Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

School of Engineering and Sciences

FEATURE SELECTION FROM BIOLOGICAL BIGDATA: IDENTIFICATION OF SIGNIFICANT ASSOCIATIONS APPLYING MULTIVARIATE MACHINE LEARNING ALGORITHMS TO GENOME-WIDE ASSOCIATION STUDIES (GWAS)

A dissertation presented by

DÉBORA GARZA HERNÁNDEZ

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Computers Sciences

Monterrey Nuevo León, May, 2022

Dedication

I dedicate my dissertation work to my family. A special feeling of gratitude to my loving husband, and my little baby. Thanks for all the confidence, support and encouragement. You were my motivation to finish this work.

Acknowledgements

First of all, I would like to thank my family, my aunt Carmen, for her unconditional support and love for me to fulfill my goals, my aunts Sandra and Susy, my sisters Talia and Aglaia, for all the help and advices to finish this work. To my mother in law Maria de los Angeles for all the help in the last year.

I would like to express my special thanks of gratitude to my adviser's Dr Victor Treviño and Dr Karol Estrada who gave me the opportunity to do this l project on the application of multivariate methods to GWAS data, which also helped me to improve my research skills, I am really thankful to them.

To the committee members, Dr Luis Ángel Trejo, Dr. Marion Emilie Genevieve Brunck and Dr. Edgar Emmanuel Vallejo. I'm, grateful for all the comments and suggestions which helped me to improve my presentation skills and the delivery of the research message. Unfortunately, Dr. Edgar Vallejo passed away but left forever imprint of wisdom, dedication, and good will.

Also, to the professors of the computer's science department, who helped me with doubts and concerns about the computer's science field, Dr José Carlos Ortiz Bayliss, Dr Hugo Terashima, Dr Raul Monroy, Dr José Gerardo Tamez and Dr. Emmanuel Martinez.

Additionally, I would also like to thank my friends who helped me a lot by encouraging me to finish the project and to my classmates, professors and collaborators in the group of Bioinformatics, for all the support and advises.

And finally, I would like to thank to the Tecnológico de Monterrey for 100% support on tuition and CONACyT with the support for living.

FEATURE SELECTION FROM BIOLOGICAL BIGDATA: IDENTIFICATION OF SIGNIFICANT ASSOCIATIONS APPLYING MULTIVARIATE MACHINE LEARNING ALGORITHMS TO GENOME-WIDE ASSOCIATION STUDIES (GWAS)

By

DÉBORA GARZA HERNÁNDEZ

ABSTRACT

Crohn's Disease (CD) is a type of Inflammatory Bowel Disease (IBD) affecting the gastrointestinal tract with diverse symptoms. At present, Genome-Wide Association Studies (GWAS) have discovered over 140 genetic loci associated with CD. Usual univariate GWAS methods have allowed the discovery of minor effects from common variants. It assumes independence among them, which can lead to missing subtle combinatorial signals. Considering the importance of CD, multivariate approaches can aid to elucidate the etiology of the disease and facilitate the identification of novel associations. However, current univariate-based and multivariate CD models have a broad performance spectrum and have been assessed in different datasets under diverse methodological settings. Other multivariate methods and models (LASSO, XGBoost, Random Forest, BSWiMS, and LDpred) were compared under a strict sub-sampling and cross-validation approach to predict CD risk in a GWAS dataset (de Lange et al. 2017). The predictions were explored and compared to whether the generated models could provide additional information about variants and genes associated with CD. Additionally, the effect of common strategies was assessed by increasing and decreasing the number of SNP markers (using genotype imputation and LD-clumping). The LDpred model without imputation appears to be the best model among all tested models to predict Crohn's disease risk (AUROC = 0.667 ± 0.024) in this dataset. The best models were validated in a second dataset (NIDDK IBD Genetics), where LDpred was also the best method with similar performance (AUROC = 0.634 ± 0.009). Finally, based on the importance of the variants yielded by the multivariate models, an unnoticed region was identified within chromosome 6, SNP rs4945943, close to gene MARCKS, which appears to contribute to CD risk.

LIST OF FIGURES

LIST OF TABLES

Content

1. CHAPTER 1: INTRODUCTION

1.1 OVERVIEW

The recent advances in DNA sequencing have boosted the necessity of statistical methods to analyze the gathered information. Specifically, Genome-Wide Association Studies (GWAS) are statistical genetic methods that allow the identification of alleles controlling a specific trait (Alqudah et al., 2020). GWAS are observational studies where a genome-wide set of genetic variants in different individuals are analyzed to find any association with a trait. These studies typically focus on associations between Single Nucleotide Polymorphism (SNP) and usually complex human traits (Bush & Moore, 2012).

GWAS are based on the assumption that a marker allele (i.e., SNP), spaced through the genome in linkage disequilibrium LD (i.e., non-random association) with a "causal variant," would be associated with the trait of interest (Stranger et al., 2011). This state of non-random association can be caused by selection, genetic drift, genetic distance, and other effects. However, recombination and gene conversion can break this state (Wigginton, 2005).

To date, GWAS has allowed the identification of common SNP with mainly large effects on phenotype, identifying several novel susceptibility loci (McCarthy et al., 2008). Also, the markers present on the SNP arrays have been selected to be common, to facilitate the variants discovery among different populations. GWAS are biased in terms of what is found; this can be caused by the allele frequencies, affecting the strength of statistical association between alleles. A rare variant in a low LD with a common variant will have fewer probabilities of

being detected in the analysis. Therefore, allowing the detection of causal variants common in the population but has a problem detecting rare variants (Visscher et al., 2012).

Nevertheless, there are statistical challenges to be addressed when a wide genomic study is applied, specifically those that can lead to spurious relationships. The most commonly focused are sample failures, genotyping errors, and population structure (Teo, 2008). Also, gene-gene and gene-environment interactions are of great interest for GWAS development, but understanding them and their relationship with GWAS results is still a significant challenge.

The analysis of genome-wide association data involves a series of single-locus statistic tests, which examines the independent SNP association to the phenotype (Bush & Moore, 2012). However, evidence has shown that many complex traits are highly polygenic, implying that multiple causal variants contribute simultaneously to genetic susceptibility. Thus, examining genetic scores rather than individual SNPs may lead to better insights when studying the genetic contributions for complex traits (Levine et al., 2009).

Most of the existing feature selection approaches, for big data applications, focus on univariate analysis to screen features based on the estimated "individual" effects on the outcome of interest, which is the case of GWAS. But for many complex traits, the underlying mechanisms are neither static nor linear. Therefore, identifying interaction effects among variables will help obtain more accurate phenotype prediction results and reveal functional interactions (Xu et al., 2018).

Currently, the existing implementations of machine learning methods pose several limitations for application to genome-wide data. Further research is needed to identify and

evaluate variable selection procedures that are especially suited for genetic data (Szymczak et al., 2009).

Here I propose that machine learning methods be adapted to GWAS data to improve the predictions' performance and identify significant SNP-phenotype Associations. This can be done by selecting the most predictive SNPs, testing combinations between them, and identifying potential pathways involved in developing the phenotype. Also, I intend to validate the results by fitting a typical univariate analysis and obtaining the standard polygenic scores to compare the accuracy of the multivariate models. This will be addressed using GWAS data of Crohn's Disease (CD).

1.2 PROBLEM DEFINITION

The main problem to be addressed in this work is detailed. The proportion of heritability explained by common variation for most common diseases to date is modest because traditional GWAS do not have the power needed to detect minimal individual effects for determining variants (Ferns et al., 1986). Also, many complex traits are driven by enormously large numbers of variants of minor effects (Boyle et al., 2017). GWAS' assumption that causal variants are in LD with tag SNPs allows the identification of significant associations for those markers, potentially leading to the identification of the causal gene/mutation (Visscher et al., 2012). However, in a genome-wide random SNP approach, many disease-causing genes are missed due to an incomplete or null LD among the variants (Ferns et al., 1986).

GWAS has identified more than 1,200 loci associated with more than 165 common human diseases. However, the heritability of these traits has been poorly explained; this has been

called “missing heritability” (Zuk et al., 2012). Understanding that combinations of rare variants rather than single variants contribute to a significant proportion of the missing heritability is a big challenge. Also, one crucial challenge requires integrating interaction models to determine variants associated with a specific phenotype. This will require the application of statistical and computational methods that detect patterns of interactions among the variants (Eichler et al., 2010). The need for computationally efficient methods is rising, particularly to analyzing exome and whole-genome sequence data (Eichler et al., 2010).

This investigation considers that machine learning approaches are likely complements to standard single-and multi-SNP analysis methods for understanding the overall genetic architecture of complex traits (Szymczak et al., 2009). However, the existing implementations of machine learning methods face several limitations for application to genome-wide data. Further research is needed to identify and evaluate variable selection procedures that are especially suited for genetic data (Szymczak et al., 2009).

1.3 MOTIVATION

The motivation to utilize a GWAS dataset of Crohn’s Disease (CD) relies on the impact of this disease on scientific research. The proposed solution model will test multiple variants together by a feature selection algorithm and evaluate its score as a measure to classify the analyzed phenotypes.

Crohn's Disease (CD) is a type of Inflammatory Bowel Disease (IBD) that affects the gastrointestinal tract with diverse symptoms depending on disease severity (Baumgart & Sandborn, 2012). CD incidence and prevalence in developing countries is considered high;

it has a reported incidence in North America of 6.3 to 23.8 per 100,000 (Ng et al., 2017). As with other complex traits, CD incidence has been theorized to be related to environmental and genetic factors (Feuerstein & Cheifetz, 2017) (Liu & Anderson, 2014).

Additionally, this work considers comparing the performance of multivariate machine learning algorithms with the univariate-based algorithms methods through the polygenic risk scores. To determine which method performs the best for the identification of risk groups. Finally, this work aims to identify potential pathways involved in developing the phenotype by using the information collected from the multivariate models and validating the multivariate machine learning algorithms.

The differences between this proposal and the actual efforts rely on the application of feature selection algorithms combined with filtering strategies that have not been applied to the De Langer et al. CD dataset and their adaptation to reduce the complexity of the analysis.

The following was studied during the investigation:

- Adaptation of feature selection algorithms to apply them to genome-wide data: The general strategies used are: filtering by a threshold of several variants for genetic imputed, non-imputed, and a combination of imputed and non-imputed datasets.
- Implementation of multivariate algorithms adapted to deal with GWAS data, to identify significant associations.
- Comparison of the performance achieved for CD-risk prediction between univariate and multivariate models applied to GWAS data.

- Validation of the multivariate derived' variants associated with CD risk by analyzing the potential biological pathways involved in developing the phenotype.

1.4 HYPOTHESIS, OBJECTIVES, AND CONTRIBUTIONS

1.4.1 Hypothesis

The research is conducted under the hypothesis that applying multivariate machine learning algorithms to genome-wide SNP data will improve the performance and allow the identification of significant association variants for GWAS.

The proportion of heritability explained by common variation for most common diseases to date is modest. Thus, identifying variants interactions concerning the outcome of interest can improve phenotypic prediction. Also, the hypothesis is based on the assumption that applying multivariate methods will generate a discrimination score that will perform better than the traditional polygenic risk scores. This, while controlling for common confounders such as population structure, sample size, etc.

Furthermore, a feature selection algorithm aims to identify relevant features according to a definition of relevance. Concerning GWAS, the feature selection problem can be seen as a search in a set of hundreds of possible solutions, for which adaptation should be performed to reduce the computational complexity which could be derived.

According to this, the following research questions are generated:

- How different are the results obtained by univariate-based and multivariate methods for GWAS analysis?

- Which are the required adaptations for a feature selection algorithm, which must be implemented to obtain reliable results to identify significant association variants for GWAS.
- Is it possible to use multivariate analysis to generate an alternative to the univariate-based polygenic risk scores?
- How is the performance of multivariate models for the phenotype prediction?
- Is it possible to generate a multivariate rank, to order the variants according to the importance achieved within the multivariate models?
- Is it possible to identify novel variants or genes by applying a variant's multivariate ranking?

1.4.2 Objectives

This work aims to identify significant SNP-phenotype associations and polygenic scores through multivariate machine learning algorithms from Genome-Wide Association Studies (GWAS).

This research intended to achieve these particular goals:

- Acquire phenotypic and genotypic data (GWAS) for CD.
- Filter genome-wide SNP data according to established criteria.
- Estimate the classic univariate analysis to validate the GWAS methodology.
- Increment the features by imputing genotypes in LD from a reference dataset.
- Reduce the GWAS dataset by filtering and LD-clumping to decrease computational complexity in subsequent analyses.

- Apply the multivariate and univariate-based methods to predict CD risk at the GWAS data.
- Identify SNPs from reduced GWAS data using "adapted" multivariate methods.
- Compare results from multivariate and univariate methods on GWAS data.
- Establish a multivariate rank for the variants belonging to the best CD-risk prediction model.
- Identify likely key biological sources on variability by using gene ontology and pathways data, using as input the list of genes retrieved from the best CD-risk prediction model.

Particular goals for the "Adapted" multivariate methods

- Adapt multivariate algorithms for feature selection on genome-wide SNP data.
    a. Filter by a threshold of the number of variants for both imputed and not imputed data.
    b. Filter by LD-clumping for both imputed and not imputed data.
- Implement a feature selection algorithm on GWAS data.
    a. Multivariate
        i. Random Forest (RF), LASSO, XGBoost, BSWiMS, and LDpred
    b. Univariate-based
        i. Polygenic Risk Score P+T and Polygenic Risk Score Unadjusted
- Evaluate the performance of multivariate methods for SNP-phenotype Associations.
- Evaluate the adaptation made to the multivariate algorithms using ROC AUC (AUROC).

- Assign variant importance for the multivariate methods, defined by the feature weight observed for each variant within the train set.

1.4.3 Contributions

The contributions of this research are listed below.

- It reviews the impact and applications of multivariate analysis on GWAS to predict the risk of developing a disease.
- Comparisons of multivariate analysis for CD-risk prediction are presented to relate with the results achieved in this investigation.
- A strategy for adding information to the model's trough genotypes imputation is tested and compared with the application of multivariate models on not imputed data.
- LD-clumping was evaluated as a measure of feature reduction by pruning the non-independent variants.
- A robust 10x cross-validation methodology was implemented to select variants by their importance within the models.
- Five multivariate approaches were evaluated for the CD dataset, LASSO, XGBoost, Random Forest, BSWiMS, and LDpred.
- Two common PRS approaches were evaluated for the CD dataset, PRS unadjusted and PRS P+T.
- LDpred outperformed both the common PRS and the other multivariate models, with a mean AUROC of 0.667 ± 0.024 in the testing set.
- A validation dataset was used to evaluate the performance of the models, where the LDpred models also achieved the best performance (AUROC = 0.634 ± 0.009).

- A multivariate rank was constructed based on the importance of the variants within the best multivariate model.
- Based on the importance of the variants yielded by the multivariate models, the unnoticed region within chromosome 6, SNP rs4945943, close to gene MARCKS, was identified.
- Functional analysis for the 402 genes within the CD-risk prediction's best model was implemented, where it shows CD-risk genes linked to well-known inflammatory processes.

1.5 DISSERTATION ORGANIZATION

The present document is structured as Chapter 1 presents an overview of the research, including the problem definition, motivation, hypothesis, objectives, and contributions. Chapter 2 comprises the theoretical framework, including GWAS review, univariate and multivariate methods, and limitations and problems. Chapter 3 describes the solution model and methodology. Chapter 4 presents the results obtained through this investigation, including the performance of the models and the model validation analysis, with its biological interpretation. Finally, Chapter 8 mentions the research's discussion and conclusions, describes the limitations, and proposes the future work to be applied to this topic.

2. CHAPTER 2: THEORETICAL FRAMEWORK

2.1 GENOME-WIDE ASSOCIATION STUDIES

Genome-Wide Association Studies (GWAS) are observational studies, where genomic-wide variants within a group of individuals are tested against a phenotype of interest, usually a major human disease, to find a statistical association between them (Alqudah et al., 2020). A single nucleotide polymorphism (SNP) is a variation at a single position in a DNA sequence among individuals. GWAS are based on the assumption that a variant (i.e., SNP), spaced through the genome and in linkage disequilibrium (LD) (i.e., non-random association of alleles of different loci) with a "causal variant," would be associated with the trait of interest (Stranger et al., 2011).

GWAS had allowed the identification of several alleles associated with different traits of medical interest (Stranger et al., 2011). Due to the improvement of SNP genotyping and microarray SNP analysis, acquiring genotype data has become extremely simple and quick, revolutionizing gene identification and applying GWAS (Stranger et al., 2011).

2.1.1 GWAS LIMITATIONS AND CHALLENGES

GWAS has allowed the identification of several common SNP-based variants, as novel susceptibility loci, with large effects on phenotype (McCarthy et al., 2008). According to the GWAS catalog, up to date, there are 5,457 publications for GWAS, involving 190,974 reported SNPs, which trends for the identification of complex traits' genetic variants (MacArthur et al., 2017).

However, according to Visscher and collaborators in their 2017 review, about 10,000 independent associations (**Figure 1**), with a genome-wide p-value, have been reported in the last ten years between genetic variants and several complex traits (Visscher et al., 2017). This reflects the great gap in identifying significant associations for complex traits through GWAS.

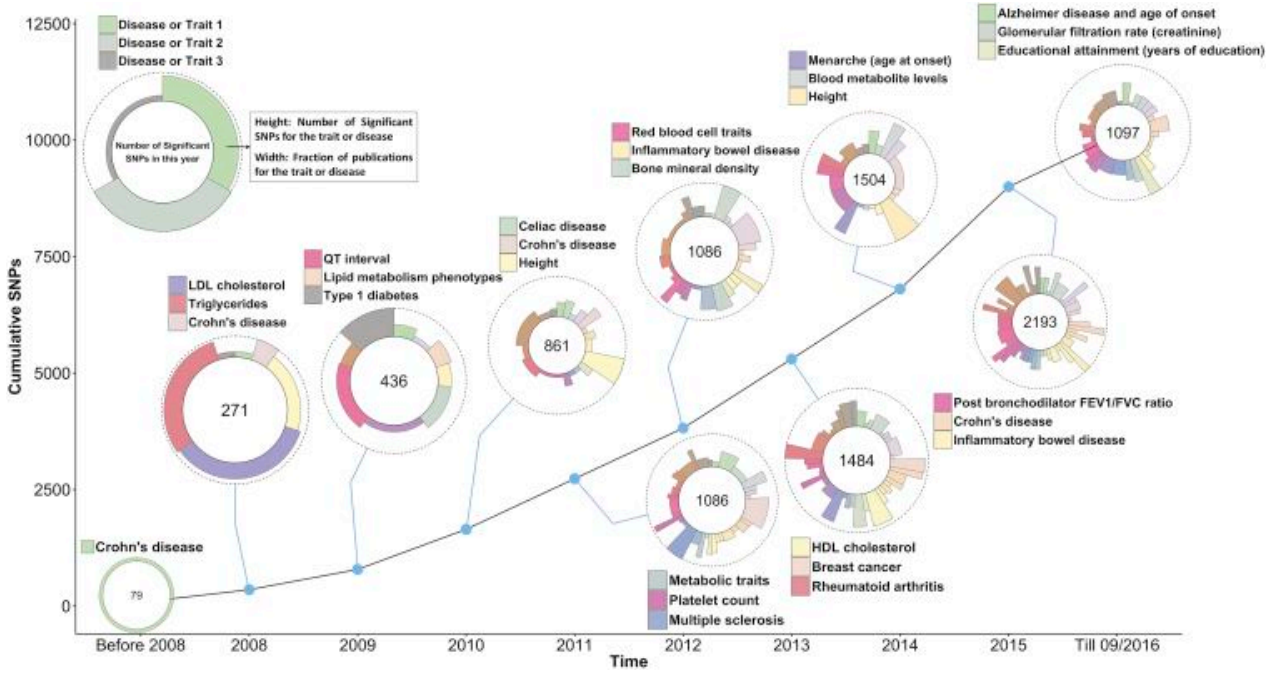

Figure 1. GWAS SNP-Trait Discovery Timeline, by (Visscher et al., 2017)

Despite GWAS having allowed the detection of several loci associated with complex traits, results with a p-value less stringent coming out from GWAS can fail to replicate; this could be affected by different reasons such as missing genotypes, genetic heterogeneity, unexpected LD, minor effects size, low allele frequency, or complex genetic architectures (Korte & Farlow, 2013).

While GWAS had detected hundreds of associated genetic variants, it can only explain a small proportion of the phenotypic variance attributed to additive genetic factors (Hofuku et al., 2013). Also, GWAS results can be biased in detecting variants due to the impact of allele frequencies on the strength of statistical association between alleles and traits. Additionally, the level of LD can affect the GWAS results, as a rare variant in low LD with a common variant will have fewer probabilities of being detected in the analysis (Visscher et al., 2012). Thus, many disease-causing genes could be missed even at high density in a genome-wide random SNP approach.

The markers present on the SNPs platforms have been selected to be common in the genome. Therefore, allowing the detection of causal variants common in the population but has problems detecting rare variants (Visscher et al., 2012). Similarly, the effect sizes for the GWAS associations are generally relatively modest (Witte, 2010), which indicates the need to use methods more robust, which could deal with high dimensional data and minimal frequencies.

Case-control is the most common and most straightforward study design for GWAS. However, this design assumes that any difference in the allele frequencies on the studied dataset relates directly to the measured phenotype (Cardon & Palmer, 2003). Thus, measures for confounders control must be implemented.

2.1.2 GWAS QUALITY CONTROL

Quality control (QC) procedures are essential criteria to be considered in GWAS. This step is carried out to remove low-quality samples or markers and reduce the spurious associations in later analysis. However, it can be computationally intensive, technically challenging, and

constant development (M. H. Wang et al., 2018). Among the criteria, the QC addresses in GWAS are: examining for potential sample identity problems, samples' genotyping efficiency, genotypes' call rate, minor allele frequency filtering for rare variants, testing for Hardy-Weinberg Equilibrium (HWE), because departure from this equilibrium can be indicative of potential genotyping errors, population stratification, and can lead to false associations, and finally batch effect can also be considered (Turner et al., 2011).

2.1.3 ASSOCIATION WITH A SINGLE MARKER (UNIVARIATE ANALYSIS)

The most common approach for testing the association between a genetic variant and a phenotype is a single-locus test. This strategy assumes that in a random mating population with no population structure, the association between a marker and a trait can be tested with a single marker regression (Hayes, 2013).

For a single marker test association, different genetic models can be fitted; a dominant model indicates that one specific allele will increase the risk of disease in equal amount for the homozygous most frequent and the heterozygous compared to the baseline risk for the homozygous less frequent, i.e., For allele "A" will translate the genotypes (AA, Aa, aa) to (1, 1, 0). An additive or co-dominant model will indicate that each additional copy of the "disease" allele will increase the disease risk, i.e., For allele "A" will code (AA, Aa, aa) as (2, 1, 0). A recessive genetic model will mean that two copies of the "disease" allele will be required to express the phenotypic characteristic related to this allele, i.e., For allele "A" will code (AA, Aa, aa) as (1, 0, 0). The most common genetic model used in GWAS analysis is the additive or co-dominant model (M. H. Wang et al., 2018).

The null hypothesis under the univariate GWAS test establishes no association between the genotype and phenotype, while the alternative hypothesis is that the marker affects the trait. For binary traits, this is commonly addressed by a Chi-squared test on 1° of freedom, odds ratio test, Fisher's Exact Test, and Armitage's trend test. However, when the phenotype is a continuous variable, ANOVA or t-test can be used. Also, when testing a single locus, simple linear regression and logistic regression would give an identical result to the tests mentioned above, as all of them are regression models with one predictor variable (M. H. Wang et al., 2018).

2.1.4 MULTIPLE TESTING

As the number of variables in a univariate analysis increases, the probability of detecting false positives increments; in GWAS, thousands of tests are conducted, each with its false positive probability. Therefore, the probability of finding more false positives over the entire GWAS analysis is much higher; this is referred to as multiple testing (Bush & Moore, 2012).

The most common and simplest method to correct for multiple testing is the Bonferroni correction. The Bonferroni correction adjusts the alpha value from $\alpha=0.05$ to $\alpha=(0.05/k)$, where k is the number of hypothesis tests conducted. Another method used to correct multiple testing is determining the false discovery rate (FDR). The FDR (Benjamini Hochberg) method sorts and ranks the P-values and multiply each P-value by the total number of the hypothesis tested, to finally divide the value by its assigned rank, obtaining the adjusted P-values (Hochberg & Benjamini, 1990).

Multiple testing can also be managed with permutation testing. The phenotype status of the individuals is randomly permuted, with the maximum test statistic calculated for the original

status and each permuted status. Then, based on comparing the test statistics, the p-values are adjusted (Besag & Clifford, 1991).

2.1.5 POPULATION STRATIFICATION

Population stratification (PS) refers to the state where populations are distinguishable by specific genotypes caused by differences in allele frequencies through the genome. This state is caused by colonization, migration, and random mating. PS can be a confounder within an association study by highlighting false associations between a genotype and the trait of interest (Hellwege et al., 2017). Several methods have been proposed to deal with PS; one of them is the genomic control method, which estimates a genomic (variance) inflation factor, given by equation 1.

$$\hat{\lambda} = median(X_1^2, X_2^2 \dots X_p^2)/0.456$$

Equation 1. Genomic inflation factor.

Where X^2 is a chi-squared distributed statistic calculated from the genome-wide scan of p SNPs, the test statistics are adjusted for the genomic inflation as indicated in equation 2. A $\hat{\lambda}$ value around 1 will be considered a measure of no population structure (M. H. Wang et al., 2018).

$$Y^2 = X^2/\hat{\lambda}$$

Equation 2. Test statistics adjustment.

Another approach to managing population structure is adjusting the individual genotypes and phenotypes through linear regression on Principal Component Analysis (PCA). This

approach consists in first, apply PCA to genotype data to infer the genetic variation. Second, the genotypes and phenotypes are adjusted by amounts attributable to ancestry, and third, the association statistics using ancestry-adjusted genotypes and phenotypes are computed (Price et al., 2006).

2.1.6 POWER OF GWAS

The statistical power refers to the probability of detecting an effect, meaning that the probability of making a type II error (false negative) will be low. This parameter depends on the effect size and the study sample size. In GWAS, the power turns on: a) The correlation between the marker and the causal variant, b) the proportion of total phenotypic variance explained by the genetic variant, c) the sample size, d) the disease prevalence, e) the genetic architecture, f) the genotyping array and haplotype reference panel used for imputation, g) the allele frequencies and h) the significance level threshold set for the study (Ferreira, 2018; Hayes, 2013; Visscher et al., 2017).

2.2 POLYGENIC RISK SCORE

Many complex traits are highly polygenic, implying that multiple causal variants contribute simultaneously to genetic susceptibility. Thus, examining "genetic scores" rather than individual SNPs may lead to better findings when studying the genetic contributions for complex traits (Boyle et al., 2017; Levine et al., 2009). A polygenic score, commonly called Polygenic Risk Score (PRS) on biomedical analysis, is an estimation based on the variation in multiple genetic loci. This approach assumes that phenotypic variation can be explained by the ensemble of markers (Dudbridge, 2013). A PRS is usually calculated as a weighted sum of the number of risk alleles carried by an individual. The risk alleles and weights are

defined by the loci and their measured effects as detected by GWAS. In some instances, a lower threshold than genome-wide statistical significance may be used to improve the phenotypic variance explained by the model, which allows incorporating variants that are not perfectly correlated with the causal genetic factors (Torkamani et al., 2018).

Some alternatives have been applied to PRS estimation. The primary approach called PRS P+T performs an LD-clumping to the variants to remove variants in LD and to keep the most significant for each clump and then estimates the score based on the GWAS summary statistics (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015). And, the PRS unadjusted, where all the markers, without LD clumping, are used to compute the PRS (i.e., the sum of all genetic markers across the genome, weighted by their marginal effect size estimates) (Ge et al., 2019).

The PRS relies on allele frequencies, varying across populations (Reisberg et al., 2017). Yet, PRS commonly fails when applied to different populations to the discovery set. However, polygenic analyses can be robust despite their disadvantages while still including many non-significant markers (Dudbridge, 2013).

2.3 MULTIVARIATE GWAS ANALYSIS

Non-linear effects that control variation in phenotypes can be caused by interactions (Epistasis), either between SNPs, genes, or quantitative trait loci (QTLs) (H. Zhang et al., 2017). Thus, the future impact of GWAS relies not only on the identification of the association signals against a specific trait but on the proper identification of gene-gene interactions effects on complex traits (McCarthy et al., 2008).

Machine learning approaches had been proposed as a mechanism to capture the cumulative effect of variants and their overall contribution to the outcome in disease prediction (Behravan et al., 2018; Curbelo Montañez et al., 2018). They promise complements to standard SNP analysis for understanding the overall genetic architecture of complex traits (Szymczak et al., 2009).

Multivariate methods had allowed the identification of complex additive effects on specific loci. However, the task of identifying valid combinations of genetic variants by multivariate search strategies can be highly computationally intensive due to the high number of models to be explored (Malovini et al., 2016).

2.3.1 FEATURE SELECTION ALGORITHMS

A feature selection (FS) algorithm is a computational solution driven by the relevance of the features and which can yield a weighted order of features (Molina et al., 2002). An FS method consists of four steps (**Figure 2**), 1) subset generation, where a candidate feature subset will be chosen, 2) subset evaluation, where the feature subset is evaluated according to an evaluation criterion, 3) stopping criterion, which is reached after founding the subset that best fits the evaluation criterion and 4) result validation, where the subset is validated using a validation set (Tang et al., 2014).

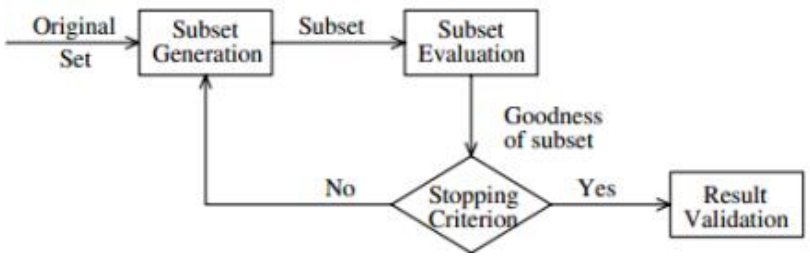

Figure 2. A General Framework of Feature Selection for Classification (Oreski & Novosel, 2014).

There are two basic feature selection techniques: filter and wrapper techniques. Filter methods **(Figure 3)** are generally used as a preprocessing step. Features are commonly selected based on correlations with the outcome variable. Wrapper methods **(Figure 4)** require a predetermined data mining algorithm in feature selection; it uses a subset of features and trains a model using them (Oreski & Novosel, 2014).

Figure 3. A General Framework of Filter Method for Feature Selection

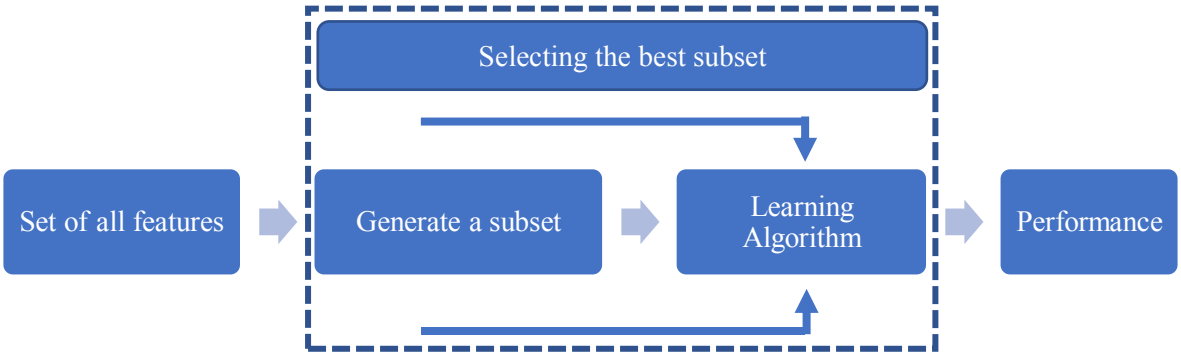

Figure 4. A General Framework of Wrapper Method for Feature Selection

Most of the existing feature selection approaches, for big data applications, focus on univariate analysis to screen features based on the estimated "individual" effects on the trait of interest, which is the case of GWAS. But its underlying mechanisms are neither static nor linear for many complex traits. Thus, identifying interaction effects among variables will help obtain more accurate results for phenotype prediction and reveal functional interactions (Xu et al., 2018).

2.3.2 MULTIVARIATE METHODS APPLIED TO GWAS ANALYSIS

Multivariate feature selection methods allow researchers to identify a subset or a combination of genetic variants that underlies the risk of developing a phenotype (Malovini et al., 2016). They can be categorized as regression-based (Dinu et al., 2012), LD-based (Paré et al., 2017), Bayesian approaches (Y. Zhang, 2012), machine learning methods (Behravan et al., 2018; X. Wang et al., 2016) and a combination of machine learning and statistical approaches (Mieth et al., 2016)

In 2014, an implementation of the correlation learning method Sure Independence Screening (SIS) was applied, as a genome-wide interaction analysis, to screen the most associated SNP-SNP interactions affecting regional brain volumes (Hibar et al., 2015). This method performs a dimensionality reduction as it generates a subset of SNPs n/ln(n), based on the correlation between the SNP and trait.

Another attempt to identify two-marker interaction effects was applied to amyloid imaging phenotype, using the Alzheimer's Disease Neuroimaging Initiative data. They selected the top significant 10,000 SNPs from univariate analysis to fit into a subsequent 2-marker interaction analysis (Li et al., 2015). This method was able to test up to two SNP combinations.

The algorithm COMBI, published in 2016, consists of a two-step process combining machine learning and statistical testing. First, a support vector machine (SVM) is trained to determine a subset of candidate SNPs, and then it performs hypothesis tests for these SNPs and an adequate threshold correction (Mieth et al., 2016).

Genetic algorithms (GAs) have also been applied to GWAS, and they have proved to be promising to detect multi-locus associations. A study published in 2014 used a GA to discover groups of SNPs (of size 2, 3, or 4) jointly associated with bipolar disorder. They found that their algorithm could realize significant multi-locus associations, even among SNPs that were not strongly associated with the disease (Mooney et al., 2011).

Support Vector Regression with Pearson Universal kernel (SVR PUK), as a fitness function, has also been applied for GWAS data. This method selected the most relevant SNPs for a continuous variable. Also, it generated groups of markers, considering all features with significant effects on the phenotype, and allowed the entry of less significant markers. Then it evaluated the performance of the group of SNPs and found that this method increased the Pearson correlation for the models and reduced the number of SNPs used to make the predictions (de Oliveira et al., 2014).

Random Forest has been shown to perform better than univariate tests in real GWAS data, but the probability of detecting interacting SNPs drops as the total number of SNPs increases. A proposed method based on the Random Forest algorithm, Trees inside Trees (T-Trees), considers the correlation structure among the genetic markers implied by linkage disequilibrium in GWAS data. This method showed significant improvement in terms of predictive power. However, this method is sensitive to rare variants and markers deviating from HWE (Botta et al., 2014).

A Bayesian genomic risk prediction method known as LDpred, proposed in 2015, infers each marker's posterior mean effect size by integrating LD information from an external reference panel (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015). This method has

improved the prediction of schizophrenia and multiple sclerosis (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015).

A Finnish study in 2018 designed a machine learning approach to identify interactions among SNPs contributing to breast cancer risk. They applied a gradient tree boosting method followed by an adaptive iterative SNP search and used a support vector machine (SVM) as the classification method. Their approach performed better than the results obtained using a PRS model (Behravan et al., 2018). This method was later applied to a second dataset and added demographic factors (Behravan et al., 2020).

Another study published in 2019 applied random forest to evaluate the risk of individual susceptibility to asthma using SNPs with a p-value $<1e^{-3}$ from a statistical association analysis. K-nearest neighbor (kNN) and SVM algorithms were trained to classify the individuals according to their susceptibility to asthma, showing that the occurrence of a multifactorial disease such as asthma can be predicted with RF-SVM (Gaudillo et al., 2019).

In 2021, gene-gene or gene-environment interactions impacting Drug-induced liver injury susceptibility were studied. For this, Multivariate Adaptive Regression Splines (MARS) and Multifactor Dimensionality Reduction (MDR) were applied to SNP data, and a decision tree model was successfully used to predict Drug-induced liver injury (Moore et al., 2021).

2.4 CROHN'S DISEASE: GENETICS AND PREDICTION

Crohn's Disease (CD) is a type of Inflammatory Bowel Disease (IBD) that affects the gastrointestinal tract with diverse symptoms depending on disease severity. These symptoms involve abdominal pain, fever, diarrhea, and bleeding (Baumgart & Sandborn, 2012). CD incidence and prevalence in developing countries is considered high; it has a reported incidence in North America of 6.3 to 23.8 per 100,000 (Ng et al., 2017).

As with other complex traits, CD incidence has been theorized to be related to environmental factors, including H. pylori exposure, occupation, microbiota, diet, lifestyle, medications, pollution, and genetic factors (Feuerstein & Cheifetz, 2017) (Liu & Anderson, 2014). The population within industrial urbanized societies is the most affected by this inflammatory bowel condition attributed, mainly, to a westernized lifestyle (Koloski et al., 2008; M'Koma, 2013).

At present, over 140 genetic loci associated with CD have been discovered by Genome-Wide Association Studies (GWAS) (de Lange et al., 2017; Liu et al., 2015).

The heritability of liability for CD, calculated from GWAS, has been estimated to be 0.37, which contrasts against the estimated pooled twin data (h=0.67) (Gordon et al., 2015). NOD2, IL23R, and ATG16L1 are among the well-known CD-risk genes identified. These genes are involved in inflammation and the immune system's response (Gajendran et al., 2018).

Heritability for Crohn's disease has been estimated, from pooled twin data, to be around 0.75, which contrasts GWAS's heritability estimate, which is about 0.37 (Gordon et al., 2015).

Identifying genetic variants or susceptibility genes for CD has allowed the development of more efficient and disease-directed drugs (Grenier & Hu, 2019), thus evidencing the importance of investigating CD risk genes.

Current CD models, including univariate-based and multivariate models, have a broad spectrum of performance, which can vary from 0.59 to 0.84 (AUROC, area under the receiving operating characteristic curve) (Kooperberg et al., 2010; Mittag et al., 2015; Newcombe et al., 2019; Romagnoni et al., 2019; Wei et al., 2013). The variations reflect the complexity of the disease and the dependencies of datasets and methodologies. The methods that have been tested are LDpred (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015), LASSO (Kooperberg et al., 2010; Newcombe et al., 2019; Wei et al., 2013), gradient boosting (Romagnoni et al., 2019), support vector machines (SVM) (Mittag et al., 2015), k-nearest-neighbors (KNN) (Mittag et al., 2015), multi-layer perceptron (MLP) (Mittag et al., 2015), Bayesian methods (G.-B. Chen et al., 2017) and random forest (Mittag et al., 2015). However, the performance of these risk prediction methods has not been tested on the same CD dataset under similar, robust, and stringent methodological conditions. In addition, the de Lange et al. dataset (de Lange et al., 2017) has not been assessed for risk prediction models. **Table 1** shows the studies where CD has been studied with different machine learning approaches, with their respective performances and models information.

Table 1. Crohn's Disease prediction studies: Multivariate methods.

Publication	Dataset	# of Samples	Method	Impute	Threshold for pre-selection	SNPs / Pre-selected SNPs	ROC AUC
Wei et al., 2013 (Wei et al., 2013)	IIBDGC' Immunochip project	~17,000 CD ~22,000 HC	LASSO	No	1e-4	573/10,799	0.86
Romagnoni et al., 2019 (Romagnoni et al., 2019)	IIBDGC	18,227 CD 34,050 HC	LR	Only NAs	1e-4	2,575/21,896	~0.80
			GBT				
			ANN				
G. B. Chen et al., 2017 (G. B. Chen et al., 2017)	IIBDGC	16,850 CD 27,050 HC	BayesR inmunochip	No	?	?/42,534	0.75
Wang et al., 2019 (Y. Wang et al., 2019)	PopGen Biobank WTCCC panel GTEx panel (phs000424)	115 CD and 62 HC 2,678 CD 635 HC	AVA,Dx	No	DKMcost	125 genes / 13,957 genes / 173,013 variants	0.75
Song et al., 2020 (Shuang Song, Wei Jiang, Lin Hou, 2020)	IIBDGC	6,333 CD 15,056 HC	EB-PRS	No	?	?/871,743	0.69
			PRS		?		0.63
	WTCCC	1,689 CD 2,891 HC	PRS P+T		1e-7		0.68
			Ldpred-inf		?		0.62
			Ldpred		?		0.66
			So's		?		0.69
			Mak's		?		0.68
Newcombe et al., 2019 (Newcombe et al., 2019)	WTCCC	1,684 CD 2,836 HC	LASSOSum	No	?	255,781	0.65
			Ldpred				0.69
			JAM				0.69
Vilhjálmsson et al, 2015 (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015)	WTCCC	1,687 CD 2,867 HC	PRS	No	All	376,901	0.62
			PRS P+T		1e-4	? / 376,901	0.63
			Ldpred-inf		CF = 0.01	~3,769 / 376,901	0.63
			Ldpred		CF = 0.01		0.67
Kooperberg et al., 2010 (Kooperberg et al., 2010)	WTCCC	2,000 CD 3,000 HC	LASSO	Only NAs	3000 top SNPs	33/100	0.64
	NIDDK	792 CD 932 CD					
Mittag et al., 2015 (Mittag et al., 2015)	WTCCC	2,000 CD 1,500 HC	SVM, KNN,RF, MLP	No	1e-3	??/1,560	~0.59

IIBDGC International Inflammatory Bowel Disease genetic consortium, WTCCC Wellcome Trust Case Control Consortium, NIDDK National Institute of Diabetes and Digestive and Kidney diseases, HC healthy controls, LR Logistic regression, GBT Gradient Boost Trees, ANN Artificial Neural Networks.

2.5 LIMITATIONS AND PROBLEMS

Currently, the existing implementations of machine learning methods pose several limitations for application to genome-wide data. Further research is needed to identify and evaluate variable selection procedures suited for genetic data (Szymczak et al., 2009). The study of epistatic interactions at the whole genome level has been limited due to the complexity of conducting pairwise statistical tests. The difficulty of testing interactions is caused by computational complexity issues, selection of multiple testing thresholds, and LD (M. H. Wang et al., 2018).

Different multivariate methods have been successfully applied to various genetic datasets; however, they can be affected by computational complexity issues in GWAS (McKinney et al., 2006). The next chapter explains a solution approach for detecting SNP interaction based on feature selection by a multivariate algorithm.

For GWAS, the computational complexity can be estimated as O(np), where n refers to the input size and p to the number of variants (SNPs). Whereas for multivariate analysis accounting for pairwise comparisons due to the interaction between the variants, the computational complexity would be as high as $O(n^2p)$.

3. CHAPTER 3: SOLUTION MODEL AND METHODOLOGY

3.1 SOLUTION MODEL

This research aimed to solve common problems for GWAS wherein a genome-wide random SNP approach many disease-causing genes are missed. This, caused by an incomplete or null LD among the markers with the causal variants (Ferns et al., 1986). The SNPs replication is also an issue in these studies, which could be affected by missing genotypes, genetic heterogeneity, unexpected LD, minor effects size, low allele frequency, or even complex genetic architectures (Korte & Farlow, 2013). The typical approach to conducting GWAS (univariate analysis) has allowed the discovery of large effect common variants. Still, it has limitations to deal with the advent of big data and more efficient genotyping technologies. Therefore, more efficient methods are needed to analyze multi-locus interactions for GWAS.

Machine learning approaches are considered good complements to standard SNP analysis methods for understanding the overall genetic architecture of complex traits (Szymczak et al., 2009). However, the difficulty of testing interactions is caused by computational complexity issues, selection of multiple testing thresholds, and LD issues (Wang et al., 2018). Some works have been performed to detect SNP-SNP interactions. However, algorithmic development is still ongoing due to its mathematical and computational complexities. Nevertheless, it is essential to consider that the performance of the established computational or statistical methods will vary depending on the type of data to be analyzed. Thus, robust strategies for SNP association identification involving interactions among locus are needed.

The number of combinations can become computationally challenging when many-variable interactions need to be contemplated, as is the case of a GWAS. Thus, this work considered the adaptation of multivariate machine learning algorithms to deal with genome-wide data through techniques of features reduction. Imputation of SNPs was evaluated to investigate the effect of adding information for non-genotyped markers to the prediction performance. And, dealing with the independence of data reflected by the LD among the SNPs was explored with and without LD clumping. The data for this research is "publicly" available genotypic and phenotypic data for CD. This data was requested by the standard methods on their respective database, i.e., NCBI dbGaP or EGA EMBL-EBI. The dataset used for the multivariate analysis has not been used for prediction analysis for either univariate-based or multivariate analysis.

The solution model proposed for this research (**Figure 5**) consisted in replicating the GWAS published by de Lange et al. in 2017 (de Lange et al., 2017) to validate the subsequent methodology. This consisted in adapting multivariate feature selection algorithms such as LDpred, Random Forest, XGBoost, BSWiMS and LASSO, and the common univariate-based PRS approaches such as PRS P+T and PRS unadjusted. The adaptations consisted of adding information through SNPs imputation, filtering strategies, such as filtering by a threshold or causal fraction of some variants for imputed and non-imputed data and reducing the number of markers by evaluating its independence through LD clumping. Filtering was necessary to reduce the computational complexity issues. Univariate-based analyses were attempted to compare the common polygenic score strategy with the proposed multivariate machine learning approach. And, finally the validation of the results with its biological interpretation.

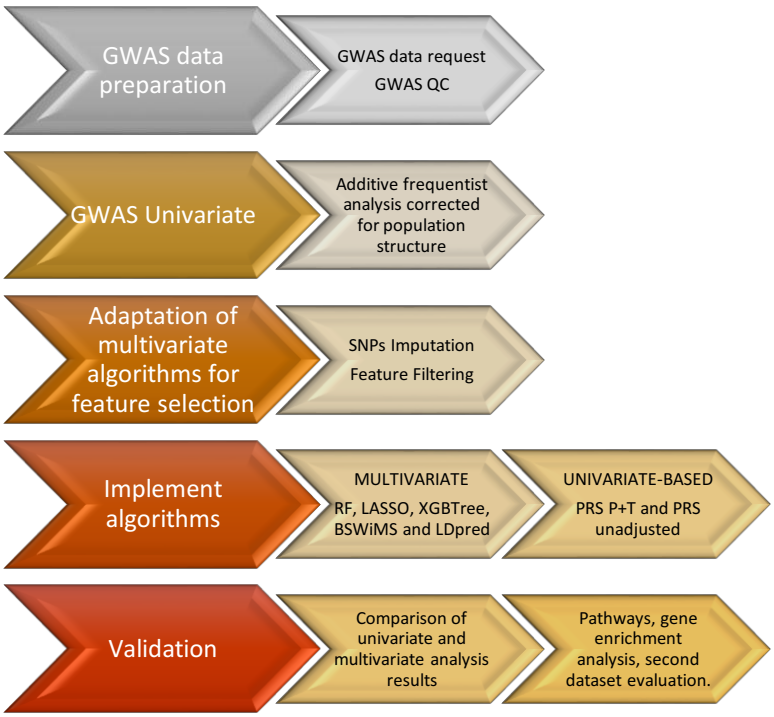

Figure 5. General Framework of the solution model for Feature Selection from Biological BigData: Identification of Significant Associations Applying Multivariate Machine Learning Algorithms to Genome-Wide Association Studies (GWAS)

3.2METHODOLOGY

Figure 6 presents a scheme of the overall methodology to compare the performances for CD-risk prediction. First, two datasets were selected, the raw and the imputed data, to assess the effect on prediction models of increasing features and complexity by SNP imputation. Second, to reduce dimensionality and facilitate further computational analysis, a univariate analysis was performed on 40% of the samples to select the top 5% of features maintaining the other model generation steps blind. Third, an LD clumping was applied to decrement the number of SNPs filtering highly correlated SNPs. Forth, robust 10-fold-cross-validation geometric mean p-values (*GMP*) were estimated from the 60% of remaining samples. Fifth, prediction models were built to select SNPs at diverse *GMP* thresholds and assessed in a 10-fold-cross-validation manner.

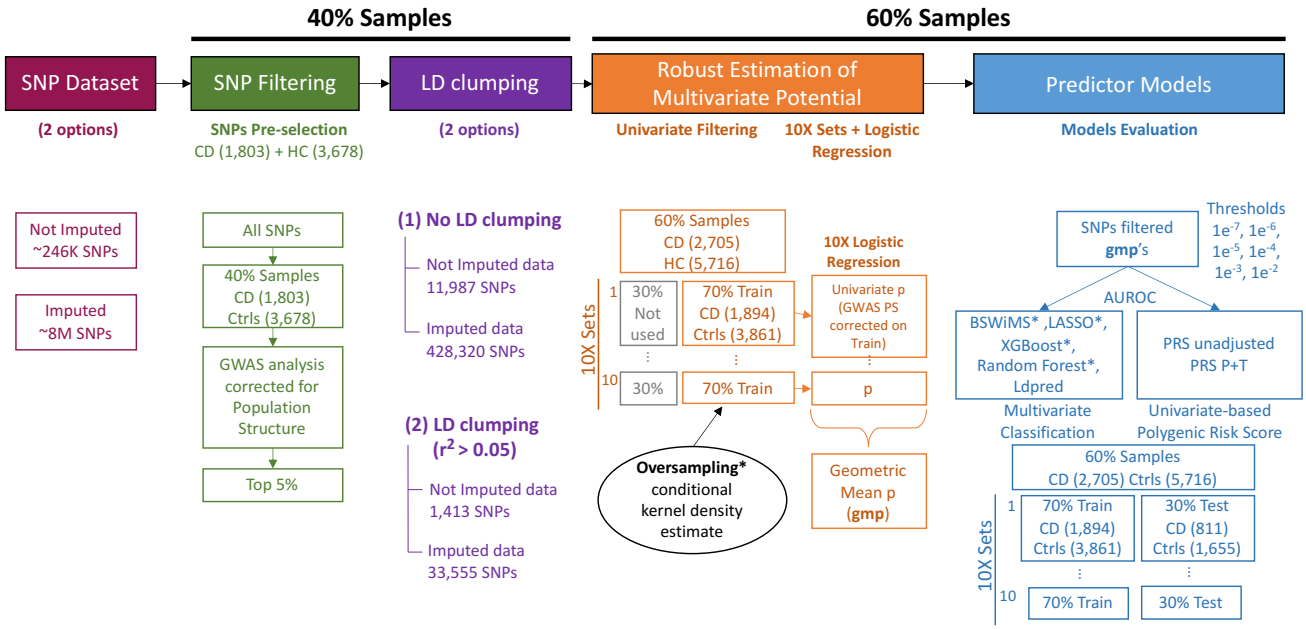

Figure 6. Approach designed to evaluate multivariate and univariate-based models to predict CD risk in the UKIBDGC and UK10K GWAS dataset.

3.2.1 GWAS DATASETS ACQUISITION AND QUALITY CONTROL

UKIBDGC and UK10K GWAS raw data was requested from the European Genome-phenome Archive (EGA) website under accession EGAS00001000924 (de Lange et al., 2017). The dataset contained 4,508 UK CD cases, diagnosed using accepted endoscopic, histopathological, and radiological criteria, and genotyped on the Human Core Exome v12.1. On the other hand, 9,944 population control samples genotyped on the Human Core Exome v12.0 were obtained from the Understanding Society Project under the accession number EGAC00001000205.

Quality control (QC) for genotypes and sample were conducted as implemented in the original published data (de Lange et al., 2017), which consisted of removing variants that were not present on both versions of the genotyping platforms, had missing values >5%, had a significant difference in call rate between cases and controls ($P < 10^{-5}$), deviated from Hardy–Weinberg equilibrium in controls ($P < 10^{-5}$) or were affected by a genotyping batch effect (significant association ($P < 10^{-5}$).

After this QC process, 246,735 variants remained. For samples, the criteria were to keep the samples that passed the QC in the original study (information provided in the dataset). Also, the top 10 principal components provided in the dataset were used to correct for population structure.

Data from the NIDDK IBD Genetics Consortium Crohn's Disease was obtained through dbGaP accession phs000130.v1.p1 (Duerr et al., 2006) and was used to evaluate the models in an external dataset for validation. The dataset contained 513 CD cases and 515 control samples from European ancestry. SNPs were excluded for call rates less than 90% and MAF less than 1%. After this QC process, 313,752 SNPs remained. Only SNPs overlapping with the UKIBDGC and UK10K GWAS dataset were used for analysis.

3.2.2 GENOTYPE IMPUTATION

Imputation was performed remotely using the Michigan Imputation Server (S et al., 2016). For this process, the European 1000 genomes were used as reference data. QC for genotypes consisted of removing variants with MAF <1%, INFO SCORE <0.4 (a measure of imputation quality), and deviated from Hardy–Weinberg equilibrium in controls ($P < 10^{-7}$).

3.2.3 UNIVARIATE ANALYSIS REPLICATION FOR ALL SAMPLES

An additive frequentist analysis corrected for population structure was implemented for all QC-pass GWAS data, which consisted of 9,194 healthy controls and 4,508 CD cases. An additive model, with the top 10 first principal components as covariates, was implemented as a logistic regression in R (R Core Team, 2021) for the non-imputed dataset and SNPTEST for the imputed datasets.

3.2.4 ESTIMATION OF MARKERS WITH DISEASE POTENTIAL

An additive analysis, including the initially reported top 10 PC as covariates, was performed using 40% of samples (1,803CD and 3678 HC) for the imputed (8,755,412 SNPs) and the original dataset (not imputed, 246,735 SNPs). The additive genetic model is represented by **equation 3**. Where β_0 refer to the intercept, β_n, to the effect size of the *nth* marker, X is the number of copies of the reference allele, and ε, the residual error.

$$\ln(odds) = \beta_0 + \beta_n X + \varepsilon$$

Equation 3. Additive Genetic Model

A preselection step using the 40% of samples was implemented to deal with the large number of variants yielded by the imputation process. Then, after removing duplicated data (by position), the top 5% of the top associated markers, 428,320 and 11,987 SNPs for the imputed

and not imputed data, respectively, were selected as potential markers and were used for the subsequent analysis.

3.2.5 LINKAGE DISEQUILIBRIUM CLUMPING

An LD clumping process was set (Fig. 1D), using *plink* (Purcell et al., 2007), removing all the SNPs with an r^2 >0.05 keeping the most significant markers of de clump set. After this process, 1,413 variants remained within the not imputed dataset and 33,555 within the imputed dataset.

3.2.6 ROBUST ESTIMATION OF MULTIVARIATE POTENTIAL

The remaining 60% of samples, was used to test the multivariate and univariate-based methods. The models were trained using 70% of the samples and tested on the remaining 30%. The experimental setting for this analysis consisted of performing 10 repeated random sub-sampling validation to correct for sampling, which is represented in **Figure 7**.

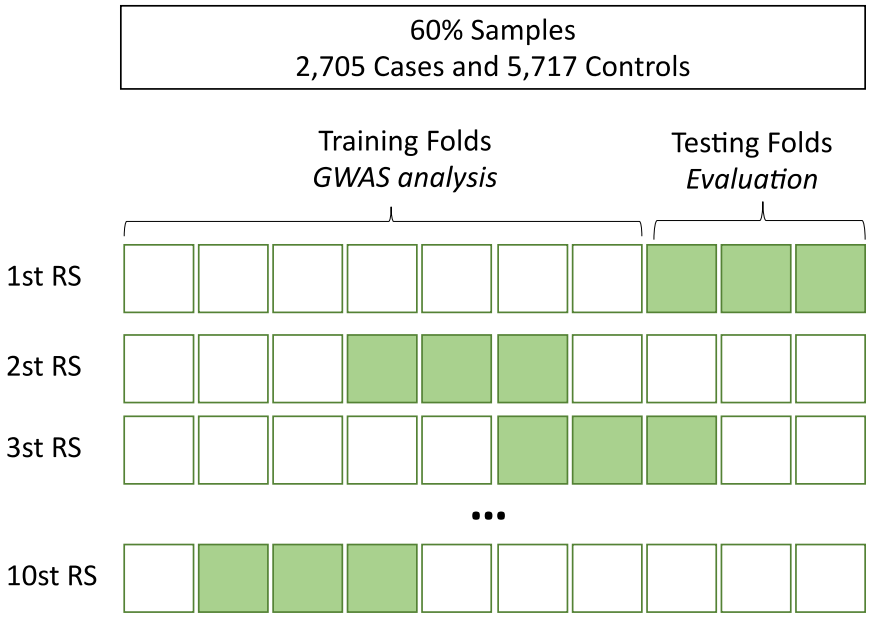

Figure 7. Random sampling' ten-fold cross-validation diagram, used to apply multivariate and univariate-based methods.

Then, a univariate GWAS analysis was performed for each RS (random sampling) in the training set, and those SNPs with a p-value < 0.01 were subsequently used. An additive analysis corrected for population structure was implemented for QC-pass GWAS data. An additive frequentist model was used as is the most common approach in univariate studies. The model was implemented as a logistic regression in R using the 10 first principal components as covariates.

3.2.7 PREDICTOR MODELS

The training set (42% of the data resulting from 70% of the 60%) and the SNPs filtered by p-value threshold were used to evaluate the multivariate and univariate-based methods. The test set (18% of the data resulting from 30% of the 60%) was finally used to assess the prediction performance for all the models. The performance average and standard deviation across the 10-folds sets were used (**Figure 8**).

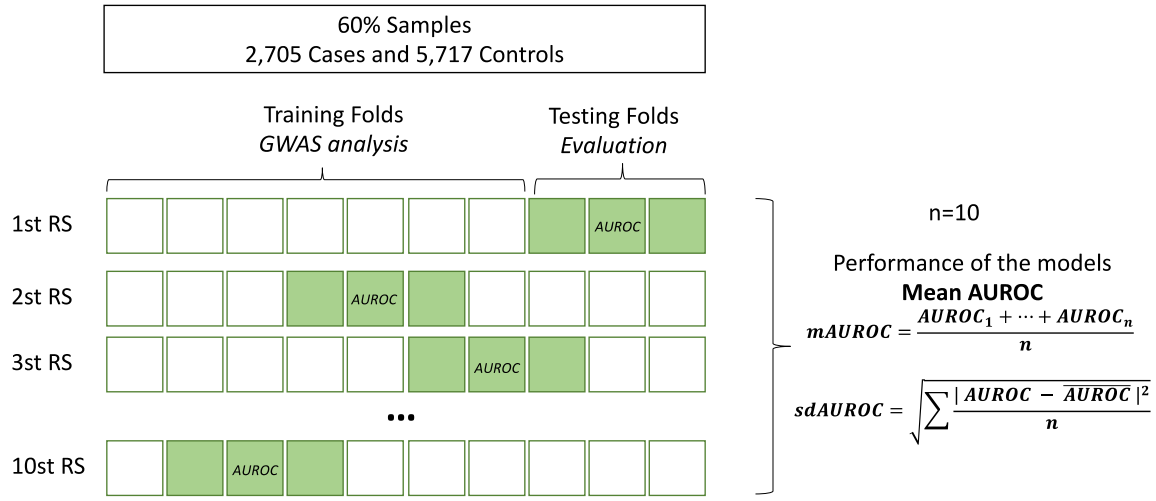

Figure 8. Overall methodology for evaluating the performance of the models for the CD dataset.

3.2.7.1 UNIVARIATE-BASED METHODS

A PRS was estimated with the log of the odds ratio of the effect sizes (used as weights), using various p-value thresholds from $1e^{-7}$ to $1e^{-2}$. This score is represented in **equation 4**.

$$PS_i = \beta_{1g1i} + \cdots + \beta_{mgsi}$$

Equation 4. PRS estimation

Where g_{mi}, is the number of effect alleles (0, 1, or 2) of SNP m for individual *i*, and βm denotes the allele "risk" effect of SNP *m* (Behravan et al., 2018; Torkamani et al., 2018). The PRS adjusted and unadjusted were estimated. PRS is considered unadjusted when all the markers within a determined threshold are selected without considering possible detrimental effects due to neighbor variants in LD (Li et al., 2015). Adjusted PRS was estimated using both *plink P+T* (Li et al., 2015). PRS P+T scores were generated by first clumping all the markers with an LD $r^2 > 0.05$ using *plink* v1.9 (Purcell et al., 2007), thus keeping only the independent variants.

3.2.7.2 MULTIVARIATE METHODS

Embedded methods, like LASSO, are characterized by including a feature selection process in the training stage of the model. LASSO (least absolute shrinkage and selection operator) is a regression analysis method (**Figure 9**) that enhances the prediction accuracy by performing variable selection and L1 regularization (Tibshirani, 1996).

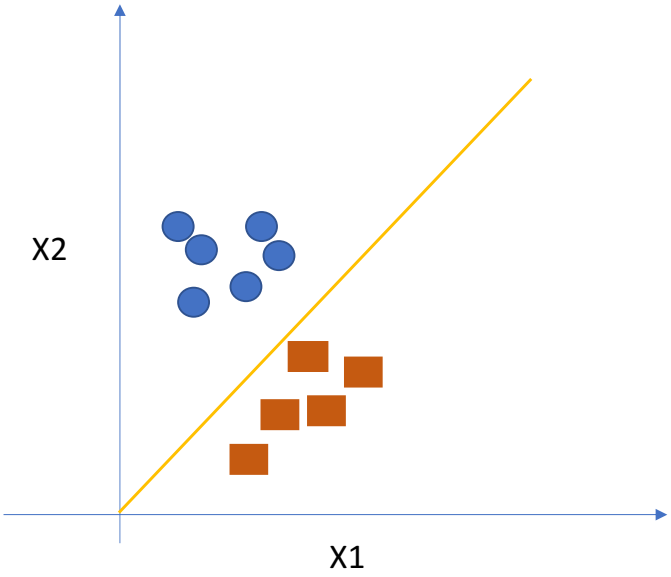

Figure 9. The schematization of a classification problem with a binary trait.

Due to, L1 regularization some coefficients can become zero and eliminate from the model. (Tibshirani, 1996). The regression algorithm minimizes the error (loss function) when training a model. Lasso (L1) regularization modifies the loss function as presented in **equation 5**. LASSO was implemented with the R package *glmnet* (Friedman et al., 2010)

$$L = \sum (y - \hat{y})^2 + \alpha \sum |m|$$

Equation 5. Lasso (L1) regularization, L = Sum of squared residuals + Penalty.

Random forest (RF) is a classification tree-based strategy that ranks the features by how well they improve the purity of the node (**Figure 10**). This method is an ensemble learning method for classification where the output of the random forest analysis is the class selected by most trees (Ho, 1998). The Gini Impurity of a node is a measurement of the likelihood of incorrect classification of a sample given a variable (Ho, 1998).

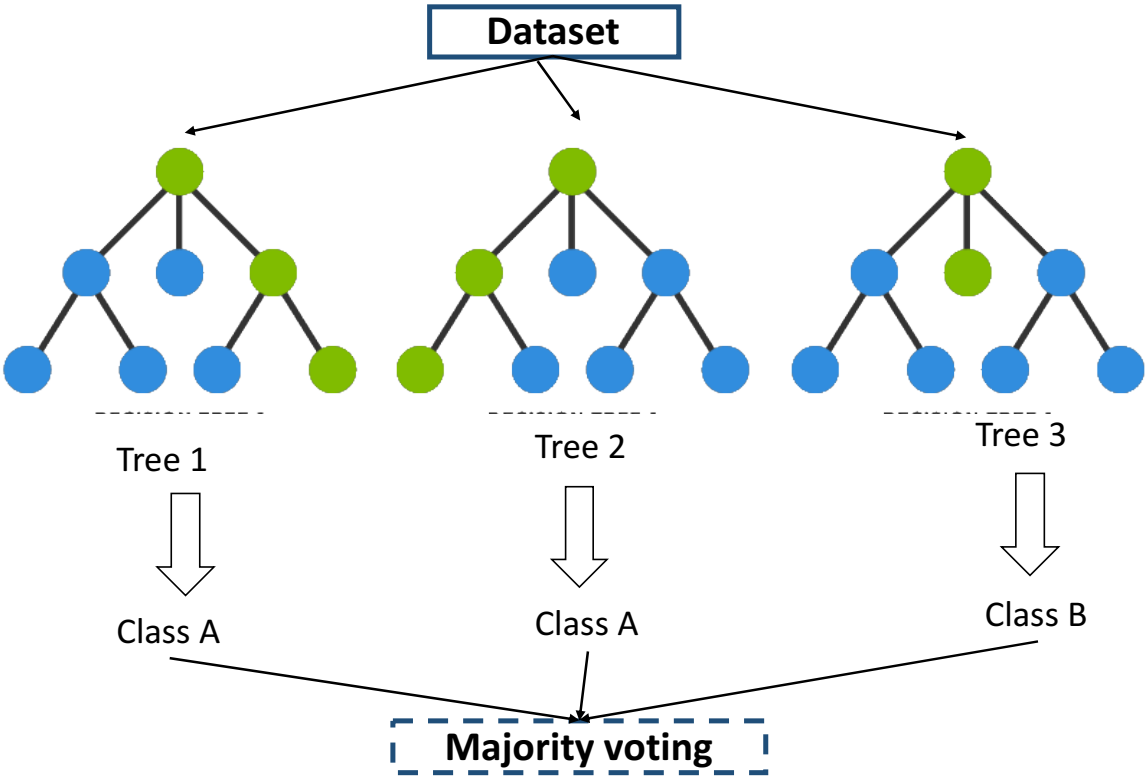

Figure 10. Random Forest representation for a binary class classification problem.

In a classification problem, the Gini impurity is represented by **equation 6**, where *fi* is the frequency of sample *i* at a node and *C* is the number of unique samples. RF was used from the R package *caret* (Max Kuhn, 2021) and *ordinalForest* (Roman Hornung, 2021).

$$G = \sum_{i=1}^{C} -f_i\,(1 - f_i)$$

Equation 6. Gini impurity to decide the best split for the RF problem

XGBoost provides a regularizing gradient boosting framework, which constructs boosted trees with a score indicating how useful or valuable each feature was in creating the boosted decision trees within the model (Tianqi Chen et al., 2021). In gradient boosting, the learning procedure consecutively fits new models to classify the variable (**Figure 11**); the idea is to

add new models to the ensemble sequentially (Natekin & Knoll, 2013). XGBoost was used from the R package *caret* (Max Kuhn, 2021) and *xgboost* (Tianqi Chen et al., 2021).

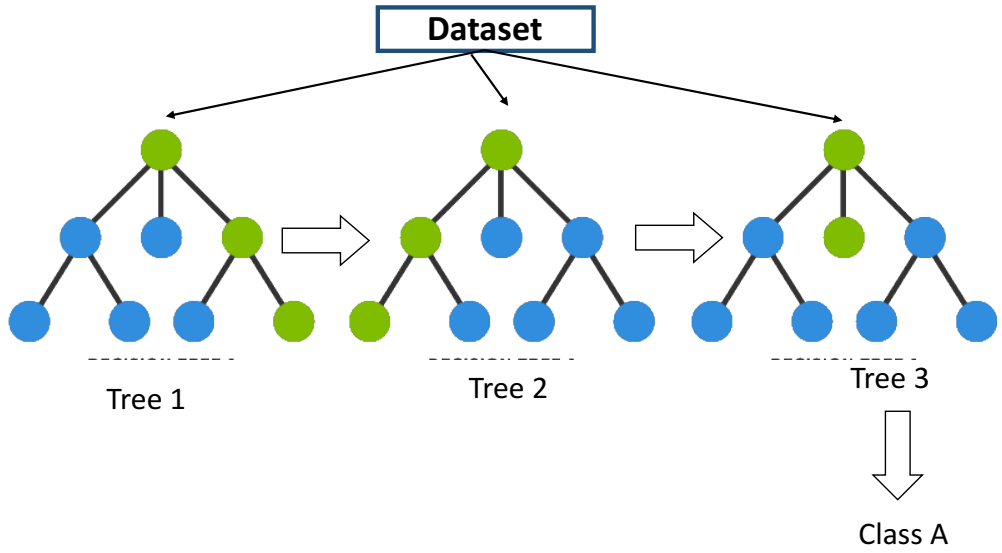

Figure 11. XGBoost, a gradient boosting framework, representation. Model built sequentially.

Bootstrap Stage Wise Model Selection (BSWIMS) is based on forwarding, and backward selection (**Figure 12**) coupled to logistic models, using the R package *FRESA.CAD* (A Martinez-Torteya et al., 2018). BSWIMS extracts those SNPs whose all terms are statistically significant, repeating the model generation until no significant SNPs are added.

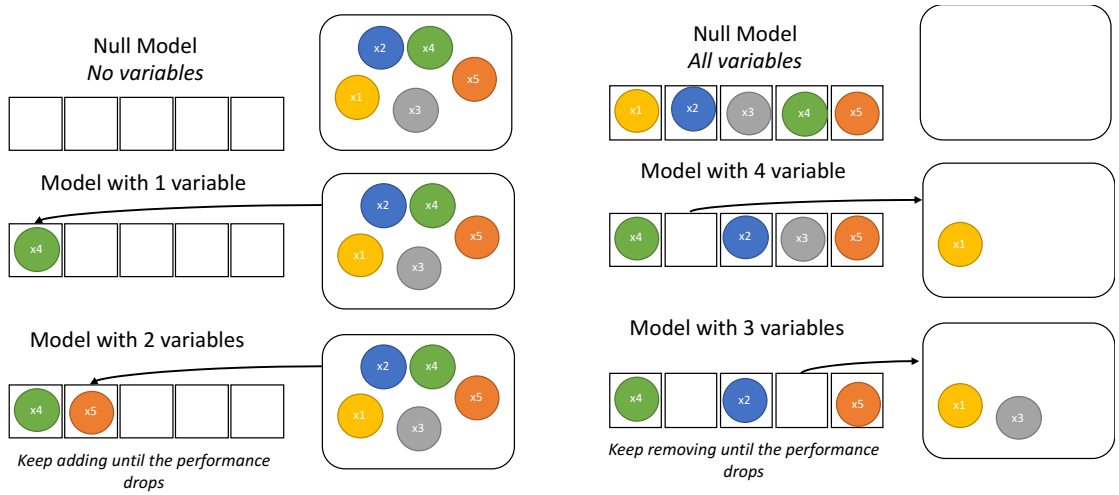

Figure 12. Schematization of forwarding and backwards selection

LDpred is a Python-based software package that adjusts GWAS summary statistics for the effects of linkage disequilibrium (LD) (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015). This method infers each marker's posterior mean effect size using prior effect sizes and LD information from an external reference panel (**Figure 13).** Ldpred applies an approximate MCMC (Marko Chain Monte Carlo) Gibbs sampler to infer the posterior mean. This approximate Gibbs sampler sample the update of the effect sizes from the posterior distribution (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015).

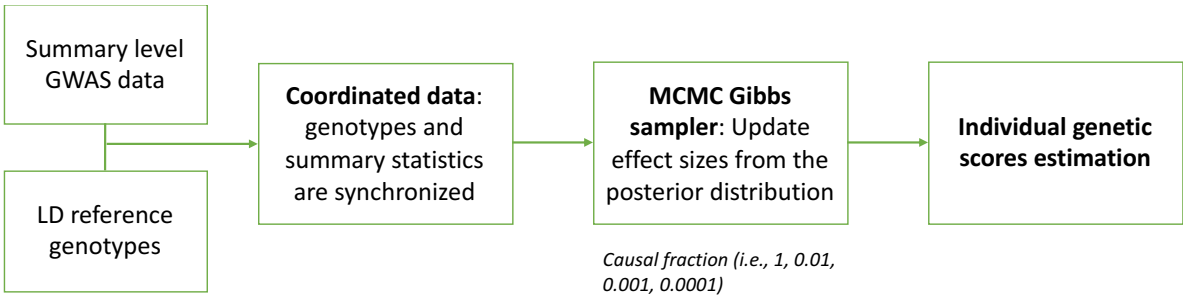

Figure 13. Diagram for Ldpred methodology, as established by (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015)

Assuming that distant markers are unlinked, Vilhjálmsson et al. mention that the posterior mean for the effect sizes within a small region *l* under an infinitesimal model are approximated by **equation 7**. Where D_l denotes the regional LD matrix within the region of LD, and $B\sim1$ represents the least-squares-estimated effects within that region. Ldpred can estimate the PRS by assigning a causal fraction for the data.

$$E(\beta_l||\tilde{\beta}_l,D)\approx(MNh^2_g I+D_l)^{-1}\tilde{\beta}_l.$$

Equation 7. LDpred: Bayesian Approach in the Presence of LD as described by (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015)

The *LDpred* method was also applied to the data, but instead of selecting the markers to build the PRS by p-value thresholding, this was performed by a causal fraction. Version 1.0.10,

available in https://github.com/bvilhjal/ldpred, was used. Due to the conditions for which *LDpred* has been developed (to relay in the LD among the markers), these analyses were only performed with the datasets without the LD clumping.

Most of the classification methods assume the dataset is evenly distributed and are optimized to perform better in those circumstances. Thus different methods to balance the data class distribution have been proposed (Ali et al., 2015). For this research, an oversampling step was integrated to balance the data classes for the multivariate methods, using the package ROSE (Lunardon et al., 2014). Random Oversampling involves supplementing the training data with additional data for the minority classes (Figure 14) (Ling & Li, 1998). This approach creates a sample of synthetic data for the minority class, where the new examples are drawn from a conditional kernel density estimate of the two classes. The unknown density function is estimated by averaging over a set of homogeneous kernel functions centered at each sample point and then generating new sample points based on the estimated density function (Lunardon et al., 2014).

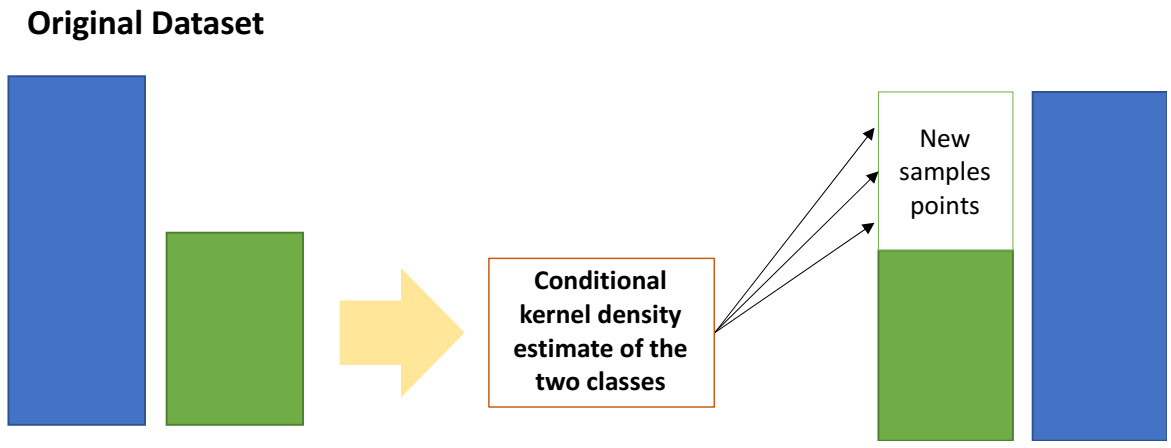

Figure 14. Schematization of random oversampling for a minority class

3.2.7.3 PERFORMANCE OF THE MODELS

The predicted probability of disease was evaluated on the test data for each model. Receiver operator characteristic curves (ROC) and the area under the ROC curve (AUROC) were used to assess the models. The AUC score reduces the ROC curve to a single measure performance metric. The AUROC is the probability that a random sample for a case is ranked more likely to be diseased than a random sample for control (Hanley & McNeil, 1982). This is given by **equations 8 to 10**.

$$TPR = \frac{TP}{TP + FN}$$

Equation 8. True Positive Rate estimation

$$FPR = \frac{FP}{FP + TN}$$

Equation 9. False Positive Rate estimation

$$AUC = \int TPR\, d(FPR)$$

Equation 10. AUC estimation

3.2.7.4 IMPORTANCE OF THE VARIANTS FOR MULTIVARIATE MODELS

For XGBoost and Random Forest, the models estimate variance importance for the SNPs, which are averaged scaled class-specific scores (Max Kuhn, 2021). The importance is a measure of the reduction in the statistic when each predictor's feature is added to the model. For LASSO, BSWiMS, and LDpred, the measure of variance importance is given by the size of the re-estimated effects (betas) of the variants (A Martinez-Torteya et al., 2018; Tibshirani, 1996; Vilhjálmsson, Yang, Finucane, Gusev, Price, et al., 2015).

The rank of importance for these methods was derived by ranking the absolute value of the re-estimated betas. For XGBoost and Random Forest, the rank was performed based on the accumulative importance, obtained by adding the mean importance of the variant and the number of CVs selected.

3.2.8 VALIDATION AND BIOLOGICAL INTERPRETATION

3.2.8.1 FUNCTIONAL ANALYSIS OF GENES AND VARIANTS

Functional analysis for genes associated with variants from the models was performed using *ENRICHR* and *DAVID's* bioinformatic tools (Huang et al., 2009), which conducts an over-representation test to determine if the selected genes have a non-random presence in a biological pathway. The criteria for clustering terms consisted of selecting those terms statistically significant (after Bonferroni correction for $p < 0.05$) and that involved a certain number of genes (≥10 for diseases, ≥ 4 for GO, and KEGG).

The pathways and gene ontology terms for molecular function and biological process were analyzed. Because many highly related terms were observed, the significant terms were grouped by similarity using hierarchical clustering separately for GO and KEGG and by groups of similar disorders or diseases. Groups were generated by averaging the presence of the gene among the diseases/terms merged in the group. HLA genes were removed in this analysis to simplify cluster generation of terms. Finally, the gene list from those reported in open targets was compared (Carvalho-Silva et al., 2019) to identify novel or unreported genes for CD.

3.2.8.2 PERMUTATION OF CD PHENOTYPES

The experimental setting of the prediction models was repeated, but adding a permutation of phenotypes (class: case or control), thus randomizing the samples. This was performed to evaluate the efficacy of the models to predict the CD status (**Figure 15**).

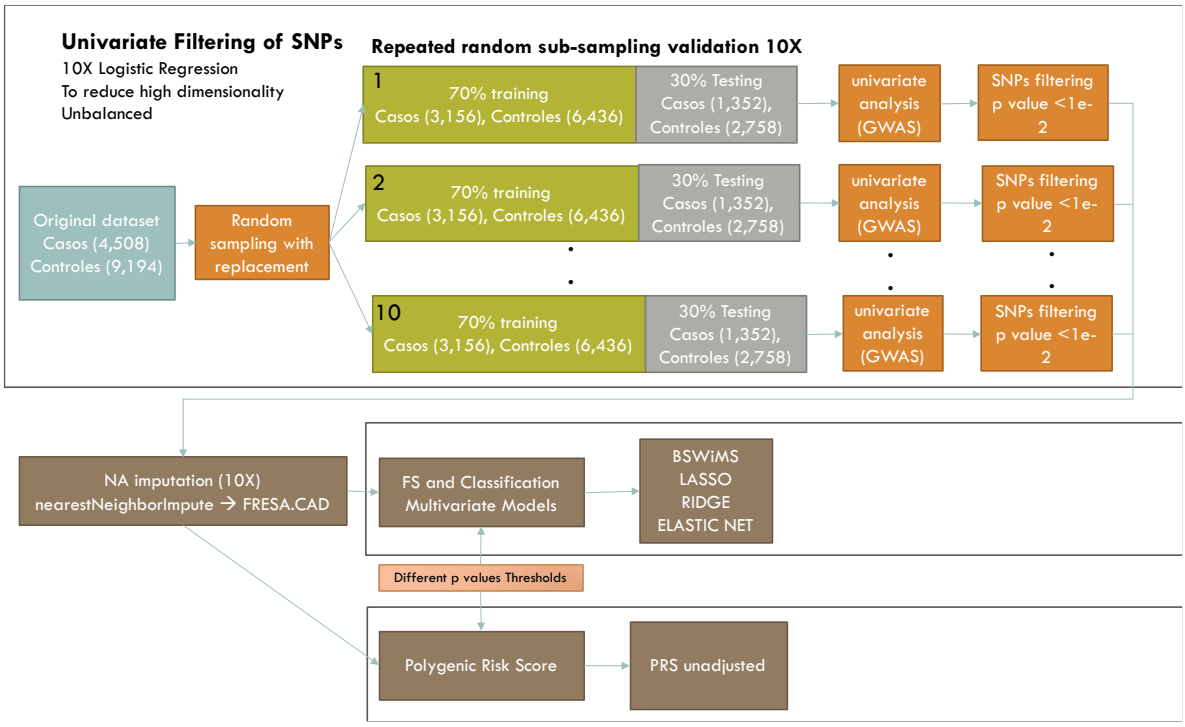

Figure 15. Workflow for comparing multivariate and univariate models with the permutation of samples phenotypes.

3.2.8.3 MISCLASSIFIED EXPERIMENT

An additional model building was conducted for the poorly or misclassified samples for LASSO and XGBoost, selecting misclassified samples from the not imputed data, with no LD clumping and a p-value < 1e-3 (**Figure 16**). For this, three different GWAS were implemented for each method: 1) misclassified CD cases vs. correctly classified CD cases, 2) misclassified healthy controls vs. correctly classified healthy controls, and 3) misclassified CD cases vs. misclassified healthy controls. The same methodology previously mentioned was applied, but with the SNPs identified on the 3 GWAS for the misclassified samples.

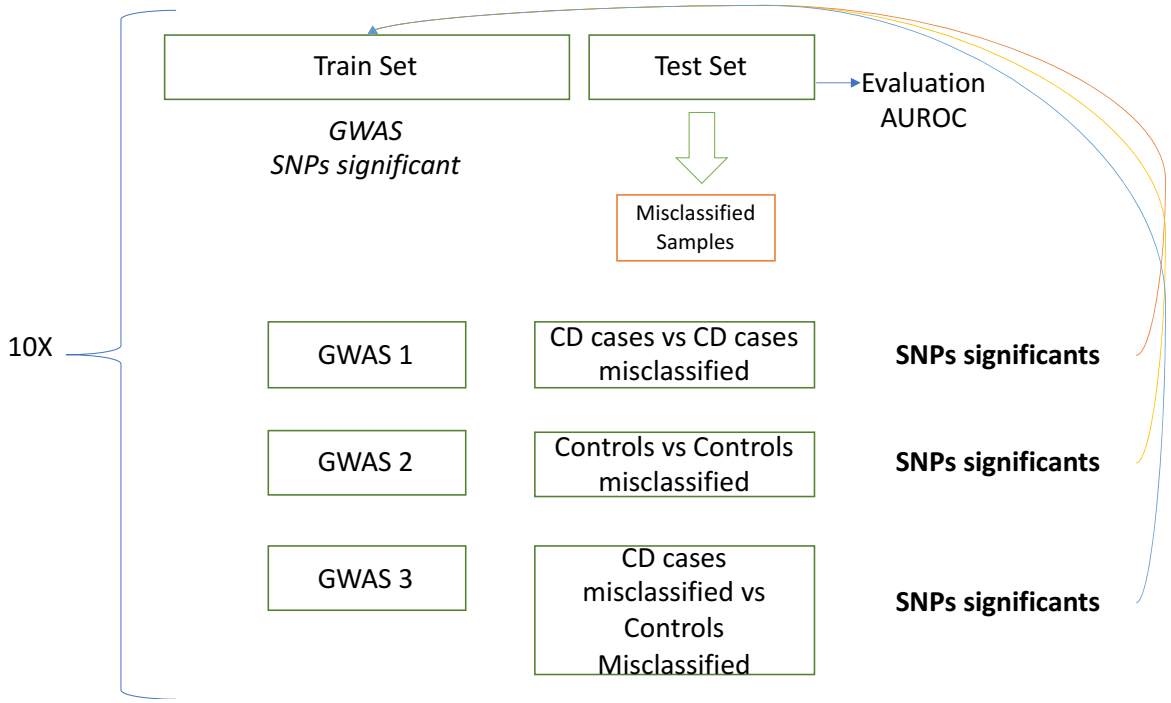

Figure 16. Approach for misclassified samples. 3 additional GWAS were performed for the misclassified samples.

3.2.8.4 GENOTYPE SUBTYPING FOR "NOVEL" MARKER

Fourth, to further challenge rs4945943 polymorphisms attempting to explain its importance in CD and to explore possible links between this association and known CD genes, it was reasoned that CD might present highly diverse genotypes. Thus, surged the question of what the genome-wide associations could be when cases and controls are pre-selected for the same specific rs4945943 polymorphism. This procedure could imply an epistatic effect between the resulting associations and a particular polymorphism. Therefore, a GWAS analysis was performed, comparing controls and cases having AA, AB, or BB polymorphisms in rs4945943 (**Figure 17**). To compensate for decreases in the number of samples, only the 1205 SNPs whose p-value < 10^{-2} in the entire dataset were used. The number of cases was 251, 1,655, and 2,601, correspondingly to AA, AB, and BB polymorphisms, and the number of controls was 397, 3,154, and 5,643, respectively.

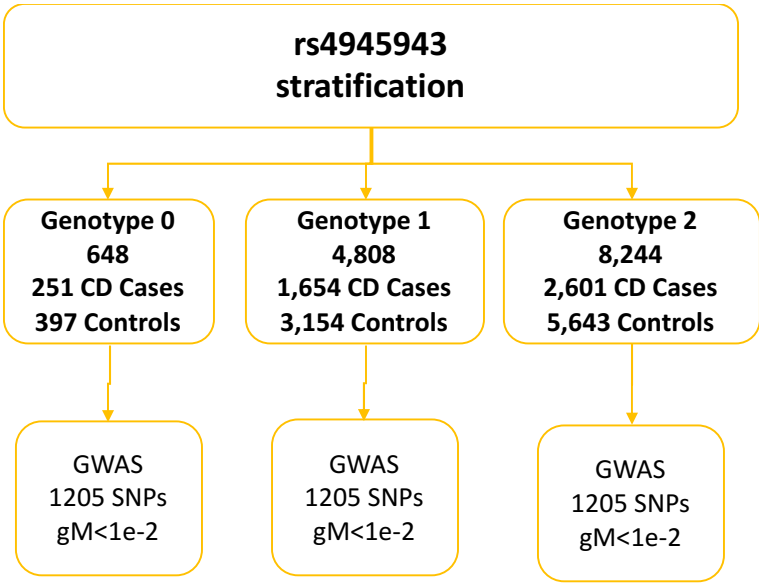

Figure 17. Representation of the genotype subtyping analysis for the rs4945943 SNP.

4. CHAPTER 4: RESULTS

4.1 GWAS DATA AND QC FILTERING

For the discovery and testing, the GWAS raw data was obtained from the European Genome-phenome Archive (EGA) website under accession EGAS00001000924 and EGAC00001000205 (de Lange et al., 2017), corresponding to the UK IBD Genetics (UKIBDGC) and UK10K GWAS dataset. The dataset contained 4,508 UK CD cases (EGAS00001000924), diagnosed using accepted endoscopic, histopathological, and radiological criteria, and genotyped on the Human Core Exome v12.1. 9,944 population control samples genotyped on the Human Core Exome v12.0 were obtained from the Understanding Society Project (EGAC00001000205). Quality control (QC) for genotypes and sample were conducted as implemented in the original published data (de Lange et al., 2017), which consisted of removing variants that were not present on both versions of the genotyping platforms, had missing values >5%, had a significant difference in call rate between cases and controls ($P < 10^{-5}$), deviated from Hardy–Weinberg equilibrium in controls ($P < 10^{-5}$) or were affected by a genotyping batch effect (significant association ($P < 10^{-5}$). For samples, the criteria were to keep the samples that passed the QC in the original study (information provided in the dataset information). For this non-imputed dataset, after quality control for genotypes, data were available for 4,508 Crohn's disease cases and 9,194 controls for 246,735 variants (**Figure 18).**

Finally, the top 10 principal components provided in the dataset were used to correct for population structure for the subsequent analysis (**Figure 19**).

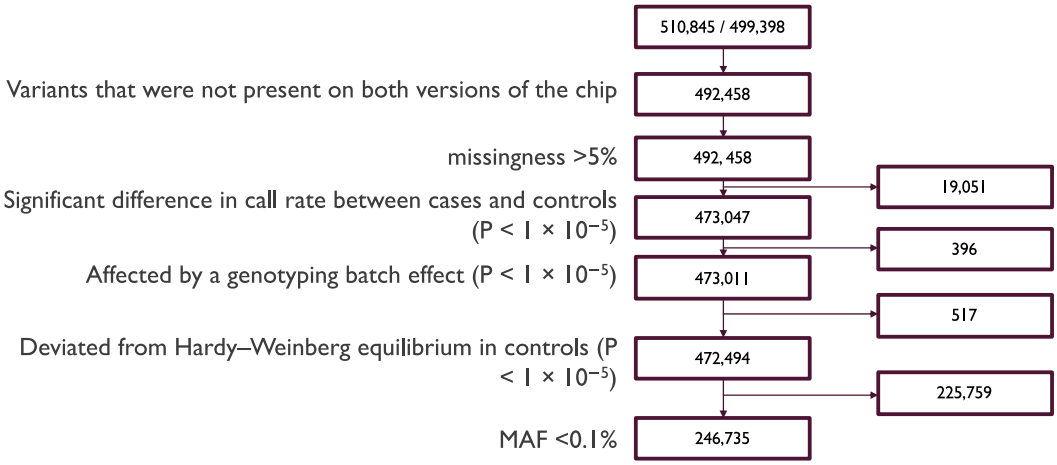

Figure 18. Summary of non-imputed variants QC

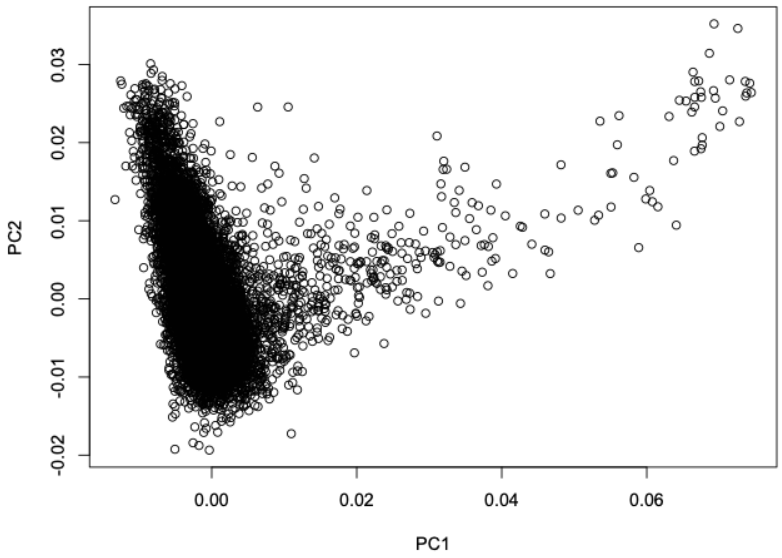

Figure 19. Top 2 Principal components plot

For the validation analysis, the NIDDK IBD Genetics Consortium Crohn's Disease data was requested, obtained through dbGaP accession phs000130.v1.p1 (Duerr et al., 2006). This dataset contained 513 CD cases and 515 control samples from European ancestry. SNPs were excluded for call rates less than 90% and MAF less than 1%. After this QC process, 313,752 SNPs remained. Only SNPs overlapping with the UKIBDGC and UK10K GWAS dataset were used for analysis.

4.2 GENOTYPE IMPUTATION

Imputation was performed remotely using the Michigan Imputation Server (S et al., 2016). For this process, the European 1000 genomes reference data was applied. 47,077,455 variants were retrieved for autosomal chromosomes. For this dataset, QC for genotypes consisted of removing variants with MAF <1%, INFO SCORE <0 .4 (which is a measure of imputation quality), and deviated from Hardy–Weinberg equilibrium in controls ($P < 10^{-7}$). The process resulted in 8,755,412 variants after QC for the imputed dataset (**Table 2).**

Table 2. Summary of imputed variants QC

Filtering Criteria	Chr	1-22
	SNPs Imputation	47,077,455
	Mismatching calls	160
	Duplicate Vars	2,603
	Qualilty < 0.4	32,861,186
	HWE $p < 1e^{-7}$	588
	Maf < 1%	5,457,506
Final Data	Variant postQC	**8,755,412**
	% Recovery	19%

4.3 UNIVARIATE GWAS REPLICATION (100% SAMPLES)

4.3.1 NON-IMPUTED DATASET

An additive frequentist analysis corrected for population structure was implemented for all QC-pass GWAS data. With 100% of samples, the univariate analysis yielded a correlation of 95% (**Figure 20**) against the original data. Among the causes of the differences between this replica and the original results are the samples used for the analysis, as they removed some additional samples to perform the metanalysis.

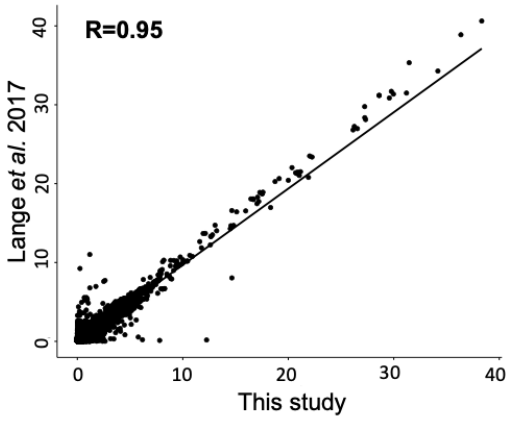

Figure 20. Correlation plot of GWAS replication in log10 p-values scale. Only SNPs not imputed and used for metanalysis within Langer et al. 2017 are considered for this plot.

The Manhattan plot (**Figure 21**) and QQ plot (**Figure 22**) showed CD-associated locus in chromosomes 1, 5, and 16, previously associated with disease susceptibility. The thresholds for identifying the significant variants were $1e^{-7}$ for genome-wide significance and $1e^{-5}$ for suggestive associations.

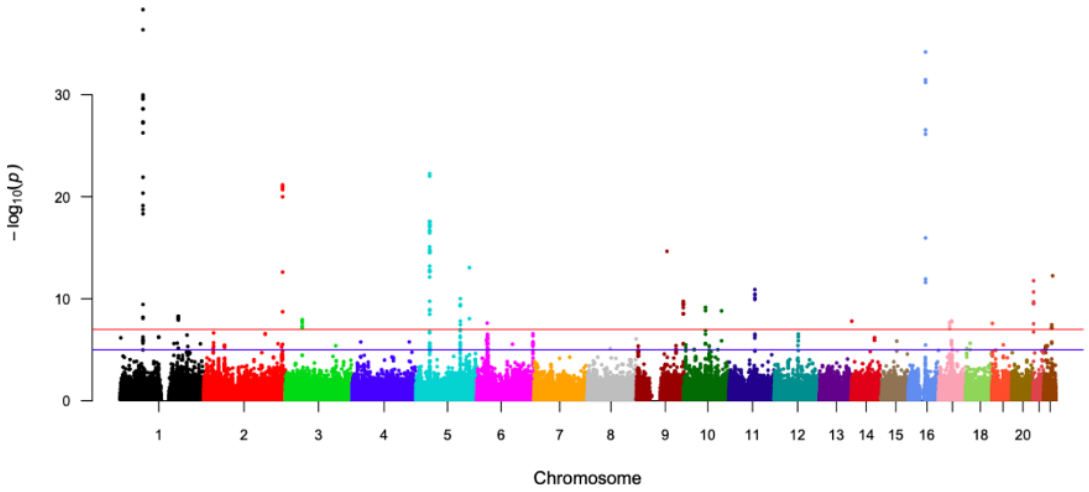

Figure. 21. Manhattan plot of CD GWAS, for 100% sample and not impute dataset, x-axis refers to chromosomes and y-axis to -log10 p values of logistic regression. Red line: genome-wide significant threshold $1e^{-7}$. Blue line: Suggestive association threshold $1e^{-5}$.

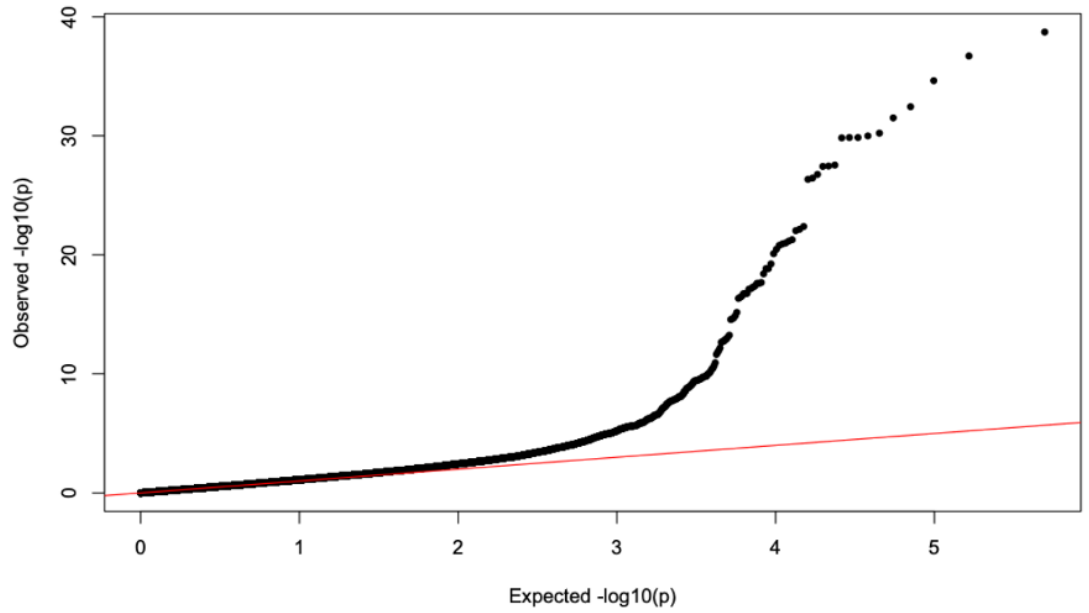

Figure 22. QQ plot of GWAS for the not imputed dataset, the x-axis represents expected -log10 and y-axis refer to the observed -log10 of each SNP.

4.3.2 IMPUTED DATASET

An additive frequentist analysis, corrected for population structure, was set in SNPTEST v2.0 (Marchini et al., 2007). This was implemented for all QC-pass GWAS data with 100% samples and yielded a Pearson correlation of 87% (**Figure 23**) against the original data. Among the causes for the differences between this replica and the original results are: 1) the samples used for the analysis were slightly different because they removed some duplicated samples, with other studies, to perform the metanalysis, and 2) they used an additional (not provided) set of whole-genome sequencing to enrich the imputation reference dataset. The Manhattan plot (**Figure 24**) shows known CD-associated locus at thresholds of 1e-7 for genome-wide significant variants and 1e-5 for variants with a suggestive association. The QQ plot (**Figure 25**) represents the deviation of the observed P values from the null hypothesis. These results provide confidence for this subsequent analysis due to the reproducibility of associated variants.

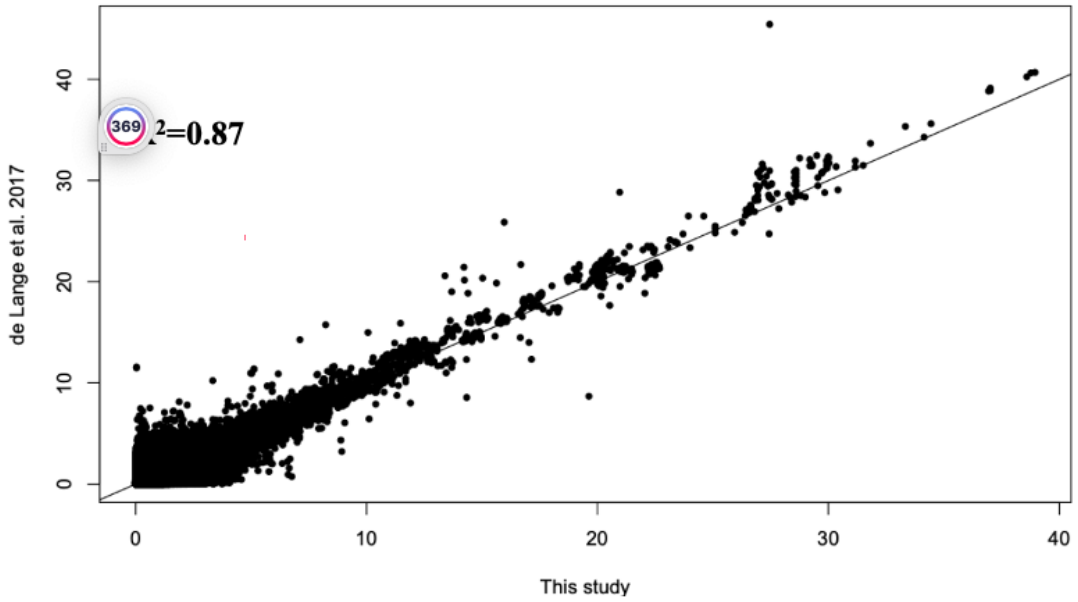

Figure. 23. Correlation of -log10 P values of replica and CD GWAS.

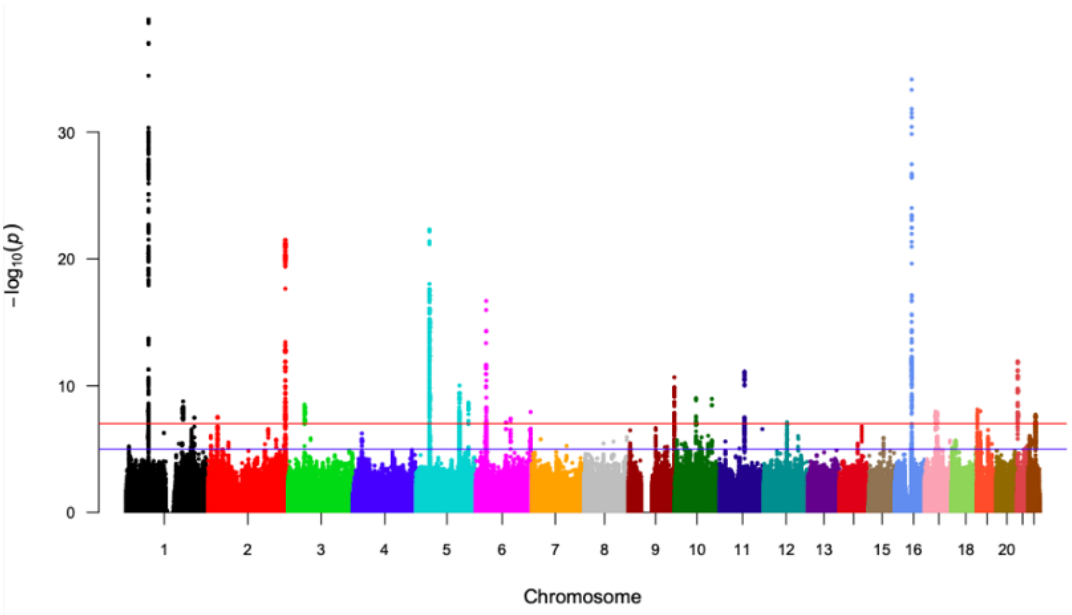

Figure 24. Manhattan plot of CD GWAS for 100% samples and imputed dataset, x-axis refers to chromosomes and y-axis to -log10 p values of logistic regression. Red line: genome-wide significant threshold $1e^{-7}$. Blue line: Suggestive association threshold $1e^{-5}$.

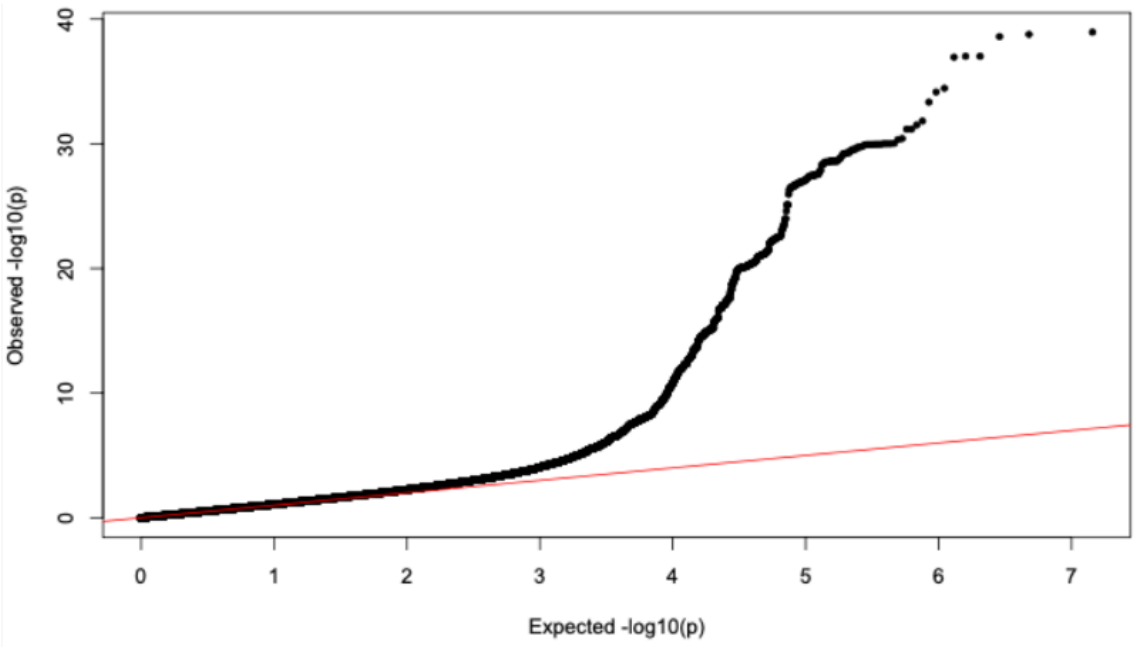

Figure 25. QQ plot of GWAS for the imputed dataset, the x-axis represents expected -log10 and y-axis refer to the observed -log10 of each SNP.

4.4 APPROACH DESIGNED TO EVALUATE MULTIVARIATE AND UNIVARIATE-BASED MODELS TO PREDICT CD RISK IN THE UKIBDGC AND UK10K GWAS DATASET.

The approach described within the methods section compared different methods to predict CD risk from a GWAS dataset. For this, four pipeline versions were derived from the combination of original data with and without SNP imputation together with and without LD clumping. To facilitate computational analysis, a univariate analysis on 40% of the samples was performed to select the top 5% of the features, including more than genome-wide significant SNPs. The remaining 60% of the samples were used to robustly estimate a p-value for each SNP and the error for multivariate risk models. The robust estimation was performed in 10-fold-subsamples where 70% of the samples were used for training and 30% for testing when evaluating multivariate models.

The robustly estimated geometric mean p-value was used to filter markers at specific p-value thresholds and then fed into multivariate algorithms to generate the multivariate models.

4.4.1 SNPS DATASETS

The UKIBDGC AND UK10K GWAS dataset was selected to train and test the CD risk, prediction models. In addition, the NIDDK IBD Genetics Consortium Crohn's Disease was selected as the validation set, where the best models were also tested. **Figure 26** describes the analysis that was conducted in each dataset.

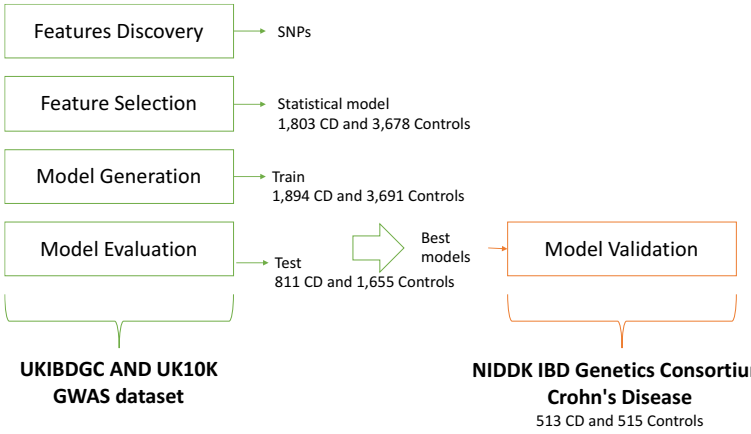

Figure 26. Datasets selected for CD risk prediction methodology.

4.4.2 SNPS FILTERING: PRE-SELECTION GWAS (40% OF SAMPLES)

For both not imputed and imputed datasets, a GWAS was performed on the 40% of samples to decrease the number of features and facilitate the subsequent analysis. A frequentist additive analysis was implemented for both sets. The not imputed dataset contained 246,735 SNPs, and the imputed dataset consisted of 8,755,412 SNPs. The samples were the same for both analyses, corresponding to 1,803 CD cases and 3,676 healthy controls. These analyses were performed within SNPTEST v2.0 (Marchini et al., 2007).

For this 2 GWAS, 3,846 and 111,653 markers were found to be significant at a p-value < $1e^{-2}$, however, to add more features, SNPs for the cross-validated univariate analysis, a top 5% of SNPs were selected for both imputed and not imputed sets. **Figure 27** shows the distribution of markers with a p-value <1e-2 for both not imputed and imputed datasets.

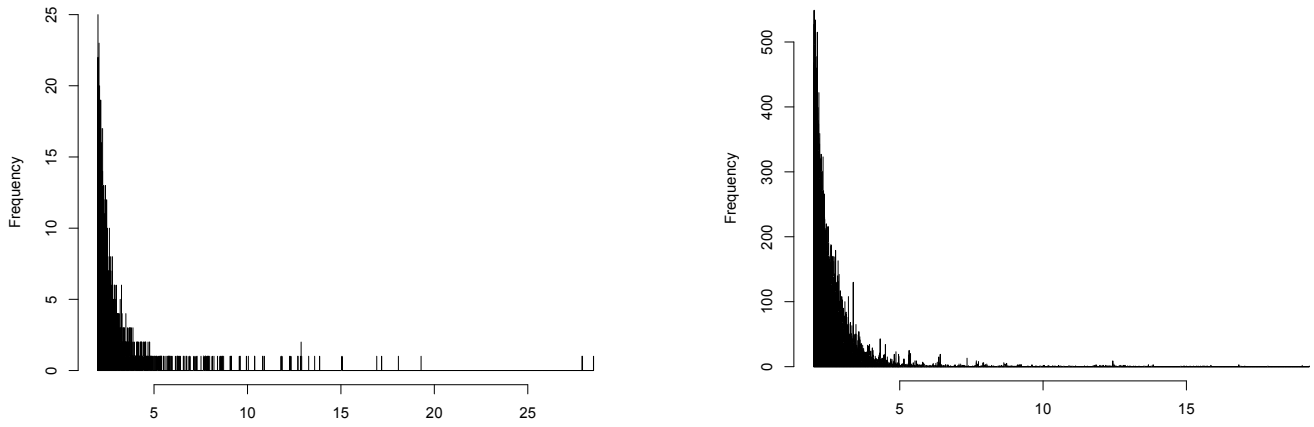

Figure 27. Distribution of markers with a p-value <1e-2, for both A) not imputed and B) imputed dataset

Finally, LD-clumping was performed with plink for both not imputed and imputed datasets. 11,987 and 428,320 markers were selected as the top 5% of each non-imputed and imputed dataset. **Table 3** displays the number of SNPs remaining for each dataset version.

Table 3. Number of SNPs remaining in each dataset version.

Dataset version	**#SNPs LD-clumping**	**#SNPs No LD-clumping**
Not imputed	11,987	1,413
Imputed	428,320	32,142

4.4.3 PRE-SELECTION OF SNPS FOR UNIVARIATE ANALYSIS

The groups of markers were filtered based on the p-value threshold from $1e^{-7}$ to $1e^{-2}$. However, for the imputed dataset, the analysis for some multivariate methods was possible only up to $1e^{-4}$ because of the complexity of the models. **Table 4** shows the number of markers selected for a predefined threshold of $1e^{-7}$, $1e^{-6}$ $1e^{-5}$, $1e^{-4}$, $1e^{-3}$, and $1e^{-2}$, for either the imputed dataset or the not imputed dataset and for the application of LD clumping.

Table 4. Mean number of SNPs selected in each threshold. ± standard deviation.

Dataset version / Filtering Threshold	**$1e^{-7}$**	**$1e^{-6}$**	**$1e^{-5}$**	**$1e^{-4}$**	**$1e^{-3}$**	**$1e^{-2}$**
Not Imputed No LD clumping	39 ± 6	46 ± 5	60 ± 6	88 ± 6	168 ± 15	417 ± 25
Not Imputed with LD clumping	4 ± 1	5 ± 1	6 ± 1	8 ± 1	13 ± 2	34 ± 5
Imputed No LD clumping	516 ± 133	694 ± 108	990 ± 143	1,555 ± 135	3,249 ± 401	10,324±485
Imputed with LD clumping	7 ± 2	9 ± 2	13 ± 2	25 ± 4	74 ± 9	439±23

4.4.4 SNPS REPLICATED AMONG THE 10X RANDOM SAMPLING GWAS

A 10X random subsampling approach was implemented to consider the "sampling effect" for the marker's pre-selection from the univariate analysis. **Figures 28 and 29** show the number of variants replicated among the 10X GWAS for the not imputed and imputed dataset, most of them reproduced with higher thresholds. The lack of replication is evident for p-values less stringent as $p<1e^{-2}$. This lack of replication, for SNPs not genome-wide associated, is one of the documented pitfalls of GWAS. Also, these results show that the replication of markers is slightly better for the imputed dataset than the non-genotyped SNPs (imputed set).

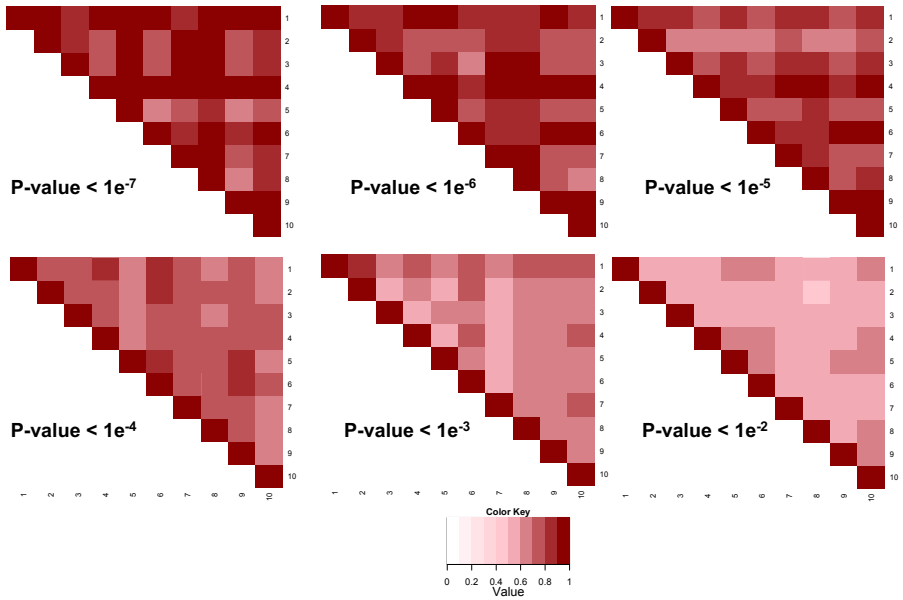

Figure 28. Percentage of markers replicated in each GWAS at different thresholds for the not imputed dataset.

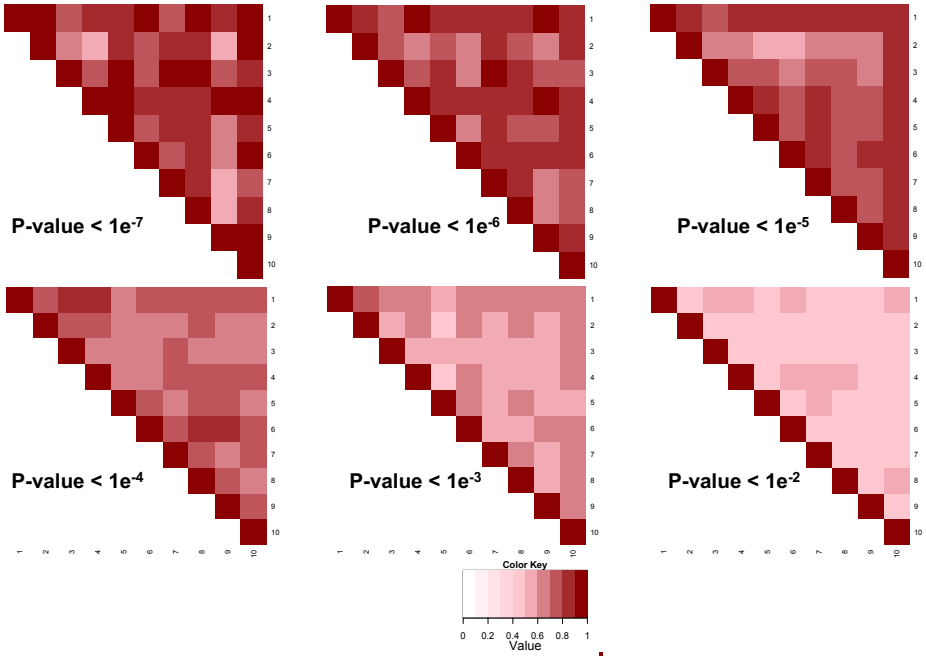

Figure 29. Percentage of markers replicated in each GWAS at different thresholds for the imputed dataset.

4.4.5 ROBUST ESTIMATION OF MULTIVARIATE POTENTIAL

The 10X test mean and standard deviation AUROC for each combination of the dataset, filtering, LD clumping, imputation, and predictor model is shown in **Table 5** for non-imputed

datasets and **Table 6** for imputed datasets. The AUROC mean and deviation for each model are also shown in **Figure 30**. Due to LDpred building models by filtering the markers by a causal fraction (using a reference LD panel to model the LD among SNPs), these 10X models were tested using various causal fractions cutoffs for the two versions of the No LD clumped (top 5% markers) datasets. The best performance observed was obtained by us all the Top 5% SNPs in the not imputed dataset (mean AUROC=0.667, **Figure 30A** and **Figure 30B**, **Table 7**, **Figure 31**). Compared to both the multivariate and PRS univariate-based models, this performance was the best. All markers with no filtering were fitted, 2M markers in imputed data and 100K markers in no imputed data, and with a variation in the causal fraction. This, to explore a further increment in AUROC. However, the results did not show an increment for AUROC in both analyses (**Table 7** and **Table 8**).

PRS P+T showed the second-best performance (AUROC=0.656). However, the multivariate models were further analyzed, specifically the XGBoost models, due to its evaluation of variants (the variant importance estimation), which is a measure that cannot be directly estimated from a typical PRS analysis.

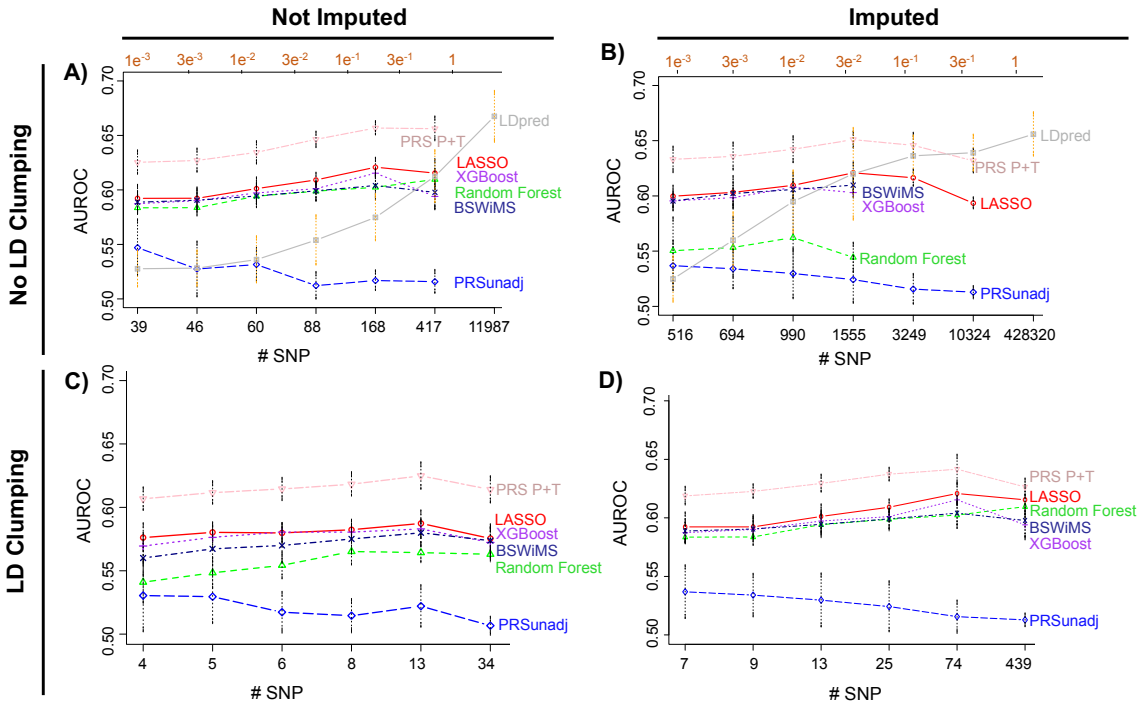

Figure 30. Mean AUROC of multivariate and univariate-based models in testing sets across datasets. (A) For not imputed with no LD clumping. (B) For not imputed with LD clumping. (C) For imputed with no LD clumping. (D) For imputed with LD clumping. Multivariate models: BSWIMS, LASSO, Random Forest, and XGBoost. Univariate-based models: PRS unadjusted, PRS P+T, and LDpred. Vertical bars around mean dots represent the standard deviation. The upper axis (orange) for the No LD clumped set only corresponds to LDpred causal fractions (gray).

For the multivariate models, LDpred (AUROC=0.667), LASSO (AUROC=0.621), and XGBoost (AUROC=0.615) showed better performances than the other models along dataset versions and number of SNP used. LASSO was slightly better than XGBoost in the four datasets versions. LASSO, BSWiMS, and LDpred were also analyzed because even though they have an effect size, as the PRS methods, and do not provide a variant analysis, to assign importance for the markers compared to methods as Random Forest and XGBoost, these effects sizes are readjusted according to the importance of the features within the models.

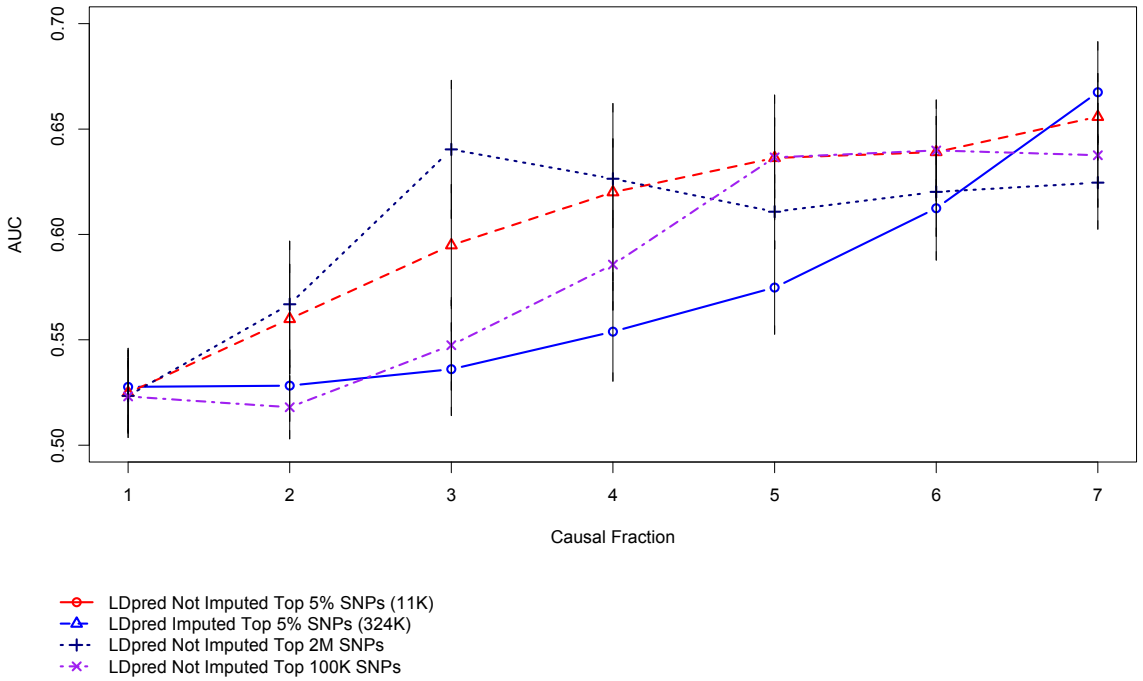

Figure 31. Mean ROC AUC of LDpred models in the testing set for not imputed and imputed sets. Vertical bars show the standard deviation.

Overall the performances of most algorithms do not show a detrimental tendency when increasing the number of SNPs, even when they barely carry CD causal information (when less significant SNPs are added). Only PRS unadjusted showed a decrement for the AUROC when less informative SNPs were added (**Figure 30**).

Table 5. Mean AUROC for testing dataset, for 10X RS for multivariate models and univariate-based models, for the not imputed dataset. Data in bold refer to the highest AUROC value. +/- indicate standard deviation

Threshold (for p-value filtering)	**$1e^{-7}$**	**$1e^{-6}$**	**$1e^{-5}$**	**$1e^{-4}$**	**$1e^{-3}$**	**$1e^{-2}$**
#SNPs No LD clumping	**39**	**46**	**60**	**88**	**168**	**417**
LASSO	0.592 ±0.013	0.592 ±0.011	0.601 ±0.011	0.609 ±0.007	**0.621 ±0.009**	0.615 ±0.013
RF	0.584 ±0.006	0.584 ±0.008	0.594 ±0.009	0.599 ±0.010	0.602 ±0.012	**0.606 ±0.007**
XGBoost	0.587 ±0.009	0.590 ±0.007	0.597 ±0.01	0.601 ±0.005	**0.615 ±0.004**	0.594 ±0.013
BSWiMS	0.589 ±0.011	0.591 ±0.011	0.594 ±0.011	0.599 ±0.009	**0.604 ±0.011**	0.598 ±0.015
PRSunadj	**0.547 ±0.027**	0.527 ±0.026	0.532 ±0.016	0.512 ±0.013	0.517 ±0.010	0.516 ±0.011
PRS P+T	0.625 ±0.012	0.627 ±0.012	0.634 ±0.011	0.646 ±0.008	**0.657 ±0.007**	0.656 ±0.012
#SNPs LD clumping r2 > 0.05	**4**	**5**	**6**	**8**	**13**	**34**
LASSO	0.57 6±0.012	0.580 ±0.008	0.580 ±0.008	0.582 ±0.008	**0.587 ±0.011**	0.575 ±0.012
RF	0.541 ±0.017	0.548 ±0.013	0.554 ±0.011	**0.565 ±0.011**	0.564 ±0.009	0.563 ±0.006
XGBoost	0.569 ±0.011	0.576 ±0.010	0.580 ±0.008	0.580 ±0.006	**0.583 ±0.008**	0.572 ±0.012
BSWiMS	0.560 ±0.015	0.567 ±0.014	0.570 ±0.014	0.575 ±0.012	**0.580 ±0.011**	0.574 ±0.011
PRSunadj	**0.530 ±0.029**	0.530 ±0.022	0.517 ±0.017	0.515 ±0.014	0.522 ±0.017	0.507 ±0.008
PRS P+T	0.607 ±0.010	0.612 ±0.010	0.615 ±0.009	0.618 ±0.010	**0.625 ±0.011**	0.614 ±0.011

Table 6. Mean AUROC for the testing dataset, for 10X RS for multivariate models, and univariate-based models, for the imputed dataset. Data in bold refer to the highest AUROC value. +/- indicate standard deviation

Threshold (for p-value filtering)	**$1e^{-7}$**	**$1e^{-6}$**	**$1e^{-5}$**	**$1e^{-4}$**	**$1e^{-3}$**	**1e-2**
#SNPs No LD clumping	**516**	**694**	**990**	**1555**	**3249**	**10324**
LASSO	0.600 ±0.010	0.603 ±0.015	0.610 ±0.012	**0.621 ±0.010**	0.616 ±0.008	0.593 ±0.006
RF	0.550 ±0.031	0.553 ±0.028	**0.562 ±0.037**	0.544 ±0.014	NA	NA
XGBoost	0.596 ±0.010	0.598 ±0.009	**0.608 ±0.005**	0.603 ±0.006	NA	NA
BSWiMS	0.595 ±0.007	0.602 ±0.011	0.606 ±0.010	**0.610 ±0.008**	NA	NA
PRSunadj	**0.537 ±0.023**	0.534 ±0.019	0.530 ±0.023	0.524 ±0.022	0.516 ±0.014	0.513 ±0.006
PRS P+T	0.633 ±0.012	0.636 ±0.013	0.642 ±0.013	**0.651 ±0.008**	0.646 ±0.012	0.632 ±0.012
#SNPs LD clumping r2 > 0.05	**7**	**9**	**13**	**25**	**74**	**439**
LASSO	0.583 ±0.009	0.586 ±0.007	0.590 ±0.011	0.594 ±0.006	**0.594 ±0.005**	0.573 ±0.008
RF	0.558 ±0.023	0.553 ±0.016	0.550 ±0.020	0.551 ±0.019	**0.560 ±0.020**	0.550 ±0.013
XGBoost	0.578 ±0.004	0.585 ±0.011	0.588 ±0.013	**0.591 ±0.007**	0.589 ±0.016	0.567 ±0.010
BSWiMS	0.569 ±0.008	0.575 ±0.010	0.579 ±0.007	**0.584 ±0.008**	0.583 ±0.006	0.564 ±0.011
PRSunadj	0.532 ±0.024	**0.533 ±0.025**	0.528 ±0.021	0.5184 ±0.011	0.519 ±0.018	0.511 ±0.010
PRS P+T	0.619 ±0.008	0.623 ±0.007	0.629 ±0.008	0.637 ±0.006	**0.642 ±0.013**	0.626 ±0.008

Table 7. Mean AUROC for testing dataset, for 10X RS for LDpred models, for the imputed and not imputed dataset. Data in bold refer to the highest AUROC value. +/- indicate standard deviation

Threshold (for Causal Fraction)	**$1e^{-3}$**	**$3e^{-3}$**	**$1e^{-2}$**	**$3e^{-2}$**	**$1e^{-1}$**	**$3e^{-1}$**	**1**
~ #SNPs	12	36	118	355	1,184	3,552	11,839
LDpred Not Imputed Top 5%	0.528 ±0.017	0.528 ±0.017	0.536 ±0.022	0.554 ±0.024	0.575 ±0.022	0.612 ±0.025	**0.667 ±0.024**
~ #SNPs	88	265	882	2,645	8,818	26,454	88,181
LDpred Not Imputed Top 100K	0.523 ±0.017	0.518 ±0.015	0.547 ±0.021	0.586 ±0.022	0.637 ±0.030	**0.640 ±0.024**	0.638 ±0.025
~ #SNPs	324	973	3,244	9,732	32,439	97,317	324,390
LDpred Imputed Top 5%	0.525 ±0.021	0.560 ±0.026	0.595 ±0.029	0.620 ±0.042	0.636 ±0.019	0.639 ±0.017	**0.656 ±0.021**
~# SNPs	1,480	4,439	14,798	44,395	147,983	443,948	1,479,826
LDpred Imputed Top 2M	0.523 ±0.018	0.567 ±0.030	**0.640 ±0.033**	0.626 ±0.019	0.611 ±0.018	0.620 ±0.021	0.625 ±0.022

Table 8. Mean AUROC for testing dataset, for 10X RS for LDpred models, for the not imputed dataset of 100K SNPs, with causal fractions from $1e^{-1}$ to $3e^{-1}$. Data in bold refer to the highest AUROC value. +/- indicate standard deviation.

Threshold *(for Causal Fraction)*	***$1e^{-1}$***	***$1.5e^{-1}$***	***$2e^{-1}$***	***$2.5e^{-1}$***	***$3e^{-1}$***
~ #SNPs	8,818	13,227	17,636	22,045	26,454
LDpred Not Imputed Top 5%	0.642 ±0.030	**0.660 ±0.023**	0.659 ±0.025	0.653 ±0.025	0.637 ±0.018

Imputation of non-genotyped markers was performed to increase the number of features. In addition, LD clumping was also used to remove redundancy and decrease the data dimensionality with filtering based on the correlation of the markers, which is known to be high in imputed datasets [43]. However, for all the methods, the imputed dataset had only a slight increase in AUROC compared to the original, not imputed dataset (**Figure 32**), except for Random Forest. In LASSO, for example, the mean AUROC=0.610 using 88 SNPs, but for the imputed dataset, the mean AUROC=0.621 using 1555 SNPs (both results without LD clumping). Similarly, LD clumping had an overall detrimental effect on the performance of

the models for both the not imputed and imputed datasets. This is most evident in the not imputed datasets (**Figure 32**).

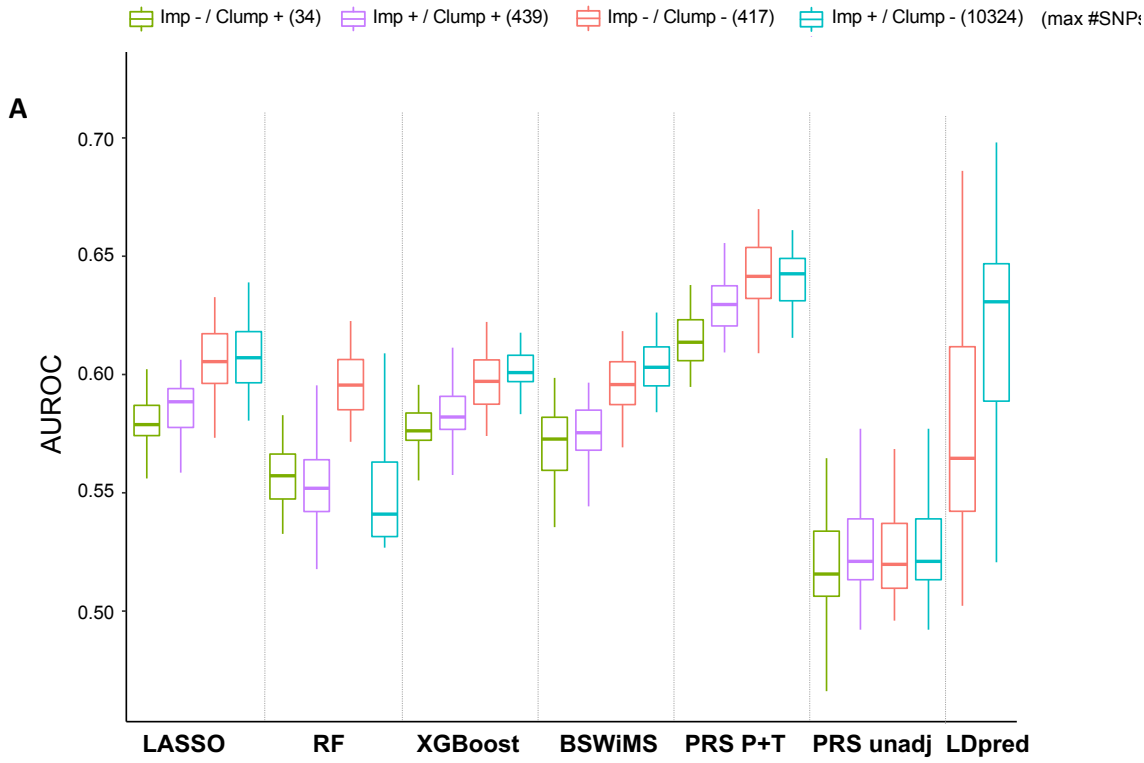

Figure 32. AUROC values for the seven methods and their respective thresholds or causal fraction (LDpred). AUROC values for each method across the 4 datasets. The maximum number of features for each class is indicated in parenthesis.

4.4.6 STATISTICAL ANALYSIS OF CD RISK PREDICTION ASSOCIATED METHODS

4.4.6.1 CRITICAL DIFFERENCE

A critical difference analysis highlighted the differences among methods for the results of both imputed and not imputed datasets and both with and without LD clumping (**Figure 33A**). This analysis fits a Nemenyi post-hoc test to rank methods based on their AUROC. However, this diagram did not compare LDpred results because its AUROC is based on causal fractions rather than p-value thresholds. **Figure 33A** shows that when LDpred is not present, PRS P+T had the best results, and PRS unadjusted had the lowest-ranked results.

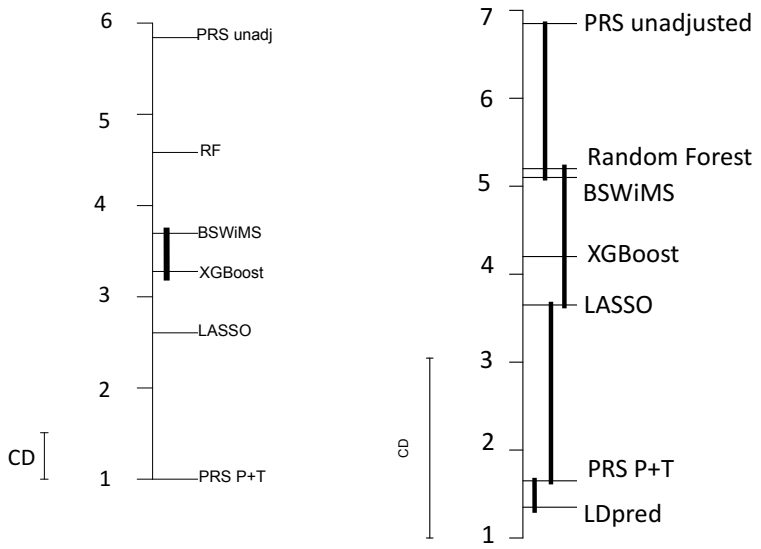

Figure 33. Critical difference of the multivariate and univariate-based methods shows the statistical comparison of all models against each other. Classifiers that are not connected by a bold line of length equal to CD (critical difference) have significantly different mean ranks (Confidence level of 95%) A) All 10X models and thresholds for imputed and not imputed datasets, without LDpred method. B) Best 10X models for every method against each other.

Another critical difference analysis was performed to fit LDpred AUROC estimates and compare them against the other multivariate and univariate analyses. The best 10X models for each method were selected to fit the critical difference test for imputed and not imputed data. **Figure 33B** shows the statistical comparison of all 10X models against each other. This diagram ranks LDpred as the best method, with no significant difference with PRS P+T. This validates the use of LDpred models to identify variants and genes associated with CD risk.

4.4.6.2 LINEAR REGRESSION

The prediction of CD risk was performed under different conditions such as imputation status, LD clumping application, threshold or causal fraction filtering, fold repetition, and multivariate and univariate-based methods. Thus, a linear regression analysis was performed to identify the effect of each condition on the prediction (AUROC) is represented by **equation 11**.

$$AUROC \sim \beta_o + \beta_I + \beta_L + \beta_T + \beta_F + \beta_M + \beta_{IL} + \beta_{IT} + \beta_{IF} + \beta_{IM} + \beta_{LT} + \beta_{LF} + \beta_{LM} + \beta_{TF} + \beta_{TM} + \beta_{FM}$$

Equation 11. Linear regression model for AUROC

Where **AUROC** refers to the area under the curve reached by each model, **0** to the intercept, **I** to the imputation status, **L** to the LD clumping application, **T** to the threshold or causal fraction filtering, **F** to the fold number, **M** to the multivariate or univariate-based methods, and the interactions among them. This analysis was performed in the R-package. The estimates and significance p-values for all the variables are displayed in **Table 9**. LD clumping status, method, and the interaction between imputation and either LD clumping status, threshold, or method were significant after a Bonferroni correction (p-value < 0.05/40), accounting for the 40 predictors fitted in the model. As expected, the fold variable was not significant, meaning that the random sampling method did not affect the models.

Table 9. Summary of linear regression model for all the model's variables (imputation and LD clumping status, Thresholds, Folds, and Methods).

Variable	Estimate	Std. Error	t value	Pr(>|t|)	
Imputed	0.0080	3.14E-03	2.431	0.01517	*
LD	0.0243	3.22E-03	7.544	7.90E-14	***
TH	-0.535	3.744E-01	-1.429	0.15314	
Fold	-0.0004	4.728E-04	-0.896	0.37026	
Method XGBoost	0.0090	4.51E-03	1.990	0.04676	*
Method LASSO	0.008	4.41E-03	1.832	0.06713	.
Method LDpred	-0.060	5.37E-03	-11.288	< 2e-16	***
Method PRS P+T	0.0460	4.413E-03	10.350	< 2e-16	***
Method RF	-0.009	4.510E-03	-2.125	0.03375	*
Method PRS unadjusted	-0.049	4.41E-03	-11.169	< 2e-16	***
Imputed * No LD clumping	-0.010	1.958E-03	-5.225	1.99E-07	***
Imputed * TH	-0.053	8.934E-03	-6.037	1.98E-09	***
Imputed * Fold	0.0005	3.21E-04	1.771	0.0767	.
Imputed * Method XGBoost	-0.002	3.472E-03	-0.444	0.65693	
Imputed * Method LASSO	-0.001	3.379E-03	-0.419	0.67513	
Imputed * Method LDpred	0.0430	4.45E-03	9.763	< 2e-16	***
Imputed * Method PRS P+T	0.0010	3.38E-03	0.319	0.74999	
Imputed * Method RF	-0.025	3.472E-03	-7.224	8.05E-13	***
Imputed * Method PRS unadjusted	-0.003	3.379E-03	-0.787	0.4314	
No LD clumping * TH	0.8350	2.81E-01	2.967	0.00305	**
No LD clumping * Fold	0.0009	3.37E-04	2.657	0.00797	**
No LD clumping * Method XGBoost	-0.007	3.472E-03	-2.035	0.04202	*
No LD clumping * Method LASSO	-0.003	3.379E-03	-0.831	0.40628	

No LD clumping Method PRS + PT	-0.007	3.379E-03	-2.181	0.02936	*
No LD clumping * Method RF	-0.004	3.472E-03	-1.029	0.30365	
No LD clumping * Method PRS unadjusted	-0.023	3.379E-03	-6.845	1.12E-11	***
TH * Fold	-0.003	1.555E-03	-2.061	0.03952	*
TH * Method XGBoost	-0.665	5.110E-01	-1.300	0.19371	
TH * Method LASSO	-0.395	4.824E-01	-0.820	0.41257	
TH * Method LDpred	-0.148	4.068E-01	-0.363	0.71645	
TH * Method PRS P+T	0.4030	4.82E-01	0.836	0.40314	
TH * Method RF	1.3900	5.11E-01	2.721	0.00659	**
TH * Method PRS unadjusted	-1.422	4.824E-01	-2.948	0.00325	**
Fold * Method XGBoost	-0.00006	5.946E-04	-0.098	0.92228	
Fold * Method LASSO	0.0006	5.83E-04	1.053	0.29239	
Fold * Method LDpred	0.0010	7.68E-04	1.461	0.14426	
Fold * Method PRS P+T	0.0003	5.83E-04	0.585	0.55831	
Fold * Method RF	0.0002	5.95E-04	0.338	0.7354	
Fold * Method PRS unadjusted	0.0005	5.83E-04	0.855	0.39277	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Residual standard error: 0.01791 on 1480 degrees of freedom. Multiple R-squared: 0.8036, Adjusted R-squared: 0.7984

4.4.7 MISCLASSIFIED SAMPLES EXPERIMENT

Additional model building was conducted for the poorly or misclassified samples for both LASSO and XGBoost, selecting misclassified samples from the not imputed data, with no LD clumping and a p-value < 1e-3. For this, three different GWAS were implemented for each method. **Table 10** shows the mean number of samples selected for each GWAS and the assigned code for each group.

Table 10. Mean number of misclassified samples for each GWAS conducted for both LASSO and XGBoost models. CD (Crohn's Disease), HC (Healthy controls)

		CD vs mcCD	HC vs mcHC	mcCD vs mcHC
LASSO	**0**	1,252	2,570	1,291
	1	641	1,291	641
XGBoost	**0**	1,224	2,739	1,122
	1	669	1,122	669

The number of markers added to the model building analysis for both LASSO and XGBoost is displayed in **Table 11**.

Table 11. Number of SNPs added to the new LASSO and XGBoost models (LASSO | XGBoost) from each GWAS.

	CD vs mcCD	HC vs mcHC	mcCD vs mcHC
1e-7	53 | 28	60 | 67	41 | 21
1e-6	60 | 37	65 | 79	47 | 26
1e-5	68 | 47	66 | 87	54 | 29
1e-4	71 | 54	67 | 89	60 | 38
1e-3	81 | 65	74 | 100	68 | 46
1e-2	111 | 157	133 | 188	109 | 163

However, the new models derived from adding SNPs associated with both CD or HC from misclassified samples did not reach the maximum AUROC obtained in the original approach (**Figures 34 and 35**). Still, an increment in the AUROC from the most significant p-values was observed, meaning that this strategy allows improving models where the features have been discovered with a high significance, but as the variants with less significance are added to the models (p-value <1e-3), they perform even worse than the original approach.

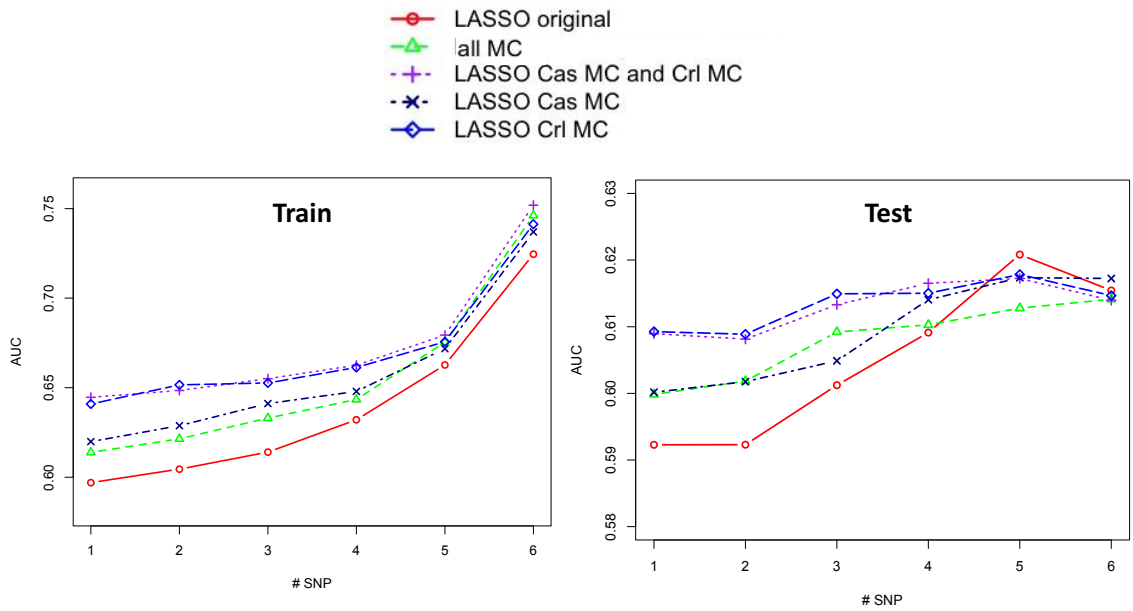

Figure 34. Mean AUROC for the LASSO misclassified approach. All MC (SNPs from mcCD vs. mcHC), CAS MC and Crl MC (SNPs from CD vs. mcCD and HC vs. mcHC), Cas MC (SNPs from CD vs. mcCD), Crl MC (SNPs from HC vs. mcHC). X-axis, p-value thresholds (1e-7 to 1e-2).

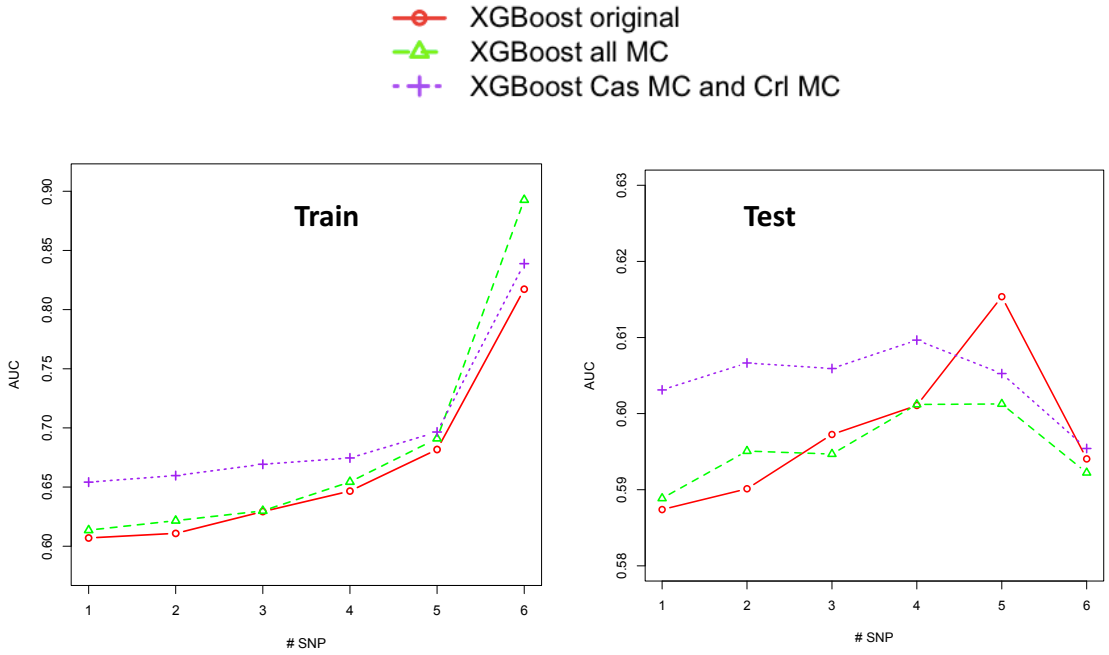

Figure 35. Mean AUROC for the XGBoost misclassified approach. All MC (SNPs from mcCD vs. mcHC), CAS MC and Crl MC (SNPs from CD vs. mcCD and HC vs. mcHC), Cas MC (SNPs from CD vs. mcCD), Crl MC (SNPs from HC vs. mcHC). X-axis, p-value thresholds (1e-7 to 1e-2).

4.4.8 RANDOM PERMUTATION OF PHENOTYPES

A random permutation of phenotypes was implemented to determine if the multivariate model performed better than random. The permutation of phenotypes for LASSO, BSWIMS, and PRS unadjusted, showed that for the testing set, a mean area under the curve around 0.5 was found (**Figure 36**), which means that the multivariate models indeed perform better than random.

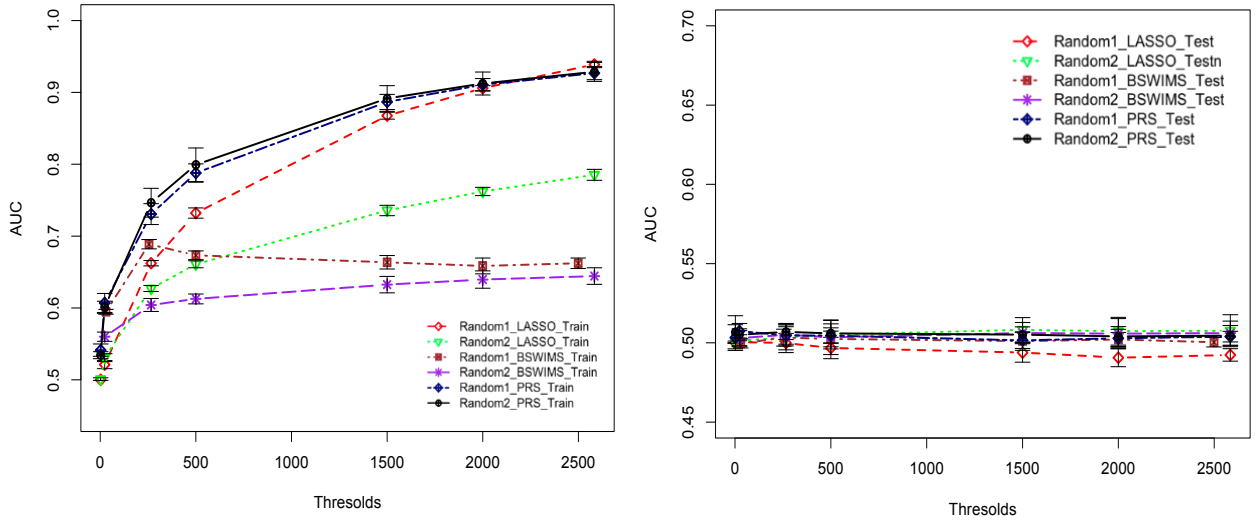

Figure 36. The area under the curve for training (left) and testing set (right) for BSWIMS, LASSO, and PRS unadjusted with permutated data. 2 repetitions were performed, referred to as random 1 and random 2. Multivariate models: BSWIMS and LASSO. Univariate-based models: PRS unadjusted. Vertical bars show the standard deviation.

4.5 MODEL VALIDATION ANALYSIS

4.5.1 GENE-BASED ANALYSIS

Because this analysis might include potential novel markers in the predictor model building step by incorporating SNPs close to the genome-wide significant criteria, the gene and variant content for possible discovery was explored. The different methods evaluated in this research had other techniques to assign a value to each SNP/variant. For example, BSWiMS, LASSO,

and LDpred had a re-estimation of the beta or effect size values, and RF and XGBoost had a measure of importance, which was described within the methods section. On the other hand, PRS P+T does not measure the SNPs' importance (it takes the estimated effects from the GWAS summary statistics) to further analysis, making decisions for the variant and gene identification.

The models varied mainly in performance and the number of SNP used (4 to 11,839). Therefore, the best model regarding AUROC and variants was selected to analyze further the markers and genes associated with CD. The best model was generated by LDpred using no SNP imputation (mean AUROC=0.667), with a CD of 1 and including 11,839 SNPs, corresponding to 4,677 genes (**Table A1**). LDpred method is designed and used for prediction rather than variants or gene identification (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015). Then, to facilitate the application of LDpred models to identify variants and genes associated with CD-risk, filtering on effect sizes (re-estimated betas) was performed. All the variants with an effect higher than 0.01 were selected, corresponding to 516 (283 intragenic and 233 intergenic) from 402 genes (**Figure 37**).

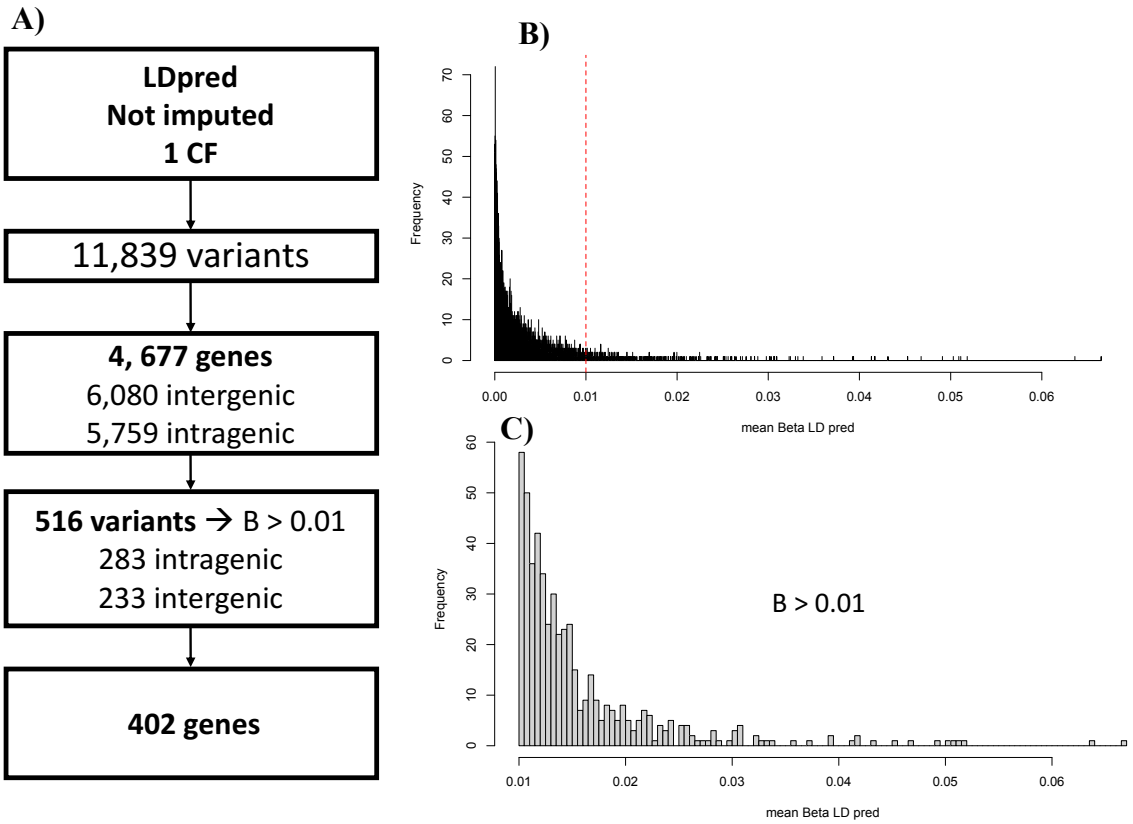

Figure 37. A) Variants and genes within selected LDpred model. B) Distribution of absolute mean values of effect sizes (re-estimated betas) within 10X 1 CF not imputed LDpred models. C) Distribution of absolute mean values of effect sizes (re-estimated betas) higher than 0.01, within 10X 1 CF not imputed LDpred models.

However, for the remaining methods, it was observed that for the not imputed and not clumped data, the best AUROC were obtained for the p-value threshold of $1e^{-3}$. This threshold selects 168 variants in each 10xCV, integrating 418 unique variants among the 10 CVs, which refer to 223 genes (**Table A2**). These variants and genes were also analyzed to back and get more confidence for the results obtained for the LDpred method.

The procedure to annotate variants to genes was performed with R package BioMart (Smedley et al., 2009) and variants or genes of interest were further reviewed within dbNCBI and OpenTargets genetics. Also, GWAS were benchmarked to find if a gene or variant could be considered a novel finding (**Table A2**).

4.5.2 VARIANT IMPORTANCE TO FINDING "NOVEL" MARKERS

Univariate-based methods assume the importance of the variants in a test model is the same as the reached within the discovery set. These models do not weigh the variables (SNPs), thus affecting the replication process. The prediction models for CD risk elucidated here found that even though LDpred was the best method, there was no significant difference between PRS P+T (the univariate-based most common PRS approach) and PRS P+ T and LASSO. However, PRS P+T relies only on the summary statistics from the GWAS, thus cannot aid to distinguish among the 418 variants (SNPs) to rank them and hence make asserted conclusions. To further analyze the variants to identify those contributing the most to the prediction, the best model (LDpred for 1 CF and not imputed data) was analyzed together with the best model for the threshold filtering (LASSO and XGBoost for p-value <1e-3, not imputed and not LD-clumped data). LDpred models are used mainly for prediction (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015); thus, this second model was also reviewed to back and compare the results obtained from the LDpred model.

The mean variants rank, provided by the method at the CF of 1 for LDpred (516 variants and 402 genes) and the threshold 10^{-3} (418 variants and 223 genes) or equivalent for LDpred, was analyzed to generate an average method rank for the multivariate methods. Moreover, variants and genes present in many sets suggest robust participation in the model, highlighting its relevance in CD. Thus, the frequency of relevant variants across the 10X internal subsampling sets was analyzed for LASSO and XGBoost. The most highly relevant variants were in well-known CD genes (**Table A1 and Table A2**). For example, top variants were associated with NOD2 and IL23R, well-known CD genes. Next, the importance of variants (rank) and the occurrences in the best model that contains 10X CV models was

assessed (**Figure 38**). A mean multivariate rank was constructed to facilitate the variants and genes analysis (**Figure 38D**). Here, a novel variant, rs4945943, was identified as relevant for the LDpred model (rank 46) and, on average, for 8 out of the 10 sampling sets for XGBoost and LASSO methods been ranked as the top 18, for the named multivariate rank.

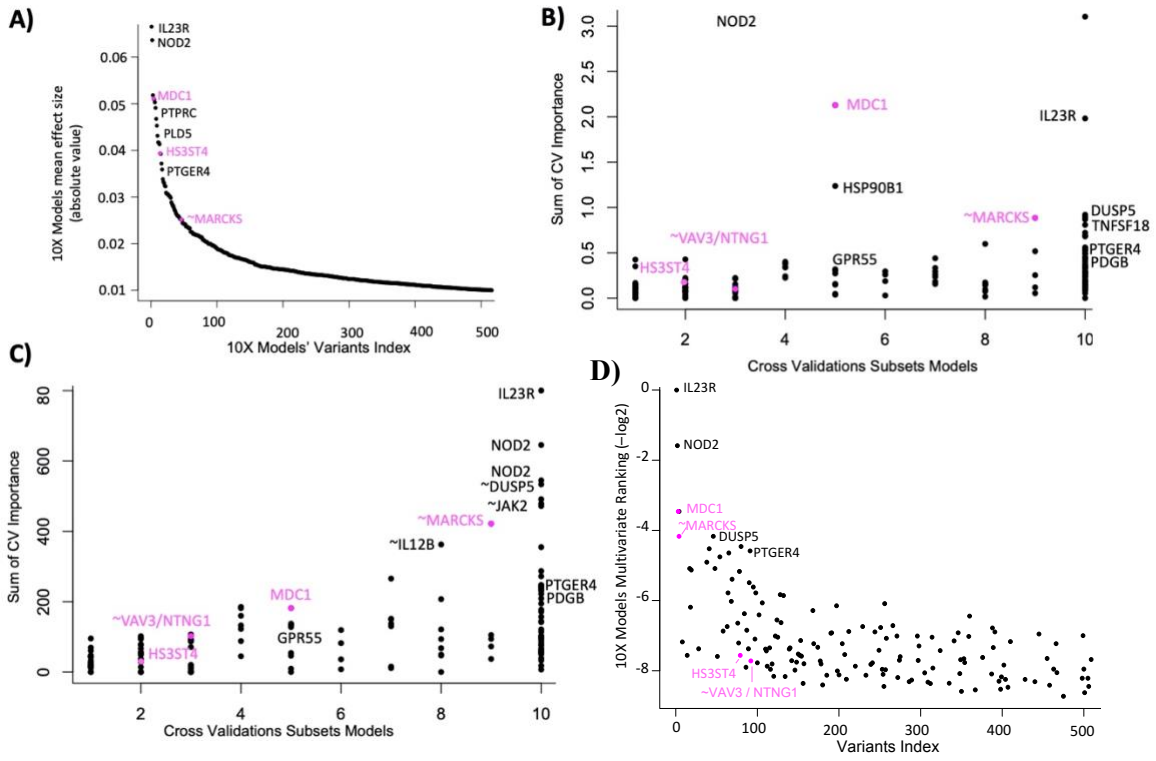

Figure 38. Variants importance and MARCKS. A) Importance of variants among the 10 subsets models relative to the number of appearing models A) LDpred, B) LASSO, C) XGBoost, D) Multivariate rank. Some well-known CD genes are marked in black. The genes with ~ are close to the observed variant. The top 3 non-associated genes in CD and related diseases are highlighted in magenta.

Besides MARCKS, other variants not previously associated with CD or related diseases were noted, such as HS3ST4, VAV3/NTNG1, and MDC1 (**Figure 38**), among others (**Table A1 and Table A2**). However, these variants identified in an average of 2, 3, and 5 CVs from the multivariate models (XGBoost and LASSO) had a mean multivariate rank of 189, 211, and 11, respectively, and might be subject to future analyses. However, due to is low replication

within the cross-validated models, these variants were not subject to subsequent analysis (Genotype stratification and interaction analysis).

4.5.3 rs4945943AS ALLOWS THE IDENTIFICATION OF MARCKS AS A PUTATIVELY NOVEL MARKER

rs4945943 variant is close to coding gene MARCKS. As described below, evidence suggests that MARCKS may be important in CD. First, MARCKS encodes for a Myristoylated alanine-rich C-kinase substrate, regulating proinflammatory cytokine expression in macrophages (Lee et al., 2015). Interestingly, MARCKS is upregulated within a murine model of colitis, where its regulation relies on non-coding RNAs (Mo et al., 2016). Second, rs4945943 is located at chromosome 6 at 460Kbp from the MARCKS gene within a putative enhancer region (**Figure 39**). The GeneHancer tool from GeneCards (Stelzer et al., 2011) shows that this enhancer region appears to affect MARCKS expression and two other long noncoding RNA (lncRNA) sequences at a high probability. MARCKS classical GWAS p-value was marginally significant (mean $p=8.8x10^{-5}$, for the 10X univariate analysis), contrasting the variant relevance in 8 out of 10 subsampling sets suggesting regulation of MARCKS as putative important in CD.

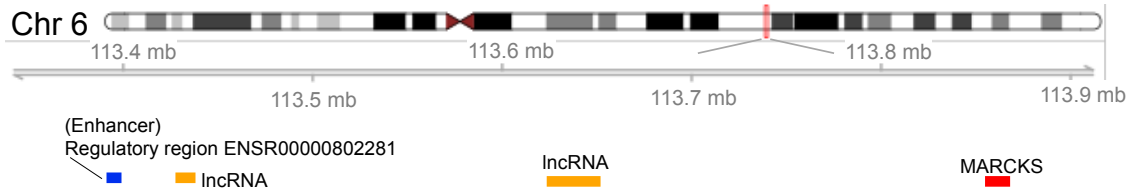

Figure 39. Mapping representation of variant rs4945943 in chromosome 6 (hg38).

Third, an analysis performed in the STRING database (Szklarczyk et al., 2019) using the top CD-associated genes suggests a possible link with well-known CD genes, such as NOD2,

through the protein kinases (PKCA and PKCB) and the NEMO(IKBKG) genes (**Figure 40**). MARCKS gene encodes a rod-shaped protein of 35 kDa, which is susceptible to phosphorylation by protein kinases (i.e., PKC, ROCK) (Amri et al., 2018). MARCKS regulates human neutrophil migration and adhesion, also promoting neutrophil secretion of inflammatory cytokines (Amri et al., 2018). On the other hand, NOD2 activation leads to ubiquitinoylation of NEMO, a key component of the NF-kB signaling complex (Abbott et al., 2004) connected to MARCKS.

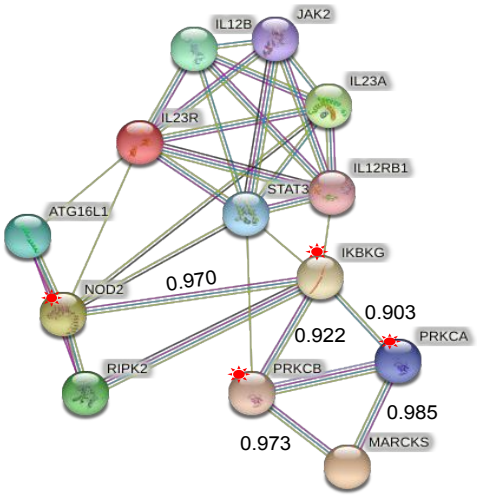

Figure 40. STRING analysis for the top CD associated genes (i.e., NOD2, ATG16L1, IL23R) relative to MARCKS

Fourth, to further challenge rs4945943 polymorphisms attempting to explain its importance in CD and to explore possible links between this association and known CD genes, it was reasoned that CD might present highly diverse genotypes. Thus, surged the question of what the genome-wide associations could be when cases and controls are pre-selected for the same specific rs4945943 polymorphism. This procedure could imply an epistatic effect between the resulting associations and a particular polymorphism. Therefore, a GWAS analysis was performed, comparing controls and cases having AA, AB, or BB polymorphisms in rs4945943. To compensate for decreases in the number of samples, only the 1205 SNPs

whose p-value < 10^{-2} in the entire dataset were used. The number of cases was 251, 1,655, and 2,601, correspondingly to AA, AB, and BB polymorphisms, and the number of controls was 397, 3,154, and 5,643, respectively.

88 significant associations were found ($p < 10^{-6}$) where 40 were originally genome-wide significant in the whole dataset, and 49 were novel rs4945943-dependent calls. Logistic regression was applied to these 88 markers in the entire dataset to test interactions with rs4945943. For the 39 known markers, there were 4 significant interactions ($p<0.05$), all at the IL23R gene located in chromosome 1 (**Table 12**). However, only two interactions were significant after a Bonferroni correction using only independent markers, 4 from the 39 significant SNPs ($p<0.05/4$). For the other 49 rs4945943-dependent calls, there were 12 significant interactions (**Table 13** ($p<0.05$), distributed at the genes PTPN22 (Chr 1), ZNF365 (Chr 10), USP25 (Chr 21), and ADO (Chr 10). However, none reached Bonferroni significance after correcting multiple testing on the 13 independent markers. Except for ADO, the other three genes show significant associations with CD or inflammatory bowel disease in previous studies (de Lange et al., 2017; Liu et al., 2015). Moreover, ADO is located at 300Kbp downstream of the gene ZNF365, thus perhaps representing a CD-associated region.

Table 12. Logistic regression results, for interaction models between rs4945943 genotype and the significant SNPs from the genotype stratification analysis. SNPs originally genome-wide significant in the full dataset

SNP_X	$P_{RS4945943}$	P_{SNPX}	$P_{INTERACTION}$	$B_{RS4945943}$	B_{SNPX}	$B_{INTERACTION}$	$DELTA\ R^2$
1:67681669:T:G	$4.4e^{-6}$	0.029	**0.006**	-0.290	0.160	0.120	0.010
1:67670916:G:A	$8.1e^{-6}$	0.028	**0.010**	-0.280	0.160	0.120	0.009
1:67753508:C:T	0.430	0.120	0.031	-0.044	-0.110	-0.097	0.006
1:67688349:T:C	$3.1e^{-5}$	0.024	0.040	-0.250	0.170	0.091	0.008

Delta R^2. R^2 Change, compared with the logistic model without the interaction and SNP_x terms. Data in bold refer to significance after Bonferroni correction, using independent markers.

Table 13. Logistic regression results for interaction models between rs4945943 genotype and the significant SNPs from the genotype stratification analysis. SNPs novel rs4945943-dependent calls.

SNP_X	$P_{RS4945943}$	P_{SNPX}	$P_{INTERACTION}$	$B_{RS4945943}$	B_{SNPX}	$B_{INTERACTION}$	$DELTA\ R^2$
10:64408367:C:T	$3.2e^{-6}$	0.430	0.008	-0.290	-0.059	0.120	0.002
21:16817938:G:A	$1.2e^{-4}$	0.540	0.034	-0.300	0.049	0.100	0.003
21:16817051:A:G	$1.3e^{-5}$	0.720	0.026	-0.260	0.026	0.100	0.003
10:64445760:T:C	$8.8e^{-5}$	0.750	0.022	-0.320	-0.027	0.120	0.002
10:64398466:C:T	0.730	0.840	0.027	-0.022	0.015	-0.100	0.002
10:64438486:G:C	0.590	0.870	0.036	-0.033	-0.012	-0.095	0.002
21:16813212:T:C	$5.2e^{-6}$	0.940	0.012	-0.280	-0.006	0.110	0.003
21:16812552:C:A	$7.8e^{-6}$	0.970	0.016	-0.280	0.003	0.110	0.003
21:16805220:T:C	$7.7e^{-6}$	0.980	0.016	-0.280	0.002	0.110	0.003
1:114377568:A:G	0.003	0.990	0.050	-0.430	-0.002	0.150	0.002
1:114303808:C:A	0.003	0.990	0.050	-0.430	-0.002	0.150	0.002
10:64445564:A:G	$1.1e^{-5}$	0.990	0.021	-0.260	0.0005	0.100	0.002

Delta R^2. R^2 Change, compared with the logistic model without the interaction and SNP_x terms.

Fifth, to further review the importance of rs4945943, perhaps modulating MARCKS, an analysis of the rs4945943 region was performed, specifically in the data statistics from the de Langer *et al.* 2017 cohort (authors dataset in **Figure 41**). In the de Langer *et al.* study, the

rs4945943 marker was not evaluated in the meta-analysis presumably because it was not present in all datasets used for the meta-analysis but reported in the specific dataset generated by de Langer *et al.* and used in this study (the rs4945943 marker position was added to the **Figure 41**, to facilitate the comparison). It was observed that around the enhancer region, other SNPs also showed a potential association with CD risk even at higher significance (**Figure 41**). Thus, in summary, the region is important in the de Langer *et al.* cohort but lost in the meta-analysis and therefore unnoticed (**Table 14**).

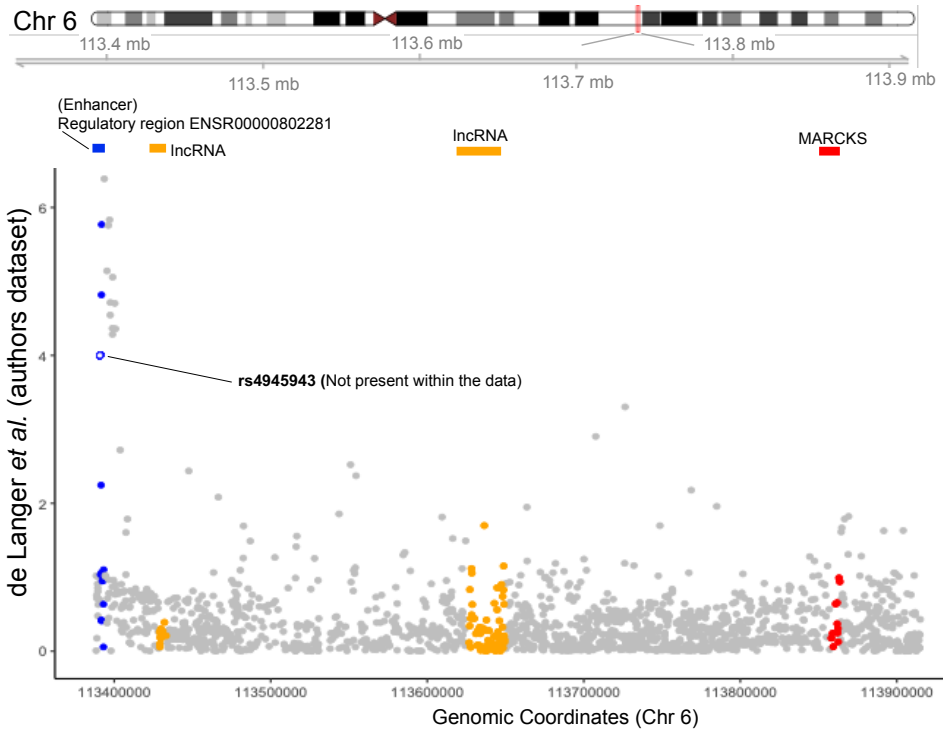

Figure 41. Mapping representation of variant rs4945943 in chromosome 6 (hg38). Region of enhancer, ENSR00000802281, lncRNA regions, and MARCKS genes are highlighted. De Langer et al. univariate summary statistics of 1,701 SNPs around rs4945943 region, for de Langer et al. 2017 authors dataset (univariate analysis). The authors' dataset represents the specific patients assayed by de Langer et al. (instead of the meta-analysis). rs4945943 is not present in the de Langer et al. meta-analysis (presumably because it was not present in all datasets from the meta-analysis) but present in the assayed data generated by de Langer et al. and used in this study (4,474 Cases and 9,500 Healthy Controls).

Table 14. Summary statistics from de Langer et al. 2017, for markers within chromosome 6 region tagged by rs4945943 in this study.

Position (hg37)	P.value	Pval_IBDseq	Pval_IIBDGC	Pval_GWAS3	Region
113713065	0.0651400	0.643083	0.4158	$1.52e^{-5}$	Enhancer ENSR00000802281
113713163	0.0353600	0.633853	0.4060	$1.70e^{-6}$	Enhancer ENSR00000802281
113714876	0.0067800	0.810792	0.5206	$4.10e^{-7}$	Intergenic region
113716570	0.0516900	0.695421	0.4177	$7.20e^{-6}$	Intergenic region
113717459	0.0295800	0.713358	0.4422	$1.75e^{-6}$	Intergenic region
113718188	0.0136500	0.922356	0.4975	$1.47e^{-6}$	Intergenic region
113718667	0.0762200	0.650426	0.4343	$2.85e^{-5}$	Intergenic region
113718725	0.0665500	0.663029	0.4345	$1.93e^{-5}$	Intergenic region
113720008	0.0862700	0.647798	0.4340	$4.31e^{-5}$	Intergenic region
113720073	0.0917500	0.637362	0.4344	$5.20e^{-5}$	Intergenic region
113720180	0.0210200	0.928552	0.5629	$8.76e^{-6}$	Intergenic region
113721570	0.0759100	0.630312	0.4011	$1.98e^{-5}$	Intergenic region
113722011	0.0232600	0.673220	0.8838	$4.37e^{-5}$	Intergenic region
113734952	0.0001628	0.646900	0.9244	$3.56e^{-8}$	Intergenic region

Figure 42 represents the SNPs around the rs4945943 region, for de Langer et al. 2017 data in this study, specifically in the 40% samples GWAS, performed at the initial stage of the analysis. Whereas **Figure 43** shows the gpmean of –log10 p-value of 13 SNPs around rs4945943 region, for 60% of data, from the 10X CV data, with the standard deviation of p-values.

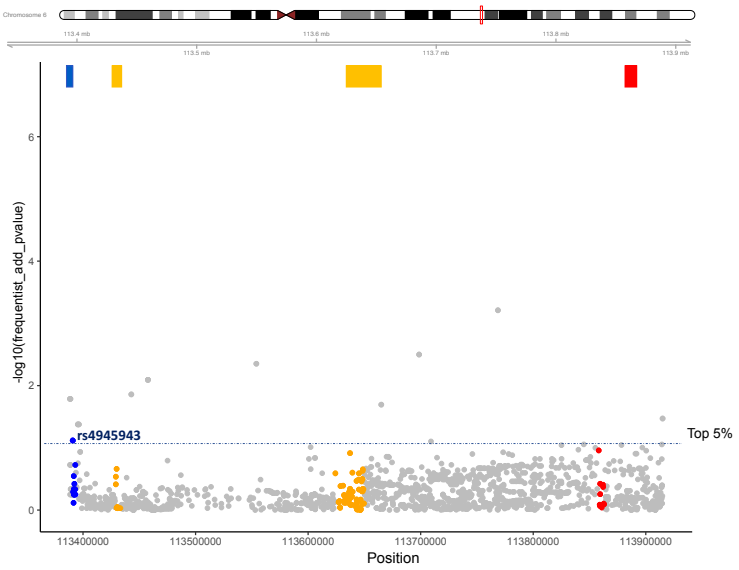

Figure 42 –log10 p-value of 1,701 SNPs around rs4945943 region, for de Langer et al. 2017 data (univariate analysis). Region of enhancer, ENSR00000802281, lncRNA regions, and MARCKS genes are highlighted. rs4945943 is not present in the original de Langer (metanalysis) analyzed data. 4,474 Cases and 9,500 Healthy Controls.

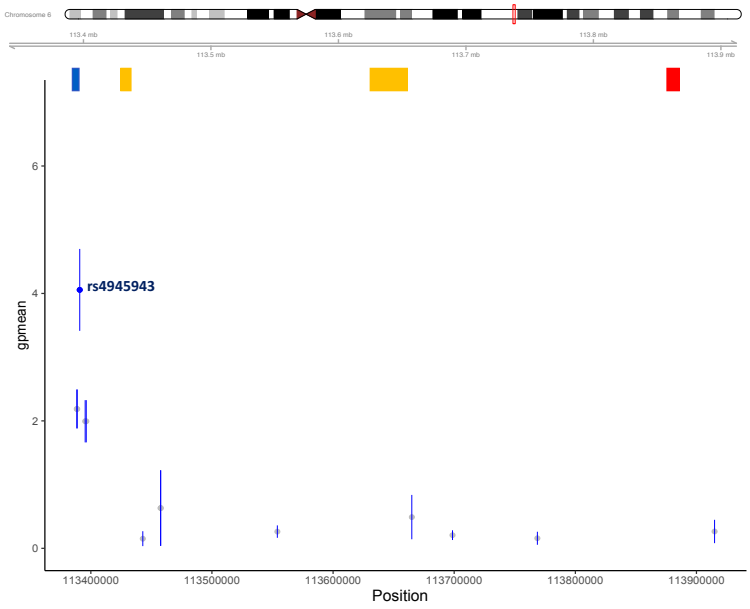

Figure 43. gpmean of –log10 p-value of 13 SNPs around rs4945943 region, for 60% of data, from the 10X CV data. Region of enhancer, ENSR00000802281, lncRNA regions, and MARCKS genes are highlighted. 2704 Cases and 5516 Healthy Controls.

4.5.4 RS9262151 IDENTIFICATES MDC1 AS A LESS ROBUST MARKER FOR CD RISK

The rs9262151 variant is located within the MDC1 gene. The same analysis performed to rs4945943 (~MARCKS) was applied to this SNP to evaluate the shreds of evidence for its importance related to CD risk. MDC1 classical GWAS p-value was marginally significant (mean $p=5x10^{-4}$, for the 10X univariate analysis). This SNP showed relative importance for Ldpred and LASSO models' variant was only identified in 4 out of 10 subsampling sets, showing less replication evidence than rs4945943. The analysis performed in the STRING database (Szklarczyk et al., 2019) using the top CD-associated genes suggests a possible link with well-known CD genes, such as ATG16L1, through the RB1CC1 and TP53 genes, affecting autophagy, one of the processes related to CD development and which states that CD arises from a defective innate immune response to enteric bacteria (Henderson & Stevens, 2012).

Also, a GWAS analysis was performed, comparing controls and cases having AA, AB, or BB polymorphisms in rs9262151, for the 1205 SNPs with a p-value $< 10^{-2}$ in the entire dataset. The number of cases was 5, 340, and 13,357, correspondingly to AA, AB, and BB polymorphisms, and the number of controls was 397, 3,154, and 5,643, respectively. 134 significant associations were found ($p < 10^{-6}$) where 100 were originally genome-wide significant in the whole dataset, and 34 were novel rs9262151-dependent calls. Logistic regression was applied to these 134 markers in the entire dataset to test interactions with rs9262151. Only two interactions within TNXB were found ($p<0.05$), and both were significant after Bonferroni correction (p-value 1.1e-6 and p-value 1.6e-6). However, due to the small number of samples for nG0 and nG1, these results are not reliable. Finally, at

summary statistics from De Lange et al., 2017 a variant located 2kb, rs35743249, was associated with CD with a p-value of $3e^{-4}$.

4.5.5 GENE ENRICHMENT AND PATHWAYS ANALYSIS

To further validate the variants and genes identified with the established approach and to further investigate the pathways and biological terms associated with those genes, a functional annotation of the variants in the two models (LDpred 1 CF and LASSO and XGBoost as p-value $<1e^{-3}$ filtering) were performed. The variants were re-annotated using BioMart (Smedley et al., 2009), dbNCBI, and OpenTargets platform (Carvalho-Silva et al., 2019; Sherry et al., 2001) and functionally analyzed using DAVID and ENRICHR (Ashburner et al., 2000). The analysis was performed on the 402 1CF LDpred models genes and the 223 p-value $<1e^{-3}$ genes.

Similar functional terms were collapsed and weighted for each gene depending on their relative presence among collapsed terms to summarize the findings. Next, each functional analysis is described.

4.5.5.1 P-VALUE <1E-3 MODELS (223 GENES)

Among the terms identified in the functional analysis for the p-value filtered genes are those related to *inflammatory processes*, *response to organic stimulus, transport, innate immune response, cell migration, signaling pathways such as cytokines, TNF, jak-stat, AGE-RAGE,* and *adhesion*, also with the *regulation of transcription* and diseases as *autoimmune* and *inflammatory bowel disease*. It was found that the genes most enriched were those associated with the *inflammatory bowel diseases*, as expected, and with the biological process of the *immune response*, regulation of immune response, *cytokine, jak-stat, and AGE-RAGE*

signaling, which are related to the nature of Crohn's disease. Also, the molecular function involving *DNA and protein binding*, together with *regulation of gene expression, cellular adhesion, transport,* and *immune response*, shows a general overview of the elements related to the etiology of the disease. **Figure 44** shows the hierarchical clustering of the terms and highlights genes commonly associated with CD, such as NOD2, ATG16L1, IL23R, and PTGER4 (**Figure 44**, red lines). This functional analysis observed that CD, UC, and IBD were collapsed together because they were clustered together at the disease's hierarchical clustering, meaning that the variants had a less stringent p-value for CD risk (p-value <1e-3) are less specific for CD. On the other hand, MARCKS and MDC1 were identified as associated with the focal-adhesion term from the GO Cellular Component Adhesion cluster. Thus, this gene enrichment analysis helped validate the standard approach to identifying terms enriched by genes identified through GWAS and the proposed random sampling and multivariate setting methodology.

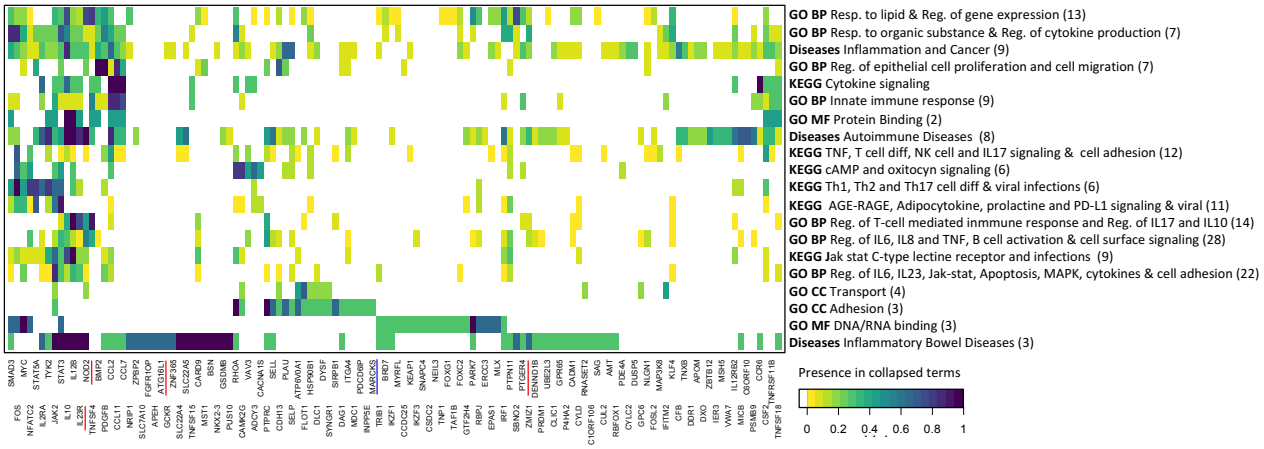

Figure 44. Functional analysis of P-value <1e-3 filtered genes. Columns show genes, and rows refer to collapsed terms from DAVID and ENRICHR analysis for GO, KEGG, and Diseases. Gene names were divided into two labeling rows for clarity. Red lines highlight ATG16L1, IL23R, and NOD2, whereas the blue lines highlight MARCKS. Numbers in parenthesis represent the number of terms collapsed within each general term.

4.5.5.2 LDPRED MODELS WITH 1 CF (402 GENES)

LDpred is used for trait prediction rather than to identify novel variants or genes (Vilhjálmsson, Yang, Finucane, Gusev, Price, et al., 2015); this is due to its mechanism where up to 2,000,000 variants can be fitted to generate a prediction. However, this study aimed to use the results from LDpred prediction for a variant/gene analysis. From the 11,839 markers within the LDpred 1 CF, there were 516 SNPs from 402 genes, which had an effect size higher than 0.01. These genes were used to perform second functional enrichment analysis.

Among the terms identified in the functional analysis for LDpred 1 CF genes are those related to the *regulation of immune response (interleukins 10, 12 and 17, B and T cell),* a*ctivation of cellular immunity, regulation of cytokines and TNF, viral infection & MAPK and TNF signaling, cell migration and phagocytosis & NK cell cytotoxicity, bacterial infection and immune response, adhesion*, *transport,* and *ion channels*, also with diseases as *autoimmune, coronary* and *inflammatory bowel disease*. Here, the genes most enriched were those associated with the *inflammatory bowel diseases and autoimmune diseases*, and with the pathways *cell migration and phagocytosis & NK cell cytotoxicity* and *viral infection & MAPK and TNF signaling* and the biological process of the *cell proliferation and cytokine signaling*, which are related to the nature of Crohn's disease. **Figure 45** shows the hierarchical clustering of the terms and highlights genes commonly associated with CD, such as NOD2, ATG16L1, IL23R, and PTGER4 (**Figure 45**, red lines). However, in this functional analysis CD, was only collapsed with IBD, being UC collapsed with other diseases (Psoriasis, Behcet syndrome, and viral infections) at the diseases hierarchical clustering, meaning that the variants selected with LDpred models are more specific for CD, thus, being able to distinguish between CD and UC. On the other hand, MARCKS and MDC1 were identified. Still, only MARCKS had a biological interpretation with a significant term (*Fc*

gamma R-mediated phagocytosis within *KEGG Cell migration and phagocytosis & NK cell cytotoxic* cluster). MDC1 was enriched only for *autoimmune diseases*. Thus, this gene enrichment analysis helped validate genes identified through GWAS and the proposed random sampling and multivariate setting methodology.

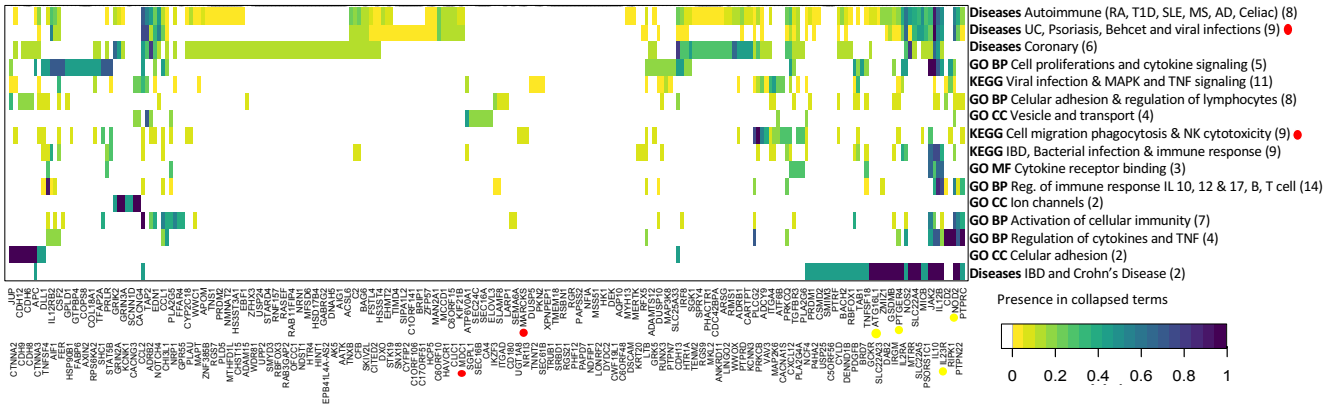

Figure 45. Functional analysis of LDpred 1 CF genes. Columns show genes, and rows refer to collapsed terms from DAVID and ENRICHR analysis for GO, KEGG, and Diseases. Gene names were divided into two labeling rows for clarity. Yellow circles highlight ATG16L1, IL23R, and NOD2, whereas the red circles highlight MARCKS and MDC1. Numbers in parenthesis represent the number of terms collapsed within each general term.

4.5.6 VALIDATION OF LDPRED IN NIDDK IBD GENETICS CONSORTIUM DATASET

The SNP data from the validation dataset was converted from the hg35 genome version to the correspondent in the hg37 version, then used for the CD risk prediction approach. Among the 11,839 SNPs from LDpred models, 2,360 SNPs were genotyped at the NIDDK CD dataset. And, among the 418 SNPs for the $1e^{-3}$ threshold for LASSO, XGBoost, Random Forest, and PRS P+T, 88 were genotyped in the validation dataset. Thus, the 10 cross-validated CD risk prediction models were evaluated in the NIDDK dataset (**Table 15 and Figure 46**).

Table 15. Mean AUROC of multivariate methods and PRS P+T for the validation dataset. Data for 2,360 common genotyped SNPs. SD standard deviation.

Method	Mean	SD
PRS P+T	0.588	0.016
LASSO	0.516	0.017
XGBoost	0.526	0.012
Random Forest	0.526	0.013
BSWIMS	0.512	0.016
LDpred	**0.632**	0.009

LDpred method had the highest AUROC (0.632 ± 0.009), which is highly similar to that reached in the test evaluation in the UKIBDGC and UK10K GWAS dataset. The best AUROC was obtained for the 10th model for LDpred with an AUROC of 0.644.

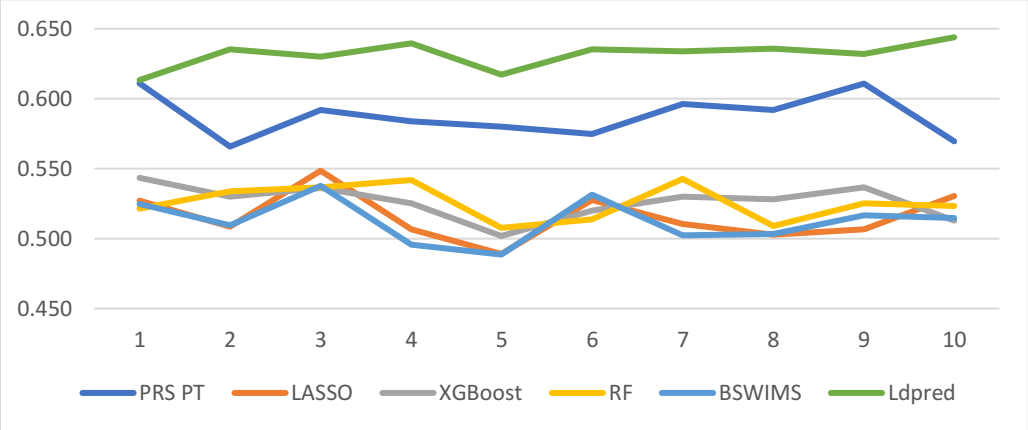

Figure 46. AUROC of multivariate methods and PRS P+T for the validation dataset. Data for 2,360 common genotyped SNPs.

These results confirmed that the LDpred model and this approach are robust and valuable. Also, in the validation dataset, the rs4945943 (~MARCKS) variant ranked top, having a mean rank of 5, from the 2360 SNPs of the 10X cross-validate models, thus validating its contribution to the CD-risk prediction with the LDpred model.

A second approach was applied to determine if adding additional markers closest to the non-genotyped variants could improve the prediction. **Table 16** and **Figure 47** show that even though LDpred again reached the highest performance among all the methods, it did not pass the one achieved by using only the genotyped markers.

Table 16. Mean AUROC of multivariate methods and PRS P+T for the validation dataset. Data for 2,360 common genotyped SNPs and 9,468 nearest-SNPs. SD standard deviation.

Method	Mean	SD
PRS PT	0.543	0.022
LASSO	0.515	0.014
XGBoost	0.510	0.019
Random Forest	0.479	0.010
BSWiMS	0.504	0.013
LDpred	0.583	0.010

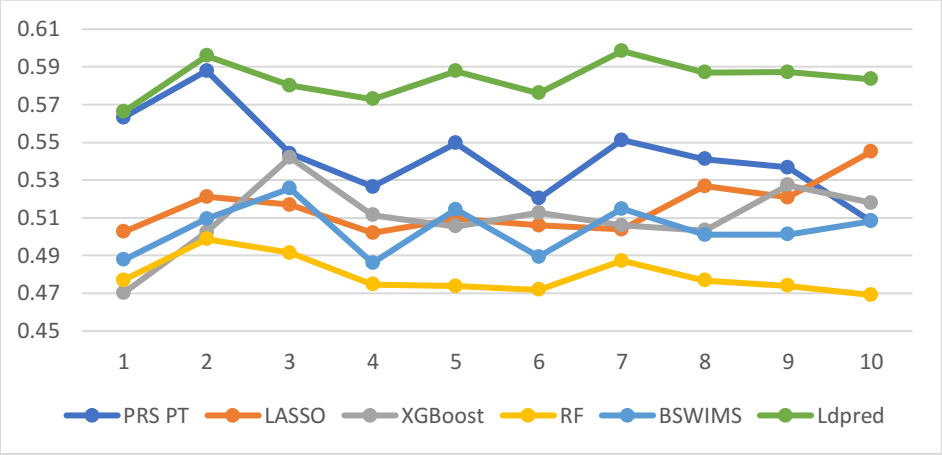

Figure 47. AUROC of multivariate methods and PRS P+T for the validation dataset. Data for 2,360 common genotyped SNPs and 9,468 nearest-SNPs.

For this, 20 LDpred analyses were performed by each 50 pb up to 1000 pb. A gradient of nearest-SNPs distance was evaluated to determine if there was an optimal distance between the nearest marker and the non-genotyped SNP, which could help to improve the AUROC of

the prediction CD risk models in the validation data. However, **Figure 48** shows that even using a gradient of nearest markers, the AUROC was not improved.

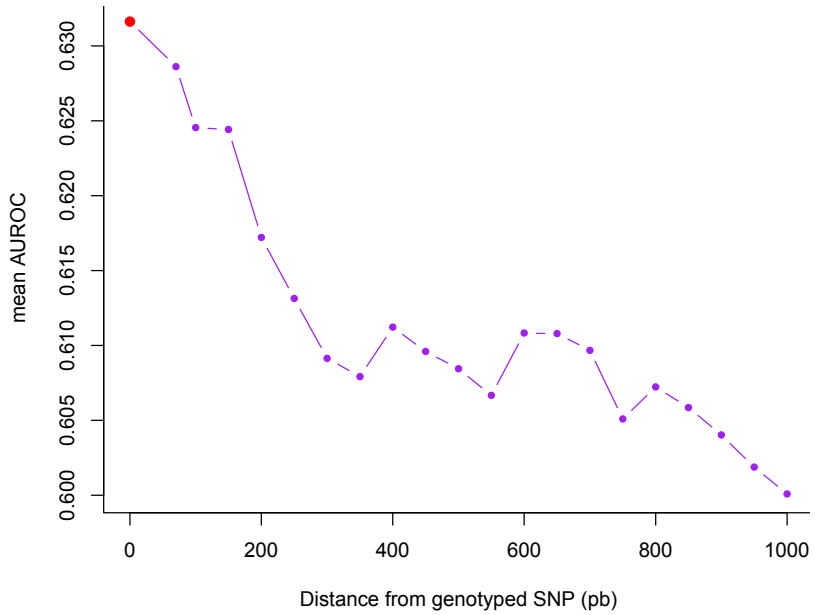

Figure 48. LDpred Mean AUROC for the gradient of the nearest genotyped marker from the non-genotyped SNP within the validation dataset.

5. CHAPTER 5: DISCUSSION AND CONCLUSIONS

5.1 DISCUSSION

5.1.1 ROBUST ESTIMATION OF MULTIVARIATE POTENTIAL: LDPRED APPEARS TO BE THE BEST MODEL

The necessity of finding models to predict disease based on genotype has incremented over the years, and even that the PRS has been proposed to understand the genetic risk of developing a disease, its clinical application remains limited because it is known that genetic factors only contribute part of the disease risk, and more data is needed to allow PRS development (Wray et al., 2021).

Nevertheless, attempting to improve the genetic risk prediction, several methods have been proposed to replace or improve PRS, such as EB-PRS (Shuang Song, Wei Jiang, Lin Hou, 2020), LDpred (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015), PRS-CS (Ge et al., 2019), and multivariate models, for example, regression-based as LASSO, RIDGE, and ElasticNET (Romagnoni et al., 2019; Wei et al., 2013) or classification-based as SVM and Random Forest (Goldstein et al., 2010).

GWAS has allowed the identification of common variants for complex diseases. However, the contribution of rare or less frequent variants and solving the missing heritability remains a challenge (Eichler et al., 2010). Crohn's disease is a well-studied complex trait, with an annual incidence of 20 cases per 100,000, and is affected by a combination of environmental and genetic factors. However, its exact etiology is still unknown (Liu & Anderson, 2014). CD has a high heritability derived from pooled twin studies (0.75), which contrasts with the reached by GWAs (0.37) (Gordon et al., 2015). This, together with the importance of finding

novel targets for treatment and drugs development, makes CD an eligible model to test a multivariate methodology.

The methods that have been implemented in CD GWAS datasets are LASSO (Kooperberg et al., 2010; Newcombe et al., 2019; Wei et al., 2013), gradient boosting (Romagnoni et al., 2019), SVM (Mittag et al., 2015), KNN (Mittag et al., 2015), MLP (Mittag et al., 2015), Bayesian methods (G.-B. Chen et al., 2017) and random forest (Mittag et al., 2015). These studies vary in particularities of the analysis, such as sample size, SNPs platforms, imputation strategies, reduction of features, and methods applied. Thus, for CD, in particular, there is no agreement on which method could be the best to predict the disease risk under similar methodological conditions. Therefore, different multivariate methods were compared to generate models to predict CD risk in a GWAS dataset. For this, CD GWAS data was requested from the European Genome-phenome Archive (EGA), followed by an imputation process (to increment the data features with non-independent information), a pre-filtering process (LD clumping and p-value filtering) was implemented on 40% of samples. Then a set of markers of a specific p-value threshold were fitted in multivariate and univariate-based models for the remaining 60% of samples.

These results showed that LDpred with no imputation yielded the most efficient performance (AUROC of 0.667 with 1 as a causal fraction using 11,839 SNPs) than univariate-based models such as common PRS and other multivariate models with GWAS data. The performance reported here is the highest for the de Lange et al. 2017 dataset. It was observed, as expected, that imputing genotypes to increase the number of features slightly improved the performance of the models (Bargelloni et al., 2021) but only on p-value thresholds less than 1e-4. Random forest did not increase the performance under imputation and clumping; additionally, the number of model markers increases, complicating interpretations. Also,

decreasing the number of features through LD clumping proved detrimental, mainly for the non-imputed dataset. Even if the pruning and thresholding approach is simple and computationally efficient, this approach (LD clumping) discards some information that could be useful (Paré et al., 2017).

PRS usually yields performances around 0.65 for diseases with SNPs of strong effects (i.e., age-related macular degeneration and Crohn's disease) (Richardson et al., 2018). This performance for CD prediction was validated with this experimental setting.

LDpred showed the best AUROC for the not imputed data (mean AUROC=0.667), using a causal fraction of 1 (~11k variants). The results for this method did not drastically change along with the number of features (imputed sets) or variations in causal fractions. Ldpred was not tested with LD-clumping because LDpred made internal LD adjustments (Paré et al., 2017). LDpred generally improves predictions over traditional PRS (Vilhjálmsson, Yang, Finucane, Gusev, Zheng, et al., 2015), which was also observed here. The critical difference analysis showed that Ldpred ranked as the best model with no significant difference with PRS P+T, which was also not different than LASSO. This showed that the multivariate methods were, in fact, equivalent to PRS, validating its subsequent use to the identification of variants associated with CD.

These results also showed a detrimental tendency for the AUROC when adding SNPs with a p-value less than 1e-2. This result confirms that adding many SNPs with a less suggestive association with the disease can be disadvantageous for the prediction.

It was expected to have better results for LASSO models based on what is reported in previous research where this method had reached AUROC from 0.64 (Kooperberg et al., 2010) to 0.80 (Wei et al., 2013). However, the sample size of these studies was much larger (> 40,000 subjects), and the SNP used was more specific (Immunochip). Instead, an AUROC

of 0.621 was reached for LASSO, with a univariate threshold of 10-3, using 168 SNPs for the not imputed data and an AUROC of 0.621 for 1,555 SNPs, with a univariate threshold of 10-4. Alternatively, SVM, RBF, KNN, RF, and MLP have been reported to be less efficient in predicting CD with AUROC around 0.59 (Mittag et al., 2015) for WTCCC data. These performances were improved using the XGBoost method (mean AUROC 0.608 and 0.615 for imputed and not imputed datasets) and LDpred (mean AUROC 0.656 and 0.667 for imputed and not imputed datasets).

The random permutation experiment reflected that the markers identified within the GWAS were, in fact, associated with the phenotype rather than a random association. The validation analysis with the NIDDK IBD Genetics Consortium data confirmed the application of the LDpred model with a different dataset, reaching similar performance (AUROC = 0.634). This approach differs from other studies by introducing a robust 10x cross-validation estimation for selecting variants in models and a preselection strategy on different samples. At the same time, the standard methods use only one set of samples and no cross-validation (Yan et al., 2021). Thus, this approach considers the subsets' variability and the samples' influence on the feature's significance.

5.1.2 MODEL VALIDATION ANALYSIS: MULTIVARIATE RANK AND "NOVEL" MARKERS

The different methods evaluated in this research had different techniques to assign a value to each SNP/variant. BSWiMS, LASSO, and LDpred had a re-estimation of the beta or effect size values, and RF and XGBoost had a measure of importance. The best model, generated by LDpred using no SNP imputation. Because of, LDpred method is designed and used for prediction rather than variants or gene identification (Vilhjálmsson, Yang, Finucane, Gusev,

Zheng, et al., 2015), to facilitate the application of LDpred models to identify variants and genes associated with CD-risk, filtering on effect sizes was performed and 516 variants, corresponding to 402 genes, with an effect size >0.01 were selected. Also, the variants corresponding to the best p-value threshold model (p<1e-3) were also evaluated, which integrated 418 unique variants among the 10 CVs from 223 genes.

A rank for the variants was established for each method, and a mean multivariate rank was finally generated by merging the ranks information. This, to provide a robust validation for the SNPs identification. Variants and genes present in many sets and methods suggest robust participation in the model, highlighting its relevance in CD. Most variants showing high relevance were located in well-known CD genes such as NOD2 (rank 3), IL23R (rank 1), and PTGER4 (rank 30) (de Lange et al., 2017; Liu et al., 2015; Michail et al., 2013). Nevertheless, there are other less-studied genes, but have also been reported to be associated with CD and are annotated at Open Targets as IL12B (rank 57), TNFSF18 (rank 25) and, DUSP5 (rank 78), PDGFB (rank 83) (de Lange et al., 2017; Liu et al., 2015) among others.

However, other genes had been suggestively associated with either CD or IBD (p-value <10-5) in previous studies (de Lange et al., 2017; L et al., 2012; Liu et al., 2015). Some of these were highlighted in this prediction model, such as GPR55 (rank 132), whose expression is different among CD patients (Włodarczyk et al., 2017), PLD5, which encodes the phospholipase D family member 5 (rank 208), which has found CD association by using neighborhood information (Yang et al., 2011). The RNA-binding protein RBFOX1 (rank 219), which has been linked to CD (Elding et al., 2013); and TSBP1 (testis expressed basic protein 1 from the C6orf10 region, rank 286), which has been basally associated with CD and RA (Zheng & Rao, 2015).

The multivariate rank highlighted a novel variant, rs4945943, which was identified as relevant for the LDpred model (rank 46) and, on average as the top 18, for the named multivariate rank. This SNP is located within a regulatory region, an enhancer, which could putatively affect the MARCKS gene. MARCKS gene encodes a rod-shaped protein of 35 kDa, which is susceptible to phosphorylation by protein kinases (i.e., PKC, ROCK) (Amri et al., 2018). The suppression of MARCKS expression in macrophage cell lines blocks LPS-induced expression of TNF-α at the transcriptional level (Lee et al., 2015). Also, these genes have been suggested to contribute to the tumorigenesis of colorectal carcinomas (Kim et al., 2002). Recently, it has been shown that a miRNA regulates MARCKS expression in a model of colitis in mice (Mo et al., 2016). When reviewing the gene network for the top associated CD-genes NOD2, IL23R, and ATG16L1 with MARCKS, it was found that MARCKS links to NOD2 through the interaction between protein kinases and NF-kappa-B essential modulator (IKBKG known as NEMO gene). NOD2 activation leads to ubiquitinylation of NEMO, a key component of the NF-kB signaling complex (Abbott et al., 2004). Also, when testing the interaction for rs4945943, a Bonferroni significant interaction with SNPs markers of IL23R was found. These findings point to a MARCKS enhancer with the evidence of expression changes in colitis, and its links to other CD-related genes suggest that regulation of the expression of MARCKS is critical in CD. Although the function of MARCKS is still not well understood, the above pieces of evidence strongly suggest that MARCKS play a role in CD. Future experimental validation would also be necessary.

There were two additional polymorphisms linked to genes that have not been associated with CD, rs8050730 close to HS3ST4 (univariate p value= 2.73e-3), rs11185129 close to VAV3 (univariate p value= 3.78e-3), and rs9262151 a missense variant for MDC1 (univariate mean p-value=5e-4). HS3ST4 encodes the enzyme heparan sulfate D-glucosaminyl 3-O-

sulfotransferase 4 (rank 189) and is considered a pro-tumoral gene for colon cancer (Denys & Allain, 2019). VAV3 (Vav Guanine Nucleotide Exchange Factor 3, rank 211) is an oncogene expressed in colorectal cancer whose overexpression could dysregulate the expression of cell cycle control by activating the PI3K-AKT signaling pathway (Uen et al., 2015). MDC1 is a mediator of DNA damage checkpoint (rank 11), determining cell survival fate. MDC1 is expressed lowly in various cancers, including lung cancer, breast carcinomas, and gastric carcinoma (Bo et al., 2014). Also, MDC1 is considered a potential therapeutic target for diagnosing and treating human gastric cancer (Qin et al., 2018). For MDC1, there is a polymorphism at ~2Kb, associated with CD in De Lange et al., 2017. Also, opentargets reports an association between MDC1 and Ulcerative Colitis, identified within a Japanese population (Asano et al., 2009).

However, these last polymorphisms were not confirmed by all multivariate methods by being selected in fewer CVs compared to rs4945943 (2, 3, and 5 CVs, respectively).

5.1.2.1 MODEL VALIDATION: GENE ENRICHMENT AND PATHWAYS ANALYSIS

The genetic overlap between CD and other immune diseases has been reported in several studies (Liu & Anderson, 2014), mainly due to GWAS, where UC, type 1 diabetes, coeliac disease, or rheumatoid arthritis are among the immune diseases with a reported genetic overlap with CD (Zhernakova et al., 2009).

Functional annotation of the 402 genes highlighted in this study found the pathways of *inflammatory bowel disease, autoimmune diseases, regulation of immune response, activation of cellular immunity, signaling pathways, cell migration, and phagocytosis, NK cell cytotoxicity, adhesion,* and *transport*. This analysis shows CD-risk genes linked to well-known inflammatory processes (Feuerstein & Cheifetz, 2017) validating this strategy and the

application of multivariate models to identify genes associated with CD-risk. In this functional analysis CD, was only collapsed with IBD, being UC collapsed with other diseases (Psoriasis, Behcet syndrome, and viral infections) at the diseases hierarchical clustering, meaning that the variants selected with LDpred models are more specific for CD; thus, being able to distinguish between CD and UC.

For the functional annotation of the 223 genes highlighted by the p-value threshold, the following terms were found: pathways of *inflammatory bowel disease*, *autoimmune diseases*, *signaling pathways, adhesion, transport, immune response, and regulation of immune response* and *gene expression* were identified. This analysis showed more general terms related to the immune response. Here, CD, UC, and IBD collapsed together, meaning that the variants had a less stringent p-value for CD risk (p-value <1e-3) are less specific for CD.

5.2 CONCLUSIONS

LDpred performed better in predicting CD-risk than other multivariate and standard PRS analyses. Also, multivariate methods allowed the identification of markers with their feature importance ranking. rs4945943 SNP, putatively connected to MARCKS, contributed to the CD-risk prediction.

The hypothesis of this research was partially achieved (75%) because the prediction performance estimated for CD risk cannot be compared with the current literature since the data set used for this investigation has not been used for other prediction analyses yet. Also, the sample size and the differences in SNP platforms difficult the comparison with other multivariate prediction methods. Yet, the performances achieved in this research are better than the ones reported in the literature for datasets of similar size to the dataset used in this investigation. However, this methodology successfully identified variants previously

associated with CD and highlighted putatively "novel" markers, which would need additional experimental validation.

5.3 FUTURE WORK

To improve the performance of the prediction models, parameter optimization, random search approaches, and neural networks are strategies that could further be directed in CD GWAS data.

This methodology was intended to be applied to a Mexican dataset of T2D (Type 2 Diabetes). However, the limitations of the sample size of the data, reflected in a lack of statistical power, limited the application of the methodology. Nevertheless, the methods proposed here can be applied to other complex diseases, such as T2D, which could be addressed in future work with a dataset of proper dimensions.

6. APPENDIX

Table A1: 402 Genes from LDpred models

Table A2: 418 Genes from p-value < 1-3 models

7. BIBLIOGRAPHY

A Martinez-Torteya, Alanis, I., & Tamez-Pena, J. (2018). FeatuRE Selection Algorithms for Computer-Aided Diagnosis: an R package. *Submitted*, .

Abbott, D. W., Wilkins, A., Asara, J. M., & Cantley, L. C. (2004). The Crohn's disease protein, NOD2, requires RIP2 in order to induce ubiquitinylation of a novel site on NEMO. *Current Biology*, *14*(24), 2217–2227. https://doi.org/10.1016/j.cub.2004.12.032

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A Review. *Int. J. Advance Soft Compu. Appl*, *7*(3).

Alqudah, A. M., Sallam, A., Stephen Baenziger, P., & Börner, A. (2020). GWAS: Fast-forwarding gene identification and characterization in temperate Cereals: lessons from Barley – A review. In *Journal of Advanced Research* (Vol. 22, pp. 119–135). Elsevier B.V. https://doi.org/10.1016/j.jare.2019.10.013

Amri, M. El, Fitzgerald, U., & Schlosser, G. (2018). MARCKS and MARCKS-like proteins in development and regeneration. *Journal of Biomedical Science*, *25*(1). https://doi.org/10.1186/S12929-018-0445-1

Asano, K., Matsushita, T., Umeno, J., Hosono, N., Takahashi, A., Kawaguchi, T., Matsumoto, T., Matsui, T., Kakuta, Y., Kinouchi, Y., Shimosegawa, T., Hosokawa, M., Arimura, Y., Shinomura, Y., Kiyohara, Y., Tsunoda, T., Kamatani, N., Iida, M., Nakamura, Y., & Kubo, M. (2009). A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nature Genetics*, *41*(12), 1325–1329. https://doi.org/10.1038/NG.482

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. In *Nature Genetics* (Vol. 25, Issue 1, pp. 25–29). https://doi.org/10.1038/75556

Bargelloni, L., Tassiello, O., Babbucci, M., Ferraresso, S., Franch, R., Montanucci, L., & Carnier, P. (2021). Data imputation and machine learning improve association analysis and genomic prediction for resistance to fish photobacteriosis in the gilthead sea bream. *Aquaculture Reports*, *20*, 100661. https://doi.org/10.1016/J.AQREP.2021.100661

Baumgart, D. C., & Sandborn, W. J. (2012). Crohn's disease. *The Lancet*, *380*(9853), 1590–1605. https://doi.org/10.1016/S0140-6736(12)60026-9

Behravan, H., Hartikainen, J. M., Tengström, M., Kosma, V. –M, & Mannermaa, A. (2020). Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Scientific Reports 2020 10:1*, *10*(1), 1–16. https://doi.org/10.1038/s41598-020-66907-9

Behravan, H., Hartikainen, J. M., Tengström, M., Pylkäs, K., Winqvist, R., Kosma, V.-M., & Mannermaa, A. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-31573-5

Besag, J., & Clifford, P. (1991). Sequential Monte Carlo p-values. *Biometrika*. https://doi.org/10.1093/biomet/78.2.301

Bo, W., Lisha, Z., Fuman, Q., Wenxiang, F., Jieqiong, D., Yifeng, Z., Jiachun, L., & Lei, Y. (2014). A newfound association between MDC1 functional polymorphism and lung cancer risk in Chinese. *PLoS ONE*, *9*(9), 106794. https://doi.org/10.1371/journal.pone.0106794

Botta, V., Louppe, G., Geurts, P., & Wehenkel, L. (2014). Exploiting SNP Correlations within Random Forest for Genome-Wide Association Studies. *PLoS ONE*, *9*(4), e93379. https://doi.org/10.1371/journal.pone.0093379

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An Expanded View of Complex Traits: From

Polygenic to Omnigenic. *Cell*, *169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038
Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology*, *8*(12). https://doi.org/10.1371/journal.pcbi.1002822
Cardon, L. R., & Palmer, L. J. (2003). Population stratification and spurious allelic association. *The Lancet*, *361*(9357), 598–604. https://doi.org/10.1016/S0140-6736(03)12520-2
Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., Mcauley, E., Miranda, A., Peat, G., Spitzer, M., Barrett, J., Hulcoop, D. G., Papa, E., Koscielny, G., & Dunham, I. (2019). Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research*, *47*. https://doi.org/10.1093/nar/gky1133
Chen, G.-B., Lee, S. H., Montgomery, G. W., Wray, N. R., Visscher, P. M., Gearry, R. B., Lawrance, I. C., Andrews, J. M., Bampton, P., Mahy, G., Bell, S., Walsh, A., Connor, S., Sparrow, M., Bowdler, L. M., Simms, L. A., Krishnaprasad, K., Radford-Smith, G. L., & Moser, G. (2017). Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. *BMC Medical Genetics*, *18*(1), 94. https://doi.org/10.1186/s12881-017-0451-2
Chen, G. B., Lee, S. H., Montgomery, G. W., Wray, N. R., Visscher, P. M., Gearry, R. B., Lawrance, I. C., Andrews, J. M., Bampton, P., Mahy, G., Bell, S., Walsh, A., Connor, S., Sparrow, M., Bowdler, L. M., Simms, L. A., Krishnaprasad, K., Radford-Smith, G. L., & Moser, G. (2017). Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. *BMC Medical Genetics*, *18*(1), 1–11. https://doi.org/10.1186/s12881-017-0451-2
Curbelo Montañez, C. A., Fergus, P., Curbelo Montañez, A., & Chalmers, C. (2018). *Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs*. https://arxiv.org/pdf/1804.03198.pdf
de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., … Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, *49*(2), 256–261. https://doi.org/10.1038/ng.3760
de Oliveira, F., Borges, C. C. H., Almeida, F., e Silva, F., da Silva Verneque, R., da Silva, M. V. G., & Arbex, W. (2014). SNPs selection using support vector regression and genetic algorithms in GWAS. *BMC Genomics*, *15*(Suppl 7), S4. https://doi.org/10.1186/1471-2164-15-S7-S4
Denys, A., & Allain, F. (2019). The Emerging Roles of Heparan Sulfate 3-O-Sulfotransferases in Cancer. *Frontiers in Oncology*, *9*(JUN), 507. https://doi.org/10.3389/FONC.2019.00507
Dinu, I., Mahasirimongkol, S., Liu, Q., Yanai, H., Sharaf Eldin, N., Kreiter, E., Wu, X., Jabbari, S., Tokunaga, K., & Yasui, Y. (2012). SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics. *PloS One*, *7*(10), e43035. https://doi.org/10.1371/journal.pone.0043035
Dudbridge, F. (2013). Correction: Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, *9*(4).
Duerr, R. H., Taylor, K. D., Brant, S. R., Rioux, J. D., Silverberg, M. S., Daly, M. J., Steinhart, A. H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L. W., Kistner, E. O., Schumm, L. P., Lee, A. T., Gregersen, P. K., … Cho, J. H. (2006). A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science (New York, N.Y.)*, *314*(5804), 1461. https://doi.org/10.1126/SCIENCE.1135245
Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., & Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics 2010 11:6*, *11*(6), 446–450. https://doi.org/10.1038/nrg2809

Elding, H., Lau, W., Swallow, D. M., & Maniatis, N. (2013). REPORT Refinement in Localization and Identification of Gene Regions Associated with Crohn Disease. *The American Journal of Human Genetics*, *92*, 107–113. https://doi.org/10.1016/j.ajhg.2012.11.004

Ferns, G. A. A., Shelley, C. S., Stocks, J., Rees, A., Paul, H., Baralle, F., & Galton, D. J. (1986). A DNA polymorphism of the apoprotein AII gene in hypertriglyceridaemia. *Human Genetics*, *74*(3), 302–306. https://doi.org/10.1007/BF00282553

Ferreira, M. A. (2018). Ten years of genome-wide association studies of immune-related diseases. *Clinical & Translational Immunology*, *7*(6), e1022. https://doi.org/10.1002/cti2.1022

Feuerstein, J. D., & Cheifetz, A. S. (2017). Crohn Disease: Epidemiology, Diagnosis, and Management. *Mayo Clinic Proceedings*, *92*(7), 1088–1103. https://doi.org/10.1016/j.mayocp.2017.04.010

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1. /pmc/articles/PMC2929880/

Gajendran, M., Loganathan, P., Catinella, A. P., & Hashash, J. G. (2018). A comprehensive review and update on Crohn's disease. *Disease-a-Month*, *64*(2), 20–57. https://doi.org/10.1016/j.disamonth.2017.07.001

Gaudillo, J., Joseph Russell Rodriguez, J., Nazareno, A., Rigi Baltazar, L., Vilela, J., Bulalacao, R., Domingo, M., & Albia, J. (2019). Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLOS ONE*, *14*(12), e0225574. https://doi.org/10.1371/JOURNAL.PONE.0225574

Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A., & Smoller, J. W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nature Communications*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-09718-5

Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: Methodological considerations and new findings. *BMC Genetics*, *11*. https://doi.org/10.1186/1471-2156-11-49

Gordon, H., Moller, F. T., Andersen, V., & Harbord, M. (2015). Heritability in Inflammatory Bowel Disease: From the First Twin Study to Genome-Wide Association Studies. *Inflammatory Bowel Diseases*, *21*(6), 1428. https://doi.org/10.1097/MIB.0000000000000393

Grenier, L., & Hu, P. (2019). Computational drug repurposing for inflammatory bowel disease using genetic information. *Computational and Structural Biotechnology Journal*, *17*, 127–135. https://doi.org/10.1016/j.csbj.2019.01.001

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. https://doi.org/10.1148/radiology.143.1.7063747

Hayes, B. (2013). *Overview of Statistical Methods for Genome-Wide Association Studies (GWAS)* (pp. 149–169). Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0_6

Hellwege, J. N., Keaton, J. M., Giri, A., Gao, X., Velez Edwards, D. R., & Edwards, T. L. (2017). Population Stratification in Genetic Association Studies. *Current Protocols in Human Genetics*, *95*(1), 1.22.1-1.22.23. https://doi.org/10.1002/cphg.48

Henderson, P., & Stevens, C. (2012). The Role of Autophagy in Crohn's Disease. *Cells*, *1*(3), 492. https://doi.org/10.3390/CELLS1030492

Hibar, D. P., Stein, J. L., Jahanshad, N., Kohannim, O., Hua, X., Toga, A. W., McMahon, K. L., de Zubicaray, G. I., Martin, N. G., Wright, M. J., Alzheimer's Disease Neuroimaging Initiative, the A. D. N., Weiner, M. W., & Thompson, P. M. (2015). Genome-wide interaction analysis reveals replicated epistatic effects on brain structure. *Neurobiology of Aging*, *36 Suppl 1*(0 1), S151-8. https://doi.org/10.1016/j.neurobiolaging.2014.02.033

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hochberg, Y., & Benjamini, Y. (1990). More powerful procedures for multiple significance testing.

Statistics in Medicine, *9*(7), 811–818. http://www.ncbi.nlm.nih.gov/pubmed/2218183
Hofuku, I., Yokoi, T., & Oshima, K. (2013). An introduction of the node-clustering algorithm using the PH algorithm. *Information (Japan)*, *16*(12 B), 8597–8610. https://doi.org/10.1038/nature08494
Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211
Kim, N. G., Rhee, H., Li, L. S., Kim, H., Lee, J. S., Kim, J. H., Nam, K. K., & Kim, H. (2002). Identification of MARCKS, FLJ11383 and TAF1B as putative novel target genes in colorectal carcinomas with microsatellite instability. *Oncogene*, *21*(33), 5081–5087. https://doi.org/10.1038/sj.onc.1205703
Koloski, N. A., Bret, L., & Radford-Smith, G. (2008). Hygiene hypothesis in inflammatory bowel disease: A critical review of the literature. In *World Journal of Gastroenterology* (Vol. 14, Issue 2, pp. 165–173). Baishideng Publishing Group Co. https://doi.org/10.3748/wjg.14.165
Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology*, *34*(7), 643–652. https://doi.org/10.1002/gepi.20509
Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods*, *9*(1), 1. https://doi.org/10.1186/1746-4811-9-29
L, J., S, R., RK, W., RH, D., DP, M., KY, H., JC, L., LP, S., Y, S., CA, A., J, E., M, M., K, N., I, C., E, T., SL, S., S, R., P, G., Z, W., … JH, C. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, *491*(7422), 119–124. https://doi.org/10.1038/NATURE11582
Lee, S. M., Suk, K., & Lee, W. H. (2015). Myristoylated alanine-rich C kinase substrate (MARCKS) regulates the expression of proinflammatory cytokines in macrophages through activation of p38/JNK MAPK and NF-κB. *Cellular Immunology*, *296*(2), 115–121. https://doi.org/10.1016/j.cellimm.2015.04.004
Levine, M. E., Langfelder, P., & Horvath, S. (2009). *Protein Networks and Pathway Analysis*. *563*(m), 277–290. https://doi.org/10.1007/978-1-60761-175-2
Li, J., Zhang, Q., Chen, F., Yan, J., Kim, S., Wang, L., Feng, W., Saykin, A. J., Liang, H., & Shen, L. (2015). Genetic Interactions Explain Variance in Cingulate Amyloid Burden: An AV-45 PET Genome-Wide Association and Interaction Study in the ADNI Cohort. *BioMed Research International*, *2015*, 647389. https://doi.org/10.1155/2015/647389
Ling, C. X., & Li, C. (1998). Data Mining for Direct Marketing: Problems and. *KDD'98: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. www.aaai.org
Liu, J. Z., & Anderson, C. A. (2014). Genetic studies of Crohn's disease: past, present and future. *Best Practice & Research. Clinical Gastroenterology*, *28*(3), 373–386. https://doi.org/10.1016/j.bpg.2014.04.009
Liu, J. Z., Van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., Ripke, S., Lee, J. C., Jostins, L., Shah, T., Abedian, S., Cheon, J. H., Cho, J., Daryani, N. E., Franke, L., Fuyuno, Y., Hart, A., Juyal, R. C., Juyal, G., … Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, *47*(9), 979–986. https://doi.org/10.1038/ng.3359
Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A Package for Binary Imbalanced Learning. *The R Journal*, *6*, 79–89.
M'Koma, A. E. (2013). Inflammatory bowel disease: an expanding global health problem. *Clinical Medicine Insights. Gastroenterology*, *6*, 33–47. https://doi.org/10.4137/CGast.S12731
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*.

https://doi.org/10.1093/nar/gkw1133
Malovini, A., Bellazzi, R., Napolitano, C., & Guffanti, G. (2016). Multivariate Methods for Genetic Variants Selection and Risk Prediction in Cardiovascular Diseases. *Frontiers in Cardiovascular Medicine*, *3*, 17. https://doi.org/10.3389/fcvm.2016.00017
Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics 2007 39:7*, *39*(7), 906–913. https://doi.org/10.1038/ng2088
Max Kuhn. (2021). *caret: Classification and Regression Training* (R package version 6.0-88).
McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344
McKinney, B. A., Reif, D. M., Ritchie, M. D., & Moore, J. H. (2006). Machine learning for detecting gene-gene interactions: a review. *Applied Bioinformatics*, *5*(2), 77–88. http://www.ncbi.nlm.nih.gov/pubmed/16722772
Michail, S., Bultron, G., & Depaolo, R. W. (2013). Genetic variants associated with Crohn's disease. *The Application of Clinical Genetics*, *6*, 25–32. https://doi.org/10.2147/TACG.S33966
Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., Farré, X., Marigorta, U. M., Fehr, E., Dickhaus, T., Blanchard, G., Schunk, D., Navarro, A., & Müller, K.-R. (2016). Combining Multiple Hypothesis Testing with Machine Learning Increases the Statistical Power of Genome-wide Association Studies. *Scientific Reports*, *6*(1), 36671. https://doi.org/10.1038/srep36671
Mittag, F., Römer, M., & Zell, A. (2015). Influence of feature encoding and choice of classifier on disease risk prediction in genome-wide association studies. *PLoS ONE*, *10*(8), 1–18. https://doi.org/10.1371/journal.pone.0135832
Mo, J.-S., Alam, K. J., Kim, H.-S., Lee, Y.-M., Yun, K.-J., & Chae, S.-C. (2016). MicroRNA 429 Regulates Mucin Gene Expression and Secretion in Murine Model of Colitis. *Journal of Crohn's and Colitis*, *10*(7), 837–849. https://doi.org/10.1093/ECCO-JCC/JJW033
Molina, L. C., Belanche, L., Nebot, À., Girona, J., & C, C. N. (2002). Feature Selection Algorithms: A Survey and Experimental Evaluation. *Proceeding ICDM '02 Proceedings of the 2002 IEEE International Conference on Data Mining*, 306.
Mooney, M., Wilmot, B., Bipolar Genome Study, T., & McWeeney, S. (2011). The GA and the GWAS: using genetic algorithms to search for multilocus associations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(3), 899–910. https://doi.org/10.1109/TCBB.2011.145
Moore, R., Ashby, K., Liao, T. J., & Chen, M. (2021). Machine Learning to Identify Interaction of Single-Nucleotide Polymorphisms as a Risk Factor for Chronic Drug-Induced Liver Injury. *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 10603*, *18*(20), 10603. https://doi.org/10.3390/IJERPH182010603
Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(DEC), 21. https://doi.org/10.3389/FNBOT.2013.00021/BIBTEX
Newcombe, P. J., Nelson, C. P., Samani, N. J., & Dudbridge, F. (2019). A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genetic Epidemiology*, *43*(7), 730–741. https://doi.org/10.1002/gepi.22245
Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., Panaccione, R., Ghosh, S., Wu, J. C. Y., Chan, F. K. L., Sung, J. J. Y., & Kaplan, G. G. (2017). Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*, *390*(10114), 2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0
Oreski, D., & Novosel, T. (2014). *Comparison of Feature Selection Techniques in Knowledge*

Discovery Process (Vol. 3, Issue 4). www.temjournal.com

Paré, G., Mao, S., & Deng, W. Q. (2017). A machine-learning heuristic to improve gene score prediction of polygenic traits. *Scientific Reports*, *7*(1), 12665. https://doi.org/10.1038/s41598-017-13056-1

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics*, *81*(3), 559. https://doi.org/10.1086/519795

Qin, Y., Zhuang, S., Wen, J., & Zheng, K. (2018). Long non-coding RNA MDC1-AS inhibits human gastric cancer cell proliferation and metastasis through an MDC1-dependent mechanism. *Experimental and Therapeutic Medicine*, *15*(1), 191. https://doi.org/10.3892/ETM.2017.5370

R Core Team. (2021). *A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Reisberg, S., Iljasenko, T., Läll, K., Fischer, K., & Vilo, J. (2017). Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLOS ONE*, *12*(7), e0179238. https://doi.org/10.1371/journal.pone.0179238

Richardson, T. G., Harrison, S., Hemani, G., & Smith, G. D. (2018). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *BioRxiv*, 1–24. https://doi.org/10.1101/467910

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., Hugot, J. P., Peyrin-Biroulet, L., Chamaillard, M., Colombel, J. F., Cottone, M., D'Amato, M., D'Incà, R., Halfvarson, J., Henderson, P., Karban, A., Kennedy, N. A., Khan, M. A., Lémann, M., Levine, A., Massey, D., … Whittaker, P. (2019). Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. *Scientific Reports*, *9*(1), 1–18. https://doi.org/10.1038/s41598-019-46649-z

Roman Hornung. (2021). *ordinalForest: Ordinal Forests: Prediction and Variable Ranking with Ordinal Target Variables* (R package version 2.4-2).

S, D., L, F., S, S., C, S., AE, L., A, K., SI, V., EY, C., S, L., M, M., D, S., D, S., PR, L., WG, I., A, S., LJ, S., F, C., F, K., M, B., … C, F. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/NG.3656

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*, 308–311.

Shuang Song, Wei Jiang, Lin Hou, H. Z. (2020). Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Computational Biology*, *16*(2), e1007565.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart - Biological queries made easy. *BMC Genomics*, *10*(1), 1–12. https://doi.org/10.1186/1471-2164-10-22/TABLES/4

Stelzer, G., Dalah, I., Stein, T. I., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I., Krug, H., Perco, P., Mayer, B., Kolker, E., Safran, M., & Lancet, D. (2011). In-silico human genomics with GeneCards. *Human Genomics*, *5*(6), 709–717. https://doi.org/10.1186/1479-7364-5-6-709

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, *187*(2), 367–383.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. von. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional

discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*(D1), D607–D613. https://doi.org/10.1093/NAR/GKY1131

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., Kö Nig, I. R., Zhang, H., & Sun, Y. V. (2009). Machine Learning in Genome-Wide Association Studies. *Genetic Epidemiology*, *33*, 51–57. https://doi.org/10.1002/gepi.20473

Tang, J., Alelyani, S., & Liu, H. (2014). *Feature Selection for Classification: A Review*. https://pdfs.semanticscholar.org/310e/a531640728702fce6c743c1dd680a23d2ef4.pdf

Teo, Y. Y. (2008). Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Current Opinion in Lipidology*, *19*(2), 133–143.

Tianqi Chen, He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., & Li, Y. (2021). *xgboost: Extreme Gradient Boosting* (R package version 1.4.1.1).

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methological)*, *58*(1), 267–288. https://www.jstor.org/stable/2346178

Torkamani, A., Wineinger, N. E., & Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, *19*(9), 581–590. https://doi.org/10.1038/s41576-018-0018-x

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., de Andrade, M., Doheny, K. F., Haines, J. L., Hayes, G., Jarvik, G., Jiang, L., Kullo, I. J., Li, R., Ling, H., Manolio, T. A., Matsumoto, M., McCarty, C. A., McDavid, A. N., … Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Current Protocols in Human Genetics*, *Chapter 1*, Unit1.19. https://doi.org/10.1002/0471142905.hg0119s68

Uen, Y.-H., Fang, C.-L., Hseu, Y.-C., Shen, P.-C., Yang, H.-L., Wen, K.-S., Hung, S.-T., Wang, L.-H., & Lin, K.-Y. (2015). VAV3 Oncogene Expression in Colorectal Cancer: Clinical Aspects and Functional Characterization. *Scientific Reports 2015 5:1*, *5*(1), 1–8. https://doi.org/10.1038/srep09360

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H., Kathiresan, S., Pato, M., Pato, C., Tamimi, R., Stahl, E., Zaitlen, N., Pasaniuc, B., … Zheng, W. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *The American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P. R., Bhatia, G., Do, R., Hayeck, T., Won, H. H., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, *97*(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

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. *The American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Wang, M. H., Cordell, H. J., & Van Steen, K. (2018). Statistical methods for genome-wide association studies. *Seminars in Cancer Biology*. https://doi.org/10.1016/J.SEMCANCER.2018.04.008

Wang, X., Peng, Q., & Fan, Y. (2016). Detecting Susceptibility to Breast Cancer with SNP-SNP Interaction Using BPSOHS and Emotional Neural Networks. *BioMed Research International*, *2016*, 1–7. https://doi.org/10.1155/2016/5164347

Wang, Y., Miller, M., Astrakhan, Y., Petersen, B. S., Schreiber, S., Franke, A., & Bromberg, Y.

(2019). Identifying Crohn's disease signal from variome analysis. *Genome Medicine*, *11*(1), 1–15. https://doi.org/10.1186/s13073-019-0670-6

Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., Frackelton, E., Kim, C., Mentch, F., Van Steen, K., Visscher, P. M., Baldassano, R. N., & Hakonarson, H. (2013). Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *American Journal of Human Genetics*, *92*(6), 1008–1012. https://doi.org/10.1016/j.ajhg.2013.05.002

Wigginton, J. E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *American Journal of Human Genetics*, *77*(5), 754–767. https://doi.org/10.1086/497345

Witte, J. S. (2010). Genome-wide association studies and beyond. *Annual Review of Public Health*, *31*, 9-20 4 p following 20. https://doi.org/10.1146/annurev.publhealth.012809.103723

Włodarczyk, M., Sobolewska-Włodarczyk, A., Cygankiewicz, A. I., Jacenik, D., Krajewska, W. M., Stec-Michalska, K., Piechota-Polańczyk, A., Wiśniewska-Jarosińska, M., & Fichna, J. (2017). G protein-coupled receptor 55 (GPR55) expresses differently in patients with Crohn's disease and ulcerative colitis. *Scandinavian Journal of Gastroenterology*, *52*(6–7), 711–715. https://doi.org/10.1080/00365521.2017.1298834

Wray, N. R., Lin, T., Austin, J., McGrath, J. J., Hickie, I. B., Murray, G. K., & Visscher, P. M. (2021). From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry*, *78*(1), 101–109. https://doi.org/10.1001/JAMAPSYCHIATRY.2020.3049

Xu, E. L., Qian, X., Yu, Q., Zhang, H., & Cui, S. (2018). Feature selection with interactions in logistic regression models using multivariate synergies for a GWAS application. *BMC Genomics*, *19*(Suppl 4), 170. https://doi.org/10.1186/s12864-018-4552-x

Yan, Q., Jiang, Y., Huang, H., Swaroop, A., Chew, E. Y., Weeks, D. E., Chen, W., & Ding, Y. (2021). Genome-wide association studies-based machine learning for prediction of age-related macular degeneration risk. *Translational Vision Science and Technology*, *10*(2), 1–8. https://doi.org/10.1167/tvst.10.2.29

Yang, C., Zhou, X., Wan, X., Yang, Q., Xue, H., & Yu, W. (2011). Identifying disease-associated SNP clusters via contiguous outlier detection. *Bioinformatics*, *27*(18), 2578–2585. https://doi.org/10.1093/BIOINFORMATICS/BTR424

Zhang, H., Yu, J.-Q., Yang, L.-L., Kramer, L. M., Zhang, X.-Y., Na, W., Reecy, J. M., & Li, H. (2017). Identification of genome-wide SNP-SNP interactions associated with important traits in chicken. *BMC Genomics*, *18*(1), 892. https://doi.org/10.1186/s12864-017-4252-y

Zhang, Y. (2012). A novel bayesian graphical model for genome-wide multi-SNP association mapping. *Genetic Epidemiology*, *36*(1), 36–47. https://doi.org/10.1002/gepi.20661

Zheng, W., & Rao, S. (2015). Knowledge-based analysis of genetic associations of rheumatoid arthritis to inform studies searching for pleiotropic genes: a literature review and network analysis. *Arthritis Research & Therapy*, *17*(1), 202. https://doi.org/10.1186/s13075-015-0715-1

Zhernakova, A., Van Diemen, C. C., & Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews. Genetics*, *10*(1), 43–55. https://doi.org/10.1038/NRG2489

Zuk, O., Hechter, E., Sunyaev, S. R., & Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(4), 1193–1198. https://doi.org/10.1073/PNAS.1119675109/-/DCSUPPLEMENTAL

Appendix 1

SNPname	Ch	Position	Gene	Mean Multi-variate Rank	Ldpred Rank	XGBTree Rank	RF Rank	LASSO Rank	BSWiMS Rank	Trait CD | IBD | UC at OpenTargets
rs2659046	17	79145891	AATK		457	-	-	-	-	UC
rs440970	5	131336287	ACSL6	49	95	34	38	21	55	CD
exm105654	1	155033308	ADAM15		366	-	-	-	-	CD
rs9791011	5	33622158	ADAMTS12		322	-	-	-	-	
rs10051817	5	33679999	ADAMTS12		330	-	-	-	-	
rs6555299	5	4534191	ADAMTS16		308	-	-	-	-	
rs17256169	16	3993253	ADCY9		372	-	-	-	-	
rs2050395	10	115801595	ADRB1		488	-	-	-	-	
rs12654778	5	148205741	ADRB2		325	-	-	-	-	
rs7529090	1	247069232	AHCTF1		427	-	-	-	-	
exm-rs9348876	6	31575276	AIF1		416	-	-	-	-	CD
rs2271696	10	71874683	AIFM2		203	-	-	-	-	
rs9496563	6	143616271	AIG1		50	-	-	-	-	
rs11162331	1	77850800	AK5		313	-	-	-	-	
exm24765	1	19201919	ALDH4A1		510	-	-	-	-	
exm2257537	6	135128858	ALDH8A1	341	458	232	350	249	418	CD
rs4421638	1	105681907	AMY1C		52	-	-	-	-	
rs16872693	5	74946455	ANKDD1B		130	-	-	-	-	
rs12928649	16	89333342	ANKRD11		26	-	-	-	-	CD
rs13188321	5	111914106	APC		375	-	-	-	-	IBD
exm-rs3117582	6	31620520	APOM	329	156	336	318	418	418	CD
rs1685633	1	154291718	AQP10		167	-	-	-	-	IBD
exm1315364	17	36614485	ARHGAP23		35	-	-	-	-	
rs7342975	17	66391232	ARSG		55	-	-	-	-	
rs6501429	17	66393689	ARSG		114	-	-	-	-	
rs3785613	17	66275830	ARSG		206	-	-	-	-	
exm1327986	17	42254236	ASB16		16	-	-	-	-	
exm3709	1	1430985	ATAD3B		347	-	-	-	-	IBD
exm-rs3830076	6	32096244	ATF6B		454	-	-	-	-	CD
rs1045100	2	234203597	ATG16L1	110	388	21	37	74	32	CD
exm-rs10210302	2	234158839	ATG16L1	119	244	143	57	113	37	CD
rs2241879	2	234183468	ATG16L1	124	445	22	44	73	34	CD
exm-rs3792109	2	234184417	ATG16L1	128	499	73	28	25	14	CD

exm-rs3828309	2	234180410	ATG16L1	144	410	89	48	46	128	CD
exm276398	2	234183368	ATG16L1	202	353	114	35	92	418	CD
rs6663281	1	63221191	ATG4C	271	348	85	240	264	418	CD
rs4782612	16	84384326	ATP2C2		82	-	-	-	-	Appendicitis
exm-rs1032757	5	81939318	ATP6AP1L		187	-	-	-	-	
rs1032070	17	40618251	ATP6V0A1	280	233	219	276	255	418	CD
rs4791171	17	63541497	AXIN2		448	-	-	-	-	
rs2837102	21	41004152	B3GALT5		340	-	-	-	-	CD
exm1371085	17	80923546	B3GNTL1		12	-	-	-	-	CD
rs7774238	6	91107018	BACH2		23	-	-	-	-	CD
exm-rs2844463	6	31615167	BAG6		228	-	-	-	-	CD
rs9922832	16	50444159	BRD7		195	-	-	-	-	CD
rs7204069	16	50451650	BRD7		326	-	-	-	-	CD
rs16945643	17	59893990	BRIP1		234	-	-	-	-	
exm74874	1	92554283	BTBD8		232	-	-	-	-	IBD
exm1196960	16	613344	C16orf11		299	-	-	-	-	
exm1238987	16	49430534	C16orf78		25	-	-	-	-	UC
rs7217052	17	21452282	C17orf51	280	185	298	261	240	418	CD
rs1052227	17	54906137	C17orf67	209	297	121	301	192	133	CD
exm1338509	17	54872439	C17orf67		238	-	-	-	-	CD
rs12139580	1	244373559	C1orf100		283	-	-	-	-	
exm-rs7554511	1	200877562	C1orf106	273	387	118	227	215	418	CD
rs10489182	1	169710669	C1orf112	184	150	99	151	103	418	CD
rs2902440	1	67670916	C1orf141	23	41	6	7	20	42	CD
exm-rs7517847	1	67681669	C1orf141	24	91	8	6	9	5	CD
exm-rs11805303	1	67675516	C1orf141	55	98	63	17	48	48	CD
exm-rs11209003	1	67601132	C1orf141	65	68	102	27	72	58	CD
rs10489630	1	67662622	C1orf141	99	129	42	36	133	154	CD
rs4655690	1	67659896	C1orf141	107	63	288	55	56	73	CD
rs7539625	1	67672765	C1orf141	117	58	29	26	53	418	CD
exm67171	1	67560956	C1orf141	141	153	167	157	141	88	CD
rs3762318	1	67597119	C1orf141	163	140	50	88	121	418	CD
rs10789224	1	67605134	C1orf141	167	110	156	29	120	418	CD
rs10749771	1	67573730	C1orf141	219	241	202	117	117	418	CD
rs2064689	1	67653010	C1orf141	229	402	66	40	217	418	CD
rs1885276	1	67568824	C1orf141	293	272	319	167	289	418	CD
rs10489631	1	67601115	C1orf141	321	309	328	215	333	418	CD
exm-rs497309	6	31892484	C2	349	255	418	340	312	418	CD
exm-rs558702	6	31870326	C2	369	398	418	223	386	418	CD
exm1612908	22	42089623	C22orf46	349	506	214	352	256	418	CD

rs6895072	5	133144207	C5orf15		160	-	-	-	-		
rs7715716	5	147289799	C5orf46		444	-	-	-	-		
exm-rs12521868	5	131784393	C5orf56	150	169	165	140	137	140		CD
rs2522051	5	131797578	C5orf56	194	275	314	109	144	129		CD
rs2548993	5	131808869	C5orf56	201	455	148	139	151	111		CD
exm535230	6	32261252	C6orf10	286	120	302	222	368	418		CD
rs2894179	6	31066671	C6orf15		429	-	-	-	-		CD
exm-rs9368699	6	31802541	C6orf48		413	-	-	-	-		CD
exm1342077	17	58235763	CA4		99	-	-	-	-		CD
rs3767498	1	201020727	CACNA1S	135	258	54	150	78	137		CD
rs9921802	16	24297521	CACNG3		492	-	-	-	-		IBD
exm2264647	17	65000219	CACNG4		262	-	-	-	-		
rs11684413	2	85627714	CAPG	304	428	181	273	222	418		IBD
exm1196761	16	597764	CAPN15		85	-	-	-	-		
rs16871475	5	71020547	CARTPT		245	-	-	-	-		UC
rs738469	22	39510995	CBX7		108	-	-	-	-		CD
exm-rs12470505	2	219908369	CCDC108		323	-	-	-	-		
exm458879	5	68616079	CCDC125		344	-	-	-	-		
rs210837	17	32735169	CCL1		135	-	-	-	-		CD
rs159297	17	32706889	CCL1		143	-	-	-	-		CD
rs991804	17	32587725	CCL2	129	296	49	96	83	120		CD
rs10515854	5	162737689	CCNG1		43	-	-	-	-		
rs2059849	5	66610910	CD180		133	-	-	-	-		CD
rs798029	1	117337524	CD2		352	-	-	-	-		
rs17122383	1	100860339	CDC14A		33	-	-	-	-		
exm454994	5	54468450	CDC20B		13	-	-	-	-		
exm154203	1	227182033	CDC42BPA		449	-	-	-	-		
rs1038843	5	21257067	CDH12		70	-	-	-	-		IBD
rs10492861	16	82866767	CDH13	275	254	82	320	302	418		CD
rs16960006	16	83297261	CDH13		101	-	-	-	-		CD
rs7716554	5	31200936	CDH6		281	-	-	-	-		
rs7200019	16	62518621	CDH8		451	-	-	-	-		
rs6450652	5	28621295	CDH9		355	-	-	-	-		
rs10045332	5	28560977	CDH9		415	-	-	-	-		
rs9292272	5	28561546	CDH9		476	-	-	-	-		
rs16941336	17	20575697	CDRT15L2		292	-	-	-	-		
rs1139056	22	17661178	CECR1		277	-	-	-	-		
exm148351	1	214818548	CENPF		9	-	-	-	-		CD
rs6710428	2	169368019	CERS6		34	-	-	-	-		
rs4668082	2	169497129	CERS6		73	-	-	-	-		

exm-rs1270942	6	31918860	CFB	385	350	418	360	378	418	CD
rs903358	1	203147139	CHI3L1		482	-	-	-	-	
exm862774	10	125780647	CHST15		119	-	-	-	-	
rs3765265	16	840597	CHTF18		59	-	-	-	-	
rs4895566	6	139878255	CITED2		205	-	-	-	-	UC
exm-rs3131383	6	31704294	CLIC1	318	286	418	243	227	418	CD
rs7199467	16	80951145	CMC2		496	-	-	-	-	IBD
exm1579312	21	46875874	COL18A1		418	-	-	-	-	
rs7563419	2	237742697	COPS8		252	-	-	-	-	
exm1620076	22	50315363	CRELD2		267	-	-	-	-	UC
rs25887	5	131416061	CSF2	172	324	160	114	157	106	CD
rs2069616	5	131408077	CSF2	218	354	64	182	361	131	CD
rs10914850	1	34496094	CSMD2		260	-	-	-	-	
rs7557987	2	81469658	CTNNA2	266	467	69	231	145	418	Allergic rhinitis
rs953458	10	67082879	CTNNA3		447	-	-	-	-	
exm849080	10	101993033	CWF19L1		96	-	-	-	-	
rs809601	10	44789602	CXCL12		218	-	-	-	-	
rs652267	5	156795204	CYFIP2		516	-	-	-	-	
rs1420872	16	50807779	CYLD	161	174	210	104	195	121	CD
rs4785452	16	50842077	CYLD	315	396	418	106	239	418	CD
rs11863019	16	50847819	CYLD		405	-	-	-	-	CD
rs13333062	16	50922786	CYLD		503	-	-	-	-	CD
exm843816	10	96447920	CYP2C18		62	-	-	-	-	
rs4796803	17	76630164	CYTH1		265	-	-	-	-	UC
rs2939378	5	40042245	DAB2		94	-	-	-	-	CD
exm519094	6	18256625	DEK		56	-	-	-	-	IBD
exm-rs12134279	1	197781198	DENND1B	297	498	80	255	235	418	CD
rs4927176	1	55354335	DHCR24	297	505	141	236	184	418	CD
rs4558075	10	6401625	DKFZP667F0711		450	-	-	-	-	
rs903630	6	170428032	DLL1		230	-	-	-	-	
exm472888	5	118480316	DMXL1		316	-	-	-	-	
rs16970950	16	21145657	DNAH3		223	-	-	-	-	
rs1992711	5	13510918	DNAH5		11	-	-	-	-	
exm206363	2	84745113	DNAH6		22	-	-	-	-	
rs2150431	21	42291678	DSCAM		306	-	-	-	-	
rs6690208	1	212295571	DTL		329	-	-	-	-	
rs4589119	1	221915967	DUSP10		288	-	-	-	-	
rs4433402	1	221946172	DUSP10		489	-	-	-	-	
rs11195128	10	112186148	DUSP5	22	80	4	13	6	6	CD
rs10444086	10	112179167	DUSP5	78	168	84	60	26	54	CD

exm-rs474534	6	31938107	DXO		310	-	-	-	-	CD
exm533591	6	31938412	DXO		334	-	-	-	-	CD
exm837666	10	82126600	DYDC2		149	-	-	-	-	CD
rs6869051	5	158345923	EBF1		446	-	-	-	-	CD
rs1034237	6	12395762	EDN1		261	-	-	-	-	
rs7578234	2	233543537	EFHD1		189	-	-	-	-	
exm533062	6	31864538	EHMT2	73	18	241	30	49	29	CD
exm851598	10	103988265	ELOVL3		480	-	-	-	-	IBD
rs1428556	5	73856207	ENC1		406	-	-	-	-	UC
exm38093	1	29320013	EPB41		20	-	-	-	-	
rs341295	5	111848890	EPB41L4A-AS2		165	-	-	-	-	CD
rs3926393	5	159667613	FABP6		161	-	-	-	-	
rs2901520	5	159613503	FABP6		460	-	-	-	-	
rs16872345	5	74149478	FAM169A		377	-	-	-	-	UC
exm896844	11	22646398	FANCF		441	-	-	-	-	
rs17114146	10	103401516	FBXW4		166	-	-	-	-	
rs4661028	1	157322310	FCRL5		83	-	-	-	-	
rs13361304	5	108013371	FER		141	-	-	-	-	
rs6875865	5	108556802	FER		181	-	-	-	-	
exm842900	10	95347041	FFAR4		29	-	-	-	-	
rs7189414	16	86620191	FOXL1	234	112	130	307	201	418	
rs733023	5	132655625	FSTL4		284	-	-	-	-	
rs312305	5	162095410	GABRG2		456	-	-	-	-	
rs2290949	16	81413389	GAN		343	-	-	-	-	IBD
exm181733	2	27730940	GCKR	119	391	37	83	54	30	CD
exm1543134	20	42891917	GDAP1L1		393	-	-	-	-	CD
rs10788959	1	54068567	GLIS1		321	-	-	-	-	
rs3013760	1	54036593	GLIS1		423	-	-	-	-	
rs9426298	1	28990922	GMEB1		333	-	-	-	-	
rs252200	5	141400028	GNPDA1		473	-	-	-	-	CD
rs2312489	1	167040788	GPA33		280	-	-	-	-	
kgp7075915	6	24473399	GPLD1		219	-	-	-	-	
rs1848728	2	231839240	GPR55	132	314	58	110	61	116	CD
rs17062729	6	102432315	GRIK2		214	-	-	-	-	UC
rs12596342	16	9939750	GRIN2A		317	-	-	-	-	
rs10117679	9	104378479	GRIN3A		424	-	-	-	-	
rs11198881	10	121087219	GRK5		57	-	-	-	-	
exm1317567	17	38064469	GSDMB	121	191	199	105	87	25	CD
rs11078927	17	38064405	GSDMB	171	249	71	121	280	132	CD
rs7094905	10	1015247	GTPBP4		338	-	-	-	-	

exm546690	6	42153428	GUCA1B		179	-	-	-	-		IBD
exm1593009	22	24039410	GUSBP11		270	-	-	-	-		
rs7707445	5	156449748	HAVCR1		282	-	-	-	-		
exm-rs2395029	6	31431780	HCP5		163	-	-	-	-		
exm530706	6	31379043	HCP5		300	-	-	-	-		
rs3891636	5	130439091	HINT1		370	-	-	-	-		CD
rs11596587	10	71113988	HK1		486	-	-	-	-		
rs2523619	6	31318144	HLA-B	289	399	154	252	221	418		CD
exm-rs2844586	6	31318024	HLA-B		240	-	-	-	-		CD
exm-rs9264942	6	31274380	HLA-C	105	266	20	69	30	141		CD
exm-rs2524229	6	31275231	HLA-C		273	-	-	-	-		CD
exm-rs10484554	6	31274555	HLA-C		472	-	-	-	-		CD
exm-rs1480380	6	32913246	HLA-DMA		479	-	-	-	-		CD
exm-rs7758736	6	32758394	HLA-DOB	197	376	225	190	107	86		CD
exm-rs2187668	6	32605884	HLA-DQA1	374	367	418	251	418	418		CD
rs7775228	6	32658079	HLA-DQB1		356	-	-	-	-		CD
rs7749057	6	32448904	HLA-DRA	304	404	273	201	225	418		CD
exm-rs2395175	6	32405026	HLA-DRA		436	-	-	-	-		CD
rs7736962	5	135764923	HNRNPA1P13		303	-	-	-	-		
rs10521233	17	13559080	HS3ST3A1		395	-	-	-	-		
rs8050730	16	25965289	HS3ST4	189	14	198	142	171	418		
rs780433	5	118990871	HSD17B4		328	-	-	-	-		
exm1031652	12	104332224	HSP90B1	115	87	61	5	5	418		CD
rs4585392	5	62670313	HTR1A		304	-	-	-	-		
rs9686886	5	148010913	HTR4		285	-	-	-	-		
rs10476898	5	148056656	HTR4		443	-	-	-	-		
exm-rs907092	17	37922259	IKZF3	129	122	117	115	182	109		CD
exm-rs9303277	17	37976469	IKZF3	210	142	122	156	213	418		CD
exm-rs3024493	1	206943968	IL10	234	220	418	287	172	71		CD
exm-rs3024505	1	206939904	IL10	257	198	175	296	200	418		CD
rs4921227	5	158849837	IL12B	57	128	12	65	18	61		CD
exm-rs10045431	5	158814533	IL12B	224	178	149	235	138	418		CD
exm-rs6871626	5	158826792	IL12B		37	-	-	-	-		CD
rs7720046	5	158884535	IL12B		369	-	-	-	-		CD
rs1495965	1	67753508	IL12RB2	173	111	127	64	147	418		CD
exm-rs924080	1	67760140	IL12RB2	267	305	129	95	387	418		CD
exm67254	1	67705958	IL23R	1	1	1	2	2	1		CD
exm-rs10889677	1	67725120	IL23R	42	69	27	42	50	22		CD
rs10889676	1	67722567	IL23R	122	190	192	58	76	95		CD
rs61839660	10	6094697	IL2RA	186	207	258	267	132	67		CD

rs58736	5	62560237	IPO11	137	97		113	266	108	102	CD
exm-rs11117432	16	86019270	IRF8		182	-		-	-		CD
exm-rs7714584	5	150270420	IRGM		81	-		-	-		CD
exm-rs13361189	5	150223387	IRGM		173	-		-	-		CD
exm-rs11747270	5	150258867	IRGM		373	-		-	-		CD
rs1449263	2	182319301	ITGA4	130	359		60	99	69	63	CD
rs6740847	2	182308352	ITGA4	206	349		65	128	68	418	CD
exm1236563	16	31418975	ITGAD		31	-		-	-		IBD
exm-rs10758669	9	4981602	JAK2	87	360		7	34	12	20	CD
exm2273550	17	39880545	JUP		425	-		-	-		
rs895767	2	224023296	KCNE4		437	-		-	-		
rs1693229	1	233738342	KCNK1		311	-		-	-		
rs6681392	1	154796712	KCNN3		144	-		-	-		IBD
rs11264268	1	154796520	KCNN3		400	-		-	-		IBD
exm-rs11584383	1	200935866	KIF21B	328	335		216	324	345	418	CD
exm1319729	17	39041052	KRT20		259	-		-	-		
rs4958756	5	154045825	LARP1		468	-		-	-		
exm142334	1	205353492	LEMD1		208	-		-	-		
rs17017451	1	211557908	LINC00467		216	-		-	-		
rs11580269	1	169005304	LINC00970		301	-		-	-		
rs7865479	9	27845760	LINGO2		319	-		-	-		
exm216435	2	100915330	LONRF2		379	-		-	-		
exm838275	10	85984444	LRIT2		242	-		-	-		
exm-rs769177	6	31547611	LTB		32	-		-	-		CD
rs7605137	2	150092739	LYPD6B		362	-		-	-		
rs17162589	5	109384131	MAN2A1		508	-		-	-		
rs4968857	17	67420433	MAP2K6		61	-		-	-		
rs753173	10	30778738	MAP3K8	248	500		242	264	155	81	CD
rs11008080	10	30802799	MAP3K8	395	501		310	365	381	418	CD
exm1330831	17	43922897	MAPT		67	-		-	-		
exm2270128	5	126356042	MARCH3		263	-		-	-		
rs4945943	**6**	**113712036**	**MARCKS**	**18**	**46**		**9**	**18**	**8**	**10**	**Acute apendicitis**
exm530865	6	31496949	MCCD1		105	-		-	-		CD
rs9262151	**6**	**30672353**	**MDC1**	**11**	**4**		**38**	**4**	**3**	**4**	**UC**
exm565825	6	90362783	MDN1		3	-		-	-		Periodontitis
exm220981	2	112725747	MERTK		66	-		-	-		
rs7562674	2	172259687	METTL8	303	209		300	311	279	418	CD
exm1606112	22	37866063	MFNG		72	-		-	-		

rs11695439	2	191261942	MFSD6		512	-	-	-	-		
exm-rs3099844	6	31448976	MICB	339	180	285	392	418	418		CD
exm530818	6	31474820	MICB		461	-	-	-	-		CD
rs12090955	1	67412261	MIER1		113	-	-	-	-		CD
rs8081387	17	5395383	MIS12		497	-	-	-	-		
rs9935021	16	14132484	MKL2		371	-	-	-	-		
exm1620696	22	50584130	MOV10L1		172	-	-	-	-		UC
exm914444	11	60183189	MS4A14		201	-	-	-	-		
rs792369	17	55446071	MSI2		434	-	-	-	-		CD
exm834193	10	75184444	MSS51		103	-	-	-	-		CD
exm586580	6	151270231	MTHFD1L		88	-	-	-	-		
rs6665978	1	238575867	MTRNR2L11		126	-	-	-	-		CD
rs2487091	1	238255771	MTRNR2L11		194	-	-	-	-		CD
rs10512951	5	8303911	MTRR		49	-	-	-	-		
rs10512943	5	8151979	MTRR		183	-	-	-	-		
rs6721218	2	177414915	MTX2		137	-	-	-	-		
exm1280666	17	4457116	MYBBP1A		36	-	-	-	-		
rs9909522	17	10235790	MYH13	278	200	177	393	202	418		
rs7217738	17	59630944	NACA2		162	-	-	-	-		
rs2072711	22	37268555	NCF4		337	-	-	-	-		CD
rs6718462	2	183795633	NCKAP1		211	-	-	-	-		
rs11749731	5	141500436	NDFIP1		215	-	-	-	-		CD
rs9324864	5	141469000	NDFIP1		459	-	-	-	-		CD
rs3095902	5	149926998	NDST1		243	-	-	-	-		CD
rs322388	5	172156062	NEURL1B		269	-	-	-	-		
rs3890764	1	61961953	NFIA		278	-	-	-	-		
rs12144629	1	183397899	NMNAT2		45	-	-	-	-		
exm1239948	16	50745926	NOD2	3	2	2	3	4	3		CD
exm-rs5743289	16	50756774	NOD2	27	54	5	10	31	33		CD
rs11647841	16	50743331	NOD2	94	124	53	70	105	119		CD
rs2066843	16	50745199	NOD2	107	229	133	41	80	52		CD
exm-rs2076756	16	50756881	NOD2	150	204	18	53	59	418		CD
exm1239874	16	50744624	NOD2	216	148	272	86	156	418		CD
rs2297516	17	26095730	NOS2		430	-	-	-	-		CD
exm-rs3830041	6	32191339	NOTCH4		431	-	-	-	-		CD
rs12448862	16	18068642	NPIPA8		307	-	-	-	-		
rs12562860	1	161211678	NR1I3		414	-	-	-	-		CD
rs9405897	6	6052055	NRN1	355	407	297	335	317	418		Spondylosis without

											myelopathy
rs10827245	10	33639488	NRP1		339	-	-	-	-		
rs11244602	10	126064956	OAT		363	-	-	-	-		IBD
rs9358619	6	9282211	OFCC1		453	-	-	-	-		
exm114185	1	159283746	OR10J3		109	-	-	-	-		
exm912261	11	58034651	OR10W1		293	-	-	-	-		IBD
exm-rs12517906	5	180170819	OR2Y1		279	-	-	-	-		
rs10903267	5	180171716	OR2Y1		392	-	-	-	-		
rs16942771	17	56252489	OR4D2	100	76	124	141	99	62		IBD
rs7211774	17	56257984	OR4D2	239	86	418	163	111	418		IBD
exm912309	11	58125774	OR5B17		426	-	-	-	-		IBD
exm166706	1	247875313	OR6F1		40	-	-	-	-		
rs162907	5	131580152	P4HA2	243	315	418	189	197	96		CD
rs156109	5	131626611	P4HA2		248	-	-	-	-		CD
exm1226533	16	23634293	PALB2		196	-	-	-	-		IBD
rs4408174	1	99946091	PALMD		60	-	-	-	-		
rs12060602	1	99994648	PALMD		358	-	-	-	-		
rs3859118	16	50252235	PAPD5	318	440	418	136	179	418		CD
rs4312853	5	7006254	PAPD7		186	-	-	-	-		
rs2302404	10	89474072	PAPSS2		199	-	-	-	-		
exm488660	5	140763665	PCDHGA1		15	-	-	-	-		Appendicitis
exm159815	1	233397909	PCNXL2		27	-	-	-	-		
exm-rs968451	22	39670851	PDGFB	58	132	24	32	28	72		CD
exm-rs2413583	22	39659773	PDGFB	83	84	97	116	44	76		CD
rs9369484	6	12661620	PHACTR1		452	-	-	-	-		
exm1307233	17	27238135	PHF12		42	-	-	-	-		
rs1932758	1	88685730	PKN2		397	-	-	-	-		
rs12038869	1	188469106	PLA2G4A		327	-	-	-	-		CD
rs656755	1	20424167	PLA2G5		477	-	-	-	-		IBD
exm1608133	22	38528888	PLA2G6		295	-	-	-	-		
rs2227552	10	75669319	PLAU	129	271	55	134	94	90		CD
exm1262411	16	81922813	PLCG2		138	-	-	-	-		IBD
rs318843	5	41373205	PLCXD3		351	-	-	-	-		CD
rs390956	1	242398403	PLD5	208	274	152	247	233	135		CD
rs10926652	1	242380948	PLD5	249	253	161	237	174	418		CD
rs2342271	1	242663273	PLD5		5	-	-	-	-		CD
rs12406043	1	242940149	PLD5		121	-	-	-	-		CD
rs12075764	1	242907976	PLD5		382	-	-	-	-		CD
rs34052165	2	132057166	PLEKHB2		213	-	-	-	-		

rs7769470	6	106487292	PRDM1	109	251	59	119	55	59	CD
rs2495060	1	14032007	PRDM2		474	-	-	-	-	
rs10852259	16	24068314	PRKCB		515	-	-	-	-	IBD
rs12570350	10	6795493	PRKCQ		30	-	-	-	-	CD
exm449399	5	35072712	PRLR		24	-	-	-	-	
rs1584468	5	119937691	PRR16		93	-	-	-	-	
rs6909321	6	31093190	PSORS1C1		147	-	-	-	-	CD
rs7730267	5	40548545	PTGER4	30	38	23	8	15	64	CD
rs9687958	5	40496423	PTGER4	34	48	48	12	38	24	CD
exm-rs4613763	5	40392728	PTGER4	34	17	31	11	29	82	CD
exm-rs17234657	5	40401509	PTGER4	35	19	123	9	16	8	CD
rs4286721	5	40497604	PTGER4	45	90	77	15	27	16	CD
rs7720838	5	40486896	PTGER4	55	64	79	24	86	23	CD
exm-rs11742570	5	40410584	PTGER4	142	341	17	82	180	92	CD
exm-rs10440635	5	40490790	PTGER4	148	77	52	31	164	418	CD
rs6869535	5	40597618	PTGER4	163	115	111	54	118	418	CD
rs13163402	5	40607910	PTGER4	166	28	81	75	228	418	CD
rs10077544	5	40484938	PTGER4	167	139	153	33	93	418	CD
rs4957138	5	40622940	PTGER4	193	51	140	73	281	418	CD
exm-rs6896969	5	40424426	PTGER4	205	509	330	72	71	43	CD
rs1876143	5	40521648	PTGER4	330	312	418	185	315	418	CD
exm85427	1	114377568	PTPN22		225	-	-	-	-	CD
exm138344	1	202128601	PTPN7		432	-	-	-	-	
rs10801677	1	198628483	PTPRC	145	8	203	62	33	418	CD
rs10758997	9	8829567	PTPRD		487	-	-	-	-	
rs1905339	17	40582296	PTRF	157	298	96	170	98	122	CD
rs9912887	17	29827405	RAB11FIP4		403	-	-	-	-	
rs12948477	17	29928492	RAB11FIP4		478	-	-	-	-	
rs375085	17	29943842	RAB11FIP4		485	-	-	-	-	
rs4272630	1	220486349	RAB3GAP2		342	-	-	-	-	
rs3738091	1	21997282	RAP1GAP		368	-	-	-	-	
rs11139654	9	85220698	RASEF	224	118	269	174	143	418	
rs9921862	16	5658177	RBFOX1	219	100	75	304	198	418	UC
rs8063739	16	5778928	RBFOX1	223	157	110	241	187	418	UC
rs12933690	16	5758532	RBFOX1		394	-	-	-	-	UC
rs7193708	16	8084800	RBFOX1		411	-	-	-	-	UC
rs4523953	17	77373331	RBFOX3		159	-	-	-	-	UC
exm96669	1	151316324	RFX5		175	-	-	-	-	IBD
rs11200948	10	86023610	RGR		79	-	-	-	-	
rs10754006	1	192390633	RGS21		438	-	-	-	-	

rs2686226	1	241376512	RGS7		514	-	-	-	-		
rs1877823	17	63226943	RGS9		264	-	-	-	-		
rs9896245	17	63173756	RGS9		417	-	-	-	-		
rs497915	1	182406654	RGSL1		217	-	-	-	-		
exm560271	6	72984123	RIMS1		462	-	-	-	-		
exm513179	6	3111166	RIPK1		107	-	-	-	-		CD Irritable bowel syndrome
exm1357589	17	74154496	RNF157		7	-	-	-	-		
rs12627970	22	39721745	RPL3	71	197	40	46	32	41		CD
rs137603	22	39694225	RPL3	162	389	106	138	85	91		CD
rs11655133	17	72130512	RPL38		490	-	-	-	-		
rs3778409	6	166954244	RPS6KA2		155	-	-	-	-		CD
rs10864040	1	213468908	RPS6KC1		345	-	-	-	-		
exm-rs6679677	1	114303808	RSBN1	253	146	176	342	181	418		CD
rs10489161	1	25338799	RUNX3		421	-	-	-	-		IBD
rs11649472	16	51457948	SALL1		385	-	-	-	-		CD
rs7742658	6	28600492	SCAND3		357	-	-	-	-		
exm2250	1	1226757	SCNN1D		117	-	-	-	-		IBD
exm2070	1	1220954	SCNN1D		331	-	-	-	-		IBD
exm1332781	17	45915766	SCRN2		227	-	-	-	-		IBD
rs790088	17	71588183	SDK2	189	152	267	277	158	89		
rs4789155	17	71558419	SDK2		374	-	-	-	-		
exm798317	9	139369066	SEC16A		237	-	-	-	-		CD
rs10158522	1	177938712	SEC16B		222	-	-	-	-		
rs7076585	10	75503692	SEC24C	150	466	56	111	65	53		CD
exm834716	10	75523634	SEC24C		136	-	-	-	-		CD
rs431892	9	102056057	SEC61B		75	-	-	-	-		
rs6594980	5	116042843	SEMA6A		39	-	-	-	-		CD
rs4788999	17	75621704	SEPT9		481	-	-	-	-		
rs12530071	6	2886067	SERPINB9		435	-	-	-	-		
rs11968128	6	134445802	SGK1		291	-	-	-	-		
exm831814	10	72604263	SGPL1	67	106	39	112	39	38		CD
rs12414453	10	83006207	SH2D4B		502	-	-	-	-		CD
rs12076073	1	154944156	SHC1		276	-	-	-	-		CD
rs585499	1	232476942	SIPA1L2		504	-	-	-	-		
exm533505	6	31935567	SKIV2L		192	-	-	-	-		CD
exm114567	1	159799910	SLAMF8		92	-	-	-	-		CD
rs6596966	6	3407646	SLC22A23		171	-	-	-	-		CD
exm476696	5	131676320	SLC22A4	248	463	86	149	125	418		CD

exm-rs2073643	5	131723288	SLC22A5	85	102	32	81	140	70	CD
rs7550613	1	9562830	SLC25A33		290	-	-	-	-	
rs12618482	2	196206373	SLC39A10		226	-	-	-	-	
rs11589993	1	8284497	SLC45A1	192	154	93	193	102	418	CD
exm-rs35391	5	33955673	SLC45A2		469	-	-	-	-	
exm182283	2	27887160	SLC4A1AP		71	-	-	-	-	CD
rs1358175	17	38757789	SMARCE1		364	-	-	-	-	UC
exm2259809	1	152849299	SMCP		409	-	-	-	-	IBD
rs159364	5	60522414	SMIM15		365	-	-	-	-	
rs6873063	5	150187805	SMIM3		164	-	-	-	-	CD
rs2153409	1	245995920	SMYD3		221	-	-	-	-	
rs2548621	5	53902339	SNX18		6	-	-	-	-	
exm57998	1	48825355	SPATA6		53	-	-	-	-	
rs16960660	17	19965587	SPECC1		176	-	-	-	-	
rs6502788	17	4354348	SPNS3		408	-	-	-	-	
rs10040443	5	141693593	SPRY4		493	-	-	-	-	CD
rs13430952	2	45765157	SRBD1		361	-	-	-	-	
rs36921	5	65464936	SREK1		287	-	-	-	-	
rs26055	5	111040812	STARD4		151	-	-	-	-	
rs8082391	17	40398973	STAT5B	424	475	418	391	418	418	CD
exm-rs389884	6	31940897	STK19	287	134	316	293	273	418	CD
rs4716127	6	17181845	STMND1		507	-	-	-	-	
rs5750824	22	39830123	TAB1		433	-	-	-	-	CD
rs10803704	2	9957097	TAF1B	121	346	62	108	42	47	CD
exm-rs2857106	6	32787570	TAP2	308	289	163	305	364	418	CD
kgp8226585	6	32813768	TAPSAR1		170	-	-	-	-	CD
exm2948	1	1269554	TAS1R3		236	-	-	-	-	IBD
exm1362902	17	77984172	TBC1D16		44	-	-	-	-	
rs2789074	6	121126482	TBC1D32		383	-	-	-	-	
rs9398632	6	121661170	TBC1D32		464	-	-	-	-	
rs1779429	1	119460481	TBX15		390	-	-	-	-	
rs2973656	5	167280392	TENM2		131	-	-	-	-	
rs9466019	6	10356036	TFAP2A		10	-	-	-	-	Sialoadenitis
rs6717937	2	122043586	TFCP2L1		123	-	-	-	-	
rs2580359	2	122063844	TFCP2L1		401	-	-	-	-	
rs2810891	1	92154088	TGFBR3		104	-	-	-	-	IBD
rs10059560	5	156405784	TIMD4		318	-	-	-	-	
rs10114675	9	82759463	TLE4		420	-	-	-	-	
rs2724918	2	569139	TMEM18		188	-	-	-	-	
rs11165328	1	95630461	TMEM56	251	257	221	217	142	418	CD

exm1573724	21	43809092	TMPRSS3		302	-	-	-	-	IBD
exm-rs7517810	1	172853460	TNFSF18	25	65	26	19	7	7	CD
exm-rs9286879	1	172862234	TNFSF18	36	78	25	39	24	13	CD
rs10912561	1	173164565	TNFSF4	166	89	47	187	91	418	CD
rs2996494	1	201343638	TNNT2		184	-	-	-	-	CD
exm137208	1	201330429	TNNT2		336	-	-	-	-	CD
rs4788430	16	24793053	TNRC6A		294	-	-	-	-	
rs1035671	2	218674484	TNS1		511	-	-	-	-	CD
rs16858496	2	218824450	TNS1		513	-	-	-	-	CD
exm-rs1150754	6	32050758	TNXB	132	125	172	169	96	98	CD
exm534227	6	32036788	TNXB		145	-	-	-	-	CD
exm-rs3807039	6	32078373	TNXB		246	-	-	-	-	CD
exm-rs9267796	6	32023425	TNXB		380	-	-	-	-	CD
rs2077580	6	32020844	TNXB		484	-	-	-	-	CD
rs2545093	5	180682862	TRIM52		381	-	-	-	-	
rs26173	5	14479938	TRIO		332	-	-	-	-	
rs7092748	10	116730538	TRUB1		491	-	-	-	-	
rs6556280	5	176096188	TSPAN17		378	-	-	-	-	
rs4867851	5	176101823	TSPAN17		419	-	-	-	-	
exm2270040	5	40755568	TTC33	249	116	257	226	229	418	CD
exm1621207	22	50658424	TUBGCP6		495	-	-	-	-	UC
kgp9751812	2	234651880	UGT1A8		47	-	-	-	-	CD
rs17114247	21	43478135	UMODL1		494	-	-	-	-	IBD
exm234549	2	158958605	UPP2	118	212	266	61	34	17	
rs7606259	2	158959335	UPP2		202	-	-	-	-	
rs9824503	3	179451466	USP13		439	-	-	-	-	
rs207136	1	55814164	USP24		412	-	-	-	-	
rs2823286	21	16817938	USP25	68	256	13	20	14	39	CD
exm276857	2	234436069	USP40		74	-	-	-	-	CD
rs11185129	1	108070435	VAV3	211	177	74	183	204	418	
rs9623076	22	22560977	VPREB1		422	-	-	-	-	
exm966088	11	124016058	VWA5A		384	-	-	-	-	
rs1457113	5	110467074	WDR36		210	-	-	-	-	
rs8068482	17	80608149	WDR45B		320	-	-	-	-	
rs28758854	9	137015951	WDR5		193	-	-	-	-	
exm1275236	17	1636950	WDR81		247	-	-	-	-	
exm1348476	17	66449122	WIPI1		250	-	-	-	-	
rs6555796	5	167731435	WWC1		386	-	-	-	-	
rs2737290	16	78897506	WWOX		158	-	-	-	-	
rs17124963	10	110463874	XPNPEP1		483	-	-	-	-	

rs8049877	16	17667833	XYLT1		21	-	-	-	-	CD
rs16827790	2	187240797	ZC3H15		268	-	-	-	-	IBD
rs3751675	16	88688832	ZC3H18		470	-	-	-	-	
rs7189871	16	72996847	ZFHX3		224	-	-	-	-	
exm526404	6	29641514	ZFP57		127	-	-	-	-	
rs2802372	10	81047575	ZMIZ1	226	239	87	195	190	418	CD
exm2259640	10	80898075	ZMIZ1		231	-	-	-	-	CD
rs1515278	2	180503273	ZNF385B		442	-	-	-	-	
kgp5731474	1	91330027	ZNF644		471	-	-	-	-	
exm165864	1	247151487	ZNF695		465	-	-	-	-	
exm1279692	17	3981290	ZZEF1		235	-	-	-	-	

Appendix 2

Mean Rank Multivariate Models	SNPname	Chr	Position (hg37)	XGBTree Rank	RF Rank	LASSO Rank	BSWiMS Rank	Gen for Functional Analysis	Trait CD | IBD | UC at Opentargets
1.5	rs11209026	1	67705958	1	2	2	1	IL23R	CD
3	rs2066844	16	50745926	2	3	4	3	NOD2	CD
1.75	rs2066845	16	50756540	3	1	1	2	NOD2	CD
7.25	rs11195128	10	112186148	4	13	6	6	DUSP5	CD
19.75	rs5743289	16	50756774	5	10	31	33	NOD2	CD
18.75	rs2902440	1	67670916	6	7	20	42	IL23R	CD
18.25	rs10758669	9	4981602	7	34	12	20	JAK2	CD
7	rs7517847	1	67681669	8	6	9	5	IL23R	CD
11.25	**rs4945943**	**6**	**113712036**	**9**	**18**	8	10	**MARCKS**	**Acute apendicitis**
13.5	rs1039823	2	28623159	10	22	13	9	FOSL2	CD
12.25	rs4807569	19	1123378	11	16	10	12	SBNO2	CD
39	rs4921227	5	158849837	12	65	18	61	IL12B	CD
21.5	rs2823286	21	16817938	13	20	14	39	NRIP1	CD
23.75	rs8413	9	139323311	14	47	19	15	INPP5E	CD
17.25	rs2155219	11	76299194	15	21	22	11	LRRC32	CD
37.25	rs2005557	3	49701298	16	68	37	28	BSN	CD
92.75	rs11742570	5	40410584	17	82	180	92	PTGER4	CD
137	rs2076756	16	50756881	18	53	59	418	NOD2	CD
26.5	rs6478108	9	117558703	19	51	17	19	TNFSF15	CD

65.5	rs9264942	6	31274380	20	69	30	143	HLA-C	CD
41	rs1045100	2	234203597	21	37	74	32	ATG16L1	CD
43.25	rs2241879	2	234183468	22	44	73	34	ATG16L1	CD
27.5	rs7730267	5	40548545	23	8	15	64	PTGER4	CD
39	rs968451	22	39670851	24	32	28	72	PDGFB	CD
25.25	rs9286879	1	172862234	25	39	24	13	TNFSF18	CD
14.75	rs7517810	1	172853460	26	19	7	7	TNFSF18	CD
35.25	rs10889677	1	67725120	27	42	50	22	IL23R	CD
32.75	rs2048507	4	26079164	28	59	23	21	RBPJ	CD
131.5	rs7539625	1	67672765	29	26	53	418	IL23R	CD
61.75	rs7130588	11	76270683	30	45	47	125	LRRC32	CD
38.25	rs4613763	5	40392728	31	11	29	82	PTGER4	CD
80.75	rs2073643	5	131723288	32	81	140	70	SLC22A5	CD
187	rs2069235	22	39747780	33		110	418	SYNGR1	CD
37	rs440970	5	131336287	34	38	21	55	ACSL6	CD
65.75	rs10784838	12	70757341	35	77	52	99	KCNMB4	CD
140	rs9267798	6	32044834	36	66	40	418	TNXB	CD
51	rs1260326	2	27730940	37	83	54	30	GCKR	CD
12.25	rs9262151	6	30672353	38	4	3	4	MDC1	RA
57	rs12770335	10	72604263	39	112	39	38	SGPL1	CD
39.75	rs12627970	22	39721745	40	46	32	41	SYNGR1	CD
179.25	rs4409764	10	101284237	41	129	129	418	NKX2-3	CD
91.25	rs10489630	1	67662622	42	36	133	154	IL23R	CD
68.5	rs1729662	2	61391305	43	133	67	31	C2orf74	CD
66.75	rs912113	20	1342187	44	122	45	56	SIRPB1	CD
195	rs10489629	1	67688349	45	50	267	418	IL23R	CD
76	rs910050	6	32315654	46	101	63	94	TSBP1, C6orf10	CD
185.75	rs10912561	1	173164565	47	187	91	418	TNFSF4	CD
30.5	rs9687958	5	40496423	48	12	38	24	PTGER4	CD
87	rs991804	17	32587725	49	96	83	120	CCL2	CD
169.25	rs3762318	1	67597119	50	88	121	418	IL23R	CD
76.75	rs10995271	10	64438486	51	146	75	35	ADO	CD
166.25	rs10440635	5	40490790	52	31	164	418	PTGER4	CD
86.75	rs11647841	16	50743331	53	70	105	119	NOD2	CD
104.75	rs3767498	1	201020727	54	150	78	137	CACNA1S	CD
93.25	rs2227552	10	75669319	55	134	94	90	PLAU	CD
71.25	rs7076585	10	75503692	56	111	65	53	SEC24C	CD
25	rs34536443	19	10463118	57	14	11	18	TYK2	IBD
86.25	rs1848728	2	231839240	58	110	61	116	GPR55	CD
73	rs7769470	6	106487292	59	119	55	59	PRDM1	CD

72.75	rs1449263	2	182319301	60	99	69	63	ITGA4	CD
122.25	rs116891695	12	104332224	61	5	5	418	HSP90B1	CD
64.75	rs10803704	2	9957097	62	108	42	47	TAF1B	CD
44	rs11805303	1	67675516	63	17	48	48	IL23R	CD
184.5	rs2069616	5	131408077	64	182	361	131	CSF2	CD
169.75	rs6740847	2	182308352	65	128	68	418	ITGA4	CD
185.25	rs2064689	1	67653010	66	40	217	418	IL23R	CD
59	rs2242258	10	75607168	67	78	41	50	CAMK2G	CD
140.25	rs933243	6	167403873	68	224	128	141	RNASET2	CD
215.75	rs7557987	2	81469658	69	231	145	418	LRRTM1	Allergic rhinitis
101	rs4785205	16	50237938	70	152	95	87	TENT4B, PAPD5	CD
151	rs11078927	17	38064405	71	121	280	132	GSDMB	CD
77.75	rs10761659	10	64445564	72	85	51	103	ADO	CD
35	rs3792109	2	234184417	73	28	25	14	ATG16L1	CD
219.75	rs11185129	1	108070435	74	183	204	418	VAV3	
248.75	rs9921862	16	5658177	75	304	198	418	RBFOX1	
187.25	rs1063169	14	75747118	76	166	89	418	FOS	CD
33.75	rs4286721	5	40497604	77	15	27	16	PTGER4	CD
103.25	rs224120	10	64445760	78	135	115	85	ZNF365	CD
53	rs7720838	5	40486896	79	24	86	23	PTGER4	CD
247	rs12134279	1	197781198	80	255	235	418	DENND1B	CD
200.5	rs13163402	5	40607910	81	75	228	418	PTGER4	CD
280.5	rs10492861	16	82866767	82	320	302	418	CDH13	CD
217.5	rs2188962	5	131770805	83	118	251	418	IRF1	CD
56	rs10444086	10	112179167	84	60	26	54	DUSP5	CD
251.75	rs6663281	1	63221191	85	240	264	418	ATG4C	CD
194.5	rs1050152	5	131676320	86	149	125	418	SLC22A4	
222.5	rs2802372	10	81047575	87	195	190	418	ZMIZ1	CD
153.25	rs1480382	6	32740895	88	253	148	124	HLA-DQB2	CD
77.75	rs3828309	2	234180410	89	48	46	128	ATG16L1	CD
96.75	rs2872507	17	38040763	90	87	97	113	IKZF3	CD
83.75	rs10487279	7	102516744	91	120	58	66	FBXL13	CD
235	rs10781510	9	139279173	92	143	287	418	SNAPC4	CD
201.5	rs11589993	1	8284497	93	193	102	418	PARK7	CD
92	rs4945087	11	76136482	94	79	77	118	GVQW3	CD
96.5	rs10852936	17	38031714	95	100	79	112	ZPBP2	CD
121.5	rs1905339	17	40582296	96	170	98	122	MLX	CD
83.25	rs2413583	22	39659773	97	116	44	76	PDGFB	CD
94.25	rs6127152	20	52872215	98	145	90	44	PFDN4	UC
192.75	rs10489182	1	169710669	99	151	103	418	SELL	CD

227.5	rs744166	17	40514201	100	216	176	418	STAT3	CD
143.5	rs17293632	15	67442596	101	168	178	127	SMAD3	CD
64.75	rs11209003	1	67601132	102	27	72	58	C1orf141	CD
192.25	rs7068361	10	64408367	103	89	159	418	ZNF365	CD
112.25	rs6583623	8	143480669	104	147	88	110	TSNARE1	CD
207.75	rs4643314	16	50375955	105	196	112	418	BRD7	CD
105	rs137603	22	39694225	106	138	85	91	SYNGR1	CD
60.25	rs7927997	11	76301375	107	49	36	49	LRRC32	CD
188.5	rs6478109	9	117568766	108	94	134	418	TNFSF15	CD
200.25	rs1736020	21	16812552	109	125	149	418	NRIP1	CD
239	rs8063739	16	5778928	110	241	187	418	RBFOX1	UC
175.25	rs6869535	5	40597618	111	54	118	418	PTGER4	CD
112.75	rs1736148	21	16813212	112	126	153	60	NRIP1	CD
147.25	rs58736	5	62560237	113	266	108	102	LRRC70	CD
164.75	rs2241880	2	234183368	114	35	92	418	ATG16L1	CD
147.75	rs114543649	6	32261158	115	23	35	418	TSBP1, C6orf10	CD
225	rs4354332	8	115492603	116	230	136	418	CSMD3	CD
130.75	rs907092	17	37922259	117	115	182	109	IKZF3	CD
244.5	rs7554511	1	200877562	118	227	215	418	INAVA, C1orf106	CD
248.75	rs10853952	19	1163934	119	291	167	418	SBNO2	CD
263.25	rs2236866	1	169596313	120	295	220	418	SELP	CD
186.75	rs1052227	17	54906137	121	301	192	133	DGKE	CD
227.25	rs9303277	17	37976469	122	156	213	418	IKZF3	CD
39	rs17234657	5	40401509	123	9	16	8	PTGER4	CD
106.5	rs16942771	17	56252489	124	141	99	62	OR4D1	IBD
244.25	rs890432	17	46629232	125	265	169	418	HOXB3	
142.75	rs10882091	10	94374377	126	219	109	117	KIF11	CD
189	rs1495965	1	67753508	127	64	147	418	IL23R	CD
126	rs12537821	7	102624395	128	165	146	65	FBXL13	CD
257.25	rs924080	1	67760140	129	95	387	418	IL12RB2	CD
264	rs7189414	16	86620191	130	307	201	418	FOXC2	
163.75	rs4833071	4	38582859	131	256	122	146	KLF3	CD
221.5	rs11190140	10	101291593	132	124	212	418	NKX2-3	CD
76.5	rs2066843	16	50745199	133	41	80	52	NOD2	CD
178.25	rs2305480	17	38062196	134	80	81	418	GSDMB	CD
188.75	rs7927894	11	76301316	135	67	135	418	LRRC32	CD
256.5	rs8005161	14	88472595	136	206	266	418	GPR65	CD
219.75	rs12523160	5	40385790	137	93	231	418	PTGER4	CD
198.25	rs10781500	9	139269338	138	113	124	418	CARD9	CD
72	rs6451493	5	40410935	139	43	60	46	PTGER4	CD

228	rs4957138	5	40622940	140	73	281	418	PTGER4	CD
244.75	rs4927176	1	55354335	141	236	184	418	DHCR24	CD
175	rs1866316	15	67441997	142	288	126	144	SMAD3	CD
87.5	rs10210302	2	234158839	143	57	113	37	ATG16L1	CD
71	rs10043301	5	40530886	144	25	70	45	PTGER4	CD
266.5	rs16826094	3	116814285	145	292	211	418	LSAMP	
303.75	rs463426	22	21809185	146	329	322	418	UBE2L3	CD
111.25	rs1172294	2	25169200	147	137	104	57	DNAJC27	CD
137.25	rs2548993	5	131808869	148	139	151	111	IRF1	CD
235	rs10045431	5	158814533	149	235	138	418	UBLCP1	CD
151.5	rs713586	2	25158008	150	164	139	153	ADCY3	CD
233.75	rs1074664	14	29068832	151	203	163	418	FOXG1	CD
191.75	rs390956	1	242398403	152	247	233	135	PLD5	CD
174.25	rs10077544	5	40484938	153	33	93	418	PTGER4	CD
261.25	rs2523619	6	31318144	154	252	221	418	MICB	CD
279.25	rs4625	3	49572140	155	200	344	418	DAG1	CD
180.75	rs10789224	1	67605134	156	29	120	418	IL23R	CD
296	rs652157	13	102764516	157	332	277	418	FGF14	CD
297.75	rs13035268	2	127937551	158	363	252	418	ERCC3	CD
256.5	rs6451494	5	40411291	159	159	290	418	PTGER4	CD
134.25	rs25887	5	131416061	160	114	157	106	SLC22A4	CD
247.5	rs10926652	1	242380948	161	237	174	418	PLD5	CD
95.5	rs1060962	3	49708502	162	107	62	51	BSN	CD
312.5	rs2857106	6	32787570	163	305	364	418	HLA-DOB	CD
131.25	rs1297265	21	16817051	164	91	119	151	NRIP1	CD
146	rs12521868	5	131784393	165	140	137	142	IRF1	CD
76.75	rs3091315	17	32593665	166	71	43	27	CCL7	CD
138.25	rs11208997	1	67560956	167	157	141	88	C1orf141	CD
134.25	rs10883365	10	101287764	169	123	130	115	NKX2-3	CD
211.25	rs10124038	9	110872527	170	312	223	140	KLF4	Acute infective polyneuritis/guillain-barre syndrome
287.25	rs2197465	14	48572632	171	330	230	418	MDGA2	MS
133.75	rs1150754	6	32050758	172	169	96	98	TNXB	CD
81.25	rs1373692	5	40431183	173	52	64	36	PTGER4	CD
324	rs6702421	1	197559324	174	325	379	418	DENND1B	CD
272.25	rs3024505	1	206939904	175	296	200	418	IL10	CD
279.25	rs6679677	1	114303808	176	342	181	418	PHTF1	CD
297.5	rs9909522	17	10235790	177	393	202	418	MYH13	
317.5	rs4917129	7	50323174	178	411	263	418	IKZF1	CD

213	rs11159833	14	88476004	179	171	84	418	AMT	CD
315.25	rs4345993	11	82222510	180	381	282	418	MIR4300HG	CD
273.5	rs11684413	2	85627714	181	273	222	418	CAPG	IBD
246.75	rs11957215	5	40445681	182	153	234	418	PTGER4	CD
329.5	rs3130649	6	30803254	183	380	337	418	FLOT1	CD
266.25	rs10952869	7	75846453	184	260	203	418	SRRM3	CD
204	rs10941516	5	40522212	185	63	150	418	PTGER4	CD
139.5	rs1516975	8	129548193	186	180	123	69	MYC	CD
280.75	rs1177283	2	61348804	187	232	286	418	KIAA1841	CD
168.75	rs7848647	9	117569046	188	132	210	145	TNFSF15	CD
168.25	rs10500264	19	33750314	189	208	175	101	SLC7A10	CD
290.25	rs4029774	17	40428961	190	268	285	418	STAT5A	CD
280.5	rs2823269	21	16793467	191	218	295	418	NRIP1	CD
105.25	rs10889676	1	67722567	192	58	76	95	IL23R	CD
264.75	rs1000141	2	234242347	193	92	356	418	SAG	CD
298	rs7105981	11	4980153	194	334	246	418	APEH	UC
168.5	rs11178234	12	70819638	195	175	165	139	KCNMB4	CD
322.75	rs3902025	17	38119254	196	336	341	418	GSDMA	CD
311.25	rs5751086	22	41768862	197	383	247	418	TEF	CD
232.25	rs8050730	16	25965289	198	142	171	418	HS3ST4	
104	rs11078928	17	38064469	199	105	87	25	GSDMB	CD
279.25	rs10398	11	308180	200	254	245	418	IFITM2	CD
184.75	rs747063	20	50102203	201	238	193	107	NFATC2	Acute periodontitis
213.5	rs10749771	1	67573730	202	117	117	418	C1orf141	CD
179	rs10801677	1	198628483	203	62	33	418	PTPRC	CD
254.5	rs1479008	4	57017493	204	234	162	418	KIAA1211	IBD
224.25	rs10513140	3	141218057	205	349	209	134	RASA2	CD
164.75	rs35263917	3	41952852	206	248	131	74	ULK4	CD
327.75	rs3865452	19	41211056	207	331	355	418	COQ8B	CD
180	rs2513638	11	115412993	208	249	170	93	CADM1	CD
320.75	rs7955946	12	66912923	209	303	353	418	GRIP1	Gerd
157.5	rs1420872	16	50807779	210	104	195	121	CYLD	CD
207.75	rs2284176	6	30875622	211	317	219	84	GTF2H4	CD
325.75	rs7725523	5	40372223	212	308	365	418	PTGER4	CD
302.5	rs17745066	2	46488081	213	319	260	418	EPAS1	Vitiligo
310	rs739134	22	42089623	214	352	256	418	C22orf46	CD
281	rs13003464	2	61186829	215	250	241	418	PUS10	CD
325.75	rs11584383	1	200935866	216	324	345	418	GPR25	CD
290.5	rs8090824	18	57147798	217	269	258	418	CCBE1	CD
294.5	rs876187	14	98478564	218	294	248	418	C14orf177	CD

292	rs1032070	17	40618251	219	276	255	418	ATP6V0A1	CD
291.5	rs10177578	2	217885659	220	290	238	418	TNP1	Gerd
249.5	rs11165328	1	95630461	221	217	142	418	TLCD4, TMEM56	CD
139.25	rs3742704	14	88477882	222	161	106	68	GPR65	CD
313	rs59278059	8	49589181	223	337	274	418	EFCAB1	CD
234.5	rs735277	8	27534727	224	346	216	152	CCDC25	RA
152	rs7758736	6	32758394	225	190	107	86	HLA-DQA2	CD
304	rs4572134	11	42916859	226	271	301	418	API5	CD
340.25	rs7919913	10	5926216	227	389	327	418	ANKRD16	CD
260.25	rs1944564	18	75071036	228	210	185	418	GALR1	Spondylosis and allied disorders
336.5	rs7704367	5	158821493	229	281	418	418	IL12B	CD
318.75	rs2711981	4	39039258	230	390	237	418	TMEM156	CD
186.75	rs11066301	12	112871372	231	228	188	100	PTPN11	CD
312.25	rs9483751	6	135128858	232	350	249	418	ALDH8A1	CD
302.75	rs1551398	8	126540051	233	298	262	418	TRIB1	CD
201.75	rs1536780	13	94708442	234	242	205	126	GPC6	CD
327.25	rs1015563	20	6690101	235	326	330	418	BMP2	CD
194.75	rs10822047	10	64424284	236	191	244	108	ADO	CD
133.75	rs12547052	8	84492735	237	176	82	40	RALYL	
330.75	rs605790	6	142879589	238	347	320	418	ADGRG6	UC
332.25	rs682666	13	101706701	239	353	319	418	NALCN	
371.25	rs9842389	3	173276758	240	409	418	418	NLGN1	
87.25	rs115884658	6	31864538	241	30	49	29	ATP6V0A1	CD
185.5	rs753173	10	30778738	242	264	155	81	MAP3K8	
302.5	rs2823289	21	16841195	243	199	350	418	NRIP1	CD
283.25	rs13064993	3	149403563	244	280	191	418	WWTR1	CD
293.75	rs451686	12	70840890	245	209	303	418	KCNMB4	CD
241.75	rs4532399	5	40467272	246	97	206	418	PTGER4	CD
247.5	rs7589485	2	28645120	247	148	177	418	FOSL2	CD
311.25	rs34762726	3	49689210	248	202	377	418	BSN	CD
232.5	rs1860180	17	32628064	249	282	261	138	CCL11	CD
247.75	rs7076156	10	64415184	250	127	196	418	ZNF365	CD
284.5	rs1396733	2	28642747	251	197	272	418	FOSL2	CD
339.5	rs7786795	7	51721558	252	370	318	418	COBL	Acute infective polyneuritis/guillain-barre syndrome
359.75	rs1886684	1	92975938	253	398	370	418	EVI5	CD
263.75	rs6651252	8	129567181	254	184	199	418	MYC	CD
300.5	rs61736408	11	35456061	255	297	232	418	PAMR1	Spondylosis and allied disorders

168.25	rs3091316	17	32593974	256	102	218	97	CCL2	CD
282.5	rs3805495	5	40755568	257	226	229	418	TTC33	CD
181	rs61839660	10	6094697	258	267	132	67	IL2RA	CD
343.25	rs2301436	6	167437988	259	356	340	418	CEP43, FGFR1OP	CD
260	rs10077785	5	131801158	260	194	168	418	IRF1	CD
228.25	rs6845304	4	88280502	261	279	224	149	HSD17B11	UC
305.75	rs11178010	12	70366794	262	300	243	418	MYRFL	CD
267.25	rs8113472	19	10608064	263	205	183	418	KEAP1	CD
154.25	rs1051738	19	10577843	264	154	116	83	PDE4A	
365.75	rs10484530	6	167461562	265	362	418	418	CEP43, FGFR1OP	CD
94.5	rs6710480	2	158958605	266	61	34	17	UPP2	
197.75	rs790088	17	71588183	267	277	158	89	SDK2	
345	rs7469903	9	105524913	268	373	321	418	CYLC2	MS
251	rs11139654	9	85220698	269	174	143	418	RASEF	
325	rs9303898	18	2312881	270	358	254	418	METTL4	CD
366	rs11808092	1	93073228	271	416	359	418	EVI5	CD
233	rs2066842	16	50744624	272	86	156	418	NOD2	CD
279.25	rs7749057	6	32448904	273	201	225	418	HLA-DRA	
285.5	rs7138344	12	70837122	274	179	271	418	KCNMB4	CD
303.5	rs1402246	3	34589121	275	262	259	418	PDCD6IP	CD
318	rs11244	6	32780724	276	285	293	418	HLA-DOB	CD
317.25	rs6073315	20	42718605	277	309	265	418	JPH2	Allergy or anaphylactic reaction to drug
342	rs7608910	2	61204856	278	284	388	418	PUS10	CD
124.75	rs9292777	5	40437948	279	74	66	80	PTGER4	CD
225.5	rs4263839	9	117566440	280	103	101	418	TNFSF15	CD
206	rs740495	19	1124835	281	221	186	136	SBNO2	CD
199.5	rs12624279	2	28634790	282	186	207	123	FOSL2	CD
306.5	rs11745587	5	131796922	283	275	250	418	IRF1	CD
356.25	rs1736168	2	71880211	284	371	352	418	DYSF	
378.25	rs3099844	6	31448976	285	392	418	418	MICB	CD
330.5	rs6913309	6	32339840	286	314	304	418	HLA-DRB5	CD
372.25	rs1926554	10	35344969	287	366	418	418	CUL2	CD
118	rs4655690	1	67659896	288	55	56	73	IL23R	CD
324.5	rs9981974	21	16759963	289	257	334	418	NRIP1	CD
302.75	rs6441841	3	44450564	290	198	305	418	TCAIM	Acute gastritis
268.75	rs2241874	2	234247627	291	130	236	418	SAG	CD
383.75	rs181362	22	21932068	292	407	418	418	UBE2L3	CD
309	rs4871611	8	126537570	293	272	253	418	TRIB1	CD
330.75	rs13251655	8	119760559	294	323	288	418	TNFRSF11B	CD

337	rs2844702	6	30912481	295	338	297	418	MUCL3, DPCR1	CD
335	rs6850861	4	178236402	296	348	278	418	NEIL3	
341.75	rs9405897	6	6052055	297	335	317	418	NRN1	Spondylosis without myelopathy
304.25	rs7217052	17	21452282	298	261	240	418	C17orf51	CD
337.75	rs263156	6	142907515	299	306	328	418	ADGRG6	UC
327	rs7562674	2	172259687	300	311	279	418	METTL8	CD
220.5	rs17399261	2	222994084	301	278	189	114	CCDC140	CD
327.5	rs7775397	6	32261252	302	222	368	418	TSBP1, C6orf10	CD
291.5	rs4378078	9	139375962	303	188	257	418	SEC16A	CD
342.25	rs1046974	2	234255547	304	274	373	418	SAG	CD
376.75	rs12985380	19	51861176	305	400	384	418	ETFB	CD
309.5	rs17215589	3	41831203	306	245	269	418	BMS1P4	CD
346.25	rs2293158	17	40447558	307	321	339	418	STAT5A	CD
229.25	rs593400	10	30762088	308	310	194	105	MAP3K8	CD
382.75	rs7756521	6	30848253	309	386	418	418	DDR1	CD
368.5	rs11008080	10	30802799	310	365	381	418	MAP3K8	CD
123	rs6576498	15	26223788	311	98	57	26	ATP10A	Psoriasis
307	rs7736920	5	40520217	312	204	294	418	PTGER4	CD
316.25	rs1383261	6	32765451	313	259	275	418	HLA-DQB1	CD
174	rs2522051	5	131797578	314	109	144	129	IRF1	CD
367	rs170773	8	19008265	315	355	380	418	PSD3	CD
325	rs389884	6	31940897	316	293	273	418	DXO	CD
154.25	rs1992660	5	40415067	317	56	114	130	PTGER4	CD
266.25	rs9822268	3	49719729	318	177	152	418	APEH	CD
298.25	rs1885276	1	67568824	319	167	289	418	C1orf141	CD
383.5	rs12240347	10	35359475	320	378	418	418	CUL2	CD
209	rs10995251	10	64398466	321	211	154	150	ADO	CD
302	rs12789493	11	76275703	322	158	310	418	LRRC32	CD
351	rs4821116	22	21973319	323	357	306	418	UBE2L3	CD
319.5	rs1884444	1	67633812	324	162	374	418	IL23R	CD
361.25	rs7246953	19	10621108	325	339	363	418	TYK2	CD
325	rs11209030	1	67737775	326	213	343	418	IL12RB2	CD
181.25	rs3785142	16	50787147	327	90	160	148	CYLD	CD
323.5	rs10489631	1	67601115	328	215	333	418	C1orf141	CD
315	rs1321157	1	67654110	329	181	332	418	IL23R	CD
129	rs6896969	5	40424426	330	72	71	43	PTGER4	CD
266.25	rs10781499	9	139266405	331	155	161	418	CARD9	CD
371	rs6456156	6	167522300	332	376	358	418	CCR6	CD
327.25	rs2675677	10	75648249	333	212	346	418	PLAU	CD

327	rs6981209	8	13416666	334	258	298	418	DLC1	Ulcerative colitis (chronic)
324	rs7563345	2	234143244	335	220	323	418	ATG16L1	CD
372.5	rs3117582	6	31620520	336	318	418	418	APOM	CD
174.5	rs2201841	1	67694202	418	76	100	104	IL23R	CD
191.5	rs3197999	3	49721532	418	144	127	77	MST1	CD
211	rs4077515	9	139266496	418	178	173	75	CARD9	CD
225	rs162907	5	131580152	418	189	197	96	PDLIM4	CD
237	rs3024493	1	206943968	418	287	172	71	IL10	CD
244.25	rs9611613	22	41961831	418	315	166	78	CSDC2	CD
256.75	rs12540583	7	102760511	418	322	208	79	NAPEPLD	CD
273.5	rs11955347	5	131567924	418	233	296	147	P4HA2	CD
277.5	rs7211774	17	56257984	418	163	111	418	TSPOAP1	IBD
287.75	rs3859118	16	50252235	418	136	179	418	TENT4B, PAPD5	CD
295.25	rs4785452	16	50842077	418	106	239	418	CYLD	CD
325	rs7936562	11	76278258	418	172	292	418	LRRC32	CD
326.25	rs2136187	5	131577894	418	160	309	418	P4HA2	CD
326.5	rs3131383	6	31704294	418	243	227	418	CLIC1	CD
329.25	rs80043692	17	56247306	418	173	308	418	OR4D2	
331.5	rs3849969	10	75525999	418	214	276	418	SEC24C	CD
331.75	rs1736135	21	16805220	418	131	360	418	NRIP1	CD
334	rs1876143	5	40521648	418	185	315	418	PTGER4	CD
338	rs3131379	6	31721033	418	246	270	418	MSH5	CD
338.5	rs10065787	5	131436486	418	207	311	418	SLC22A4	CD
343.75	rs7381376	6	32767673	418	225	314	418	HLA-DQB1	CD
348.25	rs11715915	3	49455330	418	343	214	418	AMT	CD
351.25	rs10740418	10	75519322	418	286	283	418	SEC24C	CD
356	rs2167566	2	61519408	418	263	325	418	USP34	CD
359.25	rs3117577	6	31727474	418	302	299	418	MSH5	CD
360	rs6431655	2	234162415	418	229	375	418	ATG16L1	CD
360.25	rs4746143	10	75477298	418	289	316	418	BMS1P4	CD
360.5	rs181359	22	21928641	418	299	307	418	UBE2L3	CD
361.25	rs558702	6	31870326	418	223	386	418	ZBTB12	CD
361.5	rs2518934	2	61794498	418	368	242	418	XPO1	CD
363	rs3117574	6	31725230	418	244	372	418	MSH5	CD
363.25	rs9276644	6	32745043	418	333	284	418	PSMB9	CD
364.75	rs11800409	1	93181013	418	397	226	418	EVI5	CD
369	rs4795893	17	32574448	418	316	324	418	CCL2	CD
369.25	rs9366076	6	167373708	418	328	313	418	RNASET2	CD
371.5	rs11799915	1	197582780	418	359	291	418	DENND1B	CD
371.75	rs2241876	2	234186734	418	313	338	418	ATG16L1	CD

372	rs497309	6	31892484	418	340	312	418	DXO	CD
373	rs3816769	17	40498273	418	388	268	418	STAT3	CD
373.25	rs915652	6	31749142	418	239	418	418	VARS1, VARS	CD
375.5	rs272885	5	131667736	418	283	383	418	SLC22A4	CD
375.75	rs7004689	8	119769036	418	367	300	418	TNFRSF11B	CD
376.25	rs2187668	6	32605884	418	251	418	418	HLA-DQA1	
380.25	rs3095352	6	30805921	418	354	331	418	IER3	CD
381	rs3115671	6	31734345	418	270	418	418	VWA7	CD
382	rs11706370	3	49441091	418	345	347	418	RHOA	CD
382.25	rs3101017	6	31733466	418	344	349	418	VWA7	CD
382.5	rs460106	22	21806401	418	327	367	418	UBE2L3	CD
385	rs3130655	6	30823710	418	369	335	418	IER3	CD
388.25	rs2858331	6	32681277	418	341	376	418	HLA-DRB5	CD
389.25	rs17582416	10	35287650	418	385	336	418	CUL2	CD
389.5	rs3909130	6	30874165	418	351	371	418	GTF2H4	CD
390.25	rs4430924	2	61703856	418	396	329	418	USP34	CD
390.5	rs9276711	6	32757297	418	375	351	418	PSMB9	CD
392.5	rs12537160	7	51733827	418	408	326	418	COBL	
393.5	rs1270942	6	31918860	418	360	378	418	CFB	CD
394.5	rs126092	22	42178441	418	394	348	418	MEI1	CD
397.25	rs9468845	6	30869593	418	399	354	418	GTF2H4	CD
397.5	rs139553	22	42187199	418	412	342	418	MEI1	CD
398	rs3812594	9	139368953	418	374	382	418	SEC16A	CD
399.5	rs4821112	22	21964761	418	405	357	418	UBE2L3	CD
403.25	rs7099036	10	35349574	418	415	362	418	CUL2	CD
403.75	rs8408	6	30867666	418	361	418	418	GTF2H4	CD
404.5	rs2694642	2	61596180	418	413	369	418	C2orf74	CD
404.5	rs5754217	22	21939675	418	364	418	418	UBE2L3	CD
405	rs2074506	6	30890483	418	418	366	418	VARS2	CD
405.5	rs916920	6	30877202	418	401	385	418	GTF2H4	CD
407.75	rs2517449	6	30919701	418	377	418	418	MUCL3, DPCR1	CD
408.25	rs4618569	6	30855251	418	379	418	418	DDR1	CD
409	rs12529876	6	167461501	418	382	418	418	CEP43, FGFR1OP	CD
409.5	rs11498	2	61370819	418	384	418	418	C2orf74	CD
410.25	rs8139993	22	41995335	418	387	418	418	DESI1	CD
411.25	rs8082391	17	40398973	418	391	418	418	C22orf46	CD
412.25	rs2239517	6	30865115	418	395	418	418	DDR1	CD
414	rs2252760	6	30892377	418	402	418	418	VARS2	CD
414.25	rs2535327	6	30826904	418	403	418	418	DDR1	CD
414.5	rs1264309	6	30875899	418	404	418	418	VARS2	CD

415	rs1264303	6	30882513	418	406	418	418	VARS2	CD
416	rs1264333	6	30844314	418	410	418	418	DDR1	CD
417	rs2844654	6	30838688	418	414	418	418	VARS2	CD
417.75	rs7738138	6	30887344	418	417	418	418	VARS2	CD