Genetic feature engineering enables characterisation of shared risk factors in immune-mediated diseases

Oliver S Burren[1,2], Guillermo Reales[1,2], Limy Wong[1,2], John Bowes[3,4], James C Lee[1,2], Anne Barton[3,4], Paul A Lyons[1,2], Kenneth GC Smith[1,2], Wendy Thomson[3,4], Paul DW Kirk[1,5,6], Chris Wallace* [1,2,5]

[1]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), Jeffrey Cheah Biomedical Centre, Cambridge Biomedical Campus, University of Cambridge, Puddicombe Way, Cambridge CB2 0AW, UK.

[2]Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge, CB2 0QQ,UK.

[3]National Institute of Health Research Manchester Biomedical Research Centre, Manchester Academic Health Science Centre, Manchester University NHS Foundation Trust, Manchester, UK.

[4]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, The University of Manchester, UK.

[5]MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge, CB2 0SR, UK.

[6]Cancer Research UK Cambridge Centre, Ovarian Cancer Programme, University of Cambridge Li Ka Shing Centre, Robinson Way, Cambridge, CB2 0RE, UK.

* Corresponding author cew54@cam.ac.uk

Background Genome-wide association studies (GWAS) have identified pervasive sharing of genetic architectures across multiple immune-mediated diseases (IMD). By learning the genetic basis of IMD risk from common diseases, this sharing can be exploited to enable analysis of less frequent IMDs where, due to limited sample size, traditional GWAS techniques are challenging.
Methods Exploiting ideas from Bayesian genetic fine-mapping, we developed a disease-focused shrinkage approach to allow us to distill genetic risk components from GWAS summary statistics for a set of related diseases. We applied this technique to 13 larger GWAS studies of common IMD, deriving a reduced-dimension `basis' that summarised the multidimensional components of genetic risk. We used independent datasets including the UK Biobank to assess the performance of the basis and characterise individual axes. Finally we projected summary GWAS data for smaller IMD studies, with less than 1000 cases, to assess whether the approach was able to provide additional insights into genetic architecture of less common IMD or IMD subtypes, where cohort collection is challenging.
Results We identified 13 IMD genetic risk components. The projection of independent UK Biobank data demonstrated the IMD-specificity and accuracy of the basis even for traits with very limited case-size (e.g. vitiligo, 150 cases). Projection of additional IMD-relevant studies allowed us to add biological interpretation to specific components, e.g. related to raised eosinophil counts in blood and serum concentration of the chemokine CXCL10 (IP-10). On application to 22 rare IMD and IMD subtypes we were able to not only highlight subtype-discriminating axes (e.g. for juvenile idiopathic arthritis) but also suggest eight novel genetic associations.
Conclusions Requiring only summary level data, our unsupervised approach allows the genetic architectures across any range of clinically-related traits to be characterised in fewer dimensions. This facilitates the analysis of studies with modest sample size by matching shared axes of both genetic and biological risk across a wider disease domain, and provides an evidence-base for possible therapeutic repurposing opportunities.
Word count = 324 (Max is 350).

Background

The collected summary data of genome-wide association studies (GWAS) represent, in a compressed form, assays of thousands of phenotypes across millions of common genetic variants. Analysed individually, GWAS have elucidated the polygenic component of common human diseases[1] and comparative studies of summary GWAS results have highlighted a shared genetic aetiology across different diseases[2]. Evidence for such sharing can highlight opportunities for therapeutic repurposing [3]. However, comprehensive overviews of sharing between multiple diseases are made difficult by the dimension of these statistics (100,000s of SNPs), the complex patterns that exist, and the limitation that while all dimensions carry information about technical differences between studies (DNA storage, processing, and population sampling), only a minority carry information about disease risk. Therefore, integrative analyses have typically been approached from one of two angles: a variant-by-variant analysis across multiple diseases focusing on individual variants in turn[4,5], or pairwise analysis of diseases across multiple variants at a regional or genome-wide level[6,7]. Both approaches have limitations. Different variants reflect different patterns of sharing across diseases, making generalisations about inter-disease relationships difficult, while disease-pairwise approaches make comparison of more than two diseases challenging. Thus, a need exists for a framework to study shared genetic architectures across multiple variants and between multiple diseases simultaneously.

The GWAS approach explicitly accounts for the number of tests (SNPs) by requiring successively larger samples (tens of thousands). Large samples present an insurmountable barrier for rare diseases, where efforts have instead focused on searching for rare variants of high penetrance through whole exome[8] or whole genome[9,10] sequencing. Despite this, moderate-sized GWAS-style studies of rare diseases have found both polygenic association with common variants[10,11] and evidence for differential genetic associations between clinical subtypes of these rare diseases[12]. Thus, a need exists to democratise GWAS to less common diseases, which may be possible by considering them in the context of more common, clinically-related diseases.

We propose summarising the multifactorial genetic risks of related diseases in an informed dimension-reduction approach. Matrix decomposition, for example via principal component analysis (PCA), expresses a matrix as the product of two smaller matrices, and has been used extensively as a dimension reduction tool in genetics to summarise population structure and address its confounding effects in association studies[13]. It has also been used to explore structure in genetic association with multiple traits, either from different studies aggregating signals across nearby SNPs[14], or using a linkage disequilibrium (LD) independent subset of SNPs from a single cohort[15]. In either case, the reduced dimensional space was used to explore the same datasets as used to define it, with two implications. First, GWAS summary statistics are a composite of biological signal, technical noise, and sampling variation. Decomposition aims to find axes that maximise variance explained in the input datasets, and cannot distinguish between these three sources of variability. We therefore expect it to magnify technical and random differences as well as biological, a problem related to over-fitting in high-dimensional datasets. Second, in this reduced dimension space there is no treatment of uncertainty, so whilst we can measure the distance between diseases, we are unable to formally assess whether that distance significantly differs from 0.

Here we propose augmenting PCA of GWAS summary statistics by a Bayesian shrinkage approach that mitigates overfitting. Our central aim is to define a reduced dimension space, with components that describe different patterns of genetic susceptibility corresponding to underlying biological risk factors. In a transfer learning paradigm, we can project independent datasets into this space, allowing us to study the distinct and shared genetic contributions to related diseases, and use standard statistical techniques to test for genetic association of rare diseases or genetic differences between disease subtypes. We use immune-mediated diseases (IMD) as an example of a set of traits with established aetiological overlap[2] to highlight the potential uses of this method.

Methods

Method for constructing a common genetic basis for related diseases

We aimed to decompose common components underlying susceptibility to a set of related diseases using PCA. There are three particular challenges with performing PCA on GWAS summary statistics. First, the SNP effect estimates (e.g. log odds ratios, denoted $\hat{\beta}$) must be on the same scale; second, we must deal with variable correlation between input dimensions (SNPs) due to LD; and third, while all SNPs are expected to show small deviations between studies due to random noise, different genotyping platforms and data processing decisions, only a minority of SNPs will be truly related to the diseases of interest.

The uncertainty attached to $\hat{\beta}$ depends on both study sample size and SNP minor allele frequency (MAF). We adjusted for the variance in $\hat{\beta}$ due to MAF, σ^{2}_{MAF}❑, as this varies between SNPs, but not variance due to sample size, as this would overly shrink smaller studies relative to larger. To ensure the disease-relevance of the basis, we wanted to preferentially use information from truly associated SNPs, while avoiding double counting evidence from SNPs in LD. We therefore dealt with the latter two challenges simultaneously, using a Bayesian fine mapping technique which calculates the “posterior probability” that each SNP is causal for each trait, under the assumption that at most one causal variant exists in each recombination hotspot-defined block of SNPs[16,17]. Note that the method also assumes the causal variant is in the dataset, an assumption likely to be violated without dense GWAS data. We thus use the method not to interpret the output as genuine probabilities, but for its side effect of generating a shrinkage weight that naturally adjusts for LD. At each SNP, we computed a weighted average of the posterior probabilities across input studies to create an overall weight for that SNP, w. w will be close to zero when there is no association in a region, limiting the influence of technical noise between studies, and will otherwise act to weight associated SNPs according to the extent of LD in a region. The final input for basis creation is a matrix of $\hat{\gamma}=w\hat{\beta}/\sigma_{MAF}$.

A mathematically detailed summary is given in Additional File 1, and a summary of the method is shown in Fig. 1.

Construction of IMD basis

We identified 13 IMD GWAS studies with >6,000 samples of European ancestry for which full summary statistics were publicly available (Additional File 2: Table S1). Studies were chosen to balance the competing aims of maximising the number of studies, the number of SNPs common to all studies, and the number of samples in each study (to minimise noise in $\hat{\beta}$). We selected SNPs present in all 13 studies, with MAF>1% in the 1000 Genomes Phase 3 EUR data. We excluded all variants within the major histocompatibility complex (MHC, GRCh37 Chr6:20-40Mb) due to its long and complex LD structure, and because SNPs in the MHC have a profound involvement in IMD susceptibility, and thus the potential to dominate the basis. We also excluded SNPs for which the unambiguous assignment of the effect allele was impossible (e.g. palindromic SNPs). We harmonised all effect estimates to be with respect to the alternative allele relative to the reference allele as defined by the 1000 Genomes reference genotype panel. After filtering, harmonised effect estimates were available for 265,887 SNPs across all 13 selected `basis' traits (Additional File 3: Fig. S1), and additional analyses of a subset of six datasets with dense genotyping showed that these 265,887 SNPs adequately tagged the information available in the full SNP data (Additional

File 1). In order to provide a baseline for subsequent analyses we created an additional synthetic control trait, for which effect sizes across all traits were set to zero. This can be thought of as the limit of a simulated null GWAS as the number of cases and controls tend to infinity (Additional File 1). We used these to construct two matrices M and M' where elements reflect raw ($\hat{\beta}$) and shrunk effect sizes ($\hat{\gamma} = w\hat{\beta}/\sigma_{MAF}$ respectively, such that rows and columns reflect traits (n=14) and SNPs (p=265,887). After mean centring columns we used the R command *prcomp* to carry out PCA of both M and M' to generate naive and "shrunk" IMD bases. It is likely that the trailing components of any PCA represent noise, so to assess the maximal subset of informative components, we examined the mean squared reconstruction error and found that the fewest components needed to minimise this error was m=n-1=13 (Additional File 3: Fig. S2). We therefore discarded the final 14th component. As in conventional PCA, this basis consists of orthogonal principal components (PCs), constructed as linear functions of input $\hat{\beta}$, which together provide a lower dimensional representation of genetic associations with IMD.

Driver SNPs

We noted that the majority of entries in the p x m PCA rotation matrix, Q, were close to 0, and chose to hard threshold these to 0 for computational efficiency and to identify which *driver SNPs* were relevant to each component. To do this, using Q_k to represent the k^{th} column of Q, we define $Q_k(\alpha)=Q_k \times I(|Q_k|>\alpha)$ where I () is an indicator function and "x" represents element-wise multiplication. We quantify the distance between projection with Q_k and $Q_k(\alpha)$ by

$$D_k(\alpha) = 1 - \mathrm{cor}(M^C Q_k, M^C Q_k(\alpha)).$$

where M^C is the centered matrix of shrunk effect sizes M', defined above. We chose the threshold for each component, α_k, as the largest value α such that $D_k(\alpha) < 0.001$. Finally, we defined the sparse basis rotation matrix as the matrix constructed from the column vectors Q_k, k=1,...,m. This identified both driver SNPs which define the support for each component, and enabled computationally efficient examination of many traits in the reduced dimension space defined.

Projection of independent datasets

We constructed a compendium of publicly available GWAS summary statistics across a wide range of traits including UK Biobank (UKBB) self-reported traits (http://www.nealelab.is/uk-biobank, http://geneatlas.roslin.ed.ac.uk/ - Additional File 2: Tables S2-S3), IMD relevant GWAS studies (Additional File 2: Table S4), and GWAS of quantitative measures from blood count data,[18] immune cell counts,[19] and cytokine levels[20] (Additional File 2: Tables S5-S7). Disease GWAS data were obtained from the URL given or via request to study authors, with the exception of anti-neutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV), juvenile idiopathic arthritis (JIA) and psoriatic arthritis (PsA) which are described in Additional File 4 and data given in Additional File 5, Table S9.

Prior to projection, effect alleles were aligned to the 1000 Genomes reference genotype panel. For traits sensitive to missing data (studies of neuromyelitis optica, NMO,[10] and 8 by Aterido[21] see Additional File 1), we imputed missing variants using ssimp[22] (v 0.5.6 ----ref 1KG/EUR --impute.maf 0.01), otherwise we set effect estimates to zero. Data were then shrunk as for the basis traits (multiplying by w/σ_{MAF}), and projected into basis space by multiplying by the sparse basis rotation matrix Q. We report projected results as $\hat{\delta}$, the difference between the projected $\hat{\beta}$ and a projected synthetic control with all entries 0, which allows us to make statistical inference about whether its estimand, δ, differs from control. We calculated variance of $\hat{\delta}$ as described in Additional File 1.

GWAS test multiple null hypotheses of the form $\beta = 0$ to identify disease-associated SNPs. This approach has been extended to test genetic correlation through cross trait polygenic score tests. A SNP set and weights are learnt to optimise genetic prediction of a trait of interest, and the weighted sums of β are constructed in a second dataset, and tested for association with a second trait of interest[23]. We consider each component in the basis to be a polygenic score for an uncharacterised factor contributing to one or more basis input traits. We looked for an association of the projected traits to any component by testing the null hypothesis that the vector $\delta = 0$ across all 13 components using a chisquare test (Additional File 1: equation 2). This null hypothesis is related to the global GWAS null hypothesis of no association, but is restricted to the small number of components identified in the basis, which are formed as a weighted linear function of a subset of variants. Failure to reject this null could reflect either a lack of power (as with all GWAS studies), or a lack of genetic association with the common components shared by the basis diseases. We called significant associations according to FDR < 0.01, calculated using the Benjamini-Hochberg approach, run independently within the broad categories: primary analysis (UKBB self reported disease and cancer, plus IMD-relevant GWAS); blood cell counts; cytokines; immune cell counts. This was our primary measure of significance. We took the same strategy to independently calculate FDR for each component individually for additional annotation, and traits were considered “component-significant” if they were significant (component FDR < 0.01) on that component *and* overall.

Classification of diseases according to autoantibody status was performed by a specialist clinician using available medical literature. This assignment was blinded to the PC1 results.

Clustering

We used the *hclust ()* function in R to cluster diseases in the basis using agglomerative hierarchical clustering according to Ward's criterion (method="Ward.D2") on the Euclidean distance between projected locations of each disease in the basis.

Consistency

We would like to interpret significant results as representing a composite of many small effects working in consistent directions. However, false positives could also occur if a single SNP with a large weight in the basis is in LD with a SNP with a large effect on the projected trait due to chance. To guard against this, we used weighted Spearman rank correlation which is robust to such outlier observations to test the "consistency" of each projection on a subset of driver SNPs in low LD (r^2<0.01), with weights w/σ_{MAF} and significance determined by permuting the projected values. All projected values are given in Additional File 6: Table S10.

Candidate significant driver SNPs

For each of 10 diseases or subtypes with < 2000 cases and significant on at least one component (Myasthenia gravis, late onset; eosinophilic granulomatosis with polyangiitis [EGPA], myeloperoxidase positive [MPO+], ANCA-negative [ANCA-] and combined; JIA, extended oligoarticular [EO], persistent oligoarticular [PO], and polyarticular rheumatoid factor positive/negative [RF+ and RF-, respectively]), we selected all driver SNPs on any significant component, and calculated the FDR within this set of SNPs as a subset-selected FDR.[24] We ordered SNPs by increasing values of ssFDR, and deleted any SNPs in the list that were in LD (r^2>0.1) with a higher placed SNP, leaving a set of unlinked SNPs associated with each trait shown in Table 1. These were annotated through literature searches.

Results

A genetic basis for immune mediated diseases

To illustrate the importance of our informed shrinkage procedure, we built four bases, with GWAS summary statistics for the 13 IMDs shrunk differently in each case. We assessed their relative performance by projection of matching self-reported diseases (SRD) from UK BioBank (UKBB)[25] using summary statistics from a compendium provided by the Neale lab [http://www.nealelab.is/uk-biobank/], and used hierarchical clustering to examine whether expected patterns of similarities between diseases are captured in each reduced dimension space. The first was a naive approach without any shrinkage. Here, the UKBB SRD clustered with each other rather than their GWAS comparator, suggesting that the structure identified related to between study differences other than disease (Fig. 2). In contrast, in the basis created with continuous shrinkage, all selected UKBB SRD clearly clustered with their GWAS comparators (Fig. 2), suggesting that the structure captured is disease-relevant, such that UKBB data from relatively infrequent diseases such as type 1 diabetes (T1D) (318 cases) and vitiligo (105 cases) are projected onto the same vectors as their larger comparator GWAS.

To illustrate the importance of using continuous shrinkage, we compared it to hard-thresholding, as used in the single-dataset decomposition approach, DeGAs,[15] which replaced $\hat{\beta}$ by Z scores, and set Z=0 when the associated p > 0.001. As Z scores are standardised $\hat{\beta}$, this has the effect of shrinking $\hat{\beta}$ towards 0 when uncertainty is high, such as when allele *or* disease frequencies are low, which means information from more common diseases will dominate. We generated hard-thresholded, LD-thinned bases using either Z scores or $\hat{\beta}$. For these, some of the structure identified was disease-related for the larger GWAS of more common traits (asthma, multiple sclerosis [MS], Crohn’s disease, ulcerative colitis [UC]), but the smaller diseases were dominated by dataset-specific structure (Additional File 3: Fig. S3).

We projected data from three classes of study onto the basis with shrinkage. First, we used all self-reported disease and cancer traits from UKBB to characterise the basis components, to examine specificity to IMD, and to assess power as a function of sample size: case numbers for UKBB self reported IMD range from 41,000 (asthma) to 105 (vitiligo). Second, we used IMD GWAS with smaller sample sizes than used in basis construction, including diseases studied in multiple ancestral backgrounds to explore robustness to ancestry differences. Third, we used the basis to analyse studies of IMD that are too rare or clinically heterogeneous to build large GWAS cohorts.

Genetic analysis of multiple IMD in reduced dimensions

Across all 312 projected UKBB traits (Additional File 2: Table S2), 27 had significantly non-zero $\hat{\delta}$ (FDR < 1%). These were overwhelmingly immune-related traits (Fig. 3): no significance was observed for traits such as coronary artery disease, stroke, or obstructive sleep apnea, confirming the immune-mediated specificity of our basis. Significant results were detected with as few as 105 cases for vitiligo, emphasising the potential of this approach to unlock the genetics of rare IMD GWAS.

Of 28 traits from target (non-UKBB) IMD GWAS, including JIA, NMO, vasculitis and their clinical subtypes, 16 were significant (FDR < 1%, Additional File 2: Table S3, Additional File 3: Fig. S4-S16). We found, reassuringly, that increasing evidence for non-zero δ on any component correlated with increasing consistency on that component (see Methods) amongst disease traits (Additional File 3: Fig. S17), suggesting that significant results were produced by an average effect over many driver SNPs rather than random overlap of a small number of driver SNPs with trait-associated SNPs.

We clustered all 28 target traits and all 27 significant UKBB self reported traits to generate a visual overview of IMD and associated traits (Fig. 4). Hierarchical clustering solutions are generally unstable, and dependent on the composition of the items to be clustered, as well as the method used for clustering[26]. While clustering provides only a visual overview rather than a formal statistical analysis of trait similarity, it highlighted two small disease groups, inflammatory bowel disease (IBD) and EGPA, and two larger groups, one comprising autoimmune diseases and the other a heterogeneous cluster containing subgroups centred on MS, ankylosing spondylitis (AS), atopy, and traits with only weak or non-significant signals. Notably, three studies of AS all clustered together, despite only one having sufficient sample size for significant results and the three studies representing different ancestries (UK-European, International and Turkish/Iranian).

While our basis was created from predominantly European GWAS, there is an imperative to increase ancestry diversity in GWAS[27]. We undertook a search for available IMD GWAS data with coverage of non-European ancestry and identified 6 studies of asthma, RA, UC and CD in African and/or East Asian ancestry populations (Additional File 2: Table S8). Projecting these onto the basis, we find that all significant points have the same sign of delta for any given ancestry and PC combination (Additional File 3: Fig. S18). Thus, results are consistent across GWAS studies of the same traits in populations with different ancestry backgrounds. A broader examination comparing projections of all ~452,000 UKBB subjects to the European subset of 360,000 subjects found that while the mixed ancestry GWAS tended to result in slightly attenuated estimates of $\hat{\delta}$, the increased sample size also led to increased power compared to smaller European GWAS (Additional File 3: Fig. S19).

Most disease subtypes clustered together (Fig. 4). For example, myasthenia gravis, a chronic, autoimmune, neuromuscular disease characterized by muscle weakness, has been shown to have a bimodal incidence pattern by age, and some genetic associations have been identified only for the late onset subtype[28]. However, both subtypes fell in very similar locations across all components, and cluster together with several subtypes of JIA.

For two other diseases, however, subtypes clustered apart. NMO is a rare (prevalence 0.03–0.4:10,000) disease affecting the optic nerve and spinal cord for which HLA association is established[10] and which can be divided according to aquaporin 4 autoantibody seropositivity status (IgG+ or IgG-). The projections of seropositive and seronegative NMO showed non-significant differences on several components, leading to differential clustering. While seropositive NMO clustered with the classical autoimmune diseases, most closely with SLE and Sjögren’s disease, IgG- NMO clustered away from the classic seropositive diseases, most closely with MS. This finding mirrors analysis which directly compared NMO subtypes to each of SLE and MS via polygenic scores[10], and strengthens the findings by specifically suggesting SLE and MS as the nearest neighbours of IgG+ and IgG- NMO respectively, out of all IMDs considered for clustering.

JIA is a heterogeneous paediatric disease, with an overall childhood prevalence in Europe of 20/10,000[29], and with seven recognised subtypes[30]. While studies have begun to identify distinct genetics of the systemic subtype[31] and have shown subtype-specific differences in the MHC[32], systematic comparison between subtypes has been underpowered. Although, the systemic and enthesitis-related arthritis (ERA) subtypes did not significantly differ from controls (despite relatively moderate sample sizes of 219 and 267 cases respectively), they clustered with MS and AS respectively, and away from the other JIA subtypes, which clustered with the other autoimmune diseases.

Association of driver SNPs to rare IMD or subtypes

Given that most of the IMD and subtypes with small GWAS studies have few established genetic associations, we sought to exploit the component-level associations above to detect new disease associations. Our basis has only 13 dimensions. If genetic susceptibility to rare IMD and IMD subtypes overlaps that of common IMD, we can increase power by focusing on these dimensions. Of 22 diseases or disease subtypes with < 1000 cases, 12 were significant (FDR<1%), even with as few as 132 cases (NMO IgGPos).

Although not a specific goal, the basis generated is naturally sparse (Additional File 3: Fig. S20), enabling us to identify 107-373 “driver SNPs” that are required to capture genetic associations on any individual component. We found a strong enrichment for small GWAS p values at driver SNPs on trait-significant components (Additional File 3: Fig. *S*21). Using a “subset-selected” FDR approach[24], we analysed driver SNPs for 22 significant trait-component pairs (12 unique traits), and identified 25 trait-SNP associations (subset-selected FDR < 1%, Table 1) after pruning SNPs in LD. Twelve of these were genome-wide significant ($p < 5x10^{-8}$) either in this study (4 associations) or in other published data (8 associations) and a further five were significant in other published analysis that levered external data. These included, for example, the non-synonymous *PTPN22* SNP rs2476601 which was associated with myasthenia gravis (overall and the late onset subset) by subset-selected FDR < 0.01. This SNP was previously associated with myasthenia gravis in a different study[33], and lack of clear replication in the data analysed here ($P=6x10^{-5}$) was attributed to differences in population structure. Eight associations (five variants) were not previously reported to our knowledge, including associations near *IRF1/IL5* for myasthenia gravis, near *TNFSF11* for RF- JIA and near *CD2/CD2*8 for EGPA.

Component interpretation

PC1, which explained the greatest variation in the training datasets, appears to represent an autoimmune / (auto)inflammatory axis[34], also characterised by whether diseases are considered antibody ‘seropositive’ or ‘seronegative’ (Fig. 5). The exception is vitiligo, in which, despite strong evidence of T cell autoimmunity, autoantibodies are reported but are not consistent features of disease[35]. Weaker but significant association of psoriatic arthritis (PsA) among the other seropositive IMD is also consistent with a recent report of novel pathogenic antibodies in PsA[36]. On the inflammatory/seronegative side, we also saw weaker but still significant signals for atopy, basal cell carcinoma and malignant melanoma. Both malignant melanoma and non-melanoma skin cancer incidence is increased in IBD, but the relative role of treatment or IBD itself in driving this is hard to determine[37,38]. On the seropositive side, we saw significant results for pernicious anemia, a disease strongly associated with anti-gastric parietal cell and anti-Intrinsic Factor antibodies, as well as with autoimmune thyroiditis, T1D and vitiligo[39].

To help characterise the biology captured by individual components we projected additional datasets: blood counts[18], immune cell counts[19], and serum cytokine concentrations[20] (Additional File 2: Tables S5, S6 and S7). Testing for consistency identified outliers in the blood count data, which had been generated from a much larger sample, and so we additionally filtered on consistency in that dataset. These data aided interpretation of two further components.

PC13 was striking for the general association of many diseases across all four main clusters in a concordant direction, and was the only component for which any projected trait was more extreme than any original basis trait (Fig. 6). The most extreme was EGPA, both ANCA+ and ANCA- subtypes. EGPA is a rare form of AAV (annual incidence 1-2 cases per million) for which genetic differences relating to autoantibody status have been identified[12]. We found PC13 was strongly associated with higher eosinophil counts in a population cohort[18] (FDR<10^{-200}), suggesting that this component describes eosinophilic involvement in IMD. This is consistent with the extreme projection of EGPA which is classified as an eosinophilic form of AAV with both asthma and raised eosinophil count included in its diagnostic criteria.

Eosinophils are pro-inflammatory leukocytes with an established role in atopic diseases such as asthma[40], inflammatory diseases such as IBD[41], and autoimmune diseases such as RA[42]. Mendelian Randomization (MR) analysis of blood cell traits had previously further associated eosinophils with celiac disease (CEL), asthma and T1D[18]. Our analysis thus supports earlier findings and extends the list of IMD with genetically supported involvement of eosinophils to include EGPA, JIA subtypes, AS, ATD, MS, hayfever and eczema, in agreement with other recent findings[43].

PC3 (Fig. 7) was the only component which showed a significant relationship with any serum cytokine concentration. Higher concentrations of CXCL9 (MIG) and CXCL10 (IP-10), Th1 chemoattractants and ligands to the regulator of leukocyte trafficking CXCR3, were both significant in the same direction as several autoimmune diseases, with strongest signals for myasthenia gravis, several JIA subtypes, as well as IBD, CEL, AS and sarcoidosis. IP-10 and MIG are chemokines, secreted by epithelial and dendritic cells (amongst others), which act as chemoattractants for immune cells which express the receptor CXCR3, including Th1 cells. Both MIG and IP-10 expression at the site of autoimmune target have been implicated in the development of autoimmunity[44,45] and IP-10 has been observed to be upregulated in follicular cells of patients with myasthenia gravis[46]. Serum IP-10 has also been found to be raised in patients with recent-onset T1D[47,48] and Graves' disease (hyperthyroidism)[45], and to correlate with increased disease activity in SLE[49] and AS[50].

Discussion

Our motivation in this work was threefold. First, to overcome the problems of dimensionality and allow an overview of genetic association patterns from multiple related diseases without over-simplification. While previous efforts to relate different traits through GWAS statistics have focused on large studies and shown that they can distinguish broad classes of immune-mediated, cardiovascular and metabolic diseases,[6,14] we have tackled the problem of finding structure *within* a single class of diseases. Unlike other applications of PCA to genetics, we split our datasets into "training" and "test" sets, enabling standard statistical hypothesis testing and providing robustness against overfitting. Importantly, our method allows synthesis of knowledge from different studies, allowing large numbers of cases from different diseases to contribute to the constructed dimensions.

Our second motivation was to generate new knowledge in rare IMDs. The number of polymorphic human genetic variants together with our understanding that genetic effects on human disease are generally modest has led to massive GWAS to overcome the penalty that must be applied for multiple testing. This is simply not possible for rare diseases. One of the tools which has enhanced rare disease GWAS is the borrowing of information from larger GWAS of aetiologically related diseases[12] and our basis serves a similar function here. By leveraging information about a SNP's potential to be IMD-associated, we can both increase genetic discovery and place less common diseases in the context of their more prevalent counterparts. More generally, studies of SRD are being enabled on a massive scale by UKBB[51] and 23andMe,[52] although studies of such cohorts tend to focus on the more common diseases such as type 2 diabetes (T2D) and coronary heart disease. Our results provide reassurance that SRD associations are consistent with those from targeted GWAS studies, and extend their utility to IMD and other diseases which are generally found at a lower frequency.

Our final motivation was to extract different axes underlying IMD genetic risk. Work in metabolic[53] and psychiatric[54] diseases have attempted to learn composite factors underlying risk of these related diseases through deeper phenotyping of patients before testing these factors for genetic association. Alternatively, decomposition of estimated effects at 94 T2D risk variants, together with their effects on 46 metabolic traits was used to cluster variants into 5 groups, three focused on insulin resistance and two on beta cell function.[55] Here, we hoped to learn the same sorts of factors by decomposing only summary GWAS data on clinical disease endpoints. Our continuous shrinkage weight learnt across all training datasets enables us to extract disease-relevant structure, with projected traits lying close to their training data counterparts, something achieved with disease-specific hard thresholded weights[15] for only the largest datasets.

There are limitations with the method. The assumption of a single causal variant per disease, and per LD-defined region, in generating SNP weights is obviously unrealistic. However, it is this simple assumption that allows us to process summary GWAS data from multiple studies without accurate LD-estimates from each study. The assumption, whilst simplistic, has nonetheless been used in both fine-mapping and colocalisation analyses, because in most cases it means only the strongest signal in each region is considered per disease[56]. More sophisticated fine mapping methods which can cope with multiple causal variants in LD will be required to adapt our method to the MHC which harbours many of the strongest IMD effects. A more impactful limitation is likely to be that signals in projected datasets can only be discovered if they are also captured in the diseases used to build the basis. Thus, the careful selection of plausibly relevant traits is important, and a negative result for a projected dataset only means no detected association with the identified components, and not an absence of genetic association. For example, the relative under-representation of atopic diseases in our input datasets may underlie the relative lack of associations seen for allergy and eczema. The number of available input datasets also limits the number of components that may be distinguished to the rank of the matrix of shrunk effect sizes, which cannot be greater than the number of datasets. For both these reasons, future work will seek to expand the number of datasets included to develop a more comprehensive IMD basis.

We found components defined using the largest GWAS studies of IMD we could access showed different patterns of association with different disease subsets, emphasising the utility of a multi-dimensional view. The autoimmune / (auto-)inflammatory axis in IMD represented by PC1 is well documented, with the gradient along PC1 corresponding to a shift from auto-antibody seronegative to seropositive diseases. Significant IMD on the MIG/IP-10-associated PC3 included both ‘seropositive’ and ‘seronegative’ diseases, although not atopy, while all three groups were represented on the eosinophil-associated PC13. While these observations support a link between certain IMD and serum cytokine levels or blood cell counts, our results do not directly implicate these as causal. Both cytokines and blood count data were measured in unselected population cohorts which will include individuals with IMD, such that the association with IMD may be causal or consequential. For example, we can conclude only that PC3 represents an IMD-related process that contributes to serum cytokine levels. Nonetheless, clinical efficacy of MDX1100, a monoclonal antibody to IP-10, has been demonstrated in RA[57] and a dose-response relationship observed in UC.[58] Our results suggest IP-10 blockade might also be considered in patients with myasthenia gravis, JIA, AS, and sarcoidosis.

Conclusions

Our proposed approach may be considered a form of feature engineering. We represent genetic associations for aetiologically related traits using radically fewer features, with attached estimates of uncertainty. This enabled us to identify clusters of IMD and nominate involvement of IP-10 and eosinophil counts as involved in a wider range of IMD than previously suggested. Such observations provide a rationale for potential therapeutic repurposing opportunities. Beyond these uses, we expect that reduced dimensional representation of multiple genetic association data sets will offer a foundation for other novel cross-disease analyses within and beyond the immune-mediated focus here.

List of abbreviations

AAV ANCA-associated vasculitis

ANCA anti-neutrophil cytoplasmic antibody

AS ankylosing spondylitis

CEL celiac disease

EGPA eosinophilic granulomatosis with polyangiitis

ERA enthesitis-related arthritis

GWAS genome-wide association studies

IBD inflammatory bowel disease

IMD immune-mediated disease(s)

JIA juvenile idiopathic arthritis

JIA subtypes:

EO extended oligoarticular

PO persistent oligoarticular

RF+/RF- polyarticular rheumatoid factor positive/negative, respectivelyDeclarations

LD linkage disequilibrium

MAF minor allele frequency

MHC major histocompatibility complex

MPO myeloperoxidase

MPO myeloperoxidase

MS multiple sclerosis

NMO neuromyelitis optica

PCA principal component analysis

PsA psoriatic arthritis

SRD self-reported diseases

T1D type 1 diabetes

T2D type 2 diabetes

UC ulcerative colitis

Declarations

Ethics approval and consent to participate

All PsA patients provided written informed consent (UK PsA National Repository MREC 99/8/84). Ethical approval was given by the HRA and North West - Haydock Research Ethics Committee.

JIA participants were recruited with ethical approval and provided informed consent, from the North West Multi-centre for Research Ethics Committee (MREC:02/8/104 and MREC:99/8/84), West Midlands Multi-centre Research Ethics Committee (MREC:02/7/106), North West Research Ethics Committee (REC:09/H1008/137) and the NHS Research Ethics Committee (REC:05/Q0508/95).

The research conformed to the principles of the Helsinki declaration.

All other datasets were publicly available, and ethical approval was not required for our use of them.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests

Additional Files

Additional File 1: Supplementary Note: Mathematical exposition of basis construction

Additional File 2: Tables S1-S8: Summary of input datasets, sources, and sample sizes

Additional File 3: Fig. S1-S21: Distribution of SNPs across the basis components, reconstruction plot error, hierarchical clustering of basis diseases and their basis counterparts using different thresholds, delta plots for the 13 principal components in the IMD basis, test for consistency across trait groups, comparison of projections across different ancestries, distribution of entries in the rotation matrix for each component in the basis, and QQ plots of p-values for driver SNPs on trait-significant components.

Additional File 4: Methods for GWAS analysis of individual level datasets: vasculitis, JIA and PsA

Additional File 5: Table S9: Input summary statistics for SNPs needed for basis projection for JIA and PsA. Beta refers to the effect of allele a2 compared to a1. se.beta is the standard error of beta. SNPs are identified by chromosome, position, reference and alternative alleles.

Additional File 6: Table S10: Projection results for each studied trait, giving the delta value for each PC, its variance, and the Benjamini-Hochberg adjusted p value (“fdr.delta”) together with an overall test of significance, both raw (“p.overall”) and Benjamini-Hochberg adjusted (“fdr.overall”)

Availability of data and materials

All data generated or analysed during this study are included in this published article and its additional information files:

Input datasets, and sources are summarised in Additional File 2, Tables S1-S8.

Input summary for datasets not yet publicly available (JIA and PsA) are given in Additional File 5, Table S9.

Projection results for each studied trait are given in Additional File 10, Table S10.

An R implementation of the method is available from https://github.com/ollyburren/cupcake/[60]**.** Code to run the analyses presented here is available from https://zenodo.org/record/4069214[61].

We also created an online tool to allow other researchers to project their own data into the basis https://grealesm.shinyapps.io/IMDbasisApp/[62]. Code underlying this tool is available at https://github.com/GRealesM/IMDbasisApp[63].

Funding

Wellcome Trust: WT107881 (Wallace, Burren, Reales), 105920/Z/14/Z (Lee), 110303/Z/15/Z (Wong), 083650/Z/07/Z (Smith) MRC: MC_UU_00002/4 (Wallace), MC_UU_00002/13 (Kirk). Funders had no role in the design of the study, collection, analysis, interpretation of data or in writing the manuscript.

Authors' contributions

Conceived the study, drafted paper: CW, OB.

Wrote the software: OB.

Performed analyses: OB, CW, LW, JB.

Interpreted data: OB, CW, JL, PDWK, KGCS.

Acquired data: OB, JB, WT, JB, KGCS, PAL, AB, CW.

Created online projection tool: GR, OB.

All authors read and approved the final manuscript.

Acknowledgements

We thank the following for sharing data from their studies:

Ann Morgan and Jennifer Barrett for methotrexate response in RA, on behalf of the CARDERA, IACON, PAMERA, and RAMS Consortia[64]

Jonas Kuiper and Bobby Koeleman for the birdshot chorioretinopathy GWAS[65]; and all researchers who made their GWAS summary data available on the GWAS catalog.

We thank Urs Christen for helpful discussions on IP-10.

This study acknowledges the use of the following UK JIA cohort collections: British Society of Paediatric and Adolescent Rheumatology (BSPAR) study group, Childhood Arthritis Prospective Study (CAPS) (funded by Versus Arthritis, grant reference number 20542), Childhood Arthritis Response to Medication Study (CHARMS) (funded by Sparks UK, reference 08ICH09, and the Medical Research Council, reference MR/M004600/1), United Kingdom Juvenile Idiopathic Arthritis Genetics Consortium (UKJIAGC). Genotyping of the UK JIA and PsA case samples was supported by the Versus Arthritis grants reference numbers 20385 and 21754. This research was funded by the NIHR Manchester Biomedical Research Centre and supported by the Manchester Academic Health Sciences Centre (MAHSC). The views expressed are those of the author (s) and not necessarily those of the NHS, the NIHR or the Department of Health. We would like to acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester. Understanding Society: The UK Household Longitudinal Study is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. The survey was conducted by NatCen and the genome-wide scan data were analysed and deposited by the Wellcome Trust Sanger Institute. Information on how to access the data can be found on the Understanding Society website https://www.understandingsociety.ac.uk/.

References

1. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

2. Cotsapas, C. & Hafler, D. A. Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends Immunol.* **34**, 22–26 (2013).

3. Bovijn, J., Censin, J. C., Lindgren, C. M. & Holmes, M. V. Using human genetics to guide the repurposing of medicines. *Int. J. Epidemiol.* (2020) doi:10.1093/ije/dyaa015.

4. Majumdar, A., Haldar, T., Bhattacharya, S. & Witte, J. S. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS Genet.* **14**, (2018).

5. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).

6. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

7. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat. Genet.* **47**, 839+ (2015).

8. Yang, Y. *et al.* Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).

9. Ouwehand, W. H. Whole-genome sequencing of rare disease patients in a national healthcare system. *Nature* doi:10.1101/507244.

10. Estrada, K. *et al.* A whole-genome sequence study identifies genetic risk factors for neuromyelitis optica. *Nat. Commun.* **9**, 1929 (2018).

11. Li, J. *et al.* Association of CLEC16A with human common variable immunodeficiency disorder and role in murine B cells. *Nat. Commun.* **6**, 6804 (2015).

12. Lyons, P. A. *et al.* Genome-wide association study of eosinophilic granulomatosis with polyangiitis reveals genomic loci stratified by ANCA status. *Nat. Commun.* **10**, 5120 (2019).

13. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

14. Chang, D. & Keinan, A. Principal component analysis characterizes shared pathogenetics from genome-wide association studies. *PLoS Comput. Biol.* **10**, e1003820 (2014).

15. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).

16. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P - values. *Genet. Epidemiol.* **33**, 79¬ﾀﾓ86 (2009).

17. The Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294¬ﾀﾓ1301 (2012).

18. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016).

19. Roederer, M. *et al.* The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. *Cell* **161**, 387–403 (2015).

20. Ahola-Olli, A. V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am. J. Hum. Genet.* **100**, 40–50 (2017).

21. Aterido, A. *et al.* Genetic variation at the glycosaminoglycan metabolism pathway contributes to the risk of psoriatic arthritis but not psoriasis. *Ann. Rheum. Dis.* **78**, (2019).

22. Rüeger, S., McDaid, A. & Kutalik, Z. Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet.* **14**, e1007371 (2018).

23. Power, R. A. *et al.* Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat. Neurosci.* **18**, 953–955 (2015).

24. Yekutieli, D. *et al.* Approaches to multiplicity issues in complex research in microarray analysis. *Stat. Neerl.* **60**, 414–437 (2006).

25. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a

wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

26. Smith, S. P. & Dubes, R. Stability of a hierarchical clustering. *Pattern Recognit.* **12**, 177–187 (1980).
27. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
28. Renton, A. E. *et al.* A genome-wide association study of myasthenia gravis. *JAMA Neurol.* **72**, 396–404 (2015).
29. Thierry, S., Fautrel, B., Lemelle, I. & Guillemin, F. Prevalence and incidence of juvenile idiopathic arthritis: a systematic review. *Joint Bone Spine* **81**, 112–117 (2014).
30. Petty, R. E. *et al.* International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J. Rheumatol.* **31**, 390–392 (2004).
31. Ombrello, M. J. *et al.* Genetic architecture distinguishes systemic juvenile idiopathic arthritis from other forms of juvenile idiopathic arthritis: clinical and therapeutic implications. *Ann. Rheum. Dis.* **76**, 906–913 (2017).
32. Hinks, A. *et al.* Fine-mapping the MHC locus in juvenile idiopathic arthritis (JIA) reveals genetic heterogeneity corresponding to distinct adult inflammatory arthritic diseases. *Ann. Rheum. Dis.* **76**, 765–772 (2017).
33. Gregersen, P. K. *et al.* Risk for myasthenia gravis maps to a (151) Pro→Ala change in TNIP1 and to human leukocyte antigen-B*08. *Ann. Neurol.* **72**, 927–935 (2012).
34. McGonagle, D. & McDermott, M. F. A proposed classification of the immunological diseases. *PLoS Med.* **3**, e297 (2006).
35. Boniface, K., Seneschal, J., Picardo, M. & Taïeb, A. Vitiligo: Focus on Clinical Aspects, Immunopathogenesis, and Therapy. *Clin. Rev. Allergy Immunol.* **54**, 52–67 (2018).
36. Yuan, Y. *et al.* Identification of Novel Autoantibodies Associated With Psoriatic Arthritis.

Arthritis Rheumatol **71**, 941–951 (2019).

37. Singh, H., Nugent, Z., Demers, A. A. & Bernstein, C. N. Increased risk of nonmelanoma skin cancers among individuals with inflammatory bowel disease. *Gastroenterology* **141**, 1612–1620 (2011).
38. Singh, S. *et al.* Inflammatory bowel disease is associated with an increased risk of melanoma: a systematic review and meta-analysis. *Clin. Gastroenterol. Hepatol.* **12**, 210–218 (2014).
39. Toh, B.-H. Pathophysiology and laboratory diagnosis of pernicious anemia. *Immunol. Res.* **65**, 326–330 (2017).
40. Busse, W. W. & Sedgwick, J. B. Eosinophils in asthma. *Ann. Allergy* **68**, 286–290 (1992).
41. Al-Haddad, S. & Riddell, R. H. The role of eosinophils in inflammatory bowel disease. *Gut* vol. 54 1674–1675 (2005).
42. Hällgren, R., Feltelius, N., Svenson, K. & Venge, P. Eosinophil involvement in rheumatoid arthritis as reflected by elevated serum levels of eosinophil cationic protein. *Clin. Exp. Immunol.* **59**, 539–546 (1985).
43. Diny, N. L., Rose, N. R. & Čiháková, D. Eosinophils in Autoimmune Diseases. *Front. Immunol.* **8**, 484 (2017).
44. Christen, U., McGavern, D. B., Luster, A. D., von Herrath, M. G. & Oldstone, M. B. A. Among CXCR3 chemokines, IFN-gamma-inducible protein of 10 kDa (CXC chemokine ligand (CXCL) 10) but not monokine induced by IFN-gamma (CXCL9) imprints a pattern for the subsequent development of autoimmune disease. *J. Immunol.* **171**, 6838–6845 (2003).
45. Romagnani, P. *et al.* Expression of IP-10/CXCL10 and MIG/CXCL9 in the thyroid and increased levels of IP-10/CXCL10 in the serum of patients with recent-onset Graves' disease. *Am. J. Pathol.* **161**, 195–206 (2002).

46. Meraouna, A. *et al.* The chemokine CXCL13 is a key molecule in autoimmune myasthenia gravis. *Blood* **108**, 432–440 (2006).

47. Shimada, A. *et al.* Elevated serum IP-10 levels observed in type 1 diabetes. *Diabetes Care* **24**, 510–515 (2001).

48. Antonelli, A. *et al.* Serum Th1 (CXCL10) and Th2 (CCL2) chemokine levels in children with newly diagnosed Type 1 diabetes: a longitudinal study. *Diabet. Med.* **25**, 1349–1353 (2008).

49. Kong, K. O. *et al.* Enhanced expression of interferon-inducible protein-10 correlates with disease activity and clinical manifestations in systemic lupus erythematosus. *Clin. Exp. Immunol.* **156**, 134–140 (2009).

50. Wang, J. *et al.* Circulating levels of Th1 and Th2 chemokines in patients with ankylosing spondylitis. *Cytokine* **81**, 10–14 (2016).

51. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

52. Tian, C. *et al.* Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.* **8**, 599 (2017).

53. Avery, C. L. *et al.* A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS Genet.* **7**, e1002322 (2011).

54. Mallard, T. T. *et al.* Not just one p: Multivariate GWAS of psychiatric disorders and their cardinal symptoms reveal two dimensions of cross-cutting genetic liabilities. *bioRxiv* 603134 (2019) doi:10.1101/603134.

55. Udler, M. S. *et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med.* **15**, e1002654 (2018).

56. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic

association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).

57. Yellin, M. *et al.* A phase II, randomized, double-blind, placebo-controlled study evaluating the efficacy and safety of MDX-1100, a fully human anti-CXCL10 monoclonal antibody, in combination with methotrexate in patients with rheumatoid arthritis. *Arthritis Rheum.* **64**, 1730–1739 (2012).

58. Mayer, L. *et al.* Anti-IP-10 antibody (BMS-936557) for ulcerative colitis: a phase II randomised study. *Gut* **63**, 442–450 (2014).

59. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature Publishing Group* **45**, 664–669 (2013).

60. Burren, O.S., Wallace, C. cupcake. Github. https://github.com/ollyburren/cupcake/ (2020).

61. Burren, O. S., Wallace, C. R code to support "Genetic feature engineering enables characterisation of shared risk factors in immune-mediated diseases". Zenodo. https://zenodo.org/record/4069214 (2020).

62. Reales, G., Burren, O.S. IMD basis App. shinyapps.io. https://grealesm.shinyapps.io/IMDbasisApp/ (2020).

63. Reales, G., Burren, O.S. IMD basis App. Github. https://github.com/GRealesM/IMDbasisApp (2020).

64. Taylor, J. C. et al. Genome-wide association study of response to methotrexate in early rheumatoid arthritis patients. Pharmacogenetics J. 18(4), 528-538 (2018).

65. Kuiper J.J.W. et al. A genome-wide association study identifies a functional ERAP2 haplotype associated with birdshot chorioretinopathy. Hum. Mol. Genet. 23(22), 6081-6087 (2014).

Figure Legends

Fig. 1. Schematic of basis creation and projection. Basis creation: GWAS summary statistics for related traits are combined to create a matrix, M (n x m), of harmonised effect sizes ($\hat{\beta}$) and a learned vector of shrinkage values for each SNP. After multiplying each row of M by the shrinkage vector, PCA is used to decompose M into component and loading matrices. Basis projection: For an independent set of studies, trait effects are harmonised with respect to the basis, shrinkage applied and the resultant vector is multiplied by the basis loading matrix to obtain component scores. These component scores can be used for testing hypotheses of the form that a weighted average of effect sizes in the test GWAS is non-zero, because the weights (basis loading matrix) are learnt from an independent set of large GWAS.

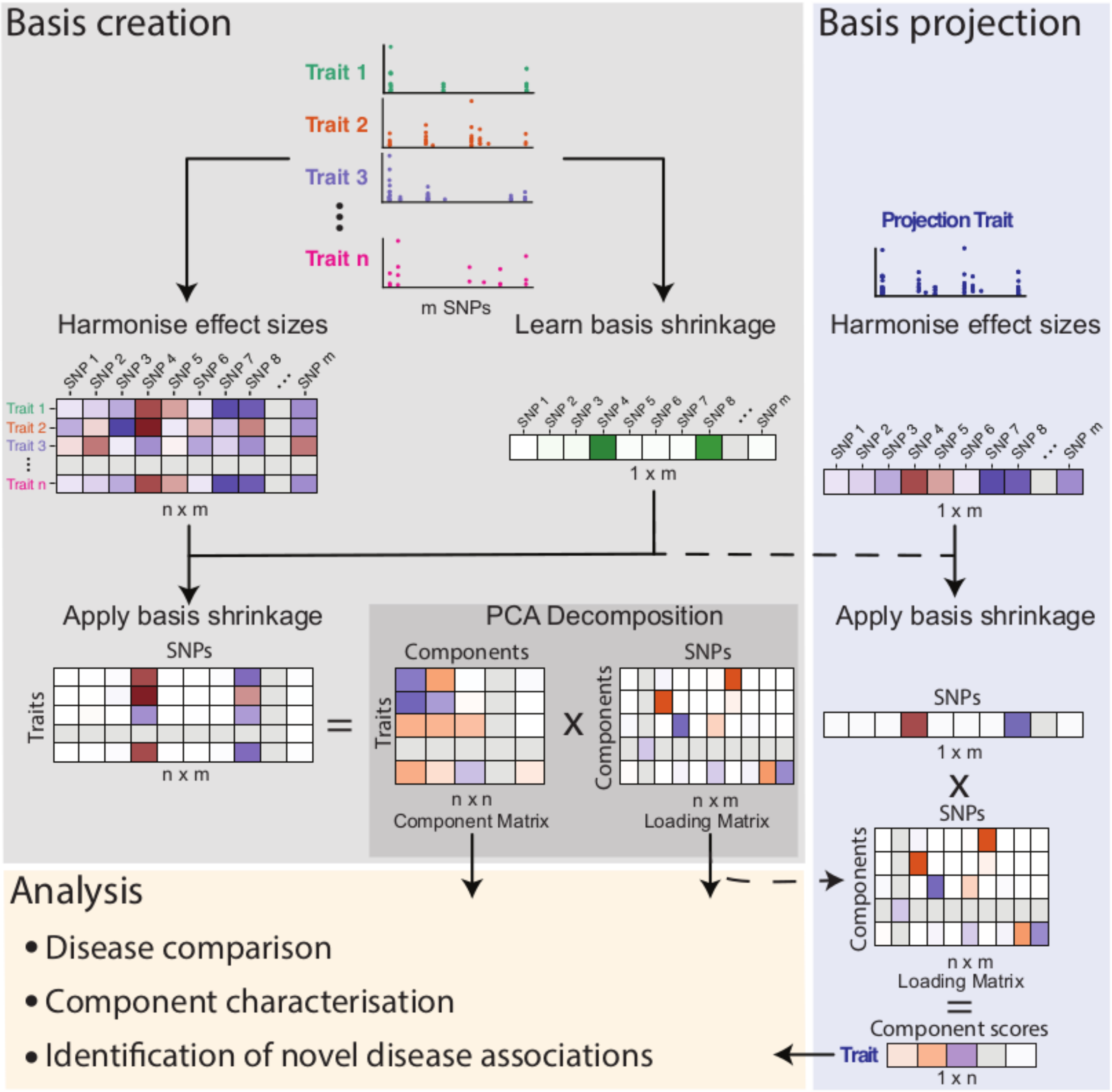


Fig. 2. Hierarchical clustering of basis diseases and their UKBB counterparts in basis space

a unweighted basis constructed using $\hat{\beta}$**b** basis constructed using continuous shrinkage applied to $\hat{\beta}$. Heatmaps indicate projected $\hat{\delta}$ for each disease on each component PC1-PC13, with grey indicating 0 (no difference from control), and darker shades of green or magenta showing departure from controls in one direction or the other. GWAS datasets: T1D = type 1 diabetes, CEL= celiac disease, asthma, MS = multiple sclerosis, UC = ulcerative colitis, CD = Crohn's disease, RA = rheumatoid arthritis, VIT = vitiligo, SLE = systemic lupus erythematosus, PSC = primary sclerosing cholangitis, PBC= primary biliary cholangitis, LADA= latent autoimmune diabetes in adults, IgA_NEPH= IgA nephropathy. UKBB_ prefixed diseases correspond to self reported disease status in UK Biobank.

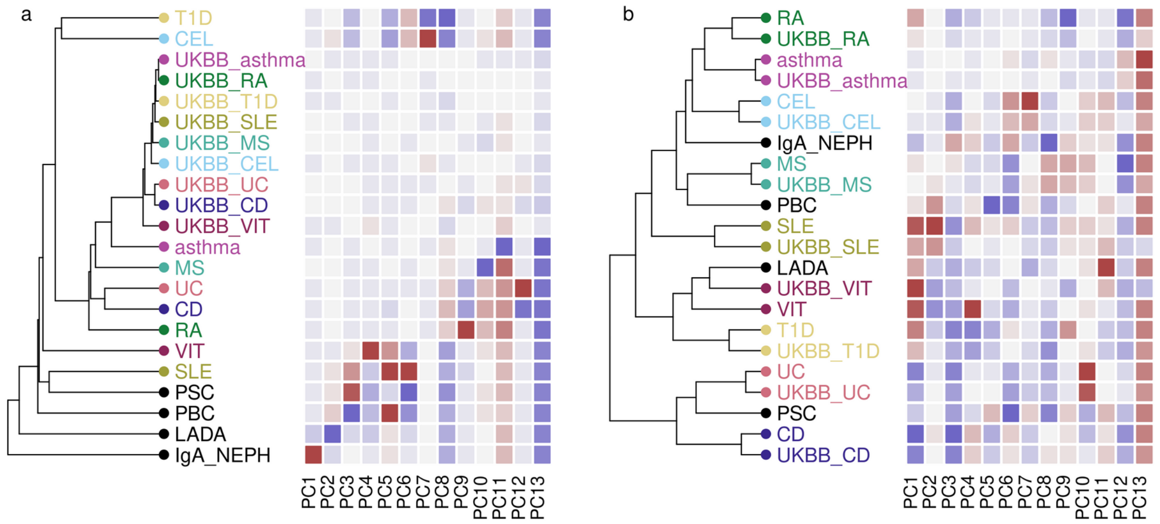


Fig. 3. Of 312 UKBB self-reported traits projected onto the basis, 27 were significant at FDR < 1%, and IMD were enriched amongst this set, with 63% of IMD showing significance compared to <3% of non-IMD traits. Each trait projected is shown according to FDR (-log10 scale, axis truncated at FDR=10^{-6} for display) and number of cases. All IMD (yellow) and all significant non-IMD traits (grey) are labelled.

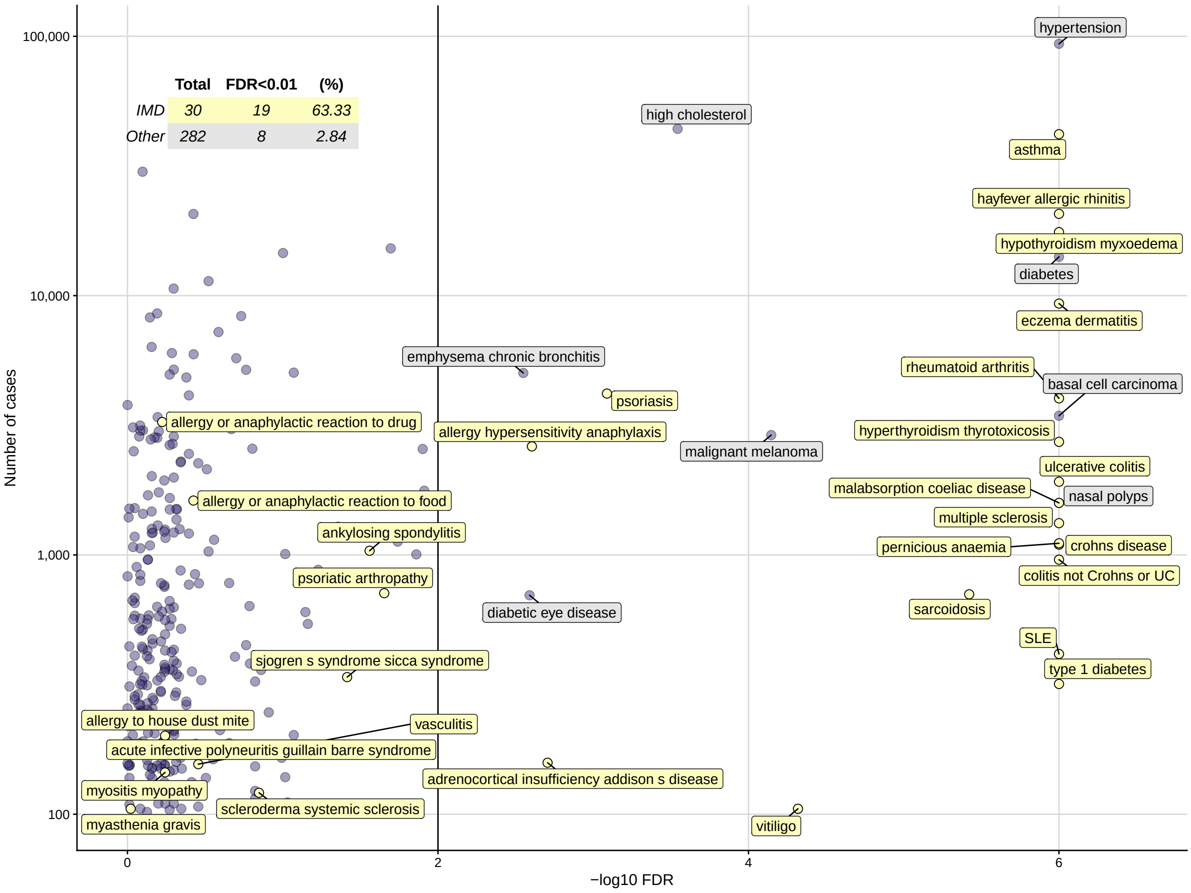


Fig. 4. Hierarchical clustering of projected diseases significantly different from control (FDR < 1\%) or of small sample size. Coloured labels are used to distinguish UKBB (grey) and other GWAS (green) datasets. Heatmaps indicate delta values for each disease on each component PC1-PC13, with grey indicating 0 (no difference from control), and darker shades of blue or magenta showing departure from controls in one direction or the other. An overlaid * indicates delta was significantly non zero (FDR<0.01). Roman numerals indicate clusters described in the text. Abbreviations: ANCA- = anti-neutrophil cytoplasmic antibody negative, Ank. Spond = ankylosing spondylitis, EGPA = eosinophilic granulomatosis with polyangiitis, EO = extended oligo, ERA = juvenile enthesitis-related arthritis, IgGPos = IgG positive, JIA = juvenile idiopathic arthritis, MPO+ = myeloperoxidase positive NMO = neuromyelitis optica,

PO = persistent oligo, PR3+ = proteinase 3 positive, PsA = psoriatic arthritis, RF +/- = polyarticular rheumatoid factor positive/negative, SLE = systemic lupus erythematosus, UC = ulcerative colitis.

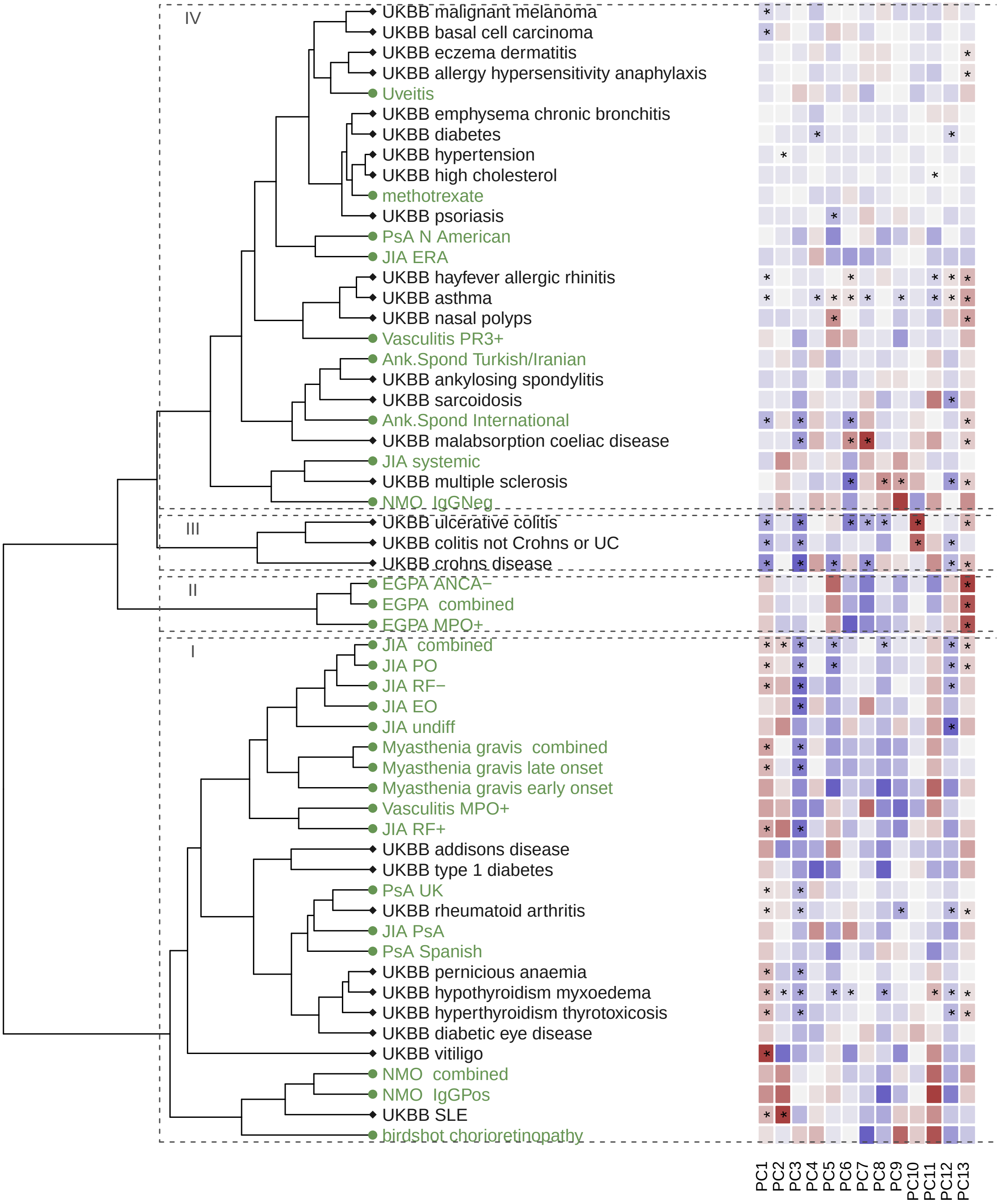


Fig. 5. Forest plots showing projected values for diseases significant overall and on components 1. Grey squares dots indicate projected data and 95% confidence intervals.

Red dots indicate the 13 IMD used for basis construction and for which no confidence interval is available. Points to the right of each line indicate disease classification according to whether they have specific autoantibodies that are either directly implicated in disease pathogenesis ("pathogenic") or which are specific to the disease, but not involved in pathogenesis ("non-pathogenic"). Diseases that are not associated with specific autoantibodies were classified as "none". Abbreviations: ANCA- = anti-neutrophil cytoplasmic antibody negative, Ank. Spond = ankylosing spondylitis, EGPA = eosinophilic granulomatosis with polyangiitis, EO = extended oligo, ERA = juvenile enthesitis-related arthritis, IgGPos = IgG positive, JIA = juvenile idiopathic arthritis, LADA = latent autoimmune diabetes in adults, NMO = neuromyelitis optica, PO = persistent oligo, PsA = psoriatic arthritis, RF +/- = polyarticular rheumatoid factor positive/negative, SLE = systemic lupus erythematosus, UC = ulcerative colitis.

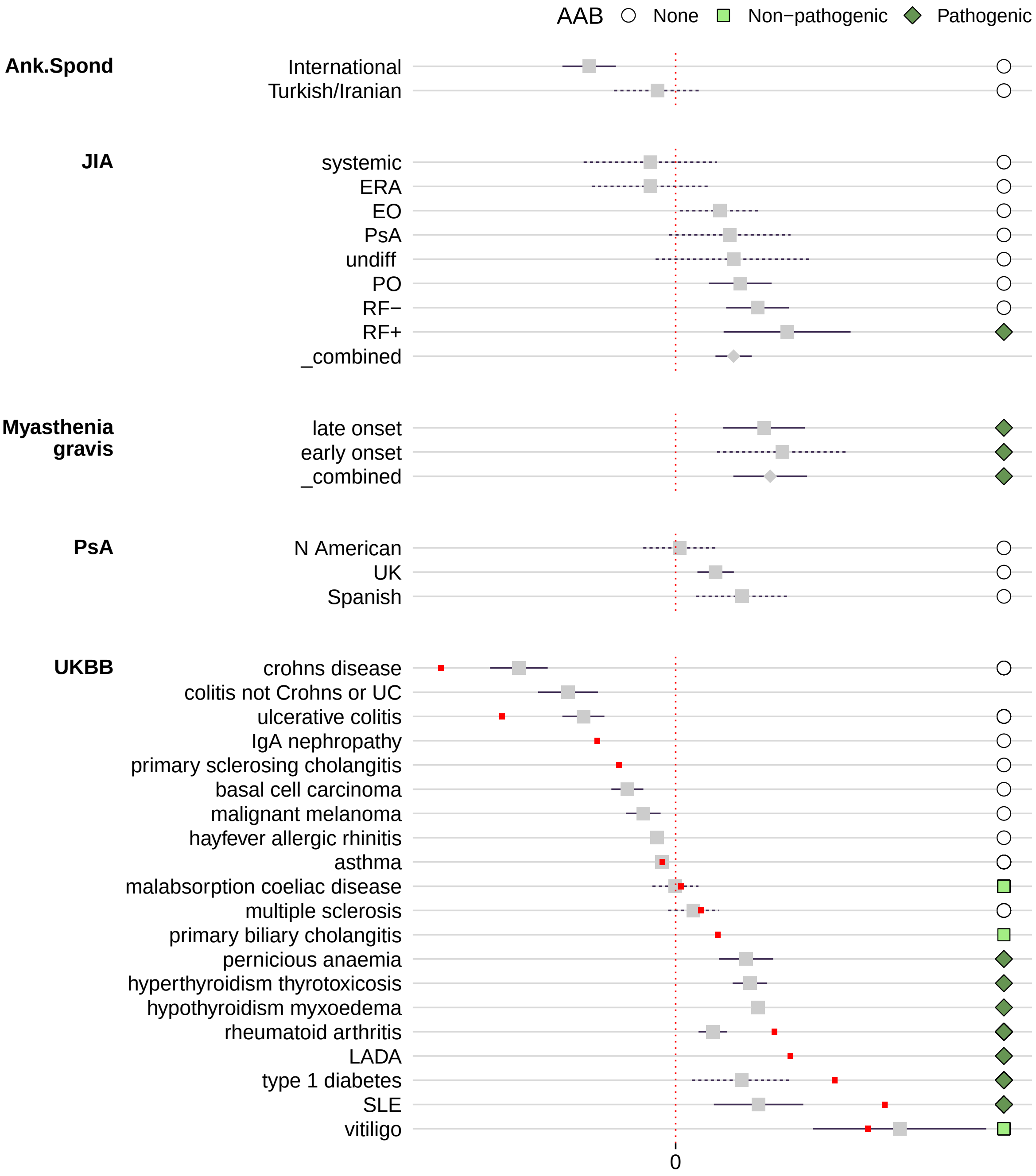


Fig. 6. Forest plot of significant traits on PC13 which also shows association with eosinophil counts in blood. Abbreviations: ANCA- = anti-neutrophil cytoplasmic antibody negative, Ank. Spond = ankylosing spondylitis, EGPA = eosinophilic granulomatosis with polyangiitis, JIA =

juvenile idiopathic arthritis, MPO+ = myeloperoxidase-positive, PO = persistent oligo, RF- = polyarticular rheumatoid factor negative.

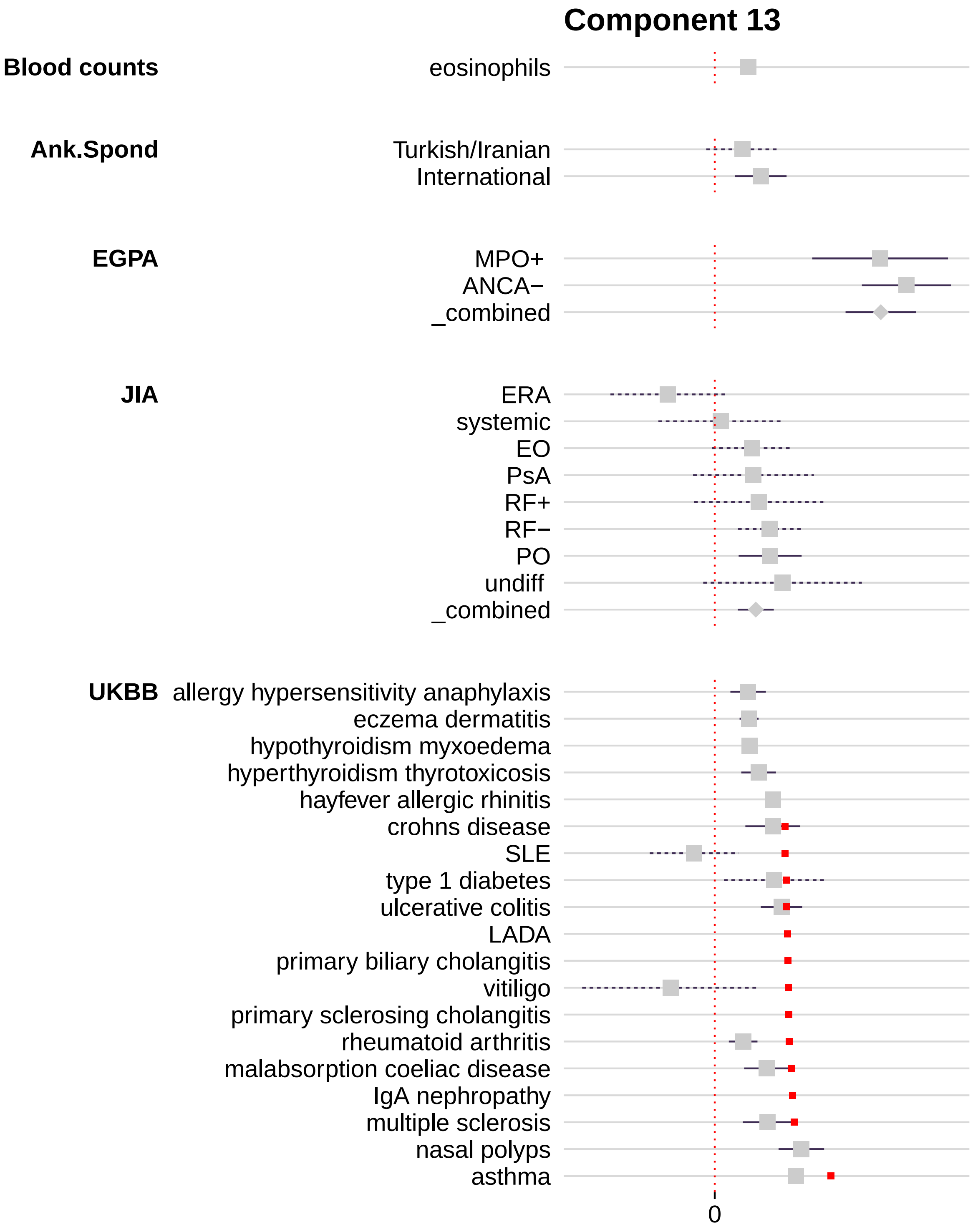
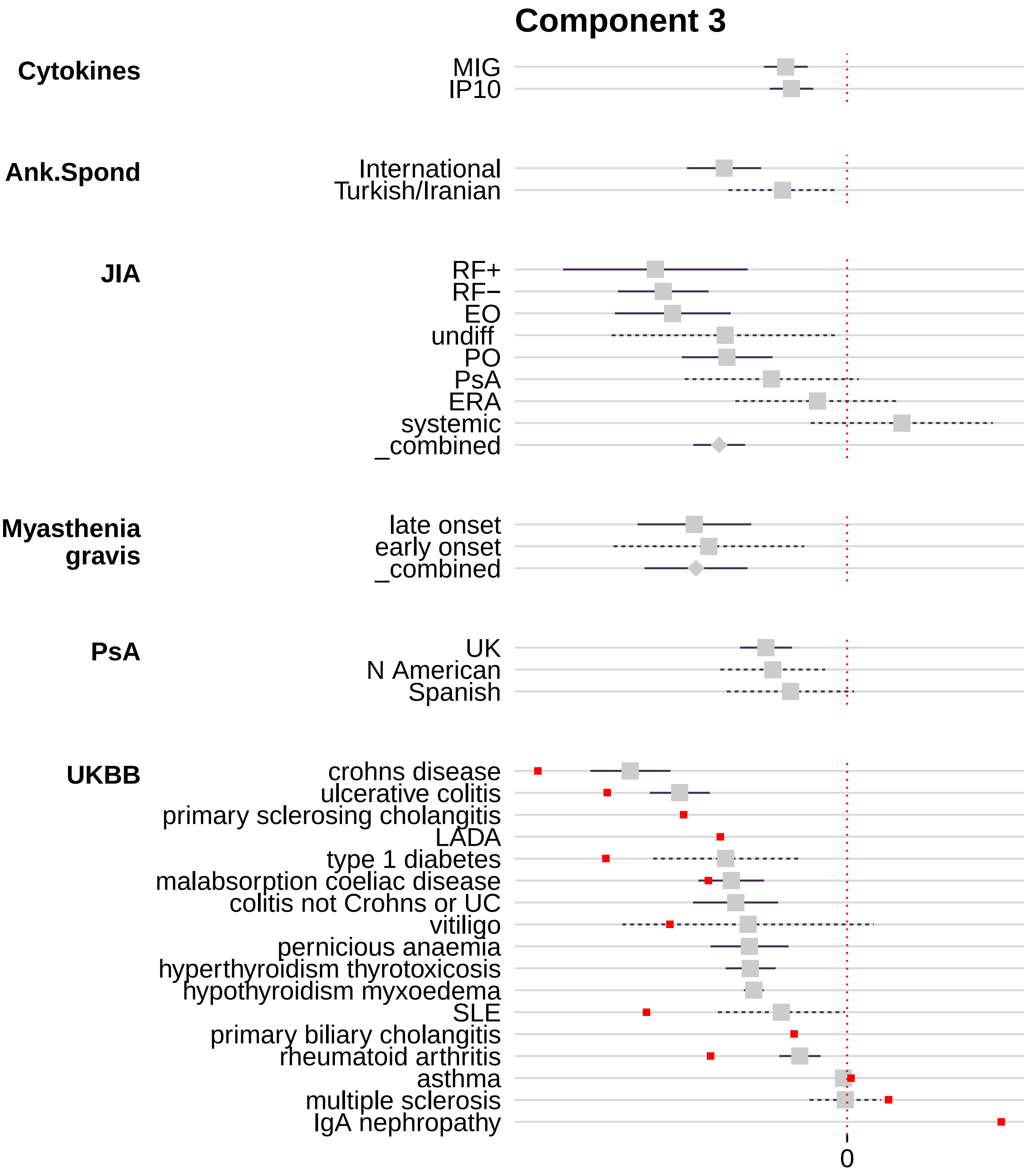


Fig. 7. Forest plot of significant traits on PC3 which also shows association with serum cytokine levels of IP-10 (CXCL10) and MIG (CXCL9). Abbreviations: EO = extended oligo, PO = persistent oligo, RF +/- = polyarticular rheumatoid factor positive/negative, UC = ulcerative colitis.


Component 3
Cytokines
MIG
IP10
Ank.Spond
International
Turkish/Iranian
JIA
RF+
RF−
EO
undiff
PO
PsA
ERA
systemic
_combined
Myasthenia gravis
late onset
early onset
_combined
PsA
UK
N American
Spanish
UKBB
crohns disease
ulcerative colitis
primary sclerosing cholangitis
LADA
type 1 diabetes
malabsorption coeliac disease
colitis not Crohns or UC
vitiligo
pernicious anaemia
hyperthyroidism thyrotoxicosis
hypothyroidism myxoedema
SLE
primary biliary cholangitis
rheumatoid arthritis
asthma
multiple sclerosis
IgA nephropathy
0