My PhD Thesis

My PhD subtitle

Omar El Garwany

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Churchill College October 2023

I would like to dedicate this thesis to my loving parents . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Omar El Garwany
October 2023

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Chapter 1

HLA-DRB1*01:03 is associated with risk of perianal Crohn’s Disease

1.1 Introduction

Perianal Crohn’s disease (pCD) is a sub-phenotype of Crohn’s disease, a chronic inflammatory disease of the gut that affects 1% of the population worldwide. pCD represents a major burden on both patients and healthcare providers, and is estimated to affect 20-40% of CD patients worldwide, with a higher prevalence in Asia than in Western countries [1]. As their disease progresses, CD patients become more likely to develop perianal symptoms. Twenty years after their CD diagnosis, CD patients have a 42% cumulative probability of developing pCD. Timing of pCD diagnosis, however, varies significantly between healthcare systems. Previous studies from countries including France, Sweden and Japan have reported that between 4%-68% of pCD patients present with perianal symptoms before or at the time of CD diagnosis [2–4].

pCD patients present with a variety of perianal symptoms. These include perianal skin tags, fissures, ulcers, faecal incontinence, rectal discharge and bleeding, perianal abscess, and fistulas. Perianal fistulas are the most common form of pCD, followed by perianal abscess [5]. Likewise, the impact of pCD on patients is multi-faceted. In addition to physical manifestations, patients report impaired quality of life as well as social and emotional complications of pCD. Furthermore, surgical interventions that aim to treat pCD and restore normal ano-rectal functions, such as seton insertion and fistula drainage, often impact essential functions such as walking and sitting [6]. Moreover, pCD patients, who often require multiple surgical interventions, suffer from high recurrence and relapse rates. In fact, only one third of pCD patients are estimated to achieve remission [7, 8].

Despite the diversity in presentation and course, pCD patients share a number of Crohn's disease characteristics. pCD is more common among patients with distal than proximal disease. Patients with colonic or rectal CD are more likely to develop or initially present with pCD. Moreover, pCD tends to drive Crohn's disease towards a more invasive behaviour. Initially, two thirds of patients have inflammatory manifestations, but over time, the majority of pCD patients display increasingly stricturing and penetrative characteristics of CD [9, 10]. However, invasive distal CD does not always precede pCD, which can sometimes present a diagnostic challenge in clinical setting. Although 95% of patients will eventually develop luminal disease, an estimated 17.2% of patients initially present with pCD only [11].

At a more fundamental level, our biological understanding of pCD fistula formation and progression mechanisms is still markedly lacking. One proposed pathophysiological mechanism is epithelial-to-mesenchymal transformation (EMT). EMT is a well-studied biological process, whereby polarised epithelial cells that line the gastrointestinal tract gain mesenchymal functions, such as enhanced cell invasion, and migration. The EMT hypothesis is supported by the presence of transitional cells which express both epithelial and mesenchymal cell markers in fistula tracts. These include epithelial markers cytokeratin 8 and 20, and mesenchymal markers vimentin and actin. Transforming growth factor β and interleukin-13, which have been associated with the initiation of EMT, have also been identified in transitional cells lining pCD fistula tracts [12].

In this chapter, I will describe a within-case genome-wide association study (GWAS) meta-analysis that I performed between two IBD cohorts to understand the genetic underpinnings of pCD. The two cohorts are: the NIHR IBD-BioResource (IBD-BR), and the UK IBD Genetics Consortium (UKIBDGC).

The NIHR IBD-BR is a UK-wide collaborative project between the UK IBD Genetics Consortium and NIHR Bioresource, with the aim of recruiting 50,000 patients with Crohn's disease, ulcerative colitis or unclassified IBD. The IBD-BR collects phenotypic and epidemiological information (both clinical and self-reported) as well as DNA samples for both array genotyping and whole-genome and whole-exome sequencing. The aims of the IBD-BR are wide-ranging. They include understanding the genetics of IBD response to therapy, disease mechanism as well as determinants of disease course [13–15]. So far, the IBD-BR has recruited over 31,000 patients, with detailed clinical phenotypes on different epidemiological characteristics, clinical picture, extra-intestinal manifestations, prescribed medication and treatment history, surgical history and disease behaviour and complications. The recruitment process starts by an expression of interest by volunteers who visit participating recruitment centres. Interested volunteers are then provided with an invitation letter and a patient information sheet that provides information on study requirements. Patients who agree to take part are then provided with an informed consent form, and subsequently asked to complete a health and lifestyle questionnaire. After these initial steps, the clinical team then proceeds to collect clinical data from hospital records. A clinician or research nurse extracts core information including IBD type, location and behaviour, complications, comorbidities, family history, smoking history, surgical data and drug therapy outcomes [13].

The UK IBD Genetics Consortium (UKIBDGC), part of the International IBD Genetics Consortium, is a large IBD case-control consortium, with patients recruited from multiple UK centres in Cambridge, Edinburgh, Manchester, Newcastle, Exeter, Oxford, London, Dundee and Nottingham [16].

Although data on perianal disease are recorded for both cohorts, the depth of clinical phenotyping is different. For example, the IBD-BR, contains specific manifestations of pCD. Clinicians and clinical nurses who complete the IBD-BR questionnaire perform an automated search of hospital records for clinical IBD information, including perianal manifestations [13]. If the search is unsuccessful, they ask patients about perianal involvement: *"Ever had perianal involvement? 1) Yes 2) No 3) Unknown"* and record the answer in the clinical questionnaire [14]. A follow-up question about the type of perianal involvement is then asked: *“If Yes - What type of perianal lesion has the patient had? (Select all that apply): 1) Tags/fissures/ulcers 2) Perianal abscess 3) Simple fistula 4) Complex fistula 5) Other”*. Clinicians may report one or more perianal involvement manifestations. Unlike IBD-BR, the specific manifestations of pCD , such as fissures, ulcers or fistulas are not recorded for UKIBGC participants and only a binary phenotype is recorded (pCD+ or pCD-).

In this chapter, I describe several analyses I conducted to identify genetic variants associated with pCD risk, leveraging these two IBD cohorts. Additionally, I perform several follow-up analyses to ensure the robustness of the genome-wide association signal I found. Finally, I conclude by mapping the associated locus to an HLA allele that partly explain the genome-wide significant association and recommend that higher-resolution HLA alleles may explain this association more completely.

1.2 Methods

1.2.1 UK IBD Genetics Consortium Genotype Quality Control

UKIBDGC samples were genotyped with two genotyping arrays: Affymetrix Human Mapping 500K Array (I will refer to this as GWAS1; number of variants before QC=469,281), and Illumina Human Core Exome-12v1.0 or its newer version Illumina Infinium Core Exome-24v1.1 (I will refer to this as HCE; number of variants before QC=535,434 and 557,662 respectively). Quality control for UKIBDGC genotype data was performed as part of the International IBD Genetics Consortium cases-control meta-analysis. QC was performed using a combination of Plink (v1.9 and v2), bcftools (v1.16), and KING (v2.2.4).

1.2.2 Variant-level QC

Variants that met the following criteria were excluded:

- Low call rate (<0.95 for variants with minor allele frequency (MAF) > 0.01 or < 0.98 for variants with MAF ≤ 0.01).

- Significant difference in genotype call rate (P-value $< 10^{-4}$) between IBD cases and controls.

- Large allele frequency (AF) differences between UKIBDGC and Gnomad (Non-Finnish Europeans), or TOPMed (global) using the following formula:

  $$\frac{(P_1 - P_0)^2}{(P_1 + P_0)(2 - P_1 - P_0)} > \varepsilon$$

  where $\varepsilon = 0.025$ or 0.125, for Gnomad and TOPmed respectively, P_0 is the minor allele frequency (MAF) in Gnomad or TOPMed and P_1 is UKIBDGC MAF. This formula accounts for larger AF differences between UKIBDGC and population references in common than in low-frequency variants. The TOPMed global AF difference cutoff is higher to account for AF computed across diverse populations

- Hardy Weinberg Equilibrium (HWE) P-value < 10-5 in IBD controls or < 10^{-12} in IBD cases. HWE P-value was estimated in EUR ancestry samples. For chromosome X, HWE was estimated in genotypically-inferred EUR females, or

- Monomorphic variants.

1.2.3 Sample-level QC

Samples that meet the following criteria were excluded:

- Missing genotyping rate > 0.05
- Heterozygosity estimate $\pm$ 4 standard deviations from the EUR mean, or
- mismatch between recorded gender and genotypically-inferred sex.

1.2.4 Imputation to TOPMed

The TOPMed imputation server (imputationserver at 1.5.7) was used for UKIBDGC imputation. Directly genotyped variants with empirical imputation $R < -0.5$ were flipped, and variants with empirical $R^2 \leq 0.5$ were excluded after imputation. After their exclusion, imputation was repeated, and variants with HWE P-value $\leq 10^{-5}$ in IBD controls were excluded.

1.2.5 IBD-BR Genotype QC and Imputation

The cohort was genotyped with two different versions of the UKBiobank ThermoFisher genotyping array. The same genotype QC steps as UKIBDGC were applied to IBD-BR, except for 1) The AF difference check, where 1000 Genomes Panel (1000GP) was used as a reference panel 2) Imputation, where the Sanger Imputation Server was used [17], with two imputation reference panels: UK10K+1000GP and HRC. Imputed genotypes from both panels were combined. For variants that existed in both panels, HRC imputed genotypes only were retained. Imputed variants with large AF difference between the two panels, or with HWE P-value $\leq 10^{-5}$ in IBD controls were removed.

Genotypic principal components (PC) were estimated for all participants, using a set of genotyped variants that were also available in the 1000 Genomes Project (100GP; excluding variants associated with IBD susceptibility; P-value < 10^{-4}, and variants in long LD regions (as defined in [18]). This final list was pruned with the following parameters: window size = 50 kbp; step size = 5; R^2 = 0.2. PCs were then projected to 1000GP PCs. Samples within the European ancestry group were retained for the subsequent analyses.

1.2.6 Identification of overlapping samples between UKIBDGC and IBD-BR

Identification of duplicate individuals between UKIBDGC and IBD-BR genotyping data was performed with KING [19]. Based on the distribution of kingship coefficients, duplicates were defined as having a kinship coefficient > 0.354 as recommended in [19]. Estimation of kinship coefficient was performed using post-QC directly genotyped SNPs (number of variants used for kinship inference between IBD-BR and GWAS1=42,292:, and between IBD-BR and HCE=53,431).

1.2.7 Genome-wide association analysis

All genome-wide association analyses were performed using REGENIE v3.2.5 [20] following a 2-step procedure. Briefly, in step 1, a whole-genome regression model is fit using directly-genotyped variants in order to estimate a set of genome-wide predictors that capture a large fraction of phenotypic variance. These predictors are then used as covariates in step 2 where a larger set of variants are tested for association. I used post-QC genotyped variants in step 1 and both genotyped and imputed variants in step 2 from all autosomal chromosomes. Additionally, I enabled Firth approximation for all variants with P-value < 0.01 (`--firth --approx --pThresh 0.01`).

1.2.8 1000GP LD calculation

Population LD estimates were calculated using the 1000 Genomes Project high coverage dataset [21]. R^2 values where calculated using plink v1.9 and after retaining non-Finnish European individuals only.

1.2.9 χ^2 comparison between different pCD definitions

In order to compare associations from meta-analyses using different pCD+ case definitions to the meta-analysis with a broad pCD+ case definition, I adjusted the the broad-definition χ^2 values using this formula:

$$\chi^2_{Broad,n} = \frac{n}{N}\chi^2_{Broad}$$

Where n is the sample size of the meta-analysis being assessed, and χ^2_{Broad} is the broad-definition observed association statistic and $\chi^2_{Broad,n}$ is the broad-definition association statistic adjusted for sample size.

1.3 Results

Among 30,894 IBD-BR participants, 15,152 were diagnosed with Crohn's disease, 14,819 of which had perianal involvement data: 4,448 answered "*Yes*" to "*Ever had perianal involvement?*" (pCD+; 30%), 9,751 answered "*No*" (pCD-; 65.8%), and 620 answered "*Unknown*" (4.1%), matching previous pCD prevalence estimates. Perianal simple or complex fistula was the most common manifestation (2327; 52.3% pCD+ participants), followed by perianal abscess (1806; 40.5% of pCD+ participants).

From 26,327 UKIBDGC patients, a total of 8,977 were diagnosed with CD. 7106 CD patients had perianal involvement information. pCD prevalence was lower than IBD-BR. 18.2% of CD patients answered "*Yes*" to the question "Perianal involvement?", 61% answered "No", and 20.8% answered "*Unknown*". UKIBDGC does not report specific manifestations of pCD.

1.3.1 Epidemiological characteristics

Epidemiological characteristics of pCD+ and pCD- patients were largely similar in both cohorts (Table 1.3.1). Males were more likely than females to report perianal involvement in both cohorts (P-value=7×10^{-4} and 8×10^{-6} in IBD-BR and UKIBDGC respectively). pCD+ was not associated with a family history of CD, while smoking was slightly less common in pCD+ patients (P-value=0.006 and 0.003).

	IBD-BR		UKIBDGC	
	pCD+	pCD-	pCD+	pCD-
Male	2115 (47.5)	4339 (44.5)	807 (49.5)	2363 (43.2)
Female	2333 (52.5)	5412 (55.5)	824 (50.5)	3112 (56.8)
Family History	1325 (34.7)	2795 (34.2)	290 (27.1)	598 (24.6)
Surgery	2971 (68.8)	3636 (38.3)	896 (63.1)	1935 (42.6)
Smoking	656 (16.4)	1572 (18.2)	363 (30.1)	913 (29.6)

1.3.2 Clinical characteristics

pCD is associated with lower age-of-CD-diagnosis and rectal CD

Previous pCD studies have reported an association between pCD and distal penetrating CD as well as pCD and an earlier age of CD diagnosis [22, 23]. Compared to pCD- patients, pCD+ patients were significantly younger at diagnosis (P-value $< 2 \times 10^{-16}$; median age of CD diagnosis for pCD+ patients was 24 versus 29 for pCD- patients in IBD-BR; Figure 1.1). Additionally, pCD+ patients were at least twice as likely to have penetrating disease behaviour. In IBD-BR, 19.1% of pCD+ patients had disease behaviour classified as B3 versus 8.1% in pCD-. This enrichment was stronger in UKIBDGC (28.5% versus 10.6%, respectively).

In both cohorts, more patients reported ileal than colo-rectal CD (68.7% versus 56.3% in IBD-BR). Ileal and colo-rectal CD were either isolated, or extended to other parts of the gut. In IBD-BR, patients with an isolated colo-rectal CD were 2.4 times as likely to develop pCD, compared to patients with an isolated ileal CD (59.3% versus 24.8%). Despite the lower pCD prevalence in UKIBDGC, patients with isolated colo-rectal CD were similarly enriched for pCD+ patients as IBD-BR (26.6% versus 10.9%).

In IBD-BR, where colonic and rectal involvement are reported as separate indicators, rectal involvement accounted for this enrichment. Patients with isolated colonic disease were not significantly more enriched for pCD+ than patients with isolated ileal disease (28.3% versus 24.8%; Figure 1.1).

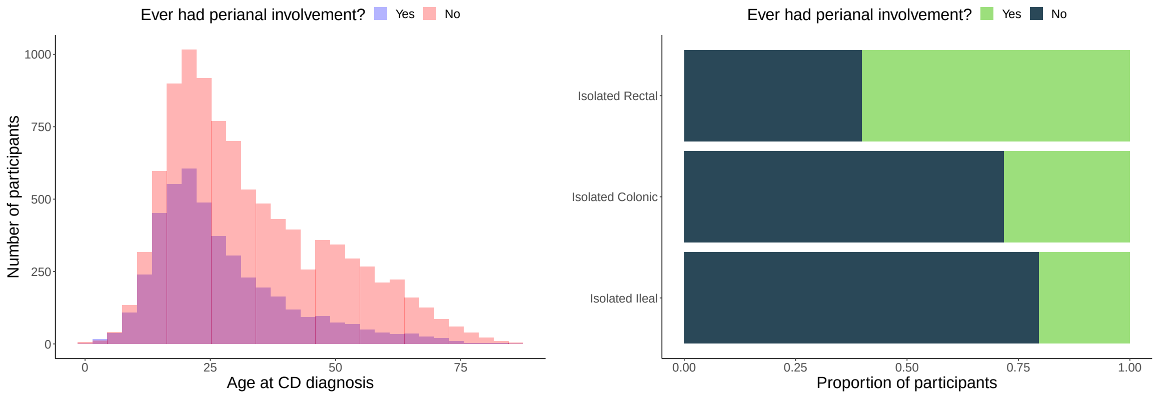

Fig. 1.1 age at diagnosis and macroscopic extent of Crohn's disease in pCD+ and pCD- patients in the IBD-BR.

Lower rates of drug intake in pCD+ patients in IBD-BR

pCD+ patients were less likely to be actively prescribed a number of CD medications: oral steroids, Infliximab, Adalimumab, Vedolizumab, and Mesalazine (P-value < 0.0036; Table 1.1). Anti-TNF therapies, including Infliximab and Adalimumab, are among the first-line drugs for perianal fistulas, and lead to fistula healing in 50% of pCD patients (in combination with other surgical procedures) [24–27]. On the other hand, oral steroids are known to be ineffective for fistula closure and may even exacerbate perianal abscess [28]. Lower drug intake could therefore be attributed to drug inefficacy, but it could also be a contributing factor to pCD.

pCD+ patients are enriched for six extraintestinal manifestations

pCD+ patients were enriched for extra-intestinal manifestation compared to pCD- patients (26.3% versus 19.6%; P-value < 0.05). Enteropathic arthritis was the most most prevalent extraintestinal manifestation among pCD+ patients, followed by serious infections and psoriasis. In total, six extraintestinal manifestations showed significant enrichment in pCD+ versus pCD- patients (P-value < 0.005; Table 1.1). The association between extraintestinal manifestations was stronger in female participants, possibly since extraintestinal manifestations were more prevalent in female participants overall (odds ratio=1.3 in males versus 1.6 in females).

Surgical burden of pCD

Combined surgical and medical interventions represent some of the few effective interventions available to pCD patients. Different surgical options are available to perianal disease patients depending on its anatomical features, complications and disease severity. Exploration under anaesthesia and seton insertion are the typical first-line management options, and further medical or surgical interventions are based on initial exploration [29]. As expected, 2,971 pCD patients (66.8%) had undergone any type of surgical intervention compared to 3,637 (37.2%) of pCD- participants. In total, almost half the pCD+ patients with operative history had undergone one of three pCD-related surgical procedures (1431 patients; 48.2%): drainge of perianal abscess, insertion of seton, or drainage of fistula. Perianal abscess drainage was the most common: 808 pCD+ patients (27.2% of surgically-operated patients) underwent at least one perianal abscess drainage operation, followed by insertion of a seton suture (744 pCD+ patients; 25%), followed by perianal fistula repair operation (438 pCD+ patients; 14.7%).

	pCD+(%)	pCD-(%)	P-value
Extraintestinal Manifestations			
Primary Sclerosing Cholangitis	25 (0.6)	72 (0.8)	0.29
Enteropathic Arthritis	413 (9.7)	635 (6.7)	2.1×10^{-9}
Erythema Nodosum	199 (4.6)	222 (2.3)	7.7×10^{-13}
Iritis	183 (4.2)	242 (2.5)	1.1×10^{-7}
Orofacial Granulomatosis	153 (3.6)	162 (1.7)	2.4×10^{-11}
Psoriasis	311 (7.2)	518 (5.4)	6×10^{-5}
Ankylosing Spndylitis	110 (2.6)	238 (2.5)	0.89
Multiple Sclerosis	9 (0.2)	27 (0.3)	0.53
Lymphoma	18 (0.4)	36 (0.4)	0.85
Serious Infections	320 (7.4)	465 (4.9)	3.2×10^{-9}
Drugs			
Azathioprine	1373 (41.4)	2649 (42.1)	0.49
Mercaptopurine	271 (35.6)	564 (36.1)	0.83
Methotrexate	192 (28.6)	404 (35.4)	4×10^{-3}
Ciclosporin	3 (8.1)	3 (7.5)	1
Infliximab	1207 (50.9)	1622 (55.7)	6×10^{-4}
Adalimumab	740 (47.8)	1368 (54.2)	8×10^{5}
Golimumab	9 (56.2)	9 (50)	0.98
Vedolimumab	236 (67.8)	424 (77.4)	2×10^{-3}
Ustekinumab	171 (69.2)	209 (72.6)	0.45
Mesalazine	528 (30)	1649 (43.7)	$< 2.2 \times 10^{-16}$
Oral Steroids	304 (11.3)	823 (14.2)	3×10^{-4}

Table 1.1 Drug intake and extraintestinal manifestations in pCD+ and pCD- patients in the IBD-BR. Percentage of patients are shown between paranthese.

pCD prevalence decreased over time

Understanding pCD prevalence over time is important to understand the impact of pCD on patients and healthcare providers. Previous work has shown reduced pCD incidence over the last decade [ref], which was partly attributed to improved treatment options. To investigate this observation in the IBD-BR, I partitioned participants according to their year of CD diagnosis into two-year groups (e.g. 2006-2008), and calculated prevalence estimates in each period. As expected, the number of diagnosed participants varied between year groups, with fewer participants diagnosed in the early years. 100-230 participants per two years were diagnosed in the years from 1980 to 2000, but this rose to 497-734 per two years in the years from 2010 to 2020. Notably, point prevalence estimates decreased starting from the year 2010 onwards. The mean point prevalence between 1980 to 2010 decreased significantly from 35.9% to 25.1% between 2010 to 2020 (P-value $< 2 \times 10^{-16}$; Figure 1.2). The decrease in prevalence remained significant when mean point prevalence was calculated between 2010 to 2016 only (mean prevalence=26.8%), or between 2010 to 2014 only (mean prevalence=28.1%).

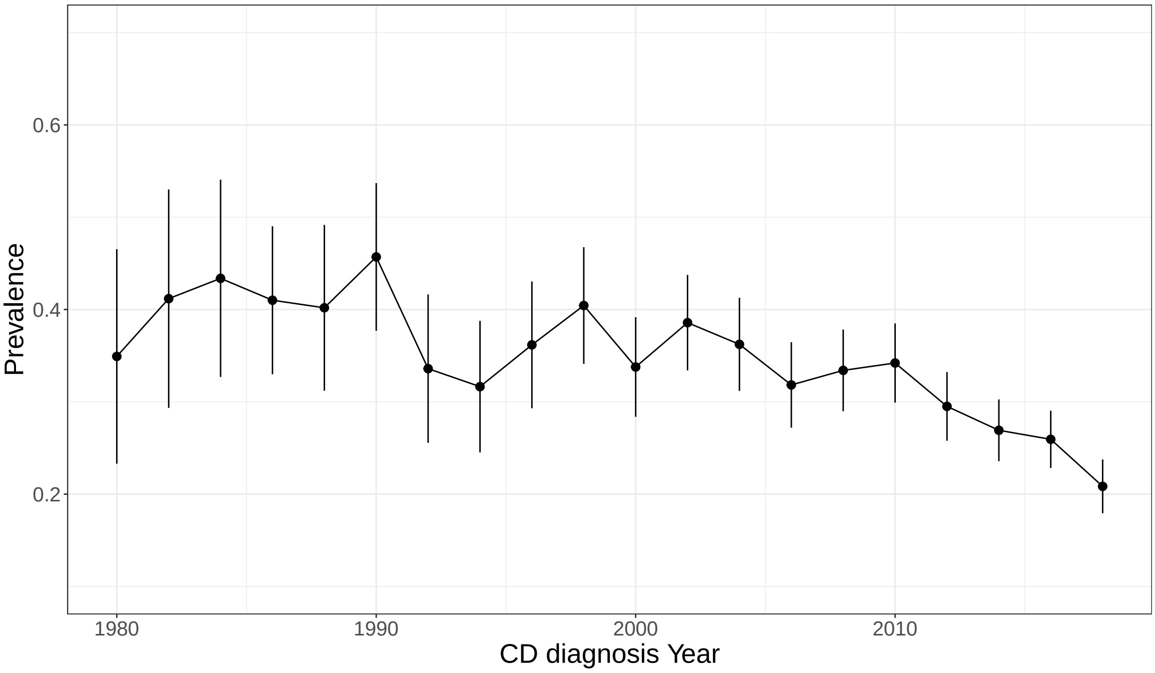

Fig. 1.2 Prevalence per year of CD diagnosis, partitioned into two-year groups. 95% confidence intervals around point estimates are calculated using a bootstrap procedure, whereby participants where resampled 1,000 times within each two-year group.

A decrease in pCD prevalence has been observed previously [30]. However, such a decrease has not been precisely quantified, partly due to the relatively smaller sample sizes of most studies [31–34]. A potential limitation of this analysis is that censored data may contribute to the observed decrease in pCD prevalence in later years. pCD does not always present at the time of diagnosis, which may bias prevalence estimates downwards in the years after 2010. However, the consistent decrease in prevalence even when I excluded patients diagnosed after 2014 indicates that the contribution of censored observation is likely minimal. Although some patients may develop perianal symptoms up to 20 years after diagnosis, the cumulative probability of developing pCD does not increase significantly after 5 years. It is therefore unlikely that pCD prevalence for IBD-BR patients diagnosed between 2010 to 2014 will increase significantly [34].

1.3.3 UKIBDGC and IBD-BR definitions of pCD are similar

Similar to IBD-BR, the UKIBDGC can be used to perform a GWAS between pCD+ and pCD-. Although designed as a case-control cohort for IBD patients and healthy controls, the UKIBDGC reports a number of clinical and phenotypic characteristics of IBD patients. For each IBD participant, disease subtype diagnosis, and location (inclduding perianal diseasel) are recorded.

As a large consortium of data collected from several hospitals and recruitment centres, it is unclear whether the criteria for assigning pCD status is consistent between the different centers, and more importantly if it matches the criteria used to assign pCD status to IBD-BR patients. Heterogeneity in pCD status definition can often arise from different diagnostic criteria being applied, or different times of phenotype update between recruitment centres. Ensuring the consistency of pCD status between UKIBDGC and IBD-BR is crucial to minimise the heterogeneity of case/control definition between the two cohorts and maximise the statistical power from a meta-analysis between the two cohorts. To assess this, participant overlap between IBD-BR and UKIBDGC can be leveraged to understand the level of agreement in pCD status assignment. Since participant identifiers are not mapped across studies, genetic similarity of individuals across cohorts can instead be leveraged to identify overlapping participants (Methods).

Out of 971 overlapping CD participants, only one was assigned pCD+ in IBD-BR and pCD- in UKIBDGC. A total of 432 individuals had missing or "Unknown" perianal involvement, 406 of which were"Unknown" in both cohorts (Table 1.2). This strong agreement in pCD status indicates that both cohorts assign pCD status in a similar fashion.

Patients who answered “Yes” in both studies were not enriched in any particular type of perianal involvement (e.g. 51.9% of overlapping individuals reported either simple or complex fistula versus 52% in IBD-BR). 27% of overlapping patients reported only skin tags, fissures or ulcer, indicating that milder forms of pCD were also included in the UKIBDGC assignment of pCD+ status.

Table 1.2 Number of overlapping individuals between UKIBDGC and IBD-BR who answered Yes, No or Unknown to *Ever had perianal involvement?*

	UKIBDGC		
IBD-BR	Yes	No	Unknown
Yes	201	0	1
No	1	337	6
Unknown	6	13	406

1.4 Genome-wide association analysis of pCD

1.4.1 Defining pCD+ cases

Unlike GWAS of CD, where robust diagnostic criteria are applied to clearly demarcate cases and controls, in GWAS of disease pCD it is not obvious which perianal manifestations should be considered pCD+ cases. The IBD-BR questionnaire reports several types of pCD manifestations, including skin tags, fissures or ulcers, perianal abscess, and simple and complex fistulas. In this chapter, my aim is to perform a pCD meta-analysis between IBD-BR and UKIBDGC, and therefore similarity in pCD+ case definition across the cohorts is an important consideration to ensure the the robustness of genome-wide significant hits. In the previous section, I showed that leveraging the overlapping individuals can give an insight into the composition of the UKIBDGC pCD+ cases. This showed that UKIBDGC pCD+ cases were not particularly enriched in any particular type of perianal manifestations. Additionally, when I inspected the UKIBDGC questionnaire used to collect perianal manifestations data, I found that the relevant question appeared to include all types of perianal manifestations: *"Ever had perianal fistula (incl recto-vaginal), abscess, anal ulcer or significant anal stenosis?"*.

1.4.2 IBD-BR

Although clinical and phenotypic data are available for all participants, not all participants have been genotyped in the current release (04/04/2022). From a total of 15,152 participants with CD diagnosis, 9,458 European ancestry participants with perianal involvement data were genotyped. To ensure that pCD- controls do not include recently diagnosed CD patients who may develop perianal disease in the near future, I excluded pCD- controls diagnosed with CD less than 5 years before the last clinical review. This choice was informed by previous studies that showed that the cumulative risks of developing perianal disease 5 years and 10 years after diagnosis are similar [34]. This resulted in a total of 6833 participants (2,664 pCD+ cases and 4,169 pCD- controls). After these filters were applied, the composition of genotyped pCD+ cases cohort matched the overall composition of all participants with perianal involvement information. 53.6% (1480) of genotyped pCD+ individuals had either a simple or complex perianal fistula, and 41.2% (1098) had perianal abscess. Together, patients with perianal fistula or abscess account for 74.9% (1995) of genotyped pCD+ cases.

I performed GWAS between pCD+ cases and pCD- controls using REGENIE and used four European-ancestry genotypic principal components and sex as covariates. I removed variants with imputation INFO score < 0.4 and minor allele frequency (MAF) < 0.01, leaving 9,777,139 variants for association analysis (see Methods for detailed genotype and imputation QC). None of the tested variants achieved genome-wide significant association (P-value $< 5 \times 10^{-8}$). There was moderate evidence of genomic inflation (median χ^2=0.49; λ_{GC}=1.08).

1.4.3 UKIBDGC

A total of 8,078 patients of European ancestry were diagnosed with CD in the UKIBDGC, of which 6550 had perianal involvement information. To minimise sample overlap with the IBD-BR, I removed overlapping individuals between UKIBDGC and IBD-BR (Methods), and performed GWAS with the remaining individuals (1303 pCD+ and 4761 pCD-).

I performed GWAS similar to the IBD-BR analysis, with the difference that UKIBDGC samples genotyped on different arrays were analysed seperately (HCE and GWAS1; Methods).A total of 8,916,200 and 8,897,554 variants were tested in HCE and GWAS1, respectively. No variants achieved genome-wide significant association (P-value $< 5 \times 10^{-8}$). There was no evidence of genomic inflation (HCE: median χ^2=0.47; λ_{GC}=1.04; GWAS1: median χ^2=0.46; λ_{GC}=1.01).

1.4.4 Meta-analysis between UKIBGC and IBD-BR: a genome-wide significant locus at 6p21.32

I used METAL to perform a fixed-effects meta-analysis between summary statistics from IBD-BR, and the two UKIBDGC summary statistics HCE and GWAS1, with a total of 3,967 pCD+ cases and 8,930 pCD- controls. There was a total of 8,473,930 overlapping variants across the meta-analysed summary statistics, and an additional 1,645,123 were unique to one of the studies, 42.7% of which were indels. Given that 16% of variants were unique to one of cohorts, I did not remove them from their respective summary statistics file. It is important to note, however, that this choice may favour variants that are available in all studies. Additionally, I enabled METAL's `GENOMICCONTROL ON` option to correct genomic inflation in each of the two summary statistics before performing the meta-analysis. The resulting meta-analysed summary statistics showed no evidence of genomic inflation (λ_{GC}=1.03; Figure 1.3).

Studies	SNP	Indel	Total
pcd.ibdbr	8,626,072	1,150,933	9,777,005
pcd.gwas1	8,307,857	589,198	8,897,055
pcd.allhce	8,325,721	589,997	8,915,718

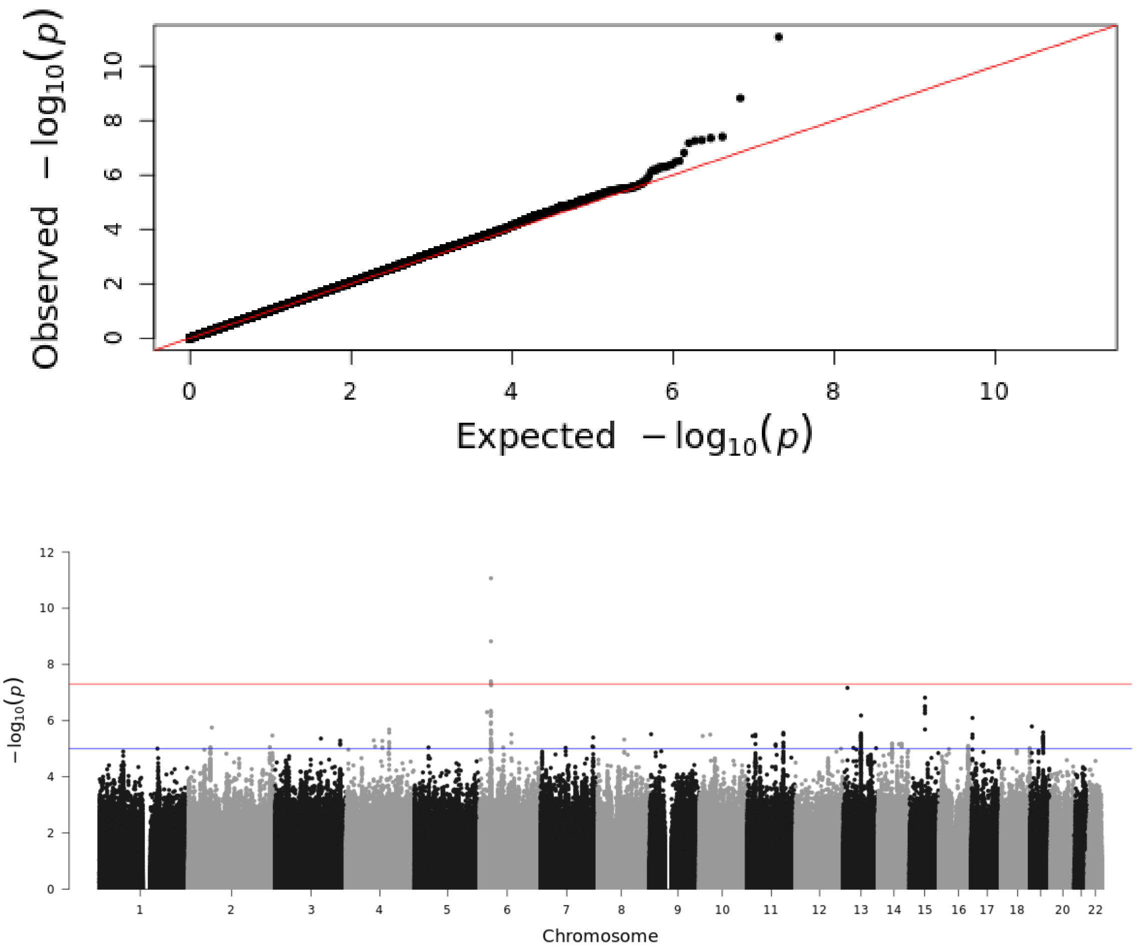

Fig. 1.3 (a) Quantile-quantile plot for the meta-analysis between UKIBDGC and IBD-BR cohorts, suggesting a good fit to the uniform distribution, and showing no evidence of genomic inflation (median χ^2=0.47; λ_{GC}=1.03; median χ^2 was calculated by converting P-values to χ^2 values using the function `qchisq(P, df=1,lower.tail=F)` in R v4.1.0). (b) Manhattan plot of meta-analysis between IBD-BR and UKIBDGC. pCD+ cases are defined as CD patients with any type of perianal involvement and pCD- controls are defined as CD patients with no perianal involvement.

Four variants in the 6p21.32 locus showed genome-wide significant association (P-value $< 5 \times 10^{-8}$; index variant rs115378818 P-value=8.6×10^{-12}; Table 1.3). None of the variants showed significant heterogeneity of effect size between the constituent cohorts (Bonferroni-corrected P_{het} < 0.008). All six variants were well-imputed across the constituent cohorts (INFO score > 0.7).

Table 1.3 genome-wide significant variants in the 6p21.32 locus. Odds ratio and their 95% confidence intervals are shown. MAF=minor allele frequency.

CHROM	GENPOS	Allele1	OR	P	MAF
6	32,205,822	C	1.45 (1.27 - 1.66)	4.0e-08	0.05
6	32,243,461	C	1.38 (1.23 - 1.55)	4.4e-08	0.08
6	32,279,268	G	1.57 (1.36 - 1.82)	1.5e-09	0.05
6	32,333,650	T	1.78 (1.51 - 2.1)	8.6e-12	0.04

All four variants had a low minor allele frequency (MAF), ranging from 0.04 to 0.08 in the constituent cohorts (Table 1.4). Over the last 15 years, hundreds of GWAS studies have emphasised the importance of guarding against spurious associations driven by low-frequency variants. In some studies, these associations were later found to be false positives [35]. To minimise this risk and ensure the robustness of the associated variants, one may compare the within-cohort genetic properties of these variants to their expected genetic properties in a matching general population. A mismatch between the general population and the cohorts may indicate that the detected association may indeed be artefactual. To this end, I performed a number of checks to ensure the robustness of the association at 6p21.32. First, I checked whether the variants' MAF is comparable between the constituent cohorts (IBD-BR and UKIBDGC). Second, I compared cohort MAFs to the general population MAFs (in non Finnish Europeans). Third, I investigated whether the association strength is consistent with the expected LD structure in non-Finnish Europeans.

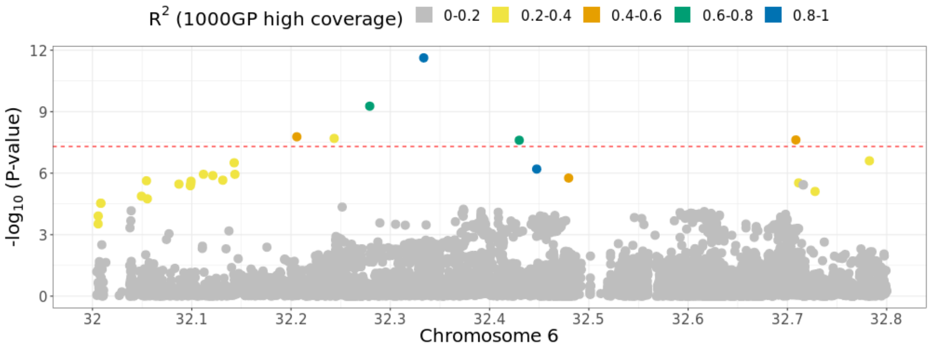

Fig. 1.4 Regional association plot showing meta-analysis association P-values in the 6p21.32 regions. Variants are coloured by R^2 with the index variant, derived from 1000 Genomes Project High Coverage study (Non-Finnish Europeans)

Associated variants at 6p21.32 match population allele frequencies

None of the variants showed significant differences in MAF within the constituent cohorts or between the cohorts and non-Finnish Europeans in the general population (Table 1.4; 1000 Genomes Project and GnomAD; see Methods for how MAF deviation was assessed). Despite no significant deviation from expected population MAF, all variants had consistently higher cohort than population MAFs. Such a systematic increase in MAF may be expected if these variants are also associated with CD risk, since the constituent cohorts are composed entirely of CD patients. In support of this, I found that three of the six variants are significantly associated with CD risk in an independent CD GWAS study, partly explaining the higher cohort MAFs (de Lange et al. 2017; the three other variants were not imputed in the CD study). Moreover, the minor alleles had a risk-increasing effect on both CD and pCD, but the magnitude of effect was significantly higher for CD ($P_{het} < 0.008$).

Table 1.4 Minor allele frequencies for the genome-wide significant variants in the 6p21.32 locus in meta-analysis constituent cohorts, and in two population cohorts (1000 Genomes Project and GnomAD; Non-Finnish Europeans). Association P-value with Crohn's Disease from de Lange et al. 2017 are shown for variants reported in the study.

	Study Cohorts			Population cohorts		
SNP	IBD-BR	HCE	GWAS1	GnomAD	1000GP	CD P-value
6:32333650_C_T	0.033	0.041	0.042	0.012	0.012	NA
6:32279268_T_G	0.044	0.050	0.050	0.020	0.015	4.3×10^{-42}
6:32205822_T_C	0.053	0.054	0.053	0.028	0.021	NA
6:32243461_G_C	0.074	0.076	0.081	0.052	0.036	NA
6:32708532_A_C	0.038	0.042	0.041	0.024	0.020	5×10^{-9}
6:32429885_C_A	0.036	0.050	0.049	0.020	0.016	4.2×10^{-31}

Association signal at 6p21.32 matches expected linkage disequilibrium pattern in non-Finnish Europeans

Although the six variants spanned a 500 kbp region, they displayed high LD with rs115378818, as the 6p21 region is known to exhibit long-range LD (Figure 1.4). Overall, the LD structure and association signal in 6p21.32 matched the expected LD structure in non-Finnish Europeans. First, I found that R^2 values derived from the cohorts were highly correlated with R^2 derived from non-Finnish Europeans (1000 Genomes Project; Pearson correlation coefficient $R^2 > 0.6$). Second, the association signal for the six variants matched the expected LD pattern in 1000GP. Specifically, P-values of nominally associated variants (P-value < 5×10^{-4}) were correlated with their LD with rs115378818 (Figure 1.5).

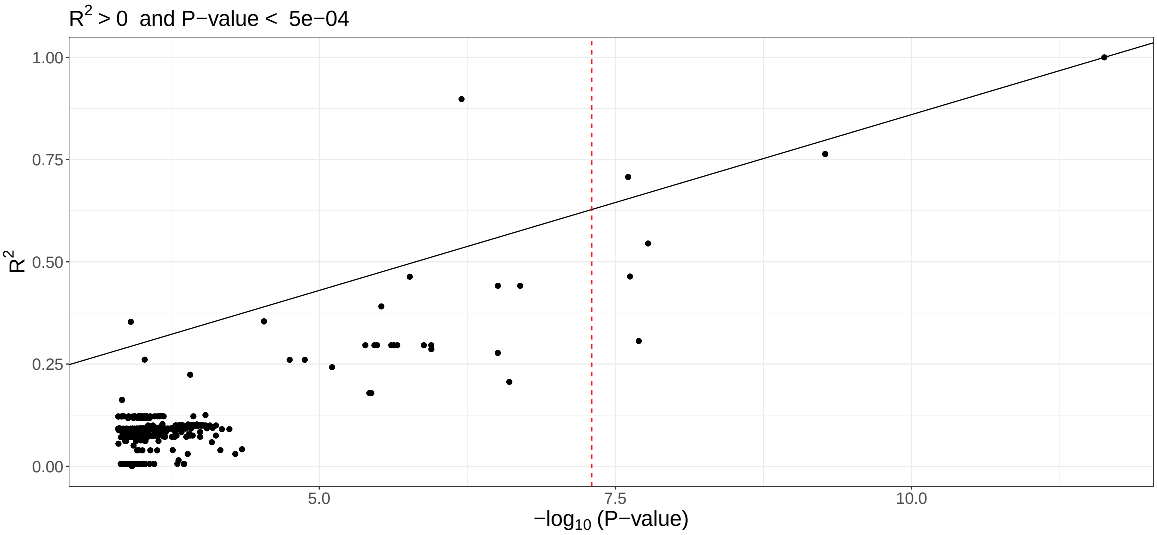

Fig. 1.5 association P-value for all variants in a 1mbp window around rs115378818 (P-value < 5×10^{-4}) on the x-axis, and R^2 of variants with rs115378818 on the y-axis (derived from 1000GP). Vertical red-line indicates the genome-wide significant threshold (5×10^{-8}).

1.4.5 Association at 6p21.32 is robust to more severe pCD+ definitions

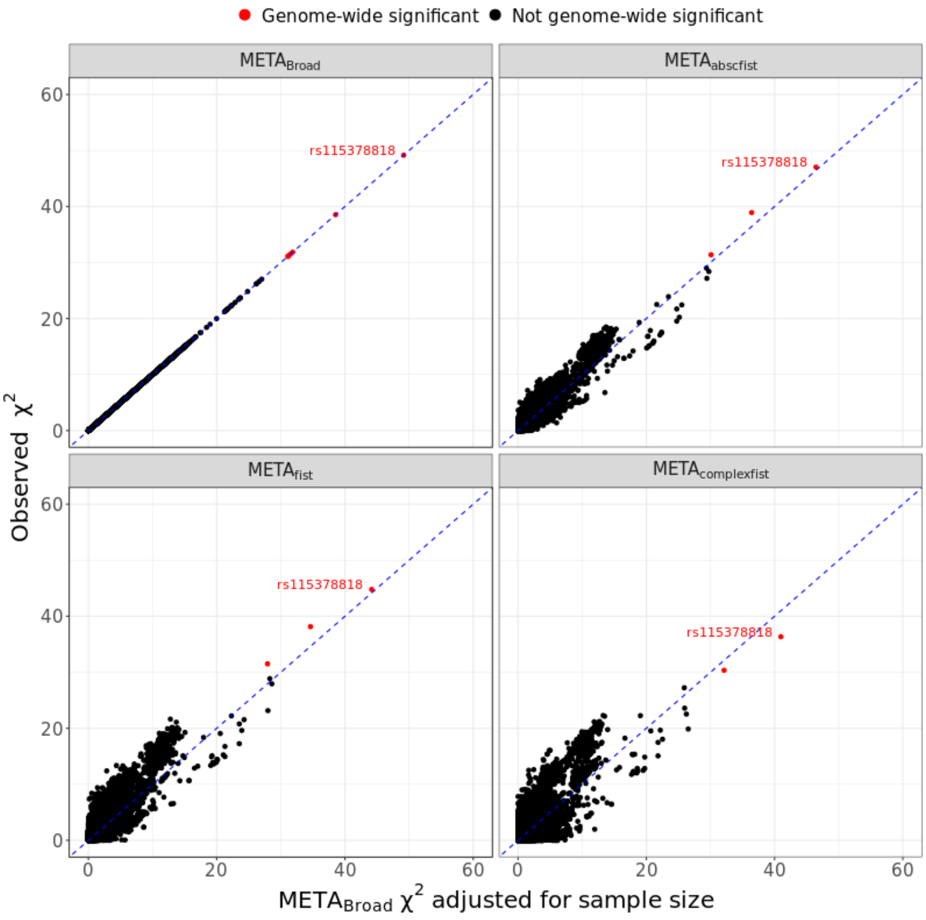

Fig. 1.6 association P-value for all variants in a 1mbp window around rs115378818 (P-value $< 5 \times 10^{-4}$) on the x-axis, and R^2 of variants with rs115378818 on the y-axis (derived from 1000GP). Vertical red-line indicates the genome-wide significant threshold (5×10^{-8}).

The IBD-BR provides information about the type of perianal involvement each patient presents with. To understand the effect of different definition criteria, I investigated how the meta-analysed association signal at 6p21.32 is sensitive to different definitions of pCD+ cases in the IBD-BR. In addition to the broad-definition mata-analysis described in the previous section ($META_{broad}$), I performed a meta-analysis between UKIBDGC and IBD-BR using three additional pCD+ definitions that have an increasingly severe impact on patients: pCD+ as abscess or simple or complex fistula only ($META_{abscfist}$), as simple or complex fistula only ($META_{fist}$), and as complex fistula only ($META_{complexfist}$).

Since the stricter definitions resulted in a reduction in the number of pCD+ cases, a proportional decrease in association test statistic (χ^2) should also be expected. Under the hypothesis that the stricter-definition meta-analyses are simply a subset of $META_{broad}$, the χ^2 observed in any definition meta-analysis should match χ^2 from $META_{broad}$ adjusted for the reduction in sample size (I will refer to this as $\chi^2_{Broad,n}$; see Methods for how this adjustment was performed).

In a 1mbp window centred around rs115378818, I compared the χ^2 observed in each of the three stricter-definition meta-analyses to $\chi^2_{Broad,N}$. All six genome-wide significant variants achieved the expected association in the stricter definition meta-analyses. For example, rs115378818 remained genome-wide significant in $META_{fist}$ despite the decrease in sample size (1,234 fewer pCD+ cases; observed P-value=2.2 × 10^{-11}; broad-definition P-value adjusted for sample size=3 × 10^{-11}; Figure 1.6). More broadly, across all variants in 6p21.32, I observed strong correlation between observed χ^2 in the stricter-definition meta-analyses and $\chi^2_{Broad,n}$, which shows the robustness of the association signal against different definitions of pCD+ cases in IBD-BR ($META_{abscfist}$=0.95; $META_{fist}$=0.92,$META_{complexfist}$=0.84).

1.4.6 pCD is associated with HLA allele DRB1*01:03

To link the pCD-associated locus to an HLA allele, I performed association analyses between pCD status and class I and II HLA alleles, both at the allele group and specific allele levels (2-digit and 4-digit resolutions; see Methods for HLA imputation). Similar to the genome-wide association analysis, I performed the HLA association analyses separetely for IBD-BR and UKIBDGC and subsequently meta-analysed the summary statistics (effect sizes and standard errors).

None of the tested HLA alleles achieved genome-wide significance (P-value < 5 × 10^{-8}) within the cohorts or in the meta-analysis. At the allele group level, DRB1*01 had the strongest association. At a specific allele level, HLA-DRB1*01:03 had the most significant association, and had a stronger association compred to its allele group ($P_{DRB1*01:03} = 1.8 \times 10^{-6}$; $P_{DRB1*01} = 1.4 \times 10^{-3}$). I tested both dominant and additive modes of inheritance and found that the dominant model achieved better model fit at both the allele group and specific allele levels ($AIC_{dominant} < AIC_{additive}$; Table 1.5).

Table 1.5 Top HLA allele associations with pCD status. Both allele groups (2-digit resolution; first two rows) and specific alleles (4-digit resolution; second two rows) are shown. Meta-analysed P-values and odds ratios between UKIBDGC and IBD-BR cohorts are shown (and their 95% confidence intervals). Both dominant and additive modes of inheritance for DRB1*01 and DRB1*01:03 were tested. Akaike Information Content (AIC), a measure of model fit, is shown in the last theree columns for each of the three constituent cohorts, and shows a better fit for the dominant model (lower AIC).

HLA Allele	Inheritance	Odds Ratio	P-value	AIC (IBDBR)	AIC (HCE)	AIC (GWAS1)
DRB1*01	Dominant	1.2 (1.1 - 1.3)	9.4e-04	9127.887	3901.092	1783.907
DRB1*01	Additive	1.1 (1.1 - 1.2)	1.5e-03	9128.485	3901.413	1783.907
DRB1*01:03	Dominant	1.6 (1.3 - 1.9)	5.3e-07	9122.007	3651.987	1615.647
DRB1*01:03	Additive	1.5 (1.3 - 1.8)	1.8e-06	9122.825	3653.500	1615.647

Conditioning association signal on DRB1*01:03

After conditioning on rs115378818, I did not observe an association with DRB1*01:03, indicating that the DRB1*01:03 association is completely accounted for by rs115378818 ($P_{DRB1*01:03|rs115378818}$=0.61). Conversely, DRB1*01:03 did not completely account for the rs115378818 association. When I conditioned the rs115378818 association on DRB1*01:03, rs115378818 remained nominally associated with pCD ($P_{rs115378818} = 2.4 \times 10^{-12}$ and $P_{rs115378818|DRB1*01:03} = 1.1 \times 10^{-5}$; Figure 1.7). Taken together, this evidence suggests that DRB1*01:03 is indeed associated with pCD and that it partly explains the observed genome-wide association signal, but that another HLA allele may additionally contribute to the observed association signal.

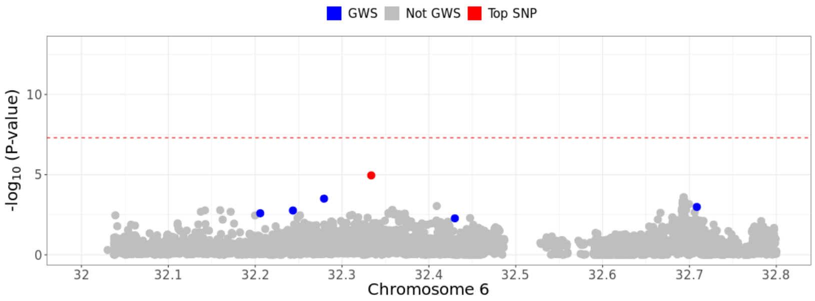

Fig. 1.7 Residual association signal after including DRB1*01:03 as a covariate. Blue points represent variants with a genome-wide significant association in the model that does not include DRB1:01*03. The red point indicates rs115378818 (index variant).

1.5 Discussion

In this chapter, I have used two IBD cohort (UKIBDGC and IBD-BR) with rich clinical data to identify the genetic variants associated with risk of pCD, a CD sub-phenotype that affects up to 40% of CD patients, with significant health and lifestyle impact on patients. Both cohorts report similar epidemiological characteristics of pCD+ patients compared to pCD- patients. Furthermore, IBD-BR with its deeper clinical phenotyping shows lower anti-TNF intake and higher extraintestinal manifestations in pCD+ patients, two observations that should be further investigated in future work. More importantly, the availability of genotypic data enabled the identification of overlapping individuals between the two studies, and confirming a remarkable consistency in pCD status assignment despite the large number of hospital and recruitment centers that contributed patient data and samples to these two cohorts.

By meta-analysing pCD GWAS data from both cohorts, I have identified a genome-wide significant in the highly polymorphic Major Histocompatibility Locus (MHC) at 6p21.32 associated with pCD risk. Given the relative low frequency of genome-wide significant variants, I have performed a number of "post-GWAS" checks to ensure the veracity of the assocation. Genome-wide significant variants allele frequencies matched expected non-Finnish Europeans allele frequencies, and the P-values at the locus matched their expected LD pattern in non-Finnish Europeans. Moreover, the association was robust to pCD+ status defined with increasingly stricter criteria of perianal involvement, showing its robustness to heterogeneous case definitions.

Previous GWAS studies have attempted to identify pCD-associated loci, but none have achieved genome-wide significance [36, 37]. For example Akhlaghpour et al. identified a Complement Factor B (*CFB*) coding variant that was nominally associated with pCD risk (rs4151651;P-value=9.35×10^{-6}). This variant is unlikely to account completely for our meta-analysis association. First, the rs115378818 and rs4151651 are in weak LD (R^2=0.24). Second, Upon conditioning on the *CFB* signal (rs114969413; R^2 with index variant = 0.99), there was a residual association in our meta-analysis (rs115378818 P-value=2.1×10^{-6}; Figure 1.8).

To follow up on our association signal, I tested the association between pCD status and different HLA alleles imputed from genotype data. HLA alleles can be mapped at different resolutions from allele groups (2-digit) to specific alleles (4-digit) to amino acid resolution. The current release of our HLA imputation data includes 2-digit and 4-digit HLA alleles.

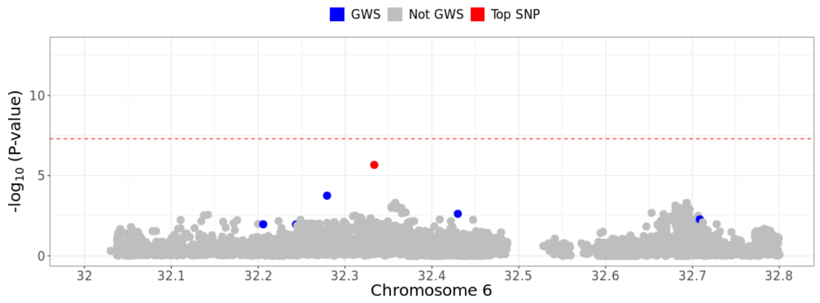

Fig. 1.8 Residual association signal after including rs114969413 (a *CFB* locus variant) as a covariate. Blue points represent variants with a genome-wide significant association in the model that does not include rs114969413. The red point indicates rs115378818 (index variant).

HLA-DRB1*01:03 showed a strong (but not genome-wide significant) association with pCD risk. Conditioning on the index variant (rs115378818) completely explained the association of HLA-DRB1*01:03, but HLA-DRB1*01:03 did not completely account for the index variant association.

Several reasons may explain this. First, the association signal at 6p21.32 may be explained by multiple HLA alleles, and therefore conditioning on a signle HLA allele cannot completely explain the 6p21.32 signal. The highly polymorphic nature of most HLA genes often makes it difficult to completely attribute disease risk to a single HLA allele. This has been previously observed for Rheumatoid Arthritis (RA), where several *HLA-DRB1* alleles confer risk to RA [38]. Second, higher HLA allele resolution at the amino acid level may better account for the pCD signal [39]. Interestingly, using only three amino acid positions in a predictive model of RA risk provided identical prediction with a model that included all HLA-DRB1 allleles [40]. Therefore, a follow-up to my work should explore the association of HLA-DRB1 alleles with pCD at an amino acid level. The presence of a small number of amino acid substitution is a reasonable explanation for the 6p21.32 association that was not completely account by HLA-DRB1*01:03.

References

[1] Wee Khoon Ng, Sunny H Wong, and Siew C Ng. Changing epidemiological trends of inflammatory bowel disease in asia. *Intest. Res.*, 14(2):111–119, April 2016.

[2] Tsunekazu Mizushima, Mihoko Ota, Yasushi Fujitani, Yuya Kanauchi, and Ryuichi Iwakiri. Diagnostic features of perianal fistula in patients with crohn's disease: Analysis of a japanese claims database. *Crohns Colitis 360*, 3(3):otab055, July 2021.

[3] Javier Salgado Pogacnik and Gervasio Salgado. Perianal crohn's disease. *Clin. Colon Rectal Surg.*, 32(5):377–385, September 2019.

[4] Pauline Wils, Ariane Leroyer, Mathurin Fumery, Alonso Fernandez-Nistal, Corinne Gower-Rousseau, and Benjamin Pariente. Fistulizing perianal lesions in a french population with crohn's disease. *Dig. Liver Dis.*, 53(5):661–665, May 2021.

[5] Tim W Eglinton, Murray L Barclay, Richard B Gearry, and Frank A Frizelle. The spectrum of perianal crohn's disease in a population-based cohort. *Dis. Colon Rectum*, 55(7):773–777, July 2012.

[6] Samuel O Adegbola, Lesley Dibley, Kapil Sahnan, Tiffany Wade, Azmina Verjee, Rachel Sawyer, Sameer Mannick, Damian McCluskey, Nuha Yassin, Robin K S Phillips, Philip J Tozer, Christine Norton, and Ailsa L Hart. Burden of disease and adaptation to life in patients with crohn's perianal fistula: a qualitative exploration. *Health Qual. Life Outcomes*, 18(1):370, November 2020.

[7] Julian Panes, Walter Reinisch, Ewa Rupniewska, Shahnaz Khan, Joan Forns, Javaria Mona Khalid, Daniela Bojic, and Haridarshan Patel. Burden and outcomes for complex perianal fistulas in crohn's disease: Systematic review. *World J. Gastroenterol.*, 24(42):4821–4834, November 2018.

[8] G C Braithwaite, M J Lee, D Hind, and S R Brown. Prognostic factors affecting outcomes in fistulating perianal crohn's disease: a systematic review. *Tech. Coloproctol.*, 21(7):501–519, July 2017.

[9] Laurent Peyrin-Biroulet, Edward V Loftus, Jr, Jean-Frederic Colombel, and William J Sandborn. The natural history of adult crohn's disease in population-based cohorts. *Am. J. Gastroenterol.*, 105(2):289–297, February 2010.

[10] Michael Scharl, Gerhard Rogler, and Luc Biedermann. Fistulizing crohn's disease. *Clin. Transl. Gastroenterol.*, 8(7):e106, July 2017.

[11] Tim W Eglinton, Murray L Barclay, Richard B Gearry, and Frank A Frizelle. The spectrum of perianal crohn's disease in a population-based cohort. *Dis. Colon Rectum*, 55(7):773–777, July 2012.

[12] Michael Scharl, Sandra Frei, Theresa Pesch, Silvia Kellermeier, Joba Arikkat, Pascal Frei, Michael Fried, Achim Weber, Ekkehard Jehle, Anne Rühl, and Gerhard Rogler. Interleukin-13 and transforming growth factor β synergise in the pathogenesis of human intestinal fistulae. *Gut*, 62(1):63–72, January 2013.

[13] The IBD BioResource. The ibd bioresource protocol version 8, 2021.

[14] The IBD BioResource. The ibd bioresource questionnaire version 7, 2021.

[15] The IBD BioResource. What is the ibd bioresource?, 2022.

[16] The UK IBD Genetics Consortium. Uk ibd genetics consortium aims, 2023.

[17] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[18] PLINK. Plink qc high ld regions, 2023. Accessed on 20/10/2023.

[19] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, November 2010.

[20] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell, and Jonathan Marchini. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.*, 53(7):1097–1103, July 2021.

[21] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, Susan Fairley, Alexi Runnels, Lara Winterkorn, Ernesto Lowy, Human Genome Structural Variation Consortium, Paul Flicek, Soren Germer, Harrison Brand, Ira M Hall, Michael E Talkowski, Giuseppe Narzisi, and Michael C Zody. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, September 2022.

[22] Dana Duricova, Johan Burisch, Tine Jess, Corinne Gower-Rousseau, Peter L Lakatos, and ECCO-EpiCom. Age-related differences in presentation and course of inflammatory bowel disease: an update on the population-based literature. *J. Crohns. Colitis*, 8(11):1351–1361, November 2014.

[23] Manreet Kaur, Deepa Panikkath, Xiaofei Yan, Zhenqiu Liu, Dror Berel, Dalin Li, Eric A Vasiliauskas, Andrew Ippoliti, Marla Dubinsky, David Q Shih, Gil Y Melmed, Talin Haritunians, Phillip Fleshner, Stephan R Targan, and Dermot P B McGovern. Perianal crohn's disease is associated with distal colonic disease, stricturing disease behavior, IBD-associated serologies and genetic variation in the JAK-STAT pathway. *Inflamm. Bowel Dis.*, 22(4):862–869, April 2016.

[24] Miguel Regueiro and Houssam Mardini. Treatment of perianal fistulizing crohn's disease with infliximab alone or as an adjunct to exam under anesthesia with seton placement. *Inflamm. Bowel Dis.*, 9(2):98–103, March 2003.

[25] Paulo Gustavo Kotze, Idblan Carvalho de Albuquerque, André da Luz Moreira, Wanessa Bertrami Tonini, Marcia Olandoski, and Claudio Saddy Rodrigues Coy. Perianal complete remission with combined therapy (seton placement and anti-TNF agents) in crohn's disease: a brazilian multicenter observational study. *Arq. Gastroenterol.*, 51(4):284–289, October 2014.

[26] A Haennig, G Staumont, B Lepage, P Faure, L Alric, L Buscail, B Bournet, and J Moreau. The results of seton drainage combined with anti-TNFα therapy for anal fistula in crohn's disease. *Colorectal Dis.*, 17(4):311–319, April 2015.

[27] Wolfgang B Gaertner, Alejandra Decanini, Anders Mellgren, Ann C Lowry, Stanley M Goldberg, Robert D Madoff, and Michael P Spencer. Does infliximab infusion impact results of operative treatment for crohn's perianal fistulas? *Dis. Colon Rectum*, 50(11):1754–1760, November 2007.

[28] J H Jones and J E Lennard-Jones. Corticosteroids and corticotrophin in the treatment of crohn's disease. *Gut*, 7(2):181–187, April 1966.

[29] Samuel Adegbola. Medical and surgical management of perianal crohn's disease. *Ann. Gastroenterol.*, 2018.

[30] Sang Hyoung Park, Satimai Aniwan, W Scott Harmsen, William J Tremaine, Amy L Lightner, William A Faubion, and Edward V Loftus. Update on the natural course of fistulizing perianal crohn's disease in a population-based cohort. *Inflamm. Bowel Dis.*, 25(6):1054–1060, May 2019.

[31] Charlène Brochard, Marie-Laure Rabilloud, Stéphanie Hamonic, Emma Bajeux, Maël Pagenault, Alain Dabadie, Agathe Gerfaud, Jean-François Viel, Isabelle Tron, Michel Robaszkiewicz, Jean-François Bretagne, Laurent Siproudhis, Guillaume Bouguen, and Groupe ABERMAD. Natural history of perianal crohn's disease: Long-term follow-up of a population-based cohort. *Clin. Gastroenterol. Hepatol.*, 20(2):e102–e110, February 2022.

[32] Annecarin Brückner, Katharina J Werkstetter, Jan de Laffolie, Claudia Wendt, Christine Prell, Tanja Weidenhausen, Klaus P Zimmer, and Sibylle Koletzko. Incidence and risk factors for perianal disease in pediatric crohn disease patients followed in CEDATA-GPGE registry. *J. Pediatr. Gastroenterol. Nutr.*, 66(1):73–78, January 2018.

[33] Kevin W A Göttgens, Steven F G Jeuring, Rosel Sturkenboom, Mariëlle J L Romberg-Camps, Liekele E Oostenbrug, Daisy M A E Jonkers, Laurents P S Stassen, Ad A M Masclee, Marieke J Pierik, and Stéphanie O Breukink. Time trends in the epidemiology and outcome of perianal fistulizing crohn's disease in a population-based cohort. *Eur. J. Gastroenterol. Hepatol.*, 29(5):595–601, May 2017.

[34] Lester Tsai, Jeffrey D McCurdy, Christopher Ma, Vipul Jairath, and Siddharth Singh. Epidemiology and natural history of perianal crohn's disease: A systematic review and meta-analysis of population-based cohorts. *Inflamm. Bowel Dis.*, 28(10):1477–1484, October 2022.

[35] Meredith E Tabangin, Jessica G Woo, and Lisa J Martin. The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proc.*, 3 Suppl 7(S7):S41, December 2009.

[36] Marzieh Akhlaghpour, Talin Haritunians, Shyam K More, Lisa S Thomas, Dalton T Stamps, Shishir Dube, Dalin Li, Shaohong Yang, Carol J Landers, Emebet Mengesha, Hussein Hamade, Ramachandran Murali, Alka A Potdar, Andrea J Wolf, Gregory J Botwin, Michelle Khrom, International IBD Genetics Consortium, Ashwin N Ananthakrishnan, William A Faubion, Bana Jabri, Sergio A Lira, Rodney D Newberry, Robert S Sandler, R Balfour Sartor, Ramnik J Xavier, Steven R Brant, Judy H Cho, Richard H Duerr, Mark G Lazarev, John D Rioux, L Philip Schumm, Mark S Silverberg, Karen Zaghiyan, Phillip Fleshner, Gil Y Melmed, Eric A Vasiliauskas, Christina Ha, Shervin Rabizadeh, Gaurav Syal, Nirupama N Bonthala, David A Ziring, Stephan R Targan, Millie D Long, Dermot P B McGovern, and Kathrin S Michelsen. Genetic coding variant in complement factor B (CFB) is associated with increased risk for perianal crohn's disease and leads to impaired CFB cleavage and phagocytosis. *Gut*, April 2023.

[37] Manreet Kaur, Deepa Panikkath, Xiaofei Yan, Zhenqiu Liu, Dror Berel, Dalin Li, Eric A Vasiliauskas, Andrew Ippoliti, Marla Dubinsky, David Q Shih, Gil Y Melmed, Talin Haritunians, Phillip Fleshner, Stephan R Targan, and Dermot P B McGovern. Perianal crohn's disease is associated with distal colonic disease, stricturing disease behavior, IBD-associated serologies and genetic variation in the JAK-STAT pathway. *Inflamm. Bowel Dis.*, 22(4):862–869, April 2016.

[38] Vincent van Drongelen and Joseph Holoshitz. Human leukocyte antigen–disease associations in rheumatoid arthritis. *Rheum. Dis. Clin. North Am.*, 43(3):363–376, August 2017.

[39] Julio E Molineros, Loren L Looger, Kwangwoo Kim, Yukinori Okada, Chikashi Terao, Celi Sun, Xu-Jie Zhou, Prithvi Raj, Yuta Kochi, Akari Suzuki, Shuji Akizuki, Shuichiro Nakabo, So-Young Bang, Hye-Soon Lee, Young Mo Kang, Chang-Hee Suh, Won Tae Chung, Yong-Beom Park, Jung-Yoon Choe, Seung-Cheol Shim, Shin-Seok Lee, Xiaoxia Zuo, Kazuhiko Yamamoto, Quan-Zhen Li, Nan Shen, Lauren L Porter, John B Harley, Kek Heng Chua, Hong Zhang, Edward K Wakeland, Betty P Tsao, Sang-Cheol Bae, and Swapan K Nath. Amino acid signatures of HLA Class-I and II molecules are strongly associated with SLE susceptibility and autoantibody production in eastern asians. *PLoS Genet.*, 15(4):e1008092, April 2019.

[40] Soumya Raychaudhuri, Cynthia Sandor, Eli A Stahl, Jan Freudenberg, Hye-Soon Lee, Xiaoming Jia, Lars Alfredsson, Leonid Padyukov, Lars Klareskog, Jane Worthington, Katherine A Siminovitch, Sang-Cheol Bae, Robert M Plenge, Peter K Gregersen, and Paul I W de Bakker. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.*, 44(3):291–296, January 2012.