Molecular Signatures Database (MSigDB) 3.0

Arthur Liberzon¹, Aravind Subramanian¹, Reid Pinchback¹, Helga Thorvaldsdóttir¹, Pablo Tamayo¹ and Jill P. Mesirov¹,∗
¹Broad Institute, Cambridge, MA 02142, USA

Associate Editor: Dr. 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 6,700 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

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 coordinate expressed sets of genes, our seminal paper (Mootha, 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, 2009). Typically, such databases focus on a particular domain of knowledge or annotation procedure. For example, Gene Ontology (GO) (Ashburner, 2000) represents a hierarchy of controlled terms to describe individual gene products, while TRANSFAC (Matys, 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, 2006), GeneSigDB (Culhane, 2009), CCancer (Dietmann, 2010), L2L and LOLA (Cahan, 2007).

MSigDB differs from these resources in several distinguishing aspects. (1) 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, 2005). (2) 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 & Goto, 2000), TRANSFAC and L2L. (3) MSigDB gene sets are acquired both through manual curation and by automatic computational means, whereas other databases emphasize only one of these approaches. (4) Finally, MSigDB contains the largest number of gene sets overall.

The initial MSigDB database, released in 2005 with GSEA software, contained 1,325 sets. By contrast, MSigDB 3.0, released in September 2010, includes 6,769 sets and a richer set of annotations. We describe here 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, 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 ex-
ternal resources are especially important as they allow researchers to place the sets in the context of a specific study and facilitate decisions on follow-up experiments.

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C1: positional</td>
<td>319</td>
<td>386</td>
<td>326*</td>
</tr>
<tr>
<td>C2: curated (total)</td>
<td>522</td>
<td>1,892</td>
<td>3,272</td>
</tr>
<tr>
<td>C2: chemical &amp; genetic perturbations</td>
<td>50</td>
<td>1,186</td>
<td>2,392</td>
</tr>
<tr>
<td>C2: canonical pathways</td>
<td>472</td>
<td>639</td>
<td>880</td>
</tr>
<tr>
<td>C2: uncategorized</td>
<td>0</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>C3: motifs (total)</td>
<td>57</td>
<td>837</td>
<td>836*</td>
</tr>
<tr>
<td>C3: transcription factor targets</td>
<td>57</td>
<td>500</td>
<td>615</td>
</tr>
<tr>
<td>C3: micro-RNA targets</td>
<td>0</td>
<td>222</td>
<td>221*</td>
</tr>
<tr>
<td>C3: uncategorized</td>
<td>0</td>
<td>115</td>
<td>0</td>
</tr>
<tr>
<td>C4: computational</td>
<td>427</td>
<td>883</td>
<td>881*</td>
</tr>
<tr>
<td>C5: GO terms</td>
<td>0</td>
<td>1,454</td>
<td>1,454</td>
</tr>
<tr>
<td>MSigDB Total</td>
<td>1,325</td>
<td>5,452</td>
<td>6,769</td>
</tr>
</tbody>
</table>

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

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, 2009) and ArrayExpress (Parkinson, 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 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.

Fig. 1. A typical gene set page on the MSigDB Web site. The list of genes has been abbreviated from 41 to two 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: This work was supported by the National Cancer Institute [5R01CA121941].

REFERENCES