Project Proposal (STAT 605)

Che Tian, Edward Peng, Huajun Gao, Thomas Li, Shu Cao

2025-11-04

1. Question of Interest

Morningness describes an individual's preference of waking up early. Among 10 sleep/health traits around morningness, what are (i) SNP-heritabilities, (ii) pairwise genetic correlations, and (iii) causal effects of morningness?

The github link of our final project is https://github.com/EdwardyfPeng/STAT605-Final-Project.

2. Data

We plan to collect **GWAS (genome-wide association study) data** of 10 traits: morningness, sleep duration, insomnia, daytime sleepiness, daytime napping, snoring, BMI, resting heart rate, physical activity, depression. These GWAS data can be obtained from GWAS Catalog. Take GWAS data of morningness as example, there contains 11 variables:

Table 1: Variables in morningness GWAS summary statistics

Variable	Description
SNP	RSID of variant (or CHR:POS_A1_A0 if no RSID available)
CHR	Chromosome number
BP	Chromosomal position (b37)
ALLELE1	Effect allele that LOGOR corresponds to
ALLELE0	Non-effect allele
A1FREQ	Frequency of ALLELE1
INFO	Information metric indicating imputation quality
LOGOR	Log odds ratio for ALLELE1
LOGOR_SE	Standard error of log odds ratio
P_BOLT_LMM	Non-infinitesimal mixed model p-value
HWE_P	Hardy-Weinberg equilibrium test p-value (~380k samples)

The code of reading data is like that:

```
morning_gwas <- read.table(
  "morning_person_BOLT.output_HRC.only_plus.metrics_maf0.001_hwep1em12_info0.3_logORs.txt")
```

3. Methods

We will use **LD Score Regression** (LDSC) to estimate SNP-heritability for each of the 10 traits. There will be 10 independent tasks on CHTC. And then we will compute pairwise genetic correlations (rg) between all trait pairs using bivariate LDSC. This yields $\binom{10}{2} = 45$ jobs. Next we will perform two-sample **Mendelian Randomization analysis** to assess causal effects between morningness and the other 9 traits, which are also 10 tasks on CHTC. Finally we will aggregate results and generate summary tables/figures such as forest plots, heatmaps, and scatter plots.