

Предобработка данных: тримминг и фильтрация

CNE 8972 Биоинформатика · 6B05105 Биотехнология

Неделя 9 · Жаниязов Ж.А., старший преподаватель

План лекции

- Зачем нужна предобработка
- Что удаляют из прочтений
- Trimmomatic и его шаги
- Запуск для парных прочтений
- Проверка результата и ошибки

9 / 15

неделя · предобработка

Зачем нужна предобработка

- Сырые прочтения содержат адаптеры и некачественные концы
- Ошибки ухудшают сборку и поиск вариантов
- Предобработка идёт между FastQC и сборкой
- Цель: сохранить как можно больше надёжных данных

Очистка

качество перед сборкой

Что удаляют

Адаптеры

Технические последовательности библиотеки

Низкокачественные концы

Позиции с низким Q

Короткие прочтения

Слишком короткие после обрезки

Неопределённые основания

Позиции с N

Тримминг и фильтрация

Понятие	Смысл
Тримминг	Обрезка концов и адаптеров внутри прочтения
Фильтрация	Удаление прочтений целиком
Скользящее окно	Среднее качество по окну позиций
Минимальная длина	Порог для оставшихся прочтений

Trimmomatic

- Программа на Java для данных Illumina
- Режимы: SE (одиночные) и PE (парные чтения)
- Шаги выполняются в заданном порядке
- Установка: `conda install -c bioconda trimmomatic`

PE / SE

режимы работы

Основные шаги

Шаг	Действие
ILLUMINACLIP	Удаляет адаптеры
LEADING:3	Обрезает начало, пока $Q < 3$
TRAILING:3	Обрезает конец, пока $Q < 3$
SLIDINGWINDOW:4:20	Окно 4 нуклеотида, средний Q не ниже 20
MINLEN:50	Оставляет прочтения не короче 50

Команда для парных чтений

Элемент	Значение
Режим и шкала	trimmomatic PE -phred33
Вход	R1.fastq.gz R2.fastq.gz
Выход	4 файла: R1_paired, R1_unpaired, R2_paired, R2_unpaired
Параметры	ILLUMINACLIP ... SLIDINGWINDOW ... MINLEN

Результат режима PE

- Получаются 4 файла: 2 парных и 2 непарных
- Для сборки используют парные файлы
- В журнале смотрите долю Both Surviving
- Если остаётся слишком мало данных, смягчите параметры

4 файла
результат PE

Параметры в цифрах

4

размер окна,
нуклеотидов

20

пороговое среднее
качество

50

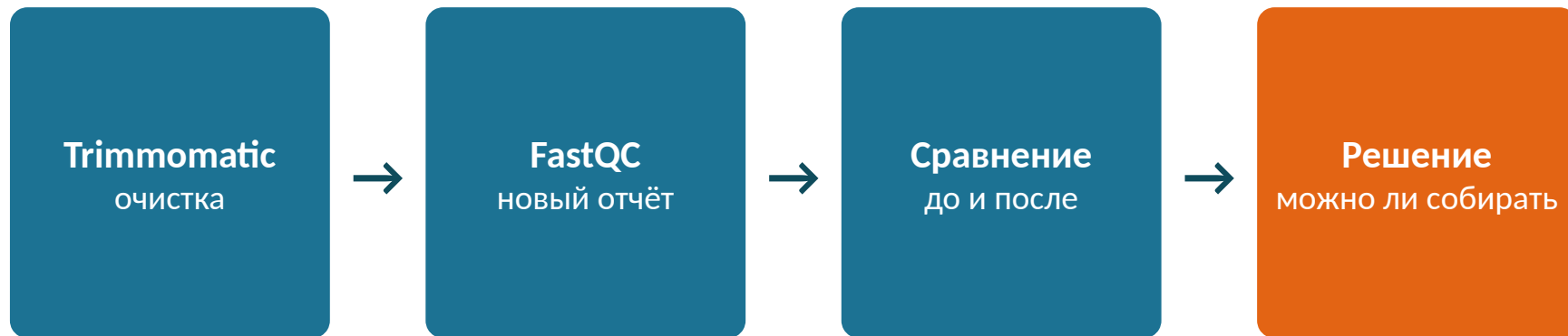
минимальная длина
после обрезки

33

шкала качества
Phred+33

Параметры подбирают по отчёту FastQC, а не наугад.

Проверка после тримминга



Если проблемы остались, меняем параметры и повторяем.

Типичные ошибки

Неверный файл адаптеров

Адаптеры остаются в данных

Слишком строгие параметры

Теряется много данных

Не указан -phred33

Неверная шкала качества

Перепутаны R1 и R2

Нарушается парность

Итоги и контрольные вопросы

- Чем тримминг отличается от фильтрации?
- Что делает шаг SLIDINGWINDOW:4:20?
- Сколько файлов создаёт Trimmomatic в режиме PE?
- Практика: ПР №9 «Trimmomatic»

141–159

стр. · [1] Lesk