Arkesh Das
CMSE 410
Advisor: Dr. Jianrong Wang

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

Problem and Goals

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) investigated how human genetic variation influences antibody responses to common pathogens and vaccines using genome-wide association studies (GWAS) in a cohort of 1000 healthy individuals. To identify genetic associations, they applied a stringent Bonferroni correction, setting a genome-wide significance threshold of $P < 2.6 \times 10^{-9}$ to adjust for the large number of variants and phenotypes tested. While Bonferroni effectively limits false positives, it is often overly conservative in high-dimensional settings like GWAS, where tests are not independent due to linkage disequilibrium (LD) among SNPs. As a result, no SNPs met this threshold in their primary GWAS analyses, even though some showed strong, biologically plausible signals, particularly in the HLA region. This suggests that true associations may have been missed due to the conservative nature of the correction.

For this project, I set out to explore how different false discovery rate (FDR) correction methods influence SNP-level significance in genome-wide association studies (GWAS), using immune response data from the Milieu Intérieur project. My motivations were both technical and personal. Beyond the research objective, this project was also a learning opportunity for me. I have limited prior experience with biostatistics, because I come from a biochemistry background. Through this project, I aimed to deepen my understanding of GWAS methodology, high-dimensional inference, and FDR control. I also wanted to sharpen my R programming skills and apply concepts from previous data science courses to a real-world genomics problem by replicating a real GWAS workflow and deepening my statistical toolkit.

This project was originally aimed to just explore how alternative FDR control methods influence the discovery of genetic associations in this same dataset. At the beginning of the semester, I was not very familiar with various FDR methods, and so the first thing I did was try to learn more about them. I read a couple of papers to familiarize myself with the field of GWAS as a whole (Bush et al. 2012, Visscher et al. 2017). Then, as the semester progressed, I shifted my focus towards learning more about the specific FDR methods I would choose to implement.

By the end of the semester, after reading the documentation for various FDR methods, I realized that some were more applicable to my problem than others. In the end, I chose to select 4 FDR methods, ranging from most conservative to least: Benjamini-Yekutieli (BY), Benjamini-Hochberg (BH), Benjamini-Krieger-Yekutieli (BKY) and Storey-Tibshirani (q-values). After obtaining my results, I shifted towards understanding the biological relevance of the top SNPs flagged by BKY.

Datasets

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 (Milieu Intérieur. (n.d.). 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, especially since the original authors found no significant SNPs from the GWAS despite their other analysis methods leading them to find biologically significant SNPs. The data is also 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.

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. Of these 12 million SNPs, only the 5.6 million SNPs with a minor allele frequency greater than 5% were kept (Scepanovic et al. 2018).

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

For my analysis, I was only concerned with the traits *#CHROM, POS, ID and P*. I stored the entire dataset in a dataframe, however I conducted my analyses using a subset of 100k SNPs (using a set seed for replicability), because my computer struggled to analyze the full dataset.

Computational Methods / Approach

Originally Proposed Flowchart:

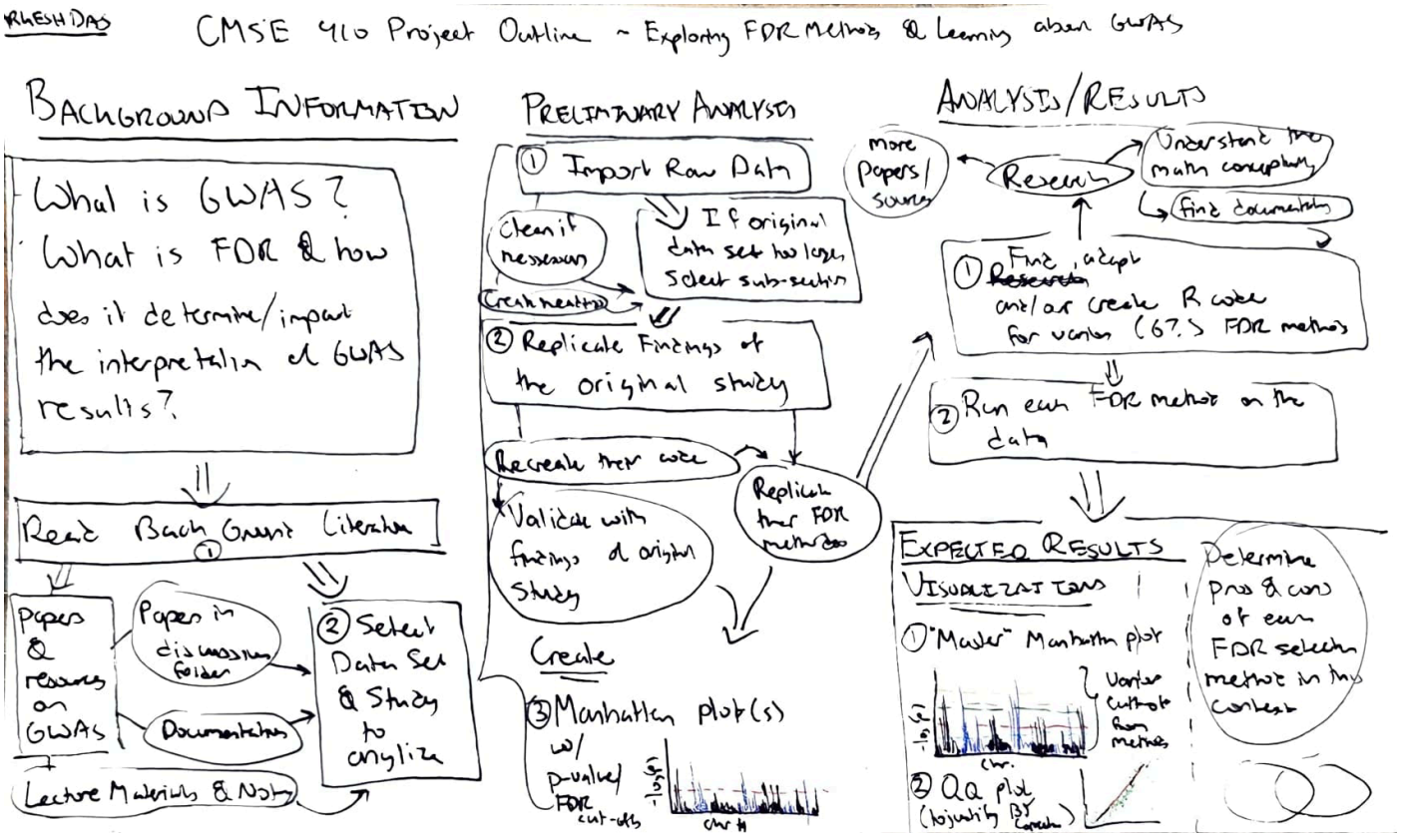

Figure 1: Original Flow Chart

Original Analysis Plan:

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 Bonferroni correction and generate Manhattan and QQ 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) are significant

Steps Completed after the Midterm Presentation:

Step 3: Alternative FDR approaches

- I created a Manhattan plotting function, as well as a QQ plotting function prior to realizing that the qqman package in R already contained built-in functions for these things (Turner 2014). However, these functions ended up being valuable because I recycled my Manhattan plotting function to plot the adjusted p-values using each method (without $\log_{10}$ scaling).
- I then used the p.adjust function to adjust the p-values using the BY and BH corrections. No SNPs passed the 0.05 significance level using either test, so I realized I needed to try an even less conservative approach.
- I originally wanted to use the mutoss package to adjust my p-values using BKY (Sauerbrei et al). However, the package is depreciated in modern versions of R, so I had to instead implement the function myself based on the procedures outlined in the original Benjamini-Krieger-Yekutieli paper (Benjamini 2006). I was able to identify 301 unique SNPs, however they all had very similar adjusted p-values. Therefore, I decided that I should see if I can implement a more granular method, so that I could identify and easily rank the most significant SNPs.
- To do this, I used the qvalue package in R to apply the Storey-Tibshirani FDR method (Storey et Tibshirani 2003). However, when I conducted this analysis, I found π_0 to be 0.98, which suggests that 98% of the tested SNPs are likely null. As a result, none of the q-values fell below the 0.05 significance threshold.

Step 4: Visualization

- To visualize the results of each FDR correction, I used the qqman package's Manhattan Plot function, as well as my custom plotting function. These graphs are shown in the Summary of Key Findings section, as well as on my github page.

Step 5: Biological Interpretation:

- Taking a second look at the p-values that were marked as being significant from BYK, they appear to be randomly clustered throughout the genome. However, the second most significant SNP is located on chromosome 6, the same chromosome as the HLA gene cluster that was found to be significant in the original study. So, I think to conclude, I will try to figure out where this SNP is located and its potential biological function.
- I used the NCBI Human Genome Data Viewer to figure out the genomic context of chr6:129948488. This SNP lies within an intron region of L3MBTL3, situated between a nearby HLA class II gene and an uncharacterized gene. L3MBTL3 has been implicated in several GWAS as influencing hematopoietic traits and immunoglobulin A (IgA) levels. Variants in L3MBTL3 have been associated with natural variation in IgA concentrations, suggesting a potential role in humoral immune regulation. These findings suggest that SNPs in or near L3MBTL3 could indirectly modulate immune responses, potentially by influencing the development or regulation of immune effector cells (Arai et al. 2005).

Summary of Key Findings and Take-homes

<u>Key Finding 1: The original Bonferroni Correction used by Scepanovic et al. was too conservative</u>

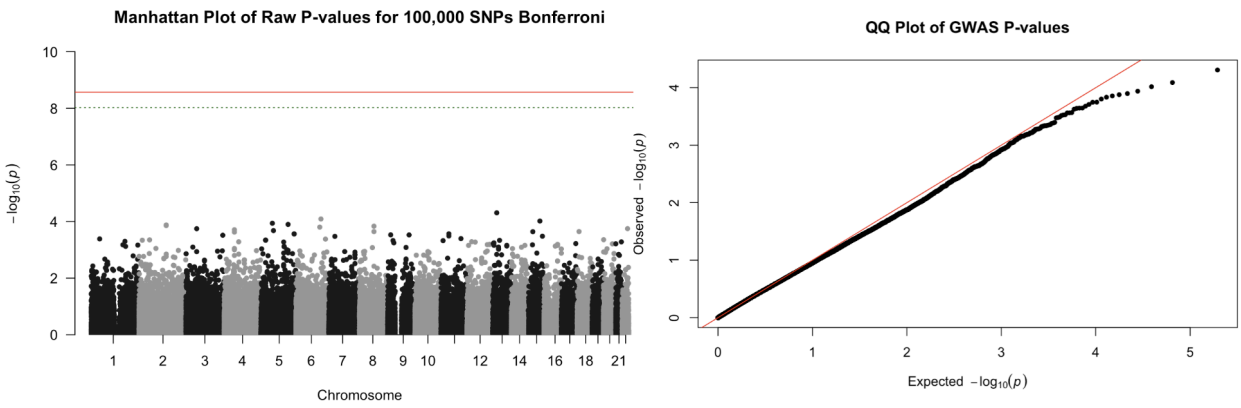

Figure 2: QQ plot and Manhattan plots of original dataset. Green dashed line in Manhattan plot represents the significance threshold they chose to use, and the red line represents a classic Bonferroni Correction ($P < 9.45 \times 10^{-9}$).

The original study by Scepanovic et al. (2018) applied a modified Bonferroni correction to control for the multiple comparisons inherent in GWAS. Since they tested approximately 1 million SNPs across 19 phenotypes, they set an extremely stringent genome-wide significance threshold of $P < 2.6 \times 10^{-9}$. In my analysis, I also implemented the more traditional Bonferroni correction using a simplified approximation: $P < 0.05 / 5{,}290{,}000 \approx 9.45 \times 10^{-9}$, based on the number of common SNPs (MAF > 5%) retained in the cleaned dataset.

Even SNPs with low raw p-values (e.g., $< 10^{-6}$) were deemed non-significant using these significance thresholds, as shown in Figure 2. Therefore, Bonferroni is likely too conservative for GWAS datasets, especially when tests are not independent due to linkage disequilibrium (LD).

Figure 1 also shows that the SNPs deviate from a null distribution. The SNPs deviate from the diagonal early, meaning that there is fewer extreme p-values than expected, even if we assumed that the p-values followed a null distribution.This flattening indicates that there are fewer low p-value SNPs than expected under a null distribution, suggesting a lack of strong signal in the data. This, along with the Manhattan plot, supports that a less conservative significance threshold is needed.

Key Finding 2: BY is more conservative than BH, and BKY is the least conservative

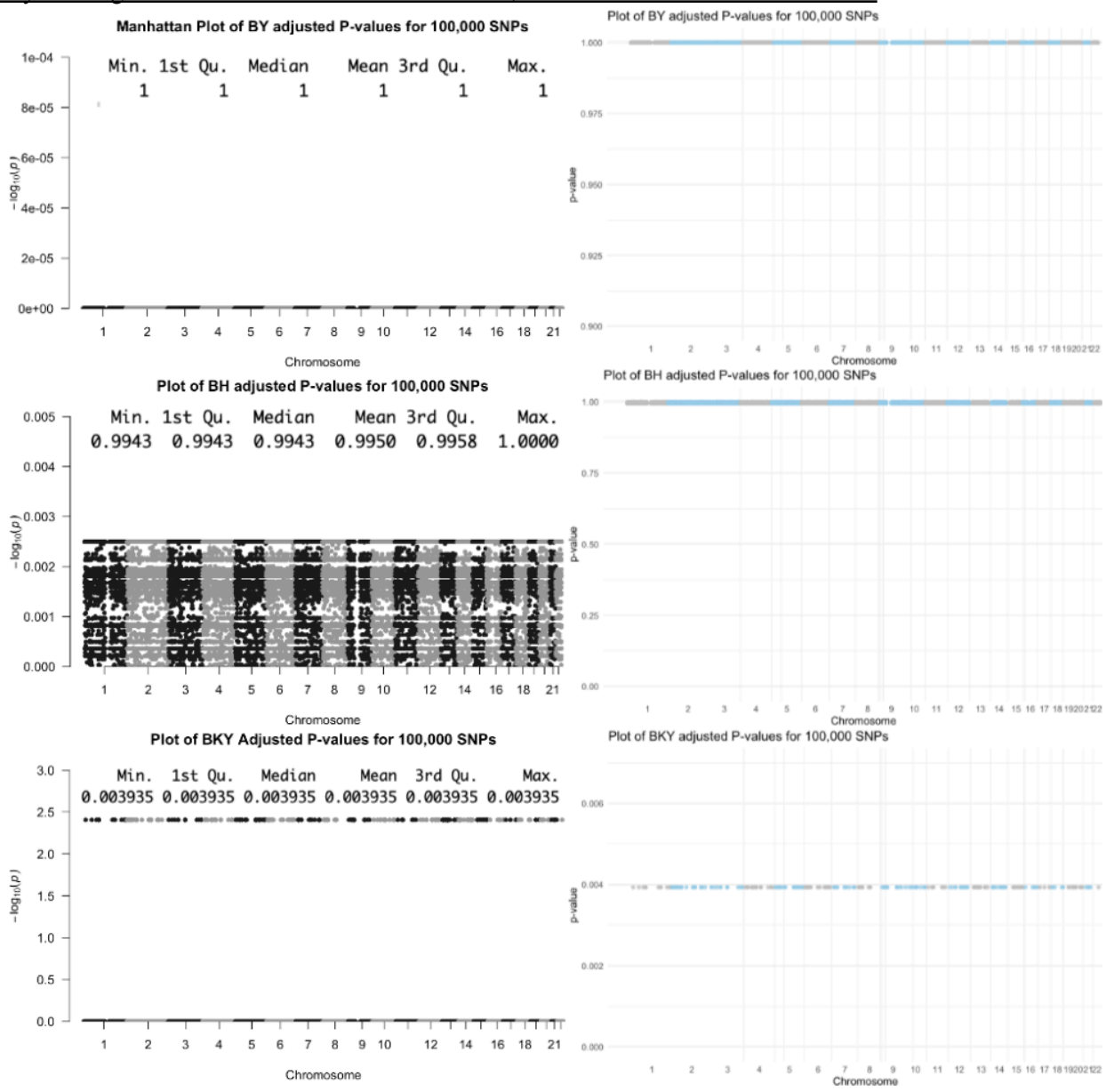

Figure 3: Summary Statistics, Manhattan Plots and P-value plots for each FDR method.

After confirming that the Bonferroni correction was too strict to detect meaningful SNPs, I next applied two widely used false discovery rate (FDR) control methods: Benjamini-Hochberg (BH) and Benjamini-Yekutieli (BY). These methods are specifically designed to be less conservative than Bonferroni. Using the p.adjust() function in R, I applied both methods to the 100,000 SNPs. However, neither BH nor BY produced any significant results at a 0.05 threshold (Figure 3). BY, which adjusts for potential dependence among tests, was expectedly more conservative, but I was still surprised that it produced uniformly adjusted p-values of 1, as it did in this case. BH did create some range in the adjusted p-values, but no SNP was declared significant.

Then, when I applied the Benjamini-Krieger-Yekutieli (BKY) procedure—by manually implementing the algorithm, I was able to identify 301 SNPs that passed the adjusted significance threshold (Figure 3). This marked a substantial improvement in discovery rate compared to BH and BY. The SNPs identified by BKY had nearly identical adjusted p-values, indicating the step-down nature of the method but limiting the ability to rank them by strength of association.

Key Finding 3: High π_0 Value in q-Value Test Reinforced Need to Revisit BKY Results

pi0	double [1]	0.9804612
qvalues	double [5157447]	0.974 0.974 0.974 0.974 0.974 0.974 ...
pvalues	double [5157447]	0.2022 0.3799 0.3309 0.4781 0.0786 0.0786 ...
lfdr	double [5157447]	1 1 1 1 1 1 ...
pi0.lambda	double [19]	1.01 1.01 1.01 1.01 1.01 1.01 ...
lambda	double [19]	0.05 0.10 0.15 0.20 0.25 0.30 ...
pi0.smooth	double [19]	1.01 1.01 1.01 1.01 1.01 1.01 ...

Figure 4: Summary of the q-value object showing parameters such as π_0 and the list of q-values.

After exploring the BKY method, I turned to the Storey-Tibshirani q-value procedure. I used the qvalue package in R, which estimates the proportion of null hypotheses (π_0) to inform its calculation of q-values. When I applied this method to my dataset, I found that π_0 was estimated at 0.98, meaning that 98% of the tested SNPs are expected to be truly null. This extremely high estimate led to uniformly large q-values, and ultimately, no SNPs passed the significance threshold of $q < 0.05$. Looking at Figure 4, the histogram of raw p-values supported the high π_0 estimate: it was largely flat, showing no enrichment of low p-values that would suggest underlying signal.

Given the absence of any significant hits under the q-value approach, I returned to the 301 SNPs identified under the BKY method for further exploration. A key takeaway is that in this context, the BKY-adjusted SNPs served as a more productive foundation for downstream interpretation, offering a focused list of candidates for further biological validation. This highlights an important point: no single FDR method is universally optimal. Depending on the nature of the data and research question, adaptive or hybrid approaches may be more informative than standard pipelines.

Key Finding 4: BKY-Identified SNPs Reveal Potential Immune-Relevant Loci Missed by Original Study

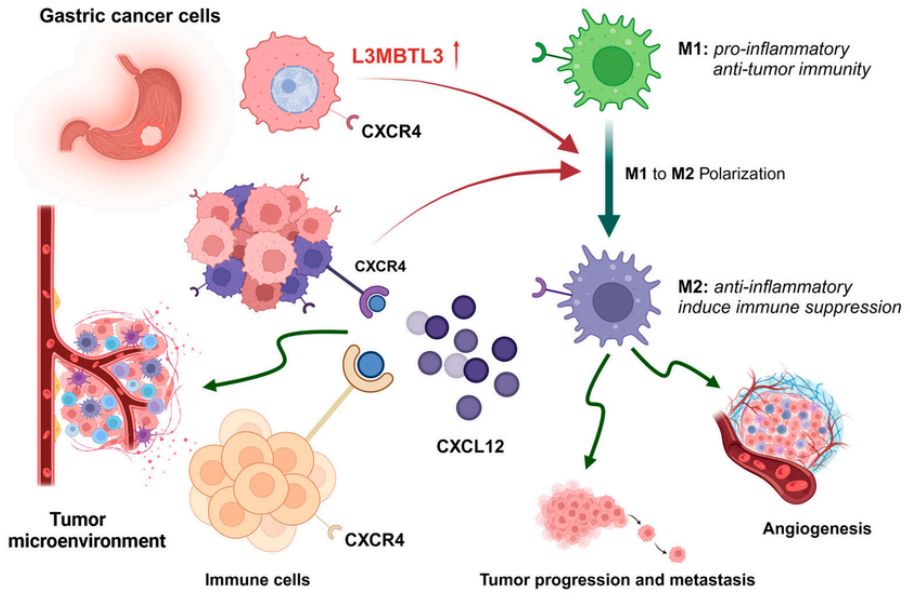

Figure 5: Image showcasing the role of L3MBTL3. Adapted from Gan et al. 2023.

While the original study by Scepanovic et al. (2018) reported no genome-wide significant SNPs under their stringent Bonferroni correction, revisiting the data using the Benjamini-Krieger-Yekutieli (BKY) procedure allowed me to recover a subset of SNPs that may hold biological relevance. One of the top-ranking SNPs stood out: it was located on chromosome 6, the same chromosome where the HLA class II gene cluster is, which is a region previously highlighted in the original study for its immunological significance.

I used the NCBI Genome Data Viewer (National Center for Biotechnology Information, n.d.) to investigate the genomic context of this SNP, specifically chr6:129948488. This variant lies within an intronic region of L3MBTL3, a gene positioned between an HLA class II gene and an uncharacterized open reading frame. Importantly, L3MBTL3 has been implicated in multiple GWAS as influencing hematopoietic traits and levels of immunoglobulin A (IgA). Variants in this gene have been associated with natural variation in serum IgA levels, suggesting a regulatory role in humoral immunity (Arai & Miyazaki, 2005).

Despite the absence of Bonferroni-significant findings in the original study, adaptive FDR methods like BKY can uncover variants with plausible biological relevance. The identification of a SNP within L3MBTL3, in close proximity to HLA genes, suggests a potential regulatory mechanism worth further study in the context of immune response variation. The main takeaway for me was that further analysis of BKY-identified SNPs may yield novel loci that were missed due to overly conservative statistical filters—especially in datasets where weak or polygenic signals dominate.

Code and Data Availability

All code and figures generated for this project are available on GitHub at: https://github.com/arkeshdas/CMSE-410-Semester-Project/tree/main. The repository includes an R notebook containing modular R scripts covering the entire analysis pipeline, from data preprocessing to statistical testing and visualization. Instructions are provided for reproducing all plots and results, as well as documenting my progress and interpretations throughout the process. The dataset is open access and available at http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST006001-GCST007000/GCST006334/ to download to rerun the analysis using the included scripts and environment setup.

The Glossary of Figures and code is located inside of the Github repo.

Challenges

I successfully achieved the broader goal of this project: exploring how alternative FDR correction methods influence the identification of genetic associations in GWAS. I was able to test and compare multiple methods—including BH, BY, BKY, and Storey-Tibshirani q-value—and generate meaningful visualizations and interpretations of the results.

A key technical challenge was the implementation of the Benjamini-Krieger-Yekutieli (BKY) method. The mutoss package is deprecated and non-functional in current R versions. As a result, I had to manually implement the algorithm based on the original 2006 paper. Troubleshooting the implementation and confirming its correctness required a detailed understanding of the method's step-down structure.

Another scientific challenge was interpreting the output of these multiple testing corrections in a dataset where signal appeared to be weak or sparsely distributed. Estimating π_0 using the q-value method, for example, returned an extremely high value (~0.98), indicating that most tests likely reflected null hypotheses. This raised questions about statistical power, effect sizes, and how best to extract meaningful biological insight from a dataset with such a high null burden.

On the practical side, computational limitations forced me to work with representative subsets of the full dataset rather than all 5.7 million variants. Memory usage and slow processing speeds made it infeasible to run full-scale tests on my laptop, particularly for repeated visualization, troubleshooting and testing across FDR methods.

Reflection and Future Directions

If I were to start this project over, I would consider narrowing the scope earlier, like by focusing on a specific phenotype like IgA levels or limiting the analysis to a known immune-related region such as the HLA locus. This more targeted approach may have allowed for deeper biological interpretation and more efficient use of computational resources.

Methodologically, I would also be interested in exploring Bayesian FDR methods or permutation-based approaches as alternatives to classical corrections. These may better capture the correlation structure inherent in GWAS data, especially when signals are subtle. I'd also consider pathway-based aggregation methods or dimension-reduction techniques to detect associations that may be missed at the single-SNP level.

I think one of the most promising extensions of this work is to further investigate the 300 remaining SNPs identified as significant by the BKY procedure. Many of these SNPs appear to be scattered across the genome, but some show signs of clustering (Figure 3). A useful next step would be to determine the genomic locations and associated genes of these variants. Performing pathway enrichment analysis or network-based clustering could reveal functional groupings or regulatory hotspots that influence immune phenotypes.

If I had access to the full dataset, I would also extend this analysis to include SNPs with minor allele frequencies (MAF) below 5%, which were excluded in this project. Although rare variants are often filtered out due to statistical concerns like inflated p-values or low power, the conservative nature of the FDR methods I tested may actually make their inclusion more reasonable. Moreover, in the context of holistic or personalized medicine, understanding the role of rarer genetic variants could be crucial for characterizing less common immune responses and informing individualized care.

There is a growing movement in biomedical science toward personalized and integrative healthcare. Projects like the Milieu Intérieur study are at the forefront of this shift, emphasizing a systems-level view of health that accounts for genetic, environmental, and phenotypic diversity. In that spirit, I hope my work contributes to the broader goal of tailoring medical decisions to the unique biology of each person.

Acknowledgements

Thank you to Dr. Jianrong Wang for giving me the idea of analyzing different FDR methods, and for suggesting that I work with a dataset from the NHGRI-EBI GWAS Catalog.

I would also like to thank Dr. Jason Mezey, whose online lecture slides helped me wrap my head around a lot of these GWAS Concepts.

References

Arai S, Miyazaki T. Impaired maturation of myeloid progenitors in mice lacking novel Polycomb group protein MBT-1. EMBO J. 2005 May 18;24(10):1863-73. https://doi.org/10.1038/sj.emboj.7600654

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive Linear Step-up Procedures That Control the False Discovery Rate. Biometrika, 93(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

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

Gan, L., Yang, C., Zhao, L., Wang, S., Ye, Y., & Gao, Z. (2023). L3MBTL3 is a potential prognostic biomarker and correlates with immune infiltrations in gastric cancer. *Cancers, 16*(1), 128. https://doi.org/10.3390/cancers16010128

Mezey, J. (2013, March 7). *Multiple testing and false discovery rate* [Lecture slides]. Weill Cornell Medical College. https://physiology.med.cornell.edu/people/banfelder/qbio/resources_2013/2013_1_Mezey.pdf

Milieu Intérieur. (n.d.). *The Milieu Intérieur*. Institut Pasteur. https://www.milieuinterieur.fr/en/about-us/the-milieu-interieur/

National Center for Biotechnology Information. (n.d.). *Genome Data Viewer*. U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/gdv/

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

Sauerbrei, W., Binder, H., Royston, P., Schmid, M., & Hothorn, T. (n.d.). *mutoss: Unified multiple testing procedures*. CRAN. https://search.r-project.org/CRAN/refmans/mutoss/html/multiple.down.html

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

Turner, S. D. (2014). *qqman: Q-Q and Manhattan plots for GWAS data* (R package version 0.1.2) [Software documentation]. RDocumentation. https://www.rdocumentation.org/packages/qqman/versions/0.1.2/topics/manhattan

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