Published in final edited form as:
Curr Protoc Hum Genet. 2019 January ; 100(1): e79. doi:10.1002/cphg.79.

Strategies for Pathway Analysis using GWAS and WGS Data

Marquitta J. White[1], **Brian L. Yaspan**[2], **Olivia J. Veatch**[3], **Pagé Goddard**[1], **Oona S. Risse-Adams**[1], and **Maria G. Contreras**[1,4]

[1]Department of Medicine, Lung Biology Center, University of California San Francisco, San Francisco, California

[2]Genentech Inc., 1 DNA Way, South San Francisco, California

[3]Center for Sleep and Circadian Neurobiology, Department of Medicine, University of Pennsylvania

[4]MARC, San Francisco State University, San Francisco, CA, USA

Abstract

Single allele study designs, commonly used in genome-wide association studies (GWAS) as well as the more recently developed whole genome sequencing (WGS) studies, are a standard approach for investigating the relationship of common variation within the human genome to a given phenotype of interest. However, single-allele association results published for many GWAS studies represent only the tip of the iceberg for the information that can be extracted from these datasets. The primary analysis strategy for GWAS entails association analysis in which only the single nucleotide polymorphisms (SNPs) with the strongest *p-values* are declared statistically significant due to issues arising from multiple testing and type I errors. Factors such as locus heterogeneity, epistasis, and multiple genes conferring small effects contribute to the complexity of the genetic models underlying phenotype expression. Thus, many biologically meaningful associations having lower effect sizes at individual genes are overlooked, making it difficult to separate true associations from a sea of false-positive associations. Organizing these individual SNPs into biologically meaningful groups to look at the overall effects of minor perturbations to genes and pathways is desirable. This pathway-based approach provides researchers with insight into the functional foundations of the phenotype being studied and allows testing of various genetic scenarios.

Keywords

GWAS; genome-wide association; pathway analysis; genetic epidemiology

INTRODUCTION

This unit focuses on available pathway databases, software analysis packages, and points to consider when performing pathway analysis from genome-wide association studies (GWAS) or whole genome sequencing (WGS) data. While some of the rationale and hypotheses behind pathway analysis from high-throughput genomic data (GWAS/WGS) are described, the emphasis is on the logistic issues of choosing a database and algorithm. Pathway analysis algorithms are optimized for different study designs, thereby it is essential to understand and select the statistically appropriate model. A targeted study, such as a candidate gene study, is not amenable to most of the pathway analysis methods outlined; the methods presented in this chapter require genome-wide data unless otherwise stated. Genotyping data from custom platforms that probe for candidate single nucleotide polymorphisms (SNPs) are not appropriate for pathway-based GWAS/WGS analysis. This unit assumes that the GWAS/WGS dataset under study has been generated and proper quality control methods have been applied. It should be noted that while WGS data may generate information for rare variants, minor allele frequency ($MAF < 0.01$), this chapter will focus solely on assessing common variation.

KEY CONCEPTS

Biological Pathways

Biological pathways capture our understanding of biological processes and can be thought of as groups of genes that are functionally related. These pathways represent a series of events leading to an ultimate functional result, be it assembly of a new molecule, completion of a necessary cellular process, or turning genes on and off at specific developmental time points. Each gene in the pathway encodes a molecule that functions to carry out a biochemical reaction. Consequently, pathways are often defined based on either the cellular organelle where the steps are carried out or the type of process being accomplished (i.e., metabolic, signaling, gene regulation). Generally, the initial gene in the pathway encodes the molecular product necessary for successful completion of all subsequent steps in the pathway. The final step in a pathway is often the reaction that generates the molecular product necessary to conclude the process of interest. Phenomena such as feedback inhibition and compensatory reactions often occur, contributing to the complexity of pathway definitions. Also, various reactions in many uniquely defined pathways generate by-products necessary to the function of other pathways. This means that dysfunction of a gene in one pathway has the potential to affect the function of numerous biological processes. For the purpose of this unit, a pathway refers to a set of biologically related genes found in a collated database.

Gene Sets

Gene sets are collections of genes having some functional or evolutionary relationship other than contributing to a shared biological pathway. This definition could include a group of genes that are in the same phylogenetic class, subclass, or family. A gene set could also be defined as a group of genes located tandemly on the same chromosome. For example, genes comprising one of the Hox clusters could be considered a gene set since each cluster is located on the same chromosomal region. All Hox genes also have a highly conserved homeobox sequence that may perform similar functions during organism development, and are thought to have evolved from the same ancestral gene. Additionally, a gene set could also include groups of genes with evidence for epistatic interactions. For the purpose of this unit, a gene set refers to a personalized list of genes collated by an individual investigator.

Pathway Database Curation

Database curation is the process of evaluating and selecting items to be entered into the database. Pathway databases are curated in two ways: manually or computationally. Manual curation relies on manual transfer of knowledge from scientific publications. This process allows expert review of every publication related to a pathway prior to submission in the database. Most databases with manual curation procedures employ a scientific review boards to assess and discuss each pathway before adding it to the database. Usually, curators perform periodic reviews of all annotations for accuracy and completeness, updating as necessary. This process involves adding new annotations to reflect advances in knowledge and removing any annotations that are no longer supported by the literature. Expert curation is a powerful tool for producing and maintaining biologically relevant and current databases; however, users should be aware of the update intervals and criteria when using a manually curated database.

Computationally curated database submissions are not reviewed by scientific curators. Submissions are made by a variety of computational procedures, such as sequence similarity methods and keyword mapping files. Computational curation is less expensive and much faster than manual curation. However, pathways in databases that are solely computationally curated may be less scientifically accurate since computational predictions do not require experimental work other than the genome sequence. Computational curation procedures are usually repeated on a regular basis to keep up with improvements to computational methods and changes to genome sequence annotation (Costanzo et al., 2011). As manual and computational curations capture unique aspects, databases that combine both methods seem to provide the most accurate, current, and complete functional annotation of the genome.

Data Access

Most pathway databases provide the user with information in the form of downloadable files. Data in these files can be written in various computational languages. Knowledge of the language used is important since proper data format is essential to accurate evaluation of pathways in the analysis program being used. Three computational languages are commonly used to represent biological pathways and gene sets at the molecular and cellular levels. These standards are Biological Pathway Exchange (BioPAX), Proteomics Standard Initiative-Molecular Interactions (PSI-MI), and Systems Biology Markup Language (SBML). BioPAX was developed to provide well-defined semantics for pathway representation, allowing pathway databases and software to interact more efficiently (Demir et al., 2010). PSI-MI is a community standard data model for the representation and exchange of protein interaction data. This data model was jointly developed by members of the Proteomics Standards Initiative (PSI), a work group of the Human Proteome Organization (HUPO), and is supported by many major protein interaction data providers (Hermjakob et al., 2004; Deutsch et al., 2017). SBML is a free, open, XML-based format for representing biochemical reaction networks. SBML is a software-independent language for describing models common to research in many areas of computational biology, including cell signaling pathways, metabolic pathways, and gene regulation (Hucka et al., 2003; Hermjakob et al., 2004).

Genome-Wide Pathway Analysis (GWPA)

GWPA is an agnostic, approach that harnesses the wealth of information from GWAS or WGS data to identify the additive effects of single variants aggregating in particular gene sets or pathways. The general concept behind pathway analysis of GWAS data can be thought of as data reduction and aggregation. First, the data on millions of individual sites around the genome are collected in the form of SNP data. These SNPs are then analyzed for allele frequency differences between case and control groups (as one example) and an assessment of statistical inference is obtained. The SNPs are placed within genes by genomic position, and the genes are placed into pathways or gene sets based on the information provided by the selected database(s). Thus, individual SNPs are now collections of SNPs within the framework of a group of genes. Most of the algorithms then look for a preponderance of single-allele *p-values* meeting a specific criterion within the pathway, or gene set, relative to what one would expect to see by chance (Fig. 1).

Candidate Pathway Analysis

Candidate pathway analysis is a hypothesis-driven approach to pathway-level investigation. Pathways, or gene-sets, of interest are preselected based on prior knowledge or investigation, and association testing is performed on the preselected loci of interest instead of the full genome. Rather than scanning an entire database of pathway information as with GWPA, which can be computationally intensive and requires stringent multiple-testing corrections, candidate pathway analysis is less demanding and therefore lends itself more readily to large data sets such as imputed and WGS data.

STRATEGIC APPROACH & ALGORITHM SELECTION

There is an ever-growing list of pathway analysis algorithms. While they ostensibly claim to achieve the same objective, there are well-defined differences between these algorithms. Algorithm selection should be guided by underlying methodology and applicability to the specific question. Selection will depend heavily on study design and the type of data that is available. This section discusses the primary deciding factors in designing a robust pathway analysis study and selecting the most appropriate algorithm: research question, input data type, database selection, and program options.

Research Question

The research question itself should indicate if a candidate or agnostic pathway analysis is most appropriate. A hypothesis-driven question, such as “are oxidative stress pathways genetically different in individuals with cancer?” would benefit from a candidate pathway study design, while the broader question, “what pathways are these genetic associations aggregating in?” would benefit more from an agnostic approach. Additionally, investigator-curated gene sets can be a useful intermediate. For instance, if a researcher wishes to attempt a gene-level replication of previously implicated variants, they may wish to perform gene-set analysis with their GWAS results and a manually curated list of previously associated genes. Establishing the overall study design best suited to the research question, in conjunction with the available input data, is essential for selecting both the biological database(s) of interest and the analysis algorithm.

Input Data Considerations

Pathway analyses can be performed with multiple types of input data, including raw genotype data from individual samples, SNP-level *p-values* such as those from a GWAS, and gene-level *p-values* (for example from RNA-seq differential expression analysis). The selection of input data should be based not only on what data are available, but also on what criteria are most important in the downstream analysis. Algorithms that utilize genotype data can correct for linkage disequilibrium (LD) but often cannot correct for covariates in the analysis.

Pre-computed *p-values*, on the other hand, can certainly be adjusted for covariates and population structure, although many pathway analysis algorithms that use pre-computed SNP *p-values* require post-hoc correction for LD. It is especially important to consider the LD correction options when working with populations of non-European descent; LD patterns vary by race/ethnicity and not accounting for this variation could adversely affect SNP-gene mapping (Wall & Pritchard., 2003) Most algorithms have been designed and optimized for analysis of European-descent populations, but some allow the investigator to specify population-specific reference panels from available databases, such as the 1000 Genomes Project (1KGP) or, better yet, include a user-generated LD file for more accurate mapping (The 1000 Genomes Project Consortium., 2015). Additionally, many SNP-based pathway analysis algorithms rely on curated lists of reference SNPS (refSNPs) with accompanying refSNP or “rs” identification (rsID) numbers from the National Center for Biotechnology Information (NCBI) database for annotation and mapping. Consequently, research questions that include novel loci not found within the NCBI database, or other public databases, should consider algorithms, such as PARIS, that can accept position files in lieu of an rsID list.

Pathway Database Considerations

Appropriate database selection should be heavily informed by the research question. For instance, candidate pathway analyses can manage more detailed pathway annotations and may benefit from specialized databases that include plausible recently identified attributes, while GWPA studies may find canonical information more useful and computationally manageable. Additionally, some pathway analysis programs allow for incorporation of user-defined gene sets. Numerous pathway databases are freely available (Ooi et al., 2010). A selection of commonly used databases appears in Table 1. These databases serve to group genes based on biological function while providing information on defined gene networks in humans. These databases may provide two main pathway database structures: discrete pathway maps and functional hierarchies (ontologies).

As for all of the options presented in this unit, the choice of protocol is dependent on your study design and dataset. For example, if a gene of interest is not annotated in a certain database, it would be best to choose another database where it is annotated. The Pathguide resource (http://www.pathguide.org) helps identify the database most suited to the underlying biology of your phenotype of interest. This resource provides the user with information for each listed database regarding access, types of pathways defined, computational language used to represent the information, and links to the original research papers describing the database (Bader et al., 2006).

Algorithm Considerations

At the minimum, pathway analysis algorithms may be thought of as data aggregators. To move from individual SNPs to pathways, there must be a way of systematically grouping the SNPs. Pathway databases (or personally selected gene sets) act as the framework. The algorithm assigns the SNPs to genes and the genes to the framework, and then determines whether the statistically significant signals are over-represented in the set. A selection of commonly used pathway analysis algorithms is presented in Table 2. **However, when calculating over-representation, the philosophy behind pathway-based analysis is subject to multiple biases; gene size, pathway size, density of SNP coverage, and linkage disequilibrium (LD) patterns are all factors that must be considered and appropriately addressed.** At a standard type I error rate of $\alpha = 0.05$, each SNP tested has a 5% chance of being associated with a disease by chance alone. Testing more SNPs therefore increases the number of false-positive associations. Thus, genes with more SNPs tested have an *a priori* increased likelihood of having a greater number of SNPs associated by chance. Larger pathways with more genes similarly increase this potential bias. Furthermore, any type I error will likely extend across all SNPs that are in LD with each other. Fortunately, many of the available algorithms employ methodology to help reduce these biases.

SNP to Gene mapping—Once the input data is determined, the pathway analysis algorithm must map the input to genes and gene-sets. The following is true regardless of the approach (GWPA or Candidate) selected. In the simplest approach, input SNPs or positions are mapped to genes (defined in the software database or by user input) by identifying variants whose base pair positions fall within the reported gene boundaries. However, linkage disequilibrium (LD) differences can vary significantly between genes, and cis regions outside the transcript start and end sites can play significant roles in gene regulation. It has been shown that 90% of SNPs affecting expression quantitative trail loci were observed within 15kb from the 5’ and 3’ gene boundary (Pickrell et al., 2010). Thus, restricting mapping protocols to include only SNPs that fall within the strict boundaries of the gene (5’ – 3’ region) may exclude regulatory SNPs with biologically relevant information.

To account for LD, many programs will sort SNPs into high-LD regions, or blocks, before scanning for positional overlap with gene regions. Thus, SNPs that are not within transcriptional boundaries but are in high LD with a genic SNP will be considered in that gene’s enrichment score. If genotype data is accessible, high LD regions can be calculated from the study cohort by the algorithm (if raw genotype data is accepted), prior to analysis in an external program like PLINK and passed through the algorithm as a specified input (if custom LD region files are accepted). If genotype data is not available, many programs rely on HapMap or 1000 Genome reference panels to estimate LD blocks; it is imperative that the selected LD reference panel be representative of the study cohort. Gene boundaries can typically be manipulated manually by the user to include additional upstream and downstream regions, if desired, to include potential regulatory variants.

It is important to note, however, that the assignment of disease-associated variants to their closest mapped genes may not be the most accurate approach for predicting functional relationships. Evidence has suggested that SNPs may regulate gene expression over broad genomic regions, resulting in SNPs whose gene-level impacts involve distal genes rather than their closest mapped or encompassing genes (Heintzman & Ren, 2009). For instance, intronic variants in the *FTO* locus demonstrate strong association with obesity; however, chromatin-capture sequencing (Hi-C) and expression Quantitative Trait Loci (eQTL) data later revealed that these variants did not regulate the expression of *FTO*, but of *IRX3/5*, a gene about half a mega base away (Herman & Rosen, 2015). IRX3/5 was later shown to play a role in white and beige adipocyte differentiation. Thus, eQTL and Hi-C mapping can be powerful approaches for estimating gene-level *p-values* based on putative functional relationships with disease-associated SNPs. A recently released web based platform, FUMA, incorporates positional, Hi-C, and eQTL data as a part of its SNP-to-gene mapping approach (Watanabe et al., 2017). FUMA also incorporates the pathway analysis tool, MAGMA (de Leeuw et al., 2015); enabling it to combine its integrative SNP-to-gene annotation scheme with a powerful gene set analysis method to produce gene, pathway, and tissue enrichment results. Of the algorithms listed in Table 1.20.2, only GSA-SNP allows users to upload a custom SNP-to-gene mapping file; this feature enables GSA-SNP (Nam et al., 2010) to incorporate Hi-C and eQTL data through pre-processing using programs such as FUMA, thereby increasing the likelihood that that gene/pathway analysis will yield functionally relevant results. While the other algorithms mentioned in Table 1.20.2 still hold merit as capable pathway analysis algorithms, all would benefit from software updates that would allow for the inclusion of Hi-C and eQTL data as a part of the SNP mapping procedures.

Gene p-value calculation and corrections—Previously, several common pathway analysis algorithms employed some form of *sentinel SNP* approach, in which the most significant SNP mapped to a given gene is selected as the representative p-value for the gene, to assign gene-level significance. Some programs, like ICSNPathway (Zhang et al., 2011) couple this approach with functional annotation to further prioritize SNPs, while others, like GSA-SNP2 (Nam et al., 2010; Yoon et al., 2018) use the *k*th best SNP as the representative *p*-value to limit the false positive effects of spurious GWAS hits. However, selecting only one SNP's *p*-value to represent each gene may not capture the additive effects of multiple SNPs within a gene with smaller effect sizes.

Algorithms like PARIS avoid the potential confounding of spurious GWAS associations with a *threshold overrepresentation* approach; a blanket *p*-value threshold for moderate association (*p*-value < 0.05 by default) is paired with permutation testing to determine the likelihood that the number of associated features mapped to the gene or gene set by chance. Such a method effectively tests for the additive effects of moderately associated SNPs on phenotype susceptibility without bias from randomly significant hits. However, in the case of true causative GWAS hits with highly significant *p-values*, it may be more powerful to use an approach that incorporates relative SNP *p-values*.

Finally, algorithms such as VEGAS2 (Liu et al., 2010; Mishra & MacGregor, 2015; Mishra & MacGregor, 2017) and INRICH (Lee et al., 2012) employ a *ranked enrichment* test in which a statistical test is performed to calculate the enrichment of a gene or gene set's SNPs at the significant end of a list of ranked SNP *p-values*. This approach is powerful because it can account for all SNPs, reducing the effects of false positive GWAS results in favor of enrichment of moderate SNPs, without negating potential importance of extremely significant vs. moderately significant SNPs. However, ranking and testing enrichment using all the SNPs from a GWAS can be computationally intensive, so these tests are often used with a threshold (e.g. top 5% or top 1% of SNPs) to reduce computation time.

Genes can vary by size, SNP density, and LD patterns. The latter is addressed during SNP mapping but can also affect *p*-value calculations when the number of significant SNPs/gene is of interest. Gene size and SNP density can skew the likelihood of significant SNPs occurring in a given gene by chance and should be addressed during gene *p*-value calculation. A large gene has a higher probability of containing significant SNPs than genes that span fewer base pair; likewise, a gene with high genotyping density (i.e. genotype arrays contain a higher proportion of SNPs for one gene compared to others) has a higher chance of containing significant SNPs. To account for variation in genomic structure is yet another aspect of pathway based analysis that must be considered when both performing analysis and interpreting the results.

Pathway p-value calculation and corrections—Pathway based analysis (PBA) algorithms essentially test the enrichment or likelihood of gene *p-values* in gene lists that represent pathways or curated gene sets ("pathway" and "gene set" will be used interchangeably). Again, there are many mathematical models employed to evaluate the significance of a pathway given the distribution of its constituent gene *p-values*. Algorithms can adopt a competitive or self-contained approach. *Competitive* methods compare the gene statistics for the constituents of a given pathway to those of the rest of the genome, evaluating if the genes in the set have the same magnitude of trait association as the rest of the genome. *Self-contained* methods directly test gene set association with the trait of interest, often utilizing permutations, or other randomizations, to correct for pathway size and structure, without depending on the rest of the genome. Competitive methods effectively test the "enrichment" of associated genes within a gene set compared to the rest of the genome while self-contained approaches test for the "existence" of associated genes in each pathway and calculate significance based on the likelihood of a pathway containing that many significant constituents. Self-contained approaches tend to be more sensitive and thus more powerful for finding novel pathways, and are better powered when SNPs in multiple gene sets are associated with the trait. However, genes often have multiple functions, such that the mere presence of an associated gene in a pathway does not necessarily confer a pathway-level aberration, which is accounted for in the competitive enrichment model. Thus, competitive and self-contained pathway analysis algorithms serve as complementary approaches.

Like genes, pathway size and annotation density can impact the likelihood of association by chance; a large pathway or a heavily annotated pathway is more likely to be enriched with significant genes by chance than a small or under-documented pathway that may actually be associated. In addition to database selection, pathway analysis algorithms leverage a variety of techniques to correct for structural confounders within and between pathways and gene sets.

Summary: Algorithm Considerations—In this section, we discussed common methods employed in pathway-based analysis as well as the potential biases and proposed solutions to keep in mind when selecting a method. In brief, several common gene-level calculations can rely on *sentinel SNPs*, *threshold overrepresentation* testing, *ranked enrichment* of SNP *p-values*, or some combination while pathway-level assessment can be *self-contained* or *competitive* depending on the study design's null hypothesis. When addressing important corrections, population structure and covariates are easiest to correct for in GWAS or initial testing prior to PBA. Accounting for LD between SNPs and genes is performed during the mapping stages of PBA and the permutation/ randomization stages, where applicable. When working with non-European populations, read the algorithm options carefully to ensure that you can either provide your own LD regions or, at the very least, select a reference population that is representative of the study population. Finally, gene and pathway size and density can skew association results and must be corrected for with permutation testing for some statistical modification.

In summary, there are several biases involved with performing pathway-based analysis from GWAS data. Many of these biases are size-related (the number of SNPs tested within a gene boundary, the number of genes in a pathway, etc.). In most cases, the algorithms themselves provide unique and robust ways to control for these biases.

COMMENTARY

After Running the Algorithm

After the algorithm and database are chosen and the data processed, assuming you are running a test of more than one pathway or gene set, you will have a list of pathways significantly associated with your phenotype. At this point, it is important to point out that, while many algorithms correct for multiple testing in the original genome-wide SNP data, they may not correct for testing of more than one pathway or gene set. Often, several pathways will be seen as significantly associated at the chosen *p-value* significance threshold. Thus, it is helpful to outline steps that simplify the results into a manageable list of pathways. This section describes examples of possible ways to sort through a list of significantly associated pathways. A simple flowchart for one possible method appears in Figure 2.

First, as we are interested in a *pathway* and not an individual gene, it is recommended to remove any pathways where the signal is driven by a single gene. There are many ways to investigate this. As an example, the SNP Ratio Test (SRT) can be used to test each associated pathway in a step-method that selectively deleted one gene at a time (Anney et al., 2011). If the significance of the whole pathway dropped due to the exclusion of one gene, than that one gene was driving the signal for the entire pathway, thus reducing interest in that pathway. Theoretically, this method could be applied to any algorithm that allows the use of custom gene sets. As another example, PARIS (Yaspan et al., 2011; Butkiewicz et al., 2016) contains functionality for assessing the contribution of SNPs within each individual gene as they relate to the significance of a pathway as a whole.

Many different biological pathways contain similar sets of genes. While this may lead to a large list of associated pathways, it is possible to use this to your advantage to identify subsets of genes within those pathways that truly drive the signal. Consider a scenario where three pathways are associated. The largest pathway has 100 genes and a *p-value* of 0.04, the second largest pathway has 50 genes and a *p-value* of 0.01, and the smallest pathway has 10 genes and a *p-value* of 0.001. Further inspection of the genes in these pathways identifies five genes common to all three pathways. In this scenario, it would be recommended to check the *p-values* of the SNPs within these genes for significance, as it is possible that it is not the three pathways that are of interest, but rather the five genes that are common between them.

For hypothesis-independent investigations, it is recommended to consider prioritizing biologically relevant pathways when sorting through a substantial list of associated pathways; leveraging prior biological knowledge when sifting through the list of associated pathways and can be a very powerful approach to identifying interesting results. A final recommendation for hypothesis-independent GWPA is to utilize more than one algorithm and prioritize pathways and gene sets identified by both approaches. Dual-pronged pathway analysis approaches that provide diverse, but convergent, evidence of an association may prove effective for prioritizing putative pathways for further functional investigation.

Visual Representation

The inclusive concept of pathway analysis lends itself to several types of diagram or network visualization (e.g., KEGG database). Once a pathway of interest is identified, an overlay of information on the visual framework can be very illuminating. Visual overlay can encompass many types of information ranging from sentinel SNP *p-values* or gene mutational load (e.g. ratio of significant SNPs to total SNPs) to other functional information. Figure 3 shows an example of a pathway diagram with genes shaded by sentinel SNP for rapid visual assessment of patterns between genes; for instance, genes that interact directly would be obvious candidates for gene-gene interaction analysis to examine possible epistatic effects.

It may also be helpful to visualize gene-sets and pathways with alternate paradigms to investigate additional aspects of pathway enrichment output. For instance, *treeplots* and *dotplots* can efficiently summarize information about several pathways simultaneously. In a treeplot, pathway groups, sized relative to pathway size or pathway significance, can be further subdivided by their gene content with each gene box colored by its *p-values* for visualization of the gene ratio (# significant genes to # total genes) for each pathway. A dotplot provides a visual summary of pathway-level information including multiple pathways (y-axis), gene ratio (x-axis), pathway size (dot size) and pathway significance (dot color), which can be especially helpful for identifying relationships between pathway size and summary statistics or gene ratio and significance. Evaluation of pathway overlap can be visualized effectively with a *heatplot* in which pathways are clustered according to the similarity of their genic content and genes are color-coded by their *p-values* within the pathways. Finally, a useful and intuitive visualization technique when extrapolating gene and pathway enrichment from GWAS data is the *gene-level Manhattan plot*; each dot on a chromosome represents a gene rather than a SNP, ordered by transcription start site location, -log(p_{gene}) on the y-axis and pathway genes represented in a contrasting color scheme to non-pathway genes (e.g. i-GSEA4GWASv2 output). Alternative visualizations can summarize multiple pathways at once, effectively represent the relationship with between pathway size and significance, and help the investigator identify potential confounding gene sets driving the significance of multiple pathways.

Summary

In this unit, we have discussed the basic tenets and principles behind pathway analysis from GWAS data. We have examined several different algorithms and pathway databases through the logistical lens, including range of methodology and approaches. We have also provided some commentary on which algorithm, database, and design could be appropriate given a particular dataset or analysis plan. Finally, we have outlined some beginning steps toward making sense of the results after analysis. When used, properly, pathway analysis of high throughput genomic data (GWAS/WGS) is a powerful technique for expanding the utility of a GWAS dataset, and can be performed to both answer and generate lines of scientific theory.

Acknowledgement

This work was supported in part by the Sandler Family Foundation, the National Institutes of Health R01HL117004, R01HL128439, R01HL135156, 1X01HL134589, R01HL141992, K01-LM-012870–01, K01HL140218, T34-GM08574, the National Institute of Health and Environmental Health Sciences R01ES015794, R21ES24844, the National Institute on Minority Health and Health Disparities P60MD006902 RL5GM118984.

LITERATURE CITED

Anney RJ, Kenny EM, O'Dushlaine C, Yaspan BL, Parkhomenka E, Buxbaum JD, Sutcliffe J, Gill M, Gallagher L, Bailey AJ, Fernandez BA, Szatmari P, Scherer SW, Patterson A, Marshall CR, Pinto D, Vincent JB, Fombonne E, Betancur C, Delorme R, Leboyer M, Bourgeron T, Mantoulan C, Roge B, Tauber M, Freitag CM, Poustka F, Duketis E, Klauck SM, Poustka A, Papanikolaou K, Tsiantis J, Gallagher L, Gill M, Anney R, Bolshakova N, Brennan S, Hughes G, McGrath J, Merikangas A, Ennis S, Green A, Casey JP, Conroy JM, Regan R, Shah N, Maestrini E, Bacchelli E, Minopoli F, Stoppioni V, Battaglia A, Igliozzi R, Parrini B, Tancredi R, Oliveira G, Almeida J, Duque F, Vicente A, Correia C, Magalhaes TR, Gillberg C, Nygren G, Jonge MD, Van EH, Vorstman JA, Wittemeyer K, Baird G, Bolton PF, Rutter ML, Green J, Lamb JA, Pickles A, Parr JR, Couteur AL, Berney T, McConachie H, Wallace S, Coutanche M, Foley S, White K, Monaco AP, Holt R, Farrar P, Pagnamenta AT, Mirza GK, Ragoussis J, Sousa I, Sykes N, Wing K, Hallmayer J, Cantor RM, Nelson SF, Geschwind DH, Abrahams BS, Volkmar F, Pericak-Vance MA, Cuccaro ML, Gilbert J, Cook EH, Guter SJ, Jacob S, Nurnberger JI, Jr, McDougle CJ, Posey DJ, Lord C, Corsello C, Hus V, Buxbaum JD, Kolevzon A, Soorya L, Parkhomenko E, Leventhal BL, Dawson G, Vieland VJ, Hakonarson H, Glessner JT, Kim C, Wang K, Schellenberg GD, Devlin B, Klei L, Minshew N, Sutcliffe JS, Haines JL, Lund SC, Thomson S, Yaspan BL, Coon H, Miller J, McMahon WM, Munson J, Estes A, and Wijsman EM 2011 Gene-ontology enrichment analysis in two independent family-based samples highlights biologically plausible processes for autism spectrum disorders. Eur. J. Hum. Genet 19(10)

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G 2000 Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet 25:25–29. [PubMed: 10802651]

Bader GD, Cary MP, and Sander C 2006 Pathguide: A pathway resource list. Nucleic Acids Res 34:D504–D506. [PubMed: 16381921]

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, and Eddy SR 2004 The Pfam protein families database. Nucleic Acids Res 32:D138–D141. [PubMed: 14681378]

Butkiewicz M, Cooke Bailey JN, Frase A, Dudek S, Yaspan BL, Ritchie MD, Pendergrass SA, Haines JL 2016 Pathway analysis by randomization incorporating structure-PARIS: an update. Bioinformatics 32(15):2361–3. [PubMed: 27153576]

Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, and Sander C 2011 Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res 39:D685–D690. [PubMed: 21071392]

Costanzo MC, Park J, Balakrishnan R, Cherry JM, and Hong EL 2011 Using computational predictions to improve literature-based Gene Ontology annotations: A feasibility study. Database 2011:bar004.

Dall'olio GM, Jassal B, Montanucci L, Gagneux P, Bertranpetit J, and Laayouni H 2011 The annotation of the asparagine N-linked glycosylation pathway in the Reactome Database. Glycobiology. 21(11):1395–400. [PubMed: 21199820]

Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, Schacherer F, Martinez-Flores I, Hu Z, Jimenez-Jacinto V, Joshi-Tope G, Kandasamy K, Lopez-Fuentes AC, Mi H, Pichler E, Rodchenkov I, Splendiani A, Tkachev S, Zucker J, Gopinath G, Rajasimha H, Ramakrishnan R, Shah I, Syed M, Anwar N, Babur O, Blinov M, Brauner E, Corwin D, Donaldson S, Gibbons F, Goldberg R, Hornbeck P, Luna A, Murray-Rust P, Neumann E, Reubenacker O, Samwald M, van IM, Wimalaratne S, Allen K, Braun B, Whirl-Carrillo M, Cheung KH, Dahlquist K, Finney A, Gillespie M, Glass E, Gong L, Haw R, Honig M, Hubaut O, Kane D, Krupa S, Kutmon M, Leonard J, Marks D, Merberg D, Petri V, Pico A, Ravenscroft D, Ren L, Shah N, Sunshine M, Tang R, Whaley R, Letovksy S, Buetow KH, Rzhetsky A, Schachter V, Sobral BS, Dogrusoz U, McWeeney S, Aladjem M, Birney E, Collado-Vides J, Goto S, Hucka M, Le NN, Maltsev N, Pandey A, Thomas P, Wingender E, Karp PD, Sander C, and Bader GD 2010 The BioPAX community standard for pathway data sharing. Nat. Biotechnol 28:935–942. [PubMed: 20829833]

Deutsch EW, Orchard S, Binz PA, Bittremieux W, Eisenacher M, Hermjakob H, Kawano S, Lam H, Mayer G, Menschaert G, Perez-Riverol Y, Salek RM, Tabb DL, Tenzer S, Vizcaíno JA, Walzer M, Jones AR 2017 Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. J Proteome Res 16(12):4288–4298. [PubMed: 28849660]

de Leeuw CA, Mooij JM, Heskes T, Posthuma D 2015 MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 11(4).

Efron B, and Tibshirani R 2007 On testing the significance of sets of genes. *Ann*. Appl. Stat. 1:107–129

Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, and Bateman A 2010 The Pfam protein families database. Nucleic Acids Res 38:D211–D222. [PubMed: 19920124]

Heintzman ND, Ren B 2009 Finding distal regulatory elements in the human genome. Curr Opin Genet Dev 19(6): 541–549. [PubMed: 19854636]

Herman MA, Rosen ED 2015 Making Biological Sense of GWAS Data: Lessons from the FTO Locus. Cell Metab 22(4):538–9. [PubMed: 26445508]

Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von MC, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, and Apweiler R 2004 The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat. Biotechnol 22:177–183. [PubMed: 14755292]

Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P Wellcome Trust Case-Control Consortium, Owen MJ, O'Donovan MC, and Craddock N 2009 Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am. J. Hum. Genet 85:13–24. [PubMed: 19539887]

Hong MG, Pawitan Y, Magnusson PK, and Prince JA 2009 Strategies and issues in the detection of pathway enrichment in genome-wide association studies. Hum. Genet 126:289–301. [PubMed: 19408013]

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le NN, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, and Wang J 2003 The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531. [PubMed: 12611808]

Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de BB, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, and Stein L 2005 Reactome: A knowledgebase of biological pathways. Nucleic Acids Res 33:D428–D432. [PubMed: 15608231]

Kandasamy K, Mohan SS, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C, Gollapudi SK, Tattikota SG, Mohan S, Padhukasahasram H, Subbannayya Y, Goel R, Jacob HK, Zhong J, Sekhar R, Nanjappa V, Balakrishnan L, Subbaiah R, Ramachandra YL, Rahiman BA, Prasad TS, Lin JX, Houtman JC, Desiderio S, Renauld JC, Constantinescu SN, Ohara O, Hirano T, Kubo M, Singh S, Khatri P, Draghici S, Bader GD, Sander C, Leonard WJ, and Pandey A 2010 NetPath: A public resource of curated signal transduction pathways. Genome Biol 11:R3. [PubMed: 20067622]

Kanehisa M and Goto S 2000 KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30. [PubMed: 10592173]

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, and Hirakawa M 2006 From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res 34:D354–D357. [PubMed: 16381885]

Kanehisa M, Goto S, Furumichi M, Tanabe M, and Hirakawa M 2010 KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38:D355–D360. [PubMed: 19880382]

Liu JZ, Mcrae AF, Nyholt DR, Medland SE, Wray NR, Brown KM, AMFS Investigators, Hayward NK, Montgomery GW, Visscher PM, Martin NG, Macgregor S 2010 A Versatile Gene-Based Test for Genome-wide Association Studies. Am J Hum Genet 87(1): 139–145. [PubMed: 20598278]

Lee PH, O'Dushlaine C, Thomas B, Purcell SM 2012 INRICH: interval-based enrichment analysis for genome-wide association studies. Bioinformatics 28(13):1797–9. [PubMed: 22513993]

Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, Al-Shahrour F and Dopaso J 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. [PubMed: 19502494]

Mishra A, Macgregor S 2015 VEGAS2: Software for More Flexible Gene-Based Testing. Twin Res Hum Genet18(1):86–91. [PubMed: 25518859]

Mishra A, Macgregor S 2017 A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. Twin Res Hum Genet 20(1):1–9. [PubMed: 28105966]

Nam D, Kim J, Kim SY, Kim S 2010 GSA-SNP: a general approach for gene set analysis of polymorphisms. Nucleic Acids Res 38(Web Server issue):W749–54. [PubMed: 20501604]

O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, and Corvin A 2009 The SNP ratio test: Pathway analysis of genome-wide association datasets. Bioinformatics 25:2762–2763. [PubMed: 19620097]

Ooi HS, Schneider G, Lim TT, Chan YL, Eisenhaber B, and Eisenhaber F 2010 Biomolecular pathway databases. Methods Mol. Biol 609:129–144. [PubMed: 20221917]

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC 2007 PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet 81:559–575. [PubMed: 17701901]

The 1000 Genomes Project Consortium. 2015 A global reference for human genetic variation. Nature 526: 68–74. [PubMed: 26432245]

Wall JD, Pritchard JK 2003 Haplotype blocks and linkage disequilibrium in the human genome. Nat. Rev. Genet 4: 587–597 [PubMed: 12897771]

Wang K, Li M, and Hakonarson H 2010 Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet 11:843–854. [PubMed: 21085203]

Watanabe K, Taskesen E, van Bochoven A, Posthuma D 2017 Functional mapping and annotation of genetic associations with FUMA. Nat. Comm 8:1826.

Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, and Eisenberg D 2000 DIP: The database of interacting proteins. Nucleic Acids Res 28:289–291. [PubMed: 10592249]

Yang W, de las Fuentes L, Davila-Roman VG, and Gu CC 2011 Variable set enrichment analysis in genome-wide association studies. Eur. J. Hum. Genet 19:893–900. [PubMed: 21427759]

Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, Sutcliffe JS, and Haines JL 2011 Genetic analysis of biological pathway data through genomic randomization. Hum. Genet 129:563–571. [PubMed: 21279722]

Yoon S, Nguyen HCT, Yoo YJ, Kim J, Baik B, Kim S, Kim J, Kim S, Nam D 2018 Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Res 46(10):e60. [PubMed: 29562348]

Zhang K, Chang S, Cui S, Guo L, Zhang L, Wang J 2011 ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework. Nucleic Acids Res 39(suppl 2): W437–W443. [PubMed: 21622953]

Zhang KL, Cui SJ, Chang SH, Zhang LY, Wang J 2010 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(Web Server issue):W90–5. [PubMed: 20435672]

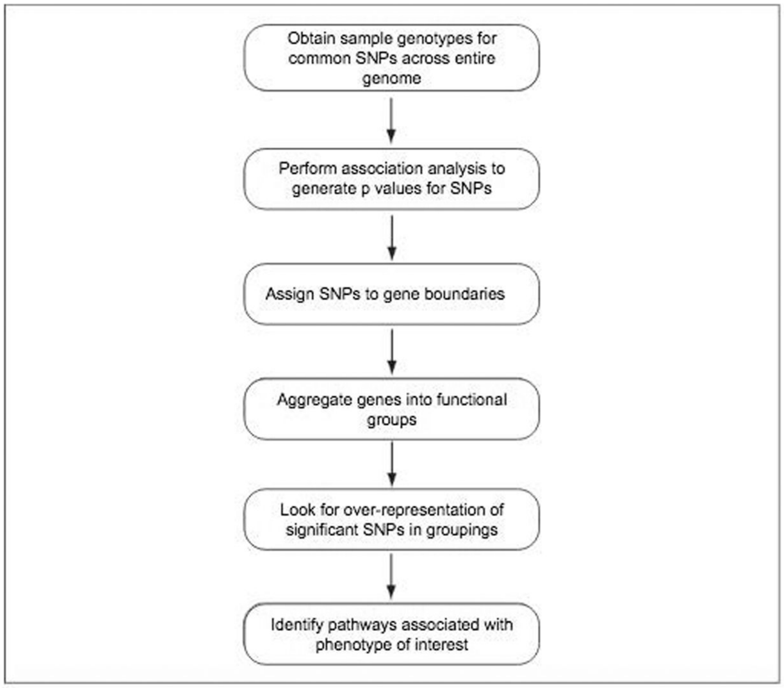
Obtain sample genotypes for common SNPs across entire genome
Perform association analysis to generate p values for SNPs
Assign SNPs to gene boundaries
Aggregate genes into functional groups
Look for over-representation of significant SNPs in groupings
Identify pathways associated with phenotype of interest

Figure 1:

Overall protocol for pathway-based GWAS / WGS analysis.

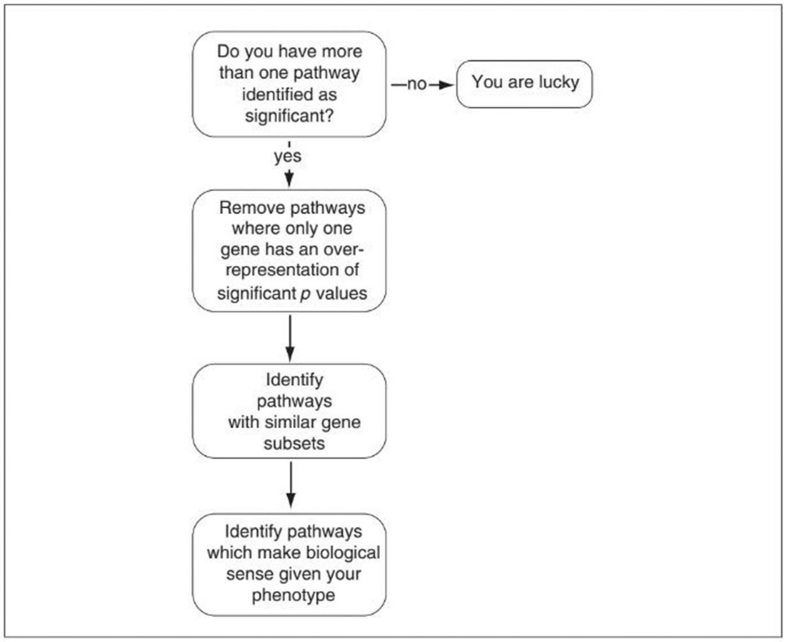
Do you have more than one pathway identified as significant?
no
You are lucky
yes
Remove pathways where only one gene has an over-representation of significant p values
Identify pathways with similar gene subsets
Identify pathways which make biological sense given your phenotype

Figure 2:
Flowchart for sorting through a list of pathways after the algorithm has finished.

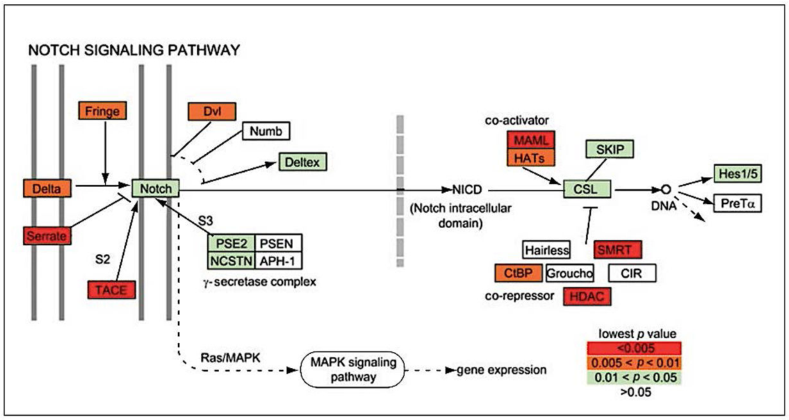

Figure 3:

Notch signaling pathway as seen in the KEGG database (hsa:04330). The most significant SNP in the gene from the single-allele analysis is colored. A graphic overlay such as this can highlight genes most likely to be of interest in follow-up studies such as gene-gene interaction analyses. Reprinted with permission from Kanehisa Laboratories.

Table 1

Pathway Databases Available for Evaluation

PATHWAY DATABASE	AVAILABILITY	STANDARD LANGUAGE FORMAT	FUNCTIONAL FOCUS	CURATION	DATABASE STRUCTURE	REFERENCE	DATABASE LINK
KEGG	Free to academic users	BioPAX	Metabolic pathways	Manual	Both	Kanehisa et al. (2006)	http://www.genome.jp/kegg/
GO	Free to public	No	Protein-protein interactions, metabolic pathways, signaling pathways	Both	Ontology	Ashburner et al. (2000)	http://www.geneontology.org
Reactome	Free to public	BioPAX, SBML	Diverse cellular level metabolic processes	Manual	Discrete	Joshi-Tope et al. (2005)	http://www.reactome.org
BioCarta	Free to public	No	Diverse biological processes	Manual	Discrete	n/a	http://www.biocarta.com/genes/index.asp
Ingenuity Knowledge Base	License purchase required	No	Protein-protein interactions, metabolic pathways, signaling pathways, transcription factors/gene regulatory networks, protein-compound interactions	Manual	Ontology	n/a	http://www.ingenuity.com/
NetPath	Free to public	BioPAX, PSI-MI, SBML	Signal transduction pathways (immune and cancer)	Manual	Ontology	Kandasamy et al. (2010)	http://www.netpath.org/
Pfam	Free to public	No	Protein families	Manual	n/a	Bateman et al. (2004)	http://pfam.janelia.org/
DIP	Free to academic users	PSI-MI, SQL	Protein-protein interactions	Both	n/a	Xenarios et al. (2000)	http://dip.doe-mbi.ucla.edu/
Pathway Commons	Free to public	BioPax	Protein-protein interactions, metabolic pathways, signaling pathways, protein-compound interactions	Both	Both	Cerami et al. (2011)	http://www.pathwaycommons.org

Table 2

Pathway-Based Analysis from GWAS Data Algorithm Comparison

METHOD USED	PATHWAY DATABASES	REQUIRES ORIGINAL DATASET	LD CORRECTION	GENE SIZE CORRECTION	STUDY DESIGNS ALLOWED	REFERENCE	DOWNLOAD LINK	NOTES
FUMA	see MAGMA	No	Pre-computed by PLINK	see MAGMA	Any	Watanabe et al. (2017)	http://fuma.ctglab.nl	Incorporates 18 biological data repositories for improved annotation
GSA-SNP2	GO, custom	No	Re-standardization method by Efron et al. (2007)	No	Any	Nam et al. (2010)	https://sourceforge.net/projects/gsa-snp/	use the kth best SNP as the gene level p-value to limit false positives
i-GSEA4GWASv2	KEGG, GO, BioCarta, custom	No	Variant Label Permutation	Yes	Any	Zhang et al. (2010)	http://gsea4gwas-v2.psych.ac.cn/inputPage.jsp	Web-based interface
ICSN Pathway	KEGG, GO, BioCarta, custom	No	LD analysis using information from HapMap populations	Yes	Any	Zhang et al. (2011)	http://icsnpathway.psych.ac.cn/	Web-based interface
Ingenuity Pathway Analysis	Ingenuity Knowledge Base	No	Done prior to IPA	Done prior to IPA	Any	http://www.ingenuity.com	http://www.ingenuity.com/products/pathways_analysis.html	Commercial product; web-based interface
INRICH	KEGG, GO, Custom	No	Permutation of genomic intervals	Yes	Any	Lee et al. (2012)	http://atgu.mgh.harvard.edu/inrich/ https://bitbucket.org/statgen/inrich/downloads/	
MAGMA	Any	No	Permutation	Yes	Any	de Leeuw et al. (2015)	https://ctg.cncr.nl/software/magma	Can perform Gene × Environment analyses
PARIS	KEGG, GO, Reactome, NetPath, Pfam, DIP, custom	No	Genomic randomization	Yes	Any	Yaspan et al. (2011) Butkiewicz et al. (2016)	https://ritchielab.psu.edu/software/paris-download	
PLINK	Any	Yes	Case-control status randomization	No	Any	Purcell et al. (2007)	http://pngu.mgh.harvard.edu/~purcell/plink/	
proxyGeneLD	Any	No	Bonferroni-type based on # of LD blocks and LE SNPs	Yes	Any	Hong et al. (2009)	https://github.com/Rundmus/ProxyGeneLD	Currently only for Caucasian datasets
SRT	Any	Yes	Case-control status randomization	Yes	Case-control[a]	O'Dushlaine et al. (2009)	https://sourceforge.net/projects/snpratiotest/	Designed to work with PLINK input/output
VEGAS2	GO, BIOCARTA, Reactome, KEGG, PANTHER, custom	No	Simulations from the multivariate normal distribution	Yes	Any	Liu et al. (2010) Mishra & MacGregor (2015) Mishra & MacGregor (2017)	https://vegas2.qimrberghofer.edu.au/ https://vegas2.qimrberghofer.edu.au/zVEGAS2offline.tgz	
VSEA	Any	Yes	Gene-score normalization	Yes	Case-control[a]	Yang et al. (2011)	Component of R package	Produces normalized gene score and uses GSEA to calculate enrichment of gene set

[a] Modifiable for use with family-based datasets by using non-transmitted alleles to create pseudo-controls.