



MSIB32500 Advanced Bioinformatics Fall 2018

RNAseq Data Analysis and Clinical Applications, Part III

Riyue Bao, Ph.D.

Research Assistant Professor (Bioinformatics)

Center for Research Informatics & Department of Pediatrics

The University of Chicago



Outline

- Part I (11/24/2018)
 - Introduction to RNAseq technology and clinical applications
 - Hands on: From raw data to gene expression quantification
- Part II (12/01/2018)
 - Differential gene expression analysis and data visualization
 - Hands on: Identification of genes and pathways significantly changed under condition
 - ***Homework assignment***
- Part III (12/08/2018)
 - How to associate gene expression data with clinical outcome
 - Hands on: Use gene expression data to discover tumor subtypes and survival analysis



Evaluation forms

- ▶ You will receive a link to the end-of quarter evaluation end of today from Graham School.
- ▶ Feedbacks will be appreciated!



Class materials

- GitHub

- <https://github.com/MScBiomedicalInformatics/MSIB32500>

- This lecture note contains the same contents as the notebook. In addition, the notebook also contains hands-on materials

- **lecture10.pdf**

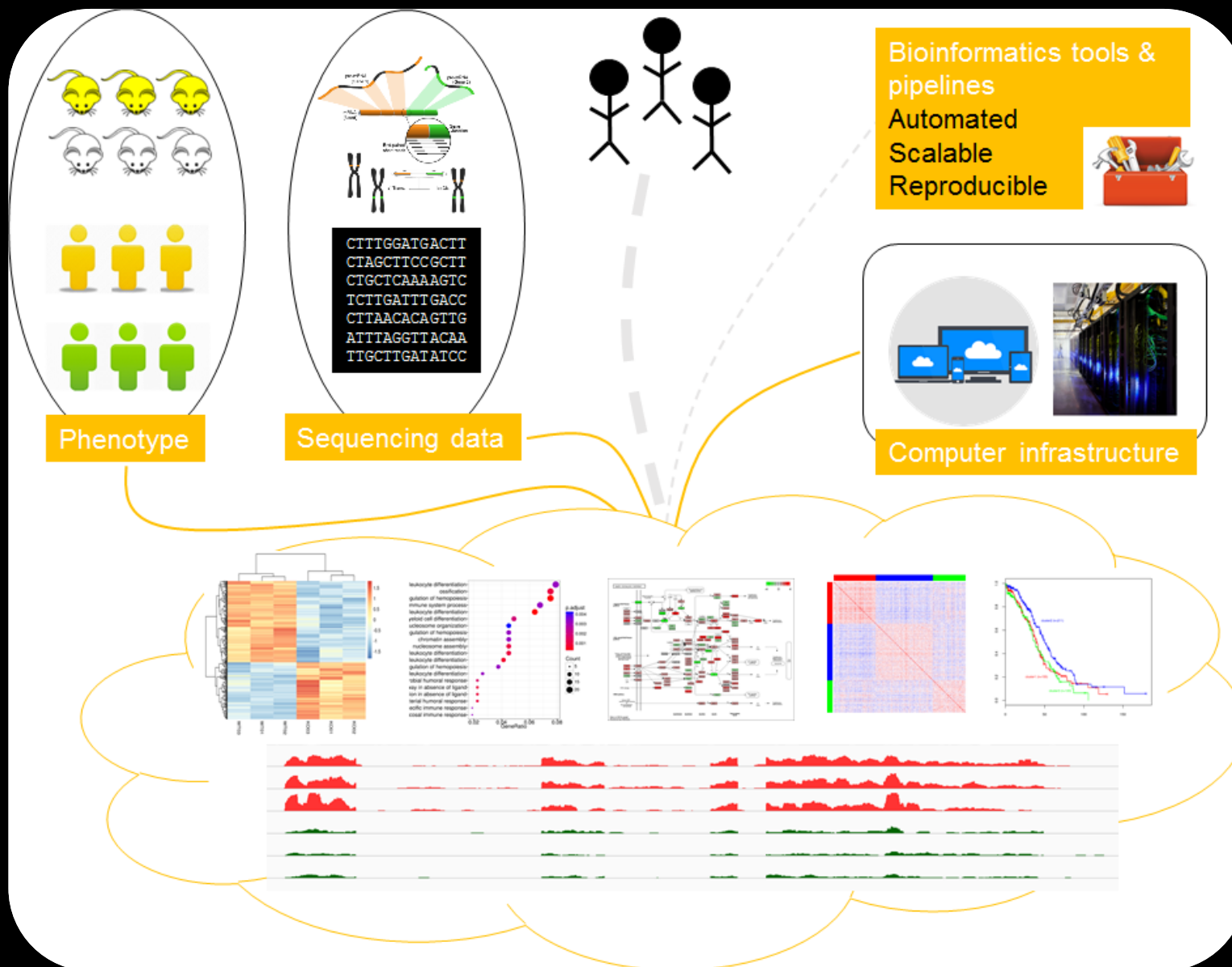
- **Handson10.Rmd**

- Rstudio (or R console) on personal computers (hands on practice)

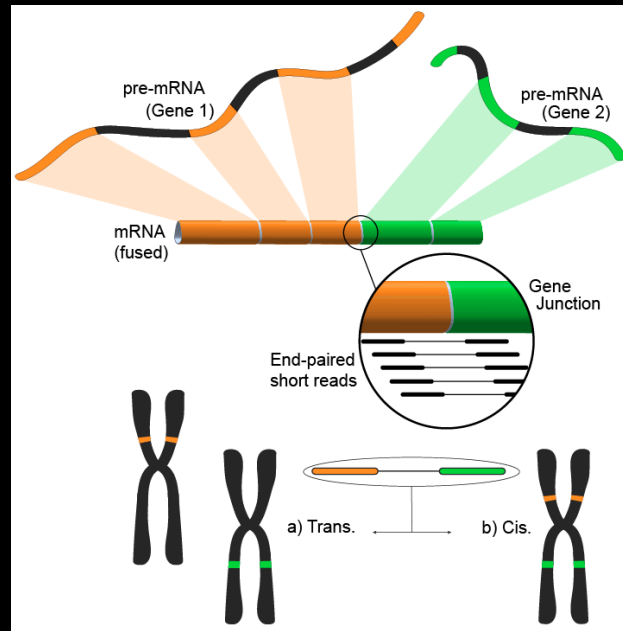


Objective

- ▶ Learn the background and application of The Cancer Genome Atlas (TCGA)
- ▶ Learn the structure and access of Genomics Data Commons (GDC)
- ▶ Explore datasets hosted on GDC
- ▶ Practice how to associate gene expression with clinical data
 - ▶ Use gene expression to identify tumor subtype
 - ▶ Detect survival difference between subtypes
 - ▶ Produce high-quality plots for publication



Background

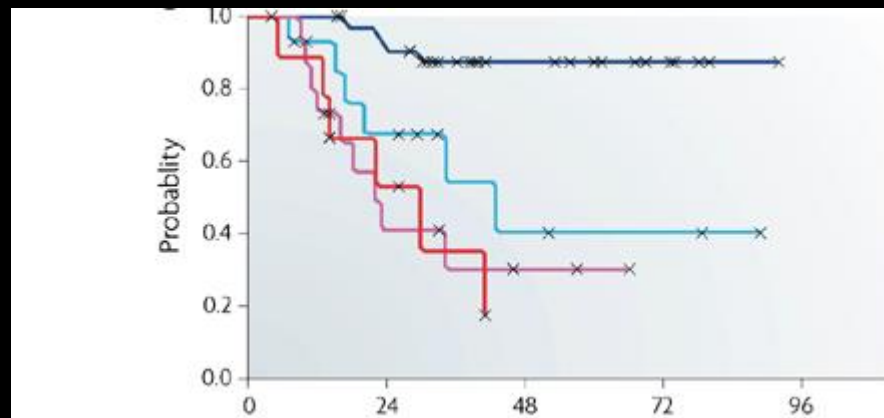
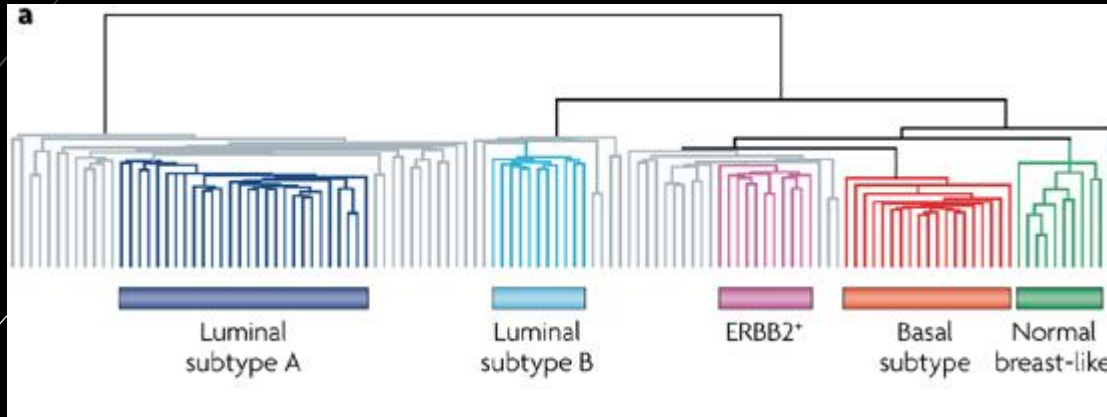


Gene Expression



Patient's clinical data

Background



- Correlate gene expression with clinical data (tumor stage, tumor grade, time to death, time to relapse, etc.)
- Identify tumor subtypes through sample clustering
- Detect survival difference between tumor subtypes
- Discover gene signatures to predict patient classes

Sample Data

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

A catalogue of molecular aberrations that cause ovarian cancer is critical for developing and deploying therapies that will improve patients' lives. The Cancer Genome Atlas project has analysed messenger RNA expression, microRNA expression, promoter methylation and DNA copy number in 489 high-grade serous ovarian adenocarcinomas and the DNA sequences of exons from coding genes in 316 of these tumours. Here we report that high-grade serous ovarian cancer is characterized by *TP53* mutations in almost all tumours (96%); low prevalence but statistically recurrent somatic mutations in nine further genes including *NF1*, *BRCA1*, *BRCA2*, *RBI* and *CDK12*; 113 significant focal DNA copy number aberrations; and promoter methylation events involving 168 genes. Analyses delineated four ovarian cancer transcriptional subtypes, three microRNA subtypes, four promoter methylation subtypes and a transcriptional signature associated with survival duration, and shed new light on the impact that tumours with *BRCA1/2* (*BRCA1* or *BRCA2*) and *CCNE1* aberrations have on survival. Pathway analyses suggested that homologous recombination is defective in about half of the tumours analysed, and that NOTCH and FOXM1 signalling are involved in serous ovarian cancer pathophysiology.

Sample Data

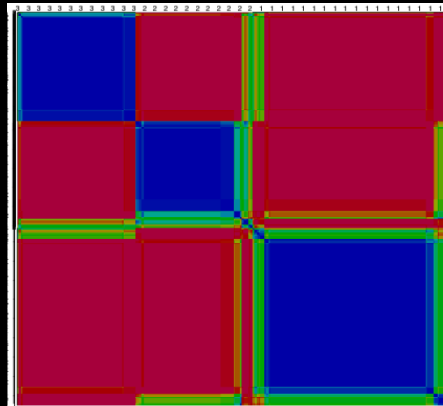
ARTICLE

doi:10.1038/nature10166

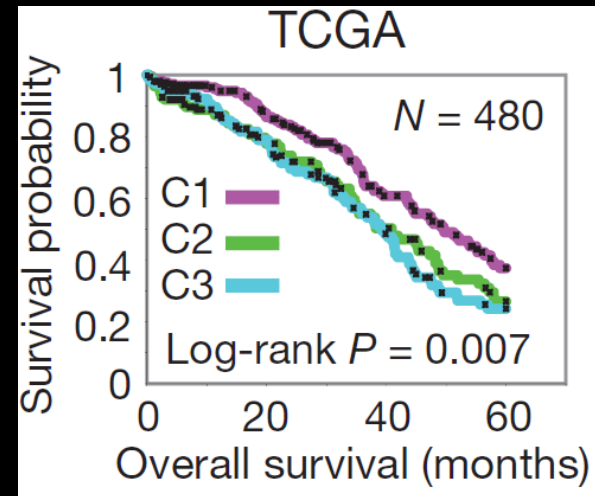
Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

150 most variable
miRNAs for
sample clustering



Cluster 1 has significantly
better survival



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

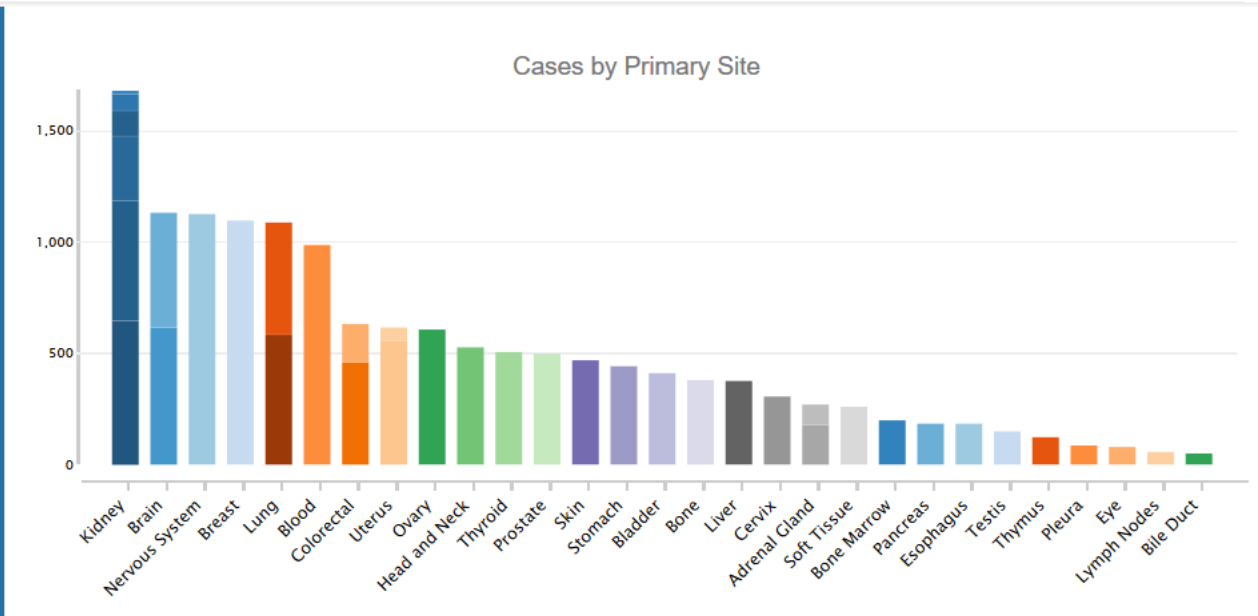
Get Started by Exploring:

Projects

Data

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below	736 Cases	1,519 Files
CNV data of female brain cancer cases	459 Cases	1,788 Files
Gene expression quantification data in TCGA-GBM project	166 Cases	522 Files



DATA PORTAL SUMMARY

[Data Release 6.0 - May 9, 2017](#)

PROJECTS

39

PRIMARY SITE

29

CASES

14,551

FILES

274,724

Infrastructure

Data is continuously being processed and harmonized by the GDC.
[View GDC system statistics:](#)

Compute Infrastructure	12,800 Cores	87.96 TB RAM
Storage Infrastructure	4.08 PB Used	5.42 PB Total

Documentation

Learn how to use the GDC Data Portal to its full potential with common topics such as:

- [Browse Data using Facet Search](#)
- [Search Data with Advanced Search Technology](#)

GDC Applications


The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

Cases Files

« Hide Filters


[Add a Case/Biospecimen Filter](#)

Case



Search for Case Id

Case Submitter ID Prefix



Search for Submitter Id

Primary Site

☐ Kidney

☐ Brain

☐ Nervous System

☐ Breast

☐ Lung

1,681

1,133

1,127

1,098

1,089

[24 More...](#)

Cancer Program

☐ TCGA

☐ TARGET

11,315

3,236

Project

☐ TARGET-NBL

1,127

Start searching by selecting a facet or try the Advanced Search

Advanced

Summary

Cases (14,551)

Files (274,724)

[Browse Annotations](#)

Add all files to the Cart

Download Manifest

FILES

274,724



CASES

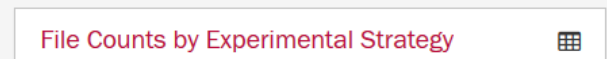
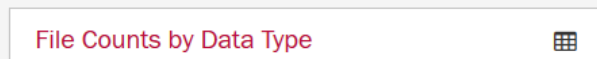
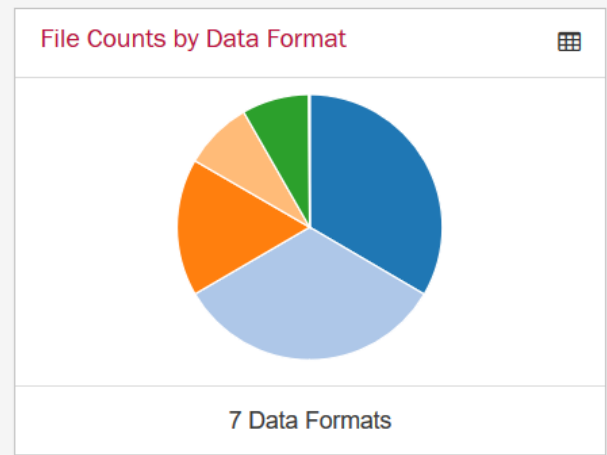
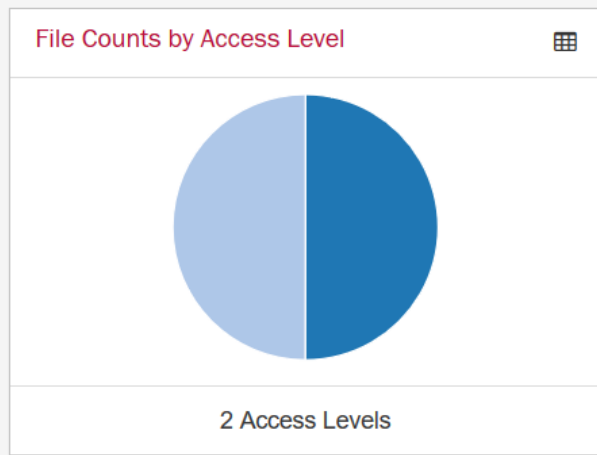
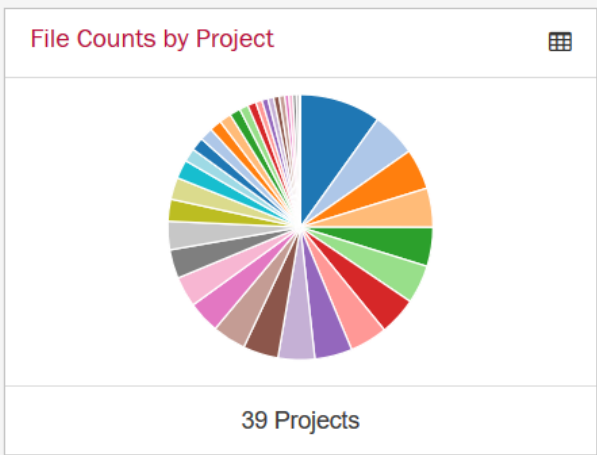
14,551



FILE SIZE

470.58 TB





https://portal.gdc.cancer.gov

The Cancer Genome Atlas (TCGA)
Ovarian cancer

?

Can't find your data? [Click here for more information.](#)

NIH

NATIONAL CANCER INSTITUTE

GDC Data Portal

Home

Projects

Data

Analysis

login

Cart 0

GDC Apps

Cases

Files

« Hide Filters

Add a File Filter

File

Search for File Id

Data Category

☐ Simple Nucleotide Variation

4,880

☐ Copy Number Variation

2,292

☐ Transcriptome Profiling

2,135

☐ Raw Sequencing Data

1,929

☐ DNA Methylation

623

☐ Biospecimen

608

☐ Clinical

587

Less...

Data Type

☐ Annotated Somatic Mutation

2,436

☐ Raw Simple Somatic Mutation

2,436

☐ Aligned Reads

1,929

☐ Copy Number Segment

1,146

☐ Masked Copy Number Segment

1,146

8 More...

Clear

Program Name

IS

TCGA

AND

Project Id

IS

TCGA-OV

Summary

Cases (608)

Files (13,054)

Add all files to the Cart

Download Manifest

FILES

13,054

File Counts by Project

1 Project

CASES

608

File Counts by Access Level

2 Access Levels

File Counts by Primary Site

File Counts by Data Type

File Counts by Experimental Strategy

Data Portal

Website

API

Data Transfer Tool

Documentation

Data Submission Portal

Legacy Archive

GDC cBio Portal

6 Data Formats

ID	Disease Type	Primary Site	Program	Cases
TCGA-BRCA	Breast Invasive Carcinoma	Breast	TCGA	1,098
TCGA-GBM	Glioblastoma Multiforme	Brain	TCGA	617
TCGA-OV	Ovarian Serous Cystadenocarcin...	Ovary	TCGA	608
TCGA-LUAD	Lung Adenocarcinoma	Lung	TCGA	585
TCGA-UCEC	Uterine Corpus Endometrial Carci...	Uterus	TCGA	560
TCGA-KIRC	Kidney Renal Clear Cell Carcinoma	Kidney	TCGA	537
TCGA-HNSC	Head and Neck Squamous Cell C...	Head and Neck	TCGA	528
TCGA-LGG	Brain Lower Grade Glioma	Brain	TCGA	516
TCGA-THCA	Thyroid Carcinoma	Thyroid	TCGA	507
TCGA-LUSC	Lung Squamous Cell Carcinoma	Lung	TCGA	504
TCGA-PRAD	Prostate Adenocarcinoma	Prostate	TCGA	500
TCGA-SKCM	Skin Cutaneous Melanoma	Skin	TCGA	470
TCGA-COAD	Colon Adenocarcinoma	Colorectal	TCGA	461
TCGA-STAD	Stomach Adenocarcinoma	Stomach	TCGA	443
TCGA-BLCA	Bladder Urothelial Carcinoma	Bladder	TCGA	412
TCGA-LIHC	Liver Hepatocellular Carcinoma	Liver	TCGA	377
TCGA-CESC	Cervical Squamous Cell Carcino...	Cervix	TCGA	307
TCGA-KIRP	Kidney Renal Papillary Cell Carci...	Kidney	TCGA	291
TCGA-SARC	Sarcoma	Soft Tissue	TCGA	261
TCGA-LAML	Acute Myeloid Leukemia	Bone Marrow	TCGA	200
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas	TCGA	185
TCGA-ESCA	Esophageal Carcinoma	Esophagus	TCGA	185
TCGA-PCPG	Pheochromocytoma and Paragan...	Adrenal Gland	TCGA	179
TCGA-READ	Rectum Adenocarcinoma	Colorectal	TCGA	172
TCGA-TGCT	Testicular Germ Cell Tumors	Testis	TCGA	150
TCGA-THYM	Thymoma	Thymus	TCGA	124
TCGA-KICH	Kidney Chromophobe	Kidney	TCGA	113
TCGA-ACC	Adrenocortical Carcinoma	Adrenal Gland	TCGA	92
TCGA-MESO	Mesothelioma	Pleura	TCGA	87
TCGA-UVM	Uveal Melanoma	Eye	TCGA	80
TCGA-DLBC	Lymphoid Neoplasm Diffuse Larg...	Lymph Nodes	TCGA	58
TCGA-UCS	Uterine Carcinosarcoma	Uterus	TCGA	57
TCGA-CHOL	Cholangiocarcinoma	Bile Duct	TCGA	51
Total				11,315

TCGA cancer types (n=33)

- TCGA raw data were harmonized by NCI's GDC team
 - Release 6 June 2017
 - Release 7 available soon
- Data types: from raw files to compiled results
- Result access: public or protected
- Apply for access: dbGap
- Download: GDC

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

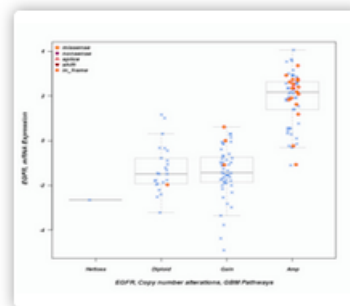
New modern-looking GDC visualization website available soon!!

The GDC cBioPortal is being retired. In its place, the GDC will soon release new Data Analysis, Visualization, and Exploration (DAVE) Tools that will be integrated into the existing GDC Data Portal.

During this transition the GDC cBioPortal will no longer be updated. Starting with **Data Release 6.0** the content of the GDC cBioPortal will not reflect the updated MAF files found in the GDC Data Portal. For the most up-to-date mutation information, please refer to data found in the [GDC Data Portal](#).

The GDC cBioPortal is an implementation of cBioPortal that supports the visualization of mutation data. For information on cBioPortal, please refer to cBioPortal.org.

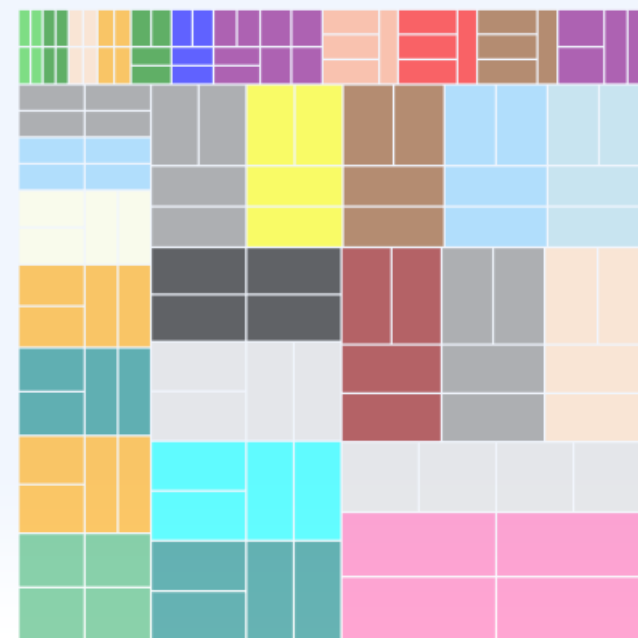
DISCLAIMER: This application has been reviewed for compliance with [Section 508 accessibility standards](#) and we know that there are accessibility issues. If you experience any accessibility issues when using this application, please contact the GDC Help Desk (support@nci-gdc.datacommons.io) for assistance.



What's New

Data Sets

The Portal contains **132 cancer studies**. [\[Details\]](#)



Query Download Data

Select Cancer Study:

Search...

Access GDC data

```
lrbao@cr116in002 ~]$ gdc-client -h
usage: gdc-client [-h] [--version] {download,upload,interactive} ...

The Genomic Data Commons Command Line Client

optional arguments:
  -h, --help            show this help message and exit
  --version             show program's version number and exit

commands:
  {download,upload,interactive}
                        for more information, specify -h after a command
  download              download data from the GDC
  upload               upload data to the GDC
  interactive          run in interactive mode
```

Bioconductor packages

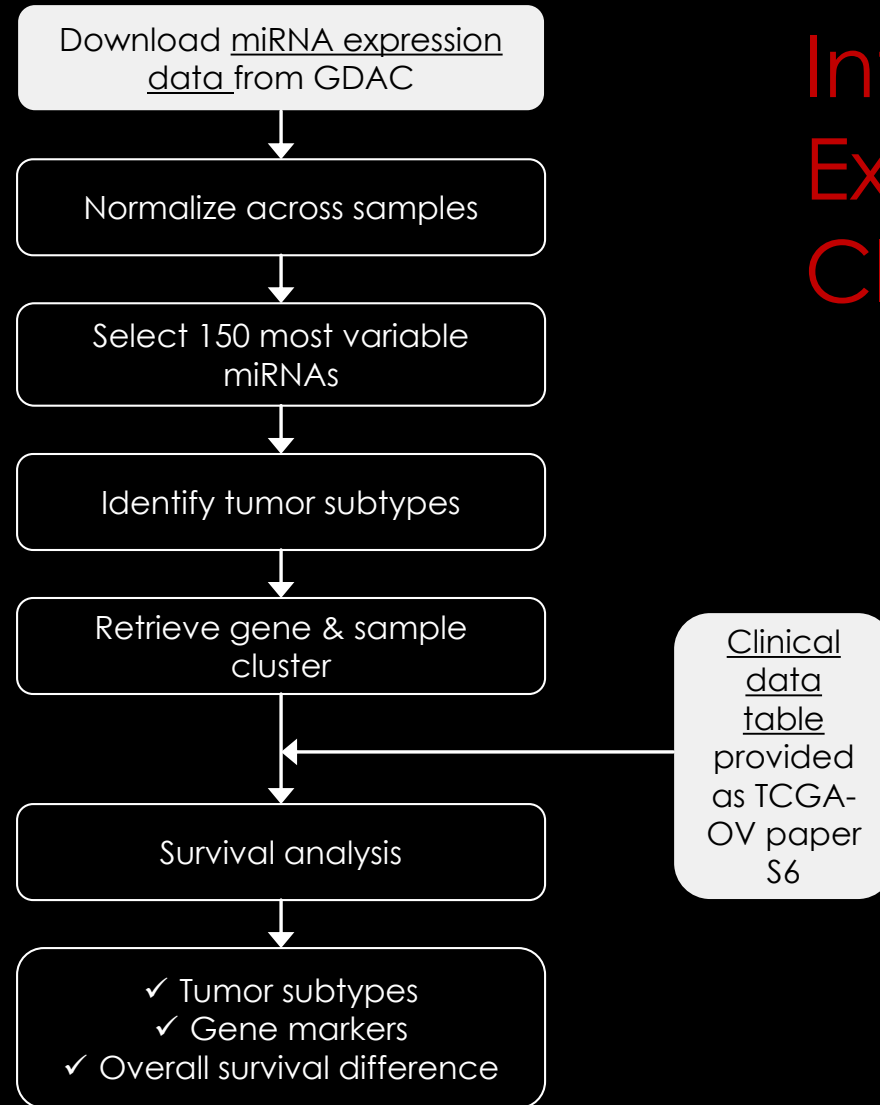
- library(TCGAbiolinks)
 - <https://goo.gl/ytPe07>
- library(GenomicDataCommons)
 - <https://goo.gl/cbUfp3>

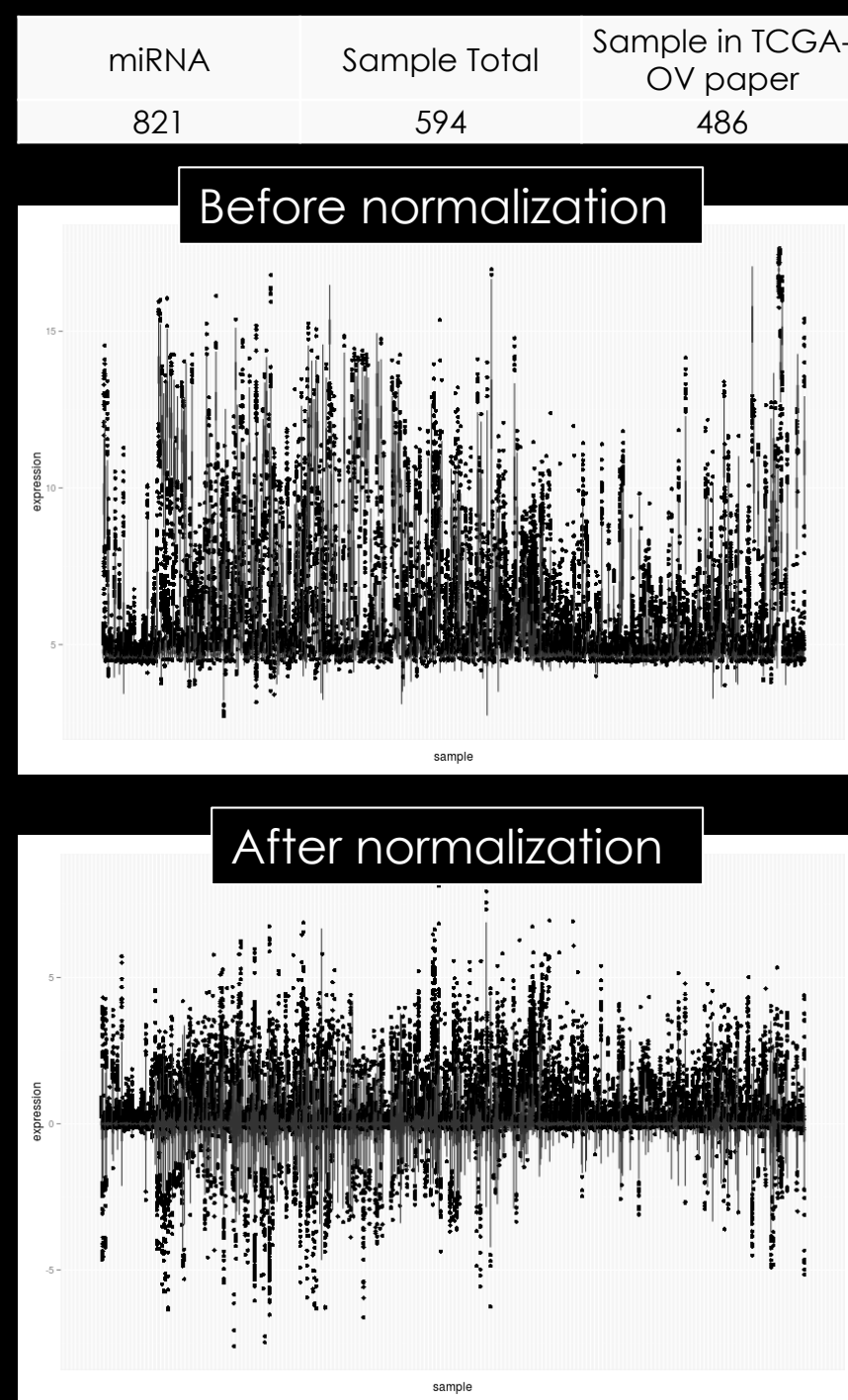
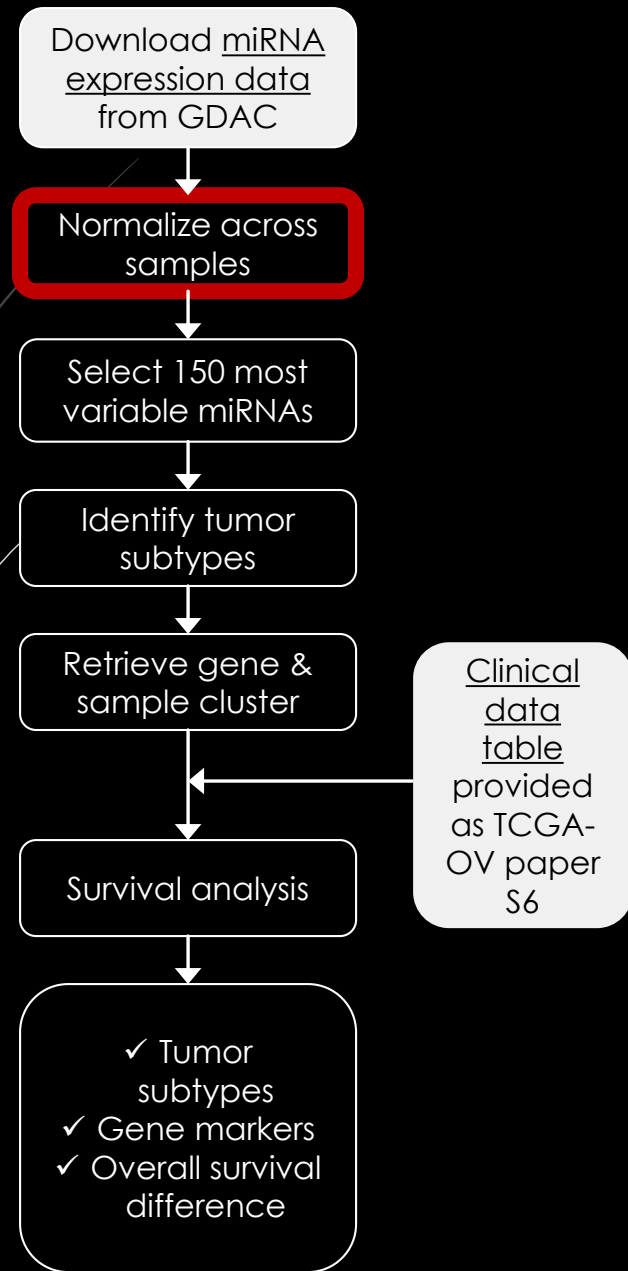
The GDC Application Programming Interface (API): An Overview

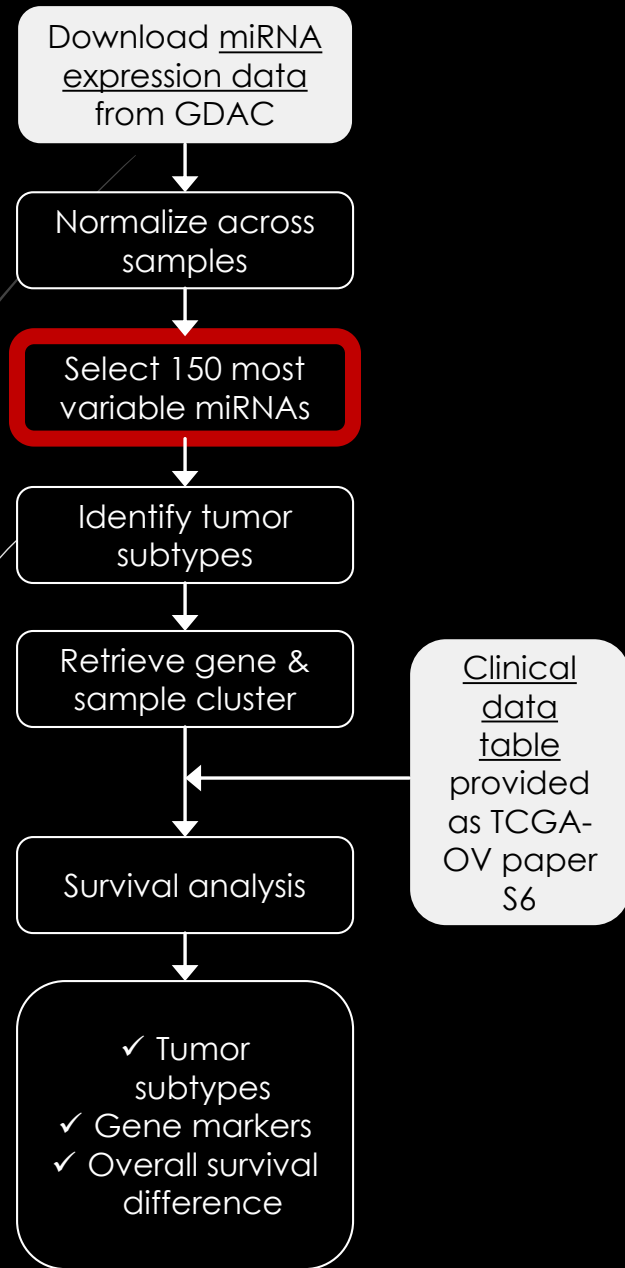
The GDC API drives the GDC Data and Submission Portals and provides programmatic access to GDC functionality. This includes searching for, downloading, and submitting data and metadata. The GDC API uses JSON as its communication format, and standard HTTP methods like `GET`, `PUT`, `POST` and `DELETE`.

```
curl https://api.gdc.cancer.gov/files/d853e541-f16a-4345-9f00-88e03c2dc0bc?pretty=true
```


Integrate Expression with Clinical Data

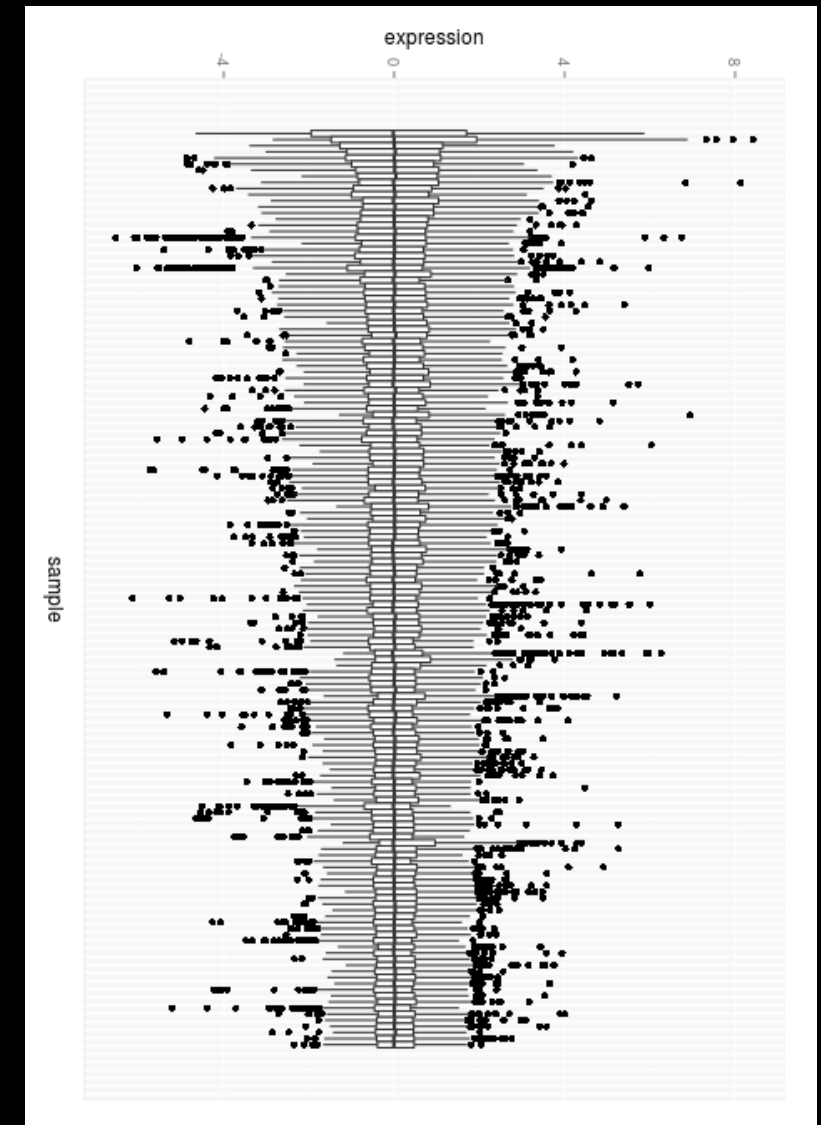






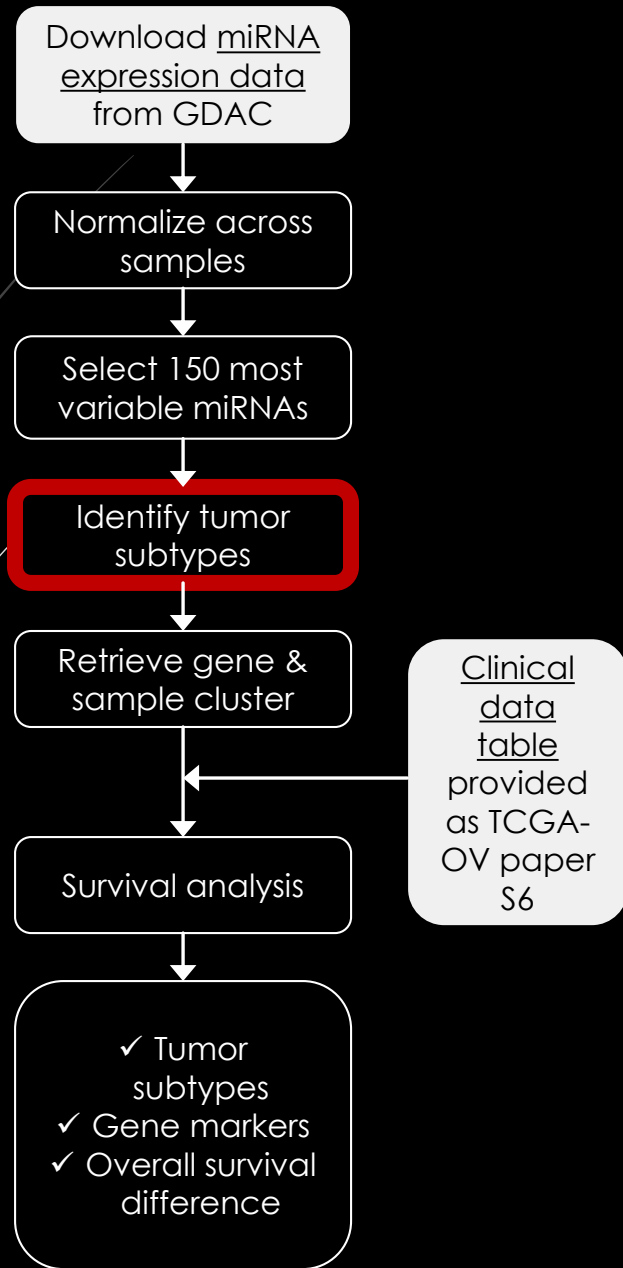
	mad
HSA-MIR-205	2.6824804
HSA-MIR-449A	2.3890068
HSA-MIR-31	1.8097683
HSA-MIR-224	1.6650244
HSA-MIR-451	1.6289215
HSA-MIR-10A	1.4811346
HSA-MIR-10B	1.4523870
HSA-MIR-31*	1.4370410
HSA-MIR-363	1.3751314
HSA-MIR-96	1.3709655
HSA-MIR-203	1.3489193
HSA-MIR-494	1.2932659

Showing 1 to 13 of 150 entries



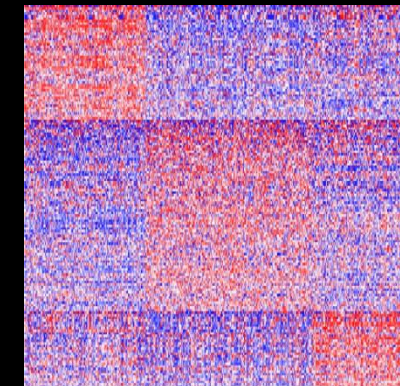
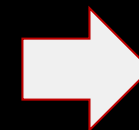
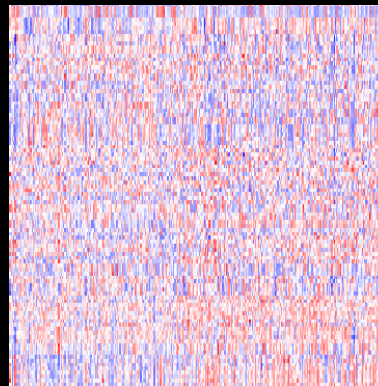
Median absolute deviation (MAD)

$$\text{MAD} = \text{median}_i (|X_i - \text{median}_j(X_j)|)$$



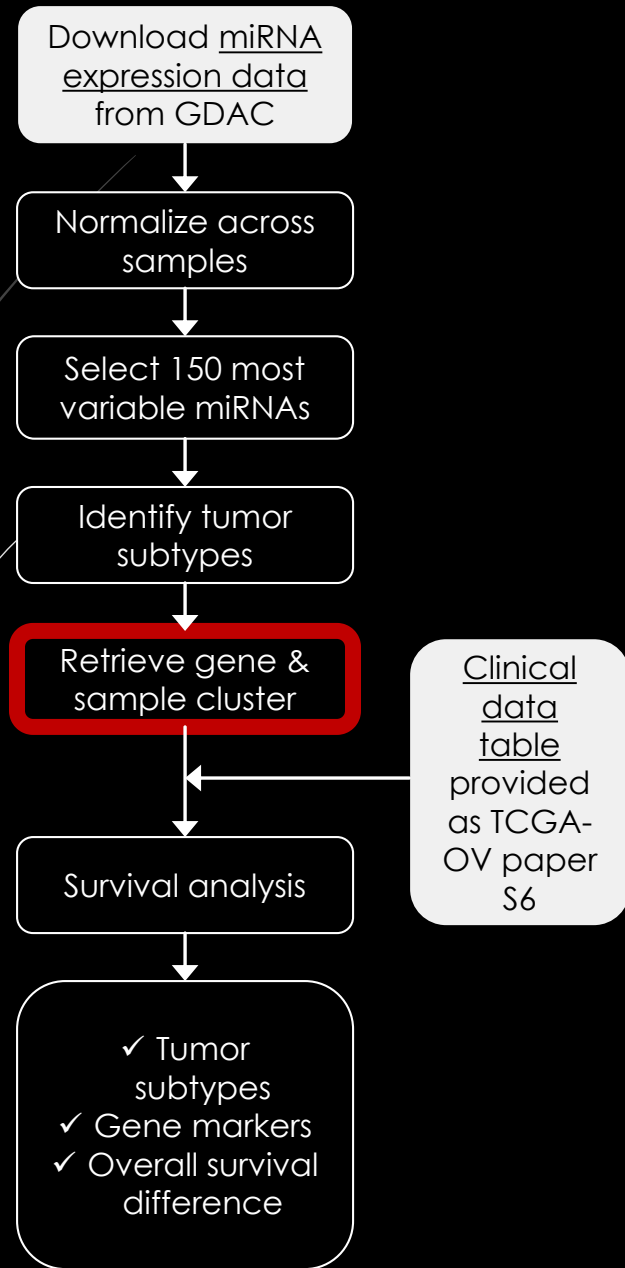
Non-negative matrix factorization (NMF), also non-negative matrix approximation is a group of algorithms in multivariate analysis and linear algebra where a matrix V is factorized into (usually) two matrices W and H , with the property that all three matrices have no negative elements.

$$\begin{bmatrix} W \end{bmatrix} \times \begin{bmatrix} H \end{bmatrix} \approx \begin{bmatrix} V \end{bmatrix}$$



Gene

Samples

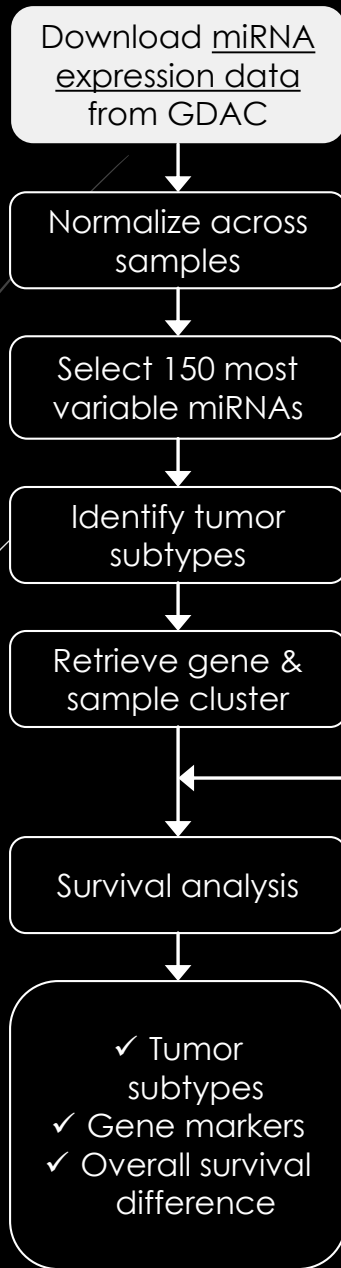


sample	cluster
TCGA.04.1331	1
TCGA.04.1338	1
TCGA.04.1341	1
TCGA.04.1343	1
TCGA.04.1348	1
TCGA.04.1362	1
TCGA.04.1365	1
TCGA.04.1530	1
TCGA.04.1542	1
TCGA.04.1648	1
TCGA.04.1649	1
TCGA.04.1651	1

Sample clusters

gene	cluster
HSA-MIR-205	1
HSA-MIR-494	1
HSA-MIR-144	1
HSA-MIR-142-5P	1
HSA-MIR-181A	1
HSA-MIR-151-3P	1
HSA-MIR-1225-5P	1
HSA-MIR-222	1
HSA-MIR-638	1
EBV-MIR-BART19-3P	1
HSA-MIR-21*	1
HSA-MIR-630	1

Gene clusters

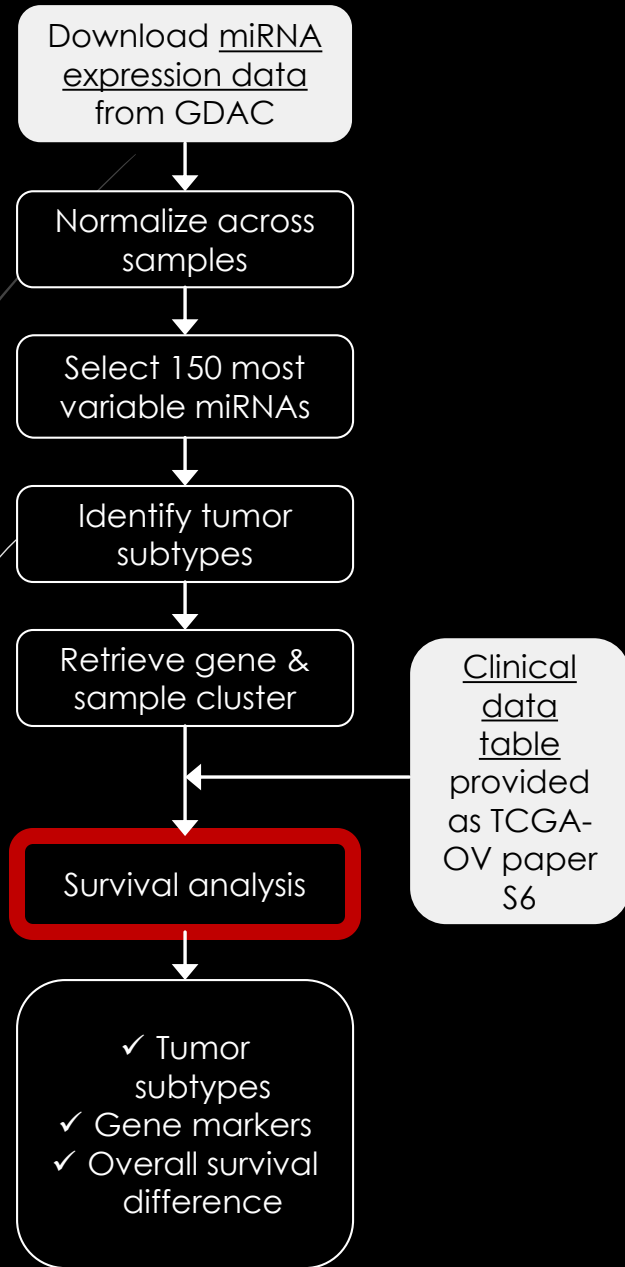


Clinical data table
provided as TCGA-OV paper S6

	sample	vital.status	overall.survival.month	age.at.diagnosis.year	tumor.stage	tumor.grade
1	TCGA.04.1331	DECEASED	43.80	79.04	IIIC	G3
5	TCGA.04.1338	LIVING	46.49	78.87	IIIC	G3
6	TCGA.04.1341	LIVING	NA	85.52	NA	G3
8	TCGA.04.1343	DECEASED	11.84	72.41	IV	G3
11	TCGA.04.1348	DECEASED	48.62	44.48	IIIB	G3
17	TCGA.04.1362	DECEASED	44.20	59.58	IIC	G3
19	TCGA.04.1365	LIVING	76.33	87.47	IIIB	G3
25	TCGA.04.1530	DECEASED	118.75	68.53	IIIC	G3
26	TCGA.04.1542	DECEASED	83.97	52.78	IIIB	G2
29	TCGA.04.1648	DECEASED	28.56	57.84	IIIC	G2
30	TCGA.04.1649	LIVING	64.46	74.42	IIIC	G3
31	TCGA.04.1651	DECEASED	36.07	53.78	IIIC	G3

Showing 1 to 13 of 486 entries

- Sample ID
- Vital status
- Overall survival
- Age at diagnosis
- Tumor stage
- Tumor grade
- **Sample cluster** (from the previous step)



► Survival analysis

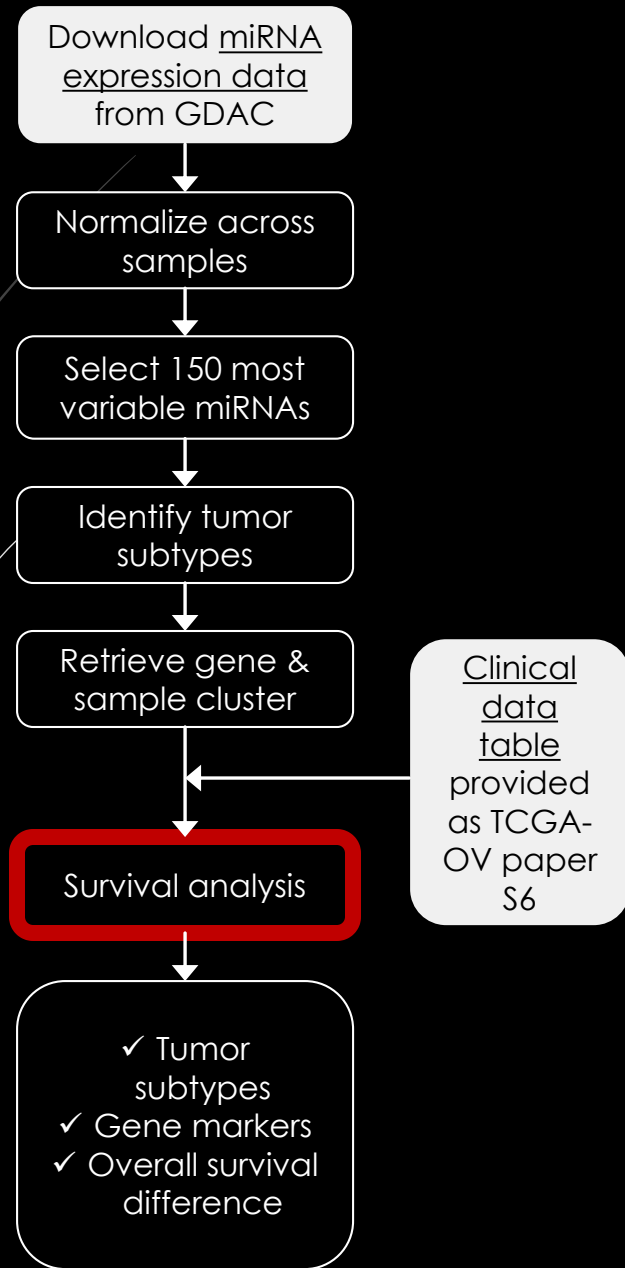
- Study the time between entry to a study or an event (such as death)
- Calculate survival/risk difference and detect significance between groups
- Build models to predict prognosis

► Survival data

- Time to event (in year, month, etc.)
- Status (whether the event has happened?)
 - Censoring: only some individuals have experienced the event by the last follow up, while for others, the time is unknown
- “cumulative” survival time

► Survival methods

- Kaplan-Meier estimator
- Log-rank test (Mantel-Haenzel test)
- Cox regression model (proportional hazard model)



Right Censoring

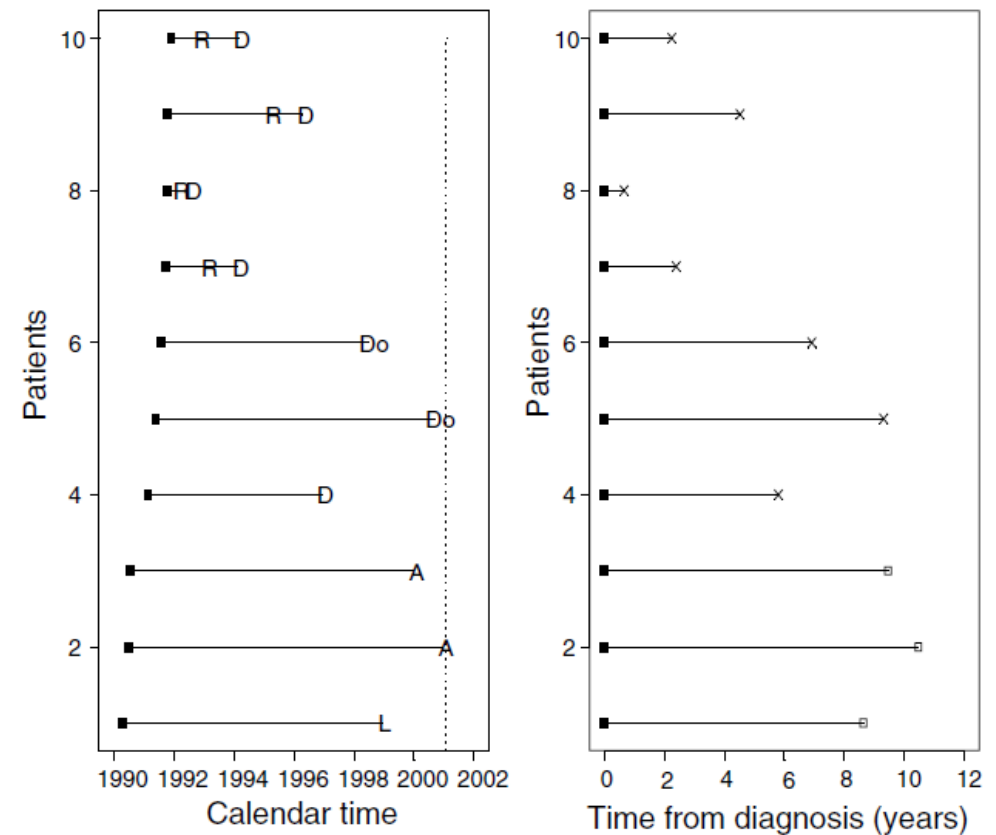
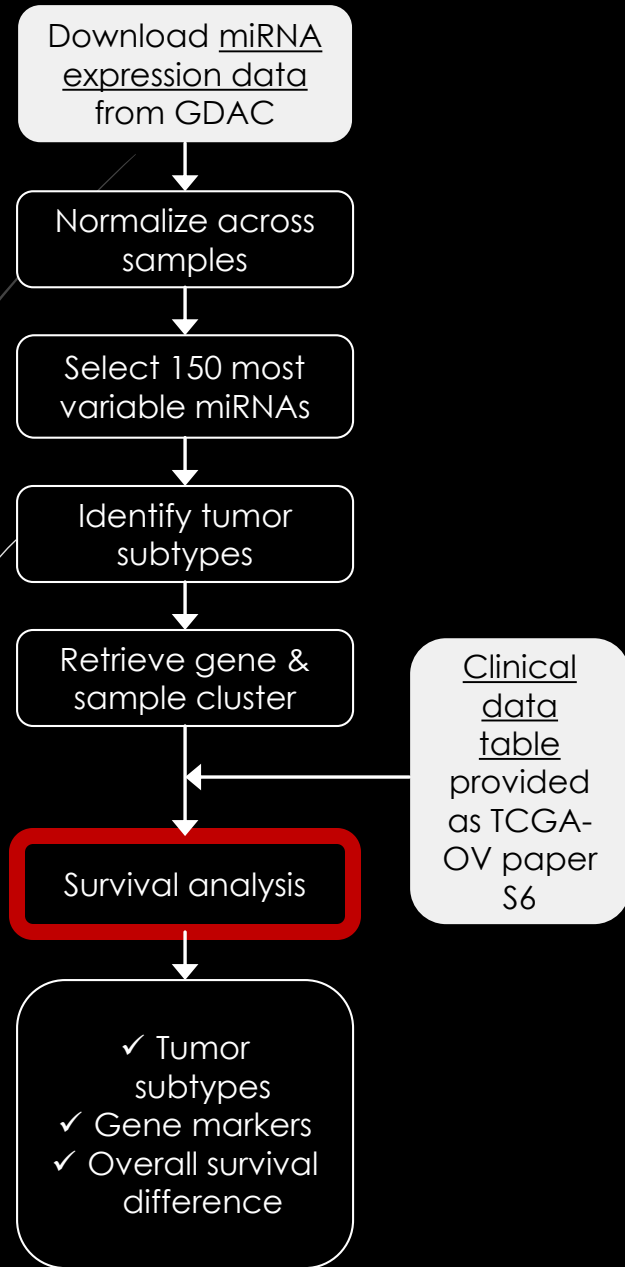
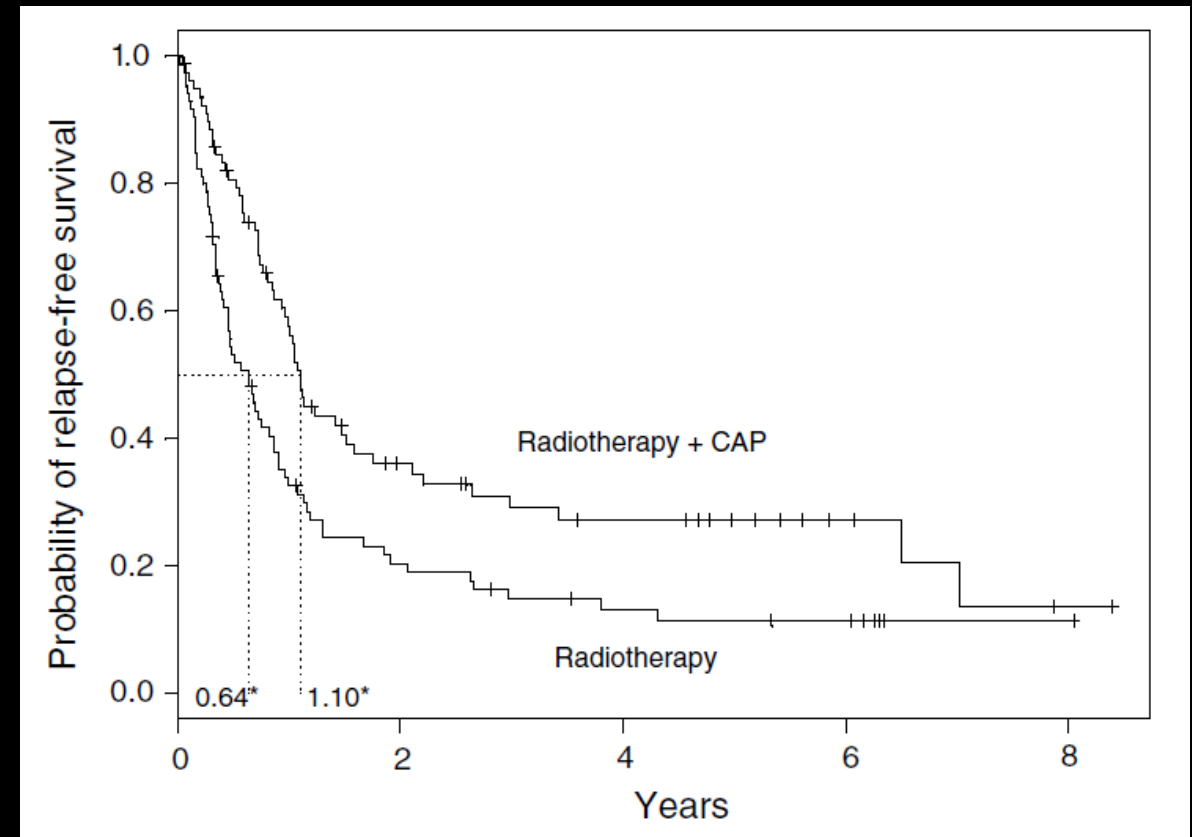
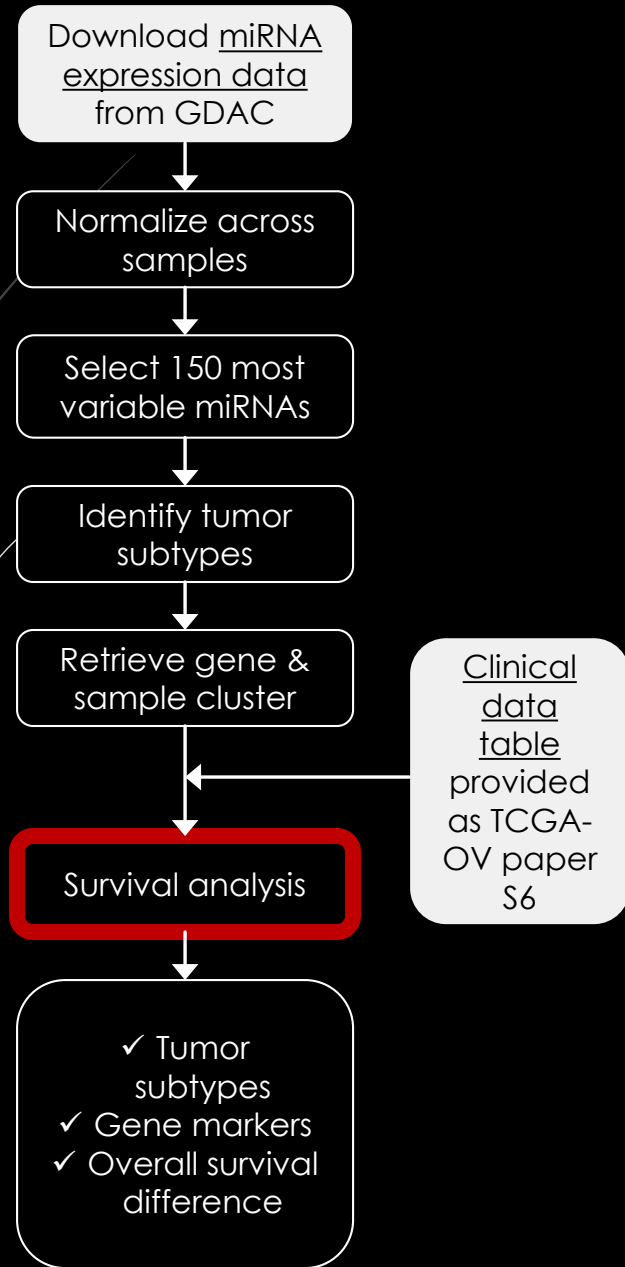


Figure 1 Converting calendar time in the ovarian cancer study to a survival analysis format. Dashed vertical line is the date of the last follow-up, R = relapse, D = death from ovarian cancer, Do = death from other cause, A = attended last clinic visit (alive), L = loss to follow-up, X = death, □ = censored.



- Kaplan-Meier estimator
 - Stepwise function
 - Does not account for effect of other covariates (univariate test)





- Log-rank test

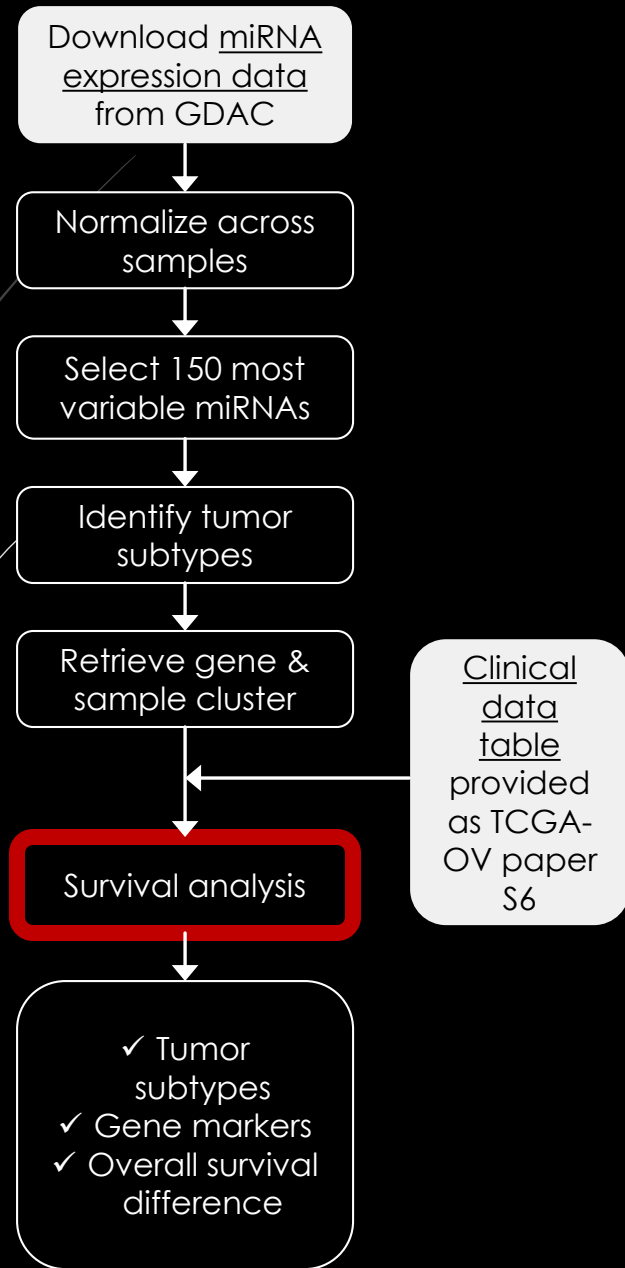
- Chi-square test
- Efficient in comparing groups differed by categorical variables, but not continuous ones
- Univariate test

$$HR = \frac{O_1/E_1}{O_2/E_2}$$

The hazard ratio (HR) is a measure of the relative survival experience in the two groups

O: Observed

E: Expected (if no difference between group 1 and 2)



- Cox Proportional hazard model

- Conveniently access the effect of continuous and categorical variables
- Test the significance of factor of interest adjusting for other factors
- Multivariate test!

```
S ~ sample.cluster +  
    patient.age +  
    tumor.grade
```

```
> summary(m3)  
Call:  
coxph(formula = surv ~ (clinical.sub$cluster + clinical.sub$age.at.diagnosis.year))  
  
n= 558, number of events= 292  
  
              coef exp(coef) se(coef)      z Pr(>|z|)  
clinical.sub$cluster2    0.451874  1.571254 0.145955  3.096 0.001962 **  
clinical.sub$cluster3    0.309421  1.362636 0.138750  2.230 0.025744 *  
clinical.sub$age.at.diagnosis.year 0.019574  1.019766 0.005354  3.656 0.000257 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cox Proportional hazard model

- β : coefficient of explanatory variables or predictors
- $\exp(\beta)$: the ratio of the hazards between two individuals whose values of x differ by one unit when all other covariates are held constant (**hazard ratio**, analogous to an **odds ratio** in the setting of multiple logistic regression analysis)
- Z : Wald statistics calculated by dividing β by its standard error
- P : P-value that corresponds to Z statistics. If $P < 0.05$, then the null hypothesis of β equal to zero can be rejected at 95% confidence level

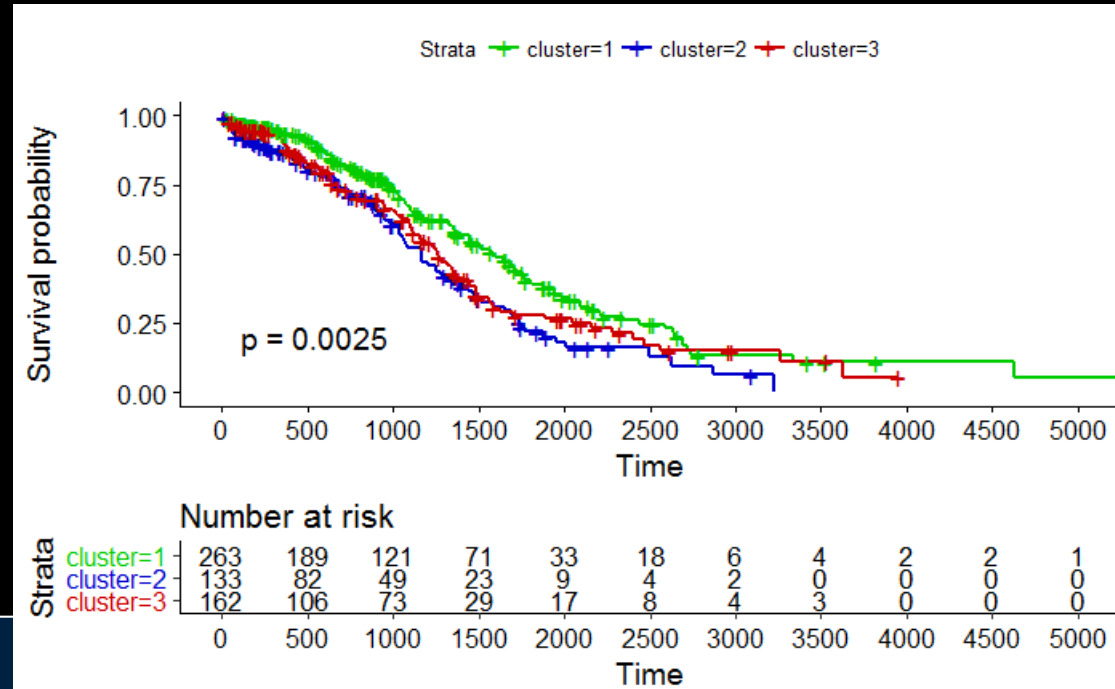
```
> summary(m3)
Call:
coxph(formula = surv ~ (clinical.sub$cluster + clinical.sub$age.at.diagnosis.year))

n= 558, number of events= 292
```

	coef	exp(coef)	se(coef)	z	Pr(> z)	
clinical.sub\$cluster2	0.451874	1.571254	0.145955	3.096	0.001962	**
clinical.sub\$cluster3	0.309421	1.362636	0.138750	2.230	0.025744	*
clinical.sub\$age.at.diagnosis.year	0.019574	1.019766	0.005354	3.656	0.000257	***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cox Proportional hazard model



```
> summary(m3)
```

```
Call:
```

```
coxph(formula = surv ~ (clinical.sub$cluster + clinical.sub$age.at.diagnosis.year))
```

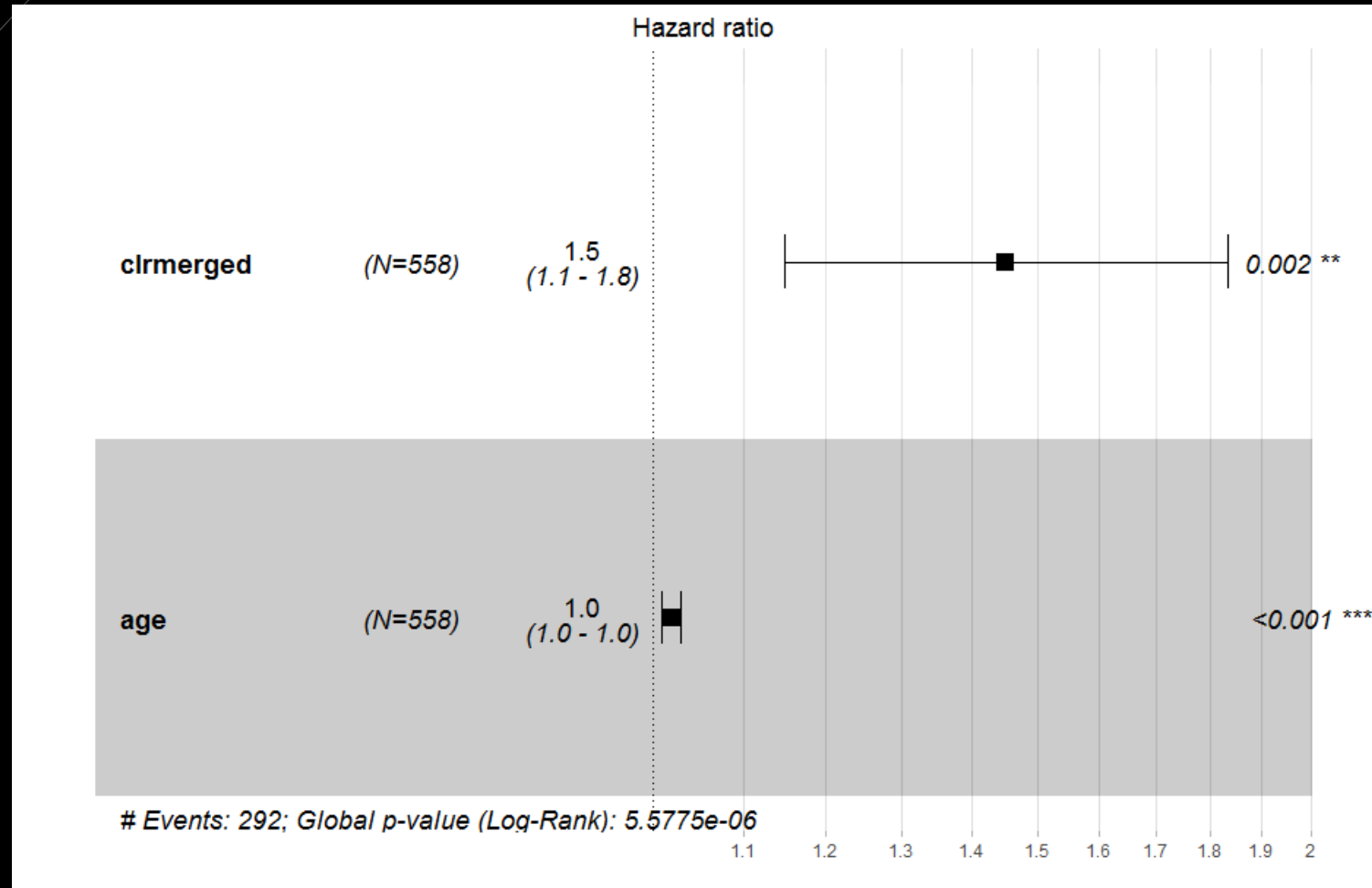
```
n= 558, number of events= 292
```

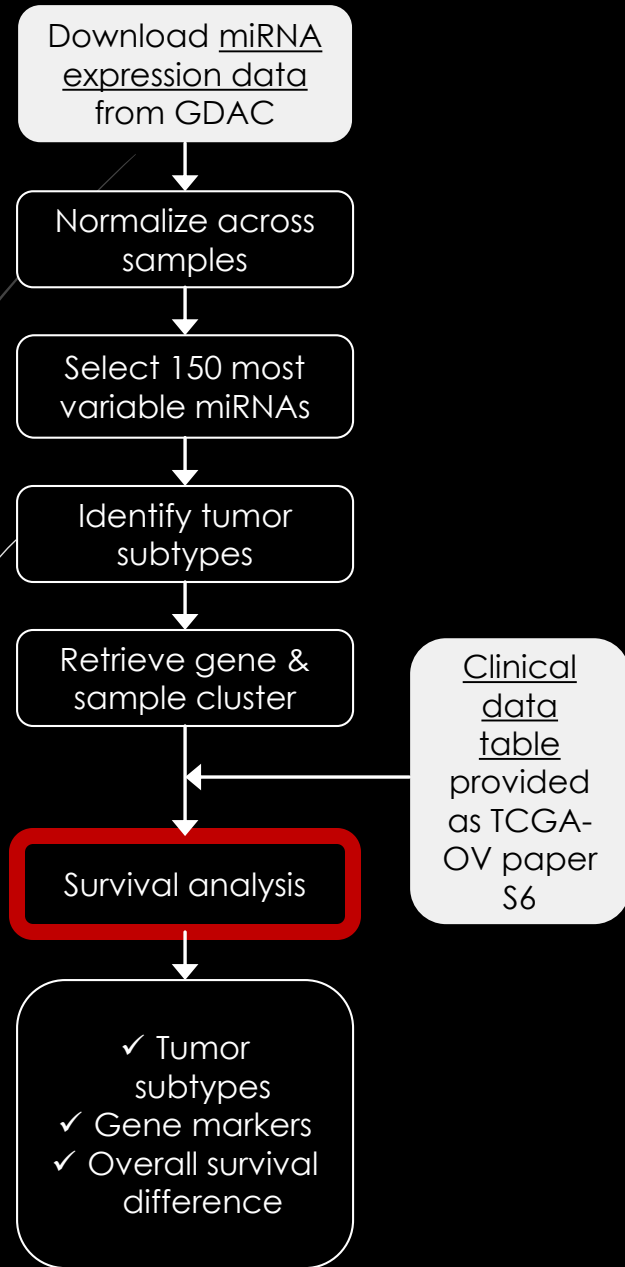
	coef	exp(coef)	se(coef)	z	Pr(> z)	
clinical.sub\$cluster2	0.451874	1.571254	0.145955	3.096	0.001962	**
clinical.sub\$cluster3	0.309421	1.362636	0.138750	2.230	0.025744	*
clinical.sub\$age.at.diagnosis.year	0.019574	1.019766	0.005354	3.656	0.000257	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Visualize Cox PH model: forest plot





► Survival methods

- Kaplan-Meier estimator
- Log-rank test (Mantel-Haenzel test)
- Cox regression model (proportional hazard model)

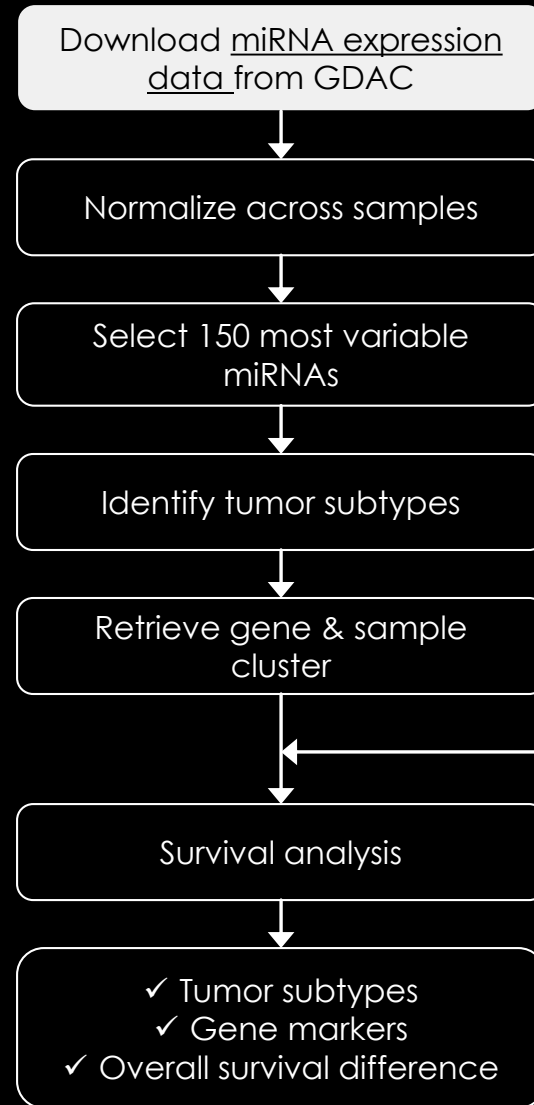
Library (survival)

survfit

survdiff

coxph

Integrate Expression with Clinical Data



lecture10.Rmd

Clinical data table
provided
as TCGA-
OV paper
S6

Integration of multi-omics data

- Multiple types of genomic data, e.g. mRNA, miRNA, methylation

- SNF: similarity network fusion

<http://compbio.cs.toronto.edu/SNF/SNF/Software.html>

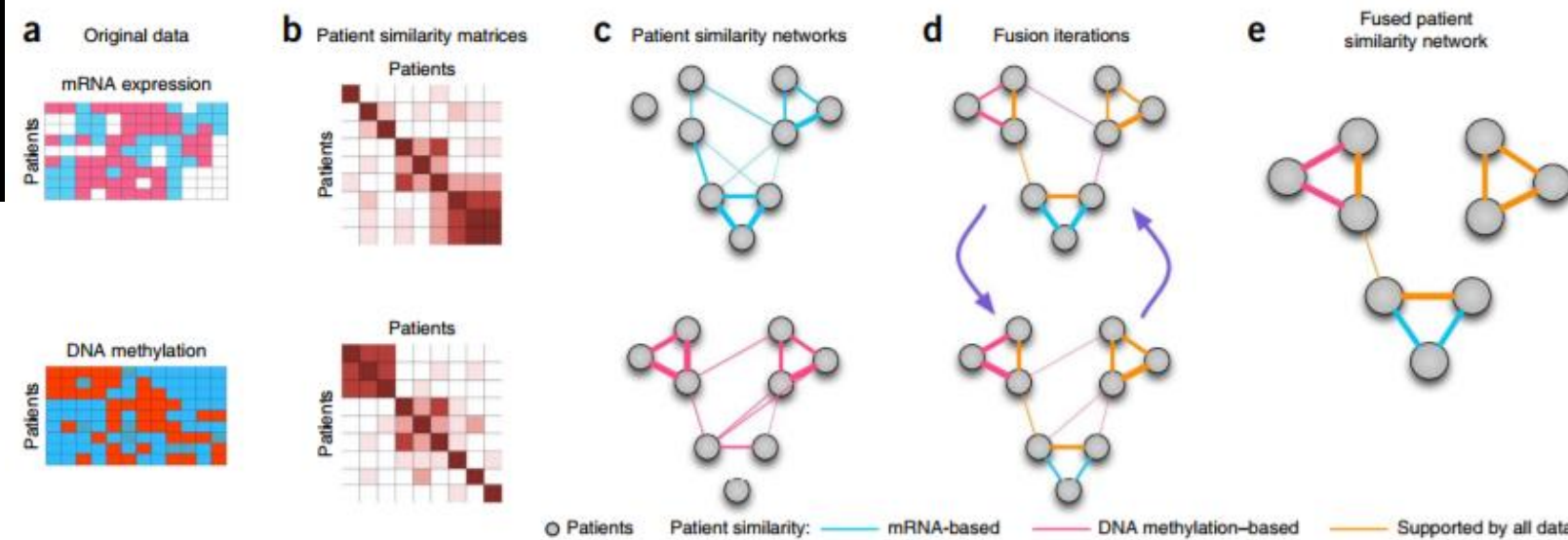
- iClusterPlus: a joint latent variable model for integrative clustering

<https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html>

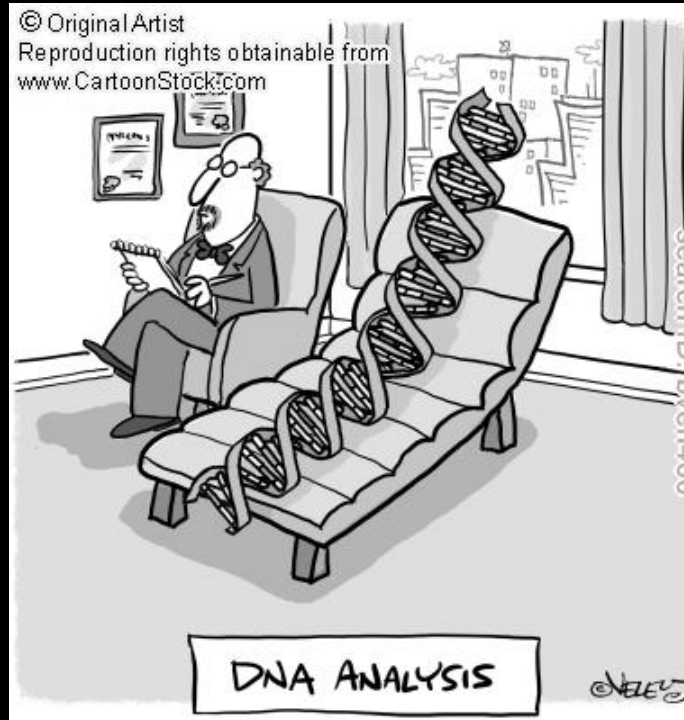
Similarity network fusion for aggregating data types on a genomic scale

Bo Wang^{1,5}, Aziz M Mezlini^{1,2}, Feyyaz Demir^{1,2}, Marc Fiume², Zhuowen Tu³, Michael Brudno^{1,2}, Benjamin Haibe-Kains^{4,5} & Anna Goldenberg^{1,2}

Recent technologies have made it cost-effective to collect diverse types of genome-wide data. Computational methods are needed to combine these data to create a comprehensive view of a given disease or a biological process. Similarity network fusion (SNF) solves this problem by constructing networks of samples (e.g., patients) for each available data type and then efficiently fusing these into one network that represents the full spectrum of underlying data. For example, to create a comprehensive view of a disease given a cohort of patients, SNF computes and fuses patient similarity networks obtained from each of their data types separately, taking advantage of the complementarity in the data. We used SNF to combine mRNA expression, DNA methylation and microRNA (miRNA) expression data for five cancer data sets. SNF substantially outperforms single data type analysis and established integrative approaches when identifying cancer subtypes and is effective for predicting survival.



Thank you!



Questions



