

ARCHIVED: This documentation refers to an old version of the gene sets database. Please refer to the MSigDB website (<http://www.broad.mit.edu/msigdb> for the current version).

Version 1.0 March 2005


The gene sets described here correspond to the database used in (Subramanian & Tamayo et al., 2005)


GENE SETS ARE AVAILABLE FOR DOWNLOAD FROM:

http://www.broad.mit.edu/msigdb/archived_releases/

RELEASE NOTES

The power of GSEA is a function of how well the gene sets used to assess enrichment in a condition of interest represent meaningful coordinated or concordant gene expression behavior that reflects actual biological processes or states. The better they represent specific transcriptional programs or processes relevant for a particular cell state the better they will work as GSEA "probes". Because of this the definition and curation of gene sets is of paramount importance and deserves a systematic effort. This page contains an initial release of a collection of molecular signatures used in our [2005 PNAS GSEA paper](#).


 **Database C1 (chromosomal location):** This database consists of 24 sets corresponding to the genes on each of the 24 human chromosomes, as well as 301 sets corresponding to cytogenetic bands. This database can be helpful in identifying effects related to epigenetic silencing, dosage compensation, copy number polymorphisms, and aneuploidy or other chromosomal deletions/amplifications.


 **Database C2 (functional):** This database includes 475 metabolic and signaling pathways gleaned from the following 10 publicly available manually curated databases:

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. Gene ontology: <http://www.geneontology.org/>
8. Sigma Aldrich pathways:
http://www.sigmaaldrich.com/Area_of_Interest/Biochemicals/Enzyme_Explorer/Key_Resources.html

9. Gene arrays, BioScience corporation: <http://www.superarray.com/>
10. Human cancer genome anatomy consortium: <http://cgap.nci.nih.gov/>

In addition, there are 51 sets representing gene expression signatures of genetic and chemical perturbations that have been culled from experimental results in the literature.

 **Database C3 (motif-based):** 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.

 **Database C4 (correlational):** Correlation gene sets are groups of genes defined by computationally mining large-scale experimental datasets for co-expressed genes.

The databases contain one line for each gene set (probe id's separated by tabs) and are microarray type specific.

These will become part of a new resource database for "molecular signatures" that we call MSigDB (Molecular Signatures Database). This resource database will be expanded over time to include additional gene sets and other more general types of molecular signatures (ranked and signed marker lists, extracted features, itemsets, neighborhoods, probabilities, weighted vectors, matrices, graphs etc.).

REFERENCES

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). From the Cover: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-15550.