Chapter 9

A Description of the Molecular Signatures Database (MSigDB) Web Site

Arthur Liberzon

Abstract

Annotated lists of genes help researchers to prioritize their own lists of candidate genes and to plan follow-up studies. The Molecular Signatures Database (MSigDB) is one of the most widely used knowledge base repositories of annotated sets of genes involved in biochemical pathways, signaling cascades, expression profiles from research publications, and other biological concepts. Here we provide an overview of MSigDB and its online analytical tools.

Key words Bioinformatics, Database, Genomics, Molecular sequence annotation, Gene expression profiling

1 Introduction

Many genomic and molecular biology studies report their findings in the form of gene lists. For example, the list could consist of genes belonging to a biochemical pathway, or of genes associated with a disease phenotype, etc. Png genes in such lists helps to figure out what they are doing in the cell. In addition, considering many genes as a group increases power of many popular analytical methods, as has been pioneered in Gene Set Enrichment Analysis (GSEA) [1]. Tools like GSEA rely on well-annotated collections of gene sets. In fact, we have originally developed the Molecular Signatures Database (MSigDB) to supply gene sets for GSEA. MSigDB quickly gained popularity as a stand-alone knowledge-base resource that can also be used independently of GSEA. Here, we will review main features of MSigDB and the accompanying suite of Web tools.

MSigDB is one of the largest and most widely used databases of gene sets [2]. Its most recent version, v4.0, released on May 31, 2013, contains 10,925 gene sets. Gene sets in MSigDB are lists of genes (in no particular order, each gene occurs only once in the set) with annotations and links to external sources.

Benjamin L. Kidder (ed.), *Stem Cell Transcriptional Networks: Methods and Protocols*, Methods in Molecular Biology, vol. 1150, DOI 10.1007/978-1-4939-0512-6_9, © Springer Science+Business Media New York 2014

GSEA/MSigDB Web site has several key components listed below. In this manual, our primary focus will be on the MSigDB home page.

1.1 Login/Register

The link is located in the upper right corner, near logo of the Broad Institute. If you have not registered before, you can do it by clicking this link. Registration is free for noncommercial users—its only purpose is to track usage for reports to our funding agencies. You can also register by following the corresponding link in Subheading 3.1 on GSEA or MSigDB home pages.

1.2 Navigation Tools

1.2.1 GSEA/MSigDB Navigation Strip

The horizontal navigation strip serves to quickly move throughout the entire site. Use links in this strip to move between these pages: GSEA home, Downloads, MSigDB home, Documentation, and Contact information. This navigation strip runs through all pages, except for the Documentation wiki. The navigation strip has links to the following pages:

GSEA Home

http://www.broadinstitute.org/gsea/index.jsp

This page provides a registration link and displays latest news about GSEA and MSigDB; information about members of the GSEA team, members of the scientific advisory board, and about funding agencies; and contact and acknowledgements to our contributors. Finally, there is also a note explaining how to cite use of GSEA.

Downloads

From this page, you can download GSEA software and MSigDB database files.

Molecular Signatures Database

http://www.broadinstitute.org/gsea/msigdb/index.jsp

This page provides an overview of MSigDB Web site with links to its Web analytics. It has an additional navigation menu on the upper left side to move around MSigDB pages. This page also includes directions for registration; notes about current version of MSigDB database and the Web site; information about members of the GSEA team, members of the scientific advisory board, and about funding agencies; contact and acknowledgements to our contributors; overview of MSigDB collections; contact information; and a note explaining how to cite use of MSigDB (*see GSEA Home* in Subheading 1.2.1 above).

Documentation

This link leads to pages with detailed documentation about the GSEA/MSigDB resource. It contains GSEA User Guide and Tutorial, FAQs, descriptions of data formats, release notes, and other documents.

Contact

Here you will find how to send us questions, suggestions, and other feedback information. For convenience, this information is reproduced at the end of GSEA and MSigDB home pages.

1.2.2 MSigDB Side Bar

This navigation menu appears on every MSigDB Web page, is restricted to these pages only, and allows you to quickly move around the entire MSigDB web site.

MSigDB Home

Clicking this link will bring you to the MSigDB home page described *Molecular Signatures Database* in Subheading 1.2.1 above.

About Collections

This link will bring up detailed information about individual collections and subcollections of MSigDB. The page is organized as a table with three columns, where the first column contains collection names and a link showing total number of sets in the collection. The middle column contains detailed description of each collection and recommendations for its use. The third column contains links to download gene sets as GMT files (*see* **Note 1**).

Browse Gene Sets

Clicking this link will let you browse gene sets by their name or collection (*see* Subheading 3.2 for details).

Search Gene Sets

Clicking this link will bring you to the search tool (*see* Subheading 3.3 for details).

Investigate Gene Sets

Click this to navigate to a suite of gene set analysis tools (*see Gene families* in Subheading 3.4).

View Gene Families

This link will direct you to a table with a functional overview of all MSigDB sets categorized into a small number of gene families. Gene families are special collections of gene products that share a common feature such as homology or molecular function.

Help

Follow this link to learn more about features listed in the MSigDB navigation side menu.

2 Materials

To explore MSigDB, you will need a computer with the Internet connection and a Web browser. To access MSigDB, registration is required. Registration is free for noncommercial users. Its only purpose is to help us track usage for reports to our funding agencies. After registering, you can log in at any time using your e-mail address.

3 Methods

3.1 Registration

Register to view and explore MSigDB contents. To register, go to the MSigDB home page (http://www.broadinstitute.org/gsea/msigdb/index.jsp) or GSEA home page (http://www.broadinstitute.org/gsea/index.jsp) and click on the word **register** in the Registration section. Alternatively, you can click on the word **register** at the top of these pages, near the Broad Institute logo (*see* Subheading 1.1).

3.2 Browse

Here you can browse gene sets by their name or collection. You can get there by clicking **Browse** in the **Overview** section of the MSigDB home page or by clicking **Browse Gene Sets** at the side bar navigation menu. This page has three parts.

In the top part, you can search by **gene set name**. For example, to search for sets with names containing word **stem**, type it in the form and click the **search** button. To see the results, scroll down to the end of the page. There are Web links to pages matching the search term in set names.

Alternatively, you can browse sets by first letter or number. For example, click on the letter **B** and then scroll down to the end of the page. There will be Web links to pages of sets with names starting from letter **B**. Note that the second set there is named **B_CELL_DIFFERENTIATION**.

Finally, you can browse sets by collection. To see a short info about each (sub) collection, move mouse over the question mark icon near it. For example, go to **C5** and click on **BP**. In the long list of links displaying all sets in the C5 BP subcollection, locate the link to **B_CELL_DIFFERENTIATION**.

3.3 Search

Here you can find gene sets by keyword, gene set name, collection, organism, or contributor. You can get to this tool by clicking the **Search** link in the Overview section of MSigDB home page. Alternatively, you can get there by clicking **Search Gene Sets** at the side bar menu. Accessing this tool requires registration. You can do the search by typing a keyword in the form below the word **Keywords** and click the **search** button. For details about keyword search, move your mouse over the question mark icon. You can apply a number of filters to your search using the scroll bar forms under **Search Filters**. For example, to search for sets containing word “stem,” type it in the keywords form and click “search.” Search results appear below the red line which reads “found 439 gene sets” in this case. The results appear in a table with up to 10 rows by default and six columns. To see next pages of the results, you can use navigation page numbers above the table. You can vary the number of hits per page from 10 (default) to 20, 50, or 100. For each gene set, the columns display its **name**, number of genes (**#genes**), **description**, its collection code (**collections**), **organism**, and **contributor** organization. Click on rows to select gene sets for subsequent action. For example, click on the row with **B_CELL_DIFFERENTIATION**. This set has 12 genes and comes from C5 BP subcollection. Notice that the results now show **1** gene set selected. To un-select, click the row again. Note that the number of sets selected reverts to 0. Click the set to select it again.

At the top of the table, there is a menu **Select An Action...** Clicking and holding it shows a number of action options available. Top five choices allow you to export selected gene set in a variety of standard GSEA formats.

3.4 Examine

To get detailed information about a single gene set, click a link with the set's name. You can come across these during your explorations in Subheadings 3.2 or 3.3. For example, go to the **Browse** tool by clicking **Browse Gene Sets** on the side menu. Then type **B_CELL_DIFFERENTIATION** in the form **Gene set name** and click search button. Scroll down the page to see the results—there should be only one link to set named **B_CELL_DIFFERENTIATION**. To examine the set, click on this link. This will bring you to a standard gene set page that has detailed description of the individual set. The page has four major sections with all fields described in detail in the documentation (*see* **Note 2**).

The first section consists of the set's annotations. All sets have unique database identifiers and names and include brief and full descriptions. We use HUGO gene symbols and human Entrez Gene IDs as universal gene identifiers. We also preserve original identifiers as they appeared in the gene set source. Other annotations depend on the type of gene set. Annotations linking to external resources are particularly important as they allow researchers to place the sets in the context of their origins [2].

The second section describes how to use MSigDB analysis tools to further investigate the set.

3.4.1 Download Gene Set

It allows you to download the set in one of our standard file formats. Gene set file formats are described in the **Documentation** section here: http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#Gene_Set_Database_Formats. The file contains a single gene set made of human gene symbols (*see* **Note 3**). Navigate to the gene set page for **B_CELL_DIFFERENTIATION** as described in Subheading 3.4 and download the set in GRP format by clicking on "format: grp." This should have downloaded the **B_CELL_DIFFERENTIATION** set in the GRP format. The specific location depends on the download settings of your browser.

3.4.2 Compute Overlaps

Examining genes shared by two sets can highlight common processes and reveal other useful relationships. This tool evaluates the overlap of a query gene set and an estimate of the statistical significance of the overlap with one or more MSigDB collections or subcollections (*see* **Note 4**). Overlaps can only be done against current version of MSigDB gene sets. Overlaps are computed using human gene symbols and the tool does any required conversion automatically.

Follow the steps in the first paragraph of Subheading 3.4 to navigate to the gene set page for **B_CELL_DIFFERENTIATION**. To display the results, click on a link to a gene set collection, e.g., CP:canonical pathways (*see* **Note 5**).

The results have three parts. At the top, there is a table summarizing conversion details from input gene identifiers to human gene symbols. The next part lists summary statistics for the overlaps as a table with rows corresponding to the number of MSigDB collections chosen for the analysis and the following columns:

Collections—indicates collection of sets selected

# Overlaps—lists the number of overlapping gene sets (*see* **Note 6**)

# Gene sets in collections—lists the total number of sets tested

# Genes in comparison (n)—lists the number of genes in your set

# Genes in Universe (N)—list the number of all known human gene symbols

The following table reports detailed overlap statistics for each gene set. A link above this table allows you to export the results as an Excel file. The table lists one set per row and has the following columns:

Gene Set Name [# Genes (K)]—link to the gene set page [number of genes]

Description—brief description of the set

# Genes in overlap k/k (k)—number of genes in the overlap

p-value—the significance of the overlap according to the hypergeometric distribution [3]

FDR *q*-value—the significance estimate after correcting for multiple hypothesis testing [4]

The final part contains the overlap matrix where rows are genes from the query set and columns are links to the overlapping sets.

3.4.3 Compendia Expression Profiles

This tool displays a profile of the gene set based on a selected compendium of expression data, such as human tissue compendium (Novartis) [5], global cancer map (Broad Institute) [6], or NCI-60 panel of cell lines (National Cancer Institute) [7].

Follow the steps in the first paragraph of Subheading 3.4 to navigate to the gene set page for **B_CELL_DIFFERENTIATION**. To display the results, click on a link to a compendium.

Alternatively, you can submit the set to this tool by clicking the further investigate link near the **Advanced query**. This will lead you to the **Investigate Gene Sets** page and your set will be pasted to the **Gene Identifiers** box. Choose one of the available compendia and click on “display expression profile.” The resulting heat map includes dendrograms clustering gene expression by gene and samples. Genes are indicated by probe set id, gene symbol, description, and gene family.

3.4.4 Gene Families

Gene families are special collections of gene products that share a common feature such as homology or molecular function. This feature highlights particularly interesting members of a set.

Follow the steps in the first paragraph of Subheading 3.4 to navigate to the gene set page for **B_CELL_DIFFERENTIATION**. Click on the Categorize link to retrieve an overview of the set with its members categorized into the gene families.

Alternatively, you can submit the set to this tool by clicking the further investigate link near the **Advanced query**. This will lead you to the **Investigate Gene Sets** page and your set will be pasted to the **Gene Identifiers** box. Click on “show gene families” to retrieve an overview of the set with its members categorized into the gene families.

3.4.5 Advanced Query

This action submits the set to a suite of analysis tools at **Investigate Gene Sets** page. See Subheading 3.4 for details.

4 Notes

1. We provide three kinds of GMT files depending on what type of gene identifiers is used to make gene sets. Thus, “original identifiers” correspond to whatever gene identifiers were used in the original source of the set. We provide these files for reference only and do not recommend using them for standard analyses because various sets can have different types of original identifiers. On the other hand, “gene symbols” correspond to the GMT file with sets made of human gene symbols. We use our own system to map all kinds of original identifiers to the space of human gene symbols. When the original identifiers stand for genes from species other than human, we map them to the corresponding orthologous human gene symbols. Biologists are familiar with human gene symbols and we thus recommend using GMT built from human genes symbols for most purposes. Computationally, working with gene symbols can present certain challenges because different genes can have the same symbols and the same gene can have a number of alternative gene symbols. For programmatic access, it is safer to work with more robust gene identifiers, such as NCBI Entrez Gene IDs. We thus provide a version of GMT files with genes made of human Entrez Gene IDs as well.

2. Detailed description of this page and its features is here: on the top horizontal navigation strip, click on **Documentation**. This will bring you to the documentation wiki pages. On the left side of this wiki page, go to section **msigdb** and click on **Guide to a GeneSetCard**.
3. To download many gene sets, click **Downloads** on the navigation strip and then locate a desired GMT file that contains your sets of interest. There you can also choose what kind of gene identifiers to have in the GMT file. Alternatively, click **Molecular Signatures Database** on the navigation strip and then go to **Search**. After the search is done, select as many sets as needed and then click on **Select An Action…** and export the sets in desired file format. This option will export sets as human gene symbols only.
4. To compute overlaps between several collections of gene sets, you should submit the set to this tool by clicking the further investigate link near the **Advanced query**.
5. From the **Investigate Gene Sets** page, click on the **compute overlaps** button to display the results.
6. By default, the report displays the 10 gene sets in the collection that best overlap with your gene set. If you compute overlaps from the **Investigate Gene Sets** page, you can choose the number of overlapping gene sets to display in the report by varying the top number of sets in the pull-down menu (10, 20, 50, or 100) or by changing the FDR *q*-value threshold.

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102(43):15545–15550
2. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27(12):1739–1740
3. Rivals I, Personnaz L, Taing L, Potier MC (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 23(4):401–417
4. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57(1):289–300
5. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101(16):6062–6067
6. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98(26):15149–15154
7. Shankavaram UT, Reinhold WC, Nishizuka S, Major S, Morita D, Chary KK, Reimers MA, Scherf U, Kahn A, Dolginow D, Cossman J, Kaldjian EP, Scudiero DA, Petricoin E, Liotta L, Lee JK, Weinstein JN (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. Mol Cancer Ther 6(3):820–832