Publicly Available Data

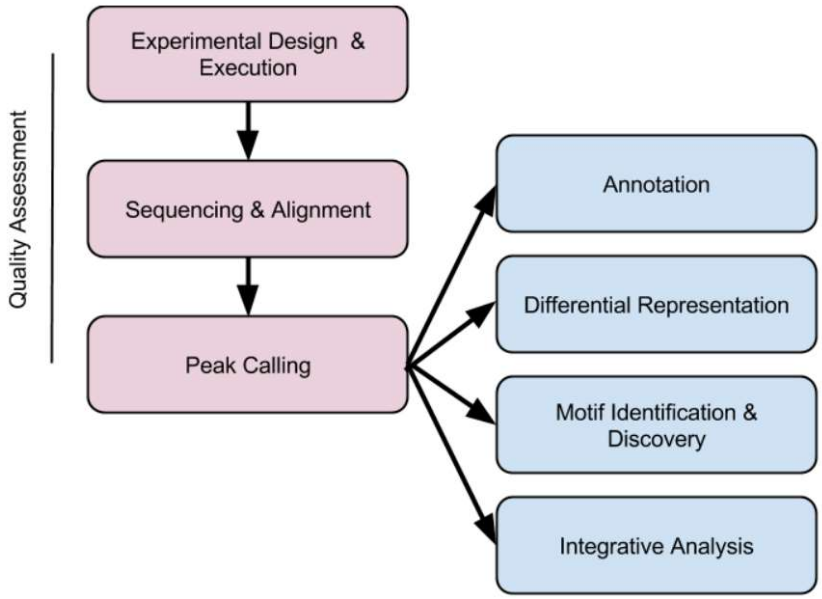
Lauren Vanderlinden

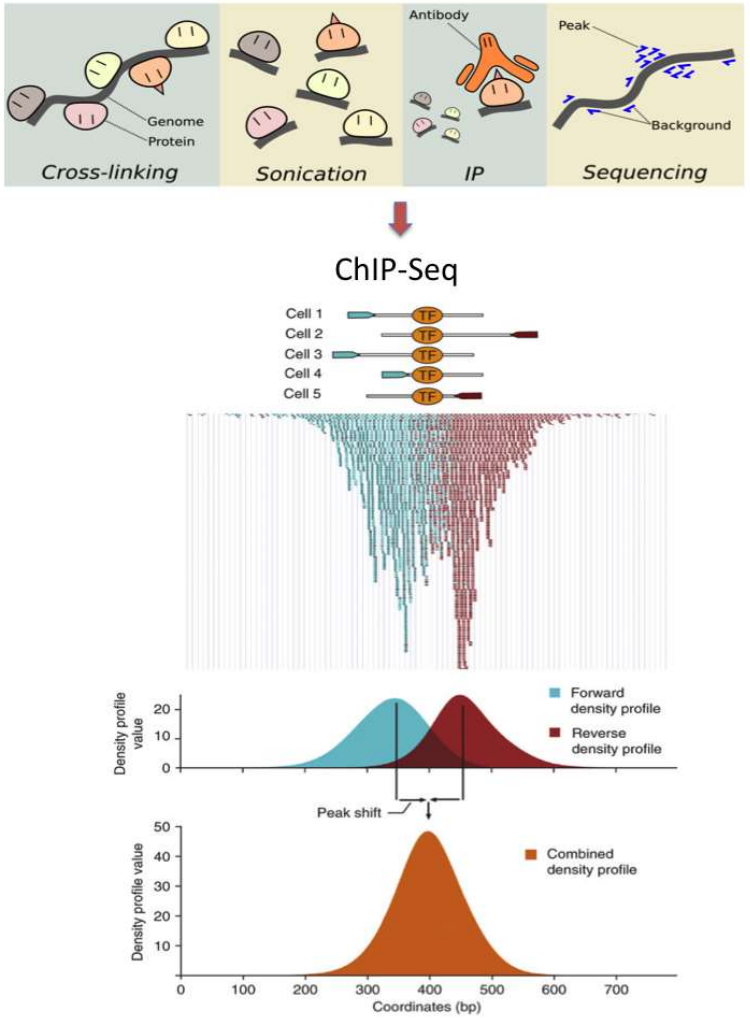
BIOS 6660

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

NCBI

Gene Expression Omnibus

EMBL-EBI

TCGA

GTEx Portal

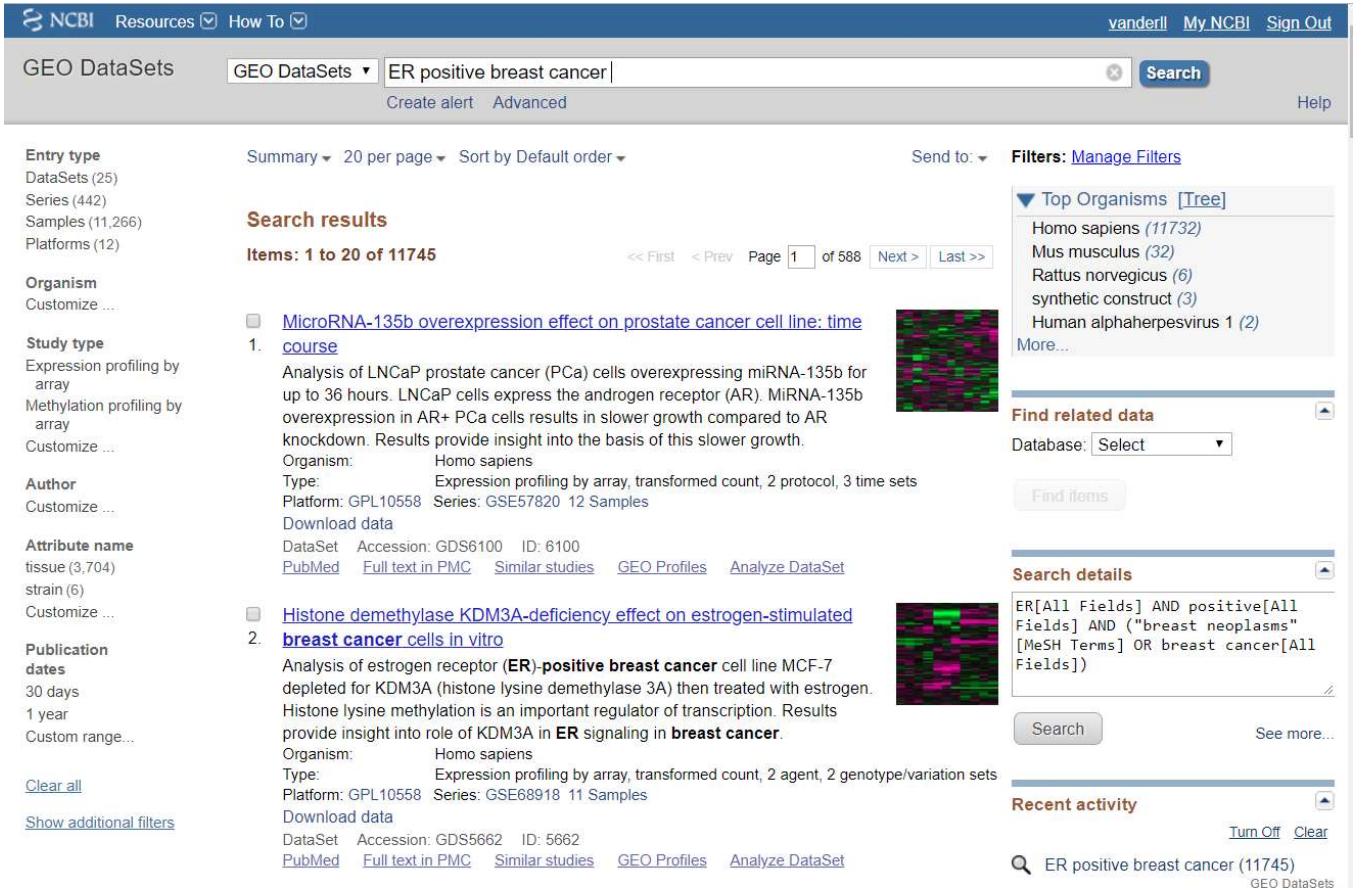
PhenoGen Informatics
The site for quantitative genetics of the transcriptome.

GeneNetwork
University of Tennessee: www.genenetwork.org

RGD

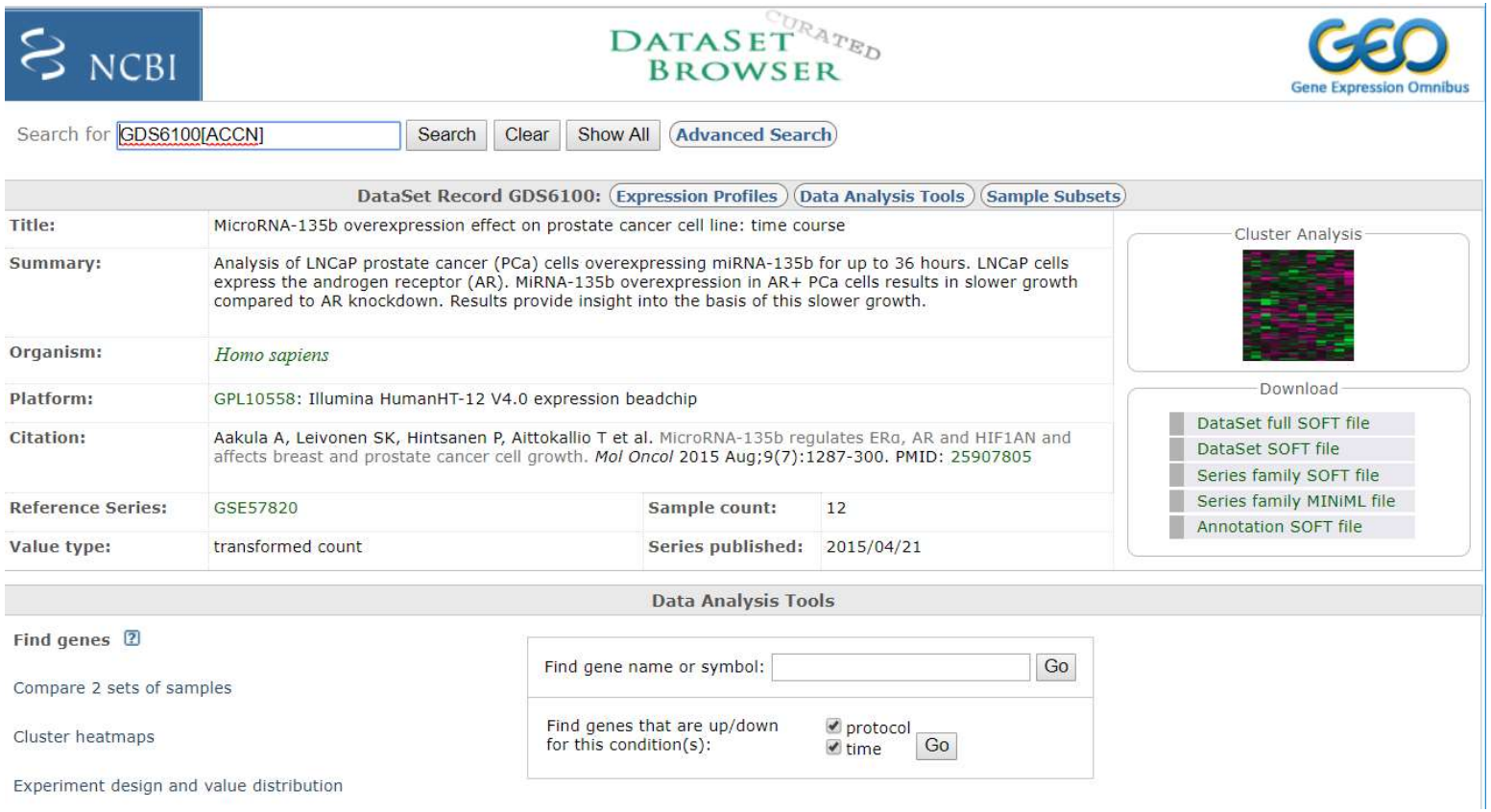
GEO Database

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

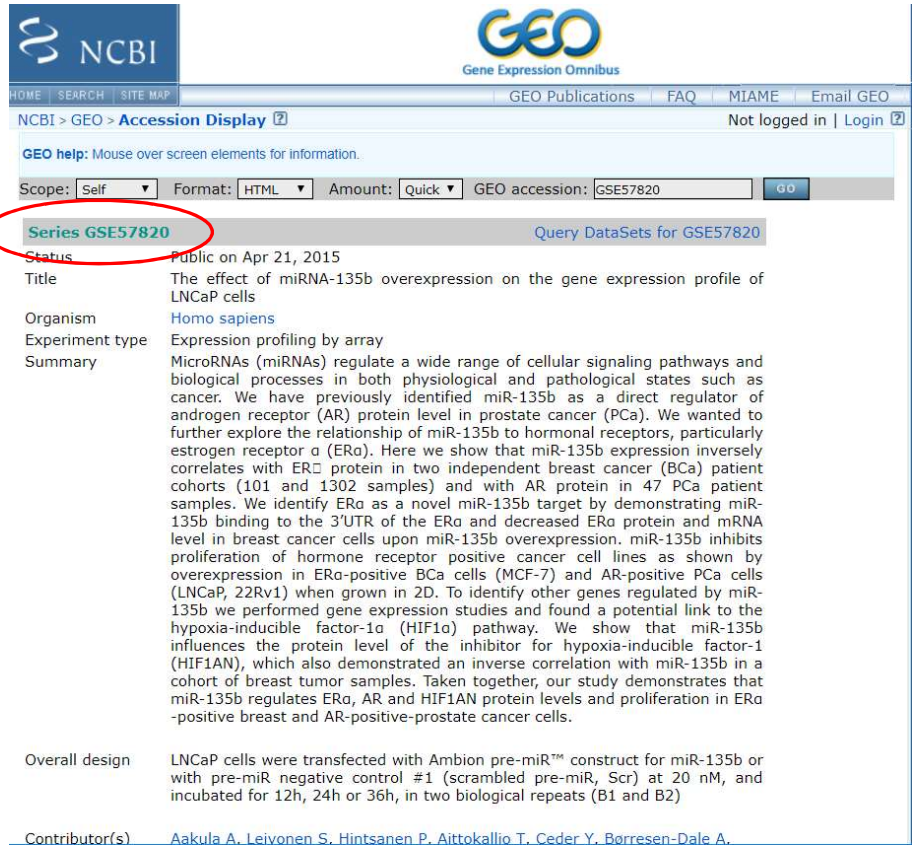


GEO Datasets – Selected Title

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



GEO Datasets – Select Reference Series



NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display** | Not logged in | Login

GEO help: Mouse over screen elements for information.

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

Series GSE57820 | Query DataSets for 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 (ERα). Here we show that miR-135b expression inversely correlates with ER protein in two independent breast cancer (BCa) patient cohorts (101 and 1302 samples) and with AR protein in 47 PCa patient samples. We identify ERα as a novel miR-135b target by demonstrating miR-135b binding to the 3'UTR of the ERα and decreased ERα 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 ERα-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-1α (HIF1α) 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 ERα, AR and HIF1AN protein levels and proliferation in ERα -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 S, Hintsanen P, Aittokallio T, Ceder Y, Børresen-Dale A

Citation(s)	Aakula A, Leivonen SK, Hintsanen P, Aittokallio T et al. MicroRNA-135b regulates ERα, 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	Tukholmankatu 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
More...	GSM1394595 LNCaP_miR-135b_12h_B2
	GSM1394596 LNCaP_miR-135b_24h_B1

Relations

BioProject PRJNA248178

Analyze with GEO2R

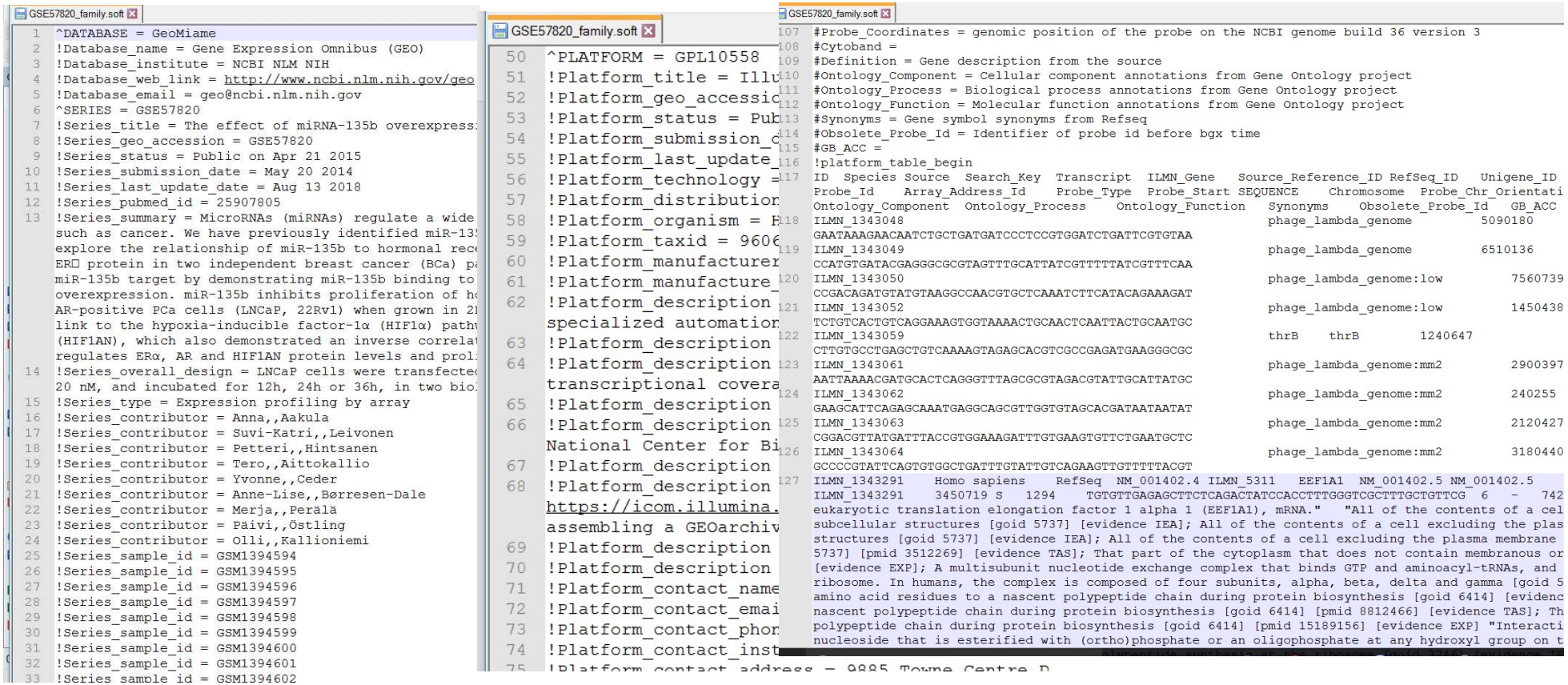
SOFT: Simple Omnibus Format in Text
Has both array data and meta data

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	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



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\vanderll\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
> pdata <- pData(exampleData) # data.frame of meta data
> dim(pdata)
[1] 12 45
```

47,323 probe sets & 12 samples

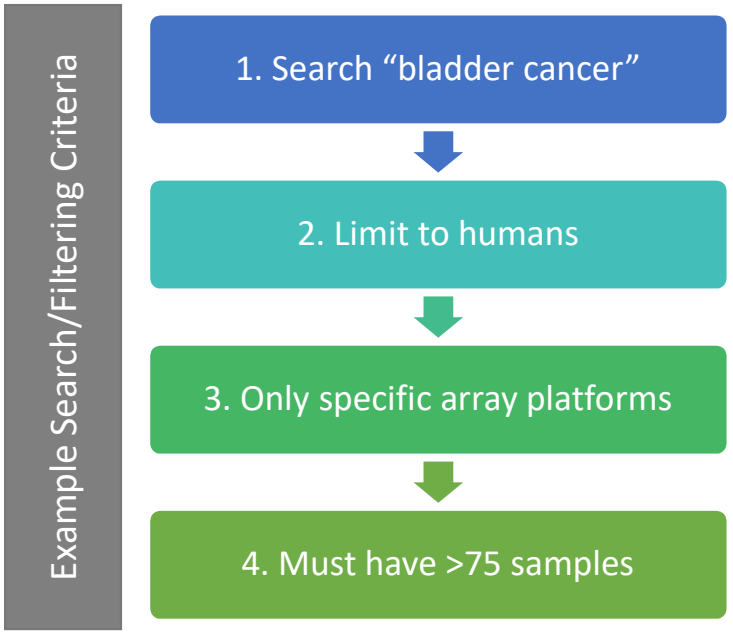
45 phenotypes for 12 samples

- 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)
```




```
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

humanKM.flowChart.Rmd* | bladder.4

Filter

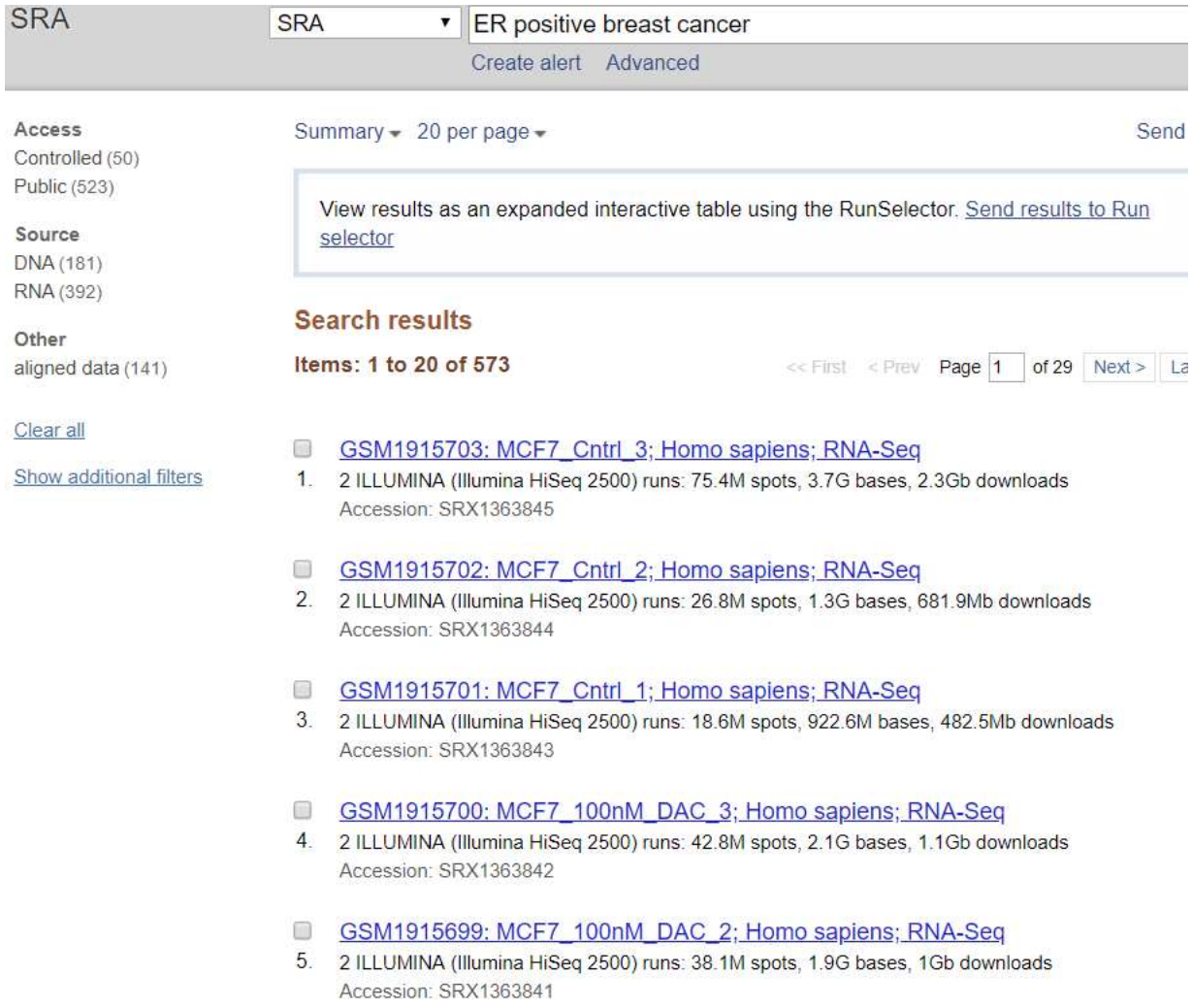
	Series	Organism	Title	Type	Platform	Sample.Number	Description	Term
162	GSE31189	Homo sapiens	Molecular Biomarker Signatur...	Expression profiling by array	GPL570	92	In this study we applied differe...	bladder cancer
181	GSE31684	Homo sapiens	Combination of a novel gene ...	Expression profiling by array	GPL570	93	Urothelial carcinoma of the bla...	bladder cancer
198	GSE13507	Homo sapiens	Predective 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"
  "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 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)

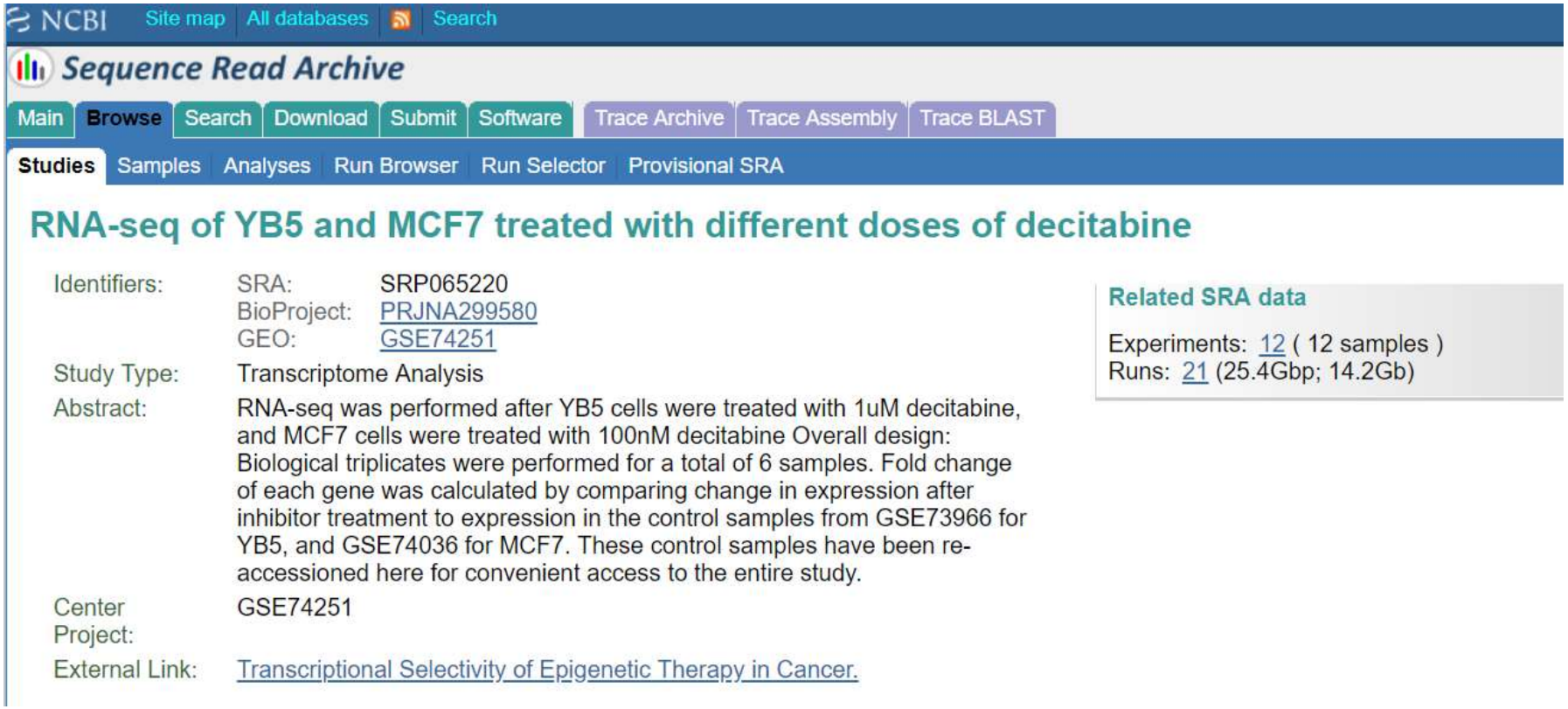
Experiment attributes:
GEO Accession: GSM1915703

Links:

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

Run	# of Spots	# of Bases	Size	Published
SRR2753169	37,479,476	1.9G	1.1Gb	2017-01-31
SRR2753170	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

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.

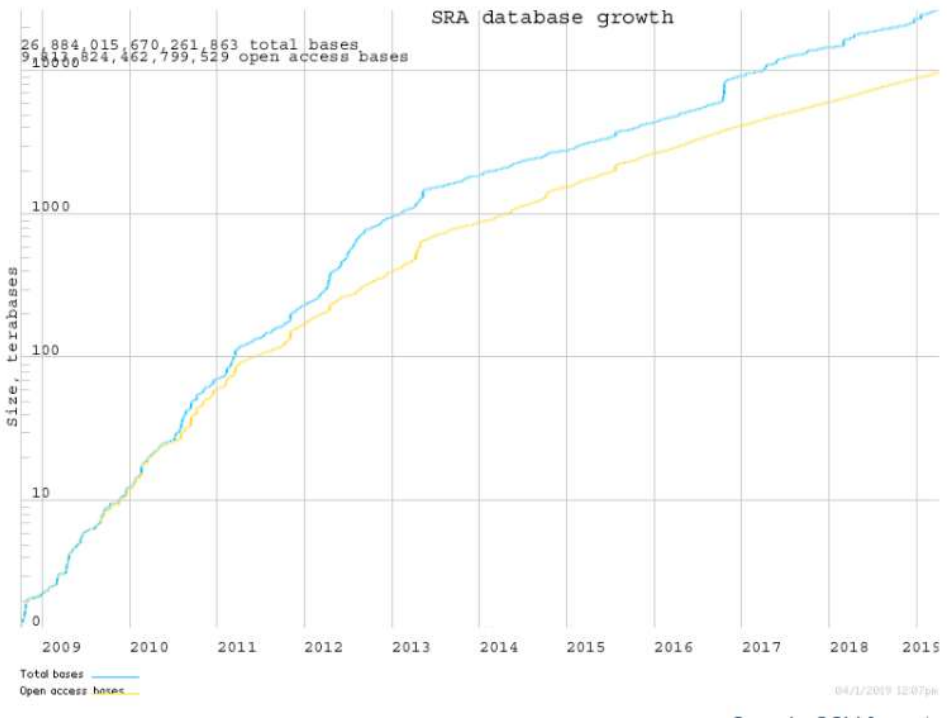
SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

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

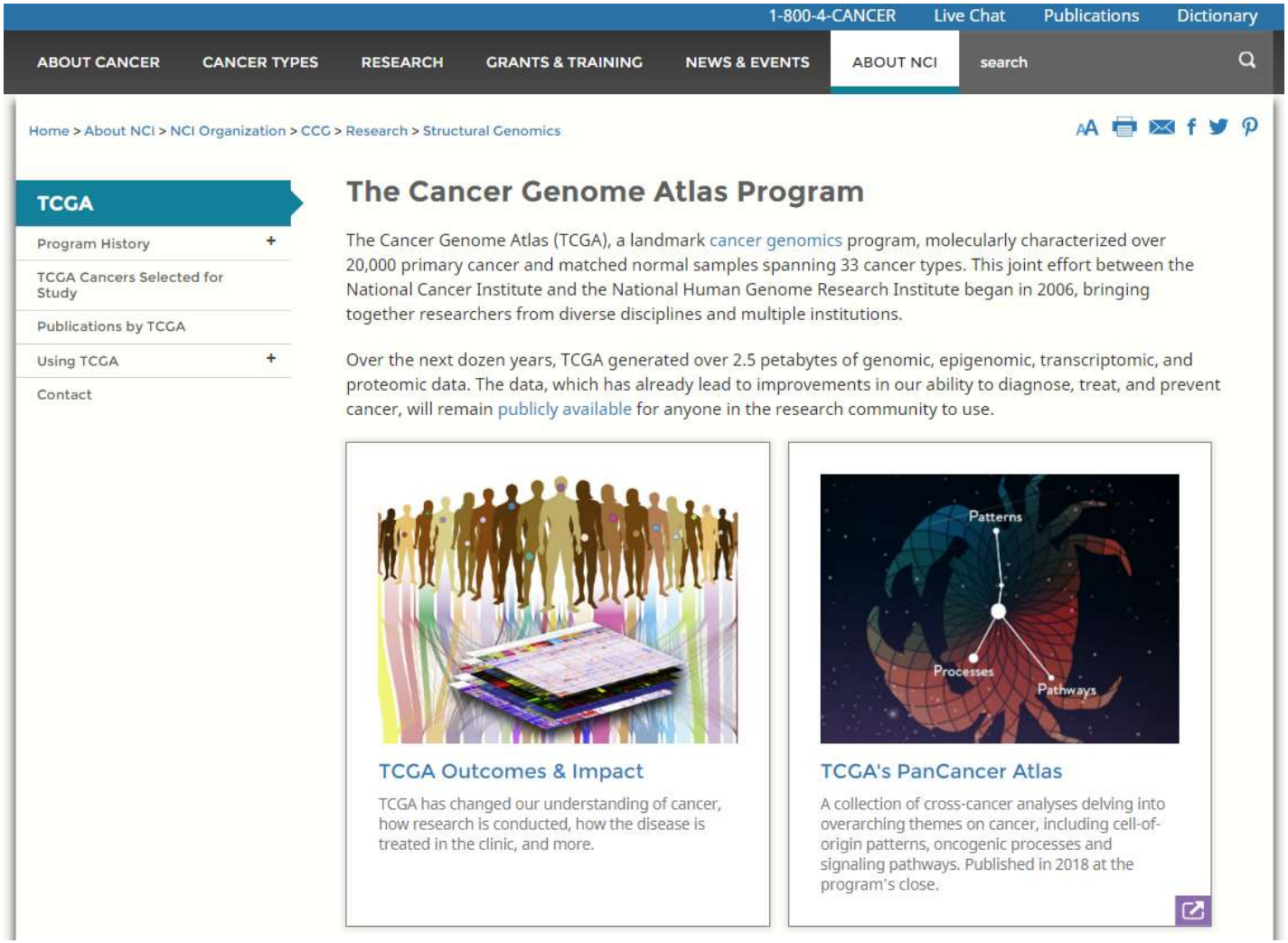
Search: Go

What can be entered in this field?

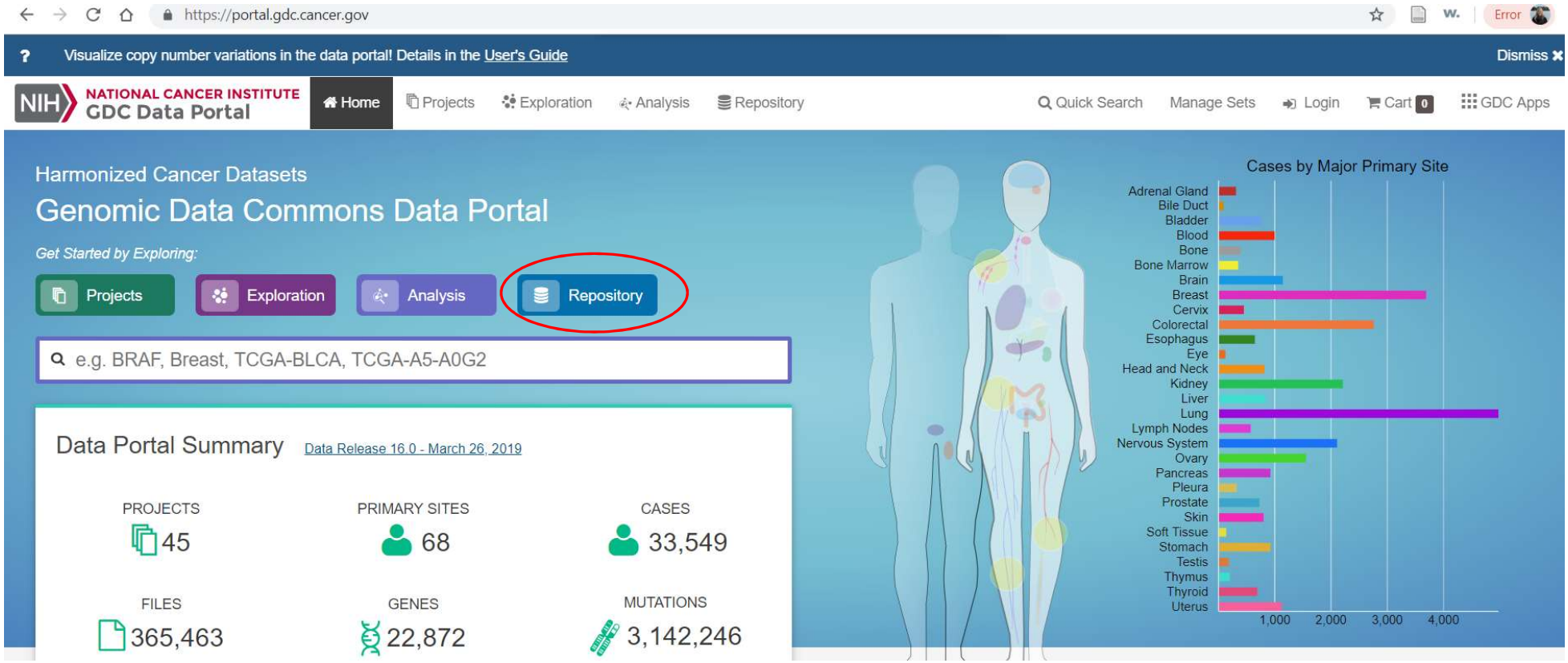
List of SRA Samples. 4916785 found.

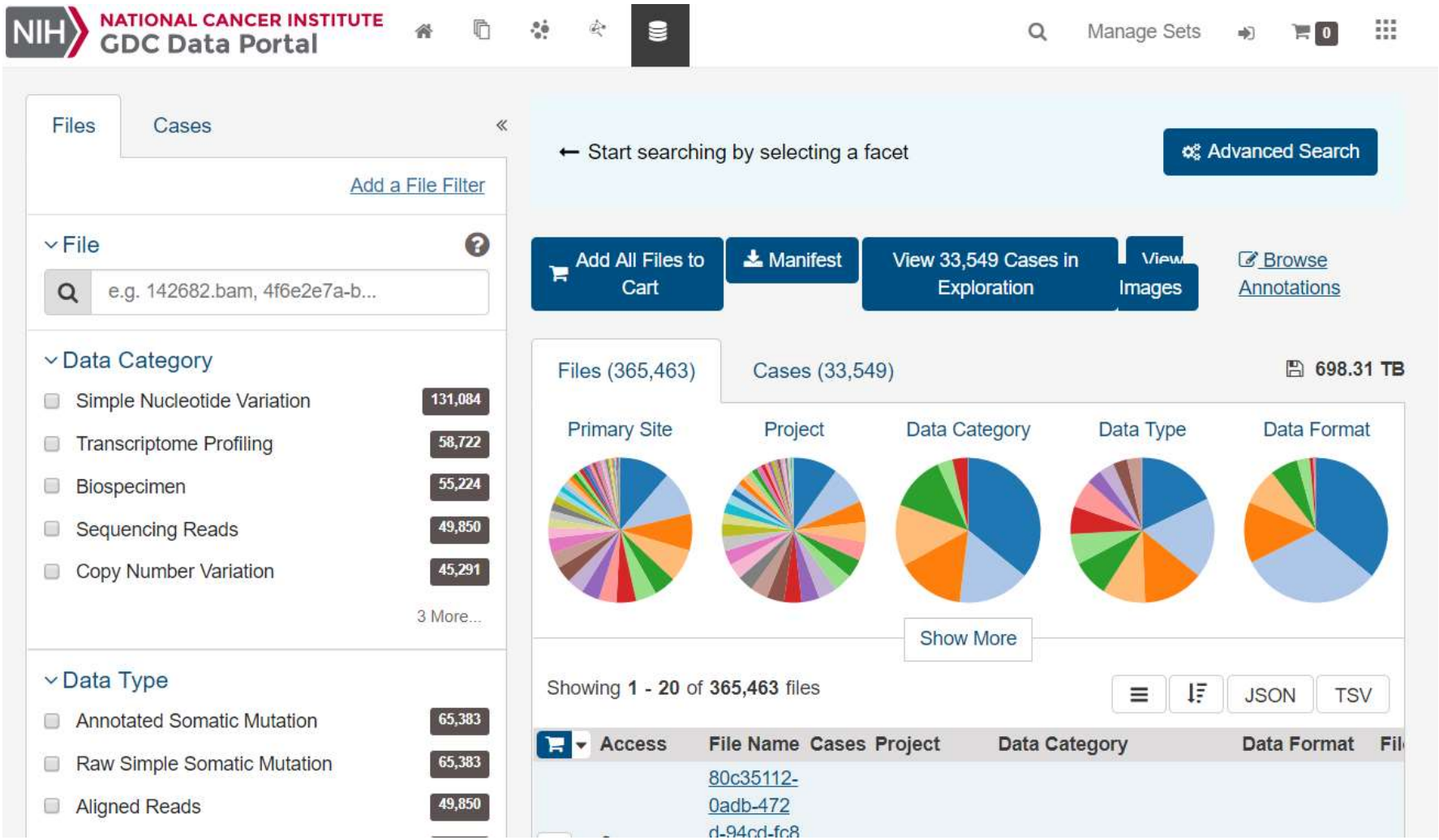
#	Accession	Organism	Title	Attributes
1.	SRS4512763	Homo sapiens	T-9_R1.fastq	isolate: **T-9-R1** age: **missing** biomaterial_provider: **david.tulasne@ibl.cnrs.fr** sex: **-** tissue: **LUNG** cell_line: **-** cell_subtype: **-** cell_type: **-** culture_collection: **-** dev_stage: **-** disease: **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
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.
Patterns
Processes
Pathways
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




NIH NATIONAL CANCER INSTITUTE
GDC Data Portal
Manage Sets
0
Files
Cases
Add a File Filter
File
e.g. 142682.bam, 4f6e2e7a-b...
Data Category
Simple Nucleotide Variation 131,084
Transcriptome Profiling 58,722
Biospecimen 55,224
Sequencing Reads 49,850
Copy Number Variation 45,291
3 More...
Data Type
Annotated Somatic Mutation 65,383
Raw Simple Somatic Mutation 65,383
Aligned Reads 49,850
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
JSON
TSV
Access
File Name
Cases
Project
Data Category
Data Format
80c35112-0adb-472

R/TCGAbiolinks

TCGAbiolinks

platforms all | rank 97 / 1649 | posts 5 / 1 / 1 / 0 | in Bioc 3.5 years

build warnings | updated < 1 month

DOI: 10.18129/B9.bioc.TCGAbiolinks

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

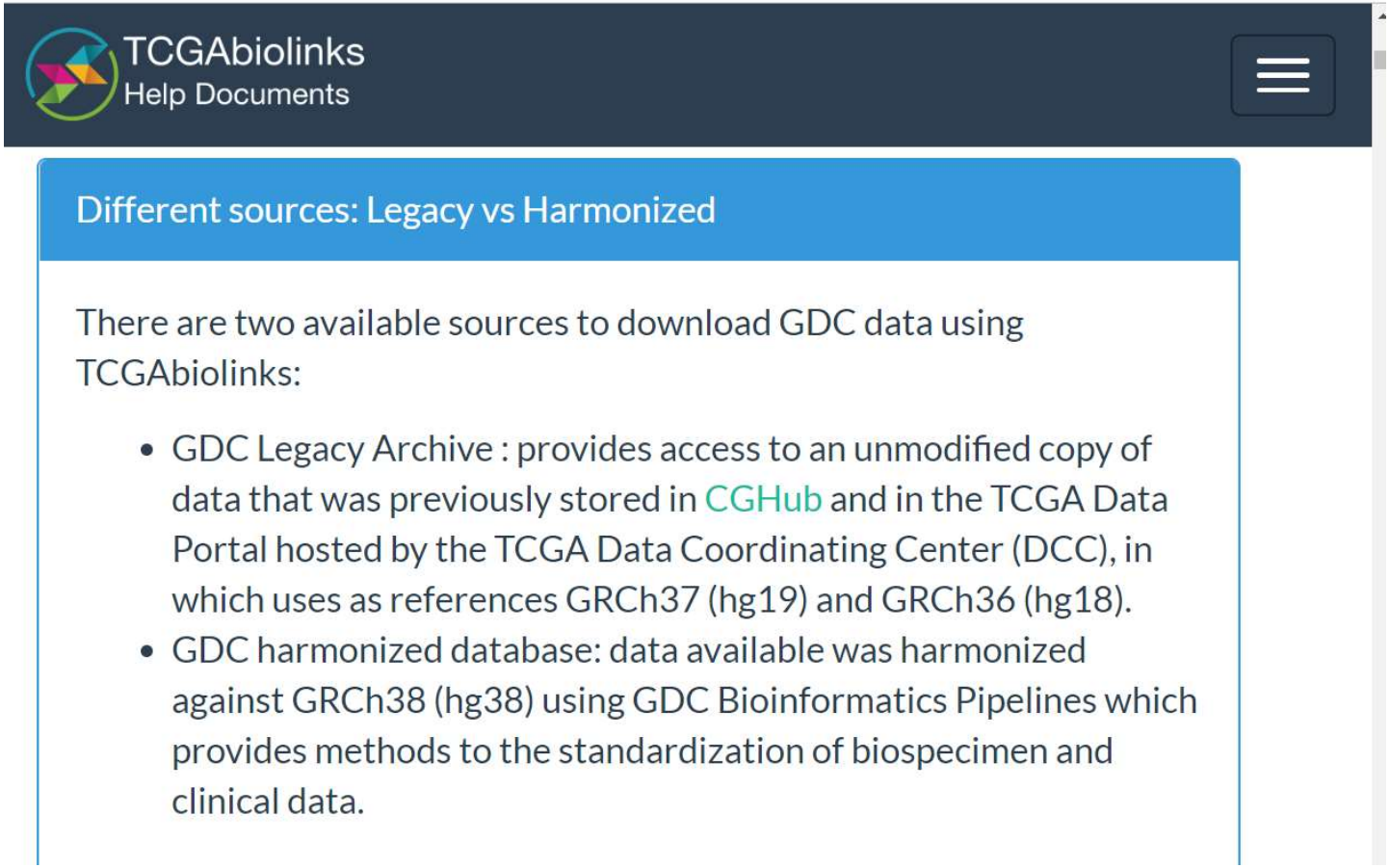
```
browseVignettes("TCGAbiolinks")
```

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



GDCquery()

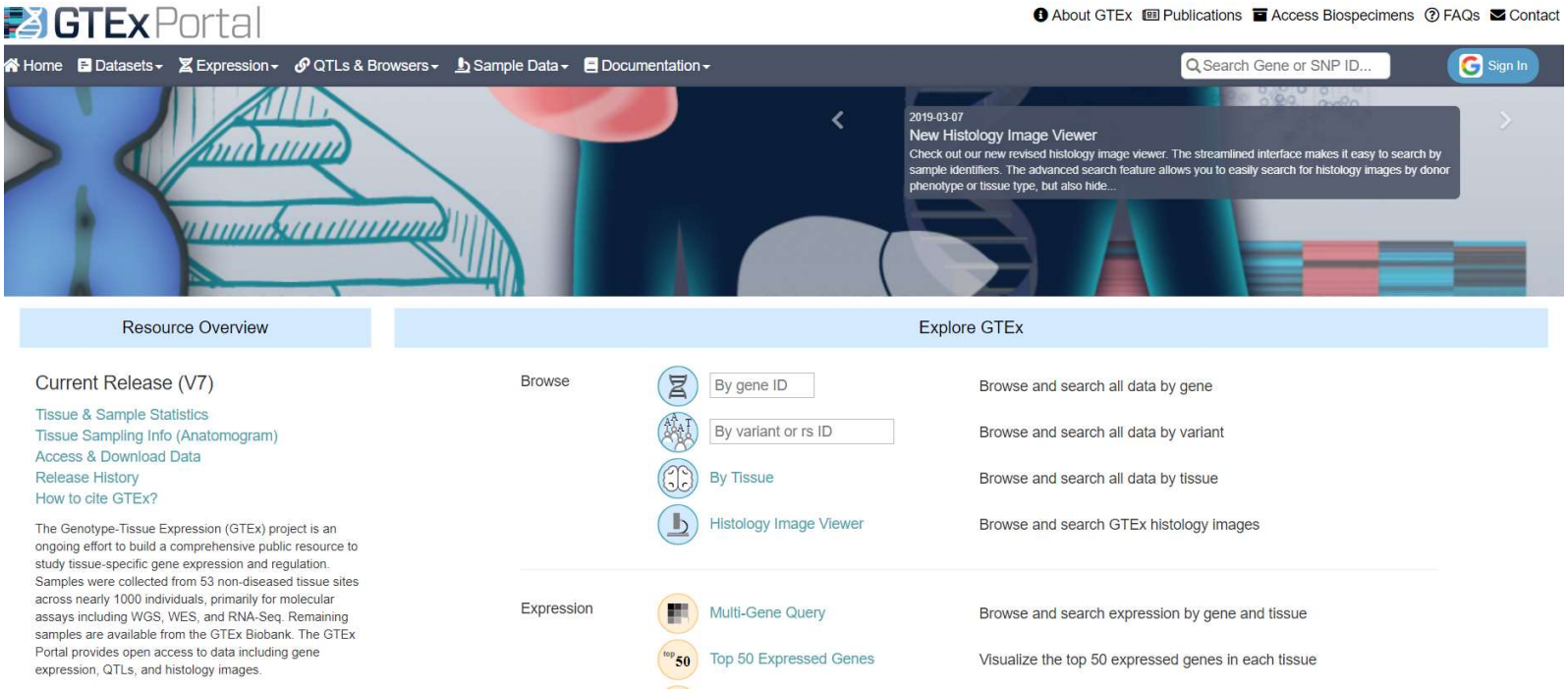
```
query <- GDCquery(project = c("TCGA-GBM", "TCGA-LGG"),
                  data.category = "DNA Methylation",
                  legacy = FALSE,
                  platform = c("Illumina Human Methylation 450"),
                  sample.type = "Recurrent Solid Tumor"
)
```

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

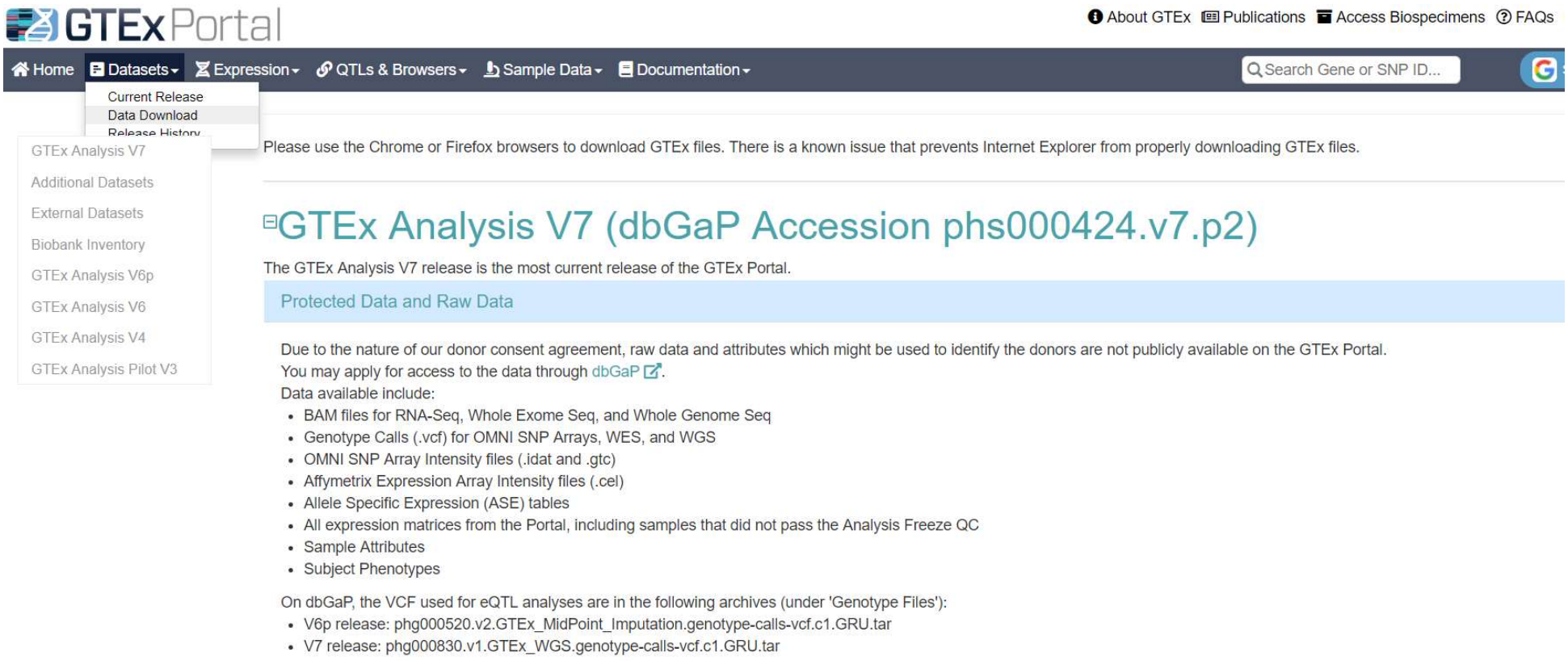
GTEx

https://gtexportal.org/home/

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



GTEx


GTEx Portal
About GTEx
Publications
Access Biospecimens
FAQs
Home
Datasets
Expression
QTLs & Browsers
Sample Data
Documentation
Search Gene or SNP ID...
Current Release
Data Download
Release History
GTEx Analysis V7
Additional Datasets
External Datasets
Biobank Inventory
GTEx Analysis V6p
GTEx Analysis V6
GTEx Analysis V4
GTEx Analysis Pilot V3
Please use the Chrome or Firefox browsers to download GTEx files. There is a known issue that prevents Internet Explorer from properly downloading GTEx files.
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

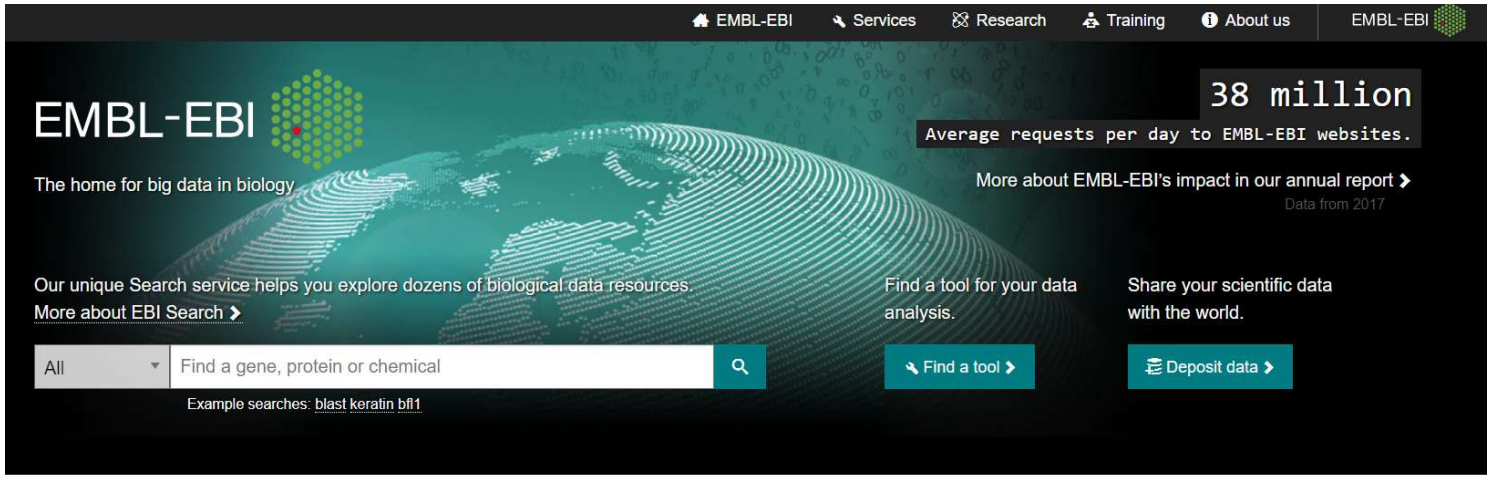
GTEx RNA-Seq Downloads

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.

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

Training

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

Industry

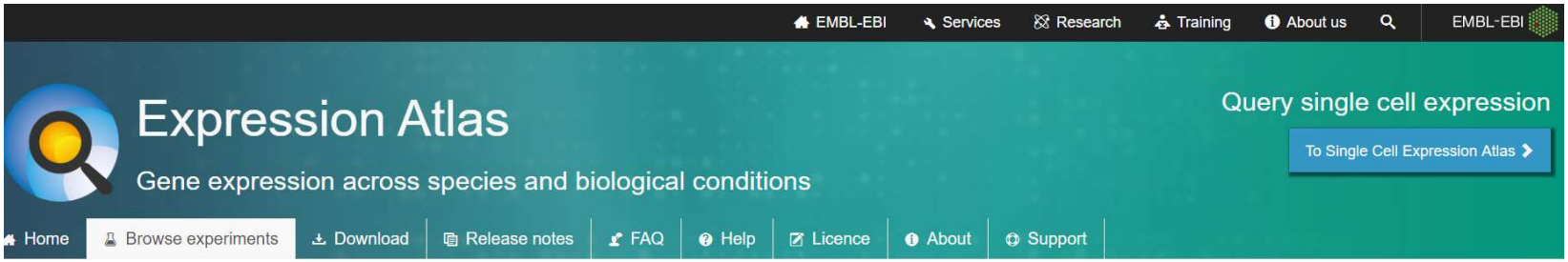
Explore our knowledge-exchange Industry Programme and take part in translational partnerships and

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



Experiments in Expression Atlas

Kingdom:

All

All
Plants
Animals
Fungi

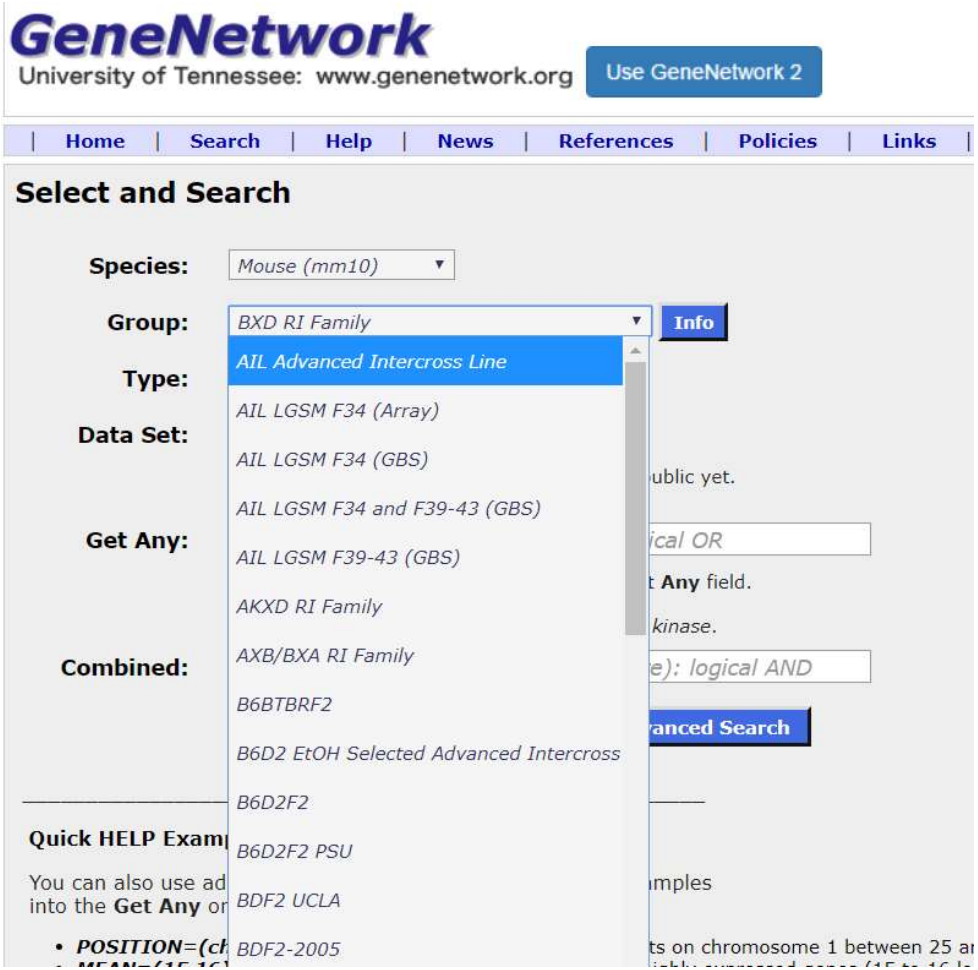
Search all columns:

	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

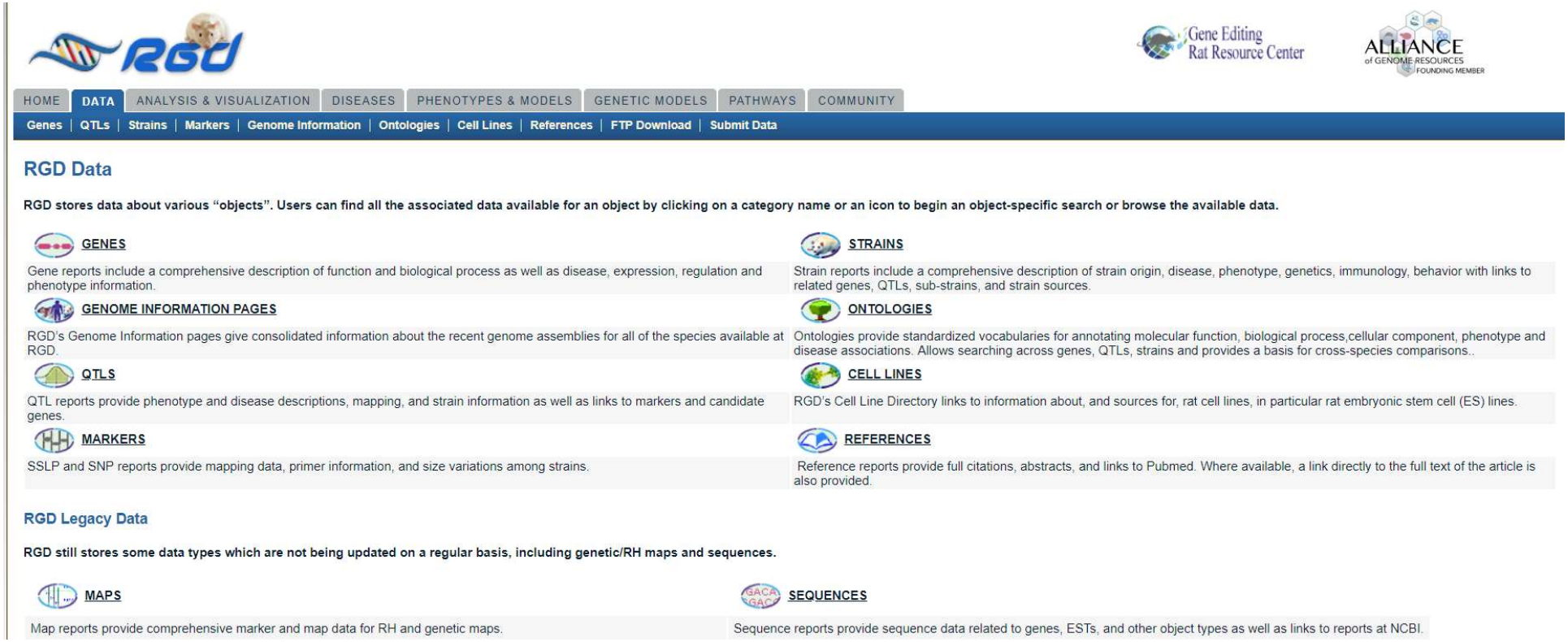
- Extensive database on rodents
- Group is panel or cohort
- Type: Phenotype or Tissue Specific (Adispose)
- Dataset: Specific experiment
- Get Any and Combined is a way to filter the previous selected fields

http://www.genenetwork.org



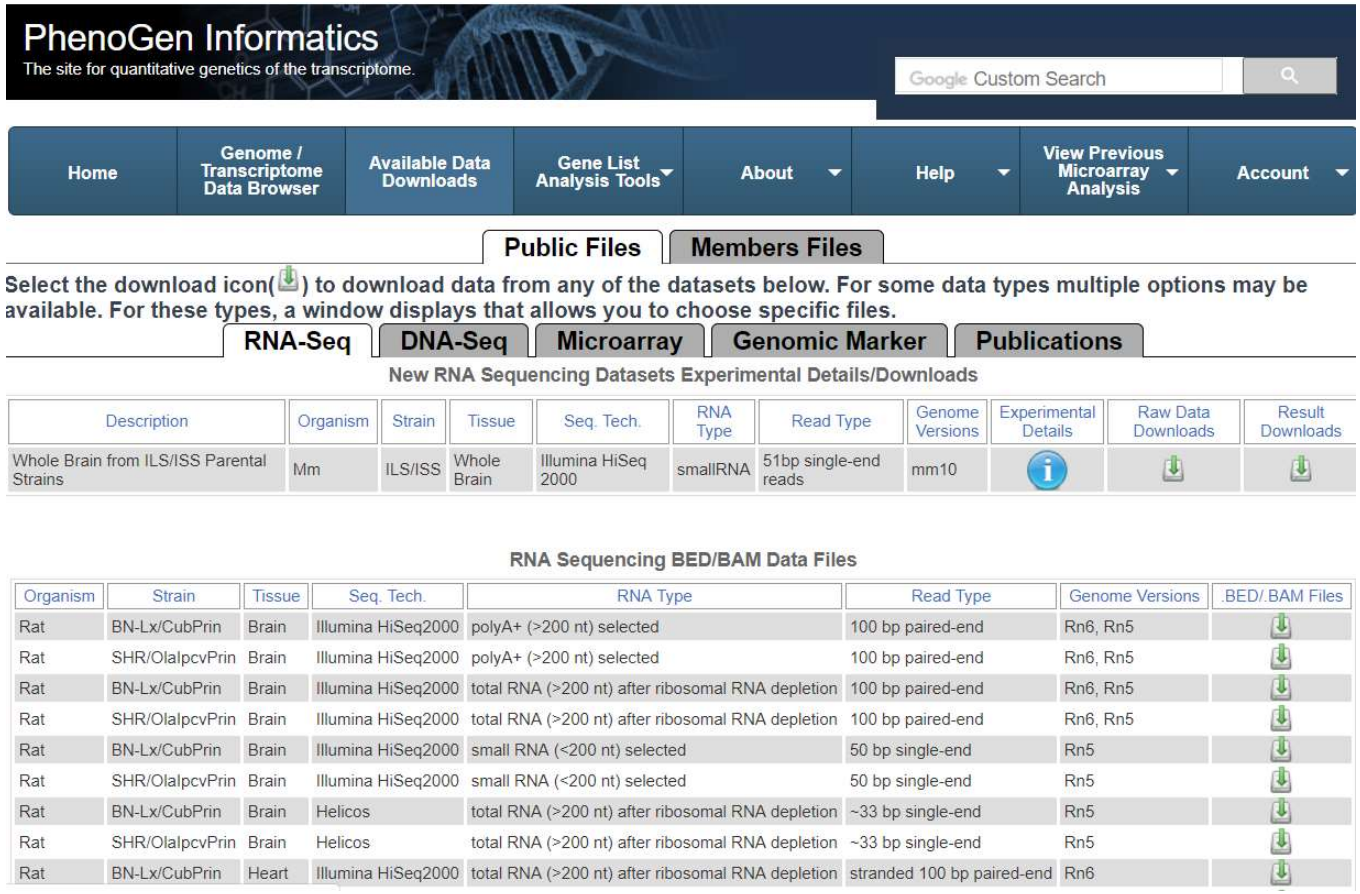
RGD

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



PhenoGen

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



PhenoGen Informatics

The site for quantitative genetics of the transcriptome.

Google Custom Search

Home | Genome / Transcriptome Data Browser | Available Data Downloads | Gene List Analysis Tools | About | Help | View Previous Microarray Analysis | Account

Public Files | Members Files

Select the download icon( ) 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.

RNA-Seq | DNA-Seq | Microarray | Genomic Marker | Publications

New RNA Sequencing Datasets Experimental Details/Downloads

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			

RNA Sequencing BED/BAM Data Files

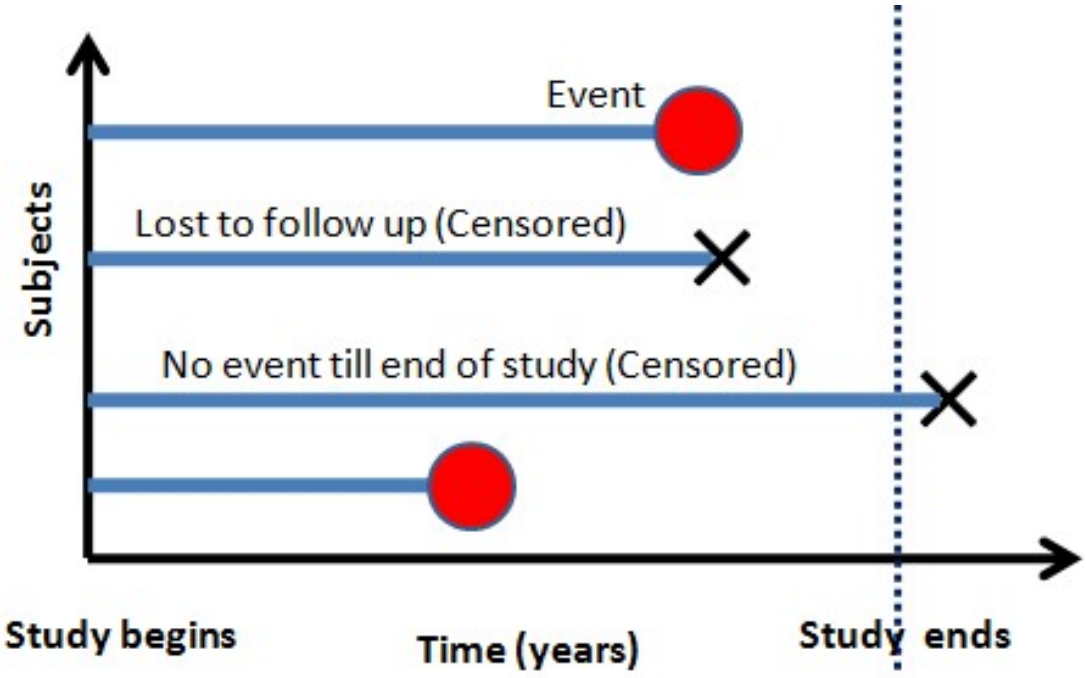
Organism	Strain	Tissue	Seq. Tech.	RNA Type	Read Type	Genome Versions	.BED/.BAM Files
Rat	BN-Lx/CubPrin	Brain	Illumina HiSeq2000	polyA+ (>200 nt) selected	100 bp paired-end	Rn6, Rn5	
Rat	SHR/OlaIpcvPrin	Brain	Illumina HiSeq2000	polyA+ (>200 nt) selected	100 bp paired-end	Rn6, Rn5	
Rat	BN-Lx/CubPrin	Brain	Illumina HiSeq2000	total RNA (>200 nt) after ribosomal RNA depletion	100 bp paired-end	Rn6, Rn5	
Rat	SHR/OlaIpcvPrin	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/OlaIpcvPrin	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/OlaIpcvPrin	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

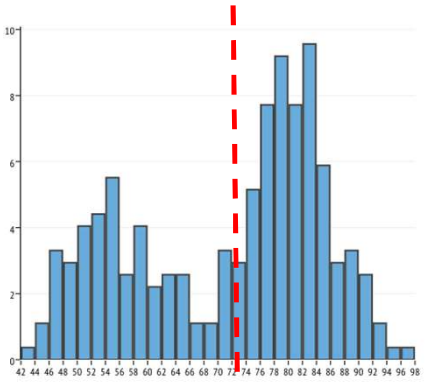
1. For each gene, perform a cox proportional hazard model and obtain resulting coefficient.
2. Calculate a score for each subject by summing gene*βcoefficient.
3. Classify each subject as having a high/low signature by determining if individuals score is above/below median score value.
4. Perform cox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

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

Baseline hazard function

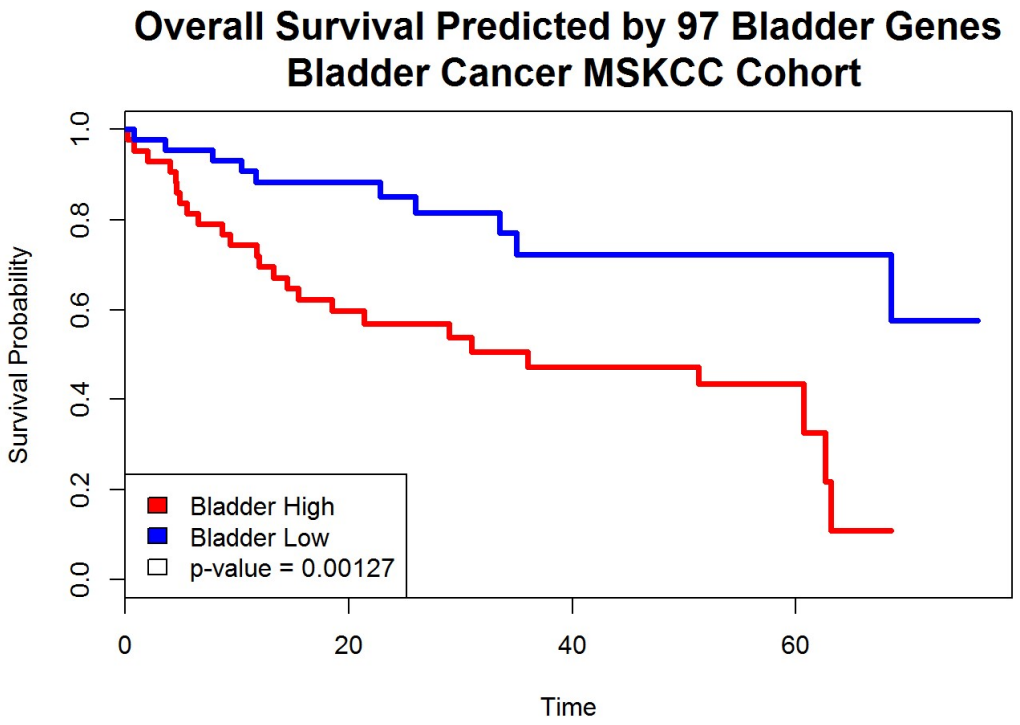
Linear function of set of predictors

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



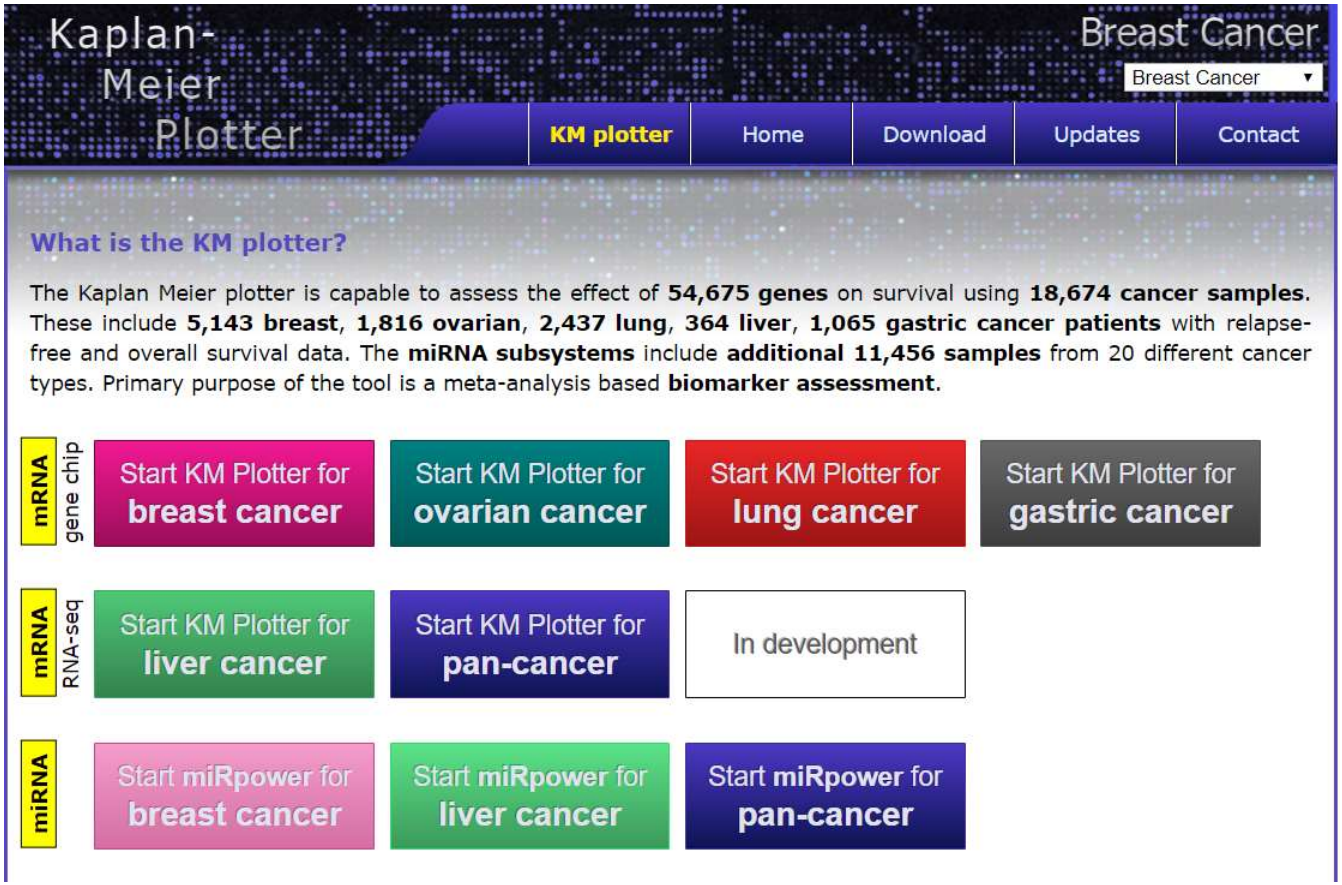
Testing Gene Signature

- 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

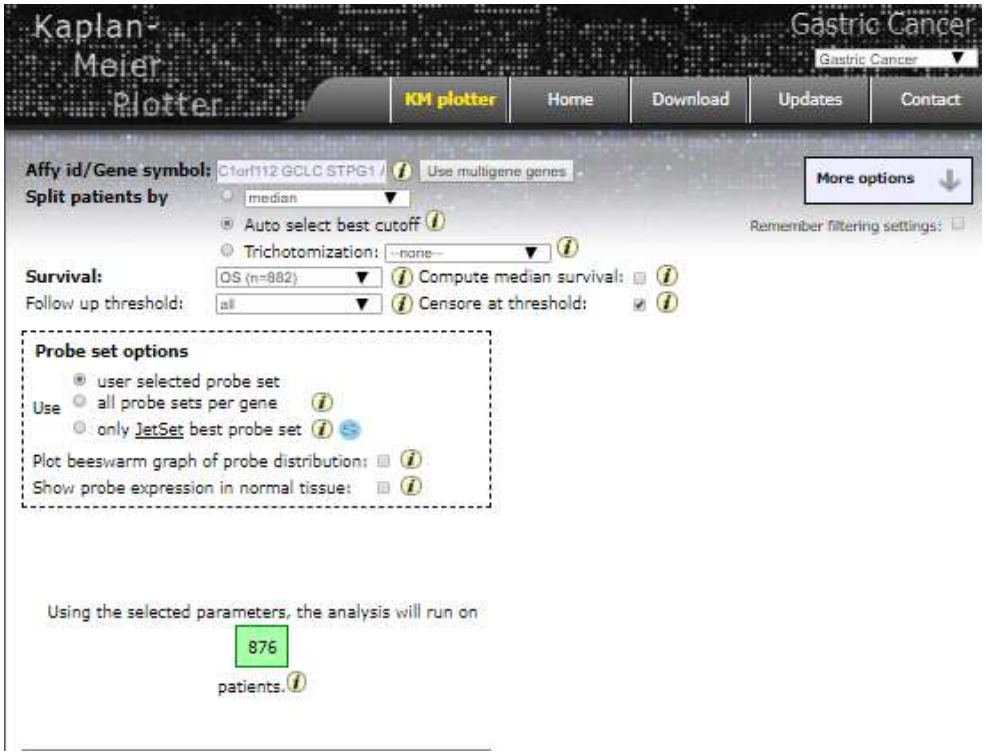


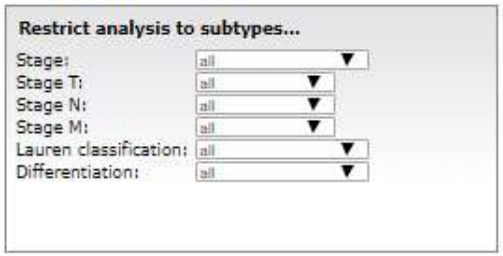
KM Plotter

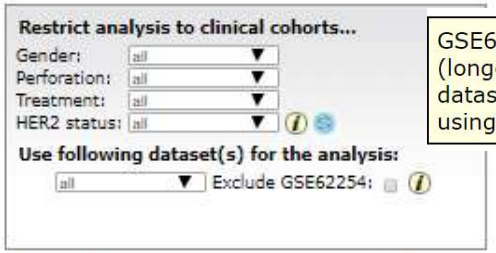
http://kmplot.com/analysis/



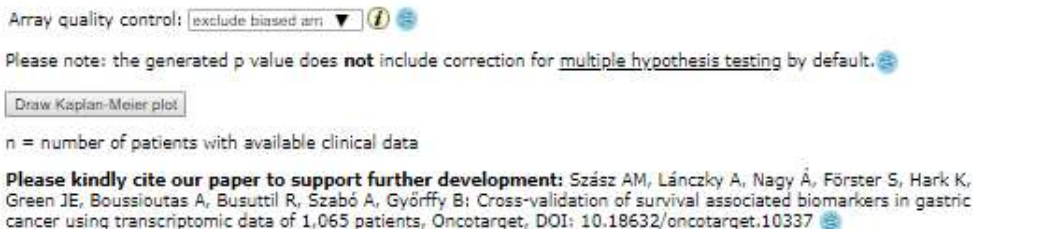
KM Plotter – Selecting Data for Analysis







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



KM Plotter - Results

Plotter | KM plotter | Home | Downloa

The desired is valid: 205942_s_at (ACSM3),

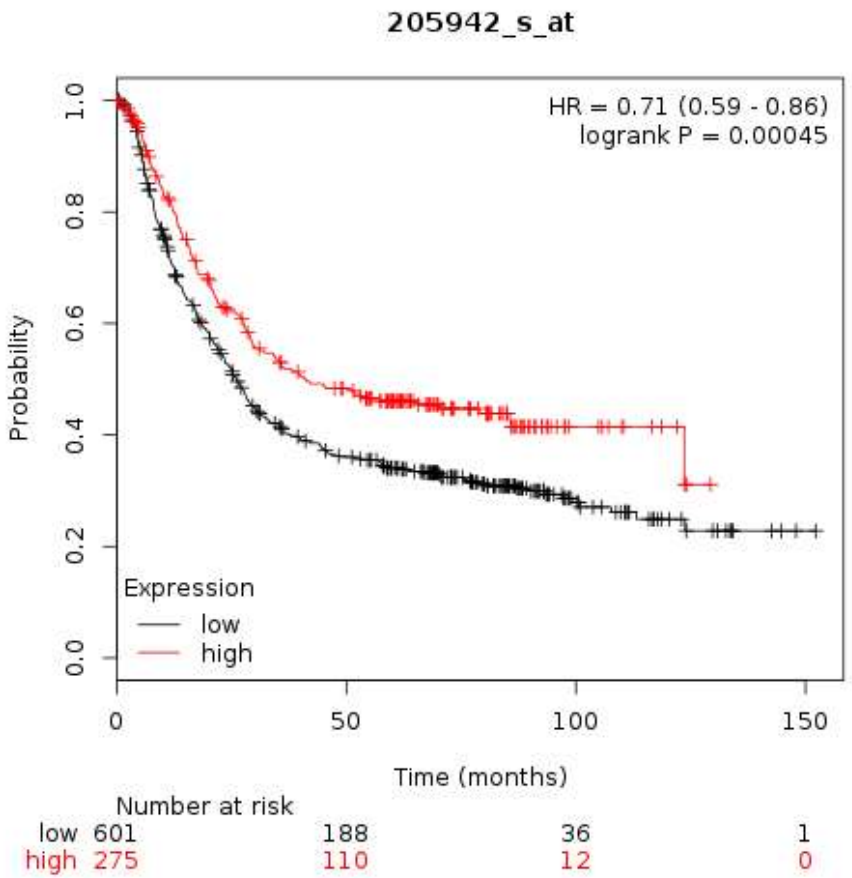
Affy ID:	205942_s_at	ACSM3, SA, SAH
Survival:	OS	
Auto select best cutoff:	checked	
Follow up threshold:	all	
Censore at threshold:	checked	
Compute median over entire database:	false	
Cutoff value used in analysis:	230	
Expression range of the probe:	2 - 2185	
Probe set option:	user selected probe set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

Restrictions

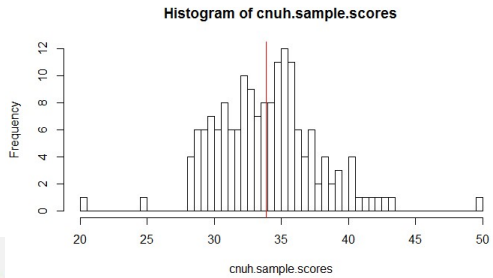
Stage:	all
Stage T:	all
Stage N:	all
Stage M:	all
Lauren classification:	all
Differentiation:	all
Gender:	all
Perforation:	all
Treatment:	all
HER2 status:	all
Dataset:	all
Exclude GSE62254:	not checked

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")
```

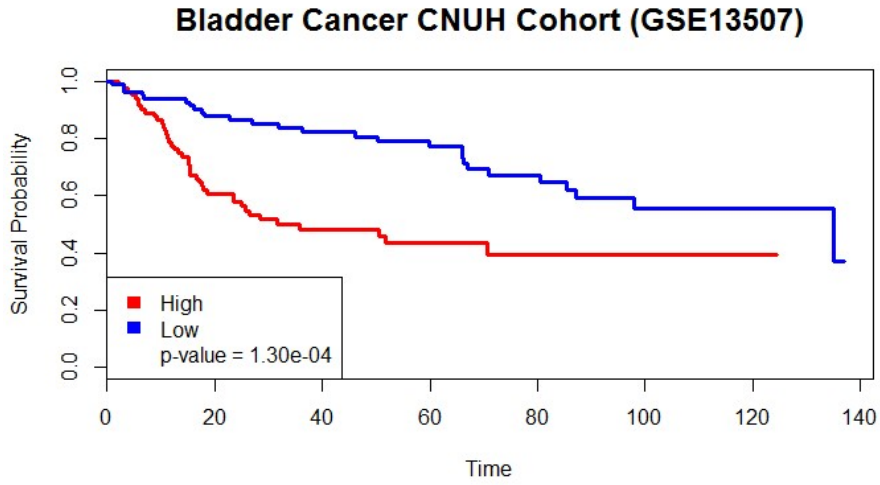
Gene Signature R Code

```
##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