FEBS Open Bio 2 (2012) 76–82

FEBS openbio

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

GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis

Hiromitsu Araki [a], Christoph Knapp [b], Peter Tsai [b], Cristin Print [a,b,*]

[a] Department of Molecular Medicine & Pathology, School of Medical Sciences, Faculty of Medical and Health Sciences, The University of Auckland, Private Bag 92019, Auckland, New Zealand
[b] Bioinformatics Institute, The University of Auckland, Private Bag 92019, Auckland, New Zealand

ARTICLE INFO

Article history:
Received 16 January 2012
Revised 4 April 2012
Accepted 12 April 2012

Keywords:
Database
Gene set
Enrichment analysis
Functional genomics

ABSTRACT

Most "omics" experiments require comprehensive interpretation of the biological meaning of gene lists. To address this requirement, a number of gene set analysis (GSA) tools have been developed. Although the biological value of GSA is strictly limited by the breadth of the gene sets used, very few methods exist for simultaneously analysing multiple publically available gene set databases. Therefore, we constructed GeneSetDB (http://genesetdb.auckland.ac.nz/haeremai.html), a comprehensive meta-database, which integrates 26 public databases containing diverse biological information with a particular focus on human disease and pharmacology. GeneSetDB enables users to search for gene sets containing a gene identifier or keyword, generate their own gene sets, or statistically test for enrichment of an uploaded gene list across all gene sets, and visualise gene set enrichment and overlap using a clustered heat map.

1. Introduction

With the rapid development of high-throughput measurement technologies such as next generation sequencing and microarrays, biologists can easily analyse cells and tissues on a whole-genome scale. To make biological sense of the results of these analyses, biologists usually need to comprehensively interpret the biological meaning of gene lists. These gene lists may for example represent mRNAs co-regulated by a drug or experimental condition. Determining whether the members of a gene list share biological features has been made more important by the recent realisation that the expression of transcription factor target sets [1] or RNAs encoding proteins of similar function [2,3], are more often correlated than would be expected by chance. These correlated groups of functionally related RNAs appear to be tissue-specific and conserved across evolution [4,5]. For these reasons, genomics researchers find it valuable to analyse gene sets as well as individual genes [6,7] and gene set analysis (GSA) is frequently employed when interpreting genomic data. GSA statistically assesses whether experimentally identified gene lists have a larger intersection with biologically relevant gene sets than expected due to chance.

A number of GSA tools have been proposed and used successfully over the past decade [7]. However, most of these tools focus on identifying or visualising statistically significant gene set enrichment in one gene set database at a time. There are a fewer published reports of simultaneous analysis of multiple biologically distinct gene sets databases and their cross-visualisation. This is despite the fact that the gene sets employed critically limit the results of any GSA, and affect these results just as much as the statistical analysis methodologies used [7]. At present, Gene Ontology (GO) [8] is used as a "gold standard" gene set by many GSA tools [7]. However, many biologists would prefer to perform GSA using a single meta-database that allowed the statistically robust interrogation of GO and many other types of gene sets databases simultaneously, followed by a cross-visualisation of the results. A few GSA meta-databases (some embedded within specific GSA tools) have already been generated, like ConceptGen [9], DAVID [10], GATHER [11], GeneSigDB [12], MSigDB [13], and WhichGenes [14]. These are very useful resources/tools for GSA. In addition to these tools, there are several commercial products that allow variations of GSA but due to their cost these are not available to many academic researchers. However there remain no databases that are specifically designed for GSA and provide a searchable interface with full coverage of the available pathway, medical and pharmacological datasets.

Therefore, in order to allow comprehensive GSA across multiple databases of different types, we have constructed GeneSetDB, a

Abbreviation: GSA, gene set analysis

* Corresponding author at: Department of Molecular Medicine & Pathology, School of Medical Sciences, Faculty of Medical and Health Sciences, The University of Auckland, Private Bag 92019, Auckland, New Zealand. Fax: +64 9 373 7492.

E-mail addresses: h.araki@auckland.ac.nz (H. Araki), c.knapp@auckland.ac.nz (C. Knapp), p.tsai@auckland.ac.nz (P. Tsai), c.print@auckland.ac.nz (C. Print).

2211-5463 © 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.fob.2012.04.003

Table 1
Sources databases included in GeneSetDB.

Subclass Name	Sources database	Reference/URL
Pathway	Biocarta	http://www.biocarta.com
	EHMN	[15]
	HumanCyc	[16]
	INOH	[17]
	NetPath	[18]
	PID	[19]
	Reactome	[20]
	SMPDB	[21]
	Wikipathways	[22]
Disease/Phenotype	CancerGenes	[23]
	HPO	[24]
	KEGG Disease	[25]
	MethCancerDB	[26]
	MethyCancer	[27]
	MPO	[28]
	SIDER	[29]
Drug/Chemical	CTD	[30]
	DrugBank	[31]
	MATADOR	[32]
	STITCH	[33]
	T3DB	[34]
Gene Regulation	MicroCosm Targets	[35]
	miRTarBase	[36]
	Rel/NF-κB target genes	http://bioinfo.lifl.fr/NF-KB
	TFactS	[37]
GO	Gene Ontology	[8]

comprehensive meta-database integrating 26 public databases. GeneSetDB allows users to identify and download the intersection between an individual gene or a gene list and gene sets in 26 databases. Moreover it allows users to statistically analyse the degree of enrichment of their gene list in gene sets and cross-visualise this enrichment in a clustered heatmap based on the overlap between the enriched gene sets.

2. Materials and methods

2.1. Gene set building

Data was downloaded from each source database with permission. Source databases were classified into five subclasses based on the database content: Pathway, Disease/Phenotype, Drug/Chemical, Genes Regulation and Gene Ontology (Table 1). Since different gene/protein identifiers are used in each database, Entrez gene ID was used as a representative identifier in GeneSetDB. The *Bioconductor* (http://www.bioconductor.org/) or *biomaRt* [38] bioinformatic resources were used in this identifier conversion. GeneSetDB is based primarily on human data; however, it supports mouse and rat gene lists by using the information in NCBI HomoloGene.

2.2. Enrichment analysis

In general, enrichment analysis/overrepresentation analysis is the statistical assessment of whether input gene list has a larger intersection with biologically relevant gene sets than expected by chance. GeneSetDB uses the hypergeometric distribution to calculate the probability of overrepresentation (shown as *P*-value). The calculation of this *P*-value is followed by multiple testing correction use the Benjamini and Hochberg method [39], the result of which is shown as a false discovery rate (FDR). Gene sets with less than 10 or more than 500 genes are not used in the enrichment analysis. The reference (background) gene set was the set of Entrez gene IDs that have at least one annotation in the union of the gene sets used in the analysis (e.g. Subclass Pathway, GO, etc.). The gene sets shown on the results page can be filtered based on FDR. GeneSetDB allows the use of several types of identifier for the input gene list, including official gene symbols and commercial microarray probe IDs. Each input identifier is converted to an Entrez gene ID using the *Bioconductor* or *biomaRt* resources [38]. GeneSetDB allows visualisation of gene set overlap with the submitted gene list in a clustered heatmap. The heatmap colors show the proportion of overlap between the gene sets.

2.3. Implementation

All gene sets are stored in a MySQL database management system, and the web interface is implemented using Apache, PHP, Javascript and HTML. The statistical package *R* is used for statistical calculations and for drawing clustered heatmaps. To use GeneSetDB users can paste gene lists into the web interface or upload

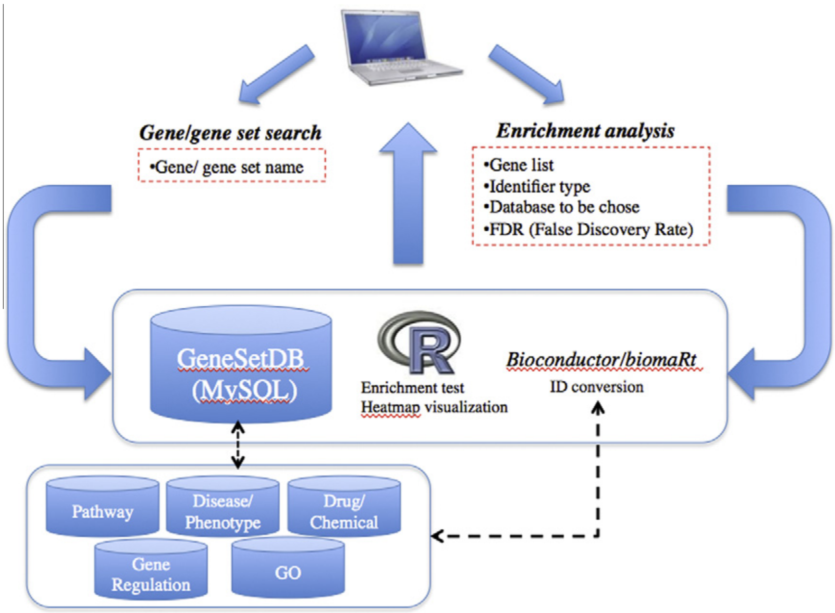

Fig. 1. Database structure and analysis scheme. The gene sets are downloaded from source databases and deposited into a MySQL database. All gene identifiers of both the source databases and the input gene list are converted into Entrez Gene ID using *Bioconductor* or *biomaRt*.