Journal of Bioinformatics and Computational Biology
Vol. 17, No. 4 (2019) 1950024 (12 pages)

DOI: 10.1142/S0219720019500240

GEREDB: Gene expression regulation database curated by mining abstracts from literature

Tinghua Huang*, Xiali Huang†, Bomei Shi‡ and Min Yao§

College of Animal Science, Yangtze University
Jingzhou, Hubei 434025, P. R. China
**thua45@yangtzu.edu.cn*
†*huangxiali36@126.com*
‡*shibomei1991@sina.cn*
§*minyao@yangtzu.edu.cn*

Received 4 December 2018
Revised 17 April 2019
Accepted 18 April 2019
Published 29 August 2019

Understanding how genes are expressed and regulated in different biological processes are fundamental and challenging issues. Considerable progress has been made in studying the relationship between the expression and regulation of human genes. However, it is difficult to use these resources productively to analyze gene expression data. GEREDB (www.thua45.cn/geredb) has been developed to facilitate analyses that will provide insights into the regulation of genes that govern specific biological responses. GEREDB is a publicly available, manually curated biological database that stores the data regarding relationships between expression and regulation of human genes. To date, more than 39,000 Links have been contextually annotated by reviewing more than 53,000 abstracts. GEREDB can be searched using the official NCBI gene symbol as a query, and it can be downloaded along with the GEREA software package. GEREDB has the ability to analyze user-supplied gene expression data in a causal analysis oriented manner using the GEREA bioinformatics tool.

Keywords: Gene expression; regulator; causal analysis.

1. Background

In the last decade, significant progress has been made in understanding human gene regulation systems, including the innate immune system, metabolic systems, signal transduction, and other aspects.[1,2] Numerous important gene expression regulators, including tumor necrosis factor (TNF), interferon gamma (IFNG), and mitogen-activated protein kinase (MAPK), have been discovered.[3–5] Despite these efforts, many issues remain unanswered, including the methods to use to discover how key regulators initiate distinct responses to particular stimulations.[6,7]

It is becoming increasingly clear that the gene expression response involves complex networks and is under the control of several important positive and negative effectors, which influence the composition of the transcriptome.[8,9] Several state-of-the-art databases have been developed to address this problem, including TFactS,[10,11] HTRIdb,[12] and TRRUST.[13]

In this paper, we present a database (Gene Expression Regulation Database, GEREDB) to manually mine information regarding regulation of gene expression from PubMed abstracts. The database will enable biologists to explore their data in a causal analysis oriented manner.

2. Construction and Content

2.1. *Training dataset preparation and word feature extraction*

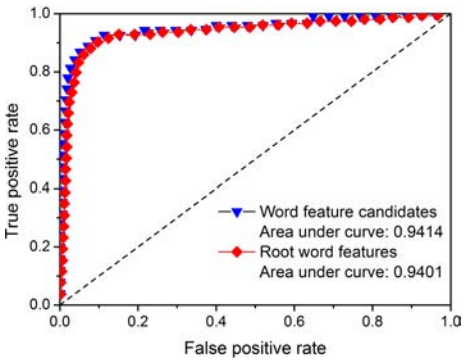
Abstract retrieval relied on the NCBI Entrez Programming Utilities search engine (E-utilities, Esearch, and Efetch) to filter related articles.[14] However, the results from the NCBI search engine had noise that significantly reduced the text-mining performance. We first filtered approximately 28 million abstracts from the Medline 2017 database using the query formulation "(gene regulation[MeSH Terms]) AND (human[MeSH Terms])" to retrieve human gene expression regulation studies.[15] This returned 461,555 abstracts. From these, we extracted 105,210 candidate sentences, which needed to contain at least two gene names based on definitions in the NCBI Gene database.[16] The candidate sentences were subjected to a manual classification procedure to establish gold-standard sets of positive and negative candidate sentences. Sentences describing the gene expression and regulation relationship were designated as "real:1" class, while others were designated as "pseudo:0" class. As a result, a training dataset containing 20,000 real sentences and 20,000 pseudo sentences were created. The word feature selection is based on Chi-square statistics,[17] and the top 2000 word features were manually inspected for their importance for classification. A total of 127 candidate word features were selected. Stemming and synonyms analysis assigned these candidate word features to 46 root word features (Table 1).

2.2. *Weka model training and testing*

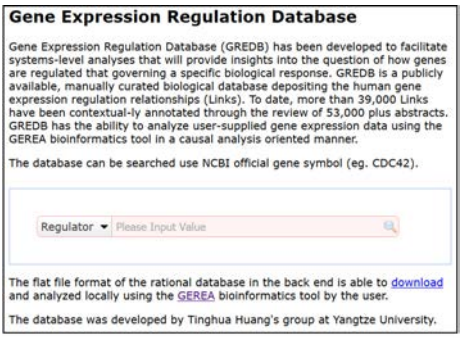
The sentences in the training datasets were converted into a set of numeric attributes termed the score datasets, representing the occurrence (value 1) or nonoccurrence (value 0) of word features presented in Table 1. The score datasets were split into 10 subsets, for training and cross validation. Each subset included 2000 positive samples and 2000 negative samples selected from the score dataset by a random procedure. The datasets were used for the J48 model training in WEKA 3.8.[18] A 10-fold cross validation technique was used for the selection of the most suitable model. The training of the model yielded an accuracy rate of 94.5% (-C 0.25 –M 2). The receiver operating curve cure analysis of the model showed that the area under the curve was

Table 1. Word features extracted from training datasets.

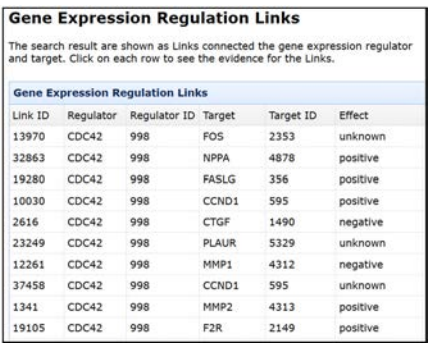
Feature	Score	Member
express	1.42E+08	express; expressed; expressing; expression; Expression; expressions
induce	2.94E+07	induce; induced; induces; inducible; inducing; induction
inhibit	2.59E+07	inhibit; inhibited; inhibiting; inhibition; Inhibition; inhibitor; inhibitors; inhibitory; inhibits
activate	2.56E+07	activate; activated; activates; activating; activation; Activation; activator
produce	2.19E+07	produced; production
transcript	1.96E+07	transcript; transcription; transcriptional; transcripts; transduction
increase	1.75E+07	increase; increased; increases; increasing
upregulate	1.50E+07	upregulate; up-regulate; upregulated; up-regulated; upregulates; up-regulates; upregulation; up-regulation
promoter	1.45E+07	promoter; promoters
regulate	1.43E+07	regulate; regulated; regulates; regulating; regulation
stimulate	1.09E+07	stimulate; stimulated; stimulates; stimulating; stimulation; Stimulation; stimulatory; stimuli; stimulus
enhance	1.05E+07	enhance; enhanced; enhancer; enhances; enhancing
suppress	8.84E+06	suppress; suppressed; suppresses; suppressing; suppression; suppressor
signal	7.13E+06	signaling
downregulate	6.88E+06	downregulated; down-regulated; down-regulates; downregulation; down-regulation
overexpress	6.65E+06	overexpression; Overexpression; over-expression
decrease	5.20E+06	decreased; decreases
reduce	5.07E+06	reduce; reduced; reduces
less	3.40E+06	less
significant	3.36E+06	significantly
mediate	3.33E+06	mediated
repress	3.16E+06	repressed; represses; repression
active	3.11E+06	active; activin; activities; activity
transactivate	2.78E+06	transactivation
more	2.56E+06	more
block	2.56E+06	blocked
attenuate	2.00E+06	attenuated
modulate	1.99E+06	modulated; modulates; modulating
promote	1.69E+06	promotes; promoting
greater	1.45E+06	greater
prevent	1.27E+06	prevented
change	1.06E+06	changes
negative	8.20E+05	negative; negatively
low	7.32E+05	low
transfect	5.61E+05	Transfection
converse	5.46E+05	Conversely
depress	4.32E+05	depression
reverse	3.47E+05	reversed
direct	2.48E+05	directed; directly
indirect	2.09E+05	indirect
inactivate	1.63E+05	inactivation
alter	1.48E+05	altered
effect	5.79E+04	effects
elicit	4.32E+04	elicited
positive	2.27E+04	positive; positively
affect	1.12E+04	affected



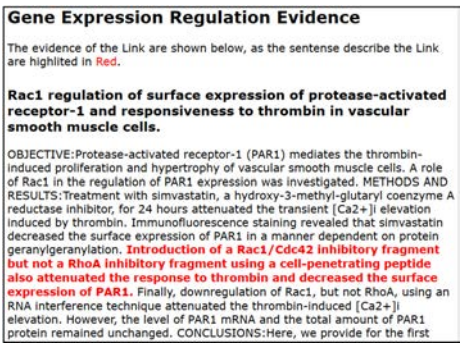
(a)



(b)



(c)



(d)

Fig. 1. Receiver operating characteristics (ROC) curve and web database screen shots. (a) The ROC curve of Weka J48 model training. (b) The GEREDB search page. (c) The Links search results page. (d) The Evidence page.

approximately 0.94 (Fig. 1(a)). There was no significant difference for accuracy models built on the root word features or all of the word feature candidates (Fig. 1(a)). These results indicated that the model had good performance in distinguishing between real and pseudo gene expression regulation sentences.

2.3. *Manual curation of gene expression regulation relationships*

We then filtered more than approximately 28 million abstracts from the Medline 2017 database using the query formulation "human[MeSH Terms]" to consider all of the human biological studies.[15] The approximately 12 million returned abstracts were first prioritized with the model (a total of 53,229 sentences passed the threshold) and then were subjected to manual extraction of gene expression regulation relationship by biologists. We used simplified gene expression regulator–target Links to represent the relationship between gene expression and regulation. The standard definition of "gene expression regulator" that we provide in the Links is a gene that

can change the expression of a target gene. The Links include three major functional elements — the gene expression regulator (regulon), the target gene, and the direct link connecting the regulon and the target.

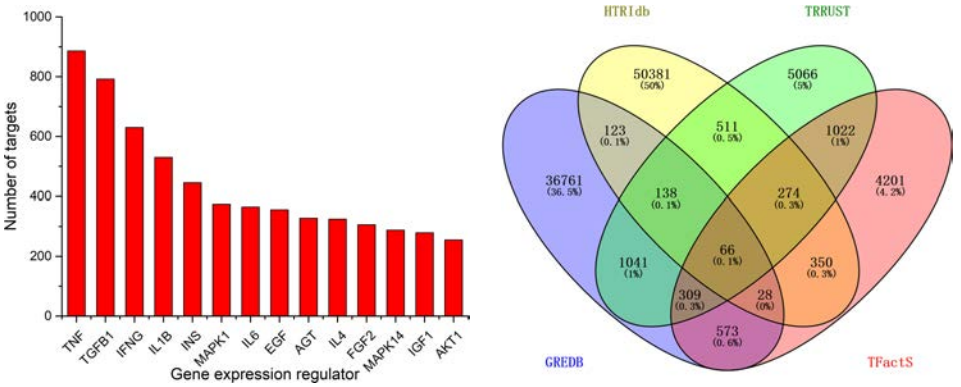
Finally, a total of 39,939 gene expression regulation interactions were curated, richly annotating them in terms of the evidence they provided and the context in which they occurred in the abstracts. The importance of manual curation is clear, the detailed manual curation permitted us to richly annotate these Links and to place them in their relevant context. This contextual annotation includes details of the supporting publication, genes studied, the species, the effect of regulation, and several other fields.

2.4. *Development of the GEREDB web database*

A flexible web-based interface (www.thua45.cn/geredb) allows simple searching of GEREDB. This interface has been developed in close collaboration with our biologists to ensure that the interface is easy to use. Moreover, the flat file format of the rational database in the back end can be downloaded and analyzed locally by the user. On the GEREDB search page, one has an option of searching the Links for "Regulator" or "Target" (Fig. 1(b)). As the data in GEREDB are organized according to genes, the Links search allows one to retrieve the data of interest using an NCBI official gene symbol (gene name). Genes are frequently known by several different names or symbols, which are termed synonyms. These synonyms of the genes stored in GEREDB are mapped to Entrez Gene name alias to avoid any ambiguous search terms. From the Links search results page (Fig. 1(c)), information related to the Links of interest can be obtained, including the official gene symbol and gene identity of the regulators and targets, and the effect of the regulator on the target in each Link. The users can also use a program to access the search the result page by passing the URL parameters of "gene_symble" and "search_for". From the Evidence page (Fig. 1(d)), details of the evidences supporting the Link, including the original abstract and the sentence supporting the Link (highlighted in red), can be obtained.

2.5. *The GEREDB database is a unique resource for gene expression regulation*

Analysis of the links in the GEREDB database identified an abundance of a set of over-represented sub-networks known as network motifs.[19] Figure 2(a) shows the regulators of the top 14 network motifs with the highest number of targets. These network motifs represent our current understanding of which motif is the most important "regulator" for gene expression when dividing a big network into small blocks. For example, the TNF can regulate a total of 886 genes in GEREDB database, while TGFB1 can regulate a total of 792 genes. More than 55 regulators in GEREDB can regulate more than 100 genes, suggesting their important role in transcriptome regulation. However, GEREDB is not the only database in the



(a) (b)

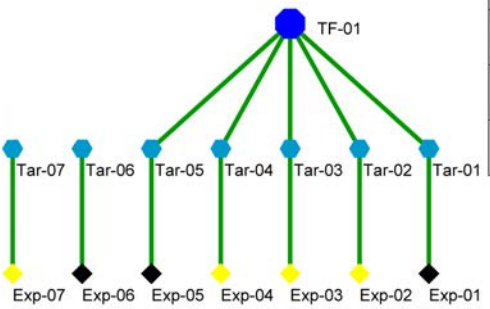
Fig. 2. Summary of the Links deposited in GEREDB. (a) Regulators of the top 14 network motifs with the highest number of targets. (b) Comparison of numbers of Links deposited in GEREDB with HIRIdb, TRRUST, and TFactS.

community. HIRIdb,[12] TRRUST,[13] and TFactS[10,11] also have curated gene expression regulation relationships for transcription factors. To date, a total of 51,871, 8427, and 6823 Links have been deposited in HIRIdb, TRRUST, and TFactS, respectively. A comparison of the number of Links deposited in GEREDB with HIRIdb, TRRUST, and TFactS showed that there were 355 overlaps with HIRIdb, 1554 overlaps with TRRUST, and 976 overlaps with TFactS. Interestingly, 94% of the Links deposited in GEREDB did not overlap with HIRIdb, TRRUST, and TFactS, indicating that GEREDB is a unique resource for gene expression regulation relationships in the community. The numbers of Links deposited in GEREDB, HIRIdb, TRRUST, and TFactS were shown in Fig. 2(b).

3. Utility and Discussion

3.1. *Development of the GEREA analyzing tool*

Building on this data, the GEREA bioinformatics tool was developed to discover gene expression regulators that will facilitate casual inference of the gene expression data. Here, we define the active regulator as a gene that has targets that occur in lists of differentially expressed genes with significantly higher frequencies than expected. After the regulator–target regulation Links were built, the gene expression profiling data was loaded on the targets of the regulators (Fig. 3). The network-based data could be organized as a two-by-two contingency table, and an expected ratio in the background (P0) and an observed ratio (P1) in the list of differentially expressed genes were calculated (Fig. 3). If P1 was significantly higher than P0, then we considered the targets of the regulator to be significantly enriched in the differentially expressed genes, indicating that the targets were regulated by the regulator.



	Differentially expressed	Not changed	Row total
Regulator targets	a (x)	b	a+b (s)
Not regulator targets	c	d	c+d
Column total	a+c (k)	b+d	a+b+c+d (n)

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

$$P0 = (a+b)/(a+b+c+d)$$

$$P1 = (a)/(a+c)$$

Fig. 3. Schema of the GEREA bioinformatics tool. TF-01 represents the gene expression regulator, Tar-01 to Tar-05 represent the five targets of TF-01, while Tar-06 and Tar-07 are non-TF-01 targets presented in the gene expression data.

Table 2. Significantly enriched gene expression regulators in LPS stimulation versus no stimulation control expression data.

Regulator name	Uploaded gene count	GEREDB targets count	Differentially expressed genes count	Differentially expressed targets count	Expected ratio (P0)	Observed ratio (P1)	FDR
TNF	2468	511	718	209	0.21	0.29	1.91E-10
IL1B	2468	275	718	119	0.11	0.17	3.69E-07
CD40	2468	69	718	42	0.03	0.06	6.62E-07
IFNG	2468	397	718	158	0.16	0.22	1.65E-06
TLR4	2468	74	718	41	0.03	0.06	4.80E-06
MAPK1	2468	207	718	90	0.08	0.13	1.18E-05
TLR2	2468	52	718	31	0.02	0.04	1.30E-05
TGFB1	2468	414	718	158	0.17	0.22	1.63E-05
IL10	2468	147	718	66	0.06	0.09	1.03E-04
IL1A	2468	93	718	46	0.04	0.06	1.15E-04
IL6	2468	200	718	84	0.08	0.12	1.49E-04
MAPK3	2468	108	718	51	0.04	0.07	1.58E-04
IFNB1	2468	92	718	45	0.04	0.06	2.04E-04
AGT	2468	165	718	74	0.07	0.10	3.41E-04
EDN1	2468	82	718	41	0.03	0.06	4.03E-04
TAC1	2468	38	718	22	0.02	0.03	6.42E-04
PTGS2	2468	50	718	27	0.02	0.04	6.61E-04
NOTCH1	2468	76	718	37	0.03	0.05	7.80E-04
VEGFA	2468	112	718	49	0.05	0.07	1.24E-03
HGF	2468	95	718	44	0.04	0.06	1.27E-03
OSM	2468	57	718	29	0.02	0.04	1.53E-03
PPARG	2468	132	718	56	0.05	0.08	1.54E-03
MAPK14	2468	166	718	68	0.07	0.09	1.63E-03
CD40LG	2468	56	718	30	0.02	0.04	1.76E-03
NFKBIA	2468	30	718	18	0.01	0.03	1.89E-03
TNFRSF1A	2468	33	718	19	0.01	0.03	1.89E-03

Table 2. (*Continued*)

Regulator name	Uploaded gene count	GEREDB targets count	Differentially expressed genes count	Differentially expressed targets count	Expected ratio (P0)	Observed ratio (P1)	FDR
TNFSF10	2468	38	718	21	0.02	0.03	1.99E-03
IL4	2468	206	718	81	0.08	0.11	2.44E-03
CCL2	2468	28	718	18	0.01	0.03	2.70E-03
REL	2468	24	718	15	0.01	0.02	2.77E-03
IL1RN	2468	21	718	14	0.01	0.02	2.80E-03
TNFRSF8	2468	6	718	6	0.00	0.01	2.82E-03
IL17A	2468	50	718	26	0.02	0.04	2.96E-03
LTA	2468	17	718	12	0.01	0.02	3.01E-03
ABL1	2468	24	718	17	0.01	0.02	3.64E-03
VIP	2468	50	718	25	0.02	0.03	3.81E-03
PPARA	2468	85	718	38	0.03	0.05	4.10E-03
CD4	2468	99	718	45	0.04	0.06	4.75E-03
PRL	2468	83	718	37	0.03	0.05	4.91E-03
FGF18	2468	11	718	9	0.00	0.01	5.07E-03
PTGER2	2468	14	718	10	0.01	0.01	5.85E-03
NFKB1	2468	16	718	11	0.01	0.02	6.05E-03
TLR7	2468	21	718	13	0.01	0.02	6.24E-03
MIF	2468	30	718	17	0.01	0.02	6.94E-03
F2	2468	90	718	40	0.04	0.06	7.97E-03
CD8A	2468	69	718	33	0.03	0.05	9.79E-03

Table 3. Significantly enriched gene expression regulators in cytokine TNF treatment versus no stimulation control expression data.

Regulator name	Uploaded gene count	GEREDB targets count	Differentially expressed genes count	Differentially expressed targets count	Expected ratio (P0)	Observed ratio (P1)	FDR
TLR4	2458	74	298	24	0.03	0.08	9.11E-06
IL13	2458	80	298	23	0.03	0.08	2.61E-04
TGFB1	2458	399	298	71	0.16	0.24	4.15E-04
IL1B	2458	274	298	54	0.11	0.18	4.26E-04
HSPD1	2458	22	298	10	0.01	0.03	8.71E-04
TLR2	2458	53	298	16	0.02	0.05	1.14E-03
TNFSF12	2458	17	298	8	0.01	0.03	1.63E-03
TNF	2458	499	298	82	0.20	0.28	1.66E-03
TLR3	2458	31	298	11	0.01	0.04	2.39E-03
IL4	2458	192	298	39	0.08	0.13	2.49E-03
RUNX2	2458	18	298	8	0.01	0.03	2.93E-03
ZC3H12A	2458	5	298	4	0.00	0.01	4.29E-03
NOX4	2458	8	298	5	0.00	0.02	6.31E-03
IL18	2458	37	298	12	0.02	0.04	6.39E-03
IL17A	2458	48	298	14	0.02	0.05	7.54E-03
KL	2458	5	298	4	0.00	0.01	7.98E-03
NME1	2458	12	298	6	0.00	0.02	8.51E-03
IL10	2458	145	298	30	0.06	0.10	8.61E-03

Statistical analysis was performed based on cumulative hypergeometric distribution[20,21] and corrected by the Benjamini and Hochberg method (FDR).[22]

3.2. *Case study: Testing GEREA with real gene expression datasets*

GEREA was tested on a published dataset of macrophage cells stimulated with lipopolysaccharide (LPS) or the TNF cytokine.[23] The raw transcriptome dataset was obtained from the NCBI GEO database with accession number GSE100382. The data were normalized using the quantile method in Bioconductor.[24] Comparisons were made between the no stimulation control and LPS stimulation or TNF cytokine treatment. The GEREA run took 62 s. The result indicated that targets of 46 regulators were significantly enriched in the genes that were differentially expressed between LPS stimulation versus no stimulation control (FDR < 0.01, Table 2). We also found targets of 18 regulators that were significantly enriched in the genes differentially expressed between TNF cytokine treatment versus no stimulation control (FDR < 0.01, Table 3).

4. Discussion

Different from GEREDB, over 95% of interactions (49,762 out of 51,871) deposited in HIRIdb[12] were derived from high throughput experiment such as deep sequencing or microarray. TRRUST[13] was built using similar technology with GEREDB, but the number of interactions curated was much fewer (8427 versus 39,938). TFactS[10,11] combined interactions from multiple resources, and the number of interactions was the fewest in the four databases. The regulation relationships (positive or negative) for HIRIdb and TFactS were not available. Each of those databases was build based on different rationale and have unique features.

While past studies have usually dealt with individual regulatory interactions, it has become clear that the only way to understand the regulatory activity of genes is to directly address the complex network nature of the whole ensemble of gene expression regulation elements involved in such a process.[25] Accurate gene regulator discovery plays an important role in studies of complex networks because these regulators provide a way to reveal the major effectors for expression patterns among multiple types of interactions. However, detecting gene regulators in complex networks remains challenging. Many methods exist to elucidate regulatory subnetworks, but complete information to explicitly connect expression data to gene expression regulators is still lacking.[26,27] Using the relationships between expression and regulation of human genes deposited in GEREDB and a regulator finding algorithm called GEREA, we are able to discover regulators that orchestrate a particular transcriptional profile.

5. Conclusions

GEREDB is a useful and convenient resource for investigators interested in analyzing human gene expression data. GEREDB is a unique resource in that it provides

comprehensive information of relationships between gene expression and regulation. Such data are typically difficult to obtain because of the lack of an effective method to extract them from the literature. The database is useful for exploring how differentially expressed genes detected by high throughput experiments are regulated by certain regulator genes. GEREDB can assist investigators in discovering relationships between gene expression and regulation that may lead to new hypotheses. GEREDB can be continually developed in the further, curating new abstract from human literatures as well as other organisms such as mouse. New modules should also be developed for better usage in applications such as (1) user-friendly web searching interface, (2) automatic network building and visualization, and (3) robust regulator enrichment statistical algorithms.

Acknowledgments

This project was funded by the National Natural Science Foundation of China (NSFC Grant No. 31402055), the College Students' Innovation and Entrepreneurship Training Program of Yangtze University (Grant No. 2018057), the Yangtze Youth Talents Fund (Grant No. 2015cqr12), the Yangtze Youth Fund (Grant No. 2015cqn39), and the Scientific Research Starting Foundation for the Returned Overseas Chinese Scholars of the Ministry of Education of China.

References

1. Wachter A, Gene regulation by structured mRNA elements, *Trends Genet* **30**:172–181, 2014.
2. Jones B, Gene expression: Layers of gene regulation, *Nat Rev Genet* **16**:128–129, 2015.
3. Croft M, Siegel RM, Beyond TNF: TNF superfamily cytokines as targets for the treatment of rheumatic diseases, *Nat Rev Rheumatol* **13**:217–233, 2017.
4. Green DS, Young HA, Valencia JC, Current prospects of type II interferon gamma signaling and autoimmunity, *J Biol Chem* **292**:13925–13933, 2017.
5. Arthur JS, Ley SC, Mitogen-activated protein kinases in innate immunity, *Nat Rev Immunol* **13**:679–692, 2013.
6. Subramanian A *et al.*, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proc Natl Acad Sci USA* **102**:15545–15550, 2005.
7. Shojaie A, Michailidis G, Network enrichment analysis in complex experiments, *Stat Appl Genet Mol Biol* **9**:Article22, 2010.
8. Zhao Y *et al.*, Network-based identification and prioritization of key regulators of coronary artery disease loci, *Arterioscler Thromb Vasc Biol* **36**:928–941, 2016.
9. Huang TH *et al.*, Distinct peripheral blood RNA responses to Salmonella in pigs differing in Salmonella shedding levels: Intersection of IFNG, TLR and miRNA pathways, *PLoS One* **6**:e28768, 2011.
10. Essaghir A, Demoulin JB, A minimal connected network of transcription factors regulated in human tumors and its application to the quest for universal cancer biomarkers, *PLoS One* **7**:e39666, 2012.
11. Essaghir A *et al.*, Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data, *Nucl Acids Res* **38**: e120, 2010.

12. Bovolenta LA, Acencio ML, Lemke N, HTRIdb: An open-access database for experimentally verified human transcriptional regulation interactions, *BMC Genomics* **13**:405, 2012.
13. Han H *et al.*, TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions, *Nucl Acids Res* **46**:D380–D386, 2018.
14. Coordinators NR, Database resources of the National Center for Biotechnology Information, *Nucl Acids Res*, 2017.
15. Fatehi F, Gray LC, Wootton R, How to improve your PubMed/MEDLINE searches: 3. advanced searching, MeSH and My NCBI, *J Telemed Telecare* **20**:102–112, 2014.
16. Brown GR *et al.*, Gene: A gene-centered information resource at NCBI, *Nucl Acids Res* **43**:D36–D42, 2015.
17. Jin C *et al.*, Chi-square statistics feature selection based on term frequency and distribution for text categorization, *IETE J Res* **61**:351–362, 2015.
18. Smith TC, Frank E, Introducing machine learning concepts with WEKA, *Methods Mol Biol* **1418**:353–378, 2016.
19. Boyle AP *et al.*, Comparative analysis of regulatory information and circuits across distant species, *Nature* **512**:453–456, 2014.
20. Fisher RA, On the interpretation of χ^2 from contingency tables, and the calculation of P, *J R Stat Soc* **85**:87–94, 1922.
21. Agresti A, [A Survey of Exact Inference for Contingency Tables]: Rejoinder, *Stat Sci* **7**:173–177, 1992.
22. Rivals I *et al.*, Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **23**:401–407, 2007.
23. Park SH *et al.*, Type I interferons and the cytokine TNF cooperatively reprogram the macrophage epigenome to promote inflammatory activation, *Nat Immunol* **18**:1104–1116, 2017.
24. Bolstad BM *et al.*, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, *Bioinformatics* **19**:185–193, 2003.
25. Kang T *et al.*, A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data, *BMC Bioinformatics* **18**:565, 2017.
26. Huang Y, Li S, Detection of characteristic sub pathway network for angiogenesis based on the comprehensive pathway network, *BMC Bioinformatics* **11**(1):S32, 2010.
27. Lee SA *et al.*, POINeT: Protein interactome with sub-network analysis and hub prioritization, *BMC Bioinformatics* **10**:114, 2009.

Tinghua Huang received his B.S. degree in Animal Science from Huazhong Agriculture University, Wuhan, China in 2004, and Ph.D. degree in Animal Genetics and Breeding from Huazhong Agriculture University, Wuhan, China in 2009. His research interests include genetic, bioinformatics, computational biology, algorithms, and computational learning. He is a heavy programmer and his works have been published in Journals including BMC bioinformatics, International Journal of Data Mining and Bioinformatics, Molecular Genetics and Genomics, Veterinary Research, and PLoS One. He has awarded the Yangtze Youth Talents Award 2014.

Xiali Huang received her B.S. degree in Animal Science from Yangtze University, she is currently studying for a M.S. degree in Microbiology at Yangtze University. Her research interests include algorithms, genetic, computational learning, microbiology, data mining, bioinformatics, and computational biology. Her works have been published in Journals including Canadian Journal of Animal Science, Research in Veterinary Science. She has awarded scholarships of the Yangtze University and the honor of outstanding graduates.

Bomei Shi received her B.S. degree in Veterinary Medicine from Henan Institute of Science and Technology, XinXiang, China in 2013, and M.S. degree in Basic Veterinary Medicine from GuangXi University, Nanning, China in 2016. She is currently an experimental technician at college of animal sciences, Yangtze University. Her researh interests include animal reproductive endocrinology, Salmonella infection in swine and cryopreservation technology of boar semen. Her works have been published in journals including Acta Agriculturae Boreali–Sinica, Guangxi Agricultural Sciences, Veterinary Science in China, and Canadian Journal of Animal Science.

Min Yao received her B.S. degree in Veterinary Medicine from Huazhong Agriculture University, Wuhan, China in 2004, and Ph.D. degree in Basic Veterinary Medicine from Huazhong Agriculture University, Wuhan, China in 2010. She is currently a lecturer at college of animal sciences, Yangtze University. Her research interests include animal science, Salmonella infection in swine, microRNA and bioinformatics. Her works have been published in journals including Research in Veterinary Science, Veterinary Research, PLoS One and Bioscience Report.