MSigDB November 2005 Release Notes

This is a description of the curation process, content and file formats of the November 2005 release of the Molecular Signature Database (MSigDB) – a collection of gene sets.

Note: If you are looking for annotation about a specific gene set, you will *not* find it here. Instead, look it up in the GeneSetCard web page: http://www.broad.mit.edu/msigdb

MSigDB is organized into 4 broad categories.

1. C1: Positional gene sets
2. C2: Curated gene sets
3. C3: Motif gene sets
4. C4: Computed gene sets

SUMMARY OF CHANGES

DATABASE	# OF SETS MARCH 2005 RELEASE	# OF SETS NOVEMBER 2005 RELEASE
C1	319	396
C2	522	912
C3	57	623
C4	427	427 (no change)
TOTAL	1325	2358

DATABASE SUB-CATEGORY	# OF GENE SETS
C2:PATH: Pathway	501
C2:ONTO: Ontology	30
C2:EXCU: Expert curated	66
C2:LREV: Literature review	15
C2:PERT: Pertuberation	201
C2:CLIN: Clinical	65
C2:MODL: Animal model	34
C3:REPM	174
C3:TRANSFAC	449

CURATION DETAILS

Details on the curation and content of each category follow.

C1: Positional gene sets

In this database, genes from the same location of the genome are grouped together in a gene set. Currently, the grouping is based on cytogenetic bands. There are 24 sets corresponding to the genes on each of the 24 human chromosomes, as well as 372 sets corresponding to cytogenetic bands (total 396 gene sets).

Cytogenetic annotations come from 3 sources:

1) **HUGO**: For each gene symbol we looked up its cytogenetic location in the HUGO database (Release October 2005, http://www.gene.ucl.ac.uk/nomenclature/data/gdlw_index.html)
2) **Unigene**: We downloaded the Unigene release 180 from ftp://ftp.ncbi.nih.gov/repository/UniGene/Homo_sapiens/
3) **Array annotations**: Because several probe sets in common use map to genes that are not (yet) valid HUGO gene symbols, we also downloaded cytogenetic annotations for the Affymetrix human chips from (September 19, 2005 release) http://www.affymetrix.com/support/technical/byproduct.affx?cat=arrays&Human

We merged these databases, and derived a single cytogenetic band for every gene symbol. These were then grouped into gene sets. Decimals in cytogenetic bands were ignored. For example, 5q31.1 was considered as 5q31. Therefore, a gene annotated as 5q31.2 was placed in the same gene set (5q31) as a gene annotated as 5q31.3.

There are two main issues in building this database. First, some genes have more than 1 annotated location (sometimes even on different chromosomes). We simply omitted these genes from our gene sets. The second inconsistency occurs when annotations from the various sources disagree. In such cases, we picked the Unigene annotation.

C2: Curated gene sets

This database contains gene sets curated from a wide variety of sources. These sources include pathway database, publications and expression datasets. We label every gene set in this database with one or more sub-category flag that indicate its provenance. The sub-categories are described below.

PATH: Pathway gene sets

A major component of the C2 category are gene sets curated from pathway databases. Usually, these gene sets are canonical representations of a biological process compiled by an expert curator. We term this the PATH sub-category and it includes gene sets from the following 9 sources.

1) BioCarta: http://www.biocarta.com/
2) Signaling pathway database: http://www.grt.kyushu-u.ac.jp/spad/menu.html
3) Signaling gateway: http://www.signaling-gateway.org/
4) Signal transduction knowledge environment: http://stke.sciencemag.org/
5) Human protein reference database: http://www.hprd.org/
6) GenMAPP: http://www.genmapp.org/
7) Sigma Aldrich pathways: http://www.sigmaaldrich.com/Area_of_Interest/Biochemicals/Enzyme_Explorer/Key_Resources.html
8) Gene arrays, BioScience corporation: http://www.superarray.com/
9) Human cancer genome anatomy consortium: http:/cgap.nci.nih.gov/

Compared to the previous (March 2005) release, we have made a large increase in the number of gene sets from GenMAPP (increased from 57 to 177). The number of gene sets from BioCarta increased by 232 to 272. The other pathway databases had no additions.

ONTO: Gene sets from ontologies

These are gene sets curated from structured, controlled vocabulary efforts such as the Gene Ontology (GO). The GO describes gene products in terms of their associated biological biological processes, cellular components and molecular functions in a species-independent manner. However, because GO is hierarchical, gene sets made from every ontological class would be very redundant. Some groups have proposed methods to truncate the tree and only pick some of the nodes in the tree. In our effort, we simply picked from all GO terms, those that were most relevant (as determined by in-house biologists) to Cancer. This of course, is a very simplistic use of GO. Because excellent resources already exist in this regard, our emphasis in MSigDB has been on other (non-ontological) gene sets.

EXCR: Expert curated gene sets

These are gene sets that have been curated by local experts. Typically, they represent the application of domain knowledge of collaborators and local experts. Hence, no external PubMed reference exists.

LREV: Gene sets from a review article

These are gene sets collected from a specific review article in the literature. Thus, unlike EXCR sets, these gene sets are directly traceable to a table (or listing) in a peer-reviewed review article. LREV gene sets differ from PERT gene sets in that they are not linked to a single experiment but typically represent a gestalt of an expert’s knowledge of the pathway. An example is the X-inactivation gene set from (Disteche et al., 2002).

PERT: Perturbation gene sets

These are gene sets representing gene expression signatures of genetic and chemical perturbations. Typically, these gene sets come in pairs: an xxx_UP gene set representing genes turned on by the pertuberation and an xx_DOWN gene set representing genes repressed by the treatment (up in controls). They are curated from published experimental results and hence each gene set is linked to a specific PubMed ID.

CLIN: Clinical gene sets

These are gene sets that represent a clinical phenotype. For example, a gene expression signature of resistance to a treatment regimen would constitute a clinical gene set.

MODL: Gene sets from animal models

These are gene sets derived from animal models. For example gene expression signatures of mice with targeted mutations (Sweet-Cordero et al., 2005).

C3: Motif gene sets

Each set contains genes that lie downstream of a motif that is conserved across the human, mouse, rat, and dog genomes. The motifs are catalogued in Xie, *et al.* 2005 and represent known or likely regulatory elements in promoters and 3'-untranslated regions.

In the previous release (March 2005) the clustered motif gene sets were released (57 sets). In this release, we also make available as part of MSigDB, the entire collection of gene sets discovered by Xie. For details on the how the gene sets were colected please see Xie, *et al.* 2005 and the supplementary information at http://www.nature.com/nature/journal/v434/n7031/extref/nature03441-s1.pdf

REPM: Gene sets from Representative Motifs

These motifs gene sets are derived from clustering of motifs discovered by (Xie et al., 2005). Please refer to their paper for details on the clustering (Note: clustering of motif sequences, *not* expression data). The 173 gene sets in this collection are a superset of the 57 sets included in the March 2005 release (which were only 8-mers, the current collection also includes conserved 6-mers). Note that the first motif listed in “Supplementary Table S3 Motifs discovered in promoters in clusters” of (Xie et al., 2005) is used as the “representative motif” sequence. 69 of these representative motifs have close matches to a known binding site in TRANSFAC and are hence annotated with a specific transcription factor. The rest are highly conserved motifs but it is not known what transcription factor binds to them.

TFAC: TRANSFAC motif gene sets

These are gene sets that contain a transcriptional regulator motif from the TRANSFAC (version 7.4, http://www.gene-regulation.com/) database. Hence, each motif gene set in this database is annotated with a TRANSFAC record, an n-mer motif sequence and a specific transcription factor. The computational process begins by picking a motif (n-mer) from TRANSFAC and then identifying genes that have this motif in their aligned promoter regions across all 4 mammalian species. All such genes are placed into a gene set. The process is repeated for all ~450 motifs extracted from TRANSFAC. Note that some transcription factors have > 1 motif gene set because they have > 1 motif sequence (i.e. putative binding site) in TRANSFAC. Please see Xie *et al* for details.

Note: If you use a C3 gene set for your publication, please cite (Xie et al., 2005)

C4: Computational gene sets

These gene sets are *unchanged* from the previous March 2005 release. Excerpted below are notes from that release.

We curated a list of 380 cancer associated genes internally and from a published cancer gene database (6). We then defined neighborhoods around these genes in four large gene expression data sets:

(*i*) Novartis normal tissue compendium (7),
(*ii*) Novartis carcinoma compendium (8),
(*iii*) Global cancer map (9), and
(*iv*) An internal large compendium of gene expression data sets, including many of our in-house Affymetrix U95 cancer samples (1,693 in all) from a variety of cancer projects representing many different tissue types, mainly primary tumors, such as prostate, breast, lung, lymphoma, leukemia, etc.

Using the profile of a given gene as a template, we ordered every other gene in the dataset by its Pearson correlation coefficient. We applied a cutoff of $R \geq 0.85$ to extract correlated genes. The calculation of neighborhoods is done independently in each compendium. In this way, a given oncogene may have up to four “types” of neighborhoods according to the correlation present in each compendium. Neighborhoods with <25 genes at this threshold were omitted yielding the final 427 sets.

CURATION PROCESS

The main aspects of curating gene sets are:

1) Converting a gene set into electronic form: This involves manually copying from a table on a web page or PDF file the members of the gene set.
2) We capture the gene set in terms of its original (i.e. reported) accessions. This is important because accession to gene symbol mappings change over time as.
3) We run a program that converts accessions to gene symbols.
4) We use an alias database (containing aliases compiled from HUGO, Unigene and internally) to map commonly used gene names to their valid official gene symbol (for example p53 to TP53).
5) We use a sequence accession database (compiled from HUGO and Unigene) to convert accessions to gene symbols
6) For gene sets reported in an array format (i.e. Affymetrix HG_U95Av2) we use the array annotation file to map the probe sets to gene symbols.

Note: We do *not* restrict valid gene symbols only to those in HUGO. This is because several array probe sets map to gene symbols that are in Unigene but are not yet officially approved symbols. We provide an annotation file called Gene_Symbol.chip that lists all symbols that are recognized in the gene set curation process.

FILE FORMATS

There are 4 main gene set file formats:

1) Gene set (grp): This is a plain text file that represents a single gene set. Each member needs to be on a new line. Gene matrix (gmx): each column of this file corresponds to
2) Gene matrix (gmx): This is a plain text, tab-delimited file that represents a collection of gene sets. Each column represents a gene set. The first line of the file are gene set names, the second line a field that is ignored (i.e. place a description there). The members of the gene set begin on the third line. This format is convenient to handle a small number of gene sets in MS Excel (i.e. < 256).
3) Gene matrix transposed (gmt): This is a plain text tab-delimited file that represents a collection of gene sets. Each row represents a gene set. The

first column are gene set name, the second column is a field that is ignored and the members of the gene sets begin on the third column.

4) GeneSetAnnotation.xml: This is a new file format introduced in this release. It is an XML formatted file that contains both the gene set as well as extensive annotation about the gene set. See the msigdb.dtd for details on the format.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--
        This is the DTD for MSigDB.
        It describes a format for a database of gene sets that captures
        both the content of a gene set (i.e gene members) as well as annotation about
        the gene set.
        @version 1.0 November 2005
-->

<!ELEMENT MSigDB GeneSet*             -- top level container -->
<!ATTLIST MSigDb
        NAME   CDATA  #REQUIRED       -- name of the database
>

<!ELEMENT GeneSet                     -- a single gene set -->
<!ATTLIST GeneSet
        STANDARD_NAME         CDATA   #REQUIRED      -- an english name for the gene set
        SYSTEMATIC_NAME               CDATA  #REQUIRED      -- a standardized identifier
        ORGANISM              CDATA   #REQUIRED      -- organism in which the geneset was
generated
        EXTERNAL_DETAILS_URL  CDATA   #REQUIRED      -- 3rd party url for more info
        CHIP                  CDATA   #REQUIRED      -- platform on which the geneset was
generated
        CATEGORY_CODE         CDATA   (C1|C2|C3|C4)
        SUB_CATEGORY_CODE     CDATA   (CYTO|PATH|ONTO|EXCU|PERT|MODL|CLIN|REPM|TFAC|NEGH)
        CONTRIBUTOR           CDATA   #REQUIRED      -- name of person/database that curated
the set
        PMID                  CDATA                  -- PubMed ID if available
        DESCRIPTION           CDATA                  -- a full detailed desc or abstract
        DESCRIPTION_BRIEF     CDATA   #REQUIRED      -- a brief 1 line description
        %MESH                 CDATA                  -- keywords and MESH identifiers
        %MEMBERS              CDATA   #REQUIRED      -- Genes in the original format
        %MEMBERS_SYMBOLIZED   CDATA   #REQUIRED      -- Genes after converting to gene
symbols
>
```

Up-to-date version is always at: http://www.broad.mit.edu/gsea/dtd/msigdb.dtd

Note: For the grp, gmx and gmt file formats, lines that begin with'#' (hash) are ignored (i.e. comments). White space before and after each member is trimmed.

MSIGDB GENE SETS ENTRY FORMAT

SYSTEMATIC_NAME – MSigDB ID: Unique numeric identifier assigned by us when the gene set is created.

STANDARD_NAME – Gene Set Name: This is a unique pure ASCII name (words linked with underscores) intended to capture the nature of the set in a short string. This is what appears in the GSEA output as “gene set”. It is defined according to *MSigDB’s naming convention*.

CONTRIBUTOR - the person that created the gene set in the first place (e.g. Jean, Aravind, Pablo).

REVIEWER – typically a biologist that reviews the gene set and may make changes etc. (e.g. Ben, Mike).

REVIEW_DATE: date of set review e.g. 6/29/2005

DESCRIPTION – a few lines describing in more detail the characteristics and salient features of the set.

CATEGORY_CODE – the type of set from {Experimental, Manually Curated, Database, Neighborhood, Motif, Computational or Other}. The SOURCE (see below) for each of these types is as follows:

PMID – The detailed source of the set (see above): according to the category of the set this is PUBMED ID for publication, or the name of the database, curator e-mail etc.

GEOID - the entry for the dataset where the set was defined as for example identifiers for GEO(gene Omnibus or ArrayExpress).

ORGANISM – The organism where this set was obtained

CHIP - the type of gene identifiers that the original set uses. For example Genbank ID, HUGO gene symbols, etc. or the type of probe and microarray used: Affymetrix U95av2, Agilent x, Stanford image clones etc.

MEMBERS - the list of genes in their original **GENE ID TYPE format**

MEMBERS_SYMBOLIZED - the list of genes using HUGO gene symbols (we will create this computationally).

WEIGHTS – a set of signed or unsigned weights (e.g. S2N ratios, Bayesian priors) that is associated with each gene for future applications that may use it. The order is the same as the genes. If there are no weights the entry should be NULL.

Gene Set Cards

A significant aspect of this release is the availability of single web page capturing annotation about a gene set. The web page presents a concise description of the gene set in a standard format. Hyperlinks from the Gene Set Card web page lead the user to relevant information about, for instance, the publication from which the gene set was derived.

REFERENCES

Disteche, C. M., Filippova, G. N., and Tsuchiya, K. D. (2002). Escape from X inactivation. Cytogenet Genome Res *99*, 36-43.
Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R., and Jacks, T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nat Genet *37*, 48-55.
Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.