First Year Report - Draft

Oliver S Burren

August 9, 2016

Abstract

TODO

Contents

Chapter 1

Introduction

1.1 Motivation

A key property of the adaptive immune system is the ability to recognise pathogens from self-antigens and dysregulation of this process results in damage to healthy tissues and autoimmunity. At this time over 80 diseases have been found to have an underlying autoimmune pathogenesis, with approximately half presenting as rare diseases [Hayter and Cook, 2012]. The health burden of more common autoimmune diseases such as type 1 diabetes (T1D), rheumatoid arthritis (RA), inflammatory bowel disease (IBD) and multiple sclerosis is high with approximately 7 - 9% of the European population affected [Cooper et al, 2009]. Although environment is a contributing factor in disease susceptibility, genetic heritability, defined as the proportion of phenotypic variance attributable to genetic variability, is observed ranging from 0.39 in primary billiary cirrhosis (PBC) to 0.9 in ankylosing spondilitis (AS) [Gutierrez-Arcelus et al, 2016]. Genome-wide association studies (GWAS) have identified at least 324 distinct genetic loci associated with one or more autoimmune diseases (http://www.immunobase.org). Focus is now drawn to elucidating the mechanisms by which causal variation modulates phenotypic endpoints, a prerequisite for successful therapeutic development. 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 (PCH-C) information to provide data driven approaches to prioritising causal variation for autoimmune susceptibility, the tissue contexts within which it might operate, and the genes and biological pathways which it might peturb. To do this I develop fine mapping, variant set enrichment and pathway analysis

methods that use PHi-C data from 17 cell types to analyse GWAS data for 10 autoimmune and for comparison 23 non-autoimmune traits.

1.2 Fine mapping

Fine mapping is the process of refining association signals in a 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. If raw genotyping data for a study is available then stepwise regression could be used, however this approach whilst computationally tractable doubts exist over it's validity[cite millner]. Effectively having selected a single variant that best explains the variance of a trait, stepwise regression then 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. An alternative approach is to use a stochastic search method that allows for multiple causal variants within a region. One example is GUESSFM that uses a Bayesian evolutionary stochastic search algorithm to effectively sample the model space, which has been shown to have consistently better performance when predictors (SNPs) are highly correlated Wallace et al [2015].

The drawback of both of these methods is that they require raw genotyping data which for valid privacy reasons are hard to obtain. One method of fine mapping candidate causal variants using summary statistics is Wakefields synthesis of approximate Bayes factors (ABF) [Wakefield, 2009] to compute ABF for each SNP, assuming a single causal variant in a given genetic 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].

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

Here if $\hat{\beta}$ is our estimate of the log(Odds Ratio) and $\sqrt{V}$ is the standard error of the Odds Ratio then $Z = \frac{\hat{\beta}}{\sqrt{V}}$. r is a shrinkage factor which is the ratio of the prior variance on $\hat{\beta}$ to the total variance such that $r = \frac{W}{V+W}$. This means that for a SNP where we have only a univariate p-value we can estimate Z using an inverse normal cumulative distribution function, V is approximated using the MAF and study sample size such that $V = \frac{1}{2Nf(1-f)}$ for quantitiative traits, where N is number of samples and f is the SNP minor allele frequency. In the case/control setting $V = \frac{1}{2Nf(1-f)s(1-s)}$ where s is the proportion of cases . For an estimate of the standard deviation of a normal prior $\sqrt{W}$.

Given the ABF for all SNPs in a given genetic block we can estimate the posterior probability

that a SNP i is causal using the following.

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

as $\pi_0 = 1 - \sum_j \pi_j \approx 1$. π_i is our prior probability for any SNP selected at random to be causal for a trait. As current sample sizes are underpowered and limit resolution and this has driven the the development of integrative techniques that couple fine mapping methods with functional information 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] develope a Bayesian hierarchical frameworks implemented in *fgwas* that estimates variable priors for each SNP based on other sources of annotation. For a given SNP the prior probability for association, π_{ik}, varies depending on the enrichment of annotations, λ_l.

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j \in S_k} e^{x_j}} \tag{1.3}$$

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. When combined with equation 1.2 p_{ik} 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. Whilst promising fgwas requires careful cross-validation to prevent overfitting 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 annotations available.

1.3 Variant set enrichment

Understanding whether variants associated with a trait are enriched for a particular annotation or gene set can provide global information on mechanisms, relevant tissue contexts and biological pathways. However this is complicated by correlation between variants due to LD and between 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 can be almost as large as $\mathrm{Var}(A) + \mathrm{Var}(B)$, in this case the observed variance will 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. Permutation based methods, that can use GWAS summary statistcs are desirable as they can help to overcome this inflation by estimating the variance emprically. There are two broad permution methods, firstly using a suitable reference genotype set to compute covariance matrices allows the estimation of SNP variance under the null hypothesis using the multivariate normal Burren et al [2014]; Liu et al [2010]. These approaches are computationally expensive and scale exponentially with SNP density, thus an LD pruning strategy is employed which limits applicability to larger genomic annotations such as genes. The second 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 the this approach finding enrichment for H3K4me3 marks in CD+ memory T cells.

As previously mentioned integrative approaches such as *fgwas* combine variant set enrichment with fine mapping strategies to simultaneously prioritise annotations and variants. It should be noted that all methods are dependent on the quality and coverage of input annotations. Indeed 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 our lack of knowledge of genomic function [Huang et al, 2015a].

1.4 High resolution promoter capture Hi-C

Recently 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 resuting in covalent links between spatially adjacent chromatin segments. This chromatin is then digested with a restricition enzyme and sticky ends are filled in with biotin labelled nucleotides. Ligation is then performed

under dilute conditions which favours intramolecular 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 chromain 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 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

Maps of 17 primary human cells of the haematopoetic lineage were generated collaboratively with members of the Fraser, Spivakov, Ouwehand and Diabetes and Inflammation Laboratories. Each cell type was assessed over on average 3 technical replicates (Table 6.1). 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 $\geqslant 5$ were used in all downstream analysis.

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 Table6.2. 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, 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 odds ratio or β coefficients and their standard errors 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 dont map to the reference genotype set.

2.4 Variant posterior probabilities of inclusion for 31 traits using GWAS summary statistics

To fine map candidate causal variants I used Wakefields synthesis of approximate Bayes factors (ABF) [Wakefield, 2009] and using the method described in The Wellcome Trust Case Control Consortium et al [2012] computed 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 i^{th} 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. We 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 Variant posterior probabilities of inclusion using ImmunoChip summary statistics

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 quartiile range of between 126Kb and 392Kb) that were densely genotyped on the ImmunoChip [Onengut-Gumuscu et al, 2015].

To evaluate my PMI approach I used a set of marginal posterior probabilities of inclusion across fours diseases(autoimmune thyroid disease, celiac disease , rheumatoid arthritis and type 1 diabetes) generated by Chris Wallace for which we had access for full genotyping data from ImmunoChip using a stochastic search method that allows for multiple causal variants within a region [Wallace et al, 2015]. I incorporated these into further analysis on the assumption that full gentotyping data and imputation would provide more accurate posterior probabilities.

2.6 blockshifter - A competitive test for associated variant enrichment in PCHiC interaction maps

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 (PIR) in order to overcome 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 GWAS signal overlap, or that are found in both test or control set are discarded. 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 between the test and control set . Due to correlation between GWAS signals and between PIRs the variance of δ is inflated we thus compute it empirically using permutation. Runs of one or more PIRs (separated by at most one *Hind*III fragment) are combined into blocks, that are labeled 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 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**??**). We 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}. We use δ_{null} to compute an empirical Z-score:

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

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

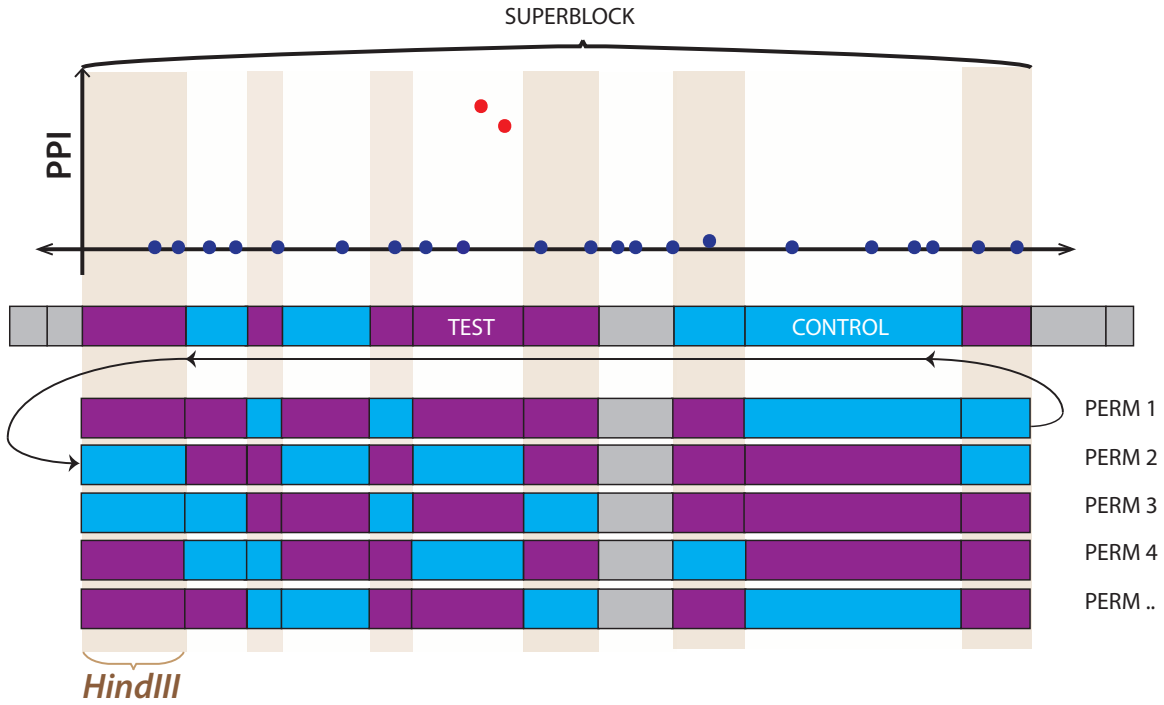


Figure 2.1: Illustration of circularised permutation method for mixed 'superblocks'. Test(Purple) and Control(Turquoise) tissue PIRs are rotated to generate permutations, *Hind*III fragments with no PIRs (grey) are fixed.

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

I developed an algorithm to compute tissue specific gene scores for each GWAS trait, taking into account linkage disequilibrium, interactions and functional SNP annotation (figure2.2). For each gene annotation, for which we have at least one significant interaction and recombination block (used by PMI (see above) to compute trait posterior probabilities) the algorithm computes a block gene score that is composed of three components.

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

2. The contribution due to promoter SNPs, which we define as SNPs that overlap a region encompassing the bait and flanking 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.

Thus 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. Assuming

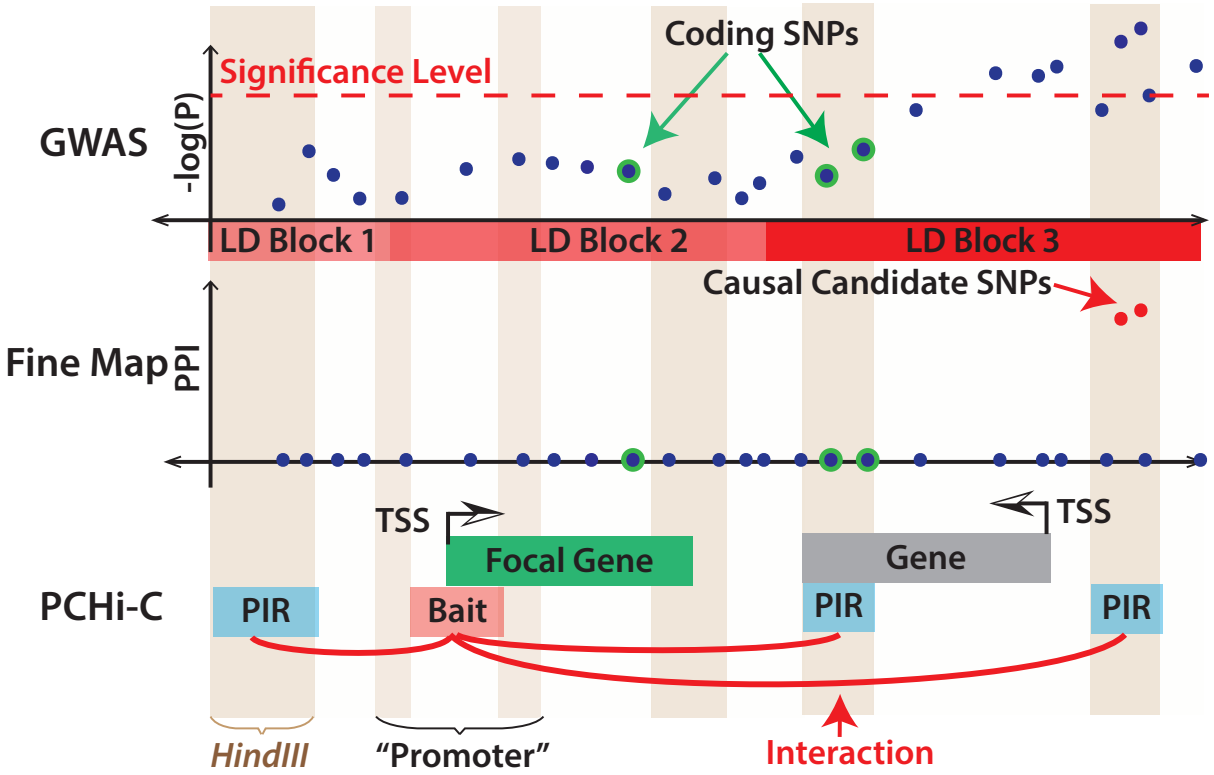


Figure 2.2: Illustration of COGS method

independence we can combine blocks to get an overall gene score such that:

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

Here $i \in R_j$ is the set of relevant SNPs i in the j^{th} region. Whilst components 1 and 2 are fixed for a given gene and trait the contribution of variants overlapping PIRs varies depending on the tissue context being examined. We developed a hierarchical heuristic method to ascertain for each target gene which was the mostly likely component and cell state. Firstly for each gene we compute the gene score due to genic effects (components 1 + 2) and interactions (component 3) using all available tissue interactions for that gene. We 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 (gene.score ratio >3) we 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 we compare interaction gene scores for activated vs non-activated cells. If at any stage no branch is substantially preferred over its competitor (ratio of gene scores < 3) we return the previous set as most likely, otherwise we continue until a single cell state/set is chosen. In this way we can prioritise genes based on the overall score and label as to a likely mechanism for candidate causal variants.

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 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 indentifiers 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.

2.9 Tissue specific HALLMARK gene set enrichment analysis

Using modified R code developed by Chris Wallace I used Wilcoxon rank sum tests to compare the distribution of gene prioritisation scores for genes in each MSigDB HALLMARK [Liberzon et al, 2015] set to its complement, within the set of genes that both had membership of at least one HALLMARK set and had a gene score. I used the p value from this test, together with the difference in mean gene prioritisation scores, to generate a signed Z score, Z_{ig} for trait i and gene set g. To test for relative enrichment in autoimmune diseases versus non-autoimmune traits, we used t tests to compare the distributions of Z_{ig} for i indexing autoimmune diseases to the Z_{jg} for j indexing non-autoimmune traits.

Chapter 3

Results

3.1 Comparison of PM with genotype level imputation

To validate performance of the PMI technique, I compared the results as imputed by PMI to those as reported by [Okada et al, 2014]. 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 . I confined my comparison to those SNPs imputed by PMI, a total of 235,412 SNPs. There was good agreement between PMI imputed p-values and those derived from classical imputation as reported in Okada, Wu, Trynka, Raj, Terao, Ikari, Kochi, Ohmura, Suzuki, Yoshida et al [2014] ($\rho = 0.9418103$, figure3.1).

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 as we expect PIRs to be enriched for regulatory regions, demonstration of robust enrichment within tissue specific PIRs is a pre-requisite for further analysis. I, in collaboration with Chris Wallace, developed a competitive test for enrichment called *blockshifter*, that takes into account correlation between GWAS and PIR data sets (see methods for details). Using the PMI imputed summary statistics from the 31 GWAS assembled(Table6.2) I found that variants associated with autoimmune disease are enriched in PIRs in lymphoid compared to myeloid tissues (figure3.2). 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 (figure3.2). I next examined whether we 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 specific to activated compared to non

activated $CD4^+$ T cells (figure3.3).

3.3 PCHiC assisted gene prioritisation across 31 traits

I next developed COGS, an algorithm to prioritise protein coding genes for each trait. This enabled me to prioritise 2,604 unique genes (Overall COGS score > 0.5) across all 31 traits examined. The prioritised genes exhibited enrichments for specific pathways in the Reactome pathway database [Fabregat et al, 2016]. 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 hemostasis (figure3.4). SNPs associated with traits generally unrelated to haematopoetic cells, such as blood pressure and bone mineral density did not show enrichment for PIRs in any of the cell types analysed.

3.4 Tissue specific PCHiC assisted gene prioritisation using dense summary statistics

blockshifter results lead us to examining autoimmune traits in the context of activated and non activated $CD4^+$ promoter interaction maps. We extended COGs to incorporate a simple binary decision tree model to allow tissue or functional category resolution (see methods). Across the combined GWAS and ImmunoChip autoimmune data sets we were able prioritise 602 distinct protein coding genes (Figure3.5). To summarise the behaviour of prioritisation, we focused on a subset of 220 input autoimmune GWAS regions with genome-wide significant signals ($P < 5 \times 10^{-8}$). We prioritised at least one gene with a COGS scores > 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 334kb and we 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 (add a table ?).

3.5 Gene set enrichment analysis of COGS prioritised genes using MSigDB HALLMARK gene sets

To check whether the list of prioritised genes corresponded to existing biological understanding of autoimmune disease, in collaboration with Chris Wallace we used gene set enrichment analysis of the genome wide COGS scores as described in the methods across 31 traits to MSigDB HALLMARK gene sets. We found significantly greater (Bonferroni $p < 0.05$) enrichment in autoimmune diseases compared to non-autoimmune traits in four gene sets relating to signalling

and receptor molecules as well as reactive oxygen species involved in early T cell activation (figure 3.6).

3.6 Comparison of COGS with TAD bounday prioritisation

3.7 Prioritised gene overlap with eQTLs

In collaboration with Oliver Stegle and Roman Kretzhuber, using SLE [Bentham et al, 2015] and rheumatoid arthritis [Okada et al, 2014] data sets, for which full imputed summary statistics were available. we performed COGS gene prioritisation and 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-localization 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 the causal variant functions in a specific tissue context and thus it's eQTL is not observable in available eQTL datasets. The implication of this is that regulatory annotations that might be used to augment fine mapping methods are similarly incomplete.

3.8 CHiCP: a genome browser for PCHiC interaction maps

In collaboration with Ellen Schofield, I developed a web based application, CHiCP `https://www.chicp.org`, to allow the visual integration of both public and private promoter capture Hi-C maps with GWAS and BLUEPRINT epigenetic data [Schofield et al, 2016]. To do this I developed an interactive circularised display, with a search interface for genes, variants and chromosomal coordinates. An interaction can be selected, allowing users to access a conventional browser view to see a zoomed in view of chromatin states and GWAS signals at promoter and promoter interacting regions.

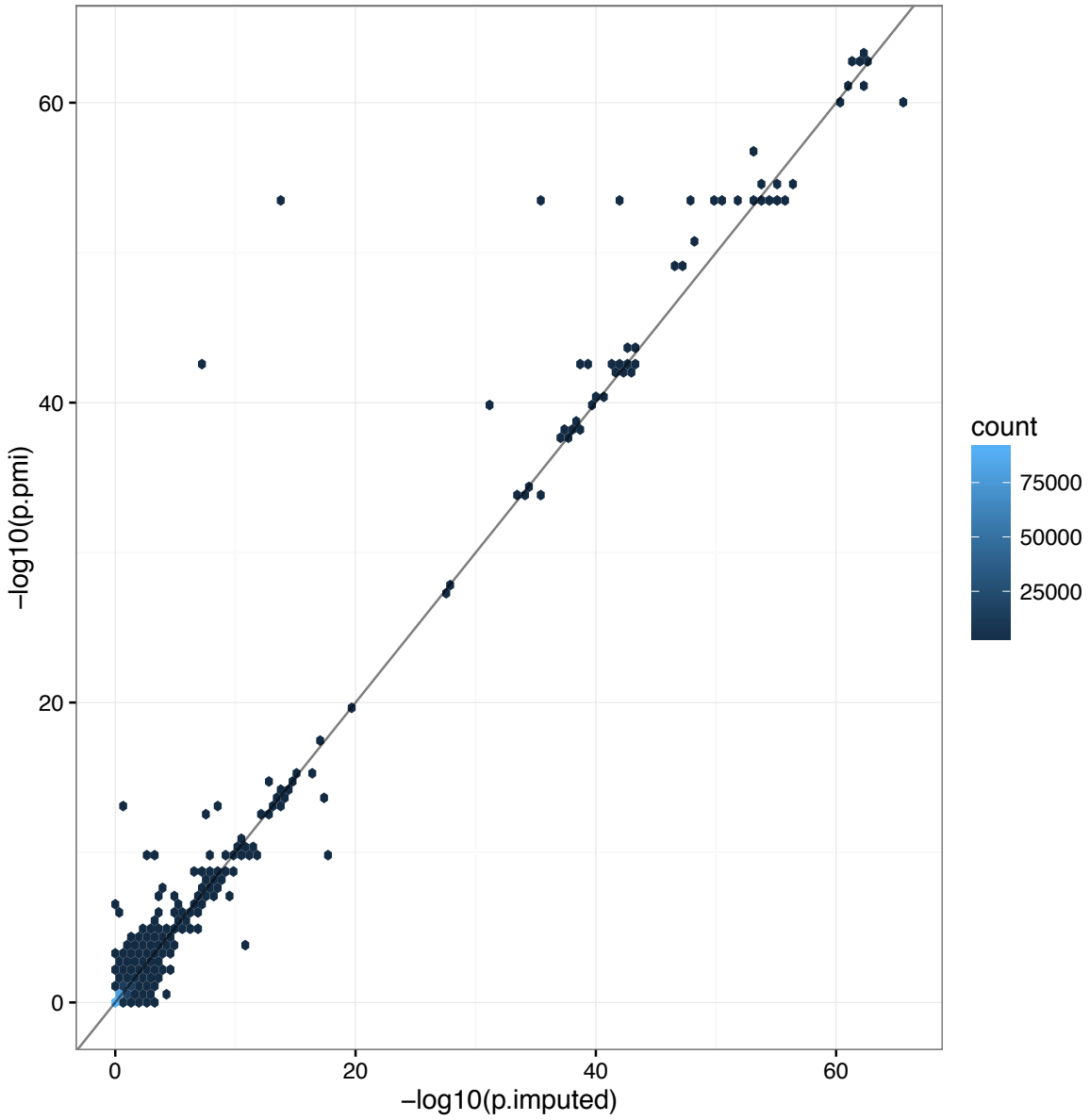


Figure 3.1: Comparison of P-values imputed by PMI versus those reported in [Okada et al, 2014] for chromosome. Axis are $-\log_{10}$ transformed P-values

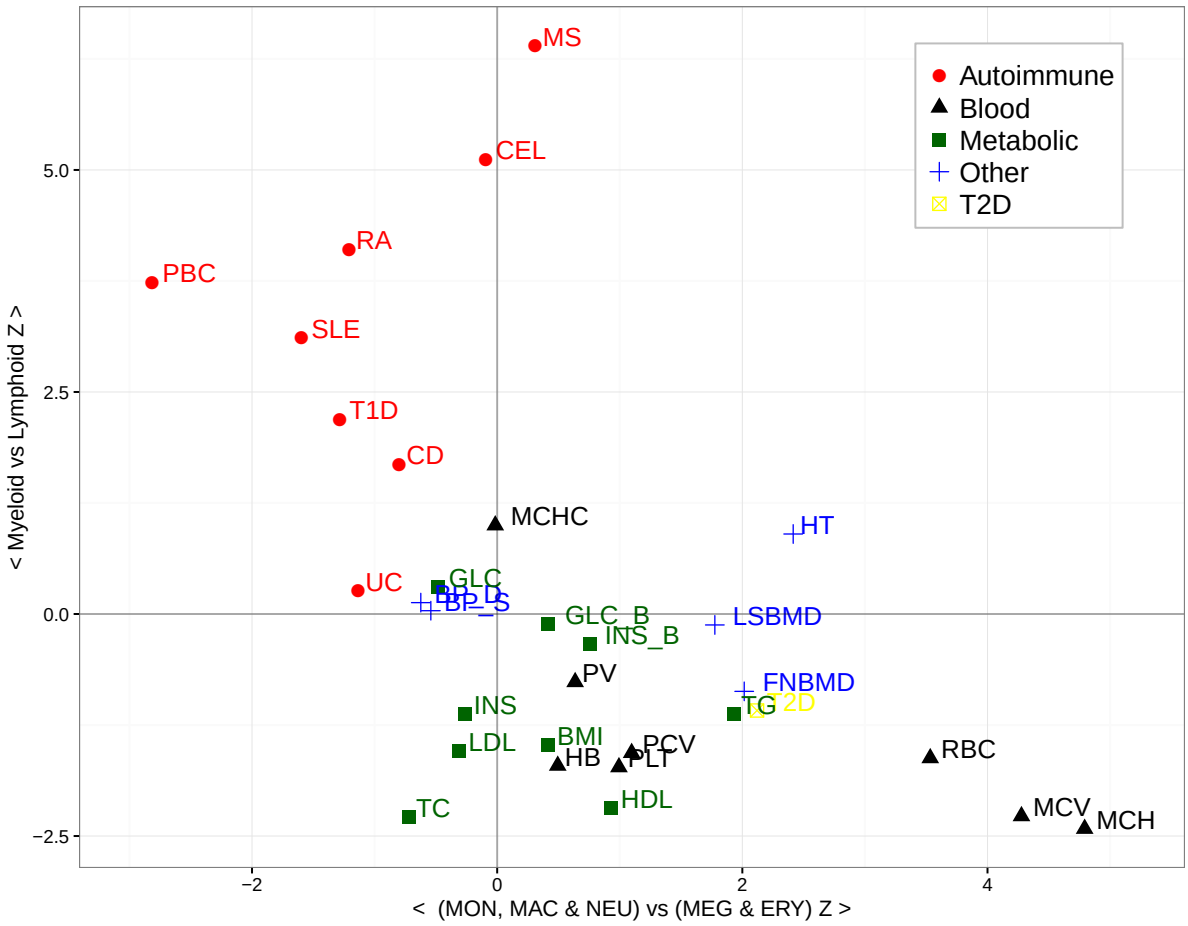


Figure 3.2: Enrichment of GWAS summary statistics at PIRs by tissue groups. Axes reflect *blockshifter* Z-scores for two different tissue group comparisons, firstly lymphoid versus myeloid and then within myeloid lineage (MON - Monocyte, MAC - macrophage and NEU - Neutrophil) versus (MEG - Megakaryocyte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

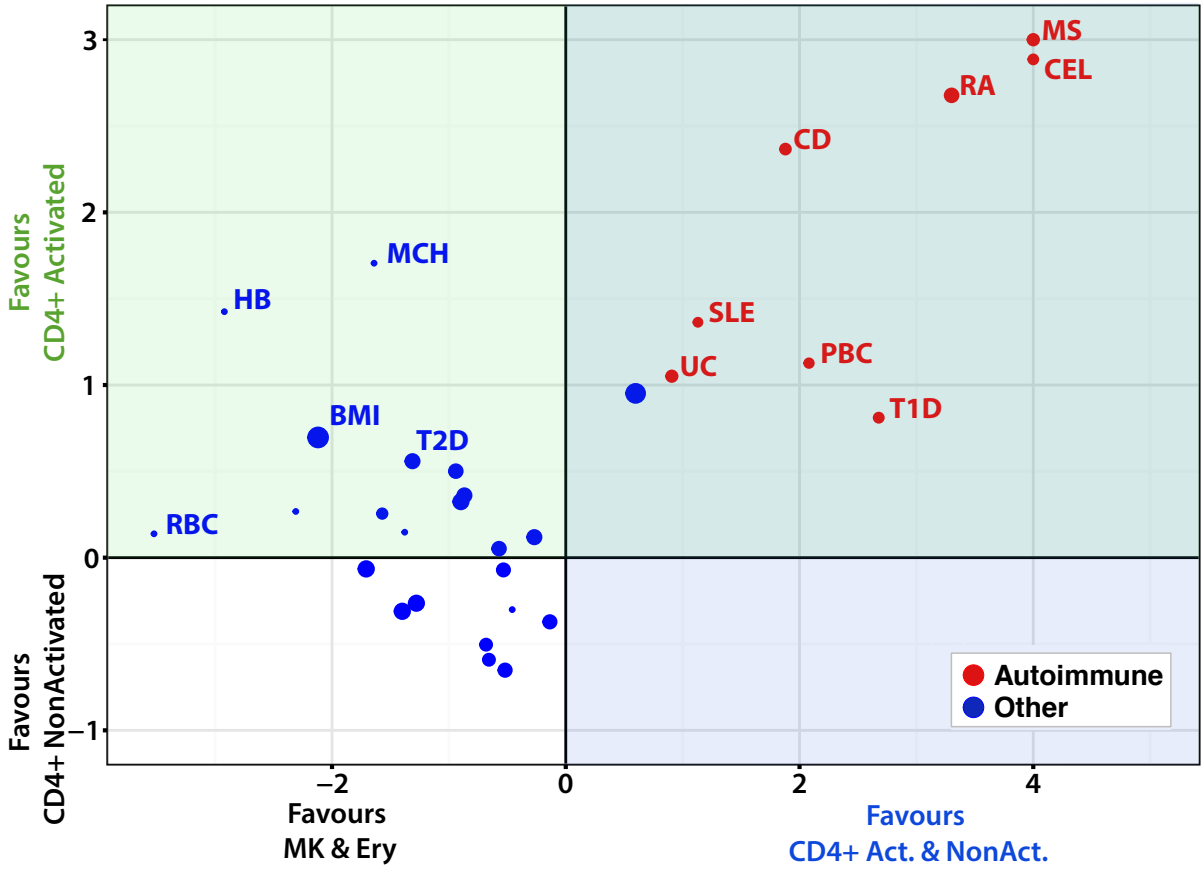


Figure 3.3: Tissue specific enrichment of autoimmune GWAS summary statistics at PIRs by tissue groups. Axes reflect *blockshifter* Z-scores for two different tissue group comparisons, firstly MK - megakaryocytes and ERY - erythroblast versus Activated/Non-activated CD4$^+$ T cells. Y-axis shows comparison 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.

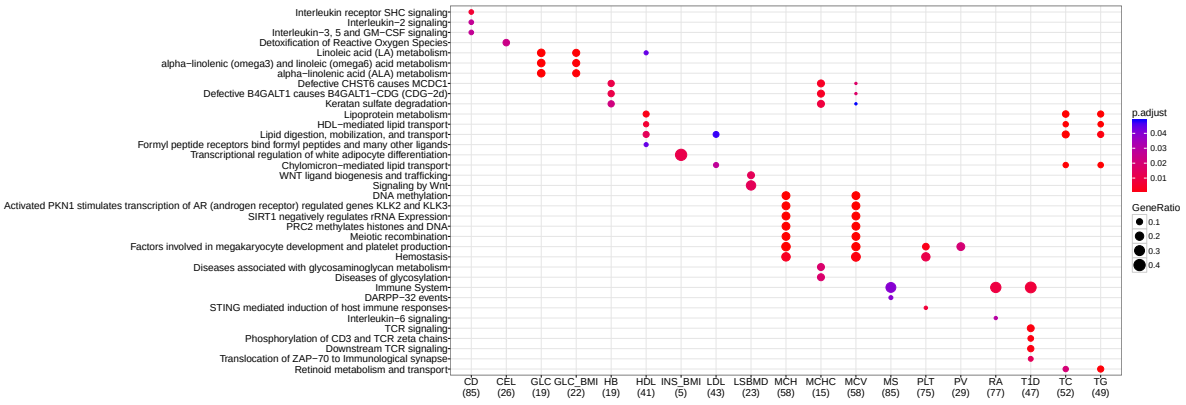


Figure 3.4: Bubble plot of traits with significant enrichment ($P_{adj} < 0.05$) in one or more reactome pathways for genes with COGS score > 0.5 across 31 traits. Number in parentheses below trait labels indicate the total number of genes analysed for each trait, bubble size indicates the ratio of test genes to those in the pathway, and blue to red shading corresponds to decreasing adjusted P-value for enrichment

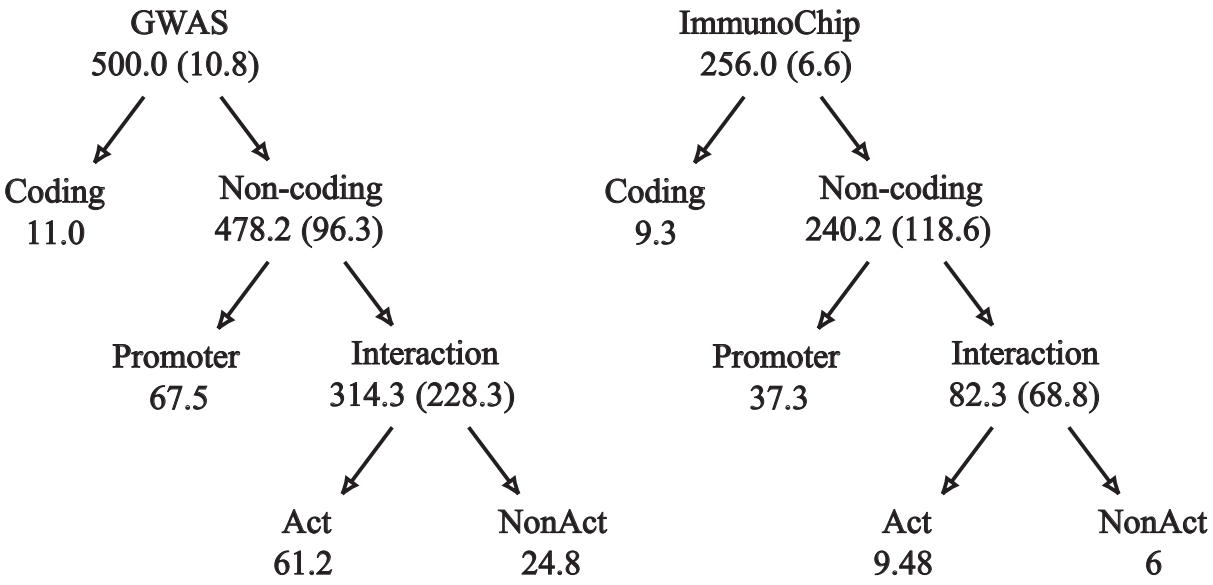


Figure 3.5: Functional gene prioritisation across 11 autoimmune diseases using genome wide (GWAS) or targeted genotyping array (ImmunoChip) data. The numbers at each node give the number of genes prioritised at that level. Where there is evidence to split into one of two non-overlapping hypotheses ($\log_{10}$ 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, we assigned a fractional count to each node, defined as the proportion of the n disease for which the gene was prioritised at that node.

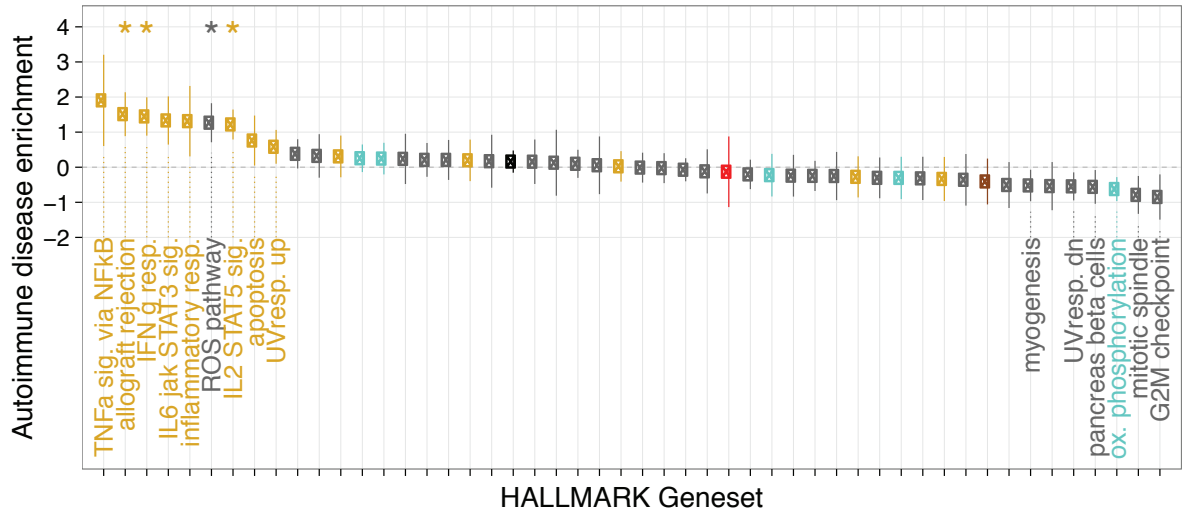


Figure 3.6: Tissue specific gene set enrichment analysis using COGS gene prioritisation scores for 11 autoimmune traits. Colours correspond to shared pathways enriched in modules identified by weighted gene correlation network analysis (WGCNA) of activated vs non-activated $CD4^+$ T cells RNA-Seq data. Points give the difference in average enrichment of Z scores across autoimmune versus non-autoimmune GWAS and bars the 95% confidence interval. Significant gene sets after Bonferroni correction are indicated by $*$.

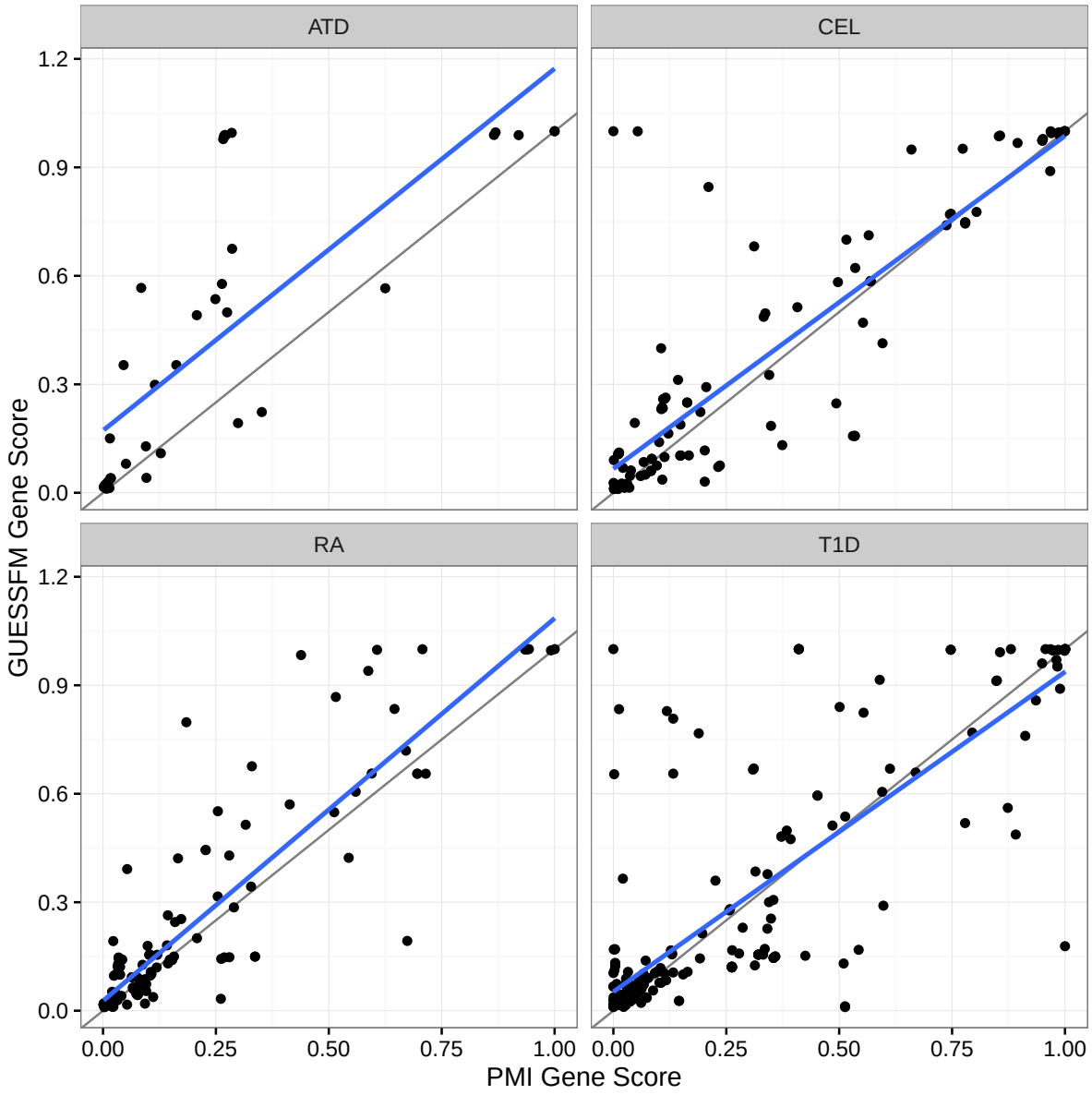
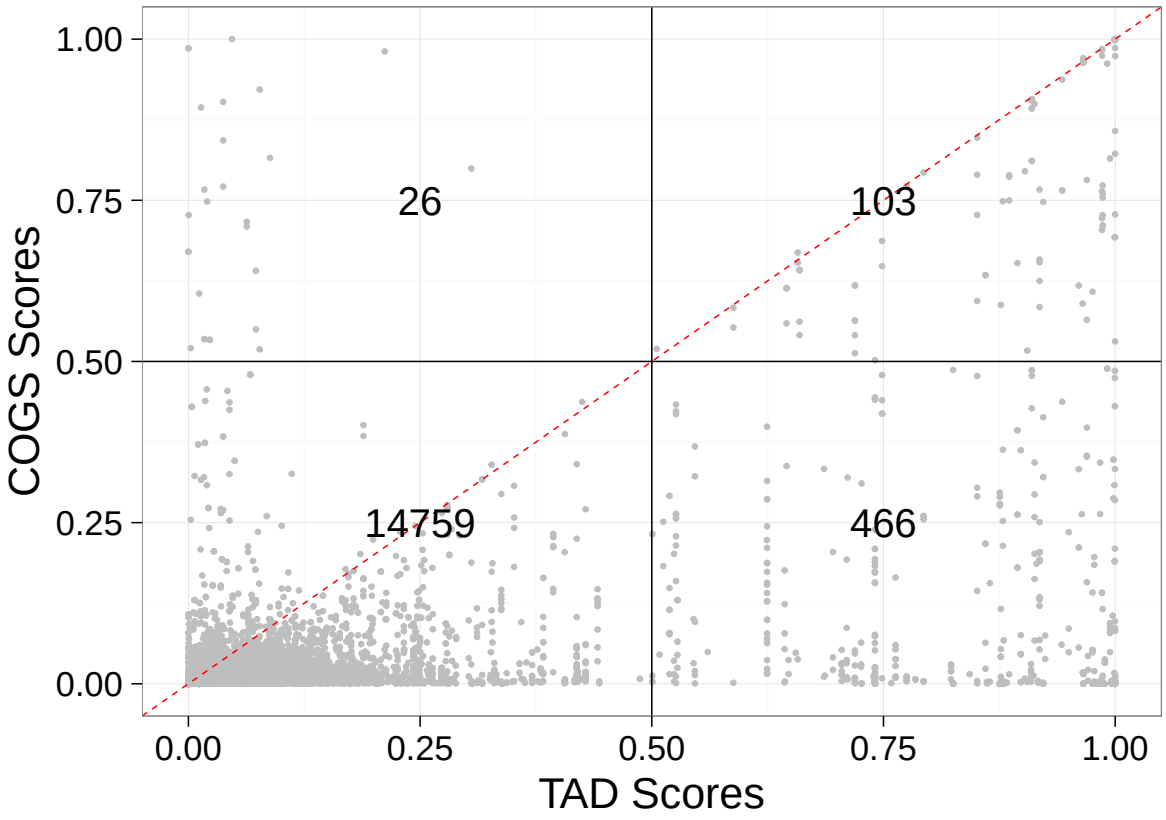
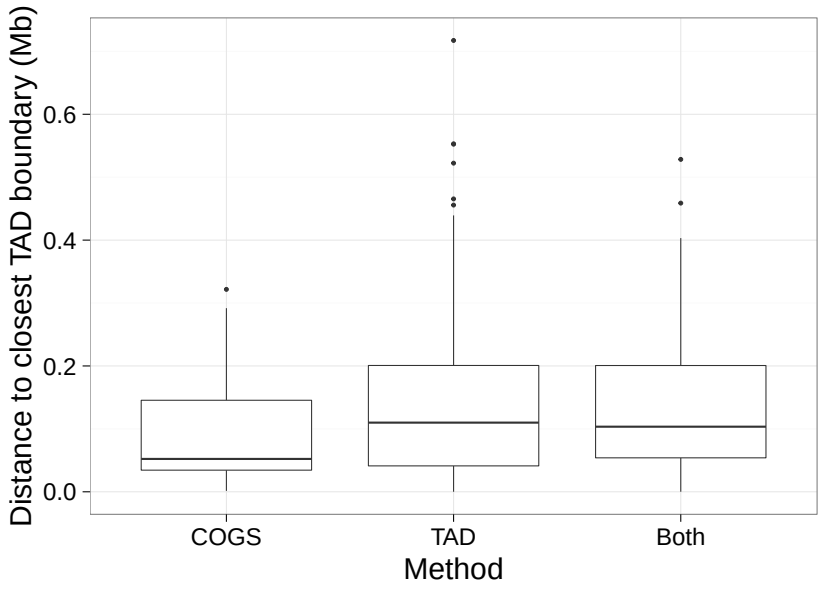


Figure 3.7: Comparison of COGS gene scores for four autoimmune traits (ATD = autoimmune thyroid disease, CEL = celiac disease, RA = rheumatoid arthritis and T1D = type 1 diabetes computed using marginal posteriors from GUESSFM method and those computed using posteriors from PMI, grey line shows $y = x$, blue lines indicate the fitting of a linear model



(a)



(b)

Figure 3.8: Comparison of PCHi-C COGS scores and Hi-C derived TAD scores, using 7 Cell types (Erythroblasts, Macrophages, Monocytes, Naive B cells, Naive $CD4^+$ T cells, Naive $CD8^+$ T cells and Neutrophils) using rheumatoid arthritis GWAS data from Okada et al. (a) Comparison of scores between the two methods across a total of 15,355 protein coding genes with sufficient coverage, a line of equivalence is marked in red, counts represent the number of genes in each quadrant. (b) 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.

Chapter 4

Discussion

4.1 Summary

Methods to integrate genetic and genomic data sets are important in order to identify causal variants, genes and pathways and the tissue contexts within which they operate. I have developed two Bayesian approaches to integrate GWAS summary statistics and targeted genotyping data (ImmunoChip) with high resolution chromatin conformation data. Firstly *blockshifter*, a method to perform competitive tissue enrichment analysis in the presence of correlation in order to prioritise relevant tissue contexts. Secondly COGS, a method to prioritise causal candidate genes based across tissue contexts and traits. I used *blockshifter* to verify the relevance of chromatin interactions with activated and unactivated $CD4^+$ T cells to autoimmune traits. Using COGS I was able to prioritise a number of genes for further study, I performed gene set analysis using a standard hypergeometric method on the Reactome database and a tissue specific GSEA using MolSigDB HALLMARK gene sets which indicated that COGS prioritised genes were enriched for both expected and unexpected pathways.

4.2 Functional validation of COGS prioritised gene *IL2RA*

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 which is 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(add references). 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**??**) identified in a previous fine mapping study [Wallace et al, 2015]. This 'A' set of SNPs(Figure**??**a) 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 using software I developed previously [Rainbow et al, 2015] we measured the relative expression of the two alleles at one of these set 'A' 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. This was further validated using rs12244380 found in the 3' UTR of *IL2RA* (Figure4.2).

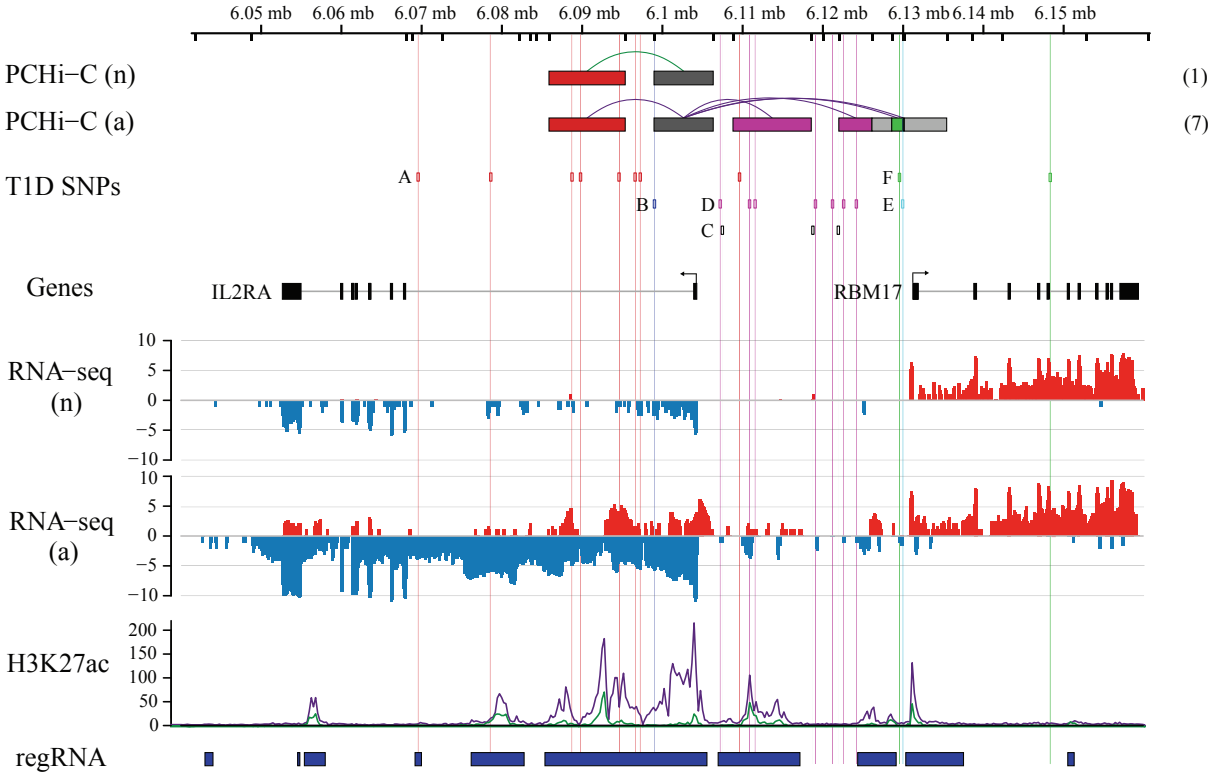


Figure 4.1: PCHi-C interactions link the *IL2RA* promoter to autoimmune disease associated genetic variation which leads to expression differences in *IL2RA* mRNA.

4.3 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. This clearly effects the output of the COGS algorithm (figure 3.7) where I compare COGS scores for genes using PMI approach and using GUESSFM which considers any number of causal variants. Whilst there is no "gold standard" in COGS gene scores to compare, in general GUESSFM seems to prioritise more genes than PMI (Table **??**). A portion of this might be better the resolution that the GUESSFM approach favours, however some of this will be offset by it's sensitivity to genotyping error.

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, in practice an interaction with a score of 4.99 will be omitted. Future approaches might investigate methods for incorporating promoter interaction scores in both *blockshifter* and COGS methods. One

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 4.1: 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.

approach might be to look at techniques that utilise CHiCAGO scores across multiple tissues for a given interaction to adjust local FDR.

COGS assumes that coding variation effects 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. With this in mind, questions include whether such variation has a cis effect on other genes by functioning as regulatory DNA as well as whether we can improve the granularity of coding variant hypothesis by further subdivision (e.g. non-synonymous and synonymous).

One significant problem 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 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. The 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 we include all transcriptional start site annotations in Ensembl (Version 75) then this rises to 6,703 (40%), in reality 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 PCHI-C data alone, one possible naive solution is to integrate context specific expression data to match genes expressed in that context to possible context specific interactions. This leads on to another possible extension which is to extend prioritisation to include annotated non coding genes as a recent study suggests that these modulate autoimmune disease susceptibility [Castellanos-Rubio et al, 2016].

Excluding single cell implementations all genomic technologies give an overview of the molecular events across the range 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.

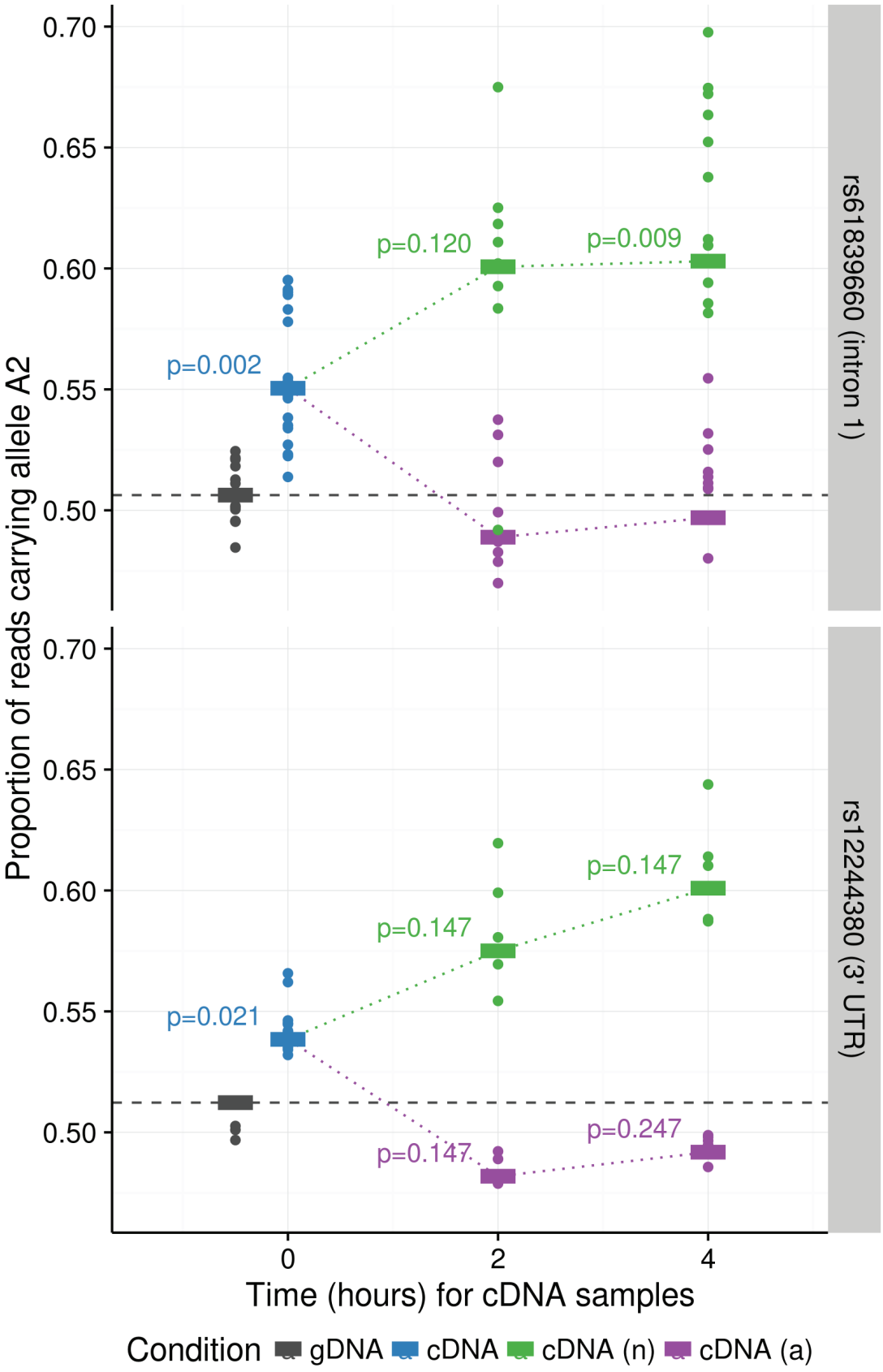


Figure 4.2: =*IL2RA* ASE data - needs caption.

Chapter 5

Future Work

5.1 Application of COGS scores to the full PCHiC dataset

Having shown evidence that the COGS tissue specific prioritisation score is useful, I have begun the process of applying it across all traits to all 17 tissue types. For each trait it should be possible to generate a decision tree as in Figure 5.1. If this is successful I would like to examine metrics for comparing tree structures and therefore inter trait relationships (Figure 5.2).

5.2 Comparison with Hi-C defined TAD boundaries

So far it is unclear as to how much information is gained in gene prioritisation using promoter capture Hi-C over classical Hi-C. Csilla Varnai and Michiel Thiecke have generated TAD domain boundaries for 8 of the 17 cell types assayed using conventional Hi-C libraries. I will use these to generate pseudo interaction matrices whereby within each TAD the promoter of all genes interacts with every fragment, excepting the baited fragment, and those bounding it for a given gene within a TAD. For these tissues for which Hi-C libraries are available I will also compute gene scores using COGS and analyse the differences to see whether information is gained from using promoter-capture Hi-C over classical Hi-C.

5.3 Integrating other annotations into COGS

PMI assumes a fixed prior for all SNPs however recent work [Pickrell, 2014] using Bayesian hierarchical frameworks demonstrates a method by which we can estimate variable priors for each SNP based on other sources of annotation. In the hierarchical method implemented by *fgwas* for a given SNP the prior probability for association, π_{ik}, varies depending on the enrichment of annotations, λ_l.

$$\pi_{ik} = \frac{e^{x_i}}{\sum_{j \in S_k} e^{x_j}} \tag{5.1}$$

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.

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il} \tag{5.2}$$

Initially I would use further information on $CD4^+$ T cells including various histone-modification marks and methylation data alongside public ally available annotations to compute λ_l using *fgwas*. Using a model including the most relevant annotations these x_i values could also be applied to compute variable priors (π_i) for dense ImmunoChip summary statistics. These more annotation aware posterior probabilities will then be input into COGS to see how this alters gene prioritisation. If this was successful I would look into the incorporation of other BLUEPRINT annotations across all 31 traits.

5.4 blockshifter development - perhaps omit

Demonstrating the enrichment of an annotation for associated variants robustly in the presence of correlation is challenging. Correlation, occurring through linkage disequilibrium between SNPs and between genomic annotations, has the effect of inflating the underlying text statistic and must be carefully adjusted for. The *blockshifter* approach does this using a mixture of circularised and weighted permutation techniques using underlying *Hind*III fragments as atomic units to define underlying block structure. I will investigate the utility of *blockshifter* outside of the PCHiC initially looking at it's performance using (un)activated $CD4^+$ T cell ChIP-Seq data sets. In this case I am not limited to *Hind*III but will assay the performance of more frequent cutters (e.g. *SspI*) in terms of results and computational efficiency.

5.5 Tissue specific gene set enrichment analysis

TODO

5.6 Data driven discovery of relevant COGS score thresholds

TODO

5.7 Future Directions

In the longer term I hope to collaborate with clinicians to build promoter capture Hi-C maps of relevant tissues in disease and healthy states. Controlling for variation between individuals in this context is paramount, thus employing strategies to overcome this within the bounds of the economic and technical limitations is important. One strategy is to compare diseased and healthy tissue within the same individual, taken at the same time point to look for differences in

interactions, an example of this might be in juvenile idiopathic arthritis (JIA), here the diseased tissue is known and samples, in the form of synovial fluid drains for are available. PCHi-C results generated from CD4+ T cells isolated from the same individual might be compared with those isolated from the periphery. Another complimentary strategy for systemic autoimmune disease where diseased tissue is challenging to collect is to take a longitudinal approach, here patients presenting with a disease are assayed, with further follow up assays collected over time taking into account any therapeutic interventions. Genomic comparison between time points is then conducted to look for differential interactions that might be related to disease prognosis. Further evidence of for a genes involvement in autoimmune disease susceptibility might come from adapting the methods previously described to work with rare variants. For this work one could use gene level variant aggregation tests [Lee et al, 2013] and then use genotype data from primary immune disorders collected as part of the BRIDGE study, PCHiC data would be used to assign variants to genes allowing the integration of both coding and non-coding variation.

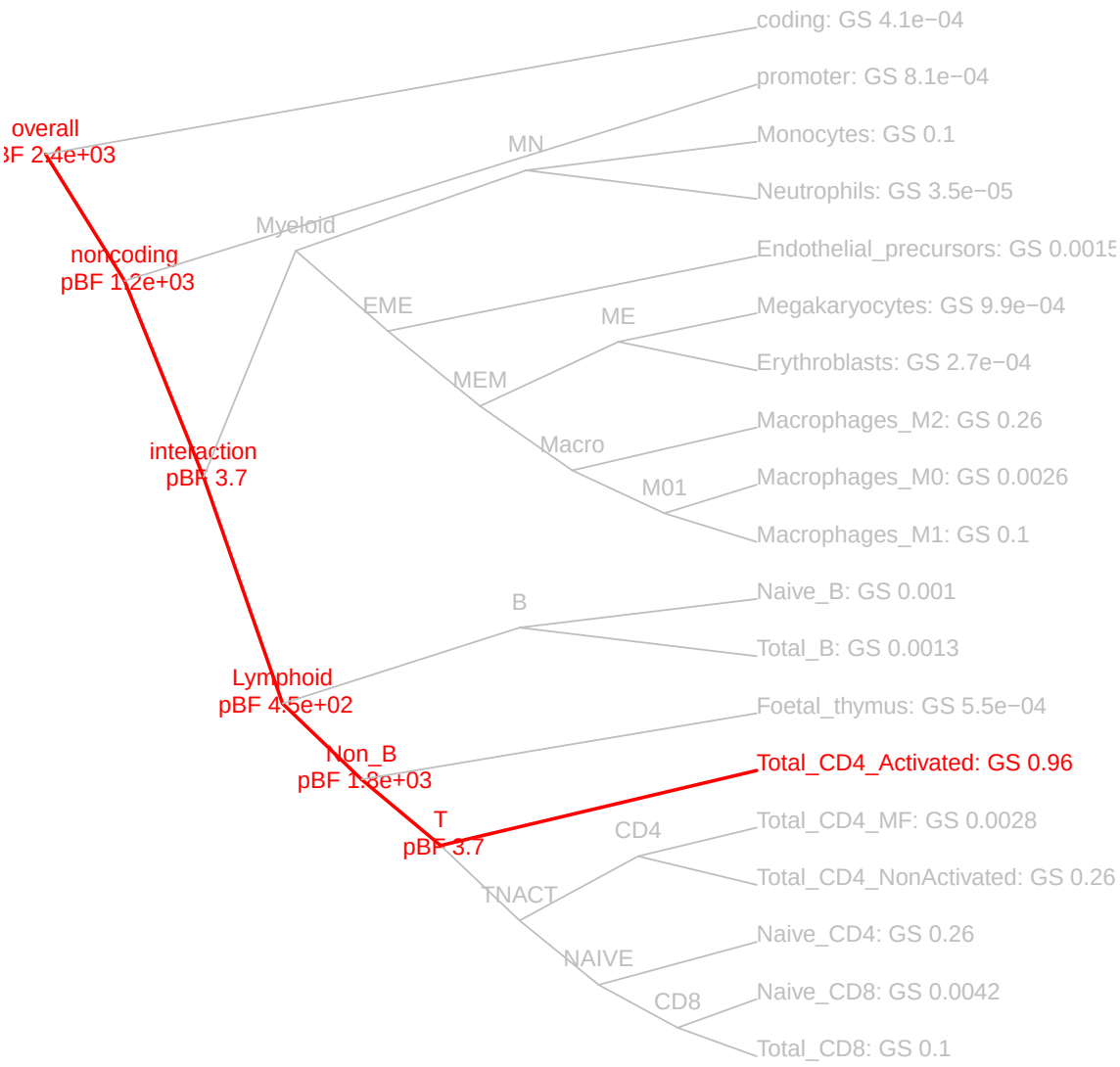


Figure 5.1: COGS decision dendrogram using rheumatoid arthritis data from Okada et al and PCHi-C maps from 17 haematopoietic cell types for the *AHR* gene. The red edges denote the path taken by COGS through the binary decision tree. Each node is labelled with pseudo Bayes factors(pBF). Terminal nodes are labelled with the tissue/annotation specific gene score(GS)

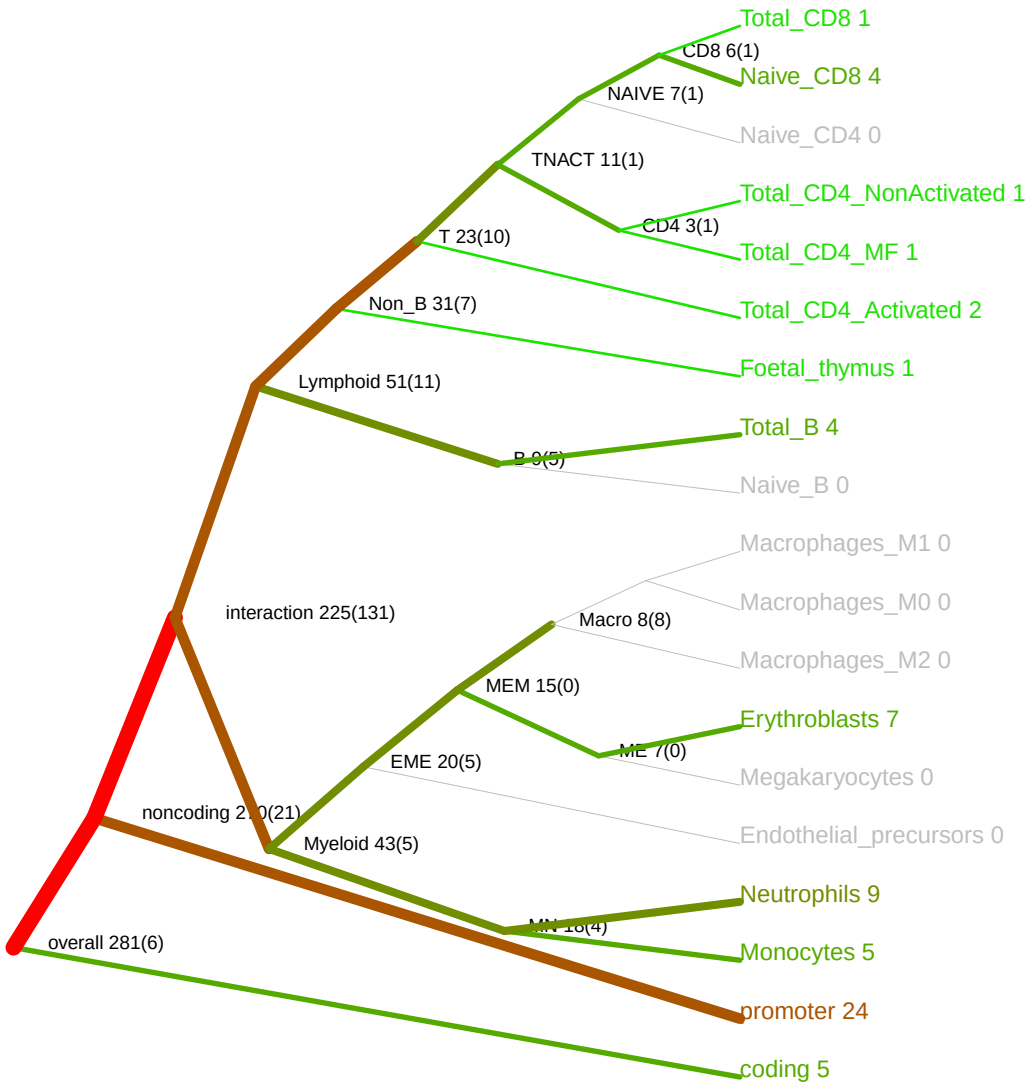


Figure 5.2: COGS decision dendrogram for all prioritised protein coding genes (GS $>$ 0.5) using rheumatoid arthritis data from Okada et al and PCHi-C maps from 17 haematopoietic cell types. Edges are coloured based on number of genes flowing between connected nodes. Nodes are marked with the total number of genes at a node and in brackets the number of genes that are assigned to that node.

Chapter 6

Appendix

Table 6.1: Summary of PCHi-C datasets generated in this study

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 6.2: 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	LS-BMD	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

Bibliography

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

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.

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.

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.

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.

Lee S, Teslovich TM, Boehnke M, and Lin X. 2013. General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet* **93**: 42–53.

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* .

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, and Tamayo P. 2015. The molecular signatures database (msigdb) hallmark gene set collection. *Cell Syst* **1**: 417–425.

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.

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.

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.* .

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.

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.

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.