

Биологические данные: форматы FASTA и FASTQ

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

План лекции

- Форматы биологических данных
- Формат FASTA
- Формат FASTQ
- Показатель качества Phred
- Базовая проверка данных

6 / 15

неделя · форматы данных

Основные форматы

Формат	Содержание
FASTA	Последовательности ДНК, РНК или белков
FASTQ	Прочтения секвенатора и их качество
VCF	Варианты относительно референса
GFF	Аннотация: гены и признаки на геноме

Формат FASTA

- Заголовок начинается с символа «>»
- Затем идёт последовательность: A, T, G, C, N
- В одном файле может быть много записей
- Расширения: .fa, .fasta, .fna, .faa



начало заголовка

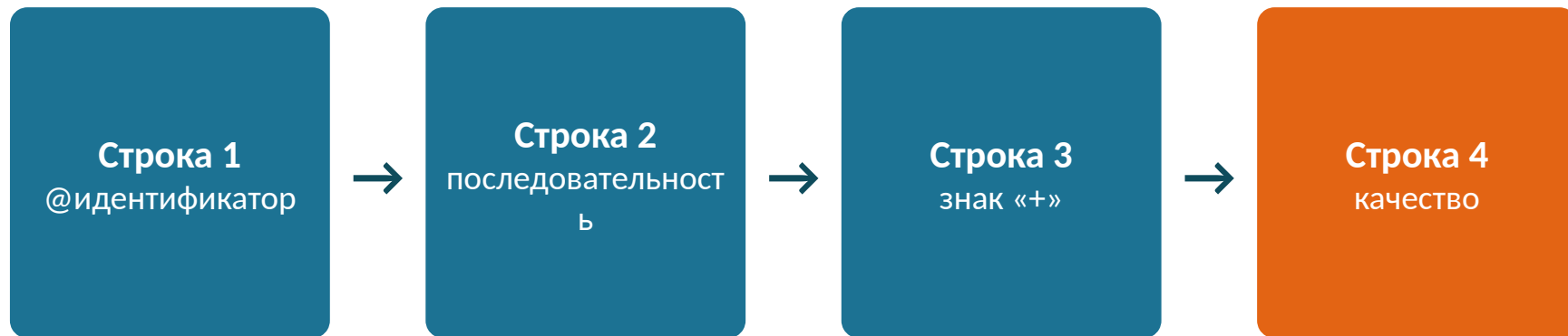
Пример FASTA

Строка	Содержимое
>seq1 E. coli recA	Заголовок: идентификатор и описание
ATGGCTATCGACGAAAACAAACAGAAAGCG	Последовательность (первые нуклеотиды)

FASTA в терминале

Задача	Команда
Число записей	<code>grep -c ">" file.fa</code>
Заголовки	<code>grep ">" file.fa head</code>
Начало файла	<code>head -n 4 file.fa</code>
Статистика (если установлен seqkit)	<code>seqkit stats file.fa</code>

Структура FASTQ



Каждое прочтение занимает ровно четыре строки.

Качество Phred

- $Q = -10 \times \lg(P)$, где P — вероятность ошибки
- Q20 соответствует точности 99%, Q30 — 99,9%
- Качество кодируется символами ASCII (Phred+33)
- Для каждого нуклеотида своё значение качества

Q30

99,9% точность

Шкала Phred

Q	Вероятность ошибки	Точность
10	1 из 10	90%
20	1 из 100	99%
30	1 из 1000	99,9%
40	1 из 10 000	99,99%

Парные чтения

- Paired-end: секвенируются оба конца фрагмента
- Два файла: _R1 и _R2, в одинаковом порядке
- Файлы обычно сжаты: .fastq.gz
- Число прочтений: `zcat файл | wc -l`, результат делится на 4

R1 + R2

парные чтения

Данные в цифрах

4

строки на одно
прочтение

100–300

типичная длина
прочтения Illumina

Q30

ориентир качества

.gz

формат хранения

Число прочтений равно числу строк, делённому на четыре.

Базовая проверка данных

Число прочтений

Хватает ли покрытия

Длина прочтений

Одинаковая или разная

Доля Q30

Сколько нуклеотидов качественные

Адаптеры и N

Признаки загрязнения и неопределённых оснований

Итоги и контрольные вопросы

- Чем FASTA отличается от FASTQ?
- Что означает Q30?
- Как посчитать число прочтений в FASTQ?
- Практика: ПР №6 «Просмотр данных, базовый анализ качества»

40–55

стр. · [2] Baxevanis