Toward comprehensive functional analysis of gene lists weighted by gene essentiality scores

Rui Fan, Qinghua Cui*

Department of Biomedical Informatics, Department of Physiology and Pathophysiology,
Center for Noncoding RNA Medicine, MOE Key Lab of Cardiovascular Sciences, School of
Basic Medical Sciences, Peking University, 38 Xueyuan Rd, Beijing, 100191, China

* To whom correspondence should be addressed:

Qinghua Cui, Email: cuiqinghua@hsc.pku.edu.cn

ABSTRACT

Gene functional enrichment analysis represents one of the most popular bioinformatics methods for annotating the pathways and function categories of a given gene list. Current algorithms for enrichment computation such as Fisher's exact test and hypergeometric test totally depend on the category count numbers of the gene list and one gene set. In this case, whatever the genes are, they were treated equally. However, actually genes show different scores in their essentiality in a gene list and in a gene set. It is thus hypothesized that the essentiality scores could be important and should be considered in gene functional analysis. For this purpose, here we proposed WEAT (<https://www.cuilab.cn/weat/>), a weighted gene set enrichment algorithm and online tool by weighting genes using essentiality scores. We confirmed the usefulness of WEAT using two case studies, the functional analysis of one aging-related gene list and one gene list involved in Lung Squamous Cell Carcinoma (LUSC). Finally, we believe that the WEAT method and tool could provide more possibilities for further exploring the functions of given gene lists.

INTRODUCTION

Gene functional enrichment analysis represents one class of the most popular tools in bioinformatics (1,2). Within a large group of methods, the hypergeometric test is one of the most frequently used ones and it has been widely used and continuously improved in many online tools including DAVID Bioinformatics Resources (3).

However, the current algorithms for gene functional analysis (e.g. hypergeometric test) are based on a hypothesis that all the genes in one gene set are identical, which is not the true cases in the various biological processes. In fact, different genes have different contributions to one biological process. For example, essential genes are normally more important than non-essential genes in molecular functions. Therefore, one critical problem is that, for any two gene lists with the same count number of hitting genes within a given gene set, current algorithms will show the same significance. But in fact, the significances of the two input gene lists should be different. That is, the list whose hitting genes with greater essentiality scores should be more important than the other list. However, all of the current algorithms fail to address the above issues.

To address the above issue, here we proposed a weighted functional enrichment method based on hypergeometric test by introducing a prior weight to each gene. We first collected a group of scores representing the gene essentiality. Then we developed Weighted Enrichment Analysis Tool (WEAT, <https://www.cuilab.cn/weat/>) algorithm and tool to perform weighted functional enrichment analysis. Finally, we showed the usefulness of WEAT using two case studies.

METHODS AND MATERIALS

Conventional Algorithms for Gene Functional Enrichment Analysis

Hypergeometric test and binomial test have been widely accepted for gene functional enrichment analysis (4). The above two types tests are quite similar and the major difference is the sampling with or without replacement. Thus, the conventional methods for gene functional enrichment analysis is described as follows. Suppose A is the intersection set of the inputted gene list and one known gene set. Let k be the size of A , that is $k = |A|$. Normally, hypergeometric test is based on hypergeometric distribution, and it is to estimate the probability of $X > k$ intersection genes in a specific gene set, which is defined as follows:

$$P(X > k) = \sum_{a=k+1}^{\min(|m|, |n|)} \frac{\binom{|m|}{a} \binom{|N|-|m|}{|n|-a}}{\binom{|N|}{|n|}} \quad (1)$$

where N is the set of genes in the background, m denotes the inputted gene list, n is a specific known gene set. Hence, we can obtain the significance of each gene set from the probability calculated by equation (1).

Weighted Algorithm for Gene Functional Enrichment Analysis

Here we presented the algorithm for weighted gene functional enrichment analysis, which is an extension of the hypergeometric test by replacing the gene counts with the sum score of those genes, which is described as follows:

$$N' = \sum_{i \in N} w_i^Q, n' = \sum_{i \in n} w_i^Q, m' = \sum_{i \in m} w_i^Q, a' = \sum_{i \in a} w_i^Q \quad (2)$$

where w_i represents the weight for gene i and Q denotes the scaling factor of scores. Nevertheless, introducing the scores into the hypergeometric test violates the discreteness of hypergeometric distribution. Because the calculation of hypergeometric distribution is based on the calculation factorial, which only works for positive integers. Hence, to solve this problem, we introduce the Gamma function as the natural extension of factorial calculation, which is defined as follows:

$$\forall x \in \mathbb{C}, \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt = (x+1)! \quad (3)$$

Gamma function can extend the domain of factorial function from positive integers to all complex numbers. Here we only use the positive real part. We then calculate the probability of weighted algorithm with positive real numbers defined as follows:

$$P'(X > k) = \int_{a'}^{\min(m', n')} \frac{\binom{m'}{a} \binom{N'-m'}{n'-a}}{\binom{N'}{n'}} da \approx \sum_{a=a'+1}^{\min(m', n')} \frac{\binom{m'}{a} \binom{N'-m'}{n'-a}}{\binom{N'}{n'}} \quad (4)$$

The discrete hypergeometric distribution will become a continuous distribution when applying the Gamma function on it. Therefore, the calculation should change from summation to integration. However, for the simplicity and consistency of the calculation, we still use the discrete form to estimate the probability as defined in equation (4). It should be noted that if all the gene weights equal to one, the weighted algorithm will reduce to the conventional algorithms as shown in equation (1).

Traditional methods (5,6) always take the scores into ranks or round the scores into integers to

prevent violation of discreteness, while it scaling up some of N , m , n , and a , which will also cause deviations from the original values of probability because of the scaling but not weighting. For example, if we scale up all genes to have a score of 2 rather than 1 (count), the expected result should have exactly the same p-value but it is not.

Score normalization

To keep the consistency between the conventional methods and the weighted method, normalization is needed to keep all the scores have the mean value of one and greater than zero as well, which is defined as follows:

$$t_i = \frac{w_i - \min(W)}{\max(W) - \min(W)} \quad (5)$$

$$w'_i = t_i - \bar{T} + 1 \quad (6)$$

where w_i represents the original weight of the i^{th} gene, $W = [w_1, w_2, \dots, w_N]$ is the original weight matrix, t_i denotes the pre-normalized weight of gene i , $T = [t_1, t_2, \dots, t_N]$ is the pre-normalized weight matrix, $\bar{T}$ represents the mean value of T , w'_i is the normalized weights. As a result, the normalized weights have the mean value of one, which makes the weighted methods comparable to unweighted one.

Inverse Document Frequency score for distinct categories

Given a large number of gene sets, the frequency of one gene turns up in different sets has a very big difference. Thus, there can be a very intuitive hypothesis: a gene frequently involved in various gene functional terms may cause less effect on individual sets. In natural language processing (NLP), there is a popular concept and algorithm named Inverse Document Frequency (IDF) which is used to evaluate the importance of a word among documents. That is, the more one word shows in other documents, the less important this word will be in any document. Hence, the calculation of the IDF of one gene can be defined as follows:

$$\text{IDF}(g_i) = \log\left(\frac{N}{n_i} + 1\right) \quad (7)$$

where N is the total number of terms in one category (refer to the concept of corpus in NLP), n_i is the count of different terms (refer to the concept of document in NLP) where gene i is included.

Log proportion normalization

IDF calculation works well for gene sets but not for gene expression profiles among tissues.

And, our purpose is trying to extract the tissue-specific genes in a reasonable way. Hence, to evaluate the importance of each gene in different tissues, a log proportion normalization of expression profile is applied, which defined as follows:

$$s_{ij} = \frac{\log(m_{i,j})}{\sum_j \log(m_{i,j})} \quad (8)$$

where $m_{i,j}$ represents the gene counts of the i^{th} gene in the j^{th} tissue. Therefore, s_{ij} can be interpreted as the importance score of a specific gene i in different tissues.

Gene sets and gene essentiality scores

For gene sets, we collected some common gene annotation sets for functional enrichment analysis including Kyoto Encyclopedia of Genes and Genomes (KEGG, <https://www.kegg.jp/>), Gene Ontology (GO) Resource (7), and gene sets in Enrichr (8). In addition, we collected and processed five groups of gene essentiality scores including evolutionary conservation, expression profiles of specific tissues, gene importance score, protein-protein interaction (PPI) score and IDF of gene set as shown in Table 1.

Web server construction

The web server is built using the Django framework as the back-end for processing and calculation. For the front-end, we are using Bootstrap for web framework design, bootstrap-table for table illustration, Plotly.js for figure visualization, and JQuery for application logic. All the algorithms are implemented in Python using packages Scipy, Numpy, and Pandas. The web server is available at <https://www.cuilab.cn/weat/>.

RESULTS

Overall design and Web server

The basic workflow of weighted gene functional enrichment analysis (WEAT) is shown in Figure 1. In the knowledge part, WEAT collected and curated 200 gene functional categories (140 for Homo Sapiens), 471,296 gene sets (315,285 for Homo Sapiens) and 654 types of gene essentiality scores (563 for Homo Sapiens). Then, the user needs to input a gene list, for example, the deregulated genes from a study of omics. Next, WEAT will calculate the p-value for both the conventional gene functional enrichment analysis methods and the weighted gene functional enrichment analysis method. After that, a comparison will be performed between the results of the two methods. Meanwhile, the gene sets which are only significant in the

weighted methods will be highlighted.

As shown in Figure S1-2, the web interface of WEAT is designed following the above workflow. Firstly, users should select the species and then input a gene list. And then users need to choose a type of gene essentiality score. Meanwhile, WEAT also provides a choice for users to upload their own gene essentiality scores. (Figure S1). Next, after choosing the scale factor Q , users can click the “Run” button to submit the task. On the brief result page (Figure S2), users can perform analyses for any category of gene sets. Meanwhile, users can search, sort, export, and visualize their results.

Case Studies

The Prion pathway is involved in the aging process in human brain

It is known that one common feature of the ageing brain is the age-associated periventricular white matter lesions (PVL), which reflects white matter injury. It is thus important to dissect the functional pathways or terms involved in PVL. For doing so, we applied WEAT to a gene expression dataset of brain PVL caused by aging (GEO: GSE157363) (9). By using the instructions of the dataset, we identified 1715 differentially expressed genes ($p\text{-value} \leq 0.05$, $|\log_2(\text{fold-change})| \geq 1.2$). Next, we selected the brain gene expression in GTEx as the gene essentiality score. The results on the BioCarta pathway category are shown in Figure 2 (Table S1 for detail information).

It should be noted that the Prion pathway shows the greatest statistical significance in the weighted method while is not significant in the conventional ones. There are only two hit genes (GFAP and DNAJB6) between the input gene list and the Prion pathway gene set (Figure 3). Therefore, the conventional methods failed to identify the Prion pathway as significance, whereas the GFAP gene showed a very high normalized gene essentiality score (7.81) in brain tissue, resulting in the greatest significance of this pathway. The conventional methods cannot identify this term because they treated GFAP and other genes equally. In fact, GFAP, the glial fibrillary acidic protein, is highly related to brain and neural function. It was reported that the cellular prion protein (PrP) is greatly associated with Alzheimer's and Parkinson's diseases, which are indeed highly related with brain aging and lesions (10,11). This finding supported that WEAT can uncover some important patterns missed by the conventional methods.

Lung Squamous Cell Carcinoma

Lung Squamous Cell Carcinoma (LUSC) is the most common histologic subtype of non-small cell lung cancer (NSCLC), highly associated with cigarette smoking (12), with poor clinical prognosis (13). Exploring the functional pathways involved in LUSC is important for understanding its mechanisms and finding new therapeutic targets. Here we downloaded the LUSC gene expression data from The Cancer Genome Atlas (TCGA) (14) including 502 tumors and 49 tumor adjacent tissues. Using the edgeR (15) package, we identified 773 differentially expressed genes. For comparison, we used different gene essentiality scores and focused on the KEGG pathways. The novel terms identified by WEAT are shown in Figure 4 (Table S2 for detail information).

As a result, we found that different gene essentiality scores have different preferences for terms without loss of plausibility. For example, there was a patient reported as LUSC with severe cardiac dysfunction similar to dilated cardiomyopathy (16) which was successfully identified by WEAT using the phastCons score. Similarly, a case with lung cancer was found to be related to the dilated phase of hypertrophic cardiomyopathy (17) which was identified by WEAT using the gene importance calculator (GIC) score. Other terms including the disorder of DNA replication (18) and cytoskeleton proteins (19) using the phastCons scores, focal adhesion (20) and ABC transporters (21) using the GIC scores, are indeed the basal pathways of cancer that were not identified by the conventional methods.

Different scores prefer different terms

As discussed above, different gene essentiality scores have different tendencies on identifying functional terms. Thus, to further investigate the preference of the gene essentiality scores, we calculated the variance of gene essentiality scores for each functional term in the GO biological process (GOBP) as an example. Given that if the gene essentiality scores in a gene set could be greatly different from each other, this set may play a more important role under the situation of the selected score, here we only focus on the terms with greatest variance (Figure S3-5).

As a result, different essentiality scores indeed produce preferences to specific functional terms. As shown in Figure S3, for essentiality scores derived from GTEx gene expression, all the tissue expressed based scores prefer to essential functional terms such as hydrogen

peroxide biosynthetic process and electron or proton transport (right part of Figure S3). Moreover, each tissue also prefers to tissue-specific functional terms (left part of Figure S3). For example, the essential score derived from the blood prefer to gas-related functions such as nitric oxide transport and oxygen transport. Meanwhile, WEAT also provided normalized GTEx essential scores. The normalization of the GTEx essential scores will eliminate the common functional terms and focus on the tissue-specific terms (Figure S4). For essentiality scores of PPI degree and PPI scores, they are very similar and have a higher preference to the functional terms that the hub genes are involved in. The conservation scores including phastCons and phyloP show preference to the evolutionary conservative pathways such as embryonic development functions; The GIC score prefers to functional terms related to essential biological processes; The IDF score are more likely to identify functional terms with unique genes. In all, different essentiality scores can make the enrichment result subtle and hence the weighted method is able to identify different reasonable functional terms by using different scores.

DISCUSSION

Gene functional enrichment analysis is one of the most popular ways for annotating the functions of a given gene list. However, the genes in one input gene list and one gene set are equally treated for both hypergeometric test and binomial test, which are the basic algorithms for the conventional methods. This would lead to biased results because genes are actually different in essentiality.

In this study, we introduced a weighted gene functional enrichment analysis method, WEAT, by considering the above truth that different genes have different essentiality in one biological process. We further implemented the WEAT algorithm, developed the online tool, and confirmed its usefulness by two case studies. As a result, WEAT shows its capacity in exploring important gene functional terms which the conventional methods failed to dissect.

WEAT could be extended to various applications. For example, if we choose a gene expression essentiality score different from the tissue of the input case, we may find some functional terms involved in the crosstalk between the two tissues. Hence, this provides a potential way to find inter-organ interactions. Moreover, the scale factor Q is positive in the current WEAT algorithm, which means higher gene essentiality leads to greater weight in a

gene functional term. While it can be negative, which could be useful to explore the trivial functional terms when using IDF scores. In addition, WEAT could be improved in a number of aspects, for example, more gene sets and gene essentiality scores in the knowledge base, and more species will be annotated in the future. Finally, the profile-level two-phenotype gene set enrichment analysis could be improved using the idea of WEAT.

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analyzed during the current study are available on our website at <https://www.cuilab.cn/weat/>. The datasets of case studies are available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157363> and <https://portal.gdc.cancer.gov/projects/TCGA-LUSC>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the grants from the National Key R&D Program (2020YFC2004704), PKU-Baidu Fund (2019BD014), the Natural Science Foundation of China (81970440/62025102/81921001/82025008), and Peking University Basic Research Program (BMU2020JC001).

Authors' contributions

Q.C. conceived the project. R.F. collected the data, designed the algorithm, performed the analyses and built the web application. R.F. and Q.C. wrote and edited the manuscript and approved the final manuscript.

REFERENCES

1. Khatri, P., Sirota, M. and Butte, A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, **8**, e1002375.
2. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, **37**, 1-13.
3. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, **4**, 44-57.
4. Rivals, I., Personnaz, L., Taing, L. and Potier, M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401-407.
5. Qureshi, R. and Sacan, A. (2013) Weighted set enrichment of gene expression data. *BMC Syst Biol*, **7 Suppl 4**, S10.
6. Stojmirovic, A. and Yu, Y.K. (2010) Robust and accurate data enrichment statistics via distribution function of sum of weights. *Bioinformatics*, **26**, 2752-2759.
7. Gene Ontology, C. (2021) The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res*, **49**, D325-D334.
8. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A. *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, **44**, W90-97.
9. Fadul, M.M., Heath, P.R., Cooper-Knock, J., Kurz, J.M., Al-Azzawi, H.A., Ali, Z., Smith, T., Matthews, F.E., Brayne, C., Wharton, S.B. *et al.* (2020) Transcriptomic Analysis of

Age-Associated Periventricular Lesions Reveals Dysregulation of the Immune Response. *Int J Mol Sci*, **21**.

10. Duyckaerts, C., Clavaguera, F. and Potier, M.C. (2019) The prion-like propagation hypothesis in Alzheimer's and Parkinson's disease. *Curr Opin Neurol*, **32**, 266-271.
11. Gasperini, L. and Legname, G. (2014) Prion protein and aging. *Front Cell Dev Biol*, **2**, 44.
12. Gandara, D.R., Hammerman, P.S., Sos, M.L., Lara, P.N., Jr. and Hirsch, F.R. (2015) Squamous cell lung cancer: from tumor genomics to cancer therapeutics. *Clin Cancer Res*, **21**, 2236-2243.
13. Tanoue, L.T. and Detterbeck, F.C. (2009) New TNM classification for non-small-cell lung cancer. *Expert Rev Anticancer Ther*, **9**, 413-423.
14. Cancer Genome Atlas Research, N. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, **489**, 519-525.
15. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139-140.
16. Ogino, H., Nishimura, N., Kitamura, A., Ishikawa, G., Okafuji, K., Tomishima, Y., Jinta, T., Yamazoe, M., Yang, Y. and Chohnabayashi, N. (2014) A patient with lung squamous cell carcinoma presenting with severe cardiac dysfunction similar to dilated cardiomyopathy with left bundle branch block induced by myocardial metastasis. *Intern Med*, **53**, 2353-2357.
17. Sakai, T., Tsushima, T., Kimura, D., Hatanaka, R., Sawada, M. and Sato, Y. (2017)

[Primary Lung Cancer Associated with Dilated Phase of Hypertrophic Cardiomyopathy; Report of a Case]. *Kyobu Geka*, **70**, 147-150.

18. Macheret, M. and Halazonetis, T.D. (2015) DNA replication stress as a hallmark of cancer. *Annu Rev Pathol*, **10**, 425-448.
19. Hall, A. (2009) The cytoskeleton and cancer. *Cancer Metastasis Rev*, **28**, 5-14.
20. Eke, I. and Cordes, N. (2015) Focal adhesion signaling and therapy resistance in cancer. *Semin Cancer Biol*, **31**, 65-75.
21. Domenichini, A., Adamska, A. and Falasca, M. (2019) ABC transporters as cancer drivers: Potential functions in cancer development. *Biochim Biophys Acta Gen Subj*, **1863**, 52-60.
22. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, **15**, 1034-1050.
23. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, **20**, 110-121.
24. Zeng, P., Chen, J., Meng, Y., Zhou, Y., Yang, J. and Cui, Q. (2018) Defining Essentiality Score of Protein-Coding Genes and Long Noncoding RNAs. *Front Genet*, **9**, 380.
25. Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M. *et al.* (2017) Defining a Cancer Dependency Map. *Cell*, **170**, 564-576 e516.

26. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*, **47**, D607-D613.
27. Consortium, G.T. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, **45**, 580-585.
28. Fagerberg, L., Hallstrom, B.M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpour, S., Danielsson, A., Edlund, K. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics*, **13**, 397-406.
29. Szabo, L., Morey, R., Palpant, N.J., Wang, P.L., Afari, N., Jiang, C., Parast, M.M., Murry, C.E., Laurent, L.C. and Salzman, J. (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol*, **16**, 126.
30. Duff, M.O., Olson, S., Wei, X., Garrett, S.C., Osman, A., Bolisetty, M., Plocik, A., Celniker, S.E. and Graveley, B.R. (2015) Genome-wide identification of zero nucleotide recursive splicing in *Drosophila*. *Nature*, **521**, 376-379.

Table 1. List of gene essentiality scores used in WEAT.

Class	Name	Preprocess	Source	Description
Conservation	phastCons	Average	UCSC	Gene evolutionary conservation score calculated by phastCons (22).
	phyloP	Average	UCSC	Gene evolutionary conservation score calculated by phyloP (23)
Gene Importance	GIC	None	GIC	A score calculated by Gene Importance Calculator (24).
	DepMap_lineage	Average	DepMap	A score describing gene essentiality provided by DepMap (25) grouped by lineages.
	DepMap_disease	Average	DepMap	A score describing gene essentiality provided by DepMap (25) grouped by diseases.
IDF	[category]	IDF	Multiple	A score can describe the gene specificity of a category.
PPI	PPI_Degree	Count	STRING	Degree count of each gene node in PPI (26) graph.
	PPI_Score	Sum	STRING	Sum edge weights of each gene node in PPI (26) graph.
Expression	GTEEx_smts	None	GTEEx V8	GTEEx (27) expression profile grouped by tissue type, the area from which the tissue sample was taken.
	GTEEx_smtsd	None	GTEEx V8	GTEEx (27) expression profile grouped by more specific detail of tissue type
	PRJEB2445	None	BioProject	RNA-seq of individual and mixture of 16 human tissues.
	PRJEB4337	None	BioProject	RNA-seq of 95 human individuals representing 27 different tissues (28).
	PRJNA270632	None	BioProject	RNA-seq of 35 human fetal samples from 6 tissues (29).
	PRJNA280600	None	BioProject	RNA-seq of 20 human tissues (30).
	[project]_log	Log proportion		The scores of all described expression profiles above after log proportion normalization.

GIC, Gene Importance Calculator; PPI, Protein-Protein Interaction; IDF, Inverse Document

Frequency; [category], all the existing gene sets in our software; [project], all the other score names under the Expression class.

Figure Legends

Figure 1. The framework of the weighted gene functional enrichment analysis.

Figure 2. WEAT analysis results of the deregulated genes in age-associated periventricular white matter lesions (PVL) in the BioCarta signaling pathway category using GTEx brain essentiality score. Rich Factor = (sum score of hit genes) / (sum score of all genes in this set)

Figure 3. The normalized gene essentiality scores of genes in the Prion pathway by the WEAT analysis of deregulated genes in age-associated periventricular white matter lesions (PVL). The hit genes are shown in orange and the missed genes are shown in green.

Figure 4. The KEGG pathways identified only by WEAT analysis results but not by the conventional methods for the deregulated genes in Lung Squamous Cell Carcinoma (LUSC) using different gene essentiality scores (phastCons score and GIC score). Rich Factor = (sum score of hit genes) / (sum score of all genes in this set)

WEAT Results in BioCarta using GTEx Brain score

WEAT Result in KEGG using phastCons and GIC scores

WEAT: Weighted Enrichment Analysis Tools

[Home](#) [About Us](#) [Services](#) [Contact Us](#)

Figure 1

[illegible]

Step 2: Enter Game List (Game Symbols, Assembly ID or PARTS ID)

TITRE
 TITRE
 NCM
 LEAF2
 SAS12
 DATE MON
 DATE MON
 LACTE


```

4: isoperidine
5: isoperidine Oltis: none
6: isoperidine Oltis: none log
7: isoperidine Oltis: sorted
8: isoperidine Oltis: sorted log
9: isoperidine R000444
10: isoperidine R000444: log
11: isoperidine R000444:
12: isoperidine R000444: log
13: isoperidine R000444:

```

- Adipose tissue
- Arterial blood
- Blood
- Blood vessel
- Blood
- Bone**
- Bone
- Connective tissue
- Endothelium
- Epithelium

Step 4: Enter Scoring Section for Game Quantifiability Score

100

WEAT: Weighted Enrichment Analysis Tools

You have submitted 1467 gene(s) to have enrichment analysis with
scores/Expression/CTC₅₀ants/3radn and scaling factor 3.

					Search
☐	Database	Counts	All	Coverage (%)	Detail
GO_biological_process	1125	17563	76.55

KEGGpath	674	28175	49.18

Human_Protein-Protein_Web_Site	323	3086	18.25

Enrich	176	14848	67.15

WikiPathways_human	408	6201	32.98

Reactome	125	1548	8.15

GO_molecular_function	1161	12159	79.41

NCI ontolite component	1171	18504	63.30