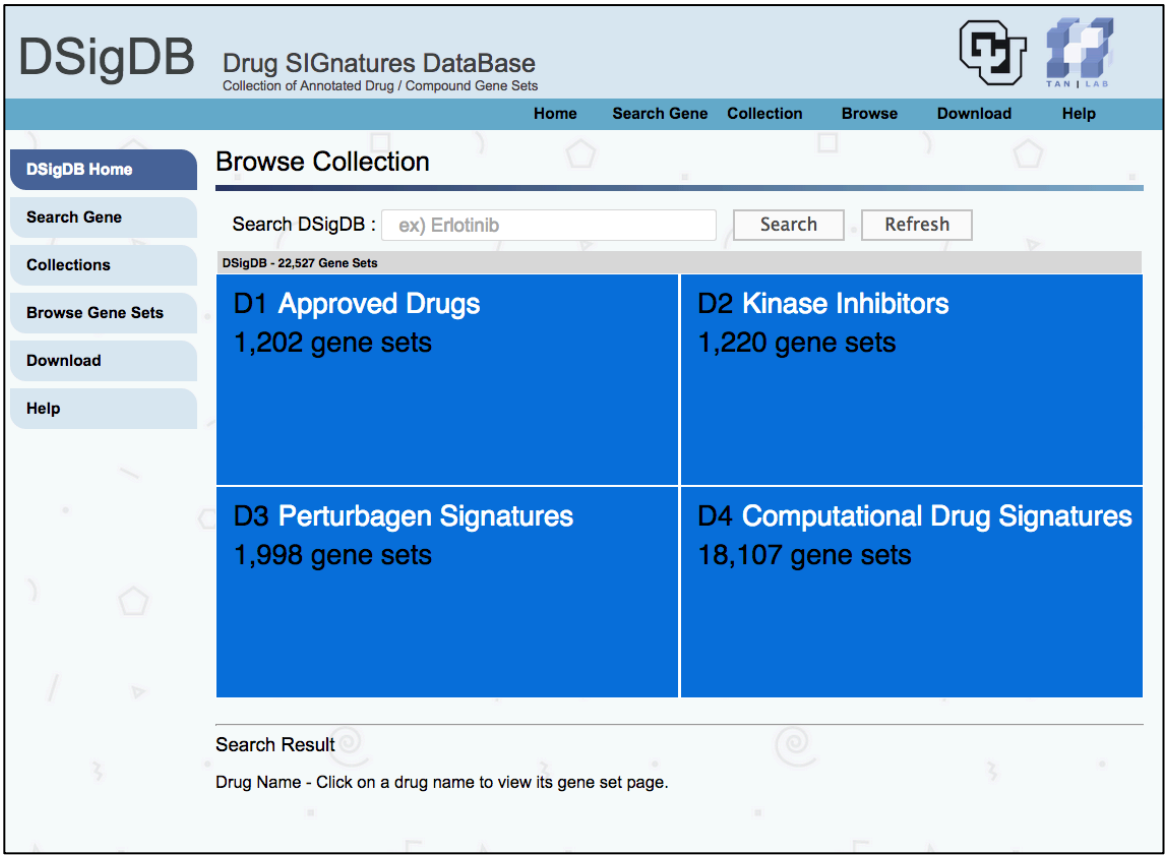
# DSigDB

## Drug Signatures Database
## Online Resource User Manual




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

**Version: 1.0 (May 2015)**

## INTRODUCTION

We report the creation of Drug Signatures Database (DSigDB), a new gene sets resource that relate drugs/compounds and their target genes, for gene set enrichment analysis. DSigDB currently holds 22,527 gene sets, representing 17,389 unique compounds covering 19,531 genes. We also develop 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.

## DEVELOPMENT

DSigDB is developed by the Translational Bioinformatics and Cancer Systems Biology Laboratory, Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus.

## AVAILABILITY

DSigDB is freely accessible: http://tanlab.ucdenver.edu/DSigDB.

## PLEASE CITE DSigDB!

Minjae Yoo, Jimin Shin, Jihye Kim, Karen A. Ryall, Kyubum Lee, Sunwon Lee Jaewoo Kang and Aik Choon Tan. (2015). **DSigDB: Drug Signatures Database for Gene Set Analysis**. *Bioinformatics*. ***In Press***.

## CONTACT

Aik Choon Tan, aikchoon.tan@ucdenver.edu
Minjae Yoo, minjae.yoo@ucdenver.edu

## ACKNOWLEDGEMENT

This work is partly supported by the Cancer League of Colorado, the National Institutes of Health P30CA046934, P50CA058187, and the David F. and Margaret T. Grohne Family Foundation. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the funders.

# TABLE OF CONTENTS


1. GETTING STARTED 4

2. SEARCHING COMPOUND IN DSigDB 5

3. SEARCHING GENE IN DSigDB 6

4. BROWSING DSigDB COLLECTION 8

5. DETAIL GENE SET WEB PAGE 10

6. DSIGDB COLLECTIONS 13

7. DOWNLOAD PAGE 15

8. HELP PAGE 15

# 1. GETTING STARTED

## STARTING POINT

DSigDB (http://tanlab.ucdenver.edu/DSigDB/) is the companion online resource for search, view and download the annotated drug/compound gene sets. Figure 1 is a snapshot of the homepage of the DSigDB online resource.

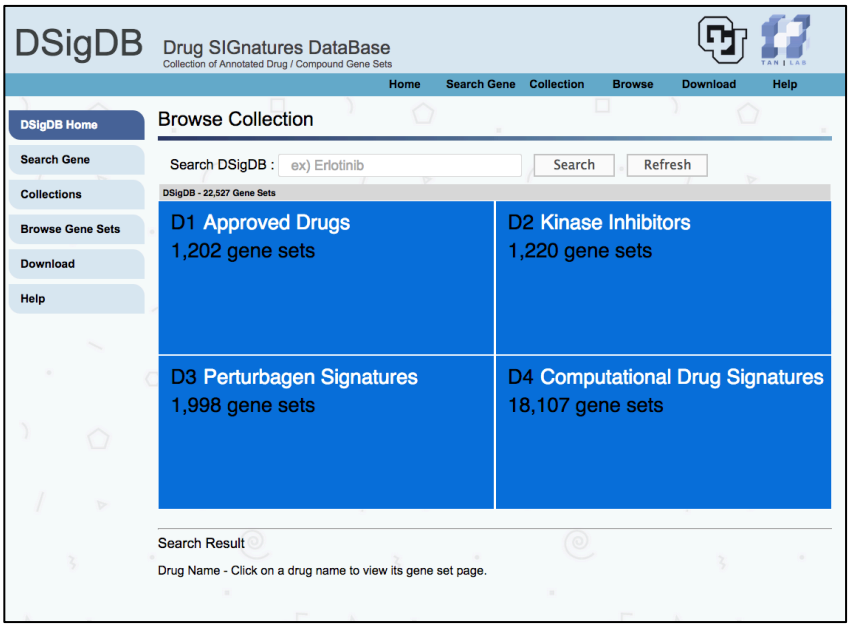



Figure 1: DSigDB Online Resource Homepage.

## ANATOMY OF THE DSIGDB HOMEPAGE

Figure 2 illustrates the anatomy of the DSigDB online resource. The top and left panels represent the menu available in this website. User could search a compound/gene set by key in the name of the compound using the search box. The blue table represents the zoomable table for user to browse the DSigDB collections. The bottom section of the table represents the results page after searching or browsing.

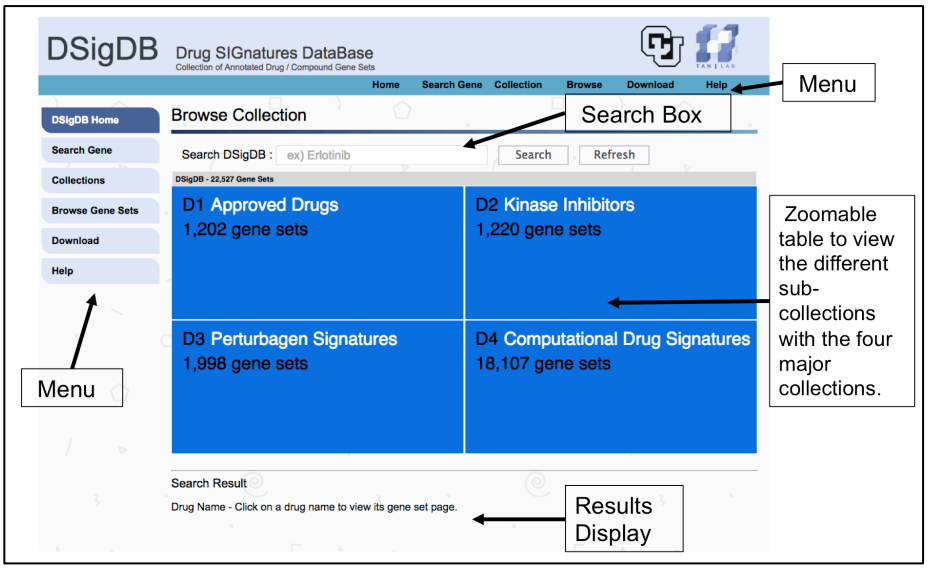



Figure 2: Anatomy of the DSigDB Online Resource Homepage.

## 2. SEARCHING COMPOUND IN DSigDB

To search a compound in the DSigDB, user could enter the name of the compound in the search box. For example, searching the compound “Erlotinib” (Figure 3). Once the name of the compound is entered, press the “Search” button to perform the search. The zoomable table will change from blue to red color, indicating that “Erlotinib” is found in these gene set collections. Figure 3 illustrates that “Erlotinib” is found in D1, D2 and D4 collections. At the bottom of the page, these gene sets are displayed at the results section. Click on the drug name will open a new webpage for the detail gene set in one of the collections. For a given compound query, DSigDB generates an integrated gene set from all the sources (D1 – D4) for download (.gmt and .txt files).

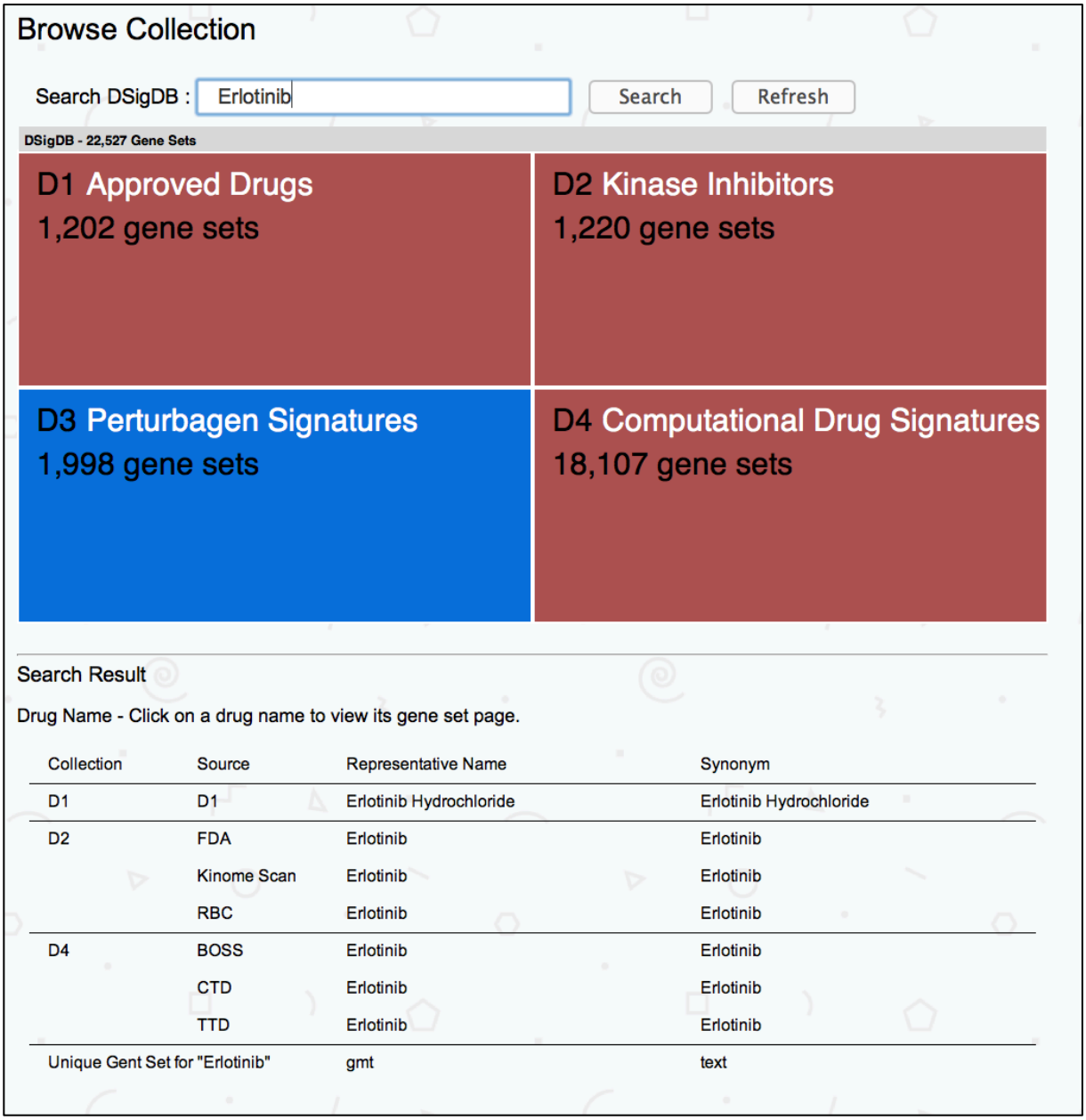



Figure 3: Searching the DSigDB.

## 3. SEARCHING GENE IN DSigDB

To search a gene in the DSigDB, user should click on the “Search Gene” button on the left menu (Figure 4) or on the top menu panels. The “Search Gene” page is illustrated in Figure 5.

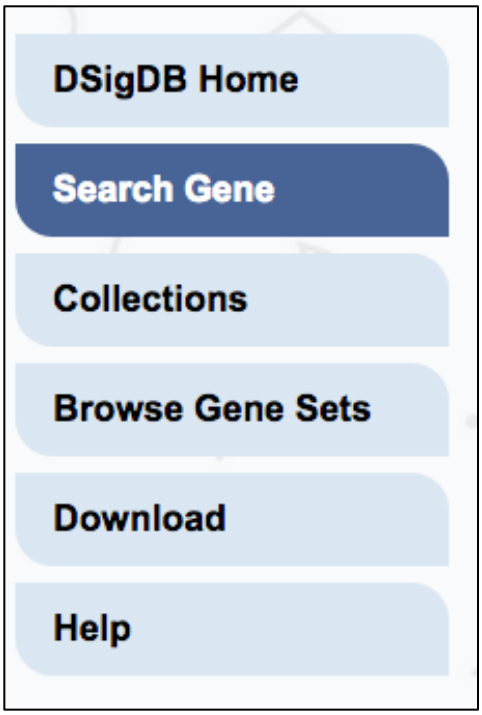



Figure 4: Search Gene option in the Left Menu.

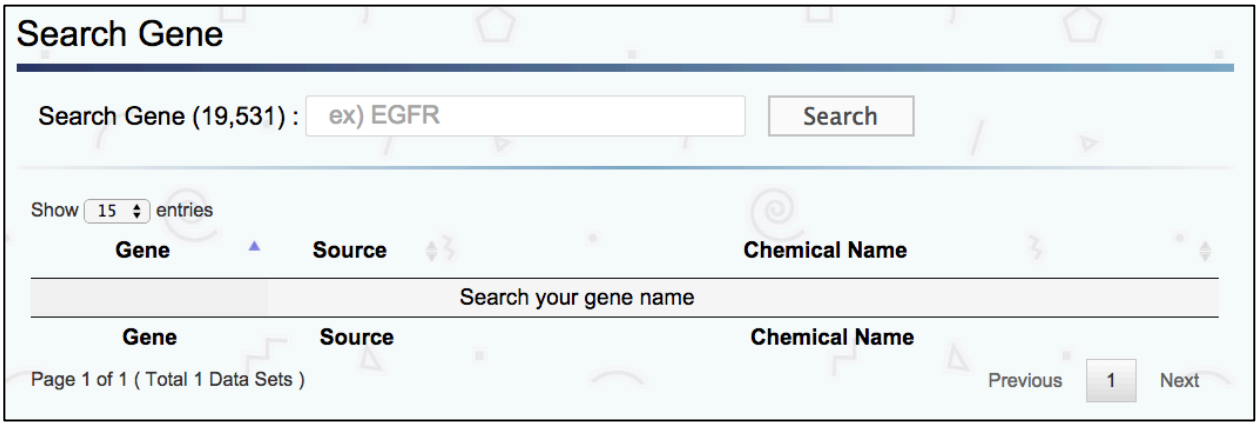



Figure 5: Screenshot of the Search Gene page.

To search for a gene that is related to a gene set in DSigDB, user could enter the official gene symbol of the gene in the search box. For example, searching the gene “EGFR” (Figure 6). Once the name of the gene is entered, press the “Search” button to perform the search. The result will refresh and display below the “Search” box. All the gene sets that contain “EGFR” as a gene member (i.e. compounds that target EGFR) will be displayed. For example, in the “EGFR” search, there are 616 gene sets that have “EGFR” as a gene member (Figure 6). Users could change the option to display the number of results per page, sort the “Source” Or “Chemical Name” by clicking the “arrow” in the results table (Figure 6).

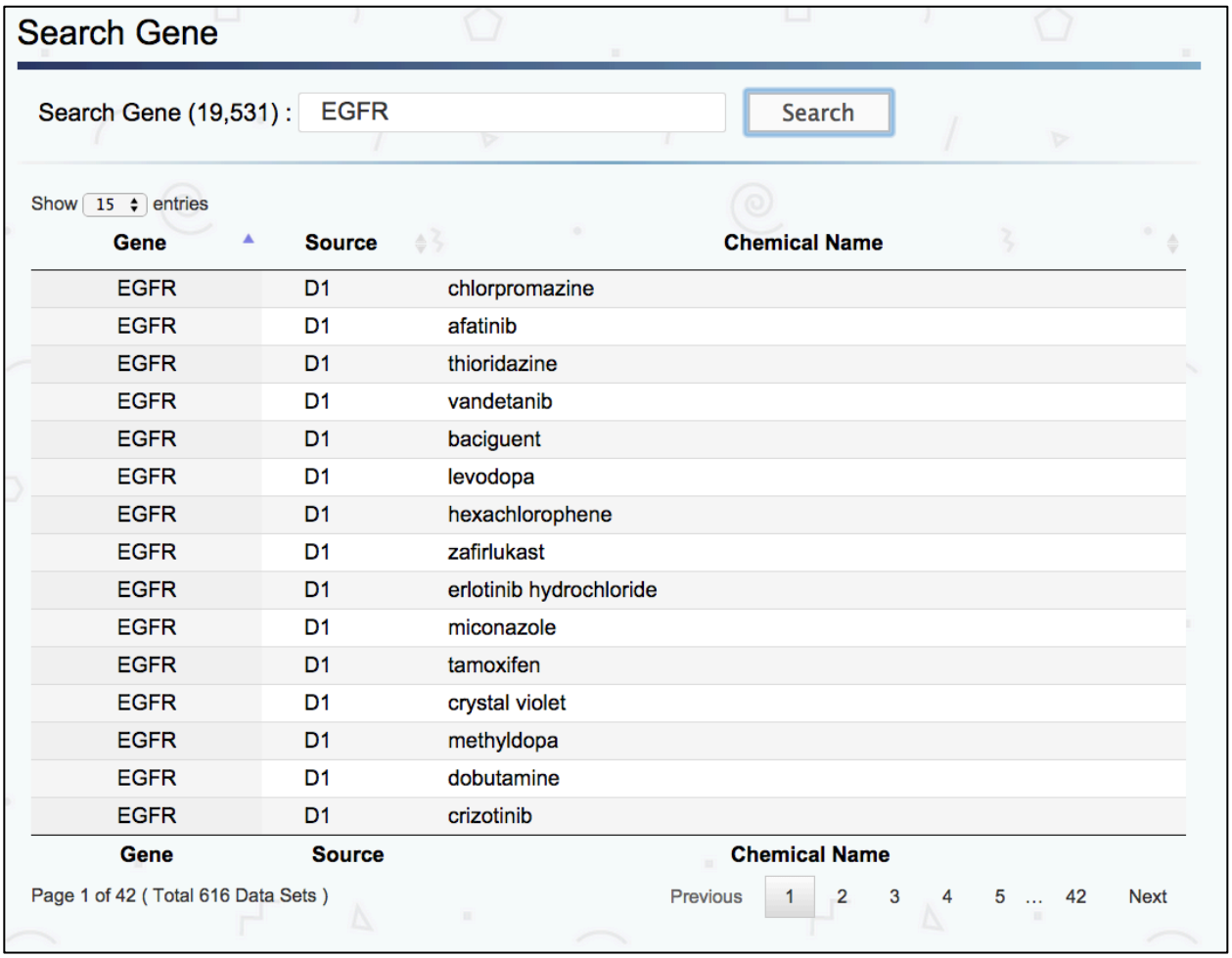


Figure 6. Search results for query “EGFR”.

## 4. BROWSING DSigDB COLLECTION

To browse the DSigDB collection, user may use the “Browse Collection” button on the left menu, or click on the DSigDB zoomable table (the blue square). For example, clicking on the D2: Kinase Inhibitors box (Figure 7A) will zoom in to the sub-collections of D2 (Figure 7B). There are currently seven sub-collections in the D2. To return to the original table, click on the top grey bar (Figure 7B red arrow).

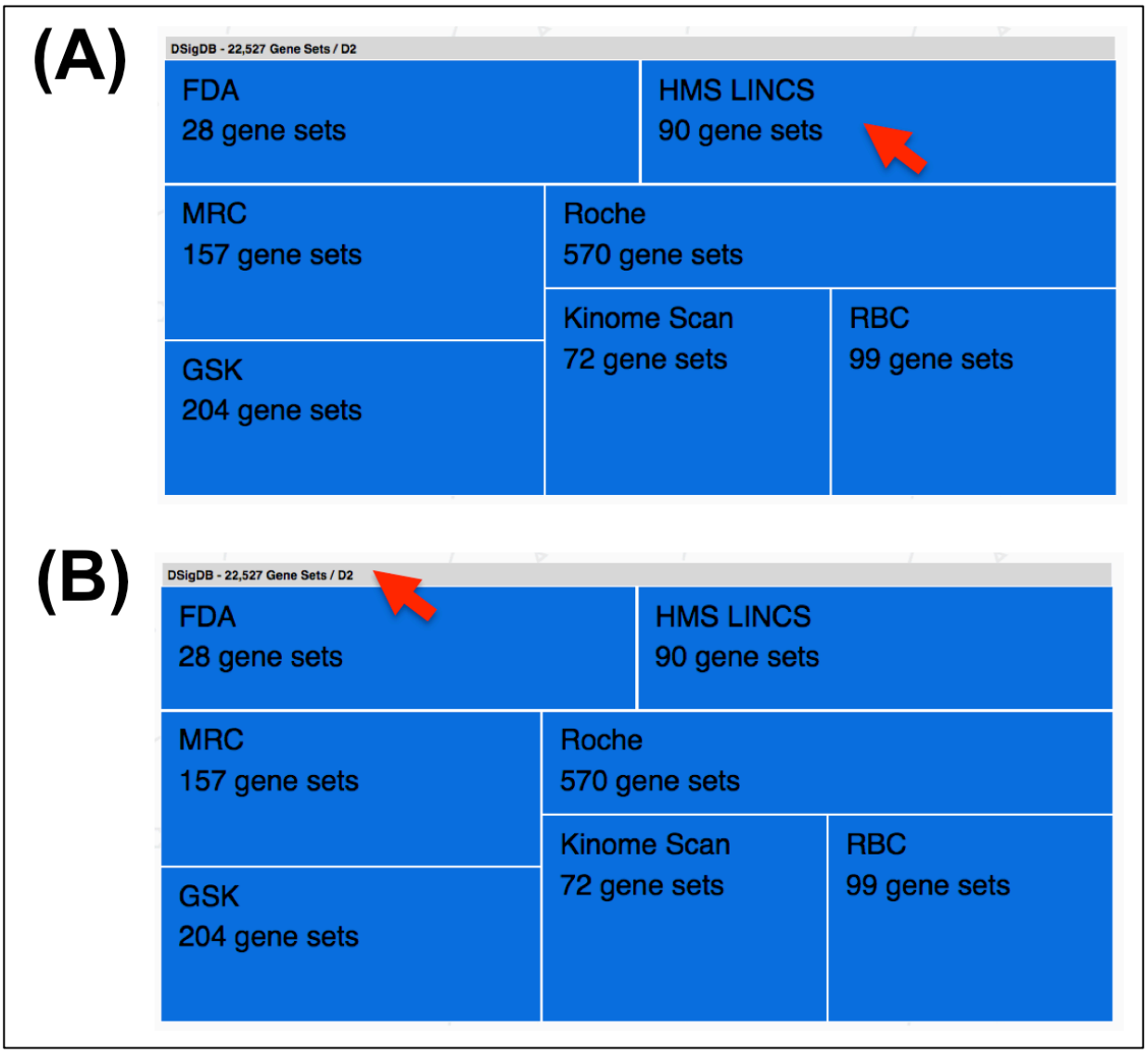



Figure 7: Browsing DSigDB using zoomable table. (A) Zooming D2: Kinase Inhibitors collection by clicking on the square. (B) There are seven sub-collections in the D2: Kinase Inhibitors. Clicking on the top grey bar will zoom out.

User could click on any of the sub-collection box. For example, by clicking the FDA (Figure 8, red arrow) of the D2: Kinase Inhibitors, the results page will list out all the FDA approved compounds that were collected in this sub-collection. Clicking on the drug will open a new window for the detail gene set page.

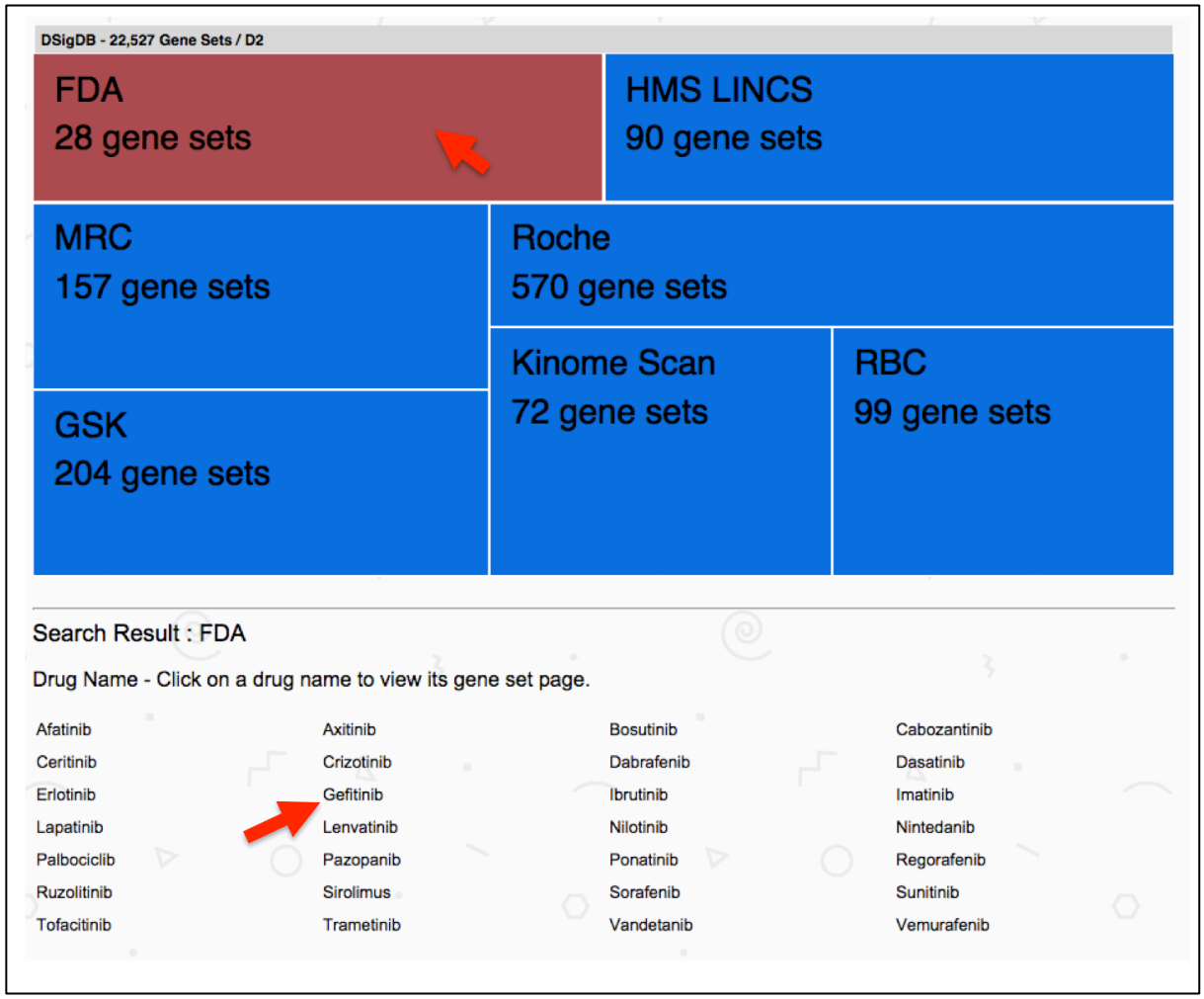



Figure 8: Browsing the FDA approved kinase inhibitors by clicking on the FDA box. At the “Results” section, it lists out the 28 kinase inhibitors and their gene sets available in DSigDB. Click on “Gefitinib” for detail view of the gene set for this drug.

## 5. DETAIL GENE SET WEB PAGE

Each gene set and all of its annotations are presented as an individual web page (Figure 9). Each web page contains four parts: 1) top part describes the clinical development of the compound (approved or clinical trials); 2) middle part indicates the molecular details of the compound including chemical structure (2D and 3D), links to PubChem or ChEMBL; 3) bottom part lists the gene memberships embedded links to source of evidence; 4) download of the gene set. Figure 10 illustrates the anatomy of the individual gene set page. All the external links are embedded in the web page.

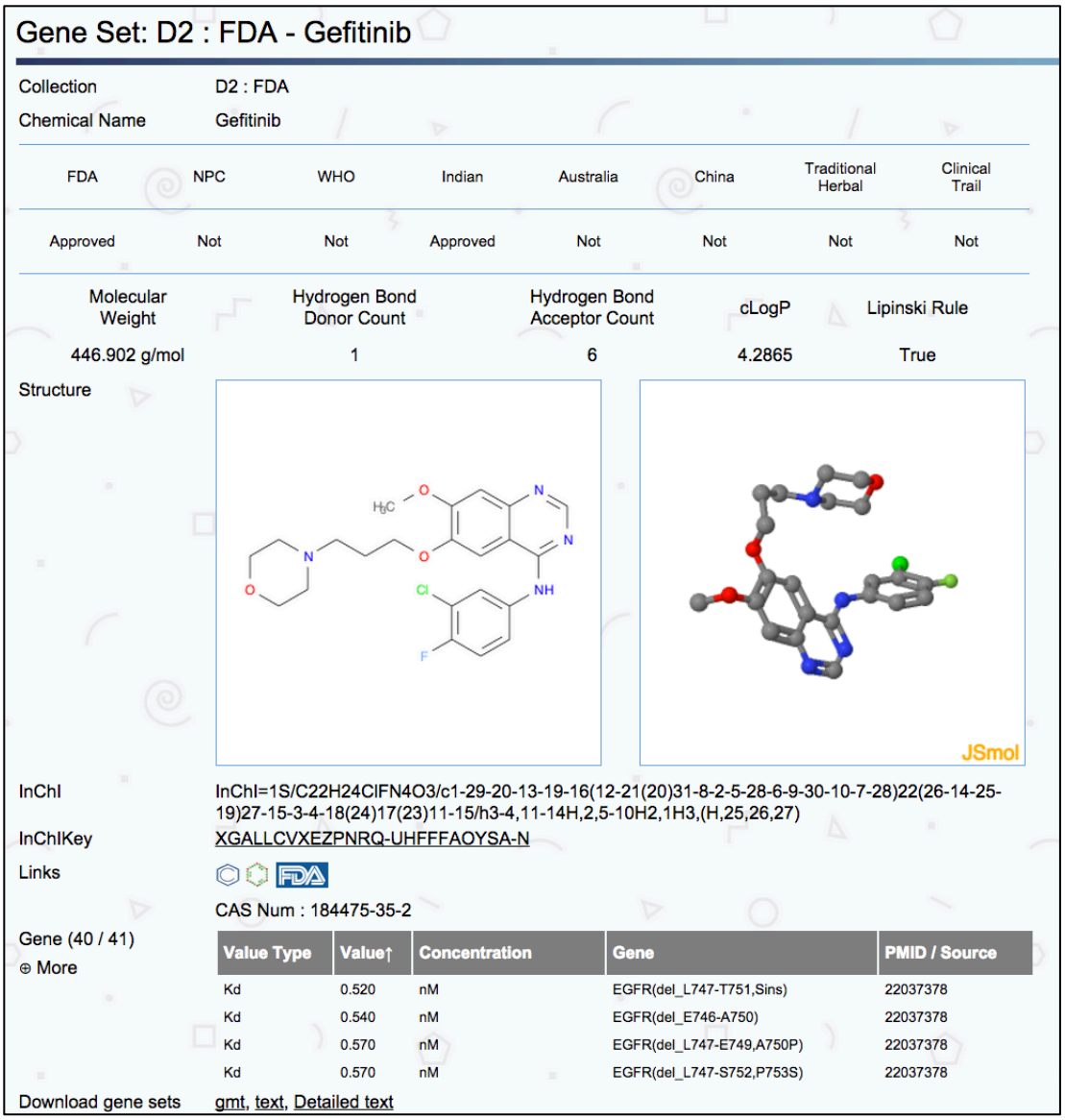



Figure 9: An example of the gene set page.

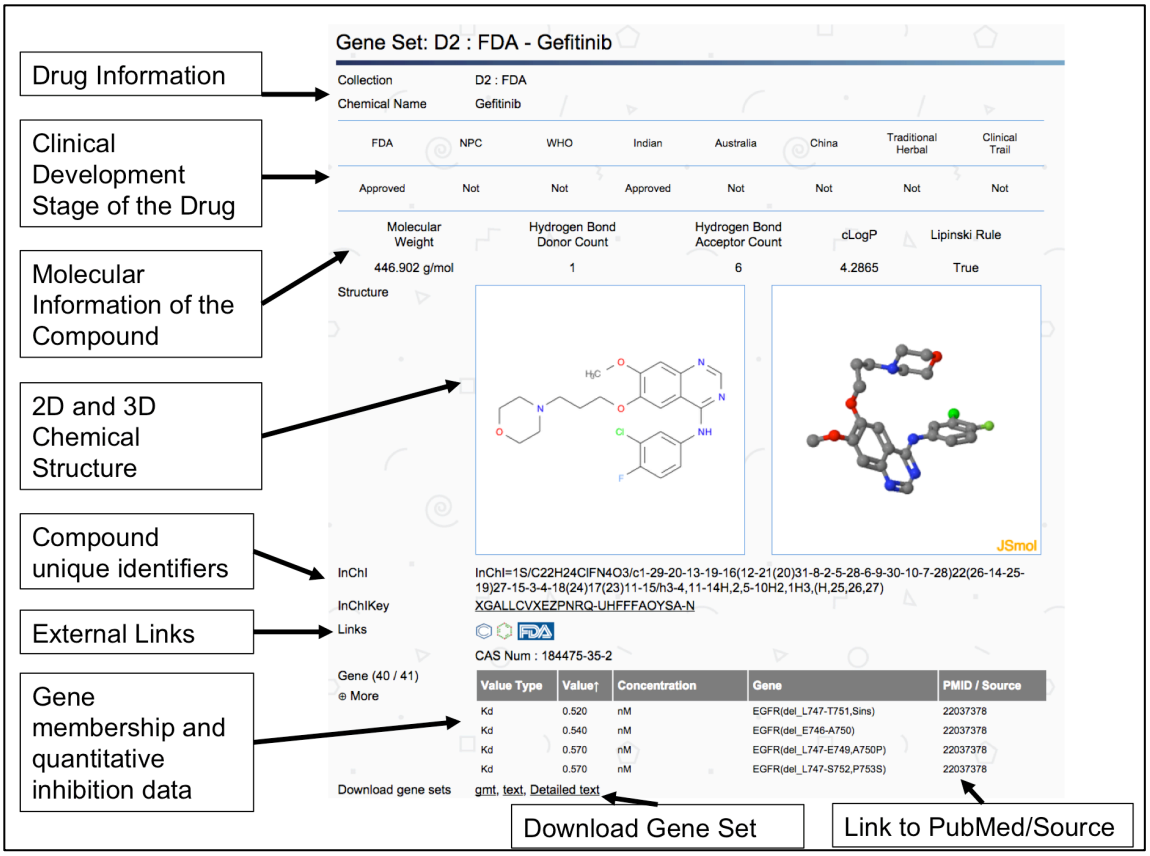



Figure 10: Anatomy of the gene set page.

DSigDB gene sets are available to download as GSEA gene set (.gmt) (Figure 11), plain text (.txt) (Figure 12) or detailed info in text (Detailed.txt)(Figure 13) formats. The .gmt file format can be directly imported into GSEA to execute the program. The gene set results generated from GSEA provides links to DSigDB online resource for detail information about the compounds.

```
Gefitinib	http://tanlab.ucdenver.edu/DSigDB/DSigDBv0.2/displayDrug.py?db=d2_fda&id=1210
	EPHA6	STK10	MKNK1	EGFR	RIPK2	MAP2K5	HIPK4	ABL1	FLT3	CSNK1E	GAK	LYN
	IRAK1	CHEK2	IRAK4	ERBB3	ERBB4	SLK	SBK1	CDK7	MAP3K19	LCK
```

Figure 11: GMT file format – Gefitinib.gmt

The plain text format provides simple listing of gene set membership of the compound. The first line contains the Compound name. The other lines represent the genes involved in this gene set. All genes are represented by their official gene symbol and separated by new line (Figure 12).

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

```
CHEK2
IRAK4
ERBB3
ERBB4
SLK
SBK1
CDK7
MAP3K19
LCK
```

Figure 12: Text file format – Gefitinib.txt

The Detailed text format provides detailed information of the relations between genes and drug. It contains four columns: Drug, Gene, Type and Source as illustrated in Figure 13. Every line represents the relation between drug and gene, the type of interactions (either quantitative binding results or qualitative), and the source of the relation.

```
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
Gefitinib       STK10   Kd=470.0(nM)    FDA
Gefitinib       ERBB4   Kd=410.0(nM)    FDA
Gefitinib       ABL1    Kd=400.0(nM)    FDA
Gefitinib       EGFR    Kd=0.57(nM)     FDA
Gefitinib       FLT3    Kd=1000.0(nM)   FDA
Gefitinib       CSNK1E  Kd=430.0(nM)    FDA
Gefitinib       EGFR    Kd=0.52(nM)     FDA
Gefitinib       EGFR    Kd=0.94(nM)     FDA
Gefitinib       EGFR    Kd=2.0(nM)      FDA
Gefitinib       RIPK2   Kd=530.0(nM)    FDA
Gefitinib       MAP2K5  Kd=600.0(nM)    FDA
Gefitinib       ABL1    Kd=520.0(nM)    FDA
Gefitinib       EGFR    Kd=1.4(nM)      FDA
Gefitinib       HIPK4   Kd=310.0(nM)    FDA
Gefitinib       EGFR    Kd=1.0(nM)      FDA
Gefitinib       EGFR    POC=2.97(0.5uM)         FDA
Gefitinib       MAP3K19 Kd=240.0(nM)    FDA
Gefitinib       EPHA6   Kd=590.0(nM)    FDA
```

Figure 13: Detailed text file format – Gefitinib_detailed.txt

## 6. DSigDB COLLECTIONS

**DSigDB Collections:** DSigDB organized drugs and small molecules related gene sets into four collections based on quantitative inhibition data:

**D1: Approved Drugs.** This collection of gene sets contains 1,202 FDA approved drugs covering 1,288 target genes. We obtained all the approved drugs from US Food and Drug Administration (FDA) website, and retrieved bioactivity data for these drugs from PubChem and ChEMBL. Genes with "active" bioassay results recorded in these databases were compiled as the drug target genes

**D2: Kinase inhibitors.** The human kinome has been a class of intensely pursued drug targets by the pharmaceutical industry. Kinases are frequently mutated in various cancers. Therefore targeting these kinases with small molecules is an attractive therapeutic approach for personalized cancer treatment. This collection of gene sets contains 1,220 kinase inhibitors (1,065 unique kinase inhibitors) covering 407 kinases. We collected large-scale *in vitro* kinase profiling assays from literature and two databases (MRC Kinase Inhibitor database and HMS LINCS database). We considered the kinase a target of a kinase inhibitor if the $IC_{50}/K_d/K_i \leq 1\mu M$ or the Percent of inhibition over Control (POC) ≤ 15% from the assays. These target kinases make up the gene sets for the kinase inhibitors.

**D3: Perturbagen Signatures.** This collection of gene sets was obtained from gene expression profiles induced by compounds. We collected 7,064 gene expression profiles from three cancer cell lines perturbed by 1,309 compounds from CMap (build 02) (Lamb *et al.*, 2006). For each compound, we compared the treated vs. control gene expression profiles for each cell line. Genes with more than 2-fold change from the control were considered as gene sets (either up or down) for that compound. We defined 1,998 gene sets (1,154 unique compounds) covering 11,137 genes in this collection.

**D4: Computational Drug Signatures.** We compiled 18,107 drug signatures extracted from literatures using a mixture of manual curation and text mining approaches. Using manual curation of targets, we compiled 10,830 and 5,163 gene sets from the Therapeutics Targets Database (TTD) (Qin *et al.*, 2014) and the Comparative Toxicogenomics Database (CTD) (Davis *et al.*, 2013), respectively. For the text mining approach, we used the Biomedical Object Search System (BOSS) (Choi *et al.*, 2012) engine to acquire 2,114 co-occurrences of compounds and genes from PubMed abstracts. In addition, we also retrieved genes with "active" bioactivity data for these drugs from PubChem and ChEMBL as in D1. These genes, with quantitative inhibition data, were integrated with the drug signatures obtained from the source to construct the final gene sets for the drug

**Gene set annotations**: Each DSigDB gene set consists of a list of target genes of a compound. The current version of DSigDB focuses on human gene sets. We used human Entrez Gene IDs to serve as universal identifiers to map across different databases. We used InChiKey to serve as the universal compound identifiers to map between PubChem and ChEMBL, and to determine the number of unique compounds within DSigDB.

## DSigDB Collections

DSigDB organized drugs and small molecules related gene sets into four collections based on quantitative inhibition and/or drug-induced gene expression changes data.

| 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 |

Figure 14: Description of the DSigDB collections.

## 7. DOWNLOAD PAGE

We provide three different options to download all the data of DSigDB. Users could download the data from the Download Page. Figure 15 illustrates the screenshot of the DSigDB Download page. The page provides the version (current release is Version 1.0, May 2015), and the three file formats (.gmt, .txt and Detailed.txt) for download.

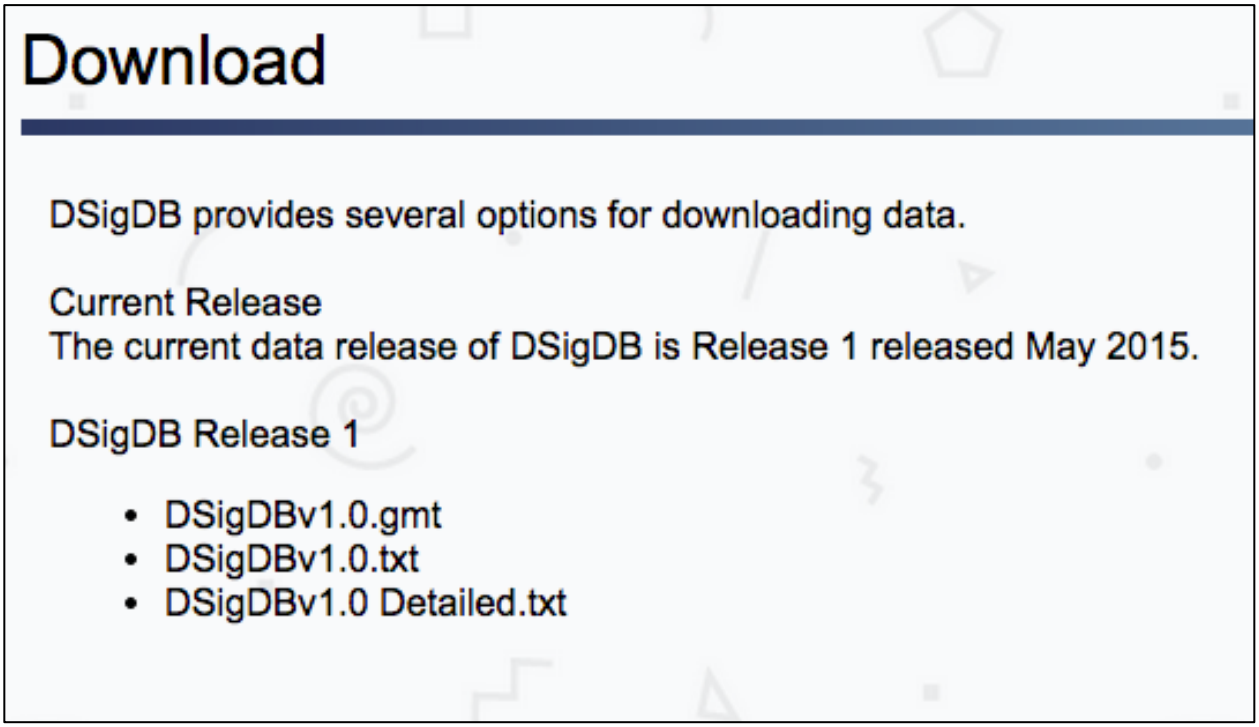



Figure 15: Screenshot of the DSigDB Download Page.

## 8. HELP PAGE

In the Help page, users could download a copy of this DSigDB User Manual. If users need more information, please contact:

Aik Choon Tan, aikchoon.tan@ucdenver.edu
Minjae Yoo, minjae.yoo@ucdenver.edu