BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 12 2011, pages 1739–1740
doi:10.1093/bioinformatics/btr260

Databases and ontologies

Advance Access publication May 5, 2011

Molecular signatures database (MSigDB) 3.0

Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir,
Pablo Tamayo and Jill P. Mesirov*

Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Well-annotated gene sets representing the universe of the biological processes are critical for meaningful and insightful interpretation of large-scale genomic data. The Molecular Signatures Database (MSigDB) is one of the most widely used repositories of such sets.

Results: We report the availability of a new version of the database, MSigDB 3.0, with over 6700 gene sets, a complete revision of the collection of canonical pathways and experimental signatures from publications, enhanced annotations and upgrades to the web site.

Availability and Implementation: MSigDB is freely available for non-commercial use at http://www.broadinstitute.org/msigdb.

Contact: gsea@broadinstitute.org

Received on January 13, 2011; revised on April 4, 2011; accepted on April 12, 2011

1 INTRODUCTION

Microarrays and other high-throughput genomic technologies typically produce long lists of potentially interesting genes, which are not always easily interpreted. Recognizing the importance of coordinately expressed sets of genes, our seminal paper (Mootha *et al.*, 2003) introduced Gene Set Enrichment Analysis (GSEA) to discover metabolic pathways altered in human type 2 diabetes mellitus. GSEA and other analytical enrichment tools summarize genomic data in prioritized lists of higher-level biological features. As underscored by a recent survey of 68 enrichment tools, they critically depend on 'backend annotation databases' (Huang *et al.*, 2009). Typically, such databases focus on a particular domain of knowledge or annotation procedure. For example, Gene Ontology (GO) (Ashburner *et al.*, 2000) represents a hierarchy of controlled terms to describe individual gene products, while TRANSFAC (Matys *et al.*, 2006) stores information about transcription factor binding sites. A growing number of databases obtain sets from gene expression signatures reported in the literature. These include SignatureDB (Shaffer *et al.*, 2006), GeneSigDB (Culhane *et al.*, 2009), CCancer (Dietmann *et al.*, 2010) and L2L and LOLA (Cahan *et al.*, 2007).

Molecular Signatures Database (MSigDB) differs from these resources in several distinguishing aspects. (i) MSigDB is explicitly designed to provide gene sets for enrichment analysis methods. As such, it is natively and seamlessly integrated with our GSEA software (Subramanian *et al.*, 2005). (ii) MSigDB covers a substantially more diverse and wider range of gene set sources and types. These include signatures extracted from original research publications, and entire collections of sets derived from specialized resources such as GO, KEGG (Kanehisa and Goto, 2000), TRANSFAC and L2L. (iii) MSigDB gene sets are acquired both through manual curation and by automatic computational means, whereas other databases emphasize only one of these approaches. (iv) Finally, MSigDB contains the largest number of gene sets overall.

The initial MSigDB database, released in 2005 with GSEA software, contained 1325 sets. In contrast, MSigDB 3.0, released in September 2010, includes 6769 sets and a richer set of annotations. Here, we describe the MSigDB 3.0 sets in more detail and the accompanying online resource.

2 RESULTS

Gene set collections: gene sets in MSigDB 3.0 are organized into five collections according to their derivation:

C1: Genes located in the same chromosome or cytogenetic band.

C2: Gene sets representing canonical pathways from pathway resources [including 430 new sets contributed by Reactome (Matthews *et al.*, 2009)], and sets corresponding to chemical and genetic perturbations from 786 scientific publications.

C3: Sets of genes sharing *cis*-regulatory motifs in their promoter (transcription factor targets) or 3′UTR (micro-RNA targets) sequences.

C4: Clusters of coexpressed modules defined by computational analysis of large gene expression compendia.

C5: Gene sets corresponding to GO terms.

Table 1 shows the growth of the MSigDB collections and database since the initial release (see also online Release Notes).

Gene set annotations: each MSigDB gene set is a list of genes with relevant annotations and links to external resources. MSigDB focuses on human gene sets. However, we do include sets from some model organisms and gene set annotations include organism identification. We use HUGO gene symbols and, as of version 3.0, human Entrez Gene IDs serve as universal identifiers. These Entrez IDs are guaranteed to be unique and stable, can easily be mapped into a variety of other identifiers and are natively integrated with the GenBank resources of primary nucleic and protein sequences. We also preserve whatever original identifiers were used in the gene set source. All sets have unique database identifiers and names, and include brief and full descriptions. Other annotations depend on the type of gene set. Annotations linking to external resources are especially important as they allow researchers to place the sets in

*To whom correspondence should be addressed.

Table 1. MSigDB versions and changes in the number of gene sets

Gene set category	1.0 (2005)	2.5 (2008)	3.0 (2010)
C1: positional	319	386	326[a]
C2: curated (total)	522	1892	3272
C2: chemical and genetic perturbations	50	1186	2392
C2: canonical pathways	472	639	880
C2: uncategorized	0	66	0
C3: motifs (total)	57	837	836[a]
C3: transcription factor targets	57	500	615
C3: micro-RNA targets	0	222	221[a]
C3: uncategorized	0	115	0
C4: computational	427	883	881[a]
C5: GO terms	0	1454	1454
MSigDB total	1325	5452	6769

[a] Decrease in number due to the removal of sets with too few genes to run GSEA.

the context of a specific study and facilitate decisions on follow-up experiments.

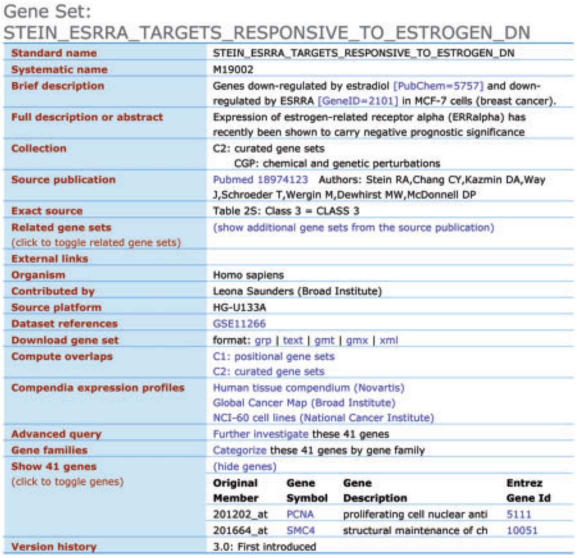
Gene sets from publications are the most richly annotated. Their annotations include the PubMed ID of the publication, pointers to other gene sets from the same publication, and now also details on the exact table or figure from which the gene set was extracted. For version 3.0, we updated the names of these gene sets to make them more descriptive and standardized and the accompanying brief descriptions to follow a more uniform and consistent format. Other annotation features introduced with version 3.0 include links to source datasets in Gene Expression Omnibus (GEO) (Barrett *et al.*, 2009) and ArrayExpress (Parkinson *et al.*, 2009). Canonical pathway sets include links to the pathway at the source web site.

File formats: MSigDB gene set files are available for download in plain text and XML formats. The plain text files contain simple listings of gene set membership, while the XML files also include the annotations. To ensure reproducibility of GSEA results, older versions of the MSigDB files are always available. Note that users of our GSEA software do not need to download the MSigDB files as the tool directly and automatically retrieves the gene sets.

3 MSigDB ONLINE RESOURCE

In version 3.0, we updated the MSigDB web site. First introduced in July 2007, the site allows users to view the annotated gene sets and perform simple search and analysis tasks. Each gene set and all of its annotations are presented on a separate web page (Fig. 1). Embedded hyperlinks connect annotations to corresponding external web resources, including PubMed, GEO and ArrayExpress, PubChem and Entrez Gene.

The MSigDB web site allows users to find gene sets by searching for keywords in the annotations. The online analysis tools allow users to: (i) compute overlaps between gene sets; (ii) view a heat map of a gene set in one of the reference expression compendia; and (iii) categorize the genes in a set by gene families. Gene families offer a quick view of a gene set by grouping its members into a small number of informative categories. We have updated the gene families and they now include: oncogenes, tumor suppressors, translocated cancer genes, transcription factors, protein kinases, homeodomain proteins, cell differentiation markers and cytokines/growth factors.



Gene Set:
STEIN_ESRRA_TARGETS_RESPONSIVE_TO_ESTROGEN_DN

<table>
<tr><td>Standard name</td><td>STEIN_ESRRA_TARGETS_RESPONSIVE_TO_ESTROGEN_DN</td></tr>
<tr><td>Systematic name</td><td>M19002</td></tr>
<tr><td>Brief description</td><td>Genes down-regulated by estradiol [PubChem=5757] and down-regulated by ESRRA [GeneID=2101] in MCF-7 cells (breast cancer).</td></tr>
<tr><td>Full description or abstract</td><td>Expression of estrogen-related receptor alpha (ERRalpha) has recently been shown to carry negative prognostic significance</td></tr>
<tr><td>Collection</td><td>C2: curated gene sets
CGP: chemical and genetic perturbations</td></tr>
<tr><td>Source publication</td><td>Pubmed 18974123 Authors: Stein RA,Chang CY,Kazmin DA,Way J,Schroeder T,Wergin M,Dewhirst MW,McDonnell DP</td></tr>
<tr><td>Exact source</td><td>Table 2S: Class 3 = CLASS 3</td></tr>
<tr><td>Related gene sets
(click to toggle related gene sets)</td><td>(show additional gene sets from the source publication)</td></tr>
<tr><td>External links</td><td></td></tr>
<tr><td>Organism</td><td>Homo sapiens</td></tr>
<tr><td>Contributed by</td><td>Leona Saunders (Broad Institute)</td></tr>
<tr><td>Source platform</td><td>HG-U133A</td></tr>
<tr><td>Dataset references</td><td>GSE11266</td></tr>
<tr><td>Download gene set</td><td>format: grp | text | gmt | gmx | xml</td></tr>
<tr><td>Compute overlaps</td><td>C1: positional gene sets
C2: curated gene sets</td></tr>
<tr><td>Compendia expression profiles</td><td>Human tissue compendium (Novartis)
Global Cancer Map (Broad Institute)
NCI-60 cell lines (National Cancer Institute)</td></tr>
<tr><td>Advanced query</td><td>Further investigate these 41 genes</td></tr>
<tr><td>Gene families</td><td>Categorize these 41 genes by gene family</td></tr>
<tr><td>Show 41 genes
(click to toggle genes)</td><td>(hide genes)
<table>
<tr><th>Original Member</th><th>Gene Symbol</th><th>Gene Description</th><th>Entrez Gene Id</th></tr>
<tr><td>201202_at</td><td>PCNA</td><td>proliferating cell nuclear anti</td><td>5111</td></tr>
<tr><td>201664_at</td><td>SMC4</td><td>structural maintenance of ch</td><td>10051</td></tr>
</table></td></tr>
<tr><td>Version history</td><td>3.0: First introduced</td></tr>
</table>

Fig. 1. A typical gene set page on the MSigDB web site. The list of genes has been abbreviated from 41 to 2 for the purposes of this figure.

ACKNOWLEDGEMENTS

We thank J. Roberston, L. Saunders and L. Kazmierski for gene set collection; H. Kuehn and J. McLaughlin for documentation; and M. Wrobel for web site development.

Funding: National Cancer Institute (5R01CA121941).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Barrett,T. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **30**, D5–D15.

Cahan,P. *et al.* (2007) Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, **401**, 12–18.

Culhane,A. *et al.* (2009) GeneSigDB – a curated database of gene expression signatures. *Nucleic Acids Res.*, **38**, D716–D725.

Dietmann,S. *et al.* (2010) CCancer: a bird's eye view on gene lists reported in cancer-related studies. *Nucleic Acids Res.*, **38** (Suppl), W118–W123.

Huang,da W. *et al.* (2009) *Nucleic Acids Res.*, **37**, 1-13.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Matthews,L. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.

Matys,V. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110.

Mootha,V. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Parkinson,H. *et al.* (2009) ArrayExpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.

Shaffer,A. *et al.* (2006) A library of gene expression signatures to illuminate normal and pathological lymphoid biology. *Immunol Rev.*, **210**, 67–85.

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**, 15545–15550.