Learning Your Data: Data Processing of Genetic Data

Workshop Lead: Benjamin Kaufman

Registration link: NA

Approximate duration: 4 hours

Prerequisites:

1. Basic understanding of genetics and genomics concepts
2. Familiarity with Python programming (beginner to intermediate level)
3. Basic statistics knowledge (helpful but not required)

Summary:

Processing open-access genetic data is often a vital step for many human genetic projects, but the process of ensuring that a dataset is updated and clean can be opaque. In this tutorial, students will walk through the process of creating their own dataset from publicly accessible sources. This tutorial will teach students how to understand the formatting of common genetic data formats, update their genomic build, perform standard quality control measures, and carry out standard genetic research techniques such as Principal Component Analyses and Genome-Wide Association Studies.

Learning Objectives:

1. Understand Genome Builds and Liftover Process
2. Perform Quality Control (QC) on Genetic Data
3. Conduct Principal Component Analysis
4. Run a Genome-Wide Association Study

Content:

1. **Module 1 (1.5 hours)**
    a. VCF Formatting and Identifying Genomic Builds (30 -45min)
        i. How to interpret a standard VCF file
        ii. How to correctly identify the genomic build of a dataset
    b. Hands-on activity 1: liftover and alignment (45 -60min)

2. **Module 2 (1 hour)**

a. Quality control measures (30 min)
    i. Missingness and its impact on data.
    ii. Minor Allele Frequency: What is it, and how do we calculate it?
    iii. Hardy-Weinberg Equilibrium, the assumptions of Hardy-Weinberg and its role in standard quality control measures.
b. Hands-on activity 2: quality control on genetic data (30 min)

3. Module 3 (1.5 hours)

a. Principal-Component Analysis (15-30 mins)
    i. What they are conceptually
    ii. How do we use them
b. Genome-Wide Association Studies (15-30 mins)
    i. How do we conduct a GWAS
    ii. What are some of the considerations with GWAS
    iii. How do we interpret a Manhattan Plot
c. Hands-on activity 3: PCA and GWAS (45 -60min)