MSigDB Collections: Details and Acknowledgments

H collection: Hallmark gene sets

We envision this collection as the starting point for your exploration of the MSigDB resource and GSEA. Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying gene set overlaps and retaining genes that display coordinate expression. The hallmarks reduce noise and redundancy and provide a better delineated biological space for GSEA. We refer to the original overlapping gene sets, from which a hallmark is derived, as its 'founder' sets. Hallmark gene set pages provide links to the corresponding founder sets for deeper follow up.

This collection is an initial release of 50 hallmarks which condense information from over 4,000 original overlapping gene sets from v4.0 MSigDB collections C1 through C6. We refer to the original gene sets as "founder" sets.

Hallmark gene set pages provide links to the corresponding founder sets for more in-depth exploration. In addition, hallmark gene set pages include links to microarray data that served for refining and validation of the hallmark signatures.

To cite your use of the collection, and for further information, please refer to *Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417-425*.

C1 collection: Positional gene sets

Gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene. Cytogenetic locations were parsed from HUGO, October 2006, and UniGene, build 197. We merged the relevant annotations from these resources and derived a single cytogenetic band location for every gene symbol. These were then grouped into sets. Decimals in cytogenetic bands were ignored. For example, 5q31.1 was considered 5q31. Therefore, genes annotated as 5q31.2 and those annotated as 5q31.3 were both placed in the same set, 5q31. When there were conflicts, the UniGene entry was used. These gene sets can be helpful in identifying effects related to chromosomal deletions or amplifications, dosage compensation, epigenetic silencing, and other regional effects.

C2 collection: Curated gene sets

Gene sets curated from various sources such as online pathway databases, the biomedical literature, and contributions from domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into the following two sub-collections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP).

> C2 sub-collection CGP: Chemical and genetic perturbations

Gene sets that represent expression signatures of genetic and chemical perturbations.

Most of the CGP sets came from the biomedical literature. Over the past several years, microarray studies have identified signatures of several important biological and clinical states (e.g. cancer metastasis, stem cell characteristics, drug resistance). The C2 collection makes many of these signatures, originally published as tables in a paper, available as gene sets. To do this, we compiled a list of microarray articles with published gene expression signatures and, from each article, extracted one or more gene sets from tables in the main text or supplementary information. A number of these gene sets come in pairs: xxx_UP (and xxx_DN) gene sets representing genes induced (and repressed) by the perturbation. The majority of CGP sets were curated from publications. They include links to the PubMed citation, the exact source of the set (e.g., Table 1), and links to any corresponding raw data in GEO or ArrayExpress repositories. When the gene set involves a genetic perturbation, the set's brief description includes a link to the gene's entry in the NCBI Entrez Gene database. When the gene set involves a chemical perturbation, the set's brief description includes a link to the chemical's entry in the NCBI PubChem Compound database.

Other CGP gene sets include:

Gene sets contributed by the **L2L database** of published microarray gene expression data at University of Washington. See *Newman JC, Weiner AM. L2L: a simple tool for discovering the hidden significance in microarray expression data. Genome Biol. 2005;6(9):R81*. See also http://depts.washington.edu/l2l.

Gene sets curated by Dr. Chi Dang from the **MYC Target Gene Database** at Johns Hopkins University School of Medicine. See *Zeller KI, Jegga AG, Aronow BJ, O'Donnell KA, Dang CV. An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets. Genome Biol. 2003;4(10):R69*.

A number of individuals have contributed gene sets to this collection. The gene set annotation includes a "contributor" field that acknowledges the contributor by name/affiliation.

> C2 sub-collection CP: Canonical pathways

The pathway gene sets are curated from the following online databases:

BioCarta: http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways. Note also special terms for these gene sets in the MSigDB license.

KEGG: http://www.pathway.jp. Note also special terms for these gene sets in the MSigDB license.

Matrisome Project: From the Hynes Lab at MIT. http://matrisomeproject.mit.edu. See also *Naba A, Clauser KR, Hoersch S, Liu H, Carr SA, Hynes RO. The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. Mol Cell Proteomics. 2012 Apr;11(4):M111.014647*.

Pathway Interaction Database: The National Cancer Institute (NCI) Pathway Interaction Database (PID) http://pid.nci.nih.gov. Now available via the NDEx database (http://www.ndexbio.org) hosted by the Ideker Lab at University of California, San Diego.

Reactome: http://www.reactome.org

SigmaAldrich: http://www.sigmaaldrich.com/life-science.html

Signal Transduction KE: 27 gene sets were derived from the AAAS/STKE Cell Signaling Database, which is no longer maintained. See http://stke.sciencemag.org/about/help/cm. See also *Westfall PJ, Ballon DR, Thorner J, High Osmolarity Glycerol (HOG) Pathway in Yeast. Science Signaling (Connections Map in the Database of Cell Signaling)*. Note also special terms for these gene sets in the MSigDB license.

Signaling Gateway: The Signaling Gateway is hosted by the San Diego Supercomputer Center at University of California, San Diego. http://www.signaling-gateway.org.

SuperArray SABiosciences: http://www.sabiosciences.com/ArrayList.php

C3 collection: Motif gene sets

Gene sets representing potential targets of regulation by transcription factors or microRNAs. The sets consist of genes grouped by short sequence motifs they share in their non-protein coding regions. The motifs represent known or likely cis-regulatory elements in promoters and 3'-UTRs. These gene sets make it possible to link changes in an expression profiling experiment to a putative *cis*-regulatory element. The C3 collection is divided into two sub-collections: microRNA targets (MIR) and transcription factor targets (TFT).

> C3 sub-collection MIR: microRNA targets

These sets consist of genes sharing 7-nucleotide motifs in their 3' untranslated regions. Each 7-mer motif matches (is complementary to) the seed (bases 2 through 8) of the mature human microRNA (miRNAs) catalogued in v7.1 of miRBase (October 2005).

> C3 sub-collection TFT: transcription factor targets

Gene sets that share upstream *cis*-regulatory motifs which can function as potential transcription factor binding sites. We used two approaches to generate these motif gene sets.

Gene sets of "conserved instances" consist of the inferred target genes for each motif ***m*** of 174 motifs highly conserved in promoters of four mammalian species (human, mouse, rat and dog). The motifs represent potential transcription factor binding sites and are catalogued in *Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005 Mar 17;434(7031):338-45*. Each gene set consists of all human genes whose promoters contained at least one conserved instance of motif ***m*** , where a promoter is defined as the non-coding sequence contained within a 4-kilobase window centered at the transcription start site (TSS).

Mammalian transcriptional regulatory motifs were extracted from v7.4 TRANSFAC database (see supplementary data of *Xie et al*). Each gene set consists of all human genes whose promoters contains at least one conserved instance of the TRANSFAC motif, where a promoter is defined as the non-coding sequence contained within a 4-kilobase window centered at the transcription start site (TSS).

C4 collection: Computational gene sets

Computational gene sets defined by mining large collections of cancer-oriented microarray data. This collection is divided into two sub-collections: Cancer gene neighborhoods (CGN) and Cancer modules (CM).

> C4 sub-collection CGN: Cancer gene neighborhoods

In our GSEA paper, Subramanian, Tamayo et al. 2005, PNAS 102, 15545-15550, we mined 4 expression compendia datasets for correlated gene sets, starting with a list of 380 cancer-associated genes curated from internal resources and *Brentani, Caballero et al. Human Cancer Genome Project/Cancer Genome Anatomy Project Annotation Consortium.; Human Cancer Genome Project Sequencing Consortium. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A. 2003 Nov 11;100(23):13418-23*. Using the profile of a given gene as a template, we ordered every other gene in the data set 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.

GNF2: Human tissue compendium (Novartis). Gene expression profiles from the Novartis normal tissue compendium, as published in *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. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A. 2004 Apr 20;101(16):6062-7*.

CAR: Novartis carcinoma compendium (Novartis). Gene expression profiles from the Novartis normal tissue compendium, as published in *Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF Jr, Hampton GM. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. 2001 Oct 15;61(20):7388-93.*

GCM: Global Cancer Map (Broad Institute). Gene expression profiles from the global cancer map, as published in *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. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A. 2001 Dec 18;98(26):15149-54*.

MORF: An unpublished compendium of gene expression data sets, including many of Broad Institute's Cancer Program in-house Affymetrix HG-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.

> C4 sub-collection CM: Cancer modules

Gene sets defined by *Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet. 2004 Oct;36(10):1090-8*. Briefly, the authors compiled gene sets ('modules') from a variety of resources such as KEGG, GO, and others. By mining a large compendium of cancer-related microarray data, they identified 456 such modules as significantly changed in a variety of cancer conditions. See also http://robotics.stanford.edu/~erans/cancer.

C5 collection: Gene Ontology (GO) gene sets

Gene sets in this collection are derived from Gene Ontology (GO) annotations. GO is a collaborative effort to develop and use ontologies to support biologically meaningful annotation of genes and their products. A GO annotation consists of a GO term associated with a specific reference that describes the work or analysis upon which the association between a specific GO term and gene product is based. A gene product might be associated with one or more GO terms. Each annotation also includes an evidence code to indicate how the annotation to a particular term is supported (http://geneontology.org/page/guide-go-evidence-codes).

The gene sets in the C5 collection are based on GO terms (go-basic.obo, downloaded on 3 May, 2016) and their associations to human genes (gene2go, downloaded from NCBI FTP server on 3 May, 2016). The GO terms in the collection belong to one of three GO ontologies: **molecular function (MF), cellular component (CC)** or **biological process (BP),** and the collection is divided into sub-collections accordingly. We omitted GO terms for very broad

categories that would produce extremely large gene sets. GO terms that produced gene sets with fewer than 10 genes have also been omitted. We defined sets as "highly similar" if their Jaccard's coefficient was > 0.85. For each pair of highly similar sets we kept only the larger set, and repeated the procedure until all such pairs were resolved.

Note to GSEA users: Gene set enrichment analysis identifies gene sets consisting of *co-regulated* genes; GO gene sets are based on ontologies and *do not* necessarily comprise co-regulated genes.

C6 collection: Oncogenic signatures

Gene sets represent signatures of cellular pathways which are often dis-regulated in cancer. The majority of signatures were generated directly from microarray data from NCBI GEO or from internal unpublished profiling experiments which involved perturbation of known cancer genes. In addition, a small number of oncogenic signatures were curated from scientific publications.

C7 collection: Immunologic signatures

Immunologic signatures collection (also called *ImmuneSigDB*) is composed of gene sets that represent cell types, states, and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology.

We first captured relevant microarray datasets published in the immunology literature that have raw data deposited to Gene Expression Omnibus (GEO). For each published study, the relevant comparisons were identified (e.g. WT vs. KO; pre- vs. post-treatment etc.) and brief, biologically meaningful descriptions were created. All data was processed and normalized the same way to identify the gene sets, which correspond to the top or bottom genes (FDR < 0.02 or maximum of 200 genes) ranked by mutual information for each assigned comparison.

The immunologic signatures collection was generated as part of our collaboration with the Haining Lab at Dana-Farber Cancer Institute and the Human Immunology Project Consortium (HIPC). To cite your use of the collection, and for further information, please refer to *Godec J, Tan Y, Liberzon A, Tamayo P, Bhattacharya S, Butte A, Mesirov JP, Haining WN, Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation, 2016, Immunity 44(1), 194-206*.

MSigDB database v6.1 updated October 2017
GSEA/MSigDB web site v6.3 released January 2018