Architecture of height, a complex trait

BIO-373: Project 2

December 7, 2020

1 Setting the environment

ggplot2 : please make sure that it is installed in your environment.

resources : download the folder from Moodle.

2 Data exploration

From the **resources** folder:

genotypes.txt is a processed VCF file containing genotyping data of 916 samples for 8722 variants. Each column corresponds to a variant, each line corresponds to a sample. The matrix components are either 0, 1 or 2. These values correspond to homozygous for the reference allele (0), heterozygous (1), homozygous for the alternate allele (2).

variants_info_openSNP.txt contains supplementary information on the variants. Each line corresponds to a variant. Columns give the position, variant ID and corresponding chromosome. The order of the variants matches the order of the variants in **genotypes.txt**

phenotypes.txt contains the height of the corresponding 916 samples.

covariates.txt contains the sex and the platform (type of genotyping array) used for the 916 samples.

1. Start your exploration by using statistical methods to look at the potential effect of covariates on the phenotype. For that, you can use linear regression function **lm()** to see the relationship between your phenotype and the covariates. Use the function **lm()** to fit a model with formula $HEIGHT \sim COVARIATE$. The function **lm()** returns a fitted model object, use it in the function **summary()** to get more statistics about your regression. Look at the Coefficient of Determination (R^2) to know how much the covariates impact your phenotype.

2. Plot the phenotype to visualize its distribution. You can use the *ggplot* function **geom_density()**.

3. Visualize the impact of the sex covariate by plotting the two sex separately on top of the previous plot. Height was obtained through a survey and not measured, it can explain some oddities in the distribution. Feel free to discuss it.

3 Genome-Wide Association Study

You will now conduct a Genome-Wide Association Study (GWAS). You will have to do multiple tests where each variant is tested for its association with the height phenotype.

1. Run a GWAS with and without covariates. You should use a **for** loop (or one of R functional counterpart such as **lapply()**) as a control structure to iterate on every variants. We will use linear regression to test for the association between the variants and the phenotype. Again, use the function **lm()** to build your model and the **summary()** function to get more statistics. Extract at each iteration of the loop the coefficient of association (β) between the variant and the phenotype , and the corresponding pvalue.

2. Using the pvalues from the previous step, produce a Manhattan plot (You should use a $-log_{10}$ scale). You can use the *ggplot* function, **geom_point()**. Using the data in file the **variants_info_openSNP.txt**, alternate two colors between each chromosome. On the Manhattan plot, add a line corresponding to the Bonferroni corrected threshold. You can use the *ggplot* function, **geom_hline()**.

3. Compare the results with and without covariates.

4 Meta-analysis with the GIANT Study

The GIANT consortium produced a GWAS on height based on 250,000 samples. We would like to see how our results based on 916 samples from the OpenSNP cohort compare to that. The file **summary_GWAS_giant.txt** contains the summary statistics from the GIANT study, including the 8702 variants present in our study.

1. There are 20 variants in the OpenSNP cohort not present in the GIANT summary statistics. Take the overlap between the variants from openSNP and the one from the GIANT study.

2. Evaluate the correlation between the coefficients of association you got from the GWAS you ran ($\beta_{openSNP}$) and the one coming from the summary statistics of the GIANT consortium GWAS (β_{GIANT}). You can use **test.cor()**.

3. Visualize this correlation. You can use the *ggplot* function **geom_point()** to plot against each other $\beta_{openSNP}$ and β_{GIANT}.

4. Plot the regression line. You can use **lm()** + **summary()** to the calculate coefficient of association and the intercept of the linear model $\beta_{openSNP} \sim \beta_{GIANT}$. Using these coefficient of association and intercept, plot the regression line on top of the previous plot with the function **geom_abline()**.

5 BONUS: Genomic prediction

You now want to compare the power of these two studies to do some genomic predictions. The files:

genotypes test.txt contains genotyping data of an additional 132 samples with the same 8722 variants

phenotypes test.txt contains the height of the corresponding 132 samples.

covariates test.txt contains the age, sex, and the genotyping arrays used for the 132 samples.

1. You must compare the accuracy of height prediction using the $\beta_{openSNP}$ you calculated in part 3 against the coefficients β_{GIANT} given by the summary statistics of the GIANT study. To do that you will do some operations on matrices as $PredictedHeight = Genotypes \times \beta$. You can use the $\%\star\%$ operator for matrix multiplications. NOTE: as in 4.1, do not forget to remove the variants present in OpenSNP and not in GIANT

2. Using linear regression, evaluate the accuracy of prediction by comparing the R^2 of the predicted height using one or the other β against the real height. As usual, you can use **lm()** function. Use the **summary()** function to get the R^2

3. To boost your prediction accuracy you can use also the covariates from **covariates test.txt** in your prediction model.