Анализ данных 16S рРНК секвенирования

Одинцова Вера,

Кномикс

Обо мне

- 7 лет опыта работы с Анализом данных (НИИ ФХМ, Атлас, Кномикс)
- Область интересов
  - кишечные, кожная, ротовая, почвенная микробиота, микробиота насекомых (около 50 проектов)
  - анализ данных 16S и shotgun секвенирования, в основном стат. анализ 16S данных
  - >10 публикаций, как с результатами различных исследований, так и по методам стат. анализа

https://scholar.google.com/citations?user=jrc2iSoAAAAJ&hl=ru&oi=sra

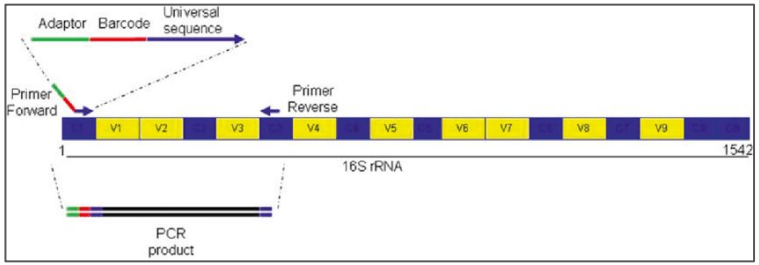
О чем расскажу

- Биоинформатический анализ
  - Как по данным с секвенатора оценить пропорции микробов в образце
  - какие еще есть характеристики образцов

- Статистический анализ
  - Проверка гипотез по выборке: есть ли связь между микробиотой и различными факторами
  - Особенности данных

Процесс получения состава образцов

1. **Пробоподготовка и секвенирование**
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
   a. Подготовка базы
   b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
   a. Альфа-разнообразие
   b. Бета-разнообразие
   c. Метаболический потенциал



Fastq

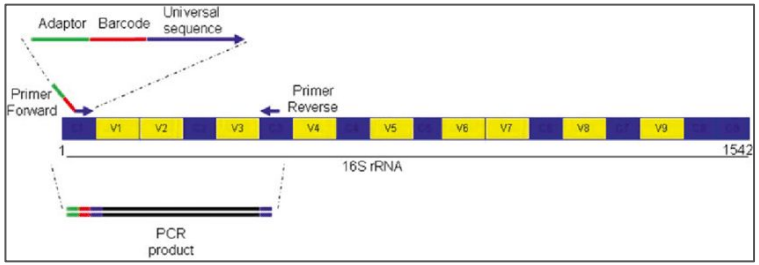
@read1
CCTGTTTGCTCCCCTAGCTTTCGCGCCTCAGCG
TCAGTTGTGGTCCAGTGCGCCGCTTTCGCCAC
AGGTGTTCCTCCCGATATCTACGCATTTCACC
GCTACACCGGGAATTCCACGCACCTCTCCCAC
ACTCAAGCACGGCAGTTTCGCCTGCAGGCTCT
GAGTTGAGCCCAGAGTTTTCACAGGTGACTT
ACCGCACCGCCTACGCGCCCTTTACGCCCAAT
AATTCCGAACAACGCTTGCCCCCCTCGTA
+
HGHHHHEEDFHHFGFEEHHHFHG?0EAEFFHH
GGGG]]]]]]]]]]S]]\IQ]TSQS]T]SUU]]]]]]T]]]]]]]
]]]]]]]T]]RU]]\]]]]]]]]]]]]]]]U]V]R]]]T]]]]T]]]]]]]]
]O]]]]]]S]]]O]]]]]]]]]]]]]S]]]]Z]]]]]]R]]]]]<]]]]]]]T
[]T]]]]]]]]]]S]]]]]]]]]]S]][<\]]U]]]]]]]F/?CCG/F
/FGBAAACCGGCGCGGHHFH
@read2
CCTGTTTGCTCCCCACGCTTTCGCGCCTCAGCG
TCAGTTACAGACCAAAAAGCCGCCTTCGCCAC
TGGTGTTCCTCCACATCTCTACGCATTTCACCG
…

Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. **Демультиплексирование и тримминг**
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
    a. Подготовка базы
    b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
    a. Альфа-разнообразие
    b. Бета-разнообразие
    c. Метаболический потенциал

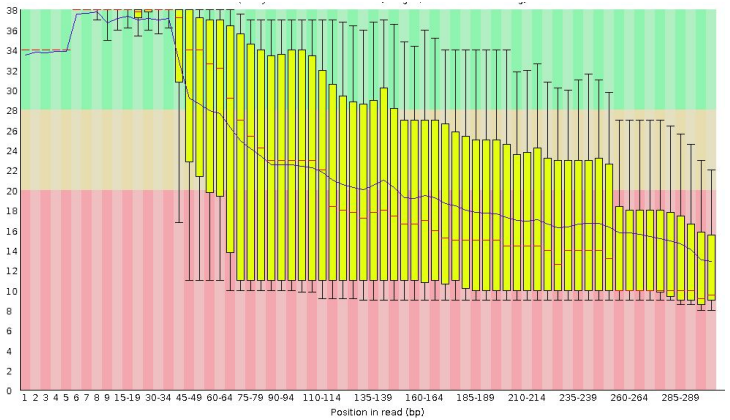
Демультиплексирование -
разделяем риды на файлы по образцам
https://cutadapt.readthedocs.io/en/stable/

Тримминг -
убираем служебные последовательности
(адаптеры, баркоды, праймеры)
https://cutadapt.readthedocs.io/en/stable/
https://docs.qiime2.org/



Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. **Анализ качества**
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
    a. Подготовка базы
    b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
    a. Альфа-разнообразие
    b. Бета-разнообразие
    c. Метаболический потенциал



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	2041709	3.355499770866574	No Hit
CTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCT	240040	0.39449998261202374	No Hit
ACACTCTTTCCCTACACGACGCTCTTCCGNNNNNN	130911	0.21514908858408033	Illumina Single End PCR Primer 1 (96% over 29bp)
TCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCT	106672	0.17531287345938093	No Hit
CTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTT	97553	0.16032601567968152	No Hit
TCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTCTCTTC	85228	0.1400701737962738	No Hit

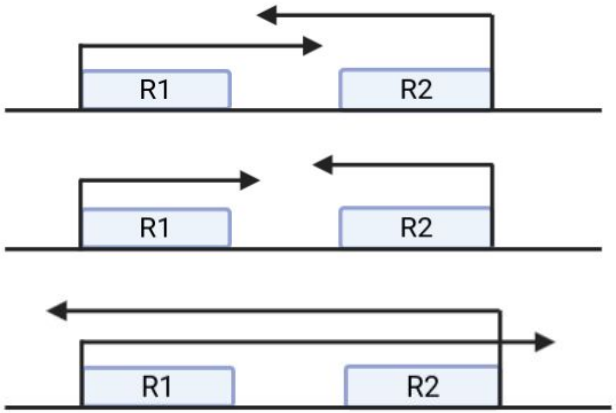
FastQC, MultiQC

- обрезаем концы плохого качества или риды с плохим качеством
- выкидываем риды с низким покрытием
- проверяем наличие адаптеров и баркодов
- смотрим гистограмму длин
- проверяем количество ридов на образец

https://support.illumina.com/bulletins/2016/12/what-sequences-do-i-use-for-adapter-trimming.html

Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. **Объединение парных ридов**
5. Исправление ошибок секвенирования
6. Картирование
    a. Подготовка базы
    b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
    a. Альфа-разнообразие
    b. Бета-разнообразие
    c. Метаболический потенциал



- следим за количеством оставшихся ридов
- иногда лучше оставить только R1 или R2

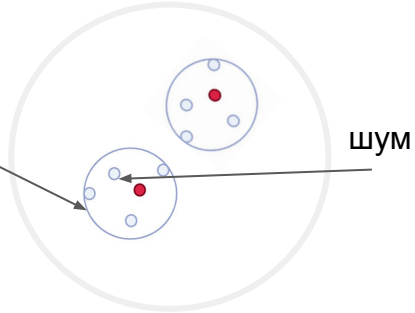
> vsearch --fastq_mergepairs (OR qiime2 fastq-join)
> pandaseq
> pear
> SeqPrep
> leeHom (useful for short amplicons)
> Dada2 - алгоритм предполагает слияние после denoising

Процесс получения состава образцов

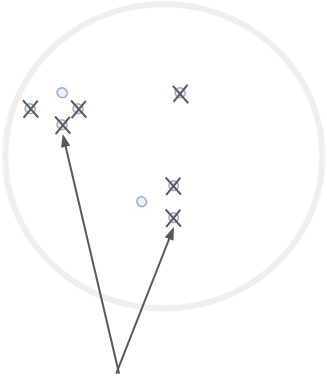
1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. **Исправление ошибок секвенирования**
6. Картирование
   a. Подготовка базы
   b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
   a. Альфа-разнообразие
   b. Бета-разнообразие
   c. Метаболический потенциал

Кластеризация de novo

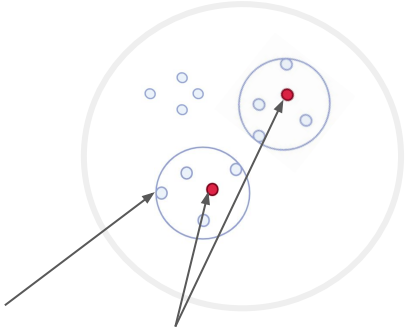
OTE (операционная таксономическая единица)



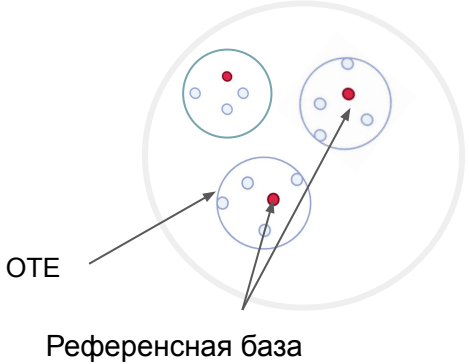
Фильтрация (Denoising): DADA2, Deblur



Картирование на базу (closed-reference)



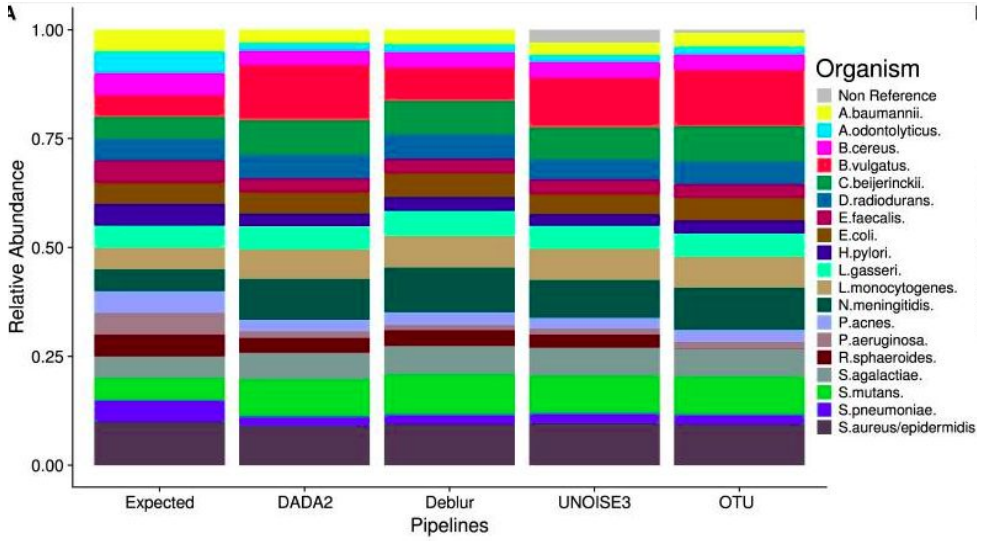
Комбинация кластеризации и картирования



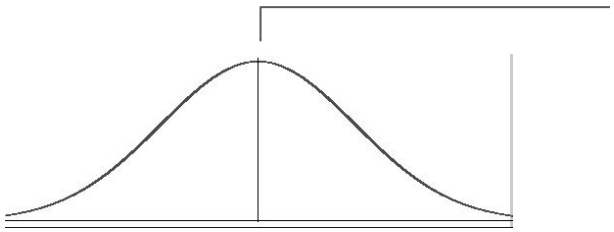
Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. **Исправление ошибок секвенирования**
6. Картирование
   a. Подготовка базы
   b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
   a. Альфа-разнообразие
   b. Бета-разнообразие
   c. Метаболический потенциал

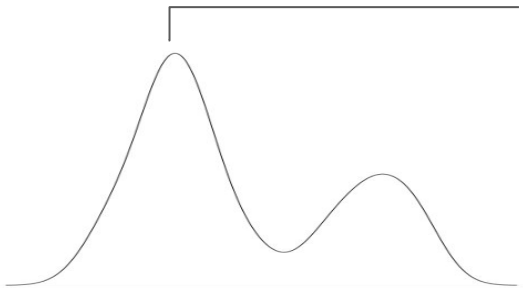
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6087418/



Как выбрать длину для алгоритмов денойзинга:



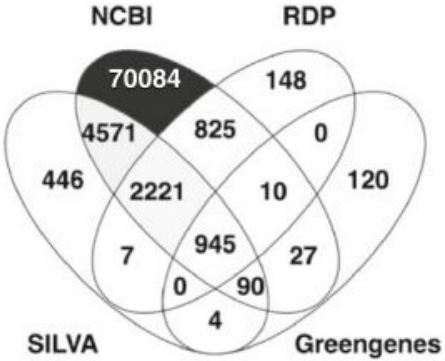
Наиболее частая длина L.
Риды разбиваются на группы длины L-5, L-4, L-3, L-2, L-1, >=L.
Алгоритмы запускаются независимо.



Развиваем на группки вокруг вершин.
Описанные выше шаги выполняются независимо для каждой из вершин.

Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. **Картирование**
   a. **Подготовка базы**
   b. **Картирование**
7. Вычисление пропорций
8. Вычисление других характеристик
   a. Альфа-разнообразие
   b. Бета-разнообразие
   c. Метаболический потенциал



https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3501-4

Подготовка базы:

- выбор базы
- обрезание по нужному региону
- подходит для целей исследования?
  - Сможем различить нужные бактерии по выбранному региону?
  - Насколько база обновленная?
  - Насколько хорошо курируется?

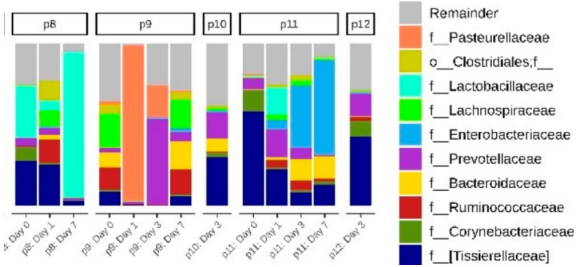
https://www.ibi.vu.nl/programs/taxmanwww/
http://bioinformatics.org/cd-hit/

Картирование:

- Illumina
  - Выравнивание (Usearch, Vsearch, blastn)
  - Классификация (RDP Naive Bayes, Qiime2 Naive Bayes)
- Oxford Nanopore
  - Emu
  - NanoCLUST
- может занимать долгое время

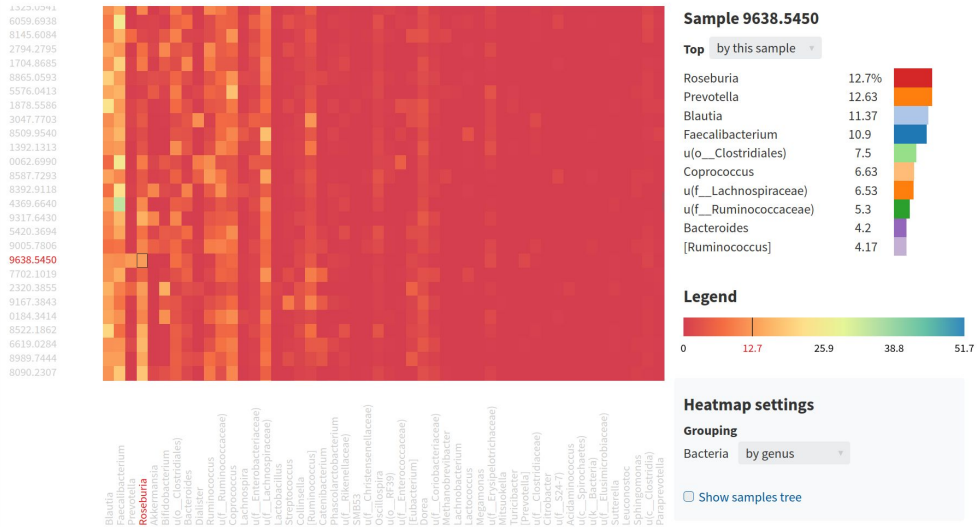
Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
    a. Подготовка базы
    b. Картирование
7. **Вычисление пропорций**
8. Вычисление других характеристик
    a. Альфа-разнообразие
    b. Бета-разнообразие
    c. Метаболический потенциал



https://doi.org/10.3390/jpm11111113

biota.knomx.com

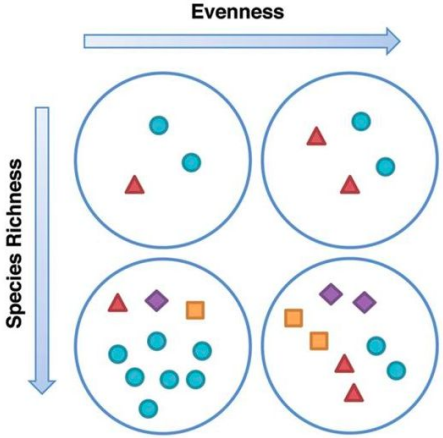


QC:

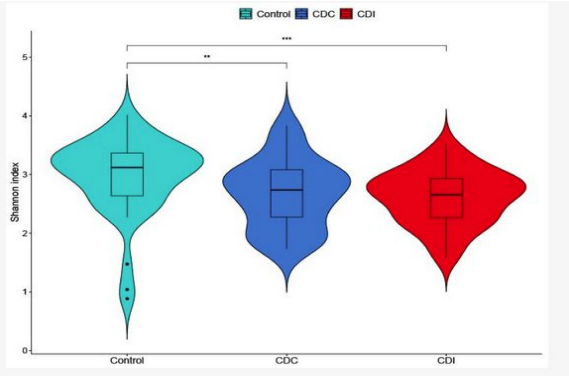
- есть ли “странные микробы”?
- есть ли “странные” образцы?
- состав положительного и отрицательного контроля?
- достаточное ли покрытие
- процент картировавшихся ридов

Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
   a. Подготовка базы
   b. Картирование
7. Вычисление пропорций
8. **Вычисление других характеристик**
   a. **Альфа-разнообразие**
   b. Бета-разнообразие
   c. Метаболический потенциал



Cox et al, Human Molecular Genetics, 2013



Crobach et al., 2020

Визуализация:

- боксплот
- violin plot

Альфа-разнообразие это:

- Характеристика 1 образца
- Насколько много в образце разных микробов (богатство)
- Насколько равномерно они представлены (ровность)

Примеры:

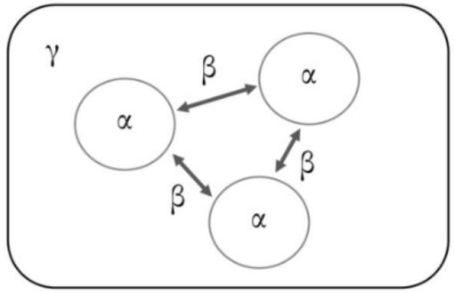
- Индекс Шеннона (равномерность)
- Индекс Chao1 (богатство)
- Индекс Симпсона (равномерность)

Что учесть:

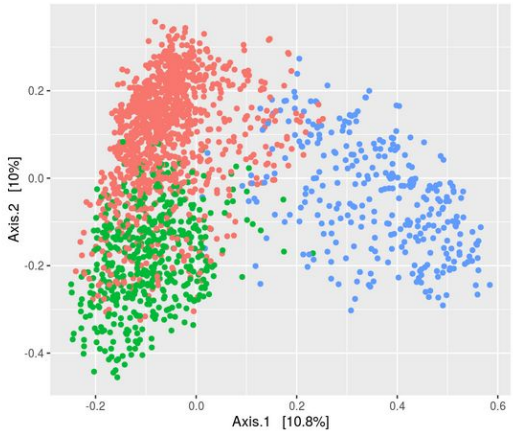
- Можно использовать несколько индексов
- Чувствительны к покрытию образца
- Стандартные методы статистики (обычно непараметрические)

Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
   a. Подготовка базы
   b. Картирование
7. Вычисление пропорций
8. **Вычисление других характеристик**
   a. Альфа-разнообразие
   b. **Бета-разнообразие**
   c. Метаболический потенциал
   d.



https://doi.org/10.3390/math6070119



Визуализация:

- PCoA

Бета-разнообразие это:

- Характеристика 2 образцов
- Насколько два микробных сообщества отличаются

Примеры:

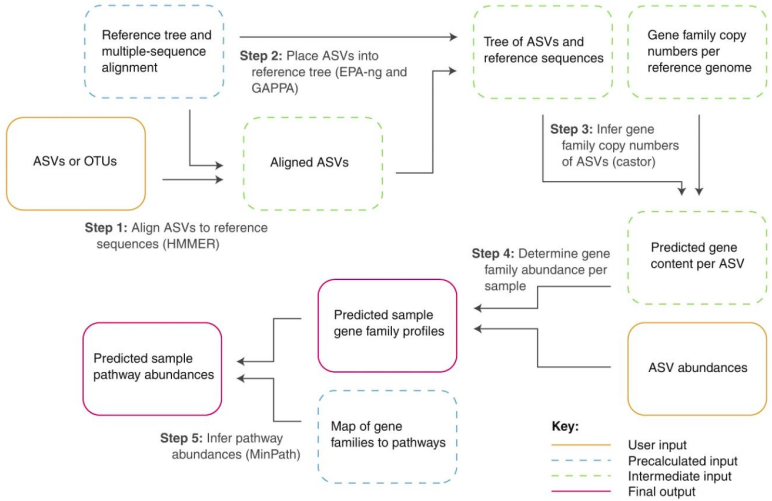
- UniFrac
- Bray-Curtis
- Aitchison

Что учесть:

- Можно использовать несколько индексов
- Чувствительны к покрытию образца
- PERMANOVA для стат. анализа

Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
   a. Подготовка базы
   b. Картирование
7. Вычисление пропорций
8. **Вычисление других характеристик**
   a. Альфа-разнообразие
   b. Бета-разнообразие
   c. **Метаболический потенциал**



PICRUSt2: требователен к объему памяти и времени при большом числе OTU

FAPROTAX: для сравнения >80 высокоуровневых функций разнообразных микробиомов (“ферментация”, “метаногенез”...).

Tax4fun: более детальный, хорошо подходит для микробиомов океана.

https://www.nature.com/articles/s41587-020-0548-6

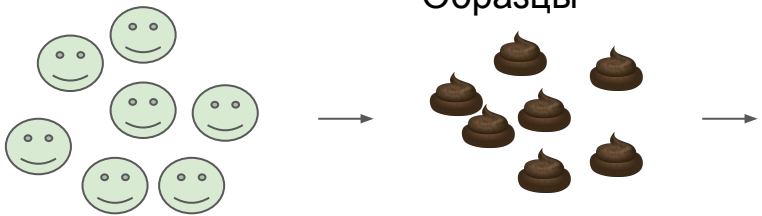
Процесс получения состава образцов

1. Пробоподготовка и секвенирование
2. Демультиплексирование и тримминг
3. Анализ качества
4. Объединение парных ридов
5. Исправление ошибок секвенирования
6. Картирование
    a. Подготовка базы
    b. Картирование
7. Вычисление пропорций
8. Вычисление других характеристик
    a. Альфа-разнообразие
    b. Бета-разнообразие
    c. Метаболический потенциал

Как все это сделать:

1. qiime2.org
2. mothur.org
3. biota.knomx.com
4. …

Статистический анализ



- Пропорции микробов
- альфа-разнообразие
- бета-разнообразие
- метаболический потенциал

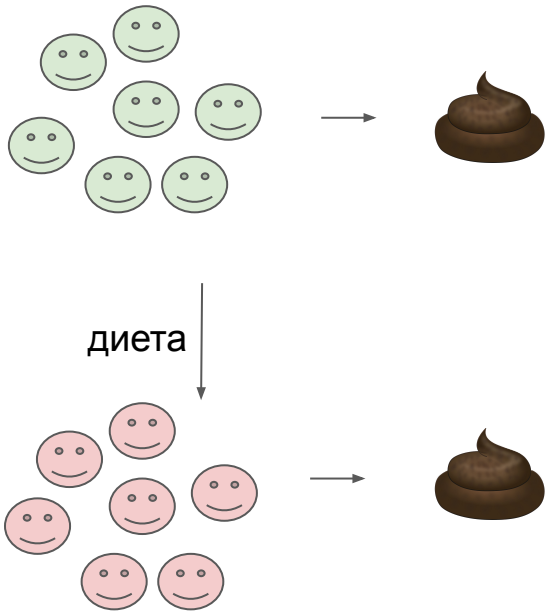
Метаданные

	Age	Sex	Visit	...
s1				
s2				
s3				
s4				
s5				
...				

? Как связаны

Отступление: композиционность данных

Пример: влияние диеты на микробиоту человека



		bact1	bact2	…	bactM
до	subj1	50	10	…	3
	subj2	0	1	…	0
	…	…	…	…	…
	…	120	260	…	127
после	subj1	70	0	…	3
	subj2	0	0	…	27
	…	…	…	…	…
	subjN	9	14	…	17

Особенности данных:

композиционные данные

Особенность	**Что это значит**
сумма по строчкам - случайная величина	оцениваем только **относительные** представленности бактерий (рис. 1)
значения - целые числа	точность оценки зависит от покрытия образца (общего количества ридов в нем)
много нулей	ноль - не всегда действительно ноль (рис. 2)

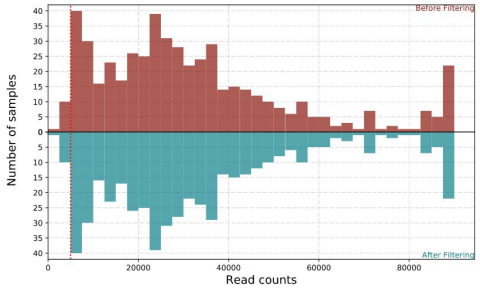


Рис. 1. Покрытие образов до и после фильтрации по качеству

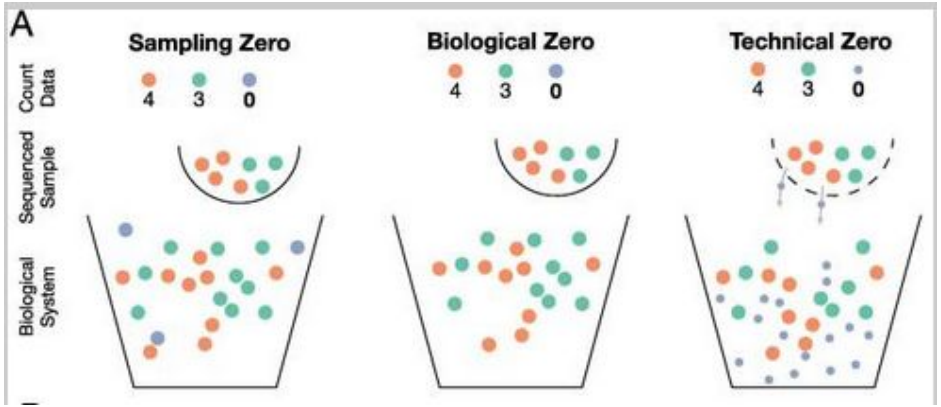
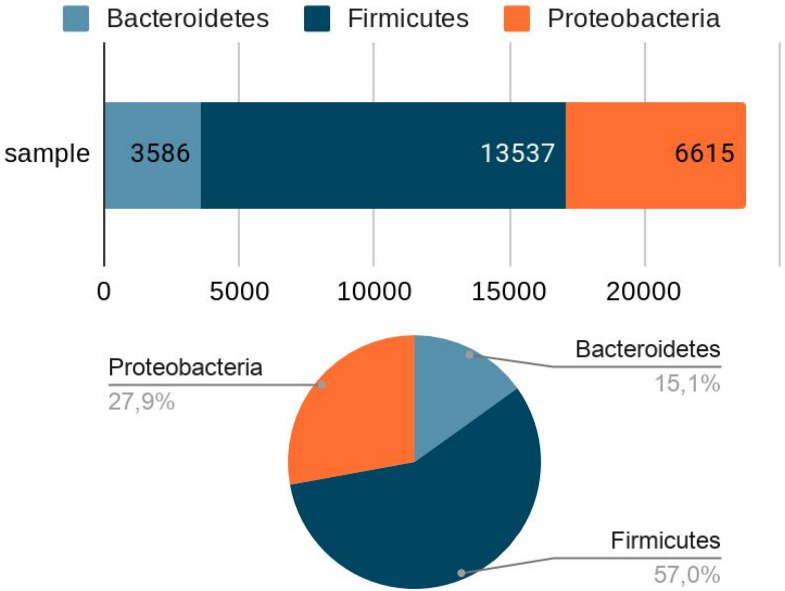
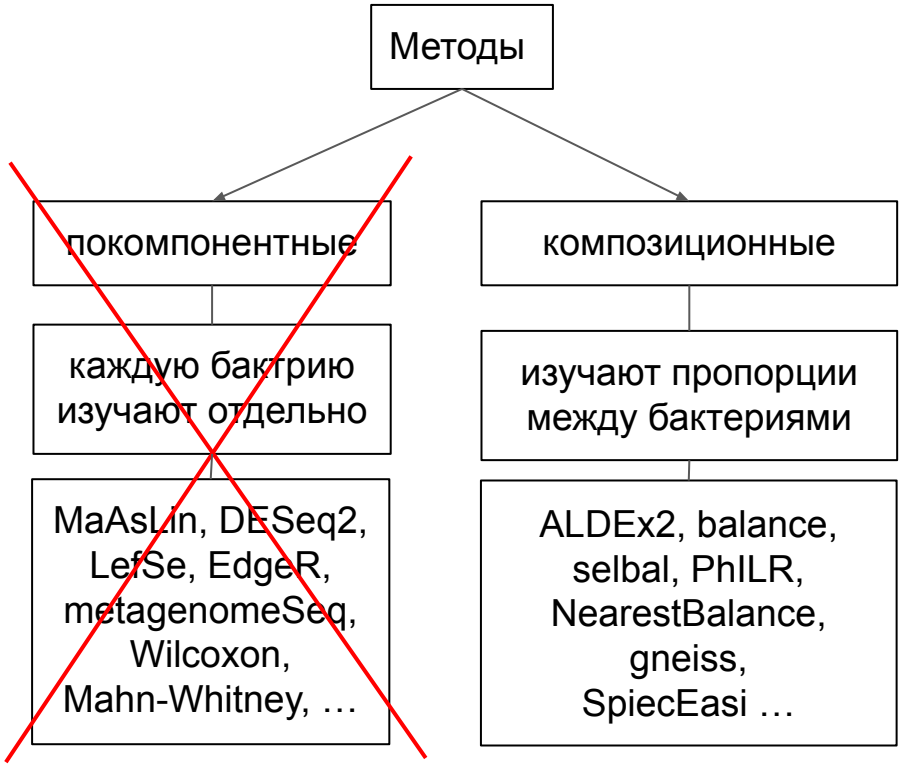


Рис. 2 - Виды нулей (https://doi.org/10.1016/j.csbj.2020.09.014)

Композиционность данных - важное свойство



невозможно поменять долю одной бактерии не поменяв долю остальных



Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1], Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]*

Пример:
заранее знаем что поменялось

	b1	b2	b3
количество ($\cdot 10^7$)			
образец 1	1	4	5
образец 2	1	4	15
пропорции			
различие			

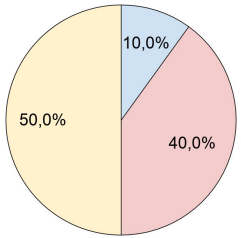
Пример:
заранее знаем что поменялось

	b1	b2	b3
количество $(\cdot 10^7)$			
образец 1	1	4	5
образец 2	1	4	15
пропорции			
образец 1	10%	40%	50%
образец 2	5%	20%	75%
различие			

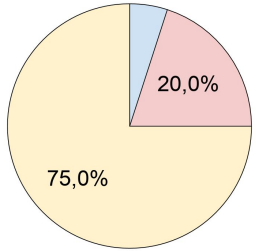
Пример:
заранее знаем что поменялось

	b1	b2	b3
количество ($\cdot 10^7$)			
образец 1	1	4	5
образец 2	1	4	15
пропорции			
образец 1	10%	40%	50%
образец 2	5%	20%	75%
различие			
покомпонентное	-5%	-20%	+25%

образец 1



образец 2



Критика покомпонентного сравнения

1. Видим изменения которых нет
2. Размер эффекта этих изменений разный
3. Теряется информация о постоянном соотношении бактерий b1 и b2

Пример:
заранее знаем что поменялось

	b1	b2	b3
количество ($\cdot 10^7$)			
образец 1	1	4	5
образец 2	1	4	15
пропорции			
образец 1	10%	40%	50%
образец 2	5%	20%	75%
различие			
покомпонентное	-5%	-20%	+25%
композиционное	1/2 (20%)	1/2 (20%)	3/2 (60%)

Композиционный анализ:

1. Наши выводы не должны меняться при переходе от абсолютных значений к долям
2. Различие доли бактерии в образцах - “во сколько раз” (не “на сколько”)
3. Расстояние Эйтчисона между образами - как сильно отличаются пропорции между компонентами (не сами компоненты)

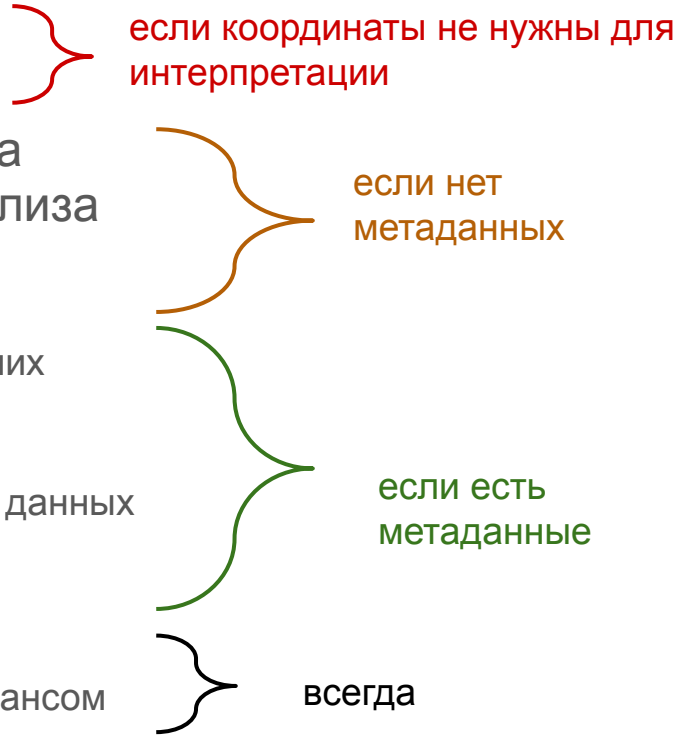
Разные идеи

	b1	b2	b3
количество ($\cdot 10^7$)			
образец 1	1	4	5
образец 2	1	4	15
пропорции			
образец 1	10%	40%	50%
образец 2	5%	20%	75%
различие			
покомпонентное	-5%	-20%	+25%
композиционное	1/2 (20%)	1/2 (20%)	3/2 (60%)

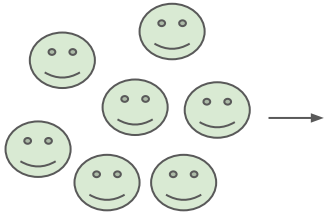
- <u>логарифмирование</u>:
  - \+ превращает сложение в умножение
  - \+ отсутствие изменений - нулевой вектор
  - \+ сохраняется расстояние Эйтчисона
  - \- остается зависимость компонент нормировки
- <u>CLR</u> (centered log-ratio): $\ln(b_i)$-mean($\ln(b_j)$)
  - \+ все преимущества логарифмирования
  - \+ нормировка
  - \- остается зависимость компонент (сумма=0)
- <u>ALR</u> (additive log-ratio): $\ln(b_i/b_0)$
  - \+ превращает сложение в умножение
  - \+ отсутствие изменений - нулевой вектор
  - \+ независимые компоненты
  - \- не сохраняется расстояние Эйтчисона
- <u>ILR</u> (isometric log-ratio) - объединяет все плюсы, но сложная интерпретация

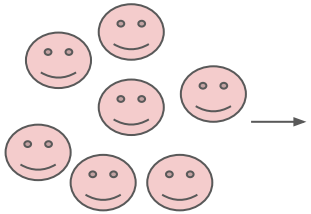
Примеры способов выбора ilr-координат

- произвольное (пакет balance)
- PCA
- PhiLR - на основе генетического сходства
- principal balance analysis - на основе анализа вариации данных
- gneiss (встроен в QIIME2)
  - кластеризация таксонов по характерным для них значениям фактора
- DBA (пакет balance)
  - для сравнения двух групп на основе вариации данных
- selbal
  - поиск только одного самого важного баланса
- NearestBalance
  - приближает любой ILR вектор ближайшим балансом



NearestBalance*

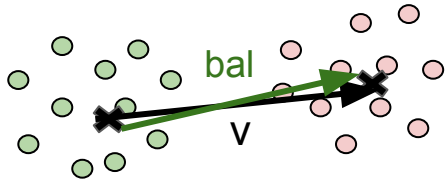




		bact1 %	bact2 %	…	bactD %
до	subj1	5	1	…	0.3
	subj2	15	0.1	…	1
	…	…	…	…	…
	…	12	26	…	0.1
после	subj1	7	8	…	3
	subj2	1	15	…	0.2
	…	…	…	…	…
	subjN	9	14	…	17

для анализа и
интерпретации векторов

ILR-координаты



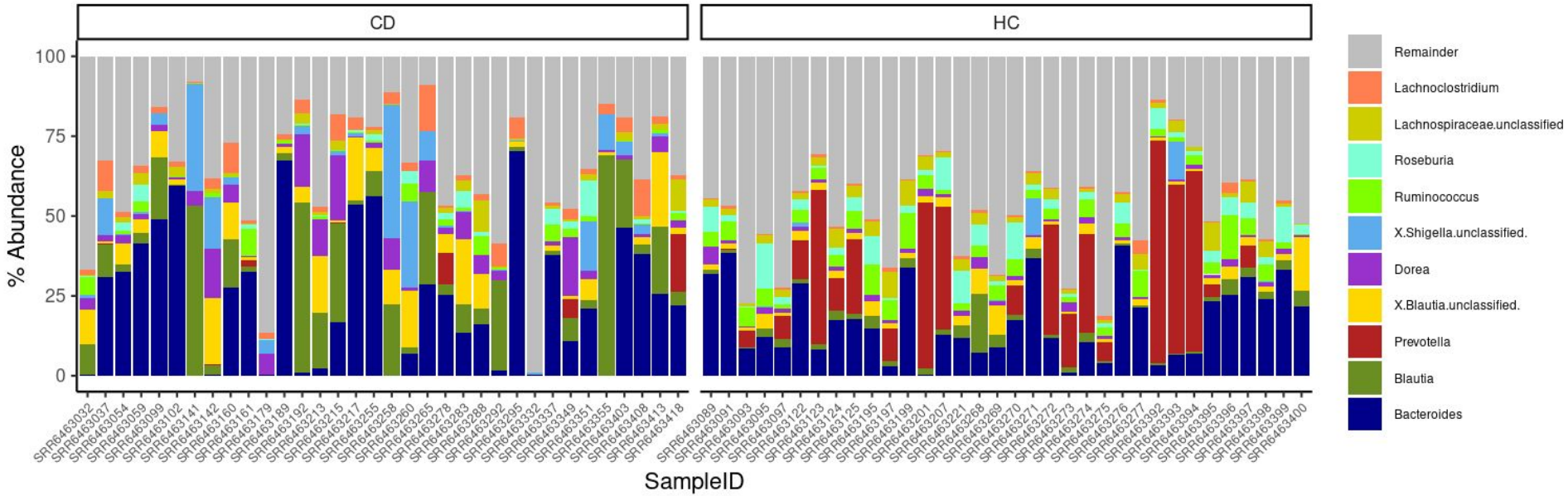
для интерпретации композиций (точек)

bal - ближайший вектор-баланс (изменение микробиоты состоит только в изменении соотношения между двумя группами бактерий)

* Odintsova VE, Klimenko NS, Tyakht AV. Approximation of a Microbiome Composition Shift by a Change in a Single Balance Between Two Groups of Taxa. Msystems. 2022 May 9:e00155-22.

Пример: болезнь Крона (34 здоровых vs 34 больных)

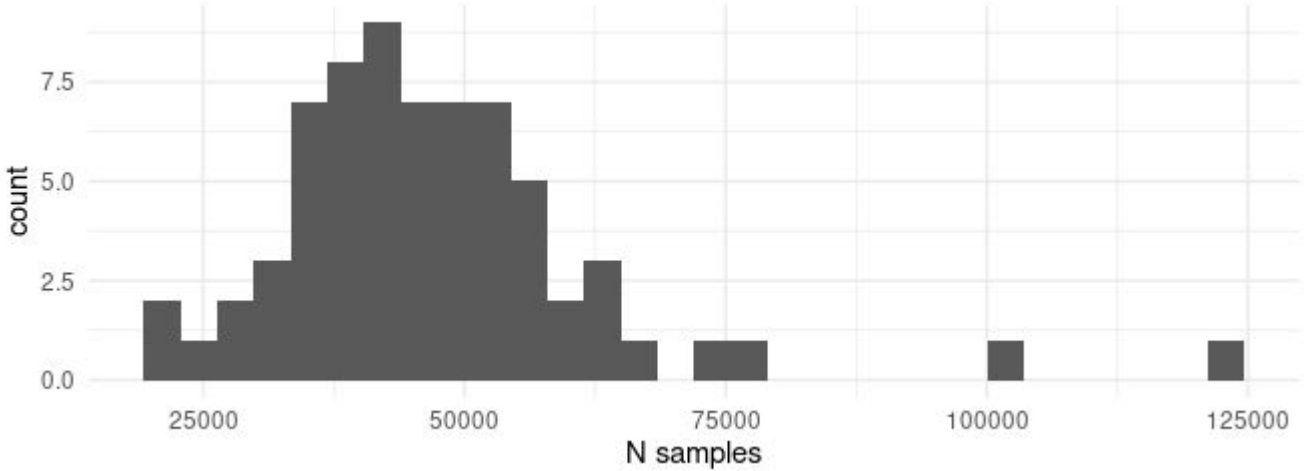
Данные из статьи https://pubmed.ncbi.nlm.nih.gov/28179361/:



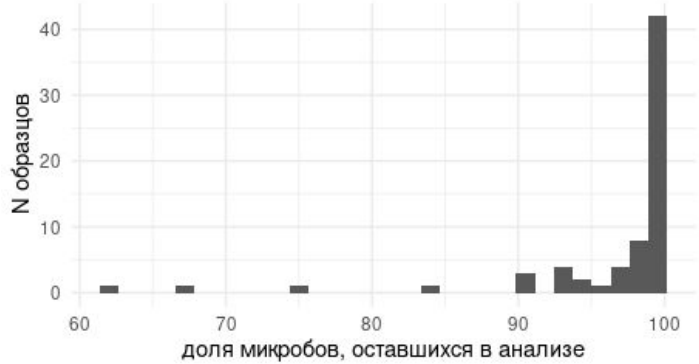
Всего 210 микробов

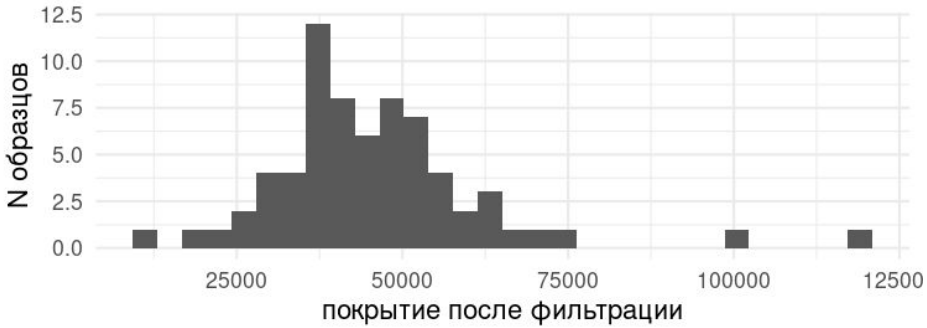
Пример: качество покрытия

Качество секвенирования достаточное



Пример: фильтрация от редких микробов





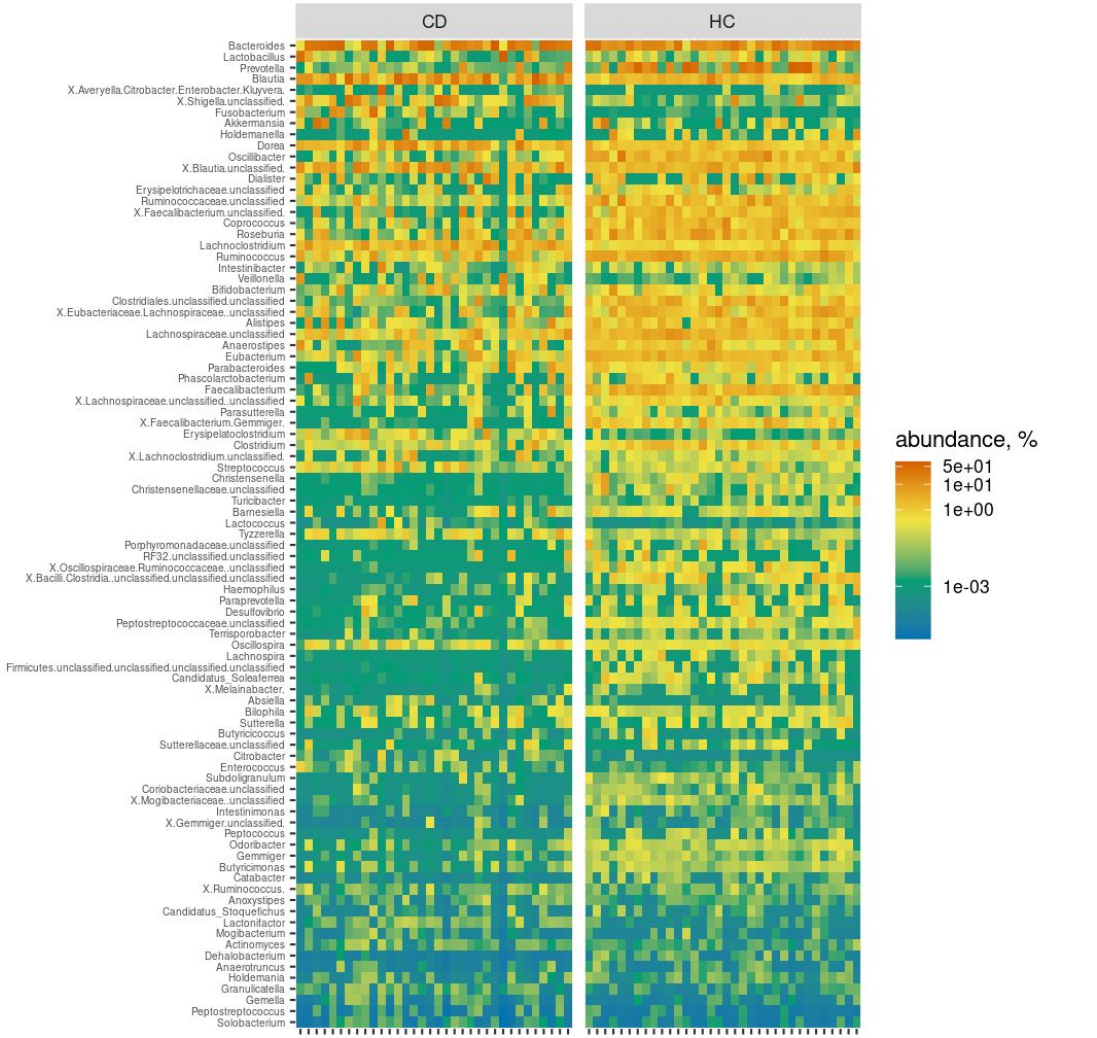
Оставляем микробы, встречающиеся в >30% образцов

	до	после
N микробов	210	89
мин. покрытие	19414	12981

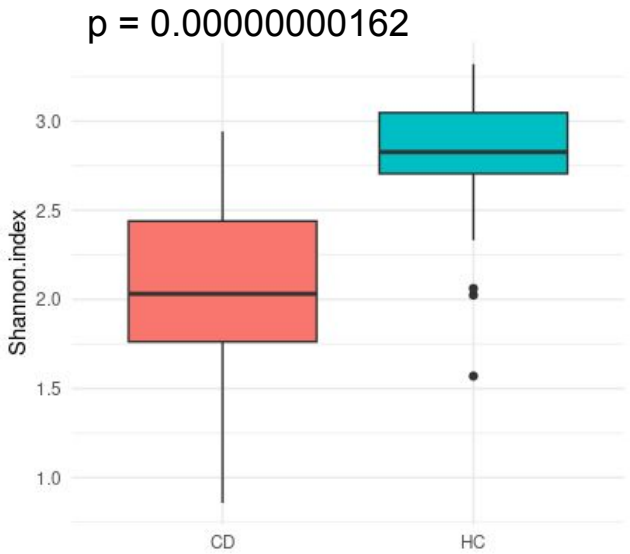
Пример:

Считаем относительные представленности основных микробов

(zCompositions)



Пример: альфа-разнообразие

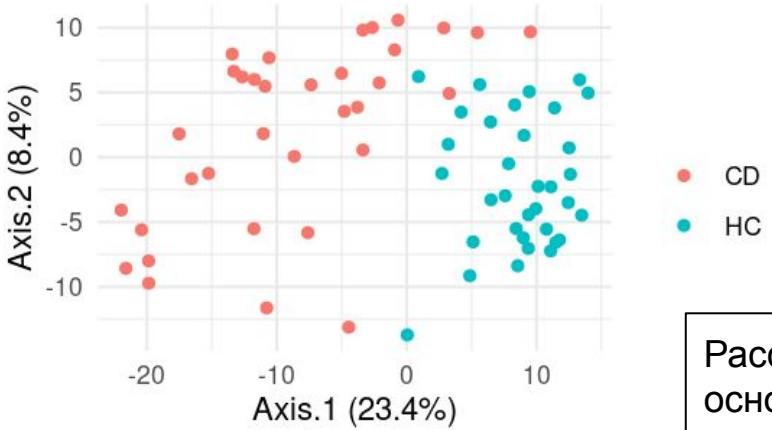


Альфа-разнообразие:

- Возьмем индекс Шеннона
- Посчитаем его после 5-кратного прореживания до 19000 ридов и усреднения
- Построим боксплот
- Проверим стат. значимость тестом Манна-Уитни

Пример: бета-разнообразие

PCoA



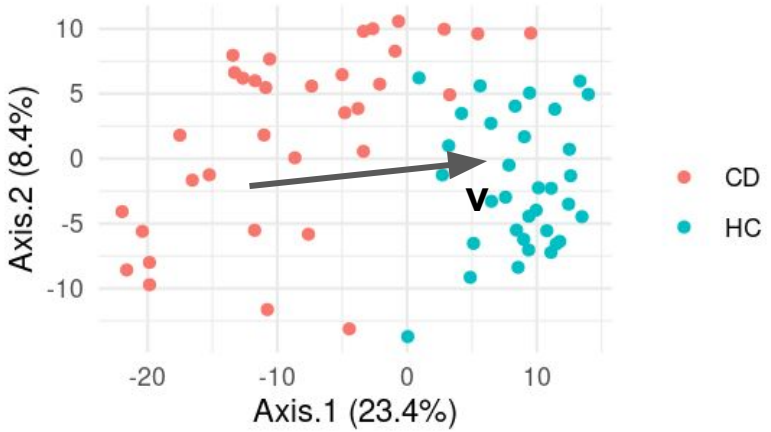
p=0.001

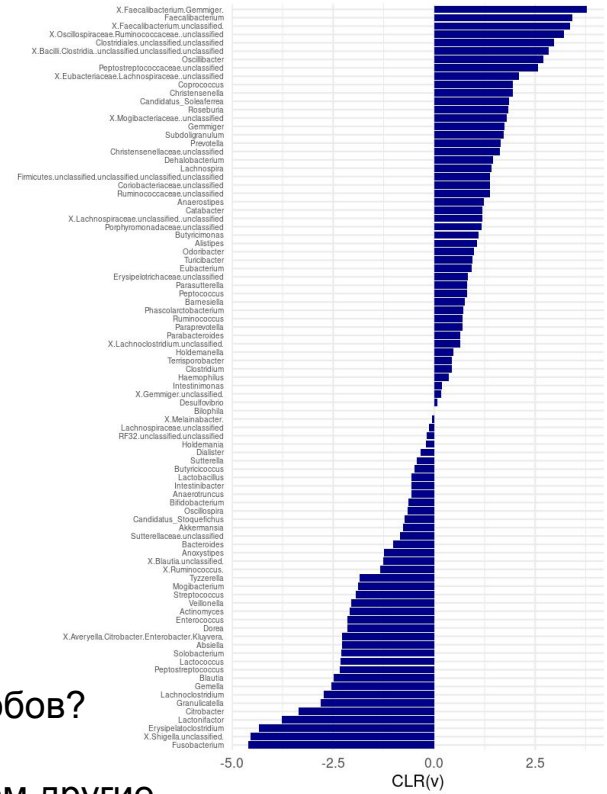
- Используем бета-разнообразие Эйтчисона
- Визуализируем с помощью PCoA
- Проверяем различие с помощью PERMANOVA

Расстояние Эйтчисона - насколько похожи пропорции основных микробов в двух образцах

PERMANOVA - образцы из одной и той же группы больше похожи, чем из разных?

Пример: в чем именно заключаются различия



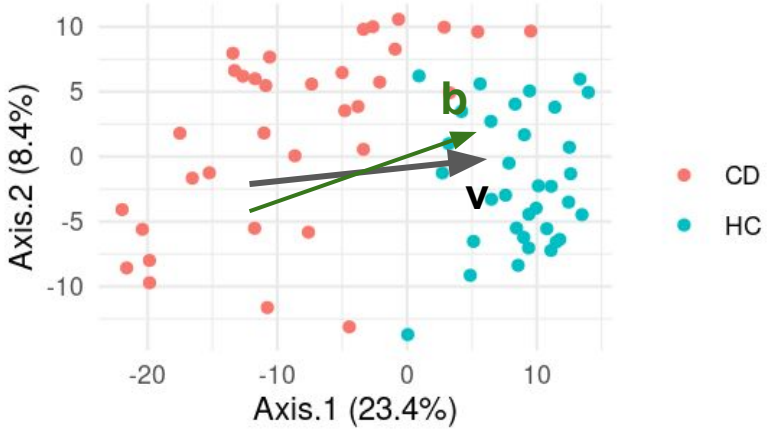


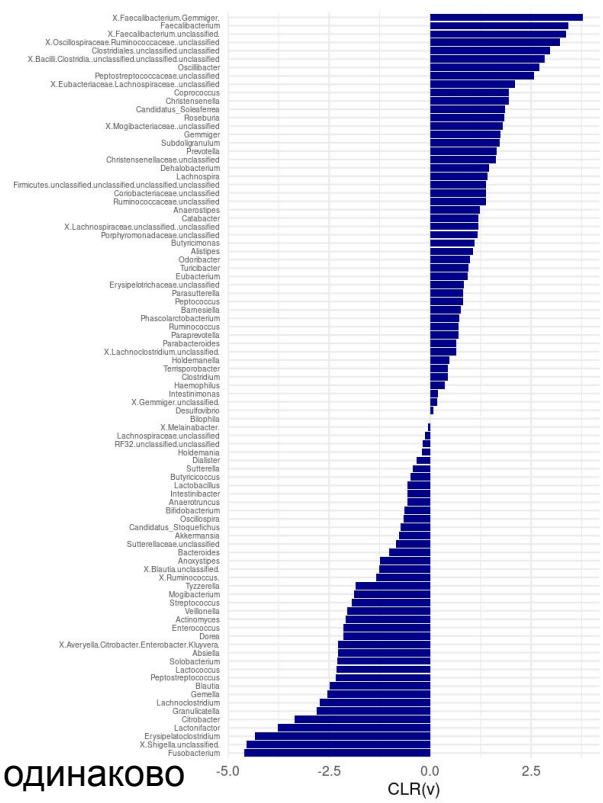
Смотреть на каждый отдельный микроб - ошибочно!
Изменение даже в абсолютном количестве одного микроба означают изменения всех долей

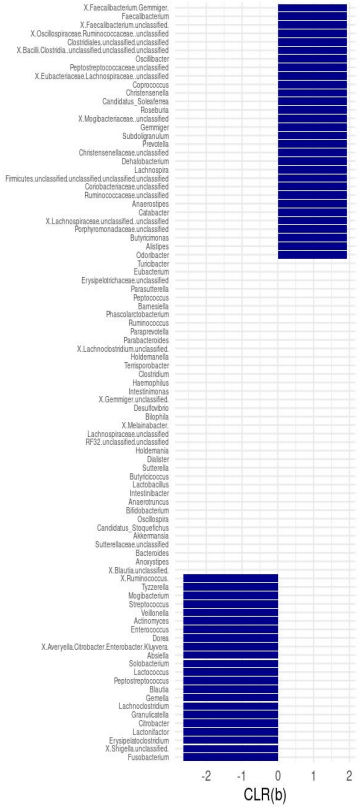
CLR:

- на сколько порядков поменялась доля микробов?
- вычитаем среднее по всем микробам
- смотрим, вырос ли он сильнее, чем в среднем другие микробы?

Пример: в чем именно заключаются различия





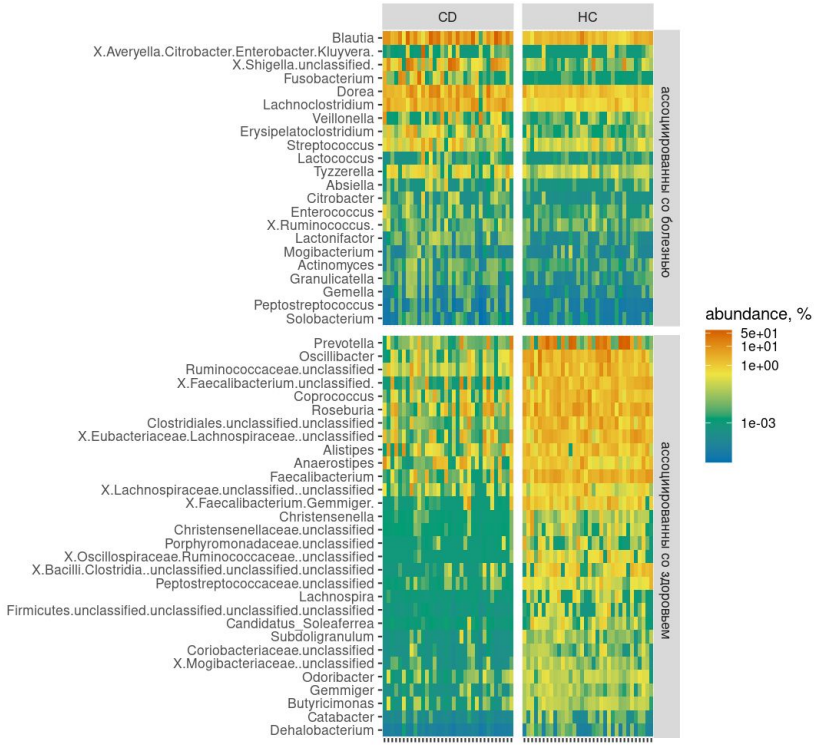


Упрощенное изменение (NearestBalance):

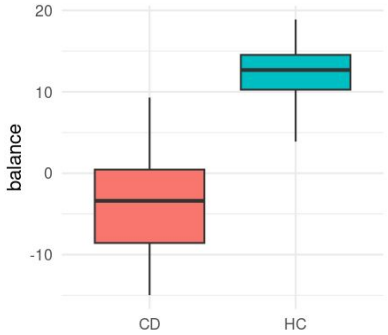
- выделяем три группы микробов
- внутри каждой микробы меняются примерно одинаково
- одна ассоциирована со здоровьем, другая - с заболеванием, третья - не зависит от него

- 83.7% всех различий описываются этим упрощением

Пример: как эти различия выражены в каждом образце



Значение баланса, $p=10^{-17}$ (критерий Манна-Уитни)



Значение баланса характеризует соотношение микробов ассоциированных со здоровьем и болезнью в образце

Что дальше?

- биологическая интерпретация
  - что объединяет микробы, оказавшиеся в одной группе?
  - нет ли антагонизма между микробами разных групп
  - Какая может быть связь с заболеванием (где причина, где следствие)?
- анализ метаболического потенциала (аналогично)
- построение новых гипотез и проверка их другими методами

Спасибо за внимание!

v.odintsova@knomx.com