MSigDB7 and Beyond

Anthony S. Castanza, PhD
Curator, Molecular Signatures Database
Mesirov Lab, Department of Medicine

ITCR PI Meeting – May 30, 2019
Overview

**Gene Set Enrichment Analysis** (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this web site, you can:

- **Download** the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- **Explore the Molecular Signatures Database (MSigDB)**, a collection of annotated gene sets for use with GSEA software.
- **View documentation** describing GSEA and MSigDB.

What's New

16-Jul-2018: MSigDB 6.2 released. This is a minor release that includes updates to gene set annotations, corrections to miscellaneous errors, and a handful of new gene sets. See the release notes for more information.

19-Oct-2017: MSigDB 6.1 released. See release notes for more information, including important corrections to gene sets in the C3 collection.

11-Aug-2017: Four new CHIP files are now available for use with data

License Terms

GSEA and MSigDB are available for use under these license terms.

Please register to download the GSEA software, access our web tools, and view the MSigDB gene sets. After registering, you can log in at any time using your email address. Registration is free. Its only purpose is to help us track usage for reports to our funding agencies.
Usage and Growth of GSEA and MSigDB

Top Level MSigDB 6.2 Downloads (Distinct Users, 2018)

- C5, 2892, 17%
- C7, 2457, 14%
- C6, 2042, 12%
- C3, 1089, 6%
- C1, 1066, 6%
- Hallmarks, 4134, 24%
- C2, 2735, 16%
- C4, 832, 5%

Global Userbase of the GSEA Ecosystem

14,720 Total Citations
167,000 Total Registered Users

Annual New Citations

Annual New User Registrations

UC San Diego
School of Medicine
GSEA; 14 years, still the gold standard

Pathway Analysis vs Gene Set Analysis: What is the Difference and When Should I Use Each?

Advaita Bioinformatics Blog - December 2018

“In the gene set analysis category, the method that came on top in our extensive testing is the Gene Set Enrichment Analysis (GSEA). As mentioned above, we ran many datasets and GSEA is unbelievably unbiased and it gives you the best results that you can get if you are willing to ignore all pathway topology and the phenomena described by the pathway.”


Other Gene Set Analysis Methods also use MSigDB Gene Sets

• GAGE
• Significance Analysis of Function and Expression (SAFE)
• sigPathway
• Correlation Adjusted Mean RAnk (CAMERA)
ITCR Sustained Support U24

Supporting and evolving Gene Set Enrichment Analysis and the Molecular Signatures Database for cancer research

PIs: Jill P. Mesirov and Pablo Tamayo, UC San Diego

Aim 1: Develop and deploy the next generation of the GSEA method and software to keep pace with the needs of the cancer research community.

Aim 2: Extend the scope and specificity of the MSigDB, and evolve the underlying technology.

Aim 3: Provide training and outreach activities for the cancer research community, and maintain and support the GSEA software and MSigDB.
Supporting and evolving Gene Set Enrichment Analysis and the Molecular Signatures Database for cancer research

PIs: Jill P. Mesirov and Pablo Tamayo, UC San Diego

Aim 1: Develop and deploy the next generation of the GSEA method and software to keep pace with the needs of the cancer research community.

Aim 2: Extend the scope and specificity of the MSigDB, and evolve the underlying technology.

Aim 3: Provide training and outreach activities for the cancer research community, and maintain and support the GSEA software and MSigDB.
**MSigDB Collections**

**C1**: genomic regions

- chr7p21
- 60 genes

**C2**: canonical pathways and publications

- C1: genomic regions
- chr7p21
- 60 genes
- C2: canonical pathways and publications
- H: Hallmarks (consensus signatures)

**C3**: computational sequence motif targets

- TF
- CDS
- microRNA

**C4**: cancer modules and neighborhoods

- C3: computational sequence motif targets
- TF
- CDS
- microRNA

**C5**: gene ontology

- C4: cancer modules and neighborhoods

**C6**: oncogenic pathway activation

- C5: gene ontology

**C7**: immunological states

- C6: oncogenic pathway activation

signatures derived directly from expression data

**Highlights**

- ImmuneSigDB: Collection of 5,000 gene sets derived from 400 immunological studies
- Includes a wide range of mouse and human immune cell states and perturbations
- Designed for use with GSEA and an approach called Leading Edge Metagene analysis
- ImmuneSigDB identified shared and unique biology in human and mouse sepsis response

**Authors**

Jernej Godec, Yan Tan, Arthur Liberzon, ..., Atul J. Butte, Jill P. Mesirov, W. Nicholas Haining

**Correspondence**

nicholas_haining@dfci.harvard.edu

**In Brief**

Meaningful interpretation of gene-expression analyses relies on identifying changes in expression of sets of genes corresponding to specific biological processes and cell states. Haining and colleagues generated a collection of 5,000 annotated, immunology-specific gene sets and uncovered shared and species-specific biology in mouse and human transcriptional responses to sepsis.

Godec et al., 2016, Immunity 44, 1–13

January 19, 2016

© 2016 Elsevier Inc.

http://dx.doi.org/10.1016/j.immuni.2015.12.006
MSigDB Online Tools

Investigate Gene Sets

Gain further insight into the biology behind a gene set by using the following tools:

- compute overlaps with other gene sets in MSigDB
- display the gene set expression profile based on a selected condition of expression data
- categorize members of the gene set by gene families

Compute Overlap Statistics

- Collectors
  - # Genes in Overlap
  - # Genes in Complement
  - # Genes in Universe

- Compute Overlaps for Selected Genes

- Gene Set Names
  - Description
  - # Genes in Overlap
  - Gene Symbol

Project on to Compendia Expression Profiles

- Gene Symbol
- Description
- geneset overlap matrix
- Gene Set Name
- Gene Symbol
From 6.2 to 7.0 – A major tune-up for a key resource
Migration to 7.0

v5.2
- Last major MSigDB content update
- Total replacement of Gene Ontology collections

v6.X
- New Creative Commons Attribution license agreement
- Updates with new curated sets to C2 and various metadata corrections.

v7+ updates
- Major roll-up update for MSigDB Gene Symbol mappings
- Depreciation of unmaintainable (eg. UniGene based) gene sets
- Merge upstream updates for external collections
**Gene Symbols and You**

- MSigDB builds from a merge of annotations from multiple sources
- MSigDB7 unified with ENSEMBL 96 – April 2019
  - Future MSigDB releases will roll-up to current ENSEMBL Transcriptome annotations
  - Continual automatic gene set refinement as ENSEMBL releases improve transcript annotations and probe mappings
- New process for ENSEMBL-based gene orthology conversion

![HGNC Symbol Changes Per Year](chart)
C2: Reactome

- Current collection based on Reactome Release 44
  - Redundancy filtered between Reactome, KEGG (non-public updates) and BioCarta (discontinued) was performed
- MSigDB 7.0 Beta built on Reactome Release 68 (March 2019)
  - ~1500 more pathways represented
    - ~4700 more genes covered
  - Update to Reactome stable identifiers
    - Allows deep linking to the Reactome pathway graph from the MSigDB website
C5: Gene Ontology Collection

GO in MSigDB 6.2 – May 2016 GO Annotations

- 5917 Gene Sets
  - 4436 Biological Process 901 Molecular Function
  - 580 Cellular Component

2019 Update:

- ~2400 new gene sets (8368 total, ~40% of all GO)
  - 6385 Biological Process 1202 Molecular Function
  - 794 Cellular Component
- ~2000 more genes with ontology annotations
C1: Positional Sets

- Introduced in MSigDB 2.0
  - Based on cytogenetic annotations from HUGO/HGNC and UniGene
- Rebase on GRCh38.p12
  - ENSEMBL 96 gene annotations (April 2019)

- Beyond 7.0: Supplement C1 with a new subcollection of genes within Topologically Associated Domains (TADs)
  - Initial set of ~3000 TADs from Human ES cells (Ren Lab, UCSD)
7.0+ – New resources for new science
New Resources

- C1: New Topologically Associated Domains subcollection
- Overhaul of Collection 3
  - Replacement of miRNA targets
  - Replacement of transcription factor target datasets
- New C2 Subcollection
  - Curation of protein target sets from the Small Molecule Pathway Database (C2:CP:SMPD)
- Preliminary collection of single-cell sequencing derived cell identity signatures
C3:miRNA 2.0 - in collaboration with miRDB (mirdb.org)

Current Release Based on miRBase 7.1
- 221 sets of miRNAs targeting 7,631 unique genes

MSigDB7: miRDB 6.0, Jan 2019 – miRBase v22
- ~2,000 miRNAs with 16,731 total “high confidence” gene-level target predictions
- 120% increase in genomic coverage
- Target prediction model experimentally validated with CLIP, RNA-seq, and miRNA knockdown studies


miRNA regulatory networks are deeply intertwined with disease pathogenesis and remain a challenge.

C3: Transcript Factor Targets (TFT) 2.0

In collaboration with the Gene Transcription Regulation Database

- Current C3:TFT:
  - 615 Gene Sets, computationally derived from cross species conserved motif prediction
  - Annotated using TransFac data – new annotations no longer public

- C3:TFT 2.0
  - Experimentally derived data curated from the Gene Transcription Regulation Database (GTRD) – Updated January 2019

- Constructed from 7,239 Experiments
  - Consensus peaks by MACS, SISSR, GEM and PICS
  - Targets for 852 Human transcription factors

Finding Druggable Pathways with C2:CP:SMPD

• ~400 MSigDB compatible pathways

• Encompasses Drug Action, Disease, Metabolic, and Signal Transduction Pathways

• All pathways annotated for both protein and small molecule participants
  • Link-out to interactive, detailed, visual pathway diagrams for each drug and small molecule
Integration of MSigDB with NDEx
Summary

With MSigDB7+:

- Gene annotations now pegged to ENSEMBL release versions.
- Major roll-up of externally sourced gene sets
  - C1 from GRCh38p13, C2:CP:Reactome, C5:GO updates
- New collaborations with miRDB, GTRD, SMPDB
- New TAD sub-collection to support genomic spatial regulation studies
- Integration with NDEx allows MSigDB to re-capture some lost pathway topology

(Some) Future Plans

- Collaboration with JAX to construct PDX gene expression signatures
- Transition away from XML files to a new relational database architecture
- Overhaul of the MSigDB website to support new resource integrations
Acknowledgements

Co-PIs
• Jill P. Mesirov
• Pablo Tamayo

Mesirov Lab
• Helga Thorvaldsdottir
• David Eby
• Barbara Hill
• Michael Reich

Tamayo Lab
• Kwat Yeerna
• Xiaojun (Max) Xu

Collaborators

Trey Ideker Lab, UCSD
Dexter Pratt

Carol Bult
Jill Recla