Arkesh Das
CMSE 410
Advisor: Dr. Jianrong Wang

Assessing the Influence of False Discovery Rate Methods on Genetic Associations in an Immune Response GWAS (Project Proposal)

Background, Goals and Significance

Genome-wide association studies (GWAS) have emerged as powerful tools for linking genetic variants with complex traits, including immune responses. A typical GWAS involves testing millions of single nucleotide polymorphisms (SNPs) for associations, which leads to a substantial multiple testing burden. Applying a simple p-value threshold (e.g., 0.05) is inadequate, as it would generate many false positives. Thus, false discovery rate (FDR) control has become an essential strategy to address this challenge.

The Milieu Intérieur project by Scepanovic et al. (2018) analyzed how human genetic variation influences antibody responses to common pathogens and vaccines. In their genome-wide association study (GWAS), they used the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR). While BH is widely adopted for multiple testing correction, it assumes independence or positive dependence among tests—a condition often violated in GWAS due to linkage disequilibrium (LD) between SNPs. This appears to be the case in their dataset, as when I created a QQ plot from their data, it suggested strong deviations from expected null distributions, indicating that many SNPs may not follow normal assumptions. As a result, the BH method may have missed true associations (being overly conservative) or allowed too many false positives (being too liberal), depending on the underlying correlation structure. These limitations highlight the need to evaluate alternative FDR methods that better accommodate the dependence and sparsity inherent in high-dimensional genomic data.

This project aims to assess how different FDR methods impact GWAS findings. After replicating the BH-based results from Scepanovic et al., I will implement and compare alternative FDR techniques—such as Bonferroni, Benjamini-Yekutieli, and the Storey-Tibshirani q-value approach—on the same dataset. I will visualize and compare the sets of significant variants under each method using Manhattan and QQ plots.

Beyond the research objective, this project is also a learning opportunity for me. I have limited prior experience with biostatistics, because I come from a biochemistry background. Through this project, I am excited to deepen my understanding of GWAS methodology, high-dimensional inference, and FDR control. I also hope to sharpen my R programming skills and apply concepts from previous data science courses to a real-world genomics problem. Learning how to implement and interpret statistical techniques in a GWAS setting will help bridge the gap between what I've learned and the real-world applications of these techniques.

Ultimately, this work could provide a deeper understanding of how different FDR correction methods impact biological discovery and reproducibility in genomics. If successful, it may also offer practical guidance for selecting appropriate statistical methods in future immune-related GWAS studies.

Beyond its technical scope, this project also aligns with a broader shift in biomedical science toward holistic and personalized medicine. As interest in individualized, integrative healthcare continues to rise, initiatives like the Milieu Intérieur project are helping lay the groundwork. By examining how genetic, environmental, and phenotypic variation contribute to immune response, the Milieu Intérieur project exemplifies a systems-level approach to health that embraces complexity rather than treating patients as averages. In this way, I hope that my work contributes to the long-term vision of tailoring medical care to the unique biology of each person.

The Dataset

Background on the data:

This project utilizes data from the Milieu Intérieur (MI) project, a comprehensive, population-based study coordinated by Prof. Lluis Quintana-Murci and Dr. Darragh Duffy at Institut Pasteur in Paris. Established under the French Government's Investissement d'Avenir – Laboratoire d'Excellence (LabEx) initiative, the MI project aims to dissect the interplay between genetics, environment, and immune variation.

It was designed to address a critical discrepancy in medicine: while immune responses vary widely among individuals, medical care and therapeutic strategies are often standardized across populations.

I chose to use this dataset because it offers a unique opportunity to evaluate the impact of false discovery rate methods on GWAS results in the context of immune variation, and the data is high-resolution and well-characterized so it's ideal for benchmarking statistical methods.

The data was accessed from the NHGRI-EBI GWAS Catalog. The Catalog is a publicly available repository of SNP-trait associations identified in published GWAS studies. It provides standardized annotations, p-values, effect sizes, and sample information for millions of SNPs across hundreds of phenotypes.

About the data:

The MI cohort consists of 1,000 healthy individuals, age- and sex-stratified across five decades (ages 20–70), and of french origin. Detailed serological data were collected, including total levels of IgA, IgE, IgG, and IgM, along with antibody responses to 15 antigens from common infectious agents and vaccines (e.g., influenza, EBV, CMV, HSV, rubella). The cohort was genotyped for over 700,000 SNPs and subsequently imputed to yield over 12 million genetic variants. Extensive phenotypic, demographic, clinical, and environmental metadata are also available, including vaccine history, infection exposure, CRP levels, lipid profiles, and lifestyle factors (Scepanovic et al 2018).

Data formatting:

The genotype-phenotype association data from the MI project is structured in standard GWAS summary statistics format. The data is written in a text file (.txt). Each row represents a SNP-trait association and includes the following fields:

- *#CHROM* – Chromosome number
- *POS* – Genomic position of the variant (in base pairs)
- *ID* – Variant identifier (e.g., rsID)
- *REF* – Reference allele
- *ALT* – Alternate allele
- *ALT_FREQ* – Frequency of the alternate allele
- *TEST* – Type of test performed (e.g., additive model)
- *OBS_CT* – Number of observations used in the test
- *OR* – Odds ratio (for binary traits) or beta coefficient (for quantitative traits)
- *SE* – Standard error of the effect size
- *T_STAT* – Test statistic value
- *P* – Raw p-value of the association

Computational Methods / Approach

The analysis will be conducted in R. I expect to use the data.table, qqman, qvalue, stats, multtest, ggplot2 and dyplr packages. Additional packages I may use are plink2, vcftools if data cleaning is needed.

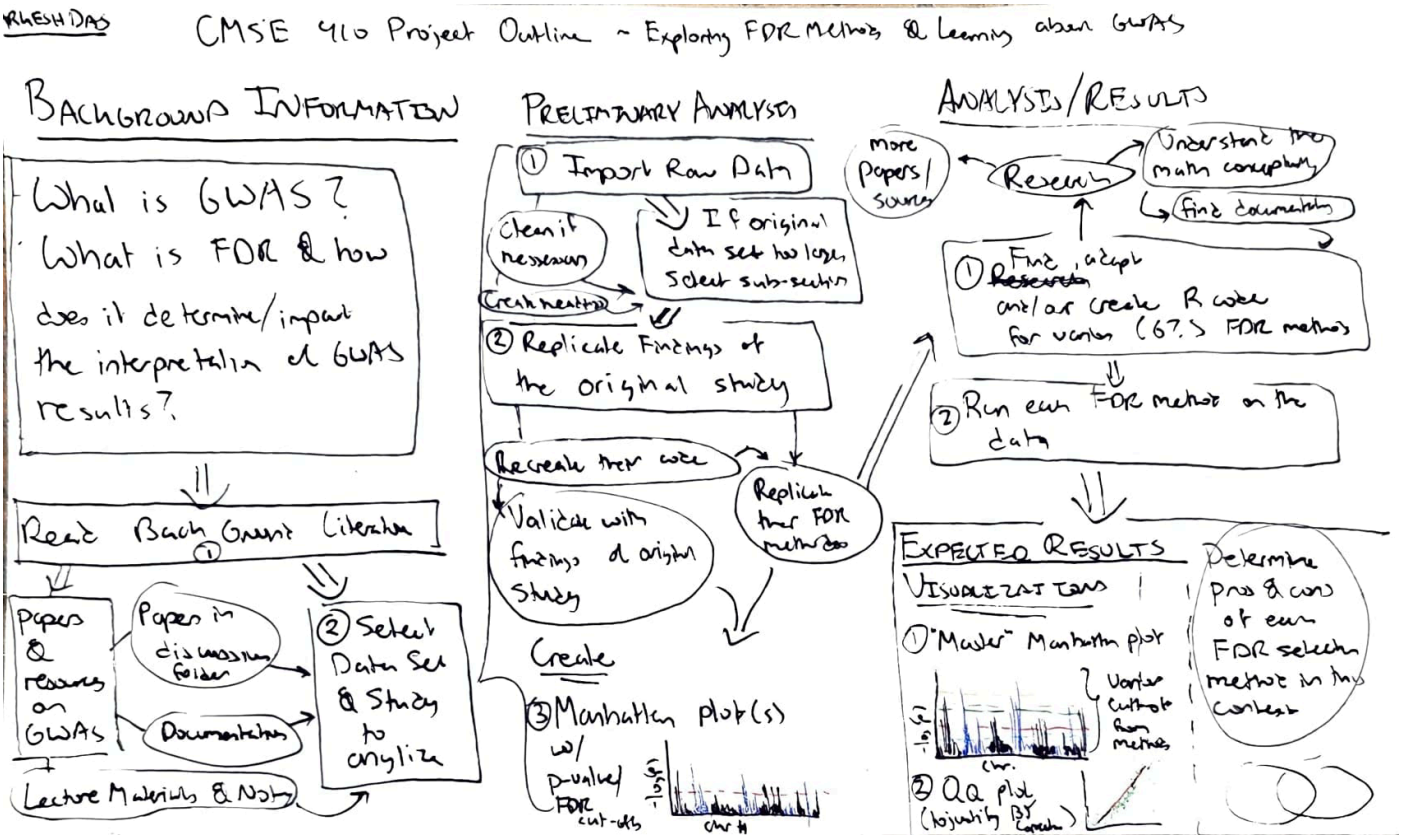
I will follow these steps:

1. Data acquisition & QC
    - Download phenotype and genotype summary data
    - Perform basic cleaning and filtering if needed
2. Reproduction of Scepanovic et al 2018 results
    - Recalculate association p-values if needed (logistic/linear models)
    - Apply BH correction and generate a Manhattan plot
3. Alternative FDR approaches
    - Apply Bonferroni, Benjamini-Yekutieli, and Storey-Tibshirani (qvalue) methods
    - Compare number and type of SNPs declared significant
4. Visualization
    - Generate Manhattan and QQ plots for each method
    - Compare overlaps in significant SNPs between methods
5. Biological interpretation
    - Focus on top SNPs under each method
    - Assess whether immune-related loci (e.g., HLA) remain significant

Evaluation Plan and Milestones

The steps listed above collectively serve as both an analytical workflow and an evaluation framework. By reproducing the original results from Scepanovic et al. (2018), I can verify the correctness of my data handling and statistical implementation. Applying multiple FDR correction methods and visualizing the outcomes through Manhattan and QQ plots will facilitate quantitative comparisons, helping assess each method's sensitivity and specificity. Additionally, the biological interpretation step (especially the focus on immune-relevant loci such as HLA), will allow me to evaluate whether the results make biological sense and remain consistent with known immunogenetic associations.

Flowchart:



Potential challenges & alternative approaches:
This project operates under several assumptions, one of which is that the statistical models used in the original analysis by Scepanovic et al. (2018) were accurate. Any bias in the original study or errors may carry over into my replication and comparison analyses.

There are also methodological challenges inherent to the statistical techniques I plan to implement. For instance, the Storey-Tibshirani q-value method relies on accurately estimating π_0, the proportion of null hypotheses, which might be difficult given some of the highly correlated genomic data. From a practical standpoint, the size of the dataset may pose computational constraints, as processing millions of SNPs with multiple correction methods and visualizations could exceed the capacity of my laptop.

To address these potential limitations, I will remain flexible in my analytical strategy. If necessary, I will restrict the analysis to a representative subset of the dataset (e.g., a single chromosome or trait). Additionally, I may explore permutation-based FDR corrections or empirical Bayes methods as alternative approaches if the standard methods yield inconsistent or unstable results. If implementation challenges persist, I will prioritize a robust comparison between the Benjamini-Hochberg and q-value methods, focusing on their practical differences in immune-relevant loci detection.

References

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology, 8*(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Milieu Intérieur Project. *Institut Pasteur*. Retrieved April 22, 2025, from https://www.milieuinterieur.fr/en/about-us/the-milieu-interieur/

NHGRI-EBI GWAS Catalog. https://www.ebi.ac.uk/gwas/

Scepanovic, P., Alanio, C., Hammer, C., et al. (2018). Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. *Genome Medicine, 10*, 59. https://doi.org/10.1186/s13073-018-0568-8

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics, 101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005