PLOS COMPUTATIONAL BIOLOGY

SOFTWARE

GeneCOCOA: Detecting context-specific functions of individual genes using co-expression data

Simonida Zehr[1,2], **Sebastian Wolf**[3], **Thomas Oellerich**[3], **Matthias S. Leisegang**[1,2], **Ralf P. Brandes**[1,2], **Marcel H. Schulz**[2,4], **Timothy Warwick**[1,2*]

1 Goethe University Frankfurt, Institute for Cardiovascular Physiology, Frankfurt am Main, Germany, **2** German Centre for Cardiovascular Research (DZHK), Partner site Rhine-Main, Frankfurt am Main, Germany, **3** Goethe University Frankfurt, University Hospital, Department of Medicine II, Haematology/Oncology, Frankfurt am Main, Germany, **4** Goethe University Frankfurt, Institute for Computational Genomic Medicine, Frankfurt am Main, Germany

* warwick@vrc.uni-frankfurt.de

Citation: Zehr S, Wolf S, Oellerich T, Leisegang MS, Brandes RP, Schulz MH. et al. (2025) GeneCOCOA: Detecting context-specific functions of individual genes using co-expression data. PLoS Comput Biol 21(3): e1012278. https://doi.org/10.1371/journal.pcbi.1012278

Editor: Wei Li, Children's National Hospital, George Washington University, UNITED STATES OF AMERICA

Received: June 26, 2024

Accepted: February 3, 2025

Published: March 31, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1012278

Data availability statement: There are no primary data in the paper; all software and code are available at

Abstract

Extraction of meaningful biological insight from gene expression profiling often focuses on the identification of statistically enriched terms or pathways. These methods typically use gene sets as input data, and subsequently return overrepresented terms along with associated statistics describing their enrichment. This approach does not cater to analyses focused on a single gene-of-interest, particularly when the gene lacks prior functional characterization. To address this, we formulated *GeneCOCOA*, a method which utilizes context-specific gene co-expression and curated functional gene sets, but focuses on a user-supplied gene-of-interest (GOI). The co-expression between the GOI and subsets of genes from functional groups (e.g. pathways, GO terms) is derived using linear regression, and resulting root-mean-square error values are compared against background values obtained from randomly selected genes. The resulting *p* values provide a statistical ranking of functional gene sets from any collection, along with their associated terms, based on their co-expression with the gene of interest in a manner specific to the context and experiment. *GeneCOCOA* thereby provides biological insight into both gene function, and putative regulatory mechanisms by which the expression of the GOI is controlled. Despite its relative simplicity, *GeneCOCOA* outperforms similar methods in the accurate recall of known gene-disease associations. We furthermore include a differential *GeneCOCOA* mode, thus presenting the first implementation of a gene-focused approach to experiment-specific gene set enrichment analysis. *GeneCOCOA* is formulated as an R package for ease-of-use, available at https://github.com/si-ze/geneCOCOA.

Introduction

Advances in sequencing technology have decreased the costs and increased the accuracy of transcriptome profiling [1]. This has resulted in an abundance of datasets generated from a

https://github.com/size/geneCOCOA. NCBI GEO accessions for the publicly available data analyzed here are as follows: GSE36980, GSE28253, GSE9006, GSE48060, GSE17048, GSE114922. RNA-sequencing data from the acute myeloid leukemia patients analyzed here is available from the European Genome-Phenome Archive under the accession EGAD0000100848. Single-cell RNA-sequencing data analyzed here is available under the NCBI GEO accession GSE143148.

Funding: This work was supported by the Goethe University Frankfurt am Main, the German Centre for Cardiovascular Research (DZHK, 81X2200165 to R.P.B.; 81Z0200101 to M.H.S.), the Deutsche Forschungsgemeinschaft (DFG) excellence cluster EXS2026 (Cardio-Pulmonary Institute, project number 390649896, to R.P.B. and M.H.S.), the DFG Project-ID 403584255 - TRR 267 (TP A04 to M.S.L.; TP A06 to R.P.B.; TP Z03 to M.H.S.), the Dr. Rolf Schwiete Stiftung (to R.P.B.) and the Hessian Center for Artificial Intelligence (hessian.AI, to M.H.S.). All authors are employees of Goethe University Frankfurt am Main. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

wide variety of experimental conditions, many of which are made publicly available [2–4]. As such, interrogation of public sequencing data has become an increasingly important step in research focused on a specific gene or gene product of interest. Normally, this is limited to detecting whether the gene-of-interest is expressed in a given dataset or whether the expression of the gene changes in a particular experimental condition [5]. However, this approach does not supply insight into any potential functions of the gene-of-interest in the data, or any regulatory mechanisms which might govern expression of the gene.

Functional enrichment analyses carried out in the course of differential gene expression analysis usually relies upon the input of one or more gene sets which are derived throughout the course of the analysis (e.g. differentially expressed genes) [6]. Curated associations between each gene and sets of annotations such as ontologies [7], pathways [8,9] and diseases [10] are then computed. These associations are subsequently statistically analyzed for over-represented terms, considering the size of the input gene set, the number of genes associated with the given term, and enrichment in hits compared to an appropriate background gene set [11–15]. The outcome of these analyses is a list of terms stratified by statistical values such as p value, adjusted p value, precision and recall. Results from these approaches have the potential to inform future research directions and wet-lab experiments. However, they cannot provide insight into the functional relevance of individual genes, especially when genes lack prior functional characterization.

One approach that can be used to examine potential function of an individual gene-of-interest (GOI) is to model the expression of the GOI against the expression of other genes present in a given dataset, in a co-expression analysis [16]. Co-expression pertains to identification of genes which display common patterns of regulation, and may therefore be subject to similar gene regulatory mechanisms (e.g. transcription factors). Methods for co-expression analysis range from simple models of linear regression between expression values of genes [17], to construction of weighted co-expression networks consisting of gene modules [18] and deep learning-based approaches [19]. Assigning functional and biological significance to an individual gene based on co-expression requires further analysis, however, the dissection and stratification of results of co-expression analyses can be challenging [20]. This means that potentially interesting insight into functions of individual genes may be lost during transitions between methods.

Methods aiming to determine the functions of individual genes are available, and implement different approaches. Some have the objective to identify genes or genetic variation relevant to certain tissues, cell types, or cell lines (e.g. *CONTENT* [21], and *ContNeXt* [22]) (Table 1). While these methods are useful for the identification of significant gene-context associations, they do not predict the biological function of the given gene. Other methods (Tables 2 and 3) use network properties (e.g. *NetDecoder* [23]) or apply coessentiality analyses (*FIREWORKS* [24]) to characterize gene-gene associations in a given context. These tools help to identify other genes significantly associated with a GOI in a context-specific manner, but again do not link these results with biological meaning. *GeneWalk* [25], *DAVID* [14,26] and *Correlation AnalyzeR* [27] (Table 4) are three tools which come closest to determining the function of individual genes, in that they aim to provide context-specific biological meaning whilst being able to focus on individual genes.

GeneWalk [25] takes a user-provided input list of genes and assembles a network composed of these genes and associated Gene Ontology (GO) terms. Network representation learning with random walks is then performed on the network. Statistical association between a given gene and GO terms is determined through comparison of node similarities between the true network and a null distribution based on node similarities in randomized networks.

Table 1. Approaches to co-expression analysis (not supporting individual gene perspective)

Method	Input data	Description
WGCNA[18]	Expression matrix	Identifies modules of highly correlated genes, identifies most relevant genes of a module, relates modules to one another and to external traits such as GO-terms.
CEMiTool[49]	Expression matrix	Identifies modules of highly correlated genes, identifies most relevant genes of a module, integrates external data (e.g. interactome, pathways).
FIREWORKS[24]	Gene list	Ranks top correlations and anti-correlations in an undirected, unweighted network and returns gene-gene associations. No knowledge distillation.
CONTENT[21]	Expression matrix + SNPs	Computes associations between SNPs and tissues by decomposing expression data across samples into context-shared and context-specific components. No knowledge distillation.
GeneFriends[51]	Gene list	Uses the gene list as a seed for the construction of co-expression network to find highly correlated genes in pre-computed expression data of a selected tissue. Thereby allows for functional annotation of a single-gene.
COXPRESdb[52]	Gene list	Queries a precomputed database to identify highly coexpressed genes, genes with the same GO annotation and genes which are co-expressed with the GOIs in a selected tissue of a selected organism.
diffcoexp[53]	Expression data of two conditions	Compares two expression data sets against each other and identifies gene pairs with significantly different correlation coefficients under the two conditions. No knowledge distillation.
HGCA[54]	GOI	Identifies top co-expressed genes for the provided GOI (precomputed on representative tissue samples), performs various built-in gene term enrichment analyses on the co-expression module.

https://doi.org/10.1371/journal.pcbi.1012278.t001

Alternatively, associations between individual genes and biological functions can be performed using *DAVID* [14,26], which takes a list of genes as input and returns GO terms, protein domain information and curated pathways which are statistically enriched in their association with a given gene, computed using Fisher’s exact test. While these approaches do provide insight into putative functions of individual genes, neither method considers the expression of the provided genes or other genes relevant to the GO terms in question. Not considering expression as a feature in these analyses could result in missing dynamic relationships between the gene-of-interest and the genes, or subsets of genes, associated with the given term. Additionally, the implementation of *GeneWalk* is limited to the use of GO terms, and cannot be implemented with other curated gene sets which may provide more relevant functional annotations in a specific context, such as disease.

Table 2. Approaches to knowledge distillation (not supporting individual gene perspective)

Method	Input data	Description
Myers et al. (2008)[55]	Various types of raw data	Trains a support vector machine classifier on the raw data and a list of GO terms using annotated genes as positive examples to predict gene function.
ClusterProfiler[56]	Differential expression analysis results	Returns relevant terms (e.g. GO terms, KEGG, ...) associated with enriched gene sets.
PANTHER.db	Gene list	Returns relevant terms (e.g. GO terms, pathways, ...) associated with enriched gene sets.
ReactomePA	Gene list	Supports hypergeometric tests and gene set enrichment analyses, returns enriched REACTOME pathways.
NOA[57]	Gene list	Infers link ontology for given gene set using associated GO terms, performs enrichment analysis on resulting network.

https://doi.org/10.1371/journal.pcbi.1012278.t002

Table 3. Gene function prediction agnostic to user-provided context.

Method	Input data	Description
GeneMANIA [58]	GOI	Builds association networks from different publicly available data types (co-expression, co-regulation, co-localisation, shared protein domains, ...). Not customisable to individual experiment.
ContNeXt [22]	Gene list	Computes gene-tissue associations across three different contexts (i.e., tissues, cell types, and cell lines). No knowledge distillation. Expression data precomputed.
GIANT [59]	GOI	Integrates thousands of datasets to predict interactions of the provided GOI and provide associated GO terms.
NewGOA [60]	GOI	Combines publicly available data on protein interactions and GO annotations in a graph and uses a random walk to predict function. Not customisable to individual experiment.
BiRWLGO [61]	GOI	Combines lncRNA-lncRNA similarity, lncRNA-protein interaction and protein-protein interaction data into hybrid graph, applies bi-random walk to predict lncRNA function. Not customisable to individual experiment.
NMFGO [62]	GOI	Builds gene-term association matrix, uses a semantic similarity approach to predict gene function.

https://doi.org/10.1371/journal.pcbi.1012278.t003

One method which considers co-expression and outputs putative gene function is *Correlation AnalyzeR* [27]. Here, weighted Pearson correlations between normalized gene expression counts are calculated between a gene-of- interest and other genes present in the

Table 4. Gene function prediction based on a user-provided context.

Method	Input data	Description
DAVID [26]	Gene list	Summarises genes based on shared categorical data from public resources, runs modified Fisher's Exact Test for gene-enrichment analysis. Individual gene-GO associations retrievable.
GeneWalk [25]	Gene list	Assembles context-specific network from provided gene list, associates GO terms using public resources, applies an unsupervised network representation learning to retrieve most relevant GO terms. Individual gene-GO associations retrievable.
NetDecoder [23]	GOI + Phenotype data of two traits	Computes differential gene-gene associations and network characteristics (e.g. genes with high flow differences between trait 1 and trait 2). No knowledge distillation.
Correlation AnalyzeR [27]	GOI + Expression matrix	Takes custom expression data or fetches public data sets. Uses genome-wide Pearson correlations as a ranking metric for GSEA algorithm, returns gene sets correlated with a gene of interest.

https://doi.org/10.1371/journal.pcbi.1012278.t004

expression data. A ranked gene list is then assembled from the resulting correlation values, which is used as input to gene set enrichment analysis, resulting in statistically enriched terms which are theoretically co-expressed with the gene-of-interest. However, the authors state that for a robust analysis, datasets of more than 30 samples and at least 4 different studies should be used, limiting the contexts in which this method can be used.

We sought to explore how co-expression and functional enrichment analyses can be combined into a single workflow which provides insight into the function of a specific GOI in a given context provided by the input data. Such a method would permit a comprehensive assessment of expression patterns and putative functions of a GOI across multiple experimental conditions using experimental data generated by the user. To this end, we propose *GeneCOCOA*, an *R* package which identifies and ranks functional gene sets which are co-expressed with a user-supplied GOI. *GeneCOCOA* may be run using either user-supplied or publicly available gene expression data from bulk or single-cell experiments, and can utilize several curated databases of gene annotations in order to compute functional enrichments in co-expression.

Design and implementation

Databases

For the functionality of *GeneCOCOA* described herein, curated gene sets from the Hallmark database [28], as well as genes annotated to the Biological Process domain of Gene Ontology [29] (GO:BP) were used.

Input data

The use cases described in this manuscript utilized publicly available transcriptome profiling data available from *Gene Expression Omnibus* [2,3] under the accession numbers GSE36980

[30], GSE28253 [31], GSE5406 [32], GSE9006 [33], GSE48060 [34], GSE17048 [35], and GSE114922 [36].

The RNA-sequencing data arising from acute myeloid leukemia patients [37] is available publicly from the *European Genome-Phenome Archive* [38] under the accession EGAD00001008484, and initial access prior to the publication of the data was provided by Prof. Dr. Thomas Oellerich and Dr. Sebastian Wolf (Goethe University Frankfurt, University Hospital Frankfurt).

Single-cell RNA-sequencing data from [39] was used for the implementation of *GeneCOCOA* at single-cell resolution. The data were pre-processed to a normalised matrix of $cell * gene$ counts using *Seurat* [40], which were used as input data to *GeneCOCOA*. The mouse MSigDB hallmark gene sets were also provided as input.

Preprocessing

Raw reads were aligned against the *hg38* genome using *Bowtie2* (v2.3.5.1) [41], with default parameters, and quantified using *Salmon* (v1.5.2) [42], with default parameters. Curated quantified and normalized expression data sets were fetched with *gemma.R* [43].

Detection of gene sets which are co-expressed with a gene-of-interest

Determining number of gene subsets. The number of gene subsets sampled from each gene set i is implemented as a user-controlled parameter. In test runs, we determined $i = 1000$ to provide an acceptable compromise between efficiency and statistical power (see S1 Fig.). Therefore, we set $i = 1000$ for all analyses in this manuscript.

Generation of gene subsets. Initially, a number of subsets (default 1000) are derived from a given gene set (e.g. pathway, GO term), as described by the following:

$$G_i \subset g_1, g_2, ..., g_N, where\ |G_i| = n\ for\ i = 1, 2, ..., 1000 \tag{1}$$

where G_i is the i-th subset of n genes $g_1, g_2, ..., g_N$ which make up the total gene set G.

Linear regression models. The dataset-specific expression values of each gene in a subset of genes serve as predictor variables in a linear regression model with the expression of a GOI being the outcome variable, as described by:

$$y = \beta_0 + \beta_1 g_1 + \beta_2 g_2 + ... + \beta_n g_n + \varepsilon \tag{2}$$

where $g_1, g_2, ..., g_n$ represent the dataset-specific expression values of the genes of the subset, $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the coefficients for each predictor variable, y represents the predicted expression of the GOI, and ε represents the error of the linear regression model.

Root-mean-square error calculation. For each gene subset, the linear regression model produces predicted values $\hat{y}_i$ based on the predictors g_i. The root-mean-square error (RMSE) for the i-th subset is then calculated as:

$$RMSE_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_{ij})^2} \tag{3}$$

where y_j is the true expression of the GOI and $\hat{y}_{ij}$ is the predicted expression from the linear regression model for the j-th observation using the subset G_i.

The same procedure is performed for a size-matched set of randomly sampled genes, resulting in two sets of RMSE values. One derived from linear regression models predicting the expression of the GOI from subsets of genes from a given gene set, and one derived from linear regression models predicting the expression of the GOI from randomly sampled subsets of genes expressed in the given dataset.

Computation of gene set-specific enrichment P values. RMSE values $RMSE_i, ..., RMSE_{1000}$ derived from subsets $G_i, ..., G_{1000}$ of a given gene set G are compared against RMSE values $_rRMSE_i, ..., _rRMSE_{1000}$ derived from randomly subset genes using a Student's t-test. The outcome is a p value describing the probability of the GOI being associated with G in the experimental condition C, denoted as $p(G, C)$. The $p(G, C)$ values are subsequently adjusted for multiple testing using the Benjamini-Hochberg method [44]. This results in an adjusted $p(G, C)$ value for each gene set in a given curated database, describing the strength of association between the genes comprising each gene set and the user-provided GOI.

Differential *GeneCOCOA* score. To estimate how the association between a GOI and gene set G changes between two conditions C_1 and C_2, a differential mode of *GeneCOCOA* was implemented. The differential score (DS) is computed as the negative logarithm of the ratio of the gene set-specific p values between two conditions:

$$DS(G, C_1, C_2) = -log_{10}\left(\frac{p(G, C_1)}{p(G, C_2)}\right) \tag{4}$$

A $DS > 0$ indicates that the GOI and gene set G are more strongly associated in C_1, while a negative DS values indicates a stronger association in C_2. To determine the statistical significance of a derived DS, we adapted the approach described in [45]. To this end, we assumed that p-values produced by *GeneCOCOA* for a GOI in the two conditions, $p(G, C_1)$ and $p(G, C_2)$, are independent of one another and are uniformly distributed between $[0, 1]$. For variables meeting these conditions, the corresponding negative decadic logarithm of the p-value ratios $(-log_{10}(p(G, C_1)/p(G, C_2)))$ will be Laplace-distributed with a mean of 0 and a standard deviation of 1 ($L(0, 1)$). In order to test this assumption, we first randomised the expression data in each condition by reshuffling the expression values per column. *GeneCOCOA* was run on the randomised datasets, resulting in $p(G, C_1)$ and $p(G, C_2)$ for each gene set. We then computed DS per gene set as described in Eq 4. A Laplace (L) distribution was fitted to the data, and Kolmogorov-Smirnov test showed no statistical difference ($p = 0.08$) between a $L(0, 1)$ and $L(data)$. We can thus assume that the distribution of P ratios resulting from a *GeneCOCOA* run with the default bootstrapping with 1000 resamples follows a Laplace-distribution. Based on this observation, $L(0, 1)$ can be used to derive significance values for the change in association between G and the GOI between the conditions:

$$P_{DS}(G, C_1, C_2) = p(DS(G, C_1, C_2)) \tag{5}$$

The probability density distribution of $L(0, 1)$ is thus applied to assign statistical significance to each observed $DS(G, C_1, C_2)$.

Comparison to similar methods

GeneCOCOA was compared to other methods which aim to annotate the functions of individual genes by testing the ability of each tool to accurately link genes associated with a given disease to GO terms implicated in the same disease. The methods considered for comparison were DAVID [14], GeneWalk [25] and Correlation AnalyzeR[27].

Definition of disease-relevant genes. In order to define a relevant gene set for each condition to be studied, the DisGeNET [46] platform was queried via web interface (https://www.disgenet.org/search) with the full name of each condition ("Amyotrophic Lateral Sclerosis", "Alzheimer's Disease", "Dilated Cardiomyopathy", "Insulin-dependent Diabetes Mellitus", "Myocardial Infarction" and "Multiple Sclerosis"). As of 14-06-2023, the top-ranked hits were the entries with the UMLS/concept IDs C0002736, C0002395, C0007193, C0011854, C0027051 and C0026769. From each summary of gene-disease associations (GDA), genes with a $Score_{gda} \geq 0.5$ were considered as substantially associated with the disease and included in the input set of disease-relevant genes.

Definition of disease-relevant terms/gene sets. To obtain disease-relevant gene sets, the MalaCards database [47] was queried via web interface (https://www.malacards.org/) with the full name of the condition ("Alzheimer's Disease", "Amyotrophic Lateral Sclerosis", "Dilated Cardiomyopathy", "Insulin-dependent Diabetes Mellitus", "Myocardial Infarction" and "Multiple Sclerosis") on 14-06-2023. The top hit was selected based on the MalaCards InFormaTion Score and the Solr relevance score provided by MalaCards. For each disease card (MalaCards IDs ALZ065, AMY091, DLT002, TYP008, MYC007 and MLT020, respectively), the complete list of Gene Ontology Biological Process terms was downloaded and treated as the ground truth collection T for the respective disease.

Construction of input gene lists for GeneWalk and DAVID. To assemble a context-specific gene network, GeneWalk requires a list of relevant genes obtained from a specific experimental assay as an input. To this end, GEO2R [3] was used to obtain a list of differentially expressed (DE) genes for each of the publicly available transcriptomic gene sets. Any gene with an adjusted $p < 0.05$ between control and disease condition, as calculated by *DESeq2* [48], was considered differentially expressed. To ensure that all disease-relevant genes obtained via DisGeNET would be included as well, the union of disease-relevant genes and DE genes was obtained. Thus, a context-set C was created for each condition.

Systematic comparison of *GeneCOCOA*, *Correlation AnalyzeR*, *GeneWalk* and *DAVID*. Each method was used to determine the association of disease-relevant genes (as per defined via DisGeNET, see subsection *Definition of disease-relevant genes*) with disease-relevant gene sets (as defined via MalaCards, see previous subsection). Since GeneWalk results are computed on Gene Ontology annotations [7], we restricted the comparison to gene sets from the GO:BP collection.

The ability of each tool to report any disease-relevant GO:BP term for a list of disease-relevant genes-of-interest across different diseases was tested. We distinguished two cases: (1) A disease-relevant gene is analyzed in a condition matching its disease. In this case, we expect the method to report a significant association between the gene and any of the disease-relevant GO:BP terms reported in MalaCards (”true positive”). (2) A disease-relevant gene-of-interest is analyzed using data arising from a separate disease where said gene is not annotated as being important in DisGeNet, therefore a significant association between the gene and the terms present in the MalaCard for the disease is not expected (”false positive”).

GeneCOCOA, *GeneWalk*, *DAVID* and *Correlation AnalyzeR* [27] were run for every combination of disease-relevant genes – Alzheimer's Disease (AD): 24, Amyotrophic Lateral Sclerosis (ALS): 16, Dilated Cardiomyopathy (DC): 12, Diabetes Mellitus (DM): 4, Myocardial Infarction (MI): 21, Multiple Sclerosis (MS): 7 – and diseases. In each case, a disease-specific expression data set was provided as input, a single disease-relevant gene was provided as the gene-of-interest, and the GO:BP ontology provided as the collection of gene sets to rank. For each disease, *GeneWalk* and *DAVID* were run with the appropriate context-set C (see previous subsection) as the input list (including the

additional genes-of-interest which are not functionally linked to the disease in question, see case (2) above). Gene-of-interest-associated GO:BP terms were parsed from the results of each method using a threshold of
$p_{adjusted} < 0.05$.

Time and memory profiling. Time and memory profiling was performed on a machine with an AMD Ryzen 9 5950X processor (16 cores, 3.40 GHz) and 128 GB of RAM. *GeneCOCOA* was run with R version 4.4.2 (2024-10-31) on the Windows subsystem for Linux Ubuntu-20.04 on a 64-bit Windows 10 (build 19045). Time and peak memory consumption were tracked running *GeneCOCOA* on a subsample (n=5) as well as the complete familial hypercholesterolemia (FH) patient data set (n=10), and on a subset (n=50) as well as the complete acute myeloid leukemia (AML) patient data set (n=135). We ran *GeneCOCOA* on the respective gene of interest (*LDLH* for FH, *FLT3* for AML) on both the Hallmark gene set collection (50 sets) as well as the gene set collection defined by the GO Biological Processes (7608 sets).

GeneCOCOA R package

GeneCOCOA is formulated as an R package and is hosted on GitHub at the URL https://github.com/si-ze/geneCOCOA.

Results

GeneCOCOA identifies functional gene sets co-expressed with a gene-of-interest

The COmparative CO-expression Analysis focused on a Gene-of-interest (*GeneCOCOA*) presented here incorporates multiple approaches which aim to functionally annotate genes following gene expression profiling (Fig 1A). Several approaches exist for the analysis of experiment-specific co-expression patterns (e.g. *WGCNA* [18], *CemiTool* [49]), the harnessing of curated knowledge (e.g. Molecular Signature Database (*MSigDB*) [28]), as well as for the integration of prior knowledge with experiment-specific co-expression patterns (e.g. *GSEA* [50], *Enrichr* [12]). Some methods also aim to apply prior knowledge to predict the functions of individual genes, most notably *DAVID* [26], *GeneWalk* [25] and *Correlation AnalyzeR* [27]. However, few methods exist which utilize co-expression and curated gene sets to predict gene function (summarized in Table 1). To our knowledge, only *Correlation AnalyzeR* [27] provides this option in *single-gene mode*. Yet, its results are based on a single correlation analysis. *GeneCOCOA* has been developed as an integrative method which aims to apply curated knowledge to experiment-specific expression data in a gene-centric manner based on a robust bootstrapping approach.

The input to *GeneCOCOA* is a list of curated gene sets (e.g. from Gene Ontology, MSigDB, pathways), a gene-of-interest (GOI) that the user wishes to interrogate, and a normalized gene expression matrix of (Fig 1B, top). The input expression matrix can either originate from bulk expression data ($sample * gene$) or from single-cell expression data ($cell * gene$). The expression matrix may also be composed of samples from a single condition or multiple conditions, depending on the user's question. If the user is unsure of the consistency with which the GOI may be regulated between conditions, then we would recommend the use of one condition at a time, or the differential implementation of *GeneCOCOA*. If the user believes that the GOI would be consistently co-regulated across conditions, then they can maximize the sample number, and therefore statistical power, and provide an expression matrix encompassing multiple experimental conditions at once. In general, our experimentation with input data

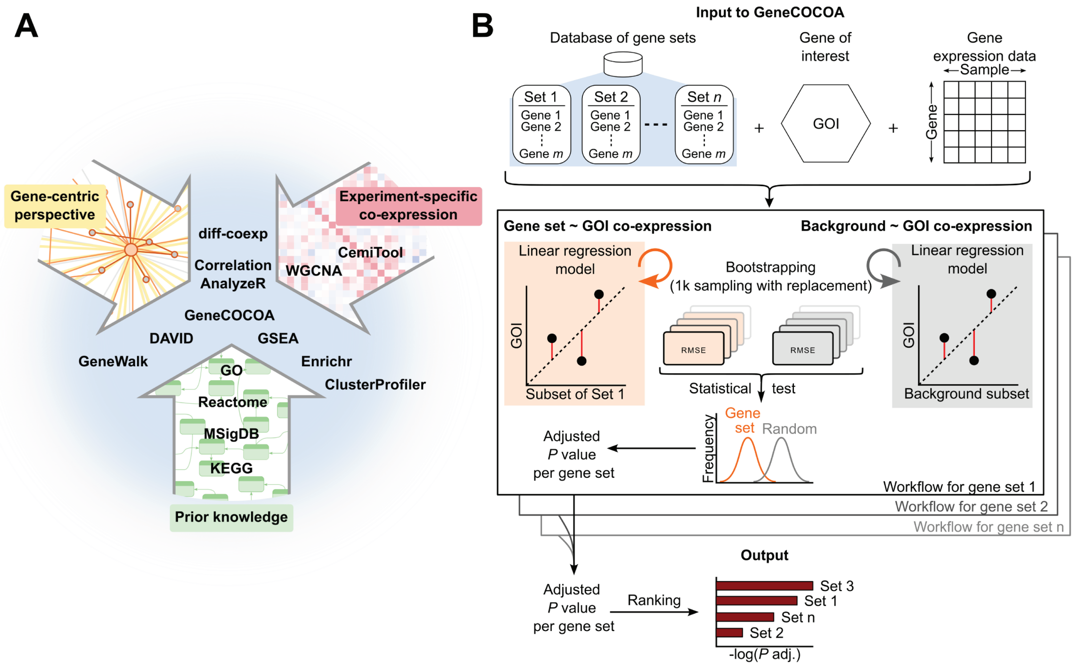

Fig 1. *GeneCOCOA* workflow for identification of functional gene sets co-expressed with a gene-of-interest. (A) Strategies and related methods for statistically associating genes to putative functions, summarized into gene-centric (*GeneWalk*, *DAVID*), prior knowledge (*GO*, *Reactome*, *MSigDB*) and co-expression (*WGCNA*, *CemiTool*) approaches. *GeneCOCOA* incorporates elements of each of these approaches into a single workflow. **(B)** Schematic representation of the *GeneCOCOA* workflow, which takes as input user-provided functional gene sets, a gene-of-interest (GOI) and gene expression data to report statistically ranked gene sets associated with the provided GOI. This is achieved by comparing root-mean-square error (RMSE) values from bootstrapped linear regression models predicting the expression of the GOI using either genes arising from a single gene set, or randomly sampled genes from the expression data. Gene set errors and random errors are statistically compared, and the resulting p values are adjusted, resulting in an output list of functional gene sets ranked statistically by the strength of their association with the provided gene-of-interest.

https://doi.org/10.1371/journal.pcbi.1012278.g001

has shown that a minimum sample number of 5 should be used with *GeneCOCOA* in order to obtain good performance.

From each gene set, n genes are sampled and used as predictor variables in a linear regression modelling the expression of the GOI as the outcome variable (Fig 1B, middle). A background model is created analogously by sampling n random genes from the complete expression data set. For bootstrapping, this procedure is repeated i times, i being a parameter that can be specified by the user. Testing different values of i, we found $i = 1000$ to provide the best tradeoff between efficiency and power (see S1 Fig.). The i gene set model errors and i random model errors are compared in a t-test. Gene sets with $p_{adjusted} < 0.05$ are considered to model the expression of the GOI better than random, and the $p_{adjusted}$ values are used to stratify and rank gene sets (Fig 1B, bottom). The results output by *GeneCOCOA* aim to provide insight into potential functions of the gene-of-interest in the specific context provided by the gene expression data.

Detection of context-specific changes in gene function using *GeneCOCOA*

To test the ability of *GeneCOCOA* to detect changes in gene function resulting from disease, it was applied to identify functions of the gene FMS-like tyrosine kinase 3 (*FLT3*) in acute myeloid leukemia (AML). AML is a malignancy of the hematopoietic system affecting the

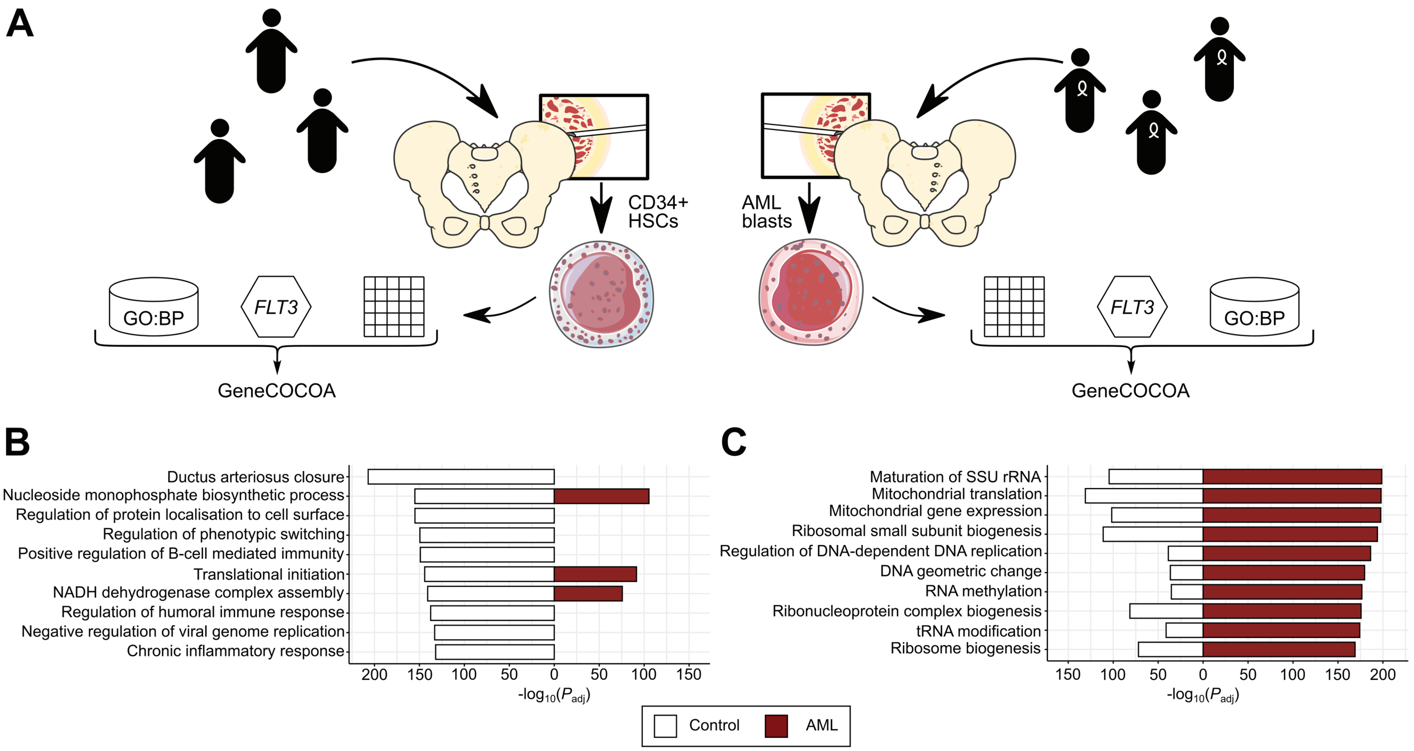

Fig 2. Example use case of *GeneCOCOA* to predict context-specific *FLT3* function using expression data from hematopoietic stem cells and acute myeloid leukemia blasts. (A) In an exemplary use case, *GeneCOCOA* was applied to study the co-expression patterns of *FLT3* with Gene Ontology Biological Process (GO:BP) terms in bulk RNA-sequencing of CD34+ hematopoietic stem cells (HSCs) from 48 healthy subjects, and blasts from 31 patients with acute myeloid leukemia (AML) positive for *FLT3*-ITD mutations. Illustrations of the pelvis and cells were adapted from vector files hosted at bioicons.com under a CC BY 4.0 license. **(B)** The 10 highest ranked GO:BP terms with *FLT3* in HSCs from healthy donors, as computed by *GeneCOCOA*. The corresponding significance values in AML blasts are provided for comparison. Ranks are annotated next to the bars; non-significant terms are not annotated. **(C)** The 10 highest-ranked GO:BP terms with *FLT3* in patients with AML and *FLT3*-ITD mutations, as computed by *GeneCOCOA*. The corresponding significance values in healthy HSCs are provided for comparison. Ranks are annotated next to the bars; non-significant terms are not annotated.

https://doi.org/10.1371/journal.pcbi.1012278.g002

differentiation and maturation of myeloid blood cells. Characterized by a complex genetic landscape, AML can be divided into various subtypes, which differ in both phenotype and prognosis. One common (25% of patients [63]) mutation linked to AML is the internal tandem duplication (ITD) of *FLT3*. Normally, expression and activation levels of *FLT3* are important for maintaining a balance of proliferation and differentiation in hematopoietic cells [64]. *FLT3*-ITD results in a constitutive activation of the kinase, promoting a hyperproliferative state and cell survival [65]. *FLT3*-ITD is associated with a higher disease burden, higher relapse rate and inferior overall survival [14].

A whole-transcriptome RNA-sequencing dataset of 136 AML patients [37] was subset for patients with *FLT3*-ITD mutations (31 patients). Taking *FLT3* as the GOI, *GeneCOCOA* was used to assess the significance of the association between *FLT3* and gene sets defined by GO Biological Processes (GO:BP). For comparison, a control set of 48 healthy *CD34+* bone marrow samples was constructed from data under the GEO accession GSE114922 [36]. Again, *GeneCOCOA* was used to detect and rank associations between *FLT3* and GO:BP terms (Fig 2A).

Physiologically, *FLT3* is involved in immune function and regulation of hematopoietic cell proliferation and differentiation [64]. Accordingly, among the GO:BP terms associated with *FLT3* by *GeneCOCOA* in healthy *CD34+* cells are terms associated with immune response (e.g. ”Regulation of humoral immune response”, ”Chronic inflammatory response”) and terms indicating both proliferative processes (e.g. ”Nucleoside monophosphate biosynthetic process”) and differentiation (e.g. ”Positive regulation of B-cell mediated immunity”, ”Regulation of phenotypic switching”) (Fig 2B). This complex profile is lost in the *GeneCOCOA* results for *FLT3* co-expression patterns in AML blasts (Fig 2C). The top 10 GO:BP terms reflect mitochondrial processes (e.g. ”Mitochondrial gene expression”) and cell growth/division (e.g. ”Regulation of DNA-dependent DNA replication”, ”Ribosome biogenesis”), reflecting the switch to a predominantly proliferative profile. The results thus replicate dysregulation of *FLT3* expression and function previously described in literature, indicating that *GeneCOCOA* may be able to detect context-dependent changes in gene function, given appropriate data.

Implementation of *GeneCOCOA* on single-cell gene expression data

To test the feasibility of *GeneCOCOA* to detect gene-function relationships from single-cell resolution data, we utilised single-cell RNA-sequencing (scRNA-seq) data generated from mouse heart tissue [39]. Following pre-processing and normalization of the data (Fig 3A), *GeneCOCOA* was used to detect gene-function relationships of *Ldlr* and *Tgfb1*, both of which were widely expressed in the data (Fig 3B and 3C). Associations between these genes and murine Hallmark gene sets reflected prior findings on the functions of these genes. *Ldlr* was most strongly associated with mTORC signaling (Fig 3D), an association previously noted in literature [66,67]. Similarly, *Tgfb2* was functionally associated to epithelial mesenchymal transition (Fig 3E), reflecting the well-studied role of TGFβ signaling in cell state transitions [68,69].

Following the successful implementation of *GeneCOCOA* on bulk and single-cell gene expression datasets, we next sought to implement a method of detecting differential gene-function associations between two conditions.

GeneCOCOA detects disease-driven alterations in gene co-expression patterns

To this end, *GeneCOCOA* was applied to gene expression datasets arising from diseases with well-studied causative genes. The co-expression patterns between respective causative GOIs and 50 MSigDB Hallmark gene sets [28] were compared between healthy controls and disease data sets using *GeneCOCOA* in differential mode (Fig 4A, see Eq 4 in *Materials and methods*). The differential mode is used to compare two conditions, C_1 and C_2. Here, for each gene set, *GeneCOCOA* compares the significance of association between G and the GOI in the two given conditions by computing the ratio of significance values $DS(G, C_1, C_2) = -log_{10} \frac{p(G,C_1)}{p(G,C_2)}$. A $DS > 0$ indicates that the GOI and G are more strongly associated in C_1, $DS < 0$ hints to stronger associations in C_2. Furthermore, a p-value $p(DS)$ can be inferred for each given DS.

One disease in which causative genes have been suggested in literature is amyotrophic lateral sclerosis (ALS). The first gene to be identified as causative for this neurodegenerative disease was superoxide dismutase 1 (*SOD1*) [70]. *SOD1* codes for Cu/Zn superoxide dismutase type-1, an enzyme crucial for cellular antioxidant defense mechanisms. Mutations of *SOD1* in ALS are known to destabilize the protein, leading to misfolding. This triggers various pathophysiological events such as protein accumulation, mitochondrial and/or proteasome dysfunction and accumulation of reactive oxygen species (ROS). This switch between contexts is reflected in the differential *GeneCOCOA* results for *SOD1* (Fig 4B), comparing disease (11 patients with ALS) and healthy (lymphocytes of 11 healthy donors) conditions. Here,

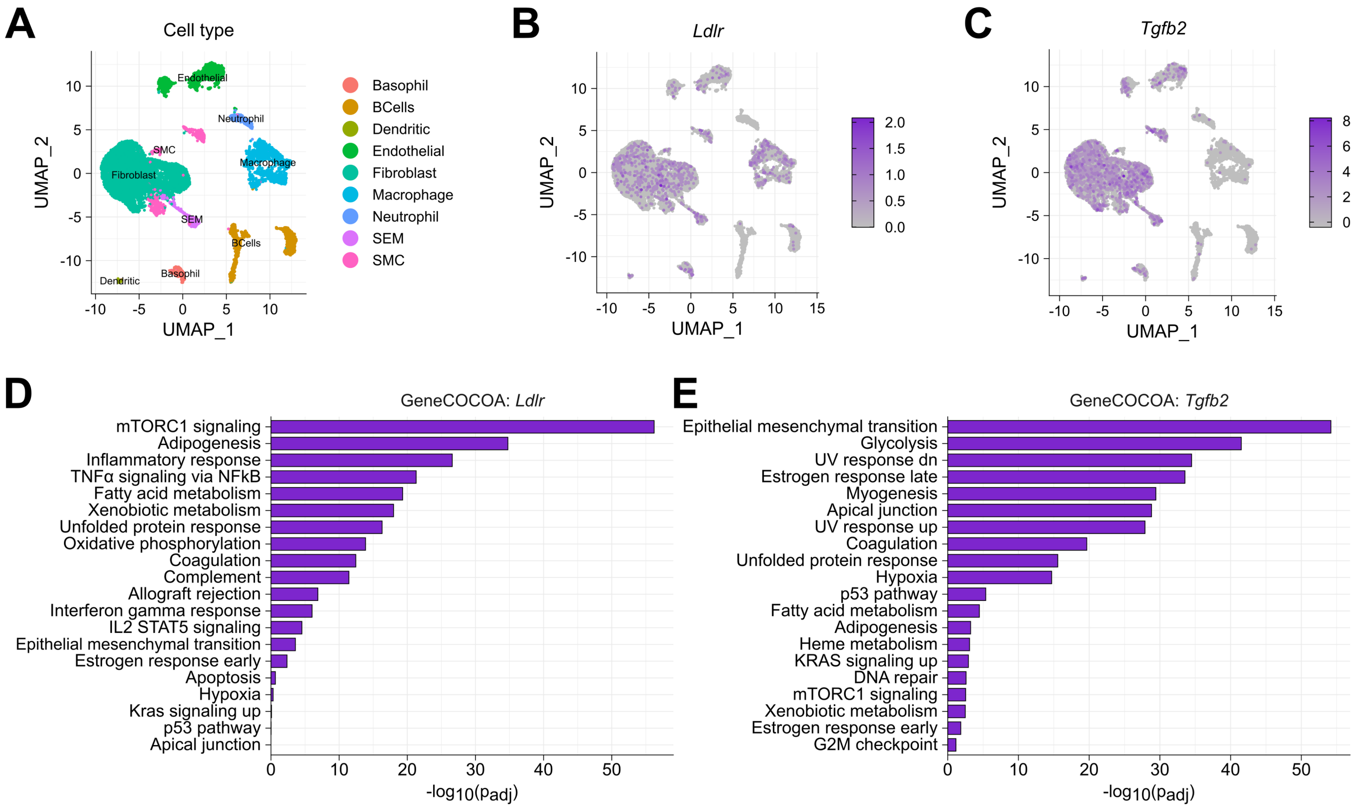

Fig 3. *GeneCOCOA* recovers functionally relevant terms from single-cell sequencing data. (A) Single cell sequencing data of endothelial cells after myocardial infarction [39] was analyzed with *GeneCOCOA*, taking **(B)** *Ldlr*, which is involved in lipid metabolism, and **(C)** *Tgfb2*, an inducer of epithelial- and endothelial-to-mesenchymal transition, as exemplary genes-of-interest. **(D)** *Ldlr* shows strong associations with Adipogenesis and mTORC1 signalling. **(E)** *Tgfb2* was linked to Epithelial-to-mesenchymal transition.

https://doi.org/10.1371/journal.pcbi.1012278.g003

SOD1 shows a stronger tendency to associate with immune function (e.g. ”Allograft rejection”, ”TNF-α signalling via NF-κB”, ”Inflammatory response”) in the healthy condition. In accordance with literature [71,72], *SOD1* is not linked with these gene sets in in the diseased transcriptomes, Instead, the GOI shows the strongest link with *Oxidative phosphorylation*, reflecting potential mitochondrial defects. Also indicative of the pathophysiology of *SOD1*-driven ALS was the association between *SOD1* expression and the Hallmark gene set ”Unfolded protein response”. The detection of this term – specifically in the ALS samples – demonstrates that *GeneCOCOA* has the potential to identify context-specific co-expression patterns with disease relevance.

In another use case, *GeneCOCOA* was run using gene expression data originating from isolated lymphocytes of 10 patients with familial hypercholesterolemia (FH), comparing them to 13 healthy control samples. FH is an autosomal dominant disorder of lipoprotein metabolism characterized by high levels of cholesterol. The most common causes are mutations in the gene coding for low-density lipoprotein receptor (*LDLR*). Physiologically, the LDL transmembrane receptor mediates the internalization and lysosomal degradation of LDL. Mutations disrupting the function of *LDLR* lead to elevated plasma levels of LDL, promoting accelerated atherosclerosis and coronary heart disease [73,74]. In correspondence with these

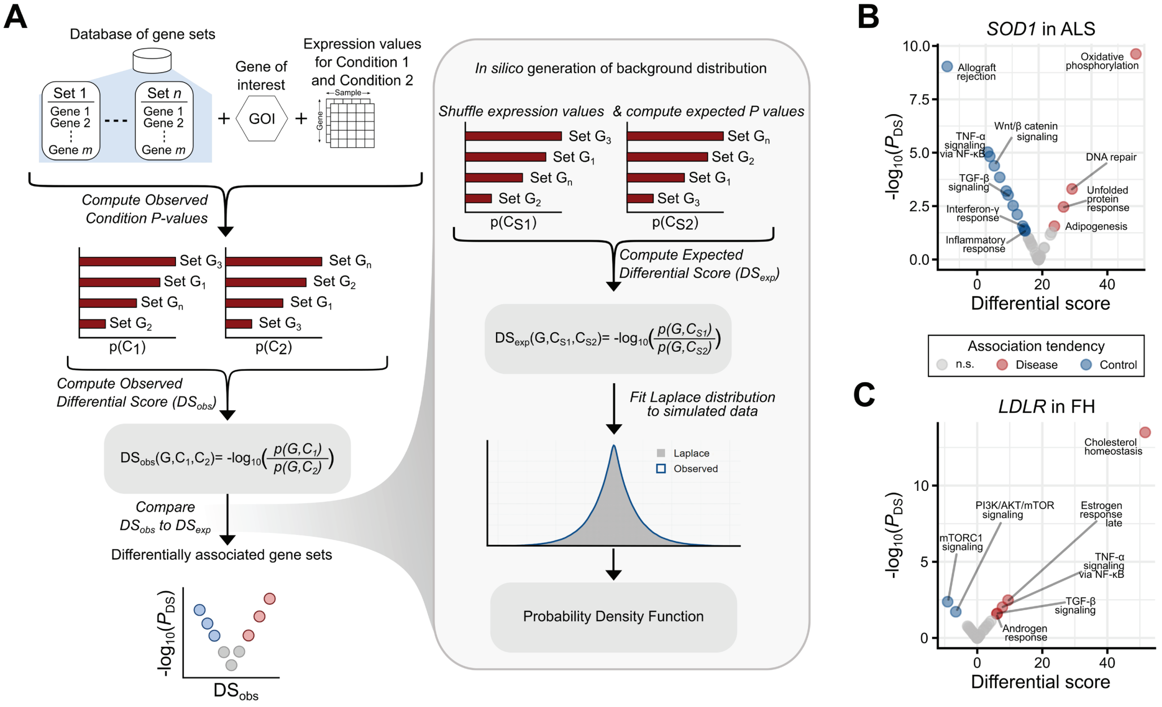

Fig 4. Differential *GeneCOCOA* detects gene-gene set associations enriched in disease. (A) A schematic overview of how the differential mode integrates two individual *GeneCOCOA* results (referred to as sets of the respective Condition *P*-values) into a volcano plot to illustrate gene-gene set associations which are enriched in one of the two conditions. The x-values in the volcano plot indicate the direction of change in association and are computed as the ratio of the Condition *P*-values. The corresponding significance in change (Differential *P*-value) is derived from a Laplace distribution fitted to the data and plotted as the y-values. Applied to diseases with monogenic signatures, *GeneCOCOA* helps detect relevant responses of a gene-of-interest in disease such as **(B)** a gain in association between *SOD1* and ”Oxidative phosphorylation” and ”DNA repair” in lymphocytes associated from patients with amyotrophic lateral sclerosis vs. healthy donors, and **(C)** a gain in association between *LDLR* and ”Cholesterol homeostasis” in monocytes isolated from patients with familial hypercholesterolemia vs. healthy donors.

https://doi.org/10.1371/journal.pcbi.1012278.g004

mechanisms described in literature, the *GeneCOCOA* results (Fig 4C) indicated that the functional association between *LDLR* and genes annotated to be important for ”Cholesterol homeostasis” became stronger in FH samples compared to control samples. Again, these results suggest that *GeneCOCOA* is able to detect changes in gene co-expression which are pertinent to disease-specific conditions.

While these results were promising, the question remained of how the approach implemented in *GeneCOCOA* compared to methods with the similar aim of functionally annotating individual genes.

GeneCOCOA provides a comprehensive gene-focused co-expression and functional analysis missing from similar methods

To our knowledge, only few approaches to the problem of inferring the function of a specific gene-of-interest (GOI) been published (Table 1), most notably *DAVID* [14], *GeneWalk* [25] and *Correlation AnalyzeR* [27].

DAVID is a web-accessible set of functional annotation tools which allows for the rapid mining of a wide range of public resources. Provided with a list of gene identifiers, *DAVID* summarizes them, based on shared categorical data in gene ontology, protein domain, and biochemical pathway membership, returning a modified Fisher Exact p-value for gene-enrichment analysis.

GeneWalk allows for the GO enrichment analysis of an experiment-specific gene set (e.g. differentially expressed genes). Using publicly available resources, *GeneWalk* first assembles a context-specific gene network which represents both interactions between the provided genes and links to GO terms, then applies an unsupervised network representation learning algorithm (*DeepWalk* [75]) to retrieve the GO terms of highest statistical relevance.

Correlation AnalyzeR [27] has been developed for the exploration of co-expression correlations in a given data set, and in *single-gene mode* also supports the prediction of individual gene functions and gene-gene relationships. In an adaption of the Gene Set Enrichment Analysis[50] (GSEA) algorithm, it employs genome-wide Pearson correlations as a ranking metric to determine the gene sets correlated with a GOI.

GeneCOCOA and *Correlation AnalyzeR* [27] exploit the user-provided expression data to gain insight into gene correlations in a context-specific manner. GeneWalk and, less explicitly, *DAVID*, require a list of input genes to assemble the context. Using *gemma.R* [43] and *GEO2R* [3] for the selection of potential input data sets, we therefore focused on sufficiently large ($n > 10$) transcriptomic data sets in which we could reliably identify a set of DE genes. Six curated data sets met our criteria. Disease-relevant GO:BP terms were then retrieved from *MalaCards* [47], and disease-relevant genes from *DisGeNET* [10,46].

In a systematic comparison, *DAVID*, *GeneWalk*, *Correlation AnalyzeR* [27] and *GeneCOCOA* were used to search for statistically significant associations between matching disease-relevant genes and disease-relevant GO:BP terms (Fig 5A). Each method was run for every combination of disease (AD: Alzheimer's Disease, ALS: Amyotrophic Lateral Sclerosis, DC: Dilated Cardiomyopathy, DM: Insulin-dependent Diabetes Mellitus, MI: Myocardial Infarction and MS: Multiple Sclerosis) and disease-relevant genes (total genes AD: 24, ALS: 16, DC: 12, DM: 4, MI: 21, MS: 7). For each method, a statistically significant ($p_{adjusted} < 0.05$) association between a given gene and a condition-relevant term was recorded. If the gene belonged to the matching disease-relevant gene set, this was considered a true positive, whereas if the gene was a member of one of the other disease sets, it was considered a false positive. Although these terms are not strictly accurate given the nature of these types of analysis, they are used here in an attempt to compare these methods in an objective and unbiased manner, and this matter is further covered in the *Discussion*.

Across all conditions, *GeneCOCOA* had a substantially higher true positive rate than either *DAVID* or *GeneWalk*, and in all but one case also a higher true positive rate than *Correlation Analyzer* (Fig 5B). In order to confirm that *GeneCOCOA* was not just returning spurious significant associations for every provided gene, the proportions of false positives across all conditions for all methods was further analyzed. Overall, *GeneCOCOA* reported more false positives than the other methods (Fig 5C, S2 and S3 Figs.). However, when considering the results in a gene-set-focused perspective, *GeneCOCOA* recalls more true positives per gene set than false positives (corresponds to the summary of row counts in Fig 5C; see also S2, S3 and S4 Figs. This is truly independent of the disease expression set provided. From a condition-wise perspective (corresponding to columns in Fig 5C), *GeneCOCOA* consistently reports a higher proportion of true positives than false positives across all conditions (S3 Fig.). For *GeneWalk* and *DAVID*, the proportions of true and false positives were negligible, resulting in both methods having high true negative rates, but accompanying high false negative rates as well. *Correlation Analyzer* managed to recover more true positives than the prior two methods, yet in the majority of cases the false positive rate was at least as high as the true positive rate (see S3 Fig.). Thus, *GeneCOCOA* recovers the most relevant disease terms whilst maintaining an acceptable level of specificity, independent of disease type.

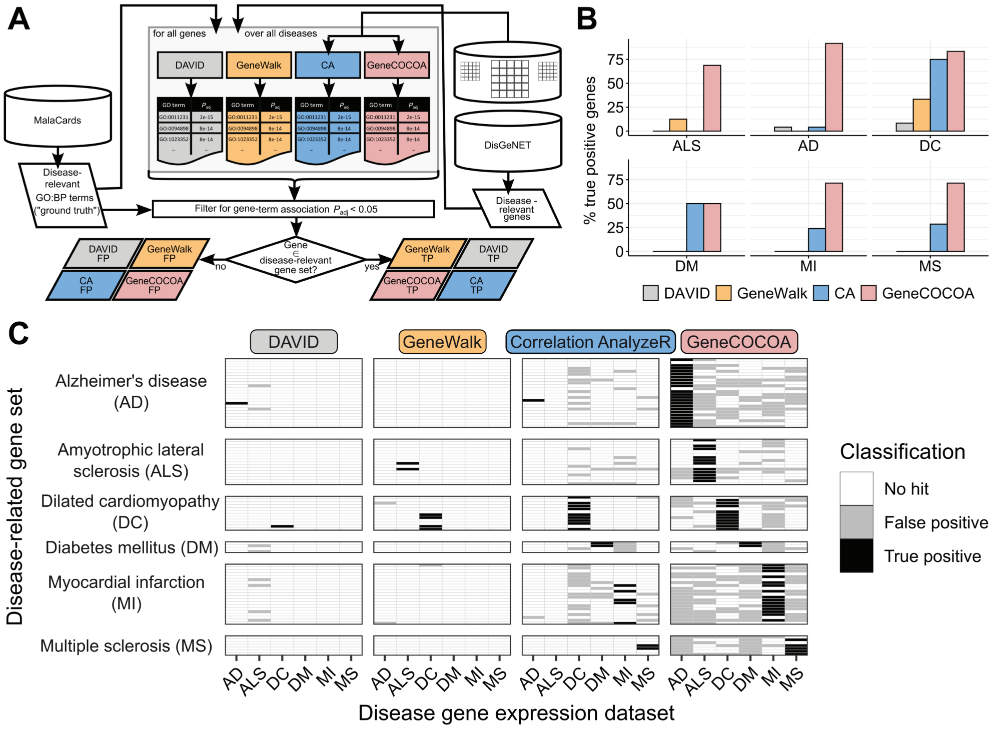

Fig 5. Systematic comparison of *GeneCOCOA*, *DAVID*, *Correlation AnalyzeR* and *GeneWalk* for their performance in statistically linking disease-relevant genes and GO:BP terms. (A) *GeneCOCOA*, *DAVID*, *Correlation AnalyzeR* (CA) and *GeneWalk* were each run to identify significantly associated disease-relevant genes from DisGeNet and disease-associated Gene Ontology Biological Process terms (GO:BP) as listed on MalaCards. Genes significantly associated to the matching disease terms were considered true positives (TP), and genes statistically linked to terms from other diseases as false positives (FP). **(B)** Proportion of true positive associations between disease-relevant genes and matching disease GO:BP terms by *GeneCOCOA*, *GeneWalk*, *Correlation AnalyzeR* and *DAVID* (AD: Alzheimer's disease, ALS: Amyotrophic lateral sclerosis, DC: Dilated cardiomyopathy, DM: Diabetes mellitus, MI: Myocardial infarction, MS: Multiple sclerosis). **(C)** Summary of true positive and false positive gene-term associations per set of disease-relevant genes across all diseases, as computed by *GeneCOCOA*, *GeneWalk*, *Correlation AnalyzeR* and *DAVID*.

https://doi.org/10.1371/journal.pcbi.1012278.g005

In the course of the benchmarking, we observed runtimes of 0.87 - 1.76 seconds per term and a peak memory consumption of 2.67 GB. In practice, the runtime and memory consumption depends on the size of the input data set and the number of terms the user is querying (see S5 Fig.).

Taken together, the results presented here demonstrate that *GeneCOCOA* is capable of identifying statistically significant functional co-expression patterns linked to a gene-of-interest. Dynamics in context also seem to be detectable, as well as gene-specific functions. *GeneCOCOA* offers a different approach to other methods, which appears to identify more biologically relevant gene functions than similar tools, although benchmarking these kinds of approaches remains highly challenging.

Discussion

This manuscript describes *GeneCOCOA*, a method designed to implement both co-expression and functional enrichment analyses focused on a gene-of-interest (GOI). Evidence of the functionality of *GeneCOCOA* was demonstrated by using bulk and single-cell transcriptome profiling data, resulting in the identification of co-expressed gene sets with a relevant gene in each scenario. The use of *GeneCOCOA* to detect context-specific alterations in gene function was illustrated using RNA-sequencing data arising from a large cohort of patients suffering from acute myeloid leukemia. Here, functional gene sets associated with disease progression and prognosis could be found to be significantly co-expressed with *FLT3*, a known driver of the disease. The performance of *GeneCOCOA* relative to similar methods was compared across several distinct contexts, and showed that *GeneCOCOA* has the potential to fill a previously underpopulated niche in the toolkit of gene expression data analysis.

Advancements in next-generation sequencing technology have resulted in an abundance of high quality, publicly available transcriptome profiling data from a wide range of species, conditions and stimuli [3]. This has shifted the experimental bottleneck from data generation towards data analysis, with a resulting requirement for robust, efficient methods to extract maximal insight from these data. This must be accomplished whilst simultaneously maintaining ease-of-use for the user, many of whom are not expert computational biologists. Another by-product of this wealth of data is that researchers with specific genes-of-interest can query these data for metrics such as co-expression. However, manually curating co-expression results to derive biological insight can be complex and time-consuming.

Herein, we demonstrated that *GeneCOCOA* is capable of providing the user with functional gene sets which are enriched in their co-expression with a GOI. The functionality of *GeneCOCOA* in conjunction with data from large cohort experiments was demonstrated with a large data set consisting of 79 RNA-sequencing samples [36,37], where the known functional role of *FLT3* could be recapitulated. In this illustrative example, the link between the gene-of-interest and experimental condition is extremely well established. This makes it difficult to truly assess the sensitivity of *GeneCOCOA* for discovering *de novo* functional roles of a GOI in a given condition.

In further illustrative use cases, *GeneCOCOA* was implemented on genes implicated as being causative for amyotrophic lateral sclerosis and familial hypercholesterolemia, specifically the GOIs *SOD1* [76] and *LDLR* [73]. In each case, *GeneCOCOA* identified functional, co-expressed enriched terms pertinent to the given disease. It should be noted, however, that in each case there were several replicates per condition (11 vs. 11, and 13 vs. 10, respectively). These replicate numbers are relatively uncommon in experimental setups designed around cell culture systems, where three biological replicates per biological condition is common [77]. The identification of robust enrichments when *GeneCOCOA* is provided with datasets of this smaller size is more challenging than when using larger datasets, and certainly represents a potential drawback of the approach. We found that a sample number of 5 was required for good *GeneCOCOA* performance, and in the case of fewer samples per condition, we recommend the combination of sample data into a single matrix prior to running *GeneCOCOA*. Happily, transcriptome profiling of larger patient cohorts is becoming increasingly common and accessible [78–80], providing ideal input for *GeneCOCOA* and similar tools. Similarly, the continuing rise of single-cell resolution data, where each cell can be considered as a sample, also provides a wealth of data which should result in high *GeneCOCOA* performance, providing the GOI is sufficiently covered in the data.

Another caveat to consider in the course of analysis of transcriptomic data with *GeneCOCOA* or any similar method, is the disconnect between expression and true function. Whilst *GeneCOCOA* is capable of using an array of curated gene annotation databases to infer potential functionality, a vast number of genes remain uncharacterized with regard to functional importance [81]. These genes are therefore excluded from the analysis, despite potentially interesting co-expression with the gene in question. Similarly, in a native co-expression analysis without any functional subsetting of genes, genes co-expressed with one another may in fact have diverse functions. For example, genes whose products make up negative feedback loops may be similarly regulated in order to provide a controlled response to a stimulus, despite having antagonistic functions [82].

In a systematic comparison of *GeneCOCOA* against similar methods (*GeneWalk*, *DAVID* and *Correlation AnalyzeR*), *GeneCOCOA* was able to identify a greater proportion of evidence-linked disease-relevant gene-GO term relationships. By computing these links across a number of diseases, it could be shown that disease-relevant associations reported by *GeneCOCOA* tended to be enriched in specificity for the diseases in question. However, it should be stated that making concrete conclusions on the relative performance of these types of methods is highly challenging, given the difficulties in ascribing true positive and true negative validation sets. This arises from the curated nature of gene sets, which rely wholly on published gene functions, as well as the extent and quality of databases used to record and document relationships between genes and functions. A consequence of this approach is that there may be genes not yet linked to a function or disease, which may just be unstudied in that capacity rather than irrelevant. For example, inflammatory genes such as *TNF* and *TGFB1* (both annotated as being important to myocardial infarction) are not included in the list of genes associated with Alzheimer's disease on *DisGeNET*. As a consequence, significant associations reported for these genes (S4 Fig.) with Alzheimer's-relevant terms were marked as quasi-false positives. Yet, dysregulations related to these genes have been linked to the development of Alzheimer's disease in prior research [83–86]. Similarly, *GeneCOCOA* also reported false positive associations in the amyotrophic lateral sclerosis (ALS) data set for the genes *BCL2* and *BAX*. While they are present in the Alzheimer's disease gene set, these apoptotic genes have also been described as mediators of motor neuron loss in ALS [87–89]. Thus, the supposedly false positive associations returned by *GeneCOCOA* might, in several cases, hint at biologically meaningful GOI-disease associations which are not reflected in our strict approach to the definitions of ground truth.

From a methodological perspective, it was interesting that the relatively simple methods employed by *Correlation AnalyzeR* and *GeneCOCOA* both outperformed the more complex method implemented in *GeneWalk*. *Correlation Analyzer*'s approach of considering entire gene sets in their enrichment analysis could result in a decreased sensitivity compared to *GeneCOCOA*, which samples subsets of gene sets. This would explain the greater sensitivity (but additionally increased false positive rate) of *GeneCOCOA*. The authors of *Correlation AnalyzeR* recommend input data with many samples in order for a robust analysis, whereas the iterative sampling approach of *GeneCOCOA* might permit increased performance on smaller datasets. It should further be mentioned that the default implementation of two predictors per regression model reduces the potential impact of multicollinearity on the *GeneCOCOA* results. What the performance of these two similar methods shows, is that using co-expression in combination with functional enrichment is a valid approach for inferring gene function, particularly of previously unstudied genes. Which specific method of co-expression analysis and functional enrichment should be used likely depends on the type and extent of the input data.

The formulation of *GeneCOCOA* to provide a functionally-resolved co-expression analysis framework is designed to minimize both data and time loss when moving data between different methods. Performance is largely determined by the iterative computation of background gene sets, the number of which may be set by the user. We aimed to maximize ease-of-use by formulating *GeneCOCOA* as an R [90] package, thereby making it simple to integrate the analysis with common workflows such as differential gene expression analysis [48,91]. In the future, we can imagine that the scope of *GeneCOCOA* could be expanded to explore the functional roles of other genomic elements such as enhancers, but this would require a robust collection of enhancer-gene links across which function could be inferred.

In summary, *GeneCOCOA* provides a method by which users can infer putative functions of a gene-of-interest based on co-expression of the given gene with curated sets of functionally-annotated genes. *GeneCOCOA* therefore empowers users to take advantage of the growing number of publicly available transcriptome profiling datasets, in order to provide greater functional insight and generate new hypotheses pertaining to the roles of individual genes in different contexts.

Availability and future directions

At the time of writing, *GeneCOCOA* is available via the GitHub repository https://github.com/si-ze/geneCOCOA, from where it can be installed as an R package. In the future, we hope to make the package available via Bioconductor. We are committed to maintaining the performance of *GeneCOCOA* for the forseeable future, and are open to developing the tool further in the face of new data types and gene set annotations.

Conclusion

- *GeneCOCOA* is a combined method for the identification of functional gene sets which are significantly co-expressed with a gene-of-interest.
- The method can be used in a highly flexible manner on user-supplied or publicly available transcriptome profiling data at bulk or single-cell resolution.
- Function gene sets can be provided by the user, or taken from curated, publicly available databases which hold information on ontologies, pathways and diseases.
- *GeneCOCOA* successfully recapitulates functional signatures of genes implicated in monogenic diseases.
- *GeneCOCOA* detects greater numbers of evidence-linked gene-disease relationships than similar methods.

Supporting information

S1 Fig. Identification of recommended number of bootstraps. With different values for number of bootstrapping rounds were tested, $i = 1000$ was found to provide the best trade off between efficiency and power. Displayed here are exemplary results for the association between *FLT3* and the 50 MSigDB hallmark gene sets in the expression data set of 136 AML patients. We inspected the results of 16 *GeneCOCOA* runs with bootstrap rounds ranging from 2 to 100,000. All terms which were identified as significant P_{adj} in any of the runs are listed as rows, while columns indicate the different *GeneCOCOA* runs. White tiles indicate that this term was not identified as significant in the respective *GeneCOCOA* run, while red indicates that it was returned as one of the terms significantly associated with *FLT3* expression.
(TIF)

S2 Fig. Comparison of true positives and false positives hits across gene sets. For each gene set, we evaluated the number of hits by method, differentiating true positives (TP hits in the original disease context) from false positives (FP hits in other disease contexts) **(A)** Across gene sets, the number of hits returned by *GeneCOCOA* in the TP condition is either higher or comparable to any other number of hits in FP contexts. *DAVID* and *GeneWalk* recover a smaller number of hits in general. While *GeneWalk* – except for the case of MI – manages to

retain a good TP:FP ratio **(C)** *DAVID* **(B)** and *Correlation AnalyzeR* **(D)** report more FP than TP hits in a third of the cases.
(TIF)

S3 Fig. Comparison of true positives and false positives across conditions. For each condition, the set of genes which are disease relevant as per DisGeNET can be defined as the true data set, all other genes are defined as other. **(A)** Comparing the proportions of true genes with disease relevant term hits against the proportion of other genes with disease relevant term hits, *GeneCOCOA* consistently manages to recover more true hits than other hits across all conditions. **(B)** *DAVID* and **(C)** *GeneWalk* show only a negligible proportion of other hits. Yet these methods also fail to recover a substantial amount of true hits. **(D)** In two cases, *Correlation AnalyzeR* shows slightly more true than other hits. Yet, in all other cases there are at least as many other as true hits. The overall percentage of true hits recovered is smaller than in the *GeneCOCOA* runs.
(TIF)

S4 Fig. False/true positive matrices for all three methods with gene symbols. Summary of true positive and false positive gene term associations per set of disease relevant genes across all diseases, as computed by *DAVID*, *GeneWalk*, *Correlation AnalyzeR* and *GeneCOCOA*.
(TIF)

S5 Fig. Runtime and memory profiling of *GeneCOCOA*. Time and peak memory consumption were assessed using expression datasets of varying sizes: *GeneCOCOA* was run on LDLR in a subsampled (n = 5, FH_ss) and complete (n = 10, FH) version of the familial hypercholesterolemia patient expression set, as well as on FLT3 in a subsampled (n = 50, AML_ss) and complete (n = 135, AML) acute myeloid leukemia data set. Profiling was conducted on a small gene set collection (MSigDB, 50 sets, "Hallmark", coloured in blue) and a large one (Gene Ontology Biological Process, 7608 sets, "GO:BP", coloured in red). **(A)** Total runtime is a function of both dataset size and size of gene set collection. **(B)** The MSigDB Hallmark collection features larger gene sets, which combinatorially allow for a greater number of unique subsets of predictor genes than small data sets featured in GO:BP, slowing down the random subsampling per term. **(C)** Peak memory consumption is significantly higher for the larg GO:BP collection compared to Hallmark, and increases with the dataset size for both collections.
(TIF)

Author contributions

Conceptualization: Marcel H. Schulz, Timothy Warwick.

Data curation: Simonida Zehr, Timothy Warwick.

Formal analysis: Simonida Zehr, Timothy Warwick.

Investigation: Simonida Zehr, Timothy Warwick.

Methodology: Simonida Zehr, Sebastian Wolf, Timothy Warwick.

Resources: Sebastian Wolf, Thomas Oellerich, Matthias S. Leisegang, Ralf P. Brandes.

Software: Simonida Zehr, Marcel H. Schulz, Timothy Warwick.

Supervision: Ralf P. Brandes, Marcel H. Schulz, Timothy Warwick.

Validation: Simonida Zehr, Timothy Warwick.

Visualization: Simonida Zehr, Timothy Warwick.

Writing – original draft: Simonida Zehr, Marcel H. Schulz, Timothy Warwick.

Writing – review & editing: Simonida Zehr, Matthias S. Leisegang, Ralf P. Brandes, Marcel H. Schulz, Timothy Warwick.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51. https://doi.org/10.1038/nrg.2016.49 PMID: 27184599
2. Edgar R, Domrachev M, Lash A. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. 2013;41(Database issue):D991-5. https://doi.org/10.1093/nar/gks1193 PMID: 23193258
4. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;49(D1):D10–7. https://doi.org/10.1093/nar/gkaa892 PMID: 33095870
5. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013;14(2):89–99. https://doi.org/10.1038/nrg3394 PMID: 23269463
6. 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. https://doi.org/10.1093/bib/bbz158 PMID: 32026945
7. Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–8. https://doi.org/10.1093/nar/gky1055 PMID: 30395331
8. 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. https://doi.org/10.1093/nar/gkac963 PMID: 36300620
9. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D):D687–92.
10. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 2020;48(D1):D845–55. https://doi.org/10.1093/nar/gkz1021 PMID: 31680165
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). 2021;2(3):100141. https://doi.org/10.1016/j.xinn.2021.100141 PMID: 34557778
12. 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):W90-7. https://doi.org/10.1093/nar/gkw377 PMID: 27141961
13. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8. https://doi.org/10.1093/nar/gkz369 PMID: 31066453
14. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4(5):P3. PMID: 12734009
15. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017;45(D1):D183–9. https://doi.org/10.1093/nar/gkw1138 PMID: 27899595
16. Fuller T, Langfelder P, Presson A, Horvath S. Review of weighted gene coexpression network analysis. Handbook of Statistical Bioinformatics. 2011. p. 369–88.
17. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinformatics. 2012;13(1):1–21. https://doi.org/10.1186/1471-2105-13-1
18. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008
19. Tasaki S, Gaiteri C, Mostafavi S, Wang Y. Deep learning decodes the principles of differential gene expression. Nat Mach Intell. 2020;2(7):376–86. https://doi.org/10.1038/s42256-020-0201-6 PMID: 32671330

20. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform. 2018;19(4):575–92. https://doi.org/10.1093/bib/bbw139 PMID: 28077403

21. Thompson M, Gordon MG, Lu A, Tandon A, Halperin E, Gusev A, et al. Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. Nat Commun. 2022;13(1):5704. https://doi.org/10.1038/s41467-022-33212-0 PMID: 36171194

22. Figueiredo RQ, Del Ser SD, Raschka T, Hofmann-Apitius M, Kodamullil AT, Mubeen S, et al. Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets. BMC Bioinformatics. 2022;23(1):231. https://doi.org/10.1186/s12859-022-04765-0 PMID: 35705903

23. da Rocha EL, Ung CY, McGehee CD, Correia C, Li H. NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities. Nucleic Acids Res. 2016;44(10):e100. https://doi.org/10.1093/nar/gkw166 PMID: 26975659

24. Amici DR, Jackson JM, Truica MI, Smith RS, Abdulkadir SA, Mendillo ML. FIREWORKS: a bottom-up approach to integrative coessentiality network analysis. Life Sci Alliance 2021;4.

25. Ietswaart R, Gyori B, Bachman J, Sorger P, Churchman L. GeneWalk identifies relevant gene functions for a biological context using network representation learning. Genome Biol. 2021;22(1):1–35.

26. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35(Web Server issue):W169-75. https://doi.org/10.1093/nar/gkm415 PMID: 17576678

27. Miller HE, Bishop AJR. Correlation AnalyzeR: functional predictions from gene co-expression correlations. BMC Bioinformatics. 2021;22(1):206. https://doi.org/10.1186/s12859-021-04130-7 PMID: 33879054

28. 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. https://doi.org/10.1016/j.cels.2015.12.004 PMID: 26771021

29. 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. https://doi.org/10.1038/75556 PMID: 10802651

30. Hokama M, Oka S, Leon J, Ninomiya T, Honda H, Sasaki K, et al. Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. Cereb Cortex. 2014;24(9):2476–88. https://doi.org/10.1093/cercor/bht101 PMID: 23595620

31. Mougeot J, Li Z, Price A, Wright F, Brooks B. Microarray analysis of peripheral blood lymphocytes from ALS patients and the SAFE detection of the KEGG ALS pathway. BMC Med Genom. 2011;4(1):1–19. https://doi.org/10.1186/1756-0381-4-1

32. Hannenhalli S, Putt ME, Gilmore JM, Wang J, Parmacek MS, Epstein JA, et al. Transcriptional genomics associates FOX transcription factors with human heart failure. Circulation. 2006;114(12):1269–76. https://doi.org/10.1161/CIRCULATIONAHA.106.632430 PMID: 16952980

33. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC. Gene expression in peripheral blood mononuclear cells from children with diabetes. J Clin Endocrinol Metab. 2007;92(9):3705–11. https://doi.org/10.1210/jc.2007-0979 PMID: 17595242

34. Suresh R, Li X, Chiriac A, Goel K, Terzic A, Perez-Terzic C, et al. Transcriptome from circulating cells suggests dysregulated pathways associated with long-term recurrent events following first-time myocardial infarction. J Mol Cell Cardiol. 2014;74:13–21. https://doi.org/10.1016/j.yjmcc.2014.04.017 PMID: 24801707

35. Gandhi KS, McKay FC, Cox M, Riveros C, Armstrong N, Heard RN, et al. The multiple sclerosis whole blood mRNA transcriptome and genetic associations indicate dysregulation of specific T cell pathways in pathogenesis. Hum Mol Genet. 2010;19(11):2134–43. https://doi.org/10.1093/hmg/ddq090 PMID: 20190274

36. Pellagatti A, Armstrong R, Steeples V, Sharma E, Repapi E, Singh S, et al. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. Blood J Am Soc Hematol. 2018;132:1225–40.

37. Jayavelu AK, Wolf S, Buettner F, Alexe G, Häupl B, Comoglio F, et al. The proteogenomic subtypes of acute myeloid leukemia. Cancer Cell. 2022;40(3):301-317.e12. https://doi.org/10.1016/j.ccell.2022.02.006 PMID: 35245447

38. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, et al. The European genome-phenome archive of human data consented for biomedical research. Nat Genet. 2015;47(7):692–5. https://doi.org/10.1038/ng.3312 PMID: 26111507

39. Tombor LS, John D, Glaser SF, Luxán G, Forte E, Furtado M, et al. Single cell sequencing reveals endothelial plasticity with transient mesenchymal activation after myocardial infarction. Nat Commun. 2021;12(1):681. https://doi.org/10.1038/s41467-021-20905-1 PMID: 33514719

40. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol 2023. https://doi.org/10.1038/s41587-023-01767-y

41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

42. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417–9. https://doi.org/10.1038/nmeth.4197 PMID: 28263959

43. Lim N, Tesar S, Belmadani M, Poirier-Morency G, Mancarci BO, Sicherman J, et al. Curation of over 10000 transcriptomic studies to enable data reuse. Database (Oxford). 2021;2021:baab006. https://doi.org/10.1093/database/baab006 PMID: 33599246

44. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc: Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

45. Baumgarten N, Rumpf L, Kessler T, Schulz MH. A statistical approach for identifying single nucleotide variants that affect transcription factor binding. iScience. 2024;27(5):109765. https://doi.org/10.1016/j.isci.2024.109765 PMID: 38736546

46. Pinero J, Bravo A, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2016;44(D1):gkw943. https://doi.org/10.1093/nar/gkw943

47. Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res. 2017;45(D1):D877–87. https://doi.org/10.1093/nar/gkw1012 PMID: 27899610

48. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

49. Russo PST, Ferreira GR, Cardozo LE, Bürger MC, Arias-Carrasco R, Maruyama SR, et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinform. 2018;19(1):56. https://doi.org/10.1186/s12859-018-2053-1 PMID: 29458351

50. 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. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

51. Raina P, Guinea R, Chatsirisupachai K, Lopes I, Farooq Z, Guinea C, et al. GeneFriends: gene co-expression databases and tools for humans and model organisms. Nucleic Acids Res. 2023;51(D1):D145–58. https://doi.org/10.1093/nar/gkac1031 PMID: 36454018

52. Obayashi T, Hayashi S, Shibaoka M, Saeki M, Ohta H, Kinoshita K. COXPRESdb: a database of coexpressed gene networks in mammals. Nucleic Acids Res. 2008;36(Database issue):D77-82. https://doi.org/10.1093/nar/gkm840 PMID: 17932064

53. 53. Wenbin Wei SA, diffcoexp. 2018. https://doi.org/10.18129/B9.BIOC.DIFFCOEXP

54. Zogopoulos VL, Malatras A, Kyriakidis K, Charalampous C, Makrygianni EA, Duguez S, et al. HGCA2.0: an RNA-Seq based webtool for gene coexpression analysis in homo sapiens. Cells. 2023;12(3):388. https://doi.org/10.3390/cells12030388 PMID: 36766730

55. Guan Y, Myers C, Hess D, Barutcuoglu Z, Caudy A, Troyanskaya O. Predicting gene function in a hierarchical context with an ensemble of classifiers. Genome Biol. 2008;9(1):1–18.

56. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7. https://doi.org/10.1089/omi.2011.0118 PMID: 22455463

57. Wang J, Huang Q, Liu Z-P, Wang Y, Wu L-Y, Chen L, et al. NOA: a novel Network Ontology Analysis method. Nucleic Acids Res. 2011;39(13):e87. https://doi.org/10.1093/nar/gkr251 PMID: 21543451

58. 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.

59. Wong AK, Krishnan A, Troyanskaya OG. GIANT 2.0: genome-scale integrated analysis of gene networks in tissues. Nucleic Acids Res. 2018;46(W1):W65–70. https://doi.org/10.1093/nar/gky408 PMID: 29800226

60. Yu G, Fu G, Wang J, Zhao Y. NewGOA: predicting New GO Annotations of proteins by bi-random walks on a hybrid graph. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(4):1390–402. https://doi.org/10.1109/TCBB.2017.2715842 PMID: 28641268

61. Zhang J, Zou S, Deng L. Gene Ontology-based function prediction of long non-coding RNAs using bi-random walk. BMC Med Genomics. 2018;11(Suppl 5):99. https://doi.org/10.1186/s12920-018-0414-2 PMID: 30453964

62. Yu G, Wang K, Fu G, Guo M, Wang J. NMFGO: Gene function prediction via nonnegative matrix factorization with gene ontology. IEEE/ACM Trans Comput Biol Bioinform. 2018;17:238–49.

63. Kennedy VE, Smith CC. FLT3 mutations in acute myeloid leukemia: key concepts and emerging controversies. Front Oncol. 2020;10:612880. https://doi.org/10.3389/fonc.2020.612880 PMID: 33425766

64. Grafone T, Palmisano M, Nicci C, Storti S. An overview on the role of FLT3-tyrosine kinase receptor in acute myeloid leukemia: biology and treatment. Oncol Rev. 2012;6(1):e8. https://doi.org/10.4081/oncol.2012.e8 PMID: 25992210

65. Friedman R. The molecular mechanisms behind activation of FLT3 in acute myeloid leukemia and resistance to therapy by selective inhibitors. Biochim Biophys Acta (BBA)-Rev Cancer. 2022;1877:188666.

66. Ai D, Chen C, Han S, Ganda A, Murphy AJ, Haeusler R, et al. Regulation of hepatic LDL receptors by mTORC1 and PCSK9 in mice. J Clin Invest. 2012;122(4):1262–70. https://doi.org/10.1172/JCI61919 PMID: 22426206

67. Bonacina F, Moregola A, Svecla M, Coe D, Uboldi P, Fraire S, et al. The low-density lipoprotein receptor-mTORC1 axis coordinates CD8+ T cell activation. J Cell Biol. 2022;221(11):e202202011. https://doi.org/10.1083/jcb.202202011 PMID: 36129440

68. Zavadil J, Böttinger EP. TGF-beta and epithelial-to-mesenchymal transitions. Oncogene. 2005;24(37):5764–74. https://doi.org/10.1038/sj.onc.1208927 PMID: 16123809

69. Piera-Velazquez S, Jimenez SA. Endothelial to mesenchymal transition: role in physiology and in the pathogenesis of human diseases. Physiol Rev. 2019;99(2):1281–324. https://doi.org/10.1152/physrev.00021.2018 PMID: 30864875

70. Bunton-Stasyshyn RKA, Saccon RA, Fratta P, Fisher EMC. SOD1 function and its implications for amyotrophic lateral sclerosis pathology: new and renascent themes. Neuroscientist. 2015;21(5):519–29. https://doi.org/10.1177/1073858414561795 PMID: 25492944

71. Saccon RA, Bunton-Stasyshyn RKA, Fisher EMC, Fratta P. Is SOD1 loss of function involved in amyotrophic lateral sclerosis?. Brain. 2013;136(Pt 8):2342–58. https://doi.org/10.1093/brain/awt097 PMID: 23687121

72. Pansarasa O, Bordoni M, Diamanti L, Sproviero D, Gagliardi S, Cereda C. SOD1 in amyotrophic lateral sclerosis: "Ambivalent" behavior connected to the disease. Int J Mol Sci. 2018;19(5):1345. https://doi.org/10.3390/ijms19051345 PMID: 29751510

73. Hobbs HH, Brown MS, Goldstein JL. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. Hum Mutat. 1992;1(6):445–66. https://doi.org/10.1002/humu.1380010602 PMID: 1301956

74. Chora JR, Medeiros AM, Alves AC, Bourbon M. Analysis of publicly available LDLR, APOB, and PCSK9 variants associated with familial hypercholesterolemia: application of ACMG guidelines and implications for familial hypercholesterolemia diagnosis. Genet Med. 2018;20(6):591–8. https://doi.org/10.1038/gim.2017.151 PMID: 29261184

75. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014. p. 701–10.

76. Kiskinis E, Sandoe J, Williams LA, Boulting GL, Moccia R, Wainger BJ, et al. Pathways disrupted in human ALS motor neurons identified through genetic correction of mutant SOD1. Cell Stem Cell. 2014;14(6):781–95. https://doi.org/10.1016/j.stem.2014.03.004 PMID: 24704492

77. Robasky K, Lewis N, Church G. The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet. 2014;15:56–62.

78. Zhu Y, Qiu P, Ji Y. TCGA-assembler: open-source software for retrieving and processing TCGA data. Nature Methods. 2014;11:599–600.

79. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648–60. https://doi.org/10.1126/science.1262110 PMID: 25954001

80. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45(W1):W98–102. https://doi.org/10.1093/nar/gkx247 PMID: 28407145

81. Wood V, Lock A, Harris MA, Rutherford K, Bähler J, Oliver SG. Hidden in plain sight: what remains to be discovered in the eukaryotic proteome?. Open Biol. 2019;9(2):180241. https://doi.org/10.1098/rsob.180241 PMID: 30938578

82. Likhoshvai V, Golubyatnikov V, Khlebodarova T. Limit cycles in models of circular gene networks regulated by negative feedback loops. BMC Bioinformatics. 2020;21(1):1–15.

83. Chang R, Yee K-L, Sumbria RK. Tumor necrosis factor α Inhibition for Alzheimer's Disease. J Cent Nerv Syst Dis. 2017;9:1179573517709278. https://doi.org/10.1177/1179573517709278 PMID: 28579870

84. Decourt B, Lahiri DK, Sabbagh MN. Targeting tumor necrosis factor alpha for Alzheimer's disease. Curr Alzheimer Res. 2017;14(4):412–25. https://doi.org/10.2174/1567205013666160930110551 PMID: 27697064

85. Caraci F, Battaglia G, Bruno V, Bosco P, Carbonaro V, Giuffrida ML, et al. TGF-β1 pathway as a new target for neuroprotection in Alzheimer's disease. CNS Neurosci Ther. 2011;17(4):237–49. https://doi.org/10.1111/j.1755-5949.2009.00115.x PMID: 19925479

86. von Bernhardi R, Cornejo F, Parada GE, Eugenín J. Role of TGFβ signaling in the pathogenesis of Alzheimer's disease. Front Cell Neurosci. 2015;9426. https://doi.org/10.3389/fncel.2015.00426 PMID: 26578886

87. Mu X, He J, Anderson D, Springer J, Trojanowski J. Altered expression of bcl-2 and bax mRNA in amyotrophic lateral sclerosis spinal cord motor neurons. Annals Neurol: Off J Am Neurol Assoc Child Neurol Soc. 1996;40(5):379–86.

88. Vukosavic S, Dubois-Dauphin M, Romero N, Przedborski S. Bax and Bcl-2 interaction in a transgenic mouse model of familial amyotrophic lateral sclerosis. J Neurochem. 1999;73(6):2460–8. https://doi.org/10.1046/j.1471-4159.1999.0732460.x PMID: 10582606

89. Hetz C, Thielen P, Fisher J, Pasinelli P, Brown RH, Korsmeyer S, et al. The proapoptotic BCL-2 family member BIM mediates motoneuron loss in a model of amyotrophic lateral sclerosis. Cell Death Differ. 2007;14(7):1386–9. https://doi.org/10.1038/sj.cdd.4402166 PMID: 17510659

90. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2022. Available from: https://www.R-project.org/

91. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616 PMID: 19910308