Gene Expression Analysis – Gene Set Analysis Approach CANB 7640

Aik Choon Tan, Ph.D.
Associate Professor of Bioinformatics
Division of Medical Oncology
Department of Medicine
aikchoon.tan@ucdenver.edu

9/18/2018

http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640/

Outline

- Methods for Identifying Gene Set Analysis
  - Motivation
  - Statistics
  - Gene Set Enrichment Analysis (GSEA)

Limitations of Candidate Gene Analysis

- Functional genomics technologies such as expression profiling using microarrays provide a global approach to understanding cellular. processes in different biological phenotypes.
- Candidate genes analyses
  - Gene lists
  - Number of genes range from hundred to thousands
  - Sifting through gene list is a daunting task to group these genes into functional groups (*ad hoc* analysis)
  - Bias and require expert knowledge

Motivation

- Genes must act in concert to drive various cellular processes.
- Gene expression alterations might be revealed at the level of biological pathways or co-regulated gene sets (functional groups).
- Gene set analysis –
  - more objective and robust.
  - able to discover sets of coordinated differentially expressed genes among pathway members and their association to a specific biological phenotype.
  - provide new insights linking biological phenotypes to their underlying molecular mechanisms.
  - suggesting new hypotheses about pathway membership and connectivity.

How to find Sets of Co-ordinately Differentially Expressed Genes

- High-throughput “omics” data (e.g. microarray gene expression, RNA-seq etc) full with noise.
- Real biological signals might be subtle.
- (*Assumption*) Genes with similar function or participate in a biological process have similar expression patterns.

- *Goal*: Find these sets of genes from high-throughput “omics” data

Gene Set Analysis

- Samples with known biological phenotypes (e.g. *class labels*)
- High-throughput measurements of data points (e.g. *gene expressions*)
- Set of genes involved in biological processes or cellular functions or pathways (e.g. *gene sets*)
- Compare the *gene expressions* of various *class labels* to find differentially expressed *gene sets*.

Gene Set Enrichment Analysis (GSEA)

- *Goal*: to detect modest but coordinated expression changes in pre-specified sets of related genes (gene sets).
- Gene set can be all the genes involved in
  - specific pathway (obtained from Pathway databases such as KEGG, BioCARTA, REACTOME etc)
  - specific gene ontology class (obtained from Gene Ontology category)
  - specific chromosome locations
  - specific transcriptional regulated targets (e.g. transcription factor targets, miRNA targets)
  - specific gene signatures (obtained from published papers or your own experiments)
  - Specific drug targets (obtained from experiments of drug-gene interactions)

Original Publication of GSEA

PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes

Vamsi K Mootha[1,2,3,10], Cecilia M Lindgren[1,4,10], Karl-Fredrik Eriksson[4], Aravind Subramanian[1], Smita Sihag[1], Joseph Lehar[1], Pere Puigserver[5], Emma Carlsson[4], Martin Ridderstråle[4], Esa Laurila[4], Nicholas Houstis[1], Mark J Daly[1], Nick Patterson[1], Jill P Mesirov[1], Todd R Golub[1,5], Pablo Tamayo[1], Bruce Spiegelman[5], Eric S Lander[1,6], Joel N Hirschhorn[1,7,8], David Altshuler[1,2,7,9,11] & Leif C Groop[4,11]

DNA microarrays can be used to identify gene expression changes characteristic of human disease. This is challenging, however, when relevant differences are subtle at the level of individual genes. We introduce an analytical strategy, Gene Set Enrichment Analysis, designed to detect modest but coordinate changes in the expression of groups of functionally related genes. Using this approach, we identify a set of genes involved in oxidative phosphorylation whose expression is coordinately decreased in human diabetic muscle. Expression of these genes is high at sites of insulin-mediated glucose disposal, activated by PGC-1α and correlated with total-body aerobic capacity. Our results associate this gene set with clinically important variation in human metabolism and illustrate the value of pathway relationships in the analysis of genomic profiling experiments.

NATURE GENETICS VOLUME 34 | NUMBER 3 | JULY 2003 267-273
[Citations: >4700]

Original Publication of GSEA

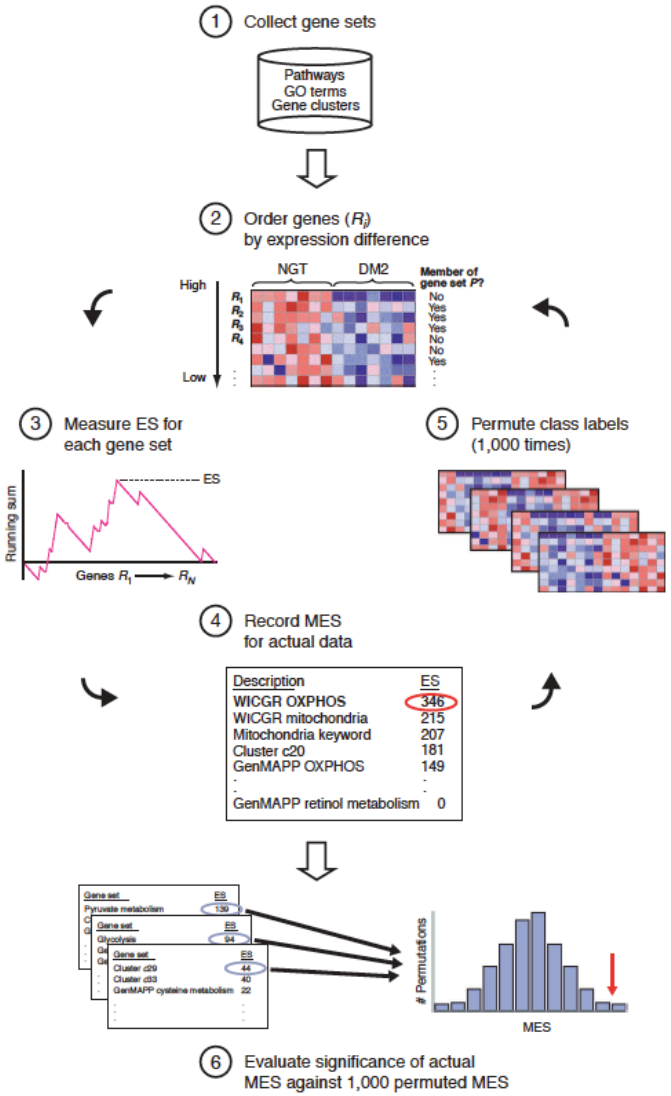


Figure 1 Schematic overview of GSEA. The goal of GSEA is to determine whether any *a priori* defined gene sets (step 1) are enriched at the top of a list of genes ordered on the basis of expression difference between two classes (for example, highly expressed in individuals with NGT versus those with DM2). Genes $R_1,\ldots R_N$ are ordered on the basis of expression difference (step 2) using an appropriate difference measure (for example, SNR). To determine whether the members of a gene set S are enriched at the top of this list (step 3), a Kolmogorov-Smirnov (K-S) running sum statistic is computed: beginning with the top-ranking gene, the running sum increases when a gene annotated to be a member of gene set S is encountered and decreases otherwise. The ES for a single gene set is defined as the greatest positive deviation of the running sum across all N genes. When many members of S appear at the top of the list, ES is high. The ES is computed for every gene set using actual data, and the MES achieved is recorded (step 4). To determine whether one or more of the gene sets are enriched in one diagnostic class relative to the other (step 5), the entire procedure (steps 2–4) is repeated 1,000 times, using permuted diagnostic assignments and building a histogram of the maximum ES achieved by any pathway in a given permutation. The MES achieved using the actual data is then compared to this histogram (step 6, red arrow), providing us with a global P value for assessing whether any gene set is associated with the diagnostic categorization.

Original Publication of GSEA

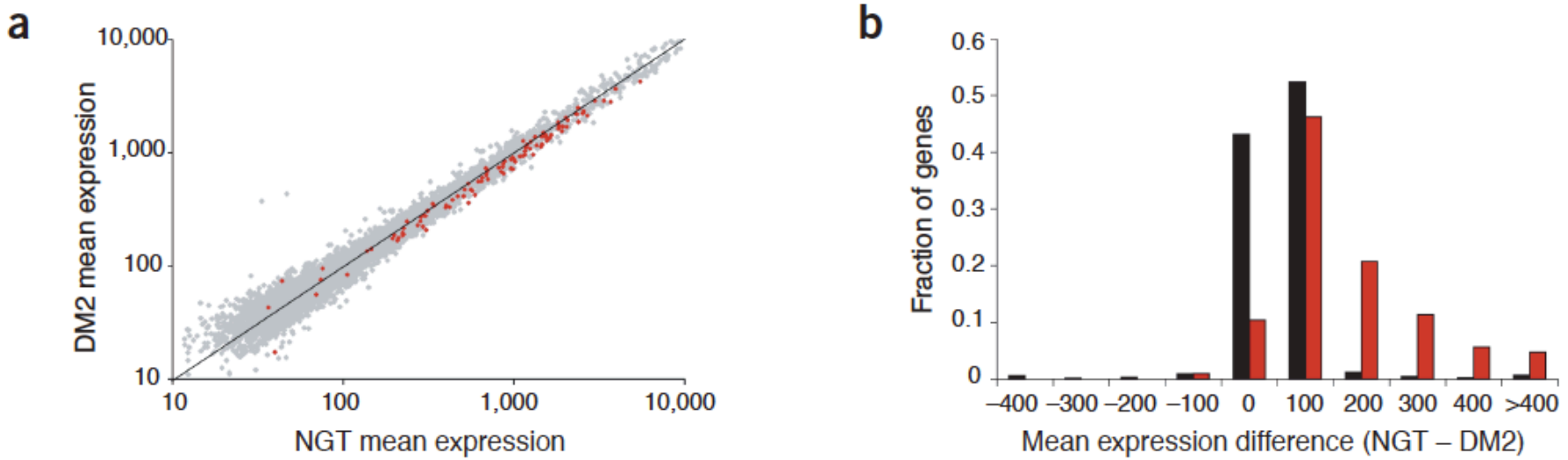


Figure 2 OXPHOS gene expression is reduced in diabetic muscle. (**a**) The mean expression of all genes (gray) and of OXPHOS genes (red) is plotted for individuals with DM2 versus those with NGT. (**b**) Histogram of mean gene expression level differences between individuals with NGT and DM2, using the data from **a**, for all genes (black) and for OXPHOS genes (red).

GSEA paper (PNAS 2005)

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

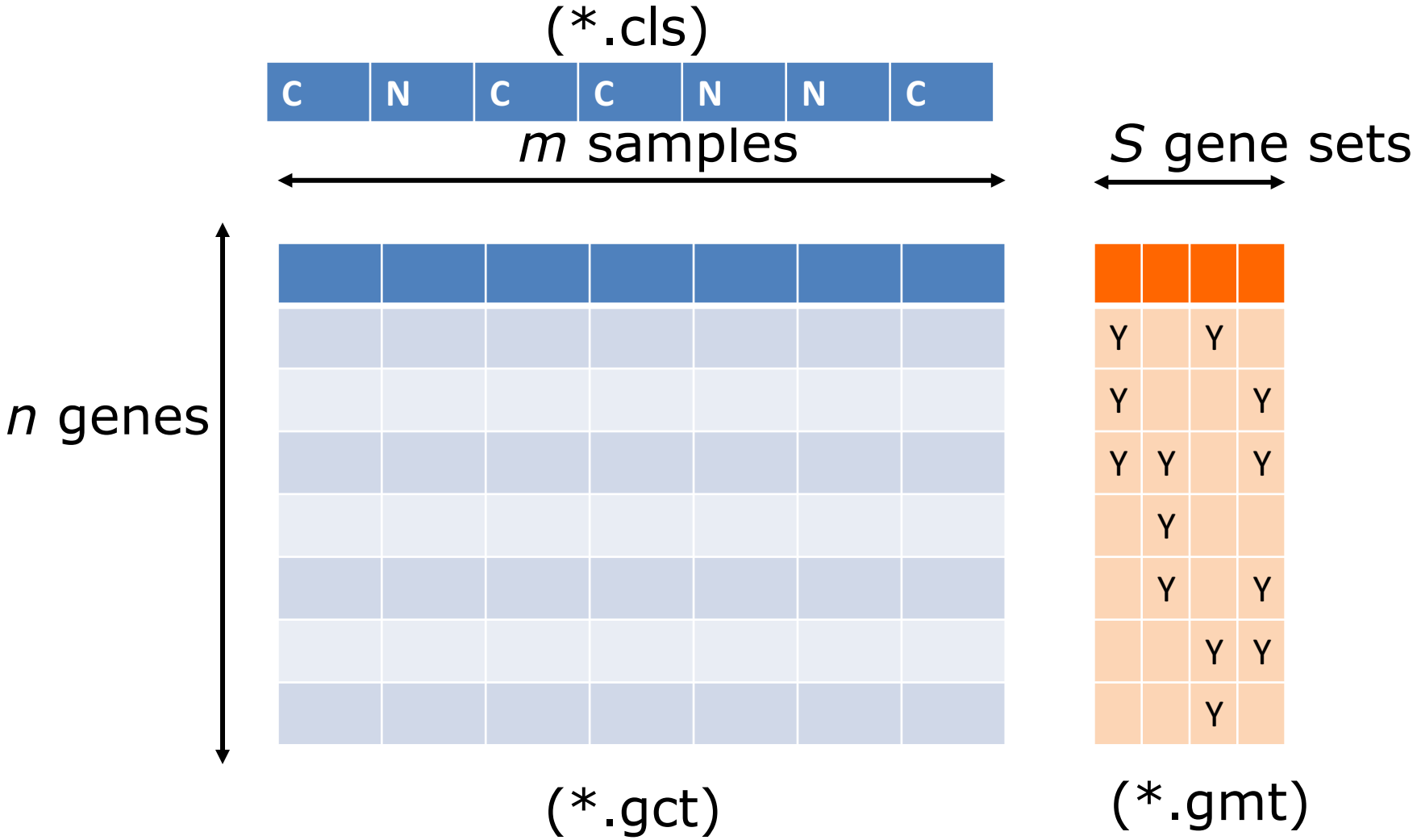
PNAS October 25, 2005 vol. 102 no. 43 15545–15550
[Citations: >14000]

GSEA Algorithm: Step 1

- Collect gene sets (*.gmt) from databases
- Compile gene expression data (*.gct)
- Define class labels for your samples (*.cls)

- Microarray chip definition file (*.chip) [*Not applicable to RNA-seq*]

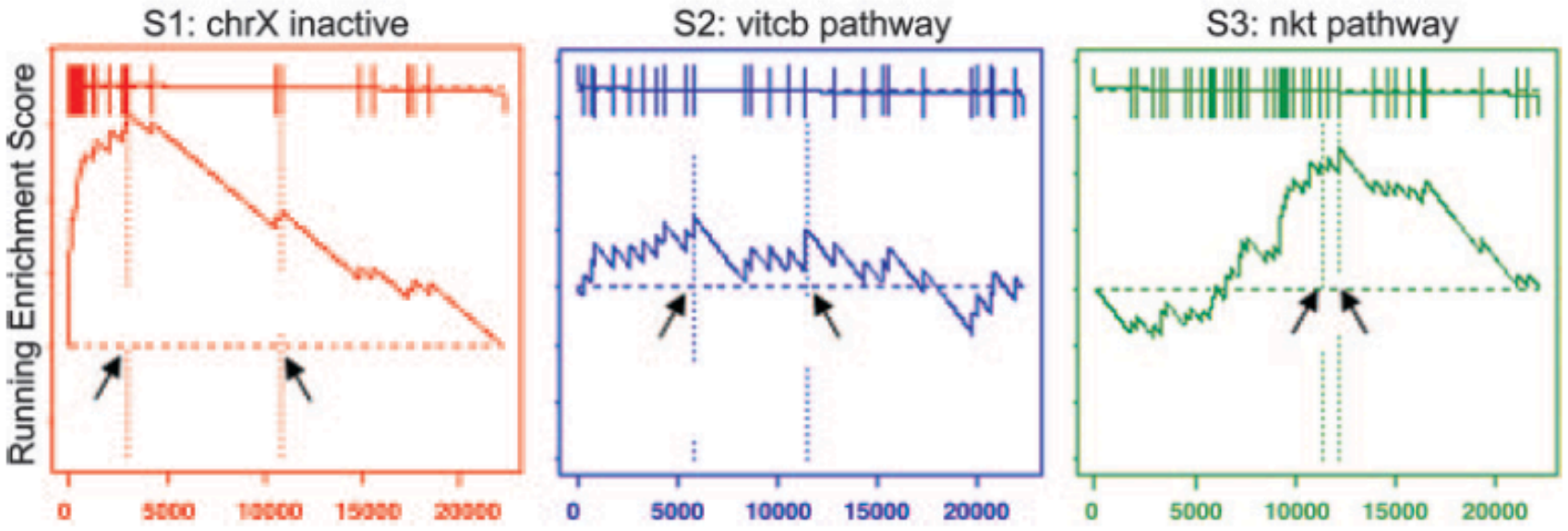
Required Files for GSEA



GSEA Algorithm: Step 2

- Rank genes based on their expression differences between the two phenotypes (in GSEA, the measurement is Signal-to-noise ratio)
- Compute Enrichment Score (ES)
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise.
    - Magnitude of increment depends on correlation of gene with phenotype.
  - Maximum deviation from zero = enrichment score

GSEA Algorithm: Step 2 Enrichment Plot



Results on SETPATHWAY (Enrichment Plot)

Max Score = ES

Table: GSEA Results S

Dataset	Leukemia_collapsed
Phenotype	Leukemia.cls#ALL_v
Upregulated in class	ALL
GeneSet	SETPATHWAY
Enrichment Score (ES)	0.7762963
Normalized Enrichment Score (NES)	1.8301569
Nominal p-value	0.0
FDR q-value	0.7549038
FWER p-Value	0.348

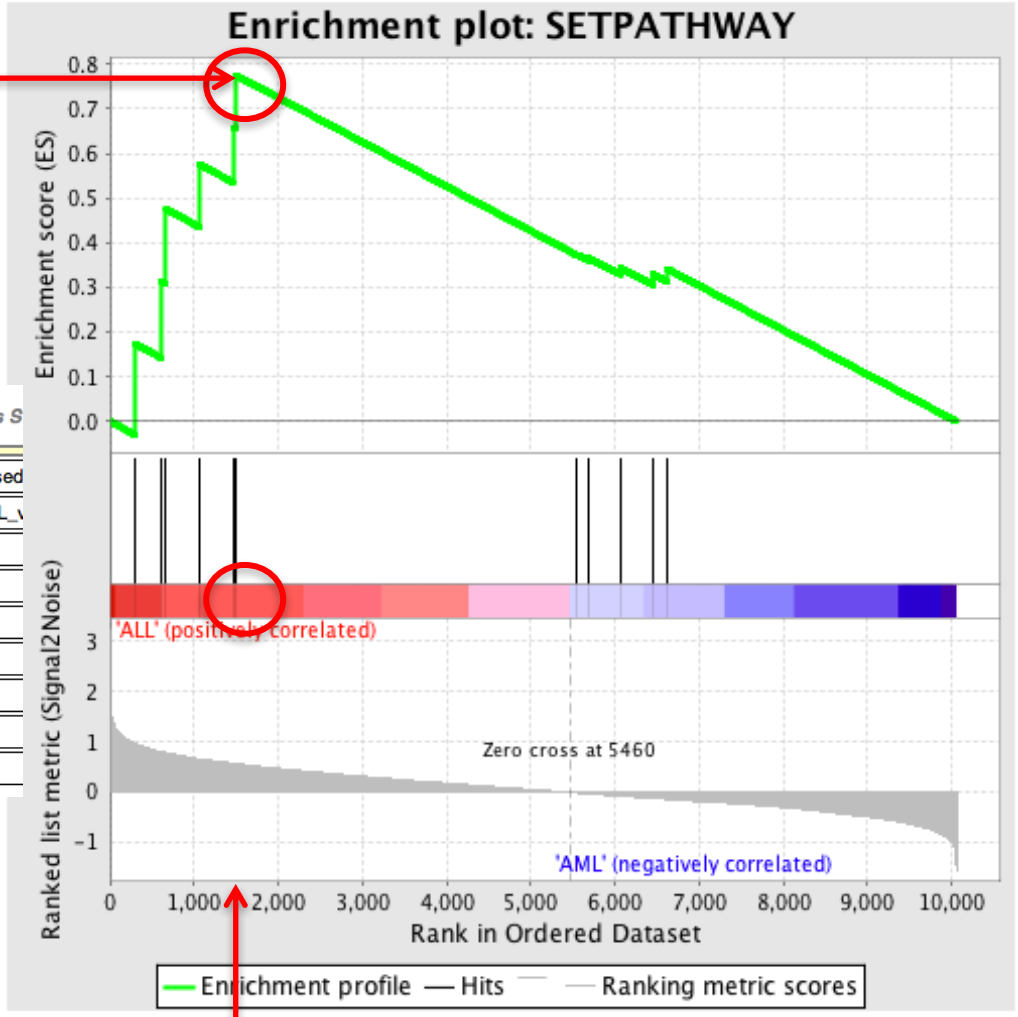


Fig 1: Enrichment plot: SETPATHWAY

Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

GSEA Algorithm: Step 3

- Compute the Significance of the Gene Sets by Permutation Test
- Permutation Test (*n* times)
  - Permute phenotype labels
  - Permute gene sets
- For each permutation, compute ES score
- Compare ES score for actual data to distribution of ES scores from permuted data

GSEA Algorithm: Step 4

- Adjustment for multiple hypothesis testing:
  - Normalize the ES accounting for size of each gene set, yielding normalized enrichment score (NES)
  - Control proportion of false positives by calculating FDR corresponding to each NES, by comparing tails of the observed and null distributions for the NES.

GSEA Leading Edge Analysis

- Genes might be involved in different pathways/gene sets.
- Selected core genes might be identified in top enriched gene sets.

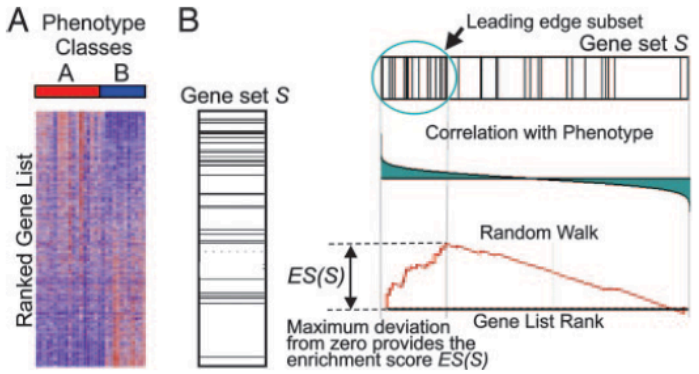
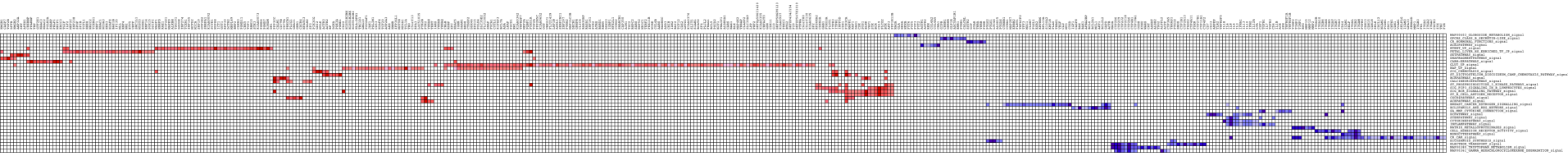


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.



GSEA Homepage

http://www.broadinstitute.org/gsea/index.jsp

GSEA
Gene Set Enrichment Analysis

GSEA Home | Downloads | Molecular Signatures Database | Documentation | Contact

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

19-Sep-2016: The Documentation section of our website is temporarily offline. We are working to resolve the issue and get it back as soon as possible.

15-Aug-2016: The first beta of the next major GSEA Desktop release is available, with SVG plots, Cytoscape 3.3+ support and much more.

17-Jun-2016: You can now follow @GSEA_MSigDB on Twitter!

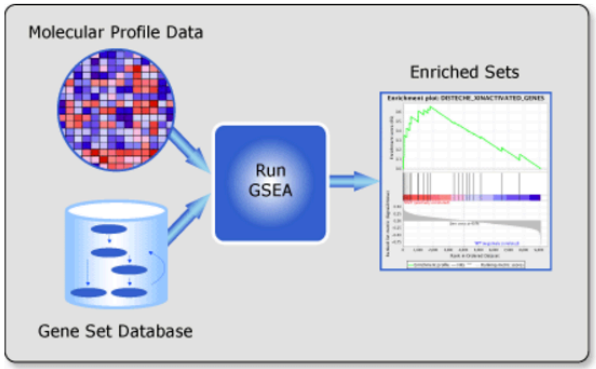
29-Feb-2016: The Sunday 28-Feb-2016 maintenance is complete on the GSEA/MSigDB website. Thanks for your patience!

13-Jan-2016: Version 5.1 of the Molecular Signatures Database (MSigDB) is now available. It includes the addition of 2,962 gene sets to the C7 collection of immunologic signatures, as well as a number of updates and corrections. See the Release Notes for details.

23-Dec-2015: Our paper describing the generation of the Hallmarks collection and examples of its use for GSEA was published in Cell Systems.

10-Dec-2015: We have confirmed that GSEA v2.2.0 and newer are compatible with Java 8 and produce equivalent results. Its use is highly recommended.

Follow @GSEA_MSigDB



Registration

Please register to download the GSEA software 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.

Contributors

GSEA and MSigDB are maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. Our thanks to our many contributors.
Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Citing GSEA

To cite your use of the GSEA software, please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273).

GSEA Implementation

http://www.broadinstitute.org/gsea/index.jsp

Software

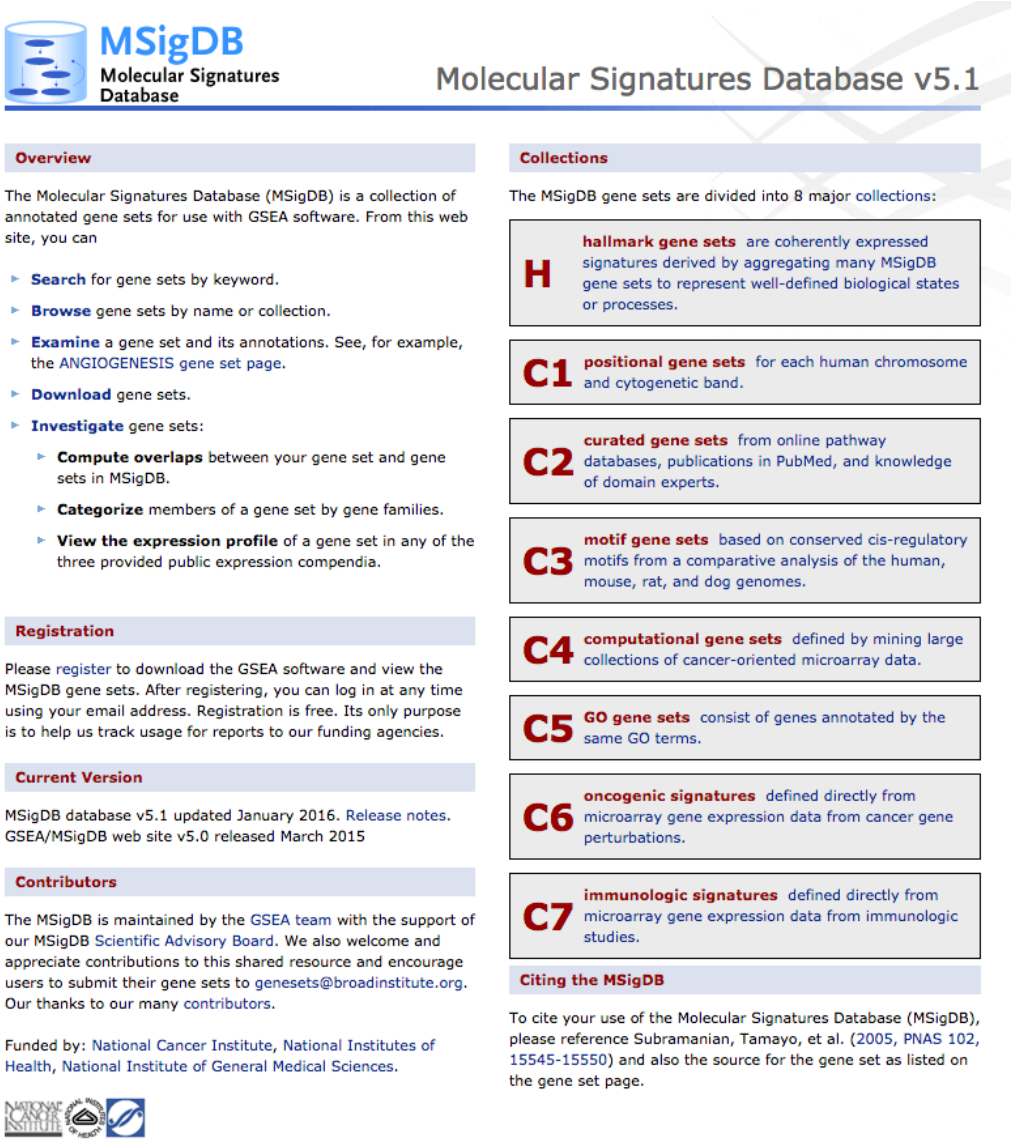
There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. Current Java implementations of GSEA require Java 7 or 8. Java 8 is recommended.

javaGSEA Desktop Application	▸ Easy-to-use graphical user interface ▸ Runs on any desktop computer (Windows, Mac OS X, Linux etc.) that supports Java 7 or 8. Java 8 is recommended. ▸ Produces richly annotated reports of enrichment results ▸ Integrated gene sets browser to view gene set annotations, search for gene sets and map gene sets between platforms	Launch with 1GB (for 32 or 64-bit Java) memory: Launch
javaGSEA Java Jar file	▸ Command line or offline usage. See our User Guide for details. ▸ Runs on any platform that supports Java 7 or 8. Java 8 is recommended. ▸ We recommend using the 'Launch' buttons above instead of this mode for most users	download gsea2-2.2.2.jar
BETA **javaGSEA Java Jar file**	▸ The first Beta version of the next major GSEA Desktop release, with SVG plots, Cytoscape 3.3+ support for the Enrichment Map, and much more. ▸ **Our tests show this Beta version produces equivalent results, but use the Production version if you have concerns. At a minimum, verification with the Production version before publication is strongly recommended.** ▸ Please contact us with bugs or other feedback. We will aim to address problems as soon as possible in future Beta releases. ▸ Runs only on the command line. See our User Guide for details. ▸ Runs on any platform that supports Java 7 or 8. Java 8 is recommended.	**BETA** download gsea2-3.0_beta_1.jar
R-GSEA R Script	▸ Usage from within the R programming environment ▸ Easily inspect, learn and tweak the algorithm ▸ Incorporate GSEA into your own data analysis pipeline ▸ Programmatically call the open source GSEA R API ▸ **Note that this script has not been updated since 2005 and may not work as-is with modern R distributions.** ▸ Click here to learn more about the R-GSEA script	download GSEA-P-R.1.0.zip
GenePattern GSEA Module	▸ Use GSEA from within GenePattern ▸ Use GSEA in concert with a large suite of other analytics found in GenePattern (a powerful and flexible analysis platform developed at the Broad Institute)	GenePattern site

For details on the GSEA algorithm and software refer to the Documentation.
For details on the latest release refer to the Release Notes.

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/



MSigDB
Molecular Signatures Database

Molecular Signatures Database v5.1

Overview

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. From this web site, you can

- **Search** for gene sets by keyword.
- **Browse** gene sets by name or collection.
- **Examine** a gene set and its annotations. See, for example, the ANGIOGENESIS gene set page.
- **Download** gene sets.
- **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 any of the three provided public expression compendia.

Registration

Please register to download the GSEA software 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.

Current Version

MSigDB database v5.1 updated January 2016. Release notes.
GSEA/MSigDB web site v5.0 released March 2015

Contributors

The MSigDB is maintained by the GSEA team with the support of our MSigDB Scientific Advisory Board. We also welcome and appreciate contributions to this shared resource and encourage users to submit their gene sets to genesets@broadinstitute.org. Our thanks to our many contributors.

Funded by: National Cancer Institute, National Institutes of Health, National Institute of General Medical Sciences.

Collections

The MSigDB gene sets are divided into 8 major collections:

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

C1 **positional gene sets** for each human chromosome and cytogenetic band.

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

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

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

C5 **GO gene sets** consist of genes annotated by the same GO terms.

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

C7 **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

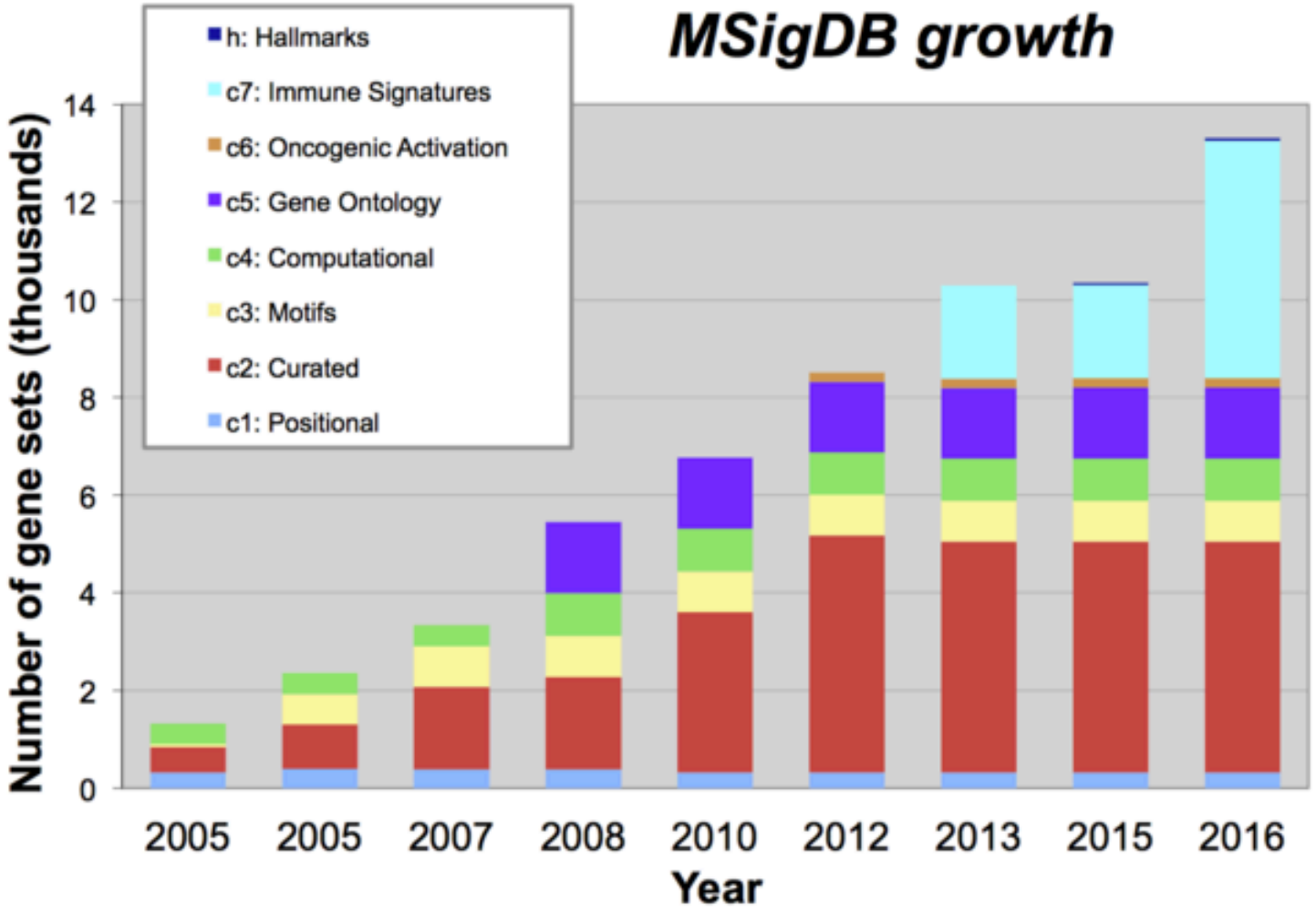
Citing the MSigDB

To cite your use of the Molecular Signatures Database (MSigDB), please reference Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) and also the source for the gene set as listed on the gene set page.

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

Currently contains 13,311 gene sets organized into 8 categories.



Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

Browse Gene Sets

Gene set name: Search

By first letter: 1 2 3 4 5 6 7 8 9 0 **A** B C D E F G H I J K L M N O P Q R S T U V W X Y Z

By collection: [about the MSigDB collections]

- **H** (hallmark gene sets, 50 gene sets)
- **C1** (positional gene sets, 326 gene sets)
  - by chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
- **C2** (curated gene sets, 4726 gene sets)
  - **CGP** (chemical and genetic perturbations, 3396 gene sets)
  - **CP** (Canonical pathways, 1330 gene sets)
  - **CP:BIOCARTA** (BioCarta gene sets, 217 gene sets)
  - **CP:KEGG** (KEGG gene sets, 186 gene sets)
  - **CP:REACTOME** (Reactome gene sets, 674 gene sets)
- **C3** (motif gene sets, 836 gene sets)
  - **MIR** (microRNA targets, 221 gene sets)
  - **TFT** (transcription factor targets, 615 gene sets)
- **C4** (computational gene sets, 858 gene sets)
  - **CGN** (cancer gene neighborhoods, 427 gene sets)
  - **CM** (cancer modules, 431 gene sets)
- **C5** (GO gene sets, 1454 gene sets)
  - **BP** (GO biological process, 825 gene sets)
  - **CC** (GO cellular component, 233 gene sets)
  - **MF** (GO molecular function, 396 gene sets)
- **C6** (oncogenic signatures, 189 gene sets)
- **C7** (immunologic signatures, 4872 gene sets)

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

Click on a gene set name to view its gene set page.

HALLMARK_ADIPOGENESIS
HALLMARK_ALLOGRAFT_REJECTION
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_ANGIOGENESIS
HALLMARK_APICAL_JUNCTION
HALLMARK_APICAL_SURFACE
HALLMARK_APOPTOSIS
HALLMARK_BILE_ACID_METABOLISM
HALLMARK_CHOLESTEROL_HOMEOSTASIS
HALLMARK_COAGULATION
HALLMARK_COMPLEMENT
HALLMARK_DNA_REPAIR
HALLMARK_E2F_TARGETS
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_FATTY_ACID_METABOLISM
HALLMARK_G2M_CHECKPOINT
HALLMARK_GLYCOLYSIS
HALLMARK_HEDGEHOG_SIGNALING
HALLMARK_HEME_METABOLISM
HALLMARK_HYPOXIA
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_INFLAMMATORY_RESPONSE
HALLMARK_INTERFERON_ALPHA_RESPONSE
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_KRAS_SIGNALING_DN
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_MITOTIC_SPINDLE
HALLMARK_MTORC1_SIGNALING
HALLMARK_MYC_TARGETS_V1
HALLMARK_MYC_TARGETS_V2
HALLMARK_MYOGENESIS
HALLMARK_NOTCH_SIGNALING
HALLMARK_OXIDATIVE_PHOSPHORYLATION
HALLMARK_P53_PATHWAY
HALLMARK_PANCREAS_BETA_CELLS
HALLMARK_PEROXISOME
HALLMARK_PI3K_AKT_MTOR_SIGNALING
HALLMARK_PROTEIN_SECRETION
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY
HALLMARK_SPERMATOGENESIS
HALLMARK_TGF_BETA_SIGNALING
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_UNFOLDED_PROTEIN_RESPONSE
HALLMARK_UV_RESPONSE_DN
HALLMARK_UV_RESPONSE_UP
HALLMARK_WNT_BETA_CATENIN_SIGNALING
HALLMARK_XENOBIOTIC_METABOLISM

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

Gene Set: HALLMARK_KRAS_SIGNALING_UP

Standard name	HALLMARK_KRAS_SIGNALING_UP
Systematic name	M5953
Brief description	Genes up-regulated by KRAS activation.
Full description or abstract	
Collection	H: hallmark gene sets
Source publication	
Exact source	
Related gene sets	(show 14 founder gene sets for this hallmark gene set)
External links	
Organism	Homo sapiens
Contributed by	Arthur Liberzon (Broad Institute)
Source platform	HUMAN_GENE_SYMBOL
Dataset references	(show 5 hallmark refinement datasets) (show 1 hallmark validation datasets)
Download gene set	format: grp | text | gmt | gmx | xml
Compute overlaps ?	(show collections to investigate for overlap with this gene set)
Compendia expression profiles ?	Human tissue compendium (Novartis) Global Cancer Map (Broad Institute) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 200 genes
Gene families ?	Categorize these 200 genes by gene family
Show members	(show 200 members mapped to 200 genes)
Version history	5.0: First introduced

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

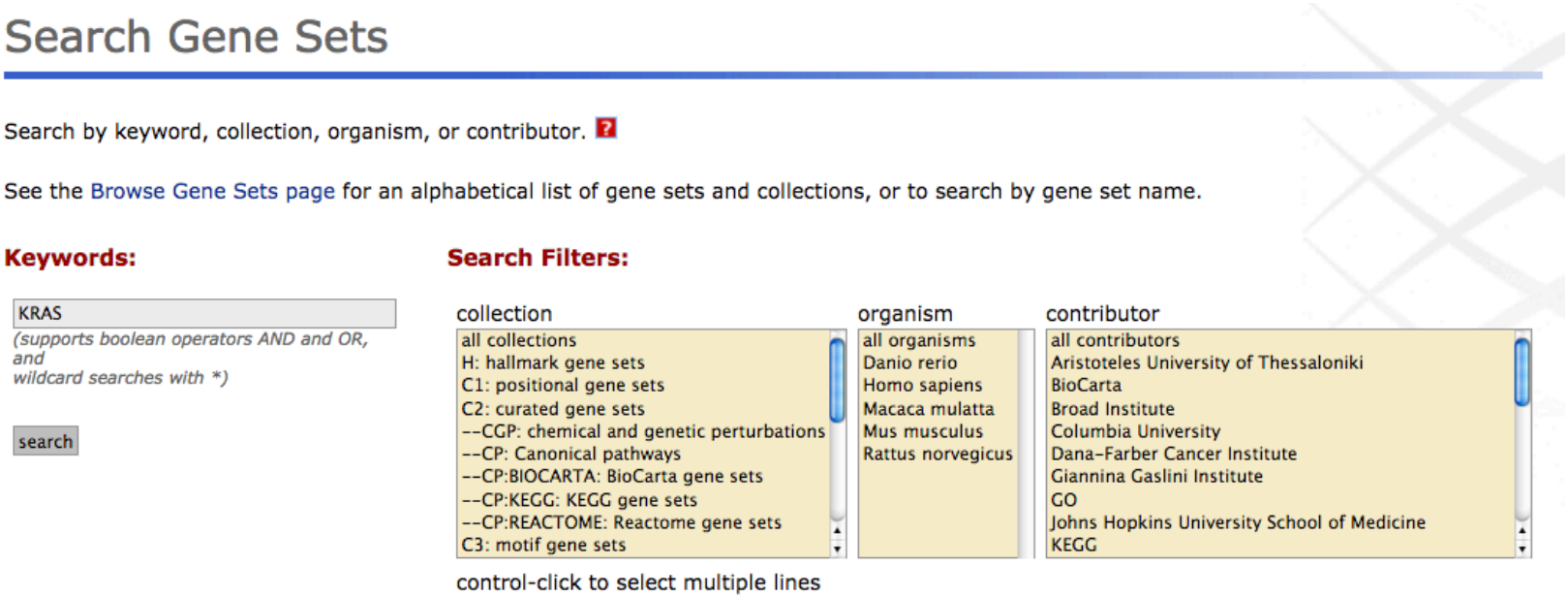
Show members

(hide 200 members mapped to 200 genes)

Original Member	Entrez Gene Id	Gene Symbol	Gene Description
ABCB1	5243	ABCB1	ATP-binding cassette, sub-family B (MD...
ACE	1636	ACE	angiotensin I converting enzyme (pepti...
ADAM17	6868	ADAM17	ADAM metallopeptidase domain 17
ADAM8	101	ADAM8	ADAM metallopeptidase domain 8
ADAMDEC1	27299	ADAMDEC1	ADAM-like, decysin 1
AKAP12	9590	AKAP12	A kinase (PRKA) anchor protein 12
AKT2	208	AKT2	v-akt murine thymoma viral oncogene ho...
ALDH1A2	8854	ALDH1A2	aldehyde dehydrogenase 1 family, membe...
ALDH1A3	220	ALDH1A3	aldehyde dehydrogenase 1 family, membe...
AMMECR1	9949	AMMECR1	Alport syndrome, mental retardation, m...
ANGPTL4	51129	ANGPTL4	angiopoietin-like 4
ANKH	56172	ANKH	ankylosis, progressive homolog (mouse)
ANO1	55107	ANO1	anoctamin 1, calcium activated chlorid...
ANXA10	11199	ANXA10	annexin A10
APOD	347	APOD	apolipoprotein D
ARG1	383	ARG1	arginase, liver
ATG10	83734	ATG10	ATG10 autophagy related 10 homolog (S....
AVL9	23080	AVL9	AVL9 homolog (S. cerevisiase)
BIRC3	330	BIRC3	baculoviral IAP repeat containing 3
BMP2	650	BMP2	bone morphogenetic protein 2
BPGM	669	BPGM	2,3-bisphosphoglycerate mutase
BTBD3	22903	BTBD3	BTB (POZ) domain containing 3
BTC	685	BTC	betacellulin

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/



Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

click on rows to select gene sets, click a gene set name to view the gene set page

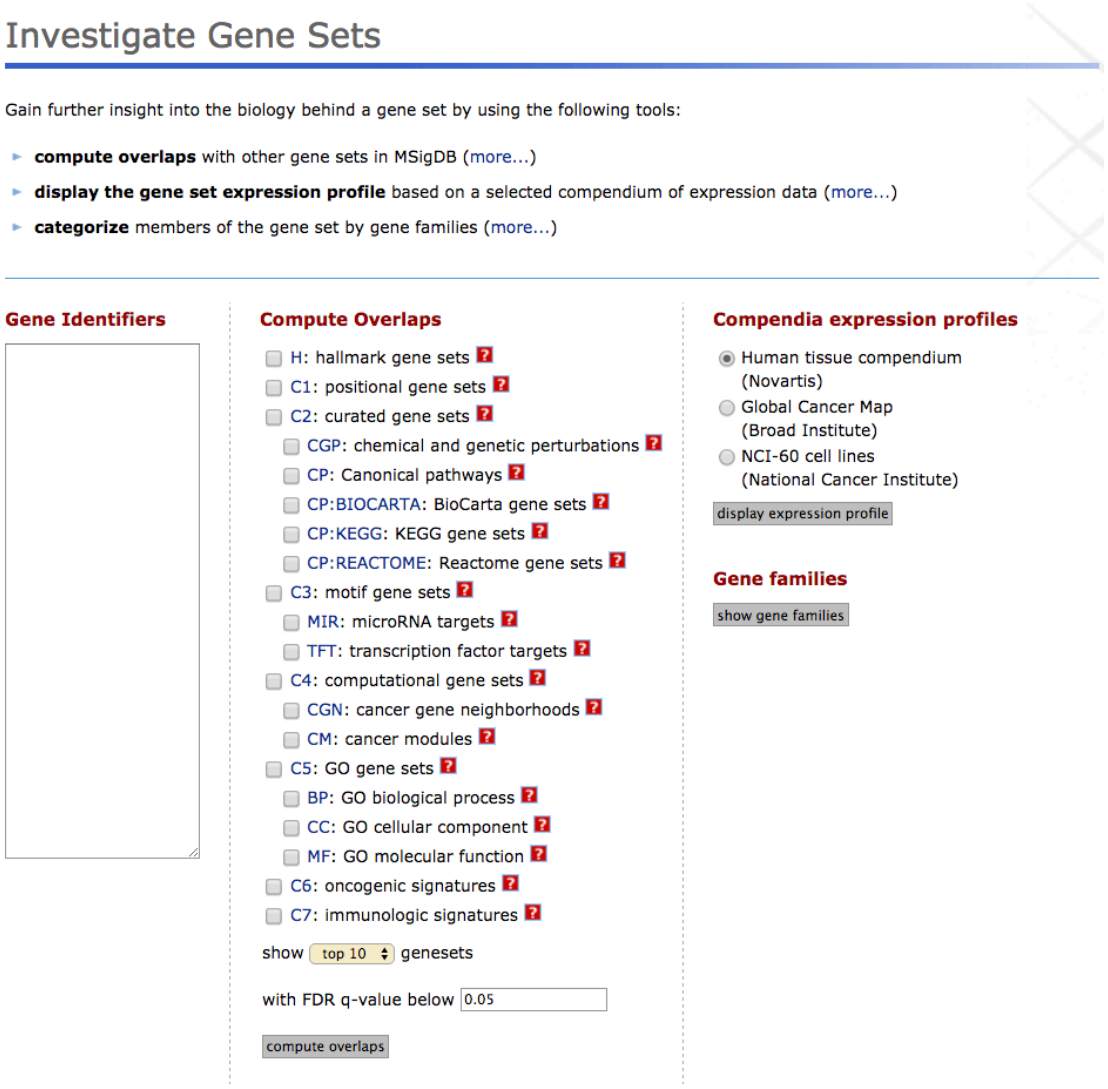
select all 305 **0** gene sets selected Select An Action...

<< < **1** 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 > >> 10

name	# genes	description	collections	organism	contributor
AACTTT_UNKNOWN	1890	Genes with promoter regions [-2kb,2kb] around transcription start site containing motif AACTTT. Motif does not match any known transcription factor	C3 TFT	Homo sapiens	Broad Institute
ABE_VEGFA_TARGETS_30MIN	29	Genes up-regulated in HUVEC cells (endothelium) at 30 min after VEGFA [GeneID=7422] stimulation.	C2 CGP	Homo sapiens	University of Washington
ACCATTT,MIR-522	160	Targets of MicroRNA ACCATTT,MIR-522	C3 MIR	Homo sapiens	Broad Institute
ACTGAAA,MIR-30A-3P,MIR-30E-3P	201	Targets of MicroRNA ACTGAAA,MIR-30A-3P,MIR-30E-3P	C3 MIR	Homo sapiens	Broad Institute
AGCATTA,MIR-155	134	Targets of MicroRNA AGCATTA,MIR-155	C3 MIR	Homo sapiens	Broad Institute
ATAGGAA,MIR-202	102	Targets of MicroRNA ATAGGAA,MIR-202	C3 MIR	Homo sapiens	Broad Institute
ATATGCA,MIR-448	212	Targets of MicroRNA ATATGCA,MIR-448	C3 MIR	Homo sapiens	Broad Institute
ATGCAGT,MIR-217	115	Targets of MicroRNA ATGCAGT,MIR-217	C3 MIR	Homo sapiens	Broad Institute
BENPORATH_CYCLING_GENES	648	Genes showing cell-cycle stage-specific expression [PMID=12058064].	C2 CGP	Homo sapiens	Broad Institute
BIOCARTA_TEL_PATHWAY	18	Telomeres, Telomerase, Cellular Aging, and Immortality	C2 CP CP:BIOCARTA	Homo sapiens	BioCarta

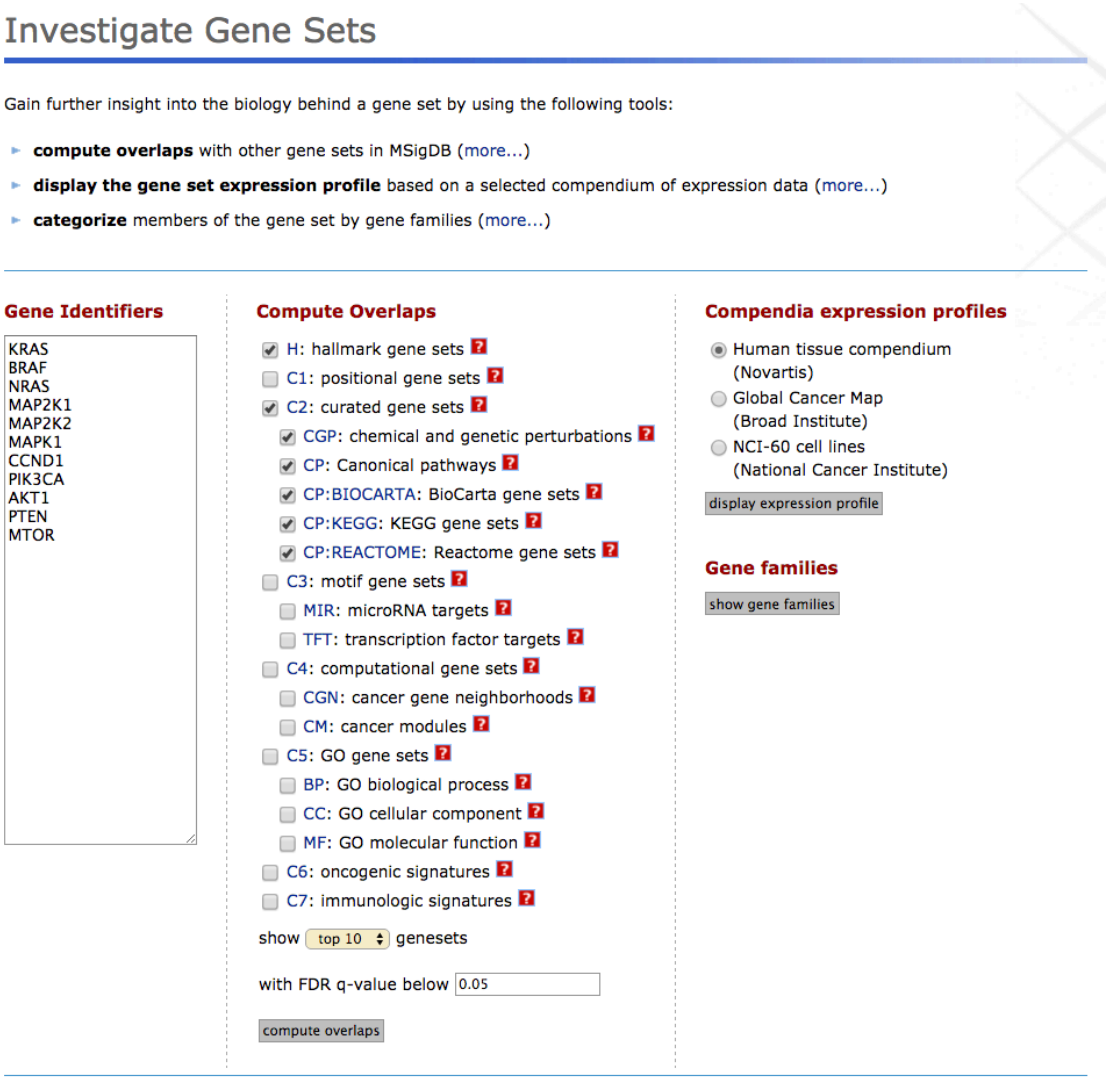
Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/



Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/



Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

Compute Overlaps for Selected Genes

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
C2, H	10	4776	11	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

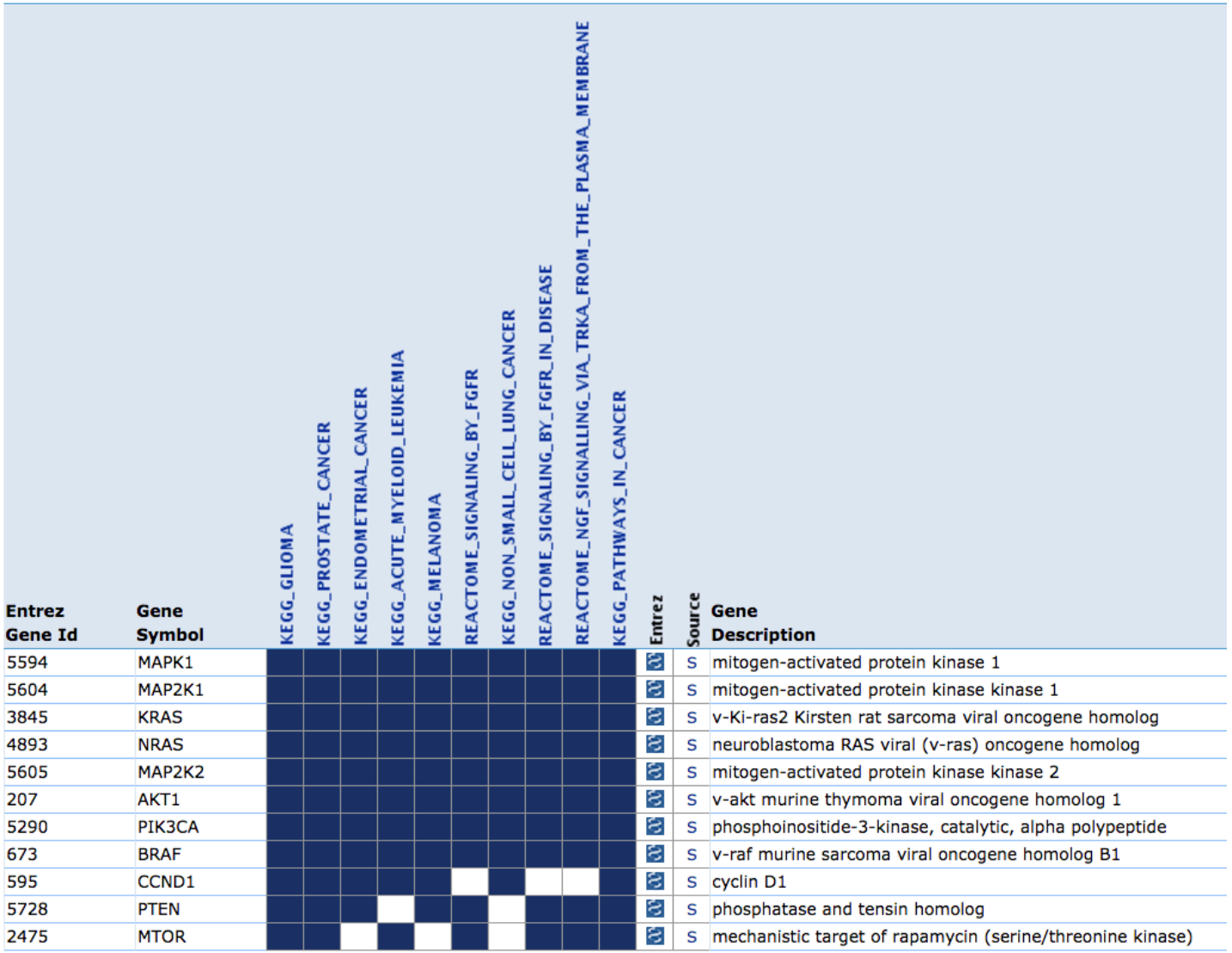
Save to: Excel | GenomeSpace

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
KEGG_GLIOMA [65]	Glioma	11		$1.85\ e^{-32}$	$8.85\ e^{-29}$
KEGG_PROSTATE_CANCER [89]	Prostate cancer	11		$7.56\ e^{-31}$	$1.8\ e^{-27}$
KEGG_ENDOMETRIAL_CANCER [52]	Endometrial cancer	10		$1.5\ e^{-29}$	$2.39\ e^{-26}$
KEGG_ACUTE_MYELOID_LEUKEMIA [60]	Acute myeloid leukemia	10		$7.16\ e^{-29}$	$8.55\ e^{-26}$
KEGG_MELANOMA [71]	Melanoma	10		$4.39\ e^{-28}$	$4.19\ e^{-25}$
REACTOME_SIGNALING_BY_FGFR [112]	Genes involved in Signaling by FGFR	10		$5.37\ e^{-26}$	$4.28\ e^{-23}$
KEGG_NON_SMALL_CELL_LUNG_CANCER [54]	Non-small cell lung cancer	9		$1.16\ e^{-25}$	$7.91\ e^{-23}$
REACTOME_SIGNALING_BY_FGFR_IN_DISEASE [127]	Genes involved in Signaling by FGFR in disease	10		$1.98\ e^{-25}$	$1.18\ e^{-22}$
REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_OM_THE_PLASMA_MEMBRANE [137]	Genes involved in NGF signalling via TRKA from the plasma membrane	10		$4.35\ e^{-25}$	$2.31\ e^{-22}$
KEGG_PATHWAYS_IN_CANCER [328]	Pathways in cancer	11		$2.07\ e^{-24}$	$9.19\ e^{-22}$

Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

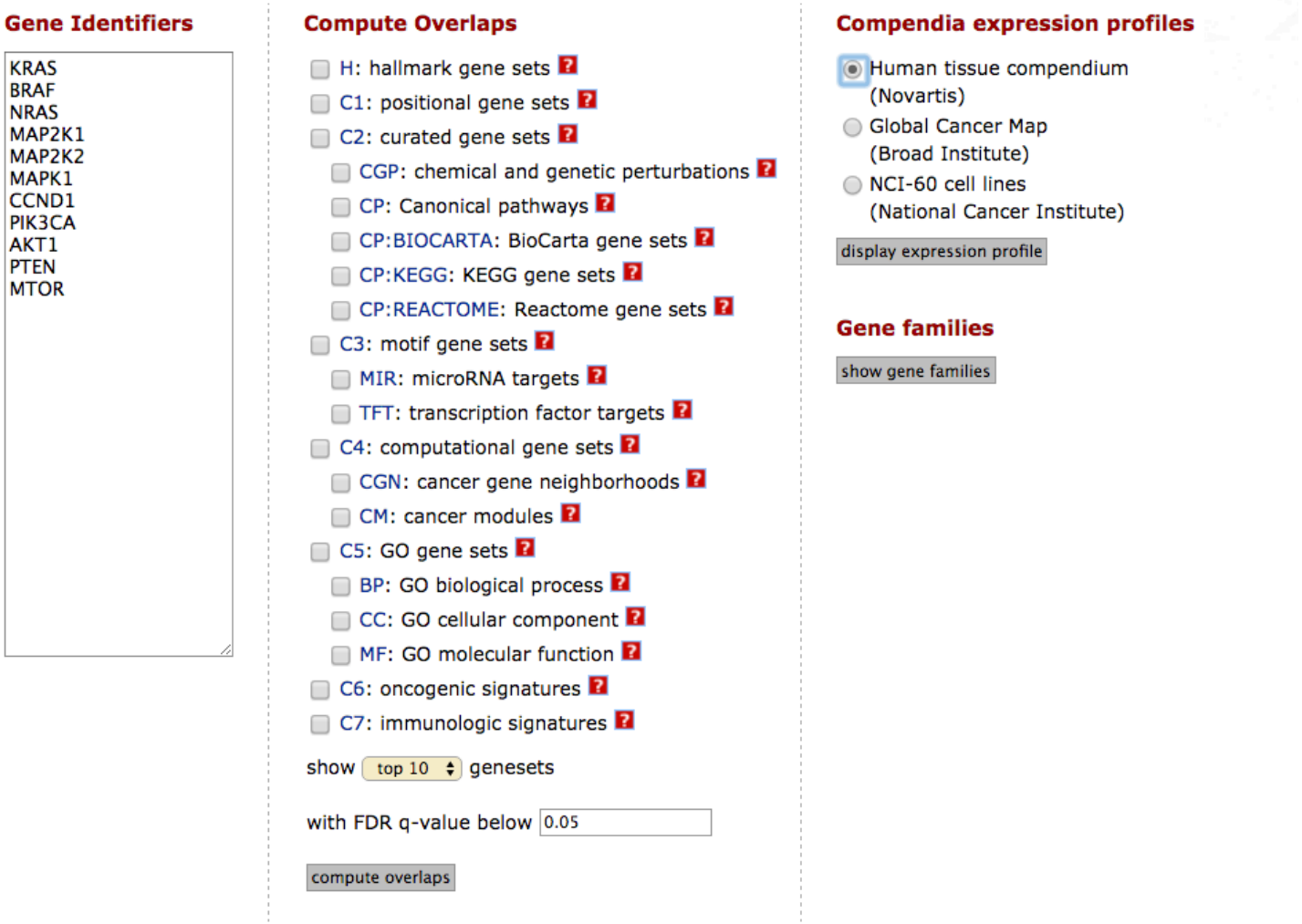
Gene/geneset overlap matrix



Entrez Gene Id	Gene Symbol	KEGG_GLIOMA	KEGG_PROSTATE_CANCER	KEGG_ENDOMETRIAL_CANCER	KEGG_ACUTE_MYELOID_LEUKEMIA	KEGG_MELANOMA	REACTOME_SIGNALING_BY_FGFR	KEGG_NON_SMALL_CELL_LUNG_CANCER	REACTOME_SIGNALING_BY_FGFR_IN_DISEASE	REACTOME_NGF_SIGNALLING_VIA_TRKA_FROM_THE_PLASMA_MEMBRANE	KEGG_PATHWAYS_IN_CANCER	Entrez	Source	Gene Description
5594	MAPK1	■	■	■	■	■	■	■	■	■	■		S	mitogen-activated protein kinase 1
5604	MAP2K1	■	■	■	■	■	■	■	■	■	■		S	mitogen-activated protein kinase kinase 1
3845	KRAS	■	■	■	■	■	■	■	■	■	■		S	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
4893	NRAS	■	■	■	■	■	■	■	■	■	■		S	neuroblastoma RAS viral (v-ras) oncogene homolog
5605	MAP2K2	■	■	■	■	■	■	■	■	■	■		S	mitogen-activated protein kinase kinase 2
207	AKT1	■	■	■	■	■	■	■	■	■	■		S	v-akt murine thymoma viral oncogene homolog 1
5290	PIK3CA	■	■	■	■	■	■	■	■	■	■		S	phosphoinositide-3-kinase, catalytic, alpha polypeptide
673	BRAF	■	■	■	■	■	■	■	■	■	■		S	v-raf murine sarcoma viral oncogene homolog B1
595	CCND1	■	■	■	■	■		■			■		S	cyclin D1
5728	PTEN	■	■	■		■	■		■	■	■		S	phosphatase and tensin homolog
2475	MTOR	■	■		■		■		■	■	■		S	mechanistic target of rapamycin (serine/threonine kinase)

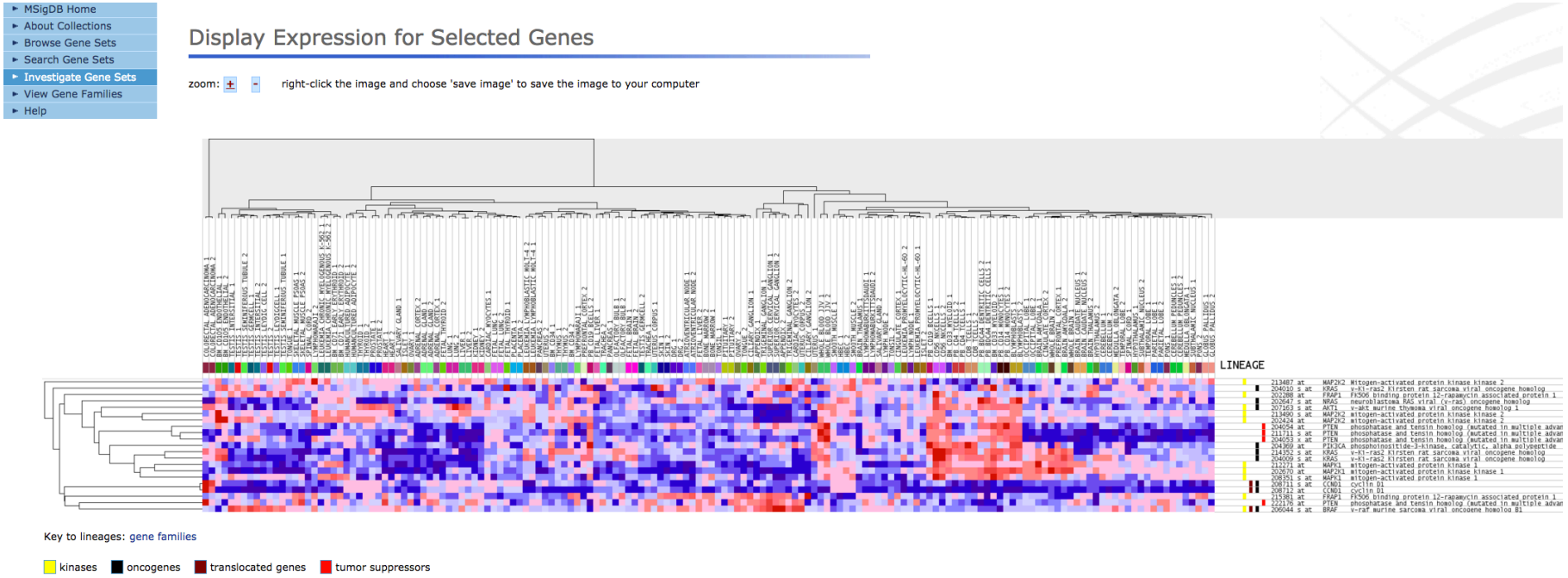
Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/



Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/



Molecular Signatures Database (MSigDB)

http://software.broadinstitute.org/gsea/msigdb/

View Gene Families for Selected Genes

The following table provides a functional overview of the MSigDB gene sets by categorizing their genes into a small number of carefully chosen "gene families". To categorize the genes in a gene set, use the gene set page or the Investigate Gene Sets page.

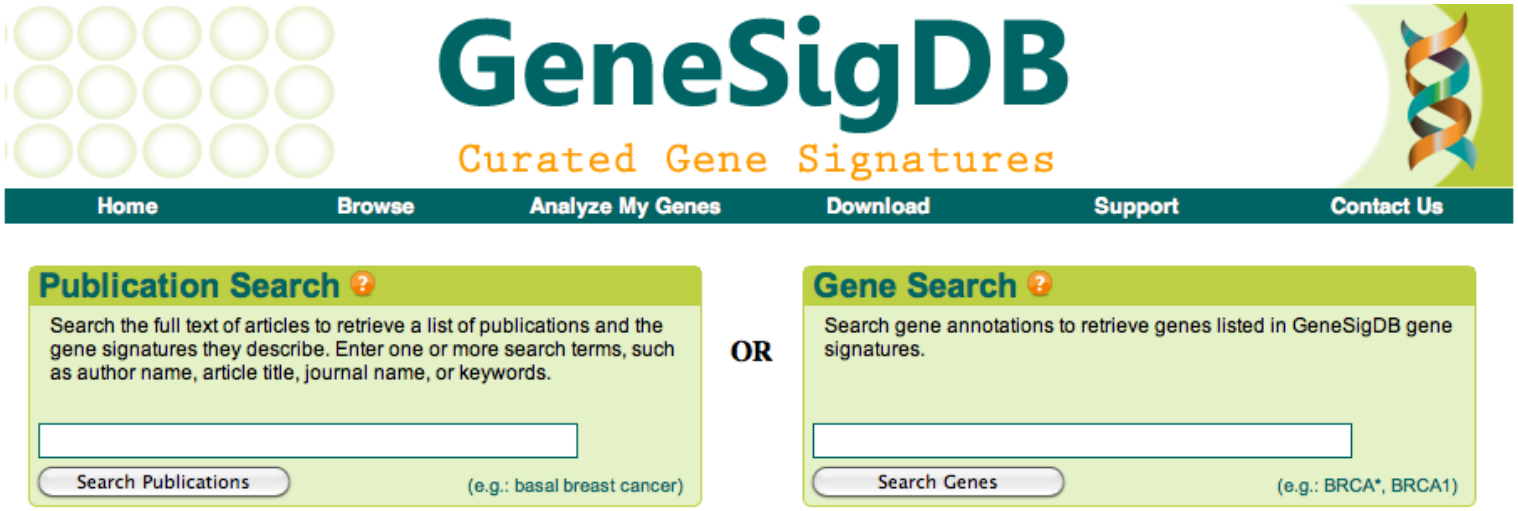
Click on a gene family or gene family intersection to retrieve annotations for those genes.

	cytokines and growth factors	transcription factors	homeodomain proteins	cell differentiation markers	protein kinases	translocated cancer genes	oncogenes	tumor suppressors
tumor suppressors	0	0	0	0	0	0	0	1
oncogenes	0	0	0	0	2	2	6	
translocated cancer genes	0	0	0	0	1	2		
protein kinases	0	0	0	0	6			
cell differentiation markers	0	0	0	0				
homeodomain proteins	0	0	0					
transcription factors	0	0						
cytokines and growth factors	0							

Members of these "gene families" share a common feature such as homology or biochemical activity. They do not necessarily have common origins. For the source of each "gene family" definition, click here.

Curated Gene Signatures (GeneSigDB)

http://genesigdb.org/genesigdb/



DSigDB

http://tanlab.ucdenver.edu/DSigDB

Bioinformatics, 31(18), 2015, 3069–3071
doi: 10.1093/bioinformatics/btv313
Advance Access Publication Date: 19 May 2015
Applications Note

OXFORD

Systems biology

DSigDB: drug signatures database for gene set analysis

Minjae Yoo[1,†], Jimin Shin[1,†], Jihye Kim[1], Karen A. Ryall[1], Kyubum Lee[2], Sunwon Lee[2], Minji Jeon[2], Jaewoo Kang[2] and Aik Choon Tan[1,2,*]

[1]Department of Medicine, Translational Bioinformatics and Cancer Systems Biology Laboratory, Division of Medical Oncology, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA and [2]Department of Computer Science and Engineering, Korea University, Seoul 136-713, South Korea

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on February 17, 2015; revised on April 30, 2015; accepted on May 13, 2015

Abstract

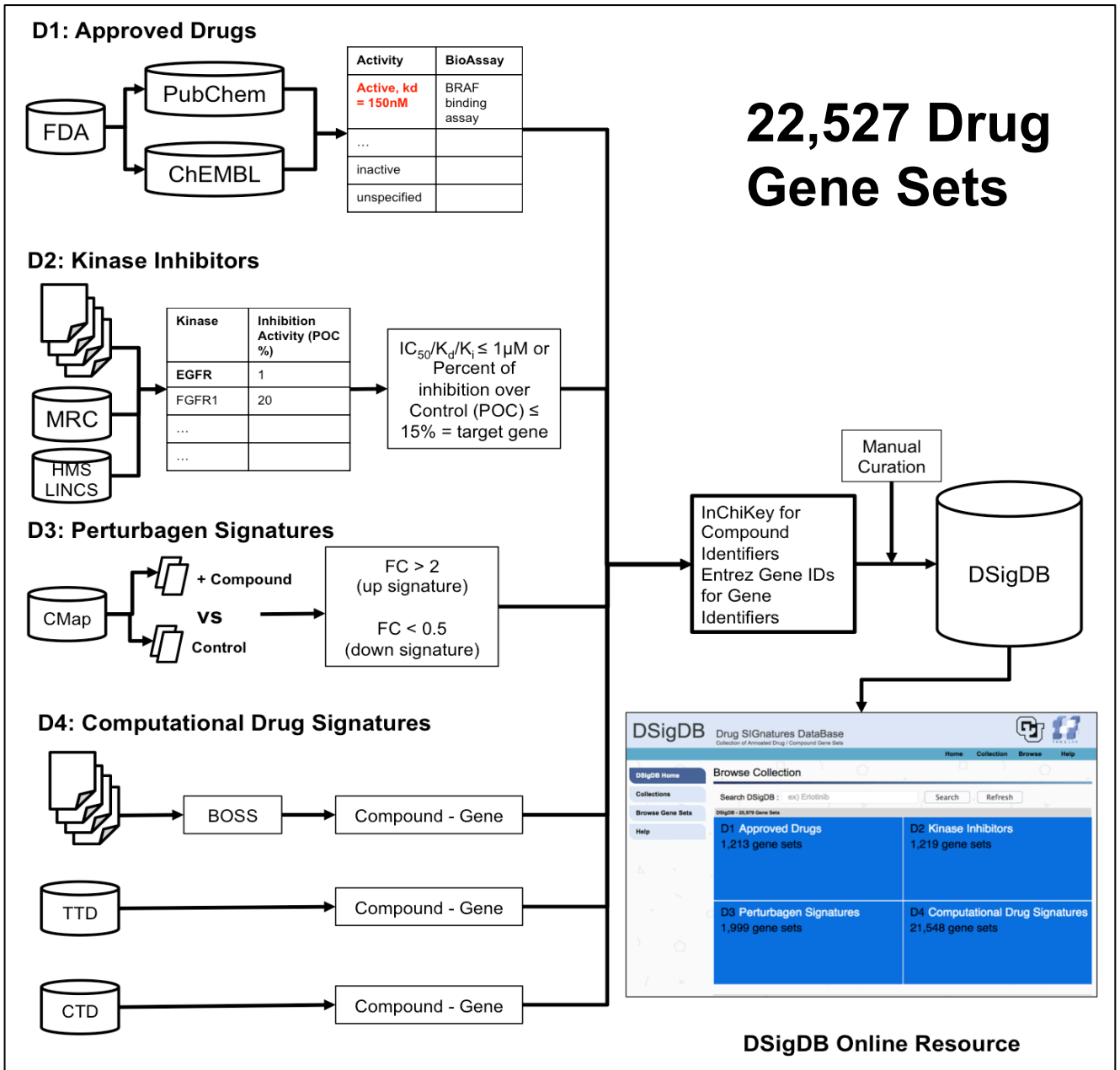
Summary: We report the creation of Drug Signatures Database (DSigDB), a new gene set resource that relates drugs/compounds and their target genes, for gene set enrichment analysis (GSEA). DSigDB currently holds 22 527 gene sets, consists of 17 389 unique compounds covering 19 531 genes. We also developed an online DSigDB resource that allows users to search, view and download drugs/compounds and gene sets. DSigDB gene sets provide seamless integration to GSEA software for linking gene expressions with drugs/compounds for drug repurposing and translational research.

Availability and implementation: DSigDB is freely available for non-commercial use at http://tanlab.ucdenver.edu/DSigDB.

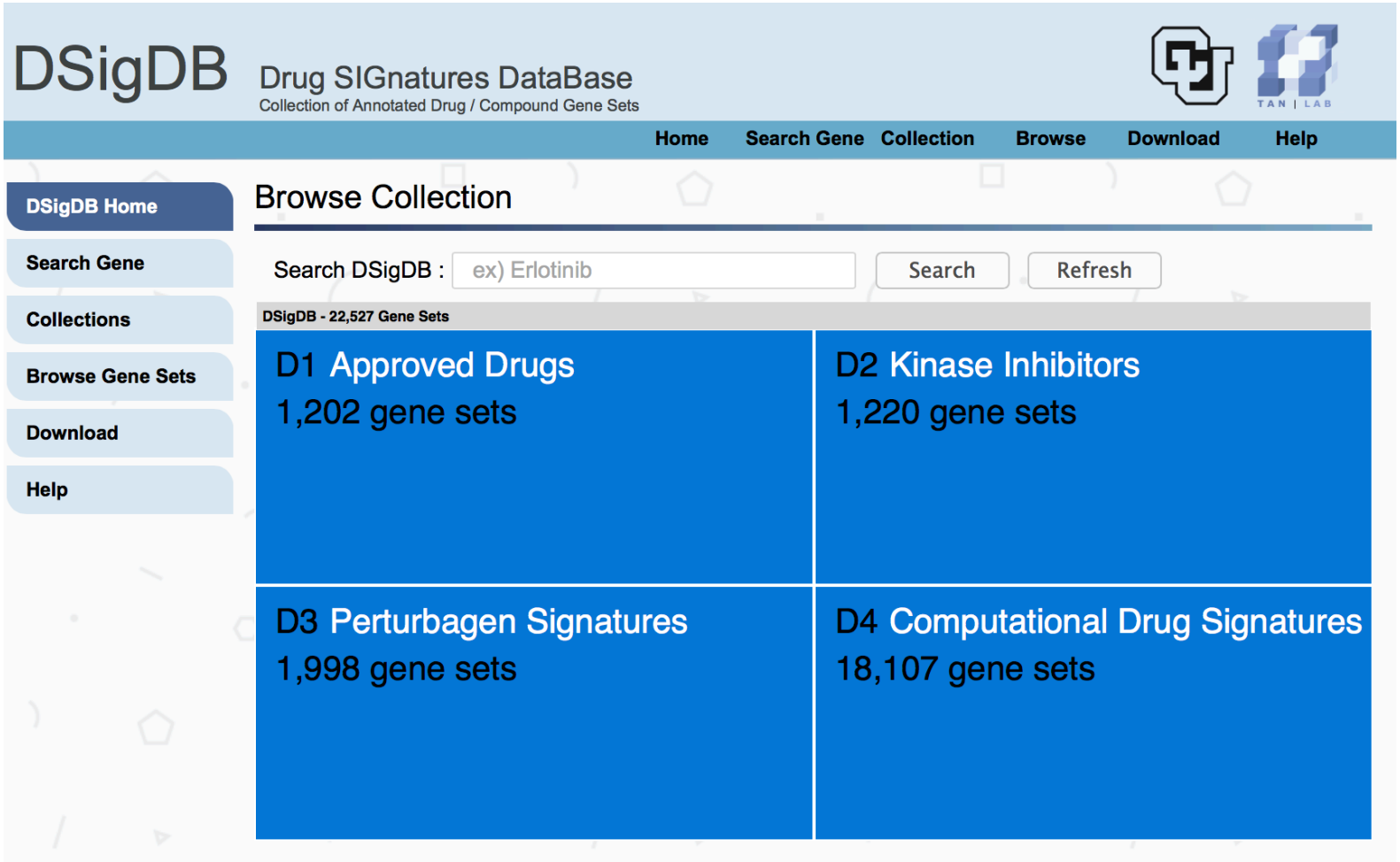
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: aikchoon.tan@ucdenver.edu

DSigDB Workflow

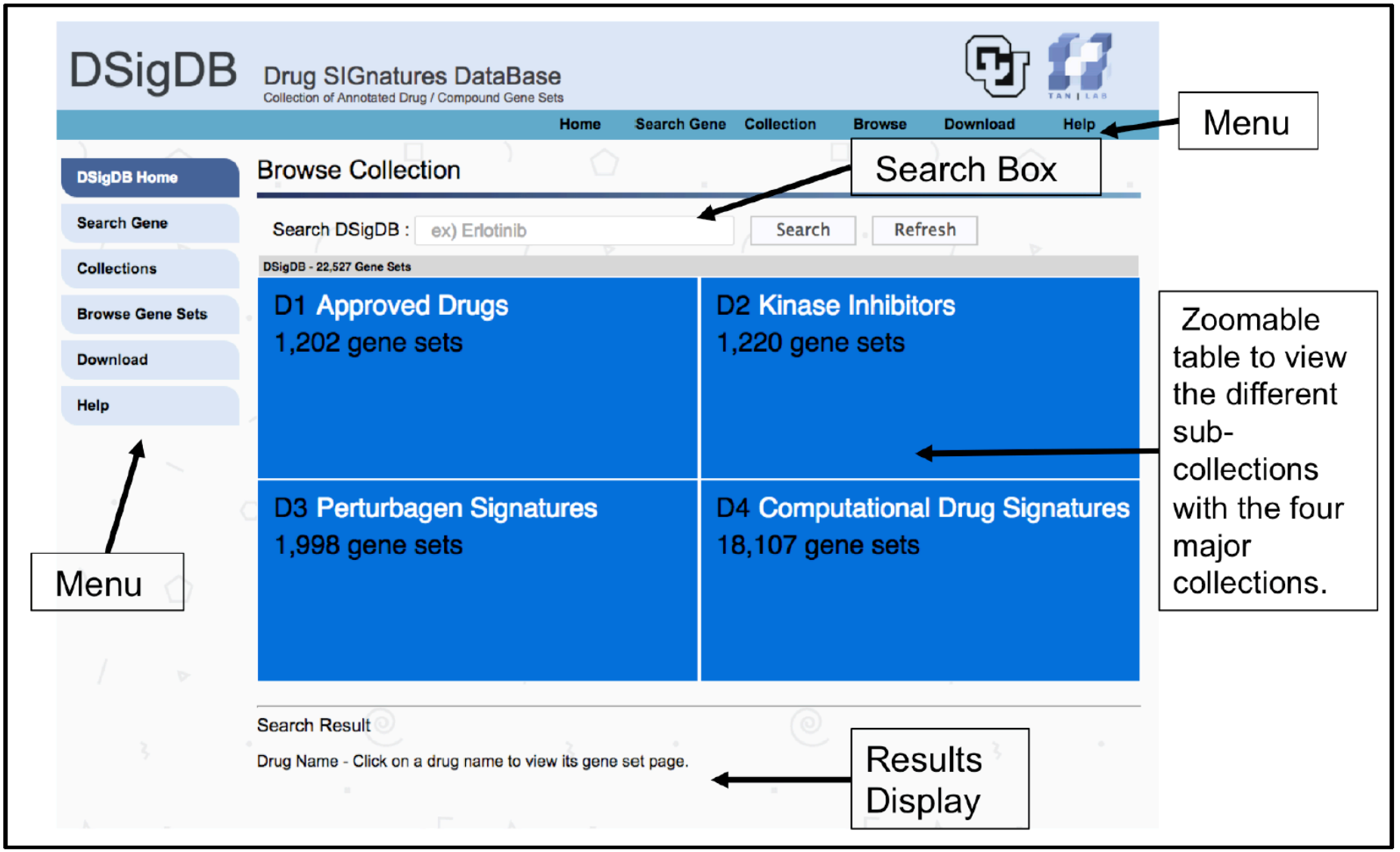


DSigDB

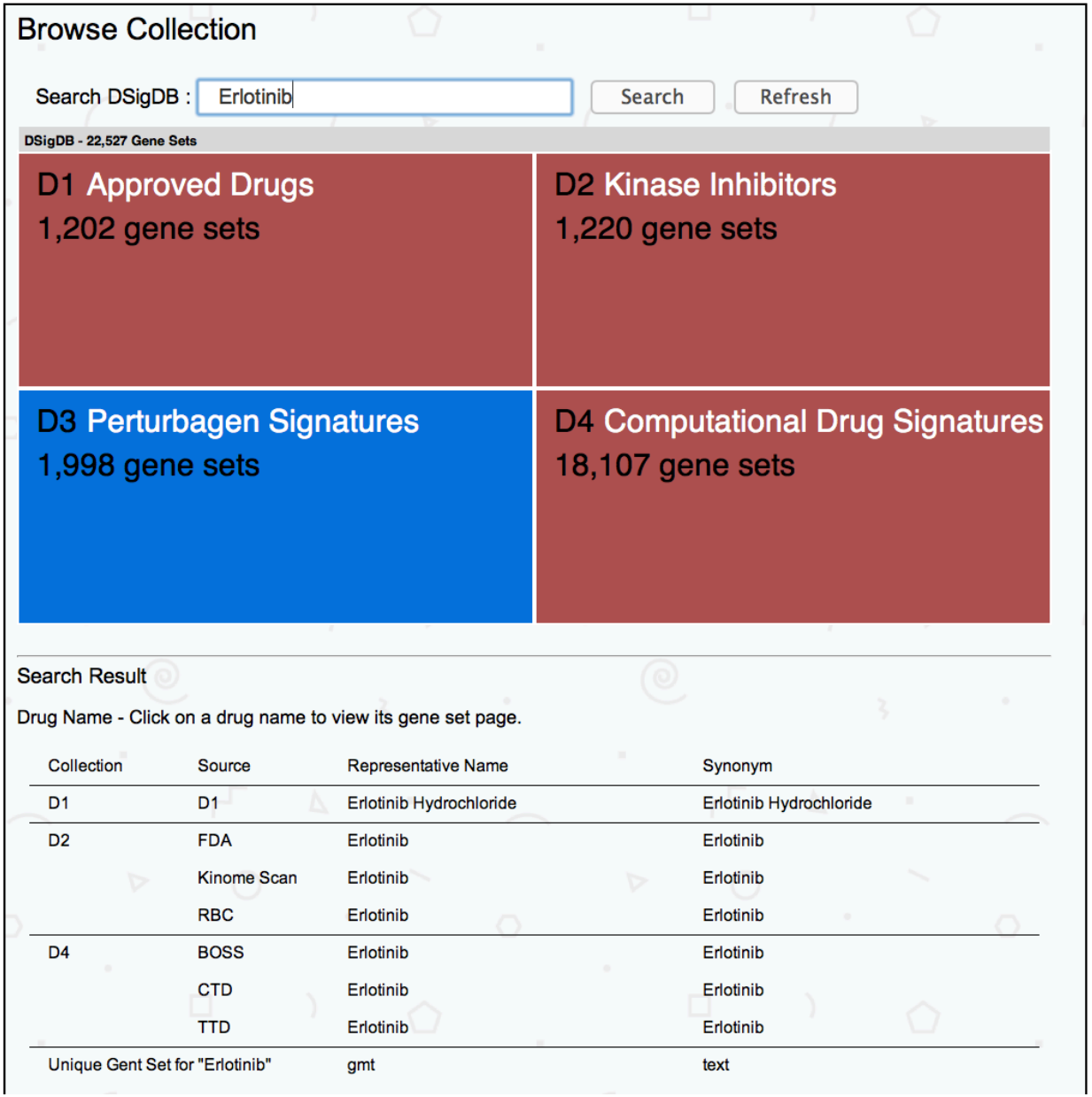


http://tanlab.ucdenver.edu/DSigDB/

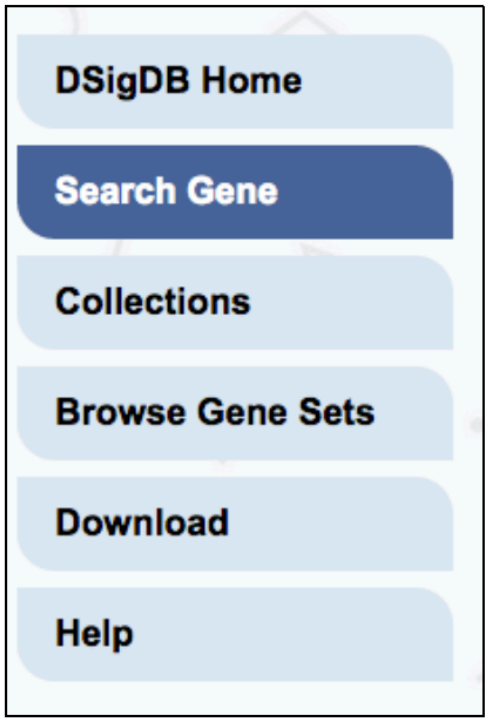
DSigDB

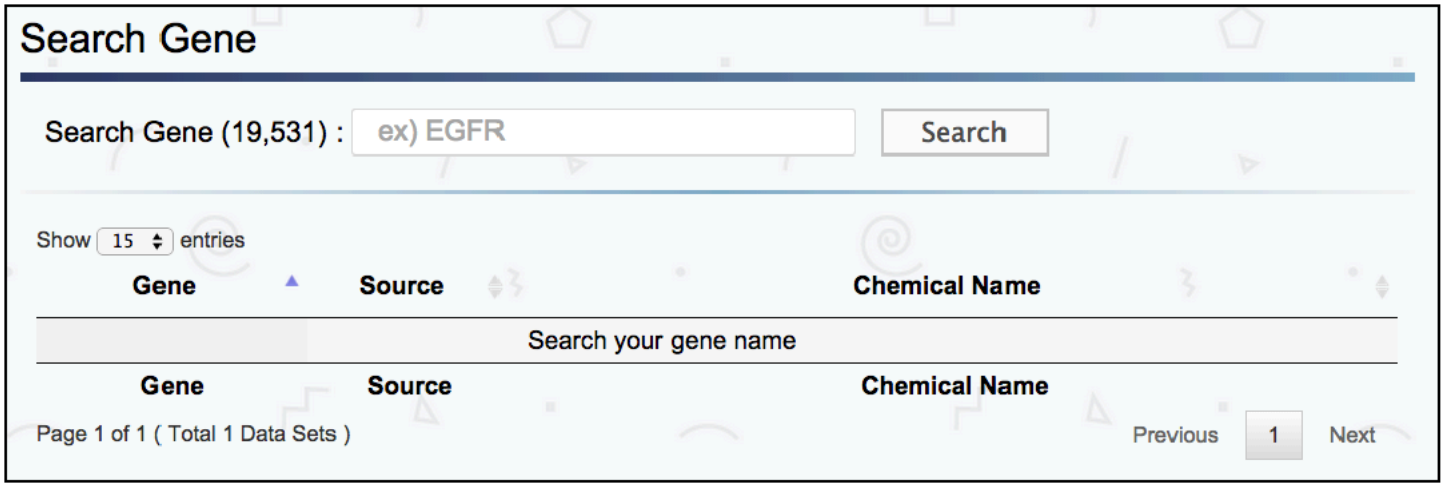

DSigDB
Drug SIGnatures DataBase
Collection of Annotated Drug / Compound Gene Sets
Home
Search Gene
Collection
Browse
Download
Help
Menu
DSigDB Home
Search Gene
Collections
Browse Gene Sets
Download
Help
Browse Collection
Search Box
Search DSigDB :
ex) Erlotinib
Search
Refresh
DSigDB - 22,527 Gene Sets
D1 Approved Drugs
1,202 gene sets
D2 Kinase Inhibitors
1,220 gene sets
D3 Perturbagen Signatures
1,998 gene sets
D4 Computational Drug Signatures
18,107 gene sets
Zoomable table to view the different sub-collections with the four major collections.
Menu
Search Result
Drug Name - Click on a drug name to view its gene set page.
Results Display

Drug Search


Browse Collection
Search DSigDB : Erlotinib
Search
Refresh
DSigDB - 22,527 Gene Sets
D1 Approved Drugs
1,202 gene sets
D2 Kinase Inhibitors
1,220 gene sets
D3 Perturbagen Signatures
1,998 gene sets
D4 Computational Drug Signatures
18,107 gene sets
Search Result
Drug Name - Click on a drug name to view its gene set page.
Collection Source Representative Name Synonym
D1 D1 Erlotinib Hydrochloride Erlotinib Hydrochloride
D2 FDA Erlotinib Erlotinib
Kinome Scan Erlotinib Erlotinib
RBC Erlotinib Erlotinib
D4 BOSS Erlotinib Erlotinib
CTD Erlotinib Erlotinib
TTD Erlotinib Erlotinib
Unique Gent Set for "Erlotinib" gmt text

Gene Search


DSigDB Home
Search Gene
Collections
Browse Gene Sets
Download
Help


Search Gene
Search Gene (19,531) :
ex) EGFR
Search
Show 15 entries
Gene
Source
Chemical Name
Search your gene name
Gene
Source
Chemical Name
Page 1 of 1 (Total 1 Data Sets)
Previous
1
Next

Gene Search Result

Search Gene

Search Gene (19,531) : EGFR Search

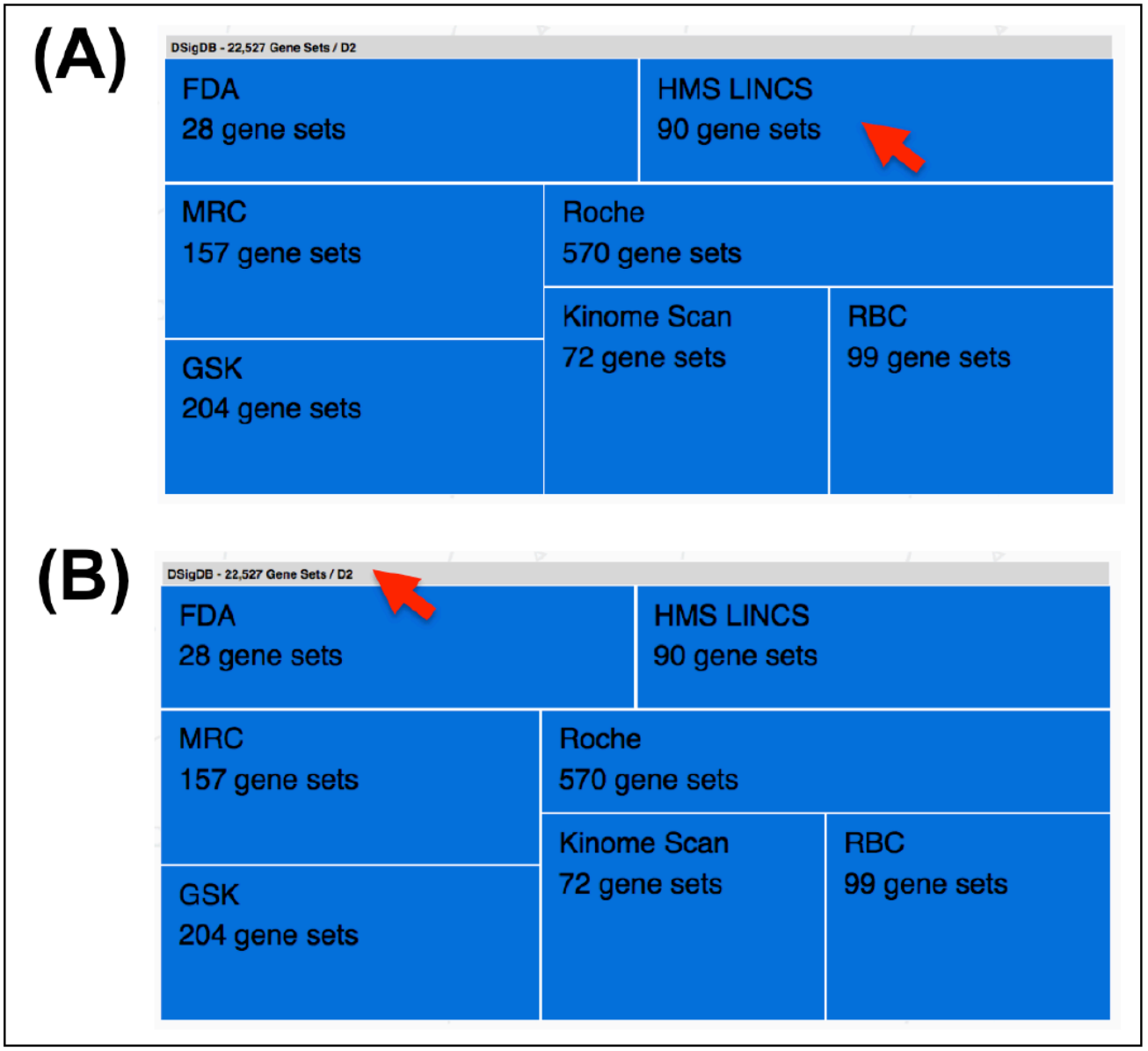
Show 15 entries

Gene	Source	Chemical Name
EGFR	D1	chlorpromazine
EGFR	D1	afatinib
EGFR	D1	thioridazine
EGFR	D1	vandetanib
EGFR	D1	baciguent
EGFR	D1	levodopa
EGFR	D1	hexachlorophene
EGFR	D1	zafirlukast
EGFR	D1	erlotinib hydrochloride
EGFR	D1	miconazole
EGFR	D1	tamoxifen
EGFR	D1	crystal violet
EGFR	D1	methyldopa
EGFR	D1	dobutamine
EGFR	D1	crizotinib
Gene	Source	Chemical Name

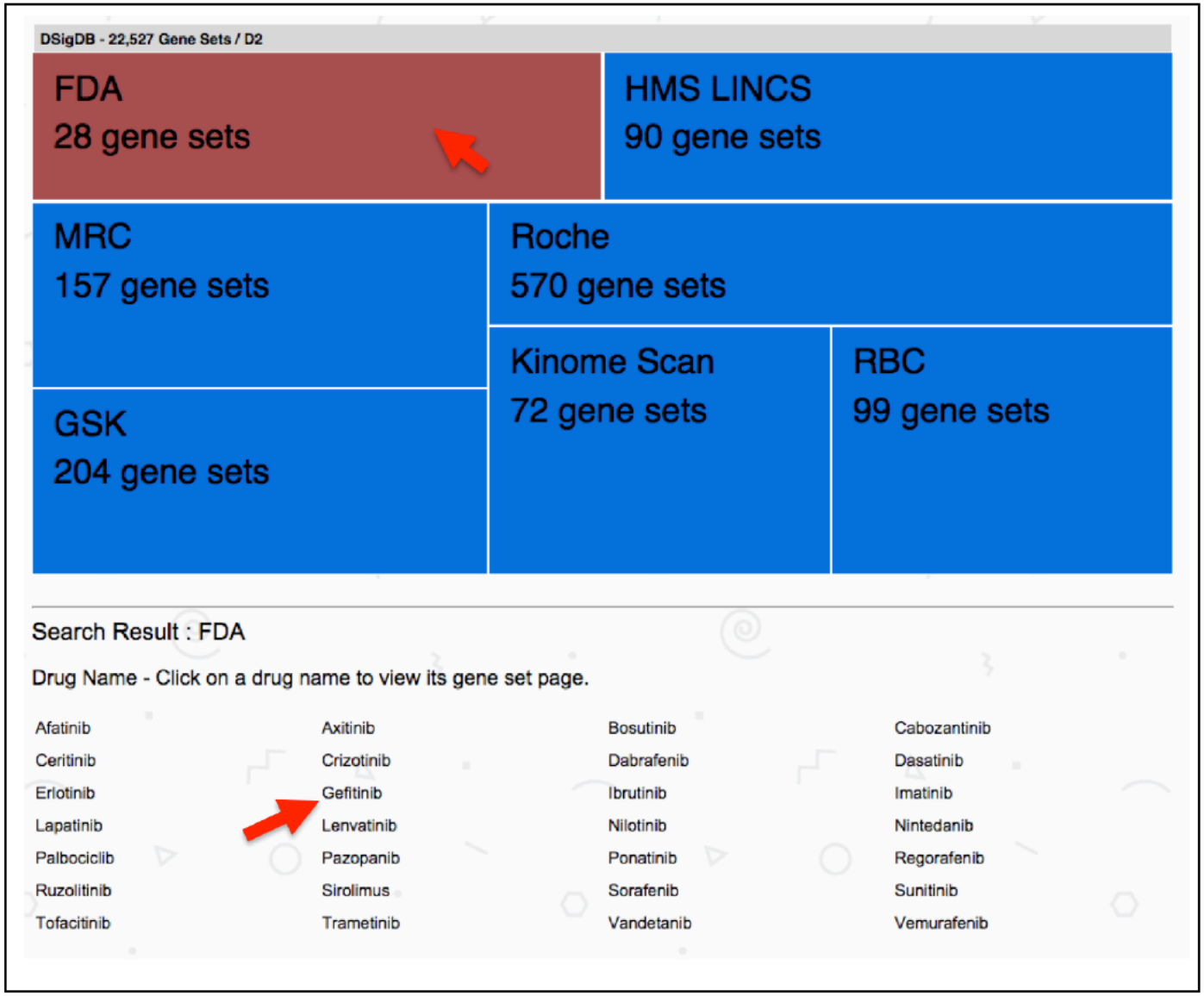
Page 1 of 42 (Total 616 Data Sets)

Previous 1 2 3 4 5 … 42 Next

Browsing Collection



Browsing Collection



Compound Webpage

Gene Set: D2 : FDA - Gefitinib

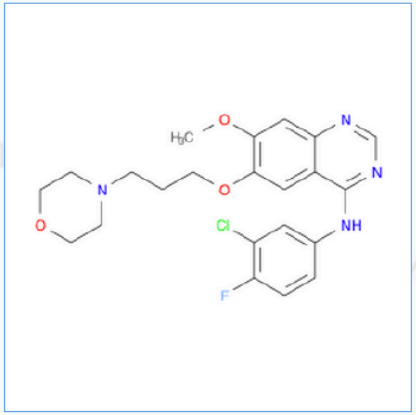
Collection: D2 : FDA

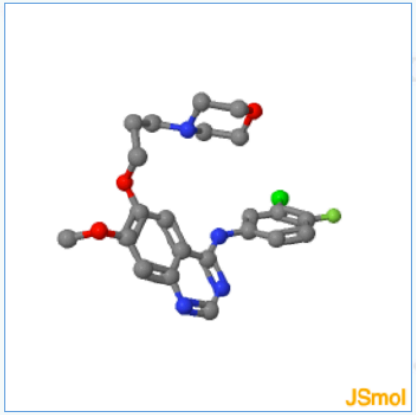
Chemical Name: Gefitinib

FDA	NPC	WHO	Indian	Australia	China	Traditional Herbal	Clinical Trail
Approved	Not	Not	Approved	Not	Not	Not	Not

Molecular Weight	Hydrogen Bond Donor Count	Hydrogen Bond Acceptor Count	cLogP	Lipinski Rule
446.902 g/mol	1	6	4.2865	True

Structure





InChI: InChI=1S/C22H24ClFN4O3/c1-29-20-13-19-16(12-21(20)31-8-2-5-28-6-9-30-10-7-28)22(26-14-25-19)27-15-3-4-18(24)17(23)11-15/h3-4,11-14H,2,5-10H2,1H3,(H,25,26,27)

InChIKey: XGALLCVXEZPNRQ-UHFFFAOYSA-N

Links

CAS Num : 184475-35-2

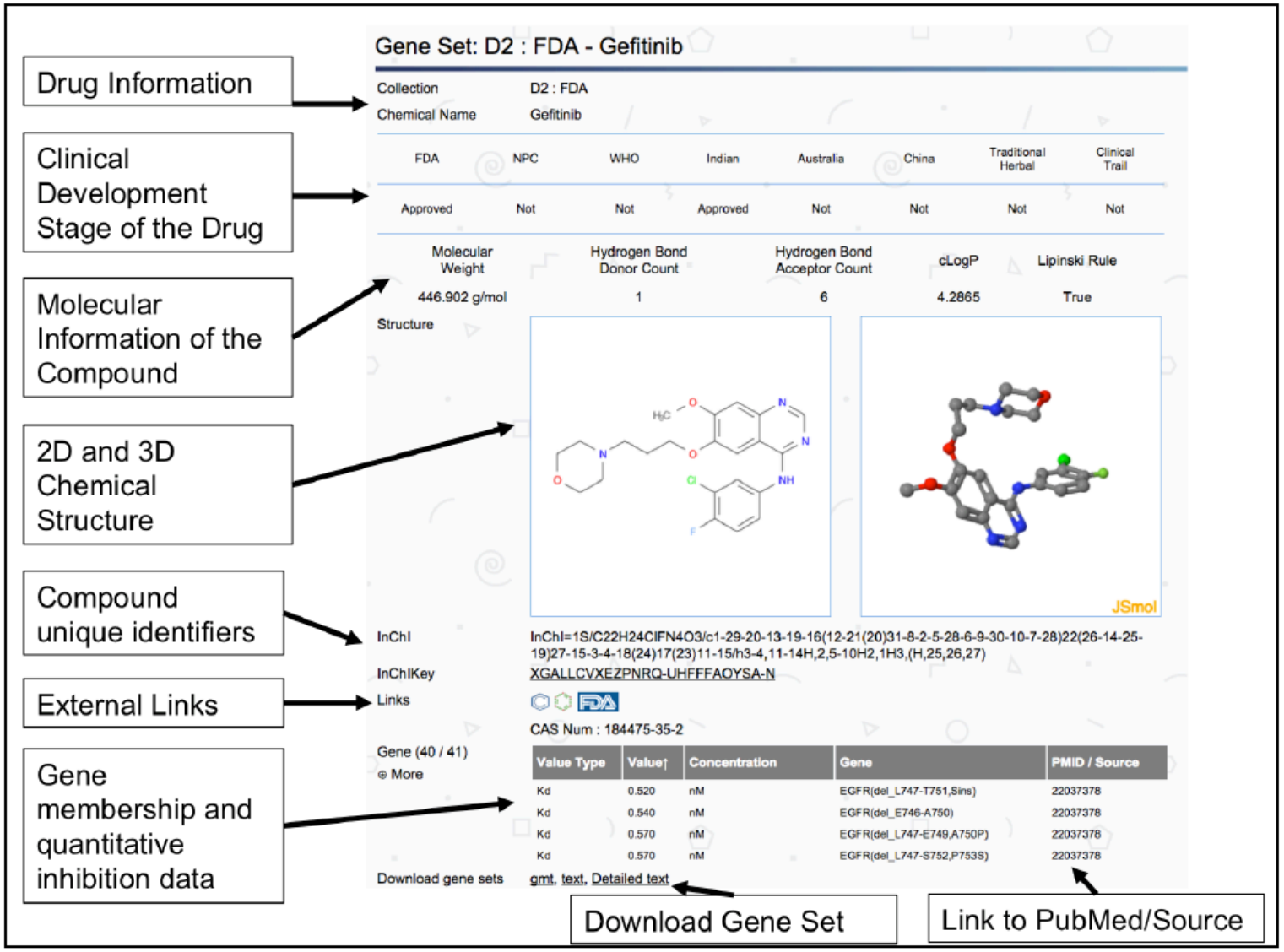
Gene (40 / 41)

⊕ More

Value Type	Value↑	Concentration	Gene	PMID / Source
Kd	0.520	nM	EGFR(del_L747-T751,Sins)	22037378
Kd	0.540	nM	EGFR(del_E746-A750)	22037378
Kd	0.570	nM	EGFR(del_L747-E749,A750P)	22037378
Kd	0.570	nM	EGFR(del_L747-S752,P753S)	22037378

Download gene sets: gmt, text, Detailed text

Compound Webpage



Collections

Collection	Description	Unique Number of Genes	Number of Gene Sets	Download
DSigDB	All Gene Sets.	19,531	22,527	GMT File
D1 : FDA Approved (browse 1,202 gene sets)	FDA Approved Drug Gene Sets.	1,288	1,202	GMT File
D2 : Kinase Inhibitors	Kinase Inhibitors Gene Sets based on in vitro kinase profiling assays.	407	1,220	GMT File
FDA (browse 28 gene sets)	FDA Approved Kinase Inhibitors.	341	28	GMT File
HMS LINCS (browse 90 gene sets)	Kinase inhibition assays extracted from HMS LINCS database.	381	90	GMT File
MRC (browse 157 gene sets)	Kinase inhibition assays extracted from MRC Kinome Inhibition database.	137	157	GMT File
GSK (browse 204 gene sets)	GSK Published Kinase Inhibitor Set (PKIS), kinase inhibitors used as chemical probes.	116	204	GMT File
Roche (browse 570 gene sets)	Kinase Inhibitors profiled by Roche.	153	570	GMT File
RBC (browse 99 gene sets)	Kinase Inhibitors profiled by Reaction Biology Corporation.	246	99	GMT File
KinomeScan (browse 72 gene sets)	Kinase Inhibitors profiled by DiscoveryRx using KinomeScan technology.	374	72	GMT File
D3 : Perturbagen Signatures (browse 1,998 gene sets)	7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02).	11,137	1,998	GMT File
CMAP (browse 1,998 gene sets)	7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02).	11,137	1,998	GMT File
D4 : Computational Drug Signatures	Drug signatures extracted from literatures using a mixture of manual curation and by automatic computational approaches.	18,854	18,107	GMT File
BOSS (browse 2,114 gene sets)	Text mining approach of drug-gene targets using Biomedical Object Search System (BOSS).	3,354	2,114	GMT File
CTD (browse 5,163 gene sets)	Curation of targets from Comparative Toxicogenomics Database (CTD).	18,700	5,163	GMT File
TTD (browse 10,830 gene sets)	Manual curation of targets from the Therapeutics Targets Database (TTD).	1,389	10,830	GMT File

Download Collections

Download

DSigDB provides several options for downloading the data.

Current Release
The current data release of DSigDB is Release 1 (released May 2015).

DSigDB Release 1 (Updated)

- DSigDBv1.0.gmt
- DSigDBv1.0.txt
- DSigDBv1.0 Detailed.txt

```
Compound : Gefitinib
EPHA6
STK10
MKNK1
EGFR
RIPK2
MAP2K5
HIPK4
ABL1
FLT3
CSNK1E
GAK
LYN
IRAK1
```

Drug	Gene	Type	Source	
Gefitinib	EGFR	Kd=40.0(nM)	FDA	
Gefitinib	EGFR	Kd=0.54(nM)	FDA	
Gefitinib	EGFR	Kd=0.98(nM)	FDA	
Gefitinib	ABL1	Kd=460.0(nM)	FDA	
Gefitinib	CDK7	Kd=610.0(nM)	FDA	
Gefitinib	EGFR	Kd=140.0(nM)	FDA	
Gefitinib	ABL1	Kd=680.0(nM)	FDA	
Gefitinib	ABL1	Kd=360.0(nM)	FDA	
Gefitinib	LCK	Kd=630.0(nM)	FDA	
Gefitinib	ABL1	Kd=480.0(nM)	FDA	
Gefitinib	MKNK1	Kd=290.0(nM)	FDA	
Gefitinib	SBK1	Kd=560.0(nM)	FDA	
Gefitinib	SLK	Kd=920.0(nM)	FDA	
Gefitinib	EGFR	Kd=1.1(nM)	FDA	
Gefitinib	ABL1	Kd=230.0(nM)	FDA	
Gefitinib	IRAK4	Kd=540.0(nM)	FDA	
Gefitinib	ERBB3	Kd=790.0(nM)	FDA	
Gefitinib	GAK	Kd=13.0(nM)	FDA	
Gefitinib	ABL1	Kd=780.0(nM)	FDA	
Gefitinib	LYN	Kd=990.0(nM)	FDA	
Gefitinib	IRAK1	Kd=69.0(nM)	FDA	
Gefitinib	CHEK2	Kd=800.0(nM)	FDA	

Use Case Example: EGFRwt NSCLC

Cell line	Histology	EGFR	KRAS	Gefitinib IC50 (μmol/L)
Sensitive				
H358	BAC	Wild-type	Mutant	0.18
H322	BAC	Wild-type	Wild-type	0.25
Calu-3	Adenocarcinoma	Wild-type	Wild-type	0.3
H1334	Large	Wild-type	Wild-type	0.3
H1648	Adenocarcinoma	Wild-type	Wild-type	0.38
HCC78	Adenocarcinoma	Wild-type	Wild-type	0.4
H2126	Large	Wild-type	Wild-type	1
HCC193	Adenocarcinoma	Wild-type	Wild-type	1.5
HCC95	Adenocarcinoma	Wild-type	Wild-type	1.9
Resistant				
H125	Adenosquamous	Wild-type	Wild-type	4.8
HCC44	Adenocarcinoma	Wild-type	Mutant	7.9
H1703	Squamous	Wild-type	Wild-type	8
HCC15	Squamous	Wild-type	Wild-type	9.4
A549	Adenocarcinoma	Wild-type	Wild-type	9.6
H157	Squamous	Wild-type	Mutant	12.8
H460	Large	Wild-type	Mutant	12.9
H520	Squamous	Wild-type	Wild-type	13.6
H1299	Large	Wild-type	Wild-type	14.7

(Adapted from Coldren et al MCR 2006)

Enriched Gene Sets

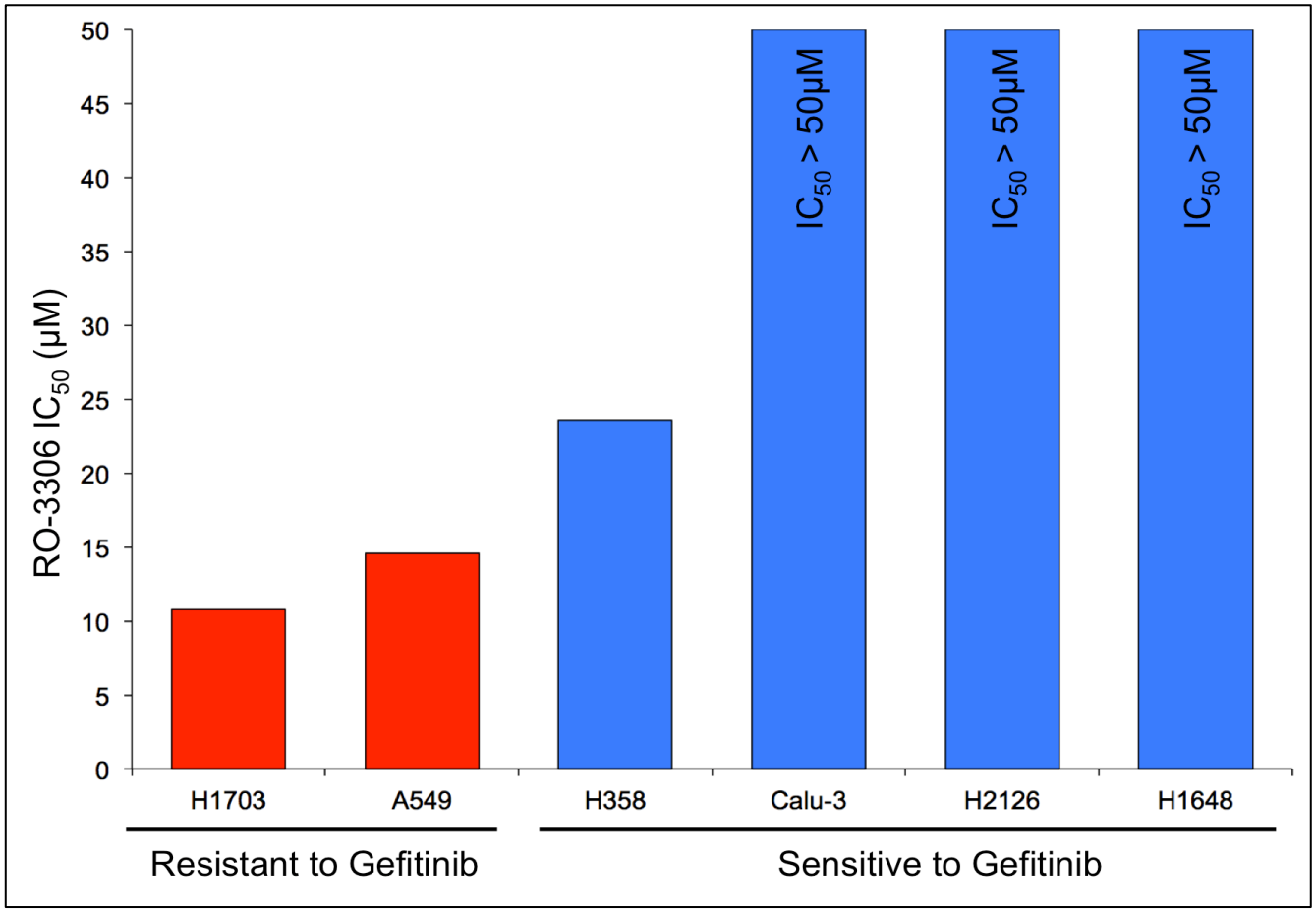
Enriched in Sensitive Group (p < 0.05)

GENE SET NAME	GENE SET SIZE	Normalized Enrichment Score	Nominal p-val	Intended targets	Inhibiting EGFR/ERBB2/ERBB3 based on Kinase Inhibition Assays		
					EGFR	ERBB2	ERBB3
CI-1033_KINOME SCAN	28	1.78	0.0000	EGFR/ERBB2	Yes	Yes	Yes
AZD-9291_LINCS	43	1.63	0.0125	EGFR	Yes	Yes	No
ZM-447439_LINCS	41	1.55	0.0285	AURKA	Yes	Yes	Yes
AZD-2171_KINOME SCAN	42	1.52	0.0271	VEGFR2/PDGFRA/PDGFRB	Yes	No	Yes
SB-203580_KINOME SCAN	18	1.52	0.0496	p38-alpha	Yes	No	No
WH-4-023_LINCS	124	1.48	0.0101	LCK	Yes	Yes	Yes
PP-242_KINOME SCAN	111	1.48	0.0116	MTOR/PIK3CA	Yes	Yes	No
CABOZANTINIB_FDA	45	1.47	0.0313	VEGFR2,MET	No	No	No
VANDETANIB_FDA	51	1.47	0.0355	RET/VEGFR2/EGFR	Yes	No	Yes
HG-9-91-01_LINCS	137	1.46	0.0179	SIK1	Yes	Yes	Yes
AZ-628_LINCS	51	1.46	0.0265	BRAF	Yes	No	No
VANDETANIB_KINOME SCAN	51	1.45	0.0448	RET/VEGFR2/EGFR	Yes	No	Yes
EXEL-2880/GSK-1363089_KINOME SCAN	131	1.45	0.0147	MET/AXL/VEGFR2	Yes	No	Yes
PD-173955_KINOME SCAN	105	1.40	0.0305	ABL1/SRC	Yes	No	Yes
BOSUTINIB_LINCS	69	1.40	0.0490	ABL1/SRC	Yes	Yes	Yes
R406_KINOME SCAN	183	1.39	0.0123	SYK,FLT3	Yes	No	No

Enriched in Resistant Group (p < 0.05)

GENE SET NAME	GENE SET SIZE	Normalized Enrichment Score	Nominal p-val	Intended targets	Inhibiting EGFR/ERBB2/ERBB3 based on Kinase Inhibition Assays		
					EGFR	ERBB2	ERBB3
KINOME_858_ROCHE	17	-1.81	0.0041	NA	No	No	No
KINOME_1901_ROCHE	17	-1.67	0.0149	NA	No	No	No
CHEMBL2062936_ROCHE	16	-1.66	0.0042	NA	No	No	No
KINOME_1242_ROCHE	16	-1.62	0.0198	NA	No	No	No
KINOME_1221_ROCHE	19	-1.62	0.0194	NA	No	No	No
KINOME_866_ROCHE	15	-1.56	0.0395	NA	No	No	No
RO-3306_MRC	29	-1.45	0.0455	CDK1	No	No	No

RO-3306 Sensitivity (From GDSC)



Take home message

- Genes don’t act alone to drive biological processes
- Gene set analysis such as GSEA can identify set of coordinately and subtle expressed genes participated in a functional group compared to Candidate Gene Analysis
- *Biology trumps statistics* – if you can validate the gene sets