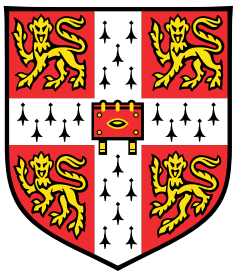
# Molecular and clinical profiling of immune disease genetic loci



**Omar El Garwany**

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Churchill College

November 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
November 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

# Epidemiological and genetic characterisation of perianal Crohn's Disease

## 1.1 Contributions

Genotype and imputation quality control was performed by Dr. Laura Fachal and kinship analysis was performed by Dr. Marcus Tutert as part of the ongoing International IBD Genetics Consrtium GWAS project that is being undertaken in the Anderson laboratory. HLA allele imputation was performed by Dr. Qian Zhang as part of his IBD-BR drug response and disease progression study. I performed all the GWAS analyses, meta-analyses and all downstream analyses described in this chapter.

## 1.2 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 32% cumulative probability of developing pCD [2]. 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 [3–5].

### Clinical picture of pCD

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 [6]. 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 [7]. Moreover, pCD patients, who often require multiple surgical interventions, suffer from high recurrence and relapse rates of pCD. In fact, only one third of pCD patients are estimated to achieve remission [8, 9].

Despite the diversity in presentation and course, pCD patients share a number of Crohn's disease characteristics. pCD is more common among patients with more distal than proximal disease, and 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 [10, 11], which are characterised by a narrowing of the lumen, and development of abdominal fistulas, inflammatory masses and abscess [12]. However, invasive distal CD does not always precede pCD, which can sometimes present a diagnostic challenge in clinical settings. Although 95% of patients will eventually develop luminal disease, an estimated 17.2% of patients initially present with pCD only [13].

### Pathogenesis of pCD

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 gain mesenchymal functions, such as enhanced cell invasion, and migration (reviewed in [14]). 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 $\beta$ and interleukin-13, which have been associated with the initiation of EMT, have also been identified in transitional cells lining pCD fistula tracts [15–17]. Despite these observations, little is understood about the causal biology of pCD. What are the drivers of this transformation? What causes variation in fistulising disease severity? Which biological pathways give rise to fistulas and what facilitates their development into complex branching structures? What is the role of genetic variation in pCD predisposition? In this regard, genome-wide association studies have improved our understanding of the pathophysiology of several complex disease [18]. In the case of pCD, a well-powered GWAS between CD patients who develop pCD and CD patients who do not can help us understand which effector genes and biological pathways are causally linked to pCD risk. Unfortunately, none of the pCD GWAS conducted so far were able to identify genome-wide significant variants associated with pCD risk. Some studies have investigated nominally significant association to better understand pCD biology, but the hypotheses about causal biology remain difficult to reconcile. For example, based on a GWAS of 1,720 CD patients with and without pCD, Kaur et al. [19] report an enrichment of nominally-associated variants in genes implicated in the JAK/STAT pathway, a proinflammatory signalling cascade that has previously been implicated in several autoimmune diseases [20, 21]. More recently, Akhlaghpour et al. [22] found a nominally-associated coding variant that impaired macrophage phagocytosis, and hypothesised that it may contribute to the pathogenesis of fistulising pCD. But overall, there is no clear consensus on the genetic underpinnings of pCD risk.

### Available pCD cohorts

The NIHR IBD-BR is a UK-wide collaborative project that is part of the 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. These aims include understanding the genetics of IBD response to therapy, disease mechanism as well as determinants of disease course [23–25]. So far, the IBD-BR has recruited over 31,000 patients, with epidemiological characteristics, clinical phenotypes, extra-intestinal manifestations, prescribed medications 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 [23]. Disease location data include details of perianal manifestations, which can be used to define a pCD case-control cohort.

Another pCD case control cohort can be defined using data from The UK IBD Genetics Consortium (UKIBDGC), a large collaborative consortium that studies the genetics of IBD susceptibility, progression and drug response. Patients are recruited from multiple UK centres in Cambridge, Edinburgh, Manchester, Newcastle, Exeter, Oxford, London, Dundee and Nottingham, and other sites across the UKB [26]. In additiona to basic epidemiological data such as sex, age, smoking and family history, data on type of IBD, location, surgery, and extraintestinal manifestations is recorded. Disease location data also include whether the disease is located in the perianal region.

Although data on perianal disease are recorded for both cohorts, the depth of clinical phenotyping is different. For example, the IBD-BR, contains information about 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 [23]. 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 [24]. 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 characterise pCD. Using the rich clinical phenotyping in the IBD-BR, I first explored the clinical characteristics of pCD+ patients, which largely conformed with what is known about the disease characteristics of pCD. Moreover, given our limited understanding of the genetic underpinnings of pCD

[27–29, 22, 19], I also performed a pCD GWAS meta-analysis leveraging the pCD cohorts of IBD-BR and UKIBDGC to identify pCD-associated variants. I conclude with an analysis that may partly explain the discovered pCD-associated hit and outline future steps for a more comprehensive understanding of the genetic underpinnings of pCD.

## 1.3 Methods

### 1.3.1 pCD prevalence estimates

IBD-BR patients were diagnosed with CD over several decades, mostly from 1980 till 2018. To investigate the temporal trends of pCD prevalence, I divided participants by year of CD diagnosis into 39 windows, and calculated pCD point prevalence in each two-year period. Additionally,to calculate 95% confidence intervals around each point estimate, I randomly subsampled CD patients 1,000 times using a bootstrap proceducre implemented in the `boot()` functions. Finally, I compared overall trends in prevalence estimate by pooling CD participants diagnosed with CD before and after 2010 and calculated point estimates and confidence intervals as mentioned before. The difference between these overall estimates was then tested using a t-test to deterime if pCD prevalence has significantly decreased before and after 2010.

### 1.3.2 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).

### 1.3.3 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 $\leq 0.01$).

- 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. or
- Monomorphic variants.

### 1.3.4 Sample-level QC

Samples that meet the following criteria were excluded:

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

### 1.3.5 Imputation to TOPMed

The TOPMed imputation server (imputationserver at 1.5.7) was used for UKIBDGC genotype imputation. Alleles at directly genotyped variants with an 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 another HWE filtering step was performed.

### 1.3.6 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 [30], 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 were retained.

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 [31]). 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.

### 1.3.7 Identification of overlapping samples between UKIBDGC and IBD-BR

Identification of duplicate individuals between UKIBDGC and IBD-BR genotyping data was performed with KING [32]. Duplicates were defined as sample pairs with a kinship coefficient > 0.354 as recommended in KING documentations [32]. Estimation of kinship coefficient was performed using post-QC genotyped SNPs (number of variants used for kinship inference between IBD-BR and GWAS1=42,292:, and between IBD-BR and HCE=53,431).

### 1.3.8 Genome-wide association analysis

All genome-wide association analyses were performed using REGENIE v3.2.5 [33] following a 2-step approach. This approach is more computationally efficiency than other approaches that account for cryptic relatedness between individuals, such as linear mixed models. Briefly, in step 1, a whole-genome regression model is fitted using a subset of high-quality genome-wide variants in order to estimate a set of genome-wide predictors that capture a large fraction of phenotypic variance. These predictors are then included as covariates in the single-variant association models tested in step 2, where a larger set of variants of interest are tested for association. I used post-QC genotyped variants in step 1 as recommended by REGENIE documentation ($N_{IBD-BR}$=338,697; $N_{UKIBDGC(HCE)}$=359,209; $N_{UKIBDGC(GWAS1)}$=436,931), and both genotyped and imputed variants in step 2, testing all autosomal chromosomes ($N_{IBD-BR}$=9,777,139; $N_{UKIBDGC(HCE)}$=,916,200; $N_{UKIBDGC(GWAS1)}$=8,897,554). The step 2 model was specified as following: pCD $\sim$ variant + sex + genotypic PCs, using first 4 genotypic PCs. Step 2 reports single-variant association summary statistics.

### 1.3.9 Meta-analysis of IBD-BR and UKIBDGC cohorts

I used METAL to perform the fixed-effects meta-analysis between the IBD-BR and UKIBDGC summary statistics. METAL can perform fixed-effects meta-analysis using one of two different well-established schemes: P-values and effective sample size, or effect sizes and standard errors. The P-value scheme is implemented to enable meta-analysis of GWAS summary statistics that do not report the effect allele, while the effect sizes scheme can be used when each variant's effect size and effect allele are reported. All my pCD GWAS analyses report the effect allele, so I used the effect size scheme of METAL (`SCHEME STDERR`).

There was a total of 8,473,930 overlapping variants across the meta-analysed summary statistics, and an additional 1,645,123 variants that 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.

Table 1.1 Number of SNPs and indels in each of the three GWAS summary statistics.

| **Studies** | **SNP** | **Indel** | **Total** |
|---|---|---|---|
| IBD-BR | 8,626,072 | 1,150,933 | 9,777,005 |
| UKIBDGC (GWAS1) | 8,307,857 | 589,198 | 8,897,055 |
| UKIBDGC (HCE) | 8,325,721 | 589,997 | 8,915,718 |

Moreover, METAL automatically aligns any variants that may be flipped between the meta-analysed summary statistics. METAL also enables filtering of variants to be meta-analysed based on their allele frequencies, which was not necessary since I previously filtered out variants with MAF < 0.01 in each summary statistics file. Finally, given the potential subpopulation stratification in the IBD-BR GWAS ($\lambda_{GC}$=1.08), I enabled a METAL option to correct genomic inflation before performing the meta-analysis (`GENOMICCONTROL ON`) as recommended in METAL's documentation website. There was no evidence of genomic inflation in the meta-analysed summary statistics ($\lambda_{GC}$=1.03).

For each variant, METAL outputs the effect allele, meta-analysed effect size, standard error, and P-values. After performing meta-analyses, it is important to compare the effect sizes between the meta-analysed cohorts. Comparison of both the direction and magnitude of effect sizes gives an indication on how similar the estimated effects of meta-analysed genetic variants are. To formally test this, I used Cochran's Q test of effect size heterogeneity

implemented in METAL. Cochran's Q test asseses two or more effect size estimates and their corresponding standard errors and reports a $\chi^2$ statistic that quantifies the deviation from the null hypothesis that the meta-analysed effect sizes are similar. Depending on the number of meta-analysed studies (in this case 3), a P-value is derived from a theoretical $\chi^2$ distribution with $N-1$ degrees of freedom, where N is the number of meta-analysed studies (heterogeneity of effect P-value $P_{het}$). I used $P_{het}$ to test if the genome-wide significant variants demonstrate heterogeity of effect size between the meta-analysed cohorts. To account for multiple variants being tested, I set a Bonferroni-corrected P-value threshold for rejecting the null hypothesis (P-value < $\frac{0.05}{k}$, where $k$ is the number of variants tested).

### 1.3.10 LD calculation from 1000GP

Reference LD panels obtained from the 1000 Genomes Project High Coverage project [34] were used in the post-GWAS check to study the relationship between LD and association strength at the genome-wide significant locus. $R^2$ values were calculated between the index variant and all variants in the locus. I downloaded VCFs from the 1000GP high coverage and used PLINK v1.9 to compute LD between all variants and the index variant in a 1mbp window. I used unrelated individuals with non-Finnish European ancestry (NFE; N=426). Relevant samples were included in the LD calculation using the following PLINK command:

```
plink --r2 --keep EUR.samples --ld-window-r2 0
```

### 1.3.11 $\chi^2$ comparison between different pCD definitions

In order to compare association statistics from the pCD meta-analysis to meta-analyses performed using more severe pCD+ case criteria (all perianal manifestations, abscess and fistula only, fistula only and complex fistula only), I adjusted the broad-definition $\chi^2$ 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, $\chi^2_{Broad}$ is the broad-definition observed association statistic and $\chi^2_{Broad,n}$ is the broad-definition association statistic adjusted for sample size. This adjustment ensure that comparison of association statistics from meta-analyses with different sample sizes is valid.

### 1.3.12 HLA allele imputation

HLA genes located in the major histocompatibility complex region (MHC) are known to contribute to immune disease susceptibility [35]. Although sequence-based typing (SBT) is the gold standard to identify HLA alleles, its relatively higher cost and the complexity of HLA sequencing has prevented scaling up of SBT methods to large cohorts [36]. HLA allele imputation methods based on SNP arrays are a reliable alternative to SBT methods of HLA typing, and can be performed using a small number of genotyped SNPs in each HLA gene [37–42] (reviewed in [43]).

HLA alleles were imputed for all IBD-BR and UKIBDGC individuals using HIBAG, a computationally efficient prediction algorithm that was pre-trained on a diverse set of haplotypes from different ancestries and is used to impute HLA alleles [40]. Model parameters that were pre-trained on SNPs from the UK Biobank Affymetrix Axiom array from European and multi-ethnic ancestries were used in HLA imputation (downloaded from [44]). After downloading the pre-trained models, HLA imputation was performed at the HLA allele and HLA allele group levels for a total of 7 HLA genes (4-digit and 2-digit resolutions; Table 1.2).

Table 1.2 Number of genotyped variants used to perform HLA imputation.

| Gene | 2 digits | 4 digits |
|---|---|---|
| HLA*A | 723 | 717 |
| HLA*B | 763 | 752 |
| HLA*C | 819 | 770 |
| HLA*DPB1 | 478 | 478 |
| HLA*DQA1 | 753 | 655 |
| HLA*DQB1 | 754 | 702 |
| HLA*DRB1 | 675 | 653 |

For each HLA allele, I performed the association test using the logistic regression model: pcd status $\sim$ HLA allele copies + covariates, with the R function `glm(family=binomial())` and using the same covariates as used in the GWAS. The association analysis were performed separately for each of IBD-BR, UKIBDGC (HCE) and UKIBDGC (GWAS1). The association effect sizes and standard errors were subsequently meta-analysed using the R package `metafor`. Additionally, I performed conditional association analysis to investigate if any HLA alleles can account for genome-wide significant SNPs. Conditional association analyses were performed by including SNP dosages as covariates in the same model.

## 1.4 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 reported perianal disease location, 61% reported a different disease location, and 20.8% answered "*Unknown*". UKIBDGC does not report specific manifestations of pCD.

### 1.4.1 Epidemiological characteristics

Epidemiological characteristics of pCD+ and pCD- patients were largely similar in both cohorts (Table 1.3). Males were more likely than females to report perianal involvement in both cohorts (P-value=$7 \times 10^{-4}$ and $8 \times 10^{-6}$ 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).

Table 1.3 Epidemiological characteristics of pCD+ and pCD- patients in IBD-BR and UKIBDGC

| | 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) |

### 1.4.2 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 [45, 19]. 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).

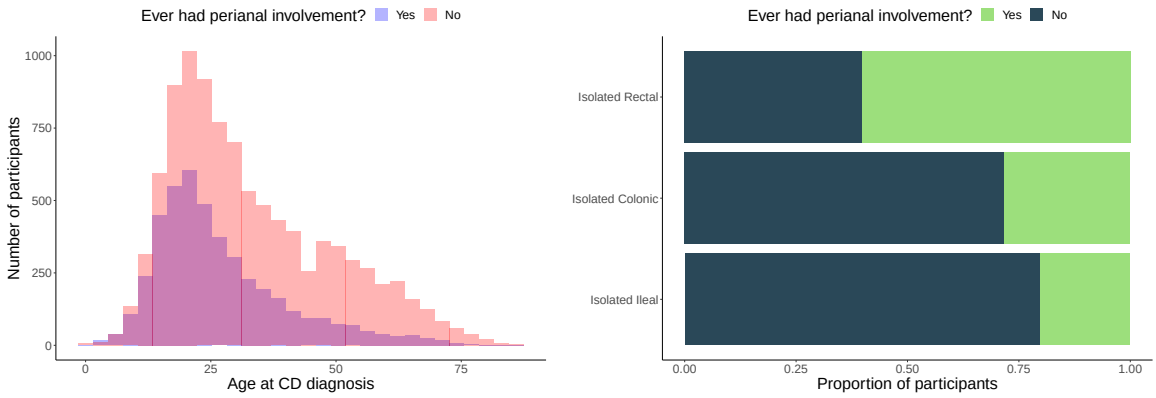



Fig. 1.1 age at diagnosis and macroscopic extent of Crohn's disease in pCD+ and pCD- patients in the IBD-BR.

### 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 1.4; odds ratio=0.83, 0.77, 0.62 and 0.55 respectively). 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) [46–49]. Additionally, the ENTERPRISE clinical trial found that Vedolizumab achieved remarkable fistula closure and healing [50]. On the other hand, oral steroids are known to be ineffective for fistula closure and may even exacerbate perianal abscess [51]. 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 1.4). 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 [52]. 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

| | 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) | $2.1 \times 10^{-9}$ |
| Erythema Nodosum | 199 (4.6) | 222 (2.3) | $7.7 \times 10^{-13}$ |
| Iritis | 183 (4.2) | 242 (2.5) | $1.1 \times 10^{-7}$ |
| Orofacial Granulomatosis | 153 (3.6) | 162 (1.7) | $2.4 \times 10^{-11}$ |
| Psoriasis | 311 (7.2) | 518 (5.4) | $6 \times 10^{-5}$ |
| 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) | $3.2 \times 10^{-9}$ |
| **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) | $4 \times 10^{-3}$ |
| Infliximab | 1207 (50.9) | 1622 (55.7) | $6 \times 10^{-4}$ |
| Adalimumab | 740 (47.8) | 1368 (54.2) | $8 \times 10^{5}$ |
| Vedolimumab | 236 (67.8) | 424 (77.4) | $2 \times 10^{-3}$ |
| 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) | $3 \times 10^{-4}$ |

Table 1.4 Drug intake and extraintestinal manifestations in pCD+ and pCD- patients in the IBD-BR. Percentage of patients are shown between parentheses. Significant differences between pCD+ and pCD- were assesed using a $\chi^2$ test and the P-value is shown in the last column.

suture (744 pCD+ patients; 25%), followed by perianal fistula repair operation (438 pCD+ patients; 14.7%).

**pCD prevalence decreased over time**

Understanding pCD prevalence over time is important to understand how the burden of pCD on patients and healthcare providers has changed. Previous work has shown reduced pCD incidence over the last decade [53], which was partly attributed to improved treatment options. In this regard, IBD-BR offers a unique opportunity to assess this trend. Although the precise time of pCD development is not avaialble, the time between CD diagnosis and the last clinical review can be used to compare how pCD prevalence among CD patients has changed in different eras. Over 93% of IBD-BR patients with pCD information were clinically re-assessed in or later than 2016, which reduces the bias introduced by potentially outdated clinical data.

To investigate this trend 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, confidence intervals around the point estimates were larger in the years between 1980-2010 since fewer IBD-BR patients were diagnosed with CD in those years (100-230 participants per two years). 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 (t-test 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%), between 2010 to 2014 only (mean prevalence=28.1%), or between 2010 to 2012 (mean prevalence=29.4%).

A decrease in pCD prevalence has been observed previously [53]. However, such a decrease has not been precisely quantified, partly due to the relatively smaller sample sizes of most studies [54–57]. 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. An important consideration when investigating the effect of data censorship is that pCD prevalence estimates should be up-to-date. For example, a patient who is diagnosed with CD in 2012 for example may be incorrectly considered pCD-free if their clinical information were not updated afterwards. In IBD-BR, this bias is mitigated by the the fact that the majority of patients were clinically assessed between 2016 and 2021. Additionally,

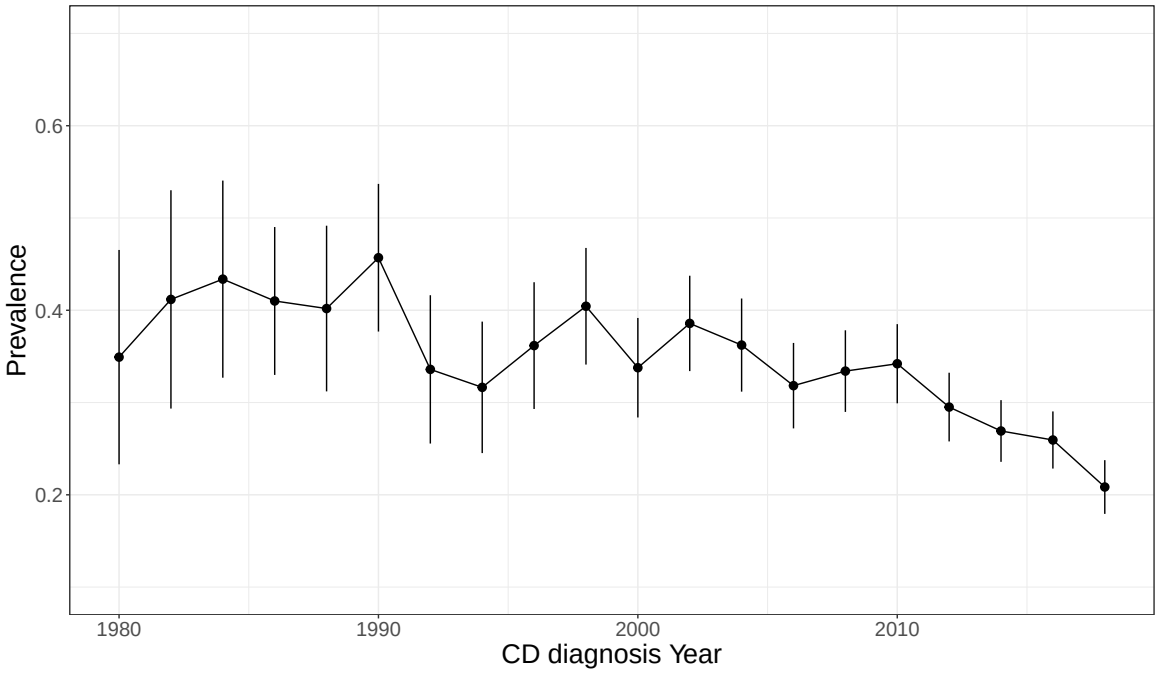



Fig. 1.2 Prevalence per year of CD diagnosis, partitioned into two-year groups. 95% confidence intervals around point estimates are calculated using a bootstrap procedure, whereby participants where resampled 1,000 times within each two-year group.

the consistent decrease in pCD prevalence even when I excluded patients diagnosed after 2016, 2014, or 2012 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 [57]. It is therefore unlikely that the decrease in pCD prevalence is affected by censored observations.

### 1.4.3 UKIBDGC and IBD-BR definitions of pCD are similar

Similar to IBD-BR, the UKIBDGC can be used to define a pCD+/pCD- cohort. The UKIBDGC reports a number of clinical and phenotypic characteristics of IBD patients. For each IBD participant, disease subtype diagnosis, and location (inclduding perianal disease) are recorded. However, 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 patients. Ensuring the consistency of pCD status between UKIBDGC and IBD-BR is crucial to minimise the heterogeneity of phenotype definition between the two cohorts and maximise

the statistical power from a meta-analysis between the two cohorts. To assess this, participants who may have taken part in both the 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 participant exhibited discordant pCD status between IBD-BR and UKIBDGC. A total of 432 participants had missing or "Unknown" perianal involvement, 406 of which were"Unknown" in both cohorts (Table 1.5). This strong agreement in pCD status indicates that both cohorts assign pCD status in a similar fashion.

I then asked if the UKIBDGC cohort was enriched for particular perianal manifestations. Since this information is not available in the UKIBDGC phenotype data, it can be obtained from the IBD-BR clinical data for patients who reported pCD+ status in both studies. I found that these patients 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). Moreover, 27% of the 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. Overall, this shows that the definition of pCD status is likely consistent between the cohorts. From the analysis of overlapping individuals it does not appear that the UKIBDGC pCD status assignment criteria were different from the criteria used in the IBD-BR questionnaire.

Table 1.5 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 |

## 1.5 Genome-wide association analysis of pCD

### 1.5.1 Defining pCD+ cases

Genome-wide studies of disease subphenotypes pose unique challenges compared to traditional case-control GWASes. Unlike GWASes of CD, for example, where robust diagnostic criteria are applied to clearly demarcate cases and controls, in GWAS of disease subphenotype such as pCD it is not obvious which specific manifestations should be considered 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 robustness of genome-wide significant hits. In the previous section, I showed that leveraging individuals who registered for both studies 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?"*. I threfore defined pCD+ cases in both cohorts as CD patients that report any type of perianal involvement,

### 1.5.2 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 risk of developing perianal disease 5 years and 10 years after diagnosis are similar [57]. 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 reported earlier. 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.

With the pCD case-control cohort defined above, 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 $\chi^2$=0.49; $\lambda_{GC}$=1.08).

### 1.5.3 UKIBDGC

As mentioned earlier, UKIBDGC only reports whether or not partipcipants report perianal involvement and does not provide specific perianal manifestations. A total of 8,078 patients of European ancestry were diagnosed with CD, of which 6550 had perianal involvement information. To minimise sample overlap with the IBD-BR, I removed UKIBDGC individuals who showed genetic similarity with individuals from the IBD-BR (see Methods for more details on how genetic similarity was assessed), 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 were genotyped with two different gentopying arrays and were therefore analysed seperately (I will refer to these two cohorts as 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 $\chi^2$=0.47; $\lambda_{GC}$=1.04; GWAS1: median $\chi^2$=0.46; $\lambda_{GC}$=1.01).

### 1.5.4 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. Four variants in the MHC region at the 6p21.32 locus showed genome-wide significant association (index variant rs115378818; P-value=$8.6 \times 10^{-12}$; Table 2.4). None of the variants showed significant heterogeneity of effect size between the constituent cohorts ($P_{het}$ < 0.008). All four variants were well-imputed across the constituent cohorts (INFO score > 0.7).

Table 1.6 Genome-wide significant variants in the 6p21.32 locus. Odds ratio and their 95% confidence intervals are shown. MAF=minor allele frequency.

| Chromosome | Position (b38) | Effect Allele | OR | P-value | MAF |
|---|---|---|---|---|---|
| 6 | 32,205,822 | C | 1.45 (1.27 - 1.66) | $4 \times 10^{-8}$ | 0.05 |
| 6 | 32,243,461 | C | 1.38 (1.23 - 1.55) | $4.4 \times 10^{-8}$ | 0.08 |
| 6 | 32,279,268 | G | 1.57 (1.36 - 1.82) | $1.5 \times 10^{-9}$ | 0.05 |
| 6 | 32,333,650 | T | 1.78 (1.51 - 2.1) | $8.6 \times 10^{-12}$ | 0.04 |

Table 1.7 Case and control minor allele frequencies of the genome-wide significant variants in the 6p21.32 locus in all constituent cohorts.

| | IBD-BR | | UKIBDGC (HCE) | | UKIBDGC (GWAS1) | |
|---|---|---|---|---|---|---|
| SNP | Cases | Controls | Cases | Controls | Cases | Controls |
| 6:32333650_C_T | 0.042 | 0.027 | 0.059 | 0.037 | 0.055 | 0.035 |
| 6:32279268_T_G | 0.054 | 0.038 | 0.067 | 0.046 | 0.062 | 0.044 |
| 6:32205822_T_C | 0.063 | 0.046 | 0.070 | 0.051 | 0.065 | 0.047 |
| 6:32243461_G_C | 0.084 | 0.066 | 0.092 | 0.073 | 0.099 | 0.072 |

All four variants had a low minor allele frequency (MAF), ranging from 0.04 to 0.08 in the constituent cohorts (Table 2.4). Low-frequency variant calling is more prone to genotyping errors than common variants, and therefore low-frequency genome-wide significant hits require additional quality checks. In some studies, these associations were later found to be false positives [58, 59]. To minimise this risk and ensure the robustness of the associated variants, I performed a post-GWAS to investigate whether the association evidence is consistent with the expected LD structure between variants in non-Finnish Europeans. This check ensures that variants in the genome-wide significant locus follow their expected LD. Mismatches between association strength and expected LD reflects a potential false positive association.

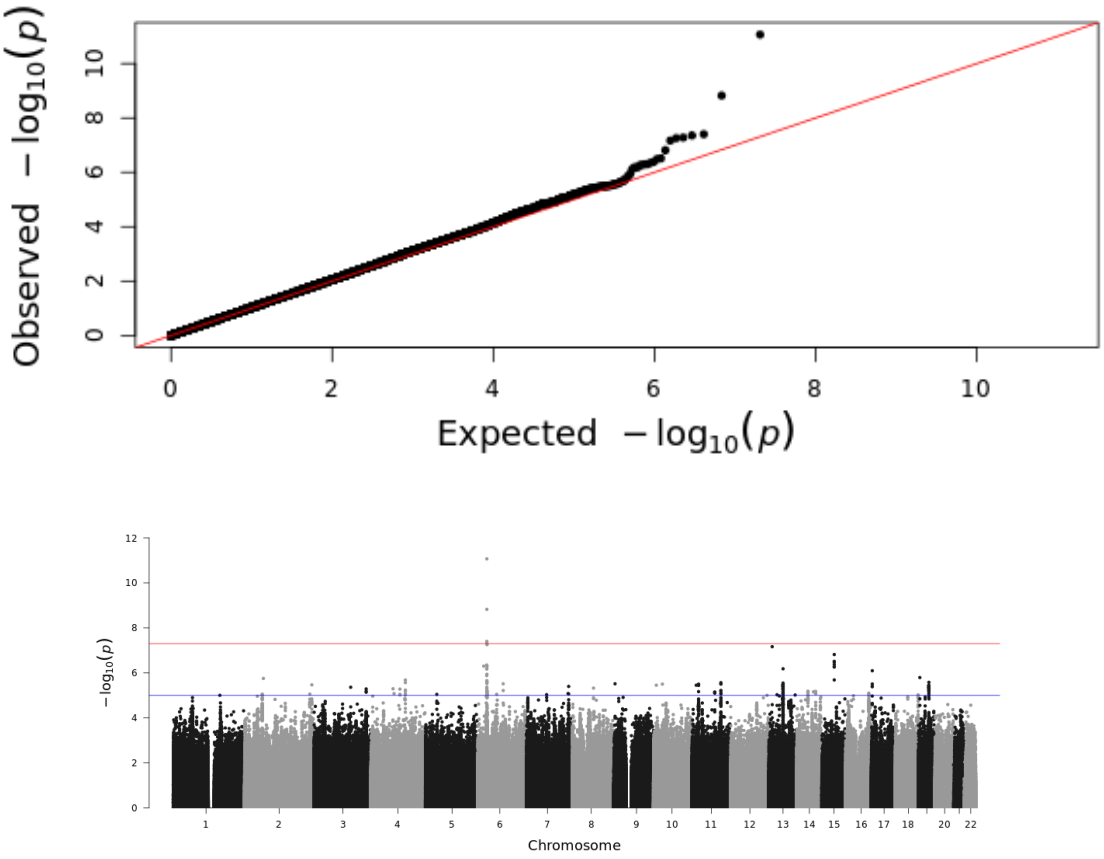



Fig. 1.3 (a) Quantile-quantile plot for the meta-analysis between UKIBDGC and IBD-BR cohorts, suggesting a good fit to the uniform distribution, and showing no evidence of genomic inflation (median $\chi^2$=0.47; $\lambda_{GC}$=1.03; median $\chi^2$ was calculated by converting P-values to $\chi^2$ values 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.

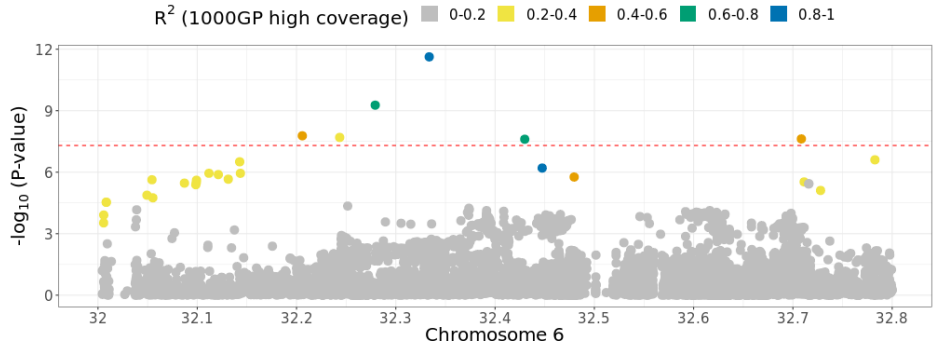



Fig. 1.4 Regional association plot showing meta-analysis association P-values in 6p21.32. Variants are coloured by $R^2$ with the index variant, derived from 1000 Genomes Project High Coverage study (Non-Finnish Europeans)

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

Although the four variants spanned a 130 kbp region, they displayed high LD with the index variant, as the 6p21 region is known to exhibit long-range LD (Figure 1.4). To better understand if the association strength matches the expected LD pattern, I measured the correlation between the association P-value and $R^2$ with the index variant. The underlying assumption is that, for a given variant in a locus, the lower its LD with the index variant, the weaker its association is expected to be. P-values that do not match this expectation may suggest a spurious association due to a genotyping or imputation error or cryptic population structure. At 6p21.32, I found that LD friends P-values were correlated with the expected $R^2$ with the index variant ($\rho$=0.45; LD friends defined as variants with $R^2 > 0.5$ with the index variant and P-value < 0.01; Figure 1.5). The relatively weak correlation is likely driven by the small number of LD friends that the index variant has (N=4). When I relaxed the $R^2$ cutoff of LD friends this correlation became increasingly stronger ($\rho$=0.65 and 0.72 at $R^2 >$ 0.4 and 0.2, respectively), showing that overall the variants at this locus follow their expected association strength given the LD structure between variants.

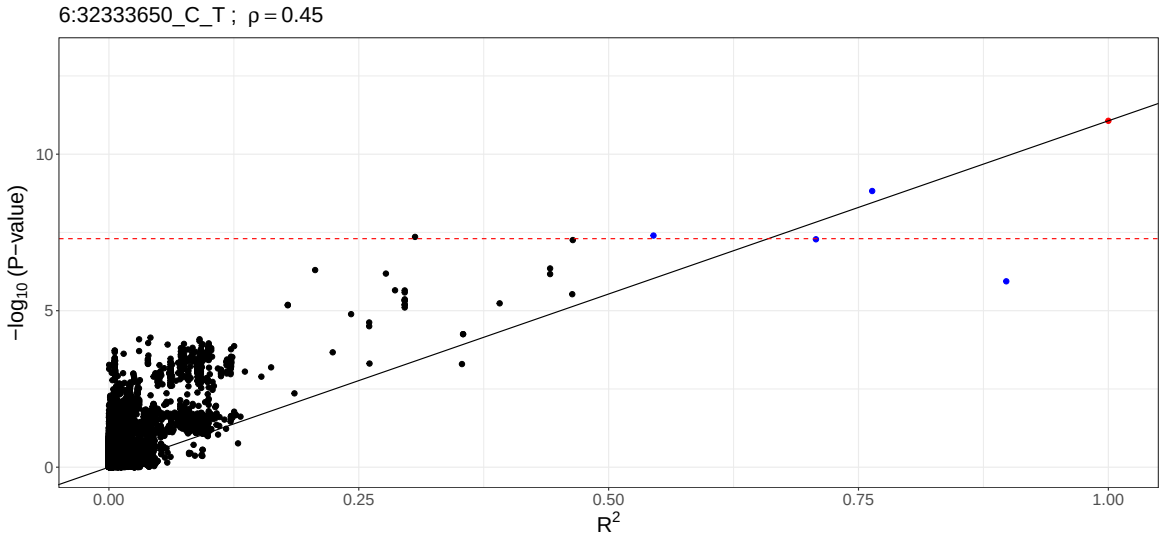



Fig. 1.5 Association P-value for all variants in a 1mbp window around rs115378818 (P-value $< 5 \times 10^{-4}$) on the x-axis, and $R^2$ of variants with the index variant on the y-axis (derived from 1000GP). Horizontal red-line indicates the genome-wide significance threshold (P-value $= 5 \times 10^{-8}$).

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

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 meta-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 ($META_{complexfist}$).

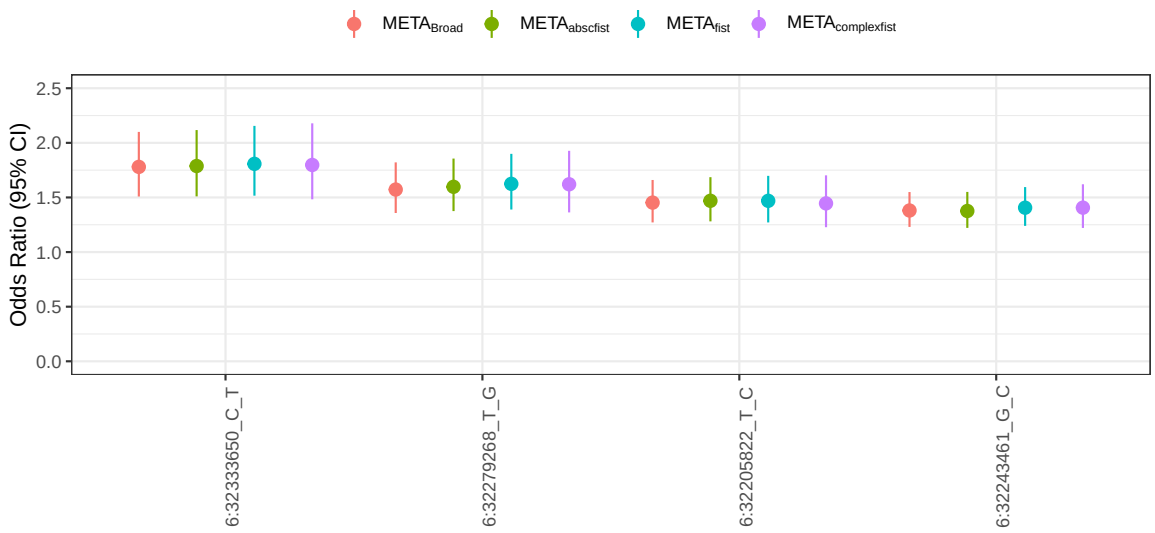



Fig. 1.6 Odds ratios of all four genome-wide significant SNPs in pCD cohorts defined with different case inclusion criteria.

I compared the effect sizes of the four-genome-wide significant SNPs and found that none of them exhibited heterogeneity of effect sizes across the different pCD definition meta-analyses ($P_{het} > 0.01$). Additionally, I compared the association statistic between different definitions. Since the stricter definitions resulted in a reduction in the number of pCD+ cases, a proportional decrease in association test statistic ($\chi^2$) may also be expected. Under the hypothesis that the stricter-definition meta-analyses are simply a random subset of $META_{broad}$, the $\chi^2$ observed in any definition meta-analysis should match $\chi^2$ 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 the index variant, I compared the $\chi^2$ statistics observed in each of the three stricter-definition meta-analyses to $\chi^2_{Broad,N}$. All four 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=$2.2 \times 10^{-11}$; broad-definition P-value adjusted for sample size=$3 \times 10^{-11}$; Figure 1.7). More broadly, across all variants in 6p21.32, I observed strong correlation between observed $\chi^2$ 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).

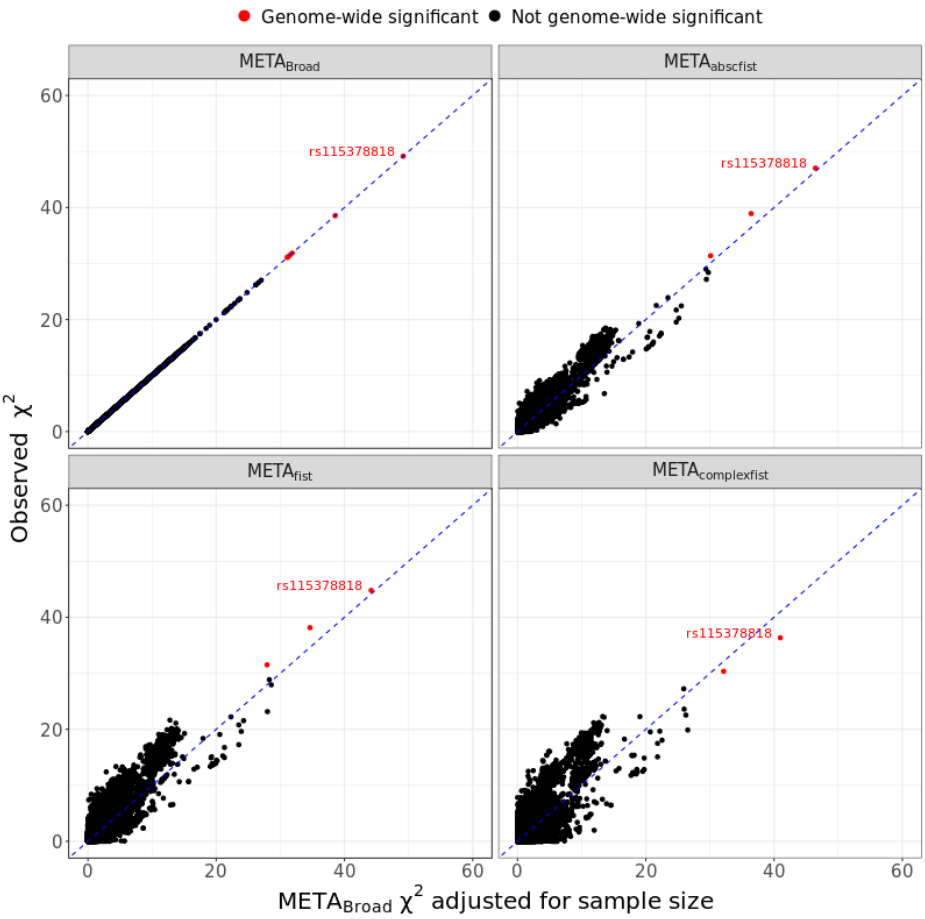



Fig. 1.7 Association P-value for all variants in a 1mbp window around rs115378818 (P-value $< 5 \times 10^{-4}$) on the x-axis, and $R^2$ of variants with the index variant rs115378818 on the y-axis (derived from 1000GP). The blue line indicates the line that passess through the points (0,0) and ($\chi^2_{Broad,n}$,$\chi^2_{observed}$).

### 1.5.6 pCD is associated with HLA allele DRB1*01:03

The MHC region is known to be highly polymorphic and to exhibit long-range LD patterns, which often span several hundred kbps. This makes the mapping of MHC associations to effector genes challenging [60]. To this end, HLA imputation based on genotyped variants can aid the interpretation of genome-wide significant hits in the MHC locus. HLA genes are broadly divided into class I and class II genes [61]. Both classes of genes are responsible for presenting antigens to T lymphocytes and natural killer cells via antigen-presenting cells in order to initiate an innate and adaptive immune response [35]. Due to the extensive polymorphism of the MHC regions, groups of HLA alleles are categorised in HLA groups which are referenced using a 2-digit naming system [62] (2-digit reolution; which I will refer

to as *allele group*). Two additional digits may be used to reference the specific HLA allele [63] (4-digit reolution; which I will refer to as *specific allele*).

To identify which HLA allele is associated with pCD, 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 how HLA imputation was performed). 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 $< 5 \times 10^{-8}$) 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 compared 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 1.8).

Table 1.8 Top HLA allele associations with pCD status. Both allele groups (2-digit resolution; first two rows) and specific alleles (4-digit resolution; third and fourth rows) are shown. Meta-analysed P-values and odds ratios between UKIBDGC and IBD-BR cohorts are shown (with 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

I then asked if the DRB1*01:03 association with pCD accounted for the genome-wide significant locus at 6p21.32. Linking the two associations can explain which HLA allele the genome-wide significant hit may map to. To this end, I repeated the association tests for all variants in the locus, including DR1*01:03 as a covariate. Additionally, I repeated the HLA allele association test conditioning on the index variant in the locus to understand if

the DRB*01:03 association is completely accounted for by the index variant. Similar to the GWAS and the HLA allele association tests, I also analysed the different cohorts separately and then meta-analysed the effect sizes and standard errors.

After conditioning on the index variant rs115378818, I did not observe an association with DRB1*01:03, indicating that the DRB1*01:03 association is completely accounted for by the index variant ($P_{DRB1*01:03|rs115378818}$=0.61). Conversely, DRB1*01:03 did not completely account for the rs115378818 association. When I conditioned the rs115378818 association on DRB1*01:03, the index variant remained nominally associated with pCD ($P_{rs115378818} = 8.6 \times 10^{-12}$ and $P_{rs115378818|DRB1*01:03} = 1.1 \times 10^{-5}$; Figure 1.8). Taken together, this evidence suggests that DRB1*01:03 is only nominally associated with pCD and that it only partly explains the observed genome-wide association signal.

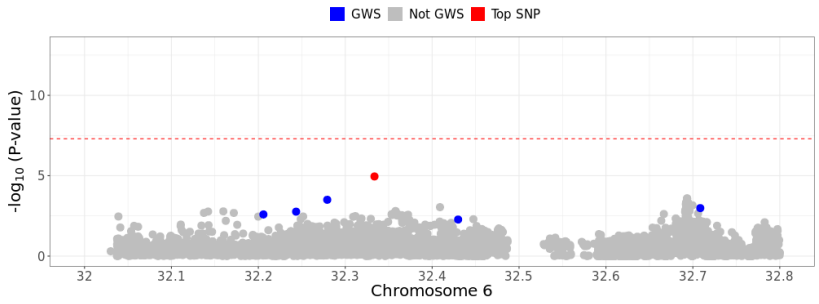



Fig. 1.8 Residual association signal after including DRB1*01:03 as a covariate. Blue points represent variants with a genome-wide significant association in the model that does not include DRB1:01*03. The red point indicates rs115378818 (index variant).

## 1.6 Discussion

In this chapter, I have used two IBD cohorts (UKIBDGC and IBD-BR) with rich clinical data to describe the clinical characteristics of pCD and identify genetic variants associated with pCD risk. With a total of 12,897 individuals, this meta-analysis represents one of the largest pCD GWAS studies to date [22]. Although others have attempted to identify pCD-associated loci, none of the studies have found genome-wide significant hits [22, 64]. The most recent pCD study by Akhlaghpour et al. identified a Complement Factor B (*CFB*) missense variant that was nominally associated with pCD risk (rs4151651; P-value=$9.35 \times 10^{-6}$). *CFB* is part of the alternative pathway responsible for the activation of the complement system, an innate immune subsystem that improves the ability of phagocytic cells to clear pathogens [65, 66].

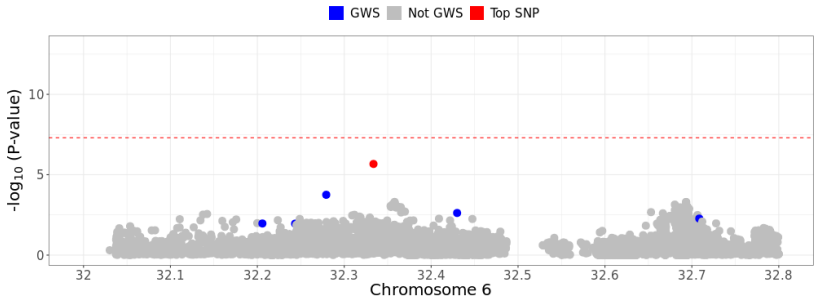



Fig. 1.9 Residual association signal after including rs114969413 (a *CFB* locus variant) as a covariate. Blue points represent variants with a genome-wide significant association in the model that does not include rs114969413. The red point indicates rs115378818 (index variant).

Functional follow-up work showed that the *CFB* missense variant impairs the phagocytic ability of macrophages. This impaired phagocytic capability was hypothesised to impact the ability of the immune system to fight bacterial strains found in the fistulas of CD patients. Although this study established a plausible factor that may contribute to pCD risk, a more complete picture of its pathogenesis is needed.

In an attempt to fill the gap in our understanding of pCD pathogenesis, I performed a meta-analysis of two pCD cohorts, with a total of 3,967 pCD+ cases and 8,930 pCD- controls. I identified a genome-wide significant locus in the highly polymorphic Major Histocompatibility Locus (MHC) at 6p21.32 that was associated with pCD risk (index variant rs115378818). Additionally, my post-GWAS checks have reasonably confirmed the veracity of this association. Furthermore, the assocaition signal revealed by our meta-analysis is likely distinct from the *CFB* association. First, 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), rs115378818 remained nominally associated suggesting that the *CFB* does not completely account for it ($P_{rs115378818|rs114969413}=2.1\times10^{-6}$; Figure 1.9), and showing that our genome-wide significant variant represents a novel pCD-associated locus.

Despite our ability to identify a novel association, case heterogeneity remains an important limitation of this study. Akhlaghpour et al. have also noted that a potentially heterogeneous composition of their cohorts may have decreased their power to detect genome-wide associations with pCD. Their cohorts, similar to ours, may have had patients with milder forms of pCD such as skin tags and ulcer. Therefore, I tested the association of our locus with

pCD+ cases redefined with increasingly severe criteria and found that the genome-wide significant association was robust to different case definitions. However, the heterogeneity of case definition may still have decreased our power to detect *additional* pCD-associated loci. In this study, a more refined definition of pCD+ as fistulising pCD was hampered by the unavailability of more granular data on the specific pCD manifestations in the UKIB-DGC cohort. It remains to be seen if larger sample sizes can overcome the heterogeneity in case definition and identify more pCD-associated loci. Increasing power will not only identify more pCD-associated loci, but it will also give us a more complete understanding of the pathogenesis of pCD, and may implicate biological pathways that are targeted by pCD-associated genetic variation.

Nonetheless, the finding that HLA-DRB1*01:03 may explain the genome-wide association signal is a promising starting point. HLA-DRB1*01:03 has been previously shown to be associated with colonic Crohn's disease, ulcerative colitis risk [67, 68] and rheumatoid arthritis (RA) [69]. However, several avenues should be explored to identify which HLA alleles completely account for the genome-wide significant association. Several reasons may explain the incomplete association of HLA-DRB*1:03 with pCD. First, the association signal at 6p21.32 may be explained by multiple HLA alleles, and therefore conditioning on a single 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 [70]. Second, multiple *HLA-DRB1* alleles with slight differences in their amino acid sequences, especially in the peptide binding regions, have different affinities to antigens being presented to T-cells [71]. Therefore, testing the pCD association with multiple HLA-DRB1 alleles that share the same amino acid sequences might better account for the pCD signal [72]. Interestingly, Raychaudhuri et al. found that only three amino acid positions in a predictive model of RA risk provided identical prediction to a model that included all *HLA-DRB1* allleles [73]. Therefore, an important follow-up to my work should explore the association of HLA-DRB1 alleles with pCD at the amino acid level.

Finally, it is important that future studies do not consider pCD as an isolated disease. The broad clinical phenotyping of IBD-BR showed lower drug intake, higher prevalence of extraintestinal manifestations, and higher prevalence of rectal CD in pCD+ patients. These observations naturally pose a question about how these disease characteristics relate to pCD. A plausible explanation is that these clinical characteristics are simply manifestations of what is collectively termed "severe CD". It is still important however to understand how severe CD

drives these seemingly unrelated manifestations. For example, is there common pathogenesis, or genetic predisposition that drives all these manifestations, including pCD? In this regard, my univariate pCD meta-analysis is limited. Future studies should quantify the genetic correlation between pCD and disease severity and location, extraintestinal manifestations, and drug response. Furthermore, multivariate GWAS methods should be employed to identify shared and distinct genetic factors driving each of these manifesations. The observation that clusters of disorders have common as well as distinct genetic factors has been previously shown to further our understanding of multiple groups of disorders such as psychiatric disorders [74].

Literature about pCD pathogenesis largely attributes the development of fistlising CD to the theory of epithelial-to-mesenchymal transformation of epithelial cells. With the ongoing growth of IBD-BR, and with more genotyped CD participants, a more nuanced understanding of which dysregulated pathways give rise to pCD risk, and whether or not these pathways corroborate the EMT theory of pCD pathogenesis is increasingly attainable.

# Chapter 2

# Genome-wide Meta-analysis of All-cause Perianal Disease

## 2.1 Contributions

UK Biobank phenotype and genotype data were obtained by Dr. Laura Fachal. Genotype quality control and principal component analysis were also performed by Dr. Laura Fachal. All the other analyses described here were performed by me.

## 2.2 Introduction

In the previous chapter, I described the characteristics and genetic underpinning of perianal Crohn's disease. However, perianal disease (pAD) is not only associated with Crohn's disease. pAD encompasses a broader set of perianal manifestations such as perianal abscess, fissures, and fistulas.

Anal fissures are tears and/or ulcers in the perianal skin that cause sharp pain associated with defecation and rectal bleeding. They are classified as acute, lasting less than 6 weeks, or chronic. The etiology of anal fissures is still debated [75], but they are thought to be caused by trauma resulting from high anal canal pressure associated with constipation or diarrhea [76]. Although recent population studies on the etiology of anal fissures are markedly lacking [76], a number of studies conducted between 1977 and 1983 found that chronic constipation accompanied fissures in only 25% of patients while diarrhea accounted for less than 7% [77–79]. Anal fissures are conservatively treated with dietary and lifestyle modifications, analgesics, and fiber supplements [80]. Anal fistula is a more serious perianal manifestation

of pAD. An anal fistula is defined as an abormal communication between the anal canal and the perianal skin [81]. It is characterised by pain, rectal discharge and bleeding, and causes significant lifestyle difficulties for patients [82]. The incidence of perianal fistulas varies per country, and a recent study in four European countries found that it ranges from 1.2 to 2.8 per 10,000 per year [83], but its incidence is likely underestimated [84]. Moreover, little is known about the pathophyiology of anal fistulas. Cryptograndualr fistulisation, a theory proposed by Parks in 1961 [85, 86], is the most accepted pathophysiological account for the origin of anal fistulas. Accoring to the cryptograndular theory, anal fistuls start as an inflammatory process in the proctodeal glands whose ducts extend to connect the perianal skin to the anal canal. Crohn's disease, tuberculosis [87], ratiotherapy [88], sexually transmitted diseases [89] and malignancy can contribute to these initial inflammatory process that resultin anal fistula formation. However, over 90% of cases remain idiopathic [90]. Management of anal fistula depends on its type, extension, and underlying cause, but the end goal of all treatments is the complete drainage of abscess, and fistula healing while preserving anal sphincter function and continence. The management plan is typically decided based on a combined clinical, radiological and/or endoscopic assessment. Surgical options of fistula management include fistulectomy, the complete excision of the fistula tract, and fistuolotomy, in which the fistula is laid open and allowed to heal. Surgical procedures that preserve anal function such as fistolotomy with sphincter reconstruction have emerged in the last 30 years as a particularly effective method of treating anal fistulas, with lower incontinence and recurrence rates compared to fistulectomy [91, 92].

Despite recent advances in the management of pAD, little is known about the biological and pathophysiological processes that lead to anal fissure and fistula formation. Despite the overall consensus on the cryptoglandular theory, it is unclear if some individuals are at higher risk of developing pAD due to their genetic predisposition. Genome-wide association studies (GWAS) have significantly improved our understanding of the genetic factors underlying several complex diseases as well as the biological pathways implicated by disease-associated genetic variants. In the last 15 years, a typical GWAS required extensive coordination between researchers, clinicians and recruitment centres to construct case-control cohorts to study a particular disease or trait of interest. In recent years, efforts to build various national biobanks where genetic, and phenotypic data are available for hundreds of thousands of participants has significantly improved our ability to conduct GWAS of previously understudied diseases and traits. Their large sample sizes have made it possible to carry out GWAS of thousands of binary and continuous traits and diseases, including relatively uncommon

diseases.

Recruitment for the UK Biobank (UKBB) started in 2006, and has so far collected genotypic, biomarker and clinical data from electronic health records as well as blood and urine samples from over 500,000 participants [93]. The UKBB queries electronic health records for various types of data including deaths, cancer registrations, and hospital inpatient episodes. Electronic health records data are provided by external sources. Upon receipt of the data, a multi-step approach is followed, where received data is subjected to further pre-processing and quality checks to ensure its alignment with the UKBB data dictionary, and that it does not contain any ambiguities. The data is then consolidated into a central UKBB database and made available to researchers [94].

FinnGen is another example of national biobanks that has made GWAS for various traits and diseases more feasible. FinnGen started in 2017 as a public-private partnership between several public Finnish institutions and thirteen international pharmaceutical companies [95]. Importantly, FinnGen restricts access to individual-lvel data to approved researchers only. However, summary statistics data from GWAS analyses are made publicly available. In its latest release (data freeze 9), GWAS are available for over 2,200 binary endpoints from 377,277 individuals [96].

The UKBB and FinnGen use slightly different clinical coding systems to register clinical data. The UKBB uses the International Classification of Disease (ICD), a hierarchical clinical framework that organises clinical diagnoses in a tree-like structure. Different groups of diseases are organised in alphabetical chapters and particular diseases within each chapter are given a numeric value (e.g. chapter K contains digestive system disorders and K60 indicates anal fissure and fistula). Further detailed diagnosis subtypes are nested within each alphanumeric code up to four levels of resolution [97]. However, the UKBB only records up to two levels of resolution (e.g. K50.0 indicates Crohn's disease of the small intestine). FinnGen, on the other hand uses an expert-curated set of endpoints that are largely parallel to ICD-10 codes. These endpoints are designed to accommodate inclusion and exclusion criteria relevant for GWAS analyses [98]. Despite these differences, both resources are valuable for studying pAD.

In this chapter, I will describe a pAD GWAS I performed in the UKBB to map pAD-associated genetic variants and I will describe the post-GWAS quality checks I carried out to ensure the validity of genome-wide significant loci. Additionally, I will outline how I used

summary statistics from the FinnGen GWAS both as a replication dataset and as a constituent cohort in a pAD meta-analysis I performed between the UKBB and FinnGen cohorts. Finally, I will describe two follow-up analyses that I performed to better understand the effects of pAD-associated variants on haemorrhoids, a closely-related disease, and identify effector genes at these loci.

## 2.3 Methods

### 2.3.1 UKBB sample preparation and data access

Hospital inpatient episode data and genotyped and imputed data were obtained under UK Biobank Application 45669. UKBB participants were genotyped using either UK Biobank Axiom Array [99] or the Affymetrix UK BiLEVE Axiom array [100]. Sample processing, genotyping and quality control were performed at the UK Biobank, Affymetrix and the Wellcome Trust Centre for Human Genetics [101]. The imputation process has been previously described [102] and consists of imputing directly genotype data to the Haplotype Reference Consortium (HRC) and UK10K, resulting in 96 million variants. Imputed data was obtained as BGEN v1.2 files [103].

### 2.3.2 Defining pAD case control cohorts

#### Case inclusion criteria

To define the case cohort, I identified all individuals with ICD-10 code K60 or ICD-9 code 565. In total, 5,257 UKBB participants had at least a single visit where they received either a primary or secondary pAD diagnosis or its corresponding ICD-9 code ("anal fissure and fistula"; 565). Six level-2 codes are nested within K60, representing two broad categories of pAD: fissures and fistulas. Three codes are used for acute and chronic fissures and three codes for acute and chronic fistulas. 92% of patients (4,858) presented with either K60.1, K60.2 or K60.3 ("chronic anal fissure", "anal fissure, unspecified" and "anal fistula", respectively; Table 2.1).

#### Control exclusion criteria

To avoid contamination of controls with lower digestive tract disorders that may be true pAD cases that were incorrectly diagnosed, I applied a set of control exclusion criteria. Specifically, I excluded from the control set any individuals who had an ICD-10 hospital diagnosis of K55-K64 or their corresponding ICD-9 codes as outlined in Table 2.2 (collectively grouped

as "Other diseases of intestines" in ICD). These ICD codes cover disorders with symptoms that may resemble pAD symptoms upon presentation, and include ano-rectal bleeding (K55 vascular disorders of the intestine, K57 diverticular disease of intestine and K64 Haemorrhoids and perianal venous thrombosis), or a change in bowel habits (K56 Paralytic ileus and K58 Irritable bowel syndrome), perianal fistula or abscess (K60 fissure and fistula of the anal region and K61 abscess of the anal and rectal region), any ano-rectal abnormalities (K62), or proximal fistulas or abscesses (K63). In total, I excluded 128,319 individuals from the control cohort (26.7%), resulting in 353,437 controls (per-code number of individuals in Table 2.1).

Table 2.1 Number of UKB participant with with a primary or secondary diagnosis for each K60 level 2 code. K60.0=Acute Anal Fissure; K60.1=Chronic Anal Fissure; K60.2=Anal Fissure; unspecified; K60.3=Anal Fistula; K60.4=Rectal Fistula; K60.5= Anorectal Fistula

| ICD-10 code | K60.0 | K60.1 | K60.2 | K60.3 | K60.4 | K60.5 |
|---|---|---|---|---|---|---|
| Number of individuals | 144 | 788 | 2,624 | 1,954 | 76 | 122 |

Table 2.2 pAD control set exclusion criteria. All ICD-10 codes had corresponding ICD-9 codes except K56 K62 and K63. For those, ICD-9 codes were obtained manually by inspecting level-2 ICD-10 codes and searching for their corresponding level-2 ICD-9 codes.

| ICD-10 code | ICD-10 meaning | ICD-9 code | ICD-9 meaning | N |
|---|---|---|---|---|
| K55 | Vascular disorders of intestine | 557 | Vascular insufficiency of intestine | 2923 |
| K56 | Paralytic ileus and intestinal obstruction without hernia | 5600, 5601, 5602, 5603, 5608A, 5608, 5609 | Intussusception, Paralytic ileus, Volvulus, Impaction of intestine,Other specified intestinal obstruction, Unspecified intestinal obstruction | 9257 |
| K57 | Diverticular disease of intestine | 562 | Diverticula of intestine | 61519 |
| K58 | Irritable bowel syndrome | 5641 | Irritable bowel syndrome | 12418 |
| K59 | Other functional intestinal disorders | 564 | Functional digestive disorders not elsewhere classified | 30087 |
| K60 | Fissure and fistula of anal and rectal regions | 565 | Anal fissure and fistula | 5079 |
| K61 | Abscess of anal and rectal regions | 566 | Abscess of anal and rectal regions | 2178 |
| K62 | Other diseases of anus and rectum | 5690, 5691, 5692, 5693, 5694 | Anal and rectal polyp, Rectal prolapse, Stenosis of rectum and anus, Hemorrhage of rectum and anus, Other specified disorders of rectum and anus | 39191 |
| K63 | Other diseases of intestine | 5695, 5696, 5697, 5698, 5699 | Abscess of intestine, Colostomy and enterostomy complications, Complications of intestinal pouch, Other specified disorders of intestine, | 33307 |

### 2.3.3 ICD code enrichment in pAD cases versus controls

The availability of a large number of clinical diagnoses and phenotypes for UKB participants enables a thorough characterisation of the pAD case cohort. I aimed to understand the cohort composition by identifying which ICD-10 codes are enriched in cases versus controls. For each ICD-10 code, I compared the prevalence in pAD cases versus controls, and I formally tested the enrichment of 1,693 codes using Fisher's exact test. For this test, I did not apply the control exclusion criteria outlined in Table 2.2.

### 2.3.4 UKBB genotype quality control

Genotyping array data from the UK Biobank dataset underwent quality control as part of the International IBD Genetics Consortium GWAS that is being undertaken in the laboratory, which resulted in 419,871 variants being retained. QC was performed using a combination of PLINK (v1.9 and v2) [104], bcftools (v1.16) [105], and KING (v2.2.4) [32]. 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 $\leq 0.01$).
- Significant difference in genotype call rate (P-value $< 10^{-4}$) between IBD cases and controls.
- Large allele frequency (AF) differences between UKBB 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 the UKBB MAF.

Genotypic principal components (PC) were estimated for all participants, using a set of genotyped variants that were also available in the 1000 Genomes Project (1000GP; excluding variants associated with IBD susceptibility (P-value < $10^{-4}$), and variants in long LD regions (as defined in [31]). This final list of variants 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.

### 2.3.5 UKBB GWAS using REGENIE

All genome-wide association analyses were performed using REGENIE v3.2.5 [33] following a 2-step approach. This approach achieves higher computational efficiency compared

to linear mixed model, which are normally used in GWAS methods to account for cryptic relatedness. Briefly, in step 1, a whole-genome regression model is fitted using a subset of high-quality genome-wide variants in order to estimate a set of genome-wide predictors that capture a large fraction of phenotypic variance. These predictors are then used in step 2 in a single-variant association testing model, where a larger set of variants of interest are tested for association. I used post-QC genotyped variants in step 1 as recommended by REGENIE documentation (N=419,871), and both genotyped and imputed variants in step 2, testing all autosomal chromosomes (N=9,705,089). Additionally, I enabled a Firth correction of effect sizes for all variants with P-value < 0.01 in order to account for the case-control imbalance in the pAD cohort (`--firth --approx --pThresh 0.01`). The Firth test corrects biased effect sizes and P-values obtained from highly unbalanced case-control designs, where such an imbalance causes unreliable P-values, inflating Type I error. Approximate Firth logistic regression is a variant of the Firth test that is more computationally tractable and is implemented in REGENIE.

With the pAD case control cohort defined above, I used REGENIE to perform a pAD GWAS with White British UKBB participants [33]. In addition to genotype QC described earlier, I excluded individuals with missing genotypes or with discordant reported and genetically-inferred sex. After this filtering a total of 4,606 pAD cases and 332,234 pAD controls remained (see Methods for genotype data quality control and imputation). In order to account for cryptic population stratification, I used 10 European-ancestry genotypic principal components, as well as sex and genotyping array as covariates in the REGENIE model. After filtering out variants with low imputation quality (INFO < 0.4) and minor allele frequency (MAF) < 0.01, a total of 9,705,089 variants were tested. After running REGENIE, I found that the summary statistics exhibited moderate genomic inflation (median $\chi^2 = 0.48$; $\lambda_{GC} = 1.06$).

### 2.3.6 LD calculation from 1000GP

Reference LD panels obtained from the 1000 Genomes Project High Coverage project [34] were used for different analysis, including genome-wide loci identification using LD clumping, and post-GWAS checks to study the relationship between LD and association strength at genome-wide significant loci. $R^2$ values were calculated between the index variant and all variants in each locus (loci were defined using LD clumping as described in the next section). I downloaded VCFs from the 1000GP high coverage and used PLINK v1.9 to compute LD between all variants and the index variant at each locus. For each GWAS check, I used unrelated individuals assigned to the relevant reference population: NFE and GBR for

UKBB and FE for Finngen (N=426, 99 and 90 respectively). Only one sample was retained from the trios found in 1000GP. Relevant samples were included in the LD calculation using the following PLINK command:

```
plink --r2 --keep EUR.samples --ld-window-r2 0
```

### 2.3.7 Defining genome-wide significant loci in UKBB

I defined genome-wide significant loci from the UKBB pAD GWAS summary statistics using PLINK v1.9 via a clumping procedure. Briefly, LD clumping identifies the most significant variant in a user-defined window to represent each locus (termed index variant). It then proceeds to define the locus boundaries by clumping neighbouring correlated variants. Specifically, any variants within the predefined window that are correlated with the index variant are considered to belong to the same locus represented by the index variant (i.e. variants in high LD). I used VCFs downloaded from the 1000GP, which are used to compute LD, and set a maximum P-value of $5 \times 10^{-8}$ for defining a genome-wide significant locus, with default values for the rest of the parameters: variants with $R^2 < 0.5$, variants outside a window of 250 kbp, or variants that have a P-value > 0.01 are not clumped with the index variant.

```
plink --clump-p1 0.00000005 --clump-r2 0.50 --clump-kb 250
--clump-p2 0.01
```

PLINK outputs each locus' index variant along with any variants that meet the clumping criteria outlined above. I then defined each locus' boundaries by sorting the clumped variants within each locus according to their genomic location: the most downstream variant defined the 5' boundary and the most upstream variant defined the 3' boundary.

### 2.3.8 Finngen summary statistics preprocessing

Publicly available FinnGen GWAS summary statistics (data freeze 7) were downloaded from the FinnGen results website `finngen.fi/en/access_results`. Similar to UKBB, variants with MAF < 0.01 were removed, but imputation quality information were not available, so I was not able to filter out variants with low imputation quality. The association summary statistics showed evidence of moderate genomic inflation ($\lambda_{GC}$=1.089). I used LD clumping with a 1000GP-derived FE LD reference panel to identify genome-wide significant loci.

### 2.3.9 Meta-analysis of UKBB and Finngen

I used METAL to perform the meta-analysis between UKBB and FinnGen GWAS summary statistics. METAL can perform fixed-effects meta-analysis using one of two different well-established schemes: P-values and effective sample size, or effect sizes and standard errors. The P-value scheme is implemented to enable meta-analysis of GWAS summary statistics that do not report the effect allele, while the effect sizes scheme can be used when each variant's effect size and effect allele are reported. Both my UKBB analysis and FinnGen's summary statistics report the effect allele, so I used the effect size scheme of METAL (`SCHEME STDERR`).

After filtering out variants with MAF < 0.01 and with low imputation score (INFO < 0.4), the two GWAS summary statistics had an intersection of 7,663,827 variants and a total of 11,096,129 variants across the two cohorts. Of these, 2,041,145 variants were specific to UKBB and 1,390,527 were specific to FinnGen. Given that 31% of variants were unique to one of the two GWAS, 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 both 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 ($\lambda_{GC}$=1.02).

For each variant, METAL outputs the effect allele, meta-analysed effect size, standard error, and P-value. After performing meta-analyses, it is important to compare the effect sizes between the meta-analysed cohorts. Comparison of both the direction and magnitude of effect sizes gives an indication on how similar the estimated effects of meta-analysed genetic variants are, which is an import post-GWAS quality control check. To formally test this, METAL uses Cochran's Q test to test for effect size heterogeneity. Cochran's Q test asseses two or more effect size estimates and their corresponding standard errors and reports a $\chi^2$ statistic that quantifies the deviation from the null hypothesis that the meta-analysed effect sizes are similar. Depending on the number of meta-analysed studies (in this case 2), a P-value is derived from a theoretical $\chi^2$ distribution with $N-1$ degrees of freedom, where N is the number of meta-analysed studies (heterogeneity of effect P-value $P_{het}$). I used $P_{het}$ to test if index variants at genome-wide significant loci demonstrate heterogeity of effect size between the two cohorts. To account for multiple index variants being tested, I set a Bonferroni-corrected P-value threshold for rejecting the null hypothesis that the effect sizes

are similar between the two studies (P-value < $\frac{0.05}{k}$, where $k$ is the number index variants tested).

### 2.3.10 Defining genome-wide significant loci in the UKBB/FinnGen meta-analysis

Meta-analysis associations are derived from two different populations with different LD structures (NFE and FE). For these associations, a representative LD reference panel that captures the true underlying LD pattern in the meta-analysis is not easy to obtain because LD will be dependent on the contribution of each population to the association signal. However, an LD reference panel is needed both to define genome-wide significant loci based on LD clumping, but also to perform post-GWAS checks. To this end, I identified genome-wide significant loci similar to UKBB and FinnGen, except that I performed LD clumping of the meta-analysis genome-wide significant loci once with an LD reference panel derived from NFE in 1000GP, and once with an LD reference panel derived from FE in 1000GP. Both LD clumping procedures identified the same 18 genome-wide significant loci, but their boundaries differed slighly. I defined consensus loci boundaries as the union of loci boundaries defined by NFE- and FE-based LD clumping.

### 2.3.11 Quality control of meta-analysis genome-wide significant loci

For all 18 genome-wide significant loci defined via the initial LD clumping procedure, I investigated the LD friends of each index variant. Specifically, for each index variant, I checked if the P-values of its LD friends decay linearly with their LD values with the index variant. In a meta-analysis between different subpopulations, it is important to ensure that P-values match expectation under LD in all the constituent cohorts. I therefore computed the correlation between P-values and LD in both NFE and FE for all 18 loci. Six loci showed that the index variant had no LD friends in NFE or FE, or that there was weak or non-existent correlation beween P-values and LD of the index variant's LD friends in NFE or FE (Pearson correlation coefficient between $-log_{10}(P)$ adn ($R^2$ $\rho < 0.2$). I removed these six loci from all downstream analysis.

### 2.3.12 Genetic correlation analysis

Genetic correlation analysis is commonly employed to understand the genetic similarities between two phenotypes or diseases of interest. I used genetic correlation to understand the

overall genetic similarity between pAD and haemorrhoids. Linkage disequillibrium score regression is a common method to compute genome-wide genetic correlation,(LDSC [106]) as it leverages association summary statistics between a pair of traits to compute a genetic correlation estimate ($r_g$). The availability of GWAS summary statistics for large numbers of traits and diseases makes genetic correlation a feasible exploratory analysis to discover genetic similarities between traits.

The fundamental concept of LDSC is that there is a linear relationship between the Z-score product ($z_1 z_2$) and the LD score of SNPs, where LD score is defined as the sum of $R^2$ values for all SNP in a pre-defined windown (the default is 1 cM) [107]. The rationale behind this relationship is that SNPs with a higher LD score are more likely to tag the causal variant at each locus, and will therefore have a larger $z_1 z_2$ value. Regressing the LD score for all SNPs can give an estimate of the overall genetic covariance between two traits. Specifically, the regression slope quantifies the gentic covariance, which can then be normalised by the sample sizes of the two traits and the number of SNPs to obtain a genetic correlation estimate between the two traits ($r_g$).

The accuracy of $r_g$ rests on the assumption that the GWAS population matches the population from which LD scores are derived. By default, LD scores are provided by LDSC, which are computed from the HapMap3 European-ancestry reference panel [108]. It is also important to note that although $r_g$ is a genome-wide measure, its computation is based on a predefined set of SNPs. The estimation of $r_g$ is based on a set of high-quality common SNPs in HapMap3, which is also provided by LDSC (MAF $\geq$ 0.05; N=1,217,312).

**Genetic correlation between pAD and haemorrhoids**

I used LDSC to compute $r_g$ between my pAD meta-analysis and two haemorrhoids GWASes (Zheng et al. 2021 [109] and the Pan-UKBB GWAS: `pan.ukbb.broadinstitute.org/`). I downloaded the Zheng et al. 2021 summary statistics via the GWAS catalogue website (study accession: GCST90014033; $N_{cases}$=218,920; $N_{controls}$=725,213). Since the Pan-UKBB performed the haemorrhoids GWAS using different ancestries, I used the summary statistics from the European-ancestry GWAS only (ICD-10 code: I84; $N_{cases}$=26,348; $N_{controls}$=394,183).

After downloading the two haemorrhoids summary statistics, I preprocessed both of them using the LDSC script `munge_sumstats.py`. The script filters the SNPs and aligns their alleles to the HapMap3 SNP list using the flag `--merge_alleles hm3.snplist`. This script also takes as input a signed summary statistic column which I provided using the flag

`--signed-sumstats effect_size,0`, where the first argument specifies the column name (effect size column) and the second argument specifies the expected value of the signed summary statistic. Each of the two "munged" summary statistics files were then provided as input to the `ldsc.py --rg`, along with the pAD summary statistics file, and $r_g$ is then computed between pAD and each of the two haemorrhoids GWASes.

### 2.3.13 Colocalisation analysis

In order to link the meta-analysis genome-wide significant loci to effector genes, I performed statistical colocalisation with a set of expression and splicing QTLs from the Genotype Expression Project (GTEx v8). Colocalisation analysis is a statistical approach that uses summary statistics from two association studies in order to make an inference whether the two association signal are likely to be driven by a shared causal variant. In this regard, five different hypothesis regarding the relationship between the two association signals are tested:

- $H_0$: none of the two signals are associated with their corresponding traits
- $H_1$: only the first signal is associated with its corresponding trait
- $H_2$: only the second signal is associated with it corresponding trait
- $H_3$: the two signals are associated with their corresponding traits, with different underlying genetic variants
- $H_4$: the two signals are associated with their corresponding traits, and share a single underlying genetic variant.

Certainty about each of these hypotheses is quantified as a posterior probability. Therefore, colocalisation analysis outputs four different posterior probabilities: $PP_0$, $PP_1$, $PP_2$, $PP_3$, and $PP_4$. Statistical colocalisation is implemented in the R package `coloc v5.1.2`.

To maximise the ability of `coloc` to identify effector genes, I downloaded summary statistics from GTEx v8, a large compendium of expression and splicing quantitative trait loci (eQTLs and sQTLs) mapped from RNA-seq samples obtained from 49 human tissues, ranging in sample sizes from 73-706 individuals. eQTLs and sQTLs were mapped in a 1mbp window centred around the transcript start site (TSS) of each gene (cis-eQTLs and cis-sQTLs) using genotyped or imputed variants with MAF $\geq 0.01$.

Within each of the meta-analysis genome-wide significant loci, I identified a list of genes and splice junctions for eQTL and sQTL colocalisation, respectively. To achieve this, for each locus I defined a 1 mbp window around the index variant at each locus, and created

a set of genes and splice junctions whose respective TSS are located within this window. Next, I performed colocalisation analysis between the meta-analysis summary statistics and each eQTLs and sQTLs summary statistics for each gene in the window. I used the `coloc.abf()`, which takes as input effect sizes and standard errors of each variant from the meta-analysis and gene or splice junction being tested. Importantly, `coloc.abf()` does not require the effect sizes to be aligned to the same effect allele, as the Bayes Factor calculation implemented in `coloc` relies on the $Z^2$ statistic to compute the posterior probabilities. Finally, I used the default priors implemented in `coloc.abf()`: prior probability a SNP is associated with pAD=$10^{-4}$, prior probability a SNP is associated with the eQTL/sQTL=$10^{-4}$.

Table 2.3 Number of eQTL and sQTL genes tested for colocalisation across all GTEx v8 tissues. All genes and splice junctions in a 1mbp around each index variant were tested.

| Index variant | Number of tested eQTL genes | Number of tested sQTL genes |
|---|---:|---:|
| 3:53034026_C_T | 41 | 18 |
| 5:64868326_TTTC_T | 13 | 4 |
| 6:1775202_G_A | 12 | 5 |
| 6:31121854_C_T | 89 | 45 |
| 6:31253340_T_C | 98 | 50 |
| 6:133260944_G_A | 18 | 7 |
| 7:2524404_G_A | 21 | 13 |
| 8:70735125_A_G | 15 | 6 |
| 8:70993166_AAGTT_A | 10 | 6 |
| 9:22124505_A_T | 19 | 5 |
| 11:10356352_C_A | 20 | 11 |
| 12:114235969_T_C | 18 | 10 |

## 2.4 Results

### 2.4.1 pAD cases are enriched in multiple disorders compared to pAD controls

In order to understand the composition of the pAD case cohort, I tested the enrichment of 1,693 ICD-10 codes in pAD cases versus controls. Overall, the tested enrichment odds ratio were inflated (median odds ratio=1.36), likely as a consequence of sampling a disease cohort within a healthy population cohort. In total, 198 codes were significantly enriched among cases versus controls (Fisher's exact P-value $< 3 \times 10^{-5}$; top 20 enriched ICD-10 codes are shown in Figure 2.1). Within digestive systems disorders, ICD-10 code K61 (abscess of anal and rectal regions), followed by K50 (Crohn's disease) were the most significantly enriched (log odds ratio=4.04 and 1.96 respectively). This is expected as many perianal fistulas start as abscess of the perianal region and later extend to form a fistula connecting the anorectal canal to the perianal skin [110], and perianal fissures and fistulas are known subphenotypes of CD [111]. Other examples indlude K57 (Diverticular disease of the intestine; log odds ratio=0.74; P-value=$1.2 \times 10^{-87}$), which is also a known cause of pAD [112]. Among non-digestive codes, haemorrhoids (I84) was the most significantly enriched diagnosis (P-value=$6 \times 10^{-98}$; 38% in pAD cases versus 6% in controls; log odds ratio=2.3). This enrichment could indicate a shared pathogenesis between haemorrhoids and pAD. However, it could also be due to a higher likelihood of diagnosing haemorrhoids in patients with more serious ano-rectal disorders such as pAD, compared to the general population where haemorrhoids patients with no other ano-rectal manifestations are less likely to seek medical advice, and may therefore remain undiagnosed.

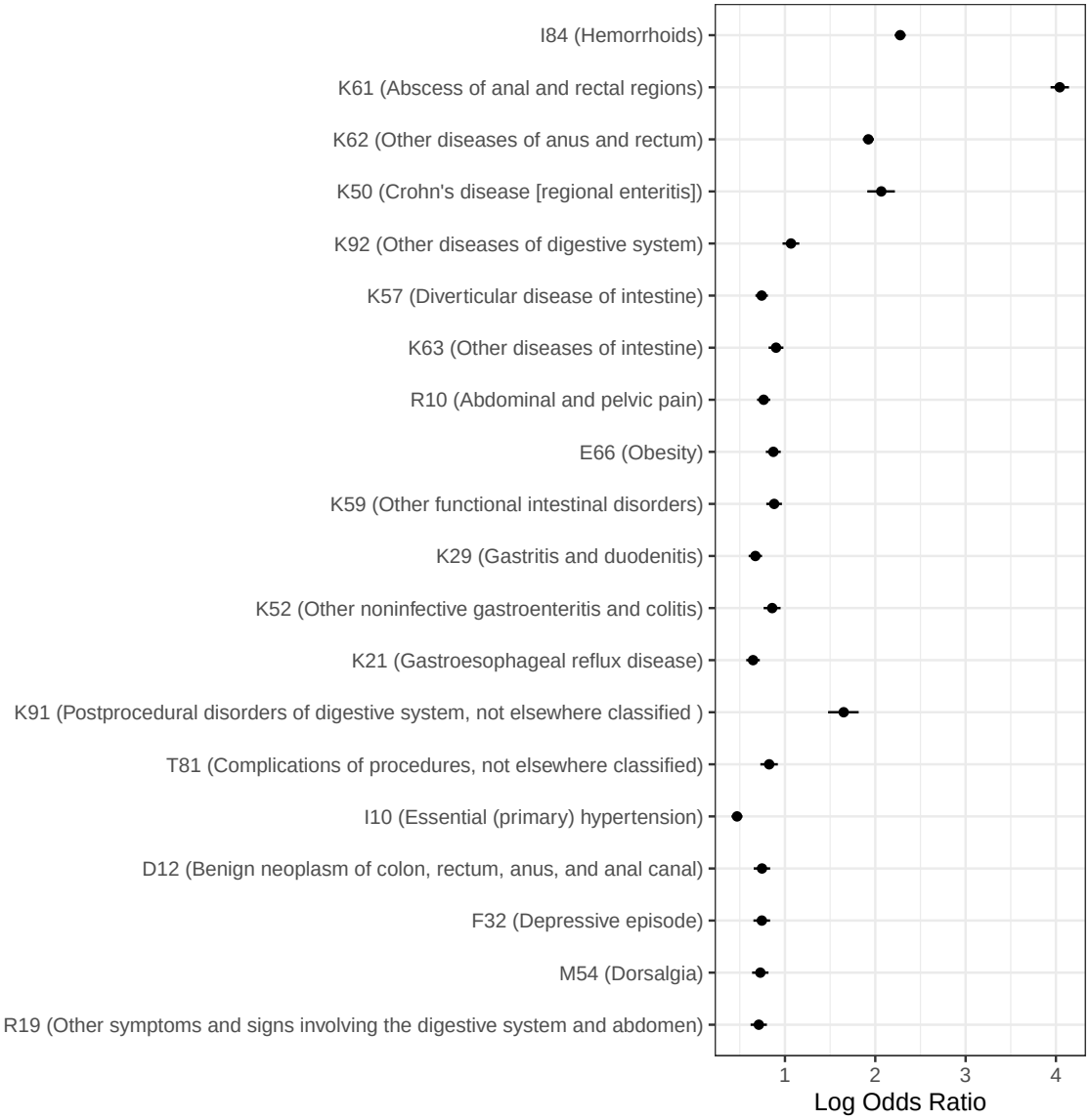



Fig. 2.1 Top 20 ICD-10 codes enriched in pAD cases versus pAD controls ordered by Fisher's exact test P-value. Odds ratios and their 95% confidence intervals were calculated using Fisher's exact test based on the number of pAD cases and controls with and without each particular ICD-10. Note that the control exclusion criteria in Table 2.2 were not applied in this analysis.

### 2.4.2 Identifying genome-wide significant loci

Defining genome-wide significant loci in GWAS studies is often complicated by the widespread correlation between proximal genetic variants (i.e. linkage disequillibrium or LD). To identify pAD-associated loci, I used an LD clumping procedure, which outputs a set of *index*

*variants*, each representing a set of highly correlated variants in a locus. Additionally, LD clumping identifies nominally-associated variants that are highly correlated with the index variant at each locus (which I will refer to as LD friends; $R^2 > 0.5$ and P-value < 0.01; Methods). In total, seven independent loci achieved genome-wide significant association (P-value $< 5 \times 10^{-8}$). All index variants were well-imputed (INFO $\geq 0.99$). I also compared the index variants MAFs to population MAFs to ensure that they did not significantly deviate from expected MAFs in non-Finnish Europeans (NFE). All index variants' MAFs matched MAFs obtained from 1000 Genomes Project (1000GP; Table 2.4; see Methods for how MAF deviation from the general population was formally assessed).

Table 2.4 Genome-wide significant index variants in the UKBB analysis. Odds ratio and their 95% confidence intervals are shown. Minor allele frequencies (MAF) in UKBB and 1000GP (NFE) are shown in the last two columns.

| Chromosome | Position (b38) | Effect Allele | Odds Ratio | P-value | MAF (UKBB) | MAF (100 |
|---|---|---|---|---|---|---|
| 3 | 52,992,368 | T | 1.13 (1.08 - 1.17) | $1.5 \times 10^{-8}$ | 0.42 | 0.44 |
| 6 | 31,044,486 | G | 1.13 (1.08 - 1.18) | $2.2 \times 10^{-8}$ | 0.37 | 0.37 |
| 6 | 31,113,288 | C | 1.13 (1.08 - 1.18) | $1.1 \times 10^{-8}$ | 0.41 | 0.44 |
| 6 | 31,113,923 | A | 1.12 (1.08 - 1.17) | $3.2 \times 10^{-8}$ | 0.49 | 0.49 |
| 6 | 31,148,469 | A | 1.12 (1.08 - 1.17) | $2.6 \times 10^{-8}$ | 0.44 | 0.45 |
| 9 | 22,119,196 | T | 0.89 (0.85 - 0.93) | $2.7 \times 10^{-8}$ | 0.48 | 0.47 |
| 11 | 10,356,352 | C | 0.88 (0.84 - 0.92) | $7.3 \times 10^{-9}$ | 0.29 | 0.30 |

Four of the seven loci were located in the major histocompatibility complex region (MHC; 6p21.33), and one locus in each of 3p21.1, 9p21.3 and 11p15.4. The MHC region is known to be highly polymorphic and exhibits complex and long-range LD patterns, which complicate the definition of independent loci. For example, two of the four MHC loci overlapped, with their two independent index variant located less than 700 bp apart. One of the two index variants (6:31113288_T_C) tagged a large number of variants ($R^2 > 0.8$) in the locus, while the other tagged no variants (6:31113923_A_G). Compared to the four MHC loci, the three non-MHC loci had less complex LD patterns. All three non-MHC index variants tagged a large number of variants and there was no overlapping independent loci in any of them. Given this complexity, I performed a number of post-GWAS checks to better understand the LD structure of all seven loci, which I will describe in the next section.

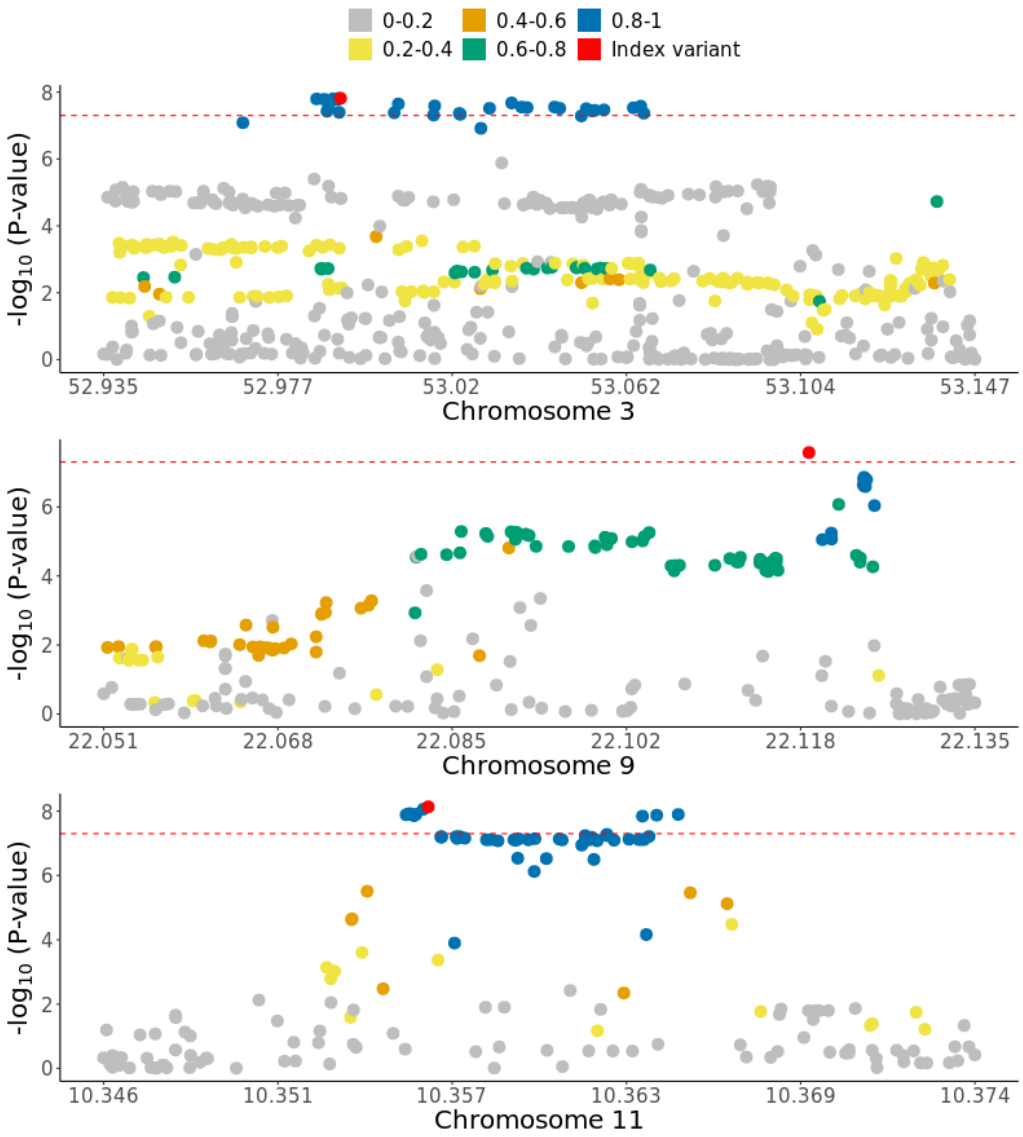



Fig. 2.2 Regional association plots for the three non-MHC loci, with position (build 38) plotted on the x-axis and $-log_{10}$ P-values shown on the y-axis for each variant. Colors indicate the $R^2$ between each variant and the index variant, and the red horizontal line indicates genome-wide significance (P-value = $5 \times 10^{-8}$).

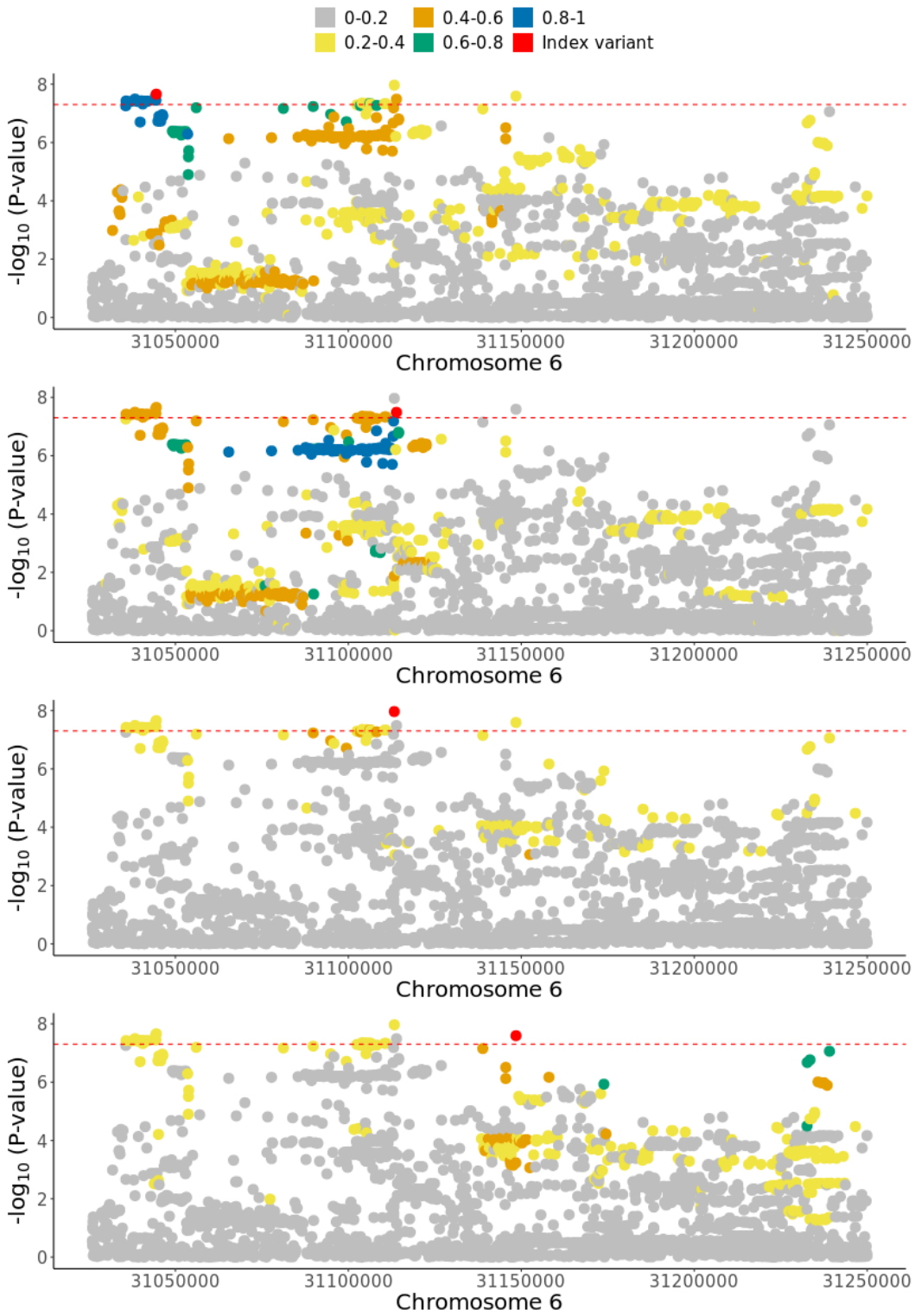



Fig. 2.3 Regional association plots for the four MHC loci, with position (build 38) plotted on the x-axis and $-log_{10}$ P-values shown on the y-axis for each variant. Colors indicate the $R^2$ between each variant and the index variant, and the red horizontal line indicates genome-wide significance (P-value = $5 \times 10^{-8}$). The second and third plots show the two overlapping MHC loci (index variants: 6:31113288_T_C and 6:31113923_A_G, respectively).

### 2.4.3 Post-GWAS quality checks

Spurious associations can seriously affect the validity of any significant results in GWAS studies. At the level of a single locus, spurious associations can be diagnosed by assessing the relationship between the index variant and its LD friends. For a given variant in a genome-wide significant locus, the lower its LD with the index variant, the weaker its association is expected to be. Loci where the association strength of LD friends does not "decay" as expected given their LD with the index variant are therefore paricularly problematic. Specifically, such a mismatch would suggest that the LD strcuture that drives the observed association strength of LD friends does not match the general population LD. A possible source of this mismatch may be crpytic subpopulation stratification, which often contributes to false positive associations in GWAS [113].

I investigated the seven genome-wide significant loci to ensure the association signal follows the expected LD pattern in the general population. For this check to be valid, LD needs to be computed from a suitable matching reference panel such as 1000GP. Additionally, each index variant needs to have a number variants LD friends. To this end, I computed $R^2$ between each variant and the index variant at each locus using NFE individuals in 1000GP as a reference panel. For each pAD-associated locus, I quantified the correlation between $R^2$ and P-values (on the $-log_{10}$ scale) of each index variant's LD friends. Additionally, I performed two follow-up assessments for the loci where this correlation is weak ($\rho < 0.2$) or cannot be computed due to a lack of LD friends.

### 2.4.4 Relationship between P-value and LD

Index variants in 3p21.1, 9p21.3 and 11p15.4 had a large number of LD friends (N=63, 66, and 49, respectively), and the P-values for each index variants' LD friends were highly correlated with $R^2$ ($\rho$ = 0.98, 0.74, 0.83, respectively), indicating that the P-values closely match the expected LD pattern in NFE. Two of the MHC loci also showed a similar LD decay pattern (index variants 6:31044486_G_C and 6:31148469_G_A in Figure 2.4), with a strong correlation between P-values and $R^2$ (Figure 2.4). However, this correlation did not hold for the two other overlapping MHC loci mentioned earlier, which motivated me to further investigate these two loci (ndex variants: 6:31113288_T_C and 6:31113923_A_G).

### A complex LD pattern at two MHC loci

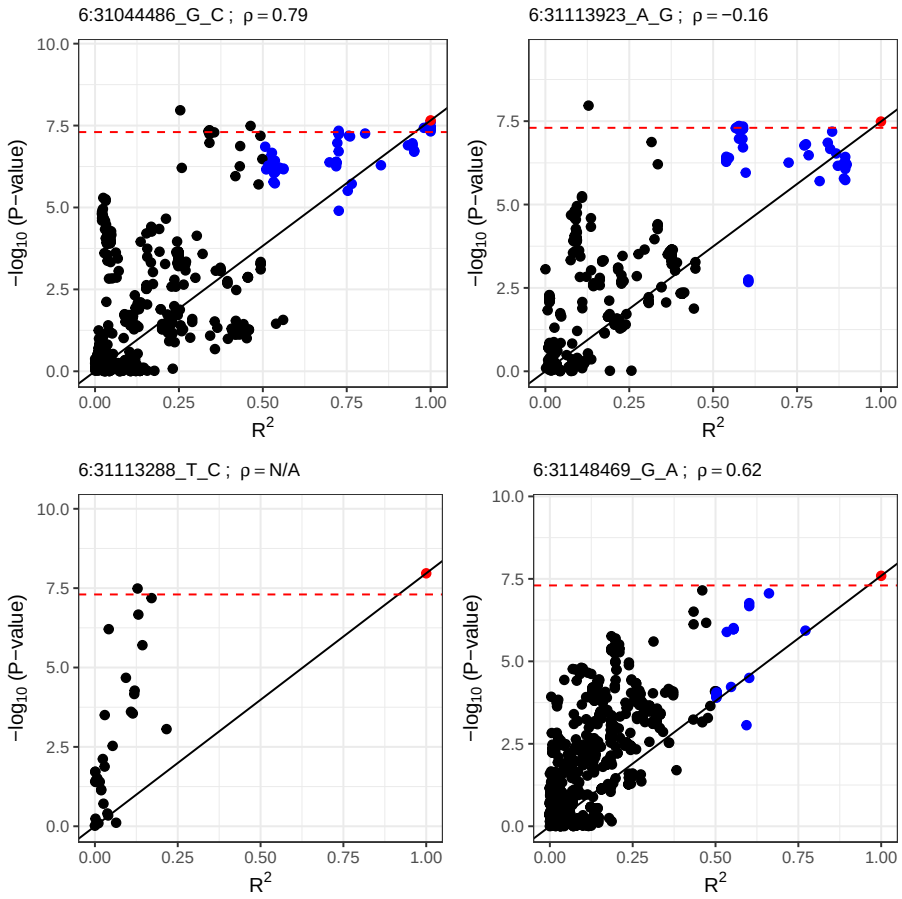



Fig. 2.4 LD decay plots showing association P-values for the four genome-wide significant loci in the MHC locus (x-axis) and each variant's $R^2$ with the index variant, derived from NFE in 1000GP (y-axis). Red dots and titles indicate the index variant in each locus. Blue dots indicate each index variant's LD friends, and the red horizontal line indicates genome-wide significance level (P-value $< 5 \times 10^{-8}$). The black line is fitted to the origin (0,0), and to the point $(1, -log_{10}(P_{index\_variant}))$, and shows the expected association strength given the LD with the index variant.

First, one of MHC index variants at the two overlapping MHC loci did not tag any LD friends and therefore the correlation between P-value and $R^2$ could not be assessed (6:31113288_T_C in Figure 2.4). It is unclear whether the absence of LD friends for 6:31113288_T_C suggests that it is a truly indepent variant, or whether it is driven by a mismatch between the LD patterns in UKBB and 1000GP. Such a mismatch may lead to an underestimateion of LD between the index variant and its LD friends. To answer this question, I recalculated the LD values in 1000GP using only British individuals (GBR; N=90), and found that the index

variant also did not tag any LD friends in 1000GP GBR as well. Given that 6:31113288_T_C is well-imputed (INFO=0.99) and common and that it is not well-tagged in both the NFE and GBR subpopulations in 1000GP, it is unlikely that the its association is driven by British-ancestry-specific LD. However, it is important to note that this does not rule out possible subpopulation stratification at this locus, which could potentially drive this association.

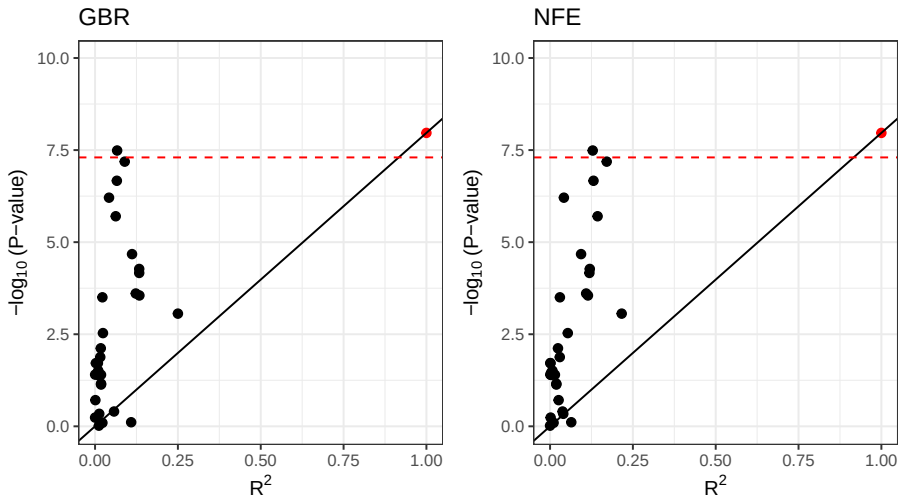



Fig. 2.5 LD decay plots showing association P-values for the locus around index variant 6:31113288_T_C. Each variant's $R^2$ with the index variant, derived from NFE and GBR in 1000GP is shown on the x-axis and $-log_{10}$ P-values on the y-axis, showing that the index variant does not tag any LD friends in both NFE and GBR. Red dots indicate the index variant adn the red horizontal line indicates genome-wide significance level (P-value $< 5 \times 10^{-8}$). The black line is fitted to the origin (0,0), and to the point (1,$-log_{10}(P_{index\_variant})$), and shows the expected association strength given the LD with the index variant.

Second, for the second overlapping locus, $R^2$ and P-values showed an inverse correlation (index variant 6:31113923_A_G in Figure 2.4; $\rho$=-0.16). This inverse correlation between $R^2$ and P-values suggest that not all the LD friends' P-values conform to their expected P-values given their LD with the index variant. I hypothesised that the reversal of correlation may be caused by the subset of LD friends with $R^2$ close to the value used for defining LD friends. This subset could lead to an inverse correlation due to a stronger-than-expected P-value given their LD with the index variant. To this effect, I found that 10 LD friends had a genome-wide significant P-value ($< 5 \times 10^{-8}$) despite all having an $R^2$ of 0.58 with the index variant. When I repeated the LD clumping procedure at this locus with a higher clumping $R^2$ cutoff (=0.6), I found that this subset of variants constituted a new genome-wide significant locus. This suggests that the identification of independent loci at this region is senstitive to the choice of

LD clumping $R^2$ cutoff, which further complicates the identification of independent loci at this region.

### 2.4.5 Finngen GWAS

Similar to UKBB, other national biobanks with genetic, clinical and phenotypic data are available. Although most national biobanks limit access to their individual-level genotype and phenotype data to approved researchers only, results from secondary analyses, including GWAS summary statistics, are made publicly available.

FinnGen is a national biobank whose aim is to collect genetic and phenotypic data for 500,000 Finnish individuals. The latest data freeze (Data Freeze 9) has genotyped over 377,000 individuals and has carried out GWAS for over 2,200 clinical endpoints. FinnGen uses a different clinical coding system from ICD to organise phenotypes into endpoints (FinnGen endpoints). There are two main differences between UKBB and FinnGen in terms of their clinical code structure. First, most FinnGen endpoints have parallel ICD codes, but additional FinnGen endpoints are created at request. Bespoke endpoints define certain inclusion or exclusion criteria based on ICD codes, or sometimes combine codes from different ICD chapters to create a new endpoint. Second, FinnGen endpoints are curated by experts in each field and are constantly reviewed in different FinnGen data freezes. They are brodaly classified as *core endpoints*, or *non-core endpoints*. Basic statistics such as prevalence and gender ratios are calculated for all FinnGen endpoints, while GWAS is conducted only for core endpoints.

ICD-10 code K60 corresponds to FinnGen endpoint K11_FISSANAL (Fissure and fistula of anal and rectal regions). K11_FISSANAL defines cases and controls similar to my UKBB cohort definition outlined in Table 2.2. However, K11_FISSANAL was considered a core endpoint only until Data freeze 7, and GWAS summary statistics for K11_FISSANAL are threfore unavailable in later data freezes.

### 2.4.6 Identification of genome-wide significant loci in Finngen

In order to investigate if the seven UKBB genome-wide significant loci replicated in an independent cohort and to identify additional loci, I downloaded GWAS summary statistics for FinnGen's clinical endpoint K11_FISSANAL. As of data freeze 7, Finngen reports 6,610 pAD cases and 253,186 controls. There was no further information regarding the subtypes of pAD (e.g. numbers of fissure and fistula cases), and it is therefore unclear if the composition

of FinnGen's pAD case cohort is similar to the UKBB pAD case cohort. Understanding the differences in subphenotype composition of each cohort is important to understand if differences in association at genome-wide significant loci is driven by genetic factors (e.g. differences in MAFs or LD structure) or by phenotypic differences between the cohorts.

After I filtered out variants with MAF < 0.01, a total of 9,054,355 variants remained. There was an acceptable level of genomic inflation (median $\chi^2$=0.495; $\lambda_{GC}$=1.089). To identify genome-wide significant loci, I used an LD clumping approach similar to the UKBB analysis, with the only difference being that I calculated LD from Finnish Europeans in 1000GP (FE; N=99). I found three genome-wide significant non-MHC loci: 1p34.2, 6p25.3 and 12q24.21 (P-value < $5 \times 10^{-8}$). Imputation quality information was not available in the downloaded summary statistics, so I was not able to confirm if the index variants had good imputation quality. However, the index variants' MAFs matched MAFs derived from FE in 1000GP, suggesting that they are imputed or genotyped with high accuracy (Table 2.5). Furthermore, I performed similar post-GWAS checks to UKBB to ensure the P-value of the index variants LD friends match their expected values given their LD with the index variant. All three showed a good decay of P-values with LD ($\rho$=0.92, 0.74 and 0.44, respectively; Figure 2.6)

Table 2.5 Genome-wide significant index variants in the FinnGen GWAS. Odds ratio and their 95% confidence intervals are shown. Minor allele frequencies (MAF) in UKBB and 1000GP (FE) are shown in the last two columns.

| Chromosome | Position (b38) | Effect Allele | Odds Ratio | P-value | MAF (Finngen) | MAF (1000 |
|---|---|---|---|---|---|---|
| 1 | 39,817,036 | T | 1.14 (1.09 - 1.19) | $7.2 \times 10^{-10}$ | 0.21 | 0.22 |
| 6 | 1,771,278 | T | 0.9 (0.87 - 0.93) | $6.7 \times 10^{-9}$ | 0.42 | 0.39 |
| 12 | 114,235,969 | T | 1.11 (1.07 - 1.15) | $7.0 \times 10^{-9}$ | 0.47 | 0.47 |

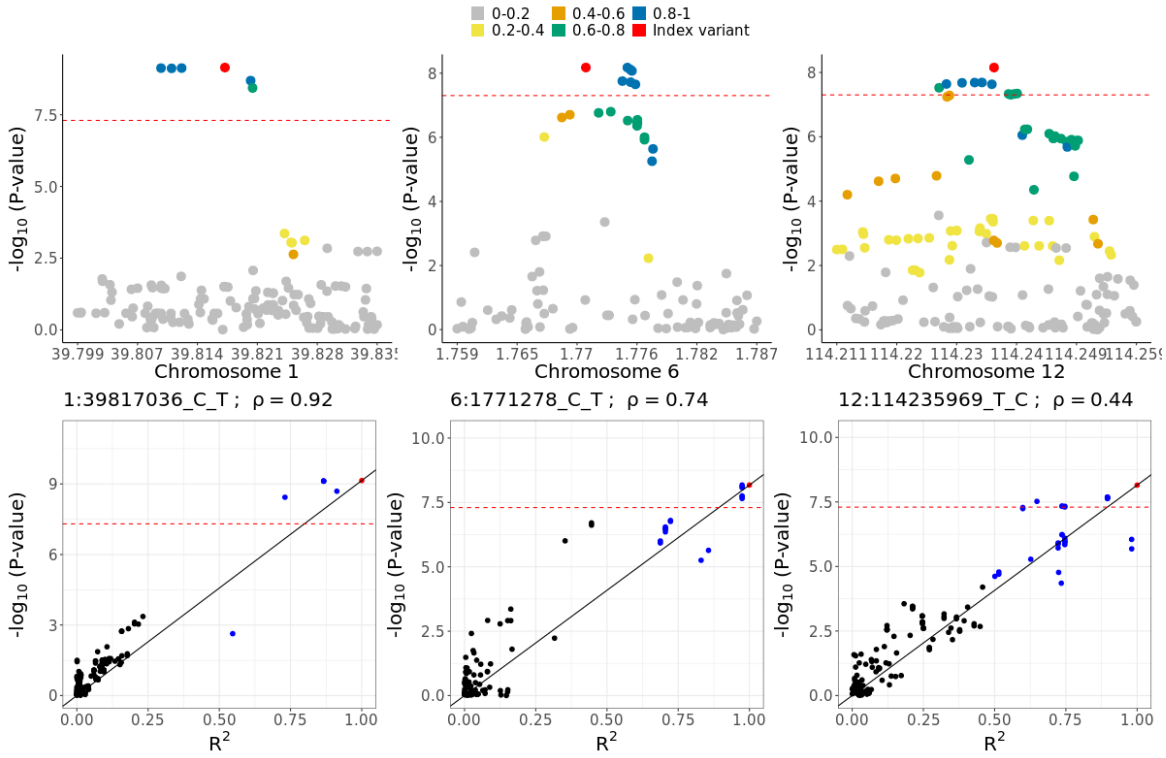



Fig. 2.6 (Top) Regional association plots for the three FinnGen loci, with position plotted on the x-axis and $-log_{10}$ P-values shown on the y-axis for each variant. Colors indicate the LD value of each variant with the index variant, and the red horizontal line indicates genome-wide significance (P-value = $5 \times 10^{-8}$). (Bottom) LD decay plots showing association P-values for the three genome-wide significant loci in FinnGen (x-axis) and each variant's $R^2$ with the index variant, derived from FE in 1000GP (y-axis). Red dots and titles indicate the index variant in each locus. Blue dots indicate each index variant's LD friends. The red horizontal line indicates genome-wide significance level, and the black line is fitted to the origin (0,0), and to the point $(1, -log_{10}(P_{index\_variant}))$, and shows the expected association strength given the LD with the index variant.

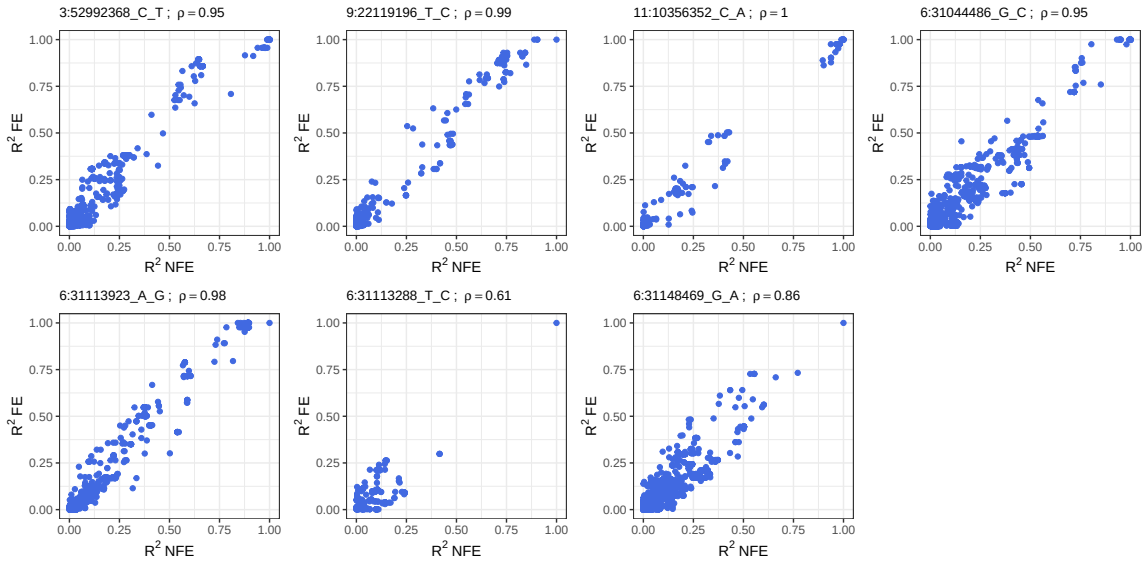
### 2.4.7 Replication of UKBB loci in Finngen

#### LD pattern in Finnish Europeans

In GWAS studies, the true causal variant in an associated locus is often unknown due to LD between variants. Moreover, it is often the case that the true causal variant may not even be genotyped in array-based GWAS studies, or may not be imputed due to different imputation protocols and QC metrics being used in different GWAS. When assessing replication of a GWAS locus between two cohorts, it is therefore important to ensure that variants that are genotyped or imputed in the two cohorts have similar LD structures. Indeed, a lack of GWAS

hit replication is sometimes driven by a difference in LD patterns between the two studies under comparison, one of which may not have genotyped or imputed any variants that tag the true causal variant in its respective population [114]. Finnish Europeans (FE) and NFE are known to exhibit systematic difference in their LD strucure, which may affect the ability to replicate the pAD-associated loci discovered in the UKBB. To compare the LD pattern between FE and NFE at the pAD-associated loci, I computed the LD between each variant and the index variant in the FE and NFE subpopulations of 1000GP.

(a)




(b)

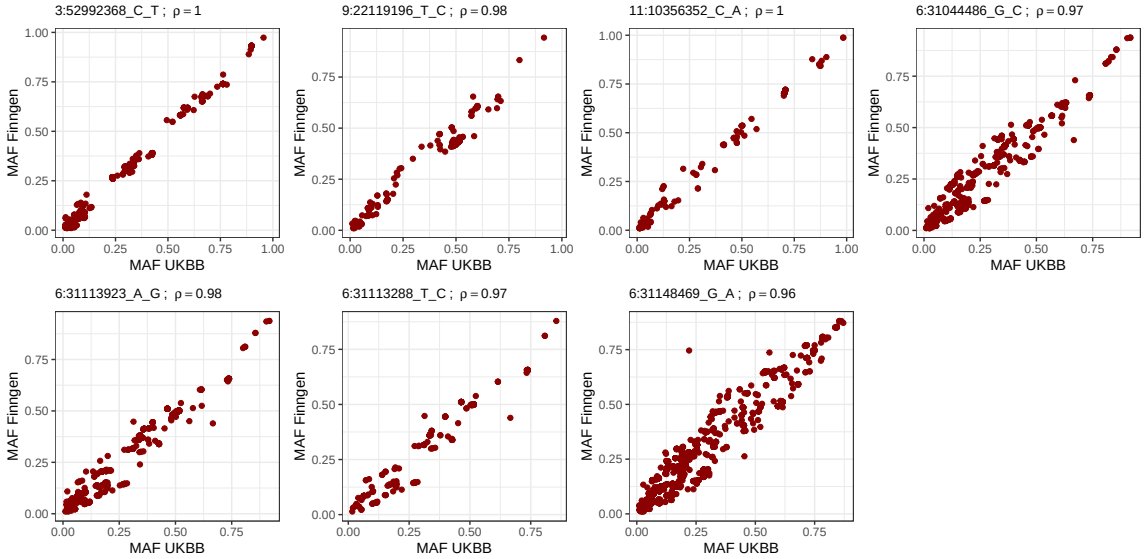



Fig. 2.7 (a) $R^2$ between all variants within each locus' boundaries and the index variant in the seven genome-wide significant loci identified in UKBB. $R^2$ are derived from non-Finnish Europeans (x-axis) and Finnish Europeans (y-axis) in 1000GP. Pearson correlation coefficients and index variants are indicated on top of each figure. (b) MAF of all variants in the UKBB (x-axis) and FinnGen (y-axis).

I found that MAF was nearly perfectly correlated in NFE and FE in all seven pAD-associated loci ($\rho > 0.94$; Figure 2.7a). $R^2$ were also strongly correlated in all loci ($\rho > 0.86$; Figure 2.7b). Notably, despite a strong $R^2$ correlation, 6:31113288_T_C did not have any LD friends in FE, similar to GBR and NFE, and the strong correlation at this locus

was driven by variants with a low $R^2$ with the index variant. Overall, witht the exception of 6:31113288_T_C which did not have any LD friends in FE, both MAF and the LD structure were consistent across all pAD-associated loci between NFE and FE. Replication of UKBB hits can therefore be reasonably assessed in the FinnGen GWAS. I found that two UKBB non-MHC index variants replicated in FinnGen (FinnGen P-value $< 7 \times 10^{-3}$ for 7 variants; Table 2.6), and that neither of them showed evidence of heterogeneity fo effect size ($P_{het}$ $< 7 \times 10^{-3}$; Table 2.6). Additionally, all the five index variants that failed to replicate also showed evidence of heterogeneity of effect sizes.

Table 2.6 Replication of the UKBB genome-wide significant index variants in FinnGen. Odds ratio and their 95% cofidence intervals are shown for both cohorts. The heterogeneity of effect P-value is shown in the last column. Only two index variants passed the replication threshold (3:52992368_C_T and 11:10356352_C_A)

| Index variant | P-value UKBB | FinnGen P-value | OR UKBB | OR FinnGen | $P_{het}$ |
|---|---|---|---|---|---|
| 3:52992368_C_T | $1.5 \times 10^{-08}$ | $1.2 \times 10^{-03}$ | 1.13 (1.08 - 1.17) | 1.06 (1.02 - 1.1) | 0.03 |
| 6:31044486_G_C | $2.2 \times 10^{-08}$ | $5.7 \times 10^{-02}$ | 1.13 (1.08 - 1.18) | 1.04 (1 - 1.07) | $1.9 \times 10^{-03}$ |
| 6:31113923_A_G | $3.2 \times 10^{-08}$ | $3.9 \times 10^{-02}$ | 1.12 (1.08 - 1.17) | 1.04 (1 - 1.07) | $3.6 \times 10^{-03}$ |
| 6:31113288_T_C | $1.1 \times 10^{-08}$ | $1.5 \times 10^{-01}$ | 1.13 (1.08 - 1.18) | 1.03 (0.99 - 1.07) | $7.2 \times 10^{-04}$ |
| 6:31148469_G_A | $2.6 \times 10^{-08}$ | $9.9 \times 10^{-01}$ | 1.12 (1.08 - 1.17) | 1 (0.97 - 1.04) | $2.0 \times 10^{-05}$ |
| 9:22119196_T_C | $2.7 \times 10^{-08}$ | $2.0 \times 10^{-02}$ | 0.89 (0.85 - 0.93) | 0.96 (0.93 - 0.99) | $6.0 \times 10^{-03}$ |
| 11:10356352_C_A | $7.3 \times 10^{-09}$ | $3.4 \times 10^{-07}$ | 0.88 (0.84 - 0.92) | 0.91 (0.87 - 0.94) | 0.27 |

### 2.4.8 Replication of FinnGen loci in UKBB

Following the same replication approach, I tested the replication of FinnGen's three genome-wide significant loci in the UKBB GWAS. I found evidence of replication for all three index variants in the UKBB (UKBB P-value $< 0.017$ for 3 variants), and none of the variants showed evidence of heterogeneity of effect sizes ($P_{het} < 0.017$; Table 2.7).

Table 2.7 Replication of the FinnGen genome-wide significant index variants in UKBB. Odds ratio and their 95% cofidence intervals are shown for both cohorts. The heterogeneity of effect P-value is shown in the last column.

| Index variant | P-value UKBB | FinnGen P-value | OR UKBB | OR FinnGen | $P_{het}$ |
|---|---|---|---|---|---|
| 1:39817036_C_T | $1.6 \times 10^{-06}$ | $7.2 \times 10^{-10}$ | 1.13 (1.08 - 1.19) | 1.14 (1.09 - 1.19) | 0.77 |
| 6:1771278_C_T | $1.5 \times 10^{-04}$ | $6.7 \times 10^{-09}$ | 0.91 (0.87 - 0.96) | 0.9 (0.87 - 0.93) | 0.63 |
| 12:114235969_T_C | $1.3 \times 10^{-03}$ | $7.0 \times 10^{-09}$ | 1.07 (1.03 - 1.12) | 1.11 (1.07 - 1.15) | 0.2 |

### 2.4.9 Meta-analysis of UKBB and FinnGen

Meta-analysis between GWAS cohorts is commonly used to increase statistical power to identify genome-wide significant loci. Practically, meta-analysis is carried out when the are constraints on sharing individual-level data, or when genotype data from several studies cannot be combined [115]. In these cases, meta-analysis of association summary statistics is the preferred analytical approach, and there is ample evidence that it achieves similar statistical power as combining genotype data from several studies [116].

#### Meta-analysis and identification of genome-wide significant loci

I performed a fixed-effects meta-analysis between UKBB and FinnGen effect sizes and standard errors using METAL (see Methods for more details). Because I performed a meta-analysis between two GWAS summary statistics from Finnish and Non-Finnish Europeans, I performed LD clumping separately with an LD panel from each of the two populations, and found 18 genome-wide significant loci (P-value $< 5 \times 10^{-8}$). I tested whether the index variants' effect size estimates were consistent between UKBB and FinnGen using Cochran's Q test, which is implemented in METAL (Methods). A strong deviation from the null hypothesis that effect sizes are similar between UKBB and FinnGen reflects uncertainty around the meta-analysed effect size estimate. To this end, I found no evidence of heterogenity for any of the 12 index variants ($P_{het} < 4 \times 10^{-3}$). Furthermore, I compared the association signal and LD for each locus using same two reference panels. Six of these loci either showed weak or inverse correlation between P-values and $R^2$ derived from either NFE and FE ($\rho <$ 0.2), and were therefore removed from the rest of the downstream analyses (more details in Methods).

Table 2.8 Meta-analysis genome-wide significant loci (P-value < $5 \times 10^{-8}$), showing the index variant at each locus, the meta-analysis P-value, and the odds ratio in the UKBB, FinnGen, and meta-analysis. 95% confidence intervals are shown for each odds ratio value. The last column shows the P-value of the effect size heterogeneity test, where $P_{het} < 4 \times 10^{-3}$ suggests evidence of heterogeneity of effects. The six loci that failed the LD decay test are highlighted in bold.

| Index variant | Meta-analysis P-value | OR UKBB | OR FinnGen | OR Meta-analysis | $P_{het}$ |
|---|---|---|---|---|---|
| **1:39809417_A_T** | $7.4 \times 10^{-15}$ | 1.13 (1.08 - 1.19) | 1.14 (1.09 - 1.19) | 1.14 (1.1 - 1.17) | 0.84 |
| **1:39836225_G_C** | $4.1 \times 10^{-08}$ | 1.09 (1.04 - 1.15) | 1.11 (1.06 - 1.16) | 1.1 (1.07 - 1.14) | 0.63 |
| 3:53034026_C_T | $7.5 \times 10^{-10}$ | 1.13 (1.08 - 1.17) | 1.07 (1.03 - 1.11) | 1.09 (1.06 - 1.12) | 0.05 |
| 5:64868326_TTTC_T | $2.0 \times 10^{-08}$ | 0.89 (0.85 - 0.93) | 0.94 (0.91 - 0.98) | 0.92 (0.89 - 0.95) | 0.05 |
| 6:1775202_G_A | $1.0 \times 10^{-11}$ | 0.91 (0.87 - 0.95) | 0.9 (0.87 - 0.93) | 0.9 (0.88 - 0.93) | 0.80 |
| 6:31121854_C_T | $4.2 \times 10^{-08}$ | 1.11 (1.07 - 1.16) | 1.06 (1.02 - 1.1) | 1.08 (1.05 - 1.11) | 0.08 |
| 6:31253340_T_C | $3.8 \times 10^{-08}$ | 1.1 (1.06 - 1.15) | 1.07 (1.03 - 1.1) | 1.08 (1.05 - 1.11) | 0.24 |
| **6:133008360_T_A** | $2.7 \times 10^{-08}$ | 1.12 (1.07 - 1.19) | 1.1 (1.05 - 1.15) | 1.11 (1.07 - 1.15) | 0.47 |
| 6:133260944_G_A | $4.7 \times 10^{-08}$ | 1.11 (1.06 - 1.17) | 1.08 (1.04 - 1.13) | 1.1 (1.06 - 1.13) | 0.42 |
| **6:133267939_T_C** | $4.5 \times 10^{-08}$ | 1.09 (1.05 - 1.14) | 1.07 (1.03 - 1.11) | 1.08 (1.05 - 1.11) | 0.42 |
| 7:2524404_G_A | $4.1 \times 10^{-08}$ | 1.14 (1.07 - 1.22) | 1.13 (1.06 - 1.2) | 1.14 (1.09 - 1.19) | 0.76 |
| 8:70735125_A_G | $3.9 \times 10^{-11}$ | 0.83 (0.77 - 0.9) | 0.82 (0.76 - 0.89) | 0.83 (0.78 - 0.87) | 0.94 |
| 8:70993166_AAGTT_A | $1.2 \times 10^{-10}$ | 0.83 (0.77 - 0.9) | 0.82 (0.75 - 0.88) | 0.82 (0.78 - 0.87) | 0.75 |
| **9:21995045_T_G** | $4.3 \times 10^{-08}$ | 1.41 (1.25 - 1.59) | NA | 1.41 (1.25 - 1.6) | 1.00 |
| 9:22124505_A_T | $2.1 \times 10^{-08}$ | 0.9 (0.86 - 0.93) | 0.95 (0.91 - 0.98) | 0.92 (0.9 - 0.95) | 0.05 |
| **10:61661180_A_G** | $2.0 \times 10^{-08}$ | 1.09 (1.04 - 1.14) | 1.08 (1.05 - 1.12) | 1.09 (1.06 - 1.12) | 0.90 |
| 11:10356352_C_A | $1.3 \times 10^{-13}$ | 0.88 (0.84 - 0.92) | 0.91 (0.87 - 0.94) | 0.89 (0.87 - 0.92) | 0.27 |
| 12:114235969_T_C | $4.2 \times 10^{-10}$ | 1.07 (1.03 - 1.12) | 1.11 (1.07 - 1.15) | 1.09 (1.06 - 1.12) | 0.20 |

### 2.4.10 Disentangling the genetic effect of pAD-associated variants on haemorrhoids

In section 2.4.1, I analysed the composition of the pAD case cohort and showed that it is significantly enriched with 198 ICD-10 clinical codes compared to pAD controls. Haemorrhoids was the most strongly eniched phenotype in pAD cases versus controls. I hypothesised that this enrichment was also reflected at the level of genetic risk predisposition. To confirm this, I carried out a genetic correlation analysis between the pAD meta-analysis and the Pan-UKBB haemorrhoids GWAS. I found strong evidence of high genetic correlation (ICD-10 code I84; P-value=$5.37 \times 10^{-26}$; $r_g$=0.66). To validate this correlation, I repeated the genetic correlation analysis using a larger haemorrhoids GWAS of over 900,000 individuals by Zheng et al. 2021 [109]. I found a similar genetic correlation estimate that was even more significant than the estimate from the Pan-UKBB analysis ($r_g$=0.63; P-value=$10^{-62}$).

The existence of a strong genetic correlation and enrichment of haemorrhoids could be explained by several factors. First, pAD could be a co-morbidity of haemorrhoids, in a similar way that Type 2 diabetes and obesity are co-morbidities. This could be a result of the same risk factors (genetic or otherwise) underlying both diseases, potentially with varying effect sizes. Alternatively, clinical diagnostic factors could also account for this overlap. Both diseases are among the differential diagnoses for patients presenting with rectal pain, swelling, bleeding and discharge. Therefore, a patient suffering from inflammed haemorrhoids is more likely to be diagnosed if they also suffer from pAD (e.g. after performing rectal examination).

Bias introduced by clinical diagnostic factors cannot be completely addressed with observational data, as this will require constructing pAD case-control cohorts where haemorrhoids cases are balanced in both cases and controls. However, the impact of such bias could also be assessed by performing a pAD GWAS where haemorrhoids cases are excluded from cases and controls (pADexclHaem), and a haemorrhoids GWAS where pAD cases are excluded from cases and controls (HaemexclpAD). Comparing the genetic association effect sizes of the previously reported 12 index variants between pADexclHaem and HaemexclpAD may give an indication as to which genetic variants are likely to underlie both diseases and which are likely to be specific to pAD.

Constructing the two cohorts requires access to a individual-level phenotypic data in both UKBB and FinnGen. Since I do not have access to FinnGen's phenotype data, I tested the hypothesis that effect sizes are different between haemorrhoids and pAD in the UKBB only. To construct the HaemexclpAD case and control cohorts, I selected individuals who

have been diagnosed with ICD-10 code I84 or ICD-9 code 455 in at least one inpatient episode as cases and excluded individuals with ICD-10 code K60 or ICD-9 code 565 from both cases and controls. Similarly, for pADexclHaem, I selected individuals who have been diagnosed with ICD-10 code K60 or ICD-9 code 565 in at least one inpatient episode and excluded individuals with ICD-10 code I84 or ICD-9 code 455 from both cases and controls. Additionally, I applied the same control exclusion criteria in Table 2.2.

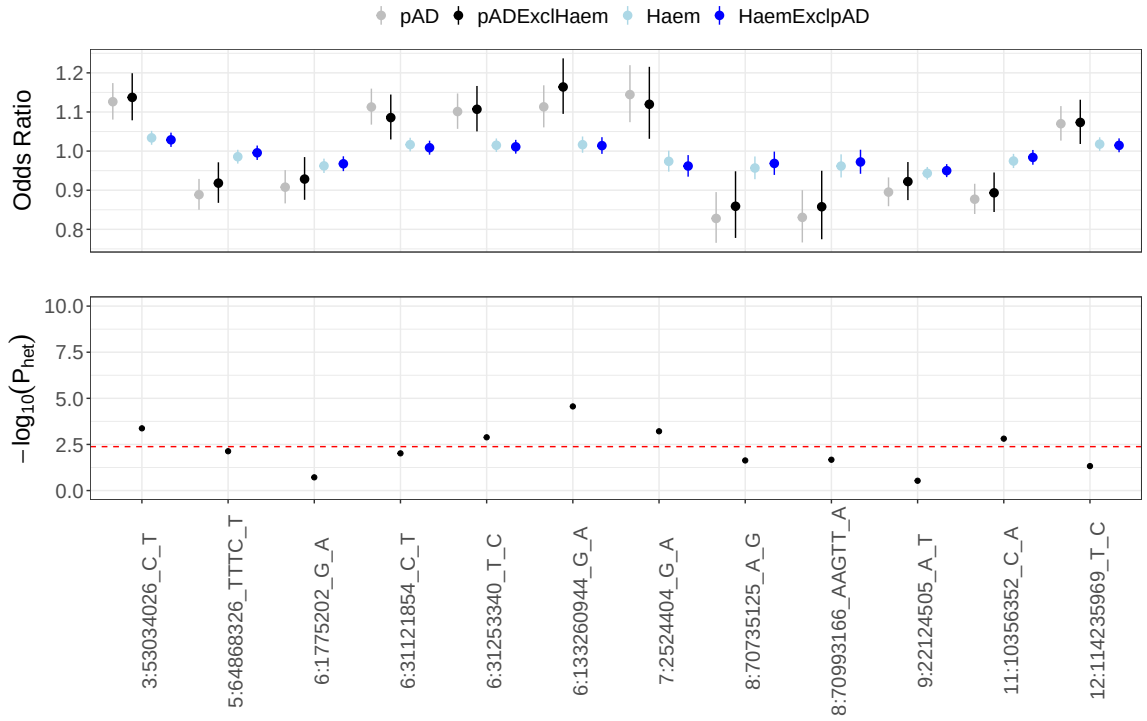



Fig. 2.8 Top plot shows the ffect sizes of the 12 pAD-assocaited index variants from four UKBB case/control cohorts: pAD in grey ($N_{case}$=4,606), pADExclHaem in black ($N_{case}$=2,799), Haem in light blue ($N_{case}$=29,285), and HaemExclpAD in blue ($N_{case}$=27,477). Bottom plot shows the heterogeneity of effect-size P-value ($P_{het}$) between the two disjoint GWAS analyses: pADExclHaem and HaemexclpAD. The red dotted line shows $P_{het}$ significance threshold ($P_{het} < 4 \times 10^{-3}$).

I tested the association of each of the 12 pAD-associated index variants with each of the four phenotypes described above. First, I examined if any of the index variants were associated with the two haemorrhoids phenotypes Haem and HaemExclpAD. Since I performed a targeted association test, I set a more permissive assocition threshold for declaring significance than normally used to declare genome-wide assoctions (P-value $< 4 \times 10^{-3}$ for 12 variants). Despite the large difference in statistical power between the two haemorrhoids cohorts and the two pAD cohorts, I found that only three of the tested index variants

achieved significant association in the better-powered haemorrhoids GWASes (index variants: 3:53034026_C_T, 6:1775202_G_A, and 9:22124505_A_T). Additionally, all three variants were significant in both Haem and HaemExclpAD, suggesting that the exclusion of pAD cases from the haemorrhoids cohort has little impact on their association. Moreover, 3:53034026_C_T showed significant evidence of effect size heterogeneity between pADExclHaem and HaemExclpAD ($P_{het} < 3 \times 10^{-3}$; Figure 2.8), with a significantly larger effect on pADExclHaem ($OR_{pADExclHaem} = 1.14 - 1.2$ and $OR_{HaemExclpAD} = 1.03 - 1.05$).

Although the other 11 index variants did not show a significant association with the two haemorrhoids definitions, four of them had significantly smaller effect sizes in HaemexclpAD than in pADexclHaem ($P_{het} < 4 \times 10^{-3}$; Figure 2.8). Moreover, five additional variants had suggestive evidence of heterogeneity of effect ($P_{het} < 0.05$), and for all five variants the effect size was larger in pADExclHaem than in HaemExclpAD.

Two conclusion can be made from this analysis. First, despite a much larger sample size in favour of the haemorrhoids cohorts, only three pAD-associated index variants were also associated with haemorrhoids, even with a relatively lenient threshold for association. Second, despite their nominal association with haemorrhoids, these three variants (and indeed all other variants) had a consistently smaller effect size on HaemExclpAD than pADExclHaem, and for 3:53034026_C_T that difference in effect size was significant. More importantly, this pattern was observed for all 12 index variants, despite the lack of power to detect a significant heterogeneity of effects. Performing a similar 'disentaglment' analysis in both FinnGen and UKBB, and subsequently identifying which variants have a significantly larger effect size in pADExclHaem than HaemExclpAD is a plausible way to validate this pattern. Such validation would more strongly establish these variants as bona fidae pAD-associated variants, with a significantly smaller effect size on haemorrhoids.

### 2.4.11 Identification of effector genes via colocalisation analysis

Many GWAS loci that have been uncovered over the last 15 years are located in non-coding regions. This complicates the task of understanding their downstream effects and linking them to effector genes. Over the last ten years, large-scale studies that map genetic variants associated with transcriptomic variation have improved our understanding of the downstream effects of disease-associated genetic variants. For example, the Gentype Expression Project (GTEx) has mapped genetic variants associated with individual variation in overall levels of gene expression (eQTL) and splicing (sQTL). Additionally, statistical methods that are able

to integrate association signals from different studies have been applied to GWAS and QTL data in order to investigate which effector genes likely underpin disease-associated GWAS signals. Colocalisation analysis, for example, quantifies the probability that two association signals are driven by a single causal variant ($PP_4$) and can therefore be used to compare GWAS and QTL association signals (more details in the Methods section).

I carried out colocalisation analysis between the 12 pAD-associated loci and eQTL and sQTL signals from GTEx v8 in a 1 mbp window centered around each locus' index variant. Across all 49 GTEx tissues, I performed the colocalisation with a total of 293 genes and all their splice junctions (see Methods for the number of genes and splice junctions tested at each locus).

The most informative output of colocalisation analysis is $PP_4$, the posterior probability of two association signals sharing a single underlying variant. Overall, I found high-confidence colocalisation evidence for seven loci, where at least one eQTL or sQTL colocalised with the association signal ($PP_4 > 0.8$; Table 2.9). All seven loci had at least a single colocalisation with an sQTL (12 sQTL genes), while five loci colocalised with at least one eQTL signal (eight eQTL genes), implicating a total of 15 genes. At many loci where both an eQTL and sQTL colocalisation were detected, distinct eQTL and sQTL genes were implicated. Moreover, QTLs in different tissues often implicated different genes. For example, the locus at index variant 7:2524404_G_A colocalised with two different genes (*BRAT1* in the liver and thyroid gland, and *LFNG* in the skin and whole blood). In fact, only three of the seven colocalised loci implicated a single gene, and only one locus implicated the same gene with high confidence in multiple tissues (index variant 5:64868326_TTTC_T). The pleiotropic nature of genetic effects on gene expression is well documented in GTEx [117], and even in other organisms [118, 119]. This pleiotropy is often attributed to the widespread gene co-expression patterns, whereby the expression of multiple genes is controlled by a single locus, sometimes termed "QTL hotspots" [120]. Co-expressed genes are often found to be functionally related via shared biological pathways [121, 122]. To explore this, I performed a gene set enrichment analysis in four databases: Reactome [123], the Gene Ontlogy (GO) Molecular Function database, GO Cellular Component and GO Biological processes [124]. I did not find any significantly enriched pathways in any of the the three GO databases or the Reactome database. Notably, 6 of the 17 genes were not found in the Reacome database, reflecting the lack of knowledge of their biological functions.

Table 2.9 Colocalisation analysis for the 12 pAD-associated index variants. The first column shows the index variants and the second and third columns shows the tissues and genes with high colocalisation $PP_4$ (> 0.8). Genes and their $PP_4$ values are shown in parentheses.

| Index SNP | Tissues (eQTL) | Tissues (sQTL) |
|---|---|---|
| 3:53034026_C_T | Kidney Cortex (ITIH4: 0.97), Colon Transverse (SFMBT1: 0.98), Esophagus Gastroesophageal Junction (SFMBT1: 0.95), Esophagus Muscularis (TMEM110: 0.82), Pancreas (TMEM110: 0.98) | Artery Aorta (ITIH4: 0.9), Artery Tibial (ITIH4: 0.97), Liver (ITIH4: 0.96), Nerve Tibial (ITIH4: 0.93), Liver (MUSTN1: 0.87), Esophagus Muscularis (RFT1: 0.84) |
| 5:64868326_TTTC_T | Testis (CWC27: 0.86) | Esophagus Muscularis (CWC27: 0.81), Testis (CWC27: 0.84) |
| 6:31121854_C_T | Lung (HLA-B: 0.86), Thyroid (HLA-B: 0.87), Adrenal Gland (POU5F1: 0.85), Brain Cerebellar Hemisphere (POU5F1: 0.89), Brain Cerebellum (POU5F1: 0.91), Brain Hypothalamus (POU5F1: 0.84) | Skin Not Sun Exposed Suprapubic (FLOT1: 0.85), Skin Not Sun Exposed Suprapubic (MICA: 0.81), Colon Transverse (PSORS1C1: 0.98), Lung (PSORS1C1: 0.99), Small Intestine Terminal Ileum (PSORS1C1: 0.89) |

Table 2.9 *(continued)*

| Index SNP | Tissues (eQTL) | Tissues (sQTL) |
|---|---|---|
| 6:31253340_T_C | Lung (HLA-B: 0.86), Thyroid (HLA-B: 0.87), Adrenal Gland (POU5F1: 0.85), Brain Cerebellar Hemisphere (POU5F1: 0.89), Brain Cerebellum (POU5F1: 0.91), Brain Hypothalamus (POU5F1: 0.84) | Skin Not Sun Exposed Suprapubic (MICA: 0.81), Colon Transverse (PSORS1C1: 0.98), Lung (PSORS1C1: 0.99), Small Intestine Terminal Ileum (PSORS1C1: 0.89) |
| 7:2524404_G_A | Liver (BRAT1: 0.94), Skin Not Sun Exposed Suprapubic (LFNG: 0.92), Skin Sun Exposed Lower leg (LFNG: 0.99), Whole Blood (LFNG: 0.85) | Thyroid (BRAT1: 0.95), Skin Not Sun Exposed Suprapubic (LFNG: 0.86), Skin Sun Exposed Lower leg (LFNG: 0.95), Whole Blood (LFNG: 0.99) |
| 11:10356352_C_A | NA | Adrenal Gland (AMPD3: 0.85) |
| 12:114235969_T_C | NA | Vagina (RBM19: 0.81) |

Only one locus consistently implicated a single gene across various tissues and with both eQTL and sQTL colocalisation evidence (*CWC27*; index variant 5:64868326_TTTC_T; $PP_4 > 0.8$). The protective allele of the index variant (odds ratio=0.91) increases the expression of *CWC27* (eQTL effect size in testis=0.29; Figure 2.9), and also changes usage of five *CWC27* splice junctions in testis. *CWC27* codes for a splicesomoal complex component. Although little is known about its role in commom complex disease, rare *CWC27* variants are known to be associated with retinitis pigmentosa with or without skeletal symptoms. Indeed, Xu et al. [125] performed whole-exome sequencing of ten individuals from seven unrelated families, nine of which suffered from retinitis pigmentosa, either alone or with a range of structural

disorders such as bradydactyly, short stature, and craniofacial defects. In all seven families, rare protein-truncating variants in CWC27 were found, establishing *CWC27* as the effector gene for this group of rare disorders.

It is not obvious how pAD-assciated *CWC27* variants and rare *CWC27* mutations that cause skeletal abnormalities converge on the same molecular pathways. However, I found that another colocalised gene, *LFNG*, is associated with skeletal abnormalities. Two reports have shown that missense variants in *LFNG* are associated with spondylocostal dysostosis, a congenital disorder characterised by short neck, short stature and scoliosis [126, 127]. *LFNG* codes for a member of the glycosyltransferase family and plays an important role in vertebral formation during embryogenesis [128]. Notably, the two reported variants affected the active site of the protein product, rendering it functionally inactive. A compelling hypothesis that emerges from the loss-of-function evidence of these two genes is that genes important for normal musculskeletal development may also implicated in pAD pathogenesis. Most relevant literature cites the cryptoglandular theory to account for the origin of anal fistulas, whereby perianal abscess that starts in the proctodeal glands extends to form fistulas [129]. But little is known so far about how genetic predisposition affects this progression and which biological pathways are responsible for this progression. In this regard, the implication of *CWC27* and *LFNG* as effector genes hints at a possible role for pathways responsible for normal skeletal development, but more robust genetic evidence is needed to identify effector genes, especially at the remaining loci, where colocalisation evidence implicates multiple genes.

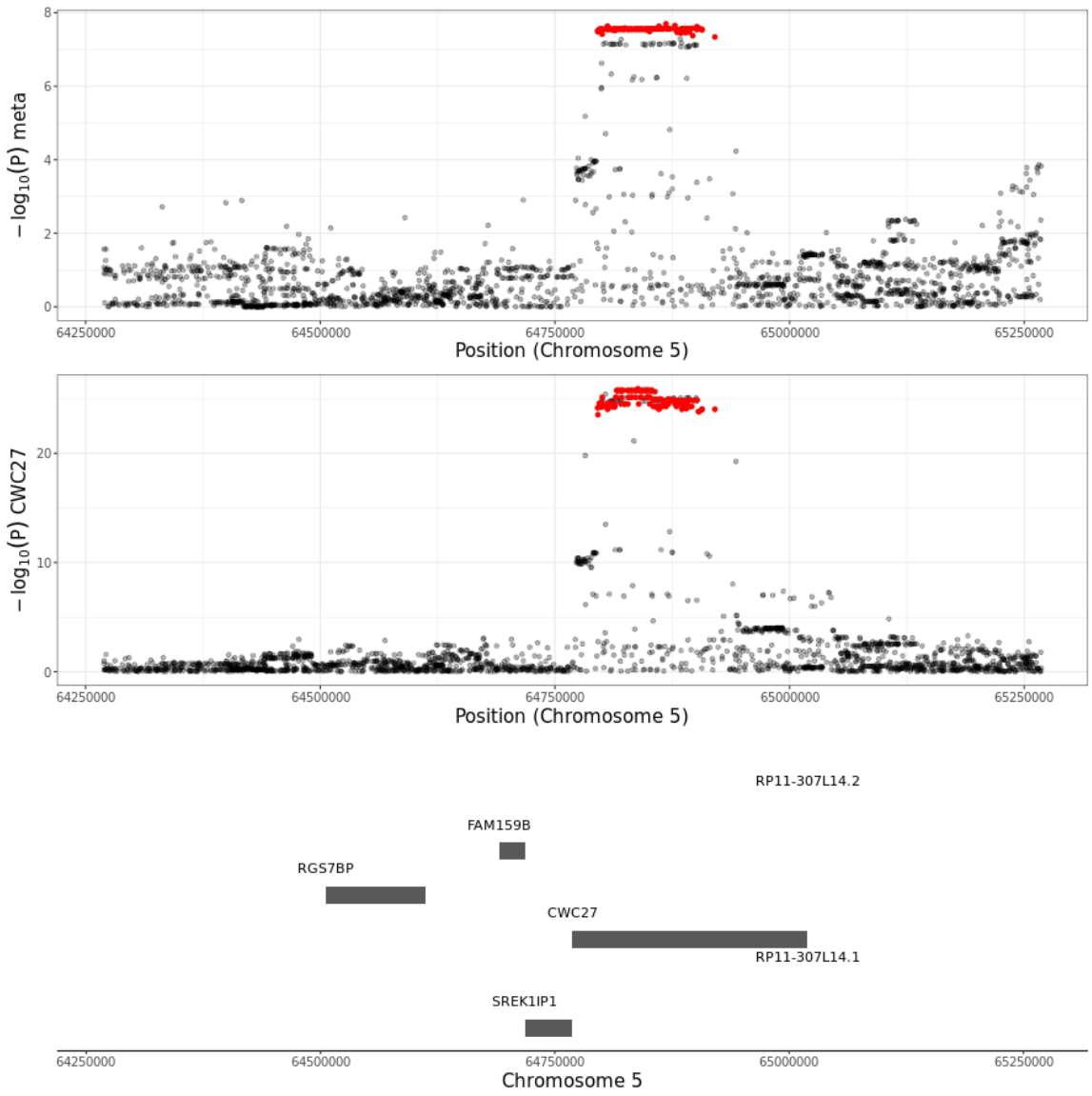



Fig. 2.9 Regional association plot for the pAD-associated signal at index variant 5:64868326_TTTC_T (top) and *CWC27* eQTL in testis (middle), showing position on the x-axis and $-log_{10}$ (P-value) on the y-axis. Red dots show pAD-associated genome-wide significant variants. Protein-coding and lincRNA gene positions are also shown (bottom).

## 2.5 Discussion

In this chapter, I have performed several analysis to map genetic variants associated with perianal disease (pAD), defined as anal and rectal fissures and fistulas. I have leveraged two large-scale national biobanks, UKBB and FinnGen, with a total of 11,216 cases and 585,420 controls. First, I performed a separate UKBB GWAS and found seven genome-wide

significant loci, that I subjected to a number of post-GWAS quality checks to ensure their veracity. Next, I downloaded FinnGen GWAS summary statistics for the same phenotype and performed similar post-GWAS quality checks. I also attempted to cross-replicate each GWAS' findings, and found that three of the seven UKBB loci replicated in FinnGen, while all FinnGen loci replicated in the UKBB. Interestingly, all the UKBB loci that failed to replicate in FinnGen were located in the MHC region, which is known to be highly polymorphic and exhibits complex LD patterns that tend to be population-specific.

There are several possible explanations for the absence of replication at the MHC loci. First, the four MHC loci discovered in the UKBB GWAS may simply be spurious associations. As I showed in Section 2.4.4, association strength and $R^2$ were not correlated in at least two of the MHC loci, which suggests that the underlying LD strcuture does not match general population LD and may lead to spurious associations at these loci. This could possibly result from cryptic population stratification, but it could also result from poor genotyping or imputation. Second, it is possible that neither the index variants nor their LD friends tag the true causal variant in FE. This is likely to be the case for the MHC loci where the index variant tags few or no LD friends. Third, a lack of statistical power might lead to an absence of replication, especially given the heterogeneity in pAD cases. The UKBB pAD case cohort is composed of perianal fissure and fistula cases, with fistulas representing 37.7% of the pAD cases and it is unclear whether the FinnGen case cohort composition is similar.

Despite the increased power afforded by meta-analysis, several limitations should be noted. First, although case and control inclusion criteria are similar between the two cohorts, it is not obvious if the composition of the pAD case cohort is also similar. Similar to the ICD codes used in the UKBB analysis to identify anal fissures and fistula cases, FinnGen's clinical endpoint covers two broad clinical diagnoses: anal fissures and fistulas. I have shown in Table 2.1 that the proportion of anal fissures and fistula cases in the UKBB case cohort is roughly 2:1. Since FinnGen's individual-level data are not publicly available, I could not confirm if the proportion of anal fissure and fistula cases are similar. Second, it is unclear if FinnGen's case cohort is enriched in any other clinical endpoints compared to FinnGen's control cohort. Showing that the pAD cases are enriched in the same disorders (e.g. haemorrhoids and anal abscess) can serve as an important phenotypic quality control check to ensure that both cohorts are as similar as possible, and maximises the ability of a meta-analysis to identify genetic variants associated with pAD risk. In this study, I was limited by the restricted access to individual-level FinnGen data, but future work should focus on

assessing the contribution of each of the pAD subtypes to each genome-wide significant locus.

I observed that pAD cases are highly enriched in haemorrhoids compared to pAD controls (38% versus 6% respectively), which may suggest that there may be shared genetic effects underlying both pAD and haemorrhoids. I therefore performed an analysis where I disentangled the effects of the pAD-associated variants on both diseases. This analysis showed that the effects of these variants were stronger and more significant on pAD than haemorrhoids despite a large difference in statistical power in favour of haemorrhoids. However, this analysis was limited to UKBB participant, and it is plausible that better powered GWAS of haemorrhoids may reveal that a larger proportion of the 12 pAD-associated variants are also associated with haemorrhoids. Indeed, when I replicated the index variants in the largest haemorrhoids GWAS (Zheng et al. [109]), I found that six of the 12 variants showed genome-wide significant association (P-value $< 5 \times 10^{-8}$). However, similar to the UKBB analysis, they all had a concordant but smaller effect sizes on haemorrhoids than pAD. A compelling interpretation of this shared genetic risk is that pAD may be a more severe form or manifestation of haemorrhoids, with the same genetic variants underlying both and with stronger effect sizes on pAD. But this observation may not be true for all haemorrhoids-associated loci. Over 100 haemorrhoids-associated loci were identified by Zheng et al., and it is plausible that most of these loci will not be associated with pAD if a genome-wide significant comparison of effect sizes was performed between the two diseases. Therefore, any conclusions made regarding the difference in effect sizes should be limited to these 12 loci.

Finally, I aimed to identify effector genes at each locus using colocalisation analysis. Although I identified several genes that colocalised with high confidence, evidence at most loci was conflicting, implicating several genes in several tissues. The role of these genes in many of the tissues where the colocalistions were detected, such as testis, thyroid and liver, was difficult to interpret given our knowledge of the pathogenesis of pAD. Although I showed evidence that two of these genes, *LFNG* and *CWC27*, were necessary for normal skeletal development, this does not constitute sufficient evidence for a novel insight into pAD pathogenesis. To this end, more follow-up work should be conducted to better interpret these loci. First, more robust methods need to be employed to establish a causal link between these loci and effector genes (e.g. Mendelian Randomisation methods [130]). Additionally, QTL studies from more relevant tissues need to be used. As discussed in 2.4.11, it is well-known that genetic variants affect the expression of different genes in different tissues. Therefore, QTLs derived from anorectal tissues will provide the best colocalisation and

mendelian randomisation evidence for effector genes. However, the anal region is composed of several tissues and cell types, and it is plausible that colocalisation with single-cell QTLs derived from anorectal biopsies will also potentially implicate different genes in different cell types. Therefore, the first step to identify the most likely effector genes is to identify the most relevant cell type via a heritability enrichment analysis (e.g. LDSC-SEG [131]). In conclusion, establishing a bona fide set of effector genes for these loci in relevant tissues will provide much stronger evidence that points to biological pathways implicated by pAD loci effector genes.

# 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] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] Christoph Gasche, Jurgen Scholmerich, Jorn Brynskov, Geert D'Haens, Stephen B Hanauer, Jan E Irvine, Derek P Jewell, Daniel Rachmilewitz, David B Sachar, William J Sandborn, and Lloyd R Sutherland. A simple classification of crohn's disease: Report of the working party for the world congresses of gastroenterology, vienna 1998. *Inflamm. Bowel Dis.*, 6(1):8–15, February 2000.

[13] 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.

[14] Raghu Kalluri and Robert A Weinberg. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.*, 119(6):1420–1428, June 2009.

[15] 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 $\beta$ synergise in the pathogenesis of human intestinal fistulae. *Gut*, 62(1):63–72, January 2013.

[16] Frauke Bataille, Christian Rohrmeier, Richard Bates, Achim Weber, Florian Rieder, Julia Brenmoehl, Ulrike Strauch, Stefan Farkas, Alois Fürst, Ferdinand Hofstädter, Jürgen Schölmerich, Hans Herfarth, and Gerhard Rogler. Evidence for a role of epithelial mesenchymal transition during pathogenesis of fistulae in crohn's disease. *Inflamm. Bowel Dis.*, 14(11):1514–1527, November 2008.

[17] F Bataille, F Klebl, P Rümmele, J Schroeder, S Farkas, P-J Wild, A Fürst, F Hofstädter, J Schölmerich, H Herfarth, and G Rogler. Morphological characterisation of crohn's disease fistulae. *Gut*, 53(9):1314–1321, September 2004.

[18] Eddie Cano-Gamez and Gosia Trynka. From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Front. Genet.*, 11:424, May 2020.

[19] 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.

[20] Xiaoyi Hu, Jing Li, Maorong Fu, Xia Zhao, and Wei Wang. The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduct. Target. Ther.*, 6(1):402, November 2021.

[21] Farhad Seif, Majid Khoshmirsafa, Hossein Aazami, Monireh Mohsenzadegan, Gholamreza Sedighi, and Mohammadali Bahar. The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. *Cell Commun. Signal.*, 15(1), December 2017.

[22] 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.

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

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

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

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

[27] T W Eglinton, R Roberts, J Pearson, M Barclay, T R Merriman, F A Frizelle, and R B Gearry. Clinical and genetic risk factors for perianal crohn's disease in a population-based cohort. *Am. J. Gastroenterol.*, 107(4):589–596, April 2012.

[28] A Latiano, O Palmieri, S Cucchiara, M Castro, R D'Incà, G Guariso, B Dallapiccola, M R Valvano, T Latiano, A Andriulli, and V Annese. Polymorphism of the IRGM gene might predispose to fistulizing behavior in crohn's disease. *Am. J. Gastroenterol.*, 104(1):110–116, January 2009.

[29] Philip J Tozer, Kevin Whelan, Robin K S Phillips, and Ailsa L Hart. Etiology of perianal crohn's disease: Role of genetic, microbiological, and immunological factors. *Inflamm. Bowel Dis.*, 15(10):1591–1598, October 2009.

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

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

[32] 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.

[33] 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.

[34] 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.

[35] Takashi Shiina, Kazuyoshi Hosomichi, Hidetoshi Inoko, and Jerzy K Kulski. The HLA genomic loci map: expression, interaction, diversity and disease. *J. Hum. Genet.*, 54(1):15–39, January 2009.

[36] Meral Beksac, editor. *Bone marrow and stem cell transplantation*. Methods in molecular biology (Clifton, N.J.). Humana Press, New York, NY, 2 edition, January 2014.

[37] Paul I W de Bakker, Gil McVean, Pardis C Sabeti, Marcos M Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, Alienke J Monsuur, Pamela Whittaker, Marcos Delgado, Jonathan Morrison, Angela Richardson, Emily C Walsh, Xiaojiang Gao, Luana Galver, John Hart, David A Hafler, Margaret Pericak-Vance, John A Todd, Mark J Daly, John Trowsdale, Cisca Wijmenga, Tim J Vyse, Stephan Beck, Sarah Shaw Murray, Mary Carrington, Simon Gregory, Panos Deloukas, and John D Rioux. A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.*, 38(10):1166–1172, October 2006.

[38] Alienke J Monsuur, Paul I W de Bakker, Alexandra Zhernakova, Dalila Pinto, Willem Verduijn, Jihane Romanos, Renata Auricchio, Ana Lopez, David A van Heel, J Bart A Crusius, and Cisca Wijmenga. Effective detection of human leukocyte antigen risk alleles in celiac disease using tag single nucleotide polymorphisms. *PLoS One*, 3(5):e2270, May 2008.

[39] Xiaoming Jia, Buhm Han, Suna Onengut-Gumuscu, Wei-Min Chen, Patrick J Concannon, Stephen S Rich, Soumya Raychaudhuri, and Paul I W de Bakker. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*, 8(6):e64683, June 2013.

[40] X Zheng, J Shen, C Cox, J C Wakefield, M G Ehm, M R Nelson, and B S Weir. HIBAG–HLA genotype imputation with attribute bagging. *Pharmacogenomics J.*, 14(2):192–200, April 2014.

[41] Seungho Cook, Wanson Choi, Hyunjoon Lim, Yang Luo, Kunhee Kim, Xiaoming Jia, Soumya Raychaudhuri, and Buhm Han. Accurate imputation of human leukocyte antigens with CookHLA. *Nat. Commun.*, 12(1):1264, February 2021.

[42] Tatsuhiko Naito, Ken Suzuki, Jun Hirata, Yoichiro Kamatani, Koichi Matsuda, Tatsushi Toda, and Yukinori Okada. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat. Commun.*, 12(1):1639, March 2021.

[43] Tatsuhiko Naito and Yukinori Okada. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. *Semin. Immunopathol.*, 44(1):15–28, January 2022.

[44] Xiuwen Zheng and Bruce S. Weir. Hibag - hla genotype imputation with attribute bagging, 01/03/2017. Accessed on 25/10/2023.

[45] 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.

[46] 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.

[47] 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.

[48] 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-TNF$\alpha$ therapy for anal fistula in crohn's disease. *Colorectal Dis.*, 17(4):311–319, April 2015.

[49] 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.

[50] David A Schwartz, Laurent Peyrin-Biroulet, Karen Lasch, Shashi Adsul, and Silvio Danese. Efficacy and safety of 2 vedolizumab intravenous regimens for perianal fistulizing crohn's disease: ENTERPRISE study. *Clin. Gastroenterol. Hepatol.*, 20(5):1059–1067.e9, May 2022.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Kristin L Ayers, Chrysovalanto Mamasoula, and Heather J Cordell. Penalized-regression-based multimarker genotype analysis of genetic analysis workshop 17 data. *BMC Proc.*, 5(S9):S92, December 2011.

[60] Vasiliki Matzaraki, Vinod Kumar, Cisca Wijmenga, and Alexandra Zhernakova. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.*, 18(1), December 2017.

[61] S G E Marsh, E D Albert, W F Bodmer, R E Bontrop, B Dupont, H A Erlich, M Fernández-Viña, D E Geraghty, R Holdsworth, C K Hurley, M Lau, K W Lee, B Mach, M Maiers, W R Mayr, C R Müller, P Parham, E W Petersdorf, T Sasazuki, J L Strominger, A Svejgaard, P I Terasaki, J M Tiercy, and J Trowsdale. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75(4):291–455, April 2010.

[62] Anthony Nolan Research Institute. Nomenclature for factors of the hla system, 20/09/2019. Accessed on 24/10/2023.

[63] Anthony Nolan Research Institute. Full list of class i proteins, 06/10/2023. Accessed on 24/10/2023.

[64] 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.

[65] Yong Huang, Peter M Krein, Daniel A Muruve, and Brent W Winston. Complement factor B gene regulation: synergistic effects of TNF-alpha and IFN-gamma in macrophages. *J. Immunol.*, 169(5):2627–2635, September 2002.

[66] Kim Goring, Yong Huang, Connie Mowat, Caroline Léger, Teik-How Lim, Raza Zaheer, Dereck Mok, Lee Anne Tibbles, David Zygun, and Brent W Winston. Mechanisms of human complement factor B induction in sepsis and inhibition by activated protein C. *Am. J. Physiol. Cell Physiol.*, 296(5):C1140–50, May 2009.

[67] Philippe Goyette, International Inflammatory Bowel Disease Genetics Consortium, Gabrielle Boucher, Dermot Mallon, Eva Ellinghaus, Luke Jostins, Hailiang Huang, Stephan Ripke, Elena S Gusareva, Vito Annese, Stephen L Hauser, Jorge R Oksenberg, Ingo Thomsen, Stephen Leslie, Mark J Daly, Kristel Van Steen, Richard H Duerr, Jeffrey C Barrett, Dermot P B McGovern, L Philip Schumm, James A Traherne, Mary N Carrington, Vasilis Kosmoliaptsis, Tom H Karlsen, Andre Franke, and John D Rioux. 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.*, 47(2):172–179, February 2015.

[68] Isabelle Cleynen, Gabrielle Boucher, Luke Jostins, L Philip Schumm, Sebastian Zeissig, Tariq Ahmad, Vibeke Andersen, Jane M Andrews, Vito Annese, Stephan Brand, Steven R Brant, Judy H Cho, Mark J Daly, Marla Dubinsky, Richard H Duerr, Lynnette R Ferguson, Andre Franke, Richard B Gearry, Philippe Goyette, Hakon Hakonarson, Jonas Halfvarson, Johannes R Hov, Hailang Huang, Nicholas A Kennedy, Limas Kupcinskas, Ian C Lawrance, James C Lee, Jack Satsangi, Stephan Schreiber, Emilie Théâtre, Andrea E van der Meulen-de Jong, Rinse K Weersma, David C Wilson, Miles Parkes, Severine Vermeire, John D Rioux, John Mansfield, Mark S Silverberg, Graham Radford-Smith, Dermot P B McGovern, Jeffrey C Barrett, and Charlie W Lees. Inherited determinants of crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet*, 387(10014):156–167, January 2016.

[69] Biljana Klimenta, Hilada Nefic, Nenad Prodanovic, Radivoj Jadric, and Fatima Hukic. Association of biomarkers of inflammation and HLA-DRB1 gene locus with risk of developing rheumatoid arthritis in females. *Rheumatol. Int.*, 39(12):2147–2157, December 2019.

[70] 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.

[71] Tianju Wang, Chunmei Shen, Hengxin Li, Liping Chen, Sheng Liu, and Jun Qi. High resolution HLA-DRB1 analysis and shared molecular amino acid signature of DR$\beta$1 molecules in occult hepatitis B infection. *BMC Immunol.*, 23(1):22, April 2022.

[72] 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.

[73] 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.

[74] Andrew D Grotzinger, Mijke Rhemtulla, Ronald de Vlaming, Stuart J Ritchie, Travis T Mallard, W David Hill, Hill F Ip, Riccardo E Marioni, Andrew M McIntosh, Ian J Deary, Philipp D Koellinger, K Paige Harden, Michel G Nivard, and Elliot M Tucker-Drob. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.*, 3(5):513–525, May 2019.

[75] Jennifer Sam Beaty and M Shashidharan. Anal fissure. *Clin. Colon Rectal Surg.*, 29(1):30–37, March 2016.

[76] Douglas W Mapel, Michael Schum, and Ann Von Worley. The epidemiology and treatment of anal fissures in a population-based cohort. *BMC Gastroenterol.*, 14(1):129, July 2014.

[77] Michael R B Keighley and Norman S Williams, editors. *Keighley & Williams' surgery of the anus, rectum and colon, fourth edition*. CRC Press, London, England, 4 edition, November 2018.

[78] P McDonald, A M Driscoll, and R J Nicholls. The anal dilator in the conservative management of acute anal fissures. *Br. J. Surg.*, 70(1):25–26, January 1983.

[79] M R Lock and J P Thomson. Fissure-in-ano: the initial management and prognosis. *Br. J. Surg.*, 64(5):355–358, May 1977.

[80] NICE NICE. Scenario: Management of an anal fissure, Apr 2021.

[81] Erica B Sneider and Justin A Maykel. Anal abscess and fistula. *Gastroenterol. Clin. North Am.*, 42(4):773–784, December 2013.

[82] Zubing Mei, Qingming Wang, Yi Zhang, Peng Liu, Maojun Ge, Peixin Du, Wei Yang, and Yazhou He. Risk factors for recurrence after anal fistula surgery: A meta-analysis. *Int. J. Surg.*, 69:153–164, September 2019.

[83] Carlo Zanotti, Carmen Martinez-Puente, Isabel Pascual, María Pascual, Dolores Herreros, and Damián García-Olmo. An assessment of the incidence of fistula-in-ano in four countries of the european union. *Int. J. Colorectal Dis.*, 22(12):1459–1462, December 2007.

[84] Chiara Eberspacher, Domenico Mascagni, Iulia Catalina Ferent, Enrico Coletta, Rossella Palma, Cristina Panetta, Anna Esposito, Stefano Arcieri, and Stefano Pontone. Mesenchymal stem cells for cryptoglandular anal fistula: Current state of art. *Front. Surg.*, 9:815504, February 2022.

[85] A G Parks. Pathogenesis and treatment of fistula-in-ano. *BMJ*, 1(5224):463–460, February 1961.

[86] Marcin Włodarczyk, Jakub Włodarczyk, Aleksandra Sobolewska-Włodarczyk, Radzisław Trzciński, Łukasz Dziki, and Jakub Fichna. Current concepts in the pathogenesis of cryptoglandular perianal fistula. *J. Int. Med. Res.*, 49(2):300060520986669, February 2021.

[87] Haig Dudukgian and Herand Abcarian. Why do we have so much trouble treating anal fistula? *World J. Gastroenterol.*, 17(28):3292–3296, July 2011.

[88] M J Johnston, G M Robertson, and F A Frizelle. Management of late complications of pelvic radiation in the rectum and anus. *Dis. Colon Rectum*, 46(2):247–259, February 2003.

[89] Roland Assi, Peter W Hashim, Vikram B Reddy, Hulda Einarsdottir, and Walter E Longo. Sexually transmitted infections of the anus and rectum. *World J. Gastroenterol.*, 20(41):15262–15268, November 2014.

[90] Jonathan Alastair Simpson, Ayan Banerjea, and John Howard Scholefield. Management of anal fistula. *BMJ*, 345(oct15 4):e6705, October 2012.

[91] Antonio Arroyo, Juan Pérez-Legaz, Pedro Moya, Laura Armañanzas, Javier Lacueva, Francisco Pérez-Vicente, Fernando Candela, and Rafael Calpena. Fistulotomy and sphincter reconstruction in the treatment of complex fistula-in-ano: long-term clinical and manometric results. *Ann. Surg.*, 255(5):935–939, May 2012.

[92] Bhupendra Kumar Jain, Kumar Vaibhaw, Pankaj Kumar Garg, Sanjay Gupta, and Debajyoti Mohanty. Comparison of a fistulectomy and a fistulotomy with marsupialization in the management of a simple anal fistula: a randomized, controlled pilot trial. *J. Korean Soc. Coloproctol.*, 28(2):78–82, April 2012.

[93] UK Biobank. Ukbb data showcase, 2023. Accessed on 20/10/2023.

[94] UK Biobank. Integrating electronic health records into the uk biobank resource version 1.0, January 2014. Accessed on 20/10/2023.

[95] FinnGen. Finngen project, 2023. Accessed on 20/10/2023.

[96] FinnGen. Finngen data freeze 9, 2023. Accessed on 20/10/2023.

[97] Centers for disease control and prevention. International classification of diseases, (icd-10-cm/pcs) transition - background, November 2015. Accessed on 20/10/2023.

[98] FinnGen. Data available in finngen, 2023. Accessed on 20/10/2023.

[99] Thermo Fisher Scientific. Axiom™ genotyping solution data analysis user guide, 18 April 2023. Accessed on 20/10/2023.

[100] Affymetrix. Uk biobank 500k samples genotyping data generation by the affymetrix research services laboratory, April 2015. Accessed on 20/10/2023.

[101] UK Biobank. Description of sample processing workflow and preparation of dna for genotyping, April 2015. Accessed on 20/10/2023.

[102] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, October 2018.

[103] Gavin Band and Jonathan Marchini. BGEN: a binary file format for imputed genotype and haplotype data. April 2018.

[104] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, September 2007.

[105] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[106] Brendan K Bulik-Sullivan, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47(3):291–295, March 2015.

[107] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Laramie Duncan, John R B Perry, Nick Patterson, Elise B Robinson, Mark J Daly, Alkes L Price, and Benjamin M Neale. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, 47(11):1236–1241, November 2015.

[108] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, October 2007.

[109] Tenghao Zheng, David Ellinghaus, Simonas Juzenas, François Cossais, Greta Burmeister, Gabriele Mayr, Isabella Friis Jørgensen, Maris Teder-Laving, Anne Heidi Skogholt, Sisi Chen, Peter R Strege, Go Ito, Karina Banasik, Thomas Becker, Frank Bokelmann, Søren Brunak, Stephan Buch, Hartmut Clausnitzer, Christian Datz, DBDS Consortium, Frauke Degenhardt, Marek Doniec, Christian Erikstrup, Tõnu Esko, Michael Forster, Norbert Frey, Lars G Fritsche, Maiken Elvestad Gabrielsen, Tobias Gräßle, Andrea Gsur, Justus Gross, Jochen Hampe, Alexander Hendricks, Sebastian Hinz, Kristian Hveem, Johannes Jongen, Ralf Junker, Tom Hemming Karlsen, Georg Hemmrich-Stanisak, Wolfgang Kruis, Juozas Kupcinskas, Tilman Laubert, Philip C Rosenstiel, Christoph Röcken, Matthias Laudes, Fabian H Leendertz, Wolfgang Lieb, Verena Limperger, Nikolaos Margetis, Kerstin Mätz-Rensing, Christopher Georg

Németh, Eivind Ness-Jensen, Ulrike Nowak-Göttl, Anita Pandit, Ole Birger Pedersen, Hans Günter Peleikis, Kenneth Peuker, Cristina Leal Rodriguez, Malte Christoph Rühlemann, Bodo Schniewind, Martin Schulzky, Jurgita Skieceviciene, Jürgen Tepel, Laurent Thomas, Florian Uellendahl-Werth, Henrik Ullum, Ilka Vogel, Henry Volzke, Lorenzo von Fersen, Witigo von Schönfels, Brett Vanderwerff, Julia Wilking, Michael Wittig, Sebastian Zeissig, Myrko Zobel, Matthew Zawistowski, Vladimir Vacic, Olga Sazonova, Elizabeth S Noblin, 23andMe Research Team, Gianrico Farrugia, Arthur Beyder, Thilo Wedel, Volker Kahlke, Clemens Schafmayer, Mauro D'Amato, and Andre Franke. Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. *Gut*, 70(8):1538–1549, April 2021.

[110] American Society of Colon and Rectal Surgery. Abscess and fistula, 2023. Accessed on 21/10/2023.

[111] Manuela Marzo, Carla Felice, Daniela Pugliese, Gianluca Andrisani, Giammarco Mocci, Alessandro Armuzzi, and Luisa Guidi. Management of perianal fistulas in crohn's disease: an up-to-date review. *World J. Gastroenterol.*, 21(5):1394–1403, February 2015.

[112] American Society of Colon and Rectal Surgery. Diverticular disease expanded information, 2023. Accessed on 21/10/2023.

[113] Jacklyn N Hellwege, Jacob M Keaton, Ayush Giri, Xiaoyi Gao, Digna R Velez Edwards, and Todd L Edwards. Population stratification in genetic association studies. *Curr. Protoc. Hum. Genet.*, 95(1):1.22.1–1.22.23, October 2017.

[114] Peter Kraft, Eleftheria Zeggini, and John P A Ioannidis. Replication in genome-wide association studies. *Stat. Sci.*, 24(4):561–573, November 2009.

[115] Evangelos Evangelou and John P A Ioannidis. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.*, 14(6):379–389, June 2013.

[116] Center for Statistical Genetics. Metal documentation, December 2017. Accessed on 21/10/2023.

[117] Diogo M Ribeiro, Simone Rubinacci, Anna Ramisch, Robin J Hofmeister, Emmanouil T Dermitzakis, and Olivier Delaneau. The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat. Commun.*, 12(1):4842, August 2021.

[118] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, April 2002.

[119] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, Peter S Linsley, Mao Mao, Roland B Stoughton, and Stephen H Friend. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, March 2003.

[120] Jianan Tian, Mark P Keller, Aimee Teo Broman, Christina Kendziorski, Brian S Yandell, Alan D Attie, and Karl W Broman. The dissection of expression quantitative trait locus hotspots. *Genetics*, 202(4):1563–1574, April 2016.

[121] Sipko van Dam, Urmo Võsa, Adriaan van der Graaf, Lude Franke, and João Pedro de Magalhães. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, page bbw139, January 2017.

[122] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, Alexandra Zhernakova, Daria V Zhernakova, Jan H Veldink, Leonard H Van den Berg, Juha Karjalainen, Sebo Withoff, André G Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A C 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B Singleton, Dena G Hernandez, Michael A Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A Gharib, Daniel A Enquobahrie, Thomas Lumley, Grant W Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C Jansen, Peter M Visscher, Julian C Knight, Bruce M Psaty, Samuli Ripatti, Alexander Teumer, Timothy M Frayling, Andres Metspalu, Joyce B J van Meurs, and Lude Franke. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.*, 45(10):1238–1243, October 2013.

[123] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Res.*, 50(D1):D687–D692, January 2022.

[124] Paul D Thomas, Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.*, 31(1):8–22, January 2022.

[125] Mingchu Xu, Yajing Angela Xie, Hana Abouzeid, Christopher T Gordon, Alessia Fiorentino, Zixi Sun, Anna Lehman, Ihab S Osman, Rachayata Dharmat, Rosa Riveiro-Alvarez, Linda Bapst-Wicht, Darwin Babino, Gavin Arno, Virginia Busetto, Li Zhao, Hui Li, Miguel A Lopez-Martinez, Liliana F Azevedo, Laurence Hubert, Nikolas Pontikos, Aiden Eblimit, Isabel Lorda-Sanchez, Valeria Kheir, Vincent Plagnol, Myriam Oufadem, Zachry T Soens, Lizhu Yang, Christine Bole-Feysot, Rolph Pfundt, Nathalie Allaman-Pillet, Patrick Nitschké, Michael E Cheetham, Stanislas Lyonnet, Smriti A Agrawal, Huajin Li, Gaëtan Pinton, Michel Michaelides, Claude Besmond, Yumei Li, Zhisheng Yuan, Johannes von Lintig, Andrew R Webster, Hervé Le Hir, Peter Stoilov, UK Inherited Retinal Dystrophy Consortium, Jeanne Amiel, Alison J Hardcastle, Carmen Ayuso, Ruifang Sui, Rui Chen, Rando Allikmets, and Daniel F Schorderet. Mutations in the spliceosome component CWC27 cause retinal degeneration with or

without additional developmental anomalies. *Am. J. Hum. Genet.*, 100(4):592–604, April 2017.

[126] Nao Otomo, Shuji Mizumoto, Hsing-Fang Lu, Kazuki Takeda, Belinda Campos-Xavier, Lauréane Mittaz-Crettol, Long Guo, Kazuharu Takikawa, Masaya Nakamura, Shuhei Yamada, Morio Matsumoto, Kota Watanabe, and Shiro Ikegawa. Identification of novel LFNG mutations in spondylocostal dysostosis. *J. Hum. Genet.*, 64(3):261–264, March 2019.

[127] D B Sparrow, G Chapman, M A Wouters, N V Whittock, S Ellard, D Fatkin, P D Turnpenny, K Kusumi, D Sillence, and S L Dunwoodie. Mutation of the LUNATIC FRINGE gene in humans causes spondylocostal dysostosis with a severe vertebral phenotype. *Am. J. Hum. Genet.*, 78(1):28–37, January 2006.

[128] Katrin Serth, Karin Schuster-Gossler, Ralf Cordes, and Achim Gossler. Transcriptional oscillation of lunatic fringe is essential for somitogenesis. *Genes Dev.*, 17(7):912–925, April 2003.

[129] Marcin Włodarczyk, Jakub Włodarczyk, Aleksandra Sobolewska-Włodarczyk, Radzisław Trzciński, Łukasz Dziki, and Jakub Fichna. Current concepts in the pathogenesis of cryptoglandular perianal fistula. *J. Int. Med. Res.*, 49(2):300060520986669, February 2021.

[130] Eleanor Sanderson, M Maria Glymour, Michael V Holmes, Hyunseung Kang, Jean Morrison, Marcus R Munafò, Tom Palmer, C Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. *Nat. Rev. Methods Primers*, 2(1), February 2022.

[131] Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, Giulio Genovese, Arpiar Saunders, Evan Macosko, Samuela Pollack, John R B Perry, Jason D Buenrostro, Bradley E Bernstein, Soumya Raychaudhuri, Steven McCarroll, Benjamin M Neale, Alkes L Price, and The Brainstorm Consortium. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.*, 50(4):621–629, April 2018.