Literature-aided interpretation of gene expression data with the weighted global test

Jelier, R.; Goeman, J.J.; Hettne, K.M.; Schuemie, M.J.; Dunnen, J. den; Hoen, P.A.C. 't

BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 518–529
Advance Access published on 22 December 2010
doi:10.1093/bib/bbq082

Literature-aided interpretation of gene expression data with the weighted global test

Rob Jelier, Jelle J. Goeman, Kristina M. Hettne, Martijn J. Schuemie, Johan T. den Dunnen and Peter A.C. 't Hoen

Submitted: 29th August 2010; Received (in revised form): 26th November 2010

Abstract

Most methods for the interpretation of gene expression profiling experiments rely on the categorization of genes, as provided by the Gene Ontology (GO) and pathway databases. Due to the manual curation process, such databases are never up-to-date and tend to be limited in focus and coverage. Automated literature mining tools provide an attractive, alternative approach. We review how they can be employed for the interpretation of gene expression profiling experiments. We illustrate that their comprehensive scope aids the interpretation of data from domains poorly covered by GO or alternative databases, and allows for the linking of gene expression with diseases, drugs, tissues and other types of concepts. A framework for proper statistical evaluation of the associations between gene expression values and literature concepts was lacking and is now implemented in a weighted extension of global test. The weights are the literature association scores and reflect the importance of a gene for the concept of interest. In a direct comparison with classical GO-based gene sets, we show that use of literature-based associations results in the identification of much more specific GO categories. We demonstrate the possibilities for linking of gene expression data to patient survival in breast cancer and the action and metabolism of drugs. Coupling with online literature mining tools ensures transparency and allows further study of the identified associations. Literature mining tools are therefore powerful additions to the toolbox for the interpretation of high-throughput genomics data.

Keywords: *text mining; gene expression profiling; pathway testing; data integration*

BACKGROUND

Gene expression profiling has become an important technology in modern molecular biology, but the interpretation of gene expression data is still a challenging task. Many gene expression studies reveal expression changes in large numbers of genes. The characterization of these changes, for instance in terms of involved biological processes, remains difficult and requires an overview of the very large amount of information on gene function currently available. Gene annotation databases are commonly used in characterization efforts. They group genes according to a shared feature, such as involvement in the same biological process. KEGG [1], Biocarta

Corresponding author. Peter A.C.'t Hoen, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. Tel: +31 71 5269421; Fax: +31 71 5268285; E-mail: p.a.c.hoen@lumc.nl

Rob Jelier did his PhD on text mining at the Erasmus MC Rotterdam, afterwards he worked as a postdoc in LUMC's Department of Human Genetics. Currently he is a postdoctoral fellow at the EMBL-CRG Systems Biology Research Unit in Barcelona.

Jelle Goeman is associate professor in LUMC's Department of Medical Statistics and Bioinformatics, where he focuses on the development new statistical algorithms for high-throughput genomics technologies.

Kristina Hettne is a PhD student shared between the Department of Medical Informatics of the Erasmus MC Rotterdam and the Department of Health Risk Analysis and Toxicology of Maastricht University. She works on text mining applications for small molecules, including metabolites and drugs.

Martijn Schuemie is a senior researcher in the Department of Medical Informatics of the Erasmus MC Rotterdam and focuses on data and text mining in medical records.

Johan den Dunnen is full professor in LUMC's Department of Human Genetics, head of the Leiden Genome Technology Center, and dedicated to the implementation of new genomics technologies.

Peter A.C.'t Hoen is assistant professor in LUMC's Department of Human Genetics and focuses on the analysis and integration of genomics, proteomics and metabolomics data.

and NetPath [2] are among the best known catalogues for metabolic and signal transduction pathways. Alternative gene annotation schemes based on protein function, localization and expression regulation are available from Gene Ontology (GO) [3] and the molecular signature database (MSigDB) [4].

The gene annotation databases are intensively used to identify functional categories, represented by gene sets, which show an association with the gene expression data. A wide variety of methods have been proposed for this task (see ref. [5] for an overview). In brief, three different types of statistical tests can be discriminated: (i) tests for the overrepresentation of a gene set in a list of differentially expressed genes using a hypergeometric or equivalent test (see ref. [6] for an overview); (ii) methods that use the *P*-values of all the genes [7, 8]. Well known is the gene set enrichment analysis (GSEA), that uses ranked *P*-values and tests whether the ranks of genes in a gene set differ from a uniform distribution [4, 9]; (iii) regression analyses that use the actual expression levels of the genes in the gene set and test whether these are associated with the studied phenotype, an example is the global test [10–12].

A serious argument can be made in favor of this last type of tests. The resulting *P*-value has a clear interpretation in the context of the experiment (the probability there are no differentially expressed genes in the gene set), confounders can be included in the model and the test procedure can be generalized from a test for a single gene (a gene set with size one) to a gene set containing all genes. For a further discussion of methodological issues we refer to the review by Goeman and Bühlmann [13].

Automated gene annotation

The construction of gene annotation databases is mostly a manual process in which genes are annotated based on information in scientific publications [14]. Due to its labor-intensive nature, manual annotation efforts struggle to keep up-to-date [15] and focus on a limited subject area. Also, these databases provide a black and white view: a gene is either part of a category or not. This implies that some inclusion criterion must be used during the annotation, and that all genes are of similar importance to what the gene set represents. However, this may not accurately reflect biology and is not very flexible.

Automated text mining can complement manual approaches, as the automation can provide a broader scope, as well as that it more easily provides up-to-date information and adaptability. The field of text mining has grown rapidly in recent years, with several applications for gene expression data analysis. A selection of web-based tools is given in Table 1. The majority of tools work with gene lists. They can retrieve concepts or terms strongly associated to the selected genes, [16] and/or cluster the genes to retrieve functionally coherent subclusters [17–30]. The main differences between methods are on the following three aspects: First, which information is retrieved from the texts. One approach is to rely on the words of the texts directly [20, 29], with some approaches using the automatic combination of related words through a factor analysis to reduce the dimensionality [24, 31]. Another approach relies on a thesaurus and a tagging engine to identify thesaurus entries in texts [18, 25, 28]. Second, methods vary in how texts are linked to genes. Some tools rely on thesaurus based approaches to identify references to genes in texts [28, 30], whereas others rely on automated Pubmed queries [27, 29] or manual gene to document links such as provided by NCBI's Entrez gene [23, 31]. Third, the way the associations between terms and genes are calculated. Some approaches focus on direct co-occurrences between genes or concepts in documents [18], whereas other approaches allow indirect relations, e.g. two genes regularly co-occur with the same term, to play a role in a variety of ways [22, 32]. Apart from the tools that focus on retrieving gene relations from a list of genes, some methods retrieve functional associations shared between gene lists of different experiments [33, 34]. Finally, text mining can be used in combination with other resources in an integrated framework to retrieve functional associations between the genes [35].

Literature-based annotation and statistical testing

The analyses performed by the mentioned text mining-based approaches are of an exploratory nature, and do not provide a statistical evaluation for the identified associations in the context of the performed experiment. However, text mining algorithms can readily be combined with the three previously mentioned classes of statistical approaches for evaluating gene annotation categories. A class one approach could simply entail the creation of gene sets, for instance by applying a threshold on the literature derived association scores between genes and biomedical concepts. Sartor *et al.* [36] provide

Table 1: Useful websites for literature-aided interpretation of gene expression profiling data

Name	Description	Website
Anni	Versatile text mining tool. Exploration of associations used in the literature weighted globaltest.	www.biosemantics.org/anni
Babelomics	Platform for the analysis of transcriptomics, proteomics and genomic data with functional profiling.	babelomics.bioinfo.cipf.es
Biocarta	Pathway database.	www.biocarta.com
ConceptGen	An enrichment testing and concept mapping tool that includes MeSH-based gene sets	conceptgen.ncibi.org
CoPub	A text-mining based enrichment testing tool	services.nbic.nl/cgi-bin/copub3/CoPub.pl
GenCLiP	Clustering of gene lists by literature profiling and constructing gene co-occurrence networks	www.genclip.com
Gene Ontology	Controlled vocabulary for the functonal annotation of genes.	www.geneontology.org
Gene2MeSH	Contains co-occurrences between genes and MeSH terms	gene2mesh.ncibi.org
Genelist Analyzer	Statistical evaluation of overrepresentation of literature concepts in gene lists	http://workerbee.igb.uiuc.edu:8080/BeeSpace/Search.jsp
Global test	Homepage of the global test R package.	www.bioconductor.org/packages/release/bioc/html/globaltest.html
Hanalyzer	Gene network visualization and reasoning tool based on literature, ontology and database mining	hanalyzer.sourceforge.net
Literature weighted global test	Described in this paper	biosemantics.org/weightedglobaltest
KEGG	Metabolic and regulatory pathway database	www.genome.jp/kegg
MILANO	Automated searches in Medline for co-occurrence with gene-based search terms	milano.md.huji.ac.il
MSigDB	Molecular signatures database, gene sets that accompany the GSEA test	www.broadinstitute.org/gsea/msigdb
NetPath	Curated signal transduction pathways in humans	www.netpath.org
Pubgene	Contains module that displays literature co-occurrence networks	www.pubgene.org

literature-based gene sets in their tool ConceptGen, which uses Gene2MeSH (http://gene2mesh.ncibi.org) to identify gene and MeSH term pairs with a significantly higher number of co-occurrences than expected by chance. Frijters *et al.* [37] and Leong and Kipling [38] calculate biomedical term over-representation for a set of regulated genes in a similar fashion to standard class one over-representation tools. Several text-mining approaches have been published that resemble the earlier mentioned class two, GSEA-like approach. Kueffner *et al.* [39] integrate the rank of the genes after sorting on *P*-value with an analysis of the literature. However, their approach is based on factorization which complicates the interpretation of their results, and does not include formally testing retrieved associations. Minguez *et al.* [40] test if a ranked list of genes shows a significant correlation with the genes' associations to a biomedical term. These associations are based on the literature and reflect the extent to which rate a gene and a biomedical term occur together in documents exceeds the rate expected by chance.

However to our knowledge, no class three, or regression analysis based text mining tool has been published. Here we introduce the literature-weighted global test to use text mining-derived associations in combination with a regression based analysis of gene expression changes.

The literature-weighted global test

The literature-weighted global test is able to identify biomedical concepts associated with gene expression changes in genome-wide expression studies. The approach integrates previously developed text mining approaches [26, 28, 41] with the global test [10, 11, 42], a statistical framework to evaluate if a set of genes shows significant changes in gene expression.

In our framework, the sources of textual information are abstracts from MEDLINE, a bibliographical database. We use a thesaurus to identify textual references to biomedical concepts in the texts. Concepts have a definition, a list of synonymous terms and can be linked to, for instance, online databases. In the thesaurus, concepts are grouped by semantic categories such as 'gene', 'drug' or 'neoplastic process' and this grouping can be used to select interesting sets of biomedical concepts. After the identification of concepts in texts, we let concepts be

represented by the set of documents in which they are mentioned. Subsequently, we use so-called concept profiles to characterize the textual information associated to concepts. A concept profile is a list of concepts with for every concept a weight to indicate its importance.

The vector product of two concept profiles is a measure for the strength of the association between two concepts. Before, association scores between concept profiles have successfully been used to infer functional associations between genes [22, 26] and between genes and GO codes [41], to infer novel genes associated with the nucleolus [43], and to identify new uses for drugs and other substances in the treatment of diseases [44]. We have developed Anni [28] to provide a versatile and user-friendly tool to work with concept profiles (www.biosemantics.org/anni). The tool can be used for a wide variety of queries, such as finding functional associations between genes or retrieving all the genes associated to a disease, and can also serve as a literature-based knowledge discovery tool.

The global test brings a whole framework for statistical testing, including analytical graphs and the ability to test for several types of response variables, such as two state, multi state and continuous variables, as well as survival. We propose to incorporate text-mining derived information in the test, by weighing the participation of genes in the test based on the match of their concept profiles with the concept profile of a biomedical concept (see the Supplementary Data for implementation details).

Figure 1 illustrates how the different resources are used and interact. The input for the global test is a

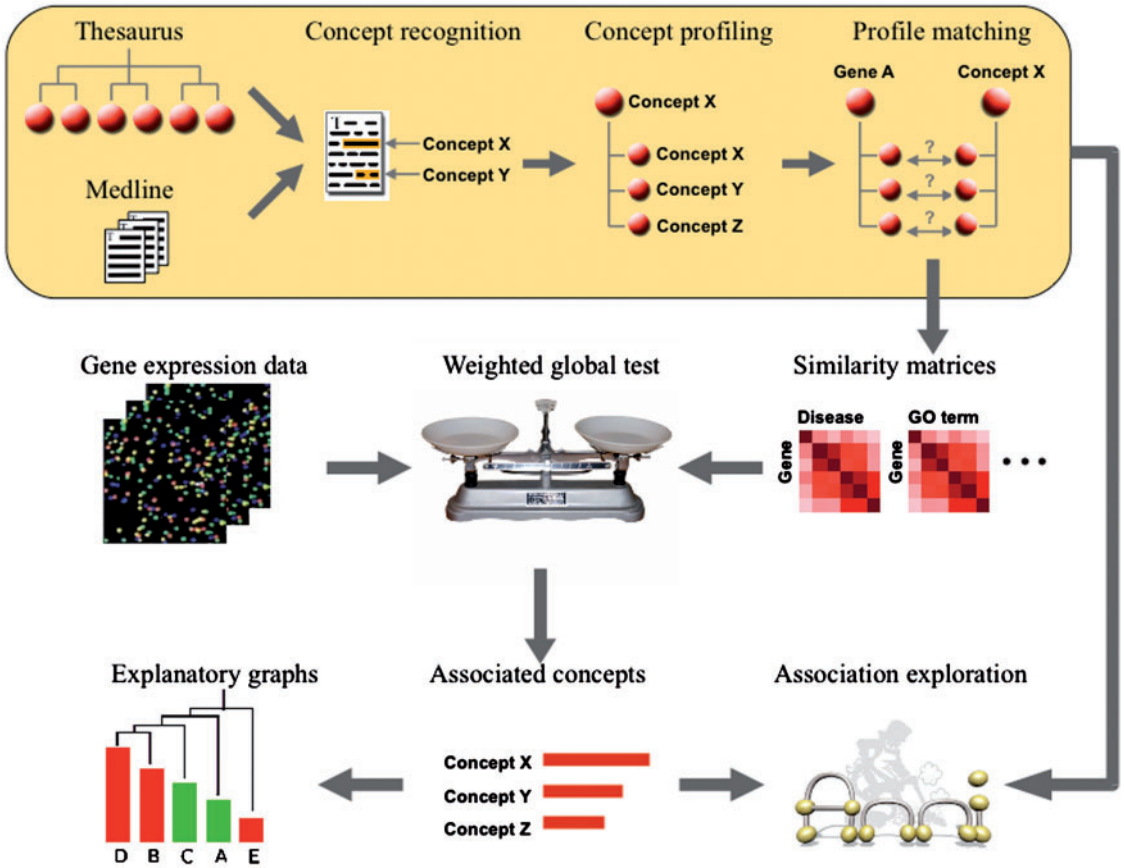

Figure 1: Overview of the literature-weighted global test framework (Description in paragraph 'The literature-weighted global test'). Concepts are characterized by concept profiles, reflecting the literature context in which a concept is mentioned. Association scores reflect the overlap in concept profiles are used to calculate association score between genes and other concepts. These association scores serve as the weights for the literature weighted global test (represented by the balance). The literature-weighted global test calculates a test-static and estimates a *P*-value for every concept evaluated and provides diagnostic plots to study the contribution of individual genes to the test statistic. Literature evidence underlying the inferred associations between a gene and a biomedical concept can be studied with the online tool Anni (www.biosemantics.org/anni).

data set of appropriately normalized gene expression measurements and the definition of the experimental variable (e.g. survival time of subjects). The literature-derived association scores used to weight the participation of genes in the test are provided in matrices that can be downloaded from our regularly updated website (http://biosemantics.org/weightedglobaltest). The literature-weighted global test calculates a test statistic and a *P*-value for every biomedical concept tested. Diagnostic plots are available, for instance to study the contribution of individual genes to the test statistic. The literature evidence underlying the inferred associations between a gene and a biomedical concept can be studied with our online tool Anni [28].

Below we will illustrate the approach, and compare it to a standard GO analysis, by analyzing three data set s in different biomedical domains: (i) cardiac arrhythmia, a biomedical domain that is poorly covered by GO; (ii) breast cancer metastasis, where we demonstrate the use of patient survival data; (iii) peroxisome proliferator alpha function, where gene expression profiles are linked to drug metabolism.

Table 2: The top 25 most significant concepts retrieved by the literature weighted global test for the category 'Pathological Function' on the *TBX3* data set

Rank	Concept		*P*-value
1	Chronic dilatation	*	1.91E-05
2	Glycogen depletion		2.07E-05
3	Neonatal bradycardia	*	3.10E-05
4	Global developmental delay		3.72E-05
5	Electrolyte imbalance	*	4.45E-05
6	Fatty infiltration		4.64E-05
7	Abnormal cardiac conduction	*	5.01E-05
8	Cardiac arrhythmia	*	5.14E-05
9	Ventricular Couplet	*	5.75E-05
10	Ventricular arrhythmia	*	5.85E-05
11	Sinus bradycardia	*	6.29E-05
12	Cardiac Arrest	*	6.43E-05
13	Neonatal hypoxia		6.52E-05
14	Upper motor neurone lesion		6.97E-05
15	Sudden cardiac death	*	7.03E-05
16	Neurogenic muscular atrophy		7.42E-05
17	Tetanic uterine contractions		7.53E-05
18	Sudden death	*	7.79E-05
19	Tachycardia, Ventricular	*	8.08E-05
20	Left coronary artery occlusion		8.33E-05
21	Progressive atrophy		8.35E-05
22	Ventricular fibrillation and flutter	*	8.59E-05
23	Ectopic atrial pacemaker	*	8.66E-05
24	Diffuse atrophy		8.77E-05
25	Premature ventricular contraction	*	8.95E-05

Concepts marked with an asterisk are associated with disturbances of cardiac conduction or ion channel activity. The *P*-values have been corrected for multiple testing according to Holm's method.

EVALUATION

Cardiac development and *TBX3*

Cardiac development and function are domains poorly covered by GO or other gene annotation databases. The literature-weighted global test can be of assistance in this type of domains as other classes of concepts can be evaluated. Here, we analyze a comparison between normal mouse atrial working myocardium and atria in which the transcription factor *TBX3* was ectopically expressed. *TBX3* expression is usually confined to the cardiac conduction system and represses properties of the working myocardium, such as fast conduction and contraction, and high level of metabolic activity [45–47]. Ectopic expression causes the spread of the conduction myocardium properties, such as pacemaker activity, slow conduction and contraction and reduced metabolic activity. In these hearts frequent spontaneous contractions and arrhythmias are observed.

We tested the semantic category 'pathologic function' for association with gene expression changes (Table 2). Fifteen of the twenty-five most significant pathological functions were found to be associated with disturbances of cardiac conduction or ion channel activity, in line with the observed phenotypes of the hearts of the mutant mice.

GO-terms can be represented by concept profiles. This enables a direct comparison of the literature-weighted global test and the standard global test based on the GO consortium gene sets, shown in Table 3 for the GO branch Biological Processes (see Supplementary Tables S1 and S2 for other branches). The literature-weighted global test retrieves more significant GO terms related to cardiac development and cardiac conduction than standard GO analysis (7 versus 2 concepts in the top 25 most significant concepts; Table 3). A review of the top 100 concepts shows that the literature-weighted global test retrieves a higher fraction of relevant terms than the standard GO analysis (0.35 versus 0.26, respectively; Supplementary Table S3). The top concept from the literature-weighted global test is 'action potential propagation' and reflects the observed action propagation defect and spontaneous contractions in the mutant mice. To gain insight into

Table 3: The top 25 most significant Biological Processes GO concepts for the standard and literature-weighted global test on the *TBX3* data set

Rank	Standard global test			Literature-weighted global test		
	GO-concept		*P*-value	GO-concept		*P*-value
1	Apoptotic program		2.69E-05	Action potential propagation	*	2.00E-05
2	Cellular component disassembly		5.84E-05	Seed development		4.41E-05
3	Monovalent inorganic cation transport		5.98E-05	Xylanase regulator		4.65E-05
4	Cellular macromolecule catabolic process		6.44E-05	Root morphogenesis		5.54E-05
5	Macromolecule catabolic process		7.08E-05	Aspartate metabolism		5.62E-05
6	Regulation of catabolic process		8.06E-05	Activ. of prog. cell death		6.19E-05
7	Nucleus organization		8.33E-05	Lipoprotein toxin		7.02E-05
8	Biopolymer catabolic process		9.80E-05	Reg. of programmed cell death		7.47E-05
9	Energy deriv. by organic comp. oxidation		9.81E-05	Suppression of hr		7.81E-05
10	DNA catabolic process		1.02E-04	Potassium conductance	*	7.92E-05
11	Catabolic process		1.13E-04	Neural crest cell development		8.01E-05
12	Lipid catabolic process		1.17E-04	Cation transport		8.11E-05
13	Regulation of lipid metabolic process		1.26E-04	Muscle hyperplasia		8.19E-05
14	Heart process	*	1.28E-04	L-glutamate transport		9.57E-05
15	Heart contraction	*	1.28E-04	Reg. of cardiac contraction	*	1.01E-04
16	DNA fragmentation, apoptosis		1.30E-04	Activation of atpase activity		1.04E-04
17	Cell struc. disassembly, apoptosis		1.30E-04	TCE metabolism		1.17E-04
18	Apoptotic nuclear changes		1.31E-04	Retrograde axonal transport		1.29E-04
19	Neg. reg. of multicell. organismal process		1.36E-04	Diaphragm contraction		1.35E-04
20	Purine nucleotide metabolic process		1.36E-04	Membrane hyperpolarization	*	1.56E-04
21	Regulation of TGF-β receptor pathway		1.74E-04	Adherens junction assembly	*	1.62E-04
22	Cation transport		1.96E-04	Generation of action potential	*	1.63E-04
23	Nucleoside triphosph. metabolic process		2.03E-04	Potassium ion conductance	*	1.74E-04
24	Regulation of cell communication		2.04E-04	GlUcose catabolism		1.82E-04
25	Purine nucleotide biosynthetic process		2.11E-04	Muscle hypertrophy		1.86E-04

The *P*-values have been corrected for multiple testing according to Holm's method. Concepts that are related to cardiac conduction and ion channel activity are indicated with an asterisk.

the workings of the test, the R-package includes several graphs. For example, Figure 2 shows the genes contributing most to the finding of this concept, among them the gap junction proteins *Gja*1 and *Scn5a*, which are under direct control of *TBX3* [45, 48–50], as well as several differentially expressed sodium and potassium channels. Using the standard GO annotations, action potential propagation is only connected to five genes on the microarray. The standard global test only retrieves the generic heart related terms 'heart process' and 'heart contraction', which were represented by an identical set of 52 genes.

The literature-weighted global test (Venn diagrams in Figure 3) finds more significant terms at the 5% confidence level than the standard test (3.6, 1.8, 2.9-times more for Biological Process, Molecular Function, Cellular Component). The extra concepts scored as significant by the literature-weighted global test are typically more specific than those found with the standard global test. This is reflected by the number of genes annotated in GO per Biological Process category: a median of 11 genes for those specific for the literature-weighted global test versus a median of 123 for the standard global test.

The top 25 list of the literature-weighted global test contains some apparently false positive hits. In Table 3, we see 'seed development', 'xylanase regulator' and 'root morphogenesis', which are plant-related concepts. Interestingly, these associations can be traced back to the functions in plants of homologues with the same name and molecular function as the genes differentially expressed in mice. For example, diacylglycerol acetyl transferase 2 (*Dgat2*) contributes to 'seed development', and is an enzyme with a conserved function in a wide range of organisms, expressed in the heart but also involved in the development of seeds [51]. On one hand, this example illustrates the power of the approach to infer relationships across species. On the other hand, it suggests that further improvements in the removal of concepts irrelevant to the domain under study should be considered.

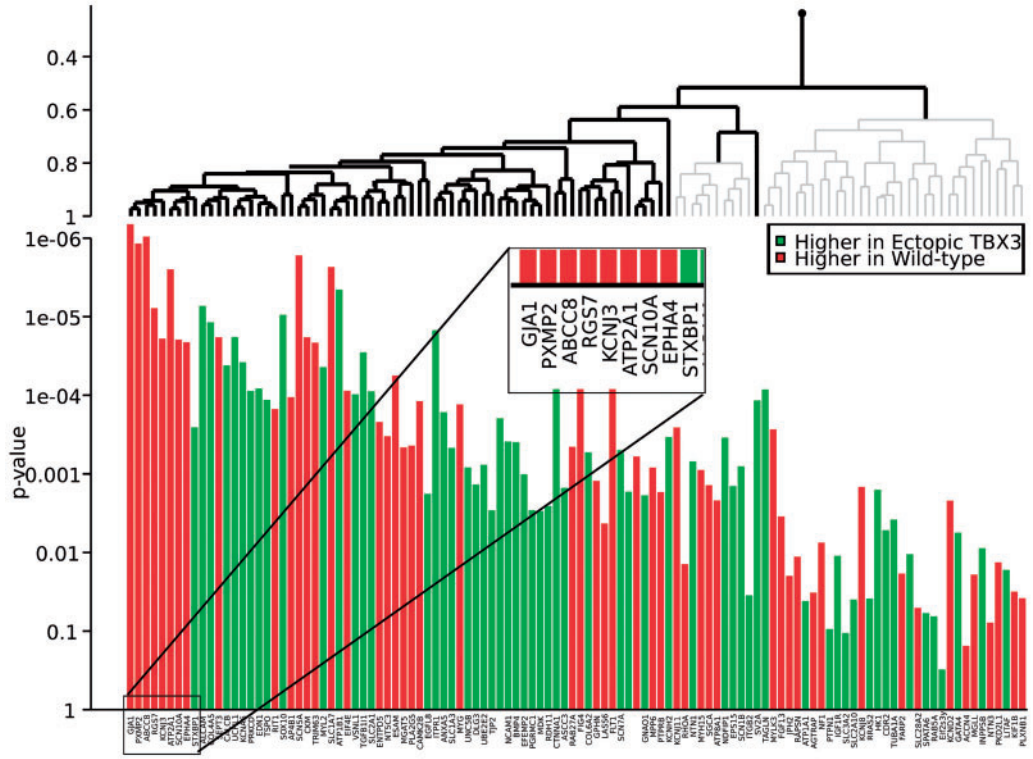

Figure 2: Features plot of differential gene expression between *TBX3* overexpressing and wild-type mice based on the importance weights for the concept 'Action Potential Propagation'. The bottom panel gives unadjusted *P*-values for differential expression of each selected gene, colored for the direction of association (green = upregulated in *TBX3*, red = downregulated in *TBX3*). The top panel gives a hierarchical clustering (average linkage) of the genes based on absolute correlation distance between their expression values. Imposed on this graph are the results of the inheritance multiple testing procedure of J.J. Goeman and L. Finos (submitted for publication), which is based on the same clustering graph and on the importance weights for the concept 'Action Potential Propagation'. Significant branches and leaves (alpha = 0.05) are shown in black, non-significant branches in gray. To facilitate presentation, all branches that are completely non-significant have been pruned from this picture, removing 1953 out of 2024 genes that have low importance weight and/or low differential expression.

Analysis of patient survival and gene expression in breast cancer

Van de Vijver *et al.* [52] investigated the association between breast cancer tissue expression profiles and patient survival and identified prognostic gene signatures. The global test is unique among the gene annotation-based methods, in that it is able to analyze survival time as a response. Biological processes previously associated with survival [53] mainly relate to cell division, and microtubule organization. The genes in these GO categories probably play a role in the development of metastases [54], an important predicting factor for survival. Accordingly, frequently prescribed drugs such as taxanes inhibit cell division and bind microtubules [55, 56]. We analyzed this 'classic' data set with the literature-weighted global test and compared the results to the global test with standard GO term assignments. Similar to the results for the *TBX3* data set, we reviewed the top 100 GO terms and found a higher fraction of relevant terms for the literature-weighted global test compared to the standard GO analysis (0.68 versus 0.48, respectively) (Supplementary Table S4). Table 4 displays the top 25 GO-terms associated with patient survival. Again, the literature-weighted global test produces more specific results than the standard global test, presenting concepts such as 'Telomerase inhibitor activity', 'polycomb group protein complex' and 'thymidine kinase activity' known to be highly relevant for cell division and cancer progression. Although these are genuine GO-categories, they contain less than 10 annotated genes and are therefore unlikely to be found with standard GO-based testing methods. The genes contributing most to the

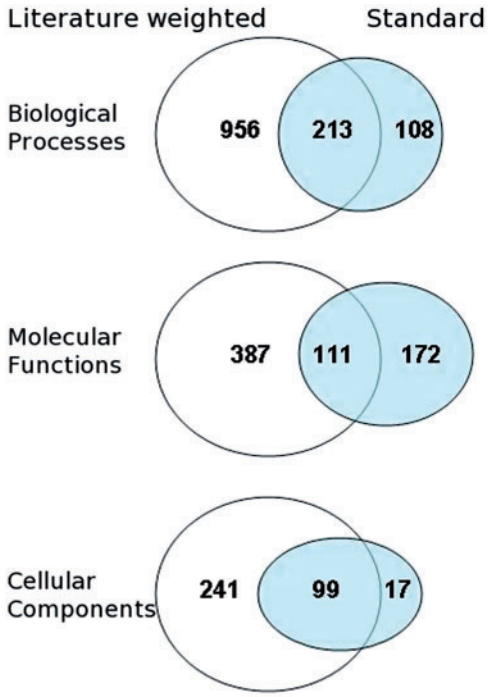

Figure 3: Venn diagrams showing the number of significant GO concepts for both the literature-weighted global test and the standard global test for the *TBX3* data set. The three GO branches are Biological processes, Molecular function and Cellular compartment.

finding of these concepts are *TSPYL5*, *EZH2* and *TK1*, respectively. *EZH2* and *TK1* were previously also associated with metastasis indicating that the molecular processes in which they participate are more generally associated with tumor metastasis and survival [57, 58]. *TSPYL5* is part of Van de Vijver's 70-gene signature to predict survival in breast cancer, but not much is known about the function of the gene. We associated *TSPYL5* with 'metastasis' in Anni and found two papers [59, 60], one of them describing its inhibitory activity on growth of tumor cells and its epigenetic silencing in gastric cancer and gliomas [59]. In addition, *TSPYL5* was associated with metastasis through the concepts 'histone deacetylation' and 'telomerase activity'. Evaluations of the Cellular Component and Molecular Function GO branches for this data set s are given in Supplementary Tables S5 and S6.

PPARalpha-mediated effects of dietary lipids on intestinal barrier gene expression

The literature-weighted global test can link gene expression data with (drug) metabolism as our thesaurus contains a large compendium of drugs and other small molecules. We demonstrate this possibility on a data set from Vogel-van den Bosch *et al.* [61] that evaluates the effect of a synthetic Peroxisome Proliferator Activated Receptor alpha (PPARalpha) stimulator, the fibrate WY14643, on the gene expression in the small intestine of both wild-type and PPARalpha-null mice. PPARalpha is a nuclear receptor highly expressed in enterocytes, and is thought to play a role in the reaction of the intestine to fatty acids. The original study reported the following manually annotated processes as responsive to PPARalpha stimulation: fatty acid oxidation, cholesterol flux, glucose transport, amino acid metabolism, intestinal motility and oxidative stress. We replicated this manual literature study in an automatic way by using the global test and weighted global test (Supplementary Table S7). With a standard global test on GO biological processes, the most significant processes found were related to fatty acid metabolism. However, GO categories corresponding to the other manual categories were not significant at the 5% level. With the weighted global test we identified significant GO categories corresponding to 5 of the manually grouped categories: 'fatty acid omega oxidation', 'regulation of cholesterol transport', 'glucose transport', 'amino acid metabolism' and 'glutathione metabolism pathway. The GO vocabulary does not contain a concept for intestinal motility. Similar to the previous examples, the literature-weighted global test produces more specific and less redundant results than the standard global test. For example, when the standard global test provided general concepts such as 'lipid metabolic process', 'fatty acid metabolic process, 'carboxylic acid metabolic process' and 'organic acid metabolic process; the literature-weighted global test provided specific concepts such as 'lauric acid metabolism', 'arachidonic acid metabolism' and 'leukotriene metabolism'.

Subsequently, we used the literature-weighted global test to evaluate the concept profiles of drugs. Known PPAR alpha agonists such as clofibrate, gemfibrozil, bezafibrate and fenofibrate were found to be significantly associated with the gene expression differences between PPARalpha-null and wild-type mice (Supplementary Table S8). The most significant drug was benazepril, which is used to treat high blood pressure and inhibits angiotensin-converting enzyme (ACE). Indeed, ACE is differentially expressed and amongst the top contributing genes. The gene with the highest contribution,

Table 4: The top 25 most significant Biological Processes GO concepts for the normal and literature-weighted global test for the Van de Vijver data set

Rank	Standard global test			Literature-weighted global test		
			P-value	GO-concept		*P*-value
1	DNA replication	*	1.28E-05	Chloroplast fission		1.40E-05
2	Protein complex localization		2.05E-05	Meiotic cell cycle regulator	*	1.47E-05
3	Chromosome segregation	*	3.75E-05	Cytokinetic process	*	1.64E-05
4	Protein–DNA complex assembly	*	3.94E-05	Bouquet formation	*	2.08E-05
5	Cytokinesis	*	4.55E-05	Meiotic recombination checkpoint	*	2.19E-05
6	Microtubule cytoskeleton organization	*	5.72E-05	Heterochromatic silencing telomere	*	2.61E-05
7	Establishment of organelle localization		6.85E-05	Stimulation of atpase activity		2.62E-05
8	DNA metabolic process		8.13E-05	Septin ring assembly		2.63E-05
9	Response to DNA damage stimulus	*	8.17E-05	Pyrimidine salvage	*	2.65E-05
10	Mitotic sister chromatid segregation	*	8.50E-05	Sister chromatid cohesion	*	2.97E-05
11	Sister chromatid segregation	*	8.50E-05	Hypusine biosynthesis		3.04E-05
12	Cell cycle	*	8.68E-05	Phosphatidylcholine Biosynthesis		3.05E-05
13	Cell division	*	9.00E-05	Regulation of dna replication	*	3.06E-05
14	Mitotic cell cycle	*	1.19E-04	Deoxycytidine metabolism	*	3.28E-05
15	Cellular macromolecular complex org.		1.47E-04	Telomere clustering	*	3.36E-05
16	M phase	*	1.52E-04	Cell tip growth		3.47E-05
17	Microtubule-based process	*	1.64E-04	Leaf morphogenesis		3.69E-05
18	M phase of mitotic cell cycle	*	1.77E-04	Telomerase inhibitor activity	*	3.71E-05
19	Nuclear division	*	1.77E-04	Heterocycle biosynthesis		3.73E-05
20	Mitosis	*	1.77E-04	Mesendoderm development		3.76E-05
21	Organelle fission		1.77E-04	Activation telomere maintenance	*	3.80E-05
22	Cell cycle phase	*	1.79E-04	TMP biosynthesis		4.02E-05
23	Cell cycle process	*	1.90E-04	Centriole replication	*	4.29E-05
24	Chromosome organization	*	2.07E-04	Double strand break repair	*	4.31E-05
25	Meiosis	*	2.09E-04	Diakinesis		4.42E-05

The *P*-values have been corrected for multiple testing according to Holm's method. Concepts indicated with an asterisk are related to disturbances in cell division and proliferation, or other cancer-related processes.

CYP4A11, showed a large change in gene expression and is directly involved in blood pressure regulation [62]. Also some of the most significant concepts found with the literature-weighted global test on GO biological processes were associated with high blood pressure. We hypothesize that PPAR alpha stimulation also regulates perfusion of the intestine.

CONCLUDING REMARKS

A standard functional analysis of gene expression data involves the testing of gene sets associated with biological processes. Though powerful, these methods mostly rely on manual annotation efforts, which are highly focused and struggle to be complete and up-to-date. Tools based on literature mining can be continuously updated and provide an essentially comprehensive scope. We combined literature mining tools with a thorough statistical framework and show that our approach compares favorably to that of a classic GO analysis, retrieving more, more relevant and more specific GO terms than an analysis based on standard GO annotations.

As expected given the ambiguous nature of gene names, a notorious problem in text mining [63], the automated literature mining comes at the cost of increased number of false positives. It is therefore an important feature of our approach that it is transparent and that results can readily be traced back to the literature that underlies a result. We nevertheless believe that the higher information content of the concepts retrieved by the literature-weighted global test, not only including GO terms but also other types of concepts such as diseases, organ structures and drugs, makes the test more suitable for the interpretation of gene expression data than other available gene set testing algorithms.

AVAILABILITY

The weighted global test is now an integral part of the R-package global test and can be obtained from www.bioconductor.org. The matrices containing the literature-derived association scores for genes with biological processes, molecular functions, cellular components, diseases, pathologic functions, tissues

and drugs and a file for the mapping between EntrezGene IDs and concept identifiers can be obtained from http://biosemantics.org/weightedglobaltest. Example R-code is available from the same website. Further investigation of the concept profiles and concepts underlying the identified associations can be performed with Anni (www.biosemantics.org/anni) [28].

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Literature mining tools complement manually curated gene annotation approaches by virtue of their broad scope and up to date information.
- The literature-weighted global test can evaluate biomedical concepts for association with gene expression changes based on text mining-derived associations. The test uses a regression analyses of the actual gene expression values and can be applied to DNA microarray as well as sequencing based expression profiling studies [64].
- A wide range of concepts can be evaluated, such as diseases, tissues and drugs, which are grouped by semantic category.
- The test is an extension of the versatile global test package. Several types of response variables can be used such as a two state, multi state and continuous variables, as well as survival. In addition, the package comes with numerous analytical graphs, which can provide insight into the results of the test.
- The presented framework is transparent so that results can readily be traced back to the underlying literature-based association score. Our online literature-mining tool can then be used to retrieve the underlying literature and further explore the association scores.

Acknowledgements

We would like to thank Drs W.M.C. Hoogaars, V.M. Christoffels and G. Hooiveld for useful discussions and critical reading of the article.

FUNDING

Centre for Medical Systems Biology within the framework of the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO); European Community's Seventh Framework Programme (FP7/2007-2013)-funded ENGAGE project; grant agreement HEALTH-F4-2007-201413; the European Union FP6 program HeartRepair LSHM-CT-2205-018630; Dutch Technology Foundation STW; applied science division of NWO; Technology Program of the Ministry of Economic Affairs.

References

1. Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.
2. Kandasamy K, Mohan SS, Raju R, *et al*. Netpath: a public resource of curated signal transduction pathways. *Genome Biol* 2010;**11**(1):R3.
3. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* 2000;**25**(1):25–29.
4. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**(43):15545–50.
5. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**(1):1–13.
6. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005;**21**(18):3587–95.
7. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005;**21**(9): 1943–9.
8. Pavlidis P, Qin J, Arango V, *et al*. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* 2004;**29**(6):1213–22.
9. Mootha VK, Lindgren CM, Eriksson KF, *et al*. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;**34**(3):267–73.
10. Goeman JJ, van de Geer SA, de Kort F, *et al*. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**(1):93–9.
11. Goeman J, van de Geer S, van Houwelingen J. Testing against a high-dimensional alternative. *J Roy Stat Soc B Stat Meth* 2006;**68**:477–93.
12. Hummel M, Meister R, Mansmann U. Globalancova: exploration and assessment of gene group effects. *Bioinformatics* 2008;**24**(1):78–85.
13. Goeman JJ, Bühlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007;**23**(8):980–7.
14. Barrell D, Dimmer E, Huntley RP, *et al*. The goa database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Res* 2009;**37**(Database issue):D396–D403.
15. Khatri P, Done B, Rao A, *et al*. A semantic analysis of the annotations of the human genome. *Bioinformatics* 2005; **21**(16):3416–21.
16. He X, Sarma MS, Ling X, *et al*. Identifying overrepresented concepts in gene lists from literature: a statistical approach based on poisson mixture model. *BMC Bioinformatics* 2010; **11**:272.
17. Shatkay H, Edwards S, Wilbur WJ, *et al*. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol* 2000;**8**:317–28.
18. Jenssen TK, Laegreid A, Komorowski J, *et al*. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001;**28**(1):21–8.

19. Blaschke C, Oliveros JC, Valencia A. Mining functional information associated with expression arrays. *Funct Integr Genomics* 2001;**1**(4):256–68.
20. Chaussabel D, Sher A. Mining microarray expression data by literature profiling. *Genome Biol* 2002;**3**(10): Research0055.1–0055.16.
21. Raychaudhuri S, Chang JT, Imam F, *et al*. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res* 2003;**31**(15): 4553–60.
22. Glenisson P, Coessens B, Vooren SV, *et al*. Txtgate: profiling gene groups with text-based information. *Genome Biol* 2004;**5**(6):R43.
23. Rubinstein R, Simon I. Milano–custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics* 2005;**6**:12.
24. Homayouni R, Heinrich K, Wei L, *et al*. Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics* 2005;**21**(1):104–15.
25. Alako BTF, Veldhoven A, van Baal S, *et al*. CoPub mapper: mining medline based on search term co-publication. *BMC Bioinformatics* 2005;**6**:51.
26. Jelier R, Jenster G, Dorssers LCJ, *et al*. Text-derived concept profiles support assessment of dna microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics* 2007;**8**(1):14.
27. Febbo PG, Mulligan MG, Slonina DA, *et al*. Literature lab: a method of automated literature interrogation to infer biology from microarray analysis. *BMC Genomics* 2007;**8**:461.
28. Jelier R, Schuemie MJ, Veldhoven A, *et al*. Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol* 2008;**9**(6):R96.
29. Huang ZX, Tian HY, Hu ZF, *et al*. Genclip: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. *BMC Bioinformatics* 2008;**9**:308.
30. Barbosa-Silva A, Soldatos TG, Magalhães ILF, *et al*. Laitor–literature assistant for identification of terms co-occurrences and relationships. *BMC Bioinformatics* 2010;**11**:70.
31. Tjioe E, Berry MW, Homayouni R. Discovering gene functional relationships using faun (feature annotation using nonnegative matrix factorization). *BMC Bioinformatics* 2010;**11**(Suppl. 6):S14.
32. Burkart MF, Wren JD, Herschkowitz JI, *et al*. Clustering microarray-derived gene lists through implicit literature relationships. *Bioinformatics* 2007;**23**(15):1995–2003.
33. Jelier R, 't Hoen PAC, Sterrenburg E, *et al*. Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. *BMC Bioinformatics* 2008; **9**:291.
34. Soldatos TG, O'Donoghue SI, Satagopam VP, *et al*. Martini: using literature keywords to compare gene sets. *Nucleic Acids Res* 2010;**38**(1):26–38.
35. Leach SM, Tipney H, Feng W, *et al*. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol* 2009;**5**(3):e1000215.
36. Sartor MA, Mahavisno V, Keshamouni VG, *et al*. Conceptgen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics* 2010;**26**(4):456–63.
37. Frijters R, Heupers B, van Beek P, *et al*. Copub: a literature-based keyword enrichment tool for microarray data analysis. *Nucleic Acids Res* 2008;**36**(Web Server issue): W406–W410.
38. Leong HS, Kipling D. Text-based over-representation analysis of microarray gene lists with annotation bias. *Nucleic Acids Res* 2009;**37**(11):e79.
39. Kueffner R, Fundel K, Zimmer R. Expert knowledge without the expert: integrated analysis of gene expression and literature to derive active functional contexts. *Bioinformatics* 2005;**21**(Suppl. 2):ii259–ii267.
40. Minguez P, Al-Shahrour F, Montaner D, *et al*. Functional profiling of microarray experiments using text-mining derived bioentities. *Bioinformatics* 2007;**23**(22): 3098–9.
41. Jelier R, Schuemie MJ, Roes PJ, *et al*. Literature-based concept profiles for gene annotation: the issue of weighting. *Int J Med Inform* 2008;**77**(5):354–62.
42. Goeman JJ, Oosting J, Cleton-Jansen AM, *et al*. Testing association of a pathway with survival using gene expression data. *Bioinformatics* 2005;**21**(9):1950–7.
43. Schuemie MJ, Chichester C, Lisacek F, *et al*. Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of medline. *Proteomics* 2007;**7**(6): 921–31.
44. Srinivasan P. Text mining: generating hypotheses from medline. *JASIST* 2004;**55**:396–413.
45. Horsthuis T, Buermans HPJ, Brons JF, *et al*. Gene expression profiling of the forming atrioventricular node using a novel tbx3-based node-specific transgenic reporter. *Circ Res* 2009;**105**(1):61–9.
46. Hoogaars WMH, Engel A, Brons JF, *et al*. Tbx3 controls the sinoatrial node gene program and imposes pacemaker function on the atria. *Genes Dev* 2007;**21**(9): 1098–112.
47. Bakker ML, Boukens BJ, Mommersteeg MTM, *et al*. Transcription factor tbx3 is required for the specification of the atrioventricular conduction system. *Circ Res* 2008; **102**(11):1340–9.
48. Simon AM, Goodenough DA, Paul DL. Mice lacking connexin40 have cardiac conduction abnormalities characteristic of atrioventricular block and bundle branch block. *Curr Biol* 1998;**8**(5):295–8.
49. Lupoglazoff JM, Cheav T, Baroudi G, *et al*. Homozygous scn5a mutation in long-qt syndrome with functional two-to-one atrioventricular block. *Circ Res* 2001;**89**(2): E16–E21.
50. Boukens BJD, Christoffels VM, Coronel R, *et al*. Developmental basis for electrophysiological heterogeneity in the ventricular and outflow tract myocardium as a substrate for life-threatening ventricular arrhythmias. *Circ Res* 2009;**104**(1):19–31.
51. Kroon JTM, Wei W, Simon WJ, *et al*. Identification and functional expression of a type 2 acyl-coa:diacylglycerol acyltransferase (dgat2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals. *Phytochemistry* 2006; **67**(23):2541–9.
52. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**(25):1999–2009.
53. Goeman JJ, Mansmann U. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics* 2008;**24**(4): 537–44.

54. van 't Veer LJ, Dai H, van de Vijver MJ, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;**415**(6871):530–6.

55. Dong X, Liu F, Sun L, *et al*. Oncogenic function of microtubule end-binding protein 1 in breast cancer. *J Pathol* 2010; **220**(3):361–9.

56. Morris PG, Fornier MN. Ixabepilone and other epothilones: microtubule-targeting agents for metastatic breast cancer. *Clin Adv Hematol Oncol* 2009;**7**(2):115–22.

57. Chen CC, Chang TW, Chen FM, *et al*. Combination of multiple mrna markers (pttg1, survivin, ubch10 and tk1) in the diagnosis of taiwanese patients with breast cancer by membrane array. *Oncology* 2006;**70**(6):438–46.

58. Varambally S, Yu J, Laxman B, *et al*. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* 2005;**8**(5):393–406.

59. Jung Y, Park J, Bang YJ, *et al*. Gene silencing of tspyl5 mediated by aberrant promoter methylation in gastric cancers. *Lab Invest* 2008;**88**(2):153–60.

60. Vachani A, Nebozhyn M, Singhal S, *et al*. A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma. *Clin Cancer Res* 2007;**13**(10):2905–15.

61. de Vogel-van den Bosch HM, Bünger M, de Groot PJ, *et al*. Pparalpha-mediated effects of dietary lipids on intestinal barrier gene expression. *BMC Genomics* 2008;**9**:231.

62. Capdevila JH, Falck JR, Imig JD. Roles of the cytochrome p450 arachidonic acid monooxygenases in the control of systemic blood pressure and experimental hypertension. *Kidney Int* 2007;**72**(6):683–9.

63. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics* 2005;**21**(2):248–56.

64. 't Hoen PAC, Ariyurek Y, Thygesen HH, *et al*. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 2008;**36**(21):e141.