Word Count: 5818

Exploring associations between cytokines/chemokines and immune mediated diseases with genetic feature engineering

Qingqing Zhou

Clare College

University of Cambridge

A thesis submitted for the degree of

MPhil in Computational Biology

August 2022

Acknowledgements

I would like to express my sincere thanks to Dr. Chris Wallace and Dr. Guillermo Reales for giving me this extraordinary opportunity to work on this project, and for their kind support and patient supervision all along.

I would also like to thank all the authors of < Genetic feature engineering enables characterization of shared risk factors in immune-mediated diseases > for their wonderful work on developing the method of basis construction and projection, without which this project would not have been possible.

Finally, I would like to thank everyone in Chris's group for their kind advice and help during this internship.

Abstract

Cytokines are a broad family of secreted proteins that play pivotal roles in regulating immunity. Previous studies on cytokines have established their extensive involvement in the pathogenesis of immune-mediated diseases (IMD). However, due to the pleiotropic effects and redundant functions of cytokines, it is challenging to systematically dissect associations between cytokines and IMDs using conventional immunochemical tools. Over the past decades, genome-wide association studies (GWAS) have successfully identified many single nucleotide polymorphisms (SNP) associated with IMDs that affect the circulating levels of cytokines/chemokines (Charles et al., 2018) (Ferkingstad et al., 2021). It is therefore tempting to integrate GWAS datasets of cytokines and IMDs to comprehensively study the associations between them.

In this project, using a method that combines Bayesian shrinkage and PCA, I constructed a genetic basis derived from 40 cytokines/chemokines GWAS datasets. Each principal component (PC) of this basis essentially represents a genetic pattern summarized from training datasets. The replicability of this basis was confirmed by the consistency of significant projection results from external cytokine/chemokine test datasets. A total of 481 IMD traits were projected to this basis. Among these, 79 showed significant projection results in at least one principal component. In particular, PC 7, 9 and 10 were the most enriched for significant IMD projections, and each of them revealed extensive associations between cytokines/chemokines and IMDs. Importantly, many of these associations are supported by direct or circumstantial evidence from experimental studies, which emphasizes the robustness of this model. Moreover, a number of novel associations between cytokines/chemokines and IMDs were identified in this study.

In conclusion, the results presented here offer valuable insights that can lead to further experimental work.

Contents

List of Figures

List of Abbreviations

IMD . . . . . Immune-mediated disease

GWAS . . . . . Genome-wide association studies

SNP . . . . . Single nucleotide polymorphisms

PC . . . . . Principal component

IL . . . . . Interleukin

TNF . . . . . Tumor necrosis factors

T1D . . . . . Type 1 diabetes

NOD . . . . . Non-obese diabetic

RA . . . . . Rheumatoid arthritis

IBD . . . . . Inflammatory bowel disease

pQTL . . . . . Protein quantative trait loci

PCA . . . . . Principal component analysis

LD . . . . . Linkage disequilibrium

BMK . . . . . Biomarker

PSD . . . . . Psychiatric disorder

INF . . . . . Infectious disease

CAN . . . . . Cancer

OTH . . . . . Others

JIA . . . . . Juvenile idiopathic arthritis

HLA . . . . . Human leukocyte antigen

VEGF . . . . . Vascular endothelial growth factor

AITD . . . . . Autoimmune thyroid disease

1 Introduction

1.1 Cytokine and chemokines

Cytokines comprise a diverse family of small (typically 5~40 kDa in size), secreted proteins that play pleiotropic functions in a variety of biological processes. These include tissue development, haematopoiesis, and most importantly, immunity. Cytokines are secreted by a broad range of cell populations, among which the most significant ones are macrophages and CD4+ T cells. Once released, cytokines can relay biological information to target cells (paracrine action and endocrine action) or regulate the activity of the same cell from which it is secreted (autocrine action). In humans, over 150 common cytokines have been identified so far. Based on their target cells and presumed functions, cytokines can be further classified into several major subtypes including interferons, interleukins, tumor necrosis factors, chemokines etc. Interferons are named after their function in interfering with viral replication. In response to a viral infection, interferons can be secreted by the infected cells or activated T cells to induce resistance to the invading pathogen. Interleukins represent some of the earliest identified cytokines that were initially thought to be secreted exclusively by leukocytes. However, more recent studies have shown that they can be produced by other cell types as well (Swiecki & Colonna, 2011). The primary function of interleukins is to regulate cell growth and migration

as well as the activation of the immune system. For example, Interleukin-1β (IL-1β), a member of the IL-1-like family of cytokines, has been shown to be a crucial mediator in the expression of various genes involved in secondary inflammation (Weber et al., 2010). Another interleukin, IL-16, is known to recruit CD4+ T cells and inhibit the CD3/TcR-dependent activation of lymphocytes (Cruikshank et al., 1996). Tumor necrosis factors (TNFs) are a class of cytokines well known for their role in mediating cell apoptosis. As the first characterized TNF, TNF-α has been shown to trigger cell death via caspase-8 activation pathways (Wang et al., 2008). In addition, it is also a key mediator in both acute and chronic inflammatory reactions (Cameron & Kelvin, 2013).

Chemokines are a special subset of chemotactic cytokines that play important roles in the development and homeostasis of the immune system. They are best characterized for their function in directing leukocyte movement through chemotaxis. Chemokines target G-protein coupled receptors (GPCRs) on the membrane of leukocytes. Such receptors in turn activate a vast number of downstream effectors via G protein mediated intracellular signaling cascades, and ultimately lead to changes in actin cytoskeleton dynamics (Wu et al., 2010) (Cancelas et al., 2006). Over the past decades, more than 50 chemokines have been identified. Based on the consensus cysteine-containing motif at the N-terminal, chemokines are classified into four major groups: C, CC , CXC and CX3C chemokines, where X denotes any single amino acid present between the cysteines.

Given their extensive involvement in both humoral and cell-mediated immunity, it is not surprising that dysregulation of cytokines and chemokines has been implicated in the pathogenesis of multiple immune mediated diseases (IMD). For example, type 1 diabetes (T1D) is an autoimmune disease characterized by loss of β-cell derived insulin due to progressive chronic inflammation in the pancreatic islets. Previous immunopathogenic studies have highlighted multiple cytokines and chemokines as causative factors in T1D. One of them, IL-1 has been reported to induce the apoptosis of pancreatic cells in non-obese diabetic (NOD) mice. Similarly, inhibition of IL-1 with antagonists significantly reduced the level of

inflammation in T1D patients (Cabrera et al., 2016) (Mandrup-Poulsen et al., 2010). Elevated circulating levels of several pro-inflammatory chemokines, such as CCL21, CCL19, CXCL6, CXCL10 and CXCL12, have also been reported by independent studies in T1D patients and animal models (Lu et al., 2020) (Gouda et al., 2018) (Alagpulinsa et al., 2019). Other well-studied IMDs, where cytokines and chemokines are known to contribute to disease etiology, include rheumatoid arthritis (RA), celiac diseases, inflammatory bowel disease (IBD), sarcoidosis and vitiligo. Rheumatoid arthritis is a chronic autoimmune and inflammatory disease characterized by the progressive destruction of joints, bones and other organs. Evidence indicates that the expression levels of CCL19/CCL21 are significantly increased in the synovial fluid of RA patients. In addition, CCL19/CCL21 are known to induce the production of proangiogenic factors by macrophages and fibroblasts (Pickens et al., 2011). In patients with IBD, the CXCL12/CXCR4 axis has been shown to be upregulated in the intestinal epithelium. The potent role of CXCL12 in the intestinal immune system is also supported by studies on mice with DSS-induced colitis. In these animals, treatment with AMD3100, an antagonist that disrupts the interaction between CXCL12 and CXCR4, can indeed attenuate colonic damage (Xia et al., 2010).

Due to their close relationship with immunity, cytokines and chemokines circulating in the serum and other biological fluids are broadly used as biomarkers in the early diagnosis and prognosis of immune mediated diseases (Monastero & Pentyala, 2017). However, most cytokines/chemokines have pleiotropic functions in different biological processes. Cross-talks among these processes make it hard to differentiate if a given cytokine is truly causative in the pathogenesis of a disease, or simply affected by it. Furthermore, screening of potential immunotherapeutic targets would require a thorough understanding of the cytokines/chemokines network that is specific to the disease of interest. Considering the large size of cytokines/chemokines pool, it would be challenging to perform such systematic analysis using immunobiochemical approaches. Luckily, due to the development of pQTL (protein quantative trait loci), it is possible to directly correlate genetic variants with protein expression

level. Furthermore, the integration of cytokine/chemokine pQTL with IMD GWAS (genome-wide association studies) data can offer valuable insights into the true determinants of IMDs (Burren et al., 2020). This approach enables convenient large-scale analysis of genetic risk factors in the pathogenesis of such diseases.

1.2 Genome-wide association studies

Genome-wide association study (GWAS) is a powerful tool used to identify genotypes associated with phenotypes by testing for differences in the allele frequency of genetic variants. Such analysis is typically carried out within populations that share similar genetic ancestry but differ in the phenotype of interest. Since its first release in 2005, GWAS has successfully identified thousands of robust associations between genetic variants and common diseases (Klein et al., 2005). For the convenience of sharing and storage, GWAS summary statistics is a commonly adapted format to publish association results. Summary statistics provide the aggregate association data for every variant analyzed in GWAS (MacArthur et al., 2021). A typical summary statistics dataset includes the following information: genomic coordinates of the variant; the effect allele that the effect estimate refers to; the reference allele, a genetic alternative found at the same loci and used as a reference; the effect allele frequency, namely the frequency of the effect allele in the control population; the beta value (or log odds-ratio), a key variable that estimates the effect size - ie the expected change in phenotype per copy of the effect allele; the p-value resulting from the hypothesis test for association between the effect allele and the trait; and the standard error, which describes the uncertainty of the beta value and is mainly determined by the sample size and the allele frequency.

In the past decades, GWAS has identified many single nucleotide polymorphisms (SNP) associated with IMDs that affect the circulating levels of cytokines/chemokines (Charles et al., 2018) (Ferkingstad et al., 2021). Meanwhile, comparative analysis of these GWAS studies suggests that a shared genetic architecture of risk factors may exist across different IMDs (Cotsapas & Hafler, 2013). Detailed understanding of

such shared genetic risks can open a window for therapeutic repurposing, and foster the development of better defined standards in patient grouping and diagnosis. However, there are several challenges when attempting to comprehensively compare multiple diseases for their shared genetic architecture. Firstly, GWAS summary statistics data typically consists of millions of dimensions (as each SNP is essentially one dimension), and only a minority of them potentially carry information relating to the disease risk. Secondly, while principal component analysis (PCA) can be used to reduce the dimensions, it would also capture technical differences, such as sampling, among the different GWAS studies. Consequently, genuine genetic signal could be masked by technical noise. Another common problem in applying PCA to summarize genetic architecture is the lack of means to perform standard statistical tests to evaluate the significance of the result. Lastly, as mentioned above, in order to distinguish driver SNPs from passengers and causative relationship from correlation, linkage disequilibrium (LD) must be taken into consideration while analyzing GWAS data.

To address the above problems, a method that combines Bayesian shrinkage and PCA is applied here to construct a basis derived from a total of 40 cytokines/chemokines GWAS datasets (Burren et al., 2020) (Figure 1.1). The basis is essentially a reduced dimension space whose components summarize different genetic patterns associated with underlying biological risks. By projecting multiple external IMD GWAS datasets into the same basis, we can identify novel associations between cytokines/chemokines and IMDs, and gain insights about both the shared and the distinct genetic architectures among the projected diseases. A detailed description of the methodology is given in the following section.

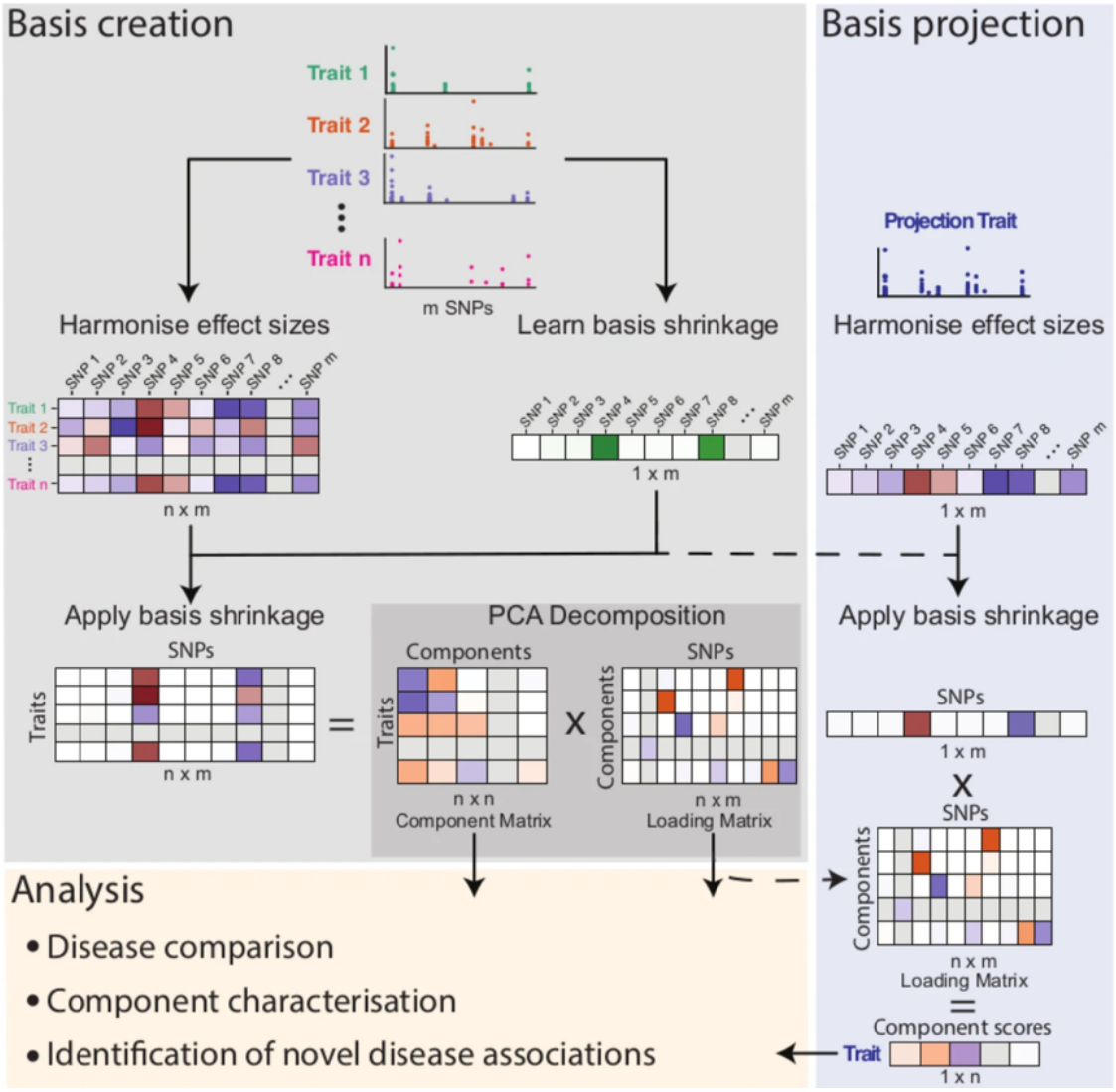

Figure 1.1: Schematic illustrating basis creation and projection. Basis creation: following QC, selected cytokines/chemokines GWAS summary statistics datasets are combined to create a matrix M (n * m) of harmonized effect sizes, where n is the number of adapted traits, m is the total count of sparse SNPs that contribute to at least one principal component (In this study, n=40, m=5519). A learnt shrinkage coefficient is computed for each sparse SNP. I applied basis shrinkage by multiplying M by the shrinkage vector, and used PCA to decompose the shrunk M to the product of a component matrix (n * n) and a loading matrix (n * m). Basis projection: independent external datasets are harmonized with respect to the basis, applied with shrinkage, and multiplied by the loading matrix to obtain component scores (delta). These component scores can be used to test the hypothesis that a weighted average of effect sizes of the projected trait is non-zero. This figure is adapted from https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00797-4 [@burren2020genetic]

2 Method

2.1 Input datasets and QC

A total of 104 cytokines/chemokines GWAS summary statistic datasets were downloaded from https://www.decode.com/summarydata/ (Ferkingstad et al., 2021). Non-specific binding of the aptamers used in this SomaScan dataset to quantify cytokines is a recognised source of false positives (Ferkingstad et al., 2022). Because most proteins under genetic control will include a cis effect, I used the presence of a cis effect as a quality control measure. Prior to basis construction, the traits were scanned for cis QTL signal, i.e. a peak with p value $< 10^{-8}$ located within 100 kb up/downstream of the locus of their target genes. After this quality control screening, the resulting 40 traits were used to construct the basis. Only SNPs that were present in all these 40 traits were then picked, and aligned to the 1000 Genomes reference genotype panel for later analysis.

2.2 Construction of cytokine/chemokine basis

2.2.1 Overview

We first simplify the model by assuming that 1) a maximum of only one SNP is causal for each trait at each recombination hotspot-defined block (LD block).

(Maller et al., 2012) (Wakefield, 2009) 2) the causal SNPs are included in our GWAS datasets. Then, for each input GWAS dataset, we use a Bayesian fine-mapping method to calculate a posterior probability for each SNP to be causal within the same LD block. At each SNP, we create an overall weight (w) by computing the weighted average of posterior probabilities across our input studies, i.e. the traits that we used to construct the basis. By using these probabilities as weights before computing our PCA, we can effectively prevent double counting of SNPs in high LD, while shrinking (possibly noisy) effect estimates of truly unassociated SNPs towards 0, and thus avoid capturing random technical noise in the components. This procedure was demonstrated to allow independent datasets that were projected into the basis space to sit close to the input datasets of the same traits (Burren et al., 2020). Without weighting, the independent datasets did not sit close to their input counterparts, suggesting that PCA with weighting captures much more of the true biological signal.

SNP effect estimates ($\hat{\beta}$) are then weighted by the above computed w and adjusted for the variance due to MAF (σ^2_{MAF}). It is noteworthy to mention that here I do not adjust for the variance of caused by sample size, as this would overly shrink input studies that are relatively small in size. The matrix of the shrunk effect size is therefore computed as :$\hat{r} = \frac{w\hat{\beta}}{\sigma_{MAF}}$. An additional control trait was also combined with the shrunk matrix to serve as a baseline for subsequent analysis. In this control trait, all effect sizes were set to zero. After mean centering this combined matrix $\hat{M}$ $(41 * 641079)$ by column, we perform a PCA with R function prcomp() to generate our cytokine/chemokine basis. In order to assess the maximal subset of informative components, I made a scree plot which revealed that the final 41st component could be safely discarded (Figure B.1).

After PCA, I obtained a $641079 * 40$ loading matrix as my complete basis. However, most entries in this matrix were very close to 0. To highlight the driver SNPs that are true contributors to each component and for computational efficiency, I performed a by-PC search to find the minimum quantile of SNPs that could be retained while maintaining a correlation value above 0.999 with the complete

matrix. SNPs that did not abide by this criteria were then set to 0. In the end, I found 5519 unique SNPs that were non-zero in at least one PC, and this downsized $5519 * 40$ loading matrix constituted my final sparse basis.

2.2.2 Mathematical exposition

GWAS summary statistics datasets provide us with the regression coefficient $\hat{\beta}_{ti}$ (i.e. the aforementioned effect size), and its corresponding standard error σ^2_{ti}, where $t \in \{1, ...T\}$ indexes traits and $i \in \{1, ...p\}$ indexes SNPs. In step 1.1, I only kept SNPs with the complete data across all traits, therefore there is no missing value for any t and i. To generate trait-specific weights for each SNP, we then compute the Bayes factor BF_{ti} as (Wakefield, 2009) :

$$BF_{ti} = \frac{P(\hat{\beta}_{ti}|\hat{\beta}_{ti} \sim N(0, \sigma^2_{ti} + (W * \sigma_t)^2))}{P(\hat{\beta}_{ti}|\hat{\beta}_{ti} \sim N(0, \sigma^2_{ti}))}$$

where σ^2_t is the population variance of the trait, and W denotes a scalar representing the standard deviation of a prior on the true effect. We set W to the default value of 0.15, corresponding to a prior belief that true β exceeds $0.15 * \sigma_t$ with a probability of about 5%.

Let $R_r \in \{R_1, ...R_R\}$ denote R non-overlapping sets of SNPs, and each region R_r is located at one unique LD block. Based on the assumption that a maximum of only one SNP is causal for each trait at one LD block, and that this causal SNP (if it exists) is already included in the input dataset, we can compute the posterior probability for each SNP $i \in R_r$ to be causal as :

$$pp_{ti} = \frac{\pi BF_i}{(1 - m_r\pi) + \sum_{i \in R_r} \pi BF_i}$$

where m_r is the total number of SNPs in R_r , and π is a scalar set to be 10^{-4} , representing a prior belief that the frequency of a causal SNP across the genome is 1 in 10000 (Giambartolomei et al., 2014). We can then estimate the probability that region R_r contains a causal SNP by marginalizing over all SNPs in that region :

$$v_{tr} = \sum_{i \in R_r} pp_{ti}$$

The final SNP weight is therefore a weighted average of pp_{ti} over all input traits:

$$w_i = \frac{\sum_t pp_{ti} v_{tr(i)}}{\sum_t v_{tr(i)}}$$

Note that each SNP has a different variance σ_i^2 contributed by the minor allele frequency f_i :

$$\sigma_i = \frac{1}{\sqrt{2f_i(1-f_i)}}$$

By weighting the effect size at each SNP with w_i and adjusting them with the population variance, we can get the $T * P$ matrix of the shrunk effect size :

$$(\hat{\gamma_{ti}}) = (\frac{w_i \beta_{ti}}{\sigma_i})$$

An additional row of zero was combined with this matrix as a baseline reference. We then center this $(T+1) * P$ matrix and perform PCA decomposition :

$$\boldsymbol{G}^C = (\hat{\gamma_{ti}}^c) = (\hat{\gamma_{ti}} - \frac{1}{T+1} \sum_t \hat{\gamma_{ti}}) = (\hat{\gamma_{ti}} - C_i)$$

$$\boldsymbol{P} = \boldsymbol{G}^C \boldsymbol{Q}$$

where Q is our loading matrix whose columns represent the first T eigenvectors of $(G^C)'G^C$.

2.3 Projection of the cytokine/chemokine basis

2.3.1 Overview

A total of 10541 external GWAS summary statistic datasets of different classes were projected to this cytokine/chemokine basis, including IMD, BMK (biomarker), PSD (psychiatric disorder), INF (infectious disease) etc. Prior to projection, effect

alleles were aligned to the basis SNPs and shrunk by multiplying $\frac{w}{\sigma_{MAF}}$. These shrunk vectors of the external traits were then projected to the basis by multiplying with the loading matrix Q. The projected result is denoted as $\hat{\delta}$, which estimates the difference between the projected $\hat{\beta}$ and control. The variance of $\hat{\delta}$ is computed as described in 2.3.2.

Analogous to standard GWAS hypothesis tests, where $\hat{\beta}$=0 serves as the null hypothesis, we can test if the vector $\hat{\delta} = 0$ across all 40 component with a chi-square test. Multiple testing correction was applied within each class of traits (e.g. IMD) using the Benjamini-Hochberg approach, and association was considered significant with an overall FDR < 0.01. For each component, similarly, we calculate FDR individually and traits are only considered significant on that PC with component FDR < 0.05 and overall FDR < 0.01.

2.3.2 Mathematical exposition

Let $\hat{\beta}_0 = (\hat{\beta_{0i}}, i = 0, ..., p)$ denote the regression coefficient (the effect size) of a test trait across the same set of SNPs. $\hat{\beta_{0i}}$ is set to zero for any SNP which is absent in the test trait, yet listed in our basis. Only traits with a coverage of $>$ 80% of the basis SNPs will be processed as meaningful projection. Each $\hat{\beta_{0i}}$ also comes with an estimate variance $\hat{v_{0i}}$. The trait is then projected to our basis with the following function:

$$\boldsymbol{P}_0 = (\boldsymbol{D}\hat{\boldsymbol{\beta}}_{\mathbf{0}} - \boldsymbol{C})'\boldsymbol{Q}$$

where D is a $p*p$ diagonal matrix with entries $\frac{w_i}{\sigma_i}$ for each SNP, Q is the loading matrix, $C = (C_i, i = 1, ..., p)$ is the vector of basis centers from 2.2. To calculate the variance of $\hat{\beta}_0$ and the projection, we define Σ as the correlation matrix of SNPs in $\hat{\beta}_0$ (Burren et al., 2014) and V_0 as the diagonal matrix of V_{0i} :

$$var(\hat{\boldsymbol{\beta}}_{\mathbf{0}}) = \boldsymbol{V_0}\boldsymbol{\Sigma}\boldsymbol{V_0}$$

$$var\boldsymbol{P_0} = \boldsymbol{Q'DV_0\Sigma V_0DQ}$$

As $X_0 = -C'Q$ represents the projection result of the control trait (the aforementioned vector of 0), we take X_0 as the null location in our space, and the difference between the test trait and control in the projected space is therefore $\hat{\delta} = P_0 - X_0$. We then test the null hypotheses of $\delta = 0$ across all components with $X^2_{overall}$ defined as below:

$$X^2_{overall} = \hat{\delta}' var(\boldsymbol{P_0})^{-1}\hat{\delta} \sim \chi^2_T$$

Using a similar strategy, we can also test if $\delta = 0$ at component j :

$$X^2_j = \hat{\delta}[j]^2 / var(\boldsymbol{P_0})[j,j] \sim \chi^2_1$$

3 Results

3.1 A genetic basis of cytokine/chemokine

Using the method described above, I constructed a genetic basis that includes 40 GWAS summary statistics datasets of cytokines/chemokines, as listed in Table 3.1. These 40 traits cover major cytokine and chemokine classes including: interleukins, TNFs, hematopoietins and C/CC/CXC/CX3C chemokines. They are pQTL datasets that reflect the sequence determinants for cytokine/chemokine levels found in the serum. All datasets were collected from 35,559 Icelanders and generated with SomaScan version 4 (Ferkingstad et al., 2021). After dimension reduction and sparse SNP quantile search, I generated a 40-dimension space that comprises the genetic information of 5519 SNPs.

To test the validity of our basis and to investigate the associations between cytokines/chemokines and IMDs, I projected a total of 5610 external GWAS datasets, which include classes of: IMD, BMK (biomarker), PSD (psychiatric disorder), INF (infectious disease), CAN (cancer) and OTH (others). Multi-testing correction was performed by trait class to keep significant projections with FDR < 0.01. Interestingly, after filtering, only the proportions of IMD and BMK traits increased, while projections of other groups were vastly filtered out, indicating a specificity of this basis in enriching immune related signals (Figure 3.1). In order to test if

Table 3.1: Cytokines and chemokines used in basis construction.

Name	Amino Acids	Location	Molecular Weight(Da)	Receptor	Uniprot	Class
IL-1RA	177	2q14.2	20055	CD121a	P18510	Interleukins
IL-9	144	5q31.1	15909	IL-9R, CD132	P15248	Interleukins
IL-16	631	15q24	66694, homotetramer	CD4	Q14005	Interleukins
CD154	261	Xq26	29273, homotrimer	CD40	P29965	TNF
CD153	234	9q33	26017, trimer?	CD30	P32971	TNF
LIGHT	240	16p11.2	26351, trimer?	LTbR, HVEM	O43557	TNF
TALL-1	285	13q32-q34	31222, trimer?	BCMA, TACI	Q9Y275	TNF
TWEAK	249	17p13.3	27216, trimer?	Apo3	O43508	TNF
Epo	193	7q21	21306	EpoR	P01588	Miscellaneous hematopoietins
Tpo	353	3q26.3-q27	37822	TpoR	P40225	Miscellaneous hematopoietins
MSP	711	3p21	80379	CDw136	P26927	Miscellaneous hematopoietins
XCL2	114	1q23	12567	XCR1	Q9UBD3	C Chemokines
CCL3	92	17q11-q21	10085	CCR1, CCR5	P10147	CC Chemokines
CCL5	91	17q11.2-q12	9990	CCR1, CCR3, CCR5	P13501	CC Chemokines
CCL7	99	17q11.2-q12	11200	CCR1, CCR2, CCR3	P80098	CC Chemokines
CCL8	99	17q11.2	11246	CCR3	P80075	CC Chemokines
CCL11	97	17q21.1-q21.2	10732	CCR3	P51671	CC Chemokines
CCL14	93	17q11.2	10678	CCR1	Q16627	CC Chemokines
CCL15	113	17q11.2	12248	CCR1, CCR3	Q16663	CC Chemokines
CCL16	120	17q11.2	13600	CCR1	O15467	CC Chemokines
CCL17	94	16q13	10507	CCR4	Q92583	CC Chemokines
CCL18	89	17q11.2	9849	?	P55774	CC Chemokines
CCL21	134	9p13	14646	CCR7	O00585	CC Chemokines
CCL22	93	16q13	10580	CCR4	O00626	CC Chemokines
CCL23	120	17q11.2	13443	CCR1	P55773	CC Chemokines
CCL25	150	19p13.2	16639	CCR9	O15444	CC Chemokines
CXCL1	107	4q21	11301	CXCR1, CXCR2	P09341	CXC Chemokines
CXCL4	101	4q12-q13	10845	?	P02776	CXC Chemokines
CXCL5	114	4q13-q21	11972	CXCR2	P42830	CXC Chemokines
CXCL6	114	4q21	11897	CXCR1, CXCR2	P80162	CXC Chemokines
CXCL8	99	4q12-13	11098	CXCR1, CXCR2	P10145	CXC Chemokines
CXCL9	125	4q21	14019	CXCR3	Q07325	CXC Chemokines
CXCL10	98	4q21	10856	CXCR3	P02778	CXC Chemokines
CXCL11	94	4q21.2	10365	CXCR3	O14625	CXC Chemokines
CXCL12	93	10q11.1	10666	CXCR4	P48061	CXC Chemokines
CXCL13	109	4q21	12664	CXCR5	O43927	CXC Chemokines
CX3CL1	397	16q13	42202	CX3CR1	P78423	CX3C Chemokines

the genetic patterns captured by this basis truly reflect biological differences (as opposed to technical noise that varies between studies), I summarized the projection results of external cytokines/chemokines datasets from different resources in a heatmap (Figure 3.2). Notably, the majority of cytokines/chemokines are clustered together closely with their GWAS comparators rather than other proteins from the same study. For example, despite the huge difference in sample size, CCL11 from three independent studies clustered as nearest neighbors. Although a few cytokines/chemokines are clustered apart, such as IL-1RA, on the PCs where different IL-1RA studies all projected as significant, the directions of the delta values were still consistent. In general, cytokines/chemokines of my basis showed good consistency with respect to external studies, and I used this heatmap as a reference

of replicable signals of cytokines/chemokines in interpreting the projections.

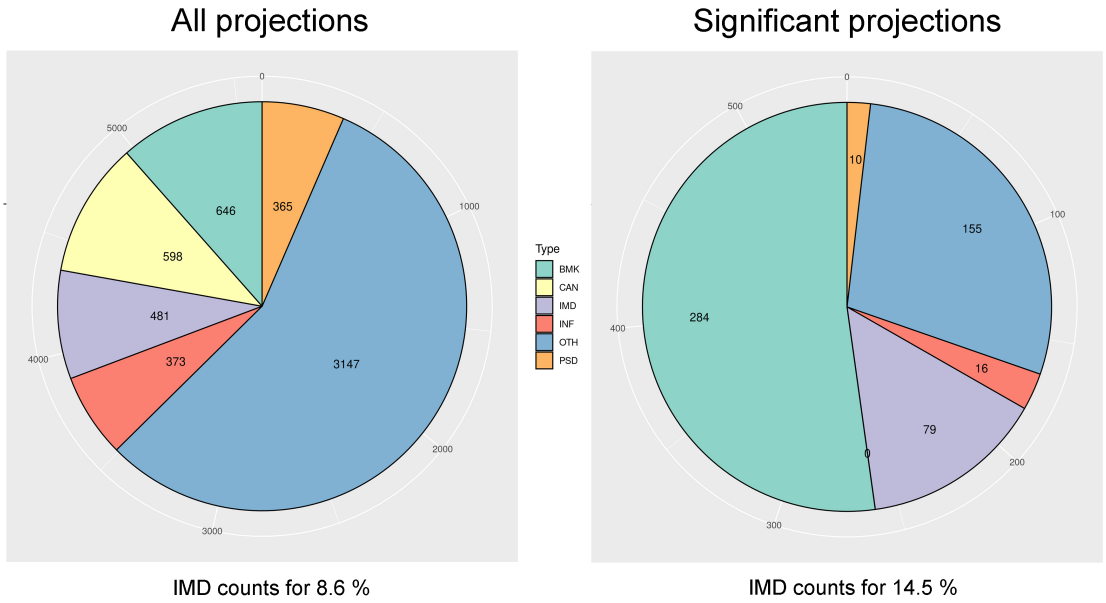

Figure 3.1: Count of projected datasets from different classes before and after filtering with FDR < 0.01. Left: All 5610 successfully projected traits. Right: After FDR adjustment, only 544 projections were considered significant. Note that the proportions of IMD and BMK traits increased after the significance test, while the size of other classes (i.e. CAN, PSD, INF, OTH) either decreased or reduced to null.

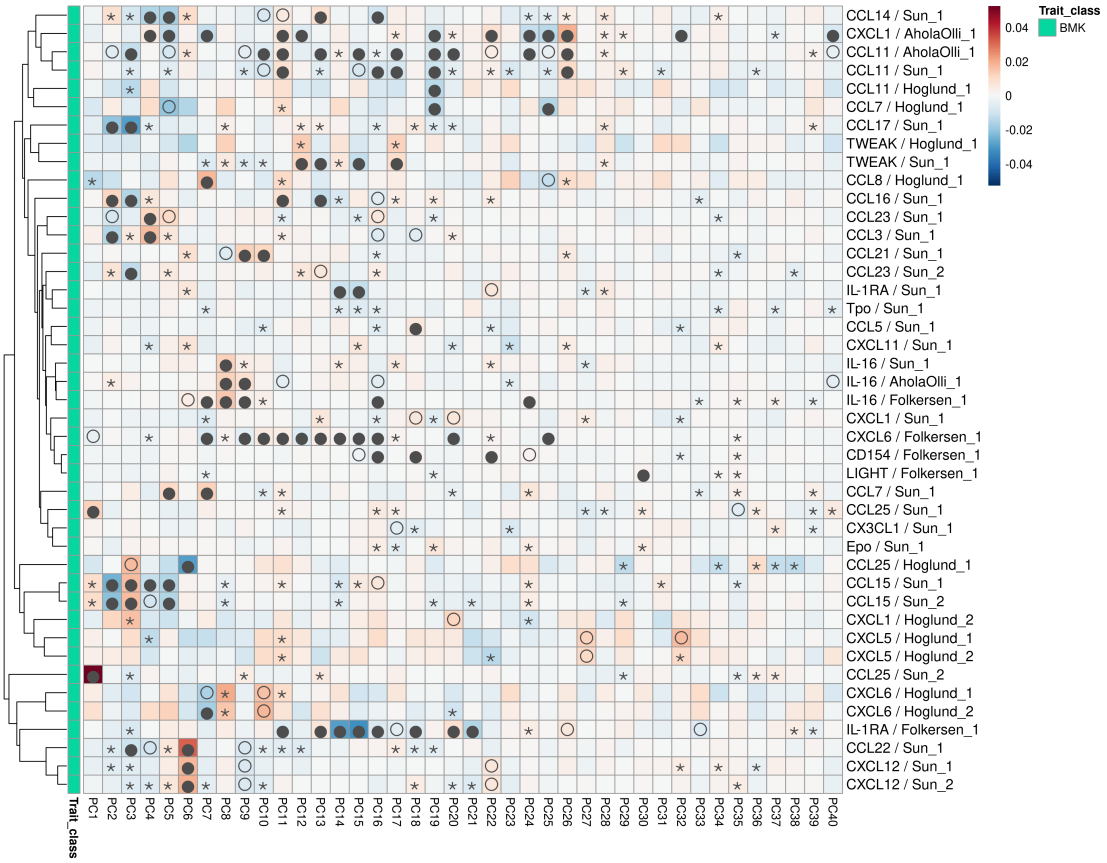

Figure 3.2: Heatmap illustrating the replicability of projected cytokines/chemokines from external studies. Solid circles and hollow circles highlight traits with adjusted P values at < 0.01 and <0.05 on that PC respectively. The asterisk symbol indicates projections with the pre-adjusted P value < 0.05. Only projections marked by solid or hollow circles were considered significant for the corresponding PC.

3.2 Interpretation of IMD projection results by components

To illustrate the associations between cytokines/chemokines and IMDs, I plotted significant IMDs together with 40 cytokines/chemokines that were used for basis construction on each PC. It is noteworthy that PC 7, 9 and 10 have the largest number of significant IMD projections, so I decided to focus on these PC for detailed investigation. For each of these PCs, I examined whether there is support in the literature for relationships between the diseases and cytokines jointly associated.

3.2.1 PC 7

In PC 7, CXCL1, CXCL6, CCL8, CCL7, IL-16 and CXCL12 generated replicable extreme signals, i.e. signals of these proteins are replicated in significant projections of external cytokine/chemokine datasets. A variety of IMDs projected with significant delta values on PC 7 (Figure 3.3). For IMDs that landed on the same side with aforementioned proteins, such as T1D and CXCL12, this projection result suggests that increased levels of corresponding cytokine/chemokines are positively associated with increased risk for such diseases. Indeed, extensive evidence from previous studies support my proposed associations. In non-obese diabetic (NOD) mice, reduction in CXCL12 level was found to cause significant delays in the onset of T1D, while overexpression of CXCL12 is associated with alterations in T cell trafficking and T1D development (Leng et al., 2008). Similarly, positive correlation with T1D has also been reported for CXCL1. In adult T1D patients, circulating levels of CXCL1 were observed to be significantly increased compared to both healthy control group and type 2 diabetes group (Takahashi et al., 2011). The potential pathogenic role of CXCL12 in vitiligo has been suggested by two independent studies. Overexpression of exogenous CCL12 in mice leads to the recruitment of leukocytes to skin and an induction of vitiligo symptoms. Significant elevations in CXCL12 serum levels were also observed in vitiligo patients (Rezk et al., 2017) (Gharib et al., 2021). In addition, a number of studies have highlighted associations between: CXCL6 and both T1D and Crohn's disease; CXCL1 and systemic lupus erythematosus; CXCL12 and sarcoidosis. (Antoniou et al., 2009) (Zeng et al., 2021) (Gijsbers et al., 2004) Reassuringly, same traits derived from independent studies (for example, T1D from both Chiou (Chiou et al., 2021) and FinnGen R7) generated similar projection results, which emphasizes the reproducibility of my model.

While not specifically mentioned in the literature, many of the associations I found on PC 7 are in line with the biomedical properties of the cytokines/chemokines analyzed. For example, CXCL6 has been reported to be selectively expressed in inflamed intestinal tissues and potentially involved in the pathogenesis of Crohn's dis-

ease (Gijsbers et al., 2004). Considering that celiac disease is also an autoimmune disorder that targets the intestine, it is reasonable to see that CXCL6 was projected closely with celiac disease suggesting a strong genetic link may exist. Altogether, an extensive number of well evidenced associations support the credibility of PC 7. Moreover, my study also proposes novel and likely biologically relevant associations (for example, between CXCL6/CXCL1 and Celiac disease/vitiligo/autoimmune thyroid disease) that are yet to be characterized experimentally.

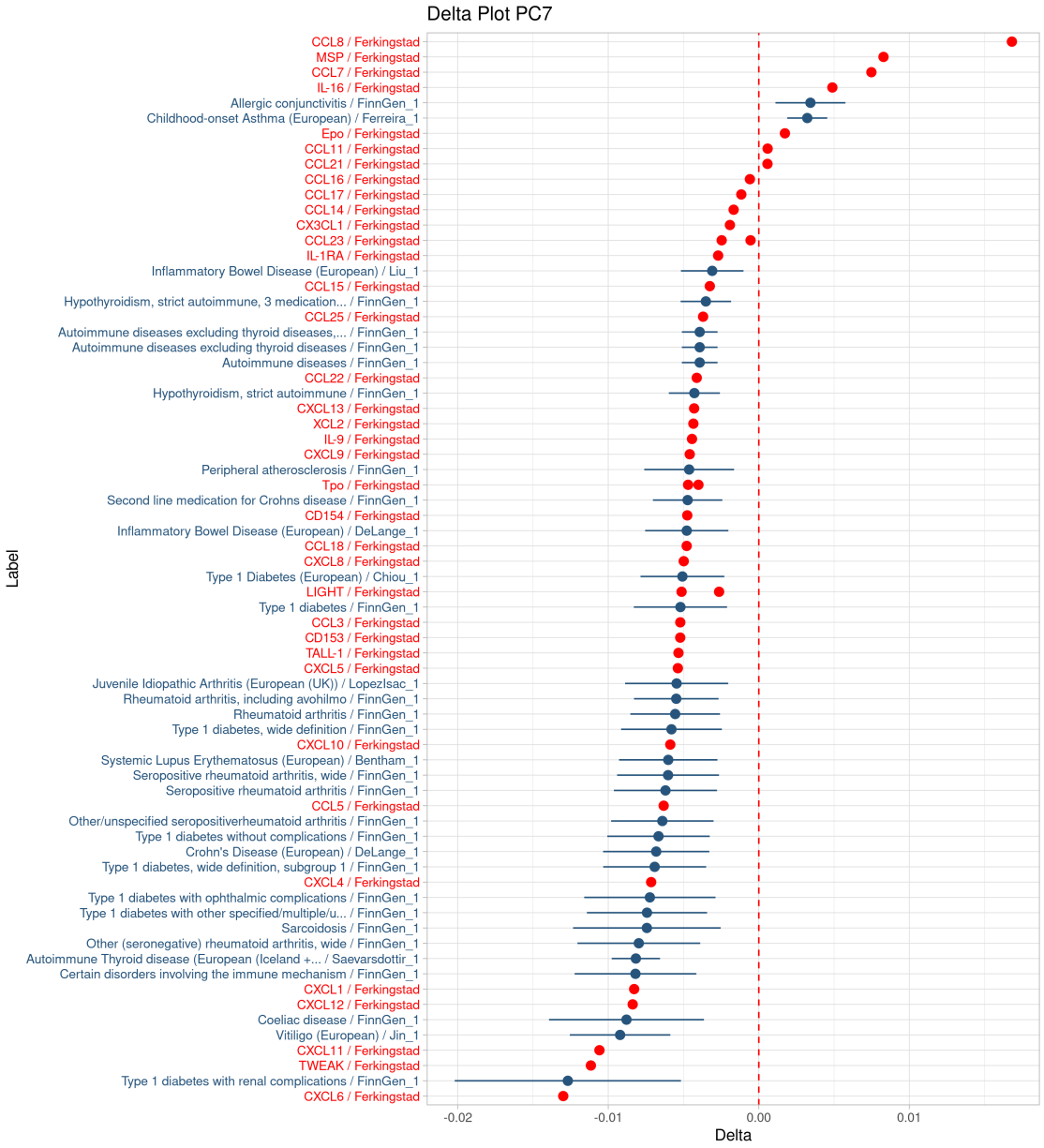

Figure 3.3: Forest plot illustrating significant projections of IMD traits on PC7. Ferkingstad datasets used in basis construction are colored in red as reference. Projections with adjusted P < 0.05 on PC 7 are shown and lines indicate 95% confidence interval.

3.2.2 PC 9

Applying the same strategy to PC 9, CXCL6, CCL22, CXCL12, CCL11, IL-16 and CCL21 were found to generate similar delta values in external cytokine/chemokine traits as in the training datasets (Figure 3.4). Amongst them, CCL21 appears to be projected with the strongest positive signal, which indicates its association

with an elevated genetic risk of juvenile idiopathic arthritis (JIA), coeliac disease, type 1 diabetes, rheumatoid arthritis (RA), autoimmune thyroid disease etc. These findings are extensively supported by previous studies: CCL21 has been found to be highly expressed in the mucosal venule endothelium of coeliac patients (Capitano et al., 2021). In NOD mice, CCL21 levels show an age-dependent increase in the pancreas, and this cytokine has been suggested to be involved in the functional change of lymphatic endothelial cells during T1D development (Qu et al., 2005). In addition to CCL21, association between IL-16 and T1D has also been reported in NOD mice. An elevated level of IL-16 in the islets highly correlates with the progression of insulitic lesion, and inhibition of IL-16 with an antibody has been shown to offer significant protection against T1D development (Meagher et al., 2010).

JIA and RA are both autoimmune diseases that cause joint inflammation and stiffness. Previous studies have suggested that while JIA shares some clinical and pathological features with RA, the genetic factors that contribute to the onset of JIA can be different than those responsible for RA (Prahalad & Glass, 2002). For example, both diseases associate with HLAs (human leukocyte antigens). However, while HLA-B27 has been consistently identified as a major risk contributor specific to pauciarticular JIA, it is HLA-DR4 that was considered play a causative role in the pathogenesis of RA (Pryhuber et al., 1996) (Fugger & Svejgaard, 2000). Association between CCL21 and RA has already been supported by multiple independent studies: CCL21 expression level is known to be highly increased in the synovial fluid of RA patients, leading to the recruitment of monocytes into the inflamed joints (Van Raemdonck et al., 2020). CCL21 has also been found to induce the production of VEGF (vascular endothelial growth factor) by RA ST fibroblasts and IL-8 by macrophages respectively, presumably through the CCL21-CCR7 axis (Pickens et al., 2011). In contrast, the relationship between CCL21 and JIA has not been reported before. Given that JIA traits were projected as the closest neighbor to CCL21 on PC 9, with an even more extreme signal than RA, there is potentially a more specific association between CCL21 levels and the risk of JIA.

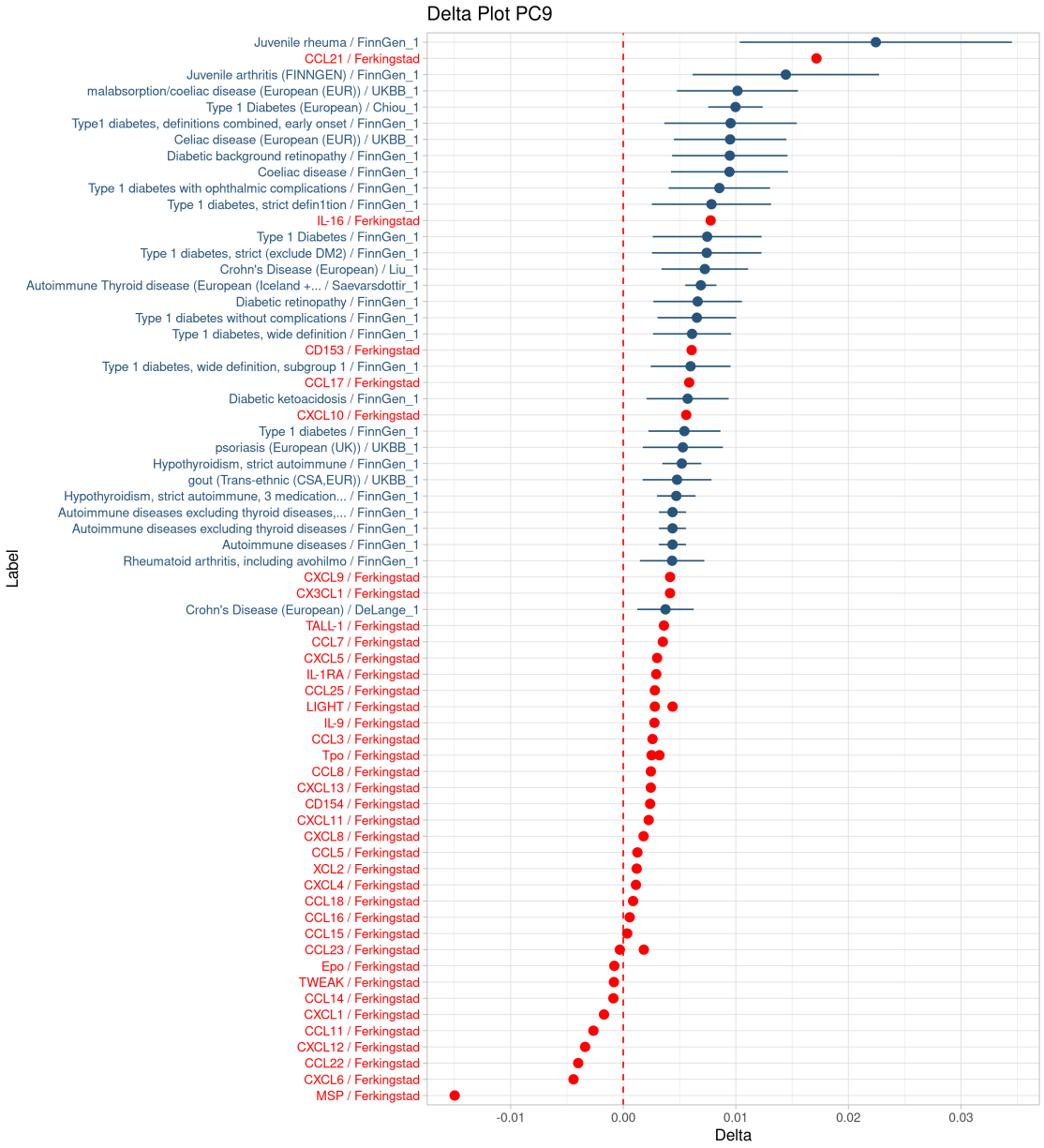

Figure 3.4: Forest plot illustrating significant projections of IMD traits on PC9. Projections with adjusted P < 0.05 on PC 9 are shown and lines indicate 95% confidence interval.

3.2.3 PC 10

In PC 10, chemokines CCL21, CXCL6 and CCL14 showed the most prominent replicable signals (Figure 3.5). In a similar fashion to PC 9, CCL21 positively correlated with T1D and coeliac disease in PC 10. In addition, PC 10 also suggested CCL21 associations with vitiligo and autoimmune thyroid disease (AITD), both of

which have been evidenced in previous studies. Specifically, deficiencies in CCL21-dependent homing of Treg (regulatory T) cells to the skin have been shown to correlate with the onset of vitiligo (Tembhre et al., 2015). CCL21 has also been reported to be expressed in the thyroid of patients with AITD. In transgenic mice models that mimic AITD, CCL21 alone is sufficient to induce lymphocyte recruitment to the thyroid through CCR7 (Martin et al., 2004). Besides CCL21, experimental evidence also support CXCL6 associations to both T1D and psoriasis. Expression levels of CXCL6 are reported to be significantly increased in patients with diabetic nephropathy (Sun et al., 2019). Notably, this CXCL6-T1D association was also highlighted in PC 7. In relation to the pathogenesis of psoriasis, CXCL6 is thought to be involved in the recruitment of myeloid dendritic cells, neutrophils and Th17 cells to the lesion sites, presumably under the regulation of IL-17 (Girolomoni et al., 2012).

To summarize, in this study I constructed a 40-dimension cytokine/chemokine basis where each principal component represents a genetic pattern summarized from our training datasets. The replicability of this basis was confirmed by the consistency of the significant projection results from external cytokine/chemokine traits. By projecting IMD test datasets into this basis, I identified extensive associations between cytokines/chemokines and IMDs. Importantly, most of these associations, especially where pairs showed extreme delta values, are supported by direct or circumstantial evidence from experimental studies. Therefore, given the robustness of our projection results, it is likely that even those associations that are not yet verified experimentally, do represent genuine, biologically significant risk factors.

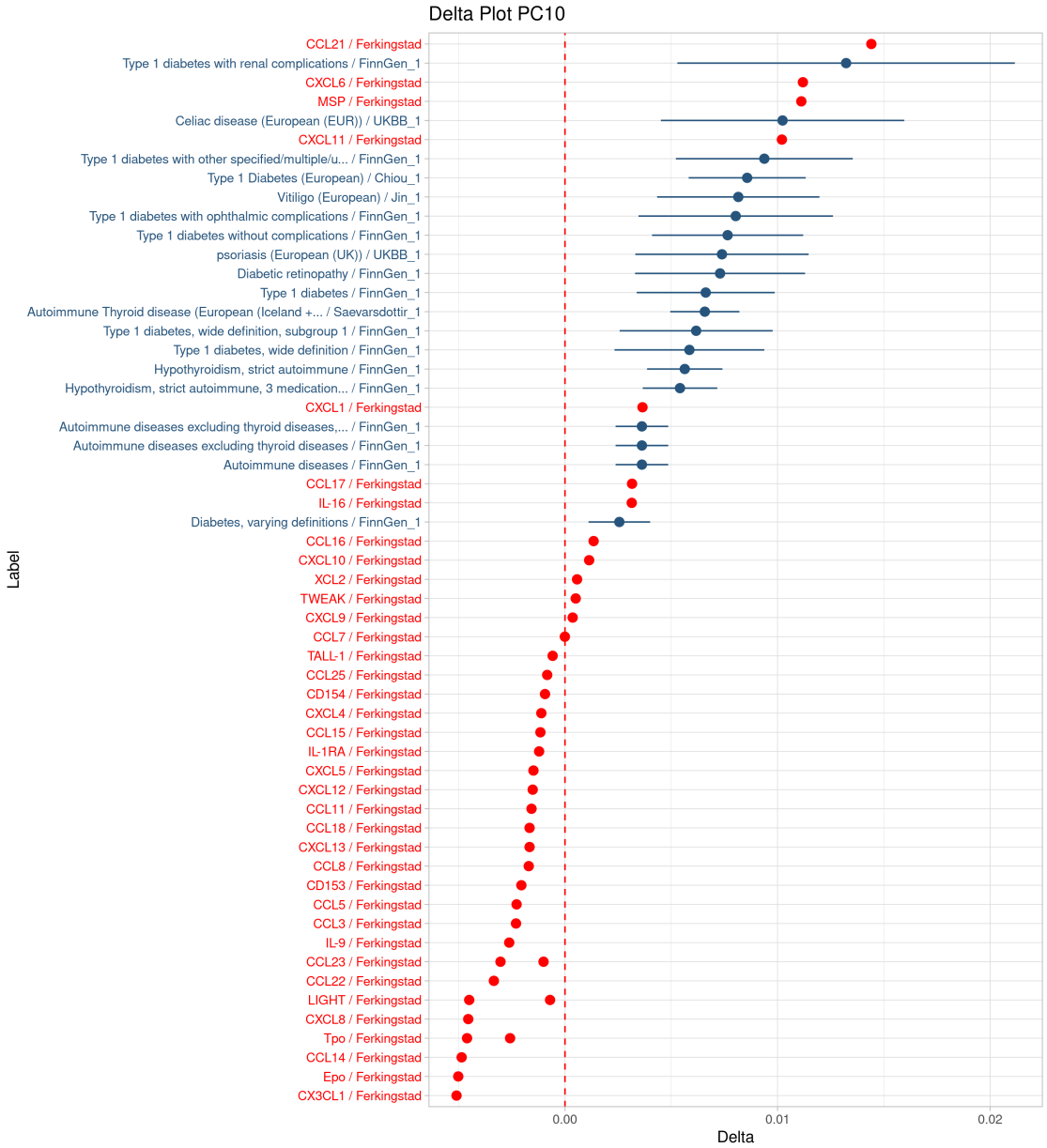

Figure 3.5: Forest plot illustrating significant projections of IMD traits on PC10. Projections with adjusted $P < 0.05$ on PC 10 are shown and lines indicate 95% confidence interval.

4 Discussion

In general, compared to negative correlations, it is relatively easier to find experimental evidence in support of a positive correlation between projected cytokine/chemokine and IMD. This is probably because most of the training datasets that are used in this basis construction are pro-inflammatory cytokines/chemokines. And the protective roles of cytokines/chemokines in IMD pathogenesis are relatively rare. That said, in this study, an experimentally verified negative association was identified in PC15. Specifically, as shown in the forest plot (Figure B.3), increased levels of cytokine IL-1RA strongly correlated with a decrease in sarcoidosis risk. Indeed, this correlation is in agreement with evidence from multiple studies: As an endogenous antagonist of IL-1, IL-1RA competitively binds to IL-1 receptor and inhibits the effect of IL-1. In patients with sarcoidosis, the expression level of IL-1 is known to be elevated, while IL-1RA levels are significantly reduced (Mikuniya et al., 2000). Moreover, a higher IL-1RA/IL-1 ratio has been shown to be associated with a better prognosis in pulmonary sarcoidosis (Hutyrová et al., 2002). It is therefore worth mentioning that, although the main focus of this report is on PC 7, 9 and 10, associations identified in other PCs, such as PC 15, are also of considerable interest and warrant further investigation.

It is noteworthy that in each PC, the majority of supportive evidence often tend to converge on proteins/diseases that display a more extreme delta value, i.e. they are found at the two ends of the forest plot. This is because generally the stronger the association, the more easily it is identifiable by different studies. For the same reason, although all protein/disease pairs found on the same PC plot can be considered as potential associations, this study mainly focuses on associations of extreme pairs, as they are more likely to be genuine novel associations. For example, in PC7, the novel associations between CXCL6/CXCL1 and celiac disease/vitiligo/autoimmune thyroid disease would be interesting to explore.

In this project I used pQTL datasets generated from SomaScan version 4 platform to construct the basis. A recent comparative analysis on the performance of SomaScan and Olink has shown that SomaScan outputs have a higher chance to produce false positive signals -i.e. aptamers could non-specifically bind to non-target proteins, leading to an inaccurate evaluation of target protein levels (Ferkingstad et al., 2022). Such false positive signals were found to undermine the specificity of the basis that was built. Therefore, a quality control step was introduced to discard traits that did not display the expected cis QTL signal nearby/within the locus of their target genes. Both replicability and specificity of the resulting basis were indeed vastly improved (data not shown). However, this also led to a considerable reduction in the number of input datasets included, which inevitably narrowed the scope of our cytokines/chemokine basis. Furthermore, when the same protein traits derived from different resources showed disagreement in the projections, it remained ambiguous whether such signal truly could not be replicated, or simply there were inconsistencies between the phenotyping methods used. Nevertheless, despite these limitations, it should be noted that most of the cytokines/chemokines were still able to replicate, indicating an overall strong reliability of these replicable cytokines/chemokines.

In future work, GO enrichment analysis should be performed to understand whether the genetic pattern behind each PC is characterized by any specific cellular function. Also, although the discussion here mainly concerns IMD associations,

projections of other classes, such as non-cytokine biomarkers, are also of relevance as they can aid in the interpretation of biological information captured by each PC. Finally, it is essentially differentially weighted SNPs that contribute to the PC. Therefore, it would be interesting to analyse the top SNPs at each dimension, and isolate SNPs that are specifically located at gene regulatory sequences or protein encoding regions. If such SNPs carry information that could explain its effect on highlighted cytokines/chemokines or IMDs, we would be able to gain more insights into the pathogenic roles of these proteins in IMDs.

Appendices

A

Code availability

All R scripts used in this study are publicly available in my Github repository . Please see README.md for a detailed guide.

B

Supplementary figures

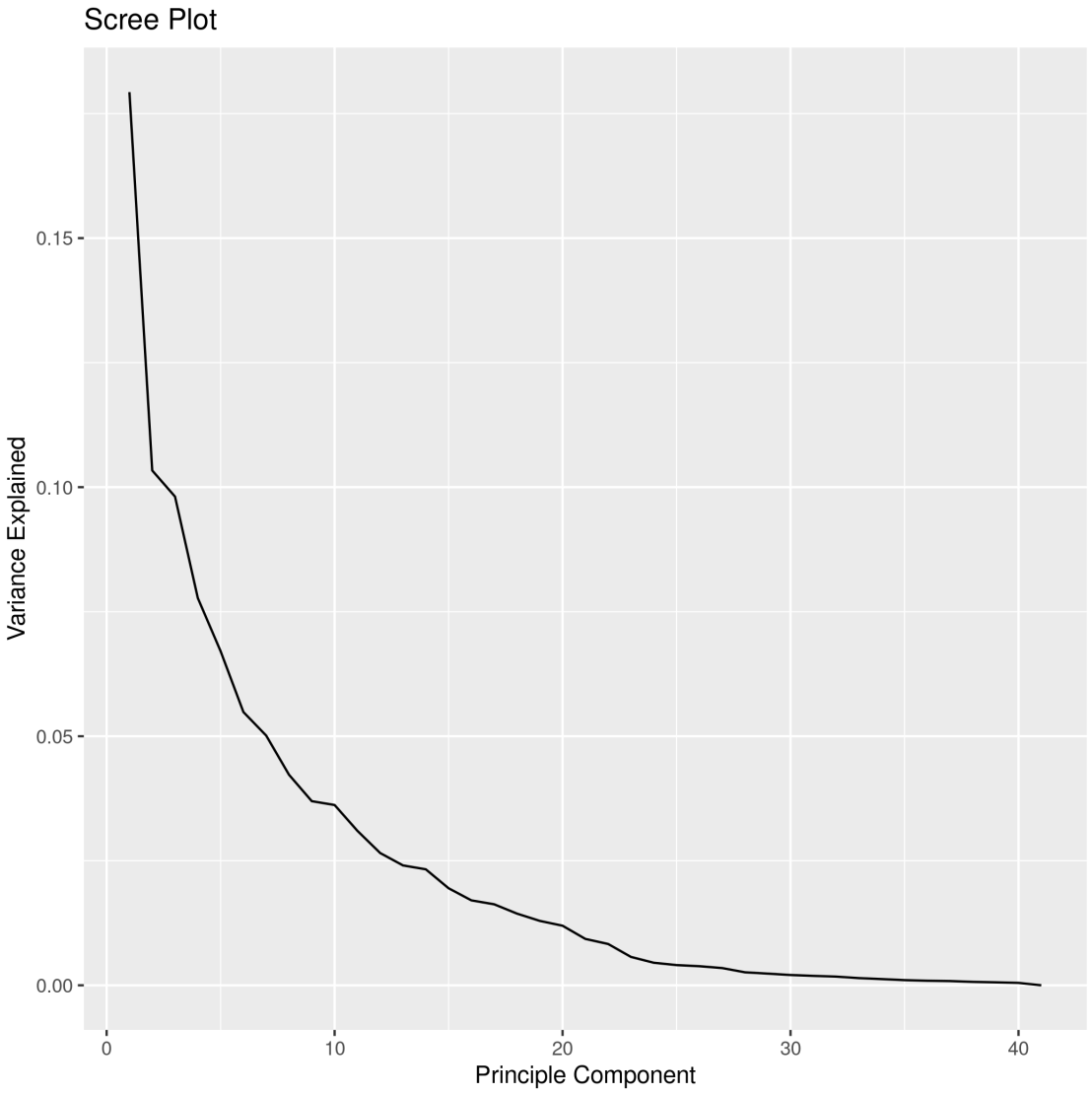

Figure B.1: Scree plot of constructed basis illustrating variance explained by PC1-41

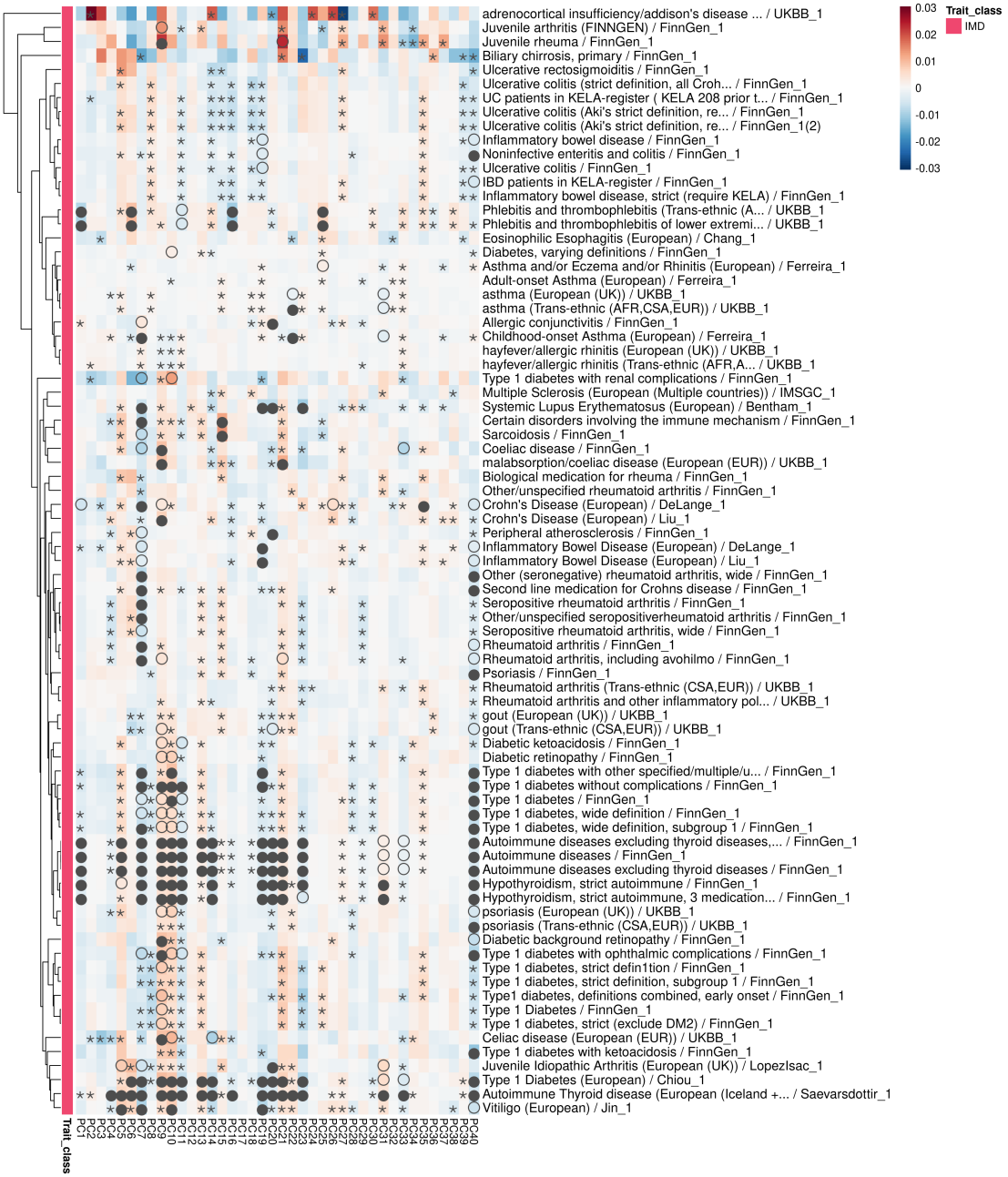

Figure B.2: Heatmap of significant IMD projections. Solid circles and hollow circles highlight traits with adjusted P values at < 0.01 and <0.05 on that PC respectively. The asterisk symbol indicates projections with the pre-adjusted P value < 0.05. Only projections marked by solid or hollow circles were considered significant for the corresponding PC. Note that the same type of diseases from different resources are clustered together.

Figure B.3: Forest plot illustrating significant projections of IMD traits on PC15. Projections with adjusted P < 0.05 on PC 15 are shown and lines indicate 95% confidence interval.

References

Alagpulinsa, D. A., Cao, J. J., Sobell, D., & Poznansky, M. C. (2019). Harnessing CXCL12 signaling to protect and preserve functional β-cell mass and for cell replacement in type 1 diabetes. *Pharmacology & Therapeutics*, *193*, 63–74.

Antoniou, K. M., Soufla, G., Proklou, A., Margaritopoulos, G., Choulaki, C., Lymbouridou, R., Samara, K. D., Spandidos, D. A., & Siafakas, N. M. (2009). Different activity of the biological axis VEGF-flt-1 (fms-like tyrosine kinase 1) and CXC chemokines between pulmonary sarcoidosis and idiopathic pulmonary fibrosis: A bronchoalveolar lavage study. *Clinical and Developmental Immunology*, *2009*.

Burren, O. S., Guo, H., & 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*(23), 3342–3348.

Burren, O. S., Reales, G., Wong, L., Bowes, J., Lee, J. C., Barton, A., Lyons, P. A., Smith, K. G., Thomson, W., Kirk, P. D.others. (2020). Genetic feature engineering enables characterisation of shared risk factors in immune-mediated diseases. *Genome Medicine*, *12*(1), 1–17.

Cabrera, S. M., Wang, X., Chen, Y.-G., Jia, S., Kaldunski, M. L., Greenbaum, C. J., Group, T. 1. D. T. C. S., Mandrup-Poulsen, T., Group, A. S., & Hessner, M. J. (2016). Interleukin-1 antagonism moderates the inflammatory state associated with type 1 diabetes during clinical trials conducted at disease onset. *European Journal of Immunology*, *46*(4), 1030–1046.

Cameron, M. J., & Kelvin, D. J. (2013). Cytokines, chemokines and their receptors. In *Madame curie bioscience database [internet]*. Landes Bioscience.

Cancelas, J. A., Jansen, M., & Williams, D. A. (2006). The role of chemokine activation of rac GTPases in hematopoietic stem cell marrow homing, retention, and peripheral mobilization. *Experimental Hematology*, *34*(8), 976–985.

Capitano, M. L., Jaiswal, A., Broxmeyer, H. E., Pride, Y., Glover, S., Amlashi, F. G., Kirby, A., Srinivasan, G., Williamson, E. A., Mais, D.others. (2021). A humanized monoclonal antibody against the endothelial chemokine CCL21

for the diagnosis and treatment of inflammatory bowel disease. *Plos One*, *16*(7), e0252805.

Charles, B. A., Hsieh, M. M., Adeyemo, A. A., Shriner, D., Ramos, E., Chin, K., Srivastava, K., Zakai, N. A., Cushman, M., McClure, L. A.others. (2018). Analyses of genome wide association data, cytokines, and gene expression in african-americans with benign ethnic neutropenia. *PloS One*, *13*(3), e0194400.

Chiou, J., Geusz, R. J., Okino, M.-L., Han, J. Y., Miller, M., Melton, R., Beebe, E., Benaglio, P., Huang, S., Korgaonkar, K.others. (2021). Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature*, *594*(7863), 398–402.

Cotsapas, C., & Hafler, D. A. (2013). Immune-mediated disease genetics: The shared basis of pathogenesis. *Trends in Immunology*, *34*(1), 22–26.

Cruikshank, W. W., Lim, K., Theodore, A. C., Cook, J., Fine, G., Weller, P. F., & Center, D. M. (1996). IL-16 inhibition of CD3-dependent lymphocyte activation and proliferation. *The Journal of Immunology*, *157*(12), 5240–5248.

Ferkingstad, E., Lund, S., Helgason, H., Magnusson, O., Halldorsson, B., Olason, P., Zink, F., Gudjonsson, S., Sveinbjornsson, G., Magnusson, M.others. (2022). *Large-scale comparison of immunoassay-and aptamer-based plasma proteomics through genetics and disease*.

Ferkingstad, E., Sulem, P., Atlason, B. A., Sveinbjornsson, G., Magnusson, M. I., Styrmisdottir, E. L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B. V.others. (2021). Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics*, *53*(12), 1712–1721.

Fugger, L., & Svejgaard, A. (2000). Association of MHC and rheumatoid arthritis: HLA-DR4 and rheumatoid arthritis-studies in mice and men. *Arthritis Research & Therapy*, *2*(3), 1–5.

Gharib, K., Gadallah, H., & Elsayed, A. (2021). Chemokines in vitiligo pathogenesis: CXCL10 and 12. *The Journal of Clinical and Aesthetic Dermatology*, *14*(9), 27.

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, *10*(5), e1004383.

Gijsbers, K., Van Assche, G., Joossens, S., Struyf, S., Proost, P., Rutgeerts, P., Geboes, K., & Van Damme, J. (2004). CXCR1-binding chemokines in inflammatory bowel diseases: Down-regulated IL-8/CXCL8 production by leukocytes in crohn's disease and selective GCP-2/CXCL6 expression in inflamed intestinal

tissue. *European Journal of Immunology*, *34*(7), 1992–2000.

Girolomoni, G., Mrowietz, U., & Paul, C. (2012). Psoriasis: Rationale for targeting interleukin-17. *British Journal of Dermatology*, *167*(4), 717–724.

Gouda, W., Mageed, L., Abd El Dayem, S. M., Ashour, E., & Afify, M. (2018). Evaluation of pro-inflammatory and anti-inflammatory cytokines in type 1 diabetes mellitus. *Bulletin of the National Research Centre*, *42*(1), 1–6.

Hutyrová, B., Pantelidis, P., Drábek, J., Zurkova, M., Kolek, V., Lenhart, K., Welsh, K. I., Du Bois, R. M., & Petrek, M. (2002). Interleukin-1 gene cluster polymorphisms in sarcoidosis and idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine*, *165*(2), 148–151.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T.others. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science*, *308*(5720), 385–389.

Leng, Q., Nie, Y., Zou, Y., & Chen, J. (2008). Elevated CXCL12 expression in the bone marrow of NOD mice is associated with altered t cell and stem cell trafficking and diabetes development. *BMC Immunology*, *9*(1), 1–12.

Lu, J., Liu, J., Li, L., Lan, Y., & Liang, Y. (2020). Cytokines in type 1 diabetes: Mechanisms of action and immunotherapeutic targets. *Clinical & Translational Immunology*, *9*(3), e1122.

MacArthur, J. A., Buniello, A., Harris, L. W., Hayhurst, J., McMahon, A., Sollis, E., Cerezo, M., Hall, P., Lewis, E., Whetzel, P. L.others. (2021). Workshop proceedings: GWAS summary statistics standards and sharing. *Cell Genomics*, *1*(1), 100004.

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A.others. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301.

Mandrup-Poulsen, T., Pickersgill, L., & Donath, M. Y. (2010). Blockade of interleukin 1 in type 1 diabetes mellitus. *Nature Reviews Endocrinology*, *6*(3), 158–166.

Martin, A. P., Coronel, E. C., Sano, G., Chen, S.-C., Vassileva, G., Canasto-Chibuque, C., Sedgwick, J. D., Frenette, P. S., Lipp, M., Furtado, G. C.others. (2004). A novel model for lymphocytic infiltration of the thyroid gland generated by transgenic expression of the CC chemokine CCL21. *The Journal of Immunology*, *173*(8), 4791–4798.

Meagher, C., Beilke, J., Arreaza, G., Mi, Q.-S., Chen, W., Salojin, K., Horst, N.,

Cruikshank, W. W., & Delovitch, T. L. (2010). Neutralization of interleukin-16 protects nonobese diabetic mice from autoimmune type 1 diabetes by a CCL4-dependent mechanism. *Diabetes*, *59*(11), 2862–2871.

Mikuniya, T., Nagai, S., Takeuchi, M., Mio, T., Hoshino, Y., Miki, H., Shigematsu, M., Hamada, K., & Izumi, T. (2000). Significance of the interleukin-1 receptor antagonist/interleukin-1β ratio as a prognostic factor in patients with pulmonary sarcoidosis. *Respiration*, *67*(4), 389–396.

Monastero, R. N., & Pentyala, S. (2017). Cytokines as biomarkers and their respective clinical cutoff levels. *International Journal of Inflammation*, *2017*.

Pickens, S. R., Chamberlain, N. D., Volin, M. V., Pope, R. M., Mandelin, A. M., & Shahrara, S. (2011). Characterization of CCL19 and CCL21 in rheumatoid arthritis. *Arthritis & Rheumatism*, *63*(4), 914–922.

Prahalad, S., & Glass, D. N. (2002). Is juvenile rheumatoid arthritis/juvenile idiopathic arthritis different from rheumatoid arthritis? *Arthritis Research & Therapy*, *4*(3), 1–8.

Pryhuber, K., Murray, K., Donnelly, P., Passo, M., Maksymowych, W., Glass, D., Giannini, E., & Colbert, R. (1996). Polymorphism in the LMP2 gene influences disease susceptibility and severity in HLA-B27 associated juvenile rheumatoid arthritis. *The Journal of Rheumatology*, *23*(4), 747–752.

Qu, P., Ji, R.-C., & Kato, S. (2005). Expression of CCL21 and 5-nase on pancreatic lymphatics in nonobese diabetic mice. *Pancreas*, *31*(2), 148–155.

Rezk, A. F., Kemp, D. M., El-Domyati, M., El-Din, W. H., Lee, J. B., Uitto, J., Igoucheva, O., & Alexeev, V. (2017). Misbalanced CXCL12 and CCL5 chemotactic signals in vitiligo onset and progression. *Journal of Investigative Dermatology*, *137*(5), 1126–1134.

Sun, M.-Y., Wang, S.-J., Li, X.-Q., Shen, Y.-L., Lu, J.-R., Tian, X.-H., Rahman, K., Zhang, L.-J., Nian, H., & Zhang, H. (2019). CXCL6 promotes renal interstitial fibrosis in diabetic nephropathy by activating JAK/STAT3 signaling pathway. *Frontiers in Pharmacology*, *10*, 224.

Swiecki, M., & Colonna, M. (2011). Type i interferons: Diversity of sources, production pathways and effects on immune responses. *Current Opinion in Virology*, *1*(6), 463–475.

Takahashi, K., Ohara, M., Sasai, T., Homma, H., Nagasawa, K., Takahashi, T., Yamashina, M., Ishii, M., Fujiwara, F., Kajiwara, T.others. (2011). Serum CXCL1 concentrations are elevated in type 1 diabetes mellitus, possibly reflecting activity of anti-islet autoimmune activity. *Diabetes/Metabolism Research and Reviews*, *27*(8), 830–833.

Tembhre, M., Parihar, A., Sharma, V., Sharma, A., Chattopadhyay, P., & Gupta, S. (2015). Alteration in regulatory t cells and programmed cell death 1-expressing regulatory t cells in active generalized vitiligo and their clinical correlation. *British Journal of Dermatology*, *172*(4), 940–950.

Van Raemdonck, K., Umar, S., & Shahrara, S. (2020). The pathogenic importance of CCL21 and CCR7 in rheumatoid arthritis. *Cytokine & Growth Factor Reviews*, *55*, 86–93.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: Comparison with p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, *33*(1), 79–86.

Wang, L., Du, F., & Wang, X. (2008). TNF-α induces two distinct caspase-8 activation pathways. *Cell*, *133*(4), 693–703.

Weber, A., Wasiliew, P., & Kracht, M. (2010). Interleukin-1β (IL-1β) processing pathway. *Science Signaling*, *3*(105), cm2–cm2.

Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., Brooun, A., Wells, P., Bi, F. C.others. (2010). Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science*, *330*(6007), 1066–1071.

Xia, X.-M., Wang, F.-Y., Xu, W.-A., Wang, Z.-K., Liu, J., Lu, Y.-K., Jin, X.-X., Lu, H., & Shen, Y.-Z. (2010). CXCR4 antagonist AMD3100 attenuates colonic damage in mice with experimental colitis. *World Journal of Gastroenterology: WJG*, *16*(23), 2873.

Zeng, Y., Lin, Q., Yu, L., Wang, X., Lin, Y., Zhang, Y., Yan, S., Lu, X., Li, Y., Li, W.others. (2021). Chemokine CXCL1 as a potential marker of disease activity in systemic lupus erythematosus. *BMC Immunology*, *22*(1), 1–10.