Module 5 Statistical Genomics 3 – Polygenic prediction

Instructor: Dr Jian Zeng
Tutors: Dr Tian Lin, Dr Moksedul Momin, Dr Xuemin Wang

This module covers key concepts and statistical methods for polygenic prediction of complex traits and diseases. We will learn about the concept, utility and evaluation of polygenic scores (PGS), the methodologies to construct PGS using individual-level data or GWAS summary statistics, data quality controls, and challenges and pitfalls in the application. Additionally, participants will gain insights into Bayesian methods, a methodology widely used in the area, and we will introduce our in-house pipeline for polygenic prediction analysis. In practical sessions, we will use R and command line software, including PLINK, PRSice, GCTA and GCTB.

Prerequisite: If you are not familiar with GWAS, we highly recommend attending Module 1 Genetic Mapping in advance. While attending Module 3 Heritability Estimation is not mandatory, a basic knowledge of statistics and programming in R and Linux is required for this module.

Goal: By the end of the module, you will acquire the skills to compute the polygenic risk scores for the trait of interest from scratch, and will have a general understanding of the complex statistical methods involved in polygenic prediction.

Thursday PM

1-1:40pm **Lecture 1: Fundamentals of polygenic prediction.** We will introduce the concept, utilities, and applications of polygenic prediction, and a basic method to predict polygenic scores (clumping and P-value thresholding (C+PT)).

5 min break

1:45-2:30pm **Practical 1: Calculation of polygenic score using clumping and P-value thresholding.** Software: R, PLINK, PRSice.

2:30-2:45pm Afternoon tea break

2:45-3:15pm **Lecture 2: Best linear unbiased prediction (BLUP) and pitfalls in prediction analysis**. We will learn BLUP for PGS prediction and discuss common pitfalls in practice.

5 min break

3:20-4pm **Practical 2: Calculation of polygenic score using BLUP.** Software: R, GCTA.

Friday AM

9-9:40am **Lecture 3: Bayesian methods for polygenic prediction.** We will learn Bayesian methods that use individual-level data and summary statistics. We will also learn Gibbs sampler, a Markov chain Monte Carlo (MCMC) algorithm for obtaining posterior samples.

5 min break

9:45-10:30am **Practical 3: BayesR and SBayesR.** Software: R, GCTB.

10:30-10:45am Morning tea break

10:45-11:15am **Lecture 4: A Bayesian model incorporating functional annotations**. We will introduce a Bayesian model (SBayesRC) that integrates GWAS summary statistics with functional genomic annotations.

5 min break

11:20-12:00pm **Practical 4: SBayesRC.** Software: GCTB.

Friday PM

1-1:30pm **Lecture 5: Evaluation of PGS for diseases**. We will introduce statistics to assess the prediction accuracy in disease traits and ways of visualization.

5 min break

1:35-2:10pm **Practical 5: Calculation of prediction accuracy for disease.** Software: R, PLINK.

2:10-2:30pm **Lecture 6 (part 1): Summary data quality control.** We will showcase common issues in summary data processing and our solutions.

2:30-2:45pm Afternoon tea break

2:45-3:10pm **Practical 6: Summary data quality control.** Software: R, PLINK.

5 min break

3:15-4pm **Lecture 6 (part 2): In-house PGS workflow and wrap-up.** We will introduce our workflow for PGS prediction using data generated from the lab. We will summarise key points, address final questions, and discuss major challenges in the near future.