

Publicly Available Data

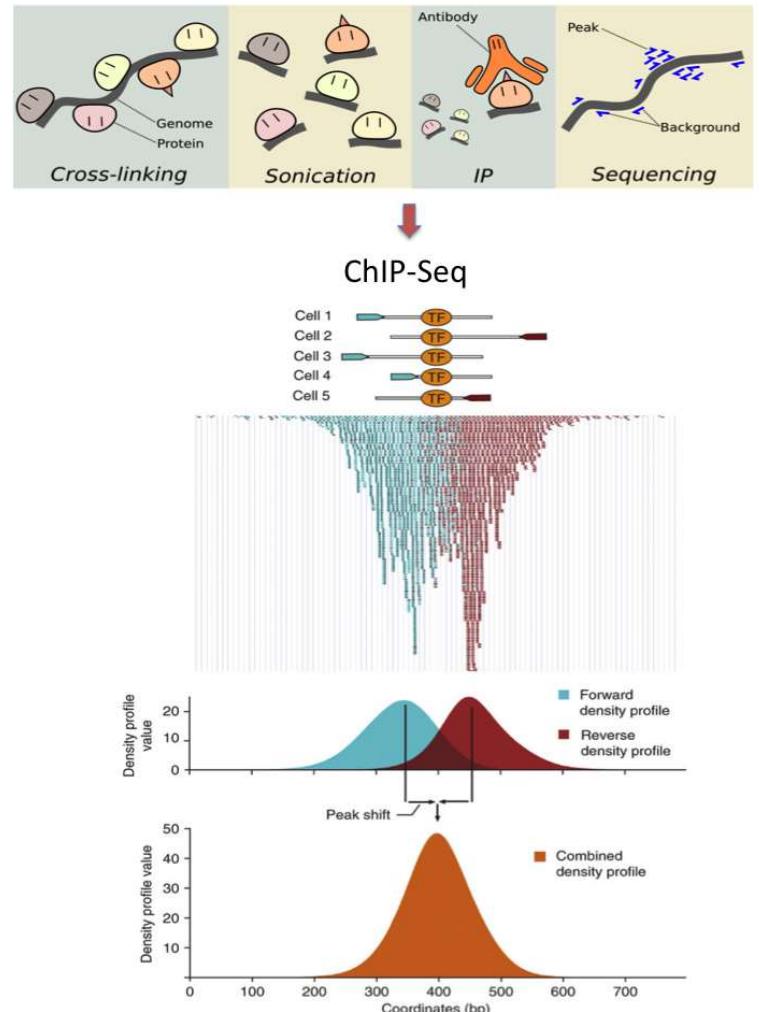
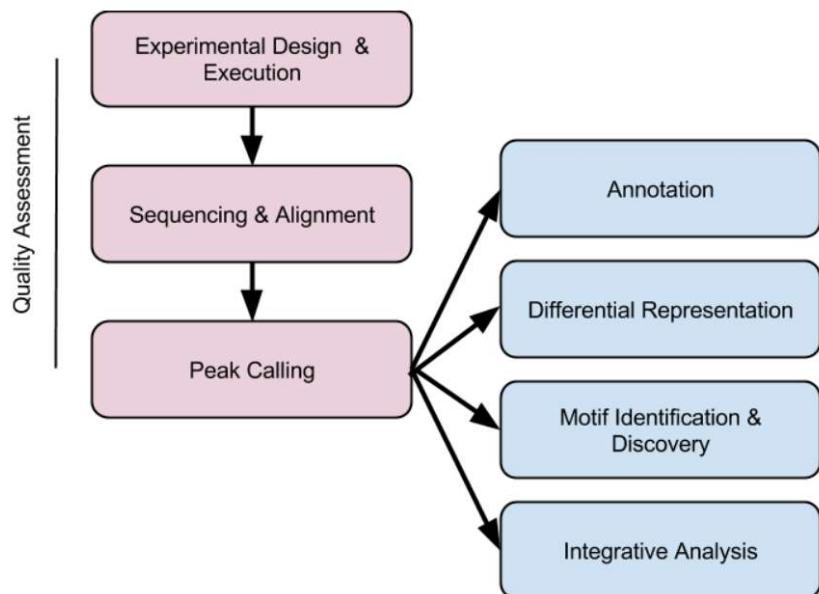
Lauren Vanderlinden

BIOS 6660

Spring 2019

Overview From Last Time

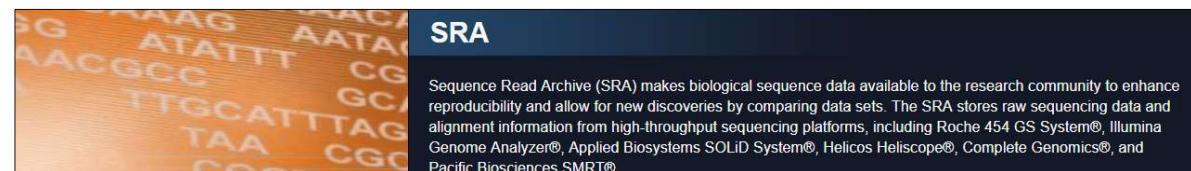
- Have wrapped up ChIP-Seq Analysis



Publicly Available Data



- Free data!
- Validate your results
 - Data generated under same/similar conditions
 - How consistent are the results?
- Statistical method development
 - Real world data
- Compare to a different experimental design
 - Tissue differences



SRA
Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos HeliScope®, Complete Genomics®, and Pacific Biosciences SMRT®.



GEO Database

- NCBI Gene Expression Omnibus
- Microarray database
 - mRNA arrays
 - Methylation arrays
 - SNP arrays
 - miRNA arrays
- Query like a literature search

NCBI Resources How To

GEO DataSets GEO DataSets ER positive breast cancer Create alert Advanced

Entry type DataSets (25) Series (442) Samples (11,266) Platforms (12)

Organism Customize ...

Study type Expression profiling by array Methylation profiling by array Customize ...

Author Customize ...

Attribute name tissue (3,704) strain (6) Customize ...

Publication dates 30 days 1 year Custom range... Clear all Show additional filters

Summary 20 per page Sort by Default order

Send to: Filters: Manage Filters

Search results Items: 1 to 20 of 11745 << First < Prev Page 1 of 588 Next > Last >

Top Organisms [Tree]
Homo sapiens (11732)
Mus musculus (32)
Rattus norvegicus (6)
synthetic construct (3)
Human alphaherpesvirus 1 (2)
More...

Find related data Database: Select Find items

Search details ER[All Fields] AND positive[All Fields] AND ("breast neoplasms" [MeSH Terms] OR breast cancer[All Fields])

Recent activity Turn Off Clear

ER positive breast cancer (11745) GEO DataSets

MicroRNA-135b overexpression effect on prostate cancer cell line: time course
Analysis of LNCaP prostate cancer (PCa) cells overexpressing miRNA-135b for up to 36 hours. LNCaP cells express the androgen receptor (AR). MiRNA-135b overexpression in AR+ PCa cells results in slower growth compared to AR knockdown. Results provide insight into the basis of this slower growth.
Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 protocol, 3 time sets
Platform: GPL10558 Series: GSE57820 12 Samples
Download data
DataSet Accession: GDS6100 ID: 6100
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

Histone demethylase KDM3A-deficiency effect on estrogen-stimulated breast cancer cells in vitro
Analysis of estrogen receptor (**ER**-positive breast cancer cell line MCF-7 depleted for KDM3A (histone lysine demethylase 3A) then treated with estrogen. Histone lysine methylation is an important regulator of transcription. Results provide insight into role of KDM3A in ER signaling in **breast cancer**.
Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 2 agent, 2 genotype/variation sets
Platform: GPL10558 Series: GSE68918 11 Samples
Download data
DataSet Accession: GDS5662 ID: 5662
PubMed Full text in PMC Similar studies GEO Profiles Analyze DataSet

GEO Datasets – Selected Title

- GSM: Sample level
- GDS: Dataset
- GSE: Series
- GPL: Platform

The screenshot shows the NCBI GEO Dataset Browser interface. At the top, there is a header with the NCBI logo, the word "CURATED", and the GEO logo. Below the header, a search bar contains the identifier "GDS6100[ACCN]". There are buttons for "Search", "Clear", "Show All", and "Advanced Search". The main content area displays a "DataSet Record" for GDS6100. The record includes fields for Title, Summary, Organism, Platform, Citation, Reference Series, Sample count, Value type, and Series published date. To the right of the record, there is a "Cluster Analysis" section showing a heatmap and a "Download" section with links to various file formats. Below the record, there is a "Data Analysis Tools" section with options for finding genes, comparing samples, and creating cluster heatmaps.

DataSet Record GDS6100: [Expression Profiles](#) [Data Analysis Tools](#) [Sample Subsets](#)

Title:	MicroRNA-135b overexpression effect on prostate cancer cell line: time course		
Summary:	Analysis of LNCaP prostate cancer (PCa) cells overexpressing miRNA-135b for up to 36 hours. LNCaP cells express the androgen receptor (AR). MiRNA-135b overexpression in AR+ PCa cells results in slower growth compared to AR knockdown. Results provide insight into the basis of this slower growth.		
Organism:	<i>Homo sapiens</i>		
Platform:	GPL10558: Illumina HumanHT-12 V4.0 expression beadchip		
Citation:	Aakula A, Leivonen SK, Hintsanen P, Alttokallio T et al. MicroRNA-135b regulates ER α , AR and HIF1AN and affects breast and prostate cancer cell growth. <i>Mol Oncol</i> 2015 Aug;9(7):1287-300. PMID: 25907805		
Reference Series:	GSE57820	Sample count:	12
Value type:	transformed count	Series published:	2015/04/21

Data Analysis Tools

Find genes [?](#)

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol:

Find genes that are up/down for this condition(s): protocol time

GEO Datasets – Select Reference Series

 NCBI

 GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO | Not logged in | Login

NCBI > GEO > Accession Display

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSE57820 | GO

Series GSE57820

Status: Public on Apr 21, 2015

Title: The effect of miRNA-135b overexpression on the gene expression profile of LNCaP cells

Organism: Homo sapiens

Experiment type: Expression profiling by array

Summary: MicroRNAs (miRNAs) regulate a wide range of cellular signaling pathways and biological processes in both physiological and pathological states such as cancer. We have previously identified miR-135b as a direct regulator of androgen receptor (AR) protein level in prostate cancer (PCa). We wanted to further explore the relationship of miR-135b to hormonal receptors, particularly estrogen receptor α (ER α). Here we show that miR-135b expression inversely correlates with ER α protein in two independent breast cancer (BCa) patient cohorts (101 and 1302 samples) and with AR protein in 47 PCa patient samples. We identify ER α as a novel miR-135b target by demonstrating miR-135b binding to the 3'UTR of the ER α and decreased ER α protein and mRNA level in breast cancer cells upon miR-135b overexpression. miR-135b inhibits proliferation of hormone receptor positive cancer cell lines as shown by overexpression in ER α -positive BCa cells (MCF-7) and AR-positive PCa cells (LNCaP, 22Rv1) when grown in 2D. To identify other genes regulated by miR-135b we performed gene expression studies and found a potential link to the hypoxia-inducible factor-1 α (HIF1 α) pathway. We show that miR-135b influences the protein level of the inhibitor for hypoxia-inducible factor-1 (HIF1AN), which also demonstrated an inverse correlation with miR-135b in a cohort of breast tumor samples. Taken together, our study demonstrates that miR-135b regulates ER α , AR and HIF1AN protein levels and proliferation in ER α -positive breast and AR-positive-prostate cancer cells.

Overall design: LNCaP cells were transfected with Ambion pre-miR™ construct for miR-135b or with pre-miR negative control #1 (scrambled pre-miR, Scr) at 20 nM, and incubated for 12h, 24h or 36h, in two biological repeats (B1 and B2)

Contributor(s): Aakula A, Leivonen SK, Hintsanen P, Aittokallio T, Ceder Y, Børresen-Dale A.

Citation(s): Aakula A, Leivonen SK, Hintsanen P, Aittokallio T et al. MicroRNA-135b regulates ER α , AR and HIF1AN and affects breast and prostate cancer cell growth. *Mol Oncol* 2015 Aug;9(7):1287-300. PMID: 25907805

Submission date: May 20, 2014
 Last update date: Aug 13, 2018
 Contact name: Anna Aakula
 E-mail: anna.aakula@fimm.fi
 Organization name: Institute for Molecular Medicine Finland, FIMM
 Street address: Tuholmankatu 8
 City: HELSINKI
 ZIP/Postal code: 00290
 Country: Finland

Platforms (1): GPL10558 Illumina HumanHT-12 V4.0 expression beadchip
 Samples (12):
 GSM1394594 LNCaP_miR-135b_12h_B1
 GSM1394595 LNCaP_miR-135b_12h_B2
 GSM1394596 LNCaP_miR-135b_24h_B1

Relations: BioProject: PRJNA248178

Analyze with GEO2R

SOFT: Simple Omnibus Format in Text Has both array data and meta data

Format
SOFT
MINIML
TXT

Supplementary file	Size	Download	File type/resource
GSE57820_RAW.tar	26.2 Mb	(http)(custom)	TAR
GSE57820_non_normalized.txt.gz	3.6 Mb	(ftp)(http)	TXT

Raw data is available on Series record

Processed data included within Sample table

SOFT Dataset

Three screenshots of software interface windows showing SOFT dataset files:

- GSE57820_family.soft**: A text-based file showing metadata for a GEO dataset. Lines 1-33 show general database information like name, institute, and email. Lines 40-80 show series-level details such as title, accession, status, submission date, last update date, pubmed ID, and summary.
- GSE57820_family.soft**: A text-based file showing platform metadata. Lines 50-75 describe the platform as GPL10558 (Illumina), its status as public, and its submission date as May 20 2014. Lines 107-125 provide detailed annotations for the probe, including its genomic coordinates, cytoband, and various ontology terms.
- GSE57820_family.soft**: A text-based file showing detailed probe annotations. Lines 107-125 continue from the previous window, providing specific details for each probe entry, including its ID, array address, probe type, start position, chromosome, orientation, and various ontology terms.

R/GEOquery

```
library(GEOquery)
exampleData <- getGEO("/path/file.soft.gz")
> exampleData <- getGEO("GSE57820")
Found 1 file(s)
GSE57820_series_matrix.txt.gz
trying URL 'https://ftp.ncbi.nlm.nih.gov/geo/series/GSE57nnn/GSE57820/matrix/GSE57820_series_matrix.txt.gz'
Content type 'application/x-gzip' length 4575066 bytes (4.4 MB)
downloaded 4.4 MB

Parsed with column specification:
cols(
  ID_REF = col_character(),
  GSM1394594 = col_double(),
  GSM1394595 = col_double(),
  GSM1394596 = col_double(),
  GSM1394597 = col_double(),
  GSM1394598 = col_double(),
  GSM1394599 = col_double(),
  GSM1394600 = col_double(),
  GSM1394601 = col_double(),
  GSM1394602 = col_double(),
  GSM1394603 = col_double(),
  GSM1394604 = col_double(),
  GSM1394605 = col_double()
)
File stored at:
C:\Users\vander11\AppData\Local\Temp\RtmpAXemR3/GPL10558.soft
> summary(exampleData)
Length Class Mode
GSE57820_series_matrix.txt.gz 1 ExpressionSet S4
> |> exampleData <- exampleData[[1]]
> exampleData
ExpressionSet (storageMode: lockedEnvironment)
assayData: 47323 features, 12 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM1394594 GSM1394595 ... GSM1394605 (12 total)
  varLabels: title geo_accession ... transfected with:ch1 (45 total)
  varMetadata: labelDescription
featureData
  featureNames: ILMN_1343291 ILMN_1343295 ... ILMN_3311190 (47323 total)
  fvarLabels: ID Species ... GB_ACC (30 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL10558
```

R/GEOquery

```
> exprs <- exprs(exampleData) # matrix of expression  
> dim(exprs)  
[1] 47323    12    47,323 probe sets & 12 samples  
> pdata <- pData(exampleData) # data.frame of meta data  
> dim(pdata)  
[1] 12 45      45 phenotypes for 12 samples
```

- You can use this data moving forward
- CHECK: with the array data you won't have missing data, but you may have a lot with phenotype data
- Names are super detailed and redundant, may want to change before analyses

```
> table(pdata$characteristics_ch1.5)  
transfected with: Ambion pre-miR negative control #1 (scrambled pre-miR, Scr) at 20 nM  
6  
transfected with: Ambion pre-miRâ„¢ construct for miR-135b at 20 nM  
6  
> table(pdata$characteristics_ch1.6)  
incubation time: 12h incubation time: 24h incubation time: 36h  
4           4           4  
`
```

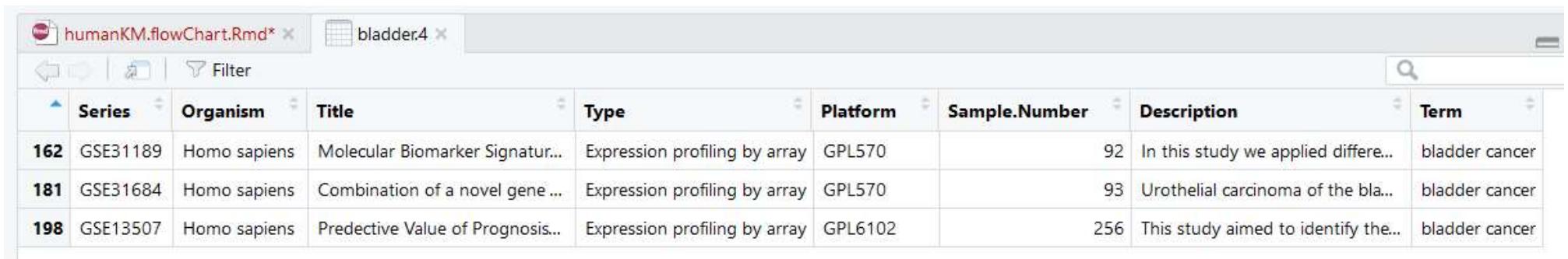
Browsing GEO in R

```
library(GEOquery)
library(GEOsearch)
library(DBI)
library(RSQLite)
library(GEOMetadb)
```

- Example Search/Filtering Criteria
1. Search "bladder cancer"
 2. Limit to humans
 3. Only specific array platforms
 4. Must have >75 samples

```
bladder.1 = GEOSearchTerm("bladder cancer")
bladder.2 = bladder.1[which(bladder.1$Organism=="Homo sapiens"),]
arraysWant = c("GPL6102", "GPL570", "GPL571")
bladder.3 = bladder.2[which(bladder.2$Platform %in% arraysWant),]
bladder.4 = bladder.3[which(bladder.3$Sample.Number>75),]
```

Browsing GEO in R



The screenshot shows the RStudio interface with two tabs open: "humanKM.flowChart.Rmd*" and "bladder.4". The "bladder.4" tab displays a data grid with the following columns: Series, Organism, Title, Type, Platform, Sample.Number, Description, and Term. The data includes three rows of bladder cancer datasets from GSE31189, GSE31684, and GSE13507.

	Series	Organism	Title	Type	Platform	Sample.Number	Description	Term
162	GSE31189	Homo sapiens	Molecular Biomarker Signatur...	Expression profiling by array	GPL570	92	In this study we applied differe...	bladder cancer
181	GSE31684	Homo sapiens	Combination of a novel gene ...	Expression profiling by array	GPL570	93	Urothelial carcinoma of the bla...	bladder cancer
198	GSE13507	Homo sapiens	Predictive Value of Prognosis..	Expression profiling by array	GPL6102	256	This study aimed to identify the...	bladder cancer

```
> dim(bladder.4)
[1] 3 8
> colnames(bladder.4)
[1] "Series"          "Organism"        "Title"           "Type"            "Platform"
[5] "Sample.Number"   "Description"    "Term"
> bladder.4$Series
[1] "GSE31189" "GSE31684" "GSE13507"
>
```

Once you identify the GEO datasets you want, use `getGEO()` for analyses

SRA Database

- NCBI
- Very similar to GEO except this is all sequencing data
 - RNA-Seq
 - DNA-Seq
 - ChIP-Seq
- Notice initial search gives sample results

SRA SRA ER positive breast cancer
Create alert Advanced

Access Summary ▾ 20 per page ▾ Send
Controlled (50)
Public (523)

Source View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)
DNA (181)
RNA (392)

Other Search results
aligned data (141)

[Clear all](#)
[Show additional filters](#)

Items: 1 to 20 of 573 << First < Prev Page 1 of 29 Next > Last

[GSM1915703: MCF7_Cntrl_3; Homo sapiens; RNA-Seq](#)
1. 2 ILLUMINA (Illumina HiSeq 2500) runs: 75.4M spots, 3.7G bases, 2.3Gb downloads
Accession: SRX1363845

[GSM1915702: MCF7_Cntrl_2; Homo sapiens; RNA-Seq](#)
2. 2 ILLUMINA (Illumina HiSeq 2500) runs: 26.8M spots, 1.3G bases, 681.9Mb downloads
Accession: SRX1363844

[GSM1915701: MCF7_Cntrl_1; Homo sapiens; RNA-Seq](#)
3. 2 ILLUMINA (Illumina HiSeq 2500) runs: 18.6M spots, 922.6M bases, 482.5Mb downloads
Accession: SRX1363843

[GSM1915700: MCF7_100nM_DAC_3; Homo sapiens; RNA-Seq](#)
4. 2 ILLUMINA (Illumina HiSeq 2500) runs: 42.8M spots, 2.1G bases, 1.1Gb downloads
Accession: SRX1363842

[GSM1915699: MCF7_100nM_DAC_2; Homo sapiens; RNA-Seq](#)
5. 2 ILLUMINA (Illumina HiSeq 2500) runs: 38.1M spots, 1.9G bases, 1Gb downloads
Accession: SRX1363841

SRA Database

- SRR: Individual run in the experiment
- SRS: sample information
- SRX: experiment level
- SRP: project level
- Many still linked to GEO

[SRX1363845](#): [GSM1915703](#): MCF7_Cntrl_3; Homo sapiens; RNA-Seq
2 ILLUMINA (Illumina HiSeq 2500) runs: 75.4M spots, 3.7G bases, 2.3Gb downloads

Submitted by: NCBI (GEO)

Study: RNA-seq of YB5 and MCF7 treated with different doses of decitabine
[PRJNA299580](#) • [SRP065220](#) • [All experiments](#) • [All runs](#)
[show Abstract](#)

Sample: MCF7_Cntrl_3
[SAMN04202263](#) • [SRS1126530](#) • [All experiments](#) • [All runs](#)
Organism: [Homo sapiens](#)

Library:
Instrument: Illumina HiSeq 2500
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Construction protocol: RNA was isolated using Rneasy Mini Kit (Qiagen) Strand-specific of RNA using TruSeq stranded total RNA with Ribo-Zero Gold (Illumina)

Experiment attributes:
GEO Accession: GSM1915703

Links:

Runs: 2 runs, 75.4M spots, 3.7G bases, 2.3Gb

Run	# of Spots	# of Bases	Size	Published
<u>SRR2753169</u>	37,479,476	1.9G	1.1Gb	2017-01-31
<u>SRR2753170</u>	37,904,723	1.9G	1.1Gb	2017-01-31

SRA Database

NCBI Site map All databases Search

 Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

RNA-seq of YB5 and MCF7 treated with different doses of decitabine

Identifiers:

SRA:	SRP065220
BioProject:	PRJNA299580
GEO:	GSE74251

Study Type:

Transcriptome Analysis

Abstract:

RNA-seq was performed after YB5 cells were treated with 1uM decitabine, and MCF7 cells were treated with 100nM decitabine Overall design: Biological triplicates were performed for a total of 6 samples. Fold change of each gene was calculated by comparing change in expression after inhibitor treatment to expression in the control samples from GSE73966 for YB5, and GSE74036 for MCF7. These control samples have been re-accessioned here for convenient access to the entire study.

Center Project:

GSE74251

External Link: [Transcriptional Selectivity of Epigenetic Therapy in Cancer.](#)

Related SRA data

Experiments: [12](#) (12 samples)
Runs: [21](#) (25.4Gbp; 14.2Gb)

NCBI Site map All databases Search

Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST

Overview

The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.

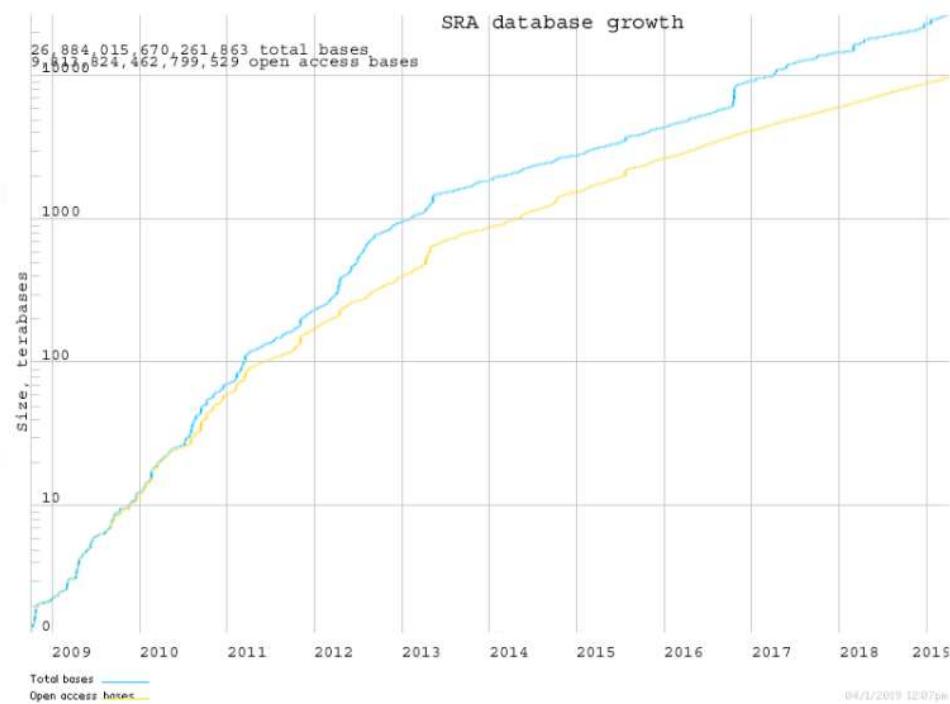
SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

Please check [SRA Overview](#) for more information.

Submitting to SRA

Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- [Submission Quick Start](#)
- [Frequently Asked Questions and Troubleshooting](#)
- [Log in to Submission Portal](#) (for submitting sequence data)
- [Log in to SRA](#) (for updating and troubleshooting submissions)



NCBI Site map All databases Search

 Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive Trace Assembly Trace BLAST

Studies Samples Analyses Run Browser Run Selector Provisional SRA

Search: Go ? What can be entered in this field?

List of SRA Samples. 4916785 found.

#	Accession	Organism	Title	Attributes
1.	SRS4512763	Homo sapiens	T-9_R1.fastq	isolate: T-9-R1 age: missing biomaterial_provider: david.tulasne@jbl.cnrs.fr sex: - tissue: LUNG cell_line: - cell_subtype: - cell_type: - culture_collection: - dev_stage: - disease: CANCER disease_stage: - ethnicity: CAUCASIAN health_state: - karyotype: - phenotype: - population: - race: - sample_type: TISSUE treatment: -
2.	SRS4512764	Homo sapiens	T-8_R1.fastq	isolate: T-8-R1 age: missing

TCGA

1-800-4-CANCER Live Chat Publications Dictionary

ABOUT CANCER CANCER TYPES RESEARCH GRANTS & TRAINING NEWS & EVENTS ABOUT NCI search Q

Home > About NCI > NCI Organization > CCG > Research > Structural Genomics

TCGA

- Program History
- TCGA Cancers Selected for Study
- Publications by TCGA
- Using TCGA
- Contact

The Cancer Genome Atlas Program

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already lead to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use.

TCGA Outcomes & Impact

TCGA has changed our understanding of cancer, how research is conducted, how the disease is treated in the clinic, and more.

TCGA's PanCancer Atlas

A collection of cross-cancer analyses delving into overarching themes on cancer, including cell-of-origin patterns, oncogenic processes and signaling pathways. Published in 2018 at the program's close.

TCGA – Data Portal

https://portal.gdc.cancer.gov

Visualize copy number variations in the data portal! Details in the User's Guide

Dismiss X

NIH NATIONAL CANCER INSTITUTE GDC Data Portal Home Exploration Analysis Repository Quick Search Manage Sets Login Cart GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Exploration Analysis Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Repository button circled in red

Data Portal Summary Data Release 16.0 - March 26, 2019

PROJECTS	PRIMARY SITES	CASES
45	68	33,549
FILES	GENES	MUTATIONS
365,463	22,872	3,142,246

Human body diagram showing cancer locations

Cases by Major Primary Site

Primary Site	Cases
Adrenal Gland	~100
Bile Duct	~50
Bladder	~1,000
Blood	~1,000
Bone	~500
Bone Marrow	~500
Brain	~1,000
Breast	~3,500
Cervix	~500
Colorectal	~2,500
Esophagus	~500
Eye	~100
Head and Neck	~1,000
Kidney	~2,000
Liver	~1,000
Lung	~4,500
Lymph Nodes	~500
Nervous System	~2,000
Ovary	~1,500
Pancreas	~500
Pleura	~100
Prostate	~500
Skin	~1,000
Soft Tissue	~100
Stomach	~1,000
Testis	~100
Thymus	~100
Thyroid	~500
Uterus	~1,000

NATIONAL CANCER INSTITUTE
GDC Data Portal

Files Cases Add a File Filter

Start searching by selecting a facet Advanced Search

Add All Files to Cart Manifest View 33,549 Cases in Exploration View Images Browse Annotations

Files (365,463) Cases (33,549) 698.31 TB

Primary Site Project Data Category Data Type Data Format

Show More

Showing 1 - 20 of 365,463 files

Access File Name Cases Project Data Category Data Format File

80c35112-
0adb-472
d-94cd-fc8

The screenshot shows the GDC Data Portal interface. On the left, there are facets for 'File' (e.g., 142682.bam), 'Data Category' (Simple Nucleotide Variation, Transcriptome Profiling, Biospecimen, Sequencing Reads, Copy Number Variation), and 'Data Type' (Annotated Somatic Mutation, Raw Simple Somatic Mutation, Aligned Reads). The main area displays summary statistics: 365,463 files, 33,549 cases, and 698.31 TB of data. It features five pie charts representing Primary Site, Project, Data Category, Data Type, and Data Format. Below this is a table showing the first 20 files from a total of 365,463, with columns for Access, File Name, Cases, Project, Data Category, Data Format, and File. The 'File Name' column shows entries like '80c35112-' and '0adb-472'. At the bottom, there are links for 'Manage Sets', a shopping cart icon with '0', and a grid icon.

R/TCGAbiolinks

TCGAbiolinks

```
browseVignettes("TCGAbiolinks")
```

DOI: [10.18129/B9.bioc.TCGAbiolinks](https://doi.org/10.18129/B9.bioc.TCGAbiolinks)



TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

[HTML](#) [R Script](#) 1. Introduction

[HTML](#) [R Script](#) 10. TCGAbiolinks_Extension

[HTML](#) [R Script](#) 2. Searching GDC database

[HTML](#) [R Script](#) 3. Downloading and preparing files for analysis

[HTML](#) [R Script](#) 4. Clinical data

[HTML](#) [R Script](#) 5. Mutation data

[HTML](#) [R Script](#) 6. Compilation of TCGA molecular subtypes

[HTML](#) [R Script](#) 7. Analyzing and visualizing TCGA data

[HTML](#) [R Script](#) 8. Case Studies

[HTML](#) [R Script](#) 9. Graphical User Interface (GUI)

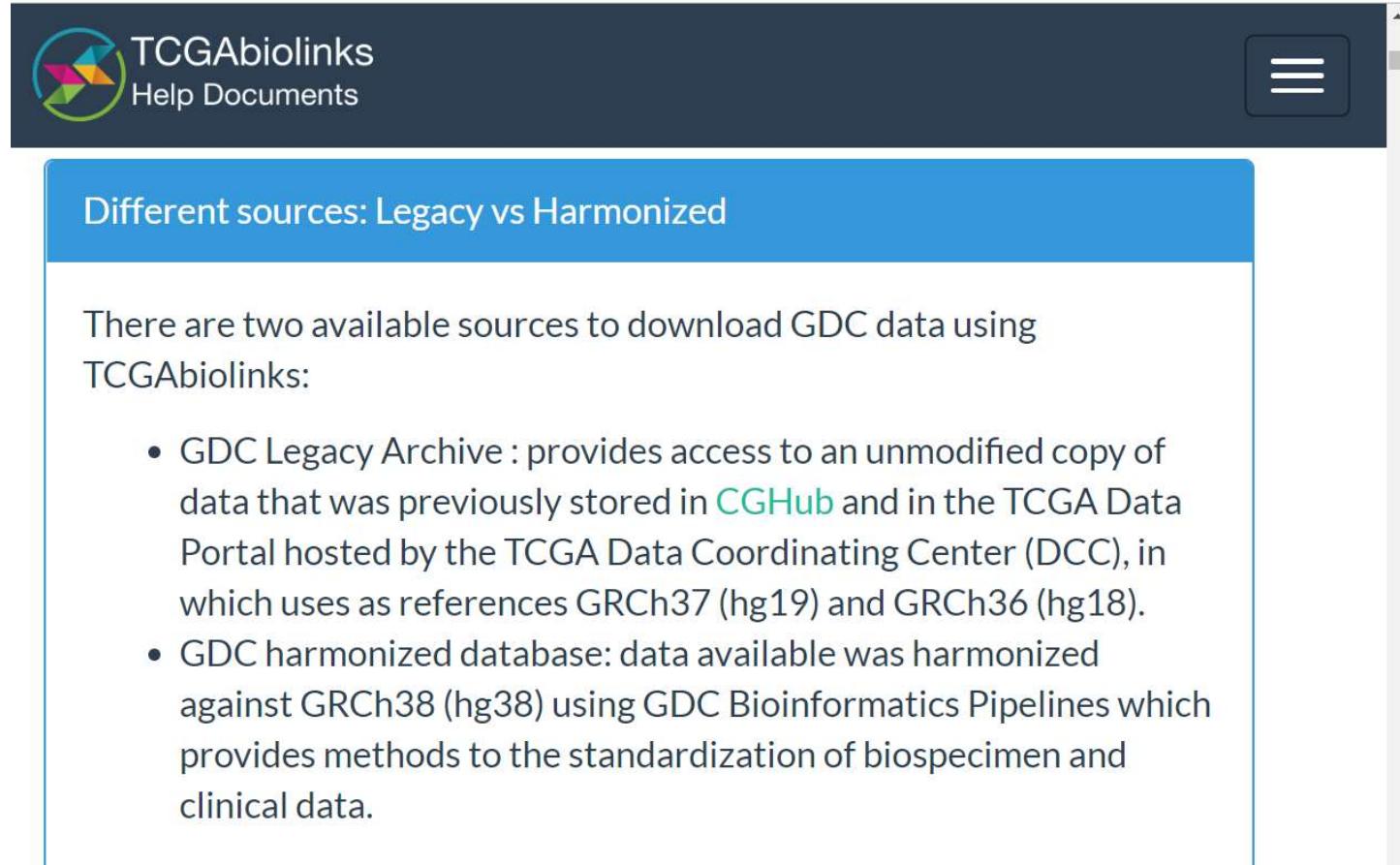
[PDF](#) Reference Manual

[Text](#) NEWS

- Naming system:
- Aliquot barcode: TCGA-G4-6317-02A-11D-2064-05
- Participant: TCGA-G4-6317
- Sample: TCGA-G4-6317-02

TCGA Legacy

- Legacy older genomes (hg19 or hg18)
- Harmonized database have standards for specimen and clinical data



The screenshot shows a dark-themed web page header with the TCGAbiolinks logo and "Help Documents". On the right is a menu icon. Below the header, a blue box contains the title "Different sources: Legacy vs Harmonized". The main content area has a white background and contains text about two available sources for GDC data: the GDC Legacy Archive and the GDC harmonized database.

Different sources: Legacy vs Harmonized

There are two available sources to download GDC data using TCGAbiolinks:

- GDC Legacy Archive : provides access to an unmodified copy of data that was previously stored in [CGHub](#) and in the TCGA Data Portal hosted by the TCGA Data Coordinating Center (DCC), in which uses as references GRCh37 (hg19) and GRCh36 (hg18).
- GDC harmonized database: data available was harmonized against GRCh38 (hg38) using GDC Bioinformatics Pipelines which provides methods to the standardization of biospecimen and clinical data.

GDCquery()

```
query <- GDCquery(project = c("TCGA-GBM", "TCGA-LGG"),  
                      data.category = "DNA Methylation",  
                      legacy = FALSE,  
                      platform = c("Illumina Human Methylation 450")  
                      sample.type = "Recurrent Solid Tumor"  
)
```

- This is dependent on SQL databases and can be computationally intensive

GTEx

<https://gtexportal.org/home/>

- Human tissue information on NON-DISEASED individuals
- Post mortem samples
- Quantitative Trait Loci (QTL)

The screenshot shows the GTEx Portal homepage. At the top, there's a navigation bar with links for Home, Datasets, Expression, QTLs & Browsers, Sample Data, Documentation, About GTEx, Publications, Access Biospecimens, FAQs, Contact, a search bar, and a sign-in button.

A prominent feature is a news card in the center-right: "2019-03-07 New Histology Image Viewer. Check out our new revised histology image viewer. The streamlined interface makes it easy to search by sample identifiers. The advanced search feature allows you to easily search for histology images by donor phenotype or tissue type, but also hide..."

The main content area has two main sections:

- Resource Overview**: Contains links to Current Release (V7), Tissue & Sample Statistics, Tissue Sampling Info (Anatomogram), Access & Download Data, Release History, and How to cite GTEx? It also includes a detailed description of the GTEx project and its data collection.
- Explore GTEx**: Contains four categories: "Browse" (By gene ID, By variant or rs ID, By Tissue, Histology Image Viewer), "Expression" (Multi-Gene Query, Top 50 Expressed Genes), and a "Sample Data" section.

GTEx

GTEx Analysis V7 (dbGaP Accession phs000424.v7.p2)

The GTEx Analysis V7 release is the most current release of the GTEx Portal.

Protected Data and Raw Data

Due to the nature of our donor consent agreement, raw data and attributes which might be used to identify the donors are not publicly available on the GTEx Portal. You may apply for access to the data through dbGaP [↗](#).

Data available include:

- BAM files for RNA-Seq, Whole Exome Seq, and Whole Genome Seq
- Genotype Calls (.vcf) for OMNI SNP Arrays, WES, and WGS
- OMNI SNP Array Intensity files (.idat and .gtc)
- Affymetrix Expression Array Intensity files (.cel)
- Allele Specific Expression (ASE) tables
- All expression matrices from the Portal, including samples that did not pass the Analysis Freeze QC
- Sample Attributes
- Subject Phenotypes

On dbGaP, the VCF used for eQTL analyses are in the following archives (under 'Genotype Files'):

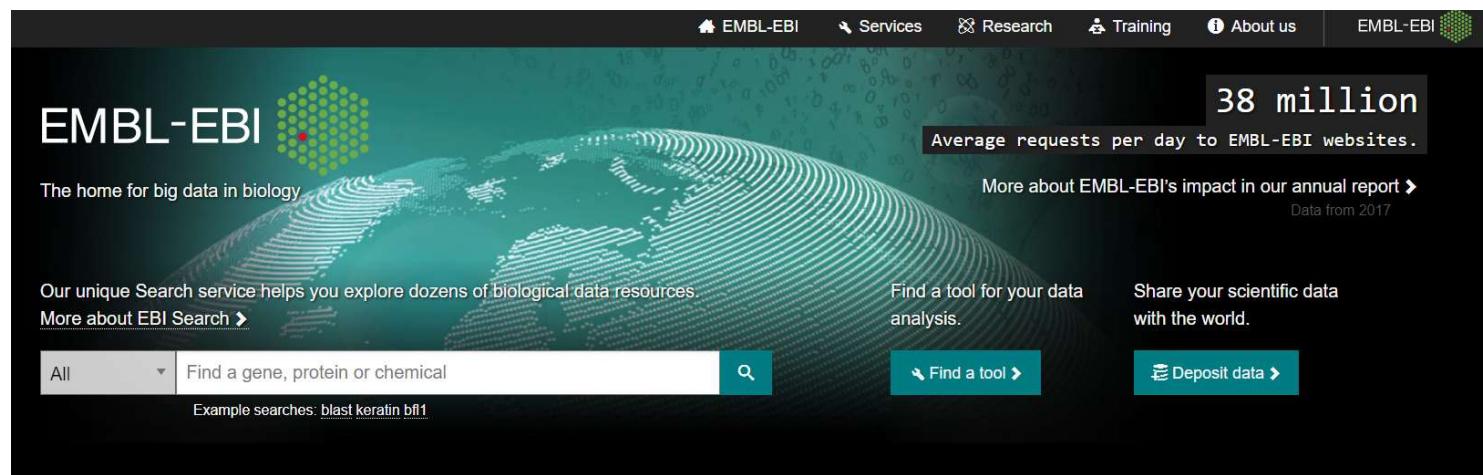
- V6p release: phg000520.v2.GTEx_MidPoint_Imputation.genotype-calls-vcf.c1.GRU.tar
- V7 release: phg000830.v1.GTEx_WGS.genotype-calls-vcf.c1.GRU.tar

GTEx RNA-Seq Downloads

RNA-Seq Data		
Description	Name	Size
Gene read counts.	GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_reads.gct.gz	496M
Gene TPMs.	GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz	827M
This file contains the median TPM by tissue. These medians were calculated directly from the file GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz.	GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz	4.8M
Junction read counts.	GTEx_Analysis_2016-01-15_v7_STARv2.4.2a_junctions.gct.gz	2.2G
Transcript read counts.	GTEx_Analysis_2016-01-15_v7_RSEMv1.2.22_transcript_expected_count.txt.gz	2.3G
Transcript TPMs.	GTEx_Analysis_2016-01-15_v7_RSEMv1.2.22_transcript_tpm.txt.gz	1.9G
Exon read counts.	GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_exon_reads.parquet	7.0G

European Bioinformatics Institute

<https://www.ebi.ac.uk/>



We are EMBL-EBI

The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact. ➤

Training

Access a wealth of world-leading training in bioinformatics and scientific service provision.

Data resources

Explore our open data resources to enrich your research. Browse data, perform analyses or share your own results. ➤

Research

Find out about our research groups, postdoctoral schemes and PhD Programme ➤

Industry

Explore our knowledge-exchange Industry Programme and take part in translational partnerships and

ELIXIR

We support, as an ELIXIR node, the coordination of biological data provision throughout Europe ➤

EBI Expression Atlas

Browse and download in R:
R/ExpressionAtlas

Home Browse experiments Download Release notes FAQ Help Licence About Support

Query single cell expression
To Single Cell Expression Atlas

Experiment	Assays	Comparisons	Species	Experimental Variables	Array Designs	ArrayExpress
RNA-seq of 20 Theobroma cacao cultivars	20		Theobroma cacao	cultivar		
RNA-seq of banana fruit development	5		Musa acuminata AAA Group	developmental stage		
RNA-seq of a E-type and a Z-type cultivar of sugar beet at five different developmental stages	10		Beta vulgaris subsp. vulgaris	cultivar developmental stage		
RNA-seq of the olfactory system of newborn mice	12		Mus musculus	organism part		

Experiments in Expression Atlas

Kingdom:	Search all columns:					
All						
All						
Plants						
Animals						
Fungi						
Experiment	Assays	Comparisons	Species	Experimental Variables	Array Designs	ArrayExpress
06-12-2017 RNA-seq of 20 Theobroma cacao cultivars	20		Theobroma cacao	cultivar		
04-01-2018 RNA-seq of banana fruit development	5		Musa acuminata AAA Group	developmental stage		
06-12-2017 RNA-seq of a E-type and a Z-type cultivar of sugar beet at five different developmental stages	10		Beta vulgaris subsp. vulgaris	cultivar developmental stage		
13-05-2016 RNA-seq of the olfactory system of newborn mice	12		Mus musculus	organism part		

Gene Network

- Extensive database on rodents
- Group is panel or cohort
- Type: Phenotype or Tissue Specific (Adipose)
- Dataset: Specific experiment
- Get Any and Combined is a way to filter the previous selected fields

<http://www.genenetwork.org>

GeneNetwork
University of Tennessee: www.genenetwork.org [Use GeneNetwork 2](#)

| Home | Search | Help | News | References | Policies | Links |

Select and Search

Species: [Info](#)

Group: [Info](#)

Type:

Data Set:

- AIL LGSM F34 (Array)
- AIL LGSM F34 (GBS)
- AIL LGSM F34 and F39-43 (GBS)
- AIL LGSM F39-43 (GBS)
- AKXD RI Family
- AXB/BXA RI Family
- B6BTBRF2
- B6D2 EtOH Selected Advanced Intercross
- B6D2F2
- B6D2F2 PSU
- BDF2 UCLA
- BDF2-2005

Get Any:

Combined:

Quick HELP Examples

You can also use advanced search operators such as AND, OR, NOT, etc. into the Get Any or Combined fields.

- **POSITION=(chr1|chr2)**
- **MEAN_GENE=100**

[Advanced Search](#)

RGD

<https://rgd.mcw.edu/wg/data-menu/>

The screenshot shows the RGD website's homepage. At the top, there is a navigation bar with links to HOME, DATA (which is highlighted in blue), ANALYSIS & VISUALIZATION, DISEASES, PHENOTYPES & MODELS, GENETIC MODELS, PATHWAYS, and COMMUNITY. Below the navigation bar, there is a sub-navigation menu with links to Genes, QTLs, Strains, Markers, Genome Information, Ontologies, Cell Lines, References, FTP Download, and Submit Data.

RGD Data

RGD stores data about various "objects". Users can find all the associated data available for an object by clicking on a category name or an icon to begin an object-specific search or browse the available data.

GENES Gene reports include a comprehensive description of function and biological process as well as disease, expression, regulation and phenotype information.	STRAINS Strain reports include a comprehensive description of strain origin, disease, phenotype, genetics, immunology, behavior with links to related genes, QTLs, sub-strains, and strain sources.
GENOME INFORMATION PAGES RGD's Genome Information pages give consolidated information about the recent genome assemblies for all of the species available at RGD.	ONTOLOGIES Ontologies provide standardized vocabularies for annotating molecular function, biological process, cellular component, phenotype and disease associations. Allows searching across genes, QTLs, strains and provides a basis for cross-species comparisons..
QTLs QTL reports provide phenotype and disease descriptions, mapping, and strain information as well as links to markers and candidate genes.	CELL LINES RGD's Cell Line Directory links to information about, and sources for, rat cell lines, in particular rat embryonic stem cell (ES) lines.
MARKERS SSLP and SNP reports provide mapping data, primer information, and size variations among strains.	REFERENCES Reference reports provide full citations, abstracts, and links to Pubmed. Where available, a link directly to the full text of the article is also provided.

RGD Legacy Data

RGD still stores some data types which are not being updated on a regular basis, including genetic/RH maps and sequences.

MAPS Map reports provide comprehensive marker and map data for RH and genetic maps.	SEQUENCES Sequence reports provide sequence data related to genes, ESTs, and other object types as well as links to reports at NCBI.
---	--

Gene Editing Rat Resource Center

ALLIANCE
of GENOME RESOURCES
FOUNDING MEMBER

PhenoGen

<https://phenogen.org/web/sysbio/resources.jsp>

The screenshot shows the PhenoGen Informatics website interface. At the top, there is a header with the title "PhenoGen Informatics" and a subtitle "The site for quantitative genetics of the transcriptome". Below the header is a search bar with a "Google Custom Search" placeholder and a magnifying glass icon. The main navigation menu includes links for "Home", "Genome / Transcriptome Data Browser", "Available Data Downloads", "Gene List Analysis Tools", "About", "Help", "View Previous Microarray Analysis", and "Account". Below the menu, there are two tabs: "Public Files" and "Members Files", with "Public Files" being the active tab. A note below the tabs says: "Select the download icon (green arrow) to download data from any of the datasets below. For some data types multiple options may be available. For these types, a window displays that allows you to choose specific files." Below this note, there is a horizontal navigation bar with tabs for "RNA-Seq", "DNA-Seq", "Microarray", "Genomic Marker", and "Publications". Under the "RNA-Seq" tab, there is a section titled "New RNA Sequencing Datasets Experimental Details/Downloads". This section contains a table with columns for "Description", "Organism", "Strain", "Tissue", "Seq. Tech.", "RNA Type", "Read Type", "Genome Versions", "Experimental Details", "Raw Data Downloads", and "Result Downloads". The first row of the table is for "Whole Brain from ILS/ISS Parental Strains" and includes a blue download icon for the genome version mm10. Below this table is another section titled "RNA Sequencing BED/BAM Data Files", which contains a similar table for various RNA sequencing datasets across different organisms, strains, tissues, and sequencing technologies.

Description	Organism	Strain	Tissue	Seq. Tech.	RNA Type	Read Type	Genome Versions	Experimental Details	Raw Data Downloads	Result Downloads
Whole Brain from ILS/ISS Parental Strains	Mm	ILS/ISS	Whole Brain	Illumina HiSeq 2000	smallRNA	51bp single-end reads	mm10			

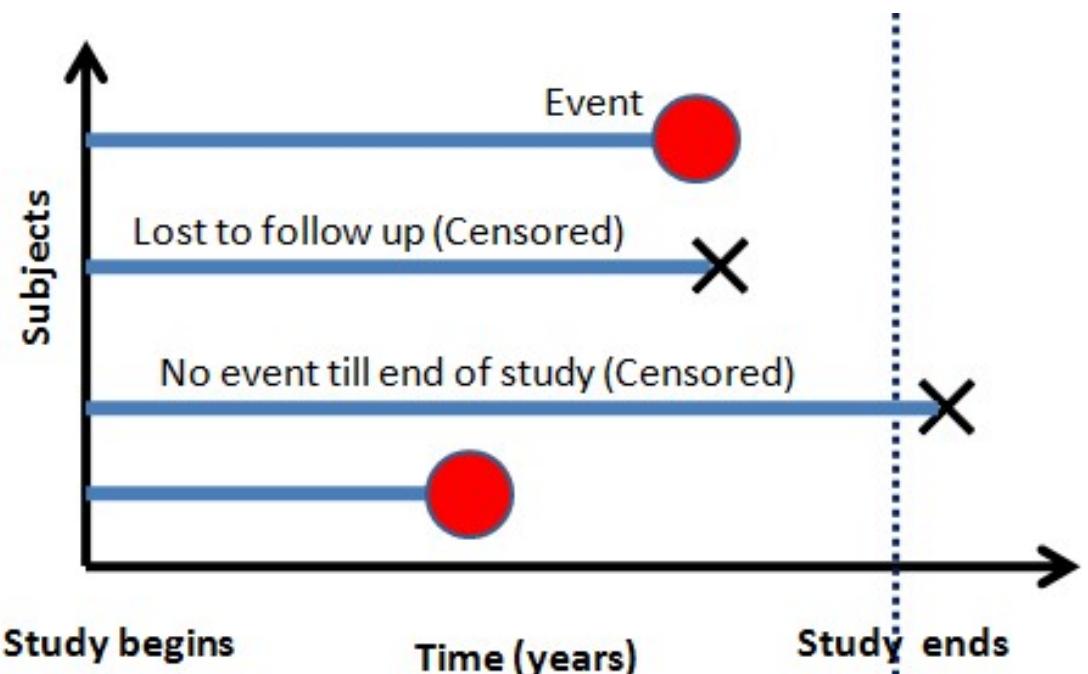
Organism	Strain	Tissue	Seq. Tech.	RNA Type	Read Type	Genome Versions	BED/ BAM Files
Rat	BN-Lx/CubPrin	Brain	Illumina HiSeq2000	polyA+ (>200 nt) selected	100 bp paired-end	Rn6, Rn5	
Rat	SHR/OlaLpcvPrin	Brain	Illumina HiSeq2000	polyA+ (>200 nt) selected	100 bp paired-end	Rn6, Rn5	
Rat	BN-Lx/CubPrin	Brain	Illumina HiSeq2000	total RNA (>200 nt) after ribosomal RNA depletion	100 bp paired-end	Rn6, Rn5	
Rat	SHR/OlaLpcvPrin	Brain	Illumina HiSeq2000	total RNA (>200 nt) after ribosomal RNA depletion	100 bp paired-end	Rn6, Rn5	
Rat	BN-Lx/CubPrin	Brain	Illumina HiSeq2000	small RNA (<200 nt) selected	50 bp single-end	Rn5	
Rat	SHR/OlaLpcvPrin	Brain	Illumina HiSeq2000	small RNA (<200 nt) selected	50 bp single-end	Rn5	
Rat	BN-Lx/CubPrin	Brain	Helicos	total RNA (>200 nt) after ribosomal RNA depletion	~33 bp single-end	Rn5	
Rat	SHR/OlaLpcvPrin	Brain	Helicos	total RNA (>200 nt) after ribosomal RNA depletion	~33 bp single-end	Rn5	
Rat	BN-Lx/CubPrin	Heart	Illumina HiSeq2000	total RNA (>200 nt) after ribosomal RNA depletion	stranded 100 bp paired-end	Rn6	

Validation - Gene Signatures

- Popular in cell culture and animal models
- Do my results translate to in vivo human?
- Very common in cancer
 - Diseased tissue availability
- Example:
 - Differential expression between treatment and placebo in diseased animals
 - Know treatment helps outcome
 - List of candidate genes that are differentially expressed
 - Find a human disease dataset
 - Perform survival analysis on your “gene signature”

Survival Data

- Often used in clinical trials
- Outcome variable is the time until the occurrence of an event of interest
 - Death (overall survival)
 - Disease specific survival (DSS)
 - Disease-free survival
 - Progression-free survival
 - Metastasis-free survival
- Censored data



Source: alphabeta statistics

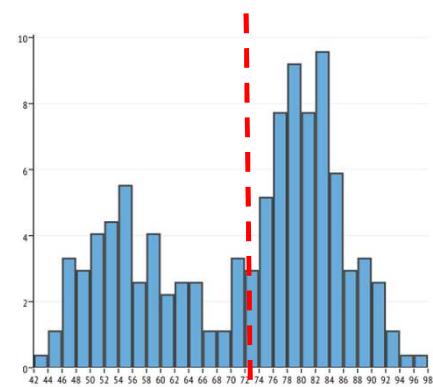
Creating a Gene Signature

1. For each gene, perform a cox proportional hazard model and obtain resulting coefficient.
2. Calculate a score for each subject by summing gene * β coefficient.
3. Classify each subject as having a high/low signature by determining if individuals score is above/below median score value.
4. Perform cox proportional hazard on high/low signature classification.

$$h_i(t) = h_0(t) e^{\{\beta_1 X_{i1} + \dots + \beta_k X_{ik}\}}$$

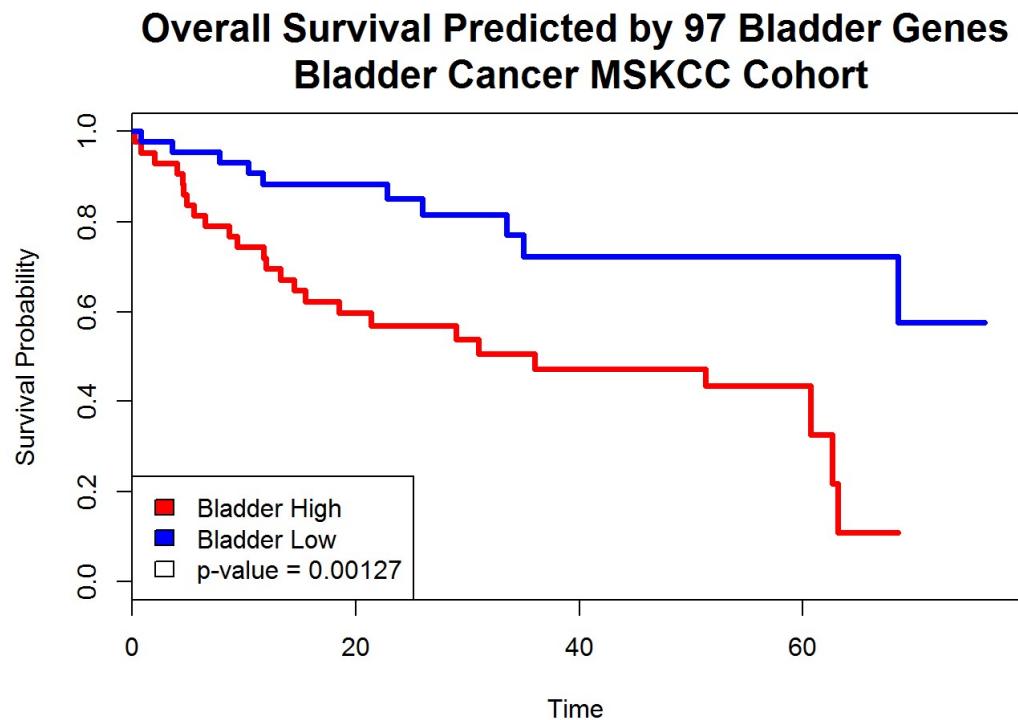
Baseline hazard Linear function of
function set of predictors

$$Score_i = \sum_{k=1}^n \beta_k * X_{ik}$$



Testing Gene Signature

- Results from your cox proportional hazard model
- Visualize using a Kaplan-Meier (KM) plot
- My list of candidate genes predicts survival in a bladder cancer cohort



KM Plotter

<http://kmplot.com/analysis/>

The screenshot shows the KM Plotter website. At the top, there's a dark header bar with the text "Kaplan-Meier Plotter" on the left and "Breast Cancer" with a dropdown menu on the right. Below the header is a navigation bar with links for "KM plotter", "Home", "Download", "Updates", and "Contact". The main content area has a purple background. It features a section titled "What is the KM plotter?" which describes the tool's capabilities. Below this, there are three rows of buttons. The first row is for mRNA gene chip data, showing buttons for breast, ovarian, lung, and gastric cancer. The second row is for mRNA-seq data, showing buttons for liver cancer and pan-cancer, with a note that pan-cancer is in development. The third row is for miRNA data, showing buttons for breast, liver, and pan-cancer.

What is the KM plotter?

The Kaplan Meier plotter is capable to assess the effect of **54,675 genes** on survival using **18,674 cancer samples**. These include **5,143 breast**, **1,816 ovarian**, **2,437 lung**, **364 liver**, **1,065 gastric cancer patients** with relapse-free and overall survival data. The **miRNA subsystems** include additional **11,456 samples** from 20 different cancer types. Primary purpose of the tool is a meta-analysis based **biomarker assessment**.

mRNA gene chip	Start KM Plotter for breast cancer	Start KM Plotter for ovarian cancer	Start KM Plotter for lung cancer	Start KM Plotter for gastric cancer
mRNA RNA-seq	Start KM Plotter for liver cancer	Start KM Plotter for pan-cancer	In development	
miRNA	Start miRpower for breast cancer	Start miRpower for liver cancer	Start miRpower for pan-cancer	

KM Plotter – Selecting Data for Analysis

Kaplan-Meier Plotter | **KM plotter** | **Home** | **Download** | **Updates** | **Contact**

Gastric Cancer

Affy id/Gene symbol: Cifrl112 GCLC STPG1 | Use multigene genes

Split patients by: median | Auto select best cutoff | Trichotomization: none | Compute median survival: | Remember filtering settings: | More options

Survival: OS (n=882) | Censore at threshold: |

Follow up threshold: all | Censore at threshold: |

Probe set options

- user selected probe set
- all probe sets per gene |
- only JetSet best probe set |

Plot beeswarm graph of probe distribution: | |

Show probe expression in normal tissue: | |

Using the selected parameters, the analysis will run on **876** patients. |

Restrict analysis to subtypes...

Stage:	all
Stage T:	all
Stage N:	all
Stage M:	all
Lauren classification:	all
Differentiation:	all

Restrict analysis to clinical cohorts...

Gender:	all
Perforation:	all
Treatment:	all
HER2 status:	all

GSE62254 has markedly different characteristics (longer survival, shifted expression) than the other datasets. We suggest to exclude this dataset when using all samples together.

Use following dataset(s) for the analysis:

all	<input type="checkbox"/> Exclude GSE62254: <input type="checkbox"/>
-----	---

Array quality control: exclude biased arm | |

Please note: the generated p value does **not** include correction for multiple hypothesis testing by default. |

Draw Kaplan-Meier plot:

n = number of patients with available clinical data

Please kindly cite our paper to support further development: Szász AM, Lánczky A, Nagy Á, Förster S, Hark K, Green JE, Boussioutas A, Busuttil R, Szabó A, Győrffy B; Cross-validation of survival associated biomarkers in gastric cancer using transcriptomic data of 1,065 patients, Oncotarget, DOI: 10.18632/oncotarget.10337 |

KM Plotter - Results

Plotter | KM plotter | Home | Download

The desired is valid: 205942_s_at (ACSM3).

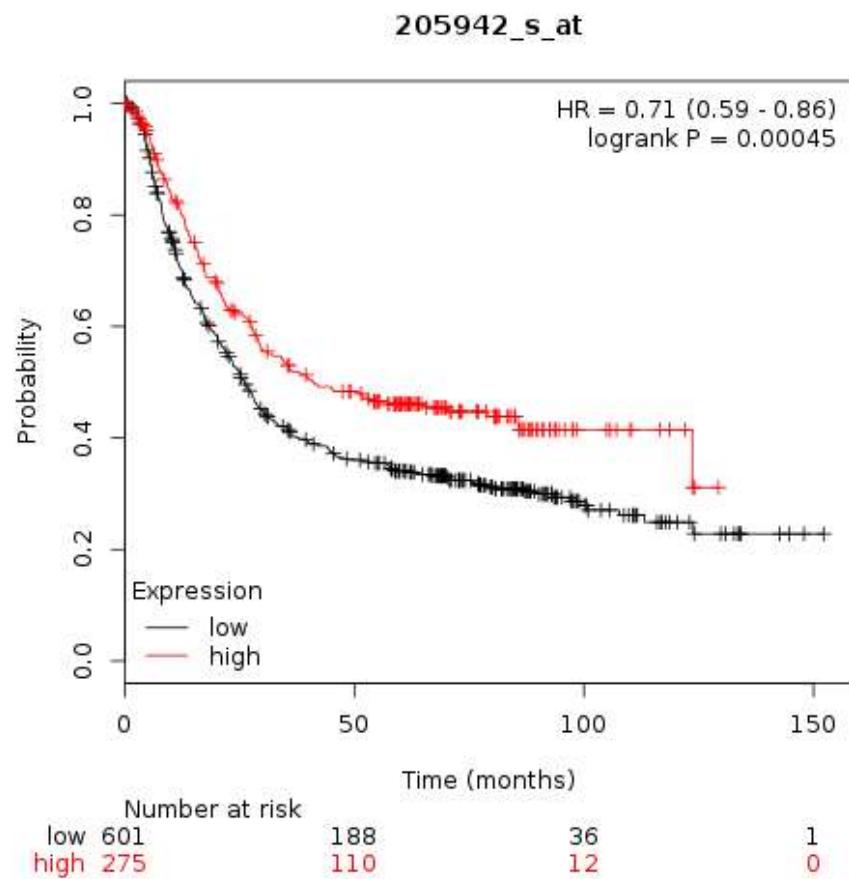
Affy ID: 205942_s_at ACSM3, SA, SAH
Survival: OS
Auto select best cutoff: checked
Follow up threshold: all
Censor at threshold: checked
Compute median over entire database: false
Cutoff value used in analysis: 230
Expression range of the probe: 2 - 2185
Probe set option: user selected probe set
Invert HR values below 1: not checked

Restrictions

Stage: all
Stage T: all
Stage N: all
Stage M: all
Lauren classification: all
Differentiation: all
Gender: all
Perforation: all
Treatment: all
HER2 status: all
Dataset: all
Exclude GSE62254: not checked

Results

P value: 0.0005
FDR: 10%



Gene Signature R Code

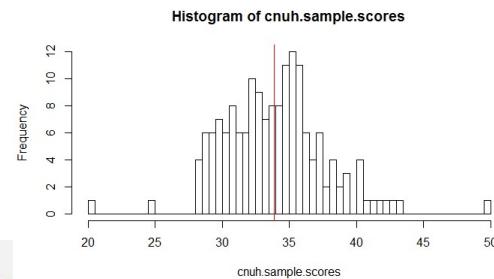
```
load(file="Y:/Sottnik/Data/BladderCancer.CNUH/Korea.Rdata")
load(file="Y:/Sottnik/Reports/v1.analysis/BC97.commonArrayPSwant.Rdata")

# get the probesets to perform Cox proportional-hazard regression on;
cnuh = GSE13507.expr[which(rownames(GSE13507.expr) %in% illumina6.BC97.ps),
-which(is.na(GSE13507.OS.time))]
cnuh.outcome = as.numeric(GSE13507.OS.outcome[!is.na(GSE13507.OS.outcome)])
cnuh.time = as.numeric(GSE13507.OS.time[!is.na(GSE13507.OS.time)])

## STEP1: Cox PH Regression;
library(survival)
library(simPH)

cox.byGene = apply(cnuh, 1, function(a) coxph(Surv(cnuh.time, cnuh.outcome) ~ a)$coefficients)

## STEP2: Get Sample Scores;
cnuh.sample.scores = apply(cnuh, 2, function(a) sum(a*cox.byGene))
hist(cnuh.sample.scores, breaks=50)
abline(v=median(cnuh.sample.scores), col="red")
```



Gene Signature R Code

```
##STEP3: Identify High vs Low Gene Score
cnuh.high = names(cnuh.sample.scores[which(cnuh.sample.scores >
median(cnuh.sample.scores))])
cnuh.low = names(cnuh.sample.scores[which(cnuh.sample.scores <=
median(cnuh.sample.scores))])

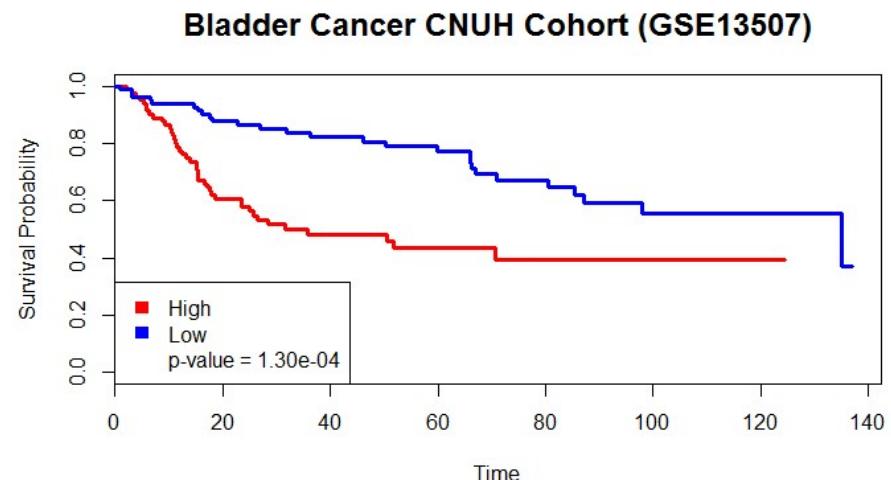
##STEP4: Perform Cox PH & Get Kaplan-Meier Plots;
bladder.signature = c()
for(i in 1:ncol(cnuh)){
  if(colnames(cnuh)[i] %in% cnuh.high){bladder.signature[i]="Bladder
High"}else{bladder.signature[i]="Bladder Low"}
}

cnuh.bladder.surv <- survfit(Surv(cnuh.time, cnuh.outcome)~
as.factor(bladder.signature), conf.type="none")
cnuh.cox <- coxph(Surv(cnuh.time, cnuh.outcome)~ as.factor(bladder.signature))
```

Gene Signature R Code

```
plot(cnuh.bladder.surv, col=c("red", "blue"), xlab="Time", ylab="Survival Probability",
main="Bladder Cancer CNUH Cohort (GSE13507)", lwd=3.5, font=1, font.lab=1,
cex.main=1.5, cex=1, cex.lab=1, cex.axis=1)
par(font=1)
legend("bottomleft", legend=c("High", "Low", paste("p-value = ",
formatC(coef(summary(cnuh.cox))[1,5], format = "e", digits = 2), sep="")),
fill=c("red", "blue", "white"), border="white", col=c("red", "blue", "black")))
```

- Beware of having large number of genes in your candidate list
- KMplots won't take more than 65 genes



References

- <http://www.karlin.mff.cuni.cz/~pesta/NMFM404/survival.html>