажмите кнопку “Upload” (рис. 10) и дождитесь полной загрузки всех файлов. Затем нажмите кнопку “Proceed” (рис. 11).

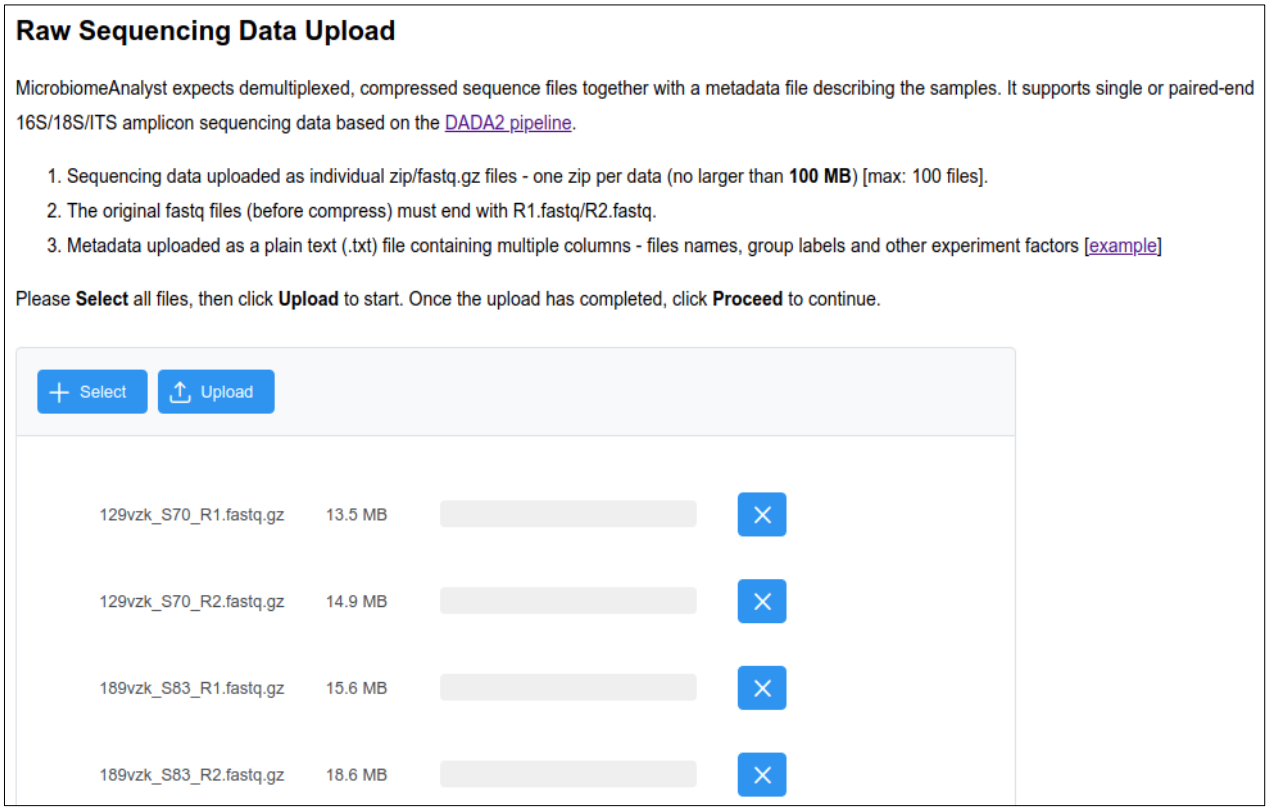

Рис. 10. Страница загрузки файлов сырых прочтений web-платформы MicrobiomeAnalyst после выбора файлов

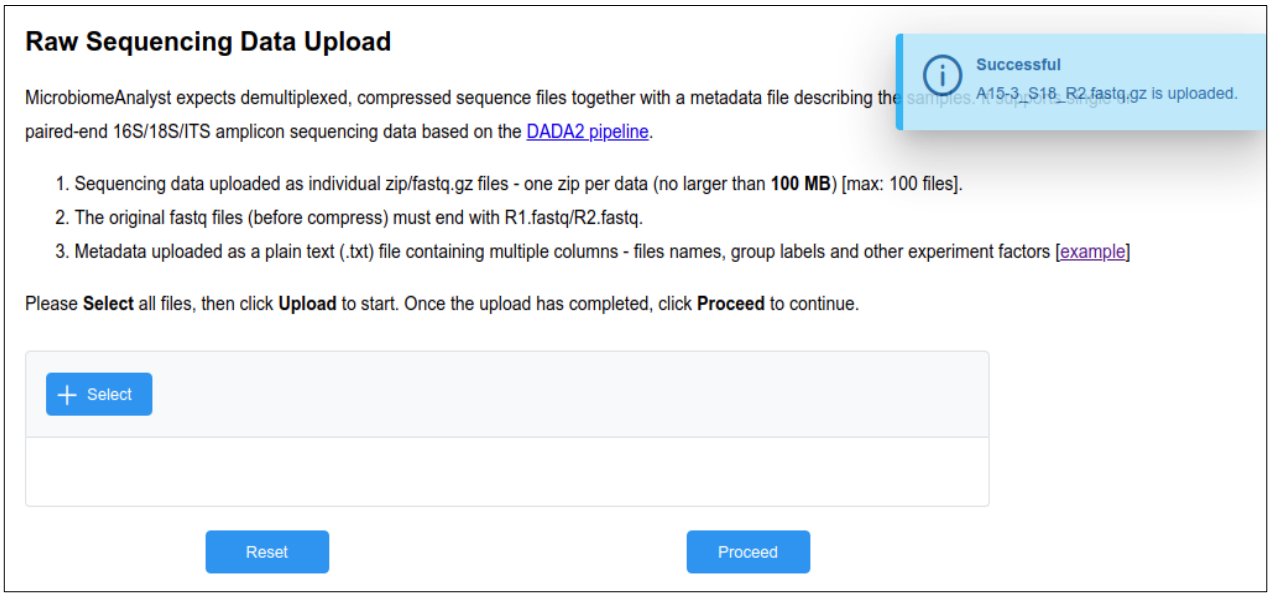

Рис. 11. Страница загрузки файлов сырых прочтений web-платформы MicrobiomeAnalyst после успешной загрузки всех файлов

2.3. Подготовка данных к анализу

После нажатия кнопки “Proceed”, возникнет всплывающее окно для подтверждения типа прочтений (парноконцевые (paired-end) или одноконцевые (single-end)). В случае выполнения учебного проекта кликните на кнопку “Yes” при возникшем вопросе “We noticed your data is paired-end, right?” (рис. 12).

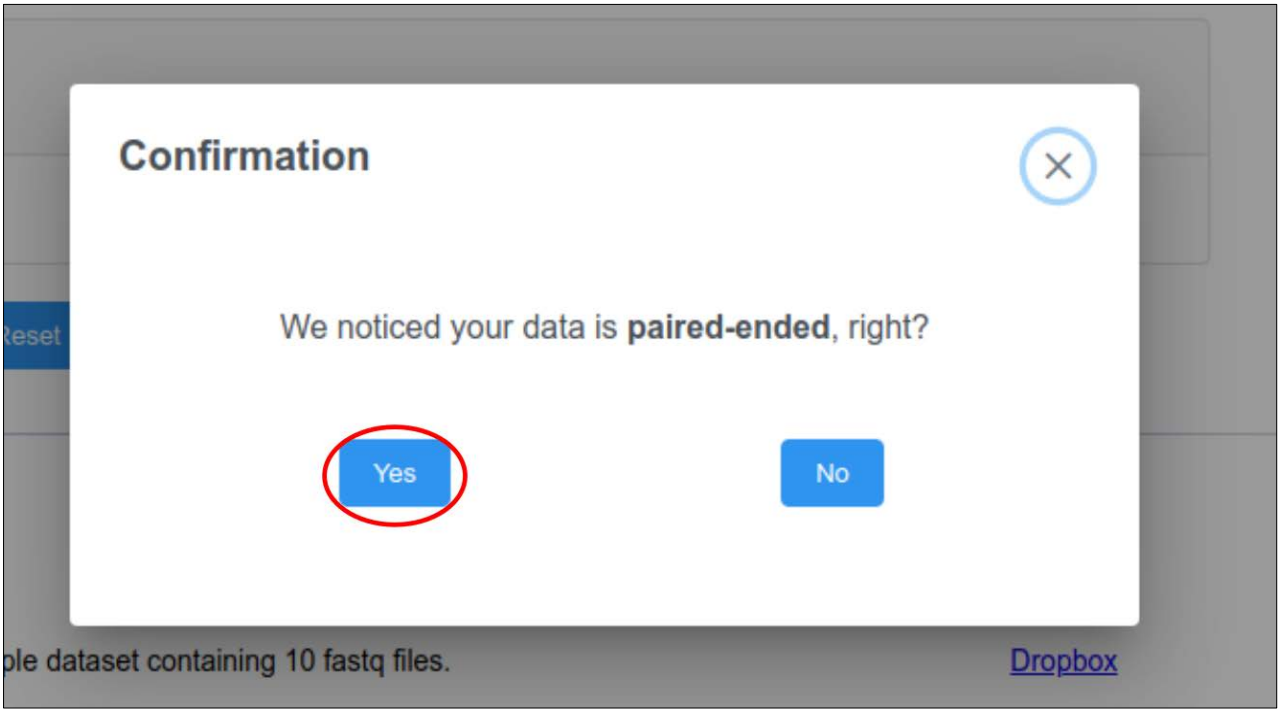

Рис. 12. Всплывающее окно выбора типа прочтений

Далее на следующей странице проверьте целостность загруженных данных (рис. 13):

- Имена прямого и обратного рида верно определились для каждого образца (написаны в одной строке).
- Количество прямых и обратных ридов в каждом образце совпадает (колонка “Reads”).
- Верно определилась группа сравнения (колонка “Group”) в соответствии с метаданными в файле metadata.txt.
- Прочтения валидны – файлы прочитались, они не повреждены (колонка “Valid”). Если все условия соблюдены, нажмите на кнопку “Proceed”.

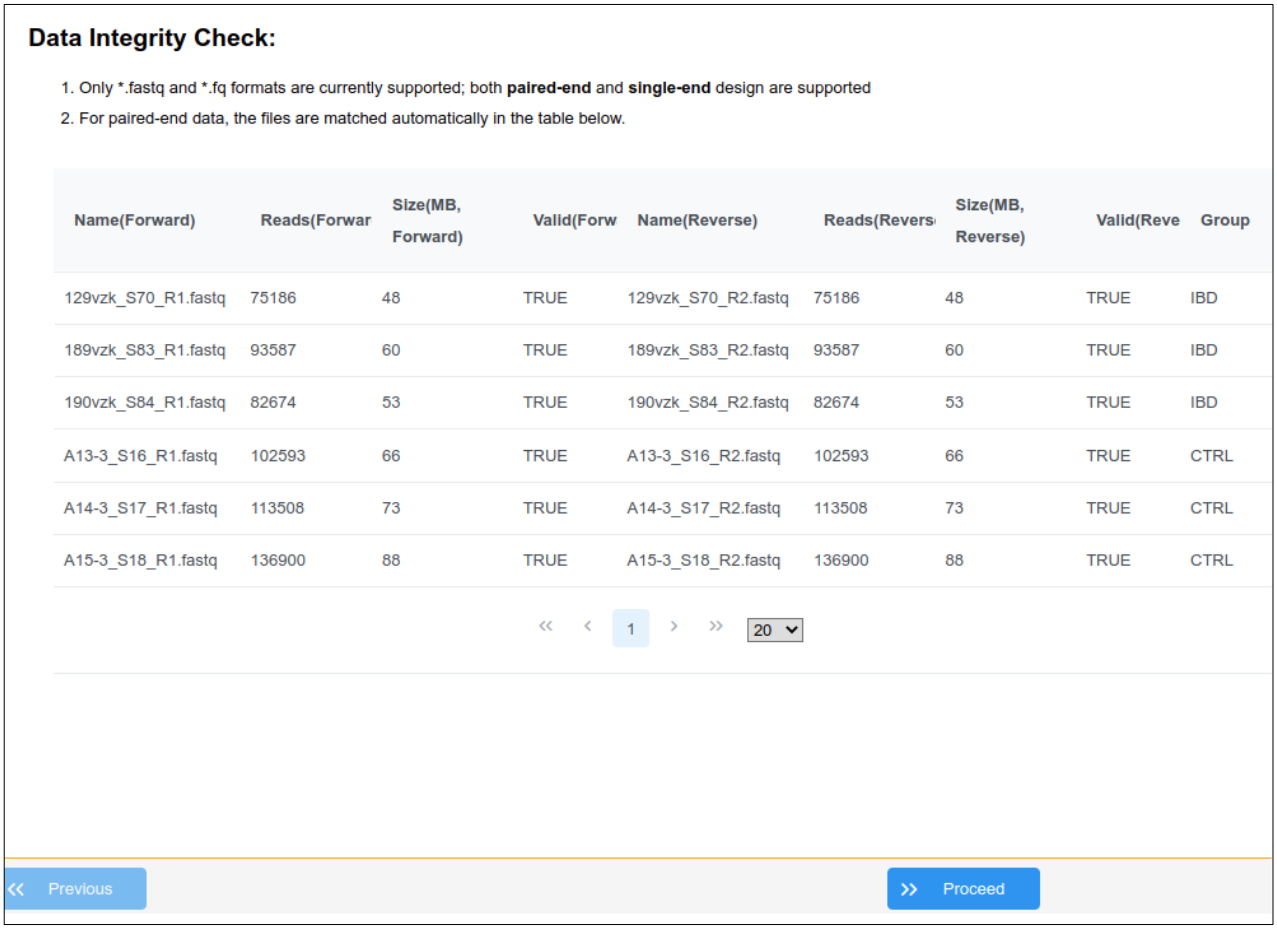

Рис. 13. Страница проверки целостности загруженных данных

2.4. Выбор параметров анализа

На странице с выбором параметров анализа также представлены результаты оценки качества прочтений двух образцов (рис. 14). Это необходимо для определения длины рида, которую нужно оставить после обрезки по качеству. Качество прочтения оценивается по шкале Phred от 0 до 40, где каждому значению соответствует вероятность ошибочного определения нуклеотида при секвенировании (табл. 1). Обратите внимание, что качество секвенирования снижается к концу прочтения (рис. 14), что является стандартной ситуацией при использовании методов NGS, причем качество обратного рида снижается заметно раньше, чем качество прямого рида. Качество прочтения ниже 30 принято считать неудовлетворительным, поэтому необходимо оставить длину ридов, где качество в среднем не будет ниже 30. Кроме того, нужно оставить запас длины для перекрывания парных ридов (около 20 п.о.), чтобы получить единый фрагмент участка маркерного гена. В учебном проекте были секвенированы ампликоны V3-V4 региона гена 16S рРНК, длина которого составляет около 460 п.о., таким образом, исходя из оценки качества, представленной на рисунке 14, нам достаточно будет длины 260 п.о. для прямых ридов и 220 п.о. – для обратных.

Таблица 1

Соответствие значения качества Phred точности определения нуклеотида при секвенировании

Phred Quality	Вероятность неправильного определения нуклеотида	Точность определения нуклеотида
0	1.0000	0.00%
5	0.3162	68.38%
10	0.1000	90.00%
15	0.0316	96.84%

Продолжение табл. 1

20	0.0100	99.00%
25	0.0032	99.68%
30	0.0010	99.90%
35	0.0003	99.97%
40	0.0001	99.99%

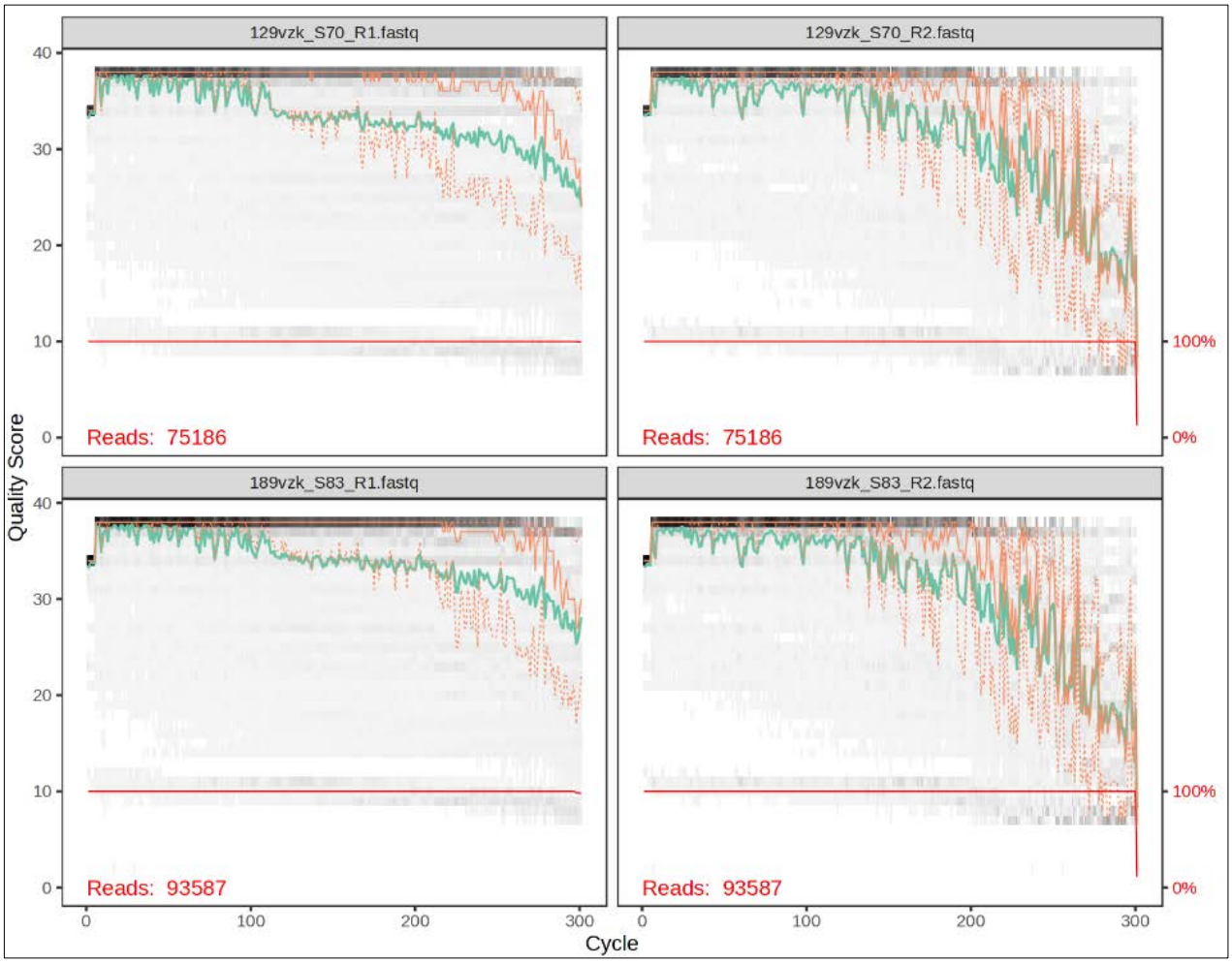

Рис. 14. Оценка качества секвенирования

Таким образом, необходимо заполнить параметры анализа, указанные на рисунке 15:

- Sequence type – 16S (в случае выполнения учебного проекта)
- Forward trunc length – 260 (см. описание выше)
- Reverse trunc length – 220

• Max EE of Forward и Max EE of Reverse оставляем без изменения

• Sequence Trimmer – TrimLeft:30 and TrimRight: 30 (позволяем обрезать также начало рида вплоть до 30 п.о. при низком качестве)

• MaxN, Min Q, Trunc Q, Remove PhiX оставляем без изменения

• Taxonomy reference databases – можно выбрать любую из доступных, отдавая предпочтение Silva (138.1) или Greengenes (13.8). Данные актуальны на 2024 год, в дальнейшем на сайте могут появиться новые, более актуальные базы данных.

После выбора всех параметров кликаем на кнопку “Submit” внизу страницы.

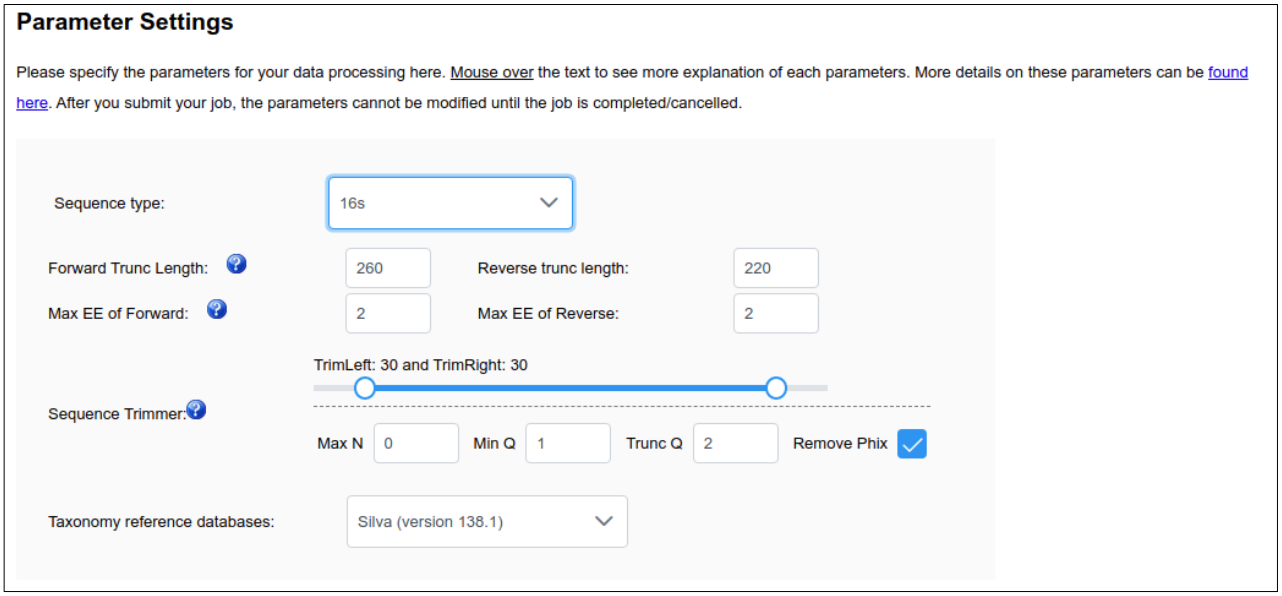

Рис. 15. Выбор параметров для анализа

2.5. Анализ данных

Начинается самый длительный процесс непосредственно анализа данных. Обычно анализ длится около часа, но иногда не происходит обновления страницы. Поэтому сразу нажмите на текст “Create Job URL”, скопируйте и сохраните ссылку на результаты вашего анализа (Рис. 16). Результаты хранятся на сервере MicrobiomeAnalyst в течение двух недель. На данном этапе происходят следующие этапы анализа:

- Контроль и обрезка ридов по качеству секвенирования

- Сравнение последовательностей всех ридов и объединение по их последовательностям в группы
- Поиск ошибок при ПЦР
- Определение конкретной последовательности группы ридов как ASV в каждом образце (прямой рид и обратный отдельно)
- Объединение парных ридов в единый фрагмент
- Удаление химерных фрагментов
- Формирование общей таблицы с информацией о представленности каждого ASV в исследуемых образцах
- Таксономическая аннотация выявленных ASV

Дождавшись успешного окончания анализа, либо перейдя по сохраненной ссылке, кликните кнопку “Proceed” (рис. 16).

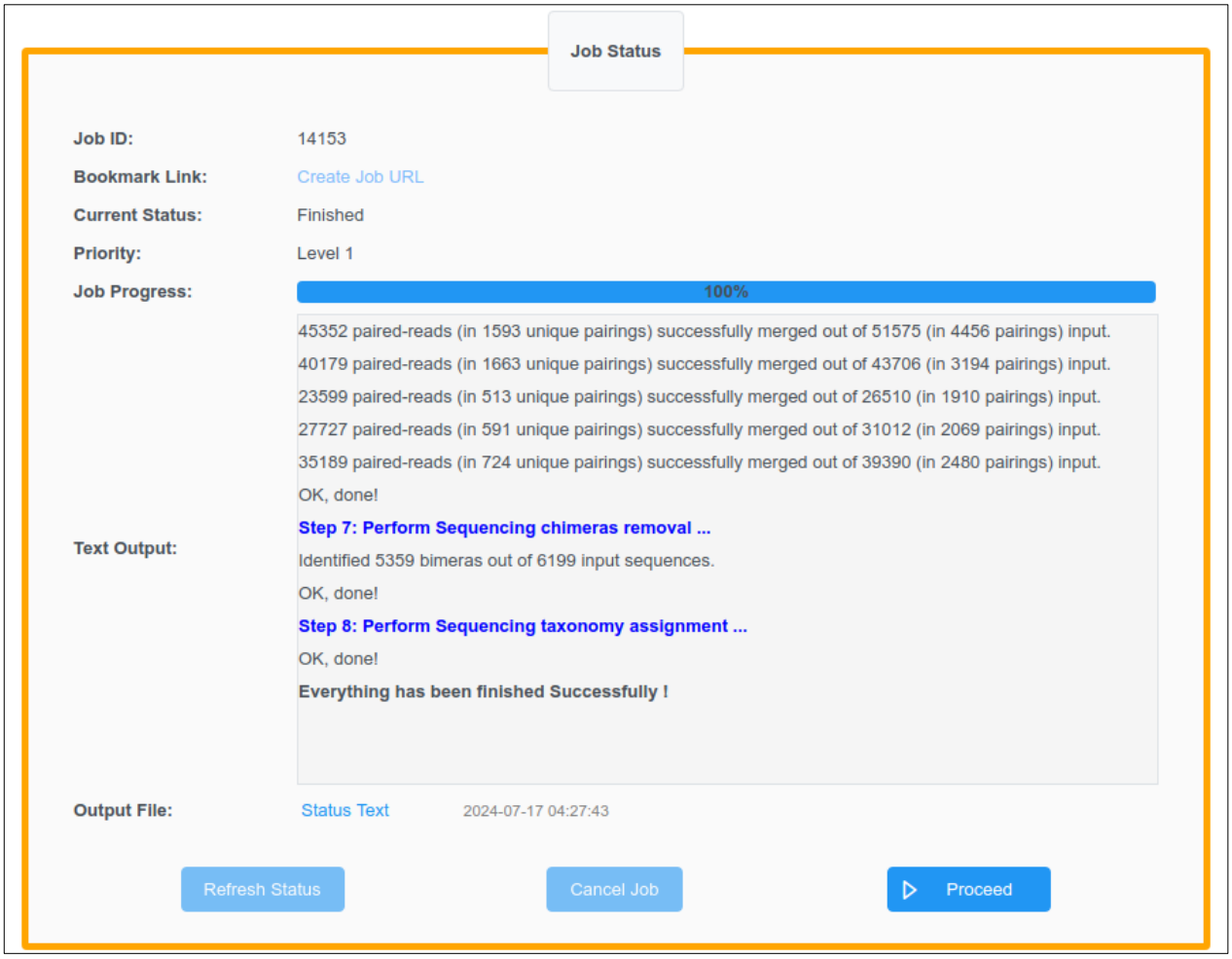

Рис. 16. Окно успешного завершения анализа

2.6. Результаты анализа

1. Финальное количество ридов в исследуемых образцах

Открывшееся окно представляет собой отчет о прошедшем анализе (рис. 17). Приводится общая информация о количестве обработанных образцов, выявленных ASV (несмотря на то, что на сайте написано OTU, это ASV, 775 на рисунке 17), а также количестве ридов, прошедших все фильтры анализа. Кроме того, размер финальных библиотек для каждого образца отражен на графике (Рис. 17) и в виде подробной таблицы (Рис. 18). Принято считать, что для надежного анализа таксономии образцов микробиоты кишечника человека достаточно 10000 прочтений на образец. Образцы из учебного проекта превысили данное количество, т.к. минимальный размер финальной библиотеки составляет более 25000 ридов (см. рис. 17 и столбец “NonChim” на рис. 18). Обратите внимание, что для анализа таксономического состава других биологических образцов требуется разное количество прочтений, так, чем потенциально более богато сообщество, тем больше ридов требуется секвенировать. Например, для микробных сообществ почв следует добиваться не менее 20000 ридов на образец, прошедших все этапы анализа.

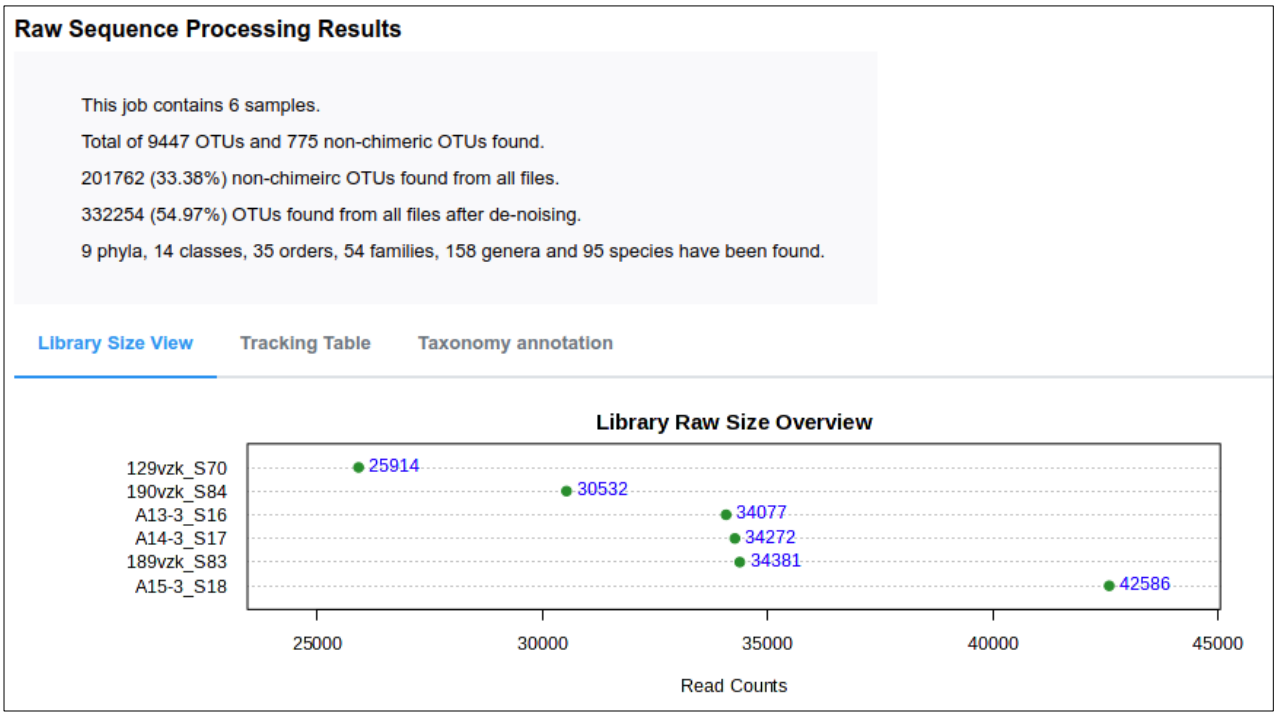

Рис. 17. Отчет о результатах анализа

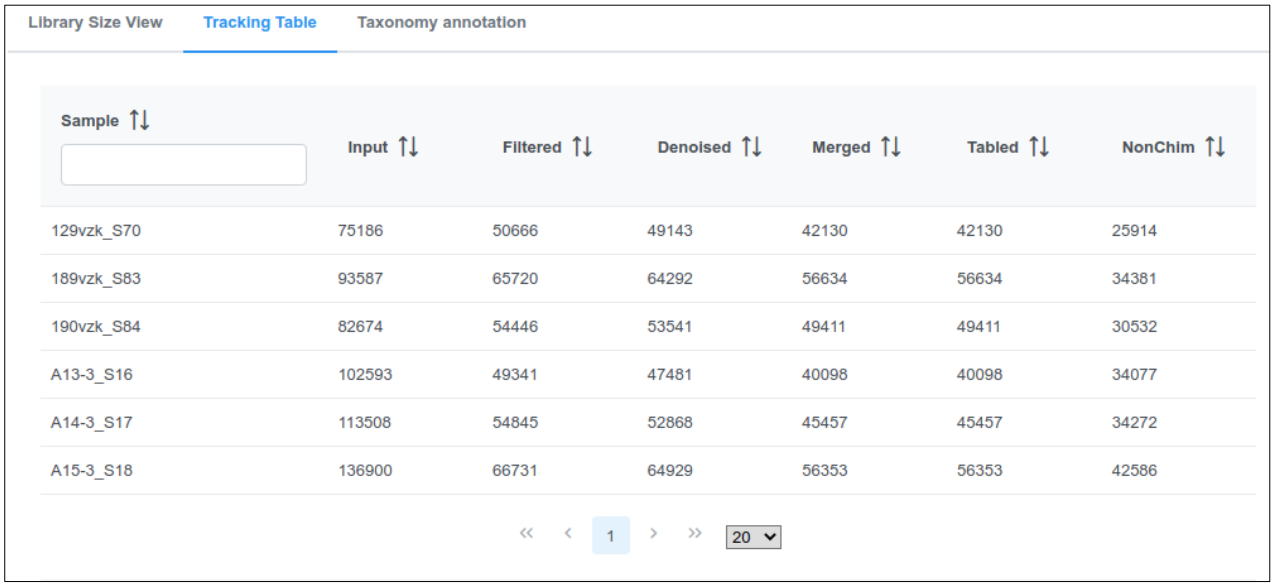

Рис. 18. Количество прочтений, оставшееся на каждом этапе анализа

2. Таксономическая аннотация выявленных ASV

Также данная страница сайта MicrobiomeAnalyst содержит информацию о таксономической принадлежности каждого ASV (рис. 19). Обратите внимание, что один и тот же вид (род, семейство) может иметь разные ASV, т.е. разные последовательности V3-V4 региона маркерного гена 16S рРНК (в случае учебного проекта).

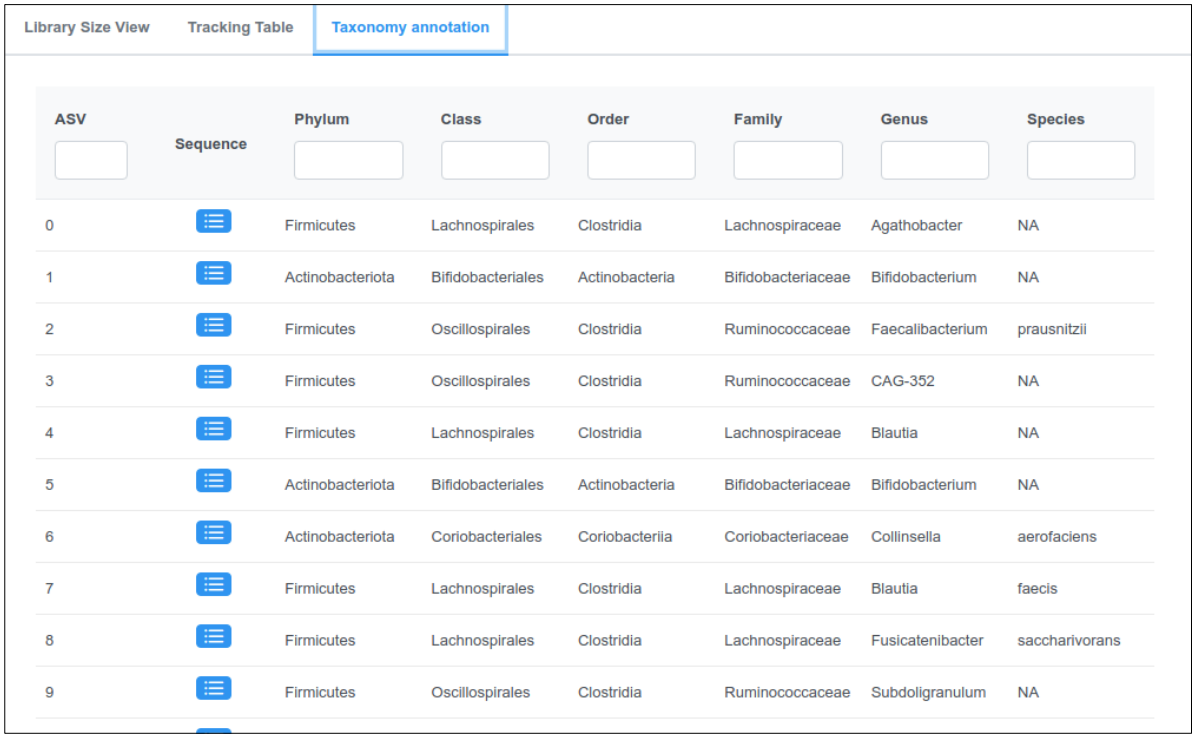

Рис. 19. Таксономическая аннотация выявленных ASV

3. Скачивание полученных результатов

Кликните на кнопку “Proceed” для перехода на страницу для скачивания результатов. В открывшемся окне представлены файлы результатов проведенного анализа. Скачайте на персональный компьютер все файлы одним архивом кликнув на “Download.zip” (рис. 20). Извлеките файлы из архива.

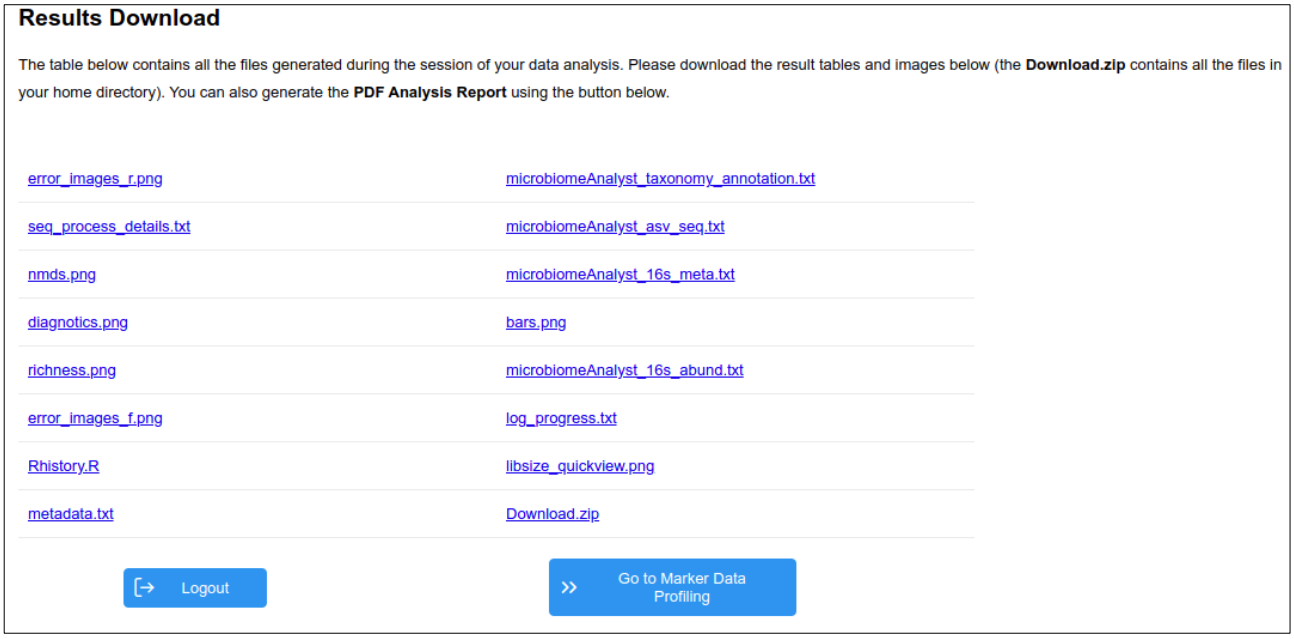

Рис. 20. Страница скачивания полученных результатов

4. Описание полученных результатов

a. Файл “seq_process_details.txt” содержит логи всех этапов анализа.

b. Файл “microbiomeAnalyst_16s_abund.txt” содержит информацию о представленности каждого ASV в исследуемых образцах вместе с таксономической аннотацией (рис. 21). Данный файл можно открыть в программе Excel для удобства восприятия. Этот файл можно использовать для дальнейших статистических расчетов, в том числе и с использованием других модулей web-платформы MicrobiomeAnalyst.

	A	B	C	D	E	F	G
1	#NAME	129vzk_S70	189vzk_S83	190vzk_S84	A13-3_S16	A14-3_S17	A15-3_S18
2	Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Agathobacter; uncultured bacterium	496	7943	0	27	1508	0
3	Bacteria; Actinobacteriota; Actinobacteria; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; uncultured bacterium	1661	0	5161	1728	0	0
4	Bacteria; Firmicutes; Clostridia; Oscillospirales; Ruminococcaceae; Faecalibacterium; prausnitzii	912	3379	655	992	593	1402
5	Bacteria; Firmicutes; Clostridia; Oscillospirales; Ruminococcaceae; CAG-352; uncultured bacterium	877	0	0	0	88	6438
6	Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Blautia; uncultured bacterium	2243	260	545	384	761	1306
7	Bacteria; Actinobacteriota; Actinobacteria; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; uncultured bacterium	31	0	408	0	4855	0
8	Bacteria; Actinobacteriota; Coriobacteriia; Coriobacteriales; Coriobacteriaceae; Collinsella; aerofaciens	552	1803	0	993	459	1206
9	Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Blautia; faecis	257	839	0	57	192	3582
10	Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Fusicatenibacter; saccharivorans	170	72	3147	305	376	594
11	Bacteria; Firmicutes; Clostridia; Oscillospirales; Ruminococcaceae; Subdoligranulum; uncultured bacterium	3954	0	0	0	376	141
12	Bacteria; Firmicutes; Clostridia; Oscillospirales; Ruminococcaceae; Subdoligranulum; uncultured bacterium	0	3703	0	28	317	401
13	Bacteria; Firmicutes; Clostridia; Oscillospirales; Ruminococcaceae; Faecalibacterium; prausnitzii	1063	0	118	218	470	1782
14	Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Ligilactobacillus; uncultured bacterium	120	0	3260	0	0	0
15	Bacteria; Actinobacteriota; Actinobacteria; Bifidobacteriales; Bifidobacteriaceae; Bifidobacterium; longum	465	1413	971	105	0	0
16	Bacteria; Firmicutes; Bacilli; Lactobacillales; Lactobacillaceae; Limosilactobacillus; uncultured bacterium	33	0	2592	0	0	0
17	Bacteria; Firmicutes; Clostridia; Oscillospirales; Ruminococcaceae; Faecalibacterium; uncultured bacterium	226	0	1346	414	326	294
18	Bacteria; Firmicutes; Clostridia; Lachnospirales; Lachnospiraceae; Anaerostipes; hadrus	648	170	0	311	633	670
19	Bacteria; Firmicutes; Negativicutes; Veillonellales-Selenomonadales; Veillonellaceae; Dialister; uncultured bacterium	0	0	0	0	2352	0
20	Bacteria; Firmicutes; Bacilli; Erysipelotrichales; Erysipelatoclostridiaceae; Erysipelotrichaceae UCG-003; bacterium	43	591	0	41	722	949
21	Bacteria; Bacteroidota; Bacteroidia; Bacteroidales; Bacteroidaceae; Bacteroides; vulgatus	0	243	29	1775	101	98

Рис. 21. Содержимое файла “microbiomeAnalyst_16s_abund.txt”

c. Файл “microbiomeAnalyst_asv_seq.txt” содержит схожую информацию о представленности каждого ASV в исследуемых образцах с файлом “microbiomeAnalyst_16s_abund.txt”, но вместо таксономической аннотации содержит последовательности обнаруженных ASV (рис. 22).

	A	B	C	D	E	F	G
1	#NAME	129vzk_S70	189vzk_S83	190vzk_S84	A13-3_S16	A14-3_S17	A15-3_S18
2	ACAATGGGCGAAAGCCTGATGCAGCGACGCCGCGTGAGCGAAGAAGTATTTCGGTATGTAAAGCTCTATCAGCAGGGA	496	7943	0	27	1508	0
3	ACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGCGGGATGACGGCCTTCGGGTTGTAAACCGCTTTTGACTGG	1661	0	5161	1728	0	0
4	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTCTTCGGATTGTAAACTCCTGTTGTTGAGG	912	3379	655	992	593	1402
5	GCAATGGGGGAAACCCTGACGCAGCAACGCCGCGTGAAGGATGAAGGTCTTCGGATTGTAAACTTCTTTTATTAAGGA	877	0	0	0	88	6438
6	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAAGGAAGAAGTATCTCGGTATGTAAACTTCTATCAGCAGGGA	2243	260	545	384	761	1306
7	ACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGCGGGATGACGGCCTTCGGGTTGTAAACCGCTTTTGATCGGG	31	0	408	0	4855	0
8	GCAATGGGGGGAACCCTGACGCAGCGACGCCGCGTGCGGGACGGAGGCCTTCGGGTCGTAAACCGCTTTCAGCAG	552	1803	0	993	459	1206
9	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAAGGAAGAAGTATCTCGGTATGTAAACTTCTATCAGCAGGGA	257	839	0	57	192	3582
10	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAGCGAAGAAGTATTTCGGTATGTAAAGCTCTATCAGCAGGGA	170	72	3147	305	376	594
11	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTTTTCGGATTGTAAACTCCTGTCGTTAGGG	3954	0	0	0	376	141
12	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTTTTCGGATTGTAAACTCCTGTCGTTAGGG	0	3703	0	28	317	401
13	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTCTTCGGATTGTAAACTCCTGTTGTTGAGG	1063	0	118	218	470	1782
14	ACAATGGACGAAAGTCTGATGGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTCGTAAAATTCTGTTGTCAGAGA	120	0	3260	0	0	0
15	ACAATGGGCGCAAGCCTGATGCAGCGACGCCGCGTGAGGGATGGAGGCCTTCGGGTTGTAAACCTCTTTTATCGGGG	465	1413	971	105	0	0
16	ACAATGGGCGCAAGCCTGATGGAGCAACACCGCGTGAGTGAAGAAGGGTTTCGGCTCGTAAAGCTCTGTTGTTAGAG	33	0	2592	0	0	0
17	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGGAGGAAGAAGGTCTTCGGATTGTAAACTCCTGTTGTTGGGG	226	0	1346	414	326	294
18	ACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAGTGAAGAAGTATCTCGGTATGTAAAGCTCTATCAGCAGGGA	648	170	0	311	633	670
19	GCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGTGATGACGGCCTTCGGGTTGTAAAGCTCTGTGATCGGG	0	0	0	0	2352	0
20	GCAATGGGGGAAACCCTGACCGAGCAACGCCGCGTGAAGGAAGAAGTAATTCGTTATGTAAACTTCTGTCATAGAGGA	43	591	0	41	722	949
21	TCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGA	0	243	29	1775	101	98

Рис. 22. Содержимое файла “microbiomeAnalyst_asv_seq.txt”

d. Файл “microbiomeAnalyst_taxonomy_annotation.txt” содержит объединенную информацию о последовательностях выявленных ASV, а также таксономическую аннотацию, разгруппированную по отдельным столбцам для каждого иерархического уровня – Царство, Фила, Класс, Порядок, Семейство, Род, Вид (рис. 23). Данная информация может быть необходима для дальнейшего статистического анализа на разных таксономических уровнях, например, при описании отличий представленности бактериальных семейств в разных группах сравнения. Обратите внимание, что не все ASV определились до вида. Это ограничение метода секвенирования ампликонов одного маркерного гена и особенно его фрагмента.

	A	B	C	D	E	F	G	H
1	#TAXONOMY	Kingdom	Phylum	Class	Order	Family	Genus	Species
2	ACAATGGGCGAAAGCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Agathobacter	NA
3	ACAATGGGCGCAAGCCTGATGCAGCGA	Bacteria	Actinobacteriota	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	NA
4	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Faecalibacterium	prausnitzii
5	GCAATGGGGGAAACCCTGACGCAGCAA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	CAG-352	NA
6	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Blautia	NA
7	ACAATGGGCGCAAGCCTGATGCAGCGA	Bacteria	Actinobacteriota	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	NA
8	GCAATGGGGGGAACCCTGACGCAGCGA	Bacteria	Actinobacteriota	Coriobacteriia	Coriobacteriales	Coriobacteriaceae	Collinsella	aerofaciens
9	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Blautia	faecis
10	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Fusicatenibacter	saccharivorans
11	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Subdoligranulum	NA
12	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Subdoligranulum	NA
13	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Faecalibacterium	prausnitzii
14	ACAATGGACGAAAGTCTGATGGAGCAAC	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Ligilactobacillus	NA
15	ACAATGGGCGCAAGCCTGATGCAGCGA	Bacteria	Actinobacteriota	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	longum
16	ACAATGGGCGCAAGCCTGATGGAGCAA	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Limosilactobacillus	NA
17	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Oscillospirales	Ruminococcaceae	Faecalibacterium	NA
18	ACAATGGGGGAAACCCTGATGCAGCGA	Bacteria	Firmicutes	Clostridia	Lachnospirales	Lachnospiraceae	Anaerostipes	hadrus
19	GCAATGGGCGAAAGCCTGACGGAGCAA	Bacteria	Firmicutes	Negativicutes	Veillonellales-Selenc	Veillonellaceae	Dialister	NA
20	GCAATGGGGGAAACCCTGACCGAGCAA	Bacteria	Firmicutes	Bacilli	Erysipelotrichales	Erysipelatoclostridiaceae	Erysipelotrichaceae	bacterium
21	TCAATGGGCGAGAGCCTGAACCAGCCA	Bacteria	Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	vulgatus

Рис. 23. Содержимое файла “microbiomeAnalyst_taxonomy_ annotation.txt”

e. Файл “microbiomeAnalyst_16s_meta.txt” содержит информацию о принадлежности к той или иной группе сравнения каждого образца. Отличается от исходного файла “metadata.txt” тем, что указаны отдельные образцы, а не риды. Данный файл может быть использован для дальнейших статистических расчетов, в том числе и с использованием других модулей web-платформы MicrobiomeAnalyst.

f. Файл “diagnostics.png” представляет собой информацию о качестве секвенирования первых двух библиотек, которая была представлена на рисунке 14. Файлы графиков “error_images_f.png” и “error_images_f.png” представляют собой частоту замен нуклеотидов друг на друга в зависимости от качества секвенирования для прямых (f) и обратных ридов (r), соответственно (рис. 24). Данные графики позволяют выявить закономерности в ошибках секвенирования.

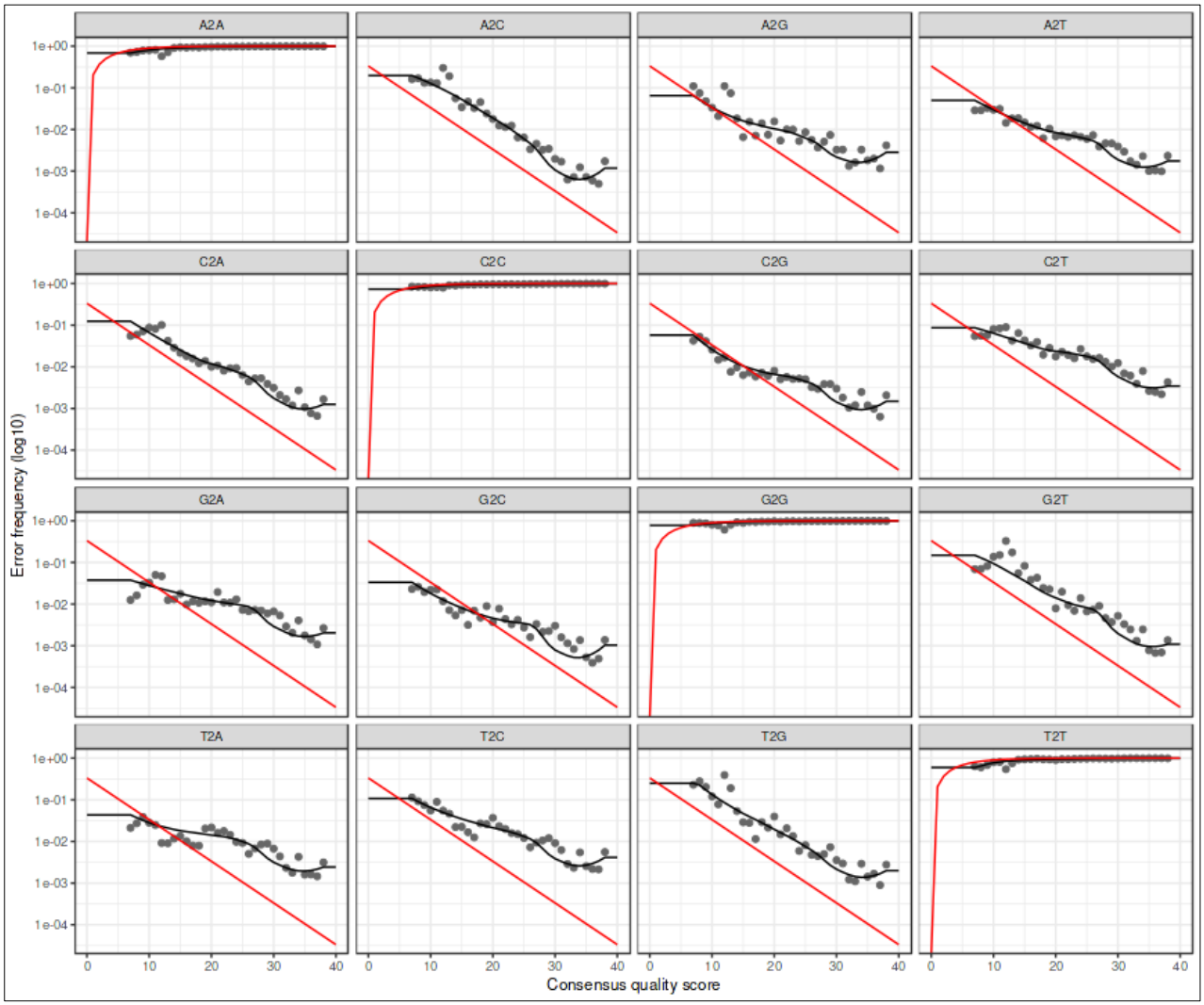

Рис. 24. График “error_images_f.png”. A2C – замена аденина на цитозин, и т.п.

g. Файл “libsize_quickview.png” представляет собой информацию о количестве ридов, прошедших все фильтры, которая была представлена на рисунке 17.

h. График “bars.png” представляет собой столбчатые диаграммы накопления, характеризующие таксономический состав в исследуемых образцах с группировкой по группам сравнения (рис. 25). Данные представлены на уровне родов с группировкой по цвету для семейств. График является описательным, служит для первичной оценки таксономического состава исследуемых образцов.

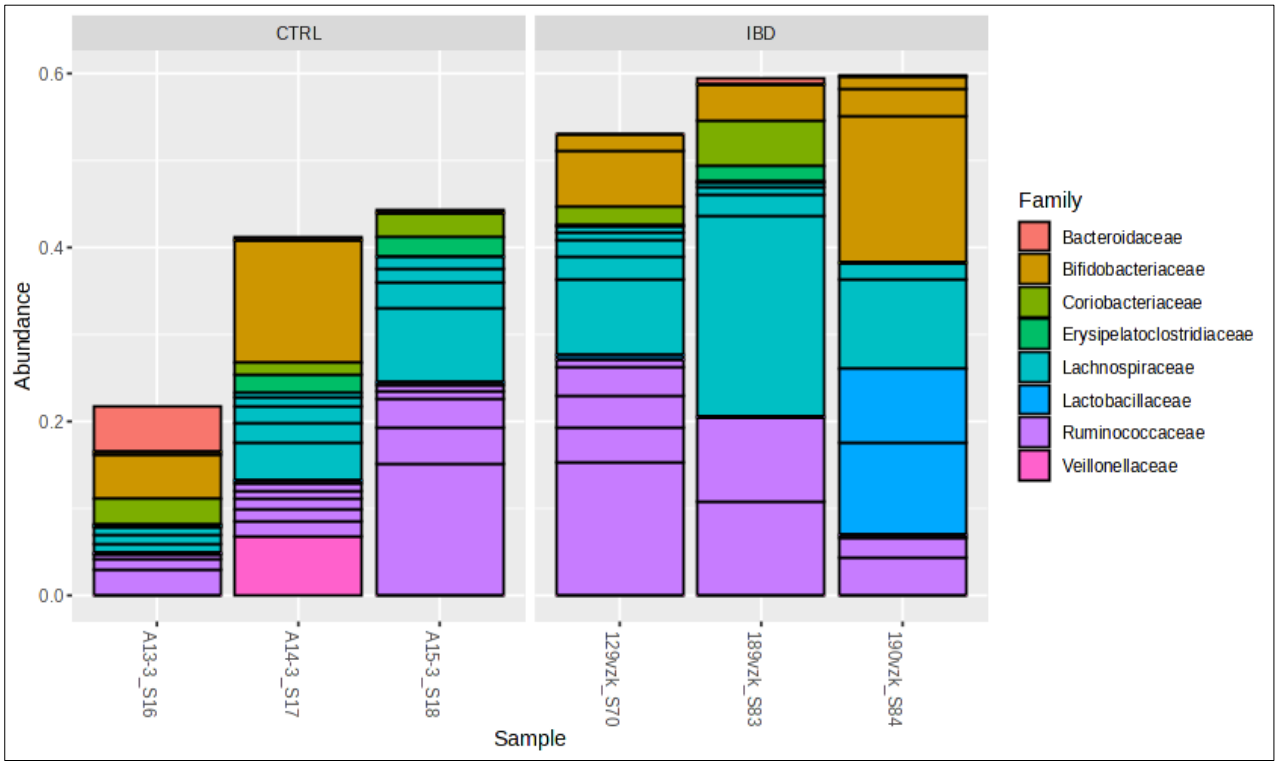

Рис. 25. График “bars.png”

i. График “richness.png” содержит информацию об индексах альфа-разнообразия Шеннона и Симпсона в каждом образце с группировкой по цвету для разных групп сравнения (Рис. 26). Обратите внимание, что по результатам учебного проекта выявлено, что образцы микробиоты кишечника здоровых пациентов имеют более высокие индексы альфа-разнообразия, чем образцы от пациентов с воспалительными заболеваниями кишечника. Это свидетельствует о снижении богатства и разнообразия микробиоты кишечника у пациентов с воспалительными заболеваниями кишечника.

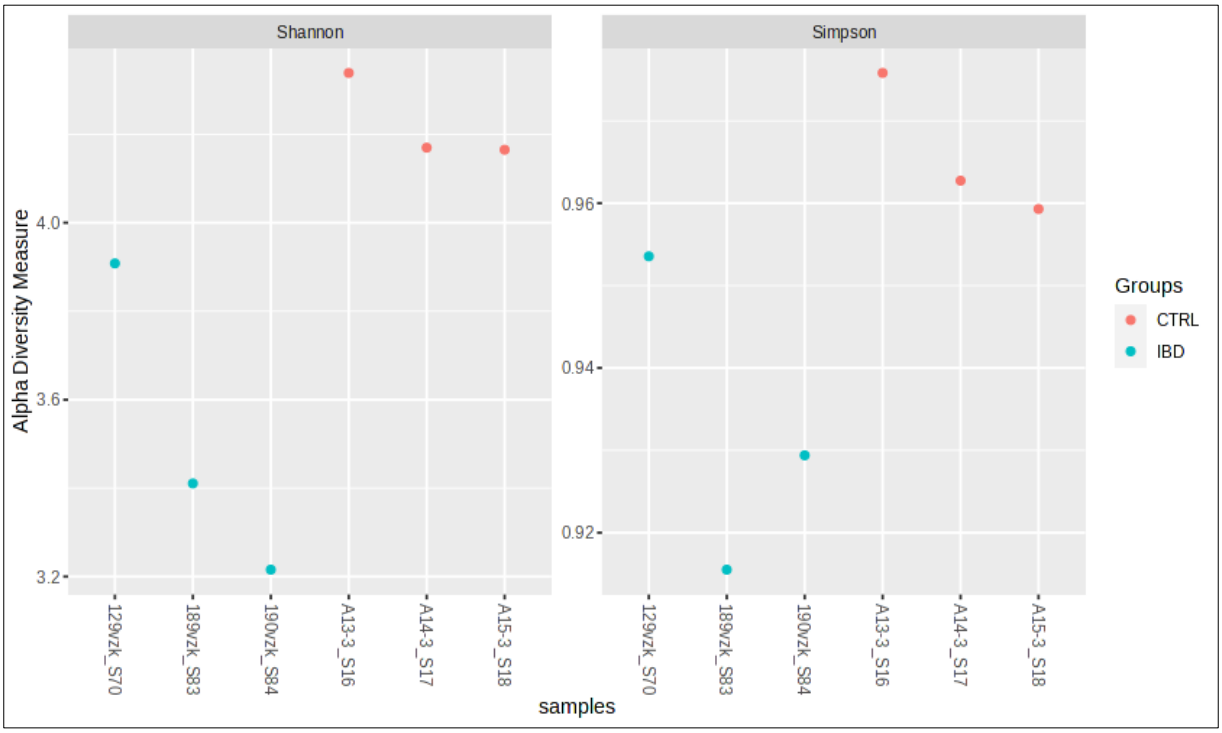

Рисунок 26. График “richness.png”

j. График “nmds.png” представляет собой график расстояний между образцами на основе несходства Брея-Кёртиса (Рис. 27). Это метод оценки бета-разнообразия, метод снижения размерности многомерных данных. Это способ понять, какие образцы больше схожи друг с другом. В более удачном случае, можно было бы увидеть, что образцы из одной группы сравнения образуют кластер, отдельный от кластера образцов из другой группы сравнения (Рис. 28). В случае анализа учебного проекта такой кластеризации не выявлено.

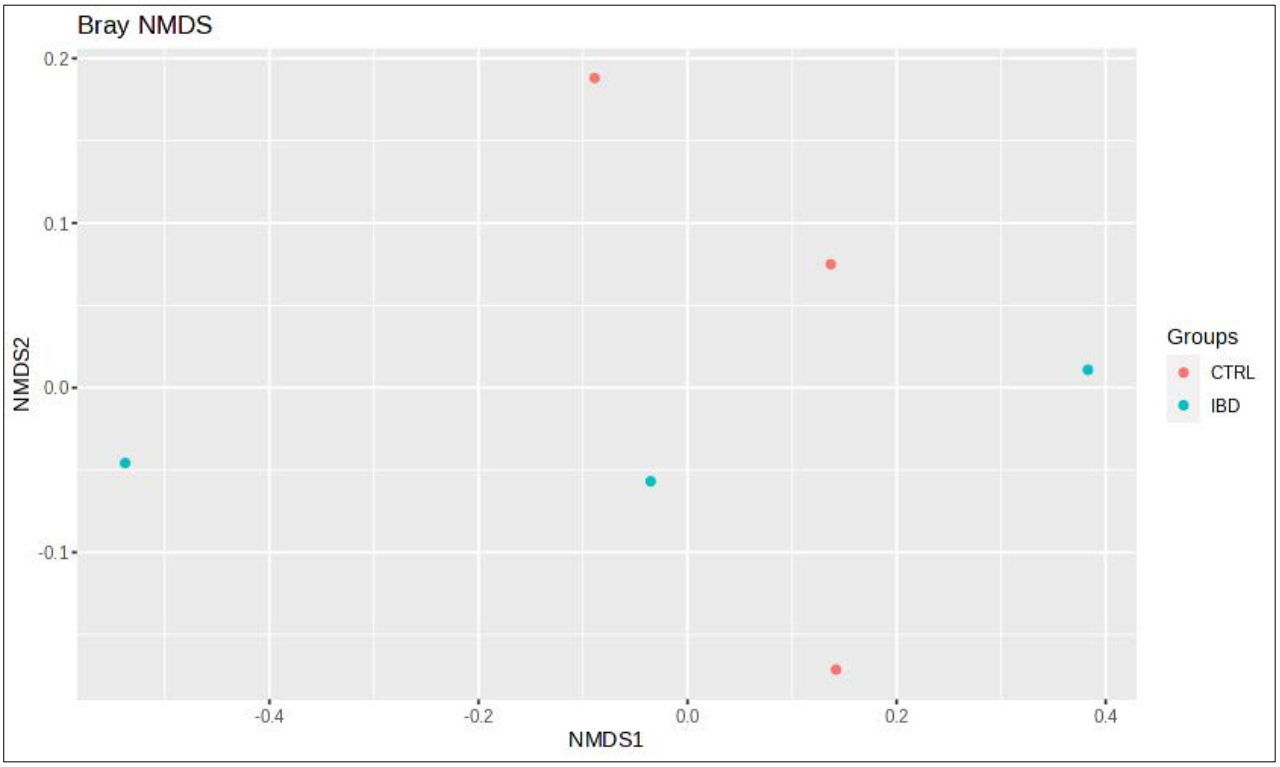

Рис. 27. График “nmds.png”

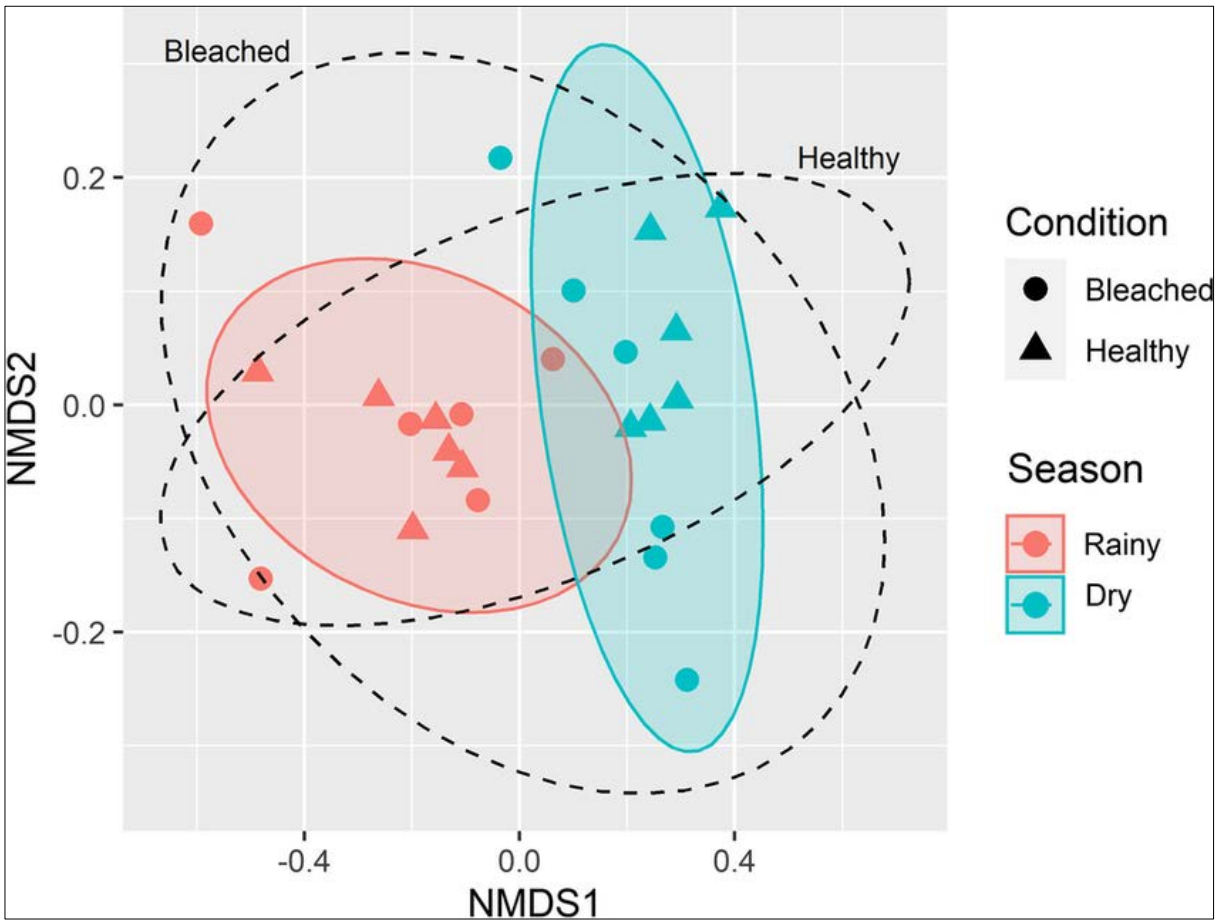

Рис. 28. Пример кластеризации образцов из разных групп сравнения [Paulino *et al.*, 2023]

СТАТИСТИЧЕСКАЯ ОБРАБОТКА И ВИЗУАЛИЗАЦИЯ ДАННЫХ

3.1. Загрузка файлов для визуализации и статистического анализа

Для последующих этапов анализа данных необходимы файлы, полученные на предыдущем шаге - “microbiomeAnalyst_16s_abund.txt” и “microbiomeAnalyst_16s_meta.txt”. Также данные файлы учебного проекта доступны для скачивания по ссылке https://disk.yandex.ru/d/05IqRb1TCa2qDQ.

После успешного завершения предыдущего этапа анализа перейдите на главную страницу сайта (рис. 7). Далее кликните на модуль “Marker Data Profiling” (рис. 29).

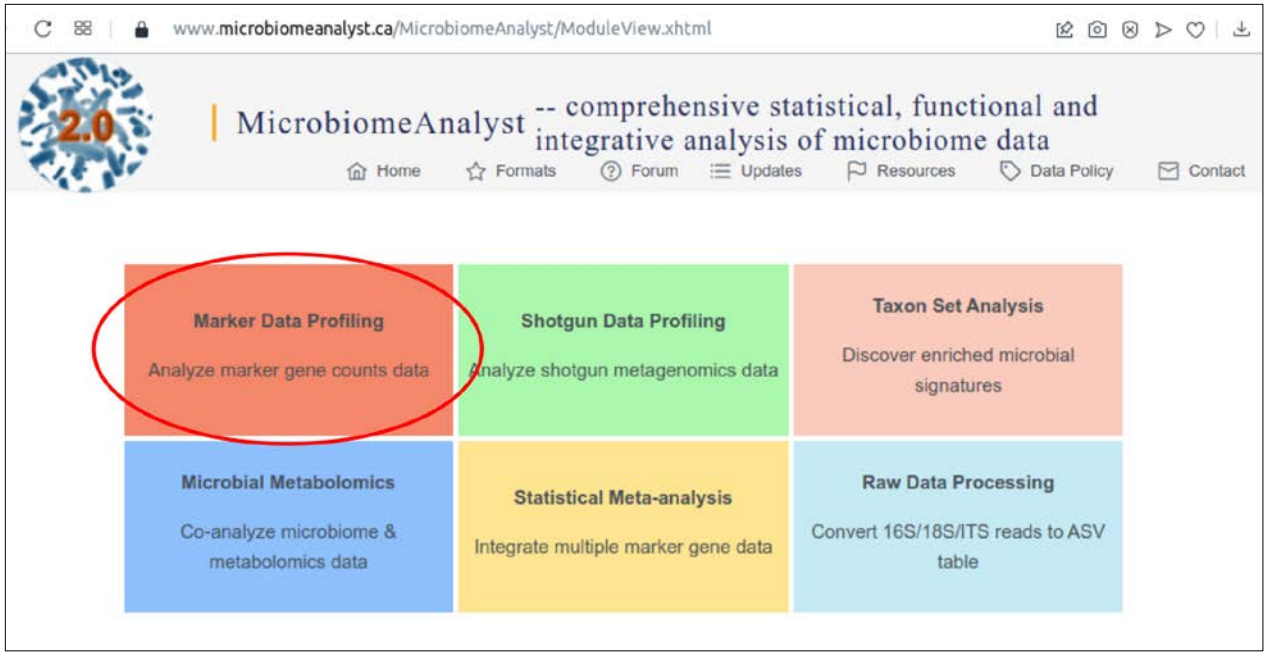

Рис. 29. Страница выбора модуля “Marker Data Profiling” web-платформы MicrobiomeAnalyst

В открывшемся окне выберите вкладку “Text table format”, отметьте галочкой “Taxonomy included” (рис. 30). Далее в пункте OTU/ASV table выберите файл “microbiomeAnalyst_16s_abund.txt” на вашем персональном компьютере, а в пункте “Metadata file” - “microbiomeAnalyst_16s_meta.txt”. В пункте “Taxonomy labels” выберите ту базу данных, которую указывали в п. 2.4. настоящего пособия.

Если используете файлы, скачанные по ссылке из п. 3.1., то отметьте “SILVA Taxonomy”. Далее кликните на кнопку “Submit”.

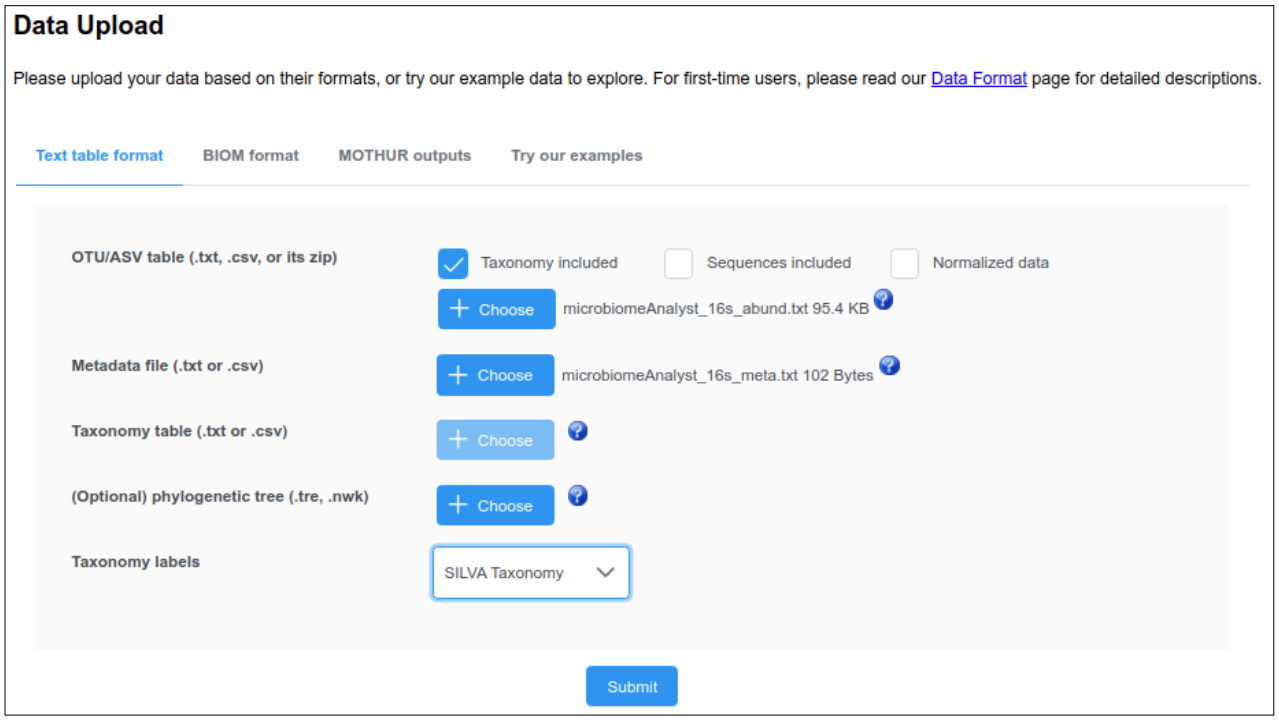

Рис. 30. Страница загрузки файлов представленности ASV и метаданных web-платформы MicrobiomeAnalyst

В открывшемся окне параметры “Default filtering” оставьте без изменения (рис. 31). Во вкладке “Microbiome data overview” проверьте основные пункты:

- “Sample names match” – Yes
- “Normalized counts detected” – No
- 4 пункта “Number of samples” совпадают и соответствуют числу образцов в проекте. При выполнении учебного проекта – 6.

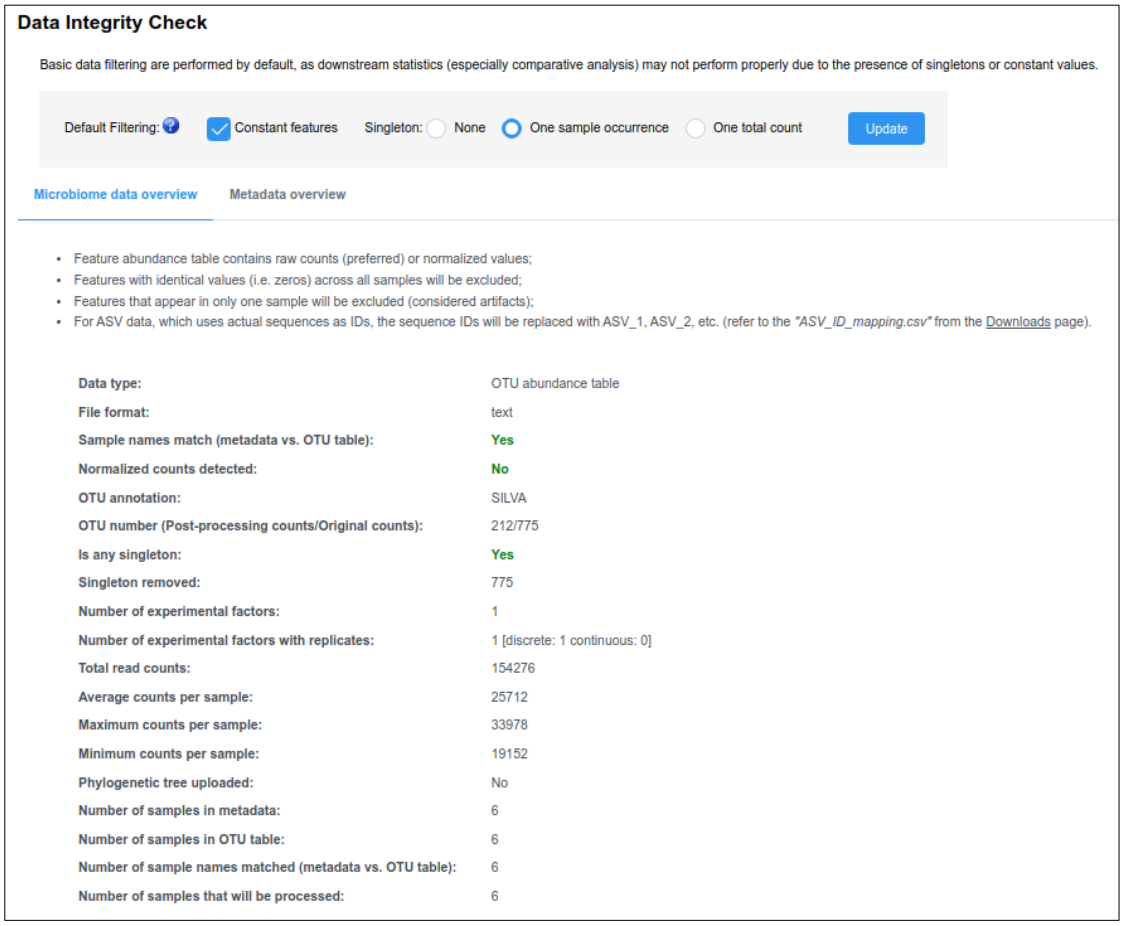

Рис. 31. Страница проверки правильности распознавания загруженных данных

Далее откройте вкладку “Metadata overview”, проверьте правильность распознавания групп сравнения, указанных в файле “microbiomeAnalyst_16s_meta.txt” (рис. 32). Выберите тип данных - “Categorical” для категориальных (дискретных) данных и “Continuous” для непрерывных данных. Категориальные данные в биологических экспериментах чаще всего обозначают категории, к которым могут быть отнесены образцы (опыт и контроль, мужчины и женщины и т.п.). Непрерывные данные могут принимать любые значения в некотором интервале (рос, вес, возраст, содержание какого-либо вещества в исследуемом образце и т.п.). Учебный проект составляют образцы микробиоты кишечника пациентов с воспалительными заболеваниями кишечника и здоровые добровольцы, что соответствует категориальному типу данных (рис. 32). Далее нажать кнопку “Proceed” внизу страницы.

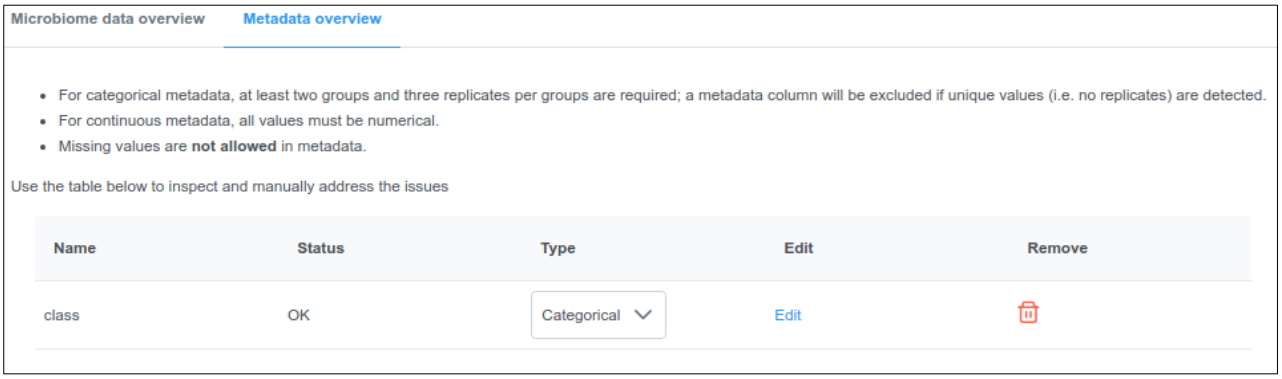

Рис. 32. Страница проверки правильности распознавания метаданных

3.2. Фильтрация и нормализация загруженных данных

Далее необходимо выбрать фильтры для низкопредставленных и низковариабельных ASV. Это позволяет избавиться от ASV, которые представляют собой либо контаминацию образцов, либо ASV, которые присутствуют во всех образцах примерно с одной представленностью. Для выполнения учебного проекта все фильтры поставьте на минимальное значение, чтобы охарактеризовать микробное сообщество полностью (рис. 33). Так, нужно изменить следующие фильтры:

- Minimum count – 0 (оставляем все ASV с представленностью ≥ 0)
- Prevalence in samples (если на предыдущем этапе выбрано “0”, то здесь оставляем любое значение; если на предыдущем этапе выбрано любое значение > 0, например 5, то выставленные в данном пункте проценты (например, 20%) означают, что для прохождения фильтра как минимум в 20% образцов представленность ASV должна быть выше 5). Кроме того, можно выбрать пороговые значения для среднего или медианы.
- Percentage to remove – 0 (оставляем все ASV вне зависимости от их вариабельности). Данный фильтр нужен для отсекания ASV, представленность которых одинакова во всех образцах вне зависимости от группы сравнения, т.е. тех ASV, которые очевидно не ассоциированы с исследуемым признаком. Далее нажать кнопки “Submit” и “Proceed” внизу страницы.

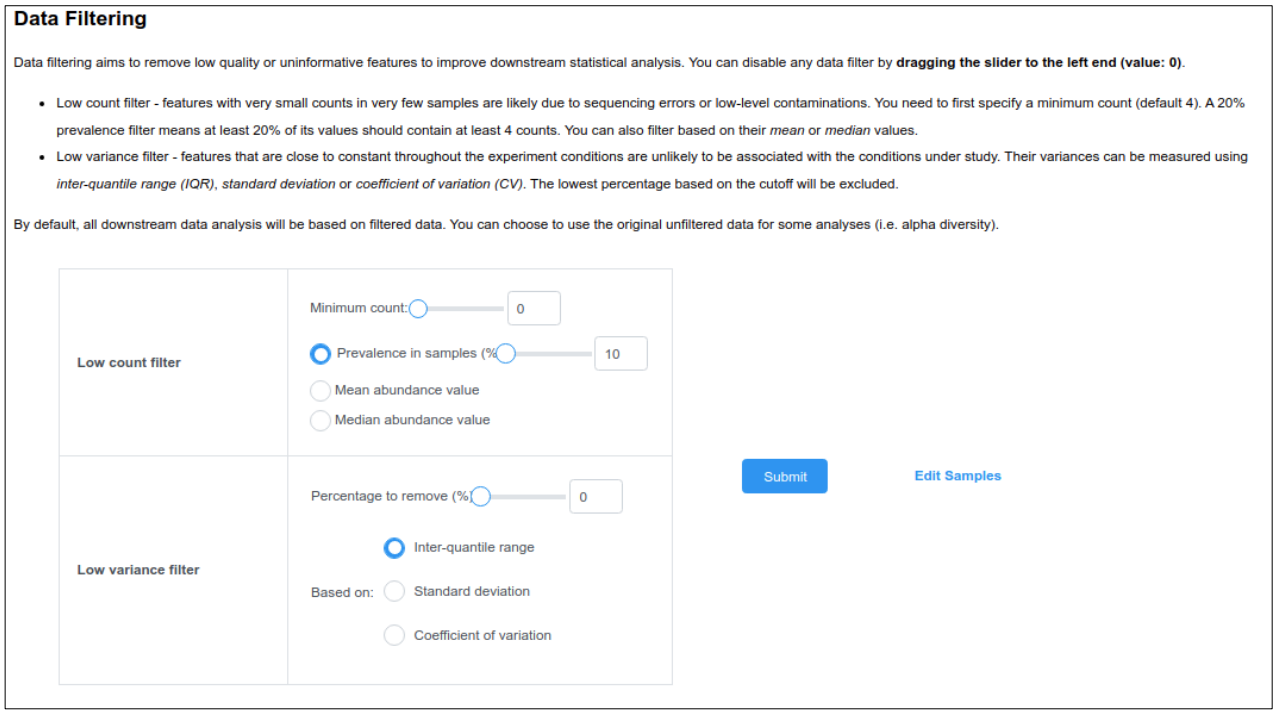

Рис. 33. Страница применения фильтров

Учитывая разную глубину прочтения библиотек (см. п. 2.6.) необходимо провести прореживание – приведение общей представленности ASV во всех образцах к единому значению. Прореживание подразумевает под собой удаление из образцов с более высокой представленностью значений для ASV до того момента, пока значения во всех образцах не будут одинаковы. Эта процедура позволяет унифицировать глубину прочтения для достижения более значимых биологических результатов при сравнении исследуемых образцов. Для этого в окне “Data Normalization” выбрать пункт “Rarefy to library size of” и выбрать минимальное значение, предложенное по умолчанию – это минимальная представленность ASV в одном из образцов после всех этапов анализа (рис. 34).

Далее необходимо выбрать метод нормализации. Обычно нормализация необходима для применения классических статистических тестов, таких как T-тест, ANOVA, Манна-Уитни и т.п. Данные тесты целесообразно применять при наличии хотя бы 10 образцов в каждой группе сравнения. Учебный проект содержит всего 6 образцов, поэтому целесообразно будет использовать более сложные статистические методы (DESeq2, edgeR и др.), которые имеют встроенные

алгоритмы нормализации. Таким образом, при выполнении учебного проекта нет необходимости дополнительно нормализовать имеющиеся данные. Выбрать пункты “Do not scale my data” и “Do not transform my data” (рис. 34). Далее нажать кнопки “Submit” и “Proceed” внизу страницы.

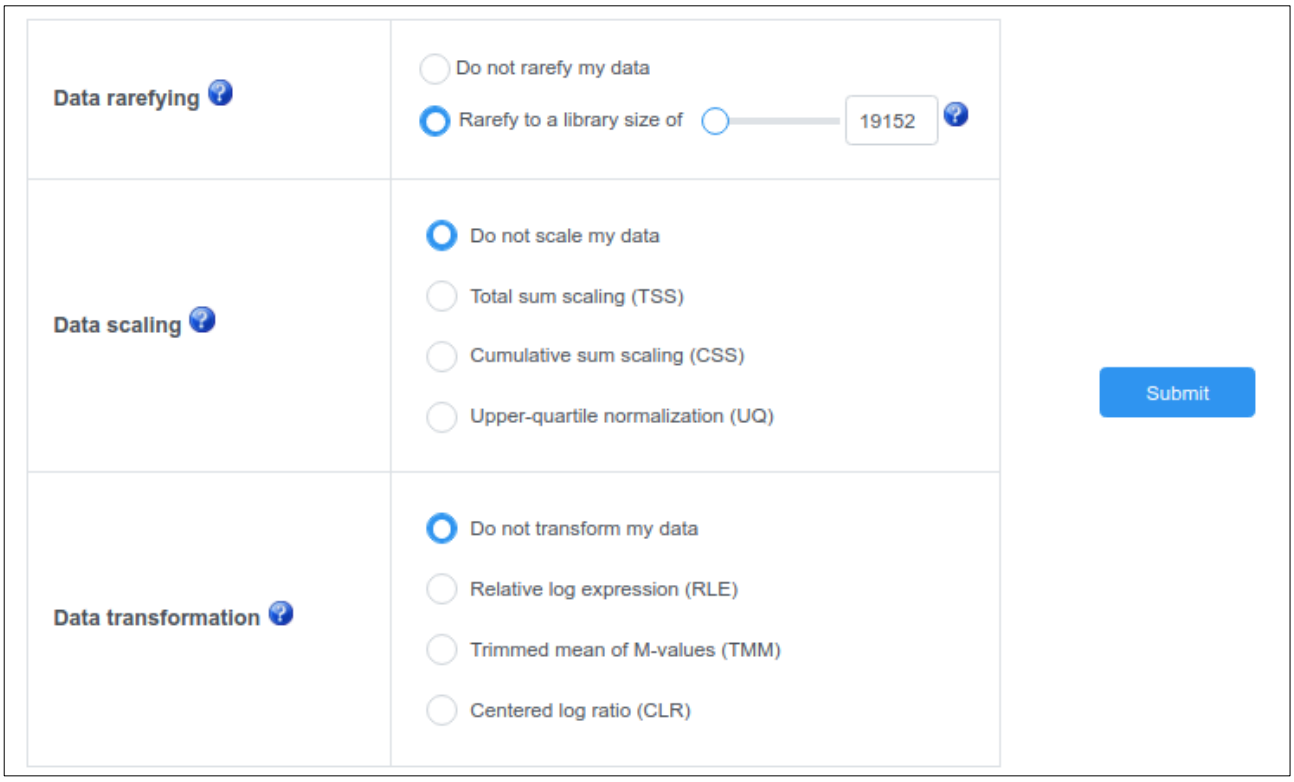

Рис. 34. Страница нормализации данных

3.3. Результаты

Используемый модуль MicrobiomeAnalyst включает в себя все популярные методы анализа данных таксономического состава микробных сообществ, которые представлены на загрузившейся странице.

3.3.1. Классическая визуализация таксономического состава

Модуль “Visual exploration” включает в себя простейшие, но очень популярные методы визуализации состава исследуемых сообществ.

1. Столбчатые диаграммы с накоплением

Выберите в модуле “Visual exploration” пункт “Stacked bar/area plot” (рис. 35). На данном этапе будут нарисованы столбчатые

диаграммы с накоплением для бактериальных фил с группировкой по группам сравнения (рис. 36). Обратите внимание, что данный рисунок является интерактивным, при наведении курсора на область построения графика выводится название таксона, соответствующего области. Это полезно при построении графика для более низких таксонов (семейств, родов, видов), что позволяет визуально отметить паттерны, характерные для того или иного образца.

Рис. 35. Выбор “Stacked bar/area plot”

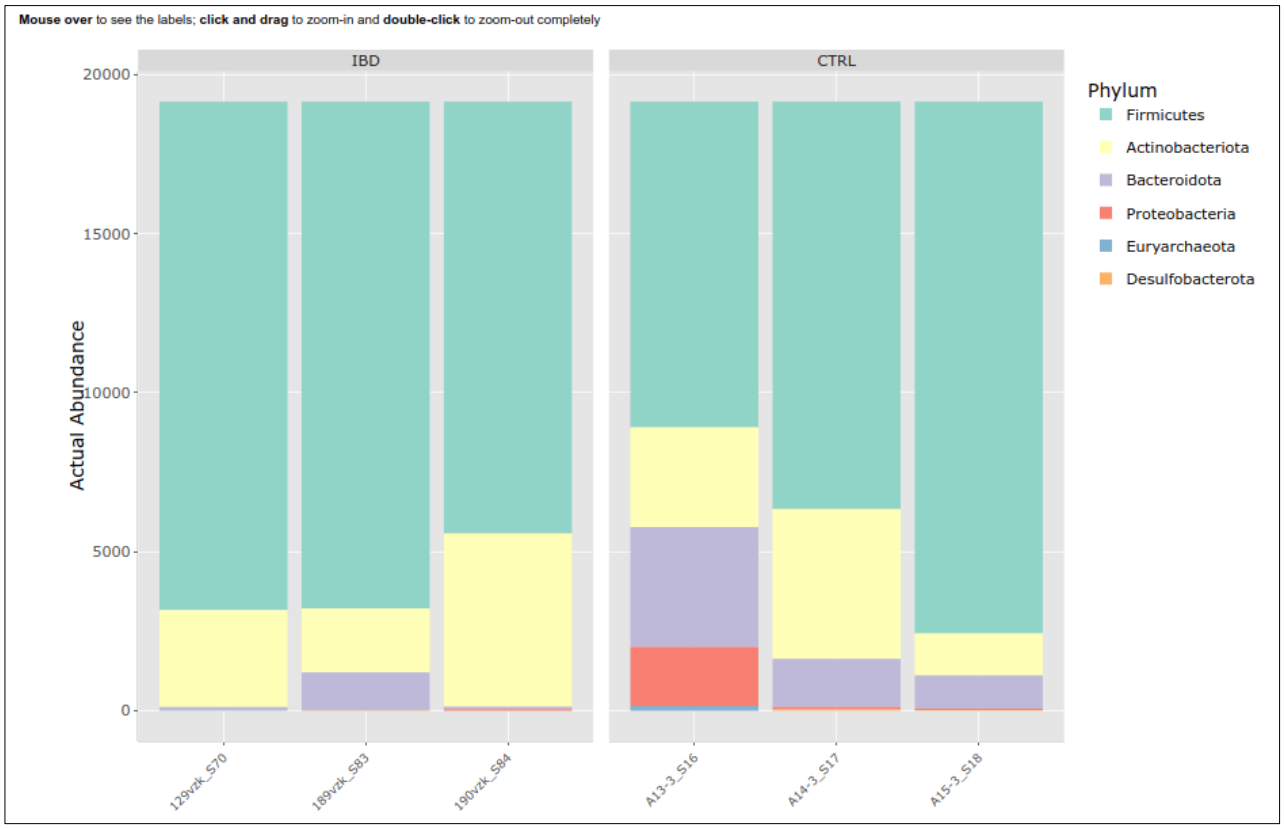

Рис. 36. Столбчатая диаграмма с накоплением для фил

Обратите внимание, что в верхнем левом углу страницы находятся кнопки сохранения табличных данных, а также представленных

на странице рисунков в растровом (png) и векторном формате (svg) (Рис. 37). Данная область будет присутствовать на каждом новом этапе анализа. Сохраняйте файлы на персональный компьютер.

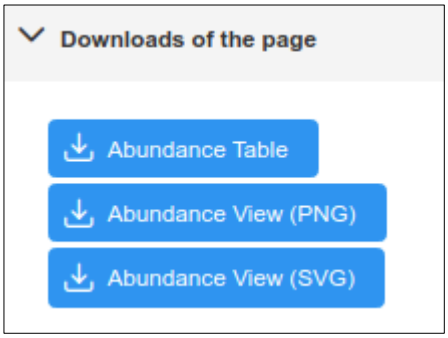

Рис. 37. Область web-страницы с кнопками сохранения результатов

Данный модуль построения столбчатых диаграмм включает в себя несколько пунктов, которые позволяют кастомизировать графики. Так, например, параметры столбчатой диаграммы для относительной представленности родов в процентном соотношении с изменением цветовой палитры соответствуют рисунку 38. Продублируйте указанные параметры и нажмите клавишу “Submit”. Результат построения отражен на рисунке 39.

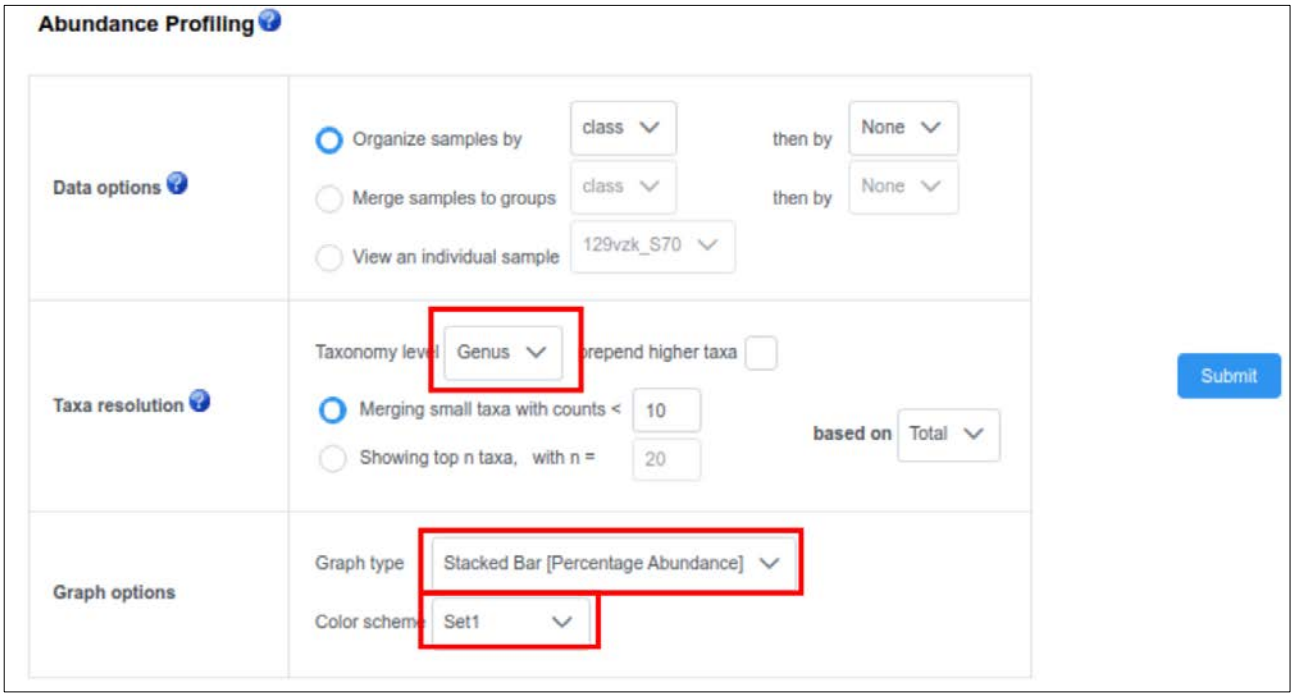

Рис. 38. Изменяемые параметры для построения столбчатых диаграмм для относительной представленности родов в % с изменением цветовой палитры

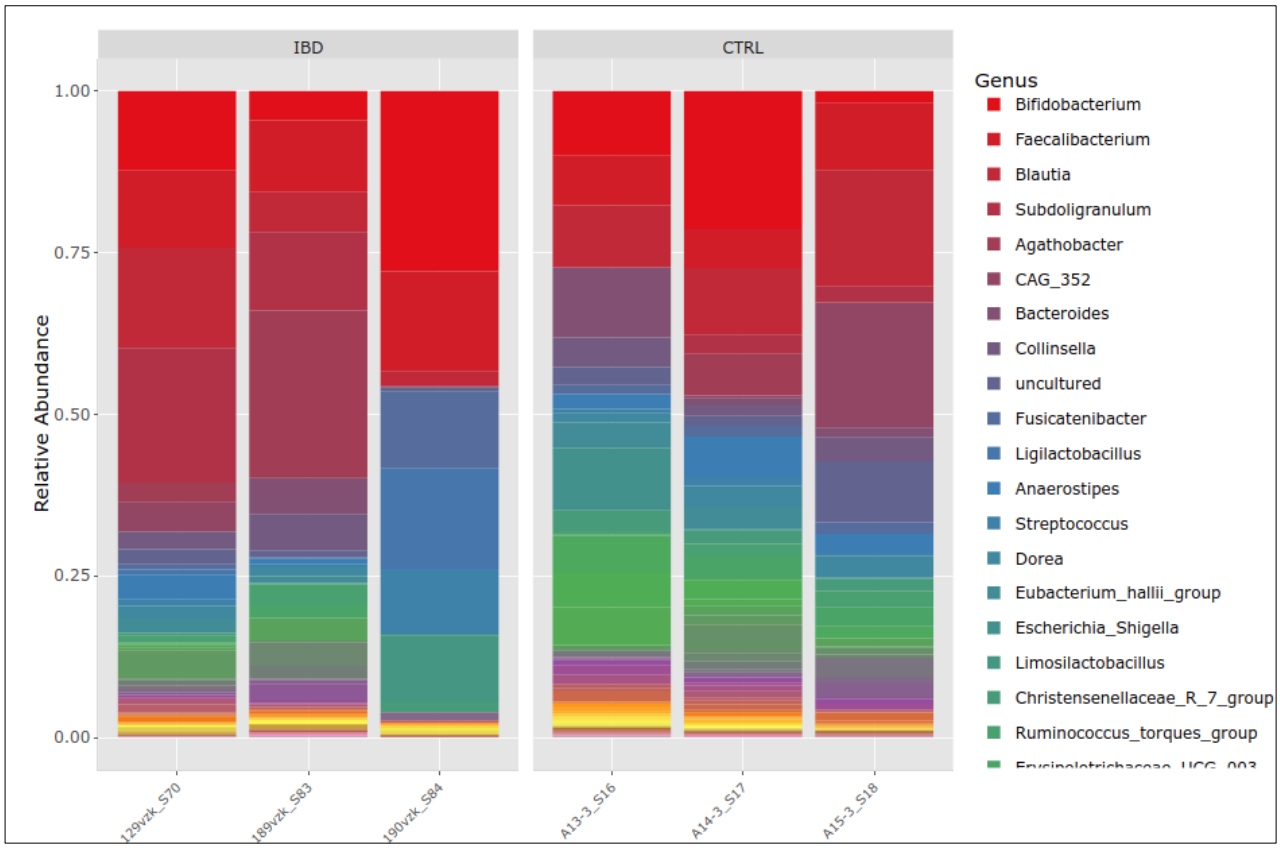

Рис. 39. Результат изменения параметров построения столбчатой диаграммы

Кроме того, можно не рисовать каждый отдельный образец в группе сравнения, а применить группировку по категории образца, а также указать только 10 наиболее представленных семейств бактерий по медиане (рис. 40). Продублируйте указанные параметры и нажмите клавишу “Submit”. Результат построения отражен на рисунке 41.

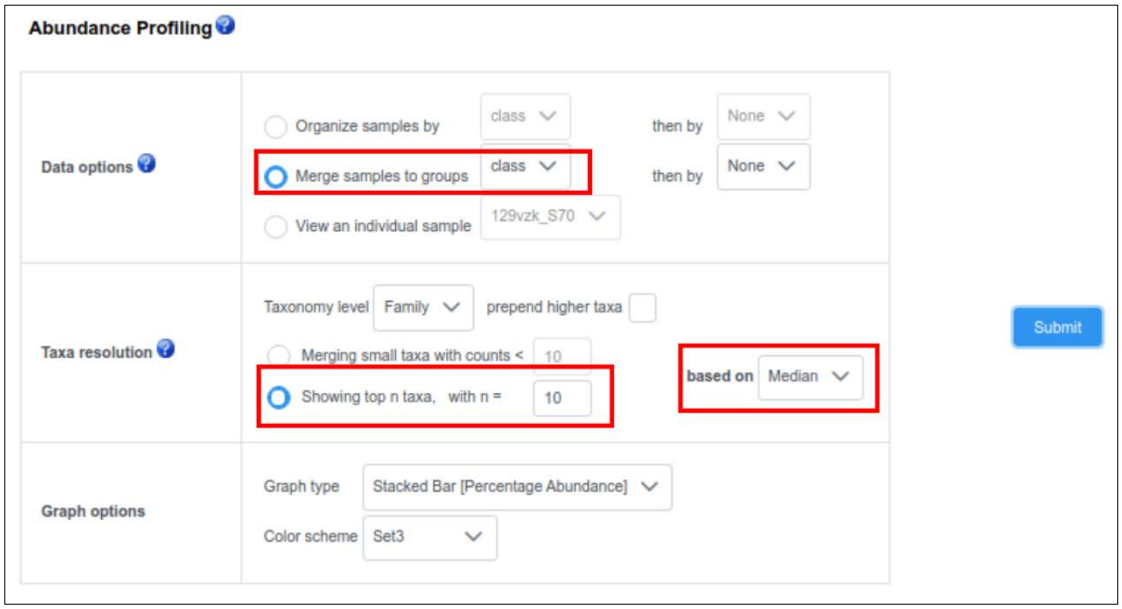

Рис. 40. Изменяемые параметры для построения столбчатых диаграмм для относительной представленности топ-10 семейств в % с группировкой по группам сравнения

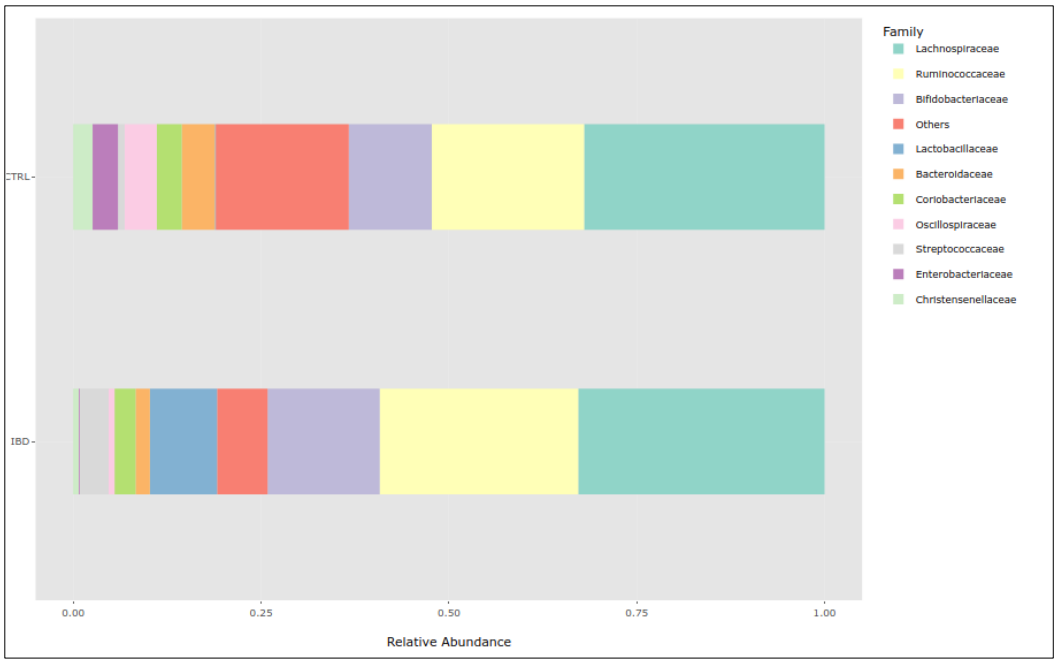

Рис. 41. Результат изменения параметров построения столбчатой диаграммы

Самостоятельно постройте график area plot, изменив параметр “Graph type” на “Stacked area plot” и отменив группировку по категориям, выбрав “Organize samples by class”.

Для выхода на страницу выбора модулей визуализации результатов нажмите на “Analysis overview” наверху web-страницы (рис. 42).

Рис. 42. Клавиша перехода на страницу выбора модулей визуализации результатов

2. Круговые диаграммы (Pie charts)

Перейдите в модуль “Interactive pie chart” (рис. 43). По умолчанию будет нарисована круговая диаграмма для представленности бактериальных фил, усредненная по всем образцам. Данный рисунок также является интерактивным, при наведении курсора на область построения графика выводится название таксона, соответствующего области.

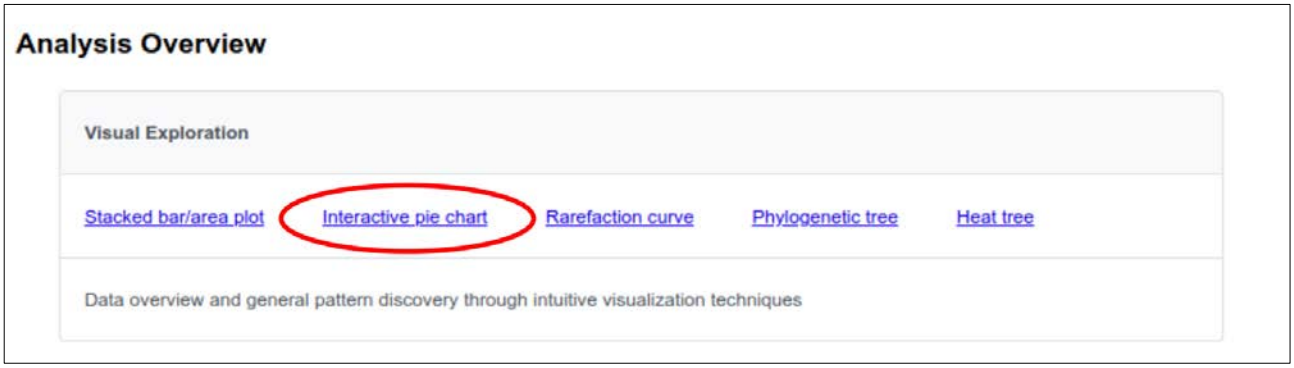

Рис. 43. Выбор модуля “Interactive pie chart”

Для построения круговой диаграммы на основе медианы представленности фил только по одной группе сравнения (например, IBD), выберете параметр, указанный на рисунке 44 и нажмите клавишу “Submit”. Результат построения отражен на рисунке 45.

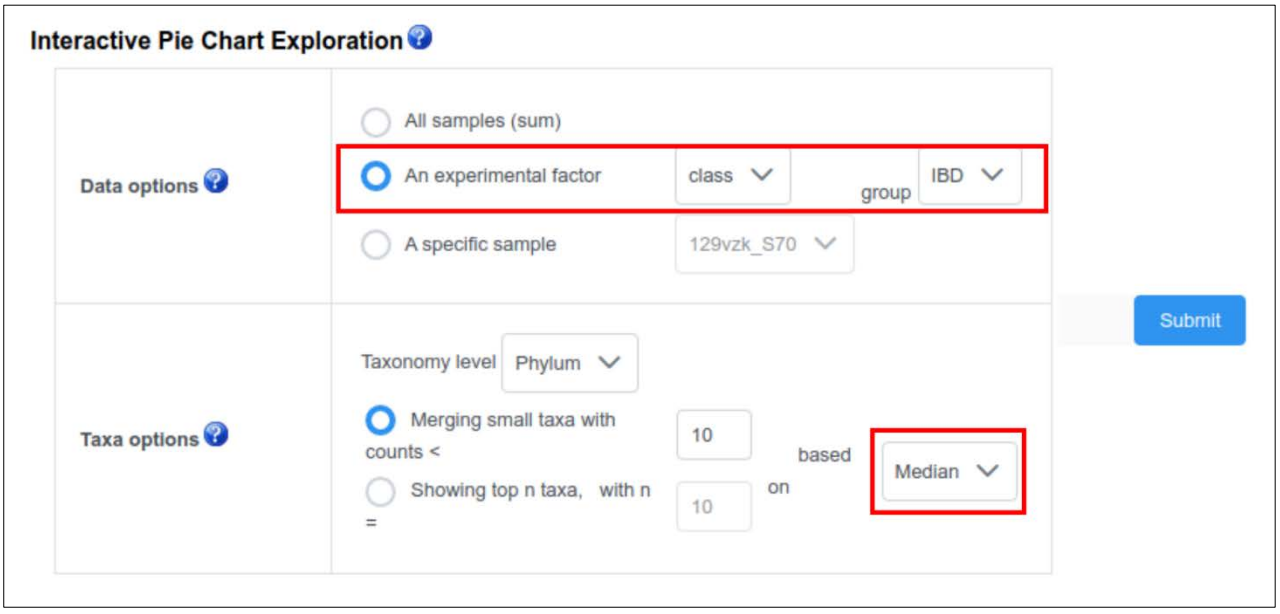

Рис. 44. Изменяемые параметры для построения круговой диаграммы по филам для группы пациентов с воспалительными заболеваниями кишечника (IBD)

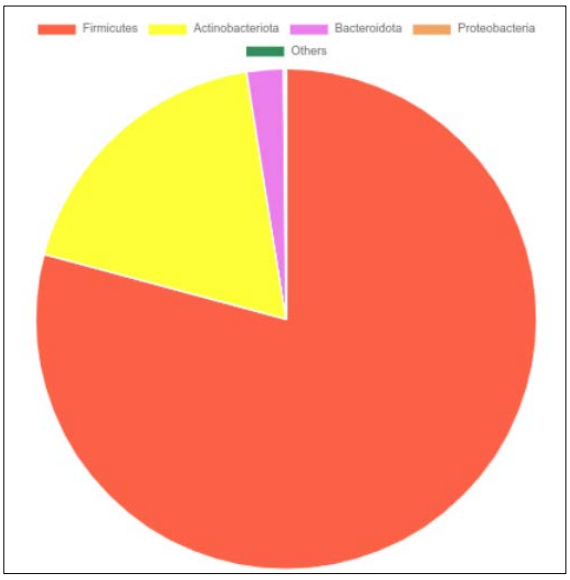

Рис. 45. Результат изменения параметров построения круговой диаграммы

Самостоятельно постройте круговую диаграмму на основе медианы представленности фил для контрольной группы.

Кроме того, постройте график для топ-10 семейств на основе медианы их представленности. Продублируйте параметры, указанные на рисунке 46 и нажмите клавишу “Submit”. Результат построения отражен на рисунке 47.

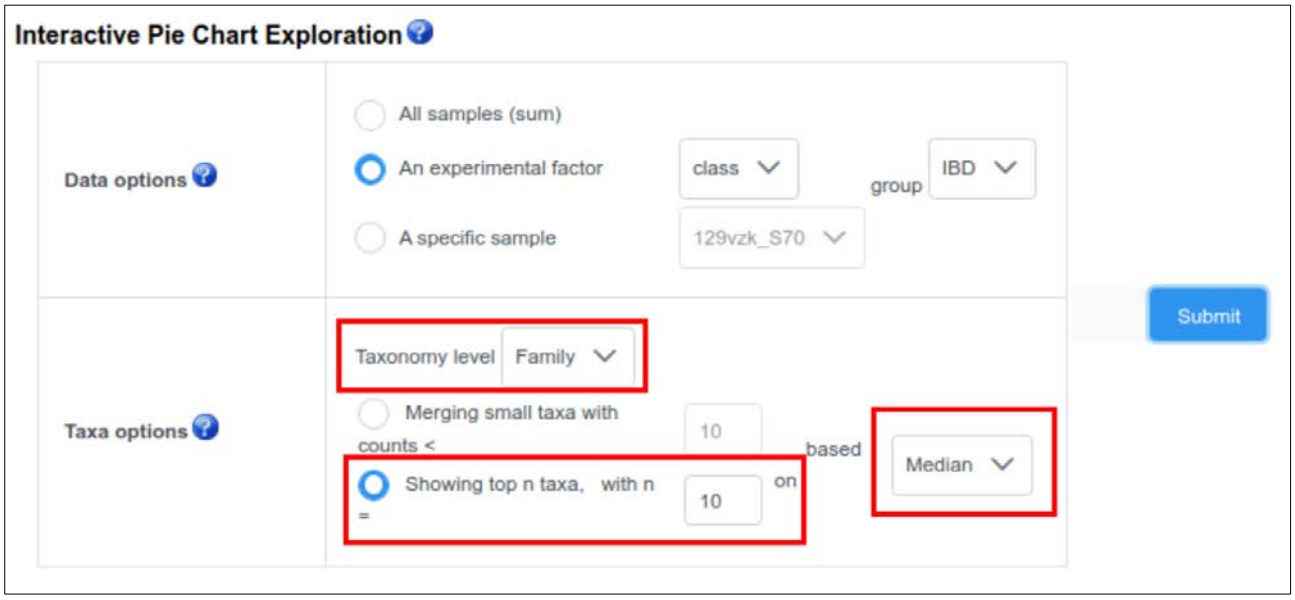

Рис. 46. Изменяемые параметры для построения круговой диаграммы по топ-10 семействам для группы пациентов с воспалительными заболеваниями кишечника (IBD)

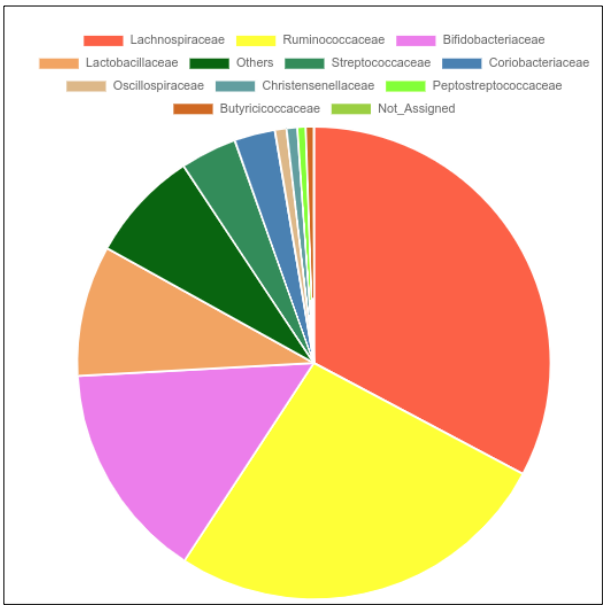

Рис. 47. Результат изменения параметров построения круговой диаграммы

3. Кривая насыщения

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Rarefaction curve” (рис. 48).

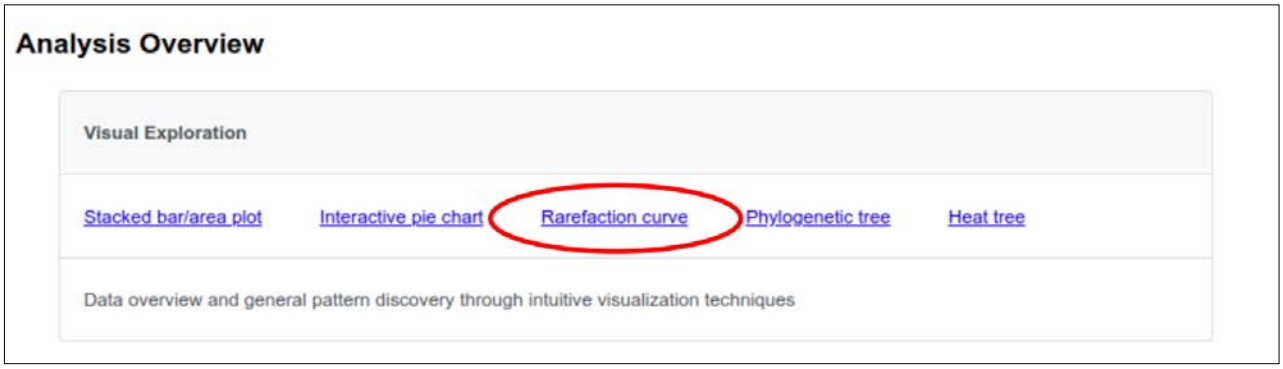

Рис. 48. Выбор модуля “Rarefaction curve”

При переходе к данному модулю автоматически строится график разрежения (насыщения) – график зависимости глубины прочтения (ось x) от количества обнаруженных видов (ось y) (рис. 49).

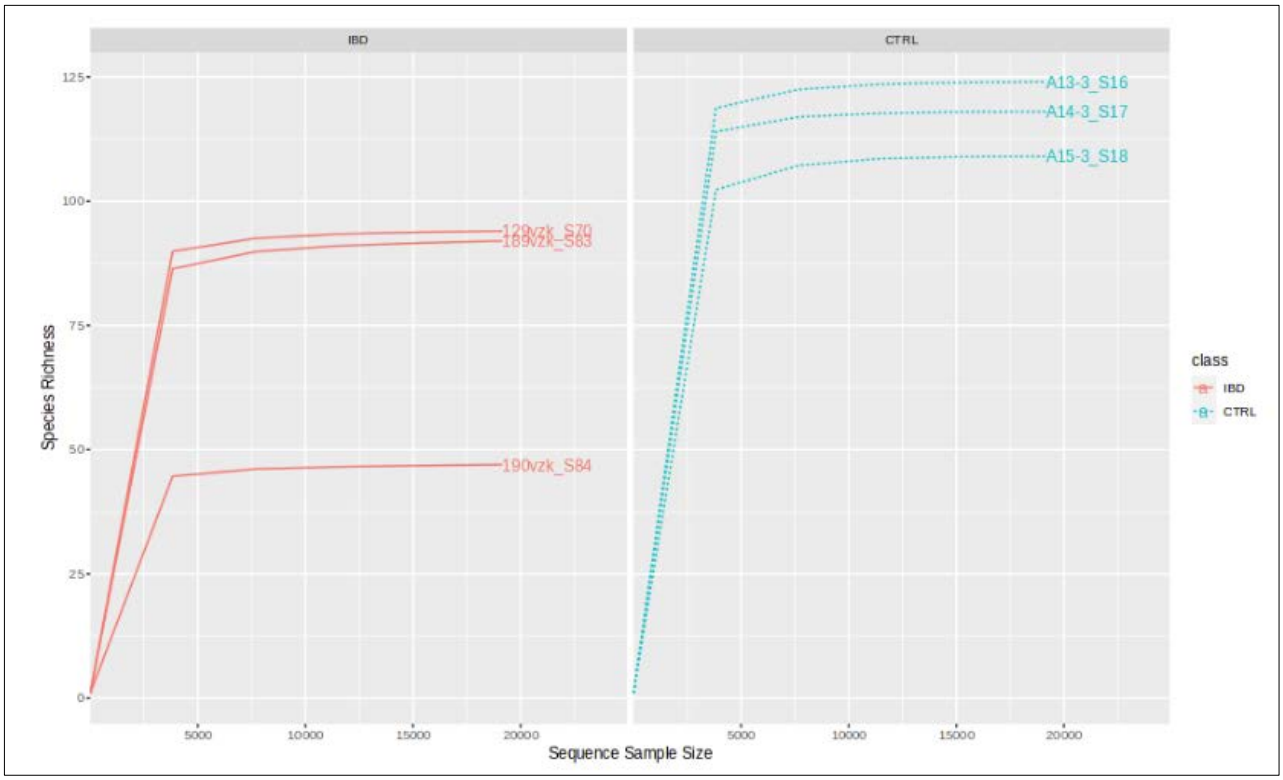

Рис. 49. Кривые насыщения

Данный график позволяет сделать вывод о достаточности/недостаточности глубины прочтения для оценки полноты сообщества. Если кривые насыщения выходят на плато – к концу кривой не

происходит непрерывного увеличения количества видов с увеличением глубины прочтения (правый конец кривой становится линией, параллельной оси x) – то глубина прочтения достаточна для анализа таксономического состава, т.к. были выявлены даже низкопредставленные виды (рис. 50, кривая 2). Для бедных микробных сообществ насыщение будет происходить быстрее, чем для богатых. Если глубины прочтения недостаточно, то кривая насыщения будет стремиться вверх (рис. 50, кривая 1). В таком случае необходимо еще раз секвенировать исследуемый образец микробного сообщества с большей глубиной покрытия.

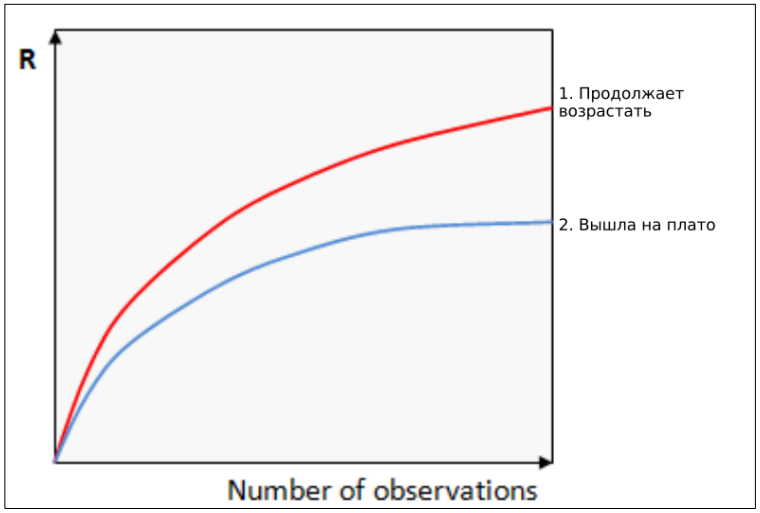

Рис. 50. Примеры кривых насыщения при недостаточной (1) и достаточной (2) глубине прочтения (https://www.drive5.com/usearch/manual/rare.html)

3.3.2. Разнообразие микробного сообщества

Неотъемлемой частью анализа микробных сообществ является оценка альфа- и бета-разнообразия. Альфа-разнообразие – мера оценки богатства и разнообразия сообщества, которая зависит от количества обнаруженных ASV (или видов) и доли каждой ASV (или вида) в сообществе. Бета-разнообразие – мера сходства образцов микробных (и не только) сообществ друг с другом – определяет, насколько сообщества двух исследуемых образцов похожи друг на друга.

1. Альфа-разнообразие

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Alpha diversity” блока “Community profiling” (рис. 51).

Рис. 51. Выбор модуля “Alpha diversity”

Данный модуль позволяет не только оценить альфа-разнообразие исследуемых сообществ с помощью разных индексов, но и применить статистические тесты для оценки достоверных отличий индексов. По умолчанию будет нарисован график для индекса Чао1 (Chao1) на уровне ASV, а статистически значимые отличия данного индекса между группами сравнения рассчитаны с помощью ANOVA. Кроме того, автоматически построен график типа boxplot (боксплот, ящик с усами). Постройте аналогичный график для индекса Шеннона, наиболее популярного индекса альфа-разнообразия. Продублируйте параметры, указанные на рисунке 52 и нажмите клавишу “Submit”. Результат построения отражен на рисунке 53.

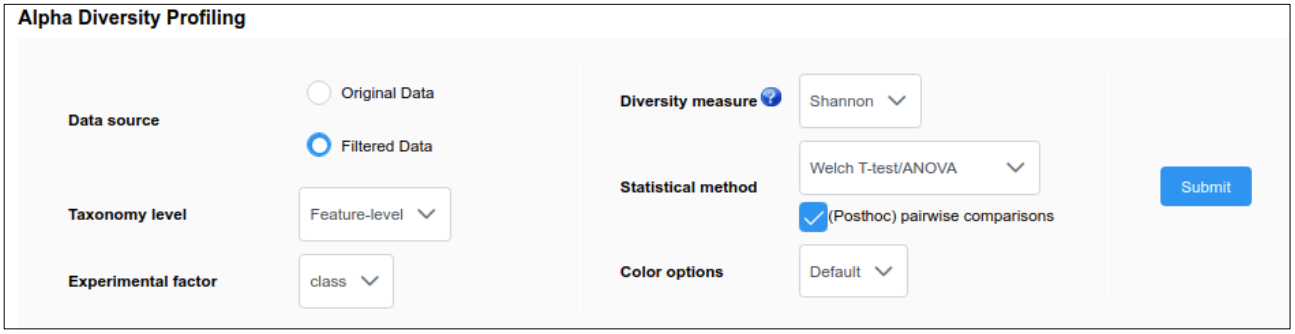

Рис. 52. Изменяемые параметры для построения графиков для индекса Шеннона

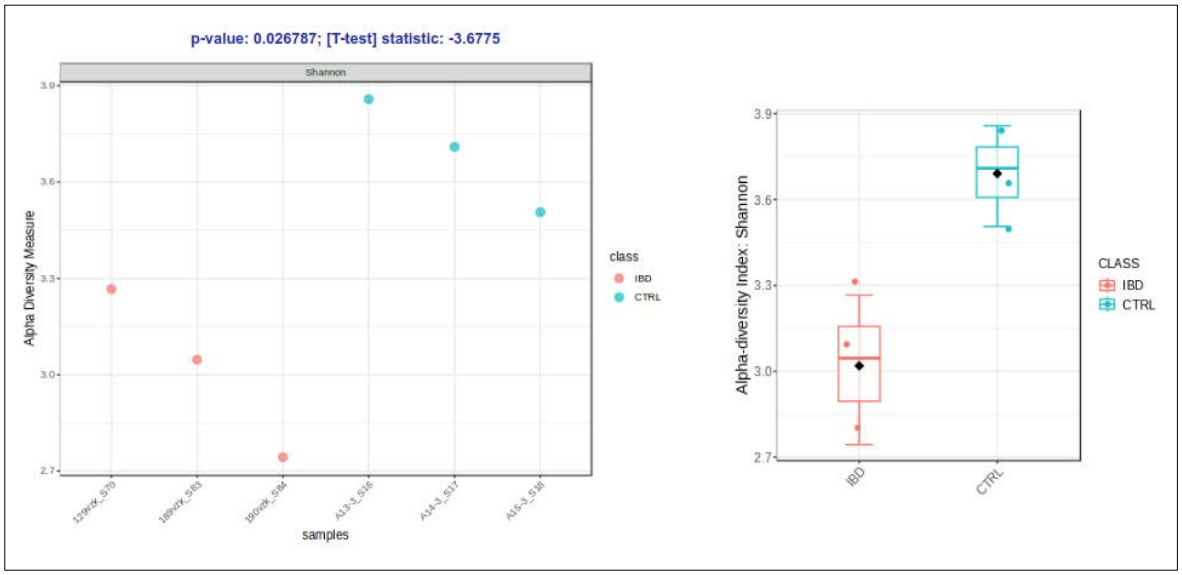

Рис. 53. Результат изменения параметров построения графиков для индекса Шеннона

Над 1 графиком на рисунке 53 указан примененный статистический тест и значение p-value = 0.0267. Таким образом выявлены статистически значимые отличия индекса Шеннона между исследуемыми группами сравнения – пациенты с воспалительными заболеваниями кишечника обладают более низким разнообразием микробиоты кишечника по сравнению с контрольной группой. В случае выполнения учебного проекта, где всего 6 образцов, первый график на рисунке 53 выглядит информативно. При анализе проекта с бóльшим количеством образцов данный график будет перегружен, в таком случае большей информативностью обладает второй график – боксплот. Данный тип графиков всегда имеет единую структуру (рис. 54).

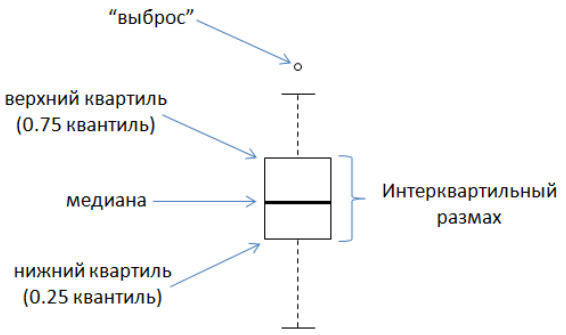

Рис. 54. Структура графика боксплот (ящик с усами) (https://r-analytics.blogspot.com/2011/11/r_08.html)

Постройте график по количеству ASV вместо индекса Шеннона (оценка таксономического богатства сообщества), достоверность отличий между группами сравнения рассчитайте с помощью непараметрического теста Манна-Уитни. Измените цветовую палитру. Продублируйте параметры, указанные на рисунке 55 и нажмите клавишу “Submit”. Результат построения отражен на рисунке 56.

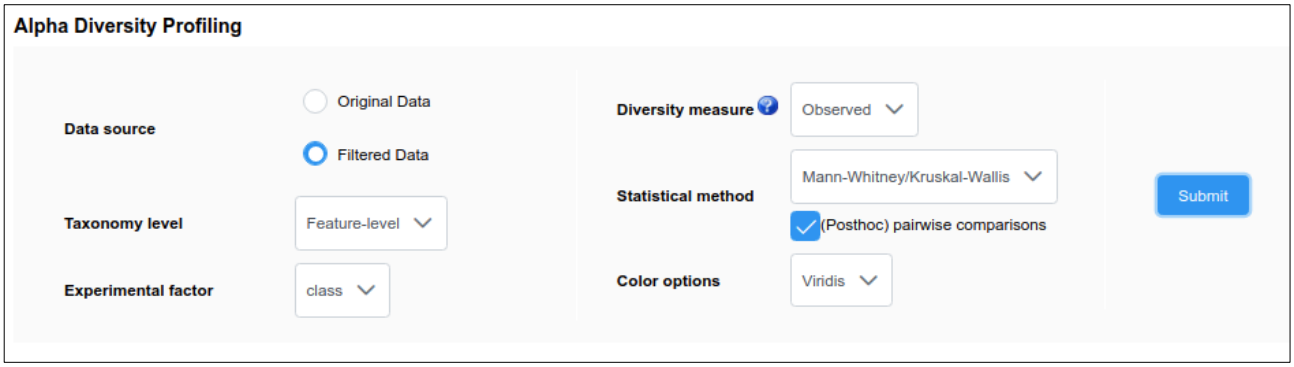

Рис. 55. Изменяемые параметры для построения графиков количества выявленных ASV

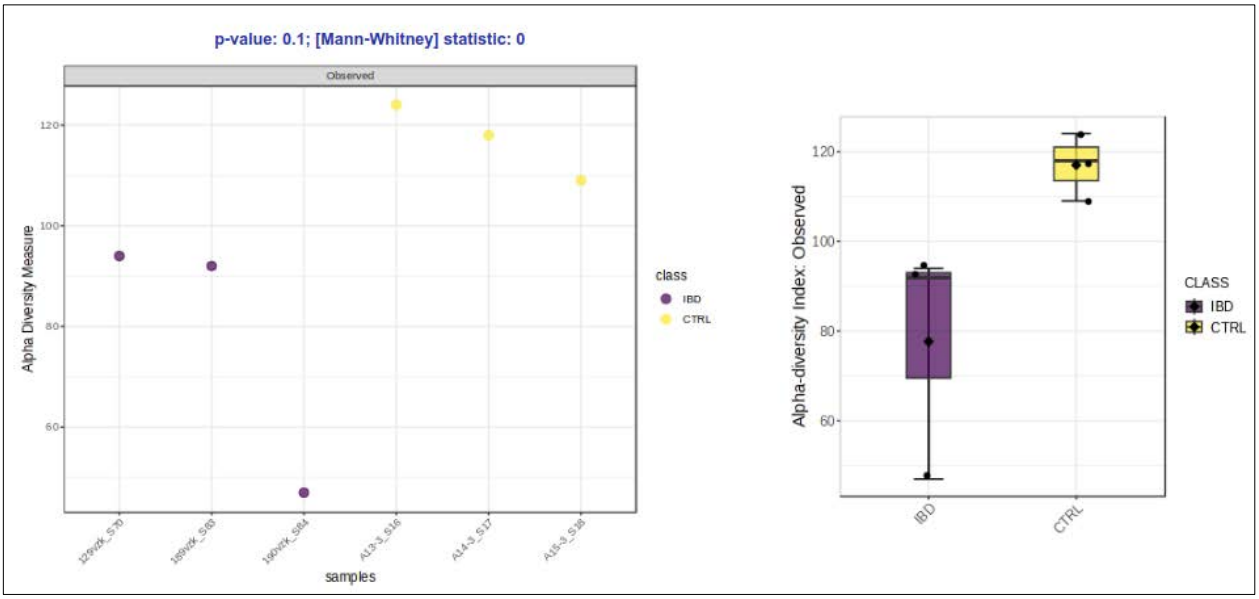

Рис. 56. Результат изменения параметров построения графиков количества выявленных ASV

2. Бета-разнообразие

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Beta diversity” (рис. 57).

Рис. 57. Выбор модуля “Beta diversity”

По умолчанию предлагается построить график PCoA (principal coordinate analysis, анализ главных координат) на основе индекса несходства Брея-Кёртиса на уровне ASV (рис. 58). Кликните на “Update”. Результат построения отражен на рисунке 59.

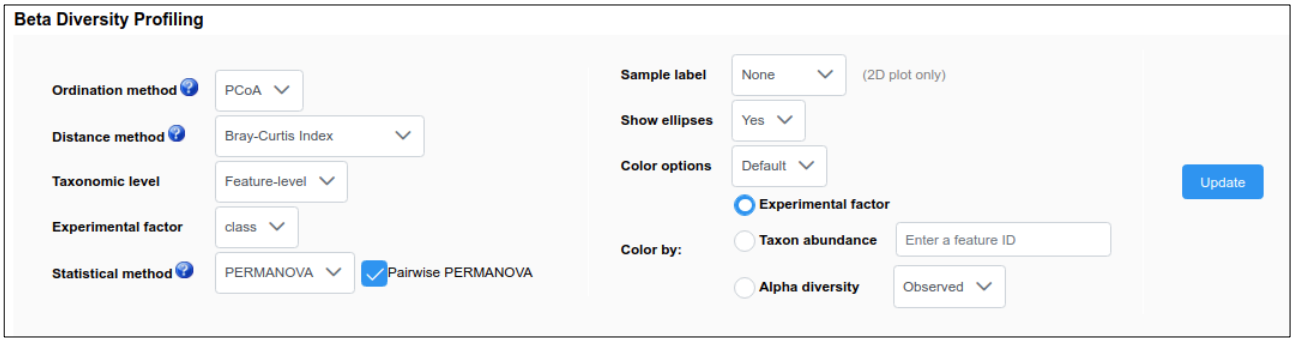

Рис. 58. Изменяемые параметры для построения графика PCoA на основе индекса несходства Брея-Кёртиса

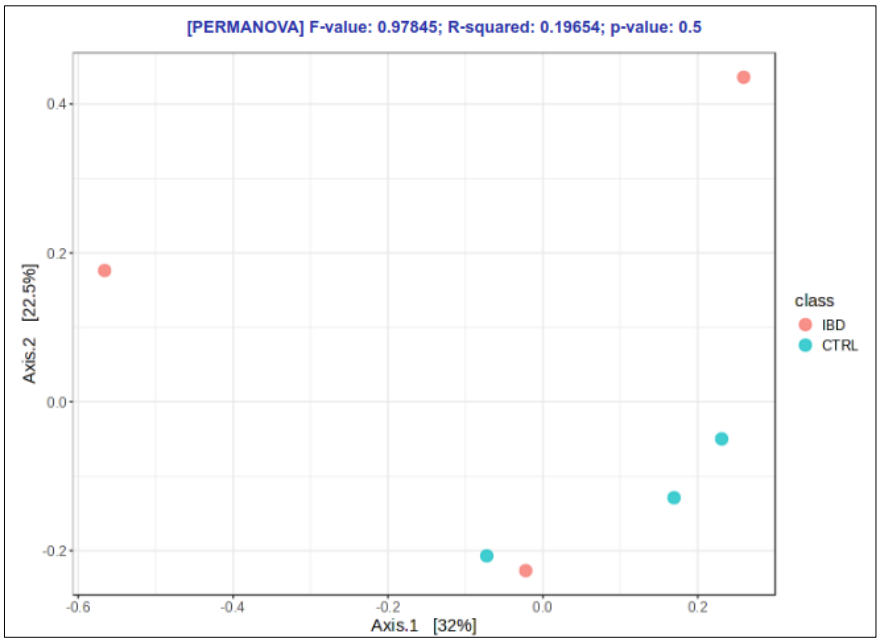

Рис. 59. Результат применения параметров построения графика PCoA на основе индекса несходства Брея-Кёртиса

Как и в пункте 2.6 настоящего пособия, данный рисунок следует трактовать как отсутствие кластеризации образцов по группам сравнения, что подтверждается результатами статистического теста PERMANOVA, p-value = 0,5. Проценты, указанные возле подписей осей обозначают вариабельность, которую они описывают – 32% и 22,5%. Т.е. данный график суммарно описывает 54,5% вариабельности исследуемых данных.

Переключитесь во вкладку “Interactive 3D Plot” для ознакомления с интерактивным трехмерным графиком (Рис. 60). Обратите внимание, что при наведении курсора на точку всплывает название образца. Данный график отражает три главные координаты – три оси, и суммарную вариабельность в 73,6%.

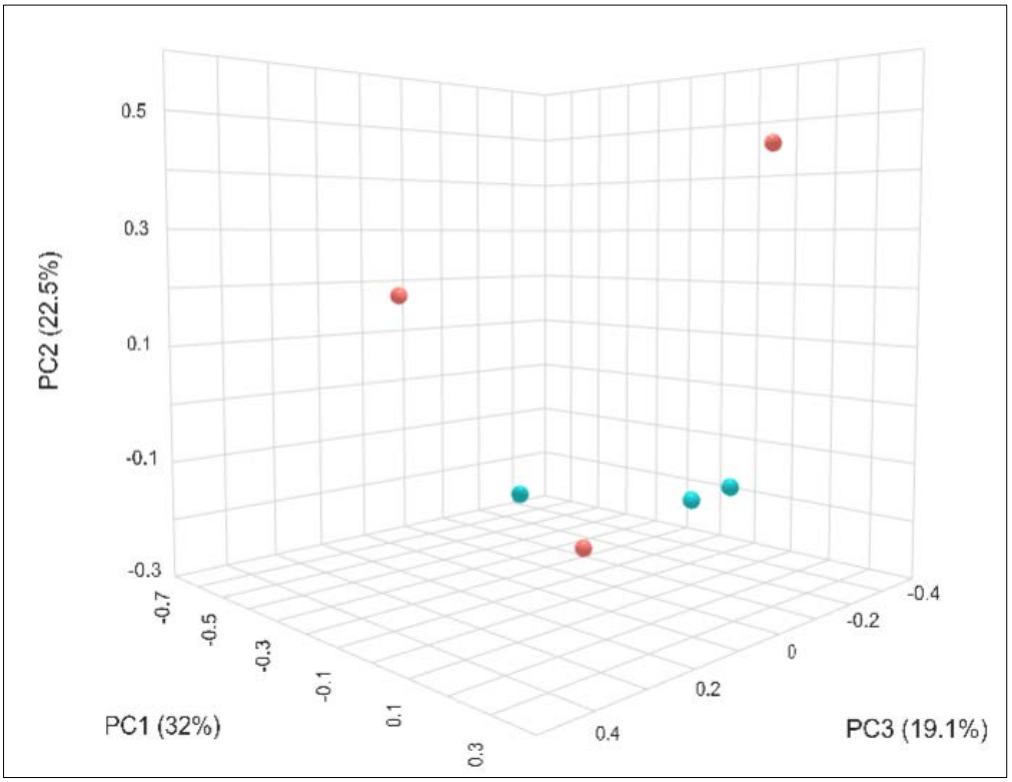

Рис. 60. Интерактивный трехмерный график PCoA на основе индекса несходства Брея-Кёртиса

Постройте график бета-разнообразия с помощью метода главных компонент (principal component analysis – PCA, не путать с PCoA) на основе представленности бактериальных семейств (рис. 61). Также добавьте названия образцов непосредственно на график (рис. 61). Метод PCA уже включает в себя способ расчета дистанций, поэтому

применять методы расчета индекса Брея-Кёртиса или Жаккарда нет необходимости. Кликните на “Update”. Результат построения отражен на рисунке 62.

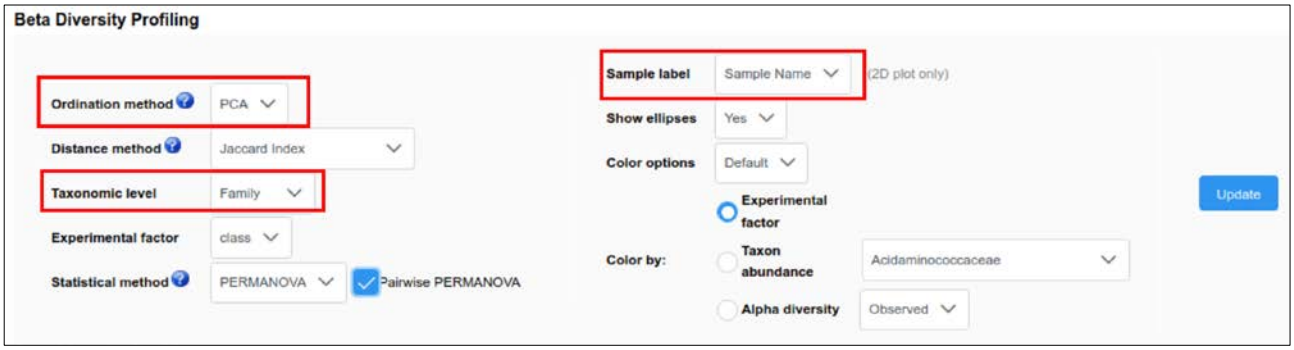

Рис. 61. Изменяемые параметры для построения графиков PCA

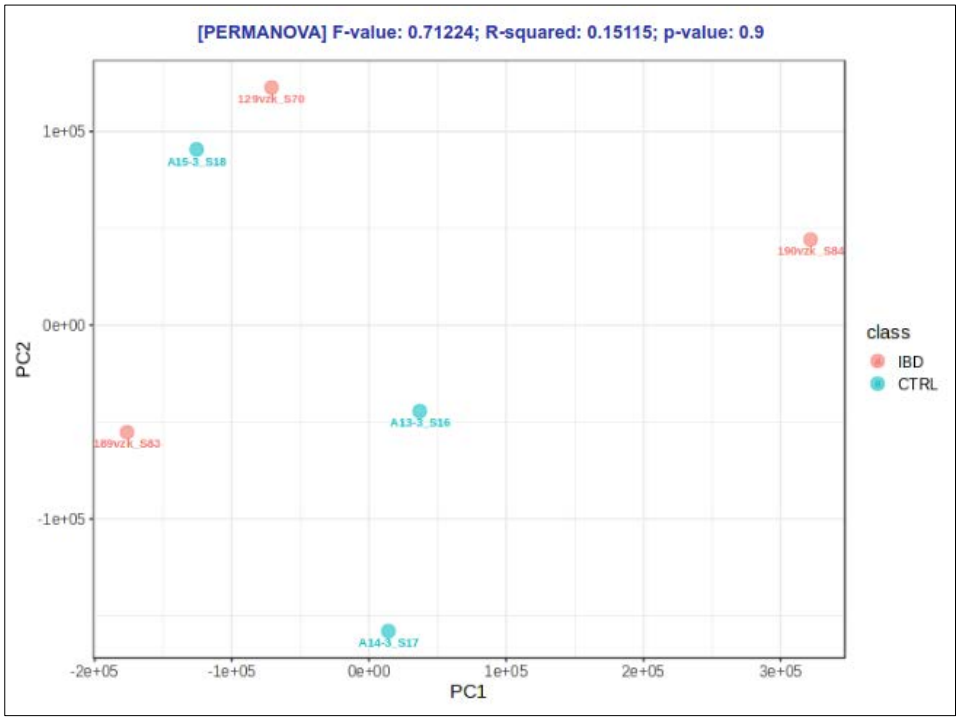

Рис. 62. Результат применения параметров построения графика PCA

Самостоятельно изучите трехмерный интерактивный график PCA.

3. Кор-микробиом (Core microbiome)

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Core microbiome” (рис. 63).

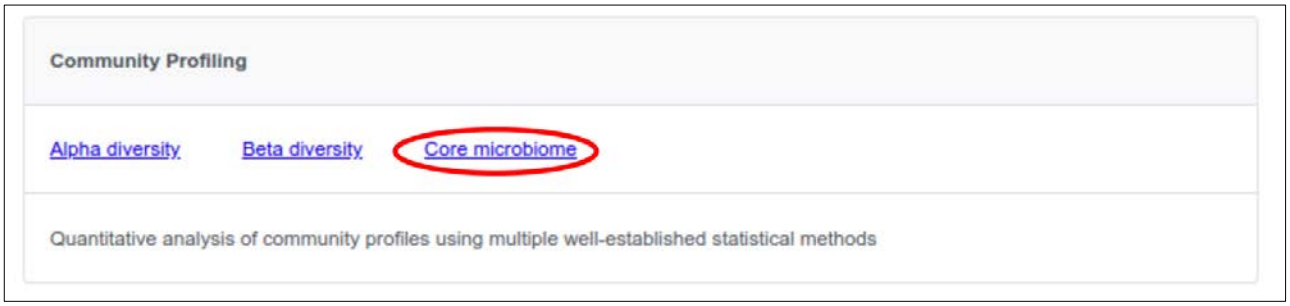

Рис. 63. Выбор модуля “Core microbiome”

Кор-микробиом или ядро микробиома представляет собой группу микробных таксонов, общих для двух или более образцов из конкретного хозяина или среды. Модуль “Core microbiome” содержит гибкие фильтры, которые позволяют выявить общие таксоны для заданного процента образцов, учитывая представленность (рис. 64).

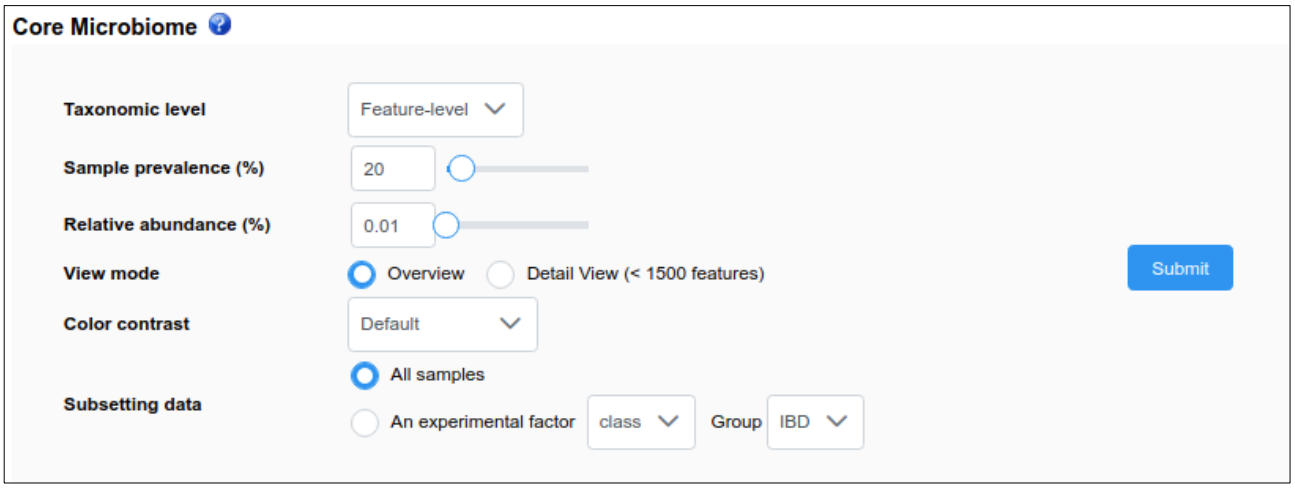

Рис. 64. Изменяемые параметры модуля “Core microbiome”

Постройте график кор-микробиома, определив общие семейства с относительной представленностью более 0,01% для 50% исследуемых образцов вне зависимости от группы сравнения. Изменяемые параметры и результат визуализации представлены на рисунке 65.

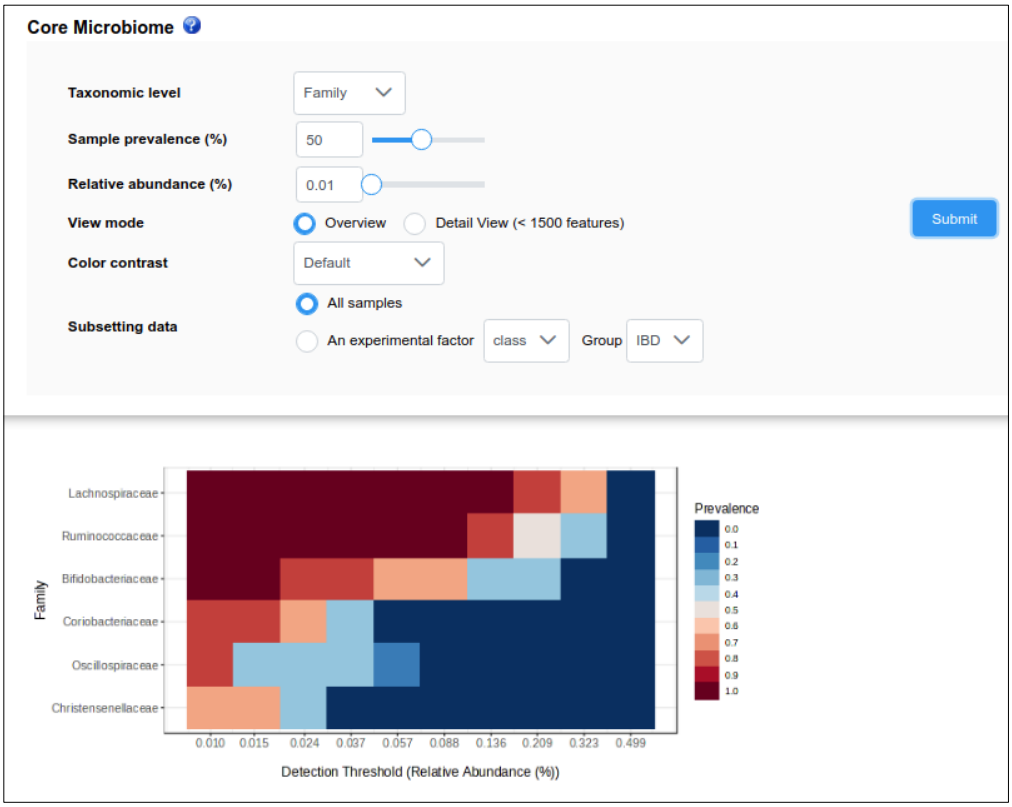

Рис. 65. Изменяемые параметры модуля “Core microbiome” и результат визуализации

Далее постройте график кор-микробиома, определив общие филы с относительной представленностью более 0,001% для 99% исследуемых образцов здоровых добровольцев. Измените цветовую палитру на “Plasma”. Изменяемые параметры и результат визуализации представлены на рисунке 66.

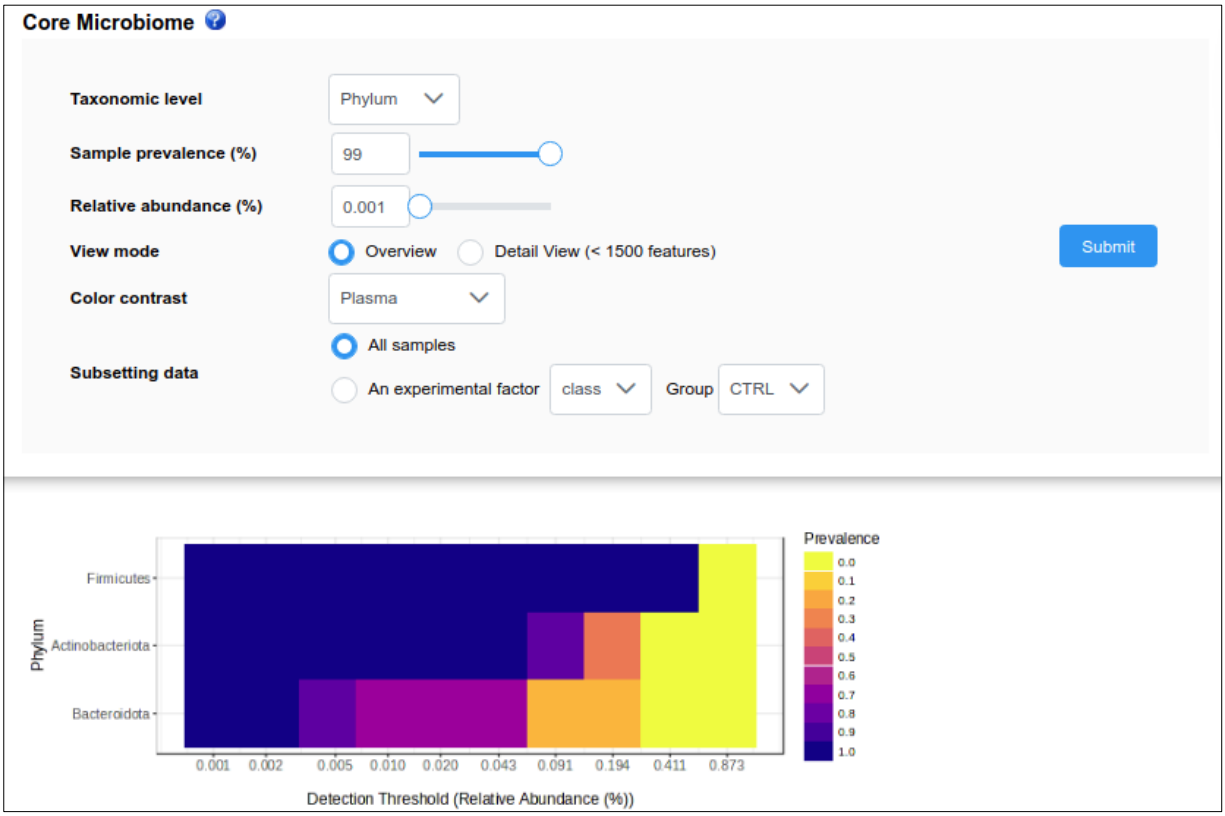

Рис. 66. Изменяемые параметры модуля “Core microbiome” и результат визуализации

3.3.3. Кластеризация и сети корреляций

1. Интерактивные тепловые карты

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Interactive heatmap” в блоке “Clustering & Correlation Network” (рис. 67).

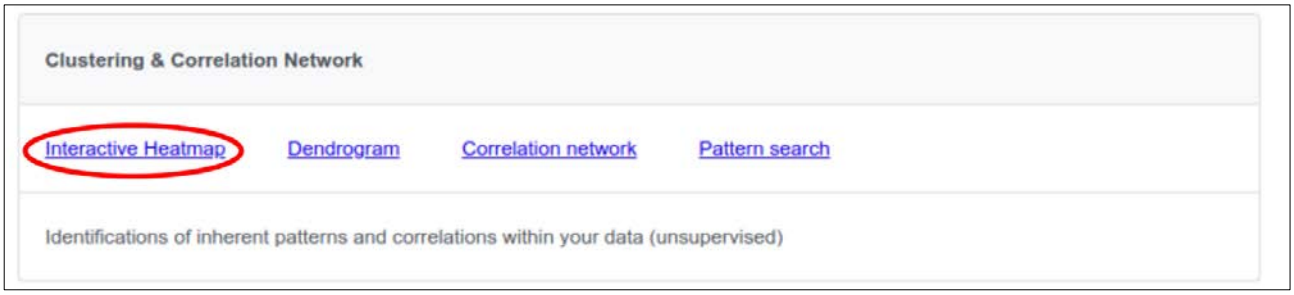

Рис. 67. Выбор модуля “Interactive heatmap”

Данный модуль позволяет отобразить одновременно представленность каждого таксона, а также иерархическую кластеризацию как образцов, так и самих таксонов. Визуально результат кластерного анализа представляется в виде дендрограмм, характеризующих дистанции

(сходство или различие) между образцами. Иерархический кластерный анализ включает в себя несколько фундаментальных этапов:

- Каждое наблюдение (каждый таксон в случае микробиоты) – отдельный кластер.
- Два соседних кластера объединяются в один (соседние = самые близкие по измеряемому расстоянию, например, по Евклидову расстоянию).
- Кластеры объединяются до тех пор, пока не останутся только два кластера.

Для построения тепловой карты по бактериальным семействам с построением дендрограммы по образцам измените параметры, указанные на рисунке 68 и нажмите клавишу “Submit”. Результат построения графика представлен на рисунке 69. Обратите внимание, что данный график интерактивный, при наведении курсора на ячейку отображается название семейства, образец и нормированная представленность. Дендрограммы рядом с названиями семейств и образцов отражают взаимосвязи, рассчитанные на основе представленности таксонов в каждом биологическом образце с использованием иерархического кластерного анализа методом Варда (применяется только вместе с Евклидовым расстоянием).

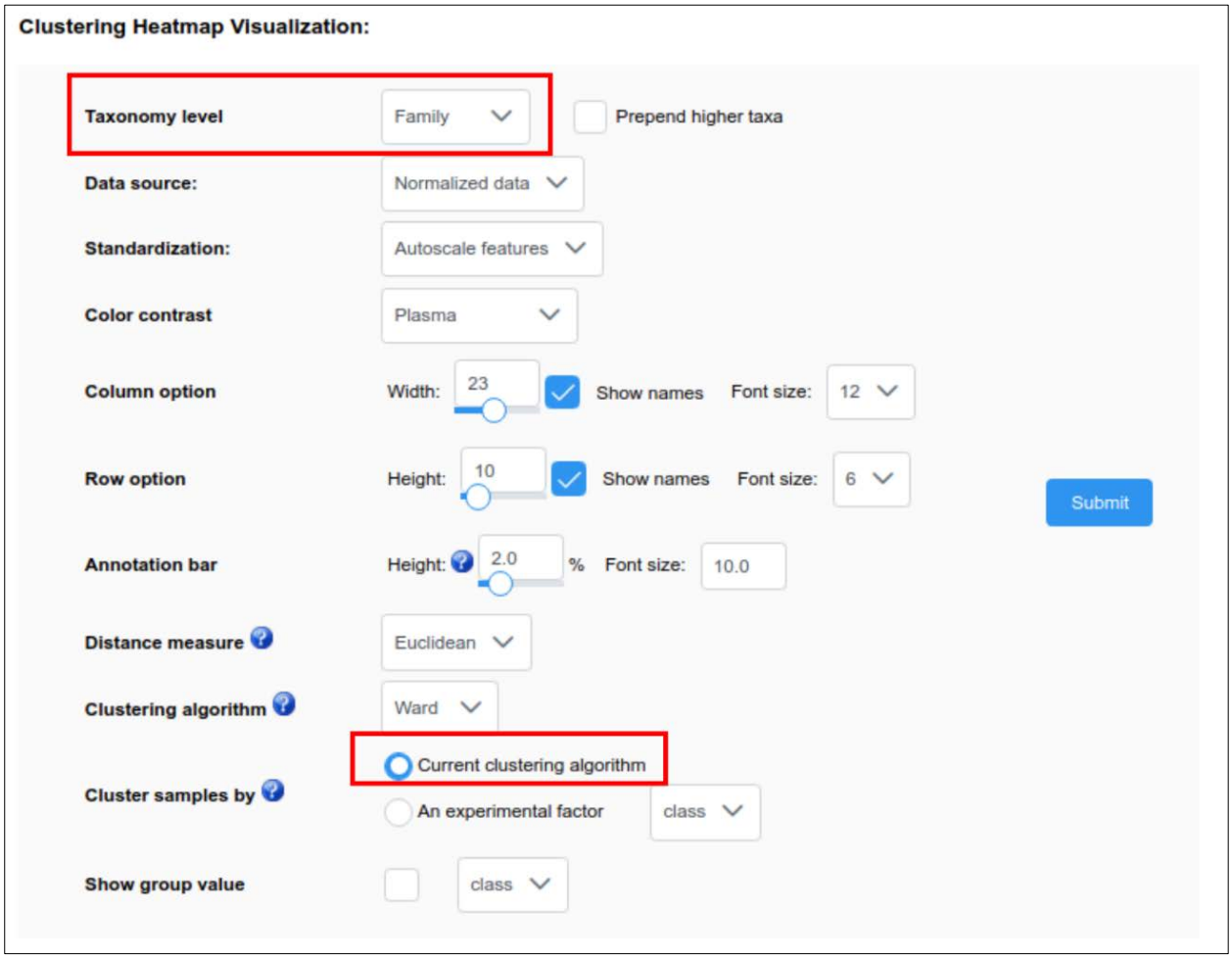

Рис. 68. Изменяемые параметры модуля “Interactive heatmap”

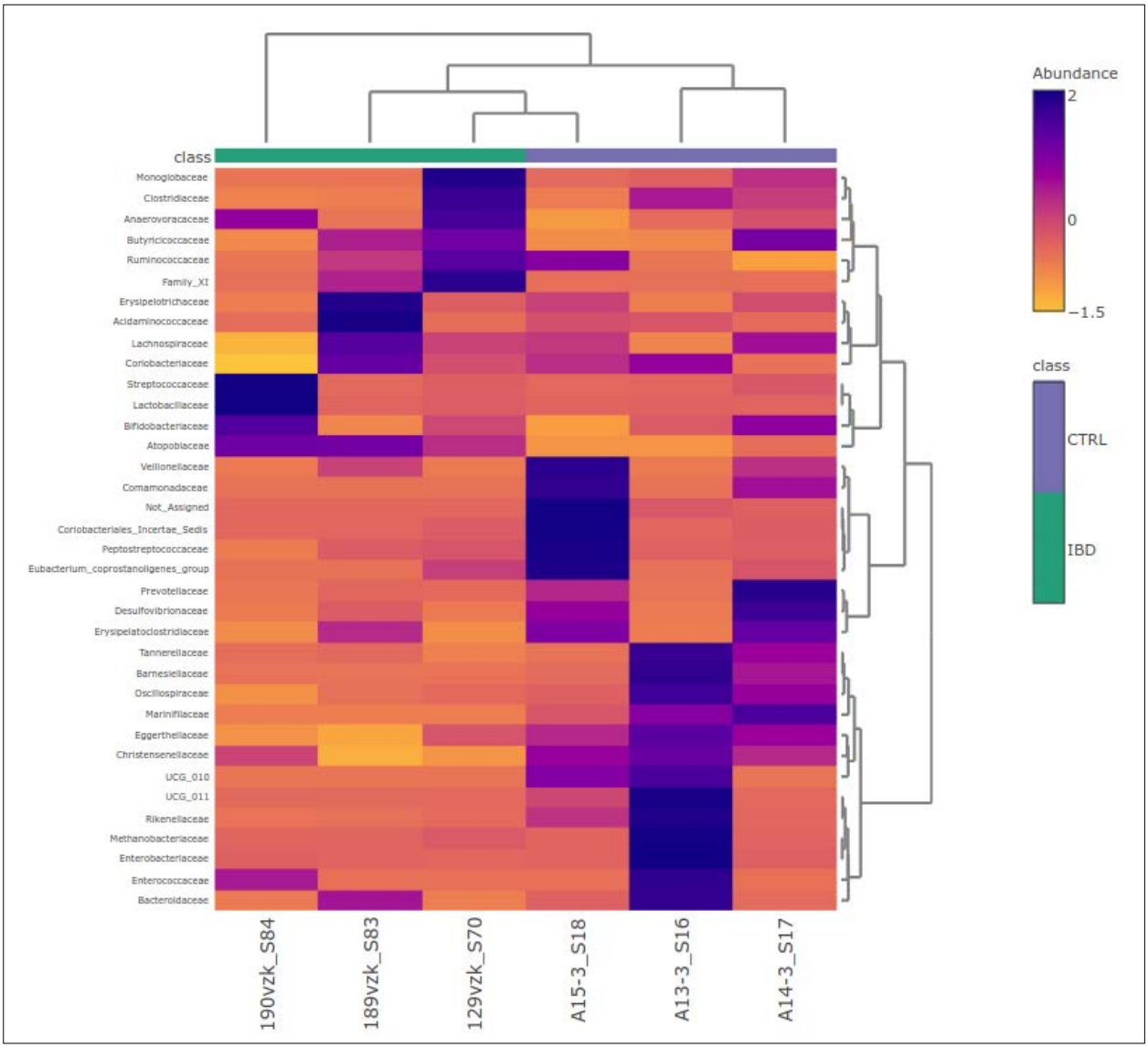

Рис. 69. Результат визуализации изменения параметров модуля “Interactive heatmap”

2. Дендрограммы

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Dendrogram” (рис. 70).

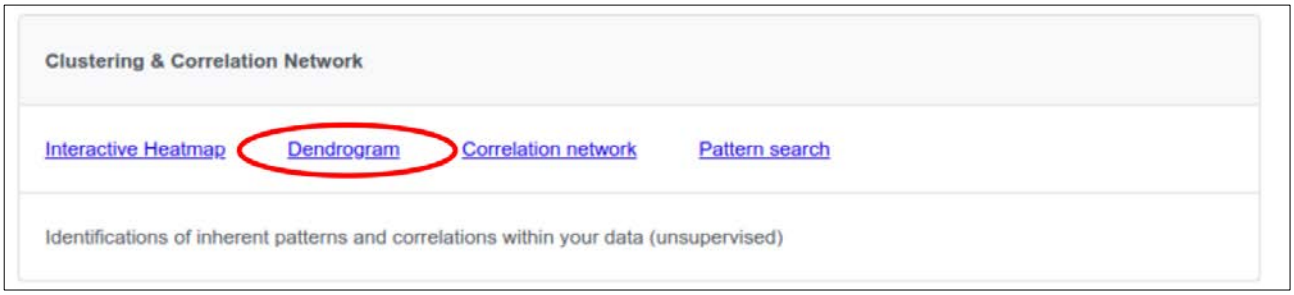

Рис. 70. Выбор модуля “Dendrogram”

Данный модуль дублирует предыдущий модуль “Interactive heatmap” по возможности проведения иерархического кластерного анализа с тем отличием, что он позволяет визуализировать дендрограммы, отражающие сходство между образцами, используя наиболее популярные в метагеномных исследованиях меры расстояний – Брея-Кёртиса, Жаккарда и Йенсена-Шеннона.

Постройте дендрограмму на основе ASV (Feature-level) с расчетом расстояния Брея-Кёртиса (параметры и результат построения отображены на рисунке 71).

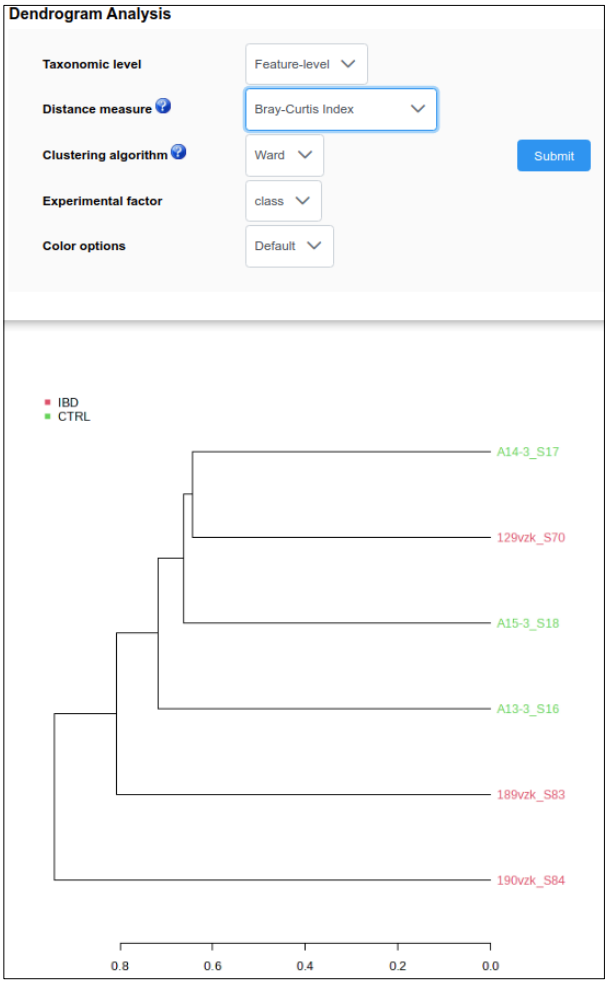

Рис. 71. Изменяемые параметры модуля “Dendrogram” и результат визуализации

Далее постройте дендрограмму на основе расстояния Йенсена-Шеннона (параметры и результат построения отображены на рисунке 72). Сравните получившиеся дендрограммы между собой и с дендрограммой на графике 69.

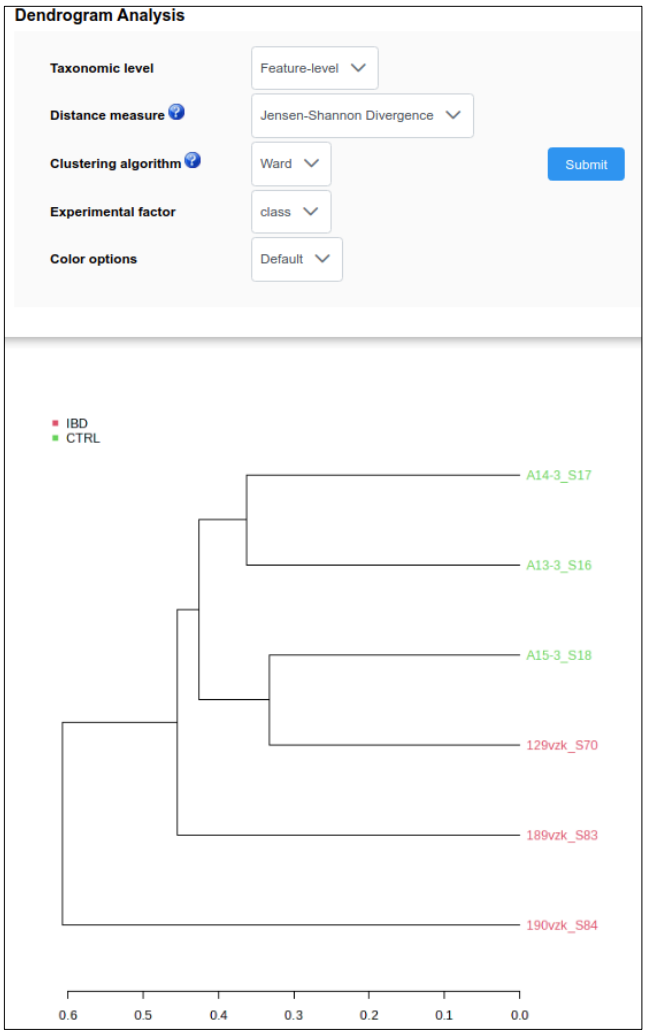

Рис. 72. Изменяемые параметры модуля “Dendrogram” и результат визуализации

3. Корреляционные сети

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Correlation network” (рис. 73).

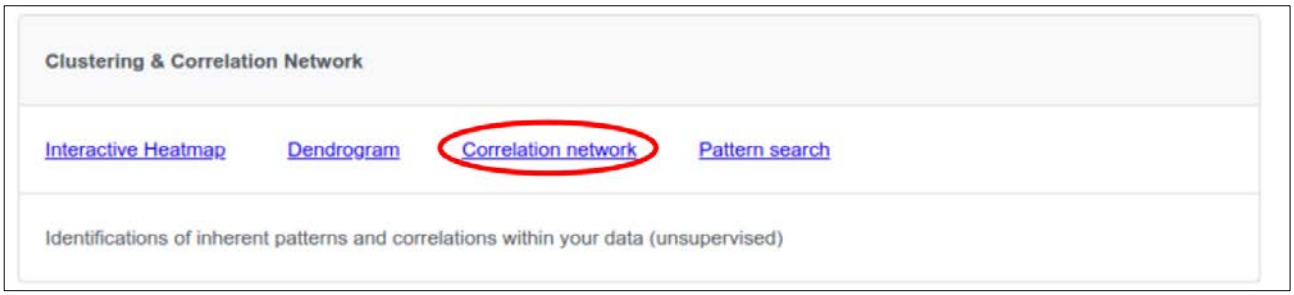

Рис. 73. Выбор модуля “Correlation network”

Данный модуль позволяет визуализировать сети корреляций между таксонами для выяснения, какие таксоны чаще встречаются

вместе (co-occurrence). Постройте сеть корреляций бактериальных семейств на основе коэффициента Спирмена с отображением относительной представленности данных семейств в виде круговых диаграмм (изменяемые параметры указаны на рисунке 74). Получившийся график отображает сеть корреляций бактериальных семейств между собой, и является интерактивным, при клике на круговую диаграмму любого семейства отображаются значения корреляций (положительные – красным цветом, отрицательные – синим) и боксплот нормированной представленности данного семейства в группах сравнения (рис. 75).

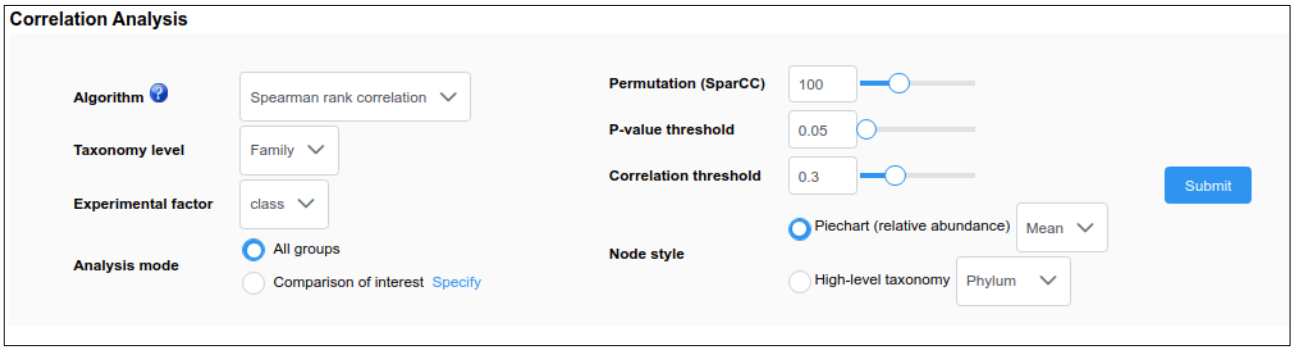

Рис. 74. Изменяемые параметры модуля “Correlation network”

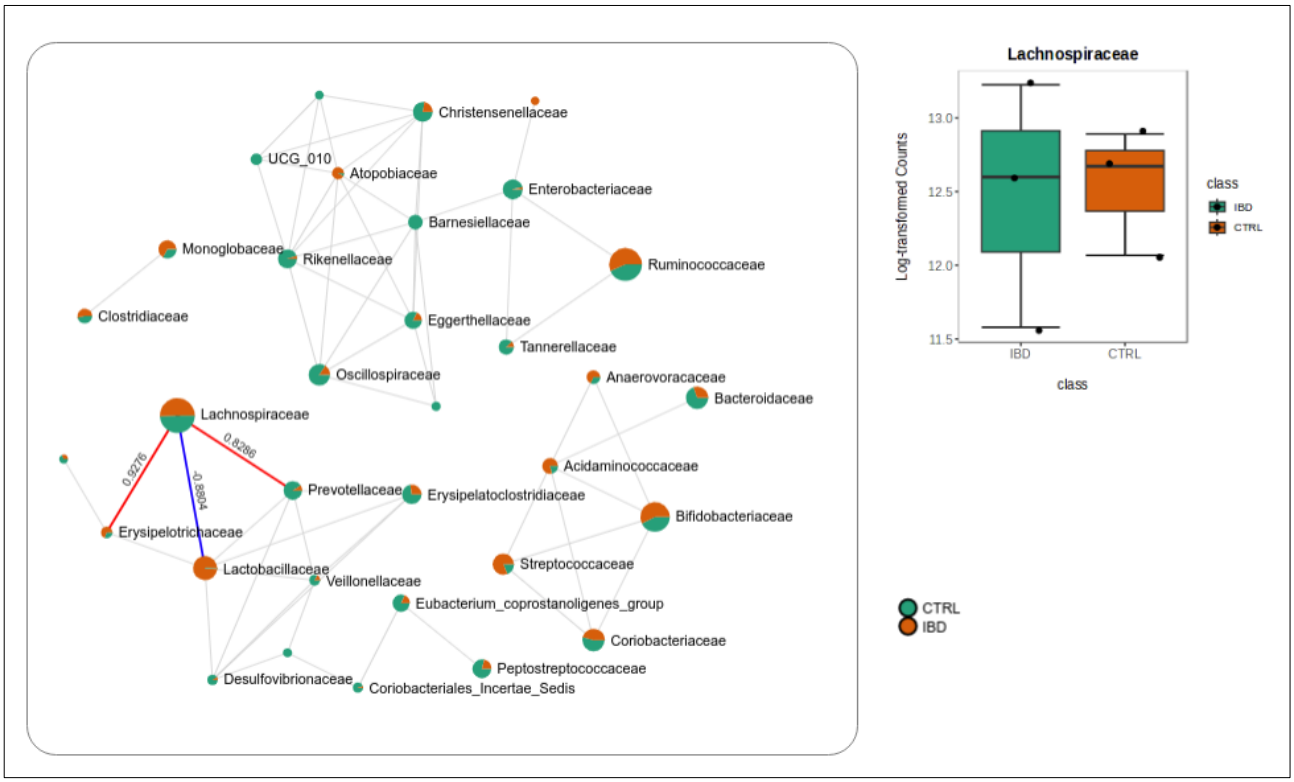

Рис. 75. Результат визуализации изменения параметров модуля “Correlation network”

4. Поиск паттернов

Вернитесь на страницу “Analysis overview”. Кликните на модуль “Pattern search” (рис. 76).

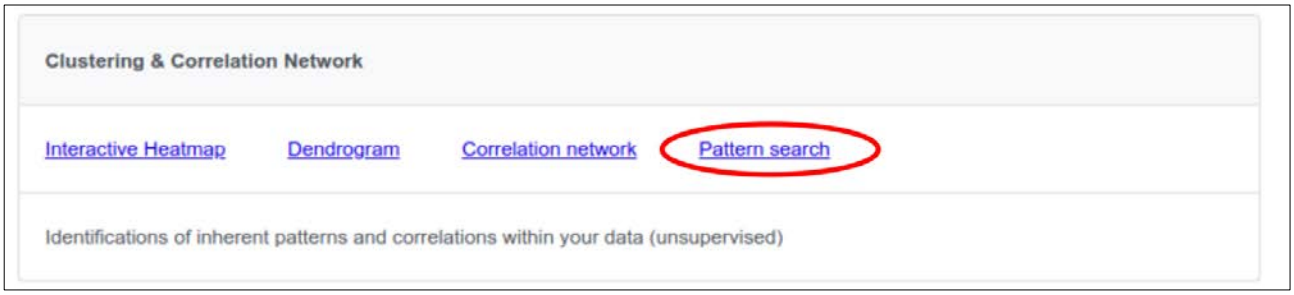

Рис. 76. Выбор модуля “Pattern search”

Данный модуль позволяет визуализировать результаты корреляционного анализа. При наличии непрерывных метаданных для исследуемых образцов (возраст, вес, уровень кальпротектина и т.п. для образцов микробиоты человека или глубина отбора пробы, pH, содержание фосфора и т.п. для образцов окружающей среды) возможно визуализировать корреляции этих показателей с относительной представленностью бактериальных таксонов. Для этого на панели изменяемых параметров нужно выбрать пункт “Continuous metadata” (рис. 77). Т.к. в учебном проекте отсутствуют такие данные, то пример визуализации не приведен. Однако данный модуль также позволяет визуализировать корреляции бактериальных таксонов друг с другом более детализировано, чем модуль “Correlation network”. Постройте график топ-25 корреляций Спирмена семейства *Bacteroidaceae* с остальными бактериальными семействами, применив параметры, указанные на рисунке 77 и нажмите клавишу “Submit”. Результат построения графика представлен на рисунке 78.

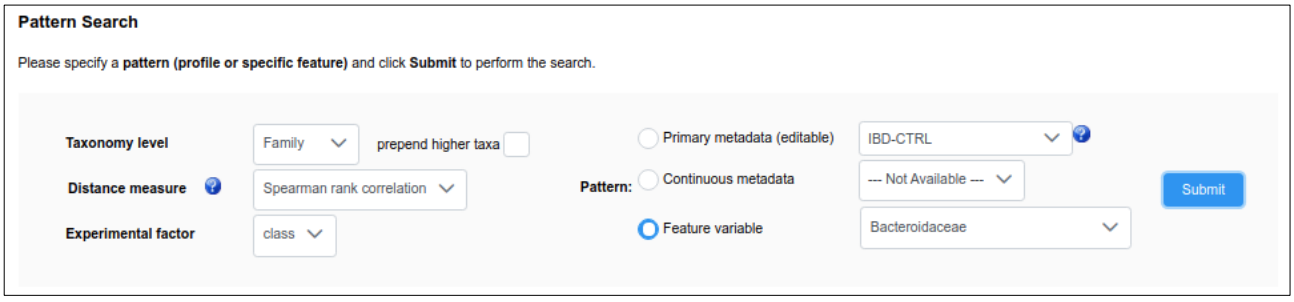

Рис. 77. Изменяемые параметры модуля “Pattern search”

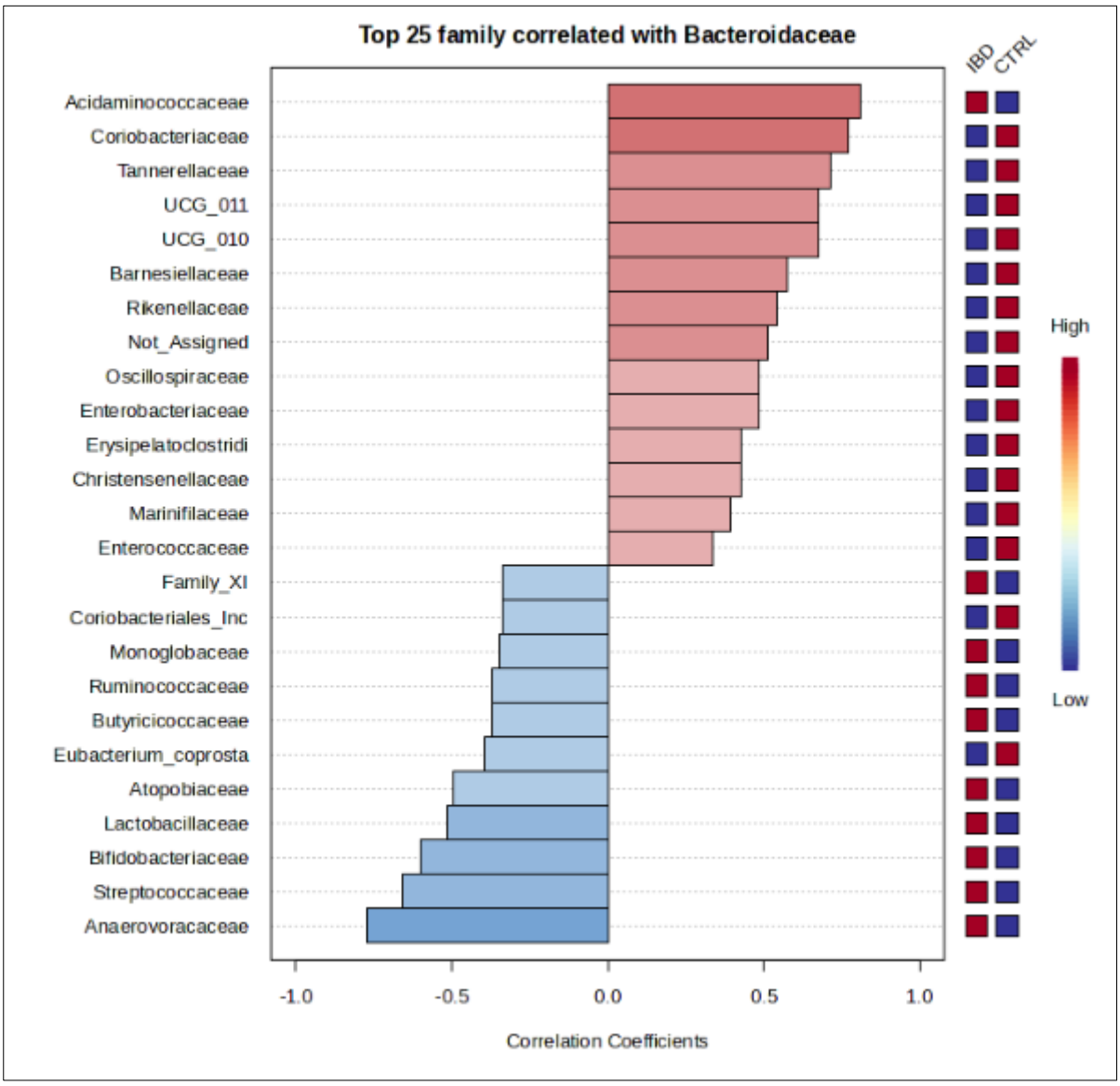

Рис. 78. Результат применения измененных параметров модуля “Pattern search”

Столбцы гистограммы с заливкой красного цвета означают положительные корреляции, а столбцы голубого цвета – отрицательные. Кроме того, справа для каждого семейства указан уровень относительной представленности в исследуемых группах сравнения (красный квадрат – в данной группе представленность семейства выше, синий квадрат – в данной группе представленность семейства ниже).

3.3.4. Сравнение и классификация

Данный модуль позволяет выявить достоверные отличия представленности бактериальных таксонов между исследуемыми группами

сравнения, а также классифицировать образцы по группам сравнения с использованием машинного обучения методом случайного леса (Random Forest).

1. Однофакторный анализ

На странице “Analysis overview” найдите блок “Comparison & Classification” и кликните на модуль “Single-factor analysis” (рис. 79). Данный модуль позволяет выявить статистически значимые отличия микробного состава исследуемых групп сравнения.

Рис. 79. Выбор модуля “Single-factor analysis”

По умолчанию данный модуль рассчитывает достоверные отличия на уровне ASV с помощью EdgeR – статистического метода, основанного на отрицательном биномиальном распределении как модели подсчета вариации. К этой же группе методов относят и DESeq2. Данные методы были разработаны для анализа транскриптомных данных, но также успешно применяются для любых типов анализа, результатами которых являются таблицы с целочисленными значениями, в том числе и для результатов метагеномных исследований. В отличие от стандартных статистических тестов (ANOVA или тест Манна-Уитни) для EdgeR и DESeq2 не требуется наличия большого числа биологических повторов (от 3 в каждой группе сравнения достаточно). В случае выполнения учебного проекта целесообразно использовать данные методы, т.к. при загрузке данных проекта не было произведено шкалирования и трансформации данных (см. п. 3.2).

Рассчитайте достоверные отличия представленности бактериальных родов в группе пациентов с воспалительными заболеваниями кишечника по сравнению со здоровыми добровольцами. Изменяемые параметры представлены на рисунке 80. Нажмите клавишу “Submit”.

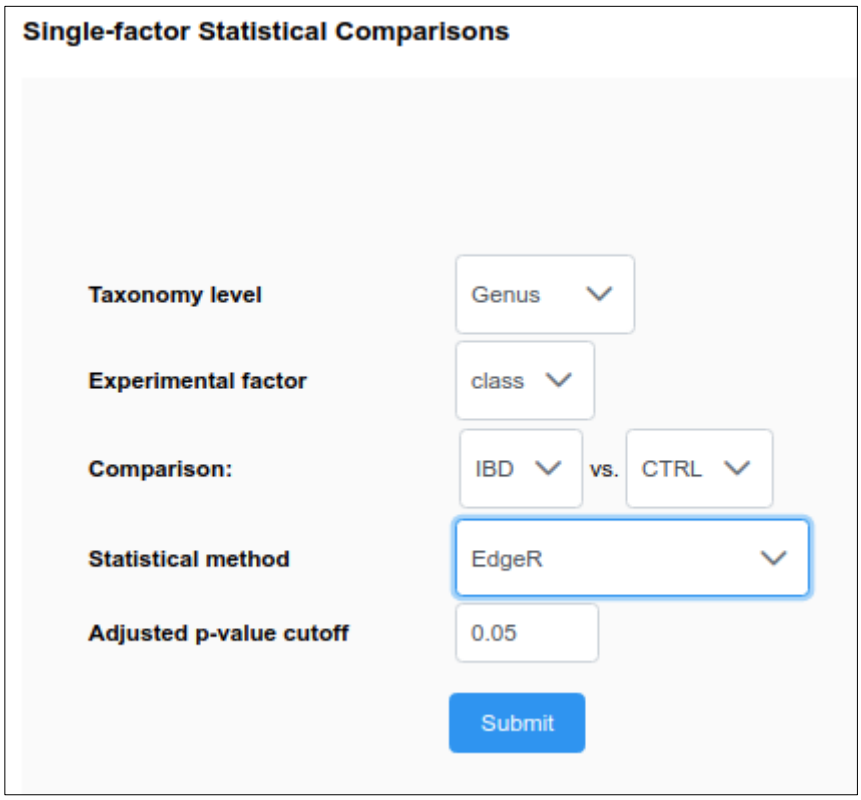

Рис. 80. Изменяемые параметры модуля “Single-factor analysis”

Результат расчета представляется в виде графика с указанием p-value (вероятности случайного определения различий) по оси y и группированием бактериальных родов по филам – по оси x (рис. 81). Значение p-value отображено в виде отрицательного десятичного логарифма, таким образом, в верхней части графика отражены самые статистически значимые результаты, цветным треугольником выделены статистически достоверные результаты (p-value с FDR < 0,05). Обратите внимание, график интерактивный, при наведении курсора на точки отображается название соответствующего бактериального рода, а при клике строится график боксплот с представленностью рода в исследуемых группах сравнения.

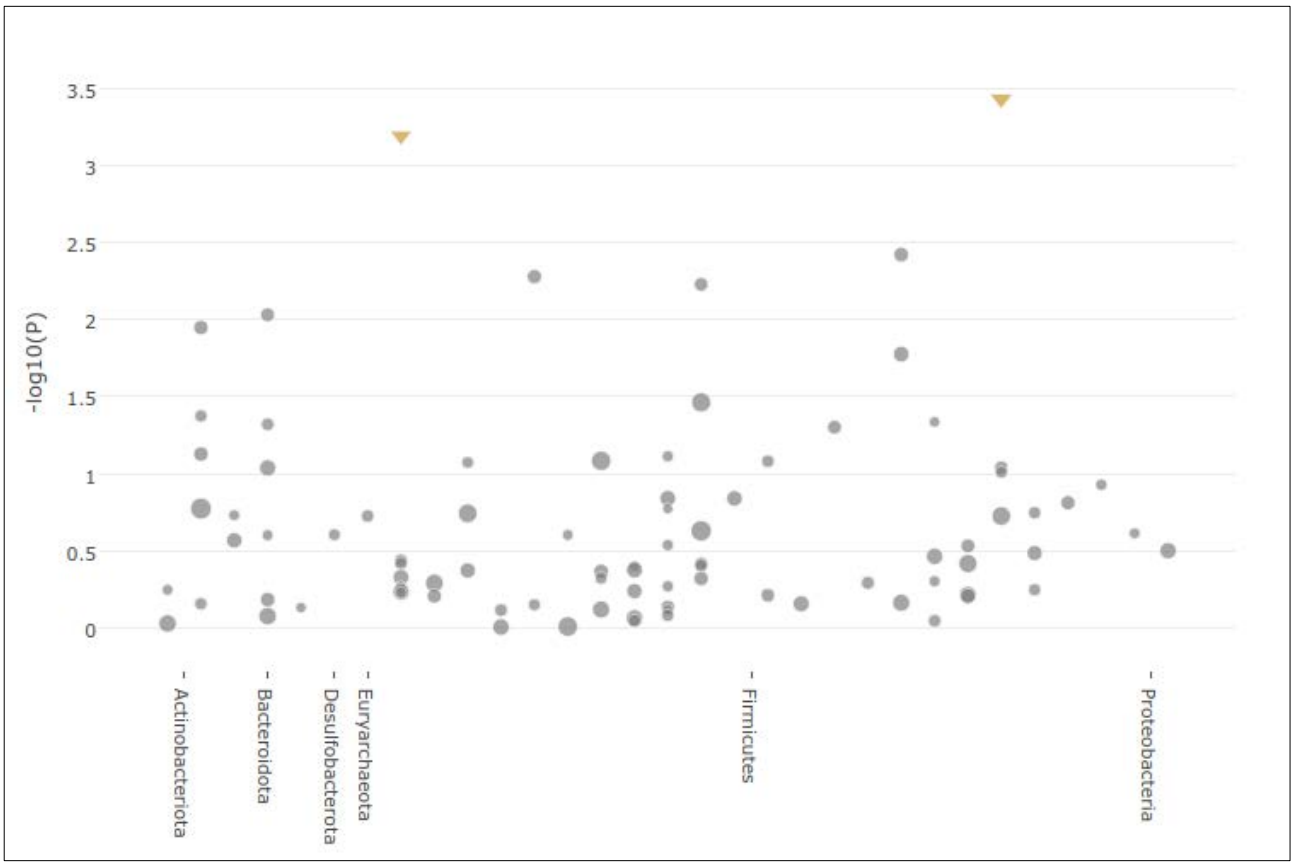

Рис. 81. Результат расчета достоверных отличий между группами сравнения с помощью EdgeR

Кроме того, результаты выявления достоверных отличий представленности бактериальных родов в группе пациентов с воспалительными заболеваниями кишечника по сравнению со здоровыми добровольцами представлены в виде таблицы (рис. 82).

Graphical Summary | Results Table

The table below shows at most 500 features ranked by their p values, with significant features highlighted in orange.

Name	log2FC	logCPM	Pvalues	FDR	View
Ligilactobacillus	13.073	16.916	3.0806E-4	0.021326	
Limosilactobacillus	12.426	16.268	4.9024E-4	0.021326	
Fournierella	8.1795	12.035	0.0028671	0.070596	
Mogibacterium	8.0412	11.898	0.0037097	0.070596	
Ruminococcus_gnavus_grc	7.3578	11.225	0.0040573	0.070596	

Рис. 82. Таблица со статистическими результатами теста EdgeR

Таблица содержит следующие поля:

- Название бактериального рода.
- $log_2FC = log_2$(Fold Change) – кратность изменения. Например, $log_2FC=3$ соответствует кратности изменения равной 8 (2^3) – в группе ВЗК представленность рода увеличивается в 8 раз по сравнению с контролем, а $log_2FC=-3$ соответствует кратности изменения равной 0.125 (2^{-3}) – группе ВЗК представленность рода снижается в 8 раз по сравнению с контролем.
- logCPM – количественно характеризует представленность каждого бактериального рода, является десятичным логарифмом от количества ридов, нормализованного на миллион. Чем выше значение logCPM, тем более представлен данный род в исследуемом проекте.
- Pvalues – значения p-value.
- FDR – значения p-value с поправкой на множественную проверку гипотез методом false discovery rate (средняя доля ложноположительных результатов). В метагеномных исследованиях мы анализируем большое количество показателей (бактериальных родов в данном случае), поэтому часть достоверных отличий может быть обнаружена ошибочно, поэтому необходимо использовать данную поправку, чтобы отбросить часть статистически значимых результатов как ложноположительные. Достоверно отличающиеся рода выделены оранжевой заливкой.
- View – кнопка построения боксплотов для каждого бактериального рода (рис. 83).

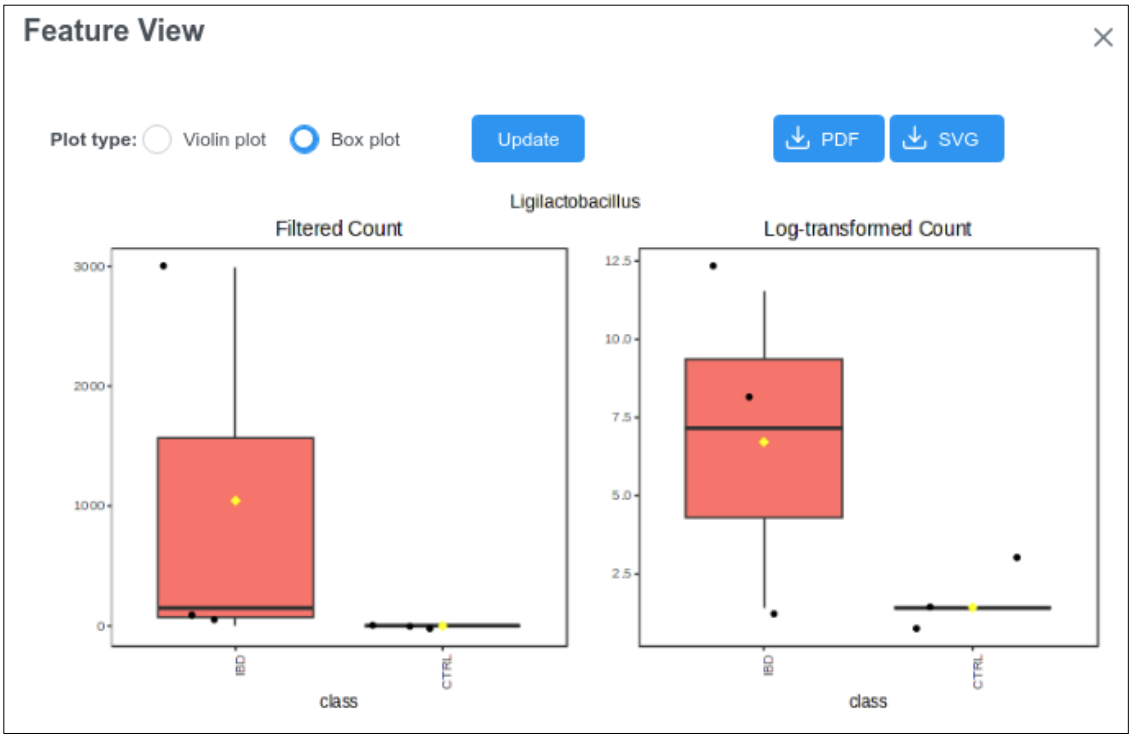

Рис. 83. Боксплоты представленности бактериального рода *Ligilactobacillus* в исследуемых группах сравнения

Произведите аналогичные расчеты достоверных отличий представленности бактериальных родов методом DESeq2 (рис. 84). Нажмите клавишу “Submit”.

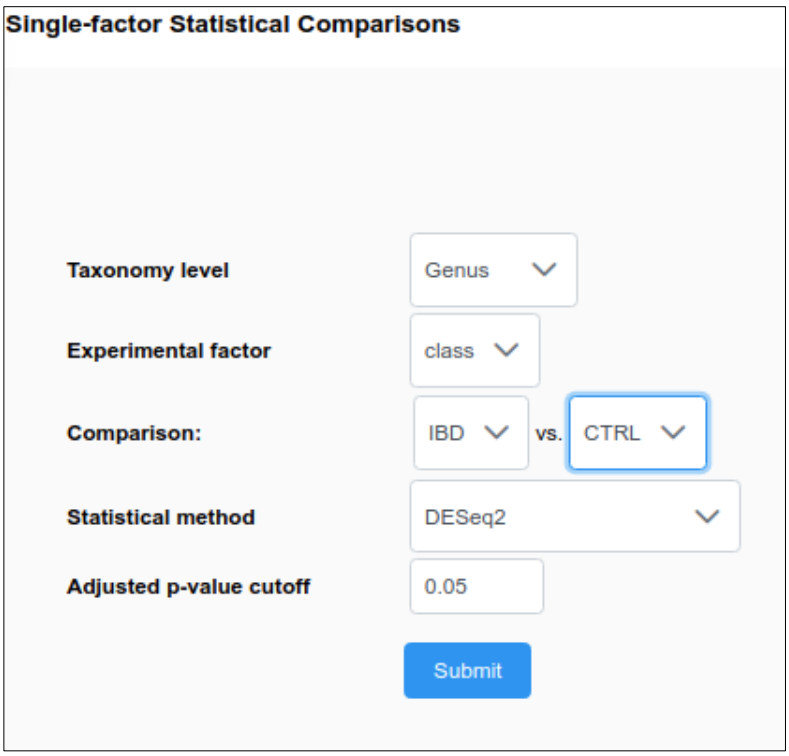

Рис. 84. Изменяемые параметры модуля “Single-factor analysis”

Графические и табличные результаты представлены на рисунках 85 и 86, соответственно. Поля в таблице отличаются от результатов EdgeR наличием колонки “lfcSE” – это значения стандартной ошибки среднего для log_2FC. Обратите внимание, что на рисунке 85 бирюзовым треугольником, направленным вниз, обозначается род *Barnesiella*, представленность которого достоверно снижена в группе ВЗК, а желтым треугольником, направленным вверх, – рода *Ligilactobacillus* и *Limosilactobacillus*, представленность которых статистически значимо повышена в группе пациентов с ВЗК. Направленность изменений можно уточнить в колонке log_2FC на рисунке 86 (положительные или отрицательные значения), а также с помощью построения боксплотов по клику на кнопку “View”.

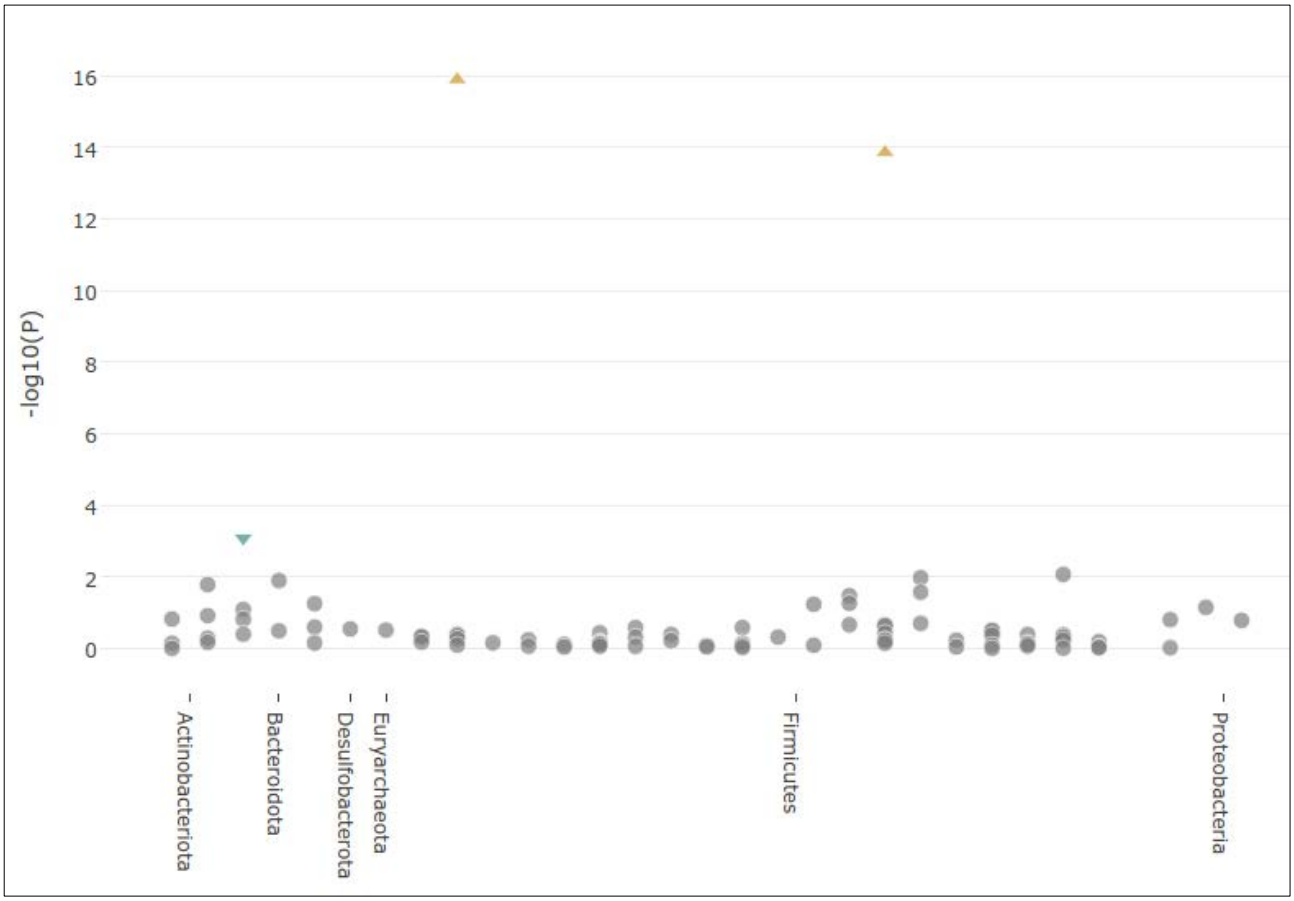

Рис. 85. Результат расчета достоверных отличий между группами сравнения с помощью DESeq2

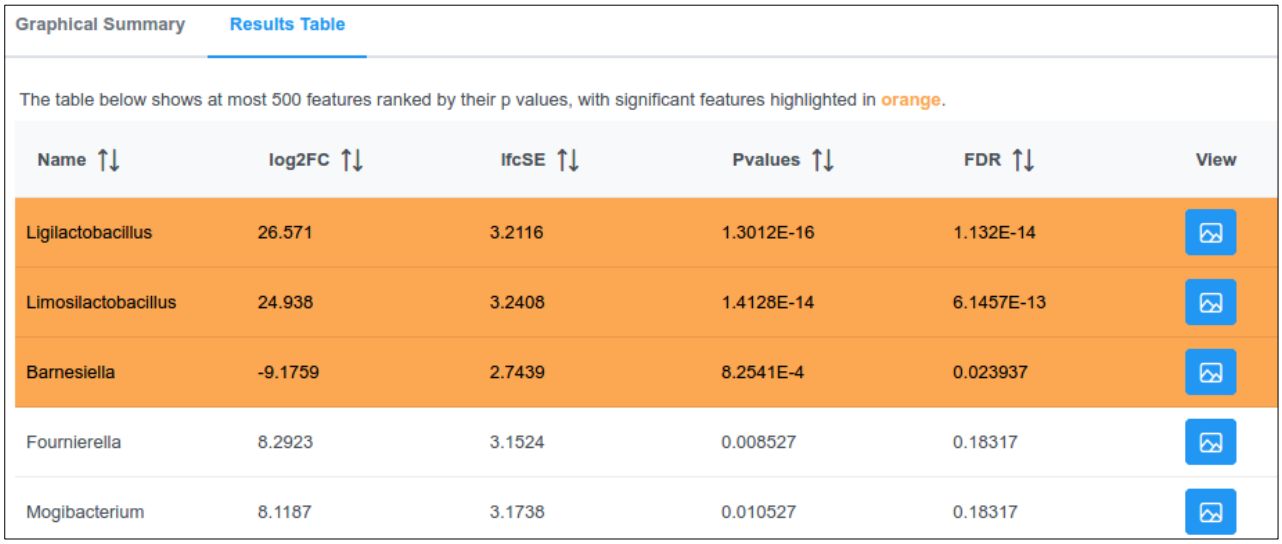

Рис. 86. Таблица со статистическими результатами теста DESeq2

Произведите аналогичные расчеты достоверных отличий представленности бактериальных родов с помощью классического непараметрического теста Манна-Уитни (рис. 87). Нажмите клавишу “Submit”.

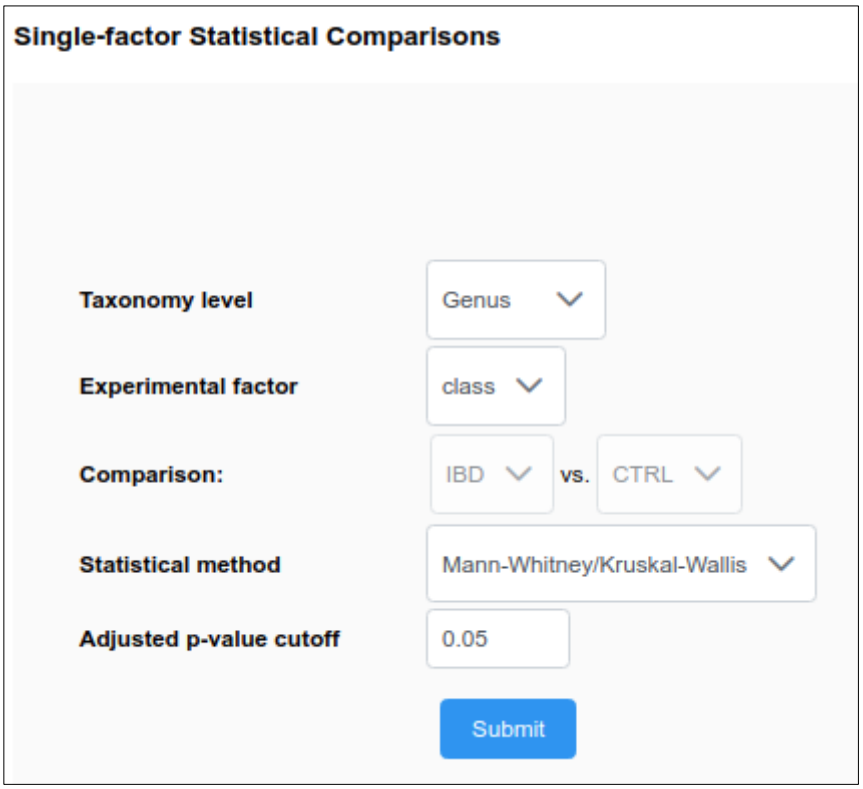

Рис. 87. Изменяемые параметры модуля “Single-factor analysis”

Результаты расчетов представлены в виде табличных результатов на рисунке 88. Очевидно, что достоверных отличий данным методом не было выявлено. Целесообразно использовать классические методы при наличии более чем 10 биологических повторов и при загрузке необходима трансформация данных (см. п. 3.2).

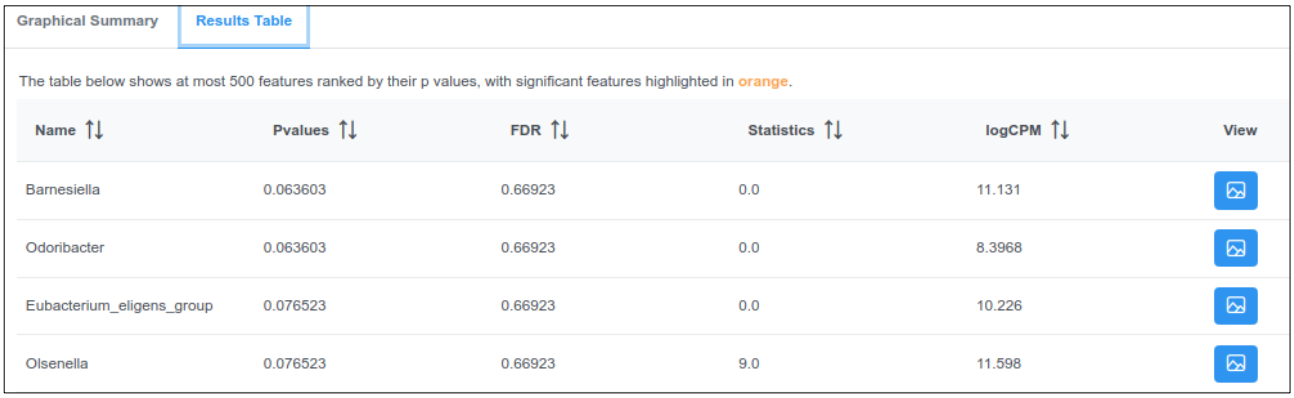

Рис. 88. Таблица со статистическими результатами теста Манна-Уитни

2. Многофакторный анализ

При наличии нескольких классов метаданных, например, диагноз и пол людей, загрязненность почвы и ее тип, водоем и сезон отбора пробы, и т.п., необходимо использовать многофакторный анализ, позволяющий учесть все факторы, которые могут повлиять на таксономический состав исследуемых образцов. На странице “Analysis overview” и кликните на модуль “Multi-factor analysis” (рис. 89).

Рис. 89. Выбор модуля “Multi-factor analysis”

Учебный проект содержит только 1 тип данных (диагноз), поэтому подробная характеристика модуля не представлена.

3. LEfSe

Метод LEfSe – Linear discriminant analysis Effect Size разработан специально для выявления таксонов, которые объясняют вариабельность состава сообщества в исследуемых группах сравнения [Segata *et al.*, 2011]. Проще говоря, данный метод позволяет выявить биомаркеры для исследуемых групп, а также количественно оценить вклад каждого биомаркера в определение принадлежности образца к группе сравнения.

На странице “Analysis overview” и кликните на модуль “LEfSe” (рис. 90).

Рис. 90. Выбор модуля “LEfSe”

На данных учебного проекта выявите бактериальные рода, которые могут являться биомаркерами кишечной микробиоты пациентов с воспалительными заболеваниями кишечника. Для этого примените параметры, указанные на рисунке 91. Обратите внимание, что P-value cutoff отмечен как “original” исключительно в образовательных целях. При анализе данных реального проекта необходимо использовать исключительно “FDR-adjusted”.

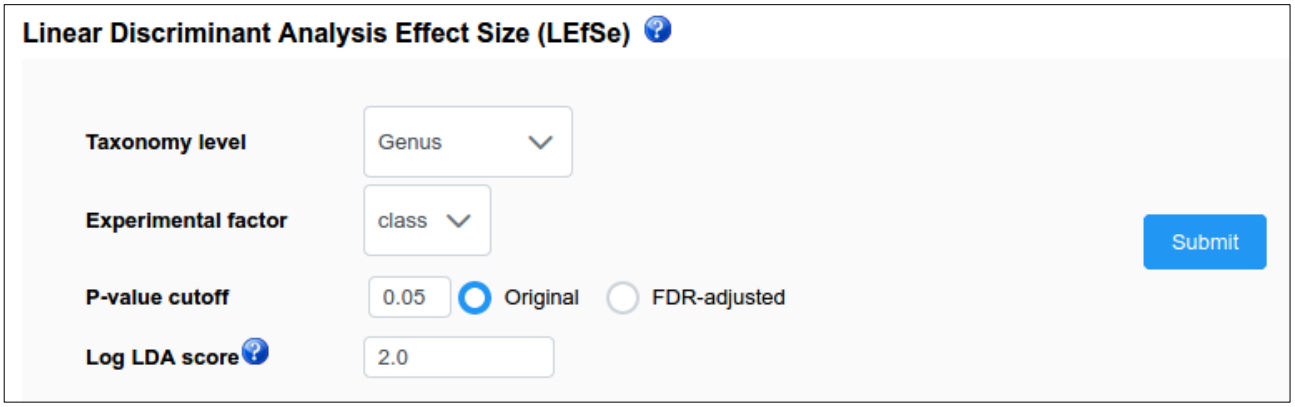

Рис. 91. Изменяемые параметры модуля “LEfSe”

Результат анализа LEfSe представлен в виде точечного графика для 9 достоверных родов-биомаркеров (рис. 92). По оси x отражены значения LDA – чем выше значение, тем большей мощностью (магнитудой) для отнесения образца к группе пациентов с ВЗК обладает данный род как биомаркер. Чем значение ниже (более отрицательное), тем большей мощностью для отнесения образца к группе здоровых добровольцев обладает данный род как биомаркер. Справа от рисунка присутствует тепловая карта, помогающая сделать вывод об увеличении/снижении каждого рода в исследуемых группах сравнения.

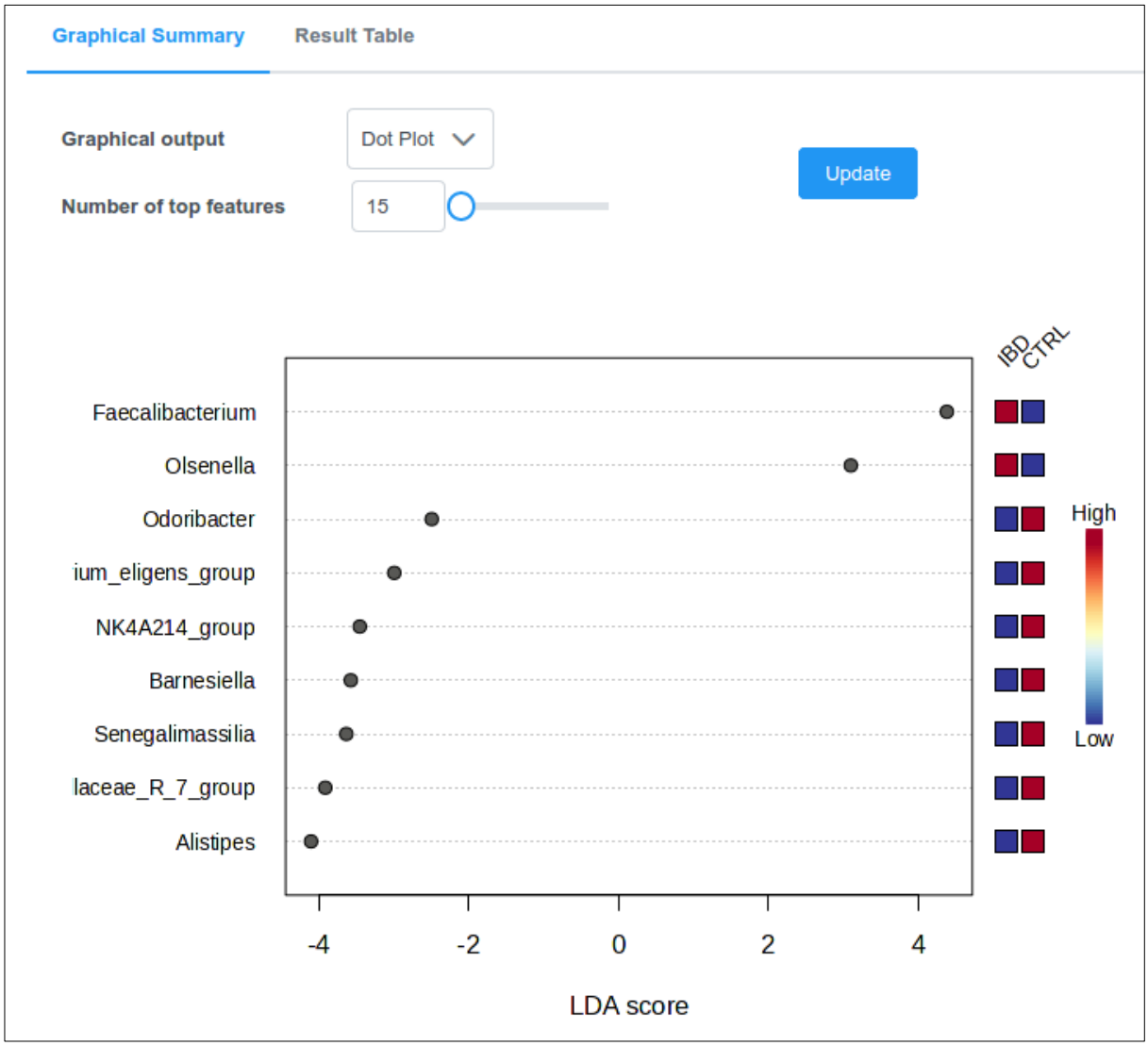

Рис. 92. Результат анализа LEfSe, точечный график

Для визуального разделения биомаркеров для разных групп сравнения можно построить гистограмму, выбрав параметр “Bar Plot” в пункте “Graphical output” (рис. 93). Такой тип построения графика

позволяет интуитивно различить биомаркеры для пациентов с ВЗК и здоровых добровольцев.

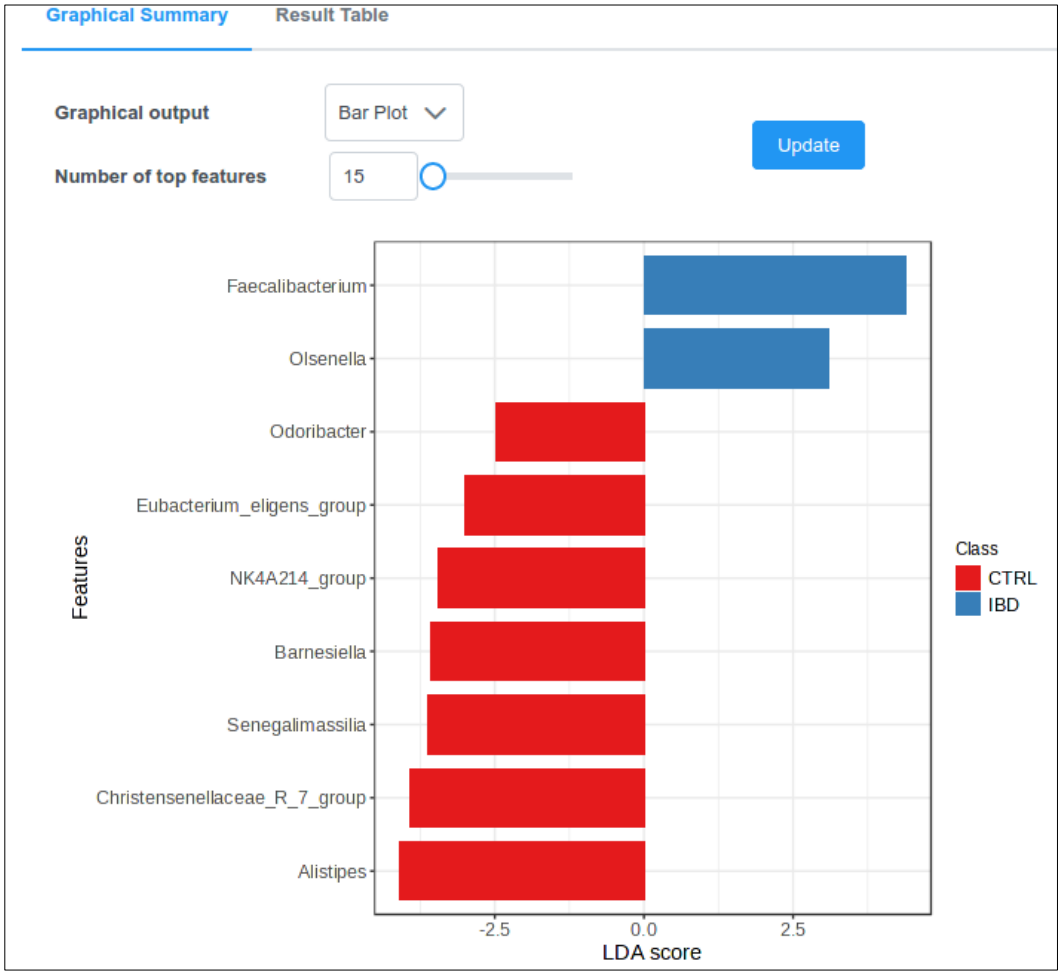

Рис. 93. Результат анализа LEfSe, гистограммы

Кроме того, результаты анализа LEfSe представлены в виде табличных данных, где представлены уже описанные поля Pvalues, FDR и LDA, а также представленность каждого рода в исследуемых группах сравнения (рис. 94). Представленность любого бактериального рода может быть отображена с помощью графика боксплот по клику на клавишу “View” (рис. 95).

Graphical Summary | Result Table

The table below shows at most 500 features ranked by their p values, with significant features highlighted in orange.

Name	Pvalues	FDR	IBD	CTRL	LDAscore	View
Odoribacter	0.036904	0.42372	0.0	609.16	-2.49	
Barnesiella	0.036904	0.42372	0.0	7449.2	-3.57	
Senegalimassilia	0.046302	0.42372	522.14	9120.0	-3.63	
Olsenella	0.046302	0.42372	2749.9	226.26	3.1	
Eubacterium_eligens_group	0.046302	0.42372	487.33	2436.6	-2.99	
Christensenellaceae_R_7_g	0.049535	0.42372	7397.0	23775.0	-3.91	
Alistipes	0.049535	0.42372	1409.8	26385.0	-4.1	
Faecalibacterium	0.049535	0.42372	128060.0	80305.0	4.38	
NK4A214_group	0.049535	0.42372	2106.0	7745.1	-3.45	
Anaerococcus	0.12134	0.42372	243.66	0.0	2.09	

Рис. 94. Результат анализа LEfSe, табличные данные

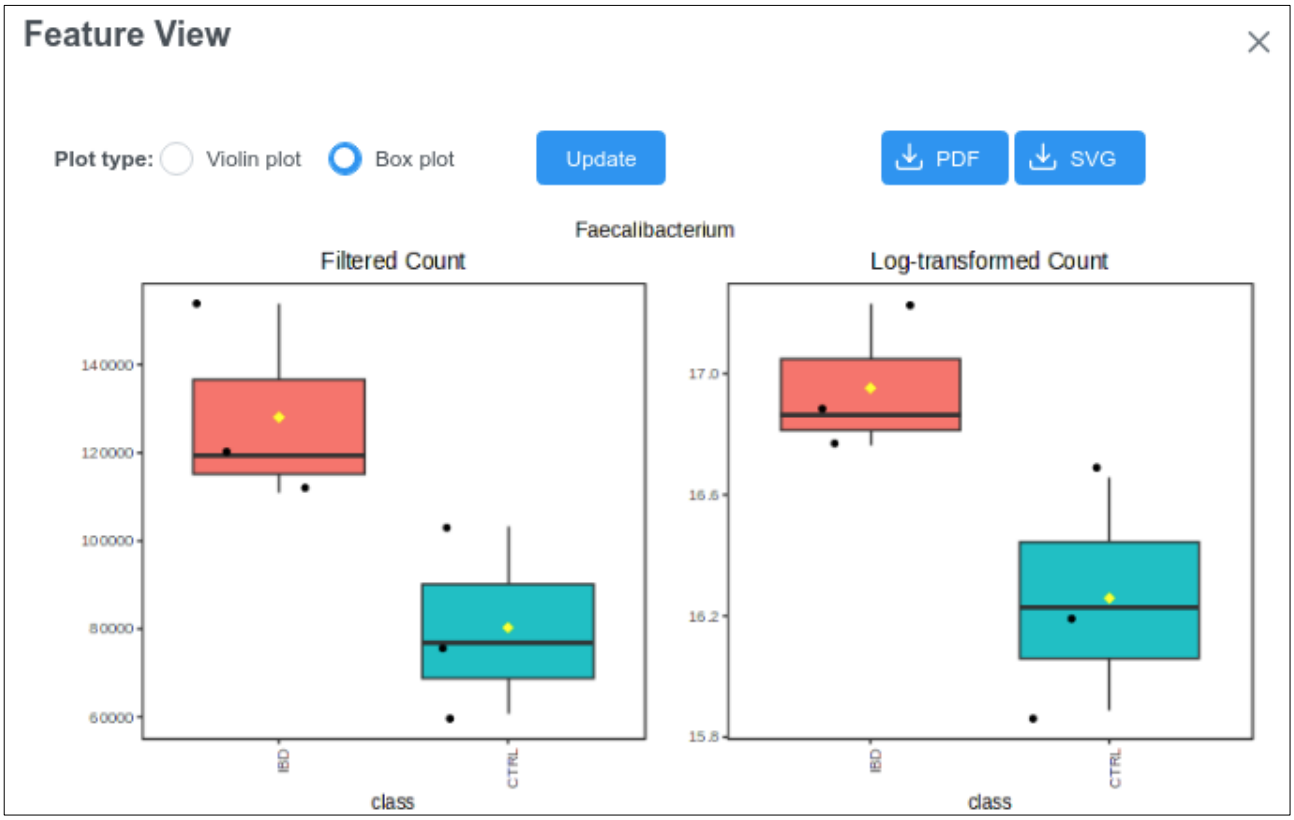

Рис. 95. Боксплоты представленности бактериального рода в исследуемых группах сравнения

4. Случайный лес

Метод машинного обучения Random Forest (случайный лес) используется для классификации образцов на основе измерения множества параметров. Принцип метода заключается в построении большого количества деревьев решений (Decision tree, классический метод машинного обучения для классификации), и выборе оптимального дерева на их основе. В случае выполнения учебного проекта данный метод позволяет выявить бактериальные таксоны (биомаркеры), предсказывающие к какой группе сравнения относится конкретный образец. Это позволяет создавать диагностические системы – классификаторы, способные относить уже новые образцы к группам сравнения – пациентам с ВЗК или здоровым добровольцам. Метод применим на данных с большим количеством биологических образцов (как и все методы машинного обучения, желательно иметь не менее сотни образцов, а лучше несколько тысяч), однако в образовательных целях будем использовать имеющиеся данные.

На странице “Analysis overview” кликните на модуль “Random Forest” (рис. 96).

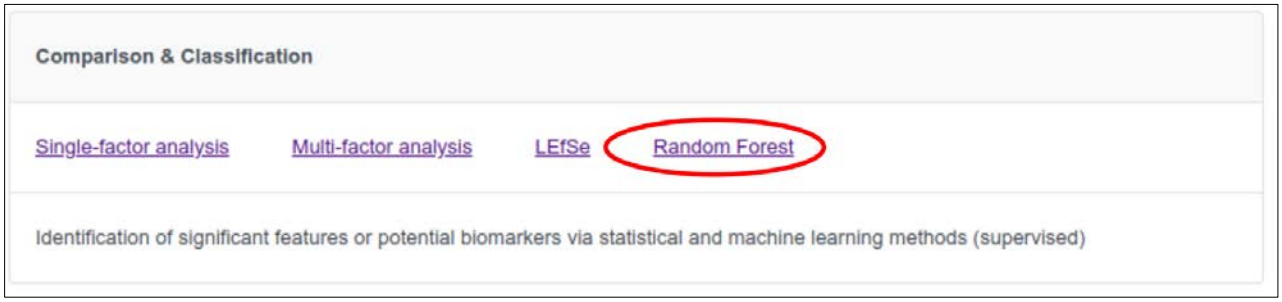

Рис. 96. Выбор модуля “Random Forest”

Постройте классификатор, который позволит отделить пациентов с ВЗК от здоровых добровольцев на основе представленности бактериальных семейств. Примените параметры, отмеченные на рисунке 97, количество деревьев (Number of trees to grow) можете выбрать на свое усмотрение. Кликните “Submit”.

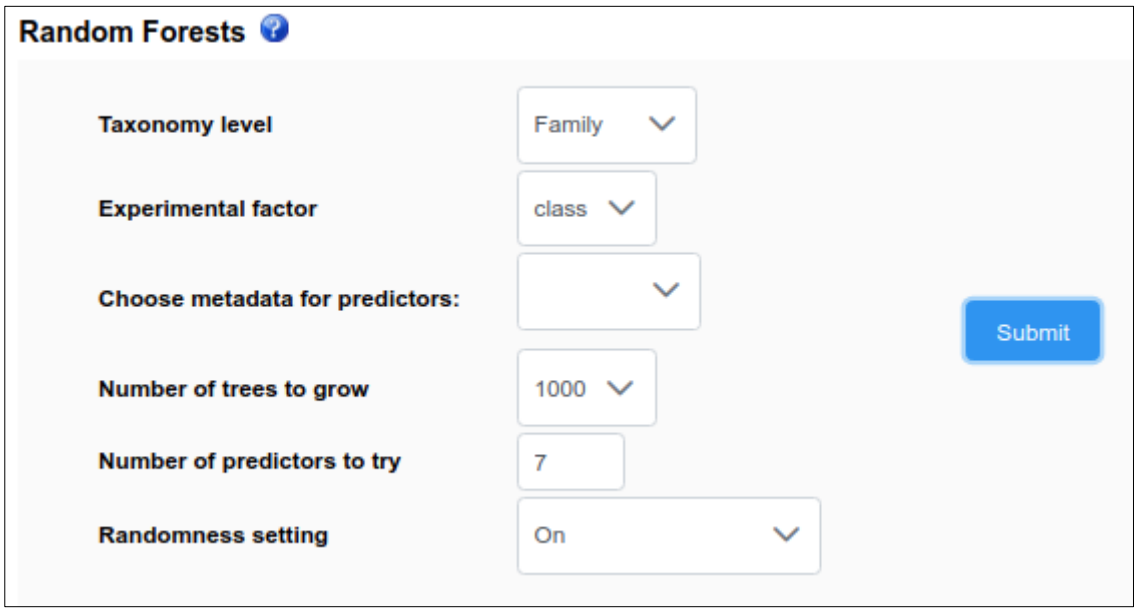

Рис. 97. Изменяемые параметры модуля “Random Forest”

Результатом анализа является график и таблица – матрица ошибок (confusion matrix) (рис. 98). На графике отображены уровни ошибок для каждого дерева решений. Основным результатом является матрица ошибок, которая отображает насколько точно полученная модель классифицирует образцы. Матрица включает в себя данные об ошибке классификатора для определения разных групп, из которых можно рассчитать следующие важные для диагностических тестов показатели:

- чувствительность – вероятность, что тест будет позитивен, если болезнь есть
- специфичность – вероятность, что тест будет негативен, если болезни нет

Обратите внимание, что полученные вами результаты, вероятно, не совпадут с результатами, представленными на рисунке 98, т.к. с каждым нажатием “Submit” строятся разные деревья и финальная модель будет разная. В данном пособии отражена модель с ошибками классификатора равными 0 (чувствительность и специфичность теста равны 100%). Постарайтесь построить модель, со схожим уровнем ошибок (кликнете “Submit” столько раз, сколько потребуется).

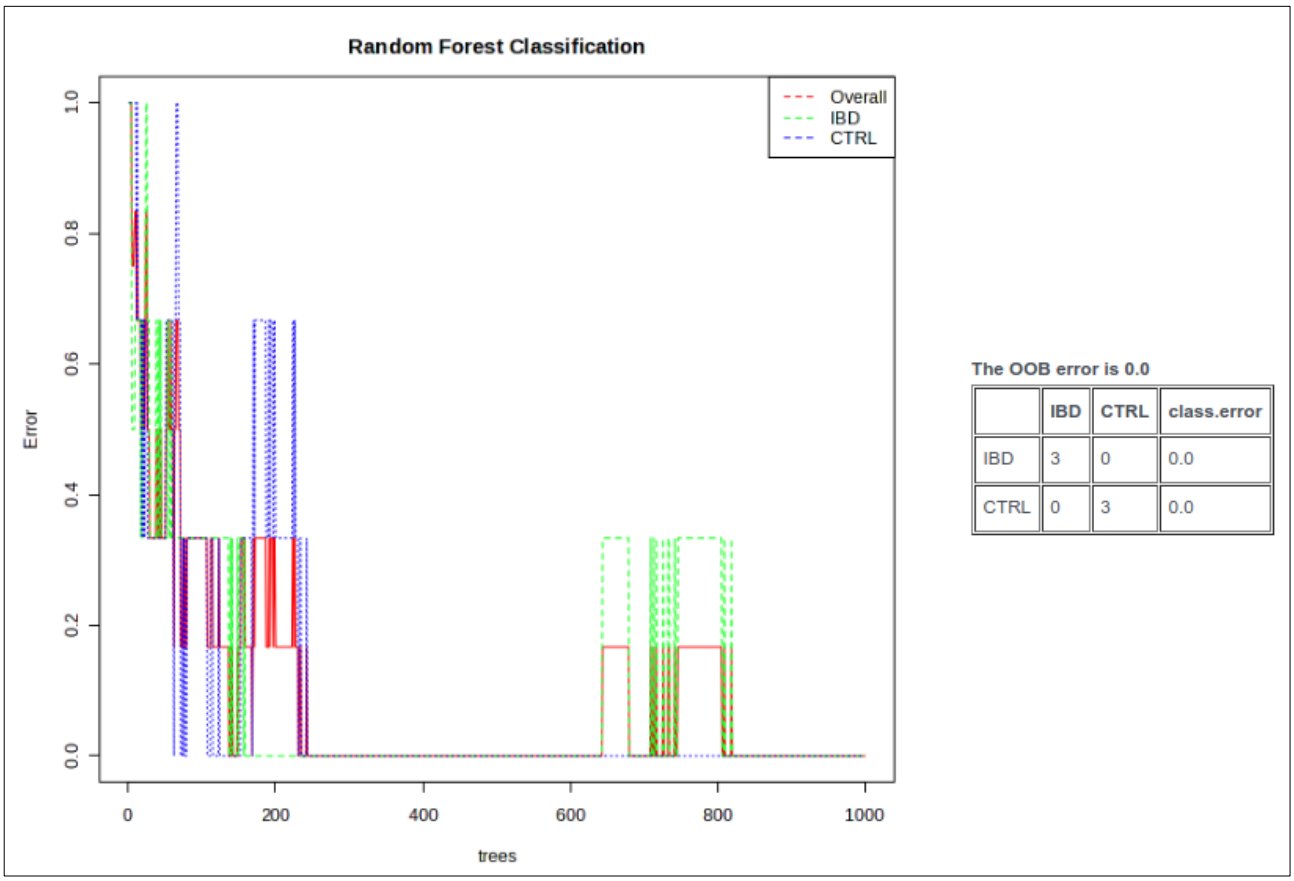

Рис. 98. Результат анализа “Random Forest”

Кроме того, можно отразить семейства-биомаркеры, кликнув на “Important Features”, причем можно отразить любое разумное их количество, изменив пункт “Top features” (рис. 99). Справа от рисунка присутствует тепловая карта, помогающая сделать вывод об увеличении/снижении каждого рода в исследуемых группах сравнения. Также обратите внимание, что ваши результаты могут не совпадать с представленными в настоящем пособии, т.к. построенные модели могут отличаться.

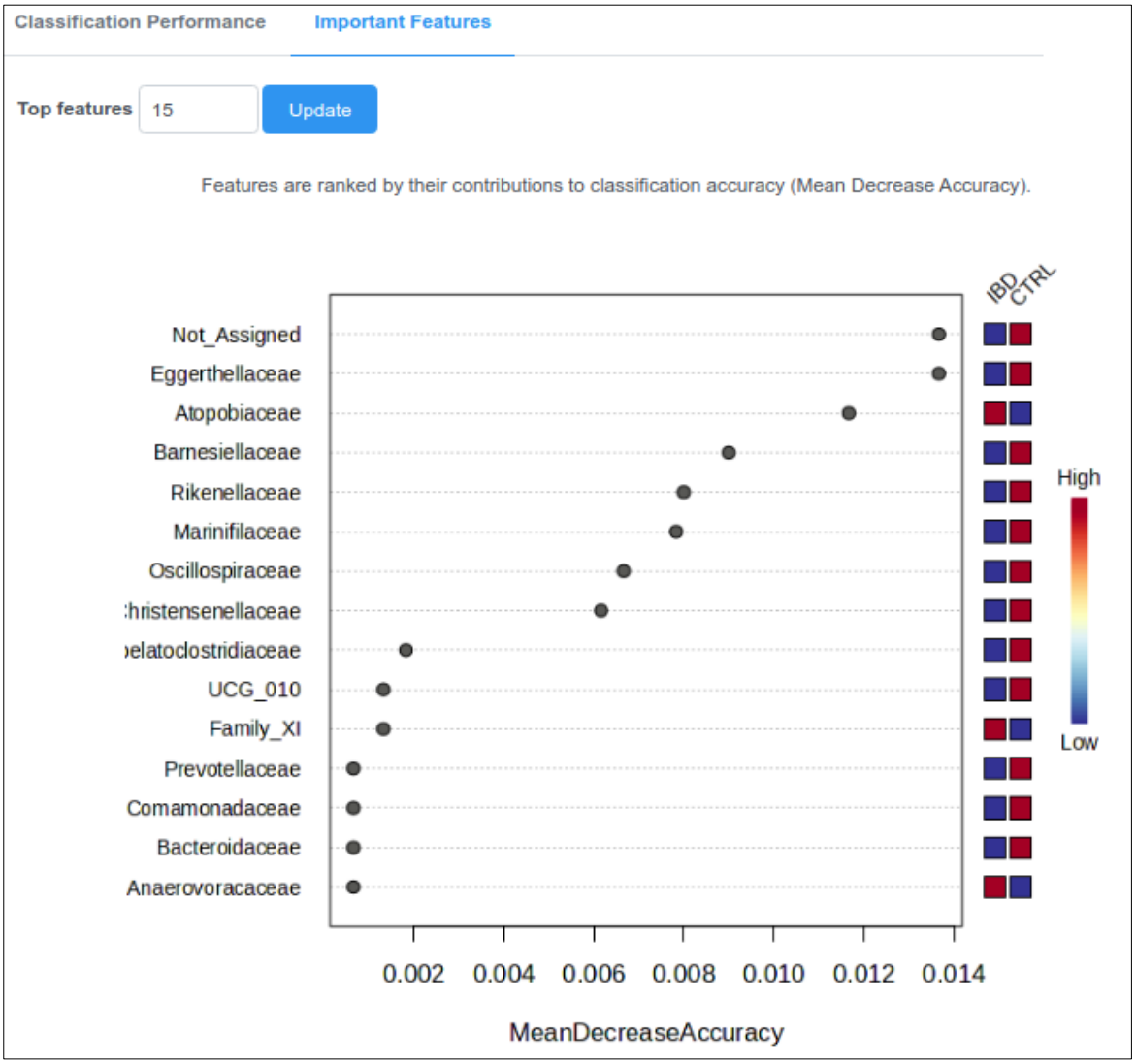

Рис. 99. Семейства-биомаркеры, выявленные с помощью модуля “Random Forest”

3.3.5. Сохранение результатов

Если не сохраняли все результаты проведенных анализов по мере их создания, web-платформа MicrobiomeAnalyst позволяет сохранить все результаты одним архивом. Для этого найдите вверху страницы кнопку “Download” и кликните на нее (рис. 100). Скачайте на персональный компьютер все файлы одним архивом кликнув на “Download.zip” (по аналогии с рис. 20).

Рис. 100. Модуль скачивания результатов

ЗАКЛЮЧЕНИЕ

Данное учебное пособие содержит общее описание методов метагеномного анализа, а также полное пошаговое описание протокола анализа данных секвенирования ампликонов вариабельных регионов V3-V4 гена 16S рРНК для оценки таксономического состава микробных сообществ с последующим статистическим анализом с помощью web-платформы MicrobiomeAnalyst.

ЛИТЕРАТУРА

1. *Aloisio I.* Evaluation of the effects of intrapartum antibiotic prophylaxis on newborn intestinal microbiota using a sequencing approach targeted to multi hypervariable 16S rDNA regions / I. Aloisio, A. Quagliariello, S. D. Fanti, D. Luiselli, C. D. Filippo, D. Albanese, L. T. Corvaglia, G. Faldella, D. D Gioia //Applied microbiology and biotechnology. – 2016. – Т. 100. – №. 12. – С. 5537–5546.

2. *Amann R. I.* Phylogenetic identification and in situ detection of individual microbial cells without cultivation / R. I. Amann, W. Ludwig, K. H. Schleifer // Microbiological reviews. – 1995. – Т. 59. – №. 1. – С. 143–169.

3. *Bolger A. M.* Trimmomatic: a flexible trimmer for Illumina sequence data / A.M. Bolger, M. Lohse, B. Usadel //Bioinformatics. – 2014. – Т. 30. – №. 15. – С. 2114–2120.

4. *Callahan B. J.* DADA2: High-resolution sample inference from Illumina amplicon data / B.J. Callahan, P.J. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, S.P. Holmes //Nature methods. – 2016. – Т. 13. – №. 7. – С. 581–583.

5. *Caporaso J. G.* QIIME allows analysis of high-throughput community sequencing data / J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight // Nature methods. – 2010. – Т. 7. – №. 5. – С. 335–336.

6. *DeSantis T. Z.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB / T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, G. L. Andersen // Appl. Environ. Microbiol – 2006. – Т. 72. – № 7. – С. 5069–5072.

7. *Fadeev E.* Comparison of two 16S rRNA primers (V3–V4 and V4–V5) for studies of arctic microbial communities / E. Fadeev, M. G. Cardozo-Mino, J. Z. Rapp, Ch. Bienhold, I. Salter, V. Salman-Carvalho, M. Molari, H. E. Tegetmeyer, P. L. Buttigieg, A. Boetius // Frontiers in microbiology. – 2021. – Т. 12. – С. 637526.
8. *Fichot E. B.* Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform / E.B. Fichot, R.S. Norman // Microbiome. – 2013. – Т. 1. – С. 1–5.
9. *Handelsman J.* Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products / J. Handelsman, M.R. Rondon, S. F. Brady, J. Clardy, R.M. Goodman // Chem. Biol. – 1998. – Т. 5. – № 10. – С. 245–249.
10. *Kumar P. S.* Target region selection is a critical determinant of community fingerprints generated by 16S pyrosequencing / P. S. Kumar, M. R. Brooker, S. E. Dowd, T. Camerlengo // PloS One. – 2011. – Т. 6. – №. 6. – С. E20956.
11. *Lu J.* Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2 / L. Lu, S.L. Salzberg //Microbiome. – 2020. – Т. 8. – №. 1. – С. 124.
12. *Martin M.* Cutadapt removes adapter sequences from high-throughput sequencing reads /M. Martin //EMBnet. journal. – 2011. – Т. 17. – №. 1. – С. 10–12.
13. *McDonald D.* Greengenes2 unifies microbial data in a single reference tree / D. McDonald, Y. Jiang, M. Balaban, K. Cantrell, Q. Zhu, A. Gonzalez, J.T. Morton, G. Nicolaou, D.H. Parks, S.M. Karst, M. Albertsen, P. Hugenholtz, T. DeSantis, S.J. Song, A. Bartko, A.S. Havulinna, P. Jousilahti, S. Cheng, M. Inouye, T. Niiranen, M. Jain, V. Salomaa, L. Lahti, S. Mirarab, R. Knight //Nature biotechnology. – 2024. – Т. 42. – №. 5. – С. 715–718.
14. *Paulino G. V. B.* Microbiota of healthy and bleached corals of the species Siderastrea stellata in response to river influx and seasonality in Brazilian northeast / G.V.B. Paulino, C.R. Félix, F.A. da Silva Oliveira, C. Gomez-Silvan, V.M.M. Melo, G.L. Andersen, M.F. Landell

//Environmental Science and Pollution Research. – 2023. – Т. 30. – №. 10. – С. 26496–26509.
15. *Quast C.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools / C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F. O. Glöckner // Nucleic Acids Res. – 2013. – Т. 41. – № D1. – С. D590–D596
16. *Schloss P. D.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities / P.D. Schloss, S.L. Westcott, T. Ryabin, J.R. Hall, M. Hartmann, E.B. Hollister, R.A. Lesniewski //Applied and environmental microbiology. – 2009. – Т. 75. – №. 23. – С. 7537–7541.
17. *Segata N.* Metagenomic biomarker discovery and explanation / N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W.S. Garrett, C. Huttenhower //Genome biology. – 2011. – Т. 12. – С. 1–18.
18. *Soriano-Lerma A.* Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples / A. Soriano-Lerma, V. Pérez-Carrasco, M. Sánchez-Marañón, M. Ortiz-González, V. Sánchez-Martín, J. Gijón, J. M. Navarro-Mari, J. A. García-Salcedo, M. Soriano // Scientific reports. – 2020. – Т. 10. – №. 1. – С. 1–13.
19. *Wang Q.* Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy / Q. Wang, G. M. Garrity, J. M. Tiedje, J. R Cole // Appl. Environ. Microbiol. – 2007. – Т. 73. – № 16. – С. 5261–5267.
20. *Wu Y. W.* MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm /Y.W. Wu, Y.H. Tang, S.G. Tringe, B.A. Simmons, S.W. Singer // Microbiome. – 2014. – Т. 2. – С. 1–18.
21. *Yarza P.* Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences /P. Yarza, P. Yilmaz, E. Pruesse, F.O. Glöckner, W. Ludwig, K.H. Schleifer, W.B. Whitman, J. Euzéby, R. Amann, R. Rosselló-Móra //Nature Reviews Microbiology. – 2014. – Т. 12. – №. 9. – С. 635–645.

Список электронных источников

1. EzBioCloud. – URL: https://www.ezbiocloud.net/ (дата обращения 01.08.2024).
2. FastQC. – URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc (дата обращения 01.08.2024).
3. FastX-Toolkit. – URL: http://hannonlab.cshl.edu/fastx_toolkit (дата обращения 01.08.2024).
4. MicrobiomeAnalyst. – URL: https://www.microbiomeanalyst.ca/MicrobiomeAnalyst/home.xhtml (дата обращения 02.08.2024).

М.И. МАРКЕЛОВА, Е.А. БУЛЫГИНА, М.Н. СИНЯГИНА, А.М. СЕНИНА, Т.В. ГРИГОРЬЕВА

АНАЛИЗ ДАННЫХ СЕКВЕНИРОВАНИЯ АМПЛИКОНОВ ГЕНА 16S РРНК С ПОМОЩЬЮ WEB-ПЛАТФОРМЫ MICROBIOMEANALYST

Учебное пособие