Master's thesis @ BiRC

Before you start your thesis, you must make a thesis contract. The thesis contract must be completed and approved by **February 1** (or **September 1**) depending on whether you started your studies in summer or winter. You can read about how to submit the contract on https://studerende.au.dk/en/studies/subject-portals/bioinformatics/masters-thesis/masters-thesis/. As part of the thesis contract, you must attach a pdf file containing project description, project goals, activity plan, and supervision plan using the below template. See http://birc.au.dk/studies/masters-thesis/for more information.

Student ID	202101646
Student name	Zhang Leyi
Group members	
Supervisor	Doug Speed
Project title	An evaluation of software for performing GWAS mixed model association analysis
Hand in date	June 15

Problem statement / project description:

10-20 lines describing the overall aim of the project. Make it clear what the objectives are, e.g. analyse data sets, implement an algorithm, develop or use theory. Remember that the project should amount to 30 ECTS of work (for each group member).
Think of the text as how you would explain your project and its objectives to others.

The most common way to test SNPs for association with a phenotype is via classical association analysis (i.e., least squares regression). However, it is now increasingly common to instead use mixed-model analyses (MMA). The main advantage of MMA is that it reduces false positives due to population structure and relatedness. Further, it can increase power to detect (because instead of testing SNPs individually, it uses a multi-SNP model).

Objective 1 – This project will first evaluate different MMA software (e.g., Bolt-LMM, SAIGE, REGENIE) on large scale data (e.g., 300k individuals and 600k SNPs from UK Biobank) for a wide variety of traits (both simulated and real). Methods will be compared based on type 1 error, power and speed.

Objective 2 – This project will compare strategies for analyzing multi-ancestry datasets. In particular, it will quantify the power increase (or decrease) of MMAA (analyzing all individuals together) with per-ancestry analyses (e.g., analyze European, African and Asian separately, then meta-analyse the three sets of results). In particular, it will investigate whether the optimum strategy depends on consistency of genetic architecture across ancestries.

Objective 3 – This project will construct prediction models (PRS) computed using the results of MMA, and compare their accuracy (and transferability across populations) with PRS computed from classical association analysis.

Problem statement / project goals:

A brief and clear presentation of what you should be able do to after the project. Formulated as 5-7 items, e.g.:
- The student should be able to describe ...
- The student should be able to implement ...
- The student should be able to analyse ...
- The student should be able to discuss ...

…
Think of these items as what you and your project will be evaluated by at the exam.

Methods:
The student should have a detailed understanding of a genome-wide association study (GWAS).
The student should be able to describe the mathematics of MMA. The student should be able to summarize the differences between alternative MMA software.
The student should be able to explain a PRS, and the differences between popular PRS software.

Analyses:
The student should learn how to apply MMA and PRS software to large-scale data from UK Biobank.
The student should be able to devise analyses to test the main objectives of this project (i.e., to compare MMA software on real and simulated data, and to compare PRS software).

Discussion:
The student should be able to evaluate the performance of MMA software on real and simulated data.
The student should be able to provide recommendations for how best to analyze multi-ancestry data.

Activity plan:

10-20 lines describing the overall timeline of your project activities, e.g. formulated bi-weekly milestones. **Think of the text as how you plan to do the project outline in the problem statement.**

1. Literature review on GWAS software and method comparison.
2. Get used to software: BOLT, REGENIE, SAIGE
3. Run association test on mixture data, and perform regressions (logistic, ridge, etc.) on it.
4. Perform Mixed Linear Model on quantitative traits in UKBB data, by REGENIE, BOLT, SAIGE.
5. Binary traits analysis by REGENIE, SAIGE, BOLT
6. Cross Validation and LOOCV on quantitative and binary traits, and compare the accuracy and performance(time and memory).
7. Simulation, to investigate false-positive and power by several software and regressions.
   Consider three situations: 1. Unrelated individuals; 2. Random individuals; 3. Half related and half unrelated individuals.
8. Quantify power, using the mean chi-square test at causal SNPs.
9. Sample a small amount of data, and perform steps 4 and 5 again, to investigate the performance of different software under different data size.
10. SNP LD matrix estimation, for PRS construction.
11. Prediction methods comparison, using the results from MMAA: PRS, LDpred2, PRScs.
12. Write the project.

Each one should be done in one week.

Supervision plan:

A few lines describing the overall structure of your supervision as agreed upon together with your supervisor, e.g. “We plan bi-weekly meetings of ~45 minutes. Specific questions to be addressed at the meeting must be e-mailed to the supervisor at least a day before the meeting in order to give proper time for preparation.”. **Think of the text as an alignment of expectations between you and your supervisor.**

We plan to have 2 meetings each week(on Monday and Thursday). Specific questions to be addressed at the meeting must be e-mailed to the supervisor at least a day before the meeting in order to give proper time for preparation. The student should prepare slides or PDF for presentation of the work in every meeting.