Домашнее задание – Практикум 13

Общая задача: поиск и аннотация вариантов одного человека по данным экзомного секвенирования на примере одной хромосомы

Задача практикума: получить список вариантов на основании полученного ранее bam файла и аннотировать их средствами VEP

1. **Получение вариантов**

Воспользуемся средствами bcftools.

bcftools mpileup -f chrN.fa file.bam | bcftools call -mv -o file.vcf

(~20-30 минут).

Как устроен полученный vcf файл? Загляните внутрь и опишите его.

Проанализируйте vcf файл с помощью программы **bcftools stats**. Запишите выходную информацию в файл.

a) Сколько получилось вариантов?

b) Сколько из полученных вариантов являются однонуклеотидными заменами?

c) Сколько получилось коротких вставок и делеций?

d) (*) Посмотрите на выход первой части команды (**bcftools mpileup)**. Опишите его.

2. **Фильтрация вариантов.**

Критериев фильтрации существует множество. Наиболее важные - покрытие и качество варианта. Bcftools умеет фильтровать по-разному, попробуем некоторые опции.

К полученному vcf файлу примените:

bcftools filter -i'QUAL>30 && DP>50'.

Обратите внимание, что на выходе вы получите тоже vcf файл.

К фильтрованному vcf файлу примените уже знакомую команду **bcftools stats**.

a) Сколько осталось вариантов после фильтрации (в штуках и в процентах)?

b) Сколько осталось однонуклеотидных замен (в штуках и в процентах)?

c) Сколько осталось коротких вставок и делеций (в штуках и в процентах)?

3. **Аннотация вариантов**

Проаннотируйте полученные выше профильтрованные варианты с помощью сервиса VEP (см. лекцию).

Подайте на вход vcf файл.

a) В отчет включите и подробно опишите информацию из раздела Summary statistics. Обратите внимание (не только в этом пункте задания, но и в целом в отчете), что описание – это не только перечисление выданных программой результатов, а еще и ваше мнение на этот счет. Не стесняйтесь писать, что вас удивляет, настораживает, что не понятно, какие возникли проблемы, чем вам не хватило и т.д.
b) Сколько получилось вариантов с IMPACT HIGH?
c) (*) Как охарактеризованы варианты с импактом HIGH:

- относительно генов (попали в ген или нет)
- относительно структуры генов (экзон\интрон)
- относительно приносимых изменений (missense, splice, …)