Chapter 5. Immune-mediated disease risk loci are enriched for differentially expressed genes from tissue-relevant functional genomic datasets

5.1 Introduction

Identifying the causal variants that are tagged by complex disease risk loci remains challenging. Blocks of linkage disequilibrium often contain multiple correlated association signals that are statistically indistinguishable from each other, and can span dozens of genes with multiple functional candidates. It is clear that the majority of common risk variants do not reside in protein coding regions (Hindorff et al.), suggesting that important aspects of disease etiology are driven by gene expression. While identifying specific causal variants is difficult, approaches that integrate GWAS association results with disease relevant functional genomic datasets may help in narrowing down potential candidate genes and the cell types in which they act.

Expression quantitative trait loci (eQTLs) provide a direct bridge between GWAS and gene expression. These studies measure gene expression across many individuals (typically in a genome-wide approach using microarrays or RNAseq), and then treat the expression level of each gene as a separate quantitative trait to test for association with SNPs – either at the same locus (cis-eQTLs) or genome-wide (trans-eQTLs). Loci that are associated with both gene expression and disease risk implicate particular genes as potential biologically relevant candidates. A limitation of eQTL studies is difficulty in obtaining large sample

sizes in relevant tissues. The largest eQTL studies in over 1000 individuals have generally focused on easy-to-obtain tissue such as heterogeneous cell types within peripheral blood (Hemani *et al.,* 2014; Westra *et al.,* 2013), while smaller studies (typically with sample sizes in the hundreds) have been performed in cell types such as lymphoblastoid cell lines (LCLs), monocytes (Fairfax *et al.,* 2014), dendritic cells (Lee *et al.,* 2014) and heterogeneous tissues such as liver, adipose tissue, skin and brain (Gibbs *et al.,* 2010; Grundberg *et al.,* 2012; Schadt *et al.,* 2008). Despite having identified hundreds of eQTLs, the majority of the heritability of gene expression remains to be uncovered, much like the case with complex disease risk loci. For instance, in a large eQTL study of LCLs, adipose tissue and skin in 856 twins, the reported cis-eQTLs explain on average only 9-12% of the total genetic variance at each gene (Grundberg *et al.,* 2012). Nevertheless, these studies are an invaluable tool for interpreting the findings from GWAS. Indeed, in Chapters 2-4, eQTL datasets were used to prioritise candidate genes at PBC, PSC and IBD risk loci.

Enrichment analysis provides a complementary approach to linking GWAS risk loci with gene expression. These types of analyses ask whether disease risk loci are found disproportionately more often overlapping certain genomic annotations (for example, coding variants, UTRs, or epigenetic marks) than by chance. For instance, GWAS loci across a range of phenotypes appear to be enriched for known eQTLs (Nicolae *et al.,* 2010). Under the further assumption that disease loci act in only a small number of cell types and under certain cell states, questions about the relative importance of specific cells and disease states in disease pathogenesis can also be studied using the enrichment approach. These studies have an advantage over eQTL studies in that genomic annotations can be generated from only a small number of individuals. Such enrichment studies of gene regulatory annotations or genes that are expressed in specific cell types are now common place in the literature (Cowper-Sallari *et al.,* 2012; Ernst *et al.,* 2011; Hu et al.; Liu *et al.,* 2012; Maurano *et al.,* 2012; Trynka *et al.,* 2013).

An important consideration in these types of approaches is the estimation of the null distribution – what amount of overlap, given the number risk loci and

frequency of genomic annotations, is expected just by chance? It is incorrect to assume that functional annotations and risk loci are both randomly distributed across the genome – both are more likely to be found nearer to genes than away from them (Hindorff et al.). Hence it is possible that sets of risk loci associated with any number of traits will be enriched for functional elements purely because of their colocalisation around genes rather than their functional relevance. For this reason, parametric approaches assuming independence or permutation approaches that randomly resample SNPs (while not accounting for LD) or switch case/control labels to construct "null" GWAS datasets may be upwardly biased in their enrichment estimation.

In this study, I combined GWAS results for four immune-mediated and two non-immune related quantitative traits with two differential expression datasets that are relevant to intestinal inflammatory diseases (e.g. Crohn's disease, ulcerative colitis and coeliac disease). The first dataset consists of a gene expression experiment of four intestinal T cell populations and their blood counterparts in healthy individuals (Raine *et al.,* 2014). T cells are the dominant population of immunocytes in the gastrointestinal tract, and display distinct characteristics in their cell surface marker expression, activation pathways and function compared with the blood counterparts. The expression of genes that drive these differences and maintain intestinal homeostasis may be prime candidates to also modulate risk immune-mediated diseases of the gastrointestinal tract.

The second dataset consists of differentially expressed transcripts in mice following infection with the whipworm *Trichuris muris*. Gene expression levels were measured in infected and uninfected populations of heterogeneous cells in cecum tissue (Foth *et al.,* 2014). High dose infections of *T. muris* in mice typically generates a T_H2 response characterised by eosinophil activation, macrophage inhibition and the production of antibodies, such that immunity is acquired. Low dose infection generates a T_H1 response, characterised by macrophage activation other cellular immunity response, ultimately leading to chronic infection. These low dose infections have been used to model the response in humans to infection

by *Trichuris trichuira*, which exhibit striking phenotypic similarities to IBD (Levison *et al.,* 2013; Levison *et al.,* 2010). Early exposure to whipworms in humans is also thought to be protective against IBD, and the hygiene hypothesis suggests that a lack of exposure to pathogens has contributed to the increasing incidences of immune-mediated disorders in developed countries (Elliot *et al.,* 2000; Okada *et al.,* 2010). Furthermore, there is some evidence that by triggering an immune response, whipworms are an effective treatment for IBD (Croese *et al.,* 2006; Summers *et al.,* 2005a; Summers *et al.,* 2005b). For these reasons, if genes that are differentially expressed upon infection are enriched in risk-loci for IBD and other immune-mediated diseases, they may be excellent candidates through which disease is mediated.

5.1.1 Contributions

Generation of gene expression datasets and identification of differential expressed genes were performed by Tim Raine, Adam Reid and others, and are described in Raine *et al.* (2014) and Foth *et al.* (2014). All other analyses were performed by myself.

5.2 Methods

5.2.1 Human T cell transcripts

Differential gene expression data were obtained from Raine *et al.* (2014). Briefly, six healthy subjects underwent biopsy collection at the terminal ileum. These samples were sorted using fluorescence activated cell sorting (FACS), and total RNA from four major T effector memory cell populations isolated: $CD4^+$ and $CD8^+$ expressing intraepithelial lymphocytes (IELs), and $CD4^+$ and $CD8^+$ expressing lamina propria lymphocytes (LPLs). Paired reference $CD4^+$ and $CD8^+$ T cells from the peripheral blood were also isolated. Gene expression was measured using the Affymetrix Gene ST 1.0 microarrays. After QC filtering, expression of 9,468 transcripts that passed in all six cell populations were obtained. Differential expression was analysed pairwise with each gut T cell population paired with its corresponding peripheral blood population taken from the same individual

($CD4^+$ IEL vs. $CD4^+$ blood, $CD4^+$ LPL vs. $CD4^+$ blood, $CD8^+$ IEL vs. $CD8^+$ blood, and $CD8^+$ LPL vs. $CD8^+$ blood). Transcripts that were significantly up-or-down-regulated in either IEL or LPLs vs. blood were taken forward for enrichment analysis.

5.2.2 Mouse cecum transcripts

Differential expression data were obtained from Foth et al. (2014). Briefly, 14 male C57BL/6 were infected with a low dose of *T. muris* (25 eggs by oral gavage) at 6-8 weeks of age. The section of the cecum where the worms reside and those without infection were extracted. Transcriptome libraries for RNA-seq were created following standard Illumina protocols and sequencing was performed on Illumina HiSeq 2000 machines. The number of reads per gene was calculated by summing over all transcripts that map to the gene. Genes that showed differential expression between the infected cases and uninfected controls were estimated at a false discovery rate of 5% using DESeq (Anders and Huber, 2010). Only protein coding genes and those with a unique human orthologue were included for downstream analysis. After filtering, 15,278 genes remained.

5.2.3 GWAS enrichment

The SNP with the strongest association signal (the lead SNP) in each of the associated loci (reported at $P < 5\times10^{-8}$) from the largest published genome-wide association studies (GWAS) were extracted for four immune-mediated complex diseases: Crohn's disease (CD), ulcerative colitis (UC), celiac disease (CeD) and type 1 diabetes (T1D) (Barrett *et al.,* 2009; Jostins *et al.,* 2012; Trynka *et al.,* 2011b), as well as two complex traits: height and body mass index (BMI) (Lango Allen *et al.,* 2010; Speliotes *et al.,* 2010). The two complex traits are unlikely to be strongly influenced by immune-related genes and were included as effective negative controls for the method. For each lead SNP, an associated locus was defined as the genomic region spanning a 0.2cM window either side of the lead SNP, estimated from HapMap Phase II genotypes (The International HapMap Consortium 2007). Where SNPs showed overlapping windows, only the window assigned to the SNP with the most significant p-value was considered.

For each differentially expressed gene, I defined its gene-region spanning ±50kb window from the gene's transcription start/stop site. To account for potential non-random clustering of genes with similar expression patterns and function (Hurst *et al.,* 2004), groups of differentially expressed genes that have overlapping windows were combined into a single window.

For each GWAS phenotype, the number of times a risk locus overlaps with at least one differentially expressed gene-window was counted. To assess the statistical significance of this overlap, I randomly sampled the same number of differentially express genes from the full list of expressed genes. If a sampled gene has a ±50kb window overlapping that of another previously sampled gene, then the windows are merged and these genes are only counted once. I then calculated the number of associated loci that overlap at least one of these randomly sampled lists of genes. The sampling process was repeated 100,000 times for each disease/trait, and the empirical p-value was the number times the overlap with the randomly sampled genes exceeds the overlap with the observed differentially expressed genes, divided by 100,000.

5.3 Results

5.3.1 Human T cell transcripts

Using a 1.4-fold change (adjusted $P < 0.05$), 246, 275, 115 and 142 genes were identified to be upregulated in LPL $CD4^+$, LPL $CD8^+$, IEL $CD4^+$ and IEL $CD8^+$ T cells respectively compared with their counterparts in the blood. Using a P-value cut-off of $P < 2\times10^{-3}$ (equivalent to a 5% Bonferroni correction for 24 tests), a significant enrichment among T1D risk loci were identified for genes upregulated in LPL $CD4^+$ ($P = 10^{-5}$) and LPL $CD8^+$ cells ($P = 10^{-5}$), with 17 and 18 respectively of the 54 associated risk loci overlapping at least one upregulated gene. Strong suggestive evidence for enrichment was also identified for upregulated genes in LPL $CD4^+$ cells in CD ($P = 0.0053$) and CeD ($P = 0.0045$), LPL $CD8^+$ cells in CD ($P = 0.0038$), IEL $CD4^+$ cells in T1D ($P = 0.0053$) and IEL CD8 cells in T1D ($P = 0.001$) (Table 5.1). Only modest levels of enrichment were identified in for LPL T cells in UC ($P = 0.037, 0.029$), almost all of which is driven

by UC risk loci that are also associated with CD (Table 5.3). The lack of enrichment in UC may reflect that fact that inflammation occurs in the colon, while the experiments described here were on cells extracted from small bowel biopsies.

		LPL upregulated vs. blood				IEL upregulated vs. blood			
		CD4+ (246)		CD8+ (275)		CD4+ (115)		CD8+ (142)	
Phenotype	Risk loci	Overlap	P	Overlap	P	Overlap	P	Overlap	P
Crohn's disease	140	23	0.0053	25	0.0038	7	0.56	10	0.33
Ulcerative colitis	133	21	0.0368	23	0.0291	5	0.88	8	0.69
Celiac disease	38	10	0.0045	10	0.0104	5	0.0161	5	0.0841
Type 1 diabetes	54	17	10^{-5}	18	10^{-5}	7	0.0053	10	0.0010
Body mass index	73	2	0.98	3	0.94	5	0.55	6	0.044
Height	192	21	0.87	18	0.99	5	0.98	7	0.99

Table 5.1. Enrichment of genes that are upregulated in gut T cells compared with blood T cells in loci associated with six phenotypes. The numbers in parentheses next to each cell type is the number of upregulated genes in that gut cell type vs. its equivalent in blood.

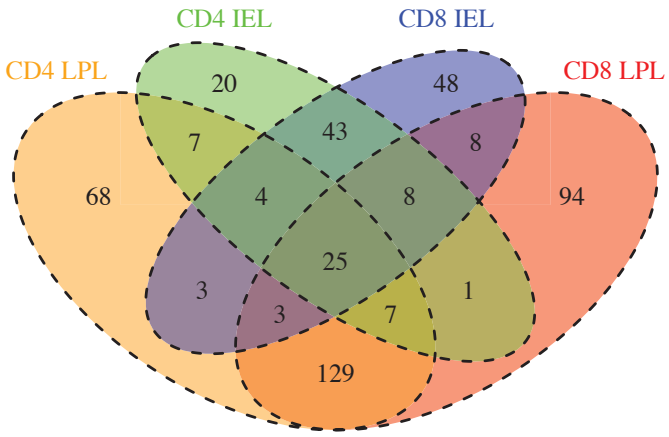

Figure 5.1. Number of upregulated genes that overlap among CD4+ LPL, CD8+ LPL, CD4+ IEL and CD8+ IEL T cells vs. counterparts in blood.

Genes that were downregulated in LPL or IEL T cells compared with their blood counterparts were also tested for enrichment, though no evidence was found for any of the phenotypes ($P > 0.01$) (data not shown). As expected, height and BMI also showed no evidence for enrichment for any of the gene sets tested. These two traits were selected as they include a similar number of associated loci as the immune-mediated diseases tested and, given that immune-related processes are unlikely to play a strong role in these traits, any enrichment

observed in these traits may have been the result of biases in the method that were unaccounted for.

5.3.2 Mouse cecum transcripts

After filtering, 824 genes showed evidence for differential expression (FDR = 5%) between infected and uninfected cecum tissue in C57BL/6 mice. A unique human ortholog was taken forward for 454 of these genes. Significant evidence for enrichment of differentially expressed genes and GWAS risk loci were found for all four immune-related diseases ($P < 0.0024$), the strongest of which were seen in Crohn's disease ($P = 2.0\times10^{-4}$) and ulcerative colitis ($P = 5.7\times10^{-4}$). As with the case for the IEL and LPL T cells, no evidence for enrichment was identified across height or BMI associated loci.

Phenotype	Risk loci	Overlap	P
Crohn's disease	140	34	2.0×10^{-4}
Ulcerative colitis	133	33	6.7×10^{-4}
Celiac disease	38	11	0.0012
Type 1 diabetes	54	15	0.0024
Body mass index	73	6	0.33
Height	192	23	0.52

Table 5.2. Enrichment of genes that are differentially expressed between infected and uninfected cecum tissue among loci associated with six phenotypes.

		CD4+ LPL				CD8+ LPL				CD4+ IEL		CD8+ IEL	Cecum			
Chr.	BP window (Mb)	CD	UC	CeD	T1D	CD	UC	CeD	T1D	CeD	T1D	T1D	CD	UC	CeD	T1D
1	7.78 - 8.27					*ERRFI1*	*ERRFI1*									
1	25.18 - 25.32			*RUNX3*				*RUNX3*								
1	67.90 - 67.9	*IL12RB2, IL23R*	*IL12RB2, IL23R*										*IL12RB2*	*IL12RB2*		
1	77.91 - 78.99	*DNAJB4*				*DNAJB4*										
1	113.82 - 114.62	*PTPN22*			*PTPN22*	*PTPN22*			*PTPN22*		*BCL2L15*	*PTPN22*	*BCL2L15, PTPN22*			*BCL2L15, PTPN22*
1	154.97 - 156.21	*LMNA*	*LMNA*			*LMNA*	*LMNA*						*FDPS*	*FDPS*		
1	160.68 - 161.15												*LY9, SLAMF7*	*LY9, SLAMF7*		
1	172.46 - 172.94									*FASLG*						
1	192.46 - 192.55			*RGS1*	*RGS1*			*RGS1*	*RGS1*	*RGS1*	*RGS1*	*RGS1*				
1	197.19 - 197.94					*DENND1B*	*DENND1B*									
1	206.79 - 207.04					*IL10*	*IL10*		*IL10*							
2	28.58 - 28.67	*FOSL2*	*FOSL2*			*FOSL2*	*FOSL2*									
2	43.45 - 44.05					*ZFP36L2*	*ZFP36L2*									
2	60.78 - 62.12	*REL*	*REL*	*REL*		*REL*	*REL*	*REL*								
2	68.51 - 68.89														*PLEK*	
2	102.69 - 103.27	*IL1R1*	*IL1R1*	*IL1R1*									*IL18RAP*	*IL18RAP*	*IL18RAP*	
2	191.81 - 192.01												*STAT1, STAT4*	*STAT1, STAT4*	*STAT1, STAT4*	*STAT1, STAT4*
2	198.14 - 199.11						*COQ10B*									
2	204.17 - 204.82			*CTLA4*	*CTLA4*			*ICOS*	*ICOS*						*ICOS, CTLA4*	*ICOS, CTLA4*
2	219.30 - 219.3												*CXCR2, SLC11A1*	*CXCR2, SLC11A1*		
2	241.70 - 241.7	*GPR35*	*GPR35*													
3	45.89 - 46.71			*LZTFL1*	*LZTFL1*			*LZTFL1*	*LZTFL1*	*LZTFL1*	*LZTFL1*	*CCR9*			*CCRL2*	*CCRL2*
3	48.17 - 51.83												*BSN, GMPPB, SHISA5, AMT*	*BSN, GMPPB, SHISA5, AMT*		
4	103.37 - 104.25		*NFKB1*													
4	122.90 - 123.77	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*	*IL2*				
5	0.40 - 0.79		*TPPP*													
5	40.18 - 40.98	*PTGER4*	*PTGER4*			*PTGER4*	*PTGER4*									
5	72.38 - 72.59												*TMEM171*			
5	131.28 - 132.14					*KIF3A*	*KIF3A*						*SLC22A4*	*SLC22A4*		
5	150.40 - 150.4	*C5orf62*	*C5orf62*			*DCTN4*	*DCTN4*									
6	0.36 - 0.47			*IRF4*				*IRF4*								
6	127.79 - 128.34												*THEMIS*		*THEMIS*	
6	137.81 - 138.29	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*	*TNFAIP3*					
6	167.34 - 167.55	*CCR6*	*CCR6*			*CCR6*	*CCR6*									
7	26.63 - 27.22	*SKAP2*			*SKAP2*											
7	50.36 - 50.75															*DDC*
7	98.71 - 99.37					*ZNF394*	*ZNF394*									
7	107.60 - 107.6													*SLC26A3*		
7	116.78 - 117.45												*CFTR, CTTNBP2*	*CFTR, CTTNBP2*		
7	128.55 - 128.82													*IRF5*		
8	126.44 - 126.63	*TRIB1*	*TRIB1*													
9	93.86 - 94.17												*NFIL3*	*NFIL3*		
10	6.00 - 6.18	*IL2RA*	*IL2RA*		*IL2RA, PFKFB3*	*IL2RA*	*IL2RA*		*IL2RA, PFKFB3*							
10	30.67 - 30.83	*MAP3K8*	*MAP3K8*			*MAP3K8*	*MAP3K8*									
10	35.10 - 35.97	*CREM*	*CREM*			*CREM*	*CREM*									
10	64.30 - 64.76	*EGR2*	*EGR2*			*EGR2*	*EGR2*									
10	80.94 - 81.15	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*	*PPIF*							
11	60.55 - 60.97												*SLC15A3, CD6*	*SLC15A3, CD6*		
11	61.37 - 61.76	*FTH1*	*FTH1*			*FTH1*	*FTH1*						*FEN1*	*FEN1*		
11	63.81 - 64.58												*STIP1, FERMT3*	*STIP1, FERMT3*		
11	65.13 - 66.08												*LTBP3, CST6, CTSW*	*LTBP3, CST6, CTSW*		
11	95.97 - 96.47						*MAML2*									
11	118.90 - 118.9												*HYOU1*	*HYOU1*	*PHLDB1*	
12	9.47 - 10.02				*KLRB1, CLEC2B*				*CD69*		*CD69*	*CD69*				
12	12.53 - 12.73	*DUSP16*	*DUSP16*			*DUSP16*	*DUSP16*									
12	56.23 - 56.84				*RPL41*				*IL23A*							
12	57.75 - 58.53				*ARHGAP9*				*ARHGAP9*		*ARHGAP9*	*ARHGAP9*				
12	68.32 - 68.63	*IFNG*	*IFNG*			*IFNG*	*IFNG*									
12	111.54 - 113.13											*C12orf51*				
13	99.61 - 100.11	*GPR183*	*GPR183*		*GPR183*	*GPR18*	*GPR18*		*GPR18*			*GPR18*	*GPR18*	*GPR18*		
14	69.14 - 69.36					*ZFP36L1*	*ZFP36L1*	*ZFP36L1*	*ZFP36L1*							
14	75.42 - 75.75	*FOS*	*FOS*			*FOS*	*FOS*									
14	88.19 - 88.73					*GPR65*	*GPR65*						*GPR65*	*GPR65*		
15	74.55 - 75.94														*SEMA7A*	
15	78.91 - 79.26				*MORF4L1*				*MORF4L1*							*CTSH*
16	11.00 - 11.0												*SOCS1*	*SOCS1*	*CIITA, SOCS1*	*CIITA, SOCS1*
16	28.28 - 29.03											*APOBR*				
16	29.89 - 31.36													*CORO1A*		
16	75.03 - 75.53											*ZNRF1*				
16	85.96 - 86.04												*IRF8*	*IRF8*		
17	32.70 - 32.7												*CCL2, CCL7*	*CCL2, CCL7*		
17	37.35 - 38.25												*NR1D1*	*NR1D1*		*NR1D1*
17	40.29 - 41.05	*STAT3*	*STAT3*										*WNK4*	*WNK4*		
17	58.20 - 58.2					*MIR21*	*MIR21*									
18	46.34 - 46.51					*SMAD7*	*SMAD7*									
19	10.37 - 10.63	*ICAM1*	*ICAM1*		*ICAM1*	*ICAM1*	*ICAM1*		*ICAM1*				*ICAM1*	*ICAM1*		*ICAM1*
19	47.14 - 47.33				*SLC1A5*											
20	44.34 - 44.82												*CD40*	*CD40*		
20	62.18 - 62.49												*PTK6*	*PTK6*		
22	21.71 - 22.27												*CCDC116, SDF2L1*	*CCDC116, SDF2L1*	*CCDC116, SDF2L1*	
22	29.81 - 30.87					*SF3A1*	*SF3A1*		*SF3A1*				*UQCR10*	*UQCR10*		*UQCR10*
22	37.50 - 37.57				*IL2RB*				*IL2RB*							*CYTH4, IL2RB, RAC2*

Table 5.3. Annotation of disease-associated loci that are show nominal levels of enrichment ($P < 0.05$) for genes that show differential expression in healthy gut vs. blood T cells and in infected vs. uninfected mouse cecum tissue. The BP (base pair) window denotes a ±0.2cM around an associated SNP. Windows that overlap were combined into a single window.

5.4 Discussion

The broad patterns of enrichment among disease risk loci and genes expressed in both healthy and in inflamed tissues points to the importance of multiple biological pathways involved in disease risk. The lack of overlap between the expression of genes upregulated in healthy human T cell populations and infected/uninfected mouse cecum samples (Table 5.3) reflects both the different cell composition of the samples and biological processes involved in maintaining homeostasis and responses to infection. That differentially expressed genes in T cells from the gut compared with those from peripheral blood appear to play a role in disease risk serves as an important reminder of the limitations of inferring biology from easily accessible blood cell types. Ideally, further understanding of how gene expression modulates disease risk will involve efforts that combine expression patterns multiple immune cell types under both healthy conditions and disease states.

A major utility of gene expression experiments in relevant tissue types is to identify potential candidate genes among GWAS risk loci. Many of the candidate genes listed here (Table 5.3) were also implicated in other *in silico* approaches reported in the original locus discovery projects. For instance the IBD associated SNP rs1819333 lies 160kb upstream of *CCR6*, a gene that is upregulated $CD4^+$ and $CD8^+$ LPL T cells. *CCR6* is an important regulator of lymphocyte homeostasis in the mucosa (Cook *et al.,* 2000), and was implicated as a candidate gene through the text-mining-based GRAIL network analysis in the original IBD GWAS (Jostins *et al.,* 2012; Raychaudhuri *et al.,* 2009). Similarly, at the IBD associated SNP rs11209026, *IL12RB* was differentially expressed in both $CD4^+$ LPL T cells and cecum tissue. This gene was also implicated in the original IBD GWAS via DAPPLE, a method identifies candidate genes based on reported protein interaction networks (Rossin *et al.,* 2011).

At other loci, the approach also offers new leads at loci with no obvious candidate gene, or alternative candidate genes to those previously proposed. For instance, at the IBD-associated SNP rs35675666, GRAIL analysis originally

suggested *TNFRSF9* as the sole candidate gene at this locus. Here, another nearby gene, *ERRFI1*, was highly expressed in $CD8^+$ LPL T cells. *ERRFI1* belongs to a family of epidermal growth factor receptors that share a common signal transduction pathway through ERK-MAPK with the T cell receptor. This growth factor-mediated signalling has been suggested to modulate intestinal T cell regulation in a murine colitis model (Zaiss *et al.*, 2013), highlighting *ERRFI1* as an alternative candidate gene at this locus. Similarly, at rs17391694, the nearby gene *DNAJB4* was highly expressed in LPL $CD4^+$ and $CD8^+$ T cells. No candidate genes were reported in the original IBD GWAS at this locus, partly reflecting the fact that *DNAJB4* has only recently been described.

Notably, T1D loci also appeared to be enriched for genes differentially expressed among the intestinal tissue described. Even though T1D does not manifest itself in the intestines, part of this enrichment may be a reflection of risk loci that are shared between T1D and the other intestinal diseases tested here. However, several genes residing near T1D-specific risk loci were also observed to be differentially expressed across all the experiments (Table 5.3). There is evidence to suggest that intestinal microbiota not only modulates local inflammation, but also systemic immune-mediated pathologies (Kamada *et al.,* 2013). Moreover, interactions between gut microbiota and the innate immune system have been suggested to partly modulate risk for T1D in mice (Wen *et al.,* 2008). The genes here that appear differentially expressed in populations of intestinal cell types may offer insights in the host-environment interactions across systemic immune-mediated disorders.

The method I described for estimating the degree of enrichment is in line with similar approaches that look to test whether a set of genes is overrepresented by genes from another pre-defined and biologically relevant gene set. Perhaps the most popular of these, Gene Set Enrichment Analysis (GSEA), was developed to estimate whether a set of genes identified from microarray experiments were enriched for genes involved in various biological pathways (Subramanian *et al.,* 2005). The advent of GWAS has spawned a

number of GSEA-type methods for analysing biological pathways that are enriched among GWAS risk loci (reviewed in Wang *et al.* (2010b)).

In the original GSEA approach, a set of genes is first identified and ranked (e.g. according to differential expression P-value between a set of cases and controls), and then tested to see if this rank correlates with a set of genes from another set (e.g. a particular biological pathway) via Kolmogorov-Smirnov-like statistics (Subramanian *et al.,* 2005). Significance is then assessed via permutation of the case-control status and repeating the original analysis in order to obtain a null distribution of correlations. In the context of GWAS, this approach is analogous to permuting case-control status and repeating the GWAS many times – which is both time-consuming and not possible without individual-level genotype data. GWAS adaptations to GSEA have sought to overcome this by only permuting SNP labels on summary GWAS statistics (Zhang *et al.,* 2010), however, this does not account for the correlated structure of SNPs due to LD. Furthermore, neither the phenotype-label nor SNP-label permutation approach takes into account the fact that SNPs that are associated with a complex trait are not randomly distributed throughout the genome, but are rather more likely to be found near functional elements such as genes or regulatory regions.

The approach described here tries to overcome these biases by permuting the set of differentially expressed genes rather than risk loci. While this accounts for both LD and the non-random distribution of risk loci, our method may also be biased by gene size and correlation of expression patterns of certain genes. Larger genes are more likely to overlap with an associated risk locus, such that permuting sets of genes will not be a true reflection of the null distribution. In the T cell datasets, there was modest evidence that differentially expressed genes were longer than the total set of genes tested, potentially inflating enrichment estimates (Figure 5.2 A). The opposite appeared to be the case for the cecum tissue, where the length of differentially expressed genes were shorter than expected, potentially making the test more conservative (Figure 5.2 B).

Similarly, the permutation approach will not truly estimate a null distribution in situations of gene-gene expression correlations. Genes with

coordinated expression are often clustered in areas of low recombination (Hurst *et al.,* 2004), and cis eQTLs may affect the expression of multiple nearby genes. I try to overcome this by combining genes that have overlapping windows (±50kb from the transcript start/stop sites) into a single window. Moreover, the empirical P-value is calculated on the number of risk loci that overlap at least one gene region, not the number of gene regions that overlap at least one risk locus. This distinction is subtle, but in situations where a risk locus overlaps more than one differentially expressed gene region, the test is conservative since these genes only count towards a single overlap, yet multiple genes are sampled during the permutations. Had the empirical P-value been calculated instead on the number of genes that overlap a risk locus, the empirical P-value may have been inflated as now multiple genes can potentially overlap with a single risk locus (Dixson *et al.,* 2014). Nevertheless, the approach will not account for situations where coexpressed genes lie far away from each other.

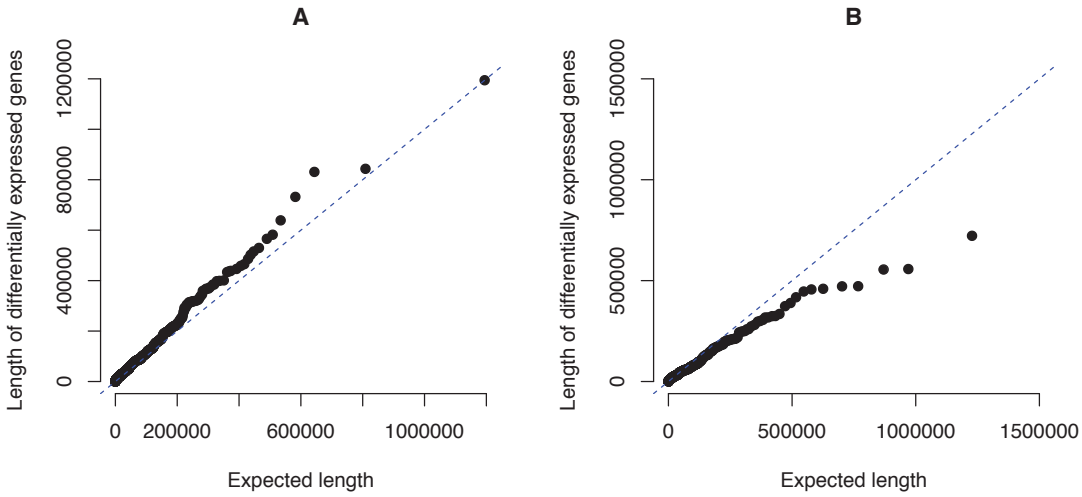

Figure 5.2. Quantile-quantile plots of gene length of differentially expressed genes in (A) gut T cells vs. blood and (B) infected vs. uninfected cecum tissue. The distribution of the expected length was the empirical distribution of all genes tested for differential expression in the respective experiments.

The choice of thresholds when defining locus and gene boundaries is often subjective. In this study, a ±0.2cM window around an associated SNP and a ±50kb window around a gene's transcript start/stop positions were used to

define whether an associated locus overlaps with a gene. The 0.2cM window describes the boundaries in which a causal variant that is tagged by an associated SNP may lie. The same window size was also used in the design of Immunochip high density regions (Tsoi *et* al., 2012 and Jostins, 2012). Similarly, the 50kb gene boundary region was chosen to adequately encompass regions where variants that affect that gene's expression may reside. This window size captures the majority (>93%) identified cis-eQTLs (Veyrieras *et al.*, 2008), though there are examples of some genes with cis-eQTLs greater than 100kb away from a transcription start site (Stranger *et al*., 2012 and Veyrieras *et al*., 2008). Larger windows may lead to more SNPs incorrectly assigned to genes, as well as a greater chance that independent loci overlap. In this study, if SNPs are incorrectly assigned to genes, power will decrease as more noise is introduced. A larger gene-boundary window will also mean that more differentially expressed genes will overlap each other and merged together. Since the resampling process cannot explicitly take this overlap into account, the results may be upwardly biased. On the other hand, using more stringent boundaries may also reduce power if truly regulatory SNPs are not assigned to its corresponding gene.

In Hu *et al.,* (2011) a similar approach looking at the overlap between gene expression in a set of immune cells and GWAS risk loci is described. Promisingly, they try to overcome the potential biases described by estimating the null distribution of enrichment by randomly selecting SNPs from a predefined, LD-pruned set of SNPs that have similar properties to disease-associated SNPs in terms of the number of genes that are located nearby. The accuracy of this approach of course depends on how this set of null SNPs is estimated, and will be more accurate for diseases where there are a large number of associated loci, such that a more representative set of null SNPs can be generated.

5.4.1 Conclusions

In summary, this study describes an approach testing whether disease risk loci are enriched for a set of functionally relevant genes. Evidence for enrichment provides additional candidate genes at associated loci, as well as generating hypotheses as to how these genes mediate disease. There was evidence for

enrichment among risk loci in four immune-mediated disorders with two differential expression datasets – the first comparing T cell subsets in healthy gut tissue with blood counterparts, and the second from samples in the cecum of mice in the presence or absence of *T. muris* infection, implicating processes in both maintaining intestinal homeostasis and response to infection in disease risk. There is a great deal of potential in these integrative approaches as a greater number of functional genomic datasets are generated for a range human tissue across multiple disease states, though care must be taken to ensure that methods employed are unbiased and statistically robust.