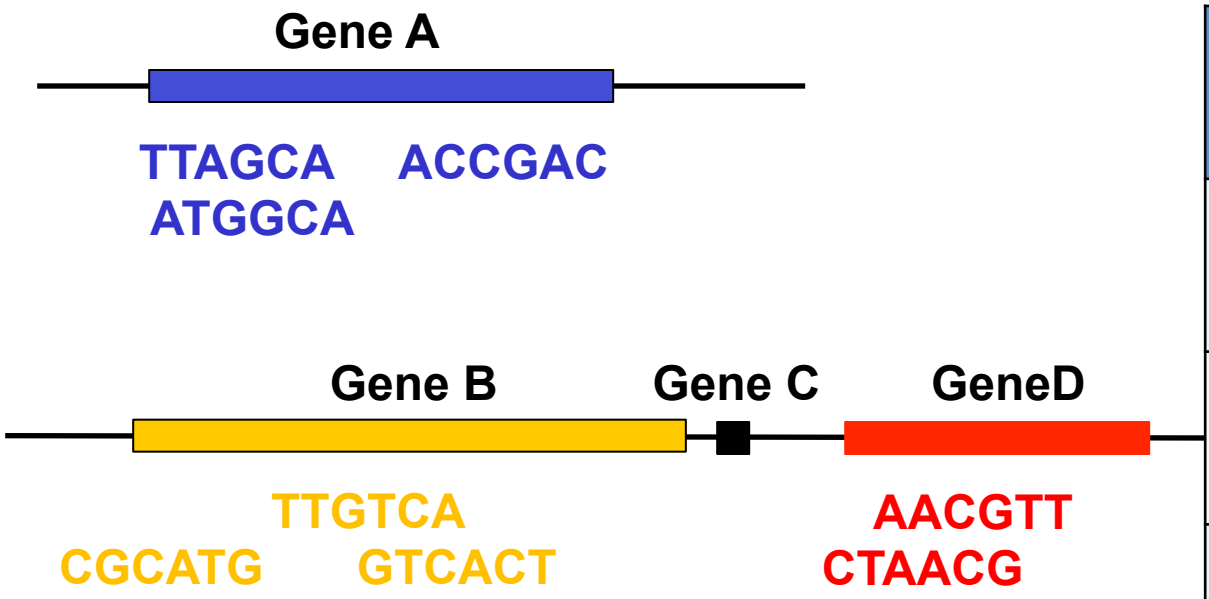
Next-generation Sequencing

Lecture 11

Steps involved on RNA-seq analysis

- Step 1: Preprocess
- Step 2: Map reads to the reference genome
- Step 3: Count how many reads fall within each feature of interest (gene, transcript, exon etc).
- Step 4: Normalization
- Step 5: Identify differentially expressed genes.

Align reads to Genome and count

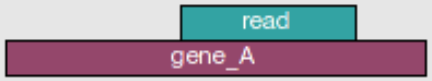
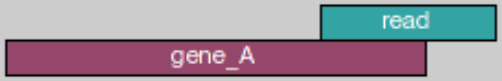
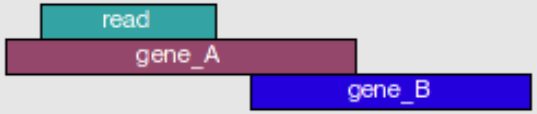
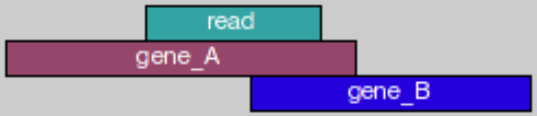
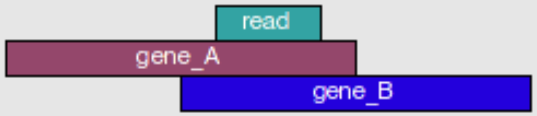


Gene ID	Sample1
A	3
B	3
C	0
D	2

For a given gene, the number of reads aligned to the gene measures its expression level.

Determine Abundance

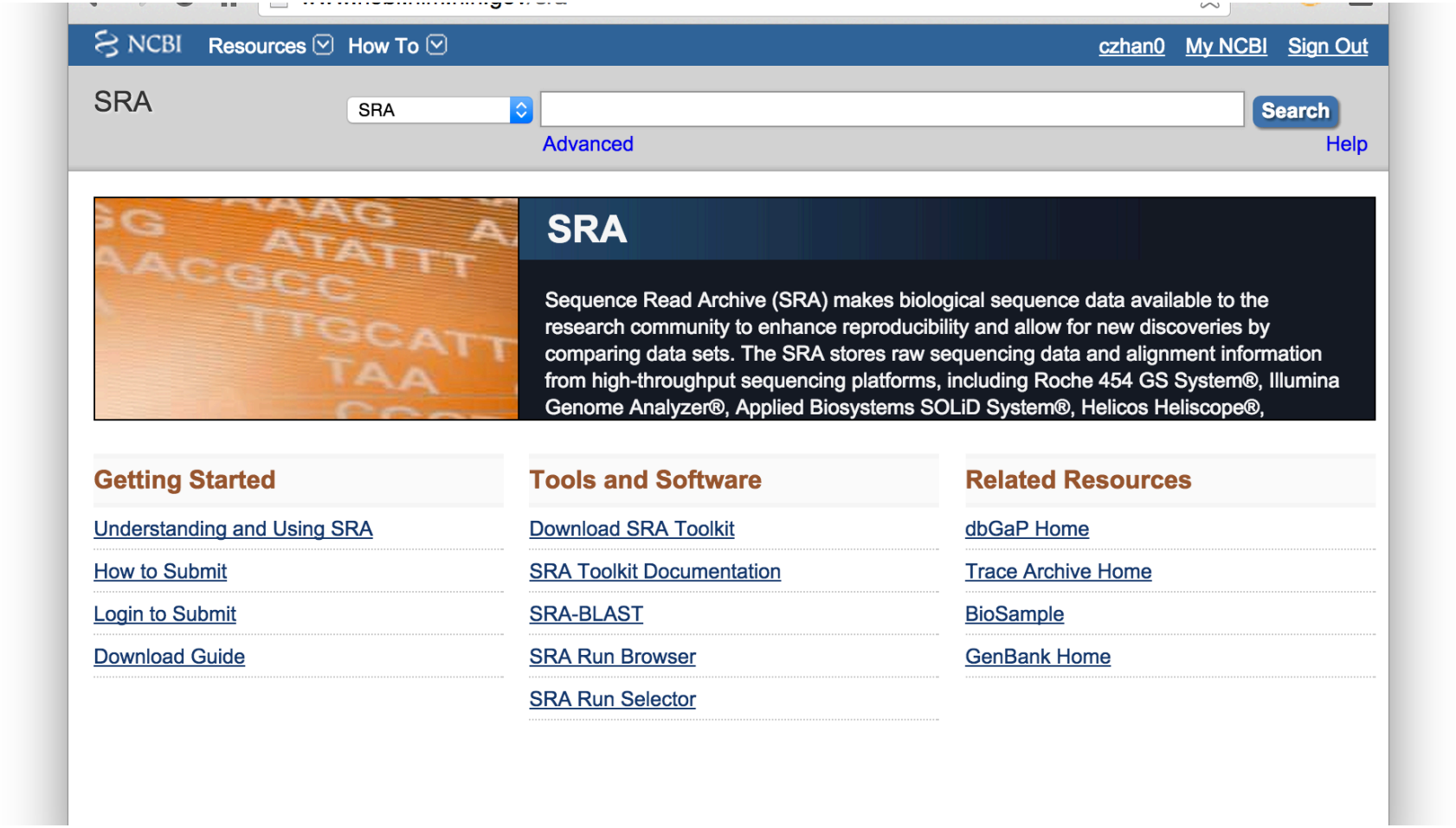
- Count reads in gene, coding area, or exons.
- Need gene annotation files in GFF (General Feature Format) format, which gives complete gene, RNA transcript or protein structures
- Tools:
  - Cufflinks (http://cufflinks.cbcb.umd.edu/)
  - Sam2counts (https://github.com/vsbuffalo/sam2counts)
  - **HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)**

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

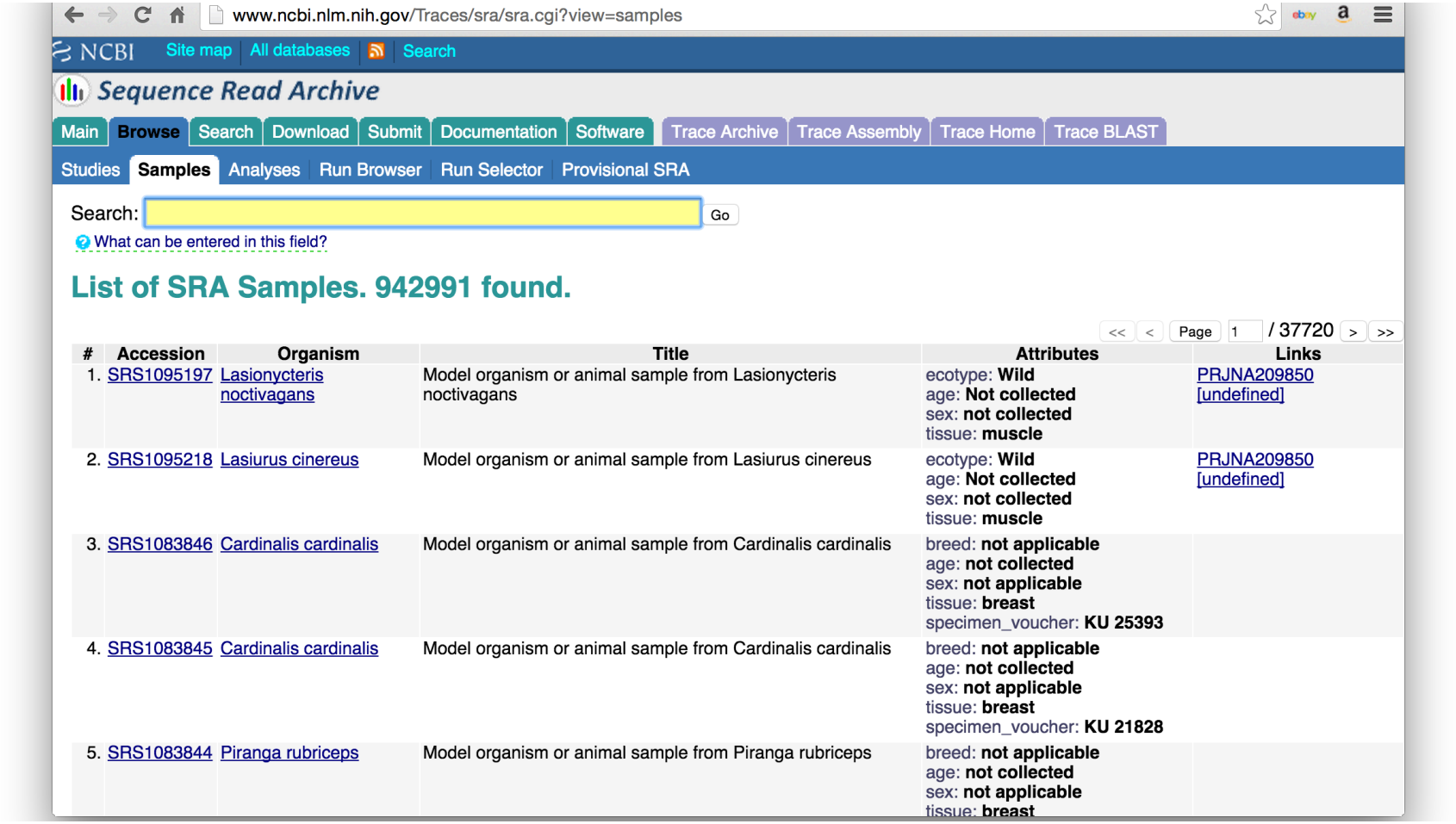
RNA-seq databases

- NCBI SRA
- NCBI GEO
- NIH TCGA

SRA


NCBI
Resources
How To
czhan0
My NCBI
Sign Out
SRA
SRA
Search
Advanced
Help
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®,
Getting Started
Understanding and Using SRA
How to Submit
Login to Submit
Download Guide
Tools and Software
Download SRA Toolkit
SRA Toolkit Documentation
SRA-BLAST
SRA Run Browser
SRA Run Selector
Related Resources
dbGaP Home
Trace Archive Home
BioSample
GenBank Home

SRA



www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=samples

NCBI Site map | All databases | Search

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly | Trace Home | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

Search: Go

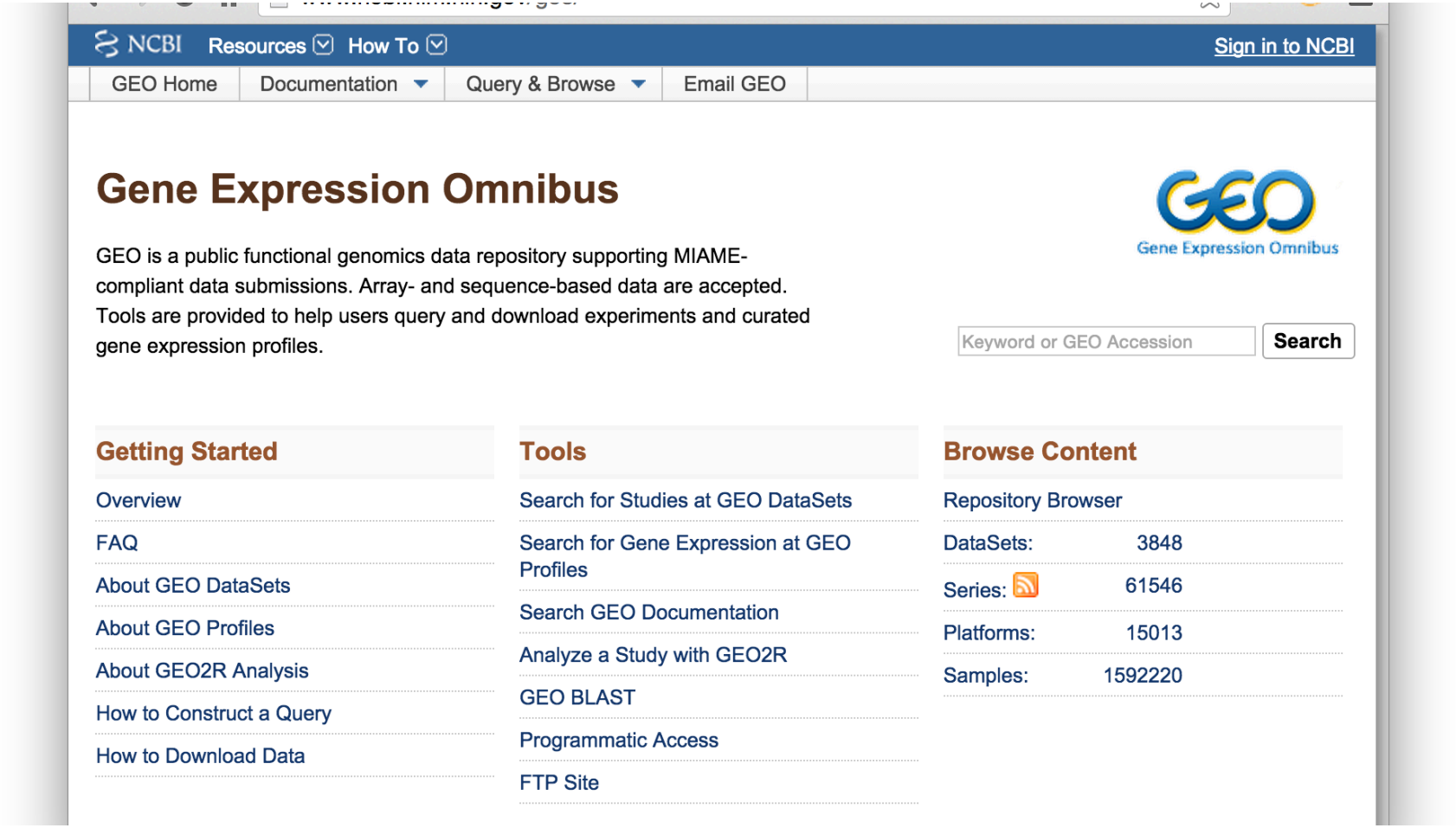
What can be entered in this field?

List of SRA Samples. 942991 found.

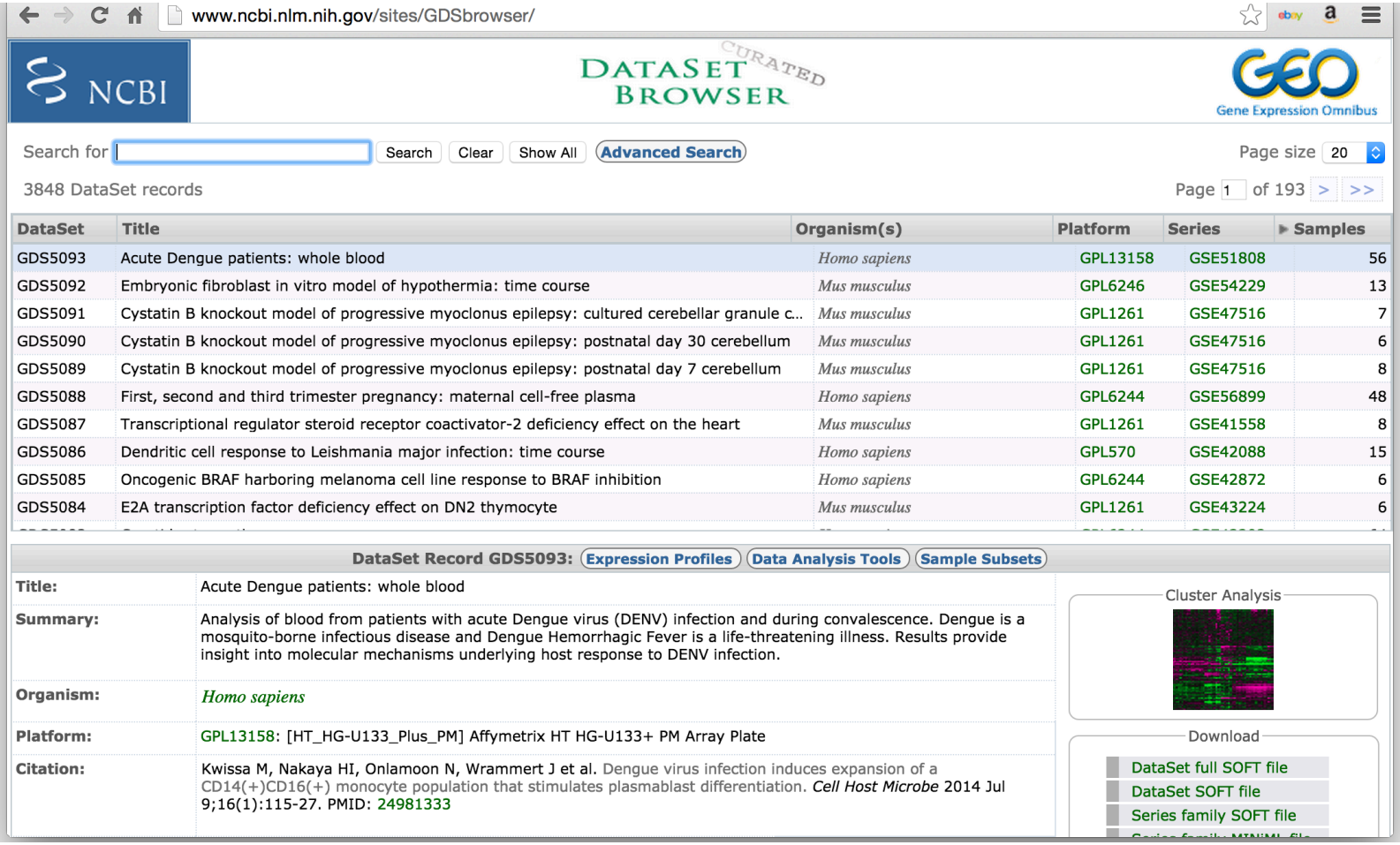
<< < Page 1 / 37720 > >>

#	Accession	Organism	Title	Attributes	Links
1.	SRS1095197	Lasionycteris noctivagans	Model organism or animal sample from Lasionycteris noctivagans	ecotype: **Wild** age: **Not collected** sex: **not collected** tissue: **muscle**	PRJNA209850 [undefined]
2.	SRS1095218	Lasiurus cinereus	Model organism or animal sample from Lasiurus cinereus	ecotype: **Wild** age: **Not collected** sex: **not collected** tissue: **muscle**	PRJNA209850 [undefined]
3.	SRS1083846	Cardinalis cardinalis	Model organism or animal sample from Cardinalis cardinalis	breed: **not applicable** age: **not collected** sex: **not applicable** tissue: **breast** specimen_voucher: **KU 25393**	
4.	SRS1083845	Cardinalis cardinalis	Model organism or animal sample from Cardinalis cardinalis	breed: **not applicable** age: **not collected** sex: **not applicable** tissue: **breast** specimen_voucher: **KU 21828**	
5.	SRS1083844	Piranga rubriceps	Model organism or animal sample from Piranga rubriceps	breed: **not applicable** age: **not collected** sex: **not applicable** tissue: **breast**	

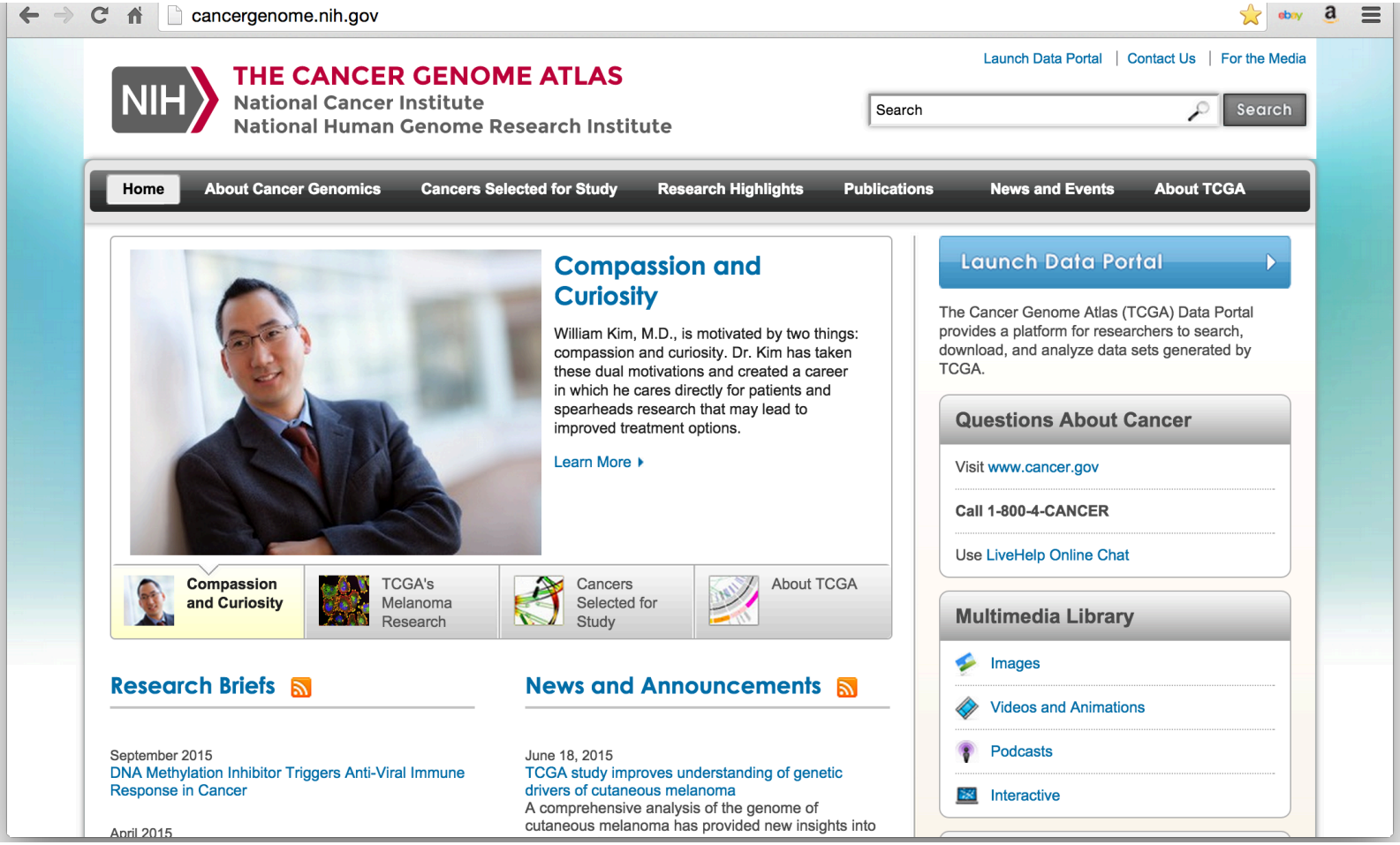
GEO


NCBI
Resources
How To
Sign in to NCBI
GEO Home
Documentation
Query & Browse
Email GEO
Gene Expression Omnibus
GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.
Gene Expression Omnibus
Keyword or GEO Accession
Search
Getting Started
Overview
FAQ
About GEO DataSets
About GEO Profiles
About GEO2R Analysis
How to Construct a Query
How to Download Data
Tools
Search for Studies at GEO DataSets
Search for Gene Expression at GEO Profiles
Search GEO Documentation
Analyze a Study with GEO2R
GEO BLAST
Programmatic Access
FTP Site
Browse Content
Repository Browser
DataSets: 3848
Series: 61546
Platforms: 15013
Samples: 1592220

GEO


www.ncbi.nlm.nih.gov/sites/GDSbrowser/
NCBI
DATASET BROWSER
CURATED
GEO
Gene Expression Omnibus
Search for
Search
Clear
Show All
Advanced Search
Page size 20
3848 DataSet records
Page 1 of 193 > >>
DataSet Title Organism(s) Platform Series Samples
GDS5093 Acute Dengue patients: whole blood Homo sapiens GPL13158 GSE51808 56
GDS5092 Embryonic fibroblast in vitro model of hypothermia: time course Mus musculus GPL6246 GSE54229 13
GDS5091 Cystatin B knockout model of progressive myoclonus epilepsy: cultured cerebellar granule c... Mus musculus GPL1261 GSE47516 7
GDS5090 Cystatin B knockout model of progressive myoclonus epilepsy: postnatal day 30 cerebellum Mus musculus GPL1261 GSE47516 6
GDS5089 Cystatin B knockout model of progressive myoclonus epilepsy: postnatal day 7 cerebellum Mus musculus GPL1261 GSE47516 8
GDS5088 First, second and third trimester pregnancy: maternal cell-free plasma Homo sapiens GPL6244 GSE56899 48
GDS5087 Transcriptional regulator steroid receptor coactivator-2 deficiency effect on the heart Mus musculus GPL1261 GSE41558 8
GDS5086 Dendritic cell response to Leishmania major infection: time course Homo sapiens GPL570 GSE42088 15
GDS5085 Oncogenic BRAF harboring melanoma cell line response to BRAF inhibition Homo sapiens GPL6244 GSE42872 6
GDS5084 E2A transcription factor deficiency effect on DN2 thymocyte Mus musculus GPL1261 GSE43224 6
DataSet Record GDS5093: Expression Profiles Data Analysis Tools Sample Subsets
Title: Acute Dengue patients: whole blood
Summary: Analysis of blood from patients with acute Dengue virus (DENV) infection and during convalescence. Dengue is a mosquito-borne infectious disease and Dengue Hemorrhagic Fever is a life-threatening illness. Results provide insight into molecular mechanisms underlying host response to DENV infection.
Organism: Homo sapiens
Platform: GPL13158: [HT_HG-U133_Plus_PM] Affymetrix HT HG-U133+ PM Array Plate
Citation: Kwissa M, Nakaya HI, Onlamoon N, Wrammert J et al. Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. Cell Host Microbe 2014 Jul 9;16(1):115-27. PMID: 24981333
Cluster Analysis
Download
DataSet full SOFT file
DataSet SOFT file
Series family SOFT file

TCGA


cancergenome.nih.gov
THE CANCER GENOME ATLAS
National Cancer Institute
National Human Genome Research Institute
Launch Data Portal | Contact Us | For the Media
Search
Search
Home
About Cancer Genomics
Cancers Selected for Study
Research Highlights
Publications
News and Events
About TCGA
Compassion and Curiosity
William Kim, M.D., is motivated by two things: compassion and curiosity. Dr. Kim has taken these dual motivations and created a career in which he cares directly for patients and spearheads research that may lead to improved treatment options.
Learn More
Compassion and Curiosity
TCGA's Melanoma Research
Cancers Selected for Study
About TCGA
Launch Data Portal
The Cancer Genome Atlas (TCGA) Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA.
Questions About Cancer
Visit www.cancer.gov
Call 1-800-4-CANCER
Use LiveHelp Online Chat
Multimedia Library
Images
Videos and Animations
Podcasts
Interactive
Research Briefs
September 2015
DNA Methylation Inhibitor Triggers Anti-Viral Immune Response in Cancer
April 2015
News and Announcements
June 18, 2015
TCGA study improves understanding of genetic drivers of cutaneous melanoma
A comprehensive analysis of the genome of cutaneous melanoma has provided new insights into

Steps involved on RNA-seq analysis

- Step 1: Preprocess
- Step 2: Map reads to the reference genome
- Step 3: Count how many reads fall within each feature of interest (gene, transcript, exon etc).
- Step 4: Normalization
- Step 5: Identify differentially expressed genes.

Example Dataset after Aligning Reads

Gene	Control			Treatment 1		
1	14	18	10	47	13	24
2	10	3	15	1	11	5
3	1	0	10	80	21	34
4	0	0	0	0	2	0
5	4	3	3	5	33	29
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
53256	47	29	11	71	278	339
Total	22910173	30701031	18897029	20546299	28491272	27082148

Biodonductor

- There already exists an extensive package of microarray analysis tools, called BioConductor, written in R.
- R and BioConductor are open source and free.
- Where is it?

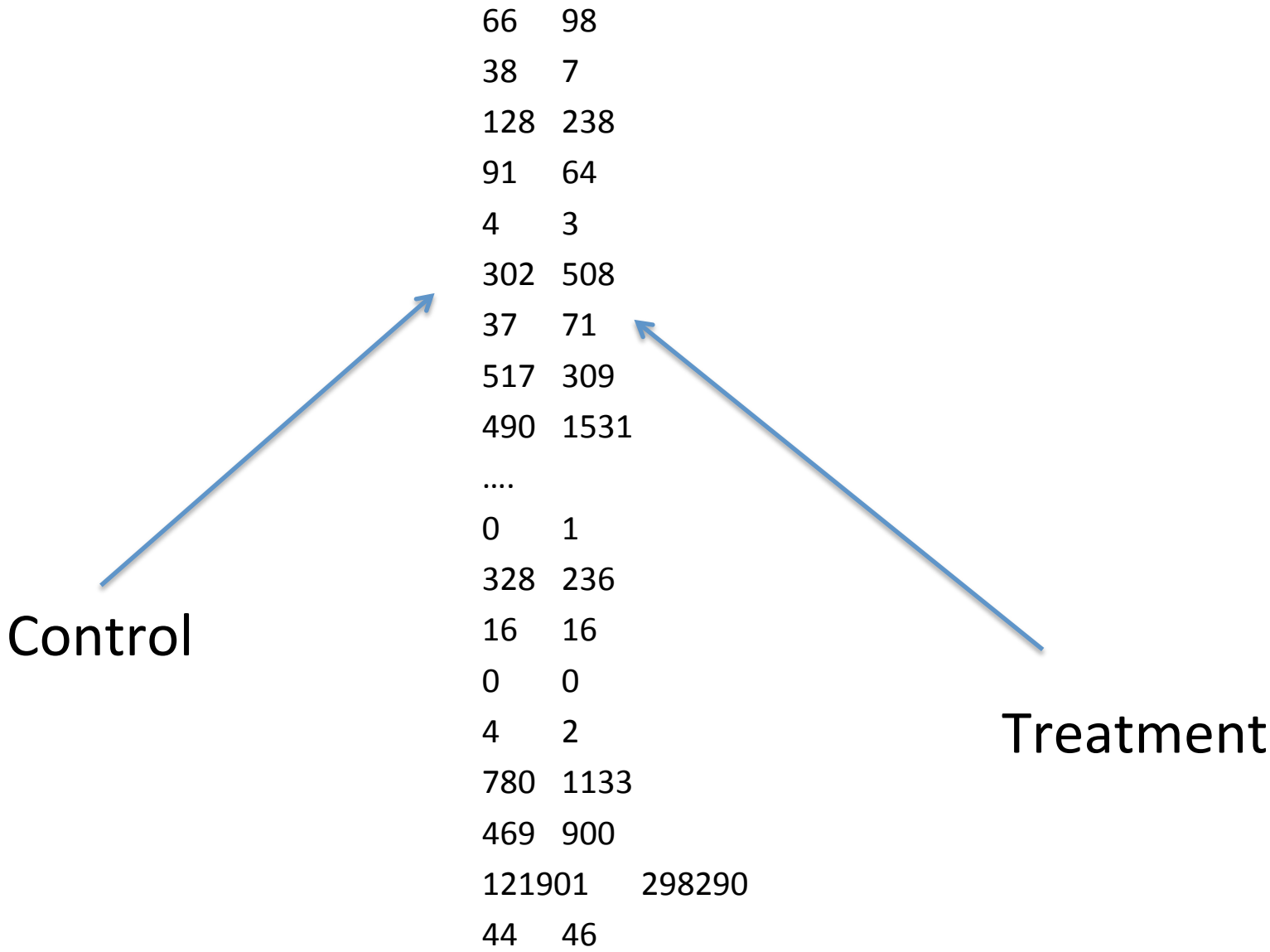
  http://www.bioconductor.org
- Installation
  - source("http://bioconductor.org/biocLite.R")
  - biocLite()

Using edgeR

Installation of edgeR

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("edgeR")
```

Input Data file



Using edgeR

```
> library(edgeR)
> library(stats)        #edgeR needs this lib
> y <- as.matrix(read.table("your_data_file", header=TRUE, row.names="Gene"))
> g=factor(c("Ctr","Ctr","Ctr","Tr","Tr","Tr"))
> d<- DGEList(counts=y, lib.size=colSums(y), group=g, remove.zeros = TRUE)
> d = calcNormFactors(d)
> d <- estimateCommonDisp(d)
> d <- estimateTagwiseDisp(d)
> ms <-  exactTest(d)
> result=topTags(ms, n=1000, adjust.method= "fdr")
```

Absent genes

- Remove absent genes (zero counts in all samples). It reduces the number of tests and the false discovery rate correction.
- Add 1 pseudocount (prevent dividing by 0).

Results

Comparison of groups: WT - Tr

	logConc	logFC	PValue	FDR
GRMZM2G304548	-3.040312	-0.8187952	0.000000e+00	0.000000e+00
GRMZM2G337229	-5.020442	0.6845438	0.000000e+00	0.000000e+00
GRMZM2G085260	-7.547953	-1.0193668	0.000000e+00	0.000000e+00
GRMZM2G060429	-2.328527	0.2535723	0.000000e+00	0.000000e+00
GRMZM2G092125	-8.148289	1.2663013	0.000000e+00	0.000000e+00
GRMZM2G123212	-10.028034	-2.1860488	0.000000e+00	0.000000e+00
GRMZM2G102356	-7.632246	1.5565755	0.000000e+00	0.000000e+00
GRMZM2G007256	-10.076032	-2.2233622	0.000000e+00	0.000000e+00
GRMZM2G168651	-4.621981	1.4587438	0.000000e+00	0.000000e+00
GRMZM2G322819	-30.122305	-39.7874987	0.000000e+00	0.000000e+00
GRMZM2G331701	-11.805408	-3.7628796	7.313298e-313	2.022459e-310
GRMZM2G068202	-11.807386	3.4861947	9.541348e-297	2.418732e-294
GRMZM2G162284	-9.951298	-1.7878318	3.360447e-254	7.863447e-252
GRMZM2G010783	-8.502862	-1.0699938	9.945640e-251	2.161045e-248

Default: it is sort by P-value

Using edgeR without replicate

```
> g=factor(c("Ctr”,"Tr"))
> d<- DGEList(counts=y, lib.size=colSums(y), group=g,
remove.zeros = TRUE)
> d = calcNormFactors(d)
> ms <-  exactTest(d, dispersion=0.04)
```

DESeq

- Installation?

```
> library(?)
> read file
> groups=c(“c”,”c”,”T”,”T”)
> d<-newCountDataSet(your_matix, groups)
> d<-estimateSizeFactors(d) #normalization
> sizeFactors(d)  # to see size factors
> d <- estimateDispersions (d)
> plotDispEsts(d) # to visualize dispersion
> res <- nbinomTest (d, "c", “T")
> write.csv(res,file=”your_file_name.csv")
```

DESeq without replicates

> groups=c(“c”,”T”)

> d<-newCountDataSet(your_matix, groups)

> d<-estimateSizeFactors(d) #normalization

> d <- estimateDispersions (d, method=“blind”, sharingMethod=“fit-only”)

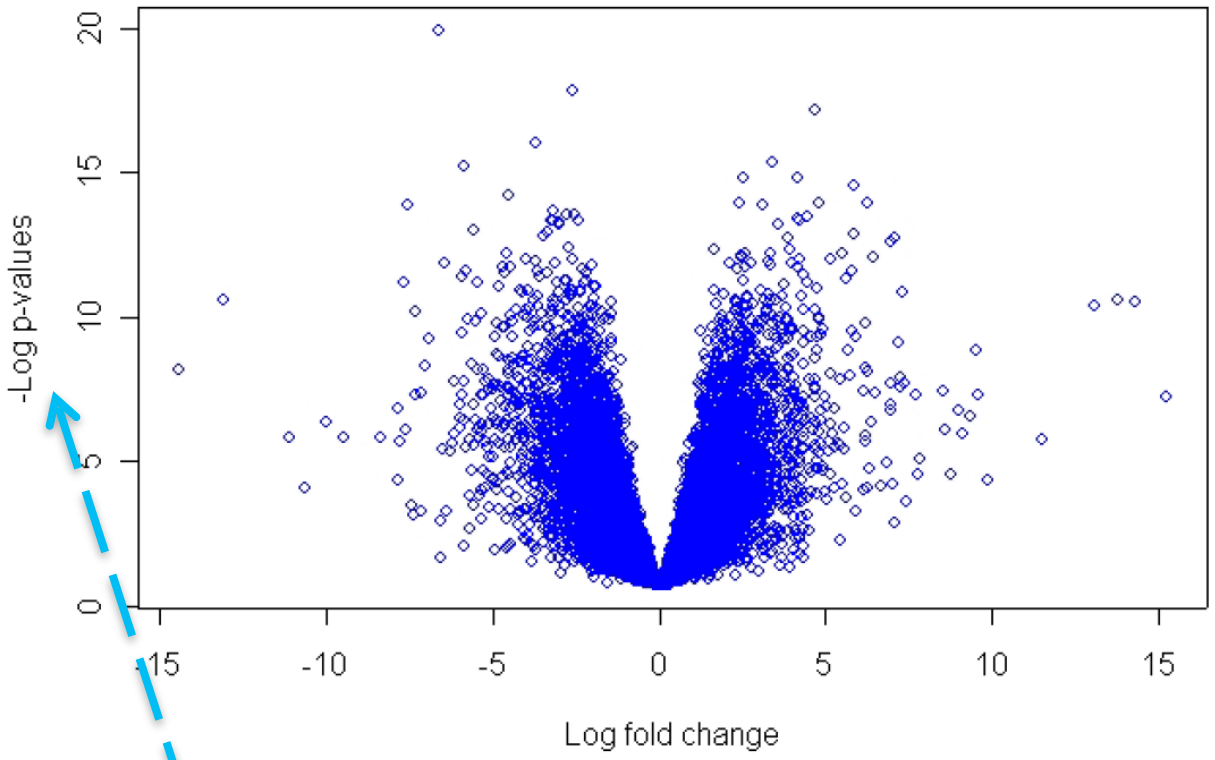
> res <- nbinomTest (d, "c", “T”)

Volcano plot

- In statistics, a **volcano plot** is a type of scatter-plot that is used to quickly identify changes in large datasets composed of replicate data.

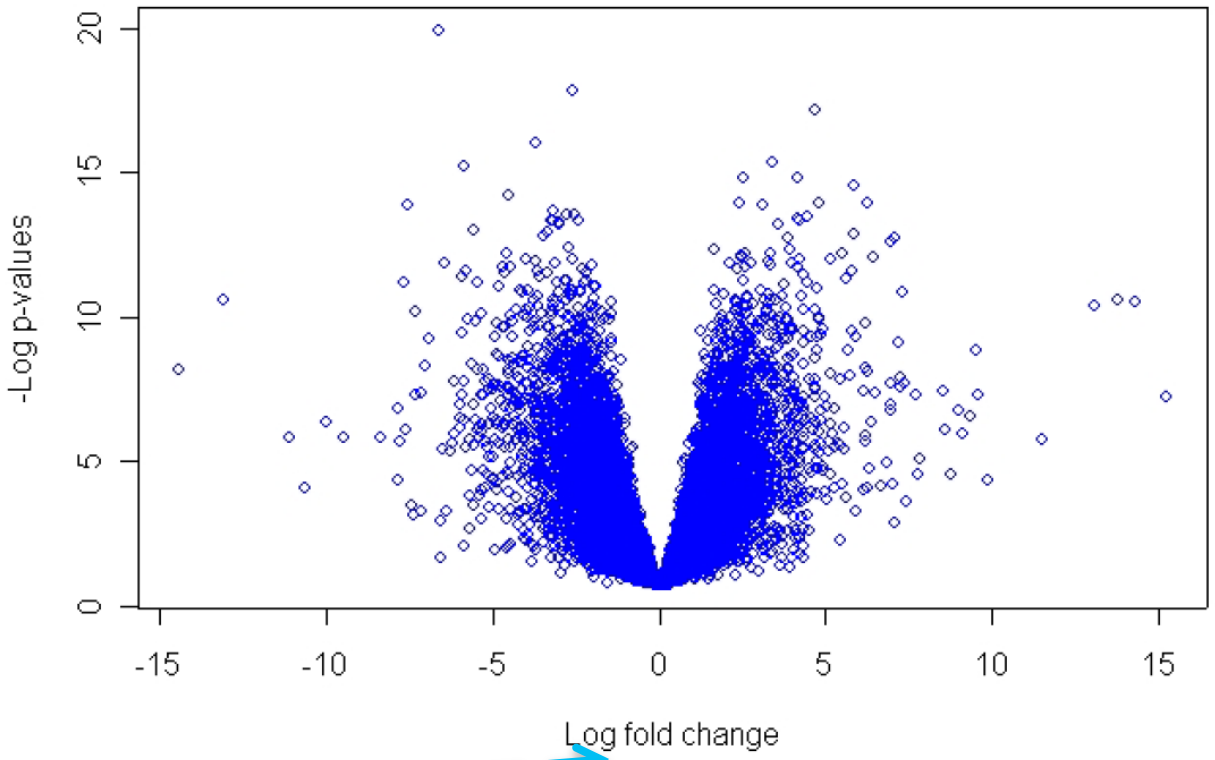
- It plots significance versus fold-change on the y- and x-axes, respectively.

Volcano plot



- A volcano plot is constructed by plotting the negative log of the p-value on the y-axis (usually base 10). This results in data points with low p-values (highly significant) appearing towards the top of the plot.

Volcano plot

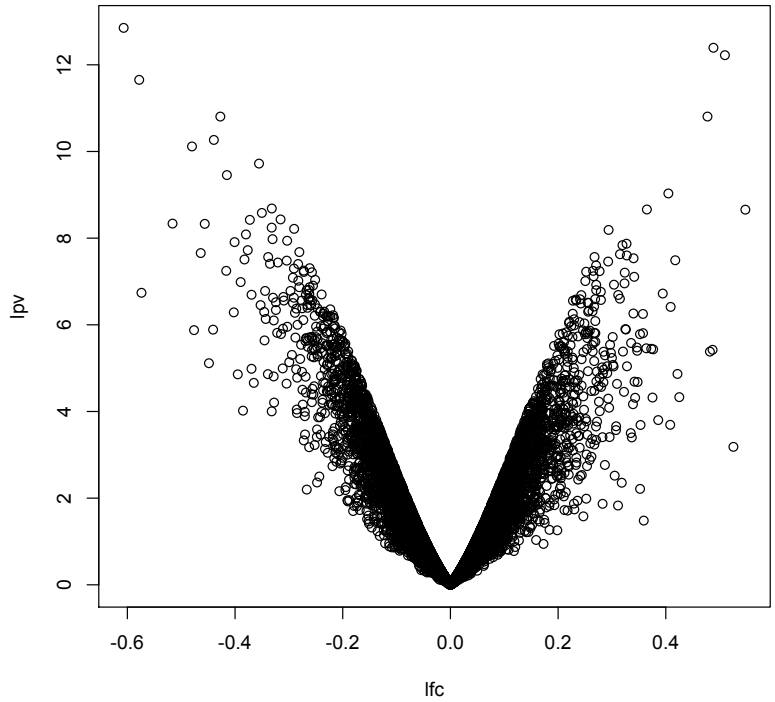


- The x-axis is the log of the fold change between the two conditions. The log of the fold-change is used so that changes in both directions (up and down) appear equidistant from the center.

Volcano plot: R

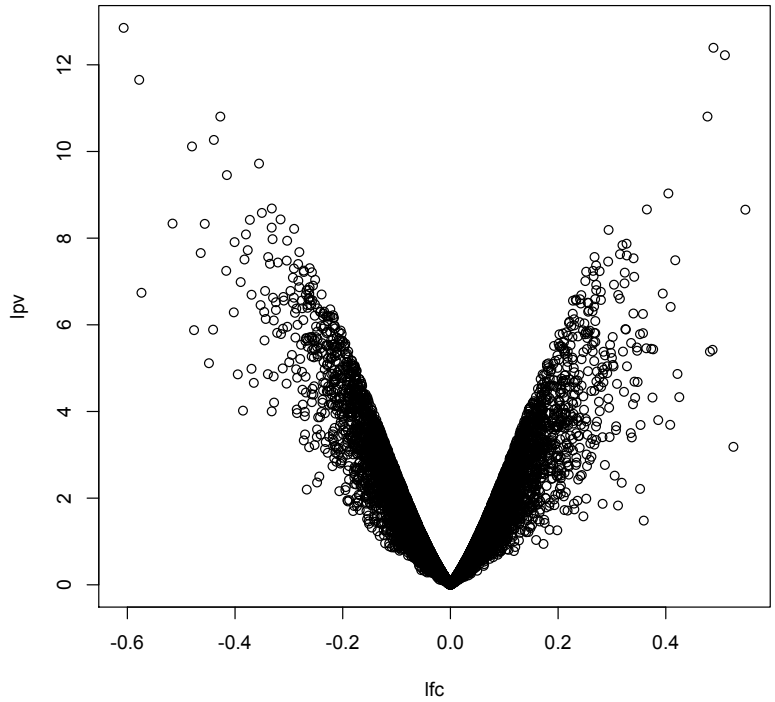
topTags

```
> results=topTags(ms, n=20000, adjust.method="fdr”)
> logfc=results[,2]
> logpvalue=-log2(results[,3])
> plot(logfc,logpvalue)
```



Volcano plot: R

```
> ms = exactTest(d)
> logfc=ms$table$logFC
> logpvalue=-log2(ms$table$PValue)
> plot(logfc,logpvalue)
```



HW5

- Using edgeR, DESeq, and volcano plot
-