JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2012) **9**, 1063–1072
doi:10.1098/rsif.2011.0551
Published online 13 October 2011

A novel method to quantify gene set functional association based on gene ontology

Sali Lv,[1,†], Yan Li[1,†], Qianghu Wang[1,†], Shangwei Ning[1], Teng Huang[1], Peng Wang[1], Jie Sun[1], Yan Zheng[1], Weisha Liu[1], Jing Ai[2] and Xia Li[1,*]

[1]*College of Bioinformatics Science and Technology and Bio-pharmaceutical Key Laboratory of Heilongjiang Province, and* [2]*Department of Pharmacology* (*State-Province Key Laboratories of Biomedicine-Pharmaceutics of China*), *Harbin Medical University, Harbin 150081, People's Republic of China*

Numerous gene sets have been used as molecular signatures for exploring the genetic basis of complex disorders. These gene sets are distinct but related to each other in many cases; therefore, efforts have been made to compare gene sets for studies such as those evaluating the reproducibility of different experiments. Comparison in terms of biological function has been demonstrated to be helpful to biologists. We improved the measurement of semantic similarity to quantify the functional association between gene sets in the context of gene ontology and developed a web toolkit named Gene Set Functional Similarity (GSFS; http://bioinfo.hrbmu.edu.cn/GSFS). Validation based on protein complexes for which the functional associations are known demonstrated that the GSFS scores tend to be correlated with sequence similarity scores and that complexes with high GSFS scores tend to be involved in the same functional catalogue. Compared with the pairwise method and the annotation method, the GSFS shows better discrimination and more accurately reflects the known functional catalogues shared between complexes. Case studies comparing differentially expressed genes of prostate tumour samples from different microarray platforms and identifying coronary heart disease susceptibility pathways revealed that the method could contribute to future studies exploring the molecular basis of complex disorders.

Keywords: gene annotation; gene set functional association; improved semantic similarity

1. INTRODUCTION

In the post-genomic era, many experiments have been designed to explore the cellular basis of complex human disorders [1–5]. The most common experimental strategy is to compare the molecular signatures of cells in normal and anomalous samples and to construct a functional set of genes with differential activities. These functional gene sets are distinct but—in many cases—related, and life scientists are often interested in comparing or finding associations between two of these gene sets. Here are some typical examples. (i) Scientists compare results from different microarray platforms using independent RNA samples to evaluate the reproducibility, specificity, sensitivity and accuracy of the platforms [6,7]. (ii) Scientists compare genes associated with one disease to the genes associated with another disease to evaluate the comorbidity of the diseases [8,9]. (iii) Scientists compare gene sets associated with specific subtypes to find the molecular pattern of each disease subtype for diagnosis [10]. (iv) Scientists compare genes associated with a disease, and genes involved in a biological pathway to identify the pathways disrupted by the disease [11].

Several methods have been developed for gene set comparison. Most methods use a common core strategy to statistically analyse the gene annotation overlap between gene sets, such as GOStats and the Database for Annotation, Visualization and Integrated Discovery (DAVID), which have allowed researchers to evaluate the associations between gene sets [12–14]. These methods offer several advantages but also pose a number of challenges. Even when the two gene sets contain no common genes, the sets may be related to each other due to common biological pathways. For instance, there are four major groups of mitogen-activated protein kinases in mammalian cells. These proteins are activated by specific stimuli on the cell surface, and all terminate when the cell proliferates, differentiates or migrates. However, the genes involved in the four groups are different [15–17].

*Author for correspondence (lixia@hrbmu.edu.cn).
†The first three authors contributed equally to the study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2011.0551 or via http://rsif.royalsocietypublishing.org.

Received 15 August 2011
Accepted 22 September 2011

Additionally, several methods, such as FatiGO and Martini, allow for comparison of gene sets based on biological functions and separation of the significantly enriched biological categories into one set with respect to the other set [18,19]. However, these methods are still more of an exploratory data-mining procedure rather than a quantifying solution and fail to provide a quantitative value and statistical analysis that enhances reliability. Analysis of the shared biological processes related to two gene sets provides only a hint that the gene sets are related. In fact, randomly generated large gene sets may overlap falsely associated biological categories, and the form of statistical analysis is critical [20]. Therefore, a method must be constructed that can be widely used to perform a systematic comparison and statistical analysis to identify associations between gene sets.

This paper proposes a novel method for comparing two gene sets, named Gene Set Functional Similarity (GSFS), in the context of gene annotations (figure 1). The method quantifies the functional association of two gene sets using an improved semantic similarity measure and evaluates the significance of the score. To validate the performance of the method, 51 protein complexes whose functional associations are known in the functional catalogue (http://mips.helmholtz-muenchen.de/genre/proj/corum/) were extracted and used as the benchmark dataset. The results show that the functional similarity scores determined by this method were highly consistent with the sequence similarity scores. Using clustering analysis based on GSFS scores, the complexes were divided into five major clusters corresponding to the known functional catalogue. Comparison with the pairwise method and the annotation method verified that our method could better reflect the functional association between gene sets. Then, two case studies, one comparing the differentially expressed genes (DEGs) of prostate tumour samples from different microarray platforms and the other identifying coronary heart disease (CHD) susceptibility pathways, were performed. We believe this method, and its associated web toolkit will enable future studies to analyse the molecular signatures of complex human diseases and will be helpful in exploring the molecular basis of complex human disorders.

2. MATERIAL AND METHODS

2.1. Data sources

In the present study, gene annotation data were based on gene ontology (GO), which assigns biological categories to genes based on the properties of their encoded proteins [21]. The GO database is widely adopted by the life sciences community to study gene products at the functional level, which is crucial for a variety of applications [22,23]. Gene annotation data were downloaded from the official GO Consortium website (http://archive.geneontology.org/full/2010-01-01/). In the original downloaded dataset, genes are assigned to the most specific GO categories. To implement gene set enrichment analysis and semantic similarity measurements, we compiled the original dataset and inferred gene associations from the lower-level to higher-level GO categories according to the ontology structure. In our study, the functional similarity score between two gene sets was calculated based on the biological process, which is a series of events accomplished by one or more ordered molecular functions. Regardless of the associated evidence codes, we used all GO–gene associations to build comprehensive functional profiles of input gene sets.

Protein complexes were used as gene sets with known functions in this study. These sets were compiled from the latest Comprehensive Resource of Mammalian protein complexes (CORUM) database (release 02.09.2009), which contains 1343 human protein complexes (each protein complex can be regarded as a gene set) [24]. Information on these complexes was obtained from individual experiments published in the primary literature and was used to assign a functional catalogue. We disregarded protein complexes containing fewer than 15 genes because a small number of genes may yield a null-enriched category set, which would result in 51 human protein complexes (see electronic supplementary material, table S1). We used these 51 protein complexes, whose functional associations are known from the functional catalogue, as the benchmark dataset for validation and comparison of our method.

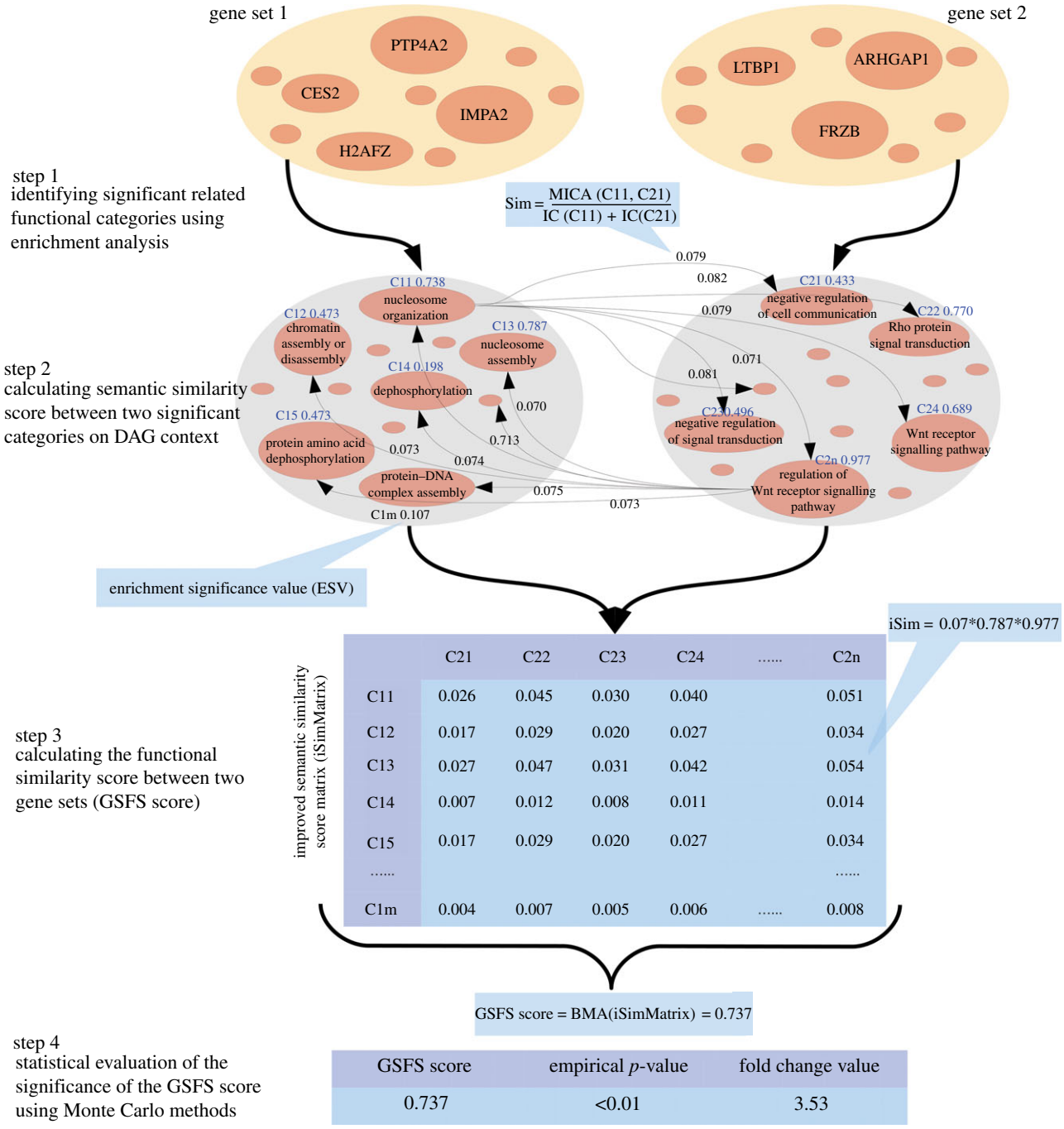
2.2. Calculating functional similarity score between two gene sets

First, we identified significant category sets using enrichment analysis for each gene set. When a gene set fell into a category, enrichment analysis was adopted to measure the enrichment significance values (ESVs) of the category. The lower the ESV, the more relevance a category has to the gene set [25]. Here, the cumulative hypergeometric test was used as follows:

$$P(X \geq q) = 1 - \sum_{x=0}^{q-1} \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}},$$

where N is the number in the GO background, M is the number of given genes, n is the number of genes annotated in a certain category and q is the number of given genes that are annotated in this category.

Second, the semantic similarity between two significant categories was then calculated [26,27]. The information content (IC) of a category was computed as $-\log\, p(c)$. The $p(c)$ calculation was based on determining the number of times that a specific GO category or any directly or indirectly related offspring appeared in annotated genes. This value is the number of genes annotated in that category divided by the number of all genes annotated to the GO domain. One of the most well-known semantic similarity measures was introduced by Resnik [28]. This measure relies on the minimum subsumer of the two categories, which is their common ancestor with the most informative content in the GO-directed acyclic graph most informative common ancestor (MICA), $\max_{c \in a(c_i, c_j)} (-\log\, p(c))$. Taking into account the differences between categories,



	C21	C22	C23	C24	.....	C2n
C11	0.026	0.045	0.030	0.040		0.051
C12	0.017	0.029	0.020	0.027		0.034
C13	0.027	0.047	0.031	0.042		0.054
C14	0.007	0.012	0.008	0.011		0.014
C15	0.017	0.029	0.020	0.027		0.034
.....						.....
C1m	0.004	0.007	0.005	0.006	.....	0.008

GSFS score	empirical *p*-value	fold change value
0.737	<0.01	3.53

Figure 1. The GSFS algorithm. The method for identifying the association of two gene sets is based on gene annotations. Step 1: enrichment analysis is applied to identify significantly enriched functional category sets for each gene set. Steps 2 and 3: the improved semantic similarity measure is used to calculate the functional similarity score. Step 4: a randomization test is used to give a *p*-value and a fold-change value to determine the statistical significance of the GSFS score.

a normalized measure was developed by Lin [29]. It is given by the formula

$$\mathrm{Sim}(c_i, c_j) = \frac{\mathrm{MICA}(c)}{\mathrm{IC}(c_i) + \mathrm{IC}(c_j)} = \frac{2 \max\limits_{c \in a(c_i, c_j)} (-\log\, p(c))}{-\log\, p(c_i) - \log\, p(c_j)}.$$

Considering that these categories have different levels of relevance to a gene set, the functional similarity score between two gene sets is influenced by ESVs. Thus, we defined a function $w(p_i, p_j)$ as the weight of the semantic similarity of categories c_i and c_j. As the relevance of the category increases with the decrease in ESV, the semantic similarity between two categories is weighted by $(1 - (p_i/\alpha))(1 - (p_j/\alpha))$, where p_i, and p_j represent the ESVs of categories and α represents the threshold of enrichment analysis. The application of the formula provides a more appropriate result for two categories from two gene sets. Thus, the improved semantic similarity between two categories is defined as

$$\begin{aligned}\mathrm{iSim}(c_i, c_j) &= \mathrm{Sim}(c_i, c_j) \cdot w(p_i, p_j) \\ &= \frac{2 \max\limits_{c \in a(c_i, c_j)} (-\log p(c))}{-\log p(c_i) - \log p(c_j)} \cdot \left(1 - \frac{p_i}{\alpha}\right) \\ &\quad \times \left(1 - \frac{p_j}{\alpha}\right).\end{aligned}$$

Third, based on the improved semantic similarity measure for the two categories, we implemented measures to investigate the functional similarity of the two category sets identified from the gene sets. For the two category sets A and B with sizes n_A and n_B, a similarity matrix $\mathrm{iSim}_{ij}(i=1,\ldots,n_A,\ j=1,\ldots,n_B)$ is calculated. The matrix contains the similarity scores of each category pair $(c_i,\ c_j)$, where category c_i is in set A, and c_j is in set B. The final similarity score between two category sets can be calculated by the best-matched average measure (BMA) using the scores in matrix iSim_{ij} [30]. It is given by the formula

$$\mathrm{GSFS\ score} = \frac{1/n_A)\sum_{i=1}^{n_A}\max_{1\le i\le n_B}\mathrm{iSim}_{ij} \quad + (1/n_B)\sum_{j=1}^{n_B}\max_{1\le j\le n_A}\mathrm{iSim}_{ij}}{2}.$$

Indeed, there are several alternative measures such as Jiang & Conrath's measure and graph information content measure [30,31]. These measures were implemented and evaluated in Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) [32,33]. In order to reason our choice of Lin measure with the BMA method, we have compared our choice with other frequently used measurements and validated our choice with CESSM. The detailed results are presented in the electronic supplementary material, table S2.

Thus, one can implement an improved semantic similarity measure as a scoring method for the functional similarity between two gene sets. The GSFS method was used to investigate the association of two gene sets at the functional level. Scores close to 1 indicate high functional similarity, whereas scores close to 0 indicate low similarity.

2.3. The evaluation of the significance of the similarity score

The similarity score is affected by the size of the gene sets. It thus requires further statistical analysis. Therefore, we constructed a statistical model to determine the significance of the similarity score of two gene sets. An empirical *p*-value (EP) is estimated for each similarity score by Monte Carlo random sampling that is obtained by randomly assigning the genes to each gene set of the same size. For each pair of genes in a random gene set, the functional similarity scores are calculated; the EP is the probability of getting the same or higher similarity score. Assuming that the GSFS score is the similarity score between two gene sets given by the method and n similarity scores, random GSFS $\mathrm{score}_i(i=1,\ldots,n)$ values are calculated from random gene sets. The estimate of the EP for the test is computed as

$$\mathrm{EP} = \sum_{i=1}^{n}\frac{\sigma_i}{n},$$

where

$$\sigma_i = \begin{cases} 0, & \text{if random GSFS score}_i \le \text{GSFS score} \\ 1, & \text{if random GSFS score}_i > \text{GSFS score} \end{cases}$$

is an indicator function.

The fold-change value is computed as

$$f = \frac{\mathrm{GSFS\ score}}{\mathrm{avg(random\ GSFS\ score)}},$$

where

$$\mathrm{avg(random\ GSFS\ score)} = \frac{1}{n}\sum_{i=1}^{n}\text{random GSFS score}_i$$

is the average score for random gene sets.

3. RESULTS

3.1. Validation of gene set functional similarity based on the sequence similarity of protein complexes

It is a biological tenet that genes with similar sequences exhibit similar functions; therefore, gene sets with highly similar sequences should have high functional similarity [34]. To test the efficacy of the GSFS in quantifying the association between gene sets, we decided to determine how the GSFS performs with respect to sequence similarity. We assumed that if the GSFS were successful in quantifying association between gene sets, then the GSFS scores would be correlated with the sequence similarity scores. To test this hypothesis, protein complex data were analysed by calculating the GSFS scores of each pair and correlating these scores with the sequence similarity scores. Here, 1275 protein complex pairs generated by the 51 complexes were compared.

We used the NCBI BLAST suite program 'BLASTN' (Basic Local Alignment Search Tool for searching Nucleotide databases) to analyse the similarity between the nucleotide sequences. This program uses the 'bit score' as a measure of sequence similarity between two proteins. Then an integrated method is given for the average bit score between each nucleotide in a complex and its most similar nucleotide in another complex, averaged with its reciprocal to obtain a symmetric score. Therefore, the sequence similarity score between two gene sets can be calculated.

The functional similarity score was generated by the GSFS (see §2). The GSFS and sequence similarity scores for 1275 complex pairs were determined and assessed for correlation. The distribution of sequence similarity scores in different groups of complex pairs was analysed with respect to different functional similarity scores. Figure 2 shows the distribution of sequence similarity scores for different functional similarity score bins of complex pairs. As expected, the complex pairs with high sequence similarity scores tended to have high GSFS scores. A greater sequence similarity for two complexes resulted in a more similar function as determined by the GSFS computation. We calculated Pearson's correlation between sequence similarity scores and functional similarity scores for 1275 complex pairs. The obtained correlation coefficient is 0.4519. To test the null hypothesis that the coefficient is equal to 0 versus the alternative hypothesis that the coefficient is not equal to 0, we get a *p*-value 2.2×10^{-16} using one sample *t*-test. Given this result, we would be inclined to reject the null hypothesis

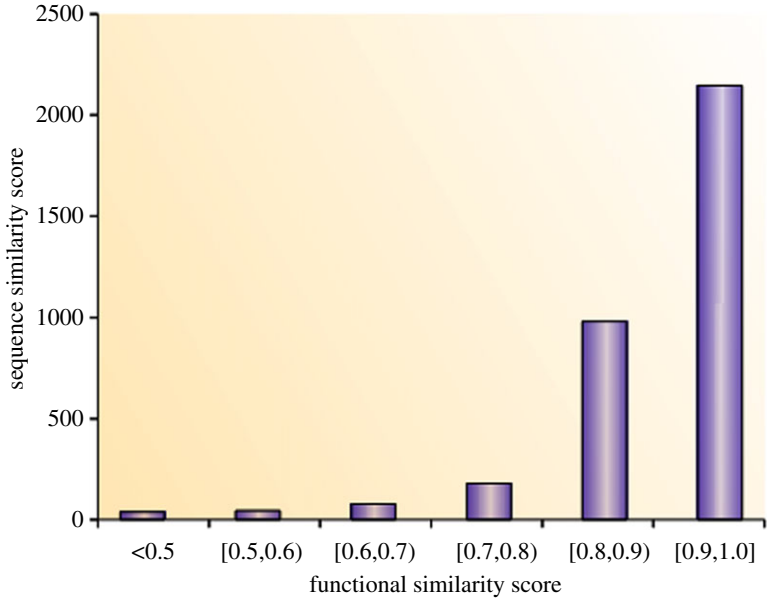


Figure 2. Distribution of the sequence similarity scores for different functional similarity score bins of protein complex pairs. A BLAST sequence analysis was performed to calculate a sequence similarity score for each gene pair in the respective complexes for which sequence data were available. The sequence scores for complex pairs were the average of the max bit score in each gene pair. The functional similarity scores were calculated using the GSFS method. (Online version in colour.)

and conclude that there exists correlation between functional similarity and sequence similarity.

3.2. Validation of gene set functional similarity on complex functional catalogue

We used the known functional annotations for protein complexes in CORUM to evaluate the efficacy of the GSFS method for quantifying gene set associations. We assumed that a relatively high score from GSFS for two protein complexes indicated that these complexes should have the same or a similar functional catalogue. Therefore, we analysed the functional similarity score matrix between all protein complex pairs. We applied a complete-linkage hierarchic cluster analysis on this score matrix using Cluster + TreeView to group functionally similar complexes [35]. These complexes were separated into seven clusters by setting the height cutoff value to 3 and then into two clusters that included one complex and five clusters that included all multiple complexes (figure 3). Remarkably, the clustered complexes were involved in the same or similar functional catalogues as the CORUM annotation. For example, the PA700 complex, PA700–20S–PA28 complex, PA28–20S proteasome, PA28 gamma–20S proteasome and 26S proteasome were gathered into a cluster that was annotated under the functional catalogue 'proteasomal degradation'. The clustering results reveal that the GSFS scores can favourably distinguish functional associations between complexes.

3.3. Comparison with other methods using complex functional catalogue

To illustrate the advantages of determining the association of two gene sets using the GSFS method, we carried out a comparison of the GSFS method with the gene pairwise method and the gene set annotation method. The gene pairwise method defines a summary for the gene set score that is the best-match average of the separate gene pair similarity scores. Gene pair similarity scores were implemented by GOSim [36]. The gene set annotation method is based on the best-match average of the similarity scores between two annotated category sets for each score, which was implemented by GOSemSim in this study [37]. We happened to choose GOSim and GOSemSim, but the semantic similarity analysis of categories and genes by other tools would be similar.

First, according to the annotation information for the complexes in the CORUM Functional Catalogue, 1275 pairs of complexes were divided into four groups. The first group consisted of 176 pairs in which the complexes were not annotated with any common functional catalogue. There is a low functional association between two complexes in any pair of the first group. The second group consisted of 241 pairs in which the complexes were annotated with a common functional catalogue in the first level. The third group consisted of 521 pairs in which the complexes were annotated with a common functional catalogue in the first and the second levels. The fourth group consisted of 337 pairs in which the complexes were annotated with a common functional catalogue in at least three levels. There is a high functional association between two complexes in any pair of the fourth group, according to the CORUM database.

We computed the functional similarity scores for each complex pair using three methods and classified the scores into four groups, as described earlier. By analysing the scores of the four groups, the functional association between complexes can be explored. It is reasonable to expect that the complexes in closer functional catalogue would likely have larger functional

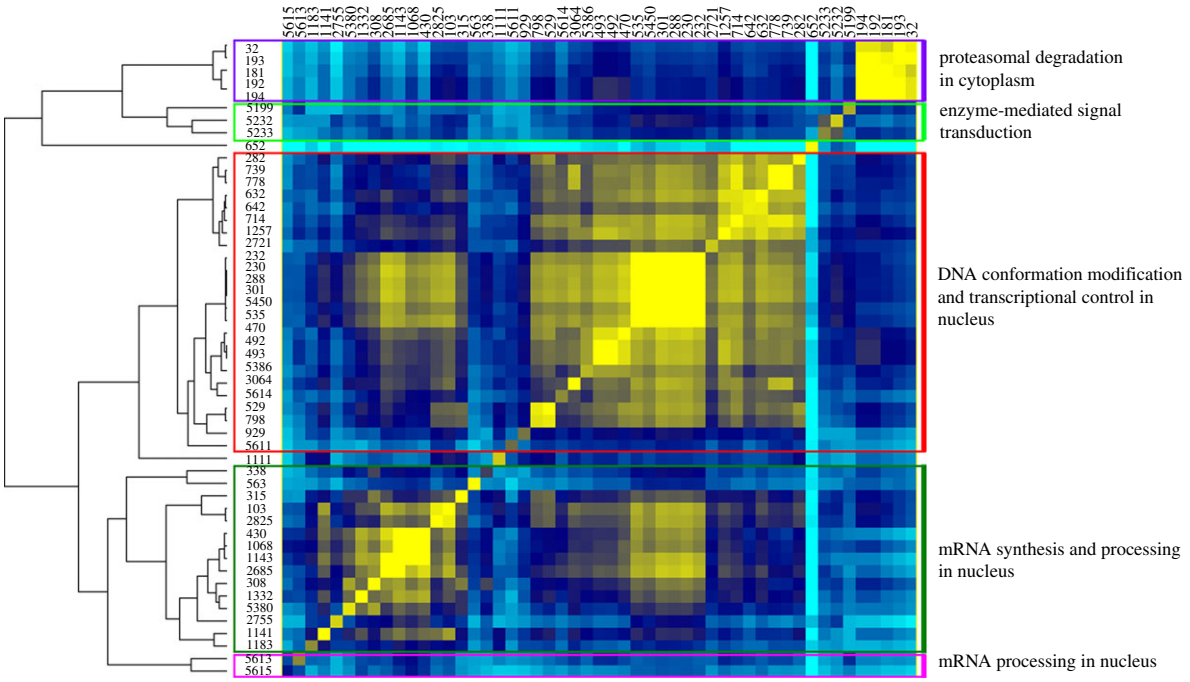


Figure 3. Clustering visualization of the GSFS scores between complexes. The clustering result is displayed as a heat map, and the value of GSFS is indicated by colour intensity, with yellow representing high functional similarity and blue representing low functional similarity.

similarity scores. The scores from the four groups are compared in figure 4. As shown in figure 4*b*,*c*, the results of the gene pair method and the gene annotation method do not agree with the expected results for groups 1, 2 and 3, and these methods yielded relatively high median scores for group 1. In figure 4*a*, the GSFS score increases from the first group to the fourth group, showing that the GSFS method has a high consistency with the functional catalogue information for protein complexes. Further, we calculated the average variance rate (AVR) for evaluating the discrimination of the three methods. Based on the median, we normalized the variance of the similarity scores from each method. The AVR was defined as follows:

$$\mathrm{AVR} = \sum_{i=1}^{3} \frac{M(g_{i+1}) - M(g_i)}{3M(g_i)},$$

where $M(g_i)$ is median of the ith group. The AVR of our method was 16.44 per cent, and the AVRs of the other methods were 8.79 and 2.22 per cent (figure 5). This finding suggests that the GSFS method would be the most discriminatory. The comparison between the GSFS and the other methods clearly shows that our method could better reflect the functional association between complexes.

3.4. Analysis results from two prostate tumour microarray datasets

Microarray experiments often produce a single gene set associated with a disease phenotype, but scientists have found that gene sets with the same or similar phenotype from different platforms have few common genes. Some authors have speculated that these different genes may share the same or similar functions that cause the same phenotype. Here, DEG sets with few overlapping genes from two microarray platforms that had functional consistency were compared. The cDNA and the oligo microarray datasets for prostate tumour and adjacent prostate tissue samples were used for analysis [38,39]. A comparison of the tumour characteristics of the two datasets is provided in table 1. The two prostate cancer datasets are very similar with respect to patient clinical characteristics, and the two DEG sets yielded a high functional similarity score.

DEGs by significance analysis of microarrays were detected with 1 per cent false discovery rate (FDR) control [5]. The cDNA microarray DEG set consists of 3342 genes, while the oligo microarray DEG set contains 2724 genes. The overlap between the two DEG sets detected from different microarray studies is 591 genes, and the average number of overlapping genes in a randomization test with 1000 repeats is 538, which indicates that it is possible to generate a similar degree of overlap, even in a randomly generated gene set. Next, the GSFS was used to compare the functional similarity of the two DEG sets. The average value of GSFS scores in a randomization test with 1000 repeats was 0.200, and the observed score was 0.505; the difference between these two scores is statistically significant. Our method yields a similarity score corresponding to the sample similarity with respect to clinical characteristics. The comparison results are shown in figure 5. The case study analysis shows that GSFS is a useful and reasonable method for comparing experimental datasets.

3.5. Identification coronary heart disease susceptibility pathways using gene set functional similarity

Pathways frequently exhibit specific biological functions. Therefore, each pathway is defined as a gene

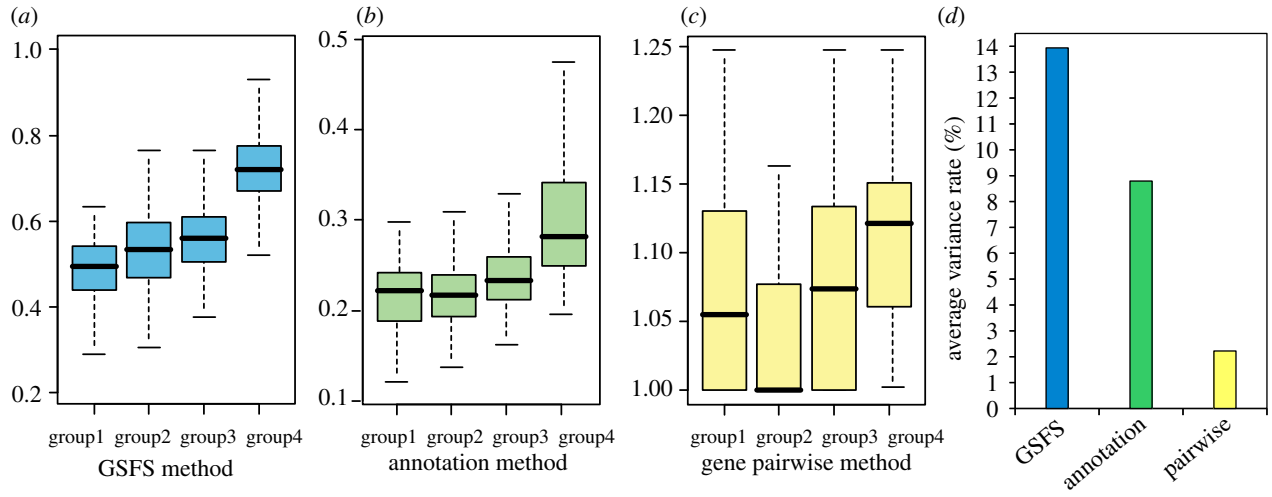


Figure 4. (*a*–*c*) Box-plots of functional similarity scores of protein complexes. The *x*-axes represent four groups of complex pairs, while the *y*-axes represent the functional similarity scores for each method. Thick horizontal lines indicate medians, boxes indicate interquartile ranges and whiskers are drawn at 1.5 times the quartile, or the maximum. (*d*) Bar graph of subgroup comparisons among the three methods derived from the AVR. (Online version in colour.)

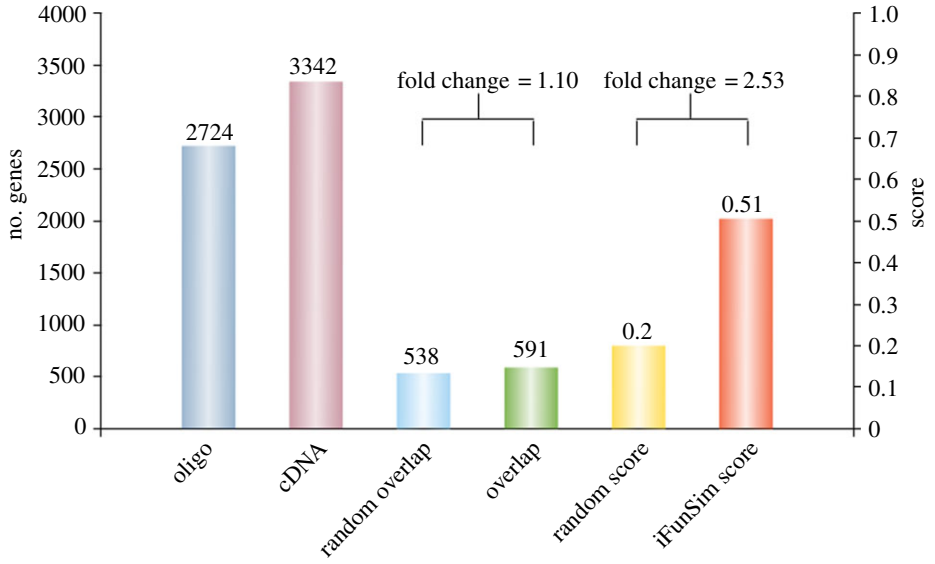


Figure 5. The *y*-axis (left) represents the number of genes, and the *x*-axis corresponds to oligos, cDNAs and overlap. The *y*-axis (right) represents the functional similarity scores, and the *x*-axis corresponds to the random score and the GSFS score. (Online version in colour.)

set, and a test based on functional similarity for a query gene set would retrieve related pathways. Thus, all pathways were compared with a gene set for CHD. First, according to the CHD classification of the World Health Organization (WHO), we manually extracted 39 CHD genes (see electronic supplementary material, table S3) from the Online Mendelian Inheritance in Man (OMIM) database involved in myocardial infarction, angina, arrhythmia, sudden cardiac arrest and ischaemic heart disease [40]. Then we downloaded 165 biological pathway datasets from the Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes and Organismal Systems catalogues in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [41].

The functional similarity scores between the CHD candidate gene set and each pathway were calculated by the GSFS method. The 165 biological pathways were ranked based on their similarity scores. The top 10 biological pathways are shown in table 2. Of the 10 pathways seven (which are marked with superscript 'a' in the table), reportedly play important roles in CHD. Additionally, the remaining three pathways (which are marked with superscript 'b'), have recently been associated with CHD [43,45–47,49–51]. For example, the sigma-1 receptor causes the pathway for neuroactive ligand–receptor interaction to malfunction, leading to memory process and cognitive disorders [42]. Sympathetic changes lead to disorders in the neurotrophin signalling pathway [44]. Similarly and equally importantly, the Fc epsilon RI signalling pathway can be detected in injured cardiac mast cells [48]. All of these disorders can cause cardiovascular diseases.

As a comparison, we also carried out a biological pathway enrichment analysis on these 39 CHD genes

Table 1. Comparison of tumour characteristics of two prostate tumour datasets.

characteristics	cDNA microarray	oligo microarray	*p*-value
tumour sample	62	52	
normal sample	41	50	
age			
mean	59.2	58.5	
range	45–72	47–72	
Gleason grade			
2–6	24 (39%)	24 (46%)	>0.05
7	22 (35%)	22 (42%)	
8–10	15 (24%)	6 (12%)	
stage			
T2a	6 (10%)	7 (13%)	>0.05
T2b	23 (37%)	25 (48%)	
T3a	19 (31%)	16 (31%)	
T3b	9 (15%)	4 (8%)	
other	5 (8%)	0 (0%)	

Table 2. Top 10 pathways identified by the gene set functional similarity method for coronary heart disease.

pathway name	GSFS score	DAVID (*p*-value)	reference
neuroactive ligand–receptor interaction[b]	0.6217	>0.1	Waterhouse *et al.* [42]
calcium signalling pathway[a]	0.61739	>0.1	Fareh *et al.* [43]
focal adhesion[a]	0.61697	>0.1	Luther & Birren [44]
neurotrophin signalling pathway[b]	0.61333	>0.1	Tedgui & Mallat [45]
cytokine–cytokine receptor interaction[a]	0.61221	0.062	Aukrust *et al.* [46]
chemokine signalling pathway[a]	0.6096	>0.1	Amasyali *et al.* [47]
adipocytokine signalling pathway[a]	0.60692	>0.1	Amasyali *et al.* [47]
Fc epsilon RI signalling pathway[b]	0.60537	>0.1	Marone *et al.* [48]
toll-like receptor signalling pathway[a]	0.59985	>0.1	Satoh *et al.* [49]
Jak-Stat signalling pathway[a]	0.59953	>0.1	Barry *et al.* [50]

[a]Seven of the 10 pathways play important roles in CHD.
[b]The remaining three pathways have recently been associated with CHD.

using DAVID, which is frequently used in other studies. In this case study, only the cytokine–cytokine receptor interaction pathway was identified as a CHD-associated pathway by DAVID (p-value = 0.062, adjusted p-value = 0.81 with the Benjamini FDR correction method).

4. DISCUSSION

Comparison between gene sets can provide further knowledge relevant to the results of biological experiments. For two gene sets, simply calculating the reproducibility of the genes may not give the expected results and does not indicate their association at the functional level. This article shows that the GSFS can be used to quantify the functional association between two gene sets. The results of enrichment analysis are integrated into the computation of semantic similarity between categories and give a functional similarity score for gene sets. The significance evaluation shows whether two sets are functionally associated.

From the view of enrichment analysis, the sets of significant categories are a rather defined functional association. The advantage of the enrichment method is that it addresses the biology of the gene sets and takes into account the significance of the biological relevance of the associations between gene sets and categories. Therefore, weighting for functional categories, the GSFS could be more suitable for a functional comparison. Additionally, the similarity score is influenced by the number of gene sets. We have developed a statistical model based on the distribution of similarity that is obtained by randomly generating two gene sets of the same size. The probability is used to estimate the association for two gene sets. This computational comparison and systematic analysis method provides a new way to evaluate the associations between gene sets.

The correctness of the method is not completely justified because there is no gold standard for evaluating the functional association of gene sets. To evaluate the GSFS algorithm, we developed several tests based on protein complex data that were annotated with functional categories in CORUM; thus, the associations are known. First, the GSFS was used to quantify functional similarity, and BLAST was used to calculate sequence similarity between complex pairs. The GSFS scores were strongly correlated with the sequence similarity scores. Second, an analysis was performed to determine

whether the cluster result using the GSFS score was consistent with the complex-annotation catalogue. As expected, complexes with high functional relatedness in CORUM were assigned to a cluster. Third, the gene pairwise method and the gene set annotation method were compared with the GSFS, and the results showed that GSFS tended to be better for exploring functional associations between complexes.

In previous studies, the gene sets of the DEGs detected from different microarray studies for a phenotype or a condition have often been highly inconsistent. We used the GSFS method to analyse the association between two DEG sets for prostate tumour cells from different platforms and verified that the GSFS method is functionally consistent with the sample. Finally, by ranking the functional similarity scores from GSFS, we identified susceptibility pathways of CHD. The GSFS method may be a novel and efficient means for further studying the reproducibility of and for meta-analysis of different gene sets, which may be useful to identify the underlying functions of gene sets and to explore the molecular basis of complex human disorders.

This research was supported by the National Natural Science Foundation of China (grant nos. 31100948, 30871394, 61073136 and 91029717), the National High Tech Development Project of China, the 863 Programme (grant no. 2007AA02Z329), National Science Foundation of Heilongjiang Province (grant no. QC2009C23), the Science Foundation of Educational Commission of Heilongjiang Province (grant no. 11551233), the Graduate Innovation Fund of Heilongjiang Province (grant nos. YJSCX2009-226HLJ, YJSCX2011-334HLJ).

REFERENCES

1 Samani, N. J. *et al.* 2007 Genomewide association analysis of coronary artery disease. *N Engl. J. Med.* **357**, 443–453. (doi:10.1056/NEJMoa072366)
2 Cohen, A. M. & Hersh, W. R. 2005 A survey of current work in biomedical text mining. *Brief Bioinform.* **6**, 57–71. (doi:10.1093/bib/6.1.57)
3 Kim, T. *et al.* 2005 Downregulation of lipopolysaccharide response in Drosophila by negative crosstalk between the AP1 and NF-kappaB signaling modules. *Nat. Immunol.* **6**, 211–218. (doi:10.1038/ni1159)
4 Choi, Y. & Kendziorski, C. 2009 Statistical methods for gene set co-expression analysis. *Bioinformatics* **25**, 2780–2786. (doi:10.1093/bioinformatics/btp502)
5 Tusher, V. G., Tibshirani, R. & Chu, G. 2001 Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121. (doi:10.1073/pnas.091062498)
6 Shi, L. *et al.* 2006 The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161. (doi:10.1038/nbt1239)
7 Guo, L. *et al.* 2006 Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24**, 1162–1169. (doi:10.1038/nbt1238)
8 Li, Y. & Agarwal, P. 2009 A pathway-based view of human diseases and disease relationships. *PLoS ONE* **4**, e4346. (doi:10.1371/journal.pone.0004346)
9 Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. & Barabasi, A. L. 2007 The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690. (doi:10.1073/pnas.0701361104)
10 Basso, G., Case, C. & Dell'Orto, M. C. 2007 Diagnosis and genetic subtypes of leukemia combining gene expression and flow cytometry. *Blood Cells Mol. Dis.* **39**, 164–168. (doi:10.1016/j.bcmd.2007.05.004)
11 Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C. & Romero, R. 2007 A systems biology approach for pathway level analysis. *Genome Res.* **17**, 1537–1545. (doi:10.1101/gr.6202607)
12 Huang da, W., Sherman, B. T. & Lempicki, R. A. 2009 Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13. (doi:10.1093/nar/gkn923)
13 Falcon, S. & Gentleman, R. 2007 Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258. (doi:10.1093/bioinformatics/btl567)
14 Subramanian, A. *et al.* 2005 Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15 545–15 550. (doi:10.1073/pnas.0506580102)
15 Huang, E. J. & Reichardt, L. F. 2003 Trk receptors: roles in neuronal signal transduction. *Annu. Rev. Biochem.* **72**, 609–642. (doi:10.1146/annurev.biochem.72.121801.161629)
16 Chang, L. & Karin, M. 2001 Mammalian MAP kinase signalling cascades. *Nature* **410**, 37–40. (doi:10.1038/35065000)
17 Chen, Z., Gibson, T. B., Robinson, F., Silvestro, L., Pearson, G., Xu, B., Wright, A., Vanderbilt, C. & Cobb, M. H. 2001 MAP kinases. *Chem. Rev.* **101**, 2449–2476. (doi:10.1021/cr000241p)
18 Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. 2004 FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics* **20**, 578–580. (doi:10.1093/bioinformatics/btg455)
19 Soldatos, T. G., O'Donoghue, S. I., Satagopam, V. P., Jensen, L. J., Brown, N. P., Barbosa-Silva, A. & Schneider, R. 2001 Martini: using literature keywords to compare gene sets. *Nucleic Acids Res.* **38**, 26–38. (doi:10.1093/nar/gkp876)
20 Osier, M. V., Zhao, H. & Cheung, K. H. 2004 Handling multiple testing while interpreting microarrays with the gene ontology database. *BMC Bioinformatics* **5**, 124. (doi:10.1186/1471-2105-5-124)
21 Ashburner, M. *et al.* 2000 Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**, 25–29. (doi:10.1038/75556)
22 Pesquita, C., Faria, D., Falcao, A. O., Lord, P. & Couto, F. M. 2009 Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.* **5**, e1000443. (doi:10.1371/journal.pcbi.1000443)
23 Xu, T., Gu, J., Zhou, Y. & Du, L. 2009 Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to gene ontology. *BMC Bioinformatics* **10**, 240. (doi:10.1186/1471-2105-10-240)
24 Ruepp, A. *et al.* 2008 CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **36**, D646–D650. (doi:10.1093/nar/gkm936)
25 Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. & Krawetz, S. A. 2003 Global functional profiling of gene expression. *Genomics* **81**, 98–104. (doi:10.1016/S0888-7543(02)00021-6)
26 Khatri, P. & Draghici, S. 2005 Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595. (doi:10.1093/bioinformatics/bti565)

27 Bastos, H., Pesquita, B. T. C., Faria, D. & Couto, F. 2011 Application of gene ontology to gene identification, *in silico* tools for gene discovery. *Methods Mol. Biol.* **760**, 141–157.

28 Resnik, P. 1999 Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130.

29 Lin, D. 1998 An information-theoretic definition of similarity, semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 15th Int. Conf. on Machine Learning*, pp. 296–304. San Francisco, CA: Morgan Kaufmann.

30 Jiang, J. J. & Conrath, D. C. 1997 Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th Int. Conf. on Research on Computational Linguistics* (*ROCLING X*), Taiwan.

31 Pesquita, C., Pessoa, D., Faria, D. & Couto, F. 2009 CESSM: collaborative evaluation of semantic similarity measures. *Jornadas de Bioinformática 2009 Conference: Challenges in Bioinformatics, 3–6 November, Lisbon, Portugal.*

32 Pesquita, C., Faria, D., Bastos, H., Falcão, A. O. & Couto, F. 2007 Evaluating GO-based semantic similarity measures. In *ISMB/ECCB 2007 SIG Meeting Program Materials, Int. Society for Computational Biology.* Vienna, Austria: International Society for Computational Biology.

33 Couto, F. M. & Silva, M. J. 2011 Disjunctive shared information between ontology concepts: application to gene ontology. *J. Biomed. Semant.* **2**, 5. (doi:10.1186/2041-1480-2-5)

34 Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. 2003 Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275–1283. (doi:10.1093/bioinformatics/btg153)

35 Saldanha, A. J. 2004 Java Treeview: extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248. (doi:10.1093/bioinformatics/bth349)

36 Fröhlich, H., Speer, N., Poustka, A. & Beissbarth, T. 2007 GOSim: an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics* **8**, 166. (doi:10.1186/1471-2105-8-166)

37 Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. & Wang, S. 2010 GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978. (doi:10.1093/bioinformatics/btq064)

38 Singh, D. *et al.* 2002 Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1**, 203–209. (doi:10.1016/S1535-6108(02)00030-2)

39 Lapointe, J. *et al.* 2004 Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. Sci. USA* **101**, 811–816. (doi:10.1073/pnas.0304146101)

40 Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. 2000 Online mendelian inheritance in man (OMIM). *Hum. Mutat.* **15**, 57–61. (doi:10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G)

41 Kanehisa, M. & Goto, S. 2000 KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. (doi:10.1093/nar/28.1.27)

42 Waterhouse, R. N., Chang, R. C., Atuehene, N. & Collier, T. L. 2007 In vitro and in vivo binding of neuroactive steroids to the sigma-1 receptor as measured with the positron emission tomography radioligand [^{18}F]FPS. *Synapse* **61**, 540–546. (doi:10.1002/syn.20369)

43 Fareh, J., Touyz, R. M., Schiffrin, E. L. & Thibault, G. 1997 Cardiac type-1 angiotensin II receptor status in deoxycorticosterone acetate-salt hypertension in rats. *Hypertension* **30**, 1253–1259.

44 Luther, J. A. & Birren, S. J. 2009 Neurotrophins and target interactions in the development and regulation of sympathetic neuron electrical and synaptic properties. *Auton. Neurosci.* **151**, 46–60. (doi:10.1016/j.autneu.2009.08.009)

45 Tedgui, A. & Mallat, Z. 2006 Cytokines in atherosclerosis: pathogenic and regulatory pathways. *Physiol. Rev.* **86**, 515–581. (doi:10.1152/physrev.00024.2005)

46 Aukrust, P., Halvorsen, B., Yndestad, A., Ueland, T., Oie, E., Otterdal, K., Gullestad, L. & Damas, J. K. 2008 Chemokines and cardiovascular risk. *Arterioscler. Thromb. Vasc. Biol.* **28**, 1909–1919. (doi:10.1161/ATVBAHA.107.161240)

47 Amasyali, B., Kilic, A., Celik, T. & Iyisoy, A. 2010 A new frame in thromboembolic cardiovascular disease: adipocytokine. *Int. J. Cardiol.* **139**, 100–102. (doi:10.1016/j.ijcard.2008.06.082)

48 Marone, G., de Crescenzo, G., Adt, M., Patella, V., Arbustini, E. & Genovese, A. 1995 Immunological characterization and functional importance of human heart mast cells. *Immunopharmacology* **31**, 1–18. (doi:10.1016/0162-3109(95)00037-3)

49 Satoh, M., Ishikawa, Y., Minami, Y., Takahashi, Y. & Nakamura, M. 2008 Role of Toll like receptor signaling pathway in ischemic coronary artery disease. *Front Biosci.* **13**, 6708–6715. (doi:10.2741/3183)

50 Barry, S. P., Townsend, P. A., Latchman, D. S. & Stephanou, A. 2007 Role of the JAK-STAT pathway in myocardial injury. *Trends Mol. Med.* **13**, 82–89. (doi:10.1016/j.molmed.2006.12.002)

51 Davies, M. J. 1990 A macro and micro view of coronary vascular insult in ischemic heart disease. *Circulation* **82**, II38–II46.