CellPress

Trends in Genetics

Review

Interpreting omics data with pathway enrichment analysis

Kangmei Zhao[1,*] and Seung Yon Rhee[1,*]

Pathway enrichment analysis is indispensable for interpreting omics datasets and generating hypotheses. However, the foundations of enrichment analysis remain elusive to many biologists. Here, we discuss best practices in interpreting different types of omics data using pathway enrichment analysis and highlight the importance of considering intrinsic features of various types of omics data. We further explain major components that influence the outcomes of a pathway enrichment analysis, including defining background sets and choosing reference annotation databases. To improve reproducibility, we describe how to standardize reporting methodological details in publications. This article aims to serve as a primer for biologists to leverage the wealth of omics resources and motivate bioinformatics tool developers to enhance the power of pathway enrichment analysis.

Highlights

Omics technologies enable holistic understanding of biological processes and establish relationships between genotypes and phenotypes.

Pathway enrichment analysis has become a standard method to interpret various types of omics data by identifying significantly impacted biological pathways.

Understanding the intrinsic features of omics data and selecting appropriate background sets and reference annotation databases are essential for generating reliable results.

A classic approach to interpreting omics data: pathway enrichment analysis

Advances in omics approaches enabled profiling transcripts, proteins, metabolites, and epigenetic modifications at genome scale, which can facilitate establishing holistic views of how organisms function and evolve [1–4]. The output of omics approaches is usually represented by long lists of genes or their downstream products, the interpretation of which remains challenging to many biologists [2]. Pathway enrichment analysis identifies metabolic pathways enriched in a dataset more than expected by chance, which has become a routine method employed to interpret omics datasets [3–5]. This approach was originally developed to analyze transcriptomic profiles generated by microarrays or RNA sequencing (RNA-seq), then expanded to explore a wide range of omics datasets, such as epigenomics, genome-wide association studies (GWAS), single-cell RNA-seq (scRNA-seq), and integrated omics data [6–10]. Pathway enrichment analysis has become integral to revealing patterns underlying various types of omics data and formulating hypotheses for downstream experimental investigations.

Despite the popularity of the technique, principles underlying pathway enrichment analysis remain obscure to many scientists, which leads to inappropriate statistical tests and unreliable results. Implementing correct methods requires understanding the intrinsic mathematical features of various omics data, particularly those generated by emerging technologies, such as single-cell and spatial transcriptomics [11,12]. Over the past few decades, more than 100 tools have been developed for enrichment analysis, whose performance is assessed by various benchmarking studies [13]. Yet, there is no practical guideline on choosing methods on the basis of intrinsic features of various types of omics datasets [13–15]. Besides statistical methods, selection of input set, **background gene sets** (see Glossary), and **reference annotation databases** dramatically influence the results of pathway enrichment analysis [13]. Failing to use appropriate background and up-to-date annotations has become a prominent issue for studies conducting enrichment tests. For example, a recent investigation surveyed nearly 200 peer-reviewed research articles reporting pathway enrichment analysis, and over 90% of the publications failed to

[1]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 94025, USA

*Correspondence: kzhao@carnegiescience.edu (K. Zhao) and srhee@carnegiescience.edu (S.Y. Rhee).

implement correct background gene sets [11]. Thus, discussions about best practices of pathway enrichment analysis are needed to establish standards for conducting the test and reporting results.

In this review, we provide a holistic view for performing pathway enrichment analysis using various omics datasets and interpreting results from a biologist's perspective (Figure 1). Specifically, we explain the statistical features of different omics datasets and discuss how they should be accounted for when designing pathway enrichment analysis. We further discuss the impacts of background gene sets and annotation databases on pathway enrichment analysis and provide guidelines for how best to choose these input data. Finally, we emphasize the importance of reporting methodological details needed for justifying the results of pathway enrichment analysis and provide guidelines for method documentation in publications.

Glossary

Background gene sets: represent the list of genes detected by omics approaches, such as total genes or proteins detected in the transcriptomic or proteomic profiling.
Network topology: represents the arrangement patterns of nodes and edges in a network, such as position of a metabolite in a metabolic pathway or a gene in a gene regulatory network.
Reference annotation databases: refer to publicly available knowledge bases that provide annotations for enzymes and pathways for different organisms, such as MetaCyc, Kyoto Encyclopedia of Genes and Genomes (KEGG), PlantReactome, and the Plant Metabolic Network (PMN).

Overview of pathway enrichment analysis methodology

Many tools have been developed for conducting pathway enrichment analysis in the past few decades, which can be grouped into three categories based on statistical approaches: (i) over-representation-based, (ii) functional scoring system (ranking)-based, and (iii) pathway topology-based methods [13,16–19] (Figure 1). The over-representation-based method requires a gene list of interest and tests whether any pathways are observed in this list more than expected by chance against a predefined background gene set [18,20]. Ranking-based methods consider functional information generated by different omics datasets, such as levels of gene expression [6,21,22]. This type of tool first ranks the total gene set based on detected signals in omics studies, such as transcript abundance, then tests whether genes annotated to the same pathway tend to cluster together at the top (or bottom) of the ranked list. Topology-based methods aim to account for additional information that impacts pathway activity by integrating scores measuring gene positions within a pathway and gene–gene interactions into the enrichment tests [19,23,24]. These three types of methods establish foundations for pathway enrichment analysis tools to interpret transcriptomic profiles and other types of omics data.

Classic input data: transcriptomics

Transcriptomic profiles generated by microarray or RNA-seq quantify the abundances of genes at a system scale, which facilitates characterizing genes of unknown function [25,26]. The

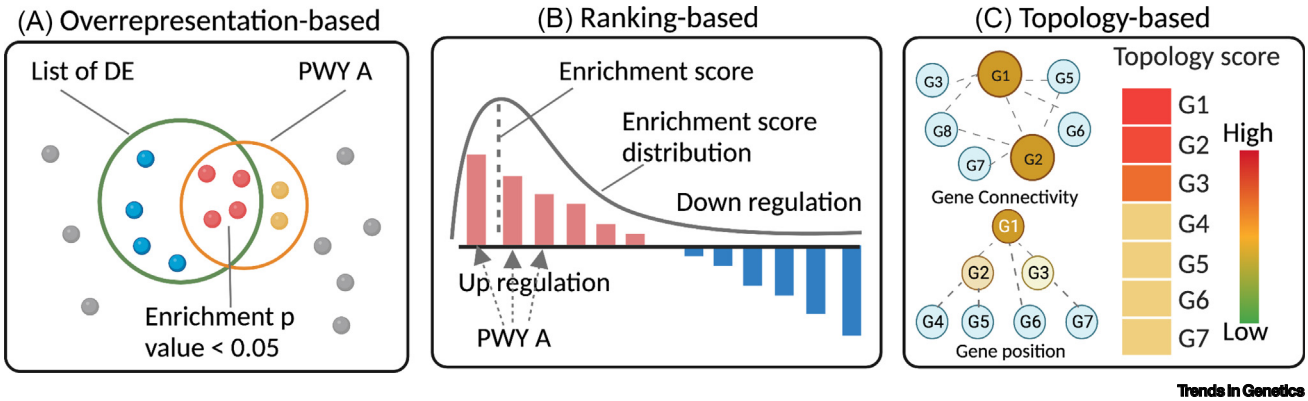

Figure 1. Overview of three types of methods for pathway enrichment analysis. (A) Over-representation-based methods examine whether any pathways are observed in a gene list of interest more than expected by chance compared with a background set. (B) Ranking-based methods first rank the total gene set on the basis of detected signals, such as change of gene expression, then tests whether genes annotated to the same pathway tend to cluster together at the top (or bottom) of the ranked list. (C) Topology-based methods integrate scores measuring gene positions within a pathway and gene–gene interactions into the enrichment tests. DE represents differentially expressed genes. PWY represents a pathway. G1 to G7 represent gene 1 to gene 7.

standard workflow of analyzing transcriptomic datasets is first to identify differentially expressed genes under stress conditions or genetic perturbation. Then a list of differentially expressed genes or the entire transcriptomic profile can be used to perform pathway enrichment analysis to identify significantly impacted biological processes [6,27] (Figure 2).

Pathway enrichment analysis tools implementing all three types of statistical approaches described above are available to interpret transcriptomic datasets. A classic example of employing over-representation-based methods in enrichment analysis is to search a list of differentially expressed genes against a reference annotation database, such as Gene Ontology (GO) (hosted at PANTHER), or Kyoto Encyclopedia of Genes and Genomes (KEGG) [17,28]. These tools will identify pathways whose constituent genes are more abundant in the input gene list based on Fisher's exact test [29]. Over-representation-based methods are conceptually straightforward but have several limitations, such as assuming independence of each gene and requiring an arbitrary cutoff to define differentially expressed gene sets. To alleviate these challenges, ranking-based methods are widely applied, and gene set enrichment analysis (GSEA) represents a popular method to analyze transcriptomic datasets. In GSEA, genes are generally ranked by values of the phenotype of interest, such as gene expression levels. GSEA computes the enrichment score for a pathway by first scanning the ranked gene list from top to bottom, followed by quantifying the distances of all genes annotated to this pathway to the middle of the rank [30,31]. Topology-based methods aim to increase the sensitivity of pathway enrichment analysis by considering the topological features of genes, such as gene position within a pathway and connectivity with other genes in coexpression or other functional networks [32]. A method called topology-based pathway enrichment analysis (TPEA) developed a scoring system to measure gene location within a pathway, number of interactions, and expression levels [24]. Though having potentially high performance, TPEA requires experimental evidence for pathway structures and gene–gene interactions, which is largely unavailable for many organisms. In summary, users can select tools for pathway enrichment analysis on the basis of available datasets and interpret the results while acknowledging the principles of each method.

Special considerations of interpreting other types of omics data

Besides transcriptomic profiling, various types of omics approaches have been developed to describe different facets of biological systems, such as proteomics, metabolomics, scRNA-

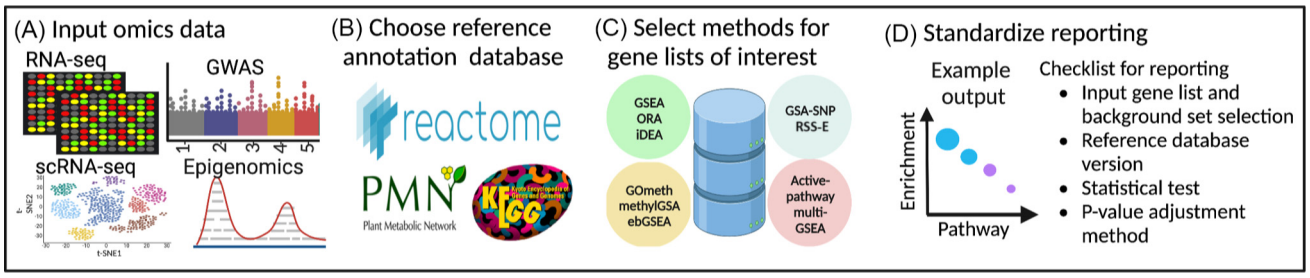

Figure 2. General workflow for interpreting various types of omics data with pathway enrichment analysis and documenting results. (A) Various types of omics data that can be interpreted using pathway enrichment analysis. (B) Choose annotation databases that provide genome-scale mapping for enzymes and pathways for different organisms. (C) Select methods that are suitable for identifying enriched pathways using various types of omics data. (D) Standardize reporting by including a list of major components of a pathway enrichment analysis method and results in publications. Abbreviations: ebGSEA, empirical Bayes gene set enrichment analysis; GOmeth, Gene Ontology testing for methylation profiles; GSA-SNP, gene set analysis for SNPs; GSEA, gene set enrichment analysis; GWAS, genome-wide association studies; iDEA, integrative differential expression and gene set enrichment analysis; methylGSA, methylation gene set analysis; multi-GSEA, multiomics gene set enrichment analysis; ORA, over-representation analysis; RNA-seq, RNA sequencing; RSS-E, regression with summary statistics enrichment analysis; scRNA-seq, single-cell RNA sequencing.

seq, GWAS, and epigenomic profiling. Pathway enrichment analysis serves as a classic method to identify patterns using these types of omics profiles. However, proteomics, metabolomics, scRNA-seq, GWAS, and epigenomic datasets have statistical distributions different from gene expression data, and how to accommodate these unique features may still be obscure to most of the scientific community. In the following sections, we explain how intrinsic features of these omics data impact pathway enrichment analysis and discuss recent advances in methods for performing pathway enrichment analysis using different types of omics datasets.

Proteomics

Proteomics help discover proteins functioning in the same biological processes, characterize subunits of a protein complex, and identify post-translational modifications responding to different conditions [33,34]. As proteins are not amplifiable, quantification of the proteome is more challenging than it is for the transcriptome. For example, proteomic datasets generated by mass spectrometry may bias toward detecting highly expressed proteins [34,35]. Moreover, proteins often work as complexes, and the quantification of a given peptide may depend on the coelution of their partners [34,35]. These biases may introduce variation between replicates and experiments, which should be considered during protein quantification and pathway enrichment analysis using proteomic datasets.

Several strategies have attempted to identify enriched pathways while accommodating the high variability in proteomic datasets. Inspired by the ranking-based pathway enrichment analysis method GSEA, a method called protein set enrichment analysis (PSEA) identifies enriched pathways using protein differential expression score [36]. PSEA takes in protein relative abundance represented by spectral counts as input and ranks them on the basis of the fraction of abundance change between conditions. Then, it computes the enrichment score of a pathway by taking the sum of distance of its constituent proteins to the middle of the ranked list [36]. PSEA helps consider the variation of protein quantification between conditions, but it is not designed to analyze datasets obtained with label-based techniques of protein quantification. To fill this gap, another tool, called PSEA-Quant, can identify enriched pathways using proteomic datasets generated by both label-based and label-free quantification methods. PSEA-Quant first computes an enrichment score for each protein by integrating their average abundance and variation between replicates [37]. It then ranks all the proteins detected in the datasets using this score and assigns higher weights to the ones showing high abundance and low variation. The enrichment score of a pathway is represented by the sum of weighted enrichment scores of its constituent proteins [37]. The statistical significance of enrichment is determined by comparing the enrichment score of each pathway to a null distribution assembled by randomly sampling proteins in the dataset [37]. These tools may help alleviate variation in proteomic datasets and identify pathways enriched with robust protein abundance measurements.

Metabolomics

Metabolomics systematically quantifies small molecules in a biological system, which is essential to identify metabolic responses to diseases and environmental signals and discover the biosynthetic routes of economically important compounds [38–40]. Quantification of metabolites relies on chemical standards or analyzing fragmentation patterns of compounds generated by tandem mass spectrometry [38]. High-throughput compound annotation represents a major bottleneck in analyzing metabolomic datasets, which makes metabolomic profiles sparser and more ambiguous than transcriptomics.

Two types of pathway enrichment analysis methods are available to interpret metabolomic datasets. The first strategy requires annotating compounds on the basis of chemical standards

or searching metabolic features, such as mass-to-charge ratio (*m/z*) or fragmentation patterns generated by tandem mass spectrometry, against metabolite libraries, before conducting pathway enrichment analysis. A representative method is called Metabolomics Pathway Analysis (MetPA), which relies on compound annotations and **network topology** to identify enriched pathways [41]. This method first converts pathways to a metabolic network with metabolites as nodes and reactions as edges. Then, it calculates the 'importance' for compounds by considering (i) the abundance change measured by metabolomic profiling, and (ii) their connectivity in the network using relative betweenness centrality and degree centrality measures. Enriched pathways are identified by comparing the frequency of 'important' metabolites annotated to a pathway with the frequency expected by chance using Fisher's test [41]. This method is limited to organisms with a prior knowledge of network topology and compound annotations. To further harness the wealth of information generated by untargeted metabolomics, a method named 'Mummichog' was developed to bypass compound annotation and directly predict pathways enriched with significantly impacted spectral features represented by *m/z* and retention time [40]. This method assumes that if a list of spectral features represents biological activities, then they should be more likely to be annotated to functionally related compounds than to be randomly distributed in the metabolic network. Mummichog first identifies a list of spectral features that are significantly altered between samples from the total features captured by untargeted metabolomics [40]. Then it identifies enriched pathways by comparing the probability of annotating the significantly altered spectral features to a pathway with the probability of randomly assigning features to compounds in this pathway [40]. This method can serve as an initial functional analysis to interpret datasets generated by high-throughput untargeted metabolomics. However, it may be less accurate than conventional enrichment analysis tools and requires downstream analytical chemistry to identify specific metabolites associated with phenotypes of interest [40]. Nonetheless, pathway enrichment analysis converts metabolic profiles to biological processes, which may help characterize enzymes and metabolites associated with phenotypes of interest.

Single-cell RNA-seq

scRNA-seq generates high-resolution gene expression measurements, which facilitates discovering novel mechanisms contributing to cell type specificity, cellular response, and communications between cells [42–45]. Due to the low abundance of RNA within each cell, scRNA-seq data are sparse and noisy compared with conventional bulk transcriptomic profiling [46]. The high dropout rate (e.g., many below-detection threshold expression values) alters the distribution of gene expression data, which influences the statistical analysis for identifying differentially expressed genes and enriched pathways [10,47].

One strategy to handle sparse datasets is to exploit expression patterns between similar genes. A method called 'iDEA' implements this idea by jointly performing differential gene expression and pathway enrichment analyses for all genes detected in scRNA-seq profiles. Bayesian hierarchical modeling was implemented to compute whether the probability of a gene being differentially expressed was higher than expected by chance [47]. Then, enriched pathways were identified by computing the probability that a pathway contains more differentially expressed genes than the null distribution where these genes were randomly assigned to the pathway [47]. Applying this method to scRNA-seq profiles of human embryonic stem cells showed that iDEA can identify pathways that are functionally related to embryo development more accurately and comprehensively than popular methods developed for bulk RNA-seq [47]. These resources facilitate identifying heterogeneity of pathway activity at a single-cell resolution, which may help engineer biological processes and develop novel therapeutic strategies.

Genome-wide association studies

Besides transcriptomic profiling, GWAS are widely applied to establish relationships between genotypes and phenotypes [48,49]. GWAS predict whether SNPs are associated with phenotypes of interest by computing the correlation between genetic diversity and phenotypic variation [50,51]. This analysis leads to a long list of candidate SNPs, which hampers downstream experimental validation. Pathway enrichment analysis can function as an orthogonal method to GWAS by identifying the biological processes that are annotated to SNPs associated with certain traits more frequently than expected by chance [7,52]. SNP-GSEA represents a tool implementing this strategy to identify enriched pathways using SNPs as input [53]. The intention is to eliminate random association between SNPs and traits and prioritize biological processes for functional characterization. Though conceptually straightforward, SNPs have unique features, which should be accounted for when performing pathway enrichment analysis. First, unlike gene expression datasets that directly measure the abundance of individual transcripts, SNPs cannot always be mapped to genes, because they can be in both coding and noncoding regions [48]. Moreover, SNPs are not evenly distributed across the genome, which impacts how statistical methods should be selected for pathway enrichment analysis [48].

Several methods are available to perform pathway enrichment analysis using GWAS data while considering the features described above. All these methods require first mapping SNPs to genes. Recently, a method called regression with summary statistics enrichment analysis (RSS-E) has been developed to perform pathway enrichment analysis using SNPs by implementing Bayesian multiple regression models [8]. This approach examines whether SNPs mapped to the same pathway are more likely to be associated with the same trait than the baseline distribution of SNPs randomly assigned to traits [8]. This analysis uses all SNPs as input, regardless of their effect size, which helps identify novel genes associated with a trait. The performance of this tool was benchmarked using the 1.1 million HapMap3 SNPs for 31 traits in humans, and it recovered known associations between pathways and traits. It also made novel pathway–trait connections, which may help develop new lines of research to further dissect the genetic basis of these traits [8].

Epigenomics

Epigenetic modifications on histone proteins and DNA play essential roles in growth and adaptation in many organisms [54,55]. Genome-scale maps of epigenetic modifications are integral to understanding the regulatory mechanisms of gene expression. The current workflow to process epigenomic datasets first identifies genomic loci that have significantly altered epigenetic modifications, then annotates these regions to the nearest genes [56,57]. After generating gene lists, pathway enrichment analysis can be performed to identify the biological processes enriched within differentially modified genes. However, these conventional approaches may not yield robust results, because they ignore the inherent features of epigenomic datasets. For example, several studies demonstrate that identification of differentially modified genes is biased toward long genes because they yield more reads during sequencing [58,59]. In addition, for DNA methylation profiles, about 10% of the methylation hotspots, named CpG islands, can be mapped to multiple genes [9,58].

Several strategies attempt to account for the intrinsic features of epigenomic datasets when conducting pathway enrichment analysis, most of which are designed to process DNA methylation profiles. A method called ‘ebGSEA’ implements a ranking-based method to identify enriched pathways using DNA methylation profiles as inputs [59]. It first ranks all the genes on the basis of their overall change in methylation abundance, then performs enrichment analysis with this ranked gene list to identify pathways containing more differentially methylated genes than expected by chance [59]. ebGSEA considers the impact of extent of DNA methylation change

on pathway enrichment. However, it does not address the issue that the same CpG sites can be mapped to multiple genes. Another method, called 'GOmeth', addresses the biases caused by gene length and multigene mapping by performing empirical analysis with CpG sites [9]. Specifically, the contribution of each CpG site on a gene is normalized by the number of genes this CpG is annotated to, which eliminates the impacts of multigene mapping [9]. The weights of all CpGs mapped to the same gene are summed and normalized by gene length to account for the gene size bias. Enriched pathways are identified on the basis of Wallenius' noncentral hypergeometric distribution, which is a generalized version of the hypergeometric distribution where corrections for sample biases can be considered [9]. This approach was benchmarked using both simulated and publicly available methylome data and could identify the most biologically relevant terms and pathways compared with several other existing methods [9]. Taken together, better understanding of the statistical features associated with epigenomic datasets can improve the sensitivity and robustness of pathway enrichment analysis.

Integrated omics

Different omics datasets describe distinct aspects of the central dogma in molecular biology, and integrating these resources may provide novel insights to dissect the genetic basis underlying development and disease [60,61]. For example, discovering the driver mutations causing different types of cancer is essential for developing effective therapeutic approaches. Integrating genetic variation and transcriptomic profiling helped generate a high-confidence catalog of driver mutations for different types of cancer [62]. Integrated omics represents an emerging and powerful strategy for generating hypotheses through data fusion. However, the integration process will generate high-dimensional datasets, which makes data interpretation and downstream analysis more challenging. New tools are required to further leverage the wealth of the information embedded in multiomics datasets, such as identifying enriched pathways within a set of genes showing certain genomic features. A method named 'ActivePathway' was developed recently to perform pathway enrichment analysis using multiomics datasets [63]. This method directly aggregates *p* values generated by processing individual omics data, such as differential gene expression and gene essentiality analyses, then generates a representative *p* value using Fisher's combined probability test to represent the significance of each gene using integrated omics datasets [63]. Then the representative *p* values are ranked and filtered on the basis of user-defined cutoffs. Pathway enrichment analysis is performed using a ranked hypergeometric test to examine whether a given pathway containing genes with low representative *p* values is more than expected by chance. This method was applied to identify pathways associated with prognosis in breast cancer by integrating transcriptome and gene copy number alterations. The results showed that immune activity in breast tumor cells and in the surrounding microenvironment affects prognosis [63]. Implementing multiomics datasets identified the most comprehensive lists of functionally related pathways as compared with solely using individual omics datasets [63]. These studies demonstrate the power of integrated omics for discovering systematic understanding of biological processes. With the rapid accumulation of omics datasets, developing new tools with high efficiency and computational power will be critical to further exploit the wealth of omics resources.

Input sets, background sets, and reference annotation databases

In addition to choosing appropriate methods for various types of omics data, accurate input and background sets and suitable reference annotation databases represent additional major components of robust pathway enrichment analysis. Both the input and background sets should be defined on the basis of biological questions of interest and reflect the actual number of genes captured by omics approaches [6,64]. Reference annotation databases provide infrastructures that organize genes to pathways, which determines how pathways are defined and the number of constituent genes each pathway has [6]. In this section, we explain how these two

components impact pathway enrichment analyses and provide practical guidelines for selecting proper background sets and reference annotation databases.

General considerations for input data

Assembling a meaningful input set is a prerequisite to interpreting omics datasets using pathway enrichment analysis. The general process of input set selection includes preprocessing raw signals measured by different omics approaches, then identifying the list of genes (or proteins, metabolites, genomic regions) by considering experimental design and applying rigorous statistical tests [6]. Refining the input gene set by eliminating irrelevant genes may help improve the performance of pathway enrichment analysis [6,64]. For example, if the researcher aims to identify novel pathways associated with a disease using enrichment analysis, filtering the input gene list to focus only on the ones correlated or physically interacting with genes known to be associated with this phenotype may yield more accurate results [64]. These practices are generally applicable to analysis of omics datasets using pathway enrichment analysis.

Impact of the background set

Selecting an appropriate background set used in the enrichment analyses is key to answering the biological question of interest. The most frequently used background set in pathway enrichment analysis is the total number of genes (or proteins) annotated in a species. This is inaccurate because only a subset of these genes (or proteins) can be captured in the experiments [2,11,65]. To demonstrate the impact of the background set on results, pathway enrichment analysis was performed using seven RNA-seq profiles with total genes in the genome and the correct background, respectively. The results showed that, on average, only 44% of enriched pathways were found in common between these analyses that used the two different background sets [11]. Thus, an accurate background set for pathway enrichment analysis should include only the number of genes whose transcript (or protein) levels are above the noise threshold rather than total annotated genes (or proteins) in a genome.

Background sets should be further evaluated and customized on the basis of biological questions of interest. If the study focuses on only a subset of genes in the genome, such as metabolic or signaling genes, then total genes detected in the omics study may not be appropriate to serve as the background. For example, enrichment analysis was performed to examine whether genes associated with specialized metabolism were associated with specific epigenetic modification patterns relative to other types of metabolic genes [66]. In this analysis, total metabolic genes were chosen as the background set instead of total genes detected in epigenomic profiling [66]. In summary, background sets should be defined on the basis of the number of genes (or other targets) detected by different omics approaches and the scope of biological questions [2].

Impact of reference annotation databases

Reference annotation databases provide the infrastructure to annotate genes to pathways, which describes how biological processes are organized in an organism. Several reference databases provide genome-scale annotations for enzymes, metabolites, and pathways [2,67]. These annotation databases use different ontologies to define pathways, which leads to a dramatic variation in pathway number and size for the same species [17,28,67,68]. For example, *Escherichia coli* metabolic genes are annotated to fewer pathways in KEGG, which leads to bigger pathways containing more constituent genes per pathway, than the annotations in EcoCyc [68]. Pathway size affects pathway enrichment results because larger pathways require more differentially expressed genes in order to be identified as significantly enriched/depleted than analogous yet smaller pathways [68]. In addition, large pathways may not provide sufficient granularity to dissect the actual biological processes responding to the perturbation [68]. Thus, implementing various

reference annotation databases in pathway enrichment analysis may yield different biological processes, and the results should be interpreted by understanding the inherent structures of each database. Besides performing pathway enrichment analysis using individual reference databases, integrating annotations from different databases can be achieved by taking the union of enzymes and metabolites associated with the same reactions, which may help expand the list of experimentally characterized enzymes and metabolites [40]. To further leverage the wealth of information to improve the robustness of pathway enrichment analysis, a uniform annotation is required to make pathways comparable between different databases.

Different versions of annotations provided by the same reference database also dramatically influence pathway enrichment analysis. Researchers should be particularly cautious when performing pathway enrichment analysis using online tools. A systematic survey about web-based pathway enrichment tools suggested that 16 out of 25 tools provided only pathway annotations that have been outdated for several years [69]. This impact of outdated annotations has been assessed by benchmarking different versions of GO in enrichment tests [69,70]. In general, using different versions of GO annotations results in low consistency in enrichment analyses using the same transcriptomic dataset. For example, 74% of enriched GO terms using the annotations generated in 2016 were missing from the

Table 1. Representative tools for performing pathway enrichment analysis using different types of omics datasets[a]

Method	Suitable omics data	Advantages	Limitations	Refs
Fisher's test	Transcriptomics	Easy to implement	Ignores gene expression changes, assumes gene independence	[73]
PADOG	Transcriptomics	Considers genes mapped to multiple pathways	Ignores gene expression changes, assumes gene independence	[58]
GSEA	Transcriptomics	Considers gene dependence and level of gene expression change	The position of genes on the ranked list impacts enrichment scores.	[28]
TPEA	Transcriptomics	Considers network topology and gene expression changes	Requires experimental evidence for network topology and gene–gene interactions as input	[23]
PSEA	Proteomics	Considers variation between samples	Only analyzes datasets generated by label-free protein quantification	[78]
PSEA-Quant	Proteomics	Considers variation between samples, a user-friendly interface	Requires many replicates to estimate robust variation coefficients	[76]
MetPA	Metabolomics	Considers network topology and metabolite abundance	Requires metabolite annotation and experimental evidence for network topology	[41]
Mummichog	Metabolomics	More efficient than tools requiring upfront compound annotation	Lacks metabolite annotations	[40]
iDEA	scRNA-seq	Integrates differential expression and enrichment analyses to increase reproducibility of scRNA-seq analysis	Requires cell type annotation as input	[36]
RSS-E	Genome-wide association studies	Performs enrichment analysis directly with SNPs without assigning them to individual genes	Only analyzes one gene at a time	[8]
ebGSEA	Epigenomics	Considers the impact of the extent of DNA methylation changes per pathway	Ignores the impacts of CpG sites that map to multiple genes	[47]
GOmeth	Epigenomics	Handles CpG sites that map to multiple genes	Requires a threshold to select differentially methylated probes as input	[9]
ActivePathway	Integrated omics	Directly integrates *p* values from various omics data without creating high-dimensional datasets	Only integrates transcriptomic and proteomic datasets and ignores metabolomic data	[51]

[a]Abbreviations: ebGSEA, empirical Bayes gene set enrichment analysis; GOmeth, Gene Ontology testing for methylation profiles; GSEA, gene set enrichment analysis; iDEA, integrative differential expression and gene set enrichment analysis; MetPA, Metabolomics Pathway Analysis; PADOG, pathway analysis with downweighting of overlapping genes; PSEA, protein set enrichment analysis; RSS-E, regression with summary statistics enrichment analysis; scRNA-seq, single-cell RNA sequencing; TPEA, topology-based pathway enrichment analysis.

2009 annotation version [69]. Thus, using the most recent annotations in the analysis and reporting both versions and dates of software and annotation releases in publications is important for reflecting the current understanding of biology using pathway enrichment analysis.

All widely applied reference annotation databases allow annotating the same genes to multiple pathways because biological processes are often interconnected. The same genes may result in enrichment of multiple pathways, which may overestimate the number of biological processes responding to the treatment. Several methods have been developed to mitigate the overlapping gene annotation problem in enrichment analyses. A common strategy is to weight genes on the basis of the number of annotated pathways, such as pathway analysis with downweighting of overlapping genes (PADOG) for interpreting transcriptomic datasets and ebGESA for DNA methylation profiles. Representative tools that perform pathway enrichment analysis are summarized in Table 1. In summary, the results of enrichment analyses should be interpreted by acknowledging the potential inflation caused by genes that are annotated to multiple pathways.

Standardizing reporting pathway enrichment analysis methods

Interpreting the ever-growing types and quantities of omics data requires establishing best practices for performing pathway enrichment analysis and standardizing method documentation in publications. Lack of sufficient information in studies conducting enrichment analysis hampers justification of the results during the peer review process and by the general scientific community [11]. To help achieve this goal, all the major components in a pathway enrichment analysis should be clearly documented, including different gene sets used to perform enrichment analysis, how gene sets of interest were generated, how background sets were defined, and the versions of reference annotations (Figure 2).

Moreover, methodological details used in performing enrichment tests are required, including statistical tests, user-defined cutoff for significance, and multitest correction strategy (Figure 2). Failing to perform or report the adjustment of *p* values generated by statistical tests is a prevalent problem in publications describing pathway enrichment analysis [2,11,13]. Because pathway enrichment analysis simultaneously tests hundreds or thousands of pathways annotated in the selected organism, the rate of false-positive results will be dramatically inflated as the number of comparisons increases [71,72]. For example, a benchmark study evaluated the impact of multitest correction on the results of enrichment analysis using six RNA-seq datasets. Lack of *p* value adjustment resulted in 25–40% more pathways identified as significant, which dramatically hampers the accuracy of pathway enrichment analysis [11]. Thus, reporting sufficient methodological details can help justify the results of pathway enrichment analysis and improve the robustness and reproducibility of publications.

Concluding remarks and future perspectives

With the rapid advancement of sequencing technology and computational tools, the volume and complexity of big data are accumulating faster than ever before [67,73–76]. Pathway enrichment analysis is essential to interpreting omics data by identifying the predominant biological pathways driving patterns observed in the massive datasets [2,6,13,17]. To further leverage the wealth of omics resources, we will need to better document and manage both raw and processed datasets (see Outstanding questions). Well-annotated and appropriately analyzed omics datasets can improve the reproducibility of pathway enrichment analysis and facilitate tool development. Findable, accessible, interoperable, reusable (FAIR) principles are recommended to organize big datasets for publications [77,78]. In addition, emerging evidence suggests that integrating different types of omics data may increase the power of statistical analysis for discovering patterns missing in individual omics datasets [60,63]. However, integrated omics will yield massive datasets with high

Outstanding questions

What functional datasets and statistical methods are needed to benchmark publicly available tools developed for enrichment analyses?

What new methods are required to effectively integrate different types of omics datasets (e.g., transcription factor binding, epigenomics, chromatin accessibility assays) to provide a holistic view of information flow from genes to phenotypes?

What user-friendly tools and interfaces are required to interpret omics resources without extensive knowledge of bioinformatics?

dimensions, which makes data interpretation more challenging. Novel methods with sufficient computational capability are required to harness the power of data fusion. Besides method development, we need to increase the accessibility of computational tools to bench scientists to mitigate the discrepancy between data generation using omics technology and interpretation of results (see Outstanding questions). One possible solution is to develop more web applications and user-friendly open-source packages to enable biologists to perform large-scale analysis using different types of omics datasets [47,63]. The growing list of genome-scale resources can help guide traditional molecular genetic and biochemical studies to facilitate sustainable agriculture, drug discovery, and therapeutic innovation.

Acknowledgments

This work was supported, in part, by U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program, grant nos. DE-SC0018277, DE-SC0023160, DE-SC0020366, and DE-SC0021286 and by U.S. National Science Foundation grants DBI-2213983, MCB-1916797, MCB-2052590, and IOS-1546838. We thank the Plant Cell Atlas (plantcellatlas.org) community and members of the Rhee lab (rheelab.org) for helpful discussions, particularly Dr Karine Prado, Dr Charles Hawkins, and Elena Del Pup. This work was done on the ancestral land of the Muwekma Ohlone Tribe, which was and continues to be of great importance to the Ohlone people.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Perez-Riverol, Y. *et al.* (2019) Quantifying the impact of public omics data. *Nat. Commun.* 10, 3512
2. Mubeen, S. *et al.* (2022) On the influence of several factors on pathway enrichment analysis. *Brief. Bioinform.* 23, bbac143
3. Khatri, P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8, e1002375
4. Mishra, P. *et al.* (2014) Gene set analysis: limitations in popular existing methods and proposed improvements. *Bioinformatics* 30, 2747–2756
5. Tamayo, P. *et al.* (2016) The limitations of simple gene set enrichment analysis assuming gene independence. *Stat. Methods Med. Res.* 25, 472–487
6. Reimand, J. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* 14, 482–517
7. Yoon, S. *et al.* (2018) Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Res.* 46, e60
8. Zhu, X. and Stephens, M. (2018) Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* 9, 4361
9. Maksimovic, J. *et al.* (2021) Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biol.* 22, 173
10. Chawla, S. *et al.* (2021) UniPath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic Acids Res.* 49, e13
11. Wijesooriya, K. *et al.* (2022) Urgent need for consistent standards in functional enrichment analysis. *PLoS Comput. Biol.* 18, e1009935
12. Longo, S.K. *et al.* (2021) Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* 22, 627–644
13. Geistlinger, L. *et al.* (2021) Toward a gold standard for benchmarking gene set enrichment analysis. *Brief. Bioinform.* 22, 545–556
14. Liu, L. *et al.* (2017) Pathway enrichment analysis with networks. *Genes* 8, 246
15. Zhang, Y. *et al.* (2020) Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.* 18, 2953–2961
16. Nguyen, T.-M. *et al.* (2019) Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.* 20, 203
17. Mi, H. *et al.* (2019) Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* 14, 703–721
18. Maleki, F. *et al.* (2020) Gene set analysis: challenges, opportunities, and future research. *Front. Genet.* 11, 654
19. Ihnatova, I. *et al.* (2018) A critical comparison of topology-based pathway analysis methods. *PLoS One* 13, e0191154
20. Das, S. *et al.* (2020) Fifteen years of gene set analysis for high-throughput genomic data: a review of statistical approaches and future challenges. *Entropy (Basel)* 22, 427
21. Mathur, R. *et al.* (2018) Gene set analysis methods: a systematic comparison. *BioData Min.* 11, 8
22. Zyla, J. *et al.* (2017) Ranking metrics in gene set enrichment analysis: do they matter? *BMC Bioinformatics* 18, 256
23. Bayerlová, M. *et al.* (2015) Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics* 16, 334
24. Yang, Q. *et al.* (2019) Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Brief. Bioinform.* 20, 168–177
25. Morozova, O. *et al.* (2009) Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet.* 10, 135–151
26. Lowe, R. *et al.* (2017) Transcriptomics technologies. *PLoS Comput. Biol.* 13, e1005457
27. Siavoshi, A. *et al.* (2022) Gene expression profiles and pathway enrichment analysis to identification of differentially expressed gene and signaling pathways in epithelial ovarian cancer based on high-throughput RNA-seq data. *Genomics* 114, 161–170
28. Kanehisa, M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361
29. Jung, S.-H. (2014) Stratified Fisher's exact test and its sample size calculation. *Biom. J.* 56, 129–140
30. 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
31. Maciejewski, H. (2013) Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.* 15, 504–518
32. Ma, J. *et al.* (2019) A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics* 20, 546

33. Wu, X. *et al.* (2014) Pathway and network analysis in proteomics. *J. Theor. Biol.* 362, 44–52
34. Schölz, C. *et al.* (2015) Avoiding abundance bias in the functional annotation of posttranslationally modified proteins. *Nat. Methods* 12, 1003–1004
35. Fu, X. *et al.* (2008) Spectral index for assessment of differential protein expression in shotgun proteomics. *J. Proteome Res.* 7, 845–854
36. Cha, S. *et al.* (2010) In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. *Mol. Cell. Proteomics* 9, 2529–2544
37. Lavallée-Adam, M. *et al.* (2014) PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. *J. Proteome Res.* 13, 5496–5509
38. Wieder, C. *et al.* (2021) Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput. Biol.* 17, e1009105
39. Marco-Ramell, A. *et al.* (2018) Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics* 19, 1
40. Li, S. *et al.* (2013) Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* 9, e1003123
41. Xia, J. and Wishart, D.S. (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* 26, 2342–2344
42. Ogbeide, S. *et al.* (2022) Into the multiverse: advances in single-cell multiomic profiling. *Trends Genet.* 38, 831–843
43. Cole, B. *et al.* (2021) Plant single-cell solutions for energy and the environment. *Commun. Biol.* 4, 962
44. Hwang, B. *et al.* (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* 50, 1–14
45. Wang, J. *et al.* (2018) Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 115, E6437–E6446
46. Vento-Tormo, R. *et al.* (2018) Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353
47. Ma, Y. *et al.* (2020) Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nat. Commun.* 11, 1585
48. Tam, V. *et al.* (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484
49. Uffelmann, E. *et al.* (2021) Genome-wide association studies. *Nat. Rev. Methods Primers* 1, 59
50. Manolio, T.A. (2010) Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363, 166–176
51. Marees, A.T. *et al.* (2018) A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods Psychiatr. Res.* 27, e1608
52. White, M.J. *et al.* (2019) Strategies for pathway analysis using GWAS and WGS data. *Curr. Protoc. Hum. Genet.* 100, e79
53. Holden, M. *et al.* (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 24, 2784–2785
54. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33, 245–254
55. Gibney, E.R. and Nolan, C.M. (2010) Epigenetics and gene expression. *Heredity* 105, 4–13
56. O'Geen, H. *et al.* (2011) Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods Mol. Biol.* 791, 265–286
57. Nakato, R. and Sakata, T. (2021) Methods for ChIP-seq analysis: a practical workflow and advanced applications. *Methods* 187, 44–53
58. Phipson, B. *et al.* (2016) missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* 32, 286–288
59. Dong, D. *et al.* (2019) ebGSEA: an improved gene set enrichment analysis method for epigenome-wide-association studies. *Bioinformatics* 35, 3514–3516
60. Misra, B.B. *et al.* (2019) Integrated omics: tools, advances, and future approaches. *J. Mol. Endocrinol.* 62, R21–R45
61. Karczewski, K.J. and Snyder, M.P. (2018) Integrative omics for health and disease. *Nat. Rev. Genet.* 19, 299–310
62. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020) Pan-cancer analysis of whole genomes. *Nature* 578, 82–93
63. Paczkowska, M. *et al.* (2020) Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.* 11, 735
64. Chicco, D. and Agapito, G. (2022) Nine quick tips for pathway enrichment analysis. *PLoS Comput. Biol.* 18, e1010348
65. Timmons, J.A. *et al.* (2015) Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol.* 16, 186
66. Zhao, K. *et al.* (2021) A novel bivalent chromatin associates with rapid induction of camalexin biosynthesis genes in response to a pathogen signal in Arabidopsis. *eLife* 10, e69508
67. Zhao, K. and Rhee, S.Y. (2022) Omics-guided metabolic pathway discovery in plants: resources, approaches, and opportunities. *Curr. Opin. Plant Biol.* 67, 102222
68. Karp, P.D. *et al.* (2021) Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics* 22, 191
69. Wadi, L. *et al.* (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* 13, 705–706
70. Tomczak, A. *et al.* (2018) Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci. Rep.* 8, 5115
71. Jafari, M. and Ansari-Pour, N. (2019) Why, when and how to adjust your P values? *Cell J.* 20, 604–607
72. Altman, N. and Krzywinski, M. (2016) P values and the search for significance. *Nat. Methods* 14, 3–4
73. Hawkins, C. *et al.* (2021) Plant Metabolic Network 15: a resource of genome-wide metabolism databases for 126 plants and algae. *J. Integr. Plant Biol.* 63, 1888–1905
74. Zhao, K. and Bartley, L.E. (2014) Comparative genomic analysis of the R2R3 MYB secondary cell wall regulators of Arabidopsis, poplar, rice, maize, and switchgrass. *BMC Plant Biol.* 14, 135
75. Chen, B. *et al.* (2020) Harnessing big 'omics' data and AI for drug discovery in hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.* 17, 238–251
76. Leonelli, S. (2019) The challenges of big data biology. *eLife* 8, e47381
77. Fischer, M. and Hoffmann, S. (2022) Synthesizing genome regulation data with vote-counting. *Trends Genet.* 38, 1208–1216
78. Wilkinson, M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018