Review

Pathway analysis of genomic data: concepts, methods, and prospects for future development

Vijay K. Ramanan[1,2,3], Li Shen[3,4], Jason H. Moore[5,6] and Andrew J. Saykin[1,3,4]

[1] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA
[2] Medical Scientist Training Program, Indiana University School of Medicine, Indianapolis, IN, 46202, USA
[3] Center for Neuroimaging, Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, 46202, USA
[4] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA
[5] Department of Genetics and Institute for Quantitative Biomedical Sciences, Dartmouth Medical School, Lebanon, NH, 03756, USA
[6] Department of Community and Family Medicine, Dartmouth Medical School, Lebanon, NH, 03756, USA

Genome-wide data sets are increasingly being used to identify biological pathways and networks underlying complex diseases. In particular, analyzing genomic data through sets defined by functional pathways offers the potential of greater power for discovery and natural connections to biological mechanisms. With the burgeoning availability of next-generation sequencing, this is an opportune moment to revisit strategies for pathway-based analysis of genomic data. Here, we synthesize relevant concepts and extant methodologies to guide investigators in study design and execution. We also highlight ongoing challenges and proposed solutions. As relevant analytical strategies mature, pathways and networks will be ideally placed to integrate data from diverse -omics sources to harness the extensive, rich information related to disease and treatment mechanisms.

The search for pathways in complex diseases: a seminal moment

Since 2005, over 1000 human genome-wide association study (GWAS) publications have described genetic associations to a wide range of diseases and traits [1]. However, extending GWAS findings to mechanistic hypotheses about development and disease has been a major ongoing challenge. In particular, the focus on single loci has been confounded by two insights: (i) most GWAS-implicated common alleles and differentially expressed genes on expression arrays have exhibited modest effect sizes; and (ii) genes function within biological pathways and interact within biological networks [2]. As such, genome-wide data sets are increasingly viewed as foundations for discovering pathways and networks relevant to phenotypes [3]. This trend is vital, given that pathway mechanisms are natural sources for developing strategies to diagnose, treat and prevent complex diseases. In this context, it is not surprising that pathway-based analyses have exploded in use during the past 3–5 years (Figure 1).

In pathway analysis, gene sets corresponding to biological pathways (Box 1) are tested for significant relationships with a phenotype. Primary data for pathway analysis are commonly sourced from genotyping or gene expression arrays, although in theory any data elements that could be mapped to genes or gene products could be used. Importantly, analyzing genomic data through functionally derived gene sets can reveal larger effects that are otherwise concealed from gene- or single nucleotide polymorphism (SNP)-based analysis. For example, high-profile studies in breast cancer [4], Crohn's disease [5] and type 2 diabetes [6] demonstrate that functionally related genes can collectively influence disease susceptibility, even if individual loci do

Glossary

Bootstrapping: a method that assesses the uncertainty of a statistical estimate through recalculation of the statistic using repeated, random sampling of the original data set.
Commercial pathway database: a collection of pathway annotation data that is available for private purchase by investigators.
Covariate: a variable that is possibly predictive of the outcome under study; for example, genetic analyses often attempt to account for the effects of variables such as age and gender to precisely determine statistical relationships between genetic factors and phenotypes.
Freeware pathway database: a collection of pathway annotation data that is publically available without cost to the user.
Genome-wide association study (GWAS): a large-scale study that assays genetic variants across the entire genome along with quantitative or categorical phenotype status to detect genotype–phenotype associations.
Genomic inflation: the systematic increase of association statistics from a genome-wide study owing to population stratification or other confounding factors.
Genotype imputation: the process of probabilistically predicting genotypes that are not directly assayed (by not being represented on that genotyping platform or via localized experimental failure) with a particular array.
Granularity: a description of the scale or level of detail in a set of data.
Linkage disequilibrium (LD): the non-random association of alleles at two or more loci; in other words, the occurrence of combinations of alleles at different frequencies than would be expected through a random formation of haplotypes.
Permutation: the process of calculating the distribution of a test statistic under the null hypothesis through repeatedly rearranging the labels in a data set; for example, in case-control studies, phenotype statuses of subjects are randomly rearranged to assess the distribution of an association statistic under the null hypothesis of no significant association between a marker and phenotype status.
Replication: the repetition of a research study in an independent sample to verify first-line results and to determine whether effects can be generalized beyond the initial sample.

Corresponding author: Saykin, A.J. (asaykin@iupui.edu).

Keywords: pathway analysis; gene set; enrichment methods; genome-wide association study; functional annotation; complex diseases.

0168-9525/$ – see front matter © 2012 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2012.03.004

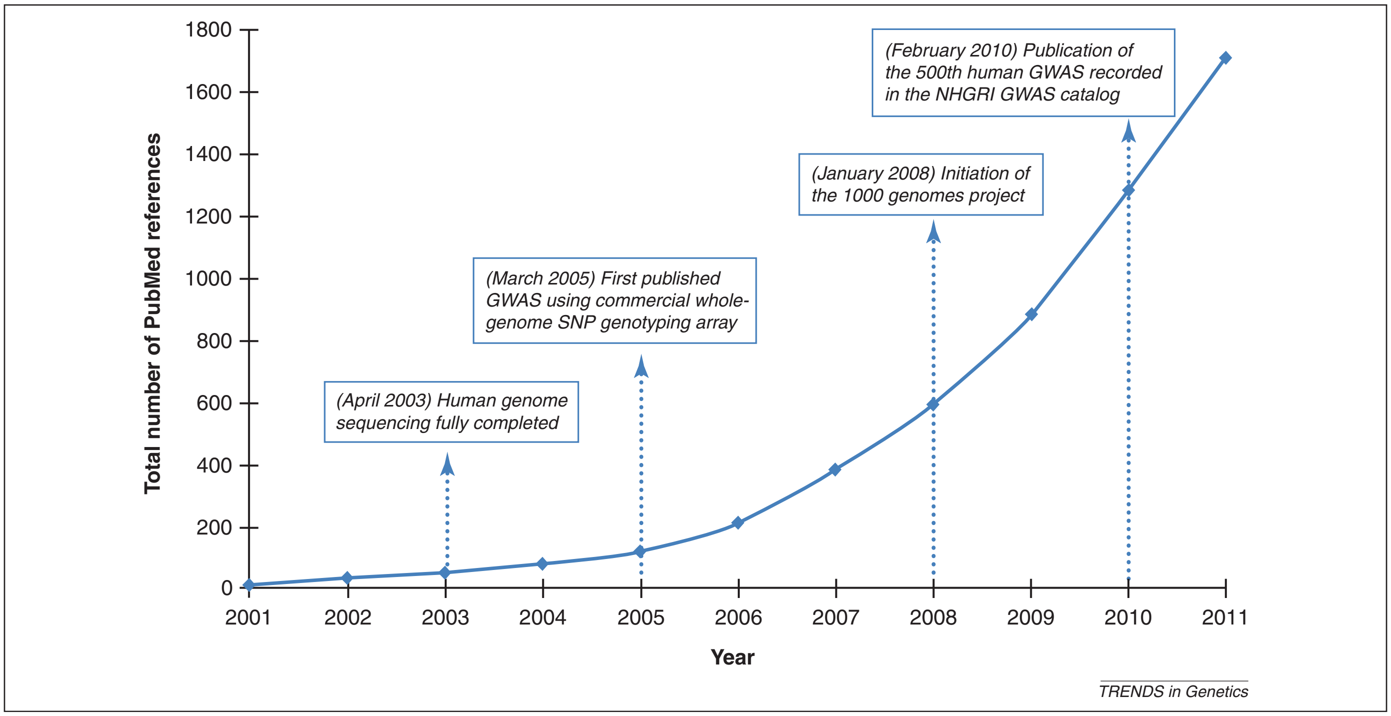

Figure 1. PubMed citations for 'pathway analysis' from 2001 to 2011. The use of pathway analysis has grown exponentially in the past 3–5 years. This explosion in use has followed major developments (shown in boxes) in characterizing the human genome and in performing genome-wide studies of complex diseases and traits. Data points represent the total number of references displayed through a PubMed search for 'pathway analysis', using date limits of January 1, 2001 and December 31 of the calendar year denoted on the *x*-axis. Abbreviations: GWAS, genome-wide association study; NHGRI, National Human Genome Research Institute; SNP, single nucleotide polymorphism.

not exhibit genome-wide significant association. As such, pathway analysis represents a potentially powerful and biologically oriented bridge between genotypes and phenotypes.

Despite their popularity and potential, strategies for pathway-based studies have progressed in the absence of guidelines, leading to ambiguity regarding optimal methods, high variability in results and barriers to further application. With surging interest in pathway analysis and the emergence of next-generation sequencing data, which will inevitably broaden its application, this is an ideal moment for a critical synthesis of current approaches and an outlining of targets for future development. Here, we clarify fundamental concepts about pathways and networks and their relationships to study design and execution. We also review extant strategies to detect pathway–phenotype association and highlight methodological challenges. Finally, we describe how pathways and networks are ideal vehicles for leveraging multi-omics data for discovery.

Selecting an overall study design

Broadly, there are two approaches to pathway-based genomic studies. Candidate pathway analysis is hypothesis driven: pathways are preselected based on prior knowledge and insight. Although the number of candidate pathways may vary with study goals (e.g. different effects may be seen within a large, complex pathway compared with numerous, smaller pathways), this approach is marked by its use of a biologically targeted subset of genomic data. The other approach, genome-wide pathway analysis (GWPA), interrogates a complete genomic data set through pathways representing an extensive range of biology. Notably, the line between 'targeted' and 'extensive' biological coverage is not precisely drawn. Although methods limited to GWPA have been used on data sets with only 1000 genes (approximately 5% of the total number of human genes) [7], the optimal point of delineation between these two approaches warrants further examination.

There are several advantages to the candidate pathway approach. Focusing the scope of analysis can enable otherwise intensive procedures, such as genotype imputation and manual pathway curation; by maximizing annotation coverage and quality, these procedures can bridge differences in genotyping platforms across cohorts for replication or meta-analysis. Unfortunately, targeted biological coverage may fail to detect unexpected relationships, such as the association between inflammatory pathways and age-related macular degeneration [8]. Furthermore, poor annotation of one pathway can be particularly limiting when only a few pathways are assessed. These traits make candidate pathway analysis most appropriate where computational resources are limited and where specific pathways are of *a priori* interest.

By contrast, GWPA maximally utilizes the available genomic data. As a result, this approach can more readily detect unexpected relationships, including those across diseases operating in different body systems [9]. However, GWPA is computationally intensive, requiring more stringent corrections for multiple comparisons and making procedures, such as imputation, more challenging. Although strategies to reduce the dimensionality of genome-wide data for pathway analysis are in active development [10,11], they will need to be evaluated further ahead of widespread use. Finally, GWPA benefits from

Box 1. Fundamental concepts about biological pathways and networks

Although unstated notions pre-date it, the first explicit description of a pathway as the events by which intermediates are processed in a defined sequence was provided in 1973 [84]. Recently, broader notions of pathways as collections of biologically related genes [24] have attempted to fit evolving scientific theories and analyses. A systematic conceptualization of biological pathways (Figure Ia) posits that pathways are vector driven toward an essential goal (i.e. their constituents as a whole are directed to a common, specific end point). Viewed this way, molecular pathways have an essential goal of basic biochemical action on molecules or compounds. Overarching this are cellular pathways that regulate global cellular status and organ and/or system pathways that execute broader physiological functions. The constituents of pathways are typically connected through known or proposed mechanisms. Of note, the particular constituents of a pathway may be context dependent, specifically, in relation to the biological outcome an investigator wishes to study.

In addition, two other types of pathway are important in the study of genetically complex diseases (Figure Ia). Disease pathways have an essential goal of the pathogenesis of a disease and its features. For example, the Alzheimer's disease pathway plausibly includes components from the organ and/or system pathway of memory, which itself has cellular and molecular underpinnings. By contrast, intervention pathways are defined within the setting of a therapy that targets disease features or pathogenesis, as in a pathway-based study of cisplatin sensitivity in ovarian cancer [85]. Importantly, disease and intervention pathways may include constituents with documented associations to a phenotype, but whose precise mechanistic roles are not yet known.

Networks can also collect genes and other biological elements for quantitative and visual assessment of relationships [86]. Unlike pathways, biological networks are not vector driven toward an essential outcome (Figure Ib). Instead, networks are characterized by nodes that are connected by edges representing defined relationships. In a particular network, nodes may represent almost any biological element, including genes, gene products, non-gene DNA sequences, pathways, diseases, therapies, or combinations thereof. Common examples of network relationships include binding in protein interaction networks and regulation by common factors in gene interaction networks. Finally, statistical networks display relationships, such as correlation, that are inferred from computational analyses [70]. A central outstanding question involves understanding the degree of connection between statistically inferred networks and biological networks [87]. Software platforms for network analysis include IPA (Ingenuity Systems, http://www.ingenuity.com/) and Cytoscape (http://www.cytoscape.org/); two recent reviews discuss these and other network-based tools in detail [88,89].

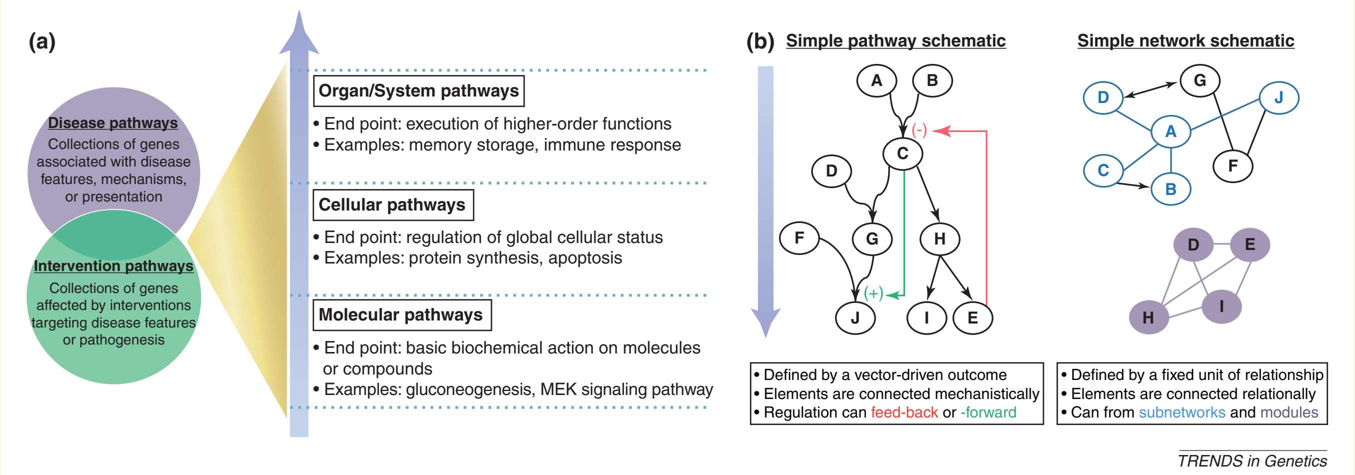
(a)
Disease pathways
Collections of genes associated with disease features, mechanisms, or presentation
Intervention pathways
Collections of genes affected by interventions targeting disease features or pathogenesis
Organ/System pathways
• End point: execution of higher-order functions
• Examples: memory storage, immune response
Cellular pathways
• End point: regulation of global cellular status
• Examples: protein synthesis, apoptosis
Molecular pathways
• End point: basic biochemical action on molecules or compounds
• Examples: gluconeogenesis, MEK signaling pathway
(b) Simple pathway schematic
A
B
(-)
C
D
F
G
H
(+)
J
I
E
Simple network schematic
D
G
J
A
C
B
F
D
E
H
I
• Defined by a vector-driven outcome
• Elements are connected mechanistically
• Regulation can feed-back or -forward
• Defined by a fixed unit of relationship
• Elements are connected relationally
• Can from subnetworks and modules
TRENDS in Genetics

Figure I. A primer on biological pathways and networks. **(a)** The major types of biological pathway are shown along with a representation of their relationships among each other. Each type of pathway is defined by its essential goal. Molecular pathways have an essential goal of basic biochemical action (biosynthesis, biodegradation, translocation, transformation, activation or inactivation) on molecules or compounds. Cellular pathways regulate global cellular status, whereas organ and/or system pathways execute higher-order physiological functions. **(b)** Pathways and networks, although complementary sets of biological elements, differ in key respects. Pathways can include directional regulation (shown in red and green) and branching, but are nevertheless vector driven to an essential outcome. Although elements in pathways are typically connected mechanistically, network elements are connected through shared relationships that may not indicate an action. As such, networks are not vector driven from a starting point to an essential outcome. Networks can be divided into subnetworks (shown in blue) exhibiting all elements connected to a central node ('A' in this example) or into modules (shown in purple) that exhibit a high density of connections.

systematic follow-up to deal with the often high overlap of genes across multiple pathways and to evaluate results in view of prior knowledge.

Obtaining input genomic and pathway annotation data

Pathway analyses can utilize raw genotype data for individual subjects [6,12,13] or a list of *P*-values relating genes or SNPs to a phenotype [14–16]. Pathway-based tools for raw genotypes do not effectively include covariates but can naturally correct for linkage disequilibrium (LD) through permutation. By contrast, *P*-value distributions are readily accessible via other researchers and can be generated with application of covariates, but require corrections for LD based on reference populations. Investigators should consider their resources and study goals when selecting the most appropriate genomic data source.

In parallel, a pathway analysis is only as good as the functional information underlying its pathway definitions. Prominent pathway annotation databases exhibit diverse features (Table 1; also see the online resource *Pathguide* [17]). The ideal choice of database depends on several variables and their impact on study goals. For example, freeware databases are commonly used because of their ease of access, transparency of features and visibility in publications. Commercial databases may require a significant investment; however, they are typically linked to user-friendly statistical analysis software and often include high-quality pathway graphics that can be exported to manuscripts. Investigators should weigh the relative importance of these factors during selection.

Pathway curation methods can also impact analyses. Most databases rely on expert review for pathway curation;

Table 1. Prominent pathway annotation databases

Name	Curation[a]	Major features	URL
Biocarta	M	Driven by user input with expert review of some pathways	http://www.biocarta.com/
DAVID	M/E	Augments and integrates annotations from other databases	http://david.abcc.ncifcrf.gov/
GO	M/E	Largest database; hierarchical structure; can filter data by evidence codes	http://www.geneontology.org/
Ingenuity	M/E	Large collection of canonical pathways; high-quality pathway maps	http://www.ingenuity.com/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	M	Reference pathways (mosaics from several organisms) and organism-specific annotations; pathway maps link to closely related genes	http://www.genome.jp/kegg/
MetaCore	M	Extensive disease pathways; can edit pathway maps for publication	http://www.genego.com/
MetaCyc	M	Metabolic pathways; can visualize connections among pathways	http://metacyc.org/
Molecular Signatures Database (MSigDB)	M/E	Can download pathways from several other databases as a collection for input to analytical software; novel groupings (e.g. motif gene sets)	http://www.broadinstitute.org/gsea/msigdb/index.jsp/
PANTHER	M	Can predict protein functions from sequence and evolutionary data	http://www.pantherdb.org/
Pathway Interaction Database (PID)	M/E	Broad range of cellular pathways with special focus on cancer signaling; can generate interaction maps from a list of genes	http://pid.nci.nih.gov/
Reactome	M	Pathways are extensively cross-referenced to PubMed, HapMap and other resources; can overlay expression or other data onto pathway maps	http://www.reactome.org/ReactomeGWT/entrypoint.html/
ResNet Series	M/E	Regular updates through web server; optional user editing or text scanning of user documents; links to reference articles	http://www.ariadnegenomics.com/

[a] Abbreviations: M, manual; M/E, manual and electronic.

however, users of these databases should be aware of their update intervals and criteria used as evidence for inclusion in pathways. Alternatively, electronic curation uses text-searching algorithms to infer functional relationships. Although these inferred annotations can be useful for hypothesis generation, their accuracy is unreliable [18], making them unsuited to many pathway analyses. Finally, targeted manual curation can be particularly appropriate when an investigator has expertise in a biological realm that is poorly annotated in databases. Although potentially time-consuming, manual curation can synthesize recent results with established relationships to produce novel candidate pathways [19,20] or gene sets representing positive controls for pathway analysis [21].

Lastly, the biological coverage of pathway annotations should be considered. Across databases, similarly named pathways can exhibit vast differences in constitution, whereas differently named pathways can exhibit significant overlap. As a result, investigators should attempt to match study goals with database coverage. For example, specialized, high-granularity databases are most useful for candidate studies of intricate signaling pathways, whereas canonical pathway collections (representing well-established pathways) provide a broad biological scope that is well suited for screening-oriented studies.

This collective diversity of features is a major factor in explaining why different databases can yield divergent results from the same input data [22]. As such, an early discussion of pathway analysis recommended the use of multiple databases for each analysis [23]. This approach can balance the relative characteristics of each database used and can yield a measure of validation when different databases yield similar results. However, this strategy is most effective when it is supplemented by a systematic review of the results. Alternatively, further analyses can reveal broader findings that drive association signals across multiple smaller pathways: for example, one study analyzed pathway sets obtained through hierarchical clustering and identified an association between the canonical MAPK (mitogen-activated protein kinase) signaling pathway and breast cancer [4].

Preparing data for association testing

Systematic processing of input genomic data and pathway annotation data is vital for pathway analyses. Although some relevant methods are actively evolving, optimized approaches to major issues can minimize variation in results and interpretation.

Pathway size

Most pathway analyses place constraints on pathway size: small pathways can exhibit false positive associations because of large single-gene or single-SNP effects [24], whereas large pathways are more likely to show association by chance alone [22]. The most common minimum threshold for pathway size appears to be ten genes [4,6,13,25]. It is important for analysts to note that this threshold may exclude highly specific and potentially informative functional sets, including those involving protein complexes and DNA sequence motifs. Frequently used maximum thresholds for pathway size include 100 [4] and 200 [6,25] genes. Notably, in the latter two studies, upper limits of 300 [6] and 400 [25] genes did not alter the results. However, larger pathways are relatively rare and often derive their size from being more general in scope; thus, their exclusion may not significantly affect analyses or downstream biological interpretation. Overall, investigators should consider their study goals when applying such thresholds and should evaluate results in that context. Although future efforts might develop size-dependent statistical corrections, at present the reporting of pathway size and related summary statistics (e.g. [26]) alongside association data can aid interpretation.

Pathway overlap

Genes and their products typically act in multiple pathways [2], and each role is potentially important to a disease or treatment mechanism. As a result, analyses can expect to have some degree of pathway overlap. However, high pathway overlap can obscure the true source of an association signal. Although this problem can exist with any pathway analysis, Gene Ontology (GO) annotations are particularly susceptible because of the large, hierarchical structure of the database [27]. Some studies have restricted analysis of GO terms to certain levels in the hierarchy [13,28], whereas a new Bayesian method incorporates the structure of the hierarchy as prior information into its pathway association metric [29]. However, users of these approaches should be aware that the information content at particular GO levels is unpredictable [30]. Pathway overlap can also be addressed during post-analysis to prioritize related pathways for further exploration. Extant strategies include hierarchical clustering in a study of breast cancer [4], overlap-based network creation in the visualization tool ‘Enrichment Map’ [31] and the listing of overlapping pathways alongside results in the analytical software PARIS [32].

Assigning data elements to genes

Genomic data have historically been integrated into pathways by mapping assayed elements to genes. For SNP-based genotyping arrays, this is not straightforward because many array SNPs are not located in known coding or regulatory regions. In one study, all SNPs that were not mapped to a single gene through a reference genome build were discarded, but this resulted in a loss of more than 25% of assayed SNPs [33]. Alternatively, each unmapped SNP can be assigned to its nearest gene [34]. However, evolving theories suggest that sequences are not associated to genes based on closest proximity, and may not even be solely associated to one gene [35,36]. Hence, many studies assign unmapped SNPs to all genes within a distance window, ranging from 10 kb to 500 kb [13,25,26,37]. Studies taking this approach should be aware that some SNPs may not be functionally related to their assigned gene(s). In addition, SNPs that map to multiple genes in the same pathway can yield spurious pathway association. This issue is particularly important for genes (such as the major histocompatibility complex/human leukocyte antigen (MHC/HLA) genes) that cluster in the genome and belong to the same pathway, because variants in those genomic regions can potentially map to all genes in the pathway. Finally, given the importance of SNP-to-gene mapping for pathway analyses, investigators should be aware that imputation can increase gene coverage by characterizing SNP genotypes that are not directly available in a particular data set. Imputation can be particularly useful for bridging differences in genotyping platforms across cohorts for replication and meta-analysis, and can also enable investigation of rare alleles and copy number variants (CNVs) that are less represented on standard platforms [38].

Calculating gene significance and accounting for LD

Most pathway analysis tools utilize one association signal per gene. Whereas expression arrays yield a single *P*-value for each gene, SNP arrays include multiple signals per gene, some of which are correlated. As such, some studies use the minimum SNP-level *P*-value within a gene as the operative signal [4,25,33,34]; however, this approach will not detect additive effects among SNPs with moderate individual association. For methods that combine SNP-level signals, including those based on the truncated product method [14], LD must be accounted for to prevent highly correlated SNPs from biasing gene-level significance. Strategies to accomplish this include discarding SNPs that depart from LD at a pre-set threshold [25,26,39] and adapting principal component analysis to extract the most independent signals within a gene [10,11,26]; unfortunately, these methods can eliminate substantial information. Alternatively, the SNP ratio test [40] and the ‘set-based analysis’ in PLINK [41] use phenotype permutation to correct naturally for biases introduced by LD and gene size; however, these tools require raw genotype data and are computationally demanding, making them better suited for studies of candidate pathways with relatively few genes. Notably, recently developed methods that accept *P*-values as input and account for LD through simulations [42,43] or genotype permutation [32] are computationally efficient and may represent new paradigms as their power is honed and evaluated.

Analytical methods to detect pathway–phenotype relationships

Following data processing, analytical methods can be applied to test for significant pathway–phenotype relationships. Prominent examples of pathway-based analytical tools and their salient features are provided in Table 2. Notably, one class of tools uses text mining of published abstracts to identify potential pathway–phenotype relationships. These tools query a list that may include SNPs meeting a *P*-value threshold, genes from candidate pathways, or pathways themselves, among other possibilities. Text-mining approaches have efficiently identified potential interactions among genes associated with neurodegenerative brain changes [20] and have equally been applied to generate a candidate pathway based on regulation or interaction with *BRCA2* (*breast cancer 2, early onset*) [44].

By contrast, pathway-enrichment tools assess for a statistically significant distribution of association within a pathway. Competitive enrichment methods compare the collective association within a pathway to the collective signal among genes not in the pathway [45]. As a result, competitive methods are not suitable for candidate pathway analyses that do not have an appropriate complement of data from outside of the candidate pathways. Meanwhile, self-contained enrichment methods test the signal within a pathway against simulated data sets that are expected to have no significant phenotype association [45,46]. Self-contained methods can be challenging to use in a screening-oriented GWPA because of the computational demand of

Table 2. Examples of publically available pathway-based analytical tools

Name	Type[a]	Input data	Analytical method	Corrections included	Refs
Chilibot	TM	Word list	Searches PubMed abstracts for relationships among word list; can distinguish biological concepts (e.g. activation or inhibition)	N/A	[90]
GenGen	C	Raw genotype data	Uses best *P*-value as gene-wide score and calculates rank-based Kolmogorov–Smirnov-like pathway statistic with permutation	LD, pathway size, gene size, FDR	[49]
GeSBAP	C	Gene or SNP *P*-values	Uses best *P*-value as gene-wide score and performs rank-based Fisher's exact test to detect pathway enrichment	FDR	[91]
GRAIL	TM	SNPs or genomic regions	For multiple disease-associated regions, identifies functionally related genes that probably highlight causal pathways	Number of genes per region	[92]
GRASS	SC	Raw genotype data	Uses principal component analysis to select representative eigenSNPs for each gene for pathway-based ridge regression	LD, gene size, FDR	[93]
GSA-SNP	C	SNP *P*-values	Uses -log (*k*th best *P*-value) as gene-wide score and calculates a z-score, iGSEA or MAXMEAN statistic for the pathway	Pathway size, FDR	[52]
GSEA-P	C	Gene *P*-values	Calculates rank-based Kolmogorov–Smirnov-like pathway statistic with phenotype permutation	LD, pathway size, FDR	[94]
GSEA-SNP	SC	Raw genotype data	Uses all SNPs for a pathway MAX-test (maximum of Cochran–Armitage trend tests under three genetic models) with permutation	LD, pathway size, gene size, FDR	[50]
MAGENTA	C	SNP *P*-values	Modified approach based on GSEA-SNP for meta-analytic data	LD, gene size, FDR	[95]
PARIS	SC	SNP *P*-values	Identifies the significant genomic features within a pathway and performs genomic permutation to assess pathway significance	LD, pathway size, gene size, FDR	[32]
PLINK set test	SC	Raw genotype data	For SNPs passing a *P*-value threshold, calculates the average test statistic for the independent SNPs within a pathway	LD, pathway size, gene size, FDR	[41]
SNP ratio test	SC	Raw genotype data	Calculates the ratio of significant SNPs to all SNPs in a pathway and uses phenotype permutation to calculate empirical *P*-value	LD, pathway size, gene size, FDR	[40]

[a]Abbreviations: C, competitive enrichment; SC, self-contained enrichment; TM, text-mining.

generating simulated data sets. In addition, self-contained approaches are particularly susceptible to false positives through genomic inflation, as each pathway is evaluated independently from any other data on the source assay. Although one study [47] normalized all association statistics to a genomic inflation factor calculated by PLINK, best practices in this area have not yet been settled. Competitive tests are more robust in controlling genomic inflation, but they can also relinquish power in data sets with diffuse association signals [45]. As such, the optimal method depends on study goals, data set properties and computational resources.

Among extant competitive enrichment methods, three analytical frameworks predominate. In the first of these, threshold-based approaches, hypergeometric, chi-square or Fisher's exact test statistics are used to identify pathways that are overrepresented among the 'significant' markers under study. Notably, the threshold for 'significance' is arbitrary and can affect results [48]; observed SNP-level thresholds have ranged from $P < 0.05$ [37] to $P < 5{\times}10^{-8}$ [34]. By contrast, rank-based approaches order all of the markers being studied by their significance and then test for pathways that have lower rankings than the overall distribution. Whereas the rank-based tools GenGen [49] and GSEA-SNP [50] use a Kolmogorov–Smirnov-like running sum that gives greater weight to more significant markers, others rely on MAXMEAN-related statistics as potentially powerful and efficient alternatives [51–53]. Compared with threshold-based methods, rank-based approaches more naturally account for differences in significance among markers [24] but may also be heavily influenced by a few highly significant markers [54]. Finally, z-score methods infer enrichment based on deviation from a normal distribution that accounts for the size of each pathway [52,55]; although these methods are sensitive and fast, their error rates have not been well characterized. Self-contained enrichment methods use even more diverse statistical methods to combine the *P*-values within a pathway into an aggregated measure (Table 2). However, in the absence of large-scale power comparisons among related methods across several well-characterized data sets, the choice of a particular enrichment tool may be less important than understanding the relative strengths and limitations of these broader categories.

An alternative to enrichment methods are module-based approaches, which examine sets defined by other biological characteristics for meaningful pathways contained therein. For example, one study used hierarchical clustering to form modules of coexpressed genes across multiple inflammatory diseases; subsequent analysis of these modules suggested a role for interferon-inducible signaling in TB [56]. Gene modules can also be defined through protein interaction networks, as in a study that associated genetic variants in glutamate pathways to brain glutamate concentration in multiple sclerosis [57]. Importantly, recent studies are combining enrichment and module-based methods to point to broader findings. For example, network analysis of enriched pathways revealed major roles for antigen presentation and interferon signaling in rheumatoid arthritis [58].

Finally, developing strategies are targeting specific pathway-based challenges. For example, machine-learning approaches [11,59] attempt to identify the most informative subsets of genes within pathways for association. Networks have been effective in studies of rare variants, as with the identification of a synaptogenesis gene network affected by rare copy number variants (CNVs) in autism [60]. Pathway-based methods for studying rare variants using genomic region-based mapping and self-contained tests are also evolving [61,62]. Indeed, the appeal of pathways and networks will continue to expand as their associated tools progress to analyze a variety of data through user-friendly platforms.

Post-analysis considerations

Following pathway analysis, appropriate data reporting and interpretation are imperative. Currently, bias introduced by gene size is less commonly addressed than is bias from pathway size. In particular, large genes containing many SNPs are more likely to contain significant SNPs by chance alone [63]; for analyses, this can favor pathways containing large genes. Analytical tools that use permutations naturally control for gene size by comparing the actual association data with the distribution of association statistics generated from randomly permuted data sets expected to reflect chance-based confounding effects. Other approaches [41,42] allow users to restrict analysis to a subset of the most significant SNPs in each gene: for large genes, this may eliminate some spuriously associated SNPs and thus limit their impact on the pathway analysis. As a minimum, studies should discuss the potential impacts of gene and pathway size on their results. Other sources of bias that should be addressed include the capacity for strongly associated markers to drive pathway association and the possible effects of SNPs being assigned to multiple genes.

Correction for multiple comparisons must also be applied to pathway *P*-values to control for false positives. As in other areas of statistical genomics, optimizing methods for correction is a work in progress. Bonferroni-related methods seem too conservative for pathway analyses because they do not allow for dependence across pathways. False discovery rate (FDR) approaches [64] are frequently applied in pathway analyses [6,26,48], whereas newer FDR-based [65] and bootstrapping [39] methods that permute on raw genotypes can better account for pathway overlap but require large computational capacity.

Fundamentally, these approaches to bias are best complemented by replication of pathway analysis findings in independent data sets. Strategies for pathway analyses can flexibly adapt to differences across data sets and, although these differences might impact SNP- or gene-level statistics [66], legitimately associated pathways would be expected to exhibit significance or a strongly trending signal across multiple studies. In this effort, a systematic framework illustrating key choices in pathway analyses (Figure 2) will limit major contributors of variance across studies and will guide investigators in selecting approaches that fit their study goals.

Future developments in genomic data analysis

Development of methods and tools related to pathway analysis is ongoing and dynamic. In particular, because pathways are of broad interest, targeted adaptations to their associated databases would expand their utility for investigators from a variety of backgrounds. These adaptations might include simpler search and download mechanisms, consistency in pathway names and classifications, and methods for describing pathway overlap. In addition, a universal format for annotation files might encourage interoperability among analytical tools, allowing investigators flexibility to match precisely their databases and statistical methods of choice.

Two recent trends among databases are also promising. Specialized disease databases, such as AlzGene [67] and the UCSC Cancer Genomics Browser [68], can aggregate

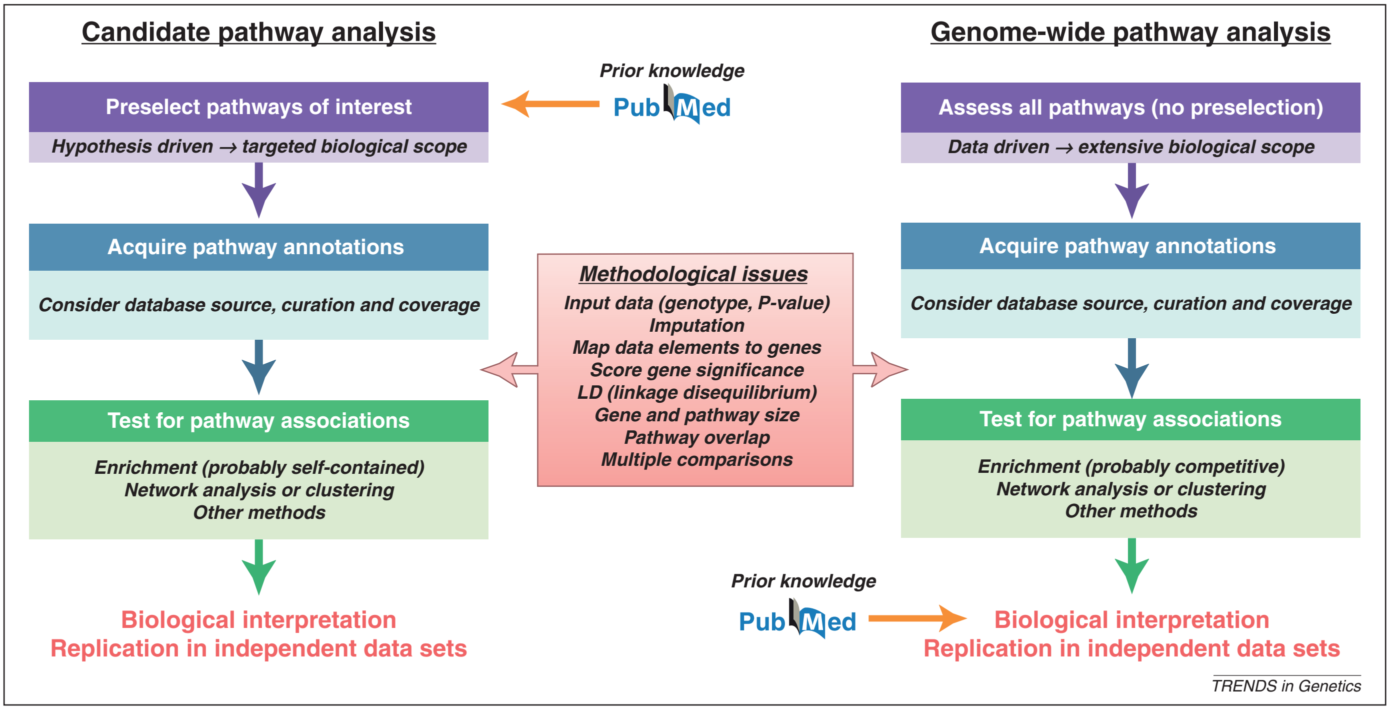
Candidate pathway analysis
Genome-wide pathway analysis
Prior knowledge
PubMed
Preselect pathways of interest
Hypothesis driven → targeted biological scope
Assess all pathways (no preselection)
Data driven → extensive biological scope
Acquire pathway annotations
Consider database source, curation and coverage
Acquire pathway annotations
Consider database source, curation and coverage
Methodological issues
Input data (genotype, P-value)
Imputation
Map data elements to genes
Score gene significance
LD (linkage disequilibrium)
Gene and pathway size
Pathway overlap
Multiple comparisons
Test for pathway associations
Enrichment (probably self-contained)
Network analysis or clustering
Other methods
Test for pathway associations
Enrichment (probably competitive)
Network analysis or clustering
Other methods
Prior knowledge
PubMed
Biological interpretation
Replication in independent data sets
Biological interpretation
Replication in independent data sets
TRENDS in Genetics

Figure 2. An informed guide to pathway analysis. Broadly, there are two approaches to pathway analysis. In candidate pathway analysis, prior knowledge is used to select pathways hypothesized to have a relationship with a phenotype. By contrast, genome-wide pathway analysis is designed to uncover significant pathway–phenotype relationships within a large data set; insight and prior knowledge are then used to interpret the findings. In both approaches, care must be taken in acquiring pathway annotations and in selecting an appropriate analytical test for association. In addition, other methodological issues (red box) guide the choice of approach and impact strategies for confounding factors. Finally, replication of pathway analysis findings in independent data sets is imperative in validating results to extend their impact.

salient information from diverse studies on a particular disease. These targeted resources are particularly up-to-date and can facilitate collaboration within highly investigated diseases. Functional annotation of genes is also becoming prominent. These annotations draw on experimental data that indicates function, location of action, or physiological region of association [69], and can allow investigators to develop candidate pathways related to localized anatomical or physiological derangements. Extensions of this concept across disciplines will probably be a prime area of advancement.

In future pathway analysis platforms, computational efficiency will be highly valued, given the impressive granularity of next-generation sequencing data. In addition, investigators may wish to use different genomic data sets, pathway annotation databases and analytical parameters, depending on study resources and goals; as such, tools that are flexible to various study approaches will maximize their impact. Finally, given that genes constitute only 1–2% of the human genome, strategies to leverage both genic and non-genic data for pathway analysis may provide increased power to detect meaningful functional sets.

Meanwhile, complementary methods can extend the biological reach of pathway-based results. For example, it is not yet understood whether gene interactions are more likely within a given pathway or across different pathways in a network. A comparative study of epistasis in pathways and networks, perhaps utilizing novel techniques for its detection within population data [70–73], could inform future strategies in this area. A related area of development involves using known protein interactions to generate subnetworks from enriched pathways; these subnetworks can highlight novel candidate genes [74] or regulatory relationships [75] from significant pathways.

Nevertheless, the ongoing development of pathway-based tools would benefit from further empirical evaluation of current approaches. For example, a creative meta-analysis might examine how various association metrics affect the likelihood of replication of findings. In addition, testing association methods against well-calibrated positive and negative control data sets might illuminate their relative capabilities. Notably, one study employed multiple pathway analysis algorithms using an extensively explored Crohn's disease data set [76]; however, the algorithms chosen were highly disparate in their null hypotheses and approaches to LD, making it difficult to compare their results uniformly. Alternatively, multi-site collaborations might simultaneously analyze several large data sets using a small number of analytical tools in the same conceptual category; comparisons of the results would advance the underlying science and critically evaluate tools against closely related options.

Finally, methods for integrating different types of association signal are being developed. A nascent view proposes that combining genome-wide expression and genotyping data into a joint quantitative signal can increase power for discovery [6,37,77,78]. One particularly attractive feature of this view is that it augments structure (genotype) with function (expression). Indeed, one study demonstrated that SNPs correlated with gene expression changes [expression quantitative trait loci (eQTLs)] were more likely to show disease association than were other SNPs from a GWAS array [79]. Relatedly, visualization tools can graphically overlay association metrics onto other data to prioritize markers. Visualization has been used to integrate SNP association with quantitative imaging phenotypes [80], among other examples.

Pathways and networks: bridging multi-omics data

As pathway analysis of genomic data has exploded in use, its methods have matured, its results are beginning to meet its potential and points of consensus are emerging for its continued application and future development. In the coming years, we anticipate that pathways and networks will assume a farther-reaching role in view of the need to integrate multi-omics data through systems biology approaches [81,82]. A variety of large-scale strategies are being used to study complex diseases, including genomic, transcriptomic, proteomic and metabolomic approaches, and data from all of these sources can be analyzed through pathways and networks representing coordinated functions and relationships. Importantly, although gene associations do not always indicate therapeutic targets [83], pathways and networks implicated by analyses at multiple levels would be prime targets for therapies. Integrating large-scale data assayed through diverse strategies related to structure and function would provide a fertile process for exploring connections between replicable, statistical association and meaningful biology. As such, the role of pathways and networks as the hub for this integration will be vital in the years to come.

Acknowledgments

Support for this work was provided by the Indiana University Medical Scientist Training Program grant NIGMS GM077229-02 (V.K.R.), the National Science Foundation grant IIS-1117335 (L.S.) and the National Institutes of Health grants AG024904 (A.J.S.), AG032984 (A.J.S.), AG036535 (A.J.S.), AG10133 (A.J.S.), AG19771 (A.J.S.), LM010098 (J.H.M.), LM009012 (J.H.M.) and AI59694 (J.H.M.). The authors would also like to thank the two anonymous reviewers for their thoughtful comments and suggestions to improve the manuscript.

References

1 Hindorff, L.A. *et al.* (2011) *A Catalog of Published Genome-wide Association Studies*, National Human Genome Research Institute

2 Schadt, E.E. (2009) Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223

3 Hirschhorn, J.N. (2009) Genomewide association studies – illuminating biologic pathways. *New Engl. J. Med.* 360, 1699–1701

4 Menashe, I. *et al.* (2010) Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer Res.* 70, 4453–4459

5 Wang, K. *et al.* (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843–854

6 Zhong, H. *et al.* (2010) Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am. J. Hum. Genet.* 86, 581–591

7 Abatangelo, L. *et al.* (2009) Comparative study of gene set enrichment methods. *BMC Bioinform.* 10, 275

8 Telander, D.G. (2011) Inflammation and age-related macular degeneration (AMD). *Semin. Ophthalmol.* 26, 192–197

9 Eleftherohorinou, H. *et al.* (2009) Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE* 4, e8068

10 Chen, X. *et al.* (2010) Pathway-based analysis for genome-wide association studies using supervised principal components. *Genet. Epidemiol.* 34, 716–724

11 Zhao, J. *et al.* (2011) Pathway-based analysis using reduced gene subsets in genome-wide association studies. *BMC Bioinform.* 12, 17

12 Chen, M. *et al.* (2011) Incorporating biological pathways via a markov random field model in genome-wide association studies. *PLoS Genet.* 7, e1001353

13 Wang, K. *et al.* (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* 81, 1278–1283

14 Askland, K. *et al.* (2011) Ion channels and schizophrenia: a gene set-based analytic approach to GWAS data for biological hypothesis testing. *Hum. Genet.* 1–19

15 Lascorz, J. *et al.* (2011) Consensus pathways implicated in prognosis of colorectal cancer identified through systematic enrichment analysis of gene expression profiling studies. *PLoS ONE* 6, e18867

16 Peng, G. *et al.* (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* 18, 111–117

17 Bader, G.D. *et al.* (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.* 34, D504–D506

18 Camon, E.B. *et al.* (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinform.* 6 (Suppl. 1), S17

19 Swaminathan, S. *et al.* (2012) Amyloid pathway-based candidate gene analysis of [(11)C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. *Brain Imaging Behav.* 6, 1–15

20 Sloan, C.D. *et al.* (2010) Genetic pathway-based hierarchical clustering analysis of older adults with cognitive complaints, amnestic mild cognitive impairment using clinical, neuroimaging phenotypes. *Am. J. Med. Genet. Part B: Neuropsychiat. Genet.* 153B, 1060–1069

21 Zhang, M. *et al.* (2011) Pathway analysis for genome-wide association study of basal cell carcinoma of the skin. *PLoS ONE* 6, e22760

22 Elbers, C.C. *et al.* (2009) Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genet. Epidemiol.* 33, 419–431

23 Cantor, R.M. *et al.* (2010) Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* 86, 6–22

24 Holmans, P. (2010) Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.* 72, 141–179

25 Perry, J.R. *et al.* (2009) Interrogating type 2 diabetes genome-wide association data using a biological pathway-based approach. *Diabetes* 58, 1463–1467

26 Ballard, D. *et al.* (2010) Pathway analysis comparison using Crohn's disease genome wide association studies. *BMC Med. Genomics* 3, 25

27 Yon Rhee, S. *et al.* (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* 9, 509–515

28 Higareda-Almaraz, J.C. *et al.* (2011) Proteomic patterns of cervical cancer cell lines, a network perspective. *BMC Syst. Biol.* 5, 96

29 Zhang, S. *et al.* (2010) GO-Bayes: gene ontology-based overrepresentation analysis using a Bayesian approach. *Bioinformatics* 26, 905–911

30 Alterovitz, G. *et al.* (2010) Ontology engineering. *Nat. Biotech.* 28, 128–130

31 Merico, D. *et al.* (2010) Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* 5, e13984

32 Yaspan, B. *et al.* (2011) Genetic analysis of biological pathway data through genomic randomization. *Hum. Genet.* 129, 563–571

33 Askland, K. *et al.* (2009) 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

34 Sawcer, S. *et al.* (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* 476, 214–219

35 Kapranov, P. *et al.* (2007) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* 8, 413–423

36 Portin, P. (2009) The elusive concept of the gene. *Hereditas* 146, 112–117

37 Edwards, Y.J. *et al.* (2011) Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. *PLoS ONE* 6, e16917

38 Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11, 499–511

39 Holmans, P. *et al.* (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.* 85, 13–24

40 O'Dushlaine, C. *et al.* (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics* 25, 2762–2763

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

42 Liu, J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* 87, 139–145

43 Huang, H. *et al.* (2011) Gene-based tests of association. *PLoS Genet.* 7, e1002177

44 Gaudet, M.M. *et al.* (2010) Common genetic variants and modification of penetrance of BRCA2–associated breast cancer. *PLoS Genet.* 6, e1001183

45 Goeman, J.J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23, 980–987

46 Fridley, B.L. *et al.* (2010) Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS ONE* 5, e12693

47 Moskvina, V. *et al.* (2009) Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol. Psychiatry* 14, 252–260

48 Lambert, J.C. *et al.* (2010) Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J. Alzheimers Dis.* 20, 1107–1118

49 Wang, K. *et al.* (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am. J. Hum. Genet.* 84, 399–405

50 Holden, M. *et al.* (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24, 2784–2785

51 Tintle, N. *et al.* (2009) Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC Proc.* 3, S96

52 Nam, D. *et al.* (2010) GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res.* 38, W749–W754

53 Wang, L. *et al.* (2011) Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* 98, 1–8

54 Hung, J-H. *et al.* (2011) Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief. Bioinform.* DOI: 10.1093/bib/bbr049

55 Kim, S-Y. and Volsky, D. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinform.* 6, 144

56 Berry, M.P.R. *et al.* (2010) An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466, 973–977

57 Baranzini, S.E. *et al.* (2010) Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis. *Brain* 133, 2603–2611

58 Lee, H.M. *et al.* (2011) Abnormal networks of immune response-related molecules in bone marrow cells from patients with rheumatoid arthritis as revealed by DNA microarray analysis. *Arthritis Res. Ther.* 13, R89

59 Pang, H. *et al.* (2011) Pathway-based identification of SNPs predictive of survival. *Eur. J. Hum. Genet.* 19, 704–709

60 Gilman, S.R. *et al.* (2011) Rare *de novo* variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70, 898–907

61 Yang, H-C. and Chen, C-W. (2011) Region-based and pathway-based QTL mapping using a p–value combination method. *BMC Proc.* 5, S43

62 McLean, C.Y. *et al.* (2010) GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotech.* 28, 495–501

63 Hong, M-G. *et al.* (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* 126, 289–301

64 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B: Methodol.* 57, 289–300

65 Sabatti, C. (2007) Avoiding false discoveries in association studies. In *Methods in Molecular Biology: Linkage Disequilibrium and Association Mapping* (Collins, A.R., ed.), pp. 195–211, Humana Press

66 Luo, L. *et al.* (2010) Genome-wide gene and pathway analysis. *Eur. J. Hum. Genet.* 18, 1045–1053

67 Bertram, L. *et al.* (2007) Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* 39, 17–23

68 Zhu, J. *et al.* (2009) The UCSC Cancer Genomics Browser. *Nat. Methods* 6, 239–240

69 Brown, S.D.M. *et al.* (2009) The functional annotation of mammalian genomes: the challenge of phenotyping. *Annu. Rev. Genet.* 43, 305–333

70 Hu, T. *et al.* (2011) Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinform.* 12, 364

71 Cowper-Sal·lari, R. *et al.* (2011) Layers of epistasis: genome-wide regulatory networks and network approaches to genome-wide association studies. *Wiley Interdisciplinary Rev. Syst. Biol. Med.* 3, 513–526

72 McKinney, B.A. and Pajewski, N.M. (2011) Six degrees of epistasis: statistical network models for GWAS. *Front. Genet.* 2, 109

73 Bandyopadhyay, S. *et al.* (2010) Rewiring of genetic networks in response to DNA damage. *Science* 330, 1385–1389

74 Sun, J. *et al.* (2011) Application of systems biology approach identifies and validates GRB2 as a risk gene for schizophrenia in the Irish Case Control Study of Schizophrenia (ICCSS) sample. *Schizophr. Res.* 125, 201–208

75 Edvardsson, K. *et al.* (2011) Estrogen receptor beta induces antiinflammatory and antitumorigenic networks in colon cancer cells. *Mol. Endocrinol.* 25, 969–979

76 Gui, H. *et al.* (2011) Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's Disease dataset. *BMC Res. Notes* 4, 386

77 Gorlov, I.P. *et al.* (2009) GWAS meets microarray: are the results of genome-wide association studies and gene-expression profiling consistent? Prostate cancer as an example. *PLoS ONE* 4, e6511

78 Myers, A.J. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.* 39, 1494–1499

79 Nicolae, D.L. *et al.* (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6, e1000888

80 Liang, W.S. *et al.* (2008) Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4441–4446

81 Le-Niculescu, H. *et al.* (2009) Convergent functional genomics of genome-wide association data for bipolar disorder: comprehensive identification of candidate genes, pathways and mechanisms. *Am. J. Med. Genet. B: Neuropsychiatr. Genet.* 150B, 155–181

82 Ala-Korpela, M. *et al.* (2011) Genome-wide association studies and systems biology: together at last. *Trends Genet.* 27, 493–498

83 Penrod, N.M. *et al.* (2011) Systems genetics for drug target discovery. *Trends Pharmacol. Sci.* 32, 623–630

84 Jarvik, J. and Botstein, D. (1973) A genetic method for determining the order of events in a biological pathway. *Proc. Natl. Acad. Sci. U.S.A.* 70, 2046–2050

85 Marchion, D.C. *et al.* (2011) BAD phosphorylation determines ovarian cancer chemo-sensitivity and patient survival. *Clin. Cancer Res.* 17, 6356–6366

86 Newman, M.E.J. (2010) *Networks: An Introduction*, Oxford University Press

87 Price, N.D. and Shmulevich, I. (2007) Biochemical and statistical network models for systems biology. *Curr. Opin. Biotechnol.* 18, 365–370

88 Ghosh, S. *et al.* (2011) Software for systems biology: from tools to integrated platforms. *Nat. Rev. Genet.* 12, 821–832

89 Thomas, S. and Bonchev, D. (2010) A survey of current software for network analysis in molecular biology. *Hum. Genomics* 4, 353–360

90 Chen, H. and Sharp, B. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinform.* 5, 147

91 Medina, I. *et al.* (2009) 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

92 Raychaudhuri, S. *et al.* (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 5, e1000534

93 Chen, L.S. *et al.* (2010) Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* 86, 860–871

94 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550

95 Segrè, A.V. *et al.* (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* 6, e1001058