KU LEUVEN

Evolutionary and Quantitative Genetics – I0D53A
Quantitative Genetics – Assignment 2
GWAS - Association

Due date of the assignment: 29.12.2023

Your last and first name *Plewa Daria*
Your student number r0976669

The info file is your first place to be. Follow the steps in the info file, and then compile your answers into this document. Try to be as specific and concise as possible in your answers. If you feel you would like to add additional information (screen shots, code…) related to specific answers, you can do so (no obligation!), but please then do this as supplementary information at the end of the document, and referring to it in the respective answer.

Important note: the dataset you are working on is a case control dataset, where the cases are patients with inflammatory bowel disease.

Part 1: Testing for association under an additive model
##################################

1. Give the full Plink command used to perform your logistic regression analysis.

plink --bfile gwa_clean --logistic --ci 0.95 --out logistic

Part 2: Accounting for confounders as covariates in a logistic regression model
##

2. Give the full Plink command used to perform the association analysis accounting for age and sex.

plink --bfile gwa_clean --logistic --ci 0.95 --covar gwa_clean.covar --covar-name AGE,SEX --adjust --out logistic_clean

3. Give the full Plink command you would use when you want to do a GWAS with AGE as an outcome trait instead of the affection status (case-control) as outcome.

plink --bfile gwa_clean --linear --ci 0.95 --pheno gwa_clean.covar --pheno-name AGE --out logistic_clean_age

---

Please upload your answer in a pdf format under Toledo > Assignments

KU LEUVEN

Part 3: Identification of individuals of divergent ancestry
##################################

4. Give the full Plink command used to perform the association analysis accounting for the first 5 PCs

plink --bfile gwa_clean --logistic hide-covar --ci 0.95 --covar gwa_clean.pca.eigenvec --covar-number 3-7 --adjust --out gwa_clean.pca

5. Give the full Plink command used to obtain the genomic control inflation factor lambda. What was the lambdaGC value you obtained?

plink --bfile gwa_clean --logistic --assoc --gc --adjust --out gwa_inflation_factor

lambdaGC: 1.3392

The value 1.14971 was obtained from basic association analysis while the 1.3392 was from the analysis containing the logistic regression.

Part 4: Exploring the results
##################

6. Which SNP is your most significantly associated SNP, and what is its genomic location according to dbSNP? Does this correspond with the location in your *.map (*.bim) file? Why (not)?

rs2066844

It does not correspond with the location in my *.map (*.bim) file. The location according to dbSNP obtained 50712015, while in my *.bim 50745926. It could be influenced by the different reference used by the database for the alignment.

Please upload your answer in a pdf format under Toledo > Assignments

7. Check the degree of LD between your most significant SNP and the second most significantly associated SNP from the same locus. Give both SNP IDs; and the D' and r² value between them. Can you speculate on why the D' and r² values are so different?

rs2066848
chr16:50715185

rs2066844 chr16:50745926	A	G		
C	73	111	184	(0.929)
T	0	14	14	(0.071)
	73 (0.369)	125 (0.631)	198	

Haplotypes		Statistics	
C_G:	111 (0.561)	D':	1.0
C_A:	73 (0.369)	R^2:	0.0444
T_G:	14 (0.071)	Chi-sq:	8.7981
T_A:	0 (0.0)	p-value:	0.003

rs2066844 and rs2066848 are in linkage equilibrium

Two of the most significant SNPs returned the error (rs2066844 and rs5743293), consequently for the analysis I used the most significant SNP and the third one (rs2066844 and rs2066848). D' obtaining 1, results in a great level of linkage disequilibrium between SNPs rs2066844 and rs2066848, while the r^2 obtaining the level of 0.444 (which is quite low) represents the low correlation between two SNPs and in conclusion, there is allelic variation present between them - they show relatively low proportion of shared genetic variation.

The difference between D' and r^2 could be influenced by substantial differences in allele frequencies between 2 SNPs, recombination events, rare variants, or population stratification.

Please upload your answer in a pdf format under Toledo > Assignments

KU LEUVEN

Part 5: Synthesizing the results
##################

8. Write a **brief <u>scientific</u> report** of maximally 3000 characters (excluding title) about the GWA study you have done (from part 3 of the assignment onwards!).

   The report needs to adhere to the following structure:

   - Title
   - Background and aim
   - Methods
   - Results
   - Discussion/Conclusion

   Some tips of what you can include:

   - Something on your dataset (number of samples, variants…)
   - Something on your methods (QC, statistics, tools used…). Even though you did not do the actual QC on this dataset, you know the steps from the first assignment.
   - How many variants/loci were found to be significantly associated, and what are their effect sizes? Which gene(s) do they point to? You can use findings from part 4 (“Exploring the results using web-based tools”) to help with this.
   - What does this tell you about the pathogenesis of the disease under study?

   The report <u>further</u> needs to include the **Manhattan plot**, the **QQ plot**, and a **table with your main results**. You <u>may</u> additionally include two figures and 1 table (so max four figures and two tables in total). All figures and tables need to fit on one page. Don’t forget to also include a brief figure title and legend.

Please upload your answer in a pdf format under Toledo > Assignments

KU LEUVEN

GWAS analysis of patients with inflammatory bowel disease

Background and aim

The dataset containing case-control data coming from patients indisposed by Inflammatory Bowel Disease (IBD) was analyzed using GWAS (Genome-Wide Association Study). IBD is the term for two conditions - Crohn's disease and ulcerative colitis which affect the gastrointestinal tract and can be caused by the inappropriate immune response to the intestinal flora in genetically susceptible individuals. The dataset contained 2017 individuals - 888 males and 1129 females, 1334 cases, and 683 controls with a total of 154873 variants passing the QC. GWAS is a genetic study, which goal is to identify genetic variation associated with particular traits. The paper aims to lead the GWAS analysis of patients indisposed by Inflammatory Bowel Disease and identify potential SNPs responsible for the susceptibility of individuals to IBD. [1]

Methods

For processing the data, and obtaining high-quality results, was performed quality control (QC). QC helps with identifying individuals with discordant sex information, high genotyping failure rates, and outlying heterozygosity rates. SNPs failing at least one of those criteria are eliminated from further analysis. The Principal Component Analysis (PCA) is also performed to retain individuals with common ancestry. Furthermore, the association analysis using Plink version 1.90 was performed. The analysis runs the logistic regression for each SNPs studying the data under 5 principal components, resulting in odds ratios, standard errors, and p-values that later are used in visualizing results in Manhattan and QQ plots. Later Plink is also used for the identification of significantly associated SNPs. For the analysis of significant SPNs, and for obtaining more information about the found loci and associated SNPs, LDlink is used. Finally, additional information about the significant SNPs was procured on sites such as dbSNP, gnomAD, and LDlink.

Results

QC resulted in returning a total of 154873 variants, that later were studied. Below the Manhattan plots represent the significant SNPs. The QQ plots show the deviation from expected p-values. Finally were obtained 116 loci with significance under 5x10-8 and 5 under 1x10-8 with the presentation in the table below.

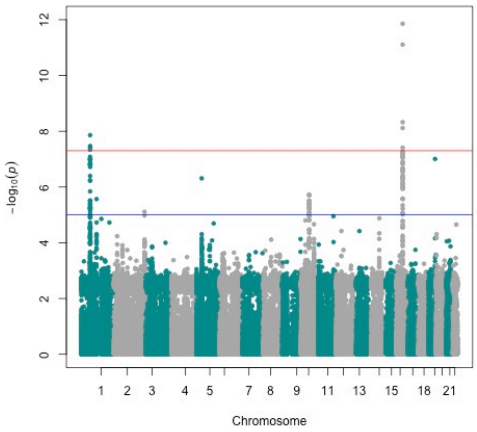
Please upload your answer in a pdf format under Toledo > Assignments

KU LEUVEN

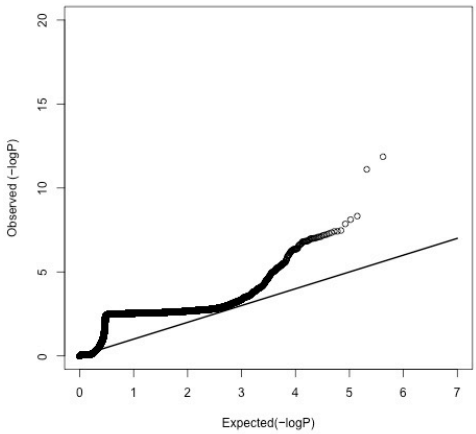
Summary of significant associated loci

Loci	Chr	P-value	OR	Associated Genes
rs2066844	16	1.387e-12	2.4080	NOD2
rs5743293	16	7.781e-12	3.5560	NOD2, CYLD-AS1
rs2066848	16	4.730e-09	0.6694	-
rs7194886	16	7.644e-09	0.6748	SNX20, NOD2
rs11209026	1	1.383e-08	0.4091	IL23R

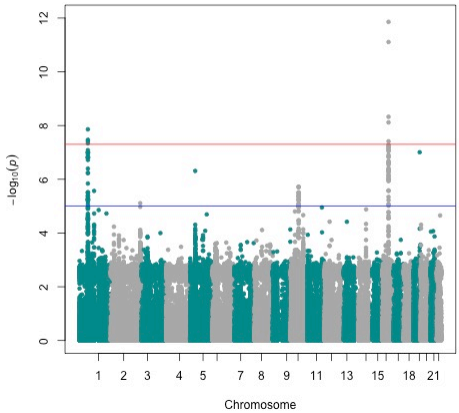
Manhattan plot with systematic bias



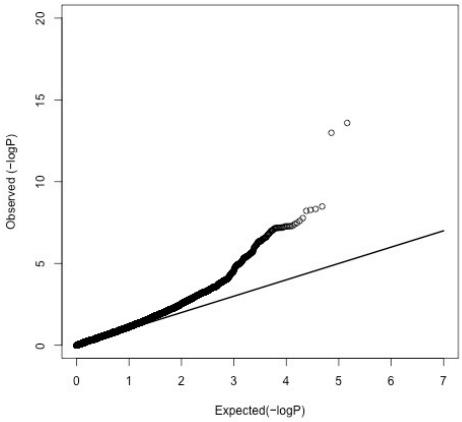
QQplot with systematic bias



Manhattan plot with Bonferroni-adjusted p-values



QQplot with Bonferroni-adjusted p-values



Please upload your answer in a pdf format under Toledo > Assignments

KU LEUVEN

Discussion/Conclussion

The study identified SNPs that could influence the susceptibility of IBD. The rs2066833 could in particular affect the appearance of Inflammatory Bowel Disease. Performed LD analysis showed a strong linkage disequilibrium and low correlation between significant SNPs. The most significant SNPs often were mapped with the NOD2 gene (Nucleotide Binding Oligomerization Domain Containing 2) which is responsible for the recognition of bacterial molecules and inducing proinflammatory and antimicrobial responses.[2] The other mapped genes CYLD-AS1 and IL23R are responsible for oxidative stress-related and inflammatory functions as well as for creating the defenced against bacterial, fungal, and viral infections.[3,4] The mapped genes suggest the chance of a great impact on the appearance of IBD. However, based on the highest significance and the amount of simultaneously mapped genes, I would suggest focusing further analyses on 2 SNPs: rs2066844 with the highest significance from all SNPs, and 1 mapped gene also responsible for immune response and also high OR (2.4080), and rs5743293 with OR (3.5560) which has also very high significance, and 2 mapped genes that can also impact the susceptibility of IBD.

References

[1] McDowell, C.; Farooq, U.; Haseeb, M. Inflammatory Bowel Disease. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2023.

[2] Caruso, R.; Warner, N.; Inohara, N.; Núñez, G. NOD1 and NOD2: Signaling, Host Defense, and Inflammatory Disease. *Immunity* **2014**, *41* (6), 898–908. https://doi.org/10.1016/j.immuni.2014.12.010.

[3] Oxidative Stress-induced lncRNA CYLD-AS1 Promotes RPE Inflammation via Nrf2/miR-134-5p/NF-κB Signaling Pathway. https://doi.org/10.1096/fj.202200887R.

[4] Protective role of R381Q (rs11209026) polymorphism in IL-23R gene in immune-mediated diseases: A comprehensive review. https://www.tandfonline.com/doi/epdf/10.3109/1547691X.2015.1115448?needAccess=true (accessed 2023-12-29).