Published online 23 September 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 **1**
https://doi.org/10.1093/nargab/lqab087

DecoPath: a web application for decoding pathway enrichment analysis

Sarah Mubeen [1,2,3,*], Vinay S. Bharadhwaj [1,2], Yojana Gadiya [1,2], Martin Hofmann-Apitius [1,2], Alpha T. Kodamullil [1] and Daniel Domingo-Fernández [1,3,4,*]

[1]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53757, Germany, [2]Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany, [3]Fraunhofer Center for Machine Learning, Germany and [4]Enveda Biosciences, Boulder, CO 80301, USA

Received June 10, 2021; Revised August 31, 2021; Editorial Decision September 07, 2021; Accepted September 14, 2021

ABSTRACT

The past decades have brought a steady growth of pathway databases and enrichment methods. However, the advent of pathway data has not been accompanied by an improvement in interoperability across databases, hampering the use of pathway knowledge from multiple databases for enrichment analysis. While integrative databases have attempted to address this issue, they often do not account for redundant information across resources. Furthermore, the majority of studies that employ pathway enrichment analysis still rely upon a single database or enrichment method, though the use of another could yield differing results. These shortcomings call for approaches that investigate the differences and agreements across databases and methods as their selection in the design of a pathway analysis can be a crucial step in ensuring the results of such an analysis are meaningful. Here we present DecoPath, a web application to assist in the interpretation of the results of pathway enrichment analysis. DecoPath provides an ecosystem to run enrichment analysis or directly upload results and facilitate the interpretation of results with custom visualizations that highlight the consensus and/or discrepancies at the pathway- and gene-levels. DecoPath is available at https://decopath.scai.fraunhofer.de, and its source code and documentation can be found on GitHub at https://github.com/DecoPath/DecoPath.

INTRODUCTION

In recent years, high-throughput (HT) technologies have given rise to a perpetual influx of *-omics* data, requiring pragmatic approaches to sift out meaning. One of the most common applications of HT technologies is gene expression profiling to simultaneously determine the expression patterns of thousands of genes at the transcription level under certain conditions (1). While a host of statistical techniques are available to identify genes that differ in expression depending on a particular condition, gene set or pathway enrichment analysis methods represent a major class of tools researchers employ to group lists of genes into defined pathways and understand the functional roles of genes for any given set of conditions (2). To date, almost a hundred different pathway enrichment methods have been proposed, including the popular over-representation analysis (ORA) and gene set enrichment analysis (GSEA) (3). Though these methods may vary based on the overarching categories they fall into (e.g. topology versus non-topology-based) or the statistical techniques used, they have widely shown their ability to deconvolute biological pathways dysregulated in a given state (4).

Numerous pathway databases have been developed which aim at representing biological pathways from various vantage points (e.g. differing scopes, contexts, boundaries or pathway types). The existence of several hundreds of these databases reflects the inherent complexity and variability of biological processes that occur in living organisms (5). Further compounding this complexity is the fact that biological pathways housed in these databases are human constructs, delimited based on abstract boundaries defined by a researcher or the consensus of the community. This implies that a well-studied pathway could contain different biological entities depending on the boundaries defined by the databases that store it. These differences across databases can manifest in variability in the results of pathway enrichment analysis (6,7), in a similar way as methods can impact results (4,8–10).

Recent approaches to pathway enrichment analysis have focused on the integration of multiple datasets across different platforms to ensure a broader coverage of significantly enriched pathways (11–13). Other techniques attempt to

*To whom correspondence should be addressed. Tel: +49 2241 14 4036; Email: sarah.mubeen@scai.fraunhofer.de
Correspondence may also be addressed to D. Domingo-Fernández. Email: daniel.domingo.fernandez@scai.fraunhofer.de

account for potential differences that may arise in the results of pathway enrichment analysis by combining gene sets from several pathway databases. For instance, (14) presented an approach that leverages GSEA to calculate a combined enrichment score for multiple -*omics* layers using several databases. However, performing pathway enrichment analysis using multiple databases to increase the number of pathways covered can only partially address the challenges associated with variability in results. This is because such an approach falls short of leveraging the substantial overlap of pathway knowledge across databases which could provide more comprehensive results (15–17) or shed light on inconsistencies across pathway databases (18). Furthermore, combining several databases can result in redundant pathways, an issue tackled by the SetRank algorithm which discounts significant gene sets if their significance can be explained by their overlap with another gene set (19). Finally, a possible, natural solution to better connect and structure redundant information across databases lies in leveraging pathway ontologies (20) or pathway mappings with database cross-references (17). By connecting related pathways across databases, we can, in turn, investigate the consensus, or lack thereof, of the results of pathway enrichment analysis between databases or methods as demonstrated by several recent benchmarks (4,8–10).

Here, we present DecoPath, a web application that provides a user-friendly and interactive application to compare and interpret the results of pathway enrichment analysis yielded by different pathway databases. To facilitate the comparison of results across databases and bring to light possible contradictory results, we present several interactive visualization tools designed to better interpret the results of pathway enrichment at both the pathway and gene-level. While these visualizations can generally be used for any pathway enrichment method, DecoPath also integrates standard pathway enrichment methods in its pipeline, thus, enabling users to conduct an entire enrichment analysis on the web application (from data submission to interpretation). Finally, although DecoPath provides four default databases, it also allows users to upload gene sets and mappings such that analyses can be run on their independently curated gene sets.

MATERIALS AND METHODS

Implementation

The server-side was implemented in the Python programming language using the Django framework (https://www.djangoproject.com/). This framework operates using a Model-View-Controller (MVC) architecture and was integrated with Celery (http://www.celeryproject.org) and RabbitMQ (https://www.rabbitmq.com) for asynchronous task execution. The front-end of DecoPath comprises several interactive visualizations implemented using a collection of powerful Javascript libraries, including jQuery (https://jquery.com), D3.js (https://d3js.org/) and DataTables (https://datatables.net/). Furthermore, DecoPath relies on Bootstrap 4 (https://getbootstrap.com/) for the main design of the website. The web application is containerized using Docker for reproducibility purposes and easy deployment. We strongly recommend the use of DecoPath on Chrome, Firefox or Safari browsers and on Mac or Linux operating systems.

Pathway resources

DecoPath enables users to compare the results of enrichment analysis yielded using various pathway databases. As mentioned in the Introduction, pathways in different databases can substantially overlap, such that a pathway in one database can have counterparts in several others. Leveraging equivalent pathway mappings across several widely-used databases, DecoPath aims at highlighting the consensus, or lack thereof, of enrichment analysis results for each equivalent pathway. Expanding upon our previous work (17), we added novel equivalent pathway mappings as well as mappings for an additional database (i.e. PathBank (21)) (Supplementary Text). Thus, the released version of DecoPath provides users with the following pathway databases: KEGG (22), Reactome (23), WikiPathways (24) and PathBank (Retrieved 3 August 2020). Additionally, as integrative resources can lead to more biologically consistent results in enrichment analysis (6), a DecoPath-specific gene set database containing merged gene sets of equivalent pathways across the aforementioned databases is also provided, as described in the following section. Finally, in order to ensure that regular updates to these pathway resources are reflected in DecoPath, the software is updated with the latest gene sets annually.

Generating a pathway hierarchy

The consolidation of each of the pathway databases into a pathway meta-database was conducted in order to generate a pathway hierarchy. In doing so, equivalent representations of pathways across KEGG, PathBank, Reactome and WikiPathways were combined. The pathway hierarchy contains a total of 644 pathways from these four databases and can be found at https://github.com/ComPath/compath-resources/blob/master/mappings/decopath_ontology.xlsx (dated 13 January 2021). The hierarchy comprises eight major categories: metabolism, immune, signaling, communication and transport, cell death, disease, DNA repair and replication, and others. All pathways in the hierarchy retained their original identifiers except equivalent pathways which were merged and given unique names and identifiers. The pathway hierarchy is a directed acyclic graph with a maximum depth of 4, in which relation types between pathways can be either *is-part-of* or *equivalent-to* relations. The curation process to generate the hierarchy is described in the Supplementary Text. Periodic updates to the pathway hierarchy are made on an annual basis.

Pathway enrichment methods

DecoPath comprises two of the most widely used pathway enrichment methods (25–27): over representation analysis (ORA) and gene set enrichment analysis (GSEA) (3). ORA aims at identifying pathways (i.e. gene sets) that are over-represented within a list of genes of interest. A pathway is considered enriched (over-represented) if the *P*-value arising from a one-sided Fisher's exact test (28)

is lower than a specified threshold, typically 0.05. As this test is conducted for each pathway in the database, DecoPath's implementation of ORA corrects the *P*-value by applying multiple hypothesis testing correction with the Benjamini–Yekutieli method under dependency (29). The second method, GSEA, determines whether a pathway or a gene set significantly differs between two groups. A pathway is considered significantly regulated in that condition if genes of that pathway appear in the top or bottom ranking of a list of differentially expressed genes (DEGs) more than expected by chance. An alternative version of GSEA, namely GSEA Pre-Ranked (3), is also available if users wish to run GSEA on a pre-ranked list of genes. DecoPath uses implementations of GSEA and GSEA Pre-Ranked from *gseapy* (https://gseapy.readthedocs.io/en/latest). Additionally, DecoPath enables conducting differential gene expression (DGE) analysis between groups through DESeq2 (version 1.22.2). Apart from these methods, DecoPath also provides the option to include additional pathway enrichment methods into the web application.

Installation

Although we provide a freely available instance of DecoPath at https://decopath.scai.fraunhofer.de/, in the case of large datasets or cases where the compute capacity of the server may be insufficient depending on the type of analysis, users can install and use DecoPath in their own system. We offer two options to install DecoPath depending on the needs of the user. The first and easiest method for those unfamiliar with Django-based web applications is to install Docker and deploy the Docker container which will install required components and run the web application. Detailed instructions are provided on GitHub (https://github.com/decopath/decopath). Alternatively, DecoPath can be directly deployed following the instructions in the GitHub repository.

Runtime considerations

Computation time is dependent on the type of analysis, size of the datasets as well as the device specifications. ORA can be run on a gene list on a timescale of seconds and requires the relatively lowest usage of memory. A DGE analysis task has a timescale of several minutes, while GSEA on a typical expression dataset with two experimental groups and four databases can also be done within minutes with a dual-core Intel Core i5 CPU and 16 GB RAM.

Case scenario

Using each of the available enrichment methods, we demonstrate a typical workflow in DecoPath with the The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) dataset (30). Gene expression data from this dataset was retrieved from the Genomic Data Commons (GDC; https://gdc.cancer.gov) portal through the R/Bioconductor package, TCGAbiolinks (version 2.16.3; (31)) on 4 August 2020. To run GSEA, we employed RNA-Seq expression data normalized using Fragments Per Kilobase of transcript per Million mapped reads upper quartile (FPKM-UQ). DGE analysis using read counts from the TCGA-LIHC dataset (retrieved from the GDC; https://gdc.cancer.gov) was performed between normal and tumor samples to derive a gene list to conduct ORA. This final list of genes was restricted to genes that exhibited an adjusted *P*-value < 0.05. Specifications of the parameter settings for ORA and GSEA are listed in Supplementary Table S1.

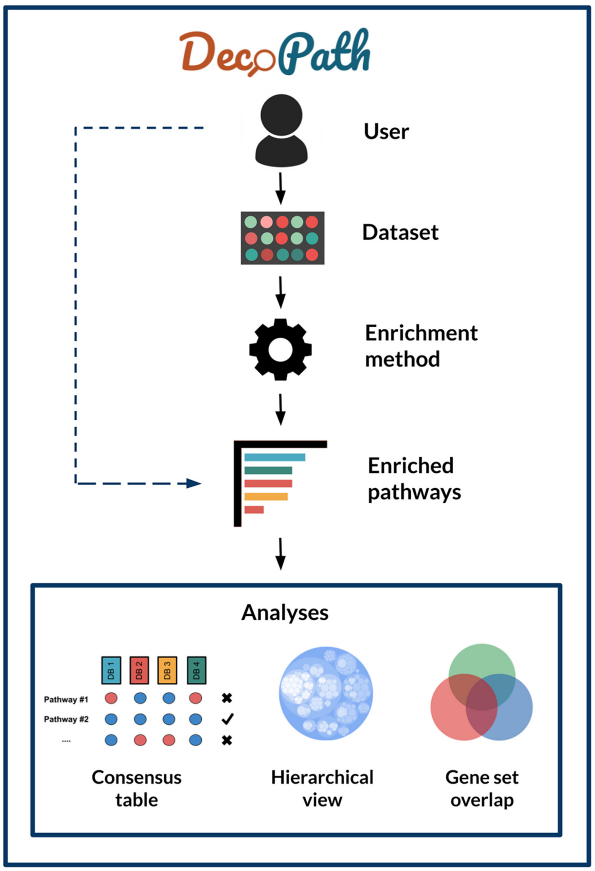

Figure 1. DecoPath workflow. Users can upload datasets to run pathway enrichment analysis or directly upload enrichment results from their own experiments. Once results have been loaded, DecoPath offers users several visualizations designed to evaluate pathway consensus at the database, hierarchy and gene set level. Users can also opt to directly upload results generated from varying enrichment methods across to visualize variations from these against a set of pathway databases.

RESULTS

Here, we describe the DecoPath web application. A typical workflow of the web application involves the submission of an experiment, generation of results, and the subsequent exploration and visualization of these results (Figure 1). In the following, we provide a detailed description for each of the steps in the workflow.

Submission form

Once a user has logged into DecoPath, on the Homepage, the input form allows them to upload their files and select parameters to run different analyses or upload results from them (Figure 2). For users opting to run analyses using De-

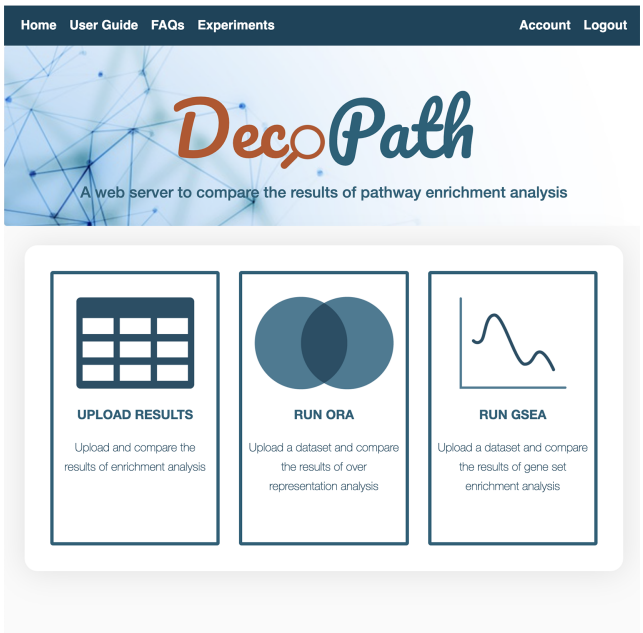

Figure 2. DecoPath homepage. Once a user has logged in, on the homepage, they are provided with the option to either run or submit the results of a pathway analysis. If a user opts to submit the results of an analysis, they can upload their data, select the databases they wish to include, choose the parameter settings for each experiment and optionally perform a concurrent DGE analysis. Once the form has been submitted, users are directed to the Experiments page where they can find visualizations and functionalities to compare and explore the consensus around different pathway databases.

coPath, the workflow depends on the analysis they select. Briefly, GSEA requires the submission of datasets, such as from RNA-Seq, microarray or ChIP-Seq, accompanied by a design matrix denoting the class labels (e.g. normal and tumor) for samples in the dataset. To run ORA, users need only submit a list of genes of interest. For either method, users can select which of the four pathway databases they would like to include in the analysis. By default, genesets from DecoPath which contain merged equivalent pathways are also included in the analysis.

These pathway enrichment methods can also be supplemented by DGE analysis to generate visualizations and identify genes that are differentially expressed according to a fold change cutoff. In order to run DGE analysis, unnormalized read counts in the form of a matrix of integer values is required, as is a design matrix, analogous to the one required for GSEA. For each of these analyses, gene identifiers should be in the form of HUGO Gene Nomenclature Committee (HGNC) symbols. Alternatively, users can opt to download gene set files for pathway databases included in DecoPath, run GSEA, ORA and/or DGE analysis, and upload the results of the analysis to the website. By directly uploading the results, users can also analyze the results of alternative enrichment methods such as EnrichNet (32) and Signaling Pathway Impact Analysis (SPIA) (33) using DecoPath. Detailed descriptions of the input files can be found in the User Guide and FAQs sections on our website.

Visualizations and analyses

Once users have submitted their query, they are directed to the Experiments page where they can view the status as well as details of their experiments, and explore and visualize their results (Figure 3). To interpret the results of enrichment analysis, we implemented multiple, customized tools intended to provide insights on the consensus across databases, each of which we detail below.

Exploring the consensus across pathway databases

The first visualization summarizes the consensus results of pathway enrichment analysis on multiple databases. For each pathway (row), the table shows the concordance across databases, reflected in terms of the significance value, specifically for ORA, and both the significance value and directionality of the normalized enrichment score (NES) for GSEA (Figure 4). Using this visualization, users can rapidly identify concordant (i.e. a given pathway is reported as significantly enriched in a gene list across all databases) and contradictory (i.e. a given pathway is reported as significantly enriched in a gene list in one or more databases, but not in the others [or vice versa]) pathways and directly compare their results.

We conducted a case scenario to investigate the results for ORA and GSEA using four pathway databases on the TCGA-LIHC dataset. Among the pathways enriched in ORA which could be found in more than one pathway database, we found 88 concordant pathways and 41 contradictory ones. Similarly, the results of GSEA revealed 70 concordant and 45 contradictory pathways. Among the contradictory pathways we observed in GSEA, the majority of contradictions pertained to whether or not the pathway was significantly enriched, while 12 pathways also differed in the sign of the NES (i.e. the same pathway was reported as enriched at the top of a ranked gene list for one database and at the bottom for another). Additionally, 53 concordant pathways were common between the results of GSEA and ORA; however, as expected, differences based on the pathway enrichment method were observed. Overall, the results of the LIHC-TCGA dataset for both methods showed that approximately one-third of equivalent pathways were contradictory across the two methods. Thus, the selection of databases, as well as the enrichment method, are important aspects in the experimental design of pathway enrichment analysis. We have observed that the use of one over another can yield discordant results, leading to different interpretations of results depending on the database choice. In the following sections, we illustrate why these results may be discrepant by analyzing the gene sets of a given pathway.

Visualizing consensus through the pathway hierarchy

In the second visualization, users can explore the results of their analysis within the context of a pathway hierarchy (see Materials and Methods section). This user-friendly and interactive visualization represents the different levels of the pathway hierarchy as circles, each of which represent a child or a parent pathway. In the case of GSEA, pathways that do not show statistically significant (adjusted P-value <0.05)

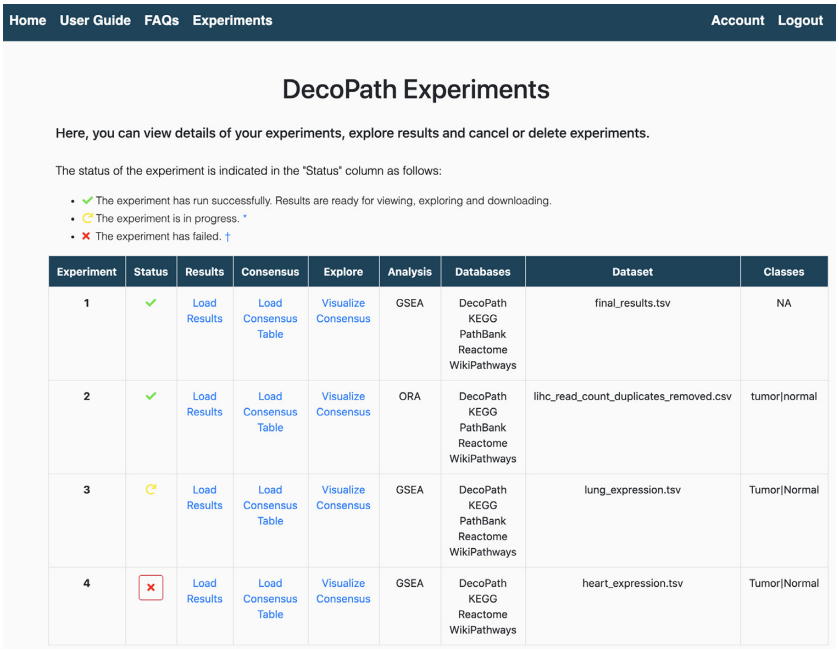

Figure 3. Experiments page. The Experiments page lists details of each of the experiments that were run or uploaded. The status of the experiment is given in the 'Status' column, indicating whether the experiment was successfully run, if it is pending or has failed. Through this page, users can then navigate to each of the different visualizations to explore the results of their analysis.

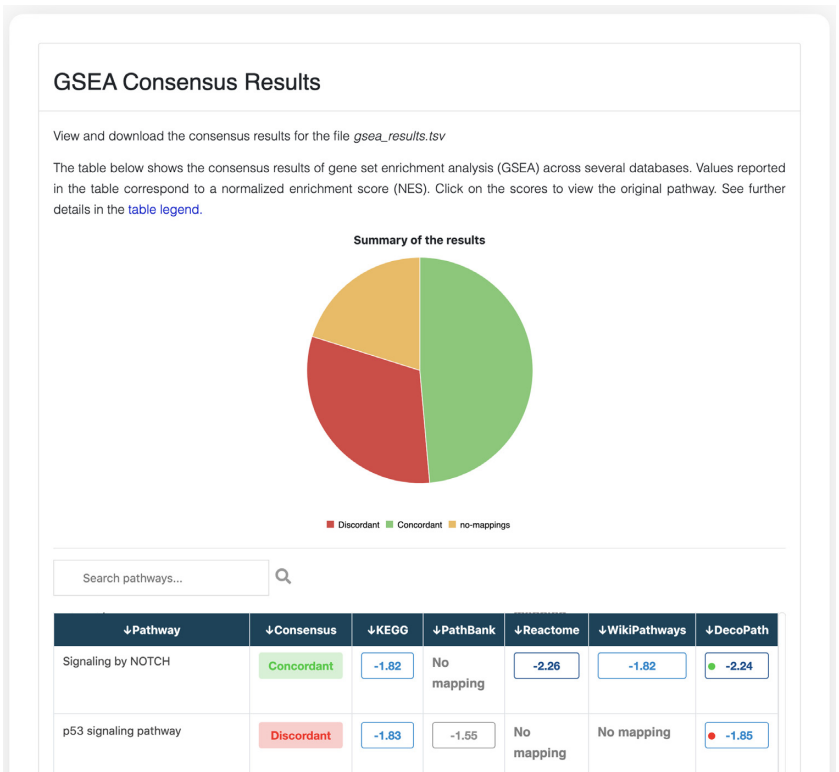

Figure 4. Consensus page. The Consensus page visualization shows the consensus of the results of enrichment analysis across databases at the pathway level. In the case of GSEA, the table displays the NES for a given pathway across each database as well as the NES of the merged gene sets of all equivalent pathways, the latter of which is indicated in the column 'DecoPath'.

differences between groups are colored gray, while statistically significant ones are colored red or blue based on the sign of the NES, and shaded by a gradient based on the magnitude of the NES. In the case of ORA, pathways are colored gray if they are not significant with an adjusted *P*-value < 0.05 and red otherwise. Additionally, the size of the gene sets for each of the pathways is proportional to the size of the circles. Furthermore, interactive visualizations also offer zoom and search functionalities to easily identify pathways of interest. In summary, with this tool, users can not only explore the enrichment results through the entire pathway hierarchy but also intuitively evaluate equivalent pathways and the size of the pathways, both of which are known to affect results (6,34).

Continuing the case scenario on the LIHC datasets, this visualization was used to identify major pathways that were enriched in both ORA and GSEA (Figure 5). The organization of pathways into eight major categories allows users to intuitively navigate through the hierarchy and identify pathway groups in which several pathways are enriched. For instance, among all pathways pertaining to metabolism, we observed that lipid and purine metabolism pathways were significantly enriched in both GSEA and ORA, indicating that there was a consensus across both methods and databases. Among other examples of consensus, we found cytokine signaling within the immune system pathways as well as MAP kinase signaling within the signaling pathways significantly enriched in all methods and databases. Finally, contrasting colors of this hierarchical view allow for the rapid identification of contradictory pathways which can then be further analyzed at the gene-level, aided by the following visualization.

Analyzing equivalent pathways at the gene level

The third visualization is an interactive Venn diagram that shows the overlap for equivalent pathways at the gene-level. In this visualization, we provide a means to analyze exactly which genes may explicate the findings of the pathway analysis. By clicking on the subsets of the Venn diagram, users can display the genes in each of the gene sets. Thus, users can pinpoint the specific genes of the pathway that might contribute to the contradictions observed in the results of the enrichment analysis. If fold changes have additionally been uploaded of DEGs or DGE analysis has been performed, users can also view the distribution of fold changes of genes in the dataset in an accompanying histogram.

To demonstrate this visualization, we explored both a pathway showing concordant results (i.e. DNA replication pathway) and another showing contradictory results (pyruvate metabolism) from the results of pathway enrichment on the TCGA-LIHC dataset. In the case of the DNA replication pathway, the results showed that the KEGG, Reactome and WikiPathways equivalent representations consistently reported NES over 2.0, suggesting that the pathway is regulated in the liver cancer dataset. We then explored the overlap of the gene sets of the DNA replication pathway from the three databases, observing that the $\log_2$ fold change values for the vast majority of genes in the pathway were positive. As GSEA finds the pathways which are nearest to the top (or bottom) of the ranked list of DEGs, this can account for the observance of the high NES (Figure 6A). Similarly, we explored a pathway (i.e. pyruvate metabolism), which had contradictory results in KEGG, Reactome and PathBank. In this case, these pathway databases disagreed in the direction of regulation of the NES; while the NES of pyruvate metabolism was positive in KEGG and PathBank, the sign of the NES was negative in Reactome. The consensus between KEGG and PathBank is not surprising as the gene sets of the pathway largely overlap (Figure 6B), while only 13 of the 31 genes in the Reactome pathway overlap with the other two gene sets. By plotting the distribution of the other 18 genes that are uniquely present in the Reactome pathway, we found that these genes were largely over-expressed, explaining the observed differences in the NES between them. Thus, this example illustrates how this tool can be used to assist in the interpretation of the discrepant results of pathway enrichment analysis.

DISCUSSION

While the popularity of pathway enrichment analysis for the interpretation of *-omics* data has grown over the past two decades and led to the development of over a hundred different methods, recent benchmarks have shown that the selected method can influence results (4,8,9,27). Furthermore, the majority of pathway enrichment analyses tend to be conducted on a single pathway database, the choice of which can also impact results of an analysis (6). While several tools have been implemented to run enrichment analysis on multiple platforms and methods (see Introduction), tools that facilitate the direct comparison of results yielded using different databases or enrichment methods at the pathway- and gene-levels are lacking. To address this issue, we have presented DecoPath, the first web application designed to assist in the interpretation of the results of pathway enrichment methods. DecoPath provides users with a broad range of built-in tools and visualization to conduct enrichment analyses and guide them in the interpretation of the results using multiple pathway databases.

Nonetheless, the presented web application is not without its limitations. First, while multiple enrichment methods exist, DecoPath only enables running two of the most popular pathway enrichment analyses. Similarly, DecoPath exclusively contains four pathway databases given the substantial curation effort required to map and harmonize pathway databases. To address these limitations, we enable users to directly upload results from other enrichment methods or pathway mappings from additional databases. Another limitation is the computational power of the server required to run experiments on datasets with a large sample size, or depending on the type of analysis conducted, may not be enough. However, since the source code of the web application is available (https://github.com/DecoPath/DecoPath) and DecoPath can be containerized in Docker, users can deploy the web application as per their needs to run more computationally demanding analyses.

In the future, we plan to map and integrate additional databases into DecoPath, as well as more enrichment methods. Furthermore, we envision the implementation of a consensus algorithm to combine the results obtained across multiple databases into a single score, in line with ap-

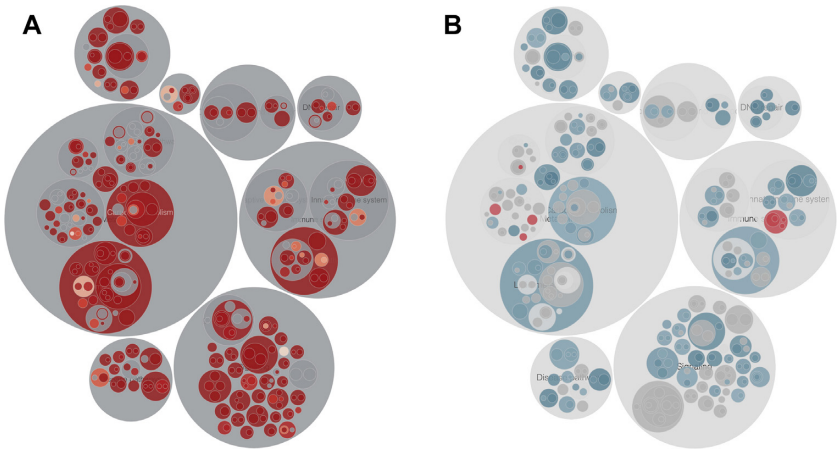

Figure 5. Circle pack visualization of the pathway hierarchy using different pathway enrichment methods. The figure corresponds to the interactive visualizations displaying the results of running ORA (**A**) and GSEA (**B**) on the LIHC dataset. In this visualization, results are customized based on the pathway enrichment method. In the case of Functional Class Scoring (FCS) and Pathway Topology (PT) based methods, the visualization highlights the direction of the dysregulation for each significantly dysregulated pathway as well as for the adjusted *P*-value (B). On the other hand, for ORA, the visualization highlights pathways that are significantly enriched based on an adjusted *P*-value (A).

Figure 6. Overlap of gene sets for a given pathway. Venn diagrams display the overlap of gene sets for equivalent pathways across user selected databases. By running DGE analysis, users can also view a histogram of the distribution of $\log_2$ fold changes for DEGs in their dataset to identify which genes are leading to either consistent or contradictory results for their pathway analysis. (**A**) Venn diagram of the overlap of gene sets for the DNA replication pathway from KEGG, Reactome and WikiPathways is shown above, while a histogram of $\log_2$ fold changes for DEGs from this pathway is shown below (in this example, the pathway representation from Reactome). (**B**) Venn diagram of the pyruvate metabolism pathway from KEGG, Reactome and PathBank and a histogram of $\log_2$ fold changes for DEGs for the pyruvate metabolism pathway Reactome are displayed.

proaches which integrate results obtained by an ensemble of enrichment methods, such as CGPS (35) and EGSEA (36), whilst taking into account variables such as gene set size and the magnitude of the enrichment score and/or *P*-value. Finally, we hope that our curation effort lays the groundwork for a future overarching pathway ontology with cross-references to databases that could be leveraged and extended by the pathway community.

DATA AVAILABILITY

A freely available instance of DecoPath can be found at https://decopath.scai.fraunhofer.de/.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We are very grateful to the curators of KEGG, Reactome, WikiPathways and PathBank for generating the raw content which was used in this work. Furthermore, we would like to thank Vasco Asturiano for developing *circle-packing,* the JavaScript library which is the basis of one of the visualizations of DecoPath.
Authors' contributions: D.D.F. conceived and designed the study. S.M. implemented the web application and analyzed the data with help from VSB and DDF. Y.G., S.M. and D.D.F. curated the pathway mappings. S.M. and D.D.F. wrote the paper. A.T.M., M.H.A., S.M. and D.D.F. acquired the funding.
All authors have read and approved the final manuscript.

FUNDING

This work was developed in the Fraunhofer Cluster of Excellence 'Cognitive Internet Technologies'.
Conflict of interest statement. D.D.F. received salary from Enveda Biosciences.

REFERENCES

1. Dillies,M.A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief Bioinform.*, **14**, 671–683.
2. Reimand,J., Isserlin,R., Voisin,V., Kucera,M., Tannus-Lopes,C., Rostamianfar,A., Wadi,L., Meyer,M., Wong,J., Xu,C. *et al.* (2019) Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, cytoscape and enrichmentmap. *Nat. Protoc.*, **14**, 482–517.
3. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
4. Nguyen,T.M., Shafi,A., Nguyen,T. and Draghici,S. (2019) Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol.*, **20**, 1–15.
5. Bader,G.D., Cary,M.P. and Sander,C. (2006) Pathguide: a pathway resource list. *Nucleic Acids Res.*, **34**, D504–D506.
6. Mubeen,S., Hoyt,C.T., Gemünd,A., Hofmann-Apitius,M., Fröhlich,H. and Domingo-Fernández,D. (2019) The impact of pathway database choice on statistical enrichment analysis and predictive modeling. *Front. Genet.*, **10**, 1203.
7. Bateman,A.R., El-Hachem,N., Beck,A.H., Aerts,H.J. and Haibe-Kains,B. (2014) Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci. Rep.*, **4**, 4092.
8. Geistlinger,L., Csaba,G., Santarelli,M., Ramos,M., Schiffer,L., Turaga,N., Law,C., Davis,S., Carey,V., Morgan,M. *et al.* (2020) Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform.*, **22**, 545–556.
9. Zyla,J., Marczyk,M., Domaszewska,T., Kaufmann,S.H., Polanska,J. and Weiner,J. 3rd (2019) Gene set enrichment for reproducible science: comparison of CERNO and eight other algorithms. *Bioinformatics*, **35**, 5146–5154.
10. Mathur,R., Rotroff,D., Ma,J., Shojaie,A. and Motsinger-Reif,A. (2018) Gene set analysis methods: a systematic comparison. *BioData Min.*, **11**, 1–19.
11. Griss,J., Viteri,G., Sidiropoulos,K., Nguyen,V., Fabregat,A. and Hermjakob,H. (2020) ReactomeGSA-Efficient multi-omics comparative pathway analysis. *Mol. Cell Proteomics*, **19**, 2115–2124.
12. Paczkowska,M., Barenboim,J., Sintupisut,N., Fox,N.S., Zhu,H., Abd-Rabbo,D., Mee,M.W., Boutros,P.C. and Reimand,J. (2020) Integrative pathway enrichment analysis of multivariate omics data. *Nat. Commun.*, **11**, 1–16.
13. Zhou,Y., Zhou,B., Pache,L., Chang,M., Khodabakhshi,A.H., Tanaseichuk,O., Benner,C. and Chanda,S.K. (2019) Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.*, **10**, 1–10.
14. Canzler,S. and Hackermüller,J. (2020) multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinform.*, **21**, 1–13.
15. Stobbe,M.D., Houten,S.M., Jansen,G.A., van Kampen,A.H. and Moerland,P.D. (2011) Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Syst. Biol.*, **5**, 165.
16. Belinky,F., Nativ,N., Stelzer,G., Zimmerman,S., Iny Stein,T., Safran,M. and Lancet,D. (2015) PathCards: multi-source consolidation of human biological pathways. *Database*, bav006.
17. Domingo-Fernández,D., Hoyt,C.T., Bobis-Álvarez,C., Marín-Llaó,J. and Hofmann-Apitius,M. (2018) ComPath: An ecosystem for exploring, analyzing, and curating mappings across pathway databases. *npj Syst. Biol. Appl.*, **4**, 43.
18. Mora,A. and Donaldson,I.M. (2012) Effects of protein interaction data integration, representation and reliability on the use of network properties for drug target prediction. *BMC Bioinform.*, **13**, 1–17.
19. Simillion,C., Liechti,R., Lischer,H.E., Ioannidis,V. and Bruggmann,R. (2017) Avoiding the pitfalls of gene set enrichment analysis with setrank. *BMC Bioinform.*, **18**, 151.
20. Petri,V., Jayaraman,P., Tutaj,M., Hayman,G.T., Smith,J.R., De Pons,J., Laulederkind,S.J., Lowry,T.F., Nigam,R., Wang,S.J. *et al.* (2014) The pathway ontology–updates and applications. *J. Biomed. Semant.*, **5**, 7.
21. Wishart,D.S., Li,C., Marcu,A., Badran,H., Pon,A., Budinski,Z., Patron,J., Lipton,D., Cao,X., Oler,E. *et al.* (2020) PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res.*, **48**, D470–D478.
22. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
23. Fabregat,A., Korninger,F., Viteri,G., Sidiropoulos,K., Marin-Garcia,P., Ping,P., Wu,G., Stein,L., D'Eustachio,P. and Hermjakob,H. (2018) Reactome graph database: Efficient access to complex pathway data. *PLoS Comput. Biol.*, **14**, e1005968.
24. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K., Miller,R.A., Digles,D., Lopes,E.N., Ehrhart,F. *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
25. García-Campos,M.A., Espinal-Enríquez,J. and Hernández-Lemus,E. (2015) Pathway analysis: state of the art. *Front. Phys.*, **6**, 383.
26. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.*, **8**, e1002375.

27. Xie,C., Jauhari,S. and Mora,A. (2021) Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinform.*, **22**, 1–16.
28. Fisher,R.A. (1992) Statistical methods for research workers. *Breakthroughs in Statistics*. Springer, New York, NY, pp. 66–70.
29. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
30. The Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113.
31. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D. and Noushmehr,H. (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71–e71.
32. Glaab,E., Baudot,A., Krasnogor,N., Schneider,R. and Valencia,A. (2012) EnrichNet: network-based gene set enrichment analysis. *Bioinformatics*, **28**, i451–i457.
33. Tarca,A.L., Draghici,S., Khatri,P., Hassan,S.S., Mittal,P., Kim,J.S., Kim,C.J., Kusanovic,J.P. and Romero,R., (2008) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.
34. Karp,P.D., Midford,P.E., Caspi,R. and Khodursky,A. (2021) Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics*, **22**, 1–11.
35. Ai,C. and Kong,L. (2018) CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *J. Genet. Genomics*, **45**, 489–504.
36. Alhamdoosh,M., Ng,M., Wilson,N.J., Sheridan,J.M., Huynh,H., Wilson,M.J. and Ritchie,M.E. (2017) Combining multiple tools outperforms individual methods in gene set enrichment analyses. *Bioinformatics*, **33**, 414–424.