Домашняя

28.01.2025

Задача 1:
Для Набора из трех позиций и четырех сэмплов, сделайте PCA. Спроектируйте на главные компоненты, мотивировав чем-то их выбор.

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Задача 2: Дано:

1. Файлы с вариантами chr21.pca.txt, chr22.pca.txt.
2. Текстовый файл со списком сэплов IBS.YRI.MEX.txt
3. файл со списком сэплов и их принадлежностью к популяция IBS.YRI.MEX.info.txt
4. Ноутбук pop.model.l3.ipynb с PCA по 22ой хромосоме для популяций IBS, YRI, MEX.

- Какой процент дисперсии описывает первая главная компонента для 22ой хромосомы?
- Объедините варианты 21ой и 22ой хромосом. Постройте PCA. Сравните результат с предыдущими.
- Нарисуйте трехмерную проекцию на три главных компоненты для 22 хромосомы.
- Случайным образом уменьшите количество маркеров на 50 и на 80 процентов. Визуализируйте и проанализируйте.
- Уравняйте количество в каждой популяции. Изменится ли от этого PCA?
- Попытка работы с данными *. Выберете 4 ваших любимых популяции из проекта 1000GP Phase 3. Список популяций и сэмплов можно найти тут

https://www.internationalgenome.org/data-portal/data-collection/phase-1

Скачать vcf файлы chr22 можно (если не открывается в браузере, использовать wget).

http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

Сделайте два текстовых файла: один со списком сэмплов, расположенных в столбик; второй состоит из двх столбцов Сэмпл, Популяция.

Установить plink и bcftools.

Скачайте скрипт pca.sh 22 list.of.samples.txt (пропишите в нем путь до переменной VCF0 и путь до plink). Запустив его, вы получите текстовый файл, в котором данные отфильтрованы. Нарисуйте проекцию на две главные компоненты. Сопоставьте с тем, что известно про эти популяции из истории и географии.