

SUPPORTING ONLINE MATERIALS FOR:

SigCom LINCS: Data and Metadata Search Engine for a Million Gene Expression Signatures

John Erol Evangelista^{1,+}, Daniel J.B. Clarke^{1,+}, Zhuorui Xie^{1,+}, Alexander Lachmann^{1,+}, Minji Jeon¹, Kerwin Chen¹, Kathleen M. Jagodnik¹, Sherry L. Jenkins¹, Maxim V. Kuleshov¹, Megan L. Wojciechowicz¹, Stephan Schurer², Mario Medvedovic³, Avi Ma'ayan^{1,*}

¹Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

²Department of Biomedical Informatics, University of Cincinnati College of Medicine, Cincinnati, OH 45267, USA

³Department of Pharmacology, Miller School of Medicine, University of Miami, Miami, USA

+Contributed equally

*To whom correspondence should be addressed: avi.maayan@mssm.edu

SUPPORTING FIGURES

Fig. S1

The figure consists of two side-by-side panels, each representing a validation step in a software interface. Both panels have a 'Validate' button at the top left.

Left Panel:

- Validation status: 1 valid entry, 1 needs review.
- Input: A green pill-shaped button containing the variant 'chr13:g.32944582T>C' with 'x' icons on either side.
- Action: 'RESOLVE VARIANT NAME TO:'
- Result: 'BRCA2' is displayed in blue text.
- Output: A grey pill-shaped button containing 'STAT3' with a checkmark on the left and an 'x' icon on the right.

Right Panel:

- Validation status: 1 invalid entry, 1 needs review.
- Input: A yellow pill-shaped button containing 'S' with a warning icon on the left and an 'x' icon on the right.
- Action: 'DID YOU MEAN:'
- Result: 'CDSN' is displayed in blue text.
- Output: A red pill-shaped button containing 'SATS' with 'x' icons on either side.

Fig. S1 Gene Names Validation. SigCom LINCS validates user-inputted gene names against all the gene names registered in the SigCom LINCS metadata. The validation function provides an option for users to resolve synonymous IDs and gene names that do not have an exact match with the registered name. This function facilitates harmonization of the input sets with the gene metadata within SigCom LINCS.

Fig. S2

A

Identify reversers and mimickers from over 1.5 million signatures by entering up and down gene sets or by performing single gene search for co-expressed genes using the input form at the bottom.

Input a set of newline-separated Entrez gene symbols (up gene set).

Input a set of newline-separated Entrez gene symbols (down gene set).

MAPK1
MAPK10
MAPK11
MAPK12
MAPK13
MAPK14
MAPK15
MAPK1IP1L

STAT3 / MAPK1 / ACE2 / rs28897756 / chr13:g.32944582T>C

Identify reversers and mimickers from over 1.5 million signatures by entering up and down gene sets or by performing single gene search for co-expressed genes using the input form at the bottom.

Validate 97 valid entries 3 invalid entries

CRKL
CAB39
FA2
STXB5
MTPN
PPP2R1
ZBTB34
MAP3K4
FAM6A1
OSBP1B
MEGF9
SYNJ1

Validate 97 valid entries 3 invalid entries

EIF2B4
WBSCR22
RDM1
IFT43
HOCR2
POU2H
SHFM1
SICCSA1
TEC8
SLBP
MTG1
ACY1

Single Gene Set Up/Down Gene Sets

Try this example up and down gene set

MAPK1

Query a single gene or a variant to identify other genes that mostly positively or negatively correlate with the queried gene based on RNA-seq co-expression data from ARCHS4:

Enter gene

STAT3 / MAPK1 / ACE2 / rs28897756 / chr13:g.32944582T>C

B

Identify perturbations from over 1.5 million signatures that maximally up- or down-regulate the expression of a gene set. Enter a set of genes or fetch a gene set using the term search form at the bottom.

Input a set of newline-separated Entrez gene symbols.

GO Molecular Function 2017b

transmembrane receptor protein **tyrosine kinase** activity (GO:0004714)

protein **tyrosine kinase** activator activity (GO:0030296)

non-membrane spanning protein **tyrosine kinase** activity (GO:0004715)

protein serine/threonine/**tyrosine kinase** activity (GO:0004712)

protein **tyrosine kinase** inhibitor activity (GO:0030292)

GO Biological Process 2021

tyrosine kinase

breast cancer / tyrosine kinase

Identify perturbations from over 1.5 million signatures that maximally up- or down-regulate the expression of a gene set. Enter a set of genes or fetch a gene set using the term search form at the bottom.

Validate 114 valid entries

SCYL1
KDR
FGFR4
ETC
INSRR
DYRK1A
STAT5A
EPHA2
BLK
SYK
FGR
FES

Single Gene Set Up/Down Gene Sets

Try this example gene set

GO Molecular Function 2017b: transmembrane recep

Fetch annotated gene sets from Enrichr by querying any search term:

Enter search term

breast cancer / tyrosine kinase

Fig. S2 Fetching signature search on co-expression and Enrichr gene sets. SigCom LINCS can retrieve gene sets from other resources by (A) converting a single gene or a variant to up and down gene sets using ARCHS4 co-expression data, and (B) fetching Enrichr gene sets via its search term search API.

Fig. S3

Signature Similarity Search Results

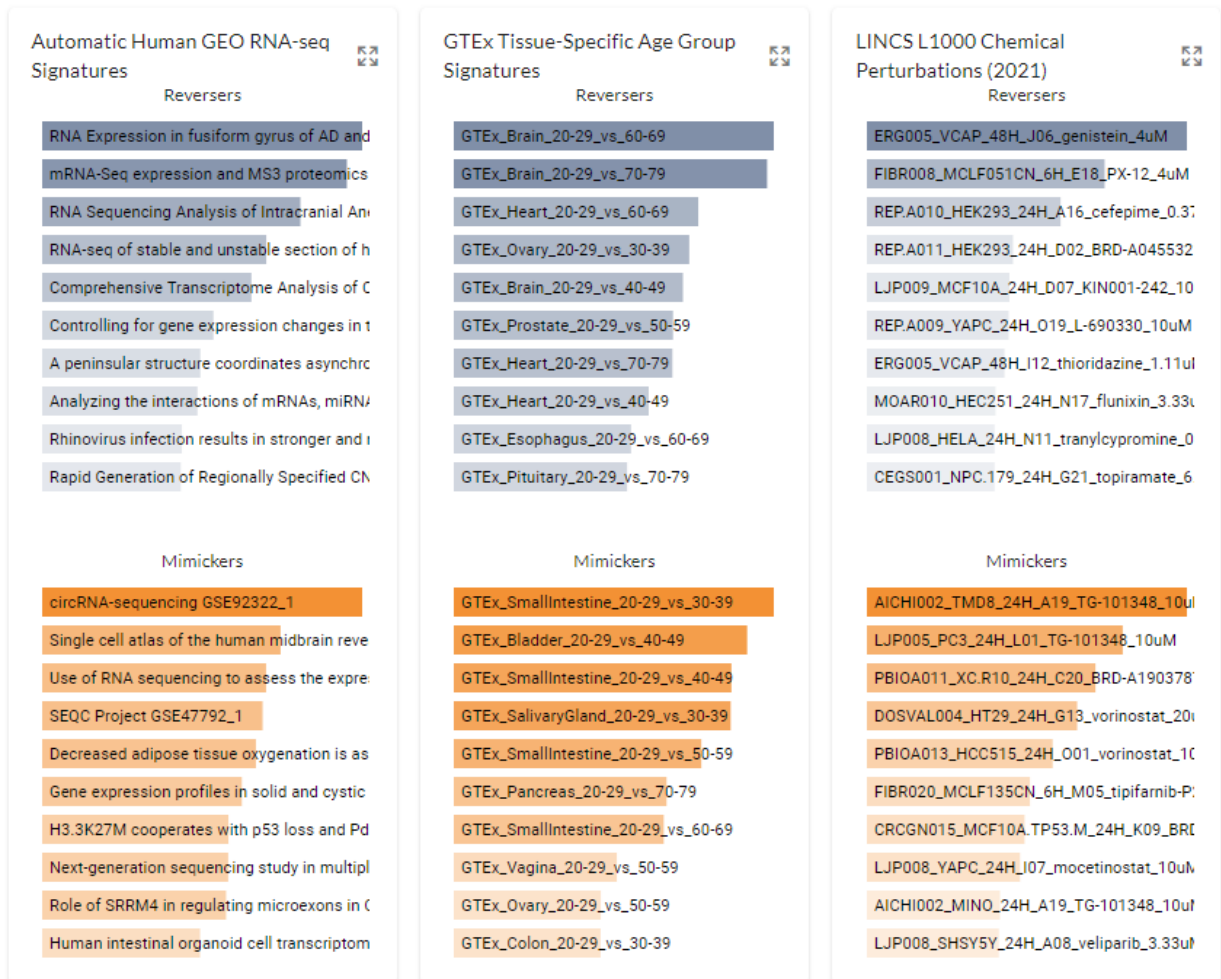


Fig. S3 SigCom LINCS search results overview. Once the user presses submit, the initial results page displays bar charts for each collection of signatures.

Fig. S4

Reversers									
FILTERS EXPORT TABLE EXPORT GENE RANK MATRIX									
<input type="checkbox"/>	Perturb...	Dose	Tissue	Cell Line	Timepoi...	log p-val...	log p-val...	z-score (...)	p-value ...
<input type="checkbox"/>	genistein	4 uM	prostate gland	VCAP	48 h	31.21	12.87	-10.39	1
<input type="checkbox"/>	PX-12	4 uM		MCLF051CN	6 h	26.34	12.34	-9.509	1
<input type="checkbox"/>	cefepime	0.37 uM	kidney	HEK293	24 h	25.00	8.298	-9.036	1.9386e-2
<input type="checkbox"/>	BRD-A04553...	3.33 uM	kidney	HEK293	24 h	21.69	10.21	-8.525	1
<input type="checkbox"/>	KIN001-242	10 uM	breast	MCF10A	24 h	21.66	9.392	-8.486	1
<input type="checkbox"/>	L-690330	10 uM	pancreas	YAPC	24 h	21.46	10.29	-8.478	1
<input type="checkbox"/>	thioridazine	1.11 uM	prostate gland	VCAP	48 h	21.23	10.36	-8.433	1
<input type="checkbox"/>	flunixin	3.33 uM	endometrium	HEC251	24 h	21.00	9.077	-8.337	1
<input type="checkbox"/>	tranlycypromi...	0.12 uM	uterine cervix	HELA	24 h	20.88	9.427	-8.327	1

Mimickers									
FILTERS EXPORT TABLE EXPORT GENE RANK MATRIX									
<input type="checkbox"/>	Perturb...	Dose	Tissue	Cell Line	Timepoi...	log p-val...	log p-val...	z-score (...)	p-value ...
<input type="checkbox"/>	TG-101348	10 uM	lymphoid syst...	TMD8	24 h	28.28	13.78	9.905	1
<input type="checkbox"/>	TG-101348	10 uM	prostate gland	PC3	24 h	25.60	12.56	9.371	1
<input type="checkbox"/>	BRD-A19037...	10 uM	epithelium	XCR10	24 h	25.52	8.489	9.145	1
<input type="checkbox"/>	vorinostat	20 uM	intestine	HT29	24 h	24.38	9.107	8.993	8.0642e-2
<input type="checkbox"/>	vorinostat	10 uM	lung	HCC515	24 h	23.41	8.782	8.789	1.5429e-1
<input type="checkbox"/>	tipifarnib-P2	4 uM		MCLF135CN	6 h	21.96	10.95	8.601	1
<input type="checkbox"/>	BRD-A19037...	1.11 uM	breast	MCF10A	24 h	22.32	8.401	8.556	1
<input type="checkbox"/>	mocetinostat	10 uM	pancreas	YAPC	24 h	25.14	4.612	8.518	4.2123e-4
<input type="checkbox"/>	TG-101348	10 uM	blood	MINO	24 h	20.99	9.455	8.352	1

Fig. S4 Tabular view of the top signatures. Tabular views of the top signatures are provided for users to inspect the metadata and scores of the results. These tables are downloadable and are provided to the users in TSV format.

Fig. S6

Query SigCom LINCS signatures with any search term

dexamethasone

Example queries: dexamethasone / MAPK1 / STAT3 / blood

Signature Search

Genes Datasets Signatures (652)



 CPC006_A549_24H_F09_dexamethasone_10uM LINCS L1000 Chemical Perturbations (2021) Perturbagen: dexamethasone Perturbation Type: Chemical Dose: 10 uM Tissue: lung Cell Line: A549 Disease: lung cancer Timepoint: 24 h Data Level: 5 Creation Time: 2021-05-08	<ul style="list-style-type: none">Data and Signature Generation CenterCell LineDataset
 CPC004_A375_6H_E13_dexamethasone_10uM LINCS L1000 Chemical Perturbations (2021) Perturbagen: dexamethasone Perturbation Type: Chemical Dose: 10 uM Tissue: skin of body Cell Line: A375 Disease: melanoma Timepoint: 6 h Data Level: 5 Creation Time: 2021-05-09	<ul style="list-style-type: none"><input checked="" type="checkbox"/> LINCS L1000 Chemical Perturbations (2021) (652)<input type="checkbox"/> Small Molecules (213)<input type="checkbox"/> L1000 Dataset -small molecule, CRISPR perturbagens- LINCS Phase 2 (December 2015) (99)<input type="checkbox"/> L1000 Dataset -small molecule perturbagens- LINCS Phase 1 (94)

Fig. S6 Metadata search results. The SigCom LINCS metadata search interface enables users to query the SigCom LINCS database for retrieving signatures using any search term, for example, a cell line, a drug, or a gene. Once the search is complete, filters can be used to narrow down the search.

Fig. S7

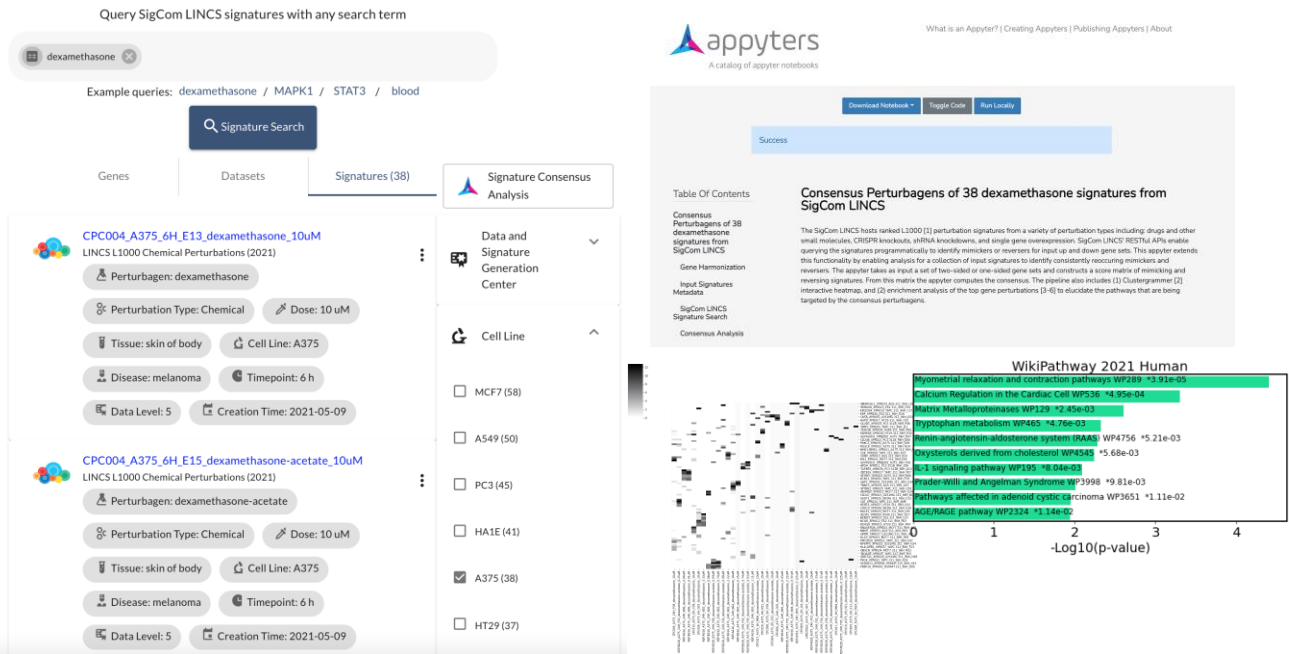



Fig. S7 Consensus analysis. SigCom LINCS provides an option to perform consensus analysis for a collection of signatures within the SigCom LINCS database. Metadata search queries that return less than 50 signature can be piped for analysis with the Signature Consensus Appyter. The SigCom LINCS Consensus Appyter produced a report that provides insights about the common mimickers and reversers across the collection of input signatures.

Fig. S8

The image shows a user interface for accessing signature data. On the left, a card displays the signature ID 'CPC014_A375_6H_H21_dexamethasone_10uM' and 'LINCS L1000 Chemical Perturbations (2021)'. Below this, several metadata tags are shown in rounded rectangles: 'Perturbagen: dexamethasone', 'Perturbation Type: Chemical', 'Dose: 10 uM', 'Tissue: skin of body', 'Cell Line: A375', 'Disease: melanoma', 'Timepoint: 6 h', 'Data Level: 5', and 'Creation Time: 2021-05-14'. On the right, a white box with a shadow contains a list of actions: 'Perform Signature Search' (with a magnifying glass icon), 'Download Up and Down Genes' (with a downward arrow icon), 'Download Full Signature' (with a downward arrow icon), 'Top 100 Up to Enrichr' (with a red upward arrow icon), and 'Top 100 Down to Enrichr' (with a red downward arrow icon).


Fig. S8 Accessing signature data. Individual signatures can be downloaded as a JSON file or a GMT file that contain the up and down gene sets for the signature. Users can also use the signature data for SigCom LINCS signature search. Furthermore, the up or down gene sets can be sent to Enrichr for enrichment analysis.

Fig. S9

	CPC006_A549_24H_F09_dexamethasone_10uM LINCS L1000 Chemical Perturbations (2021)
md5	d47b5c01a7925cf80084b7828a2bd9d4
doid	DOID:1324
sha256	3ea8d45e3f143b663144d0733b17df8c51051a1eeffab45e85140191463e963a
tissue	lung
anatomy	UBERON:0002048
cmap id	CPC006_A549_24H:BRD-A35108200-001-04-7:10
disease	lung cancer
version	1
filename	L1000_LINCS_DCIC_CPC006_A549_24H_F09_dexamethasone_10uM.tsv
local id	CPC006_A549_24H_F09_dexamethasone_10uM
cell line	A549
pert dose	10 uM
pert name	dexamethasone
pert time	24 h
pert type	Chemical
data level	5
creation time	2021-05-08
persistent id	https://lincs-dcic.s3.amazonaws.com/LINCS-sigs-2021/cd/cp/L1000_LINCS_DCIC_CPC006_A549_24H_F09_dexamethasone_10uM.tsv
size in bytes	345398
uncompressed size in bytes	345398

up down

CPC006_A549_24H_F09_dexamethasone_10uM has 100 up genes.

 **S100A9**
S100 calcium binding protein A9
Gene ID: 6280 Taxon ID: 9606
Synonyms: 60B8AG, CAGB, CFAG, CGLB, L1AG, LIAG, MAC387, MIF, MRP14, NIF, P14


 **NEAT1**
nuclear paraspeckle assembly transcript 1 (non-protein coding)
Gene ID: 283131 Taxon ID: 9606
Synonyms: LINC00084, NCRNA00084, TncRNA, VINC

Fig. S9 Metadata pages Clicking on metadata search results opens the metadata landing page which features the expanded metadata of the entry.

Fig. S10

A.

Find Signatures that up- or down-regulate a single gene

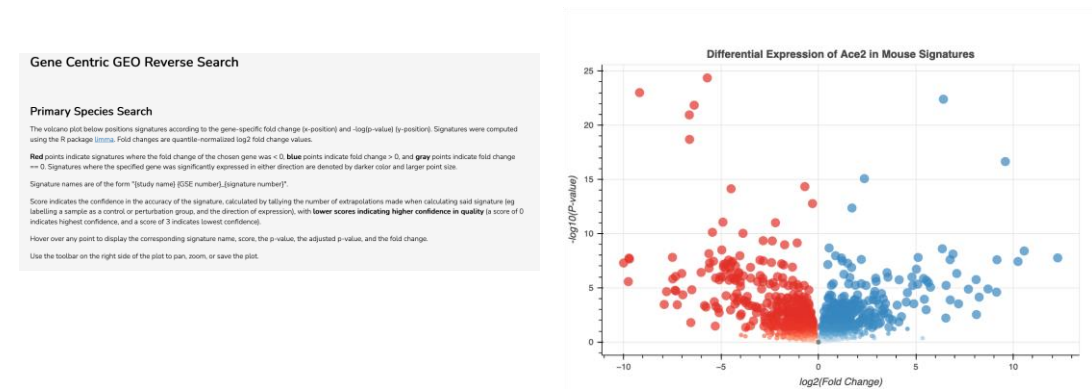
🔍 Query single genes to receive signatures that maximally up- or down-regulate its expression

Example queries: ACE2 / STAT3 / MAPK1

GEO Reverse Search Appyter RNA-seq-like Reverse Search Appyter

🔍 Search

B.



C.

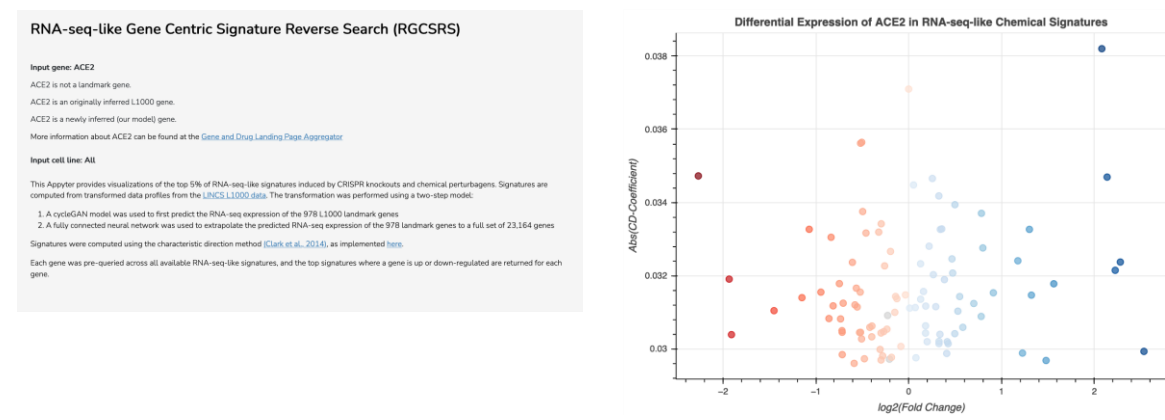


Fig. S10 Single gene reverse search Appyters A) The SigCom LINCS gene search page can be used to perform reverse signature search for a single human gene. The search identifies signatures that maximally up- or down-regulate the gene. Performing a gene search redirects the users to one of Appyters. B) The identified automated GEO signatures, and C) the RNA-seq-like signatures from L1000 signatures are visualized as volcano plots.

Fig. S11

RNA-seq Transcriptomics Signatures (Level 5)

label	date	size	shape	file type	downloads	download
Automatic Human GEO RNA-seq Signatures	2021-09-16	1.42GB	4269x35239	tsv	4	
Automatic Mouse GEO RNA-seq Signatures	2021-09-16	1.49GB	4216x32545	tsv	0	
GTEx Tissue Age Comparison Signatures	2021-09-16	123.2MB	135xVaries	tsvZip	1	

RNA-seq Transcriptomics Gene Sets (Level 5)

label	date	size	terms	file type	downloads	download
Automatic Human GEO RNA-seq Signatures	2021-09-20	7.6MB	8538	gmt	3	
Automatic Mouse GEO RNA-seq Signatures	2021-09-20	6.9MB	8432	gmt	1	
GTEx Tissue Age Comparison Signatures	2021-09-20	199KB	270	gmt	2	
CREEDS Manual Disease Signatures	2021-10-04	3.2MB	1656	gmt	0	
CREEDS Manual Single Drug Perturbations	2021-10-04	3.4MB	1750	gmt	0	
CREEDS Manual Single Gene Perturbations	2021-10-04	8.4MB	4352	gmt	0	
SARS-CoV-2 Differential Gene Expression Signatures	2021-10-05	846KB	248	gmt	0	

L1000 Characteristic Direction Coefficient Tables (Level 5)

All L1000 data that is provided here are derivative work that is significantly different from what is provided at cloud.lincs.org, the original source of these data. The original data and signatures profiled and computed by the LINCS Broad Transcriptomics Data and Signature Generation Center (DSGC) are available from cloud.lincs.org.

label	date	size	shape	file type	downloads	download
LINC15 L1000 Chemical Perturbations (2021)	2021-06-23	33.3GB	12327x720216	gctx	2	
LINC15 L1000 shRNA Perturbations (2021)	2021-06-23	7.3GB	12327x177263	gctx	2	
LINC15 L1000 CRISPR Perturbations (2021)	2021-06-23	6.5GB	12327x141368	gctx	0	
LINC15 L1000 Overexpression Perturbations (2021)	2021-06-23	1.6GB	12327x34171	gctx	0	
LINC15 L1000 Ligand Perturbations (2021)	2021-06-23	357MB	12327x7546	gctx	2	
LINC15 L1000 Antibody Perturbations (2021)	2021-06-23	27.7MB	12327x575	gctx	3	
LINC15 L1000 siRNA Perturbations (2021)	2021-06-23	9.6MB	12327x162	gctx	2	

Computed Signatures for CycleGAN Predicted RNA-Seq-Like Profiles of L1000 Samples (Level 5)

All L1000 data that is provided here are derivative work that is significantly different from what is provided at cloud.lincs.org, the original source of these data. The original data and signatures profiled and computed by the LINCS Broad Transcriptomics Data and Signature Generation Center (DSGC) are available from cloud.lincs.org.

label	date	size	shape	file type	downloads	download
LINC15 L1000 Chemical Perturbations (2021)	2021-10-11	68.07GB	718055x23614	gctx	0	
LINC15 L1000 shRNA Perturbations (2021)	2021-10-11	14.97GB	158003x23614	gctx	2	
LINC15 L1000 CRISPR Perturbations (2021)	2021-10-11	13.35GB	148945x23614	gctx	1	
LINC15 L1000 Overexpression Perturbations (2021)	2021-10-11	3.24GB	34364x23614	gctx	0	
LINC15 L1000 Ligand Perturbations (2021)	2021-10-11	716.0MB	7546x23614	gctx	0	
LINC15 L1000 Antibody Perturbations (2021)	2021-10-11	55.49MB	575x23614	gctx	0	
LINC15 L1000 siRNA Perturbations (2021)	2021-10-11	16.33MB	162x23614	gctx	0	

LINC15 Small Molecules Metadata

label	date	size	file type	downloads	download
LINC15 Small Molecules Metadata	2021-09-20	3.7MB	tsv	7	

Other Data Packages

Search: Datasets (411) Signatures

ASSAY CENTER DATA FILE SIZE

- MGH (CMT) Growth Inhibition Assay Protocol (9 compound doses: 8 uM to 0.03125 uM) - DNA Staining
MGH Center for Molecular Therapeutics (Massachusetts General Hospital)
Assay: Fluorescence Imaging cell growth inhibition assay
Date: 2014-09-28
File Size: 345.30 KB
Downloads: 0
- Eriolab KINOMEscan
HMS LINC15 (Harvard Medical School)
Assay: LINC15 KinomeScan kinase small molecule binding assay
Date: 2013-10-16
File Size: 12.44 KB
Downloads: 0
- MGH (CMT) Growth Inhibition Assay Protocol (3 compound doses) - DNA Staining (Dm)
MGH Center for Molecular Therapeutics (Massachusetts General Hospital)
Assay: Fluorescence Imaging cell growth inhibition assay
Date: 2014-09-28
File Size: 288.44 KB
Downloads: 0

L1000 Characteristic Direction Up and Down Gene Sets (Level 5)

All L1000 data that is provided here are derivative work that is significantly different from what is provided at cloud.lincs.org, the original source of these data. The original data and signatures profiled and computed by the LINCS Broad Transcriptomics Data and Signature Generation Center (DSGC) are available from cloud.lincs.org.

label	date	size	terms	file type	downloads	download
LINC15 L1000 Chemical Perturbations (2021)	2021-07-01	2.25GB	1436110	gmt	5	
LINC15 L1000 shRNA Perturbations (2021)	2021-07-01	501.7MB	316006	gmt	1	
LINC15 L1000 CRISPR Perturbations (2021)	2021-07-01	438.4MB	281890	gmt	2	
LINC15 L1000 Overexpression Perturbations (2021)	2021-07-01	106MB	68238	gmt	0	
LINC15 L1000 Ligand Perturbations (2021)	2021-07-01	23.6MB	15092	gmt	2	
LINC15 L1000 Antibody Perturbations (2021)	2021-07-01	1.8MB	1150	gmt	2	
LINC15 L1000 siRNA Perturbations (2021)	2021-07-01	506KB	324	gmt	2	

CycleGAN Predicted RNA-Seq-Like Profiles of L1000 Samples (Level 3)

All L1000 data that is provided here are derivative work that is significantly different from what is provided at cloud.lincs.org, the original source of these data. The original data and signatures profiled and computed by the LINCS Broad Transcriptomics Data and Signature Generation Center (DSGC) are available from cloud.lincs.org.

label	date	size	shape	file type	downloads	download
LINC15 L1000 Chemical Perturbations (2021)	2021-08-16	170.8GB	1805898x23614	gctx	0	
LINC15 L1000 shRNA Perturbations (2021)	2021-08-16	42.86GB	453175x23614	gctx	2	
LINC15 L1000 CRISPR Perturbations (2021)	2021-08-16	39.23GB	414816x23614	gctx	2	
LINC15 L1000 Controls (2021)	2021-11-30	14.07GB	148678x23614	gctx	1	
LINC15 L1000 Overexpression Perturbations (2021)	2021-08-16	12.46GB	131668x23614	gctx	0	
LINC15 L1000 Ligand Perturbations (2021)	2021-08-16	2.3GB	24301x23614	gctx	0	
LINC15 L1000 Antibody Perturbations (2021)	2021-08-16	158.4MB	1655x23614	gctx	1	
LINC15 L1000 siRNA Perturbations (2021)	2021-08-16	46.4MB	472x23614	gctx	2	

Fig. S11 The SigCom LINCS download page. All links to the Level 5 full signatures data, Level 3 expression profiles, as well as GMT files of the up and down gene sets are accessible on the download page. Users are also provided links to the predicted RNA-seq-like profiles from L1000 samples as well as legacy LINCS datasets from the original LINCS data portal.

Fig. S12

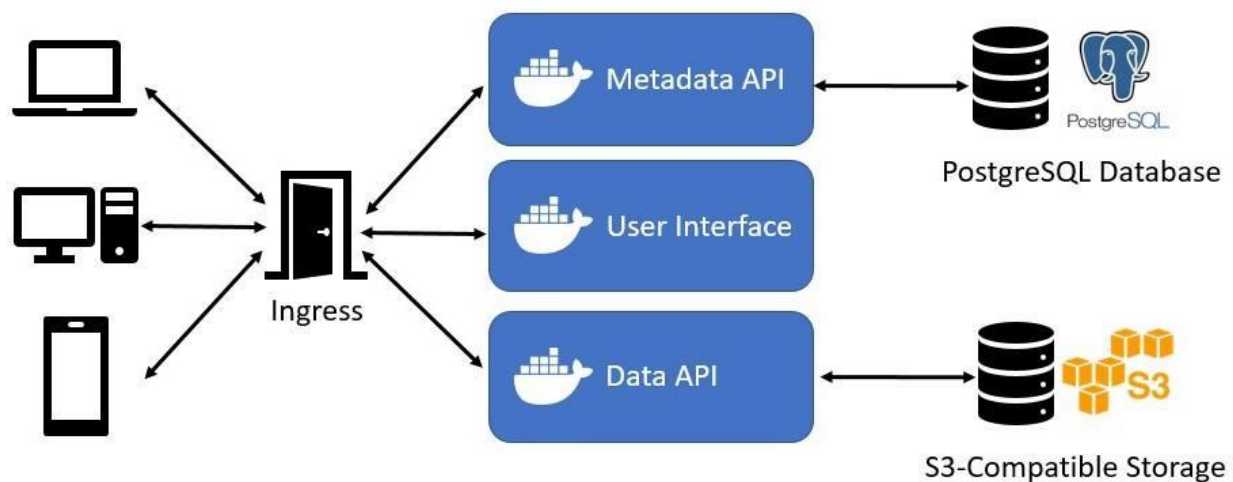


Fig. S12 The Signature Commons architecture. Signature Commons is composed of three microservices that are containerized with Docker. The SigCom LINCS repository provides images for ingress, as well as PostgreSQL and S3 storage. These are replaceable with services offered by cloud providers.

Fig. S13

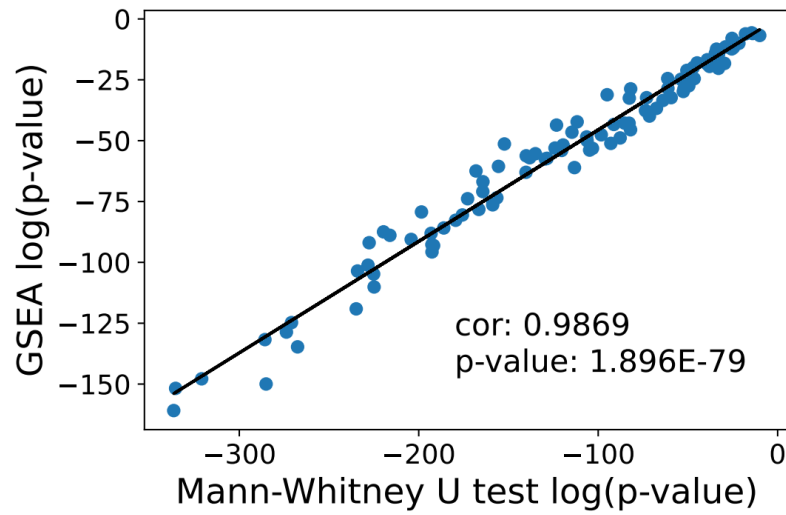


Fig. S13 Comparing the MWU test with the GSEA test. P-values produced by MWU and blitzGSEA for 100 gene sets with varying degrees of significance. To generate the random gene set ranks, we uniformly sample random rank positions in ranges of the full rank for different sizes of gene sets.

Fig. 14

The image displays three dataset entries from the LINCS Center for Transcriptomics (Broad Institute). Each entry includes a FAIRshake insignia, a title, a description, an assay type, a date, a file size, and a download count. The entries are:

- LINCS L1000 Overexpression Perturbations (2021)**: Assay: L1000 mRNA profiling assay, Date: 2021-06-10, File Size: 1.69 GB, Downloads: 0.
- LINCS L1000 CRISPR Perturbations (2021)**: Assay: L1000 mRNA profiling assay, Date: 2021-06-10, File Size: 6.98 GB, Downloads: 0.
- LINCS L1000 shRNA Perturbations (2021)**: Assay: L1000 mRNA profiling assay, Date: 2021-06-10, File Size: 7.83 GB, Downloads: 0.

Fig. S14 FAIR Assessments of the datasets and signatures within SigCom LINCS. The datasets and signatures in SigCom LINCS were assessed for FAIRness with FAIRshake with a unique rubric. These assessments are displayed near each dataset and signature entry as the FAIRshake insignia. Each square in the insignia represents a FAIR metric which asserts the adherence of the dataset and signature to the FAIR guiding principles.

SUPPORTING TABLES

Table S1

dataset	Signatures	Perturbations (drugs, genes, diseases, age groups)	Tissues of origin	Cell lines
LINCS L1000 Chemical Perturbations	718055	28564	34	196
LINCS L1000 shRNA Perturbations	158003	4917	15	22
LINCS L1000 CRISPR Perturbations	140945	5157	18	27
LINCS L1000 Overexpression Perturbations	34164	4040	10	20
LINCS L1000 Ligand Perturbations	7546	329	8	15
Automatic Human GEO RNA-seq Signatures	4269	-	-	-
Automatic Mouse GEO RNA-seq Signatures	4216	-	-	-
CREEDS Manual Single Gene Perturbations	2176	870	1364	-
CREEDS Manual Single Drug Perturbations	875	271	648	-
CREEDS Manual Disease Signatures	828	333	501	-
LINCS L1000 Antibody Perturbations	575	4	8	12
LINCS L1000 siRNA Perturbations	162	23	1	1
GTEX Aging Signatures	135	5	28	-

Table S1. Signature types in SigCom LINCS. SigCom LINCS catalogs over one million gene expression signatures along with their metadata generated from LINCS L1000 data, cDNA microarrays and RNA-seq from GEO, and RNA-seq data from GTEx.

Table S2

Label	Frequency	Category	num_gse
rep	23641	0	1733
rna	16840	0	2040
seq	13681	0	1829
replicate	6047	0	465
control	6023	1	1212
sample	5762	0	266
rnaseq	4654	0	363
patient	4557	0	173
cd	4244	0	261
cells	4094	0	395
dms0	3885	1	632
hr	3318	0	158
day	3317	0	180
wt	3081	1	619
tissue	3064	0	50
lib	2744	0	16
from	2727	0	1059
post	2677	2	59
donor	2642	0	159
+	2389	0	268
with	2115	0	169
mcf	2029	0	223
treated	2013	2	247
hra	1943	0	4
cgnd	1942	0	3
ctrl	1939	1	351
ko	1935	2	314
enclb	1858	0	983
shrna	1751	2	625
mortem	1706	0	1
seqc	1692	0	2

ilm	1691	0	1
mrna	1642	0	177
tumor	1629	2	91
x	1553	0	97
blood	1512	0	40
fgc	1358	0	2
fcc	1349	0	3
nm	1311	0	101
human	1305	0	135
non	1279	0	121
cell	1260	0	198
sirna	1259	2	220
il	1243	0	100
ad	1234	0	61
tr	1209	0	26
healthy	1197	1	84
vumc	1169	0	3
tb	1165	0	18
week	1152	0	21
hd	1119	0	36
acxx	1111	0	5
treatment	1090	2	80
total	1083	0	320
kd	1074	2	209
subject	1064	0	29
lb	1059	0	10
mc	1013	0	17
of	1008	0	116
ms	1000	0	31
clone	995	0	92
hek	978	0	94
ac	902	0	31
platelets	899	0	5
untreated	891	1	176
na	889	0	13
pre	888	0	57
mock	887	1	191

hepg	877	0	323
sh	877	0	135
individual	867	0	16
gfp	865	0	132
um	849	0	93
in	846	0	83
dox	820	2	111
w	818	0	76
vehicle	782	1	138
exp	781	0	43
biological	765	0	59
normal	764	1	107
mda	751	0	104
line	748	0	76
subjects	748	0	2
mb	747	0	105
bd	745	0	13
pc	737	0	95
hcc	732	0	71
mp	731	0	9
ir	719	0	18
hela	716	0	107
at	716	0	41
aml	707	0	33
incap	690	0	79
pf	667	0	17
case	662	2	19
baseline	661	1	25
ipsc	658	0	71
no	642	0	89
lps	638	2	69
pt	635	0	52
for	632	0	81
nt	619	0	105
male	615	0	26
primary	614	0	51
hours	604	0	42

af	594	0	34
batch	593	0	36
lung	589	0	55
cb	586	0	45
hc	579	0	48
hct	578	0	85
su	570	0	30
progressor	567	0	3
liver	562	0	49
pbmc	560	0	34
biopsy	555	0	14
jq	547	0	69
hrs	545	0	41
time	542	0	18
mir	533	0	83
cancer	533	0	36
infected	527	2	72
fgf	519	2	7
con	515	1	86
hs	513	0	58
gbm	509	0	22
risk	509	0	3
si	504	0	78
bc	492	0	27
plus	490	0	66
us	489	0	15
disease	484	2	18
female	482	0	25
nc	474	0	127
knockdown	473	2	84
subjectsle	468	0	1
high	461	0	68
sl	458	0	12
eu	457	0	6
polya	456	0	98
ctl	456	1	64
mm	454	0	53

vector	449	1	121
oe	444	2	70
naïve	441	1	20
fgfr	432	0	3
parental	429	0	84
neg	424	1	76
pd	422	0	48
sam	419	0	15
sle	414	0	15
tn	413	0	8
dm	412	0	39
npc	412	0	37
ba	411	0	21
molm	404	0	37
cn	403	0	9
ln	402	0	31
veh	397	1	78
repl	397	0	26
ns	395	0	47
ra	394	0	42
low	393	1	68
hesc	393	0	53
sw	389	0	53
mut	382	2	44
patients	377	0	5
re	372	0	27
lane	370	0	7
gm	368	0	45
after	366	2	26
uc	361	0	17
tdimi	361	0	1
ips	359	0	37
yri	359	0	1
panc	356	0	40
pos	355	0	46
and	352	0	70
uninfected	349	1	39

mv	349	0	33
transfected	345	2	33
ls	344	0	22
bgi	342	0	3
imr	339	0	43
fc	339	0	15
fibroblasts	338	0	35
ex	338	0	24
pa	335	0	25
fulv	335	0	3
cajmrnxx	335	0	2
crohn's	334	0	7
bt	333	0	52
on	333	0	31
nsclc	333	0	6
dx	331	2	4
repeat	330	0	40
th	325	0	24
whole	322	0	18
may	322	0	2
naive	320	1	46
days	319	0	30
ni	319	0	16
msc	318	0	38
bulk	316	0	26
thp	312	0	51
vu	311	0	5
ht	308	0	51
rv	306	0	45
mg	304	0	40
os	303	0	44
empty	301	0	73
sictrl	301	1	68
sum	301	1	29
cnl	301	1	2
pbmcs	299	0	14
cohort	295	0	3

tet	293	0	31
aid	290	0	18
kidney	289	0	27
anxx	289	0	5
pdac	289	0	2
min	287	0	35
ct	287	1	34
sicontrol	286	1	66
etoh	286	0	42
pat	286	0	8
sc	285	1	41
prostate	285	0	13
ca	283	0	34
sk	282	0	50
palbo	281	0	4
colon	280	0	31
tgfb	278	0	35
type	277	0	62
jc	276	0	4
jia	274	0	5
cgga	274	0	1
technical	270	0	12
derived	269	0	33
negative	268	1	57
mo	268	0	16
jacxx	268	0	3
ctr	267	1	47
pdx	267	0	28
wild	265	1	60
fg	262	0	4
nvs	262	0	2
input	260	0	52
gsk	259	0	38
rpe	259	0	36
stimulated	257	2	26
old	257	0	21
fibroblast	256	0	45

agr	256	0	1
hypoxia	255	0	43
muscle	255	0	26
st	255	0	26
rseq	255	0	2
huvec	252	0	44
fetal	251	0	25
tgf	251	0	22
rs	251	0	21
crpc	250	0	4
tnf	249	0	33
lv	249	0	17
tu	248	0	9
ev	247	0	70
cc	247	0	29
up	247	0	5
scc	246	0	24
mk	246	0	18
hiv	245	0	17
sgrna	244	0	24
young	243	0	23
controls	243	1	4
fiin	241	0	1
ifn	239	0	33
nuclear	239	0	24
dgm	239	0	11
cr	238	0	31
znf	237	0	18
resistant	236	2	42
ttr	236	0	3
copd	236	0	2
cm	235	0	42
nd	233	0	27
dld	231	0	20
macrophages	231	0	17
dht	230	0	35
skin	229	0	37

du	228	0	26
wm	227	0	23
br	226	0	18
ut	225	0	35
jurkat	224	0	31
analyzed	224	0	8
diff	222	0	22
hi	220	0	22
crispri	219	0	82
pb	218	0	21
overexpressed	218	2	13
ec	217	0	33
sp	216	0	31
mutant	213	2	35
dc	211	0	36
bm	211	0	35
id	211	0	13
ab	210	0	25
cl	210	0	24
tatagcct	210	0	1
mrnaseq	209	0	19
ggctctga	209	0	1
nl	208	0	15
analysis	208	0	12
atagaggc	208	0	1
hipsc	206	0	18
al	206	0	8
poly	204	0	37
to	204	0	19
gy	204	0	17
nasal	204	0	7
point	203	0	5
moi	202	0	10
sn	202	0	6
meningioma	202	0	2
paxgeneday	202	0	1
µm	201	0	15

fd	201	0	12
infection	200	2	15
ml	199	0	31
hour	199	0	16
ileal	199	0	1
ribo	198	0	17
group	198	0	11
differentiation	197	0	12
arthritis	197	0	3
rectal	197	0	2
globin	197	0	1
block	197	0	1
minus	195	0	30
msi	195	0	7
cctatcct	195	0	1
pbs	194	0	38
tc	194	0	18
bx	193	0	7
ng	192	1	22
brain	191	0	34
sb	191	0	21
tm	191	0	17
scramble	190	1	47
egf	190	0	20
rheumatoid	190	0	2
cytoplasmic	189	0	14
sm	188	0	13
amc	188	0	4
huh	185	0	33
ar	185	0	27
fnanxx	185	0	2
shctrl	184	1	48
sa	184	0	33
follow	184	0	1
hap	183	0	13
hf	183	0	13
hmec	183	0	7

tpm	183	0	2
experiment	182	0	18
dlpfc	182	0	4
hl	180	0	27
mt	180	2	23
participant	180	0	2
cnhi	180	0	1
sr	179	0	22
transduced	179	2	18
hfanxx	179	0	2
adult	178	0	32
rh	178	0	21
la	178	0	20
vcap	177	0	19
es	176	0	28
lm	176	0	25
rp	176	0	18
hpi	175	0	16
ov	175	0	12
ta	175	0	11
leicester	175	0	1
longitudinal	175	0	1
cs	174	0	22
timepoint	174	0	5
nk	173	0	25
xenograft	173	0	13
aza	172	0	24
as	171	0	29
sf	171	0	24
htert	171	0	19
monocyte	171	0	16
zt	171	0	2
knockout	170	2	31
set	170	0	16
bj	169	0	28
nb	169	0	27
unstimulated	169	1	27

co	167	0	29
all	167	0	20
bs	167	0	7
wb	167	0	4
only	166	0	17
msn	166	0	2
adipose	165	0	17
run	165	0	7
validation	165	0	4
gs	164	0	8
rpmi	163	0	25
arid	163	0	19
culture	162	0	17
stim	162	0	11
nutlin	161	2	21
ovcar	161	0	19
scr	160	1	38
ifng	160	2	21
discovery	160	0	2
uacxx	160	0	1
aypacxx	160	0	1
aytacxx	160	0	1
neuron	159	0	18
bl	159	0	9
year	158	0	5
normoxia	157	0	27
neurons	157	0	12
nsc	156	0	29
hep	155	0	32
yap	155	0	23
targeting	154	0	28
mel	154	0	26
obese	154	0	8
mdamb	153	0	18
hfwb	153	0	1
gr	152	0	11
sub	152	0	10

stem	151	0	22
fb	151	0	19
kras	151	0	10
mgh	151	0	5
nki	151	0	2
gdanxx	151	0	1
norm	150	1	9
lesional	150	0	3
expression	149	0	10
lc	149	0	10
epithelial	148	0	22
ociaml	148	0	9
heart	147	0	27
alpha	147	0	17
bone	147	0	17
auxin	147	0	15
epi	147	0	11
dose	147	0	6
acp	147	0	4
ha	146	0	24
lu	146	0	9
diagnosis	145	0	4
cil	144	0	13
endo	144	0	11
combo	142	0	23
differentiated	141	2	21
tgacca	141	0	20
cf	141	0	20
cgatgt	141	0	17
cg	141	0	5
study	141	0	5
lh	141	0	5
bp	140	0	20
condition	140	0	11
dt	140	0	9
cagatc	139	0	18
ly	138	0	24

dex	138	2	22
acagtg	138	0	17
yr	138	0	5
cttgta	137	0	18
gccaat	137	0	17
undiff	137	1	14
dn	136	0	18
hmsc	136	0	11
covid	136	0	9
oci	135	0	16
dcas	135	0	13
colorectal	135	0	4
lab	135	0	4
pool	134	0	18
aggcgaag	134	0	1
zzz	133	0	68
virus	133	0	15
stimulation	133	2	6
sequencing	133	0	4
hokt	133	0	2
wildtype	132	1	29
null	132	1	25
positive	132	2	23
pr	132	0	19
serum	132	0	17
atcacg	132	0	16
miapaca	132	0	16
tp	132	0	16
cutll	132	0	15
peripheral	132	0	12

Table S2 List of frequent terms identified in sample labels of GEO GSMs

Feature	SigCom LINCS	L1000 FWD	clue.io	iLINCS	L1000 CDS2	LDP1	LD2	LINCS Canvas Browser
2021 Release	X		X					
Access Indv. Sigs.	X	X		X	X		X	
Sig. Search	X	X	X	X	X		X	X
# of Sigs.	~1M	~16K	~1M	~200K	~20K	N/A	~500K	~100K
Ontology Metadata	X			X		X	X	
Gene Pages	X							
Drug Pages		X	X				X	
Metadata Search	X	X	X	X		X	X	
Gene to Sigs.	X							
Terms to Sig.	X							
Open API	X	X	X	X		X	X	
Regist. Required			X					
Heat maps	X		X		X			X
UMAP PCA t- SNE	X	X						

Table S3. Comparison of SigCom LINCS to other tools serving LINCS data and signatures.