

NGS: данные микроорганизмов, оценка качества чтений

CHE 8972 Биоинформатика · 6B05105 Биотехнология

Неделя 7 · Жаниязов Ж.А., старший преподаватель

План лекции

- Что такое NGS
- Этапы получения данных и pipeline курса
- Данные микроорганизмов и покрытие
- FastQC: модули и интерпретация
- MultiQC и типичные проблемы

7 / 15

неделя · NGS и качество

Что такое NGS

- Секвенирование нового поколения: параллельное чтение миллионов фрагментов
- Illumina: короткие прочтения 100–300 нуклеотидов
- Nanopore и PacBio: длинные прочтения
- Применение: геномы, метагеномы, транскриптомы

NGS

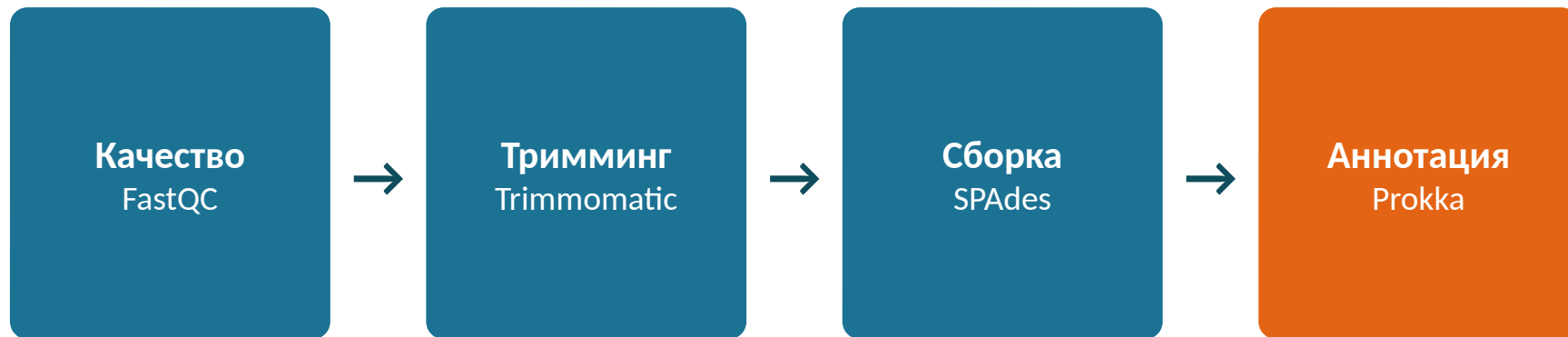
массовое параллельное
чтение

Этапы получения данных



Качество данных закладывается на каждом из этих этапов.

Pipeline курса



После сборки качество оценивают программой QUAST.

Данные микроорганизмов

1–10 Мб

размер генома
большинства
бактерий

4,6 Мб

геном E. coli K-12

≈4300

белок-кодирующих
генов E. coli

≥50×

желательное
покрытие генома

Небольшие геномы бактерий удобно собирать на обычном сервере.

Покрывтие генома

- $\text{Покрывтие} = \text{число прочтений} \times \text{длина прочтения} / \text{размер генома}$
- Пример: 2 млн прочтений по 150 н. дают 300 Мб данных
- Для генома 4,6 Мб покрывтие составит около 65x
- Чем выше покрывтие, тем надёжнее сборка и поиск вариантов

≈65x

покрывтие в примере

FastQC: основные модули

Модуль	Что показывает
Basic Statistics	Число прочтений, длина, GC-состав
Per base sequence quality	Качество по позициям прочтения
Per base sequence content	Состав нуклеотидов по позициям
Duplication levels	Уровень повторяющихся прочтений
Overrepresented sequences	Доминирующие последовательности и адаптеры
Adapter content	Доля адаптеров

Как читать качество по позициям

- Зелёная зона: $Q \geq 28$, жёлтая: 20–28, красная: ниже 20
- Качество обычно снижается к 3'-концу прочтения
- Если медиана падает ниже Q20, нужен тримминг
- Разные образцы сравнивают в одном масштабе

Q28

граница зелёной зоны

Типичные проблемы

Падение качества

Концы прочтений: обрезают тримминг

Адаптеры

Остатки технических последовательностей

Высокая дупликация

Мало исходной ДНК или избыток ПЦР

Смещённый GC

Возможное загрязнение другим видом

Запуск FastQC

Задача	Команда
Один образец	<code>fastqc sample_R1.fastq.gz -o qc/</code>
Все файлы, 4 потока	<code>fastqc *.fastq.gz -t 4 -o qc/</code>
Результат	файлы .html (отчёт) и .zip

MultiQC

- Объединяет отчёты многих образцов в один HTML-файл
- Запуск: `multiqc qc/ -o multiqc_report`
- Позволяет быстро найти образец с проблемой
- Поддерживает вывод многих программ



Контрольный список проверки

Число прочтений

Достаточно ли для покрытия

Качество

Нет ли провалов в середине прочтений

Адаптеры

Присутствуют ли и в какой доле

GC-состав

Соответствует ли виду

Итоги и контрольные вопросы

- Что такое покрытие и как его рассчитать?
- Какие модули FastQC вы знаете?
- Зачем нужен MultiQC?
- Практика: ПР №7 FastQC, MultiQC. СРО №2: групповой проект «Разработка приложения на Python»

130–135

стр. · [1] Lesk