6.881 Computational Personal Genomics: Making sense of complete genomes

Lab 7: Estimating heritability from GWAS

Goal

In this lab, we learn how to estimate heritability for a quantitative trait using genome-wide association studies (GWAS). More specifically, we learn how to calculate heritability using GCTA and LD score regression.

Introduction

We have learned a definition of heritability (narrow sense) in Lab 2. Broadly speaking, heritability is the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population. The overall heritability can be estimated from twin study because monozygote twins share the whole genome, and thus the correlation between their phenotypes is attributed to all factors originated from their genetic similarity. The overall heritability estimated from twin study contains all types of genetic variations such as single nucleotide polymorphisms (SNPs), copy number variation (CNV), etc, while the heritability estimated from GWAS contains only SNPs. The problem of missing heritability arises when the heritability of a phenotype estimated from twin study is not consistent with that estimated from GWAS.

1. Estimate heritability using genotype and phenotype data

We can estimate heritability using Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011) if individual data (e.g., genotype and phenotype data) are available. The estimation method is based on a random effect model that decomposes the genetic and non-genetic components (Yang et al., 2010). Specifically, we assume the following additive genetic model for the phenotype

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{Z}$ is the genotype matrix of causal variants, $\boldsymbol{u}$ is the effect size of causal variants, and $\boldsymbol{e}$ is error term. Taking variance on both sides, we obtain

$$\mathrm{Var}(\boldsymbol{y}) = \mathrm{Var}(\mu\mathbf{1}) + \mathrm{Var}(\boldsymbol{Zu}) + \mathrm{Var}(\boldsymbol{e}) = \boldsymbol{ZZ}'\mathrm{Var}(\boldsymbol{u}) + \mathrm{Var}(\boldsymbol{e}).$$

Therefore, if we knew all causal variants, the heritability would be

$$h^2 = \boldsymbol{ZZ}'\mathrm{Var}(\boldsymbol{u})/\mathrm{Var}(\boldsymbol{y}).$$

But in general, we have little knowledge of the true causal variants. Note that $\boldsymbol{G} = \boldsymbol{ZZ}'$ is the genetic correlation of the causal variants between the individuals. If we assume that the genetic correlation $\boldsymbol{G}$ of the causal variants is the same of the overall genetic correlation $\boldsymbol{A}$, we can substitute $\boldsymbol{G}$ with $\boldsymbol{A}$ calculated using all variants**.** However, causal variants tend to has lower minor allele frequency (MAF), GCTA uses a modified version of $\boldsymbol{A}$ to adjust for the bias (Please check (Yang et al., 2010) for more details). In Assignment 1, we learn how to estimate heritability for body mass index (BMI) using GCTA with a real dataset.

2. Estimate heritability using summary statistics

Individual data are often not available but we have only summary statistics from GWAS. Heritability can also be estimated from summary statistics using LD score regression, which is based on an assumption that the more genetic variation an index variant tags, the higher the probability that this index variant will tag a causal variant (Bulik-Sullivan et al., 2015). Bulik-Sullivan et al. show that under a polygenic model, in which effect sizes for variants are drawn independently from distributions with variance proportional to 1/(p(1 – p)), where p is the MAF, the expected χ^2 statistic of genetic variant j given its LD score $l_j = \sum_{k=1}^{M} r_{jk}^2$, where r_{jk} is the correlation between genetic variants j and k, is

$$E(\chi^2|l_j) = \frac{Nh^2 l_j}{M} + Na + 1,$$

where N is the sample size; M is the number of SNPs, such that $\frac{h^2}{M}$ is the average heritability explained per SNP; a measures the contribution of confounding biases, such as cryptic relatedness and population stratification; and l is the LD Score of the variant, which measures the amount of genetic variation tagged. Regressing χ^2 on l_j gives an estimate of h^2. The detailed derivation of $E(\chi^2|l_j)$ is given in (Bulik-Sullivan

et al., 2015). In Assignment 2, we learn how to estimate heritability for BMI using LD score regression with a summary statistics dataset.

Assignment

Please delete all the data after you finish the assignments because these datasets are collected from real individuals.

1. Estimate heritability of BMI attributed to the SNPs on chromosome 19 using GCTA
    a. Download genotype and phenotype data in plink format (bmi.bed, bmi.bim, bmi.fam, bmi.pheno). The phenotype is residuals of BMI obtained from a linear regression adjusted for sex, age and study site. The genotype data contains 10972 SNPs only on chromosome 19.
    b. Download GCTA via http://cnsgenomics.com/software/gcta/#Download
    c. Run GCTA to calculate the genetic correlation matrix $\boldsymbol{A}$ using
    gcta64 --bfile bmi --autosome --maf 0.01 --make-grm --out bmi
    d. Run GCTA to estimate heritability using
    gcta64 --grm bmi --pheno bmi.pheno --reml --out bmi
    e. Report the estimated heritability of BMI attributed to the SNPs on chromosome 19
2. Estimate heritability of BMI attributed to chromosome 19 using LD score regression
    a. Download summary statistics for the 10972 SNPs on chromosome 19 (sumstats_bmi_adj.txt). The summary statistics are obtained from a linear regression adjusted for sex, age and study site.
    b. Download the LD score regression package via https://github.com/bulik/ldsc. You need to have Anaconda to run the package.
    c. Follow the instructions via the following link to download LD scores
    https://github.com/bulik/ldsc/wiki/Heritability-and-Genetic-Correlation
    d. Reformat the summary statistics using
    munge_sumstats.py --sumstats sumstats_bmi_adj.txt --out bmi_adj
    e. Estimate heritability using
    ldsc.py --h2 bmi_adj.sumstats.gz --ref-ld-chr eur_w_ld_chr/ --w-ld-chr eur_w_ld_chr/ --out bmi_adj_h2
    f. Report the estimated heritability of BMI attributed to the SNPs on chromosome 19
3. Answer the following questions based on the results
    a. Are the two heritability estimates consistent with each other?
    b. Does the heritability estimated using LD score regression make sense? What can be the possible problem? (Hint: https://github.com/bulik/ldsc/wiki/FAQ)
    c. Do you see any evidence of population stratification in the result using LD score regression?
    d. Are the individual-based method (GCTA) and the summary-based method (LD score regression) equivalent in some sense? (Hint:
    https://www.biorxiv.org/content/biorxiv/early/2017/10/31/211821.full.pdf)

Reference

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Consortium, S.W.G. of the P.G., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.