ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Uncovering the Physiological Impacts of Omics Changes for *Oryza sativa* with Gene Set Linkage Analysis

Pengcheng Chen[1], Ling Wang[2] and Xin Chen[3,*]

[1]College of Life Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou, P.R.China
[2]College of Chemical and Biological Engineering, Zhejiang University, 866 Yuhangtang Road, Hangzhou, P.R.China
[3]Institute of Pharmaceutical Biotechnology, Zhejiang University, 866 Yuhangtang Road, Hangzhou, P.R.China
[*]Corresponding author

Abstract—**High throughput omics profiling technologies enable comprehensive and accurate measurement of plant physiology. However, currently, most high-throughput results are only used for discovery of individual significantly changed genes. High-level coordination of physiological mechanisms is rarely discovered from omics results. This is because of the lack of a powerful tool that can interpret the physiological impacts of an observed omics change. This work presents the gene set linkage analysis (GSLA) tool for rice, which uses a functional interaction network to extend the functional impacts of observed individual gene changes and to finally obtain insights on the potential physiological impacts of an observed omics change. The GSLA service for rice is freely available at http://service.synergylab.info/gsla/faces/gsla_rice.xhtml.**

Keywords-oryza sativia; biological process; omics data; functional interaction network; gene set annotation

I. Introduction

Rice (Oryza sative) is one of the most important crops and extensively studied model plants across the world [1]. Therefore, how to improve the economical traits of rice, such as the resistance to plant disease and inset pest, tolerance to salinity and drought, and detoxification of herbicide, become the primary subject in this field. With the help of high-throughput technologies, there are plenty of studies that measured rice genomes [2, 3], transcriptomes [4], proteomes [5] with the aim to discover genes or biological processes that are relevant to improving these economical traits. When studied at the individual gene level, diversities between different rice lines that are pooled in one study may lead to noise and difficulty in data analysis. Therefore, a robust and effective analysis tool for rice omics is desired to decipher the overall physiological changes from omics data.

Many tools have emerged for omics data analysis. Among these tools, some are widely used, such as GSEA [6], David [7] and Enrichr [8]. These tools are based on the annotation enrichment strategy. They use well-defined biological concepts to annotate the collective functions of multiple observed changes in the omics data. However, when the observation (i.e. the actual biological process) cannot be accurately described by an existing concept, these tools tend to report no biological process or very general biological processes, which do not help researchers to understand the data or to suggest directions for further investigation. On the other hand, innovative research tends to explore previously uncharted areas of life mechanisms, where there were no established concepts to accurately describe the observed changes.

To meet this challenge, we developed the gene set linkage analysis (GSLA) method [9], which relies on a functional association network to evaluate whether observed omics changes will collectively interfere with functions of known biological processes. Even when an omics change itself cannot be accurately described by an existing concept, its functional impacts may still be described by well-established concepts. In 2013, we published the GSLA algorithm with an implementation for analyzing human omics data [9]. This tool has shown improved analytical capability to anticipate the functional impacts of observed omics changes. For example, GSLA has been shown capable of distinguishing the lovastatin-sensitive and insensitive cell lines based on their transcriptome changes measured before the onset of apoptosis, as well as identifying a key therapeutic impact that stem cells exert to rescue fulminant hepatic failure in pigs. In these applications, traditional annotation enrichment-based tools did not provide similar insights.

To facilitate the analysis of potential physiological impacts of observed omics changes, we deployed an online GSLA web service for *Oryza sativia* with a built-in rice functional interactome. This web service can be freely accessed on the website
http://service.synergylab.info/gsla/faces/gsla_rice.xhtml.

II. Method

A. Funcitonal Linkage Analysis

GSLA uses two hypotheses to measure the strength and biological significance of functional linkages between two gene sets, i.e., one with observed omics changes and the other with well-defined biological processes. The first hypothesis (Q1) expects that interactions between two functionally linked gene sets are tighter than two random gene sets. The second hypothesis (Q2) expects that interactions between two functionally linked gene sets observed in the biologically correct interactome is higher than interactions observed in random interactomes consisting of the same genes and same topology. Q1 is designed to measure if the functional linkages between two gene sets are strong enough to make them affect each other, while Q2 ensures the biological significance of these functional linkages. These two hypotheses are connected to each other, and both of them ensure the sensitivity and specificity of GSLA.

B. *Method for Calculating DENSITY and P Value*

Formulae for calculating the density and P value are shown in Figure I. Figure I. Density is calculated as the number of interactions divided by the number of genes in gene set A and gene set B. To calculate P value, first, as shown in Figure II, a randomized interaction network is generated from the origin interaction network by replacing a gene with the other one which has the same number of edges (e.g., replacing gene A with gene X, Figure II). Second, count the number of interactions between two gene sets in the randomized network. Third, repeat the second step for 100,000 times, and count how many times the number of interactions between two gene sets in the randomized network is higher than that in the original interaction network. P value is computed as this number divided by 100,000.

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

$$P\ value(Q2) = \frac{Count_{10000}(IntNum_{A'B'} > IntNum_{AB})}{10000}$$

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 I. FORMULAE FOR CALCULATING DENSITY AND P VALUE

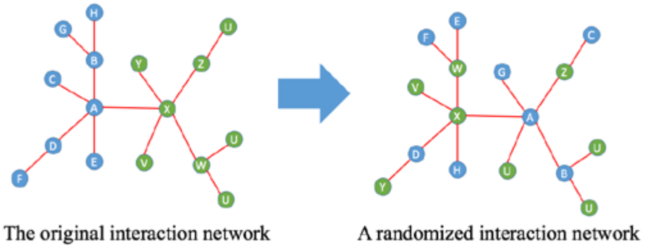

The original interaction network
A randomized interaction network

FIGURE II. SCHEME FOR GENERATING A RANDOMIZED INTERACITON NETWORK

C. *Reference Data*

For the deployment of GSLA, a functional interactome and annotation gene sets are needed. Here we used a probabilistic functional gene network: RiceNet [10], with 1,518,212 interactions that can cover 70.1% of coding genes. Functional gene sets are collected from GO[11] and RGAP [12].

To profile phenotype changes from omics data, we suggest to use the most significantly changed 20-200 genes, e.g., the top 20-200 changed genes from transcriptome analysis. In the core algorithm of GSLA, each gene has the same weight. Too few genes will miss some important functional linkages while excessive genes will bring in too much noise. After multiple tests of using different number of genes, we found that a gene set consisting of 20–200 genes may best represent an omics change for use in GSLA.

III. RESULT

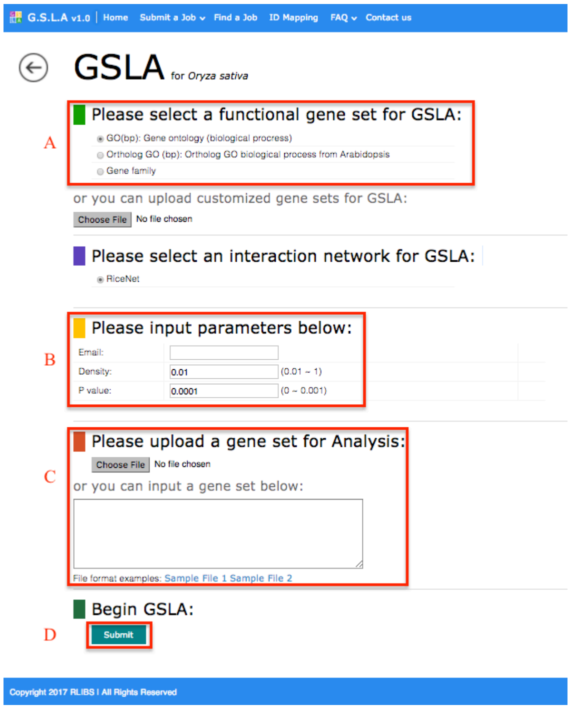
The main GSLA web interface is shown in Figure III. Steps of using GSLA to annotate a query gene set are as follows:

a. Select a functional gene set for annotation (Figure III, A);

b. Set the cutoff value of density and P value, and provide an email address (Figure III, Figure III. B);

c. Upload a query gene set (Figure III, C);

d. Click the “Submit” button (Figure III, D).



GSLA for *Oryza sativa*

Please select a functional gene set for GSLA:

● GO(bp): Gene ontology (biological process)
○ Ortholog GO (bp): Ortholog GO biological process from Arabidopsis
○ Gene family

or you can upload customized gene sets for GSLA:

No file chosen

Please select an interaction network for GSLA:

● RiceNet

Please input parameters below:

Email:		
Density:	0.01	(0.01 ~ 1)
P value:	0.0001	(0 ~ 0.001)

Please upload a gene set for Analysis:

No file chosen

or you can input a gene set below:

File format examples: Sample File 1 Sample File 2

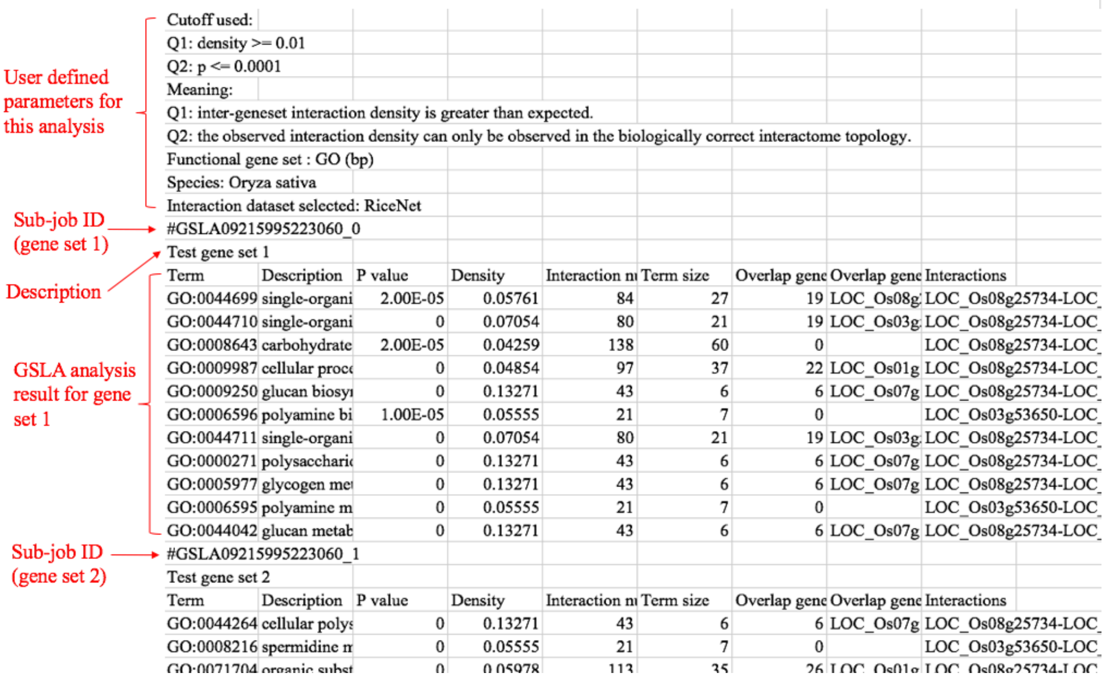
Begin GSLA:

Copyright 2017 RLIBS I All Rights Reserved

FIGURE III. GSLA WEB INTERFACE FOR RICE

When the computation is finished, the results will be sent to the provided email address automatically.

Figure IV is an example of GSLA analysis result file. The top 9 lines are parameters set by users, including cutoff (density and P value), functional gene set, species and interaction network. Lines starting with “#GSLA” are sub-job IDs. Content between two sub-job IDs is the analysis result for the query gene set, which includes a description line (starting with “#”) and the functionally linked gene sets. Each line shows a functionally linked gene set, which provides the following information.


User defined parameters for this analysis
Sub-job ID (gene set 1)
Description
GSLA analysis result for gene set 1
Sub-job ID (gene set 2)

Cutoff used:
Q1: density >= 0.01
Q2: p <= 0.0001
Meaning:
Q1: inter-geneset interaction density is greater than expected.
Q2: the observed interaction density can only be observed in the biologically correct interactome topology.
Functional gene set : GO (bp)
Species: Oryza sativa
Interaction dataset selected: RiceNet
#GSLA09215995223060_0
Test gene set 1

Term	Description	P value	Density	Interaction nı	Term size	Overlap gene	Overlap gene	Interactions
GO:0044699	single-organi	2.00E-05	0.05761	84	27	19	LOC_Os08g	LOC_Os08g25734-LOC_
GO:0044710	single-organi	0	0.07054	80	21	19	LOC_Os03g	LOC_Os08g25734-LOC_
GO:0008643	carbohydrate	2.00E-05	0.04259	138	60	0		LOC_Os08g25734-LOC_
GO:0009987	cellular proc	0	0.04854	97	37	22	LOC_Os01g	LOC_Os08g25734-LOC_
GO:0009250	glucan biosy	0	0.13271	43	6	6	LOC_Os07g	LOC_Os08g25734-LOC_
GO:0006596	polyamine bi	1.00E-05	0.05555	21	7	0		LOC_Os03g53650-LOC_
GO:0044711	single-organi	0	0.07054	80	21	19	LOC_Os03g	LOC_Os08g25734-LOC_
GO:0000271	polysacchari	0	0.13271	43	6	6	LOC_Os07g	LOC_Os08g25734-LOC_
GO:0005977	glycogen me	0	0.13271	43	6	6	LOC_Os07g	LOC_Os08g25734-LOC_
GO:0006595	polyamine m	0	0.05555	21	7	0		LOC_Os03g53650-LOC_
GO:0044042	glucan metab	0	0.13271	43	6	6	LOC_Os07g	LOC_Os08g25734-LOC_

#GSLA09215995223060_1
Test gene set 2

Term	Description	P value	Density	Interaction nı	Term size	Overlap gene	Overlap gene	Interactions
GO:0044264	cellular polys	0	0.13271	43	6	6	LOC_Os07g	LOC_Os08g25734-LOC_
GO:0008216	spermidine m	0	0.05555	21	7	0		LOC_Os03g53650-LOC_
GO:0071704	organic subst	0	0.05978	113	35	26	LOC_Os01g	LOC_Os08g25734-LOC

FIGURE IV. FORMAT OF GSLA ANALYSIS RESULT

- Column 1 (Term): ID of this functional gene set, such as "GO:0006421";
- Column 2 (Description): description or biological definition of this functional gene set;
- Column 3 (P value): biological significance of functional linkages between two gene sets.
- Column 4 (Density): strength of functional linkages between two gene sets.
- Column 5 (Interaction number): number of interactions between two gene sets.
- Column 6 (Term size): number of genes in this functional gene set.
- Column 7 (Overlap gene number): number of genes shared by the query gene set and the functional gene set.
- Column 8 (Overlap genes): shared genes between the two gene sets. For example, if geneA and geneB both exist in the query gene set and the functional gene set, it will be shown as: geneA|geneB.
- Column 9 (Interactions): interactions between the two gene sets. For example, if there are two pairs of interactions between the query gene set and the functional gene set: geneA interacting with geneB, geneC interacting with geneD, they will be shown as: geneA-geneB|geneC-geneD.

IV. CONCLUSION

We deployed an online GSLA web service for Oryza sativia. It may facilitate omics analysis, providing valuable research clues at the biological process-level.

ACKNOWLEDGMENT

This work is supported by Zhejiang Provincial Natural Science Foundation of China grant (LR13C020001) and National Natural Science Foundation of China grant (31571356).

REFERENCES

[1] Arce, A., The living fields: our agricultural heritage. Journal of the Royal Anthropological Institute, 2002. 8(3): p. 574-574.

[2] Goff, S.A., et al., A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science, 2002. 296(5565): p. 92-100.

[3] Yu, J., et al., A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science, 2002. 296(5565): p. 79-92.

[4] Xu, H., Y. Gao, and J.B. Wang, Transcriptomic Analysis of Rice (Oryza sativa) Developing Embryos Using the RNA-Seq Technique. Plos One, 2012. 7(2).

[5] Agarwal, G.K., et al., Proteome analysis of differentially displayed proteins as a tool for investigating ozone stress in rice (Oryza sativa L.) seedlings. Proteomics, 2002. 2(8): p. 947-959.

[6] Subramanian, A., 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): p. 15545-50.

[7] Dennis, G., et al., DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology, 2003. 4(9).

[8] Kuleshov, M.V., et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res, 2016. 44(W1): p. W90-7.

[9] Zhou, X., et al., Human interactome resource and gene set linkage analysis for the functional interpretation of biologically meaningful gene sets. Bioinformatics, 2013. 29(16): p. 2024-31.

[10] Ashburner, M., et al., Gene Ontology: tool for the unification of biology. Nature Genetics, 2000. 25(1): p. 25-29.

[11] Lee, T., et al., RiceNet v2: an improved network prioritization server for rice genes. Nucleic Acids Res, 2015. 43(W1): p. W122-7.

[12] Ohyanagi, H., et al., The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. Nucleic Acids Res, 2006. 34(Database issue): p. D741-4.