

# Integrated enrichment analysis of variants and pathways in genome-wide association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling genes in Crohn's disease

Peter Carbonetto<sup>1,\*</sup>, Matthew Stephens<sup>1,2</sup>

<sup>1</sup> Dept. of Human Genetics, University of Chicago, Chicago, IL, USA

<sup>2</sup> Dept. of Statistics, University of Chicago, Chicago, IL, USA

\* E-mail: pcarbo@uchicago.edu

## Abstract

Pathway analyses of genome-wide association studies aggregate information over sets of related genes, such as genes in common pathways, to identify gene sets that are *enriched* for variants associated with disease. We develop a model-based approach to pathway analysis, and apply this approach to data from the Wellcome Trust Case Control Consortium (WTCCC) studies. Our method offers several benefits over existing approaches. First, our method not only interrogates pathways for enrichment of disease associations, but also estimates the level of enrichment, which yields a coherent way to promote variants in enriched pathways, enhancing discovery of genes underlying disease. Second, our approach allows for multiple enriched pathways, a feature that leads to novel findings in two diseases where the major histocompatibility complex (MHC) is a major determinant of disease susceptibility. Third, by modeling disease as the combined effect of multiple markers, our method automatically accounts for linkage disequilibrium among variants. Interrogation of pathways from eight databases yields strong support for enriched pathways, indicating links between Crohn's disease (CD) and cytokine-driven networks that modulate immune responses; between rheumatoid arthritis (RA) and "Measles" pathway genes involved in immune responses triggered by measles infection; and between type 1 diabetes (T1D) and IL2-mediated signaling genes. Prioritizing variants in these enriched pathways yields many additional putative disease associations compared to analyses without enrichment. For CD and RA, 7 of 8 additional non-MHC associations are corroborated by other studies, providing validation for our approach. For T1D, prioritization of IL-2 signaling genes yields strong evidence for 7 additional non-MHC candidate disease loci, as well as suggestive evidence for several more. Of the 7 strongest associations, 4 are validated by other studies, and 3 (near IL-2 signaling genes *RAF1*, *MAPK14*, and *FYN*) constitute novel putative T1D loci for further study.

## Author Summary

Genome-wide association studies have helped locate gene variants that affect our susceptibility to common diseases. The analysis of these studies is typically straightforward: systematically test each genetic variant in isolation whether it is correlated with predisposition to disease. This approach often works well for identifying commonly occurring variants with moderate effects on disease risk, but the effects of many variants are so small they fail to register statistically significant correlations. This is a concern because many common diseases are modulated by a large number of genetic factors with small effects on disease risk. An alternative strategy is to examine groups of variants, such as variants sharing a common biological pathway, and to assess whether these groups are "enriched" for correlations with disease. This can be a more effective approach to identifying genetic factors relevant to disease. However, it does not tell us whether individual genes are associated with disease, which remains an important question. To address this limitation, we describe an approach that integrates enrichment analysis with tests for disease-variant correlations within a single statistical framework. We illustrate this approach in genome-wide association studies of seven complex diseases. We show that our approach supports enriched pathways in several diseases, and uncovers disease-susceptibility genes in these pathways that are not identified in conventional analyses of the same data.

## Introduction

By surveying genetic variation throughout the genome, and systematically searching for variants correlated with disease phenotypes, genome-wide association studies (GWAS) have led to the discovery of genes and genetic loci that were not previously suspected of playing a role in complex diseases [1–4]. Although the variants associated with disease usually have small effects on susceptibility to disease, they nonetheless have the potential to yield important insights into the biology of disease. In particular, finding common features of genetic loci implicated by GWAS can identify key biological processes involved in disease etiology. For example, the discovery of disease-correlated variants in GWAS of Crohn’s disease, a common form of inflammatory bowel disease, has helped draw links to genes that regulate autophagy and innate immune responses [5–10].

Recognizing that insights into disease can emerge by exploring the functional relationships among genes implicated in GWAS, researchers have attempted to assess these relationships in a systematic way by developing “pathway analysis” approaches to GWAS [11–28]. Motivated by the theory that complex disease arises from the accumulation of genetic effects acting within common biological pathways [29–32], these methods aim to identify pathways that are *enriched* for disease—that is, groups of related genes that preferentially harbour disease-associated variants compared to arbitrary regions of the genome. Not only is identifying enriched pathways an important aim in itself, but doing so can also overcome an important shortcoming of standard genome-wide mapping approaches that test each marker one at a time for association with disease: they lack power to map genetic factors of small effect [33–36]. Pathway analysis can improve power to uncover genetic factors relevant to disease because identifying the accumulation of small genetic effects acting in a common pathway is often easier than mapping the individual genes within the pathway that contribute to disease susceptibility.

Despite the considerable potential of pathway analysis approaches to GWAS, existing methods have an important limitation: they do not tell us which genes within an enriched pathway are most likely relevant to disease. While identifying enriched pathways is often useful, many pathways contain large numbers of genes with only loosely interrelated functions, so identifying the genes and variants within the pathway that are driving the enrichment is likely to yield additional insights into disease. This could be tackled in a two-stage process: first identify the enriched pathways, then gauge support for associated variants within these pathways. In the second stage, significance thresholds for association could be relaxed relative to a genome-wide scan, reflecting the increased likelihood that variants near genes in the pathway are associated with disease. This is called *prioritizing* variants within the pathway [29, 37–45]. The question is how to implement this in a systematic way; to what extent can we relax significance thresholds while keeping the rate of false positives at an acceptable level?

To address this question, we develop a model-based approach for integrated analysis of pathways and genetic variants, in which we interpret enrichment as a parameter of the model. We begin with a large-scale multivariate regression that models disease risk as the combined effect of multiple markers. Unlike single-marker disease mapping, the multi-marker approach accounts for correlations between variants that arise due to linkage disequilibrium. Within this framework, we introduce an enrichment parameter that quantifies the increase in the probability that each variant in the pathway is associated with disease susceptibility. This model-based approach not only estimates the level of enrichment, but also adjusts the evidence for disease associations in light of estimated pathway enrichments—and in so doing, tackles the problem of how to prioritize variants related to certain genes or pathways.

Though we focus on incorporating pathways—and, more broadly, biologically related gene sets—into analysis of GWAS, our methods could be applied to other types of genome annotations, such as Gene Ontology categories [46] and regions where proteins are recruited to regulate gene transcription [47–51]. In this respect, our method is related to other model-based approaches that leverage prior knowledge about variants to better discern the causal variants that affect disease risk, or to estimate enrichment of genome-wide association signals across functionally related regions [29, 52–62]. One distinguishing feature of our approach is that we have an efficient procedure to evaluate hypotheses about enrichment, which allows us to interrogate support for enrichment of thousands of candidate pathways in genome-wide data.

Another feature that distinguishes our analysis is that we use multiple pathway databases in an attempt to

interrogate pathways as comprehensively as possible—the more pathways we consider, the greater chance we have of drawing new connections between pathways, genes within these pathways, and complex disease. We demonstrate how using our approach to comprehensively interrogate pathways results in increased evidence for enrichment, and is robust to inclusion of a large number of irrelevant pathways. In our case studies, we include ~3600 candidate gene sets drawn from eight pathway databases available on the Web [44, 63, 64].

We illustrate our approach in a detailed analysis of genome-wide data from the Wellcome Trust Case-Control Consortium studies of 7 complex diseases [65]. These studies provide an opportunity to gauge the added value of our approach because genetic associations have already been published based these data [65], and pathway analyses of these data have found evidence for enriched pathways [11, 13, 16, 22, 23, 66–74]. Our methods highlight several pathways that have not been previously identified in pathway-based analyses, but which are known to be linked to these diseases. And by prioritizing variants within enriched pathways, our methods identify disease-susceptibility candidates that are not deemed significant in conventional analyses of the same data. These results demonstrate the potential for our methods to yield novel biological insights into complex disease.

## Overview of statistical analysis

Our approach builds on previous work that casts simultaneous analysis of genetic variants as a *variable selection problem*—the problem of deciding which variables (the genetic variants) to include in a multivariate regression of the phenotype. We begin with a method that assumes each variant is equally likely to be associated with the phenotype [75, 76], then we modify this assumption to allow for enrichment of associated variants in a pathway.

The data from the GWAS are the genotypes  $\mathbf{X} = (x_1, \dots, x_n)^T$  and phenotypes  $y = (y_1, \dots, y_n)^T$  from  $n$  study participants. Here we assume the genetic markers are SNPs, and the phenotype is disease status: patients with the disease (“cases”) are labeled  $y_i = 1$ , and disease-free individuals (“controls”) are labeled  $y_i = 0$ . Entries of the  $n \times p$  matrix  $\mathbf{X}$  are observed minor allele counts  $x_{ij} \in \{0, 1, 2\}$ , or expectations of these counts estimated using genotype imputation [77, 78], for each of the  $n$  samples and  $p$  SNPs.

We assume an additive model of disease risk, in which the log-odds for disease is a linear combination of the minor allele counts:

$$\log \left\{ \frac{p(y_i = 1)}{p(y_i = 0)} \right\} = \beta_0 + x_{i1}\beta_1 + \dots + x_{ip}\beta_p. \quad (1)$$

Under this additive model,  $e^{\beta_j}$  is the *odds ratio*, the multiplicative increase in odds of disease for each copy of the minor allele at locus  $j$ . We do not consider dominant or recessive effects on disease risk, but it would be straightforward to include them; see [79]. This method is also easily adapted to quantitative traits by replacing (1) with a linear regression for  $y$ .

Although the log-odds for disease is expressed in (1) as a linear combination of all SNPs, our framework is guided by the assumption that most SNPs  $j$  have no effect on disease risk ( $\beta_j = 0$ ). While there is some debate over this assumption [80], an advantage of this choice is that a SNP “included” in the multi-marker disease model—that is, a SNP  $j$  that has a non-zero coefficient ( $\beta_j \neq 0$ )—indicates that the SNP is relevant to disease, or that it is in linkage disequilibrium with other, possibly untyped, variants that contribute to disease risk. Therefore, the main goal of the analysis is to identify regions of the genome that contain SNPs included in the disease model with high posterior probability, or identify SNPs within these regions that have a high “posterior inclusion probability,”  $\text{PIP}(j) \equiv p(\beta_j \neq 0 | \mathbf{X}, y)$ . A high PIP is the analogue of a small  $p$ -value in a conventional single-marker analysis.

To obtain these posterior probabilities, we must first specify a prior for the coefficients  $\beta_j$ . A standard assumption, and the assumption made in previous approaches [75, 76], is that SNPs are equally likely to be associated with the phenotype *a priori*; that is,  $\pi_j \equiv p(\beta_j \neq 0)$  is the same for all SNPs.

To model enrichment of associations within a pathway, we modify this prior. Precisely, the prior inclusion probability for SNP  $j$  depends on whether or not it is assigned to the enriched pathway:

$$\log_{10} \left( \frac{\pi_j}{1 - \pi_j} \right) = \theta_0 + a_j \theta, \quad (2)$$

where the pathway indicators  $a_j$  keep track of which SNPs are assigned to the enriched pathway,  $a_j = 1$  when SNP  $j$  is assigned to the enriched pathway, otherwise  $a_j = 0$ . (In brief, a SNP is assigned to a pathway if it is near a gene in the pathway; see Methods.) We refer to  $\theta_0$  as the *genome-wide log-odds*, since it reflects the background proportion of SNPs that are included in the multi-marker disease model. (More precisely, it is the proportion for SNPs not assigned to the pathway, but this is usually most SNPs.) We refer to  $\theta$  as the *log-fold enrichment parameter* because it corresponds to the increase in probability, on the log-odds scale, that a SNP assigned to the pathway is included in the model. For example,  $\theta_0 = -4$  and  $\theta = 2$  indicates that 1 out of every 10,000 SNPs outside the pathway is included in the multi-marker model, but for SNPs assigned to the pathway, 1 out of every 100 is included. If  $\theta = 0$ , this reduces to the standard prior assumption made by previous methods. We expect  $\theta$  to be 0, or close to 0, for most pathways.

We assess enrichment by framing each hypothesis test as a model comparison problem. To weigh the evidence for the hypothesis that candidate pathway with indicators  $a = (a_1, \dots, a_p)$  is enriched for disease associations, we evaluate a *Bayes factor* [81, 82]:

$$\text{BF}(a) = \frac{p(y | \mathbf{X}, a, \theta > 0)}{p(y | \mathbf{X}, \theta = 0)}. \quad (3)$$

This Bayes factor (BF) is the ratio of likelihoods under two models, the model in which the candidate pathway is enriched for SNPs included in the multi-marker model ( $\theta > 0$ ), and the null model with no enrichment ( $\theta = 0$ ). A larger BF implies stronger evidence for enrichment. We compute each BF by averaging, or *integrating*, over the unknown parameters, and over multi-marker models with different combinations of SNPs, employing appropriate prior distributions for  $\theta_0$ ,  $\theta$  and the coefficients  $\beta_j$  (see Methods).

Note that the Bayes factor (3) does not allow for a negative  $\theta$ —that is, pathways that are *underrepresented* for associations with the phenotype. While it could be useful to investigate negative log-enrichments in other settings, in most GWAS of complex disease where there are generally few significant associations to begin with, reduced rates of disease associations in pathways would be difficult to find, and would be unlikely to have a useful interpretation.

We use the same approach to test for joint enrichment of multiple candidate pathways. We compute  $\text{BF}(a)$  as before (eq. 3), except that we set  $a_j$  to 1 whenever SNP  $j$  is assigned to at least one of the pathways. In this case,  $\theta$  represents the increased level of associations (on the log-odds scale) among SNPs assigned to one or more of the pathways. This is equivalent to assuming that all enriched pathways have the same level of enrichment, which greatly simplifies the analysis. We allow for different enrichment levels only when accounting for enrichment of the MHC in the analysis of RA and T1D. In this case, we have good reason to treat the MHC differently, given the predominant contribution of MHC alleles to RA and T1D [83, 84].

To assess evidence for association of individual variants with the phenotype, we compute  $\text{PIP}(j)$  for each variant  $j$ . These posterior probabilities depend on which pathways are enriched, and on the log-fold enrichment  $\theta$ , because these factors affect the prior probabilities  $\pi_j$ , which in turn affect the posterior probabilities  $\text{PIP}(j)$ , following Bayes' rule. (In practice, we account for uncertainty in  $\theta_0$  and  $\theta$  when calculating the posterior probabilities by averaging over  $\theta_0$  and  $\theta$ ; see Methods.) Since enrichment leads to higher prior inclusion probabilities for SNPs in the enriched pathway, an association that is not identified by a conventional genome-wide analysis may become a strong candidate in light of its presence in an enriched pathway. Because we estimate  $\theta$  from the data, the extent to which we prioritize variants is determined by the data. In this way, our framework integrates the problem of identifying enriched pathways with the problem of prioritizing variants near genes in enriched pathways.

## Results

We illustrate our methods in a detailed analysis of genome-wide marker data from case-control studies of seven common diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D) [65]. After steps to ensure data quality (see Methods), the data for each disease consist of ~440,000 SNPs genotyped for 1748–1963 cases and

2938 controls (Table S1). Many of the genetic associations based on these data [65] have been replicated in follow-up studies [85–89]. We compare our results to previously reported associations, and to earlier pathway analyses of these data [11, 13, 16, 22, 23, 66–74].

We analyze the data in three stages. First, we compute a BF for each candidate pathway to assess whether it is enriched for disease associations, ranking the pathways according to their BFs. (Throughout this paper, we use the term “pathway” to refer to a collection of functionally related genes.) Second, we investigate whether prioritizing variants within enriched pathways can help locate disease associations beyond those identified in analyses that ignore information about pathways. Finally, for diseases with evidence of pathway enrichment, we re-examine the data for models in which two or more pathways are enriched, and investigate whether prioritizing combinations of enriched pathways yields further genetic loci associated with disease.

## Selection of candidate pathways

We assemble a comprehensive list of candidate pathways to test for enrichment, drawing from a variety of publicly accessible collections (Fig. 3). We do not filter pathway candidates based on their potential relevance to disease. In total, we interrogate 3158 candidate pathways for each disease, plus the MHC and xMHC gene sets described below. Most candidate pathways were curated by domain experts, and others are based on experimental evidence in non-human organisms and inferred via gene homology. For full details of pathway databases used, and steps taken to compile gene sets from pathway data, refer to Fig. 3, Methods, Supplementary Materials, and links to source code implementing our analysis.















## The major histocompatibility complex

In two of the seven diseases, RA and T1D, several disease associations map to the major histocompatibility complex (MHC) region on chromosome 6. Consequently, pathway analyses for RA and T1D tend to highlight pathways that involve MHC genes. When we apply our method to these diseases, the top pathways for T1D and RA are “Allograft rejection” and “Asthma,” respectively. Both gene sets include multiple MHC genes and exhibit very strong evidence for enrichment ( $\text{BF} = 1.8 \times 10^{39}$  and  $6.8 \times 10^{15}$ ). Other pathways with the strongest enrichment signals also contain MHC genes.

Most of the support for enrichment of these pathways is likely driven by disease associations that map to the MHC. To confirm this, we create a “pathway” containing all genes within the MHC [90], and test this gene set for enrichment. The MHC gene set shows more support for enrichment than any other pathway by several orders of magnitude ( $\text{BF} = 2.7 \times 10^{54}$  and  $2.4 \times 10^{21}$  in T1D and RA, respectively), and it is accompanied by a high estimate of enrichment ( $\theta_{\text{MHC}} = 4.6$  and  $3.7$ ). Performing a similar enrichment analysis for all genes within the “extended” MHC (xMHC) [91] yields smaller BFs (Table S3), suggesting that the genetic contribution to RA and T1D risk lies mostly within the class I, II and III subregions of the MHC.

Our finding that the MHC is enriched for associations with T1D and RA is unsurprising considering the MHC is estimated to account for over half the genetic contribution to T1D risk, and at least a third for RA [83, 84, 92]. (By contrast, the genetic contribution of the MHC is estimated to be ~10% for CD [83], and the BFs for the MHC and xMHC in CD are 8 and 4, respectively.) In light of these findings, a reasonable question to ask is whether pathways show enrichment for disease associations beyond enrichment of the MHC. A strength of our model-based approach is that it can address this question by computing a BF for enrichment of each candidate pathway, conditioned on the estimated enrichment of the MHC. Thus, in our subsequent analysis of RA and T1D, we account for enrichment of disease associations within the MHC in this way. As far as we are aware, no other pathway analyses of these data incorporate enrichment of the MHC, which may explain why previous studies have highlighted mostly MHC-related pathways and gene categories.

**Table 1.** Diseases show a wide range of support for enrichment of disease associations in pathways, as measured by their Bayes factors.

disease	enriched pathway	database	Bayes factor	posterior probability**	number of genes	number of SNPs
T1D	IL-2 signaling*	PID (BS)	$1.2 \times 10^{12}$	1.000	 52	 1964
CD	Cytokine signaling	Reactome (BS)	$9.0 \times 10^5$	0.997	 225	 6711
RA	Measles*	KEGG (BS)	2576	0.449	 130	 3488
T2D	Incretin regulation	Reactome (PC)	259	0.076	 21	 689
CAD	Arf inhibits ribosomal RNA	BioCarta	144	0.044	 17	 546
BD	Transport of connexons	Reactome (PC)	15	0.005	 2	 61
HT	Ala biosynthesis	PANTHER	5	0.002	 2	 115

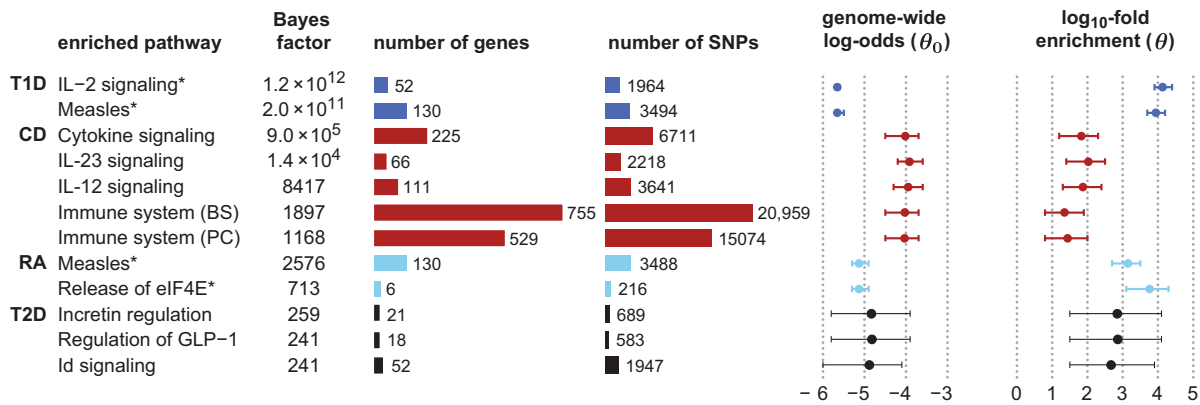
Each row shows the pathway with the largest BF for enrichment of disease associations among 3158 candidate gene sets. Columns left to right: (1) disease; (2) enriched pathway; (3) pathway database, and repository where pathway is retrieved if different from database; (4) BF for hypothesis that disease associations are enriched among SNPs assigned to pathway; (5) posterior probability of enrichment hypothesis; (6) number of genes assigned to pathway; (7) number of SNPs near these genes assigned to pathway; (8,9) posterior mean and 95% credible interval of genome-wide log-odds ( $\theta_0$ ) and log-fold enrichment ( $\theta$ ) given that pathway is enriched ( $\theta > 0$ ). Note that enrichment level is defined on log-scale (eq. 2), so  $\theta > 0$  indicates enrichment. Credible interval is smallest interval about mean that contains parameter with 95% posterior probability, calculated to nearest 0.1 using a numerical approximation. Abbreviations used in table: PID = NCI Nature Pathway Interaction Database [93], BS = NCBI BioSystems [94], PC = Pathway Commons [95]. Databases and database identifiers for pathways listed here: “Transport of connexons to the plasma membrane” (Reactome 11050, PC); “Tumor suppressor Arf inhibits ribosomal biogenesis” (BioCarta); “Cytokine signaling in immune system” (Reactome 75790, BS 366171); “Alanine biosynthesis” (PANTHER P02724); “Measles” (KEGG hsa05162, BS 213306); “IL2-mediated signaling events” (PID il2\_1pathway, BS 137976); “Incretin synthesis, secretion, and inactivation” (Reactome 23974, PC). \*Null and enrichment hypotheses for RA and T1D include enrichment of disease associations in MHC, in which SNPs within MHC are enriched at a different level than non-MHC SNPs in pathway;  $\theta_{\text{MHC}} = 3.7$  and 4.6 for RA and T1D, respectively. Number of genes/SNPs for RA and T1D count only non-MHC genes assigned to pathway. \*\*Illustrative posterior probability assuming a “conservative” prior (see text).

## Bayes factors for enrichment

To give an initial impression of our enrichment results, we show the pathway with the largest BF for each disease in Table 1. The seven diseases exhibit a wide range of support for the strongest enrichment signal. For example, the top pathway for T1D, IL-2 signaling, has a BF of  $1.2 \times 10^{12}$ , whereas the largest BF for HT is only 5.

To address whether these top pathways constitute “significant” evidence for enrichment, the BF for enrichment must be weighed against the prior probability of the pathway being enriched to obtain a posterior probability of enrichment (see “Interpretation of Bayes factors” in Methods). While specification of a prior probability of enrichment is subjective, this subjectivity is unavoidable; similar issues arise when specifying significance thresholds for  $p$ -values, though these issues are usually hidden (0.05 is a common threshold, but it is subjective and not universally appropriate [97]). If we apply a “conservative” value of  $1/3158$  to the prior probability for all candidate pathways, so that one pathway is expected to be enriched among the 3158 candidates, then CD and T1D show compelling evidence for enrichment (posterior probability  $> 0.99$ ), and RA shows suggestive evidence (posterior probability = 0.45). Considering the plausible connection between Measles pathway genes and RA (discussed below), we view this as a compelling enrichment as well. The top pathway for T2D, Incretin regulation, shows only modest evidence for enrichment if we use the conservative prior, but might be considered “significant” if we use a less conservative prior to account for the known connection of this pathway to insulin resistance and diabetes. Based on these results, we do not investigate BD, CAD and HT further, and focus on the four diseases showing strongest evidence for enrichment: CD, RA, T1D and T2D.

Table 2 shows an expanded list of pathways with the strongest support for enrichment in CD, RA, T1D and

**Table 2.** Top-ranked candidate pathways for enrichment of disease associations in CD, RA, T1D and T2D.

Refer to Table 1 for legend, abbreviations, and meaning of asterisk (\*). Database identifiers for pathways listed here and not previously mentioned: “IL23-mediated signaling events” (PID il23pathway, PC); “IL12-mediated signaling events” (PID il12\_2pathway, PC); “Immune system” (Reactome 6900, BS 106386); “Release of eIF4E” (Reactome 6836, PC); “Asthma” (KEGG hsa05310, BS 83120); “Allograft rejection” (KEGG hsa05330, BS 83123); “Synthesis, secretion, and inactivation of glucagon-like peptide-1” (Reactome 24019, PC); “Id signaling pathway” (WikiPathways WP53 [96], BS 198871). See Table S3 for more gene set enrichment results.

T2D (see Table S3 for a longer list), together with estimated enrichment levels. Beyond these top results, the vast majority of candidate pathways show little or no evidence for enrichment (Fig. S7), demonstrating that our method is robust to inclusion of many pathways that are most likely irrelevant to the disease.

Before discussing the biological relevance of these pathways, we point out three general features of Table 2. First, some of the estimated enrichments are extremely large; for example, IL-2 signaling genes show more than a 10,000-fold enrichment of T1D risk factors. In contrast, the top pathway for CD, “Cytokine signaling in immune system,” has roughly a 100-fold enrichment. (Enrichment of this pathway nonetheless yields a large BF, partly because it implicates over 6700 SNPs; the BFs depend not only on the level of enrichment, but also on the number of SNPs assigned to the pathway.) Second, some of the top pathways overlap or are subsets of one another. For example, “Cytokine signaling in immune system” is a subset of “Immune system.” The Immune system pathway from NCBI BioSystems (BS) overlaps with the Pathway Commons (PC) version of the same pathway (510 genes are common to both gene sets). This raises the question whether enrichment of just one pathway would suffice to explain the genome-wide association signal; we use our methods to investigate this question below. Third, 5 different pathway databases are represented in Table 1, and all 8 pathway databases included in our analysis appear among the top pathways (Table 2), illustrating the benefits of interrogating pathways from multiple sources.

## Biological relevance of enriched pathways

The top-ranked pathway for CD (“Cytokine signaling”) is a collection of cytokine-driven networks that exhibit a complex relationship to autoimmunity, promoting inflammatory and immune responses, while also playing an important role in suppressing immunity [98]. Cytokine signaling implicates a broad class of 225 genes, suggesting that a collection of related gene networks explains the pattern of genetic associations better than any one signal transduction pathway. Enrichment of cytokine signaling is consistent with the accumulating evidence that points to cytokines, and the signaling cascades initiated by these cytokines, in a range of autoimmune disorders, including inflammatory bowel disease (IBD) [9, 99, 100].

Previous findings from GWAS have linked autophagy genes *ATG16L1* and *IRGM* to CD [65, 101, 102]. Our pathway analysis does not provide additional support for autophagy in CD because pathways reflecting current

models of autophagy [9, 103] have not yet been incorporated, as far as we are aware, into any of the publicly available pathway databases.

Once we account for enrichment of the MHC, the top pathway for T1D is IL-2 signaling. Cytokine IL-2 and its interacting partners are indispensable to activation, development and maintenance of T regulatory cells, and disruption of IL2-mediated pathways promotes progression of autoimmune disorders [104–106]. Treatments of T1D targeting the IL-2 signaling pathway are currently undergoing clinical trials [107]. Additionally, studies in non-obese diabetic (NOD) mice suggest that defects in IL-2 signaling induce susceptibility to T1D [106, 108, 109]. Our findings support this hypothesis.

The top pathway for RA, the KEGG “Measles” pathway, contains genes involved in immune response cascades triggered by infection of measles virus, including the cellular receptors expressed for measles virus such as *SLAM* and *CD46* [110–112]. (Again, this result is conditioned on enrichment of the MHC.) While studies have associated the measles virus with RA [113], other viral and bacterial infections have also been linked to incidence of RA [114, 115], and enrichment of the Measles pathway could reflect a larger class of genes involved in regulation of immune function during infection, rather than the measles virus specifically. The large BF for this pathway in both RA and T1D supports previous indications of a shared genetic basis [100, 116], and is consistent with observations that RA and T1D, along with other autoimmune diseases, recur in the same families [117].

All CD, RA and T1D pathways in Table 2 implicate key actors in responses to pro-inflammatory stimuli and in regulation of innate and adaptive immunity. These include members of the NF- $\kappa$ B/Rel family, T-cell receptors (TCRs), members of the protein tyrosine phosphatase family (PTPs), mitogen-activated protein (MAP) kinases such as c-Jun NH<sub>2</sub>-terminal kinases (JNKs), and chemokine receptors (CXCRs) [118–121].

## Comparison with previous pathway analyses

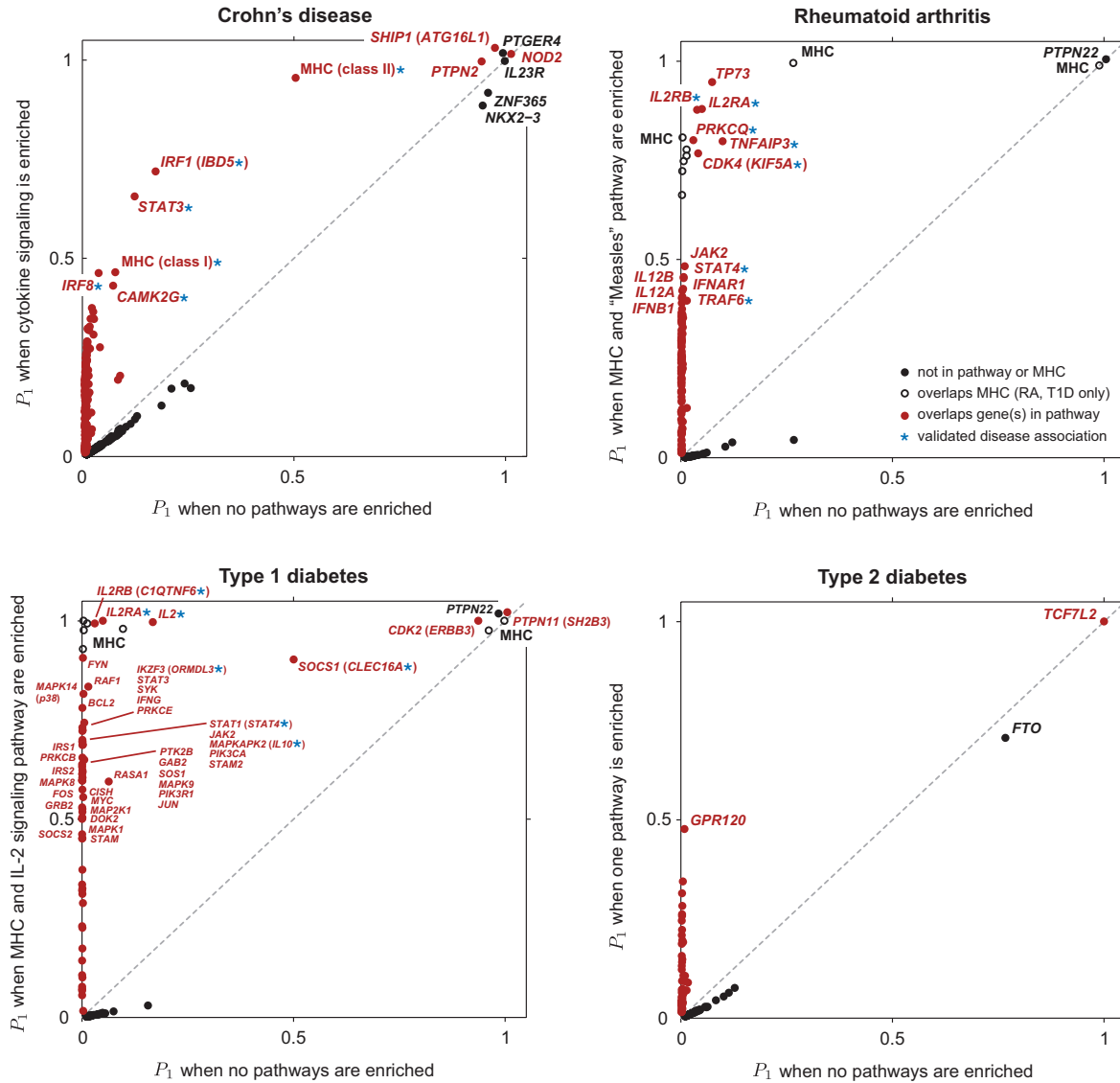
None of the top-ranked pathways for CD, RA and T1D in our analysis have been identified in previous pathway-based analyses of these diseases [11, 13, 16, 23, 67–70, 73, 122–125].

An important difference between our methods and previous pathway analyses of RA and T1D is that we incorporate enrichment of the MHC into models of enrichment. A previous analysis of RA [67] highlighted pathways “Bystander B cell activation” (BioCarta, their  $p$ -value =  $10^{-371}$ ) and “Type 1 diabetes mellitus” (KEGG,  $p$ -value =  $10^{-349}$ ). However, both these pathways contain MHC genes, and our results suggest that enrichment of the MHC offers a better explanation of the association signal; in our analysis, support for enrichment of these pathways is several orders of magnitude less than support for enrichment of the MHC (BF =  $1.6 \times 10^7$  and  $3.4 \times 10^{13}$  versus  $2.4 \times 10^{21}$ ), and the support vanishes once we account for enrichment of the MHC (BF = 0.69 and 0.57). Similarly, previous analyses of T1D [16, 125] have highlighted the same “Type 1 diabetes mellitus” pathway, but again support for enrichment is driven mostly by the association signal in the MHC, as our methods yield only modest support for this pathway after accounting for MHC enrichment (BF = 43).

It is also notable that the top-ranked pathways we identify for RA and T1D, Measles and IL-2 signaling, show strong support in our analysis *only after accounting for enrichment of disease associations within the MHC*; the BFs without MHC enrichment are 104 and 11, whereas the BFs are  $2.6 \times 10^3$  and  $1.2 \times 10^{12}$  after conditioning on enrichment of the MHC. This may explain why these pathways have not been identified in previous pathway analyses of these diseases. These results illustrate the benefits of estimating enrichment conditioned on the MHC and, more generally, quantifying support for models with multiple enriched pathways.

Another aspect that differs between our results and previous studies is that we interrogate a comprehensive set of pathway databases, which may partly explain why the BF for the top pathway in CD, “Cytokine signaling in immune system” from Reactome, eclipses the BFs corresponding to previously reported pathways. For example, Wang *et al* [73] interrogated BioCarta, KEGG and Gene Ontology [46] (and not Reactome) gene sets for enrichment of CD associations, and reported the smallest  $p$ -value for BioCarta pathway “IL12 and Stat4 dependent signaling in Th1 development” ( $p$ -value =  $8 \times 10^{-5}$ , FDR = 0.045). This pathway showed little evidence for enrichment in our analysis (BF = 20) compared to cytokine signaling (BF =  $9.0 \times 10^5$ ). (When we combine this pathway with cytokine signaling genes, we obtain stronger evidence for enrichment in CD; the BF is 81% the size of the largest BF for 2 enriched pathways in Table 4.)





**Figure 1.** Scatterplots showing  $P_1$ , the posterior probability that a small (50-SNP) region of the genome contains variants associated with disease risk, given different hypotheses about enriched pathways in CD, RA, T1D and T2D. Each point corresponds to a contiguous region of the genome containing 50 SNPs. Posterior probabilities on vertical axis for CD, RA and T1D are conditioned on enrichment of pathway with largest BF (Table 1). For T2D, since no single pathway stands out in ranking (Tables 2 and S3),  $P_1$  along vertical axis is obtained by averaging over top 5 pathways (see Methods). Points highlighted in red correspond to segments overlapping SNPs assigned to the enriched pathway (for T2D, at least 1 out of 5 top pathways). In RA and T1D, 50-SNP segments overlapping the MHC are drawn as open circles (note these SNPs are not assigned to the pathway). Overlapping segments sharing the same association signal are not shown. Some segments are labeled by gene(s) in pathway and/or most credible gene of interest based on prior studies (most credible gene is shown in parentheses if different from pathway gene). An asterisk (\*) indicates an appreciable increase in the probability of a disease association, and this association is validated by other GWAS for same disease (see Table 3).

## Associations informed by enriched pathways

An important feature of our model-based approach is that pathway enrichments can help to map additional disease associations by prioritizing variants within enriched pathways. This is particularly useful for broad groups of

**Table 3.** Regions of the genome exhibiting moderate to strong evidence for disease risk factors in CD and RA, and strong evidence for T1D risk factors, given feedback from enriched pathways.

	chr.	region (Mb)	$P_1$		$P_2$		trend	candidate	SNP	PIP	LOR (95% CI)	MAF		corroborating references
			null	alt.	null	alt.						ctrls	cases	
CD	1p31	67.30–67.48	1.00	1.00	0.05	0.03	$6.45 \times 10^{-13}$	<i>IL23R</i>	rs11805303	1.00	0.25 (0.18–0.33)	0.318	0.391	**
CD	2q37	233.92–234.27	1.00	1.00	0.01	0.05	$7.10 \times 10^{-14}$	<i>ATG16L1 (SHIP1)</i>	rs10210302	1.00	-0.27 (0.20–0.35)	0.481	0.402	**
CD	5p13	40.32–40.66	1.00	1.00	0.46	0.39	$2.13 \times 10^{-13}$	<i>PTGER4</i>	rs17234657	1.00	0.30 (0.20–0.39)	0.124	0.181	**
* CD	5q23	129.54–132.04	0.28	<b>0.81</b>	0.04	0.40	$5.40 \times 10^{-07}$	<i>IBD5 (IRF1 + 3)</i>	rs274552	0.37	-0.14 (0.03–0.26)	0.166	0.128	6, 126, 127
* CD	6	25.52–33.76	0.66	<b>1.00</b>	0.22	<b>0.95</b>	$8.65 \times 10^{-07}$	<b>MHC</b>	rs9469220	0.94	-0.16 (0.08–0.23)	0.519	0.465	6, 83, 128
CD	10q21	64.0–64.43	0.96	0.92	0.04	0.02	$2.68 \times 10^{-07}$	<i>ZNF365</i>	rs10995271	0.92	0.20 (0.13–0.28)	0.386	0.440	**
CD	10q24	101.26–101.32	0.95	0.89	0.00	0.00	$1.41 \times 10^{-08}$	<i>NKX2-3</i>	rs7095491	0.88	0.19 (0.11–0.26)	0.470	0.527	**
CD	16q12	49.0–49.4	1.00	1.00	0.16	<b>0.62</b>	$9.36 \times 10^{-12}$	<b>NOD2</b>	rs17221417	1.00	0.24 (0.16–0.31)	0.287	0.356	**
* CD	17q21	37.5–38.3	0.13	<b>0.67</b>	0.01	0.19	$7.19 \times 10^{-06}$	<b>STAT3</b>	rs744166	0.51	-0.14 (0.07–0.22)	0.439	0.392	6, 7, 129
CD	18p11	12.76–12.91	0.94	1.00	0.00	0.11	$4.56 \times 10^{-08}$	<b>PTPN2</b>	rs2542151	1.00	0.23 (0.13–0.31)	0.163	0.209	**
* RA	1p36	3.5–3.7	0.07	<b>0.94</b>	0.00	0.16	$1.08 \times 10^{-05}$	<b>TP73</b>	rs12027041	0.93	0.16 (0.08–0.23)	0.414	0.459	–
RA	1p13	113.53–114.36	1.00	1.00	0.01	0.00	$4.90 \times 10^{-26}$	<i>PTPN22</i>	rs6679677	1.00	0.49 (0.39–0.60)	0.096	0.169	**
RA	6	25.52–33.76	1.00	1.00	1.00	1.00	$3.44 \times 10^{-76}$	<b>MHC</b>	rs9268560	1.00	-0.31 (0.24–0.39)	0.483	0.306	**
* RA	6q23	138.00–138.47	0.10	<b>0.80</b>	0.00	0.26	$4.99 \times 10^{-06}$	<b>TNFAIP3</b>	rs11970411	0.71	0.21 (0.08–0.31)	0.080	0.107	88, 130–132
* RA	10p15	6.07–6.26	0.05	<b>0.87</b>	0.00	0.29	$7.02 \times 10^{-06}$	<b>IL2RA</b>	rs2104286	0.82	-0.16 (0.08–0.24)	0.286	0.244	130, 132, 133
* RA	10p15	6.36–6.49	0.03	<b>0.76</b>	0.00	0.10	$5.55 \times 10^{-05}$	<b>PRKCG</b>	rs1570527	0.73	-0.17 (0.08–0.26)	0.190	0.157	85, 130, 132, 134
* RA	12q13	55.77–56.82	0.05	<b>0.77</b>	0.00	0.09	$3.44 \times 10^{-06}$	<i>KIF5A (CDK4)</i>	rs10876991	0.74	-0.15 (0.07–0.22)	0.347	0.307	85, 130, 132, 134
* RA	22q13	35.57–35.90	0.05	<b>0.87</b>	0.00	0.14	$7.92 \times 10^{-06}$	<b>IL2RB</b>	rs743777	0.85	0.16 (0.07–0.23)	0.294	0.336	85, 130, 132
T1D	1p13	113.53–114.36	1.00	1.00	0.03	0.01	$1.17 \times 10^{-26}$	<i>PTPN22</i>	rs6679677	1.00	0.52 (0.40–0.61)	0.096	0.170	**
* T1D	3p25	12.44–12.85	0.02	<b>0.85</b>	0.00	<b>0.57</b>	$4.97 \times 10^{-05}$	<b>RAF1</b>	rs299651	0.12	-0.09 (0.01–0.16)	0.512	0.475	–
* T1D	4q27	123.3–123.93	0.17	<b>1.00</b>	0.00	0.27	$5.01 \times 10^{-07}$	<b>IL2</b>	rs17388568	1.00	0.19 (0.10–0.27)	0.260	0.307	135, 136
T1D	6	25.52–33.76	1.00	1.00	1.00	1.00	$2.4 \times 10^{-134}$	<b>MHC</b>	rs9273363	1.00	0.66 (0.59–0.73)	0.305	0.709	**
* T1D	6p21	35.86–36.38	0.00	<b>0.82</b>	0.00	0.41	$> 10^{-4}$	<b>p38</b>	rs2237093	0.58	-0.18 (0.07–0.30)	0.098	0.077	–
* T1D	6q21	112.02–112.47	0.00	<b>0.95</b>	0.00	<b>0.77</b>	$> 10^{-4}$	<b>FYN</b>	rs12910	0.66	-0.13 (0.06–0.21)	0.487	0.454	–
* T1D	10p15	6.07–6.26	0.05	<b>1.00</b>	0.00	<b>0.70</b>	$7.96 \times 10^{-06}$	<b>IL2RA</b>	rs2104286	1.00	-0.20 (0.12–0.29)	0.286	0.245	84, 135, 137–139
T1D	12q13	54.63–54.91	0.94	1.00	0.00	0.25	$1.14 \times 10^{-11}$	<i>ERBB3 (CDK2)</i>	rs1873914	1.00	0.23 (0.15–0.30)	0.414	0.471	**
T1D	12q24	109.8–111.74	1.00	1.00	0.01	<b>0.64</b>	$2.17 \times 10^{-15}$	<i>SH2B3 (PTPN11)</i>	rs17696736	1.00	0.32 (0.24–0.39)	0.424	0.505	**
* T1D	16p13	10.9–11.41	0.51	<b>0.95</b>	0.01	<b>0.76</b>	$9.24 \times 10^{-08}$	<i>CLEC16A (SOCS1)</i>	rs149310	0.61	0.14 (0.05–0.22)	0.248	0.284	89, 135, 140
* T1D	22q13	35.68–36.0	0.03	<b>0.99</b>	0.00	<b>0.75</b>	$6.79 \times 10^{-05}$	<i>C1QTNF6 (IL2RB)</i>	rs3218253	0.97	0.18 (0.09–0.26)	0.251	0.286	135, 141

For all CD and RA loci in this table, there is at least a 0.5 probability that 1 or more SNPs in the region are included in the multi-marker disease model ( $P_1 \geq 0.5$ ); for all T1D loci,  $P_1 \geq 0.8$ . Support for disease associations is conditioned on enrichment of the pathways listed in Table 1. Rows marked with \* are selected only after accounting for pathway enrichment, or show substantial increase in support due to feedback from enrichment. Right-most column cites published GWAS findings that corroborate majority of \* rows. In this column, \*\* indicates that validation is not required as disease association is already strongly supported without pathways; these rows recapitulate the strongest associations reported in the original study [65] (see Supplementary Materials for details). Genes in enriched pathways are written in bold. Table columns from left to right are: (1) disease phenotype; (2) chromosomal locus; (3) region most likely containing the risk-conferring variant(s), in Megabases (Mb); (4) posterior probability that 1 or more SNPs in region are included in model under null, and (5) under enrichment hypothesis; (6) posterior probability that 2 or more SNPs are included under null, and (7) under enrichment hypothesis; (8) smallest trend  $p$ -value in region from original analysis [65], when available (note that some smallest  $p$ -values derived from imputed SNPs are not available in our data); (9) established genes in disease pathogenesis, or most credible genes of interest based on prior studies, corresponding to locus (when the most credible gene differs from gene assigned to pathway, pathway gene is given in parentheses); (10) refSNP identifier of SNP in critical region with largest PIP (this SNP is likely in linkage disequilibrium with the causal variant rather than being causal itself, and may not match the SNP reported in [65] with smallest  $p$ -value); (11) the PIP of that SNP; (12) posterior mean (and 95% credible interval) of  $\beta_j$  or, equivalently, additive effect of minor allele count on log-odds of disease ("log-odds ratio") in multi-marker disease model conditioned on SNP being included in model; (13) frequency of minor allele for that SNP in controls, and (14) in cases. Numbers in bold in  $P_1$  and  $P_2$  columns highlight appreciable increase in support for disease associations within region after feedback from enriched pathway. Credible interval is smallest interval about posterior mean that contains  $\beta_j$  with 95% posterior probability. The "critical region" at each locus is estimated by inspecting single-SNP BF's [79], and bounding the region by areas of high recombination rate, inferred using data from Phase I, release 16a of the HapMap study [142], and visualized in UCSC Genome Browser [143]. Note that  $P_1$  statistic for critical region may be slightly different than  $P_1$  for overlapping segment shown in Fig. 1 due to different numbers of SNPs in regions and segments. All SNP information and genomic positions are based on human genome assembly 17 (NCBI build 35).

enriched genes such as “Cytokine signaling in immune system,” which contains 225 genes, as only a small portion of these genes may actually harbour genetic variants that affect CD risk. Prioritization occurs automatically within our statistical framework; the enrichment parameter affects the prior probability of association for SNPs in the pathway, which in turn increases the posterior probability of association for these SNPs.

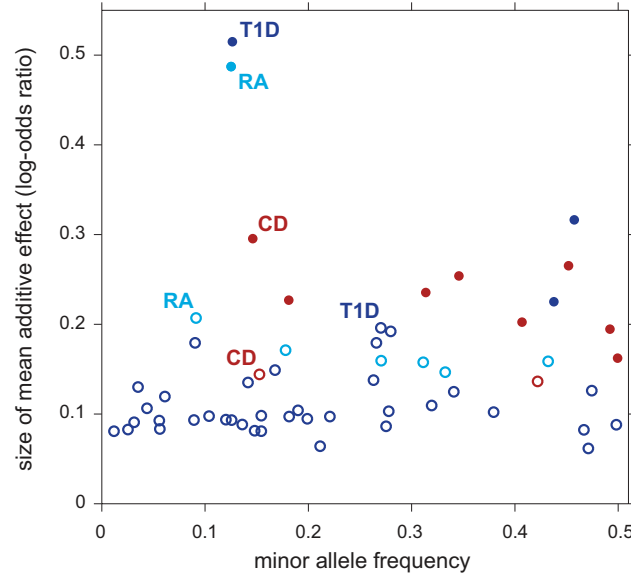
We therefore examine how re-interrogation of SNPs for association in light of inferred enrichments in CD, RA, T1D and T2D can reveal additional associations across the genome. We assess evidence for associations across genomic regions, rather than individual SNPs. The rationale is that genome-wide mapping using a multi-marker disease model sometimes spreads the association signal across nearby SNPs when they are correlated with one another, thereby diluting the signal at any given SNP [76]. We divide the genome into overlapping segments of 50 SNPs, with an overlap of 25 SNPs between neighbouring segments. For each segment, we compute  $P_1$ , the posterior probability that at least one SNP in the segment is included in the multi-marker disease model. (We use  $P_n$  to denote the posterior probability that at least  $n$  SNPs are included.) We use segments with an equal number of SNPs so that, under the null hypothesis of no enrichment, the prior probability that at least one SNP is included is the same for each segment. A segment spans, on average, 307 kb (98% of the segments are between 100 kb and 1 Mb long), so calculating  $P_1$  for these segments provides only a low-resolution map of genetic risk factors for disease. Still, this resolution suffices for our study.

Table 3 summarizes the regions of the genome showing strongest evidence for association after pathway prioritization, and Fig. 1 compares support for disease associations under the null hypothesis of no enrichment with support under the model in which the pathway with the largest BF is enriched.

Overall, prioritization of SNPs in enriched pathways increases support for disease risk factors in many regions, often substantially; these regions correspond to points above the diagonal in the scatterplots (Fig. 1). In CD and RA, 8 disease susceptibility loci with  $P_1 \geq 0.5$ , not including segments overlapping the MHC, are revealed only after prioritizing SNPs in enriched pathways. In T1D, prioritization of SNPs in the IL-2 signaling pathway yields a total of 37 associated regions outside the MHC with  $P_1 \geq 0.5$ . This dramatic result reflects the high estimated enrichment for IL-2 signaling genes. The majority of the additional disease regions with the strongest support, including many of the loci with weaker association signals, are validated by other studies; in CD and RA, 7 of the 8 additional disease susceptibility loci with  $P_1 \geq 0.5$  are corroborated by other GWAS and large-scale meta-analyses, and in T1D, 4 of the 7 additional disease regions with  $P_1 \geq 0.8$  are similarly corroborated (see Table 3 for references).

Prioritization yields many new candidate disease susceptibility loci not previously implicated by GWAS, which will require followup studies to be validated. Three unconfirmed T1D susceptibility loci with strong support ( $P_1 \geq 0.8$ ) are regions containing IL-2 signaling genes *RAF1*, *MAPK14* and *FYN*: gene *RAF1* is a critical target of insulin in primary  $\beta$ -cells, and variants of this gene may modulate loss of  $\beta$ -cell mass in forms of diabetes [144]; *MAPK14* (*p38*) encodes at least 4 distinct isoforms, and deficiencies in one isoform have been shown in mice knockout studies to improve glucose tolerance and protect against insulin resistance, pointing to a role in development of T1D [145]; *FYN* interacts with *PTPN22* to regulate T cell receptor signaling, and *PTPN22* alleles are strongly correlated with predisposition to T1D [146]. The one novel candidate region for RA contains Measles pathway gene *TP73*, whose homolog, *TP53*, is suspected to impair regulation of inflammation in RA patients [147, 148]. Finally, conditioning on enrichment of top-ranked pathways in T2D yields a single novel candidate region at 7q32 with moderate probability of containing a disease association ( $P_1 = 0.48$ ). This region contains *GPR120* (also *OSFAR1*), which is assigned to both the Incretin regulation and GLP-1 pathways. It has recently been shown that *GPR120*-deficient mice develop obesity and reduced insulin signaling, and in humans *GPR120* expression is significantly higher in obese individuals [149], so the effect of this gene may be similar to the reported T2D association with *FTO*, in which variants near *FTO* increase T2D risk through regulation of weight [150, 151].

In addition to these associated regions, several promoted regions also lie within the MHC (in the scatterplots for RA and T1D, these regions correspond to open circles above the diagonal). However, given the complexity of this region, which contains a high density of genes and long-range correlations between SNPs, disentangling the association signal in the MHC will likely require higher density SNP data, and lies beyond the capacities of our current implementation (see Discussion).


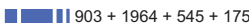

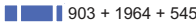





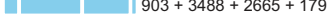














**Figure 2.** Variants in non-MHC disease susceptibility regions informed by enriched pathways have smaller effects on disease risk. Each point in scatterplot corresponds to a 50-SNP segment outside the MHC for which  $P_1 \geq 0.5$ . Filled circles correspond to selected regions containing disease risk factors without feedback from enriched pathways ( $P_1 \geq 0.5 \mid \theta = 0$ ); open circles correspond to selected regions conditioned on enrichment ( $P_1 \geq 0.5 \mid \theta > 0$ ). For each segment, minor allele frequency and posterior mean additive effect of minor allele count on log-odds of disease (“log-odds ratio”) are taken from SNP in 50-SNP segment with highest probability of being included in multi-marker disease model.

Figure 2 compares the effect sizes of variants in regions selected only after accounting for pathway enrichment to the effect sizes from regions identified without the benefit of feedback from pathway enrichment. As expected, pathway prioritization uncovers many disease-associated variants with smaller effects than we would otherwise be able to map reliably. This could explain, at least in part, why many of the putative T1D associations uncovered in our analysis are not yet confirmed; the largest meta-analysis of T1D to date, with a combined sample of size ~16,000 [135], still has limited power to detect associations within this range of effect sizes and minor allele frequencies. (In contrast, much larger meta-analyses exist for CD and RA, with 47,000 and 75,000 samples, respectively [129, 130].)

Finally, Fig. 1 offers the opportunity to remark on four other features of our results. First, the strongest association signals without feedback from pathways stay strong whether or not they are related to the enriched pathway—these correspond to the points in the top-right corner of the scatterplots. (The segments in the top-right corner of each scatterplot recapitulate the strongest associations reported for CD, RA, T1D and T2D in the original study [65]. See Supplementary Materials for a detailed comparison to single-marker  $p$ -values for all diseases.) Second, many segments show slightly decreased support for association under the enrichment hypothesis (points below the diagonal in the scatterplots). This occurs because the estimated prior inclusion probability for SNPs outside the pathway is lowered to reflect the fact that pathway enrichment helps to explain an appreciable portion of the genome-wide association signal. Third, although not evident from the figure due to over-plotting, most segments show little or no evidence for associations under either hypothesis; in each scatterplot, 98–99.7% of segments lie near the bottom-left corner. Fourth, associations with strong support under the null are not necessary for establishing evidence for enriched pathways; none of the RA regions in the top-right corner of the scatterplot contribute to evidence for enrichment of the Measles pathway.

**Table 4.** Enrichment hypotheses in which multiple pathways are enriched show increased support from the data.

enriched pathway(s)	Bayes factor	number of genes	number of SNPs
<b>T1D</b> MHC + IL-2 + ErbB + IPn biosynthesis	$6.9 \times 10^{71}$	 120 + 52 + 13 + 7	 903 + 1964 + 545 + 175
MHC + IL-2 + ErbB receptor signaling	$5.0 \times 10^{69}$	 120 + 52 + 13	 903 + 1964 + 545
MHC + IL-2 signaling	$3.2 \times 10^{66}$	 120 + 52	 903 + 1964
MHC	$2.7 \times 10^{54}$	 120	 903
<b>RA</b> MHC + Measles + Wnt + IPn biosynthesis	$2.0 \times 10^{28}$	 120 + 130 + 77 + 7	 903 + 3488 + 2665 + 179
MHC + Measles + Wnt	$7.0 \times 10^{26}$	 120 + 130 + 77	 903 + 3488 + 2665
MHC + Measles	$6.3 \times 10^{24}$	 120 + 130	 903+3488
MHC	$2.4 \times 10^{21}$	 120	 903
<b>CD</b> IL-23 + cytokine signaling + Met salvage	$5.1 \times 10^{10}$	 225 + 22 + 6	 6711 + 727 + 262
Cytokine + IL-23 signaling	$3.4 \times 10^8$	 225 + 22	 6711 + 727
Cytokine signaling	$9.0 \times 10^5$	 225	 6711

Each row gives the pathway, or combination of 2 or 3 pathways, with the largest BF for enrichment of disease associations. See Table 1 for legend and abbreviations used. All enrichment hypotheses for RA and T1D shown here also include enrichment of the MHC, allowing for a different level of enrichment within the MHC. Unlike the BFs given in Tables 1 and 2, BFs in this table are all defined relative to null hypothesis of no enrichment, so that they can be easily compared. Counts of genes and SNPs only include those that are not already assigned to other enriched pathways; for example, 37 genes belong to the IL-23 pathway, and of those 15 are cytokine signaling genes, so inclusion of IL-23 signaling adds 22 more to the set of genes. Databases and database identifiers for pathways in this table: “IL2-mediated signaling events” (PID il2\_1pathway, BS 137976); “ErbB receptor signaling network” (PID erbb\_network\_pathway, BS 138016); “Inositol pyrophosphates biosynthesis” (HumanCyc 6369, PC); “Measles” (KEGG hsa05162, BS 213306); “Wnt” (Cancer Cell Map, PC); “Cytokine signaling in immune system” (Reactome 75790, BS 366171); “IL23-mediated signaling events” (PID il23pathway, BS 138000); “Methionine salvage pathway” (Reactome 75881, BS 366245).

## Assessing combinations of pathways for enrichment

Above, we obtained evidence for enriched pathways in CD, RA and T1D. The question remains whether a combination of several enriched pathways offers a better fit to the data. A benefit of our approach is that we can compare support for enrichment of different combinations of pathways by comparing their BFs (assuming the same prior for these enrichment hypotheses).

We assess support for combinations of pathways in CD, RA and T1D by computing BFs for models in which 2 and 3 pathways are enriched. Since it is impractical to consider all combinations of 2 and 3 pathways, we tackle this in a “greedy” fashion by selecting combinations of pathways based on the initial ranking; see Methods. Table 4 gives the combinations of 2 and 3 pathways that yield the largest BFs for these diseases. Again, to properly interpret these results we must weigh these BF gains against the relative prior plausibility of the models. Using a “conservative” prior for any pair of pathways being enriched (see “Interpretation of Bayes factors” in Methods), we interpret Table 4 as providing considerable, if short of compelling, support for the hypothesis that 2 pathways are enriched for disease associations in CD, RA and T1D. For example, in CD the BF for enrichment of both cytokine signaling and IL-23 signaling genes is 377 times greater than the BF for enrichment of cytokine signaling genes alone. The BFs for models in which 3 pathways show further increases, but not enough to constitute strong evidence for enrichment of 3 pathways.

We also examine whether enrichment of multiple pathways can lead to identification of additional loci affecting susceptibility to disease. Figure S9 shows that allowing for 2 enriched pathways in CD, RA and T1D does not yield strong support for genetic associations beyond what is already revealed by enrichment of the single top pathway. We do, however, find that a single segment near *IL12B* shows a substantial gain in support for association with CD ( $P_1$  increases from 0.03 to 0.44), and this region is confirmed by other GWAS [6, 7, 129, 152].

## Discussion

Motivated by the observation that it is easier, at least in principle, to identify associations within an enriched pathway, we developed a data-driven approach to simultaneously assess support for enrichment of disease associations in pathways and prioritize variants in enriched pathways. We investigated the merits and limitations of this approach in a detailed analysis of data sets for seven complex diseases. We interrogated thousands of candidate pathways from multiple pathway databases, finding strong evidence linking pathways to pathogenesis of several diseases. By promoting variants within the enriched pathways identified in our analysis, we mapped disease susceptibility loci beyond those identified by a conventional analysis.

The CD and RA results provided some validation for our methods, as all but one of the additional disease associations identified by pathway prioritization are corroborated by other studies. The T1D results also provided some validation for our methods, as several of the strongest associations informed by enrichment of the IL-2 signaling pathway are confirmed in other GWAS for T1D. Prioritizing IL-2 signaling genes revealed other regions relevant to T1D that could not be corroborated by other GWAS, and this may be because the largest GWAS for T1D to date does not match the scale of the largest studies for CD and RA. All the disease associations informed by enriched pathways had smaller effects on disease susceptibility, illustrating how pathway prioritization can help overcome some of the constraints on our ability to reliably detect disease-conferring variants with small effects in GWAS.

Our approach builds on methods that use multi-marker models for simultaneous mapping of genetic variants in GWAS [61, 75, 76, 153–161]. In contrast to single-marker regression approaches, these methods model susceptibility to disease by the combined effect of multiple variants, and use sparse multivariate regression techniques to fit multi-marker (*i.e.* polygenic) models to the data. Within a multi-marker model of disease, estimating enrichment of a candidate pathway effectively reduces to counting, inside and outside the pathway, variants associated with disease (more precisely, we are counting the variants that are included in the multi-marker disease model). Our approach to combining multi-marker modeling with pathway analysis offers several benefits. First, unlike many pathway analysis methods that test for enrichment of significant SNPs or genes within a pathway [24, 25], we have no need to select a threshold to determine which  $p$ -values are significant; instead, we use the association signal from all variants to assess enrichment. Second, by analyzing variants simultaneously, we avoid exaggerating evidence for enrichment from disease-associated variants that are correlated with each other (*i.e.* in linkage disequilibrium), while still allowing multiple independent association signals near a gene to contribute evidence for enrichment. Third, and most importantly, quantifying enrichment within this framework gives us feedback about associations within enriched pathways, potentially leading to discovery of novel genetic loci underlying disease.

In contrast to many pathway analysis methods, we modeled enrichment of disease associations at the level of variants, rather than genes. While there are arguments for both approaches, a feature of the variant-based approach is that, when there are multiple variants near a gene that affect disease susceptibility, all these signals contribute to the evidence for enrichment of the pathways containing this gene.

Another important feature of our approach is that it can be used to assess models in which multiple pathways are enriched. Examining combinations of pathways for enrichment may highlight pathways that would otherwise not be highly ranked, and the results on RA and T1D provided vivid examples of this; evidence for enrichment of the Measles and IL-2 pathways only became compelling once we assessed support for enrichment of these pathways together with enrichment of the MHC.

Our results focused on the regions showing the strongest evidence for association with disease. However, the large number of points approaching the middle of the vertical axes in Fig. 1 suggests that many other gene variants in the enriched pathways may contribute to risk of CD, RA and T1D; from our estimates of  $\theta_0$  and  $\theta$  (Table 2), approximately 38, 45 and 59 independent risk variants are (in expectation) hidden among Measles, cytokine signaling and IL-2 signaling genes. This suggests that more associations in these pathways remain to be discovered.

Several selected disease susceptibility regions listed in Table 3 contain multiple candidate genes, including cases where the gene in the enriched pathway is not the same as the most credible gene suggested in prior studies. It

is possible that the pathway annotations would be useful to help pinpoint, or *fine-map*, the genes or variants relevant to disease within these regions. However, investigating this would require advances to our current methodology, as the approximations we made to improve the efficiency of our approach, building on earlier work [75], are less appropriate for refining the location of association signals, and they will need to be modified to accommodate this goal. We note, however, that some of these regions may contain multiple variants that disrupt or regulate different genes relevant to disease, and our methods can help assess this possibility. For example, we calculate that multiple independent risk variants reside at the 16p13 locus with probability  $P_2 = 0.75$  (Table 3), so it is possible that both *C1QTNF6* and *IL2RB* at this locus are associated with T1D risk variants.

A limitation of our current approach is that the prior variance of additive effects on disease risk must be chosen beforehand. We based our choice on the distribution of odds ratios reported in published genome-wide association studies, and checked that the ranking of enriched pathways was robust to different prior choices (see Methods). One problem with this prior is that published associations typically have the largest effects on disease risk, as these are the associations we typically have adequate power to identify, resulting in a prior that places too much weight on larger additive effects. It would be preferable to estimate this prior from the data instead, but we found that this worked poorly in practice. The likely cause of this problem is that the non-zero effects on disease are not normally distributed, contrary to our assumptions. One possible solution would be to use a more flexible prior that is better able to capture the distribution of additive effects, such as a mixture of two or more normals [80].

In summary, our results on a range of complex diseases illustrate how an integrated approach to identification of enriched pathways, and prioritization of variants within enriched pathways, can uncover additional disease associations beyond standard statistical procedures based on single-marker regression. Our results point to the potential for applying our methods to other common diseases, and larger studies, to uncover genetic loci that have not yet been identified as risk factors for disease.

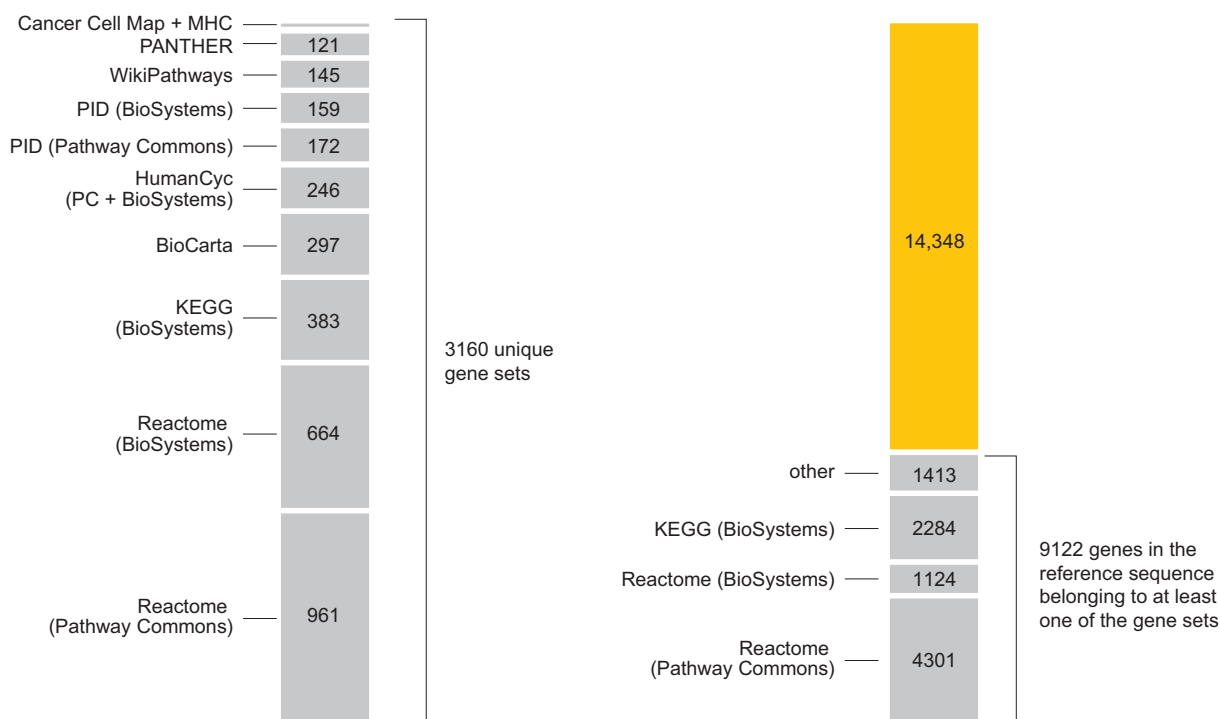
## Methods

### Samples

Results on all seven diseases are based on genome-wide marker data from the case-control studies described in the original WTCCC study [65]. For all diseases, the control samples come from two groups: 1480 individuals from the 1958 Birth Cohort (58BC), and 1458 individuals from the UK Blood Services (UKBS) cohort. All subjects are from Great Britain, and are of self-described European descent. Genetic associations from these studies were first reported in [65].

All study subjects were genotyped for roughly 500,000 SNPs on autosomal chromosomes using a commercial version of the Affymetrix GeneChip 500K platform. We estimate missing genotypes at the SNPs using the mean posterior minor allele count from BIMBAM [79, 162], with SNP data from Phase II of the International HapMap Consortium project [163]. To be consistent with the original analysis, refSNP identifiers and locations of SNPs are based on human genome reference assembly 17 (NCBI build 35).

We apply quality control filters as described in [65], and remove SNPs that exhibit no variation in the sample. For all diseases, we include an additional quality control measure to filter out potentially problematic SNPs. Some SNPs with high minor allele frequencies (MAFs) show moderate evidence for association based on our calculations—“single-SNP” BFs [79] in which prior standard deviation of the log-odds ratios is set to 0.1—but because they do not appear to be supported by nearby SNPs upon inspecting the single-SNP BFs for these SNPs, we cannot rule out the possibility of genotyping errors. Based on this criterion, we discard 2 additional SNPs in CD, rs1914328 on chromosome 8 at 69.45 Mb (BF =  $6.6 \times 10^3$ , MAF = 0.43), and rs6601764 on chromosome 10 at 3.85 Mb (BF =  $4.3 \times 10^3$ , MAF = 0.43). No nearby SNPs have single-SNP BFs greater than 46. For CAD, we discard SNP rs6553488 on chromosome 4 at 171.4 Mb (BF =  $1.4 \times 10^3$ , MAF = 0.46). No nearby SNPs have a single-SNP BF greater than 11. Following the same quality control criterion, we do not filter out any additional SNPs in the other data sets. Table S1 summarizes the data used in our analysis after following these steps to better ensure data quality.



**Figure 3.** Summary of pathways used in the analysis. Chart on left shows the number of unique gene sets obtained from the following pathway databases, in this order: Reactome [164], Kyoto Encyclopedia of Genes and Genomes (KEGG) [165], BioCarta ([www.biocarta.com](http://www.biocarta.com)), HumanCyc [166, 167], NCI Nature Pathway Interaction Database (PID) [93], WikiPathways [168, 169], PANTHER [170] and Cancer Cell Map ([cancer.cellmap.org](http://cancer.cellmap.org)). The majority of these pathways are retrieved from the Pathway Commons (PC) [95] and NCBI BioSystems [94] repositories, and we include gene sets from both repositories when gene sets from same pathway differ (see Supplementary Materials). We include 2 additional gene sets for the “classical” and “extended” MHC [90, 91]. Right-hand chart shows gains in gene coverage by including additional databases in the analysis, where “gene coverage” is defined as any genes in reference sequence that are assigned to the at least one pathway. From the total of 3160 gene sets (including the MHC and xMHC), we achieve coverage of ~39% of genes in the reference sequence (see Fig. S1).

## Pathways, and assignment of SNPs to genes in pathways

We aim for a comprehensive evaluation of pathways accessible on the Web in standard, computer-readable formats [63, 64, 171]. Since the results hinge on the quality of the pathways used in our analysis, we restrict the analysis to curated, peer-reviewed pathways based on experimental evidence, and pathways inferred via gene homology. We draw candidate pathways from the collections listed in Fig. 3 (see also Supplementary Materials). KEGG [165] and HumanCyc [166] are primarily databases of metabolic pathways, and are unlikely to be relevant to some autoimmune diseases, but we include them in the analysis of all diseases for completeness. We create 2 additional gene sets to assess support for enrichment of disease associations within the MHC and the “extended” MHC [90, 91]. We treat each candidate pathway as a set of genes, ignoring details such as molecules involved in biochemical reactions, and cellular locations of these reactions.

Many of the pathways in the databases are arranged hierarchically; we include all elements of the hierarchy in our analysis. Elements in upper levels of the hierarchy refer to groups of pathways with shared attributes or a common function. Some gene groups have a broad definition, such as “Immune system” in Reactome (ID 6900), which includes pathways involved in adaptive and innate immune response. Enrichment of a large gene set is unlikely to provide much insight into disease pathogenesis. However, a key step in our analysis is to re-interrogate SNPs for association in light of inferred enrichments. Thus, enrichment of a broad physiological target such as “Immune system” can be useful if subsequent re-interrogation reveals associations that were not significant in a



conventional analysis.

Since we combine pathways from different sources, we encounter pathways with inconsistent definitions [172, 173]; see Supplementary Materials. There is no single explanation for the lack of consensus in pathway definitions, and we have no reason to prefer one definition over another, so we include multiple versions of a pathway in our analysis.

Based on findings that the majority of variants modulating gene expression lie within 100 kb of the gene’s transcribed region [174–176], we assign a SNP to a gene if it is within 100 kb of the transcribed region. Others have opted for a 20 kb window [69, 73] based on findings that *cis*-acting expression QTLs are rarely more than 20 kb from the gene [62]. We chose a broader region since the benefit of including potentially relevant SNPs in a pathway when the association signal is sparse seems likely to outweigh the cost of including a larger number of irrelevant markers.

## Selection of combinations of pathways

In our case studies on CD, RA and T1D, we compute BF<sub>s</sub> to assess support for models in which 2 and 3 pathways are enriched for disease associations. Since it is impractical to consider all combinations of 2 and 3 pathways, we tackle this in a “greedy” fashion by selecting combinations of pathways based on the initial ranking. Our strategy is to select the pathway with the largest BF (Table 1), and assess support for this pathway in combination with pathways from a larger set of candidates (we take all pathways with BF > 10). This greedy heuristic makes it feasible to evaluate many combinations of pathways that could plausibly be jointly enriched, though it does not consider all combinations, so may miss a combination with stronger evidence for enrichment. In total, we compute BF<sub>s</sub> for 85, 24 and 408 pairs of pathways in CD, RA and T1D, respectively. (Note that the models for RA and T1D also include enrichment of the MHC.) For completeness, we extend the analysis to models with 3 enriched pathways. Following the same greedy strategy, we take the top pair of pathways (Table 4) and combine it with individual pathways with BF > 10.

## Statistical analysis

The Bayesian variable selection approach to simultaneous interrogation of SNPs involves fitting a multi-marker disease model to the data with different combinations of SNPs. By accounting for correlations between markers, fitting all markers simultaneously allows us to identify those that are *independently associated*—these markers each signal a variant that contributes to disease risk independently of other risk-conferring variants.

### Likelihood

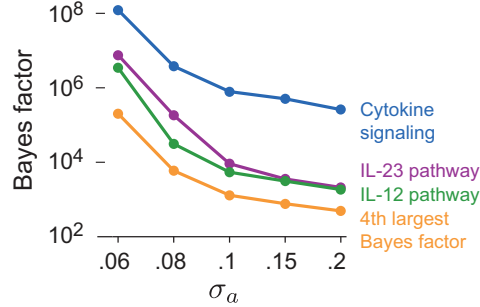
The likelihood specifies the probability of observing disease (case-control) status  $y$  given the genotypes  $\mathbf{X}$ , the intercept  $\beta_0$ , and the regression coefficients  $\beta = (\beta_1, \dots, \beta_p)$ . From the additive model for the log-odds of disease (eq. 1),  $p_i = \phi(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j)$  is the probability that  $y_i = 1$ , in which  $\phi(x) = 1/(1 + e^{-x})$  is the sigmoid function. Assuming independence of the observations  $y_i$ , the likelihood is

$$p(y | \mathbf{X}, \beta_0, \beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}. \quad (4)$$

### Priors

Next we specify prior distributions for the genome-wide log-odds  $\theta_0$ , log-fold enrichment parameter  $\theta$ , intercept  $\beta_0$ , and coefficients  $\beta_j$  of SNPs included in the multi-marker disease model.

Since inferences strongly depend on  $\theta_0$ , and since  $\theta_0$  is unknown and will be different in each setting, we estimate this parameter from the data. Following [75, 76], we assign a uniform prior to  $\theta_0$ . We restrict  $\theta_0$  to



**Figure 4.** Top 4 BF's for each setting of  $\sigma_a$ . In each case, the 3 largest BF's correspond, in order, to cytokine signaling in immune system, IL23-mediated signaling events, and IL12-mediated signaling events (these are the top 3 pathways for CD given in Table 2). The pathway with the fourth-largest BF differs across settings of  $\sigma_a$ .

$[-6, -2]$ , so as few as 0 and as many as  $\sim 4400$  SNPs are expected to be included *a priori*. We assign a uniform prior to  $\theta$  on interval  $[0, 5]$ , which allows for a wide range of enrichments.

For the prior on the non-zero coefficients  $\beta_j$ , we follow standard practice that assumes they are *i.i.d.* normal with zero mean and standard deviation  $\sigma_a$  [177]. Ordinarily, to combat sensitivity of the results to the choice of  $\sigma_a$ , we would place a prior on  $\sigma_a$  and integrate over this parameter to let the data drive selection of  $\sigma_a$ . This approach is taken in [75, 76]. But in our case we find that the heterogeneity of the odds ratios in complex diseases presents a problem: although we expect most odds ratios for a common disease—and specifically odds ratios in a pathway relevant to disease pathogenesis—to be close to 1, the odds ratios corresponding to the strongest disease associations drive estimates of  $\sigma_a$  toward larger values, and a normal distribution that puts too little weight on modest odds ratios. One possible strategy would be to redo the analysis after removing associated regions with the largest odds ratios, but this is an unattractive solution because SNPs with large odds ratios would not contribute to the evidence for enrichment. Instead, we fix  $\sigma_a$ , grounding the choice on typical odds ratios reported in published GWAS, and we assess the robustness of our findings to this choice. Our choice is  $\sigma_a = 0.1$ , which favours odds ratios close to 1 (95% of the odds ratios lie between 0.82 and 1.22 *a priori*), while being large enough to capture a significant fraction of the odds ratios for common alleles reported in genome-wide association studies of complex disease traits. According to a recent review [178], approximately 40% of estimated odds ratios are between 1.1 and 1.2, and an additional 10% of odds ratios are smaller than 1.1. This prior closely corresponds to a survey of odds ratios reported in genetic association studies of common diseases [179]. Since there may be justification for a slightly smaller or slightly larger  $\sigma_a$ , we also try different values for  $\sigma_a$ , and examine how these choices affect the ranking of enriched pathways in the CD data set (see below).

To complete the probability model, we assign an improper uniform prior to the intercept,  $\beta_0$ . In general, one must be careful with use of improper priors in Bayesian variable selection because they can result in improper posteriors. A sufficient condition for a proper posterior, and a well-defined BF, with logistic regression (eq. 1) is that the maximum likelihood estimator of  $\beta$  conditioned on which variables are included in the model, and on the other model parameters, is unique and finite [180]. This condition is difficult to check exhaustively in Bayesian variable selection, but we can at least guarantee that the posterior is proper under the variational approximation (see Supplementary Materials) so long as the coordinate ascent steps converge to a unique solution.

### Sensitivity of pathway ranking to prior distribution of odds ratios

A concern with our choice of prior for the regression coefficients is that slightly smaller or slightly larger settings of  $\sigma_a$  could also be justified, and these choices could produce different results. Associations are unlikely to accumulate at a greater rate in pathways that are not related to the disease, even associations with small effects on disease risk, so we predict that the ranking of enriched pathways is largely robust to the choice of  $\sigma_a$ . Here we verify this claim on the CD data set. We assess the sensitivity of our results to  $\sigma_a$  by recomputing the BF's for all

candidate pathways with prior choices that favor slightly smaller ( $\sigma_a = 0.06, 0.08$ ) and slightly larger coefficients ( $\sigma_a = 0.15, 0.2$ ). Fig. 4 shows that smaller settings of  $\sigma_a$  yield substantially more support for enrichment of CD-related pathways, as expected. But the pathways with the largest BFs remain IL-23, IL-12 and cytokine signaling regardless of the choice of  $\sigma_a$ . In the Supplementary Materials, we show that the BFs for most of the other candidate pathways do not change noticeably at different settings of  $\sigma_a$ .

## Bayes factors

We take the *Bayesian model averaging* strategy [82, 154, 181] to account for possible uncertainty in  $\theta_0$  and  $\theta$  when evaluating the BFs (eq. 3). The likelihood under the enrichment hypothesis ( $\theta > 0$ ) and the likelihood under the null ( $\theta = 0$ ) are each expressed as an average over possible assignments to  $\theta_0$  and  $\theta$ :

$$\text{BF}(a) = \frac{\iint p(y | \mathbf{X}, a, \theta_0, \theta) p(\theta_0) p(\theta) d\theta d\theta_0}{\int p(y | \mathbf{X}, a, \theta_0, \theta = 0) p(\theta_0) d\theta_0}. \quad (5)$$

Each instance of  $p(y | \mathbf{X}, a, \theta_0, \theta)$  in (5) expands as an average over possible assignments to the intercept  $\beta_0$  and regression coefficients  $\beta$ :

$$p(y | \mathbf{X}, a, \theta_0, \theta) = \iint p(y | \mathbf{X}, \beta_0, \beta) p(\beta_0) \prod_{j=1}^p p(\beta_j | a_j, \theta_0, \theta) d\beta_0 d\beta. \quad (6)$$

The factors in this equation are all determined by the model and priors specified above. Factor  $p(\beta_j | a_j, \theta_0, \theta) = \pi_j N(0, \sigma_a^2) + (1 - \pi_j) \delta_0$  is the “spike and slab” prior [177, 182], in which  $\pi_j = p(\beta_j \neq 0)$  is specified by eq. 2. Here,  $\delta_0$  denotes the delta mass, or “spike”, at zero, and  $N(\mu, \sigma^2)$  is the normal density with mean  $\mu$  and variance  $\sigma^2$ . Factor  $p(\beta_0)$  is the (improper) uniform prior, and  $p(y | \mathbf{X}, \beta_0, \beta)$  is the logistic regression likelihood (4). Computation of the BFs is described in the Supplementary Materials.

## Interpretation of Bayes factors

Given that enrichment analyses typically proceed by computing  $p$ -values and assessing “significance,” one may reasonably ask whether a given BF represents “significant” evidence for enrichment. Specifying an appropriate threshold for a BF to be considered significant, however, is context-dependent, and subjective. This is because the *posterior odds* for a pathway being enriched, relative to the null hypothesis that no pathways are enriched, is equal to the Bayes Factor times the prior odds for enrichment, and the prior odds for each pathway depends on how plausible it is, *a priori*, that the pathway is relevant to the disease. (Similar issues arise when specifying significance thresholds for  $p$ -values; for example, the false discovery rate at a given  $p$ -value threshold depends on the prior probability of enrichment [183, 184]. But in practice significance thresholds of 0.05 or 0.01 are often used without attending to this concern.) Nonetheless, we can make the following observations. First, if we are willing to assume the pathways in Tables 1 and 2 are all equally plausible candidates for enrichment *a priori*, then the ratio of the BFs indicates the relative support for the enrichment hypotheses; for example, if we must choose between enrichment of cytokine signaling genes and the IL-23 signaling pathway, the data overwhelmingly favour the former by a factor of  $\frac{9.0 \times 10^5}{1.4 \times 10^4} \approx 64$  (Table 2). Second, even under a “conservative” prior for enrichment in which we expect 1 pathway to be enriched among the 3158 candidates, corresponding to a prior odds of 1/3158, the top pathways in CD and T1D are large enough to strongly support enrichment (*i.e.* they are all much greater than 3158). Weighing this prior against the BF for the top pathway in RA does not yield strong support for this pathway, but given its plausible connection to RA, we view this enrichment result as compelling.

We can apply similar reasoning to weigh the evidence for hypotheses in which 2 or more pathways are enriched. For example, the model in which both cytokine signaling and IL-23 pathway genes are enriched for CD associations has a BF that is about 400 times greater than the BF for enrichment of cytokine signaling genes alone, which indicates that the best model with 2 enriched pathways provides a much better fit to the CD data than the best

model with any one enriched pathway. However, to properly interpret this result we must weigh this increase in the BF against the relative prior plausibility of the models. A naive argument using a “conservative” prior for any pair of pathways being enriched might suggest a prior odds of  $1/(3157 \times 3158)$ . This prior would make a 400-fold increase in the BF appear to be relatively insignificant. However, this argument not only depends on the earlier prior, which may be overly conservative, but also assumes independence of enriched pathways, which seems unwise considering that many pathways mentioned here have related roles in immunity; *a priori*, one might expect that a pathway is more likely to be enriched when a related pathway is enriched.

### Posterior inclusion probabilities and other posterior quantities

In this section, we define posterior inclusion probabilities (PIPs) and other posterior quantities used in the results. In all cases, posterior statistics under the null hypothesis are obtained by setting  $\theta = 0$ .

Like the Bayes factors, the PIPs are obtained by averaging over  $\theta_0$  and  $\theta$ . Taking  $\mathcal{D} = \{\mathbf{X}, y, a\}$  as shorthand for the GWAS data, we have

$$\text{PIP}(j) \equiv p(\beta_j \neq 0 | \mathcal{D}) = \iint p(\beta_j \neq 0 | \mathcal{D}, \theta_0, \theta) p(\theta_0, \theta | \mathcal{D}) d\theta_0 d\theta, \quad (7)$$

where  $p(\beta_j \neq 0 | \mathcal{D}, \theta_0, \theta)$  is the posterior inclusion probability for SNP  $j$  given hyperparameter setting  $(\theta_0, \theta)$ , and  $p(\theta_0, \theta | \mathcal{D})$  is the posterior probability of hyperparameter setting  $(\theta_0, \theta)$ . Calculation of these posterior probabilities is described in Supplementary Materials.

To identify regions of the genome associated with disease risk, we calculate for each region the posterior probability that at least 1 SNP in the region is included in the multi-marker disease model. Let  $S = n$  represent the event that exactly  $n$  SNPs in a given region are included in the multi-marker disease model, so that  $P_n \equiv p(S \geq n | \mathcal{D})$ . These posterior probabilities are easily calculated from the PIPs (7) using the variational approximation; see Supplementary Materials.

Since no single pathway stands out in Table 2 as having greatest support for enrichment of T2D associations, we compute posterior quantities by averaging over different enrichment models with the largest BFs, weighting these models by their BFs. We do the same for models in which 2 or 3 pathways are enriched. (Implicitly, this assumes that all models equally plausible *a priori*.) The ability to average across models in this way is an advantage the Bayesian model comparison approach, because it allows us to assess genetic associations in light of the enrichment evidence without having to choose a single enrichment model. Suppose we have  $m$  enrichment models, specified by their SNP annotations  $a^{(1)}, \dots, a^{(m)}$ , with corresponding BFs,  $\text{BF}(a^{(1)}), \dots, \text{BF}(a^{(m)})$ . Then  $P_n$  is given by

$$P_n = \frac{p(S \geq n | \mathcal{D}, a^{(1)}) \text{BF}(a^{(1)}) + \dots + p(S \geq n | \mathcal{D}, a^{(m)}) \text{BF}(a^{(m)})}{\text{BF}(a^{(1)}) + \dots + \text{BF}(a^{(m)})}. \quad (8)$$

See Supplementary Materials for further details about computation of relevant posterior quantities.

### Population stratification

Pathway enrichment analysis should be robust to population stratification because spurious associations that arise from population structure are unlikely to accumulate at a greater rate in the pathway. Further, the original report [65] and subsequent analyses [6, 185] affirm that cryptic population structure does not have a substantive impact in these data. Thus we did not correct for population structure in our analysis.

### Modifications to analysis to account for large contributions of MHC alleles to RA and T1D risk

Recent work by Pirinen *et al* [186] has shown that when analyzing a case-control study with a prospective model, as we do here, controlling for non-confounding covariates—particularly those with large effects—can reduce power to detect associations. Their work implies that joint modeling of the effects of multiple SNPs on disease risk could actually reduce power to detect associations compared with single-marker tests, particularly for diseases such as

RA and T1D in which MHC variants are known to have large effects on disease risk. We briefly explain how we modify our approach to address this issue.

Pirinen *et al* [186] recommend omitting non-confounding covariates when the goal is to discover new loci associated with disease. SNPs in the MHC associated with disease are non-confounding covariates of large effect, so we could adhere to their advice simply by ignoring all MHC SNPs when analyzing the non-MHC SNPs. However, omitting the MHC would prevent us from effectively interrogating pathways that contain MHC genes. Instead, we develop the following two-stage procedure: first, we fit our model using only SNPs outside the extended MHC to estimate their effects,  $\beta_{\overline{\text{MHC}}}$ ; second, fixing  $\beta_{\overline{\text{MHC}}}$  to the estimates obtained from the first step, we fit the multi-marker model to the full data to estimate  $\beta_{\text{MHC}}$ , the effects of SNPs within the xMHC. This procedure is equivalent to assuming that  $p(\beta_{\overline{\text{MHC}}} | \mathcal{D}, \theta_0, \theta, \beta_{\text{MHC}}) \approx p(\beta_{\overline{\text{MHC}}} | \mathcal{D}, \theta_0, \theta)$ . (See Supplementary Materials for details on how this is incorporated into the variational approximation to efficiently compute BFs and posterior statistics.) Thus, in the first stage we ignore the MHC when deciding which non-MHC SNPs to include in the multi-marker model, following the advice from [186], but in the second stage we allow that the MHC may contribute evidence toward pathway enrichment when the pathway includes MHC genes.

We found that addressing these issues was important for analyzing the RA and T1D data; with these modifications, our results with the multi-marker model more closely replicate the original single-marker association analysis (Figures S5 and S6), and our methods yield more support for enrichment of pathways that overlap the MHC.

### Analyses conditional on MHC enrichment

Once we establish that the MHC has far greater support for enrichment of disease associations in RA and T1D than any other candidate pathway, we condition on enrichment of the MHC, and search for gene sets with evidence for enrichment beyond the MHC. To speed up computation, we fix the MHC enrichment parameter,  $\theta_{\text{MHC}}$ , to its *maximum a posteriori* estimate, and we assume that the posterior distribution of  $\beta_{\text{MHC}}$  is unaffected by enrichment of the candidate pathway; that is,  $p(\beta | \mathcal{D}, \theta_0, \theta_{\text{MHC}}, \theta) \approx p(\beta_{\text{MHC}} | \mathcal{D}, \theta_0, \theta_{\text{MHC}}) p(\beta_{\overline{\text{MHC}}} | \mathcal{D}, \theta_0, \theta)$ . (Note that we make a similar approximation to improve the efficiency of calculating BFs for all pathways. We assume that the coefficients  $\beta_j$  for all SNPs  $j$  outside the enriched pathway are unaffected by the pathway enrichment *a posteriori*; see Supplementary Materials.) To assess the support for enrichment of a candidate pathway in RA and T1D, we compare the likelihood given the model in which the candidate pathway and the MHC are enriched to the likelihood given the model in which only the MHC is enriched. Thus, unless otherwise specified, all BFs for enrichment of pathways in RA and T1D are defined relative to the “null” that the MHC is enriched, rather than the null of no enrichment.

### Software availability

MATLAB and R implementations of the statistical methods described here, and the MATLAB scripts used to implement the steps in our analysis, are available for download through <http://github.com/pcarbo/bmapathway>.

## Acknowledgments

Thanks to Yongtao Guan for assistance with the data sets, Kevin Bullaughey and John Zekos for expert technical support, Emek Demir, Lewis Geer, Benjamin Cross and the rest of the Pathway Commons team for help with pathway databases, and Gorka Alkorta-Aranburu, Niall Cardin, Anna Di Rienzo, Hariklia Eleftherohorinou, Timothée Flutre, Stoyan Georgiev, Ron Hause, Bryan Howie, Ellen Leffler, Dan Nicolae, Heejung Shim, Xiaoquan Wen and Xiang Zhou for helpful discussions and advice. We also thank the three reviewers for their detailed comments and suggestions. This work was supported by a grant from the National Institute of Health (HG02585), and a cross-disciplinary postdoctoral fellowship from the Human Frontiers Science Program.

## References

1. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322: 881–888.
2. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* 10: 241–251.
3. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 9: 356–369.
4. Pearson TA, Manolio TA (2008) How to interpret a genome-wide association study. *Journal of the American Medical Association* 299: 1335–1344.
5. Abraham C, Cho JH (2009) Inflammatory bowel disease. *New England Journal of Medicine* 361: 2066–2078.
6. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, et al. (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics* 40: 955–962.
7. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics* 42: 1118–1125.
8. Khor B, Gardet A, Xavier RJ (2011) Genetics and pathogenesis of inflammatory bowel disease. *Nature* 474: 307–317.
9. Stappenbeck TS, Rioux JD, Mizoguchi A, Saitoh T, Huett A, et al. (2011) Crohn disease: a current perspective on genetics, autophagy and immunity. *Autophagy* 7: 355–374.
10. Van Limbergen J, Wilson DC, Satsangi J (2009) The genetics of Crohn's disease. *Annual Review of Genomics and Human Genetics* 10: 89–116.
11. Ballard D, Abraham C, Cho J, Zhao H (2010) Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Medical Genomics* 3: 25.
12. Braun R, Buetow K (2011) Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. *PLoS Genetics* 7: e1002101.
13. Chen X, Wang L, Hu B, Guo M, Barnard J, et al. (2010) Pathway-based analysis for genome-wide association studies using supervised principal components. *Genetic Epidemiology* 34: 716–724.
14. Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, et al. (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *American Journal of Human Genetics* 86: 860–871.
15. De la Cruz O, Wen X, Ke B, Song M, Nicolae DL (2010) Gene, region and pathway level analyses in whole-genome studies. *Genetic Epidemiology* 34: 222–231.
16. Eleftherohorinou H, Wright V, Hoggart C, Hartikainen A, Jarvelin M, et al. (2009) Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE* 4: e8068.
17. Holden M, Deng S, Wojnowski L, Kulle B (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24: 2784–2785.
18. Jia P, Wang L, Fanous AH, Chen X, Kendler KS, et al. (2012) A bias-reducing pathway enrichment analysis of genome-wide association data confirmed association of the MHC region with schizophrenia. *Journal of Medical Genetics* 49: 96–103.
19. Lee PH, O'Dushlaine C, Thomas B, Purcell SM (2012) INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics* 28: 1797–1799.
20. Ramanan VK, Shen L, Moore JH, Saykin AJ (2012) Pathway analysis of genomic data: concepts, methods and prospects for future development. *Trends in Genetics* 28: 323–332.
21. Ruano D, Abecasis GR, Glaser B, Lips ES, Cornelisse LN, et al. (2010) Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *American Journal of Human Genetics* 86: 113–125.
22. Shahbaba B, Shachaf CM, Yu Z (2012) A pathway analysis method for genome-wide association studies. *Statistics in Medicine* 31: 988–1000.
23. Torkamani A, Topol EJ, Schork NJ (2008) Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 92: 265–272.
24. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics* 11: 843–854.
25. Wang L, Jia P, Wolfinger RD, Chen X, Zhao Z (2011) Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* 98: 1–8.

26. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, et al. (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics* 86: 929–942.
27. Yaspan BL, Veatch OJ (2011) Strategies for pathway analysis from GWAS Data, John Wiley and Sons, Inc., volume 71, chapter 1.20. pp. 1–15.
28. Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, et al. (2009) Pathway analysis by adaptive combination of P-values. *Genetic Epidemiology* 33: 700–709.
29. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *American Journal of Human Genetics* 86: 6–22.
30. Hartwell L (2004) Robust interactions. *Science* 303: 774–775.
31. Hirschhorn JN (2009) Genomewide association studies—illuminating biologic pathways. *New England Journal of Medicine* 360: 1699–1701.
32. Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. *Nature* 461: 218–223.
33. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* 11: 446–450.
34. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461: 747–753.
35. Rioux JD, Abbas AK (2005) Paths to understanding the genetic basis of autoimmune disease. *Nature* 435: 584–589.
36. Ropers HH (2007) New perspectives for the elucidation of genetic disorders. *American Journal of Human Genetics* 81: 199–207.
37. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. *Nature Biotechnology* 24: 537–544.
38. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, et al. (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human Molecular Genetics* 18: 2078–2090.
39. Chen M, Cho J, Zhao H (2011) Incorporating biological pathways via a Markov random field model in genome-wide association studies. *PLoS Genetics* 7: e1001353.
40. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics* 78: 1011–1025.
41. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology* 25: 309–316.
42. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, et al. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genetics* 5: e1000534.
43. Saccone SF, Saccone NL, Swan GE, Madden PAF, Goate AM, et al. (2008) Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics* 24: 1805–1811.
44. Tranchevent L, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, et al. (2011) A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics* 12: 22–32.
45. Wu X, Jiang R, Zhang MQ, Li S (2008) Network-based global inference of human disease genes. *Molecular Systems Biology* 4: 189.
46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25–29.
47. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* 22: 1790–1797.
48. Degner JF, Pai AA, Pique-Regi R, Veyrieras J, Gaffney DJ, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482: 390–394.
49. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* 6: e1000888.
50. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. *Genome Research* 22: 1748–1759.
51. Ward LD, Kellis M (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology* 30: 1095–1106.

52. Capanu M, Orlow I, Berwick M, Hummer AJ, Thomas DC, et al. (2008) The use of hierarchical models for estimating relative risks of individual genetic variants: an application to a study of melanoma. *Statistics in Medicine* 27: 1973–1992.
53. Capanu M, Concannon P, Haile RW, Bernstein L, Malone KE, et al. (2011) Assessment of rare BRCA1 and BRCA2 variants of unknown significance using hierarchical modeling. *Genetic Epidemiology* 35: 389–397.
54. Chen GK, Witte JS (2007) Enriching the analysis of genome-wide association studies with hierarchical modeling. *American Journal of Human Genetics* 81: 397–404.
55. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* 12: 628–640.
56. Fridley BL, Serie D, Jenkins G, White K, Bamlet W, et al. (2010) Bayesian mixture models for the incorporation of prior knowledge to inform genetic association studies. *Genetic Epidemiology* 34: 418–426.
57. Fridley BL, Lund S, Jenkins GD, Wang L (2012) A Bayesian integrative genomic model for pathway analysis of complex traits. *Genetic Epidemiology* 36: 352–359.
58. Gaffney DJ, Veyrieras J, Degner JF, Roger P, Pai AA, et al. (2012) Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology* 13: R7.
59. Lee S, Dudley AM, Drubin D, Silver PA, Krogan NJ, et al. (2009) Learning a prior on regulatory potential from eQTL data. *PLoS Genetics* 5.
60. Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC (2007) Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic Epidemiology* 31: 871–883.
61. Swartz MD, Kimmel M, Mueller P, Amos CI (2006) Stochastic search gene suggestion: a Bayesian hierarchical model for gene mapping. *Biometrics* 62: 495–503.
62. Veyrieras JB, Kudaravalli S, Kim SY, Dermizakis ET, Gilad Y, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics* 4: e1000214.
63. Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Research* 34: D504–D506.
64. Bauer-Mehren A, Furlong LI, Sanz F (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular Systems Biology* 5: 290.
65. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
66. Askland K, Read C, Moore J (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Human Genetics* 125: 63–79.
67. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJM (2011) Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Human Molecular Genetics* 20: 3494–3506.
68. Freudenberg J, Lee AT, Siminovich Ka, Amos CI, Ballard D, et al. (2010) Locus category based analysis of a large genome-wide association study of rheumatoid arthritis. *Human Molecular Genetics* 19: 3863–3872.
69. Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, et al. (2009) Gene Ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics* 85: 13–24.
70. Luo L, Peng G, Zhu Y, Dong H, Amos CI, et al. (2010) Genome-wide gene and pathway analysis. *European Journal of Human Genetics* 18: 1045–1053.
71. O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, et al. (2011) Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Molecular Psychiatry* 16: 286–292.
72. Perry JRB, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, et al. (2009) Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes* 58: 1463–1467.
73. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, et al. (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. *American Journal of Human Genetics* 84: 399–405.
74. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE (2010) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *American Journal of Human Genetics* 86: 581–591.
75. Carbonetto P, Stephens M (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* 7: 73–108.
76. Guan Y, Stephens M (2011) Bayesian variable selection regression for genome-wide association studies, and other large-



- scale problems. *Annals of Applied Statistics* 5: 1780–1815.
77. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. *Annual Review of Genomics and Human Genetics* 10: 387–406.
  78. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* 11: 499–511.
  79. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* 3: e114.
  80. Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* 9: e1003264.
  81. Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association* 90: 773–795.
  82. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics* 10: 681–690.
  83. Fernando MMA, Stevens CR, Walsh EC, De Jager PL, Goyette P, et al. (2008) Defining the role of the MHC in autoimmunity: a review and pooled analysis. *PLoS Genetics* 4: e1000024.
  84. Polychronakos C, Li Q (2011) Understanding type 1 diabetes through genetics: advances and prospects. *Nature Reviews Genetics* 12: 781–792.
  85. Barton A, Thomson W, Ke X, Eyre S, Hinks A, et al. (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nature genetics* 40: 1156–1159.
  86. Cooper JD, Walker NM, Smyth DJ, Downes K, Healy BC, et al. (2009) Follow-up of 1715 SNPs from the Wellcome Trust Case Control Consortium genome-wide association study in type I diabetes families. *Genes and Immunity* 10: S85–S94.
  87. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, et al. (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genetics* 39: 830–832.
  88. Thomson W, Barton A, Ke X, Eyre S, Hinks A, et al. (2007) Rheumatoid arthritis association at 6q23. *Nature Genetics* 39: 1431–1433.
  89. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, et al. (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genetics* 39: 857–864.
  90. MHC Sequencing Consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401: 921–923.
  91. Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, et al. (2004) Gene map of the extended human MHC. *Nature Reviews Genetics* 5: 889–899.
  92. Imboden JB (2009) The immunopathogenesis of rheumatoid arthritis. *Annual Review of Pathology* 4: 417–434.
  93. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research* 37: D674–D679.
  94. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, et al. (2010) The NCBI BioSystems database. *Nucleic Acids Research* 38: D492–D496.
  95. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, et al. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research* 39: D685–D690.
  96. Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GSS, et al. (2010) NetPath: a public resource of curated signal transduction pathways. *Genome Biology* 11: R3.
  97. Berger JO, Sellke T (1987) Testing a point null hypothesis: the irreconcilability of P values and evidence. *Journal of the American Statistical Association* 82: 112–122.
  98. O'Shea JJ, Ma A, Lipsky P (2002) Cytokines and autoimmunity. *Nature Reviews Immunology* 2: 37–45.
  99. Godessart N, Kunkel SL (2001) Chemokines in autoimmune disease. *Current Opinion in Immunology* 13: 670–675.
  100. Zhernakova A, van Diemen CC, Wijmenga C (2009) Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nature Reviews Genetics* 10: 43–55.
  101. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, et al. (2007) A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nature Genetics* 39: 207–211.
  102. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, et al. (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genetics* 39: 596–604.
  103. Homer CR, Richmond AL, Rebert NA, Achkar J, McDonald C (2010) ATG16L1 and NOD2 interact in an autophagy-

- dependent antibacterial pathway implicated in Crohn's disease pathogenesis. *Gastroenterology* 139: 1630–1641.
104. Garg G, Tyler JR, Yang JHM, Cutler AJ, Downes K, et al. (2012) Type 1 diabetes-associated IL2RA variation lowers IL-2 signaling and contributes to diminished CD4+CD25+ regulatory T cell function. *Journal of Immunology* 188: 4644–4653.
  105. Sakaguchi S, Yamaguchi T, Nomura T, Ono M (2008) Regulatory T cells and immune tolerance. *Cell* 133: 775–787.
  106. Todd JA (2010) Etiology of type 1 diabetes. *Immunity* 32: 457–467.
  107. van Belle TL, Coppieters KT, von Herrath MG (2011) Type 1 diabetes: etiology, immunology, and therapeutic strategies. *Physiological Reviews* 91: 79–118.
  108. Chistiakov DA, Voronova NV, Chistiakov PA (2008) The crucial role of IL-2/IL-2RA-mediated immune regulation in the pathogenesis of type 1 diabetes, an evidence coming from genetic and animal model studies. *Immunology Letters* 118: 1–5.
  109. Hulme MA, Wasserfall CH, Atkinson MA, Brusko TM (2012) Central role for interleukin-2 in type 1 diabetes. *Diabetes* 61: 14–22.
  110. Dörig RE, Marciel A, Chopra A, Richardson CD (1993) The human CD46 molecule is a receptor for measles virus (Edmonston strain). *Cell* 75: 295–305.
  111. Naniche D (2009) Human immunology of measles virus infection. *Current Topics in Microbiology and Immunology* 330: 151–171.
  112. Tatsuo H, Ono N, Tanaka K, Yanagi Y (2000) SLAM (CDw150) is a cellular receptor for measles virus. *Nature* 406: 893–789.
  113. Rosenau BJ, Schur PH (2009) Association of measles virus with rheumatoid arthritis. *Journal of Rheumatology* 36: 893–897.
  114. McInnes IB, Schett G (2011) The pathogenesis of rheumatoid arthritis. *New England Journal of Medicine* 365: 2205–2219.
  115. Mehraein Y, Lennerz C, Ehlhardt S, Remberger K, Ojak A, et al. (2004) Virus antibodies in serum and synovial fluid of patients with rheumatoid arthritis and other connective tissue diseases. *Modern Pathology* 17: 781–789.
  116. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, et al. (2011) Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genetics* 7: e1002254.
  117. Tait KF, Marshall T, Berman J, Carr-Smith J, Rowe B, et al. (2004) Clustering of autoimmune disease in parents of siblings from the Type 1 diabetes Warren repository. *Diabetic Medicine* 21: 358–362.
  118. Bonizzi G, Karin M (2004) The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends in Immunology* 25: 280–288.
  119. Charo IF, Ransohoff RM (2006) The many roles of chemokines and chemokine receptors in inflammation. *New England Journal of Medicine* 354: 610–621.
  120. Dong C, Davis RJ, Flavell RA (2002) MAP kinases in the immune response. *Annual Review of Immunology* 20: 55–72.
  121. Pao LI, Badour K, Siminovitch KA, Neel BG (2007) Nonreceptor protein-tyrosine phosphatases in immune cell signaling. *Annual Review of Immunology* 25: 473–523.
  122. Ballard DH, Aporntewan C, Lee J, Lee J, Wu Z, et al. (2009) A pathway analysis applied to Genetic Analysis Workshop 16 genome-wide rheumatoid arthritis data. *BMC Proceedings* 3: S91.
  123. Beyene J, Hu P, Hamid JS, Parkhomenko E, Paterson AD, et al. (2009) Pathway-based analysis of a genome-wide case-control association study of rheumatoid arthritis. *BMC Proceedings* 3: S128.
  124. Peng G, Luo L, Siu H, Zhu Y, Hu P, et al. (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. *European Journal of Human Genetics* 18: 111–117.
  125. Wang L, Jia P, Wolfinger RD, Chen X, Grayson BL, et al. (2011) An efficient hierarchical generalized linear mixed model for pathway analysis of genome-wide association studies. *Bioinformatics* 27: 686–692.
  126. Silverberg MS, Duerr RH, Brant SR, Bromfield G, Datta LW, et al. (2007) Refined genomic localization and ethnic differences observed for the IBD5 association with Crohn's disease. *European Journal of Human Genetics* 15: 328–335.
  127. Van Limbergen J, Russell RK, Nimmo ER, Satsangi J (2007) The genetics of inflammatory bowel disease. *American Journal of Gastroenterology* 102: 2820–2831.
  128. Rioux JD, Goyette P, Vyse TJ, Hammarström L, Fernando MMA, et al. (2009) Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proceedings of the National Academy of Sciences* 106: 18680–

- 18685.
129. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491: 119–124.
  130. Eyre S, Bowes J, Diogo D, Lee A, Barton A, et al. (2012) High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics* 44: 1336–1340.
  131. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, et al. (2007) TRAF1-C5 as a risk locus for rheumatoid arthritis—a genomewide study. *New England Journal of Medicine* 357: 1199–1209.
  132. 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. *Nature Genetics* 42: 508–514.
  133. Barton A, Eyre S, Ke X, Hinks A, Bowes J, et al. (2009) Identification of AF4/fmr2 family, member 3 (AFF3) as a novel rheumatoid arthritis susceptibility locus and confirmation of two further pan-autoimmune susceptibility genes. *Human Molecular Genetics* 18: 2518–2522.
  134. Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, et al. (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nature Genetics* 40: 1216–1223.
  135. 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. *Nature Genetics* 41: 703–707.
  136. Cooper JD, Walker NM, Healy BC, Smyth DJ, Downes K, et al. (2009) Analysis of 55 autoimmune disease and type II diabetes loci: further confirmation of chromosomes 4q27, 12q13.2 and 12q24.13 as type I diabetes loci, and support for a new locus, 12q13.3-q14.1. *Genes and Immunity* 10: S95–S120.
  137. Huang J, Ellinghaus D, Franke A, Howie B, Li Y (2012) 1000 genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 data. *European Journal of Human Genetics* 20: 801–805.
  138. Lowe CE, Cooper JD, Brusko T, Walker NM, Smyth DJ, et al. (2007) Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature Genetics* 39: 1074–1082.
  139. Vella A, Cooper JD, Lowe CE, Walker N, Nutland S, et al. (2005) Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *American Journal of Human Genetics* 76: 773–779.
  140. Hakonarson H, Grant SFA, Bradfield JP, Marchand L, Kim CE, et al. (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* 448: 591–594.
  141. Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, et al. (2008) Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics* 40: 1399–1401.
  142. McVean GAT, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304: 581–584.
  143. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, et al. (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Research* 40: D918–D923.
  144. Alejandro EU, Kalynyak TB, Taghizadeh F, Gwiazda KS, Rawstron EK, et al. (2010) Acute insulin signaling in pancreatic beta-cells is mediated by multiple Raf-1 dependent pathways. *Endocrinology* 151: 502–512.
  145. Sumara G, Formentini I, Collins S, Sumara I, Windak R, et al. (2009) Regulation of PKD by the MAPK p38 $\delta$  in insulin secretion and glucose homeostasis. *Cell* 136: 235–248.
  146. Bottini N, Vang T, Cucca F, Mustelin T (2006) Role of PTPN22 in type 1 diabetes and other autoimmune diseases. *Seminars in Immunology* 18: 207–213.
  147. Bartok B, Firestein GS (2010) Fibroblast-like synoviocytes: key effector cells in rheumatoid arthritis. *Immunological Reviews* 233: 233–255.
  148. Murphy SH, Suzuki K, Downes M, Welch GL, De Jesus P, et al. (2011) Tumor suppressor protein (p)53 is a regulator of NF-kappaB repression by the glucocorticoid receptor. *Proceedings of the National Academy of Sciences* 108: 17117–17122.
  149. Ichimura A, Hirasawa A, Poulain-Godefroy O, Bonnefond A, Hara T, et al. (2012) Dysfunction of lipid sensor GPR120 leads to obesity in both mouse and human. *Nature* 483: 350–354.
  150. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316: 889–894.
  151. McCarthy MI (2010) Genomics, type 2 diabetes, and obesity. *New England Journal of Medicine* 363: 2339–2350.

152. Fisher SA, Tremelling M, Anderson CA, Gwilliam R, Bumpstead S, et al. (2008) Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nature Genetics* 40: 710–712.
153. Bottolo L, Richardson S (2010) Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis* 5: 583–618.
154. Fridley BL (2009) Bayesian variable and model selection methods for genetic association studies. *Genetic Epidemiology* 33: 27–37.
155. He Q, Lin D (2011) A variable selection method for genome-wide association studies. *Bioinformatics* 27: 1–8.
156. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics* 4: e1000130.
157. Hung RJ, Baragatti M, Thomas D, McKay J, Szeszenia-Dabrowska N, et al. (2007) Inherited predisposition of lung cancer: a hierarchical modeling approach to DNA repair and cell cycle control pathways. *Cancer Epidemiology, Biomarkers and Prevention* 16: 2736–2744.
158. Logsdon BA, Hoffman GE, Mezey JG (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* 11: 58.
159. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics* 44: 825–830.
160. Yi N, Xu S (2008) Bayesian Lasso for quantitative trait loci mapping. *Genetics* 179: 1045–1055.
161. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics* 25: 714–721.
162. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. *PLoS Genetics* 4: e1000279.
163. International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.
164. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, et al. (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* 39: D691–D697.
165. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* 38: D355–D360.
166. Caspi R, Altman T, Dale JM, Dreher K, et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 38: D473–D479.
167. Romero P, Wagg J, Green M, Kaiser D, Krummenacker M, et al. (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology* 6: R2.
168. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, et al. (2008) WikiPathways: pathway editing for the people. *PLoS Biology* 6: e184.
169. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, et al. (2012) WikiPathways: building research communities on biological pathways. *Nucleic Acids Research* 40: D1301–D1307.
170. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, et al. (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research* 38: D204–D210.
171. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, et al. (2010) The BioPAX community standard for pathway data sharing. *Nature Biotechnology* 28: 935–942.
172. Soh D, Dong D, Guo Y, Wong L (2010) Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinformatics* 11: 449.
173. Stobbe M, Houten S, Jansen G, van Kampen A, Moerland P (2011) Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology* 5: 165.
174. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* 10: 184–194.
175. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, et al. (2007) A genome-wide association study of global gene expression. *Nature Genetics* 39: 1202–1207.
176. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. *Nature Genetics* 39: 1217–1224.
177. George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88: 881–889.

178. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genetics* 40: 695–701.
179. Ioannidis JPA, Trikalinos TA, Khoury MJ (2006) Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. *American Journal of Epidemiology* 164: 609–614.
180. O'Brien SM, Dunson DB (2004) Bayesian multivariate logistic regression. *Biometrics* 60: 739–746.
181. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. *Statistical Science* 14: 382–401.
182. Mitchell TJ, Beauchamp JJ (1988) Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 83: 1023–1032.
183. Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* 31: 2013–2035.
184. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100: 9440–9445.
185. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, et al. (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genetics* 40: 638–645.
186. Pirinen M, Donnelly P, Spencer CCA (2012) Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* 44: 848–851.