

Анализ генотипов и вариантов: SNP, INDEL

CHE 8972 Биоинформатика · 6B05105 Биотехнология
Неделя 13 · Жаниязов Ж.А., старший преподаватель

План лекции

- SNP и INDEL: что это
- Эффект вариантов на белок
- Программа PLINK и её форматы
- Базовые команды и фильтрация
- Фильтрация VCF и типичные ошибки

13 / 15

неделя • варианты и PLINK

Генетическая вариация

- SNP — замена одного нуклеотида
- INDEL — вставка или делеция нуклеотидов
- В геноме человека в среднем 4–5 млн вариантов относительно референса
- Большинство вариантов нейтральны

SNP
и INDEL

Эффект вариантов в гене

Тип варианта	Эффект
Синонимичный	Аминокислота не меняется
Миссенс	Замена аминокислоты
Нонсенс	Преждевременный стоп-кодон
INDEL, кратный 3	Вставка или потеря аминокислот
INDEL, не кратный 3	Сдвиг рамки считывания

Ключевые понятия

MAF

частота минорного
аллеля

HWE

равновесие Харди-
Вайнберга

Call rate

доля образцов с
генотипом

0/1

гетерозиготный
генотип

Эти показатели лежат в основе фильтрации данных.

PLINK

- Программа для анализа геномных данных
- Работает в командной строке, очень быстрая
- Читает VCF и собственные форматы
- Выполняет фильтрацию, расчёт частот и ассоциативные тесты



Форматы PLINK

Файл	Содержание
.bed	Генотипы в бинарном виде
.bim	SNP: хромосома, ID, позиция, аллели
.fam	Образцы: семья, ID, пол, фенотип
.ped / .map	Текстовый формат

Импорт VCF

Команда	Назначение
<code>plink --vcf file.vcf.gz --make-bed --out data</code>	VCF → бинарный формат
<code>--double-id</code>	Идентификатор образца для семьи и образца
<code>--allow-extra-chr</code>	Нестандартные названия хромосом
<code>plink --bfile data --freq</code>	Частоты аллелей

Базовые команды PLINK

Опция	Что делает
--freq	Частоты аллелей
--missing	Доля пропущенных данных
--hardy	Проверка равновесия Харди–Вайнберга
--recode	Перекодировка формата
--extract, --keep	Выбор SNP и образцов

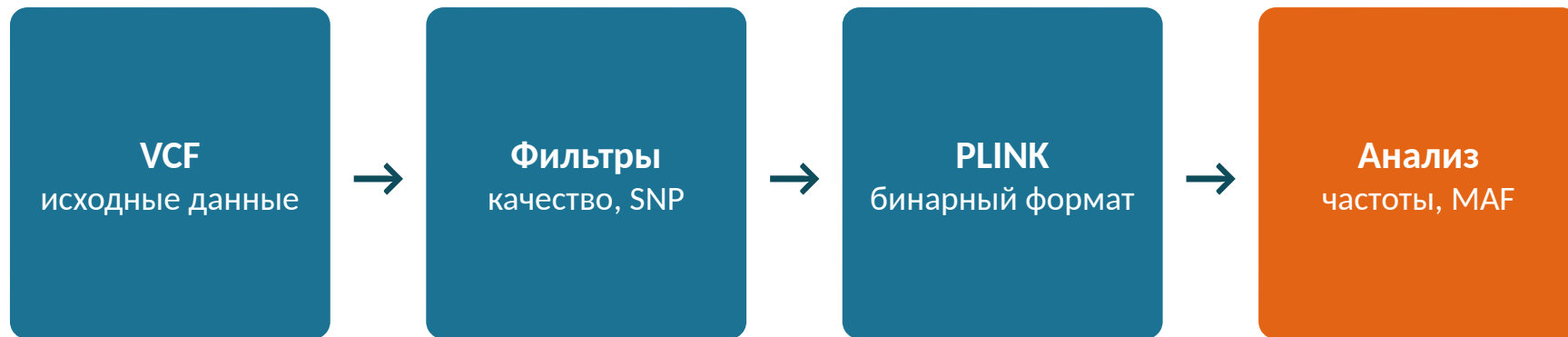
Фильтры PLINK

Опция	Смысл
--maf 0.05	Частота минорного аллеля не ниже 5%
--geno 0.1	Убрать SNP с более чем 10% пропусков
--mind 0.1	Убрать образцы с более чем 10% пропусков
--hwe 1e-6	Убрать SNP с нарушением равновесия
--chr 22	Оставить только хромосому 22

Фильтрация VCF: bcftools

Задача	Команда
Только SNP	<code>bcftools view -v snps file.vcf.gz</code>
Только INDEL	<code>bcftools view -v indels file.vcf.gz</code>
По качеству	<code>bcftools view -i 'QUAL>30' file.vcf.gz</code>
Только прошедшие фильтр	<code>bcftools view -f PASS file.vcf.gz</code>
Биаллельные	<code>bcftools view -m2 -M2 -v snps file.vcf.gz</code>

Схема анализа



Записывайте каждый шаг в скрипт, чтобы анализ можно было повторить.

Типичные ошибки

Нет --out

Результаты перезаписывают друг друга

Разные сборки генома

Не смешивайте hg19 и hg38

Слишком строгие фильтры

Теряется много вариантов

Не читают лог

Проверяйте число образцов и SNP

Итоги и контрольные вопросы

- Чем SNP отличается от INDEL?
- Какие файлы входят в бинарный формат PLINK?
- Что делает опция `--maf 0.05`?
- Практика: ПР №13 «PLINK: базовые команды, фильтрация VCF»

201–214

стр. · [1] Lesk