



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



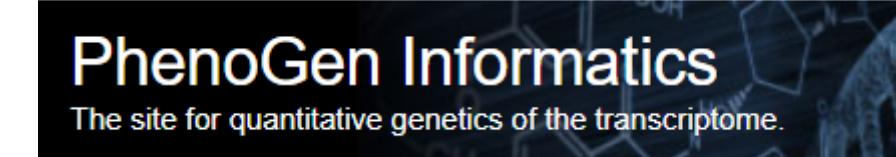
- Validate your results
 - Data generated under same/similar conditions
 - How consistent are the results?
- Statistical method development
 - Real world data
- Compare to a different experimental design
 - Tissue differences

Students: always have 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®.



Human Repositories

GEO Database

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

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

NCBI Resources How To

GEO DataSets GEO DataSets ER positive breast cancer Create alert Advanced Help

Entry type Summary 20 per page Sort by Default order Send to: Filters: Manage Filters

Organism Customise ...

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

Author Customise ...

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

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

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

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

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

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

Find related data Database: Select Find items

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

Recent activity Turn Off Clear
ER positive breast cancer (11745) GEO DataSets

GEO Datasets – Selected Title

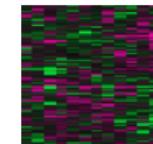
- GSM:
Sample
level
- GDS:
Dataset
- GSE: Series
- GPL:
Platform

NCBI DATASET CURATED BROWSER GEO 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:	<i>Homo sapiens</i>		
Platform:	GPL10558: Illumina HumanHT-12 V4.0 expression beadchip		
Citation:	Aakula A, Leivonen SK, Hintsanen P, Aittokallio T et al. MicroRNA-135b regulates ER α , AR and HIF1AN and affects breast and prostate cancer cell growth. <i>Mol Oncol</i> 2015 Aug;9(7):1287-300. PMID: 25907805		
Reference Series:	GSE57820	Sample count:	12
Value type:	transformed count	Series published:	2015/04/21

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

The screenshot shows the NCBI GEO Accession Display page for dataset GSE57820. At the top, there's a search bar with 'Scope: Self', 'Format: HTML', 'Amount: Quick', and 'GEO accession: GSE57820'. Below the search bar, the dataset title is 'Series GSE57820' (circled in red). The page displays various metadata fields: 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 (A detailed paragraph about miRNAs regulating cellular pathways and their effect on cancer), Overall design (LNCaP cells transfected with Ambion pre-miR™ construct for miR-135b or pre-miR negative control #1 at 20 nM for 12h, 24h, or 36h), and Contributor(s) (Aakula A, Leivonen S, Hintsanen P, Aittokallio T, Ceder Y, Børresen-Dale A.). The right side of the page shows a 'Query DataSets for GSE57820' section with a button to 'Analyze with GEO2R'.

Citation(s)	Aakula A, Leivonen SK, Hintsanen P, Aittokallio T et al. MicroRNA-135b regulates ER α , AR and HIF1AN and affects breast and prostate cancer cell growth. <i>Mol Oncol</i> 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
	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		
MINIMI formatted family file(s)	MINIMI		
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
<i>Raw data is available on Series record Processed data included within Sample table</i>			

SOFT Dataset

GSE57820_family.soft

```
107 #Probe_Coordinates = genomic position of the probe on the NCBI genome build 36 version 3
108 #Cytoband =
109 #Definition = Gene description from the source
110 #Ontology_Component = Cellular component annotations from Gene Ontology project
111 #Ontology_Process = Biological process annotations from Gene Ontology project
112 #Ontology_Function = Molecular function annotations from Gene Ontology project
113 #Synonyms = Gene symbol synonyms from Refseq
114 #Obsolete_Probe_Id = Identifier of probe id before bgx time
115 #GB_ACC =
116 !Platform_table_begin
117 ID Species Source Search_Key Transcript ILMN_Gene Source_Reference_ID RefSeq_ID Unigene_ID
118 Probe_Id Array_Address_Id Probe_Type Probe_Start SEQUENCE Chromosome Probe_Chr_Orientati
119 Ontology_Component Ontology_Process Ontology_Function Synonyms Obsolete_Probe_Id GB_ACC
120 ILMN_1343048 GAATAAAAGAACAAATGCTGATGATCCCTCCGTGGATCTGATTCTGTAA phage_lambda_genome 5090180
121 ILMN_1343049 CCATGTGATACGAGGGCGCGTAGTTGCATTATCGTTTATCGTTCAA phage_lambda_genome 6510136
122 ILMN_1343050 CCGACAGATGTATGTAAGGCCAACGTGCTCAAATCTTCATACAGAAAAGAT phage_lambda_genome:low 7560739
123 ILMN_1343052 TCTGTCACTGTCAAGGAAAGTGGTAAAAGTGCACACTCAATTACTGCAATGC phage_lambda_genome:low 1450438
124 ILMN_1343059 CTTGTGCGCTGAGCTGTCAAAAGTAGAGCACGTCGCCAGATGAAGGGCGC thrB thrB 1240647
125 ILMN_1343061 AATTAAAACGATGCACTCAGGGTTAGCGCGTAGACGTATTGCATTATGC phage_lambda_genome:mm2 2900397
126 ILMN_1343062 GAAGGCAATTGAGGCAAATGAGGCAGCGTTGGTGTAGCACGATAATAATAT phage_lambda_genome:mm2 240255
127 ILMN_1343063 CGGACGTTATGATTACCGTGGAAAGATTGTGAAGTGTTCTGAATGCTC phage_lambda_genome:mm2 2120427
128 ILMN_1343064 ILMN_1343064 GCCCCGTATTCACTGTTGGCTGATTGTATTGTCAGAAGTTGGTTTACGT phage_lambda_genome:mm2 3180440
129 ILMN_1343291 Homo sapiens RefSeq NM_001402.4 ILMN_5311 EEF1A1 NM_001402.5 NM_001402.5
130 ILMN_1343291 3450719 S 1294 TGTGTTGAGAGCTCTCGACAGTATCCACCTTGGGTCGCTTGTCTG 6 - 742
131 eukaryotic translation elongation factor 1 alpha 1 (EEF1A1), mRNA." "All of the contents of a cel
132 subcellular structures [goid 5737] [evidence IEA]; All of the contents of a cell excluding the plas
133 structures [goid 5737] [evidence IEA]; All of the contents of a cell excluding the plasma membrane
134 [goid 5737] [pmid 3512269] [evidence TAS]; That part of the cytoplasm that does not contain membranous or
135 [evidence EXP]; A multisubunit nucleotide exchange complex that binds GTP and aminoacyl-tRNAs, and
136 ribosome. In humans, the complex is composed of four subunits, alpha, beta, delta and gamma [goid 5
137 amino acid residues to a nascent polypeptide chain during protein biosynthesis [goid 6414] [evidenc
138 nascent polypeptide chain during protein biosynthesis [goid 6414] [pmid 8812466] [evidence TAS]; Th
139 polypeptide chain during protein biosynthesis [goid 6414] [pmid 15189156] [evidence EXP] "Interacti
140 nucleoside that is esterified with (ortho)phosphate or an oligophosphate at any hydroxyl group on t
141 !Platform_contact_address - 9995 Towne Centre Dr
142 !Platform_contact_email - jason.schaeffer@illumina.com
143 !Platform_contact_phone - 979-242-2222
144 !Platform_contact_institution - Illumina, Inc.
```

R/GEOquery

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

Parsed with column specification:
cols(
  ID_REF = col_character(),
  GSM1394594 = col_double(),
  GSM1394595 = col_double(),
  GSM1394596 = col_double(),
  GSM1394597 = col_double(),
  GSM1394598 = col_double(),
  GSM1394599 = col_double(),
  GSM1394600 = col_double(),
  GSM1394601 = col_double(),
  GSM1394602 = col_double(),
  GSM1394603 = col_double(),
  GSM1394604 = col_double(),
  GSM1394605 = col_double()
)
File stored at:
C:\Users\vander11\AppData\Local\Temp\RtmpAXemR3/GPL10558.soft
> summary(exampleData)
Length Class      Mode
GSE57820_series_matrix.txt.gz 1   ExpressionSet S4
>
```

```
> exampleData <- exampleData[[1]]
> exampleData
ExpressionSet (storageMode: lockedEnvironment)
assayData: 47323 features, 12 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM1394594 GSM1394595 ... GSM1394605 (12 total)
  varLabels: title geo_accession ... transfected with:ch1 (45 total)
  varMetadata: labelDescription
featureData
  featureNames: ILMN_1343291 ILMN_1343295 ... ILMN_3311190 (47323 total)
  fvarLabels: ID Species ... GB_ACC (30 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
Annotation: GPL10558
> |
```

R/GEOquery

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

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

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

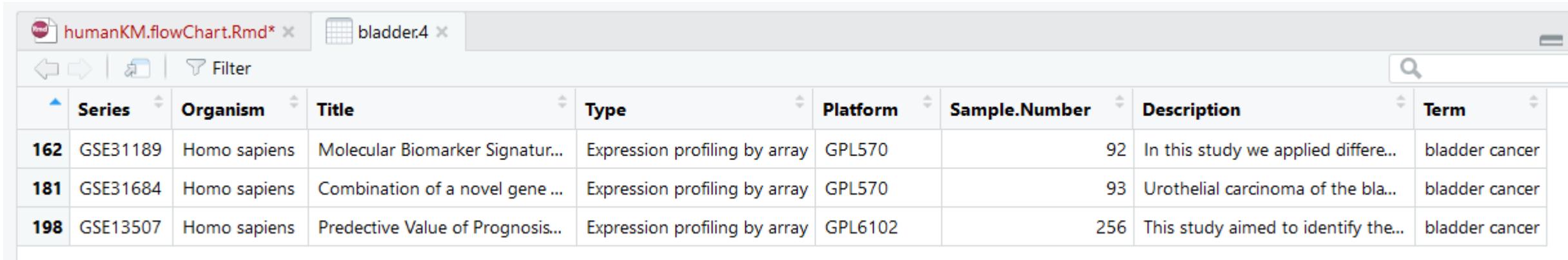
Browsing GEO in R

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

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

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

Browsing GEO in R



The screenshot shows a RStudio environment with two tabs open: 'humanKM.flowChart.Rmd*' and 'bladder.4'. The 'bladder.4' tab displays a data grid with the following columns: Series, Organism, Title, Type, Platform, Sample.Number, Description, and Term. The data grid contains the following rows:

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

```
> dim(bladder.4)
[1] 3 8
> colnames(bladder.4)
[1] "Series"          "Organism"        "Title"           "Type"            "Platform"
[6] "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 SRA ER positive breast cancer
Create alert Advanced

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

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

Other Search results
aligned data (141) Items: 1 to 20 of 573
<< First < Prev Page 1 of 29 Next > Last

[Clear all](#)

[Show additional filters](#)

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

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

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

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

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

SRA Database

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

SRX1363845 : GSM1915703 : MCF7_Cntrl_3; Homo sapiens; RNA-Seq 2 ILLUMINA (Illumina HiSeq 2500) runs: 75.4M spots, 3.7G bases, 2.3Gb downloads															
Submitted by: NCBI (GEO)															
Study: RNA-seq of YB5 and MCF7 treated with different doses of decitabine PRJNA299580 • SRP065220 • All experiments • All runs show Abstract															
Sample: MCF7_Cntrl_3 SAMN04202263 • SRS1126530 • All experiments • All runs Organism: Homo sapiens															
Library: <i>Instrument:</i> Illumina HiSeq 2500 <i>Strategy:</i> RNA-Seq <i>Source:</i> TRANSCRIPTOMIC <i>Selection:</i> cDNA <i>Layout:</i> SINGLE <i>Construction protocol:</i> RNA was isolated using Rneasy Mini Kit (Qiagen) Strand-specific of RNA using TruSeq stranded total RNA with Ribo-Zero Gold (Illumina)															
Experiment attributes: <i>GEO Accession:</i> GSM1915703															
Links:															
Runs: 2 runs, 75.4M spots, 3.7G bases, 2.3Gb															
<table border="1"><thead><tr><th>Run</th><th># of Spots</th><th># of Bases</th><th>Size</th><th>Published</th></tr></thead><tbody><tr><td>SRR2753169</td><td>37,479,476</td><td>1.9G</td><td>1.1Gb</td><td>2017-01-31</td></tr><tr><td>SRR2753170</td><td>37,904,723</td><td>1.9G</td><td>1.1Gb</td><td>2017-01-31</td></tr></tbody></table>	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
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

National Library of Medicine
National Center for Biotechnology Information

Log in

SRA SRA Advanced Search Help

Full ▾ Send to: ▾

SRX1363851: GSM1915709: YB5_Cntrl_3; Homo sapiens; RNA-Seq
1 ILLUMINA (Illumina HiSeq 2500) run: 37.6M spots, 1.9G bases, 1.1Gb downloads

Submitted by: NCBI (GEO)

Study: RNA-seq of YB5 and MCF7 treated with different doses of decitabine
[PRJNA299580](#) • [SRP065220](#) • [All experiments](#) • [All runs](#)
[hide Abstract](#)
RNA-seq was performed after YB5 cells were treated with 1uM decitabine, and MCF7 cells were treated with 100nM decitabine Overall design: Biological triplicates were performed for a total of 6 samples. Fold change of each gene was calculated by comparing change in expression after inhibitor treatment to expression in the control samples from GSE73966 for YB5, and GSE74036 for MCF7. These control samples have been re-accessioned here for convenient access to the entire study.

Sample: YB5_Cntrl_3
[SAMN04202269](#) • [SRS1126524](#) • [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 RNA libraries were generated from 1µg of RNA using TruSeq stranded total RNA with Ribo-Zero Gold (Illumina)

Experiment attributes:
GEO Accession: [GSM1915709](#)

Links:
NCBI link: [NCBI Entrez \(gds\)](#)

Runs: 1 run, 37.6M spots, 1.9G bases, [1.1Gb](#)

Run	# of Spots	# of Bases	Size	Published
SRR2753179	37,627,007	1.9G	1.1Gb	2017-01-31

Related information

BioProject

BioSample

GEO DataSets

PMC

PubMed

Taxonomy

Recent activity

Turn Off Clear

SRP065220 (12) SRA

breast cancer (167692) SRA

GPAT4 glycerol-3-phosphate acyltransferase 4 [Homo sapiens] Gene

137964[uid] (1) Gene

See more...



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.

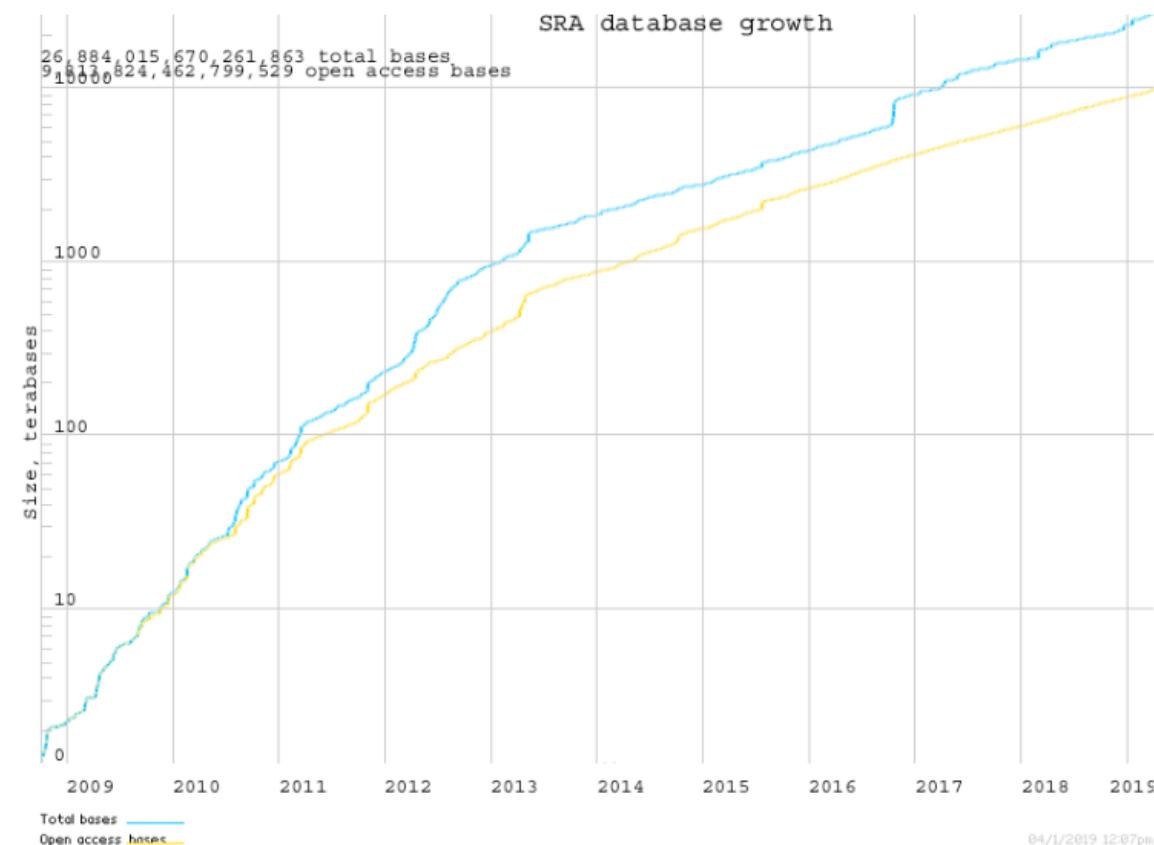
SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

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)



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: [? 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

[Program History](#)[TCGA Cancers Selected for Study](#)[Publications by TCGA](#)[Using TCGA](#)[Contact](#)

The Cancer Genome Atlas Program

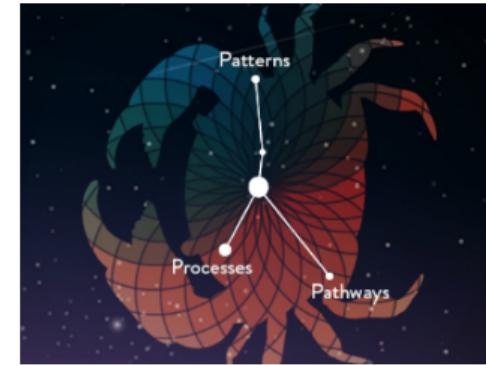
The Cancer Genome Atlas (TCGA), a landmark [cancer genomics](#) program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between the National Cancer Institute and the National Human Genome Research Institute began in 2006, bringing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

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



TCGA Outcomes & Impact

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



TCGA's PanCancer Atlas

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



TCGA – Data Portal

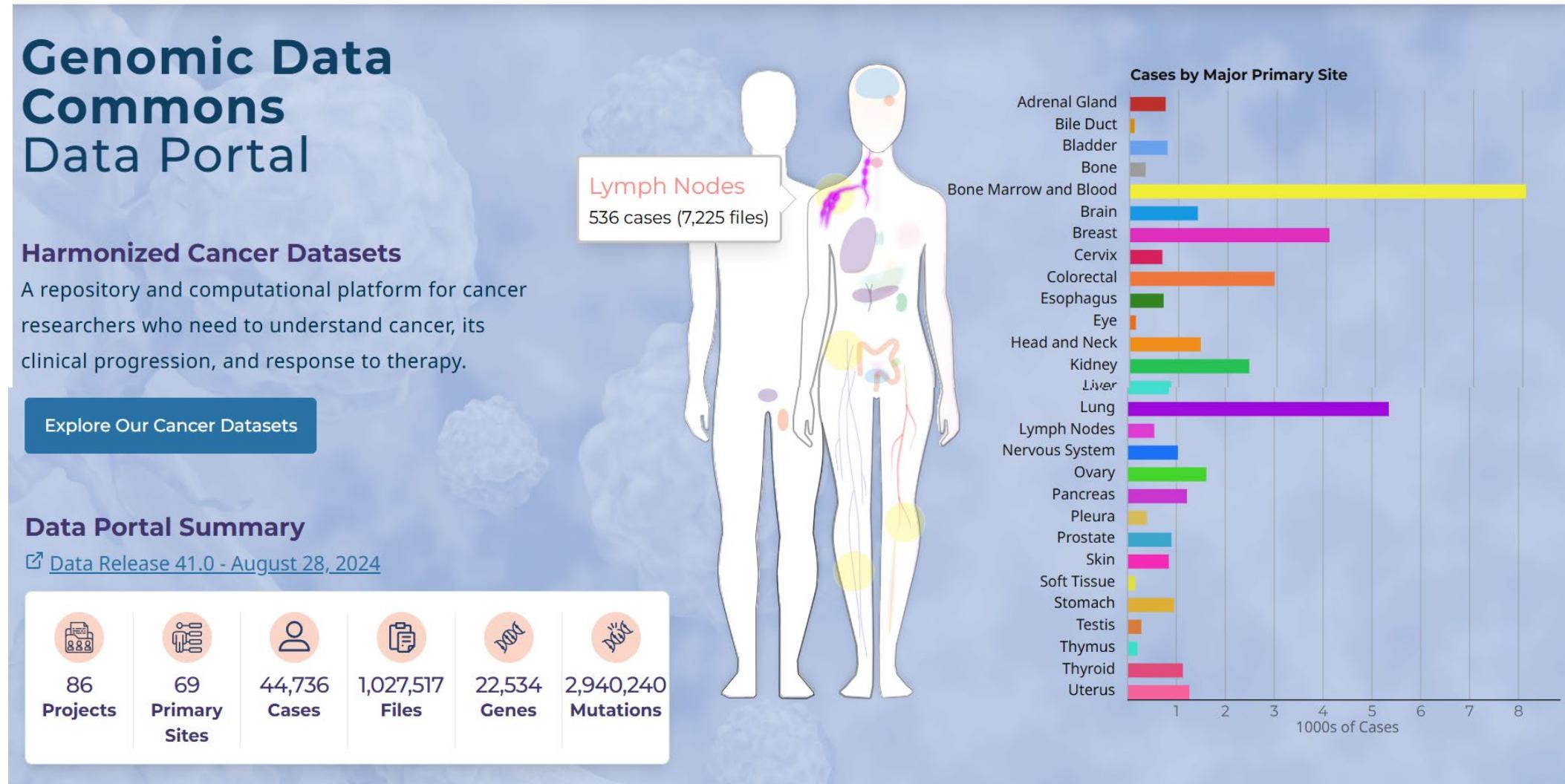
<https://portal.gdc.cancer.gov/>

NATIONAL CANCER INSTITUTE
GDC Data Portal

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

Analysis Center Projects Cohort Builder Repository

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



Analysis Center

Projects

Cohort Builder

Repository

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

Unsaved_Cohort



44,736 CASES



Cohort not saved

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

NATIONAL CANCER INSTITUTE
GDC Data Portal

Analysis Center Projects Cohort Builder Reports

Video Guides Send Feedback

Logout GDC Apps

Unsaved_Cohort Cohort not saved

Download Associated Remove All From Cart

+ Add a Custom Filter

TOTAL OF 1,027,517

JSON TSV

Cart Access File

Controlled Open

63ab7e0b-7264-4cde-a35d-94c30ec3942e.mirnaseq.isoforms.quantification.txt

Name Files

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 Sequenc...	86,977 (8.46%)
Tissue Slide	21,711 (2.11%)
WGS	71,871 (6.99%)
WXS	299,706 (29.17%)

44,736 CASES

Cases Project

Unsaved_Cohort

Cohort not saved

44,736 CASES

ANALYSIS TOOLS

BAM Slicing Download ▾
25,269 Cases

Clinical Data Analysis ▾
44,736 Cases

Cohort Comparison ▾
44,736 Cases

Cohort Level MAF ▾
17,771 Cases

Gene Expression Clustering ▾
20,712 Cases

Mutation Frequency ▾
18,640 Cases

OncoMatrix ▾
18,640 Cases

ProteinPaint ▾
16,508 Cases

Sequence Reads ▾
25,269 Cases

Set Operations ▾

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](https://doi.org/10.18129/B9.bioc.TCGAbiolinks) [f](#) [t](#)

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



TCGAbiolinks
Help Documents

Different sources: Legacy vs Harmonized

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

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

GDCquery()

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

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

GTEx

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



[*i* About Adult GTEx](#) [Publications](#) [Access Biospecimens](#) [*?* FAQs](#) [Contact](#)

Home Downloads ▾ Expression ▾ Single Cell ▾ QTL ▾ IGV Browser Tissues & Histology ▾ Documentation ▾ About ▾

Search Gene or SNP ID.

Gene expression visualizations now show V10 data. We are working on integrating V10 into the remaining visualizations. [See the News Item for more.](#)



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.



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 »](#)



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 »](#)



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

[Open Access Data](#)[Protected Access Data](#)

Gene expression datasets 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/>

 EMBL-EBI home  Services  Research  Training  About us EMBL-EBI 

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical

All



Search

Example searches: [blast](#) [keratin](#) [bfl1](#) | [About EBI Search](#)

Find data resources 

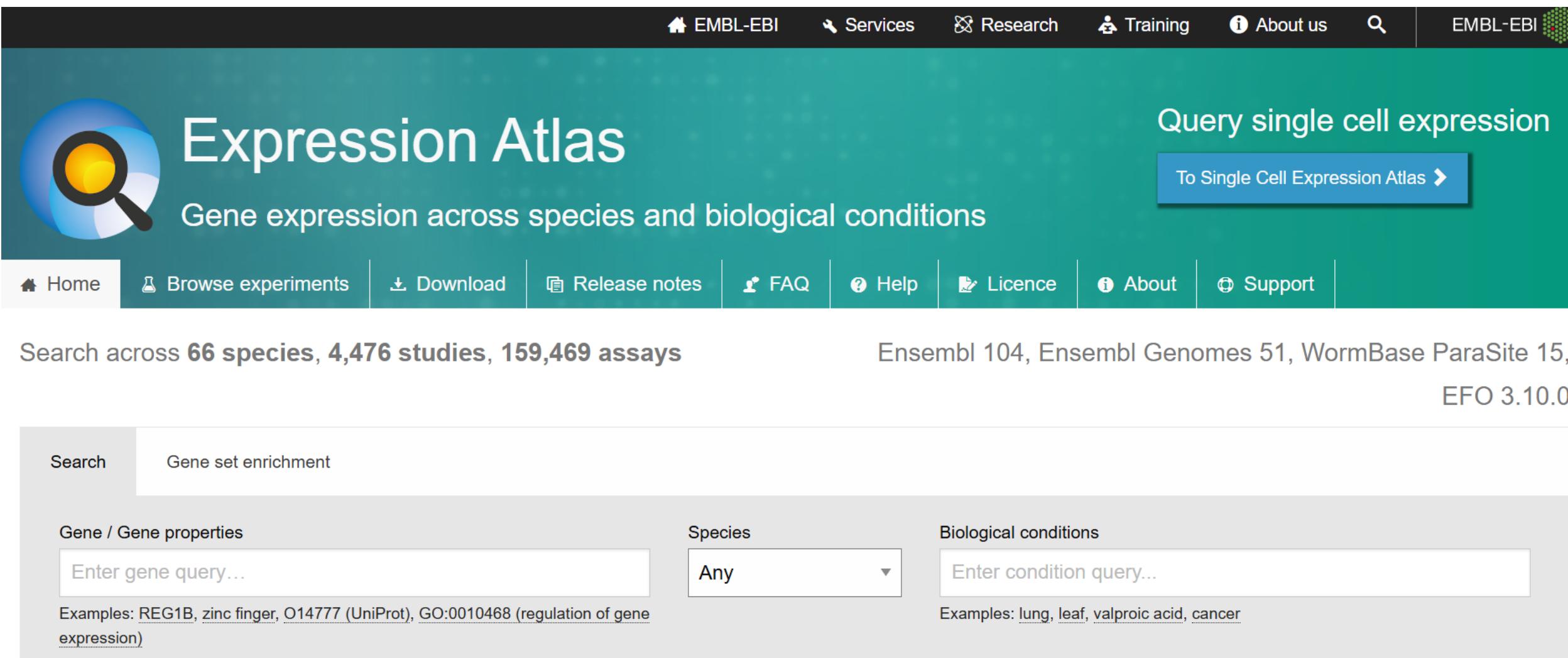
Submit data 

Explore our research 

Train with us 

EBI Expression Atlas

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



The screenshot shows the main interface of the EBI Expression Atlas. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search icon. The EMBL-EBI logo is also present. Below the header, a large teal banner features the "Expression Atlas" logo (a magnifying glass over a blue globe) and the text "Expression Atlas" and "Gene expression across species and biological conditions". To the right of the banner is a call-to-action button: "Query single cell expression" with a link to "To Single Cell Expression Atlas". The main content area includes a search bar at the top, followed by sections for "Gene / Gene properties" (with a query input field), "Species" (set to "Any"), and "Biological conditions" (with a query input field). Below these are examples of search terms for each category.

EMBL-EBI Services Research Training About us EMBL-EBI

Expression Atlas

Gene expression across species and biological conditions

Query single cell expression

To Single Cell Expression Atlas ➔

Home Browse experiments Download Release notes FAQ Help Licence About Support

Search across **66 species, 4,476 studies, 159,469 assays**

Ensembl 104, Ensembl Genomes 51, WormBase ParaSite 15, EFO 3.10.0

Search Gene set enrichment

Gene / Gene properties

Enter gene query...

Examples: REG1B, zinc finger, Q14777 (UniProt), GO:0010468 (regulation of gene expression)

Species

Any

Biological conditions

Enter condition query...

Examples: lung, leaf, valproic acid, cancer

GWAS Catalog

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



The screenshot shows the main landing page of the GWAS Catalog. At the top, there's a navigation bar with links for Diagram, Submit, Download, Learn, About, Blog, EMBL-EBI, and NIH. To the left of the main content area is a circular graphic representing a genome or association study results. The central title "GWAS Catalog" is displayed in large, bold, black font. Below it, a subtitle reads "The NHGRI-EBI Catalog of human genome-wide association studies". A search bar with the placeholder "Search the catalog" and a magnifying glass icon is positioned below the subtitle. Below the search bar, a text box contains examples of study identifiers: "Examples: Parkinson disease, rs3093017, Yao, 2q37.2, HBS1L, 6:167120000-167130000, GCST90132222, PMID:35241825". In the top right corner of the main area, there are logos for ELIXIR and GCBR (Global Core Biobank Resource).

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

GWAS / Search / rheumatoid arthritis

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*

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

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

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
rs2124992-A	4 x 10 ⁻⁷	-	0.22	1.00	-	-	IL2RA	Rheumatoid arthritis	Rheumatoid arthritis	-	GCST90131439	10:6058760

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

Metabolomic Datasets

<https://hmdb.ca/>

The screenshot shows the HMDB website homepage. At the top, there is a navigation bar with links for "Browse", "Search", "Downloads", "About", and "Contact Us". To the right of the navigation bar is a search bar with a magnifying glass icon and a dropdown menu for "metabolites". Below the navigation bar is a banner featuring the TMIC logo ("Wishart Node TMIC The Metabolomics Innovation Centre") and text about quantitative metabolomics services for biomarker discovery and validation. The main content area features a large image of bookshelves in a library. Overlaid on this image are three orange call-to-action buttons: "Browse Metabolites >>", "Learn More >>", and "What's New >>". In the bottom left corner, the HMDB logo is displayed, consisting of the lowercase letters "h mdb" in blue with a yellow flame icon above the "m", and the full name "The Human Metabolome Database" in smaller text below.

HMDB

Browse ▾ Search ▾ Downloads About ▾ Contact Us

?

Search

metabolites ▾

Search

Wishart Node
TMIC
The Metabolomics Innovation Centre

Quantitative metabolomics services for biomarker discovery and validation.

Browse Metabolites >>

Learn More >>

What's New >>

h mdb
The Human Metabolome Database

MicrobiomeDB

<https://microbiomedb.org/mbio/app>

The screenshot shows the MicrobiomeDB homepage. At the top is a dark header with the logo (a stylized DNA helix) and the text "MicrobiomeDB A data-mining platform for interrogating microbiome experiments". Below the header is a navigation bar with links for "Studies", "Workspace", "About", and "Contact Us". To the right of the navigation is a search bar, social media icons for X, YouTube, and GitHub, and a "Guest" user icon. The main content area features two blue-bordered boxes with information. The first box contains a blue info icon and text about joining a Discord community page. The second box contains a blue info icon and text about an important announcement regarding the beta release of the MicrobiomeDB R package.

JOIN OUR NEW DISCORD COMMUNITY PAGE! Interested in talking directly to the our developers or other members of the MicrobiomeDB community? Have questions or want to request a new feature? Interested in our new R package, but don't know where to start? What are you waiting for?! Join the MicrobiomeDB Discord page by clicking [here](#).

IMPORTANT ANNOUNCEMENT: We're excited to announce the beta release of our new [MicrobiomeDB R package](#)! This package is intended to be companion to our website and provides convenient, programmatic access to all of our curated 16S and 'shotgun' metagenomic datasets. In addition, the package allows you to run many of the same types of analyses that you're already familiar with from the site (e.g. diversity analyses, finding top taxa, differential testing, correlations, and more). As an R package, it is now easy for you to customize your analyses and integrate with other plotting or statistical tools in the R/Bioconductor environment. We're interested to hear your feedback, so don't hesitate to reach out to us on our [Discord page](#)!

Explore the Studies

Analyze data from the publicly available studies below.

Study summaries table	
Anopheles albimanus	Study Details
10.1038/s41396-019-0445-5	<ul style="list-style-type: none">This study assessed the impact of pyrethroid exposure on the internal and cuticle microbiome of <i>Anopheles albimanus</i>.125 samples, each pool of 3 mosquitos. V3-V4 region of 16S rRNA gene.Adult and larval <i>Anopheles albimanus</i> collected in and around Las Cruces, Guatemala.
Bangladesh 5yr	Study Details
DOI: 10.1126/science.aau4735	<ul style="list-style-type: none">This study set out to define the normal maturation of the gut microbiome during the first 5 years of postnatal life.55 members of a birth cohort with consistently healthy anthropometric scores living within the Mirpur district of Dhaka, Bangladesh.2415 stool samples; V4 region of 16S rRNA gene.Prospective cohort design with monthly sampling for the first ~5 years of life.
BONUS-CF	Study Details
DOI: 10.1038/s41591-019-0714-x	<ul style="list-style-type: none">The Baby Observational and Nutrition St (BONUS) set out to identify microbial correlates of poor growth observed in infants with cystic fibrosis (CF).207 infants diagnosed with cystic fibrosis during newborn screening.Shotgun metagenomic sequencing of 12 samples collected from healthy controls and 1157 stool samples from infants with CF collected at months 3, 4, 5, 6, 8, 10 and 12 life.

News

[MicrobiomeDB 34 Released](#)

FRI DEC 29 2023

We are pleased to announce the release of MicrobiomeDB 34. New data in this release. There are no new datasets in this release. New fea... [read more](#)

[MicrobiomeDB 33 Released](#)

THU SEP 14 2023

We are pleased to announce the release of MicrobiomeDB 33. This release was focused on site maintenance. We have some very exciting and major changes coming to virtually every aspect ... [read more](#)

[MicrobiomeDB 32 Released](#)

TUE MAY 30 2023

[See all news](#)

MicrobiomeDB Release 37
7 May 2024

©2024 The VEuPathDB Project Team



Please [Contact Us](#) with any questions or comments

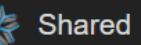
PennVet
Center for Host-Microbial Interactions

Immunology Database and Analysis Portal (ImmPort)

<https://www.immport.org/home>



 Upload  Shared



Analysis



>Data Management and Sharing Plan

Search www.import.org



Documentation

About ▾



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 at 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



National Institute of
Diabetes and Digestive
and Kidney Diseases

Search the Website...



NIDDK Central Repository

Sign In / Register

[Repository Resources](#) ▾

[Helpful Information](#) ▾

[Requests](#) ▾

[About NIDDK-CR](#) ▾

Studies

[Home](#) > Studies

Search



Search for Studies using Study Metadata...

Search

Filters

Research Area

Network

Target Population

Data Availability

Specimen Availability

Condition

1 - 50 of 188 Results [1](#) [2](#) ... [4](#) [Page](#)

Show: 50 ▾

Download Results

List Table

Manage Columns

Study Name	Study Acronym	Research Area	Data Availability	Specimen Availability	Network
Adult Living Donor Liver Transplantation Studies	A2ALL	Liver Disease	Data Available for Request	Specimens Available for Request	Adult Living Donor Liver Transplantation Studies (A2ALL Network)
Acute Liver Failure Study Group: Adult Acute Liver Failure Study	AALF	Liver Disease	Data Available for Request	Specimens Available for Request	Acute Liver Failure Study Group (ALFSG)
African American Study of Kidney Disease and Hypertension Cohort Study	AASK Cohort	Multidisciplinary Research; Kidney Disease	Data Available for Request	Specimens Available for Request	African American Study of Kidney Disease Study Group (AASK)
African American Study of Kidney Disease and Hypertension Study (Clinical Trial)	AASK Trial	Multidisciplinary Research; Kidney Disease	Data Available for Request	Specimens Available for Request	African American Study of Kidney Disease Study Group (AASK)

Institute Specific Repositories

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

Institute or Center	Repository Name	Repository Description	Access to Data	Access Type
ICOs filter		<input type="text" value="Keyword Filter"/>		
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

Genes / Molecules ▾

Enter your search term here (ex: cytochrome AND P450)

Search

?

Select and Search

Species: Mouse (Mus musculus, mm10)

Group: BXD Family

Info

Type: Hippocampus mRNA

Info

Dataset: Hippocampus Consortium M430v2 (Jun06) PDNN

Get Any:

Enter terms, genes, ID numbers in the **Search** field.

Use * or ? wildcards (Cyp*a?, synap*).

Use **quotes** for terms such as "tyrosine kinase".

[see more hints](#)

Tutorials

Webinars & Courses

In-person courses, live webinars and webinar recordings

Tutorials

Tutorials: Training materials in HTML, PDF and video formats

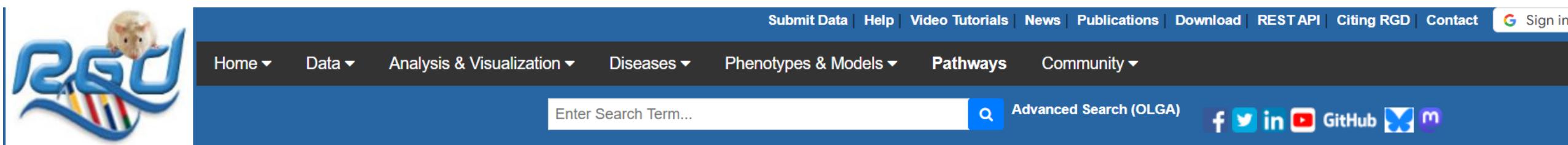
Documentation

Online manuals, handbooks, fact sheets and FAQs

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

Rat Genome Database

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



The screenshot shows the RGD homepage. At the top left is the RGD logo with a cartoon mouse. To its right is a horizontal menu bar with links: Submit Data, Help, Video Tutorials, News, Publications, Download, REST API, Citing RGD, Contact, and Sign in. Below the menu is a secondary navigation bar with Home, Data, Analysis & Visualization, Diseases, Phenotypes & Models, Pathways, and Community. A search bar with placeholder text "Enter Search Term..." is positioned below the navigation. To the right of the search bar are links for Advanced Search (OLGA), followed by social media icons for Facebook, Twitter, LinkedIn, YouTube, GitHub, and a blue butterfly icon.

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/>

The screenshot shows the main interface of the KM Plotter website. At the top, there's a dark header bar with the text "Kaplan-Meier Plotter" on the left and "Breast Cancer" on the right, accompanied by a dropdown menu. Below the header is a navigation bar with links for "KM plotter", "Home", "Download", "Updates", and "Contact". The main content area has a background of a colorful dot pattern. It features a section titled "What is the KM plotter?" followed by a detailed description of the tool's capabilities. Below this, there are three rows of colored buttons. The first row is for mRNA gene chip analysis: "Start KM Plotter for breast cancer" (pink), "Start KM Plotter for ovarian cancer" (teal), "Start KM Plotter for lung cancer" (red), and "Start KM Plotter for gastric cancer" (dark grey). The second row is for mRNA-seq analysis: "Start KM Plotter for liver cancer" (green), "Start KM Plotter for pan-cancer" (dark blue), and a button labeled "In development" (white). The third row is for miRNA analysis: "Start miRpower for breast cancer" (pink), "Start miRpower for liver cancer" (green), and "Start miRpower for pan-cancer" (dark blue).

Kaplan-Meier Plotter

Breast Cancer

Breast Cancer

KM plotter Home Download Updates Contact

What is the KM plotter?

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

mRNA gene chip Start KM Plotter for **breast cancer**

mRNA Start KM Plotter for **ovarian cancer**

mRNA Start KM Plotter for **lung cancer**

mRNA Start KM Plotter for **gastric cancer**

mRNA RNA-seq Start KM Plotter for **liver cancer**

mRNA Start KM Plotter for **pan-cancer**

In development

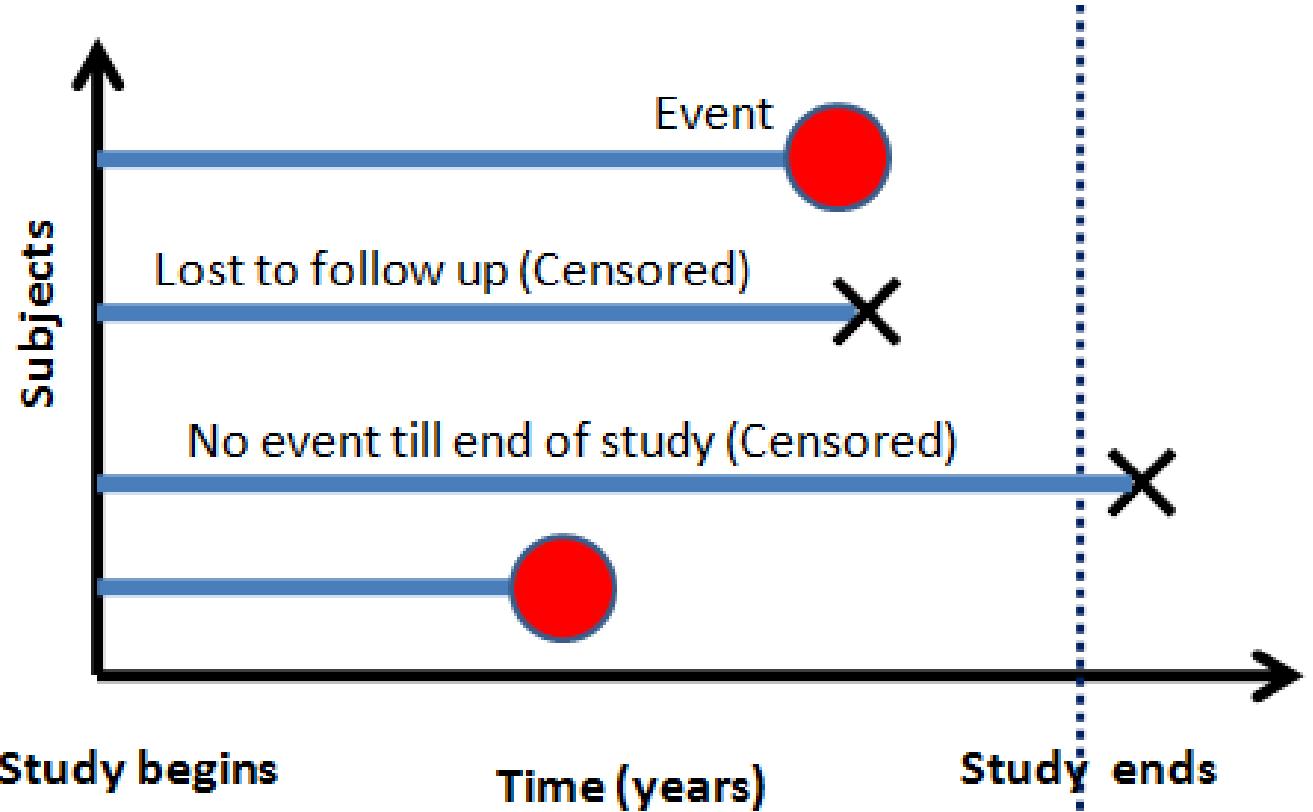
miRNA Start miRpower for **breast cancer**

miRNA Start miRpower for **liver cancer**

miRNA Start miRpower for **pan-cancer**

Quick Sidebar - Survival Data

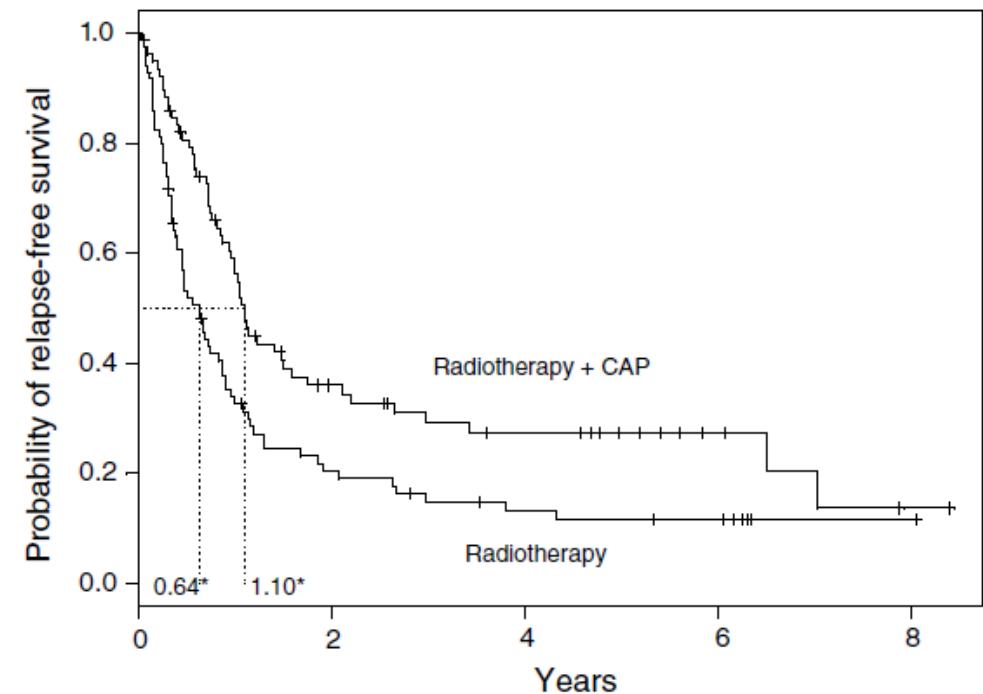
- Often used in clinical trials
- Outcome variable is the time until the occurrence of an event of interest
 - Death (overall survival)
 - Disease specific survival (DSS)
 - Disease-free survival
 - Progression-free survival
 - Metastasis-free survival
- Censored data



Source: alphabeta statistics

Kaplan-Meier survival curves

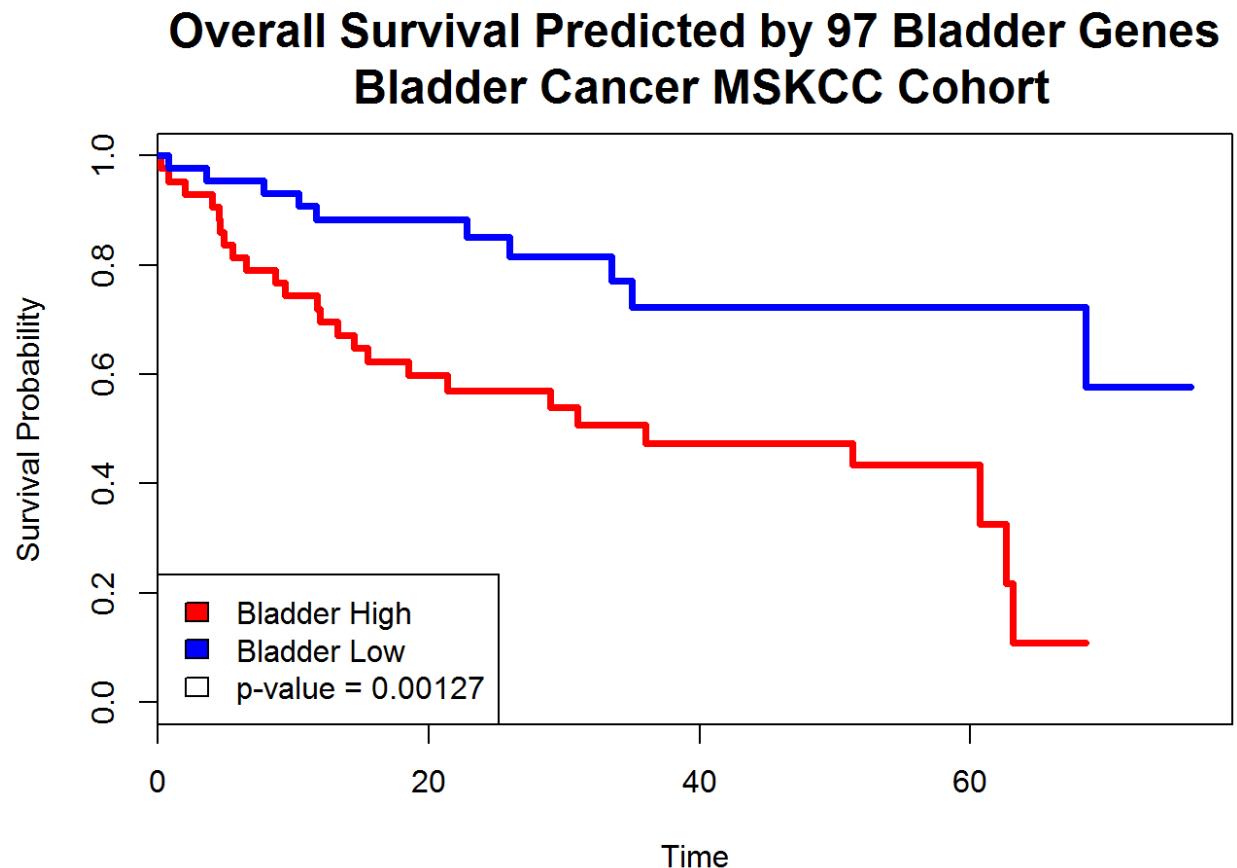
- $S(t)$ = probability of surviving (event-free) from the beginning of the study to time t
- $R(t)$ = cumulative risk of having the event from the beginning of the study to time t
- Cumulative risk = $1 - \text{cumulative survival}$
- Kaplan-Meier method:
 - Nonparametric
 - Calculate interval-specific 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

Kaplan-Meier Plotter | KM plotter | Home | Download | Updates | Contact | Gastric Cancer

Affy id/Gene symbol: C1orf112 GCLC STPG1 | ⓘ Use multigene genes

Split patients by: ⚡ median | ⓘ Auto select best cutoff ⓘ | ⓘ Trichotomization: none ⓘ

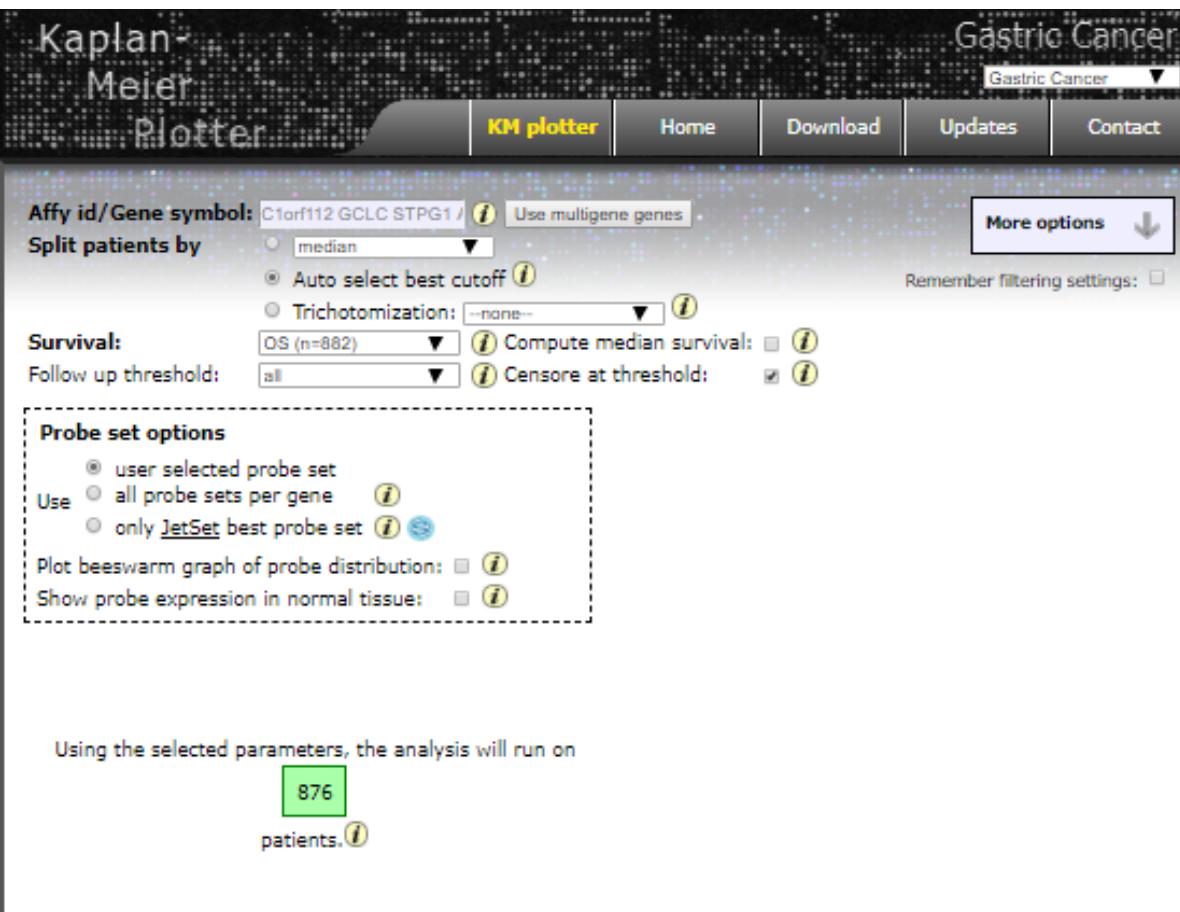
Survival: OS (n=882) ⓘ Compute median survival: ⓘ | ⓘ Follow up threshold: all ⓘ Censore at threshold: ⓘ

Probe set options:

- user selected probe set (selected)
- all probe sets per gene ⓘ
- only JetSet best probe set ⓘ ⓘ

Plot beeswarm graph of probe distribution: ⓘ ⓘ | Show probe expression in normal tissue: ⓘ ⓘ

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



Restrict analysis to subtypes...

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

Restrict analysis to clinical cohorts...

Gender: all | Perforation: all | Treatment: all | HER2 status: all | ⓘ ⓘ

Use following dataset(s) for the analysis:

all | Exclude GSE62254: ⓘ ⓘ

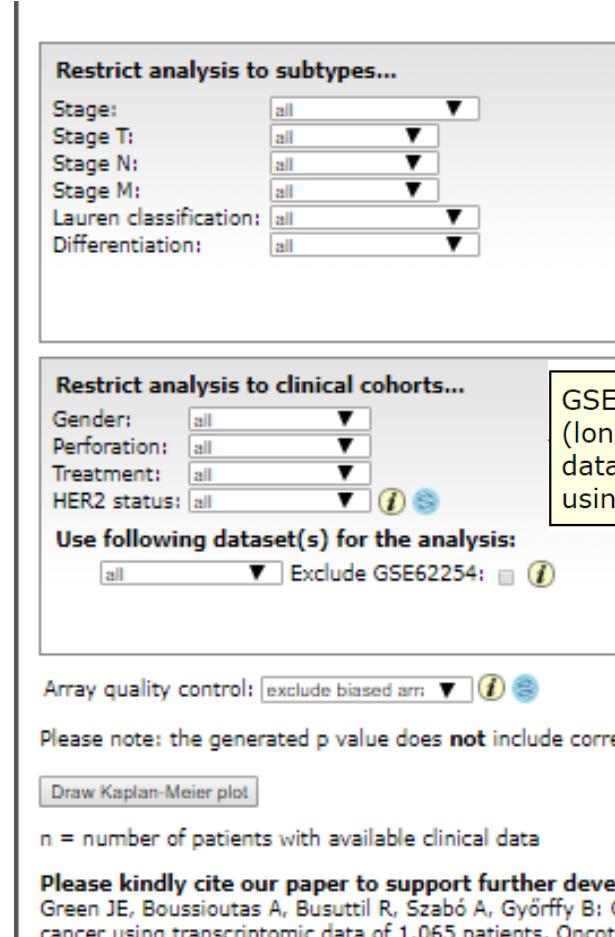
Array quality control: exclude biased arm | ⓘ ⓘ

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

Draw Kaplan-Meier plot

n = number of patients with available clinical data

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



KM Plotter - Results

Plotter | KM plotter | Home | Download

The desired is valid: 205942_s_at (ACSM3),

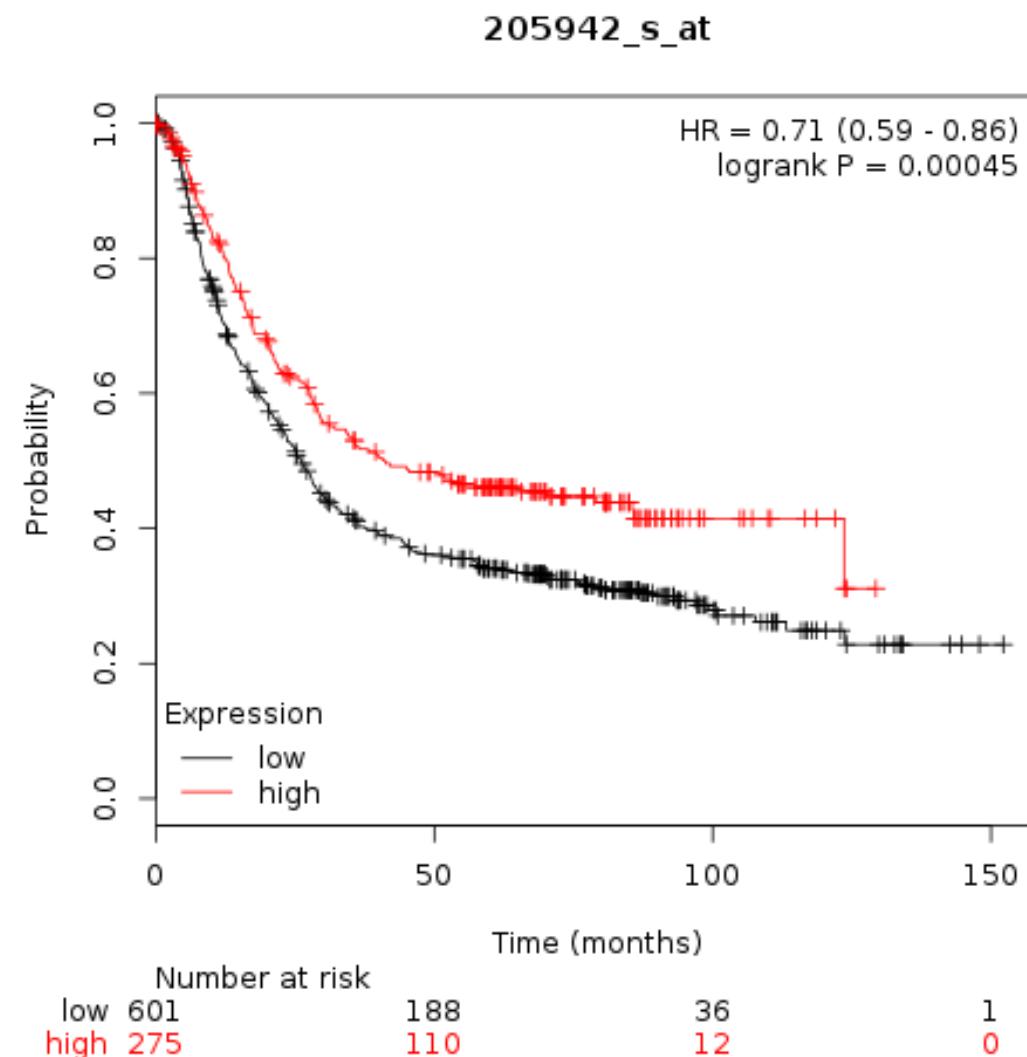
Affy ID: 205942_s_at ACSM3, SA, SAH
Survival: OS
Auto select best cutoff: checked
Follow up threshold: all
Censor at threshold: checked
Compute median over entire database: false
Cutoff value used in analysis: 230
Expression range of the probe: 2 - 2185
Probe set option: user selected probe set
Invert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

Restrictions

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

Results

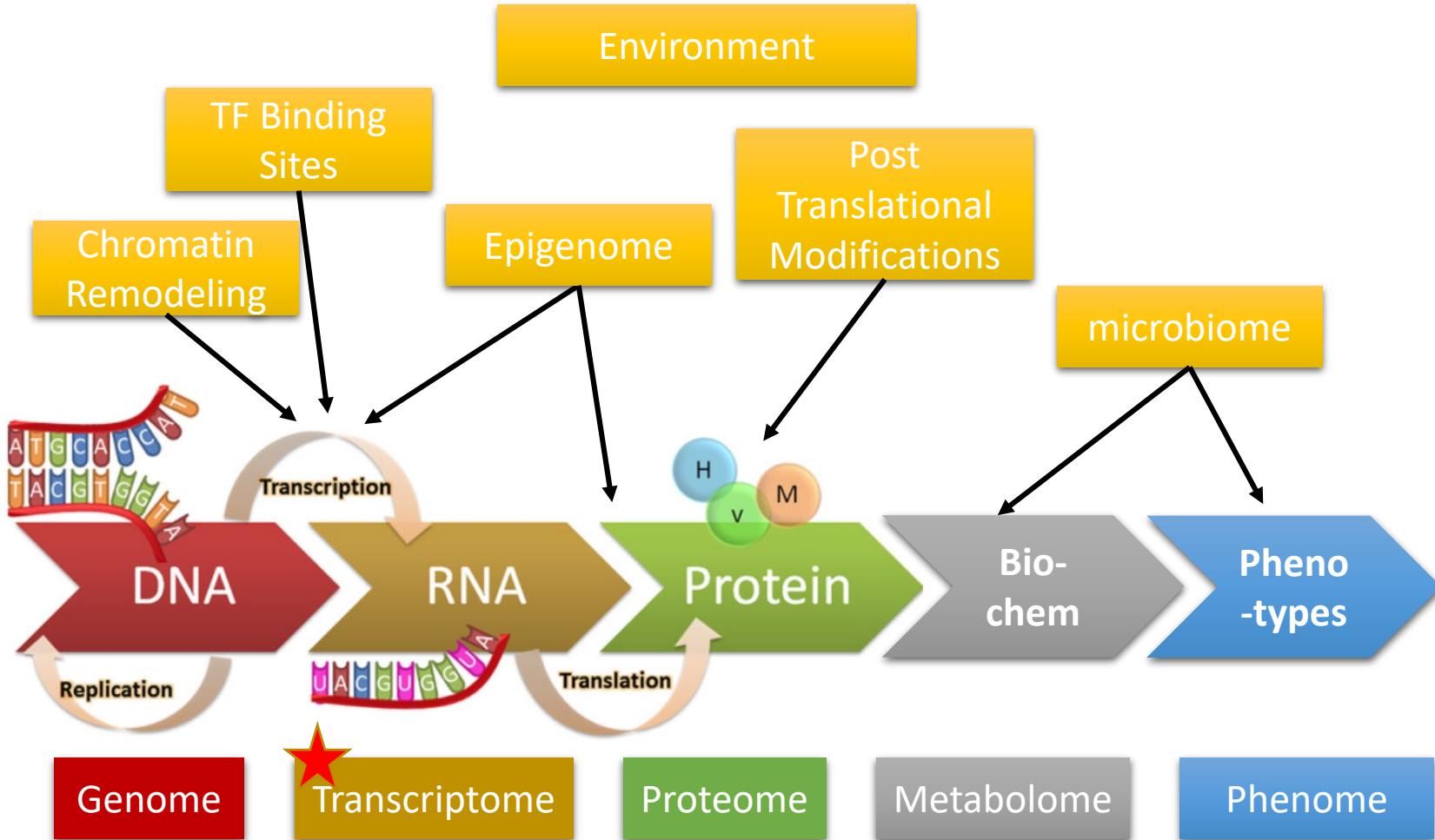
P value: 0.0005
FDR: 10%

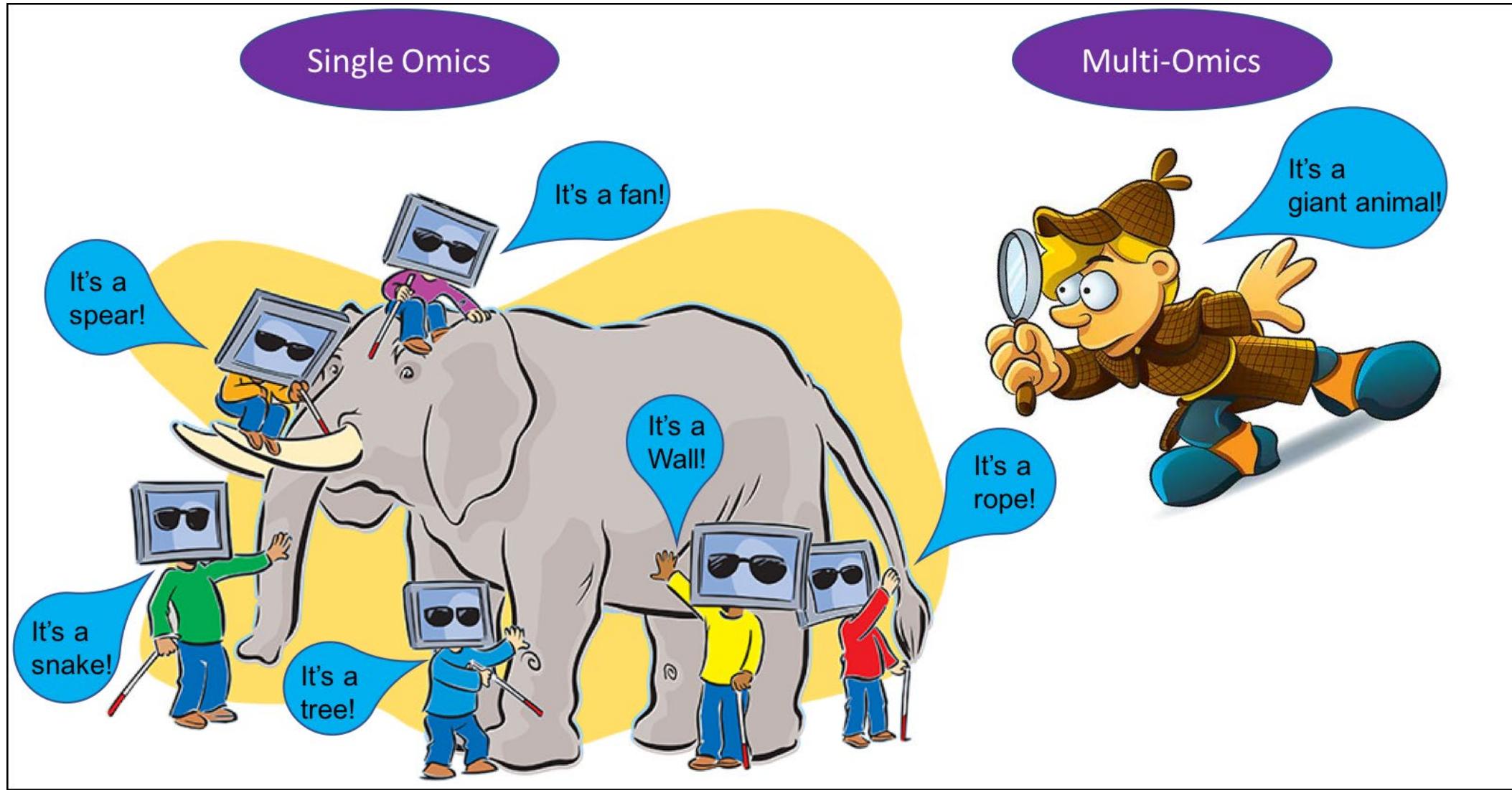


Data Integration

Multiple 'Omics Datasets

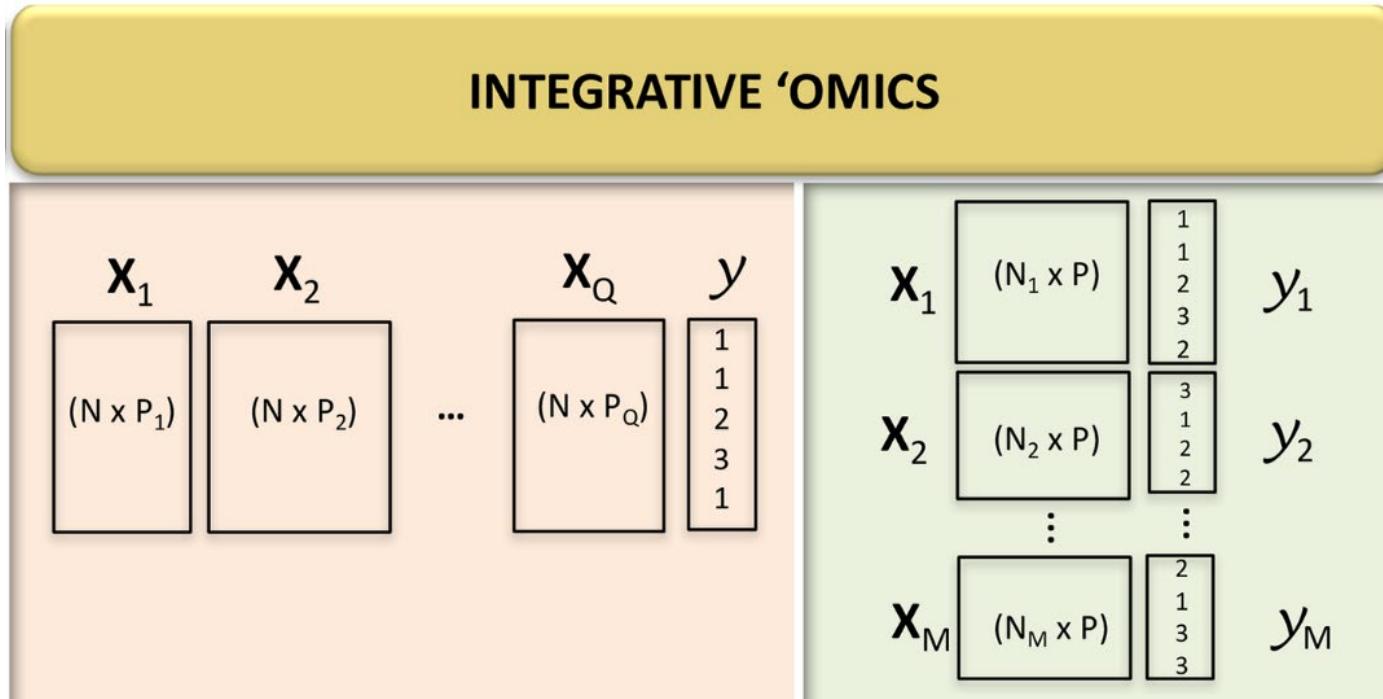
- Want to get a bigger picture on what is happening
- Multiple 'omics datasets





Source: Melgen

Horizontal vs Vertical Data Integration



Horizontal

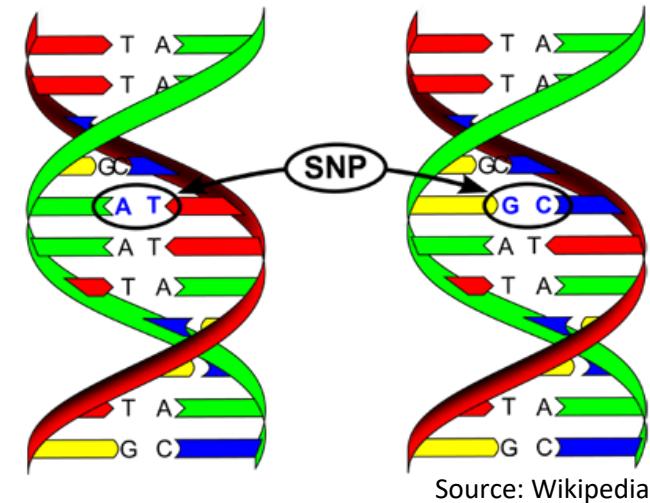
Same technology
Different samples

Vertical

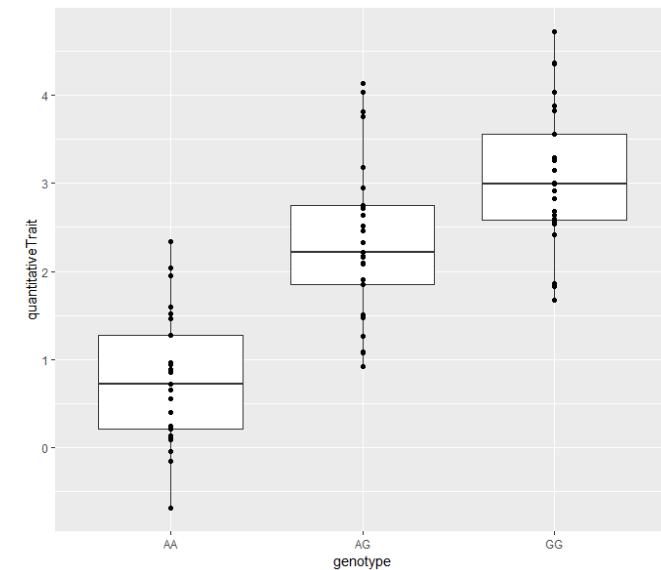
Different technology
Same samples

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)
 - Homozygous 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

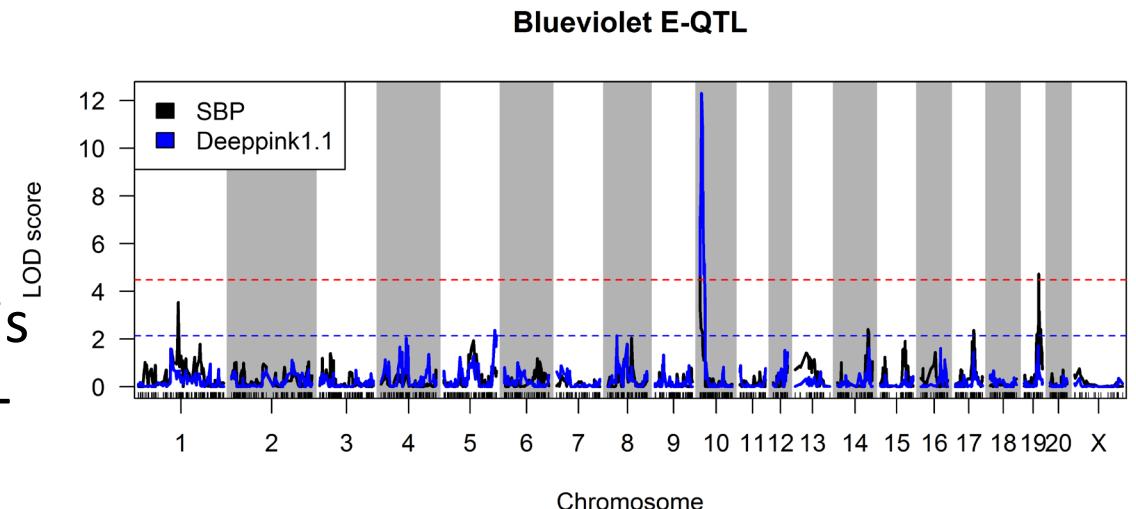


Quantitative Trait Loci (QTL) Mapping

- 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

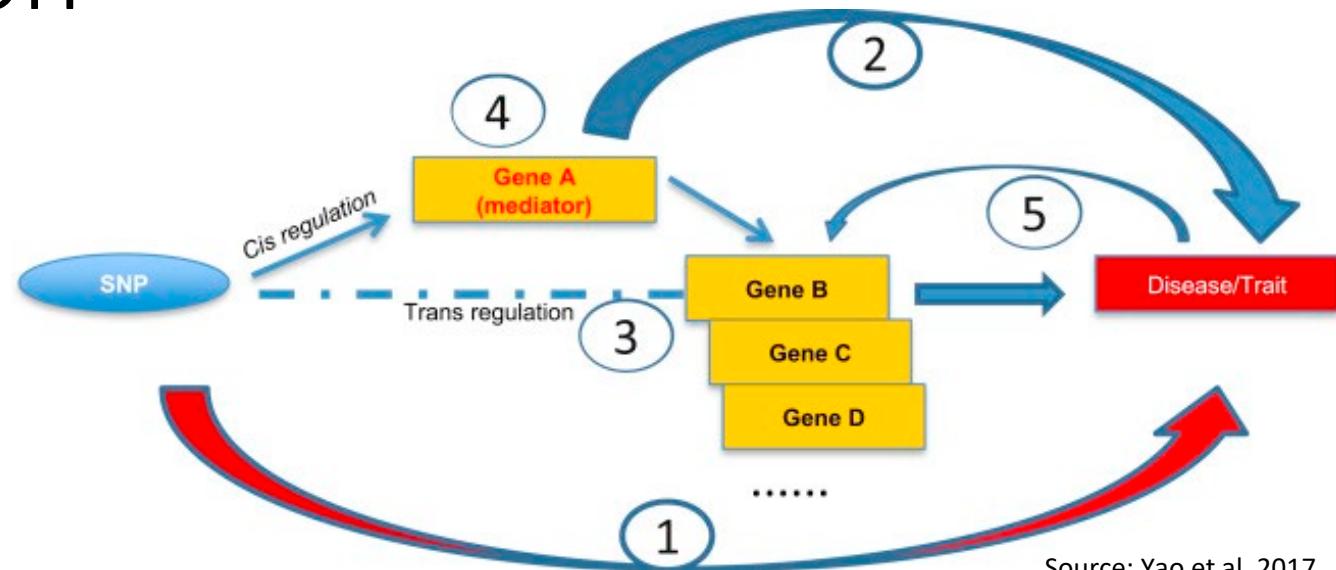
- p-values reported and plotted (Manhattan plot)
- Logarithm of odds “LOD” Score:
 - log₁₀ likelihood ratio comparing hypothesis of a QTL at position λ versus that of no QTL
 - $LOD(\lambda) = \log_{10} \left\{ \frac{P(y|QTL \text{ at } \lambda)}{P(y|no \text{ QTL})} \right\}$
 - This is common among 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

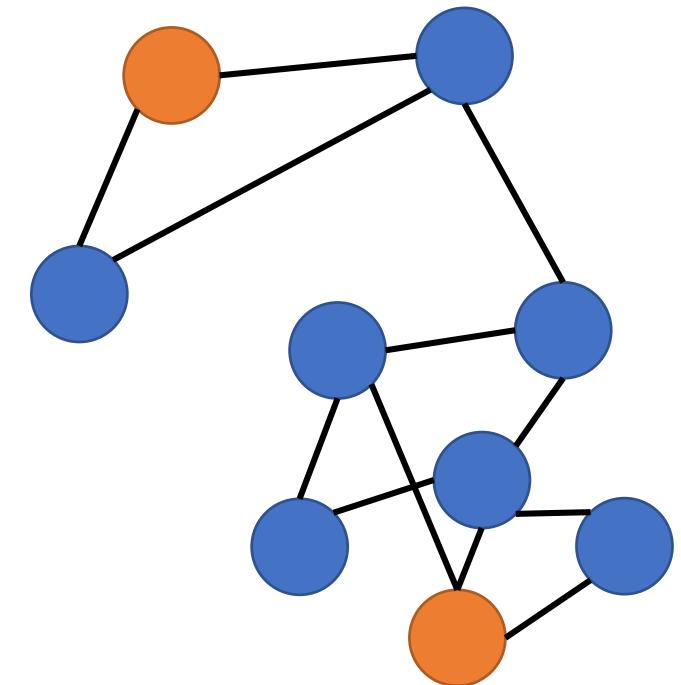
- (1) missense SNP affects protein structure/function
- (2) non-coding SNP affects gene expression (*cis*)
- (3) non-coding SNP affects remote (*trans*) gene expression directly or by
- (4) *cis*-eGene mediation of the *trans*-eQTL-*trans*-eGene association; or
- (5) reverse causality (trait has feedback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

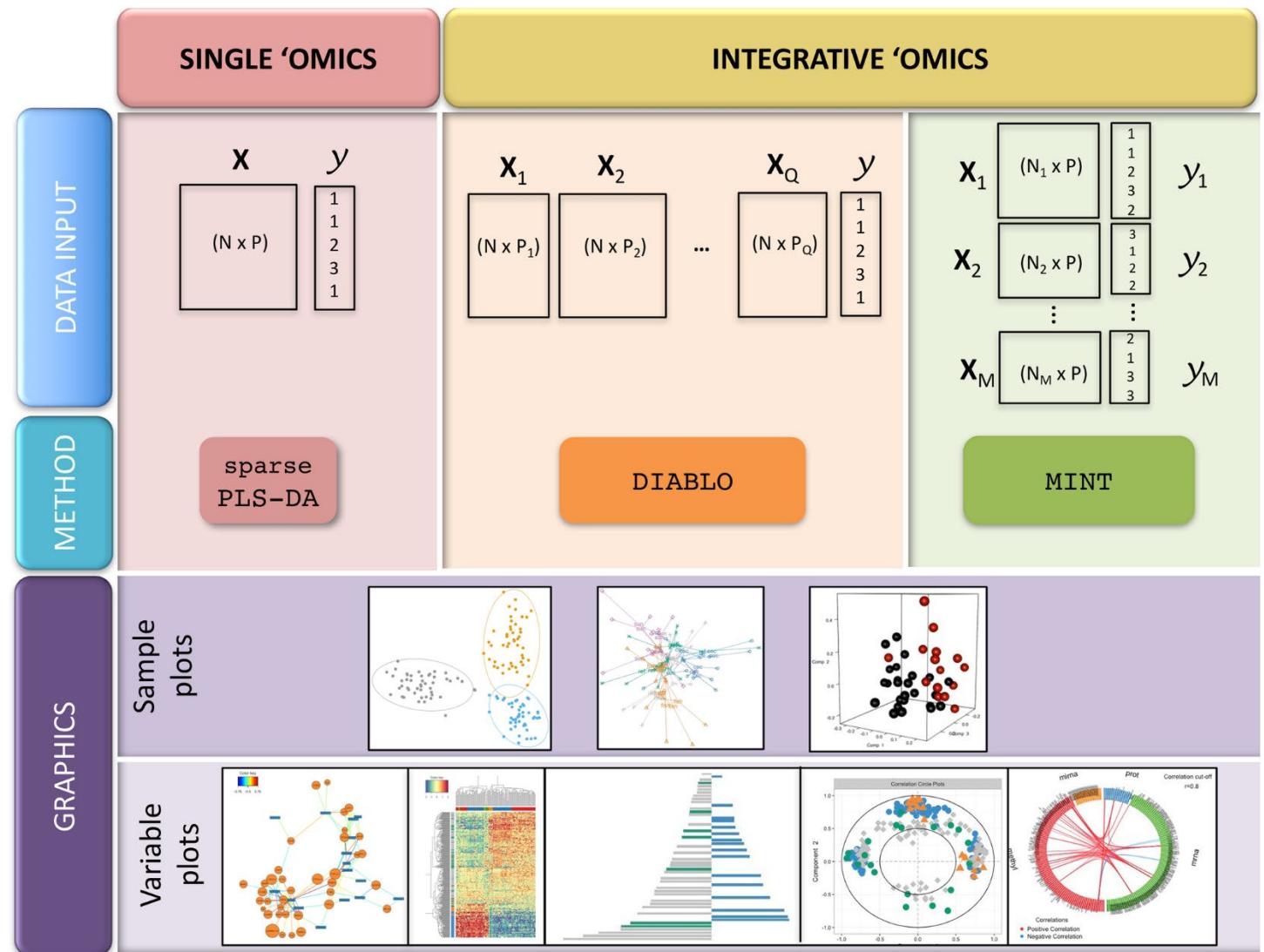
LASSO to get miRNA-mRNA network

- 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



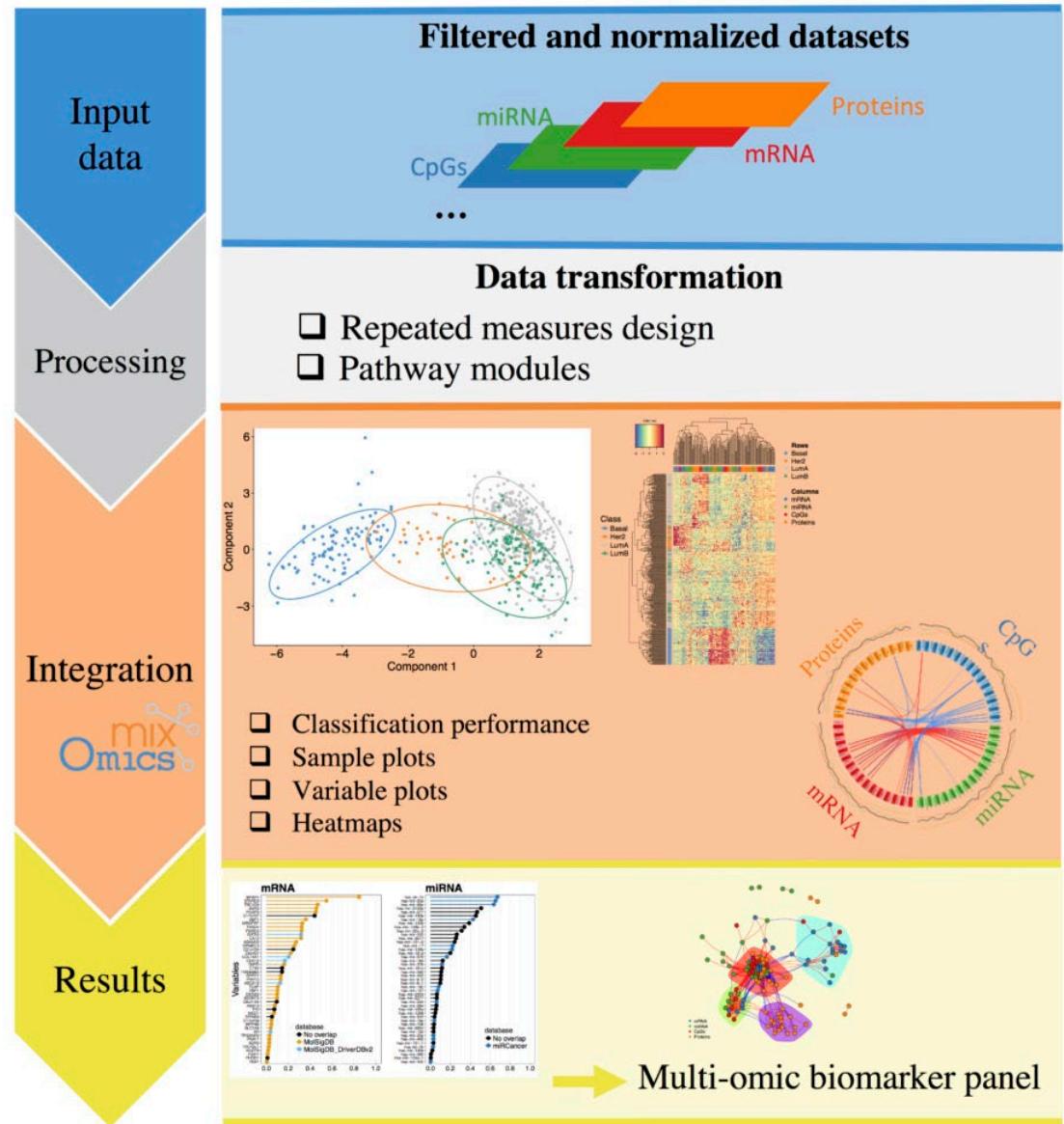
Methods Available in mixOmics

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

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

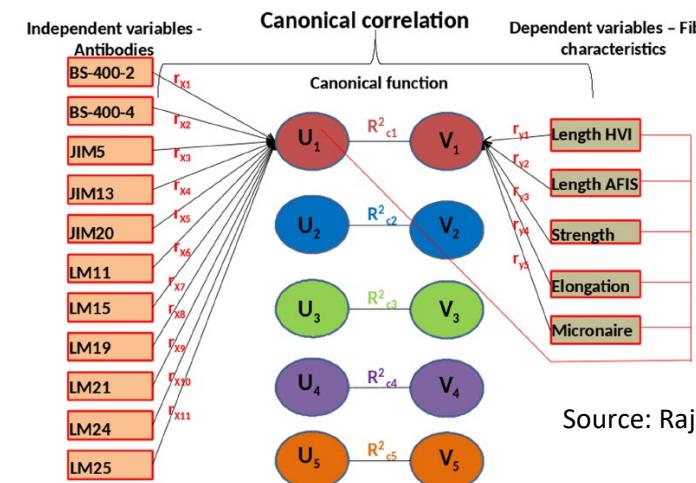
DIABLO

- 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



Canonical Correlation Analysis (CCA)

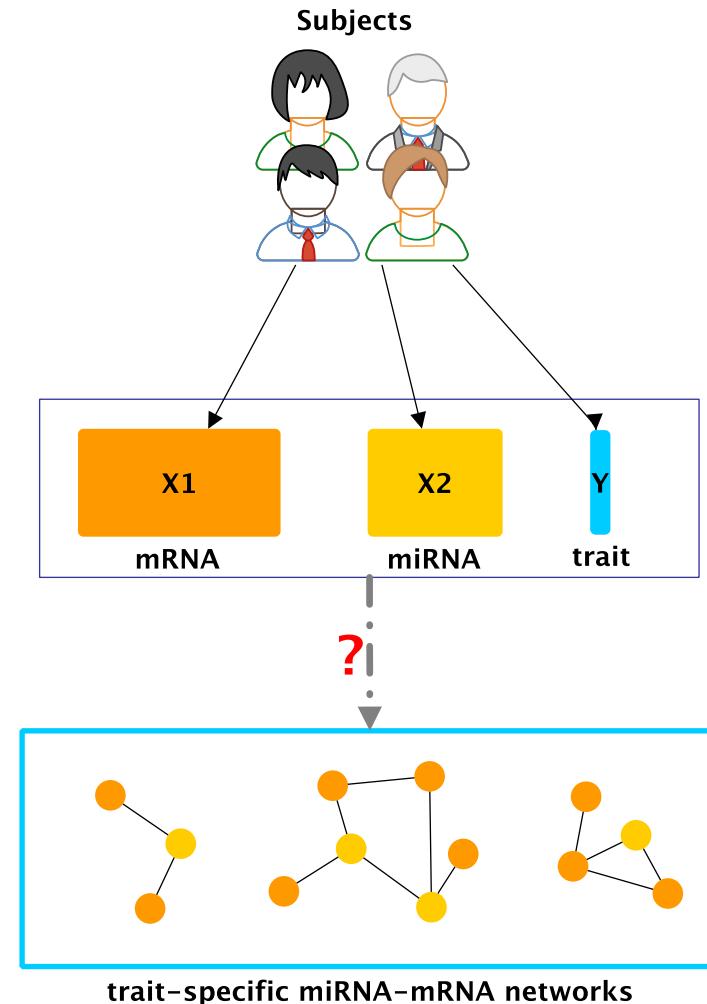
- Multivariate correlation
- Accounts for multi-collinearity
 - Features in a dataset not independent
 - Know not all genes are independent from each other
- Compare sets of variables to sets of variables
- Canonical loadings
 - Variables relationship to own set
 - E.g. gene A expression to gene B expression
- 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 Kechrис & 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 Kechrис

CCA vs Sparse CCA

Set Correlation $R = \text{Cor}(X_1, X_2)^*$

$$= \text{Cor}(w_1 \times X_1, w_2 \times X_2)$$

Sparse Set Correlation $R' = \text{Cor}(X_1, X_2)^*$

$$= \text{Cor}(w_1 \times X_1, w_2 \times X_2)$$

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

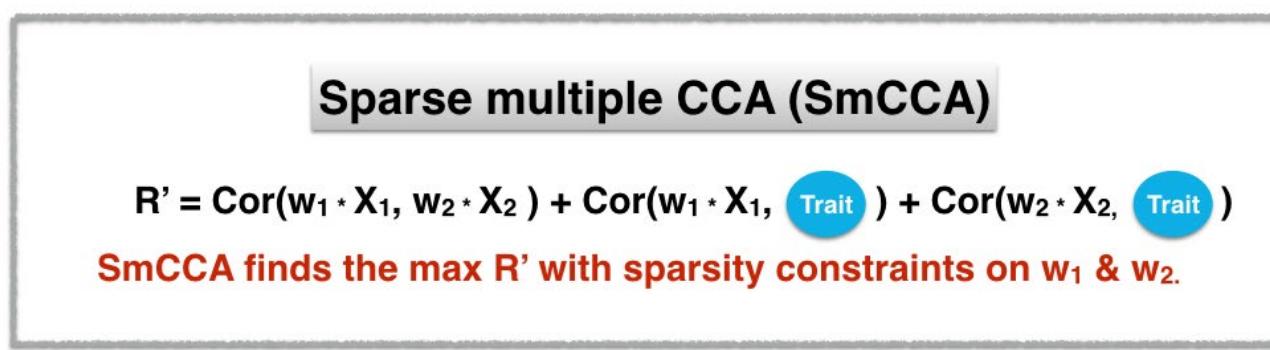
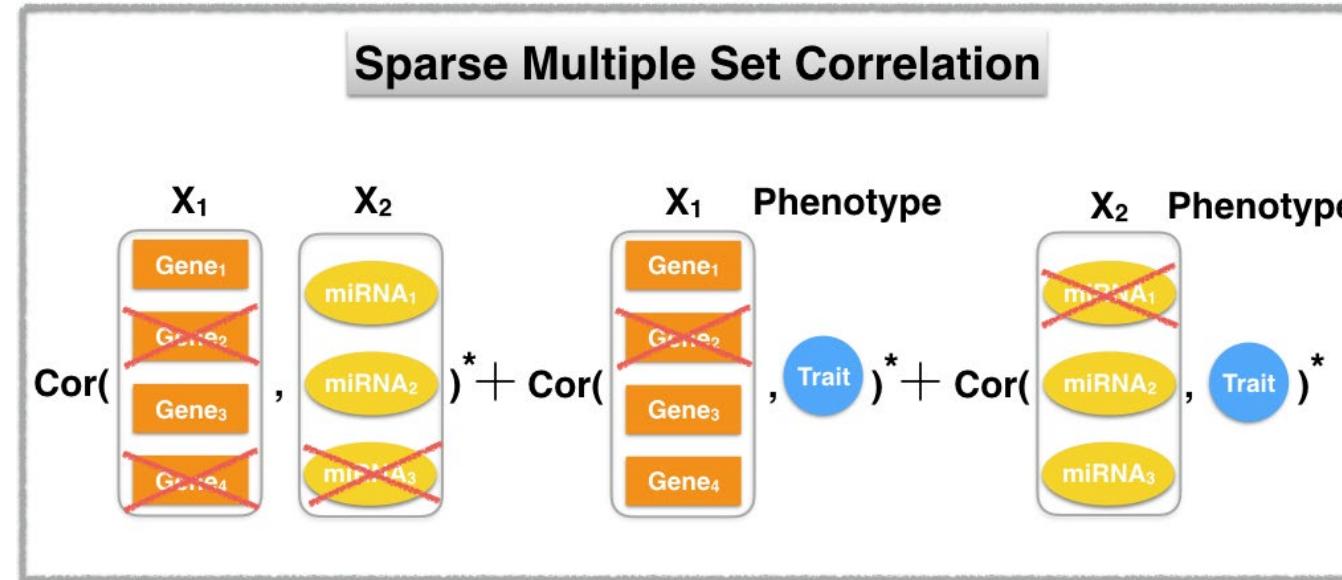
Sample canonical weights:
 $w_1 = (0.10, -0.48, 0.83, 0.22)^t$,
 $w_2 = (0.67, 0.14, 0.73)^t$.

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

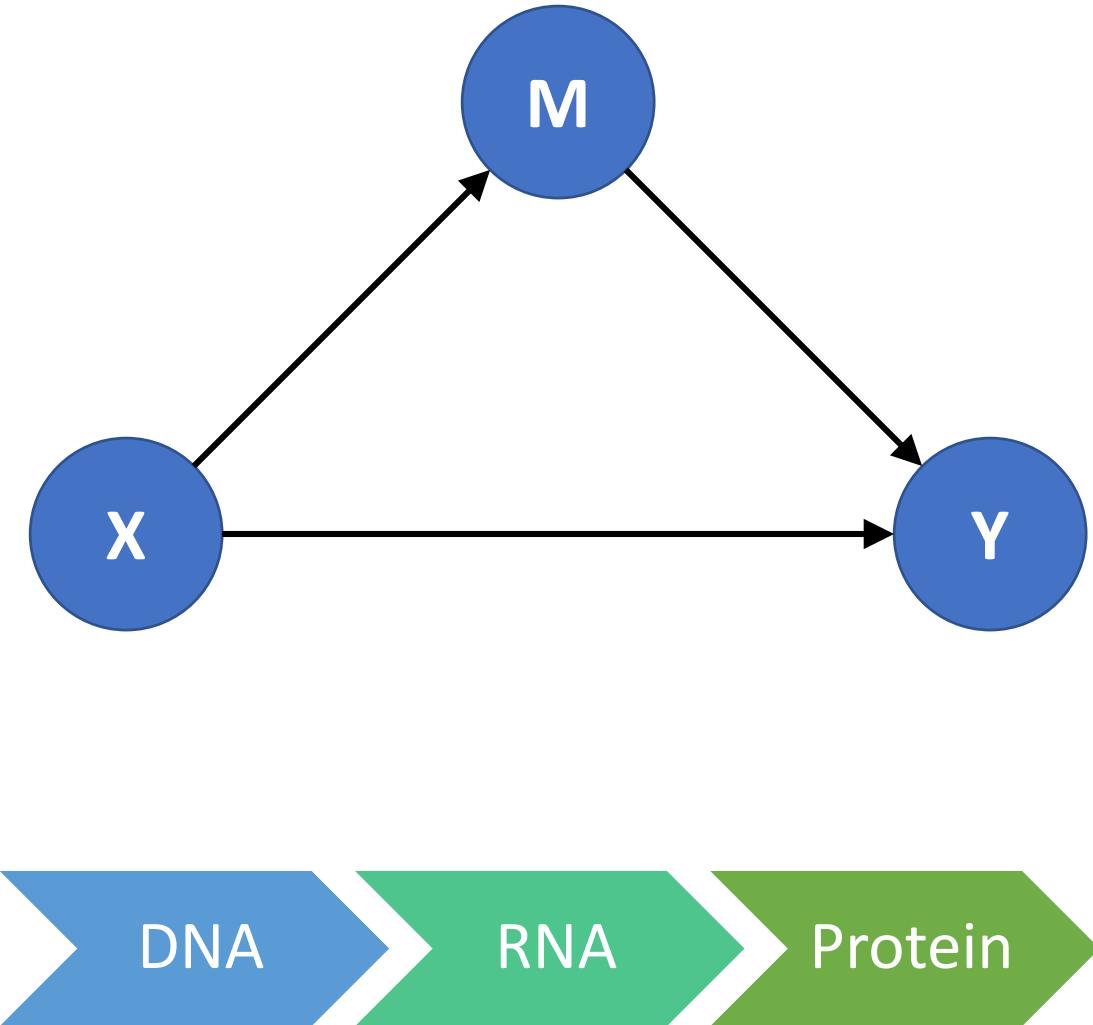


Subsampling of miRNA/mRNA & cross-validation of samples for sparse penalties on weights

Source: Katerina Kechris

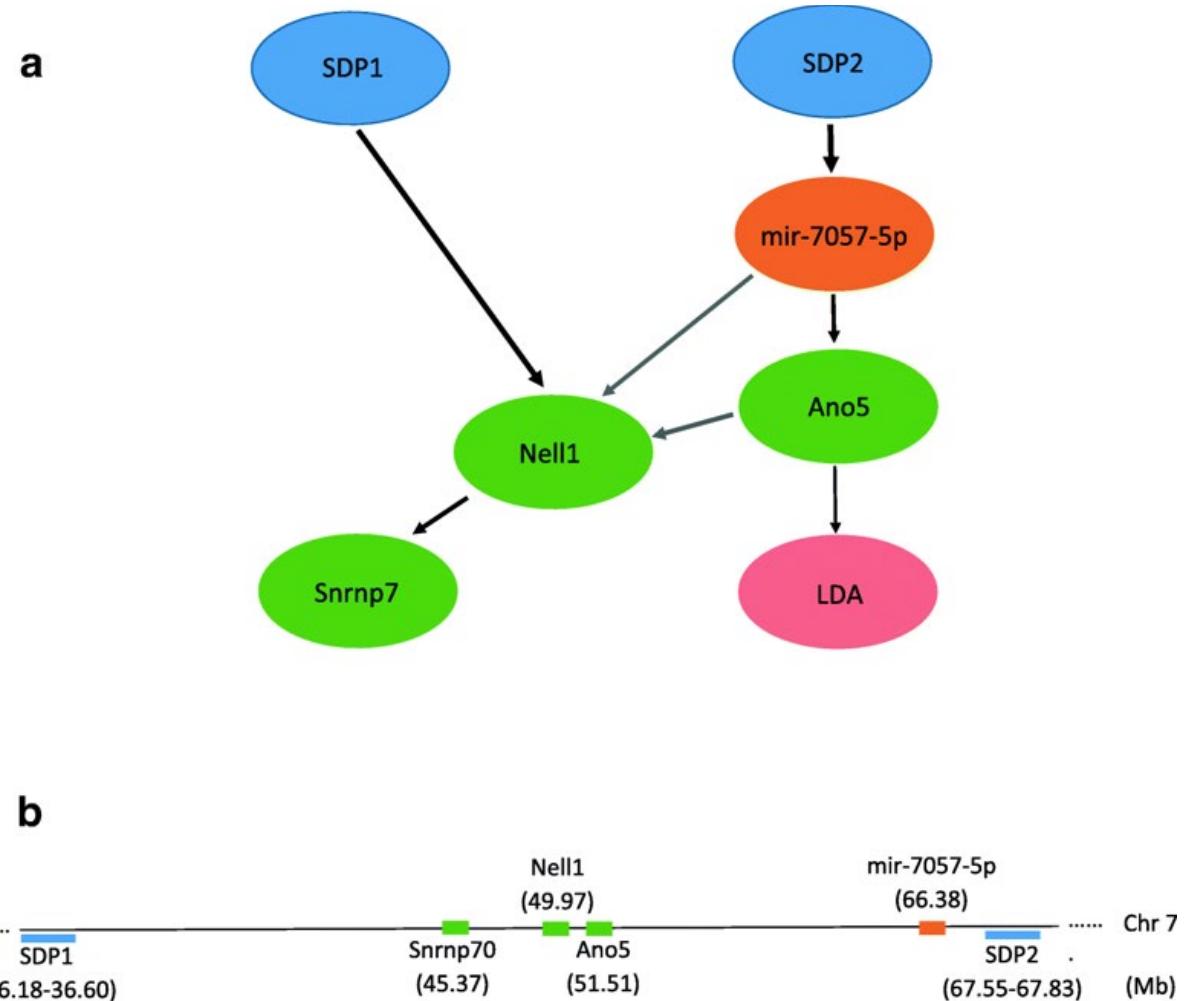
Mediation Analyses

- 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 and 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



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.

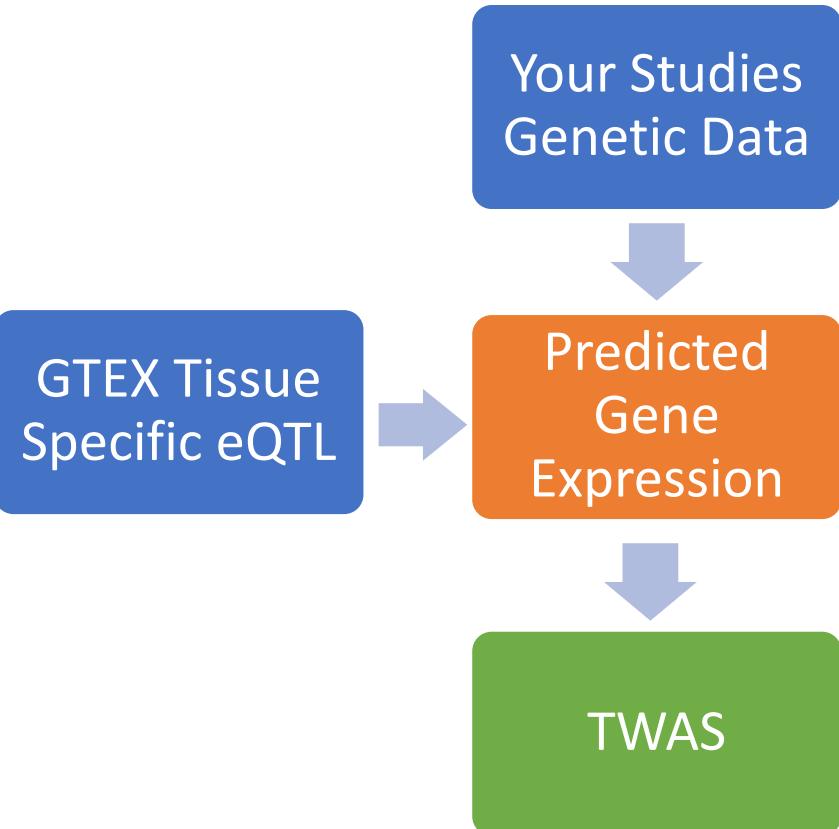
Tools

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

PrediXcan Overview



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