Chapter 4

Investigating IBD genetics using the Immunochip

4.1 Introduction

This chapter describes a set of studies carried out using a custom genotyping platform named Immunochip. This genotyping chip was collaboratively designed by a large number of researchers in the genetics of complex immune and inflammatory disease, in order to offer an affordable way of performing very large locus discovery and fine-mapping studies. This chapter describes the application of this genotyping chip to the large number of samples collected by the component research groups of the International IBD Genetics Consortium (IIBDGC).

Both the Immunochip in general, and the IIBDGC study in particular, have been very successful in uncovering the genetics of immune-mediated disease. One study described in this chapter increased the number of associated loci known for IBD to 163, which is more than for any other complex disease. The very large number of associations has also necessitated a change in the way we interpret these results, from a locus-by-locus examination of genes to a large-scale bioinformatic interrogation of all loci. Much of this chapter will be dedicated to applying these techniques to the results of the Immunochip studies.

4.1.1 Overview of this chapter

I will begin (Section 4.2) with a discussion of the design of the Immunochip. This section starts with a discussion of the economics and power considerations of large-scale locus discovery and fine-mapping projects. It also contains a brief investigation into the biology of the fine-mapping loci submitted to the Immunochip, and what they tell us about the shared biology of immune-mediated diseases.

Section 4.3 will discuss the IIBDGC Immunochip data itself, and how calling, quality control and association analyses were carried out. It will describe the large number of novel loci this study has uncovered. Section 4.4 describes a detailed set of bioinformatic analyses to transform this locus list into biological insights. These analyses draw on a range of external data, such as associations with other phenotypes, gene networks, gene annotations, population genetic data and expression analyses. This section also sets out the main biological conclusions that can be drawn from these analyses, as I see them.

Finally, I will discuss two smaller studies carried out on this dataset. Section 4.5 discusses an association study of Y chromosome haplogroups in IBD, and reports a novel association with a Northern European Y haplogroup. Section 4.6 discusses a pilot fine-mapping project, investigating coding and non-coding causal variants in the important *NOD2* locus in CD, which will act as a template for larger Immunochip fine-mapping efforts.

Genotyping method	Cost/sample	Number of variants
Sequenom genotyping (1 plex)	£1.25	25-30
Illumina OmniExpress GWAS array	£160	800,000
Agilent and HiSeq targeted sequencing	£90	All in 6Mbp target region
Illumina Infinium iSelect HD custom genotyping	£25[a]	90,000-250,000

Table 4.1: The costs and capacities of various genotyping technologies. All costs are approximate, and assume large order numbers (>5000 individuals). [a] Assuming an order of >100,000 chips.

4.2 An overview of the Immunochip

4.2.1 The economics of the Immunochip

The economics of deep replication

The 30 novel loci discovered by the last International IBD Genetics Consortium's GWAS meta-analysis of Crohn's disease (Franke et al., 2010) have a median odds ratio of 1.1. The total discovery and replication dataset in this study contained 22,441 cases and 29,496 controls, and thus had a 90% power to establish such loci at genome-wide significance ($p < 5 \text{ x } 10^{-8}$), assuming an allele frequency of 0.25 and an additive genetic model. However, a limitation of this study was that the discovery cohort only had ~29% power to detect these signals with a p-value less than the significance threshold required to be taken forward into the replication ($p < 5 \text{ x } 10^{-6}$). This means that we have likely only discovered 29% of the variants that the total collection is well-powered to detect, suggesting another 70 loci that could be discovered. How can we map these loci in an affordable manner?

One option for uncovering some of these associations would be to expand the GWAS collection. Doubling the number of cases on a low-cost genotyping chip such as the Illumina OmniExpress would cost around £160 x 6333 = £1,013,280 (all costs shown in Table 4.1). This would increase the proportion of true associations taken forward for replication to 65%, and would likely result in around 50 new loci for follow-up. Replication on two Sequenom plexes would then cost around £76,370. This would thus involve spending a total of £1,089,650 to discover approximately 33 new loci, at a cost of £29,450 per locus.

Instead of expanding the GWAS collection, we could instead expand the replication genotyping effort (a so-called deep replication experiment). For a replication set containing all SNPs with $p < 10^{-4}$ would contain around 800 SNPs (or 32 Sequenom plexes), and would include 54% of true associations. This would cost £1.25 x 32 x 30,548 = £1,221,920 to uncover approximately 26 loci, or £46,997 per locus.

There is a third option: custom microarray genotyping. Designing a custom genotyping array allows the genotyping of a large number of SNPs at a lower cost than GWAS arrays. For instance, the Illumina iSelect Infinium HD custom genotyping chips can genotype up to 250,000 markers. For small numbers of samples the cost is relatively high (starting at around £100/sample). However, if a very large number of chips are ordered the price can fall substantially, and for orders measured in the hundreds of thousands the price falls to under £25/sample.

At this price, the entire IIBDGC replication cohort can be genotyped for £763,700. Additionally, because tens of thousands of SNPs can be taken forward for replication, we can perform very deep replication. For instance, taking forward the approximately 5000 SNPs that show $p < 10^{-3}$ would allow us to test 76% of true associations. This would allow us to discover 44 new loci at a cost of £17,357 each.

The economics of fine-mapping

Most of the associations that have been established during the IIBDGC meta-analyses are still poorly understood. For all but the most long-established associations the causal variant is unknown, and in many cases the gene or genes that are being acted on are also unknown. Bioinformatic techniques, such as those discussed in section 4.4.3, can shed light some light on the causal genes. However, the gold standard for establishing causation is genetic fine-mapping, i.e. demonstrating that a single variant, and no others, is capable of explaining the observed association.

In general, fine-mapping is not easy to achieve. To take a simple example, consider a common association (allele frequency of 50%) with a small effect size (odds ratio of 1.2), with the lead SNP in high LD ($r^2 = D' = 0.95$) with another variant of the same frequency. To have 80% power to identify the causal variant with high certainty (i.e. posterior > 0.99), we would require genotypes at 20,000 cases and 20,000 controls. In practice, the structure of the genome, and the biases of GWAS detection, will lead to most associations having many variants in high LD. To fine-map these associations we need a large number of samples, genotyped for a large number of SNPs. The IIBDGC cohort, with an effective sample size of around 25,000 cases, has enough power to fine-map a significant fraction of the CD associations detected by GWAS. However, designing this experiment in an affordable manner is difficult.

A basic fine-mapping effort will involve genotyping a limited set of candidate causal variants. If we examine the 40 CD loci that have not been previously fine-mapped with the lowest degree of LD, we find that there are

536 SNPs with $r^2 > 0.8$ to the hit SNP in the 1000 Genomes pilot. This set of SNPs could be genotyped using around 19 Sequenom plexes, and would cost £1,233,504 to genotype the entire IIBDGC cohort. However, if the causal variant has $r^2 < 0.8$ to the hit SNP, we will not find it (and indeed may end up with a false positive causal variant). Additionally, only the primary signal at the locus can be fine-mapped in this fashion.

The ideal fine-mapping experiment involves sequencing entire regions, as this allows us to assay all variants that could drive the association, as well as allowing us to identify new (potentially low-frequency) associations. A pull-down array designed to capture DNA from CD loci, combined with low-cost next-generation sequencing would allow us to perform this. However, while the cost of sequencing is now extremely low, the cost of sample preparation and the pull-down arrays is still relatively high. Even if we restricted sequencing to 20K cases and 20K controls, such a project would still cost in excess of £3,600,000.

Again, a powerful third solution comes in the form of custom genotyping, and in particular via a combined deep replication and fine-mapping array. The same genotyping array that is being used for deep replication (and thus is already being run on a significant fraction of the IIBDGC cohort) can also used to genotype variants in IBD-associated regions taken from the 1000 Genomes project and dbSNP. This allows the primary signal and any secondary signals to be fine-mapped, and also allows any low-frequency variation that is in the SNP databases to be assayed as well. This approach has less full coverage than would be achieved by sequencing, but for common variation the coverage should be nearly as high, at a much lower cost. In essence, the fine-mapping and deep replication efforts are combined on a single chip.

An immune-mediated disease chip

We have seen that custom genotyping is an affordable way to discover and fine-map new loci using existing collections, providing that a large enough purchase is made. If the IIBDGC alone purchased 40,000 chips (enough to genotype all CD and UC cases, and all controls), this would still be too small an order to be cost effective. However, by including deep replication studies from other disease consortia, we can rapidly increase the total number of chip users, and reduce the price to affordable levels.

It was these economic considerations that led to the creation of the Immunochip. This custom genotyping platform was designed for deep replication and fine-mapping in a wide range of studies, with particular focus on immune-mediated diseases (Table 4.2). Along with the reduction in price, there are a number of additional advantages to this cross-consortium collaboration. Firstly, it greatly reduces the costs of control genotyping, as common control sets can be used. Secondly, because there is a high degree of genetic overlap in immune-mediated diseases (see section 4.2.3) a high proportion of deep replication SNPs and fine-mapping regions will be associated to multiple diseases, reducing redundancy and increasing the power to detect new shared associations. Finally, because the chip contains almost all known immune-mediated disease loci at time of creation, and because it is being run on a range of different immune-mediated diseases, it makes a perfect platform for performing cross-phenotype analyses of immune diseases.

4.2.2 The content of the Immunochip

The Immunochip is an Infinium iSelect HD custom genotyping chip, manufactured by Illumina. It contains 196,524 variants (largely SNPs, plus 718 small

Immune-mediated diseases		Other diseases
Autoimmune Thyroid Disease[a] (AITD)		Barrett's oesophagus
Ankylosing Spondylitis (AS)		Bipolar Disease (BD)
Bacteraemia susceptibility (BS)		Glaucoma
Crohn's Disease (CD)		Ischaemic stroke
Coeliac Disease (Coeliac)		Parkinson's Disease
IgA deficiency[a] (IgAD)		Pre-eclampsia
Multiple sclerosis (MS)		Psychosis endophenotypes
Primary Biliary Cirrhosis[a] (PBC)		Statin response
Psoriasis (PS)		Reading and mathematics abilities
Rheumatoid arthritis (RA)		Schizophrenia
Sarcoidosis		
Systemic lupus erythematosus (SLE)		
Type 1 Diabetes (T1D)		
Ulcerative colitis (UC)		
Vasculitis		
Visceral leishmaniasis		

Table 4.2: The diseases involved in the Immunochip design [a]Fine-mapping only, no deep replication.

indels), picked specifically for the purpose of discovering and fine-mapping genetic associations with immune-mediated disease. The variants are selected based on three criteria: deep replication of variants implicated by GWAS, fine-mapping of established disease associations and variants submitted as wildcards. In total, approximately 240,000 SNPs were selected for inclusion, with an assay design success rate of ∼80%.

Deep replication

Approximately 50,000 SNPs are included on the Immunochip as deep replication for the diseases shown in Figure 4.2. These SNPs showed suggestive evidence in GWAS, and are intended to be replicated in a large set of samples in order to discover novel associations. Many of these (including all replication for non-immune-mediated traits) were included as part of the second Wellcome Trust Case Control Consortium project. While these SNPs make up only a quarter of the total, they represent the larger proportion of the genome tagged, as they are largely independent (in contrast to the high level of redundancy in the fine-mapping regions).

Fine-mapping regions

A total of 290 established disease associated loci were included on the Immunochip for fine-mapping. 196 of these came from studies that were submitted, accepted or published when the Immunochip was designed (listed in Table 4.3). An additional 94 loci were included on the basis of personal communication with researchers carrying out GWAS and GWAS meta-analyses that were not yet submitted for publication at the time of chip design (listed in Table 4.4). All but one of these studies have now been published. However, many of the fine-mapping loci included were not included in the final publication for these studies. Some of these loci were subsequently discovered in other studies, but there are still 13 “false” loci that are included on the Immunochip and have never been reported in a publication (Table 4.4). Many of these loci are actually true associations; for instance, three of the four “false” IBD loci are confirmed in the IIBDGC Immunochip data (see section 4.2.2).

Fine-mapping regions were defined by taking 0.2cM on either side of the hit SNP, using the combined HapMap2 genetic map. SNPs for fine-mapping were chosen from the 1000 Genomes pilot 1 two-of-three way SNP site set (dated 10/11/2009), and from dbSNP build 130.

The 290 fine-mapping regions include a high degree of overlap. Exactly

Phenotype	Study	Loci
AITD	Kavvoura et al. (2007)	1
AITD	Brand et al. (2009)	1
AS	Burton et al. (2007)	2
BD	Ferreira et al. (2008)	2
BD	O'Donovan et al. (2008)	1
CD	Barrett et al. (2008)	30
CD	Kugathasan et al. (2008)	2
Coeliac	Hunt et al. (2008)	3
Coeliac	Dubois et al. (2010)	27
IgAD	Ferreira et al. (2010)	1
MS	Booth et al. (2008)	3
MS	De Jager et al. (2009)	5
MS	Bahlo et al. (2009)	1
MS	Esposito et al. (2010)	3
MS	Jakkula et al. (2010)	1
MS	McCauley et al. (2010)	2
MS	Mero et al. (2010)	1
PBC	Hirschfield et al. (2009)	1
PS	Capon et al. (2008)	1
PS	Nair et al. (2009)	6
PS	Zhang et al. (2009)	2
RA	Raychaudhuri et al. (2009b)	23
SLE	Harley et al. (2008)	3
SLE	Kozyrev et al. (2008)	1
SLE	Han et al. (2009)	14
SLE	Gateva et al. (2009)	7
T1D	Cooper et al. (2008)	1
T1D	Smyth et al. (2008)	1
T1D	Barrett et al. (2009a)	34
T1D	Qu et al. (2009)	1
T1D	Wallace et al. (2010)	2
UC	Franke et al. (2008)	1
UC	Kugathasan et al. (2008)	2
UC	Imielinski et al. (2009)	1
UC	Asano et al. (2009)	1
UC	Silverberg et al. (2009)	3
UC	Barrett et al. (2009b)	4
UC	Festen et al. (2009)	1

Table 4.3: Fine-mapping regions included on the Immunochip as a result of studies published or submitted at the time of chip design. The "Loci" column gives the total number of fine-mapping regions on the Immunochip from this study.

Disease	Study	On chip	In study (Confirmed)	“False”
AS	Reveille et al. (2010)	4	3 (1^c)	0
AS	Evans et al. (2011)	2	1 (1^d)	0
CD	Franke et al. (2010)	34	32 (1^a)	1
MS	Sawcer et al. (2011)	11	10	1
PS	Strange et al. (2010)	11	9	2
PS	Stuart et al. (2010)	3	2	1
RA	Stahl et al. (2010)	1	2	1
RA	Freudenberg et al. (2011)	1	1	0
SLE	NA^b	10	0^b (3^e)	7
T1D	Swafford et al. (2011)	1	1	0
T1D	Heinig et al. (2010)	1	1	0
UC	Anderson et al. (2011)	15	13 (2^a)	0

Table 4.4: Fine-mapping regions included on the Immunochip as a result of studies that were not completed at the time of chip design. “On chip” is the total number of loci included on the Immunochip from this study, “In study” is the number of these loci that were subsequently included in the final locus list for that study, “Confirmed” is the number of loci that were not included in the study have subsequently been confirmed elsewhere, and “False” is the number of loci included on the Immunochip from this study that have never been published. ^{a}These loci are confirmed in the study described in this chapter, ^{b}I do not believe that this study has been published yet. ^{c}Confirmed by Evans et al. (2011) ^{d}Confirmed by Danoy et al. (2010) ^{d}Confirmed by Guerra et al. (2012)

how many independent regions exists depends on exactly what parameters are used, but merging any regions with boundaries that lie within 50kb of each other, and excluding the two BD regions, gives 186 separate immune-mediated disease regions.

In addition to the regions included due to established associations, a total of 6378 SNPs from across the MHC were included to allow fine-mapping and imputation of HLA alleles.

Wildcard variants

Many groups with the contributing consortia submitted “wildcard” SNPs. Each contributor was given an allocation of SNPs that could be picked based on criteria not directly related to deep replication or fine-mapping.

Many researchers submitted wildcard variants in candidate genes. For instance, the IBD consortium added three SNPs in the gene *XBP1*, implicated as involved in IBD by a functional and candidate gene study (Kaser et al., 2008). The most associated SNP in the original study, rs35873774, had an odds ratio interval of 0.66-0.84 in 4389 cases and 5322 controls. In the 22,442 cases and 30,837 controls of the IIBDGC Immunochip data, it had an odds ratio interval of 0.92-1.02, suggesting that this association is not real, or at least has been overestimated. A more powerful example is an attempted replication via wildcard genotypes of an association between variants in the gene SIAE and autoimmune disease. The original study that reported the association tested 923 cases and 648 controls (Surolia et al., 2010), but an Immunochip-based study in over 60 thousand individuals failed to replicate the results (Hunt et al., 2012). Often candidate gene studies are expensive to replicate, and many false associations are not disproved. These wildcard replication efforts can allow us to confirm or falsify associations that would not be tested in ordinary circumstances.

Some groups submitted candidate SNPs generated from sequencing experiments. For instance Manny Rivas and colleagues submitted 260 low-frequency SNPs that had been identified through resequencing of IBD GWAS regions, many of which replicated successfully in the IIBDGC Immunochip cohort (Rivas et al., 2011).

Other sets of SNPs were added for other purposes. 100 SNPs within the Killer cell Immunoglobulin-like Receptor (KIR) gene cluster were added, to allow development of techniques to impute KIR serological alleles. 1735 Y chromosome SNPs were included to allow Y haplogroup analyses (discussed in section 4.5 below), and a further 848 SNPs were added from the NHGRI GWAS catalogue to allow testing of GWAS hits from non-immune-mediated diseases.

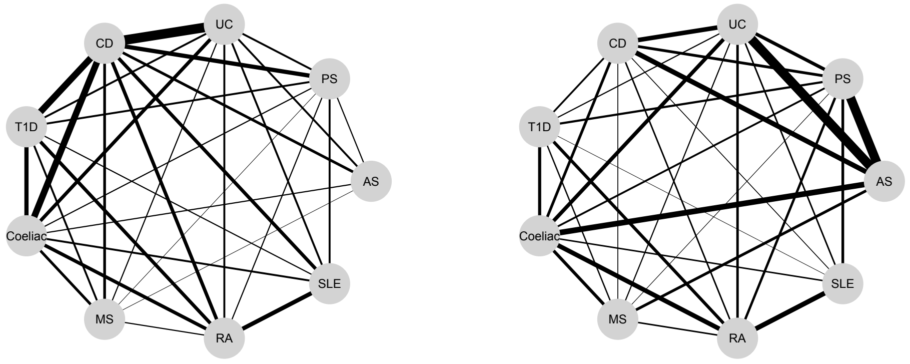
Unpicking “false” IBD fine-mapping regions

There are four IBD fine-mapping regions that were included on the Immunochip despite not appearing in either the Franke et al. (2010) nor Anderson et al. (2011) meta-analysis papers. These include two CD and two UC regions.

In the UC meta-analysis, the first “false” SNP (rs1518070) showed genome-wide significant evidence ($p_{combined}$ = 7.9 x 10^{-9}), leading to its inclusion on the Immunochip. However, final replication did not meet $p_{replication} < 0.05$ due to a high rate of technical failure. The second “false” SNP (rs1569501) showed genome-wide significant evidence of association in the UC GWAS alone, but failed assay design during replication and was thus not included in the final study.

In CD, one “false” SNP (rs1536833) met genome-wide significance in the replication datasets available when the Immunochip was first designed ($p_{combined}$ = 2.6 x 10^{-8}), but dropped just below genome-wide significance when the final replication cohorts were included ($p_{combined}$ = 9.5 x 10^{-8}). The second, rs2098112, showed a significant value of $p_{combined}$ (leading to its inclusion on the Immunochip), but the entire signal was entirely driven by association in the GWAS data, and was excluded from the final list due to lack of signal in the replication.

Of the four IBD fine-mapping regions included “in error”, three were



(a) Unweighted **(b)** Weighted

Figure 4.1: Locus sharing between immune-related diseases, using Immunochip fine-mapping regions. Connecting line width represents number of loci shared, either a) unweighted or b) weighted by square root of the product of the number of associations in both phenotypes.

found to be truly associated in the IIBDGC Immunochip study described in this chapter. The only association that failed to show signal in the Immunochip was rs2098112. Additionally, the improved GWAS imputation described below reduced the association signal from p = 4.5 x 10^{-15} to p = 0.35, showing that this association was driven entirely by poor imputation.

4.2.3 The biology of the Immunochip

The fine-mapping regions on the Immunochip represent a complete survey of the known genetics of immune-mediated disease (or at least, a relatively complete survey of the loci known in mid-2010). What can this list of loci tell us about the shared biology of the diseases that the Immunochip was designed to study?

Locus sharing between immune-mediated diseases

Of the 186 fine-mapping loci, 61 were submitted for more than one disease, including 9 loci shared by at least 4 diseases. Highly shared loci include loci that been traditionally considered important in immunity such as *IL23R*/*IL12RB2* (5 diseases) and *PTPN22* (4 diseases), and other loci that do not have well-understood roles in immunity such as *KIF21B* (5 diseases).

We can use these shared loci to construct a locus sharing network for 9 autoimmune diseases (excluding diseases with 2 or fewer loci). An unweighted network (Figure 4.1a) shows strong connectivity between CD, UC, T1D and Coeliac. However, these diseases are also those with the largest number of discovered loci, so this connectivity is unsurprising. If we weight the network edges by the geometric mean number of associations in the two diseases, we get a very different network (Figure 4.1b). The strongest connection here is between UC and AS (two comorbid diseases).

Network analyses of Immunochip loci

We can place the Immunochip loci in the context of gene networks, and ask which loci seem to play a central role in these networks. I used two gene network tools (GRAIL and DAPPLE) to construct networks using genes inside Immunochip regions. The first, GRAIL (Raychaudhuri et al., 2009a) (Gene Relationships Across Implicated Loci), is a network connectivity tool that uses text mining to calculate a network distance between genes in different implicated loci. Each gene is measured for enrichment of connectivity to genes in other associated loci, and a p-value is calculated. The second, DAPPLE (Rossin et al., 2011) (Disease Association Protein-Protein Link Evaluator), is a network connectivity tool that uses protein-protein interactions. Each gene is measured for enrichment in either direct or indirect (i.e.

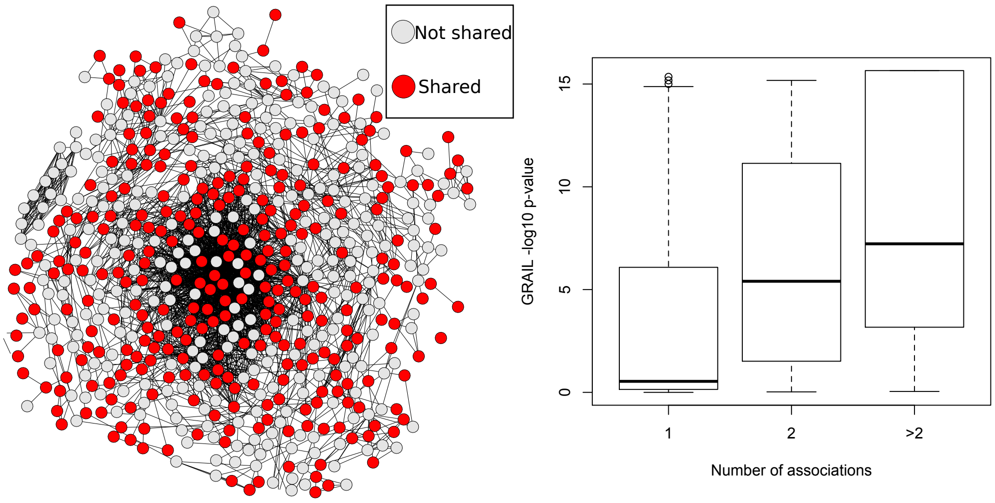


Figure 4.2: The relationship between GRAIL network connectivity and number of associations for Immunochip fine-mapping regions. a) The GRAIL gene network, with genes in shared loci highlighted in red. b) The relationship between GRAIL connectivity p-value and degree of locus sharing

via other proteins) interactions with genes in other loci, and an empirical p-value is calculated by permutation.

Looking at the GRAIL literature network (Figure 4.2), genes that tend to be closest to the centre of the network also tend to be in regions associated with more than one phenotype. In general, there is a correlation between connectivity p-value and number of associations for both GRAIL (Spearman ρ = -0.39, p = 1.45 x 10^{-7}) and DAPPLE (ρ = -0.31, p = 1.15 x 10^{-4}) networks. As intuition might lead us to believe, that loci that play a more central role in the pathways of immune disease are more likely to impact multiple diseases.

The 10 most connected Immunochip fine-mapping loci are shown in Table 4.5. Nine of these regions are associated to more than one disease, though the most significantly connected region, the *TNFSF4* locus, is only associated with SLE. *TNFSF4* (also called *OX40L*) is expressed by dendritic cells and promotes Th2 differentiation and thus humoral immunity, and has

Chrom:Pos (MB)	GRAIL p-value	DAPPLE p-value	Genes	Phenotypes
1:171.4-171.6	3.61 x 10^{-20}	0.23	*TNFSF4*	SLE
1:7.6-8.1	8.32 x 10^{-20}	0.07	*TNFRSF9*	CD, UC, Coeliac
2:204.2-204.5	1.73 x 10^{-19}	<0.002	*ICOS*, *CD28*, *CTLA4*	RA, AITD, T1D, Coeliac
16:28.2-28.9	1.64 x 10^{-18}	<0.002	*IL27*, *NFATC2IP*, *CD19*	CD, T1D
21:44.4-44.5	2.47 x 10^{-18}	0.44	*ICOSLG*	CD, Coeliac
2:191.6-191.7	4.65 x 10^{-18}	<0.002	*STAT4*, *STAT1*	SLE, RA,CD[a],UC[a]
1:67.4-67.7	9.97 x 10^{-18}	<0.002	*IL12RB2*, *IL23R*	PS, CD, UC, AS, PBC
20:44.0-44.2	1.11 x 10^{-17}	0.22	*CD40*	RA, CD[a], UC[a]
3:161.1-161.2	3.03 x 10^{-17}	<0.002	*IL12A*	MS, Coeliac
12:54.6-55.1	3.10 x 10^{-17}	<0.002	*IL23A*, *STAT2*	PS, T1D

Table 4.5: The top 10 most connected Immunochip fine-mapping regions, according to a GRAIL network analysis. [a]New associations discovered in the IBD Immunochip analysis.

been investigated as a drug target in allergic diseases (Wang and Liu, 2007). If this gene were truly associated only to SLE, and not to other immune-mediated diseases, it would suggest a good starting point for investigating deep-rooted differences between immune diseases. However, we can also use the Immunochip itself to investigate this possibility. The SLE-associated SNP, rs1234315, shows a low but sub-genome-wide-significant signal in the Crohn's disease IIBDGC data (p = 2.03 x 10^{-4}), suggesting that this locus is active in other diseases, but has too small an effect size to be reliably detected in GWAS.

Chrom:Pos (MB)	GRAIL p-value	DAPPLE p-value	Genes[b]	Phenotypes
1:199.1-199.3	0.90	0.63	*KIF21B*	MS, AS, UC, CD, Coeliac
2:162.7-163.1	0.38	0.68[c]	*IFIH1*	IgAD, T1D, PS, CD[a], UC[a]
6:90.9-91.1	0.08	0.96	*BACH2*	T1D, Coeliac, CD, UC[a]

Table 4.6: Immunochip fine-mapping regions associated with at least 3 phenotypes, but with no evidence of connection via either DAPPLE or GRAIL. [a]New associations discovered in the IBD Immunochip analysis. [b]The stated genes are the standard candidate genes given the in the literature [c]*IFIH1* is not included in the protein network used by DAPPLE

As well as highlighting highly connected genes, this analysis can also highlight loci that are associated to many different immune-mediated diseases, but do not show evidence of network centrality. Table 4.6 shows three loci that are associated with at least three diseases, but show $p > 0.05$ in both the GRAIL and DAPPLE analyses. One of these genes, *IFIH1*, was not present in the DAPPLE interaction dataset, so may represent a simple lack of data. One of the others, *KIF21B*, was originally discovered in MS, and was believed to act via its role in neuronal transport (McCauley et al., 2010). However, associations to AS, CD, UC and Coeliac disease suggest a more general role in immunity. All three of these regions are associated in IBD, and two contain candidate genes identified by the IBD-specific gene prioritisation approach described in section 4.4.3. *IFIH1* shows a marginal GRAIL association ($p = 0.032$), and *KIF21B* was prioritised by a gene co-expression network approach.

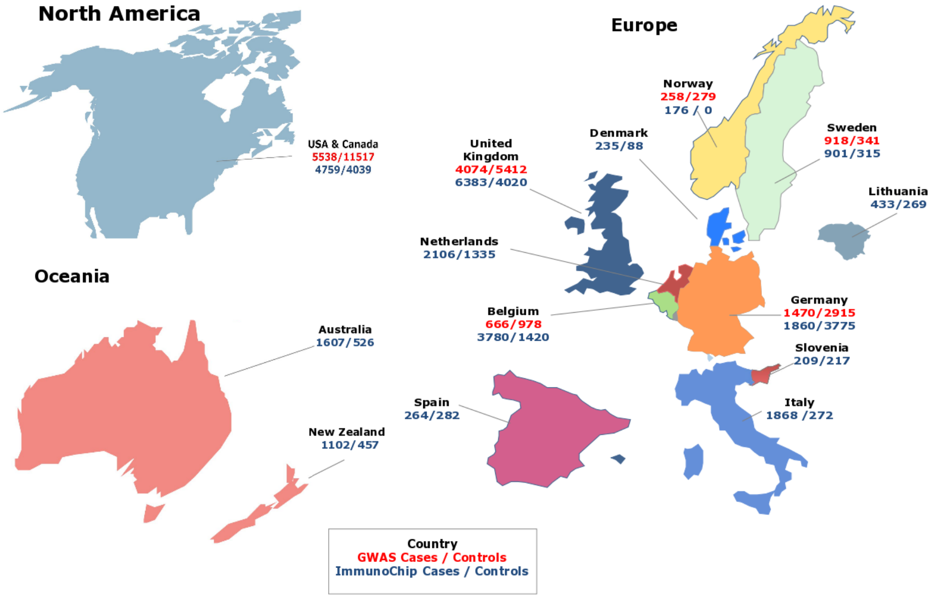


Figure 4.3: Numbers of IBD and control samples passing quality control, from each country participating in this study. The numbers for the Immunochip samples (numbers in blue) only include samples that are not also present in the GWAS (numbers in red).

4.3 QC and association analysis of the IIBDGC Immunochip dataset

4.3.1 The IIBDGC Immunochip dataset

As part of the International IBD Genetics Consortium (IIBDGC), research groups from 15 countries (Figure 4.3) collected Crohn's disease (CD) and Ulcerative colitis (UC) samples and genotyped them using the Immunochip. These data were combined with the GWAS meta-analysis collection to create a large dataset for locus discovery.

The GWAS meta-analysis dataset consists of seven Crohn's disease collections and eight ulcerative colitis collections with genome-wide SNP genotype

Cohort	Countries	Chip	Case / control (unique)
CD cohorts			
BEL1	Belgium, France	ILMN317	513 / 884 (884)
BEL2	Belgium	ILMN317	153 / 94 (94)
CEDARS	USA	ILMN317	835 / 2881 (1364)
CHOP	USA, Canada, Italy, UK	ILMN550	1495 / 6090 (3054)
GERMAN	Germany	ILMN550	480 / 1114 (573)
NIDDK	USA, Canada	ILMN317	759 / 929 (462)
WTCCC	UK	AFFX500	1721 / 2935 (1612)
Total			5956 / 14927 (8043)
UC cohorts			
CEDARS	USA	ILMN317	836 / 2928 (1566)
CHOP	USA, Canada, Italy, Scotland, Canada	ILMN550	664 / 6091 (3038)
GERMANY	Germany	AFFX6	990 / 2915 (2383)
NIDDK1	USA, Canada	ILMN550	498 / 1070 (624)
NIDDK2	USA, Canada	ILMN550	451 / 1428 (1420)
NORWEGIAN	Norway	AFFX6	258 / 279 (279)
SWEDISH	Sweden	ILMN317	918 / 341 (341)
WTCCC	UK	AFFX6	2353 / 5412 (4076)
Total			6968 / 20464 (13727)

Table 4.7: GWAS cohorts, with country of origin, genotyping chip and size. Case and control numbers are after QC, and the number in brackets in the number of unique controls after duplicates between CD and UC have been removed.

data (Table 4.7). The CD cohorts contained a total of 6,299 cases and 15,148 controls, and the UC cohorts contained a total of 7,211 cases and 20,783 controls (the control sets contain largely overlapping samples). Four different chips were used: two produced by Affymetrix (the GeneChip Human Mapping 500K Array and the Genome-Wide Human SNP Array 6.0) and two produced by Illumina (the HumanHap300 BeadChip and the HumanHap550 BeadChip). The majority of these samples were used in the published IIB-

Center	Nationality	CD / UC / control
Bonn	Germany	0 / 0 / 1494
Cedars Sinai	USA	1156 / 822 / 0
Feinstein Institute	Australia	844 / 706 / 464
	Canada	610 / 506 / 305
	New Zealand	422 / 420 / 0
	Netherlands	140 / 157 / 0
	USA	743 / 364 / 2288
	Total	2759 / 2153 / 3057
Kiel	Denmark	66 / 169 / 88
	Germany	1062 / 261 / 1490
	Italy	1273 / 595 / 272
	Lithuania/Baltic	129 / 304 / 269
	New Zealand	260 / 0 / 457
	Norway	122 / 54 / 0
	Spain	264 / 0 / 282
	Sweden	669 / 0 / 0
	Total	3845 / 1383 / 2858
Leuven	Belgium	1434 / 783 / 721
Munich	Germany	0 / 0 / 286
U of Pittsburgh	Australia	0 / 57 / 62
	Canada	0 / 25 / 20
	Germany	0 / 537 / 505
	Netherlands	0 / 327 / 346
	Sweden	0 / 232 / 315
	USA	315 / 218 / 388
	Total	315 / 1396 / 1636
U de Liege	Belgium	1015 / 548 / 699
UMC Groningen	Slovenia	171 / 38 / 217
	Netherlands	1116 / 366 / 989
	Total	1287 / 404 / 1206
UVA	UK	0 / 0 / 2441
Sanger Institute	UK	2952 / 3431 / 1579
Total		14763 / 10920 / 15977

Table 4.8: Immunochip cohorts, broken down by genotyping centre and country of origin. Case and control numbers are after QC, and after samples that overlap the GWAS cohorts have been removed.

DGC meta-analyses (Franke et al., 2010; Anderson et al., 2011).

The Immunochip dataset consists of collections from 15 countries genotyped in 11 different genotyping centres (Table 4.8). Genotyping was performed in 20 batches, with each centre processing between one and three batches. A total of 60,828 samples were genotyped on the Immunochip, comprising 20,076 CD cases, 15,307 UC cases and 25,445 controls. These numbers include many samples that were also present in the GWAS cohorts, which are to be used for fine mapping and not for locus discovery.

Overall, after QC and removing overlapping samples (see below), this dataset has 20,700 CD cases, 17,865 UC cases and 37,747 controls. This is the first time a large meta-analysis has analysed CD and UC together, allowing very high power for variants shared across both phenotypes. For instance, the dataset has an 80% power to detect common IBD associations with an odds ratio greater than 1.06. It is also well-powered to detect low-frequency variants (MAF of 1%) with an odds ratio of >1.35, and rare (MAF = 0.1%) variants with an odds ratio of >2.3.

4.3.2 Genotyping, imputation and quality control

GWAS data

In addition to the quality control performed by individual studies before submission, each GWAS study was subject to the following QC:

1. missing rate per SNP < 0.05
2. missing rate per individual < 0.02
3. heterozygosity per individual ± 0.2
4. missing rate per SNP < 0.02 (after sample removal)

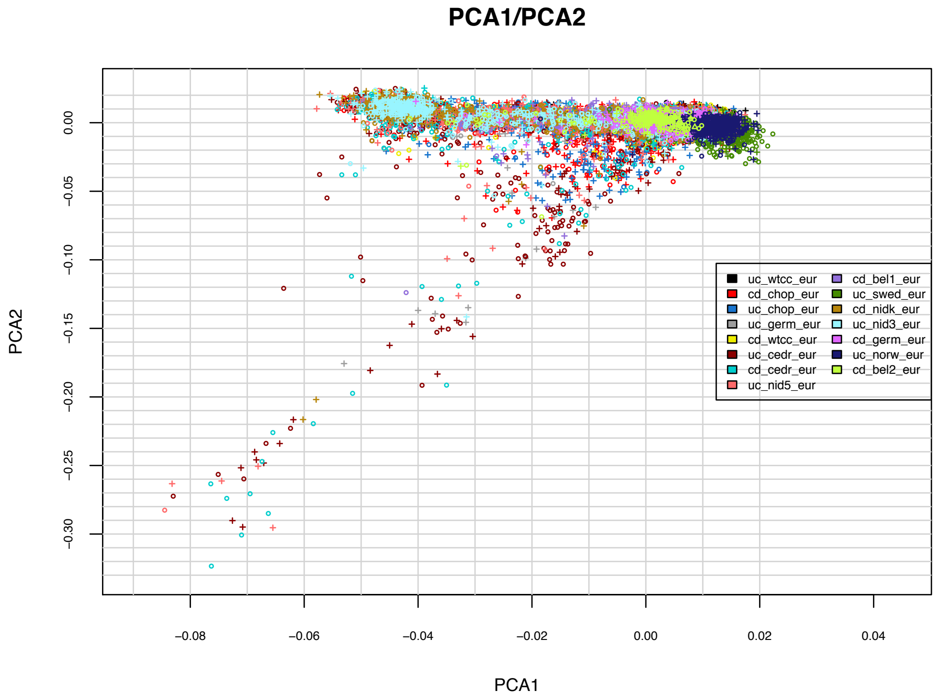


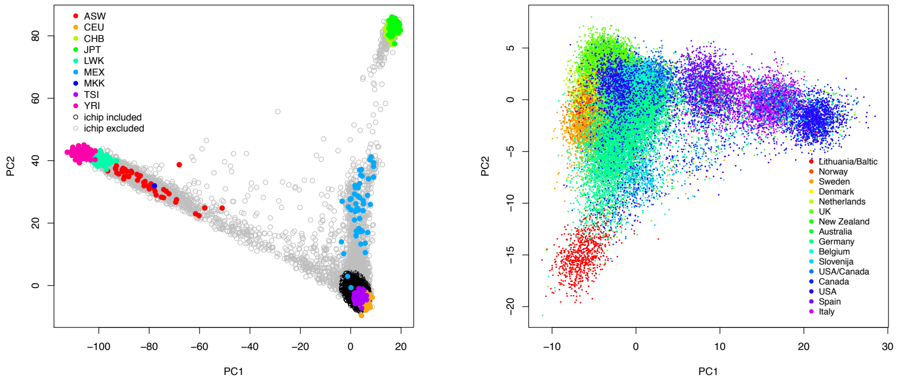
Figure 4.4: All GWAS samples plotted on the first two principal components, coloured by study. Circles are cases, crosses controls

5. missing rate per SNP difference in cases and controls < 0.02

6. Hardy-Weinberg equilibrium (controls) P $< 10^{-6}$

7. Hardy-Weinberg equilibrium (cases) P $< 10^{-10}$.

A set of 17,385 high-frequency SNPs (MAF $> 5\%$) in linkage equilibrium ($r^2 < 0.05$ for all SNP pairs) was generated. Plink was used to calculate relatedness statistics (the estimated coefficient of relatedness $\hat{\pi}$), and individuals with $\hat{\pi} > 0.2$ to another sample were removed. Samples duplicated between CD and UC control datasets were recorded: these samples are kept in for single-phenotype tests, but removed for combined tests. Principal component analysis was performed (Figure 4.4), and principal components that



(a) Continental PCA with ethnic outliers

(b) Within Europe PCA by country

Figure 4.5: a) Principal component projection of Immunochip samples onto a) continental axes fitted from HapMap samples and b) European axes fitted from Immunochip controls

correlated with disease phenotype were recorded for use as covariates.

Imputation of untyped SNPs was performed within each study in batches of 300 individuals. These batches were randomly drawn in order to keep the same case-control ratio as in the total sample from that study. Imputation was performed with the CEU+TSI HapMap3 reference set (containing 1,252,901 polymorphic SNPs), using Beagle 3.13 with a chunk size of 10Mb and default parameters.

Immunochip data

Because many of the variants on Immunochip do not meet the manufacturer's quality standards set for GWAS products, rigorous QC is essential. Furthermore, because samples with poor quality DNA or with other genome-wide problems can adversely affect the genotype calls at high quality samples, I performed a first stage of "coarse" QC on genotypes called using Illumina's GenomeStudio program. I exclude samples with >5% missing data, genome-wide heterozygosity outside a 95% confidence interval in each batch, samples of non-European ancestry (via PCA, see below) or with abnormal mean intensity values from further analysis.

For all remaining samples, I used the optiCall clustering program (Shah et al., 2012) (v0.3.0) to call genotypes, with a no-call cutoff of 0.7 and HWE blanking disabled. I identified duplicate and related samples ($\hat{\pi} > 0.1$) using PLINK with the same set of SNPs used for PCA (details below for details), and removed the duplicate or related sample with the higher missing data rate. I used a set of 692 SNPs present on both the Immunochip and all four GWAS chips to remove Immunochip samples that were also present in the GWAS. I removed samples without a phenotype definition of Crohn's disease, ulcerative colitis or healthy control, and finally removed all samples with $>$ 2% missing data in this improved call-set.

I performed SNP QC in this filtered dataset, removing SNPs with >2% missing data or HWE p-value $< 10^{-10}$ in controls. However, a relatively large number of SNPs still showed poor clustering, driving many false positive associations. To further ensure the quality of genotype calls in our analysis, I selected 3,356 variants for manual inspection, including those with meta analysis p$<10^{-5}$ which fulfilled at least one of the following criteria:

1. Cochran heterogeneity $p < 0.01$ between GWAS and Immunochip (N=871)

2. lie outside fine-mapping regions known to be associated with immune-mediated disease (N=797)

3. are one of the 3 most significantly associated SNPs in a region (N=851)

4. any SNP with $p < 5\text{x}10^{-8}$ which did not fit those criteria (N=195)

5. random SNPs as a comparator (N=642)

I distributed intensity data for these SNPs to 16 members of the IIBDGC for manual inspection. Included was a version of the manual inspection program Evoker (Morris et al., 2010) optimised for multi-cohort inspection, and a document describing the protocol for manual inspection. Each SNP was inspected by three individuals, and was considered to have passed inspection if three individuals passed the SNP, or two passed it and one marked it as a "Maybe". 1015 SNPs were removed during this process. A further 29 SNPs had genotypes manually adjusted (blind to phenotype and association statistics) to correct recoverable errors.

I used principal component analysis to identify ethnic outliers, and to generate covariates to control for population stratification. To identify outliers on the continental scale I constructed a reference set consisting of 662 HapMap founder samples genotyped on the Illumina Human1M, the Affymetrix Human SNP Array 6.0, and the Illumina Omni2.5 for the HapMap3 and 1000 Genomes Projects. This reference set was designed to maximise overlap with the Immunochip, and has a total of 3,268,731 SNPs, of which 83,689 are present on the Immunochip. I used PLINK to LD prune the data such that no pair of SNPs had $r^2 > 0.2$, and I also removed GC/AT SNPs, SNPs within known high LD regions (Price et al., 2008) and SNPs with MAF $< 5\%$. I projected the Immunochip samples on the principal component axes generated using these 17,891 SNPs from the 662 reference samples using the R package snpMatrix (Clayton and Leung, 2007). All samples that did not cluster with the European samples were excluded (Figure 4.5a).

To resolve within-Europe relationships, I performed PCA within the remaining Immunochip samples. LD pruning was performed within European

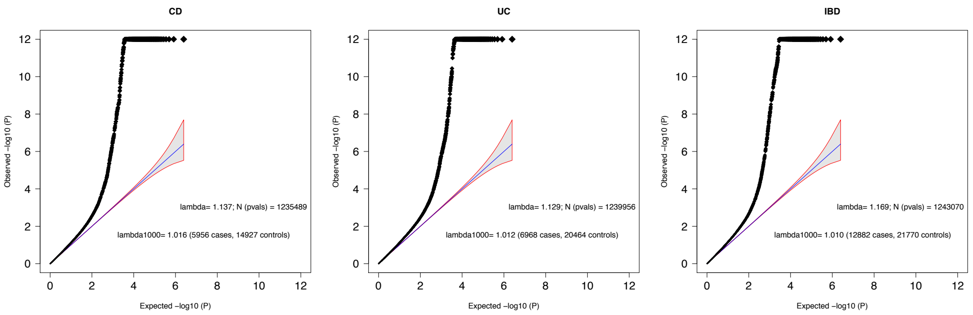


Figure 4.6: QQ plots, λ and λ_{1000} values for the CD, UC and IBD GWAS analyses. Grey shapes show 95% confidence interval under the null.

controls (this was performed three times, to properly break up the LD in fine-mapping regions), and SNPs present in high LD regions or with MAF $< 5\%$ were removed, leaving a total of 19,111 SNPs. I generated principal component axes within the controls, and projected the cases onto these axes to generate PCs for all samples. The first four principal component axes seemed to capture significant population structure (Figure 4.5b), and addition of components beyond the fourth as association covariates in a subset of the Immunochip data did not further reduce the genomic inflation factor.

4.3.3 Association analyses

4.3.4 GWAS and Immunochip analyses

Three association scans were performed for both GWAS and Immunochip. These included a CD analysis (Crohn's disease vs controls), a UC analysis (ulcerative colitis vs controls) and an IBD analysis (combined CD and UC vs controls).

For the GWAS, the CD scan had a total of 5,956 QC+ cases and 14,927 QC+ controls, the UC scan had 6,968 cases and 20,464 controls, and the IBD scan had 12,882 cases and 21,770 controls. For the IBD scan, controls

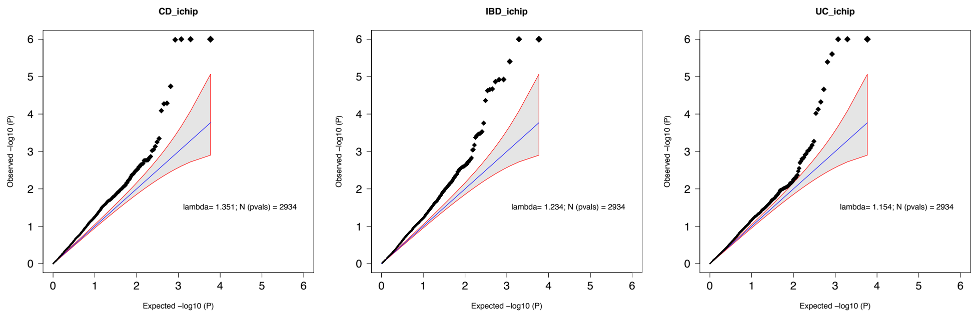


Figure 4.7: QQ plots and λ values for the CD, UC and IBD Immunochip analyses. Grey shapes show 95% confidence interval under the null.

that overlapped between the CD and UC control cohorts were removed from whichever dataset had a greater excess of controls. Association testing was carried out in PLINK, using the dosage data from the imputation and using 10, 7, 15 principal components for CD, UC, IBD respectively as covariates (all PCs that correlated with case-control status). The CD, UC and IBD scans had genomic inflation (λ_{GC}) values of 1.137, 1.129, and 1.169 respectively (Figure 4.6). These inflation figures are substantially lower than the figures for the previous CD and UC meta-analyses.

For the Immunochip analysis, the CD, UC and IBD scans all used the entire control dataset. The CD scan had a total of 14,763 QC+ cases, the UC scan had 10,920 cases, the IBD scan had 25,683 cases, and all scans used the 15,977 QC+ controls. I performed association testing using additive logistic regression in PLINK conditioned on the first four principal components. Test statistic inflation was computed from a set of 3120 SNPs chosen based on GWAS of schizophrenia, psychosis and reading/mathematics ability. Genomic inflation factors were relatively low, given the large sample size and presence of polygenic risk: $\lambda_{GC_{CD}} = 1.353$, $\lambda_{GC_{UC}} = 1.154$, $\lambda_{GC_{IBD}} = 1.234$ (Figure 4.7).

For comparison, I also performed an association test on all IBD samples using the Cochran-Mantel-Haenszel method to stratify by country of origin of the samples. This is one of the standard methods used to analyse GWAS replication data, where population stratification correction via principal components are usually not available. The genomic inflation value for the IBD all analysis was $\lambda_{GC_{IBD}} = 2.00$, showing that without the genome-wide SNP data on the Immunochip this replication analysis would have shown severe inflation.

This also has some worrying implications for the GWAS field, as it suggests that most standard international replication datasets will suffer from test statistic inflation. This in turn could mean that combined GWAS-replication p-values may be too liberal. In the future, it seems prudent that large replication analyses should include a number of ancestry-informative SNPs to control for stratification. Exactly how many such SNPs would be required to reduce inflation is unknown, and the Immunochip provides a platform to investigate this.

4.3.5 Deep replication meta-analysis

A combined analysis was performed using both the GWAS and the Immunochip association results comprising 20,700 Crohn's disease, 17,865 ulcerative colitis cases and 37,747 healthy controls.

All SNPs in GWAS association results with $p < 0.01$ in the CD, UC or IBD scans were selected for replication in the Immunochip dataset (a total of 25,075 SNPs). A fixed-effect meta-analysis was performed using odds ratios and standard errors from the GWAS hit and the Immunochip tag with the highest r^2 to the hit SNP, providing a tag with $r^2 > 0.4$ was available. The Cochran heterogeneity p-value was also calculated (none of the final association signals showed significant heterogeneity after correcting for multiple testing).

SNPs with p $<$ 5 x 10^{-8} in any of the three phenotypes in this analysis were combined into clumps if they had $r^2 > 0.1$. SNPs within these clumps were tested for evidence of association independent of the strongest signal in the clump. Because the tag SNP meta-analysis approach makes standard methods for conditional analysis impossible to carry out, so we used an approximate conditional Z-score

$$Z_i' = Z_i - r_{i,hit} Z_{hit} \tag{4.1}$$

Where Z_i is the Z score of the SNP being tested, Z_{hit} is the Z score of the strongest signal in the clump, and $r_{i,hit}$ is the correlation coefficient between the strongest signal and the SNP being tested. If P($Z_i' > 0$) $<$ 5 x 10^{-8} then this clump is considered to have a secondary signal, and the SNP with the Z_i' largest in magnitude is recorded as a secondary signal in this clump. All other SNPs in the clump are then tested for a tertiary signal independent of the first two, using

$$Z_i' = Z_i - r_{i,hit} Z_{hit} - r_{i,2nd} Z_{2nd} \tag{4.2}$$

We do not test for additional signals after the third. Theoretically, this could be extended to an arbitrary number of signals, but the approximation will become less accurate as additional signals are tested for.

This approach yielded 193 genome-wide significant independent signals of association. None of these signals had significant heterogeneity of effect size, and all had their Immunochip intensity cluster plots manually inspected to ensure that they were well clustered.

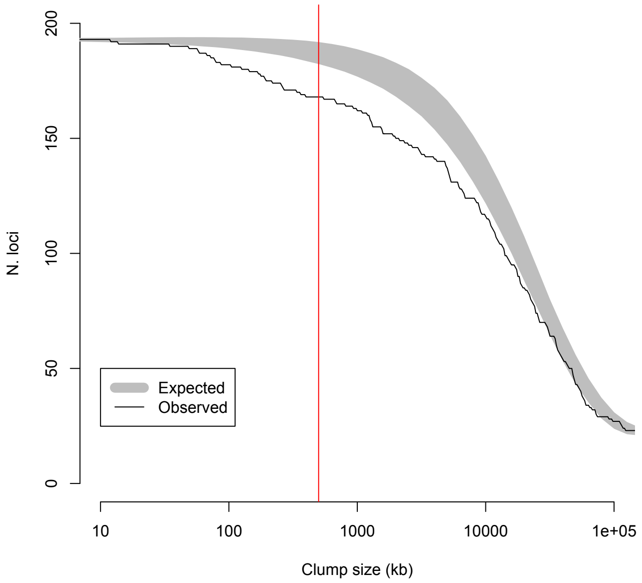


Figure 4.8: The results of a null simulation of association clumping. The x-axis shows varying thresholds of proximity for two statistically independent signals to be considered in the same locus. The y-axis shows the number of loci for a particular threshold, from 193 (the total number of independent signals) at the left when no signals are combined to fewer than 50 when even extremely distant signals 100Mb apart are combined. The grey shaded area shows the 95% confidence interval from simulations of 193 random signals, and the black line shows the true number of loci for a given clumping value. The red line is 500kb, the actual clumping distance we used.

4.3.6 Combining signals into loci

The large number of independent signals (193) makes categorising them into functionally separate loci problematic. We conventionally define signals as coming from the same locus if their lead SNPs lie within a certain physical or genetic distance of each other. However if this physical distance parameter is too large functionally independent signals that are adjacent by chance may be incorrectly combined. Conversely, selecting too small a distance parameter could cause variants that act relatively proximately on the same gene to be split into independent loci.

To test the effect of this distance parameter on classifying signals into loci, I performed a null simulation. I selected randomly from the PCA SNPs to simulate null signals, and examined what proportion of signals are incorrectly merged together for a given distance parameter value. Based on this, I decided to define a locus as a 500kb unit: 250kb on either side of the hit SNP. This results in between 95% and 99% of null loci being correctly separated (Figure 4.8).

Each independent signal had a region defined around it, which was 250kb on either side of the hit SNP, or the extent of LD (defined as the positions of the furthest up-and-downstream variants with $r^2 > 0.5$ to the hit SNP). Overlapping regions were merged together, providing that they were associated to compatible phenotypes under the likelihood analysis (see below); i.e. loci were not merged if one was uniquely associated with CD, and the other uniquely associated with UC. The final merged regions were defined as loci, with their extents being the maximum extent of their component signals. A total of 163 independent loci were thus defined (Table 4.9).

4.3.7 Crohn's disease/Ulcerative colitis likelihood modelling

We used a likelihood modelling approach to classify signals into four categories according to their relative strength of association to CD and UC. We used a multinomial logistic regression model with additive log-odds ratio parameters β_{CD} and β_{UC}. The model was fitted to the Immunochip genotypes using the `mlogit` package in R.

We fit this model with four sets of parameter constraints:

1. CD-specific model: $\beta_{UC} = 0$ (i.e. UC cases and controls have the same frequency), β_{CD} fitted by maximum likelihood

2. UC-specific model: $\beta_{CD} = 0$, β_{UC} fitted by maximum likelihood

3. IBD unsaturated (same-effect size) model: $\beta_{CD} = \beta_{UC} = \beta_{IBD}$ (i.e. frequency is the same in CD and UC cases), β_{IBD} fitted by maximum likelihood

4. IBD saturated (different effect sizes) model: β_{CD} and β_{UC} both fitted by maximum likelihood

Note that models 1-3 are all constrained versions (1 d.f.) of model 4 (2 d.f.).

We calculated likelihoods for each model, and performed a likelihood ratio test of each of models 1-3 against model 4. If the likelihood ratio test had p < 0.05 for all 3 models (the 2 d.f. model is nominally significantly a better fit than any of the 1 d.f. models), we classified the signal as "saturated" (i.e. associated to both CD and UC, but with evidence of different effect sizes). Otherwise, we classified the signal according to which of the first three models had the largest likelihood. Note that being classified as IBD unsaturated should be interpreted as "associated to both CD and UC, without significance evidence of differing effect sizes".

In Table 4.9 below, the "IBD" section contains all loci where the main signal was classified as IBD unsaturated or IBD saturated. An exception was made for the CD associations at *PTPN22* and *NOD2*, where the correct model was "IBD saturated", as there were significant UC associations that went in the opposite direction to the CD effect.

Even within these classifications there is a significant variation in the balance of CD and UC effect sizes (Figure 4.9). To capture this we also used polar-transformed log odds ratios as a continuous measure of CD vs UC effect size balance. This is defined as $\theta = \text{atan2}(\log(\text{ORCD}), \log(\text{ORUC}))$. Large values of θ correspond to associations with a stronger UC component, smaller values correspond to a stronger CD component.

4.3.8 Comparison of this locus list to previous CD and UC lists

Because this study has access to raw genotype data from both CD and UC for the first time, it has allowed us to clarify several aspects of the 99 previously reported associations:

- While previously suspected, we have confirmed that the associations in the MHC are distinct for CD and UC, and therefore should be split into two phenotype specific associations, rather than a single IBD locus.

- Conversely our improved imputation has re-localised the CD association previously reported as *VAMP3* to be the same effect as the adjacent previous UC association to *TNFRSF9*, making this a single IBD locus.

- Two previously independent associations on chromosome 2 near 102Mb (one CD, one UC) have both been shown to be IBD, and accordingly have been merged into independent effects in a single IBD locus. Similarly, a previous CD SNP (chromosome 2 near 198Mb), which is now associated to UC as well, was incorporated into a new nearby UC locus.

- Five previous associations (Chr2@198Mb, Chr5@36Mb, Chr6@3Mb, Chr6@44Mb, Chr13@42Mb) are no longer genome-wide significant. In four cases, our improved PCA-corrected analysis is >2 orders of magnitude less significant than the previous country-stratified analysis, suggesting that these associations may have been driven in part by uncorrected population structure. In the final instance the key SNP failed Immunochip design.

Thus, from 99 previously reported loci, one was split, three were merged and five were lost, leaving 92 established and 71 novel loci. This highlights both the overall robustness of our previous analyses as well as potential pitfalls in small-scale replication genotyping, for which correction for population stratification is difficult.

We also compared the total phenotypic variance of CD and UC explained by our loci compared to previously published estimates. In ulcerative colitis we improved from 3.9% of phenotypic variance explained by known loci to 7.0% explained by our 193 signals. For Crohn's disease we improved from 7.6% to 12.0%. Two additional comments are necessary: first, I have decided here to report phenotypic variance explained, rather than heritability, due to the difficulties in measuring narrow-sense heritability discussed in Chapter 2. Second, the odds ratios estimated from the Immunochip are smaller than previous estimates for several key loci in CD, including *NOD2*, *IL23R* and *ATG16L1*. This difference was not explained by an abnormal degree of stratification or differential ancestry at these sites. Our new odds ratios are estimated in replication samples in this project, so this effect may reflect less severe disease than the samples previously collected for GWAS.

Table 4.9: The 163 inflammatory bowel disease loci, split into Crohn's Disease specific, Ulcerative Colitis specific, and shared across Inflammatory Bowel Disease. Key genes are those identified by one of the candidate gene prioritisation analyses described in the text, and bold genes are identified by more than one bioinformatic approach. Loci shaded grey are newly identified in this study. SNP IDs marked * denote the presence of a second genome-wide significant alternative signal at this locus, and ** denotes the presence of two or more additional signals. Odds ratios marked with a †show evidence of heterogeneity of effect size between CD and UC.

Chrom:Pos (Mb)	SNP	P-value	RAF	OR	Key Genes (+N additional in locus)
Crohn's Disease					
1:78.37-78.87	rs17391694	2.96 x 10^{-9}	0.889	1.134	(5)
1:114.05-114.55	rs6679677	2.03 x 10^{-15}	0.907	1.196†	*PTPN22*, *DCLRE1B*, (7)
1:120.2-120.7	rs3897478	1.97 x 10^{-11}	0.891	1.161	*ADAM30*, (5)
1:172.6-173.1	rs9286879	5.53 x 10^{-22}	0.249	1.125	***FASLG***, *TNFSF18*, (0)
2:27.38-27.88	rs1728918	4.86 x 10^{-16}	0.299	1.123	*UCN*, (23)
2:62.3-62.8	rs10865331	9.77 x 10^{-10}	0.396	1.098	(3)
2:230.84-231.34	rs6716753	1.17 x 10^{-16}	0.196	1.134	*SP140*, (5)
2:233.87-234.42	rs12994997	4.14 x 10^{-70}	0.523	1.233	***ATG16L1***, *INPP5D*, (7)
4:48.11-48.61	rs6837335	1.75 x 10^{-8}	0.647	1.086	*TXK*, *TEC*, *SLC10A4*, (3)
4:102.61-103.11	rs13126505	1.84 x 10^{-12}	0.096	1.172	(1)
5:55.18-55.68	rs10065637	3.68 x 10^{-12}	0.773	1.123	***IL6ST***, *IL31RA*, (1)
5:72.29-72.79	rs7702331	5.63 x 10^{-10}	0.621	1.088	(4)

5:173.09-173.59	rs17695092	4.68 x 10^{-9}	0.703	1.095	*CPEB4*, (2)
6:21.17-21.67	rs12663356	4.01 x 10^{-12}	0.533	1.095	(3)
6:31.02-31.52	rs9264942	4.96 x 10^{-28}	0.378	1.145	*HLA-C*, *PSORS1C1*, *NFKBIL1*, (19)
6:127.2-127.7	rs9491697	3.79 x 10^{-10}	0.439	1.077	(3)
6:127.99-128.49	rs13204742	8.38 x 10^{-15}	0.124	1.173	(2)
6:159.24-159.74	rs212388	3.04 x 10^{-14}	0.410	1.105	*TAGAP*, (5)
7:26.63-27.13	rs10486483	3.48 x 10^{-8}	0.247	1.089	(2)
7:27.92-28.42	rs864745	3.65 x 10^{-9}	0.497	1.087	*CREB5*, *JAZF1*, (1)
8:90.62-91.12	rs7015630	1.42 x 10^{-8}	0.739	1.075	*RIPK2*, (4)
8:129.31-129.81	rs6651252	1.45 x 10^{-16}	0.865	1.185	(0)
13:44.2-44.7	rs3764147	2.19 x 10^{-21}	0.248	1.155	*LACC1*, (3)
15:38.64-39.14	rs16967103	3.88 x 10^{-9}	0.203	1.088	***RASGRP1***, *SPRED1*, (2)
16:50.31-51	rs2066847**	5.86 x 10^{-209}	0.024	3.103†	***NOD2***, *ADCY7*, (5)
17:25.59-26.09	rs2945412	8.68 x 10^{-17}	0.587	1.137	*LGALS9*, *NOS2*, (3)
19:0.87-1.37	rs2024092	8.26 x 10^{-22}	0.215	1.156	*GPX4*, *HMHA1*, (20)
19:46.6-47.1	rs4802307	2 x 10^{-10}	0.706	1.099	(9)
19:48.95-49.45	rs516246	1 x 10^{-15}	0.483	1.107	*DBP*, *SPHK2*, *IZUMO1*, *FUT2*, (22)
21:34.52-35.02	rs2284553	2.14 x 10^{-16}	0.599	1.123	***IFNGR2***, ***IFNAR1***, *IFNAR2*, *IL10RB*, (9)

Ulcerative Colitis					
1:2.25-2.75	rs10797432	2.62×10^{-12}	0.522	1.078	***TNFRSF14***, *MMEL1*, *PLCH2*, (8)
1:19.88-20.42	rs6426833**	2.39×10^{-68}	0.542	1.265	(9)
1:199.84-200.34	rs2816958	1.98×10^{-17}	0.887	1.23	(3)
2:198.18-199.12	rs1016883	2.87×10^{-8}	0.817	1.1	*RFTN2*, *PLCL1*, (7)
2:199.27-200.12	rs17229285*	1.73×10^{-13}	0.496	1.117	(0)
3:52.8-53.3	rs9847710	1.05×10^{-8}	0.416	1.064	*PRKCD*, *ITIH4*, (8)
4:103.26-103.76	rs3774959	3.66×10^{-12}	0.358	1.118	***NFKB1***, *MANBA*, (2)
5:0.34-0.84	rs11739663	1.81×10^{-8}	0.760	1.071	*SLC9A3*, (8)
5:134.19-134.69	rs254560	2.55×10^{-9}	0.397	1.056	(6)
6:32.33-32.86	rs6927022	4.71×10^{-133}	0.535	1.444	***HLA-DQB1***, *-DRB1*, -DQA1 (13)
7:2.53-3.03	rs798502	6.09×10^{-17}	0.709	1.127	***CARD11***, *GNA12*, *TTYH3*, (4)
7:26.97-27.47	rs4722672	2.06×10^{-8}	0.183	1.091	(14)
7:107.18-107.72	rs4380874*	2.07×10^{-26}	0.405	1.137	*DLD*, (9)
7:128.32-128.82	rs4728142	4.37×10^{-14}	0.444	1.104	***IRF5***, *TNPO3*, *TSPAN33*, (11)
11:95.77-96.27	rs483905	1.21×10^{-8}	0.292	1.056	*JRKL*, *MAML2*, (2)

11:114.13-114.63	rs561722	5.15 x 10^{-17}	0.663	1.12	*FAM55A*, *FAM55D*, (5)
15:41.29-41.81	rs28374715	2.43 x 10^{-8}	0.738	1.082	*ITPKA*, *NDUFAF1*, *NUSAP1* ,(8)
16:30.22-30.72	rs11150589	6.04 x 10^{-10}	0.463	1.09	***ITGAL***, (20)
16:68.33-68.83	rs1728785	3.71 x 10^{-8}	0.767	1.075	*ZFP90*, (6)
17:70.39-70.89	rs7210086	1.89 x 10^{-9}	0.797	1.111	(3)
19:46.87-47.37	rs1126510	1.55 x 10^{-9}	0.363	1.075	*CALM3*, (14)
20:33.55-34.05	rs6088765	2.21 x 10^{-8}	0.437	1.079	*PROCR*, *UQCC*, *CEP250*, (8)
20:42.81-43.31	rs6017342	1.43 x 10^{-43}	0.530	1.228	*ADA*, *HNF4A*, (9)
Inflammatory Bowel Disease					
1:0.99-1.49	rs12103	7.66 x 10^{-13}	0.182	1.099	*TNFRSF18*, *TNFRSF4*, (30)
1:7.77-8.27	rs35675666	1.12 x 10^{-15}	0.838	1.112	*TNFRSF9*, (6)
1:22.45-22.95	rs12568930	1.26 x 10^{-17}	0.821	1.095†	(3)
1:67.4-67.95	rs11209026**	8.12 x 10^{-161}	0.933	2.013†	***IL23R***, *IL12RB2*, (4)
1:70.74-71.24	rs2651244	2.29 x 10^{-8}	0.599	1.015†	(3)
1:151.54-152.04	rs4845604	3.52 x 10^{-16}	0.857	1.144†	*RORC*,(14)
1:155.22-156.12	rs670523	5.79 x 10^{-11}	0.324	1.06†	*UBQLN4*, *IT1*, *STO1*,(28)
1:160.6-161.1	rs4656958	6.8 x 10^{-9}	0.686	1.061	***CD48***, *SLAMF1*, *ITLN1*, *CD244*, (12)

1:161.22-161.72	rs1801274	2.12 x 10^{-38}	0.509	1.124†	***FCGR2A***, ***FCGR2B***, ***FCGR3A***, **HSPA6** (11)
1:197.33-197.87	rs2488389	8.45 x 10^{-13}	0.220	1.115	*C1orf53*,(2)
1:200.62-201.12	rs7554511	1.24 x 10^{-32}	0.725	1.164	*KIF21B*,(6)
1:206.68-207.18	rs3024505	6.66 x 10^{-42}	0.160	1.208†	***IL10***, ***IL20***, ***IL19***, ***IL24***, (7)
2:24.87-25.37	rs6545800	6.14 x 10^{-16}	0.445	1.109†	*ADCY3*, (6)
2:28.36-28.86	rs925255	2.67 x 10^{-15}	0.557	1.092†	*FOSL2*, *BRE*, (1)
2:43.56-44.06	rs10495903	8.03 x 10^{-12}	0.130	1.086†	(5)
2:60.95-61.45	rs7608910	8.65 x 10^{-32}	0.394	1.138	***REL***, *C2orf74*, *KIAA1841*, *AHSA2*, (6)
2:65.42-65.92	rs6740462	2.35 x 10^{-8}	0.739	1.081	*SPRED2*, (1)
2:102.41-103.31	rs917997*	3.12 x 10^{-20}	0.231	1.103†	*IL1R2*, *IL18RAP*, *IL18R1*, *IL1R1*, (5)
2:162.85-163.35	rs2111485	1.93 x 10^{-8}	0.404	1.066	*IFIH1*, (5)
2:191.67-192.17	rs1517352	3.28 x 10^{-11}	0.600	1.077	***STAT1***, *STAT4*, (2)
2:218.89-219.39	rs2382817	3.7 x 10^{-12}	0.408	1.073	***SLC11A1***, ***CXCR2***, ***CXCR1***, *PNKD*, (11)
2:241.31-241.83	rs3749171*	3.07 x 10^{-21}	0.167	1.135†	***GPR35***, (12)
3:18.51-19.01	rs4256159	9 x 10^{-15}	0.140	1.107†	(0)

3:47.96-49.96	rs3197999**	1.01 x 10^{-47}	0.296	1.18	***MST1***, ***PFKFB4***, *MST1R*, *UCN2*, (61)
4:74.6-75.1	rs2472649	2.57 x 10^{-8}	0.824	1.095†	***CXCL5***, ***CXCL1***, ***CXCL3***, *IL8*, (7)
4:122.91-123.53	rs7657746	2.76 x 10^{-13}	0.753	1.116	***IL2***, *IL21*, (2)
5:10.44-10.94	rs2930047	1.03 x 10^{-8}	0.382	1.065	***DAP***, (2)
5:40.02-40.74	rs11742570**	1.81 x 10^{-82}	0.605	1.198†	*PTGER4*, (1)
5:95.99-96.49	rs1363907	5.62 x 10^{-13}	0.411	1.068	*ERAP2*, *ERAP1*, *LNPEP*, (2)
5:129.75-130.26	rs4836519	4.24 x 10^{-10}	0.768	1.072†	(1)
5:130.36-132.01	rs2188962*	1.35 x 10^{-52}	0.425	1.158†	***IRF1***, ***IL13***, ***CSF2***, ***SLC22A4***, (14)
5:141.26-141.76	rs6863411	3.59 x 10^{-14}	0.630	1.089†	*SPRY4*, *NDFIP1*, (5)
5:150.02-150.52	rs11741861	2.94 x 10^{-37}	0.093	1.249†	*TNIP1*, *IRGM*, *ZNF300P1*, (8)
5:158.53-159.07	rs6871626**	1.43 x 10^{-42}	0.337	1.181†	***IL12B***, (3)
5:176.54-177.04	rs12654812	1.68 x 10^{-8}	0.335	1.068	***DOK3***, (17)
6:14.46-14.96	rs17119	3.08 x 10^{-11}	0.786	1.071	(0)
6:20.47-21.06	rs9358372*	8.66 x 10^{-14}	0.379	1.089†	(2)
6:90.71-91.21	rs1847472	1.57 x 10^{-10}	0.655	1.06	(1)
6:106.18-106.68	rs6568421	8.24 x 10^{-20}	0.301	1.108†	(2)

6:111.55-112.09	rs3851228	1.08 x 10^{-13}	0.073	1.153	***TRAF3IP2***, *FYN*, *REV3L*, (2)
6:137.75-138.25	rs6920220	1.4 x 10^{-21}	0.206	1.102†	*TNFAIP3* ,(1)
6:143.65-144.15	rs12199775	1.99 x 10^{-8}	0.929	1.129	*PHACTR2*, (5)
6:167.12-167.62	rs1819333	6.76 x 10^{-21}	0.523	1.081†	***CCR6***, ***RPS6KA2***, *RNASET2*, (3)
7:49.94-50.55	rs1456896*	7.28 x 10^{-15}	0.688	1.088	*ZPBP*, *IKZF1*, (4)
7:98.5-99	rs9297145	8.21 x 10^{-12}	0.265	1.082	*SMURF1*, (6)
7:100.06-100.61	rs1734907	1.67 x 10^{-13}	0.149	1.114†	***EPO***, (21)
7:116.64-117.14	rs38904	1.31 x 10^{-8}	0.532	1.054†	(6)
8:126.28-126.78	rs921720	8.3 x 10^{-20}	0.609	1.081†	*TRIB1*, (1)
8:130.37-130.87	rs1991866	1.65 x 10^{-9}	0.422	1.054	(2)
9:4.73-5.23	rs10758669	7.88 x 10^{-45}	0.349	1.174	***JAK2***, (4)
9:93.67-94.17	rs4743820	3.6 x 10^{-9}	0.702	1.056†	*NFIL3*, (2)
9:117.3-117.89	rs4246905**	2.8 x 10^{-32}	0.709	1.142	*TNFSF8*, *TNFSF15*, *TNC*, (2)
9:138.99-139.64	rs10781499*	4.38 x 10^{-56}	0.412	1.188†	***CARD9***, *PMPCA*, *SDCCAG3*, (19)
10:5.83-6.33	rs12722515	3.76 x 10^{-10}	0.849	1.102†	***IL2RA***, ***IL15RA***, (6)
10:30.47-30.97	rs1042058	5.93 x 10^{-11}	0.592	1.075†	***MAP3K8***, (3)
10:35.04-35.55	rs11010067	2.49 x 10^{-25}	0.346	1.115†	***CREM***, (3)
10:59.74-60.24	rs2790216	8.07 x 10^{-9}	0.778	1.066	*CISD1*, *IPMK*, (2)
10:64.12-64.89	rs10761659**	6.37 x 10^{-46}	0.543	1.166†	(3)
10:75.42-75.92	rs2227564	6.75 x 10^{-10}	0.770	1.082†	(13)

10:80.78-81.28	rs1250546	3.15 x 10^{-18}	0.604	1.096†	(5)
10:82-82.5	rs6586030	9.24 x 10^{-16}	0.847	1.115†	*TSPAN14* , *10orf58*, (4)
10:94.18-94.68	rs7911264	2.98 x 10^{-8}	0.519	1.066	(4)
10:101.03-101.53	rs4409764	1.03 x 10^{-54}	0.491	1.182	*NKX2-3*, (6)
11:1.62-2.12	rs907611	2.7 x 10^{-10}	0.315	1.068	***TNNI2***, *LSP1*, (17)
11:58.08-58.58	rs10896794	6.8 x 10^{-10}	0.762	1.08	*CNTF*, *LPXN*, (8)
11:60.52-61.02	rs11230563	9.03 x 10^{-13}	0.654	1.085	*CD6*, *CD5*, *PTGDR2*, (12)
11:61.31-61.81	rs4246215	1.93 x 10^{-15}	0.338	1.079†	*C11orf9*, *FADS1*, *FADS2*, (12)
11:63.85-64.39	rs559928	4.19 x 10^{-11}	0.821	1.101	***CCDC88B***, *RPS6KA4*,(20)
11:65.4-65.9	rs2231884	2.91 x 10^{-10}	0.157	1.083†	*RELA*, *FOSL1*, *CTSW*, *SNX32*, (22)
11:76.04-76.54	rs2155219	4.24 x 10^{-36}	0.509	1.151†	(5)
11:86.87-87.37	rs6592362	2.32 x 10^{-8}	0.248	1.083	(1)
11:118.49-118.99	rs630923	7.07 x 10^{-9}	0.846	1.074†	*CXCR5*, (17)
12:12.4-12.9	rs11612508	1.06 x 10^{-8}	0.267	1.058†	*LOH12CR1*, (8)
12:40.5-41.03	rs11564258*	6.38 x 10^{-29}	0.025	1.334†	*MUC19*, (1)
12:47.95-48.45	rs11168249	7.78 x 10^{-9}	0.467	1.054†	*VDR*, (8)
12:68.24-68.74	rs7134599	8.51 x 10^{-32}	0.378	1.096†	***IFNG***, *IL26*, *IL22*, (1)
13:27.27-27.77	rs17085007	2.79 x 10^{-19}	0.183	1.106†	(2)
13:40.45-41.26	rs941823**	2.07 x 10^{-14}	0.758	1.071†	(3)
13:99.7-100.2	rs9557195	2.37 x 10^{-14}	0.772	1.112	*GPR183*, *GPR18*,(6)

14:69.02-69.52	rs194749	2.7 x 10^{-10}	0.226	1.075†	*ZFP36L1*, (4)
14:75.45-75.95	rs4899554	2.71 x 10^{-8}	0.819	1.083†	***FOS***, *MLH3*, (6)
14:88.22-88.72	rs8005161	2.35 x 10^{-14}	0.089	1.153	***GPR65***, *GALC*, (1)
15:67.18-67.68	rs17293632	5.97 x 10^{-16}	0.235	1.067†	*SMAD3*, (2)
15:90.92-91.42	rs7495132	9.48 x 10^{-11}	0.891	1.134	*CRTC3*, (3)
16:11.12-11.95	rs529866*	1.73 x 10^{-16}	0.803	1.124†	***SOCS1***, ***LITAF***, *RMI2*, (10)
16:23.61-24.11	rs7404095	9.68 x 10^{-10}	0.572	1.06	***PRKCB***, (5)
16:28.26-28.93	rs26528	9.65 x 10^{-22}	0.451	1.099†	*RABEP2*, *IL27*, *EIF3C*, *SULT1A1*, (11)
16:85.75-86.25	rs10521318	1.41 x 10^{-9}	0.915	1.155†	*IRF8*, (4)
17:32.34-32.84	rs3091316	1.22 x 10^{-26}	0.722	1.122†	***CCL13***, ***CCL2***, *CCL11*, (4)
17:37.66-38.16	rs12946510	4.1 x 10^{-38}	0.465	1.157	*IKZF3*, *ZPBP2*, *GSDMB*, *ORMDL3*, (13)
17:40.28-40.78	rs12942547	5.51 x 10^{-22}	0.580	1.103†	***STAT3***, *STAT5B*, *STAT5A*, (13)
17:57.71-58.21	rs1292053	8.85 x 10^{-13}	0.446	1.076†	*TUBD1*, *RPS6KB1*, (9)
18:12.55-13.05	rs1893217	3.05 x 10^{-26}	0.157	1.171†	(6)
18:46.14-46.64	rs7240004	1.31 x 10^{-9}	0.616	1.057†	*SMAD7*, (2)
18:67.28-67.78	rs727088	4.65 x 10^{-9}	0.484	1.077	***CD226***, (2)
19:10.22-10.76	rs11879191*	2.04 x 10^{-18}	0.797	1.136	***TYK2***, *PPAN-P2RY11*, *ICAM1*, (25)
19:33.48-33.98	rs17694108	5.85 x 10^{-15}	0.282	1.1	*CEBPG*, (8)

19:55.13-55.63	rs11672983	6.5 x 10^{-11}	0.392	1.087	*NLRP7*, *NLRP2*, *KIR2DL1*, *LILRB4*, (15)
20:30.47-31.03	rs6142618	6.05 x 10^{-10}	0.564	1.072†	*HCK*, (10)
20:31.12-31.62	rs4911259	1.2 x 10^{-9}	0.383	1.075	*DNMT3B*, (8)
20:44.49-44.99	rs1569723	9.95 x 10^{-14}	0.259	1.091†	***CD40***, *MMP9*, *PLTP*, (11)
20:48.7-49.2	rs913678	4.59 x 10^{-8}	0.662	1.056	*CEBPB*, (5)
20:57.57-58.07	rs259964	1.01 x 10^{-12}	0.464	1.085	*ZNF831*, *CTSZ*, (5)
20:62.09-62.59	rs6062504	1.09 x 10^{-23}	0.684	1.104	*TNFRSF6B*, *LIME1*, *SLC2A4RG*, (24)
21:16.56-17.06	rs2823286	9.28 x 10^{-30}	0.708	1.157†	(0)
21:40.21-40.71	rs2836878	4.62 x 10^{-48}	0.733	1.18†	(3)
21:45.37-45.87	rs7282490	2.35 x 10^{-26}	0.391	1.105	*ICOSLG*, (9)
22:21.67-22.17	rs2266959	1.39 x 10^{-16}	0.186	1.105	*MAPK1*, *YDJC*, *UBE2L3*, *RIMBP3*, (9)
22:30.12-30.73	rs2412970	2.7 x 10^{-14}	0.457	1.08	***LIF***, ***OSM***, *MTMR3*, (8)
22:39.4-39.97	rs2413583*	4.4 x 10^{-33}	0.833	1.209†	*ATF4*, TAB1, APOBEC3G, (16)

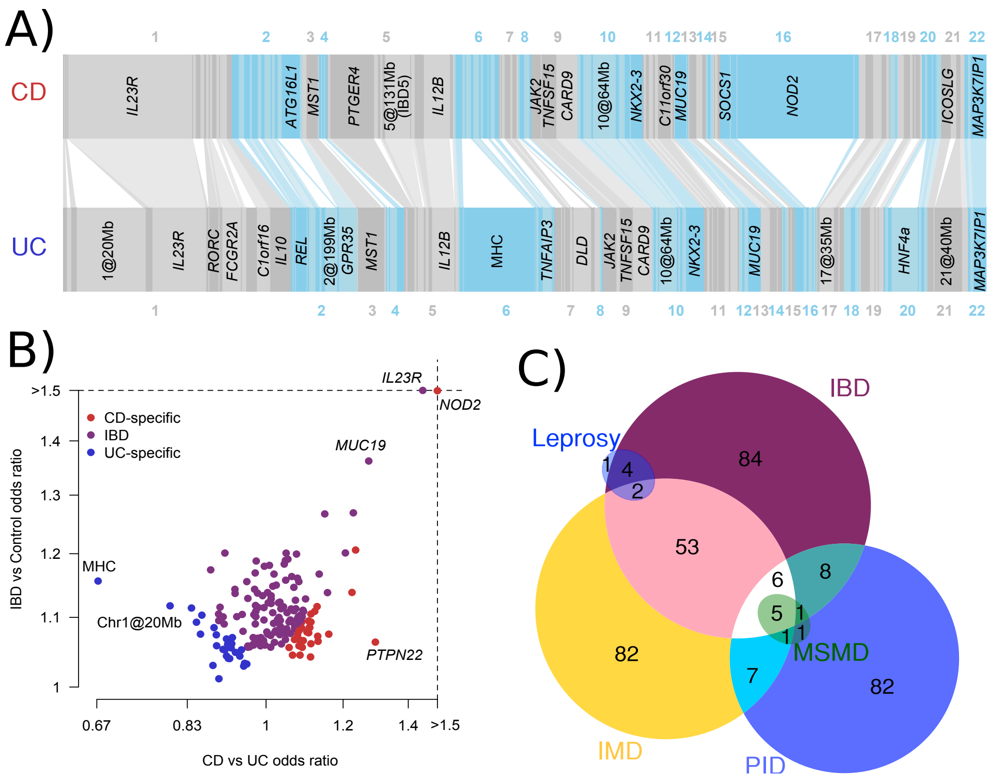


Figure 4.9: The IBD genome. A) The 163 IBD loci identified in this study. Each bar, ordered by genomic position, represents an independent locus, and the width of the bar is proportional to the variance explained by that locus in CD and UC. Bars are connected together if they are identified as being associated with both phenotypes. Loci are labelled if they explain more than 1% of the total variance explained by all loci for that phenotype. B) The 193 independent signals, plotted by total IBD odds ratio and phenotype specificity (measured by the odds ratio of CD relative to UC), and coloured by their IBD phenotype classification from Table 1. C) Number of overlapping IBD loci with other immune-mediated diseases (IMD), leprosy, and Mendelian primary immunodeficiencies (PID). Within PID, we highlight Mendelian susceptibility to mycobacterial disease (MSMD).

4.4 Biological and bioinformatic interpretation of 163 IBD loci

Our meta-analysis of the GWAS and Immunochip data identified 193 statistically independent signals of association at genome-wide significance ($P < 5 \times 10^{-8}$) in at least one of the three phenotypes (CD, UC, IBD). These signals were merged into 163 regions, of which 71 have never been reported before (Table 4.9). This is more loci than has ever been recorded for a complex disease, and the number of loci, and the large number of genes they contain, make a locus-by-locus interpretation difficult. To go from a list of regions to a set of specific biological hypotheses we have to use computational techniques, and make use of external datasets.

In this section I will discuss a number of ways in which this can be achieved, starting with a brief overview of the IBD loci. I will go on to use genetic data from other disease loci (both complex and Mendelian) to place IBD genetics in the context of other immune diseases. Next I will describe a range of methods for identifying candidate causal genes within the identified loci using gene networks and functional information. I will then describe a detailed analysis of the identified candidate genes in terms of Gene Ontology (GO) terms and canonical pathways, followed by an analysis of the IBD loci in the context of natural selection. Finally, I will describe a number of functional analyses based on gene expression data carried out by other members of the IIBDGC Immunochip analysis group.

4.4.1 Global patterns in the "IBD genome"

A traditional Manhattan plot of this study does not provide much information, due to the large number of peaks and high variation in significance between them. Instead, I have developed an alternative visualisation, which I call the "Belgravia plot" (by analogy with the flat, regular Regency terraces in Belgravia, London). This plot (Figure 4.9A) shows the relative contribution of each locus to the total variance explained in UC and CD using width, rather than height. This gives us an intuitive overview of the importance of the global structure of IBD. For instance, CD is more dominated by the two major loci (*NOD2* and *IL23R*), with UC having a more even distribution.

The likelihood-based model selection analysis described in Section 4.3.7 gives us information on the global level of genetic sharing between the two IBD phenotypes. The vast majority of loci (110) are associated with both disease phenotypes, of which 62 have an indistinguishable effect size in UC and CD, while 48 show evidence of heterogeneous effects (highlighted in Table 4.9). Of the remaining loci, 30 are classified as specific for CD and 23 for UC, but notably, 43 of these 53 show the same direction of effect in the non-associated disease (overall P = 2.8 x 10^{-6}), suggesting that only a few of the loci may be truly disease specific.

While likelihood-based approaches for the classification of IBD loci are instructive, it should be noted that there is continuous variability in the CD-UC balance of effect sizes among loci (Figure 4.9B). While locus sharing is almost exclusively in the same direction, risk alleles at two CD loci, *PTPN22* and *NOD2*, show significant ($P < 0.005$) protective effects in UC, highlighting them as particularly informative about biological differences between these related diseases.

4.4.2 IBD genetics in the context of autoimmunity and infection

To place the IBD loci in the context of other immune-related diseases, I generated lists of associations with other immune-related disease. I included complex autoimmune and immune-mediated and diseases (IMD), and autosomal dominant or recessive primary immunodeficiencies (PID).

I took autosomal dominant and recessive genes identified as causing PID from Notarangelo et al. (2009). Genes that lie within 250kb of each other were merged together into regions, giving 135 genes across 121 independent

Disease	Locus overlap	Fold-enrichment	Enrichment OR	95% CI	P-value
PS	14	13.5	14.71	8.5-25.5	4.15×10^{-12}
AS	8	12.56	13.18	6.5-26.8	3.22×10^{-7}
AD[a]	3	12.1	12.32	3.9-38.6	0.002
PBC	13	10.99	11.88	6.7-21.0	3.12×10^{-10}
PSC[b]	1	10.93	11.00	1.5-78.6	0.085
RA	12	10.92	11.74	6.5-21.1	1.64×10^{-9}
Celiac	16	10.57	11.64	7.0-19.5	4.56×10^{-12}
T1D	20	9.99	11.28	7.1-19.0	2.35×10^{-14}
SLE	12	9.75	10.47	5.8-18.9	5.87×10^{-9}
All AI	66	8.62	13.94	10.2-19.1	5.15×10^{-44}
MS	17	8.19	9.06	5.5-15.0	5.11×10^{-11}
Asthma	7	7.61	7.91	3.7-16.9	4.90×10^{-5}
All PID	20	4.85	5.42	3.4-8.7	8.52×10^{-9}

Table 4.10: Enrichment in overlap between IBD loci and loci for other immune-mediated diseases. The enrichment OR is measured on the logistic scale (as described in section 4.4.4). [a]Atopic dermititus. [b]Primary sclerosing cholangitis

regions. I took associated regions for the IMD list from the NHGRI GWAS catalogue, and included the following diseases: Primary sclerosing cholangitis, primary biliary cirrhosis, rheumatoid arthritis, type 1 diabetes, multiple sclerosis, celiac disease, atopic dermatitis, psoriasis, ankylosing spondylitis, asthma and systemic lupus erythematosus. All SNPs in the catalogue with p $< 5 \times 10^{-8}$ were included. As with the IBD loci, I defined a region as 250kb on either side of the hit SNP, and overlapping regions were merged together into loci. This generated a total of 156 independent IMD loci. I assessed overlap between lists (IBD, PID and IMD) using the method described in Section 4.4.4.

A large proportion (113 of 163) of the IBD loci are shared with other complex diseases or traits. Sixty-six of these 113 are among the 154 loci previously associated with other immune-mediated diseases (Hindorff et al., 2009), which is 8.6 times the number that would be expected by chance (Figure 4.9C, $P < 10^{-16}$). Comparing overlaps with specific diseases (Table 4.10) is confounded by the differential power in studies of different diseases. For instance, while type 1 diabetes (T1D) shares the largest number of loci (20/39, 10-fold enrichment), this is partially driven by the large number of known T1D associations. Indeed, seven other immune-mediated diseases show stronger enrichment of overlap, with the largest being ankylosing spondylitis (8/11, 14-fold) and psoriasis (14/17, 13-fold).

In addition to this well-established genetic overlap between IBD and other complex immune mediated diseases, we can now show that IBD loci are also markedly enriched (4.9-fold, $P < 10^{-4}$) in genes involved in primary immunodeficiencies (PIDs, Figure 2C). Consistent with an important role for T-cells in IBD, most of the PIDs overlapping with IBD loci are characterised by reductions in levels of circulating T-cells (*ADA*, *CD40*, *TAP1/2*, *NBS1*, *BLM*, *DNMT3B*), levels of Th17 (*STAT3*), memory T-cells (*SP110*) or regulatory T-cells (*STAT5B*), rather than reduced levels of circulating B-cells (cell count characteristics taken from Notarangelo et al. (2009)).

Compared to the overlap between PID genes and IBD loci, the subset of PIDs leading to Mendelian susceptibility to mycobacterial infection (MSMD) (Notarangelo et al., 2009; Bustamante et al., 2011; Patel et al., 2008) are enriched still further. Of the eight known autosomal genes that increase susceptibility to MSMD, six are located within IBD loci (46-fold enrichment, $P = 1.3 \text{ x } 10^{-6}$), and a seventh, *IFNGR1*, narrowly missed genome-wide significance ($P = 6 \text{ x } 10^{-8}$). A further relationship to MSMD is seen in the new association near the gene *CD40*, which is involved in MSMD induced by mutations in the X chromosome gene *NEMO* (Filipe-Santos et al., 2006). Furthermore, genetic defects in *STAT3* (Holland et al., 2007; Minegishi et al.,

2007) and *CARD9* (Glocker et al., 2009b), also within IBD loci, lead to PIDs involving skin infections with staphylococcus and candidiasis, respectively, further underscoring the importance of host-microbe interactions in IBD.

This mycobacterial disease overlap is not limited to Mendelian susceptibility. We also find IBD associations in 7/8 loci known to be associated with complex susceptibility to leprosy GWAS (Zhang et al., 2011), including 6 cases where the same SNP is implicated (Figure 4.9C).

There appears to be a shared biology underlying these all these overlapping mycobacterium susceptibility loci. All of the MSMD mutations that overlap with IBD cause defects in interferon signalling pathways, which are known to be important in mycobacterium infection (Flynn et al., 1993). Additionally, the six MSMD genes, four of the leprosy genes and *CD40* fit together into a single well-connected subnetwork within the GRAIL and DAPPLE networks described below (Figure 4.10A). This subnetwork also includes many important signalling proteins involved in IBD and bacterial defence, including *IFNG*, *IL10* and *NFKB1*.

4.4.3 Prioritising candidate genes in IBD loci

We used various methods to reduce the 1438 genes in our locus list to a more limited list of candidate variants. We used both gene network analyses, and analyses of SNP function, to implicate candidate genes.

We used GRAIL and DAPPLE (discussed in Section 4.2.3) to prioritise genes based on network connectivity. In both cases, we removed the HLA region (due to its large size), and fixed four well-established IBD genes as causal (*NOD2*, *IL23R*, *ATG16L1* and *PTPN22*). We took any gene with $p < 0.05$ as implicated. We also included genes from the gene expression network discussed in section 4.4.6.

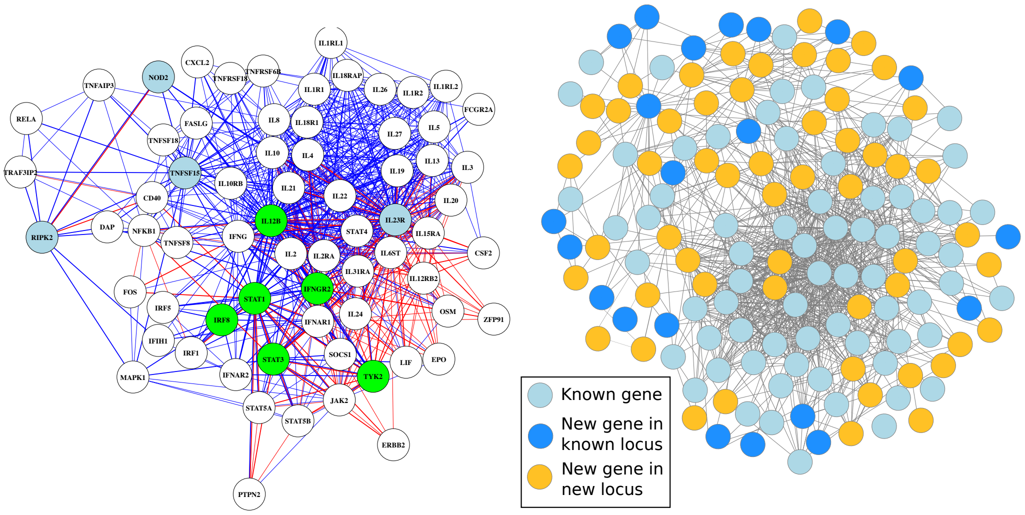


Figure 4.10: a) A combined network graph including GRAIL (blue lines) and DAPPLE (red lines) connections, consisting of all genes connected to MSMD or leprosy genes (highlighted in green and teal respectively) b) The GRAIL network for all genes with GRAIL P < 0.05. Genes included in our previous GRAIL networks in CD and UC are shown in light blue, newly connected genes in previously identified loci in dark blue, and genes from newly associated loci in gold.

Compared to previous analyses that identified candidate genes in 35% of loci (Anderson et al., 2011; Franke et al., 2010) our updated GRAIL-connectivity network identifies candidates in 53% of loci, including increased statistical significance for 58 of the 73 candidates from previous analyses. The new candidates come not only from genes within newly identified loci, but also integrate additional genes from previously established loci (Figure 4.10B). The joint-IBD loci are more likely to contain GRAIL connected genes than CD- or UC-specific loci (P = 0.005), pointing to the shared core of genetic risk between the two diseases.

We also used existing annotations of variant function to search for likely causal mechanisms. We used SeattleSeq to annotate all variants in high LD ($r^2 > 0.8$) with missense or nonsense SNPs, producing 29 IBD associations that caused a protein code change. We also collected eQTL data from a range of studies, including lymphoblastoid cell lines of asthmatic children (Dixon et al., 2007), various tissues from obese patients (Greenawalt et al., 2011), and a collection of eQTL studies from the Chicago eQTL browser. We found evidence that 64 IBD associations altered the expression of at least one gene.

Overall, our network analyses and functional annotations highlighted a total of 300 candidate genes in 125 loci, of which 37 contained a single gene supported by two or more methods.

4.4.4 Testing for enrichment of functional terms within IBD loci

Gene Ontology (GO) terms and canonical pathways are a natural way to ask questions about the function of the genes in the identified IBD loci. We can ask whether there is an enrichment of certain functional terms in IBD loci, as well as whether these functional loci are associated with a particular type of locus (e.g. CD loci). Below I outline a method for performing tests for enrichment and association of functional terms. I then go on to apply this to the IBD data, to find functional terms associated with IBD, as well as identifying terms associated with CD-UC balance, and are more strongly enriched in IBD relative to other immune-mediated diseases. Finally, I use this methodology to investigate potential functional biases introduced by the structure of Immunochip, and by using genes identified by the prioritisation approaches described above.

A methodology for testing functional enrichment in IBD loci

Basic framework

We wish to assess the enrichment of a particular functional term (e.g. a GO term) in causal IBD genes. Given a list of causal genes, we could easily calculate an enrichment odds ratio λ_i of a functional term i in IBD genes relative to the genome as a whole, and perform a statistical test of $\lambda_i = 1$ vs $\lambda_i > 1$. However, we do not know the causal variant for most IBD regions, and most IBD regions contain multiple genes. To compensate for this, we use an extension of the standard odds ratio method that takes into account the presence of non-causal genes.

Assume that we have M loci, designated by $j = (1, ..., M)$ each of which contains N_j genes. For each associated locus j we set an indicator variable δ_{ij} to 1 if the functional term i is present in locus j, and 0 otherwise. We also calculate a genome-wide frequency for term f_i that is equal to the proportion of all genes that contain the term i.

We calculate g_i, the frequency of term i in causal genes, given an enrichment odds ratio λ_i as

$$g_i = \left(1 + \frac{1 - f_i}{\lambda_i f_i}\right)^{-1}. \tag{4.3}$$

We then assume that all other genes have a frequency of the term f_i. Assuming that there is exactly one causal gene in the region, the log likelihood L_i is given by

$$L_i = \sum_j \delta_{ij} log\left(1 - (1 - f_i)^{N_j}(1 - g_i)\right) + \sum_j (1 - \delta_{ij}) log\left((1 - f_i)^{N_j}(1 - g_i)\right). \tag{4.4}$$

We fit the parameter λ_i by maximum likelihood using the Nelder-Mead optimisation method, implemented in the statistical package R. We assess the significance of the parameter λ_i by performing a likelihood ratio test of $\lambda_i = 1$ vs $\lambda_i \neq 1$.

Extension to arbitrary predictors

We can extend the method above to include arbitrary per-locus predictors $X = \{x_{jk}\}$ that correlate with level of enrichment of a function term. We can extend the definition of g_i to take the form of a generalised logistic model

$$g_i = \left(1 + \frac{1 - f_i}{f_i} exp(-\beta_0 - \vec{\beta}X)\right)^{-1} . \tag{4.5}$$

We keep the enrichment odds ratio (in this case as $\lambda_i = exp(\beta_0)$), but also include regression coefficients for the other predictors $\vec{\beta}$. The predictors X can be discrete (e.g. $x_{jk} = 0$ if locus j is a UC locus, and $x_{jk} = 1$ if it is a CD locus), or continuous (e.g. $x_{jk} = \theta_j$, where θ_j is the polar-transformed odds ratio described in section 4.3.7). The model is fitted by maximum likelihood in the same way as the simple enrichment model, and likelihood ratio tests for $\beta_k = 0$ can be used to assess the significances of the parameters.

Extension to interval overlap

We can extend the above methodology to look for an enrichment in overlap between a set of genomic intervals (e.g. a set of wide linkage peaks) and our IBD loci. We assume that we have a set of genomic intervals $k = 1.,,.R$, each of length l_k. We will also assume that the length of each locus is l_j and the length of the whole genome is l_g. We can thus modify equations 4.3, 4.4 and 4.5 above by setting

$$f_i = \frac{1}{l_g} \sum_k (l_k + l_i). \tag{4.6}$$

It was this extension that enabled me to evaluate the significance of overlap between our IBD loci and GWAS associations discussed in section 4.4.2.

Functional term associations in the IBD loci

I tested the 300 genes prioritised in section 4.4.3 for enrichment in 15,526 human GO terms (27/02/2012 release) and 833 canonical pathways (taken from KEGG, Reactome and Biocarta). I identified 286 GO terms and 48 pathways demonstrating significant enrichment in genes contained within IBD loci. The top associations are shown in Table 4.11, though the large number makes interpreting the entire list difficult.

We can use the hierarchical nature of the GO terms to bring some order to these terms. For instance, cytokine production is the most significantly enriched term, but within that four child terms drive this: IFNγ, IL12, TNFα and IL10. These cytokines are all produced by the cells of the innate immune system (including macrophages, dendritic cells and natural killer cells) in response to bacterial stimulation. This immediately suggests that the IBD risk alleles are, as a whole, interfering with the correct response to bacteria by altering the resulting rates of cytokine production.

The second strongest enrichment was in immune system processes (P = 2.6 x 10^{-23}), with activation of T-, B-, and NK-cells being the strongest contributors to this signal (P = 1.6 x 10^{-22}). Strong enrichment was also seen for response to molecules of bacterial origin (P = 9.6 x 10^{-20}), further evidence for a close relationship between IBD risk and bacterial exposure.

We can test whether any of these enriched functional terms show evidence of differential enrichment between CD and UC phenotypes, both by using the

Term	Description	Loci	P-value
GO:0002376	immune system process	69	$3.45 \text{ x } 10^{-26}$
GO:0002682	regulation of immune system process	60	$2.61 \text{ x } 10^{-25}$
GO:0001817	regulation of cytokine production	39	$2.65 \text{ x } 10^{-24}$
GO:0046649	lymphocyte activation	32	$1.77 \text{ x } 10^{-23}$
GO:0031347	regulation of defence response	39	$4.78 \text{ x } 10^{-23}$
GO:0048518	positive regulation of biological process	90	$3.23 \text{ x } 10^{-22}$
GO:0050865	regulation of cell activation	36	$1.63 \text{ x } 10^{-21}$
GO:0045321	leukocyte activation	33	$1.84 \text{ x } 10^{-21}$
GO:0048522	positive regulation of cellular process	83	$9.27 \text{ x } 10^{-21}$
GO:0002237	response to molecule of bacterial origin	28	$2.41 \text{ x } 10^{-20}$
GO:0050776	regulation of immune response	46	$2.90 \text{ x } 10^{-20}$
GO:0002684	positive regulation of immune system process	45	$3.05 \text{ x } 10^{-20}$
GO:0042110	T cell activation	24	$1.56 \text{ x } 10^{-19}$
GO:0006955	immune response	51	$1.76 \text{ x } 10^{-19}$
GO:0002694	regulation of leukocyte activation	33	$3.09 \text{ x } 10^{-19}$
GO:0001775	cell activation	38	$3.40 \text{ x } 10^{-19}$
GO:0032496	response to lipopolysaccharide	26	$5.36 \text{ x } 10^{-19}$
GO:0051249	regulation of lymphocyte activation	31	$8.13 \text{ x } 10^{-19}$
GO:0070663	regulation of leukocyte proliferation	24	$8.67 \text{ x } 10^{-19}$
GO:0080134	regulation of response to stress	43	$1.55 \text{ x } 10^{-18}$
KO:04630	Jak-STAT signalling pathway	20	$4.80 \text{ x } 10^{-15}$
KO:05140	Leishmania infection	16	$3.89 \text{ x } 10^{-14}$
KO:04060	Cytokine-cytokine receptor interaction	25	$1.66 \text{ x } 10^{-13}$
BI	Th1/Th2 differentiation	10	$1.64 \text{ x } 10^{-12}$
BI	NO2-dependent IL12 pathway	7	$3.25 \text{ x } 10^{-10}$
RE:690 0	Signalling in immune system	24	$3.35 \text{ x } 10^{-10}$
KO:04062	Chemokine signalling pathway	16	$1.10 \text{ x } 10^{-9}$
BI	IL12-dependent signalling pathway	7	$7.73 \text{ x } 10^{-9}$
KO:05330	Allograft rejection	9	$2.34 \text{ x } 10^{-8}$
KO:04660	T-cell receptor signalling pathway	13	$2.49 \text{ x } 10^{-8}$

Table 4.11: The top 20 most enriched GO terms, and top 10 canonical pathways, in IBD loci. Terms starting "GO" are Gene Ontology terms, those starting "KO" are KEGG pathways, "RE" are Reactome pathways and "BC" are Biocarta pathways

Term	Description	Direction	p_θ	$p_{CD/UC}$
GO:0007243	intracellular protein kinase cascade	CD	0.0046	0.0005
GO:0051241	negative regulation of multicellular organismal process	UC	0.0796	0.0039
GO:0000165	MAPK cascade	CD	0.0058	0.0086
GO:0002237	response to molecule of bacterial origin	CD	0.0099	0.0140

Table 4.12: Pathways that show evidence of differential enrichment ($p < 0.01$) in CD and UC. The "direction" shows which phenotype has the higher enrichment of this term. p_θ is the evidence of association between functional term and CD-UC balance parameter θ. $p_{CD/UC}$ is evidence of differential enrichment in CD and UC loci (as defined in Table 4.9)

Term	Description	p_{IMD}	p_{PID}	p_{axis}
KO:04350	TGFβ signalling pathway	0.015	0.004	0.001
BI	Erythropoietin signalling pathway	0.03	0.04	0.004

Table 4.13: Pathways that show evidence of enrichment ($p_{axis} < 0.01$) in IBD loci relative to other immune-mediated disease loci. p_{IMD} and p_{PID} is the enrichment p-value relative to complex immune-mediated diseases and Mendelian primary immunodeficiencies respectively, and p_{axis} is the enrichment p-value of IBD relative to both IMD and PID.

phenotype-specific loci defined in Table 4.9, and using the continuous CD-UC balance parameter θ defined in section 4.3.7. Neither analysis produced any results that met Bonferroni-corrected significance, but results that showed nominal ($p < 0.01$) significance are shown in Table 4.12. Perhaps the most interesting is the evidence that CD may have a larger enrichment of terms involved in response to bacterial products, as this reinforces the opposite direction of effect we see at the *NOD2* locus (itself responsible for responding to the bacterial product MDP).

We can perform a similar analysis comparing IBD to the set of immune-mediated complex diseases and primary immunodeficiencies described in sec-

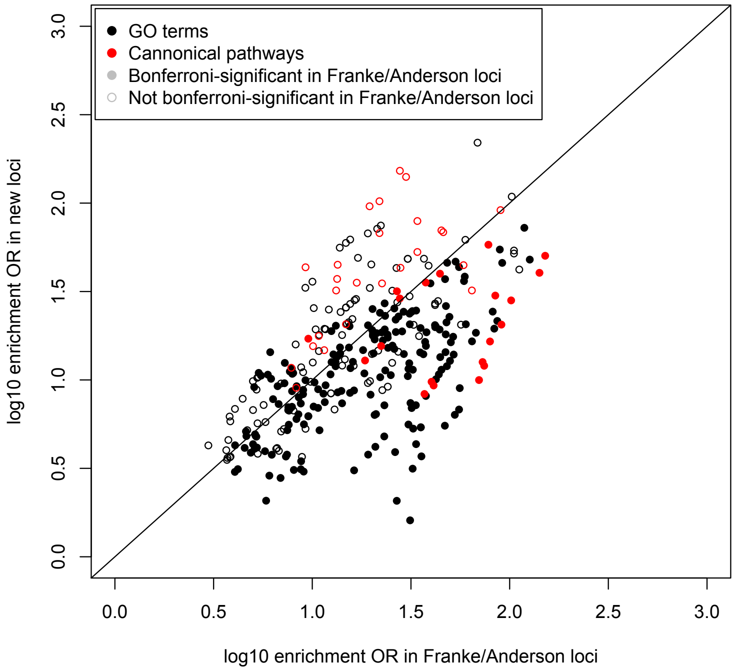


Figure 4.11: GO enrichment in known vs. new loci. The enrichment odds ratios for enriched GO terms are plotted for loci discovered via GWAS and for new loci identified in the current Immunochip analysis. The circled are filled if they were significant in the GWAS loci, and empty if they are only significant when all loci are combined.

tion 4.4.2. Again, no functional term produced a Bonferroni-significant result, but the strongest enrichment was in the TGFβ signalling pathway (Table 4.13). TGFβ is a widely expressed protein that has been implicated in many diseases. However, knock-out mice develop colorectal cancer, potentially suggesting a particular role for TGFβ in the intestinal immune system (Sterner-Kock et al., 2002).

Testing for functional biases in Immunochip genes

The Immunochip was constructed using variant lists submitted by immune-related disease association consortia. We may therefore expect there to be a bias towards discovering loci that are associated to both IBD and other immune-related diseases. This would, in turn, cause an artificial inflation in enrichment of immune-related GO terms. To test this hypothesis, I re-calculated enrichment odds ratios for the 286 enriched GO terms and 48 canonical pathways in two non-overlapping subsets of the 163 loci: (i) the 92 loci described in our previous meta-analyses, and (ii) the 71 newly discovered loci. If our analysis for identifying new IBD loci were biased (via the Immunochip design) toward loci shared across autoimmune diseases we would expect larger enrichment odds ratios in set (ii) compared to (i). Figure 4.11 shows that in fact, the opposite is true: the previous loci are, on average, slightly more strongly enriched than our new loci ($p = 2.2\text{x}10^{-9}$). This difference might suggest that the strongest IBD loci (i.e. those already known) play a more central role in key immune functions than our new discoveries.

This lack of observable bias, while initially surprising, can largely be explained by our experimental design, and the specifics of the SNP selection process for the Immunochip. As part of that design we included the top 2000 most associated SNPs each from the earlier CD and UC GWAS meta-analyses regardless of function or association with other phenotype (corresponding to $p < 0.0009$ for CD and $p < 0.0004$ for UC). This subset of SNPs therefore represents a functionally unbiased, genome-wide replication set that includes 147 (55 new, 92 known) of our 163 reported loci. Therefore the non-IBD part of the Immunochip contributed to only 16 of our loci, of which only 8 are known to be also associated with another immune-mediated disease. This number is too small to strongly bias enrichment analyses, as demonstrated

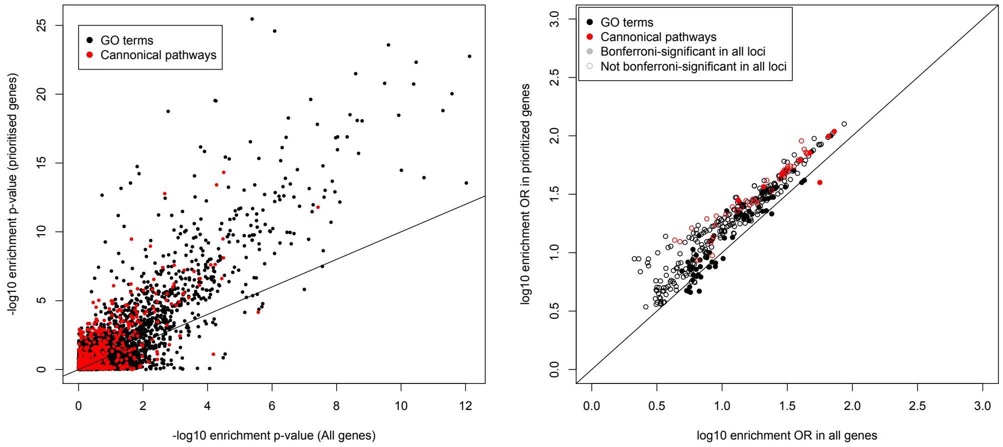


Figure 4.12: Enrichment p-values (a) and odds ratios (b) for GO terms (black dots) and canonical pathways (red dots) calculated on all 1438 genes in IBD loci (x-axis) and just the 300 prioritised genes (y-axis).

above.

Another potential source of bias is the use of the 300 genes selected by our gene prioritisation procedure. There is good reason to use these genes, as doing so grants a large increase in power to detect associations for both GO terms and canonical pathways (Figure 4.12a). However, this procedure is also likely to produce a bias towards the classes of genes and pathways that can be easily detected using gene prioritisation methods. To measure this effect, I calculated enrichment odds ratios for the selected GO term and canonical pathways using just the prioritised genes, and then using the entire set of genes inside the loci. Figure 4.12b shows that the estimated odds ratios are indeed higher when using the prioritised genes, suggesting that this introduces a detectable bias towards the detection of well-studied pathways. However, this bias is relatively small.

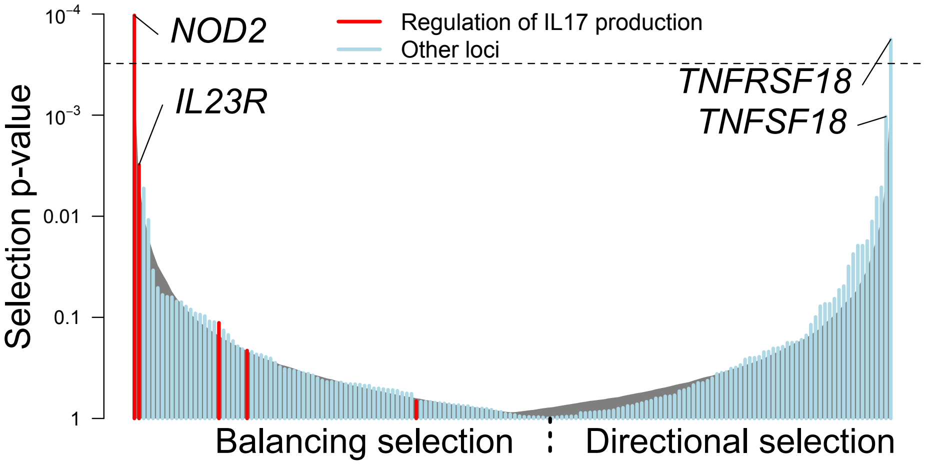


Figure 4.13: Signals of selection at IBD SNPs, from strongest balancing on the left to strongest directional on the right. The grey curve shows the 95% confidence interval for randomly chosen frequency-matched SNPs, illustrating our overall enrichment (p = 5.5 x 10^{-6}), while the dashed line represents the Bonferroni significance threshold. SNPs highlighted in red are annotated as involved in regulation of IL17 production, a key IBD functional term related to bacterial defence, and are enriched for balancing selection.

4.4.5 Natural selection in IBD loci

Infectious organisms are known to be among the strongest agents of natural selection (Lederberg, 1997). It seems logical to ask whether the strong genetic relationship between infection and IBD that emerges from the above analyses also suggests a role for natural selection in the evolutionary history of IBD susceptibility. There are many plausible types of selection that may be acting on IBD risk variants. The risk alleles may be under directional selection, either positive (if the decrease in fitness due to infection outweighs the increase in fitness due to inflammation), and negative (if vice versa). They may also be under balancing selection, indicative of an allele frequency dependent scenario typified by host-microbe co-evolution, as can be observed with parasites (Lederberg, 1997).

To test selection on IBD loci, I used data, provided by Joe Pickrell, generated using the TreeMix method developed by Pickrell and Pritchard (Pickrell and Pritchard, 2012) They constructed population trees from the Human Genetic Diversity Panel (HGDP) data (Li et al., 2008), and produced a per-variant score that measures the extent to which population allele frequencies at that site are over-dispersed relative to this tree. The most over-dispersed sites are likely to have been subjected to directional (positive or negative) selection, whereas those that match the tree most closely are likely to have been subjected to balancing selection.

I picked the best HDGP proxy SNP for each of our associated variants (picking only the UC associated variant from the HLA), and extracted the scores for these variants. Because the score is confounded with allele frequency, I calculated empirical p-values for each variant as follows: pick all variants with an allele frequency within 1 percentage point of the hit variant's allele frequency, and measure the proportion of variants with a score greater than the score of the hit variant. I calculated p-values for directional selection (the proportion of variants with a score higher than the hit variant), and p-values for balancing selection (the proportion with scores lower than the hit variant), as well as two-tailed p-values.

Two SNPs show Bonferroni-significant selection: the most significant signal, in *NOD2*, is under balancing selection (P = 5.2 x 10^{-5}), and the second most significant, in the receptor *TNFRSF18*, showed directional selection (P = 8.9 x 10^{-5}). The next most significant variants were in the ligand of that receptor, *TNFSF18* (directional, P = 5.2 x 10^{-4}), and *IL23R* (balancing, P = 1.5 x 10^{-3}). As a group, the IBD variants show significant enrichment in

Term	Description	Direction	P-value
GO:0032660	regulation of interleukin-17 production	Balancing	0.00014*
GO:00327	positive regulation of interleukin-17 production	Balancing	0.00020
GO:0009897	external side of plasma membrane	Directional	0.0018
GO:0008283	cell proliferation	Directional	0.0020
GO:0032653	regulation of interleukin-10 production	Balancing	0.0020

Table 4.14: Top 5 pathways that show evidence of natural selection in the IBD loci. *Significant after Bonferroni-correction for 334 enriched GO terms and pathways.

selection (Figure 4.13) of both types (P = 5.5 x 10^{-6}).

In order to assess whether extent or direction of selection was correlated with specific functions, I used the GO term enrichment method described above. I converted the selection p-values to Z scores using an inverse normal transformation, and tested for association between these scores and GO terms. The top five associations are shown in Table 4.14. The top result was the GO term “regulation of interleukin-17 production”, which met Bonferroni-corrected significance (Figure 4.13). The important role of IL17 in both bacterial defence and autoimmunity suggests a key role for balancing selection in maintaining the genetic relationship between inflammation and infection, and this is reinforced by a nominal enrichment of balancing selection in loci annotated with the broader GO term “defence response to bacterium” (p = 0.007).

4.4.6 Gene expression analyses of IBD loci

Gene expression datasets provide a powerful resource to interpret GWAS results. Two other groups within the IIBDGC Immunochip analysis group

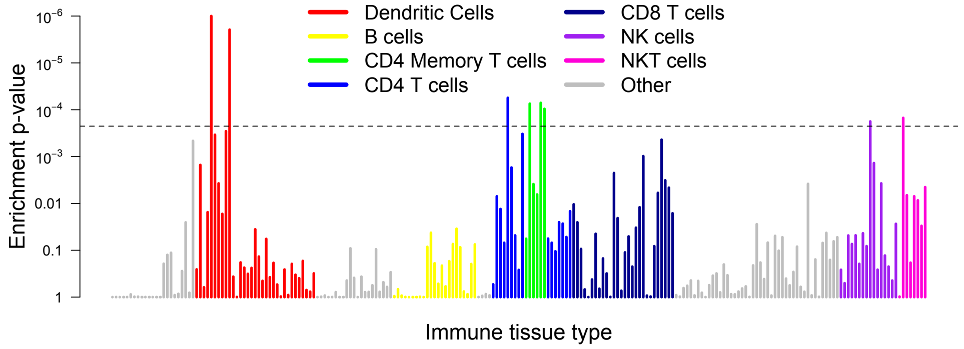


Figure 4.14: Evidence of enrichment in IBD loci of differentially expressed genes from various immune tissues. Each bar represents the empirical P-value in a single tissue, and the colours represent different cell type groupings. The dashed line is Bonferroni-corrected significance for the number of tissues tested.

used gene expression to investigate the new IBD locus list.

Xinli Hu and Soumya Raychaudhuri used enrichment of cell-type specific genes to study the cell types implicated by our locus list, using a previously described method (Hu et al., 2011). They tested for enrichment of cell-type expression specificity of genes in IBD loci in 223 distinct sets of sorted, mouse-derived immune cells from the Immunological Genome Consortium (Hyatt et al., 2006). Dendritic cells showed the strongest enrichment, followed by weaker signals that support the GO analysis, including CD4+ T, NK and NKT cells (Figure 4.14). Notably, several of these cell types express genes near our IBD associations much more specifically when stimulated; our strongest signal, a lung-derived dendritic cell, had $p_{stimulated} < 10^{-6}$ compared with $p_{unstimulated} = 0.0015$, consistent with an important role for cell activation.

Ken Hui and Eric Shadt used gene expression networks and eQTL data to infer causality in IBD associations. They screened genes in IBD loci against 211 co-expressed “modules” (sets of genes) previously identified by weighted

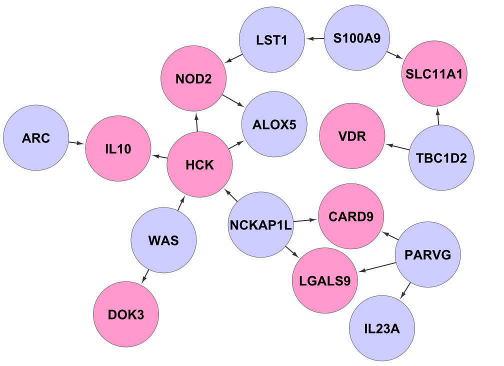


Figure 4.15: *NOD2*-focused cluster of the IBD causal subnetwork. Pink genes are in IBD associated loci, blue are not. Arrows indicate inferred causal direction of expression regulation.

gene co-expression network analyses (Zhang and Horvath, 2005) performed on multiple tissues (Greenawalt et al., 2011; Emilsson et al., 2008; Schadt et al., 2008), and identified a significantly enriched module in omental adipose tissue from obese patients ($p < 10^{-12}$). They then used gene expression and genotype data from these patients to construct a causal network using a Bayesian network reconstruction algorithm (Zhu et al., 2007). To illustrate the power of this approach, Figure 4.15 shows a small subset of this network around the gene *NOD2*, which also contains many important bacterial interaction genes including *IL10* and *CARD9*. This network implicates a number of new IBD genes as playing a part in response to bacteria, and in particular highlights the new IBD gene *HCK* as a potential regulator of the important IBD genes *NOD2* and *IL10*.

4.4.7 Take home messages about the biology of IBD

We have used a range of bioinformatic analyses to attempt to extract biological insight from the 163 loci and 1438 genes implicated by the Immunochip analysis. This has in turn produced a large amount of data, which itself needs to be interpreted. Below I will distil what I believe to be the major biological lessons that these analyses have taught us about the aetiology of IBD.

CD and UC show a very high degree of genetic overlap, with almost all of the 163 loci showing some degree of association to both. Likewise, there does not appear to be any strong differences in the function of CD and UC specific loci. However, many loci show a significant heterogeneity of odds ratio between the two phenotypes, with many having differing (or occasionally opposite) effects on CD and UC risk. Perhaps in the future we need to think about genetic differences between CD and UC not in terms of different loci, but as differently weighted combinations of the same loci. The same property may apply to subphenotypes of IBD (such as ileal verses colonic disease), and possibly even to the relationship between IBD and other immune-mediated diseases.

IBD shows genetic overlap with almost all diseases of immunity. However, **there is a startling overlap between IBD and susceptibility to both complex and Mendelian mycobacterial disease**. This is further highlighted by the large number of loci that contain genes involved in interferon gamma, including both the *IFNG* gene itself and its receptor *IFNGR2*, which is known to play a vital role in defence against *Mycobacterium tuberculosis* (Flynn et al., 1993). This relationship appears to have led to **significant natural selection on IBD alleles during human history**, and in particular balancing selection on the regulation of pro- and anti-inflammatory cytokines IL17 and IL10.

Cell types of both the innate and adaptive immune system play an important role in IBD. Gene expression data implicated dendritic and natural killer cells on the innate side, and CD4+ T-cells on the adaptive side. The

Gene Ontology analysis, however, implies a different mode of action of these two cell types. **IBD risk alleles seem to lead to defects of bacteria-induced cytokine production in innate immunity and defects of cell activation and signal transduction in the adaptive immune system**. This is not an exclusive relationship (one innate immune cell type has an activation-related GO term, "regulation of natural killer cell activation"), but it does seem to hold as a rule of thumb.

4.5 IBD and Y haplogroups

There is much suggestive evidence of a relationship between sex chromosomes and immunity. Most autoimmune diseases are more prevalent in females than in males (Whitacre, 2001), and individuals with Turner syndrome (a partial or total absence of one sex chromosome) are at higher risk of developing various immune-related diseases (Lleo et al., 2012). There is also evidence that the progression of HIV infection can vary between carriers of different Y haplogroups (Sezgin et al., 2009). However, large systematic studies of the effect of Y chromosome variation on human immune disease are relatively rare.

As mentioned in the introduction, 1735 Y chromosome variants were placed on the Immunochip for the purpose of assigning Y haplogroups. This gives us an opportunity to make a detailed and well powered study of the relationship between IBD risk and Y haplogroups. In this section I will describe the analysis of these variants in the IIBDGC Immunochip dataset.

4.5.1 Calling Y SNPs and assigning haplogroups

I selected males from the QC+ Immunochip sample set based on their mean normalised intensity at Y chromosome sites. There were a total of 22,129 males available, with 9,811 controls, 6,204 CD cases and 6,114 UC cases.

Because (at the time this study was carried out) the optiCall method used for genotype calling on the autosomes had not yet been adapted to run on sex chromosomes, I used the calling software Illuminus (Teo et al., 2007). The calls were generally of low quality, so I selected 150 haplogroup informative marker (Karafet et al., 2008) and manually inspected and fixed clusters using the program Evoker (Morris et al., 2010).

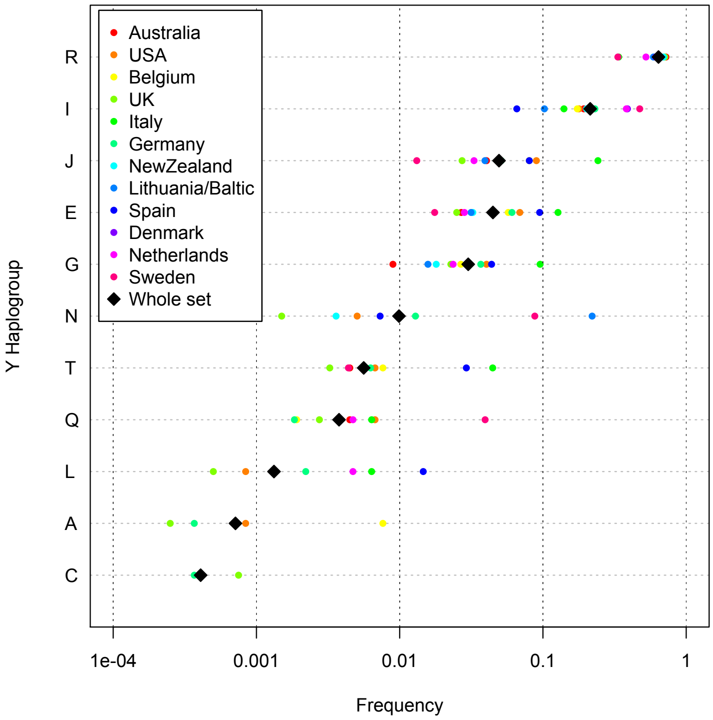


Figure 4.16: Y haplogroup frequencies in controls across the IIBDGC Immunochip dataset.

I developed a novel maximum likelihood method to assign haplogroups to these individuals (implemented in C++ as the program YFitter (Luke Jostins, 2011)). All but 9 males could be unambiguously assigned to a major haplogroup. The dataset contained samples from 10 major haplogroups, including 6 haplogroups with a frequency of greater than 1% (Figure 4.16).

4.5.2 Association analyses and controlling for stratification

I used logistic regression to assess association across these 6 common major haplogroups. The frequency spectrum differs between IBD cases and controls, even after including country-of-origin, sample collection and four autosomal principal components as covariates ($\chi^2 = 14.2$, df = 5, p = 0.014). The per-

Haplogroup	OR (95% CI)	P-value
E	1.07 (0.92 - 1.24)	0.393
G	1.20 (0.99 - 1.20)	0.059
I	1.00 (0.93 - 1.10)	0.837
J	0.85 (0.76 - 1.03)	0.112
N	1.53 (1.12 - 2.07)	0.006
R	0.96 (0.89 - 1.03)	0.229

Table 4.15: Association statistics for the Y chromosome haplogroups

haplogroup results show that this association is largely driven by a strong association between haplogroup N and IBD (Table 4.15).

Haplogroup N shows significant variation in frequency between European populations (Figure 4.16). This may lead us to suspect that the association results are due to population stratification. There are two major sources of stratification in IBD: a higher incidence in Ashkenazi Jewish, and an increasing incidence in Northern Europe compared to Southern Europe (Shivananda et al., 1996). We can rule out the former as haplogroup N has a lower frequency in Ashkenazim (Behar et al., 2004), which would produce the opposite direction of association to that observed. However, haplogroup N is at a significantly higher frequency in Northern Europe, so this is a plausible source of stratification. While I conditioned on country of origin and principal components, it is possible that additional stratification is driving the haplogroup N association.

To attempt to remove such stratification, I selected two homogeneous cohorts with over 10% frequency of haplogroup N (one Swedish and one Lithuanian). To ensure the population was homogeneous, I used principal component analysis to remove 136 individuals that lay more than two standard deviations from the mean on any of the first four PCs. Even within these two highly homogeneous populations, the results were very similar to

Collection	Cases/controls	OR (95% CI)
Sweden	165/193	1.19 (0.60 - 2.37)
Sweden (PC corrected)		1.27 (0.62 - 2.58)
Lithuania	192/109	1.94 (1.13 -3.33)
Lithuania (PC corrected)		1.81 (1.02 - 3.19)
Meta-analysis	228/228[a]	1.61 (1.05 - 2.47)
Meta-analysis (PC corrected)		1.57 (1.01 - 2.45)

Table 4.16: Association of haplogroup N with IBD in two homogenous populations. Studies were combined using variance-weighted fixed-effect meta-analysis. [a]Effective sample size

the across-Europe results (Table 4.16).

In these homogeneous groups, case-control status was correlated with principal components, weakly in Sweden (omnibus p = 0.050) and strongly in Lithuania (p = 3.6 x 10^{-4}). Equally, haplogroup N shows evidence of population stratification via a correlation between the haplogroup and principal components (p = 0.032 and p = 0.014). However, conditioning on the first four principal components within these countries does not significantly alter the results (Table 4.16), providing further evidence that this association is not driven by stratification.

4.5.3 Identifying candidate causal variants

Because the Y chromosome does not undergo recombination, the haplogroup association does not implicate a genomic region in the same way as an autosomal association does, and thus does not immediately suggest candidate genes or mutations.

To understand potential biological underpinnings of the haplogroup N enrichment in IBD, I used sequence data from the 1000 Genomes Project (specifically, from the Complete Genomics high-coverage sequencing) to iden-

Gene	Number	Location
AMELY	1	Intron
CD24	2	Upstream of TSS
KDM5D	3	CDS (missense)
NLGN4Y	8	Intron
PRKY	1	Intron
RPS4Y1	1	Intron
RPS4Y2	1	CDS (synonymous)
TBL1Y	5	Intron
TTTY10	2	Intron
TTTY14	2	Intron
TTTY15	3	Intron
USP9Y	4	Intron
UTY	13	Intron
ZFY	2	Intron

Table 4.17: Candidate genic mutations that may underlie the haplogroup N IBD association.

tify Y chromosome mutations specific to that haplogroup. A total of 379 mutations lie on or within the N haplogroup branch. 50 of these were present in or near genes, implicated 15 candidate genes (Table 4.17). These included a mutation 3kb upstream of *CD24*, a cell adhesion gene known to be up-regulated in inflammatory bowel disease, and a missense mutation in *KDM5D*, which encodes for a MHC antigen known to be involved in male-to-female tissue graft rejection.

4.6 Fine-mapping the *NOD2* locus

The IIBDGC has an ongoing project to fine-map IBD loci using the Immunochip. This project uses both the large European dataset discussed above, and an additional set of approximately 12,000 transethnic samples. It also aims to incorporate functional information from external datasets, such as gene expression and functional sequencing. It is aimed at both establishing the causal risk variants that underlie GWAS associations, and investigating the biological mechanisms through which these risk variants act. Calling and analysis of these datasets are currently ongoing.

In this section, I will describe the results of a pilot project carried out to investigate the methods and resources that could be used in such a fine-mapping project. This pilot project was focused on a single Crohn's disease fine-mapping region, the long-established *NOD2* locus. I will show how the Immunochip data can be used to infer new biological insights on both coding and non-coding associations at this locus.

4.6.1 Characterising coding mutations in *NOD2*

There are 24 polymorphic missense mutations in *NOD2* on the Immunochip. Six of these have been previously established as associated with IBD (Rivas et al., 2011). By performing stepwise logistic regression, I found that eight of these coding mutations show independent associations that are significant after correcting for the number of coding variants tested (i.e. $p < 0.002$), including the six known mutations and two that have not been reported before (Asn289Ser and Ala918Asp). With 8/24 mutations showing association, it is clear that a large proportion of the *NOD2* mutation space is associated with IBD. However, the Immunochip data can allow us to investigate in more

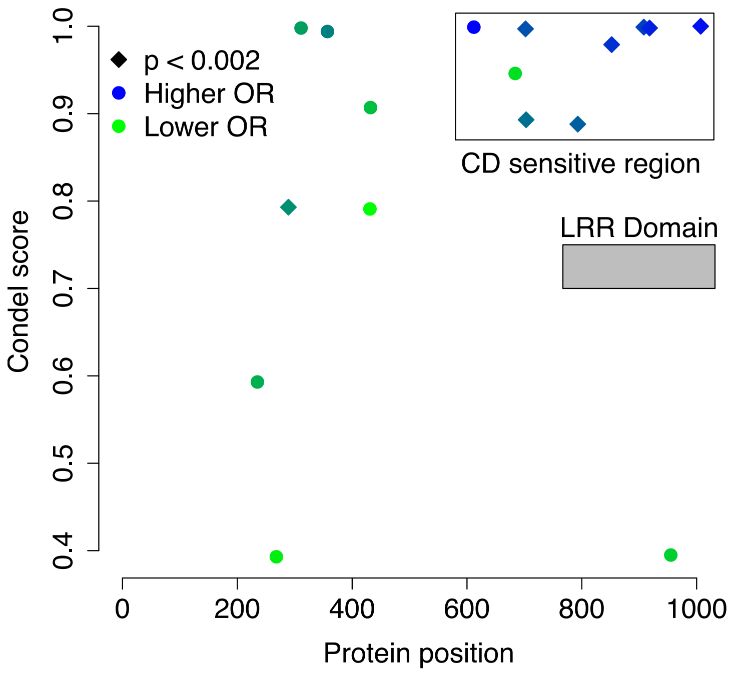


Figure 4.17: Functional characterisation of coding signals at NOD2. 17 coding variants are shown on a plot of their position in the protein and their Condel score, with colouring used to show their odds ratio. The LRR domain (responsible for bacterial sensing) is also shown.

detail what drives certain mutations to increase CD risk, while others appear to be benign from the point of view of IBD.

I took 17 of the highest frequency (MAF > 0.0005) non-synonymous variants and calculated independent odds ratios for each (conditioning on the six established *NOD2* coding mutations, plus the common regulatory signal discussed below). I also produced a Condel score (Gonzalez-Perez and Lopez-Bigas, 2011) for each mutation, which combines various measures of conservation and protein structure to estimate the probability that the mutation is pathogenic. Figure 4.17 shows the relationship between Condel score, position in the protein, and odds ratio. We can see a striking relationship:

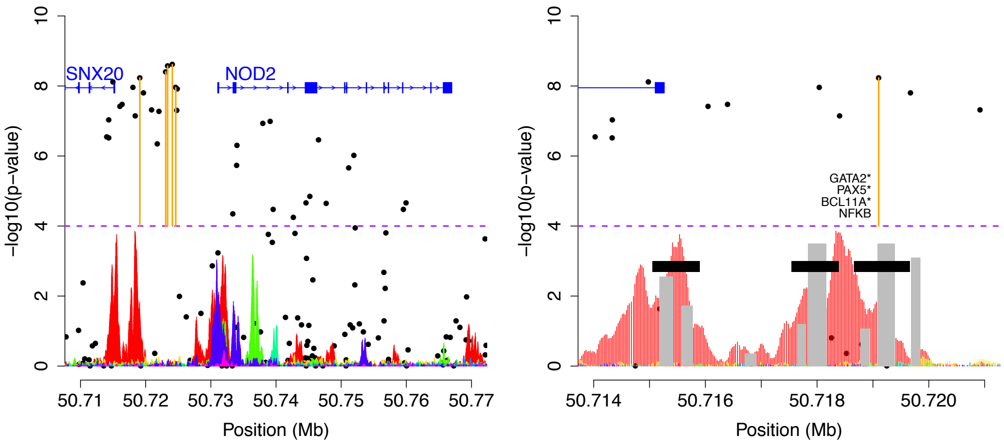


Figure 4.18: Fine-mapping and functional characterisation of a common regulatory signal at *NOD2*. Variants in orange are candidate causal variants. The coloured spikes under at the bottom of the plot show H3K27Ac histone modification levels in various tissues, with red being lymphoblastoid cell lines. Grey blocks are open chromatin and black blocks are transcription factor binding sites, with binding sites within 20bp of the candidate causal variant named in panel b).

mutations with a high Condel score, towards the end of the protein, almost invariably increase the risk of IBD. However, variants towards the start of the protein, or with a low Condel score, are rarely associated. It is likely that this "CD sensitive region" of *NOD2* represents mutations that disrupt the Leucine-Rich Repeat (LRR) domain. The LRR domain is responsible for detecting the bacterial product MDP, and is known to play a key role in Crohn's disease risk (Abraham and Cho, 2006).

4.6.2 Characterising a common regulatory signal at *NOD2*

Once we condition on the coding mutations mentioned above, a genome-wide significant signal remains around 50kb upstream of *NOD2* (Figure 4.18a). This signal is the same signal (but in the opposite direction) as the common *NOD2* association with leprosy susceptibility (Zhang et al., 2011), and is also associated with expression of both *NOD2* and *SNX20* in monocytes (Zeller et al., 2010). However, the association with IBD has not been reported before, as it can only be detected at genome-wide significance after conditioning on the coding variants.

Again, we are interested in the function of this association. The first step is to establish the set of SNPs that could plausibly be causal. To do this, we test the association statistics for the hit SNP conditional on each variant in LD with it, and rule out all SNPs that still show conditional association ($p < 0.01$). After performing this test, a total of 5 SNPs remain that could plausibly be the causal variant.

The next step is to establish what functional impact these candidate causal variants may have. Establishing the function of non-coding variants is difficult, but we can make some headway by using epigenetic sequencing data from the Encyclopaedia of DNA elements (ENCODE) (The ENCODE Project Consortium, 2012; Myers et al., 2011). In Figure 4.18b, I have overlaid H3K27Ac histone modification levels in various tissues: this is known to be an indicator of active enhancers (Creyghton et al., 2010). We can see that one of the candidate causal variants overlaps a peak that is specific to the lymphoblastoid cell line, suggesting an immune-tissue specific enhancer region. Looking closer at this region, we can see multiple sites of open chromatin and transcription factor binding (Figure 4.18b), with the candidate variant lying within one of these. The variant is nearby to binding sites for transcription factors involved in erythropoiesis (GATA2, PAX5 and BCL11A), as well as the protein NFκB, which regulates inflammation.

Taken together, this evidence points towards the existence of a common Crohn's disease risk variant in an upstream enhancer of *NOD2*. The upstream enhancer is active only in immune tissues, and appears to regulate expression of both *NOD2* and the neighbouring gene *SNX20*. This risk variant may act by interfering with a transcription factor binding, possibly a transcription factor involved in haemopoiesis.

4.7 Conclusions

The majority of this chapter has been focused on the use of the Immunochip to discovery new IBD loci in Europeans. The scale of the project has necessitated new approaches to both data handling and results interpretation, requiring a greater range of both techniques and expertise than previous IIBDGC analyses. Overall this has been a successful project, delivering both many new loci and biological information.

However, this project is only the first of many Immunochip analyses. At the time of writing we have just produced a new release of IBD Immunochip data, including over 86 thousand samples from both European and East Asian sample collections. This dataset will be used in a range of projects, including those that will fine-map existing loci, to study the contribution of IBD loci across different populations, and investigate the genetics of IBD subphenotypes. It will also be combined with Immunochip datasets from other diseases, in order to make a detailed investigation into the shared genetics of immune-mediated disease.

The results described in this chapter have taught us a number of lessons that will aid these future projects. Some of these are important, but perhaps uninteresting matters of quality control and data handling. For instance, the large manual inspection effort described in section 4.3.2 has given us many insights into the behaviour of Immunochip intensity readings, as well as setting up a framework for large, collaborative cluster plot inspection. Other lessons will have wider ranging consequences. For instance, the joint analysis of CD and UC demonstrated that two diseases can have an extremely high degree of genetic overlap (110 of 163 loci shared), but remain genetically distinct due to a high degree of effect size heterogeneity. We have learned that the relative balance of contribution of each locus can be just as important as the overall degree of locus sharing.

One of the strongest lessons to emerge from this analysis is the potential for integrating functional datasets into genetic studies. Gene expression, protein-protein interaction, canonical pathways and literature networks all added a great degree of value to the locus discovery effort. Most striking, the *NOD2* pilot fine-mapping project demonstrated the power of functional sequencing data in aiding the identification and understanding of non-coding causal variants. As a result of these successes, ongoing Immunochip projects are integrating, and in some cases specifically generating functional datasets as an integral part of their respective studies.