

SOFTWARE

Open Access



GeDi: simplifying gene set distances for enhanced omics interpretation in R/Bioconductor

Annekathrin Silvia Nedwed^{1*}, Arsenij Ustjanzew¹, Najla Abassi¹, Leon Dammer¹, Alicia Schulze¹, Sara Salome Helbich^{2,3}, Michael Delacher^{2,3}, Konstantin Strauch¹ and Federico Marini^{1,3,4*}

*Correspondence:

Annekathrin Silvia Nedwed
anedwed@uni-mainz.de

Federico Marini
marinif@uni-mainz.de

¹Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University, Mainz, Germany

²Institute of Immunology, University Medical Center Mainz, Mainz, Germany

³Research Center for Immunotherapy (FZI), Mainz, Germany

⁴Center for Thrombosis and Hemostasis (CTH), Mainz, Germany

Abstract

Background Functional enrichment analysis is a standard component in many omics data analysis workflows, supported by a variety of methods and algorithms. However, despite their utility and wide application, these methods often return the results as an extensive and redundant list of gene sets, impeding interpretation and hypothesis generation. Moreover, network based information can provide additional biological context through functional interaction data, yet this is often overlooked by existing tools.

Results We developed GeDi, an R/Bioconductor package designed to streamline and standardize the interpretation of functional enrichment results. GeDi aggregates gene sets into biologically meaningful clusters using a suite of gene set distance metrics and clustering algorithms, aimed to reduce redundancy and improve clarity. GeDi also enables the integration of protein–protein interaction (PPI) data, through the implementation of a weighted distance metric, providing a richer biological context by capturing functional connectivity between pathways and their components. The package offers visualizations, aggregation, and automated reporting, and is available as both a stand-alone R-package and an interactive Shiny application.

Conclusion GeDi facilitates clearer, faster interpretation of enrichment results by combining clustering and network context. Application to a public RNA-seq dataset revealed coherent biological themes, supporting both experimental and computational research. GeDi is freely available in the Bioconductor project under the MIT license (<https://bioconductor.org/packages/GeDi>), and a demo instance is accessible on the Shiny server (<http://shiny.imbei.uni-mainz.de:3838/GeDi>).

Keywords RNA-sequencing, Functional enrichment analysis, Data interpretation, Data visualisation, R package, Shiny

Background

One of the primary objectives in biological research is to uncover the molecular mechanisms driving phenotypes, traits and disease states [1, 2]. In bioinformatics, this is often done by means of functional enrichment analysis, which identifies pathways or biological



processes that are statistically overrepresented in gene or protein sets defined by shared characteristics (e.g. differential expression or genetic variation) [2–4]. As functional enrichment analysis became a fundamental step of modern omics workflows [5], various computational strategies have emerged, typically classified into over-representation analysis (ORA), functional class scoring (FCS), and pathway topology (PT)-based methods [2, 6, 7]. In general, ORA focuses on identifying gene sets with an overrepresentation of differentially expressed genes, while FCS methods examine the distribution of these genes within a ranked list [4]. PT-based methods additionally incorporate pathway structure information to evaluate the impact of expression changes in genes and proteins [7].

In the broader scientific landscape, ORA methods remain the most widely adopted functional enrichment approaches due to their simplicity, efficiency, and broad applicability across biological research [2]. While pathway-topology methods offer deeper mechanistic insights by incorporating pathway structure [6], the widespread availability and support for ORA has established it as the standard in many workflows. Popular tools such as topGO [8], g:Profiler [9, 10] and clusterProfiler [11] demonstrate this, and offer access to diverse annotation databases like Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome and the Molecular Signatures Database (MSigDB) [12–15].

However, the output of most ORA methods typically consists of long, redundant gene lists, often complicating interpretation [16, 17]. Despite the existence of tools such as EnrichmentMap [17], GeneSetCluster [18] or GScluster [19], which focus on the aggregation of similar gene sets into clusters, interpretation is still mostly done manually due to the absence of biological context, which could be obtained from network data, such as protein–protein interactions (PPIs). While tools such as GScluster [19] address this through the implementation of a PPI-weighted gene set distance score, the tool is limited by the amount of species supported in the analysis.

In this work, we present GeDi, an R/Bioconductor package designed to streamline the aggregation and interpretation of functional enrichment analysis results. This is achieved by computing Gene Set Distance scores to quantify relationships between reported gene sets, which then serve as the basis for clustering to group similar sets, enhance clarity, and reduce redundancy. In order to accommodate a variety of different research scenarios, GeDi offers a range of distance metrics and clustering methods, as well as the integration of biological network data (e.g., PPI) to provide additional biological context. The package can be used programmatically or interactively via a Shiny app (provided within the package itself), using plain text files, spreadsheets or common R/Bioconductor data structures as input [20]. This makes GeDi directly compatible with commonly used functional enrichment implementations, allowing for a seamless integration into various analysis workflows. As an essential downstream data exploration step, GeDi enables a more structured interpretation of results through visualization and aggregation of the large-scale enrichment results, ultimately reducing manual effort and facilitating hypothesis generation.

GeDi is freely available under the MIT license on Bioconductor (<https://bioconductor.org/packages/GeDi>), with the development version accessible on GitHub (<https://github.com/AnnekathrinSilvia/GeDi>) and a demo instance available at <http://shiny.imbei.uni-mainz.de:3838/GeDi>.

Implementation

General design

The GeDi package is implemented in the R programming language within the Bioconductor project. We used the Shiny framework [21] to implement an application which can be run locally on the infrastructure of each individual user. Additionally, GeDi can be hosted on a server, as showcased in the demo instance (<http://shiny.imbei.uni-mainz.de:3838/GeDi>), making it suitable for internal group use by multiple users, or in situations where the exchange of sensitive personal data is necessary.

GeDi is designed to integrate seamlessly into existing omics workflows by facilitating the interpretation of functional enrichment results, such as those typically produced in bulk or single-cell RNA-seq analyses [22–24]. To support a broad usability, GeDi accepts input in plain text, spreadsheets, R data.frames or as a GeneTonicList object [25] (Fig. 1A). The only mandatory input requirement is a table with two columns named “Genesets” and “Genes”. While the format is flexible, we recommend using standardized identifiers such as HGNC [26], ENSEMBL [27], or Gencode [28] for genes, and pathway IDs from GO [12], KEGG [14], Reactome [13], or MSigDB [29] for gene sets.

Additionally, GeDi can accept information on the protein–protein interactions as optional input (Fig. 1B). While the PPI information can be fetched and cached locally during runtime, providing the data upon launching of the app can reduce manual intervention, thus enhancing the reproducibility of the analysis.

The GeDi package provides nearly all of its functionality through a single entry point: the GeDi() function. This launches a web application that guides users through the exploration of functional enrichment results using a clear, section-based layout (Fig. 1C). Each panel of the app focuses on a specific aspect of the data, offering tailored visualizations

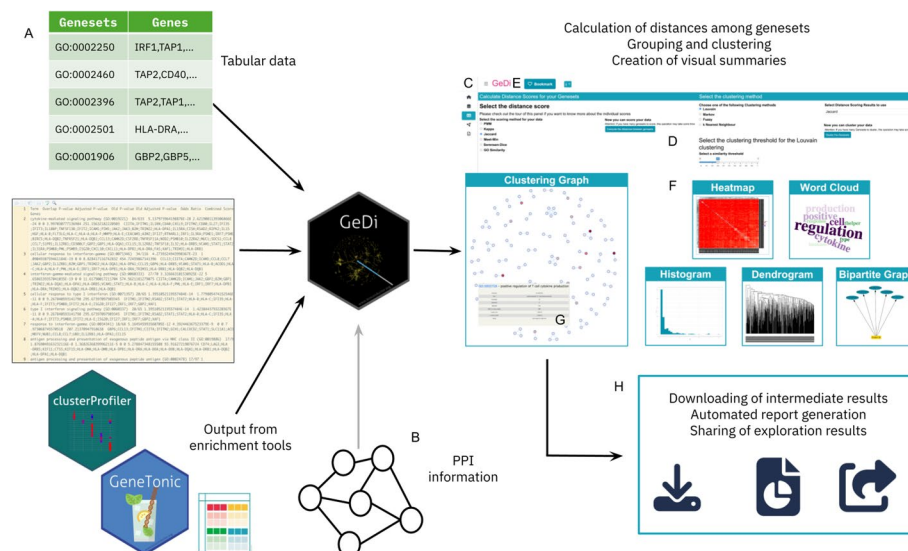


Fig. 1 Overview of analysis workflow with GeDi. **A** GeDi supports multiple input formats, including tabular data, enrichment tool outputs (e.g. topGO, clusterProfiler, etc.) and the GeneTonicList object format. **B** Users may optionally provide their own PPI data instead of downloading it through the app. **C** Overview of the individual interface panels. **D** The user interface features intuitive selectors and buttons to guide users through the application. **E** The Bookmark button allows marking of interesting genes and gene sets. **F** GeDi offers a variety of visualisation options. **G** A hover feature enables detailed inspection of gene sets. **H** GeDi provides options to save intermediate results and generate a report of the exploration session for reproducibility and result sharing

and summaries. The interface is built using the bs4Dash [30] package, ensuring a modern, responsive user experience.

To support intuitive onboarding and enhance user engagement, GeDi includes interactive, guided tours implemented with the rintrojs package [31]. These tours accompany users through the different components of the app, providing context-sensitive help and highlighting individual UI elements through modals.

To prevent overwhelming users, the app design emphasizes clarity and flow, with panels using collapsible elements, often revealed sequentially. This structure not only makes the app easier to use but also supports a more effective, step-by-step exploration of complex datasets (Fig. 1D).

Users can provide input data at the beginning of a session, or interactively during usage. The application also features a bookmarking system, allowing users to flag genes and gene sets of interest throughout their exploration (Fig. 1E). Bookmarked items are summarized in a dedicated *Report* panel and automatically included in the downloadable report generated at the end of the session. Buttons for bookmarking are consistently located in the panel headers, making the process straightforward.

While designed for interactive use, most functions are also exported and can be reused in scripted analyses (e.g., within RMarkdown or Jupyter notebooks), making GeDi flexible for both hands-on exploration and reproducible workflows.

Typical usage workflow

GeDi requires as input the results of a functional enrichment analysis, comprising at least two columns named “Genesets” and “Genes”. These can be supplied directly via the function call `GeDi()` with a corresponding `data.frame`, or uploaded through the *Data Upload* panel in the GeDi interface, supporting plain text, spreadsheet, or RDS formats.

GeDi can be readily integrated as a downstream step in workflows that include functional enrichment analysis, such as for example (bulk) RNA-sequencing data analysis workflows such as those demonstrated in Additional File 1 and 2 of the Supplementary Material. For this, GeDi implements the `path_to_GeDi()` function, which converts results from common enrichment tools, such as `topGO`, `clusterProfiler`, `ReactomePA`, `enrichR` and `fgsea`, into the expected input format without the need for manual adjustments. This enables an even smoother integration of the package into existing analysis pipelines.

In addition to the core input, GeDi optionally accepts a matrix of protein–protein interaction scores (`ppi_df`) and a distance matrix between gene sets (`distance_scores`). Both can be generated within GeDi but may incur significant computation time; saving and reusing these objects is therefore advisable. Download options for these are available within the interface. To facilitate broad applicability, GeDi currently supports outputs from widely used R packages for enrichment analysis, including `topGO` [8] and `clusterProfiler` [11]. Alternatively, users are encouraged to follow the GeneTonic workflow [25], which would generate and preprocess results in a format that is directly supported by GeDi.

While GeDi only requires gene set and gene information, the inclusion of additional metadata from enrichment analysis tools — such as gene set descriptions, p-values, or set sizes — can significantly enhance the user experience, e.g., through coloring of the visualisations according to the metadata (Fig. 1F, G). For example, such metadata enables

richer visualizations, including node coloring in graphs, and improves interpretability within the *Clustering Graph* panel. GeDi supports iterative data exploration through a range of interactive visual summaries, including heatmaps, gene set graphs, bipartite graphs, and word clouds (Fig. 1F, G). The computed distance scores serve as the basis for clustering, helping to group related gene sets and reveal broader functional patterns. Users can also hover on the visualisations in GeDi to explore additional information, for example through dynamic links to external databases such as AmiGO [32], KEGG [14], or Reactome [13], which support deeper biological interpretation.

Although GeDi is primarily accessed via its interactive web interface, all functions operate on standard R objects, allowing straightforward integration into existing R/Bioconductor workflows (e.g., [20, 23, 24]). A recommended final step in a typical analysis with GeDi is the generation of an HTML report, which facilitates both result sharing and computational reproducibility (Fig. 1H).

Results

GeDi implements distances and clustering approaches to summarize enrichment results

GeDi facilitates the interpretation of functional enrichment results by grouping gene sets into functionally coherent clusters based on their pairwise (dis)similarity. This is achieved through the computation of gene set distance scores, which quantify how similar or dissimilar gene sets are in terms of gene content, semantic meaning, or network context. GeDi implements a range of distance metrics, which can be broadly categorized into: (1) set-based metrics (Jaccard [33, 34], Sørensen-Dice [35], Kappa [36], and Meet-Min (MM) [19]), relying solely on gene set overlap; (2) hybrid metrics (protein-protein interaction weighted Meet-Min distance score (pMM, [19])), which combine gene set overlap with protein-protein interaction (PPI) context; and (3) semantic metrics (GO dissimilarity [37]), which exploit the structure of the Gene Ontology to assess functional similarity. Each metric captures a different aspect of relatedness: while set-based scores are sensitive to direct gene set overlap, pMM adds biological relevance by considering interaction networks, and semantic metrics provide a broader, ontology-driven similarity measure.

Based on the computed distances, GeDi applies clustering algorithms to aggregate gene sets into groups that represent shared biological functions. Available algorithms include Louvain [38] and Markov [39, 40] clustering, which detect communities in similarity networks, PAM (Partitioning Around Medoids [41]), a centroid-based method, and Fuzzy clustering [42, 43], which allows overlapping cluster membership. The choice of clustering algorithm affects the structure and granularity of the resulting clusters: Louvain and Markov typically yield discrete, non-overlapping groups, while Fuzzy clustering can capture functional cross-talk by assigning gene sets to multiple clusters. These complementary options enable users to explore different aspects of the functional landscape represented in enrichment results.

A detailed description and definition of the distance scores as well as of the clustering algorithms can be found in Additional File 3.

Comparison of Meet-Min score and pMM score

In a further inspection and analysis of the methods and functionality provided by GeDi, we inspected how the additional network information incorporated in the pMM can

leverage the score and results. For this purpose, we investigated how a random removal of genes from the gene sets would influence the resulting scores and clusters.

Hence, we first removed one to 15 genes from gene sets with more than 10 genes, removing at most half of the genes in any given set. Then we scored the data using either the pMM or the MM score. Afterwards, we iterated 100 times the community-detection Louvain clustering on the original data as well as on the randomly reduced data, as this algorithm is inherently non-deterministic. Afterwards, we compared the clustering results using the Adjusted Rand Index (ARI) coefficient as a measure of similarity between two data clusterings [44].

In this specific use case, we used the ARI coefficient to compare the clustering results of the reduced datasets to the original clustering result on the whole set of gene sets. To compare the two distance metrics, we plotted the resulting ARI scores for the MM and pMM score, which can be observed in Supplementary Fig. 1. Upon closer inspection, we observed that the pMM ARI is almost exclusively higher than the MM ARI score. This indicates that the network information included in the pMM score can help in retaining the information that would be lost by the removal of genes, which would significantly impact the computed MM scores.

GeDi enables the quantification and navigation of gene set similarities after functional enrichment analyses

Following the differential expression and functional enrichment analysis of various omics workflows, GeDi provides an ideal additional step in facilitating the exploration and interpretation of the resulting enriched pathway data. The functional enrichment results of several tools can be directly provided as input to GeDi, provided that they follow the expected input format of the package. However, if the input format does not adhere to the requirements, GeDi also implements functionalities to transform and adapt it.

As described above, GeDi provides a variety of gene set distance scores to quantify their (dis)similarity between individual gene sets. In addition, to represent these calculated distance scores, GeDi offers three main visualisation types: a heatmap, a dendrogram, and a network-like representation. The heatmap offers a complete view of all pairwise distances, allowing users to detect global patterns through color gradients, while the dendrogram adds a hierarchical perspective revealing nesting groups. Lastly, in the network-like representation, gene sets are represented as nodes and edges are drawn between gene sets with a distance below a user defined threshold. The network can highlight functional modules and local neighbourhoods, whose granularity can be adjusted through the threshold parameter. Together, these visualisation options provide some initial global and focused interpretation of the overarching patterns of regulation in the enrichment results.

These patterns can be further inspected through the application of one of the available clustering algorithms in GeDi. The results of the clustering process can be visualised in different ways such as in a graph network, a bipartite graph, or word clouds. For the graph network, nodes represent gene sets and edges are drawn between gene sets belonging to the same cluster, while in the bipartite graph, nodes represent gene sets and clusters, with edges connecting gene sets to their respective cluster(s). Lastly, in the word cloud representation, users can select a specific cluster and the word cloud will

visualise the most prominent biological terms associated with the gene sets in the cluster. However, this is dependent on the availability of a (biological) description of each gene set in the input meta data. The interpretation of groups and communities of gene sets can be supported by network-based statistics, as also suggested by [45], complementing the visual summaries in a more quantitative manner.

Additionally, GeDi provides functionality to export the exploration session in a ready-to-share HTML report. This report includes tables of the original input data, as well as the selected parameters during the session (such as distance scoring method, clustering algorithm and thresholds) as well as the visualisation of the calculated distance scores and clustering results. This report does not only serve as a documentation on the conducted exploration, but also as a means to share and reproduce the analysis.

Applying GeDi to publicly available data to enhance the interpretation of functional enrichment results

We use a publicly available bulk RNA-seq dataset, published by Delacher et al. [46] (GEO accession: GSE130842; <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130842>), to demonstrate GeDi's functionality. The dataset comprises 56 samples from various murine tissues, including distinct regulatory T cell (Treg) subpopulations. We focused on Klr $g1^{-}Nfil3(GFP)^{-}$ and Klr $g1^{+}Nfil3(GFP)^{+}$ Tregs, yielding a subset of eight samples. For the preprocessing, we followed the workflow proposed by [24], documented in Additional File 1 and on Github (https://github.com/AnnekathrinSilvia/GeDi_supplement). Ultimately, we generated a GeneTonicList object [25], which we used as an input for GeDi.

Providing the data to GeDi

We initiated the analysis by launching GeDi via the GeDi() command and loading the prepared GeneTonicList object into the launched app. In a preprocessing step, we first filtered out GO terms associated with ≥ 200 genes, reducing the total from 500 to 421 gene sets. Subsequently, we retrieved the protein–protein interaction matrix from STRING (version 12.0.0), using GeDi's built-in functions. A detailed workflow of this can be found in Additional File 1 or on Github (https://github.com/AnnekathrinSilvia/GeDi_supplement).

Quantifying gene set similarity using gene set distance scores

In this study, we systematically evaluated all distance metrics available in GeDi, ultimately focusing on the pMM and GO semantic distances, which consistently produced medium-sized, biologically meaningful clusters.

In Fig. 2, we visualised the resulting distance scores using the heatmap visualisation available in GeDi. Both heatmaps show that independent of the used distance metric, most of the gene sets had a high distance close to 1, with distinct clusters of smaller distances visible along the diagonal (Fig. 2A, B). However, the pMM heatmap (Fig. 2A) revealed more concentrated blocks of similar gene sets, especially in the upper-left and lower-right corners of the heatmap, compared to the GO heatmap (Fig. 2B). This was already a first indicator of the additional information gained through the incorporation of network-information from PPI-networks.

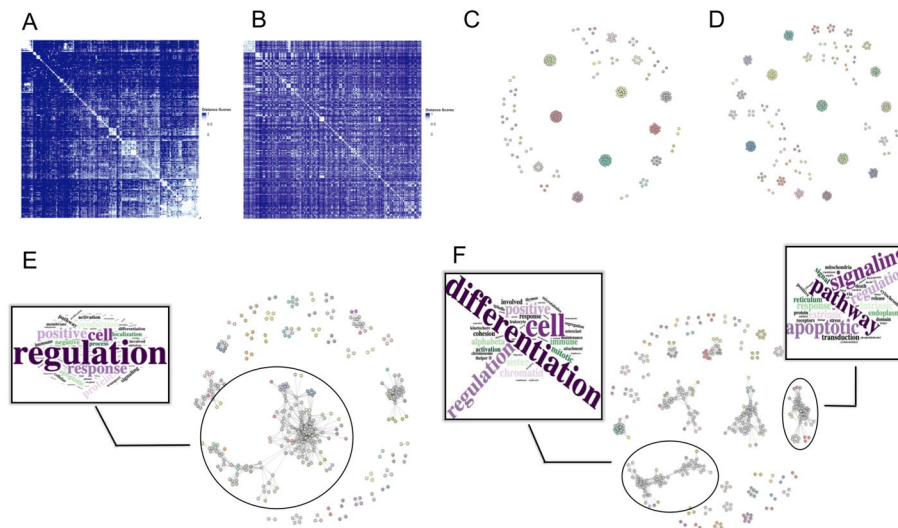


Fig. 2 Exemplary output generated with GeDi based on pMM and GO distance score for the Delacher et al. dataset. **A** Heatmap of the calculated pMM distance scores. **B** Heatmap of the calculated GO distance scores clustered by the row order of the pMM heatmap in A. **C** Louvain clustering results based on the pMM distance scores. **D** Louvain clustering results based on the GO distance scores. **E** Fuzzy clustering results based on the pMM distance scores. **F** Fuzzy clustering results based on the GO distance scores

Aggregating gene sets into similar groups using clustering algorithms

Following the distance computation, we selected two clustering algorithms, Louvain and Fuzzy clustering, to showcase the capabilities of GeDi. For the thresholds, we consistently chose 0.5 for the Louvain clustering as well as the Fuzzy similarity, membership and clustering threshold.

In Fig. 2, the resulting clustering graphs of the two clustering approaches can be seen, with the clustering of the pMM scores in Fig. 2C and E and the GO distances in Fig. 2D and F. We can observe that the Louvain clustering algorithm resulted in similar numbers of clusters for both the pMM and GO scores (41 vs. 49 clusters, Fig. 2C, D). Both graphs show an overall composition of smaller clusters, with some larger clusters in between. Using the hovering tooltip of GeDi (Fig. 1G), we identified that these larger clusters captured biological coherent themes, including T cell differentiation, cell division, and protein localization.

The Fuzzy clustering on the other hand resulted in an overall higher number of clusters (127 clusters for both distance metrics), of which many shared cluster members. The largest clusters in the middle of the graphs showed to be enriched with genes involved in immune regulation and T cell response (Fig. 2E). We could also find biological processes of T cell differentiation and apoptotic signaling pathways in these clusters, which was more prominently observable in the clustering results on the GO distances, where these biological processes were split into two larger clusters (Fig. 2F).

Overall, the recurring biological themes—most prominently, T cell differentiation and proliferation, as well as protein localisation and apoptotic signaling pathways—align well with the original findings and biological context of the used dataset. Delacher et al. [46] identified $\text{Klrg1}^-\text{Nfil3}^-$ and $\text{Klrg1}^+\text{Nfil3}^+$ Treg cells as developmental stages within the lineage of tissue-resident Tregs, a subpopulation involved in tissue regeneration [46, 47]. The observed enrichment of differentiation- and proliferation-related pathways reflects

the transition from Klr $g1^{-}Nfil3^{-}$ to Klr $g1^{+}Nfil3^{+}$ late precursors cells, driven by increased expression of Klr $g1$ and Nfil 3 . Concurrently, changes in surface protein and antigen expression during this process explain the strong representation of pathways linked to protein localization and transport. With the help of GeDi, we were able to enhance and highlight these biological processes inherent in the data, ultimately allowing for a facilitated and refined interpretation of the original functional enrichment data.

Discussion

Functional enrichment remains a fundamental step of modern omics analysis pipelines, yet the interpretability of its results continue to pose challenges. Standard methods often yield long, redundant gene set lists that obscure biological signals, ultimately hindering hypothesis generation. While similarity- and clustering-based summarization approaches exist, they are frequently limited in scope or interactivity (Supplementary Table 1), and many enrichment tools still rely on manual inspection, lacking meaningful integration of biological context. Additionally, in a recent article, Wijesooriya et al. [48] found that researchers tend to use these tools in an improper manner, while also not reporting methodological details appropriately.

GeDi addresses these limitations by offering a unified framework that integrates multiple gene set distance metrics, various clustering algorithms, and interactive visualization tools to help users explore and interpret enrichment results in a structured manner. While we currently support a broad range of metrics, the modular architecture of GeDi will allow us to incorporate future methods, ensuring long-term compatibility and relevance as enrichment methodologies evolve. Rather than aiming to outperform existing summarization or visualization tools, we developed GeDi to unify these functionalities within the familiar R/Bioconductor environment. By integrating multiple distance metrics and clustering options for genesets into a single, interoperable framework used for data preprocessing, differential analysis, and visualization, GeDi minimizes the workflow fragmentation and enhances reproducibility, ultimately enabling researchers to explore and interpret enrichment results more effectively.

Designed with both flexibility and usability in mind, GeDi can be used programmatically within notebooks and scripts, or interactively through its Shiny application. The app offers powerful drilldown capabilities that allow users to interrogate specific gene sets, explore annotation details, and access external resources, facilitating deeper biological interpretation. While providing network-based statistics as a useful level of summarization and a starting point for biological interpretation, we want to acknowledge that the best approach to take is still specific to each individual dataset, within iterations that can be simplified thanks to our tool.

Sessions can be concluded with automated report generation, enabling reproducible documentation of each analysis step. Intermediate results are passed along transparently, allowing users to resume or branch off from any point in the workflow. This design supports both exploratory and structured research scenarios, which has been shown essential in other works [24, 25] and can be seen as a democratizing factor in interdisciplinary efforts.

While there is a wide variety of exploration tools for enrichment results, such as for example EnrichmentMap [17], most of these tools only offers functionality to visualise

the results, without any functionality offered to reduce the complexity of the result list (Supplementary Table 1). While there are other tools, such as GScluster or GeneSetCluster, which have addressed these aspects of gene set aggregation (Supplementary Table 1), they often impose stricter input requirements or lack integration with the broader R/Bioconductor ecosystem. In contrast, GeDi requires minimal input—only a two-column table of genesets and genes—yet seamlessly accepts widely used outputs (e.g. from topGO, clusterProfiler or other apps such as GeneTonic which aim to reuse standardized objects and classes). This positions GeDi as an accessible and robust post-enrichment summarization tool for a broad user base.

While GeDi relies on curated gene annotations and known interaction data to define relationships between gene sets, alternative approaches have been proposed that infer inter-set similarity directly from expression data using multivariate methods such as canonical correlation analysis, gene set coexpression analysis (GSCA), or gene set network correlation analysis (GSNCA) [49–51]. These data-driven methods can capture novel, context-specific regulatory relationships that may not yet be represented in existing knowledge bases. However, annotation-based strategies like those implemented in GeDi remain the most widely adopted due to their interpretability, efficiency, and compatibility with the majority of enrichment workflows [2]. Moreover, the use of high-quality, curated resources such as GO, KEGG, and Reactome ensures biologically grounded results even in scenarios where raw expression data are unavailable.

Looking forward, we aim to extend GeDi's applicability to GSEA-style enrichment results, further broadening its scope. Expanding supported sources of PPI and functional interaction data is another key direction, with the goal of offering richer network-based summaries.

Given the increasingly widespread adoption of tools based on large language models and single cell foundation models for the interpretation of biological data, we anticipate that GeDi can provide better suited summaries that would in turn enhance the answers returned by such agents [52, 53].

In summary, GeDi serves as a scalable, user-friendly solution for the systematic summarization of enrichment results, efficiently bridging the gap between statistical detection and meaningful biological insight.

Conclusion

In our work we presented GeDi, a new R package developed to streamline and simplify the analysis and interpretation of functional enrichment results. GeDi provides several different distance metrics, which can be used to quantify the similarity of the gene sets in the input functional enrichment results, as well as a variety of clustering algorithms to aggregate the gene sets to groups based on their distance scores.

GeDi can reduce the extensive lists of gene sets arising from enrichment analyses to compact sets of functional clusters, fostering via different distance measurements and clustering approaches the careful inspection of various aspects of the data at hand. We expect that this approach, well rooted in the Bioconductor ecosystem of packages yet simple to adopt for many omics analysis workflows, will empower many scientists to generate viewpoints for clearer and better hypothesis generation and interpretation.

Availability and requirements

Project name: GeDi.

Project home page: <https://bioconductor.org/packages/GeDi> (release), <https://github.com/AnnekathrinSilvia/GeDi> (development version).

Archived version: <https://doi.org/10.5281/zenodo.16315835>, package source as gzipped tar archive of the version reported in this article.

Project documentation: rendered at <https://annekathrinsilvia.github.io/GeDi/>.

Operating systems: Linux, MacOS, Windows. Programming language: R. Other requirements: R-4.4.0 or higher, Bioconductor 3.19 or higher. License: MIT. Any restrictions to use by non-academics: none.

Abbreviations

ARI	Adjusted Rand Index
FCS	Functional Class Scoring
GEO	Gene Expression Omnibus
GO	Gene Ontology
GSCA	Gene Set Coexpression Analysis
GSEA	Gene Set Enrichment Analysis
GSNCA	Gene Set Network Correlation Analysis
HGNC	HUGO (HUMAN Genome Organisation) Gene Nomenclature Committee
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MSigDb	Molecular Signatures Database
ORA	Over-Representation Analysis
PPI	Protein-Protein Interaction
RNA-seq	RNA sequencing
Treg	Regulatory T cells

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-025-06335-6>.

Supplementary Material 1.

Acknowledgements

This work has been supported by the computing infrastructure provided by the Core Facility Bioinformatics at the University Medical Center Mainz, used also for deploying the demo instance. The authors thank the Bioconductor community and the attendees of the EuroBioC2024 for valuable feedback and suggestions.

Author contributions

ASN—conceptualization, data curation, formal analysis, methodology, project administration, resources, software, visualization, writing—original draft, writing—review and editing. AU—data curation, software, writing—review and editing. NA—data curation, software, writing—review and editing. LD—software, writing—review and editing. AS—data curation, software, writing—review and editing. SSH—data curation, interpretation, writing—review and editing. MD—data curation, interpretation, writing—review and editing. KS—conceptualization, funding acquisition, resources, supervision, writing—original draft, writing—review and editing. FM—conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, resources, software, supervision, visualization, writing—original draft, writing—review and editing. All authors read and approved the final version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. The work of FM, NA and MD is supported by the German Research Foundation (DFG) under project number 318346496-SFB 1292/2 TP19N. The funding body had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The datasets used in this manuscript and its supplements are available from the following articles: The dataset on the regulatory T cell subpopulations is included in PubMed ID: 31924477 (<https://doi.org/10.1016/j.immuni.2019.12.002>). Dataset deposited at the Gene Expression Omnibus (GEO) (GSE130884, project id: PRJNA541901) and accessed from the https://github.com/AnnekathrinSilvia/GeDi_supplement repository. The GeDi package can be downloaded from its Bioconductor page <https://bioconductor.org/packages/GeDi/> or the GitHub development page <https://github.com/AnnekathrinSilvia/GeDi>. The repository available at https://github.com/AnnekathrinSilvia/GeDi_supplement/ contains the code used to generate the supplemental material and the required input data to replicate the analyses presented in the use cases.

Declarations

Competing interests

The authors declare no competing interests.

Received: 23 July 2025 / Accepted: 18 November 2025

Published online: 07 December 2025

References

1. Cano-Gamez E, Trynka G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases. *Front Genet.* 2020;11:424.
2. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform.* 2021;22(1):545–56.
3. Garcia-Moreno A, López-Domínguez R, Villatoro-García JA, Ramirez-Mena A, Aparicio-Puerta E, Hackenberg M, et al. Functional enrichment analysis of regulatory elements. *Biomedicines.* 2022;10(3):590.
4. 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):15545–50.
5. Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinform.* 2021;22(1):191.
6. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.* 2012;8(2):e1002375.
7. Ma J, Shojai A, Michailidis G. A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinform.* 2019;20(1):546.
8. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics.* 2006;22(13):1600–7.
9. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Res.* 2023;51(W1):W207–12.
10. 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):W90–97.
11. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. ClusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (Camb (Mass)).* 2021;2(3):100141.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25–9.
13. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2018;46(D1):D649–55.
14. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* 2023;51(D1):D587–92.
15. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* 2015;1(6):417–25.
16. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37(1):1–13.
17. 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.
18. Ewing E, Planell-Picola N, Jagodic M, Gomez-Cabrero D. GeneSetCluster: a tool for summarizing and integrating gene-set analysis results. *BMC Bioinform.* 2020;21(1):443.
19. Yoon S, Kim J, Kim S-K, Baik B, Chi S-M, Kim S-Y, et al. GScluster: network-weighted gene-set clustering analysis. *BMC Genomics.* 2019;20(1):352.
20. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods.* 2015;12(2):115–21.
21. Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B. Shiny: Web Application Framework for R. 2024. <https://shiny.posit.co/>.
22. Amezquita RA, Lun ATL, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. *Nat Methods.* 2020;17(2):137–45.
23. den Van Berge K, Hembach KM, Sonesson C, Tiberi S, Clement L, Love MI, et al. RNA sequencing data: hitchhiker's guide to expression analysis. *Annu Rev Biomed Data Sci.* 2019;2(1):139–73.
24. Ludt A, Ustjanzew A, Binder H, Strauch K, Marini F. Interactive and reproducible workflows for exploring and modeling RNA-Seq data with *pcaExplorer*, *Ideal*, and *GeneTonic*. *Curr Protoc.* 2022;2(4):e411.
25. Marini F, Ludt A, Linke J, Strauch K. GeneTonic: an R/Bioconductor package for streamlining the interpretation of RNA-Seq data. *BMC Bioinform.* 2021;22(1):610.
26. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, et al. Genenames.Org: the HGNC resources in 2023. *Nucleic Acids Res.* 2023;51(D1):D1003–9.
27. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Jorge Alvarez-Jarreta M, et al. Ensembl 2020. *Nucleic Acids Res.* 2020;48(D1):D682–8.
28. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D766–73.
29. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinform (Oxf, Engl).* 2011;27(12):1739–40.
30. David G. bs4Dash: A 'Bootstrap 4' Version of 'Shinydashboard'. 2024. <https://rinterface.github.io/bs4Dash/index.html>.
31. Ganz C. Rintrojs: a wrapper for the Introjs library. *J Open Sour Softw.* 2016;1(6):63.

32. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* 2019;47(D1):D330–8.
33. Jaccard P. The distribution of the flora in the alpine zone¹. *New Phytol.* 1912;11(2):37–50.
34. Levandowsky M, Winter D. Distance between sets. *Nature.* 1971;234(5323):34–5.
35. Sorensen T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish Commons. *Biol Skr/K Danske Vidensk Selsk.* 1948;5:1–34.
36. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46.
37. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. Gosemsim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics.* 2010;26(7):976–8.
38. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech-Theory Exp.* 2008;2008(10):P10008.
39. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002;30(7):1575–84.
40. van Dongen S. A Cluster Algorithm for Graphs. CWI. 2000.
41. Kaufman L, Rousseeuw P. Clustering by means of medoids. *Statistical data analysis based on the L1 norm and related methods.* 1987.
42. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4(1):44–57.
43. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, et al. The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 2007;8(9):R183.
44. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc.* 1971;66(336):846–50.
45. Zheng C, Wang M, Yamada R, Okada D. Delving into gene-set multiplex networks facilitated by a K-nearest neighbor-based measure of similarity. *Comput Struct Biotechnol J.* 2023;21:4988–5002.
46. Delacher M, Imbusch CD, Hotz-Wagenblatt A, Mallm J-P, Bauer K, Simon M, et al. Precursors for nonlymphoid-tissue Treg cells reside in secondary lymphoid organs and are programmed by the transcription factor BATF. *Immunity.* 2020;52(2):295–312.e11.
47. Ito M, Komai K, Mise-Omata S, Iizuka-Koga M, Noguchi Y, Kondo T, et al. Brain regulatory T cells suppress astrogliosis and potentiate neurological recovery. *Nature.* 2019;565(7738):246–50.
48. 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.
49. Culhane AC, Perrière G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinform.* 2003;4:59.
50. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential co-expression test for gene sets. *Bioinform (Oxf, Engl).* 2014;30(3):360–8.
51. Tsai C-A, Chen JJ. Gene set correlation analysis and visualization using gene expression data. *Curr Bioinform.* 2021;16(3):406–21.
52. Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods.* 2024;21(8):1470–80.
53. Li L, Dong L, Zhang H, Xu D, Li Y. SpaLLM: enhancing spatial domain analysis in multi-omics data through large language model integration. *Brief Bioinform.* 2025. <https://doi.org/10.1093/bib/bbaf304>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.