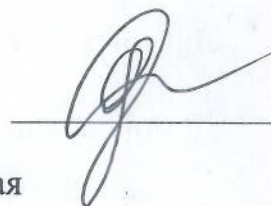


Министерство науки и высшего образования Российской Федерации  
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ,  
МЕХАНИКИ И ОПТИКИ»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**МЕТА-АНАЛИЗ МИКРОБИОТЫ ПИВА**

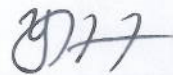
Автор: Копелиович Анна Юрьевна



Направление подготовки: 01.03.02 Прикладная  
математика и информатика

Квалификация: Бакалавр

Руководитель: Ульяновцев В.И., канд. техн. наук



**К защите допустить**

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

« \_\_\_\_ » \_\_\_\_\_ 20 \_\_\_\_ г.

Санкт-Петербург, 2019 г.

Студент Копелиович А.Ю.

Группа М3438 Факультет ИТиП

Направленность (профиль), специализация

Математические модели и алгоритмы в разработке программного обеспечения

Консультанты:

а) Тяхт А.В., канд. биол. наук

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 Копелиович А.Ю.

Группа М3438 Факультет ИТиП

Руководитель Ульяновцев В.И., канд. техн. наук, научный сотрудник ФИТиП

**1 Наименование темы:** Мета-анализ микробиоты пива

**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:** « \_\_\_\_ » \_\_\_\_\_ 20\_\_ г.

**3 Техническое задание и исходные данные к работе**

Развитие прототипа пайплайна для обработки метагеномных данных ITS-секвенирования на базе аналитической системы «Кномикс-Биота» (<http://biota.knomics.ru>) и статистическая обработка массивов опубликованных данных этого типа. Обработка новых данных от российских пивоваров.

**4 Содержание выпускной работы (перечень подлежащих разработке вопросов)**

1. Обзор существующих алгоритмов обработки метагеномных данных ITS-секвенирования.
2. Обработка и анализ массивов опубликованных данных.
3. Обработка и анализ новых данных.
4. Развитие существующей системы путем добавления бескластерной классификации.

**5 Перечень графического материала (с указанием обязательного материала)**

Графические материалы и чертежи работой не предусмотрены


**6 Исходные материалы и пособия**

- a) QIIME allows analysis of high-throughput community sequencing data / J. G. Caporaso [et al.] // Nature methods. 2010. Vol. 7, no. 5. P. 335
- б) Knomics-Biota - a system for exploratory analysis of human gut microbiota data / D. Efimova [et al.] // BioData Mining. 2018. Vol. 11, no. 1



7 Дата выдачи задания « \_\_\_\_ » \_\_\_\_\_ 20\_\_ г.

Руководитель ВКР



Задание принял к исполнению



« \_\_\_\_ » \_\_\_\_\_ 20\_\_ 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:** Копелиович Анна Юрьевна

**Наименование темы ВКР:** Мета-анализ микробиоты пива

**Наименование организации, где выполнена ВКР:** Университет ИТМО

**ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ**

1 Цель исследования: Развитие прототипа конвейера для обработки метагеномных данных ITS-секвенирования на базе аналитической системы «Кномикс-Биота».

2 Задачи, решаемые в ВКР:

- а) анализ микробиоты пива для формирования аналитической базы;
- б) адаптация бескластерной классификации.

3 Число источников, использованных при составлении обзора: 32

4 Полное число источников, использованных в работе: 44

5 В том числе источников по годам:

| Отечественных      |                   |                 | Иностранных        |                   |                 |
|--------------------|-------------------|-----------------|--------------------|-------------------|-----------------|
| Последние<br>5 лет | От 5<br>до 10 лет | Более<br>10 лет | Последние<br>5 лет | От 5<br>до 10 лет | Более<br>10 лет |
| 1                  | 0                 | 1               | 14                 | 9                 | 19              |

6 Использование информационных ресурсов Internet: да, число ресурсов: 7

7 Использование современных пакетов компьютерных программ и технологий:

| Пакеты компьютерных программ и технологий   | Раздел работы |
|---|---------------|
| Пакет QIIME для биоинформатического анализа | 2, 3          |
| Интерпретатор языка Python                  | 3             |
| Среда разработки PyCharm                    | 3             |
| Система контроля версий Git                 | 3             |

8 Краткая характеристика полученных результатов

Были проанализированы опубликованные ITS данные микробиоты пива. Помимо этого были обработаны новые данные российского пива. Конвейер анализа на базе системы «Кномикс-Биота» был улучшен и расширен путем добавления бескластерной классификации.

9 Гранты, полученные при выполнении работы

Отсутствуют

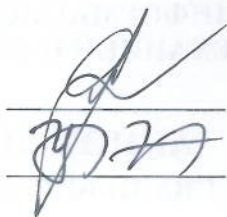
10 Наличие публикаций и выступлений на конференциях по теме работы

- а) Был представлен стендовый доклад «Food microbial consortium analysis: from dairy to beer» на конференции «Emerging applications of microbes», проходившей в Лёвене, Бельгия, 3-4 июня 2019 года

Студент            Копелиович А.Ю.

Руководитель    Ульяновцев В.И.

« \_\_\_\_ » \_\_\_\_\_ 20 \_\_\_\_ г.



## СОДЕРЖАНИЕ

|  |    |
|--|----|
| ВВЕДЕНИЕ .....   | 3  |
| 1. Обзор предметной области .....                                | 5  |
| 1.1. Термины и определения .....                                 | 5  |
| 1.2. Микробиота .....  | 5  |
| 1.2.1. Микробиота продуктов питания .....                        | 6  |
| 1.2.2. Связь микробиоты продуктов питания и человека .....       | 7  |
| 1.3. Дрожжи .....  | 8  |
| 1.4. Существующие подходы к анализу биологического материала ... | 9  |
| 1.4.1. Методика посева .....                                     | 9  |
| 1.4.2. Секвенирование .....                                      | 10 |
| 1.4.3. Полимеразная цепная реакция .....                         | 11 |
| 1.4.4. Масс-спектрометрия .....                                  | 12 |
| 1.5. Подходы к анализу результатов секвенирования .....          | 12 |
| 1.5.1. Классификация на основе выравнивания .....                | 13 |
| 1.5.2. Бескластерная классификация .....                         | 13 |
| 1.5.3. Оценка богатства сообщества .....                         | 14 |
| 1.5.4. Оценка попарного различия между сообществами .....        | 14 |
| 1.5.5. Методы понижения размерности и визуализации .....         | 14 |
| 1.5.6. Интерактивные платформы .....                             | 15 |
| 1.6. Пивоварение .....   | 15 |
| 1.6.1. Процесс производства .....                                | 15 |
| 1.6.2. Роль дрожжей .....  | 17 |
| 1.7. Постановка цели .....                                       | 17 |
| 1.7.1. Задачи .....  | 18 |
| 1.7.2. Актуальность .....  | 18 |
| Выводы по главе 1 .....  | 19 |
| 2. Аналитическое исследование .....                              | 20 |
| 2.1. Анализ данных BeerDecoded .....                             | 20 |
| 2.2. Анализ данных из статьи Bokulich'a .....                    | 23 |
| 2.3. Анализ новых данных российских малых пивоварен .....        | 24 |
| Выводы по главе 2 .....  | 30 |

|   |    |
|---|----|
| 3. Реализация бескластерной классификации ампликонных данных по уникальным прочтениям микробиоты пива ..... | 31 |
| 3.1. Описание формата данных .....  | 31 |
| 3.2. Подготовка данных .....  | 31 |
| 3.3. Deblur анализ и интерпретация результатов .....  | 32 |
| 3.4. Интерактивная визуализация и проверка на реальных данных....   | 32 |
| Выводы по главе 3 .....   | 33 |
| ЗАКЛЮЧЕНИЕ .....  | 34 |
| СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ .....  | 36 |



## ВВЕДЕНИЕ

В биохимическом составе пивной продукции ключевую роль играют дрожжи. Именно микробно-дрожжевое сообщество пива определяет его питательные и органолептические свойства. Пиво – самый употребляемый алкогольный напиток в мире. Последнее время в России увеличилось число небольших независимых пивоварен. Производимое ими пиво принято называть крафтовым. Количество потребляемого крафтового продукта растет. И, в отличие от крупных массовых производителей пива, которые в большинстве своем используют одноштаммовые рецептуры, в малом пивоварении часто экспериментируют с составами и пробуют рецепты, содержащие сложные микробные сообщества с малоизученными компонентами. Немалую роль в таких рецептах играют спонтанные брожения и добавление инновационных ингредиентов.

Проблемой является слабый контроль качества в малых пивоварнях. Чаще всего используют методики посева, имеющие ряд существенных недостатков. Самым главным является детектирование лишь немногих целевых культивируемых видов дрожжей. Подробнее этот и другие недостатки будут рассмотрены в первой главе.

Низкое качество мониторинга приводит к тому, что зачастую пивовары не знают точный состав ни во время приготовления, ни по окончании. Остается нераскрытым потенциал улучшения качества продукта и воспроизводимости целевых органолептических свойств. Незнание состава микробиоты продукта и ее связи с химическим составом снижает эффективность создания новых рецептур. Также остается открытым вопрос контроля качества и пищевой безопасности.

Основная цель работы – продвинуться в разработке интерактивной среды для коллаборативного анализа данных пивоварами и академическими исследователями из области пищевой микробиологии. Существование такой среды позволит усилить контроль качества. Также система будет помогать в открытии новых рецептур. Одновременно с этим интересно пытаться находить ассоциации свойств продукта с дрожжами и предсказывать различные параметры итогового продукта на основании его состава. Для этого необходимо иметь аналитическую базу и реализовать конвейер для автоматической обработки. Именно на это и нацелена данная работа.

В первой главе данной работы дан обзор предметной области: рассмотрены метагеномика, микробиота, способы ее изучения и методы обработки отсеквенированных данных. Помимо этого дан краткий обзор на дрожжи и пивоварение. После чего сформирована цель и поставлены задачи для ее достижения.

Во второй главе освещена аналитическая часть работы. Подробно представлен результат изучения опубликованных ранее микробиотных данных пива. Помимо этого представлен результат обработки новых данных российских пивоварен.

Третья глава посвящена рассмотрению способа реализации бескластерной классификации прочтений в системе «Кномикс-Биота». Рассказано про особенности анализа применительно к микробиоте пива. Рассмотрены результаты проверки на реальных данных.

В заключении рассказано про дальнейшие перспективы проекта в целом и про непосредственные дальнейшие шаги для его развития.

## ГЛАВА 1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Основной областью интереса данной работы является метагеномика. В первой главе будут разобраны необходимые термины и существующие подходы в области изучения микробиоты.

### 1.1. Термины и определения

**Метагеномика** — раздел молекулярной генетики, изучающий генетический материал, взятый из образцов окружающей среды. Основной задачей **метагеномного анализа** является определение видового разнообразия образца. Наиболее важным отличием от альтернативных подходов является возможность выделять и культивируемые, и некультивируемые микроорганизмы. Вместе с тем метагеномика позволяет определить метаболические взаимодействия и детально выявить механизмы функционирования всего разнообразия сообществ. **Метагеномом** принято называть совокупность геномных последовательностей, полученных непосредственно из образцов среды.

**Секвенирование** — получение из метагенома некоторого набора последовательностей нуклеотидов, называемых прочтениями.

**QIIME** — биоинформационная платформа, нацеленная на анализ микробных сообществ, полученных в результате секвенирования [30]. Является проектом с открытым исходным кодом.

**ITS** — internal transcribed spacer, внутренний транскрибируемый спейсер.

**UNITE** — публичная база для молекулярного детектирования грибных последовательностей. [38]

**ABV** — alcohol by volume, содержание алкоголя в напитке.

**IBU** — international bitterness units, шкала горечи напитка, обычно принимает значения от 0 до 100.

**Deblur** — бескластерная классификация ампликонных данных по уникальным прочтениям [12].

### 1.2. Микробиота

**Микробиота** — организованное микробное сообщество. Разнообразные микробные сообщества представлены практически везде. Интересно то, что в условиях экстремальных сред: гейзеры, атомные реакторы, — выживает не

единственный, сильнейший вид, а смешанное сообщество. Это позволяет задуматься о наличии сильных экологических связей между членами сообщества.

Концентрация бактериальных клеток сильно разнится между средами. Так, в водах щелочного озера примерно  $10^2$  клеток на миллилитр. А в почве примерно  $10^9$  клеток на грамм. В состав микробиоты входят бактерии, археи, вирусы и простейшие [2].

В рамках микробиоты человека чаще всего говорят о микробиоте кишечника. В нашем организме находится порядка сотни триллионов различных микроорганизмов. Установлено, что в случае нарушения баланса микробиоты кишечника человек страдает от различных заболеваний, как то: ожирение, болезнь Крона.

### **1.2.1. Микробиота продуктов питания**

Известно, что микробиотный состав продукта питания оказывает непосредственное влияние на вкус и аромат продуктов питания. Также известно, что отдельные виды микроорганизмов могут оказывать влияние на качество пищевого продукта: в совсем небольших количествах не представляют опасности, однако в случае нарушения обработки или хранения продукта они могут размножиться и привести не только к порче продукта, но и непосредственно к пищевому отравлению. А именно может произойти брожение, гниение или плесневение.

Стоит заметить, что часть из этих процессов под контролем может быть непосредственно частью производства продукта, микробы добавляются в продукт питания специально. Так спиртовое брожение используется в производстве алкогольной продукции. Кисломолочное брожение применяется для приготовления кисломолочных продуктов. А уксуснокислое брожение приводит к прокисанию вина и пива.

Интересен с точки зрения микробиоты и ее влияния на производственный процесс чайный гриб (его же по-другому называют комбуча). Дрожжи и микробы в нем вовлечены в метаболическое взаимодействие. В результате брожения дрожжей из фруктозы образуется этанол, а в процессе гидролиза сахара преобразуется в глюкозу. И именно этанол является источником для образования уксусной кислоты при участии углекислых бактерий. А из глюкозы, в свою очередь, получается глюконовая кислота [3].



Принято условно делить микробиотное сообщество пищевых продуктов на специфическое и неспецифическое. Штаммы микроорганизмов, применяющиеся в процессе технологического производства, принято относить к специфической микробиоте.

Случайно занесенные микробы, попавшие в процессе заготовки, доставки, переработки или хранения, относят к неспецифичной микробиоте. Источником таких микробов является животное, сырье, воздух, вода, оборудование или человек. Особую опасность представляют продукты питания, инфицированные патогенными микроорганизмами.

Помимо производства продуктов питания микробиота также важна для некоторых способов сохранения продуктов. А именно в результате специальной ферментации происходит сбраживание скоропортящихся веществ. В результате образуются консерванты – кислоты и другие микробные метаболиты. К такой ферментации относятся квашение, брожение в производстве вина и уксуса. Сыр сам по себе является примером биологической консервации. Сформировавшие его бактерии препятствуют развитию в нём других микроорганизмов.

Известно, что ферментация может увеличить питательные свойства продукта. Микробы внутри ферментированных продуктов вносят новые соединения в продукты. А также многие виды ферментированных продуктов богаты пробиотическими штаммами или же микроорганизмами, генетически подобными пробиотикам [19].

### **1.2.2. Связь микробиоты продуктов питания и человека**

Микробиота человека начинает формироваться с самого его рождения. Изначально важным фактором является, естественным ли образом рождается ребёнок, или же матери делают кесарево сечение. В случае естественного рождения обычно говорят о более богатой микробиоте. После этого сильно влияет, естественное или искусственное вскармливание у ребёнка. С молоком матери ребёнок получает и часть ее микробиоты. В случае же искусственного вскармливания часть необходимых представителей микробиоты не поступает в организм [37].

Далее же то, что ест человек в детстве, формирует его доминирующую микробиоту как напрямую, так и путем формирования необходимых бактерий для перерабатывания поступивших в организм веществ. С растительной пи-

щей, например, мы потребляем сахара, для расщепления которых у человека нет собственных ферментов. Однако бифидобактерии и лактобациллы успешно расщепляют сахара до лактата и ацетата. Эти метаболиты создают более кислые условия, что приводит к снижению размножения чувствительных к кислоте бактерий.

В рамках обсуждения микробиоты продуктов питания важно упомянуть проблему наличия антибиотиков в современных промышленных продуктах питания. Зачастую животным дают антибиотики в том или ином виде для скорейшего набора массы. В результате антибиотики не полностью покидают организм животного. Остаточные следы есть и в мясе, которое затем употребляет в пищу человек [5]. Антибиотики, в свою очередь, ведут к снижению разнообразия микробиоты человека с последующим очень длительным восстановлением – вплоть до нескольких лет [26].

Полноценная микробиота способна почти полностью перерабатывать энергию клетчатки. Микробиота со сниженным разнообразием же, как правило, способна расщеплять клетчатку до промежуточных метаболитов, а вот бактерии, специфически расщепляющие промежуточные метаболиты, в ней уже отсутствуют.

### 1.3. Дрожжи

**Дрожжи** — внетаксономическая группа одноклеточных грибов, утративших мицелиальное строение. Как и все грибы, являются эукариотами — имеют оформленное клеточное ядро, в котором располагается генетический аппарат, защищенный ядерной оболочкой. ДНК эукариот линейна в отличие от кольцевой прокариотной ДНК.

Дрожжи являются **хемоорганогетеротрофами** — способны использовать различные источники углерода. Используются органические соединения как для получения энергии, так и в качестве источника углерода. Кислород необходим им для дыхания, но в его отсутствие многие виды способны получать энергию за счет брожения с выделением спирта. И даже при доступе кислорода в случае высокого содержания глюкозы в среде дрожжи начинают ее сбрасывать [32].

Уже несколько тысяч лет люди активно используют разнообразные дрожжи. Например, дрожжи *Saccharomyces cerevisiae* используются в пекарном деле и в производстве алкоголесодержащих напитков. Ввиду этого люди

довольно давно обнаружили, что заражение неферментированных продуктов небольшим количеством ферментированных приводит к ускорению и стабилизации процесса брожения. Результат становится более предсказуем [14]. Соответственно, долгое время происходила селекция дрожжей и сформировались новые физиологические расы, не встречающиеся в природе.

*Saccharomyces cerevisiae* является первым эукариотом, геном которого был полностью секвенирован [23]. Потому по сей день является модельным организмом в генетических исследованиях.

Также дрожжи активно используются в современной медицине. Например, известно, что *Saccharomyces boulardii* снимает приступы острой диареи, и в целом благоприятно воздействует на микробиоту кишечника, поддерживая и восстанавливая ее. [16]. Помимо этого пивные дрожжи в жидком и твердом виде используют для производства различных лекарственных средств.

И, помимо всего вышеперечисленного, последнее время изучается возможность использования дрожжей в очистных процессах. Так, известно, что *Yarrowia lipolytica* разлагает алифатические, ароматические, нитроароматические и галогенированные соединения. [44]

#### **1.4. Существующие подходы к анализу биологического материала**

Чаще всего при проверке пищевых продуктов для изучения состава, в том числе микробиоты, используется методика посева. Однако благодаря достижениям последних лет доступнее становится секвенирование, позволяющее гораздо точнее оценить подробный состав микробиоты.

##### **1.4.1. Методика посева**

На первом этапе происходит забор образца с помощью микробиологической петли, после чего происходит культивирование на питательной среде. Посевы инкубируют в течение нескольких суток при комнатной температуре. В случае, когда речь идет о посеве дрожжей, на используемых питательных средах хорошо растут мицелиальные грибы, колонии которых сильно затрудняют учет и выделение дрожжей. Для замедления роста мицелиальных инкубировать посевы можно при низких температурах, однако тогда инкубация увеличивается до 3-5 недель.

После инкубации происходит анализ выросших на среде колоний. На первом этапе визуально выделяют основные типы по макроморфологическим

особенностям. После чего представителей каждого типа колоний выделяют в чистую культуру и отдельно изучают при помощи микроскопии с целью проверки однородности микробов в колонии. После установления чистоты обычно изучают культуральные признаки, морфологию, биохимические свойства. Чаще всего используют сахаролитические, протеолитические, пептолитические и гемолитические свойства.

Помимо затруднений, связанных с посторонними представителями, растущими на питательных средах, очевидным недостатком методики посева является ее применимость только для культивируемых бактерий. Однако большинство бактерий тяжело культивируются в искусственной среде. Также посев позволяет изучить свойства и состав отдельного типа бактерий, однако при изучении микробиоты немаловажна взаимосвязь между членами сообщества.

### 1.4.2. Секвенирование

Современным способом для изучения состава микробиоты является ДНК-секвенирование. Современное секвенирование активно развивается. Существующие на данный момент типы метогеномного секвенирования можно разделять по-разному.

- По составу:
  - ампликонное;
  - полногеномное;
- по производительности:
  - по Сенгеру;
  - высокопроизводительное;
- по длине ДНК-прочтений:
  - длинные (400-1000 — неск. тысяч);
  - короткие (50 — 100).

При ампликонном секвенировании читается последовательность не всей ДНК из образца, а лишь **ампликонов** — геномных фрагментов, выделенных посредством селективной амплификации. На некоторых платформах длина прочтения ампликона достигает порядка сотен пар нуклеотидов. Это позволяет в одно прочтение вместить большую часть типичного бактериального гена. Этого обычно достаточно для детальной филогенетической классификации бактерий. В качестве филогенетического маркера в случае метагеномики наиболее часто выбирают подпоследовательности гена 16S рРНК или ITS. Данный



вид секвенирования достаточно дешевый и подходит для количественного филогенетического профилирования микробиоты. Также важной особенностью является то, что результат секвенирования не нуждается в удлинении.

В качестве альтернативы существует полногеномное секвенирование. При таком подходе прочитываются фрагменты тотальной ДНК, выделенной из образца. Другим распространенным названием является «метод дробовика». В процессе секвенирования получается случайная массивная выборка клонированных фрагментов ДНК выбранного организма, на основе которых может быть восстановлена исходная последовательность ДНК. Данный метод секвенирования всё еще очень дорог, потому нечасто используется в метагеномике. Однако, в отличие от ампликонного секвенирования, полногеномное секвенирование позволяет проводить не только количественный анализ, но и функциональный.

Первым глобально распространенным автоматизированным методом секвенирования стал метод обрыва цепи, **секвенирование по Сенгеру** [33]. Основан на присоединении прямого или обратного праймера к одноцепочной последовательности ДНК и синтезе *de novo* комплементарной цепи с использованием дидезоксинуклеозидтрифосфатов в качестве терминаторов.

Следующим прорывом в области стало появление секвенирования следующего поколения [6]. Также известно как высокопроизводительное секвенирование. Позволяет прочесть одновременно сразу несколько участков генома, и именно это отличает его от более ранних методов. Некоторые методы: пиросеквенирование (Roche/454 Life Sciences), секвенирование с помощью синтеза (Illumina), секвенирование коротких прочтений, основанное на лигировании (Applied Biosystems/SOLiD).

В результате любого из подходов полученные данные пригодны для проведения биоинформатического анализа. Именно по его результатам из прочтений восстанавливается состав образца путем его детектирования по последовательности.

### **1.4.3. Полимеразная цепная реакция**

**Полимеразная цепная реакция** — ПЦР, метод молекулярной биологии, позволяющий создать копию определенного фрагмента ДНК из исходного образца, повысив его содержание в пробе на несколько порядков.

С помощью метода ПЦР в режиме реального времени можно определить уровень представленности бактерий одного или близких видов за счет таксон-специфичных праймеров. При этом является наиболее точным методом определения концентрации бактерий в образце. Классическая ПЦР может идти исключительно на матрице ДНК, поэтому когда для эксперимента используется матричная РНК, то сначала ее обращают в ДНК с помощью реакции обратной транскрипции. Амплификаторы для ПЦР в реальном времени значительно дешевле секвенаторов, а потому ПЦР в реальном времени является доступным способом анализа пищевых продуктов [4].

#### 1.4.4. Масс-спектрометрия

**Масс-спектрометрия** — метод, при котором для определения, что это за молекула, измеряют отношение ее массы к заряду в ионизированном состоянии. **МАЛДИ** — матрично-активированная лазерная десорбция/ионизация. В результате воздействия импульсов лазерного излучения на матрицу с образцом срабатывает десорбционный метод мягкой ионизации [15].

МАЛДИ масс-спектрометрия чаще всего применяется для анализа нелетучих высокомолекулярных соединений. На первом этапе на подложке масс-спектрометра смешивают биоматериал от колонии бактерий и специальную матрицу. После этого образец помещают в прибор и подвергают воздействию, в результате которого программное обеспечение прибора может получить спектр молекулярных масс. Масс-спектр сравнивается со спектрами из базы данных, и на основании сведений о массах характеристических белков происходит идентификация микроорганизмов. Метод применим в пищевой промышленности для выявления нежелательных микроорганизмов.

### 1.5. Подходы к анализу результатов секвенирования

В результате секвенирования могут быть получены миллиарды прочтений за несколько дней, что ставит перед биоинформатикой задачу анализа большого объема данных. В ручную обрабатывать большое количество образцов не представляется возможным, и потому перед биоинформатикой также стоит задача создания простых механизмов обработки данных в заданном формате.

В рамках обработки результатов секвенирования применяются методы прикладной информатики для проведения обработки, анализа больших дан-

ных, их классификации и хранения, а также методы прикладной математики для статистического анализа.

### **1.5.1. Классификация на основе выравнивания**

В результате секвенирования получается много коротких прочтений. После этого каждое прочтение сравнивают с базой проаннотированных ранее последовательностей. Это можно делать разными способами. Например, довольно популярным является способ хешировать прочтения. Также часто прибегают к слайдеру [34], который выравнивает, используя сортировку слиянием на референсных подпоследовательностях. Другим популярным способом является использование преобразования Барроуза-Уилера для сравнения строк, который применяется для выравнивания на референсную последовательность [22]. При любом из методов результатом выравнивания на референсную базу является состав микробиоты — вектор относительной представленности таксонов или генов.

### **1.5.2. Бескластерная классификация**

Тот факт, что в результате секвенирования в данных присутствует шум, ограничивает точное определение близкородственных бактерий. Однако некоторое время назад был предложен новый метод анализа субоперационно-таксономических единиц (sOTU), называемый Deblur [12]. Он использует профиль ошибок для получения безошибочных последовательностей. Основывается алгоритм на расстоянии Хемминга между рассматриваемыми последовательностями вместе с жадным алгоритмом.

Непосредственно алгоритм Deblur работает следующим образом. Все прочтения обрезаются до одной длины. Те, что короче необходимой длины, отбрасываются. Далее прочтения фильтруются одним из двух способов: отрицательная фильтрация (удаляются все последовательности, поданные пользователем) или положительная фильтрация (оставляются только последовательности, встреченные в одной из известных баз; задается пользователем). После этого последовательности выравниваются.

На следующем этапе последовательности сортируются по численности. После для последовательностей от наиболее к наименее населенной число предсказанных прочтений, полученных с ошибкой, вычитается из соседних относительно расстояния Хэмминга прочтений (верхняя граница профиля оши-

бок используется как функция расстояния Хэмминга). Далее все последовательности, населенность которых падает до 0, удаляются (классифицируются как шум). После этого происходит очистка данных от химер.

### 1.5.3. Оценка богатства сообщества

Богатство сообщества – по-другому **альфа-разнообразие**. Так как в случае анализа последовательности ITS сложно оценить число видов, классическое определение альфа-разнообразия как числа видов неуместно в данном случае. В связи с этим есть множество разных метрик для оценки альфа-разнообразия микробиоты. Общим для большинства является то, что учитывается влияние числа прочтений на число задетектированных видов. Строится график зависимости числа видов от числа используемых прочтений. И предсказание асимптоты для этой кривой дает оценку богатства сообщества. Наиболее применимы следующие оценщики разнообразия: филогенетическое разнообразие [17], оценщики Chao 1 [10] и ACE [35].

### 1.5.4. Оценка попарного различия между сообществами

Попарное различие между сообществами принято оценивать **бета-разнообразием**. Евклидова метрика не подходит из-за того, что занижает расстояние между образцами, у которых не представлены одни и те же таксоны. Обычно используют специальные меры, разработанные для анализа данных по составу сообщества. Используют **качественные** и **количественные** меры. Качественные учитывают только наличие таксона, а количественные также учитывают долю таксона. Одной из наиболее примечательных мер является UniFrac [24], в которой для пары сообществ строятся филогенетические деревья. На выходе обычно получается матрица расстояний.

### 1.5.5. Методы понижения размерности и визуализации

Именно матрица расстояний, полученная в результате расчета бета-разнообразия, используется для визуализации данных. Зачастую для начала на матрице применяются различные методы понижения размерности для упрощения визуализации, так как каждый метагеном описывается сотнями признаков. Самыми распространенными являются анализ главных координат (PCoA) [43] и многомерное шкалирование (MDS).



### 1.5.6. Интерактивные платформы

Для проведения биоинформатических исследований помимо отдельных программ и алгоритмов разрабатываются и целые системы. Основная их задача — собрать описанные выше методы воедино и облегчить метагеномные исследования.

Одной из таких систем является «Кномикс-Биота» [20]. Это веб-ресурс для исследовательского анализа метагенома. В первую очередь он нацелен на исследование кишечника человека, однако помимо этого есть возможность исследовать и микробиоту грибов. Пользователи в ней могут генерировать аналитические отчеты и обмениваться ими. Есть возможность подготовить базовый отчет, сделать case-control анализ или провести попарное сравнение. Интерактивная визуализация и статистический анализ представляются в контексте тысячи общедоступных наборов данных, объединенных в тематические коллекции. Помимо этого у пользователя есть возможность загрузить дополнительные метаданные. Каждый модуль в отчете сопровождается деталями реализации, чтобы пользователь мог самостоятельно повторить результат при необходимости или же описать методы в своей научной публикации.

В результате получается полноценная система для анализа состава микробиоты, на основе которой можно делать рекомендации по диете и впоследствии контролировать ее влияние на микробиоту кишечника.

Помимо этого существуют и другие системы для анализа метагеномных данных. Например, MG-RAST — нацелен на автоматический филогенетический и функциональный анализ [42]. Или Nephela — платформа для получения таксономической карты микробиоты [28].

## 1.6. Пивоварение

Важно понимать, что для каждого конкретного пива процесс может отличаться. Однако в общем и целом можно выделить главные, общие этапы производства [13].

### 1.6.1. Процесс производства

На первом этапе зерно (ячмень, пшеницу, овес или рожь) **проращивают** для получения солода — замачивают семена, сливают воду, а затем регулярно промывают семена, пока они не прорастут. После этого зерно сушат и в редких случаях обжаривают (обычно для получения карамельных сортов). После этого для раскрытия оболочки зерна его пропускают через дробилку.

После этого дробленый солод смешивают с водой и нагревают — **затирают**. В процессе затирания ферменты, содержащиеся в солоде, расщепляют крахмал, превращая его в сахара, которые впоследствии в результате брожения станут спиртом. Температура зависит от рецептуры, так как при разных температурах активируются различные ферменты. Получившаяся в процессе затирания жидкость называется суслом.

Следующий этап **сцеживание** — фильтрация сусла, отделение его от пивной дробины. Состоит из трёх этапов: мэш-аута (нагревание до 76 градусов, что останавливает ферментативные реакции и сохраняет в сусле сбраживаемые сахара), рециркуляции (приводит к формированию естественного фильтрующего слоя) и промывания дробины теплой водой, чтобы получить как можно больше сахара для сусла.

После того, как получено сусло, его стерилизуют при помощи кипячения. В процессе жидкость испаряется, а активность ферментов приостанавливается. Именно в процессе варки добавляют хмель.

Дальше начинается наиболее важный для органолептических свойств этап — **охмеление**. Очень важно понимать связь момента добавления хмеля и итоговых вкуса, аромата и горечи. Если хмель будет добавлен в самом начале варки, то напиток будет наиболее горьким. Если хотят получить наиболее яркий вкус, то добавляют хмель в середине процесса варки. Добавленный в конце хмель способствует насыщенному аромату. Также хмель можно добавлять на одном из последующих этапов приготовления.

Когда варка окончена, осуществляется **вихревое перемешивание** — этап, позволяющий сделать сусло прозрачным за счет удаления белков и частичек хмеля. Если делать это не в вихревом чане, а в хмелеотделителе, то сусло приобретает более яркий хмелевой вкус и аромат.

Далее сусло охлаждают до температуры, пригодной для **ферментации** и насыщают воздухом, необходимым для размножения дрожжей. Именно после этого добавляются дрожжи, и начинается основное брожение — сахара превращаются в спирт и углекислый газ. Продолжительность брожения и температура зависят от того, какое именно пиво хотят получить — верхового или низового брожения. Также иногда на этом этапе происходит добавление хмеля — сухое охмеление для усиления аромата.

После этого, в зависимости от стиля, могут быть отфильтрованы оставшиеся дрожжи.

Наконец, последний этап перед попаданием в бутылки — **созревание**. На этом этапе также может быть произведено сухое охмеление. Иногда созревание происходит в бочках, что добавляет во вкус разные необычные оттенки, особенно, если перед этим в этих бочках выдерживался какой-то другой напиток. Процесс может длиться от одной до шести недель.

В случае, если пиво хранится в холоде в течение примерно 30 дней, этот процесс называют **лагерированием**. Именно это делает лагер более прозрачным и оказывает немаловажное влияние на вкус. Также зачастую в производстве лагеров имеет место вторичное брожение, в процессе которого к ферментированному пиву добавляют другое, ферментируемое пиво с добавленными в него дрожжами. Дополнительная порция дрожжей активирует процесс диоксида углерода и способствует устранению продуктов главного брожения (диацетила и других соединений).

Далее пиво разливается в кеги или бутылки и газифицируется.

### 1.6.2. Роль дрожжей

Именно этап ферментации определяет, что же за пиво в итоге получится. Если смотреть наиболее обще, то есть два варианта развития событий: верховое брожения, происходящее обычно в тепле, и низовое, происходящее в холоде.

Верховое брожение проходит с использованием *Saccharomyces cerevisiae*. Продуктами верхового брожения являются эль, стаут, ламбик, вит, таппистен, гёз, альтбир и пшеничное пиво. Пиво, произведенное с помощью верхового брожения, хранится от 3 до 6 месяцев.

Лагеры сбраживаются с использованием *Saccharomyces pastorianus*. Продуктом низового брожения является лагер. Хранится до 2 лет. При низовом брожении гораздо меньше риск заражения сусла из-за низких температур непосредственно брожения.

### 1.7. Постановка цели

Основной целью является разработка среды для коллаборативного анализа данных пивоварами и академическими исследователями из области пищевой микробиологии и ее применение для анализа реальных данных (опубликованных и новых) по дрожжевому составу пива.

В рамках работы была поставлена цель расширить прототип среды для анализа данных, существующий сейчас в рамках освещенной выше системы «Кномикс-Биота» [20].

### **1.7.1. Задачи**

1. Основная задача — добавить в систему новый способ анализа данных. Было принято решение добавлять модуль бескластерной классификации на основе алгоритма Deblur, адаптированного к изучаемым данным.
2. После реализации нового модуля его необходимо было проверить на опубликованных ранее данных и сравнить полученные результаты. Помимо этого на обработанных результатах также получится отвалидировать реализованный ранее прототип системы.
3. Кроме этого работа подразумевала обработку и анализ новых данных российских крафтовых пивоварен с использованием получившейся системы.
4. Помимо этого было необходимо провести пилотное исследование с проведением факторного анализа на наборе метаданных, содержащих важные для пива характеристики, как то: ABV, IBU, рейтинг на популярных сайтах любителей напитка [41] [31], калорийность, стиль напитка и аромат.

### **1.7.2. Актуальность**

Прямо сейчас не существует доступного и понятного способа переиспользовать имеющиеся результаты в области биоинформатики конкретно для задач малого пивоварения. Однако общение с представителями области показало, что пивовары заинтересованы в улучшении качества своего продукта.

Также немаловажно то, что в США, например, рынок малого пивоварения это несколько десятков процентов от общего рынка пива. А в России последние годы доля крафтового пива растёт каждый год на рынке примерно на 50% [1]. Следовательно, можно предположить, что в ближайшее время рост российского крафта не остановится. А потому вопрос контроля качества на растущем рынке будет стоять всё острее.

## **Выводы по главе 1**

В данной главе были рассмотрены основные термины и понятия исследуемой области. Также были рассмотрены существующие подходы к анализу состава пивного продукта, недостатки существующего подхода посева. Помимо этого, были описаны различные способы дополнительного анализа на основе данных, полученных в процессе секвенирования, которые будут использоваться в конвейере анализа микробиоты пива.

Помимо этого была сформулирована цель работы и поставлены задачи для её достижения: расширение существующей системы за счет добавления нового модуля и проведение анализа отсеквенированных данных пива. Была обоснована актуальность и новизна работы.

## ГЛАВА 2. АНАЛИТИЧЕСКОЕ ИССЛЕДОВАНИЕ

В предыдущей главе были рассмотрены существующие способы анализа результатов секвенирования продукта.

В этой главе некоторые из них будут применены для анализа данных, представленных в посвященных текущей тематике работах. Также в данной главе будет рассказано о новых данных, обработанных в рамках данной работы, и о полученных результатах анализа.

### 2.1. Анализ данных BeerDecoded

Наибольший интерес для изучения представляла статья, посвященная проекту BeerDecoded [9]. В рамках этой статьи было обработано и проанализировано 39 различных образцов пива: 30 из Швейцарии, пять из Бельгии, два из Италии и по одному из Франции и Австрии. Так как отсеквенированные данные образцов находятся в открытом доступе, получилось независимо провести анализ и сравнить результат. Для начала была произведена качественная фильтрация с использованием скрипта *split\_libraries\_fastq.py* из QIIME [30], состоящая из обрезания низкокачественных прочтений (порог качества = 20) и отбрасывания прочтений, длина которых в результате обрезания составляет менее 75% от изначальной длины. Видно на рисунке 1, что ничего из данных образцов не было отфильтровано.

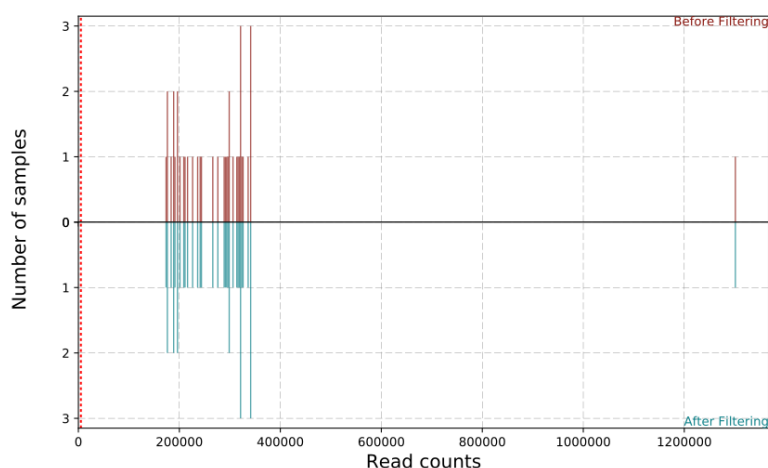


Рисунок 1 – Результат фильтрации образцов; верхняя половина рисунка показывает количество прочтений до фильтрации по длине, а нижняя – после неё; горизонтальная ось – число прочтений в образце, вертикальная ось – число соответствующих образцов



После этого отбрасываются образцы, у которых осталось менее пяти тысяч прочтений. В данном случае не было таких образцов, и все остались для дальнейшего исследования.

Далее прочтения с использованием алгоритма BWA-MEM [21] классифицировались с использованием базы UNITE версии 7.2 [38]. Все прочтения во всех образцах были успешно классифицированы (рис. 2). После этого отбрасываются образцы, у которых классифицировалось менее 70% прочтений, но в данном случае таких образцов не было.

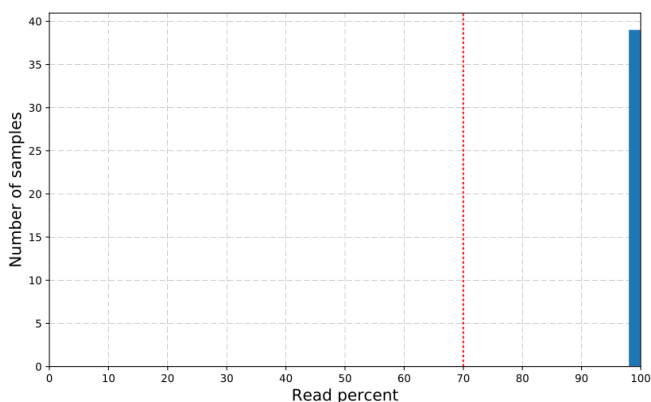


Рисунок 2 – Результат классификации образцов; горизонтальная ось процент успешно классифицированных прочтений в образце, вертикальная ось – число образцов

На следующем шаге был произведен таксономический анализ и построена интерактивная тепловая карта таксономического состава. Карта на рисунке 3 отображает относительное содержание основных микробных таксонов (столбцы) в образцах (строки) [7]. Можно выбрать разную детализацию для подробного изучения – состав по типу, классу, порядку, семейству, роду, виду и OTU. Для каждого разложения показаны несколько (максимум 10) самых представленных наименований и их процентное содержание в образцах.

После этого для имеющихся образцов были собраны метаданные для проведения факторного анализа. Для проверки процесса были использованы простые метаданные: IBU, ABV, название, стиль пива и рейтинг на сайте [41]. В результате была получена интерактивная таксономическая карта с возможностью фильтрации по метаданным и группировкой по стилю (рис. 4).

Отчетливо видно, что в большинстве сортов преобладают *Saccharomyces*, которые как раз и являются основой производства алкогольной

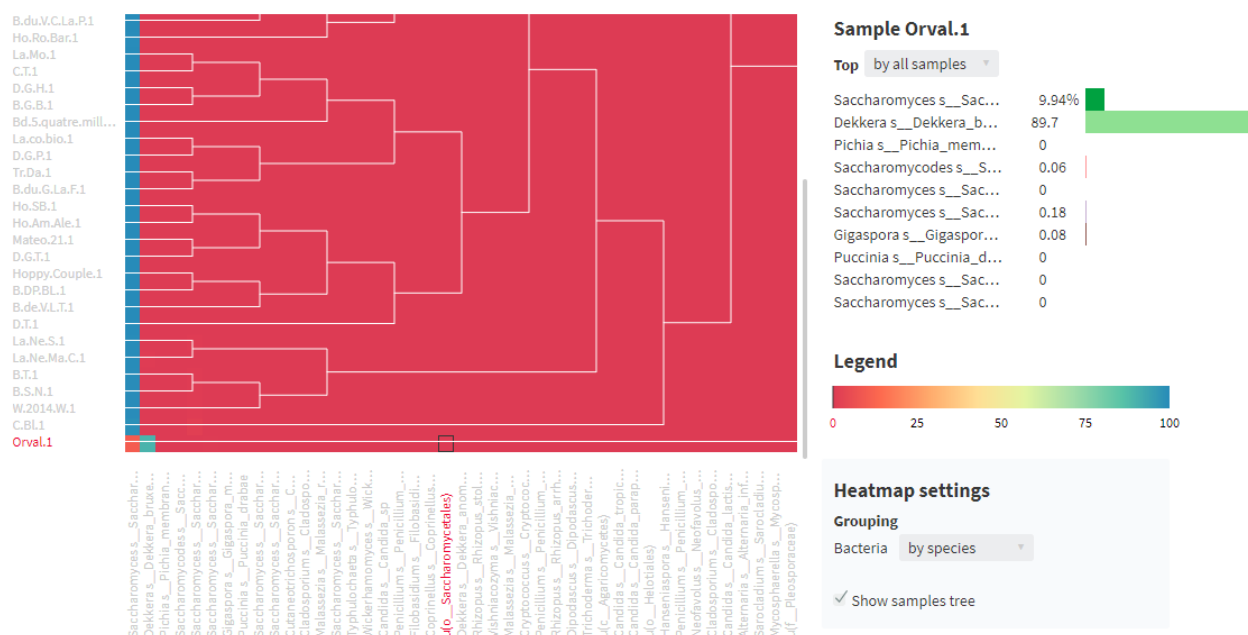


Рисунок 3 – Пример таксономического состава одного из образцов; с помощью цвета показывается процентное содержание таксона в образце

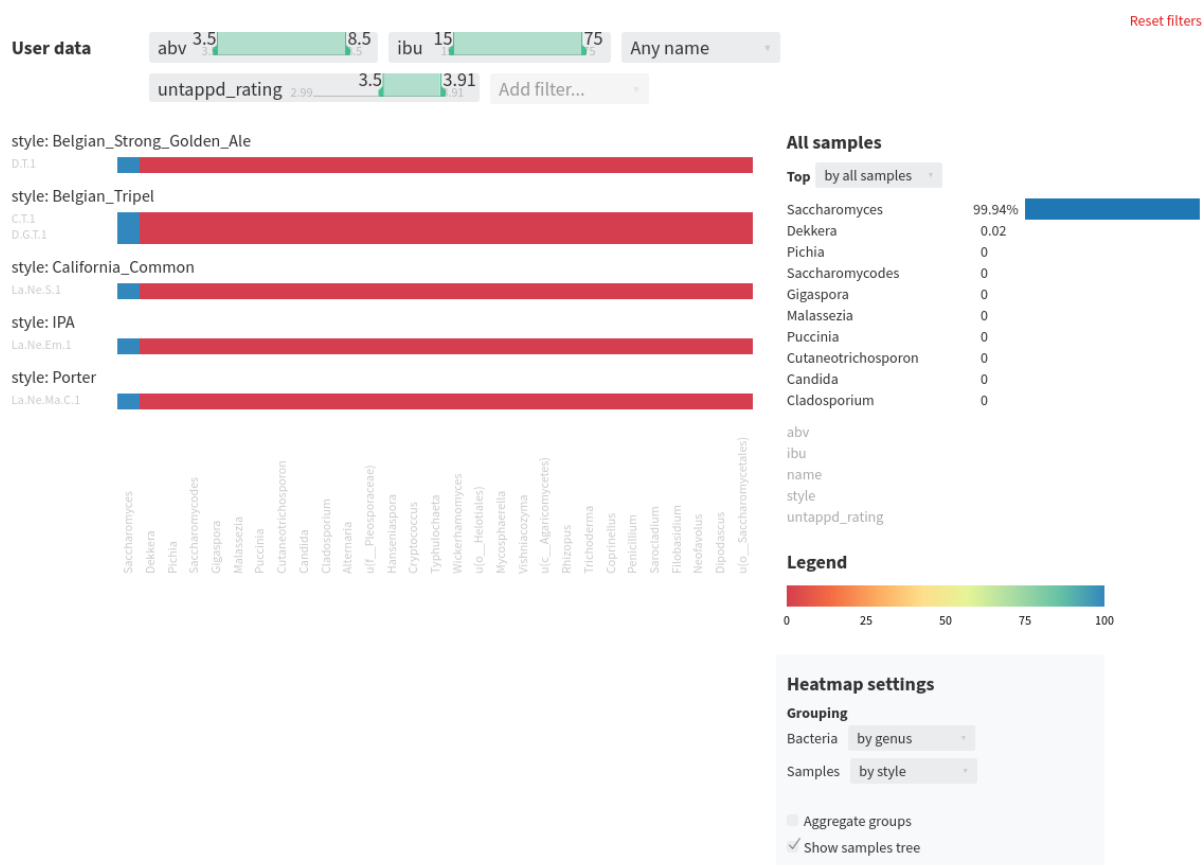


Рисунок 4 – Результат факторного анализа на таксономической карте с возможностью фильтровать образцы по признакам и группировать по стилю

продукции [40]. Однако помимо этого в одном из сортов доминирующим являлись дрожжи *Dekkera*, которые как раз характерны для бельгийских сортов пива (другое название этих дрожжей *Brettanomyces*) [39]. Также в незначительных количествах были обнаружены *Pichia* (дрожжи, считающиеся вредящими процессу брожения в виноделии, портящими качества вина; часто встречаются в орехах), *Gigaspora* (грибы, характерные для хвойных; обнаружены в продукте, анонсирующем содержание хвойных ноток в своем вкусе); *Malassezia* (дрожжи, обитающие на теле человека или животного); *Puccinia* (ржавчатые грибы, вызывающие заболевания пшеницы; поражает различные злаковые культуры); *Cutaneotrichosporon*; *Candida* (используются при приготовлении хлеба, пива, вина и кваса; постоянно проживают в кишечнике человека); *Cladosporium* (обитает повсюду в окружении человека; является возбудителем заболеваний).

Полученные результаты соответствуют результатами, приведенным в статье. Одновременно с тем данные из статьи являются хорошими представителями пивной продукции. Поэтому было принято решение использовать образцы для дальнейших сравнительных исследований. И именно этот набор данных был первым включен в аналитическую базу для Deblur анализа.

## 2.2. Анализ данных из статьи Bokulich'a

Дальше было принято решение проанализировать данные, посвященные микробиоте пивоваренного процесса в целом и пива в частности [25]. Однако в этом проекте были обнаружены только 16S рРНК данные, поэтому для них был проведен немного другой анализ по сравнению с предыдущими данными.

Из всего набора были использованы для анализа данные, относящиеся напрямую к пиву. Таких образцов в наборе оказалось пять. После этапа качественной фильтрации один образец был отброшен. А на остальных была проведена классификация. Далеко не все прочтения внутри образцов удалось классифицировать (рис. 5).

И по результатам классификации была построена таксономическая карта [8]. В трёх из четырёх образцов доминировал в составе *Lactobacillus*. В четвёртом образце (рис. 6) доминирует не классифицированная на уровне рода бактерия, относящаяся к семейству *Lactobacillaceae*. Это кисломолочные бактерии, используемые в пищевом производстве и часто встречающиеся в кис-

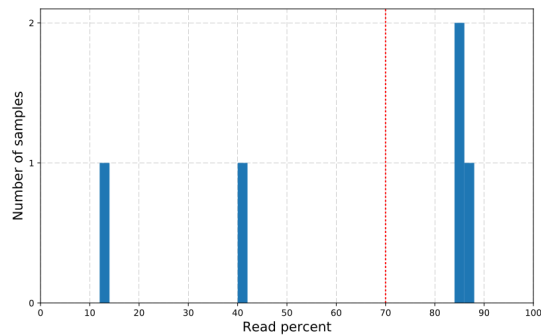


Рисунок 5 – Результат классификации образцов; горизонтальная ось процент успешно классифицированных прочтений в образце, вертикальная ось – число образцов

ломолочных продуктах, полученных в результате брожения. Встречаются и в пиве как продукт этапа затиария [29].

В тех же трёх образцах вторым крупным семейством является *Enterobacteriaceae*. В четвертом образце она также представлена, но занимает лишь небольшую часть. К этому семейству относятся многие патогенные бактерии [36].

В четвертом образце вторая по представленности последовательность не была определена даже на уровне семейства. Также в одном из образцов больше десяти процентов занимали бактерии, относящиеся к семейству *Aeromonadaceae*. Бактерии, относящиеся к этому семейству, часто встречаются в воде, некоторые из них являются патогенными [11].

В целом лишь небольшое число видов удалось соотнести с известными видами, как видно из графика 7. Бокс-графики представляют распределение относительной численности для 25 наиболее распространенных таксонов по всем выборкам (для каждого таксономического ранга). Для правильного отображения на логарифмической шкале нулевые значения заменялись псевдосчетом, не превышающим минимального значения относительного обилия основных таксонов.

### 2.3. Анализ новых данных российских малых пивоварен

В марте были получены первые опытные образцы продукции российских крафтовых пивоваров. Для исследования был один образец сидра и один образец французского крепкого эля. Так как для более качественного анализа лучше сравнивать больше образцов одновременно, было принято сделать не

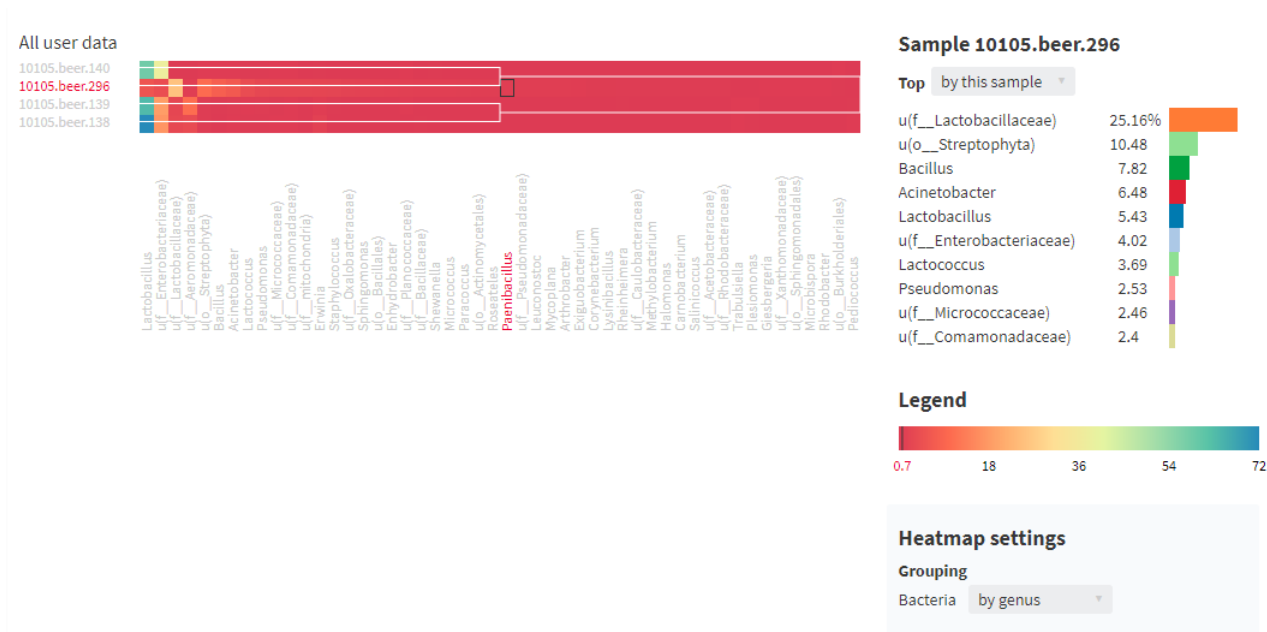


Рисунок 6 – Таксономический состав образца, больше всего отличающегося от остальных

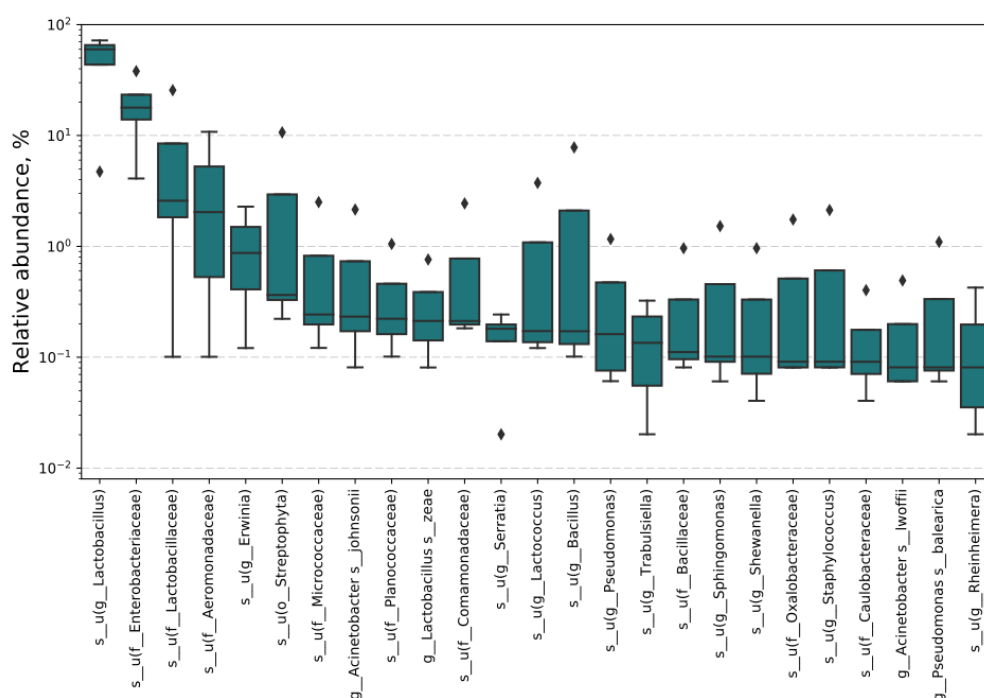


Рисунок 7 – 25 наиболее распространенных видов

только отдельный анализ образцов, но и вместе с образцами от BeerDecoded. Образец французского крепкого эля по составу был очень похож на остальные образцы – в его таксономии преобладали *Saccharomyces cerevisiae* и лишь незначительно содержались *Dekkera bruxellensis* (рис. 8).

Второй образец – сидр – имел более разнообразный таксономический состав (рис. 9): он совсем не содержал *Saccharomyces*, преобладал в нем *Dekkera*

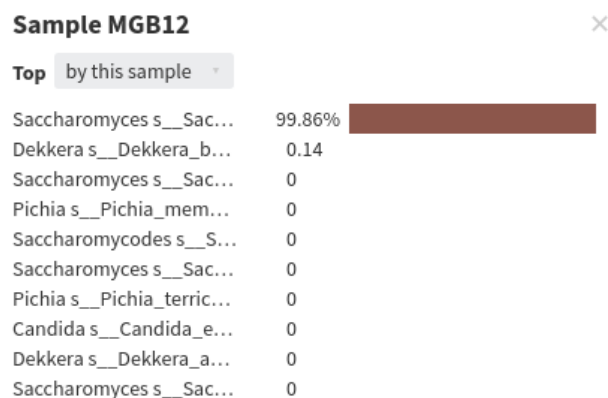


Рисунок 8 – Таксономический состав MGB12; представленные в образце таксоны и их процентное содержание

*bruxellensis*. В незначительных количествах были обнаружены *Pichia terricola* (часто встречаются во фруктовых соках); *Candida ethanolica* (ранее были обнаружены, например, в африканском пиве из сарго; встречается во многих консервированных итальянских оливках); *Dekkera anomala* (фактически испорченные дрожжи); *Kregervanrija fluxuum* (часто встречаются в винах, пиве, сидрах и засоленных продуктах; образуются на поверхности и относятся к плёночным дрожжам).

В совсем малом количестве (порядка нескольких сотых процента) были найдены также *Metschnikowia pulcherrima* (используется в виноделии для снижения крепости), *Issatchenkia orientalis* (используется в виноделии для снижения количества яблочной кислоты; часто обнаруживается в ферментированных фруктах), *Aspergillus* (содержится во многих продуктах питания), *Aureobasidium* (содержатся как правило во влажных гниющих субстратах, в сидре появился, скорее всего, ввиду того, что сидр был диким, из свежего сока яблок без пастеризации). Такое разнообразие возможно из-за немного иного производственного процесса относительно сидра.

Помимо получения таксономической карты был также произведен Deblur анализ каждого образца в отдельности, и совместно с образцами из BeerDecoded. Во время проведения Deblur только по нашим образцам два образца обрабатывались отдельно, потому что сидр состоял из коротких прочтений (96), а пиво из длинных прочтений (365). Результаты по выявленным таксонам можно увидеть в таблице 1.

На второй стадии пилотного проекта было проведено расширение выборки. Были отобраны еще десять интересных с точки зрения состава образ-

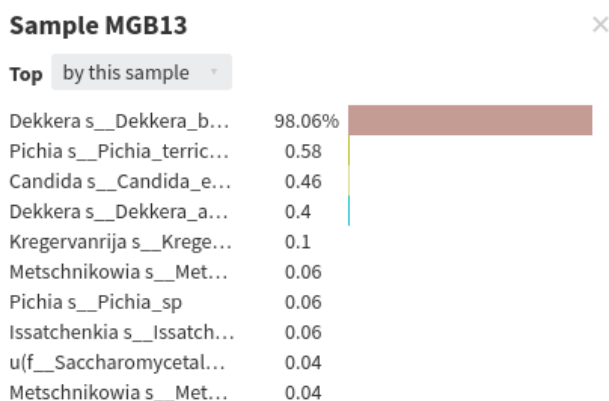


Рисунок 9 – Таксономический состав MGB13; наиболее представленные в образце таксоны и их процентное содержание

Таблица 1 – Таблица Deblur анализа двух российских образцов

| Taxon                    | MGB12 |
|--------------------------|-------|
| Saccharomyces cer. 1     | 2045  |
| Saccharomyces cer. 2     | 379   |
| Saccharomyces cer. 3     | 123   |
| Saccharomyces cer. 4     | 25    |
| Taxon                    | MGB13 |
| Brettanomyces br. 1      | 11130 |
| Brettanomyces br. 2      | 5130  |
| Brettanomyces anomalus 1 | 28    |
| Kregervanrija delftensis | 22    |
| Pichia kudriavzevii 1    | 10    |
| Brettanomyces anomalus 2 | 7     |
| Pichia kudriavzevii 2    | 5     |

цов и отправлены на секвенирование. Также на обработку был отправлен образец домашнего сидра. Однако в процессе секвенирования для трех образцов просто не удалось выделить ДНК ввиду того, что они были сильно отфильтрованными. В будущем это будет решаться за счет анализа небутилированных образцов, взятых до фильтрации. В результате для итогового анализа имелось два обработанных ранее образца, один домашний сидр и семь новых образцов российского крафта.

Их также последовательно обработали. Результат оказался очень интересным с точки зрения таксономического состава [27]. В разных образцах было много разнообразных грибных дрожжей. Был произведен факторный анализ с целью привязать таксоны к стилю пива (рис. 10).





Рисунок 10 – Таксономическая карта образцов российских пивоварен

Например, в саhti – финский натуральный рецепт – в единственном бы- ла представлена *Pichia norvegensis* (встречается как результат ферментации овсяной соломы). Образец, являющийся сидром домашнего приготовления, единственный содержал *Pichia kudriavzevii*. Помимо этого в нем обнаружено *Saccharomyces cariocanus*, который так же хорош с точки зрения пивоварения, как и *Saccharomyces cerevisiae*, но используется значительно реже.

Также для всех российских образцов был проведен совместный ана- лиз методом бескластерной классификации. Результаты представлены в таб- лицах 2, 3, 4. По результатам Deblur хорошо видно, что *Candida ethanolica* и *Pichia cecembensis* (в текущих исследованиях был обнаружен только в папайе) встретились только в sour ale, выдержанном в дубовых бочках, и, в меньшей степени, в диком сидре. Пиво стиля «Sour - Flanders Red Ale» имело наибольшее разнообразие различных таксонов, а преобладали в нем *Pichia fermentans* (зачастую используется для контроля процесса гниения фруктов; также часто является продуктом почти продуктов питания).

Таблица 2 – Таблица Deblur анализа российских образцов на коротких прочтениях

| Taxon                            | MGB2 | MGB9 | MGB10 | MGB13 |
|----------------------------------|------|------|-------|-------|
| <i>Candida ethanolica</i> v1     | 520  | 0    | 0     | 0     |
| <i>Pichia membranifaciens</i> v1 | 57   | 0    | 0     | 0     |
| <i>Candida ethanolica</i> v2     | 39   | 0    | 0     | 75    |
| <i>Pichia membranifaciens</i> v2 | 4    | 77   | 0     | 4     |
| <i>Pichia terricola</i>          | 0    | 0    | 0     | 99    |
| <i>Pichia fermentans</i> v1      | 0    | 0    | 33    | 0     |
| <i>Pichia fermentans</i> v2      | 0    | 0    | 10    | 0     |

Таблица 3 – Таблица Deblur анализа российских образцов на средних прочтениях

| Taxon                             | MGB2 | MGB7 | MGB8 | MGB9 | MGB10 | MGB13 |
|-----------------------------------|------|------|------|------|-------|-------|
| <i>Issatchenkia orientalis</i> v1 | 0    | 113  | 0    | 0    | 9     | 5     |
| <i>Issatchenkia orientalis</i> v2 | 0    | 46   | 0    | 0    | 0     | 2     |
| <i>Dekkera bruxellensis</i> v1    | 1184 | 0    | 39   | 149  | 26    | 0     |
| <i>Pichia cecembensis</i>         | 156  | 0    | 0    | 0    | 0     | 10    |
| <i>Dekkera bruxellensis</i> v2    | 131  | 0    | 106  | 0    | 0     | 11165 |
| <i>Dekkera bruxellensis</i> v3    | 98   | 0    | 71   | 0    | 0     | 5152  |
| <i>Dekkera anomala</i>            | 0    | 0    | 0    | 0    | 0     | 28    |
| <i>Kregervanrija delftensis</i>   | 0    | 0    | 0    | 0    | 0     | 22    |

Таблица 4 – Таблица Deblur анализа российских образцов на длинных прочтениях

| Taxon                        | MGB2 | MGB7 | MGB8 | MGB11 | MGB12 | MGB13 |
|------------------------------|------|------|------|-------|-------|-------|
| <i>Saccharomyces</i> cer. v1 | 6    | 15   | 351  | 0     | 1111  | 9     |
| <i>Saccharomyces</i> cer. v2 | 0    | 0    | 0    | 0     | 202   | 0     |
| <i>Saccharomyces</i> cer. v3 | 0    | 0    | 0    | 0     | 67    | 0     |
| <i>Saccharomyces</i> cer. v4 | 0    | 0    | 0    | 0     | 15    | 0     |
| <i>Saccharomyces</i> cer. v5 | 0    | 0    | 77   | 0     | 0     | 0     |
| <i>Saccharomyces</i> cer. v6 | 0    | 0    | 0    | 8     | 0     | 3     |

После изучения непосредственно новых данных из интереса был проделан совместный анализ для новых данных российских пивоварен совместно с BeerDecoded данными [9]. По результатам совместного анализа видно (рис. 11), например, что всё еще преобладает *Saccharomyces* и *Dekkera*. Однако, например, в Brett DIPA российского набора и в Coudres Pale Ale набора, ис-

пользуемого для сравнения присутствует специфичная *Pichia membranifaciens* (часто обнаруживается в бродящих напитках).

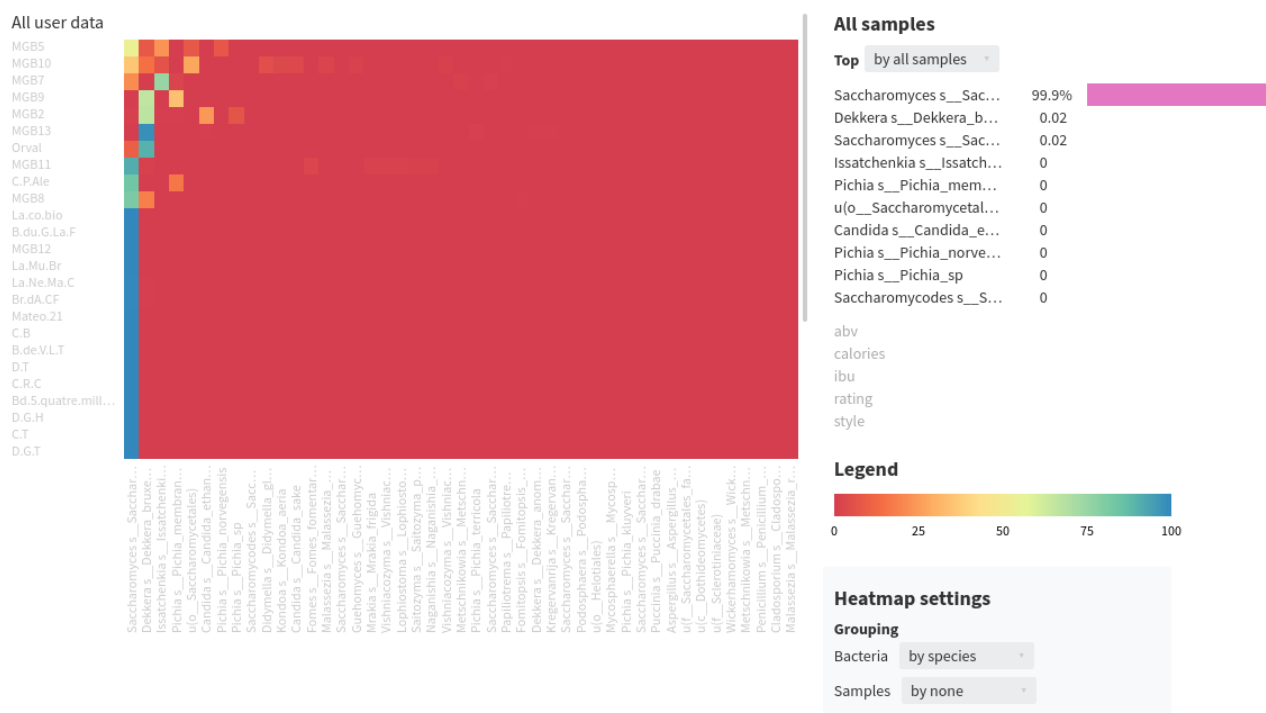


Рисунок 11 – Таксономическая карта образцов российских пивоварен совместно с BeerDecoded

Полученные результаты, описанные выше, были представлены на конференции [18].

Важно отметить, что дрожжи, найденные в российских образцах, встречаются в других продуктах питания.

## Выводы по главе 2

В данной главе были приведены результаты анализа данных микробиоты пива из ранее опубликованных работ. Помимо этого были обработаны новые, не представленные нигде ранее данные российского пива производства малых пивоварен. Для этих данных был произведен таксономический анализ, а также подготовлен анализ на основании факторов с использованием метаданных.

Теперь, после проверки на простых метаданных, в систему можно загружать более детальную метаинформацию по образцам, как то: оценку органолептических свойств экспертом, уровни летучих метаболитов, отвечающих за свойства продукта, полученные из метаболомных экспериментов, оценка органолептических свойств простым потребителем (с сайтов заинтересованного интернет-сообщества [41]).

## ГЛАВА 3. РЕАЛИЗАЦИЯ БЕСКЛАСТЕРНОЙ КЛАССИФИКАЦИИ АМПЛИКОННЫХ ДАННЫХ ПО УНИКАЛЬНЫМ ПРОЧТЕНИЯМ МИКРОБИОТЫ ПИВА

В данной главе будет рассказано про особенности технического добавления бескластерной классификации ампликонных данных по уникальным прочтениям в систему «Кномикс-Биота». Выбранный тип классификации и алгоритм, а также результаты запуска на исследуемых образцах представлены выше, а в данном разделе будут представлены непосредственно технические особенности внедрения.

### 3.1. Описание формата данных

Результатом секвенирования является множество прочтений, сопровождаемых баллами качества. Прочтения состоят из четырех символов нуклеотидов – А, С, G и Т. Также в прочтениях встречается служебный символ N (иногда его заменяют на '.') – полная неопределенность значения в данной позиции. **Балл качества** — значение, которое характеризует вероятность отсутствия ошибки в данной позиции. Вычисляется программным обеспечением секвенатора исходя из качества сигнала:

$$Q = -10 \log_{10} P$$

$P$  — вероятность ошибки в данной позиции.

Обычно прочтения и баллы качества генерируются в виде двух отдельных файлов для каждого образца — *FASTA*, либо прочтения и баллы качества объединяют в единый файл — *FASTQ*.

### 3.2. Подготовка данных

В начале происходит фильтрация прочтений предоставленных образцов. Для этого используется QIIME [30], функцией `split_libraries_fastq.py`. Фильтрация является не обязательной, однако она позволяет получить более качественные результаты анализа.

После фильтрации в случае для 16S данных часто можно сразу запускать Deblur анализ. Однако для ITS данных важен тот факт, что все прочтения будут обрезаны до одинаковой длины. Это приведет к тому, что либо много данных будет отсеяно, либо длинные прочтения будут исследованы не так качественно, как могли бы. Из-за этого было принято решение после фильтрации производить статистическое исследование прочтений.

В рамках статистического исследования данные обрабатывались с помощью программы на языке python. Подсчитывается, сколько прочтений какой именно длины встречается. После вычисления статистики производился анализ плотности данных. Данные разбиваются на короткие и длинные прочтения основываясь на кластерах.

### **3.3. Deblur анализ и интерпретация результатов**

После того, как данные готовы, каждая из групп анализируется методом Deblur [12] отдельно, а в качестве минимальной длины для Deblur подается минимальная длина в группе.

Для положительной фильтрации по дополнительной базе использовалась база UNITE [38]. После этого у нас есть статистика по каждой из групп с закодированной информацией OTU в формате \*.biom, и отдельно встреченные подпоследовательности OTU.

Для получения таксономии используется QIIME, метод feature-classifier, которому на вход подаются встреченные и не отфильтрованные в процессе анализа последовательности. После этого происходит сопоставление закодированных данных об OTU и реальных цепочек с информацией об их таксономии.

После этого происходит объединение полученных данных для обеих групп.

### **3.4. Интерактивная визуализация и проверка на реальных данных**

Полученные результаты анализа всё еще довольно плохо воспринимаются неподготовленным человеком, и потому их необходимо визуализировать (рис. 12, 13).

Для этого использовались существующие наработки системы «Кномикс-Биота», которые позволяют из таксономического состава нарисовать интерактивную тепловую таксономическую карту.

Для того, чтобы отвалидировать получившийся модуль, использовали описанные выше данные российских крафтовых пивоваров и результаты ручного проведения Deblur анализа на этих данных. На основании запусков были найдены некоторые неточности в первоначальной идее обрезать до самой широко представленной длины в кластере. Именно по результатам запусков было принято в автоматической системе использовать ту же логику, что и при ручных запусках.

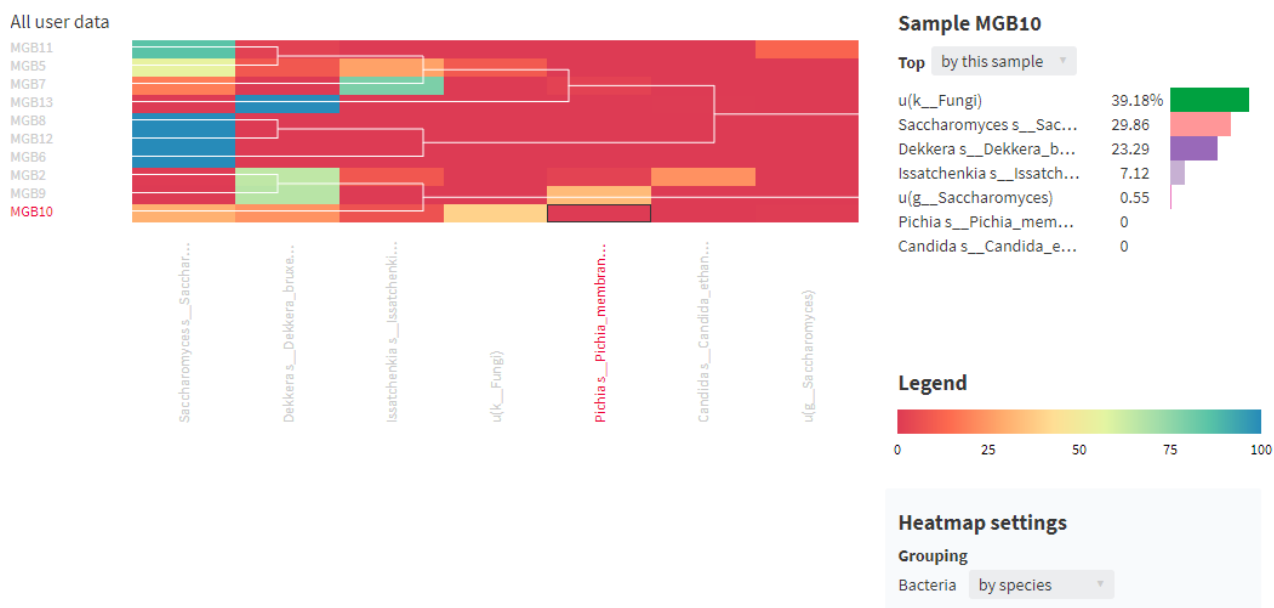


Рисунок 12 – Таксономический состав, полученный по результатам автоматического Deblur анализа из российских образцов

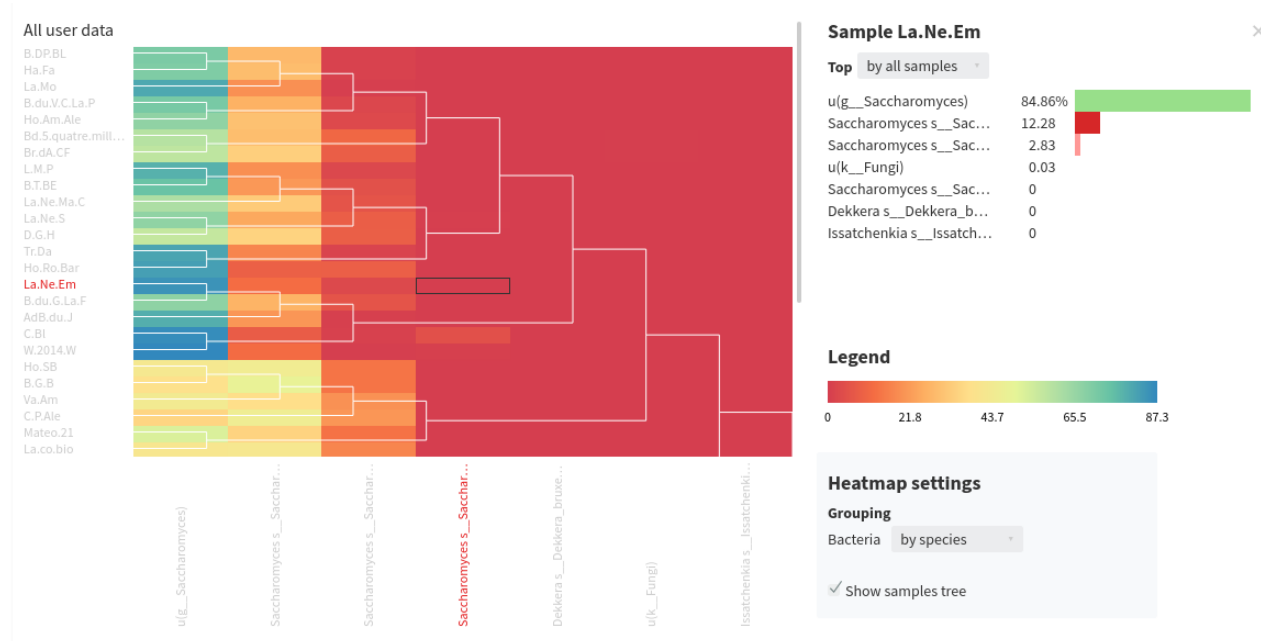


Рисунок 13 – Таксономический состав, полученный по результатам автоматического Deblur анализа из образцов Beerdecoded

### Выводы по главе 3

Был реализован и внедрен в систему «Кномикс-Биота» новый способ анализа ITS данных. Полученный метод был отвалидирован на реальных данных. Теперь в системе можно загружать свои данные и получать визуализированные результаты Deblur анализа.

## ЗАКЛЮЧЕНИЕ

В работе приведен обширный анализ предметной области. Рассмотрены особенности микробиоты продуктов питания, способы изучения микробиоты и их недостатки по сравнению с секвенированием. Помимо этого рассмотрены различные способы анализа отсеквенированных данных.

В рамках работы была достигнута поставленная цель — развитие интерактивной среды для коллаборативного анализа данных пивоварами и академическими исследователями из области пищевой микробиоты. Все задачи были успешно выполнены. Опубликованные ранее данные успешно обработаны с помощью системы «Кномикс-Биота», и полученные результаты совпадали с представленными авторами статей. В большинстве образцов преобладала типичная для пивоварения *Saccharomyces cerevisiae*. И в одном образце преобладала *Dekkera bruxellensis*, характерная для бельгийских сортов пива.

Новые данные российских пивоварен были проанализированы, а результаты были представлены на конференции «Emerging applications of microbes» в формате постерного доклада. В российских образцах преобладали всё еще *Saccharomyces cerevisiae* и *Dekkera bruxellensis*. Однако помимо этого в кислых сортах пива были обнаружены *Candida ethanolica* и *Pichia cecembensis* в больших количествах. В образце саhti было значительное количество *Pichia norvegensis*. В домашнем сидре в значительных количествах содержалась *Pichia kudriavzevi*.

Полученный результат имеет практическое применение при разработке новых рецептов, так как обнаруженные дрожжи можно попробовать специально добавить в пиво на этапе ферментации для попытки достижения аналогичных с исходным продуктом свойств.

На обрабатываемых образцах также был проведен факторный анализ с добавлением простых метаданных (например, стиль пива, IBU, ABV). Система корректно их обрабатывала и позволяла оценивать влияние состава на определенные факторы. В дальнейшем в систему можно загружать более сложные метаданные, чтобы оценивать их влияние.

В рамках работы в систему «Кномикс-Биота» была добавлена возможность автоматической бескластерной классификации ампликонных данных по уникальным прочтениям. В качестве алгоритма классификации был использован алгоритм Deblur, адаптированный к особенностям анализируемых данных.



Работа нового модуля была провалидирована на имеющихся ITS данных по микробиоте пива. Уже сейчас новый модуль доступен на сайте для свободного использования.

Дальнейшее развитие системы подразумевает обработку еще большего числа образцов и добавление новых методов анализа данных.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

- 1 Пивное дело 2-2019. Россия: рынок пива усложняется [Электронный ресурс]. — URL: <https://www.pivnoe-delo.info/2019/05/20/pivnoe-delo-2-2019-rossiya-rynok-piva-uslozhnyaetsya/>; дата обращения: 25.05.2019.
- 2 Тяхт А. В. Функциональный анализ метабенома кишечника человека : дис. ... канд. б. наук / Тяхт Александр Викторович. — 2014.
- 3 A Review on Kombucha Tea-Microbiology, Composition, Fermentation, Beneficial Effects, Toxicity, and Tea Fungus / R. Jayabalan [et al.] // Comprehensive Reviews in Food Science and Food Safety. — 2014. — Vol. 13, no. 4. — P. 538–550.
- 4 Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes / J. Vandesompele [et al.] // Genome biology. — 2002. — Vol. 3, no. 7. — research0034–1.
- 5 Animal antibiotic use has an early but important impact on the emergence of antibiotic resistance in human commensal bacteria / D. L. Smith [et al.] // Proceedings of the National Academy of Sciences. — 2002. — Vol. 99, no. 9. — P. 6434–6439.
- 6 Ansorge W. J. Next-generation DNA sequencing techniques // New Biotechnology. — 2009. — Vol. 25, no. 4. — P. 195–203.
- 7 Basic report BeerDecoded [Электронный ресурс]. — URL: [https://biota.knomics.ru/public-report?key=elm4igfPEjPgkYzBKzSPxRCKC9E\\_6XQS](https://biota.knomics.ru/public-report?key=elm4igfPEjPgkYzBKzSPxRCKC9E_6XQS); дата обращения: 18.03.2019.
- 8 Basic report for Bokulich article [Электронный ресурс]. — URL: [https://biota.knomics.ru/public-report?key=YxlrFI0rytg-zrij98\\_8YP1mMjPsduuT](https://biota.knomics.ru/public-report?key=YxlrFI0rytg-zrij98_8YP1mMjPsduuT); дата обращения: 12.06.2019.
- 9 BeerDeCoded: the open beer metagenome project / J. Sobel [et al.] // F1000Research. — 2017. — Vol. 6. — P. 1676.
- 10 Chao A. Nonparametric Estimation of the Number of Classes in a Population // Scandinavian Journal of Statistics. — 1984. — Vol. 11, no. 4. — P. 265–270. — ISSN 03036898, 14679469.

- 11 *Deák T.* Current taxonomy of common foodborne bacteria: part I. Gram-negative phyla of proteobacteria and bacteroidetes // *Acta alimentaria*. — 2010. — Vol. 39, no. 4. — P. 471–487.
- 12 Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns / A. Amir [et al.] // *mSystems* / ed. by J. A. Gilbert. — 2017. — Vol. 2, no. 2.
- 13 *Dept E.* Beer 101: The Fundamental Steps of Brewing [Электронный ресурс]. — 2016. — URL: <https://beerconnoisseur.com/articles/beer-101-fundamental-steps-brewing>; дата обращения: 08.06.2019.
- 14 Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts / B. Gallone [et al.] // *Cell*. — 2016. — Vol. 166, no. 6. — 1397–1410.e16.
- 15 *Domon B., Aebersold R.* Mass spectrometry and protein analysis // *science*. — 2006. — Vol. 312, no. 5771. — P. 212–217.
- 16 Effectiveness and safety of *Saccharomyces boulardii* for acute infectious diarrhea / E. C. Dinleyici [et al.] // *Expert Opinion on Biological Therapy*. — 2012. — Vol. 12, no. 4. — P. 395–410.
- 17 *Faith D. P.* Conservation evaluation and phylogenetic diversity // *Biological Conservation*. — 1992. — Vol. 61, no. 1. — P. 1–10.
- 18 Food microbial consortium analysis: from dairy to beer / A. Tyakht [et al.] // — *Emerging applications of microbes*. Leuven, Belgium, 2019.
- 19 Health benefits of fermented foods: microbiota and beyond / M. L. Marco [et al.] // *Current Opinion in Biotechnology*. — 2017. — Vol. 44. — P. 94–102.
- 20 Knomics-Biota - a system for exploratory analysis of human gut microbiota data / D. Efimova [et al.] // *BioData Mining*. — 2018. — Vol. 11, no. 1.
- 21 *Li H.* Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly // *Bioinformatics*. — 2012. — Vol. 28, no. 14. — P. 1838–1844.
- 22 *Li H., Durbin R.* Fast and accurate short read alignment with Burrows-Wheeler transform // *Bioinformatics*. — 2009. — Vol. 25, no. 14. — P. 1754–1760.
- 23 Life with 6000 Genes / A. Goffeau [et al.] // *Science*. — 1996. — Vol. 274, no. 5287. — P. 546–567.

- 24 *Lozupone C., Knight R.* UniFrac: a New Phylogenetic Method for Comparing Microbial Communities // *Applied and Environmental Microbiology*. — 2005. — Vol. 71, no. 12. — P. 8228–8235.
- 25 Mapping microbial ecosystems and spoilage-gene flow in breweries highlights patterns of contamination and resistance / N. A. Bokulich [et al.] // *eLife*. — 2015. — Vol. 4.
- 26 *Modi S. R., Collins J. J., Relman D. A.* Antibiotics and the gut microbiota // *Journal of Clinical Investigation*. — 2014. — Vol. 124, no. 10. — P. 4212–4218.
- 27 MolGenBrew public reports [Электронный ресурс]. — URL: <https://biota.knomics.ru/molgenbrew>; дата обращения: 01.06.2019.
- 28 Nephele/PACTs / D. Battré [et al.] // *Proceedings of the 1st ACM symposium on Cloud computing - SoCC '10*. — ACM Press, 2010.
- 29 Predominant lactic acid bacteria involved in the spontaneous fermentation step of tchapalo process, a traditional sorghum beer of Cote d'Ivoire. / D. Marcellin [et al.] // *Research Journal of Biological Sciences*. — 2009. — Vol. 4, no. 7. — P. 789–795.
- 30 QIIME allows analysis of high-throughput community sequencing data / J. G. Caporaso [et al.] // *Nature methods*. — 2010. — Vol. 7, no. 5. — P. 335.
- 31 ratebeer [Электронный ресурс]. — URL: [www.ratebeer.com](http://www.ratebeer.com); дата обращения: 20.02.2019.
- 32 Resurrecting ancestral alcohol dehydrogenases from yeast / J. M. Thomson [et al.] // *Nature Genetics*. — 2005. — Vol. 37, no. 6. — P. 630–635.
- 33 *Sanger F., Nicklen S., Coulson A. R.* DNA sequencing with chain-terminating inhibitors // *Proceedings of the National Academy of Sciences*. — 1977. — Vol. 74, no. 12. — P. 5463–5467.
- 34 Slider—maximum use of probability information for alignment of short sequence reads and SNP detection / N. Malhis [et al.] // *Bioinformatics*. — 2008. — Vol. 25, no. 1. — P. 6–13.

- 35 Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica / R. Chazdon [et al.] // Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies. — 1998. — P. 285–309.
- 36 The identification of Enterobacteriaceae from breweries: combined use and comparison of API 20E system, gel electrophoresis of proteins and gas chromatography of volatile metabolites / H. Van Vuuren [et al.] // Journal of Applied Bacteriology. — 1981. — Vol. 51, no. 1. — P. 51–65.
- 37 The microbiome in early life: implications for health outcomes / S. Tamburini [et al.] // Nature Medicine. — 2016. — Vol. 22, no. 7. — P. 713–722.
- 38 The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications / R. H. Nilsson [et al.] // Nucleic Acids Research. — 2018. — Vol. 47, no. D1. — P. D259–D264.
- 39 The wine and beer yeast *Dekkera bruxellensis* / A. J. Schifferdecker [et al.] // Yeast. — 2014. — Vol. 31, no. 9. — P. 323–332.
- 40 The yeast *Saccharomyces cerevisiae* — the main character in beer brewing / E. J. Lodolo [et al.] // FEMS Yeast Research. — 2008. — Vol. 8, no. 7. — P. 1018–1036.
- 41 UNTAPPD [Электронный ресурс]. — URL: <https://untappd.com>; дата обращения: 15.01.2019.
- 42 Using the Metagenomics RAST Server (MG-RAST) for Analyzing Shotgun Metagenomes / E. M. Glass [et al.] // Cold Spring Harbor Protocols. — 2010. — Vol. 2010, no. 1. — pdb.prot5368–pdb.prot5368.
- 43 Vidal R., Ma Y., Sastry S. Generalized principal component analysis (GPCA) // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2005. — Vol. 27, no. 12. — P. 1945–1959.
- 44 *Yarrowia lipolytica* and pollutants: Interactions and applications / S. Zinjarde [et al.] // Biotechnology Advances. — 2014. — Vol. 32, no. 5. — P. 920–933.