DOI:10.12113/j.issn.1672-5565.201710001

Enrich_analysis.pl,差异表达基因富集分析的 perl 脚本

李旭凯[1,2,3*],王钇杰[2,3,4],王俊杰[2,3,4]
(1.山西农业大学 生命科学学院,山西 太谷,030801;
2.特色杂粮种质资源发掘与遗传改良山西省重点实验室,山西 晋中,030801;
3.农业部黄土高原作物基因资源与种质创制重点实验室,太原,030031;
4.山西农业大学 农学院,山西 晋中 030801)

摘 要:如何从庞大的基因表达数据库里挖掘出有价值的信息,并做出科学的生物学诠释,是基因表达分析领域的重要挑战。富集分析为解决这一问题提出了合理的方案。用 Perl 语言写了一个脚本——Enrich_analysis.pl,可根据基因注释信息进行基因富集分析,并利用 Fisher's Exact Test 做检验。富集分析先根据生物学知识将基因归类,然后进行基因差异表达分析,提高数据的可解释性。并利用 3 篇 SCI 文章的差异基因数据,对本 perl 脚本进行了对比测试,证明本 perl 脚本算法正确可靠。

关键词:Enrich_analysis.pl;perl 脚本;富集分析;Fisher-s Exact Test;检验
中图分类号:Q811.4; R857.3 **文献标志码**:A **文章编号**:1672-5565(2018)03-178-06

Enrich_analysis.pl, a perl script for enrichment analysis of differentially expressed genes

LI Xukai[1,2,3*], WANG Yijie[2,3,4], WANG Junjie[2,3,4]
(1. *College of Life Science*, *Shanxi Agricultural University*, *Taigu* 030801, *Shanxi*, *China*;
2. *Shanxi Key Laboratory of Genetic Resources and Genetic Improvement of Minor Crops*, *Jinzhong* 030801, *Shanxi*, *China*;
3. *Key Laboratory of Crop Gene Resources and Germplasm Enhancement on Loess Plateau*, *Ministry of Agriculture*, *Taiyuan* 030031, *Shanxi*, *China*;
4. *College of Agriculture*, *Shanxi Agricultural University*, *Jinzhong* 030801, *Shanxi*, *China*)

Abstract: It is a major for challenge transcriptome data analysis to dig out research-related information and provide professional interpretation. Enrichment analysis offers a rational option. In this article, a perl script Enrich_analysis.pl was written for analyzing differentially expressed gene product annotations in biology, and the statistical test adopted Fisher's Exact Test. Enrichment analysis performed domain aggregation by combining gene expressions before testing for the differentially expressed. It helps investigators assign biological meaning to some group of genes. Moreover, using the data of differential expression gene from 3 SCI articles, the perl script was tested. The results proved Enrich analysis to be correct and reliable.

Keywords: Enrich_analysis.pl; Perl script; Enrichment analysis; Fisher's Exact Test; Statistical inferenc

在基因组和转录组迅速发展的今天,如何从庞大的基因表达数据库里挖掘出有价值的信息,并做出科学的诠释,是基因分析领域的重要挑战和关键问题。生物统计学家们利用 t 检验、比值法、固定效应模型、混合效应模型等多种统计分析方法,对筛选出的差异表达的单基因进行分析比较,推测和预测其可能的功能[1]。然而,单个基因分析在组间基因表达量差别较小、差异基因数量过大且没有统一分类信息时,容易出现假阴性和无法解释的现象。之后,生物学家们总结了单基因分析的缺陷,提出了以基因集(Gene set)为基础的基因富集分析法(Gene set enrichment analysis, GSEA)[1]。

从 Disease Ontology(DO), Gene Ontology(GO), Kyoto encyclopedia of genes and genomes(KEGG),到

收稿日期:2017-10-09;**修回日期**:2018-03-21.
基金项目:山西省主要农作物种质创新与分子育种重点科技创新平台(No.201605D151002);山西省优秀博士来晋工作奖励资金科研项目(No.SXYBKY201738).
* **通信作者**:李旭凯,男,讲师,研究方向:生物信息学;E-mail:xukai_li@sxau.edu.cn.

Reactome Pathway 等多种基因富集性分析方法不断被推出。科学家们借助各种生物信息学的数据库和分析工具进行统计分析, 对不同层次、不同来源的数据进行归纳与整合,力图找到目标基因与库中有相同或类似功能基因集之间的相关性[2-8]。具体是对一组基因(共表达或者差异表达)中某个功能类别的显著性利用超几何分布分析进行检验,这个功能类别可能就是导致样本性状差异的最重要的功能差别,其功能特征会阐明样本变化的内在生物学意义。常用的统计学方法主要是基于 Fisher test[2,3,7] 和 chi-square test[8] 检验。

目前做富集分析已经开发了很多软件工具:有 GFINDer, GOALIE, GOTM, L2L, Ontology Traverser, ProfCom, SeqExpress, STEM 等。但是绝大多数都是通过 web 和其对应的数据库中的数据进行分析,对于个人的或者特殊的物种无法分析。其中 GSEA 是最有名的软件,但是它基于芯片的分析软件,老版本对现今的 RNA-seq 数据不适用,最新的版本用起来并不是很方便。DAVID 可用 Gene ID 或 symbol 做 GO 分析,但是无法对其数据库中没有的基因 ID 进行分析。本文中介绍的 Practical Extraction and Report Language(perl)脚本可以利用研究者现有的基因表达差异和注释数据,进行分析。

1　数据库

基因集是先验的生物学知识的集合,即已被证实了的生物通道、基因共表达信息等。目前常用的定义基因集的基因注释数据库有 GO 和 KEGG 等。

GO 是基因本体联合会建立的数据库,旨在构建一个可以随着细胞中基因和蛋白质功能知识的累积和变化,进行统一描述的语义词汇标准。它是根据不同实验目的,筛选出差异表达基因,分析差异表达基因在 GO 数据库中的分布情况,从而诠释实验中样本差异在基因功能层面上的差异体现。GO 共有 3 层结构的定义方式:分子功能,生物过程和细胞组成[9]。在 GenBank 中,基因与其编码的蛋白都有着特殊的 ID 编号一一对应,也有能通过序列注释的方法找到与之对应的 GO 号[10],找到检索的全部分子功能所处的位置和关系信息。

KEGG 是整合了生物化学、基因组学以及系统功能组学的信息的基因功能系统分析知识库,这个数据库有助于科研工作者从分子通路水平来解释生物系统高层次功能的数据库。KEGG 不仅可以提供出色的整合代谢途径查询,还可以利用 Java 的图形工具访问基因组图谱,对基因组图谱、表达图谱和序列进行图形比较和通路计算。KEGG 是一个综合数据库,它们大致分为系统信息、基因组信息和化学信息三大类。进一步可细分为 16 个主要的数据库[11]。科学家们将属于同一 KEGG 基因通路(KO 分类)的基因定义为一个基因集。在生物体内,不同的生物学功能是通过不同基因间的相互协调互作而实现的,KEGG 能够通过比较通路富集的情况,推测差异表达基因可能参与的主要生化代谢和信号转导路径,从而推测其可能的路径数目及功能[12],这对寻找出不同样品的差异表达基因可能与哪些生物学通路的改变有关具有提示作用。

2　方　法

2.1　在 windows 操作系统下的安装 Perl

Perl 语言是拉里·沃尔(Larry Wall)于 1987 年 12 月 18 日发表的。它借鉴采用了 C、awk、sed、shell 等其他编程语言的特性以及优点。Perl 语言可以在 UNIX、Linux、MAC 和 Windows 等操作系统下跨系统平台运行,并由第三方代码库 CPAN[13](Comprehensive Perl Archive Network)不断更新和维护,它包含了极丰富 perl 写成的软件包和其源文件[14]。

在 Enrich_analysis.pl 脚本运行前,电脑上需事先安装一个可执行 perl 语言编辑与运行的解释器。一般的 UNIX(含 Linux 与 Mac OS)系统都含有 perl 解释器,因此在终端输入“perl-v”命令,就可以查看 perl 的版本。Windows 系统一般可与“ActivePerl”和“Strawberry Perl”兼容。ActivePerl[15] 是 ActiveState 公司开发的一个 Perl 语言运行环境,是一套面向各个平台、集成度高、易于使用的 Perl 脚本解释器。ActivePerl 目前最新版本为 5.26,可在 http://www.activestate.com/activeperl/downloads 页面下下载,其中“msi”文件是安装程序。安装完成之后,打开 cmd 窗口输入“perl-v”如有返回 perl 版本信息则表示安装成功[16]。

2.2　数据的格式

输入两个文件:第 1 个文件是各个基因对应的 GO 分类或 KEGG 分类;第 2 个文件是差异表达基因的名称。

Enrich_analysis.pl 要求输入两个数据文件。各物种 GO 注释数据可以通过 Blast2GO 获得,或者从 http://geneontology.org/page/download-annotations 与 http://bioinfo.cau.edu.cn/agriGO/download.php 网页下载(KEGG 数据目前收费),之后用 Excel 打开,删除掉多余信息,只留下两行,第一行为基因编号或基因名,第二列是对应的注释数据。数据结构见表 1。

表 1 基因分类数据格式

Table 1 Data format of gene classification

基因编号	注释信息
Gene_name_1	Annotation_1
Gene_name_2	Annotation_2
Gene_name_3	Annotation_3
…	…
Gene_name_n	Annotation_m

差异基因数据文件只有一列:差异基因的基因编号或基因名。

将 Excel 表另存为“文本文件(用制表符“Tab”分割)(*.txt)”类型的文本文件,例如 annotation.txt 和 DEGlist.txt。将这两个数据文件与 Enrich_analysis.pl 存放在同一个目录下,如 C 盘根目录下。

2.3 Enrich_analysis.pl 的运行

通过点击任务栏上的“开始-运行”,可以运行存放在数据目录下的 Enrich_analysis.pl 文件。在运行框中输入“cmd”,打开“cmd 命令提示符”,继而在窗口中键入“cd C:\”后回车,转到含 perl 命令执行程序的目录下。在“C: >”提示符下输入“perl Enrich_analysis.pl-h”回车后输出帮助信息。在“C: >”提示符下输入“perl Enrich_analysis.pl annotation.txt DEGlist.txt > Result.txt”,回车开始运行。输出的结果 Enrich_analysis.pl 会在“C:\”目录下自动生成文件“Result.txt”,并存放入其中。“Result.txt”分 8 列,依次表示:Enrich term、Ontology、Description、Number in input、Number in BG/Ref、p-value、Genes、FDR。

3 结果与分析

3.1 Enrich_analysis.pl 的结果与已发表 sci 文章比较

为了证明程序的可靠性,我们利用 Priest *et al.*, 2015 (Plos One)[17], Hu, *et al.*, 2016 (Bmc Plant Biology)[18], Wu, *et al.*, 2016 (Plant Physiology and Biochemistry)[19] 三篇 sci 文章的数据进行了比较分析(见表 2、表 3、表 4)。可以发现结果基本是相同的,但是 p-value 略有差异。这是由于我们使用的 GO 注释数据与作者使用的不完全相同。更多的比较可通过在线工具 agriGO (http://bioinfo.cau.edu.cn/agriGO)[20] 进行特定物种的分析。

表 2 Module 1 数据 GO 富集分析结果比较[17]

Table 2 Comparison of the GO enrichment results of Module 1 data[17]

Enrich_analysis.pl 结果							文章结果[17]		
Enrich	Ontology	Description	Number in input	Number in BG/Ref	p-value	FDR	query	refitem	p-value
GO:0006418	biological process	tRNA aminoacylation for protein translation	12	28	1.34×10^{-7}	8.91×10^{-6}	16	56	5.60×10^{-8}
GO:0009579	cellular component	thylakoid	21	147	2.86×10^{-5}	1.04×10^{-3}	29	262	2.80×10^{-5}
GO:0031969	cellular component	chloroplast membrane	3	30	7.20×10^{-4}	3.28×10^{-2}	9	49	8.00×10^{-4}
GO:0006457	biological process	protein folding	18	222	4.10×10^{-2}	3.2410^{-1}	15	203	5.20×10^{-2}

表 3 GO 富集分析结果比较[18]

Table 3 Comparison of the GO enrichment results[18]

Enrich_analysis.pl 结果							文章结果[18]		
Enrich	Ontology	Description	Number in input	Number in BG/Ref	p-value	FDR	query	refitem	p-value
GO:0050896	biological process	response to stimulus	166	5 099	7.70×10^{-8}	4.27×10^{-6}	166	6 928	8.30×10^{-9}
GO:0005618	cellular component	cell wall	47	896	1.13×10^{-7}	4.19×10^{-6}	47	1 179	1.70×10^{-8}
GO:0030312	Cellular component	external encapsulating structure	47	906	1.44×10^{-7}	3.19×10^{-6}	47	1 189	2.20×10^{-8}

续(表 3)

Enrich_analysis.pl 结果							文章结果[18]		
Enrich	Ontology	Description	Number in input	Number in BG/Ref	p-value	FDR	query	refitem	p-value
GO:0006629	Biological process	lipid metabolic process	51	940	9.50×10^{-9}	1.05×10^{-6}	51	1 376	3.90×10^{-8}
GO:0009628	biological process	response to abiotic stimulus	85	2 109	1.14×10^{-7}	3.16×10^{-6}	85	3 022	2.30×10^{-7}
GO:0003824	molecular function	catalytic activity	270	9 650	7.28×10^{-7}	1.35×10^{-5}	270	13 508	6.80×10^{-7}
GO:0016787	molecular function	hydrolase activity	107	3 005	8.53×10^{-7}	1.35×10^{-5}	107	4 293	1.60×10^{-6}
GO:0005623	cellular component	cell	393	15 915	1.06×10^{-4}	1.17×10^{-5}	393	22 048	3.20×10^{-5}
GO:0006950	biological process	response to stress	105	3 513	1.24×10^{-3}	8.59×10^{-3}	105	4 660	1.30×10^{-4}
GO:0005576	Cellular component	extracellular region	26	533	2.83×10^{-4}	2.61×10^{-3}	26	730	1.70×10^{-4}
GO:0009719	biological process	response to endogenous stimulus	53	1 438	3.11×10^{-4}	2.65×10^{-3}	53	2 015	2.70×10^{-4}
GO:0005975	biological process	carbohydrate metabolic process	41	931	6.01×10^{-5}	7.41×10^{-4}	41	1 439	3.00×10^{-4}
GO:0008152	biological process	metabolic process	343	13 840	8.57×10^{-4}	6.79×10^{-3}	343	19 328	6.70×10^{-4}

表 4　GO 富集分析结果比较[19]

Table 4　Comparison of the GO enrichment results[19]

Enrich_analysis.pl 结果							文章结果[19]		
Enrich	Ontology	Description	Number in input	Number in BG/Ref	p-value	FDR	query	refitem	p-value
GO:0006950	biological process	response to stress	34	3 584	1.65×10^{-15}	7.08×10^{-4}	33	4 660	1.60×10^{-15}
GO:0042592	biological process	homeostatic process	15	226	3.89×10^{-18}	3.35×10^{-16}	14	342	3.80×10^{-16}
GO:0065008	biological process	regulation of biological quality	18	647	2.64×10^{-15}	7.57×10^{-14}	17	863	2.10×10^{-14}
GO:0050896	biological process	response to stimulus	38	5 227	5.03×10^{-14}	1.08×10^{-12}	37	6 928	6.00×10^{-14}
GO:0065007	biological process	biological regulation	19	2 146	1.01×10^{-7}	1.09×10^{-6}	18	2 871	2.90×10^{-7}
GO:0044444	cellular component	cytoplasmic part	40	7 669	3.91×10^{-10}	4.80×10^{-9}	38	10 930	1.70×10^{-8}
GO:0005623	cellular component	cell	49	16 259	2.05×10^{-4}	1.26×10^{-3}	47	22 048	4.40×10^{-4}
GO:0043226	cellular component	organelle	18	3 914	2.41×10^{-3}	1.38×10^{-2}	36	12 251	7.20×10^{-6}

续(表 4)

Enrich_analysis.pl 结果							文章结果[19]		
Enrich	Ontology	Description	Number in input	Number in BG/Ref	p-value	FDR	query	refitem	p-value
GO:0044464	cellular component	cell part	47	14 094	1.76×10^{-5}	1.38×10^{-4}	45	19 532	8.90×10^{-5}
GO:0005737	cellular component	cytoplasm	42	8 344	3.40×10^{-10}	4.88×10^{-9}	40	11 866	9.50×10^{-9}
GO:0008152	biological process	metabolic process	47	14 136	1.79×10^{-5}	1.28×10^{-4}	46	19 328	1.70×10^{-5}
GO:0003824	molecular function	catalytic activity	42	9 878	1.14×10^{-7}	1.09×10^{-6}	42	13 508	2.80×10^{-8}
GO:0005739	cellular component	mitochondrion	20	1 112	1.65×10^{-13}	2.84×10^{-12}	19	1 611	3.50×10^{-12}
GO:0009628	biological process	response to abiotic stimulus	16	2 178	1.51×10^{-5}	1.30×10^{-4}	16	3 022	1.40×10^{-5}
GO:0009856	biological process	pollination	6	271	3.46×10^{-5}	2.29×10^{-4}	6	337	1.80×10^{-5}

3.2 Enrich_analysis.pl 的原理和特点

做富集分析需要两个条件:1)针对差异表达的基因(比如说基因芯片或者转录组的数据);2)需要该物种的基因组的所有基因组注释结果作为背景。

各种软件做 GO 富集性分析时一般都是使用超几何分布[21](Hypergeometric Distribution)进行计算。而 Fisher's Exact Test[22]是依据超几何概率分布得到。

在富集分析中把基因类比成非黑即白的抽球问题,用来判断和某个注释/分类是否有相关性。也可以列一个 2×2 的表(见表 5),进行独立性分析。之后有很多方法可以做检验,经典的有卡方检验和 fisher's exact test。但是卡方检验通常也只能做为近似估计值,特别是当样本量比较小时,计算并不准确。而 fisher's exact test 和超几何模式计算的 p-值是一样的。通过 FDR(false discovery rate,错误发现率)校正之计算得到 FDR 值后,以 0.05 为阈值,≤0.05的注释/分类为显著富集。

表 5 Fisher's Exact Test 2×2 表

Table 5 Fisher's Exact Test 2×2

Gene	Genes in category	Genes not in category	Total
DE genes	*n*11	*n*12	*n*1p
Not DE genes	*n*21	*n*22	*n*2p
Total	*n*p1	*n*p2	*n*pp

备注:*n*pp:物种基因总数;*n*1p:该物种有 *n*1p 个基因属于某个注释;*n*2p:该物种有 *n*2p 个基因不属于这个注释;*n*p1 表示差异表达的基因数目;

那么 *n*p1 个差异表达基因中正好有 *n*11 个基因属于这个注释的概率为:

$$P(X=n11 \mid npp, n1p, np1)=\frac{\binom{n1p}{n11}\binom{n2p}{n21}}{\binom{npp}{np1}} \quad (1)$$

Fisher's exact test 得分表示 *n*p1 个差异表达基因中至少有 *n*11 个属于这个注释的概率:

$$p=1-\sum_{i=0}^{n11-1}\frac{\binom{n1p}{i}\binom{n2p}{np1-i}}{\binom{npp}{np1}} \quad (2)$$

"Result.txt"为结果输出文档,Enrich_analysis.pl 脚本运行操作简单、快捷,所需设置的参数少。

3.2 Enrich_analysis.pl 的源代码

Enrich_analysis.pl 源代码见附件。

也可在 gitHub 平台 https://gist.github.com/xukaili/15bf411472092ebcf03b676352ec6a00 网址下载获得。

4 结 论

1)基因富集分析的主要任务是筛选出功能富集基因,有效地将基因表达信息与基因注释等信息相结合,从统计上把转录组数据与生物学意义很好地衔接起来,使得研究者们能够更加轻松、合理地解读转录组结果,还能达到降维的目的,使转录组的海量基因信息得到充分利用,但目前对生物基因注释信息的准确度和精确度还有待进一步完善。

2)本文用 Perl 语言写了 Enrich_analysis.pl,根据基因注释信息进行基因富集分析,并利用 Fisher's Exact Test 做检验,提高数据的可解释性。

3)利用 3 篇 SCI 文章的差异基因数据,对本 perl 脚本进行了对比测试,证明本 perl 脚本算法正确可靠。

4)目前各物种的注释信息并不完善,仍有很多基因尚没有注释信息,随着研究的不断深入,这些未知的基因也在不断得到注释,使得基因数据库在不断更新。但是目前现有的富集分析数据库中的数据并没有及时更新,或者有些国外富集分析网站不能访问。利用本 perl 脚本对最新的基因注释数据进行富集分析,相对于使用多年不更新的网络版工具更具有优势。

参考文献(References)

[1]SUBRAMANIAN A, TAMAYO P, MOOTHA V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proceedings of the National Academy of Sciences, 2005, 102(43):15545-15550. DOI: 10.1073/pnas.0506580102.

[2]DENNIS G, SHERMAN B T, HOSACK D A, et al. DAVID: database for annotation, visualization, and integrated discovery[J]. Genome Biology, 2003, 4(9): R60-R60. DOI: 10.1186/gb-2003-4-9-r60.

[3]HUANG D W, SHERMAN B T, LEMPICKI R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nature Protocols, 2009, 4(1): 44-57. DOI: 10.1038/nprot.2008.211.

[4]HUANG D W, SHERMAN B T, LEMPICKI R A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists[J]. Nucleic Acids Research, 2009, 37(1): 1-13. DOI: 10.1093/nar/gkn923.

[5]郭昊,朱云平,李栋,等. 肿瘤相关生物学通路的发现和建模[J]. 遗传, 2011, 33(8): 809-819. DOI: 10.3724/SP.J.1005.2011.00809.
GUO Hao, ZHU Yunping, LI Dong, et al. Identification, modeling and simulation of key pathways underlying certain cancers[J]. Hereditas, 2011, 33(8): 809-819. DOI: 10.3724/SP.J.1005.2011.00809.

[6]刘明,王米渠,丁维俊,等.表达谱芯片数据的基因功能富集分析[J]. 生物医学工程学杂志, 2010, 27(5): 1166-1168.
LIU Ming, WANG Miqu, DING Weijun, et al. Gene function enrichment analysis of microarray data[J]. Journal of Biomedical Engineering, 2010, 27(5): 1166-1168.

[7]AL-SHAHROUR F, DIAZ-URIARTE R, DOPAZO J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes[J]. Bioinformatics, 2004, 20(4): 578-580. DOI: 10.1093/bioinformatics/btg455.

[8]RIVALS I, PERSONNAZ L, TAING L, et al. Enrichment or depletion of a GO category within a class of genes: which test? [J] Bioinformatics, 2007, 23(4): 401-407. DOI: 10.1093/bioinformatics/btl633.

[9]HARRIS M A, CLARK J, IRELAND A, et al. The Gene Ontology (GO) database and informatics resource[J]. Nucleic Acids Research, 2004, 32(Database issue): D258-D261. DOI: 10.1093/nar/gkh036.

[10]YOUNG M D, WAKEFIELD M J, SMYTH G K, et al. Gene ontology analysis for RNA-seq: accounting for selection bias[J]. Genome Biology, 2010, 11(2): R14-R14. DOI: 10.1186/gb-2010-11-2-r14.

[11]KANEHISA M, GOTO S, FURUMICHI M, et al. KEGG for representation and analysis of molecular networks involving diseases and drugs[J]. Nucleic Acids Research, 2010, 38(Database issue): 355-360. DOI: 10.1093/nar/gkp896.

[12]KANEHISA M, ARAKI M, GOTO S, et al. KEGG for linking genomes to life and the environment[J]. Nucleic Acids Research, 2008, 36(Database issue): 480-484. DOI: 10.1093/nar/gkm882.

[13]THE COMPREHENSIVE PERL ARCHIVE NETWORK (CPAN) [EB/OL]. http://www.cpan.org, 2012-10-09.

[14]RANDAL L. SCHWARTZ, PHOENIX T, et al. Learning Perl. 5th Edition [M]. Sebastopol: O'Reilly Media, 2008.

[15]ACTIVESTATE SOFTWARE INC. ActivePerl for Windows [EB/OL]. http://www.activestate.com/activeperl, 2017-08-01.

[16]李旭凯,彭良才,王令强. pep_pattern.pl,搜索蛋白质序列模体的 Perl 脚本[J]. 华中农业大学学报, 2014, 33(04): 1-6.
LI Xukai, PENG Liangcai, WANG Lingqiang. pep_pattern.pl, a perl script for searching motifs in a group of related DNA/protein sequences[J]. Journal of Huazhong Agricultural University, 2014, 33(04): 1-6.

[17]PRIEST H D, FOX S E, ROWLEY E R, et al. Analysis of global gene expression in *Brachypodium distachyon* reveals extensive network plasticity in response to abiotic stress [J]. PLoS One, 2014, 9(1): e87499. DOI: 10.1371/journal.pone.0087499.

[18]HU J, CHEN G, ZHANG H, et al. Comparative transcript profiling of alloplasmic male-sterile lines revealed altered gene expression related to pollen development in rice (*Oryza sativa* L.)[J]. BMC Plant Biology, 2016, 16(1):175. DOI: 10.1186/s12870-016-0864-7.

[19]WU T, YANG C, DING B, et al. Microarray-based gene expression analysis of strong seed dormancy in rice cv. N22 and less dormant mutant derivatives[J]. Plant Physiology and Biochemistry, 2016, 99:27-38. DOI: 10.1016/j.plaphy.2015.12.001.

[20]TIAN T, LIU Y, YAN H, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update[J]. Nucleic Acids Research, 2017, 45(W1): W122-W129. DOI: 10.1093/nar/gkx382.

[21]WIKIPEDIA. Hypergeometric distribution [EB/OL]. http://en.wikipedia.org/wiki/Hypergeometric_distribution, 2015-02-23.

[22]WIKIPEDIA. Fisher's exact test [EB/OL]. http://en.wikipedia.org/wiki/Fisher%27s_exact_test, 2014-11-22.