Gene Set linkage analysis: a tool for interpreting the overall functional impacts of observed transcriptomic changes

Pengcheng Chen
Institute of Biopharmaceutical Technology, School of Basic Medical Sciences, Zhejiang University
https://orcid.org/0000-0002-7144-574X
Xi Liang
Institute of Biopharmaceutical Technology, School of Basic Medical Sciences, Zhejiang University
Yun Li
Institute of Biopharmaceutical Technology, School of Basic Medical Sciences, Zhejiang University
Xiaoxuan Wang
Institute of Biopharmaceutical Technology, School of Basic Medical Sciences, Zhejiang University
Xin Chen (xinchen@zju.edu.cn)

Software

Keywords: Gene set annotation, Interaction network, Gene expression, Gene Ontology

Posted Date: May 14th, 2019

DOI: https://doi.org/10.21203/rs.2.9595/v1

Abstract

Background To date, the majority of software tools developed for high-level interpretation of transcriptomics data were based on annotation enrichment. These tools use existing biological concepts to describe the observed omics changes. However, if an observation cannot be accurately described by an existing concept, these tools can not report any term or will report very general terms (such as “biological process”, GO:0008150), which provides limited assistance for researchers to understand the data and design further investigation. Results We present the gene set linkage analysis (GSLA) tool for interpretation of the collective functional impacts of a set of changed genes. The GSLA algorithm relies on a functional association network to evaluate whether the observed omics changes collectively interfere with functions of known biological processes. Although an omics change may not be accurately described by an existing concept, its functional impact may still be described by well-established concepts. GSLA has been shown useful in several previous studies. It derived novel insights into high-level coordination of physiological processes, where conventional annotation enrichment-based tools did not provide similar insights. This standalone version of GSLA tool integrates interaction networks for four species (i.e., A. thaliana, D. melanogaster, H. sapiens and S. cerevisiae) and using four kinds of annotation gene sets (i.e., Gene Ontology, Reactome pathway, Panther pathway and Wikipathways) for each species. Conclusions The GSLA tool is designed to interpret the collective functional impacts of a set of changed genes. Its usefulness has been demonstrated in a series of previous researches analyzing transcriptomic changes. GSLA is freely available at https://github.com/synergy-zju/gsla.

Background

Biological traits of interest, such as disease states and crop yields, are produced by interactions among different physiological functions and biological processes, and each physiological function and biological process is implemented by a network of molecular interactions to ensure its robustness and coordination with other functions and processes. In the age of omics, we know that each individual has genetic idiosyncrasies and that biological functions and processes are more robust and consistent across populations. Therefore, researchers tend to uncover high-level laws of biology from omics data at the biological process level because these laws are more robust, logical, and can be consistently applied in a population.

Various tools have been developed for omics data interpretation. Most tools are based on the annotation enrichment strategy. The widely used tools include GSEA [1], DAVID [2] and Enrichr [3]. These tools use existing concepts of biological processes to describe the observed omics changes. However, if an observation (i.e., an actual biological process) cannot be accurately described by an existing concept, these tools do not report the biological process or report very general biological processes (such as “biological process”, GO:0008150), which does not help researchers understand the data or provide direction for further investigation.

Implementation

To address this challenge, we developed the gene set linkage analysis (GSLA) method, which relies on a functional association network to evaluate whether observed omics changes collectively interfere with the functions of known biological processes. Although an omics change may not be accurately described by an existing concept, its functional impact can still be described by well-established concepts. In previous studies, GSLA has been shown to distinguish sensitive and insensitive cell lines based on the transcriptome changes measured before the onset of apoptosis. GSLA correctly anticipated that only the sensitive cell lines would undergo apoptosis [4]. GSLA was also shown to be capable of identifying a key functional impact exerted by stem cells to rescue fulminant hepatic failure in pigs. A cytokine known to regulate this function significantly improved animal survival in both rat and pig models [5]. In both cases, widely used annotation enrichment-based approaches did not provide similar insights.

This software article describes the standalone version of the GSLA tool, which is an operating system-independent Java program. This program supports 4 species: *A. thaliana, D. melanogaster, H. sapiens* and *S. cerevisiae.* It can use local computing resources to improve analyzing efficiency. This program has been successfully applied in a number of previous studies, and the results show that GSLA can provide useful insights that are valuable for designing further investigations, whereas the widely used annotation enrichment-based approaches do not provide similar insights.

Algorism description

GSLA relies on testing two hypotheses to detect substantial functional linkages between biologically meaningful gene sets (one set represents the observed changes, and the other set represents a known biological process or function). The first hypothesis (Q1) expects that the inter-gene-set interaction density between functionally linked gene sets is higher than the background interaction density between random gene sets (i.e., the interactions between two functionally linked gene sets are stronger than those between two random gene sets). The second hypothesis (Q2) expects that the observed high density between functionally linked gene sets can only be observed in a biologically correct interactome (i.e., the interactions between two functionally linked gene sets observed in a biologically correct interactome are greater than those observed in random interactomes consisting of the same genes and same topology). Q1 tests the strength of the functional linkage between two gene sets. Q2 verifies that the observed robust linkage is due to the biologically accurate network topology (i.e., our knowledge of molecular mechanisms) rather than the compositions of these two gene sets. Q2 is used to remove the confounding factor of gene set composition and ensure the “biological significance” of the detected functional linkages between gene sets. Q1 and Q2 are related but different tests. These tests complement each other to increase the sensitivity and specificity of GSLA.

To test Q1, the density of the inter-gene-set interactions is first calculated as the observed number of inter-gene-set interactions divided by the total number of possible inter-gene-set gene pairs (See Fig. 1, Q1). A common gene shared by two gene sets is treated as two distinct delegate genes, and each delegate gene belongs to only one gene set. Any gene that interacts with the shared gene is considered to interact with

both delegate genes in both gene sets. Two gene sets sharing a gene do not constitute an inter-gene-set interaction.

To test Q2, the interaction network is randomized 100000 times. Suppose that there are two interactions that involve four genes (i.e., G1–G2 and G3–G4). By replacing these two interactions with two new interactions (i.e., G1–G3 and G2–G4), the topology of the interaction network changes, but the gene composition and number of neighbors of each gene remain the same. This topology perturbation is performed 100000 times to create one random interactome. Using 100000 random interactomes, the P-value for Q2 is computed as the fraction of random interactomes in which the densities of the inter-gene-set interactions are higher than the densities observed in the correct interaction network (See Fig. 1, Q2).

Input gene set

To interpret the biological significance of an experimentally observed molecular profile, we recommend using the “most significant” 20–200 genes to represent an observed omics change. The GSLA algorithm considers all query genes equal; therefore, using too many genes will likely dilute the focus on the most important changes. However, because profiling data are intrinsically noisy and interaction networks have false positives and false negatives, using too few genes will not suppress this noise and take advantage of the corroborative strength of a gene set. In our experience, using the top 20–200 significant genes to represent a molecular profile usually produces consistent results.

Interaction network

The quality and coverage of interaction networks used in GSLA determines whether the query gene set can be correctly annotated to biologically meaningful gene sets. For *D. melanogaster* and *S. cerevisiae*, their experimentally reported protein-protein interactions [6-8] represent significant proportions of their entire interactomes. For *A. thaliana* and *H. sapiens*, because their experimentally reported interactions are limited, we used the predicted interaction networks (HIR [4] and PAIR [9]) as alternative choices.

Annotation gene sets

GSLA assesses the functional linkage between an observed gene set and a functional gene set to annotate the potential functional impacts of the observed gene set. Several types of functional gene sets are used, including GO biological processes [10], the Panther pathway [11, 12], the Reactome pathway [13], and WikiPathways [14]. For all species supported in GSLA, GO biological processes are used as the default functional gene sets to annotate the functional impacts of a query gene set. For each GO biological process, we only consider terms with 20-200 member genes that are not inferred from electronic annotations (IEA).

P value and density

According to the central limit theorem, the expected density of the interactions in a sufficiently large number of pairs of genes, N, follows a normal distribution. Among the current model organisms, the entire interactome in yeast has been experimentally verified and can be used as a reference for estimating interaction probabilities. In any species, if the frequency of gene interactions in random gene pairs equals the frequency of protein interactions in yeast (1/775) [15], the expected mean of this normal distribution is the density of the interactions among all genes (0.0008). As shown in Additional file 1, the expected standard deviation approaches 0.0012 as N increases in our simulation, where N random gene pairs are sampled 100000 times to calculate the mean and standard deviation of the interaction density. According to these data, the not very high inter-gene-set interaction density of 0.01 translates into a small P-value ($P<10\text{-}10$) if the two gene sets both have ≥ 20 members ($N \geq 400$). The functional linkages between two gene sets that are worthy of experimental investigation are anticipated to be much stronger than those linkages that marginally pass the significance threshold. Therefore, we set the cutoff value for Q1 to ≥ 0.01 (indicating that the functional linkages between two gene sets are sufficiently strong) and that for Q2 to ≤ 0.001 (indicating that the functional linkages between two gene sets are significant) as default. Users can optimize the results by adjusting the cutoff values.

Output file

The output file contains annotation gene sets that have functional linkages with the queried gene set. These annotation gene sets represent phenotype changes that are responsible for the genotype changes.

Result

We use two examples to illustrate the usage of GSLA.

Understanding the stress factors in yeast secondary fermentation

GSLA is frequently used to derive insights into omics changes that are associated with specific phenotypes at the biological process level. Here, the yeast secondary fermentation data (GEO accession GSE29273 [16]) were reanalyzed to illustrate this usage. To investigate the main factors that affect yeast physiology during its secondary fermentation, Penacho et al. performed a genome-wide expression analysis of an industrial wine yeast strain under real secondary fermentation conditions. Comparisons were performed between samples obtained at different time points to identify genes that are simultaneously up- or downregulated. The authors analyzed the biological annotations that were significantly associated with these genes using GENECODIS. Their results showed that the main factor

influencing gene transcription during secondary fermentation is alcohol respiration, but GENECODIS did not report other stress factors, such as nitrogen starvation.

In our analysis, the GEO2R online service was used to identify differentially expressed genes that satisfied “t statics >0” or “t statics < 0” and were simultaneously up- or downregulated across different time points. Genes were sorted by the absolute value of the t-statics, and the top 50 genes with significantly changed expression, containing both up- and downregulated genes (Additional file 2), were analyzed.

The GSLA results are provided in Additional file 3. In our GSLA analysis of the upregulated genes, some terms were consistent with the results of GENECODIS, such as “electron transport chain”, “cellular respiration”, and “TCA cycle”. In addition, GSLA reported another category of terms that are related to nitrogen metabolism, including “glutathione metabolic process” and “metabolism of amino acids and derivatives”. Nitrogen starvation is an important stress factor in secondary fermentation, and the consumption of glutamine is a manifestation of nitrogen starvation [17]. In the original publication, GENECODIS also identified new peripheral terms that are consistent with nitrogen starvation, such as increased ROS, Sod2p, Cat1p activities and the upregulation of autophagy. However, GENECODIS failed to identify any term directly related to nitrogen starvation.

Similarly, the analysis of the downregulated genes also showed that GSLA and GENECODIS produced consistent terms, but the terms produced by GSLA are more specific and relevant.

We also used DAVID [2] and WebGestalt [18] to analyze the same gene list. GSLA was the only tool that reported nitrogen starvation-related processes in the results (Additional file 4). These results showed that GSLA has an improved capability of understanding the functional impacts of an omics change.

Predicting functions of microRNAs

GSLA can also predict the overall functional impacts of a set of genes. MicroRNAs are known to regulate target genes to achieve their high-level functions. In this example, GSLA was used to predict high-level microRNA function via an analysis of its target genes.

The target genes of an Arabidopsis microRNA, ath-MIR417, are listed in Additional file 5, along with their GO annotations. Our GSLA analysis of these target genes identified a biological function that has not been previously annotated to any of the ath-MIR417 target genes, namely, water deprivation (Table 1). This biological process was identified because two genes functioning in water deprivation have strong functional interactions with one target gene of ath-MIR417, and these functional interactions passed the GSLA criteria. This prediction is consistent with a study showing that the expression of ath-MIR417 was significantly changed under dehydration stress [19].

Discussion

GSLA uses interaction networks and annotated functional gene sets to identify the potential functions of a query gene set. To demonstrate the utility of GSLA, we reanalyzed two published datasets, which were parallelly analyzed with widely used enrichment-based tools. In the first case, GSLA demonstrated superior capability in identifying known stress factors that were missed by enrichment-based tools. In the second case, we showed that GSLA may be used to predict the functional impacts of a biological meaningful gene set (e.g. target genes of a microRNA).

Previously, GSLA has also been used in a series of transcriptomic data analyses. It has been used in analysis of the expression profiles of multiple myeloma cells treated by lovastatin. Results show that different mechanisms led sensitive cells to apoptosis and insensitive cells to survival. The identification of disrupted apoptosis pathway was achieved using transcriptomic profiles observed at an early time point before apoptosis really took place. GSLA were compared with other enrichment-based tools. Only GSLA produced this insight [4]. In another study of human bone mesenchymal stem cells rescuing fulminant hepatic failure in pigs, GSLA found a biological process that explains a significant proportion of the transcriptomic benefits of stem cells. Further analysis confirmed that a known regulatory cytokine of this biological process can increase survival of fulminant hepatic failure in both pig and rate models [5]. In an Arabidopsis study, GSLA explained the mechanism of abscisic acid hypersensitivity in Arabidopsis roa1 mutant during its seedling stage, and identified a gene without significant expression change but with significant functional associations to genes that show expression changes. This gene was later demonstrated to have a key role in producing the phenotype [9]. In these examples, enrichment-based analysis tools were parallelly used. Only GSLA produced these insights.

A limitation of GSLA is the requirement for high quality interaction networks to infer functional synergy and high quality annotation gene sets to express the collective functional impacts of a gene set. For a number of species, lacking high quality interaction networks and high quality annotation gene sets limited the capability of GSLA to produce insights that may inspire further study. Continued development of high quality interaction networks and annotation gene sets for more model organisms will fill the gap.

Conclusion

The GSLA tool is designed for interpretation of the collective functional impacts of a set of changed genes. It takes advantage of multithread computing resources to perform a network-based interpretation of gene set function on the most common computing platforms. GSLA typically produce terms that consist of annotation enrichment tools, with additional insights into the physiology of the observed changes. Its usefulness has been demonstrated in a number of previous applications analyzing transcriptomic changes.

Abbreviations

GSLA: Gene Set Linkage Analysis

GO: Gene Ontology

HIR: Human Interactome Resource

PAIR: Predicted Arabidopsis Interactome Resource

GSEA: Gene Set Enrichment Analysis

DAVID: The Database for Annotation, Visualization and Integrated Discovery

Declarations

Availability and requirements

Project name: GSLA

Project home page: https://github.com/synergy-zju/gsla

Operating system: Platform independent

Programming language: Java

Other requirements: None

License: NPOSL-3.0

Any restrictions to use by non-academics: no restriction

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The GSLA tool is freely accessible at https://github.com/synergy-zju/gsla/blob/master/gsla_app.jar.zip. An user manual is performed at https://github.com/synergy-zju/gsla/blob/master/README.md. Relevant data are available at https://github.com/synergy-zju/gsla. The GEO data used in the yeast second fermentation analysis are available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29273.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the National Natural Science Foundation of China [31571356 and 81830073] and Zhejiang Provincial Natural Science Foundation of China [LR13C020001]. The funding body had no role in the design of the study, collection, analysis, or interpretation of data, or in writing the manuscript.

Authors' contributions

XC and PCC conceived the project and designed the manuscript. PCC developed the software, wrote the manuscript and software documentation, and carried out the performance comparisons. XL, XXW and YL test the software and proposed additional features and improvements to the software. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Reference

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005, 102(43):15545-15550.

2. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 2003, 4(9).

3. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A *et al*: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016, 44(W1):W90-97.

4. Zhou X, Chen P, Wei Q, Shen X, Chen X: Human interactome resource and gene set linkage analysis for the functional interpretation of biologically meaningful gene sets. *Bioinformatics* 2013, 29(16):2024-2031.

5. Shi D, Zhang J, Zhou Q, Xin J, Jiang J, Jiang L, Wu T, Li J, Ding W, Li J *et al*: Quantitative evaluation of human bone mesenchymal stem cells rescuing fulminant hepatic failure in pigs. *Gut* 2016.

6. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R *et al*: The BioGRID interaction database: 2019 update. *Nucleic Acids Res* 2019, 47(D1):D529-D541.

7. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U *et al*: The IntAct molecular interaction database in 2012. *Nucleic Acids Res* 2012, 40(Database issue):D841-846.

8. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E *et al*: MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 2012, 40(Database issue):D857-861.

9. Yao H, Wang X, Chen P, Hai L, Jin K, Yao L, Mao C, Chen X: Predicted Arabidopsis Interactome Resource and Gene Set Linkage Analysis: A Transcriptomic Analysis Resource. *Plant Physiol* 2018, 177(1):422-433.

10. The Gene Ontology C: The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019, 47(D1):D330-D338.

11. Nikolsky Y, Bryant J: Protein networks and pathway analysis. Preface. *Methods Mol Biol* 2009, 563:v-vii.

12. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD: PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* 2017, 45(D1):D183-D189.

13. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B *et al*: The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 2018, 46(D1):D649-D655.

14. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Melius J, Cirillo E, Coort SL, Digles D *et al*: WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res* 2018, 46(D1):D661-D667.

15. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N *et al*: High-quality binary protein interaction map of the yeast interactome network. *Science* 2008, 322(5898):104-110.

16. Penacho V, Valero E, Gonzalez R: Transcription profiling of sparkling wine second fermentation. *Int J Food Microbiol* 2012, 153(1-2):176-182.

17. Martinez-Rodriguez AJ, Carrascosa AV, Martin-Alvarez PJ, Moreno-Arribas V, Polo MC: Influence of the yeast strain on the changes of the amino acids, peptides and proteins during sparkling wine production by the traditional method. *J Ind Microbiol Biotechnol* 2002, 29(6):314-322.

18. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B: WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* 2017.

19. Jung HJ, Kang H: Expression and functional analyses of microRNA417 in Arabidopsis thaliana under stress conditions. *Plant Physiol Bioch* 2007, 45(10-11):805-811.

Table 1

Table 1

Results of a GSLA analysis of ath-MIR417 target genes.

GO Term ID	GO Description	P value	Density
GO:0042631	cellular response to water deprivation	1.00E-04	0.01282
GO:0006281	DNA repair	9.00E-04	0.00396
GO:0071103	DNA conformation change	0	0.02173
GO:0032508	DNA duplex unwinding	8.00E-04	0.02564

Figures

$$\text{Denstiy(Q1)} = \frac{IntNum_{AB}}{Size_A \times Size_B}$$

$$\text{P value(Q2)} = \frac{Count_{100000}(IntNum_{A'B'} > IntNum_{AB})}{100000}$$

A, B	Gene set A and gene set B.
$IntNum_{AB}$	The number of interactions between gene set A and B in the real interactome.
$Size_A$, $Size_B$	The number of genes of gene set A and B.
$IntNum_{A'B'}$	The number of interactions between gene set A and B in the random interactome.
$Count_{100000}$	The times of $IntNum_{A'B'} > IntNum_{AB}$ when generating 100000 times of random interactomes.

Figure 1

The formula used in GSLA to calculate the Density (Q1) and P-value (Q2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile3.xlsx
- Additionalfile1.docx

- Additionalfile5.xlsx
- Additionalfile4.xlsx
- Additionalfile2.xlsx