Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

РАЗРАБОТКА БИОИНФОРМАТИЧЕСКОГО ИНСТРУМЕНТАРИЯ ДЛЯ АНАЛИЗА МИКРОБИОМНЫХ ДАННЫХ ИЗ РАНДОМИЗИРОВАННЫХ КОНТРОЛИРУЕМЫХ ИССЛЕДОВАНИЙ

Автор: Гафаров Альберт Фаилевич _____

Направление подготовки: 01.03.02 Прикладная математика и информатика

Квалификация: Бакалавр

Руководитель: Ульянцев В.И., доцент ФИТиП, к.т.н. _____

К защите допустить

Руководитель ОП Парфенов В.Г., проф., д.т.н. _____

«____» _____ 20___ г.

Санкт-Петербург, 2019 г.

Студент Гафаров А.Ф. Группа М3438 Факультет ИТиП

Направленность (профиль), специализация
Математические модели и алгоритмы в разработке программного обеспечения

Консультанты:

а) Одинцова В.Е., без степени, без звания _____

ВКР принята «_____» _____ 20___ г.

Оригинальность ВКР _____%

ВКР выполнена с оценкой _____

Дата защиты «_____» _____ 20___ г.

Секретарь ГЭК Павлова О.Н. _____

Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

УТВЕРЖДАЮ
Руководитель ОП
проф., д.т.н. Парфенов В.Г. _____
«____» _____ 20___ г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Студент Гафаров А.Ф. **Группа** М3438 **Факультет** ИТиП

Руководитель Ульянцев В.И., доцент ФИТиП, к.т.н., главный научный сотрудник Университета ИТМО

1 Наименование темы: Разработка биоинформатического инструментария для анализа микробиомных данных из рандомизированных контролируемых исследований

Направление подготовки (специальность): 01.03.02 Прикладная математика и информатика

Направленность (профиль): Математические модели и алгоритмы в разработке программного обеспечения

Квалификация: Бакалавр

2 Срок сдачи студентом законченной работы: «31» мая 2019 г.

3 Техническое задание и исходные данные к работе

Требуется разработать алгоритм автоматической обработки для аналитической платформы обработки метагеномных данных «Кномикс-Биота», позволяющий анализировать результаты плацебо-контролируемых интервенционных исследований.

4 Содержание выпускной работы (перечень подлежащих разработке вопросов)

Необходимо выбрать статистические методы для оценки влияния интервенции на микробиоту в ходе плацебо-контролируемого исследования. Выбранные методы должны выявлять различия в изменениях таксономического, функционального состава и альфа-разнообразия между интервенционной группой и группой плацебо. Алгоритм должен быть интегрирован в платформу «Кномикс-Биота». Необходимо разработать визуализацию (в том числе интерактивную) полученных результатов.

5 Перечень графического материала (с указанием обязательного материала)

Графические материалы и чертежи работой не предусмотрены

6 Исходные материалы и пособия

Отсутствуют.

7 Дата выдачи задания «01» сентября 2018 г.

Руководитель ВКР _____

Задание принял к исполнению _____

«01» сентября 2018 г.

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Студент: Гафаров Альберт Фаилевич

Наименование темы ВКР: Разработка биоинформатического инструментария для анализа микробиомных данных из рандомизированных контролируемых исследований

Наименование организации, в которой выполнена ВКР: Университет ИТМО

ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

1 Цель исследования: Разработка компонента в рамках платформы «Кномикс-Биота» для анализа микробиомных данных из рандомизированных контролируемых исследований

2 Задачи, решаемые в ВКР:

а) Постановка задачи, обзор предметной области.

б) Разработка компонента, анализирующего микробиомные данные из рандомизированного контролируемого исследования.

в) Тестирование разработанного компонента на реальных метагеномных образцах.

3 Число источников, использованных при составлении обзора: 12

4 Полное число источников, использованных в работе: 20

5 В том числе источников по годам:

Отечественных			**Иностранных**		
Последние 5 лет	От 5 до 10 лет	Более 10 лет	Последние 5 лет	От 5 до 10 лет	Более 10 лет
1	0	0	10	3	6

6 Использование информационных ресурсов Internet: да, число ресурсов: 12

7 Использование современных пакетов компьютерных программ и технологий:

Пакеты компьютерных программ и технологий	**Раздел работы**
Интерактивная оболочка `IPython`	2.2.4
Веб-оболочка `Jupyter Notebook`	2.2.4
Модули `skbio`, `statsmodels`, `pymer4`	3.1.2
JavaScript-фреймворк `AngularJS`	3.1.3

8 Краткая характеристика полученных результатов

Разработан компонент, необходимый для анализа микробиомных данных из рандомизированных контролируемых исследований в рамках платформы «Кномикс-Биота». Компонент протестирован на реальных данных, проведено сравнение результата работы с уже имеющимися результатами.

9 Гранты, полученные при выполнении работы

При выполнении работы грантов получено не было.

10 Наличие публикаций и выступлений на конференциях по теме работы

По теме этой работы я ничего не публиковал.

Студент Гафаров А.Ф. _____

Руководитель Ульянцев В.И. _____

«____» _____ 20___ г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Микробиота кишечника человека участвует во многих процессах, важных для здоровья человека. На ее состояние влияют различные факторы, такие как диета, прием лекарственных средств, различные заболевания. Для выявления таких ассоциаций проводятся клинические исследования. Один из распространенных видов таких исследований – контролируемое интервенционное исследование. В ходе него участники делятся на две группы, одна из которых получает лекарство или следует определённой диете, а другая нет. Затем сравниваются образцы микробиоты, взятые в начале и в конце исследования в обеих группах. Такой дизайн эксперимента является стандартом формата клинических исследований, например, в фармацевтической индустрии и пищевой промышленности с целью выяснения эффективности, безопасности различных лекарственных средств, пищевых добавок, функционального питания. Обработка результатов таких исследований предполагает использование специфических методов статистического анализа и визуализации.

Целью данной работы является разработка компонента для обработки микробиомных данных, полученных в ходе контролируемых интервенционных исследований. Он разработан на основе алгоритмов платформы «Кномикс-Биота». Эта платформа представляет собой сервис для анализа данных о микробных сообществах с учетом особенностей дизайна исследования. Разработанный компонент предлагается встроить в платформу как отдельный, новый тип анализа. Для тестирования предлагается апробация разрабатываемого компонента на уже изученных микробиомных данных для сравнения результатов анализа.

ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Микробиота кишечника человека и ее исследование

1.1.1. Обзор микробиомных сообществ

Сообщества микроорганизмов широко представлены практически в любой среде обитания, включая экстремальные, такие как гейзеры или щелочные озера. В каждой такой среде нельзя выделить один наиболее приспособленный вид, только некоторое сообщество, из чего можно предположить зависимость между ними. Микробиота состоит из самых разных микроорганизмов: бактерий, вирусов, архей, также от среды сильно зависит степень концентрации: до 10^9 бактериальных клеток в грамме почвы и до 10^6 в миллилитре речной воды.

Состав микробиоты может влиять на саму среду: некоторые группы бактерий, содержащиеся в почве, имеют свойство поглощать азот из воздуха, тем самым увеличивая ее плодородность.

1.1.2. Микробиота кишечника человека

Различные сообщества микроорганизмов обитают во многих местах человеческого организма, находясь с ним в симбиотических отношениях. Кишечник человека является самой густонаселенной бактериями средой среди всех частей тела. Ранее считалось, что присутствие бактерий в кишечнике является патогенным, но в настоящее время обнаружено, что микробиота кишечника выполняет обширный набор функций: выполняет эндокринную функцию, является частью пищеварительной системы, синтезируя витамины и другие необходимые для человека вещества, поддерживает иммунитет.

Микробиота кишечника человека состоит не только из бактерий, в нее также входят простейшие, грибки и вирусы [3]. Последние исследования выявили, что ее состав не одинаков для каждого человека, и что можно выделить так называемые энтеротипы [8], типовые наборы бактерий из которых она состоит. В ее состав по примерным оценкам входит от 300 до 1000 видов, основные компоненты представлены на рисунке 1.

Дальнейшие микробиомные исследования показали, что ее состав стабилен во времени, и были выделены основные факторы, влияющие на состав, такие как прием антибиотиков, изменения рациона и образа жизни [5].

1.1.3. Модуляция состава микробиоты

Кроме поиска маркеров заболеваний интересен вопрос модуляции состава микробиоты: приведение условно ненормального состава к нормальному

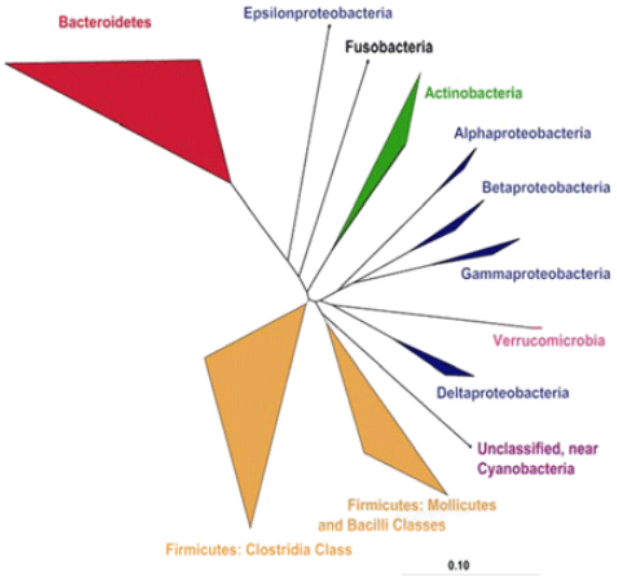

Рисунок 1 – Филогенетическое дерево разнообразия микробиоты кишечника. Адаптировано из [3].

виду, восстановления ее биоразнообразия, представленности основных, чаще всего встречаемых у здорового населения бактерий.

Одним из таких исследований является изучение восстановления микрофлоры кишечника после приема антибиотиков, в котором были сравнены прием пробиотиков – пищевых добавок, содержащих в себе живые микрокультуры, распространенными являются молочнокислые бактерии $lactobacilli$ и $bifidobacteria$, трансплатацией фекальной микробиоты и спонтанной восстановлением нормального состава микробиоты без какого-либо вмешательства. Сравнение производилось с группой не принимавших антибиотиков и не соблюдавших какую-либо специализированную диету. Результаты показали, что прием пробиотиков замедляет восстановление биоразнообразия до нормального уровня даже по сравнению со спонтанным восстановлением [18]. На графике 2 можно увидеть, что в течение эксперимента, который длился около месяца, биоразнообразие увеличилось до того же уровня только при трансплантации, а группа, принимавшие пробиотики, показала худший результат.

Кроме восстановления состава микробиоты до нормального вида, также исследуется модуляция ее состава для придания организму некоторых свойств, таких как сопротивляемость отравлению тяжелыми токсичными металлами. В качестве примера можно привести эксперимент, проводившийся в городе Мванза, Танзания, в рационе населения которого присутствует рыба из озера Виктория, подвергшееся антропогенному загрязнению. Были рассмот-

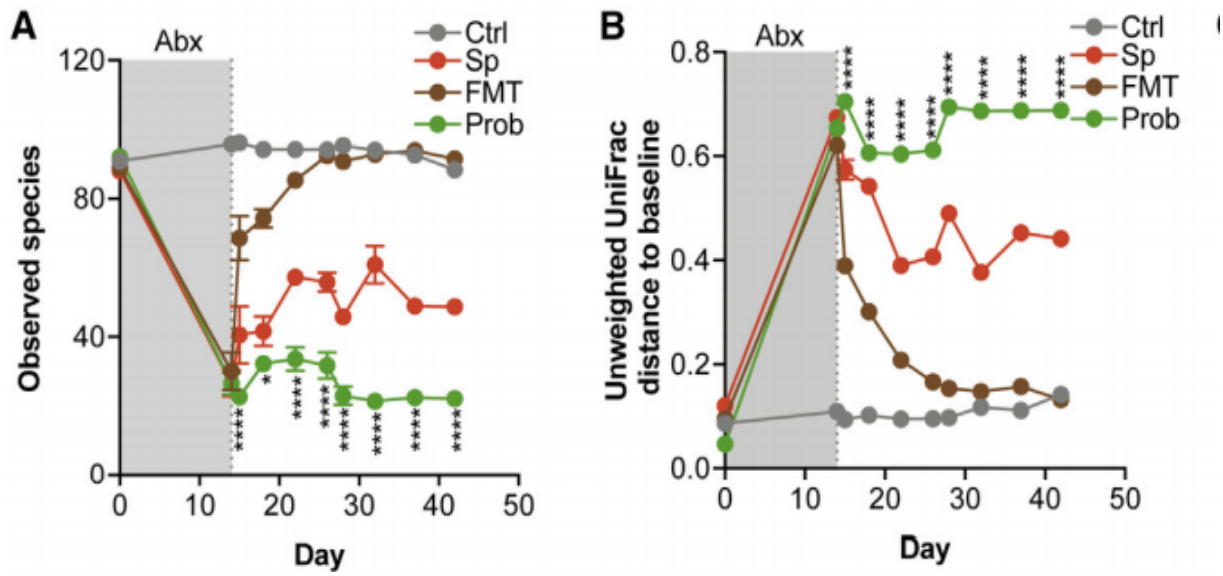

Рисунок 2 – График, отображающий среднее разнообразие кишечной микробиоты в течение исследования. Серым цветом выделен временной промежуток приема антибиотиков. Точки серого цвета – среднее образцов из группы, не принимавших антибиотик, красного – среднее образцов из группы, принимавших плацебо, коричневого – среднее образцов, для которых была проведена фекальная трансплантация, зеленого цвета – среднее для группы, принимавших пробиотик. На графике слева указано среднее количество обнаруженных видов, справа – среднее расстояние UniFrac между образцами и эталонным образцом. Адаптировано из [18]

рены две группы людей, подверженных более остальных воздействию токсичных металлов – беременных женщин и детей. В эксперименте исследовалось влияние *Lactobacillus rhamnosus*, используемого в качестве пробиотика. Результатом исследования показали, что содержание ртути и мышьяка в образцах крови у беременных женщин, принимавших плацебо выше, чем у принимавших пробиотик [20].

1.2. Анализ микробиомных данных

1.2.1. Традиционные методы исследования

При изучении микроорганизмов первым способом исследования являлось получение изолированного образца, штамма. Следующий шаг – культивирование и дальнейшее исследование, изучение свойств. При анализе целого бактериального сообщества такой метод не подходит по нескольким причинам. Большинство бактерий тяжело культивируются в искусственной среде, тем самым порождая необходимость в культуромике для кишечной микробиоты: подбора оптимальных условий культивирования для разных таксонов [4].

На рисунке 3 представлено количество культивируемых и обнаруженных видов в микробиоте человека. Также при исследовании целого сообщества нужно учитывать взаимосвязь с другими членами сообщества. Поэтому традиционные методы малоприменимы к исследованию микробиоты.

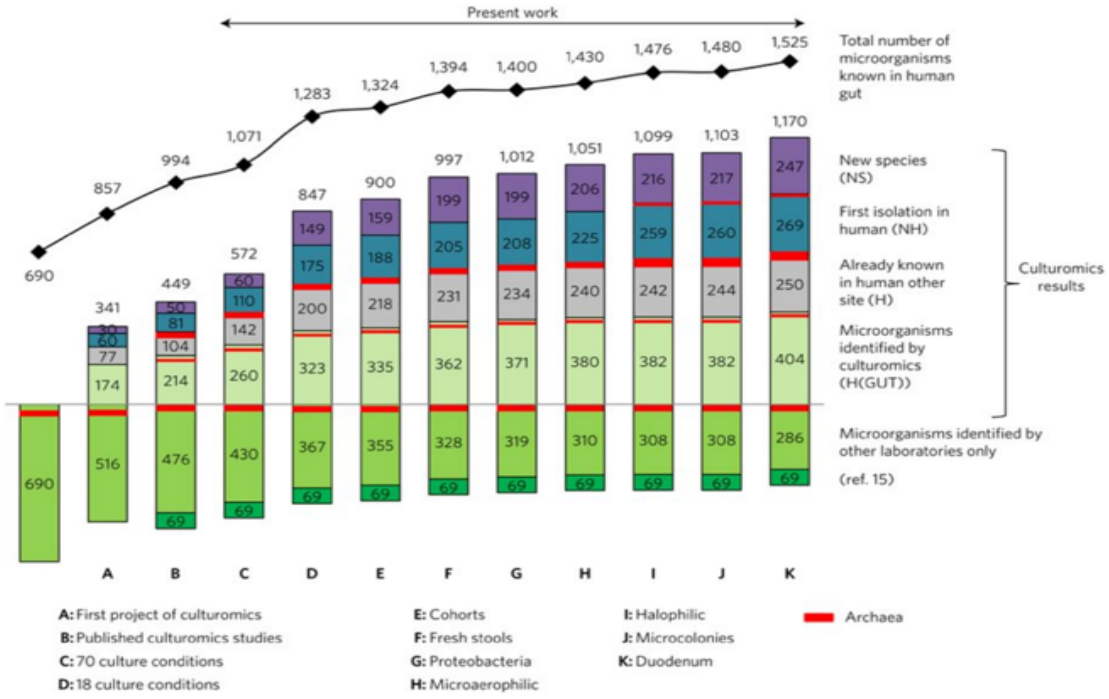

Рисунок 3 – Количество культивированных и обнаруженных видов в микробиоте человека. Адаптировано из [4]

1.2.2. Высокопроизводительное секвенирование

Современным способом выявления состава микробиоты является исследование его генома. ДНК-секвенирование – процесс получения нуклеотидной последовательности ДНК, в результате имеется набор прочтений, каждое из которых является последовательностью нуклеотидов. Размер последовательности зависит от используемого метода. Теперь данные готовы для дальнейшей обработки – биоинформатического анализа, нужного для восстановления частей генома из этих коротких прочтений, которые позволят определить наличие в составе образца какого-либо организма.

В результате появления новых методов ДНК-секвенировния, таких как исследование генома стало более доступным. Результатом явилось появлении нового направления биологии – метагеномики, в которой изучается весь генетический материал, получаемый из образца [14]. Такой анализ позволяет учитывать не только культивируемые микроорганизмы, но и не культивируемые,

получая таким образом информацию о полном многообразии организмов в образце, их функциях и экологических взаимосвязях между ними.

Секвенирование можно поделить на два типа: секвенирование полного генома или некоторых маркерных подпоследовательностей генома. Распространенным анализом метагенома является секвенирование с использованием подпоследовательностей 16S рРНК, которое входит в состав рибосомы. Использование данного гена обусловлено тем, что края этого гена похожи для большого набора организмов, но внутренние области вариативны и используются для филогенетической классификации бактерий, и тем, что данные подпоследовательности умещаются в длину одного прочтения и не требуют процедуры удлинения. Таким образом, из плюсов данного метода можно отметить его дешевизну и отсутствие необходимости в удлинении прочтений, из минусов – существование таких бактерий, что даже при совпадении подпоследовательностей 16S рРНК которых имеется заметная разница в других функциональных группах генов.

1.2.3. Цели анализа данных

Современные методы секвенировния микробиомных данных, позволяющие получать миллиарды ридов за несколько дней, поставили задачу эффективного анализа большого объема метагеномных данных, включая низкоуровневую: предобработка данных, исключающая зашумленные данные, склеивание ридов в один более длинный фрагмент, так и высокоуровневую: статистический анализ и интерактивная визуализация результатов. Вкратце, задача сводится к выполнению следующих целей:

— производительность: анализ большого объема данных должен выполняться за разумное время
— горизонтальное масштабирование: возможность увеличения вычислительной мощности путем увеличения числа вычислительных компонентов
— наглядность результатов анализа: исследователь должен иметь достаточный набор инструментов для формирования и подтверждения научной гипотезы

1.2.4. Формат данных результата секвенирования метагенома и его предобработка

Заключающий этап работы секвенатора – преобразование внутреннего представления данных в итоговый набор прочтений(ридов), каждый из которых является последовательностью нуклеотидов(A, C, G и T) или специального символа, представляющий полную неопределенность. Также для каждой позиции последовательности секвенатор предоставляет меру его качества, определяющаяся формулой:

$$Q_i = -10log_{10}P_i$$

где P_i – вероятность ошибки определения типа нуклеотида в i-той позиции. Соответствующие числа хранятся в отдельном файле для каждого образца – формат FASTA, или в одном файле вместе с ридами – формат FASTQ.

Некоторые секвенаторы позволяют обработку сразу нескольких образцов, помечая риды служебными символами, баркодами, для разделения данных по образцам. Длина прочтения не всегда жестко задана – в большинстве случаев она имеет только верхнюю оценку, поэтому имеет смысл проводить фильтрацию по минимальной длине рида. Также можно учитывать степень качества прочтения для каждого нуклеотида из рида: отбрасывать рид, если среднее качество прочтения меньше некоторого порога или удалять с концов рида прочтения с малым качеством для повышения среднего качества для всего рида.

Из нетривиальных примеров предварительной обработки ридов можно привести анализ k-меров, точечно исправляющий ошибки в прочтениях путем анализа подпоследовательностей нуклеотидов длины k, k-меров, по всем ридам.

При секвенировании генома происходит химическая обработка ДНК образца, поэтому есть возможность прочтения искусственных последовательностей. Такие риды могут влиять на результаты дальнейшего анализа, поэтому рекомендуется отбрасывать риды, содержащие шаблонные последовательности. Тем же способом происходит обработка посторонних прочтений, получившихся в следствие биологического загрязнения. В случае анализа микробиоты, средой обитания которых является некоторый организм, посторонними прочтениями являются прочтения ДНК организма хозяина.

1.2.5. Результаты секвенирования последовательности гена 16S рРНК

Следующим шагом после предобработки является количественный анализ бактериального состава микробиоты путем выявления к какому виду относится каждое прочтение 16S рРНК образца. Самые распространенные способы таксономической классификации опираются на такой термин как OTE, который расшифровывается как операционная таксономическая единица. Каждое прочтение отображается в некую OTE, которая дальше используется для определения бактериального вида. Вкратце, этот процесс можно описать следующим образом: все прочтения гена 16S рРНК группируются по некоторой мере сходства, распространенным признаком, например, является сходство в 97%, и объединяются в непересекающиеся кластеры, каждый из которых и является OTE. Для дальнейшего сравнения в каждом кластере выбирается его представитель.

После получения списка OTE и получения представителей для каждого класса наступает фаза таксономического анализа для каждого прочтения. Существует два основных подхода: поиск на основании шаблона и поиск de novo. В первом случае, описанном на рисунке 4, происходит сопоставление каждого рида с имеющейся базой OTE, состоящей из шаблонных последовательностей, для каждой из которых известна ее таксономия. При отсутствии совпадения рид отбрасывается. В результате для каждой OTE из базы имеется представленность ее в образце. Из плюсов данного метода можно отметить легкость параллелизации вычислений, легкость вычислений, из минусов – исключение не представленных в базе мало изученных видов.

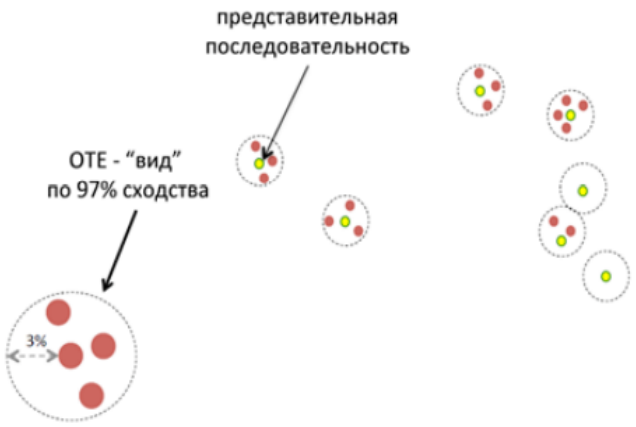

Рисунок 4 – Анализ OTE с помощью референсной базы. Адаптировано из [1].

Второй подход, описанный на рисунке 5, заключается в кластеризации всех ридов по некоторому порогу сходства. Каждый из полученных кластеров формирует OTE, а количество ридов в нем определяет его представленность в образце. Таксономическая классификация происходит с помощью специализированных алгоритмов, одним из распространенных является RDP Classifier (Ribosomal Database Project Classifier) [16]. Преимуществом данного подхода – отсутствие необходимости шаблонной базы, недостаток – сложность параллелизации и большая вычислительная сложность.

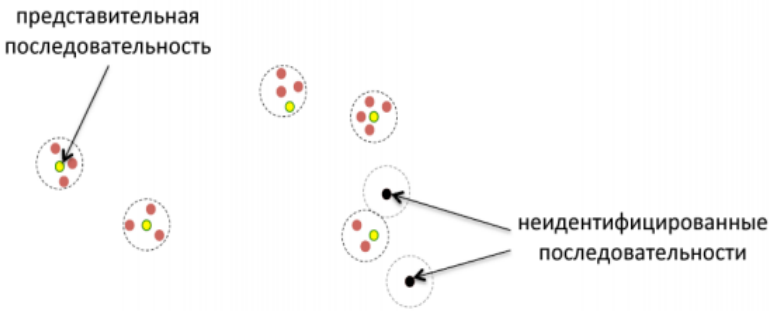

Рисунок 5 – Анализ OTE de novo. Адаптировано из [1].

Чаще используется гибридная схема, описанная на рисунке 6, заключающаяся в использовании сперва первого способа, идентификации прочтений с помощью шаблонной базы, а затем второго способа для неидентифицированных на предыдущем шаге ридов. Гибридная схема позволяет избавиться от минусов обоих методов: идентификации прочтений, не содержащихся в базе, и большой вычислительной сложности.

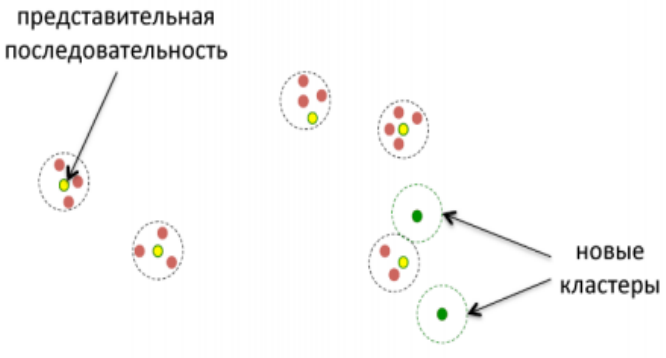

Рисунок 6 – Анализ OTE с помощью гибридной схемы. Адаптировано из [1].

При выборе гибридной схемы требуется база, содержащая шаблонные прочтения 16S рРНК с известной таксономией. Приведем самые распространенные из них: база Greengenes – база полных прочтений последовательностей гена 16S рРНК, SILVA, RDP.

Помимо представленностей видов бактерий в образце используются оценки его биоразнообразия, отличия биоразнообразия между двумя образцами и общее биоразнообразие среди всех образцов. Данные оценки удобно представлять в численном представлении для удобного сравнения, в биоинформатике эти величины обозначаются как альфа-разнообразие, бета-разнообразие и гамма-разнообразие, соответственно.

При анализе последовательностей 16S рРНК видов число обнаруженных видов зависит от количества прочтений для образца. Иными словами, чем больше прочтений на образец, тем больше вероятность обнаружить большее количество видов. Также увеличение ридов имеет нелинейный характер: происходит насыщение, и для образцов малонаслеенных сред оно происходит быстрее. Можно аппроксимировать данную зависимость, если посчитать биоразнообразие на подвыборках прочтений разного размера [7]. Результаты данного эксперимента представлены на графике 7. Данный метод полезен при сравнении двух образцов с разным количеством прочтений, так как позволяет преобразовывать количество обнаруженных ОТЕ для заданного количества прочтений. Наиболее используемые методы оценки биоразнообразия: филогенетическое разнообразие, использующее долю филогенетического дерева, найденное в образце и Chao1.

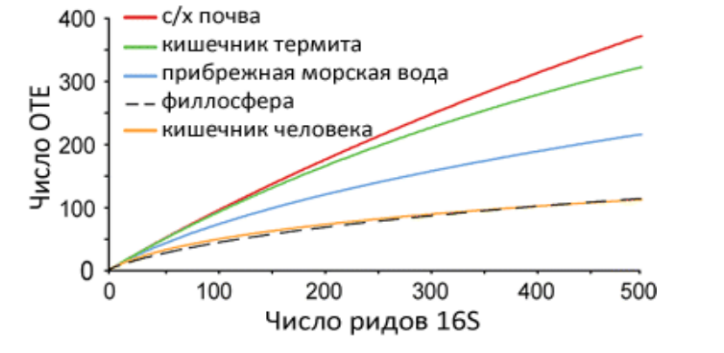

Рисунок 7 – Демострация насыщения количества обнаруженных видов с увеличением количества прочтений. Адаптировано из [7].

Бета-разнообразие является первичным признаком различия микробиомного состава между двумя сообществами. Существует большое ко-

личество методов, одним из распространенных методов является метод UniFrac [11], который схож с методом филогенетического разнообразия для подсчета альфа-разнообразия: для двух сообществ строятся их филогенетические деревья, и результатов выступает отношение длины веток одного сообщества к суммарной длине всех веток объединенного дерева.

1.3. Форматы клинических исследований

1.3.1. Обзор форматов исследований

При разработке новых лекарственных средств и исследовании уже существующих невозможно обойтись без клинических исследований. Перед проведением любого исследования следует определиться с его форматом. Клинические исследования можно классифицировать по степени вмешательства в жизнь пациента:

- Наблюдательное – исследование, при котором не происходит какого-либо вмешательства.
- Неинтервенционное – исследование, при котором выбор лекарственного средства не зависит от исследователя.
- Интервенционное – исследование, при котором исследователь проводит анализ некоторого нового лекарственного средства или других методов использования уже известных(другая дозировка или новая категория пациентов).

Также исследования классифицируют по цели исследования: профилактические, целью которых является предупреждение заболеваний у здоровых людей путем анализа, диагностическое, для выявления новых маркеров заболеваний и т.д.

Больше всего интересны так называемый дизайн клинического исследования, который представляет собой общий план или алгоритм исследования. Выделено два типа дизайнов исследования:

- Когортные исследования – дизайн исследования при котором некоторая группа людей(когорта) наблюдается в течение всего эксперимента. Пациенты разделяются на группы в зависимости от вида принимаемых лекарственных средств или других факторов. Далее происходит сравнение между группами на предмет эффективности лечения.

— Исследования «случай-контроль» – дизайн исследования при котором некоторая группа людей(случай), принимающая некое лекарственное средство, сравнивается с группой(контроль), принимающая плацебо.

При выборе группы пациентов, которым назначается плацебо, исследование называется простым слепым, если исследователь знает, каким пациентам назначено лечение, а каким плацебо, в обратном случае исследование называется двойным слепым. Последний метод позволяет избавиться от ряда некоторых субъективных факторов: назначения лекарства пациентам с более тяжелой стадией заболеваний и от приверженности трактования эффективности приема лекарства.

1.3.2. Рандомизированное контролируемое испытание

Рандомизированное контролируемое испытание на текущий момент является эталонным методом при выборе дизайна клинического исследования. При выборе такого дизайна, пациенты делятся на две группы: пациенты, принимающие плацебо, или группа «контроль», и пациенты, принимающие некий экспериментальный препарат, или группа «случай». Забор образцов производится как и до, так и после исследования, что позволяет проводить сравнение эффективности препарата на группе «случай», избегая при этом ложноположительных результатов с помощью сравнения с контрольной группой.

1.4. Интерактивные онлайн-платформы для анализа метагеномов

1.4.1. Краткое описание

Бурный рост количества метагеномных исследований и уменьшение стоимости секвенирования метагенома привело к использованию метагеномного анализа не только в академических кругах, но и в пищевой и фармакологической промышленностях. Не в каждой группе исследователей есть легкий доступ к интсрументарию для биоинформатического анализа, поэтому хотелось бы иметь такой сервис, который бы позволял проводить биоинформатический анализ. Таким образом, такой сервис должен быть удовлетворять неким требованиям:

— Вычислениядолжны производится локально – анализ прочтений может занимать много времени и быть требовательным к вычислительной мощности.

— Должны поддерживаться все дизайны исследований

— Пользователю должны быть предоставлены результаты как минимум основных статистических методов, интерактивная визуализация полученных результатов.

1.4.2. Платформа Кномикс-Биота

Одним из сервисов, предоставляющих такие возможности, является платформа «Кномикс-Биота». Функционал сервиса позволяет получать из «сырых» данных готовые аналитические отчеты, также сервис обладает интерактивной визуализацией, еще более упрощающий интерпретацию результатов клинических исследований. Данные метагеномного исследования могут сопровождаться метаданными, содержащие такую информацию о пациентах как пол, возраст, вес или вредные привычки, для дальнейшего исследования.

Все вычисления производятся удаленно, данные после загрузки хранятся в облаке, как и все результаты анализа. Платформа поддерживает большую часть дизайнов клинических исследований. Также платформа позволяет анализ результатов секвенирования гена 16S рРНК, так и прочтения полногеномного секвенирования.

Платформа представляет собой набор компонентов:

— Хранилище Amazon S3 для хранения большого количества метагеномных данных большого объема.

— Серверная часть, производящая анализ данных, который поделен на две части: базовый анализ(предобработка прочтений, получение таксономического и функционального состава микробиоты) и вторичный анализ(последующий анализ данных с помощью таких методов, как, например, статистический анализ).

— Клиентская часть для визуализации полученных результатов в браузере.

Платформа «Кномикс-Биота» не единственная платформа, позволяющая проводить метагеномные исследования. На рисунке 8 представлен сравнительный анализ с другими онлайн-платформами для анализа микробиоты. Из сравнительных преимуществ данной платформы можно отметить возможность ознакомления с результатами уже проведенных исследований, наличие интерактивной визуализации и возможность получения бесплатного доступа исследователям к инструментам платформы [10].

Pipeline name	"Raw" data analysis		Statistical analysis			External datasets availability	Data sharing
	16S rRNA sequencing	WGS	Basic statistics	Group comparison	Interactive Visualizations		
Knomics-Biota	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nephele	Yes	Yes	Yes	Yes	No	Yes (data from HMP [16] only)	Yes
MG-RAST	Yes	Yes	Yes	No	No	Yes	Yes
One Codex	No	Yes	Yes	Yes	No	No	Yes
GUSTA ME	No	No	Yes	Yes	No	No	No
CosmosID	No	Yes	Yes	Yes	No	No	Yes
QIAGEN Microbial Genomics Pro Suite	No	Yes	Yes	No	No	No	NA
Calypso	Yes	No	Yes	Yes	No	No	No

Рисунок 8 – Сравнение функционала платформы «Кномикс-Биота» и других платформ. Адаптировано из [10].

Выводы по главе 1

Увелчивающаяся с каждым днем доступность микробиомных исследований с помощью секвенирования метагенома вкупе со сравнительно недавним обнаружением важности такой части организма как микробиота кишечника человека привела к бурному росту ее изучения. Приведены примеры таких исследований: исследования пробиотиков, сравнение пробиотиков, пребиотиков и аутологичной фекальной трансплантацией. Также приведены исследования, показывающие возможность модуляции ее состава.

Подробны описаны типы и форматы клинических исследований, вдобавок приведены возможные дизайны исследований. Показано, что двойное слепое рандомизированное исследование является текущим эталоном для выбора дизайна эксперимента.

Описаны преимущества онлайн сервиса, позволяющего проводить чтение и предобратку результатов секвенирования метагенома, высокоуровненый статистический анализ и предоставлять интерактивную визуалиазцию для более интуитивной интерпретации результатов для исследователя. Описаны возможности платформы «Кномикс-Биота», ее архитектура. Также приведено ее сравнение с других платформами, показана ее преимущества и конкурентоспособность.

ГЛАВА 2. ВЫБОР МЕТОДОВ ДЛЯ АНАЛИЗА МИКРОБИОТНЫХ ДАННЫХ ИНТЕРВЕНЦИОННОГО КОНТРОЛИРУЕМОГО ИССЛЕДОВАНИЯ

2.1. Постановка задачи

2.1.1. Цели

Платформа «Кномикс-Биота» поддерживает большую часть дизайнов клинических испытаний. Текущая ее реализация уже используется в СибГМУ и МГУ для анализа результатов последних клинических исследований [9] и написания научных статей. Платформа поддерживает несколько вариантов дизайна эксперимента, однако не имеет отдельного инструмента для анализа основного дизайна клинических исследований – интервенционных плацебо-контролируемых испытаний.

Цель данной работы – разработка компонента для анализа данных клинических исследований, проведенных в формате интервенционных плацебо-контролируемых испытаний, внедрение его в платформу «Кномикс-Биота», апробация разработанного компонента на уже собранных и новых метагеномных данных.

2.1.2. Задачи

К разработке внедряемого компонента ставятся следующие требования:

— Должен проводиться анализ между двумя группами (интервенционной и контрольной), в которых образец микробиоты был взят у каждого участника два раза: в начале и в конце исследования.
— Должен быть проведен покомпонентный анализ таксономического и функционального состава, а также анализ альфа-разнообразия.
— Должна быть представлена удобная для интерпретации интерактивная визуализация таксономического состава, а также визуализация результатов.
— Проведение анализа должно быть горизонтально масштабируемым, все вычисления должны производиться в облаке.

2.1.3. Актуальность

Поддержка данного дизайна эксперимента является актуальной проблемой, на данный момент планируется проведение анализа для двух результатов клинических исследований приема пребиотиков с помощью разрабатываемого компонента и публикация получившихся результатов.

2.2. Статистический анализ

Результатом таксономического анализа является матрица представленностей, где элемент матрицы $a_{i,j}$ обозначает представленность, где i – номер образца, j – номер таксона. На основании данной таблицы производится анализ между группами, выявляются зависимости состава от различных факторов.

Одной из отличительных особенностей таких данных является несоответствие нормальному распределению: представленность большой часть бактерий близка к нулю, а бактерии, имеющие ненулевую представленность, присутствуют только в некоторых образцах [17]. Таким образом, использование методов, предполагающих нормальное распределение случайной величины, не имеет смысла. Простейшим вариантом обойти данное ограничение является использование непараметрических методов, которые не предполагают конкретный вид распределения случайной величины. Другим подходом является приведение данных к нормальному виду, так как использование непараметрических методов приводит к понижению чувствительности анализа.

2.2.1. MaAsLin

В текущих поддерживаемых форматах исследований платформы «Кномикс-Биота», таких как парное исследование или исследование «случай-контроль», многофакторный анализ, используется метод MaAsLin (от англ. Multivariate Association with Linear Models) [12], который состоит из нескольких этапов. Для каждого признака (таксона, метаболического пути, гена или реакции):

а) Фильтрация низкопредставленных признаков.

б) Отсеивание неинформативных факторов.

в) Преобразование $arcins(sqrt(x))$ представленностей признака для выравнивания дисперсии и приведения данных к распределению, более близкому к нормальному.

г) Построение множественной линейной регрессии, либо смешанной линейной модели, в которой зависимая переменная – преобразованная представленность, а независимые - исследуемые факторы. Проверка статистической значимости коэффициентов.

На рисунке 9 представлена схема работы данного алгоритма. Основываясь на данном методе, предлагается исследовать последний этап – выбор ли-

нейной модели, для получения метода для анализа данных интервенционного плацебо-контролируемого исследования.

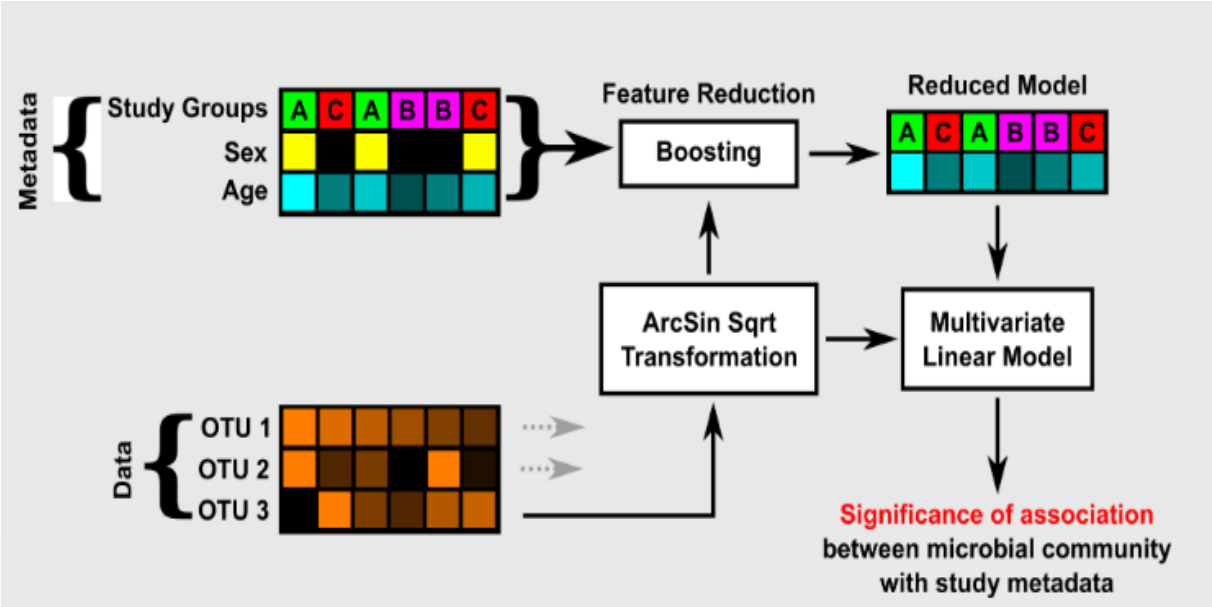

Рисунок 9 – Схема работы метода MaAsLin [12]

2.2.2. Ковариационный анализ

Ковариационный анализ или ANCOVA (от англ. Analysis of covariance) является методом математической статистики, в котором анализируется линейная модель зависимости среднего значения некоторой случайной величины от некоторых качественных факторов и количественных факторов. Линейную модель можно представить как $y \sim \mathcal{N}(X\beta + o, \sigma^2)$, где y – вектор наблюдаемых представленностей переменная, X – матрица факторов, β – вектор коэффициентов модели, o – вектор смещений, σ^2 – дисперсия.

При использовании данного метода предлагается группировать данные по параметру идентификатора пациента, получая для каждого пациента значения до и после эксперимента и фактор принадлежности к группе «случай» или «контроль». Таким образом, итоговая формула линейной модели в стиле языка R будет выглядеть так:

$$y_{after} \sim y_{before} + group,$$

где $group$ – фактор принадлежности к группе, а y_{before} и y_{after} представленность бактерии до и после эксперимента соответственно.

2.2.3. Модель со смешанными эффектами

Модель со смешанными эффектами – это статистическая модель, учитывая как и фиксированные, так и случайные переменные. Такие модели получили широкое распространение в физике и биоинформатике, одним из их достоинств является способность анализировать разреженные данные. Опишем данную линейную модель:

$$(y|B = b) \sim \mathcal{N}(X\beta + Zb + o, \sigma^2),$$

где Z – матрица случайных эффектов, B, при которых случайные факторы имеют значение b, X – матрица фиксированных факторов, β – вектор коэффициентов для фиксированных факторов.

Итоговая формула линейной модели в стиле языка R будет выглядеть так:

$$y \sim group + time + (1 \mid subject),$$

где категориальные переменные $group$ и $time$ обозначают принадлежность к группе «случай» или «контроль», принадлежность к группе до или после эксперимента, соответственно, а $subject$ – идентификатор пациента. Запись $(1 \mid x)$ при описании формулы обозначает, что для всех объектов у которых зафиксирована некоторая переменная x, свободный член будет одинаков, причем между объектами с разной величиной x свободный член будет нормально распределен с некоторым смещением. В нашем случае, данная формула обозначает предположение, что для каждой бактерии существует некоторый разброс представленности для каждого пациента, который нормально распределен.

2.2.4. Выбор статистического метода

При выборе используемого статистического метода для анализа количественного и функционального состава микробиоты было сравнены два подхода: модель со смешанными эффектами и ANCOVA. При использовании линейной модели ANCOVA, предполагается линейная зависимость между измерениями до и после эксперимента, поэтому перед сравнением модели ANCOVA и модели со смешанными факторами на данных, предлагается проверить наличие линейной зависимости. Так как данные приведены к нормальному виду, линейность проверялась с помощью коэффициента корреляции Пирсона. Для каждой бактерии проверялась гипотеза о линейности изменения представ-

ленности относительно изначального значения. На таблице 1 представлены уровни значимости отсутствия корреляции для некоторого подмножества бактерий.

Таблица 1 – Коэффициент корреляции Пирсона для каждого таксона и соответствующее p-значение.

Таксон	Коэффициент корреляции	p-значение
o_Bifidobacteriales;f_Bifidobacteriaceae;g_Bifidobacterium	-0.84	6.93e-33
o_Lactobacillales;f_Streptococcaceae;g_Lactococcus	-0.76	8.18e-23
o_Clostridiales;f_Clostridiaceae;g_Clostridium	-0.76	5.81e-24
o_Clostridiales;f_Clostridiaceae;g_SMB53	-0.75	9.96e-23
o_Clostridiales;f_Lachnospiraceae;g_Coprococcus	-0.73	1.14e-21
o_Enterobacteriales;f_Enterobacteriaceae;g_	-0.73	2.70e-19
o_Clostridiales;f_Clostridiaceae;g_	-0.72	1.41e-20
o_Rhizobiales;f_Brucellaceae;g_Ochrobactrum	-0.72	5.72e-21
o_Burkholderiales;f_Alcaligenaceae;g_Achromobacter	-0.71	3.42e-20
o_Coriobacteriales;f_Coriobacteriaceae;g_Collinsella	-0.70	2.64e-19
o_Clostridiales;f_Lachnospiraceae;g_Blautia	-0.70	2.41e-19
o_Bacteroidales;f_S24-7;g_	-0.70	1.30e-18
o_Clostridiales;f_Lachnospiraceae;g_Lachnobacterium	-0.70	1.33e-17
o_Coriobacteriales;f_Coriobacteriaceae;g_	-0.70	5.39e-19
o_Lactobacillales;f_Streptococcaceae;g_Streptococcus	-0.69	8.99e-19
o_Bacteroidales;f_Rikenellaceae;g_	-0.69	3.10e-18
o_Bacteroidales;f_[Paraprevotellaceae];g_Paraprevotella	-0.68	1.29e-17
o_Clostridiales;f_Lachnospiraceae;g_Anaerostipes	-0.68	5.81e-18
o_Coriobacteriales;f_Coriobacteriaceae;g_Adlercreutzia	-0.68	1.36e-17
o_Clostridiales;f_[Mogibacteriaceae];g_	-0.66	6.19e-17
o_Clostridiales;f_Lachnospiraceae;g_[Ruminococcus]	-0.66	3.16e-16

Используя полученные p-значения, можно провести поправку на множественные сравнения для получения p-значения гипотезы об отсутствии линейной зависимости для всех бактерий. При использовании поправки Бенджамини- Хохберга с уровнем значимости 0.05, гипотеза об отсутствии линейной зависимости отвергается. Таким образом, подтверждено наличие линейной зависимости между представленностями до и после исследований, отрицательные значения коэффициэнта на таблице 1 означают, что чем больше была представленность бактерии, тем меньше будет ее изменение после эксперимента, в качестве примера можно привести изменение бактерии Prevotella, показанное на рисунке 10.

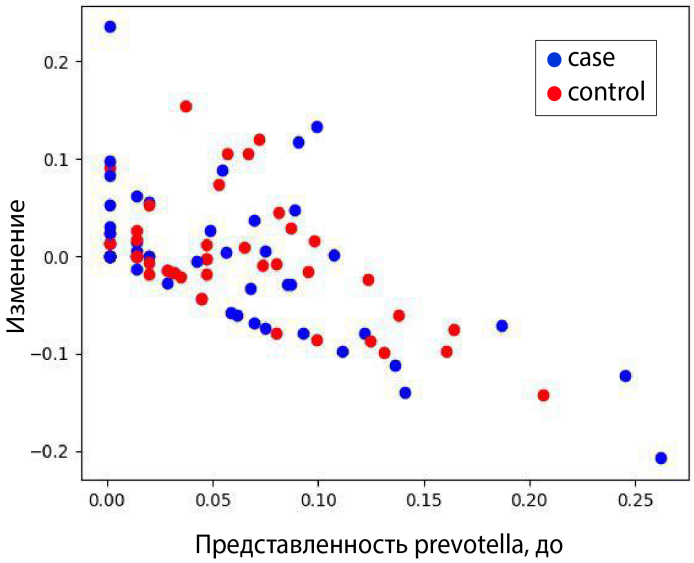

Рисунок 10 – График, отображающий изменение представленности бактерии $Prevotella$ в зависимости от изначальной представленности. На легенде группы case и control – экспериментальная и контрольная, соответственно.

Основываясь на обнаруженной линейной зависимости был выбран метод ANCOVA в качестве статистического метода для выделения бактерий, имеющих разное распределение между группами, и придания обнаруженным результатам статистической значимости. Для анализа функционального состава так же решено использовать данный метод, так как функциональный состав определяется на основании представленностей бактерий, участвующих в синтезе витаминов,

2.2.5. Поправка на множественное сравнение

При исследовании данных с помощью статистических методов существует вероятность получить ошибки второго рода, то есть получить ложное отклонение гипотезы. При множественном сравнении эта вероятность увеличивается, поэтому необходимо выполнять поправку полученных p-значений. Такая поправка позволит контролировать вероятность ошибки второго рода не для каждой бактерии, а для всего состава. При разработке компонента, отвечающего за статистический анализ таксономического и функционального состава микробиоты между группами клинического исследования, было предложено использован метод Бенджамини-Хохберга, контролирующий долю ложных отклонений нулевой гипотезы, т.е. принятия решения о том, что представленность бактерии зависит от группы исследования. Метод Бенждамини-

Хохберга требует независимости исследуемых статистик. Он заключается в отклонении нулевых гипотез H_i с p-значениями $p_i < \frac{i\alpha}{m}$, где m – количество гипотез, α – требуемый уровень значимости, а p_i - упорядоченные по возрастанию p-значения, полученный для каждой бактерии.

2.3. Интерактивная визуализация

2.3.1. Методы снижения размерности данных

Одними из основных способов визуализации данных для интерпретации результатов являются методы снижения размерности данных. Распространенным способом является отображение на двумерном графике получившихся размерностей, в надежде увидеть на итоговом графике образование таких закономерностей, как образование кластеров, разница полученных компонент между исследуемыми группами. Методы снижения размерности данных планируется использовать для визуализации представленностей бактерий.

2.3.2. Метод главных координат

Метод главных координат (PCoA, от англ. Principal Coordinates Analysis) – один из основных, широко используемых методов многомерного шкалирования. Основной идеей данного преобразования является расположение соответствующих объектов в новом пространстве меньшей размерности таким образом, чтобы евклидово расстояние между точками было максимально приближено к исходным (не обязательно евклидовым) расстояниям в первоначальном пространстве [15]. В случае, если новое пространство двумерное или трехмерное, полученные координаты точек можно использовать для визуализации.

Понижение размерности данных результата таксономического анализа с помощью метода главных координат выполнялось на стадии первичного анализа, который представляет собой отдельный тип анализа на платформе «Кномикс-Биота» и поэтому выходит за рамки этой работы. В исходных данных в новом отчете, разработанном для контролируемых исследований, принимается матрицы расстояний D, где элемент матрицы $d_{i,j}$ – мера бета-разнообразия для образцов UniFrac для i-того и j-того образца.

Далее, для каждого образца считаются координаты точек в трехмерном пространстве полученном методом главных координат, после чего массив данных сохраняется в формате JSON, который затем отображается на стороне клиента в браузере. Требовалось добавить поддержку таких функций как отобра-

жение векторов градиента представленностей наиболее представленных бактерий на получившемся графике, отображение связей между образцами для каждого пациента до начала эксперимента и после, отображение принадлежности к группе клинического испытания, выбор используемой компоненты в качестве координатной оси, фильтрация образцов, являющимися выбросами, отображение таксономического состава для каждого образца.

2.3.3. Кладограммы

Для визуализации результатов статистической проверки гипотезы о влиянии группы исследования на таксономический состав микробиоты использована кладограмма. Кладограмма представляет изображение филогенетическое дерево, в листьях которого находятся обнаруженные в процессе анализа рода, а в узлах таксоны более высокого порядка (классы, семейства и т.д.). На получившейся иерархической структуре цветом выделяются таксоны, между которыми обнаружены разница между группами клинического исследования. Цвет зависит от того, в какой группе перепредставлен конкретный таксон. Такая визуализация помогает увидеть согласование результатов анализа на разных таксономических уровнях. Пример кладограммы представлен на рисунке 15.

2.3.4. Боксплоты

Боксплот(от англ. box plot) – диаграмма для отображения распределения одномерной случайной величины на которой удобно показывать медиану, интересующие квантили. Данный вариант визуализации планируется использовать для отображения альфа-разнообразия. Пример представлен на рисунке 17.

Выводы по главе 2

В данной главе обозначены цели и задачи работы, подтверждена актуальность, был описан метод MaAsLin, на основе которого предлагается определять важность таких факторов как принадлежность к группе исследования или различная другая информация о пациента, таких как пол или возраст. Был обоснован выбор линейной модели для проведения статистического теста для таксономического и функционального состава, был выбран статистический критерий для определения разницы в альфа-разнообразии между группами.

Были сформированы необходимые требования к разработке интерактивной визуализации альфа- и бета- разнообразия, анализа таксономического и функционального состава.

ГЛАВА 3. РЕЗУЛЬТАТЫ

3.1. Детали реализации

3.1.1. Схема работы

Результаты первичного анализа, хранящиеся с помощью облачного сервиса для хранения данных Amazon S3, анализировались на сервере. Далее, полученные результаты сохранялись в JSON формате в имеющийся контейнер Amazon S3 для дальнейшей визуализиции на стороне Web-сервиса. Также формировались различные HTML компоненты, например таблицы и кладограммы, визализирующие результаты с помощью внутреннего HTML шаблонизатора. На рисунке 11 представлена схема работы платформы «Кномикс-Биота» с интегрированным в нее разработанным компонентом.

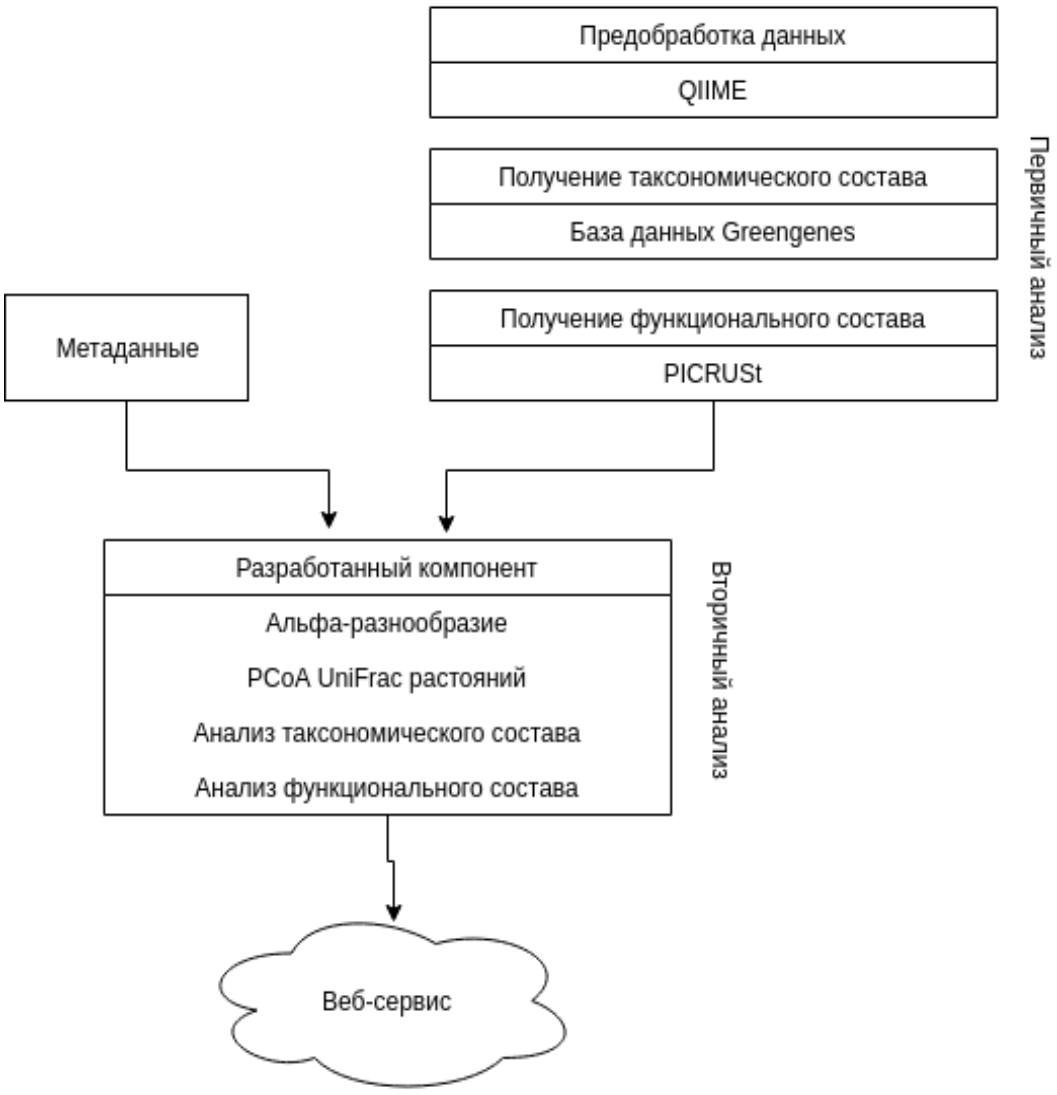

Рисунок 11 – Схема работы платформы «Кномикс-Биота».

3.1.2. Статистический анализ

Статистический анализ проводился для анализа таксономического и функционального состава, альфа-разнообразия и бета-разнообразия с помощью следующих библиотек для Python:

— модуль statsmodels для использования линейных моделей для ковариационного анализа.

— pymer4 для использования моделей со смешанными эффектами, является оберткой модуля lme4, использующегося в языке R, использующий метод максимального правдоподобия для обучения модели.

— библиотека skbio для вычисления бета-разнообразия UniFrac между образцами и последующего использования метода анализа главных координат.

3.1.3. Интерактивная визуализация

Имплементация интерактивной визуализации на стороне клиента, такой как визуализация результата метода главных компонент для таксономического состава, боксплоты для альфа-разнообразия, была реализована с помощью JavaScript-библиотеки React путем чтения результата вторичного анализа, хранящемся в формате JSON в хранилище Amazon S3.

3.2. Апробация разработанного компонента на имеющихся метагеномных данных

3.2.1. Исследование влияния пробиотика на состав кишечной микробиоты и на уровень тяжелых металлов в крови.

При тестировании разработанного компонента интересно провести анализ данных, по которым существует научная статья, для сравнения полученных результатов. Выбранный набор метагеномных данных представляет собой результаты клинического исследования, представленного в разделе 1.1.3.

На рисунке 12 представлен результат работы таксономического анализа, на котором видно, что представленность бактерий $Lactobacillus$ увеличилась в экспериментальной группе относительно контрольной. Отметим, что правильно определен таксон, который совпадает с бактериями $Lactobacillus\ rhamnosus$, содержащихся в пробиотике, который принимала экспериментальная группа, что подтверждает корректность работы разработанного компонента. Также одним из результатов данного исследования является изме-

нение концентрации тяжелых металлов в экспериментальной группе, относительно контрольной. Данный результат может объясняться многими факторами, одним из которых является факт, что структура клеточной мембраны грамположительных бактерий(в нашем случае это бактерии вида $Lactobacillus\ rhamnosus$) имеет более высокую активность в плане захвата металлов, чем грамотрицательные [2].

Overpresented in group: probiotic

taxon	taxa level	covariate	coefficient	p-value
g__Lactobacillus	genus	feature_abund_before:case_control[T.probiotic]	55.392	0.0

Рисунок 12 – Результат таксономического анализа, на котором показано увеличение представленности бактерии $Lactobacillus$ в экспериментальной группе относительно контрольной.

3.2.2. Плацебо-контролируемое исследование влияния приема пробиотического йогурта на состав микробиоты

3.2.2.1. Описание данных

Данные, на которых планируется проводить тестирование разработанного компонента представляют собой исследование кисломолочного продукта, содержащего пробиотик. В исследовании участвовало 157 пациентов, которые были поделены на две группы, принимавшую кисломолчный продукт с пробиотиком и без. Образцы биоматериала были взяты до и после эксперимента. На полученных результатах планируется написание и публикации статьи.

3.2.2.2. Анализ бета-разнообразия

Анализ бета-разнообразия проводился с помощью метрики UniFrac, полученные попарные расстояния использовались в качестве матрицы расстояний для метода анализа главных координат. На рисунке 13 изображен полученный результат. Визуализация позволяет выбирать координаты и отбрасывать образцы, являющиеся выбросами, также она отображает таксономический состав для каждого образца, связь между образцами до и после исследования, отображение групп с помощью различных форм и цвета.

Для метагеномных данных этого исследования данная визуализация бета-разнообразия не выявила каких-либо особых свойств, кроме того, что при отбросе образцов-выбросов наблюдается следующая закономерность: пациенты, образцы которых до эксперимента находятся в удалении от области

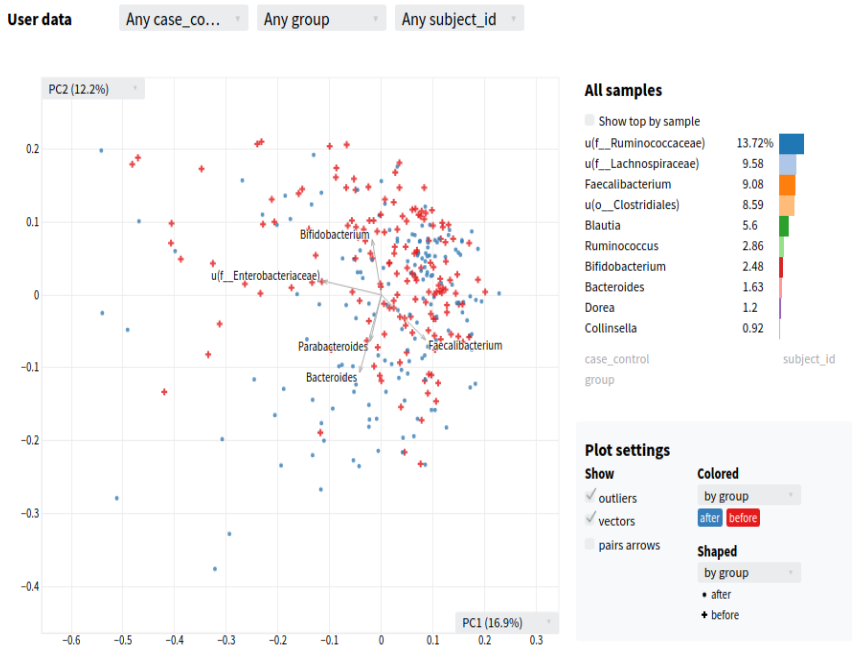

Рисунок 13 – Результат работы метода главных координат для матрицы расстояний UniFrac.

с наибольшей плотностью, имеют образцы из группы после эксперимента в области с большей плотностью. Это наблюдение можно интерпретировать как некоторую нормализацию таксономического состава пациентов. Отметим, что данное наблюдение встречается не впервые [13]. В подтверждение данный гипотезы можно привести увеличение представленностей бактерий семейства $Prevotellaceae$, что показано в разделе 3.2.2.3.

3.2.2.3. Анализ таксономического состава

На рисунке 14 представлен результат статистического анализа таксономического состава микробиоты. Как видно на получившейся таблице, в экспериментальной группе IP2 присутствуют бактерии, представленность которых увеличилась в ходе исследования по сравнению с контрольной группой. Отметим, что присутствует разница представленностей бактерий из семейств $Prevotella$ и $Collinsella$ между образцами до и после эксперимента в группе, принимавшей пробиотик относительно контрольной группы. Данные бактерии имеют большое влияние на здоровье человека [6, 19]. Также на рисунке 15 в качестве примера приведена кладограмма, на иерархической структуре которой показана разница представленностей бактерий между группами исследования.

Overpresented in group: IP2

taxon	taxa level	covariate	coefficient	p-value
f__Prevotellaceae	family	feature_abund_before:case_control[T.IP2]	0.566	0.000
g__Prevotella	genus	feature_abund_before:case_control[T.IP2]	0.570	0.000
g__Paraprevotella	genus	feature_abund_before:case_control[T.IP2]	0.441	0.004
g__Catenibacterium	genus	feature_abund_before:case_control[T.IP2]	0.331	0.000
g__Eggerthella	genus	feature_abund_before:case_control[T.IP2]	0.299	0.001
g__Prevotella s__copri	species	feature_abund_before:case_control[T.IP2]	0.655	0.000
g__Bacteroides s__uniformis	species	feature_abund_before:case_control[T.IP2]	0.450	0.003
s__u(g__Paraprevotella)	species	feature_abund_before:case_control[T.IP2]	0.446	0.004
s__u(g__Catenibacterium)	species	feature_abund_before:case_control[T.IP2]	0.332	0.000
g__Eggerthella s__lenta	species	feature_abund_before:case_control[T.IP2]	0.304	0.001

Рисунок 14 – Результат работы статистического анализа таксономического состава микробиоты.

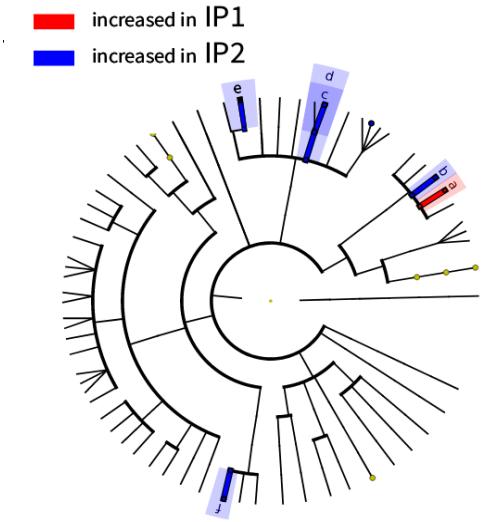

Рисунок 15 – Кладограмма для результата анализа таксономического состава. В зависимости от группы, цветом выделены таксоны, для которых обнаружено изменение представленности в течение эксперимента.

3.2.2.4. Анализ функционального состава

Статистический анализ данного набора метагеномных данных не выявил значимого влияния добавления пробиотика на функциональный состав микробиоты.

3.2.2.5. Анализ альфа-разнообразия

Анализ альфа-разнообразия, который представлен на рисунке 16, не выявил статистически значимой разницы между группами(p-значение превышает стандартный уровень значимости 0.05). Также на рисунке 17 в качестве примера представлена интерактивная визуализация альфа-разнообразия об-

разцов, визуализация позволяет выбирать группы образцов и визуализировать связи между образцами из группы до и после эксперимента.

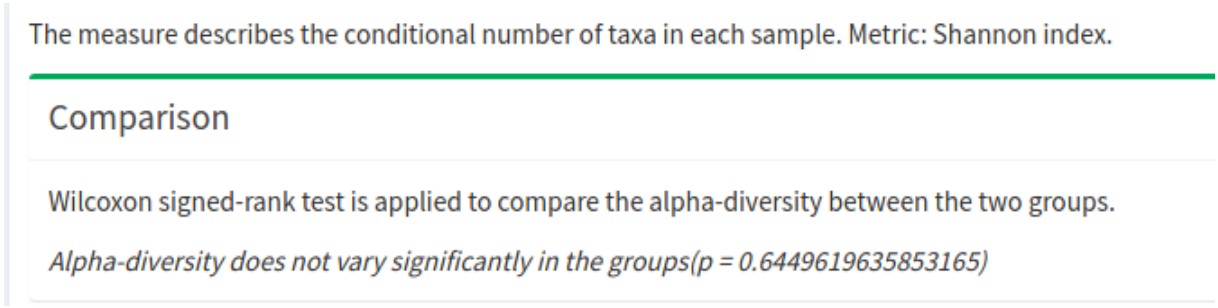

Рисунок 16 – Сравнение альфа-разнообразия между образцами.

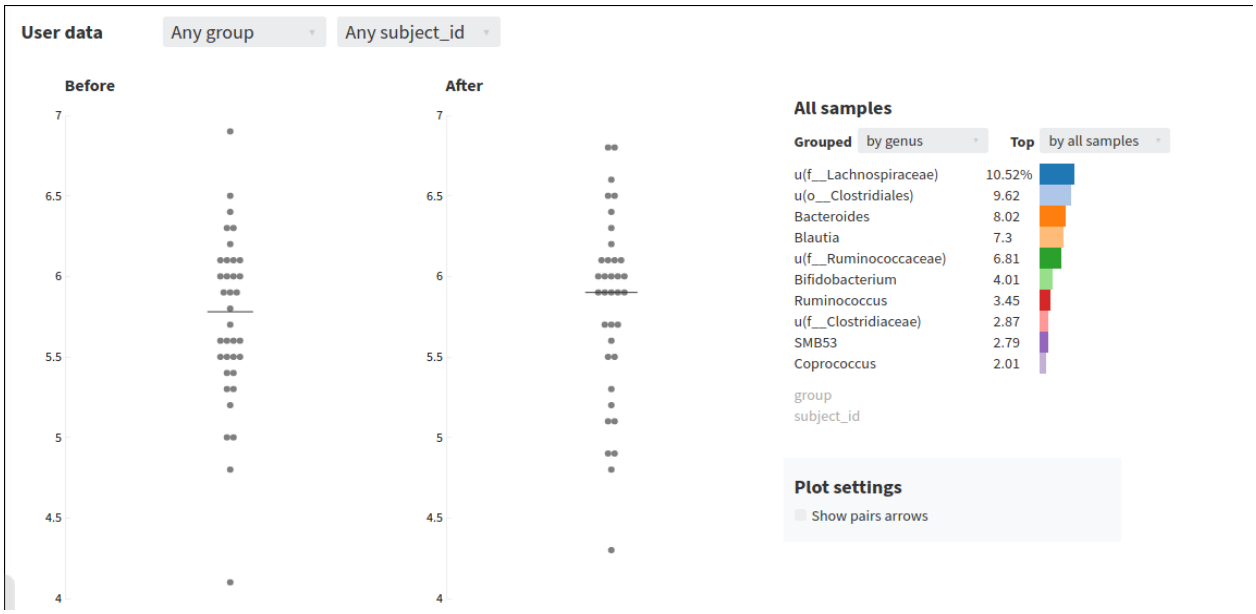

Рисунок 17 – Визуализация альфа-разнообразия между образцами.

Выводы по главе 3

В данной главе были представлены детали реализации, проведен пример работы компонента на реальных данных. Также предоставлен результат работы компонента для уже изученных и новых метагеномных данных для сравнения выводов и подтверждения корректности.

ЗАКЛЮЧЕНИЕ

Данная работа посвящена разработке алгоритма для анализа данных плацебо-контролируемых исследований влияния интервенции на микробиоту кишечника человека.

Для того, чтобы понимать характер таких данных, был сделан обзор литературы. В главе 1 приведено описание современных методов обработки метагеномных данных от секвенирования до получения таксономического и функционального состава. В качестве примера приведены последние результаты клинических исследований, связанных с модуляцией состава микробиоты. Были рассмотрены различные форматы клинических исследований, в том числе рандомизированное контролируемое испытание, принимаемое на текущий момент за эталон. Также описаны цели и задачи, стоящие перед платформами анализа микробиотных данных. Представлена разрабатываемая платформа «Кномикс-Биота» и ее сравнение с аналогами.

Основываясь на имеющихся алгоритмах этой платформы, было предложено два метода поиска ассоциаций между группой исследования и таксономическим и функциональным составом микробиоты: основанный на методе ANCOVA и на смешанной линейной модели. Проведено их сравнение на данных из данных исследования по эффективности пробиотической добавки в кисломолочном продукте. Сделан выбор в пользу метода, основанный на ANCOVA. Для визуализации исходных данных по таксономическому составу был использован метод главных координат (PCoA). Для визуализации результатов анализа таксономического и функционального состава – кладограмма.

Предложенные алгоритмы статистического анализа и визуализации были интегрированы в платформу «Кномикс-Биота» и представлены в виде отдельного, нового вида отчета – отчета по плацебо-контролируемым интервенционным исследованиям.

Разработанный компонент был протестирован на данных двух клинических исследований: по эффективности пробиотической добавки в молоко и эффективности пробиотической добавки в кисломолочном продукте. В первом исследовании выявлено увеличение по сравнению с контрольной группой представленности лактобактерий в группе, принимавшей пробиотик с лактобактериями. Кроме того, найденная ассоциация с лактобактериями косвенно

указывает на валидность выбранного метода. Во втором исследовании было выявлено увеличение представленности бактерий $Prevotella$.

Разработанный компонент расширяет общедоступный инструментарий для интерактивного анализа микробиотных данных устанавливающий связь между диетой и здоровьем человека, чем важно заниматься ввиду того, что микробиом очень важен для здоровья, и количество микробиомных данных увеличивается с каждым днем.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Тяхт А. В.* Функциональный анализ метагенома кишечника человека [Электронный ресурс]. — 2014. — URL: http://www.ibmc.msk.ru/content/thesisDocs/TyakhtAV_thesis.pdf.

2 *Beveridge T. J.*, *Fyfe W. S.* Metal fixation by bacterial cell walls // Canadian Journal of Earth Sciences. — 1985. — 22(12). — P. 1893–1898. — URL: https://doi.org/10.1139/e85-204.

3 Community proteogenomics reveals insights into the physiology of phyllosphere bacteria [Электронный ресурс] / N. Delmotte [et al.]. — 2009. — URL: https://www.ncbi.nlm.nih.gov/pubmed/19805315.

4 Culture of previously uncultured members of the human gut microbiota by culturomics [Электронный ресурс] / J. C. Lagier [et al.]. — 2016. — URL: https://www.ncbi.nlm.nih.gov/pubmed/27819657.

5 Dietary Interventions to Modulate the Gut Microbiome—How Far Away Are We From Precision Medicine [Электронный ресурс] / F. D. Filippis [et al.]. — 2018. — URL: https://academic.oup.com/ibdjournal/article-abstract/24/10/2142/4970097.

6 Distinct Genetic and Functional Traits of Human Intestinal Prevotella copri Strains Are Associated with Different Habitual Diets / F. De Filippis [et al.] // Cell Host Microbe. — 2019. — 25(3). — P. 444–453. — URL: https://www.ncbi.nlm.nih.gov/pubmed/30799264.

7 *Eckburg P.* Diversity of the human intestinal microbial flora // Science. — 2005. — 308(5728). — P. 1635–1643.

8 Enterotypes of the human gut microbiome / M. Arumugam [et al.] // Nature. — 2011. — 473(7346). — P. 1–7.

9 Human Gut Microbiome Response Induced by Fermented Dairy Product Intake in Healthy Volunteers [Электронный ресурс] / O. Volokh [et al.]. — 2019. — URL: https://www.mdpi.com/2072-6643/11/3/547.

10 Knomics-Biota - a system for exploratory analysis of human gut microbiota data [Электронный ресурс] / D. Efimova [et al.]. — 2018. — URL: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-018-0187-3.

11 *Lozupone C.*, *Knight R.* UniFrac: a New Phylogenetic Method for Comparing Microbial Communities // Applied and Environmental Microbiology. — 2005. — 12(71). — P. 8228–8235.

12 MaAsLin: Multivariate Association with Linear Models. [Электронный ресурс]. — 2012. — URL: http://huttenhower.sph.harvard.edu/maaslin.

13 Microbiome Responses to an Uncontrolled Short-Term Diet Intervention in the Frame of the Citizen Science Project [Электронный ресурс] / N. S. Klimenko [et al.]. — 2018. — URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5986456/.

14 Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. / J. Handelsman [et al.] // Chem. Biol. — 1998. — 5(10). — P. 245–249.

15 Multidimensional scaling [Электронный ресурс]. — URL: https://en.wikipedia.org/wiki/Multidimensional_scaling.

16 Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy / Q. Wang [et al.] // Appl. Environ. Microbiol. — 2007. — 16(73). — P. 5261–5268.

17 *Odintsova V.*, *Tyakht A.*, *Alexeev D.* Guidelines to Statistical Analysis of Microbial Composition Data Inferred from Metagenomic Sequencing [Электронный ресурс]. — 2017. — URL: https://www.ncbi.nlm.nih.gov/pubmed/28686566.

18 Post-Antibiotic Gut Mucosal Microbiome Reconstitution Is Impaired by Probiotics and Improved by Autologous FMT [Электронный ресурс] / J. Suez [et al.]. — 2018. — URL: https://www.ncbi.nlm.nih.gov/pubmed/30193113.

19 *Rajilic-Stojanovic M.*, *Willem M.* The first 1000 cultured species of the human gastrointestinal microbiota // Canadian Journal of Earth Sciences. — 2014. — 38(5). — P. 996–1047. — URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4262072/.

20 Randomized Open-Label Pilot Study of the Influence of Probiotics and the Gut Microbiome on Toxic Metal Levels in Tanzanian Pregnant Women and School Children [Электронный ресурс] / J. E. Bisanz [et al.]. — 2014. — URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196227.