**Genome-Wide Association Study of Crohn’s Disease**

Nandita Vaidyanathan

CGH 333 Statistical Methods for Modern Genetics and Epidemiology

**Background**

Inflammatory Bowel Disease (IBD), particularly Crohn's Disease (CD), is one of the major chronic inflammatory conditions in the western world, affecting approximately one in thousand inhabitants (Hugot et al., 1996). Its increasing prevalence in recent decades is hypothesized to be due to emerging environmental factors, though the exact causes remain unclear.

The discovery of the IBD1 susceptibility locus on chromosome 16 provided the first strong evidence for genetic factors in CD risk (Hugot et al., 2001). While early genetic studies using linkage mapping were largely unsuccessful, they did lead to the landmark identification of NOD2 as a major CD susceptibility gene (Verstockt et al., 2018). The field was subsequently transformed by Genome-Wide Association Studies (GWAS), which compare allele frequencies of Single Nucleotide Polymorphisms (SNPs) between cases and controls.

Early GWAS with limited sample sizes identified only a handful of significant associations. However, as sample sizes increased, so did the number of identified genetic determinants of CD, demonstrating the critical role of statistical power in detecting genetic associations (Verstockt et al., 2018). These studies have significantly advanced our understanding of CD's genetic architecture and potential pathogenic mechanisms.

In this study, we conducted a GWAS analysis on a simulated dataset of 2,438 individuals (801 CD cases and 1,637 controls) to explore genetic variants associated with CD. Through this analysis, we aim to understand how genetic variants might contribute to CD susceptibility and aim to validate known associations contributing to disease susceptibility.

**Methods**

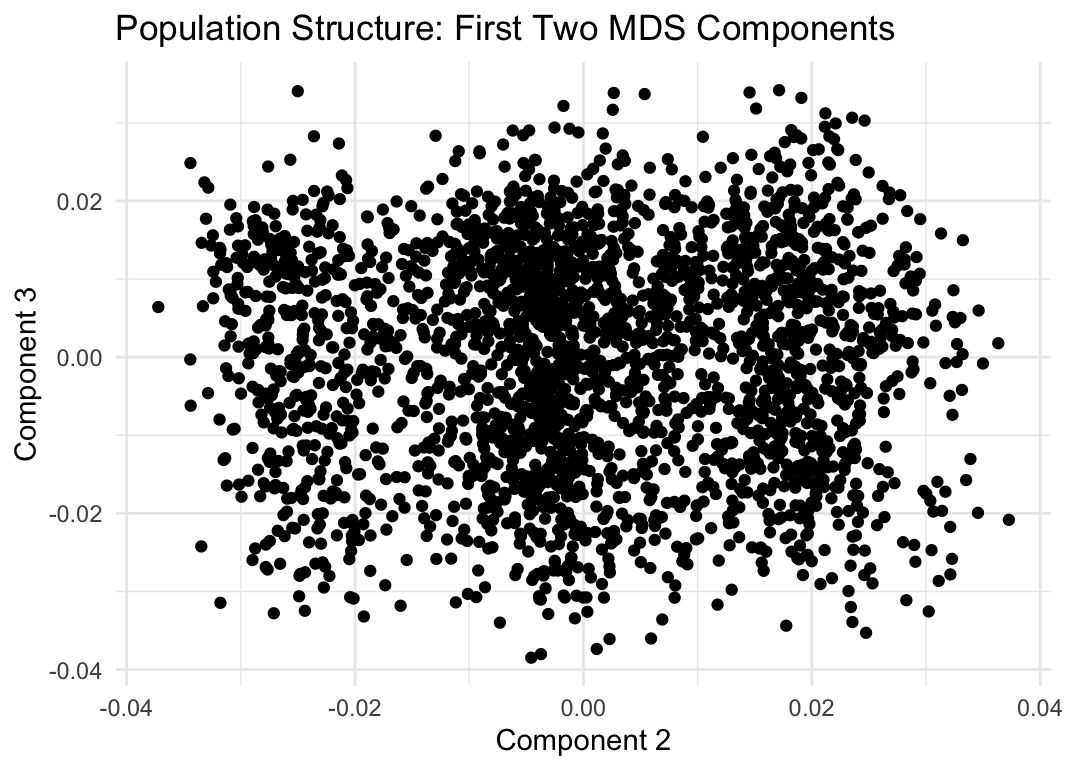
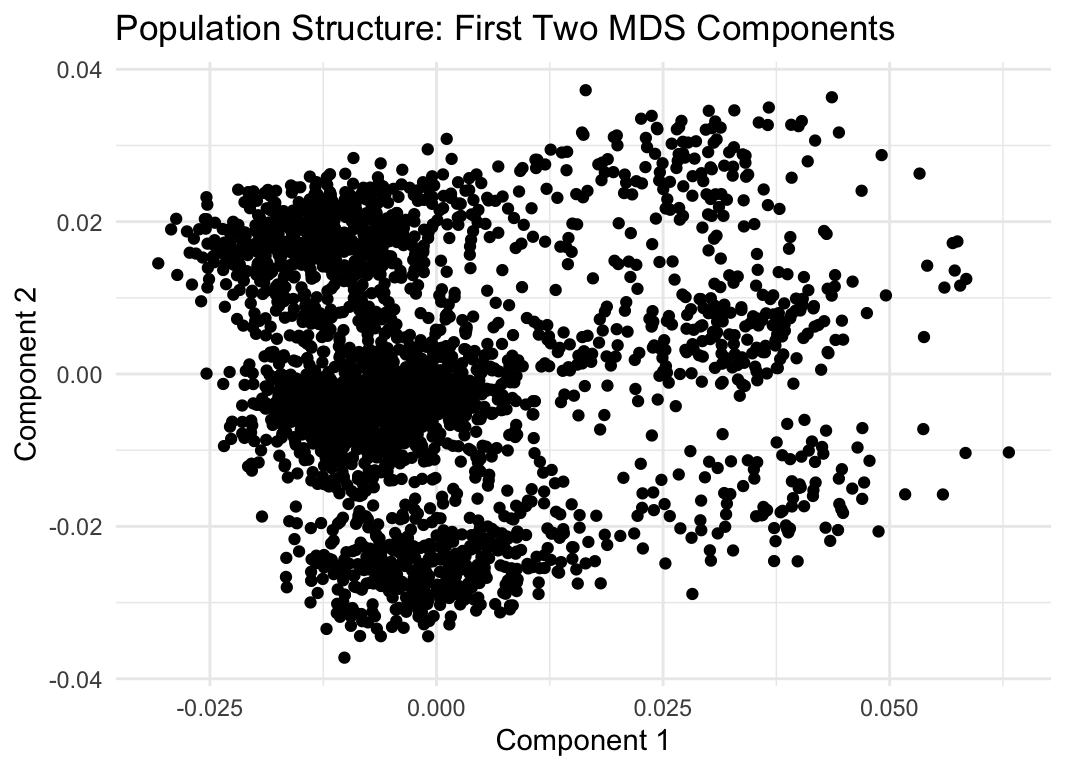
For this study, the Genome-Wide Association Study was used to compare allele frequencies of SNPs between 801 Crohn’s Disease cases and 1,637 controls. The analysis included a total of 2,438 individuals and included phenotype data that included diagnostic categories from Crohn’s Disease, Ulcerative Colitis, Inflammatory Bowel Disease Unclassified, and Controls. A stringent quality control (QC) pipeline was implemented using PLINK v1.9.

First, SNP-level QC had two iterations– initial filtering removed SNPs with missingness >5%, and secondary filtering removed SNPs with missingness >1%. This step is necessary to filter out individuals and SNPs with very high levels of missingness. A missigness test (p < 0.05) was performed to identify SNPs with differential missingness between cases and controls. To remove rare variants, the Minor Allele Frequency (MAF) threshold was set at >= 1% , and includes only SNPs that are above the set MAF threshold. Harvey Weinberg Equilibrium (HWE) was assessed separately in cases and controls to maintain quality control while avoiding removal of potential disease-associated variants. Controls had a more stringent threshold (p > 1e-6) to identify genotyping errors, and cases had a less stringent threshold (p > 1e-10) to account for potential true disease associations.

For sample-level quality control, initial filtering removed individuals with missingness> 10%, and secondary filtering removed individuals with missingness> 2%. To account for sample contamination, sample heterozygosity was assessed using the F-statistic, and individuals with heterozygosity ±3 standard deviation from the mean were flagged for removal. Sample relatedness was evaluated using identity-by-descent (IBD) analysis. Examination of the pairwise IBD metrics showed low relatedness in the dataset as the observed values were consistent with an unrelated population sample. The tight distribution of relatedness statistics (Z0 mean = 0.176 ± 0.016) further supported the quality of the dataset and absence of cryptic relatedness that could potentially bias association results.

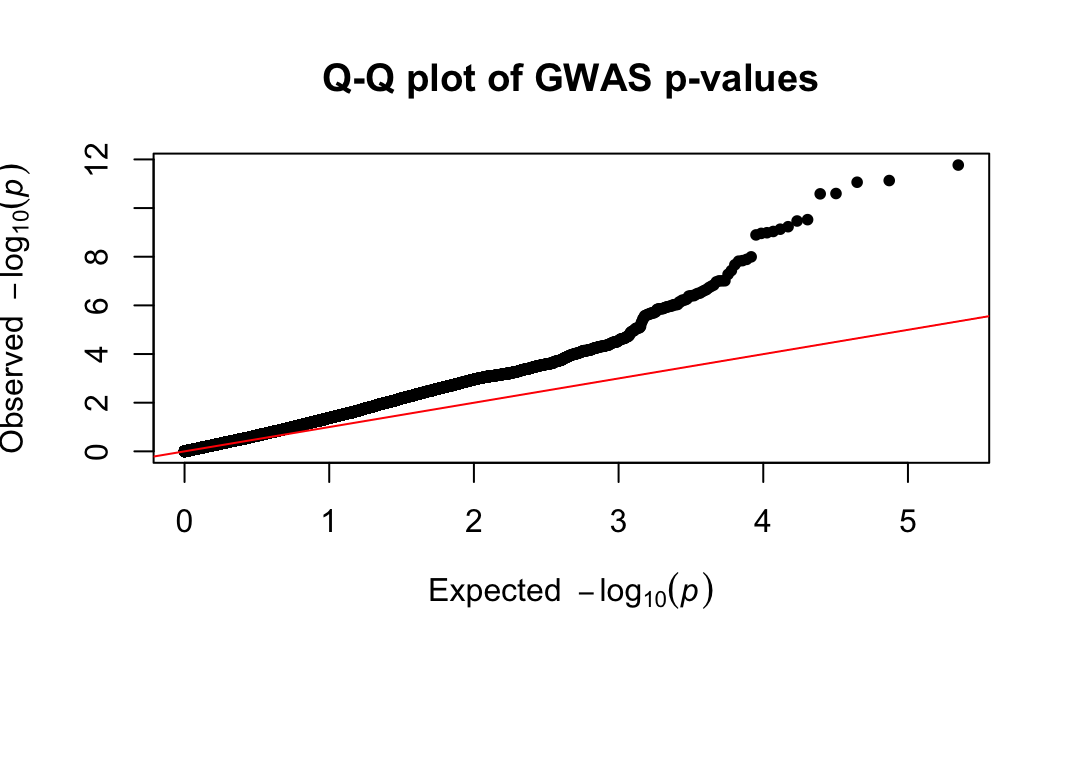
After the quality control process was implemented, population stratification was assessed using multidimensional scaling (MDS) to detect and control for population substructure in the dataset. The analysis was implemented using the --cluster and --mds-plot commands in PLINK to generate 10 dimensions of ancestry. Visual inspection of the first two MDS components revealed moderate population structure– Principal component 1 (C1) ranged from -0.030 to 0.063 and Principal component 2 (C2) showed variation from -0.04 to 0.04. The MDS plot demonstrated some clustering patterns suggestive of population subgroups.

*Figure 1:* Scatterplot of Population structure using MDS Components



Case-control association testing was performed using logistic regression under an additive genetic model. A logistic model was run since CD, the dependent variable, is binary, and it also estimates the probability of having a case.

*Figure 2*: Q-Q Plot of GWAS P-values



To account for population stratification in the association analysis, the first three MDS components were included as covariates in the logistic regression. This approach corrected for potential confounding due to ancestry differences. The quantile-quantile (QQ) plot, from figure 2, demonstrated appropriate control of genomic inflation, with points following the expected diagonal line throughout most of the distribution and deviation occurring only in the upper tail (expected -log10(p) > 4). This pattern suggests well-controlled population stratification while still detecting true genetic associations.

**Results**

After quality control, the final dataset retained 111,344 SNPs and 2,438 individuals (801 CD cases, 1,637 controls). The SNPs are distributed across all autosomes, with highest coverage on chromosomes 1 (11,784 SNPs), 2 (12,518 SNPs), and 6 (13,810 SNPs). The MDS analysis demonstrated moderate population stratification effectively controlled in the analysis.

The results of the GWAS analysis identified two significant regions. The primary signal on Chromosome 16 included thirteen genome-wide significant SNPs clustered around 49.3Mb. The SNP with the strongest association (*imm\_16\_49314275*) showed a significant risk-increasing

*Figure 3:* Manhattan Plot of GWAS Analysis for Crohn’s Disease



effect for CD (p = 1.705×10^-12, OR = 1.81). The secondary signal was identified on Chromosome 1, with three significant SNPs near 67.4Mb. The lead SNP (*imm\_1\_67443356*) showed protective effects (p = 1.012×10^-8, OR = 0.664).

The Manhattan plot in figure 3 visualizes these two distinct regions of genome-wide significant signals, with the primary peak on chromosome 16 reaching -log10(p) ≈ 12 and the secondary peak on chromosome 1 reaching -log10(p) ≈ 8. Additional suggestive signals were observed across other chromosomes but did not reach genome-wide significance.

**Discussion**

Our GWAS analysis identified multiple significant variants on chromosome 16, with the strongest signal at imm\_16\_49314275 (p = 1.705×10^-12, OR = 1.81). Chromosome 16 region corresponds to a known CD risk locus, the NOD2 region. According to GWAS studies, and subsequent meta-analyses, NOD2 region is directly linked to inflammatory gene expression and impaired autophagy (Ashton et al., 2022).

The NOD2 gene plays a vital role in making proteins that help the immune system in fighting against bacteria and viruses. Studies suggest that the particular variation of the NOD2 gene that causes CD, prevents the protein from recognizing the bacteria affecting the lower intestine and causes inflammation (*NOD2 Gene: MedlinePlus Genetics*, n.d.). The risk-increasing effect we observed aligns with the current understanding of NOD2 variants.

The identification of 13 significant SNPs clustering in this region not only validates NOD2 as a major CD susceptibility locus but also demonstrates the polygenic nature of this association. The effect size observed (OR = 1.81) is consistent with previous studies, indicating a substantial increase in CD risk for carriers of these variants (Colombel, 2003; Ashton et al. 2022; Pascoe et al.,2007). While previous studies (Mayor et al., 2012) have focused on three well-characterized NOD2 polymorphisms (rs2066844, rs2066845, rs41450053) that confer approximately two-fold increased risk in heterozygotes, our analysis identified other variants in this region. Our strongest signal (imm\_16\_49314275) shows a comparable effect size to these known variants.

We also identified significant variants on chromosome 1 region corresponding to IL23R, the lead SNP (*imm\_1\_67443356*) showing protective effects (p = 1.012×10^-8, OR = 0.664). This is confirmed by many studies that find protective effects of IL23R variants (Taylor et al., 2008). Taylor et al. (2008) further adds that while IL23R is additive to increase the risk of CD, there is no observed interaction between NOD2 variants and IL23R, suggesting them to act as “two separate pathways” to IBD.

Our GWAS analysis successfully identified significant variants in two key regions associated with CD: the NOD2 locus on chromosome 16 showing risk-increasing effects, and the IL23R region on chromosome 1 demonstrating protective effects. While these findings align with and validate previous research, it's important to note that GWAS results represent only one chapter in understanding the genetic basis of CD. As Taylor et al. (2008) points out, the identification of associated variants, while valuable, does not fully explain the functional mechanisms underlying disease pathogenesis. Additional association studies investigating other variants, complemented with functional studies examining gene-gene interactions and biological pathways, are necessary to more fully evaluate the role of these genes in CD.

**References**

Ashton, J. J., Seaby, E. G., Beattie, R. M., & Ennis, S. (2022). NOD2In Crohn’s Disease—Unfinished business. *Journal of Crohn S and Colitis*, *17*(3), 450–458. https://doi.org/10.1093/ecco-jcc/jjac124

Colombel, J. (2003). The CARD15 (also known as NOD2) gene in Crohn’s disease: Are there implications for current clinical practice? *Clinical Gastroenterology and Hepatology*, *1*(1), 5–9. https://doi.org/10.1053/jcgh.2003.50002

Hugot, J., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J., Belaiche, J., Almer, S., Tysk, C., O’Morain, C. A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Macry, J., Colombel, J., Sahbatou, M., & Thomas, G. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease. *Nature*, *411*(6837), 599–603. https://doi.org/10.1038/35079107

Hugot, J., Laurent-Puig, P., Gower-Rousseau, C., Olson, J. M., Lee, J. C., Beaugerie, L., Naom, I., Dupas, J., Van Gossum, A., Af, N. G. D. T. D., Orholm, M., Bonaiti-Pellie, C., Weissenbach, J., Mathew, C. G., Lennard-Jones, J. E., Cortot, A., Colombel, J., & Thomas, G. (1996). Mapping of a susceptibility locus for Crohn’s disease on chromosome 16. *Nature*, *379*(6568), 821–823. https://doi.org/10.1038/379821a0

Mayor, N. P., Shaw, B. E., Madrigal, J. A., & Marsh, S. G. E. (2012). NOD2 polymorphisms and their impact on haematopoietic stem cell transplant outcome. *Bone Marrow Research*, *2012*, 1–13. https://doi.org/10.1155/2012/180391

*NOD2 gene: MedlinePlus Genetics*. (n.d.). https://medlineplus.gov/genetics/gene/nod2/#conditions

Pascoe, L., Zouali, H., Sahbatou, M., & Hugot, J. (2007). Estimating the odds ratios of Crohn disease for the main CARD15/NOD2 mutations using a conditional maximum likelihood method in pedigrees collected via affected family members. *European Journal of Human Genetics*, *15*(8), 864–871. https://doi.org/10.1038/sj.ejhg.5201839

Taylor, K. D., Targan, S. R., Mei, L., Ippoliti, A. F., McGovern, D., Mengesha, E., King, L., & Rotter, J. I. (2008). IL23R haplotypes provide a large population attributable risk for Crohnʼs disease. *Inflammatory Bowel Diseases*, *14*(9), 1185–1191. https://doi.org/10.1002/ibd.20478

Verstockt, B., Smith, K. G., & Lee, J. C. (2018). Genome‐wide association studies in Crohn’s disease: Past, present and future. *Clinical & Translational Immunology*, *7*(1). https://doi.org/10.1002/cti2.1001