

Популяционная генетика: базы данных

CNE 8972 Биоинформатика · 6B05105 Биотехнология
Неделя 12 · Жаниязов Ж.А., старший преподаватель

План лекции

- Популяционная генетика и вариации
- 1000 Genomes, gnomAD, Ensembl
- Формат VCF
- Скачивание и просмотр данных
- Частоты аллелей

12 / 15

неделя · базы вариаций

Популяционная генетика на данных

- Изучает вариации в популяциях
- SNP и INDEL — основные маркёры
- Частоты аллелей описывают популяцию
- Открытые базы содержат данные тысяч геномов

SNP

однонуклеотидная замена

Основные базы

База	Данные
1000 Genomes	Геномы людей из 26 популяций
gnomAD	Частоты вариантов по экзонам и геномам
Ensembl	Геномы, аннотация, предсказание эффекта (VEP)
dbSNP	Каталог вариантов и их rsID
ClinVar	Клиническая значимость вариантов

1000 Genomes в цифрах

2504

человека в фазе 3

26

популяций

5

суперпопуляций

≈84 млн

однонуклеотидных
вариантов

Суперпопуляции: AFR, AMR, EAS, EUR, SAS.

gnomAD

- Объединяет данные экзомов и геномов многих проектов
- Показывает частоты аллелей по популяциям
- Помогает понять, насколько вариант редок
- Редкие варианты чаще подозревают в связи с болезнью



gnomAD
частоты вариантов

Ensembl

- Геномный браузер и программный интерфейс
- Аннотация генов и вариантов
- VEP предсказывает эффект варианта на ген и белок
- BioMart выгружает таблицы данных



VEP
эффект варианта

Формат VCF

Поле	Смысл
CHROM, POS	Хромосома и позиция
ID	Идентификатор (rsID)
REF, ALT	Референсный и альтернативный аллели
QUAL, FILTER	Качество и результат фильтрации
INFO	Дополнительные признаки
FORMAT и образцы	Генотипы каждого образца

Генотипы в VCF

Запись	Значение
0/0	Гомозигота по референсному аллелю
0/1	Гетерозигота
1/1	Гомозигота по альтернативному аллелю
0 1	Фазируемый генотип (знак)
./.	Нет данных

Как получить данные



Сначала посмотрите заголовок файла и список образцов.

Команды для работы с VCF

Задача	Команда
Начало файла	<code>zcat file.vcf.gz head</code>
Без метаданных	<code>zcat file.vcf.gz grep -v "^##" head</code>
Число вариантов	<code>bcftools view -H file.vcf.gz wc -l</code>
Список образцов	<code>bcftools query -l file.vcf.gz</code>

Частота аллеля

- $AF = \text{число копий аллеля} / \text{общее число аллелей}$
- Редкие варианты: AF меньше 1%
- У 2504 человек диплоидный набор даёт 5008 аллелей
- Частоту сравнивают между популяциями

5008

аллелей у 2504 человек

Итоги и контрольные вопросы

- Какие базы вариаций вы знаете?
- Что означают поля REF и ALT в VCF?
- Как расшифровать генотип 0/1?
- Практика: ПР №12 «Поиск данных в 1000 Genomes, скачивание VCF»

90–110

стр. · [2] Baxevanis