W302–W306 Nucleic Acids Research, 2011, Vol. 39, Web Server issue Published online 29 May 2011
doi:10.1093/nar/gkr392

ADGO 2.0: interpreting microarray data and list of genes using composite annotations

Sang-Mun Chi[1], Jin Kim[2], Seon-Young Kim[3,*] and Dougu Nam[4,*]

[1]School of Computer Science and Engineering, Kyungsung University, Busan, [2]School of Computer Science and Engineering, Seoul National University, Seoul, [3]Medical Genomics Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon and [4]School of Nano-Bioscience and Chemical Engineering, UNIST, Ulsan, Rep. of Korea

Received February 21, 2011; Revised April 19, 2011; Accepted May 3, 2011

ABSTRACT

ADGO 2.0 is a web-based tool that provides composite interpretations for microarray data comparing two sample groups as well as lists of genes from diverse sources of biological information. Some other tools also incorporate composite annotations solely for interpreting lists of genes but usually provide highly redundant information. This new version has the following additional features: first, it provides multiple gene set analysis methods for microarray inputs as well as enrichment analyses for lists of genes. Second, it screens redundant composite annotations when generating and prioritizing them. Third, it incorporates union and subtracted sets as well as intersection sets. Lastly, users can upload their own gene sets (e.g. predicted miRNA targets) to generate and analyze new composite sets. The first two features are unique to ADGO 2.0. Using our tool, we demonstrate analyses of a microarray dataset and a list of genes for T-cell differentiation. The new ADGO is available at http://www.btool.org/ADGO2.

INTRODUCTION

High-throughput omics experiments often produce lists of genes, and their biological interpretations have been of substantial interest. Typical approaches examine the extent of the overlap between a list of genes and predefined annotated gene sets using hypergeometric distribution, chi-square or Fisher's exact test, which may be dubbed collectively as *gene list analysis* (GLA) (1). For microarrays, each gene has its own score (e.g. two sample *t*-statistic or fold-change value) and an alternative approach, called *gene set analysis* (GSA), is applicable without selecting a list of genes (2). In many cases, the 'interpretation' of large-scale data indicates investigating the enrichment of pre-set knowledge within the given data. Accordingly, such enrichment analyses are widespread over omics research regardless of the data analyzed (microarray, mass spectrometry, ChIP-chip or next-generation sequencing). In addition, a number of algorithms and tools have been developed in this context (1–3).

In both approaches (GSA and GLA), the predefined gene sets play key roles in biological interpretations. Such gene sets are usually derived from biological databases such as Gene Ontology (4) or KEGG (5), where they share a common biological annotation for pathways, functions, cellular localizations or targets of a common transcription factor (TF), for instance. One important problem with most existing methods is that they handle only gene sets with unary annotations, thus limiting the discriminating power of the method employed. For example, suppose we want to examine whether a given list of genes is enriched with the putative targets of some TF. Because most gene sets that share a common TF binding site are dominated with false positive targets, this simple approach may not be very successful when used to uncover the relevant TFs. However, if we take intersections between the putative TF target sets and the gene sets of Gene Ontology, some of them may define biologically more relevant gene sets, which then may be enriched with the gene list. With this rationale, composite annotation gene sets were introduced for GSA (6) and GLA (7), respectively. Thereafter, several software tools were developed for GLA based on composite annotations (8–10). ADGO (6) and ProfCom (9) use Boolean set operations (intersection, union and subtraction) to generate composite gene sets, and GENECODIS (11) and COFECO (10) employ an association rule-mining algorithm to extract co-occurring annotations. In any case, the composite interpreters usually display quite a long and redundant list of significant gene sets, many of which largely overlap each other. Therefore, removing

*To whom correspondence should be addressed. D.Nam. Tel: +82 52 217 2525; Fax: +82 52 217 2509; Email: dougnam@unist.ac.kr
Correspondence may also be addressed to Seon-Young Kim. Tel: +82 42 879 8116; Fax: +82 42 879 8119; Email: kimsy@kribb.re.kr

redundancy and abstraction appear to be an important issue when utilizing composite annotations. Here, we suggest three criteria for filtering composite gene sets for GSA and GLA.

(1) If a composite set is largely overlapped with some single set over a threshold, that set should be removed *a priori*. In other words, if the members of sets are very similar to each other, the single annotation should have priority.
(2) A significant intersection or union set should be screened, if any of the single sets used to generate them are also significant.
(3) A significant subtracted set should be screened, if the single set that contains the subtracted set is also significant.

Taking into account these considerations, we constructed web-based software called ADGO 2.0 to provide composite interpretations for both microarrays and lists of genes. The previous version of ADGO was designed to illustrate the idea of using composite annotations for GSA and provided a single GSA method (6). The current version was totally rebuilt considering the automatic updating of (composite) gene sets, and extended in terms of both coverage and methods.

MATERIALS AND METHODS

Supported analyses

ADGO 2.0 currently supports analyses of eight popular organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Escherichia coli*.), and from four to seven kinds of annotations (GO terms for biological processes, cellular components, and molecular functions, KEGG pathways, chromosome and *cis*-regulatory motifs, and OMIM) are provided depending on the species selected. The user can choose one of the four methods (*Z*-statistic, gene permutation, sample permutation and Gene Set Enrichment Analysis (GSEA)) for GSA and the two methods (Fisher's exact test and hypergeometric distribution) for GLA. Only applicable methods are displayed depending on the format of input data.

Construction of annotation gene set databases

For all of the gene sets from the seven annotation categories included in ADGO 2.0, we applied three types of Boolean set operations (intersection, union and subtraction) to each pair of gene sets across different categories: the 10–20% rule was applied for intersections and subtraction (6). In other words, a pair of single gene sets was required to have at least 10 genes in common and the two subtracted sets were required to contain at least 20% of the genes in each single set. Union operations were applied if a pair of gene sets has five or more elements in common. The 'subtraction' of two annotation sets A and B is denoted by $A - B$, which is the intersection of A and the complement of B. To ensure the generated composite set is a genuinely novel set, it was compared with each single set and discarded if it has any overlap with some single set over a threshold ('Filtering Composite Sets'). The overlap (%) is computed by the portion of intersection between the composite set and a single set over the union of these two sets. For the 60% threshold, ~27% of all composite sets are screened for the three categories of Gene Ontology. All of these single and composite gene sets are prepared in the server in advance and retrieved according to the user's choice of gene set categories. One important feature of ADGO 2.0 is that the user can upload and analyze his/her own annotation gene sets. If the user chooses some of the built-in gene sets and uploads user gene set data, the server then generates *ad hoc* composite sets and shows the computation results. For this reason, it takes much more time for analyzing the user's gene sets.

Processing methods

If the user uploads microarray data or a list of genes, the server detects the file format and displays relevant analysis methods and other options. For a microarray input, four gene set analysis methods are available. Among them, 'Z-test' (12) and 'Gene permutation' (13) are gene randomization methods, and 'Sample permutation' (13) and 'GSEA' (14) are sample randomization methods. We used the average *t*-value for the set score in the gene or sample permutation methods. The *Z*-test is a parametric method and is the fastest. GSEA is the most widely used but usually takes more time for computing. This becomes problematic when analyzing composite annotations, as the number of gene sets to be handled increases in a quadratic manner against the number of usual single annotations. For this reason, we newly realized the algorithm in C++. We fixed the power of the gene score as $p = 1$ to deploy the weighted Kolmogorov–Smirnov gene set statistic. For a gene list input, the 'Fisher's exact test' and 'Hypergeometric distribution' are provided for the analysis method.

For both GSA and GLA, we provide two types of filtering methods for significant composite sets: The 'Strong Type' and the 'Weak Type'. For the strong-type option, a significant composite set is screened if a single set involved in generating the composite set is also significant. For the weak-type option, a significant composite set is displayed if it has a smaller *P*-value than those of the individual single sets. Therefore, the weak-type option yields more composite sets that are significant.

Input data types and options

For microarray input, the user can upload microarray data with two sample groups. The first column should be the header for gene IDs and the sample data values should follow in the next columns. ADGO 2.0 accepts both single and dual channel gene IDs for microarray input. For a single channel input, the probe IDs for Affymetrix, Illumina and Agilent chips are supported. For a dual channel input, five types of gene IDs (gene symbol, Ensemble, Entrez, Refseq and Uniprot) as well as the systematic names for *Saccaromyces cerevisiae* are supported. The first sample group data should appear in the first k columns of data values and the second group data should follow in the next l columns. k and l should be specified in the 'Sample Size' option. ADGO 2.0 then computes the two-sample t-statistic or average fold-change values for each gene and proceeds. We have another option for microarray input. If the user wants to use gene scores other than the two-sample t-statistic or fold-change values, he/she can directly use the gene score data (a single value for each gene).

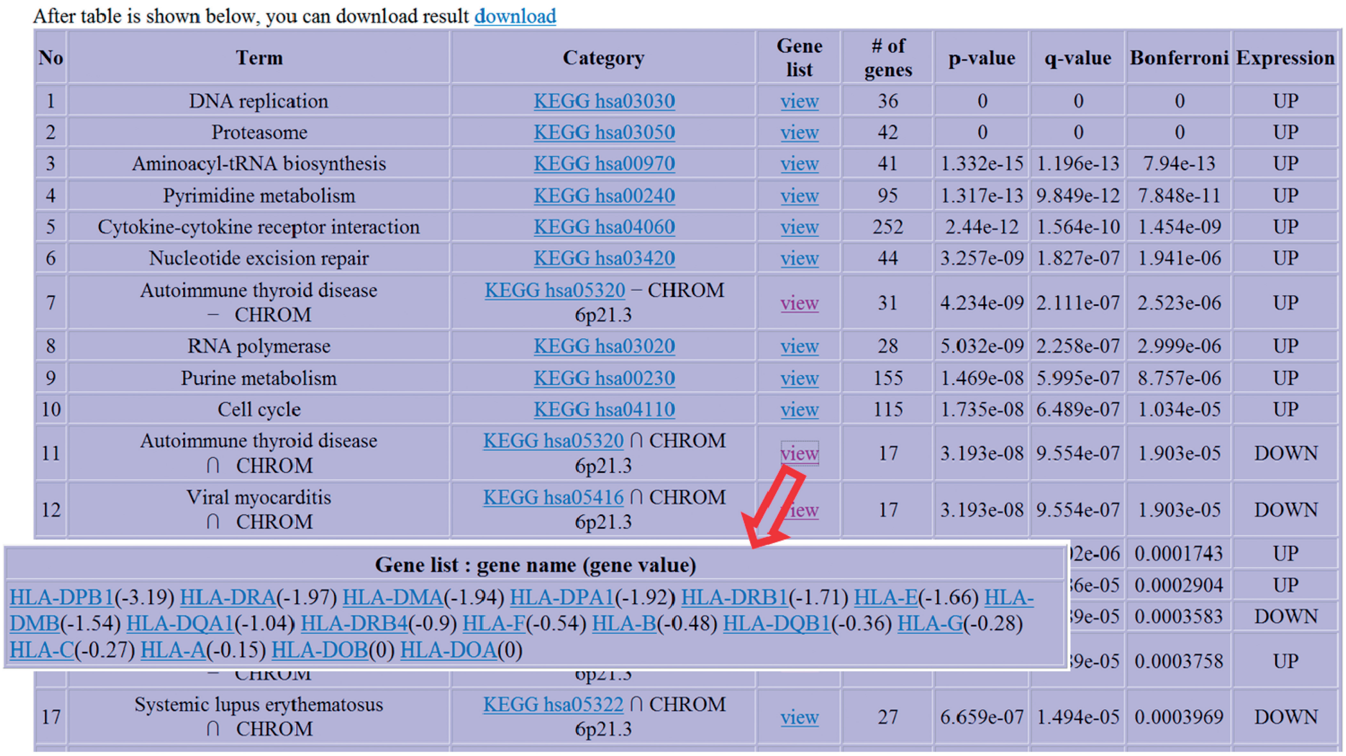
For a gene list input, the same dual channel IDs are supported. We constructed the reference (background) genes for GLA by merging all the genes contained in each annotation set. In any type of input data, the user can also paste the data into the 'Paste data panel' without uploading the data file. More detailed information is available from our web site (http://www.btool.org/ADGO2).

Outputs

If the user executes the analysis, the server shows a list of significant gene sets (single and composite) along with gene set names, gene set id's, members of each gene set, P-values, False Discovery Rate (FDR) Q-values and Bonferroni's corrected P-values (Figure 1). The members of each gene set are listed in a descending order of their association strength. The gene set list is sorted according to the FDR Q-values. Certainly, composite annotations increase the number of annotation sets to be analyzed to change the analysis results more or less. However, the FDR Q-values reflect the increased number of gene sets and provide the adjusted significance threshold. The computation results are also downloadable as a text file.

Analysis example

We present an example of gene expression data analysis by ADGO 2.0 to demonstrate its utility. Schones *et al.* (15) compared the gene expression patterns between resting and activated T cells to understand the molecular changes that occur during the T-cell differentiation (GEO number: GSE10437). We computed the fold changes of each gene in the microarray data to test a GSA. We chose the 'Z test' method (Q-value cutoff: 0.01) and two annotation categories: KEGG and Chromosome. We checked three types of composite sets: 'Single' + 'Intersection' + 'Subtraction'. Many KEGG pathways related to immune responses including 'Autoimmune thyroid disease', 'Antigen processing and presentation' and 'Graft-versus-host disease' were not significant by themselves when we used only KEGG categories. However, they were significantly induced if we excluded the genes in the chromosome 6p21.3 set. Interestingly, the same immune-related gene sets were significantly downregulated when intersected with the 6p21.3 set. This suggests that some part of the chromosome region 6p21.3 is locked during T-cell differentiation while other immune-related genes outside this region are activated. Indeed, this intersection set contained many HLA (human leukocyte antigen) genes that were mostly downregulated [i.e. HLA-DPB1 (−3.19), HLA-DRA (−1.97), HLA-DMA (−1.92), HLA-DRB1 (−1.71),



After table is shown below, you can download result download

No	Term	Category	Gene list	# of genes	p-value	q-value	Bonferroni	Expression
1	DNA replication	KEGG hsa03030	view	36	0	0	0	UP
2	Proteasome	KEGG hsa03050	view	42	0	0	0	UP
3	Aminoacyl-tRNA biosynthesis	KEGG hsa00970	view	41	1.332e-15	1.196e-13	7.94e-13	UP
4	Pyrimidine metabolism	KEGG hsa00240	view	95	1.317e-13	9.849e-12	7.848e-11	UP
5	Cytokine-cytokine receptor interaction	KEGG hsa04060	view	252	2.44e-12	1.564e-10	1.454e-09	UP
6	Nucleotide excision repair	KEGG hsa03420	view	44	3.257e-09	1.827e-07	1.941e-06	UP
7	Autoimmune thyroid disease − CHROM	KEGG hsa05320 − CHROM 6p21.3	view	31	4.234e-09	2.111e-07	2.523e-06	UP
8	RNA polymerase	KEGG hsa03020	view	28	5.032e-09	2.258e-07	2.999e-06	UP
9	Purine metabolism	KEGG hsa00230	view	155	1.469e-08	5.995e-07	8.757e-06	UP
10	Cell cycle	KEGG hsa04110	view	115	1.735e-08	6.489e-07	1.034e-05	UP
11	Autoimmune thyroid disease ∩ CHROM	KEGG hsa05320 ∩ CHROM 6p21.3	view	17	3.193e-08	9.554e-07	1.903e-05	DOWN
12	Viral myocarditis ∩ CHROM	KEGG hsa05416 ∩ CHROM 6p21.3	iew	17	3.193e-08	9.554e-07	1.903e-05	DOWN
						)2e-06	0.0001743	UP
						36e-05	0.0002904	UP
						39e-05	0.0003583	DOWN
	− CHROM	6p21.3				39e-05	0.0003758	UP
17	Systemic lupus erythematosus ∩ CHROM	KEGG hsa05322 ∩ CHROM 6p21.3	view	27	6.659e-07	1.494e-05	0.0003969	DOWN

Gene list : gene name (gene value)

HLA-DPB1(-3.19) HLA-DRA(-1.97) HLA-DMA(-1.94) HLA-DPA1(-1.92) HLA-DRB1(-1.71) HLA-E(-1.66) HLA-DMB(-1.54) HLA-DQA1(-1.04) HLA-DRB4(-0.9) HLA-F(-0.54) HLA-B(-0.48) HLA-DQB1(-0.36) HLA-G(-0.28) HLA-C(-0.27) HLA-A(-0.15) HLA-DOB(0) HLA-DOA(0)

Figure 1. Z-test results for the T-cell differentiation data set. See the text for an explanation and detailed options. If the user clicks 'view', the members of each significant annotation set as well as their scores are shown.

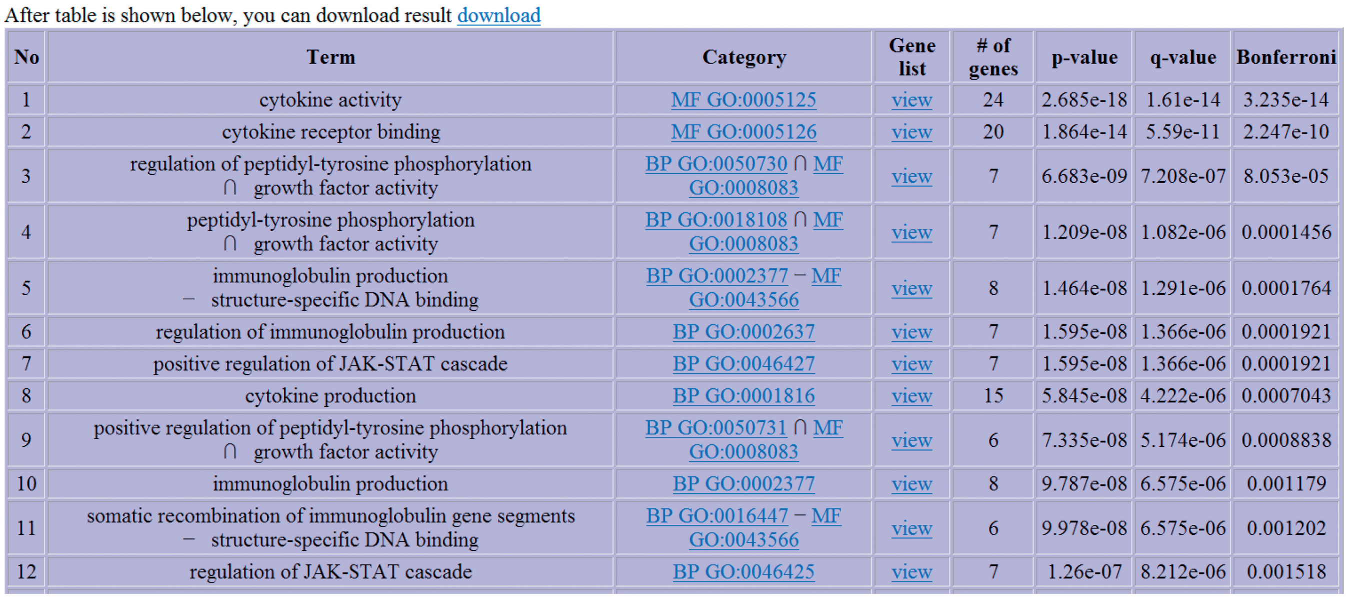
HLA-E (−1.66), HLA-DBM (−1.54)], significantly affecting the overall patterns of immune-related gene sets. See Figure 1 for the list of significant gene sets. This example illustrates how bi-directional expression patterns within a gene set can be described precisely using composite annotations.

We then selected the 200 most induced genes to test a GLA. Using Fisher's exact test and the three Gene Ontology categories, we interpreted the list. In the first trial, we chose the 'Strong Type' option, and we obtained a list of gene sets associated with the input list, many of which, as a single set, had strong relevance with T-cell differentiation. Examples include the cytokine-related processes 'JAK-STAT cascade', 'regulation of tyrosine phosphorylation', 'immunoglobulin production', 'regulation of T cell activation' and 'T-cell proliferation' (15,16). Additionally, some subtracted sets associated with 'ribosome' were detected. We then chose the 'Weak Type' to investigate more specific patterns within each single set. Figure 2 shows the computation results. Several intersected and subtracted gene sets appeared on top ranks. Most 'tyrosine phosphorylation' gene sets showed stronger patterns when intersected with 'growth factor activity'. For example, 'regulation of peptidyl-tyrosine phosphorylation' had a Q-value 1.377×10^{-4} in the strong-type analysis, but the intersection of 'regulation of peptidyl-tyrosine phosphorylation' and 'growth factor activity' had a much better Q-value of 7.208×10^{-7} in the weak-type analysis. The former single set originally contained 83 genes in total and actually included eight members from the input list, while the latter intersection set had a much smaller number of genes (25 in total), but included seven members from the input list. This feature makes the composite sets more precise descriptors of the enrichment patterns.

DISCUSSION AND CONCLUSION

ADGO 2.0 is currently a unique tool that supports GSA methods based on composite annotations. We added GLA methods to this new version. It provides several widely used GSA methods including fast GSEA (14). Several other tools [e.g. GENECODIS (11), ProfCom (9), and COFECO (10)] provide analyses via composite annotations only for GLA. GENECODIS and COFECO employ the same type of algorithm and focus on the intersection of two or more annotation gene sets, while ProfCom generates more general types of composite sets using Boolean set operations (up to five single sets). Shortcomings with these tools are that they display all the significant gene sets without filtering redundant information (GENECODIS and COFECO) or a partial list of gene sets identified by a greedy search algorithm (ProfCom). ADGO 2.0 generates composite gene sets based on Boolean operations of two overlapping single sets and screens composite sets that have redundant information for both GSA and GLA. We may also consider composite sets generated by three or more single sets as ProfCom does, but this will increase the computational complexity prohibitively and make it quite complicated to establish legitimate rules (e.g. inclusion and exclusion rules) to screen the redundant information.

Using our tool, we demonstrated how to incorporate and interpret significant composite annotations when analyzing microarray data and list of genes. Note that many significant composite terms are hard to interpret clearly. In most cases, we may not find evidence from the literature because complex biological patterns have been rarely explored so far. Therefore, our tool may be used for an explorative research. If a composite pattern is observed repeatedly across many data sets, it may be validated experimentally.



After table is shown below, you can download result download

No	Term	Category	Gene list	# of genes	p-value	q-value	Bonferroni
1	cytokine activity	MF GO:0005125	view	24	2.685e-18	1.61e-14	3.235e-14
2	cytokine receptor binding	MF GO:0005126	view	20	1.864e-14	5.59e-11	2.247e-10
3	regulation of peptidyl-tyrosine phosphorylation ∩ growth factor activity	BP GO:0050730 ∩ MF GO:0008083	view	7	6.683e-09	7.208e-07	8.053e-05
4	peptidyl-tyrosine phosphorylation ∩ growth factor activity	BP GO:0018108 ∩ MF GO:0008083	view	7	1.209e-08	1.082e-06	0.0001456
5	immunoglobulin production − structure-specific DNA binding	BP GO:0002377 − MF GO:0043566	view	8	1.464e-08	1.291e-06	0.0001764
6	regulation of immunoglobulin production	BP GO:0002637	view	7	1.595e-08	1.366e-06	0.0001921
7	positive regulation of JAK-STAT cascade	BP GO:0046427	view	7	1.595e-08	1.366e-06	0.0001921
8	cytokine production	BP GO:0001816	view	15	5.845e-08	4.222e-06	0.0007043
9	positive regulation of peptidyl-tyrosine phosphorylation ∩ growth factor activity	BP GO:0050731 ∩ MF GO:0008083	view	6	7.335e-08	5.174e-06	0.0008838
10	immunoglobulin production	BP GO:0002377	view	8	9.787e-08	6.575e-06	0.001179
11	somatic recombination of immunoglobulin gene segments − structure-specific DNA binding	BP GO:0016447 − MF GO:0043566	view	6	9.978e-08	6.575e-06	0.001202
12	regulation of JAK-STAT cascade	BP GO:0046425	view	7	1.26e-07	8.212e-06	0.001518

Figure 2. Enriched gene sets for a list of upregulated genes in T-cell differentiation. The weak-type filtering criterion is applied for significant composite sets.

One interesting future work with our tool may be to investigate the regulatory interactions between regulators (TF or miRNA) and pathways using gene expression profiles. Because sequence-based predictions of the targets of TF or miRNA inevitably include abundant false positives, taking the intersections of the putative target genes with other gene sets may be useful for exploring specific patterns in regulatory networks.

FUNDING

Basic Science Research Program through a National Research Foundation (NRF) grant funded by the Korean government (MEST) (No. 2011-0015107); National R & D Program for Cancer Control of the Ministry for Health and Welfare, Republic of Korea (1020360) (to S.-Y.K.). Funding for open access charge: NRF grant funded by Korean government (No. 2011-0015107).

Conflict of interest statement. None declared.

REFERENCES

1. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
2. Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
3. 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.
4. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
5. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–484.
6. Nam,D., Kim,S.B., Kim,S.K., Yang,S., Kim,S.Y. and Chu,I.S. (2006) ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, **22**, 2249–2253.
7. Antonov,A.V. and Mewes,H.W. (2006) Complex functionality of gene groups identified from high-throughput data. *J. Mol. Biol.*, **363**, 289–296.
8. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
9. Antonov,A.V., Schmidt,T., Wang,Y. and Mewes,H.W. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.
10. Sun,C.H., Kim,M.S., Han,Y. and Yi,G.S. (2009) COFECO: composite function annotation enriched by protein complex data. *Nucleic Acids Res.*, **37**, W350–W355.
11. Nogales-Cadenas,R., Carmona-Saez,P., Vazquez,M., Vicente,C., Yang,X., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.*, **37**, W317–W322.
12. Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
13. Tian,L., Greenberg,S.A., Kong,S.W., Altschuler,J., Kohane,I.S. and Park,P.J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.
14. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
15. Schones,D.E., Cui,K., Cuddapah,S., Roh,T.Y., Barski,A., Wang,Z., Wei,G. and Zhao,K. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.
16. O'Shea,J.J. and Murray,P.J. (2008) Cytokine signaling modules in inflammatory responses. *Immunity*, **28**, 477–487.