Archives Italiennes de Biologie, 151: 76-98, 2013.

Interpreting the gene expression microarray results: a user-based evaluation of 10 bioinformatics tools

E. MELISSARI, M. DI RUSSO, V. MARIOTTI, M. RIGHI, C. IOFRIDA, S. PELLEGRINI

Microarray Lab, Department of Surgical, Medical and Molecular Pathology and of Critical Area, University of Pisa, Italy

ABSTRACT

In recent years many tools have been developed to cope with the interpretation of gene expression results from microarray experiments. The effectiveness of these tools largely depends on their ease of use by biomedical researchers. Tools based on effective computational methods, indeed, cannot be fully exploited by users if they are not supported by an intuitive interface, a large set of utilities and effective outputs.

In this paper, ten tools for the interpretation of gene expression microarray results have been tested on eleven microarray datasets and evaluated according to eight assessment criteria: 1. interface design and usability, 2. easiness of input submission, 3. effectiveness of output representation, 4. efficacy of the downloaded outputs, 5. possibility to submit multiple gene IDs, 6. sources of information, 7. provision of different statistical tests and 8. supply of multiple test correction methods. Strengths and weaknesses of each tool are highlighted: a. to provide useful tips to users dealing with the biological interpretation of microarray results; b. to draw the attention of software developers on the usability of their tools.

Key words

Microarray • Pathway analysis • Data mining • Usability • Output representation

Abbreviations

DEG: Differentially Expressed Gene • GO: Gene Ontology

Background

Oligonucleotide and cDNA microarrays are high-throughput techniques that compare the mRNA expression of a large group of genes in different cells or experimental conditions. The final output of a microarray data analysis is a list of codes identifying differentially expressed transcripts. Interactions among genes, however, are not adequately represented by a list of codes. Therefore, the following step of a microarray experiment is the biological interpretation of results to formulate hypotheses on the molecular mechanisms underlying the investigated biological phenomenon.

To deal with this challenging task, users submit the complete list of differentially expressed genes (DEGs) to several bioinformatics tools that query specific databases of biological information. By these tools, the investigators usually expect to obtain: a) an exhaustive annotation of as many as possible DEGs, b) an effective visualization of the biological interactions among DEGs, c) some suggestions on what are the most important groups of interacting DEGs and the biological processes in which they are involved. To date, a tool that meets all these requirements does not exist and users have to cope with many different tools in order to obtain all the needed information.

Corresponding Authors: Dr. Erika Melissari, Dr. Silvia Pellegrini, Microarray Lab, Department of Surgical, Medical and Molecular Pathology and of Critical Area, University of Pisa, via Roma 55, 56126 Pisa, Italy - Email: erika.melissari@for.unipi.it, silvia.pellegrini@med.unipi.it

The first attempt to schematize biochemical interactions among genes, so-called "pathway", was done by Goto and colleagues (Goto et al., 1997) who used binary relations between genes or gene products as components of the network. They drew the first metabolic pathway thus creating the KEGG database (Ogata et al., 1999). In KEGG, any new evidence on gene interactions is subjected to a careful verification by curators.

A different approach was adopted by BioCarta founders in 2000. They developed one of the first examples of Web 2.0 technology, as information is submitted by users, who remain traceable, and is managed by the scientific community. BioCarta provides rules to govern user collaboration and guarantees two levels of information care. The first level is assigned to all contributors, who send feedbacks on the submitted pathways. The second level of information care is assured by "Gurus", a restricted community of certified contributors, who evaluate the accuracy of the most popular pathways.

Similarly to BioCarta, GenMAPP (Dahlquist et al., 2002; Salomonis et al., 2007) is a free stand-alone computer application working both as a database and software. GenMAPP organizes data in pathways, called MAPPs, and allows users to identify gene networks in their own microarray data. MAPPs are drawn based on textbooks, review papers and public databases like Ensembl.

Finally, lists of functionally related genes are also generated by the Gene Ontology Project (Gene Ontology Consortium, 2001). The Gene Ontology (GO) database derives from a joint effort of scientific community that created a dynamic framework and a unified vocabulary to annotate genes by gene products. The three ontologies, Molecular Function, Biological Process and Cellular Component, allow scientists to univocally annotate genes in different biological contexts. Although the aim of GO is not building biological networks of interactions, finding a consistent group of DEGs mapped in the same Biological Process category suggests that they are involved in the same molecular mechanism. The GO information often completes the picture provided by KEGG, BioCarta and GenMAPP in terms of biological interpretation.

"Pathway analysis" strictly refers to the analysis that uses the information contained in pathway databases like KEGG, BioCarta and GenMAPP. When GO information is used, the performed analysis is called "Functional analysis". Keeping in mind this distinction, hereafter the term "pathway analysis" will be used to indicate the analysis that exploits both pathway and GO information, as already done by other authors (Khatri et al., 2012).

Pathway analysis can be performed by two enrichment approaches (Dopazo, 2006; Huang et al., 2009): the cut-off dependent and the cut-off free approaches, respectively adopted by the over-representation (Khatri and Drăghici, 2005) and the functional class scoring (Goeman et al., 2004; Pavlidis et al., 2004) methods of analysis.

Tools adopting the first approach are focused on DEGs lists. They associate a value of statistical significance to the over-representation of groups of DEGs in some pathways or GO categories by using statistical tests (Khatri and Drăghici, 2005). Tools adopting the second approach use the complete list of gene expression data to score selected gene sets from KEGG, GO or MSigDB, filtered by nonspecific methods and ranked by their differential expression values. They calculate the gene set and DEGs enrichment scores and associate a value of statistical significance to them by sample randomization methods and Kolmogorov-Smirnov test, respectively (Nam and Kim, 2008).

In both the approaches, the statistically significant pathways or categories or gene sets are assumed to be the most important for the biological interpretation of microarray results. Universally valid statistical tests or multiple test correction methods do not exist and users have to choose them based on the experiment characteristics. For more details see (Dudoit et al., 2004; van der Laan et al.; 2004, Khatri and Drăghici, 2005; Khatri et al., 2012).

A completely different approach to build interaction networks among biological entities is provided by mining heterogeneous sources of information using data mining techniques (Jelier et al., 2003; Hoffmann et al., 2005; Krallinger et al., 2008; Jelier et al., 2011). We refer to this kind of analysis as "gene network analysis".

Data mining tools, through the analysis, extraction and visualization of information contained in free-text and/or heterogeneous sources, provide researchers with an integrated and effective view of the biological information emerging from the gene lists. MEDLINE is the primary source of published papers. A list of identifiers like gene names, keywords, etc. must be supplied to perform text mining by MEDLINE. Data mining tools usually annotate the submitted information using a variable number of databases to find synonyms and make the search as exhaustive as possible. The list of identifiers is then exploited by text mining tools to find documents and to tag in the text biological entities. Computational algorithms able to perform grammatical or morphological analyses are used to build interaction networks among entities. Databases of biological interactions, like MIPS (Mewes et al., 2011), BIND (Bader et al., 2003), HPRD (Prasad et al., 2009), IntAct (Aranda et al., 2010) and databases of pathways and ontologies are used as benchmarks to validate the literature networks from a biological point of view. At the same time, this information can be used to annotate the networks and provide a wide selection of biological evidences.

Many tools have been implemented to make easier the biological interpretation of microarray results. A full review of all the tools or classes of methods is beyond the purpose of this paper, as many technical reviews on this topic already exist (see Nam and Kim, 2008; Werner, 2008; Abatangelo et al., 2009; Huang et al., 2009; Jelier et al.; 2011, Faro et al.; 2012, Hung et al.; 2012, Khatri et al., 2012). However, up to date no author has examined these tools from the end-users point of view. Since a frequent issue in using these tools is the limited or incomprehensible accompanying documentation, this paper aims to provide to everyone dealing with the interpretation of microarray results a useful guide to make an appropriate choice among the available tools. Our purpose is also to illustrate to software developers strengths and weaknesses of tools pointed out by end-users, in order to connect users' requests and developers' strategies.

Ten freeware tools were selected, each compliant to five mandatory criteria (Table I). They represent different bioinformatics solutions to the biological interpretation of microarray data. Eight items (listed in Table II) were adopted to characterize and score each tool (see Supplementary data 1 for details on score calculation) and a hands-on evaluation using 11 DEG lists from as many microarray studies was performed (see Supplementary data 2 for details on DEG lists). In order to make the tool evaluations independent from a specific case study, the 11 datasets were selected from microarray experiments on different topics. For the data mining tools, several biological entities (keywords, gene names, etc) relevant to the experiments were chosen. Finally, we suggest a workflow to use the most effective tools among those selected.

Tools for the Biological Interpretation Task

Sixty tools were retrieved by screening 275 papers from PubMed and several links from Google. After removing tools adopting the same computational method or similar implementation solutions and those largely reviewed by other authors, 10 tools compliant with our mandatory criteria were selected: Pathway Miner (Pandey et al., 2004), WebGestalt-KEGG enrichment analysis (Zhang et al., 2005), KOBAS 2.0 (Wu et al., 2006), Pathway-Express (Draghici et al., 2007), KegArray (Wheelock et al., 2009), Onto-Express (Khatri et al., 2002) and WebGestalt-GO enrichment analysis (Zhang et al., 2005) for the pathway analysis, Coremine, FunDO (Osborne et al., 2009) and GNCPro (Liu et al., 2010) for the gene network analysis. A summary of Scores 1 and 2 totalized by each tool is shown in Tables III and IV. The complete tool evaluation is available in Supplementary data 3.

In the last few years, many tools for pathway analysis implemented the cut-off free enrichment approach. The strength of cut-off free tools is that they preserve the complete network of dependen-

Table I. - Mandatory criteria for tool selection
1. To be royalty-free.
2. To accept as input a list of DEGs.
3. To have a functional web-based application.
4. To require no programming skills to be used.
5. To be not specific to one organism (except for *Homo sapiens*) or a single microarray platform.

Table II. - Description and index definition of adopted evaluation items.

Characteristic	Description	Score
1) Interface Design and Usability	The ability of a tool to be intuitive or to allow users to learn to use it easily, that is how much the tool design is end-user oriented	Mean of users' scores
2) Easiness of Input Submission	Intuitiveness and flexibility in input data formatting and submitting	Mean of users' scores
3) Output Representation Effectiveness	"[…] despite the vastness of an information structure, the view must be small, moving around must not take too many steps and the route to any target must be discoverable […]" (31)	Mean of users' scores
4) Downloaded Output Usefulness	Usefulness of the downloaded report	Mean of users' scores
5) Opportunity to submit Multiple input ID Types	The use of a gene code conversion tool is not pre-required, as multiple input IDs are allowed	Normalized Number of multiple ID
6) Sources of Information	Databases from which the tool extracts information	Normalized Number of sources of information
7) Availability of different Statistic	Many statistics are available to allow users to make a choice depending on the size of the gene reference background of the analysis (13)	Normalized Number of statistical distribution available
8) Availability of Multiple Test Methods	Many methods are supplied to contain Type I error rate in assessing the statistical significance	Normalized Number of multiple test correction methods

cies among genes by submitting the full list of genes assayed by a microarray. However, they use re-sampling methods to achieve the statistical significance of gene-sets thus performing worse with very few samples. Concerning the experiments on the most studied organisms, *Homo sapiens*, *Rattus norvegicus* and *Mus musculus*, without taking into account the experimental design, the percentage of microarray studies in ArrayExpress realized with more than 20 samples are 40%, 37% and 25%, respectively. It follows that, for most of the available microarray experiments, the results provided by the cut-off free tools must be carefully interpreted. Moreover, microarray results obtained by complex experimental designs, like indirect comparisons or comparisons of more than two conditions, cannot be submitted to cut-off free tools due to their complex statistical framework.

Table III. - Summary of scores totalized by each tool using evaluation items 1)-4).
The values of Kendall's W (K's W), showed in round brackets for each item, were all significant and evidenced a good concordance among raters.

Tool Name		Interface Design and Usability (*K's W: 0.65*)	Easiness of Input Submission (*K's W: 0.74*)	Output Representation Effectiveness (*K's W: 0.55*)	Downloaded Output Usefulness (*K's W: 0.58*)	Score 1
Pathway Analysis Tools	*WebGestalt-KEGG*	3.5	3.75	3.75	3.5	**14.5**
	Pathway-Express	3.5	2.75	3.75	2	**12**
	Pathway Miner	3	3	2.75	3	**11.75**
	KOBAS 2.0	2	2.75	2	1	**7.75**
	KegArray	1.5	1.5	2.5	2	**7.5**
	WebGestalt-GO	3.5	3.75	3.75	3.5	**14.5**
	Onto-Express	3.25	3	3.5	3	**12.75**
Gene Network Analysis Tools	*GNCPro*	3.5	2	3	2.75	**11**
	Coremine	2.25	2.75	2.75	3	**10.75**
	FunDO	3.75	4	2	1	**10.75**

Table IV - Summary of scaled scores totalized by each tool using evaluation items 5)-8).
For each class, tools are ranked on descending Score 2. See Supplementary data 3 for detailed information.

Tool Name		ID Types	Sources of Information	Statistics	Multiple test correction methods	Score 2
Pathway Analysis Tools	*WebGestalt-KEGG*	4	3.5	0.8	4	**12.3**
	KOBAS 2.0	1.33	3	4	2.4	**10.73**
	Pathway-Express	3.73	0.25	2.4	3.2	**9.58**
	KegArray	1.6	0.25	0	0	**1.85**
	Pathway Miner	0.26	0.75	0.8	0	**1.81**
	WebGestalt-GO	4	0.25	0.8	4	**9.05**
	Onto-Express	1.86	0.25	1.6	3.2	**6.91**
Gene Network Analysis Tools	*Coremine*	1.06	4	0.8	0	**5.86**
	FunDO	0.53	0.25	0.8	0.8	**2.38**
	GNCPro	1.06	0.5	0	0	**1.56**

Cut-off dependent tools accept lists of DEGs as input and assume that there is no dependency among genes. This kind of approach could increase the number of false positive results. On the other hand, these tools manage all the microarray experimental designs and with small datasets produce more reliable results than cut-off free tools (see Goeman and Bühlmann, 2007; Rivals et al., 2007; Nam and Kim, 2008; Huang et al., 2009; Gatti et al., 2010).

We focused on web-based application tools since tool installation represents the first problematic step for many users. They are, indeed, discouraged when files, libraries or databases must be locally installed and manually connected or updated. The downside is that the usage of web-based application tools depends on the tool server maintenance downtime.

Pathway analysis tools

Pathway Miner

Pathway Miner is one of the first attempts at mining gene co-regulation networks. Annotation data available in Pathway Miner are updated from NCBI dbEST, Unigene and Entrez Gene.

Pathway Miner supplies a plain interface for microarray data or keyword submission. It accepts only the GenBank Accession Number code as input ID, the most used IDs by microarray suppliers.

It provides a clear results summary page reporting the number of mapped genes and pathways for each source of information. Detailed results are obtained browsing the pathway images, where mapped genes are highlighted. This visualization is effective to individuate groups of interacting DEGs and the automatically locally savable images provide a suitable analysis report.

For each source of information the downloaded report contains the GeneSymbol list and accession numbers of mapped genes. In addition, the complete web page is locally savable by the browser saving options, including all the hyperlinks to external databases.

The distinguishing strength of this tool is that it supplies pathway information from KEGG, BioCarta and GenMAPP, which are the databases to date available that contain information on biological pathways. However, no multiple test correction methods are provided, thus a priori hypotheses on pathways or functional categories are needed.

KOBAS 2.0

KOBAS 2.0 uses KEGG ORTHOLOGY as controlled vocabulary for gene annotation. It takes as input a list of gene codes or sequences in FASTA format that annotates using several databases. The chance to submit FASTA sequences is unique and represents the strength of KOBAS 2.0, concerning microarray studies on poorly annotated organisms.

KOBAS 2.0 results are obtained by the annotation and the identification steps. Although the interface is not fully intuitive and the annotation step does not accept gene IDs frequently used by microarray suppliers, the available heterogeneous sources of information provide an exhaustive annotation of the submitted gene list. The identification step provides the statistically enriched pathways by using as input the annotation step output. The available statistical test set is really complete. However, although the provided list of enriched pathways and biochemical interaction networks is exhaustive, no map representing the mapped genes is included. Moreover, all the hyperlinks to databases are removed from the locally saved tables and the user needs to rerun the analysis every time.

KegArray

KegArray was selected although it is a stand-alone Java application as it has been developed by KEGG developer team. This tool is designed for integrated analysis of KEGG EXPRESSION data. It also maps microarray results in KEGG pathways but only using KEGG EXPRESSION data as template. Details on the usage of its not fully intuitive interface are provided in the help documentation.

KegArray appears the most effective tool in mapping genes (see Table V), a characteristic really appreciated by users dealing with the biological interpretation task. As evidenced in (Drăghici et al., 2006), this ability is due to several elements including the pathway update frequency and the type of ID used to query the pathway database. KegArray accepts only KEGG GENE IDs that are not usual but that proved to be very effective for gene expression data, based on the performance evidenced in this paper. They are obtained by the KegArray ID conversion tool, using IDs from NCBI GI, Entrez Gene, UniGene, UniProt and IPI databases. We were not able to obtain the conversion of GenBank codes, which are among the most used gene ID.

The KegArray result is a list of pathways where groups of DEGs are mapped. Each pathway name is linked to the KEGG pathway image with the mapped genes highlighted. The output cannot be saved in file format. However, as no statistical test is provided, it should only be used in an exploratory manner since it does not give any suggestion on which are the most significant pathways.

Pathway-Express

Pathway-Express, one of the most advanced pathway analysis tools, associates to the pathways a statistic called "Impact Factor", based on a system biology approach. This approach schematizes pathways like networks and considers the different molecular effect of diverse interactions/reactions between genes and/or gene products, i.e. activation, ubiquitination, glycosylation, etc.

The calculated Impact Factor derives from:

a) the number of DEGs in the pathway compared to the total number of genes in it;
b) the extent of the displacement from the pathway signaling steady state;
c) the location of the mapped DEGs within the pathway topology.

a) determines how much the mapped genes are representative of the whole pathway; b) estimates the level of network perturbation due to gene expression changes and to the different molecular effects produced by diverse gene-gene interactions/reac-

Table V. - Normalized citation indices, percentage of mapped genes and Total Score.
For each class, tools are ranked on descending Total Score.

Tool Name		Normalized ISI Citation Index	Normalized Scopus Citation Index	Mapped genes (average %)	Total Score
Pathway Analysis Tools	*WebGestalt-KEGG*	35.16	30.66	30.27	**26.8**
	Pathway-Express	37.25	30.75	21.98	**21.58**
	KOBAS 2.0	10.6	9.4	NA	**18.44**
	Pathway Miner	7.71	6.57	28.79	**13.56**
	KegArray	4.8	2.8	37.76	**9.35**
	Web-Gestalt-GO	35.16	30.66	66.48	**23.55**
	Onto-Express	27.88	25.44	85.68	**19.66**
Gene Network Analysis Tools	*Coremine*	NA	NA	61.4	**16.61**
	FunDO	5.5	6	31.43	**13.13**
	GNCPro	NA	NA	49.88	**12.56**

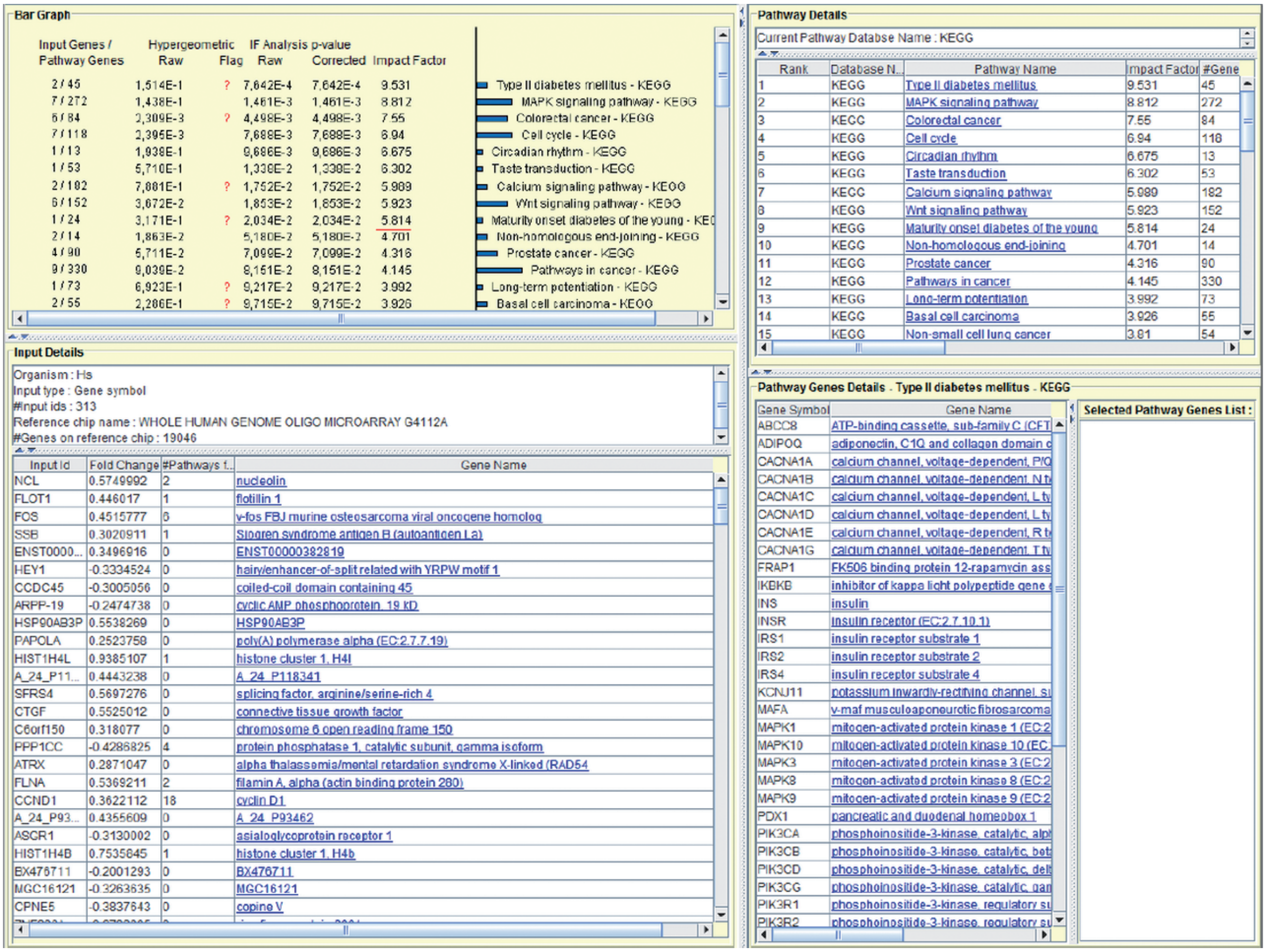
tions; c) considers that expression changes of genes upstream located in the pathway topology affect the pathway more than those of genes downstream located. To date, the Impact Analysis is the only technique that puts together statistical and biological features to provide an index of pathway perturbation due to gene expression changes.

Therefore, even if Pathway-Express implements a cut-off dependent approach, it considers dependencies among DEGs and for this reason it is more appropriately numbered among the Pathway Topology-based tools (Khatri et al., 2012).

The data submission interface of Pathway-Express is intuitive. Information about file format and ID types is available via a FAQ link. The list of genes assayed on custom or commercial arrays can be selected in the drop-down menu and used as reference for over-representation analysis. Many KEGG signaling pathways are exploited by this tool, whereas metabolic pathways are not considered. While the basic information submission does not require any particular expertise, setting advanced options is more complicated and not explained at all.

Many statistical distributions and multiple test correction methods are available.

The output consists of two windows. Four sub-windows form the first one and contain the pathway analysis results (Fig. 1), while the second one shows the retrieved pathways, in which the mapped DEGs are highlighted (Fig. 2). The sub-windows can be saved separately, but all the reciprocal links are lost. The pathway images can be locally saved in image format only one by one and losing the essential hyperlinks to KEGG. A saving option of the analysis as a project might be useful in order to not rerun the analysis every time.



Bar Graph

Input Genes /	Hypergeometric		IF Analysis p-value			
Pathway Genes	Raw	Flag	Raw	Corrected	Impact Factor	
2 / 45	1,514E-1	?	7,642E-4	7,642E-4	9.531	Type II diabetes mellitus - KEGG
7 / 272	1,438E-1		1,461E-3	1,461E-3	8.812	MAPK signaling pathway - KEGG
6 / 84	2,309E-3	?	4,498E-3	4,498E-3	7.55	Colorectal cancer - KEGG
7 / 118	2,395E-3		7,688E-3	7,688E-3	6.94	Cell cycle - KEGG
1 / 13	1,938E-1		9,686E-3	9,686E-3	6.675	Circadian rhythm - KEGG
1 / 53	5,710E-1		1,338E-2	1,338E-2	6.302	Taste transduction - KEGG
2 / 182	7,881E-1	?	1,752E-2	1,752E-2	5.989	Calcium signaling pathway - KEGG
6 / 152	3,672E-2		1,853E-2	1,853E-2	5.923	Wnt signaling pathway - KEGG
1 / 24	3,171E-1	?	2,034E-2	2,034E-2	5.814	Maturity onset diabetes of the young - KE
2 / 14	1,863E-2		5,180E-2	5,180E-2	4.701	Non-homologous end-joining - KEGG
4 / 90	5,711E-2		7,099E-2	7,099E-2	4.316	Prostate cancer - KEGG
9 / 330	9,039E-2		8,151E-2	8,151E-2	4.145	Pathways in cancer - KEGG
1 / 73	6,923E-1	?	9,217E-2	9,217E-2	3.992	Long-term potentiation - KEGG
2 / 55	2,286E-1	?	9,715E-2	9,715E-2	3.926	Basal cell carcinoma - KEGG

Input Details

Organism : Hs
Input type : Gene symbol
#Input ids : 313
Reference chip name : WHOLE HUMAN GENOME OLIGO MICROARRAY G4112A
#Genes on reference chip : 19046

Input Id	Fold Change	#Pathways f...	Gene Name
NCL	0.5749992	2	nucleolin
FLOT1	0.446017	1	flotillin 1
FOS	0.4515777	6	v-fos FBJ murine osteosarcoma viral oncogene homolog
SSB	0.3020911	1	Sjogren syndrome antigen B (autoantigen La)
ENST0000...	0.3496916	0	ENST00000382819
HEY1	-0.3334524	0	hairy/enhancer-of-split related with YRPW motif 1
CCDC45	-0.3005056	0	coiled-coil domain containing 45
ARPP-19	-0.2474738	0	cyclic AMP phosphoprotein, 19 kD
HSP90AB3P	0.5538269	0	HSP90AB3P
PAPOLA	0.2523758	0	poly(A) polymerase alpha (EC:2.7.7.19)
HIST1H4L	0.9385107	1	histone cluster 1, H4l
A_24_P11...	0.4443238	0	A_24_P118341
SFRS4	0.5697276	0	splicing factor, arginine/serine-rich 4
CTGF	0.5525012	0	connective tissue growth factor
C6orf150	0.318077	0	chromosome 6 open reading frame 150
PPP1CC	-0.4286825	4	protein phosphatase 1, catalytic subunit, gamma isoform
ATRX	0.2871047	0	alpha thalassemia/mental retardation syndrome X-linked (RAD54
FLNA	0.5369211	2	filamin A, alpha (actin binding protein 280)
CCND1	0.3622112	18	cyclin D1
A_24_P93...	0.4355609	0	A_24_P93462
ASGR1	-0.3130002	0	asialoglycoprotein receptor 1
HIST1H4B	0.7535845	1	histone cluster 1, H4b
BX476711	-0.2001293	0	BX476711
MGC16121	-0.3263635	0	MGC16121
CPNE5	-0.3837643	0	copine V

Pathway Details

Current Pathway Databse Name : KEGG

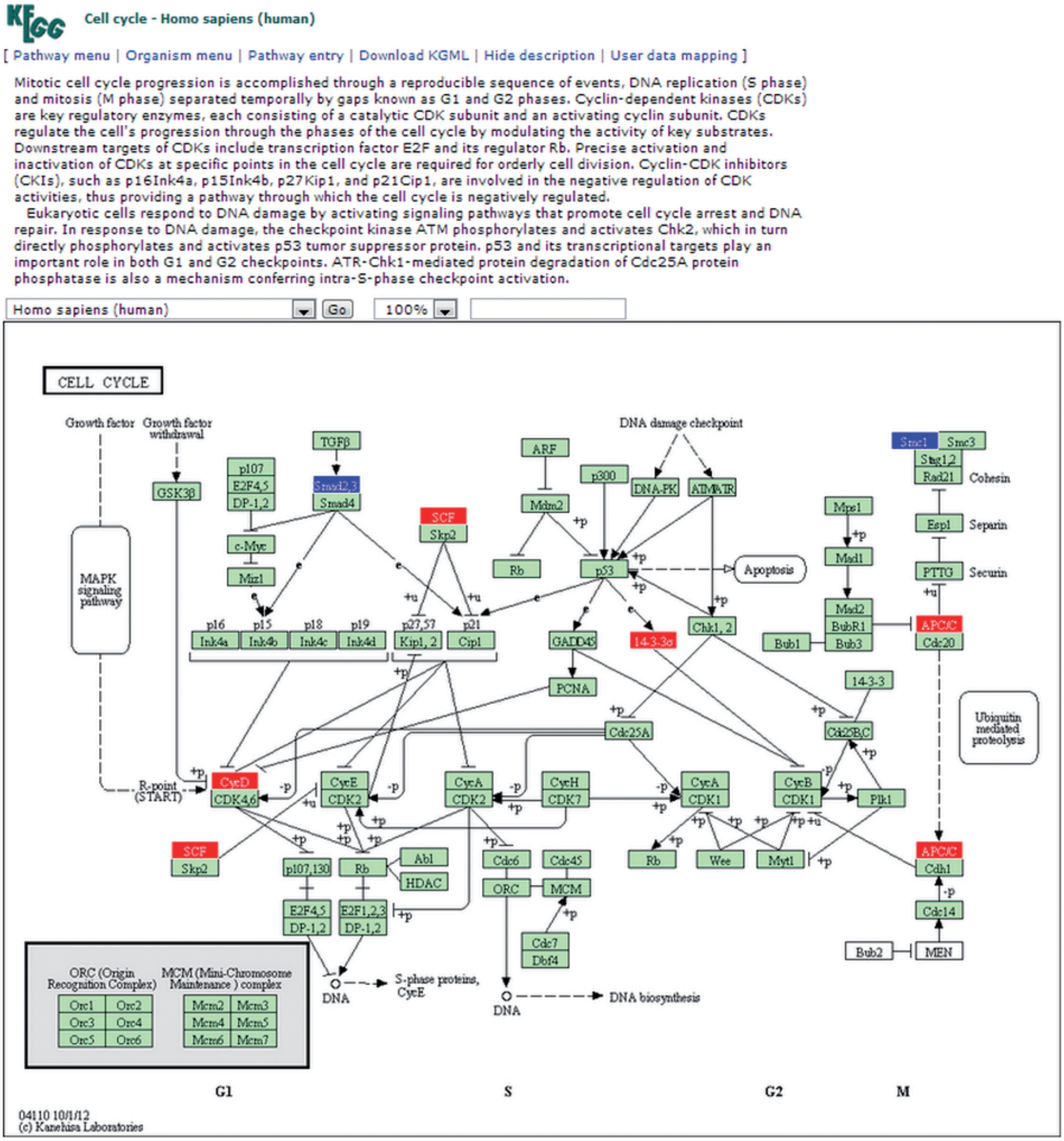
Rank	Database N...	Pathway Name	Impact Factor	#Gene
1	KEGG	Type II diabetes mellitus	9.531	45
2	KEGG	MAPK signaling pathway	8.812	272
3	KEGG	Colorectal cancer	7.55	84
4	KEGG	Cell cycle	6.94	118
5	KEGG	Circadian rhythm	6.675	13
6	KEGG	Taste transduction	6.302	53
7	KEGG	Calcium signaling pathway	5.989	182
8	KEGG	Wnt signaling pathway	5.923	152
9	KEGG	Maturity onset diabetes of the young	5.814	24
10	KEGG	Non-homologous end-joining	4.701	14
11	KEGG	Prostate cancer	4.316	90
12	KEGG	Pathways in cancer	4.145	330
13	KEGG	Long-term potentiation	3.992	73
14	KEGG	Basal cell carcinoma	3.926	55
15	KEGG	Non-small cell lung cancer	3.81	54

Pathway Genes Details - Type II diabetes mellitus - KEGG

Gene Symbol	Gene Name
ABCC8	ATP-binding cassette, sub-family C (CFT
ADIPOQ	adiponectin, C1Q and collagen domain c
CACNA1A	calcium channel, voltage-dependent, P/Q
CACNA1B	calcium channel, voltage-dependent, N t
CACNA1C	calcium channel, voltage-dependent, L ty
CACNA1D	calcium channel, voltage-dependent, L ty
CACNA1E	calcium channel, voltage-dependent, R t
CACNA1G	calcium channel, voltage-dependent, T ty
FRAP1	FK506 binding protein 12-rapamycin ass
IKBKB	inhibitor of kappa light polypeptide gene
INS	insulin
INSR	insulin receptor (EC:2.7.10.1)
IRS1	insulin receptor substrate 1
IRS2	insulin receptor substrate 2
IRS4	insulin receptor substrate 4
KCNJ11	potassium inwardly-rectifying channel, su
MAFA	v-maf musculoaponeurotic fibrosarcoma
MAPK1	mitogen-activated protein kinase 1 (EC:2
MAPK10	mitogen-activated protein kinase 10 (EC:
MAPK3	mitogen-activated protein kinase 3 (EC:2
MAPK8	mitogen-activated protein kinase 8 (EC:2
MAPK9	mitogen-activated protein kinase 9 (EC:2
PDX1	pancreatic and duodenal homeobox 1
PIK3CA	phosphoinositide-3-kinase, catalytic, alph
PIK3CB	phosphoinositide-3-kinase, catalytic, bet
PIK3CD	phosphoinositide-3-kinase, catalytic, delt
PIK3CG	phosphoinositide-3-kinase, catalytic, gam
PIK3R1	phosphoinositide-3-kinase, regulatory su
PIK3R2	phosphoinositide-3-kinase, regulatory su

Selected Pathway Genes List :

Fig. 1. - Output of *Pathway-Express*.
The four reciprocally linked sub-windows produced by *Pathway-Express* after the submission of dataset D8 are shown.



Cell cycle - Homo sapiens (human)

Mitotic cell cycle progression is accomplished through a reproducible sequence of events, DNA replication (S phase) and mitosis (M phase) separated temporally by gaps known as G1 and G2 phases. Cyclin-dependent kinases (CDKs) are key regulatory enzymes, each consisting of a catalytic CDK subunit and an activating cyclin subunit. CDKs regulate the cell's progression through the phases of the cell cycle by modulating the activity of key substrates. Downstream targets of CDKs include transcription factor E2F and its regulator Rb. Precise activation and inactivation of CDKs at specific points in the cell cycle are required for orderly cell division. Cyclin-CDK inhibitors (CKIs), such as p16Ink4a, p15Ink4b, p27Kip1, and p21Cip1, are involved in the negative regulation of CDK activities, thus providing a pathway through which the cell cycle is negatively regulated.

Eukaryotic cells respond to DNA damage by activating signaling pathways that promote cell cycle arrest and DNA repair. In response to DNA damage, the checkpoint kinase ATM phosphorylates and activates Chk2, which in turn directly phosphorylates and activates p53 tumor suppressor protein. p53 and its transcriptional targets play an important role in both G1 and G2 checkpoints. ATR-Chk1-mediated protein degradation of Cdc25A protein phosphatase is also a mechanism conferring intra-S-phase checkpoint activation.

Fig. 2. - A pathway map by *Pathway-Express*.
The map highlights the differentially expressed genes. Down-regulated genes are in blue and up-regulated genes in red.

Onto-Express

Onto-Express helps users to mine the available functional annotation data and find relevant functional categories. For each functional category, represented by a node in the trees schematizing the three ontologies, this tool calculates two different p-values. The first one is the collapsed p-value, calculated on the number of mapped genes, compared to those totally involved both in the considered node and its child nodes. The second one is the expanded p-value, calculated on the number of genes mapped, compared to those totally involved only in the considered node. This double statistical evaluation is the Onto-Express strength and allows users to decide the level of depth of their analysis. The interface design, file format, ID types and statistical options are similar to that of Pathway-Express. The full GO database or user selected ontologies can be set as reference for the analysis.

A full set of views is provided by Onto-Express to support users in result visualization and interpretation. Among the most informative there are the Tree View, the Flat View and the Single Gene View (Fig. 3).

For the Biological Process ontology, two locally savable analysis reports are available. The first one shows the tree structure and for each functional category lists the p-values and the number of mapped genes but not their names (IDs). The second one contains the same information plus the mapped genes for each functional category, but loses the tree structure. As both nesting levels and mapped genes are fundamental for the biological interpretation of DEGs, an integrated report would be more useful from an end-user point of view, especially if a large number of functional categories is statistically significant and user needs to frequently jump between the two reports to integrate the information.

WebGestalt

WebGestalt V2 is an intuitive tool that provides an easy way for researchers to have an integrated overview on large sets of genes. It consists of four modules: gene set management, information retrieval, organization/visualization and statistics.

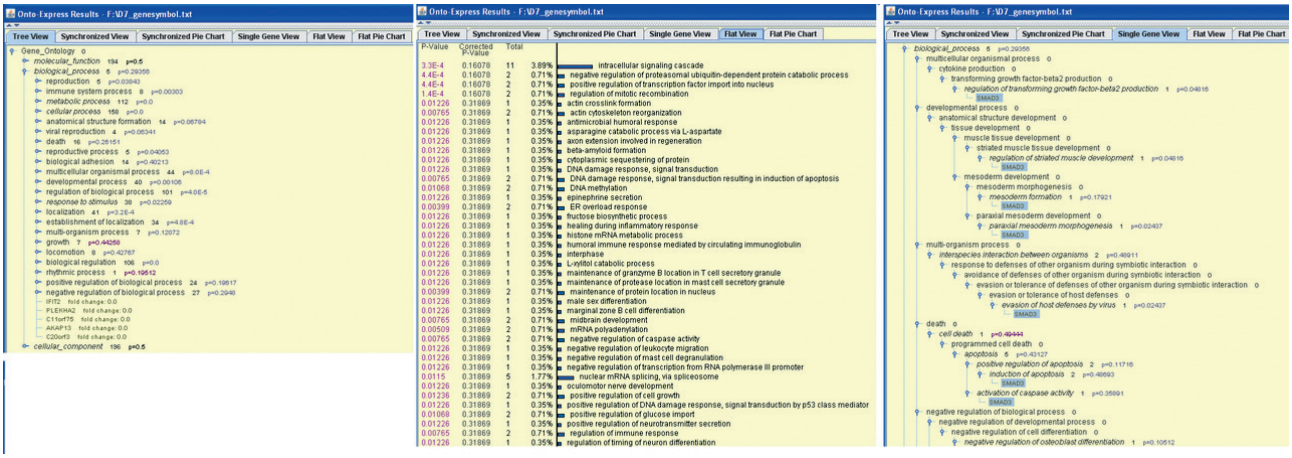
The gene set management module allows the user to easily upload lists of many different gene IDs and to manage them making the data formatting process simple. The information retrieval system retrieves annotations from up to 20 sources by the local database GeneKeyDB (Kirov et al., 2005). By exploiting the organization/visualization module, the annotated list can be subjected to several kinds of analysis, including Gene Ontology, tissue expression pattern, chromosome distribution, metabolic and signaling pathways and protein domain.

GO enrichment analysis performs the over-representation analysis of DEGs in GO categories and produces three directed acyclic graphs (Fig. 4). Each enriched category is linked to the list of mapped genes. This view is effective in helping users to retrieve at a glance the core information to be used for biological interpretation of the data. Details on the enriched categories and mapped genes are collected in a locally savable table.

Similarly, KEGG enrichment analysis performs the over-representation analysis of DEGs in KEGG pathways. For each statistically relevant pathway an exhaustive annotation of the mapped genes and a useful pathway visualization with the highlighted genes are reported. All these tables can be locally saved.

For both analyses the complete web pages can be saved, thus obtaining an effective report hyperlinked to all the available data sources.

Hypergeometric distribution and Fisher's exact test are supplied and users can select different signifi-



```
Gene_Ontology  0
  molecular_function  194  p=0.5
  biological_process  5  p=0.29356
    reproduction  5  p=0.03843
    immune system process  8  p=0.00303
    metabolic process  112  p=0.0
    cellular process  158  p=0.0
    anatomical structure formation  14  p=0.06784
    viral reproduction  4  p=0.06341
    death  16  p=0.26151
    reproductive process  5  p=0.04063
    biological adhesion  14  p=0.40213
    multicellular organismal process  44  p=8.0E-4
    developmental process  40  p=0.00106
    regulation of biological process  101  p=4.0E-5
    response to stimulus  38  p=0.02259
    localization  41  p=3.2E-4
    establishment of localization  34  p=4.8E-4
    multi-organism process  7  p=0.12072
    growth  7  p=0.44258
    locomotion  8  p=0.42767
    biological regulation  106  p=0.0
    rhythmic process  1  p=0.19512
    positive regulation of biological process  24  p=0.19617
    negative regulation of biological process  27  p=0.2948
    IFIT2  fold change: 0.0
    PLEKHA2  fold change: 0.0
    C11orf75  fold change: 0.0
    AKAP13  fold change: 0.0
    C20orf3  fold change: 0.0
  cellular_component  196  p=0.5
```

```
P-Value   Corrected  Total
          P-Value
3.3E-4    0.16078    11   3.89%  intracellular signaling cascade
4.4E-4    0.16078    2    0.71%  negative regulation of proteasomal ubiquitin-dependent protein catabolic process
4.4E-4    0.16078    2    0.71%  positive regulation of transcription factor import into nucleus
1.4E-4    0.16078    2    0.71%  regulation of mitotic recombination
0.01226   0.31869    1    0.35%  actin crosslink formation
0.00765   0.31869    2    0.71%  actin cytoskeleton reorganization
0.01226   0.31869    1    0.35%  antimicrobial humoral response
0.01226   0.31869    1    0.35%  asparagine catabolic process via L-aspartate
0.01226   0.31869    1    0.35%  axon extension involved in regeneration
0.01226   0.31869    1    0.35%  beta-amyloid formation
0.01226   0.31869    1    0.35%  cytoplasmic sequestering of protein
0.01226   0.31869    1    0.35%  DNA damage response, signal transduction
0.00765   0.31869    2    0.71%  DNA damage response, signal transduction resulting in induction of apoptosis
0.01068   0.31869    2    0.71%  DNA methylation
0.01226   0.31869    1    0.35%  epinephrine secretion
0.00399   0.31869    2    0.71%  ER overload response
0.01226   0.31869    1    0.35%  fructose biosynthetic process
0.01226   0.31869    1    0.35%  healing during inflammatory response
0.01226   0.31869    1    0.35%  histone mRNA metabolic process
0.01226   0.31869    1    0.35%  humoral immune response mediated by circulating immunoglobulin
0.01226   0.31869    1    0.35%  interphase
0.01226   0.31869    1    0.35%  L-xylitol catabolic process
0.01226   0.31869    1    0.35%  maintenance of granzyme B location in T cell secretory granule
0.01226   0.31869    1    0.35%  maintenance of protease location in mast cell secretory granule
0.00399   0.31869    2    0.71%  maintenance of protein location in nucleus
0.01226   0.31869    1    0.35%  male sex differentiation
0.01226   0.31869    1    0.35%  marginal zone B cell differentiation
0.00765   0.31869    2    0.71%  midbrain development
0.00509   0.31869    2    0.71%  mRNA polyadenylation
0.00765   0.31869    2    0.71%  negative regulation of caspase activity
0.01226   0.31869    1    0.35%  negative regulation of leukocyte migration
0.01226   0.31869    1    0.35%  negative regulation of mast cell degranulation
0.01226   0.31869    1    0.35%  negative regulation of transcription from RNA polymerase III promoter
0.0115    0.31869    5    1.77%  nuclear mRNA splicing, via spliceosome
0.01226   0.31869    1    0.35%  oculomotor nerve development
0.01236   0.31869    2    0.71%  positive regulation of cell growth
0.01226   0.31869    1    0.35%  positive regulation of DNA damage response, signal transduction by p53 class mediator
0.01068   0.31869    2    0.71%  positive regulation of glucose import
0.01226   0.31869    1    0.35%  positive regulation of neurotransmitter secretion
0.00765   0.31869    2    0.71%  regulation of immune response
0.01226   0.31869    1    0.35%  regulation of timing of neuron differentiation
```

```
biological_process  5  p=0.29356
  multicellular organismal process  0
    cytokine production  0
      transforming growth factor-beta2 production  0
        regulation of transforming growth factor-beta2 production  1  p=0.04816
          SMAD3
  developmental process  0
    anatomical structure development  0
      tissue development  0
        muscle tissue development  0
          striated muscle tissue development  0
            regulation of striated muscle development  1  p=0.04816
              SMAD3
        mesoderm development  0
          mesoderm morphogenesis  0
            mesoderm formation  1  p=0.17921
              SMAD3
          paraxial mesoderm development  0
            paraxial mesoderm morphogenesis  1  p=0.02437
              SMAD3
  multi-organism process  0
    interspecies interaction between organisms  2  p=0.48911
      response to defenses of other organism during symbiotic interaction  0
        avoidance of defenses of other organism during symbiotic interaction  0
          evasion or tolerance of defenses of other organism during symbiotic interaction  0
            evasion or tolerance of host defenses  0
              evasion of host defenses by virus  1  p=0.02437
                SMAD3
  death  0
    cell death  1  p=0.49444
      programmed cell death  0
        apoptosis  5  p=0.43127
          positive regulation of apoptosis  2  p=0.11716
            induction of apoptosis  2  p=0.48693
              SMAD3
          activation of caspase activity  1  p=0.35891
            SMAD3
  negative regulation of biological process  0
    negative regulation of developmental process  0
      negative regulation of cell differentiation  0
        negative regulation of osteoblast differentiation  1  p=0.10512
```

Fig. 3. - Tree View, Flat View of Biological Process ontology and Single Gene View of SMAD3 gene by *Onto-Express*. The Tree View displays the results in a tree diagram where the three principal branches represent the ontologies from GO Consortium. Each node can be expanded to find statistically significant terms and the mapped genes. The Flat View visualizes for a selected ontology the p-values and the percentage of DEGs by a bar graph. The Single Gene View shows GO terms where a selected gene has been found.

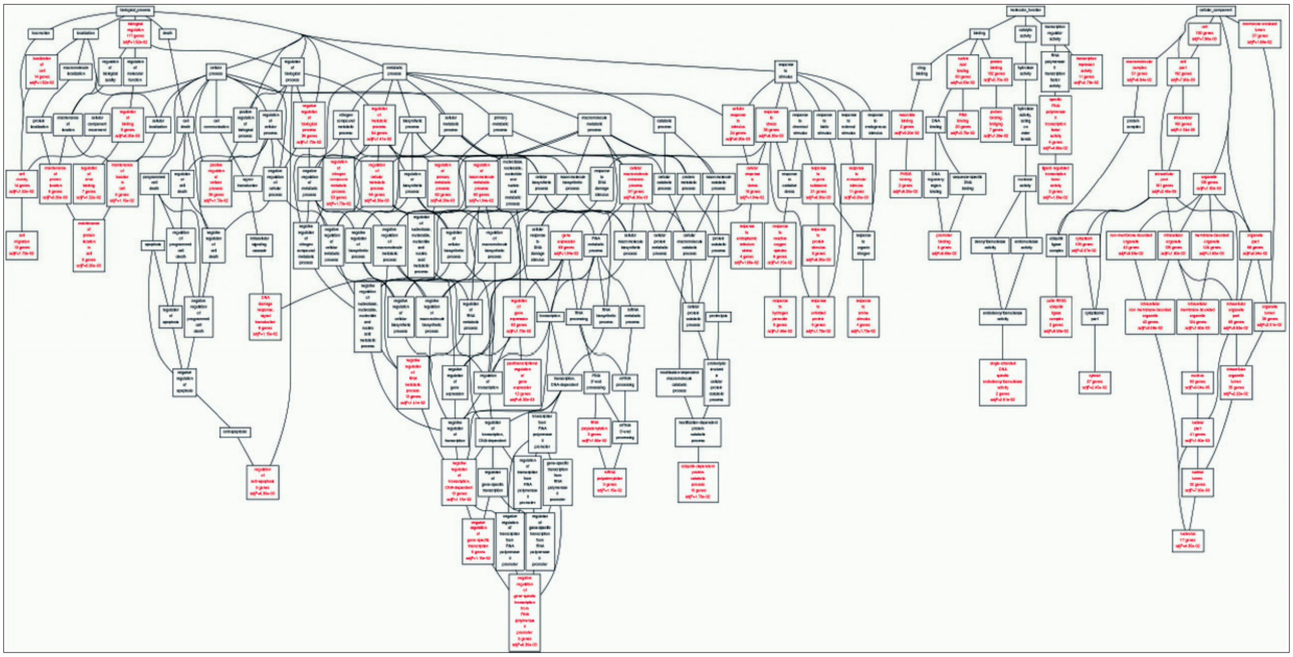


Fig. 4. - Directed Acyclic Graphs by *WebGestalt-GO enrichment analysis*.
The Directed Acyclic Graphs provided by *WebGestalt-GO enrichment analysis* after the submission of dataset D8. The statistically significant categories are highlighted in red.

cance levels for the statistical analysis and specify how many genes a category has to contain to be highlighted as statistically significant in the directed acyclic graphs.

Several multiple test correction methods are provided.

Gene Network Analysis Tools

FunDO

FunDO adopts text mining techniques to associate a list of DEGs with diseases from the Disease Ontology (DO) database. The association is performed by using the NCBI GeneRIF database (Mitchell et al., 2003) information, thus a greater number of gene-disease associations are available than that given by OMIM. This tool is mostly effective when the a priori hypotheses of the experiment concern a disease. Nevertheless, in the absence of any a priori hypothesis, the associations between genes and diseases suggest molecular mechanisms to further investigate by pathway analysis or other gene network analysis tools.

FunDO interface is plain and the list of human Entrez or GeneSymbol IDs is directly pasted in the text box input field. The output is a locally savable Java window consisting of two parts: a network, where the top five DO terms and the submitted genes associated to them are represented, and a table that summarizes the association results. The result view supplied by FunDO highlights what are the most represented DO terms in the submitted DEG list (Fig. 5). However, the edges are not linked to any URL providing the association by GeneRIF. The complete list of genes collected by GeneRIF and the subset of associated genes are shown in the table. The GeneSymbol codes are linked to their Entrez IDs. The over-representation of groups of submitted genes for each DO Term is evaluated by Fisher's exact test and corrected by Bonferroni method.

GNCPro

GNCPro is a free data integration and visualization tool, developed and maintained by SABiosciences (SABiosciences, Frederick, MD, USA). This tool creates networks of heterogeneous information built based on binary relationships between human genes. The considered relationships concern but are not limited to functional and transcriptional regulation, co-expressions, chemical modifications, physical interactions, interactions reported by literature. The results visualization is very effective in helping user to formulate biological hypotheses due to the eas-

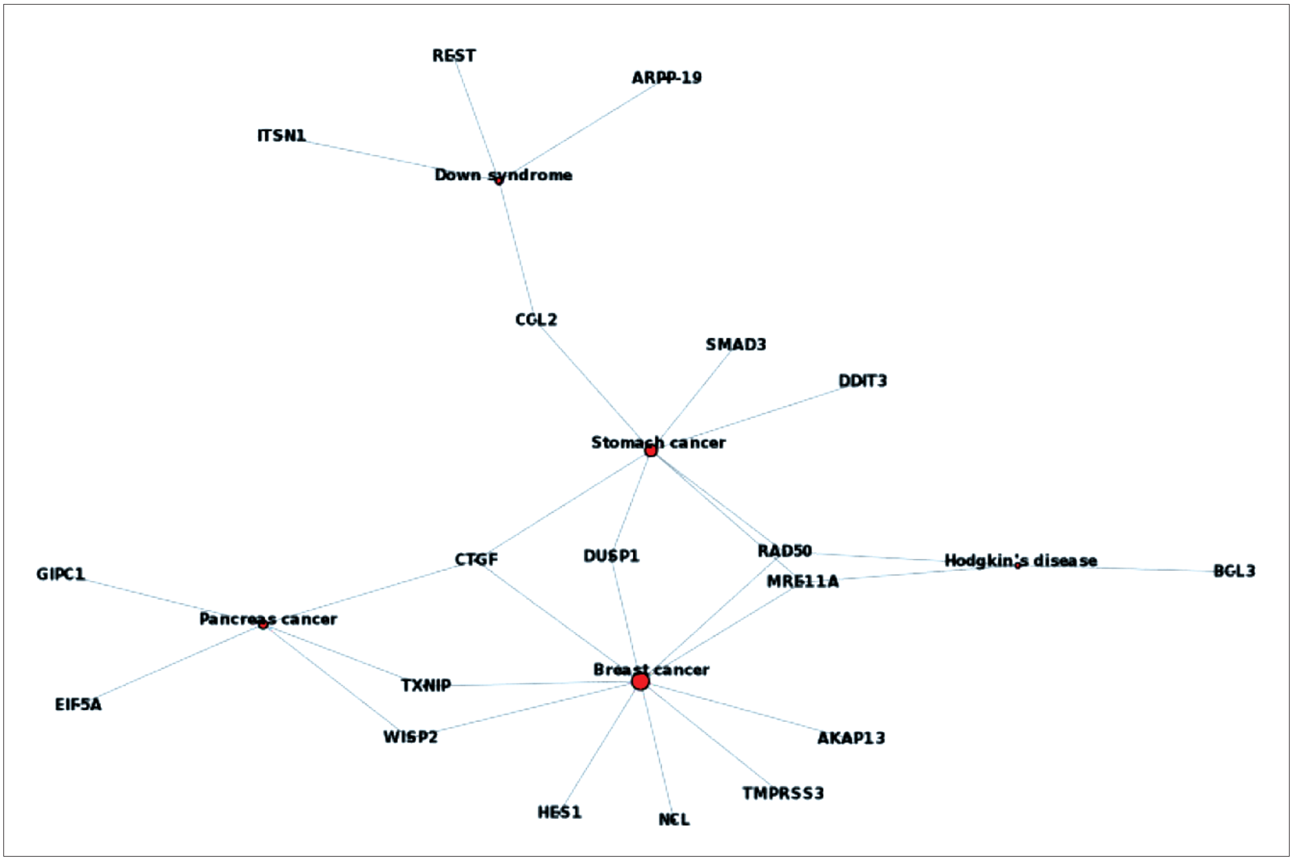


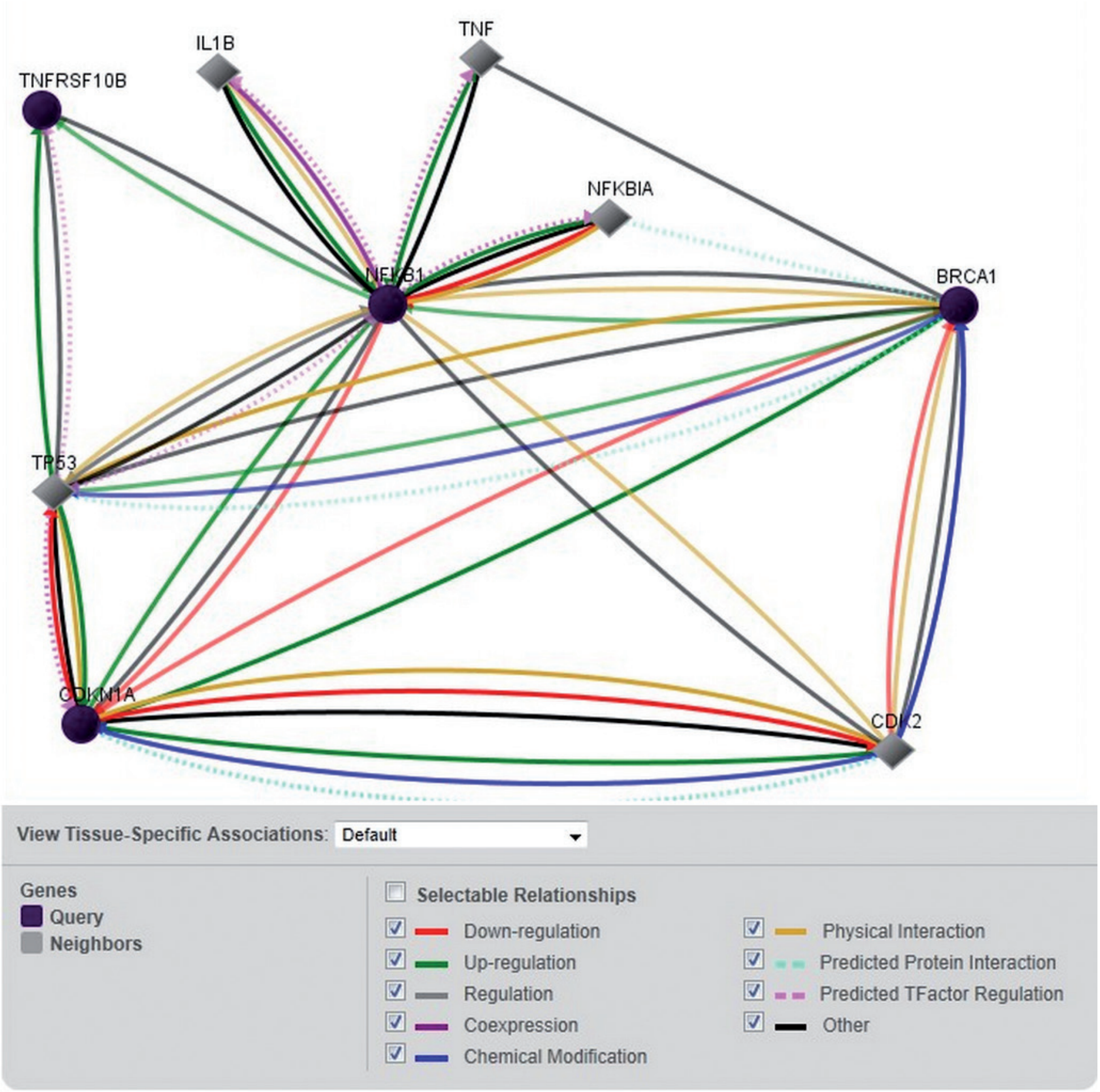
Fig. 5. - Gene-disease network by *FunDO*.
The diameters of the red spots representing DO terms suggest which are the most represented diseases, as they are proportional to the number of genes associated. Text mining analysis from *FunDo* shows that the submitted dataset D7 is primarily related to breast cancer.

ily retrievable information on groups of interacting genes, the types of relationships involved and the documentation confirming the association (Fig. 6).

GNCPro builds networks based on three types of techniques: text mining, data mining and data acquisition. Text mining splits PubMed abstracts into sentences and then into single words. A curated vocabulary of human genes is used to retrieve biological entities and a dictionary of words concerning interactions allows the identification of different kinds of relationships. An extensive manual checking of text mining results has been performed by the GNCPro team. Data mining and acquisition mine heterogeneous sources of information to build additional relationships among the submitted genes. Protein-protein interaction data from HPRD and ProLinks as well as microarray gene expression data from GEO are acquired to build interactions and annotate the network. Genes co-expression data are mined from GEMMA and from papers on microarray experiments. Predicted relationships are extracted by prediction algorithms, based on the evolution theory (Marcotte et al., 1999; Pellegrini et al., 1999).

GNCPro also suggests to the user new genes connected to those submitted (Fig. 6), which helps to bridge gaps among submitted genes and to formulate new hypotheses. Moreover, GNCPro makes available other biochemical networks, i.e. KEGG, Reactome, NCI and SABiosciences PCR array pathways, to provide further information on the submitted genes.

A set of utilities is supplied to carefully investigate each gene. Gene Network Central builds interaction networks starting from a single submitted gene. Gene Name Translator translates a gene ID in several gene codes. Gene ID Explorer summarizes NCBI Gene information. Tissue Expression Viewer provides the level of gene expression in several tissues. GNCPro interface is simple and allows users to



View Tissue-Specific Associations: Default

Genes
■ Query
■ Neighbors

☐ Selectable Relationships

☑ ▬ Down-regulation
☑ ▬ Up-regulation
☑ ▬ Regulation
☑ ▬ Coexpression
☑ ▬ Chemical Modification
☑ ▬ Physical Interaction
☑ ╍ Predicted Protein Interaction
☑ ╍ Predicted TFactor Regulation
☑ ▬ Other

Fig. 6. - Gene network by *GNCPro*.
The graphical representation of each relationship is an oriented colored edge located between two blue nodes that represent known submitted genes. Unknown submitted genes are indicated by red squares. Genes added by *GNCPro* are represented in the network as gray diamonds.

paste a list of codes into the text box input field. Up to 125 human gene codes can be submitted simultaneously, while no keyword is accepted in addition to the DEGs list. Although the network can be easily managed to draw a comprehensible view, the only possibility to exclude genes from visualization is to uncheck them one by one by the advanced options. Different kinds of information are exportable from the network: the annotated lists of submitted genes and interactions are the most useful. Finally, no statistical test is supplied to measure the strength of the interactions, thus making GNCPro a useful tool to essentially gather heterogeneous information.

Coremine

To date, no paper concerning Coremine has been published but some supporting documentation is available in the help page. Coremine derives from PubGene (Jenssen et al., 2001), which represents the first attempt to create a human gene-to-gene co-citation network by mining PubMed titles and abstracts. It annotates each gene by the MeSH (Medical Subject Heading) indices and the GO database. Assuming that a biological relationship between two genes co-mentioned exists, PubGene builds a curated literature co-occurrence network, where nodes represent the genes and edges the mined information.

The first important improvement of Coremine compared to PubGene is the number of sources and of biological entities that can be submitted. Coremine, indeed, uses 16 different sources of information to build a "map of concepts" among the biological entities derived from 13 different categories. Due to its ability to find and connect an impressive amount of information, Coremine is considered a full biomedical search engine able to perform an integrative analysis among heterogeneous biological entities.

Around 300 GeneSymbol IDs can be uploaded with an unlimited number of keywords from the available biological categories. When a DEG list is submitted, Coremine works either in a supervised or an unsupervised way to individuate issues for data interpretation. When one or more keywords are submitted together with a list of genes, Coremine links together genes, keywords and biological entities from categories chosen by the user. If no a priori assumptions are formulated and no other categories are selected, Coremine works in an unsupervised manner building the map of concepts among the genes. If no genes but only keywords are submitted, Coremine produces a map of concepts among them, the most co-mentioned genes and/or other biological entities from the selected categories.

Many tools are available. The Library tool allows user to build a map of concepts among the submitted biological entities by mining either all MEDLINE publications or a group of papers selected by the user. The Genomic HyperBrowser (Sandve et al., 2010) tool connects Coremine to Galaxy (Giardine et al., 2005), an open web-based research and analysis platform for biomedical data that allows users to perform an integrated analysis of biomedical data. It also provides lists of genes associated to the biological entities submitted by users. For each gene a p-value is calculated based on the binomial distribution and representing the gene-entity association strength. The File upload tool manages the submission of a DEG list to Coremine.

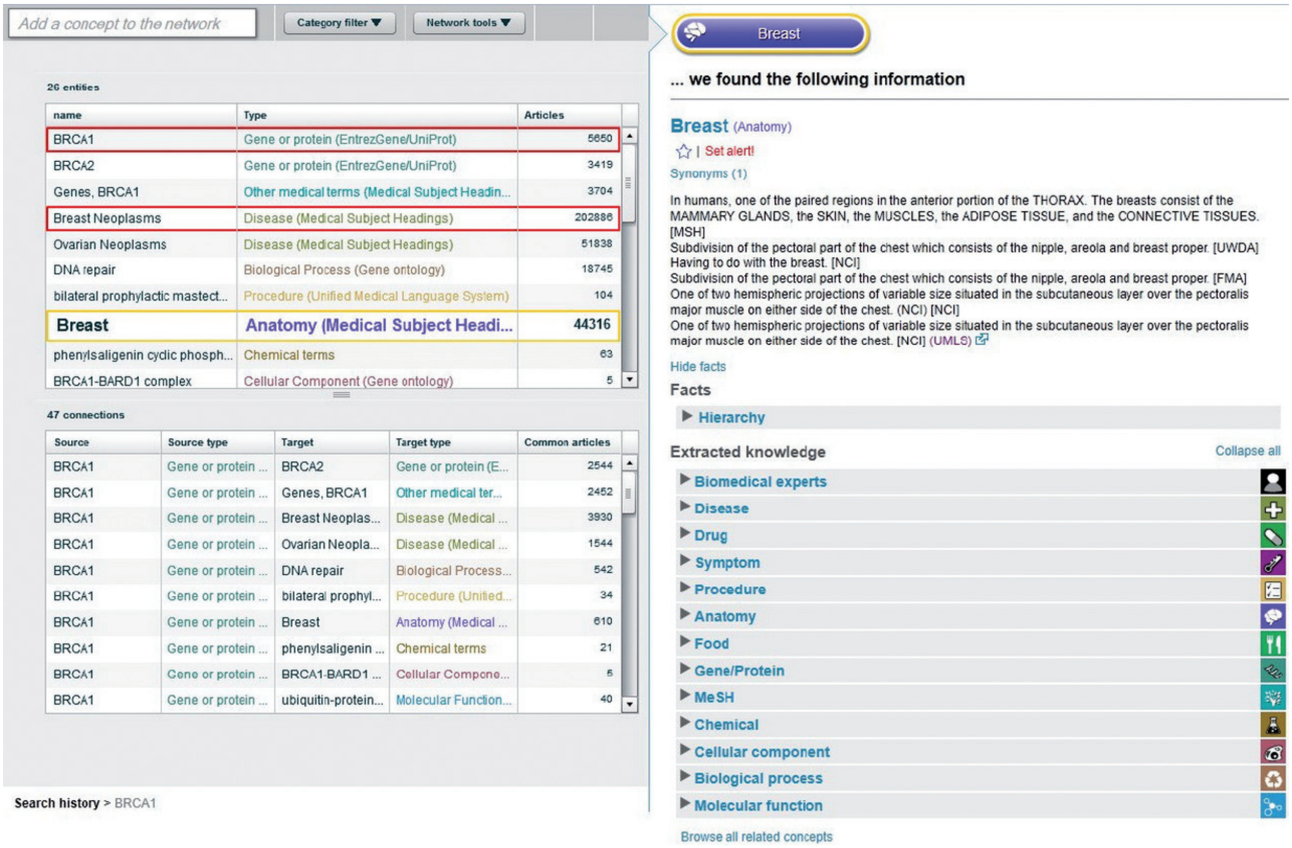
At a first glance, the Coremine interface may confuse users, as, to make a submission, many options seem necessary to select. However, for a simple search, is sufficient to type just a GeneSymbol or a part of a gene name or a keyword in the text box input field. Coremine helps users in the submission suggesting alternatives from the 13 available biological categories in the drop-down menu. However, if a quite targeted search is desired, it is advisable to uncheck the categories considered uninteresting for the study.

The Coremine output consists of two views: the network view and the grid view. In the network view, the submitted biological entities are represented by nodes and the gathered information by edges. By clicking on a specific node, the so-called Facts, short information from the NCBI databases are shown. To retrieve information underlining a connection between two nodes, the so-called Hits, all links to the 16 sources of information are shown. The network view is an effective graphical solution but it becomes difficult to manage when many biological entities are submitted. This limitation can be overcome by exploiting a priori information or hypotheses, i.e. suggested by the pathway analysis. The grid view is useful to summarize information from a crowded network. It divides into two tables showing the linked entities and the connections. As in the network view, each element is linked to Facts and Hits (Fig. 7).

Although the views are not automatically savable, the map of concepts can be saved as a project in the user's personal page, then can be reloaded and modified every time Coremine is accessed.

Performances in DEG mapping

The usefulness of a bioinformatics tool to formulate biological hypotheses depends on the body of retrieved information that, in turn, depends on the number of genes that the tool is able to annotate and map in networks of interactions. The number of



26 entities

name	Type	Articles
BRCA1	Gene or protein (EntrezGene/UniProt)	5650
BRCA2	Gene or protein (EntrezGene/UniProt)	3419
Genes, BRCA1	Other medical terms (Medical Subject Headin...	3704
Breast Neoplasms	Disease (Medical Subject Headings)	202886
Ovarian Neoplasms	Disease (Medical Subject Headings)	51838
DNA repair	Biological Process (Gene ontology)	18745
bilateral prophylactic mastect...	Procedure (Unified Medical Language System)	104
Breast	Anatomy (Medical Subject Headi...	44316
phenylsaligenin cyclic phosph...	Chemical terms	63
BRCA1-BARD1 complex	Cellular Component (Gene ontology)	5

47 connections

Source	Source type	Target	Target type	Common articles
BRCA1	Gene or protein ...	BRCA2	Gene or protein (E...	2544
BRCA1	Gene or protein ...	Genes, BRCA1	Other medical ter...	2452
BRCA1	Gene or protein ...	Breast Neoplas...	Disease (Medical ...	3930
BRCA1	Gene or protein ...	Ovarian Neopla...	Disease (Medical ...	1544
BRCA1	Gene or protein ...	DNA repair	Biological Process...	542
BRCA1	Gene or protein ...	bilateral prophyl...	Procedure (Unified...	34
BRCA1	Gene or protein ...	Breast	Anatomy (Medical ...	610
BRCA1	Gene or protein ...	phenylsaligenin ...	Chemical terms	21
BRCA1	Gene or protein ...	BRCA1-BARD1 ...	Cellular Compone...	5
BRCA1	Gene or protein ...	ubiquitin-protein...	Molecular Function...	40

... we found the following information

Breast (Anatomy)

☆ | Set alert!

Synonyms (1)

In humans, one of the paired regions in the anterior portion of the THORAX. The breasts consist of the MAMMARY GLANDS, the SKIN, the MUSCLES, the ADIPOSE TISSUE, and the CONNECTIVE TISSUES. [MSH]
Subdivision of the pectoral part of the chest which consists of the nipple, areola and breast proper. [UWDA]
Having to do with the breast. [NCI]
Subdivision of the pectoral part of the chest which consists of the nipple, areola and breast proper. [FMA]
One of two hemispheric projections of variable size situated in the subcutaneous layer over the pectoralis major muscle on either side of the chest. (NCI) [NCI]
One of two hemispheric projections of variable size situated in the subcutaneous layer over the pectoralis major muscle on either side of the chest. [NCI] (UMLS)

Fig. 7. - Grid view of *Coremine* output.
A summary of the interactions among biological entities is supplied.

mapped DEGs depends on: a) the number and kind of queried sources of information; b) their updating frequency; c) the methods to query the sources; d) the native ID type used by the tool to query the source of information. The average percentage of DEGs mapped by each of the investigated tool, for the 10 tested datasets, is reported in Table V.

The adoption of multiple tools for the analysis guarantees the fullest possible coverage of all the available information. For further details see (Drăghici et al., 2006).

What tool/tools to use?

Based on the scores (Tables III, IV and V), WebGestalt-KEGG enrichment analysis appeared to be the most effective pathway analysis tool, followed by Pathway-Express. However, only their joint utilization mapped in KEGG the 37% of DEGs on average. The addition of Pathway Miner mapped the 47% of DEGs on average. This extra 10% is because Pathway Miner queries BioCarta and GeneMAPP databases in addition to KEGG. However, although Pathway Miner totalized a high Score 1, that means a positive feedback by users, its inadequate Score 2 (see Table IV) highlights some technical limitations that should be taken into account.

KegArray showed good performance in mapping genes (38% on average), but totalized the lowest scores (see Tables III, IV and V). It should be used only as an exploratory tool, as no statistical tests are run and no suggestions on the most significant pathways are provided.

KOBAS 2.0 is the only tool that accepts sequences in FASTA format as input and identifies unknown genes from poorly studied genomes using BLAST algorithm. It appears a good performing tool if we consider its Total Score, but its high ranking is only due to Score 2 (see Table IV). Its low Score 1 suggests that KOBAS 2.0 does not address the end-user requirements, especially in terms of output effectiveness.

Onto-Express and WebGestalt-GO enrichment analysis performed in a comparable way (Tables III and IV), retrieving Biological Process GO terms for, respectively, 86% and 66% of DEGs on average. This difference is probably due to the different used GO databases (GO or GO slim see Supplementary data 3). In general, the GO tool better performances compared to the pathway analysis tools in mapping genes are due to the larger number of genes annotated in GO database.

No relevant difference among the tools for gene network analysis emerged considering their Score 1. However, the Score 2 indicated that Coremine outperformed FunDO and GNCPro in terms of availability of information sources (Table IV). Moreover, Coremine linked the 61% of DEGs on average to the submitted keywords. The statistical methods adopted by the gene network analysis tools, however, should be improved as appears from Table IV. To date, a framework for proper statistical evaluation of the associations between genes and literature concepts is still lacking, although some promising methods have been recently proposed (Jelier et al., 2011).

The normalized citation indices shown in Table V, utilized as a posteriori validation of our evaluation process, evidenced a gap similar to that identified by Score 1 and 2 between the best and the worst performing tools. Moreover, all these indices highlighted that generally the gene network analysis tools are less frequently used than the pathway analysis tools. Neither on the basis of our scores nor by considering the number of mapped genes, the best absolute performing tool or class of tools was individuated. All these bioinformatics instruments show strengths and limitations that can be overcome by their integrated use. By our experience, the combined use of WebGestalt-KEGG enrichment analysis and Pathway-Express, together with Coremine and Onto-Express plus WebGestalt-GO enrichment analysis, in a step-by-step workflow, guarantees the best results. To demonstrate the validity of our theory we applied this workflow of analysis to the D7 dataset, which refers to a microarray experiment performed in our lab (Iofrida et al., 2012). Fig. 8 shows the intersection among the genes from this dataset mapped by these tools.

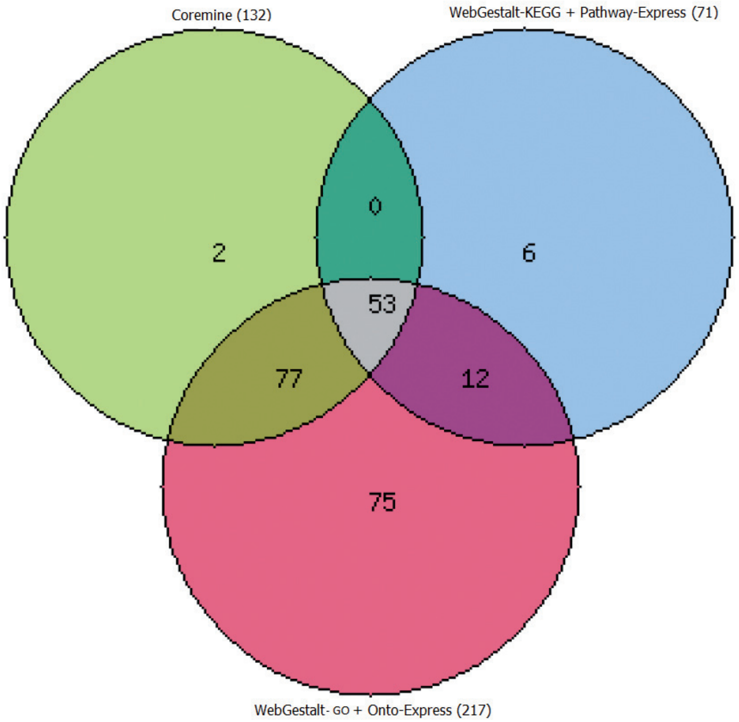


Fig. 8. - Intersection among the mapped genes.
The intersection among the genes mapped by the best performing tools is shown. The D7 dataset has been used to produce this comparison. *WebGestalt-KEGG enrichment analysis* and *Pathway-Express* results are represented in the blue circle, *Onto-Express* and *WebGestalt-GO enrichment analysis* results are in the red circle and *Coremine* results are in the green circle.

The joint use of WebGestalt-KEGG and Pathway-Express collected 71 mapped genes from D7 dataset. Fifty three out of these 71 were in common among all the used tools and may be used as core genes for a further map of concepts with Coremine, as well as the mechanisms suggested by the pathway analysis tools may be considered as preliminary biological entities. Seventy nine genes out of those annotated by Coremine were not mapped by the pathway analysis tools due to a lack of knowledge in information sources. Seventy seven of these genes, however, were recovered by the GO tools and indicated Biological Process categories that may be used as biological entities by Coremine.

Similar results were obtained with the other datasets suggesting that a joint use of these pathway and gene network analysis tools is essential to get an as comprehensive as possible biological interpretation of microarray data.

Conclusions

The biological interpretation of microarray results represents one of the most exciting challenges both for biologists and bioinformaticians. Producing knowledge from a list of genes in a fast and effective manner is a common target for these research areas. However, although the number of bioinformatics solutions increases weekly, there is still not a tool able to collect all the necessary information and to completely replace human skills in interpreting microarray results.

Pathway analysis tools are fundamental to gather molecular information emerging from microarray results. They are used to highlight the biochemical networks and to formulate hypotheses about the underlying molecular mechanisms. However, notwithstanding the continuous implementation of new tools, only a small number of known genes, shared by many pathways, is annotated in KEGG, BioCarta and GenMAPP, thus limiting the obtained information.

Extracting evidence from heterogeneous sources reveals additional biological insights that are not covered by mining only pathway and ontology databases. This information is becoming fundamental for the biological interpretation of microarray gene expression data and will be essential to interpret data obtained by the “Next Generation Sequencing” technologies (Cullum et al., 2011), like RNA-Seq (Costa et al., 2010). However, the blind use of data mining tools may be discouraging, as the quantity of information can be overwhelming and misleading, as the level of accuracy in not currently satisfying. For the biological interpretation of data, users need more friendly and accurate data mining tools to obtain complete information on the increasing number of topics uncovered by canonical pathway or ontology databases. The active engagement of content providers in making full text papers available, as well as of software developers in producing more effective and user-oriented solutions, is mandatory to increase the power of data mining tools.

In this regard, the BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) organization has been founded to evaluate text mining and information extraction systems applied to biological contexts. It also facilitates the development of new tools sharing information by Web 2.0 technologies similar to Wiki or social networks, as well as the building of additional sources like lexicons, terminology standards and ontologies, and the creation of new standardized methods for paper writing like automatically generated abstracts or summaries. Last but not least, the BioCreAtIvE organization focuses on the need of increasing quality and number of interactions between developers and users to narrow the gap between bioinformatics solutions and end-user requests.

Acknowledgements

EM was supported by a grant from the Italian Research Ministry (PRIN 2008). VM was supported by the Interdepartmental Research Center “E. Piaggio”, University of Pisa. CI was supported by IRIS Foundation (Castagneto Carducci, Livorno, Italy), a nonprofit organization for the rehabilitation of individuals with psychiatric disturbances.

References

Abatangelo L., Maglietta R., Distaso A., D’Addabbo A., Creanza T.M., Mukherjee S., et al. Comparative study of gene set enrichment methods. *BMC Bioinformatics*, **10**: 275, 2009.

Aranda B., Achuthan P., Alam-Faruque Y., Armean I., Bridge A., Derow C., et al. The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38** (Database issue): D525-531, 2010.

Bader G.D., Betel D., Hogue C.W. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31** (1): 248-250, 2003.

Costa V., Angelini C., De Feis I., Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-Seq. *J Biomed Biotechnol.*, **2010**: 853-916, 2010.

Cullum R., Alder O., Hoodless P.A. The next generation: using new sequencing technologies to analyse gene regulation. *Respirology*, **16** (2): 210-222, 2011.

Dahlquist K.D., Salomonis N., Vranizan K., Lawlor S.C., Conklin B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet.*, **31** (1): 19-20, 2002.

Dopazo J. Functional interpretation of microarray experiments. *OMICS.*, **10** (3): 398-410, 2006.

Draghici S., Khatri P., Tarca A.L., Amin K., Done A., Voichita C., et al. A systems biology approach for pathway level analysis. *Genome Res.*, **17** (10): 1537-1545, 2007.

Drăghici S., Sellamuthu S., Khatri P. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics.*, **22** (23): 2934-2939, 2006.

Dudoit S., van der Laan M.J., Pollard K.S. Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol.*, **3**: Article13, 2004.

Faro A., Giordano D., Spampinato C. Combining literature text mining with microarray data: advances for system biology modeling. *Brief Bioinform.*, **13** (1): 61-82, 2012.

Furnas G.W. Effective view navigation. In: ACM, editor. CHI'97 Human factors in computing systems, 1997.

Gatti D.M., Barry W.T., Nobel A.B., Rusyn I., Wright F.A. Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics.*, **11**: 574, 2010.

Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.*, **11** (8): 1425-1433, 2001.

Giardine B., Riemer C., Hardison R.C., Burhans R., Elnitski L., Shah P., et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15** (10): 1451-1455, 2005.

Goeman J.J., Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics.*, **23** (8): 980-987, 2007.

Goeman J.J., van de Geer S.A., de Kort F., van Houwelingen H.C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics.*, **20** (1): 93-99, 2004.

Goto S., Bono H., Ogata H., Fujibuchi W., Nishioka T., Sato K., et al. Organizing and computing metabolic pathway data in terms of binary relations. *Pac Symp Biocomput.*, 175-186, 1997.

Hoffmann R., Krallinger M., Andres E., Tamames J., Blaschke C., Valencia A. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE.*, **283**: pe21, 2005.

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

Hung J.H., Yang T.H., Hu Z., Weng Z., Delisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform.*, **13** (3): 281-291, 2012.

Iofrida C., Melissari E., Mariotti V., Guglielmi C., Guidugli L., Caligo M.A., et al. Effects on human transcriptome of mutated BRCA1 BRCT domain: a microarray study. *BMC Cancer.*, **12** (1): 207, 2012.

Jelier R., Goeman J.J., Hettne K.M., Schuemie M.J., den Dunnen J.T., 't Hoen P.A. Literature-aided interpretation of gene expression data with the weighted global test. *Brief Bioinform.*, **12** (5): 518-529, 2011.

Jelier R., Jenster G., Dorssers L.C., Van Mulligen E.M., Mons B., Kors J.A. Mining microarray datasets aided by knowledge stored in literature. *AMIA Annu Symp Proc.*, 879, 2003.

Jenssen T.K., Laegreid A., Komorowski J., Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet.*, **28** (1): 21-28, 2001.

Khatri P., Drăghici S., Ostermeier G.C., Krawetz S.A. Profiling gene expression using onto-express. *Genomics.*, **79** (2): 266-270, 2002.

Khatri P and Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics.*, **21** (18): 3587-3595, 2005.

Khatri P., Sirota M., Butte A.J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol.*, **8** (2): e1002375, 2012.

Kirov S.A., Peng X., Baker E., Schmoyer D., Zhang B., Snoddy J. GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments. *BMC Bioinformatics.*, **6**: 72, 2005.

Krallinger M., Valencia A., Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9 Suppl 2**: S8, 2008.

Liu G.G., Fong E., Zeng X. GNCPro: navigate human genes and relationships through net-walking. *Adv Exp Med Biol.*, **680**: 253-259, 2010.

Marcotte E.M., Pellegrini M., Thompson M.J., Yeates T.O., Eisenberg D. A combined algorithm for genome-wide prediction of protein function. *Nature.*, **402** (6757): 83-86, 1999.

Mewes H.W., Ruepp A., Theis F., Rattei T., Walter M., Frishman D., et al. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res.*, **39** (Database issue): D220-224, 2011.

Mitchell J.A., Aronson A.R., Mork J.G., Folk L.C., Humphrey S.M., Ward J.M. Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc.*, 460-464, 2003.

Nam D. and Kim S.Y. Gene-set approach for expression pattern analysis. *Brief Bioinform.*, **9** (3): 189-197, 2008.

Ogata H., Goto S., Sato K., Fujibuchi W., Bono H., Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27** (1): 29-34, 1999.

Osborne J.D., Flatow J., Holko M., Lin S.M., Kibbe W.A., Zhu L.J., et al. Annotating the human genome with Disease Ontology. *BMC Genomics.*, **10 Suppl 1**: S6, 2009.

Pandey R., Guru R.K., Mount D.W. Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics.*, **20** (13): 2156-2158, 2004.

Pavlidis P., Qin J., Arango V., Mann J.J., Sibille E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res.*, **29** (6): 1213-1222, 2004.

Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D., Yeates T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A.*, **96** (8): 4285-4288, 1999.

Prasad T.S., Kandasamy K., Pandey A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol Biol.*, **577**: 67-79, 2009.

Rivals I., Personnaz L., Taing L., Potier M.C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics.*, **23** (4): 401-407, 2007.

Salomonis N., Hanspers K., Zambon A.C., Vranizan K., Lawlor S.C., Dahlquist K.D., et al. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics.*, **8**: 217, 2007.

Sandve G.K., Gundersen S., Rydbeck H., Glad I.K., Holden L., Holden M., et al. The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol.*, **11** (12): R121, 2010.

van der Laan M.J., Dudoit S., Pollard K.S. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mol Biol.*, **3**: Article14, 2004.

Werner T. Bioinformatics applications for pathway analysis of microarray data. *Curr. Opin. Biotechnol.*, **19** (1): 50-54, 2008.

Wheelock C.E., Wheelock A.M., Kawashima S., Diez D., Kanehisa M., van Erk M., et al. Systems biology approaches and pathway tools for investigating cardiovascular disease. *Mol Biosyst.*, **5** (6): 588-602, 2009.

Wu J., Mao X., Cai T., Luo J., Wei L. KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34** (Web Server issue): W720-724, 2006.

Zhang B., Kirov S., Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33** (Web Server issue): W741-748, 2005.

Web

BioCarta: www.biocarta.com

BioCreAtIvE: http://biocreative.sourceforge.net/

COREMINE: www.coremine.com/medical/

dbEST: http://www.ncbi.nlm.nih.gov/nucest

Ensembl: http://www.ensembl.org/index.html

Entrez Gene: http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

GEMMA: http://www.chibi.ubc.ca/Gemma/home

Gene Ontology: http://www.geneontology.org/

GEO: http://www.ncbi.nlm.nih.gov//sites/entrez?db=geo

HPRD: http://nar.oxfordjournals.org/cgi/reprint/37/suppl_1/D767

KEGG: http://www.genome.jp/kegg

KEGG EXPRESSION: http://www.genome.jp/kegg/expression/

KEGG ORTHOLOGY: http://www.genome.jp/kegg/ko.html

MEDLINE: http://www.nlm.nih.gov/databases/databases_medline.html

MSigDB: http://www.broadinstitute.org/gsea/msigdb/index.jsp

NCI: www.cancer.gov

OMIM: http://www.ncbi.nlm.nih.gov/omim

ProLinks: http://prl.mbi.ucla.edu/prlbeta/prolinks.jsp

Reactome: http://www.reactome.org/ReactomeGWT/entrypoint.html

SABiosciences PCR array: http://www.sabiosciences.com/ArrayList.php?pline=PCRArray

Unigene: http://www.ncbi.nlm.nih.gov/unigene

Supplementary data 1

Selecting the Tools for Biological Interpretation of Microarray Results

Mandatory criteria for tool selection

The tools were retrieved by querying PubMed and the popular search engine Google with a series of Boolean queries.

The following mandatory criteria were then adopted to identify the most useful tools:

1. to be royalty-free;
2. to accept as input a list of DEGs;
3. to have a functional web-based application;
4. to require no programming skills to be used;
5. to be not specific to one organism (except for *Homo sapiens*) or a single microarray platform;

The selected tools were further analyzed to exclude "clone" tools, which are tools adopting the same computational method and similar implementation solutions.

Items for tool evaluation

The selected tools were evaluated on the basis of the following eight items: interface design and usability, easiness of input submission, effectiveness of the output presentation, usefulness of the downloaded outputs, opportunity to submit multiple types of input IDs, sources of information, availability of different statistics and multiple test correction methods. All the items are described in Table II.

Each tool was tested from a technical point of view by two bioinformaticians and by three biologists to gather their opinions as end-users. For item 1)-4) a questionnaire was drawn to detail the features evaluated by the raters. Each feature, formulated as negative sentences, was evaluated by using a discrete scale of four indices that quantify the rater's agreement to the feature. The index scale ranged from 1, meaning "I completely agree", to 4, meaning "I completely disagree". Kendall's W, as implemented in *concord* package by r-project, was adopted to evaluate the inter-rater agreement for each tool and for each item.

For each tool two Scores have been calculated:

- Score 1: the sum of evaluation indices (S_1) of items 1)-4);
- Score 2: the sum of suitably scaled (SS_2) evaluation indices (S_2) of items 5)-8).

The Score 1 indicates how much a tool is compliant to the end-user requirements. In order to calculate the average raters' agreement $(S_1)_j$ for a fixed item j, all the indices $(S_1)_i$ of raters' agreement were averaged, where i = 1,…,5 identifies the rater. Then, a tool totalizing an average index $(S_1)_j$ close to 4 can be considered compliant to the end-user expectations, whereas if that score is close to 1 the tool has to be considered insufficient by end-users. For each tool the Score S_1 equals the sum of $(S_1)_j$ scores, where j = 1,…,4 identifies the item.

The Score 2 summarizes how much information is available to the user and how much it is used to obtain the analysis results. In order to give the same range of representation to both the scores, the S_2 indices have been rescaled using the same range of Score 1 indices.

To obtain SS_2, the scaling factor (SF) was calculated by dividing the maximum index value of items 1)-4), that is 4, by the maximum index value collected from all the tools for each item 5)-8).

SF and SS_2 are defined as follows:

$$SF = 4/\max(S_2)_i$$
$$(SS_2)_i = SF * (S_2)_i$$

where i = 1,…,10 identifies the tool.

For example, if the tool *x* counts 14 sources of information and this is the maximum value from all the 10 tools, SF and $(SS_2)_x$ are 4/14 and 4, respectively.

A final score has been derived by summing the two Scores.

As the items 1)-4) may be subjective, for each tool ISI and Scopus citation indices were gathered. These indices were used as an *a posteriori* validation of our final ranking. The total number of citations was divided by the number of years elapsed from the paper publication to remove the citation bias. For each tool basic information concerning tool web-site and documentation was also collected.

The complete information is available in Supplementary data 3.

Concerning the hands-on assessment of the tools by a list of DEGs (details on the employed lists can be seen in Supplementary data 2), the performances of each tool in mapping genes were compared and, to make this final contrast as independent as possible from different statistical solutions, no statistical test or multiple test correction were used during the data submission. In order to make the performance in mapping genes independent by a specific case study, the 11 datasets have been taken from microarray experiments performed on different topics.

Onto-Translate (2) and MADGENE (3, 4) has been used to translate each gene code into the corresponding HUGO nomenclature before submitting the DEGs lists to the tools, except for *Pathway Miner* and *KOBAS 2.0*, for which GenBank Accession number and NCBI Entrez ID were respectively used. to eliminate any nomenclature bias…

Notes

[1] R-project. Available from: http://www.r-project.org/.

[2] Drăghici S., Sellamuthu S., Khatri P. Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22** (23): 2934-2939, 2006.

[3] MADGene. Available from: http://cardioserve.nantes.inserm.fr/mad/madgene/.

[4] Baron D., Bihouée A., Teusan R., Dubois E., Savagner F., Steenman M., et al. MADGene: retrieval and processing of gene identifier lists for the analysis of heterogeneous microarray datasets. *Bioinformatics*, **27** (5): 725-726, 2011.

Supplementary data 2

Datasets submitted to the 10 tools.

Dataset	Organism	Platform used	URL	# DEGs
D1	Homo sapiens	Affymetrix HG-U133A	http://www.ncbi.nlm.nih.gov/pubmed/17894889	185
D2	Homo sapiens	Illumina Sentrix BeadChip (Human-6v2)	http://www.ncbi.nlm.nih.gov/pubmed/21176028	1074
D3	Homo sapiens	Agilent Whole Genome	http://www.ncbi.nlm.nih.gov/pubmed/20406432	104
D4	Homo sapiens	GeneChip® HG-U133 Plus 2.0 arrays	http://www.ncbi.nlm.nih.gov/pubmed/18194544	2922
D5	Homo sapiens	Affymetrix HGU133A plus 2.0	http://www.ncbi.nlm.nih.gov/pubmed/19753302	386
D6	Homo sapiens	Agilent Whole Genome	http://www.ncbi.nlm.nih.gov/pubmed/22646717	173
D7	Homo sapiens	Agilent Whole Genome	http://www.ncbi.nlm.nih.gov/pubmed/22646717	201
D8	Homo sapiens	Agilent Whole Genome	http://www.ncbi.nlm.nih.gov/pubmed/22646717	313
D9	Homo sapiens	Agilent Whole Genome	http://www.ncbi.nlm.nih.gov/pubmed/17244347	58
D10	Homo sapiens	Affymetrix Human Genome U133 Plus 2.0	http://www.ncbi.nlm.nih.gov/pubmed/19596987	758
D11	Homo sapiens	Human Genome U133 Plus 2.0	http://www.ncbi.nlm.nih.gov/pubmed/18778695	1791

Supplementary data 3

Complete characterization of the 10 selected tools.

	Tool Name	Paper URL	Tool web-site	ISI Citation Index	Scopus Citation Index	ID types[a]	Sources of information[b]	Statistics[c]	Multiple test correction methods[d]	Species analyzed (KEGG organisms code)	Software version
Pathway Analysis Tools	WebGestalt-GO enrichment analysis	http://www.ncbi.nlm.nih.gov/pubmed/15980575	http://bioinfo.vanderbilt.edu/webgestalt/	211	184	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	44	1	1, 2, 3, 4, 5	hsa, mmu, rno, dre, cel, sce, cfa, dme	V 2
	WebGestalt-KEGG enrichment analysis						1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14				
	Pathway-Express	http://www.ncbi.nlm.nih.gov/pubmed/17785539	http://vortex.cs.wayne.edu/projects.htm#Pathway-Express	149	123	6, 11, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	15	1, 2, 3	6, 3, 4, 7	hsa, mmu, rno	V 6.0
	Onto-Express	http://www.ncbi.nlm.nih.gov/pubmed/17785539	http://vortex.cs.wayne.edu/projects.htm#Onto-Express	251	229	6, 9, 11, 12, 13, 14, 15	4	1, 2, 3	6, 3, 4, 7	ath, bta, cel, cjr, cal, cbu, dre, det, dme, gga, gsu, wgl, hsa, lma, lmf, mca, mmo, pst, rno, sce, spo, son, sil, tbr	V 6.0
	Pathway Miner	http://www.ncbi.nlm.nih.gov/pubmed/15145817	http://www.biorag.org/pathway.html	54	46	11	5, 27, 28	4	0	hsa, mmo	V 1.1
	KOBAS 2.0	http://www.ncbi.nlm.nih.gov/pubmed/16845106	http://kobas.cbi.pku.edu.cn/home.do	53	47	5, 26, 27, 28, 29	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26	1, 2, 3, 4, 5	1, 2, 8	All the organisms	V 2.0
	KegArray	http://www.ncbi.nlm.nih.gov/pubmed/19462016	http://www.genome.jp/download/	24	14	6, 7, 9, 10, 11, 27	5	0	0	All the organisms	V 1.2.3

Complete characterization of the 10 selected tools.

	Tool Name	Paper URL	Tool web-site	ISI Citation Index	Scopus Citation Index	ID types[a]	Sources of information[b]	Statistics[c]	Multiple test correction methods[d]	Species analyzed (KEGG organisms code)	Software version
Gene Network Analysis Tools	FunDO	http://www.ncbi.nlm.nih.gov/pubmed/19594883	http://django.nubic.northwestern.edu/fundo/	11	12	5, 6	31	1	3	hsa	NA
	GNCPro	http://www.ncbi.nlm.nih.gov/pubmed/20865508	http://gncpro.sabiosciences.com/gncpro/gncpro.php	NA	NA	5, 6, 11, 31	29, 30	0	0	NA	NA
	Coremine	NA	http://www.coremine.com/medical/	NA	NA	6, 32, 33, 34	1, 4, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43	3	0	hsa, ptr, mmo, rno, bta, ssc, cfa, gga, dra, dme, cel, osa, ath, sce, spo, eco	NA

[a]: 1) NCBI dbSNP, 2) Codelink, 3) Ensembl gene, 4) Ensembl peptide, 5) NCBI Entrez ID, 6) NCBI GeneSymbol, 7) IPI, 8) NCBI RefSeq, 9) NCBI UniGene, 10) UniProt, 11) GenBank Accession number, 12) GeneBank nucleotide ID, 13) GeneBank sequence ID, 14) LocusLink, 15) Affymetrix, 16) Agilent, 17) Amersham, 18) Clontech, 19) Illumina, 20) NIA, 21) Operon, 22) PerkinElmer, 23) Sigmagenosys, 24) Superarray, 25) Takara, 26) UNIPROTKB AC, 27) GI, 28) FASTA protein sequence, 29) FASTA nucleotide sequence, 30) KEGG gene ID, 31) NCBI GeneName, 32) MeSH, 33) GeneOntology, 34) Keywords in natural language.

[b]: 1) NCBI Gene, 2) NCBI GEO, 3) Ensembl, 4) GeneOntology, 5) KEGG PATHWAY, 6) IPI, 7) Pathways Common, 8) Wikipathways, 9) MGI, 10) SGD, 11) MSigDB, 12) Affymetrix, 13) Illumina, 14) NCBI dbSNP, 15) KEGG PATHWAY (session 2,3,4,5,6 only), 16) PID Curated, 17) PID BioCarta, 18) PID Reactome, 19) BioCyc, 20) Reactome, 21) Panther, 22) KEGG DISEASE, 23) GAD, 24) FunDO, 25) OMIM, 26) NHGRI, 27) BioCarta, 28) GenMAPP, 29) MEDLINE, 30) GEMMA, 31) Disease Ontology, 32) PubMed Chemical, 33) MeSH, 34) DrugBank, 35) Unified Medical Language System for Symptoms, 36) Unified Medical Language System for Procedure, 37) Food, 38) UniGene, 39) GoldemPath, 40) HomoloGene, 41) iProclass, 42)UniProt, 43) ENZYME, 44) GO slim.

[c]: 1) Hypergeometric, 2) Chi-square, 3) Binomial, 4) Fisher, 5) Frequency list.

[d]: 1) Benjamini and Hochberg, 2) Benjamini and Yekutieli, 3) Bonferroni, 4) Holm, 5) Hommel, 6) FDR, 7) Sidak, 8) Qvalue.