Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Enrichment Disequilibrium: A novel approach for measuring the degree of enrichment after gene enrichment test

Yongshuai Jiang *, Mingming Zhang [1], Xiaodan Guo [1], Ruijie Zhang *,[1]

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, China

ARTICLE INFO

Article history:
Received 31 May 2012
Available online 7 July 2012

Keywords:
Gene
Functional analysis
Enrichment analysis

ABSTRACT

Motivation: Commonly used gene enrichment analysis methods, such as Hypergeometric distribution, play an important role in the functional analysis of interesting gene lists. But the statistical significance obtained by these methods only represents the probability of error that is involved in accepting enrichment, and is not suitable to evaluate the degree of enrichment. Although there have been some methods to measure the enrichment degrees, such as relative enrichment factor, new methods are still needed to meet the requirements for comparing the degree of enrichment.

Results: We developed a novel method, Enrichment Disequilibrium (ED), to measure the degree of enrichment. Enrichment equilibrium means that the interesting gene set and the known functional gene set (such as a KEGG pathway) are independent (i.e. random association). ED is defined as the degree of non-independence. Compared with the relative enrichment factor, ED has a clearer biological meaning, is a standardized indicator, and has a symmetrical interval (range from −1 to +1). It is more suitable to measure the enrichment degree. For an interesting gene set, researchers can obtain some significant functional gene sets by traditional enrichment test. Then using ED, they can compare the degree of enrichment among these significant gene sets, and prioritize them.

1. Introduction

High-throughput technologies (such as DNA microarrays, proteomics, ChIP-on-CHIPs, etc.) usually produce large amounts of ‘interesting’ gene lists as their final results. However, understanding the biological meaning of the output gene lists is still a challenge [1]. The being developed genome annotation databases (such as KEGG pathway database [2–5] and GO database [6–9]) and a number of high-throughput enrichment analysis methods (such as Hypergeometric distribution, Chi-square, Binomial probability and Fisher’s exact test [10,11]) made it possible for us to understand the biological meaning of the ‘interesting’ gene set from system or functional level.

In the process of enrichment analysis, the statistical significance (p-value) obtained by enrichment test methods (such as Hypergeometric distribution) was used to identify whether there is an association between a functional pathway (or a GO category) and the ‘interesting’ gene set. In some studies, the p-value was also used to represent the degree of enrichment and prioritize significant pathways or GO categories [12]. But the p-value is not suitable to evaluate the degree of enrichment. For example, there are 300 genes in an ‘interesting’ gene list and 10,000 genes in the background distribution. Consider two known functional gene sets, such as two specific GO categories: G1, containing 50 genes, 10 of which were ‘interesting’ genes, and G2, containing 500 genes, 50 of which were ‘interesting’ genes. The two p-values, calculated using the Hypergeometric distribution, were $p1 = 1.942\text{e-}7$ for G1 and $p2 = 4.730\text{e-}12$ for G2. Though $p2 < p1$, it does not mean that the ‘interesting’ genes had a higher degree of enrichment in G2 than G1. In general, the p-value only represents the probability of error that is involved in accepting enrichment. When we calculated the percentage of the ‘interesting’ genes in the two gene set G1 and G2 (enrichment percentage, EP), we could see that EP1 = 20% (=10/50) > EP2 = 10% (=50/500). This indicates that G1 has a higher proportion of ‘interesting’ genes than G2. Though Zeeberg et al. also developed a ‘relative enrichment factor’ $R_e = (n_g/n_G)/(n_K/n_N)$ (where n_g is the number of ‘interesting’ genes in a known functional gene set, n_G is the total number of genes in the known functional gene set, n_K is the number of ‘interesting’ genes in an ‘interesting’ gene list, and n_N is the total number of genes in the background distribution) [13,14], to try to measure the degree of enrichment, it is a simple improvement of EP. New methods are still needed to meet the requirements for comparing the degree of enrichment.

In this study, we developed a novel method to measure the degree of enrichment in gene enrichment analysis. Our method mainly uses the principle of independence ($P(AB) = P(A)P(B)$) in

* Corresponding authors. Fax: +86 045186615922.
E-mail addresses: jiangyongshuai@gmail.com (Y. Jiang), zhangruijie2009@yahoo.com.cn (R. Zhang).

[1] Joint First Authors.

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.06.154

statistics. This principle has been successfully used to define the degree of linkage disequilibrium (LD) between the two Single Nucleotide Polymorphisms (SNPs). Consider two SNP (SNP-A and SNP-B), each having two alleles (A1, A2, B1, and B2). LD coefficient is defined as $D = p(A1B1) - P(A1)P(B1)$, and it was usually used to measure the degree of non-random association of alleles at the two loci [15]. In this study, we will extend the application of the principle of independence, and use it to describe the relationship between the interesting gene set and the known functional gene set.

2. Methods

2.1. Enrichment Equilibrium

Suppose that there were n_N genes in a background data set N, and total n_K genes in an 'interesting' gene set K (such as a list of differentially expressed genes between two groups of samples). Consider a known functional gene set G (such as a specific GO term), containing n_G genes, n_g of which were in the 'interesting' gene list. If there is no association between the 'interesting' gene set K and the functional gene set G, we usually think that the two sets are independent of each other. Based on the principle of independence, we will get the following expression: $P(KG) = P(K)P(G)$ (Fig. 1A). Then, if K and G satisfy the following conditions:

$$P(KG) = P(K)P(G)$$

We call that the 'interesting' gene set K is Enrichment Equilibrium in the known functional gene set G.

2.2. Enrichment Disequilibrium

Enrichment Disequilibrium (ED) can be described as the non-random association between the 'interesting' gene set K and the known functional gene set G. As we described in the above section, if there is no association between K and G, we will get $P(KG) = P(K)P(G)$. Then, if there is non-random association between K and G, we will get $P(KG) \neq P(K)P(G)$. Therefore, we define the Enrichment Disequilibrium coefficient ed as follow:

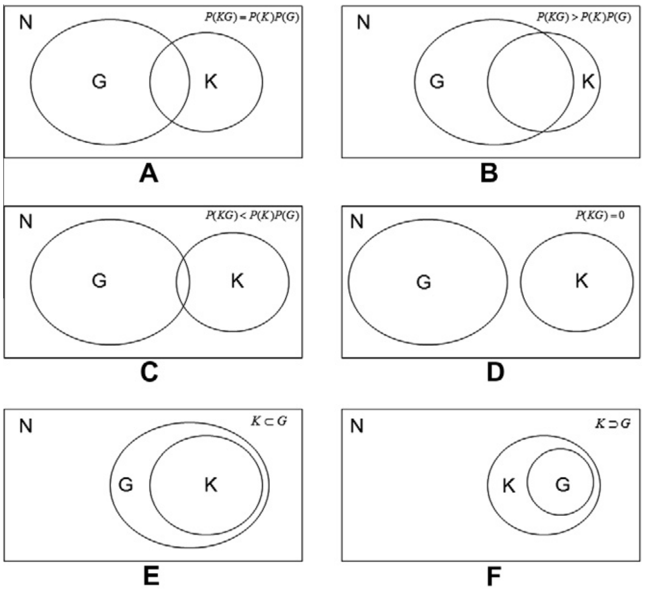

Fig. 1. The relationships between the 'interesting' gene set K and the functional gene set G. (A) K and G are independent. (B) The number of 'interesting' gene annotated in G is more than random. (C) The number of 'interesting' gene annotated in G is less than random. (D) K and G are incompatible. (E) K is a subset of G. (F) G is a subset of K.

$$ed = p(KG) - p(K)p(G)$$

ed can be used to described the degree of non-random association between the 'interesting' gene set K and the known functional gene set G. $ed > 0$ means that the number of 'interesting' genes annotated in functional gene set G is more than random (overrepresentation, Fig. 1B), and $ed < 0$ means that the number of 'interesting' genes annotated in functional gene set G is less than random (underrepresentation, Fig. 1C).

2.3. Some special ed values

In this section, we will discuss some special relationships between the 'interesting' gene set K and the functional gene set G. There were four special relationships between K and G: K and G are independent, K and G are incompatible, $K \subset G$ and $K \supset G$. We will get four different ed values:

(1) If K and G are independent (Fig. 1A), then $ed = p(KG) - p(K)p(G) = 0$.
(2) If K and G are incompatible (Fig. 1D), then $ed = p(KG) - p(K)p(G) = -p(K)p(G)$.
(3) If $K \subset G$ (Fig. 1E), then $ed = p(K) - p(K)p(G) = p(K)(1 - p(G)) = p(K)p(\bar{G})$.
(4) If $K \supset G$ (Fig. 1F), then $ed = p(G) - p(K)p(G) = p(G)(1 - p(K)) = p(\bar{K})p(G)$

2.4. Standardization of ed

We have measured the degree of enrichment using ed, but ed is a value has not been standardized. In some cases, ed is not suitable for comparing the degree of enrichment between two functional sets (such as two KEGG pathways). For example, in Fig. 1E and F, both $k \subset G$ and $K \supset G$ represent the strongest associations between the 'interesting' gene set K and the functional gene set G. But we can observe different ed values in the above section: for $K \subset G$ (Fig. 1E), $ed = p(K)p(\bar{G})$, however, for $K \supset G$ (Fig. 1F), $ed = p(\bar{K})p(G)$. To solve this problem, and convenient to compare the degree of enrichment between two functional sets, we standardize the ed as follows:

$$ED = \frac{ed}{ed_{\max}}$$

where

$$ed_{max} = \begin{cases} p(K)p(\bar{G}) & if \quad P(KG) > p(K)p(G) \quad and \quad n_G > n_k \\ p(\bar{K})p(G) & if \quad P(KG) > p(K)p(G) \quad and \quad n_G < n_k \\ p(K)p(G) & if \quad P(KG) < p(K)p(G) \end{cases}$$

ED range from -1 to +1. Table 1 shows the relationships between the 'interesting' gene set K and the functional gene set G under different ED intervals.

2.5. How to use ED

As a useful method, ED can be used in assessing the degree of enrichment after traditional enrichment analysis test. Using traditional enrichment analysis methods, such as Hypergeometric distribution and Fisher's exact test, researchers can obtain some significant functional gene sets (such as some pathways or GO terms). Then using the ED, they can measure the enrichment degree of these significant functional sets.

In addition, when researchers calculate the ED, they can use the following formula to estimate the probability: $\hat{P}(N) = 1, \hat{P}(K) = n_k/n_N, \hat{P}(G) = n_G/n_N$ and $\hat{P}(KG) = n_g/n_N$.

Table 1
The relationships between K and G under different ED intervals.

ED	Relationship between K and G
$ED = 1$	$K \subset G$ or $K \supset G$, in the two cases, the 'interesting' gene set K and the known functional gene set G have the strongest association.
$0 < ED < 1$	The number of 'interesting' genes annotated in functional gene set G is more than random. A larger ED indicates a higher degree of enrichment (overrepresentation).
$ED = 0$	The 'interesting' gene set K and the known functional gene set G are independent.
$-1 < ED < 0$	The number of 'interesting' genes annotated in functional gene set G is less than random. A smaller ED indicates a higher degree of depletion (underrepresentation).
$ED = -1$	The 'interesting' gene set K and the known functional gene set G are incompatible.

3. Results

3.1. An example of using ED

In this study, we use the differential expression results from Chin et al. study [16]. They analyzed the gene expression differences between human induced pluripotent stem cells (hiPSCs) and human embryonic stem cells (hESCs). By using a Student's t-test (p-value < 0.05) and minimum 1.5 fold expression difference between hESC and early passage hiPSC, they detected 3,947 differential expression genes. Among these genes, 3640 genes has Entrez gene ID in NCBI build 37.2.

There are total 5900 genes ($n_N = 5900$) in human KEGG pathways. These genes can be considered as a background data set N. Among 3640 differential expression genes, 1132 genes ($n_K = 1132$) can be annotated to KEGG database. These genes can be considered as an 'interesting' gene set K. There are total 220 human pathways containing at least 10 genes in KEGG database. Each pathway can be considered as a functional gene set G. Next, we will use Hypergeometric distribution to find some 'interesting' KEGG pathways, and use the Enrichment Disequilibrium coefficient ED to prioritize these pathways.

The results (Table 2) showed that there were 16 significant pathways which had $p < 0.01$ (Hypergeometric test), and the DNA replication pathway (hsa03030) has the highest ED ($ED = 0.484$). This indicated that the DNA replication pathway has the strongest association with expression differences. But the DNA replication pathway (p = 3.35E-08) did not have the lowest p-value. The pathway with the lowest p-value is RNA transport (hsa03013), but the ED is 0.272 (<0.484).

In this example, we not only find some "interesting" pathways by traditional enrichment analysis test, but also measure the degree of enrichment by Enrichment Disequilibrium coefficient ED.

3.2. Comparison between relative enrichment factor R_e and Hypergeometric test p-value

In this section, we will compare the correlation between enrichment analysis test p-value and traditional indicators that measure the degree of enrichment. Here, Hypergeometric test will be used to represent the enrichment analysis test method. EP and R_e will be used to represent the traditional indicators that measure the degree of enrichment. We calculated the EP and R_e for all above 220 human KEGG pathways. The Pearson's correlation coefficient between EP and R_e is greater than 0.99, and then we only use R_e as a traditional indicator to compare the correlation with Hypergeometric test p-value.

Fig. 2A shows a scatter plot of R_e against p-value. From Fig. 2A, we can see that with the p-value decreases, R_e tends to increase. This indicates that some correlation does exist for R_e and p-value. We calculated the Pearson's correlation coefficient between R_e and p-value. The correlation coefficient is −0.890. Though there is a high degree of negative correlation between R_e and p-value, we still found different degree of correlation in different range of p-value. For example, when p is less than 0.1 or more than 0.9 ($p = 0.9$ corresponds to the left tail probability of 0.1), R_e and p-value show a lower correlation. The Pearson's correlation coefficients are −0.613 for $p < 0.1$ and −0.482 for $p > 0.9$. In addition, 0.05 and 0.01 are often used as significance level. For pathways which have $p < 0.05$ and 0.01, we also calculated the Pearson's correlation coefficients between R_e and p-value. The two correlation coefficients are −0.549 for $p < 0.05$ and −0.533 for $p < 0.01$. This indicates that when p-value less than a certain threshold (e.g. $p < 0.01$), it is not suitable to be used to measure the degree of enrichment.

3.3. Comparison between ED and Hypergeometric test p-value

To compare the correlation between ED and Hypergeometric test p-value, we also drew a scatter plot of ED against p-value

Table 2
The enrichment analysis results of differential expression genes. 16 significant pathways were ranked by ED.

Pathway ID	Pathway name	n_N	n_K	n_G	n_g	Hypergeometric test p-vaule	ED
hsa03030	DNA replication	5900	1132	36	21	3.35E-08	0.48441
hsa03410	Base excision repair	5900	1132	34	15	0.000203	0.3085
hsa03008	Ribosome biogenesis in eukaryotes	5900	1132	84	37	3.85E-08	0.30764
hsa03013	RNA transport	5900	1132	153	63	5.93E-11	0.27211
hsa04110	Cell cycle	5900	1132	128	52	3.95E-09	0.26528
hsa03018	RNA degradation	5900	1132	71	28	1.87E-05	0.25058
hsa03430	Mismatch repair	5900	1132	23	9	0.006509	0.24679
hsa03420	Nucleotide excision repair	5900	1132	48	17	0.00228	0.20084
hsa05217	Basal cell carcinoma	5900	1132	55	18	0.005074	0.16756
hsa04115	p53 signaling pathway	5900	1132	69	22	0.003626	0.15712
hsa03022	Basal transcription factors	5900	1132	54	17	0.009511	0.15214
hsa03015	mRNA surveillance pathway	5900	1132	84	26	0.003059	0.14559
hsa04120	Ubiquitin mediated proteolysis	5900	1132	139	41	0.001055	0.12758
hsa00240	Pyrimidine metabolism	5900	1132	100	29	0.005836	0.12143
hsa00230	Purine metabolism	5900	1132	163	46	0.001651	0.11179
hsa03040	Spliceosome	5900	1132	128	36	0.004714	0.11061

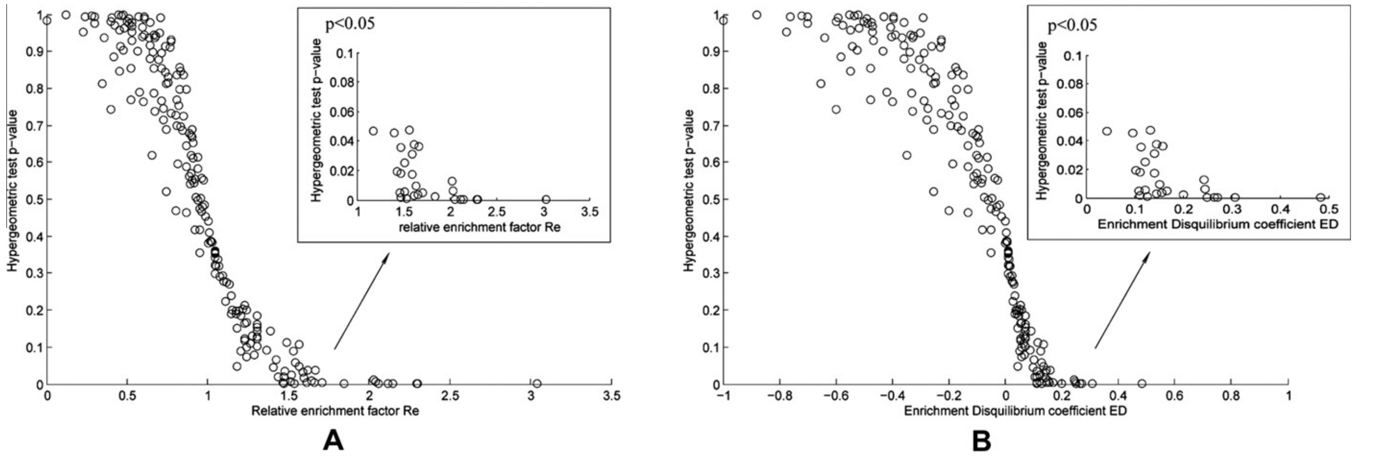

Fig. 2. Scatter plots of enrichment analysis test *p*-value against indicators that measure the degree of enrichment. (A) A scatter plot of R_e against *p*-value. (B) A scatter plot of *ED* against *p*-value. We can see that when *p*-value <0.05 (commonly used threshold), *p*-value shows a low correlation with R_e or *ED*.

(Fig. 2B). From Fig. 2B, we also observed a high degree of negative correlation between *ED* and *p*-value. The Pearson's correlation coefficient is −0.895. However, we also notice that when *p* is less than 0.1 or more than 0.9, *ED* and *p*-value show a lower correlation. This is consistent with the results described in the above section. Combining the results in this section and the above section, we found that when *p*-value < 0.1, 0.05 or 0.01 (commonly used threshold), *p*-value shows a low correlation not only with the traditional R_e , but also with our novel method *ED*. This implies that the *p*-value only represents the probability of error, and not suitable to represent the degree of enrichment.

3.4. Comparison between ED and R_e

In this section, we will describe the correlation and differences between traditional R_e and our novel method *ED*. Fig. 3 shows a scatter plot of R_e against *ED*. From Fig. 3, we can observed a high degree of positive correlation between R_e and *ED*. The Pearson's correlation coefficient is 0.921. This implies that R_e and *ED* have some similar properties, and can solve the same types of problems (measure the degree of enrichment). Although both R_e and *ED* can measure the degree of enrichment, they differ in some aspects:

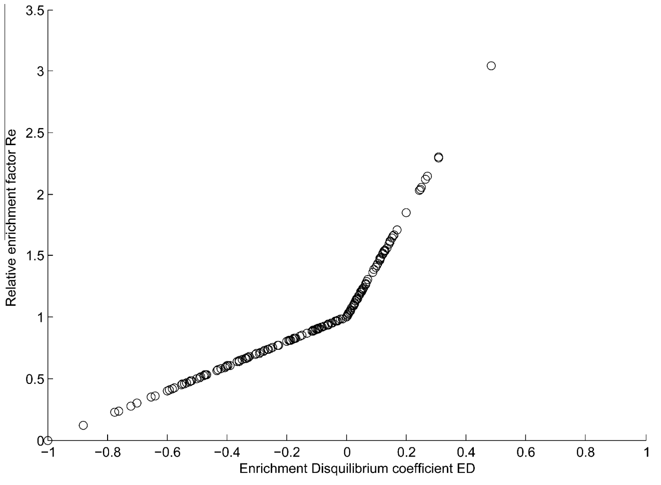

Fig. 3. Scatter plots of R_e against *ED*.

(1) *ED* reflects the degree of non-random association between the 'interesting' gene set and the known functional gene set, while R_e describes the ratio between percentage of 'interesting' genes in known functional gene set and in background gene set. Therefore, *ED* has a more specific biological meaning.

(2) *ED* is the standardized value of *ed*, and will fall within a certain range [−1,1]. The interval is symmetrical about 0. There were three special reference values: *ED* = 1 (upper bound) represents the strongest association between the 'interesting' gene set and the known functional gene set; *ED* = 0 represents the two sets are independent; *ED* = −1 (lower bound) represents the two sets are incompatible. For a given *ED*, we could know how far it is from the strongest degree of enrichment (*ED* = 1). However, for R_e, it is a value has not been standardized, and has no fixed upper bound (R_e range from 0 to +∞). The interval of R_e is non-symmetrical. Therefore, for a specific 'interesting' gene set and a known functional gene set, *ED* can help us better understand the degree of enrichment by comparing with the upper and lower bound.

(3) As we described in the Method, both $K \subset G$ and $K \supset G$ (Fig. 1E and F) represent the strongest associations between the 'interesting' gene set and the functional gene set, and they should have the same degree of enrichment. In the two cases, *ED* has the same value +1, and could represent the same degree of enrichment. But R_e has different values: for $K \subset G$ (i.e.$n_g = n_K$, Fig. 1E), $R_e = (n_g/n_G)(n_K/n_N) = n_G/n_N$, however, for $K \supset G$(i.e. $n_g = n_G$, Fig. 1F), $R_e = (n_g/n_G)(n_K/n_N) = n_K/n_N$. Therefore, *ED* is more suitable to describe the relationship between the 'interesting' gene set and the functional gene set.

4. Discussion

Enrichment analysis is important in understanding the biological interpretation of 'interesting' gene set (derived from the results of high-throughput data analysis). In the enrichment analysis process, both enrichment test and enrichment degree should be considered. They were different types of methods. Enrichment test was used to identify whether there is an association between a functional gene set and the 'interesting' gene set, while enrichment degree was used to measure the extent of the overlap. In this study, we also illustrate that when enrichment test *p*-value less than a certain threshold, enrichment test *p*-value and enrichment degree indicator (*ED* or R_e) have a lower correlation. However, some researchers usually pay no attention to the difference between enrichment test and enrichment degree, and only use *p*-value to measure the degree of enrichment and prioritize significant functional gene sets. We must emphasize that the *p*-value only represents the probability of error, and is not suitable to measure the degree of enrichment. A complete enrichment analysis should include two steps: (1) the first step is to identify which functional gene sets are significantly associated with the 'interesting' gene set by using enrichment test method, (2) the second step is to measure the degree of enrichment and prioritize significant functional gene sets by using enrichment degree indicator.

In this study, we focus on the second step. We have developed a novel method *ED* to measure the degree of enrichment. The principle of independence ($P(AB) = P(A)P(B)$) was used to describe random association between a functional gene set and the 'interesting' gene set (i.e. Enrichment Equilibrium). The degree of non-independence could reflect the strength of the association between the two sets. Therefore, the *ED* has a clear biological meaning, and is suitable to describe the relationship between the 'interesting' gene set and a functional gene set. As a standardized value, *ED* convenient to compare the enrichment degree between different functional sets, and can help to prioritize these functional sets. In addition, the symmetrical interval of *ED* (rang $[-1, +1]$) and three special reference values ($ED = -1, 0, +1$) can help us better understand the relationship between the 'interesting' gene set and a functional gene set.

At last, we hope that *ED* can be widely used to measure the enrichment degree of after enrichment test (such as Hypergeometric test).

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 81172842, 30871394, 61170154), Heilongjiang Province Health Department Fund (Grant No. 2011-250).

References

[1] W. Huang da, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (2009) 44–57.

[2] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29–34.

[3] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (2000) 27–30.

[4] J. Wixon, D. Kell, The Kyoto encyclopedia of genes and genomes–KEGG, Yeast 17 (2000) 48–55.

[5] M. Kanehisa, The KEGG database, Novartis Found Symp 247 (2002) 91-101; discussion 101-103, 119-128, 244-152.

[6] J.A. Blake, M.A. Harris, The Gene Ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis, Curr. Protoc. Bioinformatics, Chapter 7 (2002) Unit 7 2.

[7] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, R. Apweiler, The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro, Genome Res. 13 (2003) 662–672.

[8] The Gene Ontology (GO) project in 2006, Nucleic Acids Res. 34 (2006) D322–326.

[9] The Gene Ontology project in 2008, Nucleic Acids Res. 36 (2008) D440–444.

[10] W. Huang da, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, Nucleic Acids Res. 37 (2009) 1–13.

[11] D.L. Gold, K.R. Coombes, J. Wang, B. Mallick, Enrichment analysis in high-throughput genomics - accounting for dependency in the NULL, Brief. Bioinform. 8 (2007) 71–77.

[12] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, G. Sherlock, GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, Bioinformatics 20 (2004) 3710–3715.

[13] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, K.J. Bussey, J. Riss, J.C. Barrett, J.N. Weinstein, GoMiner: a resource for biological interpretation of genomic and proteomic data, Genome Biol. 4 (2003) R28.

[14] P. Khatri, S. Draghici, Ontological analysis of gene expression data: current tools, limitations, and open problems, Bioinformatics 21 (2005) 3587–3595.

[15] J.C. Mueller, Linkage disequilibrium for different scales and applications, Brief. Bioinform. 5 (2004) 355–364.

[16] M.H. Chin, M.J. Mason, W. Xie, S. Volinia, M. Singer, C. Peterson, G. Ambartsumyan, O. Aimiuwu, L. Richter, J. Zhang, I. Khvorostov, V. Ott, M. Grunstein, N. Lavon, N. Benvenisty, C.M. Croce, A.T. Clark, T. Baxter, A.D. Pyle, M.A. Teitell, M. Pelegrini, K. Plath, W.E. Lowry, Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures, Cell Stem Cell 5 (2009) 111–123.