UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Biomedicina Comparata ed Alimentazione

Corso di laurea magistrale in Biotecnologie per l'Alimentazione

"Polygenic risk score" computation for Inflammatory Bowel Disease using an Illumina "custom array"

Relatore
Prof. Luca Bargelloni
Correlatore
Dott.ssa Barbara Arredi

Laureando
Federico Tamiazzo
Matricola n. 2039047

ANNO ACCADEMICO 2022/2023

TABLE OF CONTENTS

ABSTRACT

Over the past 15 years Genome-Wide Association Studies (GWAS) have produced a huge amount of data on genetic loci associated with quantitative traits, including food intolerance and other digestive disorders. Usually, the variants identified by GWAS are not directly responsible for the phenotype, but they happen to be near the causative allele, at least in some of the haplotypes present in the population. Therefore, the contribution of a variant can be weak, depending on its actual association with the causative allele. Furthermore, quantitative traits are typically resulting from the contribution of several genes, each with a small impact to the final phenotype.

The aim of this project is to analyze public GWAS and other genetic data to extract useful information about variants involved in food intolerance and digestive disorders with a particular focus on Crohn's disease. In particular, the study will make use of a set of about 2100 variants included on an Illumina custom array designed by BMR Genomics. Algorithms for the calculation of polygenic risk scores will be defined for the variants associated with Crohn's disease and will be tested on a genotyped population.

Negli ultimi 15 anni i Genome-Wide Association Studies (GWAS) hanno prodotto un'enorme quantità di dati sui loci genetici associati a tratti quantitativi, tra cui l'intolleranza alimentare e altri disturbi digestivi. Di solito, le varianti identificate dai GWAS non sono direttamente responsabili del fenotipo, ma si trovano vicino all'allele causale, almeno in alcuni degli aplotipi presenti nella popolazione. Pertanto, il contributo di una variante può essere debole, a seconda della sua effettiva associazione con l'allele causale. Inoltre, i tratti quantitativi sono tipicamente il risultato del contributo di diversi geni, ciascuno con un piccolo impatto sul fenotipo finale.

L'obiettivo di questo progetto è analizzare GWAS pubblici e altri dati genetici per estrarre informazioni utili sulle varianti coinvolte in intolleranze alimentari e patologie del tratto gastrointestinale, con particolare attenzione al morbo di Crohn. In particolare, lo studio si avvarrà di un set di circa 2100 varianti incluse in un custom array Illumina disegnato da BMR Genomics. Gli algoritmi per il calcolo del polygenic risk score saranno definiti per le varianti associate al morbo Crohn e saranno testati su una popolazione genotipizzata.

1. INTRODUCTION

1.1 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a term that describes a chronic inflammation affecting the gastrointestinal tract. IBDs are mainly divided into ulcerative colitis (UC) and Crohn's disease (CD). They present similar symptoms, including weight loss, abdominal pain, diarrhea and rectal bleeding. Inflammatory bowel disease is most often diagnosed during young adulthood and adolescence. Infact, about a quarter of patients with IBD are diagnosed before age 20 years and there is a rising incidence in pediatric patients (Rosen et al. 2015)[1].

1.1.1 Main differences between ulcerative colitis and Crohn's disease

Although both diseases concern mainly the gastrointestinal tract with a variety of extraintestinal manifestations (mainly osteopenia/osteoporosis, arthritis, spondylitis, erythema nodosum, iritis and uveitis), they present some differences.

Crohn's disease can potentially interests all the gastrointestinal tract but most commonly involves the terminal ileum and colon. Ulcerative colitis is limited to the colon and always interests the rectum. Sometimes it's difficult to distinguish the two diseases, especially when Crohn's disease is limited only to the colon. CD produces discontinuous lesions of the walls with the potential formation of stenosis and/or fistulae. Instead UC leads to less deep but continuous damages and those lesions are limited to the colon mucosa. At the chronic stage, CD presents more severe symptoms than UC. Some of which are continuous vomiting, loss of appetite, fatigue, fever, abscess, bowel obstruction, painful diarrhea and severe abdominal pain.

1.1.2 Causes of the disease

Genetic and environmental factors are thought to play a role, but specific triggering events are yet to be identified. Podolsky and colleagues suggest that the pathogenicity in IBD depends on various factors: microflora of the intestine, patient's susceptibility and mucosal immunity (Podolsky D. K. 2002)[2].

Infact, it is demonstrated that patients with CD or UC are characterized by an alteration of microbial flora composition. An example of this dysbiosis was published by Marteau and

colleagues, showing that *E. coli* and Bacteroidetes were higher in patients with inflammatory bowel disease (Marteau P. et al. 2004)[3]. Some common pathogenic bacteria that may be involved in IBD are *Salmonella, Campylobacter, Yersinia, Shigella, Aeromonas* and the already mentioned *E. coli*. Some strains of *Bifidobacterium* and *Lactobacillus* are protective against IBD and have an important role in prevention (Marteau P. et al. 2004)[3]. Moreover, Prideaux and colleagues showed that other factors like diseases, ethnicity and geography have strong effects on the composition and diversity of the gut microbiota. So those elements may be critical in shaping emerging patterns of IBDs (Prideaux L. et al. 2013)[4]. Even considering these correlations between microflora and inflammatory bowel disease, there is still no result that attests how the microorganisms are directly involved in the development of the diseases.

Furthermore intestinal permeability plays a crucial role for the development of inflammatory bowel disease. A defective mucosal barrier leads them to the exposition to luminal content and triggers an immunological response. This promotes the characteristic inflammation found in ulcerative colitis and Crohn's disease.

Another important factor in IBD regards nutrition and our eating habits. Infact, diets rich in sugars and fatty acids compounds and poor in vegetal fibers greatly increase the prevalence of inflammatory bowel diseases (Thornton J. R. et al. 1979)[5]. Diet has also an important impact on microbial composition, the integrity of intestinal barrier and host immunity. Infact, the excessive intake of specific food groups like fat and sugars may promote gut dysbiosis. This leads to an alteration of gut barrier, immune response and tissue damage and can have a role in the development of IBD (Roncoroni L. et al. 2022)[6]. Moreover, Pedersen and colleagues demonstrated that a diet with low levels of FODMAPs (Fermentable, Oligosaccharides, Disaccharides, Monosaccharides and Polyols) can reduce gastrointestinal symptoms in IBD patients, generally improving quality of life (Pedersen N. et al. 2017)[7].

Furthermore, Corrao demonstrated that smoking and oral contraception increases the risk of developing UC and CD while breastfeeding in infancy is protective against IBD (Corrao G. et al. 1998)[8]. In particular, smoking increases the amount of CD4+ T cells that are part of the white blood cells, promoting the interferon gamma proinflammatory proteins release in the lungs. In a second moment they will move to the intestine, causing inflammation.

1.1.3 Genetic of IBD

As previously mentioned, genetics play an important role in IBD susceptibility. Some of the most important genes for Crohn's disease and ulcerative colitis are listed below (Younis N. et al. 2020)[9].

- NOD2 (nucleotide binding oligomerization domain containing 2): the first gene that has been associated with Crohn's disease. It encodes a protein with six leucine-rich repeats (LRRs) and two caspase recruitment domains (CARD). In the immune response the encoded protein recognize the muramyl dipeptide (MDP) derived from the lipopolysaccharides of the bacterial cell membrane and activate the NFKB protein (Lauro M.L. et al. 2016)[10];

- ATG16L1 (autophagy related 16 like 1): the respective protein is part of a protein complex that is necessary for autophagy: the process responsible for the degradation of intracellular components in lysosomes (Hamaoui D. et al. 2022)[11];

- IL10 (interleukin 10): encodes for a protein primarily produced by monocytes: an important cytokine for immunoregulation and inflammation. This protein enhances B cells survival, proliferation and antibody production. It is also able to stop NF-kappa B activity and down-regulates the expression of Th1 cytokines and costimulatory molecules on macrophages (Steen E. H. et al. 2019)[12];

- IL10RA (interleukin 10 receptor subunit alpha): encodes a receptor for interleukin 10. The protein regulates the synthesis of proinflammatory cytokines by the mediation of the immunosuppressive signal of interleukin 10. Moreover, this receptor promotes survival of progenitor myeloid cells with the phosphorylation of JAK1 and TYK2 kinases (Al-Abbasi F. A. et al. 2018)[13];

- IL10RB (interleukin 10 receptor subunit beta): the encoded protein is part of the cytokine receptor family . The coexpression of this and IL10RA proteins are required for IL10-induced signal transduction (Ahn D. et al. 2020)[14];

- IRGM (immunity related GTPase M): this gene encodes a member of the p47 immunity-related GTPase family. It is important for in innate immunity response by the autophagy regulation in response to intracellular pathogens (Nath P. et al. 2020)[15];

- LRRK2 (leucine rich repeat kinase 2): member of the leucine –rich repeat kinase family. The protein is largely expressed in cytoplasm and also associates with the mitochondrial outer membrane. Mutations in LRRK2 have been identified as a genetic risk factor for both sporadic and familial Parkinson's disease (Zhang X. et al. 2023)[16];

- PTPN2 (protein tyrosine phosphatase non-receptor type 2): this gene encodes for a member of the protein tyrosine phosphatase (PTP) family which protects the IEC barrier from inflammation-induced disruption and regulates macrophage functions (Spalinger M. R. et al. 2020)[17];

- IL23R (interleukin 23 receptor): the protein encoded by this gene and the protein encoded by IL12RB1 (interleukin 12 receptor subunit beta 1) are both subunits of the receptor IL23A/IL23. This protein associates with janus kinase 2 and binds to transcription activator STAT3 in IL23A signaling (Subhadarshani S. et al. 2021)[18];

- CDH1 (cadherin 1): encodes a cadherin, a calcium-dependent cell-cell adhesion protein. Loss of function of this gene is thought to contribute to cancer progression by increasing proliferation, invasion and metastasis. (Hansford S. et al. 2015)[19];

- HNF4α (hepatocyte nuclear factor 4 alpha): the encoded protein is a nuclear transcription factor which binds DNA as a homodimer, controlling in this way the expression of several genes (ie. hepatocyte nuclear factor 1 alpha which encodes for a transcription factor that regulates the expression of several hepatic genes). This gene may play a role in the development of the liver (Yu Y. et al. 2022)[20].

Genes IBD-related often encodes for proteins involved in innate or adaptive immunity and this highlights the possible correlation between genetic factors and inflammation.

Those genes, as all the genes of all genomes, can have permanent changes in the DNA sequence caused by mutations and or inherited from a parent.

The next chapter demonstrates the importance of the genetic variants of the listed genes and how they are related to Crohn's disease. For this reason DNA sequencing plays a crucial role to study genetics of complex diseases like IBDs.

1.2 Genome-Wide Association Studies

A common way to study the relation between genetic variants and a specific trait/disease is achieved by testing hundreds of thousands of variants across many genomes to find those statistically associated with the trait/disease through a Genome-Wide Association Study (GWAS). 'A Genome-Wide Association Study (abbreviated GWAS) is a research approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait' (Hutter C. M. National Human Genome Research Institute, 2023)[21]. The first study was published in 2005 and the GWAS Catalog[29] is now including 6499 publications, reporting 539949 top associations. So, GWAS strategy is based on a high throughput screening approach. These are complex and financially demanding studies because they require the typing of hundreds of thousands of gene loci in tens of thousands of patients. Since the loci responsible for genetic characteristics are not known a priori, GWAS generally do not aim at causative loci, i.e. directly responsible for genetic characteristics, but only at the identification of indirect associations.

This method can give information about genotype-phenotype correlations by testing for differences in the allele frequency of genetic variants between individuals from the same geographical area. Even if genetic loci listed in GWAS are informative on association but not on causation about pathological mechanisms, it allows the study of complex genetic diseases like CD. Moreover, variants associated with the disease and causative variants can be in linkage. As an example of this, a Genome-Wide Association Study implicated the IL-12/IL-23 pathway in the development of Crohn's disease (Luo Y. et al. 2017)[22]. Infact, interleukin 12 and 23 are important cytokines with a crucial role in the regulation of tissue inflammation. This study and the biology of IL-12/IL-23 influenced the development of therapeutic strategies and clinical trials in IBD (Moschen A. R. et al. 2019)[23].

Therefore, Genome-Wide Association Studies constitute one of the most relevant strategies of post-genomics research and are at the basis of the discovery of new genetic biomarkers.

1.2.1 Conducting GWAS

The first step is to select the population. This is important because very large sample sizes allow to identify reproducible genome-wide associations. Software like Genetic Power Calculator[30] can be used to define the correct sample size. There are a lot of public resources available, for example the UK Biobank[31], that provides access to large cohorts with both genotypic and phenotypic information. This data source is cheaper and more rapid than assembling de novo a dataset to conduct a GWAS. To avoid false positives, the population substructure can be considered including different ethnicities in the same study. Moreover, it should be considered that not all human populations and subpopulations have the same haplotype structure.

Another important point of GWAS is genotyping: the detection of small genetic differences between individuals' genotypes and a reference genome that can lead to major changes in phenotype, in our case a pathological phenotype. Human bead arrays are the ideal method for genotyping thousands of individuals for millions of known variants with relatively lower cost compared to other whole genome sequencing methods. The latter are composed by a very large set of molecular probes placed on a solid surface that can bind a complementary sequence of fragmented DNA of an individual. This allows to identify single nucleotide polymorphisms (SNPs) and to find genetic variants of each individual. Each genetic locus makes such a small contribution to disease susceptibility and this is the reason why there is a need to expand as much as possible the number of the analyzed SNPs.

Another common technique for genotyping is the whole genome sequencing (WGS), a method that leads to the determination of nearly the entirety of the DNA sequence of an individual's genome at a single time. This strategy allows to include rare variants in our study: alternative forms of a gene with a minor allele frequency (MAF) of less than 1%. This is advantageous when rare variants make a substantial contribution to the trait of interest.

After genotyping we will have the individual ID numbers, coded family relations between individuals, sex, phenotype information and genotype calls for all called variants (Uffelman E. et al. 2021)[24]. All this information is necessary as input to conduct GWAS, but only after some quality control (QC) steps, including removing variants that are not in Hardy-Weinberg (HW) equilibrium. The latter is a model in population genetics according to which in an ideal population there is equilibrium between allelic and genotypic frequencies between consecutive generations. In particular, a variant is in HW equilibrium if the frequency of observed genotypes of the variant in a population can be derived from the observed allele frequencies (Uffelman E. et al. 2021)[24]. Other

important quality control steps to filter the input GWAS data expect the removal of SNPs that are missing in some individuals, the removal of eventual genotyping errors and the assurance that phenotypes correspond to the possessed genetic data. The latter is often done with a comparison between sex based on the X and Y chromosomes and the self reported sex of the individuals ("sex check"). After QC, variants are subjected to phasing and imputation. The first one, allows to estimate which of the genotyped alleles derive from the maternal or paternal allele. The second leads to attribute missing genotypes with tools like IMPUTE2[32], MACH[33], Beagle 4.1[34], Minimac4[35] and SHAPEIT2[36], which even give a score for the quality of the imputation (INFO), useful for next post GWAS analysis. An haplotype is a set of DNA variants along a single chromosome that tend to be inherited together. Imputation needs a reference haplotype panel like TOPMed[37]. Then a principal component analysis (PCA) is conducted to identify and exclude possible outliers. Population stratification is the presence of genetically distinct subpopulations that differ in their mean phenotypic values and as we already said it should be considered in this step to avoid false positives.

1.2.2 Association testing

At this point we can finally test the associations between variants and phenotype. To reach this aim regression models are used. In our case the phenotype is binary (presence or absence of Crohn's disease), so a logistic regression model is used, which means that will be estimates the probability of the disease based on a specific dataset of independent variables. In statistics there are explanatory variables and response variables. Explanatory variables explain the variation in the response variable, but some other variables may exist that also affects the response variables: the covariates. In this case, sex, age and ancestry can be considered as covariates to increase statistical power.

If the phenotype is continuous (ie. body mass index or bloody pressure) a linear regression model is used. Those statistical methods won't be discussed, but it is important to highlight that in logistic regression model a "logit link function" is used. This one is a function that converts a linear combination of covariate values into probabilities.

After the association analysis, the effect size (beta or ES) of each variant will be obtained. This is an estimation for the association of each genetic marker with the trait of interest. When this estimate refers to a risk factor and quantifies the increased odds of having a disease per risk allele count,

this is called odds ratio (OR). The latter is an important element for the future calculation of the polygenic risk score (PRS) and this will be discussed in chapter 1.4.

Next step of GWAS can be the "genome-wide meta analysis" and it serves to increase again the sample size. It is carried out analyzing together data from multiple cohorts and applying on them the same standardized quality control pipelines. Another optional point can be the replication of GWAS increasing again the sample size with the aim to obtain more generalizable results.

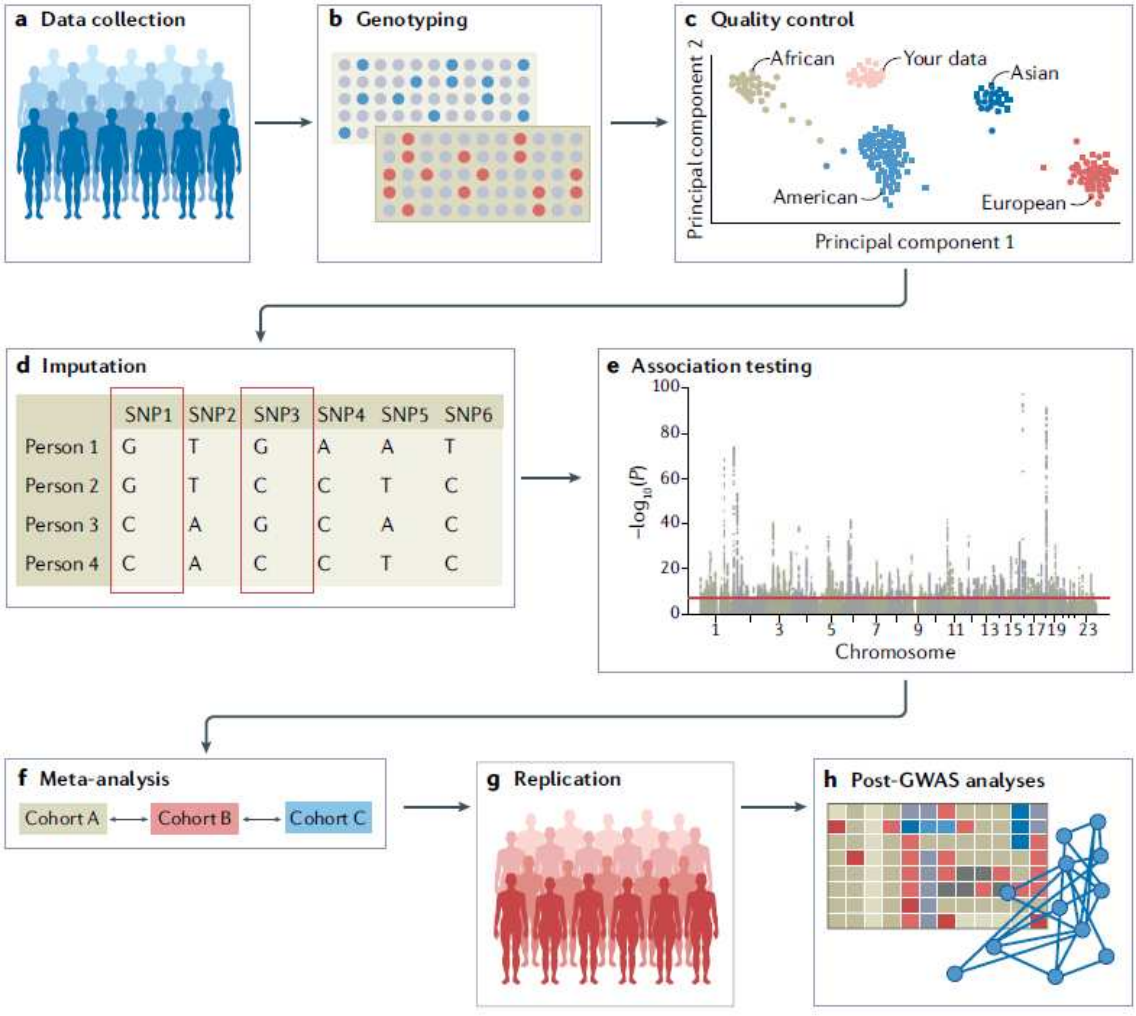

Fig. 1 | *Overview of steps for conducting GWAS (Uffelmann E. et al. 2021)*[24]

1.2.3 Summary statistics

After a Genome-Wide Association Study we will have a primary outcome called summary statistics. These contain some useful data for future analyses in which them will be also called "base data". First of all, those files show a list of SNP IDs, namely the main way to identify SNPs that passed GWAS quality control steps and associated to the disease. This nomenclature starts with "rs", which means "reference SNP", and this abbreviation is followed by a number that uniquely identifies a specific single nucleotide polymorphism. The second information always given from summary statistics is the location of the SNP on the genome followed by the genomic build. Those coordinates integrate the rd IDs and complete the basic information on the SNPs that passed all the quality control steps and filters of GWAS.

Then the possible alleles and the strand are listed. There is an "effect allele" which usually is the less frequent of all the possible alleles and the one that gives the disease or the considered trait, and a "non-effect allele". The latter is often considered protective against the analyzed phenotype. In some cases even the "effect allele" can be protective for the trait, and this will be reported in literature after GWAS.

In summary statistics there are always effect sizes (beta) or odds ratio that gives a numerical value to the association between the SNP and the disease. In addition to those values, even P-value and standard error are listed for each rs. The first one is a number that identifies the SNPs that are closer to the causal variants and their possible convergence in biological pathways. A causal variant is a single nucleotide polymorphism that is responsible for a particular trait or disease and is related to P-value: the higher is this number, the closer the SNP will be to the causal variant. Instead, standard error describes the accuracy with which our sample distribution represents the analyzed population. It is calculated with the square root of the standard deviation (a measure of how much the value deviates from the average) divided by the number of samples. The latter and the minor allele frequency (MAF) are usually the last information given in summary statistics. MAF represents the frequency at which the less common allele occurs in the population. It is calculated dividing the alleles positive for the variant by the total number of alleles screened. This value gives us an idea on the presence of the variant associated with the disease in the population. In the end, in summary statistics are reported the imputation information score, usually listed under the heading "INFO". A common way to represent the results of a GWAS is the Manhattan plot, in which the results of the associations between traits or diseases are graphically distinguished for their position (x-axes) and for the decimal logarithm of their P-value (y-axes).

chr	Bp location	SNP ID	Other allele	Effect allele	MAF	beta	Standard error	p_value	INFO
10	88766	rs55896525	T	C	0.945194	0.077287	0.07638	0.31161116	0.91642
10	90127	rs185642176	T	C	0.913977	-0.105855	0.05986	0.0769786	0.98473
10	90164	rs141504207	G	C	0.916561	-0.129943	0.06263	0.03800844	0.92349
10	94263	rs184120752	A	C	0.974493	0.0394445	0.10739	0.71339818	0.95929

Table 1 | *Example of summary statistics derived from GWAS (Zorina-Lichtenwalter L. et al. 2023)*[25]

1.3 Genotyping

In addition to summary statistics, information on the genotype of individuals of interest are needed to compute a polygenic risk score (PRS) and they will be the "target data" of the next analysis.

Molecular genetics is one of the sectors in which the technological innovations of recent years have had the greatest impact. The sequencing of the entire human genome costs over one hundred million euros in the early 2000s, when it was first carried out, while now it can be obtained for less than one thousand euros, even from BMR Genomics[38]. These technological advances open up new application perspectives which, according to many analysts, are having a great effect on medical diagnostics.

Infact, thanks to the knowledge emerging from GWAS, diagnostic Beadchips has realized and those allows to identify causative loci, directly responsible for the characteristics of interest. In this regard, Illumina[39] has created a Beadchip called "Infinium Global Diversity Array with Cytogenetics-8"[40] which includes many loci of clinical interest. In addition, Illumina[39] offers the possibility to create "Custom BeadChips", relevant in this project, to identify variants in defined loci. Taking advantage of this possibility, BMR Genomics[38] realized a custom chip called "Chrysalus" which contains more than 2000 SNPs related to nutrition traits and pathological conditions. This technology allows to obtain information about specific variants for each individual of interest and to conduct bioinformatic analyses.

1.4 Polygenic risk score

Summary statistics derived from the association testing are very important for the post-GWAS analyses made in silico using further data from external resources and the individuals haplotypes. Polygenic risk score (PRS) computation is one of them. PRS is a numeric value that predicts an individual's genetic predisposition for a certain trait or disease. It is being explored for potential clinical application in personalized medicine such as predicting disease risk, tailoring treatment approaches or implementing preventive measures. It can be applied to a wide range of traits or diseases including common complex conditions like cardiovascular diseases, diabetes and psychiatric disorders. Each variant derived from Genome-Wide Association Studies has a different small effect on the trait in question and this is represented by the effect size taken from summary statistics. The weighted sum of the effects of an individual genetic variant gives the polygenic risk score. Infact, the basic equation for the PRS of an individual *j* is:

$$\text{PRS}_j = \sum_i^N \beta i * dosageij$$

where N is the number of SNPs in the score, βi is the effect size (beta) of variant i and $dosageij$ is the number of copies of SNP i of individual j. Therefore for this calculation we should have "base data", represented by summary statistics, and "target data" in which we find information about individuals genotypes and phenotypes. To manipulate those complex data files, there are a range of software tools but this thesis is focused on PLINK (v1.90 beta)[41], the most commonly used one. With those software we can filter data with some basic quality control (QC) steps.

1.4.1 Quality control of base data

Public base data files are usually compressed to reduce storage space requirements. So first you have to make sure that the downloaded file is not corrupted and contains all the needed information for the PRS computation. The second thing to check is the genome build: target and base data must be on the same genome build. If not, tools like LiftOver[42] can be used across the datasets. After those primary filters, you can proceed with the real quality control passage. As already said, PLINK[41] can be useful and it groups them in one single step. In this phase, SNPs with

low values of MAF and INFO are removed from the dataset to increase statistical power and to avoid false positives. Usually, rs with MAF < 0.01 and INFO < 0.8 are filtered. Then "strand flipping" is recommended (here or during QC of target data) if there are SNPs with mismatching alleles to obtain their complementary. PLINK[41] also automatically removes non-resolvable mismatching SNPs.

In base data duplicate variants can be reported and these rs must be removed, because most of the PRS software do not work with duplicate SNPs. Usually it is assumed that there are no overlapping samples between datasets. In addition, if the base and target data were produced with different genotyping chips, we don't know the chromosome strand used for either and it will be unknown if the base and target data are referring to the same allele. Then it is not possible to pair-up the alleles of ambiguous SNPs across the datasets and only non ambiguous rs must be retained.

1.4.2 Quality control of target data

The aim of this section is to ensure that all individuals and variants included in the study have high quality data. As already said for GWAS, even sample size is also important to obtain reliable results and for this reason Choi suggests performing PRS analyses on target data of at least 100 individuals (Choi S. W. et al. 2020)[27].

After checking that the target data file has not changed during possible local transfers, quality control steps should be performed. As already mentioned, PLINK[41] groups them in a single passage, removing SNPs with low MAF, low genotyping rate, out of Hardy-Weinberg Equilibrium and filtering out individuals with low genotyping rate (Marees A. T. et al. 2017)[28]. With this step we remove from our analysis possible genotyping errors, SNPs that are missing in a high part of analyzed subjects and individuals who have a high rate of genotype missingness. The next check regards the heterozygosity rates of individuals which could indicate DNA contamination or high levels of inbreeding when it's value is too big or too low. So first highly correlated SNPs are removed. This "pruning" passage is usually done in PLINK[41] filtering rs with Linkage Disequilibrium (LD) r^2 higher than 0,25 (Marees A. T. et al. 2017)[28]. Then, after computing heterozygosity rates, with the help of R commands Marees suggest to remove individuals with F coefficients (estimates for assessing heterozygosity) that are more than 3 standard deviation units from the mean

(Marees A. T. et al. 2017)[28]. Even target data requires the removal of possible duplicate SNPs and a filter of the closely related individuals.

Another thing that can lead to invalid results is mislabeling or misreporting samples. To recognize them, a sex check-is usually performed as already said for GWAS.

1.4.3 PRS computation

At this point, various files should have been generated:

- The post-quality control summary statistics file;
- The filtered genotype file;
- The file in which are listed SNPs that passed QC steps;
- A file with filtered samples;
- The phenotype file;
- An eventual file that contains the covariates of the samples.

In addition, if the effect size relates to disease risk and is given as an odds ratio (or beta for continuous traits), the PRS is computed as a product of ORs. This calculation can be simplified transforming ORs in their natural logarithm so polygenic risk score can be computed using a summation. Even in this case, and for the next steps PLINK[41] can be useful, for example to simplify clumping. Linkage disequilibrium (LD) measures non-random association between alleles at different loci at the same chromosome in a population. Inter alia, SNPs are in LD when the frequency of association of their alleles is higher than expected with random assortment. Clumping is important to retain only weakly correlated SNPs and simultaneously maintaining those that are most associated with the phenotype.

At this point, with the transformed base data file, the file containing SNP IDs and their corresponding P-values, a file in which we put different P-value for inclusion of SNPs in the PRS, we can finally compute the polygenic risk score, generating a number of files which matches the number of P-value thresholds of the last file.

In particular, to compute PRS, PLINK[41] uses the following formula:

$$PRS_j = \frac{\sum_i^N S_i * G_{ij}}{P * M_j}$$

where S_i is the effect size of the variant i; G_{ij} is the number of effect alleles observed in sample j; the ploidy of the sample is P (for human is 2); N is the number of SNPs included in PRS; and M_j is the number of non-missing SNPs observed in sample j.

In conclusion, it is possible to approximate the "best fit PRS" to select the polygenic risk score that explains the highest phenotypic variance. To do this, we can perform a regression between PRS calculated at a range of P-value thresholds using R[43].

2. AIMS

A custom chip Illumina[39] called “Chrysalus” has been realized by BMR[39] Genomics. It contains more than 2000 variants related to various specific traits predominantly related to nutrition or pathological conditions related to inflammatory bowel disease. The objective of this thesis are to:

- identify between the SNPs of the custom array which can be related to inflammatory bowel diseases with a particular focus on Crohn’s disease;

- test those listed molecular markers on a wide range of samples and verify the relative probe efficiency and call rate;

- find other possible SNPs that could be included in a future update of Chrysalus to improve the reliability of the system;

- explore the possibility to extend the project and analyze GWAS data with the help of PLINK[41]. To reach this aim, various GWAS datasets related to inflammatory pathological conditions of the gastrointestinal tract should be evaluated. Moreover, those data will be used for a first polygenic risk score computation, which will be improved in future studies.

3. MATERIAL AND METHODS

3.1 CD-related SNPs and new molecular markers

"Chrysalus" is an Illumina[39] custom array realized by BMR[39] Genomics which contains 2036 different SNPs (3000 total SNPs, 964 duplicates). Those molecular markers have been selected for their specific associations with particular traits or phenotypes such as food intolerance, lipid and carbohydrates metabolism, oxidative stress, inflammation and Crohn's diseases (SNPs selected from: Sazonovs A. et al., *Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility*[26], 2022; Younis N. et al., *Inflammatory bowel disease: between genetics and microbiota*[9], 2020; https://genportal.tellmegen.eu/results/diseases/60/0[53])

Sometimes SNPs are not exclusively related to a single disease and can be linked to other traits. So the first step is to identify which of those variants have been associated with Crohn's disease in previous studies. For this purpose, some online databases such as NCBI[44] and SNPedia[45] were consulted to spot those variants. An initial "operative list" of SNPs CD-related was created which contains for each probe:

- rs ID;
- chromosome in which the variant is located;
- SNP position;
- gene symbol and name;
- variant function;
- risk alleles;
- diseases associated to the variant;
- other useful comments such as the risk increase of a specific genotype for Crohn's disease development;
- references about studies in which the variant is associated with CD.

Those information are required for the next analyses on polygenic risk score.

With the same online databases (SNPedia[45] and NCBI[44]), some other variants have been identified. Those SNPs will be included in the update of the 'Chrysalus' custom Beadchip array, so a list of "new molecular markers" was created with the same format of the previous table.

3.2 Infinium workflow

Chrysalus bead chip, which contains 3000 probes (2036 SNPs related to nutrition traits and pathological conditions and 964 duplicates/triplicates of "highly required SNPs" selected on the basis of BMR Genomics[39] and customers exigencies), has been used by BMR Genomics[39] to genotype 792 individuals by following Illumina Infinium protocol (*see 6.3*). All reagents are contained in Infinium DNA analyses assay kit and every kit allows to process 48 samples using two Beadchips. The workflow starts with DNA extraction from buccal swabs using Mag-Bind Blood & Tissue DNA HDQ 96 Kit following the attached protocol (*see 6.2*) and with a whole genome amplification performed according to the Infinium protocol. After the PCR reaction, the product of amplification is incubated overnight for 20/24 hours. Then a robust endpoint fragmentation is conducted on amplified DNA with restriction enzymes followed by alcohol precipitation and resuspension of fragments for the hybridization step. The resuspended DNA is loaded on the bead chip, each containing 24 samples. The latter is composed of microwells containing two identical probes (one for each allele) per fluorescently labelled bead type. Then hybridization of samples occurs in 17/24 hours. The following step is the extension with DNA polymerase and biotin-labelled or dinitrophenol-labelled dideoxynucleotides of samples previously annealed to locus-specific probes on Beadchip. If the probes don't match loci present in samples, no extension occurs. Staining allows the detection of the loci of interest exploiting the streptavidin-biotin/dinitrophenol (DNP) matrix formation. Biotin-labelled guanine and cytosine are associated with streptavidin with green fluorescence and anti-streptavidin-biotin. Dinitrophenol-labelled thymine and adenosine are instead recognized by anti-DNP that give red fluorescence and anti-Ab-DNP.

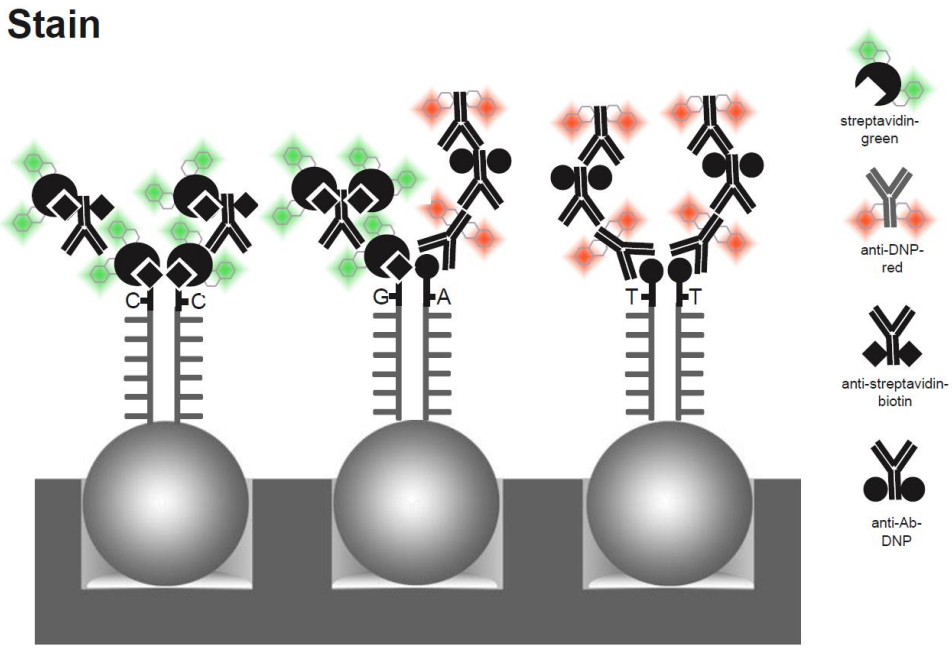

Fig. 2 | *Staining with streptavidin-biotin and anti-DNP-anti-Ab-DNP*

Intensity of the beads fluorescence is detected after placing the Beadchip in the iScan[46]. The latter is a scanner that uses lasers to excite labels and then takes an image. At this point genotypes are called automatically by Genome-Studio (v. 2.0.5)[47] by the different type and intensity of fluorescences. This software generates a final report which contains even call rates, information about quality of signals and the status of every probe related to Hardy-Weinberg equilibrium.

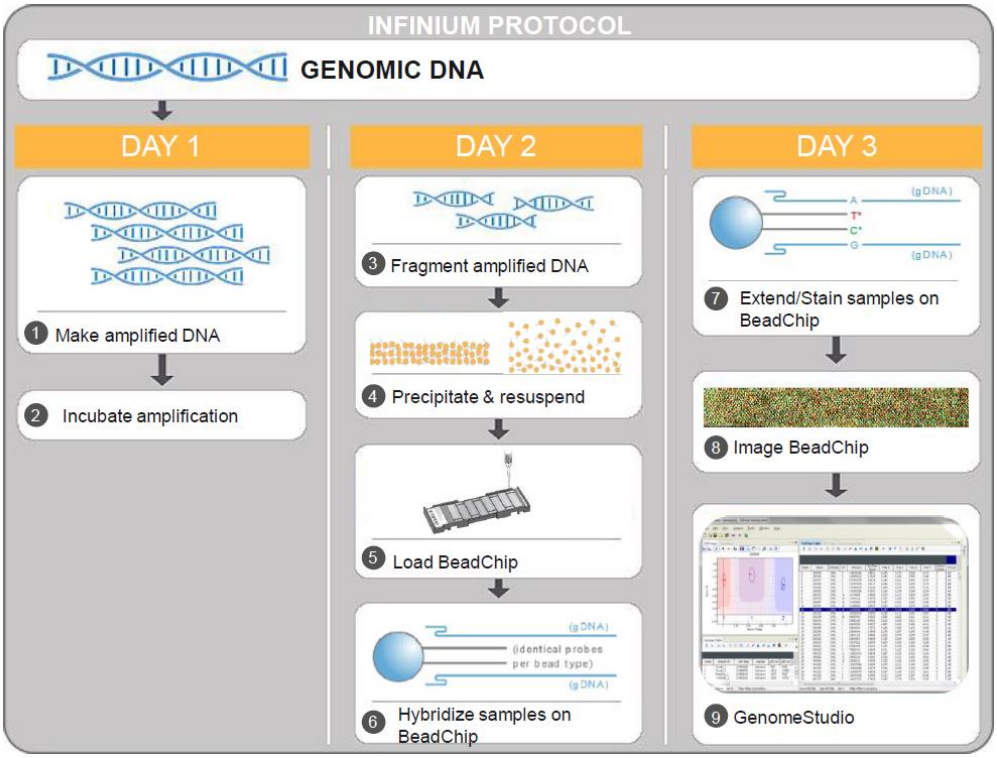

Fig. 3 | *Overview of Infinium workflow*

3.3 GWAS dataset selection and quality control

Future analyses require the selection of GWAS summary statistics which contains information about the Chrysalus SNPs associated with IBDs. To obtain this dataset, some online publicly available databases have been browsed (gwasATLAS[48], GWAS Catalog[29], GWAS Central[49], PLCO Atlas[50] and FINNGEN[51]). The selection of this summary statistics is based on the completeness and quality of information required for PRS computation. In particular SNP IDs, base pairs locations, the chromosomes, effect and non effect alleles, minor allele frequency (MAF), betas, standard errors, p-values and INFOs have been checked and evaluated for SNPs of interest.

Then, various quality control steps have been applied on the selected dataset using awk command, filtering rs for effect allele frequencies (keeping rs with MAF > 0.01), INFO (filtering out SNPs with INFO < 0.8) and eliminating duplicates and ambiguous SNPs.

3.4 Quality control of target data

In this step the command line program PLINK (v. 1.90 beta)[41] has been used in Linux terminal, so first input files (.bim, .bed and .fam) have been created using Plink Input Report Plug-in (v. 2.1.4)[52] for Genome-Studio[47] and the --make-bed command in PLINK[41]. To select just the list of rs associated with CD previously created (*see 3.1 CD-related SNPs and new molecular markers*) the option --extract is used. Then a first quality control step filtered SNPs and individuals using some PLINK[41] commands and setting various thresholds taken from literature (Choi S. W. et al. 2020)[27]. In particular the PLINK[41] functions are:

- --maf 0.01 : to prevent genotyping errors, SNPs with minor allele frequency lower than 0.01 are eliminated from the list;

- --hwe 1e-6 : removed rs with low P-value (less than $1e^{-6}$) from the Hardy-Weinberg Equilibrium Fisher's exact or chi-squared test;

- --geno 0.01 : excluded SNPs that are missing in a high fraction of individuals (more than 1%);

- --mind 0.01 : excluded individuals who have a high rate of genotype missingness (more than 1%) removing samples with low genotype calls.

Then pruning is performed, filtering out any SNPs with Linkage Disequilibrium r^2higher than 0.25 and, after estimating heterozygosity rates represented by F-coefficient (--het command in PLINK[41]), individuals with F-values that are more than 3 standard deviations (SD) from the mean are removed with R (v. 4.2.3)[43]. This step prevents high levels of heterozygosity (indication of low sample quality) and low levels of heterozygosity (may be due to inbreeding).
So a final target dataset has been created which is represented by 3 files with the PLINK[41] output format (.bed, .bim, .fam).

3.5 PRS computation

PRS is computed as a product of ORs (or Beta for continuous traits). To simplify this calculation, the natural logarithm of the Betas is calculated using R[43] so that the PRS can be computed using summation instead (and can be back-transformed afterwards). Then clumping, which preferentially retain only weakly correlated SNPs but preferentially retaining the SNPs most associated with CD, has been performed with the following commands and parameters found in literature (Choi S. W. et al. 2020)[27]:

- --clump-p1 1 : P-value threshold (1 is selected such that all SNPs are include for clumping) for a SNP to be included as an index SNP;

- --clump-r2 0.1 : removing rs having r^2 higher than 0.1 with the index SNPs;

- --clump-kb 250 : SNPs within 250kb of the index SNP are considered.

After the generation of the clumped file, index rs IDs have been extracted and another file (.txt) containing different P-value thresholds for the inclusion of SNPs in the PRS with the following content:

IBD.0.5.profile 0 0.5
IBD.0.4.profile 0 0.4
IBD.0.3.profile 0 0.3
IBD.0.2.profile 0 0.2
IBD.0.1.profile 0 0.1
IBD.0.05.profile 0 0.05
IBD.0.001.profile 0 0.001

Values indicated in this last file are inclusive so for example the 0.1 threshold will include even eventual SNPs with P-value equal to 0.1.

At this point the --score command in PLINK[41] generated seven files, one for each threshold considered, in which there are PRS values for each sample.

4. RESULTS

The initial research step led to identifying 75 variants in Chrysalus associated with Crohn's disease (Table 2).

rs ID	CHR	POSITION	GENE SYMBOL
rs1004819	1	67204530	IL23R
rs10889677	1	67259437	IL23R
rs11209026	1	67240275	IL23R
rs11465804	1	67236843	IL23R
rs17436816	1	6144101	CHD5
rs2201841	1	67228519	IL23R
rs2274910	1	160882256	ITLN1
rs2476601	1	113834946	PTPN22, AP4B1-AS1
rs2641348	1	119895261	ADAM30
rs3024493	1	206770623	IL10
rs3024505	1	206766559	NA
rs4655215	1	19811221	NA
rs6426833	1	19845367	NA
rs6679677	1	113761186	PHTF1
rs7554511	1	200908434	C1orf106
rs76418789	1	67182913	IL23R
rs80174646	1	67242472	IL23R
rs1260326	2	27508073	GCKR
rs148746268	2	25382703	DTNB

rs ID	CHR	POSITION	GENE SYMBOL
rs17229679	2	198696033	LOC105373831
rs2241880	2	233274722	ATG16L1, SCARNA5
rs35667974	2	162268127	IFIH1
rs3749171	2	240630275	GPR35
rs780094	2	27518370	GCKR
rs3197999	3	49684099	MST1
rs9307388	4	113154532	ANK2
rs1000113	5	150860514	IRGM
rs10045431	5	159387525	ENSG00000249738
rs12521868	5	132448701	C5orf56
rs13361189	5	150843825	IRGM
rs1992660	5	40414965	PTGER4
rs2188962	5	132435113	C5orf56
rs4958847	5	150860025	IRGM
rs56167332	5	159400761	NA
rs1049526	6	32981027	BRD2
rs1799964	6	31574531	TNF, LTA, LOC100287329
rs2301436	6	167024500	FGFR1OP
rs28701841	6	106082455	PRDM1
rs6908425	6	20728500	CDKAL1
rs1800795	7	22727026	IL6, LOC541472

rs ID	CHR	POSITION	GENE SYMBOL
rs4728142	7	128933913	NA
rs10758669	9	4981602	near JAK2
rs141992399	9	136365140	DNLZ, CARD9
rs4263839	9	114804160	TNFSF15
rs4986790	9	117713024	TLR4
rs10748781	10	99523573	near LINC01475, NKX2-3
rs2104286	10	6057082	IL2RA
rs224136	10	62710915	NA
rs61839660	10	6052734	IL2RA
rs7915475	10	62621908	ZNF365
rs102275	11	61790331	TMEM258
rs174535	11	61783884	MYRF
rs1793004	11	20677383	NELL1
rs630923	11	118883644	CXCR5
rs694739	11	64329761	LOC102723878
rs7927894	11	76590272	NA
rs3184504	12	111446804	SH2B3
rs3764147	13	43883789	LACC1
rs35874463	15	67165360	SMAD3
rs104895438	16	50711745	NOD2
rs2066844	16	50712015	NOD2

rs ID	CHR	POSITION	GENE SYMBOL
rs2066845	16	50722629	NOD2
rs2066847	16	50729868	NOD2
rs2076756	16	50722970	NOD2
rs5743293	16	50729868	NOD2, CYLD-AS1
rs72796367	16	50728860	NOD2
rs9889296	17	34243528	NA
rs35018800	19	10354167	TYK2
rs602662	19	48703728	FUT2, LOC105447645
rs6017342	20	44436388	NA
rs6062496	20	63697746	TNFRSF6B, RTEL1-TNFRSF6B
rs2836754	21	38919816	LOC400867
rs762421	21	44195678	NA
rs1569414	22	45331684	FAM118A
rs2143178	22	39264824	NA

Table 2 | *Variants of Chrysalus associated to Crohn's disease (NA means that rs is located in the intergenic region in NCBI[44]);* SNPs selected from: Sazonovs A. et al., *Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility*[26], 2022; Younis N. et al., *Inflammatory bowel disease: between genetics and microbiota*[9], 2020; https://genportal.tellmegen.eu/results/diseases/60/0[53].

Moreover 33 “new variants” not currently in the chip have been identified and listed below (Table 3).

rs ID	Chromosome	Position	Gene
rs10489629	1	67222666	IL23R
rs11209032	1	67274409	IL23R
rs12037606	1	172929262	NA
rs7517847	1	67215986	IL23R
rs1343151	1	67253446	IL23R
rs1495965	1	67287825	NA
rs1800872	1	206773062	IL19, IL10
rs3024496	1	206768519	IL10
rs10210302	2	233250193	ATG16L1
rs77981966	2	43550825	THADA
rs9858542	3	49664550	BSN
rs143431075	4	172130730	GALNTL6
rs1992662	5	40393750	PTGER4
rs2631367	5	132369766	SLC22A5
rs6596075	5	132406536	IBD5
rs17234657	5	40401407	NA
rs1050152	5	132340627	SLC22A4
rs9469220	6	32690533	NA
rs7753394	6	137764111	NA
rs7807268	7	148560956	IL23R
rs6601764	10	3820350	NA
rs10761659	10	62685804	NA
rs3135932	11	117993348	IL10RA
rs33995883	12	40346884	LRRK2
rs2201840	15	24899980	SNRPN, SNHG14
rs17221417	16	50705671	NOD2
rs41450053	16	50722629	NOD2
rs1728785	16	68557327	ZFP90
rs10431923	16	68805360	CDH1

rs ID	Chromosome	Position	Gene
rs2542151	18	12779948	PTPN2
rs7234029	18	12877061	PTPN2
rs8111071	19	45804148	RSPH6A
rs2834167	21	33268483	IL10RB

Table 3 | *List of variants associated to Crohn's disease that potentially could be updated in Chrysalus (NA means that the respective gene is not reported in NCBI[44])*

Genotyping 792 european healthy individuals with Chrysalus and the Infinium protocol, relative probe efficiency and call rate have been tested and listed for each SNP (indicated with the name assigned by Genome-Studio[47]) related to Crohn's disease (Table 4). In this result, duplicated SNPs are still conserved because quality control steps of target data will remove them in a next phase.

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
1	rs1004819	782	0.8837
1	rs10889677	784	0.8815
1	rs11209026	758	0.8352
1	rs11209026_Ilmn Dup1	743	0.8112
1	rs11209026_Ilmn Dup2	752	0.8283
1	rs11465804	784	0.9296
1	rs11465804_ilmndup1	781	0.9203
1	rs11465804_ilmndup2	783	0.9234
1	rs17436816	747	0.7111

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
1	rs2201841.2_F2BT	772	0.8442
1	rs2274910	782	0.9344
1	rs2476601	784	0.9062
1	rs2641348	783	0.8924
1	rs3024493.1_F2BT	771	0.9456
1	rs3024493.2_F2BT	784	0.7898
1	rs3024505	786	0.8625
1	rs4655215	784	0.9379
1	rs4655215_ilmndup1	784	0.9401
1	rs4655215_ilmndup2	785	0.9415
1	rs6426833	772	0.8526
1	rs6679677	770	0.8998
1	rs7554511	784	0.9323
1	rs76418789	771	0.9096
1	rs76418789_ilmndup1	781	0.9176
1	rs76418789_ilmndup2	778	0.9176
1	rs80174646	781	0.9155
2	rs1260326.1_F2BT	780	0.6090

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
2	rs148746268	745	0.7027
2	rs17229679	776	0.8500
2	rs17229679_ilmndup1	776	0.8457
2	rs17229679_ilmndup2	752	0.8191
2	rs2241880	783	0.8587
2	rs35667974	786	0.8698
2	rs35667974_ilmndup1	787	0.8838
2	rs35667974_ilmndup2	788	0.8655
2	rs3749171	735	0.6868
2	rs780094	783	0.8200
3	rs3197999	779	0.7421
4	rs9307388	772	0.8829
5	rs1000113	780	0.9062
5	rs1000113_ilmndup1	780	0.8997
5	rs1000113_ilmndup2	777	0.8997
5	rs10045431	781	0.8504
5	rs12521868	787	0.8450
5	rs13361189	781	0.7882

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
5	rs1992660	773	0.8731
5	rs2188962_ilmndup1	748	0.7612
5	rs2188962_ilmndup2	727	0.6227
5	rs2188962	723	0.6250
5	rs4958847	777	0.8311
5	rs56167332.2_F2BT	771	0.8791
5	rs56167332.2_F2BT_ilmndup1	770	0.9753
5	rs56167332.2_F2BT_ilmndup2	775	0.9234
6	rs1049526	785	0.7988
6	rs1799964	763	0.8027
6	rs2301436	782	0.7856
6	rs28701841	775	0.8686
6	rs28701841_ilmndup1	773	0.8667
6	rs6908425	777	0.8951
7	rs1800795	785	0.8437
7	rs1800795_IlmnDup1	785	0.8453
7	rs1800795_IlmnDup2	784	0.8419
7	rs4728142	771	0.7975

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
7	rs4728142_ilmndup1	765	0.7640
7	rs4728142_ilmndup2	778	0.8104
9	rs10758669.2_F2BT	775	0.9044
9	rs141992399.1_F2BT	774	0.9023
9	rs141992399.1_F2BT_ilmndup1	767	0.9176
9	rs141992399.2_F2BT	761	0.8430
9	rs141992399.2_F2BT_ilmndup1	766	0.8492
9	rs141992399.2_F2BT_ilmndup2	766	0.8451
9	rs4263839	778	0.7796
9	rs4263839_ilmndup1	780	0.7842
9	rs4263839_ilmndup2	780	0.7731
9	rs4986790.1_F2BT	774	0.8212
9	rs4986790.2_F2BT	772	0.8785
10	rs10748781.1_F2BT	774	0.6371
10	rs10748781.1_F2BT_ilmndup1	773	0.6406
10	rs10748781.1_F2BT_ilmndup2	779	0.5818
10	rs10748781.2_F2BT	781	0.6573
10	rs10748781.2_F2BT_ilmndup1	780	0.6552

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
10	rs10748781.2_F2BT_ilmndup2	781	0.6673
10	rs2104286	781	0.8912
10	rs224136.1_F2BT	768	0.9070
10	rs224136.2_F2BT	757	0.8618
10	rs61839660	778	0.8162
10	rs61839660_ilmndup1	771	0.8052
10	rs61839660_ilmndup2	777	0.8269
10	rs7915475	786	0.9100
10	rs7915475_ilmndup1	783	0.9147
10	rs7915475_ilmndup2	782	0.9080
11	rs102275	773	0.7897
11	rs174535.1_F2BT	767	0.9158
11	rs174535.2_F2BT	770	0.8145
11	rs1793004.3_F2BT	762	0.8497
11	rs630923	777	0.9116
11	rs630923_ilmndup1	780	0.9144
11	rs630923_ilmndup2	778	0.9256
11	rs694739	779	0.7954

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
11	rs7927894	777	0.9285
12	rs3184504.2_F2BT	785	0.9323
13	rs3764147	784	0.8828
15	rs35874463	785	0.8202
15	rs35874463_ilmndup1	784	0.8207
15	rs35874463_ilmndup2	786	0.8186
16	rs104895438.1_F2BT	751	0.8925
16	rs104895438.1_F2BT_ilmndup1	749	0.8750
16	rs104895438.1_F2BT_ilmndup2	764	0.8698
16	rs104895438.2_F2BT	772	0.9096
16	rs104895438.2_F2BT_ilmndup1	774	0.9062
16	rs104895438.2_F2BT_ilmndup2	777	0.9035
16	rs2066844	742	0.8562
16	rs2066844_ilmndup1	717	0.5713
16	rs2066845.1_F2BT	774	0.9422
16	rs2066845.1_F2BT_ilmndup1	770	0.9296
16	rs2066845.1_F2BT_ilmndup2	759	0.9179
16	rs2066845.2_F2BT	778	0.9368

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
16	rs2066845.2_F2BT_ilmndup1	785	0.9425
16	rs2066845.2_F2BT_ilmndup2	787	0.9188
16	rs2066847	779	0.8890
16	rs2076756	764	0.8273
16	rs5743293	788	0.8904
16	rs5743293_ilmndup1	788	0.8779
16	rs5743293_ilmndup2	785	0.8917
16	rs72796367	777	0.8834
16	rs72796367_ilmndup1	778	0.8090
16	rs72796367_ilmndup2	777	0.8745
17	rs9889296	785	0.9384
17	rs9889296_ilmndup1	786	0.9321
17	rs9889296_ilmndup2	746	0.7407
19	rs35018800	783	0.7796
19	rs35018800_ilmndup1	783	0.7796
19	rs35018800_ilmndup2	784	0.7864
19	rs602662	765	0.7675
20	rs6017342	777	0.9270

Chromosome	rs IDs	Calling Rate	Signal quality (GenTrain Score)
20	rs6017342_ilmndup1	782	0.9361
20	rs6017342_ilmndup2	779	0.9359
20	rs6062496	779	0.9026
20	rs6062496_ilmndup1	778	0.8936
20	rs6062496_ilmndup2	778	0.8968
21	rs2836754	770	0.8259
21	rs762421	780	0.8139
22	rs1569414	778	0.7933
22	rs2143178	762	0.7922
22	rs2143178_ilmndup1	773	0.9083
22	rs2143178_ilmndup2	769	0.8644

Table 4 | *Calling rate per rs (on the 792 total samples) and signal quality (reported as GenTrain Score from Genome-Studio[47]) for each IBD variant and eventual duplicates present in Chrysalus*

Various public summary statistics from different Genome-Wide Association Studies have been checked to select the one that contains as many Chrysalus SNPs associated with IBDs as possible. The dataset that has been selected is from Zorina-Lichtenwalter's GWAS (Zorina-Lichtenwalter K. et al. 2023)[25] in which is estimated the genetic risk shared across 24 different chronic pain conditions, including Crohn's disease.

This dataset contains 11 323 612 SNPs associated with the traits and for each rs is reported the variant ID, chromosome, base pair location, effect and non effect allele, minor allele frequency, beta, standard error, P-value and the value of INFO.

chromosome	base_pair_location	variant_id	other_allele	effect_allele	effect_allele_frequency	beta	standard_error	p_value	INFO
10	88766	rs55896525	T	C	0.945194	0.077287	0.0763821	0.311611157816631	0.916423	
10	90127	rs185642176	T	C	0.913977	-0.105855	0.059856	0.0769785985160769	0.984729	
10	90164	rs141504207	G	C	0.916561	-0.129943	0.0626302	0.038008436062007	0.923494	
10	94263	rs184120752	A	C	0.974493	0.0394445	0.107392	0.713398184854526	0.959293	
10	94426	rs10904045	T	C	0.603606	0.0265233	0.034132	0.437111265964983	0.99154	
10	94541	rs11251906	A	C	0.969679	-0.0694109	0.0978623	0.47815642964906	0.975825	
10	95074	rs6560828	A	G	0.489545	-0.00397396	0.0336246	0.905920109126875	0.977746	
10	95662	rs35849539	C	G	0.681702	-0.00794125	0.0357843	0.824376114448353	0.991624	
10	95844	rs117205301	T	C	0.961768	-0.129116	0.0866242	0.136086172665433	1	

Fig. 4 | *First ten lines of summary statistics (from Zorina-Lichtenwalter K. et al. 2023)*[25]

To create a base dataset for PRS computation, rs have been filtered as discussed in 3.3 *GWAS dataset selection and quality control*. At the end of this step there are 9 582 415 SNPs in the final base data file. In the dataset there's no duplicate rs or SNPs with MAF or INFO values below the imposed thresholds so only ambiguous SNPs have been removed.

Then quality control steps on target data have been applied and with proper PLINK[41] commands (–mind, –geno, –hwe and –maf) 113 variants and 616 individuals passed those filters. Moreover, 58 highly correlated SNPs have been removed with pruning. So the final target dataset, which is grouped in 3 PLINK[41] input files (.fam: the sample information file; .bim: extended variant information file; .bed: binary biallelic genotype table), contained 616 individuals and 55 variants.

Then the polygenic risk score is finally computed and seven different files have been generated, one for each threshold of P-value previously mentioned (*see 3.5 PRS computation*).

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	36	23	0.0275043
2	207221690004_R01C02	-9	36	24	0.0289976
3	207221690004_R02C01	-9	36	21	0.0236661
4	207221690004_R02C02	-9	36	19	0.0226682
5	207221690004_R03C01	-9	36	28	0.0329446
7	207221690004_R04C01	-9	36	24	0.0287604
8	207221690004_R04C02	-9	36	26	0.0310827
9	207221690004_R05C01	-9	36	19	0.0234582
10	207221690004_R05C02	-9	36	23	0.0281744
11	207221690004_R06C01	-9	36	21	0.0248191
13	207221690004_R07C01	-9	36	21	0.025119
14	207221690004_R07C02	-9	36	23	0.0273832
15	207221690004_R08C01	-9	36	23	0.0281426
16	207221690004_R08C02	-9	36	23	0.0267358
17	207221690004_R09C01	-9	36	26	0.0311949
18	207221690004_R09C02	-9	36	24	0.0285407
19	207221690004_R10C01	-9	36	26	0.0306987
20	207221690004_R10C02	-9	36	19	0.022724
21	207221690004_R11C01	-9	36	26	0.0301679
22	207221690004_R11C02	-9	36	20	0.0235972
23	207221690004_R12C01	-9	36	26	0.0304581
24	207221690004_R12C02	-9	36	25	0.0298206
26	207149100001_R01C02	-9	36	24	0.0280684
27	207149100001_R02C01	-9	36	22	0.0264507
28	207149100001_R02C02	-9	36	26	0.0305907
29	207149100001_R03C01	-9	36	21	0.0257074
30	207149100001_R03C02	-9	36	21	0.0258293
31	207149100001_R04C01	-9	36	23	0.0276747
33	207149100001_R05C01	-9	36	25	0.0294535
34	207149100001_R05C02	-9	36	26	0.0305375
35	207149100001_R06C01	-9	36	26	0.0311946
36	207149100001_R06C02	-9	36	26	0.0306192
37	207149100001_R07C01	-9	36	22	0.0259372
38	207149100001_R07C02	-9	36	26	0.0309126
39	207149100001_R08C01	-9	36	28	0.0329903
40	207149100001_R08C02	-9	36	23	0.0268551
41	207149100001_R09C01	-9	36	21	0.0253698
42	207149100001_R09C02	-9	36	25	0.0297925
43	207149100001_R10C01	-9	36	26	0.0309056
44	207149100001_R10C02	-9	36	26	0.0305488
45	207149100001_R11C01	-9	36	24	0.0278445

Fig. 5 | *Output of the PRS computation on the first 45 individuals that passed QC steps of target data for the threshold of P-value from 0 to 0.5. FID: family ID, IID: sample ID, PHENO: phenotype value (-9 means that in this case it is unknown), CNT: number of non-missing alleles used for scoring (represents twice the number of variants included in the considered threshold), CNT2: sum of named allele counts (each ScoreFile line names a variant, and only one of its two alleles), SCORE: PRS value.*

5. DISCUSSION AND CONCLUSIONS

The results that this first polygenic risk score calculation test for Crohn's disease gave are similar for all SNPs P-value thresholds considered. In particular, as we can see in part of the results presented below (*Fig. 6-12*), SCORE value is included between 0.01 and 0.04.

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	36	23	0.0275043
2	207221690004_R01C02	-9	36	24	0.0289976
3	207221690004_R02C01	-9	36	21	0.0236661
4	207221690004_R02C02	-9	36	19	0.0226682
5	207221690004_R03C01	-9	36	28	0.0329446
7	207221690004_R04C01	-9	36	24	0.0287604
8	207221690004_R04C02	-9	36	26	0.0310827
9	207221690004_R05C01	-9	36	19	0.0234582
10	207221690004_R05C02	-9	36	23	0.0281744
11	207221690004_R06C01	-9	36	21	0.0248191

Fig. 6 | *PRS calculated with SNPs having P-value between 0 and 0.5 (first ten samples)*

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	34	22	0.0281408
2	207221690004_R01C02	-9	34	23	0.0297219
3	207221690004_R02C01	-9	34	21	0.0250582
4	207221690004_R02C02	-9	34	18	0.0230203
5	207221690004_R03C01	-9	34	27	0.0339012
7	207221690004_R04C01	-9	34	24	0.0304522
8	207221690004_R04C02	-9	34	26	0.0329111
9	207221690004_R05C01	-9	34	18	0.0238568
10	207221690004_R05C02	-9	34	23	0.0298317
11	207221690004_R06C01	-9	34	20	0.0252977

Fig. 7 | *PRS calculated with SNPs having P-value between 0 and 0.4 (first ten samples)*

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	30	18	0.0247474
2	207221690004_R01C02	-9	30	19	0.0265394
3	207221690004_R02C01	-9	30	20	0.0266249
4	207221690004_R02C02	-9	30	15	0.0207426
5	207221690004_R03C01	-9	30	23	0.0312758
7	207221690004_R04C01	-9	30	21	0.0291413
8	207221690004_R04C02	-9	30	22	0.0301537
9	207221690004_R05C01	-9	30	14	0.0198921
10	207221690004_R05C02	-9	30	19	0.0266637
11	207221690004_R06C01	-9	30	16	0.0215252

Fig. 8 | *PRS calculated with SNPs having P-value between 0 and 0.3 (first ten samples)*

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	28	18	0.0265151
2	207221690004_R01C02	-9	28	19	0.028435
3	207221690004_R02C01	-9	28	18	0.0261444
4	207221690004_R02C02	-9	28	13	0.0198419
5	207221690004_R03C01	-9	28	22	0.0323186
7	207221690004_R04C01	-9	28	19	0.0288405
8	207221690004_R04C02	-9	28	21	0.0311164
9	207221690004_R05C01	-9	28	14	0.021313
10	207221690004_R05C02	-9	28	19	0.0285683
11	207221690004_R06C01	-9	28	16	0.0230628

Fig. 9 | *PRS calculated with SNPs having P-value between 0 and 0.2 (first ten samples)*

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	26	16	0.0242464
2	207221690004_R01C02	-9	26	17	0.026314
3	207221690004_R02C01	-9	26	17	0.0260013
4	207221690004_R02C02	-9	26	12	0.019214
5	207221690004_R03C01	-9	26	20	0.0304964
7	207221690004_R04C01	-9	26	17	0.0267506
8	207221690004_R04C02	-9	26	19	0.0292016
9	207221690004_R05C01	-9	26	12	0.0186442
10	207221690004_R05C02	-9	26	17	0.0264575
11	207221690004_R06C01	-9	26	15	0.0226827

Fig. 10 | *PRS calculated with SNPs having P-value between 0 and 0.1 (first ten samples)*

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	24	15	0.0248794
2	207221690004_R01C02	-9	24	16	0.0271193
3	207221690004_R02C01	-9	24	17	0.0281681
4	207221690004_R02C02	-9	24	12	0.0208152
5	207221690004_R03C01	-9	24	19	0.0316502
7	207221690004_R04C01	-9	24	17	0.0289799
8	207221690004_R04C02	-9	24	19	0.0316351
9	207221690004_R05C01	-9	24	12	0.0201978
10	207221690004_R05C02	-9	24	16	0.0272748
11	207221690004_R06C01	-9	24	14	0.0231854

Fig. 11 | *PRS calculated with SNPs having P-value between 0 and 0.05 (first ten samples)*

FID	IID	PHENO	CNT	CNT2	SCORE
1	207221690004_R01C01	-9	10	6	0.0222966
2	207221690004_R01C02	-9	10	6	0.0224011
3	207221690004_R02C01	-9	10	7	0.0262628
4	207221690004_R02C02	-9	10	4	0.0140897
5	207221690004_R03C01	-9	10	8	0.0295852
7	207221690004_R04C01	-9	10	6	0.0229404
8	207221690004_R04C02	-9	10	6	0.0225601
9	207221690004_R05C01	-9	10	6	0.0227827
10	207221690004_R05C02	-9	10	6	0.0227827
11	207221690004_R06C01	-9	10	7	0.0262628

Fig. 12 | *PRS calculated with SNPs having P-value between 0 and 0.001 (first ten samples)*

This primary result represents the relative risk to develop Crohn's disease for each genotyped individual. The analyzed population is probably made up of healthy individuals even if we don't have any information on the phenotype. Since every trait/disease has its own PRS range of value, we can't yet give an interpretation of the obtained PRS values. In other words, with this first genetic and statistical test we can't say when a sample is considered at high or low risk for CD development and we still can not interpret the obtained values. In addition, since IBDs are multifactorial diseases, genetics has a relative influence on the final phenotype of each person. For this reason, to investigate the role of genetics in IBDs, a PRS evaluation on CD patients must be done and compared to the PRS of healthy people. This would highlight the influence of the genetic factor on every individual and in general for the development of the considered disease. In

addition, this analysis would give more information on the relative weight of each considered SNP for the onset of IBDs.

The presented workflow can also be optimized and improved in the future to obtain even more reliable scores which surely can't be used for CD diagnosis, but that could help to prevent disease by individuating the most susceptible individuals on the genetic level. For example, this method can gain more statistical value with the introduction of more quality control steps on the target dataset, such as the sex-check, the eliminations of too closely related individuals and with the strand flipping that would help to keep even more SNPs associated to CD that PLINK[41] would automatically remove (mismatching SNPs). Another important gain of statistical power could be carried by the selection of an even more wide and accurate summary statistics file from GWAS to be used as a reference dataset for the PRS computation. The same GWAS is important for both defining the sample size and the population that should be genetically similar to the population genotyped in our laboratory. Moreover, the addition of even more CD-associated SNPs in Chrysalus chip will help to improve and refine the polygenic risk score computation. Another important future aim for BMR Genomics[38] regards the creation of a first functional report based on the previously identified SNPs that describe the genetic profile for each individual and the relative risk for the development of Crohn's disease. For this purpose the implementation of a more accurate PRS will play an important role to develop a product that will investigate the genetics of each individual and explain it in "user friendly" terms.

Moreover, the exploration of other PRS dedicated programs such as PRSice-2[54] and LDpred-2 (implemented in R package bigsnpr)[55] will help to compare various ways to obtain the score and to choose the best fitting one for our analyses.

6. SUPPLEMENTARY MATERIAL

6.1 Full list of commands

BASE DATA QC

```
#check integrity of the file
gunzip -c GCST90129433_buildGRCh37.tsv.gz | head

#total SNPs listed in summary statistics
gunzip -c GCST90129433_buildGRCh37.tsv.gz | nl | tail -10

#filtering SNPs: MAF and INFO
gunzip -c GCST90129433_buildGRCh37.tsv.gz | awk 'NR==1 || ($6 > 0.01) && ($10 > 0.8) {print}' |
gzip > MAF.INFO.gz

#checking the number of SNPs
gunzip -c MAF.INFO.gz | nl | tail -10

#filtering repeated SNPs
gunzip -c MAF.INFO.gz | awk '{seen[$3]++; if(seen[$3]==1){print}}' | gzip - > nodup.gz

#checking the number of SNPs
gunzip -c nodup.gz | nl | tail -10

#filtering ambiguous SNPs
gunzip -c nodup.gz | awk '!( ($4=="A" && $5=="T") || ($4=="T" && $5=="A") || ($4=="G" &&
$5=="C") || ($4=="C" && $5=="G")) {print}' | gzip > basedata.QC.gz

#checking the number of SNPs
gunzip -c basedata.QC.gz | nl | tail -10
```

TARGET DATA QC

#needed files: output files from Genome-Studio[47] (.ped and .map format)

#generation of PLINK input files

./plink --ped 740.ped --map 740.map --extract snplistdaestrarreprova.txt --make-bed --out target.IBD

#standard QC

./plink --bfile target.IBD --maf 0.01 --hwe 1e-6 --geno 0.01 --mind 0.01 --write-snplist --make-just-fam --out target.QC

#pruning

./plink --bfile target.IBD --keep target.QC.fam --extract target.QC.snplist --indep-pairwise 200 50 0.25 --out target.pruned.QC

#estimation of heterozygosity rates (F coefficient)

./plink --bfile target.IBD --extract target.pruned.QC.prune.in --keep target.QC.fam --het --out target.het.QC

#filtering SNPs with heterozygosity rates with more than 3 DS from the mean

##starting R

R

##read in the EUR.het file, specify it has header

dat <- read.table("target.het.QC.het", header=T)

##calculate the mean

m <- mean(dat$F)

```
    ##calculate the SD
s <- sd(dat$F)
    ##get any samples with F coefficient within 3 SD of the population mean
valid <- subset(dat, F <= m+3*s & F >= m-3*s)
    ##print FID and IID for valid samples
write.table(valid[,c(1,2)], "target.valid.sample", quote=F, row.names=F)
    ##exit R
q()


#generating a final dataset
./plink --bfile target.IBD --make-bed --keep target.valid.sample --exclude target.pruned.QC.prune.out  --extract target.QC.snplist --out targetdataset.QC


POLYGENIC RISK SCORE
#needed files:
##targetdataset.QC.bed
##targetdataset.QC.bim
##targetdataset.QC.fam
##QCedbasedata.gz


#updating effect size (OR)
    ##starting R
R
    ##reading the QCedbasedata file
dat <- read.table(gzfile("QCedbasedata.gz"), header=T)
    ##transforming betas
dat$BETA <- log(dat$beta)
    ##printing the table
write.table(dat, "QCedbasedata.Transformed", quote=F, row.names=F)
```

##exit R

q()

#clumping

./plink --bfile targetdataset.QC --clump-p1 1 --clump-r2 0.1 --clump-kb 250 --clump QCedbasedata.Transformed --clump-snp-field variant_id --clump-field p_value --out EUR

#estaction of index SNP IDs:

awk 'NR!=1{print $3}' EUR.clumped > EUR.valid.snp

#for PRS computation: we need three files:

##The base data file: QCedbasedata.Transformed

##A file containing SNP IDs and their corresponding P-values

awk '{print $3,$9}' QCedbasedata.Transformed > SNP.pvalue

##A file (rangelist.txt) containing the different P-value thresholds for inclusion of SNPs in the PRS with the following format:

IBD.0.5.profile 0 0.5

IBD.0.4.profile 0 0.4

IBD.0.3.profile 0 0.3

IBD.0.2.profile 0 0.2

IBD.0.1.profile 0 0.1

IBD.0.05.profile 0 0.05

IBD.0.001.profile 0 0.001

```
#PRS computation
./plink \
  --bfile targetdataset.QC \
  --score QCedbasedata.Transformed 3 5 8 header \
  --q-score-range rangelist.txt SNP.pvalue \
  --extract EUR.valid.snp \
  --out PRS
```

6.2 Extraction protocol

The full extraction protocol of the Mag-Bind Blood & Tissue DNA HDQ 96 Kit followed to extract DNA from buccal swabs can be downloaded at: https://ensur.omegabio.com/ensur/contentAction.aspx?key=Production.3572.S2R4E1A3.20190207.67.4686935

6.3 Infinium HD Assay Ultra Protocol

The full Infinium HD Assay Ultra Protocol can be found at:
https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium-hd-ultra/11328087_RevB_Infinium_HD_Ultra_Assay_Guide_press.pdf

7. REFERENCES

1. Rosen MJ, Dhawan A, Saeed SA. Inflammatory Bowel Disease in Children and Adolescents. *JAMA Pediatrics*. 2015;169(11):1053. doi:https://doi.org/10.1001/jamapediatrics.2015.1982
2. Podolsky DK. Inflammatory Bowel Disease. *New England Journal of Medicine*. 2002;347(6):417-429. doi:https://doi.org/10.1056/nejmra020831
3. Marteau P, Lepage P, Mangin I, et al. Gut flora and inflammatory bowel disease. *Alimentary Pharmacology & Therapeutics*. 2004;20:18-23. doi:https://doi.org/10.1111/j.1365-2036.2004.02062.x
4. Prideaux L, Kang S, Wagner J, et al. Impact of Ethnicity, Geography, and Disease on the Microbiota in Health and Inflammatory Bowel Disease. *Inflammatory Bowel Diseases*. 2013;19(13):2906-2918. doi:https://doi.org/10.1097/01.mib.0000435759.05577.12
5. Thornton JR, Emmett PM, Heaton KW. Diet and Crohn's disease: characteristics of the pre-illness diet. *BMJ*. 1979;2(6193):762-764. doi:https://doi.org/10.1136/bmj.2.6193.762
6. Roncoroni L, Gori R, Elli L, et al. Nutrition in Patients with Inflammatory Bowel Diseases: A Narrative Review. *Nutrients*. 2022;14(4):751. doi:https://doi.org/10.3390/nu14040751
7. Pedersen N, et al. "Low-FODMAP Diet Reduces Irritable Bowel Symptoms in Patients with Inflammatory Bowel Disease." *World Journal of Gastroenterology*, 14 May 2017, pubmed.ncbi.nlm.nih.gov/28566897/.
8. Corrao G, Tragnone A, Caprilli R, et al. Risk of inflammatory bowel disease attributable to smoking, oral contraception and breastfeeding in Italy: a nationwide case-control study. Cooperative Investigators of the Italian Group for the Study of the Colon and the Rectum (GISC). *International Journal of Epidemiology*. 1998;27(3):397-404. doi:https://doi.org/10.1093/ije/27.3.397
9. Younis N, Zarif R, Mahfouz R. Inflammatory bowel disease: between genetics and microbiota. *Molecular Biology Reports*. Published online February 21, 2020:1-11. doi:https://doi.org/10.1007/s11033-020-05318-5
10. Lauro ML, Burch JM, Grimes CL. The effect of NOD2 on the microbiota in Crohn's disease. *Current Opinion in Biotechnology*. 2016;40:97-102. doi:https://doi.org/10.1016/j.copbio.2016.02.028

11. Hamaoui D, Subtil A. ATG16L1 functions in cell homeostasis beyond autophagy. *The FEBS Journal*. 2021;289(7):1779-1800. doi:https://doi.org/10.1111/febs.15833
12. Steen EH, Wang X, Balaji S, Butte MJ, Bollyky PL, Keswani SG. The Role of the Anti-Inflammatory Cytokine Interleukin-10 in Tissue Fibrosis. *Advances in Wound Care*. 2020;9(4):184-198. doi:https://doi.org/10.1089/wound.2019.1032
13. Al-Abbasi FA, Mohammed K, Sadath S, Banaganapalli B, Nasser K, Shaik NA. Computational Protein Phenotype Characterization of IL10RA Mutations Causative to Early Onset Inflammatory Bowel Disease (IBD). *Frontiers in Genetics*. 2018;9:146. doi:https://doi.org/10.3389/fgene.2018.00146
14. Ahn D, Prince A. Participation of the IL-10RB Related Cytokines, IL-22 and IFN-λ in Defense of the Airway Mucosal Barrier. *Frontiers in Cellular and Infection Microbiology*. 2020;10. doi:https://doi.org/10.3389/fcimb.2020.00300
15. Nath P, Jena KK, Mehto S, et al. IRGM links autoimmunity to autophagy. *Autophagy*. 2020;17(2):578-580. doi:https://doi.org/10.1080/15548627.2020.1810920
16. Zhang X, Kortholt A. LRRK2 Structure-Based Activation Mechanism and Pathogenesis. *Biomolecules*. 2023;13(4):612. doi:https://doi.org/10.3390/biom13040612
17. Spalinger MR, Sayoc-Becerra A, Santos AN, et al. PTPN2 Regulates Interactions Between Macrophages and Intestinal Epithelial Cells to Promote Intestinal Barrier Function. *Gastroenterology*. 2020;159(5):1763-1777.e14. doi:https://doi.org/10.1053/j.gastro.2020.07.004
18. Subhadarshani S, Yusuf N, Elmets CA. IL-23 and the Tumor Microenvironment. *Advances in Experimental Medicine and Biology*. Published online January 1, 2021:89-98. doi:https://doi.org/10.1007/978-3-030-55617-4_6
19. Hansford S, Kaurah P, Li-Chang H, et al. Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncology*. 2015;1(1):23-32. doi:https://doi.org/10.1001/jamaoncol.2014.168
20. Yu Y, Zhang Q, Wu N, et al. HNF4α overexpression enhances the therapeutic potential of umbilical cord mesenchymal stem/stromal cells in mice with acute liver failure. *FEBS letters*. 2022;596(24):3176-3190. doi:https://doi.org/10.1002/1873-3468.14453
21. National Human Genome Research Institute. Genome-Wide association studies (GWAS). Genome.gov. Published 2019. https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies

22. Luo Y, de Lange KM, Jostins L, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nature Genetics*. 2017;49(2):186-192. doi:https://doi.org/10.1038/ng.3761
23. Moschen AR, Tilg H, Raine T. IL-12, IL-23 and IL-17 in IBD: immunobiology and therapeutic targeting. *Nature Reviews Gastroenterology & Hepatology*. 2018;16(3):185-196. doi:https://doi.org/10.1038/s41575-018-0084-8
24. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nature Reviews Methods Primers*. 2021;1(1). doi:https://doi.org/10.1038/s43586-021-00056-9
25. Zorina-Lichtenwalter K, Bango CI, Van Oudenhove L, et al. Genetic risk shared across 24 chronic pain conditions: identification and characterization with genomic structural equation modeling. *Pain*. Published online May 23, 2023. doi:https://doi.org/10.1097/j.pain.0000000000002922
26. Sazonovs A, Stevens CR, Venkataraman GR, et al. Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. *Nature Genetics*. 2022;54(9):1275-1283. doi:https://doi.org/10.1038/s41588-022-01156-2
27. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols*. 2020;15(9):2759-2772. doi:https://doi.org/10.1038/s41596-020-0353-1
28. Marees, Andries T., et al. "A Tutorial on Conducting Genome-Wide Association Studies: Quality Control and Statistical Analysis." *International Journal of Methods in Psychiatric Research*, vol. 27, no. 2, 27 Feb. 2018, p. e1608, onlinelibrary.wiley.com/doi/full/10.1002/mpr.1608, https://doi.org/10.1002/mpr.1608.

8. SITOGRAPHY

29. https://www.ebi.ac.uk/gwas/
30. https://zzz.bwh.harvard.edu/gpc/
31. https://www.ukbiobank.ac.uk/
32. https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
33. http://csg.sph.umich.edu/abecasis/MaCH/
34. https://faculty.washington.edu/browning/beagle/b4_1.html
35. https://genome.sph.umich.edu/wiki/Minimac4
36. https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
37. https://topmed.nhlbi.nih.gov/
38. https://www.bmr-genomics.it/
39. https://www.illumina.com/
40. https://www.illumina.com/products/by-type/clinical-research-products/infinium-global-diversity-array-cytogenetics-8.html
41. https://www.cog-genomics.org/plink2/
42. https://genome.ucsc.edu/cgi-bin/hgLiftOver
43. https://www.r-project.org/
44. https://www.ncbi.nlm.nih.gov/
45. https://www.snpedia.com/index.php/SNPedia
46. https://www.illumina.com/systems/array-scanners/iscan.html
47. https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html
48. https://atlas.ctglab.nl/
49. https://www.gwascentral.org/
50. https://dceg.cancer.gov/tools/public-data/plco-atlas
51. https://www.finngen.fi/fi
52. https://support.illumina.com/downloads/genomestudio-2-0-plug-ins.html
53. https://genportal.tellmegen.eu/results/diseases/60/0
54. https://choishingwan.github.io/PRSice/
55. https://cran.r-project.org/web/packages/bigsnpr/index.html

9. ACKNOWLEDGEMENTS

Vorrei esprimere la mia immensa gratitudine al Professor Giorgio Valle che mi ha concesso l'opportunità di lavorare a questo progetto e di acquisire esperienza in campo biotecnologico e bioinformatico.
Inoltre, vorrei ringraziare la Dott.ssa Barbara Arredi e tutto lo "SNP Team" (Dott.ssa Stefania Vendramin, Dott.ssa Silvia Pescarolo) per il grande aiuto dato per lo sviluppo di questa tesi.
Grazie ai dipendenti di BMR Genomics per avermi accolto e aiutato in questa bellissima esperienza da cui ho imparato molto.
Grazie al Professor Luca Bargelloni che mi ha aiutato nel ruolo di tutor universitario e per i preziosi insegnamenti dati durante le lezioni.
Grazie ai miei compagni tesisti e dottorandi per avermi supportato (e sopportato) durante questo periodo.

Volevo ora rivolgere un ringraziamento particolare ai miei affetti più cari. In particolar modo ai miei genitori per cui le parole non bastano per descrivere la mia immensa gratitudine e affetto; a Enrico e Marzia che mi hanno dimostrato di esserci sempre e con cui abbiamo condiviso tanti bei momenti; ad Anna che da qualche anno è entrata a far parte della mia vita e a cui non posso più fare a meno; agli amici del "Vero Storico", della pallavolo, ai miei parenti e a tutti coloro che in questi anni mi sono stati accanto.