KU LEUVEN

FACULTY OF MEDICINE
BIOMEDICAL SCIENCES

# The genetic architecture of inflammatory bowel disease in multiplex families

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Biomedical Sciences by

**Deborah JANS**

Supervisor: Prof. Dr. Isabelle CLEYNEN

Leuven, 2021-2022

KU LEUVEN

FACULTY OF MEDICINE
BIOMEDICAL SCIENCES

# The genetic architecture of inflammatory bowel disease in multiplex families

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Biomedical Sciences by

**Deborah JANS**

Supervisor: Prof. Dr. Isabelle CLEYNEN

Leuven, 2021-2022

# Preface

During this academic year, I have investigated the genetics of inflammatory bowel disease in multiplex families with this thesis as a result. Not only has my enthusiasm for genetics and bioinformatics grown in the last year, I have learned a lot about what goes on behind the scenes of a lab. Most importantly, I have learned how to conduct and improve my own research. The person I have to thank the most for this experience is my supervisor professor Isabelle Cleynen. First of all, she gave me the opportunity to carry out the research for my master's thesis in the Laboratorium for Complex Genetics. Moreover, she supported me and, at the same time, gave me the freedom to make this project my own. I am very grateful for everything she has done for me during this year and I could not have wished for a better supervisor. I would also like to thank the other students – Paula, Amine, Wenjing, Joren and Kaan - and the PhD students – Sara and Yasmina - of the lab. They made the lab a welcoming place and they always were ready to help. Lastly, I would like to thank my family and friends. They might not have entirely understood what my research was about. Nevertheless, they always supported and encouraged me.

# Table of contents

# List of abbreviations

| | |
|---|---|
| CD | Crohn's disease |
| GWAS | genome-wide association studies |
| IBD | inflammatory bowel diseases |
| IIBDGC | international inflammatory bowel disease genetics consortium |
| PRS | polygenic risk scores |
| pT | p-value threshold |
| SNP | single nucleotide polymorphism |
| UC | ulcerative colitis |
| VEO-IBD | very early onset inflammatory bowel disease |
| WES | whole-exome sequencing |
| WGS | whole-genome sequencing |

# Abstract

Inflammatory bowel disease (IBD), encompassing Crohn's disease and ulcerative colitis, has as main characteristic inflammation of the intestinal tract. Some families have many family members affected with IBD, so-called multiplex families. The reason for this familial aggregation remains unresolved. Genetic and environmental factors are involved in the development and both are shared between relatives. I will study the genetic architecture of multiplex families to investigate if and to what extent genetics can be a reason behind their familial aggregation. For 55 multiplex families, 53 of European descent, with at least three affected first-degree relatives, I calculated polygenic risk scores (PRS) based on different p-value thresholds. I found that PRS including SNPs with a p-value higher than genome-wide significant (5e-08) or suggestive (1e-05) better predicted the case-control status of sporadic cases and controls, and within multiplex families. Thus, both sporadic and familial IBD seem to be truly polygenic, and some real associations are present in the less strongly associated variants. Affected relatives have a PRS similar to sporadic cases, and unaffected relatives have a higher PRS than the population controls. Thus, many common genetic risk variants seem to be segregating in these families. Yet, the PRS of affected relatives is higher than their unaffected first-degree relatives, indicating a higher burden of common risk variants in affected relatives. Of note, when families were looked at individually, some families had a PRS lower than the mean PRS of healthy population controls, indicating a very low burden of common variants. These could be interesting to study for sequencing as they are good candidates to carry a rare variant. Furthermore, a family-based association analysis indicated novel specific risk variants associated with IBD in families. In conclusion, familial aggregation seems to be due to a high burden of common risk variants in many families, however some families have another reason for familial aggregation.

# 1 Literature overview

## 1.1 Inflammatory Bowel Disease

Inflammatory Bowel Disease (IBD) is an overarching term which encompasses the main subtypes Crohn's disease (CD) and ulcerative colitis (UC).(1) As the name implies, IBD is a disease with chronic inflammation of the intestinal tract. Onset of disease can happen at all life stages, but diagnosis during childhood is often paired with a more severe course of disease. (2)

### 1.1.1 Incidence and prevalence

The incidence and prevalence of IBD differs based on geographical region.(3) In Europe and North America, the prevalence is the highest and even raises above 0.3%. However, incidence in these regions is stabilizing or even decreasing. Interestingly, a gradient from high to low runs through Europe with the highest incidence rate in western Europe.(4) Even within countries differences are being spotted, probably due to the degree of urbanization.(5) In contrast to the western countries, an increasing incidence is observed in Asia, South America and Africa. The prevalence is not yet equal to the western countries, but is being expected to follow in the footsteps of Europe and North America.

### 1.1.2 Clinical classification

The most widely used clinical classification system for IBD is the Montreal classification (Table 1).(6) UC and CD have a separate classification based on different parameters. The three main guidelines in CD are age of onset (A), location (L) and behaviour (B). Age of onset divides into three categories: younger than sixteen years, between the age of seventeen and 40 years, and older than 40 years. These are, respectively, A1, A2 and A3. Ileal, colonic and ileocolonic are the subtypes of disease location depicted by L1, L2 or L3, respectively. An extra subtype L4 can be included if isolated upper gastrointestinal disease is present. Disease behaviour is also categorised into three groups. B1 stands for non-stricturing and non-penetrating, B2 means stricturing disease behaviour, while B3 points at penetrating behaviour. Perianal disease has a separate symbol, namely p, which can be added to the behaviour type.

The subdivision of UC is only established by two characteristics, extent (E) and severity (S).(6) If only the rectum is affected, then it is classified as proctitis (E1). Left sided UC (E2) is spoken of when more of the colon is involved but inflammation goes no further than the splenic flexure. In extensive UC (E3), the inflammation goes beyond the splenic flexure. The other characteristic, severity, is not often used in practice. Severity is defined by the amount of stool passages per day, inflammatory markers and systemic toxicity. Rooted in these measurement arise four categories: clinical remission (S0), mild UC (S1), moderate UC (S2) and severe UC (S3).

Sometimes a definitive diagnosis of CD or UC cannot be made on the basis of clinical examination and endoscopic biopsies.(6,7) If this is the case, then the diagnosis inflammatory bowel disease type unclassified (IBD-U) will be made. Some occasions are more difficult to obtain a differential diagnosis of CD or UC, for example initial onset of disease, paediatric patients, treatment interferences or very severe disease.(7) Often, biopsies taken on a later timepoint can be of interest to determine whether the patient has UC or CD. The diagnosis of IBD-U is based on the routine clinical tests for IBD. Sometimes a colectomy is necessary and a more elaborate testing can be performed. If these additional tests are also inconclusive, a certain differential diagnosis cannot be established and these patients are classified as having indeterminate colitis.(6) Thus, IBD-U is often a temporary diagnosis

which can change if tests are repeated or with additional research, while indeterminate colitis is a definite diagnosis that probably will not change anymore.

**Table 1: Clinical classification of Crohn's disease and ulcerative colitis**

**Crohn's disease**

| Categories | Abbreviation | Information |
|---|---|---|
| **Age of onset** | A1 | < 17 years |
| | A2 | 17 – 40 years |
| | A3 | > 40 years |
| **Location** | L1 | Ileal inflammation |
| | L2 | Colonic inflammation |
| | L3 | Ileocolonic inflammation |
| | L4* | Isolated upper gastrointestinal disease |
| **Disease behaviour** | B1 | Non-stricturing and non-penetrating |
| | B2 | Stricturing |
| | B3 | Penetrating |
| | p* | Perianal disease |

**Ulcerative Colitis**

| Categories | Abbreviation | Information |
|---|---|---|
| **Extent** | E1 | Proctitis: only rectum involved |
| | E2 | Left-sided UC: inflammation until the splenic flexure |
| | E3 | Extensive UC: inflammation beyond the splenic flexure |
| **Severity** | S0 | Clinical remission |
| | S1 | Mild UC |
| | S2 | Moderate UC |
| | S3 | Severe UC |

*The characteristics are depicted in the left column. The short names of the categories are presented in the middle column and some extra information is shown in the right column. * an extra category which is additional to the other categories.*

### 1.1.3 Clinical presentation and diagnosis

IBD is sometimes troublesome to diagnose because of the unclear nature of the symptoms (Figure 1). Patients are typically experiencing symptoms for weeks or months.(8,9). Chronic diarrhoea, abdominal pain and weight loss are mainly encountered and this diarrhoea can contain blood, especially in UC. Other stool problems are also frequently observed, including rectal bleeding, tenesmus, urgency and nocturnal defaecation. Fever, tachycardia, nausea, vomiting and weight loss point at systemic symptoms and are mainly associated with severe UC but might also be seen in CD patients. Symptoms can vary widely in presentation and severity between patients which makes it difficult to diagnose IBD and to differentiate between its two main subtypes.

First, a comprehensive history of the patient is ascertained to exclude other causes, e.g. traveller's diarrhoea.(8–10) Occurring symptoms need to be thoroughly discussed. Family history of IBD and colorectal cancer can provide insightful information because IBD is partly hereditary and therefore runs in families. The main goal is to gather as much information that can indicate a possible diagnosis of IBD. For example, a recent smoking cessation or the use of non-steroidal anti-inflammatory drugs are often seen previous to the onset of IBD. A general physical examination might seem normal in some cases, mainly patients with mild to moderate disease.

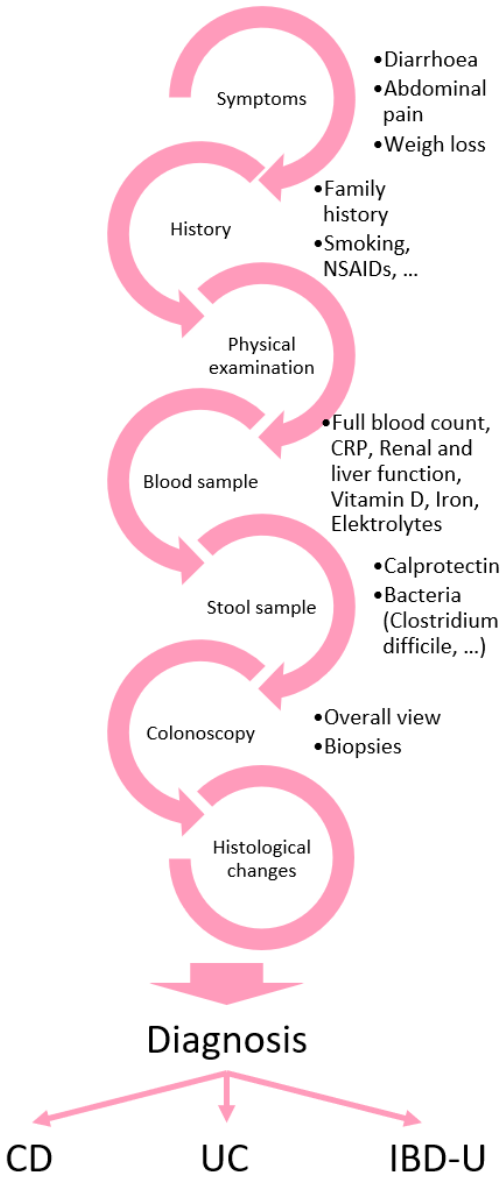



***Figure 1: Workflow to diagnose IBD***

*The general workflow to determine whether a patient has IBD and which specific subtype. Extra information is written at the side.*

Blood tests are an important part of the diagnosis but are neither specific, nor consistently abnormal.(8,9,11) Patients with a mild or moderate form of IBD can have results within the established range. The assay contains a full blood count, C reactive protein, markers for renal and liver function, vitamin D, iron and electrolytes. A stool sample is taken as well for the determination of faecal calprotectin, a marker of colonic inflammation. Infectious diseases caused by bacteria, especially *Clostridium Difficile*, can also be excluded through this stool sample. None of these tests are specific but only indicate ongoing inflammation. Therefore, a differential diagnosis of UC or CD cannot be made based solely on the results of a blood and stool sample. Although, the laboratory tests can indicate whether or not a colonoscopy might be useful, especially a high faecal calprotectin level denotes potential IBD patients.(10)

Endoscopic and histopathologic research are holding an important place in the differential diagnosis of CD and UC (Figure 1 and 2).(9–12) An overall view of the colon and ileum is performed with the

taking of several biopsies, a minimal of two biopsies on five different places. Inflammation in UC is limited to the colon and starts at the rectum, where it is often more severely present.(1) Continuous and symmetric inflammation with a clear boundary between healthy and inflamed tissue is typical for UC. In contrast, any part of the gastrointestinal tract, from mouth to anus, can be compromised in CD. The appearance of inflammation is not continuous, as is the case in UC, but is patchy and irregular. Erythema, granularity, partial or complete loss of visible vasculature, bleeding and ulcers can be found with an endoscopic examination in patients with UC.(11,12) CD has a typical cobblestone look and longitudinal ulcers.(9) The differences in endoscopic appearance might already indicate the subtype of IBD.

In UC, histological changes will only encompass the mucosal and submucosal layer (Figure 2).(1,11) The most distinguishing feature for UC is basal plasmacytosis.(12) This trait appears as one of the first in contrast with other aspects such as architectural damage and transmucosal inflammatory cell infiltrates. Inflammation in CD will extent deeper and might show granulomas.(1) Further, discontinuous chronic inflammation and crypt irregularities are also seen.(9) Typical features associated with one subtype of IBD, as diffuse crypt irregularity in UC or granulomas in CD, discriminates between the two and might make a differential diagnosis possible.

### 1.1.4 Natural history

CD is mostly a chronic intermittent disease, meaning most of the patients experience alternatively relapses and remissions.(13) Relapse rates increase each year after the diagnosis of a patient. A few patients, around 1%, endure a continuously active course over several years, but approximately half of them will go in remission within 3 years. At the other end of the spectrum, approximately 10% of patients remain in unremitting state for a longer period. This remission can be due to a surgical intervention. Surgery is often necessary.(5) One-third up to half of all CD patients require surgery

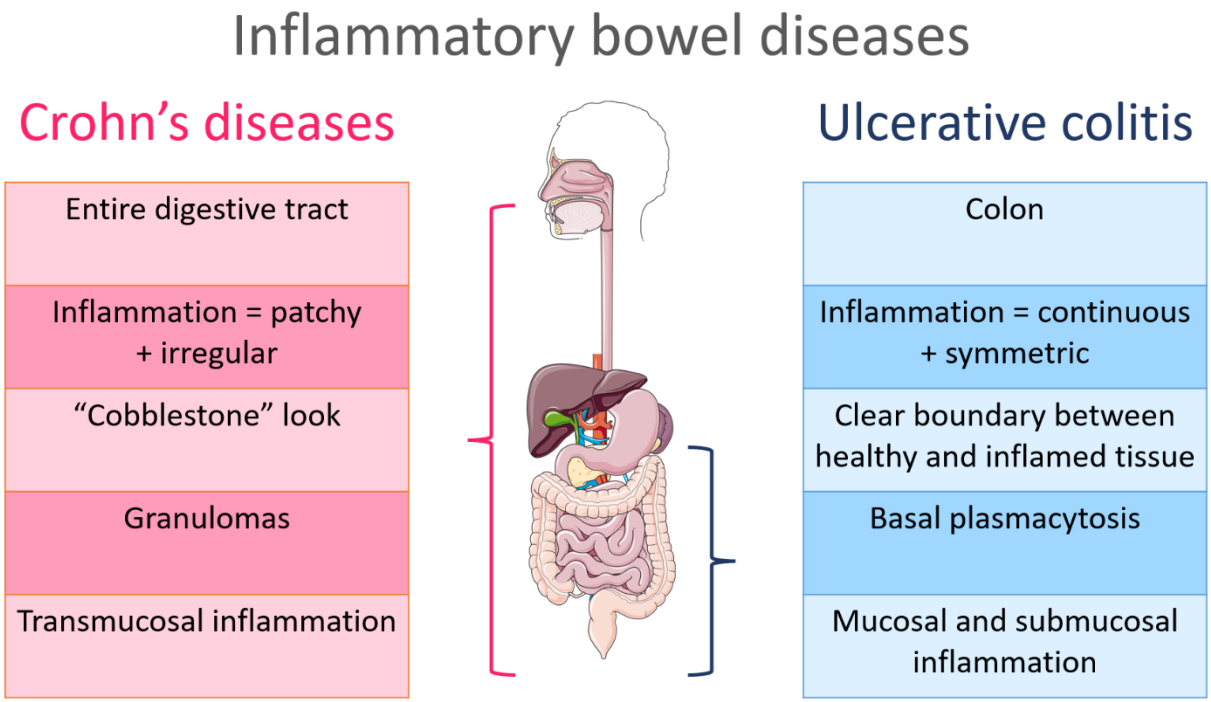



***Figure 2: The distinction between Crohn's disease and ulcerative colitis***

*Some typical characteristics for Crohn's disease (left) and ulcerative colitis (right). Accolades indicate which part of the gastrointestinal system is involved. The Figure was partly generated using Servier Medical Art, provided by Servier, licensed under a Creative Commons Attribution 3.0 unported license*

within 10 years after diagnosis. Likewise, hospitalization is a common event with the highest rate taking place in the first year after diagnosis and declining afterwards.

As described in the classification, CD is divided according to three parameters: age of onset (A), disease location (L) and disease behaviour (B). Age of onset is fixed but the other two might change over time. Disease location however is not very variable over time, most patients, 87%, remain in the same class (L1, L2 or L3) since the time of diagnosis.(5) A switch from disease behaviour is more likely to occur, from non-stricturing, non-penetrating disease (B1) to a stricturing or penetrating disease behaviour (B2 or B3).

A chronic relapsing disease with flares is also the main disease course in UC.(14) The cumulative risk for going through a relapse in a period of ten years following diagnosis is approximately 70 to 80%.(15) As in CD, some patients have a chronic continuous disease activity. Although, the percentage is higher in patients with UC, around 5%. A hospitalization is necessary at the time of diagnosis in 10-15% of the patients. Moreover, half of the patients will need a hospital stay. Once a patient required a hospitalization, the chances of a rehospitalization are very high going from one quarter within one year to three quarters within ten years after the first time. Cumulative colectomy rate is, at the moment, approximately 15% after 10 years. Although, surgery is mostly needed early, within the first year after diagnosis.(14)

Most UC patients present with left-sided colitis (E2) at the time of diagnosis.(16) Progression can only develop further proximally because it begins at the rectum. This means from proctitis to left-sided colitis or extensive colitis, or from left-sided colitis to extensive colitis. Extension is not very often seen, around 13% of patients have progression. This percentage is similar to the change of disease location in CD. At diagnosis, mild or moderate disease are mostly encountered.(14) Probably due to more effective treatments, the disease course switched from mostly moderate to a more mild course during the first five years after diagnosis.(15)

### 1.1.5 Extra-intestinal manifestations

IBD is not only restricted to the intestines but is also a systemic disease.(17) Approximately half of the patients will develop an extra-intestinal manifestation.(18) Sometimes, extra-intestinal manifestations are the first symptoms to occur and this makes it even harder to diagnose a patient with IBD. Interestingly, CD patients are more prone to extra-intestinal manifestations than UC patients. Although almost any organ can be involved in IBD, the organs which are mainly affected include the biliary tract, eyes, joints and the skin.(17) However, other organs are not entirely excluded, for example osteoporosis, pulmonary diseases and liver diseases are also frequently observed in patients.(18) Thus, a lot of heterogeneity exists in extra-intestinal manifestations and they all have their own optimal treatments.

### 1.1.6 Mortality and morbidity

An overall increased mortality is not observed in UC patients. However, an increased risk of mortality is associated with some extra-intestinal symptoms like liver diseases and pulmonary diseases. Although no elevated overall mortality is seen, a high rate of morbidities is encountered. Patients often complain about fatigue and might experience sleep difficulties. Patients report a lower quality of life, including due to fatigue. Furthermore, patients are often not able to work.(15)

CD patients have an increased mortality but only after 25 years from diagnosis.(19) Many deaths of CD patients are due to CD-specific causes. Intestinal failure, intestinal cancer, severe diseases are observed as causes for an early death in CD patients. More CD patients than UC patients seem to have

disabilities.(20) A lower quality of life is also reported by CD patients. Especially, active disease for a long time and psychological distress are associated with a lower quality of life. A higher rate of unemployment is also found in CD patients.

### 1.1.7 Treatment

The European Crohn's and Colitis Organisation (ECCO) on a regular basis publishes guidelines for the diagnosis and treatment of CD an UC.(9,21) These guidelines are made by experts based on clinical trials and meta-analyses. UC and CD have differing characteristics and therefore their treatment is not entirely the same. Although, medication currently available on the market is often given to both groups of patients. Treatments are chosen founded on several parameters including the specific disease characteristics, disease severity, benefit-risk ratio of the treatment, previous responses and individual factors.

The aim of initial treatment of CD and UC is focused on inducing remission in patients.(10) If this is achieved, further therapy needs to maintain remission and prevent another flare. Therefore, a top-down approach is applied and often biologicals and corticosteroids are prescribed first. However, some patients are not helped with medication and surgery is often necessary. For example, a colectomy is often performed in severe UC.(22)

Medical treatment of UC and CD can be divided in a few groups. First, aminosalicylates is the first-line therapy to maintain remission in UC but has no or very few effects in CD and is therefore not recommended.(10,21–23) Another large group is formed by the corticosteroids. The form of administration ranges from local, over oral, to intravenous if necessary. These are preferably not given chronically but are well suited to induce remission or manage flares. Several immunomodulators also belong to the package of possible therapy choices in the case of CD and UC. Thiopurines, azathioprine and mercaptopurine, are well-known and often applied therapies for UC and CD. Methotrexate is still listed as a treatment but efficacy is becoming more and more a point of discussion.(23) Two others are also available for UC, namely cyclosporine and tacrolimus. Biologicals make up the largest cluster of treatments. Anti-TNF therapy, Adalimumab and Infliximab, belong to the biologicals and are the preferred option if previous therapies seem inadequate. Vedolizumab, a gut selective anti-inflammatory, can be administered in both UC and CD. In patients with CD an alternative to vedolizumab is ustekinumab which binds the pro-inflammatory interleukins 12 and 23. Lastly, a janus kinase inhibitor, Tofacitinib, recently came on the market for UC.

Diet as a treatment of IBD is an ongoing research. The gut microbiome which plays an important role in IBD, is influenced through nutrition.(24) Therefore, changing or restricting food intake might alter the disease course of IBD. Evidence for restriction diets is limited and a control group is often not included in the study. Sometimes, parenteral or enteral nutrition is given to let the bowels rest. Results for total parenteral nutrition are highly variable between studies. In contrast, exclusive enteral nutrition can induce remission. As a remark, exclusive enteral nutrition is for adults a very difficult lifestyle to maintain. Partial enteral nutrition is better manageable but is only able to prolong remission, not induce it. A diet that mimics exclusive enteral nutrition but with food can partially modify the microbiome to resemble exclusive enteral nutrition and might be easier to maintain. Diet is an important environmental factor that exerts an influence on IBD, and might be an interesting treatment option.

General health should not be overlooked and is followed up regularly.(10) Patients are encouraged to stop smoking and adopt a healthy lifestyle. Deficiencies are common in patients with IBD, especially iron, folate, vitamin B12 and D. If a deficiency is detected, this will be treated with supplements. Some

therapies given to IBD patients have severe side effects. Adequate follow-up is recommended. For example, thiopurine therapy enhances the risk on skin cancer and corticosteroids lower bone density.

## 1.2 Genetics of IBD

First- and second-degree relatives of IBD patients have a higher prevalence and an increased risk of IBD than the general population.(25) This suggests a genetic component, but families also share their environments. One way to circumvent the problem of the shared environment in establishing the genetic component to disease is by using twin studies.(26) Twins typically also have a more shared environment than do regular siblings or related individuals. Dizygotic twins have approximately 50% of their genome the same, making them genetically first-degree relatives, while monozygotic twins are genetically identical. This difference can be used to calculate trait heritability. The results of twin studies pointed at a genetic component to IBD, and a heritability of CD estimated to be between 70 and 80%, while UC is predicted at 60 to 70%.(27) The genetic component of UC is thus lower than of CD but both are still fairly large.

### 1.2.1 Monogenic or polygenic

Both CD and UC are partly caused by a genetic component. The underlying genetic architecture is important for insights into disease pathogenesis, treatment, risk prediction, ... The two extreme forms of the genetic architecture of IBD are a monogenic and polygenic architecture (Figure 3). Monogenic means that a mutation in one gene is sufficient to develop the disease. Mutations are not prevalent in the population, while single nucleotide polymorphisms (SNPs) are typically shared by many individuals. The effect of one SNP is not enough to cause disease but many SNPs in multiple genes together might be sufficient. If multiple genes (and thus risk variants) are necessary to cause the disease, it is referred to as a polygenic disease. At the moment, IBD is generally seen as a polygenic disease, with each gene having a small effect that contributes to disease risk.(28)

Initially, IBD was believed to have a monogenic recessive mode of inheritance, however the first discovered susceptibility locus, IBD1 (*NOD2*), explained only a tiny fraction of the ten-fold increased risk seen in first-degree relatives of IBD patients.(29) Subsequently identified loci did not demonstrate higher effect sizes and therefore indicated a more polygenic or complex nature for IBD. Some monogenic disease forms of IBD with one causal gene exist but these are only encountered in a small group of paediatric patients. According to the study of Crowley *et al* (2020), 3% of paediatric patients have a monogenic cause of IBD.(30) Most early-onset IBD cases are polygenic and follow therefore the rules of common diseases. Thus, a small part of IBD patients indeed have a distinct genetic architecture with a monogenic cause, however the majority has a polygenic form (Figure 3).

The ‘common disease, common variant’ hypothesis says that a disease common in the population, like IBD, is caused by many common variants with a small effect size. This hypothesis is supported by the many successful genome wide association studies (GWAS), where more than 240 loci associated with IBD are identified so far (Figure 4, also see below for more details).(31–33) Chen *et al* (2014) calculated how much of the heritability of IBD can be explained by GWAS based on Immunochip, a chip with a high density of SNPs in regions associated with immune diseases, and based on imputed data from a more general GWAS chip.(27) The percentage of heritability explained based on Immunochip was 27% for CD and 21% for UC, and with the GWAS chip it was 37% and 27%, respectively. The remaining part is unexplained and therefore is referred to as the ‘missing heritability’.

On the other hand, rare variants with a moderate or high effect also contribute to disease risk.(34) This is the ‘common disease, rare variant’ model. This hypothesis predicts many risk variants with a larger

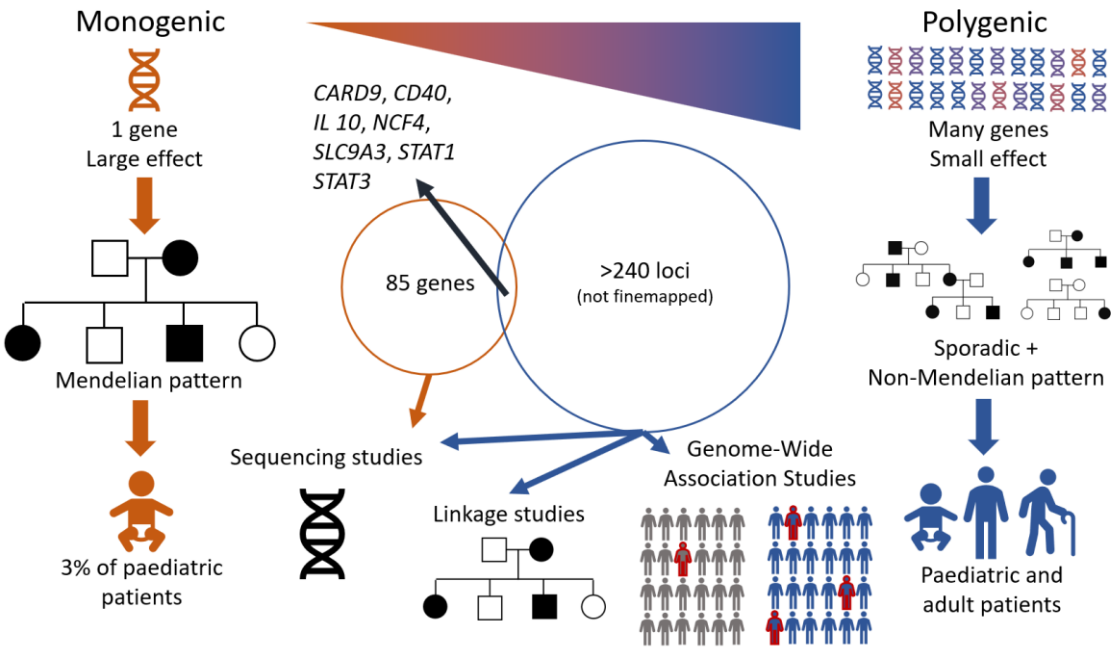



***Figure 3: Overview of the main differences between monogenic and polygenic IBD and how they are studied***

*IBD encompasses a spectrum of genetic variation. At one end of the spectrum, the monogenic form (left) is only seen in a small amount of paediatric patients.*(30) *The polygenic form (right) is localized at the other end but covers the middle part of the spectrum as well. The onset of disease can start at any age. The circles in the middle represent the number of genes or loci associated with the specific forms of IBD. A small part of genes are overlapping between the monogenic and polygenic form. Numbers and genes are derived from Jezernik et al (2020).*(39) *The different kind of studies applied to discover those genes or loci are depicted at the bottom.*

effect associated with a disease, like IBD, but with each variant having a very low population frequency. Unfortunately, these low frequency variants are difficult to pick up with standard GWAS. Very large sample sizes would be necessary to detect them, as only few individuals will carry them. Variants with a larger effect size will be a bit easier to find because then more cases will carry the risk variant. To find rare variants, targeted, whole-exome, or whole-genome sequencing studies are necessary.(35) In the study of Hunt *et al* (2013), the exome of 25 risk genes for six autoimmune disorders, including CD, were sequenced to determine the contribution of rare variants to the heritability of common diseases.(36) Eventually, they conclude from their results that the added value of rare variants to the heritability of these diseases is negligible. However, only protein-coding variants from those 25 genes were included and other rare variants were not considered. Recently, a very large study based on 269,171 individuals of European ancestry and 11,933 individuals of African, East Asian and South Asian ethnicity, was published with opposite results: many associations between protein-coding variants and phenotypes, including IBD, originated from rare variants.(37) These rare variants have also significantly higher effect sizes than common variants detected in GWAS. Indicating the importance of the contribution of rare variants to common diseases.

The exact genetic architecture of polygenic IBD is still debated. How much common and/or rare alleles contribute to disease risk is unknown. Agarwala *et al* (2013) simulated many simple models with differing parameters for the relationship between the variant's effect on fitness and its effect on a particular disease, and the mutational target size.(38) These models were compared with data from type 2 diabetes studies to test the consistency. Only the extreme models in which almost everything is explained by rare variants, or by common variants, are excluded. All other models with a variable

contribution of rare variants, ranging from less than 25% to more than 80% of heritability, and common variants are possible. Importantly, they also discovered that cohort sizes need to be sufficiently large, e.g. hundreds of thousands of individuals, in GWAS as well as sequencing studies to elucidate the genetic architecture of common diseases.

IBD thus encompasses a wide genetic spectrum going from polygenic forms with many common variants prevalent in the population and with a low effect size, to monogenic forms with rare variants having a causal effect (Figure 4). The question arises whether the involved genes are similar between the extremes of the spectrum. To date, 85 causal genes of monogenic paediatric IBD are discovered.(39) Only a very small overlap exists between the genes in or near the more than 240 risk loci of complex IBD and these 85 causal genes; the overlapping genes comprise *CARD9*, *CD40*, *IL10*, *NCF4*, *SLC9A3*, *STAT1* and *STAT3* (Figure 3). It should be noted that although the genes overlap, the specific variants involved are not the same: typically more common variants with low effect size for the polygenic form, and causal/pathogenic rare variants for monogenic IBD. While a large discrepancy exists between the two with regards to the specific genes involved, gene ontology analysis shows that 59 out of 424 terms (13.9%) were enriched in both groups.(39) Overlapping terms include Th1 and Th2 cell differentiation, Th17 cell differentiation and Jak/STAT signalling. Thus, although the genes are not the same between monogenic and polygenic IBD, the pathogeneses show some resemblance.

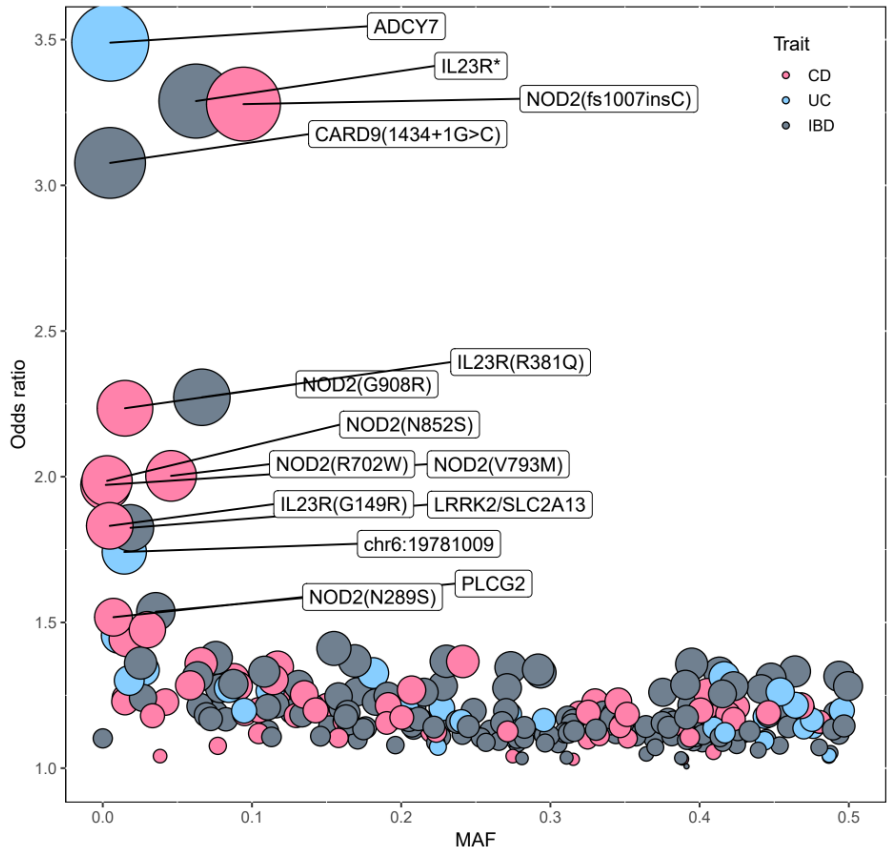



***Figure 4: Inflammatory bowel disease associated variants***

*SNPs associated with IBD are plotted according to their odds ratio (y-axis) and minor allele frequency (x-axis). Colours indicate the associated trait (CD, UC, or IBD) and the odds ratio of this particular trait is presented. Odds ratios are obtained from Huang et al (2017)* (61) *when loci could be fine mapped with a posterior probability >50%, and de Lange et al (2017)* (31)*. SNPs with an odds ratio >1.5 are labelled.*

### 1.2.2 Multiplex families

IBD is often seen in families which points to a genetic component. Importantly, the relative risk for family members of an IBD patient is increased compared to the general population.(40) Predictably, first-degree relatives share approximately 50% of their DNA, and therefore have the highest relative risk in comparison with other family members, with the exception of monozygotic twins. Interestingly, the risk for a relative of a CD patient to develop IBD seems to be higher than for a relative of a patient with UC. This observation is in line with the higher heritability of CD compared with UC, 70-80% and 60-70%, respectively.(27)

Above I briefly discussed monogenic IBD forms with a Mendelian inheritance pattern - autosomal dominant, autosomal recessive, X-linked,… - are prevalent in families and often have a young age of onset. However, some families have a remarkably high prevalence of disease compared to the expected population prevalence without showing a clear Mendelian inheritance pattern or a particularly young age at onset.(41) These are termed multiplex families - families in which more members are affected by disease than would be expected based on prevalence in the general population. Why a high appearance of IBD is seen in these families is unclear.

Linkage studies use the genetic similarity in families to pinpoint loci relevant to disease (Figure 5). Family members share parts of their genome but every related pair will have different overlapping pieces. By looking at segments of DNA shared more often between all or most affected individuals in a family, loci linked to the disease can be found. Several such loci were found for IBD using genome-wide linkage screens, termed IBD1-IBD9.(42) IBD1 was the first susceptibility locus found for CD and was replicated through a large international collaboration.(43) Subsequent fine-mapping of IBD1 led to the identification of *NOD2*, still the most strongly CD-associated risk gene.(44) Unlike for monogenic diseases where this type of linkage studies were very successful in finding the pathogenic genes, failure to replicate findings from linkage studies however was an often encountered problem with complex diseases. It turned out penetrance of the variants was too low, and the pedigrees studied hence too small, to be really successful. Therefore, research moved its attention to the more powerful GWAS, partly encouraged by technological advances.(45)

The scarce discovery of risk loci with early linkage studies does not mean that multiplex families are not relevant to study genetically. On the contrary, families with multiple affected members might provide valuable insights into IBD genetics.(41) They might carry an exceptionally high number of common risk loci, or a rare allele with a modest or high effect. One study on five multiplex IBD families by Stittrich *et al* (2016) found four families with a high genetic risk burden.(46) Interestingly, one family carried even less common risk variants than the general population. In this family, TRIM11 was proposed as a candidate risk gene because of its segregation with affected family members and its predicted function in the NF-κB signaling pathway, known to be associated with IBD.

In contrast to the study of Stittrich *et al* (2016), a study on eight multiplex Korean families found only one family which had a high burden of common risk alleles.(47) Indicating that it is not the accumulation of many common alleles but mostly a few high effect variants which contribute to familial aggregation in these families. Of note, here the inclusion criteria required more than two affected first-degree relatives while Stittrich *et al* (2016) focused on very large families containing multiple generations and at least three affected family members which were not necessarily first-degree relatives. Seventeen candidate genes which included potentially deleterious rare variants, were discovered, although further validation is necessary. Some multiplex families have a lot of common alleles which probably clarifies the high burden of IBD-affected members.(41,46) On the other hand,

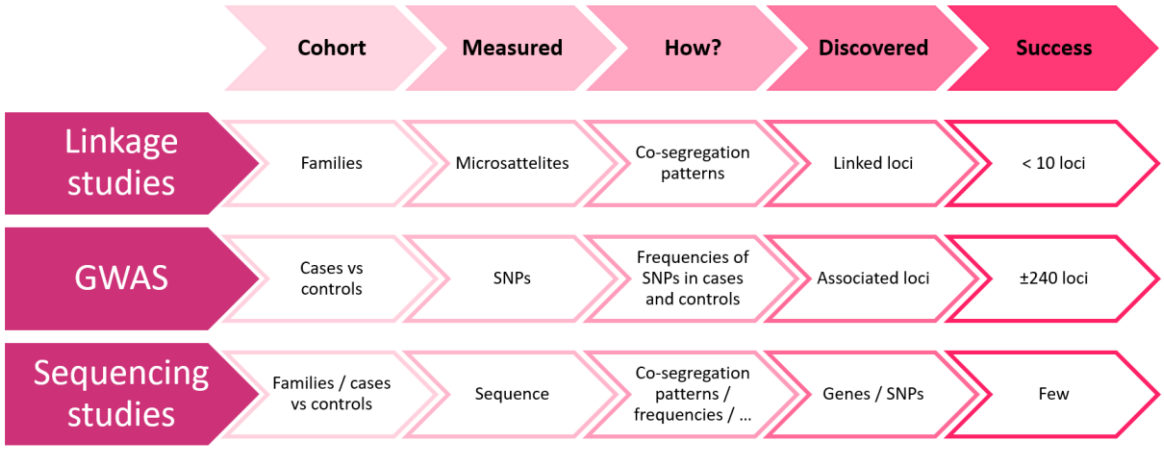



***Figure 5: Methods applied to discover associated variants***

*Many research has already been performed to detect loci associated with inflammatory bowel disease. Methods which have been used previously and currently are listed at the left side. More details about each method is provided according to the categories presented at the top.*

discovering rare variants, presumably with a higher effect size, is also possible in some families as the previous example illustrates.(46,47)

Also other studies identified rare variants in multiplex families. A recent study discovered a new variant in *NOD2* in the leucine rich repeat domain (LRR domain) with possible deleterious effects in a family with multiple members affected by CD.(48) Here, four affected members and three familial controls were sequenced, and the *NOD2* N1010K variant was found co-segregating with CD. *NOD2* L1007fs was also present in the family, and individuals who had both *NOD2* variants exhibited a more severe course and a younger age at diagnosis, adult vs paediatric age. Likewise, a *FOXP3* mutation was identified in a family with an IBD affected mother and three affected sons.(49) Interestingly, this missense mutation is located on the X-chromosome. The three sons already showed symptoms and multiple extra-intestinal manifestations at a paediatric age.

### 1.2.3 Genome-wide association studies

Genome wide association studies (GWAS, Figure 5) are frequently applied to diseases with a complex genetic architecture. A few hundreds of thousands of common variants (SNPs) throughout the genome are measured in individuals with a certain disease, the cases, and in healthy controls. The frequency of variants in both groups are compared to extract information about which variants are more or less prevalent in cases. Each variant is entrusted with an effect size based on the differences in frequencies between the cases and controls. GWAS were very successful in IBD, so a short overview cannot be missed.

Many GWAS were carried out over the years with ever increasing sample sizes (Figure 6). At the beginning of the quest to discover genes associated with IBD relatively small studies specific for CD or UC were conducted. Duerr *et al* (2006) included 547 cases with CD and 548 controls, all of European ancestry.(50) The result of this GWAS were three SNPs significantly associated with CD. Two SNPs pointed at the *NOD2* gene, already known from linkage studies, and the *IL23R* association was newly discovered. Interestingly, the identified *IL23R* variant is protective against CD. Some studies with similar cohort sizes were performed around the same time and identified only one new locus each (an intergenic region on 10q21.1, ATG16L1 and a region in a gene dessert on 5p13.1 ).(51–53)

Subsequent GWAS included more and more individuals and found more and more associated loci (Figure 6). One way to further increase the sample size without having to do a new GWAS is to perform a meta-analysis of existing GWAS.(54) A meta-analysis can be supplemented with new data but this is not a necessity. A UC meta-analysis built up of six previous GWAS and adding up to a sample size of 6,687 UC cases and 19,718 controls established an association between UC and 29 newly identified loci.(55) This at that time more than doubled the known loci from 18 to 47 loci. A similar meta-analysis was performed for CD.(56) Data from six GWAS were pooled to retrieve 6,333 CD cases and 15,056 controls, and leading to the detection of 30 new risk loci. A larger number of 41 associated loci were already known for CD. Of note, 21 of these 41 loci were derived from a previous smaller meta-analysis comprising data from three GWAS, which were also part of the meta-analysis with six GWAS.(57) A larger cohort of individuals gives more power to detect associations even if the data is not new but a combination of existing data sets.

While these first GWAS studied either CD or UC, in later studies, CD and UC cases were combined as IBD. Importantly, it are still two separate diseases and effect sizes are calculated combined but also specifically for each type. Merging the two groups of patients increases the sample size and therefore enlarges the power to find associated loci. Combining CD and UC was applied by the International IBD Genetics Consortium (IIBDGC) in a landmark study including 33,867 cases and 37,479 controls.(33) A cost-effective custom genotyping chip designed for immunogenetics studies, Immunochip, was used to genotype individuals. This study increased the number of associated loci to 163, of which 30 were specific to CD, 23 to UC, and 110 showed an association to both. The most recent GWAS conducted for IBD on 59,957 persons from European descent was published in 2017 by de Lange *et al* (2017).(31) They further expanded the list of IBD-associated loci to a total of ca 240.

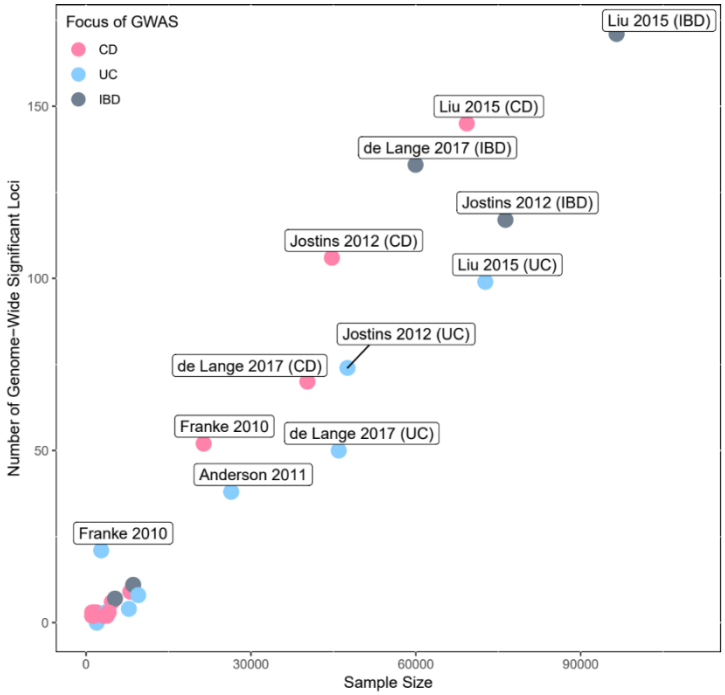



***Figure 6: Overview of genome-wide association studies and meta-analyses in IBD***

*Genome-wide association studies are plotted according to the sample size (x-axis) used in the discovery phase and how many loci (y axis) were genome-wide significant in their study. Colours of the dots represent the focus of the GWAS. If multiple phenotypes are investigated in the same study, than each trait is displayed by a separate dot. Studies with >20 genome-wide significant loci are labelled. CD = Crohn's disease, UC = ulcerative colitis and IBD = inflammatory bowel disease.*

Previously mentioned studies mostly included individuals with a European ancestry. While the European population is overrepresented in GWAS, also in those for IBD, a few studies are done including other ethnicities. The first GWAS and meta-analysis with a mixture of ethnicities, genotyped on Immunochip and several other genotype chips, is performed by the IIBDGC.(32) The study included 96,486 individuals of four ancestries, namely 86,640 European, 6,543 East Asian, 2,413 Indian and 890 Iranian participants. The majority of participants was thus still of European descent. They identified 38 new loci and increased the known associations to approximately 200 loci at the time. Another meta-analysis was conducted with Immunochip data from 9,060 Asian, Korean and East-Asian, and the 86,640 European individuals of the IIBDGC study.(58) Seven novel associations could be found of which three are located in previously undetected loci.

These meta-analyses proved that new loci can be discovered through combining samples from different ethnicities. They also showed that some loci are specific to certain ancestries. Liu *et al* (2015) performed a trans-ancestry analysis and also examined heterogeneity in effect between ancestries. Although the effect of most loci is similar across populations, differences in allele frequency and/or effect sizes were observed for a few loci, including *NOD2*, *IL23R* and *ATG16L1*.(32) A small GWAS conducted in African Americans identified two African-specific UC loci, namely *ZNF649* and *LSAM*, and one for IBD, *USP25*.(59) Remarkably, in the same study no significant SNPs were found for CD. This was a small study and larger studies will probably find more ethnicity-specific associations, also for CD.

Unfortunately, GWAS typically pinpoint genomic regions where causal gene(s) are located, but not necessarily directly identify the causal genes.(60) The variant which causes the increased risk indeed is not necessarily directly identified, but could be in linkage disequilibrium (LD) with the associated SNP. Therefore, fine-mapping studies are necessary to identify the causal variants and genes, as was for example done in the study by Huang *et al* (2017) .(61) They fine-mapped 94 IBD loci and could define 45 associations to a single variant with more than 50% certainty, 18 associations even had a certainty above 95%. This study used a very large cohort of 33,595 patients with IBD and 34,275 controls, contributing to its success. Nevertheless, many causal variants and genes are still unknown, and with them their function in pathogenesis.

### 1.2.4 Sequencing studies

Current GWAS are not well suited to detect rare variants. Rare variants are indeed not typically included on the SNP arrays used, although more recent arrays do include more low frequency variants, down to 1% minor allele frequency. Direct sequencing of DNA is the best way to discover rare variants associated with IBD (Figure 5).

Large amounts of data are generated with sequencing, and these have to be adequately processed.(35) In the early days therefore, usually targeted sequencing of specific gene regions was done as less data is obtained. In one of the first targeted sequencing studies genes within loci known to be associated with CD through GWAS – 71 loci at that time – were resequenced.(62) They successfully sequenced 56 genes present within loci associated with CD in 350 CD cases and 350 controls, pooled per 50 individuals. Several new variants associated with IBD were discovered, including four additional risk variants in *NOD2*, two protective variants in *IL23R*, another protective variant in *CARD9* and a few other variants. Interestingly, this means that common variants with low effect size and rare variants with a higher impact can reside in the same gene.

A few years later, the research team of Prescott *et al* (2015) increased the number of sequenced genes to 531, with additional candidates predicted by pathway and protein network analyses.(63) This resulted in a novel rare association in a known gene, *BTNL2*, in which two common variant associations

were already known. Further, only three suggestively associated rare variants could be detected. Although they added many more sequenced genes to their analysis, a new gene associated with IBD could not be detected.

With new algorithms being developed and technologies becoming even better, whole exome sequencing (WES) and whole genome sequencing (WGS) could be applied more easily. These technologies do not rely on previous knowledge. A small study, in today's standards, collected sequence data of 42 CD patients and 5 controls.(64) Three missense variants were detected in this limited cohort. *PRDM1* contained two of these three variants and was located in a locus previously uncovered by GWAS. Interestingly, while the variants pose a risk for CD, they are protective for UC. The other missense variant resides in *NDP52*, which was an entirely new CD association.

Distinguishing causal variants from neutral variants in sequencing is not always easy. Co-segregation of the variant with disease in families could indicate that the variant or one in linkage disequilibrium increases risk of the disease. The study of Onoufriadis *et al* (2018) is a nice example to illustrate this.(65) The most distantly related IBD-affected individuals in ten families with at least three affected first-degree relatives were whole-exome sequenced. A very stringent stepwise filtering approach was applied, including that two families had to carry a rare variant in the same gene. After this filtering, 34 rare, protein-altering variants in 17 genes remained, and only one of them could be found in a known GWAS locus, namely *NLRP7*.

GWAS used increasingly large study groups over time, and sequencing studies followed in these footsteps. Recently, a very large WES study included more than 30,000 CD cases and 80,000 controls coming from 35 centres to further identify rare variants associated with CD.(66) Only variants with a population frequency between 0.0001 and 0.1 were considered for further analyses. Association analyses found eleven newly associated variants, from which five are located in novel loci. In addition, one new gene, *ATG4C*, was implicated by the gene-based rare-variant burden tests. Thus, this study shows that adequately powered sample sizes can find rare variants associated with disease, and that these variants can reside in novel as well as in known loci.

The possibility exists that some disease-associated rare variants are located outside of the exome in non-coding regions. WGS would be necessary to pick up these rare variants. To date, not many WGS studies are performed with IBD patients. The largest is a low coverage whole genome sequencing study on 7,932 individuals, including 2,513 CD patients, 1,767 UC patients and 3,652 controls.(67) They found a missense variant in *ADCY7* that increases the risk of UC. Although this is already a large cohort, finding rare variants probably needs much larger sample sizes and a higher coverage.

Another, very recent, WGS study is performed by Somineni *et al* (2021).(68) They sequenced 1,774 individuals diagnosed with IBD and 1,644 healthy controls of African origin living in America. The main goal of the study was to compare the genetics of African-Americans with European data available. First, common variants were looked at, and 41 loci, previously found by GWAS with individuals from European descent, could be replicated in the African-American cohort. As seen in previous trans-ancestry studies, the loci largely corresponded in effect sizes and allele frequencies but some differences were seen as well. For the rare variant analysis, variants were filtered on the criteria of being likely deleterious and aggregated in sets based on their location close to a gene. *CALB2* is put forward as a possible association with UC. Although 35 variants contributed to the signal of *CALB2* in African Americans, most variants were not seen in European individuals. The authors conclude that common variants are shared between populations from different ancestry but rare variants are probably population-specific.

### 1.2.5 Genetic overlap with other diseases

Some diseases or traits have genetic overlap in variants and loci.(69). One variant can be associated with multiple diseases. Sometimes the associated variants are different but they point to the same gene or locus, which can include more than one gene. The differential effect on phenotype could be due to independent pathways. Another possibility is the gene causes one phenotype and this is in turn causal for another phenotype. Alternatively, two diseases could have a common pathway that influences disease risk. A large study compared available GWAS for 42 traits, including immune-related traits, and found 341 loci to have an association with different traits.(69) Immune-related diseases, including CD, have an elevated proportion of overlapping variants in GWAS. Although many genes are mutually causal between immune-related diseases, the effect sizes are different.

With the existence of genetic overlap between immune-mediated diseases clearly established, Ellinghaus *et al* (2016) wanted to explore the relationship between five diseases, namely CD, UC, ankylosing spondylitis, primary sclerosing cholangitis and psoriasis.(70) Their meta-analysis detected three new shared loci. The genetic similarity can be used to perform larger meta-analyses with more individuals and therefore with more power. Moreover, some new associations for specific diseases with genome-wide significance were found, including six loci for CD. Importantly, while a clear genetic overlap between the five diseases exist, they each also have their own specific risk genes.

Comorbidity of IBD and multiple sclerosis, a disease of the central nervous system, might seem less evident but it occurs.(71) The relationship is less far-fetched than it looks upon first glance because multiple sclerosis is an autoimmune disease with inflammation of myelin around neurons. Genetic correlation exists for multiple sclerosis with both CD and UC, but is clearly larger for UC than for CD. Another neurodegenerative disorder, amyotrophic lateral sclerosis (ALS), also has an association with dysregulated immunity.(72) In contrast with multiple sclerosis, a negative correlation has been established between ALS and IBD. One of the SNPs at *NOD2*, that has a strong effect size in IBD, is a shared risk factor between ALS and CD. Other common risk factors between ALS and IBD are *G2E3* and *SCFD1*.

Even less obvious diseases can also share some genes with IBD, like Parkinson's Disease.(73) *LRRK2* carries the mutation G2019S that is the best known genetic cause of Parkinson's Disease. IBD has been associated with the gene but not with that specific mutation. Another variant, N2081D, located in the same kinase domain as G2019S shows association with both CD as well as Parkinson's Disease. Moreover, the effect sizes of variants, at least nominally associated with both diseases, share the same direction of effect and are significantly correlated. Both risk as well as protective variants are found in the *LRRK2* gene. Due to the different location in the body of the two diseases, the mutated protein acts with various consequences. This genetic overlap might be of relevance for the development of new therapies which can be beneficial for Parkinson's Disease as well as IBD patients.

## 1.3 Polygenic risk scores

### 1.3.1 What are polygenic risk scores

As mentioned above, over 240 loci that have an association with IBD are discovered so far.(31–33) Looking at one such variant is not useful to estimate the risk for developing IBD because it typically only infers a very small risk. Even variants with a larger effect size, e.g. the *NOD2* frameshift variant, do not give much information about someone's risk for IBD.(60) On the other hand, combining the risks for all known loci in so-called polygenic risk scores (PRS) might provide better predictions.

Different ways from very simple to more complex models are available to compute a PRS.(60,74) The easiest model, which is not commonly used, is simply counting how many risk alleles a patient has in their genome. It does not consider the relative risks of these variants. More commonly applied models use a weighted PRS. Alleles are still summed but each allele has a certain weight allocated. Therefore, alleles with a larger effect size contribute more to the risk score than small effect alleles. The weights applied can be the effect sizes or the odds ratios of the SNPs measured in GWAS or they can be adjusted based on assumptions, like the number of causal SNPs or their potential function, to optimize the PRS.(74) Another difference is how data are shrinked to avoid poorly predicted PRS.(75) This can be done by effect size shrinkage which depending on statistics lowers the effect sizes of all SNPs, or some SNPs more than others depending on various parameters. Another option is only including SNPs that reach a specific p-value threshold of association with the disease. Some research indeed indicates the use of higher p-values than the genome wide significance level of $5x10^{-8}$.(76) That p-value is very strict and many associated SNPs will be missed. By using a higher p-value, SNPs not (yet) reaching genome-wide significance in the discovery GWAS are also included and typically improve the performance of PRS, indicating that some of those SNPs are also important for IBD development. Which method is best to calculate PRS can differ between diseases and depends on the available data and therefore should be determined case by case by testing several models.

As PRS are based on the effect size estimates from a GWAS, polygenic risk scores can be calculated for any phenotype for which a GWAS is performed. Larger GWAS detect more associations and give more accurate predictions for the effect size. Thus, the larger the discovery GWAS, the better the accuracy of the PRS. Mostly, these GWAS are carried out with individuals from European descent and other populations are underrepresented. Linkage disequilibrium and allele frequencies differ between populations and this can have consequences for the PRS. Therefore, the performance of PRS works best in a population which is closely related to the GWAS discovery population.(77) This is true for geographically different populations, e.g. Asians and Europeans, but also for other factors, e.g. a high-risk population vs the general population.(74)

### 1.3.2 Polygenic risk scores, and their added value

The first study performed with PRS in IBD applies a simple PRS calculation based on the summation of five associated genes, *NOD2*, *DLG5*, *ATG16L1*, *IL23R* and IBD5 region.(78) Here was already shown that individuals with more risk alleles, and thus a higher PRS, had a higher risk on the development of CD and a more severe disease course. This was however a small study with 1,684 CD patients and 1,350 controls based on only five associated genes.

The first study in IBD which extensively utilizes PRS based on many loci was, a large international genotype-phenotype study, investigating the association between PRS including the 163 known associated loci at the time and clinical subphenotypes age of diagnosis, time to surgery, disease location and behaviour (CD), and disease extent (UC) in almost 30,000 patients diagnosed with IBD.(79) The CD-PRS and UC-PRS showed a strong association with the disease subphenotypes age at diagnosis and disease location. Interestingly, a PRS which explored the differences between CD and UC (CD vs UC PRS) showed the strongest associations with the clinical subphenotypes. The CD vs UC PRS also hinted to a different classification of IBD: not simply into CD and UC, but into UC, and colonic CD and ileal CD as different entities. The predictive accuracy however was too low to be used in the clinic to distinguish between subtypes (AUC = 0.60). However, patients with an extremely low or high CD vs UC PRS were more often found to be misdiagnosed as CD or UC respectively. Later studies also found associations between PRS and subphenotypes. Although the study of Chen *et al* (2017) was focused on testing which model had the best prediction performance, they also investigated the association

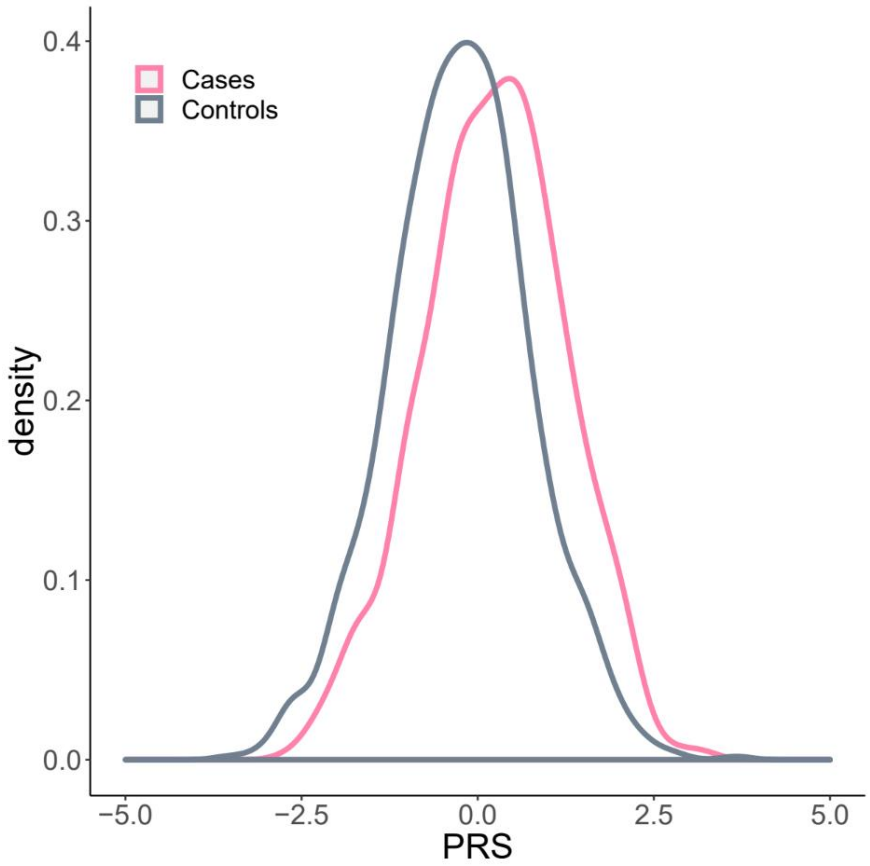



***Figure 7: Overlap of PRS between cases and controls***

*Typical density plot to show the overlap of PRS between cases and controls. In general, cases have a higher PRS than controls. However, many cases and controls have an overlapping PRS. The plot is based on simulated data.*

between PRS and some subphenotypes. (80) A younger age of onset, ileal instead of colonic CD and a higher need of bowel resections were significantly associated with an elevated PRS in patients with IBD, and confirm therefore the results of the previous study. Voskuil *et al*. (2021) studied the effects of PRS which, in contrast with Cleynen *et al* (2016) also included non-genome-wide significant SNPs, on IBD subphenotypes.(81) They replicated the results of Cleynen *et al* (2016) and further found an association between the UC-PRS and colonic disease localization in CD patients. They also observed a relation between a higher genetic risk for CD and the increased development of fibrostenotic disease and the frequency of ileocaecal resection. PRS thus seem useful to gain better insights into subphenotypes of disease and patient classification, and in differential diagnosis between CD and UC, both are interesting in context of treatment decisions.

The question then follows if PRS would also be useful to diagnose patients or predict who will develop disease. Khera *et al* (2018) aimed to identify individuals at risk for five common diseases – atrial fibrillation, breast cancer, coronary artery disease, diabetes type 2 and IBD – by applying PRS.(82) The performance of the PRS was tested in 288,978 individuals. They found that the 3.2% highest PRS group had a greater than three-fold increased risk for IBD compared to the remainder of the studied population. So being in this high-risk category does not mean a certainty to develop the disease. IBD and other complex diseases are influenced by genetic risk variants but also by environmental factors. A large percentage of heritability is still unexplained and many more variants will probably be discovered in future research. Therefore, a lot of information, environmental as well as genetic factors, are missing in PRS calculation. Furthermore, a high overlap exists between the PRS of cases and

controls (Figure 7). Predicting or diagnosing IBD using PRS might thus not be possible, however individuals at an increased risk can be distinguished. (80,82,83) While a preventive therapy for IBD is not developed yet at the time, screening all persons to detect the ones with a higher genetic risk could be useful for enrichment of clinical trial populations and/or research studies to assess novel preventive strategies.

Similarly, PRS could be helpful in prioritizing individuals for sequencing. It is hypothesized that individuals diagnosed with a disease but who have a low PRS for that disease have a higher prevalence of rare pathogenic disease variants.(84) This hypothesis was tested and found true for five common diseases, namely breast cancer, colon cancer, diabetes type 2, osteoporosis and short stature. The hypothesis held for five diseases all with slightly different underlying architecture and therefore is probably also applicable for IBD. Genotyping an individual using a SNP-array is much cheaper than sequencing. Thus, PRS calculation for IBD patients can indicate which persons improve the chance of finding a rare pathogenic variant in a cost-effective way.

## 1.4 Pathogenesis

IBD is a multifactorial or complex disease. This means that genetic as well as environmental factors have an influence on the development of IBD. The immune response, changes in gut microbiota, environmental factors and genetic variants play a role in the pathogenesis.(85,86)

### 1.4.1 From genetic variants towards understanding IBD pathogenesis

Discovering the underlying genetic architecture of IBD might improve insights into the pathogenesis. The genes associated with IBD indicate which pathways are involved. Through this method it was discovered that the innate and the adaptive immune system are both involved through several pathways, from a defective epithelial barrier to dysregulation of T- and B-cells. Moreover, discovering how a variant dysregulates the function of a gene provides invaluable information about the role of this gene or protein in the development of IBD. Many genes and their functions or dysfunctions in IBD are already clarified but the pathogenesis is still not entirely resolved.

The first gene identified to be associated with CD, *NOD2*, immediately pointed to a link with the immune system.(44) The three identified variants were all located in the LRR domain. A domain which is known to interact with bacterial lipopolysaccharides (LPS), and to inhibit the NF-κB pathway. Thus, the first discovery of an associated gene indicated that a dysregulated response to bacteria might be part of the pathogenesis.

The involvement of the immune system was further highlighted with the discovery of a protective variant in the *IL23R* gene.(50) IL23 is a proinflammatory cytokine which plays a role in the activation of effector T-cells. Furthermore, a proper response against mycobacterial infections might be mediated through IL23. Thus, here as well arises a link with the immune system and with bacteria.

The discovery of *ATG16L1* by Hampe *et al* (2007) pointed at a role for autophagy in IBD.(51) More precisely, *ATG16L1* is involved in an autophagosome pathway which handles the processing of intracellular bacteria. This gene is transcribed in the intestines and suggests that the intestinal barrier fails in IBD. Autophagy was further implicated when *IRGM* was identified.(87,88) Decreased expression or reduced function of *IRGM* would have similar consequences as for *ATG16LA*, namely persistence of intracellular bacteria.

While the first series of GWAS usually found non-synonymous coding variants as associated with disease (*NOD2, IL23R, ATG16L1*…), later GWAS mostly found associated SNPs located in non-coding

regions. The first time this occurred was when CD associated with a 250 kb region within a 1.25 Mb gene desert on chromosome 5.(52) Here, functional evidence supported the hypothesis that this region regulated the expression of nearest gene *PTGER4*. Many more associated non-coding risk variants have been detected since, and found to be causal for the association signal. The fine-mapping study of Huang *et al* (2017) indeed found that 21 of the 45 fine-mapped variants (i.e. causal with a >50% probability) were non-coding.(61) However, their exact function often remains unknown. Of the currently known 241 loci, 54 are situated in a non-coding region.

Although CD and UC are at the moment often taken together as IBD, some variants and pathways seem to be specific for one of the two subtypes. Barret *et al* (2009) discovered three genes, *HNF4A*, *CDH1* and *LAMB1*, associated with UC.(89) These three genes implicate a defective epithelial barrier function in the intestines as an important pathway in UC pathogenesis. Interestingly, CD seems not to be associated with *HNF4A* and *LAMB1*, indicating that a defective barrier function might be more important in UC than in CD.(55)

The first more systematic interrogation of pathways involved in IBD based on genetic findings was done in the landmark study by Jostins *et al* (2012).(33) Based on their genome-wide significant variants, they made a list of genes present in the associated loci prioritized based on functional annotation and gene network tools and looked for enrichment of these genes in Gene Ontology terms. Unsurprisingly, regulation of cytokine production, activation of T-cells, B-cells and Natural Killer cells, and response to molecules of bacterial origin were among the most significant results.

The latest GWAS by De Lange *et al* (2017) combined their GWAS with a fine-mapping analysis. They tried to fine-map the 25 newly and 40 previously identified loci. This resulted into two loci, *SLAMF8* and *RORC* mapped to a single variant with >99% probability of being causal. In line with the previous studies, they are also key regulators of the immune system. *SLAMF8* inhibits the migration of myeloid cells to the inflammation site and therefore downregulates inflammatory responses, and *RORC* regulates differentiation of T helper type 17 cells. Several integrin genes, that have a function in cell differentiation in inflammation, are also implicated by their location close to associated loci. Thus, more and more pathways involving the immune system are associated with IBD.

The discovery of rare variants might likewise improve insight into the pathogenesis of IBD. A sequencing study of IBD individuals found rare variants in *PRDM1* and *NDP52*.(64) An associated locus, discovered with previous GWAS, contained *PRDM1* and therefore this gene was implicated as the causal gene of that locus. The detected rare variant led to an increase in T-cell proliferation at the site of inflammation and thus enhances inflammation. *NDP52* has multiple functions in immunity. However, the loss of downregulation of the NF-κB pathways is probably the main mechanism of how this rare variant increases risk for IBD. The innate and the adaptive immune system are both involved through several pathways, from a defective epithelial barrier to dysregulation of T- and B-cells.

### 1.4.2 Environmental factors

Importantly, the underlying genetics might be interesting but the microbiome and other environmental factors, e.g. smoking and diet, should not be forgotten. Environmental factors also have an important influence. Piovani *et al* (2019) performed a large scale meta-analysis based on other meta-analyses of observational studies.(90) They divided their findings in several different categories which are associated with IBD: dietary intake and nutrients, exposure to drugs, lifestyle and hygiene, microorganisms and vaccinations, and surgeries. Not all categories have a strong association, for example only one vaccination increased the risk of IBD but this was only found in two very small studies.

Diet is currently considered the most important environmental factor. Especially, the westernized diet with its higher calorie intake, high in sugars and carbohydrates, high in saturated fats and animal proteins seems to be the culprit.(91) IBD incidence started to rise when the switch from a plant-based diet to a more animal-based diet was made. Diet mainly affects the microbiota and can increase the risk of IBD.(92) Many research is therefore focused on investigating which nutrients are emitting risk and which might be protective. Diets are also being studied as a possible treatment option.

One of the earliest known factors is smoking.(93) Current smoking has a contradictory influence on CD and UC. It increases the chance to develop CD, however it seems to be protective for UC. On the other hand, former smokers have an increased risk of UC and CD, and therefore the protective effect for UC seems to fade as someone stops smoking. The effects of smoking are probably exerted due to epigenetic alterations, immune suppression and changes in the gut microbiota.(90)

While diet and smoking are the most well-known environmental factors associated with IBD, they are certainly not the only ones. Other important factors which increase risk include the use of antibiotics, a previous appendectomy and vitamin D deficiency.(90) However, some other factors seem to have a protective effect, like an infection with Helicobacter Pylorus, physical activity or breastfeeding. Not all environmental factors are known and further research to elucidate more factors will be necessary.

# 2 Objectives/aims

Some families have many members which are affected by IBD, so-called multiplex families. Environmental as well as genetical factors are shared among (closely) related individuals. However, **the reason behind this familial clustering is unknown**. Therefore, I will study the genetics of 55 IBD multiplex families which have at least three affected first-degree relatives to find the cause of their familial aggregation. Unravelling the genetics of multiplex families might aid in risk prediction and in optimalisation of treatment for these families. However, this study has also a broader aim to further uncover the pathogenesis of IBD. Understanding the familial aggregation might provide a better insight into the overall pathogenesis of IBD.

At the moment, over 240 loci are found to be significantly associated with IBD through genome-wide association studies (GWAS).(31–33) These **common genetic risk variants** could be segregating more in multiplex families. An increase in the number of common genetic risk variants indicates an increased likelihood of developing IBD, and would therefore provide a reason for familial aggregation. **Polygenic risk scores (PRS)** accumulate the common variants, significantly and not significantly associated, to one score which reflects the genetic risk of an individual for the disease. Thus, I will use PRS to look at the burden of common genetic risk variants in these families. I will calculate PRS based on the imputed genotypes measured with Immunochip and the SNP effect sizes for IBD, CD and UC separately. PRS will not only be determined on the genome-wide significant SNPs, but higher p-value thresholds will also be considered, as probably a lot of genetic risk factors are not discovered yet. I will compare groups of affected family members, unaffected family members, sporadic cases and healthy controls to investigate whether these multiplex families have a higher burden of known genetic risk variants.

Another possibility of familial aggregation is the presence of **rare risk variants** with a higher effect. I would expect that families which have a low polygenic risk are more likely to carry such a rare variant. Thus, I will investigate the **mean PRS per family** to determine if some families have a lower value than might be expected. A mean PRS does not take into account the difference between affected and unaffected family members. One group could indeed influence the PRS and lead to distorted results. I will therefore also calculate separately the mean PRS of affected and unaffected relatives. A discrepancy between the PRS of affected and unaffected members can in addition point to (another) reason for familial aggregation. For example, if in a family the unaffected family members have a severely increased PRS in comparison with the affected family members, then the explanation for this familial aggregation is probably not a high burden of common risk variants.

The more than 240 associated loci so far are found through GWAS that are mainly based on unrelated cases and in a case-control setting. In families, other risk variants might be also important. To investigate if families have other important risk variants, I will execute a **family-based association analysis**, and an association analysis based on sporadic cases and controls. The strongest associations in both analyses will indicate if the risk variants are similar to each other or if familial IBD has some other specific risk variants.

In conclusion, I will investigate the reason behind familial aggregation by looking at the amount of common genetic risk variants through PRS. Rare variants might also be a possible explanation, especially in families which only carry few common risk variants. Thus, I will examine if some families have very few common risk variants. Furthermore, I will try to identify specific risk variants in families.

# 3 Materials and methods

## 3.1 Dataset

This study is based on two cohorts: a cohort of 55 multiplex IBD families and a sporadic case-control cohort containing 3,518 individuals. Both cohorts were recruited through the IBD unit of the University Hospital Leuven (Belgium) under the supervision of professor Séverine Vermeire, and in the framework of the IBD genetics study. Ethical approval is obtained by the Ethics Board University Hospital Leuven (study nr S53684). Multiplex families were defined here as having at least three affected first-degree relatives. The family cohort includes 337 individuals (164 CD, 32 UC, 141 Unaffected) divided over 36 CD families (125 CD, 94 Unaffected), 1 UC family (3 UC, 0 Unaffected) and 18 mixed families (39 CD, 29 UC, 47 Unaffected) (Table 2). Sporadic cases and controls have no affected relatives. The sporadic cohort includes 2,645 cases (1,705 CD, 917 UC, 23 IBD-U) and 873 healthy controls. In the analyses, the entire dataset will be divided into four groups: cases without affected relatives (sporadic cases), controls without affected relatives (healthy controls), cases within multiplex families (affected family members) and healthy first-degree relatives within multiplex families (unaffected family members)

***Table 2: Overview of the cohorts included in this study***

**Family cohort**

| | **n families** | **CD** | **UC** | **IBD-U** | **Controls** |
|---|---|---|---|---|---|
| **CD families** | 36 | 125 | | | 94 |
| **UC families** | 1 | | 3 | | 0 |
| **Mixed families** | 18 | 39 | 29 | | 47 |
| **Total** | 55 | 164 | 32 | | 141 |
| **Sporadic cohort** | | | | | |
| **Total** | | 1,705 | 917 | 23 | 873 |

*The number of individuals subdivided according to which phenotypes are present in the family. CD families only have affected family members with Crohn's diseases, UC families only have affected family members with ulcerative colitis and mixed families have both CD and UC affected family members. CD = Crohn's disease, UC = ulcerative colitis, IBD-U = IBD unclassified.*

## 3.2 Genotyping

Genotyping on both cohorts was performed previously using Immunochip (Illumina). Immunochip is a high-throughput genotyping chip based on the Illumina Infinium chip including approximately 240,000 SNPs.(94) The chip is based on GWAS of 12 autoimmune and inflammatory diseases, including CD and UC. These GWAS contributed 196,524 SNPs and small indels to the chip while approximately 25,000 SNPs from other diseases are included as control.

A subset of the sporadic cases and controls were also genotyped using the GSA chip of Illumina. GSA refers to Infinium Global Screening Array-24 Kit. This chip includes 654,027 SNPs which cover the entire genome. The GSA chip data is solely used to compare the performance between Immunochip and GSA chip.

## 3.3 Genotyping quality control and imputation

### 3.3.1 Unimputed data Immunochip

Initial quality control on the genotype data was performed for the family and sporadic dataset separately according to Jostins *et al* (2012).(33) In short, missingness per person < 0.02, heterozygosity rate within 95% interval per batch (all samples genotyped at the same time), missingness per SNP < 0.02 and Hardy-Weinberg equilibrium p-value (controls) > 1e-10. Further quality control was performed on both datasets combined before the unimputed data was analysed. Samples were not allowed to have > 0.05 missingness. Duplicate individuals were removed. Remaining SNPs after further quality control had missingness < 0.02; HWE p-value > 10e-6; MAF > 0.01. Duplicate and ambiguous SNPs were removed.

### 3.3.2 Imputed data Immunochip

Further cleaning of the dataset before imputation included the removal of insertions/deletions (indel) and ambiguous (A/T or C/G) SNPs. Imputation was done with the Michigan imputation server.(95) It should be noted that the major *NOD2* variant (rs2066847) is an indel variant and was therefore removed before imputation. Because of its importance, this SNP was reintroduced after imputation and before further analysis. Quality control after imputation included filtering per chromosome on INFO score > 0.7 and MAF > 0.01. Missingness per person and missingness per SNP was checked and no outliers were found. Duplicate individuals were removed. SNPs were further filtered as follows: Hardy-Weinberg equilibrium (all) < 10e-6; MAF < 0.01 and duplicate SNPs were excluded.

### 3.3.3 Unimputed data GSA chip

Quality control on the GSA chip genotypes was performed by the IIBDGC and according to the IIBDGC pipeline. In short, indels, monomorphic and mitochondrial variants were removed. Y-Chromosome SNPs were removed after a sex check. Individuals were removed when: missingness per individual > 0.05; sample is duplicated; heterozygosity is not within ± 4SD interval; or not of European descent. For SNPs the requirements for removal were missingness per SNP > 0.02, MAF> 0.01, Hardy-Weinberg equilibrium for controls < 1e-5, Hardy-Weinberg equilibrium for cases < 1e-12, and available in TOPMEd. Frequencies of SNPs are compared with Gnomad and TOPMEd. Only the variants of autosomal chromosomes were retained for the analyses in this study.

### 3.3.4 Imputed data GSA chip

Imputation of the GSA chip data was also performed by the IIBDGC. Duplicates, indels, monomorphic sites, data mismatching with TOPMEd and SNP call < 90% are excluded to prepare the dataset for imputation. Genotyped variants with EmpRsq < 0.5 were excluded. Imputation was performed with TOPMEd. After imputation SNPs with a Hardy-Weinberg equilibrium < 1e-5 for controls and < 1e-12 for cases were excluded. Only the variants of autosomal chromosomes were retained for the analyses in this study.

## 3.4 Principal component analysis

The 1000 genomes (1000G) dataset was used as a reference to determine the ancestry of the individuals in our Immunochip dataset. This dataset includes 261 European, 177 Asian, 22 American and 169 African individuals. I applied quality control on 1000G (missingness per person < 0.02, missingness per SNP < 0.02 and MAF > 0.01) and then merged the data with our dataset. Principal

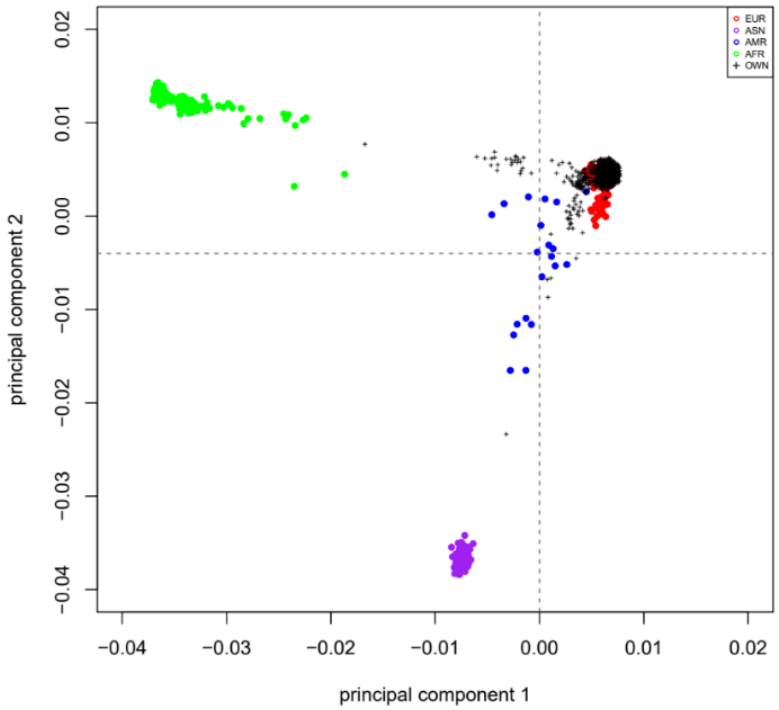



***Figure 8: PCA plot of the imputed dataset***

*Dots represent the individuals of the 1000G dataset and black crosses the individuals of the imputed sporadic and familial dataset genotyped on Immunochip. Every ancestry is shown by a different colour: red = European; blue = American; purple = Asian and green = African. The cut-off values of having a European ancestry are displayed by the dotted vertical (x= 0.0) and horizontal (y=-0.004) lines.*

components were calculated using plink v1.9 and plotted.(96) Based on visual inspection of the PCA plot (Figure 8), individuals and families with a non-European ancestry were excluded, including 20 individuals (20 cases) of the sporadic dataset and two families (family 1: 3 affected + 1 unaffected; family 2: 4 affected + 2 unaffected). After exclusion of non-European individuals, principal components were recalculated on our dataset to be included in the logistic regression analyses (see below).

## 3.5 Polygenic risk score analysis

### 3.5.1 PRS calculation

Polygenic risk scores (PRS) were calculated using PRSice 2.3.3 for different p-value thresholds (pTs) and the phenotypes IBD, CD and UC (referred to as PRS IBD, PRS CD and PRS UC) on the imputed genotypes derived from Immunochip.(97) The predefined pTs include SNPs which are genome-wide significantly associated (pT = 5e-8), genome-wide suggestively associated (pT = 1e-5) or less significantly associated (pT = 0.01, 0.05, 0.1 and 0.5). The number of SNPs are shown in Table 3. The pT of the PRS with the best goodness-of-fit (best pT) is calculated with PRSice 2.3.3 with inclusion of the principal components as covariates. The effect sizes and p-values of SNPs for IBD, CD and UC are derived from the basefiles of de Lange *et al*. (2017).(31) SNP ids were updated to match our dataset. Ambiguous SNPs were removed. The European subset of the 1000G dataset (n = 503) was used as reference to calculate linkage disequilibrium. Clumping parameters were distance to both ends from the index SNP = 250 kb, $r^2$ = 0.1 and p-value threshold = 1.

The formula used to calculate the PRS is: (–score avg in PRSice)

$$PRS_j = \sum_i \frac{S_i \times G_{ij}}{M_j}$$

$S_i$ is the effect size of the ith allele; $G_{ij}$ is the times that the ith allele is present in the jth individual; and $M_j$ is the total number of alleles included in the PRS calculation of the jth individual. (98)

Scores are z-score standardized for each phenotype per pT. The standardized PRS are further used for statistical analyses.

***Table 3: Number of SNPs per PRS based on Immunochip***

| | PRS IBD | | PRS CD | | PRS UC | |
|---|---|---|---|---|---|---|
| **pT** | **Imputed** | **Unimputed** | **Imputed** | **Unimputed** | **Imputed** | **Unimputed** |
| **5e-8** | 256 | 193 | 205 | 167 | 145 | 116 |
| **1e-5** | 537 | 407 | 455 | 350 | 337 | 253 |
| **0.01** | 3442 | 2,186 | 3,196 | 2,034 | 2,660 | 1,624 |
| **0.05** | 7,072 | 4,345 | 6,705 | 4,146 | 6,151 | 3,631 |
| **0.1** | 10,063 | 6,131 | 9,563 | 5,853 | 9,040 | 5,399 |
| **0.5** | 23,612 | 15,189 | 23,282 | 15,000 | 23,176 | 14,755 |

*The number of SNPs for all different PRS (PRS IBD, PRS CD and PRS UC) based on imputed and unimputed Immunochip data are provided for the fixed p-value thresholds (pT).*

### 3.5.2 Statistical analysis

#### *3.5.2.1 Correlation analysis between different phenotypes*

Normality of all PRS per pT and phenotype was tested using QQ-plots, density plots and Shapiro tests. Not all data were normally distributed and therefore Spearman correlations were used. Spearman correlations were calculated between PRS based on IBD effect sizes vs PRS based on CD effect sizes, PRS UC vs PRS IBD, and PRS CD vs PRS UC. Normality tests and correlation calculation were performed with R3.5.1. Correlation was considered significant as $p < 0.05/18 = 2.77e\text{-}3$.

#### *3.5.2.2 Association analysis*

PRS of four groups (affected family members, unaffected family members, sporadic cases, and healthy controls) were compared with logistic regression using R3.5.1. The first five principal components based on the individuals included in the analysis were included as covariates to correct for population substructure. This analysis was repeated for all pTs and phenotypes. Groups were considered significantly different if $p < 0.05/36 = 1.39e\text{-}3$. Odds ratios and 95%-confidence intervals were calculated based on the output of the logistic regression. The pseudo-$R^2$ for the PRS was used as the goodness-of-fit parameter and calculated according to following formula:

*Pseudo-$R^2$ PRS = Pseudo-$R^2$ (Nagelkerke) full model – pseudo-$R^2$ (Nagelkerke) null model*

The $R^2$ PRS of the best pT is calculated with PRSice 1.9 which uses the same formula. In the analyses of CD, only the CD cases – sporadic cases and affected family members – and the unaffected family members of families with at least one CD case were included. The same principle was applied for the UC PRS.

To exclude a bias of my results due to dependence of family members, this association analysis was repeated using the mean PRS of unaffected and affected family members per family instead (see also below). This analysis was not done with individuals, thus principal components could not be included here.

#### *3.5.2.3 Quantile analysis*

A quantile analysis was performed in R3.5.1. The entire dataset (familial + sporadic) and the specific datasets separately were divided into five equal quantiles based on the PRS. Each quantile is compared to the first quantile with logistic regression. The odds ratio and 95% confidence interval were calculated based on the output of the logistic regression, and this data was plotted. Quantiles are significantly different when $p < 0.05/16 = 3.13e\text{-}3$.

#### *3.5.2.4 Mean PRS per family*

R3.5.1 was used to calculate the mean PRS per family (family PRS), the mean PRS of all unaffected family members per family and all affected family members per family. A 'low PRS family' is considered as a family with a mean PRS below the mean of all unrelated healthy controls. If a family is above the threshold of the mean of all affected family members the family is labelled a 'high PRS family'.

#### *3.5.2.5 Comparison Immunochip and GSA chip*

All analyses are executed on both the unimputed and imputed datasets of both genotyping chips. The number of SNPs included in PRS IBD, PRS CD and PRS UC are presented in Table 4. Normality of all PRS per pT and genotyping chip was tested using QQ-plots, density plots and Shapiro tests in R3.5.1. Some data was not normally distributed and therefore Spearman correlations were used. Correlations were calculated on PRS IBD, PRS CD or PRS UC for each pT between the two genotyping chips. Correlation was considered significant as $p < 0.05/6 = 8.33e\text{-}3$. The overlap of the 1% highest and lowest and the 10% highest and lowest PRS is calculated with R3.5.1.

A logistic regression analysis between cases and controls is performed on all individuals for which both Immunochip and GSA chip genotypes were available. Principal components, calculated with PRSice 1.9, are included to control for population substructure. The pseudo-$R^2$ for PRS was calculated as mentioned above. The PRS of cases and controls were considered significantly different when $p < 0.05/6 = 8.33e\text{-}3$.

***Table 4: Number of SNPs per PRS based on GSA data***

| | PRS IBD | | PRS CD | | PRS UC | |
|---|---|---|---|---|---|---|
| **pT** | **Imputed** | **Unimputed** | **Imputed** | **Unimputed** | **Imputed** | **Unimputed** |
| **$5^e$-8** | 291 | 184 | 227 | 153 | 152 | 117 |
| **$1^e$-5** | 757 | 392 | 640 | 339 | 461 | 255 |
| **0.01** | 23,332 | 6,907 | 21,302 | 6,219 | 20,197 | 5,700 |
| **0.05** | 68,632 | 21,323 | 65,478 | 2,0243 | 63,217 | 19,279 |
| **0.1** | 109,149 | 35,381 | 105,351 | 34,131 | 103,141 | 328,86 |
| **0.5** | 291,460 | 111,204 | 289,479 | 110,223 | 288,703 | 109,417 |

*The number of SNPs for all different PRS (PRS IBD, PRS CD and PRS UC) based on imputed and unimputed GSA data are provided for the fixed p-value thresholds (pT).*

## 3.6 Family-based association analysis

A generalized mixed model association was performed with SAIGE (Scalable and Accurate Implementation of Generalized mixed model) on the familial and sporadic dataset separately.(99) A genetic relationship matrix (GRM) was included as covariate. To calculate the GRM, only variants with MAF > 0.01 are included. The presence or absence of IBD was used as a binary trait to fit the model. Single variant association tests were performed on genotypes for each variant in the dataset with MAF > 0.0001 and minimal allele count (MAC) > 1. Results were visualized using Manhattan plots. In addition, the association analysis results were clumped with plink 1.9 with following parameters: distance to both sides of the index SNP = 250kb, $r^2 = 0.5$, p-value (index SNPs) < 1e-4, p-value (clumped

SNPs) < 0.01. The top SNPs with a p < 1e-4 are presented and discussed. Top SNPs were annotated with Annovar.(100)

## 3.7 Plots

All plots except the Manhattan plots were made with the package ggpubr in R3.5.1. The Manhattan plots were made with the R-package qqman.

# 4 Results

## 4.1 The CD and UC PRS do not correlate well

IBD encompasses CD and UC, thus it might be expected that the PRS of IBD and CD, and IBD and UC are correlated. I therefore calculated PRS for each (sub)type (PRS IBD, PRS CD, and PRS UC) with different p-value thresholds (pT), and performed a spearman correlation analysis. The number of SNPs for each PRS can be found in table 3. When only the genome-wide significant SNPs are considered, the correlation (R) between PRS IBD and PRS CD is 0.68 (Figure 9, pink line). The correlation increases with the inclusion of additional SNPs reaching the genome-wide suggestive threshold into the score (R = 0.72, p < 2.2e-16 ), and also with the inclusion of less significant SNPs with a pT of 0.01 (R = 0.75, p < 2.2e-16). A further elevation in the number of SNPs however does not improve the correlation, with the exception of the inclusion of all SNPs below the threshold of 0.5 (R = 0.78, p < 2.2e-16). A correlation between PRS IBD and PRS UC can also be seen (Figure 9, blue line). The correlation slightly increases between the PRS based on genome-wide significant (R = 0.61, p < 2.2e-16) and genome-wide suggestive SNPs (R = 0.63, p < 2.2e-16). A larger difference in correlation is observed when moving to the next pT of 0.01 (R = 0.69, p < 2.2e-16). PRS which contain more SNPs increase the correlation between IBD and UC only slightly (pT 0.05: R = 0.73, p < 2.2e-16; pT 0.1: R = 0.74, p < 2.2e-16; pT 0.5: R = 0.76, p < 2.2e-16). All correlations are highly significant, thus a high correlation between IBD and its two subtypes exists as was expected.

While both PRS CD and PRS UC are highly correlated to PRS IBD, their correlation to each other is much lower. At a genome-wide significant p-value threshold the correlation is again lowest (R = 0.12, p = 4.53e-13) (Figure 9, grey line), while the PRS based on genome-wide suggestive SNPs has a slightly

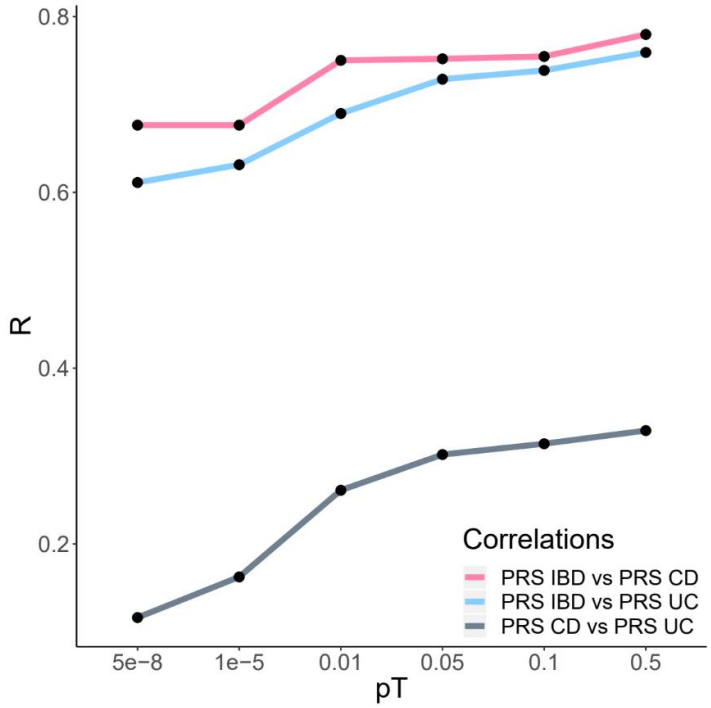



***Figure 9: Correlations of PRS for different phenotypes (CD, UC and IBD)***

*The spearman correlation (y-axis) is depicted for each threshold (x-axis). Every colour represents the correlation between PRS of two phenotypes: pink = IBD vs CD, blue = IBD vs UC and grey = CD vs UC. PRS were calculated based IBD effect sizes (PRS IBD), CD effects sizes (PRS CD) or UC effect sizes (PRS UC) and on SNPs with MAF > 0.01.*

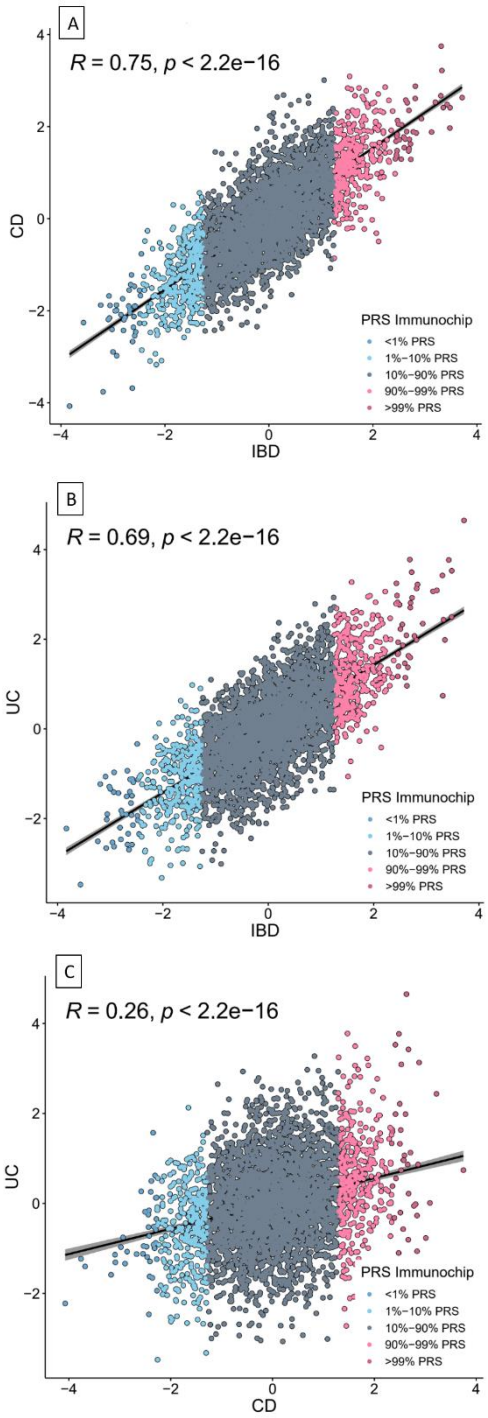



***Figure 10: Correlation plots for PRS IBD, PRS CD and PRS UC***

*Each dot represents the PRS of phenotype 1 (x-axis) and phenotype 2 (y-axis). (A) PRS IBD vs PRS CD, (B) PRS IBD vs PRS UC and (C) PRS CD vs PRS UC. Dark and light blue depicts, respectively, the 1% and 10% lowest PRS of IBD (A, B) or of CD (C). Dark and light pink depicts, respectively, the 1% and 10% highest PRS of IBD (A, B) or of CD (C). The spearman correlation is presented at the left upper corner of the plot. PRS were calculated based on pT = 0.01 and SNPs with MAF > 0.01*

better correlation (R = 0.16, p < 2.2e-16). Including the SNPs with a p-value below the 0.01 threshold highly improves the correlation (R = 0.26, p < 2.2e-16). However, the correlation does not increase much more with higher pTs (pT 0.05: R = 0.30, p < 2.2e-16; pT 0.1: R = 0.31, p < 2.2e-16; pT 0.5: R = 0.33, p < 2.2e-16). All correlations are also here highly significant and thus PRS CD and PRS UC are correlated with each other, albeit to a lesser extent than their correlation to PRS IBD.

Looking at which individuals fall in the extreme tails of the different PRS can provide additional information on how the different PRS match. Therefore, I compared the individuals with the top 1% and 10% highest and lowest PRS IBD to the PRS CD and PRS UC of the same individuals. Visually, the extremely low and high PRS CD do not deviate much from the PRS IBD, which would indicate that even the extreme values are correlating well with each other (Figure 10A and B). Respectively, 12 (30.77%) and 209 (54.15%) individuals of the top lowest 1% and lowest 10% of PRS CD are also in the top lowest 1% and 10% of PRS IBD. The number of overlapping individuals between PRS IBD and PRS CD in the top highest 1% and 10% are 13 (33.33%) and 195 (50.52%), respectively. The extreme scores seem on the plot to correspond a bit less between PRS IBD and PRS UC (Figure 10B). PRS UC has 11 (28.21%) and 176 (45.60%) individuals in the lowest 1% and 10% which are also present in these categories of PRS IBD. The top highest 1% of PRS UC has 17 (43.59%) individuals which are overlapping with PRS IBD. However, 192 (49.47%) individuals are both found in the top highest 10% of PRS UC and PRS IBD.

I also investigated how the extreme PRS individuals compared between PRS CD and PRS UC. The overlap here is much lower than for the comparison of PRS IBD vs PRS CD and PRS IBD vs PRS UC, which can also be seen in the more scattered display of the individuals with an extreme PRS value (Figure 10C). Only one (2.56%) person can both be found in the lowest 1% of PRS CD and PRS UC. In the 1% highest PRS are more persons overlapping and 6 (15.38%) individuals belong to highest 1% of both PRS CD and PRS UC. 86 and 90 individuals of the top 10% lowest and highest, respectively, PRS CD are also present in the top 10% lowest and highest PRS UC. An individual which has an extremely high or low PRS CD has not necessarily a corresponding PRS UC or vice versa.

Based on the previous results, any further analyses will be performed with PRS IBD. PRS IBD has a high correlation with both PRS CD and PRS UC. Furthermore, using the PRS IBD allows me to use the entire dataset instead of a subset. However, most analyses are also performed with PRS CD and PRS UC on the appropriate individuals and the results will be added as supplementary data.

## 4.2 The variability of IBD is better explained by PRS which include non-genome-wide significant SNPs

For each p-value threshold, I calculated how good the PRS IBD can distinguish (variance explained) between groups (sporadic cases, healthy population controls, familial cases and familial controls) as represented by the pseudo-$R^2$ - specific for the PRS. I corrected for population substructure with principal components. I also determined which PRS (i.e. with which p-value threshold) could best separate between the groups compared. I performed the same analyses for PRS CD and PRS UC, and results are added to the supplementary data (Supplementary figure 1 and 2)

First I checked how good sporadic cases can be separated from healthy population controls according to the PRS IBD. The PRS with the highest $R^2$ is constructed from SNPs with p-value < 1.45e-3 in the original GWAS (p = 1.34e-66, Figure 11A). This PRS can explain 14% of the variance between having IBD or not. It should be noted that all PRS using different p-value thresholds give a similar $R^2$ (pT 1e-5: $R^2$ = 0.12, p= 4.66e-59; pT 0.01: $R^2$ = 0.13, p= 2.11e-64; pT 0.05: $R^2$ = 0.13, p= 7.13e-63: pT 0.1: $R^2$ = 0.12, p= 1.04e-59; pT 0.5: $R^2$ = 0.12, p= 3.69e-57), although the PRS with only the genome-wide significant SNPs performs the worst to differentiate between IBD patients and controls ($R^2$ = 0.11, p= 8.58e-56).

I next looked at the difference in PRS between familial cases and their healthy first-degree relatives. Overall, the $R^2$ of these PRS is lower than in the sporadic dataset, indicating that the PRS explains less of the variance between these two groups (Figure 11B). However, the results are following largely the same trend as in the sporadic cases vs controls. All different p-value thresholds, except for the genome-

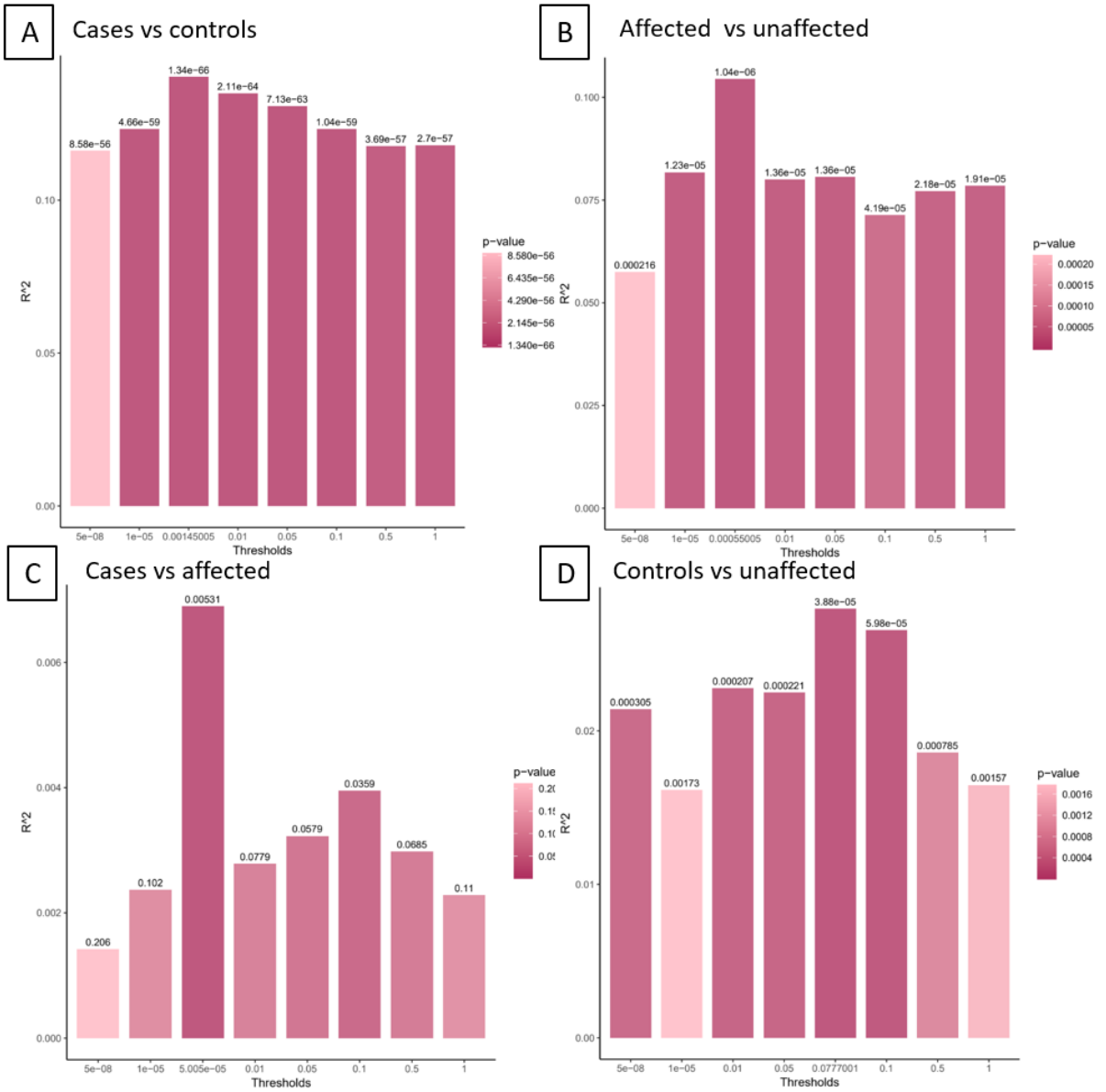



***Figure 11: Variance explained by each PRS***

*Each plot represents a different comparison of the PRS between two groups: (A) Sporadic cases vs sporadic controls; (B) Affected vs unaffected family members; (C) Sporadic cases vs affected family members; (D) Sporadic controls vs unaffected family members. Each bar depicts a separate PRS including SNPs based on different p-value thresholds. The height of the bars indicates the $R^2$ of the PRS in a logistic regression model. The p-value of the $R^2$ is represented by the colour of the bars, a darker colour indicates a more significant p-value. PRS were calculated based on the effect sizes of IBD and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

wide significant threshold, have a similar goodness-of-fit (pT 1e-5: $R^2$ = 0.08, p= 1.23e-5; pT 0.01: $R^2$ = 0.08, p = 1.36e-5; pT 0.05: $R^2$ = 0.08, p = 1.36e-5: pT 0.1: $R^2$ = 0.07, p = 4.19e-05; pT 0.5: $R^2$ = 0.08, p = 2.18e-5).The genome-wide significant PRS is also here the model that performs worst ($R^2$ = 0.06, p= 2.16e-4). On the other hand, the best model (pT = 5.50e-4) can explain 10% of the variance (p = 1.04e-6).

Differentiating between sporadic cases and cases with a familial IBD background is more difficult based on the PRS. The $R^2$ are very low for all p-value thresholds (pT 5e-8: $R^2$ = 1.42e-3, p = 0.20; pT 1e-5: $R^2$ =

2.37e-3, p = 0.10; pT 0.01: $R^2$ = 2.78e-3, p = 7.79e-2; pT 0.05: $R^2$ = 3.23e-3, p = 5.79e-2: pT 0.1: $R^2$ = 3.95e-3, p = 3.59e-2; pT 0.5: $R^2$ = 2.98e-3, p = 6.84e-2) ; even the best model (pT = 5.01e-5) only accounts for a $R^2$ of 0.69% (Figure 11C). Moreover, no PRS has a significant p-value (p < 1.39e-3). While some pTs are significant before multiple testing correction, not even the best PRS (pT = 5.01e-5, p = 5.31e-3) remains significant after correction.

Lastly, I compared healthy population controls with the unaffected family members. Interestingly, all the different threshold PRS were significant (pT 5e-8: $R^2$ = 2.14e-2, p= 3.05e-4; pT 1e-5: $R^2$ = 1.61e-2, p = 1.73e-3; pT 0.01: $R^2$ = 2.28e-2, p = 2.07e-4; pT 0.05: $R^2$ = 2.25e-2, p= 2.21e-4: pT 0.1: $R^2$ = 2.66e-2, p = 5.98e-5; pT 0.5: $R^2$ = 1.86e-2, p = 17.34e-4), indicating that PRS of healthy individuals in a multiplex family are different from PRS of healthy population controls (Figure 11D). The $R^2$ for all thresholds is around 2%, thus PRS do not explain a lot of variance. Here, the best model uses SNPs with a pT = 7.77e-2 (R = 2.79e-2, p = 3.88e-5).

Further analyses in this paper will be illustrated with pT = 0.01. This PRS had the best goodness-of-fit in the model which analysed sporadic cases vs controls, the largest dataset. In the familial cases vs first-degree relatives, this threshold is also one of the PRS which explains the most variance. Other thresholds will be mentioned if they provide additional information.

## 4.3 Affected family members do not have a higher PRS than sporadic cases

To further zoom in on how PRS differ between the four groups (sporadic cases, healthy controls, familial cases and familial controls), I compared all groups with logistic regression including principal components to correct for population stratification. For pT = 0.01, sporadic cases have significantly higher PRS than healthy population controls (p = 2.11e-64, OR = 0.44 [0.40, 0.48], Figure 12). This same observation of an increased PRS in cases is seen between affected and unaffected first-degree relatives (p = 1.36e-5, OR = 0.55 [0.41, 0.72]). Healthy controls overall have the lowest PRS. The PRS of the controls is significantly lower than those of familial cases (p = 8.04e-21, OR = 0.39 [0.32, 0.47]) as well as significantly lower than of familial controls (p = 2.07e-4, OR = 1.48 [1.21, 1.82]). However, the familial controls do not reach the high values of the PRS of the sporadic cases (p = 7.81e-5, OR = 0.66 [0.54, 0.81]). The unaffected family members thus have a PRS in between sporadic cases and controls. Remarkably, the familial cases are not significantly different from the sporadic cases (p = 7.79e-2, OR = 0.86 [0.72, 1.02]) indicating a similar PRS independent of familial history. These comparisons hold true for other p-value thresholds (Supplementary table 1). The analysis with PRS CD for CD and mixed families indicates the same results, however the analysis with PRS UC for UC and mixed families has also no significance difference between affected and unaffected family members, and unaffected family members and healthy controls (Supplementary table 2 and 3).

The multiplex families have for each family at least three affected first-degree relatives included and sometimes also multiple unaffected family members. These individuals are related and have therefore also more shared genetics. Thus, PRS might be more similar in related individuals and this in turn might influence the comparative analysis. I calculated the mean PRS of affected and unaffected family members per family and reanalysed the data (Figure 13A). No principal components are included in these logistic regression models because the PRS are not corresponding with one individual anymore, and this influences the p-values. For example, the data of sporadic cases and controls has not changed and is thus still highly significant, however the p-value is slightly different (p = 7.78e-64). The familial cases still have a significant higher PRS than their unaffected family members (p = 1.28e-3, OR = 0.33 [0.16, 0.62]) and the unrelated controls (p = 8.80e-8, OR = 0.45 [0.33, 0.60]). The familial controls also still position themselves in between the sporadic cases (p = 2.29e-3, OR = 0.61 [0.45, 0.84]) and controls (p = 9.42e-2, OR = 1.31 [0.95, 1.78]), however they are after correction for multiple testing both not

significantly different anymore. The affected family members are also not significantly different from sporadic cases regardless of the p-value threshold (p = 0.61, OR = 0.92 [0.70, 1.24], Supplementary table 4). Thus, this sensitivity analysis still shows a significant difference between affected and unaffected family members, and between affected family members and population controls. The sensitivity analyses for PRS CD and PRS UC shows only a significant difference between the PRS CD of the affected members of CD and mixed families and heatlhy controls (Supplementary table 5 and 6).

I also calculated PRS on the unimputed dataset based on the same effect sizes of the SNPs and for the same p-value thresholds to test if the imputation had an impact on the results. I performed logistic regression on the PRS of the four groups. Overall, the results are the same as the imputed dataset (Figure 13B). The healthy controls are still significantly lower than sporadic (p = 8.40e-66, OR = 1.39 [1.13, 1.71]) and familial cases (p = 3.65e-19, OR = 0.41 [0.33, 0.50]). Familial controls remain in their position between the sporadic controls (p = 2.00e-3, OR = 1.38 [1.13, 1.71]) and the sporadic cases (p = 9.30e-6, OR = 0.64 [0.52, 0.78]) and have thus an intermediate PRS. Although the difference between familial and sporadic controls is not significant for pT = 0.01, higher p-value thresholds (pT = 0.05: OR = 1.41 [1.15, 1.73], p = 1.13e-3; pT = 0.1: OR = 1.48 [1.20, 1.82], p = 2.19e-4 and pT = 0.5, OR = 1.47 [1.20, 1.81], p = 2.54e-4) are. Within families, the affected members still have a higher PRS than the unaffected members (p = 2.92e-5, OR = 0.57 [0.43, 0.73]). There is no indication of difference between sporadic and familial cases, for any of the pTs (pT = 0.01: p = 0.40, OR = 0.93 [0.79, 1.09], Supplementary table 7). PRS for specific and mixed families are provided in Supplementary tables 8 and 9.

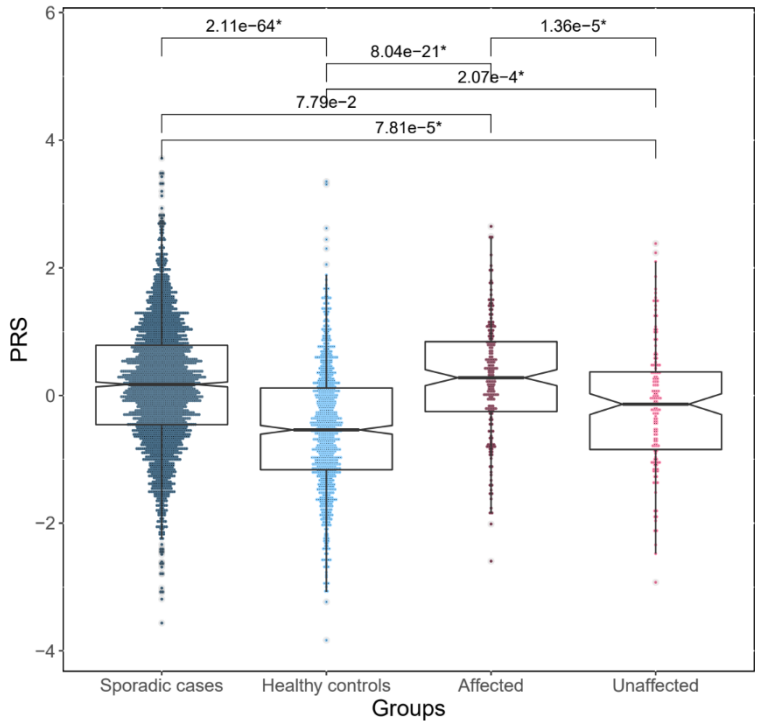



***Figure 12: Distribution plot of PRS of sporadic cases, healthy controls, affected and unaffected family members***

*The distribution of the PRS is shown, broken down into four groups: sporadic cases, healthy controls, affected family members (Affected) and unaffected family members (Unaffected). The PRS of each individual is depicted by a dot. The boxplot indicates the median and the whiskers extend to 1.5 times the interquartile range. P-values indicated are the p-values of PRS in logistic regression. PRS were calculated based on the effect sizes of IBD, pT = 0.01 and MAF = 0.01. * indicates significant p-values (p < 1.39e-3).*

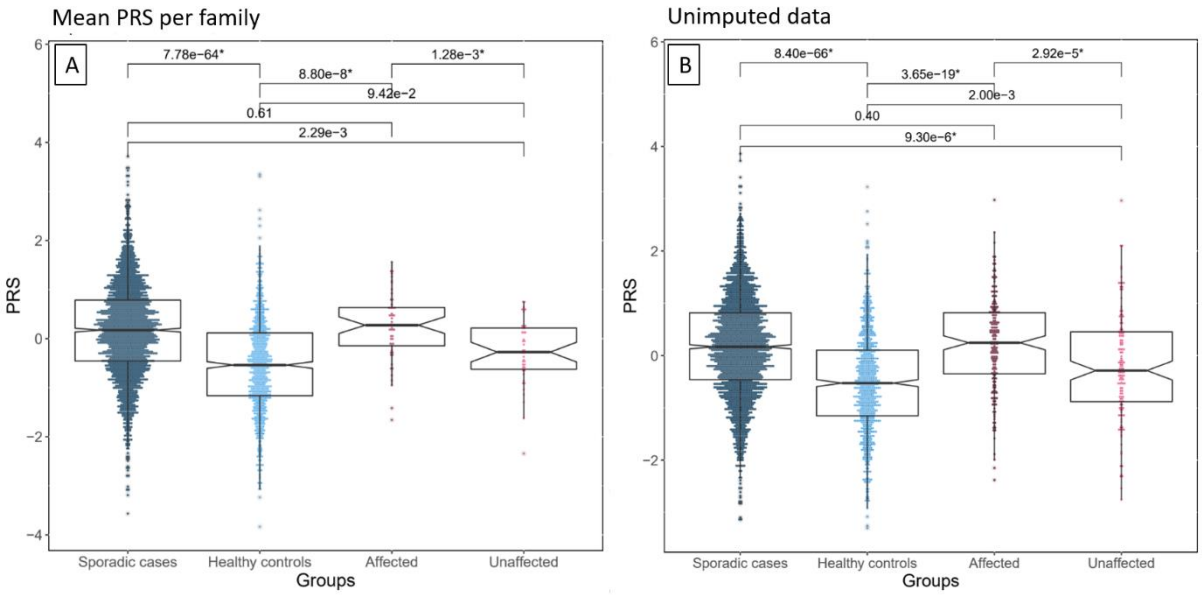


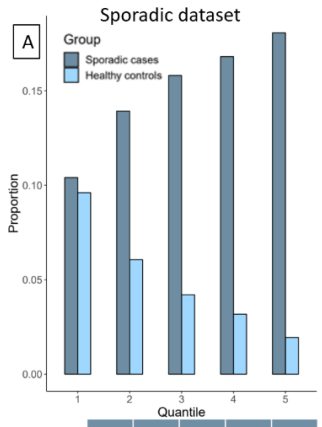

***Figure 13: Distribution plots of PRS: mean PRS per family and unimputed data***

*The distribution of the PRS is shown, broken down into four groups: sporadic cases, healthy controls, affected family members (Affected) and unaffected family members (Unaffected). (A) the affected and unaffected family members are represented by one mean PRS. (B) PRS are calculated on the unimputed dataset. The PRS of each (representative) individual is depicted by a dot. The boxplot indicates the median and the whiskers extend to 1.5 times the interquartile range. P-values indicated are the p-values of PRS in logistic regression. PRS were calculated based on the effect sizes of IBD, pT = 0.01 and MAF = 0.01. * indicates significant p-values (p < 1.39e-3).*

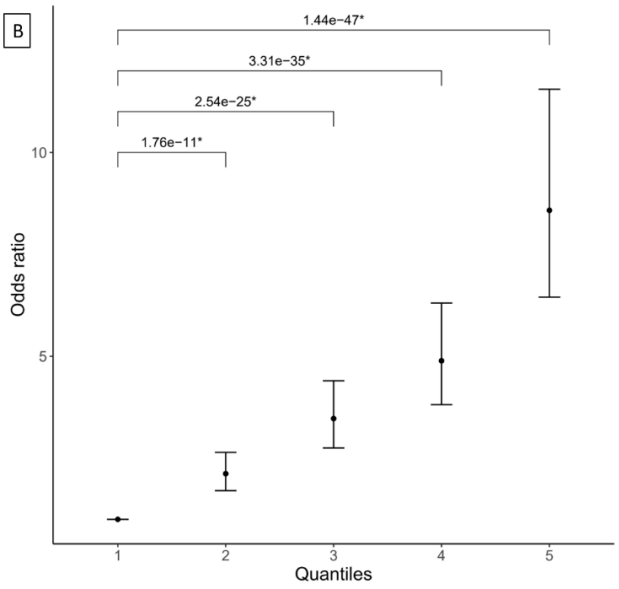
## 4.4 Individuals with a higher PRS have a higher chance to develop IBD

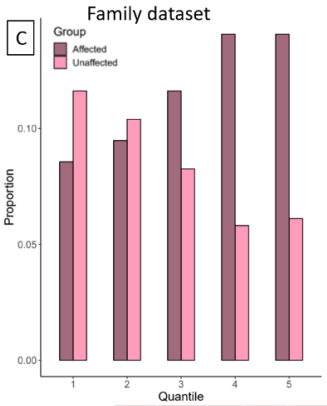
I first divided the sporadic and the family dataset separately into five quantiles. In the sporadic dataset (Figure 14A and B), the number of cases compared to controls increased in each higher quantile, as can also be seen in the increasing odds ratios (Q1vsQ2: OR = 2.12 [1.70, 2.64], p = 1.76e-11; Q1vsQ3: OR = 3.47 [2.75, 4.40], p = 2.54e-25; Q1vsQ4: OR = 4.89 [3.81, 6.30], p = 3.31e-35). Of note, more cases than controls are present in the dataset and this can be seen in the lowest quantile which also contains more cases than controls. The chances of developing the disease is 8.58 [6.45, 11.55] times higher when belonging to the highest quantile than when belonging to the lowest quantile (p = 1.44e-47).

To test if this is also the case in families, I performed the same analysis in the family dataset (Figure 14C and D). Although the dataset is approximately a tenth of the size of the sporadic dataset, the same trends can be seen. The lower quantiles include more unaffected individuals than the higher quantiles, and the reverse can be seen for affected family members. There is however no significant difference between the first quantile and Q2 or Q3 (Q1vsQ2; OR = 1.24 [0.62, 2.47] p = 0.64; Q1vsQ3: OR = 1.91 [0.96, 3.86], p = 6.76e-2). All other quantiles are significantly different from quantile one (Q1vsQ4: OR = 3.29 [1.61, 6.88], p = 1.28e-3; Q1vsQ5: OR = 3.12 [1.54, 6.49], p = 1.86e-3), although the odds ratios seem not to be increasing with every quantile.

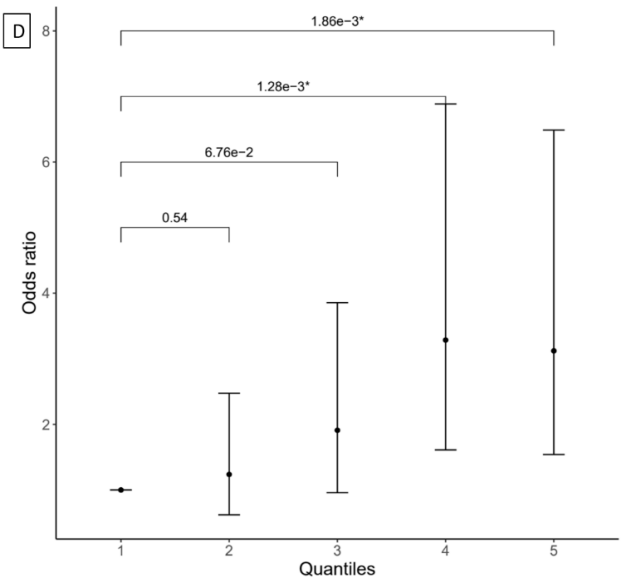


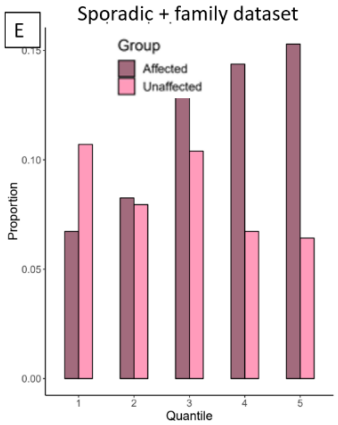

| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Cases | 364 | 487 | 553 | 588 | 632 |
| Controls | 336 | 212 | 147 | 111 | 68 |




| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Affected | 28 | 31 | 38 | 46 | 46 |
| Unaffected | 38 | 34 | 27 | 19 | 20 |




| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Cases | 369 | 500 | 540 | 588 | 627 |
| Controls | 339 | 212 | 148 | 108 | 67 |
| Affected | 22 | 27 | 43 | 47 | 50 |
| Unaffected | 35 | 26 | 34 | 22 | 21 |

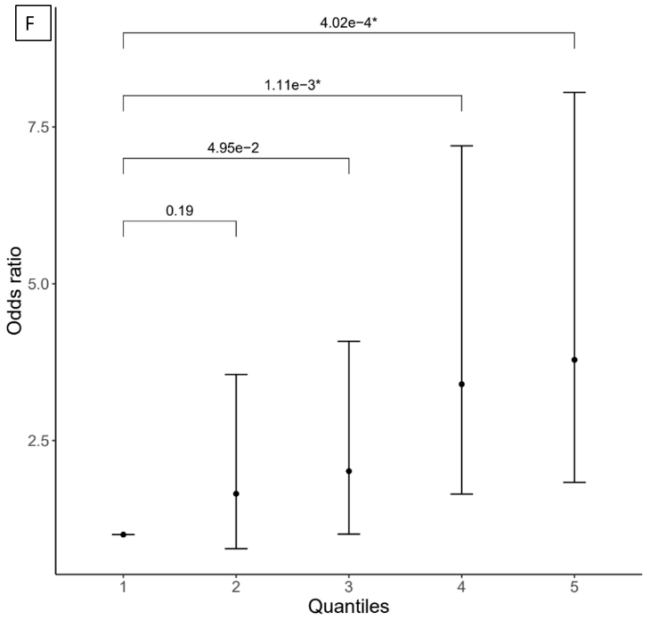


***Figure 14 Proportion plots and odds ratios of the quantile analysis***

*PRS are divided into five quantiles. (A, C, E) Bar plots which show the proportion of cases (darker colour) and controls (lighter colour) per quantile for the sporadic dataset (A), family dataset (C) and the family dataset based on quantiles of both datasets combined (E). The tables under the bar plots provide the actual number of individuals per group and quantile. (B, D, F) Plots which depict the odds ratios (y-axis) per quantile (x-axis) in comparison with the lowest quantile (Q1) for the sporadic dataset (B), family dataset (D) and family dataset based on quantiles of both datasets combined (F). Lines indicate the 95%-confidence interval. P-values are calculated using logistic regression analysis. PRS were calculated based on the effect sizes of IBD, pT = 0.01 and MAF = 0.01. Significant p-values (p < 3.13e-3) are marked with *.*

Lastly, the two datasets combined were divided into five quantiles and then the family data was taken out for further analysis (Figure 14E and F). This allows to place the PRS of family members in the context of a larger set of sporadic cases and controls. Each quantile then does not necessarily contain exactly 20% of familial individuals. Interestingly, the proportion of each group seemed to be distorted. The increase in number of cases with increasing quantile is still seen, however the higher quantiles seem to have more cases when compared to the quantile analysis based on family data only. The higher quantiles, especially Q3, have also more unaffected relatives in comparison with the family data only. Thus, more family members are found in the higher quantiles when they are divided according to the entire dataset. The odds ratios calculated for Q2 and Q3 in comparison with Q1 are not significantly different (Q1vsQ2: p = 0.19; Q1vsQ3: p = 4.95e-2) The higher quantiles have significantly different odds ratios (Q1vsQ4: p = 1.11e-3; Q1vsQ5: p = 4.02e-4). The odds ratios are increasing per quantile, although they have very wide 95% confidence intervals (Q1vsQ2: OR = 1.65 [0.78, 3.55]; Q1vsQ3: OR = 2.01 [1.01, 4.08]; Q1vsQ4: OR = 3.40 [1.65, 7.20]; Q1vsQ5: OR = 3.79 [1.83, 8.05]).

## 4.5 Some families have an extremely low PRS

A high burden of common variants, reflected by a high PRS, could be a cause of familial aggregation of IBD. However, a lot of heterogeneity between families might exist, as can also be expected from Figure 12 where a lot of variability is visible. I therefore computed the mean PRS for each family, adding the PRS of all family members, affected and unaffected, and dividing by the number of individuals in that family. When plotting the mean PRS per family in an ascending order, a slight flattening of the curve can be observed in the middle (Figure 15A), meaning that many families have a similar PRS. However, some families are having an aberrantly high or low PRS. Many families, 24 out of 55 families have a very high family PRS, defined as having a PRS above the mean of sporadic cases. A very low family PRS is characterized here as a PRS lower than the mean PRS of unrelated healthy individuals. At a pT of 0.01, seven families (13%) are below this threshold. Two families are even completely separated from this group due to an extremely low PRS.

The results shown are for the PRS IBD, and thus based on IBD effect sizes. As some families are CD-only or UC-only families, results might be different for PRS CD or PRS UC (also see part 4.2). The two families with the lowest PRS IBD, a CD-only and UC-only family respectively, continue to have the lowest scores for their particular subtype (Figure 15B and C). Moreover, the families which dive under the control threshold in the PRS CD and PRS UC also belonged to the group of very low PRS IBD, or were just slightly above the threshold.

The mean family PRS takes together the affected and unaffected family members. However, each family has a different composition and the number of affected and unaffected individuals can have an influence on the mean family PRS. Per family, a separate mean PRS for the affected and unaffected members was calculated and compared. As might have been expected from the PRS comparison analysis above, the PRS of affected individuals in general is higher than the PRS of unaffected members of the same family (Figure 16). In some families this relationship however is reversed, with the familial controls having a higher PRS than the familial cases (Figure 16, pink). A few families show a similar PRS

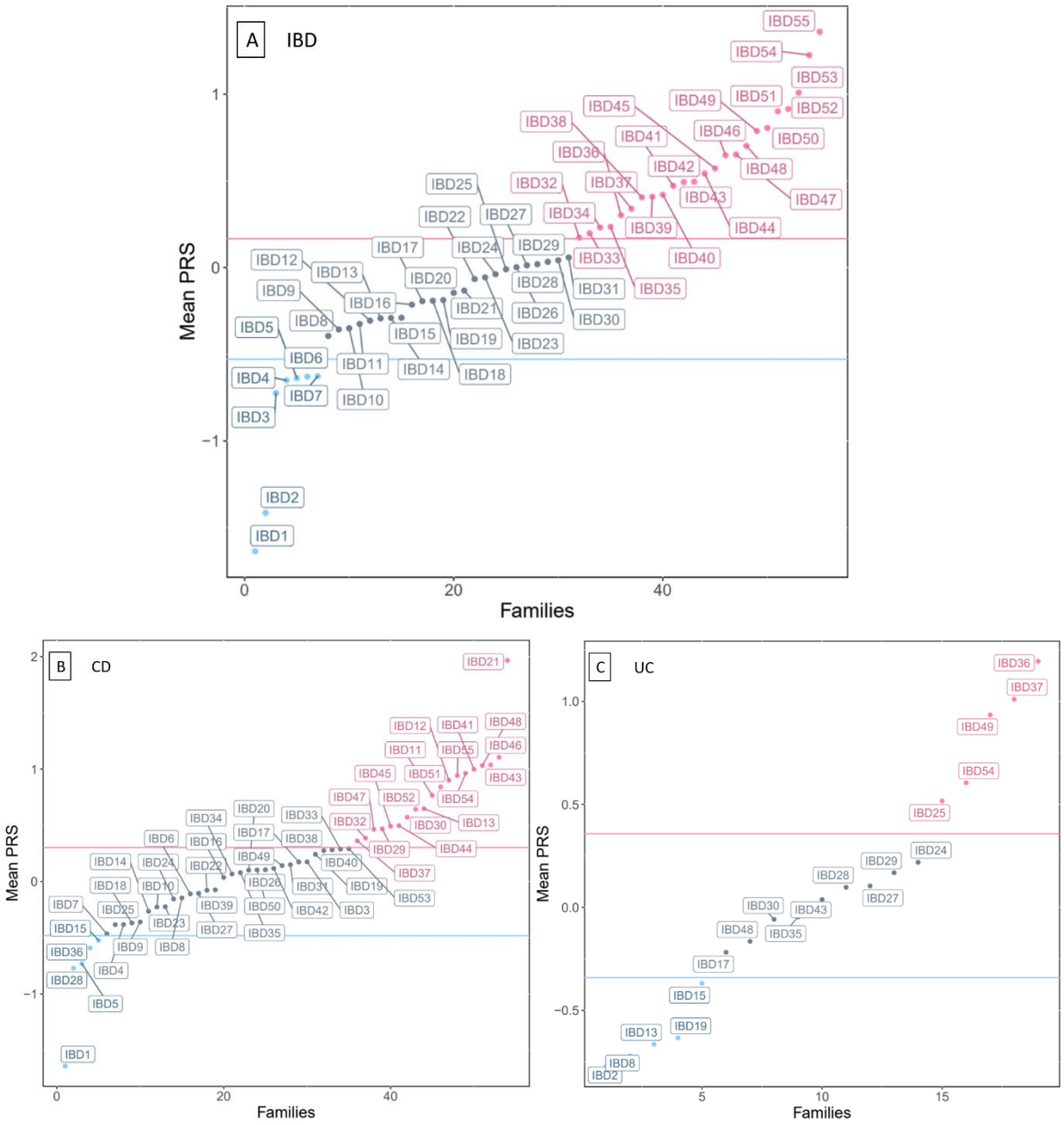



***Figure 15: Distribution of the mean PRS per family***

*The mean PRS (y-axis) per family are plotted in ascending order based on PRS IBD (A), PRS CD (B) and PRS UC (C). In (B) and (C) only CD (B) or UC (C) and mixed families are shown. Each dot represents a family, indicated by a unique family ID. The lines indicate the mean PRS of the healthy population controls (blue) and the mean PRS of the sporadic cases (pink). Families with a PRS lower than this threshold for healthy population controls are coloured blue. Families with a PRS higher than this threshold for sporadic cases are coloured pink. PRS were calculated based on pT = 0.01 and MAF = 0.01.*

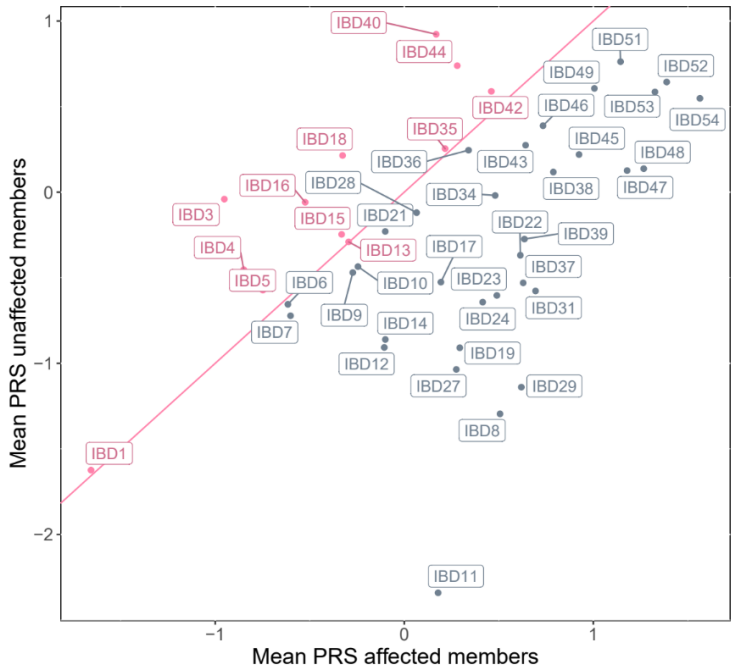



***Figure 16: Mean PRS in affected vs unaffected family members***

*The mean PRS of all affected members of a family (x-axis) is plotted against the mean PRS of all unaffected members of the same family (y-axis). The diagonal line is the x = y line. Every family with a higher mean PRS for the unaffected than the affected family members is coloured pink. PRS were calculated based on the effect sizes of IBD, pT = 0.01 and MAF = 0.01.*

between the affected and unaffected members. Interestingly, the families which have a very low family PRS are all found to have a higher or similar PRS of the unaffected members in comparison with the affected members. The mean PRS CD and PRS UC of affected and unaffected relatives are slightly different (Supplementary figures 3 and 4).

## 4.6 PRS is influenced by which genotyping chip is used

PRS are influenced by the size of the GWAS on which the effect sizes and p-values are based, and also by which and how many SNPs are available for the dataset for which you want to calculate the PRS. Thus, a larger or different set of SNPs might provide a different PRS. Some of the sporadic case-control individuals are genotyped on Immunochip as well as on the GSA chip. Immunochip is focused on several regions which are important in inflammatory and autoimmune diseases. Thus, large regions of the genome are not covered on Immunochip. The GSA chip on the other hand is meant to have a broad coverage of the whole genome and is not focused on specific regions. I computed the PRS based on Immunochip and GSA chip for the different predefined thresholds (pT = 5e-8, 1e-5, 0.01, 0.05, 0.1, 0.5) and performed a Spearman correlation analysis.

The directly genotyped SNPs, thus the unimputed data, showed a very good correlation for the PRS calculated with only genome-wide significant SNPs ($R = 0.84$, $p < 2.2e-16$). If the threshold is raised, the correlation between the two chips decreases (Figure 17A). While there is only a small drop in correlation for pT =1e-5 ($R = 0.82$, $p < 2.2e-16$), the correlation drops to 0.63 ($p < 2.2e-16$) for pT = 0.01, 0.53 ($p < 2.2e-16$) for pT = 0.05, 0.50 ($p < 2.2e-16$) for pT = 0.1, and 0.47 ($p < 2.2e-16$) for pT = 0.5.

Imputation of variants could increase the number of overlapping variants between the dataset of the two genotyping chips and thus also the correlation between PRS. Therefore, I also tested the correlation between scores based on imputed data of the two datasets. Correlation of PRS based on

genome-wide significant (R = 0.94, p < 2.2e-16) and suggestive SNPs (R = 0.90, p < 2.2e-16) is very high (Figure 17A). However, if more SNPs which are less associated are included, a drop in correlation occurs (pT 0.01: R = 0.59, p < 2.2e-16). For higher pTs, the correlations are largely similar to the unimputed data (pT 0.05: R = 0.52, p < 2.2e-16; pT 0.1: R = 0.50, p < 2.2e-16; pT 0.5: R = 0.47, p < 2.2e-16).

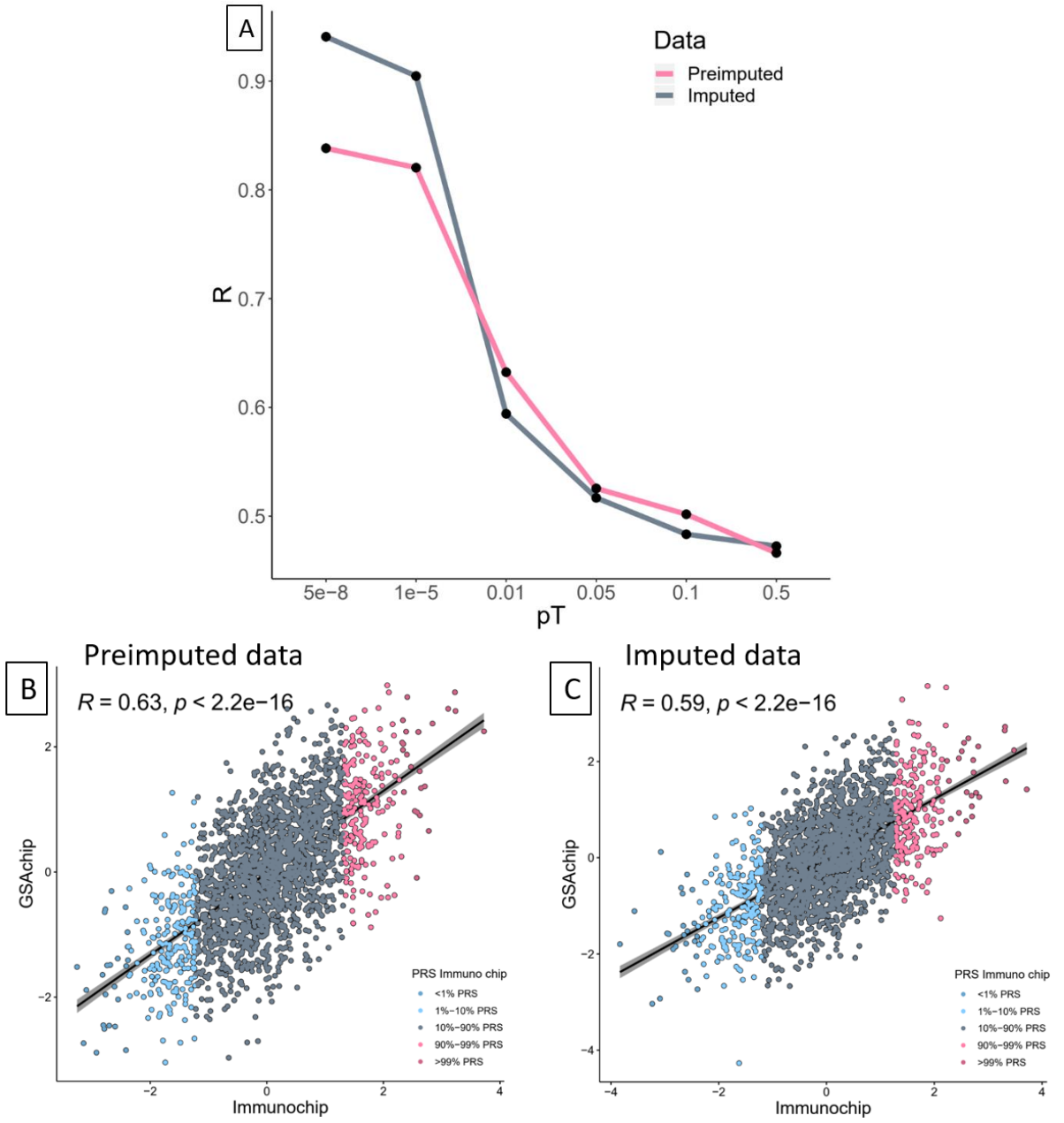



***Figure 17: Correlation GSA chip and Immunochip***

*(A) The spearman correlations (y-axis) between Immunochip and GSA chip is depicted for each threshold (x-axis). The colours indicate if the correlation is based on PRS calculated with unimputed (pink) or imputed (grey) data. (B, C) Correlation plots depicting the PRS based on Immunochip data (x-axis) and GSA chip data (y-axis) for unimputed data(B) and imputed data (C). Dark and light blue depicts, respectively, the 1% and 10% lowest PRS based on Immunochip data. Dark and light pink depicts, respectively, the 1% and 10% highest PRS based on Immunochip data. The spearman correlation is presented at the left upper corner of the plot. PRS were calculated based on effect sizes of IBD, SNPs with MAF > 0.01 and pT =0.01 (B).*

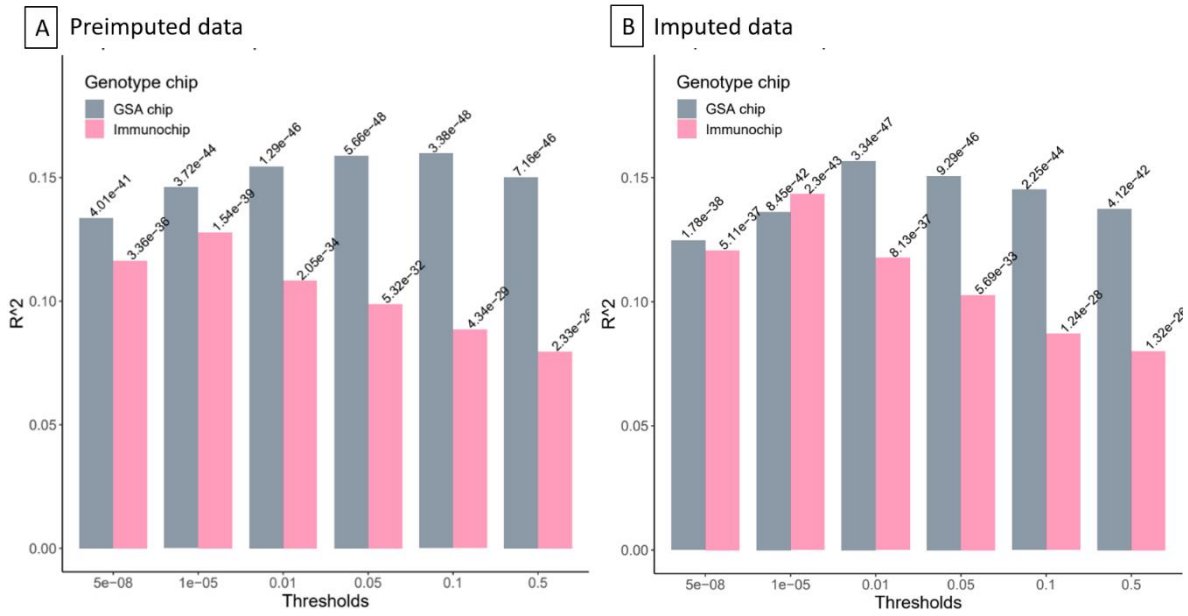



***Figure 18: Goodness-of-fit of the GSA chip and Immunochip***

*The pseudo-R2 for each pT are presented for GSA chip (grey) and Immunochip (pink) for the unimputed data (A) and the imputed data (B). PRS were calculated based on effect sizes of IBD and SNPs with MAF > 0.01. p < 8.33e-3 is considered significant.*

The individuals with an extreme PRS IBD based on SNPs genotyped on Immunochip seem visually to deviate a lot from the PRS of GSA data (Figure 17B and C). Seven (29.17%) and six (25%) individuals overlap between, respectively, the 1% lowest and highest PRS of the unimputed datasets of the two genotyping chips (Figure 17B). When the 10% lowest and highest PRS are investigated, the overlap is a bit higher with 104 (44.83%) and 97 (41.81%) individuals, respectively. The imputed datasets have a slightly lower amount of overlapping individuals. The overlap between the 1% lowest and highest PRS groups amounts to 5 (20,83%) and three (12.5%) individuals, respectively (Figure 17C). The larger group of 10% indicates the same with 93 (40.09%) individuals overlapping in the lowest 10% and 89 (38.36%) in the highest 10% of PRS. This indicates that individuals with an extreme PRS based on data of Immunochip do not necessarily also have an extreme PRS with GSA chip data.

The correlation of the PRS only indicates how similar the results are and does not indicate which chip is the best. Therefore, I calculated the pseudo-$R^2$ for each p-value threshold for both genotyping chips and compared them (Figure 18, Table 5). The PRS based on genome-wide association SNPs of the imputed data have almost the same goodness-of-fit with a slightly increased result for Immunochip (GSA: $R^2$ = 0.12, p = 5.11e-37; Immunochip: $R^2$ = 0.12, p = 1.78e-38). The GSA chip has a higher $R^2$ for the higher p-value threshold which include also genome-wide suggestive SNPs (GSA: $R^2$ = 0.14, p = 2.30e-43; Immunochip: $R^2$ = 0.14, p = 8.45e-42). However, when more SNPs are included, Immunochip achieves a better goodness-of-fit (Table 5). Moreover, PRS calculated with the preimputed datasets have even always a higher $R^2$ when individuals are genotyped on Immunochip.

***Table 5: Goodness-of-fit of PRS based on GSA chip and Immunochip***

| | Imputed | | | | Unimputed | | | |
|---|---|---|---|---|---|---|---|---|
| | GSA | | Immunochip | | GSA | | Immunochip | |
| **pT** | **$R^2$** | **p** | **$R^2$** | **p** | **$R^2$** | **p** | **$R^2$** | **p** |
| **5e-8** | 0.12 | 5.11e-37 | 0.12 | 1.78e-38 | 0.12 | 3.36e-36 | 0.13 | 4.01e-41 |
| **1e-5** | 0.14 | 2.30e-43 | 0.14 | 8.45e-42 | 0.13 | 1.54e-39 | 0.15 | 3.72e-44 |
| **0.01** | 0.12 | 8.13e-37 | 0.16 | 3.34e-47 | 0.11 | 2.05e-34 | 0.15 | 1.28e-46 |
| **0.05** | 0.10 | 5.69e-33 | 0.15 | 9.29e-46 | 0.10 | 5.31e-32 | 0.16 | 5.66e-48 |
| **0.1** | 8.7e-2 | 1.24e-28 | 0.15 | 2.25e-44 | 8.84e-2 | 4.33e-29 | 0.16 | 3.38e-48 |
| **0.5** | 8.01e-2 | 1.32e-26 | 0.14 | 4.12e-42 | 7.59e-2 | 2.33e-26 | 0.15 | 7.16e-46 |

*The pseudo-$R^2$ and corresponding p-value for the PRS based on several pTs of the cases vs controls model is presented for the imputed and unimputed genotypes of both genotyping chips, GSA chip and Immunochip.*

## 4.7 Familial cases have other specific risk variants than sporadic cases

I performed a family-based association analysis to detect variants associated with familial IBD. I compared the results with an association analysis solely executed with the sporadic dataset.

The association analysis based on sporadic cases and controls pointed at *NOD2* as the strongest association (Figure 19A, Table 6). Furthermore, the only two SNPs (chr 16 position 50763778 and position 50745926) that are genome-wide significant are both located in *NOD2*. Three more variants located in *NOD2* were genome-wide suggestively significant (p < 1e-5). Two regions were genome-wide suggestive on chromosome 1 with several independent variants: an intergenic region between *CENPF* and *KCNK2* (2 variants), and the region around *IL23R* which also includes *C1orf141* (4 variants). Lastly, one variant in *FAM83E* on chromosome 19 also reached the threshold of genome-wide suggestive significance.

In the family-based association analysis, 327 individuals were included. Genome-wide significant and genome-wide suggestive results were not found (Figure 19B, Table 6). The strongest association did just not reach suggestive significance, however nine associations had a p-value < 1e-4. The associations seen were entirely different from the associations with p < 1e-4 in the sporadic dataset. While in the sporadic dataset *NOD2* emerged as the strongest association, here *IL1RL2* ranks first. An independent variant in the same region, more specifically in *IL1RL1*, is also present in the list. There are no overlapping associated genes between the sporadic and familial dataset, indicating that within families other specific variants also play a role in the development of IBD.

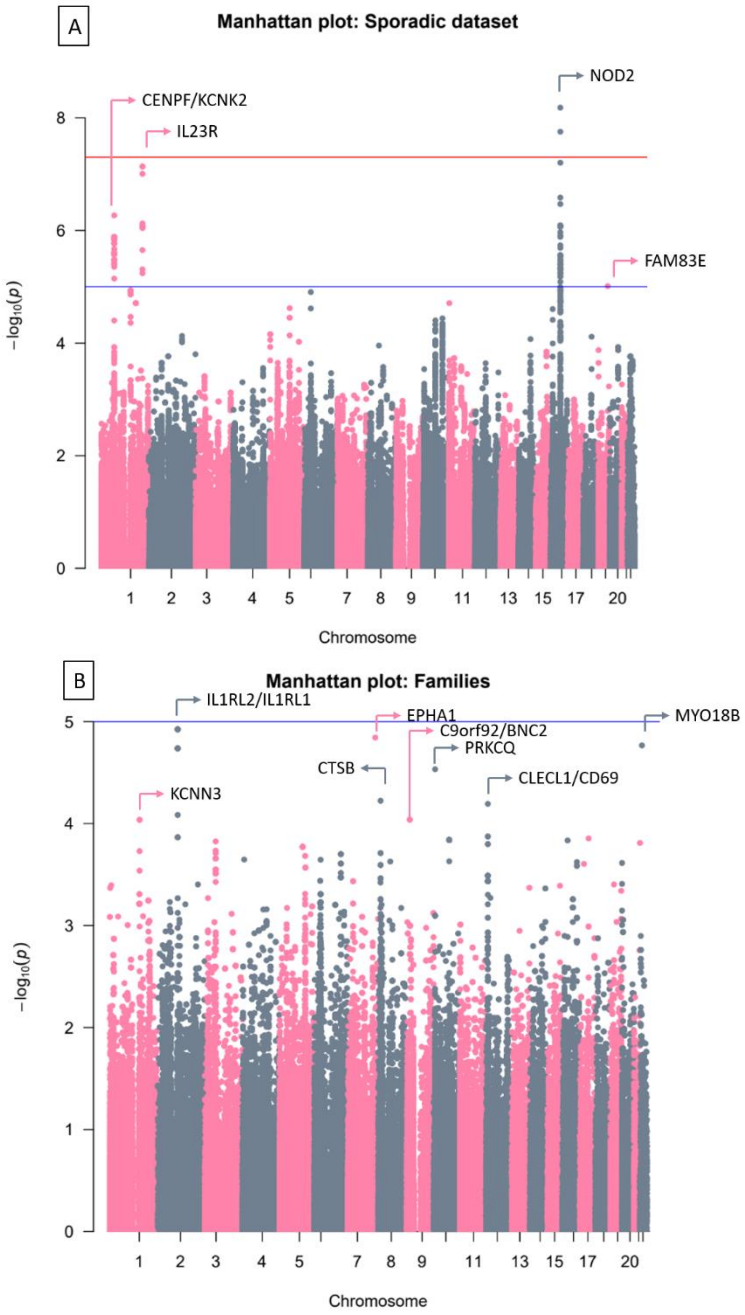



***Figure 19: Manhattan plots depicting the results of the association analysis with the sporadic dataset (A) and the familial dataset (B).***

*Each dot represents a variant with the strength of association (y-axis) presented as the negative log of the p-value and the chromosomal position (x-axis). The red line indicates the threshold for genome-wide significance (5e-8) and the blue lines indicate the threshold for genome-wide suggestive significance (1e-5). In (A) the genes where the variants are located in are shown for genome-wide significant and suggestive variants. In (B) the genes where the variants are located in are shown for associations with p < 1e-4.*

***Table 6: Top hits of the association analysis***

| Sporadic dataset | | |
|---|---|---|
| **SNP** | **p-value** | **Gene** |
| 16:50,763,778 | 6.58e-9 | *NOD2* |
| 16:50,745,926 | 1.76e-8 | *NOD2* |
| 1:215,002,708 | 7.32e-8 | Intergenic (*CENPF*, *KCNK2*) |
| 16:50,756,774 | 3.39e-7 | *NOD2* |
| 1:67,752,088 | 5.41e-7 | Intergenic (*IL23R*, *IL12RB2*) |
| 1:215,069,899 | 7.52e-7 | Intergenic (*CENPF*, *KCNK2*) |
| 16:50,737,498 | 8.10e-7 | *NOD2* |
| 1:67,705,958 | 1.29e-6 | *IL23R* |
| 1:67,596,372 | 1.65e-6 | *C1orf141* |
| 16:50,741,186 | 2.04e-6 | *NOD2* |
| 1:67,756,095 | 7.15e-6 | Intergenic (*IL23R*, *IL12RB2*) |
| 19:49,116,555 | 9.72e-6 | *FAM83E* |
| 1:153,235,837 | 1.16e-5 | Intergenic (*LOR*, *PGLYRP3*) |
| 6:32,397,662 | 1.24e-5 | Intergenic (*TSBP1-AS1*, *HLA-DRA*) |
| 1:180,633,231 | 1.95e-5 | *XPR1* |
| 11:2,223,850 | 1.95e-5 | Intergenic (*MIR4686*, *ASCL2*) |
| 5:102,664,001 | 2.40e-5 | Intergenic (*C5orf30*, *LINC02115*) |
| 16:50,769,262 | 2.49e-5 | Intergenic (*NOD2*, *CYLD*) |
| 16:11,023,868 | 2.49e-5 | *CIITA* |
| 10:101,290,301 | 3.64e-5 | *LINC01475* |
| 10:64,578,982 | 3.95e-5 | Intergenic (*EGR2*) |
| 16:50,661,273 | 4.86e-5 | *NKD1* |
| 5:1,281,693 | 6.96e-5 | *TERT* |
| 10:101,190,520 | 7.41e-5 | Intergenic (*GOT1*) |
| 16:50,912,675 | 7.43e-5 | Intergenic (*LINC02168*, *LINC02127*) |
| 2:172,368,120 | 7.46e-5 | Intergenic (*DCAF17*, *CYBRD1*) |
| 18:42,857,319 | 7.72e-5 | *SLC14A2* |
| 14:81,457,583 | 8.51e-5 | *TSHR* |
| 5:150,602,723 | 9.50e-5 | *LOC105378230* |
| **Familial dataset** | | |
| **SNP** | **p-value** | **Gene** |
| 2:102,835,706 | 1.19e-5 | *IL1RL2* |
| 7:143,093,824 | 1.43e-5 | *EPHA1* |
| 22:26,245,987 | 1.71e-5 | *MYO18B* |
| 10:6,548,841 | 2.93e-5 | *PRKCQ* |
| 8:11,708,355 | 5.98e-5 | *CTSB* |
| 12:9,896,953 | 6.41e-5 | Intergenic (*CLECL1*, *CD69*) |
| 2:102,945,121 | 8.23e-5 | *IL1RL1* |
| 9:16,378,123 | 9.16e-5 | Intergenic (*C9orf92*, *BNC2*) |
| 1:154,752,960 | 9.18e-5 | *KCNN3* |

*The variants with the strongest association in the sporadic and familial dataset. The variants are represented by chromosome:position of hg19. The strength of the association is defined by the p-value. Only associations with p < 1e-4 are presented. ). The red line indicates the threshold for genome-wide significance (5e-8) and the blue lines indicate the threshold for genome-wide suggestive significance (1e-5). Variants were annotated to genes or to the closest gene(s) if the variant was located intergenic.*

# 5 Discussion

To investigate the genetic architecture of multiplex families with IBD, I calculated polygenic risk scores (PRS) representing the genetic risk for IBD, and CD and UC separately. I did this for different p-value thresholds (pTs), indicating the certainty of the association, for each individual in the four groups: sporadic cases, healthy controls, affected family members and unaffected family members. I found that the PRS for each group was significantly different from each other, except the sporadic cases and affected family members. The healthy controls have the lowest PRS, followed by the unaffected family members. The PRS of sporadic and familial cases are similar to each other and the highest of all groups. However, a large part of mean PRS of families was above the mean of sporadic cases, indicating that the PRS of familial cases might be higher than their sporadic counterparts. Furthermore, I saw heterogeneity in PRS between the families, and in some families unaffected family members even had a higher PRS than their affected relatives. Lastly, I conducted a family-based association analysis and identified divergent associations with IBD within families.

CD and UC are different subtypes of IBD. The low correlation between PRS CD and PRS UC underlines the distinct genetic nature of the two subtypes. The first meta-analysis that combined CD and UC patients showed that 110 of the 163 loci known at the time are shared between the two subtypes.(33) A large part of the shared loci has various effect sizes in the two subtypes. Most loci have the same direction of effect, however some loci have effect sizes in the opposite direction. *NOD2*, the first gene associated with IBD, has a large risk effect for CD and a smaller protective effect in UC. PRS summarizes all associated variants and their effect sizes. Thus, differences in associated loci and slight differences in effect sizes between the two subtypes can give large discrepancies between the two PRS.

A previous genotype-phenotype association study found that IBD should be better divided into three subtypes, namely ileal CD, colonic CD and UC, instead of two subtypes, CD and UC.(79) Interestingly, this new classification was discovered with a PRS which uses the differences between CD and UC, and was based solely on loci which are genome-wide significantly associated with IBD. Thus, genome-wide significant loci can already make a distinction between CD and UC. The genetic relationship between five clinically related diseases, namely ankylosing spondylitis, Crohn's disease, psoriasis, primary sclerosing cholangitis and ulcerative colitis, was investigated in another study.(70) These five diseases have a large degree of pleiotropy – sharing of risk variants – however the variants with the highest effect sizes are often disease specific. Thus, a PRS based on genome-wide significant SNPs would probably be the most specific to the trait, and would correlate the least with the PRS of another trait. This is in concordance with my results where the PRS CD and PRS UC have the lowest correlation when only genome-wide significantly associated SNPs are included. When more SNPs which are less associated with the subtypes are included, the correlation of the PRS UC and PRS CD increases. Moreover, the PRS with extremely high values have probably an accumulation of variants with high effect sizes of one subtype, and the extremely low PRS a depletion. The effect sizes of these variants will be entirely different for the other subtype, resulting in a less high or low PRS as was seen in the small overlap of individuals with the 1% and 10% highest and lowest PRS. This indeed might indicate that the stronger associated loci are more specific to a certain subtype, while less strongly associated loci are shared between the subtypes.

Based on the above, I have suggested that the genetic nature of CD and UC is quite different. However, while the PRS of CD and UC indeed have a low correlation, the PRS IBD has a high correlation with both. Many loci are shared between CD and UC, and even have the same direction of effect.(33) Thus, the main difference is the magnitudes of the effect sizes. The very large meta-analyses with CD and UC patients together carried out in the past gave a combined effect size which will be located in between

the specific effect sizes of CD and UC. Therefore, the effect sizes and variants used will be more similar between IBD and CD, and IBD and UC, with a higher correlation as a consequence. Interestingly, also here an increase in correlation is visible when more SNPs are included in the PRS. Following the same reasoning as above, if a locus has a higher effect size and is highly specific to a disease, this locus will have a very different effect size between the two subtypes and even an intermediate value will be sizably different from the specific value, translating into various PRS for IBD and the subtypes. However, this effect will be less pronounced than between CD and UC PRS, as can be seen by the higher correlation between IBD and CD or UC, and a higher number of overlapping individuals with an extreme PRS.

With a PRS based on the five strongest risk factors, Weersma *et al* (2008) showed that sporadic cases have a higher PRS than controls.(78) In the meantime, many more genome-wide significant associated SNPs are discovered. Moreover, more heritability is explained for polygenic diseases when not only genome-wide significant SNPs but also less strongly associated SNPs are used to calculate PRS.(76) My results also indicate that the PRS of sporadic cases is higher than controls. For each PRS, I determined a goodness-of-fit measurement, $R^2$, to distinguish between the outcome groups. I saw approximately the same $R^2$ for each PRS, irrespective of the number of SNPs included, when distinguishing between sporadic cases and controls. This indicates that IBD really is a polygenic disease. If it had been an oligogenic disease, the PRS with strongly associated SNPs only, e.g. the genome-wide significant SNPs, would have been much better in separating cases and controls. The PRS IBD with only genome-wide significant SNPs even had the worst goodness-of-fit. This points to the presence of true associated SNPs among the set of less strongly associated SNPs. Although already more than 240 loci are associated with IBD, probably many more are still waiting to be identified.

A higher PRS is also seen in affected family members in comparison with their unaffected first-degree relatives. The results of Stittrich *et al* (2016) show the same trend but they do not specifically mention it.(101) This could indicate that even within families the amount of common genetic risk variants might be the reason that some family members are affected by IBD and others are not. The same conclusion can be drawn from the quantile analysis with the family dataset where the higher quantiles contain more cases than the lower quantiles. Thus, even within families, a higher PRS is connected with a higher chance to develop IBD. All PRS explain approximately the same amount of variance in the model which compares cases and controls within families. However, the best PRS seems to be towering a bit above the others, while this was not the case in the sporadic dataset. One possible explanation is that the genetic differences between affected and unaffected family members are divergent from the differences between sporadic cases and healthy controls. Sporadic cases have in general more common risk variants than the healthy controls, while affected family member might have a few common risk variants with higher effect sizes more than their unaffected relatives. If this was the case, then the best prediction PRS is the one which has more strongly associated variants. However, the genome-wide significant PRS performs the worst which could indicate that all or none of the family members have these risk variants. The best PRS has a pT slightly above the suggestively significant threshold. The less strongly associated variants might thus be the ones that determine whether or not a family member of a multiplex family develops IBD. These variants do not have a very large effect size like the genome-wide significant variants, however many of these variants might push an individual over the threshold.

Borren *et al* (2018) included 2,136 CD patients of which approximately one-third had a first- or second-degree relative with CD, the familial cases.(102) In their analysis, familial cases had a higher PRS than sporadic cases. These results are not in correspondence with my results where for all pTs the PRS of affected family members is similar to the PRS of sporadic cases. This could indicate that familial cases

are genetically similar to sporadic cases. They both could have many common variants and as a result develop IBD. However, this dataset was smaller than from Borren *et al* (2018) and a very strict correction was applied for multiple testing. Some PRS based on certain pTs are significantly higher in familial cases before correction for multiple testing and the best PRS is still almost significant after correction. This could indicate that their genetic burden of common risk variants is slightly higher than sporadic cases. Especially, the best PRS for familial IBD had a pT slightly above genome-wide suggestive significance. Cases, whether they are linked to a multiplex family or not, probably carry many genome-wide significant associated risk variants. However, the familial cases might have extra less strongly associated variants that the sporadic cases might not have, increasing their PRS. All the PRS only explain very few of the variability and none are significant after correction. Therefore, the genetics between sporadic and familial cases cannot differ greatly. Thus, I would propose that sporadic and familial cases have a similar common genetic risk burden. Although, a considerable part of families have a PRS above average in sporadic cases.

The unaffected family members have a significantly lower PRS than all cases, sporadic as well as familial, indicating that they carry less common risk variants. Moreover, the sensitivity analysis where the mean PRS of all unaffected relatives from one family is combined shows for some pTs a significantly different PRS between affected and unaffected relatives. Although they are part of a multiplex family, they have less genetic risk to develop IBD than their affected first-degree relatives. The difference in PRS could indicate that the affected family members have inherited many genetic risk factors while the unaffected relatives have not. Therefore, the familial cases were already more prone to develop IBD. However, unaffected family members still have a higher PRS than unrelated healthy controls. Thus, the healthy family members still have more common risk variants than population controls. This could indicate that in multiplex families more risk variants are segregating. Some family members inherit many common variants and in combination with a few environmental factors develop IBD, while others inherit less and do not reach the threshold to develop IBD.

Previous research is uncertain on whether familial aggregation is mostly due to a high burden of common variants or due to high effect size rare variants within families. A study of five families found four families with many common risk variants for CD or UC and only one family where the familial aggregation could not be linked to common risk variants.(101) However, another study investigated eight families and declared only one family as having a high genetic burden.(47) These are both only small studies and coincidence could play a large role. According to my results, familial cases have a similar PRS than their sporadic counterparts. However, the unaffected family members have a higher PRS than unrelated controls. In general, this indicates that multiplex families indeed have a high burden of common risk variants. Furthermore, when the entire dataset, sporadic and families, are divided into quintiles, the unaffected and affected family members are found more in the higher quintiles. Therefore, all family members seem to have a quite high PRS in comparison with the rest of the population. The higher PRS of affected and unaffected family members indicates that in multiplex families many common risk variants circulate. Thus, familial aggregation seems, in general, to be due to a high burden of common variants.

In both of the two papers that looked at PRS in IBD families, the families are a combination of families with a high burden of common risk variants and families without.(47,101) Thus, heterogeneity between families seems to exist. They also both found rare variants with a high effect size to clarify familial aggregation in some of the families, indicating another mechanism than many common variants. Another study also found some candidate rare variants in a cohort of multiplex Jewish families.(103) A large variability is also seen in the PRS of the families in my dataset. Moreover, seven families are having a lower family PRS than the mean PRS of healthy population controls, the group

with the lowest PRS. The familial aggregation in these families can probably not be attributed to common risk variants. The very low PRS indicates that only a few common risk variants are present in those families. Moreover, these low PRS families have also an unbalanced PRS between affected and unaffected family members. The affected relatives often have an equal or lower PRS than their unaffected family members, indicating that the PRS does not matter here. Of note, most families of all families still have cases with a higher PRS than their unaffected relatives, supporting the hypothesis that the affected family members have IBD due to a high burden of common risk variants. Thus, many families indeed have a high burden of common risk variants, while a few families have another reason for familial aggregation.

All analyses were performed with imputed genotype data from Immunochip. Immunochip covers mainly known regions from autoimmune and inflammatory diseases and large portions of the genome are not covered at all. Therefore, associations can be missed. Chen *et al* (2014) found that more heritability is explained with a chip that covers more of the genome and therefore recommend a chip with a GWAS backbone.(27) I compared PRS IBD based on unimputed and imputed data genotyped on Immunochip and GSA chip. The high correlation between scores with only genome-wide significant associated SNPs included indicate that both have a good coverage of the strongly associated SNPs. However, when more SNPs are included, the correlation decreases, and also the number of variants included in the PRS start to differ. Immunochip has many SNPs in the known regions, while these regions are less covered with the GSA chip. On the other hand, GSA chip has variants in regions where Immunochip has none or only a few. Therefore, the PRS are based on other variants and correlate less. Interestingly, PRS based on Immunochip is better in explaining the variance of the case-control status of IBD. One possible explanation for this is the introduced noise with the inclusion of too many non-associated SNPs with GSA chip. The GSA chip includes a broader region and therefore many more variants. In this abundance of variants some undiscovered associations might be present, however many more are probably not associated. The non-associated SNPs induce noise which overshadows the real associations in the PRS. To solve this problem, probably larger GWAS need to be performed with genome-wide chips. The most recent large meta-analyses from the IIBDGC are mainly conducted with data from Immunochip as this was financially more interesting at the time.(32,33) The most recent GWAS, the one also used in my analyses, used for most individuals the Human Core Exome Chip which has already a broader coverage.(31) However, larger GWAS with inclusion of the non-coding regions of the genome will probably be necessary to discover more common genetic risk variants.

I have mentioned that the best PRS of affected vs unaffected family members have a higher goodness-of-fit than the PRS based on my pre-defined pTs, while in the model of sporadic cases vs controls all PRS were more or less similar. The underlying genetics might differ within families and cause this slight aberration in results. To further investigate this, I performed an association analysis with the sporadic dataset, and a family-based association analysis in the family dataset. As expected, in the sporadic dataset, this analysis pointed to the genes with the strongest known effect sizes, e.g. *NOD2* and *IL23R*, as having the strongest association. Interestingly, family-based association marked other genes as having a strong association. Of note, the family-based associations are based on a few hundred individuals and therefore even the strongest associations do not reach suggestive significance (1e-5). The absence of overlap between the associations with a $p < 1e-4$ indicates that within families other genes are additionally important for developing IBD. Thus, multiplex families might have other specific variants associated with IBD.

In the family-based association the variant which emerged as the association with the lowest p-value is located in the gene *IL1R2*. Interestingly, *IL1R2* belongs to one of the 240 loci, and the implicated gene of that locus is *IL18RAP*. These variant in *IL1R2* might be in linkage disequilibrium with *IL18RAP*,

however this variant might also be independent. *IL18RAP* was discovered in 2018 by a candidate-gene study.(104) The gene translates into a part of the IL18 receptor complex which is also expressed in the intestinal epithelial cells. Another variant in the *IL1R1* gene is also included in this locus, and also has a low p-value in the family-based association. The two variants are clumped separately, thus they indicate two independent signals. Thus, the *IL18RAP* locus seems to be important in familial IBD.

A variant in *PRKCQ*, which resides in the locus of *IL2RA*, is also associated to familial IBD. *IL2RA* is a known associated locus of IBD and recently a duplication of the gene has been implicated as a cause of VEO-IBD.(105) The IL2-pathway is important in T-cell proliferation and to maintain intestinal homeostasis. The occurrence of *IL2RA* in VEO-IBD indicates that some variants might have a very high or even causal effect size. Other variants in the same gene, not causal on their own, might clarify the association with familial IBD.

All other associations with $p < 1e\text{-}4$ are not located in loci which are genome-wide significantly associated with IBD at the moment. One of these associations is a variant which is located in the gene *EPHA*. Although the ephrin receptor EphA has not been associated to IBD, the other class of ephrin receptors, EphB, is involved in an autophagy pathway implicated in UC.(106) Moreover, EphA has a modulatory role in acute inflammatory responses.(107) An antagonist of ephrins has beneficial effects in a CD mouse model, however these are due to an inhibition of ephrin B signalling.(107) Another identified gene is *CTSB*, which translates to cathepsin B. Cathepsin B is upregulated in macrophages of IBD patients, and inhibition of cathepsin B and L in mice led to an amelioration of inflammation.(108) I could not find any involvement of the other identified genes in IBD. However, the variants can be in linkage disequilibrium with the causative variant which might be located in or near other genes, or have a long-distance regulatory effect. The identified loci might be new associations with IBD, however more research is necessary to validate these findings.

A first limitation of this study is the use of Immunochip. Although I provide evidence that the Immunochip can better separate cases from controls, large regions are not covered and these might hold important information. Secondly, I take the effect sizes and strength of associations from the study of de Lange *et al* (2017).(31) This GWAS is conducted with most individuals genotyped with a chip covering the exome and thus few information about the non-coding regions is provided. However, this GWAS is the most recently published GWAS with individuals of European ancestry. The latest GWAS performed by the IIBDGC in 2015 is larger, however the sporadic dataset used in this thesis also was part of that GWAS and therefore my results would not be correct if I applied these effect sizes and p-values. I also report my analyses based on the IBD effect sizes and not separately for CD and UC. IBD had a high correlation with both of its subtypes and the combination of CD and UC patients increased the sample size. Therefore, I decided that I could draw my conclusions based on IBD. I performed most of my analyses also separately for CD and UC and these are added as supplementary data.

The reason of familial aggregation of IBD in multiplex families is unknown. In many families, a lot of common genetic risk variants are segregating. However, the affected family members still have a higher burden of common risk variants than their unaffected relatives. Thus, the familial aggregation seems to be caused by a high burden of common variants present in these families. However, some families do not carry many common variants and therefore familial aggregation in these cases cannot be attributed to the same reason. Rare variants could be a possible explanation for familial aggregation in these multiplex families. Although in many families the rule that a higher burden of common variants increases the chance to develop IBD is followed, it seems that familial IBD also has some novel specific variants associated to it.

## 5.1 Future directions

Although this study managed to put a step in the right direction, the genetic architecture of multiplex families is far from resolved. Many families seem to have a high burden of common variants present in the family. However, the family-based association marked other genes as having a strong association with familial IBD. Validation of these new variants or loci will be necessary. Further research is also necessary to see whether these variants or their associated genes also have an association with sporadic IBD, or if they are specific to familial IBD. Moreover, many different methods exist to perform family-based association analysis, e.g. PLINK, GEMMA, REGENIE.(96,109,110) It would be interesting to see if other methods point to the same loci. Some families do not have many common variants. These families must have another reason for familial aggregation and rare variants could be one of them. Families with a low polygenic risk thus probably have a higher chance to carry a rare variant with a moderate to high effect size, and make good candidates to further look for these rare variants with whole-exome or whole-genome sequencing.(111) As PRS do not clarify all heritability, the families with a moderate or high PRS might also be interesting to investigate for rare variants that could be important on top of the polygenic risk.

IBD is a complex disease and its development is the result of genetic and environmental factors. Thus, environmental factors might not be forgotten, especially not in families, where many environmental factors are also shared among relatives. However, environmental factors are difficult to investigate, thus the use of proxies might be indicated. The microbiome is partly inherited from parents and siblings.(112) Moreover, a different composition between affected and unaffected family members has already been established.(113) Therefore, studying the composition of the microbiome within families might be interesting to discover the contribution of the microbiome to familial aggregation. Another proxy can be the epigenome, which is influenced by environmental factors, e.g. smoking.(114) Through comparing the epigenomic signature of sporadic and familial cases, we might get an indication how much environmental factors contribute to familial aggregation. Furthermore, variation in the epigenome between affected and unaffected relatives might also provide useful information to discover why some family members develop IBD while others do not.

# List of tables

# List of figures

# References


1. Sairenji T, Collins K, Evans D. An Update on Inflammatory Bowel Disease. Prim Care [Internet]. 2017 Dec 1 [cited 2021 Sep 21];44(4):673–92. Available from: https://pubmed.ncbi.nlm.nih.gov/29132528/
2. Freeman HJ. Natural history and long-term clinical course of Crohn's disease. World J Gastroenterol [Internet]. 2014 Jan 7 [cited 2021 Nov 19];20(1):31–6. Available from: https://pubmed.ncbi.nlm.nih.gov/24415855/
3. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet (London, England) [Internet]. 2017 Dec 23 [cited 2021 Sep 22];390(10114):2769–78. Available from: https://pubmed.ncbi.nlm.nih.gov/29050646/
4. Burisch J, Pedersen N, Čuković-Čavka S, Brinar M, Kaimakliotis I, Duricova D, et al. East-West gradient in the incidence of inflammatory bowel disease in Europe: the ECCO-EpiCom inception cohort. Gut [Internet]. 2014 [cited 2022 May 17];63(4):588–97. Available from: https://pubmed.ncbi.nlm.nih.gov/23604131/
5. Aniwan S, Park SH, Loftus E V. Epidemiology, Natural History, and Risk Stratification of Crohn's Disease. Gastroenterol Clin North Am [Internet]. 2017 Sep 1 [cited 2021 Nov 19];46(3):463–80. Available from: https://pubmed.ncbi.nlm.nih.gov/28838409/
6. J S, MS S, S V, JF C. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut [Internet]. 2006 Jun [cited 2021 Sep 21];55(6):749–53. Available from: https://pubmed.ncbi.nlm.nih.gov/16698746/
7. K G, P VE. Inflammatory bowel disease unclassified and indeterminate colitis: the role of the pathologist. J Clin Pathol [Internet]. 2009 Mar [cited 2021 Sep 22];62(3):201–5. Available from: https://pubmed.ncbi.nlm.nih.gov/18952692/
8. Magro F, Gionchetti P, Eliakim R, Ardizzone S, Armuzzi A, Barreiro-de Acosta M, et al. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders. J Crohns Colitis [Internet]. 2017 Jun 1 [cited 2021 Oct 29];11(6):649–70. Available from: https://pubmed.ncbi.nlm.nih.gov/28158501/
9. Gomollón F, Dignass A, Annese V, Tilg H, Van Assche G, Lindsay J, et al. 3rd European Evidence-based Consensus on the Diagnosis and Management of Crohn's Disease 2016: Part 1: Diagnosis and Medical Management. J Crohns Colitis [Internet]. 2017 Jan 1 [cited 2021 Sep 21];11(1):3–25. Available from: https://pubmed.ncbi.nlm.nih.gov/27660341/
10. EK W, NS D, O N. Management of inflammatory bowel disease. Med J Aust [Internet]. 2018 Sep 1 [cited 2021 Oct 28];209(7):318–23. Available from: https://pubmed.ncbi.nlm.nih.gov/30257634/
11. Kaenkumchorn T, Wahbeh G. Ulcerative Colitis: Making the Diagnosis. Gastroenterol Clin North Am [Internet]. 2020 Dec 1 [cited 2021 Nov 9];49(4):655–69. Available from: https://pubmed.ncbi.nlm.nih.gov/33121687/
12. F M, P G, R E, S A, A A, M BA, et al. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 1: Definitions, Diagnosis, Extra-intestinal Manifestations, Pregnancy, Cancer Surveillance, Surgery, and Ileo-anal Pouch Disorders. J Crohns Colitis [Internet]. 2017 Jun 1 [cited 2021 Oct 29];11(6):649–70. Available from: https://pubmed.ncbi.nlm.nih.gov/28158501/
13. Peyrin-Biroulet L, Loftus E V., Colombel JF, Sandborn WJ. The natural history of adult Crohn's disease in population-based cohorts. Am J Gastroenterol [Internet]. 2010 Feb [cited 2021 Nov 19];105(2):289–97. Available from: https://pubmed.ncbi.nlm.nih.gov/19861953/
14. Selby W. The natural history of ulcerative colitis. Baillieres Clin Gastroenterol [Internet]. 1997 [cited 2021 Nov 23];11(1):53–64. Available from: https://pubmed.ncbi.nlm.nih.gov/9192060/
15. Fumery M, Singh S, Dulai PS, Gower-Rousseau C, Peyrin-Biroulet L, Sandborn WJ. Natural History of Adult Ulcerative Colitis in Population-based Cohorts: A Systematic Review. Clin Gastroenterol Hepatol [Internet]. 2018 Mar 1 [cited 2021 Nov 19];16(3):343-356.e3. Available from: https://pubmed.ncbi.nlm.nih.gov/28625817/
16. Duricova D, Fumery M, Annese V, Lakatos PL, Peyrin-Biroulet L, Gower-Rousseau C. The natural history of Crohn's disease in children: a review of population-based studies. Eur J Gastroenterol Hepatol [Internet]. 2017 [cited 2021 Nov 19];29(2):125–34. Available from: https://pubmed.ncbi.nlm.nih.gov/27748673/
17. Yu YR, Rodriguez JR. Clinical presentation of Crohn's, ulcerative colitis, and indeterminate colitis:

Symptoms, extraintestinal manifestations, and disease phenotypes. Semin Pediatr Surg [Internet]. 2017 Dec 1 [cited 2021 Nov 9];26(6):349–55. Available from: https://pubmed.ncbi.nlm.nih.gov/29126502/
18. Harbord M, Annese V, Vavricka SR, Allez M, Acosta MB de, Boberg KM, et al. The First European Evidence-based Consensus on Extra-intestinal Manifestations in Inflammatory Bowel Disease. J Crohns Colitis [Internet]. 2016 Mar 1 [cited 2022 Jan 10];10(3):239–54. Available from: https://pubmed.ncbi.nlm.nih.gov/26614685/
19. Yasukawa S, Matsui T, Yano Y, Sato Y, Takada Y, Kishi M, et al. Crohn's disease-specific mortality: a 30-year cohort study at a tertiary referral center in Japan. J Gastroenterol [Internet]. 2019 Jan 25 [cited 2022 May 6];54(1):42. Available from: /pmc/articles/PMC6314978/
20. Israeli E, Graff LA, Clara I, Walker JR, Lix LM, Targownik LE, et al. Low prevalence of disability among patients with inflammatory bowel diseases a decade after diagnosis. Clin Gastroenterol Hepatol [Internet]. 2014 [cited 2022 May 6];12(8). Available from: https://pubmed.ncbi.nlm.nih.gov/24361416/
21. J T, S B, G D, T K, JP G, T R, et al. ECCO Guidelines on Therapeutics in Crohn's Disease: Medical Treatment. J Crohns Colitis [Internet]. 2020 Jan 1 [cited 2021 Oct 29];14(1):4–22. Available from: https://pubmed.ncbi.nlm.nih.gov/31711158/
22. M H, R E, D B, K K, K K, U K, et al. Third European Evidence-based Consensus on Diagnosis and Management of Ulcerative Colitis. Part 2: Current Management. J Crohns Colitis [Internet]. 2017 Jul 1 [cited 2021 Oct 29];11(7):769–84. Available from: https://pubmed.ncbi.nlm.nih.gov/28513805/
23. E B, MH M, AM S, F S, G VA, P R, et al. Treatment Algorithm for Mild and Moderate-to-Severe Ulcerative Colitis: An Update. Digestion [Internet]. 2020 Sep 1 [cited 2021 Oct 29];101 Suppl(Suppl1):2–15. Available from: https://pubmed.ncbi.nlm.nih.gov/31945767/
24. AN S, AN A, M R. Diet in Treatment of Inflammatory Bowel Diseases. Clin Gastroenterol Hepatol [Internet]. 2021 Mar 1 [cited 2021 Oct 29];19(3):425-435.e3. Available from: https://pubmed.ncbi.nlm.nih.gov/31812656/
25. Orholm M, Munkholm P, Langholz E, Nielsen OH, Sørensen TIA, Binder V. Familial occurrence of inflammatory bowel disease. N Engl J Med [Internet]. 1991 Jan 10 [cited 2022 Jan 10];324(2):84–8. Available from: https://pubmed.ncbi.nlm.nih.gov/1984188/
26. Gordon H, Moller FT, Andersen V, Harbord M. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. Inflamm Bowel Dis [Internet]. 2015 Apr 3 [cited 2021 Sep 30];21(6):1428–34. Available from: https://pubmed.ncbi.nlm.nih.gov/25895112/
27. Chen G-B, Lee SH, Brion M-JA, Montgomery GW, Wray NR, Radford-Smith GL, et al. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Hum Mol Genet [Internet]. 2014 [cited 2021 Sep 30];23(17):4710–20. Available from: https://pubmed.ncbi.nlm.nih.gov/24728037/
28. Mirkov M umićević, Verstockt B, Cleynen I. Genetics of inflammatory bowel disease: beyond NOD2. lancet Gastroenterol Hepatol [Internet]. 2017 Mar 1 [cited 2021 Sep 29];2(3):224–34. Available from: https://pubmed.ncbi.nlm.nih.gov/28404137/
29. Hugot J-P, Laurent-Puig P, Gower-Rousseaut C, Olson JM, Lee JC, Weissenbach J, et al. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. 1996;
30. Crowley E, Warner N, Pan J, Khalouei S, Elkadri A, Fiedler K, et al. Prevalence and Clinical Features of Inflammatory Bowel Diseases Associated With Monogenic Variants, Identified by Whole-Exome Sequencing in 1000 Children at a Single Center. Gastroenterology [Internet]. 2020 Jun 1 [cited 2021 Nov 25];158(8):2208–20. Available from: https://pubmed.ncbi.nlm.nih.gov/32084423/
31. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet [Internet]. 2017 Jan 31 [cited 2021 Sep 30];49(2):256–61. Available from: https://pubmed.ncbi.nlm.nih.gov/28067908/
32. Liu JZ, Van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet [Internet]. 2015 Aug 27 [cited 2021 Sep 30];47(9):979–86. Available from: https://pubmed.ncbi.nlm.nih.gov/26192919/
33. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature [Internet]. 2012 Nov 1 [cited 2021 Oct 1];491(7422):119–24. Available from: https://pubmed.ncbi.nlm.nih.gov/23128233/
34. Pierre A Saint, Génin E. How important are rare variants in common disease? Brief Funct Genomics [Internet]. 2014 Sep 1 [cited 2021 Nov 25];13(5):353–61. Available from: https://pubmed.ncbi.nlm.nih.gov/25005607/

35. Venkataraman GR, Rivas MA. Rare and common variant discovery in complex disease: the IBD case study. Hum Mol Genet [Internet]. 2019 Nov 21 [cited 2021 Nov 26];28(R2):R162–9. Available from: https://pubmed.ncbi.nlm.nih.gov/31363759/
36. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature [Internet]. 2013 [cited 2021 Nov 29];498(7453):232–5. Available from: https://pubmed.ncbi.nlm.nih.gov/23698362/
37. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature [Internet]. 2021 Sep 23 [cited 2021 Nov 29];597(7877):527–32. Available from: https://pubmed.ncbi.nlm.nih.gov/34375979/
38. Agarwala V, Flannick J, Sunyaev S, Altshuler D. Evaluating empirical bounds on complex disease genetic architecture. Nat Genet [Internet]. 2013 Dec [cited 2021 Nov 29];45(12):1418–27. Available from: https://pubmed.ncbi.nlm.nih.gov/24141362/
39. Jezernik G, Mičetić-Turk D, Potočnik U. Molecular Genetic Architecture of Monogenic Pediatric IBD Differs from Complex Pediatric and Adult IBD. J Pers Med [Internet]. 2020 Nov 1 [cited 2021 Oct 12];10(4):1–18. Available from: https://pubmed.ncbi.nlm.nih.gov/33255894/
40. Russel RK, Satsangi J. IBD: a family affair. Best Pract Res Clin Gastroenterol [Internet]. 2004 Jun [cited 2021 Nov 5];18(3):525–39. Available from: https://pubmed.ncbi.nlm.nih.gov/15157825/
41. DeLisi LE. A Case for Returning to Multiplex Families for Further Understanding the Heritability of Schizophrenia: A Psychiatrist's Perspective. Mol neuropsychiatry [Internet]. 2016 [cited 2021 Dec 2];2(1):15–9. Available from: https://pubmed.ncbi.nlm.nih.gov/27606317/
42. Cleynen I, Vermeire S. The genetic architecture of inflammatory bowel disease: past, present and future. Curr Opin Gastroenterol [Internet]. 2015 [cited 2021 Nov 24];31(6):456–63. Available from: https://pubmed.ncbi.nlm.nih.gov/26444824/
43. Cavanaugh JA, Bryce ME, Stanford PM, Pavli P, Vermeire S, Peeters M, et al. International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: Crohn disease and chromosome 16. Am J Hum Genet [Internet]. 2001 [cited 2021 Dec 3];68(5):1165–71. Available from: https://pubmed.ncbi.nlm.nih.gov/11309682/
44. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cézard JP, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature [Internet]. 2001 May 31 [cited 2022 Jan 12];411(6837):599–603. Available from: https://pubmed.ncbi.nlm.nih.gov/11385576/
45. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001 22 [Internet]. 2001 Feb [cited 2022 Apr 6];2(2):91–9. Available from: https://www.nature.com/articles/35052543
46. Stittrich AB, Ashworth J, Shi M, Robinson M, Mauldin D, Brunkow ME, et al. Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. Hum genome Var [Internet]. 2016 Dec [cited 2021 Nov 5];3(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27081563/
47. Park YM, Ha E, Gu KN, Shin GY, Lee CK, Kim K, et al. Host Genetic and Gut Microbial Signatures in Familial Inflammatory Bowel Disease. Clin Transl Gastroenterol [Internet]. 2020 [cited 2021 Dec 4];11(7). Available from: https://pubmed.ncbi.nlm.nih.gov/32764209/
48. Frade-Proud'hon-Clerc S, Smol T, Frenois F, Sand O, Vaillant E, Dhennin V, et al. A Novel Rare Missense Variation of the NOD2 Gene: Evidencesof Implication in Crohn's Disease. Int J Mol Sci [Internet]. 2019 Feb 2 [cited 2021 Nov 5];20(4). Available from: https://pubmed.ncbi.nlm.nih.gov/30769939/
49. Okou DT, Mondal K, Faubion WA, Kobrynski LJ, Denson LA, Mulle JG, et al. Exome sequencing identifies a novel FOXP3 mutation in a 2-generation family with inflammatory bowel disease. J Pediatr Gastroenterol Nutr [Internet]. 2014 [cited 2021 Dec 4];58(5):561–8. Available from: https://pubmed.ncbi.nlm.nih.gov/24792626/
50. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science [Internet]. 2006 Dec 1 [cited 2021 Sep 30];314(5804):1461–3. Available from: https://pubmed.ncbi.nlm.nih.gov/17068223/
51. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. Nat Genet [Internet]. 2007 Feb [cited 2022 Jan 21];39(2):207–11. Available from: https://pubmed.ncbi.nlm.nih.gov/17200669/
52. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet [Internet]. 2007 Apr [cited 2022 Jan 21];3(4):0538–43. Available from: https://pubmed.ncbi.nlm.nih.gov/17447842/

53. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet [Internet]. 2007 [cited 2022 Jan 21];39(5):596–604. Available from: https://pubmed.ncbi.nlm.nih.gov/17435756/
54. Michailidou K. Meta-Analysis of Common and Rare Variants. Methods Mol Biol [Internet]. 2018 [cited 2022 Jan 11];1793:73–88. Available from: https://pubmed.ncbi.nlm.nih.gov/29876892/
55. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet [Internet]. 2011 [cited 2021 Sep 30];43(3):246–52. Available from: https://pubmed.ncbi.nlm.nih.gov/21297633/
56. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet [Internet]. 2010 [cited 2021 Oct 11];42(12):1118–25. Available from: https://pubmed.ncbi.nlm.nih.gov/21102463/
57. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet [Internet]. 2008 [cited 2021 Oct 18];40(8):955–62. Available from: https://pubmed.ncbi.nlm.nih.gov/18587394/
58. Hong M, Ye BD, Yang SK, Jung S, Lee HS, Kim BM, et al. Immunochip meta-analysis of inflammatory bowel disease identifies three novel loci and four novel associations in previously reported loci. J Crohn's Colitis. 2018 May 25;12(6):730–41.
59. Brant SR, Okou DT, Simpson CL, Cutler DJ, Haritunians T, Bradfield JP, et al. Genome-Wide Association Study Identifies African-Specific Susceptibility Loci in African Americans With Inflammatory Bowel Disease. Gastroenterology [Internet]. 2017 Jan 1 [cited 2021 Sep 30];152(1):206-217.e2. Available from: https://pubmed.ncbi.nlm.nih.gov/27693347/
60. Cleynen I, Halfvarsson J. How to approach understanding complex trait genetics - inflammatory bowel disease as a model complex trait. United Eur Gastroenterol J [Internet]. 2019 Dec 1 [cited 2021 Nov 23];7(10):1426–30. Available from: https://pubmed.ncbi.nlm.nih.gov/31839967/
61. Huang H, Fang M, Jostins L, Mirkov M umićević, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature [Internet]. 2017 Jul 13 [cited 2021 Sep 30];547(7662):173–8. Available from: https://pubmed.ncbi.nlm.nih.gov/28658209/
62. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet [Internet]. 2011 Nov [cited 2021 Sep 30];43(11):1066–73. Available from: https://pubmed.ncbi.nlm.nih.gov/21983784/
63. Prescott NJ, Lehne B, Stone K, Lee JC, Taylor K, Knight J, et al. Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTNL2 and implicates other immune related genes. PLoS Genet [Internet]. 2015 [cited 2021 Dec 7];11(2):1–19. Available from: https://pubmed.ncbi.nlm.nih.gov/25671699/
64. Ellinghaus D, Zhang H, Zeissig S, Lipinski S, Till A, Jiang T, et al. Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. Gastroenterology [Internet]. 2013 [cited 2021 Dec 4];145(2):339–47. Available from: https://pubmed.ncbi.nlm.nih.gov/23624108/
65. Onoufriadis A, Stone K, Katsiamides A, Amar A, Omar Y, de Lange KM, et al. Exome Sequencing and Genotyping Identify a Rare Variant in NLRP7 Gene Associated With Ulcerative Colitis. J Crohns Colitis [Internet]. 2018 Feb 28 [cited 2021 Dec 4];12(3):321–6. Available from: https://pubmed.ncbi.nlm.nih.gov/29211899/
66. Sazonovs A, Stevens CR, Venkataraman GR, Yuan K, Avila B, Abreu MT, et al. Sequencing of over 100,000 individuals identifies multiple genes and rare variants associated with Crohns disease susceptibility. medRxiv [Internet]. 2021 Jul 5 [cited 2022 Feb 9];2021.06.15.21258641. Available from: https://www.medrxiv.org/content/10.1101/2021.06.15.21258641v2
67. Luo Y, de Lange KM, Jostins L, Moutsianas L, Randall J, Kennedy NA, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nat Genet [Internet]. 2017 Jan 31 [cited 2021 Sep 30];49(2):186–92. Available from: https://pubmed.ncbi.nlm.nih.gov/28067910/
68. Somineni HK, Nagpal S, Venkateswaran S, Cutler DJ, Okou DT, Haritunians T, et al. Whole-genome sequencing of African Americans implicates differential genetic architecture in inflammatory bowel disease. Am J Hum Genet [Internet]. 2021 Mar 4 [cited 2021 Oct 12];108(3):431–45. Available from: https://pubmed.ncbi.nlm.nih.gov/33600772/

69. JK P, T B, JZ L, L S, JY T, DA H. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet [Internet]. 2016 Jul 1 [cited 2021 Sep 30];48(7):709–17. Available from: https://pubmed.ncbi.nlm.nih.gov/27182965/
70. D E, L J, SL S, A C, J B, B H, et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet [Internet]. 2016 May 1 [cited 2021 Sep 30];48(5):510–8. Available from: https://pubmed.ncbi.nlm.nih.gov/26974007/
71. Y Y, H M, S S-Y, Z Z, Y W, X L, et al. Investigating the shared genetic architecture between multiple sclerosis and inflammatory bowel diseases. Nat Commun [Internet]. 2021 Sep 24 [cited 2021 Oct 1];12(1):5641. Available from: https://pubmed.ncbi.nlm.nih.gov/34561436/
72. CY L, TM Y, RW O, QQ W, HF S. Genome-wide genetic links between amyotrophic lateral sclerosis and autoimmune diseases. BMC Med [Internet]. 2021 Dec 1 [cited 2021 Oct 12];19(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33541344/
73. Hui KY, Fernandez-Hernandez H, Hu J, Schaffner A, Pankratz N, Hsu NY, et al. Functional variants in the LRRK2 gene confer shared effects on risk for Crohn's disease and Parkinson's disease. Sci Transl Med [Internet]. 2018 Jan 10 [cited 2021 Dec 4];10(423). Available from: https://pubmed.ncbi.nlm.nih.gov/29321258/
74. Babb De Villiers C, Kroese M, Moorthie S. Understanding polygenic models, their development and the potential application of polygenic scores in healthcare. J Med Genet [Internet]. 2020 Nov 1 [cited 2022 Jan 5];57(11):725–32. Available from: https://pubmed.ncbi.nlm.nih.gov/32376789/
75. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc [Internet]. 2020 Sep 1 [cited 2022 Apr 7];15(9):2759–72. Available from: https://pubmed.ncbi.nlm.nih.gov/32709988/
76. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet [Internet]. 2013 Apr [cited 2021 Dec 11];45(4):400–5. Available from: https://pubmed.ncbi.nlm.nih.gov/23455638/
77. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet [Internet]. 2017 Apr 6 [cited 2022 Jan 5];100(4):635–49. Available from: https://pubmed.ncbi.nlm.nih.gov/28366442/
78. Weersma RK, Stokkers PCF, Van Bodegraven AA, Van Hogezand RA, Verspaget HW, De Jong DJ, et al. Molecular prediction of disease risk and severity in a large Dutch Crohn's disease cohort. 2008 [cited 2022 Apr 25]; Available from: http://gut.bmj.com/
79. Cleynen I, Boucher G, Jostins L, Schumm LP, Zeissig S, Ahmad T, et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. Lancet (London, England) [Internet]. 2016 Jan 9 [cited 2021 Dec 4];387(10014):156–67. Available from: https://pubmed.ncbi.nlm.nih.gov/26490195/
80. Chen G-B, Lee SH, Montgomery GW, Wray NR, Visscher PM, Gearry RB, et al. Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. BMC Med Genet [Internet]. 2017 Aug 29 [cited 2021 Nov 5];18(1). Available from: https://pubmed.ncbi.nlm.nih.gov/28851283/
81. Voskuil MD, Spekhorst LM, van der Sloot KWJ, Jansen BH, Dijkstra G, van der Woude CJ, et al. Genetic Risk Scores Identify Genetic Aetiology of Inflammatory Bowel Disease Phenotypes. J Crohns Colitis [Internet]. 2021 Jun 1 [cited 2021 Nov 5];15(6):930–7. Available from: https://pubmed.ncbi.nlm.nih.gov/33152062/
82. Khera A V, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet [Internet]. 2018 Sep 1 [cited 2021 Nov 5];50(9):1219–24. Available from: https://pubmed.ncbi.nlm.nih.gov/30104762/
83. Hübenthal M, Löscher B-S, Erdmann J, Franke A, Gola D, König IR, et al. Current Developments of Clinical Sequencing and the Clinical Utility of Polygenic Risk Scores in Inflammatory Diseases. Front Immunol [Internet]. 2021 Jan 29 [cited 2021 Nov 5];11. Available from: https://pubmed.ncbi.nlm.nih.gov/33633722/
84. Lu T, Zhou S, Wu H, Forgetta V, Greenwood CMT, Richards JB. Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. Genet Med [Internet]. 2021 Mar 1 [cited 2021 Nov 29];23(3):508–15. Available from: https://pubmed.ncbi.nlm.nih.gov/33110269/
85. Lee SH, Kwon J eun, Cho M La. Immunological pathogenesis of inflammatory bowel disease. Intest Res [Internet]. 2018 [cited 2022 Jan 13];16(1):26–42. Available from: https://pubmed.ncbi.nlm.nih.gov/29422795/

86. Guan Q. A Comprehensive Review and Update on the Pathogenesis of Inflammatory Bowel Disease. J Immunol Res [Internet]. 2019 [cited 2022 Jan 12];2019. Available from: https://pubmed.ncbi.nlm.nih.gov/31886308/
87. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature [Internet]. 2007 Jun 7 [cited 2022 Jan 21];447(7145):661–78. Available from: https://pubmed.ncbi.nlm.nih.gov/17554300/
88. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet [Internet]. 2007 Jul [cited 2021 Oct 7];39(7):830–2. Available from: https://pubmed.ncbi.nlm.nih.gov/17554261/
89. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nat Genet [Internet]. 2009 Dec [cited 2022 Jan 21];41(12):1330–4. Available from: https://pubmed.ncbi.nlm.nih.gov/19915572/
90. Piovani D, Danese S, Peyrin-Biroulet L, Nikolopoulos GK, Lytras T, Bonovas S. Environmental Risk Factors for Inflammatory Bowel Diseases: An Umbrella Review of Meta-analyses. Gastroenterology [Internet]. 2019 Sep 1 [cited 2022 May 7];157(3):647-659.e4. Available from: https://pubmed.ncbi.nlm.nih.gov/31014995/
91. Rizzello F, Spisni E, Giovanardi E, Imbesi V, Salice M, Alvisi P, et al. Implications of the Westernized Diet in the Onset and Progression of IBD. 2019; Available from: www.mdpi.com/journal/nutrients
92. Wark G, Samocha-Bonet D, Ghaly S, Danta M. The Role of Diet in the Pathogenesis and Management of Inflammatory Bowel Disease: A Review. Nutr 2021, Vol 13, Page 135 [Internet]. 2020 Dec 31 [cited 2022 May 7];13(1):135. Available from: https://www.mdpi.com/2072-6643/13/1/135/htm
93. Mahid SS, Minor KS, Soto RE, Hornung CA, Galandiuk S. Smoking and inflammatory bowel disease: a meta-analysis. Mayo Clin Proc [Internet]. 2006 [cited 2022 May 7];81(11):1462–71. Available from: https://pubmed.ncbi.nlm.nih.gov/17120402/
94. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther [Internet]. 2011 Feb 1 [cited 2022 Feb 25];13(1). Available from: https://pubmed.ncbi.nlm.nih.gov/21345260/
95. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet [Internet]. 2016 Oct 1 [cited 2022 May 15];48(10):1284–7. Available from: https://pubmed.ncbi.nlm.nih.gov/27571263/
96. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet [Internet]. 2007 [cited 2022 May 18];81(3):559–75. Available from: https://pubmed.ncbi.nlm.nih.gov/17701901/
97. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics [Internet]. 2015 May 1 [cited 2022 May 18];31(9):1466–8. Available from: https://pubmed.ncbi.nlm.nih.gov/25550326/
98. PRSice - PRSice-2 [Internet]. [cited 2022 Apr 30]. Available from: https://www.prsice.info/step_by_step/#prs-calculation
99. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet [Internet]. 2018 Sep 1 [cited 2022 Apr 27];50(9):1335. Available from: /pmc/articles/PMC6119127/
100. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res [Internet]. 2010 Jul 3 [cited 2022 Feb 28];38(16). Available from: https://pubmed.ncbi.nlm.nih.gov/20601685/
101. Stittrich AB, Ashworth J, Shi M, Robinson M, Mauldin D, Brunkow ME, et al. Genomic architecture of inflammatory bowel disease in five families with multiple affected individuals. Hum genome Var [Internet]. 2016 Dec [cited 2021 Dec 2];3(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27081563/
102. Borren NZ, Conway G, Garber JJ, Khalili H, Budree S, Mallick H, et al. Differences in Clinical Course, Genetics, and the Microbiome Between Familial and Sporadic Inflammatory Bowel Diseases. J Crohns Colitis [Internet]. 2018 Apr 27 [cited 2022 Apr 3];12(5):525–31. Available from: https://pubmed.ncbi.nlm.nih.gov/29145572/
103. ER S, M F, N B-Y, BE A, F S, N P, et al. Rare coding variant analysis in a large cohort of Ashkenazi Jewish families with inflammatory bowel disease. Hum Genet [Internet]. 2018 Sep 1 [cited 2021 Nov 5];137(9):723–34. Available from: https://pubmed.ncbi.nlm.nih.gov/30167848/
104. Zhernakova A, Festen EM, Franke L, Trynka G, van Diemen CC, Monsuur AJ, et al. Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring

CARD9 and IL18RAP. Am J Hum Genet [Internet]. 2008 May 9 [cited 2022 May 14];82(5):1202–10. Available from: https://pubmed.ncbi.nlm.nih.gov/18439550/

105. Joosse ME, Charbit-Henrion F, Boisgard R, Raatgeep R (H ). C, Lindenbergh-Kortleve DJ, Costes LMM, et al. Duplication of the IL2RA locus causes excessive IL-2 signaling and may predispose to very early onset colitis. Mucosal Immunol [Internet]. 2021 Sep 1 [cited 2022 May 14];14(5):1172–82. Available from: https://pubmed.ncbi.nlm.nih.gov/34226674/
106. Zhang H, Cui Z, Cheng D, Du Y, Guo X, Gao R, et al. RNF186 regulates EFNB1 (ephrin B1)-EPHB2-induced autophagy in the colonic epithelial cells for the maintenance of intestinal homeostasis. Autophagy [Internet]. 2021 [cited 2022 May 14];17(10):3030. Available from: /pmc/articles/PMC8525924/
107. Giorgio C, Allodi M, Palese S, Grandi A, Tognolini M, Castelli R, et al. UniPR1331: Small Eph/Ephrin Antagonist Beneficial in Intestinal Inflammation by Interfering with Type-B Signaling. Pharmaceuticals (Basel) [Internet]. 2021 [cited 2022 May 18];14(6). Available from: https://pubmed.ncbi.nlm.nih.gov/34074058/
108. Menzel K, Hausmann M, Obermeier F, Schreiter K, Dunger N, Bataille F, et al. Cathepsins B, L and D in inflammatory bowel disease macrophages and potential therapeutic effects of cathepsin inhibition in vivo. Clin Exp Immunol [Internet]. 2006 Oct [cited 2022 May 14];146(1):169. Available from: /pmc/articles/PMC1809720/
109. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet [Internet]. 2012 Jul [cited 2022 Feb 28];44(7):821–4. Available from: https://pubmed.ncbi.nlm.nih.gov/22706312/
110. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet 2021 537 [Internet]. 2021 May 20 [cited 2022 May 18];53(7):1097–103. Available from: https://www.nature.com/articles/s41588-021-00870-7
111. Schlafly A, Pfeiffer RM, Nagore E, Puig S, Calista D, Ghiorzo P, et al. Contribution of Common Genetic Variants to Familial Aggregation of Disease and Implications for Sequencing Studies. PLoS Genet [Internet]. 2019 [cited 2022 Feb 11];15(11). Available from: https://pubmed.ncbi.nlm.nih.gov/31730655/
112. Faith JJ, Colombel JF, Gordon JI. Identifying strains that contribute to complex diseases through the study of microbial inheritance. Proc Natl Acad Sci U S A [Internet]. 2015 Jan 20 [cited 2022 Feb 23];112(3):633–40. Available from: https://pubmed.ncbi.nlm.nih.gov/25576328/
113. Joossens M, Huys G, Cnockaert M, De Preter V, Verbeke K, Rutgeerts P, et al. Dysbiosis of the faecal microbiota in patients with Crohn's disease and their unaffected relatives. Gut [Internet]. 2011 May [cited 2022 Feb 23];60(5):631–7. Available from: https://pubmed.ncbi.nlm.nih.gov/21209126/
114. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics [Internet]. 2014 [cited 2022 May 18];9(10):1382–96. Available from: https://pubmed.ncbi.nlm.nih.gov/25424692/

# Appendix I: supplementary tables

Supplementary table 1: Group comparisons PRS based on IBD effect sizes of imputed Immunochip data

Supplementary table 2: Group comparisons PRS based on CD effect sizes of imputed Immunochip data

Supplementary table 3: Group comparisons PRS based on UC effect sizes of imputed Immunochip data

Supplementary table 4: Group comparisons PRS mean affected and unaffected PRS based on IBD effect sizes of imputed Immunochip data

Supplementary table 5: Group comparison mean affected and unaffected PRS based on CD effect sizes of imputed Immunochip data

Supplementary table 6: Group comparisons mean affected and unaffected PRS based on UC effect sizes of imputed Immunochip data

Supplementary table 7: Group comparisons PRS based on IBD effect sizes of unimputed Immunochip data

Supplementary table 8: Group comparisons PRS based on CD effect sizes of unimputed Immunochip data

Supplementary table 9: Group comparisons PRS based on UC effect sizes of unimputed Immunochip data

***Supplementary table 1: Group comparisons PRS based on IBD effect sizes of imputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.16e-4 | | |
| **Sporadic cases** | 0.21 | 3.60e-4 | |
| **Healthy controls** | 3.71e-18 | 3.05e-4 | 8.59e-56 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.23e-5 | | |
| **Sporadic cases** | 0.10 | 2.56e-5 | |
| **Healthy controls** | 4.48e-20 | 1.72e-3 | 4.66e-59 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.36e-5 | | |
| **Sporadic cases** | 7.79e-2 | 7.81e-5 | |
| **Healthy controls** | 8.05e-21 | 2.06e-4 | 2.11e-64 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.36e-5 | | |
| **Sporadic cases** | 5.79e-2 | 1.05e-4 | |
| **Healthy controls** | 1.09e-20 | 2.21e-4 | 7.13e-63 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.19e-5 | | |
| **Sporadic cases** | 3.58e-2 | 6.97e-4 | |
| **Healthy controls** | 1.24e-20 | 5.98e-5 | 1.04e-59 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.17e-5 | | |
| **Sporadic cases** | 6.85e-2 | 5.58e-5 | |
| **Healthy controls** | 1.25e-19 | 7.84e-4 | 3.69e-57 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). The imputed Immunochip data of cases (sporadic and affected), unaffected family members and healthy population controls were used to compute PRS. PRS were calculated based on the effect sizes of IBD and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

***Supplementary table 2: Group comparisons PRS based on CD effect sizes of imputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.47e-6 | | |
| **Sporadic cases** | 0.25 | 5.39e-7 | |
| **Healthy controls** | 7.56e-21 | 1.87e-3 | 4.13e-62 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.58e-06 | | |
| **Sporadic cases** | 0.16 | 1.03e-06 | |
| **Healthy controls** | 7.12e-23 | 1.68e-4 | 3.50e-67 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 8.16e-07 | | |
| **Sporadic cases** | 0.92 | 5.13e-10 | |
| **Healthy controls** | 2.59e-18 | 8.64e-3 | 5.62e-70 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.13e-5 | | |
| **Sporadic cases** | 0.96 | 8.56e-8 | |
| **Healthy controls** | 5.21e-17 | $8.61^{e}$-4 | 2.62e-66 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.09e-5 | | |
| **Sporadic cases** | 0.69 | 2.36e-7 | |
| **Healthy controls** | 5.13e-18 | 4.37e-4 | 3.13e-65 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.39e-05 | | |
| **Sporadic cases** | 0.70 | 6.56e-8 | |
| **Healthy controls** | 4.42e-17 | 2.05e-3 | 5.57e-62 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). The imputed Immunochip data of CD cases (sporadic and affected), unaffected family members of CD and mixed families and all healthy population controls were used to compute PRS. PRS were calculated based on the effect sizes of CD and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

***Supplementary table 3: Group comparisons PRS based on UC effect sizes of imputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.42e-3 | | |
| **Sporadic cases** | 0.16 | 1.42e-5 | |
| **Healthy controls** | 1.81e-6 | 0.82 | 7.36e-32 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.06-3 | | |
| **Sporadic cases** | 0.28 | 2.71e-6 | |
| **Healthy controls** | 2.27e-6 | 0.95 | 1.56e-37 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.48e-2 | | |
| **Sporadic cases** | 0.97 | 1.73e-05 | |
| **Healthy controls** | 1.07-4 | 0.61 | 3.13e-41 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.06e-2 | | |
| **Sporadic cases** | 0.55 | 9.97e-4 | |
| **Healthy controls** | 4.08e-05 | 0.23 | 2.06e-36 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.65e-2 | | |
| **Sporadic cases** | 0.88 | 1.40e-4 | |
| **Healthy controls** | 2.74e-4 | 0.51 | 1.77e-35 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 7.43e-2 | | |
| **Sporadic cases** | 0.91 | 4.66e-4 | |
| **Healthy controls** | 7.99e-4 | 0.33 | 2.68e-35 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). The imputed Immunochip data of UC cases (sporadic and affected), unaffected family members of UC and mixed families and all healthy population controls were used to compute PRS. PRS were calculated based on the effect sizes of UC and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

***Supplementary table 4: Group comparisons PRS mean affected and unaffected PRS based on IBD effect sizes of imputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.30e-2 | | |
| **Sporadic cases** | 0.54 | 2.46-2 | |
| **Healthy controls** | 9.05e-8 | 3.08-2 | 9.98e-57 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 7.63-3 | | |
| **Sporadic cases** | 0.52 | 1.85e-2 | |
| **Healthy controls** | 4.30e-8 | 2.91e-2 | 1.34e-59 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.28e-3 | | |
| **Sporadic cases** | 0.61 | 2.29e-3 | |
| **Healthy controls** | 8.80e-08 | 9.42e-2 | 7.78e-64 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.26-3 | | |
| **Sporadic cases** | 0.568 | 3.39e-3 | |
| **Healthy controls** | 1.01e-07 | 0.0811-2 | 4.79e-63 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.20e-3 | | |
| **Sporadic cases** | 0.50 | 7.66e-3 | |
| **Healthy controls** | 1.15e-07 | 6.27e-2 | 3.41e-60 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.17e-3 | | |
| **Sporadic cases** | 0.58 | 2.48e-3 | |
| **Healthy controls** | 2.75e-7 | 0.15 | 6.12e-58 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). For the affected and unaffected family members, the mean PRS of all affected or unaffected family members was calculated and used in this analysis. The imputed Immunochip data of cases (sporadic and affected), unaffected family members and healthy population controls were used to compute PRS. The PRS were calculated based on the effect sizes of IBD and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

***Supplementary table 5: Group comparison mean affected and unaffected PRS based on CD effect sizes of imputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.91e-2 | | |
| **Sporadic cases** | 3.51e-2 | 0.43 | |
| **Healthy controls** | 3.37e-10 | 1.82e-3 | 1.53e-44 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 8.45e-3 | | |
| **Sporadic cases** | 4.37e-2 | 0.34 | |
| **Healthy controls** | 1.16e-10 | 1.31e-3 | 9.17e-50 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 3.14e-3 | | |
| **Sporadic cases** | 0.16 | 5.39e-2 | |
| **Healthy controls** | 6.95e-09 | 2.42e-2 | 2.62e-52 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.04e-2 | | |
| **Sporadic cases** | 0.13 | 0.16 | |
| **Healthy controls** | 1.09e-8 | 7.51e-3 | 3.72e-51 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 9.11e-3 | | |
| **Sporadic cases** | 9.78e-2 | 0.16 | |
| **Healthy controls** | 5.95e-9 | 7.86e-3 | 4.07e-50 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 3.71e-3 | | |
| **Sporadic cases** | 0.11 | 0.11 | |
| **Healthy controls** | 1.56e-08 | 1.98e-2 | 1.80e-47 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). For the affected and unaffected family members, the mean PRS of all affected or unaffected family members was calculated and used in this analysis. The imputed Immunochip data of CD cases (sporadic and affected), unaffected family members of CD and mixed families and healthy population controls were used to compute PRS. The PRS were calculated based on the effect sizes of CD and SNPs with MAF > 0.01. $p < 1.39e-3$ is considered significant.*

***Supplementary table 6: Group comparisons mean affected and unaffected PRS based on UC effect sizes of imputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 0.21 | | |
| **Sporadic cases** | 0.51 | 0.38 | |
| **Healthy controls** | 2.03e-2 | 0.56 | 1.51e-19 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 8.31e-2 | | |
| **Sporadic cases** | 0.33 | 0.31 | |
| **Healthy controls** | 4.83e-3 | 0.55 | 1.47e-23 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 8.39e-2 | | |
| **Sporadic cases** | 0.61 | 0.20 | |
| **Healthy controls** | 9.60e-3 | 0.53 | 6.24e-32 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 5.93e-2 | | |
| **Sporadic cases** | 0.40 | 0.23 | |
| **Healthy controls** | 4.98e-3 | 0.50 | 4.77e-31 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.69e-2 | | |
| **Sporadic cases** | 0.52 | 0.13 | |
| **Healthy controls** | 8.58e-3 | 0.70 | 1.90e-31 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.02e-2 | | |
| **Sporadic cases** | 0.49 | 0.21 | |
| **Healthy controls** | 7.35e-3 | 0.54 | 1.01e-30 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). For the affected and unaffected family members, the mean PRS of all affected or unaffected family members was calculated and used in this analysis. The imputed Immunochip data of UC cases (sporadic and affected), unaffected family members of UC and mixed families and healthy population controls were used to compute PRS. The PRS were calculated based on the effect sizes of UC and SNPs with MAF > 0.01. $p < 1.39e-3$ is considered significant.*

***Supplementary table 7: Group comparisons PRS based on IBD effect sizes of unimputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.52e-4 | | |
| **Sporadic cases** | 0.73 | 1.27e-4 | |
| **Healthy controls** | 4.39e-15 | 3.21e-3 | 2.94e-54 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.77e-4 | | |
| **Sporadic cases** | 0.81 | 2.09e-5 | |
| **Healthy controls** | 1.88e-16 | 3.90e-3 | 2.28e-59 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.92e-05 | | |
| **Sporadic cases** | 0.40 | 9.30e-06 | |
| **Healthy controls** | 3.65e-19 | 2.00e-3 | 8.397e-66 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 4.06e-5 | | |
| **Sporadic cases** | 0.45 | 1.16e-05 | |
| **Healthy controls** | 3.12e-19 | 1.13e-3 | 3.63e-68 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 7.98e-05 | | |
| **Sporadic cases** | 0.26 | 1.01e-4 | |
| **Healthy controls** | 2.94e-20 | 2.19e-4 | 6.40e-67 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.74e-5 | | |
| **Sporadic cases** | 0.15 | 1.54e-4 | |
| **Healthy controls** | 2.00e-20 | 2.54e-4 | 3.72e-65 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). The unimputed Immunochip data of cases (sporadic and affected), unaffected family members and healthy population controls were used to compute PRS. PRS were calculated based on the effect sizes of IBD and SNPs with MAF > 0.01. $p < 1.39e-3$ is considered significant.*

***Supplementary table 8: Group comparisons PRS based on CD effect sizes of unimputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.18e-5 | | |
| **Sporadic cases** | 4.56e-2 | 1.70e-4 | |
| **Healthy controls** | 3.58e-20 | 3.05e-4 | 2.88e-57 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 3.53e-5 | | |
| **Sporadic cases** | 6.67e-3 | 1.45e-3 | |
| **Healthy controls** | 1.73e-22 | 1.31e-5 | 4.90e-59 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.29e-6 | | |
| **Sporadic cases** | 0.11 | 3.02e-6 | |
| **Healthy controls** | 6.26e-21 | 3.58e-4 | 2.53e-66 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.55e-5 | | |
| **Sporadic cases** | 0.13 | 2.69e-5 | |
| **Healthy controls** | 1.94e-20 | 8.80e-5 | 3.08e-65 |

**pT = 0.1**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.54e-5 | | |
| **Sporadic cases** | 6.05e-2 | 5.17e-5 | |
| **Healthy controls** | 1.96e-21 | 4.15e-5 | 1.10e-65 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.05e-05 | | |
| **Sporadic cases** | 7.80e-2 | 7.01e-5 | |
| **Healthy controls** | 1.16e-20 | 3.49e-5 | 1.49e-64 |

*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). The unimputed Immunochip data of CD cases (sporadic and affected), unaffected family members of CD and mixed families and all healthy population controls were used to compute PRS. PRS were calculated based on the effect sizes of CD and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

***Supplementary table 9: Group comparisons PRS based on UC effect sizes of unimputed Immunochip data***

**pT = $5x10^{-8}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 2.94e-3 | | |
| **Sporadic cases** | 0.63 | 9.73e-6 | |
| **Healthy controls** | 7.08e-4 | 0.40 | 2.00e-36 |

**pT = $1x10^{-5}$**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 3.16e-3 | | |
| **Sporadic cases** | 0.90 | 5.37e-6 | |
| **Healthy controls** | 1.04e-3 | 0.55 | 1.61e-41 |

**pT = 0.01**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 3.51e-3 | | |
| **Sporadic cases** | 0.51 | 2.12e-7 | |
| **Healthy controls** | 2.67e-3 | 0.43 | 1.14-49 |

**pT = 0.05**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 6.70e-3 | | |
| **Sporadic cases** | 0.56 | 9.30e-6 | |
| **Healthy controls** | 2.54e-3 | 0.91 | 1.21e-49 |

**pT = 0.1**

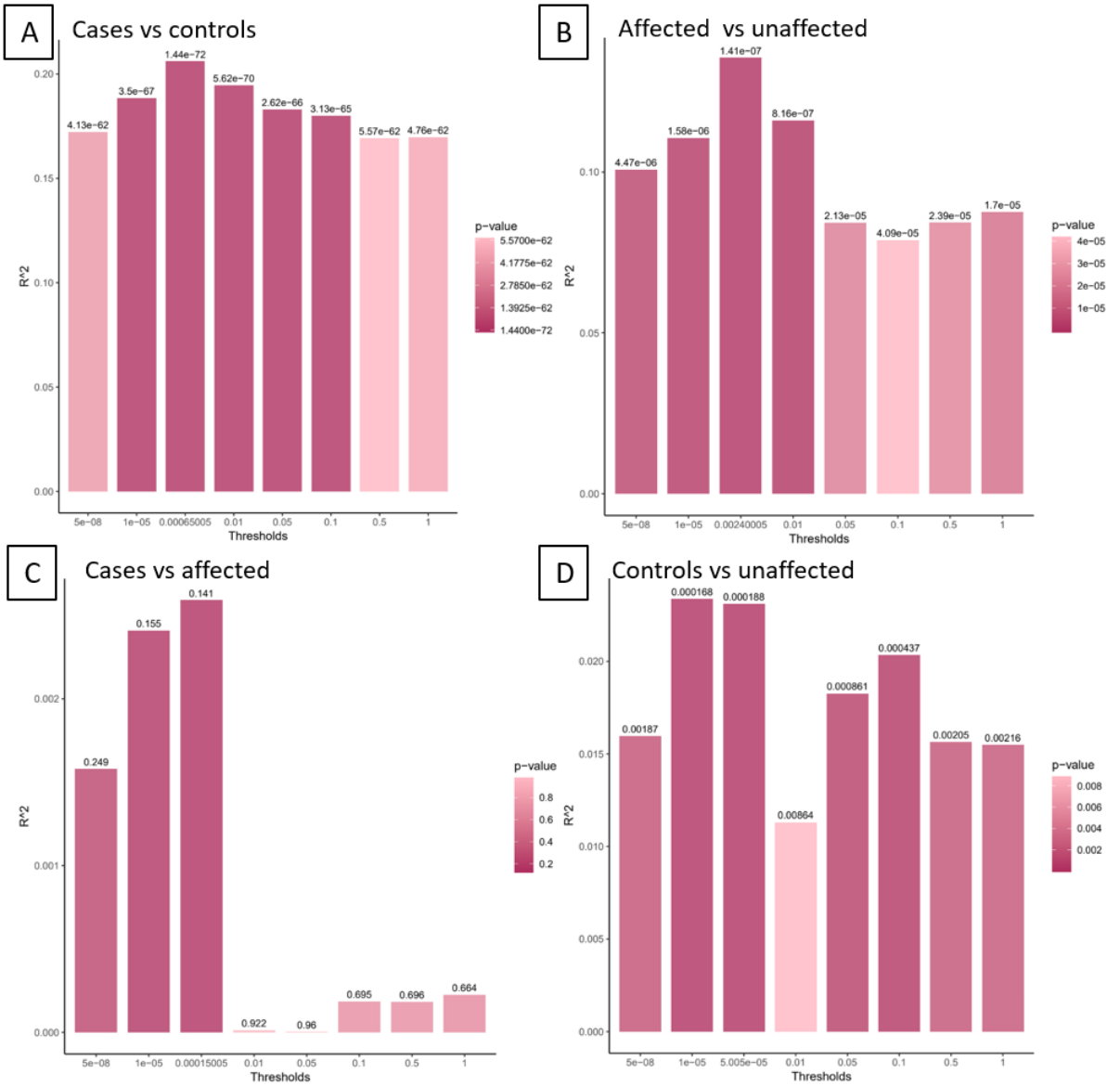
| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 9.91e-3 | | |
| **Sporadic cases** | 0.49 | 8.19e-6 | |
| **Healthy controls** | 5.16e-3 | 0.93 | 3.27e-47 |

**pT = 0.5**

| | Affected | Unaffected | Sporadic cases |
|---|---|---|---|
| **Unaffected** | 1.50e-2 | | |
| **Sporadic cases** | 0.44 | 4.47e-6 | |
| **Healthy controls** | 3.60e-3 | 0.97 | 2.12e-48 |

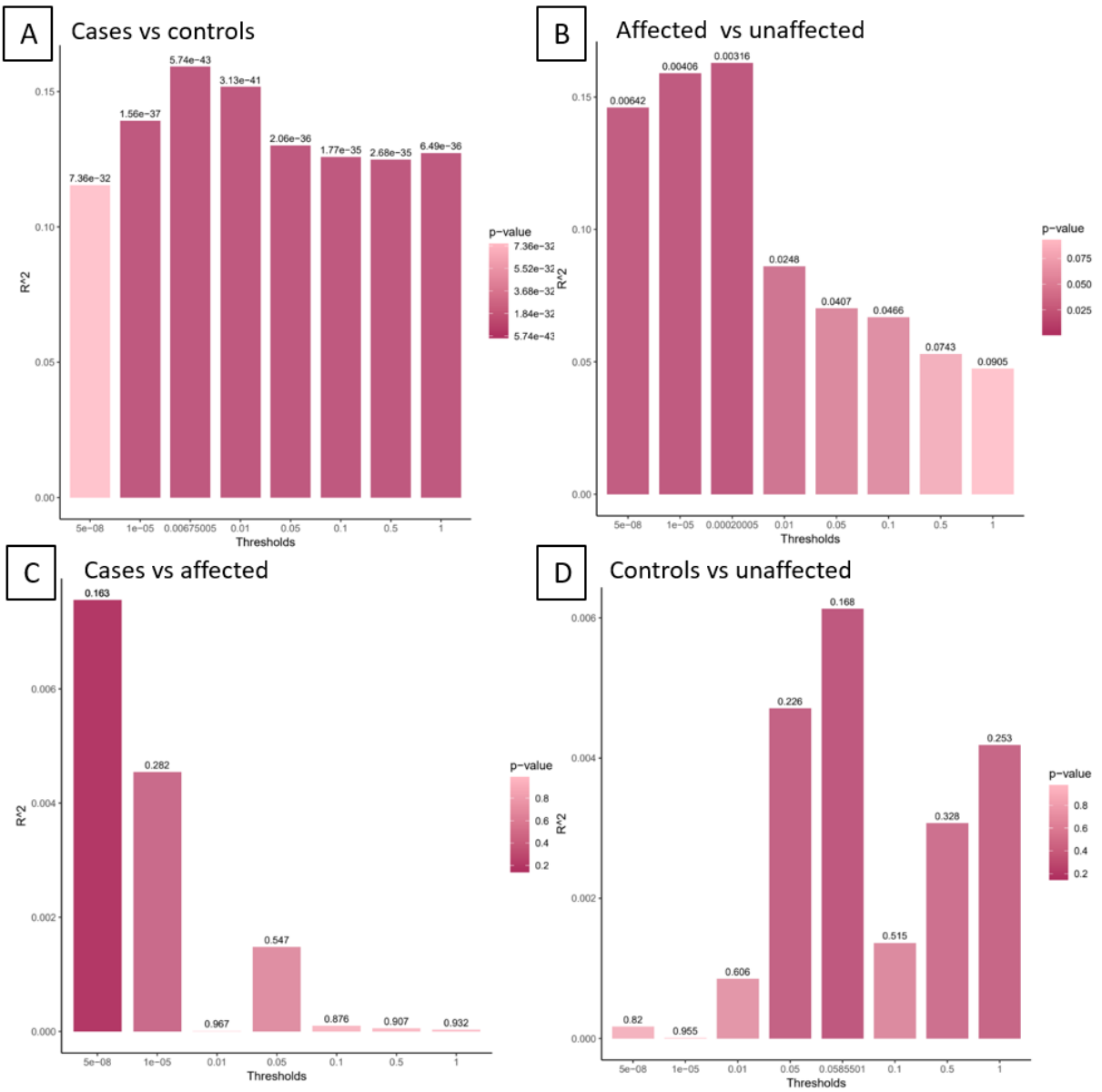
*The p-values here presented are of the PRS comparisons between the four groups: sporadic cases, healthy controls, affected family members (affected) and unaffected family members (unaffected). The unimputed Immunochip data of UC cases (sporadic and affected), unaffected family members of UC and mixed families and all healthy population controls were used to compute PRS. PRS were calculated based on the effect sizes of UC and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*

# Appendix II: supplementary figures



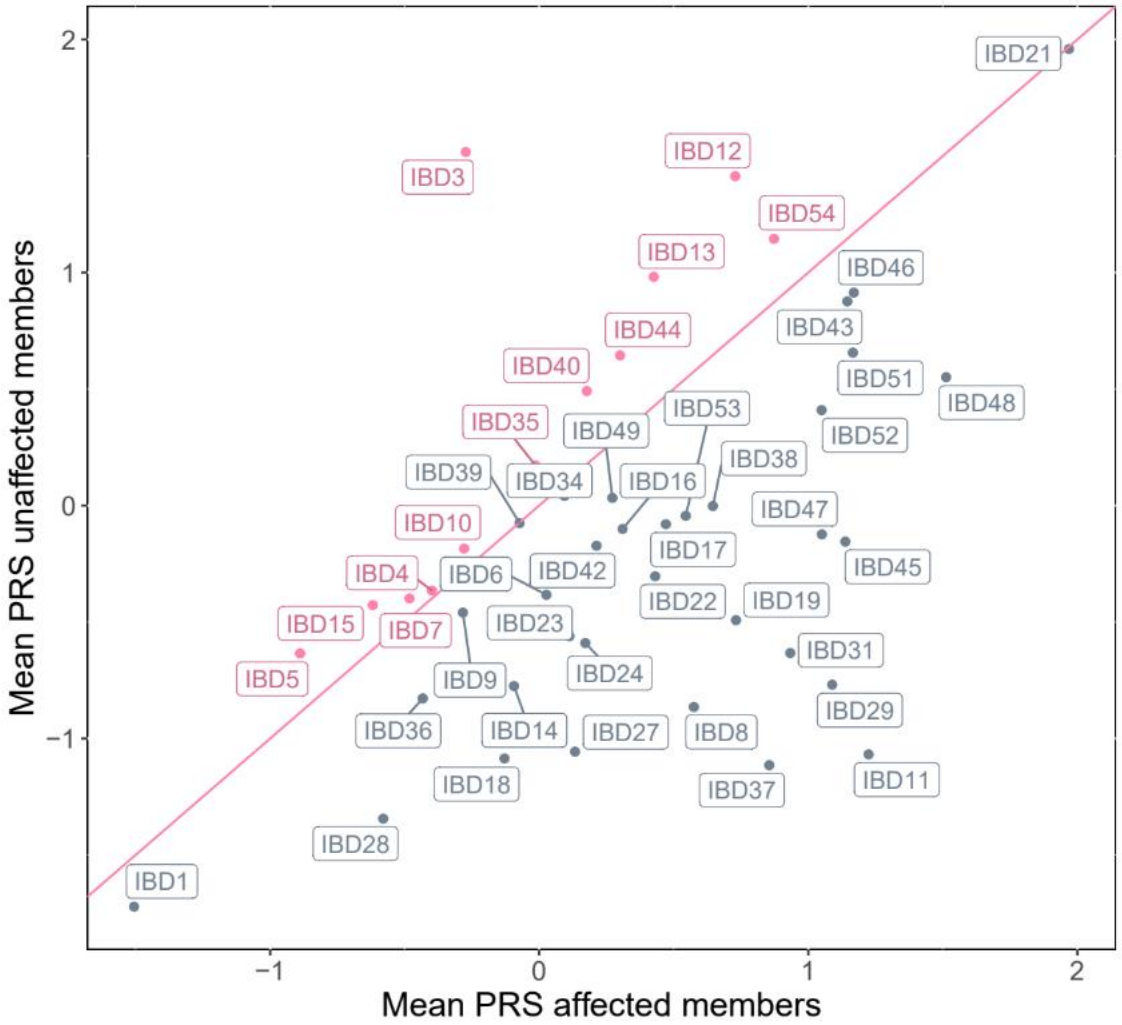

***Supplementary figure 1: Variance explained by each PRS (Crohn's disease)***

*Each plot represents a different comparison of the PRS between two groups: (A) Sporadic cases vs sporadic controls; (B) Affected vs unaffected family members; (C) Sporadic cases vs affected family members; (D) Sporadic controls vs unaffected family members. Only CD cases, affected and sporadic, and the unaffected family members of CD and mixed families are present in this analysis. All healthy controls are included. Each bar depicts a separate PRS including SNPs based on different p-value thresholds. The height of the bars indicates the $R^2$ of the PRS in a logistic regression model. The p-value of the $R^2$ is represented by the colour of the bars, a darker colour indicates a more significant p-value. PRS were calculated based on the effect sizes of CD and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*



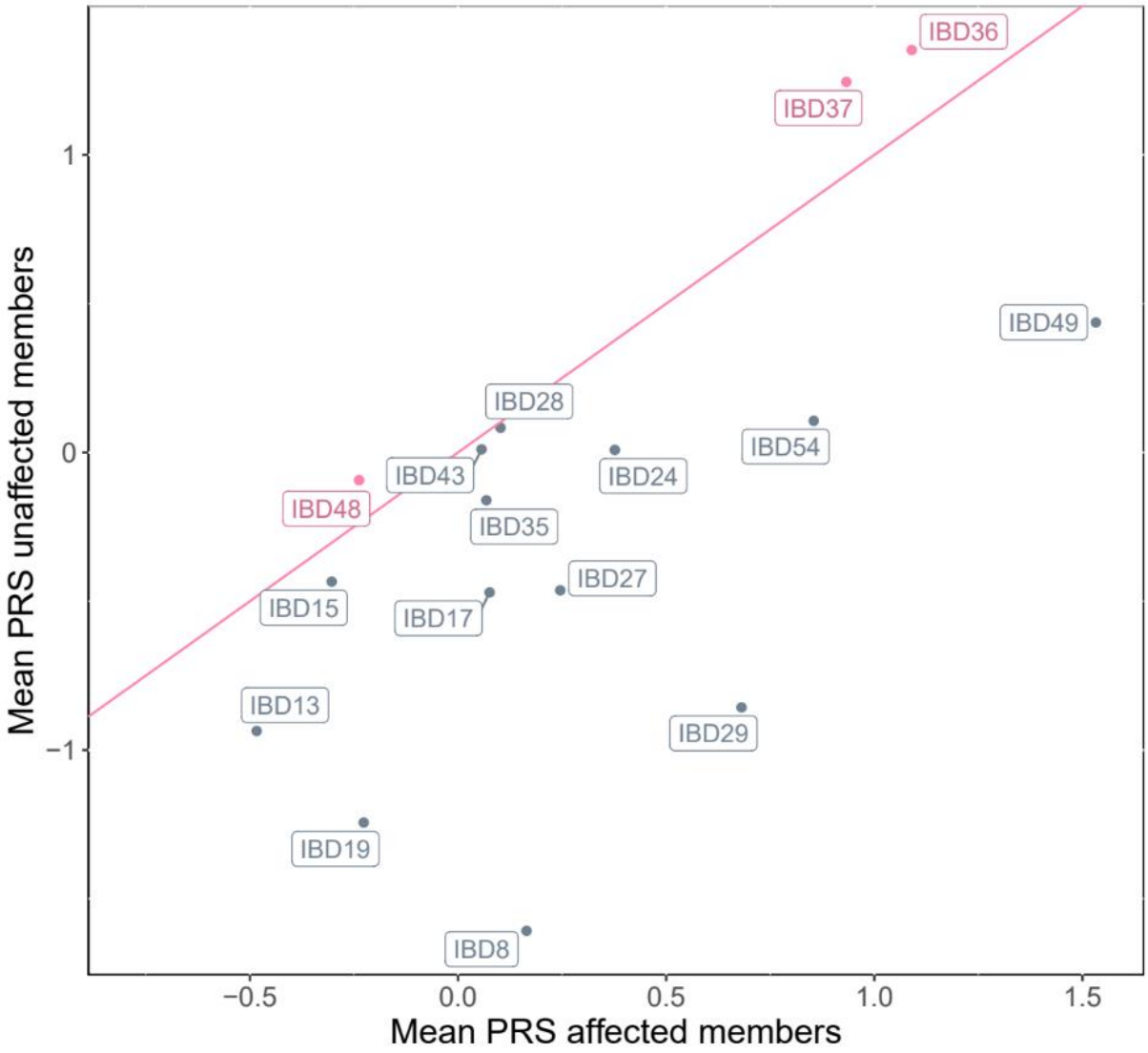

***Supplementary figure 2: Variance explained by each PRS (ulcerative colitis)***

*Each plot represents a different comparison of the PRS between two groups: (A) Sporadic cases vs sporadic controls; (B) Affected vs unaffected family members; (C) Sporadic cases vs affected family members; (D) Sporadic controls vs unaffected family members. Only UC cases, affected and sporadic, and the unaffected family members of UC and mixed families are present in this analysis. All healthy controls are included. Each bar depicts a separate PRS including SNPs based on different p-value thresholds. The height of the bars indicates the $R^2$ of the PRS in a logistic regression model. The p-value of the $R^2$ is represented by the colour of the bars, a darker colour indicates a more significant p-value. PRS were calculated based on the effect sizes of UC and SNPs with MAF > 0.01. p < 1.39e-3 is considered significant.*




***Supplementary figure 3: Mean PRS CD affected vs unaffected family members of CD and mixed families***

*The mean PRS of all affected members of a CD or mixed family (x-axis) is plotted against the mean PRS of all unaffected members of the same family (y-axis). The diagonal line is the x = y line. Every family with a higher mean PRS for the unaffected than the affected family members is coloured pink. PRS were calculated based on the effect sizes of CD, pT = 0.01 and MAF = 0.01.*




***Supplementary figure 4: Mean PRS UC affected vs unaffected family members of UC and mixed families***

*The mean PRS of all affected members of a UC or mixed family (x-axis) is plotted against the mean PRS of all unaffected members of the same family (y-axis). The diagonal line is the x = y line. Every family with a higher mean PRS for the unaffected than the affected family members is coloured pink. PRS were calculated based on the effect sizes of UC, pT = 0.01 and MAF = 0.01.*