Pervasive Sharing of Genetic Effects in Autoimmune Disease

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

As Published
http://dx.doi.org/10.1371/journal.pgen.1002254

Publisher
Public Library of Science

Version
Final published version

Accessed
Tue Oct 02 13:27:43 EDT 2018

Citable Link
http://hdl.handle.net/1721.1/69081

Terms of Use
Creative Commons Attribution

Detailed Terms
http://creativecommons.org/licenses/by/2.5/
Pervasive Sharing of Genetic Effects in Autoimmune Disease


1 Center For Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 2 Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, United States of America, 3 Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America, 4 Department of Neurology, Yale University School of Medicine, New Haven, Connecticut, United States of America, 5 Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, United States of America, 6 Health Science and Technology MD Program, Harvard University and Massachusetts Institute of Technology, Boston, Massachusetts, United States of America, 7 Harvard Biological and Medical Sciences Program, Harvard University, Boston, Massachusetts, United States of America, 8 Pediatric Surgical Research Laboratories, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 9 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark, 10 Analytical and Translational Genetics Unity, Massachusetts General Hospital, Boston, Massachusetts, United States of America, 11 Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom, 12 Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, United States of America, 13 Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom, 14 Genentech, San Francisco, California, United States of America, 15 Departments of Medicine and Genetics, Yale University School of Medicine, New Haven, Connecticut, United States of America, 16 Department of Biophysics, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, 17 Department of Dermatology, University of Michigan, Ann Arbor, Michigan, United States of America, 18 Feinstein Institute for Medical Research, North Shore-Long Island Jewish Health System, Manhasset, New York, United States of America, 19 Rheumatology Unit, Department of Medicine, Karolinska Institutet, Stockholm, Sweden, 20 Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada, 21 Biizard Institute, The London School of Medicine and Dentistry, London, United Kingdom, 22 Department of Medical Genetics, University Medical Center Groningen and Groningen University, Groningen, The Netherlands, 23 Arthritis Research UK Epidemiology Unit, School of Translational Medicine, Manchester Academic Health Sciences Centre, University of Manchester, Manchester, United Kingdom, 24 Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, United States of America

Abstract

Genome-wide association (GWA) studies have identified numerous, replicable, genetic associations between common single nucleotide polymorphisms (SNPs) and risk of common autoimmune and inflammatory (immune-mediated) diseases, some of which are shared between two diseases. Along with epidemiological and clinical evidence, this suggests that some genetic risk factors may be shared across diseases—as is the case with alleles in the Major Histocompatibility Locus. In this work we evaluate the extent of this sharing for 107 immune disease-risk SNPs in seven diseases: celiac disease, Crohn’s disease, multiple sclerosis, psoriasis, rheumatoid arthritis, systemic lupus erythematosus, and type 1 diabetes. We have developed a novel statistic for Cross Phenotype Meta-Analysis (CPMA) which detects association of a SNP to multiple, but not necessarily all, phenotypes. With it, we find evidence that 47/107 (44%) immune-mediated disease risk SNPs are associated to multiple—but not all—immune-mediated diseases (SNP-wise PCMA<0.01). We also show that distinct groups of interacting proteins are encoded near SNPs which predispose to the same subsets of diseases; we propose these as the mechanistic basis of shared disease risk. We are thus able to leverage genetic data across diseases to construct biological hypotheses about the underlying mechanism of pathogenesis.

Citation: Cotsapas C, Voight BF, Rossin E, Lægreid PMA et al. (2011) Pervasive Sharing of Genetic Effects in Autoimmune Disease. PLoS Genet 7(8): e1002254. doi:10.1371/journal.pgen.1002254

Editor: Emmanuel T. Dermitzakis, University of Geneva Medical School, Switzerland

Received January 19, 2011; Accepted July 1, 2011; Published August 10, 2011

Copyright: © 2011 Cotsapas et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors received no specific funding for this article.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mjdal@atgu.mgh.harvard.edu

† These authors contributed equally to this work.

Introduction

The human immune-mediated diseases are the result of aberrant immune responses. These immune responses may lead to chronic inflammation and tissue destruction, often targeting a specific organ site. The outcome of this process is immune-mediated inflammatory and autoimmune disease, affecting approximately 5% of the population [1]. Extensive clinical and epidemiologic observations have shown that immune-mediated inflammatory and autoimmune diseases can occur either in the same individual or in closely related family members. This clustering of multiple diseases appears more
Author Summary

Over the last five years we have found over 100 genetic variants predisposing to common diseases affecting the immune system. In this study we analyze 107 such variants across seven diseases and find that almost half are shared across diseases. We also find that the patterns of sharing across diseases cluster these variants into groups; proteins encoded near variants in the same group tend to interact. This suggests that genetic variation may influence entire pathways to create risk to multiple diseases.

frequently than expected if disease processes were independent. As each of the immune-mediated inflammatory and autoimmune diseases has strong genetic influences on disease risk [2–7], the observed clustering of multiple diseases could be due to an overlap in the causal genes and pathways [8,9].

The patterns of clustering of diseases across the population are complex [10] – each disease has a prevalence between 0.01%–3%, so direct assessment of co-aggregation within individuals or families does not result in the very large samples required for genetic or epidemiological investigation. Thus it is unsurprising that to date, these observations have yet to be translated into determinants of the shared molecular etiologies of disease.

Recent GWA studies in immune-mediated and autoimmune diseases have identified 140 regions of the genome with statistically significant and robust evidence of presence of disease susceptibility loci. A subset of these loci have been shown to modulate risk of multiple diseases [3,6,11–14]. In addition, there is evidence that loci predisposing to one disease can have effects on risk of a second disease [15], although the risk allele for one disease may not be the same as for the second [16].

Together, these observations support the hypothesis of a common genetic basis of immune-mediated and autoimmune diseases [17]. There is now the ability to estimate both the number of loci contributing to risk of multiple diseases and the spectrum of diseases that each locus influences. In addition, grouping variants by the diseases they influence should provide insight into the specific biological processes underlying co-morbidity and disease risk.

In this report, we systematically investigate the genetic commonality in immune-mediated inflammatory and autoimmune diseases by examining the contributions of associated genomic risk regions in seven diseases: celiac disease (CeD), Crohn’s disease (CD), multiple sclerosis (MS), psoriasis (Ps), rheumatoid arthritis (RA), systemic lupus erythematosus (SLE) and type 1 diabetes (T1D). We find that nearly half of loci identified in GWAS studies of an individual disease influence risk to at least two diseases, arguing for a genetic basis to co-morbidity. We also find several variants with opposing risk profiles in different diseases. Supporting the idea that common patterns of association implicate shared biological processes, we further demonstrate that loci clustered by the pattern of diseases they affect harbor genes encoding interacting proteins at a much higher rate than by chance. These results suggest that multi-phenotype mapping will identify the molecular mechanisms underlying co-morbid immune-mediated inflammatory and autoimmune diseases.

Results

We first test our hypothesis of common genetic determinants by examining evidence of association of genetic variants in known immune-mediated and autoimmune disease susceptibility loci to multiple disease phenotypes. We collated a list of 140 single nucleotide polymorphisms (SNPs) representing reported associations to at least one immune-mediated disease at genome-wide significance levels. Where data for the reported SNP itself were not available in our GWA studies (Table 1), we chose a proxy in high linkage disequilibrium to the reported marker (\(r^2>0.9\) in HapMap/CEU). We did not consider SNPs in the human Major Histocompatibility Complex (MHC) from this analysis, as its role in many of these diseases is well-established and the classically associated alleles in the HLA region are not well captured by SNPs [18]. We were able to acquire data for either the reported SNP or a good proxy in 107 of 140 cases, and assembled genotype test summaries for these from previously described GWA studies representing over 26,000 disease cases (Table 1).

We have developed a cross-phenotype meta-analysis (CPMA) statistic to assess association across multiple phenotypes. The CPMA statistic determines evidence for the hypothesis that each independent SNP has multiple phenotypic associations. Support for this hypothesis would be shown by deviations from expected uniformity of the distribution of association \(p\)-values, indicative of multiple associations. The likelihood of the observed rate of exponential decay of \(-\log_{10}(p)\) is calculated and compared to the null expectation (the decay rate should be unity) as a likelihood ratio test (see Materials and Methods for details). This CPMA statistic has one degree of freedom, as it measures a deviation in \(p\)-value behavior instead of testing all possible combinations of diseases for association to each SNP.

A total of 47 of the 107 SNPs tested have evidence of association to multiple diseases (SNP-wise \(P_{\text{CPMA}}<0.01\); expectation roughly 1 by chance; binomial probability of observing this result \(p=3 \times 10^{-64}\)). This highly significant result confirms widespread sharing of genetic loci between immune-mediated inflammatory and autoimmune diseases. Further, these “multi-phenotype” SNPs include many loci not previously known to be shared across diseases, as well as new predictions of association for previously known shared loci (Table 2).

Although our CPMA statistic is agnostic to effect direction in each disease, a subset of the 47 multi-phenotype (CPMA positive) SNPs appeared to have strong allelic effects in opposite directions in different diseases [16]. A total of 9 SNPs had strong evidence of such directional association (an association \(p<1 \times 10^{-4}\) with at least one protective and one risk effect; lower panel in Table 2). This suggests that shared associations have complex effects on disease outcomes and may be of particular importance in pathogenic processes.

Table 1. Participating studies.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Cases</th>
<th>Controls</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celiac disease</td>
<td>3796</td>
<td>8154</td>
<td>22</td>
</tr>
<tr>
<td>Crohn’s disease</td>
<td>3230</td>
<td>4829</td>
<td>1</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>2624</td>
<td>7220</td>
<td>4</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>1359</td>
<td>1400</td>
<td>5</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>5539</td>
<td>20169</td>
<td>6</td>
</tr>
<tr>
<td>Systemic Lupus Erythematosus</td>
<td>1963</td>
<td>4329</td>
<td>23</td>
</tr>
<tr>
<td>Type 1 diabetes</td>
<td>7514</td>
<td>9045</td>
<td>24</td>
</tr>
</tbody>
</table>

Data were collated for seven phenotypes from meta-analyses incorporating all known genome-wide association studies. SLE is the exception as no comprehensive meta-analysis has yet been published; data were instead obtained from a recent meta-analysis including some, but not all, known genome-wide association studies. Note that controls overlap in some cases due to the use of common shared sample genotypes.

doi:10.1371/journal.pgen.1002254.001
<table>
<thead>
<tr>
<th>SNP</th>
<th>Name</th>
<th>Gene</th>
<th>RA</th>
<th>Psoriasis</th>
<th>MS</th>
<th>SLE</th>
<th>Crohn</th>
<th>Coeliac</th>
<th>T1D</th>
<th>CPMA</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>rs10889677</td>
<td>IL23R</td>
<td>5.1</td>
<td>7.0</td>
<td>0.3</td>
<td>6.2</td>
<td>0.2</td>
<td>8.7</td>
<td>8.7</td>
<td>3.6</td>
<td>5.06E-05</td>
</tr>
<tr>
<td>2</td>
<td>rs20498118</td>
<td>IL23R</td>
<td>2.1</td>
<td>2.2</td>
<td>0.7</td>
<td>2.2</td>
<td>0.7</td>
<td>2.2</td>
<td>2.2</td>
<td>0.7</td>
<td>2.2E-01</td>
</tr>
<tr>
<td>2</td>
<td>rs11209932</td>
<td>IL23R</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1E-05</td>
</tr>
<tr>
<td>1</td>
<td>rs10889677</td>
<td>IL23R</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1E-05</td>
</tr>
<tr>
<td>1</td>
<td>rs20498118</td>
<td>IL23R</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1E-01</td>
</tr>
<tr>
<td>1</td>
<td>rs11209932</td>
<td>IL23R</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1E-05</td>
</tr>
<tr>
<td>1</td>
<td>rs13281173</td>
<td>IL23R</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1</td>
<td>5.1E-05</td>
</tr>
<tr>
<td>18</td>
<td>rs2542151</td>
<td>PTPN2</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5E-05</td>
</tr>
<tr>
<td>18</td>
<td>rs2542151</td>
<td>PTPN2</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5E-05</td>
</tr>
<tr>
<td>18</td>
<td>rs1279243</td>
<td>PTPN2</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5E-05</td>
</tr>
<tr>
<td>18</td>
<td>rs1279243</td>
<td>PTPN2</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5</td>
<td>3.5E-05</td>
</tr>
</tbody>
</table>

**Table 2:** SNPs associated with multiple phenotypes.

**Shared Genetic Effects in Autoimmune Disease**
47/107 SNPs tested showed significant evidence of association to multiple diseases ($P_{	ext{corr}} < 0.01$), where only one is expected by chance. These SNPs are therefore candidate drivers for the shared genetic architecture between diseases. The SNPs shown in the lower panel also have strong evidence of association in opposite directions across phenotypes and may be crucial decision points in pathogenesis. $Aj = \text{major allele}; Am = \text{minor allele}$. Z scores are reported from published GWA studies and arbitrarily signed relative to the direction of effect in celiac disease. Note that no MS data were available for rs2082412 as it had not been imputed accurately in the participating MS study.

Data for all SNPs is presented in Dataset S1. doi:10.1371/journal.pgen.1002254.t002

<table>
<thead>
<tr>
<th>SNP</th>
<th>Chr</th>
<th>Position</th>
<th>Aj</th>
<th>Am</th>
<th>Genes</th>
<th>RA</th>
<th>Psoriasis</th>
<th>MS</th>
<th>SLE</th>
<th>Crohn</th>
<th>Coeliac</th>
<th>T1D</th>
<th>CPMA</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs7804356</td>
<td>7</td>
<td>26644905</td>
<td>T</td>
<td>C</td>
<td>7p15.2</td>
<td>−1.3</td>
<td>1.9E-01</td>
<td>1.2</td>
<td>2.2E-01</td>
<td>1.0</td>
<td>1.5E-01</td>
<td>−0.1</td>
<td>4.8E-01</td>
<td>3.4</td>
</tr>
<tr>
<td>rs11258747</td>
<td>10</td>
<td>6512897</td>
<td>G</td>
<td>T</td>
<td>PRKCQ</td>
<td>−1.0</td>
<td>3.0E-01</td>
<td>−2.4</td>
<td>1.7E-02</td>
<td>0.1</td>
<td>5.4E-01</td>
<td>−3.4</td>
<td>3.2E-04</td>
<td>0.6</td>
</tr>
<tr>
<td>rs2476601</td>
<td>5</td>
<td>114089610</td>
<td>G</td>
<td>A</td>
<td>PTN2</td>
<td>14.2</td>
<td>9.1E-74</td>
<td>0.0</td>
<td>1.0E+00</td>
<td>0.4</td>
<td>3.5E-01</td>
<td>4.0</td>
<td>3.3E-05</td>
<td>−4.3</td>
</tr>
<tr>
<td>rs3184504</td>
<td>12</td>
<td>110347328</td>
<td>T</td>
<td>C</td>
<td>SH2B3</td>
<td>−2.9</td>
<td>3.6E-03</td>
<td>−2.0</td>
<td>4.1E-02</td>
<td>3.4</td>
<td>3.3E-04</td>
<td>−2.7</td>
<td>3.6E-03</td>
<td>−3.4</td>
</tr>
<tr>
<td>rs11865121</td>
<td>16</td>
<td>11074189</td>
<td>C</td>
<td>A</td>
<td>CLEC16A</td>
<td>0.3</td>
<td>7.7E-01</td>
<td>0.6</td>
<td>5.2E-01</td>
<td>4.3</td>
<td>8.7E-06</td>
<td>−0.7</td>
<td>2.3E-01</td>
<td>0.9</td>
</tr>
<tr>
<td>rs2816316</td>
<td>1</td>
<td>18928470</td>
<td>A</td>
<td>C</td>
<td>RGS1</td>
<td>−1.0</td>
<td>3.3E-01</td>
<td>−0.4</td>
<td>6.7E-01</td>
<td>3.1</td>
<td>9.0E-04</td>
<td>−0.1</td>
<td>5.3E-01</td>
<td>−0.5</td>
</tr>
<tr>
<td>rs1042869</td>
<td>1</td>
<td>20332899</td>
<td>G</td>
<td>A</td>
<td>IL10</td>
<td>−3.1</td>
<td>1.8E-03</td>
<td>−1.1</td>
<td>9.5E-01</td>
<td>6.2</td>
<td>3.5E-10</td>
<td>0.4</td>
<td>6.5E-01</td>
<td>−0.8</td>
</tr>
<tr>
<td>rs3024505</td>
<td>1</td>
<td>15874711</td>
<td>C</td>
<td>A</td>
<td>IL2RA</td>
<td>−1.0</td>
<td>3.3E-01</td>
<td>0.9</td>
<td>3.8E-01</td>
<td>1.5</td>
<td>7.3E-02</td>
<td>4.2</td>
<td>1.3E-05</td>
<td>2.4</td>
</tr>
<tr>
<td>rs10045431</td>
<td>5</td>
<td>138241110</td>
<td>T</td>
<td>G</td>
<td>TNFAIP3</td>
<td>−4.2</td>
<td>3.3E-05</td>
<td>4.5</td>
<td>8.0E-06</td>
<td>0.3</td>
<td>6.3E-01</td>
<td>2.4</td>
<td>8.8E-03</td>
<td>−5.8</td>
</tr>
<tr>
<td>rs61064</td>
<td>6</td>
<td>132441110</td>
<td>T</td>
<td>G</td>
<td>TNFAIP3</td>
<td>−4.2</td>
<td>3.3E-05</td>
<td>4.5</td>
<td>8.0E-06</td>
<td>0.3</td>
<td>6.3E-01</td>
<td>2.4</td>
<td>8.8E-03</td>
<td>−5.8</td>
</tr>
<tr>
<td>rs4613763</td>
<td>1</td>
<td>40428465</td>
<td>T</td>
<td>C</td>
<td>PTGER4</td>
<td>0.9</td>
<td>3.9E-01</td>
<td>0.7</td>
<td>4.7E-01</td>
<td>−4.2</td>
<td>1.1E-05</td>
<td>0.4</td>
<td>3.4E-01</td>
<td>9.6</td>
</tr>
</tbody>
</table>
We next examined the patterns of association across 47 multi-phenotype SNPs to determine evidence of either a global autoimmune process or biological pathways influencing sets of diseases. On visual inspection of these data we found a striking patterning of associations across diseases: only one SNP (rs3184504, in an exon of SH2B3) exhibited evidence of association to all seven diseases; the others appeared to associate only to subsets of diseases (Table 2).

To formalize the analysis of association patterns across diseases, we determined specific patterns of associations across SNPs by computing SNP-SNP distances based on the level of association to each disease followed by hierarchical clustering to group them (Figure 1A; see Materials and Methods section for clustering details). SNPs in loci encoding proteins known to interact clustered together: for example, the independent effects at IL12B and IL23R, which encode subunits of a ligand-receptor pair are in the same region of the dendrogram. We next partitioned the dendrogram “tree” into four clusters and summarized the cumulative association of each cluster to each disease by combining our underlying dataset of association p-values per cluster, per disease using Fisher’s omnibus test (Figure 1B; see Materials and Methods for details). Each cluster had a different pattern of associations across diseases; these patterns suggest that the clusters represent distinct co-morbid mechanisms.

Our underlying hypothesis has been that phenotype-driven clusters represent distinct molecular mechanisms. This leads to the prediction that components of these clusters/pathways are encoded in associated loci; in other words, proteins encoded around SNPs in the same cluster should interact. We test this prediction by looking for connectivity between proteins encoded around SNPs within each cluster as described elsewhere [19]. Briefly, we define a genomic region around each SNP in terms of linkage disequilibrium and consider any protein overlapping that region. We then ask if proteins encoded around SNPs in the same cluster interact using protein-protein interaction maps, excluding interactions between proteins in the same region (see Materials and Methods and [19]). We find that three of the four clusters we define by patterns of association have significant connectivity (Figure 1C; permuted P<0.05) by this method, suggesting that these represent distinct molecular mechanisms affected by genetic risk variants. Two of these groups of interacting proteins are also preferentially expressed [19] in immune cell subtypes compared to other tissue types (Figure S1), supporting our hypothesis that these represent true pathways underlying pathogenesis.

Discussion

Immune-mediated inflammatory and autoimmune diseases have been known to cluster in families, suggesting a strong genetic component to risk. The genes in the human MHC (HLA complex) have been associated with disease risk, suggesting a common immune pathway. Less clear is whether other genetic variants associated with individual diseases also form common pathways/mechanisms for autoimmunity. Recent results from GWA studies suggest that common genetic mechanisms may underlie the observed clustering of multiple autoimmune diseases within a person or family. In this work we have tested the hypothesis that immunologically relevant genetic variation will either (1) underlie risk to all immune-mediated diseases, implicating a global immunological process; (2) influence risk to a discrete subset of diseases, implicating molecular entities underlying that co-morbidity; or (3) modulate risk for only one disorder thereby implying a disease-specific process.

A central goal of complex disease genetics is to uncover the pathways perturbed in disease and shed light on the underlying
cellular processes. Despite a wealth of molecular insight into immune function few key pathways underlying genetic susceptibility to immune-mediated diseases have been elucidated. To identify these processes in immune-mediated inflammatory and autoimmune disease, we tested genetic variation contributing to seven diseases. We observed an overwhelming abundance of commonality across these phenotypes, assorting into cohesive phenotype-genotype groups that appear to underlie co-morbidities. By analyzing loci known to associate to at least one disease, we are able to identify groups of diseases that should be considered as a unified phenotype and analyzed together. We further demonstrate that this approach generates novel biological insights into pathogenesis, often difficult to obtain from genomic studies of single traits [20].

We have described a novel statistic, CPMA, which assesses evidence for multiple associations to a marker. Rather than perform a meta-analysis, which would only detect association to all phenotypes (or suffer from heterogeneity) or test all combinations of phenotypes which would increase the multiple testing burden, we look for deviation in the distribution of association \( p \) values. Our statistic thus detects markers associated to at least some, but not necessarily all, phenotypes; we note that this is a single degree of freedom test, providing high power to reject the null hypothesis. This power comes at the price of not knowing to which phenotypes the marker is associate; we overcome this with our clustering analysis, which resolves groups of markers associating to the same diseases. Thus our analytic strategy is able to both detect shared associations and identify the relevant phenotypes.

Our approach appears capable of distinguishing distinct genetic effects in the same locus in addition to validated shared associations. For example, it is now clear that the two signals in the IL2/IL21 locus on chromosome 4q27 are distinct, with T1D mapping to IL2 and other diseases to IL21 [21]. Our analysis detects this difference, clustering the two SNPs representing these associations separately (Figure 1, labeled “IL2” and “IL2/IL21”, respectively). Conversely, previous reports of an overlap in associations separately (Figure 1, labeled “IL2” and “IL2/IL21”, respectively). Our CPMA analysis relies on the expected distribution of \( p \) values for each SNP across diseases. Under the null hypothesis of no additional associations beyond those already known, we expect association values to be uniformly distributed and hence \(-ln(p)\) to be exponentially decaying with a decay rate \( \lambda = 1 \). We calculate the likelihood of the observed and expected values of \( p \) and express these as a likelihood ratio test:

\[
CPMA = -2 \times \frac{P[Data|\lambda = 1]}{P[Data|\lambda = 2]}
\]

This statistic therefore measures the likelihood of the null hypothesis given the data; we can reject the null hypothesis if sufficient evidence to the contrary is present. We note that, because we only estimate a single parameter, our test is asymptotically distributed as \( \chi^2_{-1} \). This gives us more statistical power than relying on strategies combining association statistics, which would consume multiple degrees of freedom.

**SNP–SNP distance calculation and clustering**

To compare the patterns of association for multi-phenotype SNPs we first calculate SNP-SNP distances and then use hierarchical clustering on that distance matrix to assess relative relationships between SNP association patterns.

Calculating distances based directly on \( p \) values or the underlying association statistics is problematic, as each contributing study has slightly different sample sizes and therefore different statistical power to detect associations. Thus, distance functions based on numeric data – which incorporate magnitude differences between observations – would be biased if studies have systematically different data. Normalization procedures can account for such systematic differences but may fail to remove all bias. To reduce the impact such systematic irregularities might have on our comparison, we bin associations into informal “levels of evidence” categories. We define four classes \((1 < p < 0.05, 0.05 < p < 0.001, 0.001 < p < 1 \times 10^{-8}, 1 \times 10^{-8} < p)\) and thus reduce our data to a SNP x disease matrix...
where entries are categorical variables describing these classes. We then calculate the Euclidean distances between pairs of SNPs using Gower’s method for categorical data [27], which accounts for the discrete nature of the data.

To compare the distance relationships between SNPs we use hierarchical agglomerative clustering. This process joins single entities (in this case, SNPs) or groups of entities together if certain criteria are met. Successive rounds of clustering are performed in an iterative way until all groups are joined, resulting in a tree of relationships where similar entities cluster on the same branches.

In this analysis we cluster SNPs based on the Gower distance matrix using Ward’s method for joining entities [28]. In contrast to linkage clustering methods, Ward’s method seeks to minimize the information lost during the clustering process, calculated as the error sum of squares (ESS). The higher the ESS the more information is being lost due to inaccuracy of grouping entities together. This method thus seeks compact, spherical clusters of data which are maximally similar.

All distance and clustering analysis was done using the StatMatch and stats packages in the R programming language [29].

### Cumulative association statistics

We compute per-cluster, per-disease cumulative association statistics by combining \( p \) values using Fisher’s omnibus test, where the cumulative statistic \( S_{cum} \) on \( N \) \( p \)-values is defined as:

\[
S_{cum} = -2 \times \sum_{i=1}^{N} \ln p_i
\]

and \( S_{cum} \) follows the \( \chi^2 \) distribution with \( 2N \) degrees of freedom.

### Protein–protein interaction analysis

We use previously described methodology [19] to assess whether proteins encoded around SNPs in each cluster interact. Briefly, we first compile lists of all proteins that an association may affect by defining locus boundaries around each SNP in terms of linkage disequilibrium and including all proteins overlapping this region. We then use a high-confidence protein-protein interaction map [30] as modified in [19] to ask whether proteins encoded around SNPs in each cluster interact either directly or via a common intermediary and assess the significance of such observations relative to the local structure of the protein-protein network as described elsewhere [19], using 4000 permutations. These data and methodology are publicly available for download and via a webserver (http://www.broadinstitute.org/mpg/dapple).

### Supporting Information


**Figure S1** Enrichment in immune tissue expression for interacting genes encoded close to SNPs in (A) cluster 1 and (B) cluster 4. Following Rossin et al.[19] we looked for preferential expression of significant network genes in tissue subsets. Of the genes encoded around SNPs in clusters 1 and 4 as defined in Figure 1, we found that those participating in significant networks are enriched in expression (purple circles) in immune tissues (red bars). Other genes encoded around those SNPs are not enriched in the same tissues (black circles). Thus interacting genes encoded around SNPs associated to the same immune diseases are preferentially expressed in immune tissues. Interacting genes for the remaining significant group, cluster 2, were not enriched. (PDF)

**Author Contributions**


### References


