OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–19
https://doi.org/10.1093/bib/bbac435
Advance access publication date 17 October 2022
Review

A comprehensive survey of the approaches for pathway analysis using multi-omics data integration

Zeynab Maghsoudi, Ha Nguyen, Alireza Tavakkoli and Tin Nguyen
Corresponding author: Tin Nguyen, Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA. Tel.: +1-775-784-6619; Email: tinn@unr.edu

Abstract

Pathway analysis has been widely used to detect pathways and functions associated with complex disease phenotypes. The proliferation of this approach is due to better interpretability of its results and its higher statistical power compared with the gene-level statistics. A plethora of pathway analysis methods that utilize multi-omics setup, rather than just transcriptomics or proteomics, have recently been developed to discover novel pathways and biomarkers. Since multi-omics gives multiple views into the same problem, different approaches are employed in aggregating these views into a comprehensive biological context. As a result, a variety of novel hypotheses regarding disease ideation and treatment targets can be formulated. In this article, we review 32 such pathway analysis methods developed for multi-omics and multi-cohort data. We discuss their availability and implementation, assumptions, supported omics types and databases, pathway analysis techniques and integration strategies. A comprehensive assessment of each method's practicality, and a thorough discussion of the strengths and drawbacks of each technique will be provided. The main objective of this survey is to provide a thorough examination of existing methods to assist potential users and researchers in selecting suitable tools for their data and analysis purposes, while highlighting outstanding challenges in the field that remain to be addressed for future development.

Keywords: integrative pathway analysis, multi-omics integration, multi-cohort analysis, pathway graph transformation.

Introduction

Due to the prevalence of microarray and RNA-seq data in early stages, many pathway analysis methods have been developed for the analysis of gene expression data. The first generation are the over-representation analysis (ORA) methods [1–8]. These methods take a list of differentially expressed (DE) genes as input and identify the pathways in which the DE genes are over- or under-represented. The second generation are the Functional Class Scoring (FCS) methods [9–14]. FCS methods eliminate dependency on gene selection criteria by taking all genes into consideration. The third generation are the Topology-based (TB) methods [15–27]. These methods aim to exploit pathway topology and gene interactions that are meant to capture and describe the biological phenomenon. Yet, they still focus on gene expression analysis and are not able to analyze other types of data or integrate multiple experiments.

Integrating information from different types of data or experiments has become increasingly more essential to obtain a comprehensive overview of biological systems [28, 29]. The ability to analyze together independent studies allows researchers to increase sample size and reproducibility, while integrating different types of assays will increase the ability to uncover the complete cause of diseases and their functional consequences, which no single type of assay would be able to provide. This has led to the introduction of many techniques that utilize multi-omics and multi-cohort data to identify the pathways that underly the conditions or phenotypes of interest.

In this article, we review 32 approaches that have been designed for the purpose of integrative pathway analysis in the multi-omics and/or multi-cohort setup. Figure 1 shows the timeline of integrative pathway analysis techniques developed between 2005 and 2022. The availability of many methods and their continuous development indicate the high demand for pathway analysis using multi-omics data in the research community over the past 17 years. We note that there exist many other useful tools for pathway analysis. For example, LncSEA [30], LnCompare [31] and DIANA-miRPath [32] can perform pathway analysis using noncoding RNA data. However, these methods are not within the scope of this article because they are not designed for data integration (multi-omics and/or multi-cohort integration).

Zeynab Maghsoudi is a PhD student in the Department of Computer Science and Engineering at the University of Nevada, Reno, Nevada, USA. Her research foci include pathway analysis and multi-omics data integration.
Ha Nguyen is a PhD student in the Department of Computer Science and Engineering at the University of Nevada, Reno, Nevada, USA. His research foci include single-cell data deconvolution and systems-level analysis.
Alireza Tavakkoli is an Associate Professor in the Department of Computer Science and Engineering at the University of Nevada, Reno (UNR), Nevada, USA. His research focuses on developing deep machine learning and statistical methods for a variety of domains, including medicine, biology, behavioral science and computational neuroscience.
Tin Nguyen is an Assistant Professor in the Department of Computer Science and Engineering at the University of Nevada, Reno (UNR), Nevada, USA. His research focuses on developing machine learning and statistical methods that can be applied to bioinformatics and computational biology.
Received: April 28, 2022. **Revised:** August 26, 2022. **Accepted:** September 8, 2022

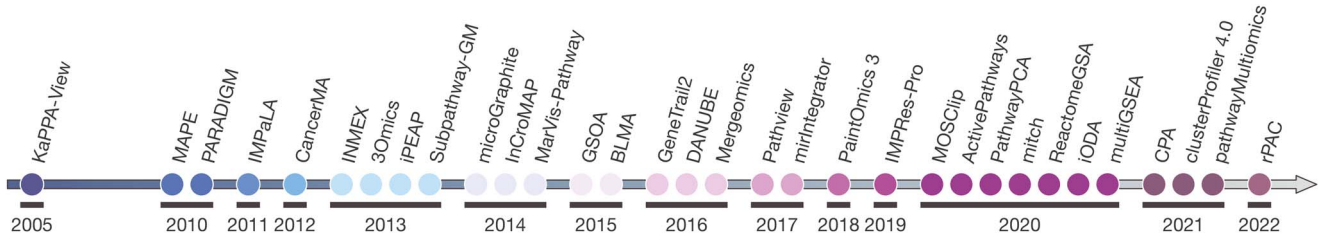

Figure 1. Timeline of the approaches developed for pathway analysis and multi-omics integration.

This will be the first article that provides such an in-depth review and discussion about integrative pathway analysis approaches using multi-omics and multi-cohort data. Many review articles [33–36] provide comprehensive assessments of pathway and subnetwork identification methods, but they are limited to the analysis of a single type of data, i.e. mRNA or scRNA-seq data. Other surveys [37–41] review techniques developed for integrating multi-omics data in general, without discussing how these can be applied in the context of pathway analysis. In contrast, here, we provide a comprehensive review of prominent methods capable of integrating multi-cohort and multi-omics data for the purpose of pathway analysis. Our survey is expected to benefit life scientists who wish to choose a method that is most suitable for their data. At the same time, this survey also benefits computational scientists who wish to know the limitations of existing methods in order to develop new methods and infrastructure addressing the current shortcomings.

In *Section Availability and Implementation* we will describe the methods' availability, their implementation and important features. In *Section Pre-analysis* we will present the steps each method follows prior to pathway analysis. In *Section Integrative Pathway Analysis* we will describe and categorize these methods according to the techniques they employ for pathway analysis and data integration while presenting their advantages and disadvantages. In *Section Method Assessment* we will assess the quality of the methods based on six different metrics. Finally, in *Section Summary and Discussion*, we will highlight outstanding challenges in the field that remain to be addressed for further enhancements. We also provide detailed description, analysis pipeline and practicality of each method in supplementary materials (SupplementaryNote.pdf, Supplementary-Table-S1.xlsx, Supplementary-Table-S2.xlsx).

Availability and implementation

Tables 1 and 2 provide a summary of the availability and important features of each of the 32 surveyed methods. Table 1 shows available hyperlinks, platforms, availability, year, number of citations (Google Scholar) and references to their publications. Among the 32 methods, R packages and web-based tools are the most common platforms used for integrative pathway analysis. Moreover, the web interface is usually intuitive and provides options for selecting the input, analyzing the data and visualizing the output. Three of the surveyed tools, i.e. CancerMA [42], DANUBE [43] and rPAC [44], are not available. Regardless, we include these approaches in the survey because we believe that understanding these methods will be beneficial to readers.

Table 2 shows the features that are currently supported by the tools: supported data types, pathway databases embedded in the tools, ID mapping (built-in/user-provided) and interactive visualization. We group the methods into four categories based on their analysis techniques and purposes: (i) gene-level, *P*-value-based integration, (ii) pathway-level, *P*-value-based integration, (iii) graph transformation-based and (iv) machine-learning-based analysis. The categories are discussed in detail in *Section Integrative Pathway Analysis*. All of the reviewed tools are capable of analyzing transcriptomics as gene expression data are the most prevalent type of available omics data. Many of the methods are capable of integrating transcriptomics with other types of data, including genomic (e.g. SNP, copy number variation), epigenomic, proteomic and metabolomic data. In addition to multi-omics integration, there are 19 methods that allow users to integrate multiple datasets/experiments (meta-analysis). Methods with interactive visualization capability provide the diagrammatic representation of omics and interactions, as well as the quantitative analysis results on their graphical user interface. Users can interactively explore the analysis results, gene-level statistics and pathway networks, and choose the set of impacted pathways based on their domain knowledge and expert opinion.

Regarding pathway knowledge, 22 methods embed pathways from known pathway databases in their software. Embedded databases include KEGG [74], GO [75], Reactome [76], STRING [77], MirTarBase [78], NCI-PID [79], Biocarta [80], PharmGKB [81], Wikipathways [82], TRRUST [83], BioCyc [84], NetPath [85], INOH [86], EHMN [87], SMPDB [88] and SignaLink [89]. For these tools, users can conveniently use the built-in pathway information for their analysis without the need of providing pathways as part of the input. As each database uses different gene/compound IDs, it is important that the molecular data are mapped to the same IDs as the gene/compound IDs in the pathways. Such mapping can be either a user-imported mapping file or based on a built-in mapping database. Many methods also allow users to import their own pathway information (mostly in the format of gene sets). Clearly, the built-in mapping database is preferable since it relieves users from technical details and mapping knowledge.

High-level workflow

Figure 2 shows the high-level workflow and overall pipeline of integrative pathway analysis methods. These methods typically consist of two analysis phases: (i) data processing and standardization (pre-analysis), and (ii) data integration and pathway analysis (integrative pathway analysis).

During the pre-analysis phase, the input data are processed to become suitable for integrative pathway analysis. Pre-analysis includes an ID mapping module that maps the input omics features to the gene/compound IDs presented in the pathways. Next, a data filtering module is designed to perform data quality control and noise reduction. Two optional modules can be also provided: data matching and pathway augmentation. In the data matching module, all expression data are matched so that the final dataset contains either common genes or samples across all original datasets. In the pathway augmentation module, validated multi-omics interactions (e.g. microRNA–gene interactions) are added to the pathway. This module is particularly important for

Table 1. Availability of 32 integrative pathway analysis approaches. The ✓ and ✗ in the Availability column indicate whether the methods are available at the time of writing this manuscript. The term 'Web' in the Platform column indicates whether the tool has a web-based graphical user interface. R, MATLAB, and Java terms indicate the programming languages used to implement the package. The three last columns, Year, Reference and Citation, indicate the year of publication, reference to the article and the number of citations

Method	URL	Platform	Avail.	Year	Ref.	Cit.
Gene-level, *P*-value-based integrative approaches						
KaPPA-View	http://kpv.kazusa.or.jp/	Web	✓	2005	[45]	175
MAPE	https://cran.r-project.org/web/packages/MetaPath/	R	✓	2010	[46]	98
CancerMA	NA	Web	✗	2012	[42]	34
INMEX	https://www.networkanalyst.ca/	Web	✓	2013	[47]	166
3Omics	https://3omics.cmdm.tw/	Web	✓	2013	[48]	136
IncroMAP	http://www.cogsys.cs.uni-tuebingen.de/software/InCroMAP/	Java	✓	2014	[49]	47
ActivePathways	https://github.com/reimandlab/ActivePathways	R	✓	2020	[50]	67
mitch	https://bioconductor.org/packages/release/bioc/html/mitch.html	R	✓	2020	[51]	14
iODA	http://www.sysbio.org.cn/iODA/	Java	✓	2020	[52]	3
Pathway-level, *P*-value-based integrative approaches						
IMPaLA	http://impala.molgen.mpg.de/	Web	✓	2011	[53]	322
iPEAP	https://drive.google.com/drive/folders/1w2RivUThk1uIqNOLz6BT1KVVrokWr2xp	Java	✓	2013	[54]	38
MarVis-Pathway	http://marvis.gobics.de/	MATLAB	✓	2014	[55]	75
BLMA	https://bioconductor.org/packages/release/bioc/html/BLMA.html	R	✓	2015	[56]	26
GeneTrail2	https://genetrail2.bioinf.uni-sb.de/	Web	✓	2016	[57]	137
DANUBE	NA	R	✗	2017	[43]	21
Mergeomics	http://mergeomics.research.idre.ucla.edu/	Web/R	✓	2016	[58]	70
Pathview	https://pathview.uncc.edu/	Web/R	✓	2017	[59]	203
PaintOmics 3	http://www.paintomics.org/	Web	✓	2018	[60]	126
ReactomeGSA	https://reactome.org/	Web/R	✓	2020	[61]	55
multiGSEA	https://bioconductor.org/packages/release/bioc/html/multiGSEA.html	R	✓	2020	[62]	20
pathwayMultiomics	https://github.com/TransBioInfoLab/pathwayMultiomics	R	✓	2021	[63]	1
CPA	http://cpa.tinnguyen-lab.com	Web	✓	2021	[64]	9
clusterProfiler 4.0	https://bioconductor.org/packages/clusterProfiler/	R	✓	2021	[65]	532
Graph-transformation-based approaches						
PARADIGM	http://paradigm.five3genomics.com	Web	✓	2010	[66]	792
Subpathway-GM	http://www2.uaem.mx/r-mirror/web/packages/iSubpathwayMiner/	R	✓	2013	[67]	110
microGraphite	http://romualdi.bio.unipd.it/micrographite	R	✓	2014	[68]	45
mirIntegrator	http://bit.ly/mirIntegrator/	R	✓	2017	[69]	15
MOSClip	https://cavei.github.io/MOSClip/	R	✓	2019	[70]	10
IMPRes-Pro	http://digbio.missouri.edu/impres	Web	✓	2020	[71]	5
rPAC	NA	NA	✗	2022	[44]	3
Machine-learning-based approaches						
GSOA	https://bitbucket.org/srp33/gsoa/src/master/	R	✓	2015	[72]	13
pathwayPCA	https://bioconductor.org/packages/release/bioc/html/pathwayPCA.html	R	✓	2019	[73]	6

methods that take into consideration interactions among genes and multi-omics layers.

The integrative pathway analysis phase is typically for identifying pathways that are significantly enriched. Most methods perform differential analysis for each data type or dataset and then combine them at the gene- or pathway-level to obtain an overall enrichment score for each pathway. These scores are often compared against the null distributions to obtain the *P*-values that represent statistical significance of the pathways.

Pre-analysis

Input data processing and filtering

The input of the surveyed methods includes: (i) omics expression matrices in .CSV format, in which rows represent genes/markers and columns represent samples/patients, and (ii) pathways matrix in the GMT format if the pathways are not embedded in the software. The analysis begins with quality control and data filtering.

The quality control step typically checks for the consistency of class labels across all datasets, the validity of gene/compound IDs and the correctness of data format. Some methods also detect outliers or perform additional checks depending on the type of omics data they support. The data filtering step aims at reducing the number of irrelevant features. To reduce the number of omics markers, most methods remove data points with low quality. The majority of methods also remove pathways that have too few genes measured. The three methods, CancerMA [42], GSOA [72] and MOSClip [70], also perform additional data filtering to suit their purpose. CancerMA only keeps cancer-related genes because it is designed for system-level analysis of 80 cancer datasets. GSOA and MOSClip filter out genes that do not belong to the pathways/conditions of interest.

Identifier (ID) mapping

At the abstract level, a pathway consists of multiple genes that work together to achieve a certain biological function. The challenge is that the same gene has different IDs in assaying

Table 2. Comparison of the 32 surveyed methods in terms of the supported omics types, meta-analysis, embedded pathways databases and mapping unit. The majority of these methods can integrate transcriptomic data with genomic, epigenomic, proteomic and metabolomic data. Among these, 19 methods are capable of performing meta-analysis (multi-cohort analysis). Further, 22 have embedded databases and thus allow users to analyze their data without the need of providing pathway/gene set information. Most methods also automatically map the gene/compound IDs of the expression data to the gene/compound IDs presented in the pathways. The column 'Inter. Visual.' indicates whether the tool has an interactive interface. Six methods provide interactive pathway diagram that allow users to interactively explore the results and inspect the statistics obtained in each omics analysis

Method	Type of omic data					Meta-analysis	Built-in Database	ID Mapping		Inter. Visual.
	Genome	Transcriptome	Epigenome	Proteome	Metabolome			Manual	Built-in	
Gene-level, *P*-value-based integrative approaches										
KaPPA-View		✓			✓		KEGG, BioCyc		✓	✓
MAPE		✓				✓		✓		
CancerMA		✓				✓	GO		✓	
INMEX		✓			✓	✓	KEGG, GO		✓	
3Omics		✓		✓	✓		KEGG, BioCyc		✓	✓
InCroMAP	✓	✓		✓	✓	✓	KEGG, Reactome, BioCarta		✓	✓
ActivePathways	✓	✓	✓	✓	✓	✓		✓		
mitch		✓	✓	✓		✓		✓		
iODA	✓	✓	✓			✓	KEGG		✓	
Pathway-level, *P*-value-based integrative approaches										
IMPaLA		✓		✓	✓		KEGG, Reactome, WikiPathways, NCI-PID, Biocarta, PharmaGKBB, INOH, EHMN, NetPath, SMPDB, BioCyc	✓		
iPEAP	✓	✓		✓	✓		KEGG		✓	
MarVis-Pathway		✓			✓	✓	KEGG, BioCyc		✓	
BLMA		✓				✓	KEGG	✓		
GeneTrail2	✓	✓	✓	✓		✓	KEGG, Reactome, WikiPathways, BioCarta, NCI-PID, PharmGKB, SignaLink, GO, SMPDB		✓	
DANUBE		✓				✓		✓		
Mergeomics	✓	✓	✓	✓		✓		✓		
Pathview	✓	✓			✓		KEGG	✓	✓	✓
PaintOmics 3	✓	✓	✓	✓	✓	✓	KEGG	✓	✓	✓
ReactomeGSA		✓		✓		✓	Reactome		✓	
multiGSEA		✓		✓	✓		KEGG, Reactome, PharmGKB, BioCyc, SMPDB, Panther, Biocarta, NCI-PID	✓	✓	
pathwayMultiomics	✓	✓	✓	✓	✓	✓		✓		
CPA		✓				✓	KEGG, GO		✓	✓
clusterProfiler 4.0		✓	✓			✓	KEGG, GO, WikiPathways	✓	✓	
Graph-transformation-based approaches										
PARADIGM	✓	✓		✓			NCI-PID		✓	
Subpathway-GM		✓			✓		KEGG		✓	
microGraphite		✓	✓				KEGG, Reactome, NCI-PID, BioCarta		✓	
mirIntegrator		✓	✓				KEGG, miRTarBase		✓	
MOSClip	✓	✓	✓						✓	
IMPRes-Pro		✓		✓			KEGG, STRING, TRRUST	✓	✓	
rPAC		✓				✓		✓		
Machine-learning-based approaches										
GSOA	✓	✓	✓	✓	✓			✓		
PathwayPCA		✓		✓		✓		✓		

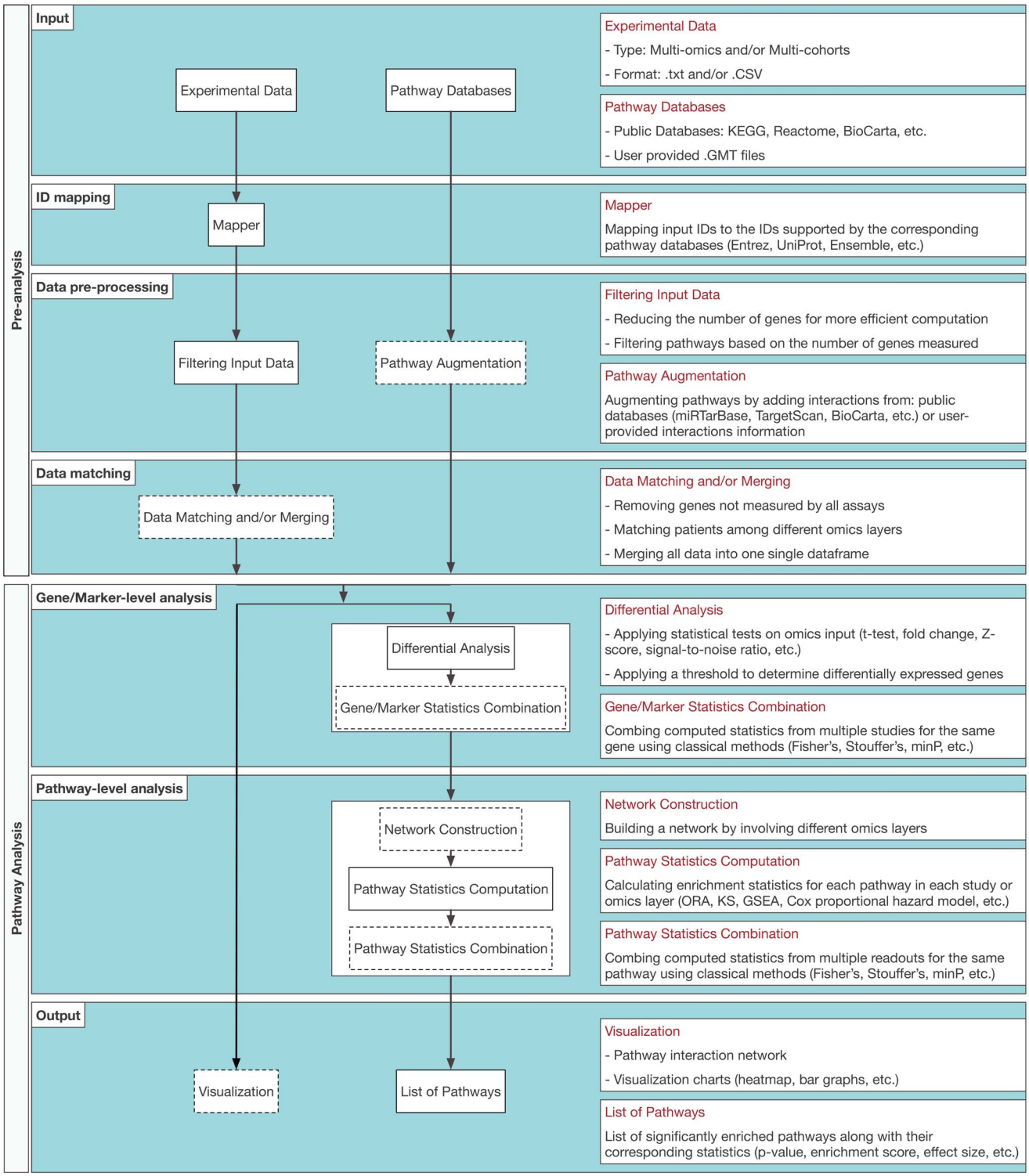

Figure 2. The high-level workflow of the surveyed approaches for integrative pathway analysis. The modules commonly present in every analysis pipeline are depicted as solid-line boxes. The dashed boxes are optional modules. The analysis process starts with processing input data (multiple multi-omics datasets). The methods first perform an integrity check to ensure the consistency among the input matrices and then map the gene/compound IDs to supported IDs of pathway databases. After pre-analysis, the methods perform differential analysis at the gene-level and then identify significant pathways using statistical hypothesis testing. The output often includes pathway *P*-values, enrichment scores and visualization plots.

platforms, pathway databases and omics layers. Crucially, ID mapping aims to integrate multi-omics data with pathway information available from different databases. Most methods perform ID mapping by either requiring users to import a mapping file, automatically mapping the IDs using a built-in mapping database, or providing both options. A complete description of supported databases and mapping IDs of all methods is provided in Supplementary Table S1.

Data matching

After importing input datasets, an optional data matching process can be employed to match or integrate the input expression

matrices. Five methods implement this module in their pre-analysis phase: CancerMA [42], mitch [51], MarVis-Pathway [55], MOSClip [70], GSOA [72]. CancerMA and mitch match the gene IDs across all expression matrices, whereas MarVis-Pathway, MOSClip, and GSOA match the sample/patient ID across all input matrices.

Pathway augmentation

The ID mapping step often leads to a one-to-many scenario where multiple IDs can represent the same gene in a pathway. For example, one single gene can code for several distinct proteins due to the alternative RNA splicing process during gene expression. Similarly, an miRNA might target several genes, or multiple miRNAs target a gene. Integrative pathway analysis methods address this challenge by either: (i) treating the duplicated gene IDs as individual data points or (ii) choosing the optimal data point among all candidates. Regardless, both solutions result in substantial information loss. To alleviate this issue, mirIntegrator [69] and microGraphite [68] augment the pathways to include all entities instead of merging them to a single gene node. Specifically, they utilize validated/predicted microRNA/target interactions and add each interaction to a pathway if the microRNA targets a gene of the pathway. The augmentation process leads to a more comprehensive representation of biological pathways and interaction networks.

Multi-cohort analysis

Over the years, a research group might have collected many datasets from different sets of samples (or patient cohorts) for a specific condition/disease. As technologies change, the assaying platforms and data types of each cohort may differ. One example scenario is as follows: (i) first dataset/cohort has gene expression, (ii) second dataset has gene expression and methylation and (iii) third cohort has gene expression and copy number variation. Even for the same omic type, the data can be generated from different platforms, including microarray and sequencing technologies. Depending on the analysis purpose, users can choose an appropriate method based on the information provided in Table 2.

The gene-level integrative approaches often combine the *P*-values of genes obtained from multiple datasets before performing pathway analysis. These methods typically remove genes that do not appear in all experiments. In other words, these methods are suitable only when the genes/features are consistently measured in all datasets. Note that *P*-value is a standardized and scaled metric, i.e. *P*-values are uniformly distributed between 0 and 1 under the null hypothesis regardless of data scaling and assaying platforms. Therefore, these methods can integrate data of different platforms as long as the *P*-values are computed correctly in each dataset and that each dataset measures a similar set of markers/genes. Table 2 also shows the types of data each method was designed and tested on. In particular, MAPE and CancerMA were designed for gene expression, while INMEX, InCroMAP, ActivePathways, mitch and iODA can analyze other types of data. Note that the two methods, KaPPA-View and 3Omics, in this category are not capable of multi-cohort analysis.

When different cohorts measure different sets of markers, users can choose among the approaches that integrate the data at the pathway-level: 10 meta-analysis methods in the pathway-level integration category (MarVis-Pathway, BLMA, GeneTrail2, DANUBE, Mergeomics, PaintOmics 3, ReactomeGSA, pathwayMultiomics, CPA and clusterProfiler 4.0), plus rPAC and PathwayPCA. These methods first analyze each dataset independently to calculate the *P*-values of each pathway and then combine the statistics (e.g. *P*-values) of the pathways across multiple cohorts. As the set of pathways is the same in each analysis, the integration at the pathway-level can almost always be performed regardless of the data type and assaying platforms being used. Compared with gene-level integration, pathway-level integrative approaches are more flexible as they do not require each dataset to measure similar markers. However, as they focus on pathway-level integration, these approaches are not capable of identifying the genes/markers that are consistently differentially expressed across the cohorts. The pros and cons of each category will be further discussed in the next sections.

Integrative pathway analysis

The pre-analysis phase prepares the data for the integrative pathway analysis phase. The 32 methods utilize different integrative techniques and hypothesis testing methods to identify pathways that are significantly different between the compared phenotypes. In the following text, we will provide a comprehensive review of these methods as follows: (i) providing a high-level description of each method, (ii) categorizing them according to their analysis techniques and purposes and (iii) providing a high-level assessment of the strategic advantages and disadvantages of each category.

At the high-level, integrative pathway analysis consists of two main steps: (i) computing the summary statistics for the genes/-markers, and (ii) performing hypothesis testing at the pathway-level to identify pathways that are significantly different between two phenotypes (e.g. disease versus control, or treated versus untreated). In the first step, these methods perform differential analysis to compute summary statistics for omics features in the experiments [90]. In the second step, integrative pathway analysis methods compute a summary statistic from omics scores of genes belonging to a pathway. Next, a statistical significance test is performed to assess how likely the computed statistic is observed by chance (*P*-value). Therefore, an enrichment score and a *P*-value for each pathway are obtained at this step, showing whether a pathway is associated with a disease phenotype. Depending on the assumption and strategy of each method, data integration (of multiple cohorts and multiple omics types) is performed in the first or second step.

Overall, the surveyed methods are divided into four categories based on their strategy used for analysis and data integration: (i) gene-level, *P*-value-based integrative approaches, (ii) pathway-level, *P*-value-based integrative methods, (iii) graph-transformation-based techniques and (iv) machine-learning-based approaches. Characteristics of each category will be described in the following sections.

Gene-level, *P*-value-based integrative approaches

Figure 3 shows the overall pipeline of these methods. Methods in this category include KaPPA-View [45], MAPE [46], CancerMA [42], INMEX [47], 3Omics [48], InCroMAP [49], ActivePathways [50], mitch [51] and iODA [52]. Consider a readout as the measurement/expression of a single omics data in a dataset (e.g. gene expression data of an experiment/dataset). The input of integrative pathway analysis approaches typically includes multiple readouts (from multiple datasets of different omics types). In each readout, rows typically represent genes/features, whereas columns represent samples. The samples are divided into two groups: disease versus control (phenotypes). Approaches in this category start by analyzing each readout independently to compute the *P*-value and effect size of each gene. For each gene, these

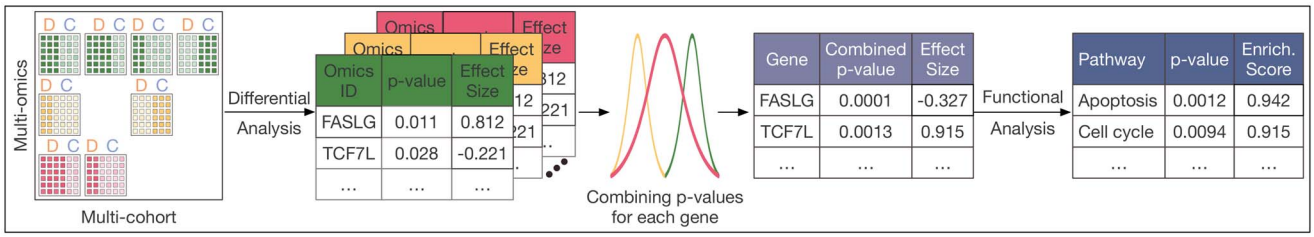

Figure 3. Overall pipeline of gene-level, *P*-value-based integrative approaches. The input includes multi-omics and/or multi-cohort data that compare two phenotypes. Methods in this category first analyze each readout independently to obtain the gene-level *P*-values and effect sizes (e.g. log fold-change). For each gene, the methods combine the *P*-values and effect sizes across multiple readouts to obtain the summary *P*-value and effect size of the gene. Finally, these approaches perform functional analysis using the summary *P*-values and statistics to identify pathways that are significantly different between the two phenotypes. The output of these methods typically include the *P*-values and enrichment scores of the pathways.

methods combine the *P*-values and statistics across all readouts to compute a summary *P*-value and statistics that represent the overall difference of the gene between the two phenotypes. These summary statistics of the genes then serve as input for enrichment methods to calculate the *P*-values and enrichment scores of pathways.

There are two main techniques that can be used to combine the *P*-values: quantile-based combination and ranked-based combination. Given multiple *P*-values for a gene p_i-s, quantile-based methods often transform the *P*-values into distributional quantiles [91], e.g. $q_i = F^{-1}(p_i)$. These methods then combine the quantiles to obtain the aggregated quantile, e.g. $Q = \sum_k q_i$. The meta *P*-value is then computed by comparing the observed Q-value against its empirical or theoretical distribution. In contrast, ranked-based methods first order the *P*-values, i.e. $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(k)}$, and then combine them using rank aggregation methods [92–95].

KaPPA-View allows users to visualize metabolomics and transcriptomics to analyze plant metabolic pathways. The software displays quantitative data changes for individual transcripts and metabolites between different experimental conditions on the same metabolic pathway maps. Furthermore, gene-to-gene and/or metabolite-to-metabolite relationships such as co-expression correlations of genes can be displayed as edges on a metabolic pathway graph.

MAPE performs meta-analysis by combining the *P*-values obtained from each input mRNA datasets. The combination is performed at both the gene-level (MAPE-G) and the pathway-level (MAPE-P). MAPE-G first computes the t-statistic for each gene in each study and then calculates the *P*-value using permutation test. For each gene, MAPE-G then combines the individual *P*-values across multiple studies using the MaxP algorithm [93] to obtain one single *P*-value for the gene. Finally, given the combined *P*-values and statistics obtained from each gene, MAPE-G then performs a Kolmogorov–Smirnov (KS) test to calculate the *P*-value for each pathway. In contrast to MAPE-G, MAPE-P combines the *P*-values at the pathway-level. In each study, MAPE-P first computes the t-statistic for each gene and then calculates the *P*-value for the gene using the permutation test. It then calculates the *P*-values for the pathways using the KS-test. MAPE-P then combines the *P*-values of a pathway across all studies using MaxP. Finally, MAPE-I combines the obtained *P*-values from MAPE-G and MAPE-P modules using MinP [92].

CancerMA was designed to analyze 80 cancer-specific gene expression datasets. This method first performs differential analysis in each dataset and then combines the *P*-values using Stouffer's method [96]. CancerMA also integrates the logFC-s using weighted linear combination [97]. By determining the differentially expressed (DE) genes based on these integrated statistics, CancerMA then performs an ORA [1] to calculate the *P*-values of GO terms.

INMEX implements differential analysis and then applies one of the five strategies to combine the computed gene *P*-values across readouts: combining *P*-values (Fisher's and Stouffer's methods), combining standardized differences, combining rank orders, vote-counting and direct data merge. After obtaining the combined *P*-values for the genes, INMEX calculates the *P*-values of the pathways using ORA or GSEA [98].

3Omics supports the analysis of proteomics, transcriptomics and metabolomics using five major modules: correlation analysis, co-expression analysis, pathway enrichment analysis and phenotypic analysis. Correlation and co-expression modules visualize connectivity and heatmaps of the two omics types. The pathway enrichment analysis component can be applied in two modes (normal and enriched). The normal mode displays user-provided metabolites via simple metabolite mapping to a pathway from the pathway databases. In the enriched mode, the tool compares two conditions to obtain the set of DE genes and enriched pathways using ORA.

InCroMAP first applies a threshold to identify the DE genes in each readout (using fold change or *P*-values). Then the intersection or union of identified DE genes are considered for pathway enrichment analysis using ORA. This tool provides two comprehensive visualizations, metabolic and pathway views. The metabolic view generates an interactive global map of cellular metabolism. The pathway view shows the integrated pathway-based data visualization from multiple omics platforms.

ActivePathways accepts a matrix of *P*-values with rows representing common genes across multiple datasets/cohort, and columns related to each omics dataset. Next, considering the dependency between omics, the *P*-values for each gene are combined across the datasets using Brown's method [99], which is an extension of Fisher's method (combining independent *P*-values). As a result, a list of combined *P*-values is generated, which is further used to identify DE genes. From the list of DE genes, this method performs a ranked ORA to compute the *P*-values for the pathways. The method also adjusts the *P*-values using the Bonferroni method. Pathways with adjusted *P*-values smaller than 5% are considered significant. The same process is also applied for each dataset to get multiple *P*-values for a pathway—one for each analysis. The output of ActivePathways is a table of pathways that are significant in at least one of the analyses. This table serves as the input of another module, named Enrichment Map [100] from Cytoscape software package [101], for the visualization of pathway analysis and data integration.

The mitch method is capable of integrating multi-cohort and multi-omics data. This method allows users to import DE genes and their statistics. Otherwise, this tool computes a

directional significance score (D) for each marker defined as: $D = -\log_{10}(P\text{-value}) \times \text{sign}(\log_2 FC)$. After ranking genes, this method divides the genes in the rank list into two groups: (i) genes that belong to the pathway and (ii) the remaining genes. For each dataset, this results in two vectors: ranking of genes belonging to the pathway and ranking of the remaining genes. Multivariate ANalysis Of VAriance (MANOVA) [102] is then applied to test for the difference between the two groups to calculate the *P*-values of the pathways. Furthermore, mitch returns an enrichment score for each pathway based on the average ranking of genes in the pathway.

iODA supports the analysis of transcriptome profiles (mRNA or miRNA expression data) and protein–DNA interactions (ChIP-Seq data). For each dataset, iODA applies six statistical methods to compute the *P*-values at the gene-level. These methods include Least Sum of Ordered Subset Squared (LSOSS) [103], Cancer Outlier Profile Analysis (COPA) [104], Maximum Ordered Subset T-statistics (MOST) [105], Outlier Robust T-statistics (ORT) [106], Outlier Sum (OS) [107] and the t-test. A gene is considered DE if four or more statistical methods report the gene as significant. The MACS [108] model is used by iODA to find significant peaks of the protein binding site in ChIP-seq data. The binding sites are then assigned to the target genes using the PeakAnalyzer [109] tool. iODA then uses two alternative strategies for pathway analysis: (i) intersecting the DE genes from all data types and then calculating the *P*-values of the pathways using ORA, or (ii) performing pathway analysis for each data type, then intersecting the significant pathways to obtain the final list.

Overall, this category provides the flexibility in adding multiple omics layers into the pathway analysis. Given that all omics markers, which match the same gene, affect the disease phenotype uniformly, the methods use well-formulated algorithms to combine the *P*-values and effect sizes from multiple readouts. Moreover, they are usually fast because pathway analysis is performed only once regardless of the number of inputs. However, they require the same set of genes across different omics datasets. Therefore, this approach may lose some important information, such as some genes may have crucial effects on pathway regulation, but they are excluded in one of the omics layers. Also, the gene statistics calculation in this approach fails to consider the topology among omics markers in the same omics layer other than transcriptomics.

Pathway-level, *P*-value-based integrative approaches

Figure 3 shows the overall workflow of pathway-level, *P*-value-based integrative approaches. Methods in this category include IMPaLA [53], iPEAP [54], MarVis-Pathway [55], BLMA [56], DANUBE [43], GeneTrail2 [57], Mergeomics [58], Pathview [59], PaintOmics 3 [60], ReactomeGSA [61], multiGSEA [62], pathwayMultiomics [63], CPA [64] and clusterProfiler 4.0 [65].

In contrast to gene-level integration, the methods in this category, for each readout, first perform differential analysis to obtain the gene-level statistics (*P*-values and effect sizes) and then perform pathway analysis using the gene-level statistics to calculate the *P*-values and enrichment scores for the pathways. Finally, for each pathway, these methods combine the *P*-values and enrichment scores from all readouts to compute a single *P*-value and enrichment score.

IMPaLA supports analyzing metabolomics and transcriptomics. This tool allows users to analyze each data type independently using ORA or Wilcoxon Enrichment Analysis (WEA) [110]. In the end, IMPaLA combines the two *P*-values into one single *P*-value by multiplying them.

iPEAP applies an existing pathway analysis method [1, 98, 111–113] on KEGG pathways for each dataset. This step returns multiple ranked lists of pathways. iPEAP then combines these ranked lists into one single rank list using the following techniques: RankAggreg, RobustRankAggre, min, median and mean.

MarVis-Pathway executes three types of enrichment analysis: entry-based, marker/feature-based and sample-based. The entry-based analysis is analogous to ORA. The marker/feature-based analysis ranks the features and then computes the pathway *P*-value using an iterative hypergeometric test [114], a rank-sum [115] or a KS test from the ranked list. Finally, sample-based analysis operates similarly to GSEA. Accordingly, the corresponding *P*-values for each pathway are returned. These *P*-values are then combined using Fisher's [116] or Stouffer's [96] method to produce a meta-*P*-value for each pathway.

BLMA was developed for the meta-analysis of multiple transcriptome datasets. The software allows users to perform pathway analysis on each dataset by multiple methods, including ORA [1], GSA [117], PADOG [118] and Impact Analysis [15]. For a chosen pathway analysis method, each pathway will therefore have multiple *P*-values—one per dataset. For each pathway, users can combine the *P*-values one of many techniques, including addCLT [56], Fisher's [116], Stouffer's [96], minP [92] or maxP [93] methods. The output of the software includes results of the data integration, as well as the results of each dataset.

DANUBE is based on the fact that pathway analysis often provides biased results toward well-studied conditions, e.g. cancer or Alzheimer's disease. This meta-analysis approach attempts to correct for method bias and integration of multiple mRNA datasets. To perform pathway analysis on each dataset, DANUBE uses one of the four techniques: GSEA, GSA, IA and PADOG. It then uses the empirical distributions to correct the *P*-values obtained for each pathway. After bias correction, each pathway has multiple *P*-values—one per dataset. Finally, DANUBE combines the resulted *P*-values using the addCLT method for each pathway.

GeneTrail2 is a web service that allows users to perform pathway analysis on each input separately, using one of these options: weighted/unweighted KS-test, Wilcoxon test, ORA, sum, mean, median, maxmean statistics and t-test. The significant pathways and their enrichment scores can be viewed online. Furthermore, for data integration, two view modes are available. The union mode displays pathways that are significant in at least one data type. The intersection mode only displays pathways that are significant in all data types.

Mergeomics consists of two independent modules: Marker Set Enrichment Analysis (MSEA) and weighted Key Driver Analysis (wKDA). MSEA focuses on identifying enriched pathways in each dataset. This approach is similar to ORA: if there are more DE genes in a pathway than what can be expected by chance, then the pathway is likely to be enriched. Chi-square-like statistics is calculated to test this assumption, and a permutation test is performed to retrieve the *P*-value for each pathway. MSEA allows users to analyze a single dataset, as well as to perform meta-analysis of multiple datasets (named 'meta-MSEA'). The meta-MSEA first calculates the *P*-values obtained from each dataset for each pathway and then combines these *P*-values using Stouffer's method [96] to obtain a single *P*-value for the pathway. wKDA focuses on identifying genes that are the potential key drivers of the enriched pathways. Mergeomics outputs enriched pathways and their potential key drivers.

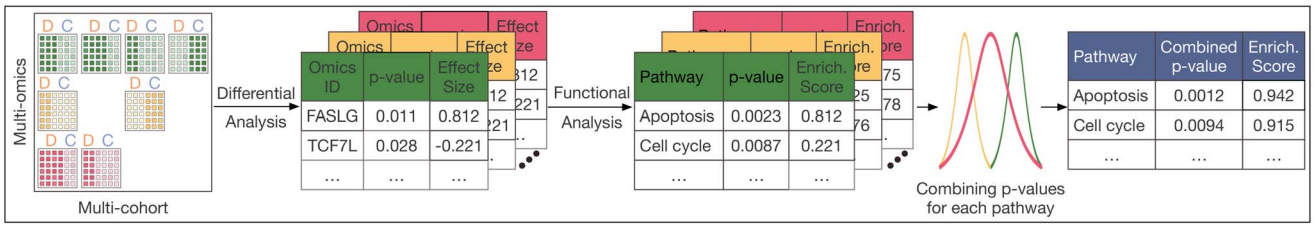

Figure 4. Overall pipeline of pathway-level, *P*-value-based integrative approaches. These methods first analyze each readout independently to calculate the *P*-values and statistics of each pathway in each readout. Next, the obtained *P*-values and statistics across all input datasets are combined to obtain the summary *P*-value and effect size for each pathway.

Pathview uses GAGE [119] (for expression data) and ORA (for compound IDs) to calculate the *P*-values of the pathways. This tool visualizes the pathway graphs with user data mapped in two views, the native KEGG pathway view and Graphviz [120] view. Graphviz view provides better control over node/edge attributes and a better view of graph topology.

PaintOmics 3 analyzes a range of different omics types: gene expression, metabolomics, region-based omics like ChIP-seq, DNase-seq, ATAC-seq, Methyl-seq, etc. and regulatory-based omics like miRNAs, transcription factors or other factors. This tool calculates the *P*-value for each pathway in each dataset using ORA and then combines the *P*-values for each pathway using Stouffer's or weighted Fisher's method. This tool provides three modules for visualizing the analysis results. In the first module, a pie chart and hierarchical structure show that KEGG pathways are organized in seven main classifications (Cellular Processes, Drug Development, Environmental Information Processing, Genetic Information Processing, Human Diseases, Metabolism and Organismal Systems). The second module is a pathways interaction network, in which the nodes represent pathways and edges indicate shared features among them. The last module allows users to explore each pathway by producing an interactive pathway diagram and a global heatmap with complementary information.

ReactomeGSA supports pathway analysis using microarray measurements, raw RNA-Seq and normalized read counts, proteomics spectral counts and intensity-based quantitative data. ReactomeGSA applies Camera [121], PADOG [118] or ssGSEA [122, 123] for pathway analysis on each data type. This tool also supports the analysis of single-cell RNA-sequencing (scRNA-seq) data. For this type of data, this tool uses either Seurat [124] or scater [125] to compute the mean expression of genes within a cluster. Then, through one of the supported analyzing methods, one pathway-level expression value per cell cluster is calculated. The obtained results are then converted to Reactome's internal data format to visualize the results. The results from different analyses can be seen and interactively explored side by side down to a single gene or protein level.

The multiGSEA software independently executes GSEA on each input data type, which can be transcriptomics, proteomics and/or metabolomics. This package allows users to utilize the pathways definition from eight different databases, including PharmGKB [81], NCI/Nature Pathway Interaction Database [126], HumanCyc [127], SMPDB [128], Panther [129], Biocarta [80], KEGG [74, 130, 131] and Reactome [76]. For each pathway, up to three adjusted *P*-values and enrichment scores are returned. The method then combines the pathway's *P*-values using Z-method, Stouffer's, Fisher's or Edgington's method [132].

The pathwayMultiomics method as input, accepts a table of pathways' *P*-values across multi-omics types. This tool, considering all combinations of pairs of *P*-values for each pathway, then picks the maximum *P*-value in each pair. Next, among the selected *P*-values, the minimum *P*-value is chosen for that specific pathway. To assess the statistical significance of chosen MinMax statistic, pathwayMultiomics assumes that all input *P*-values are independent, and the MinMax statistic for each pathway follows a Beta distribution for the rth order statistic.

CPA is a web-based platform that supports multiple inputs that can be analyzed with multiple methods in a single analysis [9, 14, 15, 110, 118, 133–135]. With input is a list of DE genes, ORA/WebGestalt [136] will be used. With input is a pre-ranked gene list (e.g. gene list with fold change), FGSEA [134], KS-test [133] and Wilcox-test [137] are the available methods. With input as an expression matrix, eight methods, including GSEA [9], GSA [14], FGSEA, PADOG [118], Impact Analysis (IA) [15, 16], ORA/WebGestalt, KS-test, and Wilcox-test, can be used. This tool provides an interactive visualization of the analysis results in both pathway- and gene-level graphs. At the pathway-level, a pathway is a node sliced into multiple parts corresponding to the analysis results. At the gene-level, a node is a gene with multiple colored parts representing the regulation direction of the gene in each data set.

The last method, clusterProfiler 4.0, allows users to perform both multi-omics integration (transcriptome and epigenome) and multi-cohort analysis. The input of the method includes multiple lists of genes and their statistics obtained from one or multiple transcriptome/epigenome datasets. The software provides embedded pathways information from KEGG, GO and WikiPathways, but it also allows users to input their customized pathways. For each dataset, the method performs pathway analysis using either ORA or GSEA. For each analysis, the method adjusts the *P*-values for multiple comparisons using one of the following methods: Holm's, Bofferoni's, Hochberg's, Hommel's, Bonferroni-Holm or Benjamini-Hochberg's false discovery rate (FDR). In addition, the method allows users to compare and contrast the results obtained from multiple analyses using the *compareCluster* function. This function returns a table of pathways with multiple adjusted *P*-values—one for each input list. Finally, it further visualizes the results in a side-by-side graph for comparison analysis.

In general, these methods confer flexibility in combining multiple omics significant signals. As a result, the statistical power of these methods is expected to increase. However, most of the mentioned statistical tests assume independence between omics types, which contradicts reality and may negatively affect the analysis accuracy. A possible solution is considering the methods that can combine dependent *P*-values (e.g. Brown's [99] and Liptak's [138]). Another critical pitfall of these methods is that they neglect the actual expression changes, i.e. effect sizes. This might result in information loss. Although *P*-value is influenced

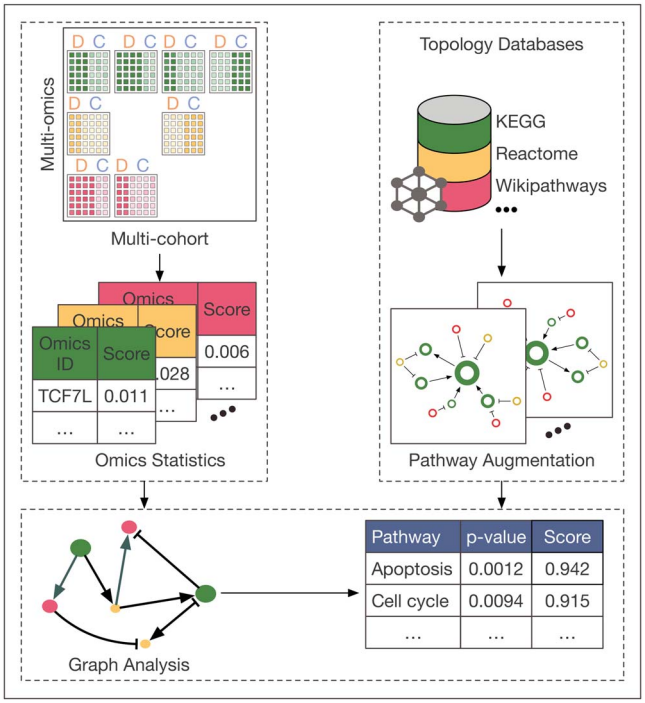

Figure 5. Overall pipeline of graph-transformation-based integrative approaches. These methods first construct pathway networks and then analyze each readout independently to obtain the summary statistics for genes and other omics entities. Finally, they perform graph-based analysis to calculate the *P*-value and network score for each pathway.

by effect size, it is also greatly affected by sample size [139]. For datasets with large sample sizes, a test for differential expression will almost always result in a significant *P*-value, unless the effect size is zero, which is very unlikely. Simply combining the *P*-values would likely produce varying degrees of false discoveries.

Graph-transformation-based approaches

Methods in graph-transformation-based category include PARADIGM [66], Subpathway-GM [67], microGraphite [68], mirIntegrator [69], MOSClip [70], IMPRes-Pro [71] and rPAC [44]. Figure 5 shows the overall pipeline of these methods. Approaches in this category emphasize the point that studying affected genes and pathways separately may hinder the understanding of the whole genome-wide picture [71]. These methods utilize different strategies to construct gene/compound network/graph from existing knowledge about pathway topology. Consequently, each pathway is represented as a network of genes and products. They then perform network-based analysis to identify significantly different pathways between the two phenotypes.

PARADIGM models each gene as a factor graph of four different biological entities: copy number, gene expression state, protein level and protein activity. From the input expression data, PARADIGM computes the scores of the nodes (observed states) and estimates the probability that a node is positive/active or negative/inactive. For each pathway, PARADIGM returns a matrix of states in which columns represent samples and rows represent nodes in the pathway factor graph. All pathways are ranked based on the average number of samples in which significant activity was detected per node.

Subpathway-GM, which is now part of the iSubpathwayMiner package [140], is capable of integrating gene expression data with metabolic data to identify metabolic pathways and subgraphs that are significantly impacted. The method first maps genes and metabolites to enzyme and metabolite nodes of KEGG pathways. The genes and metabolites are then referred to as 'signature nodes' of the pathway graph. The method next searches for shortest paths between signature nodes and then removes nodes that do not belong to these shortest pathways. Subgraphs with a number of nodes larger than a pre-defined threshold are considered important. The method calculates the enrichment score for each subgraph using the hypergeometric test. The *P*-value of the underlying pathway is the smallest *P*-value of its subgraphs. The method repeats the above process for all metabolic pathways to obtain the *P*-values for all pathways. Finally, Subpathway-GM outputs the pathways, subgraphs and their *P*-values.

Both mirIntegrator and microGraphite extend the pathways to include microRNA–gene interactions. mirIntegrator performs pathway analysis on the extended pathways using ORA and Impact Analysis and then combines the *P*-values of the two types of evidence using Fisher's method. In contrast, microGraphite decomposes the network into fully connected cliques and calculates their *P*-values. The method then builds a junction tree, having nodes as cliques and edges as connectivity, and computes the scores of every path in the graph. All the top-scored paths are combined to generate a meta-pathway. In the last step, the meta-pathway's paths are analyzed and ranked according to their involvement in the phenotype. Finally, microGraphite performs a sample permutation test to estimate the significance (*P*-value) of the pathways, meta-pathway, paths and cliques.

MOSClip focuses on identifying pathways or modules associated with patient survival. From the pathway graph, MOSClip generates all possible cliques and performs survival analysis. For each pathway or clique, MOSClip goes through the following process: (i) data filtering to keep only the genomic features that belong to the pathway/module, (ii) dimension reduction, (iii) data concatenation across data types by patient matching and (iv) survival analysis using Cox regression [141]. MOSClip outputs a Cox *P*-value for each pathway that represents how likely the pathway is significantly associated with patient survival.

IMPRes-Pro implements a step-wise active pathway detection algorithm on the background network. Starting from seed nodes, the method explores all potential paths that include as many target nodes as possible. Next, IMPRes-Pro uses the shortest path algorithm [142] with a customized penalty function to achieve the optimal pathway network. The algorithm stops when each node achieves a minimum penalty score. The final active pathway network is detected by truncating and backtracking.

The rPAC software is designed to find routes of a pathway highly associated with the disease. Routes are portions of a pathway that typically involve a transcription factor. The computed scores for each route are: (i) activity scores based on the down or up-regulation of that route, (ii) the rate that a route is altered within a cohort and (iii) the average of yielded activity scores of a route within all samples in a cohort. The *P*-values for routes are computed by testing a two-tailed hypothesis based on the scores.

The graph-transformation-based approaches have significant advantages over enrichment methods. They account for pathway topology and multi-omics interactions by modeling pathways as networks of genes and their products. As a great result, these methods present the option to identify active subnetworks, which would have more explanatory power and decrease the domain of research to the subset of biological components. Adding more

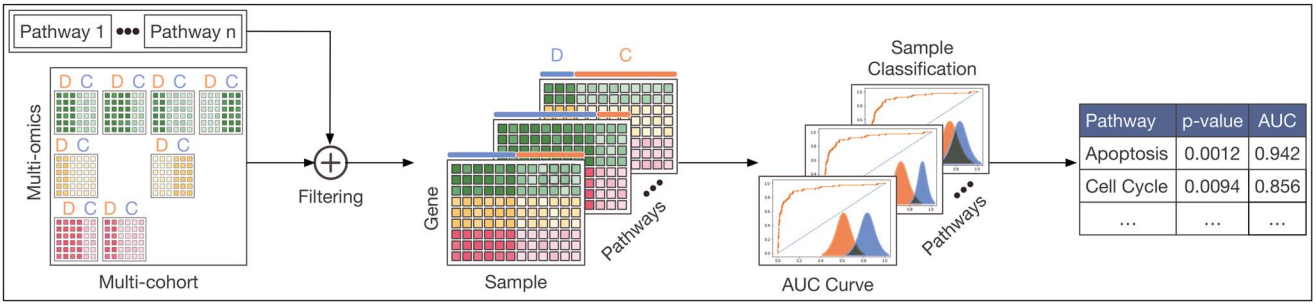

Figure 6. Overall pipeline of machine-learning-based approaches. For a given pathway, these methods filter the multi-omics data to keep only genes belonging to the pathway. Next, these methods classify each sample using the expression data and assess the accuracy using the area under the receiver operating characteristic curve (AUC). The *P*-value of the pathway is calculated by comparing the obtained AUC to its empirical distribution constructed under the null.

layers of omics might, however, result in a complex network that necessitates a significant amount of effort in terms of implementation and analysis.

Machine-learning-based approaches

Figure 6 shows the high-level description of machine-learning-based approaches. Methods in this category include GSOA [72] and PathwayPCA [73]. These methods apply machine learning techniques together with multi-omics data integration to identify the pathways strongly associated with the phenotype. The main idea is to classify samples using the genes that belong to each pathway. Next, the performance of the classification is assessed, and the importance of a pathway is assessed based on how well we can separate disease and healthy samples using the genes in the pathway.

GSOA starts with merging multi-omics data into a single data frame and then performs, by default, radial basis function support vector machine kernel (RBF SVM Kernel) [143] classification using the genes belonging to each pathway. GSOA then assesses the prediction accuracy by calculating the area under the receiver operating characteristic curve (AUC). Then, for each pathway, a *P*-value is calculated using the empirical null distribution of the AUC values.

PathwayPCA [73] extends two existing methods, AES-PCA [144] and SuperPCA [145], to perform pathway analysis and data integration. AES-PCA first performs dimension reduction and computes the latent variables for each data type and each pathway. For a given data type, the latent variables are evaluated against the phenotype either by using a regression model $g(phenotype) = alpha + betaPC1$ (default), or by using a link function g() that varies depending on the response variable (i.e. Cox Proportional Hazards, identity and logit link functions for survival, continuous and binary response variables, respectively). SuperPCA differs from AES-PCA by using the sample label to filter out genes that are not strongly associated with the underlying condition. PathwayPCA extends these methods to combine the results by intersecting the set of significant pathways from each readout. In addition, PathwayPCA implements some functions to present other information such as the obtained values for each principal components in each sample and the loading values corresponding to each gene.

One key advantage of machine-learning-based approaches is that users can incorporate many machine learning techniques into pathway analysis. Many of these techniques are publicly available and thus, require minimal implementation. However, the machine learning algorithms have some disadvantages in the scope of pathway analysis, including (i) the lack of means to take into consideration interactions among omics layers, (ii) the dependency on the chosen machine learning algorithm and parameters and (iii) computational burden due to repeated classifications, i.e. classification is performed for each pathway.

Method assessment

Figure 7 shows the quality assessment of the 32 surveyed methods using the following criteria: (i) validation, (ii) stability, (iii) installation, (iv) user-friendliness, (v) documentation and (vi) tutorial. For each criterion, a method is scored from one (worst) to five (best). Overall, there are 12 methods that have an average score above 4.0. These include ReactomeGSA, ActivePathways, PaintOmics 3, mitch, BLMA, PathwayPCA, Mergeomics, iODA, Subpathway-GM, multiGSEA, CPA and MAPE. Among these, five are GUI tools (ReactomeGSA, PaintOmics 3, Mergeomics, iODA and CPA), while the remaining are stand-alone packages. The details of the assigned scores for each method are provided in Supplementary Table S2.

First, the validation metric refers to the quality of the validation reported in the original paper of each method. A method is scored five if the corresponding paper presents an in-depth analysis using at least two case studies of high-quality, real datasets. We deduct points if the reported validation has low quality or only has one case study. Further, if a method is not validated at all, the validation score is one. Note that most methods provide at least one real case study except InCroMAP, clusterProfiler4.0 and INMEX.

Second, the stability metric indicates how stable the methods are. In order words, this metric assesses how smoothly a method executes an analysis without crashing or having bugs/errors. For each method, we create datasets that are compatible with the input required by each tool. For example, an input object of multiGSEA can have up to three types of data as this software supports the integration of transcriptome, proteome and metabolome. For each method, we create 10 such datasets of different sizes (genes and samples) and execute them to quantify their stability and investigate whether they have bugs, errors or crashes. If any unexpected problem happens during running or any failure arises in different modules of a tool, we lower the score. As shown in the plot, most methods can complete their analysis without problems. For tools that are not available at the time of writing this review (e.g. PARADIGM, rPAC, DANUBE CancerMA), we consider a score of one.

Third, the installation metric refers to how straightforward it is to install a package or software. Most of the available methods have a high score in this metric. We deduct points for some methods (e.g. GSOA, microGraphite, mirIntegrator, MOSClip) because they require users to manually install many packages that can

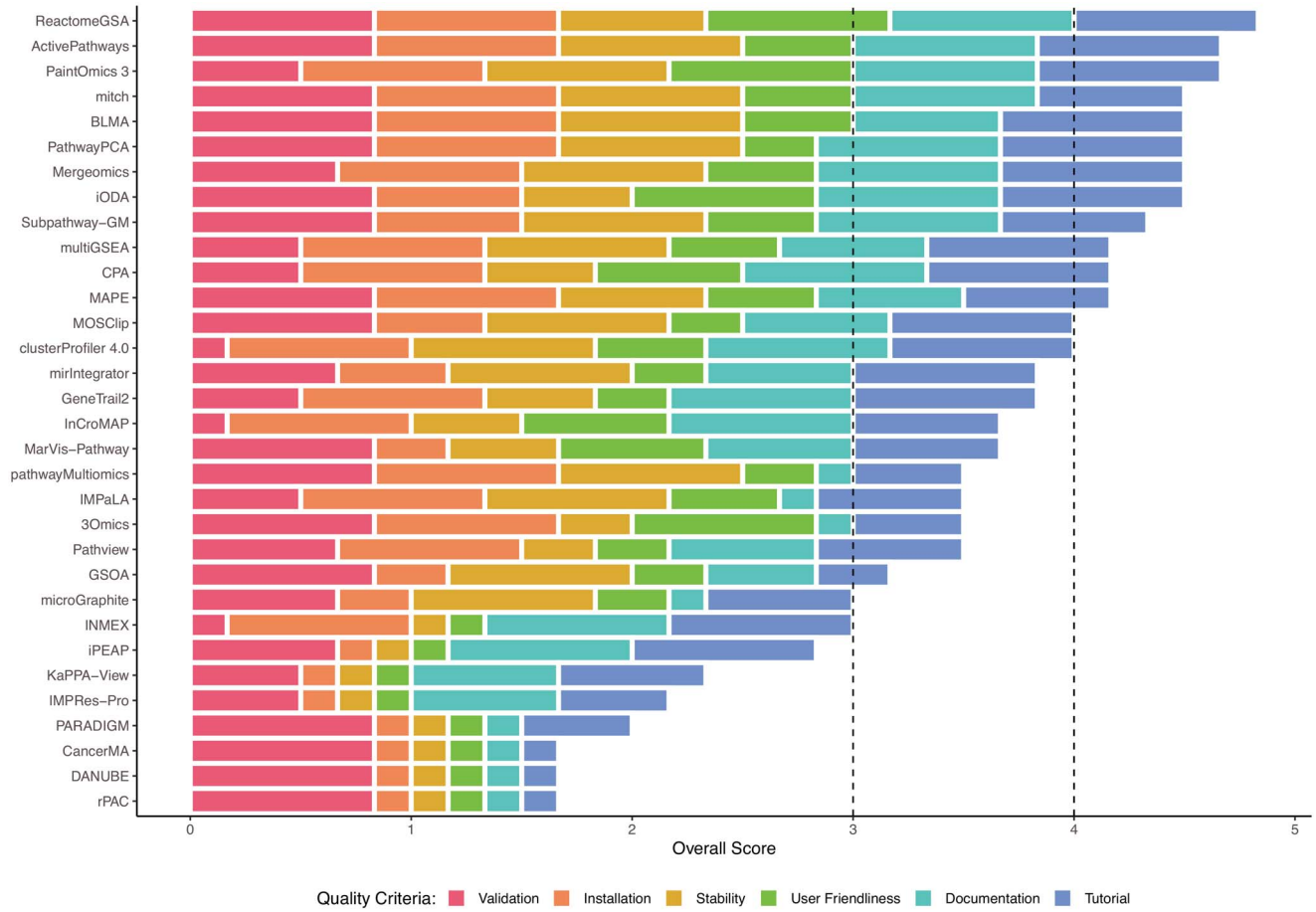

Figure 7. Assessment of 32 surveyed methods in terms of the validation, stability, installation, user friendliness, documentation and tutorial. The score of each metric ranges from one (▬) to five (▬). Each metric has a different color. The methods are sorted according to their average score in an ascending order. The horizontal axis shows the average score for each method. There are 12 methods that have an average score above 4.0: ReactomeGSA, ActivePathways, PaintOmics 3, mitch, BLMA, PathwayPCA, Mergeomics, iODA, Subpathway-GM, multiGSEA, CPA and MAPE.

potentially cause difficulties. We give a score of one to methods that cannot be installed, or require components that are not available anymore. We also give a score of one for tools that are not available or not accessible at the time of writing this manuscript.

Fourth, the user-friendliness metric shows how the tool is well designed from users' perspective. For web-based tools and software with a graphical user interface (GUI), this metric considers how well the architecture is planned so users can navigate through different tabs/modules to perform their analysis. Well-designed GUI tools receive the highest score (five) because they make it easy for users to analyze, visualize and interactively explore the results. For tools that do not have GUI, the highest score they can receive is four. For these methods, the score also takes into account the capability to plot the results or the efforts required from users to convert/pass data from one function to another. ReactomeGSA, PaintOmics 3, iODA, 3Omics, CPA and InCroMAP have the highest score in this metric.

Fifth, the documentation metric refers to the quality of software documentation. It takes into account how well each function and its parameters are documented for stand-alone software and how well different tabs/modules are explained for web-based tools and GUI software. ActivePathways, iODA, ReactomeGSA, PaintOmics 3, Mergeomics, PathwayPCA, clusterProfiler4.0, GeneTrail2, INMEX, mitch, InCroMAP and CPA are among the methods that have high-quality and thorough documentation. At the same time, microGraphite, pathwayMultiomics, PARADIGM, GSOA, IMPaLA and 3Omics are methods that do not present any documentation.

The last metric, tutorial, indicates whether the authors provided an elaborated, step-by-step tutorial that can be easily followed by users. ReactomeGSA, PaintOmics 3, ActivePathways, BLMA, iODA, Mergeomics, PathwayPCA, CPA, multiGSEA, clusterProfiler4.0, MOSClip, GeneTrail2, mirIntegrator and INMEX receive the highest score because they have a high-quality tutorial. In contrast, GSOA, 3Omics, pathwayMultiomics, MarVis-Pathway, IMPRes-Pro and PARADIGM receive low scores because they only provide a brief description of how to perform an analysis. Methods that are not available/accessible receive a score of one.

Summary and discussion

To provide readers with a comprehensive and compact view, we summarized the important details of the 32 surveyed methods in Figure 8. The summary figure emphasizes techniques used in three core modules for integrative pathway analysis: (i) network construction, (ii) pathway statistics computation and (iii) statistics combination. The network construction module refers to the way each method transforms existing pathways. The next module, pathway statistics computation, refers to techniques used to compute pathway statistics, including their enrichment score and *P*-value. The third module, statistics combination, provides the techniques used to combine the statistics across multiple

readouts. The combination can be performed either at the gene-level or pathway-level. The last column of the figure represents the overall performance score for each method (in scale of one to five as reported in *Section Method Assessment*). Additionally, the pros and cons for each category are also provided in the figure.

It is generally accepted that the emergence of a specific phenotype is not a straightforward and predetermined process going from DNA to RNA, to proteins, to downstream biological function, but rather involves complex interactions of factors playing at different levels (e.g. gene expression, variants, microRNA, long non-coding RNA and methylation). For instance, integrating miRNA and mRNA expression profiles results in a better understanding of disease phenomena, both in biomarker discovery [146–148] and pathway analysis [32, 149]. DNA methylation has also been recognized as playing a crucial role in complex diseases [150–153]. There are many other compelling pieces of evidence in mouse, cell lines and human studies demonstrating that even the integration of only two types of data (genotype and mRNA) allows one to successfully connect complex phenotypic traits to inherited and non-inherited factors [154–159]. In many cases, a multi-omics strategy is a must for a systematic understanding of affected pathways [160–162].

Compared with single-omics analysis methods, integrative pathway analysis approaches offer a number of important advantages. First, multi-omics integration allows researchers to observe the full regulatory landscape of pathways from the regulations that occur in different omics layers. Essentially, biological pathways involve a range of different biomolecules, and the flow of information as a response to a specific disease is not confined to a single-omics layer. Therefore, changes in biological pathways can only be accurately identified and comprehensively observed by analyzing multi-omics data. For example, many signaling pathways include phosphorylation reactions that are relevant for regulating the activity of proteins and are not reflected at the transcriptomics level [163]. Hence, analyzing transcriptome data alone might be insufficient. Second, multi-omics integration has the potential to reveal the key insights into underlying pathway mechanisms and uncover the cross-omics relationships, which cannot be made apparent through single-omics studies. Many integrative approaches augment the pathways to obtain a more comprehensive representation of biological pathways and multi-omics interactions. They account for pathway topology and multi-omics interactions by modeling pathways as networks of genes and their products. In this case, multi-omics integration helps researchers to better connect genotype to phenotype and provides novel scientific evidence on disease development, or treatment targets, that can be then tested in further molecular studies [164, 165]. Finally, data integration can potentially increase the statistical power and confidence level of the results [163, 165–167]. All of the integrative pathway analysis methods surveyed in this article provide case studies demonstrating the advantages of multi-omics integration over the analysis of a single omics type or experiment.

However, we would like to note that it is not always beneficial to add more omics layers to an analysis. If an omics layer is not relevant to the underlying condition, adding it to the analysis can increase noise and thus make the analysis less accurate. In an attempt to show the differences between single-omics and multi-omics analysis, Canzler *et al.* [163] provide an extensive survey. They used two case studies, mitochondrial response and murine hepatocyte datasets, to investigate the benefit of multi-omics integration (transcriptome, proteome and metabolome). In the mitochondrial study, the multi-omics integration produced more significant pathways that are relevant to the condition than analyzing transcriptome alone (single-omics). However, in the murine hepatocyte dataset, the analysis of transcriptome data produced more significant pathways that are related to the underlying condition. This shows that multi-omics integration is not always the best option in all conditions. Therefore, choosing the most relevant omics types to include is important before conducting any kind of integration.

Furthermore, there are some outstanding challenges in the context of integrative pathway analysis that necessarily need to be taken into account. First, it should be noted that the accuracy of all analysis methods is highly dependent on data processing and quality control. Multi-omics data are stochastic and heterogeneous in nature that can be attributed to a variety of factors, including technical variability (assays and data type), biological heterogeneity and study bias (sample collection and preparation, experimental design). These factors can greatly affect the results. Therefore, having consistency among input omics data in terms of applied assay, processing protocol and experiment design is very important. In order to ensure the accuracy of the analysis, each pathway analysis technique should evaluate the consistency of the provided data, either through an automatic process or by requiring users' acknowledgment.

Second, analysis results also rely on the degree of effectiveness of each type of omics. It is trivial that biological markers may be affected by a disease differently, and as a result, the amount of carried data by each type of omics could vary. However, all reviewed approaches consider the same weight for all omics types. This argument can be extended further when the input sizes between omics are considerably different, i.e. the larger the dataset, the more information can be extracted. To overcome this challenge, one can examine how each omics type affects the studied condition and then, based on the analysis results, consider a degree of effectiveness or significance for each type of omics. Some statistical tests have been designed specifically for this purpose. One such test is weighted Kolmogorov–Smirnov, which allows for applying a weighting strategy to both the effect of omics types and different input sample sizes.

Third, the inter-dependency among genes and multi-omics layers is another critical factor. These dependencies explain the natural interactions between biological compounds and their omics layers. Therefore, it is important to consider the omics dependencies in designing the pathway analyzer, as it can substantially improve the quality of omics data processing. Among the discussed analysis approaches, pathway graph transformation and visualization categories cover the omics inter-dependencies and utilize these relations in their analysis. However, despite the more powerful analysis, the model becomes more complicated. Therefore, this can be considered a trade-off between the power of the analysis and its computational complexity.

Fourth, it is undeniable that all of the reviewed approaches allow for only two experimental conditions. Depending on the underlying biological process or disease, it may be required in some experiments to have multiple conditions and study them simultaneously. One solution is to analyze every combination of pairs of experimental conditions separately and then merge the obtained results. Another solution is to search for statistical approaches that allow for the analysis of multiple conditions and then extend them to be suitable for analysis.

Fifth, combining different pathway databases has still remained challenging. The existing knowledge regarding pathway annotations is scattered among pathway databases, yet none of them

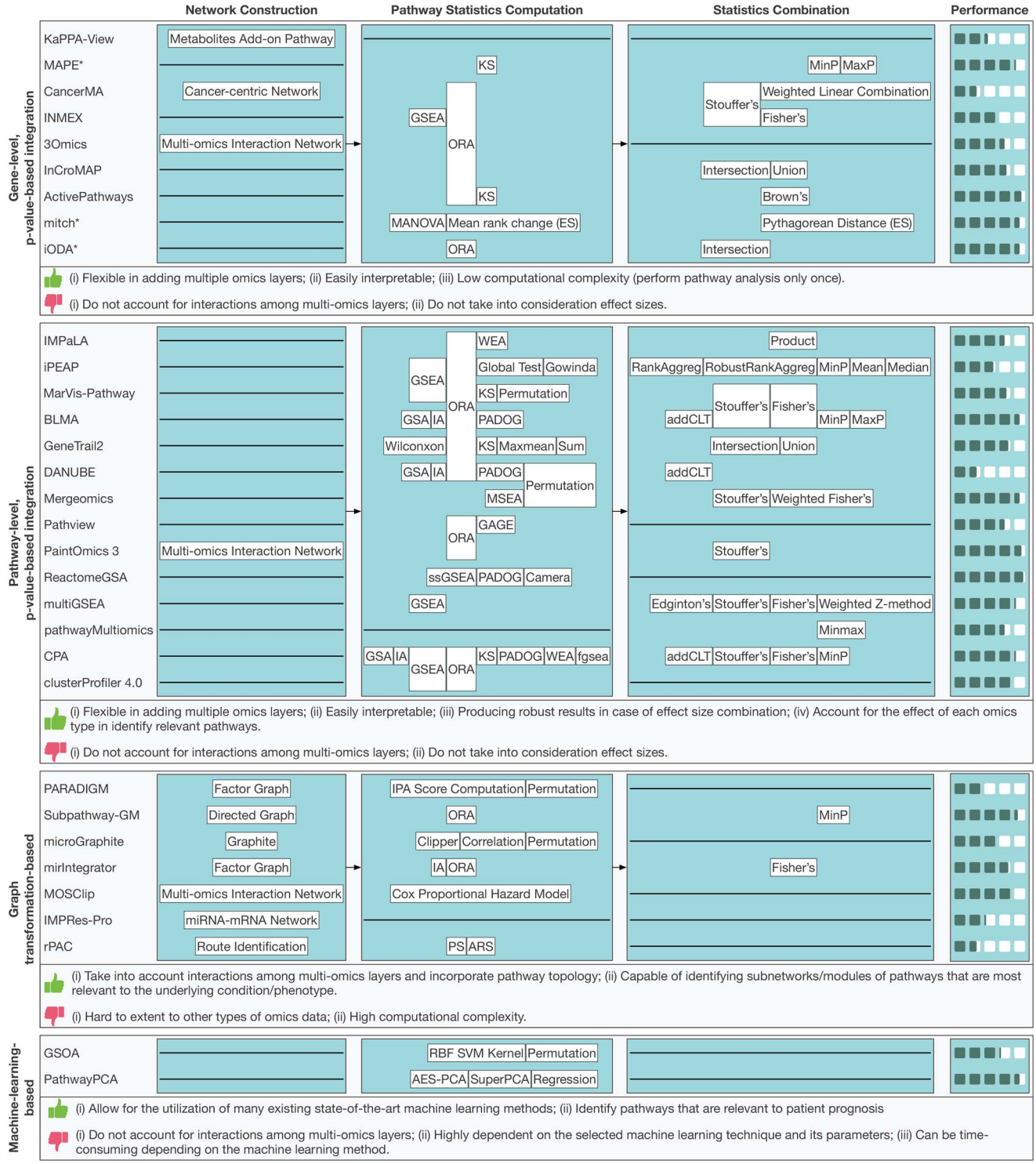

Figure 8. Three core modules of the 32 surveyed methods, their pros and cons, and overall performance score (from one to five). The network construction module represents all activities performed by each method for expanding/transforming the pathway annotation graph. The statistics computation module includes techniques employed by each method for computing statistics at the pathway-level. The score combination module includes each method's strategy in combining the computed statistics at the gene or pathway-level. Most methods combine the statistics at the pathway-level, except those in the gene-level integration category. Methods with an asterisk (*) also support pathway-level combination.

alone is complete. Therefore, exploiting the complementary knowledge available in multiple databases can improve the accuracy and statistical power. Some of the proposed pathway analysis methods attempt to merge multiple databases to cover more information and give users the power to benefit from a more comprehensive database in their analysis. However, the major drawback is that each database uses different identifiers for pathway annotations. Even in some databases, the same entity (gene, metabolite or protein) is annotated with different identifiers, causing redundancy in data.

PathwayCommons [168] is exemplary in this aspect, and it is one of the few comprehensive human pathway repositories. At the time of writing this review, this database has aggregated 5,772 pathway annotations and 2,424,055 interactions from 22 distinct pathway databases, including widely known databases such as KEGG, Reactome, BioCyc, miRTarBase and MSigDB. In comparison with the number of pathways in popular pathway databases such as KEGG (less than 400 pathways) and Reactome (less than 3,000), the number of pathways in PathwayCommons demonstrates how extensive this database is. Also, PathwayCommons annotates all pathways using HGNC or UniProt identifiers; thus, it unifies all the pathways' definitions from different repositories. Several pathway analysis and visualization tools [136, 169, 170] have been developed based on pathway information from the PathwayCommons database. However, none of them are yet capable of integrating multi-omics and/or multi-cohort. Exploring this comprehensive database can benefit researchers in their future development of integrative pathway analysis approaches.

Lastly, the ability to interpret the analysis results should be taken into consideration. Methods with interactive visualization capability may provide interactive interfaces for representing the analysis result (as noted in the last column of Table 2). However, it is still difficult to follow the chains of reactions or signaling cascades in large networks of hundreds or thousands of nodes. One solution to alleviate this problem is to generate sub-networks instead of creating a vast network. Using this strategy, users can easily follow the interactions and inspect the corresponding omics layers. Another solution is to provide a three-dimensional depiction of the entire network. This three-dimensional view will enable a more immersive interaction with the network. For example, users could rotate the perspective to better view the network components or inspect each part of the network.

Conclusion

In this survey, we systematically review 32 integrative pathway analysis methods. The overall pipeline of these methods typically includes data preprocessing, ID mapping, pathway augmentation, differential analysis, pathway analysis and visualization. We categorize these multi-omics pathway analysis approaches into four different categories based on their principal concepts and techniques: (i) gene-level, *P*-value-based integration, (ii) pathway-level, *P*-value-based integration, (iii) graph-transformation-based and (iv) machine-learning-based tools. We discuss the pros and cons of each category, as well as the overall advantages multi-omics integration over single-omics analysis. We also assess the practicality of each method using six different metrics. Our main objective is to help potential users, especially life scientists, to choose a method that is most suitable for their available data and analysis purpose. Finally, we identify the shortcomings of existing approaches with the goal of helping computational scientists to develop new methods that address the current limitations.

Key Points

- Pathway analysis is important because it provides insights into the biology underlying phenotypes beyond the detection of differentially expressed genes or proteins.
- Multi-cohort analysis (meta-analysis) increases statistical power, while multi-omics integration integrates different types of omics data to understand complex biological processes that involve multiple omics levels.
- This article reviews and discusses in-depth 32 pathway analysis methods using multi-cohort and multi-omics data: their availability, supported databases and omics types, hypothesis testing techniques and integration strategies.
- This article points out pros and cons of integrative pathway analysis methods, as well as assesses each method's practicality, and discusses outstanding challenges.
- This article will assist life scientists in selecting suitable methods for their analysis purposes, as well as computational scientists in identifying shortcomings of current methods.

Acknowledgments

This work was partially supported by NSF under grant numbers 2203236 and 2141660. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

Data Availability

The data used in Method Assessment can be found at Gene Expression Omnibus database (Accession IDs: GSE104749, GSE55945 and GSE6956). Other data are downloaded from software package provided by the authors of the papers in this review.

References

1. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**:44.
2. Dahlquist KD, Salomonis N, Vranizan K, *et al.* GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 2002;**31**:19–20.
3. Castillo-Davis CI, Hartl DL. GeneMerge – post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* 2003;**19**(7):891–2.
4. Hosack DA, Jr GD, Sherman BT, *et al.* Identifying biological themes within lists of genes with EASE. *Genome Biol* 2003;**4**:R70.
5. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* 2004;**20**(4):578–80.
6. Berriz GF, King OD, Bryant B, *et al.* Characterizing gene sets with FuncAssociate. *Bioinformatics* 2003;**19**(18):2502–4.
7. Beißbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 2004;**20**:1464–5.
8. Martin D, Brun C, Remy E, *et al.* GOToolBox: functional analysis of gene datasets based on gene ontology. *Genome Biol* 2004;**5**:R101.
9. Subramanian A, Tamayo P, Mootha VK, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceeding of The National Academy of Sciences* 2005;**102**(43):15545–50.

10. Breslin T, Eden P, Krogh M. Comparing functional annotation analyses with Catmap. *BMC Bioinformatics* 2004;**5**(193):1.
11. Goeman JJ, van de Geer, de Kort, *et al.* A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**(1):93–9.
12. Tian L, Greenberg SA, Kong SW, *et al.* Discovering statistically significant pathways in expression profiling studies. *Proceeding of The National Academy of Sciences* 2005;**102**(38): 13544–9.
13. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005;**21**(9):1943–9.
14. Efron B, Tibshirani R. On testing the significance of sets of genes. *The Annals of Applied Statistics* 2007;**1**(1):107–29.
15. Draghici S, Khatri P, Tarca AL, *et al.* A systems biology approach for pathway level analysis. *Genome Res* 2007;**17**(10):1537–45.
16. Tarca AL, Draghici S, Khatri P, *et al.* A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**(1):75–82.
17. Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. *J Comput Biol* 2010;**16**(3): 407–26.
18. Glaab E, Baudot A, Krasnogor N, *et al.* TopoGSA: network topological gene set analysis. *Bioinformatics* 2010;**26**(9):1271–2.
19. Massa MS, Chiogna M, Romualdi C. Gene set analysis exploiting the topology of a pathway. *BMC Syst Biol* 2010;**4**:121.
20. Hung J-H, Whitfield TW, Yang T-H, *et al.* Identification of functional modules that correlate with phenotypic difference: the influence of network topology. *Genome Biol* 2010;**11**:R23.
21. Greenblum SI, Efroni S, Schaefer CF, *et al.* The PathOlogist: an automated tool for pathway-centric analysis. *BMC Bioinformatics* 2011;**12**:133.
22. Geistlinger L, Csaba G, Küffner R, *et al.* From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* 2011;**27**(13):i366–73.
23. Zuguang G, Liu J, Cao K, *et al.* Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Syst Biol* 2012;**6**:56.
24. Zuguang G, Wang J. CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics* 2013;**29**(5):658–60.
25. Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. *Source Code Biol Med* 2012;**7**:10.
26. Ogris C, Helleday T, Sonnhammer ELL. PathwAX: a web server for network crosstalk based pathway annotation. *Nucleic Acids Res* 2016;**44**(W1):W105–9.
27. Nguyen T, Shafi A, Tuan-Minh Nguyen A, *et al.* NBIA: a network-based integrative analysis framework–applied to pathway analysis. *Sci Rep* 2020;**10**:4188.
28. Berger B, Peng J, Singh M. Computational solutions for omics data. *Nat Rev Genet* 2013;**14**:333–46 333.
29. Kristensen VN, Lingjærde OC, Russnes HG, *et al.* Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* 2014;**14**:299–313.
30. Chen J, Jian Zhang Y, Gao YL, *et al.* LncSEA: a platform for long non-coding RNA related sets and enrichment analysis. *Nucleic Acids Res* 2021;**49**(D1):D969–80.
31. Carlevaro-Fita J, Liu L, Zhou Y, *et al.* LnCompare: gene set feature analysis for human long non-coding RNAs. *Nucleic Acids Res* 2019;**47**(W1):W523–9.
32. Vlachos IS, Zagganas K, Paraskevopoulou MD, *et al.* DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. *Nucleic Acids Res* 2015;**43**(W1):W460–6.
33. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* 2012;**8**(2):e1002375.
34. Nguyen T-M, Shafi A, Nguyen T, *et al.* Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol* 2019;**20**:203.
35. Nguyen H, Tran D, Tran B, *et al.* A comprehensive survey of regulatory network inference methods using single-cell RNA sequencing data. *Brief Bioinform* 2021;**22**(3):1–15.
36. Nguyen H, Shrestha S, Tran D, *et al.* A comprehensive survey of tools and software for active subnetwork identification. *Front Genet* 2019;**10**:155.
37. Pinu FR, Beale DJ, Paten AM, *et al.* Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites* 2019;**9**(4):76.
38. Eicher T, Kinnebrew G, Patt A, *et al.* Metabolomics and multi-omics integration: a survey of computational methods and resources. *Metabolites* 2020;**10**(5):202.
39. Jendoubi T. Approaches to integrating metabolomics and multi-omics data: a primer. *Metabolites* 2021;**11**(3):184.
40. Subramanian I, Verma S, Kumar S, *et al.* Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights* 2020;**14**:1–24.
41. Nicora G, Vitali F, Dagliati A, *et al.* Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Front Oncol* 2020;**10**:1030.
42. Feichtinger J, McFarlane RJ, Larcombe LD. Cancerma: a web-based tool for automatic meta-analysis of public cancer microarray data. *Database* 2012;**2012**:bas055.
43. Nguyen T, Mitrea C, Tagett R, *et al.* DANUBE: data-driven Meta-ANalysis using UnBiased empirical distributions—applied to biological pathway analysis. *Proc IEEE* 2016;**105**(3):496–515.
44. Joshi P, Basso B, Wang H, *et al.* rPAC: route based pathway analysis for cohorts of gene expression data sets. *Methods* 2022;**198**:76–87.
45. Tokimatsu T, Sakurai N, Suzuki H, *et al.* KaPPA-view. A web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiol* 2005;**138**(3):1289–300.
46. Shen K, Tseng GC. Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* 2010;**26**(10):1316–23.
47. Xia J, Fjell CD, Mayer ML, *et al.* INMEX—a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res* 2013;**41**(W1):W63–70.
48. Kuo T-C, Tian T-F, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst Biol* 2013;**7**:64.
49. Eichner J, Rosenbaum L, Wrzodek C, *et al.* Integrated enrichment analysis and pathway-centered visualization of metabolomics, proteomics, transcriptomics, and genomics data by using the InCroMAP software. *J Chromatogr B* 2014;**966**:77–82.
50. Paczkowska M, Barenboim J, Sintupisut N, *et al.* Integrative pathway enrichment analysis of multivariate omics data. *Nat Commun* 2020;**11**:735.
51. Kaspi A, Ziemann M. Mitch: multi-contrast pathway enrichment for multi-omics and single-cell profiling data. *BMC Genomics* 2020;**21**:447.
52. Chunjiang Y, Qi X, Lin Y, *et al.* iODA: an integrated tool for analysis of cancer pathway consistency from heterogeneous multi-omics data. *J Biomed Inform* 2020;**112**:103605.

53. Kamburov A, Cavill R, Ebbels TMD, *et al.* Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 2011;**27**(20):2917–8.
54. Sun H, Wang H, Zhu R, *et al.* iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics* 2014;**30**(5):737–9.
55. Kaever A, Landesfeind M, Feussner K, *et al.* MarVis-pathway: integrative and exploratory pathway analysis of non-targeted metabolomics data. *Metabolomics* 2015;**11**(3):764–77.
56. Nguyen T, Tagett R, Donato M, *et al.* A novel bi-level meta-analysis approach: applied to biological pathway analysis. *Bioinformatics* 2016;**32**(3):409–16.
57. Stöckel D, Kehl T, Trampert P, *et al.* Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics* 2016;**32**(10):1502–8.
58. Shu L, Zhao Y, Kurt Z, *et al.* Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics* 2016;**17**:874.
59. Luo W, Pant G, Bhavnasi YK, *et al.* Pathview web: user friendly pathway visualization and data integration. *Nucleic Acids Res* 2017;**45**(W1):W501–8.
60. Diego R H-D, Tarazona S, Martínez-Mira C, *et al.* PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res* 2018;**46**(W1):W503–9.
61. Griss J, Viteri G, Sidiropoulos K, *et al.* ReactomeGSA-efficient multi-omics comparative pathway analysis. *Mol Cell Proteomics* 2020;**19**(12):2115–25.
62. Canzler S, Hackermüller J. multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinformatics* 2020;**21**:561.
63. Odom GJ, Colaprico A, Silva TC, *et al.* PathwayMultiomics: an R package for efficient integrative analysis of multi-omics datasets with matched or un-matched samples. *Front Genet* 2021;**12**:783713.
64. Nguyen H, Tran D, Galazka JM, *et al.* CPA: a web-based platform for consensus pathway analysis and interactive visualization. *Nucleic Acids Res* 2021;**49**(W1):W114–24.
65. Wu T, Hu E, Xu S, *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation* 2021;**2**(3):100141.
66. Vaske CJ, Benz SC, Sanborn JZ, *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* 2010;**26**(12): i237–45.
67. Li C, Han J, Yao Q, *et al.* Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res* 2013;**41**(9):e101.
68. Calura E, Martini P, Sales G, *et al.* Wiring miRNAs to pathways: a topological approach to integrate miRNA and mRNA expression profiles. *Nucleic Acids Res* 2014;**42**(11):e96.
69. Diaz D, Donato M, Nguyen T and Draghici S. MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping. In Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, and Klein TE, editors, *The Pacific Symposium on Biocomputing 2017*. Singapore: World Scientific, 2017, 390–401.
70. Martini P, Chiogna M, Calura E, *et al.* MOSClip: multi-omic and survival pathway analysis for the identification of survival associated gene and modules. *Nucleic Acids Res* 2019;**47**(14): e80–0.
71. Jiang Y, Wang D, Dong X, *et al.* IMPRes-pro: a high dimensional multiomics integration method for in silico hypothesis generation. *Methods* 2020;**173**(1):16–23.
72. MacNeil SM, Johnson WE, Li DY, *et al.* Inferring pathway dysregulation in cancers from multiple types of omic data. *Genome Med* 2015;**7**:61.
73. Odom G, Ban J, Liu L, Wang L, and Chen S. pathwayPCA: Integrative Pathway Analysis with Modern PCA Methodology and Gene Selection, 2019. R package version 1.2.0.
74. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.
75. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.
76. Matthews L, Gopinath G, Gillespie M, *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**(suppl_1):D619–22.
77. Von Mering, Huynen M, Jaeggi D, *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;**31**(1):258–61.
78. Hsu S-D, Lin F-M, Wu W-Y, *et al.* miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Res* 2010;**39**(suppl_1):D163–9.
79. Schaefer CF, Anthony K, Krupa S, *et al.* PID: the pathway interaction database. *Nucleic Acids Res* 2009;**37**(Suppl_1):D674–9.
80. Nishimura D. Biocarta. *Biotech Software and Internet Report* 2001;**2**(3):117–20.
81. Whirl-Carrillo M, McDonagh EM, Hebert JM, *et al.* Pharmacogenomics knowledge for personalized medicine. *Clinical Pharmacology & Therapeutics* 2012;**92**(4):414–7.
82. Pico AR, Kelder T, Van Iersel, *et al.* WikiPathways: pathway editing for the people. *PLoS Biol* 2008;**6**(7):e184.
83. Han H, Shim H, Shin D, *et al.* TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep* 2015;**5**:11432.
84. Karp PD, Billington R, Caspi R, *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2019;**20**(4):1085–93.
85. Kandasamy K, Mohan SS, Raju R, *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol* 2010;**11**:R3.
86. Yamamoto S, Sakai N, Nakamura H, *et al.* INOH: ontology-based highly structured database of signal transduction pathways. *Database* 2011;**2011**:bar052.
87. Ma H, Sorokin A, Mazein A, *et al.* The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol* 2007;**3**(1):135.
88. Frolkis A, Knox C, Lim E, *et al.* SMPDB: the small molecule pathway database. *Nucleic Acids Res*, *D487* 2009;**38**(suppl_1):D480.
89. Korcsmáros T, Farkas IJ, Szalay MS, *et al.* Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics* 2010;**26**(16):2042–50.
90. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;**14**:91.
91. Loughin TM. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis* 2004;**47**(3):467–85.
92. Tippett LHC. *The methods of statistics*. London: Williams & Norgate, 1931.
93. Wilkinson B. A statistical consideration in psychological research. *Psychol Bull* 1951;**48**(2):156.

94. Pihur V, Datta S, Datta S. Weighted rank aggregation of cluster validation measures: a Monte Carlo cross-entropy approach. *Bioinformatics* 2007;**23**(13):1607–15.
95. Kolde R, Laur S, Adler P, *et al.* Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* 2012;**28**(4):573–80.
96. Stouffer SA, Suchman EA, DeVinney LC, *et al. The American Soldier: Adjustment during army life*, Vol. **1**. Princeton: Princeton University Press, 1949.
97. Zaykin DV. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol* 2011;**24**(8):1836–41.
98. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 2005; **102**(43): 15545–50.
99. Brown MB. A method for combining non-independent, one-sided tests of significance. *Biometrics* 1975;987–92.
100. Merico D, Isserlin R, Stueker O, *et al.* Gary D Bader. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. 2010;**5**(11):e13984–PloS ONE.
101. Cline MS, Smoot M, Cerami E, *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2007;**2**(10):2366–82.
102. Tsai C-A, Chen JJ. Multivariate analysis of variance test for gene set analysis. *Bioinformatics* 2009;**25**(7):897–903.
103. Wang Y, Rekaya R. LSOSS: detection of Cancer outlier differential gene expression. *Biomarker Insights* 2010;**5**:BMI–S5175.
104. MacDonald JW, Ghosh D. COPA—cancer outlier profile analysis. *Bioinformatics* 2006;**22**(23):2950–1.
105. Lian H. MOST: detecting cancer differential gene expression. *Biostatistics* 2008;**9**(3):411–8.
106. Baolin W. Cancer outlier differential gene expression detection. *Biostatistics* 2007;**8**(3):566–75.
107. Tibshirani R, Hastie T. Outlier sums for differential gene expression analysis. *Biostatistics* 2007;**8**(1):2–8.
108. Zhang Y, Liu T, Meyer CA, *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;**9**:R137.
109. Salmon-Divon M, Dvinge H, Tammoja K, *et al.* PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* 2010;**11**:415.
110. Adjaye J, Huntriss J, Herwig R, *et al.* Primary differentiation in the human blastocyst: comparative molecular portraits of inner cell mass and Trophectoderm cells. *Stem Cells* 2005;**23**(10):1514–25.
111. Goeman JJ, Van De Geer, De Kort, *et al.* A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; **20**(1): 93–9.
112. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics* 2007;**81**(6):1278–83.
113. Kofler R, Schlötterer C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* 2012;**28**(15):2084–5.
114. Breitling R, Amtmann A, Herzyk P. Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics* 2004;**5**:34.
115. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947;**18**(1):50–60.
116. Fisher RA. *Statistical methods for research workers*. Edinburgh: Oliver & Boyd, 1925.
117. Efron B, Tibshirani R. On testing the significance of sets of genes. *The Annals of Applied Statistics* 2007; **1**(1): 107–29.
118. Tarca AL, Drăghici S, Bhatti G, *et al.* Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics* 2012;**13**:136.
119. Luo W, Friedman MS, Shedden K, *et al.* GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* 2009;**10**:161.
120. Ellson J, Gansner E, Koutsofios L, *et al.* (eds). Graphviz—Open Source Graph Drawing Tools. In: *International Symposium on Graph Drawing*. Springer, 2001, 483–4.
121. Ritchie ME, Phipson B, Di W, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**(7):e47–7.
122. Barbie DA, Tamayo P, Boehm JS, *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 2009;**462**(7269):108–12.
123. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 2013;**14**:7.
124. Stuart T, Butler A, Hoffman P, *et al.* Comprehensive integration of single-cell data. *Cell* 2019;**177**(7):1888–902.
125. McCarthy DJ, Campbell KR, Lun ATL, *et al.* Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 2017;**33**(8):1179–86.
126. Schaefer CF, Anthony K, Krupa S, *et al.* Pid: the pathway interaction database. *Nucleic Acids Res* 2009; **37**(suppl_1): D674–9.
127. Romero P, Wagg J, Green ML, *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* 2005;**6**:R2.
128. Jewison T, Yilu S, Disfany FM, *et al.* SMPDB 2.0: big improvements to the small molecule pathway database. *Nucleic Acids Res* 2014;**42**(D1):D478–84.
129. Mi H, Muruganujan A, Thomas PD. Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* 2012;**41**(D1):D377–86.
130. Kanehisa M, Furumichi M, Sato Y, *et al.* Kegg: integrating viruses and cellular organisms. *Nucleic Acids Res* 2021;**49**(D1):D545–51.
131. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci* 2019;**28**(11):1947–51.
132. Edgington ES. An additive method for combining probability values from independent experiments. *J Psychol* 1972;**80**(2):351–63.
133. Massey Jr FJ. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 1951;**46**(253):68–78.
134. Korotkevich G, Sukhov V, Budin N, *et al.* Fast gene set enrichment analysis. *BioRxiv*, page 060012, 2021.
135. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation-BioRxiv. 2016;**060012**.
136. Wang J, Duncan D, Shi Z, *et al.* WEB-based GEne SeT AnaLysis toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 2013;**41**(W1):W77–83.
137. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945;**1**(6):80–3.
138. Lipták T. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl* 1958;**3**:171–97.
139. Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. *J Grad Med Educ* 2012;**4**(3):279–82.

140. Li C, Li X, Miao Y, *et al.* SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res* 2009;**37**(19):e131–1.
141. Cox DR. Regression models and life-tables. *J R Stat Soc B Methodol* 1972;**34**(2):187–220.
142. Dijkstra EW. A note on two problems in Connexion with graphs. *Numerische Mathematik* 1959;**1**:269–71.
143. Vapnik VN. An overview of statistical learning theory. *IEEE Transactions on Neural Networks and Learning Systems* 1999;**10**(5):988–99.
144. Chen X. Adaptive elastic-net sparse principal component analysis for pathway association testing. *Stat Appl Genet Mol Biol* 2011;**10**(1).
145. Chen X, Wang L, Smith JD, *et al.* Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics* 2008;**24**(21):2474–81.
146. Zhao T, Ding S, Hao Z, *et al.* Integrated miRNA-mRNA analysis provides potential biomarkers for selective breeding in bay scallop (Argopecten irradians). *Genomics* 2021;**113**(4):2744–55.
147. Wotschofsky Z, Gummlich L, Liep J, *et al.* Integrated microRNA and mRNA signature associated with the transition from the locally confined to the metastasized clear cell renal cell carcinoma exemplified by miR-146-5p. *PLoS ONE* 2016;**11**(2):e0148746.
148. Volinia S, Croce CM. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci* 2013;**110**(18):7413–7.
149. Alaimo S, Giugno R, Acunzo M, *et al.* Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget* 2016;**7**(34):54572–82.
150. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. *Nature* 2019;**571**:489–99.
151. Jin Z, Liu Y. DNA methylation in human diseases. *Genes and Diseases* 2018;**5**(1):1–8.
152. Parrella P. Epigenetic signatures in breast cancer: clinical perspective. *Breast Care* 2010;**5**:66–73.
153. Esteller M. Epigenetics in cancer. *New England Journal of Medicine* 2008;**358**:1148–59.
154. Arakawa K, Tomita M. Merging multiple omics datasets in silico: statistical analyses and data interpretation. In: *Systems Metabolic Engineering*. Springer, 2013, 459–70.
155. Montague E, Stanberry L, Higdon R, *et al.* MOPED 2.5–an integrated multi-omics resource: multi-omics profiling expression database now includes transcriptomics data. *Omics: A Journal of Integrative Biology* 2014;**18**(6):335–43.
156. Kohl M, Megger DA, Trippler M, *et al.* A practical data processing workflow for multi-OMICS projects. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, **1844**(1):52–62, 2014.
157. Yoon SH, Han M-J, Jeong H, *et al.* Comparative multi-omics systems analysis of *Escherichia coli* strains B and K-12. *Genome Biol* 2012;**13**(5):R37.
158. Farrell A, McLoughlin N, Milne JJ, *et al.* Application of multi-omics techniques for bioprocess design and optimisation in chinese hamster ovary cells. *J Proteome Res* 2014;**13**(7): 3144–59.
159. Meng C, Kuster B, Culhane AC, *et al.* A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014;**15**:162.
160. Zhu B, Song N, Shen R, *et al.* Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Sci Rep* 2017;**7**:16954.
161. Snyder A, Nathanson T, Funt SA, *et al.* Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: an exploratory multi-omic analysis. *PLoS Med* 2017;**14**(5):e1002309.
162. Graw S, Chappell K, Washam CL, *et al.* Multi-omics data integration considerations and study design for biological systems and disease. *Molecular Omics* 2021;**17**(2):170–85.
163. Canzler S, Schor J, Busch W, *et al.* Prospects and challenges of multi-omics data integration in toxicology. *Arch Toxicol* 2020;**94**(2):371–88.
164. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol* 2017;**18**:83.
165. Lee AH, Shannon CP, Amenyogbe N, *et al.* Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. *Nat Commun* 2019;**10**:1092.
166. Ramasamy A, Mondry A, Holmes CC, *et al.* Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* 2008;**5**(9):e184.
167. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res* 2012;**40**(9):3785–99.
168. Cerami EG, Gross BE, Demir E, *et al.* Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res* 2010;**39**(suppl_1):D685–90.
169. Shannon P, Markiel A, Ozier O, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**(11):2498–504.
170. Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis sof RNA-Seq data. *BMC Bioinformatics* 2018;**19**:534.