Chapter 5

A meta-analysis to map loci associated with age at IBD diagnosis

5.1 Introduction

As described in the previous chapter, large GWAS meta-analyses have uncovered a total of 231 genomic signals associated with the risk of IBD, and this has substantially advanced our understanding of the processes implicated in disease development. However, disease risk is only one aspect of disease biology, and the extent to which these (or novel) association signals also influence other aspects of disease, such as disease severity, disease location, response to treatments, or age at disease onset, is still poorly understood. The identification of genes modulating these aspects of disease can be of great importance from a clinical standpoint [275], as it may ultimately have important implications for drug development, diagnosis testing and risk stratification.

5.1.1 The role of genetic variation in the age at IBD diagnosis

Contrary to other examples of complex diseases, such as Alzheimer disease and Parkinson disease [235, 284], the estimated heritability of age of onset of IBD has not yet been quantified. However, family studies have shown that age at disease diagnosis (ADD), an imperfect proxy for age at onset, is highly concordant ($r = 0.69$, $P = 0.0001$) within families [190, 380], suggesting genetic modifiers for IBD age at onset may indeed exist.

The first study aimed at identifying polymorphisms associated with age at diagnosis of CD and UC outside individual genes such as *NOD2*, focused on 332 known IBD-associated SNPs and 329 CD and 294 UC patients, respectively [93]. Using the age at diagnosis as a continuous trait, and by comparing the mean age between genotypes, the authors identified rs2076756 in *NOD2* to be associated with a younger age of onset for CD ($P = 0.0002$): patients with the AA wild-type genotype were diagnosed at 31.9±1.23 years, AG heterozygotes at 25.6± 0.99 years and GG homozygotes at 22.6±1.32 years. In addition, depending on the age subgroups further compared, SNPs in *POU5F1*, *TNFSF15* and *HLA-DRB1*501* were found to be associated with age of Crohn's disease diagnosis, and a variant in *LAMB1* with the age of UC diagnosis.

A much larger study conducted by Cleynen *et al* last year [87], made use of 16,902 CD and 12,597 UC patients genotyped on the Immunochip, a dense custom-design array of 195,806 polymorphisms located in 186 regions with known association with one or more of 12 immune-related diseases, including CD and UC [288, 375]. Apart from *NOD2*, none of the signals identified in the previous study replicated in this analysis, despite all being typed on the Immunochip. As new findings however, two loci (rs3197999 in *MST1* and rs2066847 in *NOD2*) achieved genome-wide significance for the association with age at CD diagnosis, and one SNP (rs3129891) in the major histocompatibility complex (MHC) was found to be associated at genome-wide significance with age at UC diagnosis. Together, these findings confirmed that the general timing of CD and UC onset itself is influenced by genetic variation.

5.2 Aims

The aim of the research presented in this chapter was to build on previous findings of other colleagues, who identified variation in known or immune-related regions to be associated with age at IBD diagnosis, and conduct the first association analysis to date that interrogates the entire genome of ∼5,400 CD and ∼4,400 UC individuals to identify genetic modifiers of age at disease diagnosis.

5.3 Methods

5.3.1 Association analyses

The association analysis for age at disease diagnosis was conducted using the UKIBDGC CD and UC cases for which information on age at disease diagnosis was available (5,403 CD and 4,490 UC individuals). As mentioned in the previous chapter, these samples originally came from three independent GWAS studies (GWAS1, GWAS2 or GWAS3) genotyped on different platforms or from a low-coverage whole-genome sequencing study (IBDSeq, **Table** 5.1). To leverage the whole-genome sequencing data, and thus survey lower frequency variants (1% < MAF < 5%) not well represented in the GWAS arrays, the reference panel containing haplotypes drawn from the low-coverage whole-genome IBD samples (N=4,445), as well as the UK10K (N=3,652) and 1000 Genomes (1KG) Phase 3 control sequences (N=2,505) were imputed into the GWAS cohorts [110, 295].

Studies	CD samples	UC samples
GWAS1	1,116	.
GWAS2	.	1,060
GWA3_CD	2,683	.
GWAS3_UC	.	2,165
IBDSeq_CD	1,604	.
IBDSeq_UC	.	1,265

Table 5.1 UKIBDGC sample breakdown per contributing study. The studies that contributed samples to the UKIBDGC dataset are given. Total of 5,403 CD and 4,490 UC samples.

To test for association between age at disease diagnosis and genetic variation, I carried out separate linear regression analyses within each of the three studies of each trait (CD and UC, **Table** 5.1). I tested all the variants that passed all UKIBDGC quality control procedures [110, 295] after excluding sites with MAF <1% (in UKIBDGC control samples only) and INFO <0.4, as recommended by Marchini *et al* [309]. The MAF threshold of 1% was chosen because the power to detect single-variant associations below this frequency is very low at current sample sizes [295] and because false-positives will be increased below this frequency threshold as imputation does not work as effectively at rare variant sites [309]. The INFO threshold of 0.4, as routinely used in GWAS [286, 309, 507, 539, 540], was chosen to minimise false positive associations arising from high genotype uncertainty post-imputation. **Table** 5.2 lists the total number of variants tested in each study dataset.

Studies	CD SNPs	UC SNPs
GWAS1	8,123,580	.
GWAS2	.	8,113,309
GWA3_CD	8,141,056	.
GWAS3_UC	.	8,140,904
IBDSeq_CD	7,991,854	.
IBDSeq_UC	.	7,955,914

Table 5.2 Number of high-quality SNPs tested in each UKIBDGC study. The studies that contributed samples to the UKIBDGC dataset are listed, along with the number of SNPs tested in each association analysis for age at CD or UC diagnosis. High-quality SNPs were defined as those that passed all UKIBDGC QC procedures, and had MAF >1% and INFO >0.4. For details of QC procedures see De Lange *et al* [110] and Luo *et al* [295].

Because all the UKIBDGC samples were imputed, the probabilistic nature of the genotypes meant the association testing needed to take the uncertainty of the imputed genotypes into account. To do so, I used the regression framework implemented in SNPTEST v2 [309]. This model uses well-established statistical theory for missing data problems, in which an observed data likelihood is used where the contribution of each possible genotype is weighted by its imputation probability. The test was run assuming an additive genetic model [85], where the effect is increased by β-fold for genotype *Aa* (or 1) and by 2β-fold for genotype *AA* (or 2), and contained the first 10 principal components for ancestry to adjust for potential population structure (PCs were calculated and provided by Katie De Lange):

$$E(Y_{\mathrm{i}}) = \mu + \beta_G * G_i + \eta z_i + \varepsilon \tag{5.1}$$

where E(Y_{i}) denotes the phenotypic value for each individual, μ denotes the baseline effect for the non-effect genotype, β_G, denotes the estimated effect due to each copy of the effect allele, G_i denotes the observed genotype for each individual (coded as 0, 1 or 2, according to the number of copies of the effect allele), z is a matrix of covariates and ε is a residual error.

When performing a regression on a continuous rather than in a binary phenotype (i.e. case-control), the quantitative phenotype is generally either standardized or quantile normalyzed to fit a normal distribution [37, 507]. I decided to use the quantile normalization available in SNPTEST v2 in this case, because it was the transformation previously used in the Immunochip study reported by Cleynen *et al* [87], and because I wanted to compare the effect size estimates between the two studies.

5.3.2 Meta-analysis within CD and UC studies

After performing association analysis for each SNP in each study individually, I conducted a meta-analysis to obtain pooled estimates of the effect of each SNP on the age at disease diagnosis across all studies of each trait (CD and UC). I used the fixed-effects methodology implemented in METAL [520], in which the study-specific effect estimates and standard errors derived from the regression analysis of each cohort are combined in an inverse variance-weighted fixed effects meta-analysis, the most powerful and commonly used method for discovering phenotype-associated SNPs [130, 390]. METAL assumes a given allele exerts similar effects across datasets, and calculates the combined allelic effect (B) across all studies at each marker as:

$$B = \frac{\sum_{i=1}^{k} \omega_{\mathrm{i}} \beta_{\mathrm{i}}}{\sum_{i=1}^{k} \omega_{\mathrm{i}}} \tag{5.2}$$

where k is the number of studies, β_{i} is the effect size from study i and ω_{i} represents the inverse of the variance of the estimated allelic effect, which is given by $\mathrm{SE}(\beta_{\mathrm{i}})^2$.

A fundamental principle of meta-analysis is that all studies tested the same hypothesis using near-identical procedures for QC, covariate adjustment and statistical test, for example, all of which were the case here.

5.3.3 Meta-analysis for IBD

The analysis for age-at-disease diagnosis for IBD was conducted by meta-analysing summary statistics from the CD and UC meta-analysis, similar to Cleynen *et al* [87]. This approach is also generally followed because CD and UC have slightly different age distributions (mean CD age: 27 yrs; mean UC age: 36 yrs, **Figure** 5.1), which would look bimodal if the samples were combined.

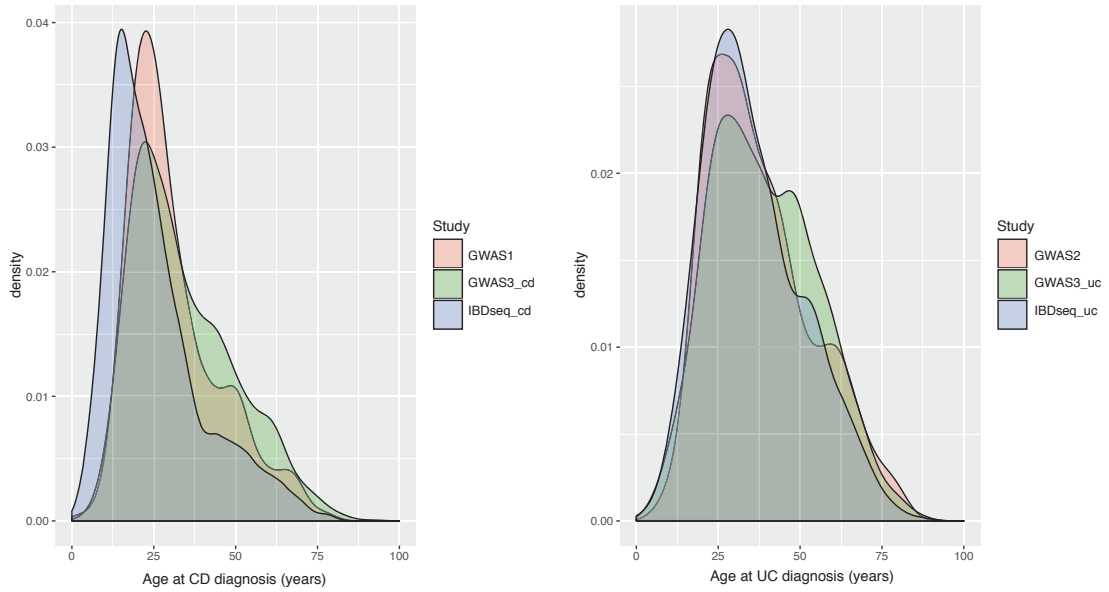

Figure 5.1 Distribution of age at disease diagnosis across the different studies. Prior to association testing, the quantile normalization was performed so that the age distributions within the CD and UC studies were forced to have the same statistical properties (mean and standard deviation), a procedure that is normally conducted when different studies are to be meta-analysed.

5.3.4 Post meta-analysis quality control

To control for between-study/traits heterogeneity in effect sizes, I excluded SNPs for which the I^2 metric was greater than 90%, similarly to what others have done [290]. Briefly, the I^2 measures the degree of inconsistency in the studies' results, and describes the percentage of total variation across the studies that is due to heterogeneity rather than chance [130, 205]. As additional filtering post-meta-analysis, I excluded SNPs that: 1) were present solely in one study/trait out of all that were meta-analysed, 2) the meta-analysis P-value (P_{META}) was greater than the individual studies' P-values and 3) the INFOs of the studies driving the signal (at $\alpha = 0.05$) were <0.6.

5.3.5 Power to detect previous Immunochip signals

Finally, I conducted an analysis to determine the statistical power of my study to detect, at genome-wide significance, the previous genome-wide signals associated with either the age at CD or UC diagnosis reported in the Immunochip study of Cleynen *et al* [87]. Because power is determined by both the frequency and the effect size of the risk allele [22], I calculated the power for each variant separately using the method

derived by Sham and Purcell [443], which assumes the non-centrality parameter (NCP) of the chi-squared distribution for a single SNP under the additive genetic model is:

$$NCP = N * h^2 \tag{5.3}$$

where N is the total number of studied individuals and h^2 is the fraction of phenotypic variance explained by the marker, which I calculated as follows:

$$h^2 = 2p(1-p)\beta^2 \tag{5.4}$$

where p is the frequency of the effect allele assuming Hardy-Weinberg equilibrium and β is its additive effect, defined as the regression coefficient of the linear model [545].

5.4 Results

Association of variants with age at disease diagnosis of CD and UC was tested using linear regression of the quantitative phenotype (**Figure** 5.1), in a total of 5,403 and 4,490 UKIBDGC cases, respectively, each split across three different studies.

Figure 5.2 shows the QQ plot for comparison of the observed and expected P-values distributions for the average of 9 million variants with MAF $> 0.5\%$ that were tested in each of the six studies. All QQ plots demonstrate evidence of genetic associations at the tail of the distribution. Importantly, the QQ plots demonstrate no evidence of population stratification, as none exhibit a global excess of higher observed p-values than expected throughout the distribution, and as measured by the inflation factor ($\lambda \sim 1$ in all studies). The λ value represents the degree of deviation from the expected distribution and was calculated as the ratio of the median association test statistic over the theoretical median test statistic of the χ^2 distribution (0.675^2).

The QQ plots resulting from the meta-analyses combining the effect sizes across the studies within each disease entity (CD, UC and IBD) are illustrated in **Figure** 5.3. No genome-wide significant signals remained after the QC procedure applied post meta-analysis (**Figure** 5.3), however, a total of four signals showed suggestive levels of association (P_{META}-value $\leq 5x10^{-7}$) with either CD, UC or IBD (**Table** 5.3). All of

these signals were driven by common variants with MAF >1% and were present in all meta-analysed studies, therefore being supported by different genotyping platforms. Moreover, all signals also showed consistency in direction and magnitude of effects across all studies within each trait. I will describe these four associations in greater detail below, however these findings should not be taken as definitive, as additional validation in independent and larger studies will be necessary. Approximately 73% of the associations with borderline significance are successfully replicated when additional data are acquired [370], therefore some of the signals I report here likely contain true associations that may be replicated in future analyses.

Disease	Signal	Locus	REF/EA	META			I^2	INFOs	EAF*	rsID	Genes in region
				Direction	β (SE)	P-value					
CD	1	2:28606778	C/T	---	-0.10 (0.01)	$1.89x10^{-7}$	47.8	0.96; 0.96; 0.99	0.450	rs2879179	*FOSL2* (intron), *BRE, PLB1, PPP1CB* (+8)
		2:28608504	C/T	---	-0.10 (0.01)	$1.95x10^{-7}$	55.2	0.96;0.96;0.99	0.448	rs4666067	
		2:28612213	G/C	---	-0.09 (0.01)	$4.16x10^{-7}$	5.4	0.96; 0.97; 0.99	0.493	rs1509396	
		2:28623047	T/C	---	-0.09 (0.01)	$3.21x10^{-7}$	30.4	0.96; 0.99; 0.99	0.476	rs4617998	
UC	2	1:245581534	C/T	+++	0.10 (0.02)	$3.60x10^{-7}$	47.4	0.97; 0.97; 0.99	0.501	rs1148919	*KIF26B* (intronic), *SMYD3, EFCAB2* (+1)
	3	22:40382249	T/C	+++	0.12 (0.02)	$2.23x10^{-7}$	0	0.90; 0.91; 0.98	0.295	rs2958654	*FAM83F, GRAP2, ENTHD1, TNRC6B* (+9)
		22:40389007	T/G	+++	0.12 (0.02)	$2.24x10^{-7}$	0	0.89; 0.92; 0.99	0.285	rs2958658	
		22:40390238	G/A	+++	0.12 (0.02)	$2.82x10^{-7}$	0	0.89; 0.92; 0.99	0.295	rs28607928	
IBD	4	20:29904377	G/A	++	0.11 (0.02)	$1.02x10^{-7}$	0	0.77; 0.79; 0.91 0.84; 0.82; 0.85	0.165	rs6141273	*DEFB115, DEFB119, DEFB116* (+17)

Table 5.3 Genetic loci associated at suggestive significance (P_{META}-value $\leq 5x10^{-7}$) with age at CD, UC or IBD diagnosis. REF: reference allele; EA: effect allele; Direction denotes either the positive (+) or negative (-) effect of the effect allele on the phenotype and it includes the direction of the effect in the three independent studies (ordered by GWAS1, GWAS3, IBDSeq for CD and GWAS2, GWAS3 and IBDSeq for UC) or, in the case of IBD, in the two traits (CD and UC); SE: standard error around the beta estimate; EAF: effect allele frequency calculated from the largest control group (GWAS3, N=9,454 individuals). I^2 measures the degree of inconsistency in the studies' results. INFOs correspond to the INFO of the individual studies that were meta-analysed (studies ordered similarly as above for CD and UC; for IBD I give the INFOs of all CD and UC studies). Location given by Ensembl VEP v75. Table is sorted by genomic location within each disease entity. All variants represent common variants and all show consistent direction of effects across studies/traits.

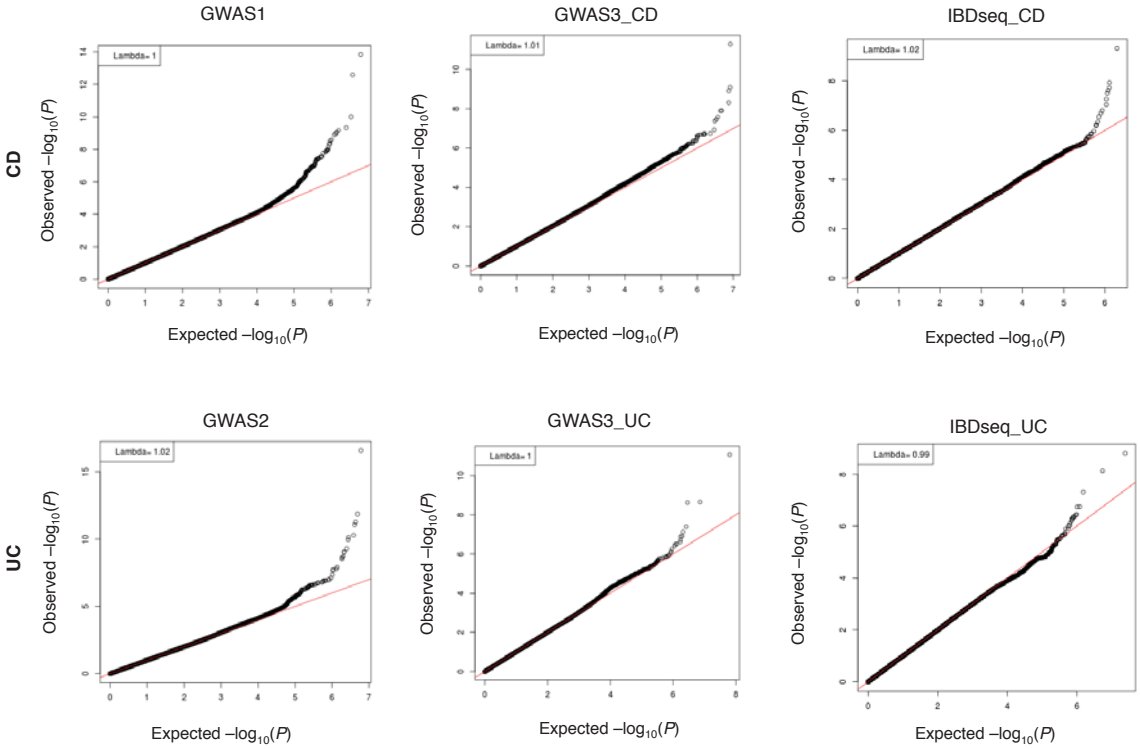

Figure 5.2 Quantile-quantile plots of the individual CD and UC association studies. The red line shows the distribution under the null hypothesis, where the observed p-values correspond exactly to the expected p-values. The inflation at the end of the tail reveals there is evidence of genetic associations. There is no evidence of inflation caused by population stratification, as all lambda values (λ) are close to 1 in all studies. Variants included in the association tests and the QQ plots are those that passed all UKIBDGC QC procedures (see [110, 295]), and had MAF >0.5% (derived from controls of each study) and INFO >0.4.

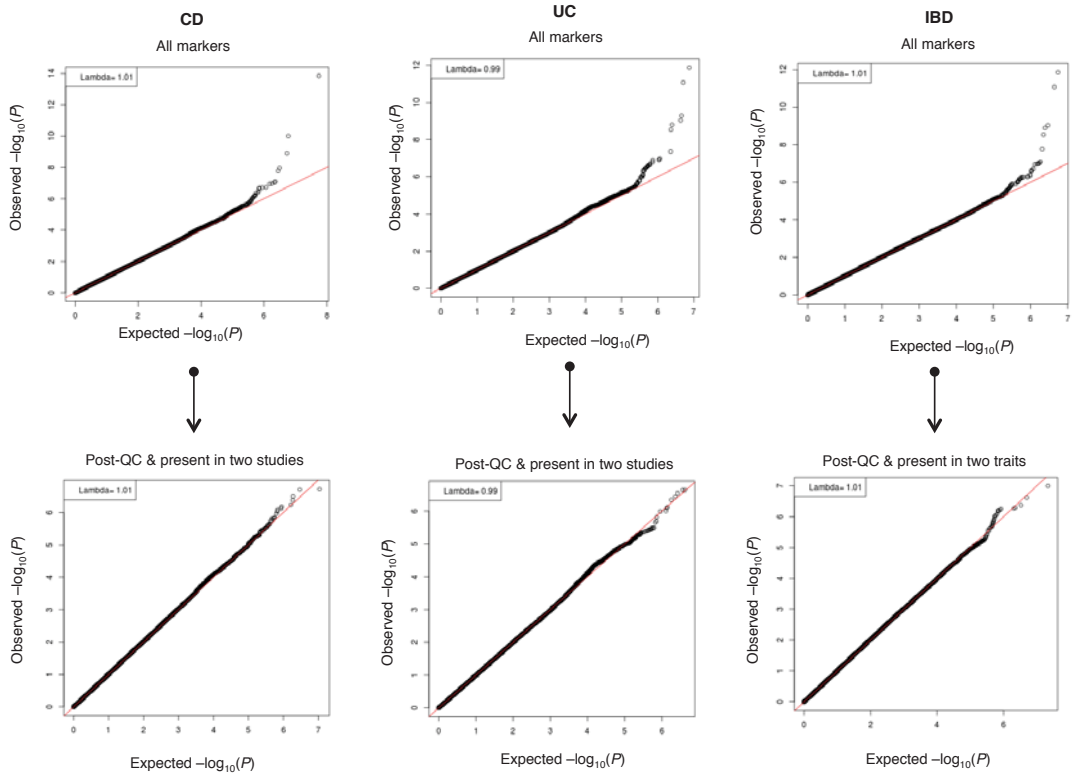

Figure 5.3 Quantile-quantile plots of the meta-analysis results for CD, UC and IBD. The red line shows the distribution under the null hypothesis, where the observed p-values correspond exactly to the expected p-values. The inflation at the end of the tail reveals there is evidence of genetic associations. There is no evidence of inflation caused by population stratification, as all lambda values (λ) are close to 1 in all studies. Variants included in QQ plots are those that passed all UKIBDGC QC procedures (see [110, 295]), had MAF >0.5% (derived from controls of each study), imputation INFO >0.4 and displayed between-study heterogeneity in effect sizes (I^2) below 90%. Note in the top panel, the significance of each marker is not necessarily supported by all the meta-analysed datasets, i.e. the p-value might be driven by a single dataset. Bottom panel illustrates the markers that are present, post meta-analysis QC, in two of the three meta-analysed studies (in the case of CD and UC) and by both traits (CD and UC) in IBD. There is no evidence of genome-wide significant signals.

5.4.1 Suggestive association for the age at CD diagnosis

Three common frequency (EAF=45%) intronic variants in *FOSL2* were associated at suggestive levels of significance with age at CD diagnosis. The lead SNP driving this signal (rs2879179, $P_{META\_CD} = 1.89x10^{-7}$, **Table** 5.3) was associated with a decrease in the age at CD diagnosis with a per-allele effect beta of -0.10 (SE=0.01). The regional plot for this signal (**Figure** 5.4), including all the SNPs within 500kb on either side of this variant, revealed multiple SNPs with varying degrees of association due to local LD patterns, which decrease the chance that genotyping artifacts are driving this suggestive association. In addition, the genotyping clusters for this SNP in all UKIBDGC individuals were well defined (**Figure** 5.5), which again argues against poor genotyping at this SNP.

Interestingly, the *FOSL2* locus has been previously reported by Jostins *et al* [232] and Liu *et al* [290] to be associated with the risk of IBD via rs925255, a SNP in high LD ($r^2 = 0.71$) with rs2879179. Both studies reported P-values of $2.67x10^{-15}$, and $1.07x10^{-16}$ for rs925255, respectively, and ORs of 1.09 (CI: 1.09 - 1.16) and 1.11 (CI: 1.09 - 1.12). The current UKIBDGC-GWAS analysis [110] replicated that signal and identified another lead SNP (rs11677002) in perfect LD ($r^2 = 1$) with that of Jostins and Liu for that IBD-risk association (**Figure** 5.4), which is actually stronger in CD (β=-0.14, SE=0.02) than in UC (β=-0.07, SE=0.02). More interestingly, this latter study also showed that rs2879179, here associated with the age at CD diagnosis, is also associated with the risk of developing IBD ($P = 2.8x10^{-9}$), again with a stronger effect on CD ($P = 2.2x10^{-12}$, **Figure** 5.4). This cross-phenotype association at the same locus is intriguing and is reminiscent of what is known for *NOD2*, which has the largest effect in susceptibility for CD while also being associated with an earlier age of CD onset (rs2066847, p.L1007fsX, β= -0.17, $P = 2.04x10^{-16}$) [87].

To contrast my *FOSL2* finding with the previous Immunochip ADD analysis, I inspected whether this locus showed nominal significance (at $\alpha = 0.05$) in the summary results kindly provided by Dr Isabel Cleynen. rs2879179 was not directly typed in the Immunochip. In fact, this whole locus was not densely represented on the chip because its association with IBD-risk was unknown at the time of design, meaning it was not included in the fine-mapping regions that were typed on the chip. Still, subsequent inspection for possible proxies of rs2879179, revealed one marker with an r^2 >0.7, and the SNP showed nominal significance (P = 0.03). The evidence of replication is perhaps not as strong as we may expect given the LD between the two variants, however the

poor representation of SNPs in higher LD with my SNP in the Immunochip prevents me to make further comparisons.

FOSL2 is part of a family of transcription factors composed of three other members (*FOSL1*, *FOSB*, *FOS*) that together form the AP-1 (activator protein-1) transcription factor complex. Amongst a plethora of functions, such as regulation of cell proliferation, death, survival and differentiation [444], AP-1 has been shown to be a positive regulator of inflammation, containing transcriptional regulator binding sites for numerous inflammatory mediators (IL6, IL8, TNF-a), and capable of binding to promoters independently of NF-kB [510].

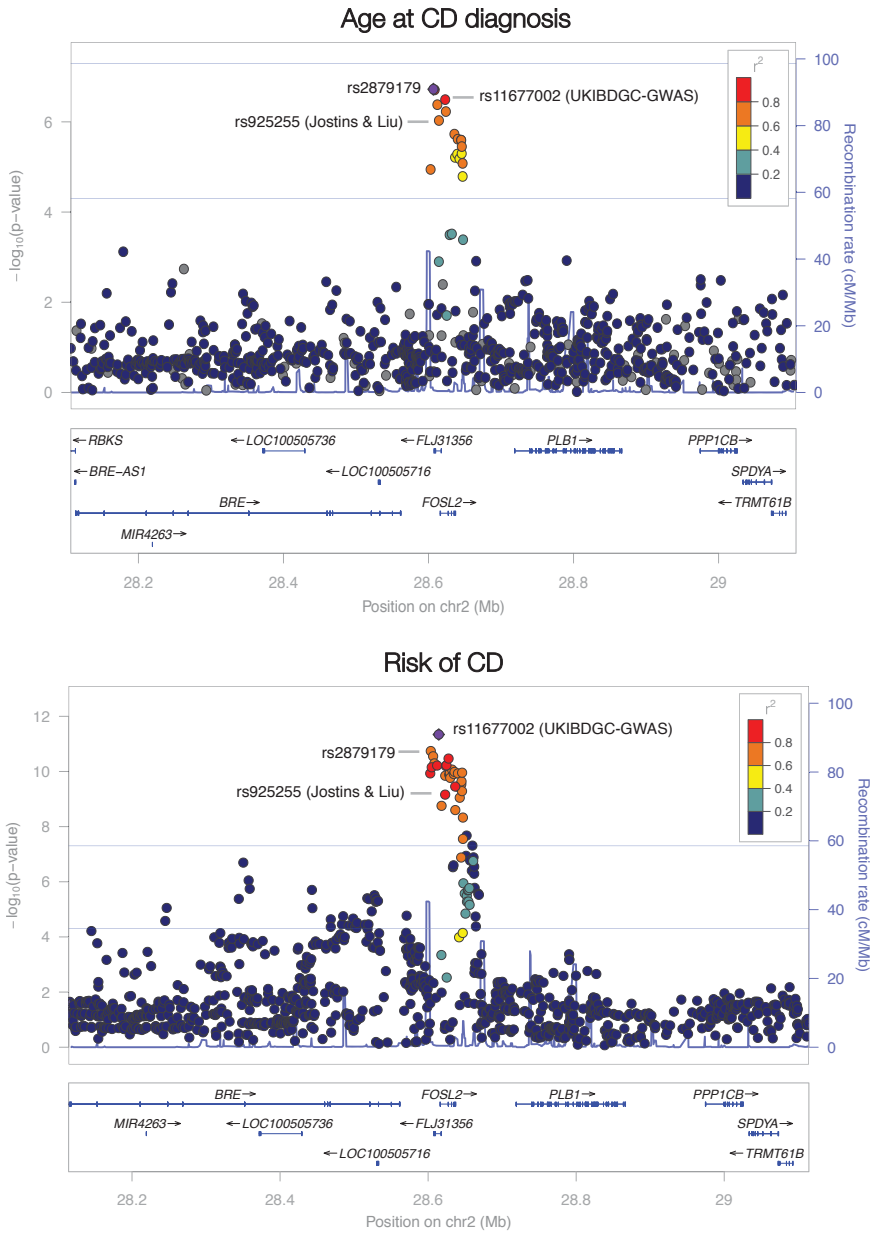

Figure 5.4 A) Regional association plot for 2p28, including the best SNP (rs2879179) for age at CD diagnosis (GWAS3_CD dataset). Plot also illustrates the SNPs of this locus that were previously reported to be associated with the risk of IBD in the analyses of Jostins Liu (rs925255) and UKIBDGC-GWAS (rs1167702). **B)** Regional association plot for 2p28, but for the SNPs associated with risk of CD in the UKIBDGC-GWAS analysis (GWAS3_CD dataset, data provided by Dr Loukas Moutsianas). Plot also shows where my SNP (rs2879179) and Jostins Liu (rs925255) lie in this associated signal. The lead SNP for age at CD diagnosis is also associated ($P = 2.2x10^{-12}$) with susceptibility to CD. The −log10 P-values for the associated SNPs are shown on the upper part of each plot. SNPs are colored based on their r^2 with the labeled lead SNP (purple symbol), which has the smallest P-value in the region. r^2 was calculated from the 1KG Phase 3 European panel. The bottom section of each plot shows the fine scale recombination rates estimated from individuals in the HapMap population, and genes are marked by horizontal blue lines. Genes within the recombination region of the hit SNPs are labeled. Figures were generated using LocusZoom [404].

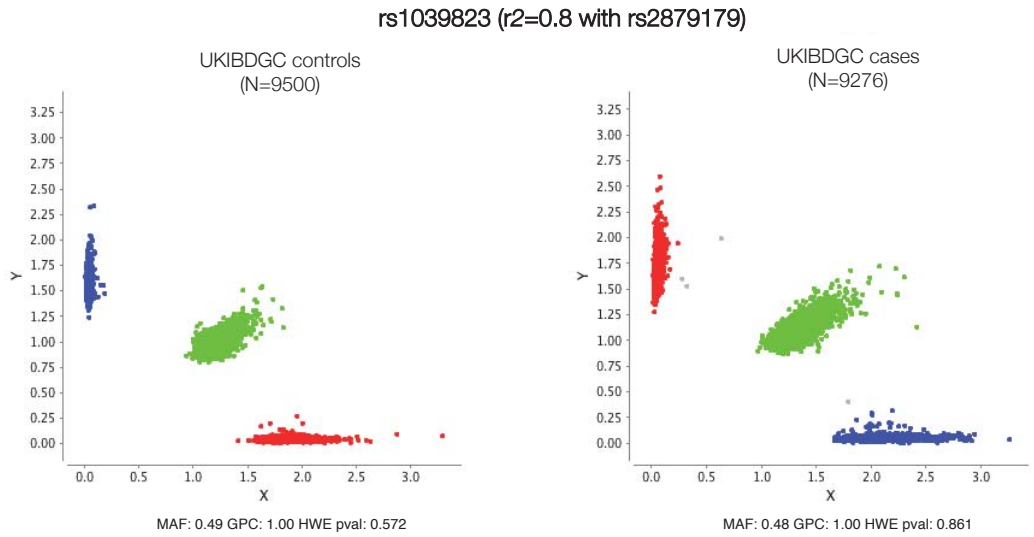

Figure 5.5 Genotype cluster plot for a directly genotyped proxy of rs2879179. The plots represent the raw intensity data from the probes used during genotyping for each UKIBDGC individual. Because rs2879179 was imputed, a proxy (rs1039823) in high LD (r^2=0.8) was chosen for plotting. The plot demonstrates genotypes are of high quality, with genotypes of the same class clustering together and with clusters consistent across UKIBDGC case and control groups. Plot generated by Daniel Rice using Evoker [337].

5.4.2 Suggestive associations for the age at UC diagnosis

Two signals driven by rs2958654 and rs1148919, respectively, were associated at suggestive significance with an increase in the age at which UC presents (**Table** 5.3). Both SNPs had high INFO scores (>0.89) in all meta-analysed datasets, and rs2958654 showed no evidence of heterogeneity between studies (I^2=0, **Table** 5.3). The genes spanning the two associated regions are illustrated in **Figure** 5.6. Again, the regional association plot demonstrates multiple correlated markers with comparable evidence of association, suggesting the signals are less likely to represent type-I errors. rs2958654 and all its proxies were imputed SNPs hence cluster plots could not be generated. The cluster plot for a proxy of rs1148919, with r^2=0.86, showed well defined genotypes (data not shown).

While the *FOSL2* signal described above overlaps with a gene with strong biological candidacy, the relevance of the loci located within these two associated regions is unclear. The closest gene to rs2958654 encodes a protein of unknown function (FAM83F) whereas rs1148919 is located in an intronic sequence of *KIF26B*, an intracellular motor protein involved in microtubule-based processes [197].

The two SNPs identified herein were not directly typed in the Immunochip study of Cleynen *et al* [87], nor were proxies with sufficient and informative LD ($r^2 > 0.1$), which precludes comparisons between the two studies.

5.4.3 Suggestive association for the age at IBD diagnosis

The search for genetic determinants for age at IBD diagnosis was conducted by meta-analysing the results from the CD and the UC meta-analyses, similar to Cleynen *et al* [87]. This approach yielded one common, imputed variant (rs6141273) with suggestive association for an increase in the age at IBD diagnosis (**Table** 5.3, β=0.11, SE=0.02, $P_{META\_CD} = 6.7x10^{-6}$ and $P_{META\_UC} = 3.0x10^{-3}$).

rs6141273 is located in a region near the centromere of chromosome 20, where a cluster of evolutionarily conserved β-defensins lie (**Figure** 5.7). These proteins are produced at a variety of epithelial surfaces, including the intestinal mucosa, and are predominantly considered to act as antimicrobial peptides that activate the NF-kB pro-inflammatory pathway [411].

Similarly as above, the associated SNP identified herein was not directly typed in the Immunochip study of Cleynen *et al* [87], nor were proxies with sufficient and informative LD ($r^2 > 0.1$), which precludes comparisons between the two studies.

Figure 5.6 Regional association plots for the common frequency signals with suggestive association with age at UC diagnosis (GWAS3_CD dataset). **A)** rs2958654. **B)** rs1148919. The −log10 P-values for the associated SNPs are shown on the upper part of each plot. SNPs are colored based on their r^2 with the labeled lead SNP (purple symbol), which has the smallest P-value in the region. r^2 was calculated from the 1KG Phase 3 European panel. The bottom section of each plot shows the fine scale recombination rates estimated from individuals in the HapMap population, and genes are marked by horizontal blue lines. Genes within the recombination region of the hit SNPs are labeled. Figures were generated using LocusZoom [404].

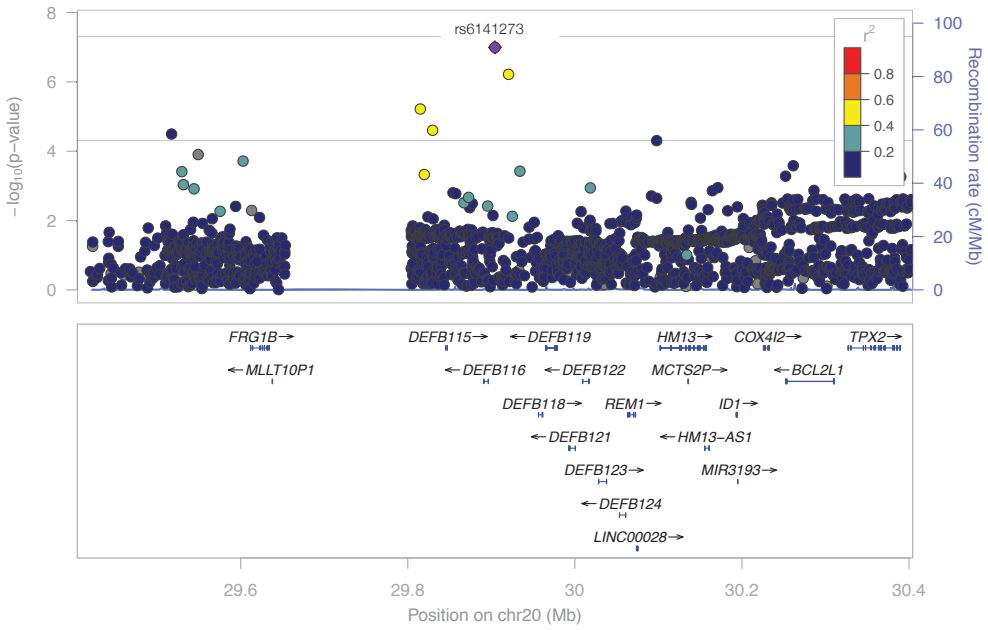

Figure 5.7 Regional association plots for the common frequency signal (rs6141273) with suggestive association with age at IBD diagnosis (GWAS3_CD dataset). The −log10 P-values for the associated SNPs are shown on the upper part of each plot. SNPs are colored based on their r^2 with the labeled lead SNP (purple symbol), which has the smallest P-value in the region. r^2 was calculated from the 1KG Phase 3 European panel. The bottom section of each plot shows the fine scale recombination rates estimated from individuals in the HapMap population, and genes are marked by horizontal blue lines. Genes within the recombination region of the hit SNPs are labeled. Figures were generated using LocusZoom [404].

5.4.4 Comparison with the previous ADD Immunochip study

As mentioned in the introduction to this chapter, three loci have been reported to be associated, at genome-wide significance, with either the age of CD (*NOD2* and *MST1*) or UC (*MHC*) diagnosis. These associations were uncovered in a well-powered Immunochip-based GWAS study comprised of 16,902 CD and 13,597 UC patients [87]. As none of these regions featured in my list of suggestive signals, I hypothesised this could potentially be attributable to the much smaller sample size available here (CD: 5,403; UC 4,490), something that would necessarily hinder the statistical power of this study, i.e. the probability of rejecting the null hypothesis when the alternative hypothesis is true [443].

According to my power calculations, the UC meta-analysis conducted here (N=4,490) was underpowered to detect the *MHC* association with age at UC diagnosis at genome-

wide significance (power = $2.3\text{x}10^{-7}$). The same was true for my CD meta-analysis (N=5,403), which had only 0.3% and 1.5% power to detect an association, at an $\alpha = 5\text{x}10^{-8}$, with *NOD2* and *MST1*, respectively (**Table** 5.4). Out of these two loci, only *NOD2* achieved nominal significance ($P_{META} = 2.08\text{x}10^{-4}$), whereas *MST1* did not. A closer look at the *NOD2* signal in my data, which showed no significant evidence of heterogeneity of effect across the studies ($I^2 = 0$), revealed my point estimate of the effect size was consistent with the previous finding, as it fell within the 95% confidence intervals reported in the more highly powered study (**Figure** 5.8). The reason why *MST1* did not show nominal significance is unclear, however the associated alleles for this region, as well as for *MHC*, showed the same direction of effect as previously reported.

			Immunochip data					Current study		
Disease	rsID	Locus	Effect allele	EAF	P-value	β (SE)	h^2	P-value	β (SE)	POWER
CD	rs3197999	3:49721532 *MST1*	A	0.281	$2.37\text{x}10^{-8}$	-0.07 (0.01)	0.20%	0.097	-0.03 (0.02)	1.5%
CD	rs5743293	16:50763778 *NOD2*	GC	0.024	$2.04\text{x}10^{-16}$	-0.17 (0.02)	0.14%	$2.08\text{x}10^{-4}$	-0.16 (0.04)	0.3%
UC	rs3129891	6:32415080 *MHC*	A	0.209	$1.43\text{x}10^{-8}$	-0.01 (0.02)	0.003%	0.323	-0.02 (0.03)	$2.3\text{x}10^{-7}$

Table 5.4 Power to detect previous loci associated with age at CD and UC diagnosis. Table lists the three loci previous detected at genome-wide significance in Cleynen *et al* [87] to be asociated with either CD or UC age at diagnosis. EAF: effect allele frequency in control samples of my study (GWAS3, N=9,459); SE: standard error of the effect size (β); h^2: phenotypic trait variance explained by the SNP. Power calculated for an $\alpha = 5\text{x}10^{-8}$.

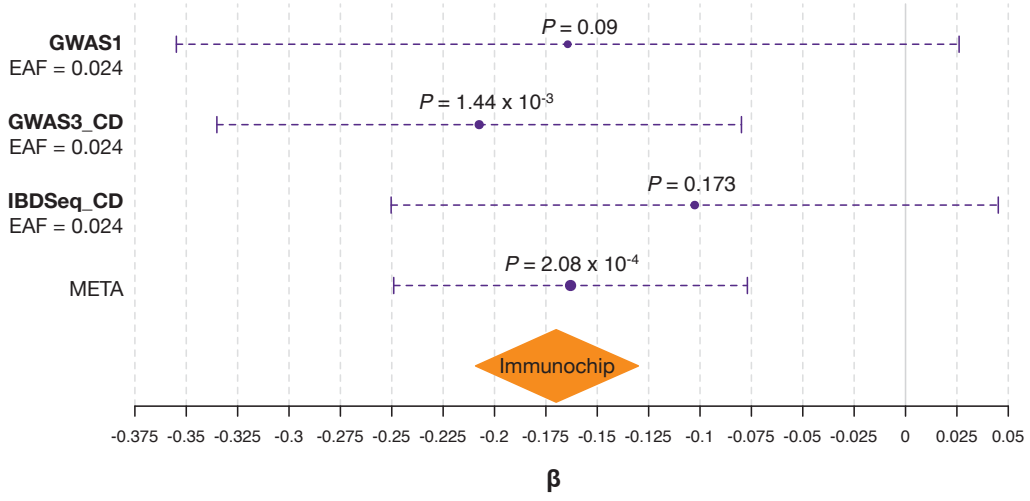

Figure 5.8 Effect size estimations for *NOD2* rs5743293 across the studies.

5.5 Discussion

To identify genetic modifiers for the age of onset of CD, UC and IBD, I conducted three GWAS studies, followed by meta-analyses, using the age at diagnosis reported across a total of 5,403 CD and 4,490 UC UKIBDGC patients. This study is the first to conduct such an analysis in a genome-wide manner, with two previous reports focusing either on 332 known IBD-risk loci [93], or on 186 known immune-associated regions that were included in the Immunochip platform at the time of design [87].

5.5.1 The advantage of imputation

While this study is one order of magnitude smaller than the previous Immunochip analysis [87], which used 16,902 CD and 12,597 UC patients, a much larger set of SNPs were available for testing after the imputation effort conducted by the UKIBDGC (9 million vs. 156,154). This imputation procedure, leveraging ∼10,600 whole-genome sequences drawn from IBD as well as from healthy individuals included in the UK10K and the 1KG projects, also meant that I could examine a much larger frequency space than previous studies, with about ∼40% of the total sites representing low-frequency variants with MAFs between 1% to 5%. In contrast, the Immunochip was designed using the early 1KG Pilot data, which has incomplete coverage particularly of lower frequency variation [1, 375]. The imputation step conducted here therefore clearly demonstrates the value of incorporating genomes of IBD patients and UK population controls and using that information to build a specific reference panel to which independent GWAS samples can be imputed in. The incorporation of UK10K haplotypes in the imputation panel was particularly beneficial, as this resource has been demonstrated to greatly increase the accuracy and coverage of low-frequency and rarer variants compared to existing panels such as the 1KG, because it contains 10-fold more European samples [507].

5.5.2 The pitfall and advantage of my genome-wide analysis

This study was underpowered, at current sample sizes, to identify associations statistically significant at the genome-wide level. This reflects a disadvantage of employing GWAS arrays instead of custom-designed platforms such as the Immunochip, which allow far more individuals to be genotyped since the cost is approximately 20% of that of contemporary GWAS chips [375]. As the number of loci identified strongly

correlates with sample size [1], using genome-wide genotyping platforms in smaller sample cohorts due to cost constraints can ultimately compromise the power of association discovery, as observed here. Despite this, however, my genome-wide analysis did yield three loci with suggestive evidence of association (P_{META}-value $\leq 5x10^{-7}$) that are worth of follow-up in additional replication studies. Importantly, none of the newly associated regions were represented in the Immunochip, which highlights the usefulness of conducting a genome-wide analysis for ADD. The three newly identified signals were of high quality and showed consistent effects across all meta-analysed studies, providing technical validation in multiple independent platforms. As expected, these three associations were driven by common-frequency variants with modest effect sizes (mean=0.10). Unsurprisingly, none of the lead SNPs represented functional variants such as missense or splice disrupting alleles, nor were they in LD with such variants, which is also reminiscent of most GWAS associations [317, 379].

5.5.3 The possible pleiotropy of *FOSL2*

A particularly intriguing result yielded by this analysis is the suggestive association observed for a variant in the *FOSL2* gene and age at CD diagnosis ($P_{META}=1.89x10^{-7}$, β=-0.1). As previously mentioned, *FOSL2* is part of a protein complex (AP-1) [510] which has been shown to upregulate genes involved in immune and pro-inflammatory responses during the pathogenesis of IBD [19, 224, 336]. More specifically, *FOSL2* is a core regulator of plasticity and a repressor of Th17 cells [81], which have emerged as major players in the tissue-specific immune pathology of IBD [162, 193, 232, 332]. Because of this, *FOSL2* has been suggested as an ideal candidate for the development of new therapeutic options aiming to target this Th17 cell population [368].

In addition to its obvious biological candidacy, this locus also showed association with IBD case/control status in three large IBD meta-analyses [110, 232, 290]. This finding is intriguing and opens up the possibility, if successfully replicated in future studies, for a locus to modulate both the risk and the age at which CD presents. A similar mechanism is already known for *NOD2*, which is associated with both the risk [217] and the age of onset of CD symptoms [87]. Another example is for Alzheimer's disease, where the major risk factor, the apolipoprotein E (*APOE*) gene, in addition to affecting the risk of Alzheimer's [100], also has a significant impact on the age at onset, explaining about 10% of its variation [235]. More generally, cross-phenotype associations, sometimes even in seemingly distinct traits [455], have been widely observed, particularly across immune diseases and psychiatric traits. Notable examples include: *IL23R* for IBD [126],

ankylosing spondylitis [132] and psoriasis [165]; *PTPN22* for rheumatoid arthritis [395], CD [33], systemic lupus erythematosus [265] and type 1 diabetes [488]; and *CACNA1C* for bipolar disorder and schizophrenia [453].

Genotyping of the *FOSL2* locus in additional IBD cases with available information on age at diagnosis is currently ongoing. If the observed association with CD-onset is successfully replicated and reaches genome-wide significance, it will be interesting to conduct further regional analysis to try and disentangle the cross-phenotype association seen at this locus. There are several possible scenarios that can underlie the (apparent) pleiotropic genetic effect observed here. One possibility is that *FOSL2* affects both the risk and age-at-onset via the same causal SNP (i.e. allelic pleiotropy). Another hypothesis is that *FOSL2* affects both phenotypes via different and independent causal variants (i.e. genetic pleiotropy). These two possibilities can, in theory, be evaluated through fine-mapping strategies conducted within each phenotype, which would help to refine the associated signals and locate the most likely causal variant (or variants) driving each association [456]. For the case of *FOSL2* however, this is likely to be challenging, because the identified SNPs are in high LD with many others, which will make their effects indistinguishable when conducting conditional analysis, preventing confident fine-mapping. An alternative approach would be to use colocalisation methods such as the one applied by Fortune *et al* [152], which is a Bayesian framework that derives the posterior support for each of five hypotheses describing the possible associations of a given region with two phenotypes. Here, the two hypothesis of greatest interest are: both traits are associated with the region via different causal variants or both traits are associated with the region and share a single causal variant.

An alternative hypothesis for the cross-phenotype association observed here could be mediated pleiotropy. Under that scenario, *FOSL2* could be indirectly associated with the risk of CD via a primary association with age at diagnosis or vice versa, which means the locus would be necessarily associated with both phenotypes if tested separately [455]. To explore this hypothesis, it will be interesting to re-test for an IBD case/control status in the UKIBDGC-GWAS samples while adjusting or stratifying the cases by the age at diagnosis of CD, for example. If the association with IBD-risk persists, then the cross-phenotype association is probably not fully mediated. Alternatively, one can also use another approach which is able to infer whether a given SNP directly influences a given phenotype through a path that does not involve a second correlated trait [501]. When conducting adjustment analyses, it will also be important to evaluate the effect of other sub-phenotypes that may equally affect the observed associations [393]. For

the case of IBD, one could account for information such as disease location at onset (i.e. ileal/colonic), disease behaviour (i.e. penetrating/stricturing/inflammatory) [87] and smoking status, a known environmental modulator [17]. However, such covariates should not be included in discovery efforts, as they can substantially reduce power for the identification of associated variants [393]. Instead, they can be accounted for afterwards, to deconvolve the associated signals. Several examples of sub-phenotype associations driving primary signals exist. For instance, an association of *NOD2* with disease behaviour has been shown to be driven almost entirely by its phenotypic correlation with location and age at diagnosis [87]. Another notable example is *FTO*. This gene was initially discovered to be associated with type 2 diabetes but subsequent correction for body-mass-index (BMI) abolished the signal, suggesting *FTO*-mediated susceptibility to type 2 diabetes was in fact driven through a relationship between *FTO* and obesity [142].