PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Nine quick tips for pathway enrichment analysis

Davide Chicco[1]*, **Giuseppe Agapito**[2]

1 Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada,
2 Dipartimento di Giurisprudenza Economia e Sociologia, Università Magna Graecia di Catanzaro, Catanzaro, Italy

* davidechicco@davidechicco.it

Abstract

Pathway enrichment analysis (PEA) is a computational biology method that identifies biological functions that are overrepresented in a group of genes more than would be expected by chance and ranks these functions by relevance. The relative abundance of genes pertinent to specific pathways is measured through statistical methods, and associated functional pathways are retrieved from online bioinformatics databases. In the last decade, along with the spread of the internet, higher availability of computational resources made PEA software tools easy to access and to use for bioinformatics practitioners worldwide. Although it became easier to use these tools, it also became easier to make mistakes that could generate inflated or misleading results, especially for beginners and inexperienced computational biologists. With this article, we propose nine quick tips to avoid common mistakes and to out a complete, sound, thorough PEA, which can produce relevant and robust results. We describe our nine guidelines in a simple way, so that they can be understood and used by anyone, including students and beginners. Some tips explain what to do before starting a PEA, others are suggestions of how to correctly generate meaningful results, and some final guidelines indicate some useful steps to properly interpret PEA results. Our nine tips can help users perform better pathway enrichment analyses and eventually contribute to a better understanding of current biology.

OPEN ACCESS

Citation: Chicco D, Agapito G (2022) Nine quick tips for pathway enrichment analysis. PLoS Comput Biol 18(8): e1010348. https://doi.org/10.1371/journal.pcbi.1010348

Editor: Francis Ouellette, McGill University, CANADA

Published: August 11, 2022

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Pathway enrichment analysis (PEA), also known as functional enrichment analysis or overrepresentation analysis, is a bioinformatics procedure that identifies specific biological pathways as being particularly abundant in a list of genes [1].

Biological pathways describe molecular activities or roles of genes of different kinds. Pathway databases can be specific (HumanCyc [2] for metabolic pathways and LIPEA for lipid functions [3], for example) or more general purpose (KEGG [4], Reactome [5], and WikiPathways [6], for example). Molecular functions can also be represented in a structured hierarchy: The Gene Ontology (GO) [7], for example, contains structured biomolecular annotations that indicate biological processes, molecular functions, or cellular components, connected in directed acyclic graphs.

Several statistical methods can be used to associate the most enriched biological pathways in the input gene list and take into account the number of genes and the likelihood of a pathway to be found enriched. g:Profiler g:GOSt [8–11], for example, uses a modified Fisher's exact test [12–15] to estimate abundance of the genes considering the frequency of the genes in the pathways' database. g:Profiler g:GOSt then proposes three different methods for computing multiple testing correction for *p*-values (g:SCS, Bonferroni correction, or Benjamini–Hochberg false discovery rate (FDR); S2 Text) [10,16].

Multiple PEA tools are available in the scientific literature, both as web tools and as standalone software programs. Some of them employ multiple databases, while others use only one, but they all have the same goal: take an input gene list and associate biological pathways with the larger gene overlap than the one obtained by chance.

Even if a PEA can be done easily, it is also easy to make mistakes that can generate overoptimistic or misleading results. We therefore propose these nine quick tips that can help beginners and inexperienced users perform a PEA properly, by avoiding common errors or pitfalls.

Other authors reported potential problems of functional enrichment analysis [17–21] and described best practices in the past [22, 23], but we believe that our guidelines are easier to follow and to understand by all users, including students and beginners.

Tip 1: Before starting, clarify which analysis you would like to perform

As simple as it might sound, the first step for a sound and robust PEA is about making up your mind: What analysis do you plan to perform? The answer to this question depends mainly on the type of scientific problem you would like to solve and on the type of data you have.

What analysis type. Several different enrichment analyses are available in the bioinformatics landscape; even if most of them have significant differences, sometimes their names are used as synonyms, increasing confusion in the scientific literature. PEA, which is the main topic of this article, is sometimes called functional enrichment analysis. These two names indicate the same procedure: the identification of enriched biological pathways (also called "biological functions") in a list of biomolecular entities (usually genes, but also microRNAs or metabolites), through statistical methods.

PEA methods can also be classified into overrepresentation analysis (ORA) and gene set enrichment analysis (GSEA) approaches. The ORA name highlights the importance of the biological functions that are overrepresented in a group of genes with respect to their role in the human genome [24]. GSEA is both the name of a bioinformatics tool developed and released by scientists at University of California San Diego (UCSD) and Broad Institute [25–27] and the name of the type of analysis they invented. The authors of Enrichr [28–30], for example, define its goal as GSEA.

Some users refer to GSEA and PEA as synonyms [31,32]. Each ORA and GSE approach can be categorized into the competitive and/or self-contained classes based on the null hypothesis. Competitive methods compute *p*-values assuming the genes independence hypothesis is not always true, whereas self-contained methods assume that genes in the gene list are equally associated with phenotype as genes not listed, yielding many relevant genes (like ROAST [33], for example). GSEA approaches are considered a mix of self-contained and competitive methods, since they permute only the genes' class labels (for example, phenotypes) into the pathway, or permute all the genes' class labels for each pathway, comparing the pathway gene set with the query gene set, depending on the parameters chosen [34].

Thus, GSEA methods can perform both self-contained and competitive hypothesis tests by altering how permutation is done for testing the null hypothesis. It is worthy to note that many PEA tools provide both options, ORA and GSEA. ORA methods differ from GSEA because they only consider the query gene set of interest and need a strict cutoff to classify genes as up- and down-regulated; thus, it is advisable to choose GSEA methods when there is uncertainty about the cutoff value. BioPAX-Parser (BiP) [35], pathDIP [36,37], SPIA [38], CePaORA [39], and PathNet [40] are competitive methods, whereas CePa [39] and GSEA [25–27] are self-contained methods. More precisely, the main difference between the GSEA approaches and the ORA approaches is the output: GSEA indicates the pathways that are enriched in genes located at both extreme ends of a ranked gene list, and a higher ranked pathway indicates that more genes are located at the very top or at the very end of this list.

Conversely, ORA outputs all pathways enriched in the query gene list as a whole, and mainly uses a nonranked list (except one option in g:Profiler g:GOst using a minimum hypergeometric value-based method). Therefore, the focus of ORA methods is the gene set, while the focus of GSEA techniques is the ranked pathways list. In this article, we will consider this distinction even if, as we mentioned earlier, the terms GSEA or PEA are often used as synonyms in the scientific literature. Furthermore, topology-based PEA (TPEA) is an advanced PEA that takes into account the hierarchical topology of the analyzed genes, such as the interactions between genes and gene products [24,41–43].

Even if these methods generally produce more precise results, they suffer from the limitation of using a gene topology based on the single cell type in use [24]. Moreover, the topologies of the genes are far from being final and might change as the general biology understanding advances. Lastly, researchers refer to chromosome region enrichment analysis or genomic enrichment analysis to PEA tools that read lists of genomic regions as input, rather than lists of genes. These analyses first associate genes to genomic regions and then retrieve their corresponding biological pathways. GREAT [44], BEHST [45], and Poly-Enrich [46] belong to this category. We report the complete list of PEA tools mentioned in this article in S1 Table. Unlike what some unexperienced PEA users think, it is important to note that PEA does not give clues about the active or inhibited status of the pathways. More appropriately, PEA provides information about how genes help carry out pathways.

Which data type. As it is easy to understand, the type of analysis depends also on the type of the data one would like to analyze. For unordered lists of genes, researchers can use g:Profiler g:GOSt [8–10], Enrichr [28,29], and BioPAX-Parser [35,47]. If the genes are ranked, g: Profiler g:GOSt can treasure this information and generate rank-based functional enrichment results. If the input data are gene expression levels, they can be analyzed through GSEA [27]. pathDIP [37], instead, can assist with curated analyses based on scientific literature.

If one would like to have topological scores to rank cross-enriched pathways using more pathway databases, cPEA [48] might be the best tool choice. GeneTrail [49–52] can be useful for results related to epigenetics, while NoRCE [53] serves well for investigating noncoding RNAs.

Another aspect to keep in mind is the format of the data one would like to analyze, and their specific representation. Different models can represent multiple biochemical reactions responsible for biological functions and pathways. Usually, signaling or metabolic pathways are considered sets of genes interacting in a coordinated way to accomplish a given biological function or process. For instance, in a standard signaling pathway, KEGG [4] uses nodes to represent genes or gene products and edges to define signals, such as activation or inhibition, going from one gene to another. A common metabolic pathway would be depicted with nodes to represent biochemical compounds and edges to represent reactions that transform one or more compounds into other compounds. Enzymes coded by genes usually accomplish these reactions. Therefore, genes or their products are associated with edges rather than nodes in a metabolic pathway, like a signaling pathway. Keep in mind that the immediate impact of this difference is that many techniques cannot be applied directly to all available pathway types (S3 Text).

To summarize, before running any analysis, spend some time studying which scientific question you would like to answer, which data you have for your study, and of which data type they consist. The answers to these questions will help you determine the most suitable enrichment analysis to use.

Tip 2: Ensure the quality of your input genes or genomic regions

The popular saying “Garbage in, garbage out” summarizes a key pillar of computer science: If the quality of data inputed into a computational system or method is poor, the output results will also be of bad quality [54,55]. No matter how efficient and robust a computational method is, to have meaningful results at the end of a computational analysis, data must be of good quality at input. This rule is valid for all computer science, including bioinformatics, and is true for PEA as well.

Before starting a functional analysis, double-check the input list of genes or genomic regions: study how that list was generated, with which tools and when. What criteria were employed in selecting those genes or chromosome regions? Was a scientific article related to that list published recently? If yes, it is a good idea to read it carefully. In a nutshell, ensure that the gene list or the genomic region list that you plan to use for your PEA was assembled in a meaningful, thorough, precise way, with a valid scientific rationale.

If you notice that the input list was generated in an obscure, odd, illogical way, discard it and focus your attention on another gene list. Let us suppose you would like to investigate a diagnostic genetic signature for breast cancer, derived from microarray gene expression. You read the article related to this signature and notice that the authors used 3 datasets generated on three different microarray platforms (Affymetrix, Illumina, and Agilent, for example), without doing any batch effect correction [56,57]. It is clear that this study contains a preprocessing mistake and its results should be discarded or at least treated with caution. In cases like this, we suggest investigating further this list of genes or even avoiding any functional enrichment analysis and look for another genetic signature.

Another red flag for a proposed gene list would be the absence of a validation on an external data cohort. If a gene list was proposed in a study involving only one dataset, it is probably not reliable enough for prognostic or diagnostic scopes.

If some gene symbols are not recognized by the PEA tool, we suggest to look for their Entrez Gene ID’s through g:Profiler g:Convert [9] or to look for their symbol aliases on Gene Cards [58–60].

Do not run a functional enrichment analysis on any gene list “just to see if anything comes out”: If “anything” comes out, it is probably misleading. Only when you are sure about the quality of your gene list you can proceed and start your PEA.

Even if the gene list is well curated, some technical issues can still occur, for example, with the gene symbols.

Tip 3: Use multiple PEA tools, not only one

What we see in multiple studies involving phases of PEA is the habit of employing a single PEA tool. Many bioinformatics and biomedical researchers, in fact, learn how to use one PEA tool well and then stick with it forever by including an analysis done with it in most of their published studies. This approach has several limitations, because using a unique PEA tool of course generates results that are relegated to the databases associated with that specific PEA software.

Even if it seems obvious, we suggest all the practitioners performing a functional enrichment analysis phase to employ at least two different PEA tools. Having results coming from different sources and methods is pivotal: Some results will be confirmatory, some will be complementary, and some might even be discordant. Seeing two sides of a PEA analysis surely can give a user the possibility to learn more about the pathways associated with the input genes.

For example, if the user had an unranked list of gene symbols, we would suggest to apply g: Profiler g:GOSt [10], Enrichr [28], and GeneTrail [52] to it, and then compare their results. Each of these three PEA tools share common databases (the Gene Ontology, for example) but also have specific ones. A user could then analyze their results first to verify if the common pathways are found by all the three methods, and then to analyze the unique terms found. The comparison between the output pathways generated by different PEA tools can be tricky but can reveal essential information about the analyzed gene list.

A quick, straightforward way to compare the enrichment results is to verify if they contain pathways with the same name. This solution could be insufficient since pathways belonging to various databases may have distinct names and might be structured in a redundant, partially overlapping manner in some other databases. This aspect is a well-known complication in PEA due to the lack of a unique standard to represent and store biological pathway data. Consequently, many available software tools can only deal with a single pathway database. To perform pathway enrichment by employing more than a single database, users can employ cPEA [48], a software tool able to deal with several pathway databases using the BioPAX language [61] to store and represent pathways. Or they can use BiP [47] by selecting the "Whole PathwayCommon Data" option that will perform cross enrichment using the whole collection of automatically downloaded locally Pathway Common databases [62] coded in BioPAX. It is worth noticing that evaluating the similarity among pathways may be helpful to compare the genes within each pathway.

Moreover, we describe two possible ways to compare, consolidate, and validate pathways in S1 Text.

Tip 4: Document all your PEA tests and their details

For each PEA software used, keep track of its version, of its parameters' arguments, and of all its details [63,64]. Write this precious information in a notebook [65], and then include it in the supplementary information of the article about the given PEA study. This step is important for you and your future analysis comparison, but also for the reproducibility of your research study [66,67].

In Box 1, we report an example of details and information regarding a functional enrichment analysis made with g:Profiler g:GOSt that should be manually written by a bionformatician in her or his notes. The user should save these pieces of information in addition to saving the full results of the PEA tests, of course. In particular, we recommend to take note of the version and last update of the databases employed: Since they change quite often, some biological annotations can become obsolete soon, with negative consequences on the scientific outcomes [68].

The last part of Box 1 regarding the output file name and location should not be included in the manuscript, of course, but should be written in the user's notebook. This piece of information will be invaluable in the future.

Tip 5: Always use the corrected *p*-value, and not the nominal one

As we explained earlier, pathway enrichment analyses include statistical steps that rank the output pathways by abundance in the gene list and express their enrichment through a

Box 1: Example of PEA test details

My test ID: 2022-02-04, h10:02 EST.

My input genes: AK4, ALDOC, EGLN1, FAM162A, MTFP1, PDK1, PGK1.

My input genes' type: gene symbols.

Source: D. Cangelosi and colleagues [69].

Disease: neuroblastoma.

Tool: g:Profiler g:GOSt.

Access: online via Google Chrome browser.

Version: e104_eg51_p15_3922dba.

URL: https://biit.cs.ut.ee/gprofiler/gost

Organism: *Homo sapiens*.

Query: unordered genes.

Statistical domain scope: only annotated genes.

Significance threshold: g:SCS threshold.

User threshold: 0.005.

Data sources: default.

All the other parameters: default.

My output file(s) name(s): gprofiler_gost_NB_2022-02-04_h1002_output.csv

My output file(s) folder: /home/davide/PEA_analyses/neuroblastoma/

My output file(s) location: bioinformatics-laptop-2021 (Dell Latitude E5420).

probability value called *p*-value. The closer to zero this *p*-value is, the more significant the result is. However, since we know that genes are unevenly annotated in the biological databases, using the simple nominal *p*-values would easily generate misleading results. Two different Gene Ontology annotations, for example, might end up having $p = 0.001$ in the PEA results, and the user might think they are enriched in this gene list. However, these *p*-values might just be related to the fact that one of the two is actually enriched, while the other is just annotated to few genes in the Gene Ontology database.

Additionally, since one *p*-value needs to be calculated for each term, it is very likely that some terms might end up having a significant *p*-value just by chance.

To alleviate this issue, we recommend using the adjusted *p*-value for multiple testing, sometimes also called corrected *p*-value and indicated as *p.adj* [70].

In the hypothesis testing, we have our hypothesis that says that some variables are correlated, and a null hypothesis that states there is no relationship between them [71]. If our test's *p*-value is significant, we can reject the null hypothesis and claim that our hypothesis is true. The issue is that, with many variables and, therefore, multiple hypotheses to test, there is a higher chance to make at least one type 1 error that is to reject the null hypothesis even if the null hypothesis is actually true. This result would be misleading and can be alleviated through methods for multiple testing error correction or adjustment.

This adjustment limits the family error rate or the FDR and therefore improves the quality of the PEA outcomes. An example of technique for the family error rate correction is the Bonferroni method [72], while a common procedure for FDR correction is the Benjamini–Hochberg procedure [73]. The terms adjusted *p-values*, *corrected p-values*, and *false discovery rate (FDR)* values are often used as synonyms in the scientific literature.

Additionally, following a recent debate on the best practices for computational statistics [74], we suggest using the adjusted *p*-value threshold at 0.005 (that corresponds to 5×10^{-3}), as recommended by Benjamin and colleagues [75].

We know that the significance of the results cannot be indicated by a single threshold for all possible PEA experiments: The significance of the pathways, instead, depends on the input data, on the size of the gene list, on the tool and method employed, on the databases used, and on the nonindependence between the genes. We therefore suggest using the $p.adj < 0.005$ threshold for a first strict analysis of the results, and then repeating the test by using a more permissive threshold such as $p.adj < 0.01$, and then again with an even higher threshold, such as the traditional $p.adj < 0.05$. Based on the characteristics of the experiment, results found by one particular threshold might be more suitable than results found with other thresholds.

In any case, the results found in this phase should be then validated through wet lab experiments or a literature review (Tip 8) and reviewed by a wet lab biologist (Tip 9), since these steps would avoid publication of many false findings [76,77].

We believe this tip is true not only for bioinformatics, but also for all the scientific studies involving statistics and probability values.

Tip 6: Keep in mind that your PEA results can be strongly affected by the statistical tests and the visualization techniques you use

Statistical tests. As we mentioned earlier, each PEA software tool uses a different statistical method to identify the biological pathways enriched in a set of input genes. These statistical techniques associate a corrected probability value (*p*-value) to each pathway, which indicates its importance: the lower the adjusted *p*-value is, the more the pathway is enriched in genes from the query gene list compared to all genes.

Pathways are not equal in their number of genes they contain, and some contain a limited number of curated genes, but therefore can be very relevant in a PEA analysis if found enriched in a high percentage of input genes.

Different statistical techniques, however, can generate different results, and this is something users should always keep in mind. We describe an example of different results obtained on g:Profiler g:GOSt when using different statistical tests in S2 Text, and we report the list of the statistical methods of the PEA tools mentioned in this article in S1 Table.

Our general advice for this task is to keep in mind that different statistical methods can generate different results, so avoid blind use of any statistical test. Study which statistical method can be more suitable for your analysis and why, and then apply it.

Visualization. Scientific visualization is a key pillar of bioinformatics and of modern scientific research [78]. Proper visualization plots do not only represent the data or the results observed in an experiment, but they can also provide alternative, new insights about the data themselves [79].

The visualization step of a PEA, although fundamental, is sometimes underrated by inexperienced users. On the contrary, we believe this phase is vital for the interpretation of the PEA results. Following what we suggested in Tip 4, we advise all PEA practitioners to employ multiple tools for this task. Moreover, the key point to keep in mind during this phase is that different visualization tools and styles can highlight different scientific aspects of the results and therefore unveil unexpected biological novelty that would have been unnoticed otherwise.

Visualization of PEA results can be useful and advantageous because it easily allows users to quickly detect the main enriched functional subjects, which they can then use to interpret the enrichment results. This identification of the functional subjects and their interpretation would be more difficult without a visualization step. Moreover, several useful PEA visualization techniques allow users to deal with redundancy of enrichment results by grouping together similar processes and pathways into common functional themes.

Enrichment Maps [80] and enrichplot [81] for biological pathways, AutoAnnotate [82] for networks, and REVIGO [83] and CirGO [84] for GO annotations are few examples of different visualization techniques and contents. Network visualization techniques can also be used to detect a lower adjusted *p*-value threshold (Tip 5).

To recap, avoid blind use of visualization techniques: understand the available ones and choose the most suitable one for your case.

Tip 7: Consider using subgroups of correlated genes instead of all your input genes

A common practice in PEA is to take all the genes derived from an experiment or a previous analysis and to use them all as input in a PEA tool. This is surely a good thing to do when the users do not know any hierarchy or relationship between the input genes, but it can also produce many gene–pathway associations that might turn out to be irrelevant or even misleading in the end. Additionally, using many input genes could produce a large number of general pathways in the results, such as "signaling pathway" available in KEGG [4] and Reactome [5], for example, which do not improve our understanding of the affected biological processes and functions. Some PEA tools give the possibility to exclude these generic terms from the results, but not all of them.

Instead of using all genes as input for the PEA, we therefore suggest bioinformatics practitioners to detect subgroups of correlated genes and perform the PEA on each of these subgroups alone.

Subgroups of correlated genes can be found, for example, through protein–protein interaction networks' tools such as IID [85–88], STRING [89–92], GeneMANIA [93–97], or Reactome Functional Interaction Network (Reactome FI) [98,99]. These software programs are able to cluster together groups of genes that might share a common physical interaction in their databases. Woodwarda and colleagues [100] recently released a GSEA tool enhanced with epistatic interactions [101], which might be of interest for this scope.

To this end, some R packages have been recently released: pathfindR [102] and netGO [103], which exploit the protein–protein interaction networks to produce more accurate PEA results.

Using these groups of genes, which are already correlated between each other by sharing the same physical interactions, would probably detect more precise biological pathways as output of the PEA.

Tip 8: Use the (recent) scientific literature to review your PEA results

The results you obtained with the PEA tools used are surely interesting and useful, but they are likely based on databases and datasets collected some months or even years ago. Therefore, your results might not be as novel as the recent scientific literature: Some new studies about pathway–gene associations might have been published between the release of the databases employed by the PEA tools used and when your PEA was executed.

To verify your findings, we therefore suggest any practitioner to manually perform a literature search and look for scientific studies published about the significant genes–pathways associations found by the functional enrichment analysis and about the role of the genes inside the enriched pathways found.

The search can be done by using the pathways and the genes as keywords on Google Scholar [104] and PubMed [105]. We also suggest to search on the preprint servers such as bioRxiv [106] and arXiv Quantitative Biology [107], although the fact that these preprint documents are not peer reviewed should be kept in mind. This phase can help alleviate the problem of outdated gene annotations [68].

However, we know that sometimes the list of genes and the list of pathways are so large that manually looking for at least one article about each of them would take too much time. As a rule of thumb, we therefore suggest the users to investigate at least the top twenty genes–pathways associations in the literature. Alternatively, the user could filter the genes by known importance: They could study their input list of genes, identify some frequently seen genes that they already saw in the literature, and investigate their pathways found by the PEA. In any case, we invite users to verify PEA results by looking at the role of the genes shared in the top PEA output pathways to precisely define the biological functions targeted by the gene list (S1 Text).

Tip 9: Ask a wet lab biologist or a clinician to review your PEA results

After looking for evidence about the results of a PEA with scientific literature (Tip 8), we believe one additional step is needed: To further validate the PEA results achieved, a wet lab biologist or a clinician should review these results and clearly say if they make sense or if they contain mistakes or inappropriate information.

Similarly to what is suggested for machine learning studies [108], we therefore suggest that all computational biologists, after performing the PEA, contact a biology researcher and ask for a review of their PEA results. This person should not be a user or a computational biologist but should have a degree in traditional biology or medicine and should be familiar with scientific results obtained in the wet lab.

The point of view of this expert will surely provide interesting considerations and feedback regarding the PEA results and will highlight some aspects that maybe the user might have overlooked. If possible, one can also consider asking this person to perform some wet lab validation of the results found through the PEA. We intended this list of quick tips for computational analyses, but it goes without saying that a precise biological validation made in a wet lab would be extremely useful and even more relevant than any literature review.

Conclusions

PEA is a pivotal step of bioinformatics studies that highlights the most relevant biological functions associated with gene lists. In the last decade, huge computational resources and numerous web tools available online made this analysis type easy for anyone to perform. However, it also became easy to make mistakes by using PEA tools, data, or results that might not address the original scientific scope properly. Following the recent debates emerged in machine learning and biomedical informatics communities [108–116], we propose these nine quick tips that can be used as a checklist for any computational user running a functional enrichment analysis.

Supporting information

S1 Table. List of the PEA tools mentioned in this study.
(PDF)

S1 Text. Description of the validation of genes inside the pathways.
(PDF)

S2 Text. Example of usage of different statistical tests in g: Profiler g:GOSt.
(PDF)

S3 Text. Description of pathway data conversion.
(PDF)

References

1. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019; 14(2):482–517. https://doi.org/10.1038/s41596-018-0103-9 PMID: 30664679
2. Trupp M, Altman T, Fulcher CA, Caspi R, Krummenacker M, Paley S, et al. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. Genome Biol. 2010; 11(1):1–1.
3. Acevedo A, Durán C, Ciucci S, Gerl M, Cannistraci CV. LIPEA: lipid pathway enrichment analysis. bioRxiv. 2018; 274969:1–5.
4. Ogata H, Goto S, Fujibuchi W, Kanehisa M. Computation with the KEGG pathway database. Biosystems. 1998; 47(1–2):119–128. https://doi.org/10.1016/s0303-2647(98)00017-3 PMID: 9715755
5. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005; 33(suppl 1):D428–D432. https://doi.org/10.1093/nar/gki072 PMID: 15608231
6. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018; 46 (D1): D661–D667. https://doi.org/10.1093/nar/gkx1064 PMID: 29136241
7. The Gene Ontology Consortium. The Gene Ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47(D1):D330–D338. https://doi.org/10.1093/nar/gky1055 PMID: 30395331
8. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007; 35(suppl 2):W193–W200.
9. Reimand J, Arak T, Vilo J. g:Profiler—a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 2011; 39(suppl 2):W307–W315.
10. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. 2016; 44(W1):W83–W89. https://doi.org/10.1093/nar/gkw199 PMID: 27098042
11. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2–an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Res. 2020; 9. https://doi.org/10.12688/f1000research.24956.2 PMID: 33564394
12. Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. J R Stat Soc. 1922; 85(1):87–94.
13. Upton GJ. Fisher's exact test. J R Stat Soc Ser A Stat Soc. 1992; 155(3):395–402.
14. Bower KM. When to use Fisher's exact test. American Society for Quality, Six Sigma Forum Magazine. vol. 2; 2003. p. 35–37.
15. Connelly LM. Fisher's exact test. MedSurg Nursing. 2016; 25(1):58–60. PMID: 27044131
16. g:Profiler. Welcome to g:Profiler; 2022. Available from: https://biit.cs.ut.ee/gprofiler/page/docs#significance_threhshold [cited 2022 Feb 2].
17. Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global-omics data. Genome Biol. 2015; 16(1):1–3. https://doi.org/10.1186/s13059-015-0761-7 PMID: 26346307
18. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2016; 25(1):472–487. https://doi.org/10.1177/0962280212460441 PMID: 23070592

19. Bleazard T, Lamb JA, Griffiths-Jones S. Bias in microRNA functional enrichment analysis. Bioinformatics. 2015; 31(10):1592–1598. https://doi.org/10.1093/bioinformatics/btv023 PMID: 25609791

20. Simillion C, Liechti R, Lischer HE, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017; 18(1):1–14.

21. Wijesooriya K, Jadaan SA, Perera KL, Kaur T, Ziemann M. Urgent need for consistent standards in functional enrichment analysis. PLoS Comput Biol. 2022; 18(3):e1009935. https://doi.org/10.1371/journal.pcbi.1009935 PMID: 35263338

22. Mubeen S, Tom Kodamullil A, Hofmann-Apitius M, Domingo-Fernández D. On the influence of several factors on pathway enrichment analysis. Brief Bioinform. 2022; 23(3):bbac143. https://doi.org/10.1093/bib/bbac143 PMID: 35453140

23. Wieder C, Frainay C, Poupin N, Rodríguez-Mier P, Vinson F, Cooke J, et al. Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. PLoS Comput Biol. 2021; 17 (9):e1009105. https://doi.org/10.1371/journal.pcbi.1009105 PMID: 34492007

24. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012; 8(2):e1002375. https://doi.org/10.1371/journal.pcbi.1002375 PMID: 22383865

25. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34(3):267–273. https://doi.org/10.1038/ng1180 PMID: 12808457

26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

27. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics. 2007; 23(23):3251–3253. https://doi.org/10.1093/bioinformatics/btm369 PMID: 17644558

28. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013; 14(1):1–14. https://doi.org/10.1186/1471-2105-14-128 PMID: 23586463

29. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016; 44(W1):W90–W97. https://doi.org/10.1093/nar/gkw377 PMID: 27141961

30. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with Enrichr. Curr Protoc. 2021; 1(3):e90. https://doi.org/10.1002/cpz1.90 PMID: 33780170

31. Mathur R, Rotroff D, Ma J, Shojaie A, Motsinger-Reif A. Gene set analysis methods: a systematic comparison. BioData Mining. 2018; 11(1):1–19. https://doi.org/10.1186/s13040-018-0166-8 PMID: 29881462

32. Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene set analysis: challenges, opportunities, and future research. Front Genet. 2020: 654.

33. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics. 2010; 26(17):2176–2182. https://doi.org/10.1093/bioinformatics/btq401 PMID: 20610611

34. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014; 15(4):504–518. https://doi.org/10.1093/bib/bbt002 PMID: 23413432

35. Agapito G, Pastrello C, Guzzi PH, Jurisica I, Cannataro M. BioPAX-Parser: parsing and enrichment analysis of BioPAX pathways. Bioinformatics. 2020; 36(15):4377–4378. https://doi.org/10.1093/bioinformatics/btaa529 PMID: 32437515

36. Rahmati S, Abovsky M, Pastrello C, Jurisica I. pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis. Nucleic Acids Res. 2017; 45(D1):D419–D426. https://doi.org/10.1093/nar/gkw1082 PMID: 27899558

37. Rahmati S, Abovsky M, Pastrello C, Kotlyar M, Lu R, Cumbaa CA, et al. pathDIP 4: an extended pathway annotations and enrichment analysis resource for human, model organisms and domesticated species. Nucleic Acids Res. 2020; 48(D1):D479–D488. https://doi.org/10.1093/nar/gkz989 PMID: 31733064

38. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Js K, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009; 25 (1):75–82. https://doi.org/10.1093/bioinformatics/btn577 PMID: 18990722

39. Gu Z, Wang J. CePa: an R package for finding significant pathways weighted by multiple network centralities. Bioinformatics. 2013; 29(5):658–660. https://doi.org/10.1093/bioinformatics/btt008 PMID: 23314125

40. Dutta B, Wallqvist A, Reifman J. PathNet: a tool for pathway analysis using topological information. Source Code Biol Med. 2012; 7(1):1–12.

41. Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol. 2019; 20(1):1–15.

42. Ma J, Shojaie A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. BMC Bioinformatics. 2019; 20(1):1–14.

43. Yang Q, Wang S, Dai E, Zhou S, Liu D, Liu H, et al. Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. Brief Bioinform. 2019; 20(1):168–177. https://doi.org/10.1093/bib/bbx091 PMID: 28968630

44. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010; 28(5):495–501. https://doi.org/10.1038/nbt.1630 PMID: 20436461

45. Chicco D, Bi HS, Reimand J, Hoffman MM. BEHST: genomic set enrichment analysis enhanced through integration of chromatin long-range interactions. bioRxiv. 2019; 168427:1–29.

46. Lee CT, Cavalcante RG, Lee C, Qin T, Patil S, Wang S, et al. Poly-Enrich: count-based methods for gene set enrichment testing with genomic regions. NAR Genome Bioinform. 2020; 2(1):lqaa006. https://doi.org/10.1093/nargab/lqaa006 PMID: 32051932

47. Agapito G, Cannataro M. Using BioPAX-Parser (BiP) to enrich lists of genes or proteins with pathway data. BMC Bioinformatics. 2021; 22(13):1–35. https://doi.org/10.1186/s12859-021-04297-z PMID: 34592927

48. Agapito G, Cannataro M. cPEA: a parallel method to perform pathway enrichment analysis using multiple pathways databases. Soft Comput. 2020; 24(23):17561–17572.

49. Keller A, Backes C, Al-Awadhi M, Gerasch A, Küntzer J, Kohlbacher O, et al. GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. BMC Bioinformatics. 2008; 9(1):1–6. https://doi.org/10.1186/1471-2105-9-552 PMID: 19099609

50. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res. 2007; 35(suppl 2):W186–W192. https://doi.org/10.1093/nar/gkm323 PMID: 17526521

51. Stöckel D, Kehl T, Trampert P, Schneider L, Backes C, Ludwig N, et al. Multi-omics enrichment analysis using the GeneTrail2 web service. Bioinformatics. 2016; 32(10):1502–1508. https://doi.org/10.1093/bioinformatics/btv770 PMID: 26787660

52. Gerstner N, Kehl T, Lenhof K, Müller A, Mayer C, Eckhart L, et al. GeneTrail 3: advanced high-throughput enrichment analysis. Nucleic Acids Res. 2020; 48(W1):W515–W520. https://doi.org/10.1093/nar/gkaa306 PMID: 32379325

53. Olgun G, Nabi A, Tastan O. NoRCE: non-coding RNA sets cis enrichment tool. BMC Bioinformatics. 2021; 22(1):1–17.

54. Rose LT, Fischer KW. Garbage in, garbage out: having useful data is everything. Measurement Interdiscip Res Perspect. 2011; 9(4):222–226.

55. Kilkenny MF, Robinson KM. Data quality:“Garbage in–garbage out”. Health Inf Manag J. 2018; 47 (3):103–105. https://doi.org/10.1177/1833358318774357 PMID: 29719995

56. Čuklina J, Pedrioli PG, Aebersold R. Review of batch effects prevention, diagnostics, and correction approaches. Mass Spectrometry Data Analysis in Proteomics. Springer; 2020. p. 373–387.

57. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE. 2011; 6(2): e17238. https://doi.org/10.1371/journal.pone.0017238 PMID: 21386892

58. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics. 1998; 14(8):656–664. https://doi.org/10.1093/bioinformatics/14.8.656 PMID: 9789091

59. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards version 3: the human gene integrator. Database. 2010; 2010. https://doi.org/10.1093/database/baq020 PMID: 20689021

60. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinform. 2016; 54(1):1–30. https://doi.org/10.1002/cpbi.5 PMID: 27322403

61. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol. 2010; 28(9):935–942. https://doi.org/10.1038/nbt.1666 PMID: 20829833

62. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. 2010; 39(suppl 1):D685–D690. https://doi.org/10.1093/nar/gkq1039 PMID: 21071392

63. Karimzadeh M, Hoffman MM. Top considerations for creating bioinformatics software documentation. Brief Bioinform. 2018; 19(4):693–699. https://doi.org/10.1093/bib/bbw134 PMID: 28088754

64. Noble WS. A quick guide to organizing computational biology projects. PLoS Comput Biol. 2009; 5(7): e1000424. https://doi.org/10.1371/journal.pcbi.1000424 PMID: 19649301

65. Schnell S. simple rules for a computational biologist's laboratory notebook. PLoS Comput Biol. 2015; 11 (9):e1004385. https://doi.org/10.1371/journal.pcbi.1004385 PMID: 26356732

66. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013; 9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285 PMID: 24204232

67. Elofsson A, Hess B, Lindahl E, Onufriev A, Van der Spoel D, Wallqvist A. Ten simple rules on how to create open access and reproducible molecular simulations of biological systems. PLoS Comput Biol. 2019; 15(1):e1006649. https://doi.org/10.1371/journal.pcbi.1006649 PMID: 30653494

68. Wadi L, Meyer M, Weiser J, Stein LD, Reimand J. Impact of outdated gene annotations on pathway enrichment analysis. Nat Methods. 2016; 13(9):705–706. https://doi.org/10.1038/nmeth.3963 PMID: 27575621

69. Cangelosi D, Morini M, Zanardi N, Sementa AR, Muselli M, Conte M, et al. Hypoxia predicts poor prognosis in neuroblastoma patients and associates with biological mechanisms involved in telomerase activation and tumor microenvironment reprogramming. Cancers. 2020; 12(9):2343. https://doi.org/10.3390/cancers12092343 PMID: 32825087

70. Jafari M, Ansari-Pour N. Why, when and how to adjust your P values? Cell J (Yakhteh). 2019; 20 (4):604. https://doi.org/10.22074/cellj.2019.5992 PMID: 30124010

71. Cornellius Yudha Wijaya. Multiple hypothesis testing correction for data scientist; 2022. Available from: https://towardsdatascience.com/multiple-hypothesis-testing-correction-for-data-scientist-46d3a3d1611d [cited 2022 Jun 1].

72. Napierala MA. What is the Bonferroni correction? AAOS Now. 2012:40–41.

73. Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. J Educ Behav Stat. 2002; 27(1):77–83.

74. Makin TR, de Xivry JJO. Science forum: ten common statistical mistakes to watch out for when writing or reviewing a manuscript. eLife. 2019; 8:e48175.

75. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018; 2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z PMID: 30980045

76. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722

77. Grosch E. Reply to "Ten simple rules for getting published". PLoS Comput Biol. 2007; 3(9):e190. https://doi.org/10.1371/journal.pcbi.0030190 PMID: 17907799

78. Hansen CD, Chen M, Johnson CR, Kaufman AE, Hagen H. Scientific visualization. Springer; 2014.

79. Pastrello C, Otasek D, Fortney K, Agapito G, Cannataro M, Shirdel E, et al. Visual data mining of biological networks: one size does not fit all. PLoS Comput Biol. 2013; 9(1):e1002833. https://doi.org/10.1371/journal.pcbi.1002833 PMID: 23341759

80. Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment Map: a network-based method for gene-set enrichment visualization and interpretation. PLoS ONE. 2010; 5(11):e13984. https://doi.org/10.1371/journal.pone.0013984 PMID: 21085593

81. Yu, Guangchuang. Biomedical knowledge mining using GOSemSim and clusterProfiler: enrichplot; 2022. Available from: https://bioc.ism.ac.jp/packages/3.7/bioc/vignettes/enrichplot/inst/doc/enrichplot.html [cited 2022 Feb 3].

82. Kucera M, Isserlin R, Arkhangorodsky A, Bader GD. AutoAnnotate: a Cytoscape app for summarizing networks with semantic annotations. F1000Res. 2016; 5:1717. https://doi.org/10.12688/f1000research.9090.1 PMID: 27830058

83. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of Gene Ontology terms. PLoS ONE. 2011; 6(7):e21800. https://doi.org/10.1371/journal.pone.0021800 PMID: 21789182

84. Kuznetsova I, Lugmayr A, Siira SJ, Rackham O, Filipovska A. CirGO: an alternative circular way of visualising Gene Ontology terms. BMC Bioinformatics. 2019; 20(1):1–7.

85. Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res. 2016; 44(D1):D536–D541. https://doi.org/10.1093/nar/gkv1115 PMID: 26516188

86. Pastrello C, Kotlyar M, Jurisica I. Informed use of protein–protein interaction data: a focus on the integrated interactions database (IID). Protein-Protein Interaction Networks. Springer; 2020:125–134. https://doi.org/10.1007/978-1-4939-9873-9_10 PMID: 31583635

87. Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. Nucleic Acids Res. 2019; 47(D1): D581–D589. https://doi.org/10.1093/nar/gky1037 PMID: 30407591

88. Kotlyar M, Pastrello C, Ahmed Z, Chee J, Varyova Z, Jurisica I. IID 2021: towards context-specific protein interaction analyses by increased coverage, enhanced annotation and enrichment analysis. Nucleic Acids Res. 2022; 50(D1):D640–D647. https://doi.org/10.1093/nar/gkab1034 PMID: 34755877

89. Von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, et al. STRING 7—Recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 2007; 35(suppl 1): D358–D362. https://doi.org/10.1093/nar/gkl825 PMID: 17098935

90. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2012; 41 (D1):D808–D815. https://doi.org/10.1093/nar/gks1094 PMID: 23203871

91. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015; 43(D1): D447–D452. https://doi.org/10.1093/nar/gku1003 PMID: 25352553

92. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2018; 47(D1):D607–D613.

93. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 2008; 9(1):1–15. https://doi.org/10.1186/gb-2008-9-s1-s4 PMID: 18613948

94. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Research. 2010; 38(suppl 2):W214–W220. https://doi.org/10.1093/nar/gkq537 PMID: 20576703

95. Zuberi K, Franz M, Rodriguez H, Montojo J, Lopes CT, Bader GD, et al. GeneMANIA prediction server 2013 update. Nucleic Acids Res. 2013; 41(W1):W115–W122. https://doi.org/10.1093/nar/gkt533 PMID: 23794635

96. Montojo J, Zuberi K, Rodriguez H, Bader GD, Morris Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. F1000Res. 2014; 3:153. https://doi.org/10.12688/f1000research.4572.1 PMID: 25254104

97. Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, et al. GeneMANIA update 2018. Nucleic Acids Res. 2018; 46(W1):W60–W64. https://doi.org/10.1093/nar/gky311 PMID: 29912392

98. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2010; 39(suppl 1):D691–D697. https://doi.org/10.1093/nar/gkq1018 PMID: 21067998

99. Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. F1000Res. 2014; 3. https://doi.org/10.12688/f1000research.4431.2 PMID: 25309732

100. Woodwarda AA, Taylorb D, Goldmuntzb E, Mitchellc LE, Agopianc A, Moorea J, et al. Gene-interaction-sensitive enrichment analysis in congenital heart disease. BioData Mining. 2022; 15(4):1–16.

101. Chicco D, Faultless T. Brief survey on machine learning in epistasis. Epistasis. Springer; 2021:169–179. https://doi.org/10.1007/978-1-0716-0947-7_11 PMID: 33733356

102. Ulgen E, Ozisik O, Sezerman OU. pathfindR: an R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front Genet. 2019: 858.

103. Kim J, Yoon S, Nam D. netGO: R-Shiny package for network-integrated pathway enrichment analysis. Bioinformatics. 2020; 36(10):3283–3285. https://doi.org/10.1093/bioinformatics/btaa077 PMID: 32083639

104. Google. Google Scholar; 2022. Available from: https://scholar.google.com [cited 2022 Jan 31].

105. US National Library of Medicine (NLM). PubMed; 2022. Available from: https://www.ncbi.nlm.nih.gov/pubmed/ [cited 2022 Jan 31].

106. bioRxiv. The preprint server for biology; 2022. Available from: https://www.biorxiv.org [cited 2022 Jan 31].

107. arXiv q-bio. arXiv. Quant Biol; 2022. Available from: https://arxiv.org/archive/q-bio [cited 2022 Jan 31].

108. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min. 2017; 10(35):1–17. https://doi.org/10.1186/s13040-017-0155-3 PMID: 29234465

109. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012; 55(10):78–87.

110. Jones DT. Setting the standards for machine learning in biology. Nature Reviews Molecular Cell Biology. 2019; 20(11):659–660. https://doi.org/10.1038/s41580-019-0176-5 PMID: 31548714

111. Kueffner R, Zach N, Bronfeld M, Norel R, Atassi N, Balagurusamy V, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. Sci Rep. 2019; 9(1):690. https://doi.org/10.1038/s41598-018-36873-4 PMID: 30679616

112. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G, Harrow J, et al. DOME: Recommendations for supervised machine learning validation in biology. Nat Methods. 2021; 18(10):1122–1127. https://doi.org/10.1038/s41592-021-01205-4 PMID: 34316068

113. Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. ESC Heart Fail. 2021; 8(1):106–115. https://doi.org/10.1002/ehf2.13073 PMID: 33205591

114. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform. 2021; 153:104510. https://doi.org/10.1016/j.ijmedinf.2021.104510 PMID: 34108105

115. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet. 2021; 23:169–181. https://doi.org/10.1038/s41576-021-00434-9 PMID: 34837041

116. Lee BD, Gitter A, Greene CS, Raschka S, Maguire F, Titus AJ, et al. Ten quick tips for deep learning in biology. PLoS Comput Biol. 2022; 18(3):e1009803. https://doi.org/10.1371/journal.pcbi.1009803 PMID: 35324884