

# Сборка генома микроорганизмов

CNE 8972 Биоинформатика · 6B05105 Биотехнология  
Неделя 10 · Жаниязов Ж.А., старший преподаватель

# План лекции

- Задача сборки генома
- Контиги, скаффолды, граф де Брёйна
- Сборщик SPAdes
- Оценка качества: N50 и QUAST
- Типичные проблемы

10 / 15

неделя · сборка генома

# Задача сборки

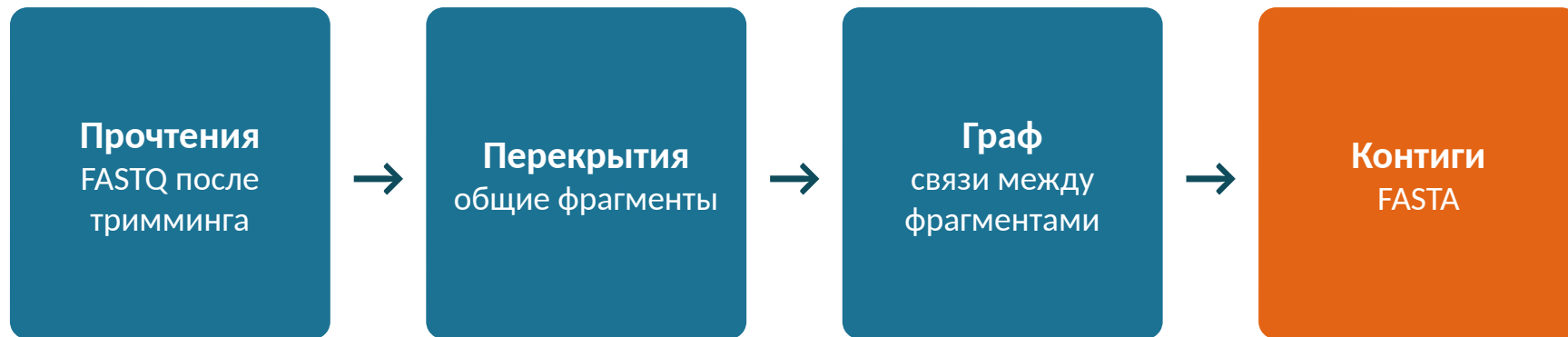
- Секвенатор читает геном короткими фрагментами
- Сборка восстанавливает исходную последовательность
- De novo: без референсного генома
- По референсу: прочтения выравнивают на известный геном



# Основные термины

| Термин       | Значение                                                    |
|--------------|-------------------------------------------------------------|
| Контиг       | Непрерывная последовательность из перекрывающихся прочтений |
| Скаффолд     | Контиги, упорядоченные с разрывами (N)                      |
| Покрытие     | Сколько раз покрыта каждая позиция                          |
| Разрыв (gap) | Место, где последовательность не восстановлена              |

# Идея сборки



Повторы длиннее прочтения разрывают сборку на отдельные контиги.

# Граф де Брёйна

- Прочтения разбиваются на  $k$ -меры
- Узлы —  $(k-1)$ -меры, рёбра —  $k$ -меры
- Путь по графу даёт последовательность
- Значение  $k$  подбирают: SPAdes использует несколько

**$k$ -мер**

короткие слова длины  $k$

# SPAdes

- Сборщик для бактерий и небольших геномов
- Принимает парные FASTQ после тримминга
- Режимы: --isolate, --meta, --plasmid
- Выполняет сборку и строит граф

**SPAdes**

сборщик геномов

# Команда SPAdes

| Параметр                | Значение                   |
|-------------------------|----------------------------|
| -1 R1.fq.gz -2 R2.fq.gz | Парные прочтения           |
| -o out_dir              | Каталог результатов        |
| --isolate               | Режим для чистого изолята  |
| -t 8 -m 32              | Потоки и лимит памяти (ГБ) |

# Результаты сборки

| Файл               | Содержание    |
|--------------------|---------------|
| contigs.fasta      | Контиги       |
| scaffolds.fasta    | Скаффолды     |
| assembly_graph.gfa | Граф сборки   |
| spades.log         | Журнал работы |

# Метрики качества

**N50**

чем больше, тем  
лучше

**L50**

число контигов для  
50% длины

**Контигов**

чем меньше, тем  
лучше

**Длина**

близка к размеру  
вида

**N50 — длина контига, при которой 50% генома лежит в контигах не короче.**

# QUAST: основные показатели

| Показатель        | Что показывает                   |
|-------------------|----------------------------------|
| # contigs         | Число контигов                   |
| Largest contig    | Самый длинный контиг             |
| Total length      | Общая длина сборки               |
| N50 / L50         | Связность сборки                 |
| GC (%)            | Доля GC                          |
| # N's per 100 kbp | Число неопределённых нуклеотидов |

# Запуск QUAST

| Задача          | Команда                                                    |
|-----------------|------------------------------------------------------------|
| Без референса   | <code>quast.py contigs.fasta -o quast_out</code>           |
| С референсом    | <code>quast.py contigs.fasta -r ref.fa -o quast_out</code> |
| Фильтр по длине | <code>--min-contig 500</code>                              |
| Результат       | report.html и report.txt                                   |

# Типичные проблемы

## Много контигов

Низкое покрытие или плохое качество данных

## Размер не тот

Возможное загрязнение другим видом

## Повторы

Разрывают сборку на фрагменты

## Плазмиды

Могут быть собраны отдельными контигами

# Итоги и контрольные вопросы

- Чем контиг отличается от скаффолда?
- Что показывает N50?
- Какие входные данные нужны SPAdes?
- Практика: ПР №10 «SPAdes, оценка сборки (QUAST)»

163–173

стр. · [1] Lesk