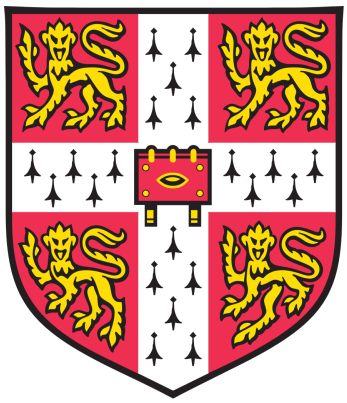


Using three dimensional chromatin structure to understand autoimmune disease mechanisms

First Year Report

Oliver S Burren

ob219@cam.ac.uk

under the supervision of Chris Wallace

August 31, 2016

This report is submitted as part requirement for the PhD Programme in Medical Science at Cambridge University. Much of the work described in this report is collaborative. My specific contributions include the assembly of a collection of GWAS summary statistics, the development, implementation of *blockshifter* and COGS algorithms and interpretation of subsequent analysis. The report may be freely copied and distributed provided the source is explicitly acknowledged.

Abstract

Genome wide association studies (GWAS) have uncovered hundreds of genetic regions that are responsible for human susceptibility to autoimmune disease. Identification and functional characterisation of causal variation, a pre-requisite for therapeutic intervention, has proved difficult. This challenge stems not only from the underlying genetic architectures that are often refractory to the statistical resolution of causal variation, but also, due to complex and tissue specific three dimensional (3D) chromatin organisation, the cryptic nature of the genes that such variation targets. Recent empirical developments have enabled high resolution maps of 3D chromatin organisation to be elucidated using promoter capture Hi-C (PCHi-C) that might allow the physical linkage of causal variants with target genes. The aim of my PhD is to develop statistical methods to integrate GWAS with these tissue specific PCHi-C maps and other genomic data in order to better understand the biology of autoimmune susceptibility.

I developed *blockshifter*, a competitive circularised permutation method for examining promoter interacting regions (PIRs) for trait associated variant enrichment, that allows for correlation between variants and interactions. By applying *blockshifter* to publicly available summary statistics for eight autoimmune and 23 non-autoimmune traits, and PCHi-C maps for 17 haematopoetic tissues I found enrichment for autoimmune variants in lymphoid tissues, which was strongest in activated $CD4^+$ T cells. Next, I developed a Bayesian method, COGS, to generate PCHi-C supported gene scores to prioritise variants, genes and tissue contexts for functional validation. I applied COGS to prioritise 2,604 genes across all 31 traits and tissues. Qualitatively, COGS gene scores were more specific and sensitive when compared to similar methods using either proximal gene variants or topologically associated domain (TAD) boundaries. I found that genes prioritised by COGS were significantly enriched for genes differentially expressed between healthy and diseased immune subsets ($P = 0.002$ - ulcerative colitis). In contrast, scores for TAD and proximal methods showed no enrichment.

I applied COGS to a set of six autoimmune traits for which dense fine mapping summary statistics were available in `http://www.immunobase.org`. To assess the effect of multiple proximal, but independent variants, on scores, I modified COGS to accept marginal posterior probabilities computed using GUESSFM, for four traits, for which full genotype data was available. Due to *blockshifter* results I limited initial analysis to PCHi-C maps for activated and non-activated $CD4^+$ T cells, prioritising 256 genes. One of these, *IL2RA*, was prioritised in $CD4^+$ T cells across multiple autoimmune diseases. Allele specific expression analysis of rs61839660 by a collaborator, provided support for a PIR in intron 1 modulating expression of *IL2RA*, in unactivated but not activated contexts.

In future work I will: a) complete detailed tissue specific analysis across all 17 tissues, b) explore approaches to the integration of other relevant genomic data sets, c) use simulated GWAS statistics to better characterise COGS scores, d) develop tissue specific gene set enrichment methods based on COGS scores.

Contents

Chapter 1

Introduction

1.1 Motivation

A key property of the adaptive immune system is the ability to recognise pathogens from self-antigens. Dysregulation of this process results in damage to healthy tissues and autoimmunity. Currently, over 80 diseases have been found to have an underlying autoimmune pathogenesis, with approximately half presenting as rare diseases [Hayter and Cook, 2012]. The collective health burden of more common autoimmune diseases such as type 1 diabetes (T1D), rheumatoid arthritis (RA), inflammatory bowel disease (IBD) and multiple sclerosis (MS) is high with approximately 7 - 9% of the European population affected [Cooper et al, 2009]. Although environment is a contributing factor in disease susceptibility, the genetic heritability, defined as the proportion of phenotypic variance attributable to genetic variability, is also important, ranging from 0.39 in primary biliary cirrhosis (PBC) to 0.9 in ankylosing spondylitis (AS) [Gutierrez-Arcelus et al, 2016].

Genome wide association studies (GWAS) have been instrumental in understanding the complex genetic architecture underlying autoimmune disease susceptibility with at least 324 distinct genetic loci robustly associated with one or more autoimmune diseases (`http://www.immunobase.org`, accessed 01/08/2016). Focus is now drawn to elucidating the mechanisms by which underlying causal variation modulates phenotypic endpoints, a prerequisite for successful therapeutic development. Interestingly, the majority of associated variants fall outside of genes [Maurano et al, 2012] and the integrative analysis of chromatin marks with GWAS highlights a tissue specific specific regulatory role [Farh et al, 2015]. Recent studies [Claussnitzer et al, 2015; Davison et al, 2012; Smemo et al, 2014] have shown that such regulatory variants might, through chromatin conformation, regulate distal genes. However, further progress in this area has been hampered by incomplete knowledge of causal variants and their target genes and the specific tissue contexts in which they operate [Albert and Kruglyak, 2015].

To date systematic methods for incorporating physical interactions between variants and their target genes have not been attempted. In this report I describe tools and analytical methods to integrate genetic and high resolution promoter capture Hi-C (PCHi-C) data sets. I focus method development on using GWAS summary statistics for input, as due to legitimate privacy concerns, access to raw genotype data is more restrictive, limiting the number of traits that can

be analysed. These methods provide a data driven approach to prioritising the putative causal variants underlying autoimmune disease susceptibility, and relevant tissue contexts, genes and biological pathways within which they operate. I apply these methods to integrate 17 PCHi-C maps for primary human cells with GWAS and fine mapping data across 31 traits. Based on these results I extend this analysis to examine summary ImmunoChip dense mapping summary statistics for ten autoimmune traits in the context of $CD4^+$ T cell activation.

1.2 Fine mapping

One of the first steps of this project is to identify optimal methods for fine mapping, that balance data availability, resolution and computational efficiency. Fine mapping is the process of refining association signals in a genomic region in order to characterise fine scale genetic architecture, a necessary step in order to identify putative causal variants. Progress in this area is challenging, due to the presence of linkage disequilibrium (LD), which in many cases means that the causal variation cannot be resolved statistically with current sample sizes [Li and Kellis, 2016].

One approach, under the simplifying assumption that a single causal variant explains the effect within a given genetic region, is to apply Wakefield's synthesis of approximate Bayes Factors (ABF) [Wakefield, 2009] to each variant within the region. These can be converted into posterior probabilities for a SNP to be causal using methods described in The Wellcome Trust Case Control Consortium et al [2012]. A sizeable benefit of this approach is that it circumvents privacy issues, as the input is limited to GWAS summary statistics rather than access to full genotyping data.

Wakefield defines the ABF as

$$ABF = \sqrt{1-r} \times \exp\left[\frac{Z^2}{2} \times r\right] \tag{1.1}$$

If a given set of summary statistics include $\hat{\beta}$, an estimate of the log(Odds Ratio) and $\sqrt{V}$, the standard error of the Odds Ratio for a variant, then we can compute $Z = \frac{\hat{\beta}}{\sqrt{V}}$. Alternatively, if only univariate p-values are supplied a Z score can be estimated using an inverse normal cumulative distribution function. If r, a shrinkage factor, is defined as the the ratio of the prior variance on $\hat{\beta}$ to the total variance, then, $r = \frac{W}{V+W}$. V is approximated using the variant minor allele frequency (MAF) and study sample size such that $V = \frac{1}{2Nf(1-f)}$ for quantitative traits, where N is number of samples and f is the MAF. In the case/control setting $V = \frac{1}{2Nf(1-f)s(1-s)}$ where s is the proportion of cases. The value of $\sqrt{W}$, the standard deviation of a normal prior for β, depends on study design considerations. In this report we use values of 0.15 and 0.2 for case/control and quantitative trait settings respectively, as discussed by Giambartolomei et al [2014].

Given the ABF for all SNPs in a given genetic block and letting π_i be our prior probability that SNP i is causal for a trait, we can estimate the posterior probability that a SNP i is causal using the following.

$$\begin{aligned} PP_i &= P(\text{SNP}_i \text{ causal}|D) \\ &= \frac{P(D|\text{SNP}_i \text{ causal})\pi_i}{\sum_j P(D|\text{SNP}_j\text{causal})\pi_j + P(D|H_0)\pi_0} \\ &= \frac{\frac{P(D|\text{SNP}_i \text{ causal})}{P(D|H_0)}\pi_i}{\sum_{j=1}^{n} \frac{P(D|\text{SNP}_j \text{ causal})}{P(D|H_0)}\pi_j + \pi_0} \\ &= \frac{\text{BF}_i\pi_i}{(\sum_{j=1}^{n} \text{BF}_j\pi_j) + \pi_0} \\ &\approx \frac{\text{BF}_i\pi_i}{(\sum_{j=1}^{n} \text{BF}_j\pi_j) + 1} \end{aligned} \tag{1.2}$$

Note that as $\pi_0 = 1 - \sum_j \pi_j \approx 1$.

However, current sample sizes are under powered and therefore have limited resolution, which has driven the the development of integrative techniques that couple fine mapping methods with functional information, such as DNase accessibility to further prioritise causal variants. Examples include *fgwas* [Pickrell, 2014], PAINTOR [Kichaev et al, 2014], RiVIERA [Li and Kellis, 2016] and the integration of eQTL data sets using Mendelian randomisation [Zhu et al, 2016]. As an example Pickrell [2014] developed a Bayesian hierarchical framework implemented in *fgwas* that estimates variable priors for each SNP based on other sources of annotation. For a given SNP i, in region j, the prior probability for association, π_{ij}, varies depending on the enrichment of annotations, λ_l.

$$\pi_{ij} = \frac{e^{x_i}}{\sum_{p \in S_j} e^{x_p}} \tag{1.3}$$

where x_i is the sum of the effect of all the annotations that the i^{th} SNP overlaps as shown below. Here λi is the effect of annotation l and I_{il} is and indicator function as to whether SNP i overlaps annotation l.

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il} \tag{1.4}$$

λ_l is itself of interest as it indicates whether a given annotation is enriched for association with a trait of interest, discussed further in Section 1.3. When combined with equation 1.2, π_{ij} allows the incorporation of functional information with genetic data allowing a modest increase in resolution and a resultant decrease in the number of causal SNPs to be considered [Pickrell,

2014]. Whilst promising *fgwas* requires careful cross-validation to prevent over fitting whereby the derived model is biased for the training set but performs poorly on a test set of data, and all such methods are limited by the input annotations available.

Where raw genotyping data for a study is available then stepwise regression could be used to fine map, however whilst this approach is attractive computationally, doubts exist over its validity [Miller, 1984]. Having selected a single variant that best explains the variance of a trait, stepwise regression looks for other variants that explain additional trait variance conditional on this 'top' variant. This is not equivalent to explaining which variants jointly explain the variance of a trait. However, approaches that search the model space exhaustively are only computationally feasible for simple models incorporating a limited number of variants. An alternative approach is to use Monte Carlo methods to sample the model space allowing the consideration of multiple causal variants within a genetic region. One example is GUESSFM that uses a Bayesian evolutionary stochastic search algorithm to effectively sample the model space, and has been shown to have consistently better performance when variants are highly correlated Wallace et al [2015].

1.3 Variant set enrichment

The next part of my project investigates whether the putative causal SNPs identified can be used to better understand the biology of autoimmune disease. If these variants are enriched for a particular annotation or gene set then this can provide global information on mechanisms, relevant tissue contexts and biological pathways. However this is complicated by both correlation between variants (due to linkage disequilibrium) and annotations.

$$\mathrm{Var}(A+B) = \mathrm{Var}(A) + \mathrm{Var}(B) + 2\mathrm{Cov}(A,B) \tag{1.5}$$

If A and B are non independent variables then the covariance term $2\mathrm{Cov}(A,B)$ can be almost as large as $\mathrm{Var}(A) + \mathrm{Var}(B)$. This causes the observed variance to be inflated compared to the theoretical variance if independence is assumed. Unfortunately most classical statistical tests based on linear sums of variables assume independence and thus the inflated variance of the computed test statistic increases type 1 error, that is the erroneous rejection of the null hypothesis. Empirical methods can be used to better estimate this variance. There are three broad sets of methods. Firstly, permuting case/control status and rerunning the original GWAS analysis, however this is very computationally expensive and requires access to the raw data [Evangelou et al, 2014]. A second approach, is to use a suitable reference genotype set to compute covariance matrices, allowing the estimation of test statistic variance under the null hypothesis using the multivariate normal [Burren et al, 2014; Liu et al, 2010]. This approach is attractive as it obviates the need for raw genotyping data, however it is computationally expensive and scales exponentially with SNP density. To overcome this an LD pruning strategy is employed, result-

ing a reduction in resolution, limiting application to larger genomic annotations such as genes. The third approach is to permute the annotations over the SNPs whilst maintaining the spatial correlation structure of the target annotation to estimate the variance under the null. The latter approach is attractive as it favours high resolution annotations and is computationally more tractable. GOSHIFTER is a recent implementation using a circular permutation strategy [Trynka et al, 2015] that demonstrates this approach finding enrichment for H3K4me3 marks in $CD4^+$ memory T cells.

As previously mentioned in Section 1.2, integrative approaches such as *fgwas* combine variant set enrichment with fine mapping strategies to simultaneously prioritise annotations and variants. This approach additionally benefits from jointly assessing multiple annotations, thus if there is additional correlation between different annotation types this can be adjusted for. It should be noted that all methods are dependent on the quality and coverage of input annotations. A recent study fine mapping IBD causal variants found that 21 variants with extremely high probability ($> 95\%$) to be causal did not overlap any functional annotations, drawing attention to the inadequacies of current genomic functional annotation [Huang et al, 2015a].

A major set of limitations inherent in the above analyses stems from the methods employed to organise genes into sets categorised by some functional annotation. Publicly available pathway databases include Gene Ontology [Ashburner et al, 2000], KEGG [Kanehisa and Goto, 2000] and Reactome [Milacic et al, 2012], that rely on domain experts for curation and are under continuous development. This expert curation can often be a confounder when they are used in the variant set enrichment setting, as genes implicated through GWAS results are often used to augment disease specific pathways. Furthermore, the complex, pleiotropic and hierarchical nature of the gene sets derived from such sources means that care needs to be taken not to double count genes that exist in multiple pathways when computing enrichment statistics. Another issue is the stability of gene sets over time, as greater biological insight is gained, this can mean that significant results obtained with one version of the database might over time be eroded. Most fundamental is the fact that these resources cannot create novel biological insight beyond the re-purposing of an existing pathway within a novel context, indeed one of the main utilities for such analysis is providing support for existing findings, as employed in Section 3.3.

Another source of gene sets is from differential analysis of high through put genomic data such as transcriptomics. In this context a specific hypothesis might be tested using either public or private data sets to see if differentially expressed genes are enriched for trait associations. Such approaches can reveal novel insights into underlying biology, for example, I have previously used transcription factor binding perturbation experiments to implicate specific factors that might have a role in type 1 diabetes [Burren et al, 2014]. Finally a more hybrid approach shows promise where gene sets are derived from the overlap of carefully curated transcriptomics experiments that are grouped on the basis of broad functional classes, for example the MSigDB Hallmark gene sets [Subramanian et al, 2005]. These provide the basis of a more exploratory analysis overcoming some of the limitations previously mentioned.

1.4 High resolution promoter capture Hi-C

The most novel aspect of this project is using a data driven approach to link putative causal variation to the target genes in relevant tissues. Techniques such as Hi-C [Lieberman-Aiden et al, 2009] and Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) [Fullwood et al, 2009] have been developed to map the genome-wide chromatin interactions in specific cell types [Rao et al, 2014]. Hi-C involves cross-linking genomic DNA with formaldehyde resulting in covalent links between spatially adjacent chromatin segments. This chromatin is then digested with a restriction enzyme and sticky ends are filled in with biotin labelled nucleotides. Ligation is then performed under dilute conditions which favours intra-molecular ligation events. The DNA is then purified and then sonically sheared and fragments are then enriched for biotinylated junctions, which then undergo paired end sequencing [van Berkum et al, 2010]. ChiA-PET again involves using formaldehyde to covalently link spatially adjacent chromatin. After sonification an immunoprecipitation step is used select a protein of interest (e.g. transcription factor) and the chromatin to which it is bound. Next biotin conjugated linker sequences are ligated to the free ends of the immunoprecipitated DNA. A further proximity based ligation takes place with enrichment of biotinylated linkers. Barcodes within the linker sequences are used to resolve chimeric intra and inter complex interactions.

Whilst both are genome-wide methods for interrogating chromatin looping, CHiA-PET targets a specific protein, and therefore, compared to Hi-C, is unable to give an unbiased overview of all chromatin interactions within a given tissue context. However, Hi-C resolution is limited by two main factors. Firstly the protocol involves a restriction enzyme digest, usually *Hind*III, and interactions are called based on the resultant fragments generated, in practice if *Hind*III is used this limits resolution to approximately 4Kb. Secondly, the complexity of the sequence libraries generated means that to increase the effective resolution of conventional Hi-C by a factor n requires n^2 sequencing reads which is prohibitive for general implementation [Jäger et al, 2015]. Promoter capture Hi-C (PCHi-C) incorporates a sequence capture extension to classical Hi-C to enrich for chromatin interactions with protein coding gene promoters [Mifsud et al, 2015]. Such PCHi-C allows for increasing resolution in an approximately linear fashion with increased sequence depth allowing for an economically viable approach for identifying promoter interactions.

Chapter 2

Materials and Methods

2.1 Promoter capture interaction maps of 17 haematopoetic primary human cells

I obtained maps of 17 primary human cells of the haematopoetic lineage that were generated collaboratively with members of the Fraser, Spivakov, Ouwehand and Diabetes and Inflammation Laboratories. Each cell type was assessed over an average of 3 biological replicates (Appendix Table 1). The bespoke capture platform employed encompassed a total of 22,076 *Hind*III fragments containing 31,253 annotated promoters for 18,202 protein-coding and 10,929 non-protein coding genes (Ensembl v75). Significantly interacting regions were called using the CHiCAGO pipeline [Cairns et al, 2016]. Interactions with a CHiCAGO score threshold of $>$ 5 were used in all downstream analysis. These maps consist of a bait *Hind*III fragment and a list of interacting *HindIII* fragments or promoter interacting regions (PIR) within a particular cellular context.

2.2 Collection and quality control of GWAS summary statistics from 31 genome wide association studies

I downloaded GWAS summary statistics, covering 8 autoimmune and 23 other traits, from online resources, as detailed in Appendix Table 2. Genotyping error can create false associations, therefore I filtered all association statistics to include only robust associations by removing those SNPs which were genome-wide significant ($p < 5 \times 10^{-8}$) but for which no variants, within 500Kb and in LD ($r^2 > 0.6$) with the lead SNP, had $p < 1 \times 10^{-5}$. Finally I removed any SNP that was genome-wide significant but was not found in the 1000 Genomes PhaseIII EUR genotype set.

2.3 Poor Man's Imputation (PMI) - Imputation of GWAS p-values to the 1000 Genome reference panel in the absence of effect size and direction

There was a high degree of variation in the coverage of GWAS summary statistics obtained, some studies contained information on approximately 5×10^5 variants where as others were imputed to 1000 genome reference genotype set and contained in excess of 7×10^6 variants. Imputation can be used to compute approximate association statistics for missing variants, however, it requires access to underlying genotype data, which in this case was unavailable for most traits. Methods exist for imputing summary statistics in the absence of genotyping data such as *GCTA* [Yang et al, 2011] and *IMPG* [Pasaniuc et al, 2014], however these rely on access to either signed Z scores, odds ratios or β coefficients and their standard errors, in order to estimate direction of effect, which are not always available. I therefore designed an alternative 'best guess' method, poor man's imputation (PMI), which requires evaluation but allows the processing of a wide range of traits for which variant coverage is heterogeneous and only univariate p values are available.

The pipeline I developed, approximates the p-value for missing SNP summary statistics for a given study using a suitable reference genotype set. Firstly the genome is split into regions based on a recombination frequency of 0.1cM using HapMap recombination rate data. For each region we retrieve from the reference genotype set (1000 genomes EUR cohort) all SNPs that have MAF $> 1\%$ and use these to compute pairwise LD. The pipeline pairs each SNP with missing p-values to the SNP with maximum pairwise r^2, ${r^2}_{max}$, if that ${r^2}_{max} > 0.6$, and impute the missing p-value as that at the paired SNP. SNPs with missing data or without a pair above threshold are discarded as are SNPs that are included in the study but do not map to the reference genotype set.

2.4 Causal variant posterior probabilities for 31 traits using GWAS summary statistics

To fine map candidate causal variants I used Wakefield's synthesis of approximate Bayes factors (ABF) [Wakefield, 2009] in connection with the method described in The Wellcome Trust Case Control Consortium et al [2012] (see Section 1.2), to compute posterior probabilities for each SNP within 0.1 cM regions, using R code adapted from the *coloc* package [Giambartolomei et al, 2014].

I set the value of prior of the any variant being causal (π_i) to that from Giambartolomei et al [2014], 10^{-4}, which means that we expect 1 in 10,000 SNPs to be causal for a trait. This framework assumes a model where either no SNPs are causal within a region or that exactly one SNP is causal. I masked the MHC region (GRCh37:chr6:25-35Mb) from all downstream

analysis due to its extended LD and known strong and complex association with autoimmune diseases

2.5 Comparison of posterior probabilities between PMI and classical imputation

Before using PMI posterior probabilities in further analyses it was necessary to assess whether they approximated posterior probabilities derived from classical imputation. Firstly I selected all SNPs mapping to Chromosome 1 as a representative sample. To create a simulated non-imputed data set I pruned these results to contain only SNPs for which p-values were reported in Stahl et al [2010]. I next ran PMI on this pruned data set and using *bedtools* [Quinlan, 2014], merged these with actual the imputed p-values from Okada et al [2014]. I confined my comparison to those SNPs imputed by PMI, a total of 235,412 SNPs

2.6 *blockshifter* - A competitive test for associated variant enrichment in PCHi-C interaction maps

In order to examine the enrichment of GWAS signals in promoter interacting regions (PIRs) I developed a method based on ideas implemented in *GOSHIFTER* [Trynka et al, 2015] to examine the enrichment of GWAS signals in the promoter interacting regions, in order to overcome both linkage disequilibrium (LD) and interaction fragment correlation. *blockshifter* implements a competitive test of enrichment between a test set of PIRs compared to a control set. Firstly the coordinates of the PIRs in the union of test and control sets are retrieved, and PIRs with no overlapping GWAS signal are discarded. As the test is competitive between PIRs in test and control sets, those overlapping both are not informative and therefore are excluded from further analysis. For the remaining PIRs we store the number and sum of overlapping GWAS posterior probabilities and these are used to compute δ, the difference in the means of posterior probabilities between the test and control set. Due to correlation between GWAS signals and between PIRs the variance of δ is inflated, I therefore estimate it empirically using permutation. Runs of one or more PIRs (separated by at most one *Hind*III fragment) are combined into blocks, that are labelled unmixed (either test or control PIRs) or mixed (block contains both test and control PIRs). Unmixed blocks are permuted in a standard fashion by reassigning either test or control labels randomly across blocks taking into account the number of blocks in the observed sets. Mixed blocks are permuted by effectively circularising each block and rotating the labels (figure 2.1). I store the mean posterior probabilities across each possible permuted block. The number of choices at each block is small, but there are many blocks. I then randomly sample from each these precomputed block permutations n times so that the proportion of underlying

PIRs labels is the same as the observed set and use this to compute the set of δ_{null}. I use δ_{null} to compute an empirical Z-score:

$$Z = \frac{\delta - \overline{\delta_{null}}}{\sqrt{V*}} \tag{2.1}$$

Where V^* is an empirical estimate of the variance of δ_{null}.

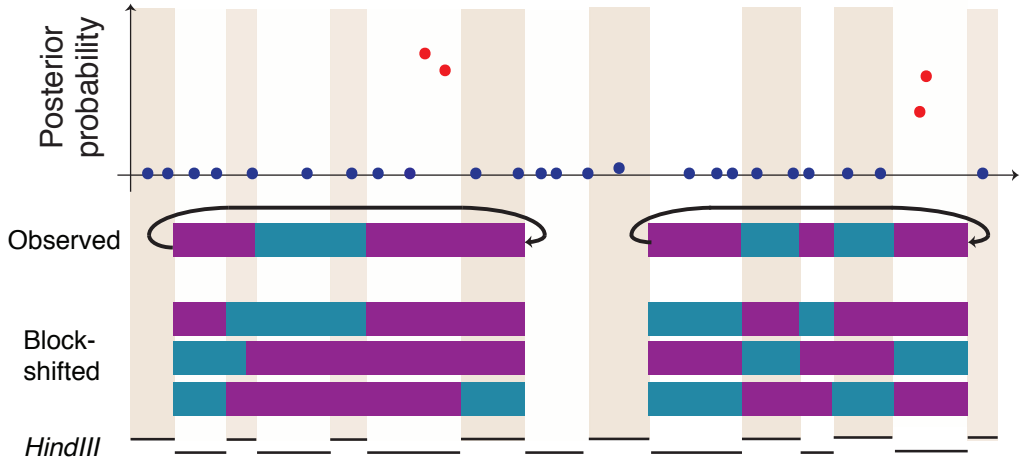


Figure 2.1: Permutation strategy for two 'mixed superblocks' in *blockshifter*. GWAS summary statistics are converted to posterior probabilities for a given SNP to be causal (Red SNPs are likely to be causal). Mixed superblocks are runs of adjacent *Hind*III PIRs found in either test (purple) or control (blue) tissue sets that are separated by two or more *Hind*III fragments for which no PIR in either set exists (grey). Posterior probabilities can be assigned to test or control labels based on PIR overlap and the sum for each category can be stored. The difference in the weighted mean of the sum of posterior probabilities between test and control sets can be used to compute enrichment. However, to control for inflation due to correlation structure between SNPs and between interactions we rotate the labels of *Hind*III fragments within the mixed superblock to generate a set test and control posterior probabilities under the null. Such a strategy is not required for unmixed superblocks with a single label. By sampling from these null test statistics across the set of mixed and unmixed superblocks (weighted so that we select similar numbers of test and control PIRs to the observed data set) we can rapidly generate an empirical null genome wide. These can be used to adjust the test statistic to account for inflation due to underlying correlation. Figure prepared with M. Spivakov.

2.7 COGS - An algorithm for PCHi-C assisted prioritisation of genes and tissues contexts

Whilst the *blockshifter* method provides information about relevant tissue contexts, the primary goal of this project is to use the PCHi-C information to link causal variation to target gene(s). To do this I developed an algorithm to compute tissue specific gene scores for each GWAS trait, taking into account linkage disequilibrium, interactions and functional SNP annotation (figure 2.2). For each gene annotation, for which we have at least one significant interaction and for every nearby recombination block, the algorithm computes a block gene score that is composed of three components.

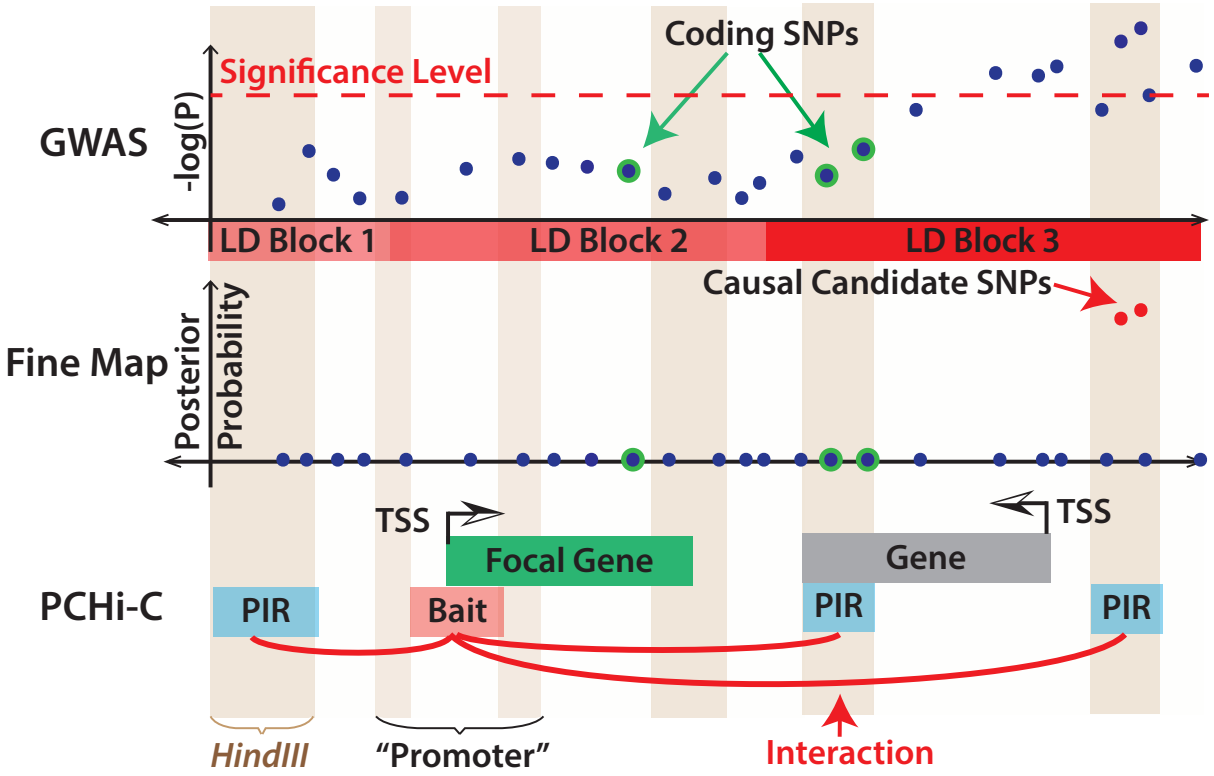


Figure 2.2: A schematic illustration of the COGS (**C**apture Hi-C **O**mnibus **G**ene **S**core) method. GWAS summary statistics are imputed by PMI and converted to posterior probabilities across three linkage disequilibrium (LD) blocks. These are then intersected with PCHi-C maps, promoter interacting regions and their linkages to bait *Hind*III fragments allow the assignment of posterior probabilities to a focal gene (green). Due to limitations in the ability to call very short range interactions we create 'Promoter' regions that consist of bait and adjoining *Hind*III fragments. We account for protein coding SNPs by considering those in the focal gene (green) and masking others. Assuming independence between LD blocks we sum the information across categories to compute an overall gene score. In the above example naive analysis might prioritise the non focal gene (grey), however PCHi-C interactions provide additional information supporting the focal gene(green).

1. The contribution due to coding SNPs in the annotated coding gene as computed by

VEP [McLaren et al, 2010].

2. The contribution due to promoter SNPs, which I define as SNPs that overlap the bait and it's pair of flanking *Hind*III fragments and not any coding SNPs.

3. The contribution due to SNPs that overlap interacting other ends for a tissue or set of tissues that do not overlap coding SNPs.

For a given target gene and recombination block, the algorithm derives a block gene score that is the sum of the posterior probabilities of SNPs overlapping each component, which is the probability that there exists a causal variant for this gene in this block, under the assumption that there is at most one causal variant in any block. Assuming independence we can combine blocks to get an overall gene score such that:

$$Gene\ score = 1 - \prod_{j} \left(1 - \sum_{i \in R_j} 1 - PP_i \right) \tag{2.2}$$

Here PP_i is the posterior probability for the i^{th} variant to be causal, $i \in R_j$ is the set of relevant SNPs i in the j^{th} region. Whilst components one and two are fixed for a given gene and trait, the contribution of variants overlapping PIRs varies, depending on the tissue context being examined. I developed a hierarchical heuristic method to ascertain for each target gene which was the mostly likely component and cell state. Firstly for each gene I compute the gene score due to genic effects (components 1 + 2) and interactions (component 3) using all available tissue interactions for that gene. I use the ratio of gene effects score to interactions score in a similar manner to a Bayes factor to decide whether one is more likely. If gene effect is more likely (ratio >3) I iterate and compare if the gene score due to coding variants (component 1) is more likely than for promoter variants (component 2). Similarly if an interaction is more likely I compare interaction gene scores for one set of tissue(s) to another. If at any stage no branch is substantially preferred over its competitor (ratio of gene scores < 3) I return the previous set as most likely, otherwise I continue until a single cell state/set is chosen. In this way I can prioritise genes based on the overall score and label a likely mechanism for candidate causal variants (Figure 4.1).

For the four traits where a stochastic search method was employed I adapted COGS to work with multiple models by aggregating the posterior probabilities for each model with a variant overlapping the PIRs, promoters or coding variants to compute marginal posterior probabilities of inclusion.

2.8 Comparison of COGS scores to non PCHi-C methods

To assess the utility of COGS scores and whether PCHi-C data sets were adding information, I generated comparative scores using two other methods orthogonal to PCHi-C. The intent behind

this analysis was to assess the utility of interaction data, therefore, as input to all methods, I supplied PMI data sets from which I had masked coding variation and variation mapping to the MHC region on chromosome six. Firstly, I had access to topological associated domain (TAD) boundary information, supplied by Csilla Varnai, Michiel Thieke and Mikhail Spivakov for six cell types (Table 2.1).

For each TAD in each cell type, I subdivided and summed posterior probabilities for each trait (excluding the MHC region) by overlap with 0.1cM recombination blocks to obtain block TAD scores, and computed an overall TAD score such that:

$$TAD.score = 1 - \prod_{\text{blocks}} (1 - TADscore.block). \tag{2.3}$$

A TAD score was assigned to each gene mapping within the respective TAD in each tissue, and the maximum score across all eight tissues was selected

Secondly, I created gene scores based on proximity of associated variants to gene promoters. I took all *Hind*III fragments within ± 0.5 Mb of a genes baited *Hind*II fragment, and using 0.1cM recombination blocks computed proximity block scores, these were combined using a similar method detailed in equation 2.3. Finally I computed a comparison COGS score using only PCHi-C maps for tissues for which TAD boundary information were available (Table 2.1).

Tissue	TAD Coverage (Gb)
Erythroblasts	1.53
Macrophages	1.68
Megakaryocytes	1.59
Monocytes	1.48
Naive B cells	1.51
Naive $CD4^+$ T cells	1.40
Naive $CD8^+$ T cells	1.51
Neutrophils	1.27

Table 2.1: Topologically associated domain coverage across eight cell types elucidated from classical Hi-C analysis

2.9 Enrichment of COGS in disease specific differentially expressed genes

As a naive method to assess the biological relevance of the genes prioritised by COGS, and to allow quantitative comparison between the TAD, Proximity and COGS methods, I examined differentially expressed genes from Peters et al [2016] for enrichment of genes prioritised on the basis of each method. To do this I used differential expression analysis of this data set supplied by Chris Wallace. The data set consists of PEER normalised microarray expression values across 49 patients with Crohn's disease, 42 with ulcerative colitis and 43 healthy controls,

across sorted $CD4^+$ T cells, CD8+ T cells, B cells, Monocytes, and Neutrophils. Differential expression was computed with a null hypothesis that expression for a given gene was the same across all three groups within a tissue. As COGS and TAD scores are derived by combining over cell types I selected the tissue with the maximum significant differential expression values across tissue for a given gene. As the differential expression analysis concerned ulcerative colitis and Crohn's disease I used as input GWAS statistics from Anderson et al [2011] and Franke et al [2010] with coding and MHC variants removed.

2.10 Reactome Pathway Analysis

Using modified R code developed by Mikhail Spivakov, for each trait I selected all protein coding genes having an overall gene score above 0.5. I converted Ensembl gene identifiers to Entrez gene identifiers using bioMaRT [Durinck et al, 2009] and used ReactomePA [Yu and He, 2016] to compute the enrichment of genes within the Reactome pathways using an FDR cutoff of 0.05, using ClusterProfiler [Yu et al, 2012] to plot a bubble plot of significant results.

2.11 Causal variant posterior probabilities using ImmunoChip summary statistics

Dense genotyping platforms target specific genomic regions to provide higher resolution genetic maps. I wanted to see, in the context of activated and non activated $CD4^+$ T cells, what extra information could be gained by integration of PCHi-C maps with genetic data from such targeted studies. Focusing on Autoimmune traits, I downloaded ImmunoChip summary statistics for six traits from `http://www.immunobase.org` carrying out QC as previously described. The ImmunoChip is a targeted genotyping platform for dense coverage of approximately 180 regions with robust demonstration of association with one or more autoimmune traits [Cortes and Brown, 2011]. Summary statistics for ulcerative colitis, Crohn's disease and psoriasis were supplied privately by study authors. I fine mapped these traits using the PMI method previously described, but replacing 0.1cM regions with the 179 regions (median size 227Kb with an inter quartile range of between 126Kb and 392Kb) that were densely genotyped on the ImmunoChip [Onengut-Gumuscu et al, 2015].

2.12 Comparison of PMI with genotype method allowing for multiple causal variants in a region

To evaluate the effect on COGS scores of the PMI assumption of a single causal variant I used a set of marginal posterior probabilities obtained by Chris Wallace using a stochastic search method, GUESSFM, that allows for multiple causal variants within a region [Wallace

et al, 2015]. I compared four diseases(autoimmune thyroid disease, celiac disease , rheumatoid arthritis and type 1 diabetes) for which we had access for full genotyping data from ImmunoChip and summary statistics. I incorporated these into further analysis on the assumption that full genotyping data and imputation would provide more accurate posterior probabilities at the targeted regions.

Chapter 3

Results

3.1 Posterior probabilities generated by PMI and genotype imputation are comparable

To validate performance of the PMI technique, I compared the results as imputed by PMI to those as reported by [Okada et al, 2014]. There was good agreement between PMI imputed $-\log_{10} p$-values and those derived from classical imputation as reported in Okada et al [2014] ($\rho = 0.9418103$, figure 3.1).

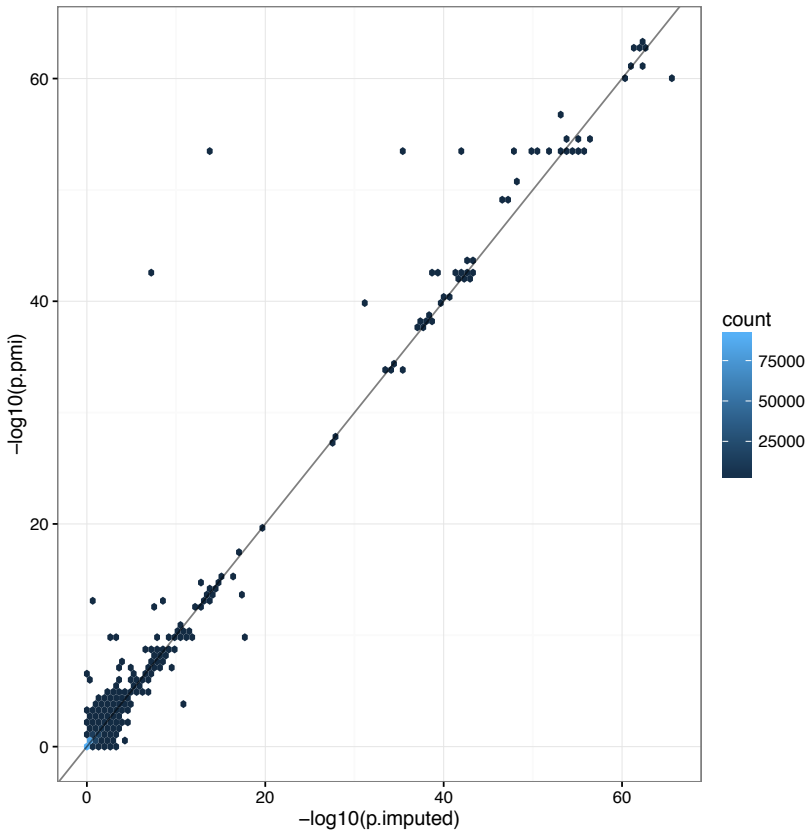


Figure 3.1: Comparison of P-values imputed by PMI versus those reported in [Okada et al, 2014] for chromosome. Axis are $-\log_{10}$ transformed p-values

3.2 Tissue specific enrichment of associated variants with PIRs across 31 traits

Enrichment of GWAS signals in tissue specific enhancers has been previously described [Maurano et al, 2012], and we expect PIRs to be enriched for regulatory regions. As such a demonstration of robust enrichment, within tissue specific PIRs, is a pre-requisite for further analysis that attempts to link causal variation to target genes in relevant tissue contexts. Using *blockshifter* with the PMI imputed summary statistics from the 31 GWAS assembled (Table 2), I found that variants associated with autoimmune disease are enriched in PIRs in lymphoid compared to myeloid tissues (Figure 3.2). In contrast, SNPs associated with erythroid traits, including mean haemoglobin concentration (MCH), mean corpuscular volume (MCV) and red blood cell count (RBC) showed a selective enrichment in erythroblasts and megakaryocytes compared to PIRs in monocytes, macrophages and neutrophils (Figure 3.2). I next examined whether I could further resolve tissue differences using *blockshifter*. I found that autoimmune traits were enriched in activated and non-activated $CD4^+$ T cells when compared to megakaryocytes and erythroblasts. This enrichment for autoimmune disease traits was stronger in activated compared to non activated $CD4^+$ T cells (Figure 3.3).

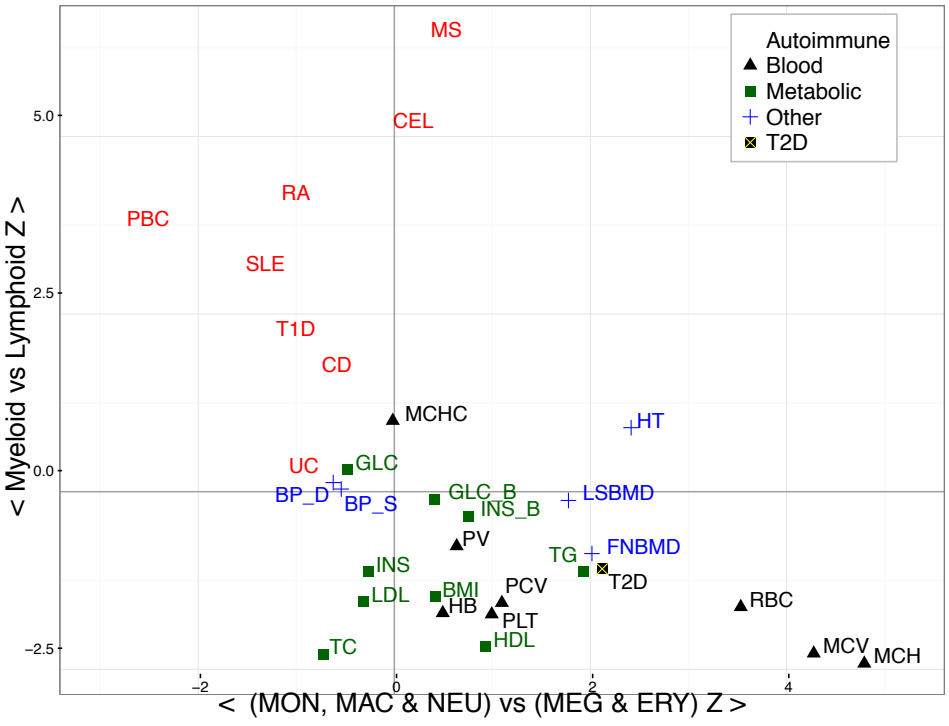


Figure 3.2: Enrichment of GWAS summary statistics at PIRs by tissue groups. Axes reflect *blockshifter* *Z*-scores for two different tissue group comparisons, firstly lymphoid versus myeloid and then within myeloid lineage (MON - Monocyte, MAC - Macrophage and NEU - Neutrophil) versus (MEG - Megakaryocyte and ERY - erythroblast). Traits are labelled and coloured by category (Appendix Table 2).

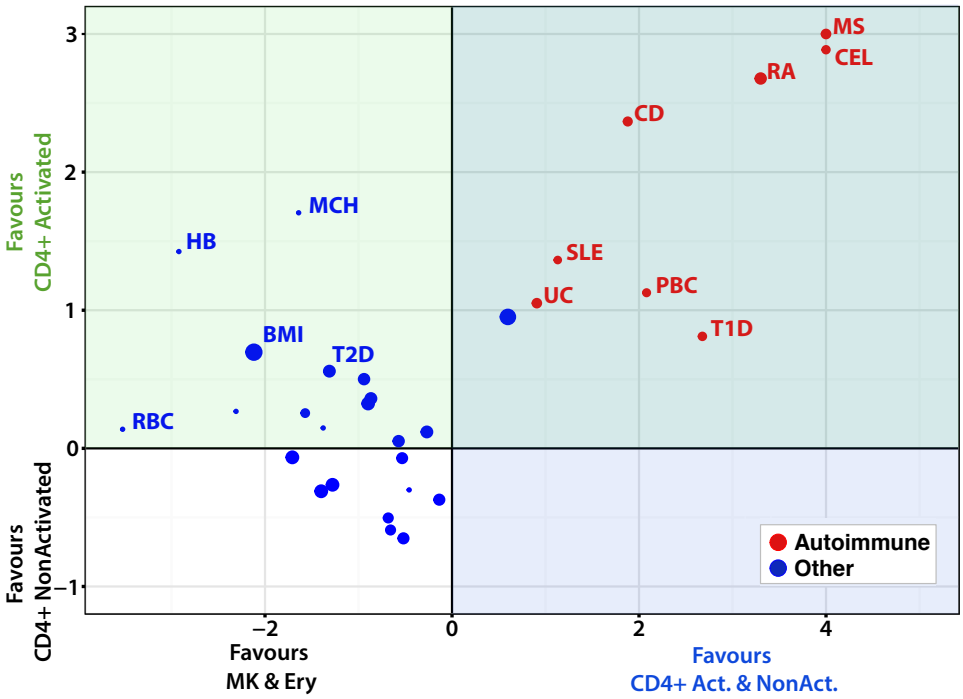


Figure 3.3: Tissue specific enrichment of autoimmune GWAS summary statistics at PIRs by tissue groups. Axes reflect *blockshifter* Z-scores for two different tissue group comparisons, firstly MK - megakaryocytes and ERY - erythroblast versus Activated/Non-activated $CD4^+$ T cells. Y-axis shows comparison of Activated versus non activated $CD4^+$ T cells. Autoimmune traits are coloured red and other traits are blue, point size reflects the log transformed sample number included in each study.

3.3 PCHi-C assisted gene prioritisation across 31 traits

Using COGS, I prioritised 2,604 unique protein coding genes with an overall COGS score greater than 0.5, across all 31 traits examined, with a median of 122 genes prioritised per trait (Figure 3.4). The mean number of protein coding genes skipped was 1.5 for genes prioritised based on promoter interactions. The prioritised genes exhibited enrichment for specific pathways in the Reactome pathway database [Fabregat et al, 2016] in a naive analysis. As expected, genes prioritised for autoimmune diseases were enriched in inflammation and immune response-related pathways, such as interleukin and T cell receptor signalling, whereas genes prioritised for platelet traits were preferentially associated with platelet production and haemostasis (Figure 3.5).

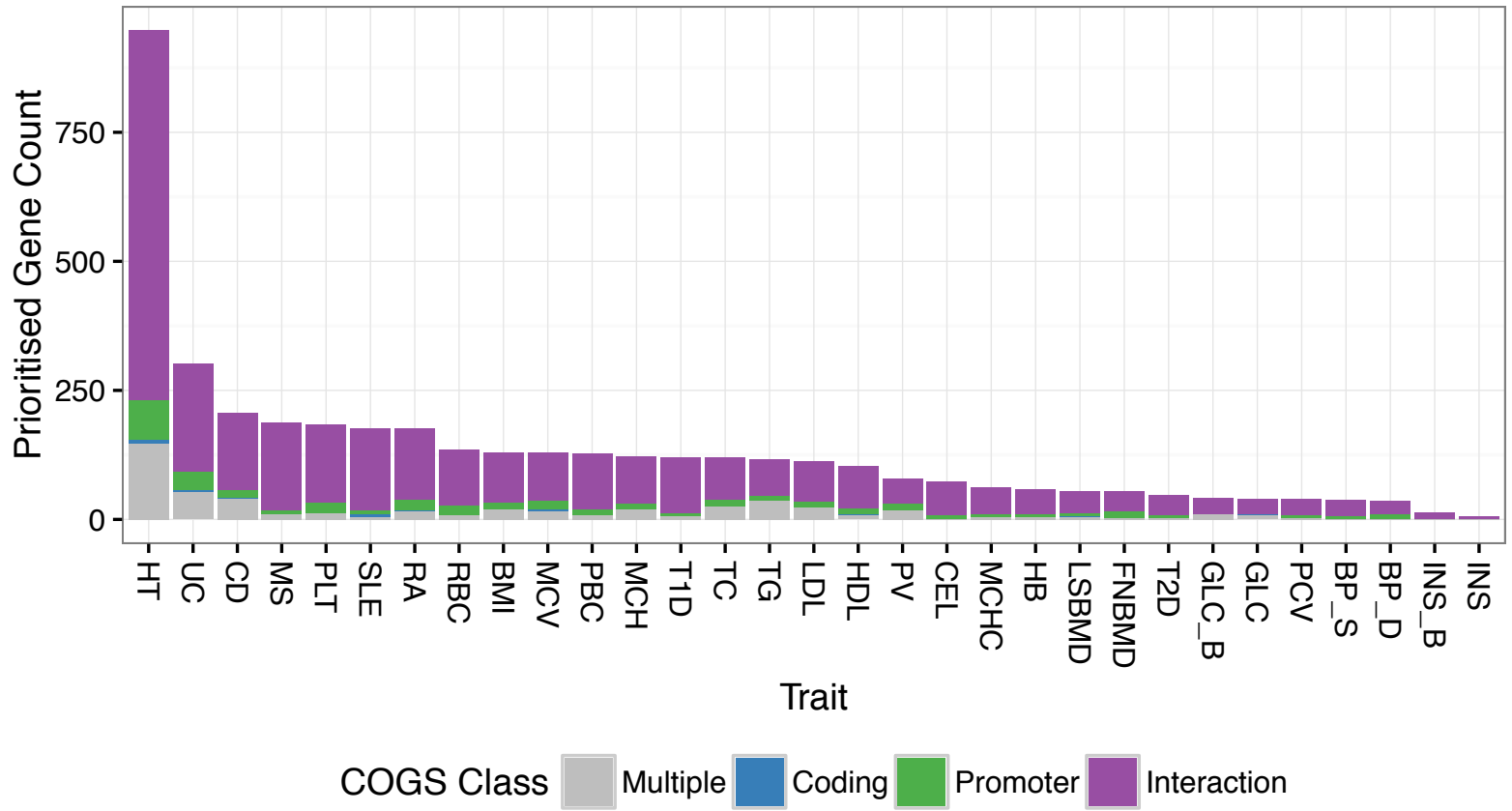


Figure 3.4: Prioritised gene counts for 31 traits. Colours indicate the proportion of genes in four broad hypotheses of causal variation mechanisms : Multiple - No specific hypothesis favoured; Coding - cSNP(s) in target gene; Promoter - cSNP(s) around TSS of target gene; Interaction - cSNP(s) in PIR of target gene.

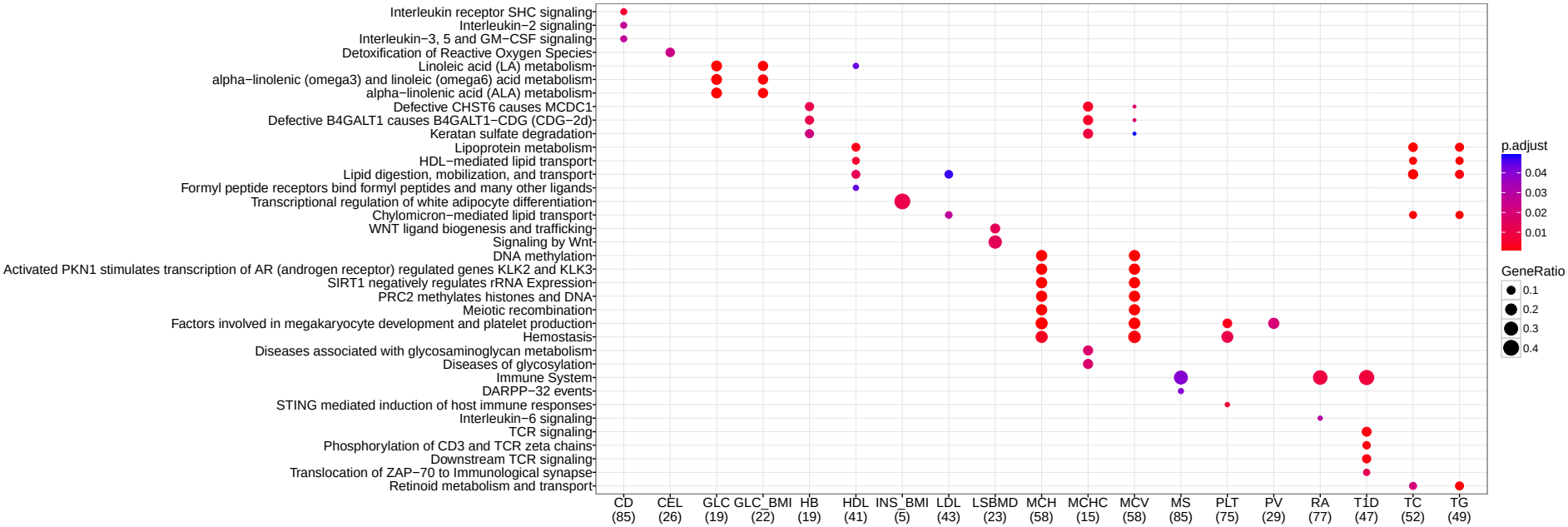


Figure 3.5: Bubble plot of traits with significant enrichment ($P_{adj} < 0.05$) in one or more reactome pathways for genes with COGS score $>$ 0.5 across 31 traits. Number in parentheses below trait labels indicate the total number of genes analysed for each trait, bubble size indicates the ratio of test genes to those in the pathway, and blue to red shading corresponds to decreasing adjusted P-value for enrichment

3.4 Prioritised genes do not overlap significantly with eQTLs

In parallel with my GWAS focused analysis Oliver Stegle and Roman Kretzhuber (RK) had been working to consider expression quantitative trail loci (eQTLs) from Fairfax et al [2012] in the context of PCHi-C. Both eQTL and GWAS are enriched in PIRs, which prompted us to consider the overlap of COGS prioritised genes in SLE [Bentham et al, 2015] and rheumatoid arthritis [Okada et al, 2014] data sets, for which full imputed summary statistics were available. I performed COGS gene prioritisation and RK looked to see if there was evidence for overlap with relevant tissue eQTLs [Fairfax et al, 2012]. Out of 456 genes that were prioritised for both traits 136 had eQTLs of which four genes (*BLK*, *RASGRP1*, *SUOX*, and *GIN1*) showed evidence for possible co-localisation in RA and two genes (*BLK* and *SLC15A4*) in SLE. Additionally the genes prioritised for RA included 5/9 candidates (*C8Orf13*, *BLK*, *TRAF1*, *FADS2* and *SYNGR1*) that were identified in a recent study [Zhu et al, 2016] that combined whole blood eQTLs with the same RA GWAS data by Mendelian randomisation. The relatively large number of GWAS prioritised genes without eQTL support agrees with previous reports of limited overlap of disease variants with eQTLs [Guo et al, 2015; Huang et al, 2015b]. One reason for this observation might be that causal variants function in a specific tissue context, which has yet to be suitably assessed by eQTL analysis. As suggested by Huang et al [2015a], this implies that regulatory annotations that might be used to augment fine mapping methods are similarly incomplete.

3.5 COGS prioritisation qualitatively performs better than TAD and distance based scores

For all eight GWAS autoimmune traits, I compared protein coding genes scores, using PCHi-C data sets (COGS), and comparative scores computed using either TAD or proximal methods. To summarise results across all diseases I selected the maximum score for a method for each gene. Using a score cut off of 0.5 I then categorised each gene as to whether it was prioritised in either a single, both or no methods (Figure 3.6). In general COGS prioritised genes sets were significantly smaller than then those from the other methods. The naive proximity based score appeared to be least specific prioritising 5,731 genes in total, however it had the greatest overlap with COGS scores implying comparatively better sensitivity. For proximity based scores 83 genes were prioritised exclusively by COGS, indicating that 12% of genes are prioritised by interactions greater than 0.5 Mb. For TAD based scores 219 genes were prioritised exclusively by COGS, indicating that approximately 30% of genes are prioritised by interactions that span TAD boundaries. I note however that examining distance metrics between gene baits and supporting PIRs that in some cases this could be due to the imprecise definition of TAD boundaries (Figure 3.7). Overall COGS appears to select genes not found by other methods and also appears to be more specific.

3.6 COGS prioritised genes are enriched for differentially expressed IBD genes

Calibrating performance metrics of the COGS method is challenging in the absence of a set of functionally validated set of scores. One possible proxy is to examine the overlap of prioritised genes with studies examining differential expression between diseased and healthy volunteers in relevant tissue types. I integrated scores from PCHi-C (COGS), proximity and TAD methods, for ulcerative colitis (UC) and Crohn's disease (CD), with differentially expressed genes (FDR 5%) from Peters et al [2016] and used Fisher's test to examine enrichment for prioritised genes (score > 0.5). I found that UC COGS genes were significantly enriched for differentially expressed genes ($P = 0.002$), for Crohn's disease there was nominal enrichment ($P = 0.04$). Conversely there was no evidence of enrichment in any data sets using proximal and TAD methods (Figure 3.9). The intersect between differentially expressed genes and COGS prioritised genes identified 67 genes (Table 4). Whilst there are many candidate genes in this list, an example on chromosome 3 for ulcerative colitis provides an illustration of the potential for hypothesis generation by integrative analysis of PCHi-C interaction maps (Figure 3.8). COGS prioritises *BCL6* which has been shown to have potent effects on Th9 cell development and IL-9 secretion both important modulators of inflammation [Bassil et al, 2014]. Putative causal variants from Anderson et al [2011] for UC reside in intron 8 of *LPP*, and previous studies have implicated a causal role for this gene, however PCHi-C results from lymphoid tissues show interactions between this region and the *BCL6* promoter (Figure 3.8). Furthermore, BLUEPRINT chromatin state maps classify this region harbouring putative causal variants as having enhancer activity across 3 out of 4 biological replicates in $CD4^+$ T cells. Expression profiles from Peters et al [2016] show that whilst *LPP* is not significantly differentially expressed between UC controls and healthy volunteers (log(Fold Change) = 0.05, $P_{adj} = 0.46$) *BCL6* is (log(Fold Change) = 0.22, $P_{adj} = 0.0018$), providing further support for it's prioritisation.

3.7 Tissue specific PCHi-C assisted gene prioritisation using dense summary statistics

The stronger enrichment of autoimmune disease GWAS traits in activated $CD4^+$ T cells lead me to examine autoimmune traits in the context of activated and non activated $CD4^+$ promoter interaction maps. I extended COGS to incorporate a simple binary decision tree model to allow tissue or functional category resolution (see Section 2.7). Across the combined GWAS and ImmunoChip autoimmune data sets I was able prioritise 602 distinct protein coding genes (Figure 3.10). To summarise the behaviour of prioritisation, I focused on a subset of 220 input autoimmune GWAS regions with genome-wide significant signals ($P < 5 \times 10^{-8}$). I prioritised at least one gene with a COGS score > 0.5 in 122 of these regions, with a median of two genes/region (inter quartile range = 1-3). The average distance from peak signal to prioritised

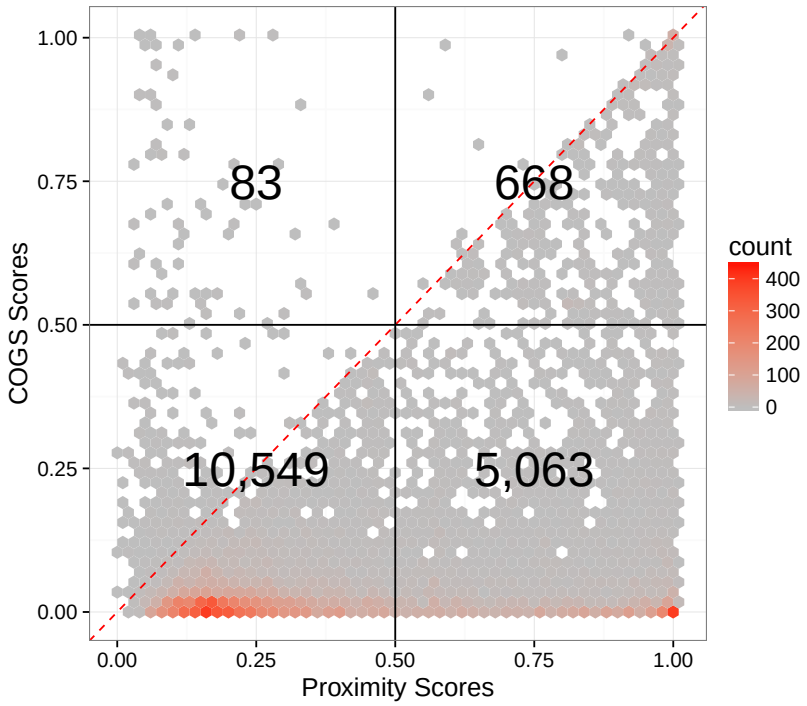
genes was 334 kb, and I observed a median of three genes 'skipped' between GWAS signal peaks and prioritised genes. Using pooled total RNA-seq expression data on the same donors (Anthony Cutler, Arcadio Rubio Garcia and Chris Wallace), I found that 457 of the prioritised genes were expressed in at least one activation state. I could relate 259 genes to GWAS significant signals of which 166 were deferentially expressed between activated and non activated states.

3.8 Allowing multiple causal variants increases number of prioritised genes

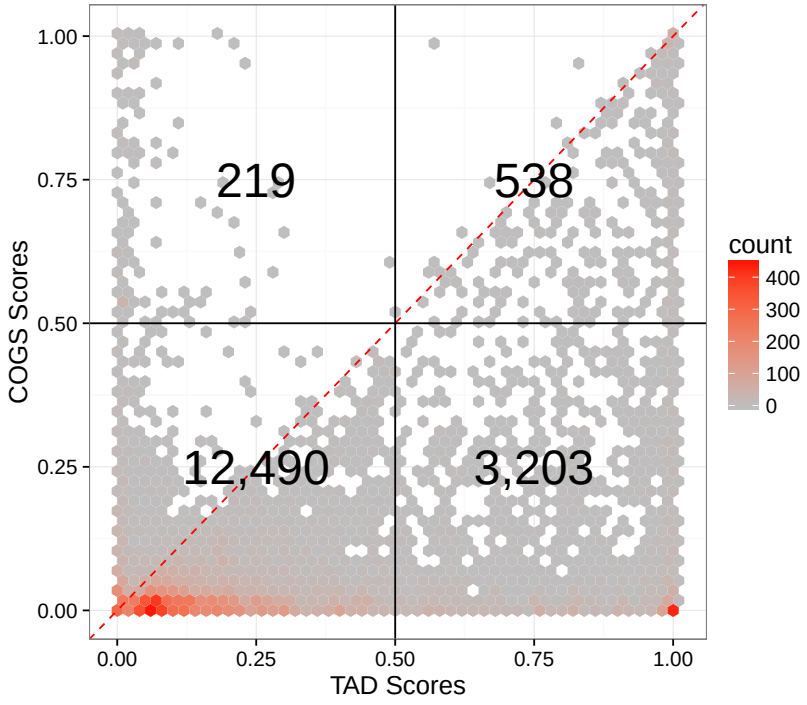
I wanted to understand the effect of assuming a single causal variant within each region on a genes COGS scores for various traits. I compared COGS scores computed from marginal posterior probabilities for dense genotype data from four autoimmune diseases (ATD, CEL, RA and T1D) , computed by GUESS FM, which allow multiple causal variants to those from PMI, which assumes a single common variant (figure 3.11). Generally there was agreement between methods, however I note that i GUESSFM prioritises more genes than PMI (Table 3.1). This increase in specificity may be down to increased resolution gained from the GUESSFM approach due to the employment of genotype based imputation and it's ability to consider different prior probabilities. However, some of this will be offset by a greater sensitivity of GUESSFM to genotyping error than the PMI based approach.

Disease	GUESSFM	PMI	Both	Total
ATD	9	0	6	38
CEL	6	5	33	114
RA	6	2	19	132
T1D	16	7	35	212

Table 3.1: Counts for protein coding genes prioritised (score > 0.5) by GUESSFM, Poor Man's Imputation (PMI) and Both methods out of Total genes with score > 0.01.



(a)



(b)

Figure 3.6: Comparison of autoimmune PCHi-C COGS scores and (a) a proximity score from assigning variants to genes within 0.5 Mb of gene promoters (b) Hi-C derived TAD scores, using seven Cell types (Erythroblasts, Macrophages, Monocytes, Naive B cells, Naive $CD4^+$ T cells, Naive $CD8^+$ T cells and Neutrophils). Counts of genes in each quadrant are shown, grey to red colour gradient indicates gene density.

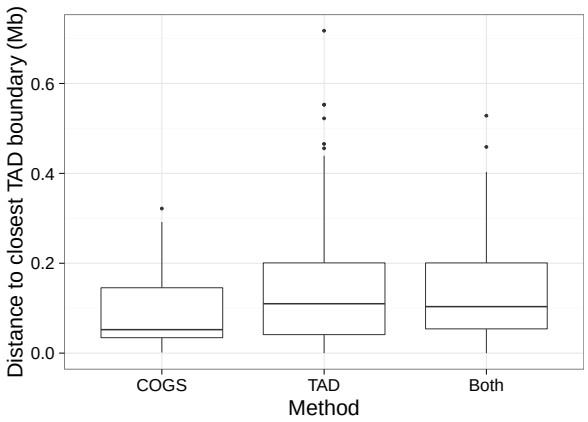


Figure 3.7: Box plot showing the distribution of distances between baits and TAD boundaries for significant (score > 0.5) genes. Both indicates that gene was significant using TAD and COGS methods

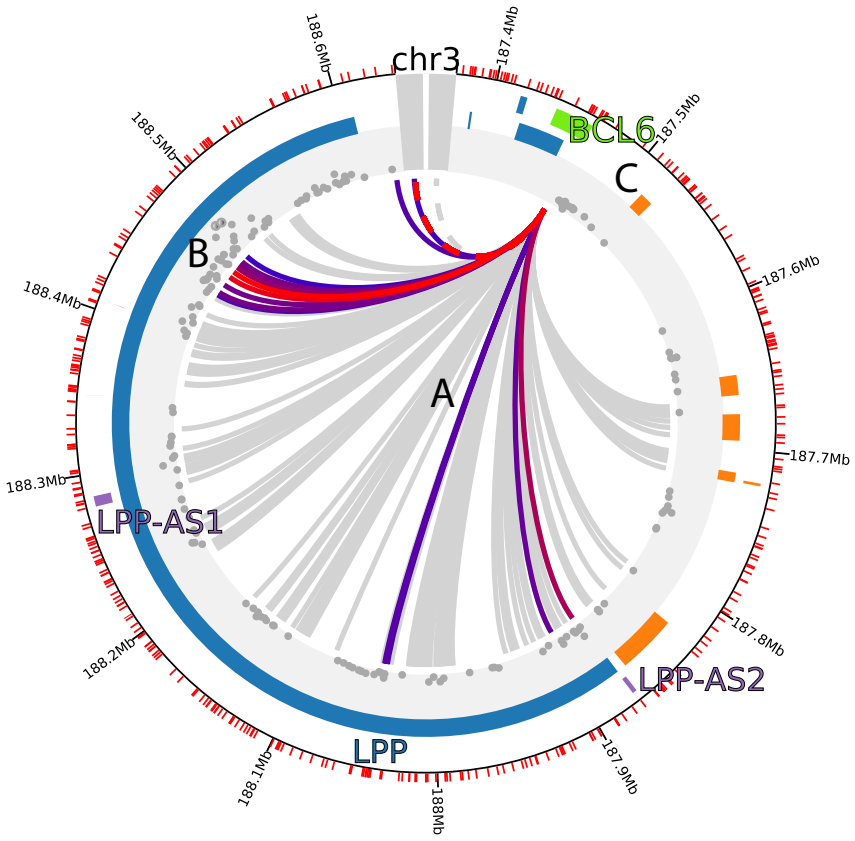


Figure 3.8: Circularised plot showing overlap between ulcerative colitis (UC) putative variants and interactions between *BCL6* gene promoter in CD4$^+$ T cells, prioritised by the COGS algorithm. A) Interactions, grey (not significant i.e. CHiCAGO score > 5 in other tissues), blue (less significant) to red(more significant) colour gradient. B) UC GWAS summary $-\log(P)$ values from Anderson et al [2011] C) Gene annotations, protein coding (blue), antisense (purple), lincRNA (orange) and *BCL6* (green). Red ticks around circumference indicate *Hind*III fragment boundaries. Figure modified from CHiCP - `http://www.chicp.org` [Schofield et al, 2016]

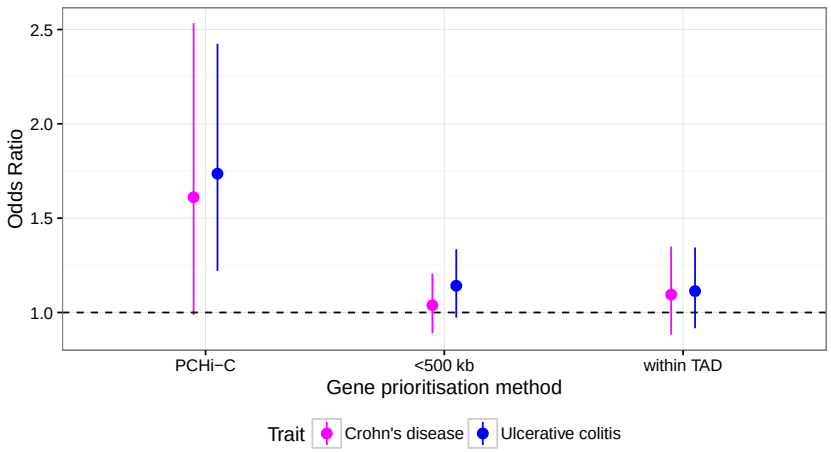


Figure 3.9: Enrichment of prioritised genes in ulcerative colitis and Crohn's differentially expressed genes from Peters et al [2016]. Methods PCHi-C (COGS), < 500 Kb (Proximity) and within TAD (topologically associated domain).

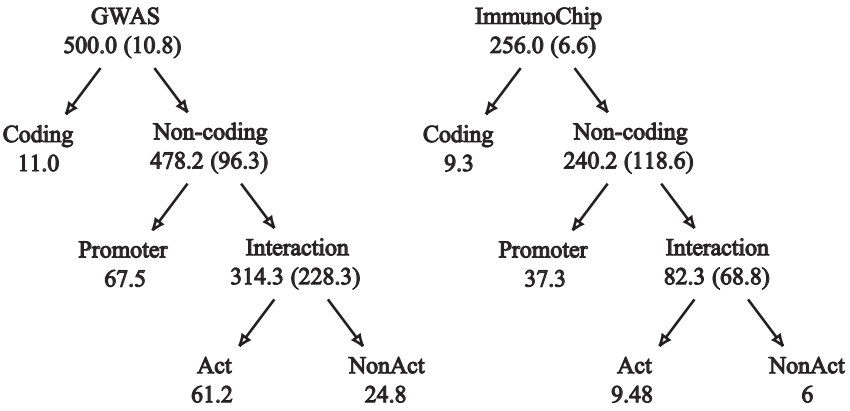


Figure 3.10: Functional gene prioritisation across 11 autoimmune diseases using genome wide (GWAS) or targeted genotyping array (ImmunoChip) data. The numbers at each node give the number of genes prioritised at that level. Where there is evidence to split into one of two non-overlapping hypotheses ($\log_{10}$ ratio of gene scores >3), the genes cascade down the tree. Where the evidence does not confidently predict which of the the two possibilities is more likely, genes are 'stuck' at the parent node (number given in brackets). When the same gene is prioritised for multiple ($n > 1$) disease, I assign a fractional count to each node, defined as the proportion of the n diseases for which the gene was prioritised at that node.

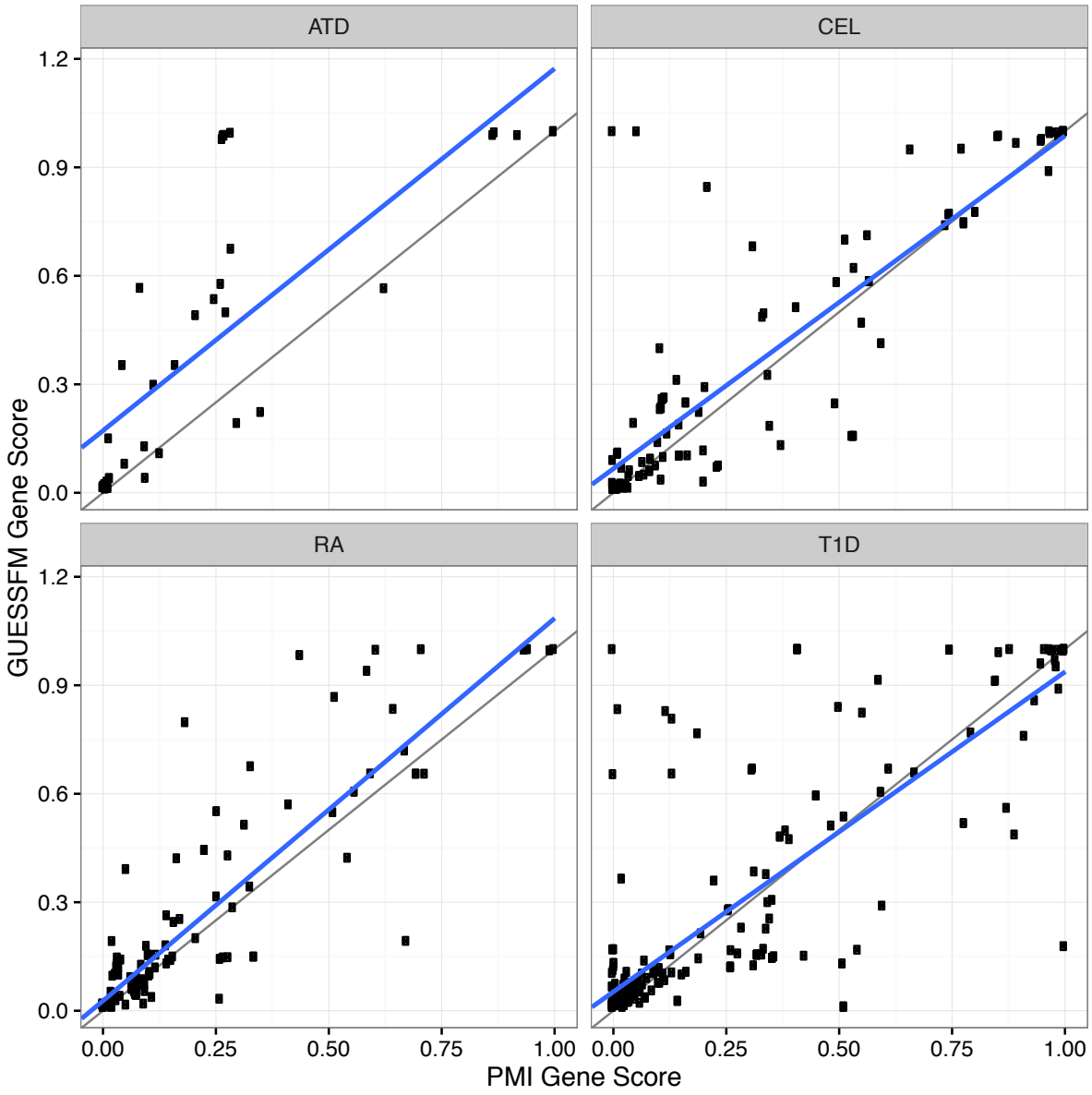


Figure 3.11: Comparison of COGS gene scores for four autoimmune traits (ATD = autoimmune thyroid disease, CEL = celiac disease, RA = rheumatoid arthritis and T1D = type 1 diabetes computed using marginal posteriors from GUESSFM and those computed using posteriors from PMI, grey line shows $y = x$, blue lines indicate the fitting of a linear model

Chapter 4

Discussion

I have presented novel methods for the integration of three dimensional chromatin data with GWAS summary stats in order to prioritise genes and target tissues. A competitive enrichment analysis, *blockshifter*, that carefully accounts for correlation, provided evidence that such a strategy was justified, implicating CD^+ T cell activation as an important tissue context, in the development of autoimmune diseases. COGS, a Bayesian framework for gene prioritisation, enabled a data driven approach to gene prioritisation, to complement current approaches. Comparisons of COGS prioritised genes with those derived from TAD and proximity based methods, show that COGS has greater sensitivity and specificity providing support for the general approach. In Section 4.1 I describe one such targeted experimental approach to validate one of the findings. However, collaboratively we have developed a reciprocal capture Hi-C platform to perform targeted validation of 949 specific PIRs identified from PCHi-C approach. Such a platform can help to overcome some technical limitations of PCHi-C. Firstly, its more targeted nature allows greater read-depth and therefore more sensitive detection of less stable or prevalent chromatin interactions. Secondly, it is not confined to promoter interactions and will allow interrogation of the inter-PIR chromatin interaction space. Indeed preliminary analysis has indicated the presence of strong interactions between the PIR detailed in section 4.1 and another PIR in a separate but proximal autoimmune region containing PRKCQ (COGS score=0.95) [Lowe et al, 2007].

A fundamental tenet of science is reproducibility, and whilst technical considerations inject variability at the level of data generation, considerable efforts should be made to develop transparent analytical methods that do not add to this. The software that I have developed over the course of this project is available `https://github.com/ollyburren/CHIGP`, however, due to the organic nature of its development its overall usability and architecture is lacking. To foster community reuse I intend to rewrite the software into a formal R package, complete with regression testing and more extensive documentation. I hope that this initial time investment will not only benefit the community but also provide a solid foundation for the future work that I set out below.

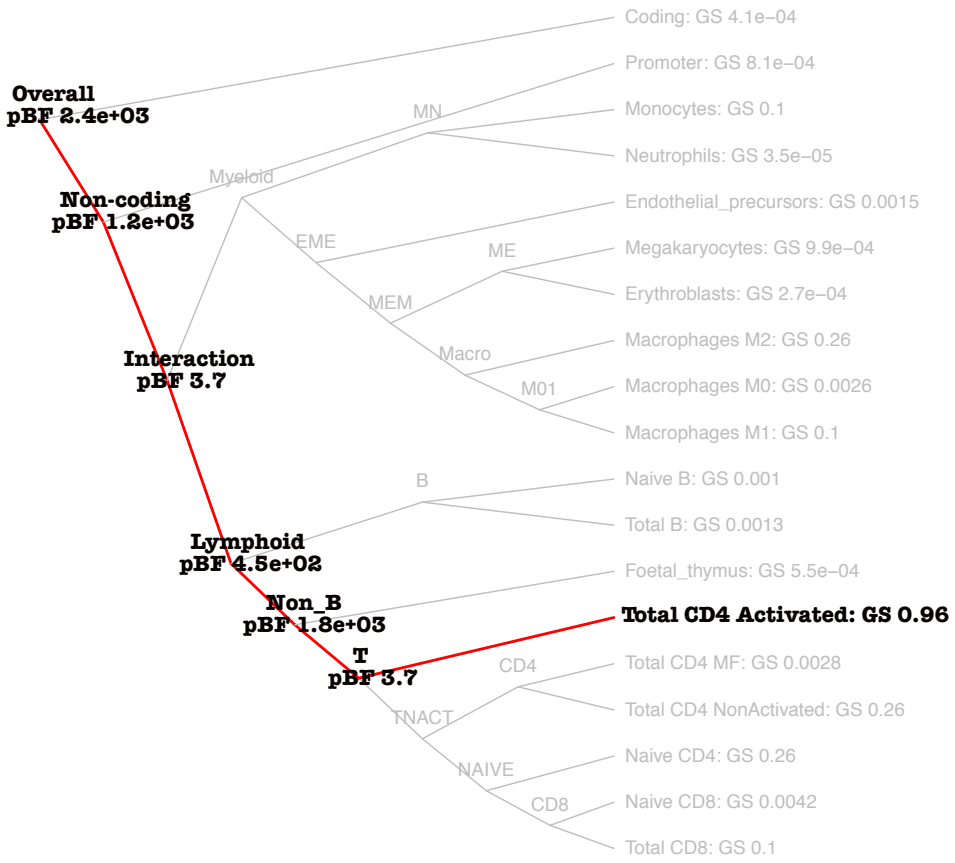


Figure 4.1: COGS decision dendrogram using rheumatoid arthritis data from Okada et al [2014] and PCHi-C maps from 17 haematopoetic cell types for the *AHR* gene. The red edges denote the path taken by COGS through the binary decision tree. Each node is labelled with pseudo Bayes factors(pBF). Terminal nodes are labelled with the tissue/annotation specific gene score(GS)

4.1 Functional validation of COGS prioritised gene *IL2RA*

I have presented *in silico* validation for genes prioritised by integrating PCHi-C and GWAS data sets, by the application of pathway and gene set enrichment analysis, however, empirical evidence is required to fully validate the strategy. Thus, in collaboration with Daniel Rainbow, Tony Cutler, Arcadio Rubio Garcia, Chris Wallace and Linda Wicker we sought to understand how we might functionally validate one such prioritised gene and whether this might lead to a greater understanding of how causal variation might modulate autoimmune disease susceptibility. COGS analysis prioritised, in multiple diseases (autoimmune thyroid disease, Crohn's disease, multiple sclerosis, rheumatoid arthritis, type 1 diabetes, and ulcerative colitis) *IL2RA* which encodes the CD25 protein, a component of the IL-2 receptor that is essential for high-affinity binding of IL-2, regulatory T cell survival and T effector cell differentiation and function [Liao et al, 2013]. I found this prioritisation to be driven by an interaction between the *IL2RA* promoter and a PIR in exon 1 known to harbour a set of type 1 diabetes putative causal SNPs (Figure 4.3) identified in a previous fine mapping study [Wallace et al, 2015].

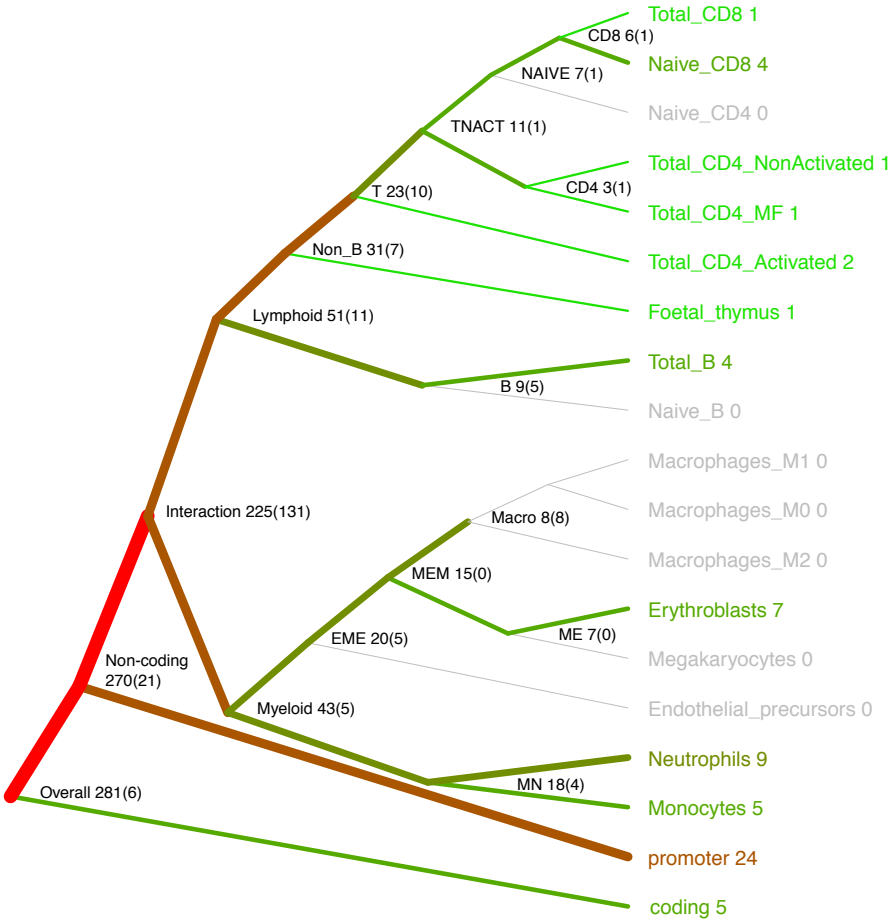


Figure 4.2: COGS decision dendrogram for all prioritised protein coding genes (GS > 0.5) using rheumatoid arthritis data from Okada et al [2014] and PCHi-C maps from 17 haematopoetic cell types. Edges are coloured based on number of genes flowing between connected nodes. Nodes are marked with the total number of genes at a node and in brackets the number of genes that are assigned to that node.

This set of SNPs (Figure 4.3) is in high LD ($r^2 > 0.8$) with rs12722495 which has been shown to affect the surface expression CD25 in memory T cells [Dendrou et al, 2009]. Using a targeted RNA-sequencing approach, and software I helped to develop previously [Rainbow et al, 2015], Daniel Rainbow measured the relative expression of the two alleles at one of these SNPs, rs61839660, in intronic cDNA, from four individuals heterozygous at rs61839660 and homozygous across most other associated SNP groups, in a four-hour activation time-course of $CD4^+$ T cells. We observed allelic imbalance in non activated $CD4^+$ T cells however on activation this was lost suggesting a context specific effect within this locus. To ensure this reflected imbalance in transcription of *IL2RA*, and not simply enhancer RNAs in intron 1, further validation was performed using rs12244380 found in the 3' UTR of *IL2RA* (Figure 4.4).

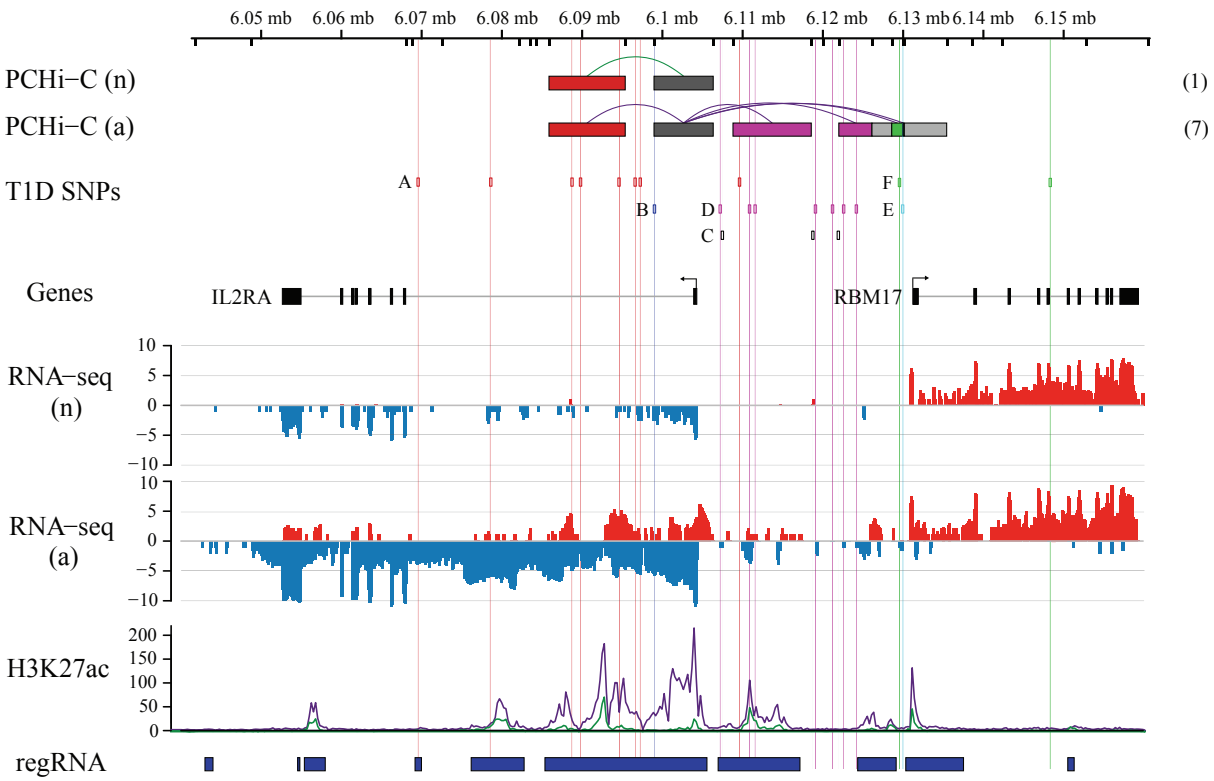


Figure 4.3: PCHi-C interactions link the *IL2RA* promoter to autoimmune disease associated genetic variation which leads to expression differences in *IL2RA* mRNA. (n) and (a) refer to non activated and activated $CD4^+$ T cells respectively. Numbers in parentheses on the right hand side represent number of interactions observed. Labels 'A' to 'F' indicate sets of SNPs likely to contain a single causal disease variant. Figure prepared by Tony Cutler, Arcadio Rubio Garcia and Chris Wallace.

4.2 Limitations

There are a number of limitations with the current approaches that I have developed. Firstly the fine mapping approach employed assumes for a given region a simplified model of a single underlying causal variant. I demonstrate that this does have an effect on gene prioritisation but further work is required to whether the genes prioritised using GUESSFM are in fact better candidates for functional follow up.

Another limitation is the thresholded approach to CHiCAGO scores that are used to call interactions. All of the methods developed so far use a threshold score of 5, so that an interaction with a score of 4.99 will be omitted. Future approaches might investigate methods for incorporating promoter interaction scores themselves in both *blockshifter* and COGS methods. One approach might be to look at techniques that utilise CHiCAGO scores across multiple tissues for a given interaction to adjust local FDR.

COGS makes the assumption that coding variation affects the gene within which it is located, however studies in model organisms [Lawrie et al, 2013] and in humans [Stergachis et al, 2013]

indicate that coding variation can fulfil a dual role in the regulation of genes. Whilst coding putative causal variants are relatively rare it remains to be seen what effect this assumption has on overall COGS scores, I hope to investigate this further using the approach detailed in Section 5.3.

Another significant challenge with these analysis is that resolution is limited by *Hind*III restriction fragment length. This manifests in two main ways. Firstly there is a blind spot for observing shorter range interactions that involve *Hind*III fragments and adjoining baited interactions. I have attempted to capture these as 'promoter' regions (Figure 2.2), however integration with functional annotation might provide further resolution and identify functional hypothesis for mechanisms for further study.

A second more pernicious issue is that many baited fragments are promiscuous in that they contain promoter regions for more than one gene. If we consider just protein coding genes then of 16,608 baits, 3,009 (18%) contain multiple promoters from different genes. If I include all transcriptional start site annotations in Ensembl (Version 75) then this rises to 6,703 (40%). In reality, even this is a conservative estimate due to the incomplete annotation of the non-coding genome. For these promiscuous baits it is impossible to resolve which promoter or promoters are involved in the chromatin looping, using the current PCHI-C data alone. One approach is to use gene expression patterns to provide a filter for target genes identified by specific chromatin interactions, which I discuss in section 5.1. Furthermore, my analysis to date has concentrated on protein coding genes as these have the most mature and complete annotation, however, recent work has suggested a role for non coding genes in the modulation of autoimmune disease susceptibility [Castellanos-Rubio et al, 2016]. The PCHi-C platform used does provide some coverage of the non protein coding genome, however this is not exhaustive by design and due to the overlapping nature of the coding and non-coding genome, the issue of promiscuous baits is exacerbated.

Excluding single cell implementations, all genomic technologies give an average of the molecular events across the (sometimes mixed) population of cells being assayed. In the case of immune subsets this is particularly relevant as broad categories, such as $CD4^+$ T cells will be heterogeneous containing further subdivisions that may or may not be relevant for disease biology. It is important to bear this in mind when considering PCHi-C maps as without single cell profiling it is impossible to resolve whether interactions are common across the assayed tissue type or are specific to an underlying and as yet unsorted subset.

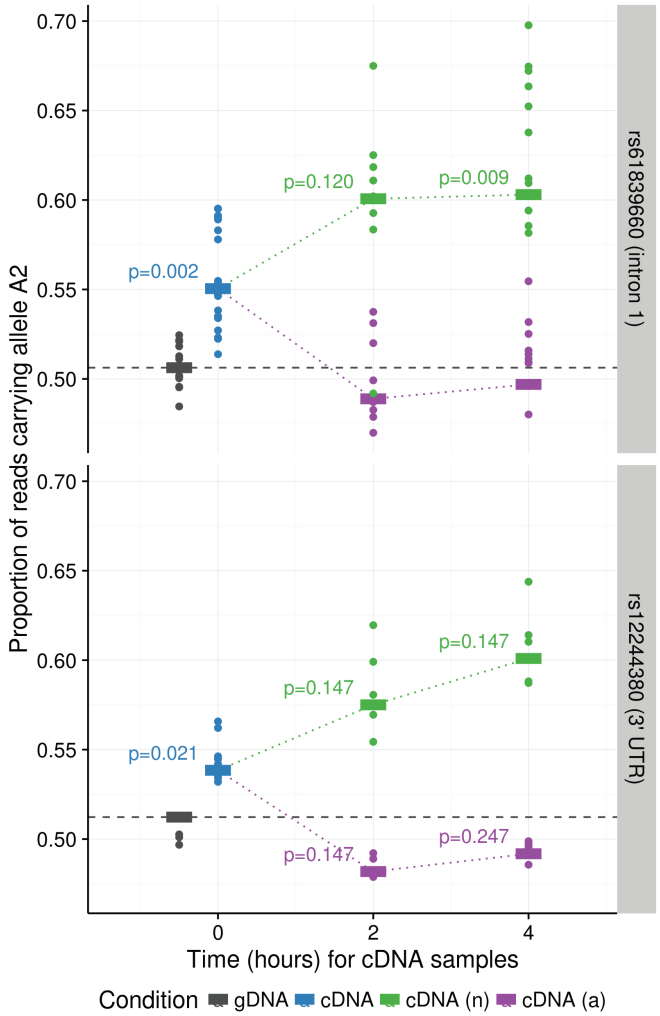


Figure 4.4: *IL2RA* allele specific expression in $CD4^+$ T cells for two putative causal variants identified by fine mapping and PCHi-C integration. Bar colour encodes DNA collected under different conditions. Black indicates genomic DNA, we expect this to be equally shared between alleles. Blue indicates cDNA collected a time zero. Green and purple bars indicate cDNA allele read counts collected under non activated and activated conditions. Figure prepared by Chris Wallace.

Chapter 5

Future Work

5.1 Creation of a catalogue of putative causal variants and genes across 17 cell types and 31 traits

Enrichment results from *blockshifter* coupled with existing results in the literature lead to a focus on understanding the tissue specificity of PCHi-C interactions within activated and non-activated $CD4^+$ T cells. The next step is to generate tissue specific COGS scores for all GWAS data sets across all 17 cell types. By way of example it should be possible to generate decision trees for all prioritised genes (e.g. Figure 4.1), enabling a deeper understanding of the distribution of tissue contexts on a disease wide scale (Figure 4.2). Further analysis will delve into the details of why a particular gene is targeted, in order to select candidates that are suitable for functional followup. Criteria might include, posterior probability concentrated on a single SNP, with an interaction observed in a specific cellular context for which relevant gene expression is observed (e.g *BCL6*). Such analysis would include further investigation of methods for more formal integration of gene expression and TAD data by, for example, taking a more tissue specific approach using the [Peters et al, 2016] data set. I would bolster this analysis by incorporating new data from the targeted validation platform, which might shed light on relevant PIR to PIR interactions which cannot be captured in the primary data set.

5.2 Genomic annotation assisted fine mapping

Increasingly high quality genomic data sets on primary human tissues are becoming available. I have shown that both non activated and activated $CD4^+$ T cells are of particular interest and have access to Chip-Seq and total RNA-Seq for these tissues. Future work will explore whether these annotations can be used as input to integrative methods such as *fgwas* to increase fine mapping resolution. I will start with autoimmune data sets for which I have access to GWAS summary statistics and use *fgwas* to integrate functional annotations, concentrating on in house ChIP-Seq and ATAC-Seq data sets available for $CD4^+$ T cells. A concern with *fgwas* is that it relies on extensive cross validation to overcome it's reliance on a single training set, to overcome this I will use the enrichment parameters generated from application to GWAS and use these to

compute variable prior probabilities for dense fine mapping information using equations 1.3 and 1.3. These prior probabilities can be used to compute 'annotation aware' posterior probabilities for a variant to be causal. I will use these as input into COGS to understand if and how this alters the genes prioritised. Depending on the outcome of this it might be worth examining RiVIERA [Li and Kellis, 2016] that can make use the covariance across related traits.

5.3 Data driven discovery of appropriate COGS score thresholds

Understanding how fine scale genetic architecture affects COGS scores is key to their interpretation, for example, ascertaining a suitable threshold for gene prioritisation. One approach requires curation of a number of non coding, causal variants with convincing functional validation for a target gene and tissue context. Unfortunately such examples are rare and so investigation of the effect of parameters such as effect size and minor allele frequency on COGS score is limited. One possible solution is to accurately simulate GWAS based on predefined parameters and causal variants and to use this as the input to assess COGS scores at specific intervals or genome wide. Using software developed by Mary Fortune that allows the accurate simulation of GWAS statistics for a given interval I will generate summary statistics to allow me to understand how various parameters affect the specificity and sensitivity of COGS and any extensions I develop.

5.4 Using COGS scores for gene set enrichment analysis

A future aim of this project is to investigate whether we can use information gained on individual genes to highlight relevant biological pathways in a tissue context specific manner. I have used the Reactome database to validate genes prioritised by COGS score, however as discussed in the introduction such analysis are imperfect. Future work will investigate two approaches using tissue specific PCHi-C, in the context of gene set enrichment analysis (GSEA). In the first approach I will extend the naive approach used to analyse Peters et al [2016] data set to use tissue specific COGS scores, a concern with this approach is the spatial correlation introduced by COGS scores and whether GWAS simulation methods I will investigate in section 5.3 can be used to account for this. The second approach will attempt to adapt the multivariate normal SNP permutation method implemented in VSEAMS [Burren et al, 2014], to use PCHi-C maps in order to assign SNPs to genes. One challenge will be to allow for the LD based SNP pruning required for computational efficiency without undermining the resolution of PCHi-C maps. As both methods will examine the problem through the prism of tissue context specificity, I will need to consider effective normalisation methods in order to accurately assess the results. One approach is to focus on autoimmune traits and use the set of 23 non autoimmune traits to assess the relative enrichment between both sets. I will assess the output of this against an extension

of the TAD method described in section 2.8. Depending on the outcome of this I can extend by using the GWAS simulation method 5.3 in order to generate sets of GWAS statistics with plausible parameters to assess enrichment. This will allow the assessment of traits without the need to normalise by trait category.

Acknowledgements

Thanks to: Chris Wallace, John Todd and Mikhail Spivakov for supervision; Tony Cutler and Linda Wicker for assistance in interpreting immunological relevance of results; Ellen Schofield, Prem Acuthan and Tim Carver for ImmunoBase and CHiCP support and development; Mary Fortune for statistical advice; Other members of the DIL, Wallace, Spivakov, Fraser, Ouwehand and Stegle Labs for data access and analytical support.

Table 1: Summary of PCHi-C datasets used in this study. Adapted from Javierre et al. (Under review)

Cell type	Acronym	Biological replicates	Unique captured read pairs	Detected interactions
Megakaryocytes	MK	4	653,848,788	150,203
Erythroblasts	Ery	3	588,786,672	144,771
Neutrophils	Neu	3	736,055,569	131,609
Monocytes	Mon	3	572,357,387	151,389
Macrophages M0	Mϕ0	3	668,675,248	163,791
Macrophages M1	Mϕ1	3	497,683,496	163,399
Macrophages M2	Mϕ2	3	523,561,551	173,449
Endothelial Precursors	EndP	3	420,536,621	141,382
Naive B cells	nB	3	629,928,642	171,439
Total B cells	tB	3	702,533,922	183,119
Fetal Thymus	FetT	3	776,491,344	145,577
Naive CD4+ T cells	nCD4	4	844,697,853	192,048
Total CD4+ T cells	tCD4	3	836,974,777	166,668
Non-Activated Total CD4+ T cells	naCD4	3	721,030,702	177,371
Activated Total CD4+ T cells	aCD4	3	749,720,649	188,714
Naive CD8+ T cells	nCD8	3	747,834,572	187,399
Total CD8+ T cells	tCD8	3	628,771,947	183,964
Total			11,299,489,740	708,007[1]

[1]Unique interactions captured in at least one cell type

Table 2: Summary of GWAS summary statistics used in this study

Trait	Acronym	Cases	Controls	Study Type	PMID	First Author(Date)	# SNPs	Source
Platelet volume	PV	18600		QUANT	22139419	Gieger(2011)	2231438	N Soranzo personal communication
Platelet count	PLT	48666		QUANT	22139419	Gieger(2011)	2206665	N Soranzo personal communication
Mean corpuscular volume	MCV	4627		QUANT	19820697	Soranzo(2009)	2156669	N Soranzo personal communication
Packed cell volume	PCV	4627		QUANT	19820697	Soranzo(2009)	2009357	N Soranzo personal communication
Red blood cell count	RBC	4627		QUANT	19820697	Soranzo(2009)	2091590	N Soranzo personal communication
Haemoglobin	HB	4627		QUANT	19820697	Soranzo(2009)	1640923	N Soranzo personal communication
Mean corpuscular haemoglobin	MCH	4627		QUANT	19820697	Soranzo(2009)	1904974	N Soranzo personal communication
Mean corpuscular haemoglobin concentration	MCHC	4627		QUANT	19820697	Soranzo(2009)	2070334	N Soranzo personal communication
Ulcerative colitis Anderson	UC	6687	19718	CC	21297633	Anderson(2011)	1399283	`http://www.immunobase.org`
Multiple sclerosis IMSGC	MS	9772	17376	CC	21833088	IMSGC(2011)	463628	`http://www.immunobase.org`
Type 1 diabetes Barrett	T1D	8000	8000	CC	19430480	Barrett(2009)	789849	`http://www.immunobase.org`

Continued on next page

Trait	Acronym	Cases	Controls	Study Type	PMID	First Author(Date)	# SNPs	Source
Celiac disease DuBois	CEL	4533	10750	CC	20190752	Dubois(2010)	509768	http://www.immunobase.org
Crohn's disease immunobase	CD	6333	15056	CC	21102463	Franke(2010)	950208	http://www.immunobase.org
Rheumatoid arthritis Okada	RA	14361	43923	CC	24390342	Okada(2014)	8513749	http://plaza.umin.ac.jp/~okada/datasource/files/GWASMetaResults/RA_GWASmeta_European_v2.txt.gz
Type 2 diabetes	T2D	12171	56860	CC	22885922	Morris(2012)	2075585	http://diagram-consortium.org/downloads.html
Height	HT	253288		QUANT	25282103	Wood(2014)	1927160	https://www.broadinstitute.org/collaboration/giant/images/0/01/GIANT_HEIGHT_Wood_et_al_2014_publicrelease_HapMapCeuFreq.txt.gz

Continued on next page

Trait	Acronym	Cases	Controls	Study Type	PMID	First Author(Date)	# SNPs	Source
Tryglycerides	TG	96598		QUANT	20686565	Teslovich(2010)	2304026	http://csg.sph.umich.edu/abecasis/public/lipids2010/TG2010.zip
High density lipoprotein	HDL	99900		QUANT	20686565	Teslovich(2010)	2322449	http://csg.sph.umich.edu/abecasis/public/lipids2010/HDL2010.zip
Low density lipoprotein	LDL	95454		QUANT	20686565	Teslovich(2010)	2298548	http://csg.sph.umich.edu/abecasis/public/lipids2010/LDL2010.zip
Total Cholesterol	TC	100184		QUANT	20686565	Teslovich(2010)	2323152	http://csg.sph.umich.edu/abecasis/public/lipids2010/TC2010.zip
Glucose sensitivity BMI adjusted	GLC_B	58074		QUANT	22581228	Manning(2012)	2622994	ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Manning_et_al_FastingGlucose_MainEffect.txt.gz

Continued on next page

Trait	Acronym	Cases	Controls	Study Type	PMID	First Author(Date)	# SNPs	Source
Glucose sensitivity	GLC	58074		QUANT	22581228	Manning(2012)	2622996	ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Manning_et_al_FastingGlucose_MainEffect.txt.gz
Insulin sensitivity BMI adjusted	INS_B	51750		QUANT	22581228	Manning(2012)	2621974	ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Manning_et_al_lnFastingInsulin_MainEffect.txt.gz
Insulin sensitivity	INS	51750		QUANT	22581228	Manning(2012)	2621977	ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Manning_et_al_lnFastingInsulin_MainEffect.txt.gz
Femoral neck bone mineral density	FNBMD	32961		QUANT	22504420	Estrada(2012)	2473840	http://www.gefos.org/sites/default/files/GEFOS2_FNBMD_POOLED_GC.txt.gz

Continued on next page

Trait	Acronym	Cases	Controls	Study Type	PMID	First Author(Date)	# SNPs	Source
Lumbar spine bone mineral density	LSBMD	32961		QUANT	22504420	Estrada(2012)	2463611	http://www.gefos.org/sites/default/files/GEFOS2_LSBMD_POOLED_GC.txt.gz
Diastolic blood pressure	BP_D	69395		QUANT	21909115	ICBP(2011)	2460847	http://www.georgehretlab.org/ICBP-summary-Nature.csv.gz
Systolic blood pressure	BP_S	69395		QUANT	21909115	ICBP(2011)	2460847	http://www.georgehretlab.org/ICBP-summary-Nature.csv.gz
Body Mass Index	BMI	322200		QUANT	25673413	Locke(2015)	1984096	https://www.broadinstitute.org/collaboration/giant/images/1/15/SNP_gwas_mc_merge_nogc.tbl.uniq.gz
Systemic Lupus Erythrematosis	SLE	4036	6959	CC	26502338	Bentham(2015)	7734064	http://www.immunobase.org
Primary Billiary Cirrhosis	PBC	2764	10475	CC	26394269	Cordell(2015)	1134133	http://www.immunobase.org

Table 3: Summary of ImmunoChip summary statistics used in this study

Trait	Acronym	Cases	Controls	Study Type	PMID	First Author(Date)	Source
Narcolepsy	NAR	1866	10421	CC	23459209	Faraco(2013)	http://www.immunobase.org
Juvenile Idiopathic Arthritis	JIA	2816	13056	CC	23603761	Hinks(2013)	http://www.immunobase.org
Autoimmune Thyroid Disease	ATD	2733	9364	CC	22922229	Cooper(2012)	http://www.immunobase.org
Multiple Sclerosis	MS	14498	24091	CC	24076602	IMSGC(2013)	http://www.immunobase.org
Primary Billiary Cirrhosis	PBC	2861	8514	CC	22961000	Liu(2012)	http://www.immunobase.org
Crohn's Disease	CRO	14763	15977	CC	23128233	Jostins(2012)	J Barrett personal communication
Rheumatoid Arthritis	RA	11475	15870	CC	23143596	Eyre(2012)	http://www.immunobase.org
Type 1 Diabetes	T1D	8000	12272	CC	25751624	Onengut-Gumuscu(2015)	http://www.immunobase.org
Ulcerative Colitis	UC	10920	15977	CC	23128233	Jostins(2012)	J Barrett personal communication
Psoriasis	PS	10588	22808	CC	23143594	Tsoi(2012)	http://www.immunobase.org
Celiac Disease	CEL	12041	12228	CC	22057235	Trynka(2011)	http://www.immunobase.org

Table 4: Summary of differentially expressed genes from Peters et al [2016] with COGS gene scores > 0.5 for ulcerative colitis (UC) and Crohn's Disease (CD)

Gene Name	COGS score	DE P_{adj}	Disease
FAIM3	1.000	0.000	UC
COX4I1	1.000	0.036	UC
RPS24	1.000	0.011	CD
IKZF1	0.999	0.001	CD
ACSS1	0.999	0.017	UC
CHD1	0.988	0.032	CD
CD274	0.987	0.002	CD
ROPN1L	0.944	0.042	UC
ADO	0.941	0.006	CD
TFAM	0.939	0.001	CD
ETS1	0.934	0.001	UC
MIDN	0.927	0.011	CD
SBNO2	0.926	0.000	CD
IPMK	0.922	0.028	CD
RQCD1	0.914	0.038	UC
STK32B	0.894	0.022	UC
CTDSP1	0.877	0.008	UC
ADAM10	0.854	0.006	UC
MYC	0.836	0.034	CD
FCGR2A	0.836	0.021	UC
FCRLA	0.835	0.018	UC
SGMS1	0.835	0.000	UC
CD244	0.823	0.003	CD
PIM3	0.822	0.001	UC
BCL6	0.811	0.001	UC
RTP2	0.800	0.043	UC
RASGRP1	0.795	0.005	CD
IKZF3	0.793	0.000	UC
LYRM7	0.771	0.001	UC
DGKD	0.758	0.004	CD
CHRNE	0.737	0.041	UC
ARRB2	0.717	0.001	UC
PIP4K2C	0.686	0.009	UC

Continued on next page

Gene Name	COGS score	DE P_{adj}	Disease
ADAM9	0.684	0.000	UC
GATA3	0.679	0.031	UC
PAPD7	0.659	0.007	UC
MFF	0.658	0.020	UC
IGF2	0.649	0.049	UC
STK36	0.634	0.013	UC
GPX4	0.631	0.000	CD
BEST1	0.627	0.001	CD
ATG4D	0.626	0.017	CD
AGAP2	0.621	0.017	UC
SMAD3	0.612	0.013	CD
MRPL4	0.599	0.017	CD
MARCH9	0.594	0.003	UC
MAN2A2	0.591	0.038	CD
GALC	0.589	0.001	CD
CDK4	0.588	0.014	UC
MLC1	0.575	0.031	CD
DSE	0.573	0.000	UC
BCL2	0.565	0.015	UC
FBL	0.561	0.001	UC
C9orf37	0.543	0.001	UC
DPP7	0.542	0.047	UC
SSNA1	0.542	0.000	UC
PHPT1	0.542	0.007	UC
CCDC183	0.542	0.012	UC
ABCA2	0.542	0.021	UC
SCAMP3	0.529	0.020	CD
TTC1	0.523	0.045	UC
ZBTB49	0.520	0.047	UC
CD55	0.517	0.002	UC
FYB	0.508	0.008	CD
ATF6	0.504	0.003	UC

Bibliography

Albert FW and Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**: 197–212.

Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, Lee JC, Goyette P, Imielinski M, Latiano A, et al. 2011. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**: 246–252.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* **25**: 25–29.

Bassil R, Orent W, Olah M, Kurdi AT, Frangieh M, Buttrick T, Khoury SJ, and Elyaman W. 2014. Bcl6 controls th9 cell development by repressing il9 transcription. *J Immunol* **193**: 198–207.

Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tombleson P, Behrens TW, Martín J, Fairfax BP, Knight JC, Chen L, et al. 2015. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* .

Burren OS, Guo H, and Wallace C. 2014. VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics* **30**: 3342–3348.

Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, Zerbino D, Schoenfelder S, Javierre BM, Osborne C, et al. 2016. Chicago: robust detection of dna looping interactions in capture hi-c data. *Genome Biol* **17**: 127.

Castellanos-Rubio A, Fernandez-Jimenez N, Kratchmarov R, Luo X, Bhagat G, Green PHR, Schneider R, Kiledjian M, Bilbao JR, and Ghosh S. 2016. A long noncoding rna associated with susceptibility to celiac disease. *Science* **352**: 91–95.

Claussnitzer M, Dankel SN, Kim KH, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V, et al. 2015. Fto obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* **373**: 895–907.

Cooper GS, Bynum MLK, and Somers EC. 2009. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *J Autoimmun* **33**: 197–207.

Cortes A and Brown MA. 2011. Promise and pitfalls of the immunochip. *Arthritis Res Ther* **13**: 101.

Davison LJ, Wallace C, Cooper JD, Cope NF, Wilson NK, Smyth DJ, Howson JMM, Saleh N, Al-Jeffery A, Angus KL, et al. 2012. Long-range DNA looping and gene expression analyses identify DEXI as an autoimmune disease candidate gene. *Hum. Mol. Genet.* **21**: 322–333.

Dendrou CA, Plagnol V, Fung E, Yang JHM, Downes K, Cooper JD, Nutland S, Coleman G, Himsworth M, Hardy M, et al. 2009. Cell-specific protein phenotypes for the autoimmune locus il2ra using a genotype-selectable human bioresource. *Nat Genet* **41**: 1011–1015.

Durinck S, Spellman PT, Birney E, and Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nat Protoc* **4**: 1184–1191.

Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, Onengut-Gumuscu S, Chen WM, Concannon P, Rich SS, et al. 2014. A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genet. Epidemiol.* **38**: 661–670.

Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, et al. 2016. The reactome pathway knowledgebase. *Nucleic Acids Res* **44**: D481–D487.

Fairfax BP, Makino S, Radhakrishnan J, Plant K, Leslie S, Dilthey A, Ellis P, Langford C, Vannberg FO, and Knight JC. 2012. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of hla alleles. *Nat Genet* **44**: 502–510.

Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, Shoresh N, Whitton H, Ryan RJ, Shishkin AA, et al. 2015. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**: 337–343.

Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nat. Genet.* **42**: 1118–1125.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**: 58–64.

Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, and Plagnol V. 2014. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**: e1004383.

Guo H, Fortune MD, Burren OS, Schofield E, Todd JA, and Wallace C. 2015. Integration of disease association and eQTL data using a bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum. Mol. Genet.* .

Gutierrez-Arcelus M, Rich SS, and Raychaudhuri S. 2016. Autoimmune diseases - connecting risk alleles with molecular traits of the immune system. *Nat Rev Genet* **17**: 160–174.

Hayter SM and Cook MC. 2012. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmun Rev* **11**: 754–765.

Huang H, Fang M, Jostins L, Mirkov MU, Boucher G, Anderson CA, Andersen V, Cleynen I, Cortes A, Crins F, et al. 2015a. Association mapping of inflammatory bowel disease loci to single variant resolution.

Huang J, Chen J, Esparza J, Ding J, Elder JT, Abecasis GR, Lee YA, Mark Lathrop G, Moffatt MF, Cookson WOC, et al. 2015b. eqtl mapping identifies insertion- and deletion-specific eqtls in multiple tissues. *Nat Commun* **6**: 6821.

Jäger R, Migliorini G, Henrion M, Kandaswamy R, Speedy HE, Heindl A, Whiffin N, Carnicer MJ, Broome L, Dryden N, et al. 2015. Capture hi-c identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**: 6178.

Kanehisa M and Goto S. 2000. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27–30.

Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, Kraft P, and Pasaniuc B. 2014. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**: e1004722.

Lawrie DS, Messer PW, Hershberg R, and Petrov DA. 2013. Strong purifying selection at synonymous sites in d. melanogaster. *PLoS Genet* **9**: e1003527.

Li Y and Kellis M. 2016. Joint bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res* .

Liao W, Lin JX, and Leonard WJ. 2013. Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity* **38**: 13–25.

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**: 289–293.

Liu JZ, McRae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, AMFSI, Hayward NK, Montgomery GW, Visscher PM, et al. 2010. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet* **87**: 139–145.

Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, Bailey R, Bourget K, Plagnol V, Field S, Atkinson M, et al. 2007. Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the il2ra region in type 1 diabetes. *Nat Genet* **39**: 1074–1082.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory dna. *Science* **337**: 1190–1195.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, and Cunningham F. 2010. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics* **26**: 2069–2070.

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, et al. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* .

Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, and Stein L. 2012. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel)* **4**: 1180–1211.

Miller AJ. 1984. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)* pp. 389–425.

Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, et al. 2014. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**: 376–381.

Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E, et al. 2015. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* .

Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, Hirschhorn J, Strachan DP, Patterson N, and Price AL. 2014. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**: 2906–2914.

Peters JE, Lyons PA, Lee JC, Richard AC, Fortune MD, Newcombe PJ, Richardson S, and Smith KGC. 2016. Insight into genotype-phenotype associations through eqtl mapping in multiple cell types in health and immune-mediated disease. *PLoS Genet* **12**: e1005908.

Pickrell JK. 2014. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**: 559–573.

Quinlan AR. 2014. Bedtools: The swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics* **47**: 11.12.1–11.1234.

Rainbow DB, Yang X, Burren O, Pekalski ML, Smyth DJ, Klarqvist MDR, Penkett CJ, Brugger K, Martin H, Todd JA, et al. 2015. Epigenetic analysis of regulatory t cells using multiplex bisulfite sequencing. *Eur J Immunol* **45**: 3200–3203.

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**: 1665–1680.

Schofield EC, Carver T, Achuthan P, Freire-Pritchett P, Spivakov M, Todd JA, and Burren OS. 2016. Chicp: a web-based tool for the integrative and interactive visualization of promoter capture hi-c datasets. *Bioinformatics* .

Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, et al. 2014. Obesity-associated variants within fto form long-range functional connections with irx3. *Nature* **507**: 371–375.

Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FAS, Zhernakova A, Hinks A, et al. 2010. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**: 508–514.

Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, Raubitschek A, Ziegler S, LeProust EM, Akey JM, et al. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**: 1367–1372.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**: 15545–15550.

The Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JMM, Auton A, Myers S, et al. 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**: 1294–1301.

Trynka G, Westra HJ, Slowikowski K, Hu X, Xu H, Stranger BE, Klein RJ, Han B, and Raychaudhuri S. 2015. Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within Complex-Trait loci. *Am. J. Hum. Genet.* **97**: 139–152.

van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, and Lander ES. 2010. Hi-c: a method to study the three-dimensional architecture of genomes. *J Vis Exp* .

Wakefield J. 2009. Bayes factors for genome-wide association studies: comparison with p-values. *Genet Epidemiol* **33**: 79–86.

Wallace C, Cutler AJ, Pontikos N, Pekalski ML, Burren OS, Cooper JD, García AR, Ferreira RC, Guo H, Walker NM, et al. 2015. Dissection of a complex disease susceptibility region using a bayesian stochastic search approach to fine mapping. *PLoS Genet* **11**: e1005272.

Yang J, Lee SH, Goddard ME, and Visscher PM. 2011. Gcta: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**: 76–82.

Yu G and He QY. 2016. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**: 477–479.

Yu G, Wang LG, Han Y, and He QY. 2012. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS* **16**: 284–287.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. 2016. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet* **48**: 481–487.