



## Pervasive Sharing of Genetic Effects in Autoimmune Disease

**Cotsapas, Chris; Voight, Benjamin F.; Rossin, Elizabeth; Hansen, Kasper Lage; Neale, Benjamin M.; Wallace, Chris; Abecasis, Goncalo R.; Barrett, Jeffrey C.; Behrens, Timothy; Cho, Judy**

*Total number of authors:*  
23

*Published in:*  
P L o S Genetics

*Link to article, DOI:*  
[10.1371/journal.pgen.1002254](https://doi.org/10.1371/journal.pgen.1002254)

*Publication date:*  
2011

*Document Version*  
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

### *Citation (APA):*

Cotsapas, C., Voight, B. F., Rossin, E., Hansen, K. L., Neale, B. M., Wallace, C., Abecasis, G. R., Barrett, J. C., Behrens, T., Cho, J., De Jager, P. L., Elder, J. T., Graham, R. R., Gregersen, P., Klareskog, L., Siminovitch, K. A., van Heel, D. A., Wijmenga, C., Worthington, J., ... Daly, M. J. (2011). Pervasive Sharing of Genetic Effects in Autoimmune Disease. *P L o S Genetics*, 7(8). <https://doi.org/10.1371/journal.pgen.1002254>

---

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Pervasive Sharing of Genetic Effects in Autoimmune Disease

Chris Cotsapas<sup>1,2,3,4,5,9</sup>, Benjamin F. Voight<sup>1,2,3,9</sup>, Elizabeth Rossin<sup>1,2,3,6,7</sup>, Kasper Lage<sup>2,8,9</sup>, Benjamin M. Neale<sup>1,2,3,10</sup>, Chris Wallace<sup>11</sup>, Gonçalo R. Abecasis<sup>12</sup>, Jeffrey C. Barrett<sup>13</sup>, Timothy Behrens<sup>14</sup>, Judy Cho<sup>5,15</sup>, Philip L. De Jager<sup>3,16</sup>, James T. Elder<sup>17</sup>, Robert R. Graham<sup>14</sup>, Peter Gregersen<sup>18</sup>, Lars Klareskog<sup>19</sup>, Katherine A. Siminovitch<sup>20</sup>, David A. van Heel<sup>21</sup>, Cisca Wijmenga<sup>22</sup>, Jane Worthington<sup>23</sup>, John A. Todd<sup>11</sup>, David A. Hafler<sup>4</sup>, Stephen S. Rich<sup>24</sup>, Mark J. Daly<sup>1,2,3,10\*</sup>, on behalf of the FOCiS Network of Consortia

**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 Biomedical 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, Massachusetts, United States of America, **13** Human Genetics, Wellcome Trust Sanger Institute, Cambridge, United Kingdom, **14** Genentech, South 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 Neurology, 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** Blizard Institute, The London School of Medicine and Dentistry, London, United Kingdom, **22** Department of 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  $P_{CPMA} < 0.01$ ). 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, Lage K, Neale BM, et al. (2011) Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLoS Genet* 7(8): e1002254. doi:10.1371/journal.pgen.1002254

**Editor:** Emmanouil T. Dermizakis, 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: mjdaly@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  $P_{CPMA} < 0.01$ ; 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.

Disease	Cases	Controls	Reference
Celiac disease	3796	8154	22
Crohn's disease	3230	4829	1
Multiple sclerosis	2624	7220	4
Psoriasis	1359	1400	5
Rheumatoid arthritis	5539	20169	6
Systemic Lupus Erythematosus	1963	4329	23
Type 1 diabetes	7514	9045	24

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.t001



**Table 2. Cont.**

SNP	Name	Chr	Position	Aj	Am	Genes	RA		Psoriasis		MS		SLE		Crohn		Coeliac		T1D		CPMA		Reference
							Z	p	Z	p	Z	p	Z	p	Z	p	Z	p	Z	p	Z	p	
rs7804356	7	26664905	T	C	7p15.2		-1.3	1.9E-01	1.2	2.2E-01	1.0	1.5E-01	-0.1	4.8E-01	3.4	5.9E-04	1.3	9.9E-02	-5.6	9.5E-09	5.6E-03	24	
rs11258747	10	6512897	G	T	PRKCQ		3.0	2.4E-03	1.9	6.0E-02	-2.5	6.7E-03	0.6	7.3E-01	0.1	9.2E-01	0.2	4.2E-01	4.8	7.7E-07	7.5E-03	24	
rs703842	12	56449006	A	G	CYP27B1		-2.7	6.1E-03	-0.1	8.9E-01	4.1	1.7E-05	-0.4	6.7E-01	-1.1	2.5E-01	2.0	2.3E-02	-2.2	1.5E-02	8.3E-03	31	
rs1990760	2	162949558	T	C	IFIH1		-1.0	3.0E-01	-2.4	1.7E-02	0.1	5.4E-01	-3.4	3.2E-04	0.6	5.6E-01	0.4	6.5E-01	-6.2	2.5E-10	8.6E-03	24	
rs2476601	1	114089610	G	A	PTPN22		18.2	9.1E-74	0.0	1.0E+00	0.4	3.5E-01	4.0	3.3E-05	-4.3	1.8E-05	1.7	4.2E-02	20.4	1.5E-92	6.3E-160	1	
rs3184504	12	110347328	T	C	SH2B3		-2.9	3.6E-03	-2.0	4.1E-02	3.4	3.3E-04	-2.7	3.6E-03	-3.4	6.2E-04	7.3	1.2E-13	-11.9	7.7E-33	4.3E-19	24	
rs11865121	16	11074189	C	A	CLEC16A		0.3	7.7E-01	0.6	5.2E-01	4.3	8.7E-06	-0.7	2.3E-01	0.9	3.8E-01	1.1	1.4E-01	-8.9	2.1E-19	1.1E-14	4	
rs2816316	1	189268470	A	C	RGS1		-1.0	3.3E-01	-0.4	6.7E-01	3.1	9.0E-04	-0.1	5.3E-01	-0.5	6.4E-01	6.9	2.7E-12	-3.9	4.2E-05	5.2E-13	3	
rs2104286	10	6139051	T	C	IL2RA		-3.1	1.8E-03	-0.1	9.5E-01	6.2	3.5E-10	0.4	6.5E-01	-0.8	4.4E-01	0.5	3.1E-01	-6.4	5.9E-11	1.2E-08	32	
rs3024505	1	203328299	G	A	IL10		-1.0	3.3E-01	0.9	3.8E-01	1.5	7.3E-02	4.2	1.3E-05	2.4	1.6E-02	1.6	5.4E-02	-4.8	6.2E-07	2.2E-06	23	
rs10045431	5	158747111	C	A	IL12B		0.4	6.5E-01	4.5	6.6E-06	0.3	6.3E-01	2.4	8.8E-03	-5.8	8.8E-09	0.5	6.8E-01	0.1	4.6E-01	6.0E-04	1	
rs610604	6	138241110	T	G	TNFAIP3		-4.2	3.3E-05	4.5	8.0E-06	0.3	3.8E-01	-1.3	9.9E-02	1.4	1.8E-01	0.4	6.5E-01	-0.2	4.3E-01	2.7E-03	5	
rs4613763	5	40428485	T	C	PTGER4		0.9	3.9E-01	0.7	4.7E-01	-4.2	1.1E-05	0.4	3.4E-01	9.6	5.0E-22	0.1	5.2E-01	-0.5	3.1E-01	4.0E-03	1	

47/107 SNPs tested showed significant evidence of association to multiple diseases ( $P_{\text{omni}} < 0.01$ ), 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.  $A_j$  = major allele;  $A_m$  = 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

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.

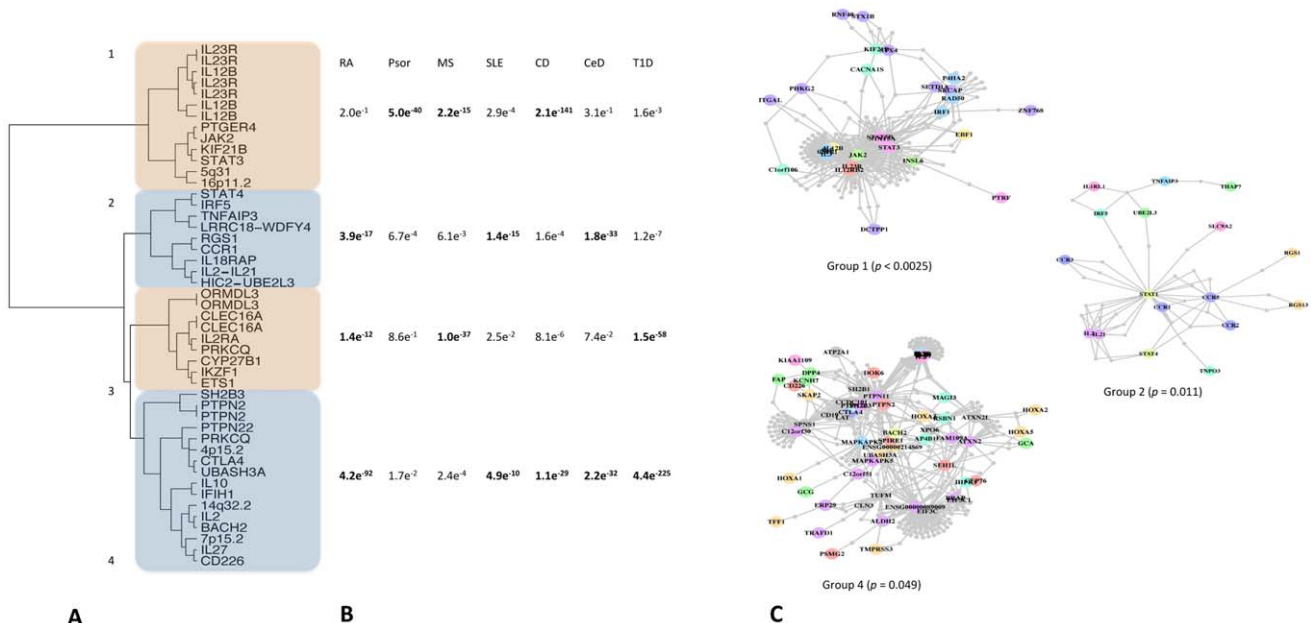
Our underlying hypothesis has been that phenotype-driven clusters represent distinct molecular mechanisms. This leads to the prediction that components of these clusters/pathways are encoded in associated loci; in other words, proteins encoded around SNPs in the same cluster should interact. We test 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



**Figure 1. Patterns of association across diseases correlate with protein-protein interactions.** A: 47 SNPs with evidence of association to multiple diseases ( $P_{cpma} < 0.01$ ) fall into groups clustered by the pattern of association across diseases. Clusters are numbered arbitrarily. B: Clusters show different patterns of association across diseases. We summarize the differential disease effects of each cluster with a cumulative association statistic (Fisher’s method for combining *p* values). These patterns are different for each cluster, suggesting each represents a different co-morbid mechanism. Note that these figures are based on the same underlying association statistics the clustering in the first panel is derived from. C: proteins encoded within the linkage disequilibrium scope around SNPs in the same cluster interact either directly or via common intermediates. Three of our four clusters have significant protein inter-connectivity (permuted  $P < 0.05$ ; see Materials and Methods and [19] for details). doi:10.1371/journal.pgen.1002254.g001

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 associated; 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 association between T1D and celiac disease [15] were in regions encoding genes highly expressed in T lymphocytes (*RGS1*, *PTPN2* and *CTLA4* in celiac; *PTPN2* and *CTLA4* in T1D). Our analysis identifies all these regions as CPMA-positive and highlights the second associations in T1D and celiac shown by Smyth *et al.* [15], indicating that our approach could be used to prioritize marginal associations for replication. We also observe other potential associations. For example, rs2816316 on near *RGS1* exhibits evidence of association to MS; rs2542151 and rs1893217 on near *PTPN2* has modest association to psoriasis. These last observations, whilst suggestive, require further investigation given the known effects of these regions on other diseases.

In summary, our multi-disease approach is applicable beyond the immune-mediated inflammatory and autoimmune diseases, to current studies of related traits in pharmacology, metabolic and psychiatric disease and in genetic studies of cellular phenotypes such as gene expression. For most studies of the genetic basis of complex human phenotypes, the pathogenic processes are still far from understood and biological pathways may be identified using these methods. Ultimately, these results will contribute to an improved molecular nosology of mechanistic definitions and, ultimately, towards improving clinical care and human health.

## Materials and Methods

### Ethics statement

All data were drawn from previously published genome-wide association studies from consortia with appropriate ethics oversight from their respective institutional review boards. As only summary

data from a small number of markers across the genome were used here no further ethical issues arise.

### Patient cohorts

Data were obtained from previously described case/control GWA studies of celiac disease [22], Crohn’s disease [2], multiple sclerosis [5], psoriasis [6], rheumatoid arthritis [7], systemic lupus erythematosus [23] and type I diabetes [24] as shown in Table 1. We note that, with the exception of psoriasis, in these cohorts diagnosis of a second immune-mediated disease is a criterion for exclusion, thereby minimizing co-morbidity as a source of bias in our study.

### Locus selection

For our analysis we selected 140 independent SNPs ( $r^2 < 0.2$ ) with reported associations to an immune-mediated disease in a genome-wide association scan and replicated in independent samples in that disease to combined genome-wide significance [25]. We then chose proxies for those SNPs present on the major versions of Affymetrix and Illumina genome-wide genotyping platforms [26]; 107 SNPs had sufficient data coverage to be included. Where possible we used the SNP originally reported; if data were unavailable for that marker, we chose a high LD proxy (HapMap/CEU  $r^2 > 0.9$ ) to represent the region.

### Cross-phenotype meta-analysis

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  $\lambda$  and express these as a likelihood ratio test:

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

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  $\chi^2_{df=1}$ . 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 ( $1 < p < 0.05$ ,  $0.05 < p < 0.001$ ,  $0.001 < p < 1 \times 10^{-6}$ ,  $1 \times 10^{-6} < p$ ) 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

**Dataset S1** Complete SNP-wise association data. Here we present the complete dataset on which we base our analysis. All data have been previously published as detailed in the main

### References

1. Vyse T, Todd J (1996) Genetic analysis of autoimmune disease. *Cell* 85: 311–318.
2. Barrett JC, Hansoul S, Nicolae D, Cho J, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 40: 955–962.
3. Graham R, Cotsapas C, Davies L, Hackett R (2008) Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. *Nat Genet* 40: 1059–1061.
4. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, et al. (2008) Newly identified genetic risk variants for celiac disease related to the immune response. *Nat Genet* 40: 395–402.
5. De Jager PL, Jia X, Wang J, De Bakker PIWd, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 41: 776.
6. Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, et al. (2009) Genome-wide scan reveals association of psoriasis with IL-23 and NF- $\kappa$ B pathways. *Nat Genet* 41: 199–204.
7. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, et al. (2010) Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 42: 508–514.
8. Criswell L, Pfeiffer K, Lum R, Gonzales B, Novitzke J, et al. (2005) Analysis of families in the multiple autoimmune disease genetics consortium (MADGC) manuscript and in the key below, and are based on chi-square (1 df) or Z association statistics. Where not provided, we computed Z scores as the square root of the cognate chi-squared statistic. Sign was assigned with reference to the minor allele declared in the psoriasis GWAS (chosen arbitrarily). CHR - chromosome. POS - physical position (hg18). major\_al - major SNP allele. minor\_al - minor SNP allele. RA.Z - association Z score for rheumatoid arthritis (Stahl *et al.* *Nat Genet* 2010) [7]. RA.P - association p value for rheumatoid arthritis (Stahl *et al.* *Nat Genet* 2010) [7]. PS.Z - ditto for psoriasis (Nair *et al.* *Nat Genet* 2009) [6]. PS.P - ditto for psoriasis (Nair *et al.* *Nat Genet* 2009) [6]. MS.Z - ditto for multiple sclerosis (De Jager *et al.* *Nat Genet* 2009) [5]. MS.P - ditto for multiple sclerosis (De Jager *et al.* *Nat Genet* 2009) [5]. SLE.Z - ditto for systemic lupus erythematosus (Gateva *et al.* *Nat Genet* 2009) [23]. SLE.P - ditto for systemic lupus erythematosus (Gateva *et al.* *Nat Genet* 2009) [23]. CD.Z - ditto for Crohn's disease (Barrett *et al.* *Nat Genet* 2008) [2]. CD.P - ditto for Crohn's disease (Barrett *et al.* *Nat Genet* 2008) [2]. CeD.Z - ditto for celiac disease (Hunt *et al.* *Nat Genet* 2008) [4]. CeD.P - ditto for celiac disease (Hunt *et al.* *Nat Genet* 2008) [4]. T1D.Z - ditto for type I diabetes (Barrett *et al.* *Nat Genet* 2009) [24]. T1D.P - ditto for type I diabetes (Barrett *et al.* *Nat Genet* 2009) [24]. Disease - disease in which the SNP was originally reported: AITD autoimmune thyroid disease; AS ankylosing spondylitis; BD CD Crohn's disease; MS multiple sclerosis; PS psoriasis; SLE systemic lupus erythematosus; T1D type 1 diabetes; UC ulcerative colitis. cpma.p - p value for CPMA statistic (chi-squared, 1 df). Genes - nearby notable genes. (TAB)

**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

Conceived and designed the experiments: C Cotsapas, BF Voight, DA Hafler, SS Rich, MJ Daly. Performed the experiments: C Cotsapas, BF Voight, E Rossin, BM Neale, MJ Daly. Analyzed the data: C Cotsapas, BF Voight, E Rossin, K Lage, MJ Daly. Contributed reagents/materials/analysis tools: BF Voight, K Lage, BM Neale, C Wallace, GR Abecasis, JC Barrett, T Behrens, J Cho, PL De Jager, JT Elder, RR Graham, P Gregersen, L Klareskog, KA Siminovitch, DA van Heel, C Wijmenga, J Worthington, JA Todd, DA Hafler, SS Rich, MJ Daly. Wrote the paper: C Cotsapas, BF Voight, JA Todd, DA Hafler, SS Rich, MJ Daly.



- collection: the PTPN22 620W allele associates with multiple autoimmune phenotypes. *The American Journal of Human Genetics* 76: 561–571.
9. Wandstrat A, Wakeland E (2001) The genetics of complex autoimmune diseases: non-MHC susceptibility genes. *Nature immunology* 2: 802–809.
  10. Eaton W, Rose N, Kalaydjian A, Pedersen M, Mortensen P (2007) Epidemiology of autoimmune diseases in Denmark. *Journal of Autoimmunity* 29: 1–9.
  11. Maier L, Lowe C, Cooper J, Downes K, Anderson D, et al. (2009) IL2RA genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genet* 5: e1000322. doi:10.1371/journal.pgen.1000322.
  12. Plenge RM, Cotsapas C, Davies L, Price AL, De Bakker PIWd, et al. (2007) Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet* 39: 1477–1482.
  13. Remmers EF, Plenge RM, Lee AT, Graham R, Hom G, et al. (2007) STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *The New England Journal of Medicine* 357: 977–986.
  14. Fung EY, Smyth DJ, Howson JM, Cooper JD, Walker NM, et al. (2009) Analysis of 17 autoimmune disease-associated variants in type 1 diabetes identifies 6q23/TNFAIP3 as a susceptibility locus. *Genes Immun* 10: 188–191.
  15. Smyth DJ, Plagnol V, Walker NM, Cooper JD, Downes K, et al. (2008) Shared and distinct genetic variants in type 1 diabetes and celiac disease. *The New England Journal of Medicine* 359: 2767–2777.
  16. Sirota M, Schaub MA, Batzoglu S, Robinson WH, Butte AJ (2009) Autoimmune disease classification by inverse association with SNP alleles. *PLoS Genet* 5: e1000792. doi:10.1371/journal.pgen.1000792.
  17. Zhernakova A, van Diemen C, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat Rev Genet* 10: 43–55.
  18. de Bakker PIWd, Mcvane G, Sabeti PC, Miretti MM, Green T, et al. (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet* 38: 1166–1172.
  19. Rossin E, Lage K, Raychaudhuri S, Xavier R, Tatar D, et al. (2011) Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet* 7: e1001273. doi:10.1371/journal.pgen.1001273.
  20. Khanna H, Davis EE, Murga-Zamalloa CA, Estrada-Cuzcano A, Lopez I, et al. (2009) A common allele in RGRIP1L is a modifier of retinal degeneration in ciliopathies. *Nat Genet* 41: 739–745.
  21. Todd JA (2010) Etiology of type 1 diabetes. *Immunity* 32: 457–467.
  22. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, et al. (2010) Multiple common variants for celiac disease influencing immune gene expression. *Nature Publishing Group* 42: 295–302.
  23. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, et al. (2009) A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systematic lupus erythematosus. *Nat Genet* 41: 1228–1233.
  24. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, et al. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 41: 703.
  25. Hindorf L, Sethupathy P, Junkins H, Ramos E, Mehta J, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 106: 9362.
  26. Johnson AD, Handsaker R, Pulit S, Nizzari M, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24: 2938–2939.
  27. Gower J (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27: 857–874.
  28. Ward Jr. J (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. pp 236–244.
  29. Team RDC (2010) R: A Language and Environment for Statistical Computing. Vienna, Austria.
  30. Lage K, Karlberg EO, Størling ZM, Ólason PÍ, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorder. *Nat Biotechnol* 25: 309–316.