



Icahn School
of Medicine at
Mount
Sinai

Department of Pharmacology and
System Therapeutics

GEOsRA database

Gene Expression Omnibus
Sequence Read Archive database

Caroline D. Monteiro

Ma'ayan Lab

Objective

Use the available SRA data from GEO studies to create a database that fulfills our current requirements for data and has a build in structure that facilitates the analysis.



LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.



LiveSlides web content

To view

Download the add-in.

liveslides.com/download

Start the presentation.

Query

```
use Data::Dumper;
use LWP::Simple;

#assemble the esearch URL with the desired time period, type of organism
my $base = 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/';
my $url = $base . 'esearch.fcgi?db=gds&term=(homo sapiens[ORGN]+OR+Mus musculus[ORGN]+OR+Rattus norvegicus[ORGN])+AND+gse[ETYP]+AND+sra[suppFile]&retmax=500000&usehistory=y';

#post the esearch URL
my $output = get($url);

#parse WebEnv and QueryKey
$web = $1 if ($output =~ /<WebEnv>(\S+)</WebEnv>/);
$key = $1 if ($output =~ /<QueryKey>(\d+)</QueryKey>/);

#assemble the esummary URL
$url = $base . "esummary.fcgi?db=$db&query_key=$key&WebEnv=$web";

#post the esummary URL
$docsums = get($url);
#print "$docsums";

#post the efetch URL
$data = get($url);
$dataout = getstore( $url, "geo_sra.txt")
print "$data";
```

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<eSearchResult>
  <Count>7876</Count>
  <RetMax>7876</RetMax>
  <RetStart>0</RetStart>
  <QueryKey>1</QueryKey>
  <WebEnv>
    NCID_1_99077948_130.14.18.34_9001_1468165660_432162525_0MetA0_S_MegaStore_F_1
  </WebEnv>
  <idlist>
    <Id>200075327</Id>
    <Id>200076783</Id>
    <Id>200077288</Id>
    <Id>200077704</Id>
    <Id>200077703</Id>
    <Id>200077702</Id>
    <Id>200077700</Id>
    <Id>200082300</Id>
    <Id>200083992</Id>
    <Id>200081682</Id>
    <Id>200081402</Id>
    <Id>200083979</Id>
    <Id>200081321</Id>
    <Id>200081329</Id>
    <Id>200081328</Id>
    <Id>200081057</Id>
    <Id>200083482</Id>
    <Id>200083356</Id>
    <Id>200083336</Id>
    <Id>200083340</Id>
    <Id>200081179</Id>
    <Id>200081181</Id>
    <Id>200081577</Id>
    <Id>200081576</Id>
    <Id>200079638</Id>
    <Id>200078996</Id>
    <Id>200070520</Id>
    <Id>200071543</Id>
    <Id>200071541</Id>
    <Id>200071539</Id>
    <Id>200063510</Id>
    <Id>200070436</Id>
    <Id>200073470</Id>
    <Id>200073528</Id>
    <Id>200081889</Id>
    <Id>200083584</Id>
    <Id>200075947</Id>
    <Id>200070150</Id>
    <Id>200081660</Id>
    <Id>200079525</Id>
    <Id>200079573</Id>
    <Id>200071376</Id>
    <Id>200079492</Id>
    <Id>200082043</Id>
    <Id>200083939</Id>
    <Id>200072196</Id>
    <Id>200071843</Id>
    <Id>200077374</Id>
    <Id>200077443</Id>
    <Id>200083888</Id>
    <Id>200093884</Id>
```

```
<eSearchResult>
  <Count>7876</Count>
  <RetMax>7876</RetMax>
  <RetStart>0</RetStart>
  <QueryKey>1</QueryKey>
  <WebEnv>
    NCID_1_99077948_130.14.18.34_9001_1468165660_432162525_0MetA0_S_MegaStore_F_1
  </WebEnv>
  <idlist>
    <Id>200077704</Id>
    <Id>200077703</Id>
    <Id>200077702</Id>
    <Id>200077700</Id>
    <Id>200082300</Id>
    <Id>200083992</Id>
    <Id>200081682</Id>
    <Id>200081402</Id>
    <Id>200083979</Id>
    <Id>200081321</Id>
    <Id>200081329</Id>
    <Id>200081328</Id>
    <Id>200081057</Id>
    <Id>200083482</Id>
    <Id>200083356</Id>
    <Id>200083336</Id>
    <Id>200083340</Id>
    <Id>200081179</Id>
    <Id>200081181</Id>
    <Id>200081577</Id>
    <Id>200081576</Id>
    <Id>200079638</Id>
    <Id>200078996</Id>
    <Id>200070520</Id>
    <Id>200071543</Id>
    <Id>200071541</Id>
    <Id>200071539</Id>
    <Id>200063510</Id>
    <Id>200070436</Id>
    <Id>200073470</Id>
    <Id>200073528</Id>
    <Id>200081889</Id>
    <Id>200083584</Id>
    <Id>200075947</Id>
    <Id>200070150</Id>
    <Id>200081660</Id>
    <Id>200079525</Id>
    <Id>200079573</Id>
    <Id>200071376</Id>
    <Id>200079492</Id>
    <Id>200082043</Id>
    <Id>200083939</Id>
    <Id>200072196</Id>
    <Id>200071843</Id>
    <Id>200077374</Id>
    <Id>200077443</Id>
    <Id>200083888</Id>
    <Id>200093884</Id>
```

```
</DocSum>
</DocSum>
<Id>200077703</Id>
<Item Name="Accession" Type="String">GSE77703</Item>
<Item Name="GDS" Type="String"/>
<Item Name="title" Type="String">
    Distinct and shared functions of ALS-associated TDP-43, FUS, and TAF15 revealed by comprehensive multi-system
    integrative analyses [RNA-Seq_mouse]
</Item>
<Item Name="summary" Type="String">
    TDP-43, FUS, and TAF15 are implicated in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia. We
    integrate CLIP-seq and RNA Bind-N-Seq technologies to discover that TAF15 binds to ~4,900 RNAs enriched for CGUA
    motifs. In the mouse brain, TAF15 and FUS, but not TDP-43, exhibit strikingly similar RNA binding profiles, yet
    they alter the expression of distinct mRNA populations upon their individual depletions. TAF15 has a minimal role
    in alternative splicing and instead affects RNA turnover, consistent with an enrichment of TAF15 binding sites in
    3' untranslated regions. In human stem cell-derived motor neurons, loss of both TAF15 and FUS affected mRNAs
    distinct from those altered by loss of either protein alone, revealing redundant roles for TAF15 and FUS in
    maintaining mRNA levels. Furthermore, concomitant rather than individual depletion of TAF15 and FUS more closely
    resembles RNA profiles of motor neurons derived from FUS R521G ALS patients or from late-stage, sporadic ALS
    patients. Our study reveals convergent and divergent mechanisms by which FUS, TAF15 and TDP-43 affects RNA
    metabolism in neurological disease.
</Item>
<Item Name="GPL" Type="String">13112</Item>
<Item Name="GSE" Type="String">77703</Item>
<Item Name="taxon" Type="String">Mus musculus</Item>
<Item Name="entryType" Type="String">GSE</Item>
<Item Name="gdsType" Type="String">Expression profiling by high throughput sequencing</Item>
<Item Name="ptechType" Type="String"/>
<Item Name="valType" Type="String"/>
<Item Name="SSInfo" Type="String"/>
<Item Name="subsetInfo" Type="String"/>
<Item Name="PDAT" Type="String">2016/07/05</Item>
<Item Name="suppFile" Type="String">CSV</Item>
<Item Name="Samples" Type="List">
    <Item Name="Sample" Type="Structure">
        <Item Name="Accession" Type="String">GSM2056813</Item>
        <Item Name="Title" Type="String">HTT Control Rep 2</Item>
    </Item>
    <Item Name="Sample" Type="Structure">
        <Item Name="Accession" Type="String">GSM2056811</Item>
        <Item Name="Title" Type="String">TAF15 Rep 2</Item>
    </Item>
    <Item Name="Sample" Type="Structure">
        <Item Name="Accession" Type="String">GSM2056812</Item>
        <Item Name="Title" Type="String">HTT Control Rep 1</Item>
    </Item>
    <Item Name="Sample" Type="Structure">
        <Item Name="Accession" Type="String">GSM2056810</Item>
        <Item Name="Title" Type="String">TAF15 Rep 1</Item>
    </Item>

```

Accession	Source link	Title	Organism	Number of Samples	Date	Summary
GSE75327	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75327	Aberrant downstream transcriptional control by oncogenic Enhancers from the epigenetic landscape	Homo sapiens	18	2016-07-09	Genes encoding the histone H3 lysine 4 methylation mark were found to be enriched in the noncoding regions of tumor cell genomes.
GSE76783	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76783	Oncogenic Enhancers from the epigenetic landscape	Homo sapiens	20	2016-07-08	The noncoding regions of tumor cell genomes contain many oncogenic enhancers.
GSE77288	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77288	Batch effects and the effect of environmental exposure on gene expression	Homo sapiens	873	2016-07-08	Single cell RNA sequencing (scRNA-seq) was used to study the effect of environmental exposure on gene expression.
GSE72609	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72609	Intergenerational consequences of environmental exposure on gene expression	Mus musculus	15	2016-07-08	The ability of environmental exposure to affect gene expression across generations.
GSE72608	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72608	Intergenerational consequences of environmental exposure on gene expression	Mus musculus	78	2016-07-08	The ability of environmental exposure to affect gene expression across generations.
GSE72607	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72607	Intergenerational consequences of environmental exposure on gene expression	Mus musculus	44	2016-07-08	The ability of environmental exposure to affect gene expression across generations.
GSE84107	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84107	Dis3l2-Mediated Decay	Mus musculus	4	2016-07-08	Mutations in the 3'x89'xdB\xaa-5'x89' sequence motif in Dis3l2 are associated with gene decay.
GSE79304	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79304	Global transcript structure	Homo sapiens	18	2016-07-08	Strand-specific Illumina RNA-Seq was used to study global transcript structure.
GSE79293	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79293	Global transcript structure	Homo sapiens	4	2016-07-08	deepCAGE was used in conjunction with RNA-Seq to study global transcript structure.
GSE79335	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79335	Global transcript structure	Homo sapiens	1	2016-07-08	Pacific Biosciences Iso-Seq was used in conjunction with RNA-Seq to study global transcript structure.
GSE81842	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81842	Genome-wide comparison of gene expression in leukemic cell lines	Mus musculus	15	2016-07-07	Analysis of gene expression of leukemia cell lines.
GSE73947	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73947	ZNF224 binding profiles	Homo sapiens	1	2016-07-07	ZNF224 is a Kruppel-associated box-containing transcription factor.
GSE73457	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73457	Genome-wide analysis of long non-coding RNAs	Mus musculus	12	2016-07-07	The purpose of this study is to investigate long non-coding RNAs.
GSE61549	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61549	A long non-coding RNA at the 3p14 locus	Homo sapiens	41	2016-07-07	Long non-coding RNAs (lncRNAs) are expressed in the brain.
GSE74994	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74994	Single Cell Expression Profiling of Human Epiblast Stem Cells	Mus musculus	2	2016-07-07	We analyzed 1 205 cells from the epiblast stem cell line.
GSE81498	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81498	Multiple mechanisms drive let-7 microRNA families	Homo sapiens	9	2016-07-06	The let-7 microRNA families are tumor suppressors.
GSE81497	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81497	Multiple mechanisms drive let-7 microRNA families	Homo sapiens	4	2016-07-06	The let-7 microRNA families are tumor suppressors.
GSE81499	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81499	Multiple mechanisms drive let-7 microRNA families	Homo sapiens	5	2016-07-06	The let-7 microRNA families are tumor suppressors.
GSE81635	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81635	Genome-wide mapping of the RAG1 targetome	Homo sapiens	2	2016-07-06	Growing number of cancer (stem) cells.
GSE72696	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72696	Integrative fine-mapping of the RAG1 targetome	Homo sapiens	24	2016-07-06	Coronary artery disease (CAD) is the leading cause of death worldwide.
GSE84052	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84052	RAG1 targeting in the genus musculus; Homo sapiens	Homo sapiens	4	2016-07-06	The RAG1/RAG2 endonuclease initiates DNA double-strand breaks.
GSE76514	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76514	Laser-capture microdissection of the mouse brain	Mus musculus; Homo sapiens	153	2016-07-05	Generally when LCM is used in diverse tissues, it can lead to artifacts.
GSE83139	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83139	Single cell RNA-seq of hippocampal neurons	Homo sapiens	635	2016-07-05	We successfully sequenced and annotated single cell RNA-seq data.
GSE81897	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81897	Shox2/TALE ChIP-Seq	Mus musculus	7	2016-07-05	ChIP-Sequencing on Shox2-HA E12.5 animals.
GSE67540	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67540	RNA polymerase in pre-mRNA processing	Homo sapiens	21	2016-07-05	[Gro-seq] Precursor B acute leukemia.
GSE70519	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70519	DNA Cytosine Hydroxymethyltransferases	Homo sapiens	7	2016-07-05	Peripheral blood leukocytes are the main source of TDP-43.
GSE77704	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77704	Distinct and shared functional roles of TAF15 and TAF15-like proteins	Homo sapiens	18	2016-07-05	TDP-43 FUS and TAF15 are implicated in various diseases.

Text Mining

TextBlob

```
import csv
from textblob.classifiers import NaiveBayesClassifier
from textblob import TextBlob
```

Creating a text classification system:

- Word category list

RNA-Seq
ChIP-Seq

- Tokenize description

```
for line in data:
    description=line[1]
    blob=TextBlob(description)
    text_words=(blob.words)
```

Transcriptome of HEK and B cells were analyzed by microarray and RNA-Seq parallelly. Both platforms were then compared in terms of sensitivity. To assess whether values were a reliable indicator of gene activity we correlated these values with hypophosphorylated RNA polymerase II (PolIIa) occupancy used as a landmark of transcription initiation. For HEK we identified PolIIa islands by chromatin immunoprecipitation and sequencing (ChIP-Seq).

```
['Transcriptome', 'of', 'HEK', 'and', 'B', 'cells', 'were', 'analyzed', 'by', 'microarray', 'and', 'RNA-Seq', 'parallelly.', 'Both', 'platforms', 'were', 'then', 'compared', 'in', 'terms', 'of', 'sensitivity.', 'To', 'assess', 'whether', 'values', 'were', 'a', 'reliable', 'indicator', 'of', 'gene', 'activity', 'we', 'correlated', 'these', 'values', 'with', 'hypophosphorylated', 'RNA', 'polymerase', 'II', '(PolIIa)', 'occupancy', 'used', 'as', 'a', 'landmark', 'of', 'transcription', 'initiation.', 'For', 'HEK', 'we', 'identified', 'PolIIa', 'islands', 'by', 'chromatin', 'immunoprecipitation', 'and', 'sequencing', '(ChIP-Seq).']
```

Transcription factors and their specific interactions with targets are crucial in specifying gene expression programs. To gain insights into the transcriptional regulatory networks in embryonic stem cells we use chromatin immunoprecipitation coupled to ultra-high-throughput DNA sequencing (ChIP-seq) to map the locations of thirteen sequence specific transcription factors (Nanog Oct4 STAT3 Smad1 Sox2 Zfx c-Myc n-Myc Klf4 Esrrb Tcfcp2l1 E2f1 and CTCF) and two transcription regulators (p300 and Suz12). These factors are known to play different roles in ES cell biology as components of the LIF and BMP signaling pathways self-renewal regulators and key reprogramming factors. Our study provides insights into the integration of the signaling pathways to the ES cell-specific transcription circuitries. Intriguingly we find specific genomic regions extensively targeted by different transcription factors. Collectively the comprehensive mapping of transcription factor binding sites identifies important features of the transcriptional regulatory networks that define ES cell identity. Keywords: Transcription factor binding sites in undifferentiated mouse embryonic stem cells

```
['Transcription', 'factors', 'and', 'their', 'specific', 'interactions', 'with', 'targets', 'are', 'crucial', 'in', 'specifying', 'gene', 'expression', 'programs.', 'To', 'gain', 'insights', 'into', 'the', 'transcriptional', 'regulatory', 'networks', 'in', 'embryonic', 'stem', 'cells', 'we', 'use', 'chromatin', 'immunoprecipitation', 'coupled', 'to', 'ultra-high-throughput', 'DNA', 'sequencing', '(ChIP-seq)', 'to', 'map', 'the', 'locations', 'of', 'thirteen', 'sequence', 'specific', 'transcription', 'factors', '(Nanog', 'Oct4', 'STAT3', 'Smad1', 'Sox2', 'Zfx', 'c-Myc', 'n-Myc', 'Klf4', 'Esrrb', 'Tcfcp2l1', 'E2f1', 'and', 'CTCF)', 'and', 'two', 'transcription', 'regulators', '(p300', 'and', 'Suz12).', 'These', 'factors', 'are', 'known', 'to', 'play', 'different', 'roles', 'in', 'ES', 'cell', 'biology', 'as', 'components', 'of', 'the', 'LIF', 'and', 'BMP', 'signaling', 'pathways', 'self-renewal', 'regulators', 'and', 'key', 'reprogramming', 'factors.', 'Our', 'study', 'provides', 'insights', 'into', 'the', 'integration', 'of', 'the', 'signaling', 'pathways', 'to', 'the', 'ES', 'cell-specific', 'transcription', 'circuitries.', 'Intriguingly', 'we', 'find', 'specific', 'genomic', 'regions', 'extensively', 'targeted', 'by', 'different', 'transcription', 'factors.', 'Collectively', 'the', 'comprehensive', 'mapping', 'of', 'transcription', 'factor', 'binding', 'sites', 'identifies', 'important', 'features', 'of', 'the', 'transcriptional', 'regulatory', 'networks', 'that', 'define', 'ES', 'cell', 'identity.', 'Keywords:', 'Transcription', 'factor', 'binding', 'sites', 'in', 'undifferentiated', 'mouse', 'embryonic', 'stem', 'cells']
```

Categorization

```
gse_category.shape  
(9244, 1)
```

Every time the word appears in the description its also added again to the category list.

```
gse_category
```

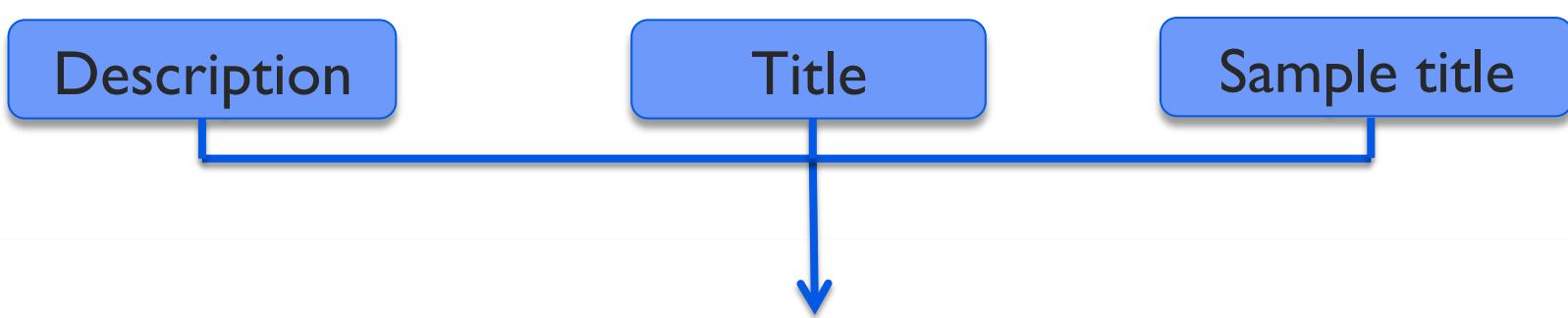
12	('GSE79293', 'RNA-Seq')
13	('GSE79293', 'RNA-Seq')
14	('GSE79293', 'RNA-Seq')
15	('GSE79293', 'RNA-Seq')
16	('GSE79293', 'RNA-Seq')
17	('GSE79293', 'RNA-Seq')
18	('GSE79293', 'RNA-Seq')
19	('GSE79293', 'RNA-Seq')
20	('GSE79335', 'RNA-Seq')
21	('GSE79335', 'RNA-Seq')
22	('GSE79335', 'RNA-Seq')
23	('GSE79335', 'RNA-Seq')
24	('GSE79335', 'RNA-Seq')

Categorization

```
lines_seen = set()
outfile = open('gse_category.csv','w')
for line in open('gse_category_rc.csv','rU'):
    if line not in lines_seen:
        outfile.write(line)
        lines_seen.add(line)
outfile.close()
```

After removing duplicates:

1236 UNIQUE accessions



~2500 UNIQUE accessions

**Control = Just
one category:**

Control
Wild Type
Untreated
DMSO
Vehicle

GSM1897882	WT_AL_ZT04_C_RFP	Control
GSM1919397	P18_lcm_normal	Control
GSM1486223	Control-RNAi/Glucose oxidase challenge 1	Control
GSM1495733	Dicer WT with vector tranfection [RiboZero-prepped library]	Control
GSM1504380	Input WT 1	Control
GSM1516108	Sample_54_WT	Control
GSM1519141	RNAseq_II4ra_WT_rep2	Control
GSM1967319	MEF WT 20 h replicate 2	Control
GSM2048426	WT_Tumor_11_5_L	Control
GSM1550723	si-/Negative control_5236	Control
GSM1551038	wt 51	Control
GSM1014632	GRO-seq_Vehicle_AC16_2	Control
GSM2052004	brain_WT_Z3	Control
GSM2055889	H1 ChIP-seq in WT Mice	Control
GSM2067941	Input WT Control 2	Control
GSM2070737	WT 10 months Female Cerebellum mRNA rep4 398979_92521...	Control
GSM1563247	WT	Control
GSM2076891	WT 6 months Female Cerebellum miRNA rep2 (Q20)	Control
GSM1573823	Control rep2	Control
GSM2090881	RN2 DMSO 2	Control
GSM2100630	CTRL-8	Control
GSM2113059	M15_Control_line_b_day_32	Control
GSM2123232	H3K27Ac_DN_WT_rep1	Control
GSM1599090	YY1_d3_Ocf_WT_Oct4_1strep	Control
GSM1150961	Control shRNA-unstim	Control

Condition = multiple categories:

Overexpression

Knockout

Knockdown

Treated

accession	sample	description	sample_type
GSE70150	GSM1717518	LNCaP DHT treated 3 h	treated
GSE69420	GSM1700950	HCT116 was treated by S2101	treated
GSE69420	GSM1700942	MHCC97H was treated by S2101	treated
GSE69420	GSM1700951	HCT116 was treated by TSA	treated
GSE69420	GSM1700945	A549 was treated by GSK126	treated
GSE69420	GSM1700946	A549 was treated by S2101	treated
GSE69420	GSM1700943	MHCC97H was treated by TSA	treated
GSE69420	GSM1700949	HCT116 was treated by GSK126	treated
GSE69420	GSM1700941	MHCC97H was treated by GSK126	treated
GSE69420	GSM1700947	A549 was treated by TSA	treated
GSE71777	GSM1845151	SGC0946 treated Rep 1	treated
GSE71777	GSM1845154	I-BET + SGC0946 Rep 2	treated
GSE71777	GSM1845149	I-BET treated Rep 1	treated
GSE71777	GSM1845152	SGC0946 treated Rep 2	treated
GSE71777	GSM1845153	I-BET + SGC0946 Rep 1	treated
GSE71777	GSM1845150	I-BET treated Rep 2	treated
GSE72023	GSM1850205	Jurkat 50 nM THZ-5-31-1 Pol2	treated
GSE72023	GSM1850206	Jurkat 500 nM THZ-5-31-1 Pol2	treated
GSE69858	GSM1711435	Hlf inducible GMLPs cultured on OP...	treated
GSE69858	GSM1711436	Hlf inducible GMLPs cultured on OP...	treated
GSE70942	GSM1823638	R1 RUNX1 overexpressing	overexpression
GSE70942	GSM1823639	R2 RUNX1 overexpressing	overexpression
GSE70942	GSM1823640	R3 RUNX1 overexpressing	overexpression
GSE74660	GSM1924996	P28KO rep1 seq	knockout
GSE78873	GSM2080025	KO Macrophage 2HR Rep1	knockout
GSE78800	GSM2077307	Fetal liver Knockout of alpha globin...	knockout
GSE78873	GSM2080023	KO Macrophage 0HR Rep1	knockout

Database

```
CREATE TABLE `accession` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT,
  `accession` varchar(20) DEFAULT NULL,
  `source_link` varchar(225) DEFAULT NULL,
  `study_title` text,
  `organism` varchar(100) DEFAULT NULL,
  `date_last_update` varchar(20) DEFAULT NULL,
  `number_of_samples` varchar(20) DEFAULT NULL,
  `study_category` varchar(20) DEFAULT NULL,
  `summary` text NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

CREATE TABLE `gsm` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT,
  `accession` varchar(20) DEFAULT NULL,
  `sample` varchar(20) DEFAULT NULL,
  `description` text,
  `sample_type` varchar(20) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;

CREATE TABLE `gsm_accession` (
  `id` int(10) unsigned NOT NULL AUTO_INCREMENT,
  `gsm_id` int(10) unsigned NOT NULL,
  `accession_id` int(10) unsigned NOT NULL,
  PRIMARY KEY (`id`),
  KEY `gsm_accession_gsm_id_foreign` (`gsm_id`),
  KEY `gsm_accession_accession_id_foreign` (`accession_id`),
  CONSTRAINT `gsm_accession_gsm_id_foreign` FOREIGN KEY (`gsm_id`) REFERENCES `gsm` (`id`),
  CONSTRAINT `gsm_acssession_accession_id_foreign` FOREIGN KEY (`accession_id`) REFERENCES `accession`
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

accession	source_link	study_title	organism	date_last_update	number_o...	study_category	summary	Curated
GSE72671	http://www.ncbi.nlm.nih.gov/geo/study/GSE72671	Gfi1 as a new predictive and therapeutical target in human glioma	Mus musculus	2016-05-12	12	RNA-Seq;ChIP-Seq	MDS is characterized by a disturbed function of Gfi1 in glioma cells	NULL
GSE57694	http://www.ncbi.nlm.nih.gov/geo/study/GSE57694	Expression profiling, Miz1 location analysis and ChIP-Seq analysis	Mus musculus	2014-12-01	16	RNA-Seq;ChIP-Seq	Expression profiling by RNA-Seq was performed	NULL
GSE81897	http://www.ncbi.nlm.nih.gov/geo/study/GSE81897	Shox2/TALE ChIP-Seq	Mus musculus	2016-07-05	7	RNA-Seq;ChIP-Seq	ChIP-Sequencing on Shox2-HA E12.5 animals	NULL
GSE67540	http://www.ncbi.nlm.nih.gov/geo/study/GSE67540	RNA polymerase in pre-B-ALL cell lines	Homo sapiens	2016-07-05	21	RNA-Seq;ChIP-Seq	[Gro-seq] Precursor B acute leukemia cell lines	NULL
GSE39406	http://www.ncbi.nlm.nih.gov/geo/study/GSE39406	Site-specific silencing of regulatory elements by H3K27me3	Mus musculus	2012-11-21	17	RNA-Seq;ChIP-Seq	The inactive X chromosome\x89\xdb\x89\xdb	NULL
GSE63204	http://www.ncbi.nlm.nih.gov/geo/study/GSE63204	Opposing Roles of STAT1 and STAT3 in IL-21-activated T cells	Mus musculus	2015-07-01	28	RNA-Seq;ChIP-Seq	Interleukin-21 (IL-21) is a type 1 cytokine	NULL
GSE42350	http://www.ncbi.nlm.nih.gov/geo/study/GSE42350	Next Generation Sequencing Analysis of Wild-type and Mice Lymphocytes	Mus musculus	2013-02-12	4	RNA-Seq;ChIP-Seq	Absence of 53BP1 influences the profile of DNA double-strand break repair	NULL
GSE56360	http://www.ncbi.nlm.nih.gov/geo/study/GSE56360	Transcriptional landscape of Rag2 -/- thymocytes	Mus musculus	2014-04-01	2	RNA-Seq;ChIP-Seq	We performed ChIP-Seq for hallmark TFs	NULL
GSE23943	http://www.ncbi.nlm.nih.gov/geo/study/GSE23943	Epigenome and transcriptome of naive pluripotent stem cells	Mus musculus	2012-04-27	41	RNA-Seq;ChIP-Seq	This study describes the epigenetic and transcriptional	NULL
GSE69095	http://www.ncbi.nlm.nih.gov/geo/study/GSE69095	Comprehensive Epigenomic Analysis Reveals the Role of G9a in Cell Differentiation	Mus musculus	2016-02-26	5	RNA-Seq;ChIP-Seq	Embryonic hematopoiesis is regulated by G9a	NULL
GSE71499	http://www.ncbi.nlm.nih.gov/geo/study/GSE71499	Study of the impact of G9a on DNA methylation in embryonic hematopoiesis	Mus musculus	2015-11-13	20	RNA-Seq;ChIP-Seq	We explored the role of the lysine-methylation	NULL
GSE73887	http://www.ncbi.nlm.nih.gov/geo/study/GSE73887	Identification of MEF2B, EBF1, and IL6R as chaperone proteins required for the formation of the PRC1 complex	Homo sapiens	2015-12-16	12	RNA-Seq;ChIP-Seq	EBNA1 is the EBV-encoded nuclear anti-sense RNA	NULL
GSE31221	http://www.ncbi.nlm.nih.gov/geo/study/GSE31221	TCF7 is a key regulator of the switch of self-renewal to differentiation in mouse ES cells	Mus musculus	2012-03-30	5	RNA-Seq;ChIP-Seq	A critical problem in biology is understanding	NULL
GSE49389	http://www.ncbi.nlm.nih.gov/geo/study/GSE49389	Genome-wide analysis of the rat colon reveals distinct gene expression programs in the colonic mucosa	Rattus norvegicus	2013-08-01	7	RNA-Seq;ChIP-Seq	Goal: Since disease susceptibility of the	NULL
GSE11892	http://www.ncbi.nlm.nih.gov/geo/study/GSE11892	A Global View of Gene Activity and Alternative Splicing in Human HEK293T and B cells	Homo sapiens	2008-07-03	11	RNA-Seq;ChIP-Seq	Transcriptome of HEK and B cells were compared	NULL
GSE54412	http://www.ncbi.nlm.nih.gov/geo/study/GSE54412	Investigation of the role of histone modifications in the regulation of gene expression	Mus musculus	2015-01-20	40	RNA-Seq;ChIP-Seq	We use ChIP-Seq and RNA-Seq technolo...	NULL
GSE57758	http://www.ncbi.nlm.nih.gov/geo/study/GSE57758	Promoter Decommissioning by the NuRD Complex	Mus musculus	2014-08-18	18	RNA-Seq;ChIP-Seq	Precise control of gene expression play...	NULL
GSE30399	http://www.ncbi.nlm.nih.gov/geo/study/GSE30399	Genome wide mapping for TF binding sites from ChIP-Seq and RNA-Seq	Homo sapiens	2011-07-09	5	RNA-Seq;ChIP-Seq	We performed ChIP-Seq for 5 different t...	NULL
GSE48085	http://www.ncbi.nlm.nih.gov/geo/study/GSE48085	Genome-scale analysis in blood progenitor cells	Mus musculus	2014-04-24	20	RNA-seq; ChIP-Seq	Despite major advances in the generati...	NULL
GSE65808	http://www.ncbi.nlm.nih.gov/geo/study/GSE65808	Combining ChIP-Seq and RNA-Seq data in early-stage placental tissue	Mus musculus	2015-12-01	6	RNA-seq; ChIP-Seq	The placenta is an understudied organ t...	NULL
GSE65807	http://www.ncbi.nlm.nih.gov/geo/study/GSE65807	Combining ChIP-Seq and RNA-Seq data in early-stage placental tissue	Mus musculus	2015-12-01	8	RNA-seq; ChIP-Seq	The placenta is an understudied organ t...	NULL
GSE44346	http://www.ncbi.nlm.nih.gov/geo/study/GSE44346	RNA-Seq and RNA Polymerase II ChIP-Seq of spermatocyte nuclei	Mus musculus	2014-01-22	12	RNA-seq; ChIP-Seq	To characterize gene expression in sper...	NULL
GSE69322	http://www.ncbi.nlm.nih.gov/geo/study/GSE69322	Gata6 potently initiates reprogramming of pluripotent stem cells	Homo sapiens;...	2015-06-30	16	RNA-seq; ChIP-Seq	Transcription factor-mediated reprogram...	NULL
GSE52657	http://www.ncbi.nlm.nih.gov/geo/study/GSE52657	Dynamic developmental signaling logic underlies the formation of the mouse brain	Homo sapiens	2013-11-22	36	RNA-seq; ChIP-Seq	Unraveling complex signaling programs...	NULL
GSE43785	http://www.ncbi.nlm.nih.gov/geo/study/GSE43785	Redefinition of Human Androgen Responsive Genes	Homo sapiens	2015-01-01	20	RNA-seq; ChIP-Seq	The androgen receptor (AR) mediates th...	NULL
GSE63282	http://www.ncbi.nlm.nih.gov/geo/study/GSE63282	ChIP-seq and RNA-seq analyses identify Wnt target genes in mouse ES cells	Mus musculus	2015-04-01	4	RNA-seq; ChIP-Seq	The Prep1 (Pknx1) homeodomain tran...	NULL
GSE48005	http://www.ncbi.nlm.nih.gov/geo/study/GSE48005	Targeted disruption of Hotair leads to homeostatic changes in gene expression	Mus musculus	2013-08-21	12	RNA-seq; ChIP-Seq	Long noncoding RNAs (lncRNAs) are th...	NULL
GSE36882	http://www.ncbi.nlm.nih.gov/geo/study/GSE36882	Critical Role of STAT5 Transcription Factor in the Regulation of Gene Expression	Mus musculus	2012-04-26	30	RNA-seq; ChIP-Seq	Cytokine-activated STAT proteins dimer...	NULL
GSE63889	http://www.ncbi.nlm.nih.gov/geo/study/GSE63889	H3K18Ac and H3K18Cr ChIP-seq and RNA-seq analysis of the mouse brain	Mus musculus	2015-03-26	10	RNA-seq; ChIP-Seq	(1) We sought to characterize the geno...	NULL

accession	sample	description	sample_type
GSE70150	GSM1717518	LNCaP DHT treated 3 h	treated
GSE69420	GSM1700950	HCT116 was treated by S2101	treated
GSE69420	GSM1700942	MHCC97H was treated by S2101	treated
GSE69420	GSM1700951	HCT116 was treated by TSA	treated
GSE69420	GSM1700945	A549 was treated by GSK126	treated
GSE69420	GSM1700946	A549 was treated by S2101	treated
GSE69420	GSM1700943	MHCC97H was treated by TSA	treated
GSE69420	GSM1700949	HCT116 was treated by GSK126	treated
GSE69420	GSM1700941	MHCC97H was treated by GSK126	treated
GSE69420	GSM1700947	A549 was treated by TSA	treated
GSE71777	GSM1845151	SGC0946 treated Rep 1	treated
GSE71777	GSM1845154	I-BET + SGC0946 Rep 2	treated
GSE71777	GSM1845149	I-BET treated Rep 1	treated
GSE71777	GSM1845152	SGC0946 treated Rep 2	treated
GSE71777	GSM1845153	I-BET + SGC0946 Rep 1	treated
GSE71777	GSM1845150	I-BET treated Rep 2	treated
GSE72023	GSM1850205	Jurkat 50 nM THZ-5-31-1 Pol2	treated
GSE72023	GSM1850206	Jurkat 500 nM THZ-5-31-1 Pol2	treated
GSE69858	GSM1711435	Hif inducible GMLPs cultured on OP...	treated
GSE69858	GSM1711436	Hif inducible GMLPs cultured on OP...	treated
GSE70942	GSM1823638	R1 RUNX1 overexpressing	overexpression
GSE70942	GSM1823639	R2 RUNX1 overexpressing	overexpression
GSE70942	GSM1823640	R3 RUNX1 overexpressing	overexpression
GSE74660	GSM1924996	P28KO rep1 seq	knockout
GSE78873	GSM2080025	KO Macrophage 2HR Rep1	knockout
GSE78800	GSM2077307	Fetal liver Knockout of alpha globin...	knockout
GSE78873	GSM2080023	KO Macrophage OHR Rep1	knockout

Website



THANK YOU

