Chapter 4

Uncovering the biological mechanisms driving association

4.1 Introduction

Next-generation sequencing represents a powerful tool for analysing the contribution of rare variation to a range of disorders, and is currently enjoying rapid growth in popularity as we usher in the so-called 'sequencing era'. But does this advance in technology mean the end of genotyping?

For low frequency and common variation, new discoveries are more likely to arise from continuing to increase sample sizes using cost-effective genotyping arrays. Indeed, this approach has proven very successful at identifying genetic risk loci for IBD. To date, 215 associated loci have been uncovered using genome-wide association studies (GWAS) and targeted follow-up using the Immunochip. However, the utility of performing these ever-larger genome-wide association studies in order to identify common variation of relatively small effect sizes has been questioned. In particular, it is notable that just 20 of these 215 IBD-associated loci have been narrowed down to a causal gene, and to date the increased biological understanding from genetic studies has not yet had a substantial impact on disease therapies.

However, recent methodological and technological advances offer the opportunity to derive more therapeutically-relevant information from these genome-wide association studies. This includes novel fine-mapping techniques that can better resolve a given association signal down to a likely causal variant, and improved statistical co-localization methods that can associate a GWAS signal with an expression quantitative trait locus (eQTL) from a variety of cell types and conditions. Such improvements, coupled with rapidly expanding databases of eQTLs and other functional annotations, may prove to be the important missing links required in order to unravel the biological mechanisms underlying many GWAS associations.

4.1.1 Chapter overview

In this chapter, I conduct a new genome-wide association study of inflammatory bowel disease in 18,355 individuals from the United Kingdom. I then meta-analyse these data with the whole genome sequences described in Chapter 3 and published GWAS summary statistics, yielding a total sample size of 59,957 subjects. This leads to the identification of 25 new IBD susceptibility loci, which are then evaluated to try to resolve the potential biological mechanisms underlying each association.

Likely causal missense variants are identified in the genes *SLAMF8*, a negative regulator of inflammation, and *PLCG2*, a gene that has been implicated in primary immune deficiency. A potentially causal variant is also observed in an intron of *NCF4*, which is another gene associated with an immune-related Mendelian disorder. In general, a significant enrichment of genes associated with Mendelian disorders of inflammation and immunity is observed for all 241 IBD-associated loci.

In addition, three novel loci lie proximal to integrin genes, which encode proteins in pathways that have been identified as important therapeutic targets in IBD. Co-localization with eQTL signals confirm that the associated IBD risk-increasing variants are also correlated with expression changes in monocytes in response to immune stimulus at two of these genes (*ITGA4* and *ITGB8*), and at two previously implicated loci (*ITGAL* and *ICAM1*). Overall, we note that new associations at common variants continue to identify genes that are relevant to therapeutic target identification and prioritization.

4.1.2 Contributions

This study was conceived and designed by the UK IBD Genetics Consortium (UKIBDGC), with case ascertainment, phenotyping and sample collection performed by the numerous clinics that contribute to this effort: please see Appendix A for a full list of contributors. DNA sample preparation and genotyping was performed by the Wellcome Trust Sanger Institute pipelines facility. Imputation of GWAS datasets using an IBD-specific reference panel was performed in collaboration with Shane McCarthy; quality control, LD score regression and conditional analysis of the resulting meta-analysis was performed by Loukas Moutsianas. Principal components were generated by Carl Anderson. Overlap with existing eQTL datasets was evaluated by Sun-Gou Ji. Fine-mapping and eQTL co-localization testing was run by Luke Jostins-Dean, but I analysed the output. Disease localisation analysis of variation in *NCF4* was performed by Jeffrey Barrett. Identification of therapeutically-relevant genes and pathways, and evaluation of the biological significance of novel findings was done in discussion with James Lee, Christopher Lamb and Nick Kennedy. Unless stated, I carried out all other analyses.

4.2 Data preparation

4.2.1 A new UK IBD genome wide association study

Sample ascertainment and genotyping

Following ethical approval by Cambridge MREC (reference: 03/5/012), 11,768 British IBD cases, diagnosed using accepted endoscopic, histopathological and radiological criteria, were consented into a new study by the UK IBD Genetics Consortium. These samples consisted of 5,695 Crohn's disease cases, 5,299 ulcerative colitis cases, and 764 inflammatory bowel disease cases of indeterminate type. In parallel, 10,484 controls were obtained by the UK Household Longitudinal Study, which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. Both cases and controls were genotyped at the Wellcome Trust Sanger Institute; controls on the Human Core Exome v12.0 chip, and cases on the Human Core Exome v12.1 chip.

Genotype calling

I called genotypes for this dataset using the software optiCall (Shah et al., 2012), run in five separate batches (four case batches, and a single control batch) to reflect the groupings by which samples were processed in the laboratory. Called genotypes were then strand aligned using files provided by William Rayner (http://www.well.ox.ac.uk/~wrayner/strand/). I removed any sites not included on both versions of the chip, leaving a total of 535,434 genotyped sites.

Sample filtering

Prior to sample quality control, sites were pruned to remove those with a missingness rate in excess of 5%. Individuals failing on one or more of the following filtering criteria were then removed from the dataset:

– Mismatching gender between that listed in the manifest, and that determined genetically. Genders were determined using PLINK v1.9 (Chang et al., 2015), which computes the inbreeding coefficient F based on data from the X chromosome. Under Hardy-Weinberg equilibrium, females should have an X-chromosome F coefficient close to zero, while for males it should be close to one.

– Heterozygosity rate ±3 standard deviations from the mean (Figure 4.1).

– Missingness rate > 1% (Figure 4.1).

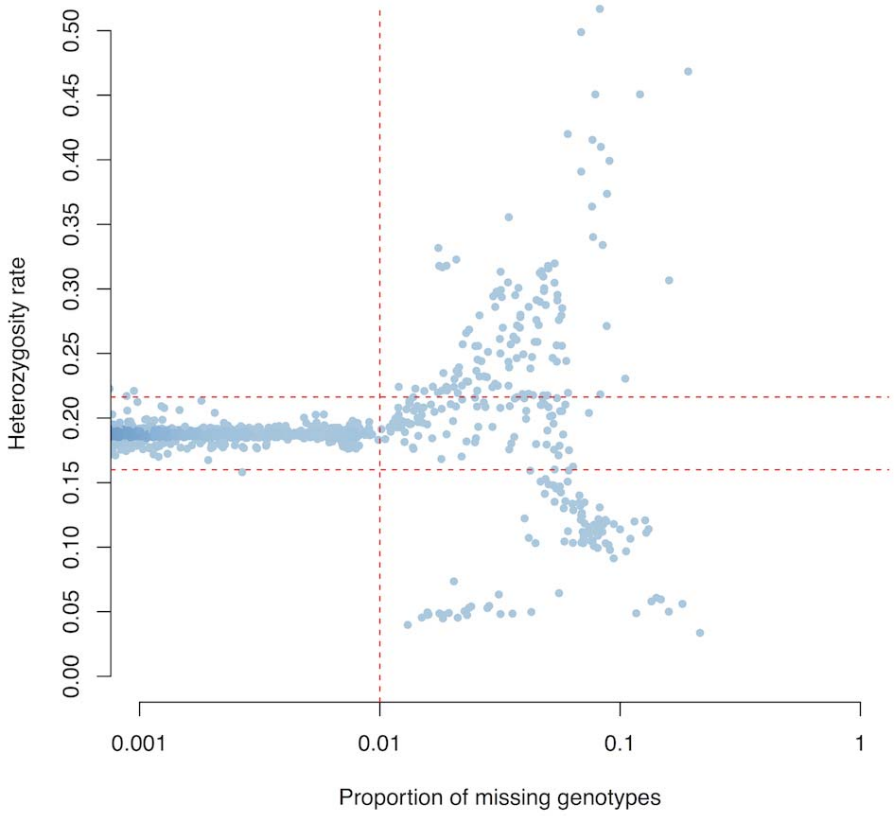

Figure 4.1: Missingness versus heterozygosity rate for samples in the new UK IBD GWAS. Samples falling outside of the dotted lines (missingness > 1% and heterozygosity rate ±3 standard deviations from the mean) were removed from the analysis. Script for figure generation available from Anderson et al. (2010).

– Duplicated or related individuals with a kinship coefficient > 0.177 (indicating first-degree relatives or closer). Kinship coefficients were calculated for samples passing the heterozygosity and missingness checks, using markers with a MAF > 0.05 and the software KING (Manichaikul et al., 2010). The sample with the lowest call rate (or mismatching gender, if applicable) of each related pair was removed.

– Non-European samples, as determined using a principal component analysis (Figure 4.2) incorporating samples from the HapMap3 project (Altshuler et al., 2010).

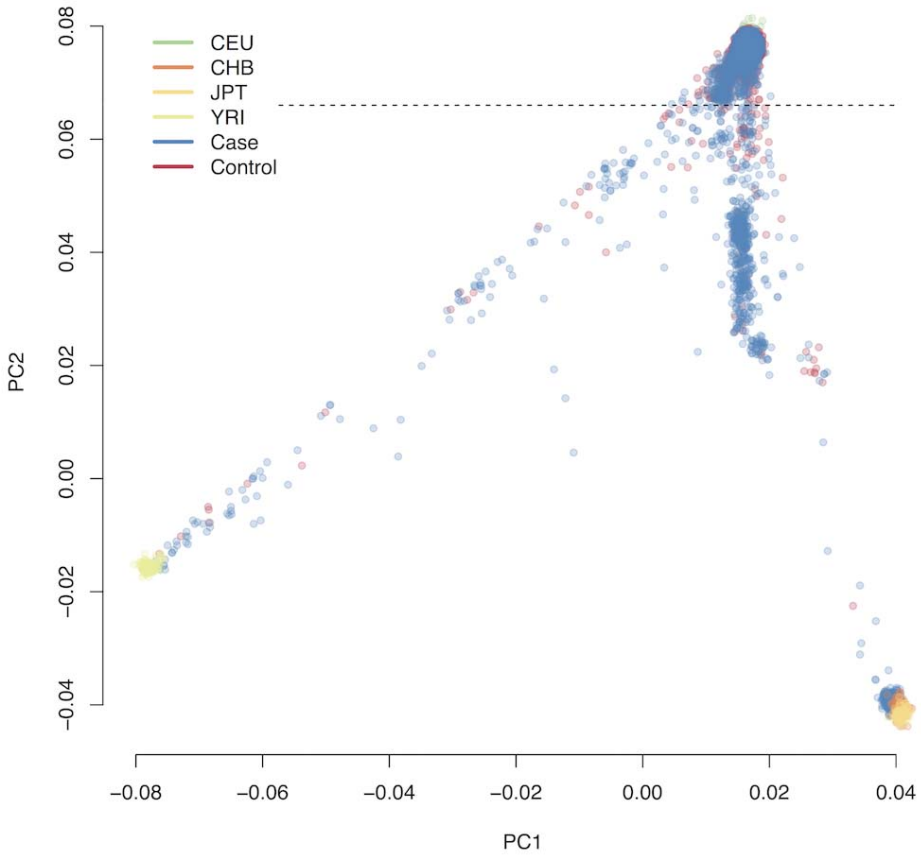

Figure 4.2: Principal component analysis of samples in the new UK IBD GWAS, analysed jointly with samples from the HapMap3 project (Altshuler et al., 2010). Samples with PC2 ≤ 0.066 (dotted line) were considered to be of non-European ancestry.

Site filtering

A final set of quality control filters were then used to remove markers still performing poorly amongst the high-quality samples, as determined by:

- Significant difference ($P < 1 \times 10^{-5}$) in call rate between cases and controls
- Evidence for a deviation from Hardy-Weinberg equilibrium in controls, where the p-value $< 1 \times 10^{-5}$
- One of 429 markers affected by a genotyping batch effect. These sites were identified by Yang Luo by computing within-sample principal components (PCs) using common variants (MAF $> 1\%$), which highlighted a clear outlier group of case samples all belonging to one genotyping batch (Figure 4.3a). PC1 was used to split cases into outliers and non-outliers, and an association test between these groups identified significant sites ($P < 1 \times 10^{-5}$). Once these sites were removed, the within-sample PCs no longer produced any outlier groups (Figure 4.3b).

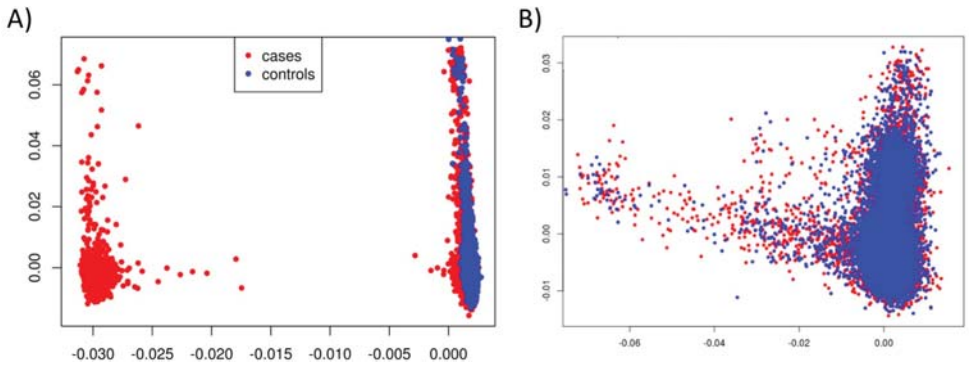

Figure 4.3: Panel A) depicts a genotyping batch effect identified in the new UK IBD GWAS using a principal component analysis, while panel B) shows the improvement after the removal of 429 sites that were significant when comparing the outlier samples from A) against the remaining samples. Figures generated by Yang Luo.

This left a high-quality dataset consisting of 510,520 genotyped sites in 9,239 cases (4,474 CD, 4,173 UC, 592 indeterminate IBD), and 9,500 controls. Before imputation, these sites were further pruned to those with a MAF $> 0.1\%$, leaving a total of 296,203 markers.

4.2.2 Imputation using an IBD-specific reference panel

Previous data has suggested that increasing the size of the reference panel used during imputation can significantly improve the accuracy of imputed low frequency variants (McCarthy et al., 2016), Therefore, as mentioned in Chapter 3, we created an expanded imputation reference panel, consisting of 4,686 low coverage IBD sequences collected by the UKIBDGC (retaining those individuals that were excluded from association analyses due to non-European ancestry), combined with 3,781 UK10K and 2,504 1000 Genomes Phase 3 control sequences. The inclusion of IBD samples helps to enrich the resulting reference panel with IBD-associated variants.

Prior to imputation, I remove any genotyped samples that were already included in the UKIBDGC low coverage sequencing study, as these would be present in the reference panel. I also remove any samples also included in the Wellcome Trust Case Control Consortium datasets (The Wellcome Trust Case Control Consortium, 2007; Barrett et al., 2009), as these samples contributed to the latest International IBD Genetics Consortium (IIBDGC) study that I shall be meta-analysing with this dataset. This left a total of 18,355 samples (4,264 Crohn's disease, 4,072 ulcerative colitis, 524 indeterminate inflammatory bowel disease, and 9,495 controls).

We then imputed whole genome sequences, down to a MAF $\sim 0.1\%$. Given the large size of both the reference and genotype panel, the computationally efficient software PBWT (Durbin, 2014) was used in order to obtain results in a tractable amount of time.

4.2.3 Meta-analysis of sequencing and imputed genomes with existing summary statistics

I tested these imputed sequences separately for association to ulcerative colitis, Crohn's disease and IBD using SNPTEST v2.5 (Marchini and Howie, 2010), performing an additive frequentist association test conditioned on the first ten principal components for each cohort. I then filtered out variants with MAF $< 0.1\%$,

INFO < 0.4, or strong evidence for deviations from Hardy-Weinberg equilibrium in controls ($P < 1 \times 10^{-7}$).

In order to increase power for the analysis of common variation, I obtained the publicly available summary statistics from the latest IIBDGC meta-analysis (Liu et al., 2015), and applied the same MAF $\geq 0.1\%$ and INFO ≥ 0.4 filters. I then used METAL (Willer et al., 2010) to perform a standard error weighted meta-analysis of the summary statistics from the UKIBDGC sequencing and imputed GWAS datasets together with the IIBDGC GWAS data.

4.2.4 Quality control

We filtered the output of this meta-analysis, removing sites with high evidence for heterogeneity ($I^2 > 0.90$) in any of the cohorts, and a meta-analysis p-value higher than all of the cohort-specific p-values. After this quality control, overall inflation of the summary statistics was still observed ($\lambda_{GC} = 1.23$ and 1.29 for Crohn's disease and ulcerative colitis, respectively). To determine if this was due to confounding population substructure that had not been properly accounted for, LD score regression was applied using LDSC v1 (Bulik-Sullivan et al., 2015) and European linkage disequilibrium (LD) scores from the 1000 Genomes Project (downloaded from https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2) on all sites with INFO > 0.95. Both intercepts were 1.09, indicating that the observed inflation is more likely to be due to broad polygenic signal.

In total, we tested 9.7 million high-quality sites across 25,042 IBD cases and 34,915 controls (Table 4.1), the largest genome-wide association test performed in inflammatory bowel disease to date. This dataset therefore offers us the opportunity to not only uncover further common variant IBD associations of small effect size, but also gives us reasonable power to perform causal variant fine-mapping in IBD-associated loci that were not covered by the Immunochip genotyping array used by Huang et al. (2015).

Table 4.1: Sample numbers and variant counts are described for each contributing dataset, at each stage of the analysis (Total = raw numbers, QC+ = post quality control, and N-o = after removing overlapping samples). Numbers are described separately for the ulcerative colitis, Crohn's disease, and inflammatory bowel disease analyses.

Study	Data	CD			UC			IBD			Controls			Grand Total		
		Total	QC+	N-o	Total	QC+	N-o	Total	QC+	N-o	Total	QC+	N-o	Total	QC+	N-o
UK sequences	Samples	2697	2513	1974	1817	1767	1326	4514	4280	3300	3910	3652	3650	8424	7932	6950
	Sites	95.1M	13.2M		95.1M	13.2M		95.1M	13.2M		95.1M	13.2M		95.1M	13.2M	
New UK GWAS	Samples	5695	4474	4264	5299	4173	4072	11758	9239	8860	10484	9500	9495	22242	18739	18355
	Genotyped	535k	296k		535k	296k		535k	296k		535k	296k		535k	296k	
	Imputed	109.5M	19.0M		109.5M	18.9M		109.5M	19.0M		109.5M	19.0M		109.5M	19.0M	
IIBDGC CD GWAS	Samples		5956	5956								14927	14927		20883	20883
	Sites	12.3M	12.0M								12.3M	12.0M		12.3M	12.0M	
IIBDGC UC GWAS	Samples					6968	6968					20464	20464		27432	27432
	Sites				12.3M	12.1M					12.3M	12.1M		12.3M	12.1M	
IIBDGC IBD GWAS	Samples								12882	12882		21770	21770		34652	34652
	Sites							12.7M	12.5M		12.7M	12.5M		12.7M	12.5M	
Meta-analysis (CD/UC/IBD)	Samples	12194			12366			25042			28072/ 33609/ 34915			40266/ 45975/ 59957		
	Sites	20.8M	9.6M		20.8M	9.6M		20.9M	9.7M		9.6M/ 9.6M/ 9.7M			9.6M/ 9.6M/ 9.7M		

4.3 Unravelling common variant associations

Overall, we identified 25 new IBD-associated loci at genome-wide significance (Table 4.2), including a number of associations of very small effect (OR < 1.1). In order to uncover causal variants, genes and mechanisms amongst these new associations, we performed a range of fine-mapping, eQTL co-localization, and gene enrichment tests as discussed in the following sections.

4.3.1 Fine-mapping and functional annotation of new and known loci

We performed a summary statistics fine-mapping analysis on the 25 novel IBD-associated loci, together with 40 previously discovered loci that reached genome-wide significance in this dataset but where fine-mapping had not previously been attempted. To do this, approximate Bayes factors were calculated from the meta-analysis effect sizes and standard errors, assuming the SNPTEST default prior variance on the log odds ratio of 0.04. These Bayes factors were then fine-mapped using the method outlined by the Wellcome Trust Case Control Consortium et al. (2012), to generate a posterior probability for each variant that reflects its likelihood of being causal in a given locus. The credible set for an association signal is defined as the smallest set of variants with posteriors that sum to at least 95%.

In order to properly resolve a GWAS signal down to the causal variant(s), it is important that all common SNPs in the locus have been directly genotyped or imputed to high quality (Spain and Barrett, 2015). This is to ensure that the truly causal SNPs are actually included in the fine-mapping comparison, when determining the relative evidence for causality of each associated SNP in the region. Therefore, to be confident about the conclusions drawn from this fine-mapping procedure, I only considered loci which had high quality imputed data for all relevant variants. This is defined as having no variants in the Phase 3 v5 release of the 1000 Genomes project (2013-05-02 sequence freeze) that are in high LD ($r^2 \geq 0.6$) with our hit SNP, but missing from our dataset, and no variants in our data within high LD ($r^2 > 0.8$) that fail during our QC procedure.

Table 4.2: Twenty five novel IBD-associated loci identified via a meta-analysis of 25,042 cases and 34,915 controls. The locus boundaries are defined by the left- and right-most variants that have an r^2 of 0.6 or more with the main variant.'RAF' refers to the risk allele frequency in the 1000 Genomes CEU and GBR populations.

Rsid	Chr	Position	Locus (Mb)	Risk Allele	Non-risk Allele	RAF	P_{Meta}	OR	95% CI	Phenotype	Implicated gene
rs34687326	1	159799910	159.80-159.80	G	A	0.900	1.06×10^{-08}	1.18	1.12-1.24	CD	*SLAMF8*
rs59043219	1	209970610	209.97-210.02	A	G	0.379	1.09×10^{-08}	1.08	1.05-1.10	IBD	-
rs6740847	2	182308352	182.31-182.33	A	G	0.508	1.22×10^{-13}	1.10	1.07-1.12	IBD	*ITGA4*
rs144344067	2	187576378	187.50-187.68	A	AT	0.895	1.29×10^{-08}	1.12	1.08-1.16	IBD	-
rs1811711	2	228670476	228.67-228.67	C	G	0.826	6.09×10^{-09}	1.14	1.10-1.18	UC	-
rs76527535	2	242484701	242.47-242.49	C	T	0.745	2.87×10^{-08}	1.09	1.06-1.12	IBD	-
rs2581828	3	53133149	53.10-53.17	C	G	0.597	6.46×10^{-09}	1.10	1.07-1.13	CD	-
rs2593855	3	71175495	71.16-71.19	C	T	0.663	2.54×10^{-09}	1.09	1.06-1.11	IBD	-
rs503734	3	101023748	100.91-101.27	A	G	0.513	2.67×10^{-08}	1.07	1.05-1.10	IBD	-
rs56116661	3	188401160	188.40-188.40	C	T	0.795	5.67×10^{-10}	1.14	1.10-1.18	CD	-
rs11734570	4	38588453	38.58-38.59	A	G	0.368	4.80×10^{-08}	1.07	1.05-1.10	IBD	-
rs17656349	5	149605994	149.59-149.63	T	C	0.466	1.54×10^{-08}	1.09	1.06-1.13	UC	-
rs113986290	6	19781009	19.72-19.83	C	T	0.989	7.59×10^{-09}	1.36	1.25-1.46	UC	-
rs67289879	6	42007403	42.00-42.01	T	C	0.179	3.04×10^{-08}	1.09	1.06-1.13	IBD	-

Continued on next page

Table 4.2 – *Continued from previous page*

Rsid	Chr	Position	Locus (Mb)	Risk Allele	Non-risk Allele	RAF	P_{Meta}	OR	95% CI	Phenotype	Implicated gene
rs11768365	7	6545188	6.50-6.55	A	G	0.816	3.88×10^{-08}	1.09	1.06-1.12	IBD	-
rs149169037	7	20577298	20.58-20.58	G	A	0.895	3.26×10^{-08}	1.14	1.10-1.19	IBD	*ITGB8*
rs243505	7	148435339	148.40-148.58	A	G	0.624	3.04×10^{-10}	1.08	1.06-1.11	IBD	-
rs7911117	10	27179596	27.16-27.18	T	G	0.871	1.84×10^{-08}	1.14	1.10-1.19	UC	-
rs111456533	10	126439381	126.32-126.55	G	A	0.829	1.18×10^{-09}	1.11	1.08-1.14	IBD	-
rs80244186	13	42917861	42.84-42.94	C	T	0.111	3.66×10^{-08}	1.13	1.09-1.18	CD	-
rs11548656	16	81916912	81.91-81.92	A	G	0.961	5.18×10^{-11}	1.27	1.20-1.34	IBD	*PLCG2*
rs10492862	16	82867456	82.87-82.92	A	C	0.308	1.26×10^{-09}	1.11	1.08-1.15	CD	-
rs4256018	20	6093889	6.08-6.10	G	T	0.250	1.23×10^{-08}	1.08	1.05-1.11	IBD	-
rs138788	22	35729721	35.72-35.74	A	G	0.418	2.95×10^{-08}	1.09	1.06-1.13	UC	-
rs4821544	22	37258503	37.26-37.26	C	T	0.321	1.76×10^{-08}	1.10	1.07-1.13	CD	-

Because of the relative sparsity with which the genome-wide microarrays cover each region (as opposed to dense genotyping arrays, such as the Immunochip), only 12 loci pass this filtering step. For 6 of these, there exists a single variant with $> 50\%$ probability of being causal (Table 4.3). For those implicated variants that were directly genotyped in the new UKIBDGC GWAS dataset, the cluster plots were manually checked to confirm quality data (Figure 4.4).

Of particular interest are two loci where a single variant had >99% probability of being causal. The first causally implicated variant, rs34687326, is a missense change predicted to affect protein function in *SLAMF8* (p.Gly99Ser, Figure 4.5a). As can be seen in Figure 4.5a, this signal was relatively easy to resolve given the low linkage disequilibrium between this lead SNP and the surrounding variation. While this sparse Manhattan plot was initially concerning, we were reassured by the very clean cluster plots produced by direct genotyping of the variant rs34687326 in the new UKIBDGC GWAS dataset (Figure 4.4a), increasing our confidence that this is a true association.

SLAMF8 is a cell surface receptor expressed by various myeloid cells (including neutrophils, macrophages and dendritic cells) after exposure to gram- or gram+ bacteria, lipopolysaccaride (LPS) or interferon (IFN)-γ, where it has been reported to inhibit the migration of these cells to sites of inflammation (Wang et al., 2015). In addition, SLAMF8 has been shown to play a role in repressing the production of reactive oxygen species (ROS) by these cells, further negatively regulating inflammatory responses (Wang et al., 2012). The risk-decreasing allele in our dataset (MAF=0.1, Table 4.2) is predicted to strongly affect protein function (CADD=32, 92nd percentile of missense variants, Kircher et al. (2014)), suggesting that further experimental follow up to evaluate a possible gain-of-function mechanism may be worthwhile.

Table 4.3: Variants fine-mapped to > 50% probability of being causal in their given signal.

Rsid	Chr	Position	P_{Causal}	Effect	Credible set size	Phenotype	P_{Meta}	Locus type
rs34687326	1	159799910	1.000	*SLAMF8* p.Gly99Ser (missense)	1	CD	1.06×10^{-08}	Novel
rs4845604	1	151801680	0.999	*RORC* (intronic)	1	IBD	7.09×10^{-14}	Known
rs1811711	2	228670476	0.914		2	UC	6.09×10^{-09}	Novel
rs56116661	3	188401160	0.561	*LPP* (intronic)	11	CD	5.67×10^{-10}	Novel
rs11548656	16	81916912	0.502	*PLCG2* p.His244Arg (missense)	3	IBD	5.18×10^{-11}	Novel
rs1143687	16	81922813	0.746	*PLCG2* p.Arg268Trp (missense)	5	IBD	3.83×10^{-08}	Novel
rs4821544	22	37258503	0.804	*NCF4* (intronic)	2	CD	1.76×10^{-08}	Novel

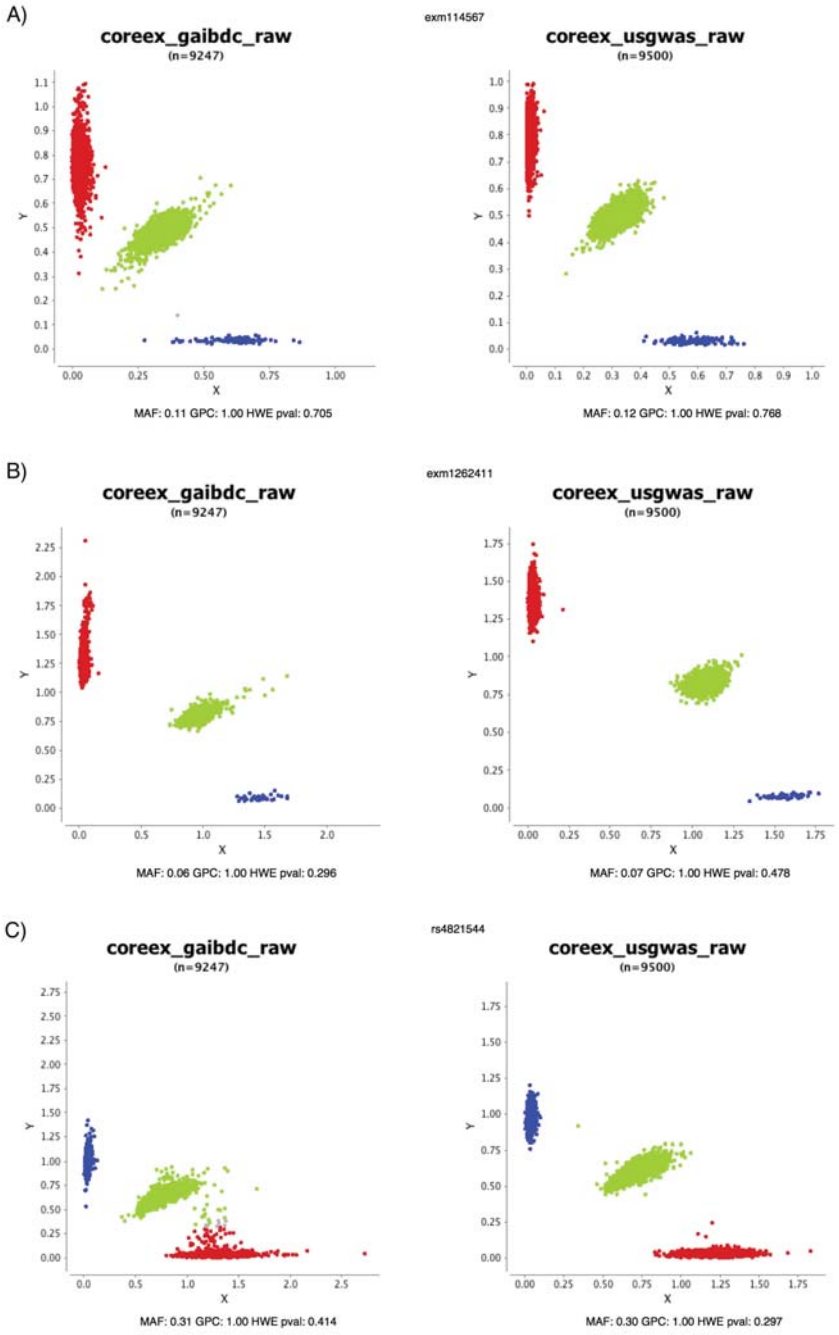

Figure 4.4: Cluster plots for A) rs34687326, B) rs1143687 and C) rs4821544 for the new UK IBD GWAS samples that passed quality control. The SNP genotypes have been assigned based on cluster formation in scatter plots of normalized allele intensities X and Y. Each circle represents one individual's genotype. Blue and red clouds indicate homozygote genotypes for the SNP (CC/AA), green heterozygote (CA) and grey undetermined. Figures generated by Daniel Rice.

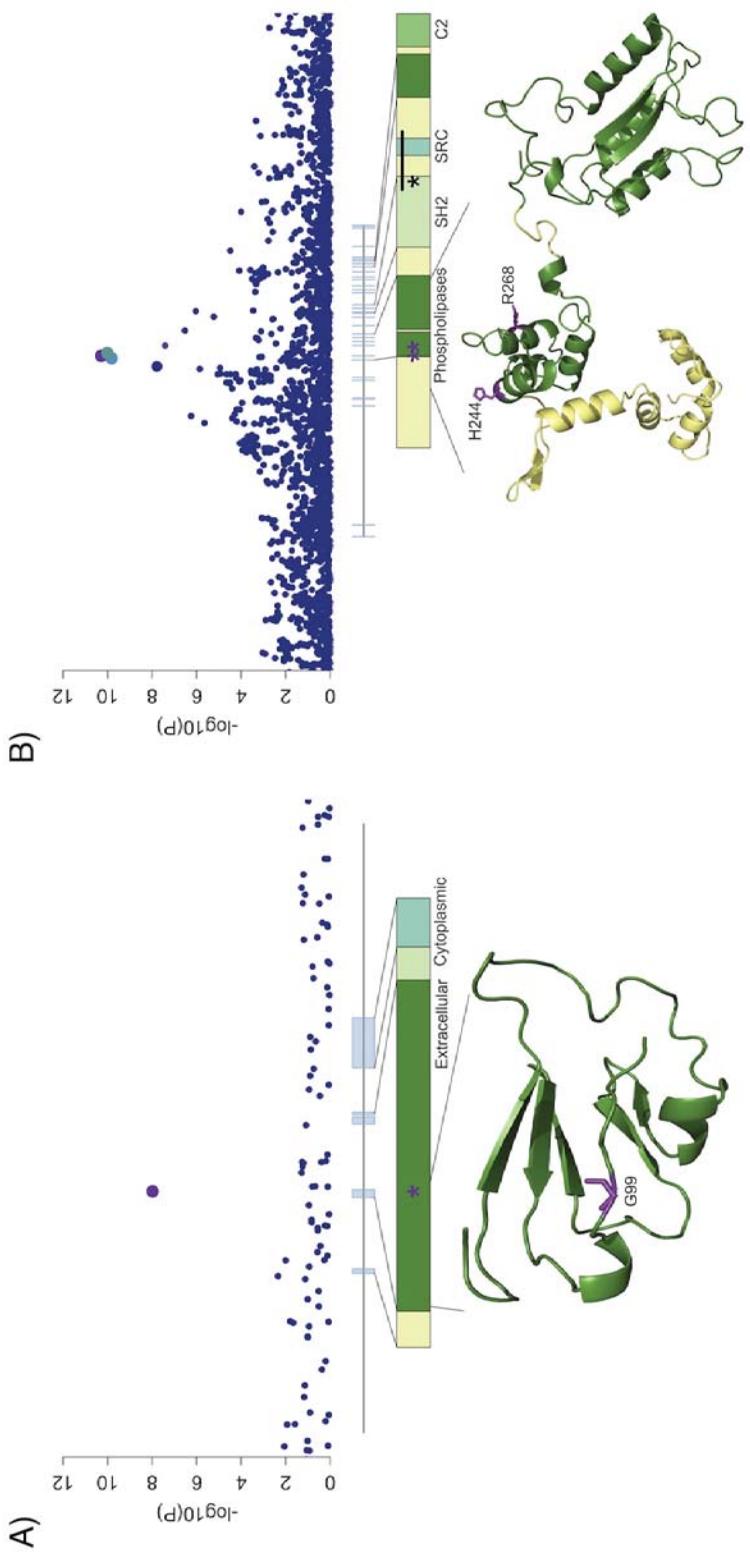

Figure 4.5: Likely causal missense variants in *SLAMF8* and *PLCG2*. For A) *SLAMF8* and B) *PLCG2*, local association results are plotted, with point size corresponding to LD to our lead variant and color to fine-mapping probability (purple $> 50\%$, intermediate blue $10 - 50\%$, navy blue $< 10\%$). Gene body diagrams and protein domain annotations are taken from ENSEMBL, and partial predicted crystal structures for both proteins are obtained from the SWISS-MODEL repository. Our fine-mapped variants (purple) and known variants leading to autosomal dominant immune disorders (black) are marked on the gene body diagrams.

The second variant with >99% probability of being causal is rs4845604, an intronic variant in the key regulator of T_H17 cell differentiation, *RORC* (Ivanov et al., 2006). *RORC* encodes RORγt, which is the master transcriptional regulator of both T_H17 cells (Ivanov et al., 2006) and group 3 innate lymphoid cells (Luci et al., 2009). These cell types both play important roles in defense at mucosal surfaces: in particular, they have been shown to help maintain homeostasis between the intestinal immune system and the gut microbiota (Yang et al., 2014c; Sawa et al., 2011). Loss of this equilibrium is often seen in inflammatory bowel disease (Gevers et al., 2014). Furthermore, pharmacologic inhibition of RORγt has been shown to be of therapeutic benefit in mouse models of intestinal inflammation, and reduces the frequency of T_H17 (but not innate lymphoid) cells isolated from primary intestinal samples of patients with inflammatory bowel disease (Withers et al., 2016).

Also of note is another likely functional variant amongst the remaining, less clearly resolved, fine-mapped loci (Table 4.3). This missense variant (CADD=16.5, 50.2% probability of causality) affects the gene *PLCG2* (Figure 4.5b). Interestingly, after conditioning on this variant we observe a second, independent missense variant in the same gene ($P = 2 \times 10^{-8}$), that is highly likely to affect protein function (CADD=34.0, 74.6% probability of causality). *PLCG2* encodes a phospholipase enzyme that plays an important role in regulating immune pathway signalling and T cell selection (Fu et al., 2012). It has also been implicated in two autosomal dominant immune disorders: intragenic deletions in the autoinhibitory domain of *PLCG2* cause antibody deficiency and immune dysregulation (familial cold autoinflammatory syndrome 3, MIM 614468), while heterozygous missense variants (e.g. p.Ser707Tyr) lead to a phenotype that includes intestinal inflammation (Ombrello et al., 2012; Zhou et al., 2012).

4.3.2 Enrichment amongst IBD loci for genes associated with Mendelian disorders of inflammation and immunity

An association is also observed between Crohn's disease and an intronic variant in *NCF4* (P=1.76×10^{-8}, 80.4% probability of causality), a gene which has also been associated with a Mendelian disorder of inflammation and immunity. In particular, *NCF4* encodes p40phox, part of the NADPH-oxidase system that destroys phagocytosed bacteria via an oxidative burst in innate immune cells (Tarazona-Santos et al., 2013). Rare pathogenic variants in *NCF4* cause autosomal recessive chronic granulomatous disease, which is characterized by intestinal inflammation and defective ROS production in neutrophils (Matute et al., 2009). Interestingly, the variant associated in our dataset, rs4821544, had previously been suggestively associated with small bowel Crohn's disease (Rioux et al., 2007; Roberts et al., 2008). When we stratified patients by disease location we found that the effect was consistently stronger for ileal disease (affecting the small bowel) compared to colonic (affecting the large bowel), as shown in Figure 4.6. This is consistent with growing genetic evidence that Crohn's disease may in fact be better defined as two distinct subtypes, ileal Crohn's disease and colonic Crohn's disease (Cleynen et al., 2016).

In order to test whether these observations in *PLCG2* and *NCF4* reflected a more general overlap between candidate IBD GWAS genes and Mendelian disorders of inflammation and immunity, I performed a gene set enrichment analysis. I defined the set of Mendelian disorder genes of interest as being those associated with primary immune deficiences according to the latest curated release by the International Union of Immunological Societies Expert Committee for Primary Immunodeficiency (Picard et al., 2015), as well as a secondary list of genes associated with rare disorders in OMIM that include inflammatory bowel disease as a clinical diagnostic. The secondary genes were obtained using a clinical synopsis search in OMIM (https://www.omim.org/search/advanced/clinicalSynopsis, as accessed on Sep 08, 2016) for the terms "Inflammatory bowel disease", "Crohn's disease" and "Ulcerative colitis", restricting the output to results where the molecular basis

has been identified. This list was then manually curated to exclude those entries corresponding to complex disorders.

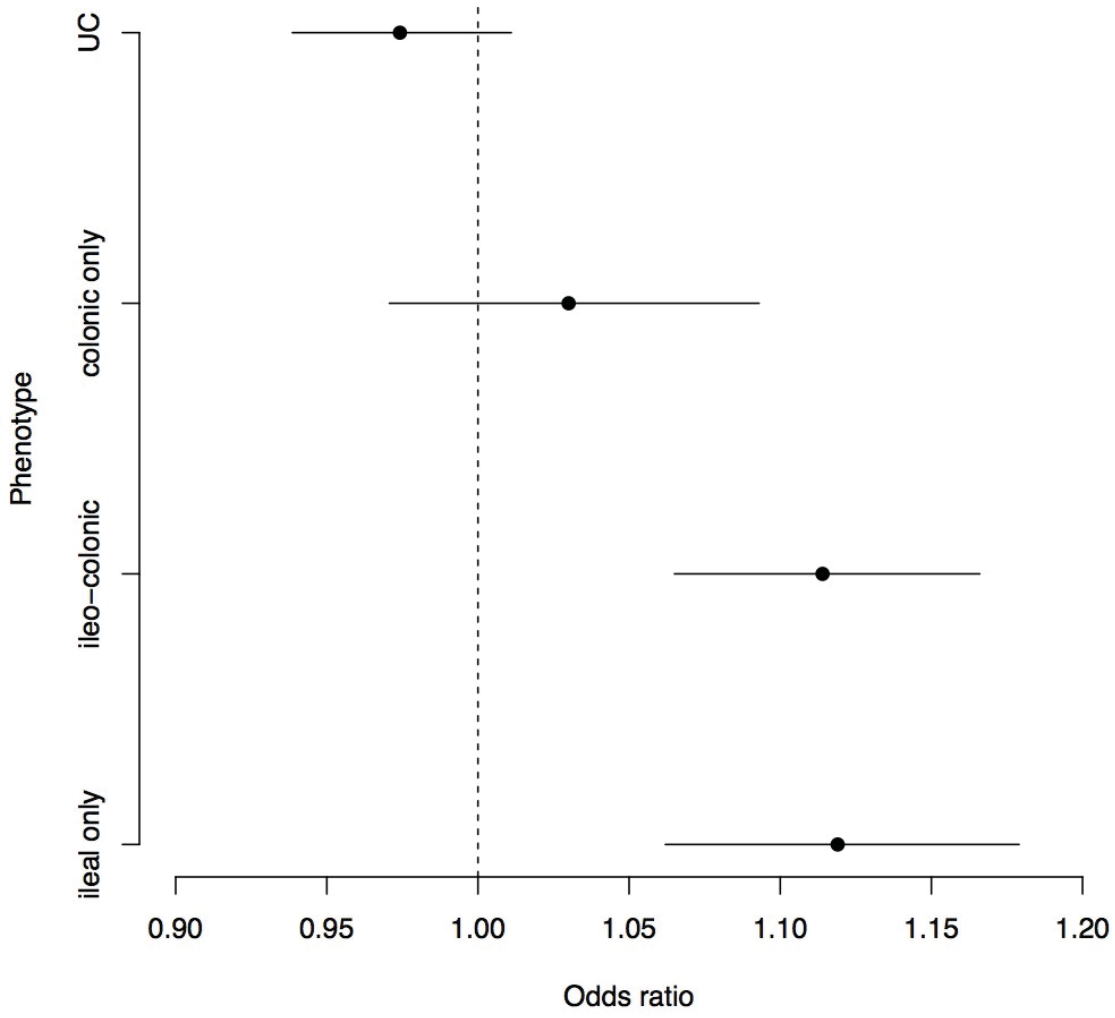

Figure 4.6: The effect of the well fine-mapped variant rs4821544, which is intronic in *NCF4*, is consistently stronger in small bowel compared to large bowel disease. Figure produced by Jeffrey Barrett.

Using the software INRICH (Lee et al., 2012a), I observe a significant enrichment ($P < 1 \times 10^{-6}$) of these genes within all 241 loci now associated with IBD (Appendix B). I then refine this test to just the 26 loci where a gene can be confidently implicated by fine-mapping to a coding variant or co-localization with an eQTL (Huang et al., 2015; Fairfax et al., 2014; Wright et al., 2014), as described in Table 4.4 and Chapter 3. Within the set of loci with a high-confidence gene assignment, the enrichment for genes associated with Mendelian disorders of inflammation and immunity is even stronger (27% vs 3%, $P = 2 \times 10^{-5}$ using a Fisher's exact test).

Table 4.4: Association of known IBD genes with Mendelian disorders of inflammation and immunity. These disorders include Primary Immune Deficiencies as defined by Picard et al. (2015), and Mendelian disorders which include IBD as a symptom, according to OMIM.

Gene	Phenotype	Primary Immune Deficiency	Additional rare disorders
CARD9	IBD	CARD9 deficiency	-
IFIH1	UC	Aicardi-Goutieres syndrome 7	Singleton-Merten syndrome 1
IL2RA	CD	CD25 deficiency	-
NOD2	CD	Blau syndrome	Early-onset sarcoidosis
PLCG2	IBD	PLAID (PLCγ2 associated antibody deficiency and immune dysregulation); Familial cold autoinflammatory syndrome 3; APLAID (autoinflammation and PLAID)	-
SMAD3	IBD	-	Loeys-Dietz syndrome 3

Remaining known IBD genes without an associated Mendelian disorder:
ADCY7 (UC), *ATG16L1* (CD), *CD6* (CD), *ERAP2* (CD), *FCGR2A* (IBD), *FUT2* (CD), *ICAM1* (IBD), *IL18RAP* (IBD), *IL23R* (IBD), *ITGA4* (IBD), *ITGAL* (UC), *ITGB8* (IBD), *MST1* (IBD), *NXPE1* (UC), *PTPN22* (CD), *SLAMF8* (CD), *SP140* (CD) and *TNFSF8* (IBD)

4.3.3 Co-localization of GWAS and eQTL associations

Among the remaining 21 novel loci, it was interesting to observe that three associations were within 150kb of integrin genes (*ITGA4*, *ITGAV* and *ITGB8*), while a previously associated locus also overlaps with a fourth integrin gene, *ITGAL*. In addition, a recent study has demonstrated that there is an IBD specific association that affects expression of *ICAM1*, which encodes the binding partner of *ITGAL* (Dendrou et al., 2016). The integrins encoded by these genes act as cell adhesion mediators that are capable of signalling across the plasma membrane in both directions, and have been shown to play a crucial role in leukocyte homing and cell differentiation in inflammation (Hynes, 2002). An overview of how integrins are involved in leukocyte homing to different tissues is given in Figure 4.7.

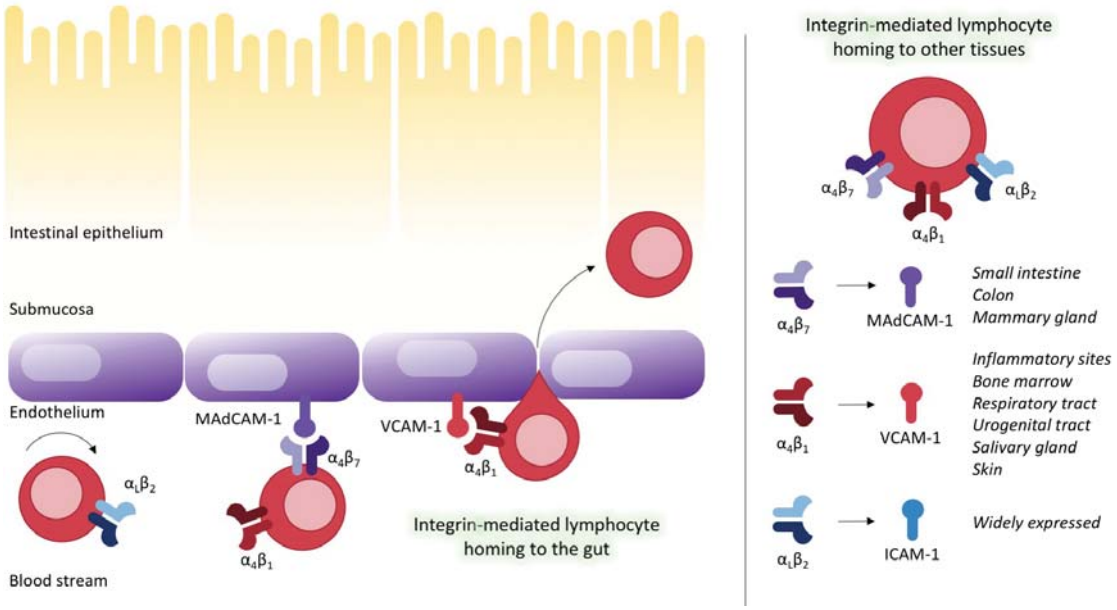

Figure 4.7: The role of integrins in leukocyte homing. Integrin complexes are expressed on the surface of leukocytes, and will bind to corresponding intercellular adhesion molecules on the surface of endothelial cells, prompting infiltration of the leukocytes into the tissue. Some of these binding partners exhibit tissue-specific expression patterns, allowing for tight control of leukocyte homing to specific sites (Kunkel and Butcher, 2003; Pals et al., 2007).

These integrin genes reflect very strong candidates for involvement in inflammatory bowel disease. However, both the gene set enrichment result discussed above, as well as the rare variant burden tests across known IBD genes described in Chapter 3, highlight the importance of using methods such as fine-mapping and eQTL association studies in order to confidently assign GWAS signals to their causal genes. As the fine-mapping analysis had already excluded the possibility that these associations could be caused by protein-coding changes, we next searched for effects of IBD risk SNPs on integrin gene expression in immune cells using a number of publicly available eQTL datasets (Zeller et al., 2010; Fairfax et al., 2012; Westra et al., 2013; Battle et al., 2014; Fairfax et al., 2014; Lee et al., 2014a; Raj et al., 2014; Ye et al., 2014; GTEx Consortium, 2015; Zhernakova et al., 2015).

While many eQTL and GWAS signals show some degree of correlation, inferences about causality require more robust statistical co-localization of the two signals. One means of obtaining this statistical support is to directly test for co-localization between IBD association signals and eQTLs using the coloc2 method (Giambartolomei et al., 2014), implemented in the R package coloc. We ran this method across our dataset, using a window size of 250kb on each side of the IBD association and default settings. Each test was repeated using two different p_{12} values ($p_{12} = 1 \times 10^{-5}$ and $p_{12} = 1 \times 10^{-6}$), which represents the prior probability of co-localization. For each gene, we test for co-localization with eQTLs in unstimulated monocytes, as well as monocytes stimulated with lipopolysaccaride (LPS) after 2 and 24 hours, monocytes stimulated with IFN-γ, and in unstimulated B cells, as described by Fairfax et al. (2014). The results of this analysis are summarised in Table 4.5.

Table 4.5: Co-localization between meta-analysis association statistics and monocyte stimulus response eQTLs. The co-localization of the meta-analysis and eQTL signals is tested with two different priors (1×10^{-5} and 1×10^{-6}) across the genes *ITGA4*, *ITGB8*, *ITGAL* and *ICAM1*. For each gene, we test co-localization with eQTLs in unstimulated monocytes, as well as monocytes stimulated with LPS after 2 and 24 hours, monocytes stimulated with IFN-γ, and in unstimulated B cells.

	Prior (p12)	Posterior probability of co-localization between GWAS association and monocyte eQTLs (after the application of stimuli)				
		Naive	LPS2HR	LPS24HR	IFN-γ	BCELL
ITGAL	1×10^{-05}	0.089	0.045	0.980	0.989	0.045
	1×10^{-06}	0.010	0.005	0.833	0.896	0.005
ITGB8	1×10^{-05}	0.061	0.057	0.712	0.051	0.178
	1×10^{-06}	0.006	0.006	0.198	0.005	0.021
ITGA4	1×10^{-05}	0.979	0.736	0.984	0.992	0.228
	1×10^{-06}	0.823	0.218	0.864	0.923	0.029
ICAM1	1×10^{-05}	0.050	0.961	0.093	0.162	0.064
	1×10^{-06}	0.005	0.713	0.010	0.019	0.007

Remarkably, three of the associations near integrin genes had $> 90\%$ probability of being driven by the same variants as monocyte-specific stimulus response eQTLs (*ITGA4*, $P_{\text{LPS\_24hr}} = 0.984$; *ITGAL*, $P_{\text{LPS\_24hr}} = 0.980$; *ICAM1*, $P_{\text{LPS\_2hr}} = 0.961$). A fourth association, *ITGB8*, is difficult to map due to extended linkage disequilibrium in the locus, but shows intermediate evidence of co-localization ($P_{\text{LPS\_24hr}} = 0.712$) in response to the same stimulus (Figure 4.8). All four of the IBD risk increasing alleles are associated with upregulated expression of their respective genes, suggesting that an increased level of pro-inflammatory cell surface markers in response to stimulus may be a consistent mechanism of action for these associations. Determining if this is indeed the case, however, would require functional follow up to prove that these IBD risk alleles causally change gene expression in response to stimulus, and indeed that changes in integrin gene expression are relevant to the inflammatory bowel disease phenotype.

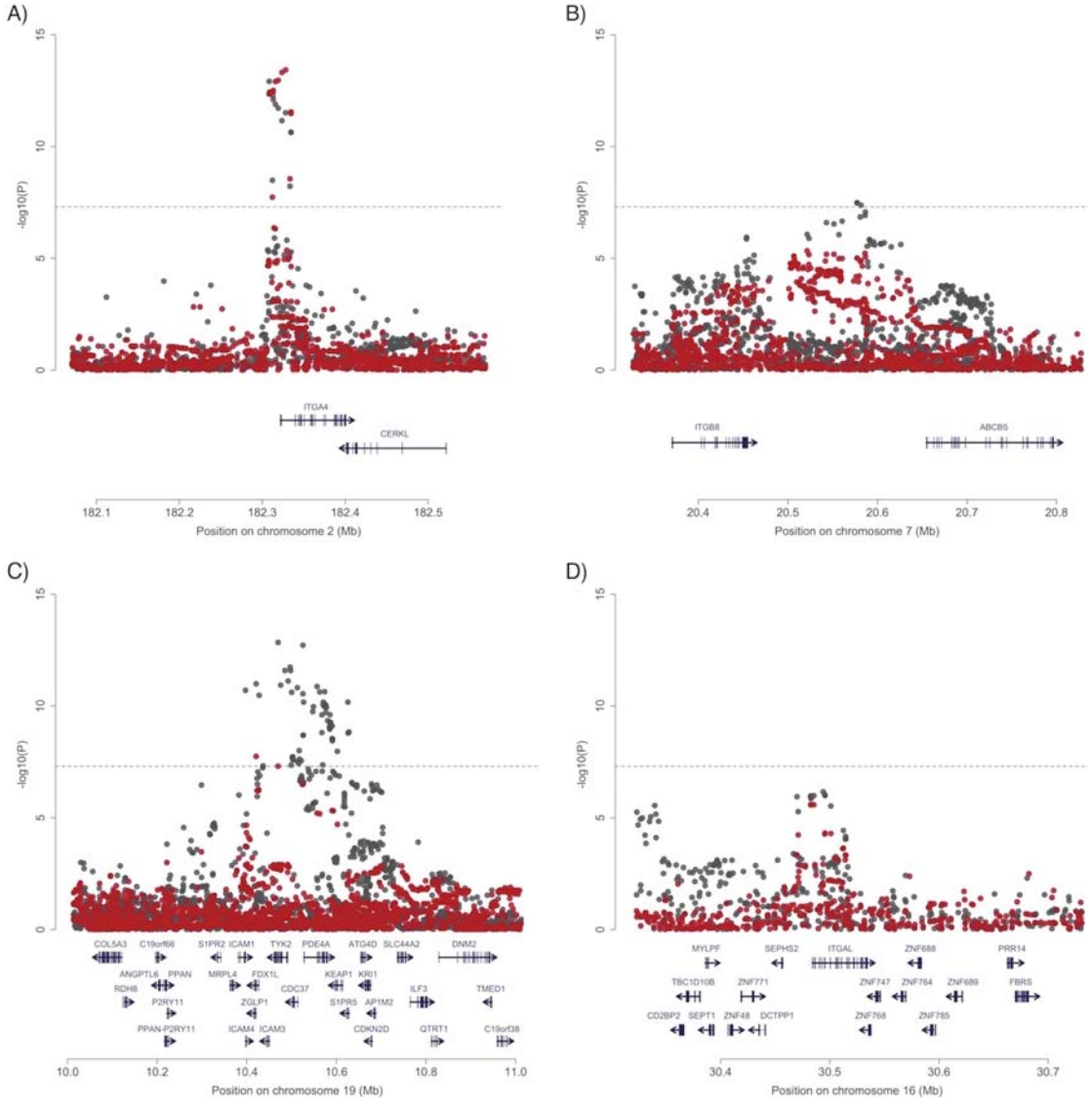

Figure 4.8: Co-localization of disease association and stimulus response eQTLs in monocytes. The local pattern of disease association (IBD: (A) *ITGA4*, (B) *ITGB8*, (C) *ICAM1*; (D) UC: *ITGAL*) in grey, and the association of that variant with response to LPS stimulation in red. Evidence of co-localization (probability > 70%) is observed for all for signals.

This second point is supported by the recent emergence of integrins and their counter-receptors as important therapeutic targets in IBD. In particular, the monoclonal antibodies vedolizumab and etrolizumab, which target the components of the $\alpha 4\beta 7$ dimer (encoded by *ITGA4* and *ITGB7*, and responsible for the gut-homing specificity of certain leukocytes), have demonstrated efficacy in both CD

and UC (Feagan et al., 2013; Sandborn et al., 2013; Vermeire et al., 2014). In addition, an antisense oligonucleotide that targets *ICAM1* has recently shown promise in the treatment of ulcerative colitis and pouchitis (Hosten et al., 2014).

Therapeutics targeting integrin αL (efalizumab) and α4 (natalizumab) have also demonstrated potential in the treatment of Crohn's disease (Sandborn et al., 2005; James et al., 2011), but have subsequently been associated with progressive multifocal leukoencephalopathy, or PML (Carson et al., 2009). This association highlights the importance of gut-selectivity in therapeutic approaches, with the potentially fatal PML condition likely to be mediated by binding to integrin dimers that are not gut-specific (leading to deficiencies in leukocyte migration to the central nervous system, and allowing for JC virus infection in the brain). Because of the risk of PML, efalizumab has been withdrawn from the market and natalizumab is not licensed for Crohn's disease in Europe.

Integrins are not only important in cell trafficking, but can also contribute to cellular signalling. For example, the αVβ8 heterodimer - both subunits of which are encoded by genes which are now within confirmed IBD loci (*ITGAV* and *ITGB8*, respectively) - is a potent activator of TGFβ. Notably, mice with dendritic-cell specific deletion of this complex had impaired regulatory T cell function and severe colitis (Travis et al., 2007), while deleting it in regulatory T cells themselves prevented the suppression of pathogenic T cell responses during active inflammation (Worthington et al., 2015). Although no therapeutics directly target αVβ8, there have been promising early results from an oral antisense oligonucleotide to the inhibitory TGFβ-signalling protein SMAD7 (Monteleone et al., 2015), itself encoded by a locus identified by genetic association studies (Jostins et al., 2012), that emphasises the therapeutic potential of modifying TGFβ in inflammatory bowel disease.

4.3.4 Therapeutic relevance of genetic associations

The associations to anti-integrin and anti-TGFβ therapies described above are just a few examples of therapeutically relevant genes that have been implicated using genetic studies of inflammatory bowel disease. To investigate these connections on a broader scale, we identified the following immune pathways as relevant to classes of approved IBD therapeutics: the IL12 and IL23 signalling pathways (ustekinumab, Sandborn et al. (2012)), the TNFα signalling pathway (infliximab, Hanauer et al. (2002); adalimumab, Colombel et al. (2007)), and the integrin signalling pathway (vedolizumab, Feagan et al. (2013) and Sandborn et al. (2013)). Genes involved in these pathways were then identified using the Molecular Signatures Database canonical pathways gene sets (C2; available at http://software.broadinstitute.org/gsea/msigdb/genesets.jsp?collection=CP), which have been curated by the Pathway Interaction Database (Schaefer et al., 2009). The integrin signalling gene list was comprised of all unique genes from the following gene sets: the integrin β1 pathway, integrin β7 pathway and integrin cell surface interactions. The list of TNFα signalling genes was obtained from the TNF pathway, and the list of IL-23/IL-12 p40 signalling genes was comprised of all unique genes from the IL12 and IL23 pathways.

Based on these gene lists, I identified genes in known IBD loci of therapeutic relevance (Table 4.6). As Figure 4.9 highlights, the importance of the biological pathways underlying associations, and their potential therapeutic significance, are not necessarily reflected in their GWAS effects sizes, with many relevant associations requiring tens of thousands of samples to identify.

Table 4.6: IBD-associated loci containing genes in immune pathways related to classes of approved therapeutics. We highlight loci that contain a gene in one of four signalling pathways related to targets of three classes of approved IBD therapeutics. In each case the relevant gene, signalling pathway, and therapeutic is marked. Genes marked with a ⋆ have been confidently implicated as the causal IBD gene.

Chr	Locus (Mb)	Relevant Gene	Pathway	Therapeutic(s)
1	67.2-68.1	*IL23R*⋆,*IL12RB2*	IL12, IL23	Ustekinumab
4	123-123.6	*IL2*	IL12, IL23	Ustekinumab
7	107.4-107.6	*LAMB1*	Integrinβ1	Vedolizumab
3	46.2-46.5	*CCR5*	IL12	Ustekinumab
14	75.7-75.7	*FOS*	IL12	Ustekinumab
16	11.3-11.7	*SOCS1*	IL12	Ustekinumab
6	149.6-149.6	*TAB2*	TNF	Infliximab, Adalimumab
4	102.7-103.5	*NFKB1*	IL12, IL23, TNF	Ustekinumab, Infliximab, Adalimumab
2	191.9-192	*STAT4*	IL12, IL23	Ustekinumab
10	75.5-75.7	*PLAU*	Integrinβ1, Integrinβ5-8	Vedolizumab
16	30.5-30.5	*ITGAL*⋆	Integrin cell interactions	Vedolizumab
17	32.6-32.6	*CCL2*	IL23	Ustekinumab
2	102.6-103.2	*IL18RAP*⋆,*IL18R1*,*IL1R1*	IL12, IL23	Ustekinumab
10	6.0-6.5	*IL2RA*⋆	IL12	Ustekinumab
5	158.7-158.9	*IL12B*	IL12, IL23	Ustekinumab
17	40.4-40.7	*STAT5A*,*STAT3*	IL12, IL23	Ustekinumab
19	10.4-10.6	*TYK2*⋆	IL12, IL23	Ustekinumab
2	182.3-182.3	*ITGA4*⋆	Integrinβ1, Integrinβ5-8, Integrin cell interactions	Vedolizumab
2	187.5-187.7	*ITGAV*	Integrinβ1, Integrinβ5-8, Integrin cell interactions	Vedolizumab
7	20.6-20.6	*ITGB8*⋆	Integrinβ5-8, Integrin cell interactions	Vedolizumab

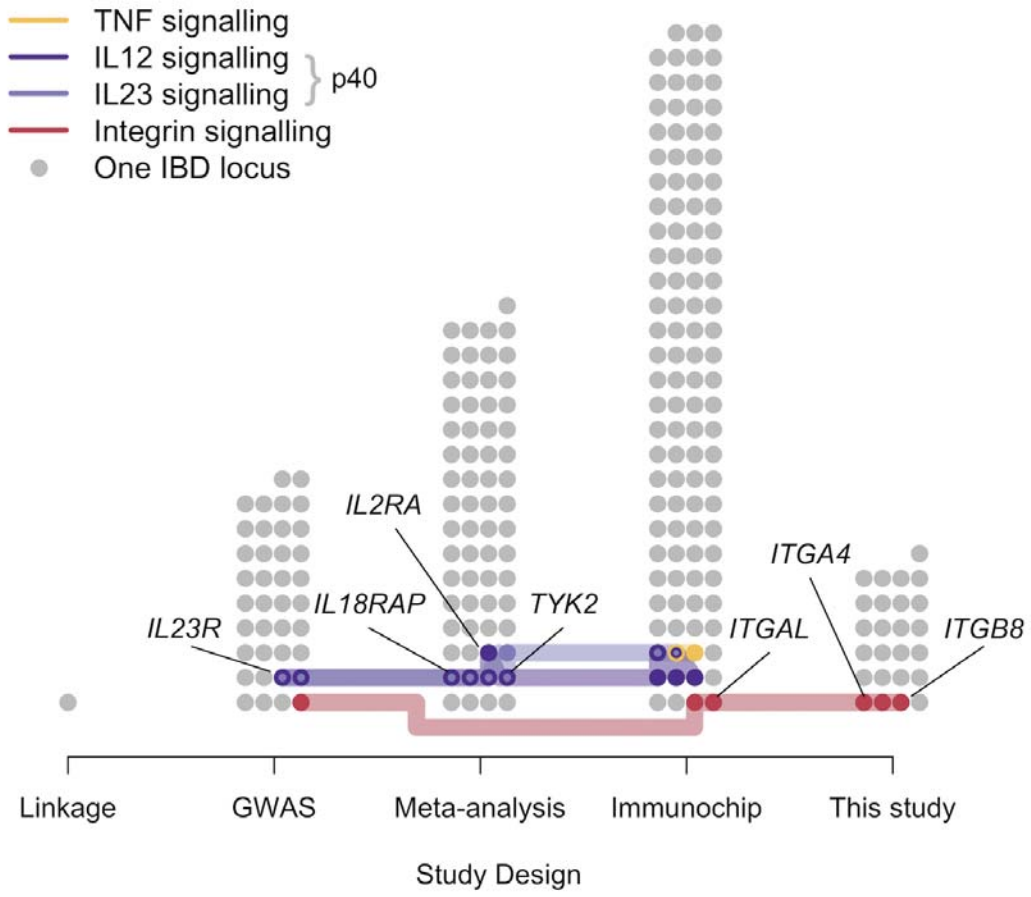

Figure 4.9: IBD-associated loci containing genes in immune pathways related to classes of approved therapeutics. All IBD loci are divided into the studies where they were first identified. Loci that contain a gene in one of four signalling pathways related to targets of three classes of approved IBD therapeutics are highlighted, with those where the pathway gene has been confidently identified as the causal IBD gene labelled. Despite the general pattern that effect size decreases from left to right, therapeutically relevant associations continue to be found.

4.4 Discussion

In this chapter I have described the discovery of 25 novel IBD-associated loci via the imputation and meta-analysis of nearly 60,000 samples, bringing the total number of loci associated with IBD to 241. Summary-statistic fine-mapping on these loci identified likely causal missense variants in the genes *SLAMF8*, a negative regulator of inflammation, and *PLCG2*, a gene implicated in primary immune deficiency. A potentially causal variant is also observed in an intron of *NCF4*, which is another gene associated with an immune-related Mendelian disorder.

A potential relationship between genes associated with Mendelian disorders of inflammation and immunity and those implicated in IBD has long been recognised, with the first Crohn's disease risk gene discovered, *NOD2*, also linked to the autosomal dominant granulomatous disorder Blau syndrome (Miceli-Richard et al., 2001). I confirm this link more generally, showing a strong enrichment for such genes amongst all known IBD loci. Furthermore, this enrichment is significantly stronger when considering just those IBD-associated loci for which a gene can be causally assigned with high confidence, either through fine-mapping or eQTL co-localization, highlighting the importance of using such methods when trying to draw conclusions about the biological mechanisms underlying an association.

Amongst the novel loci that could not be fine-mapped to a likely causal variant, three are proximal to integrin genes, which encode proteins in pathways that have been identified as important therapeutic targets in inflammatory bowel disease. Co-localization with eQTL signals confirm that the associated IBD risk-increasing variants are also correlated with expression changes in monocytes in response to immune stimulus at two of these genes (*ITGA4* and *ITGB8*), and at two previously implicated loci (*ITGAL* and *ICAM1*). This suggests that an increased level of pro-inflammatory cell surface markers in response to stimulus may be a consistent mechanism of action for these particular associations, although further functional follow up would be required to confirm this.

The discovery of this association between integrin genes and inflammatory bowel disease was particularly exciting for two key reasons. Firstly, assigning the signal

detected using GWAS to the likely causal genes would not have been possible without the ability to test for co-localization in an eQTL dataset that had analysed both the relevant cell type, and used the correct stimulus. No co-localization was observed between our data and eQTLs for integrin expression in B cells; similarly, a number of the GWAS associations failed to co-localize with integrin eQTLs from naive monocytes, or even monocytes stimulated with interferon-γ (Table 4.5). As studies that aim to uncover the specific cellular contexts in which different genes are active continue to grow in number and coverage, there is an exciting opportunity to potentially resolve the biological mechanisms underlying a number of other GWAS loci that can not be assigned to causal coding variation. Secondly, despite the relatively modest effect size of the signals near integrin genes (OR 1.10-1.12), they are of high therapeutic relevance. If we extend the idea of therapeutic relevance to other IBD-associated loci, it is clear that the importance of the biological pathways underlying genetic associations, and their potential use as drug targets, do not necessarily correlate with their GWAS effect sizes (Figure 4.9).

Overall, our findings suggest that there are still a number of potential benefits to be obtained by continuing to pursue genome-wide association studies, even in a well-studied complex disease like IBD, as valuable complementary analyses to large-scale sequencing endeavours.