GENOME-WIDE ASSOCIATION STUDIES

Analysing biological pathways in genome-wide association studies

Kai Wang‡, Mingyao Li§ and Hakon Hakonarson*||*

Abstract | Genome-wide association (GWA) studies have typically focused on the analysis of single markers, which often lacks the power to uncover the relatively small effect sizes conferred by most genetic variants. Recently, pathway-based approaches have been developed, which use prior biological knowledge on gene function to facilitate more powerful analysis of GWA study data sets. These approaches typically examine whether a group of related genes in the same functional pathway are jointly associated with a trait of interest. Here we review the development of pathway-based approaches for GWA studies, discuss their practical use and caveats, and suggest that pathway-based approaches may also be useful for future GWA studies with sequencing data.

Genome-wide association (GWA) studies have been very successful for identifying disease loci using single-marker-based association tests that examine the relationships between each SNP marker and the trait of interest[1]. Despite the success of single-marker association tests — given the hundreds of thousands of SNP markers used in most GWA studies — this strategy has limited power to identify disease genes, similar to finding needles in a haystack. Some genes may be genuinely associated with disease status but may not reach a stringent genome-wide significance threshold in any GWA study. Realizing the limitations of conventional single-marker association analysis, alternative or complementary approaches for GWA study analysis have been developed in recent years. These include association tests that use multiple SNP markers[2–8], association tests using imputed genotypes[9,10], association tests incorporating linkage information[11] and, more recently, pathway-based association approaches[12]. The pathway-based approaches typically examine whether test statistics for a group of related genes have consistent yet moderate deviation from chance, similar to finding a string of interconnected needles in a haystack. It is well known that genes do not work in isolation; instead, complex molecular networks and cellular pathways are often involved in disease susceptibility and disease progression[13]. Therefore, by taking into account prior biological knowledge about genes and pathways, we may have a better chance to identify the genes and mechanisms that are involved in disease pathogenesis.

Much of the work on pathway-based analysis of GWA study data was motivated by pathway association approaches for gene expression microarray analysis. It is well known that functionally related genes can have coordinated gene expression patterns, so examination of gene expression for a group of genes can identify pathways that have modest yet consistent changes in gene expression levels[14]. This Gene Set Enrichment Analysis (GSEA) method has been improved[15,16] and, over the past decades, dozens of alternative approaches have been proposed that have different functionalities and power levels (reviewed and compared in REFS 17,18).

Borrowing ideas from the microarray field, similar approaches can be adopted in GWA study analysis, with some modifications to address the unique challenges of GWA study data. In pathway-based association tests for GWA studies, researchers typically examine a collection of predefined gene sets for pathways based on prior biological knowledge, and the significance of each pathway can be summarized based on the disease association of markers in or near genes that are components of that pathway. Over a few years, various techniques have been proposed to summarize the significance of a biological pathway from a collection of SNPs and to adjust for multiple testing at the pathway level. Here, we review the application of pathway-based association approaches, describe and classify currently available statistical methods, discuss the pitfalls and caveats in interpreting results from them, and propose future directions and extensions, especially with respect to next-generation sequencing data.

*Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA.
‡Zilkha Neurogenetic Institute, Department of Psychiatry, Department of Preventive Medicine, University of Southern California, California 90089, USA.
§Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA.
||Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA.
Correspondence to H.H.
e-mail: hakonarson@email.chop.edu
doi:10.1038/nrg2884

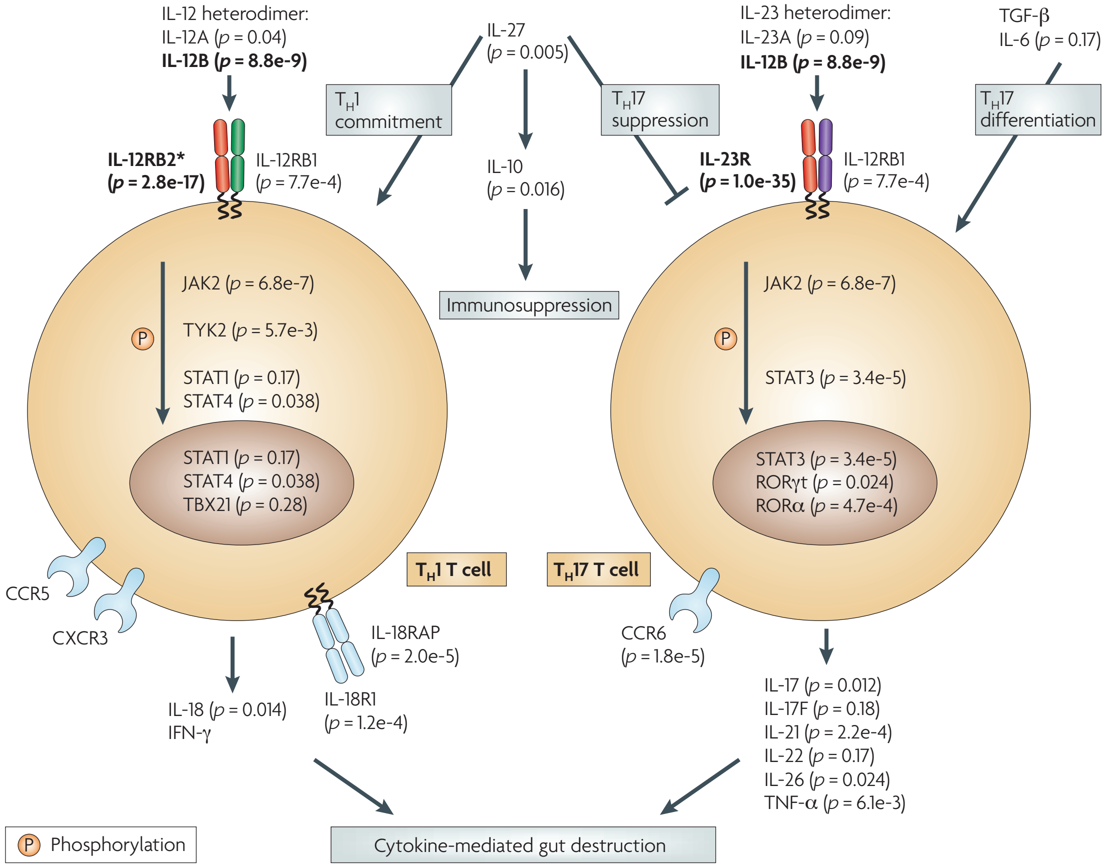

Figure 1 | **Linking pathways to disease: Crohn's disease.** As an example of how biological pathways are involved in disease pathogenesis, we illustrate a manually compiled pathway centred on interleukin (IL)-12 and IL-23 (REFS 19–22) that is important in Crohn's disease. The IL-12 and IL-23 cytokines share one subunit, their cellular receptors also share one subunit and their intracellular signalling machineries share many components. For many years, the pro-inflammatory cytokine IL-12 was thought to be a major player in Crohn's disease pathogenesis[109], which is mediated by T cells that produce T helper 1 cytokines (T$_H$1 cells). Recent genetic studies demonstrated a more important role for IL-23 (REF. 110), which activates a subset of T cells characterized by the production of the cytokine IL-17 (T$_H$17 cells)[19]. Only the main proteins in this pathway are shown. For each gene, the most significant p-value among SNPs closest to the gene (based on a large-scale meta-analysis of genome-wide association (GWA) studies on Crohn's disease[23]) was annotated. Only three genes at two loci (*IL12B* on 5q33 and *IL23R–IL12RB2* on 1p31) showed genome-wide significant signals in this GWA study (marked in bold font), but three genes (*JAK2*, C-C chemokine receptor 6 (*CCR6*) and signal transducer and activator of transcription 3 (*STAT3*)) in the IL-12–IL-23 pathway were confirmed as susceptibility genes in replication studies[23], and six genes (*STAT4*, *IL18*, IL-18 receptor accessory protein (*IL18RAP*), tyrosine kinase 2 (*TYK2*), *IL27* and *IL10*) in this pathway were reported as Crohn's disease susceptibility genes in other association and functional studies[24–29]. **IL12RB2* is located adjacent to *IL23R* so the significant marker could be tagging causal variants that target *IL23R*. CXCR3, C-X-C chemokine receptor 3; RORα, RAR-related orphan receptor A; TGF-β, transforming growth factor-β; TNF-α, tumour necrosis factor-α.

Applying pathway-based association approaches

As an example of how biological pathways are involved in disease pathogenesis, we discuss a manually compiled pathway centred on interleukin (IL)-12 and IL-23 (REFS 19–22), which are important in Crohn's disease (FIG. 1). Crohn's disease is an inflammatory disease of the gastrointestinal tract with a strong genetic component. We examined a previously published GWA study of Crohn's disease[23], mapped all assayed SNPs to their closest genes and then annotated the most significant p-values for the genes shown in FIG. 1. Only three genes at two loci showed genome-wide significant signals in this GWA study, but three genes in the IL-12–IL-23 pathway were confirmed as susceptibility genes in replication studies[23], and six genes in this pathway were reported as Crohn's disease susceptibility genes in other association and functional studies[24–29] (see FIG. 1 for details). This example clearly demonstrates that multiple related genes in the same functional pathway may work together to confer disease susceptibility and that some genes may not reach genome-wide significance in any given GWA study owing to limited power. Additionally, as the most associated gene in a pathway might not be the best candidate for therapeutic intervention, targeting susceptibility pathways might also have clinical implications for finding additional drug targets.

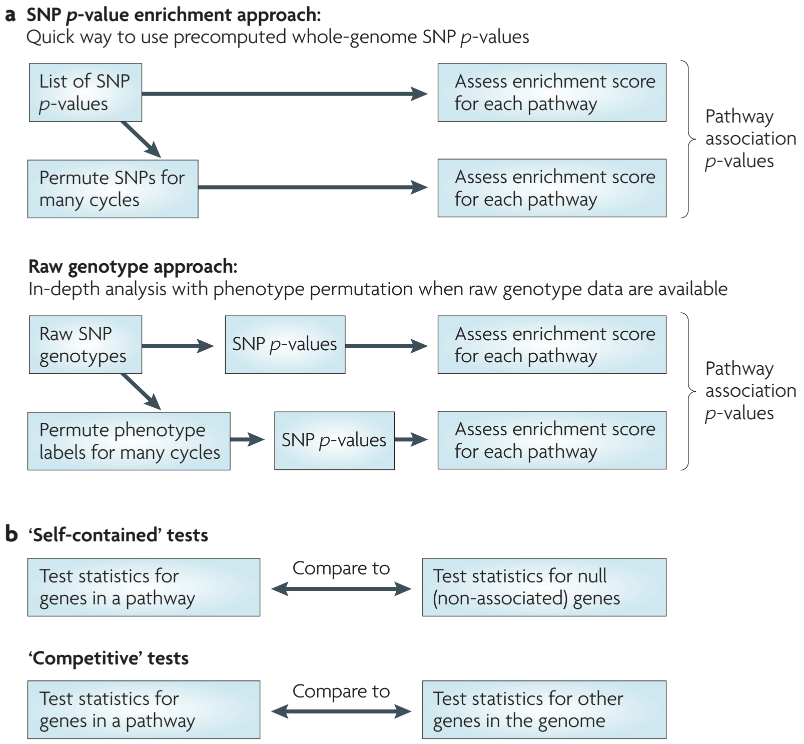

Figure 2 | **Types of pathway association method.** A summary of the two main types of pathway association approaches. The approaches can be categorized based on the input data (**a**) or the null hypotheses that are tested (**b**).

Recognizing the importance of considering SNPs in the same pathway jointly, several pathway association approaches have been developed and applied. For example, previous GWA studies demonstrated that complement factor H is a strong risk factor for age-related macular degeneration (AMD)[30–32]. However, in a pathway-based association analysis, multiple additional complement factor genes were moderately associated with AMD, which suggests that the complement pathway contributes to AMD pathogenesis[33]. Similar observations were made in other studies that examined the collective association of multiple complement factor pathway genes with AMD[12,34]. Two age-related neurological disorders, Parkinson's disease and amyotrophic lateral sclerosis, have been associated with the axon guidance pathway, although no individual SNPs in genes in this pathway reach a genome-wide significance level[35,36]. Several pathway-based studies have been conducted for neuropsychiatric disorders: for example, two studies implicated neuronal cell adhesion molecules in the aetiology of autism, schizophrenia and bipolar disorder[37,38]; and two studies on bipolar disorder implicated pathways that mediate ion channel activity, synaptic neurotransmission, and pathways that modulate transcription and cellular activity[39,40]. Furthermore, pathway-based association methods have been tested on a few autoimmune diseases, and inflammatory pathways that are well known in the pathogenesis of these diseases have been identified confidently[41,42]. Similarly, several groups at the Genetic Analysis Workshop 16 (please see Further information for a link to the website) analysed the same GWA data set on rheumatoid arthritis through various pathway-based approaches[43–47] and identified pathways known to be related to this disease. Pathway-based analysis of the Wellcome Trust Case Control Consortium (WTCCC) data sets implicated multiple pathways in disease predisposition, many of which had long been assumed to contain polymorphic genes that lead to disease risk[48]. Studies published in 2010 have also expanded the repertoire of diseases on which diverse pathway-based association approaches have been tested[37,40,49–55].

Collectively, these studies show that pathway-based approaches can provide complementary information to conventional single-marker analysis in GWA studies. Specifically, by identifying additional susceptibility genes, pathway analysis can be used to fill in part of the 'missing heritability'. In addition, pathway analysis may guide mechanistic studies, as they can help uncover the underlying disease pathways without the need to narrow down each GWA study locus to a single gene.

Methods for pathway-based association

Over recent years, dozens of different methods have been published for pathway-based association analysis[12,43,44,47,51,56–66] and some of the related issues have been discussed and reviewed[67–72]. These statistical methods can be broadly classified into two types, based on whether the required input data sets are a collection of SNP p-values or individual-level SNP genotypes (FIG. 2a). Additionally, the null hypothesis being tested in these pathway association approaches can be broadly classified as 'self-contained' versus 'competitive', based on whether comparisons were made between genes in a specific pathway and non-associated genes or other genes in the genome (FIG. 2b). Some of these published algorithms are available as software implementations or web servers (TABLE 1).

Definition and source of pathways. The term 'pathway' as used in published GWA studies typically refers to a set of related genes, rather than a diagrammatic pathway in which some genes are connected by arrows. Most pathway-based association approaches require the users to specify a predefined set of gene sets or pathways to test. The Pathguide resource provides many links to manually curated or computationally predicted pathways; some of the more commonly used pathway collections include Kyoto Encyclopedia of Genes and Genomes (KEGG)[73], BioCarta and Gene Ontology[74] (please see Further information for links to these databases). Both KEGG and BioCarta contain manually curated pathways in different biological processes, whereas Gene Ontology contains mostly electronic annotations for human genes, based on various sources of evidence such as sequence homology. The combined use of manually curated pathways and electronically compiled pathways may be employed in data analysis to ensure comprehensive coverage of pathways as well as high-quality information for well-studied pathways.

Table 1 | **Publicly available web servers or computer software for pathway analysis on genome-wide association study data sets**

Name	Input data	Hypothesis tested	Analysis strategy	URL	Ref.
ALIGATOR	SNP *p*-values	Competitive	Define significant SNPs by prespecified *p*-value cut-off, then count significant genes in each pathway	http://x004.psycm.uwcm.ac.uk/~peter	40
i-GSEA4GWAS	SNP *p*-values	Competitive	SNP label permutation, assign SNPs to genes, calculate modified Gene Set Enrichment Analysis (GSEA) enrichment score	http://gsea4gwas.psych.ac.cn	62
GenGen	Raw genotype	Competitive	Assign the best test statistic among SNPs in or near a gene to represent the gene level signal, then calculate Kolmogorov–Smirnov-like enrichment score for a pathway	http://www.openbioinformatics.org/gengen	12
GESBAP	SNP *p*-values	Competitive	Calculate enrichment score using ranked gene list, assign the best SNP *p*-value to a gene	http://bioinfo.cipf.es/gesbap	111
GRASS	Raw genotype	Self-contained	Regularized regression to select representative eigenSNPs for each gene, then assess their joint association with disease risk	http://linchen.fhcrc.org/grass.html	49
GSA-SNP	SNP *p*-values	Competitive	Use kth ($k = 1, 2, 3, 4$ or 5) best *p*-values in each gene to represent the gene	http://gsa.muldas.org	64
GSEA-SNP	Raw genotypes	Competitive	Calculate enrichment score based on all SNPs in a given pathway without calculating gene-level test statistics	http://www.nr.no/pages/samba/area_emr_smbi_gseasnp	112
PLINK set-test	Raw genotypes	Self-contained	Calculate the average of test statistics as the pathway enrichment scores, using independent and significant (by preselected *p*-value cut-off) SNPs in the pathway	http://pngu.mgh.harvard.edu/~purcell/plink	113
SNP ratio test	Raw genotypes	Competitive	Calculate the number of significant SNPs in pathway divided by the number of SNPs in pathway	http://sourceforge.net/projects/snpratiotest	58

Additionally, several databases also provide gene co-expression patterns or protein–protein interactions that can be explored in pathway analysis; for example, the Molecular Signatures Database (MSigDB) provides a list of gene sets that are defined by gene expression 'neighbourhoods' near cancer genes[15]. Several commercial providers, such as Ingenuity Pathway Analysis and GeneGo, also provide proprietary pathway databases, using multiple sources of information, including literature reviews as well as experimental evidence. More specialized pathway databases also exist to curate specific types of pathways; for example, the Science Signal Transduction Knowledge Environment and Nature Pathway Interaction Database are both manually curated databases for cell signalling, the MetaCyc is a high-quality database for metabolic pathways, the TRANSPATH is a database for transcriptional regulation and there are several databases compiled from protein–protein interaction information[75] (please see Further information for links to these databases). It should be noted that a fraction of human genes is uncharacterized and these genes are not mapped to manually curated or computationally predicted pathways, so their effects cannot be accounted for in pathway association analysis. In addition, some well-known pathways are not yet described in sufficient detail in public databases (for example, β-cell function and the IL-12–IL-23 pathway). For some specific disease areas, experts can compile more up-to-date pathways based on literature information or a prior biological hypothesis and these pathways can be tested in association analysis; for example, the complement factor pathway can be compiled and tested, given the known association of complement factor H with AMD. When neuropsychiatric diseases are studied, it may be important to manually compile candidate pathways using expert knowledge, as public databases may not have well-annotated pathways for neuronal function.

Input data for pathway association tests. The first type of pathway association approach, the '*p*-value enrichment approach', aims to determine whether a specific group of *p*-values for SNPs (or genes) is enriched for association signals. A practical advantage of this type of approach is that it only requires a list of *p*-values as input, without the need for individual-level genotypes, which eliminates many practical challenges in coordinated data analysis and data sharing. Many *p*-value-based approaches use a *p*-value cut-off (typically $p<0.05$ or $p<0.001$) for identifying a subset of significant SNPs in single-marker association tests for further pathway analysis. This means that the results are partly dependent on the user-specified cut-off. There are also several potential biases such as gene size that must be considered when using these methods (discussed further below). Nevertheless, given their practical advantages, *p*-value enrichment approaches have gained popularity in the analysis of GWA study data sets.

The second type of pathway association approach, the 'raw genotype approach', uses individual-level SNP genotypes to derive gene-level and pathway-level test statistics and usually requires phenotype permutations to adjust for statistical significance of identified pathways. The raw genotype data may be used differently in the various methods: some methods require sophisticated multi-marker tests to derive gene-level test statistics that require raw genotypes for all SNPs in the gene; some

Permutation
A strategy for assessing the probability of observing the value of a particular statistic. The probability is computed from a data set in which the data are randomly shuffled and the statistic is recomputed from the shuffled data many times and ultimately compared to the value of the statistic obtained with the non-shuffled data.

Multi-marker test
A statistical method that measures the strength of association between a trait and multiple SNP markers.

Box 1 | Technical differences among pathway-based association methods

Mapping SNPs to genes

Commonly used pathway-based association methods only examine genic SNPs (discarding all SNPs outside genes) or map SNPs to their closest genes in the genome within a certain distance threshold, such as 10 kb, 100 kb or 500 kb. Future studies may consider using recombination peaks at each locus to assign SNPs to one or several genes.

Pruning SNPs

It is important to bear in mind that different SNP arrays may have used different SNP ascertainment schemes and that some genomic regions contain more correlated SNPs than other regions. Therefore, as the coverage of SNP arrays may be uneven, identifying independent SNPs may be necessary to reduce the biases caused by different SNP density and coverage. For example, researchers may prune a set of independent SNPs based on their r^2 threshold before performing pathway-based association tests (see also BOX 2).

Calculation of gene-based test statistics

Many methods depend on having a summary test statistic for each gene, based on their representative SNPs. Commonly used methods include the minP approach that assigns the minimum *p*-value from the SNPs in or close to a gene as the *p*-value for that gene and multi-marker tests that summarize the test statistic for a gene based on many contributing markers. Some methods pool all SNPs in a pathway together without calculating test statistics for specific genes.

Calculation of pathway enrichment statistics

Commonly used methods include summary statistics that examine the shape of distributions of test statistics and hypergeometric tests that examine the categorical enrichment of test statistics.

Adjustment for gene size

Permutation approaches are typically used for adjustment of gene size, such that larger genes (or genes with more SNPs) are not more likely to generate lower *p*-values just by random chance.

Adjustment for pathway size

Depending on the enrichment score statistics used, the pathway size (number of genes in the pathway) may bias the resulting test statistics for the pathway. Most methods focus the analysis on pathways that pass specific size thresholds, such as those with at least 10 or 20 genes. Some methods resample pathways of a certain size many times and compute the distribution of observed enrichment scores to adjust for pathway size biases.

---

methods depend on single-marker *p*-values but require raw genotype data for phenotype permutation to adjust the significance of pathway enrichment scores. For example, we have developed a pathway-based association approach[12] by adopting the general framework of GSEA[16] strategies, but accounting for linkage disequilibrium (LD) among SNP markers and for the different sizes of genes and pathways, using a two-step correction procedure. The ability to perform phenotype permutation maintains the correct LD patterns among neighbouring SNPs — this is an important difference to *p*-value enrichment approaches, which may be influenced by differing LD patterns. However, it should be noted that it is not always possible to obtain raw genotype data easily and the permutation procedures are computationally intensive.

Although we broadly classify pathway-based association tests in two categories, the published methods also differ in many other aspects, including the source of precompiled pathway collections, the distance threshold to assign SNPs to nearby genes, how they summarize gene-level test statistics, how to calculate enrichment scores for each pathway and how statistical significance is assessed (BOX 1). As these differences are addressed in each published study when comparing the approach used with other approaches, we do not elaborate on them in greater detail in this review.

Null hypothesis being tested. Two types of null hypothesis can be tested in pathway analysis of genotype data and are referred to as competitive and self-contained hypotheses[76]. Competitive tests compare the statistics for genes in a given pathway with statistics for other genes to determine whether genes in a particular pathway tend to be more associated with a given phenotype. Self-contained methods only consider results in a pathway of interest and compare to the null (non-associated) genomic background. Because competitive methods require a comparison between many different pathways, these tests cannot be applied in a study that only assessed one or a few candidate pathways. By contrast, self-contained tests have the advantage that only genotypes from a collection of candidate genes are required, so this type of test can be used in many candidate gene association studies or in GWA studies that use gene-centric arrays (such as those specifically designed for cardiovascular diseases[77], metabolic traits or autoimmune conditions). However, a disadvantage of self-contained tests is that the genomic inflation of test statistics is often not monitored or adequately adjusted for, which may result in inflated type I error. Given that most GWA studies are indeed susceptible to some degrees of genomic inflation, this is a practical concern. Further comparisons of these two forms of null hypotheses in the setting of gene expression analysis have been made[76,78], and these comparisons may be relevant for GWA study analysis as well.

Other approaches using pathway information. Besides formal pathway-based association tests, investigators can also empirically identify susceptibility pathways based on a small group of SNPs that pass a genome-wide significance threshold and then examine other

SNP ascertainment
Identification of SNPs that should be placed on a genotyping array to ensure representative coverage of the genome.

Linkage disequilibrium
The non-random association of alleles at two or more closely linked loci.

Genomic inflation
The presence of excess false-positive results, measured by quantifying the ratio of the median of the empirically observed distribution of the test statistic to the expected median.

Type I error
The probability of a false-positive result from a statistical hypothesis test.

genes in these pathways that did not reach genome-wide significance. This approach is somewhat ad hoc, but for well-powered studies with large sample sizes and multiple genome-wide significant association signals, it can readily generate important biological insights based on bona fide association signals. For example, in a large-scale GWA study on body mass index with over 32,000 subjects in discovery cohorts and over 59,000 subjects in replication cohorts, multiple genome-wide significant SNPs tagged genes that are highly expressed or known to function in the central nervous system (CNS), implicating the CNS in predisposition to obesity[79]. Similarly, in a GWA study on adult height, SNPs near 20 genes were implicated with $p<5\times10^{-7}$. Interestingly, the Hedgehog signalling, extracellular matrix and cancer pathways each contain several candidate genes, suggesting that these pathways may be involved in human growth and developmental processes[80]. In a more recent GWA study on adult height, SNPs at 180 loci achieved genome-wide significance and these loci tended to be functionally connected with each other[81]. Several SNPs near genes in these pathways narrowly missed genome-wide significance but a pathway analysis implicated multiple previously reported and novel pathways, which suggested that these pathways are likely to contain additional associated variants.

Hierarchical modelling is another form of pathway association analysis. This method takes a subset of markers from the first stage of a genome-wide association scan and carries them forward to subsequent stages for testing on an independent set of subjects[82]. Rather than simply selecting a subset of the most significant markers that reach an arbitrary p-value cut-off, a prior model is used that treats each marker differently. The prior model can be based on functions of various covariates that characterize each marker, such as prior linkage, association or functional pathway data. Therefore, through hierarchical approaches, it is possible to infer whether specific pathways are associated with a phenotype of interest. In fact, hierarchical models using LASSO have been successfully applied in simultaneous multivariate analyses of all GWA study SNPs[83,84].

Challenges and considerations

Although pathway-based approaches are becoming an invaluable tool to enable powerful association tests and help formulate new hypotheses on disease susceptibility, many challenges limit their practical use.

Major-effect genes versus moderate-effect pathways. Complex diseases can have different genetic architectures that need to be taken into account in pathway analysis. Some complex diseases or traits are likely to result from the interplay of hundreds of genes in multiple pathways, and each pathway could contain several susceptibility genes that are moderately associated with disease. Pathway-based analyses are likely to be informative in such cases. However, for other traits, one strongly associated gene in a pathway can invalidate the null hypothesis and show significance at the pathway level, but such pathways are likely to be of less interest to researchers because the pathway as a whole might not influence the trait. This issue will affect both competitive tests and self-contained tests, but approaches that use SNP test statistics to calculate enrichment scores will be more susceptible to these biases. In some analyses, removing known susceptibility genes from candidate pathways and reassessing association statistics may be helpful; for example, a study removed transcription factor 7-like 2 (*TCF7L2*, a gene well known to be strongly associated with type 2 diabetes) from the Wnt signalling pathway and re-evaluated the association between the modified pathway with type 2 diabetes (the p-value dropped from 0.0007 to 0.002)[85]. Also, when major histocompatibilty complex (MHC)-linked autoimmune diseases are studied, it may be important to adjust for influences from the MHC region given its extensive LD patterns, possibly by removing all MHC genes from the data set.

Biases introduced by permutation procedures. Most pathway association approaches use permutation procedures to calculate empirically adjusted p-values for given pathways. For p-value-based approaches, permutation of SNPs is typically performed but, as discussed previously, this procedure disrupts LD patterns between SNPs and may not generate the correct null distribution. For raw genotype-based approaches, permutation of phenotypes (binary traits or quantitative traits) is typically performed but some biases can still be introduced. The reason for this is that, when phenotype permutation is performed, the 'background' distribution reflects the situation in which none of the SNPs or genes associate with the phenotype of interest; but in practice, for any trait, a proportion of SNPs will be genuinely associated with disease in unpermuted data sets. Furthermore, regardless of whether the SNPs or phenotypes are being permuted, the sampling units are assumed to be independent and identically distributed, which may not be the case, as gene–gene interactions may play an important part in disease susceptibility and study participants might be distantly related.

Adjustment of covariates. Many GWA studies use adjustment of covariates in association tests, for example, adjusting for population stratification, age, sex or environmental risk factors. These adjustments are typically necessary for refining genotype–phenotype relationships, controlling for confounding factors and reducing genomic inflation. For p-value-based pathway association approaches, it is straightforward to directly use adjusted single-marker p-values in the analysis. However, for raw genotype-based approaches, it could present some challenges. Some pathway association approaches that use chi-squared test statistics cannot handle covariates and, even if they can incorporate covariates by using a regression model to calculate test statistics, a strong assumption needs to be made about independence between pathways and specific covariates.

Testing for multiple non-independent pathways. Public pathway collections contain many pathways, including overlapping pathways that share some genes, so adjustment to correct for multiple testing needs to be considered. Some authors first perform a pairwise comparison between all pathways and then select a smaller subset of the most representative pathways for subsequent association analysis. For pathway collections with hierarchical structures (such as Gene Ontology), some authors restrict the analysis to a specific level of the hierarchy, such that pathways below this level are incorporated with their parental pathways in this level. Nevertheless, owing to the non-independent nature of many pathways, stringent Bonferroni corrections of p-values for each pathway will be over-conservative; False Discovery Rate (FDR)-based approaches may be more attractive to summarize the significance of associated pathways.

Discordant results from different approaches. In several cases, different results have been obtained when different pathway-based methods were used to analyse data on the same diseases, and sometimes even when the same GWA study data sets were analysed. For example, the bipolar disorder data set from WTCCC was analysed by four groups using different methods and they found different significant pathways[39,40,48,51] (TABLE 2). This could be due to several reasons, including the use of different pathway collections and the different properties of statistical tests on a disease architecture with no major-effect genes. By contrast, the Crohn's disease data set has also been tested by a few different methods and the interleukin- or immune-related pathways were consistently detected as significant, although the top pathways do not match exactly[40,41,48,51,55] (TABLE 2). These examples show that users should be cautious when drawing conclusions from only one pathway analysis.

The need to replicate pathway association findings. An issue that has not been stressed enough in many published pathway-based association studies is the need to replicate association results. Similar to single-marker-based association tests, pathway-based association strategies may also be susceptible to false-positive results and thus should be appropriately replicated in independent data sets. Given the unique property of pathway-based association approaches, replication studies can be flexibly conducted on GWA data from different genotyping platforms or on GWA studies from different ethnic groups (for example, REF. 41). It is likely that the ranking of individual genes or SNPs in a pathway may differ between studies but genes in genuinely associated pathways are expected to be consistently associated in replication studies.

The selection of pathway association approaches. There is no clear answer as to how the various pathway analysis methods perform against each other under different scenarios of disease architecture and sample size, and in their relative susceptibility to biases such as gene or pathway size. Nevertheless, a few recent studies have compared different approaches and these results may be used as a reference for users. For example, in one study[49], the authors demonstrated that their method performs favourably against competing approaches on simulation data sets, especially when the disease-associated variants tend to reside in large genes. In terms of selecting methods for pathway analysis, our general recommendations are to consider the data sets available and the disease architecture to be tested. When individual-level genotype data is accessible, we suggest that it is preferable to use the 'raw genotype' approaches because they are less susceptible to biases inherent in p-value-based approaches. For GWA studies on SNP panels with candidate genes, self-contained tests need to be used because some genes in any given pathway may not be interrogated in the study. Perhaps more importantly, given the distinct statistical models and analytical procedures for different approaches, it may be desirable to examine whether consistently associated pathways can be identified by more than one approach and then follow up the findings with an independent replication data set.

We caution users that pathway analysis for GWA studies is still not well developed and that results should be interpreted with care and scrutiny but we do not wish to discourage users. Indeed, the goal of pathway-based approaches is not to replace conventional single-marker analysis but to play a complementary part in identifying novel genes or sets of genes that confer disease susceptibility. The results from pathway association approaches may also lead to the formulation of new hypotheses for additional statistical validations and functional validations.

Future directions and extensions

Pathway approaches for expression microarray data were first proposed almost a decade ago[14,86] and yet novel and improved analytical methods are still being developed and reported[87–93]. As the concept of pathway analysis for GWA studies was proposed only recently[12], we expect that many more improvements will be developed in the near future to help researchers better take advantage of the large amounts of available GWA study data. In BOX 2, we summarize some of the main areas for which we think improvements are needed. There are also several potential opportunities to extend pathway analysis as genomic technologies develop, which we discuss below.

Integrative genomics that incorporates pathway information. For genetic studies, multiple data types on the same set of samples can be collected and can provide a systems view of the biological processes underlying disease susceptibility or progression. In addition to SNP genotypes, these types of data include copy number variants, gene expression, epigenetic modifications and somatic mutations, among many others. Several studies have correlated whole-genome gene expression data sets with genotype data to perform integrative genomics analysis and these analyses have led to better understanding of genetic association signals[94,95]. It is conceivable that most genome-wide data sets can be subjected to pathway-based analytical framework individually and it is also possible that one type of data can generate prior information to be tested by pathway approaches in other

Bonferroni correction
A multiple comparison adjustment approach that tests each individual hypothesis by dropping the threshold for declaring statistical significance by n-fold, when n hypotheses are being tested.

False Discovery Rate
A multiple comparison adjustment approach to control the expected proportion of incorrectly rejected null hypotheses in a list of rejected hypotheses.

Table 2 | **Top three pathways for Crohn's disease and bipolar disorder in WTCCC data sets**

Pathway	*p*-value	Statistical test	Ref.
Crohn's disease			
Antigen processing and presentation of peptide or polysaccharide antigen through MHC class II	<0.0001		40
MHC class II receptor activity	<0.0001		
MHC class II protein complex	<0.0001		
IL-12- and STAT4-dependent signalling pathway	0.00008		41
T cell receptor signalling pathway (in BioCarta)	0.0003		
T cell receptor signalling pathway (in KEGG)	0.0007		
Signal transduction — calcium signalling	0.008		48
Transcription — ChREBP regulation pathway	0.01		
Immune response — IL-3 activation and signalling pathway	0.02		
ABC transporters — general	0.0004323	Fisher's exact*	51
Extracellular matrix–receptor interaction	0.00051323		
Lck and Fyn tyrosine kinases in initiation of T cell receptor activation pathway	0.00039108		
Cytokine–cytokine receptor interaction	7.8144e–14	Simes/FDR*	51
Neuroactive ligand–receptor interaction	2.078e–05		
JAK–STAT signalling pathway	4.1916e–14		
IL-9 signalling	N.A.		55[‡]
IL-2 receptor β-chain in T cell activation	N.A.		
Bipolar disorder			
Ion channel activity	1.27e–05		39
Calcium ion binding	6.58e–05		
Potassium channel activity	0.00122		
Hormone activity	<0.0001		40
Transcription factor activity	<0.0001		
Macroautophagy	<0.0001		
Heparan sulphate and heparin metabolism	0.01		48
Cytoskeleton remodelling — α-1A adrenergic receptor-dependent inhibition of PI3K	0.01		
Niacin–HDL metabolism	0.03		
Inositol metabolism	<1e–20	Fisher's exact*	51
HOP pathway in cardiac development pathway	0.00229732		
Glycan structures — biosynthesis 1	0.000266725		
Chondroitin sulphate biosynthesis	<1e–20	Simes and FDR*	51
Glycan structures — biosynthesis 1	<1e–20		
MAPK signalling pathway	5.02134e–05		

*Reference 51 explored different statistical tests, so the testing approach is annotated for each set of results. [‡]Reference 55 listed two significant pathways in common to multiple data sets, but *p*-values were not given. ChREBP, carbohydrate response element-binding protein; FDR, False Discovery Rate; HDL, high-density lipoprotein; HOP, homeodomain only protein; IL, interleukin; KEGG, Kyoto Encyclopedia of Genes and Genomes; MHC, major histocompatibility complex; STAT, signal transducer and activator of transcription; WTCCC, Wellcome Trust Case Control Consortium.

data sets. However, the more integrated approach of using multiple data types together in the same pathway analysis may be more powerful to reveal novel biological insights. How to best correlate whole-genome SNP genotype data with many other types of data through pathway-based approaches remains a crucial issue that needs to be explored further.

Extension to high-throughput sequencing data. As improvements in high-throughput sequencing techniques enable sequencing data to be produced at ever-more rapid rates, it will become feasible to perform sequencing-based GWA studies (which here we term Seq-GWA studies) in large-scale genetic studies. Although it may seem straightforward to use genotypes

Box 2 | **Improvements for pathway-based analysis methods**

- Better summary statistics need to be developed to assess the strength of association at the gene level or at the pathway level. Many published methods use the minimum *p*-value in a gene as the representative *p*-value for that gene in pathway association tests. Despite its convenience, this approach results in loss of information and may be susceptible to genotyping errors. Some recently published pathway association studies have demonstrated that joint association tests on multiple SNPs provide better power than methods that use a collection of *p*-values for individual SNPs[49]. Shrinkage techniques such as LASSO may be needed to address the issue of many independent variables in a regression model. Alternatively, machine-learning techniques that compute discrimination between two phenotype groups may be used to model complex genetic relationships better than association[100].
- Multi-marker association tests may be adopted for pathway-based association tests with specific modifications. These tests can be used to summarize gene-level *p*-values but it is also possible that these tests may be directly used for pathway-level association by appropriately adjusting for the correlation between adjacent markers and jointly considering the significance of all markers for all genes in a pathway. A potential drawback would be the increased degree of freedom because, for modest sample sizes, many more markers are included in the regression model. Additionally, multi-marker tests are typically computationally intensive so many of them may not be easily extended to the whole genome, especially for raw genotype-based approaches.
- For SNP *p*-value enrichment approaches, it would be beneficial to develop methods that minimize biases caused by differential linkage disequilibrium (LD) patterns between different loci. Currently, some methods use SNP permutations that disrupt LD patterns; they assume that all SNPs are independent and this may lead to false-positive results. One simple correction procedure is to take a subset of relatively independent SNP markers from genome-wide association (GWA) studies for pathway-based association analysis. For example, in one study, the SNP with the most significant *p*-value in a genic region was selected and all SNPs within 1 Mb with $r^2 > 0.2$ were removed[40]. These types of procedure are effective but they may also reduce gene coverage and result in loss of information. Therefore, it would be useful to develop more powerful methods to help reduce biases caused by different LD patterns between genes while maintaining sufficient power.
- Imputation has been commonly adopted in most GWA studies to identify markers that were not directly genotyped but are associated with disease or to combine results from different genotyping platforms. The results from imputation, especially those based on the 1000 Genomes Project, may also help incorporate rare variants into association tests. However, imputation procedures raise challenges for the appropriate summarization of gene-based *p*-values and also cause vastly increased computational burdens for raw genotype approaches that require phenotype permutation. Novel approaches are needed to incorporate information from imputation to gain more comprehensive and less-biased coverage of genes, and to improve the power to identify associated pathways.
- Complex diseases can involve multiple pathways and some pathways also share genes. Joint analysis of multiple related pathways could be a new research direction to develop more powerful association strategies. Conversely, some diseases (such as Crohn's disease and ulcerative colitis) are known to share susceptibility genes so joint analysis of related GWA study data sets may help reveal shared susceptibility pathways in a more powerful manner. This would be particularly relevant for diseases for which genetic overlap is not well understood. For example, variants at *JAZF1* and *HNF1B* confer susceptibility to both type 2 diabetes and prostate cancer[101] and it is possible that these genes function in shared pathways relevant to both diseases[102], even though multiple distinct pathways may be involved in each disease.
- Several recent studies have focused on incorporating biological network structure into the analysis of GWA study data sets[103–105]. Compared to analysis of groups of distinct genes, networks provide more information on the relatedness and interconnectivity of genes. Network analysis may enable more powerful analysis when appropriate algorithms are implemented that account for the network topology, as well as gene–gene interactions. For example, Baurley *et al.*[105] presented a pathway-modelling framework that discovers plausible pathways from observational data; biological knowledge can be readily applied *a priori* on pathway structure, and this framework allows estimation of the net effect of the pathway and the types of interactions occurring among genetic risk factors. Better characterization of gene–gene relationships and further development of approaches that incorporate network topology will result in more sensitive and powerful analysis. The development of Cytoscape[106] (a software tool to visualize molecular interactions as a network diagram), as well as many plug-ins[107,108], will facilitate network-based analysis for GWA studies.

called from sequencing data in the same framework as SNP-based GWA studies, there are specific challenges for sequencing data that need to be taken into account for pathway-based analysis. We briefly summarize some of these points below.

First, genotype calling for sequencing data, especially data with low-fold coverage, remains a major challenge. In the foreseeable future, sequence-based calls will be less accurate than genotype calls generated from SNP arrays. It is likely that probabilistic genotype calls, possibly facilitated by genotype imputation from sequencing data, will be used as units of analysis in pathway-based association tests. Second, many more genetic variants will be identified from each gene. Therefore, the number of predictors may be far more than in current GWA studies and many of the methods will probably suffer from a lack of power with the dilution effects from too many null markers. Third, compared to SNP arrays used in typical GWA studies, many more rare genetic variants will be identified from Seq-GWA studies, but

Genotype imputation
A statistical method that predicts individual genotypes at ungenotyped markers from genotypes of other nearby markers, usually using the HapMap data as a reference.

the single-marker-based analysis on each of the rare variants will lead to severe loss of power owing to the low frequency of rare variants. Several methods for testing association by combining rare variants have already been developed[96–99] and they may be useful for the pathway-based analysis of future Seq-GWA studies. Finally, Seq-GWA studies allow the direct interrogation of variants rather than just variants in strong LD with each other, so the ability to directly assay functional variants may allow more powerful analysis, but only if an approximately correct interpretation of functional weights can be applied to these variants. For exonic variants, it has been shown that by collapsing multiple rare variants, each with a different functional weight, the power of the association analysis can be improved[97]. It is likely that similar approaches can also be applied to whole-genome sequencing data but the assignment of functional scores may not be straightforward as most variants lie outside protein-coding regions. In summary, these would be important research directions for which empirical data should be gathered and theoretical development performed to lead towards more powerful pathway-based analysis of Seq-GWA studies in the future.

Conclusions

Several recently published studies have clearly demonstrated the use and importance of pathway-based approaches, which complement standard single-marker analysis in extracting more biological information from existing GWA study data sets. Despite many challenges, many opportunities also lie ahead. With the development of high-throughput sequencing techniques, pathway-based approaches may also be useful in the analysis of sequencing data if appropriate functional interpretation methods are applied to unleash the true power of genetic association studies.

1. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118**, 1590–1605 (2008).
2. Li, M., Wang, K., Grant, S. F., Hakonarson, H. & Li, C. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics* **25**, 497–503 (2009).
3. Gauderman, W. J., Murcray, C., Gilliland, F. & Conti, D. V. Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.* **31**, 383–395 (2007).
4. Wang, T. & Elston, R. C. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.* **80**, 353–360 (2007).
5. Wu, M. C. *et al.* Powerful SNP-set analysis for case–control genome-wide association studies. *Am. J. Hum. Genet.* **86**, 929–942 (2010).
6. Kwee, L. C., Liu, D., Lin, X., Ghosh, D. & Epstein, M. P. A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet.* **82**, 386–397 (2008).
7. Wang, K. & Abbott, D. A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **32**, 108–118 (2008).
8. Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
9. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
10. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Rev. Genet.* **11**, 499–511 (2010).
11. Roeder, K., Bacanu, S. A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78**, 243–252 (2006).
12. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81**, 1278–1283 (2007).
This is one of the first studies to propose the use of pathway information in GWA studies. Borrowing ideas from the gene expression microarray field, the authors adapted a GSEA approach for pathway analysis and demonstrated its use in several GWA studies.
13. Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218–223 (2009).
14. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.* **34**, 267–273 (2003).
15. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007).
16. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl.* Acac. *Sci. USA* **102**, 15545–15550 (2005).
The authors proposed the GSEA approach for analysis of expression microarray data. This approach has been modified in many subsequent studies to perform pathway-based analysis on both expression data and GWA study data.
17. Song, S. & Black, M. A. Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics* **9**, 502 (2008).
18. Hedegaard, J. *et al.* Methods for interpreting lists of affected genes obtained in a DNA microarray experiment. *BMC Proc.* **3** (Suppl. 4), 5 (2009).
19. Dong, C. T_H17 cells in development: an updated view of their molecular identity and genetic programming. *Nature Rev. Immunol.* **8**, 337–348 (2008).
20. Abraham, C. & Cho, J. H. IL-23 and autoimmunity: new insights into the pathogenesis of inflammatory bowel disease. *Annu. Rev. Med.* **60**, 97–110 (2009).
21. Yoshida, H., Nakaya, M. & Miyazaki, Y. Interleukin 27: a double-edged sword for offense and defense. *J. Leukoc. Biol.* **86**, 1295–1303 (2009).
22. Abraham, C. & Cho, J. Interleukin-23/T_H17 pathways and inflammatory bowel disease. *Inflamm. Bowel Dis.* **15**, 1090–1100 (2009).
23. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
24. Glas, J. *et al.* Evidence for *STAT4* as a common autoimmune gene: rs7574865 is associated with colonic Crohn's disease and early disease onset. *PLoS ONE* **5**, e10373 (2010).
25. Martinez, A. *et al.* Association of the *STAT4* gene with increased susceptibility for some immune-mediated diseases. *Arthritis Rheum.* **58**, 2598–2602 (2008).
26. Zhernakova, A. *et al.* Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring *CARD9* and *IL18RAP*. *Am. J. Hum. Genet.* **82**, 1202–1210 (2008).
27. Leach, S. T. *et al.* Local and systemic interleukin-18 and interleukin-18-binding protein in children with inflammatory bowel disease. *Inflamm. Bowel Dis.* **14**, 68–74 (2008).
28. Wang, K. *et al.* Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Hum. Mol. Genet.* **19**, 2059–2067 (2010).
29. Sato, K. *et al.* Strong evidence of a combination polymorphism of the tyrosine kinase 2 gene and the signal transducer and activator of transcription 3 gene as a DNA-based biomarker for susceptibility to Crohn's disease in the Japanese population. *J. Clin. Immunol.* **29**, 815–825 (2009).
30. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
31. Edwards, A. O. *et al.* Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
32. Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308**, 419–421 (2005).
33. Dinu, V., Miller, P. L. & Zhao, H. Evidence for association between multiple complement pathway genes and AMD. *Genet. Epidemiol.* **31**, 224–237 (2007).
34. Ng, T. K. *et al.* Multiple gene polymorphisms in the complement factor H gene are associated with exudative age-related macular degeneration in Chinese. *Invest. Ophthalmol. Vis. Sci.* **49**, 3312–3317 (2008).
35. Lesnick, T. G. *et al.* A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.* **3**, e98 (2007).
36. Lesnick, T. G. *et al.* Beyond Parkinson disease: amyotrophic lateral sclerosis and the axon guidance pathway. *PLoS ONE* **3**, e1449 (2008).
37. O'Dushlaine, C. *et al.* Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol. Psychiatry* 16 Feb 2010 (doi:10.1038/mp.2010.7).
38. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
39. Askland, K., Read, C. & Moore, J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.* **125**, 63–79 (2009).
40. Holmans, P. *et al.* Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* **85**, 13–24 (2009).
41. Wang, K. *et al.* Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. *Am. J. Hum. Genet.* **84**, 399–405 (2009).
The authors demonstrated a successful example in which pathway-based association approaches can identify a known disease susceptibility pathway and reveal additional susceptibility genes. Furthermore, they showed that pathway association can be replicated between different genotyping platforms or different ethnicity groups.
42. Eleftherohorinou, H. *et al.* Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE* **4**, e8068 (2009).
43. Tintle, N. L., Borchers, B., Brown, M. & Bekmetjev, A. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC Proc.* **3** (Suppl. 7), 96 (2009).
44. Ballard, D. H. *et al.* A pathway analysis applied to Genetic Analysis Workshop 16 genome-wide rheumatoid arthritis data. *BMC Proc.* **3** (Suppl. 7), 91 (2009).

45. Beyene, J. *et al.* Pathway-based analysis of a genome-wide case–control association study of rheumatoid arthritis. *BMC Proc.* **3** (Suppl. 7), 128 (2009).
46. Sohns, M., Rosenberger, A. & Bickeboller, H. Integration of *a priori* gene set information into genome-wide association studies. *BMC Proc.* **3** (Suppl. 7), 95 (2009).
47. Lebrec, J. J., Huizinga, T. W., Toes, R. E., Houwing-Duistermaat, J. J. & van Houwelingen, H. C. Integration of gene ontology pathways with North American Rheumatoid Arthritis Consortium genome-wide association data via linear modeling. *BMC Proc.* **3** (Suppl. 7), 94 (2009).
48. Torkamani, A., Topol, E. J. & Schork, N. J. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* **92**, 265–272 (2008).
49. Chen, L. S. *et al.* Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* **86**, 860–871 (2010).
The authors proposed a strategy that uses representative eigenSNPs for each gene to assess their joint association with disease risk. This approach compares favourably against other approaches that examine only the most significant SNP in each gene or SNPs passing a certain *p*-value threshold.
50. Zhang, L. *et al.* Pathway-based genome-wide association analysis identified the importance of regulation-of-autophagy pathway for ultradistal radius BMD. *J. Bone Miner. Res.* **25**, 1572–1580 (2010).
51. Peng, G. *et al.* Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* **18**, 111–117 (2010).
52. Chen, Y. *et al.* Pathway-based genome-wide association analysis identified the importance of EphrinA–EphR pathway for femoral neck bone geometry. *Bone* **46**, 129–136 (2010).
53. Lambert, J. C. *et al.* Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J. Alzheimers Dis.* **20**, 1107–1118 (2010).
54. Joslyn, G., Ravindranathan, A., Brush, G., Schuckit, M. & White, R. L. Human variation in alcohol response is influenced by variation in neuronal signaling genes. *Alcohol. Clin. Exp. Res.* **34**, 800–812 (2010).
55. Ballard, D., Abraham, C., Cho, J. & Zhao, H. Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Med. Genomics* **3**, 25 (2010).
56. Yu, K. *et al.* Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* **33**, 700–709 (2009).
57. Chen, L. *et al.* Prioritizing risk pathways: a novel association approach to searching for disease pathways fusing SNPs and pathways. *Bioinformatics* **25**, 237–242 (2009).
58. O'Dushlaine, C. *et al.* The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* **25**, 2762–2763 (2009).
59. Chai, H. S. *et al.* GLOSSI: a method to assess the association of genetic loci-sets with complex diseases. *BMC Bioinformatics* **10**, 102 (2009).
60. Chasman, D. I. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet. Epidemiol.* **32**, 658–668 (2008).
61. De la Cruz, O., Wen, X., Ke, B., Song, M. & Nicolae, D. L. Gene, region and pathway level analyses in whole-genome studies. *Genet. Epidemiol.* **34**, 222–231 (2010).
62. Zhang, K., Cui, S., Chang, S., Zhang, L. & Wang, J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res.* **38** (Suppl. 2), W90–W95 (2010).
63. Schwender, H., Ruczinski, I. & Ickstadt, K. Testing SNPs and sets of SNPs for importance in association studies. *Biostatistics* 2 July 2010 (doi:10.1093/biostatistics/kxq042).
64. Nam, D., Kim, J., Kim, S. Y. & Kim, S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res.* **38** (Suppl. 2), W749–W754 (2010).
65. Luo, L. *et al.* Genome-wide gene and pathway analysis. *Eur. J. Hum. Genet.* **18**, 1045–1053 (2010).
66. Guo, Y. F., Li, J., Chen, Y., Zhang, L. S. & Deng, H. W. A new permutation strategy of pathway-based approach for genome-wide association study. *BMC Bioinformatics* **10**, 429 (2009).
67. Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).
A crucial review of current statistical approaches used in GWA studies, including meta-analysis, epistasis analysis and pathway analysis. The authors give a few recommendations for using these approaches.
68. Hong, M. G., Pawitan, Y., Magnusson, P. K. & Prince, J. A. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* **126**, 289–301 (2009).
69. Kraft, P. & Raychaudhuri, S. Complex diseases, complex genes: keeping pathways on the right track. *Epidemiology* **20**, 508–511 (2009).
The authors discuss three loosely defined approaches to pathway analysis and touch on potential pitfalls for each when applied to GWA studies. They suggest that care must be taken to avoid biases and errors that will send researchers down blind alleys.
70. Tintle, N. *et al.* Inclusion of *a priori* information in genome-wide association analysis. *Genet. Epidemiol.* **33** (Suppl. 1), 74–80 (2009).
71. Thomas, D. C. *et al.* Use of pathway information in molecular epidemiology. *Hum. Genomics* **4**, 21–42 (2009).
72. Elbers, C. C. *et al.* Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.* **33**, 419–431 (2009).
The authors present the various benefits and limitations of pathway classification tools for analyzing GWA study data. They demonstrate multiple differences in outcome between pathway tools analyzing the same data set and suggest that the limitations of pathway approaches need to be addressed.
73. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
74. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
75. Klingstrom, T. & Plewczynski, D. Protein–protein interaction and pathway databases, a graphical review. *Brief. Bioinform.* 17 Sept 2010 (doi:10.1093/bib/bbq064).
76. Goeman, J. J. & Buhlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).
77. Keating, B. J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 K SNP array for large-scale genomic association studies. *PLoS ONE* **3**, e3583 (2008).
78. Fridley, B. L., Jenkins, G. D. & Biernacka, J. M. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS ONE* **5**, e12693 (2010).
79. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genet.* **41**, 25–34 (2009).
80. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40**, 575–583 (2008).
81. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
82. Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J. & Thomas, D. C. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet. Epidemiol.* **31**, 871–882 (2007).
83. Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. & Lange, K. Genome-wide association analysis by LASSO penalized logistic regression. *Bioinformatics* **25**, 714–721 (2009).
84. Zhou, H., Sehl, M. E., Sinsheimer, J. S. & Lange, K. Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* **26**, 2375–2382 (2010).
85. Perry, J. R. *et al.* Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes* **58**, 1463–1467 (2009).
86. Mirnics, K., Middleton, F. A., Marquez, A., Lewis, D. A. & Levitt, P. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* **28**, 53–67 (2000).
This is one of the first gene expression studies demonstrating that a group of functionally related genes may show modest yet consistent expression changes between two conditions.
87. Jiang, Z. & Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* **23**, 306–313 (2007).
88. Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Ann. Appl. Stat.* **1**, 107–129 (2007).
89. Dinu, I. *et al.* Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* **8**, 242 (2007).
90. Heller, R., Manduchi, E., Grant, G. R. & Ewens, W. J. A flexible two-stage procedure for identifying gene sets that are differentially expressed. *Bioinformatics* **25**, 1019–1025 (2009).
91. Ackermann, M. & Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* **10**, 47 (2009).
92. Glazko, G. V. & Emmert-Streib, F. Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics* **25**, 2348–2354 (2009).
93. Irizarry, R. A., Wang, C., Zhou, Y. & Speed, T. P. Gene set enrichment analysis made simple. *Stat. Methods Med. Res.* **18**, 565–575 (2009).
94. Hsu, Y. H. *et al.* An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility loci for osteoporosis-related traits. *PLoS Genet.* **6**, e1000977 (2010).
95. Zhong, H., Yang, X., Kaplan, L. M., Molony, C. & Schadt, E. E. Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* **86**, 581–591 (2010).
The authors performed an analysis that leverages information from genetics of gene expression studies to identify biological pathways enriched for expression-associated genetic loci associated with disease in GWA studies. They demonstrated the utility of integrating pathway analysis and gene expression data for interpreting signals from GWA studies.
96. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
97. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
98. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
99. Han, F. & Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **70**, 42–54 (2010).
100. Wei, Z. *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.* **5**, e1000678 (2009).
101. Frayling, T. M., Colhoun, H. & Florez, J. C. A genetic link between type 2 diabetes and prostate cancer. *Diabetologia* **51**, 1757–1760 (2008).
102. Giovannucci, E. *et al.* Diabetes and cancer: a consensus report. *CA Cancer J. Clin.* **60**, 207–221 (2010).
103. Pan, W. Network-based model weighting to detect multiple loci influencing complex diseases. *Hum. Genet.* **124**, 225–234 (2008).
104. Baranzini, S. E. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.* **18**, 2078–2090 (2009).
105. Baurley, J. W., Conti, D. V., Gauderman, W. J. & Thomas, D. C. Discovery of complex pathways from observational data. *Stat. Med.* **29**, 1998–2011 (2010).
106. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
107. Zinovyev, A., Viara, E., Calzone, L. & Barillot, E. BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics* **24**, 876–877 (2008).
108. Clement-Ziza, M. *et al.* Genoscape: a Cytoscape plug-in to automate the retrieval and integration of gene expression data and molecular networks. *Bioinformatics* **25**, 2617–2618 (2009).

109. Neurath, M. F., Fuss, I., Kelsall, B. L., Stuber, E. & Strober, W. Antibodies to interleukin 12 abrogate established experimental colitis in mice. *J. Exp. Med.* **182**, 1281–1290 (1995).

110. Neurath, M. F. IL-23: a master regulator in Crohn disease. *Nature Med.* **13**, 26–28 (2007).

111. Medina, I. *et al.* Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res.* **37**, W340–W344 (2009).

112. Holden, M., Deng, S., Wojnowski, L. & Kulle, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24**, 2784–2785 (2008).

113. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

Acknowledgements

We thank D. C. Thomas (University of Southern California) for his helpful critiques which greatly improved the manuscript.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

BioCarta database:
http://www.biocarta.com
Cytoscape:
http://www.cytoscape.org
GeneGo: http://www.genego.com
Gene Ontology database:
http://www.geneontology.org
Ingenuity Pathway Analysis:
http://www.ingenuity.com
Kyoto Encyclopedia of Genes and Genomes (KEGG):
http://www.genome.jp/kegg/pathway.html
MetaCyc:
http://biocyc.org/metacyc/index.shtml
Molecular Signatures Database:
http://www.broadinstitute.org/gsea/msigdb/index.jsp
Nature Pathway Interaction Database:
http://pid.nci.nih.gov
Pathguide:
http://www.pathguide.org
Science Signal Transduction Knowledge Environment:
http://stke.sciencemag.org
TRANSPATH:
http://www.gene-regulation.com/pub/databases.html

FURTHER INFORMATION

Kai Wang's homepage: http://www.usc.edu/schools/medicine/research/institutes/zni/faculty/profile.php?fid=145
Mingyao Li's homepage: http://www.cceb.upenn.edu/faculty/?id=159
Hakon Hakonarson's homepage:
http://www.chop.edu/service/applied-genomics/meet-the-team/hakon-hakonarson-md-phd.html
ALIGATOR: http://x004.psycm.uwcm.ac.uk/~peter
Genetic Analysis Workshop 16: http://www.gaworkshop.org
GenGen: http://www.openbioinformatics.org/gengen
GESBAP: http://bioinfo.cipf.es/gesbap
GRASS: http://linchen.fhcrc.org/grass.html
GSA-SNP: http://gsa.muldas.org
GSEA-SNP:
http://www.nr.no/pages/samba/area_emr_smbi_gseasnp
i-GSEA4GWAS: http://gsea4gwas.psych.ac.cn
***Nature Reviews Genetics* series on Genome-wide association studies:** http://www.nature.com/nrg/series/gwas/index.html
***Nature Reviews Genetics* series on Study designs:**
http://www.nature.com/nrg/series/studydesigns/index.html
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink
SNP ratio test: http://sourceforge.net/projects/snpratiotest

ALL LINKS ARE ACTIVE IN THE ONLINE PDF