Published in final edited form as:
*Nat Methods*. 2023 November ; 20(11): 1619–1620. doi:10.1038/s41592-023-02014-7.

# Extending support for mouse data in the Molecular Signatures Database (MSigDB)

**Anthony S. Castanza**[1], **Jill M. Recla**[2,5], **David Eby**[1], **Helga Thorvaldsdóttir**[3], **Carol J. Bult**[2], **Jill P. Mesirov**[1,4,✉]

[1]Department of Medicine, University of California San Diego, La Jolla, CA, USA.

[2]The Jackson Laboratory for Mammalian Genomics, Bar Harbor, ME, USA.

[3]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[4]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA.

[5]Deceased: Jill M. Recla.

The rise of full transcriptome acquisition technologies has fueled the rapid proliferation of molecular-level biological data. These large datasets require interpretation beyond the single-gene level to connect them to meaningful biology and clinical impacts. In 2003, we pioneered the gene set enrichment analysis (GSEA) approach[1] to enable the identification of the coordinate activation or repression of sets of genes that share common biology, thereby distinguishing even subtle differences between phenotypes or cellular states. We first released our GSEA software and a companion resource of gene sets, the Molecular Signatures Database (MSigDB), in 2005 (ref. 2).

Historically, the GSEA/MSigDB resource (https://gsea-msigdb.org) focused on the analysis of human-specific datasets. The gene sets in MSigDB were offered exclusively in the human gene space, and analysis of mouse data was minimally supported through basic ortholog mapping to human genes. In recognition of the importance of model organisms, particularly mice, for research into the mechanisms of human disease, we recently expanded the MSigDB to address this by introducing a database of mouse-native gene sets and by substantially improving MSigDB's ortholog mapping.

✉ jmesirov@health.ucsd.edu .

Author contributions
A.S.C. and J.M.R. developed the ortholog mapping procedure. C.J.B. and J.M.R. identified and curated the gene sets in the MP:Tumor collection; J.M.R. wrote the code to generate the gene sets. J.M.R. and C.J.B. produced the ortholog translated hallmarks collection. D.E. collected existing mouse gene sets from MSigDB. A.S.C. curated all other sets and collections and produced the underlying gene mapping data. A.S.C. and D.E. produced the complete MSigDB build. A.S.C., D.E. and H.T. designed the updates to the MSigDB website; D.E. implemented the updates. A.S.C., H.T. and J.P.M. wrote the manuscript with input and feedback from all authors. C.J.B. and J.P.M. provided oversight and guidance.

Code availability
The GSEA software is open source under a BSD-style license; the source code is available at http://github.com/GSEA-MSigDB. MSigDB is available under a Creative Commons license. The full license terms can be found at http://gsea-msigdb.org.

Competing interests
The authors declare no competing interests.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-02014-7.

## Mouse-native gene sets

While ortholog mapping is useful, some degree of uncertainty exists when discriminating between orthologs and paralogs, and even orthologous genes in human and mouse can have different functions. The inclusion of all genes — that is, not excluding species-specific genes — may be key to gaining mechanistic insights and understanding species-specific biology. Therefore, we invested substantial effort to support analysis of mouse data through the release of Mouse MSigDB, which consists of gene sets curated directly from mouse-centric databases and datasets and specified in native mouse gene identifiers. These sets can be used without the need for ortholog mapping. The new Mouse MSigDB leverages the organizational paradigm of Human MSigDB and includes the following collections of sets (see Supplementary Note 1):

- M1 – A ‘positional’ collection of the genes in each cytogenetic band
- M2 – Molecular pathways mined from public mouse databases, and a subcollection of mouse signatures of chemical and genetic perturbations curated from peer-reviewed literature
- M3 – Regulatory targets, including micro-RNA (miRNA) targets and transcription factor targets for mouse
- M5 – Gene sets derived from ontology hierarchies
- M8 – Curated cell type markers from single-cell data
- MH – An ortholog-converted version of the popular MSigDB Hallmarks collection[3]; the only collection not derived from mouse native data

Although Mouse MSigDB is a new resource, it benefits from the maturity of the existing Human MSigDB and accompanying website (Fig. 1). The gene sets were prepared following the procedure that we have honed over the years for Human MSigDB (see Supplementary Note 2).

## Improved ortholog mapping

The GSEA method is species agnostic, but the software requires that the identifiers of the genes in the analyzed data match those of the querying gene sets. However, if there are mismatches, MSigDB provides a critical set of gene-identifier mapping-table files that GSEA uses to convert the identifiers in a dataset to the gene symbols namespace of MSigDB. The files can map between different name spaces and versions of human genes, as well as provide ortholog mapping between mouse genes and human genes, and vice versa. We made substantial improvements to our procedure for creating MSigDB’s ortholog mapping tables, leveraging premiere ortholog databases and considering the particular requirements of GSEA. We adopted a systematic ‘best match ortholog’ approach, using annotations from the Alliance of Genome Resources[4], which integrates annotations from multiple independent[5] methods via the *Drosophila* RNAi Screening Center’s Integrative Ortholog Prediction Tool, and supplementing with ortholog data from Ensembl[6]. The resulting files improve the analysis of mouse datasets when used with Human MSigDB, but also allow the analysis of human datasets with the mouse gene set database. (See

Supplementary Notes 3 and 4 for details of the new ortholog mapping method and an assessment of its performance.)

The mouse collections will be updated on the same frequent release schedule as Human MSigDB (see Supplementary Note 2). Our first priority is to add other mouse collections modeled on the Human MSigDB — for example, oncogenic and immunologic sets. For other future work, the infrastructural changes we have invested in developing and releasing the Mouse MSigDB open the possibility of supporting the curation of gene sets from other model organisms (for example, rat) and species of interest.

Few comprehensive resources exist for mouse native gene sets that can be used for GSEA (see Supplementary Note 5). By offering both support for native analysis of mouse data and support for mapping mouse data to use with human gene sets and human data to use with mouse-native sets, the new MSigDB enhancements represent a significant step forward in the potential translational impact of GSEA results and put the enrichment analysis of mouse and human data on equal footing.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

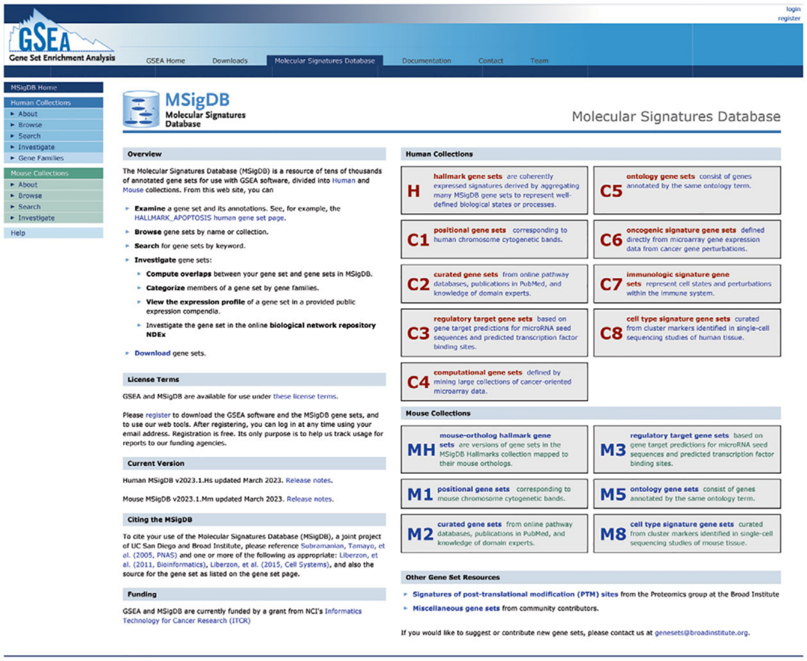
This work was supported by US National Institutes of Health grant U24CA220341 to J.P.M.

## Data availability

The MSigDB resource, including both mouse and human gene sets, is freely available at http://msigdb.org. The full GSEA/MSigDB resource is available at http://gsea-msigdb.org.

## References

1. Mootha VK et al. Nat. Genet 34, 267–273 (2003). [PubMed: 12808457]
2. Subramanian A et al. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005). [PubMed: 16199517]
3. Liberzon A et al. Cell Syst. 1, 417–425 (2015). [PubMed: 26771021]
4. Alliance of Genome Resources Consortium. Genetics 213, 1189–1196 (2019). [PubMed: 31796553]
5. Hu Y et al. BMC Bioinformatics 12, 357 (2011). [PubMed: 21880147]
6. Herrero J et al. Database (Oxford) 2016, baw053 (2016).



**Overview**

The Molecular Signatures Database (MSigDB) is a resource of tens of thousands of annotated gene sets for use with GSEA software, divided into Human and Mouse collections. From this web site, you can

- **Examine** a gene set and its annotations. See, for example, the HALLMARK_APOPTOSIS human gene set page.
- **Browse** gene sets by name or collection.
- **Search** for gene sets by keyword.
- **Investigate** gene sets:
  - **Compute overlaps** between your gene set and gene sets in MSigDB.
  - **Categorize** members of a gene set by gene families.
  - **View the expression profile** of a gene set in a provided public expression compendia.
  - Investigate the gene set in the online **biological network repository NDEx**
- **Download** gene sets.

**License Terms**

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

Please register to download the GSEA software and the MSigDB gene sets, and to use our web tools. 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.

**Current Version**

Human MSigDB v2023.1.Hs updated March 2023. Release notes.

Mouse MSigDB v2023.1.Mm updated March 2023. Release notes.

**Citing the MSigDB**

To cite your use of the Molecular Signatures Database (MSigDB), a joint project of UC San Diego and Broad Institute, please reference Subramanian, Tamayo, et al. (2005, PNAS) and one or more of the following as appropriate: Liberzon, et al. (2011, Bioinformatics), Liberzon, et al. (2015, Cell Systems), and also the source for the gene set as listed on the gene set page.

**Funding**

GSEA and MSigDB are currently funded by a grant from NCI's Informatics Technology for Cancer Research (ITCR)

**Human Collections**

**H** **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

**C5** **ontology gene sets** consist of genes annotated by the same ontology term.

**C1** **positional gene sets** corresponding to human chromosome cytogenetic bands.

**C6** **oncogenic signature gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

**C2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

**C7** **immunologic signature gene sets** represent cell states and perturbations within the immune system.

**C3** **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

**C8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of human tissue.

**C4** **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

**Mouse Collections**

**MH** **mouse-ortholog hallmark gene sets** are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

**M3** **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

**M1** **positional gene sets** corresponding to mouse chromosome cytogenetic bands.

**M5** **ontology gene sets** consist of genes annotated by the same ontology term.

**M2** **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

**M8** **cell type signature gene sets** curated from cluster markers identified in single-cell sequencing studies of mouse tissue.

**Other Gene Set Resources**

- **Signatures of post-translational modification (PTM) sites** from the Proteomics group at the Broad Institute
- **Miscellaneous gene sets** from community contributors.

If you would like to suggest or contribute new gene sets, please contact us at genesets@broadinstitute.org.

**Fig. 1 |. Home page of the MSigDB resource.**

The updated MSigDB home page at https://msigdb.org provides access to both the mouse gene set collections and the human collections. Separate color schemes are used for the two different collections throughout the website to allow the user to easily distinguish between the two. A green-and-blue color scheme indicates information about the mouse collections, blue-and-red about the human collections.