Publicly Available Data & Data Integration

Lauren Vanderlinden, PhD, MS
T15 Postdoctoral Fellow Computational Biology
Division of Rheumatology & Department of Biomedical Informatics
School of Medicine, University of Colorado Anschutz Medical Campus

CPBS 7602 - December 11, 2024

Publicly Available Data

NCBI

GEO Gene Expression Omnibus

Free data!

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®.

- 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

EMBL-EBI

TCGA

GTEx Portal

PhenoGen Informatics

The site for quantitative genetics of the transcriptome.

GeneNetwork

University of Tennessee: www.genenetwork.org

RGD

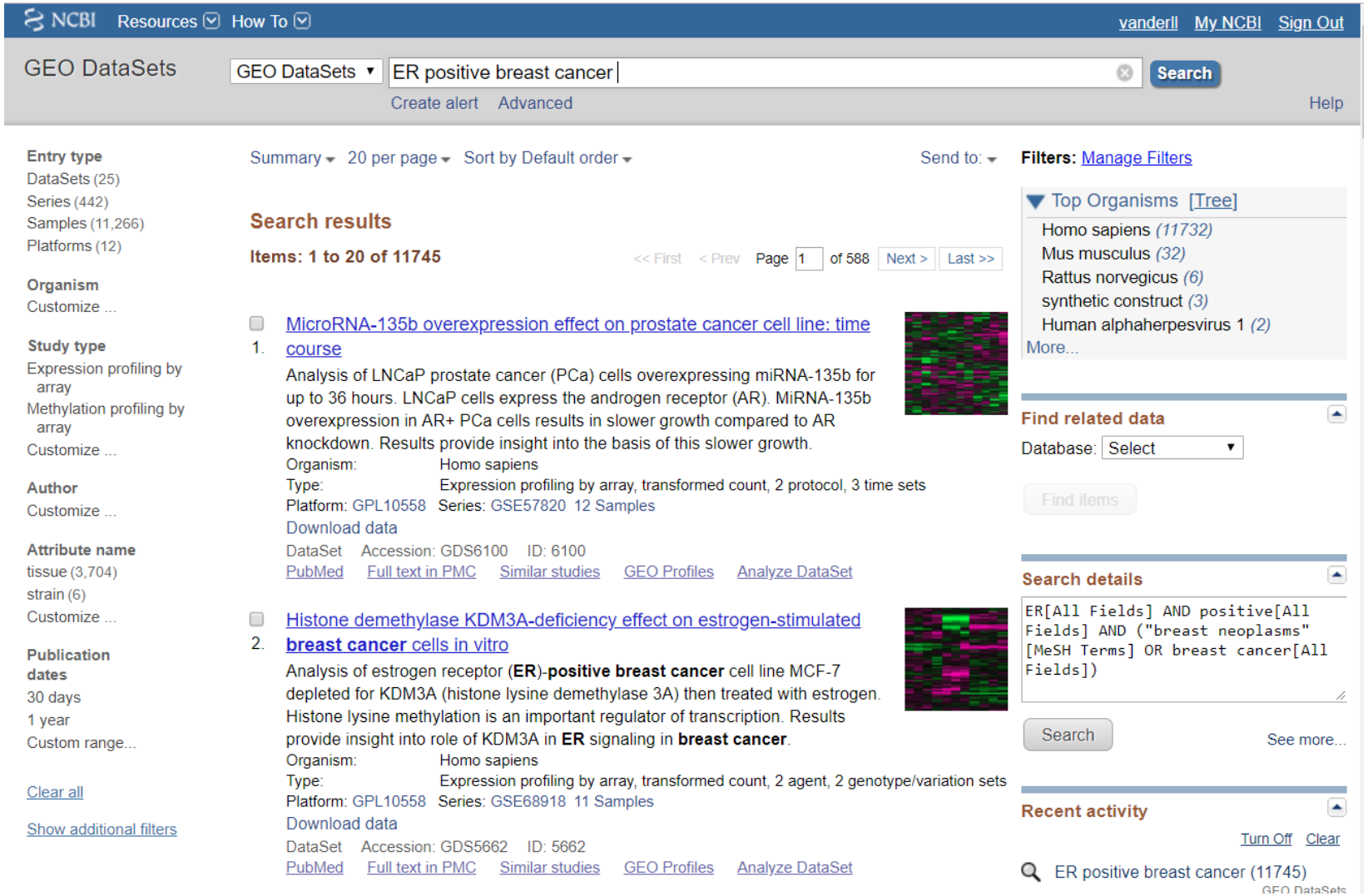
Students: always have data

Human Repositories

GEO Database

https://www.ncbi.nlm.nih.gov/geo/

- 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

NCBI

DATASET CURATED BROWSER

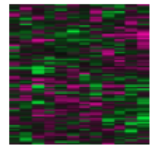
Gene Expression Omnibus

Search for GDS6100[ACCN] Search Clear Show All Advanced Search

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:	*Homo sapiens*		
Platform:	GPL10558: Illumina HumanHT-12 V4.0 expression beadchip		
Citation:	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		
Reference Series:	GSE57820	**Sample count:**	12
Value type:	transformed count	**Series published:**	2015/04/21

Cluster Analysis



Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

Data Analysis Tools

Find genes

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol: Go

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

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

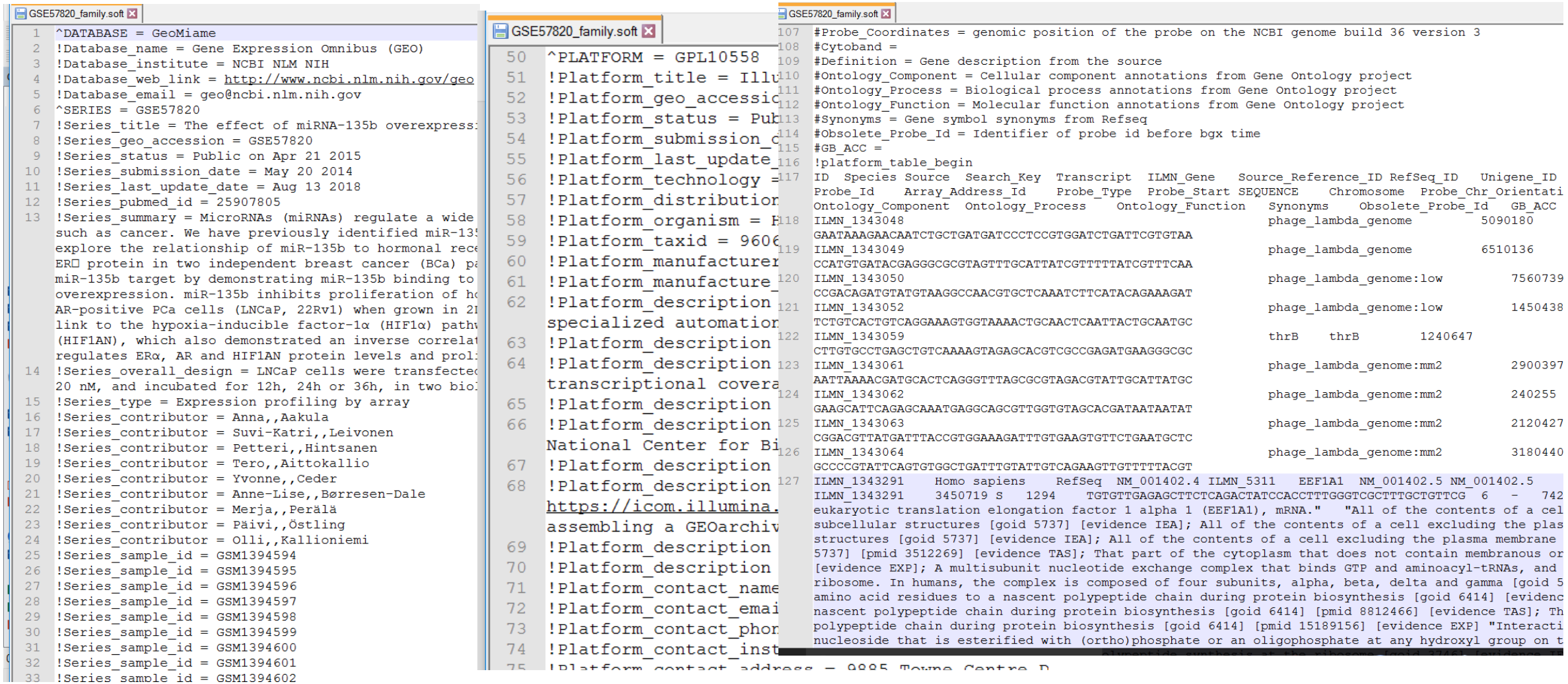
Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

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



GSE57820_family.soft

```
^DATABASE = GeoMiame
!Database_name = Gene Expression Omnibus (GEO)
!Database_institute = NCBI NLM NIH
!Database_web_link = http://www.ncbi.nlm.nih.gov/geo
!Database_email = geo@ncbi.nlm.nih.gov
^SERIES = GSE57820
!Series_title = The effect of miRNA-135b overexpress
!Series_geo_accession = GSE57820
!Series_status = Public on Apr 21 2015
!Series_submission_date = May 20 2014
!Series_last_update_date = Aug 13 2018
!Series_pubmed_id = 25907805
!Series_summary = MicroRNAs (miRNAs) regulate a wide
such as cancer. We have previously identified miR-13
explore the relationship of miR-135b to hormonal rec
ER protein in two independent breast cancer (BCa) pa
miR-135b target by demonstrating miR-135b binding to
overexpression. miR-135b inhibits proliferation of h
AR-positive PCa cells (LNCaP, 22Rv1) when grown in 2
link to the hypoxia-inducible factor-1α (HIF1α) path
(HIF1AN), which also demonstrated an inverse correla
regulates ERα, AR and HIF1AN protein levels and prol
!Series_overall_design = LNCaP cells were transfecte
20 nM, and incubated for 12h, 24h or 36h, in two bio
!Series_type = Expression profiling by array
!Series_contributor = Anna,,Aakula
!Series_contributor = Suvi-Katri,,Leivonen
!Series_contributor = Petteri,,Hintsanen
!Series_contributor = Tero,,Aittokallio
!Series_contributor = Yvonne,,Ceder
!Series_contributor = Anne-Lise,,Børresen-Dale
!Series_contributor = Merja,,Perälä
!Series_contributor = Päivi,,Östling
!Series_contributor = Olli,,Kallioniemi
!Series_sample_id = GSM1394594
!Series_sample_id = GSM1394595
!Series_sample_id = GSM1394596
!Series_sample_id = GSM1394597
!Series_sample_id = GSM1394598
!Series_sample_id = GSM1394599
!Series_sample_id = GSM1394600
!Series_sample_id = GSM1394601
!Series_sample_id = GSM1394602
```

GSE57820_family.soft

```
^PLATFORM = GPL10558
!Platform_title = Illu
!Platform_geo_accessio
!Platform_status = Pub
!Platform_submission_d
!Platform_last_update_
!Platform_technology =
!Platform_distribution
!Platform_organism = H
!Platform_taxid = 9606
!Platform_manufacturer
!Platform_manufacture_
!Platform_description
specialized automation
!Platform_description
!Platform_description
transcriptional covera
!Platform_description
!Platform_description
National Center for Bi
!Platform_description
!Platform_description
https://icom.illumina.
assembling a GEOarchiv
!Platform_description
!Platform_description
!Platform_contact_name
!Platform_contact_emai
!Platform_contact_phon
!Platform_contact_inst
!Platform_contact_address = 9885 Towne Centre D
```

GSE57820_family.soft

```
#Probe_Coordinates = genomic position of the probe on the NCBI genome build 36 version 3
#Cytoband =
#Definition = Gene description from the source
#Ontology_Component = Cellular component annotations from Gene Ontology project
#Ontology_Process = Biological process annotations from Gene Ontology project
#Ontology_Function = Molecular function annotations from Gene Ontology project
#Synonyms = Gene symbol synonyms from Refseq
#Obsolete_Probe_Id = Identifier of probe id before bgx time
#GB_ACC =
!platform_table_begin
ID  Species Source  Search_Key  Transcript  ILMN_Gene  Source_Reference_ID RefSeq_ID  Unigene_ID
Probe_Id  Array_Address_Id  Probe_Type  Probe_Start SEQUENCE  Chromosome  Probe_Chr_Orientati
Ontology_Component  Ontology_Process  Ontology_Function  Synonyms  Obsolete_Probe_Id  GB_ACC
ILMN_1343048                                                  phage_lambda_genome       5090180
GAATAAAGAACAATCTGCTGATGATCCCTCCGTGGATCTGATTCGTGTAA
ILMN_1343049                                                  phage_lambda_genome       6510136
CCATGTGATACGAGGGCGCGTAGTTTGCATTATCGTTTTTATCGTTTCAA
ILMN_1343050                                                  phage_lambda_genome:low       7560739
CCGACAGATGTATGTAAGGCCAACGTGCTCAAATCTTCATACAGAAAGAT
ILMN_1343052                                                  phage_lambda_genome:low       1450438
TCTGTCACTGTCAGGAAAGTGGTAAAACTGCAACTCAATTACTGCAATGC
ILMN_1343059                                                  thrB  thrB       1240647
CTTGTGCCTGAGCTGTCAAAAGTAGAGCACGTCGCCGAGATGAAGGGCGC
ILMN_1343061                                                  phage_lambda_genome:mm2       2900397
AATTAAAACGATGCACTCAGGGTTTAGCGCGTAGACGTATTGCATTATGC
ILMN_1343062                                                  phage_lambda_genome:mm2       240255
GAAGCATTCAGAGCAAATGAGGCAGCGTTGGTGTAGCACGATAATAATAT
ILMN_1343063                                                  phage_lambda_genome:mm2       2120427
CGGACGTTATGATTTACCGTGGAAAGATTTGTGAAGTGTTCTGAATGCTC
ILMN_1343064                                                  phage_lambda_genome:mm2       3180440
GCCCCGTATTCAGTGTGGCTGATTTGTATTGTCAGAAGTTGTTTTTACGT
ILMN_1343291  Homo sapiens  RefSeq  NM_001402.4 ILMN_5311  EEF1A1  NM_001402.5 NM_001402.5
ILMN_1343291  3450719 S  1294  TGTGTTGAGAGCTTCTCAGACTATCCACCTTTGGGTCGCTTTGCTGTTCG  6  -  742
eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), mRNA."  "All of the contents of a cel
subcellular structures [goid 5737] [evidence IEA]; All of the contents of a cell excluding the plas
structures [goid 5737] [evidence IEA]; All of the contents of a cell excluding the plasma membrane
5737] [pmid 3512269] [evidence TAS]; That part of the cytoplasm that does not contain membranous or
[evidence EXP]; A multisubunit nucleotide exchange complex that binds GTP and aminoacyl-tRNAs, and
ribosome. In humans, the complex is composed of four subunits, alpha, beta, delta and gamma [goid 5
amino acid residues to a nascent polypeptide chain during protein biosynthesis [goid 6414] [evidenc
nascent polypeptide chain during protein biosynthesis [goid 6414] [pmid 8812466] [evidence TAS]; Th
polypeptide chain during protein biosynthesis [goid 6414] [pmid 15189156] [evidence EXP] "Interacti
nucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on t
```

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

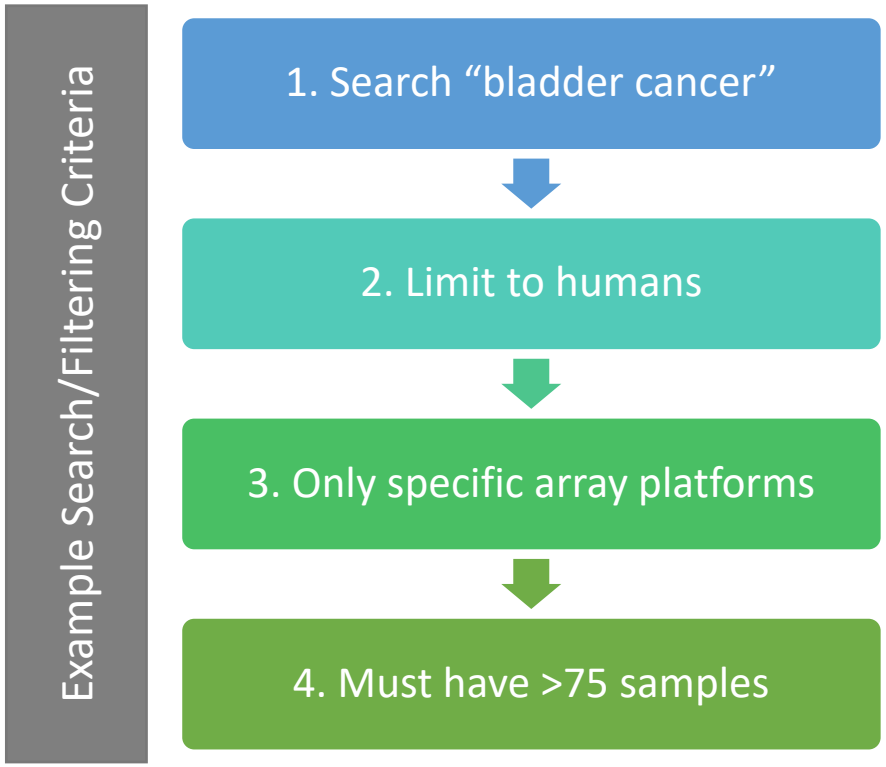
```
> table(pdata$characteristics_ch1.5)

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

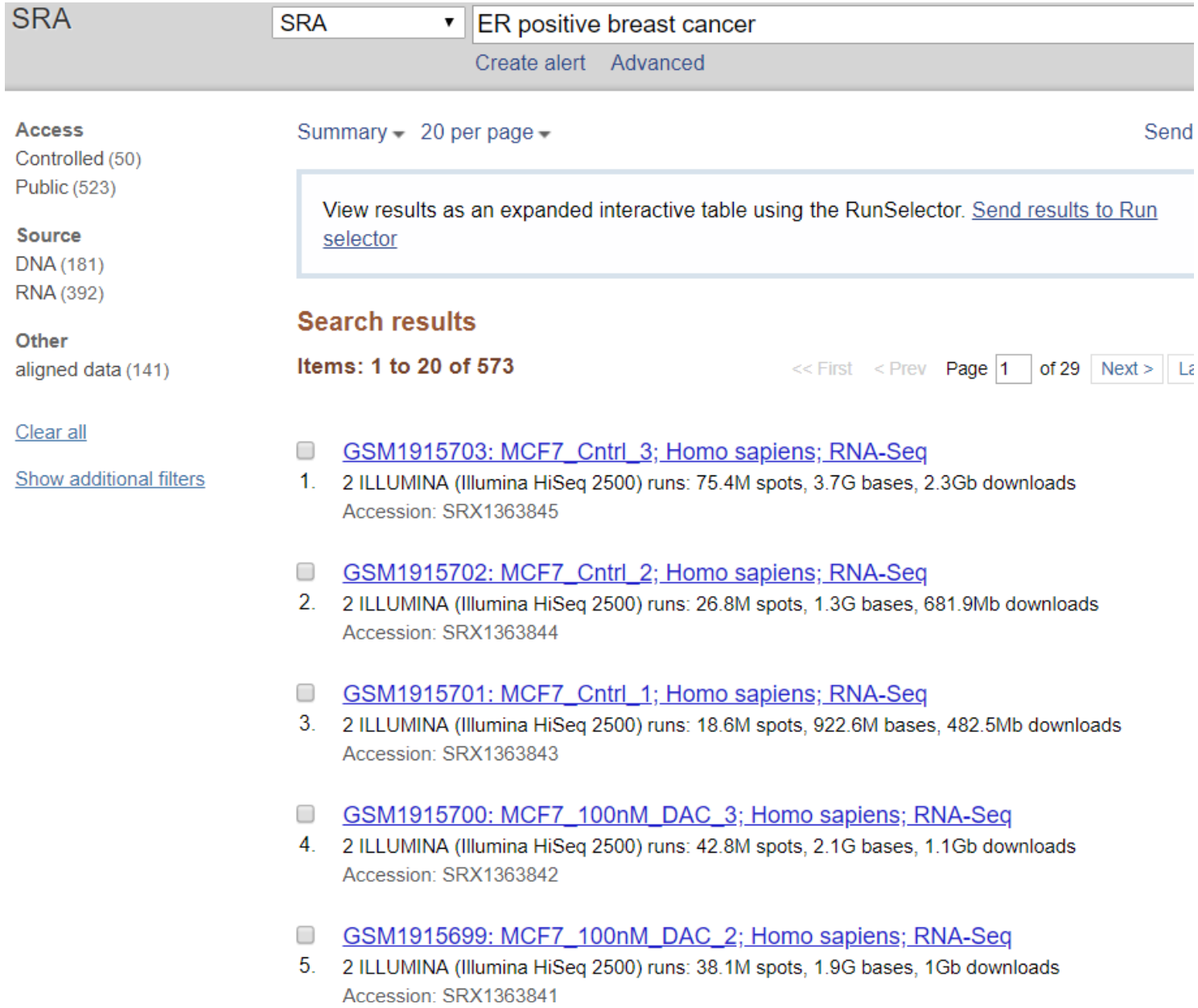
```
> dim(bladder.4)
[1] 3 8
> colnames(bladder.4)
[1] "Series"        "Organism"      "Title"         "Type"          "Platform"
  "Sample.Number" "Description"   "Term"
> bladder.4$Series
[1] "GSE31189" "GSE31684" "GSE13507"
>
```

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

SRA Database

www.ncbi.nlm.nih.gov/sra/

- NCBI Sequence Read Archive
- 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-specifi
of RNA using TruSeq stranded total RNA with Ribo-Zero Gold (Illumina)

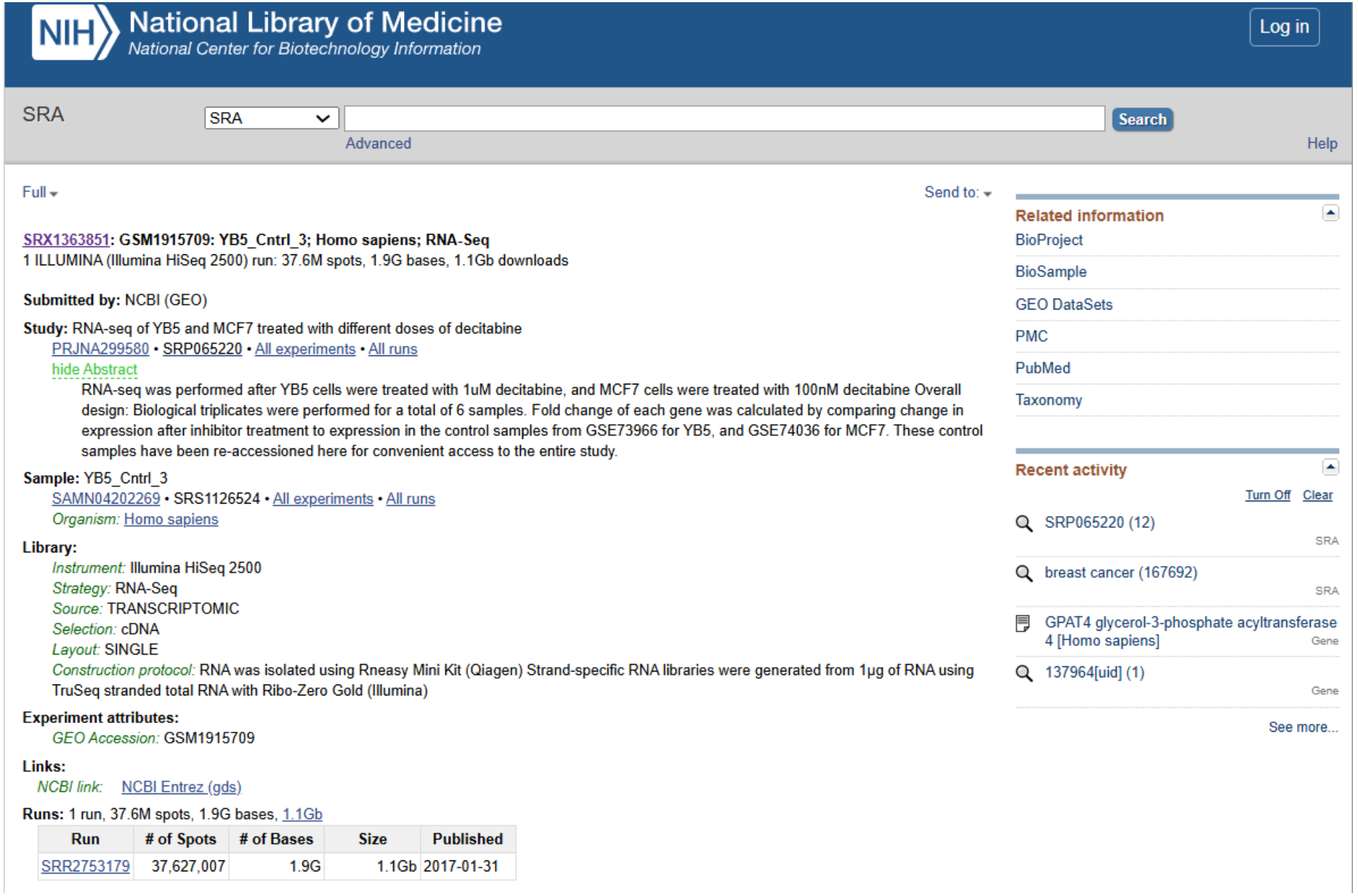
Experiment attributes:
GEO Accession: GSM1915703

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.

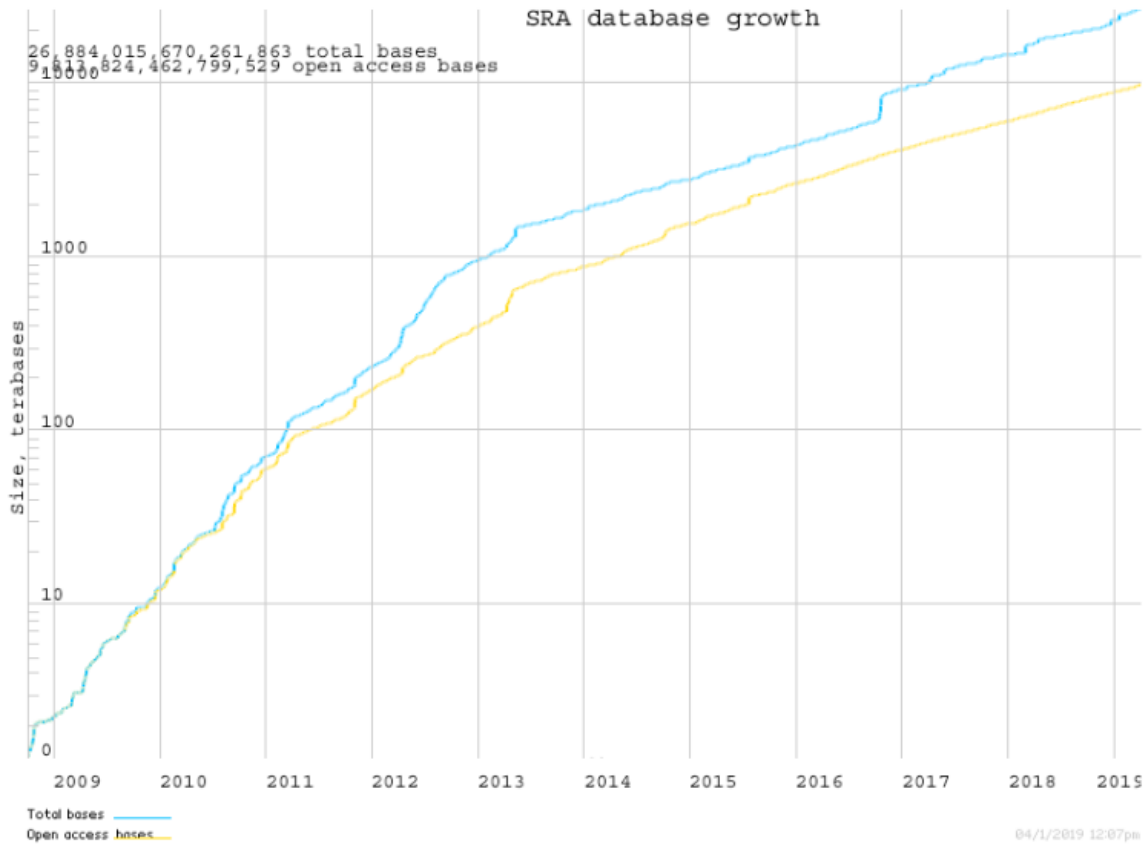
SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

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)



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

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

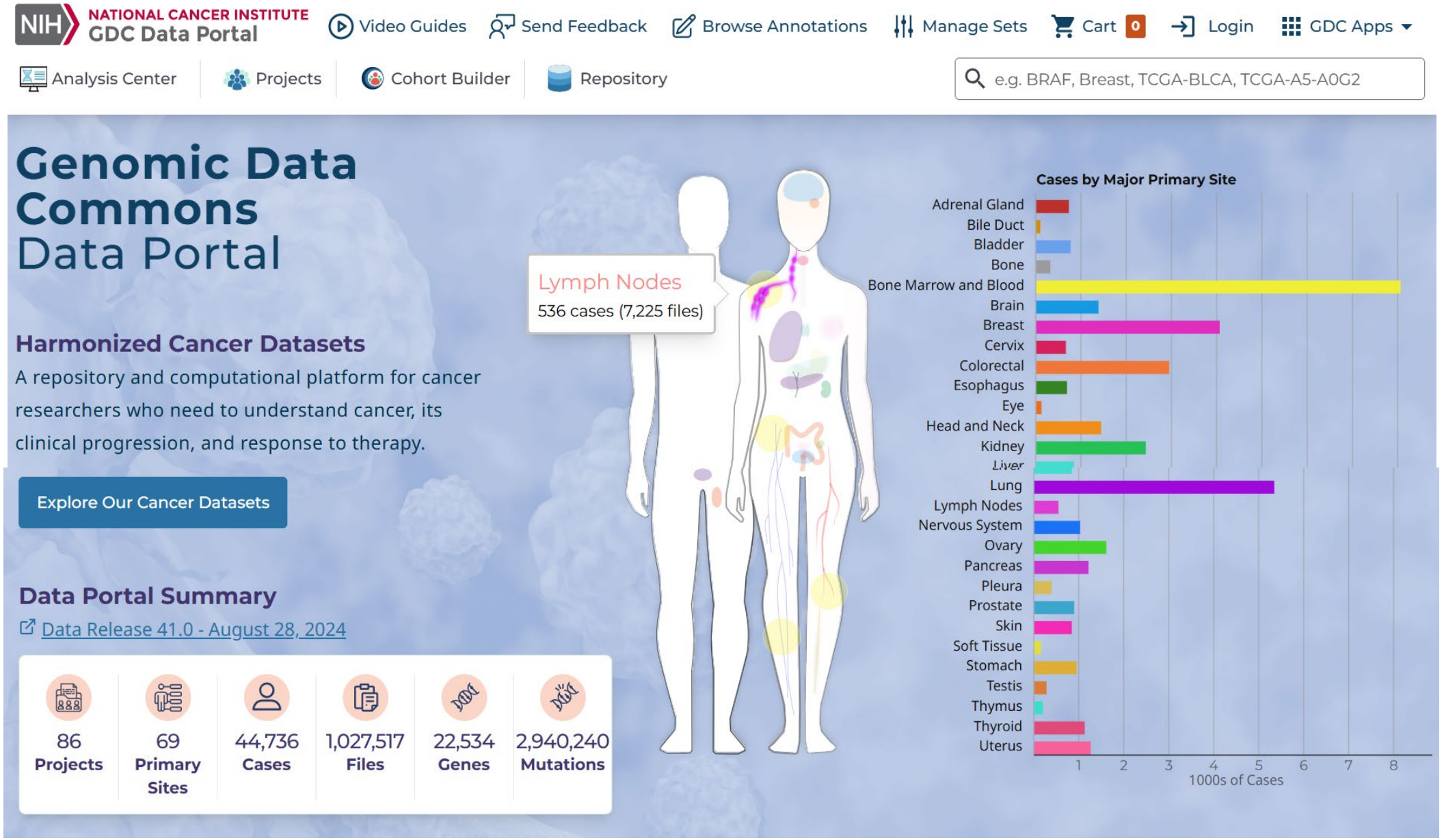
TCGA has changed our understanding of cancer, how research is conducted, how the disease is treated in the clinic, and more.

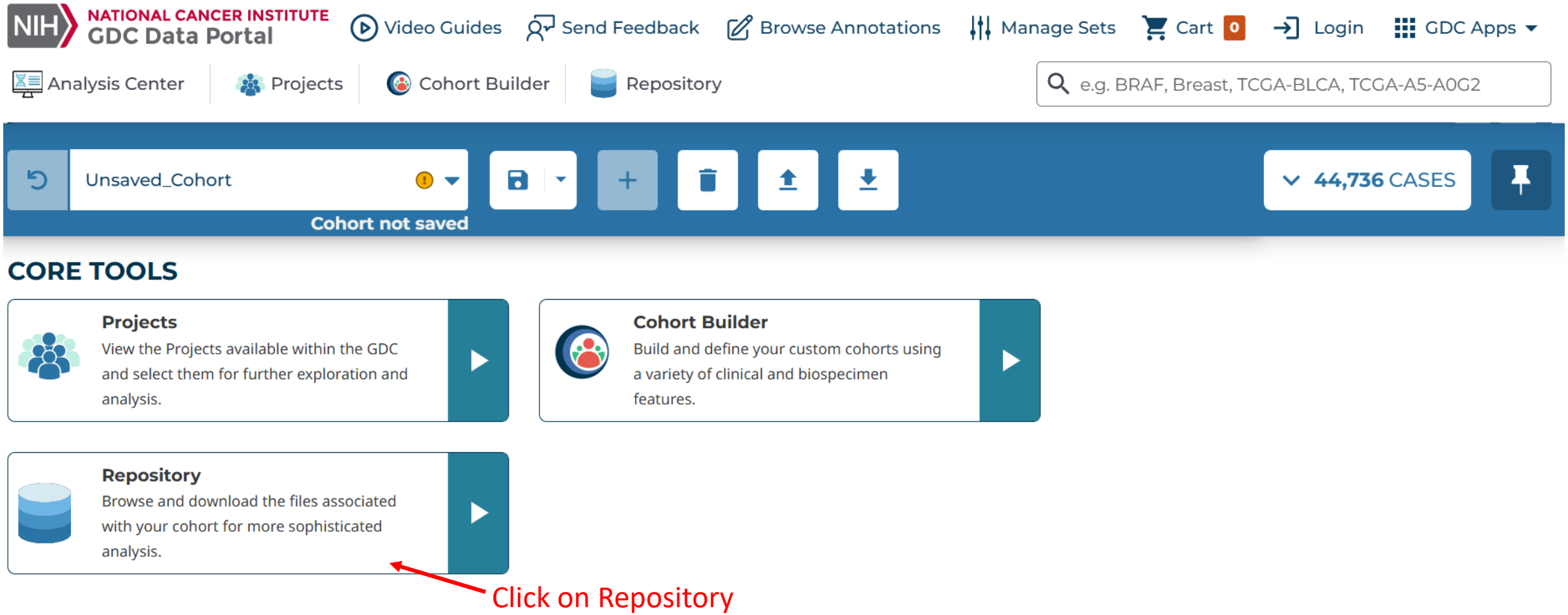
TCGA's PanCancer Atlas

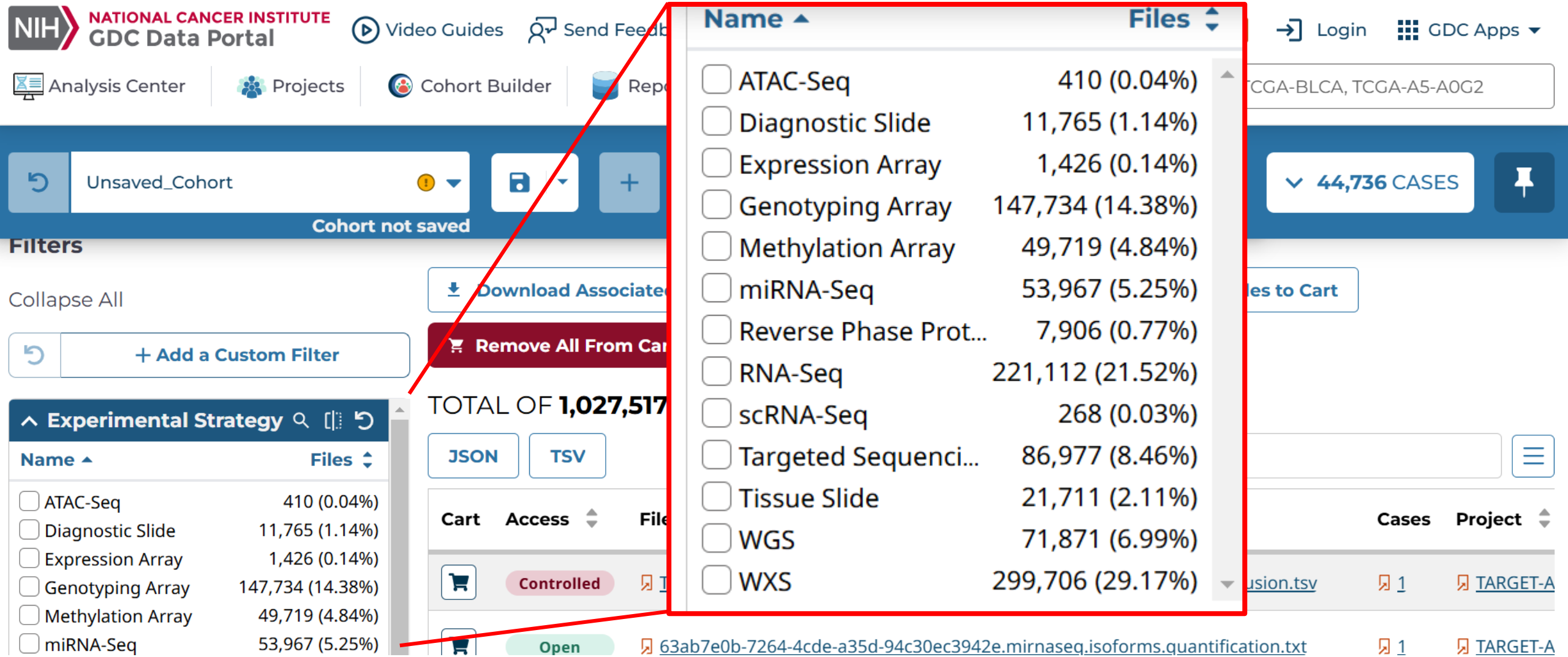
A collection of cross-cancer analyses delving into overarching themes on cancer, including cell-of-origin patterns, oncogenic processes and signaling pathways. Published in 2018 at the program's close.

TCGA – Data Portal

https://portal.gdc.cancer.gov/




NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Video Guides
Send Feedback
Browse Annotations
Manage Sets
Cart 0
Login
GDC Apps
Analysis Center
Projects
Cohort Builder
Repository
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2
Unsaved_Cohort
Cohort not saved
44,736 CASES
CORE TOOLS
Projects
View the Projects available within the GDC and select them for further exploration and analysis.
Cohort Builder
Build and define your custom cohorts using a variety of clinical and biospecimen features.
Repository
Browse and download the files associated with your cohort for more sophisticated analysis.
Click on Repository


NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal
Video Guides
Send Feedb
Login
GDC Apps
Analysis Center
Projects
Cohort Builder
Repo
TCGA-BLCA, TCGA-A5-A0G2
Unsaved_Cohort
Cohort not saved
44,736 CASES
Filters
Collapse All
+ Add a Custom Filter
Download Associate
Remove All From Car
les to Cart
TOTAL OF 1,027,517
JSON
TSV
Experimental Strategy
Name
Files
ATAC-Seq 410 (0.04%)
Diagnostic Slide 11,765 (1.14%)
Expression Array 1,426 (0.14%)
Genotyping Array 147,734 (14.38%)
Methylation Array 49,719 (4.84%)
miRNA-Seq 53,967 (5.25%)
Reverse Phase Prot... 7,906 (0.77%)
RNA-Seq 221,112 (21.52%)
scRNA-Seq 268 (0.03%)
Targeted Sequenci... 86,977 (8.46%)
Tissue Slide 21,711 (2.11%)
WGS 71,871 (6.99%)
WXS 299,706 (29.17%)
Cart
Access
File
Cases
Project
Controlled
usion.tsv
1
TARGET-A
Open
63ab7e0b-7264-4cde-a35d-94c30ec3942e.mirnaseq.isoforms.quantification.txt
1
TARGET-A

NIH NATIONAL CANCER INSTITUTE
GDC Data Portal

Video Guides | Send Feedback | Browse Annotations | Manage Sets | Cart 0 | Login | GDC Apps

Analysis Center | Projects | Cohort Builder | Repository

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

Unsaved_Cohort

Cohort not saved

44,736 CASES

ANALYSIS TOOLS

BAM Slicing Download
25,269 Cases

Clinical Data Analysis (Demo)
44,736 Cases

Cohort Comparison (Demo)
44,736 Cases

Cohort Level MAF
17,771 Cases

Gene Expression Clustering (Demo)
20,712 Cases

Mutation Frequency (Demo)
18,640 Cases

OncoMatrix (Demo)
18,640 Cases

ProteinPaint (Demo)
16,508 Cases

Sequence Reads
25,269 Cases

Set Operations (Demo)

R/TCGAbiolinks

TCGAbiolinks

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

build warnings | updated < 1 month

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

DOI: 10.18129/B9.bioc.TCGAbiolinks

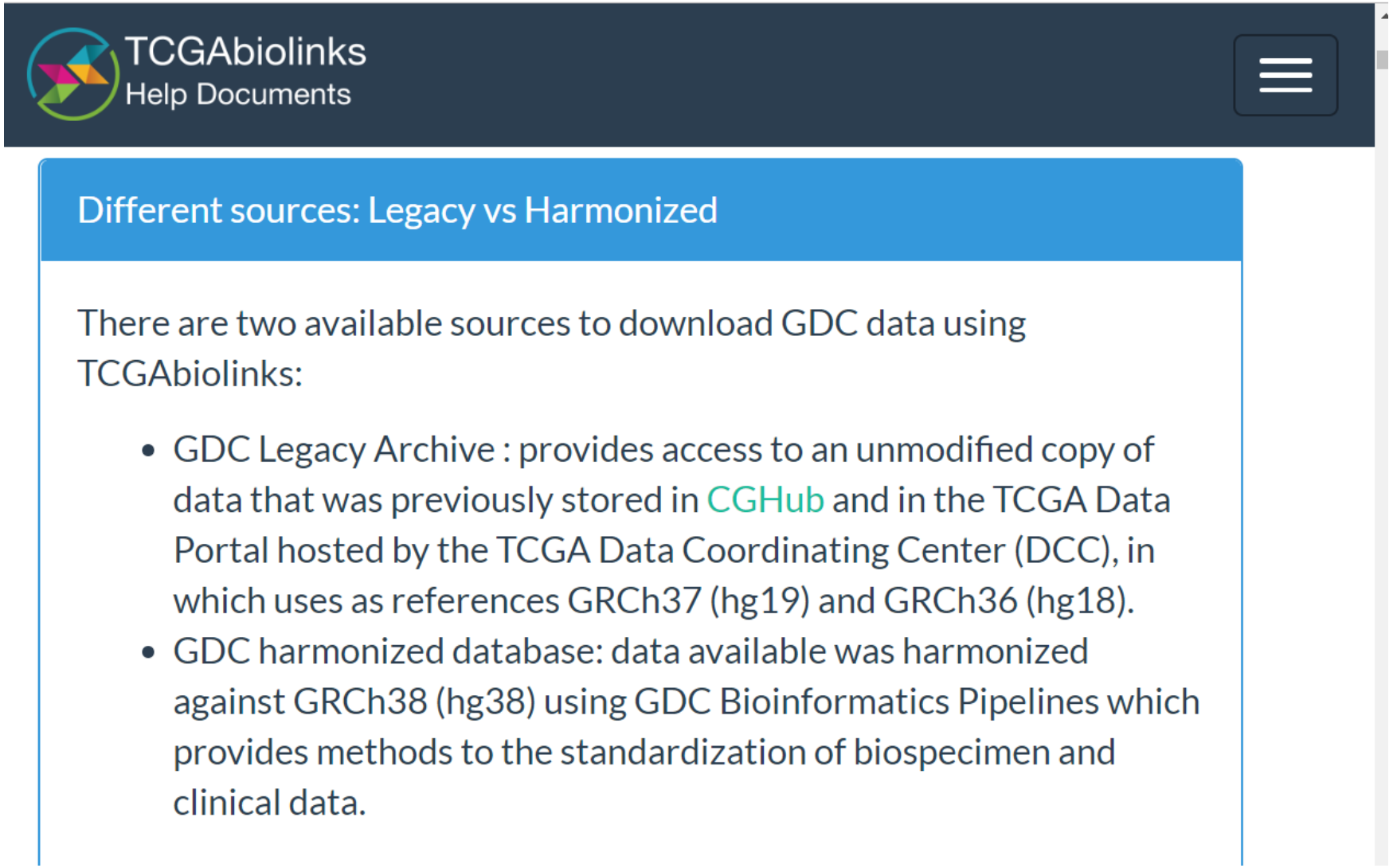
TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data

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



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/

The Genotype-Tissue Expression (GTEx) Portal is a comprehensive public resource for researchers studying tissue and cell-specific gene expression and regulation across individuals, development, and species, with data from 3 NIH projects.

GTEx

The Adult GTEx project is a comprehensive resource of WGS, RNA-Seq, and QTL data from samples collected from 54 non-diseased tissue sites across ~1000 adult individuals.

Explore »

dGTEx

The Developmental GTEx (dGTEx) project is a new effort to study development-specific genetic effects on gene expression and to establish a new data analysis and tissue biobank resource.

**Data Not Yet Available*

Explore »

NHP dGTEx

The Non-Human Primate Developmental GTEx (NHP-dGTEX) project is a complement to dGTEx in 2 translational non-human primate model species: the rhesus macaque and common marmoset.

**Data Not Yet Available*

Explore »

Adult GTEx Data and Resources

OPEN ACCESS

Expression

RNA-seq
Read counts and/or normalized values gene, exon, transcript, and junction level

Small RNA-seq
Read counts

Haplotype expression matrices
Haplotype counts from phASER

Long-read RNA-seq
ONT data from 88 RNA samples

snRNA-seq
Single nucleus RNA-seq from 24 tissues samples

Association/QTLs

Single-tissue cis-QTL
Expression, splicing, interaction, sex-biased expression, fine-mapping

Small RNA-seq cis-QTL
Expression QTLs

Single-tissue trans-QTL
Expression, splicing

Multi-tissue cis-QTL
Expression

Other

Limited donor phenotypes
Sex, 10-year age brackets, Hardy scale

Reference files
Human genome reference, gene models, variant lookup table

Histology
Aperio SVS images, histology data

PROTECTED ACCESS

Available on AnVIL and requires additional access request.

Sequencing

RNA-seq
BAM/CRAM files

WGS, WES
VCFs, CRAMs, and BAMs

Expression + SNP arrays

Allele-Specific Expression (ASE) tables

Other

All de-identified donor phenotypes
Age, race, weight, smoking status, etc.

All de-identified sample attributes

Gene e ...s now show V10 data. We are working on integrating V10 into the remaining visualizations. See the News Item for more.

Download Open Access Datasets

Select project: Adult GTEx

Overview | Bulk tissue expression | QTL | Single cell | Long read data | Haplotype expression | Variants | Reference | Metadata

Additional GTEx datasets

GTEx Analysis V10

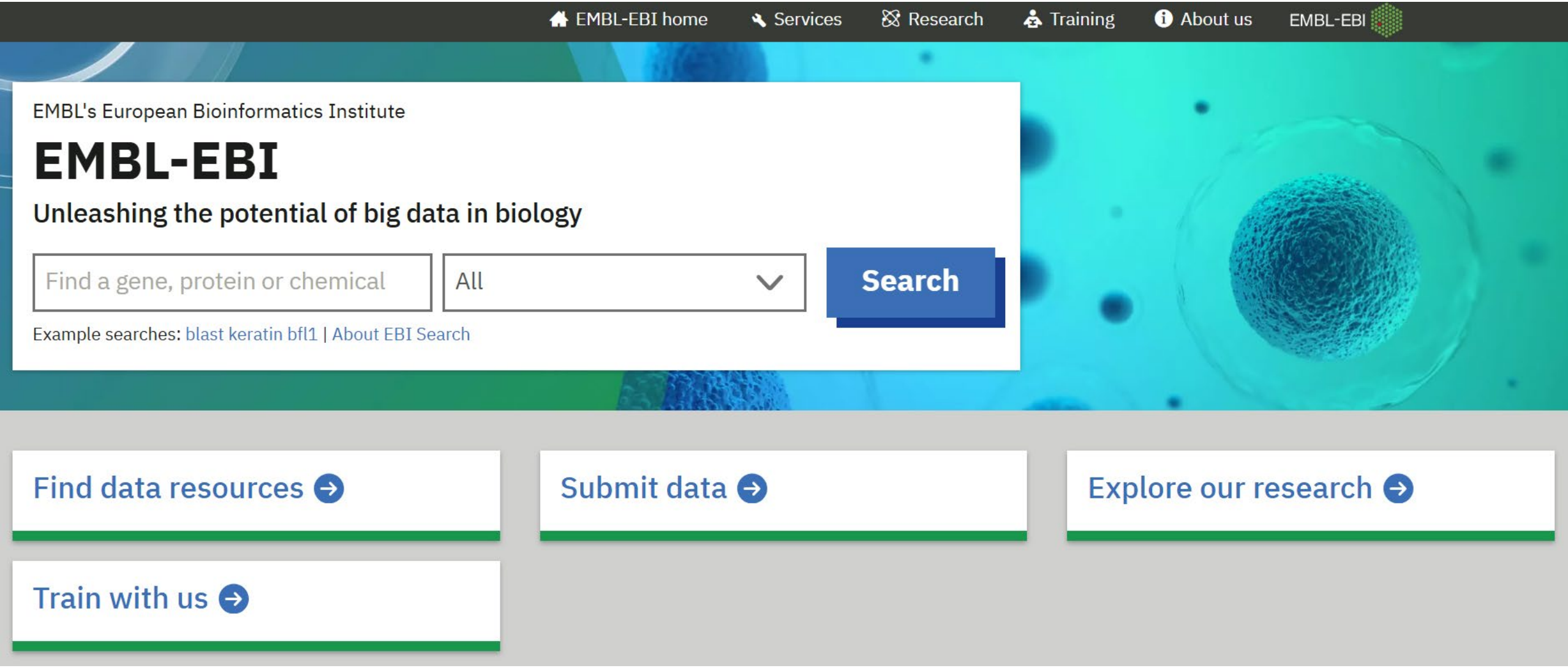
The GTEx Analysis V10 release is the most complete analyzed dataset for Adult GTEx. It includes a new data type, smallRNA-seq.

RNA-Seq

Name	Description	Size	
GTEx_Analysis_v10_RNASeQCv2.4.2_gene_reads.gct.gz	Gene Expression Read Counts from RNASeQCv2.4.2.	898 MB	

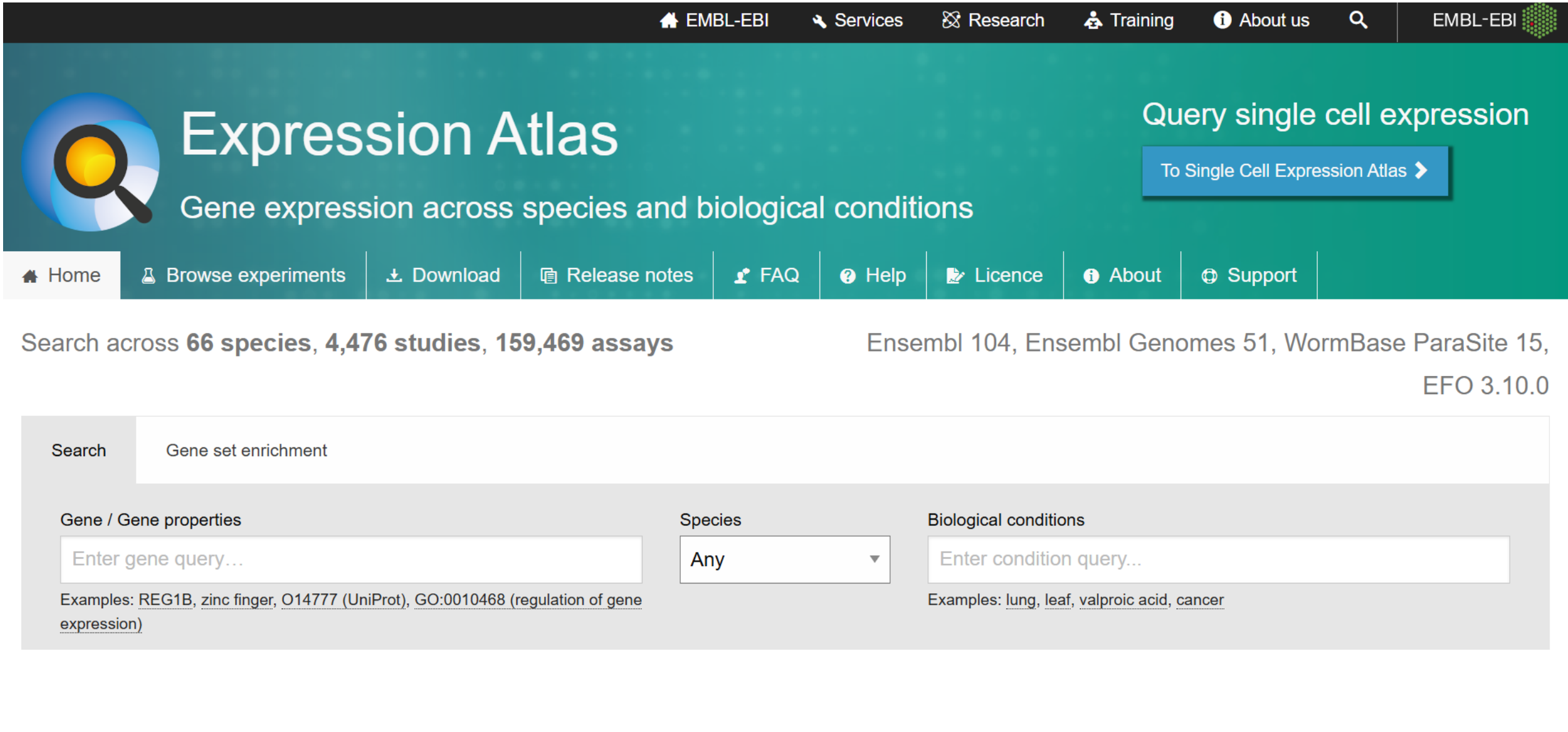
European Bioinformatics Institute

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



EBI Expression Atlas

https://www.ebi.ac.uk/gxa/home



GWAS Catalog

https://www.ebi.ac.uk/gwas/

Our latest resource update paper is now out in Nucleic Acids Research! Find out more here!

Download

Download a full copy of the GWAS Catalog in spreadsheet format as well as current and older versions of the GWAS diagram in SVG format.

Summary statistics

Documentation and access to full summary statistics for GWAS Catalog studies where available.

Submit

Submit summary statistics to GWAS Catalog.

Learn

Including FAQs, our curation process, training materials, related resources, a list of abbreviations and API documentation.

Diagram

Explore an interactive visualisation of all SNP-trait associations with genome-wide significance ($p \leq 5 \times 10^{-8}$).

Population descriptors

An introduction to our data extraction and standardisation process.

GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies

rheumatoid arthritis

Examples: Parkinson disease, rs3093017, Yao, 2q37.2, HBS1L, 6:167120000-167130000, GCST90132222, PMID:35241825

Refine search results

- Publications 74
- Traits 58

Other search filters

Browse all studies

Catalog stats

- *Last data release on 2024-11-20*

Search results for *rheumatoid arthritis*

T rheumatoid arthritis EFO_0000685

A chronic systemic disease, primarily of the joints, marked by inflammatory changes in the synovial membranes and articular structures, widespread fibrinoid degeneration of the collagen fibers in mese... Show more >

Associations 3563 Studies 168

T ACPA-negative rheumatoid arthritis EFO_0009460

A subtype of rheumatoid arthritis defined by the absence of autoantibodies that are directed against citrullinated peptides and proteins.

Associations 7 Studies 3

Associations 3563

Show 5 entries

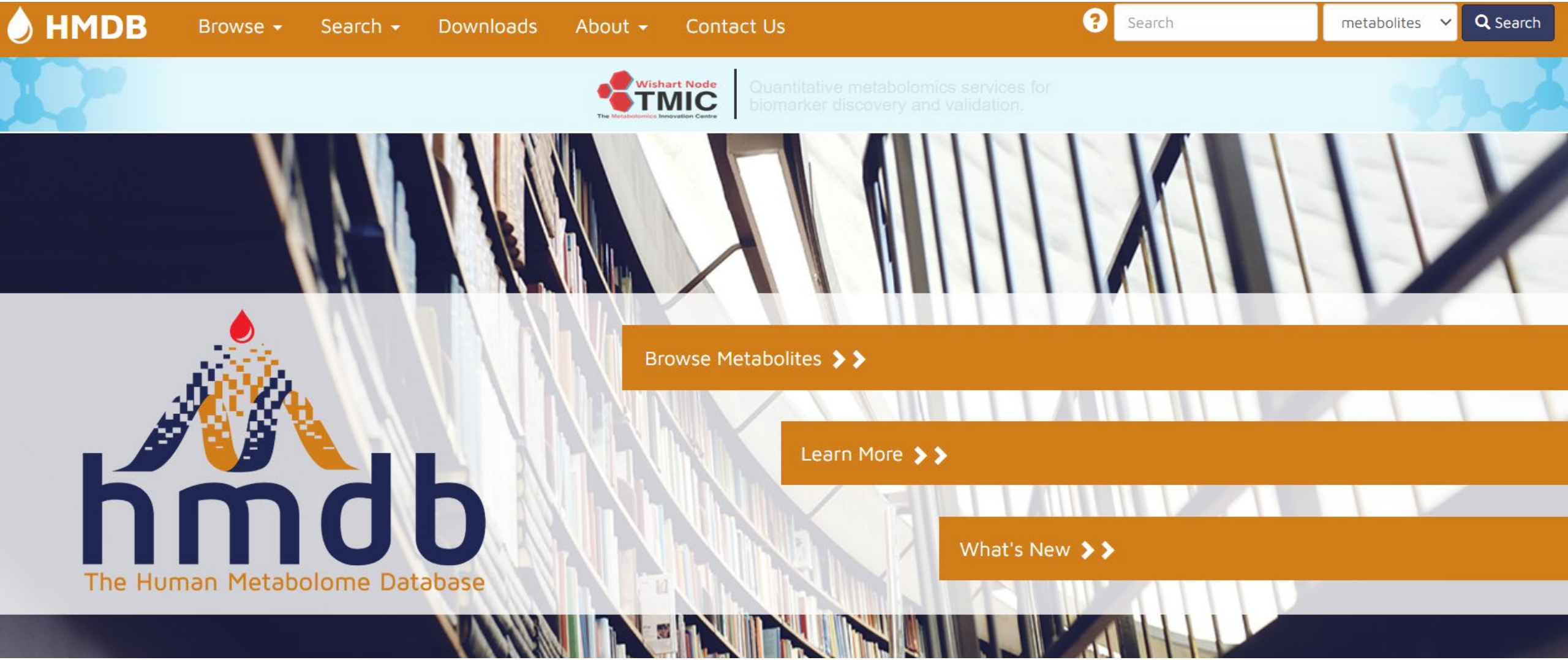
Column visibility | Export | Clear search

Variant and risk allele	P-value	P-value annotation	RAF	OR	Beta	CI	Mapped gene	Reported trait	Trait(s)	Background trait(s)	Study accession	Location
rs34536443-C	2 x 10^{-6}	-	0.04	0.81	-	-	TYK2	Rheumatoid arthritis (rheumatoid factor and anti-cyclic citrullinated peptide seronegative)	rheumatoid arthritis, ACPA-negative rheumatoid arthritis	-	GCST90131439	19:10352442
rs2476601-A	3 x 10^{-27}	-	0.0965	1.29	-	-	PTPN22	Rheumatoid arthritis (rheumatoid factor and anti-cyclic citrullinated peptide seronegative)	rheumatoid arthritis, ACPA-negative rheumatoid arthritis	-	GCST90131439	1:113834946
rs11889341-T	2 x 10^{-6}	-	0.22	1.09	-	-	STAT4	Rheumatoid arthritis (rheumatoid factor and anti-cyclic citrullinated peptide seronegative)	rheumatoid arthritis, ACPA-negative rheumatoid arthritis	-	GCST90131439	2:191079016
rs3134883-A	4 x 10^{-7}		0.33	1.09			IL2RA	Rheumatoid arthritis	rheumatoid arthritis		GCST90131439	10:6058762

Really hard to get full genetic data due to privacy and confidentiality, but summary statistics are readily available

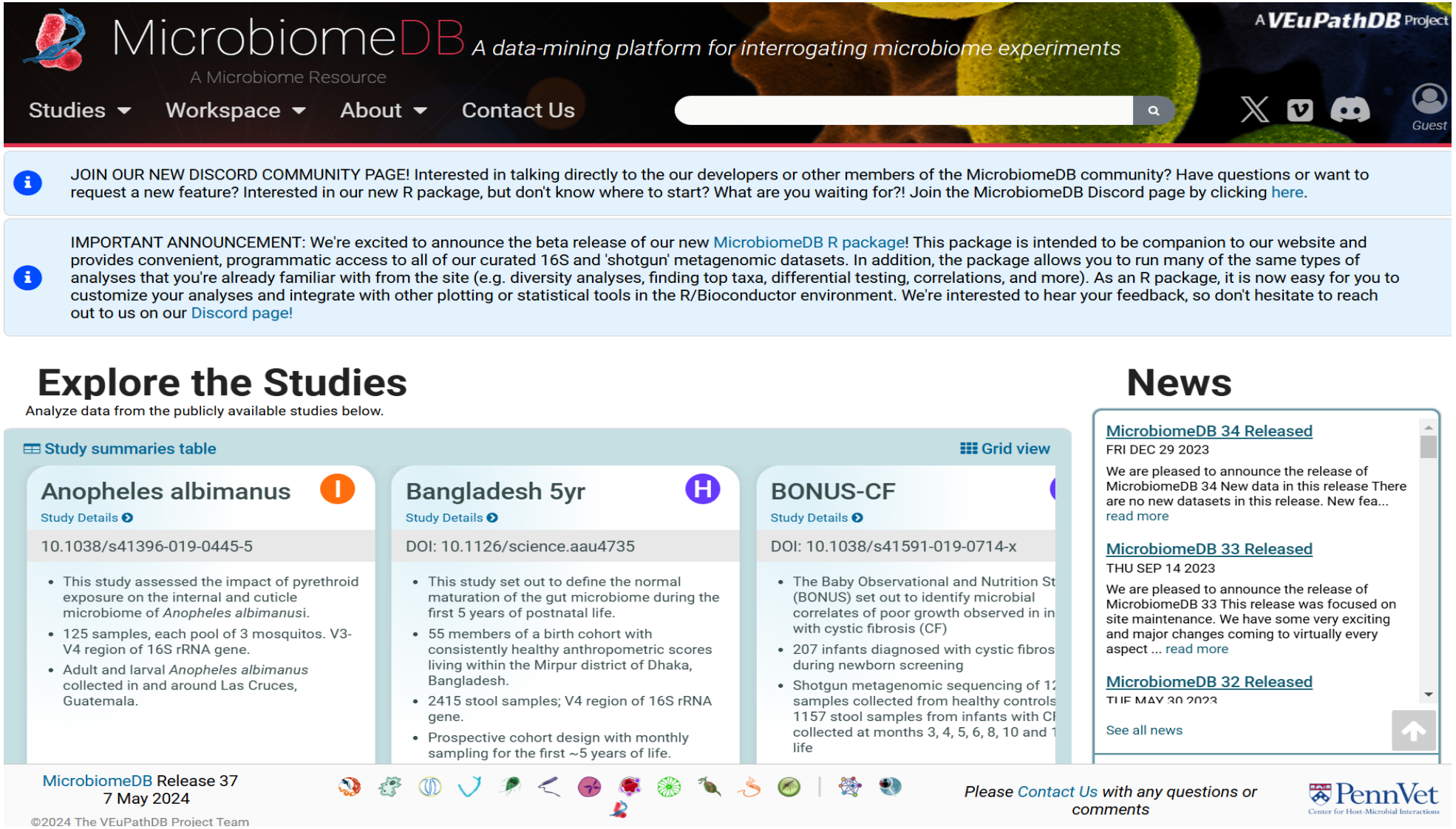
Metabolomic Datasets

https://hmdb.ca/



MicrobiomeDB

https://microbiomedb.org/mbio/app



Immunology Database and Analysis Portal (ImmPort)

https://www.immport.org/home

IMMPORT

BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

ImmPort is funded by the NIH, NIAID and DAIT in support of the NIH mission to share data with the public. Data shared through ImmPort has been provided by NIH-funded programs, other research organizations and individual scientists ensuring these discoveries will be the foundation of future research.

Data uploading or sharing questions? Please contact ImmPort_Helpdesk@immport.org.

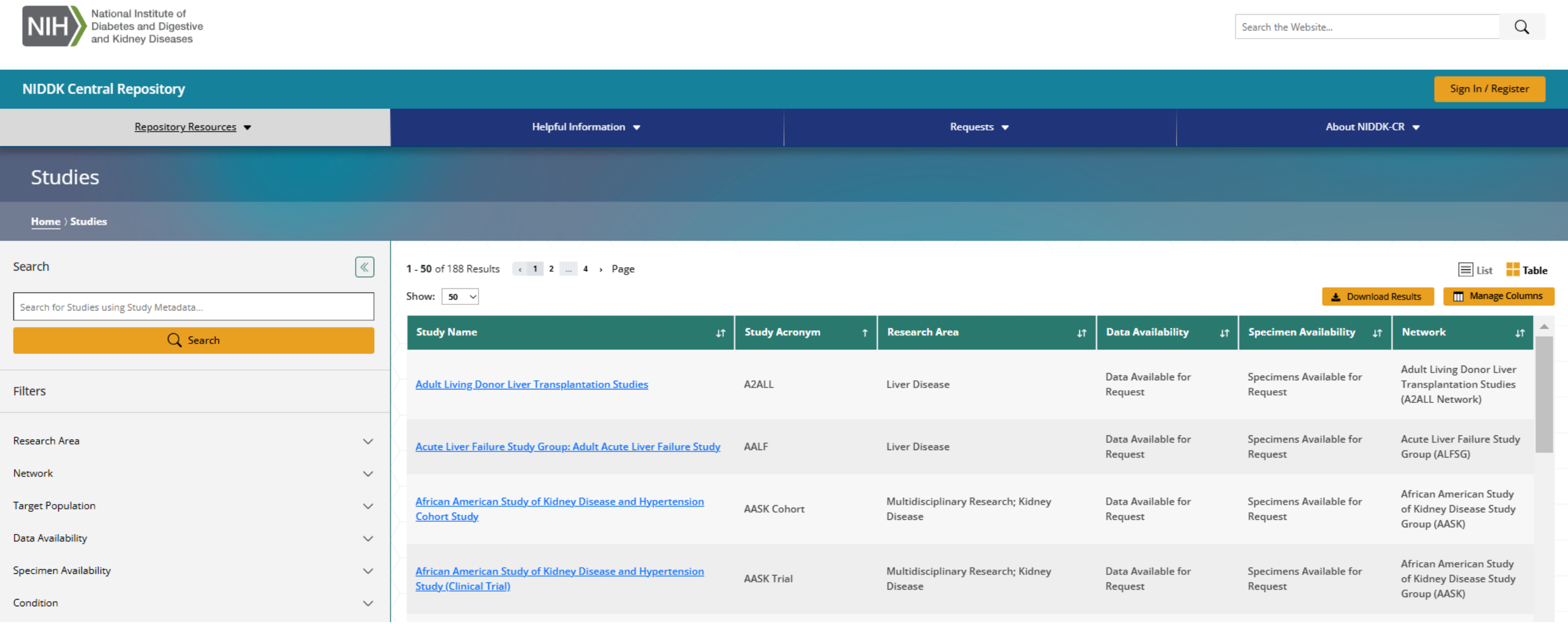
Upload Data

Shared Data

News & Events

10/24/2024 - ImmPort Data Release 53.1 is out! 26 new studies, 5 updated studies. For details please see the Data Release notes.

NIDDK Repository



Institute Specific Repositories

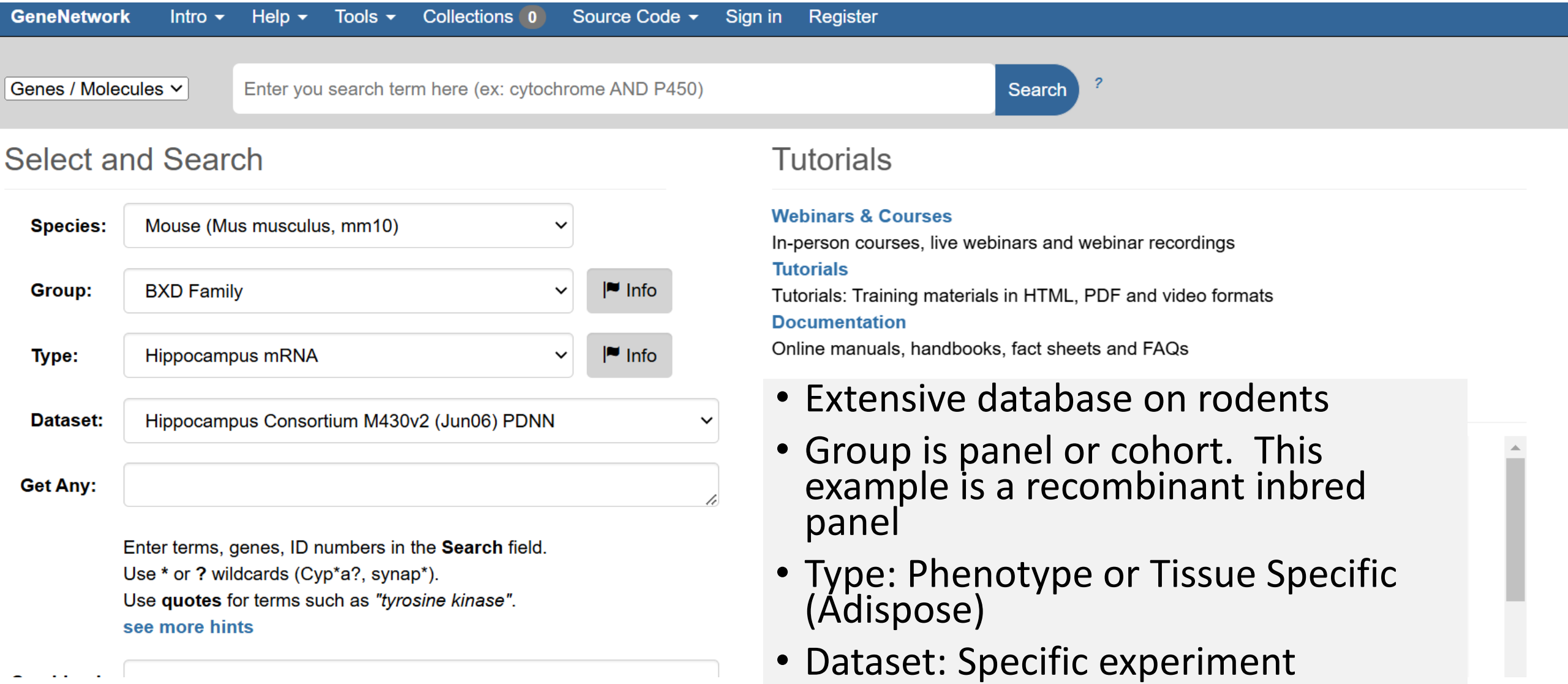
https://sharing.nih.gov/accessing-data/accessing-scientific-data

Institute or Center ICOs filter	Repository Name	Repository Description Keyword Filter	Access to Data	Access Type
ONR	Nutrition Science Data and Biospecimen Resources Portal	The National Institutes of Health (NIH) Office of Nutrition Research is committed to advancing nutrition science research through promoting access to publicly available datasets, biospecimens, and data analysis tools and resources.	NA	controlled registered open
NLM	The database for Genotypes and Phenotypes (dbGaP)	The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.	How to access	controlled open
NLM	ClinVar	ClinVar is a freely accessible, public archive of submitted reports about the relationships among human variations and phenotypes, with supporting evidence.	How to access	open

Non-human Repositories

Gene Network

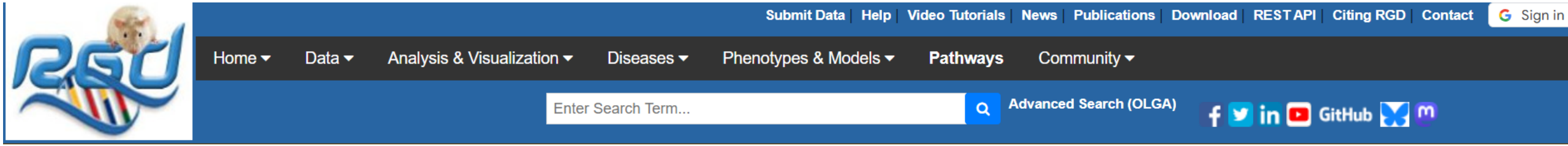
https://genenetwork.org/



- Extensive database on rodents
- Group is panel or cohort. This example is a recombinant inbred panel
- Type: Phenotype or Tissue Specific (Adispose)
- Dataset: Specific experiment

Rat Genome Database

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



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.

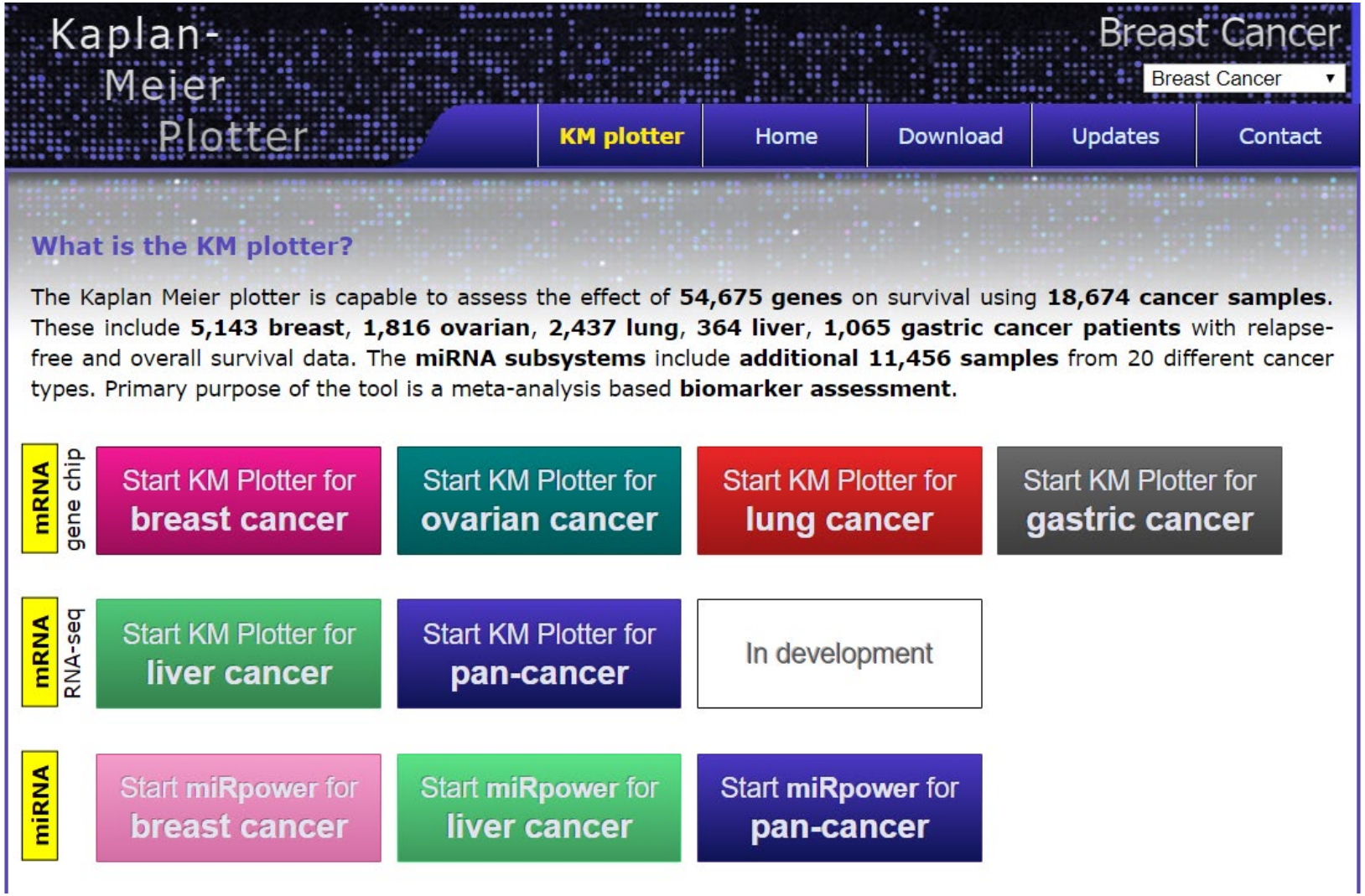
REFERENCES

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.

Validation Tools

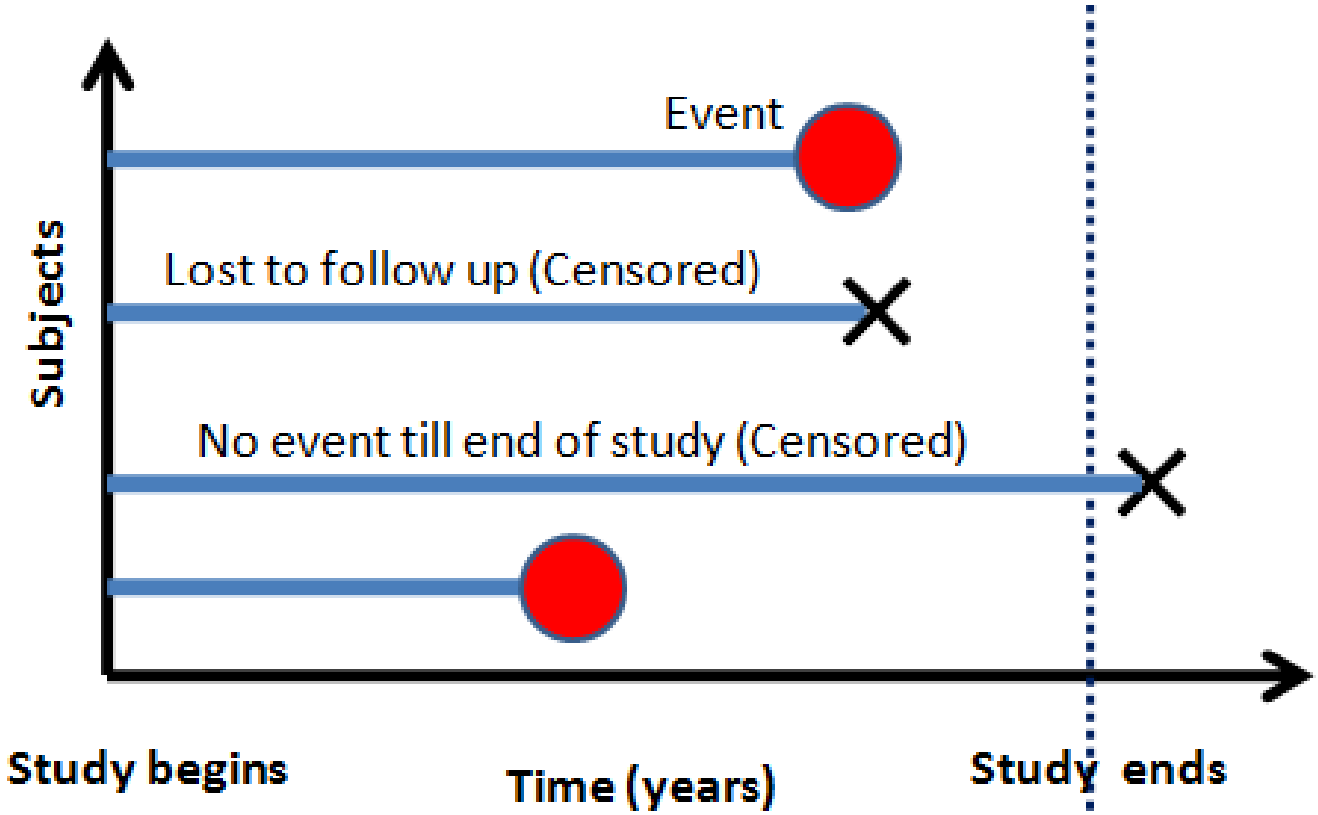
KM Plotter

http://kmplot.com/analysis/



Quick Sidebar - 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

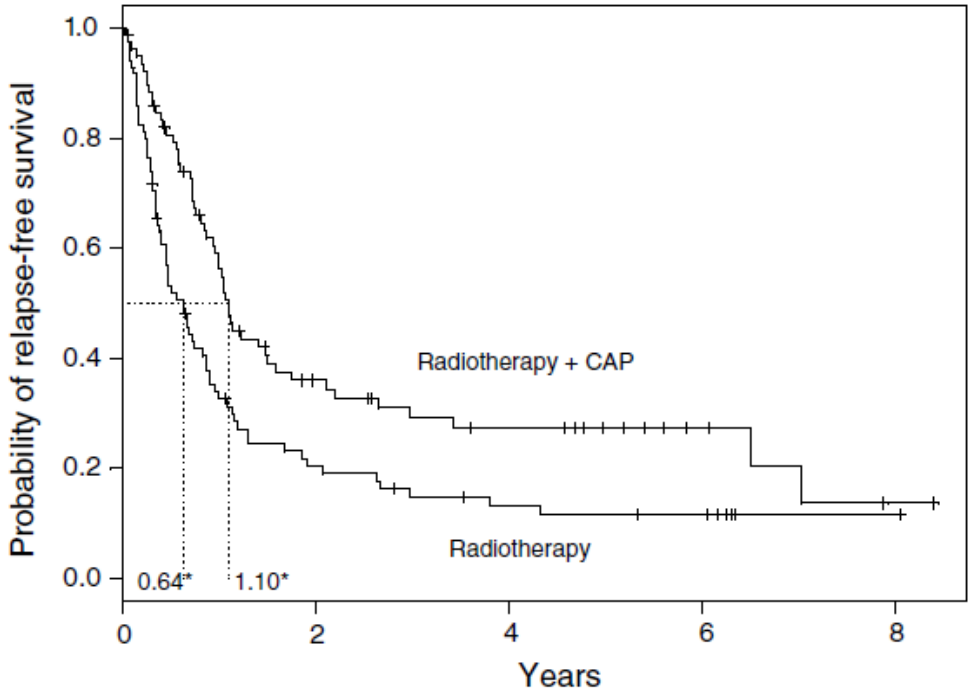


Source: alphabeta statistics

Kaplan-Meier survival curves

- S(t) = probability of surviving (event-free) from the beginning of the study to time t
- R(t) = cumulative risk of having the event from the beginning of the study to time t
- Cumulative risk = 1 – cumulative survival

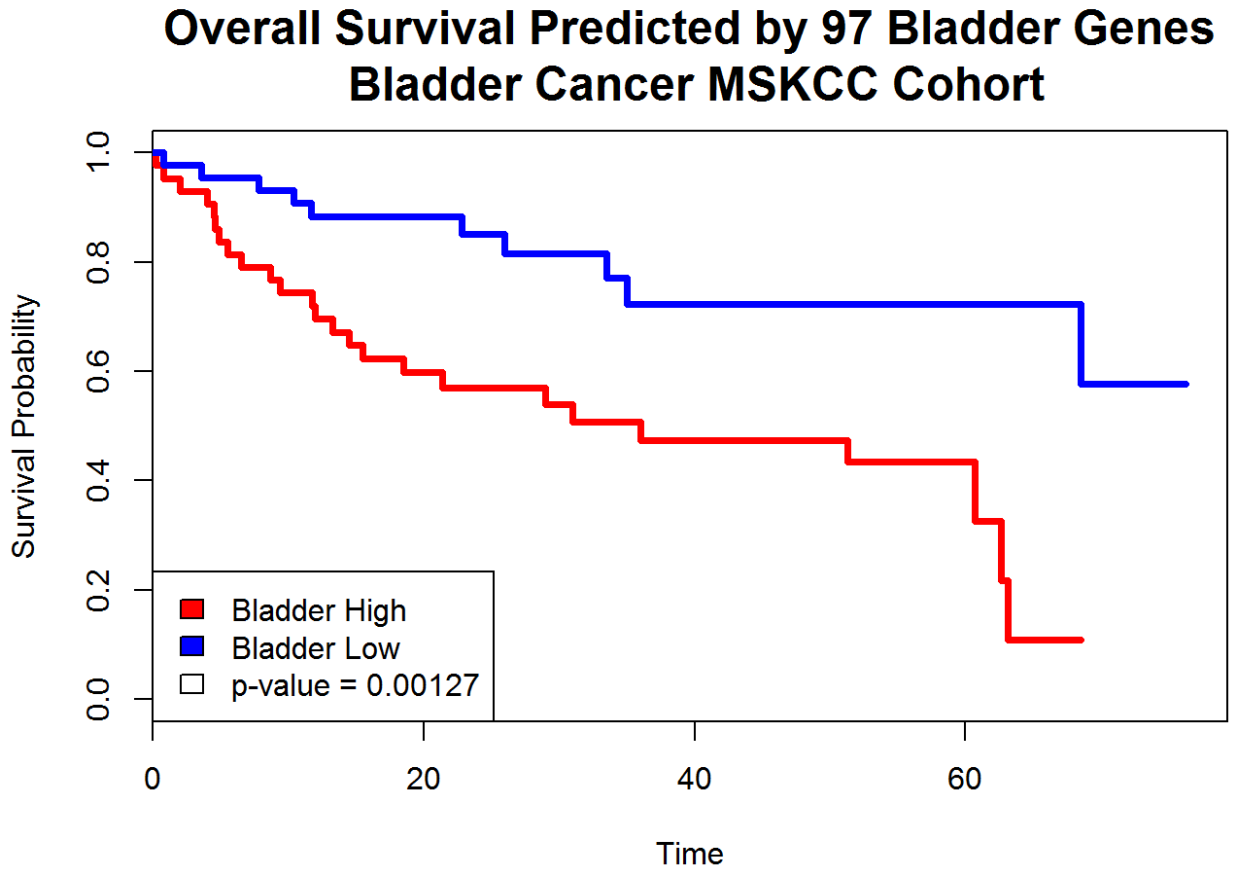
- Kaplan-Meier method:
  - Nonparametric
  - Calculate interval-specific survival probabilities
  - Generate step function (changes value only at event times)
  - Compare groups with log-rank test



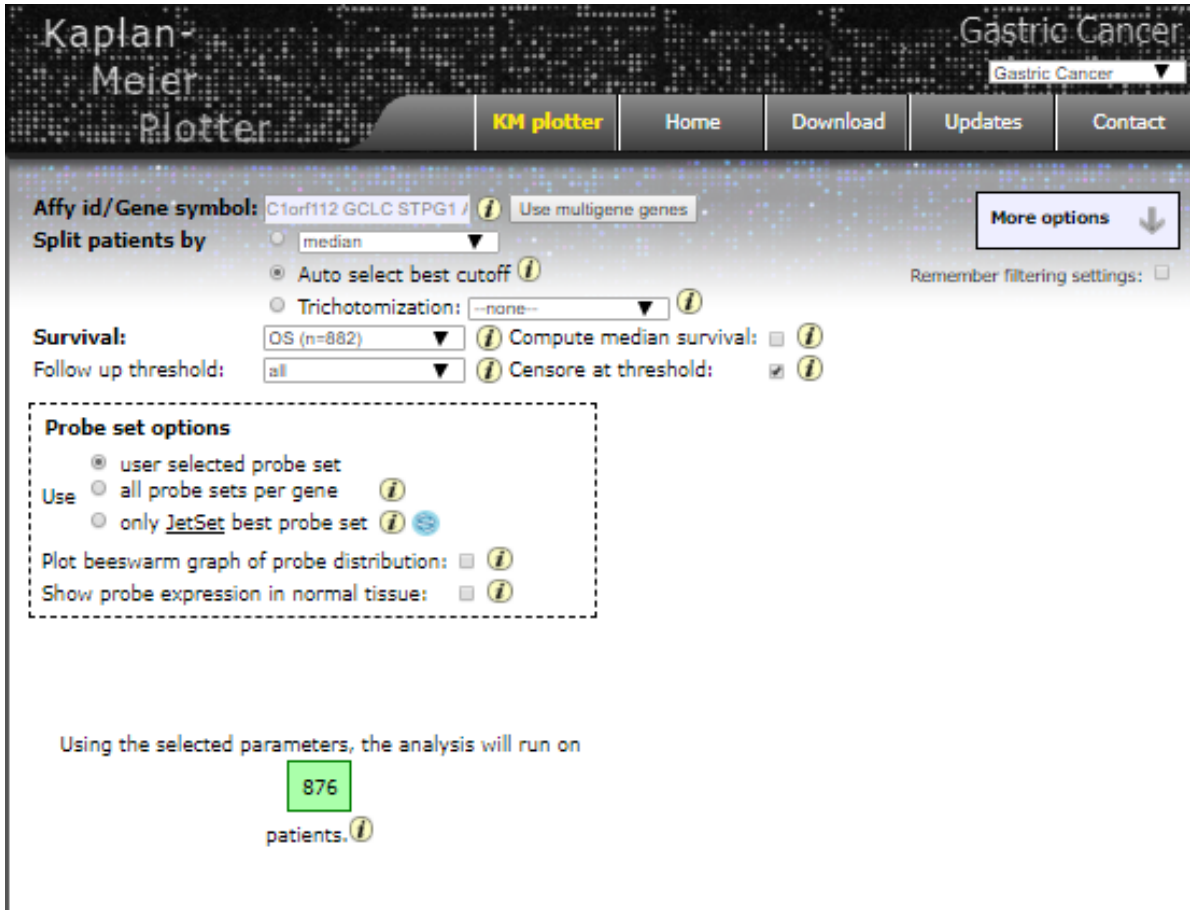
Clark et al (2003) Br J Cancer 89(2):232-238

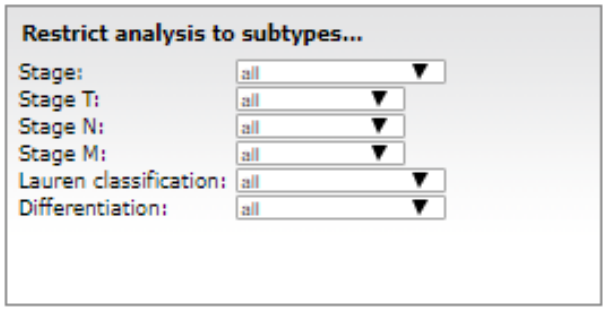
Kaplan-Meier Curves

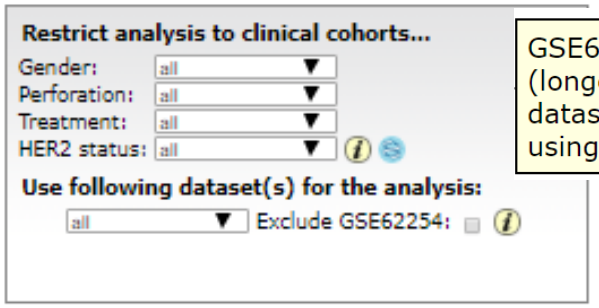
- Visualize separation using a KM curve
- For KM-plotter, can put in single gene, multiple genes (as in a gene signature)
- Not very useful for multivariable analysis, and does not estimate the magnitude of the association.



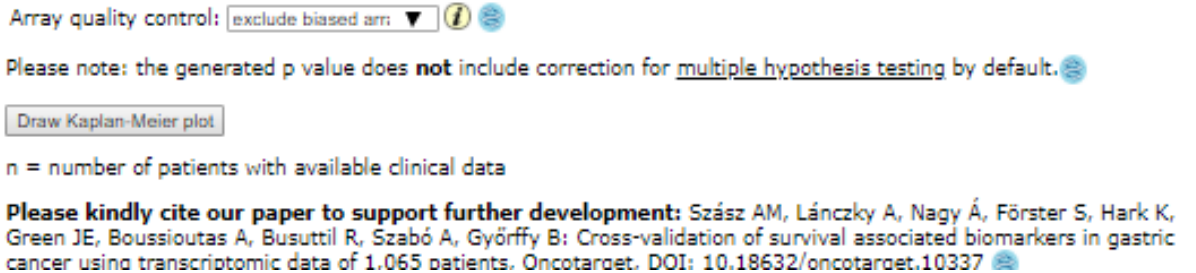
KM Plotter – Selecting Data for Analysis







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 | Downlo

The desired is valid: 205942_s_at (ACSM3),

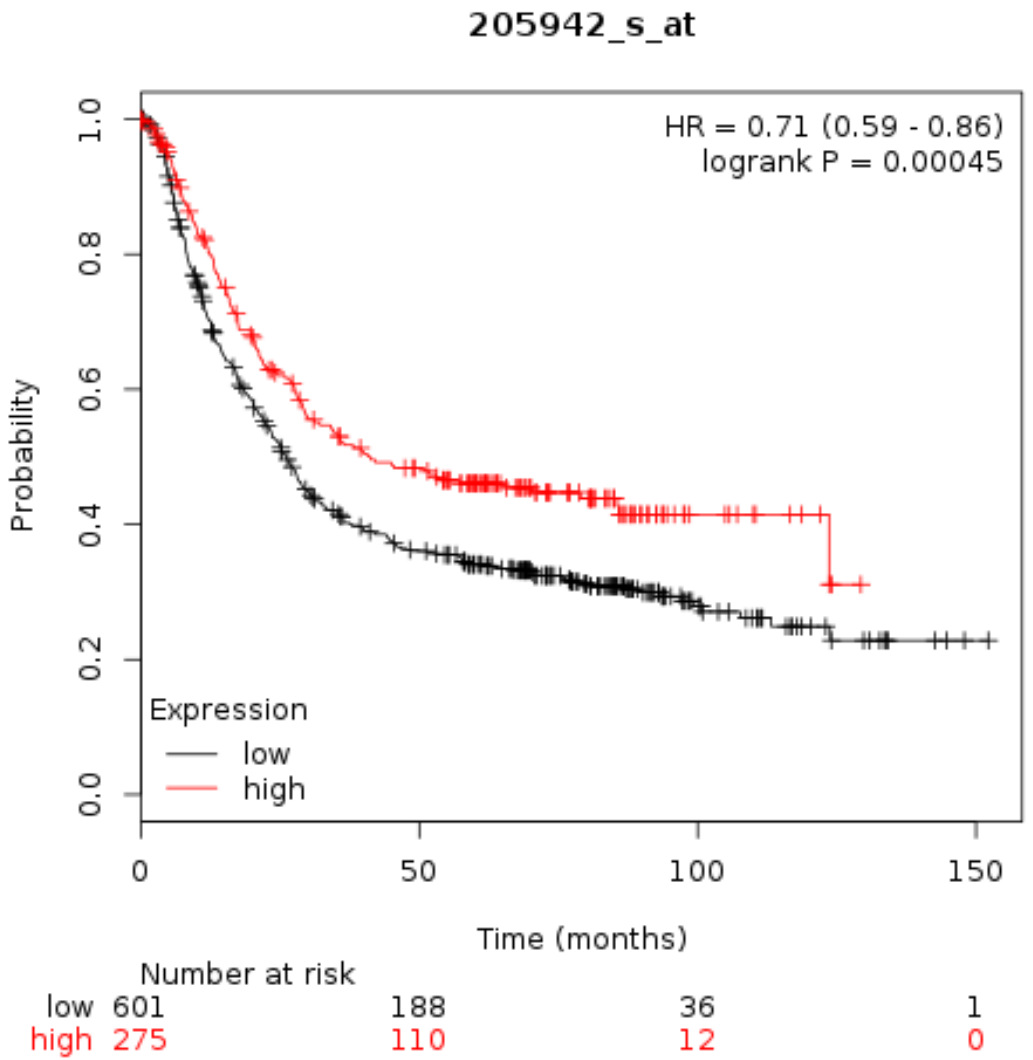
Affy ID:	205942_s_at	ACSM3, SA, SAH
Survival:	OS	
Auto select best cutoff:	checked	
Follow up threshold:	all	
Censore at threshold:	checked	
Compute median over entire database:	false	
Cutoff value used in analysis:	230	
Expression range of the probe:	2 - 2185	
Probe set option:	user selected probe set	
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

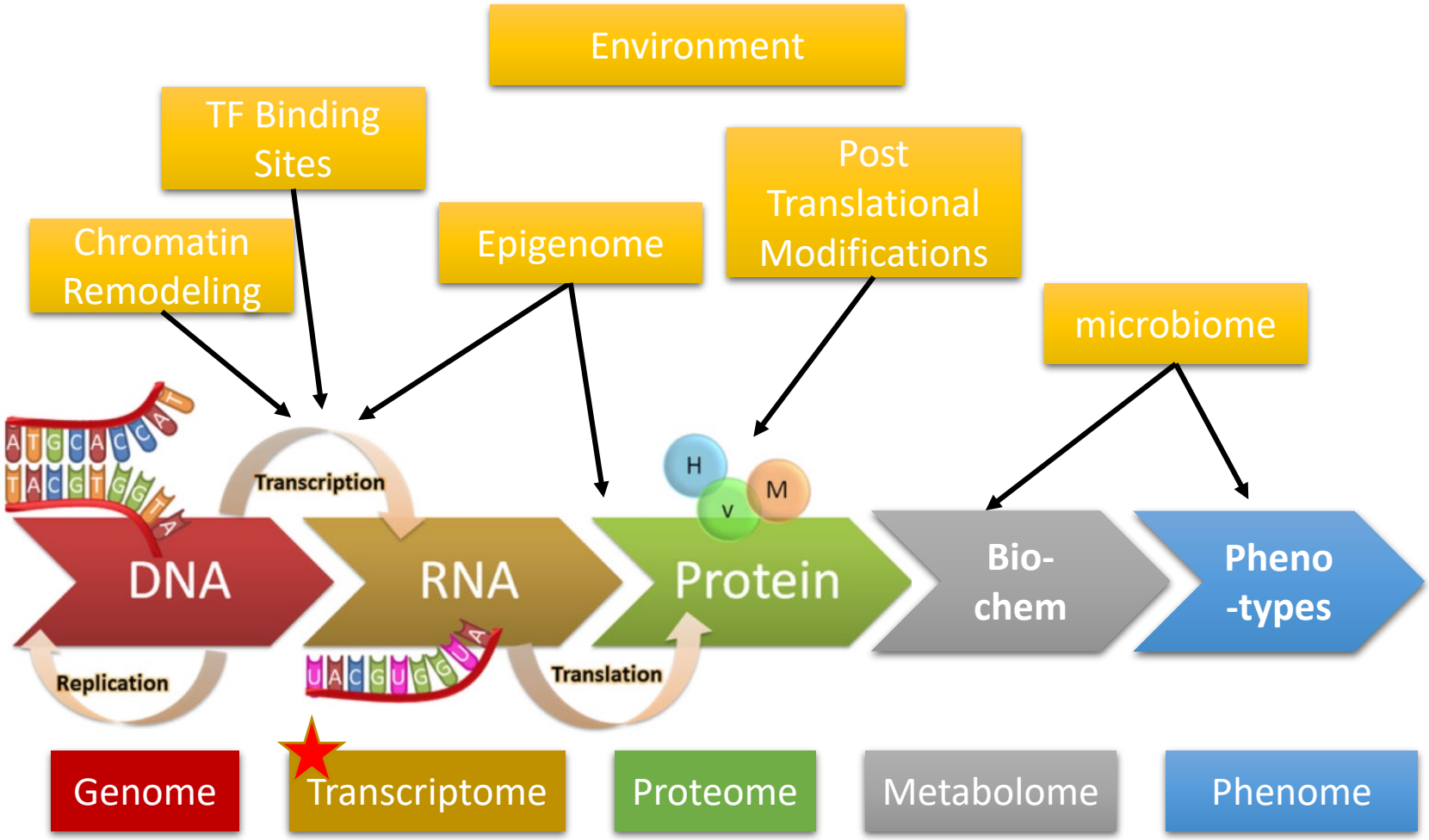
P value: 0.0005
FDR: 10%

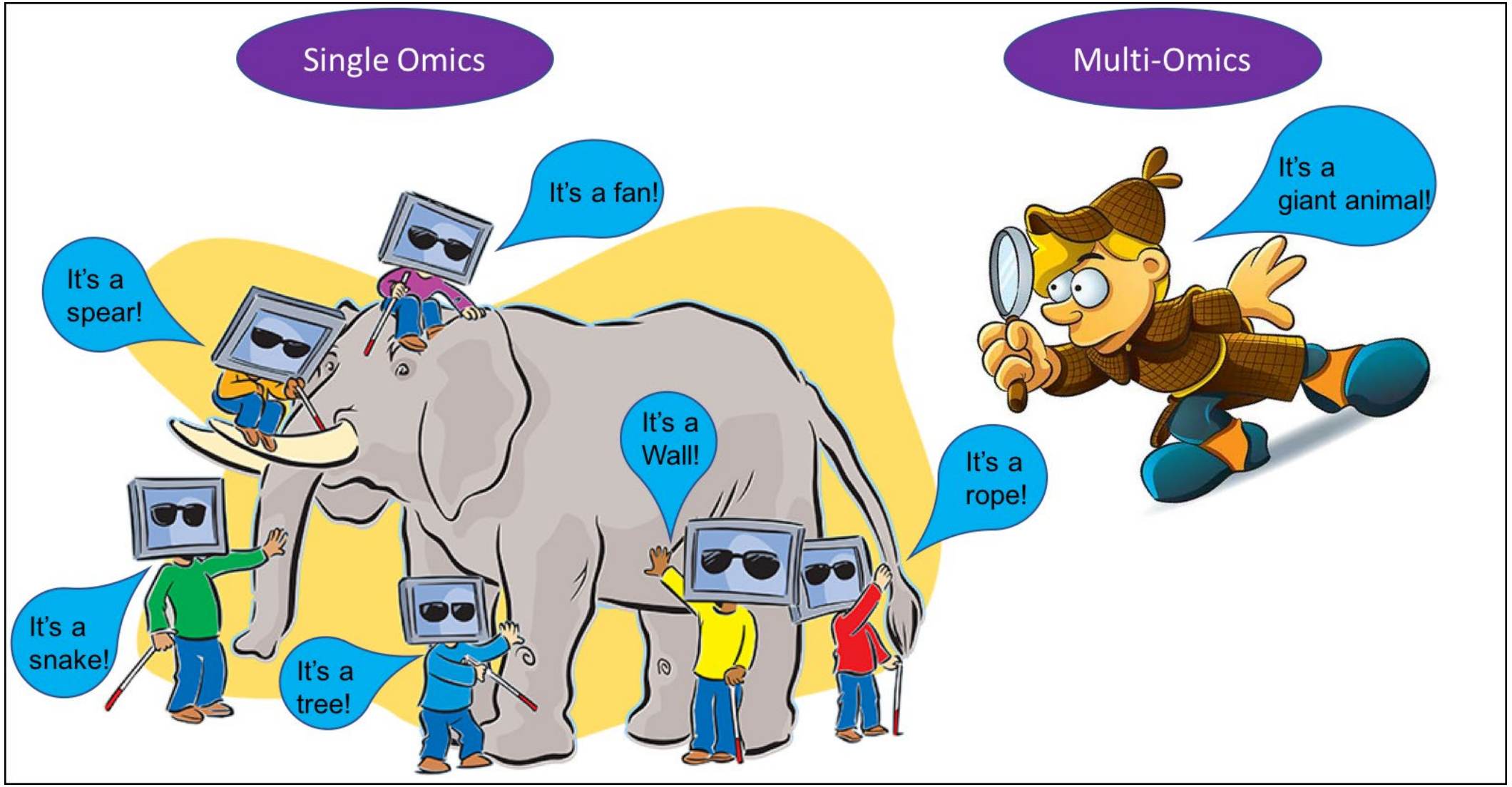


Data Integration

Multiple ‘Omics Datasets

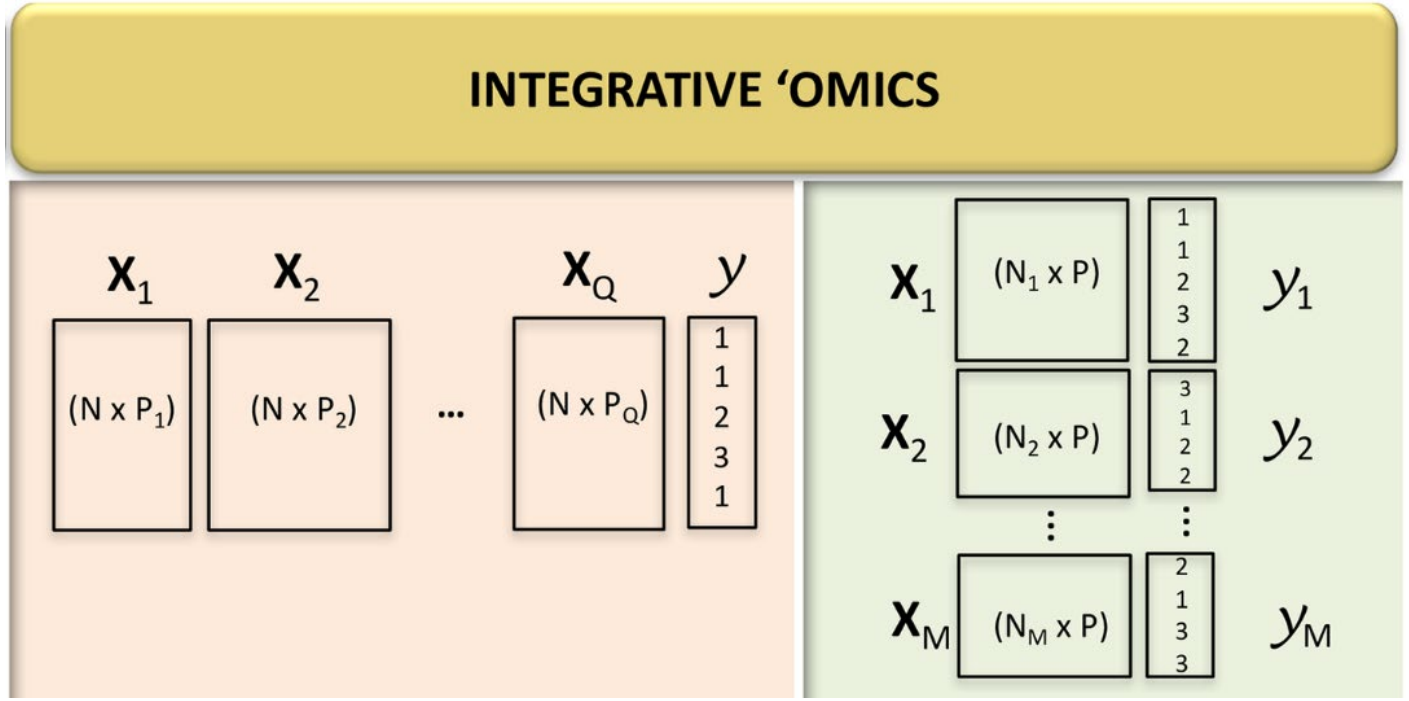
- Want to get a bigger picture on what is happening
- Multiple ‘omics datasets





Source: Melgen

Horizontal vs Vertical Data Integration

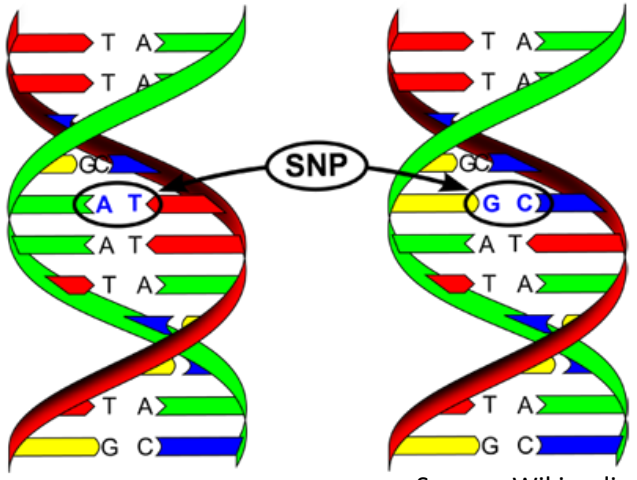


Horizontal
Same technology
Different samples

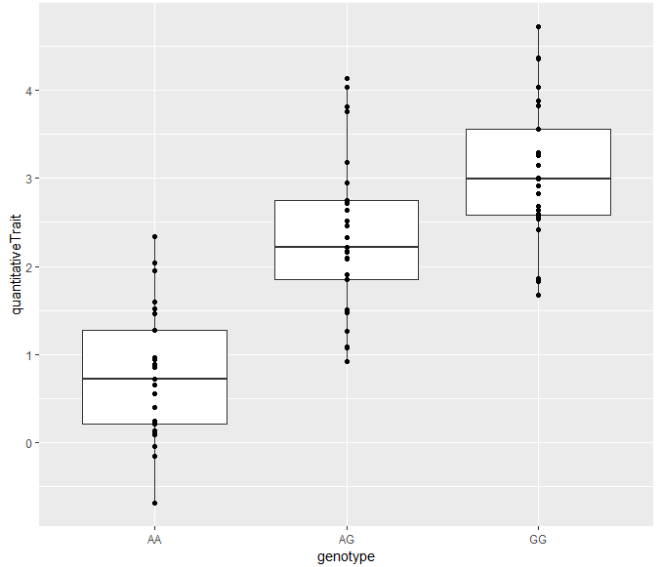
Vertical
Different technology
Same samples

Integration with DNA

- Most common form of data integration
- DNA marker set of Single Nucleotide Polymorphisms (SNPs)
  - 2 copies of chromosomes
  - Homozygous major allele (AA)
  - Heterozygous (AG)
  - Homogzygous minor allele (GG)
- 99.9% genome same among individuals
- Easy to integrate with other ‘omics datasets tied to genome
  - mRNA
  - Proteomics
  - DNA methylation
- Look at physically closest SNP or candidate marker



Source: Wikipedia

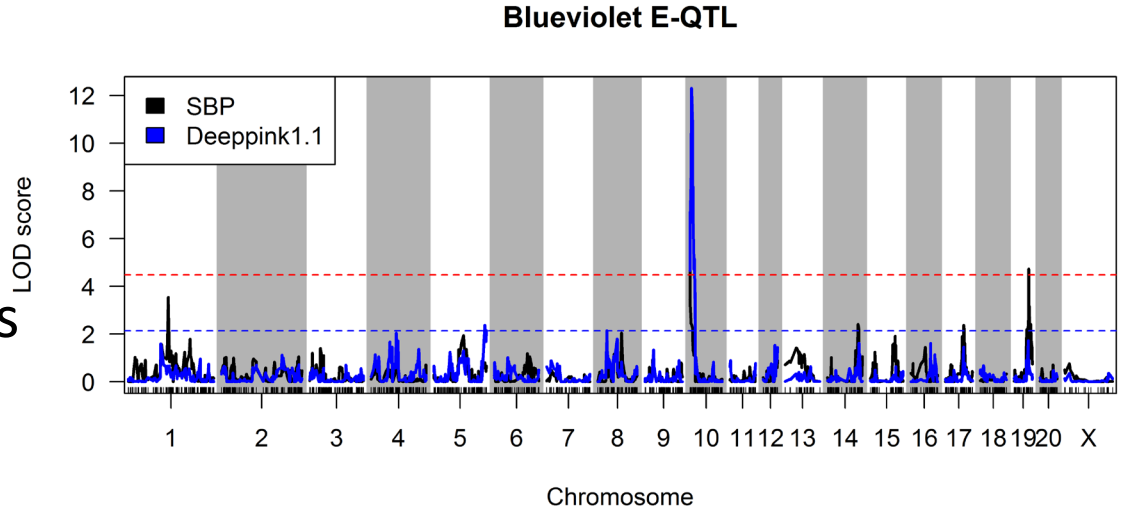


Quantitative Trait Loci (QTL) Mapping

- Attempt to explain the genetic basis of variation in complex traits
- Outcome is a continuous measure
  - Phenotype (pQTL)
  - Gene/Transcript Expression (eQTL)
  - Methylation QTL (mQTL)
  - Metabolite QTL (also can be referred to as mQTL, but metabQTL or metQTL)
- Predictor is SNP marker
- Linear modeling with the base model of outcome = SNP
  - SNP is most often modeled as additive (0, 1 or 2 copies of an allele)
  - Dominant or recessive models sometimes performed if biologically relevant

QTL Continued

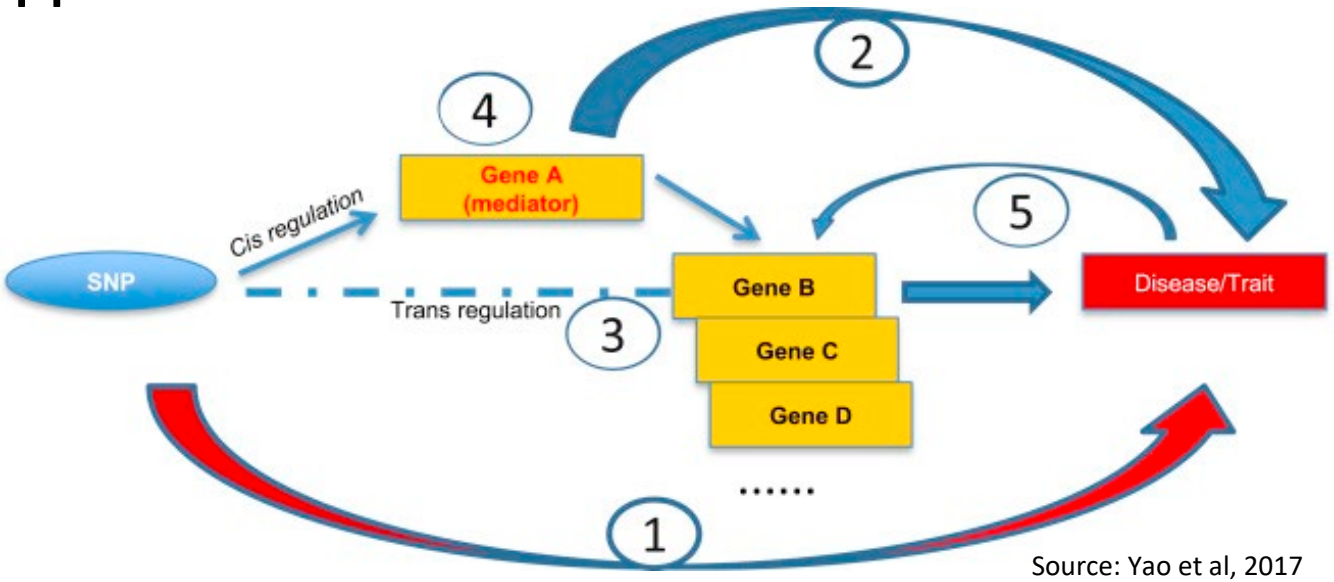
- p-values reported and plotted (Manhattan plot)
- Logarithm of odds “LOD” Score:
  - log10 likelihood ratio comparing hypothesis of a QTL at position λ versus that of no QTL
  - $LOD(\lambda) = log_{10}\left\{\frac{\mathrm{P}(y|QTL\ at\ \lambda)}{P(y|no\ QTL)}\right\}$
  - This is common among animals models
- Estimated at least 30% of gene transcripts are substantially influenced by eQTL (Romanoski et. al, 2010)



Example QTL for both a phenotype (systolic blood pressure, black trace) and a candidate eigengene expression (from WGCNA, blue trace). Red and blue dotted lines show the genome-wide significant and suggestive thresholds for the eQTL.

Cis vs Trans Regulation

- Cis is genetic control from some SNP close to gene
- Trans is genetic control from SNP far away from gene
- You do see these eQTL "hotspots"



Source: Yao et al, 2017

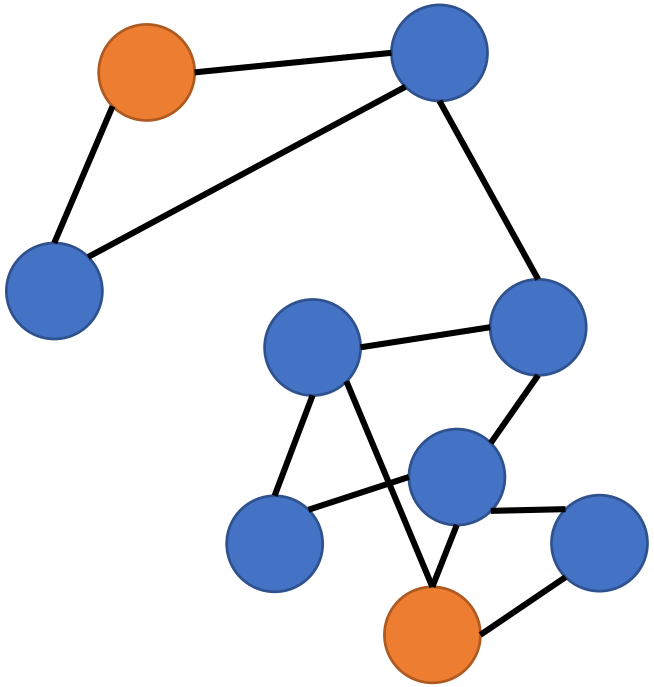
(1) missense SNP affects protein structure/function
(2) non-coding SNP affects gene expression (*cis*)
(3) non-coding SNP affects remote (*trans*) gene expression directly or by
(4) *cis*-eGene mediation of the *trans*-eQTL-*trans*-eGene association; or
(5) reverse causality (trait has feedback effect on gene expression).

Epigenetic Effects & mRNA Expression

- Easy to link because of location
- Have ChIP peak in a range of gene's TSS
- DNA methylation and gene expression
  - Same samples: correlation (expect negative correlation)
  - Different samples: take candidate list of say differential methylated positions and see if there are differential expression in corresponding gene
- miRNA and mRNA
  - Find the targets for miRNA
    - multiMiR (Dr. Katerina Kechris) http://multimir.ucdenver.edu/
  - Look at correlation or candidates (yes/no)

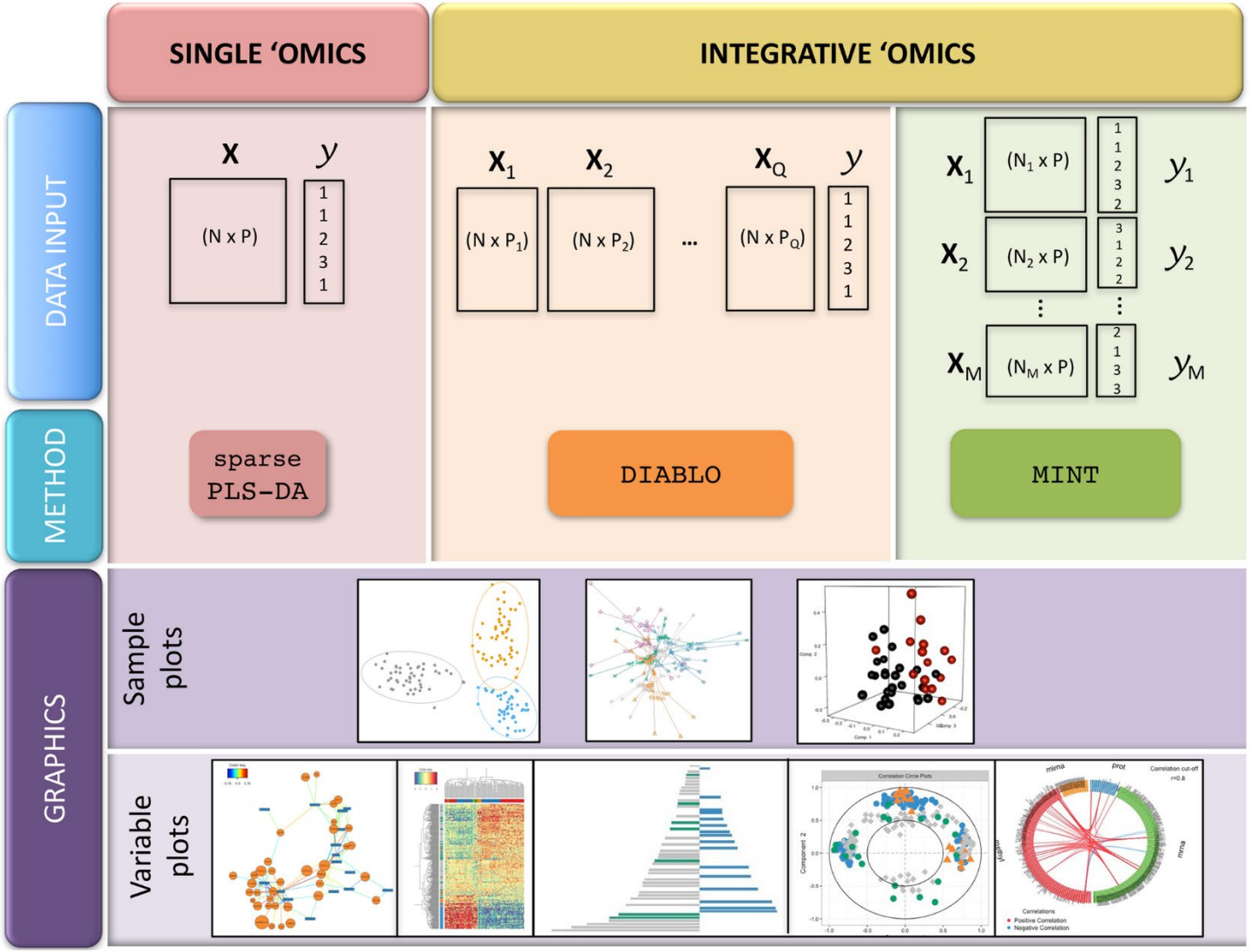
LASSO to get miRNA-mRNA network

- Datasets available:
  - miRNA dataset (~2,000)
  - mRNA dataset (~20,000)
- miRNA can target multiple genes
- Perform WGCNA on the mRNA dataset
- Identify miRNA(s) that regulate module by performing LASSO using eigengene as outcome and miRNAs as predictors
- Need to have same samples in both datasets



R/mixOmics

- Feature Selection
- Data integration
- Supervised analysis
  - Classify or discriminate sample groups
- Sparse Partial Least Squares Discriminant analysis (sPLS-DA)
  - Original 1 dataset approach
- DAIBLO
  - Integration of same biological samples (N) measured on different platforms
  - N-integration
- MINT (Meta Analysis)
  - Integration of independent datasets on measured on same predictors
  - P-integration



Methods Available in mixOmics

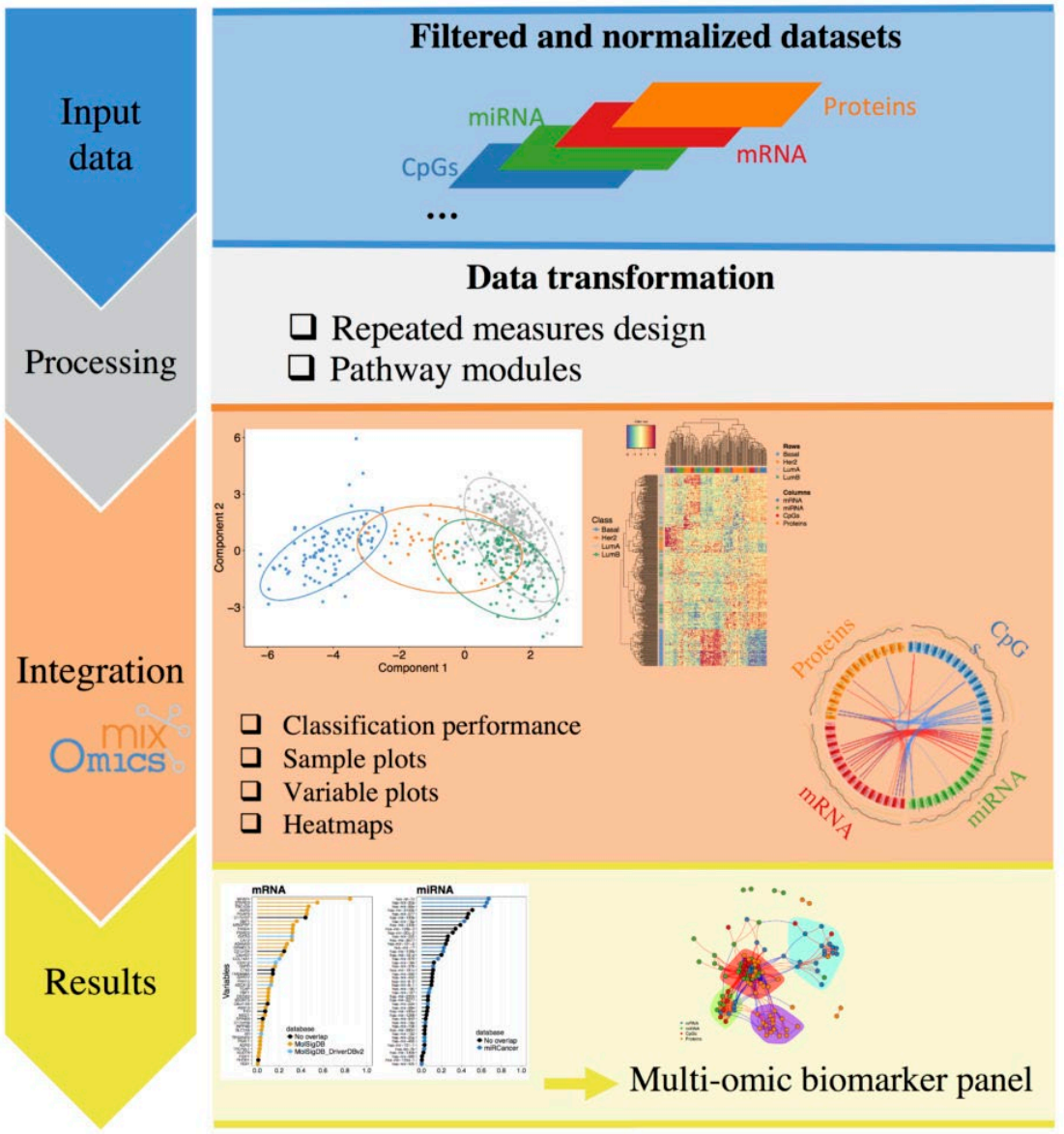
Framework		Sparse	Function name	Predictive model
Single 'omics	unsupervised	-	pca	-
		-	ipca	-
		✓	spca	-
	supervised	-	plsda	✓
		✓	splsda	✓
Two 'omics	unsupervised	-	rcca	-
		-	pls	✓
		✓	spls	✓
N-integration	unsupervised	-	wrapper.rgcca	-
		✓	wrapper.sgcca	-
		-	block.pls	✓
		✓	block.spls	✓
	supervised	-	block.plsda	✓
		✓	block.splsda (DIABLO)	✓
P-integration	unsupervised	-	mint.pls	✓
		✓	mint.spls	✓
	supervised	-	mint.plsda	✓
		✓	mint.splsda	✓

https://doi.org/10.1371/journal.pcbi.1005752.t001

DIABLO

- Data Integration Analysis for Biomarker discovery using Latent variable approaches for 'Omics studies
- Builds on

1. generalized canonical correlation analysis (CCA)
2. Sparse sGCCA method

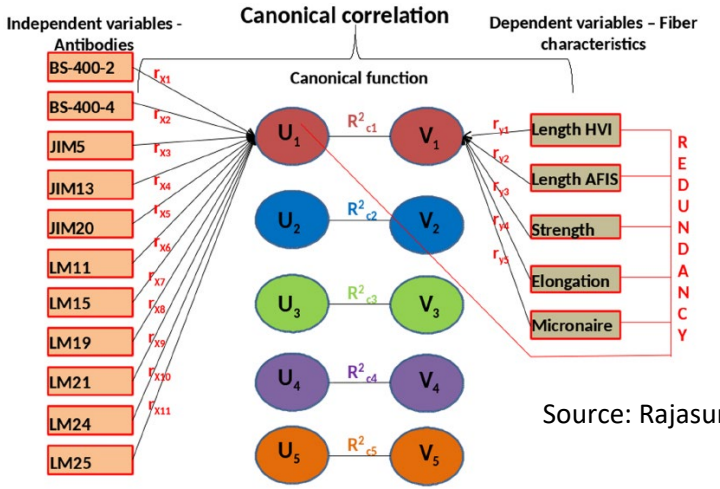


D I A B L O

Canonical Correlation Analysis (CCA)

- Multivariate correlation
- Accounts for multi-collinearity
  - Features in a dataset not independent
  - Know not all genes are independent from each other
- Compare sets of variables to sets of variables
- Canonical loadings
  - Variables relationship to own set
  - E.g. gene A expression to gene B expression
- Canonical weight
  - Variables relationship to other set
  - E.g. gene A expression to miRNA X expression
- Canonical Cross-loadings
  - Variables relationship to other set
  - E.g. gene A expression to whole miRNA set CONSIDERING all of gene A's other interaction with other genes in it's own set

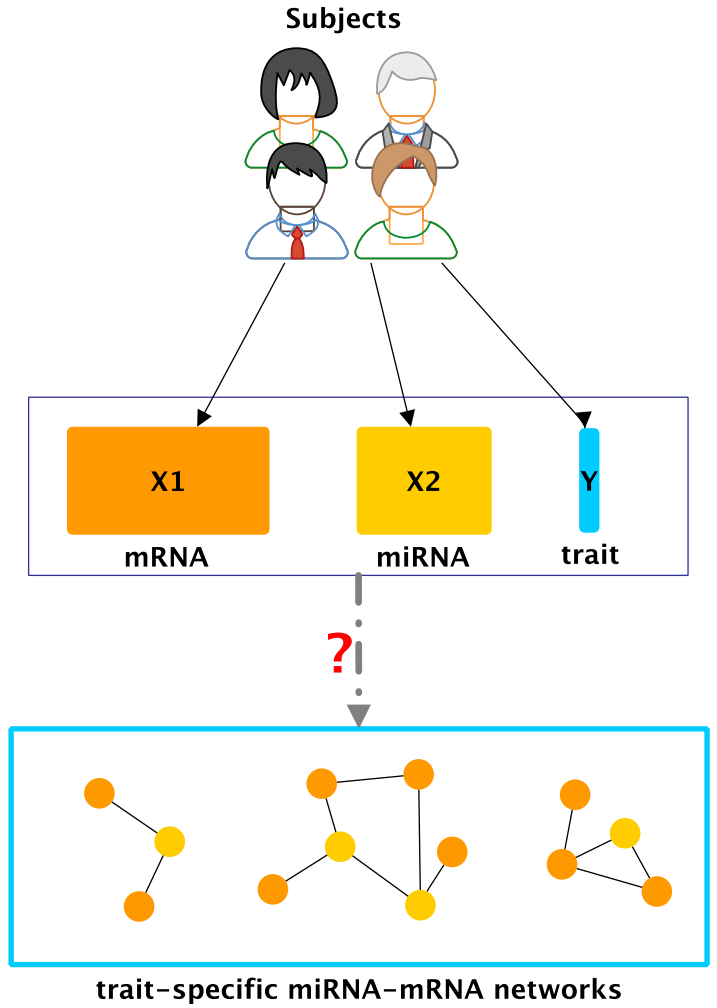
- Canonical loadings and cross-loadings are called structure coefficients
- Canonical weight is a function coefficient
- Canonical correlation is the correlation between sets
- Redundancy coefficients
  - Shared loading variance (variance explained within set)
  - Shared cross-loading variance (variance explained between sets)



Source: Rajasundaram et. Al, 2014

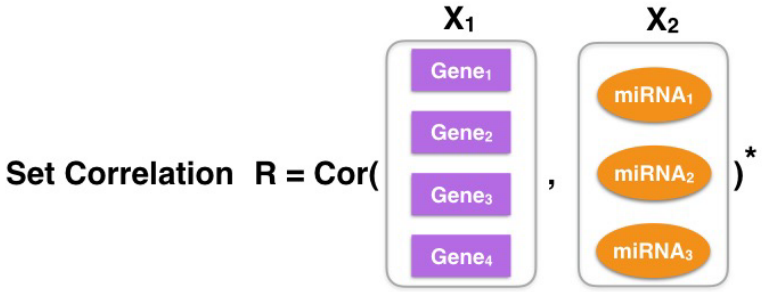
Sparse Multiple Canonical Correlation Network Analysis (SmCCNet)

- R/SmCCNet
  - Katerina Kechris & Jenny Shi
- CCA: Relationship between 2 multivariate datasets measured on same samples
  - E.g. Gene A to Gene B within mRNA dataset
- Multiple: multiple 'omics datasets
  - miRNA and mRNA
- Sparse: not expecting many connections



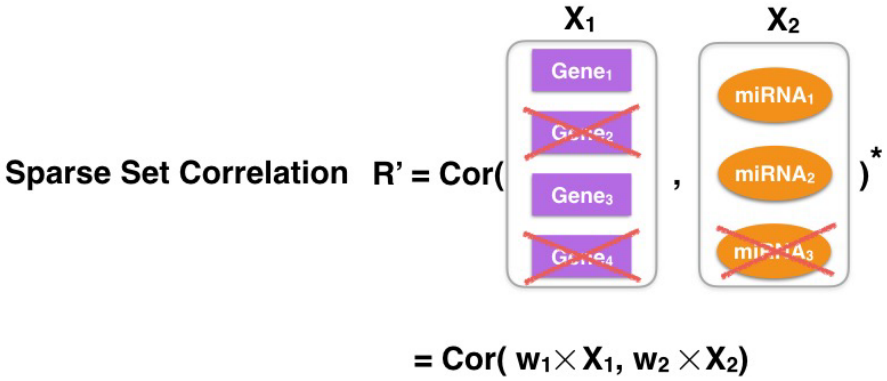
Source: Katerina Kechris

CCA vs Sparse CCA



where w_1 & w_2 are 4×1 and 3×1 unit vectors respectively.

Sample canonical weights:
$w_1 = (0.10, -0.48, 0.83, 0.22)^t$,
$w_2 = (0.67, 0.14, 0.73)^t$.

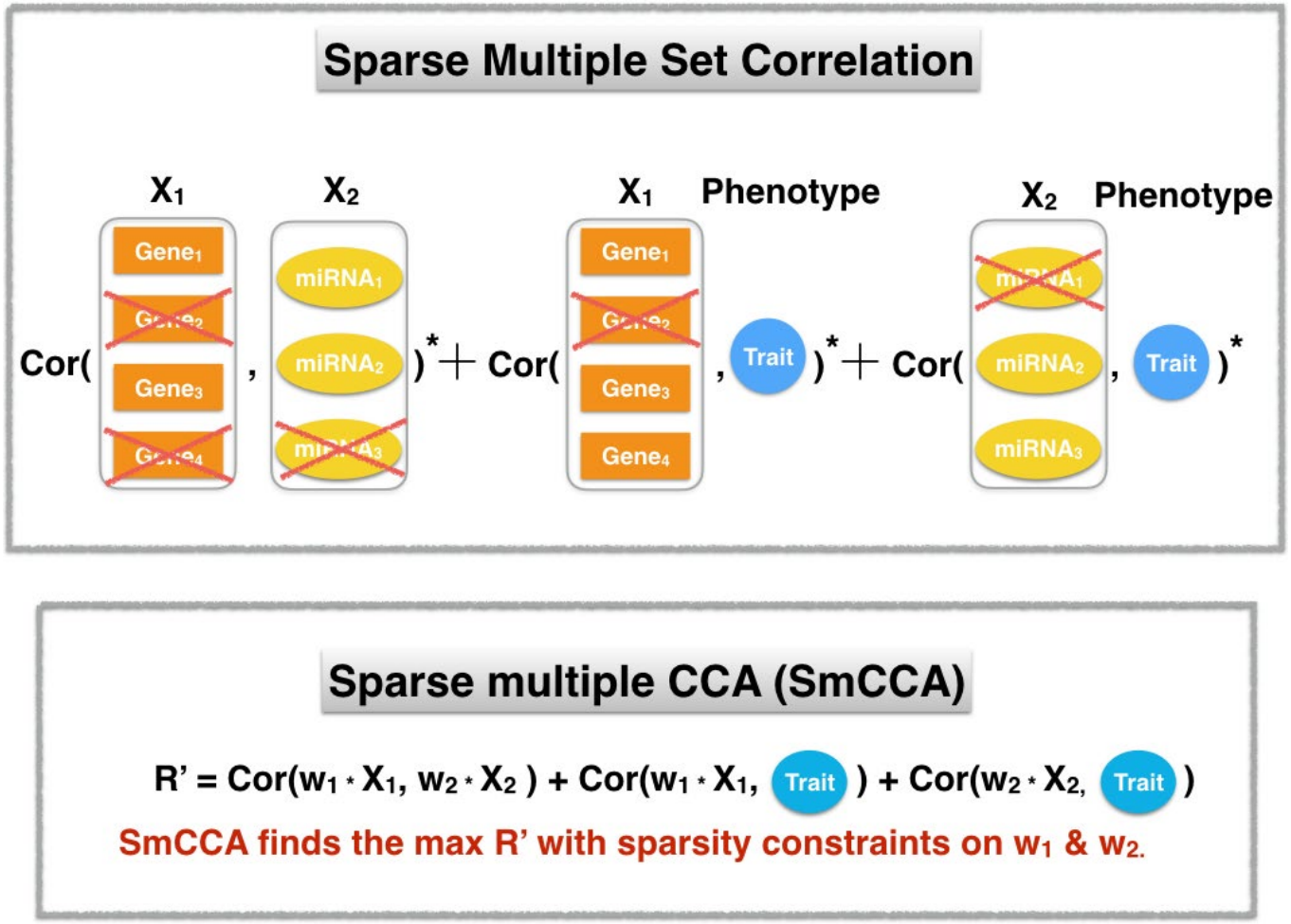


where w_1 & w_2 are 4×1 and 3×1 unit vectors, satisfying some constraints $p_1(w_1) < c_1$ and $p_2(w_2) < c_2$, respectively.

Sample canonical weights:
$w_1 = (0.17, 0, 0.37, 0)^t$,
$w_2 = (0.25, 0.28, 0)^t$.

Source: Katerina Kechris

SmCCA

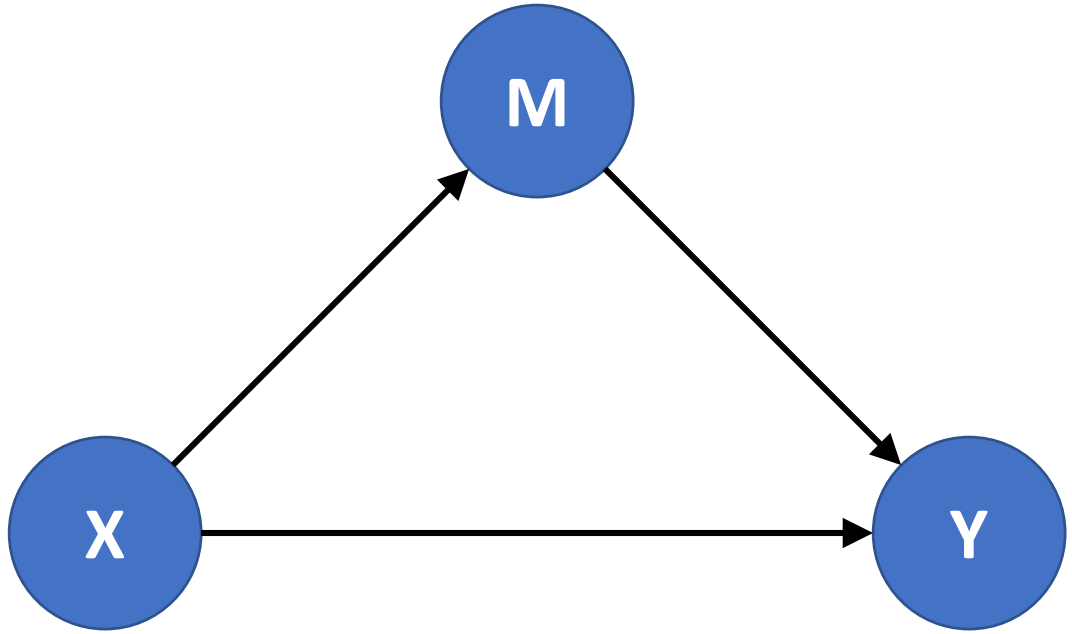


Subsampling of miRNA/mRNA & cross-validation of samples for sparse penalties on weights

Source: Katerina Kechris

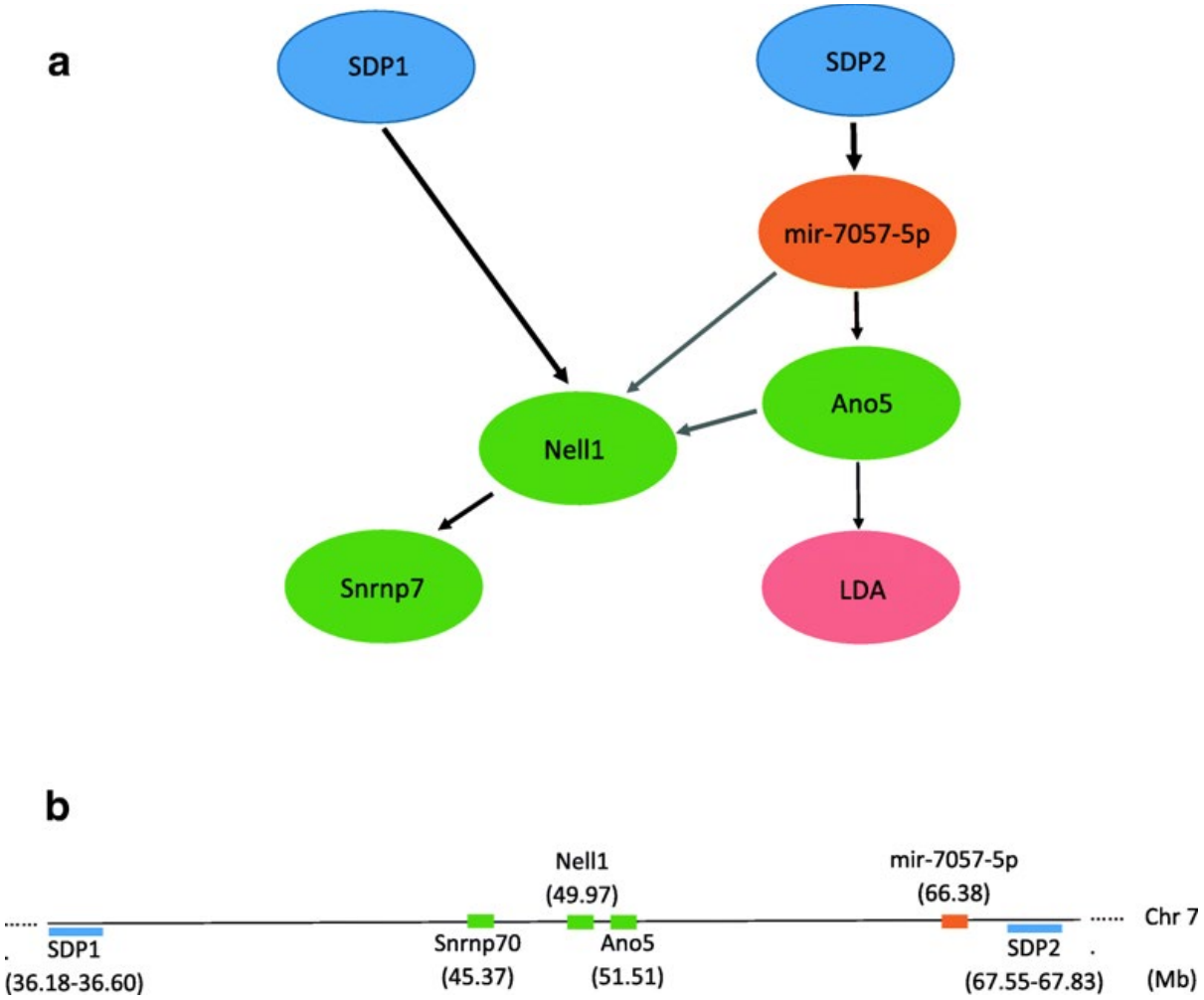
Mediation Analyses

- Observe a relationship between independent (X) and dependent variables (Y)
- Mediator (M) is in the causal pathway of X -> Y
- Complete & partial mediation
- We have seen something like this!
- Build on this an get Directed Acyclic Graph (DAG)



Mediation Example

- Bayesian networks adds the arrows between nodes of same dataset
  - Ideal for taking an event that occurred and predicting the likelihood that any one of several possible known causes was the contributing factor
  - Computationally intensive
- Once again, start small and build up



Source: Rudra et. al, 2018

Integration with Machine Learning

MetaXcan https://github.com/hakyimlab/MetaXcan

MetaXcan is a set of tools to integrate genomic information of biological mechanisms with complex traits. Almost all of the software here is command-line based.

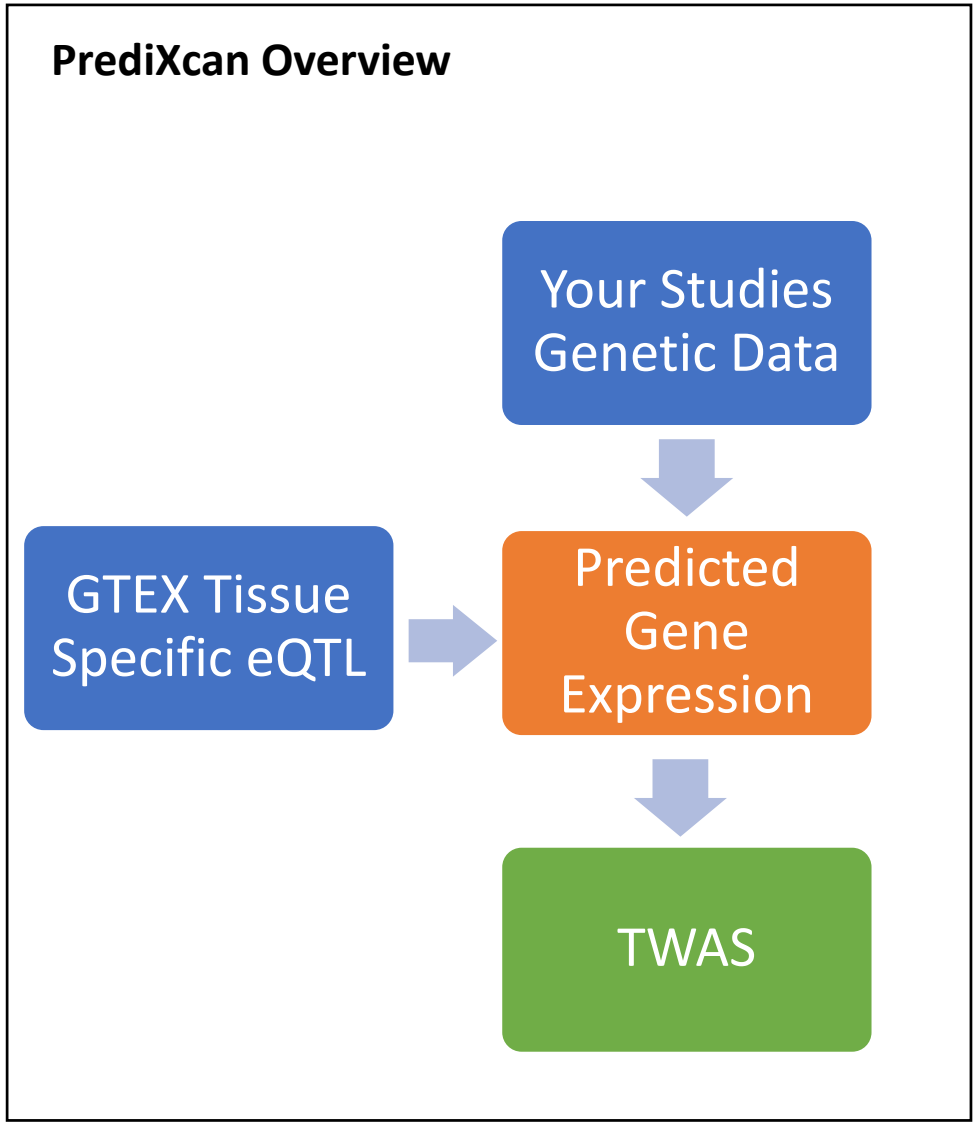
This software has been recently migrated to **python 3** as **python 2** has been sunset.

Tools

Here you can find the latest implementation of PrediXcan: **PrediXcan.py**. This uses individual-level genotype and phenotype, along a mechanism's prediction model (e.g. models predicting expression or splicing quantification), to compute associations between omic features and a complex trait.

S-PrediXcan is an extension that infers PrediXcan's results using only summary statistics, implemented in **SPrediXcan.py**. A manuscript describing S-PrediXcan and the MetaXcan framework with an application can be found S-PrediXcan.

MultiXcan (**MulTiXcan.py**) and S-MultiXcan (**SMulTiXcan.py**) compute omic associations, integrating measurements across tissues while factoring correlation. For example, if you have prediction models, each trained on different



Other Integration Software

- R/Omic
- R/integrOmics
- R/ STATegRasPLS
- R/OMICsPCA
- R/MultiAssayExperiment
- R/iCluster
- R/CNAmet
- R/OmicKriging
- matlab/JIVE
- JAVA/OmicsAnalyzer
- JAVA/VANTED
- JAVA/Lemon-Tree
- C++/DASS-GUI
- C++/GeneTrail2
- Perl/3Omics
- Perl and Python/PaintOmics

Conclusions

What did we learn?

- Many public repositories where we can download data
- Lots of them have R/packages and APIs to easily download data
- Genetic data is generally available only as summary statistics
- ‘Omics data integration is either horizontal or vertical
- Multiple ways to integrate data types