# ОБРАБОТКА ДАННЫХ ДНК-МЕТАБАРКОДИНГА ГРИБОВ: ОБЗОР МЕТОДИК

Щепин Олег Николаевич

Ботанический институт им. В.Л. Комарова РАН

Санкт-Петербург

# Методы изучения разнообразия грибов

## 1. Сбор плодовых тел в природе

- Морфологическое + молекулярное определение
- Полноте выявления мешают:
  - Охват только макромицетов
  - Нерегулярность образования плодовых тел
  - Неодновременность
  - Короткое время существования плодовых тел у многих видов

# Методы изучения разнообразия грибов

## 2. Культуральные методы (посев на среды)

- Морфологическое + молекулярное определение
- Полноте выявления мешают:
  - Наличие некультивируемых грибов
  - Избирательность питательных сред

# Методы изучения разнообразия грибов

## 3. Анализ ДНК, выделенной из природных субстратов (environmental DNA)

- Только молекулярная идентификация
- Не нужны плодовые тела и культивируемость

# Методы изучения разнообразия грибов

## 3.1 Клонирование в бактериях + секвенирование по Сэнгеру

- Последовательности любой длины
- Проблемы:
  - Низкая пропускная способность – десятки или сотни последовательностей на выходе
  - Проблема специфичности праймеров
  - Химерные последовательности

# Методы изучения разнообразия грибов

## 3.2 ДНК-метабаркодинг (металшктрихкодирование, ампликонное секвенирование)

- Огромная пропускная способность – от десятков тысяч до сотен миллионов последовательностей
- Проблем много, их и будем обсуждать :)

# ДНК-метабаркодинг

Метод оценки биологического разнообразия

Основа:

- **NGS** - технологии секвенирования следующего поколения
- **баркодинг**

# ДНК-метабаркодинг

## Что такое баркодинг?

- Баркод (штрихкод) – фрагмент ДНК, который используют как универсальный маркер для видовой идентификации организмов
- Примеры:
  - Животные – COI
  - Многие протисты – ген 18S рРНК
  - Прокариоты – ген 16S рРНК
  - Грибы - ITS
- Цель – создание референсной базы, охватывающей все организмы на планете

# ДНК-метабаркодинг

## Что такое технологии секвенирования следующего поколения?

- Одновременное секвенирование различных последовательностей в одной пробе
- Технологий много
- Дорогие, но быстро дешевеют
- Широкая сфера применений

# Технологии NGS

## 1. Пиросеквенирование (Roche 454)

- Длина до 500 п.о.
- До 1 000 000 прочтений за запуск
- Проблема – прочтение гомополимерных участков (например, AAAA)

# Технологии NGS

## 2. Полупроводниковое секвенирование (IonTorrent)

- Длина 200-400 п.о.
- До 5 000 000 прочтений за запуск
- Тоже проблема с гомополимерами и числом ошибок
- Самый дешевый прибор и реактивы

# Технологии NGS

## 3. Illumina (секвенирование путем синтеза)

- Длина до 300-350 п.о.
- Парные прочтения –> вплоть до 500-700 п.н.
- До 3 млрд. прочтений за запуск
- Гомополимеры не проблема
- Уровень ошибок почти как у сэнгеровского
- Самый дорогой :(

# ДНК-метабаркодинг: общая схема

1. Выделение ДНК из субстрата
2. ПЦР-амплификация баркода
3. Пришивание адаптеров и индексов
4. Секвенирование ДНК-библиотек методами NGS
5. **Биоинформатическая обработка**
6. Статистический анализ и осмысление

# ДНК-метабаркодинг: обработка данных

## Как выглядят данные в начале?

- Файлы FASTQ: FASTA + quality score
- Сотни тысяч или миллионы «сырых» прочтений:
  - Ошибки секвенатора
  - Инсерции, делеции и замены (ПЦР)
  - Химерные последовательности (ПЦР)

# ДНК-метабаркодинг: обработка данных

## Как выглядят данные в конце?

- Файл FASTA, сотни или тысячи уникальных последовательностей
- Большинство – корректные биологические последовательности
- Таблица OTU:
  - Распределение последовательностей по пробам
  - Таксономическая принадлежность
  - Экологическая аннотация

# ДНК-метабаркодинг: обзор инструментов



- mothur www.mothur.org
- QIIME www.qiime.org
- USEARCH www.drive5.com/usearch
- vsearch www.github.com/torognes
- MG-RAST www.metagenomics.anl.gov – полностью автоматизированный анализ на американском сервере

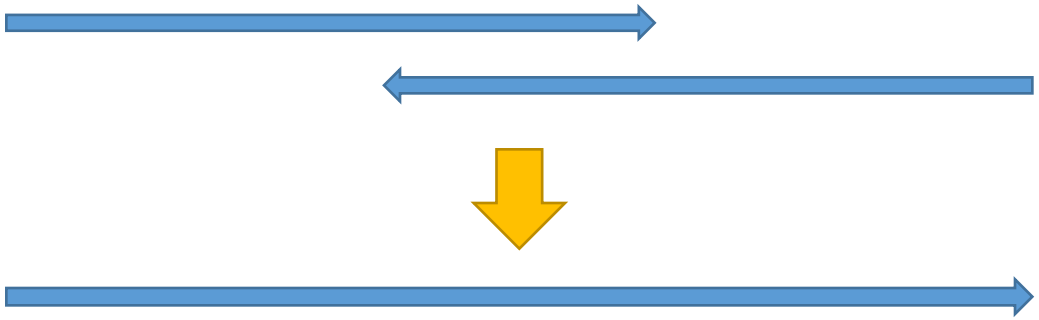
# ДНК-метабаркодинг: обработка данных

## 1. Обрезка адаптеров и демультиплексирование

- Разделение сиквенсов по пробам
- Trimmomatic или встроенное ПО секвенатора
- Сиквенсы без индексов или с ошибками

# ДНК-метабаркодинг: обработка данных

## 2. Объединение парных прочтений



- Увеличивается длина прочтений
- Увеличивается качество в участках перекрывания

# ДНК-метабаркодинг: обработка данных

## 3. Фильтрация по качеству

Цель – избавиться от ошибок секвенирования

- По среднему значению Q-score – плохо
- Обрезка прочтений – сомнительно
- **По ожидаемому числу ошибок** – самая логичная

# ДНК-метабаркодинг: обработка данных

## 4. Разворачивание

Сиквенсы могут быть ориентированы по-разному. Нужно избежать дублирования OTU.

- Все пробы сливают в один файл
- Разворачивание за счет выравнивания на референсную базу
- Часть сиквенсов отпадает

# ДНК-метабаркодинг: обработка данных

## 5. Дерепликация

- Уменьшение объема данных
- Одинаковые последовательности и подстроки более длинных удаляют
- Сохраняется информация об их обилии

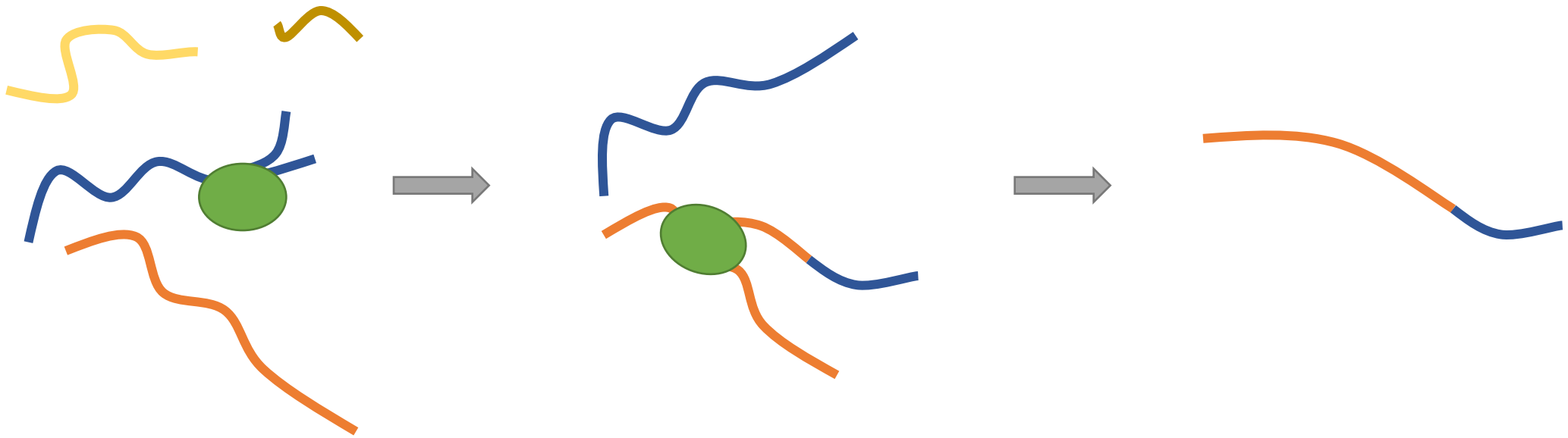
# ДНК-метабаркодинг: обработка данных

## 6. Удаление одиночных последовательностей (глобальных синглетонов)

- Одиночные сиквенсы – ошибки
- Иногда отбрасывают все малочисленные сиквенсы

# ДНК-метабаркодинг: обработка данных

## 7. Поиск химерных последовательностей



# ДНК-метабаркодинг: обработка данных

## 7. Поиск химерных последовательностей

- Поиск *de novo*
- Поиск по референсной базе
- Ложноположительные и ложноотрицательные срабатывания

# ДНК-метабаркодинг: обработка данных

## 8. Кластеризация OTU

- Вместо видов – условные единицы учета
- Оперативные (операционные) таксономические единицы – Operational Taxonomic Units, OTUs
- В идеале – примерно видового уровня

# ДНК-метабаркодинг: обработка данных

## 8. Кластеризация OTU

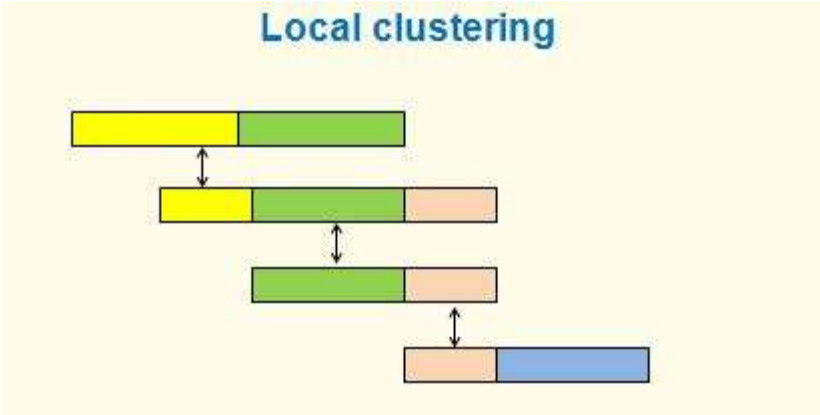
- Выбор порога сходства – проблема
- Зависит от внутривидовой изменчивости баркода
- Порог не может работать для всех грибов
- По умолчанию 97 %

# ДНК-метабаркодинг: обработка данных

## 8. Методы кластеризации OTU

### 1) Локальная кластеризация

Плохой выбор




www.drive5.com

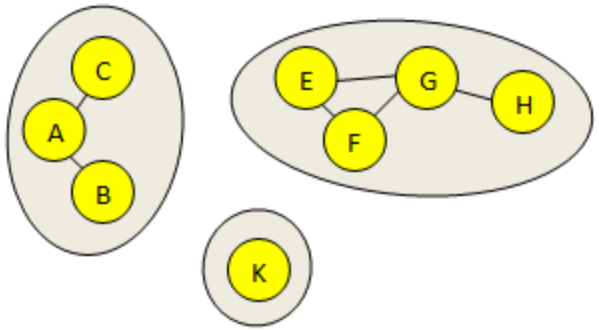
# ДНК-метабаркодинг: обработка данных

## 8. Методы кластеризации OTU

### 2) Агломеративная иерархическая кластеризация

Неравноценные OTU

Применяется редко



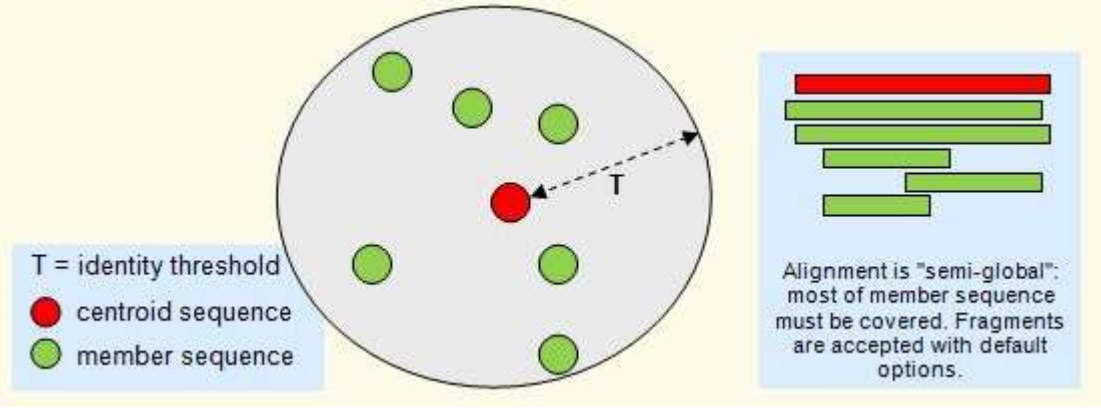

www.drive5.com

# ДНК-метабаркодинг: обработка данных

## 8. Методы кластеризации OTU

### 3) Жадная кластеризация с полуглобальным выравниванием

Самый популярный вариант



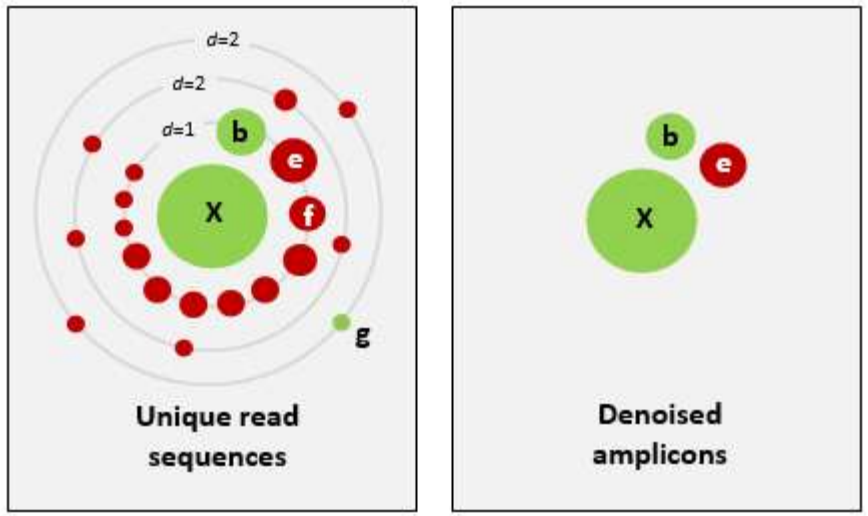

www.drive5.com

# ДНК-метабаркодинг: обработка данных

## 8. Методы кластеризации OTU

### 4) Denoising (устранение шума)

- Отказ от OTU
- Удаление всех малочисленных сиквенсов, окружающих многочисленные

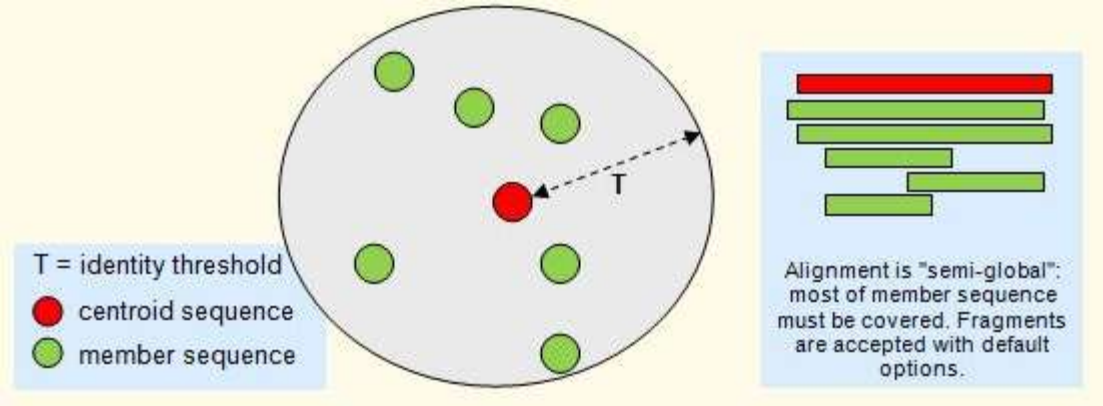



www.drive5.com

# ДНК-метабаркодинг: обработка данных

## 9. Выбор репрезентативных последовательностей

- Обычно – центроидный сиквенс
- Иногда – самый длинный сиквенс




www.drive5.com

# ДНК-метабаркодинг: обработка данных

## 10. Построение таблицы OTU

- Все «сырые» прочтения выравнивают на репрезентативные последовательности OTU
- В результате – таблица

# ДНК-метабаркодинг: обработка данных

## 10. Построение таблицы OTU

| #OTU ID | litt_june_A1 | litt_june_A2 | litt_june_A3 | litt_june_A4_( | litt_june_A5_7 | litt_june_B1_7 | litt_june_B2_7 | litt_june_B3_7 | litt_june_B4_7 |
|---|---|---|---|---|---|---|---|---|---|
| OTU0001 | 2 | 255 | 2424 | 1515 | 36 | 2980 | 2670 | 2906 | 5909 |
| OTU0002 | 107 | 5 | 643 | 1739 | 1465 | 1358 | 4 | 265 | 1440 |
| OTU0003 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| OTU0004 | 2 | 0 | 0 | 3 | 2 | 5 | 1 | 5 | 9 |
| OTU0005 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OTU0006 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| OTU0007 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 100 |
| OTU0008 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OTU0010 | 7 | 2 | 2 | 5 | 372 | 4 | 6 | 6 | 2 |
| OTU0011 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 388 | 0 |
| OTU0012 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| OTU0013 | 23 | 2 | 41 | 19 | 129 | 1 | 90 | 203 | 1 |
| OTU0014 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OTU0015 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| OTU0016 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| OTU0017 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OTU0018 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 101 |
| OTU0019 | 1 | 0 | 2 | 0 | 0 | 0 | 6 | 0 | 0 |
| OTU0020 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| OTU0021 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| OTU0022 | 0 | 0 | 61 | 1 | 632 | 717 | 11 | 433 | 624 |
| OTU0023 | 183 | 536 | 601 | 392 | 69 | 629 | 740 | 1113 | 2 |
| OTU0024 | 0 | 147 | 107 | 0 | 1 | 0 | 15 | 0 | 0 |
| OTU0025 | 0 | 0 | 2509 | 0 | 0 | 0 | 1 | 0 | 3004 |

# ДНК-метабаркодинг: обработка данных

## 11. Отсеивание нецелевых последовательностей

- Всегда присутствуют последовательности нецелевых таксонов
- Выход - BLAST

# ДНК-метабаркодинг: обработка данных

## 12. Таксономическая классификация

- Нужна качественная база референсных последовательностей
- Какие есть варианты?

# ДНК-метабаркодинг: обработка данных

## 12. Таксономическая классификация. Базы данных

### 1) GenBank NCBI

- самая большая
- сама ненадежная

# ДНК-метабаркодинг: обработка данных

## 12. Таксономическая классификация. Базы данных

### 2) BOLD (Barcode of Life Datasystem)

www.boldsystems.org

- Основное хранилище данных iBOL
- Грибы представлены скудно и плохо выверены

# ДНК-метабаркодинг: обработка данных

## 12. Таксономическая классификация. Базы данных

### 3) UNITE

www.unite.ut.ee

- Курируемая база ITS грибов
- 58 тыс. выверенных последовательностей
- Можно скачать в разных форматах
- Любой миколог может добавить или уточнить данные

# ДНК-метабаркодинг: обработка данных

## 12. Таксономическая классификация. Методы

### 1) Наивный байесовский классификатор (RDP classifier)

- Машинное обучение + теореме Байеса
- Не требует выравнивания последовательностей
- Обучается на выборке с известной таксономией
- Много ложноположительных определений

# ДНК-метабаркодинг: обработка данных

## 12. Таксономическая классификация. Методы

### 2) BLAST

- Локальное выравнивание – плохой вариант

### 3) (Полу)глобальное выравнивание

- Самый адекватный способ
- Выравнивание OTU на референсные последовательности с выбранным порогом сходства

# ДНК-метабаркодинг: обработка данных

## 13. Экологическая аннотация последовательностей

FUNGuild – база данных + ПО для автоматизированной аннотации.

Более 11 тыс. грибных таксонов.

- Способ питания (сапротроф, патотроф, симбиотроф)
- Трофическая гильдия (например, патоген животных, почвенный сапротроф или эктомикоризный вид)
- Степень надежности аннотации
- Морфологический тип (агарикоидный, болетоидный, дрожжи и т.д.)
- Ссылка на релевантную публикацию

# ДНК-метабаркодинг: обобщение

## Основные проблемы данных метабаркодинга

- Невозможно отфильтровать все ошибки
- Не все виды можно различить по короткому фрагменту ITS
- Число последовательностей, отнесенных к OTU, слабо коррелирует с числом особей
- Проблема специфичности праймеров и избирательной амплификации
- Каждая проба – это несколько грамм почвы, поэтому нужно много проб для изучения сообществ
- Неизвестно, активные это или покоящиеся формы
- Референсные базы всегда не полны

# ДНК-метабаркодинг: обобщение

## А в чем польза?

- Скорость, воспроизводимость, автоматизированность обработки
- Большие объемы данных
- Большая часть результирующих последовательностей – корректные нуклеотидные последовательности целевой таксономической группы
- Другие методы не дают настолько полной картины состава сообществ

```
vsearch --fastq_mergepairs ./RAW_DATA/1_S1_L001_R1_001.fastq.gz --
reverse ./RAW_DATA/1_S1_L001_R2_001.fastq.gz --
fastqout ./DATA/merged/1/merged.fastq --
fastqout_notmerged_fwd ./DATA/merged/1/notmerged_fwd.fastq --
fastqout_notmerged_rev ./DATA/merged/1/notmerged_rev.fastq --
fastq_allowmergestagger --fastq_maxdiffs 10 --fastq_minovlen 20


usearch -orient concatenated_ssu_nsz.fna -db
~/Documents/Myxogastria/Analysis_2016/DBs/New_DB/\#new_NGSDB_all_sequenced_cu
t_filtered_prolonged.fas -fastaout
~/Documents/Myxogastria/Analysis_2016/new_2017/concat_filt_oriented_ssu_nsz.f
na -notmatched
~/Documents/Myxogastria/Analysis_2016/new_2017/concat_filt_notoriented.fna
```

# Спасибо за внимание!

```
vsearch --derep_fulllength concat_filt_oriented_ssu_nsz.fna --sizeout --
output concatenated_derep_ssu.fna


vsearch --sortbysize concatenated_derep_ssu.fna --minsize 2 --output
concat_derep_nosing_ssu.fna


vsearch --uchime_denovo concat_derep_nosing_ssu.fna --nonchimeras
nonchimeras_denovo.fas --chimeras chimeras_denovo.fas --borderline
borderline_denovo.fas


vsearch --
usearch_global ./chimera_validation/concatenated_filtered_renamed_nsz.fna --
db ./chimera_validation/chimeras_borderline_denovo.fas --id 1 --
```