Genetic risk scores identify genetic aetiology of inflammatory bowel disease phenotypes

Parelsnoer Institute and the Dutch Initiative on Crohn and Colitis; Voskuil, M D; Spekhorst, L M; van der Sloot, K W J; Jansen, B H; Dijkstra, G; van der Woude, C J; Hoentjen, F; Pierik, M J; van der Meulen, A E

Published in:
Journal of Crohn's and Colitis

DOI:
10.1093/ecco-jcc/jjaa223

Publication date:
2020

Manuscript Doi: 10.1093/ecco-jcc/jjaa223

Genetic risk scores identify genetic aetiology of inflammatory bowel disease phenotypes

M.D. Voskuil[1,2], L.M. Spekhorst[1], K.W.J. van der Sloot[1,3], B. H. Jansen[1], G. Dijkstra[1], C.J. van der Woude[4], F. Hoentjen[5], M.J. Pierik[6], A.E. van der Meulen[7], N.K.H. de Boer[8], M. Löwenberg[9], B. Oldenbrug[10], E.A.M. Festen[1,2*], R.K. Weersma[1*], Parelsnoer Institute and the Dutch Initiative on Crohn and Colitis

* Shared last authors

1. Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands; 2. Department of Genetics, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands; 3. Department of Epidemiology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands; 4. Department of Gastroenterology and Hepatology, Erasmus Medical Centre, Rotterdam, the Netherlands; 5. Department of Gastroenterology and Hepatology, Radboud University Medical Centre, Nijmegen, the Netherlands; 6. Department of Gastroenterology and Hepatology, Maastricht University Medical Centre+, Maastricht, the Netherlands; 7. Department of Gastroenterology and Hepatology, Leiden University Medical Centre, Leiden, the Netherlands; 8. Department of Gastroenterology and Hepatology, Amsterdam University Medical Centres, Vrije Universiteit Amsterdam, AG&M research institute, Amsterdam, the Netherlands; 9. Department of Gastroenterology and Hepatology, Amsterdam University Medical Centres, Amsterdam, the Netherlands; 10. Department of Gastroenterology and Hepatology, University Medical Centre Utrecht, Utrecht, the Netherlands

Address for correspondence

Prof. Rinse K. Weersma, Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, PO Box 30.001, Hanzeplein 1, 9700 RB Groningen, the Netherlands, E: r.k.weersma@umcg.nl

Contact details

M.D. Voskuil m.d.voskuil@umcg.nl

L.M. Spekhorst l.m.spekhorst@umcg.nl

K.W.J. van der Sloot k.w.j.van.der.sloot@umcg.nl

B.H. Jansen b.h.jansen@umcg.nl

G. Dijkstra gerard.dijkstra@umcg.nl

C.J. van der Woude c.vanderwoude@erasmus.nl

F. Hoentjen Frank.Hoentjen@radboudumc.nl

M.J. Pierik m.pierik@mumc.nl

A.E. van der Meulen ae.meulen@lumc.nl

N.K.H. de Boer KHN.deBoer@vumc.nl

M. Löwenberg m.lowenberg@amc.uva.nl

B. Oldenburg boldenbu@umcutrecht.nl

E.A.M. Festen e.a.m.festen@umcg.nl

R.K. Weersma r.k.weersma@umcg.nl

Abbreviations used in the manuscript

CD	Crohn's disease
CDprog	Crohn's disease prognosis (as defined by Lee et al)[25]
FDR	false discovery rate
IBD	inflammatory bowel disease
GRS	genetic risk scores
GWAS	genome-wide association study
PSC	primary sclerosing cholangitis
SNP	single nucleotide polymorphism
UC	ulcerative colitis

ABSTRACT

Background and Aims:

Inflammatory bowel disease (IBD) phenotypes are very heterogeneous between patients, and current clinical and molecular classifications do not accurately predict the course that IBD will take over time. Genetic determinants of disease phenotypes remain largely unknown but could aid drug development and allow for personalised management. We used genetic risk scores (GRS) to disentangle the genetic contributions to IBD phenotypes.

Methods:

Clinical characteristics and imputed genome-wide genetic array data of patients with IBD were obtained from two independent cohorts (cohort A, n=1,097; cohort B, n=2,156). Genetic risk scoring was used to assess genetic aetiology shared across traits and IBD phenotypes. Significant GRS–phenotype (FDR corrected P<.05) associations identified in cohort A were put forward for replication in cohort B.

Results:

Crohn's disease (CD) GRS were associated with fibrostenotic CD (R^2=7.4%, FDR=.02) and ileocaecal resection (R^2=4.1%, FDR=1.6E-03), and this remained significant after correcting for previously identified clinical and genetic risk factors. Ulcerative colitis (UC) GRS (R^2=7.1%, FDR=.02) and primary sclerosing cholangitis (PSC) GRS (R^2=3.6%; FDR=.03) were associated with colonic CD, and these two associations were largely driven by genetic variation in *MHC.* We also observed pleiotropy between PSC genetic risk and smoking behaviour (R^2=1.7%; FDR=.04).

Conclusions:

Patients with a higher genetic burden of CD are more likely to develop fibrostenotic disease and undergo ileocaecal resection, while colonic CD shares genetic aetiology with PSC and UC that is largely driven by variation in *MHC*. These results further our understanding of specific IBD phenotypes.

Key words

Genetics, inflammatory bowel disease, phenotypes

INTRODUCTION

Inflammatory bowel disease (IBD), with ulcerative colitis (UC) and Crohn's disease (CD) as its two major forms, is a chronic, relapsing immune-mediated disease characterised by inflammation and ulceration of the gut mucosa. Patients with UC have continuous inflammation limited to the mucosal layer of the colon. In CD, the inflammation is discontinuous, may occur anywhere in the gastrointestinal tract, and involves all layers of the gut[1,2].

The clinical course of IBD is highly unpredictable and heterogeneous. Some patients have long periods of disease remission that do not even require therapy, but a significant proportion of patients experience frequent relapse of inflammation or progress to complicated CD disease behaviour such as fibrostenotic or penetrating disease[1,2]. This latter group often requires treatment escalation with potent immunosuppressive therapy, hospitalisation, or surgical resection. However, current clinical classifications cannot accurately predict IBD disease course[3].

Genome-wide association studies (GWAS) have identified around 240 independent genetic susceptibility loci for IBD and have implicated genes involved in autophagy, T-cell response and bacterial handling as important contributors to the development of IBD[4,5]. IBD also shares genetic aetiology with other immune-mediated diseases such as ankylosing spondylitis and celiac disease[6,7]. The heterogeneous character of IBD suggests that different biological mechanisms lead to inflammation, and subgroups of patients may have different effector mechanisms contributing to their disease phenotypes. Identification of these patient-specific biological mechanisms could aid drug development and allow for personalised diagnostic work-up or treatment. Although similar patterns of disease phenotypes have been observed within families, genetic determinants of these clinical aspects of disease remain largely unknown outside of their role in disease susceptibility[8-10].

Genetic risk scores (GRS) aggregate the effects of the thousands of trait-associated genetic variants discovered by GWAS. By combining the effects of many genetic variants with small effect sizes, GRS are powerful tools to identify genetic contributions to phenotypes[11,12]. GRS also have the potential to identify pleiotropic effects of genetic variants, which may aid drug discovery or drug repurposing. GRS may also help identify patients at risk for specific clinical aspects of IBD.

In this study, we performed a within-cases genotype–phenotype study using two independent cohorts of patients with IBD. We constructed GRS for thirteen traits, both related and unrelated to IBD, to reveal genetic determinants that contribute to IBD phenotypes. In contrast to previous genetic studies that used the Immunochip, we used a novel genome-wide genetic array data with the potential to capture regions of the genome not covered by the Immunochip[10].

METHODS

Phenotype data

For the discovery phase of this study, patients were included from the 1000IBD cohort (**cohort A**). 1000IBD consists of patients with IBD treated at the University Medical Center Groningen for whom detailed phenotypes are prospectively collected and multi-omics profiles are generated[13]. For the replication phase of this study, we included patients from the Dutch IBD biobank cohort (**cohort B**), a prospective nationwide biobank of patients with IBD[14]. To ensure that both cohorts were independent, patients included in cohort A were excluded from cohort B. In both cohorts, each patient was diagnosed with IBD by his or her gastroenterologist using endoscopic data, histological data, radiological data, or a combination of these, and phenotyped according to the Montreal classification[15]. For each patient, their Montreal classification, surgical history (ileocaecal resection or colectomy), presence of extra-intestinal manifestations, PSC status, and smoking status were dichotomised into binary phenotypes. Only non-missing phenotype data were used, and missing data were not imputed. Cohorts A and B were compared using either the Wilcoxon rank sum or the chi-square test. The ethical boards of each separate recruiting centre approved the study, and all patients included in this study gave written informed consent.

Genotype data

All patients were genotyped using the Global Screening Array (Infinium Global Screening Array, Illumina, San Diego, CA, USA; **Supplementary Methods**), as previously described[13,14]. In short, the Global Screening Array is a genotyping platform including over 700,000 genetic variants that comprises a multi-ethnic genome-wide backbone combined with content derived from exome-sequencing studies and meta-analyses of several phenotype-specific consortia, including the International IBD Genetics Consortium. Extensive pre-imputation quality control was performed on

the genotype data (**Supplementary Methods**), and, after pre-phasing with the Eagle2 algorithm, genetic data were imputed to the Haplotype Reference Consortium reference panel using the Michigan Imputation server[16]. After post-imputation quality control measurements were performed, 12,130,010 genetic variants with a minor allele frequency > 0.1% remained. To limit bias from population stratification, only those patients with genetic data clustering with individuals from European ancestry were included, using the 1KG European dataset as the external reference panel[17].

Genetic variant phenotype associations

Genetic variants in *MST1*, *MHC*, and *NOD2* with known associations to age at diagnosis, CD disease location, CD disease behaviour, UC disease extent, and surgical history were selected from the imputed genetic data[10] (**Supplementary Table 1A**). In total, we identified 11 out of 13 previously described variants. The remaining two genetic variants (*MST1*, rs35261698; *MHC,* rs77005575) were excluded during quality control. In cohort A, we tested for genotype–phenotype associations with CD disease location (Montreal L1 vs. L2; L3 vs. L2), CD disease behaviour (Montreal B2 vs. B1; B3 vs. B1+B2), UC disease extent (Montreal E2 vs. E1; E3 vs. E1+E2), and surgical history (ileocaecal resection or colectomy). We performed logistic regression analyses in PLINK 1.9 (CoG Genomics)[18], adjusting for the covariates age and sex and for the first 5 principal components of the genetic data. To test for association with age at diagnosis (CD, UC, and IBD), we performed linear regression analyses in PLINK, adjusting for the same covariates as above. Since we only sought to replicate previously identified associations, genotype–phenotype associations with a P-value < .05 were considered significant and no GWAS was performed. Effect sizes of genetic variants are described as odds ratio (OR) for logistic regression or beta (β) for linear regression.

***HLA* imputation**

Previous IBD genetic studies have revealed that *HLA* alleles explain substantially more of the disease variance than that explained by index genetic variants in the *MHC* region[19]. We submitted phased genotypes of 6,114 markers within the *MHC* region to the HLA*IMP:03 server, which imputed four-digit classical alleles of 11 *HLA* region genes for each individual[20]. (**Supplementary Figures 1** and **2**). Only imputed *HLA* alleles with a posterior probability ≥ 99% and a frequency ≥ 5% were included.

Genetic risk scores

We selected 13 published GWAS (or meta-analyses) that indexed traits related to IBD, gastrointestinal diseases, immunological disease, IBD phenotypes, and negative control phenotypes (**Supplementary Table 2**). We obtained summary statistics of these GWAS from publicly available repositories, or through collaboration, and these were used as 'base' data[5,21-30]. Using PRSice2 software[31], GRS were calculated for each of the base datasets. GRS were calculated by computing the sum of risk alleles corresponding to a base phenotype in each patient, weighted by the effect size estimate derived from the base GWAS. Genetic variants were pruned for linkage disequilibrium ($R^2 > 0.1$ within a 500kb window), using the 1KG European dataset as external reference panel. The optimal GRS for each phenotype in cohort A was calculated using *P*-value thresholds (pT) from 1.0E-08 to 0.5, in steps of 5.0E-08. The explained variance (Nagelkerke's R^2) was derived from a linear model in which the IBD phenotype (target phenotype) was regressed on each GRS, adjusting for the covariates age, sex, and diagnosis (CD vs. UC) and the first 5 principal components from the genetic data. In total, GRS were calculated for 13 traits – CD, UC, primary sclerosing cholangitis (PSC), rheumatoid arthritis, asthma, celiac disease, idiopathic pulmonary fibrosis, diverticulosis, ever smoking, former smoking, CD prognosis (poor prognosis defined by the need for repeated surgery or the use of two or more immunosuppressives), bone mineral density, and serum vitamin D levels – and targeted on a total of 24 phenotypes in cohort A.

Statistical analyses

To obtain the optimal pT, and thus the best predictive GRS, multiple models were fitted on each target phenotype and genetic variants were added to the model at each new pT. Although there was a high correlation between each model (i.e. only a small number of variants was added at each threshold), the significance of the best-fit GRS should be corrected for this multiple testing. To obtain the empirical P-value of each GRS–phenotype association in cohort A, 10,000 rounds of permutation were performed. GRS–phenotype associations with an empirical P-value < .05 were considered significant. Because we tested for associations with 24 phenotypes, of which 16 were independent (groups of) phenotypes, we performed Bonferroni correction for this number of phenotypes (empirical P-value * 16 = FDR). An FDR < .05 after Bonferroni correction was considered significant. All significant associations were put forward for replication in the independent cohort B. Associations were considered replicated when a GRS generated in cohort A was also significantly (P-value < .05) associated to the same phenotype in cohort B, with a consistent direction of effect.

Meta-analyses were performed to assess inter-cohort heterogeneity and obtain a meta P-value for each significant GRS–phenotype association. To facilitate interpretability, all GRS were standardised. To ensure that the significant GRS–phenotype associations were based on independent sets of genetic variants, these GRS were recalculated using a larger pruning window of 1 Mb. Next, we excluded all genetic variants within the gene regions of *NOD2*, *MHC*, and *MST1* from the finally selected GRS and tested whether the GRS remained significantly associated to the phenotype.

Significant GRS–phenotype associations, clinical factors, and previously identified genetic predictors (i.e. *NOD2, MHC,* and *MST1* for CD disease location and CD disease behaviour; *NOD2* and *MHC* for ileocaecal resection; and *MHC* for PSC) were included in a multivariate model. Finally, we repeated the multivariate analyses including *HLA* alleles previously identified as genetic predictors of CD and/or or UC instead of individual genetic variants in *MHC.*

Data availability

Raw data is (in part) available at https://ega-archive.org/studies/EGAS00001002702, or upon request.

RESULTS

Phenotype data

Clinical characteristics were obtained from 1,097 patients from cohort A and 2,156 patients from cohort B (data displayed in **Table 1**). Patients in cohort B were younger than those in cohort A (P < 1.0E-04). Patients in cohort B were less often diagnosed with CD compared to cohort A, but had younger onset of CD, more ileocolonic (Montreal L3) localisation, and more inflammatory (Montreal B1) disease behaviour (all P < 1.0E-04). Patients with UC in cohort B had more left-sided colitis compared to cohort A (P = .02). Patients in cohort B had less often undergone colonic resection, but more often undergone ileocaecal resection (both P < 1.0E-04). In cohort B, there were more current smokers but fewer former smokers compared to cohort A. In cohort A, more patients had an IBD-PSC phenotype (P < 1.0E-04). Finally, extra-intestinal manifestations differed between the two cohorts (all P < .04).

Genetic variant phenotype associations

We first sought to replicate known associations between genetic variants and IBD phenotypes. We replicated (P < .05) the associations of *NOD2* with age at diagnosis in CD and IBD, CD disease location, CD disease behaviour, and the need for surgery in CD. Consistent with previous data, the association of *NOD2* with the need for surgery in CD disease was largely mediated trough younger age at diagnosis and ileal disease location (**Supplementary Table 3**)**.** We also replicated the associations between *MHC* and UC disease extent and CD disease location. Finally, we replicated the association between *MST1* and age at diagnosis in CD (**Supplementary Table 1B**).

Genetic risk scores

We constructed a total of 24 GRS models for each base trait (GRS targeted on phenotypes). **Supplementary Table 4** gives the estimates from all GRS linear regression analyses. As displayed in **Table 2**, seven GRS models remained significantly associated with an IBD phenotype after 10,000 permutations and Bonferroni correction and were significantly replicated with a consistent effect direction in independent cohort B. The explained variance of significant GRS models across different pT are shown in **Supplementary Figure 3**. All seven GRS models remained significantly associated with an IBD phenotype when repeating the analyses with a larger pruning window of 1 Mb (**Supplementary Table 5**).

Genetic disease susceptibility predicts disease phenotypes

We reasoned that the genetic burden of CD and UC might be predictive of specific clinical phenotypes when including more genetic variants than just those that reached genome-wide significance in previous GWAS.

The composite genetic risk of CD (CD GRS) was significantly associated with CD disease behaviour (B2 vs. B1; pT = 1.0E-08, FDR = .02, R^2 = 6.9%) (**Figure 1**) and the risk of ileocaecal resection (pT = 8.0E-04, FDR = 1.6E-03, R^2 = 4.1%), and these associations remained significant after excluding *NOD2, MHC,* and *MST1* from the genetic data (both FDR = 1.6E-03). Multivariate logistic regression analyses that included genetic factors previously identified as risk factors for CD disease behaviour revealed that age (P = 1.9E-06; OR 1.09 [95% CI 1.05-1.13]), age at diagnosis (P = 1.8E-04; OR 0.93 [95% CI 0.90-0.97]), CD disease location (P = 1.8E-04; OR 3.47 [95% CI 1.84-6.76]), and CD GRS (P =

1.8E-03; OR 1.79 [95% CI 1.26-2.61]) were all independently associated with CD disease behaviour (**Supplementary Table 6A**).

Multivariate regression analyses revealed that age (P = 8.9E-04 ; OR 1.08 [95% CI 1.03-1.13]), age at diagnosis (P = .01; OR 0.94 [95% CI 0.90-0.98]), CD disease location (P = 2.3E-07; OR 25.8 [95% CI 8.5-103]), CD disease behaviour (P = 7.8E-06; OR 6.47 [95% CI 2.91-15.1]), and CD GRS (P = 3.0E-03; OR 1.99 [95% CI 1.29-3.21]) were all independently associated with the risk of ileocaecal resection in patients with CD (**Supplementary Table 6B**).

The composite genetic risk of UC (UC GRS) was significantly associated with CD disease location (L1/L3 vs. L2; pT = 1.0E-04, FDR = 0.02, R^2 = 9.1%), but this association was lost after excluding *NOD2, MHC,* and *MST1* from the genetic data (FDR = .22). Multivariate logistic regression analyses revealed that the UC GRS was independently associated to CD disease location (P = 2.0E-05; OR 0.60 [95% CI 0.48-0.76]) (**Supplementary Table 6C**).

Multivariate regression analyses were then repeated including the imputed *HLA* alleles that had previously been shown to be associated to CD or UC[19]. None of the selected *HLA* alleles were significantly associated to CD disease behaviour (**Supplementary Table 7A**). In addition to age, age at diagnosis, CD disease location, CD disease behaviour, and CD GRS, carriage of *HLA-DRB1*03:01* (P = .01; OR 6.6 [95% CI 1.53-27.9]) and *HLA-C*06:02* (P = .02; OR .21 [95% CI 0.05-0.75]) were independently associated to the risk of ileocaecal resection in patients with CD (**Supplementary Table 7B**). In addition, UC GRS and carriage of *HLA-DRB1*03:01* (P = .01; OR 0.46 [95% CI 0.27-0.81]) were both independently associated with CD disease location (**Supplementary Table 7C)**.

Shared genetic aetiology

To further our understanding of the molecular mechanisms leading to specific disease phenotypes and to aid drug-repurposing, we correlated GRS of diseases related to IBD (phenotypes) to specific IBD phenotypes.

The composite genetic risk of PSC (PSC GRS) was significantly associated with IBD-PSC (pT = 1.6E-03, FDR = 1.6E-03, R^2 = 7.5%). Multivariate logistic regression analyses revealed that age (P = 2.8E-06; OR 1.06 [95% CI 1.03-1.08]), age at diagnosis (P = 4.8E-08; OR 0.93 [95% CI 0.90-0.95]), male sex (P = .03; OR 1.79 [95% CI 1.05-3.11]), UC (P = 1.0E-08; OR 7.34 [95% CI 3.85-15.22]), and PSC GRS (P = 8.1E-09; OR 1.87 [95% CI 1.51-2.32]) were all independently associated with the risk of IBD-PSC (**Supplementary Table 6D**). Moreover, the PSC GRS were significantly associated with CD disease location (pT = .01, L1/L3 vs. L2; FDR = 0.03, R^2 = 3.6%), but this association was lost after excluding *NOD2, MHC,* and *MST1* from the genetic data (P = .09). Indeed, multivariate logistic regression

analyses showed that genetic variation in *MHC* and *MST1* and the PSC GRS (P = .02, OR 0.73 [95% CI 0.56-0.94) were all independently associated to CD disease location (**Supplementary Table 6E**). Finally, the PSC GRS showed association with the smoking history of IBD patients (ever vs. never; pT = 9.6E-03, FDR = .04, R^2 = 1.7%). Multivariate logistic regression analyses revealed that age (P = 2.9E-03; OR 1.02 [95% CI 1.01-1.04]), age at diagnosis (P = 2.9E-05; OR 1.04 [95% CI 1.02-1.05]), CD (P= 4.2E-09; OR 2.37 [95% CI 1.78-3.18]), and PSC GRS (P = 5.5E-04; OR 0.78 [95% CI 0.67-0.90]) were all independently associated with smoking history (**Supplementary Table 6F**).

Several variants in *MHC* confer risk of PSC[32], and recent GWAS have identified genetic variation in *MHC* to be associated with CD prognosis[28]. The composite genetic risk of a poor CD prognosis (CDprog GRS) was significantly associated with IBD-PSC (pT = 1.5E-06, FDR = 6.4E-03, R^2 = 5.5%). After excluding *MHC* from the genetic data, the association between CDprog GRS and IBD-PSC was lost (P = .71). Multivariate logistic regression analyses that included variants in *MHC* that explain most of the association with PSC[32] revealed that age (P = 1.9E-06; OR 1.06 [95% CI 1.03-1.08]), age at diagnosis (P = 1.4E-08; OR 0.92 [95% CI 0.90-0.95]), male sex (P = .02; OR 1.86 [95% CI 1.09-3.20]), UC (P= 4.9E-09; OR 7.55 [95%CI 3.99-15.53]), and CDprog GRS (P = 3.0E-06, OR 0.61 [95%CI 0.50-0.75]) were all independently associated with the risk of IBD-PSC (**Supplementary Table 6G**).

We identified an association between the composite genetic risk of CD and CD disease location (L1 vs. L2). However, this association failed the Bonferroni significance threshold in the discovery cohort (FDR = .10).

DISCUSSION

In this study we used GRS to study the aggregated effect of thousands of trait-associated genetic variants on IBD phenotypes. We show that increased genetic risk of CD is associated with fibrostenotic CD. We also validate the putatively shared genetic aetiology of PSC and UC with colonic CD. Finally, our results add to the existing hypothesis of an interaction between smoking and PSC.

Recent genotype–phenotype studies have explored genetic determinants of specific clinical IBD phenotypes and showed that genetic variants in the known IBD susceptibility loci *NOD2*, *MHC,* and *MST1* were associated with age at onset and CD disease location[10]. In our study, we first replicated these previously identified genetic variants as predictors of clinical IBD phenotypes[10]. Using two independent cohorts of IBD patients with genome-wide genetic array data, we then showed that the composite genetic risk of CD is associated with fibrostenotic CD in patients with CD, even after excluding *NOD2*, *MHC,* and *MST1*. Our data suggest that the variants most strongly associated with CD also play a role in fibrostenotic disease. Moreover, the composite genetic risk of CD appears to

be associated with the risk of ileocaecal resection and remains significant after correcting for CD disease location and CD disease behaviour.

We further validated the association between colonic CD and the genetic risk of UC. This observation is in line with the hypothesis that there is a continuum of phenotypes ranging from ileal CD to colonic CD to UC, rather than CD and UC being two distinct diseases[10,33]. Our data suggest that genetic variation in *MHC* is largely driving this association. Although our multivariate analyses did not identify carriage of *MHC* genetic variants as an independent predictor of colonic CD, excluding *MHC* from the genetic data leads to loss of the association. Indeed, *HLA* alleles may explain substantially more phenotypic variance than individual genetic variants in *MHC,* and we identified carriage of *HLA-DRB1*03:01* as an independent predictor of CD disease location. The fact that we could not reliably impute SNP rs77005575, the strongest *MHC* predictor of colonic CD, may explain why we did not capture additional signals in *MHC*.

We observed pleiotropy between the genetic risk of PSC and the smoking history of patients with IBD. In our study, genetic risk of PSC was negatively correlated to the risk of smoking. Indeed, recent epidemiological studies have identified a decreased risk of PSC among smokers[34], which might in part be explained by genetic factors. Shared genetic aetiology between diseases other than IBD and IBD phenotypes may point to a biology that could provide new therapeutic options or aid drug repurposing. A number of apparently unrelated disease processes may result in tissue fibrosis, including, for example, intestinal fibrosis in CD and idiopathic pulmonary fibrosis (IPF). We hypothesised we would find pleiotropy between the composite genetic risk of idiopathic pulmonary fibrosis (IPF GRS) and fibrostenotic CD, but could not identify this in our dataset. Moreover, we found no significant associations between the composite genetic risk of bone mineral density and serum vitamin D levels and the risk of osteoporosis in our IBD cohorts.

A recent study that defined poor CD prognosis by the need for repeated surgery or the use of two or more immunosuppressives found no association between CD genetic risk and CD prognosis[28]. In contrast, four novel genetic variants distinct from those that confer risk to CD susceptibility have been identified as predictors of poor CD prognosis[28]. We could not identify an association between the aggregated effects of these novel genetic variants and IBD phenotypes such as CD disease location or CD disease behaviour. Genotype–phenotype studies are, however, dependent on the criteria used to define phenotypes, such as disease prognosis. In addition, differences in local medical practice and differences in medical practices over time may introduce significant bias. This may explain why we could not identify genetic contributions to the need for surgery, which we used as a proxy for poor prognosis. More objective outcome measures, e.g. the number of flares, once

corrected for confounders, may improve the identification of genetic predictors of disease prognosis. Moreover, the relatively small sample size of the initial CD prognosis GWAS may limit the accuracy of this GRS.

Current clinical classifications fail to accurately predict IBD disease course[3]. We posit that GRS have the potential to uncover biological mechanisms contributing to disease phenotypes, and these mechanisms may in turn be used for drug development or improved patient stratification. Although current discriminative accuracy remains low, integrating other factors such as transcriptomic and microbial signatures may significantly improve discriminative power and might outperform current clinical classification systems in their ability to predict IBD disease course[3]. Patient-specific molecular profiles may be used to select more homogenous groups of patients for future clinical trials. We hypothesise that therapies in development or registered for UC might be successfully repurposed to patients with colonic CD, which is biologically associated with UC. In a clinical setting, patients with a high molecular risk of fibrostenotic disease might be treated more aggressively early in their disease course, in particular in the presence of other environmental risk factors.

We fitted GRS models on target phenotypes to obtain the optimal pT and thus the best predictive GRS for each phenotype. GRS models with relatively high optimal pT may suggest (pleiotropic) biological signals of (sets of) genetic genetics that fail to reach GW significance. Large cohort studies of patients with multi-layered molecular data are needed to explore clinically relevant molecular risk cut-off values.

The use of two independent, well-characterised clinical cohorts of IBD patients, both genotyped using the same methods, strengthens the results of this study. However, the relatively small sizes of our cohorts and the fact that we calculated GRS for only thirteen traits may have precluded comprehensive identification of genetic contributions to IBD phenotypes.

In conclusion, we show that GRS can identify genetic contributions to clinical disease heterogeneity of IBD. Molecular phenotyping, including of genetic, microbial, and environmental factors, of well-characterised cohorts of IBD patients holds promise to further our understanding of the heterogeneous character of IBD and allow clinical trials to study personalised disease-management strategies.

FUNDING

FH has received research grants from Dr Falk, Janssen-Cilag, Abbvie, and Takeda. RKW is supported by a Diagnostics Grant from the Dutch Digestive Foundation (D16-14). NKHB has received unrestricted research grants from Dr Falk, TEVA Pharma BV, MLDS, and Takeda. EAMF is supported by an MLDS Career Development grant (CDG 14-04). RKW has received unrestricted research grants from Takeda, Tramedico, and Ferring. EAMF has received an unrestricted research grant from Takeda.

ACKNOWLEGDEMENTS

The authors thank Kate Mc Intyre, Scientific Editor in the Department of Genetics, University Medical Center Groningen, for editing and formatting this manuscript

AUTHOR CONTRIBUTIONS

MDV, LMS, and KWJS had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. *Concept and design*: MDV, EAMF, RKW. *Acquisition, analysis, or interpretation of data:* MDV, LMS, KWJS, BHJ, GD, CJW, FH, MJP, AEM, NKHB, ML, BO, EAMF, RKW. *Drafting of the manuscript:* MDV, EAMF, RKW. *Critical revision of the manuscript*: GD, CJW, FH, MJP, AEM, NKHB, ML, BO.

CONFLICTS OF INTERESTS

FH has served on advisory boards or as speaker for Abbvie, Janssen-Cilag, MSD, Takeda, Celltrion, Teva, Sandoz, and Dr Falk and has received consulting fees from Celgene. NKHB has served as a speaker for AbbVie and MSD and as a consultant and principal investigator for TEVA Pharma BV and Takeda. All other authors report no conflicts of interest.

REFERENCES

1. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. *Lancet*. 2017;389(10080):1741-1755.

2. Ungaro R, Mehandru S, Allen PB, Peyrin-Biroulet L, Colombel JF. Ulcerative colitis. *Lancet*. 2017;389(10080):1756-1770.

3. Furey TS, Sethupathy P, Sheikh SZ. Redefining the IBDs using genome-scale molecular phenotyping. *Nat Rev Gastroenterol Hepatol*. 2019;16(5):296-311.

4. Liu JZ, Van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. 2015;47(9):979-986.

5. De Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. 2017;49(2):256-261.

6. Festen EAM, Goyette P, Green T, et al. A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for crohn's disease and celiac disease. *PLoS Genet*. 2011;7(1):e1001283.

7. Parkes M, Cortes A, Van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet*. 2013;14(9):661-673.

8. Colombel JF, Grandbastien B, Gower-Rousseau C, et al. Clinical characteristics of Crohn's disease in 72 families. *Gastroenterology*. 1996;111(3):604-607.

9. Halfvarson J, Bodin L, Tysk C, Lindberg E, Järnerot G. Inflammatory bowel disease in a Swedish twin cohort: A long- term follow-up of concordance and clinical characteristics. *Gastroenterology*. 2003;124(7):1767-1773.

10. Cleynen I, Boucher G, Jostins L, et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet*. 2016;387(10014):156-167.

11. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460(7256):748-752.

12. Dudbridge F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet*.

13. Imhann F, Van Der Velde KJ, Barbieri R, et al. The 1000IBD project: Multi-omics data of 1000 inflammatory bowel disease patients; Data release 1. *BMC Gastroenterol*. 2019;19(1):5.

14. Spekhorst LM, Imhann F, Festen EAM, et al. Cohort profile: Design and first results of the Dutch IBD Biobank: A prospective, nationwide biobank of patients with inflammatory bowel disease. *BMJ Open*. 2017;7(11):e016695.

15. Silverberg M, Satsangi J, Ahmad T, et al. Toward an Integrated Clinical, Molecular and Serological Classification of Inflammatory Bowel Disease: Report of a Working Party of the 2005 Montreal World Congress of Gastroenterology. *Can J Gastroenterol*. 2005;19(Suppl A):5A-36A.

16. Loh PR, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48(11):1443-1448.

17. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526:68-74

18. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7.

19. Goyette P, Boucher G, Mallon D, et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. *Nat Genet*. 2015;47(2):172-9.

20. Motyer A, Vukcevic D, Dilthey A, et al. Practical Use of Methods for Imputation of HLA Alleles from SNP Genotype Data. *BioRxiv*. 2016. 10.1101/091009

21. Ji SG, Juran BD, Mucha S, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet*. 2017;49(2):269-273.

22. Stahl EA, Raychaudhuri S, Remmers EF, et al. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*. 2010;42(6):508-514.

23. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*. 2010;363(13):1211-1221.

24. Ricaño-Ponce I, Gutierrez-Achury J, Costa AF, et al. Immunochip meta-analysis in European

and Argentinian populations identifies two novel genetic loci associated with celiac disease. *Eur J Hum Genet*. 2020;28(3):313-323.

25. Allen RJ, Porte J, Braybrooke R, et al. Genetic variants associated with susceptibility to idiopathic pulmonary fibrosis in people of European ancestry: a genome-wide association study. *Lancet Respir Med*. 2017;5(11):869-880.

26. Maguire LH, Handelman SK, Du X, Chen Y, Pers TH, Speliotes EK. Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease. *Nat Genet*. 2018;50(10):1359-1365.

27. Furberg H, Kim Y, Dackor J, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet*. 2010;42(5):441-447.

28. Lee JC, Biasci D, Roberts R, et al. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat Genet*. 2017;49(2):262-268.

29. Kemp JP, Morris JA, Medina-Gomez C, et al. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat Genet*. 2017;49(10):1468-1475.

30. Jiang X, O'Reilly PF, Aschard H, et al. Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. *Nat Commun*. 2018;9(1):260.

31. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019;8(7):giz082.

32. Lui JZ, Hov JR, Folseraas T, et al. Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* 2013:45(6):670-675.

33. Imhann F, Vich Vila A, Bonder MJ, et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut*. 2018;67(1):108-119.

34. Wijarnpreecha K, Panjawatanan P, Mousa OY, Cheungpasitporn W, Pungpapong S, Ungprasert P. Association between smoking and risk of primary sclerosing cholangitis: A systematic review and meta-analysis. *United Eur Gastroenterol J*. 2018;6(4):500-508.

TABLES

Characteristic	**Cohort A** (n = 1,097)	**Cohort B** (n=2,156)	P-values
Sex, No. (%)			.35
Female	559 (57%)	1266 (59%)	
Age, median (IQR), y	48 (36-60)	44 (33-56)	< 1.0E-04
Type of IBD diagnosis, No. (%)			< 1.0E-04
Crohn's disease	506 (52%)	1393 (65%)	
Ulcerative colitis	417 (43%)	763 (35%)*	
IBD-unclassified	48 (5%)	NA	
Montreal A			.04
A1	79 (16%)	209 (15%)	
A2	320 (65%)	977 (70%)	
A3	96 (19%)	207 (15%)	
Montreal L			< 1.0E-04
L1	179 (58%)	187 (18%)	
L2	101 (33%)	368 (34%)	
L3	26 (9%)	514 (48%)	
L4 (upper GI involvement)	51 (17%)	112 (10%)	.16
Montreal B			< 1.0E-04
B1	238 (48%)	927 (67%)	
B2	178 (36%)	270 (19%)	
B3	82 (16%)	196 (14%)	
Bp	156 (31%)	381 (27%)	.14
Montreal E			.02
E1	45 (11%)	39 (6%)	
E2	135 (32%)	230 (37%)	
E3	242 (57%)	361 (57%)	
Primary Sclerosing Cholangitis	78 (7%)	42 (2%)	< 1.0E-04
Surgery			
Colonic resection	349/980 (36%)	408 (19%)	< 1.0E-04
Ileocaecal resection	180/980 (18%)	525 (24%)	2.0E-04
Smoking status			
Current	190/942 (19%)	397 (29%)	.09
Former	516/910 (53%)	698 (36%)	< 1.0E-04
Ever	575/980 (59%)	836 (43%)	< 1.0E-04

Extra intestinal manifestations			
Ocular manifestations	35 (3%)	103 (5%)	.033
Cutaneous manifestations	147 (13%)	229 (11%)	.019
Arthropathies	304 (28%)	410 (19%)	< 1.0E-04
Arthritis	42 (4%)	150 (7%)	3.4E-04
Thromboembolism	12 (1%)	81 (4%)	< 1.0E-04
Osteoporosis	59 (5%)	452 (21%)	< 1.0E-04

Table 1. Phenotype distributions of discovery and replication cohorts

Characteristics of patients with IBD from the discovery cohort (cohort A) and replication cohort (cohort B). Percentages were calculated from non-missing data. Montreal refers to the Montreal classification[15]. GI: gastrointestinal; IBD: inflammatory bowel disease; IQR: inter-quartile range. *UC and IBDU were grouped in cohort B.

				Discovery cohort A			Replication cohort B	Meta-analyses		
Genetic risk score	**Phenotype**	**pT**	**Variants (n)**	**Empirical P-value**	**FDR**	**R^2**	**P-value**	**Z-score**	**P-value**	**Direction**
Crohn's disease	Ileocaecal resection	8.0E-04	5379	1.0E-04	1.6E-03	4.1%	2.0E-05	5.8	8.2E-09	++
excluding NOD2, MST, MHC			*3836*	*1.0E-04*	*1.6E-03*	*4.3%*				
Crohn's disease	Fibrostenotic Crohn	1.0E-08	219	1.0E-03	.02	6.9%	2.2E-03	4.3	1.6E-05	++
excluding NOD2, MST, MHC			*170*	*1.0E-04*	*1.6E-03*	*11.0%*				
Ulcerative colitis	Colonic Crohn	1.0E-04	1334	1.0E-03	.02	9.1%	6.0E-03	-4.1	3.8E-05	--
excluding NOD2, MST, MHC			*939*	*.01*	*.22*					
Primary sclerosing cholangitis	Colonic Crohn	.01	12487	2.0E-03	.03	3.6%	.04	-3.5	5.5E-04	--
excluding NOD2, MST, MHC			*10913*	*.09*						
Primary sclerosing cholangitis	IBD-PSC	1.6E-03	2365	1.0E-04	1.6E-03	7.5%	2.0E-06	6.1	8.9E-10	++
Primary sclerosing cholangitis	Smoking history	9.6E-03	9735	2.5E-03	.04	1.7%	7.6E-03	-3.9	8.5E-05	--

Manuscript Doi: 10.1093/ecco-jcc/jjaa223

Crohn's disease prognosis *excluding MHC*	IBD-PSC	1.5E-06	83	4.0E-04 .71	6.4E-03	5.5%	2.1E-05	-5.5	3.4E-08	22 --

Table 2. Overview of significant GRS–phenotype associations

Associations between genetic risk scores and clinical IBD phenotypes that remained significant after Bonferroni correction in the discovery cohort and showed positive replication with consistent direction of effect in the independent replication cohort. For IBD phenotypes with previously known genetic predictors (*NOD2*, *MHC*, and *MST1*), analyses were repeated after excluding these genes from the genetic data (depicted in grey). Variants refers to the number genetic variants included in the optimal GRS for that phenotype (most explained variance). Empirical P-value refers to the P-value after 10,000 rounds of permutation. Meta-analyses P-value refers to meta-analyses of results from both discovery (FDR) and replication (P-value) cohorts. CD: Crohn's disease; FDR: false discovery rate; GRS: genetic risk score; PSC: primary sclerosing cholangitis; pT: P-value threshold for optimal GRS.

FIGURES

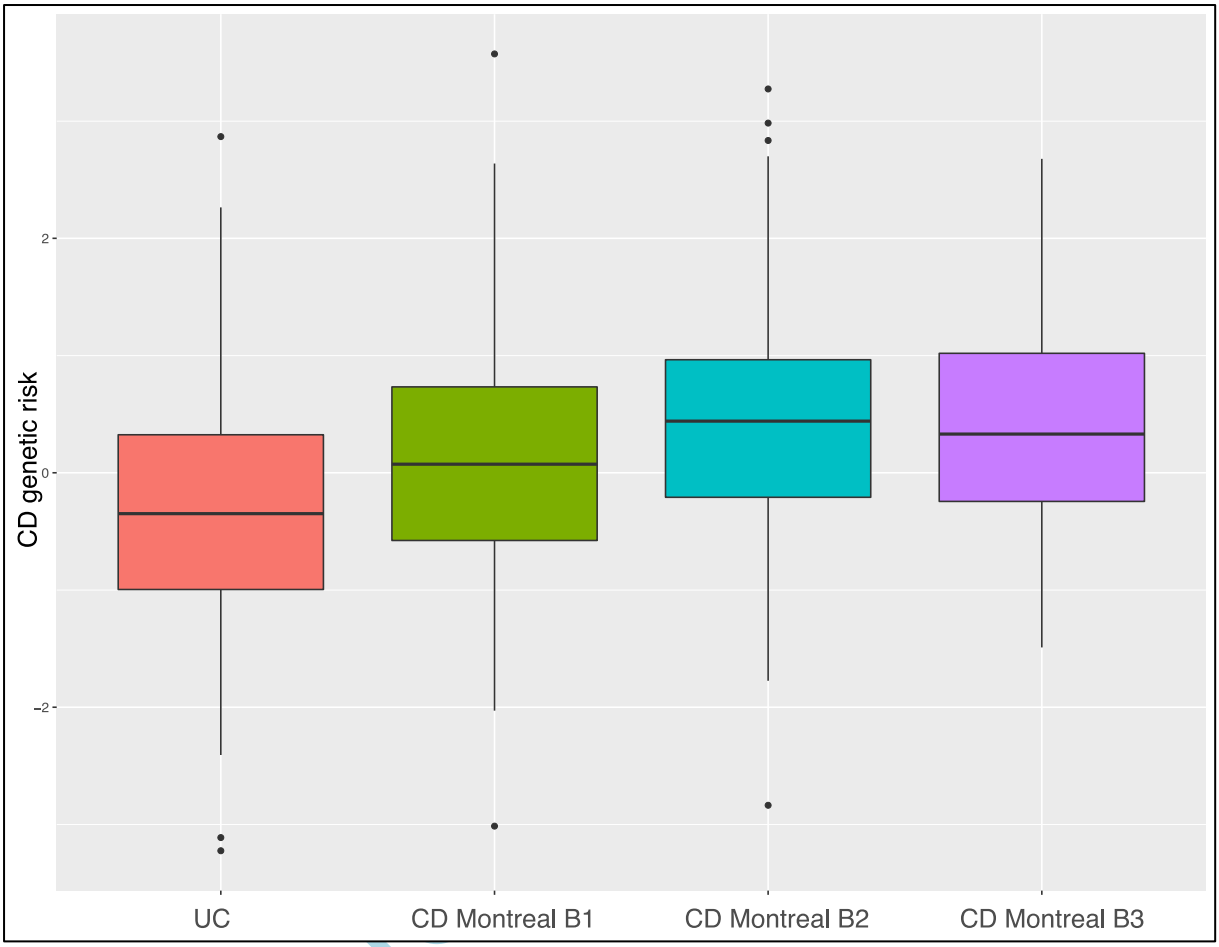

Figure 1.

Figure 1. Boxplot showing the composite genetic risk of CD by CD disease behaviour

Boxplot representing the range of the composite genetic risk of Crohn's disease for patients with UC and CD. Patients with CD are stratified by CD disease behaviour. Montreal refers to the Montreal classification[15]. Montreal B1: non-stricturing, non-penetrating Crohn's disease; B2: fibrostenotic Crohn's disease; B3: penetrating Crohn's disease. CD: Crohn's disease; UC: ulcerative colitis.

Manuscript Doi: 10.1093/ecco-jcc/jjaa223

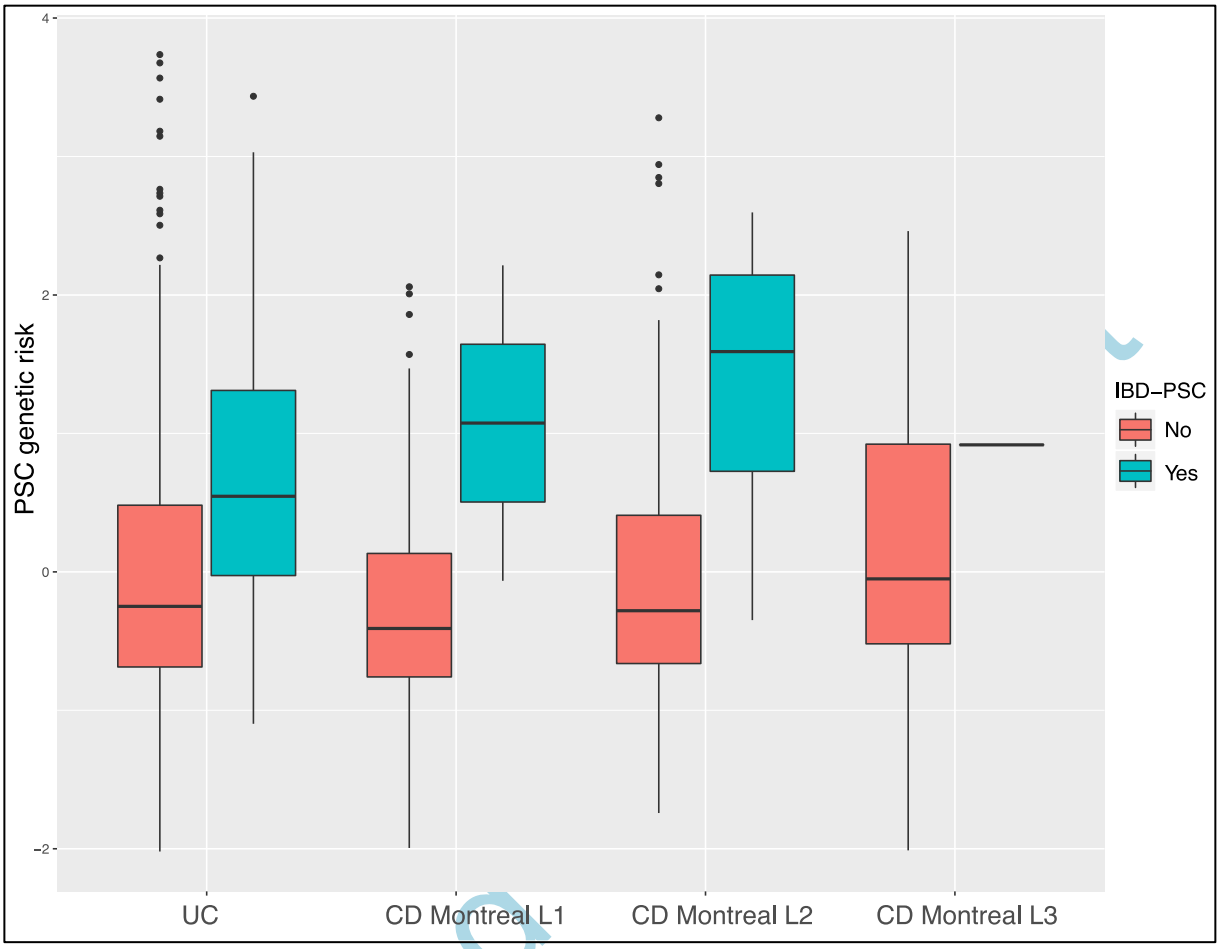

Figure 2. Boxplot showing the composite genetic risk of PSC by CD disease location

Boxplot representing the range of the composite genetic risk of PSC, for patients with UC and CD. Patients with CD are stratified by CD disease location. Montreal refers to the Montreal classification[15]. Montreal BL: ileal Crohn's disease; L2: colonic Crohn's disease; L3: ileocolonic Crohn's disease. Patients with confirmed diagnosis of PSC are displayed in red. CD: Crohn's disease; PSC: primary sclerosing cholangitis; UC: ulcerative colitis.