MSIB32500 Advanced Bioinformatics Fall 2017

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 (11/18/2017)
  - Introduction to RNAseq technology and clinical applications
  - Hands on: From raw data to gene expression quantification
- Part II (11/25/2017)
  - Differential gene expression analysis and data visualization
  - Hands on: Identification of genes and pathways significantly changed under condition
  - ***Homework assignment***
  - ***Thanksgiving week – Gleacher center physically closed. Class will be on WebEx.***
- Part III (12/02/2017)
  - How to associate gene expression data with clinical outcome
  - Hands on: Use gene expression data to discover tumor subtypes and survival analysis
  - ***Homework due tomorrow (12/03) 11:59 PM***

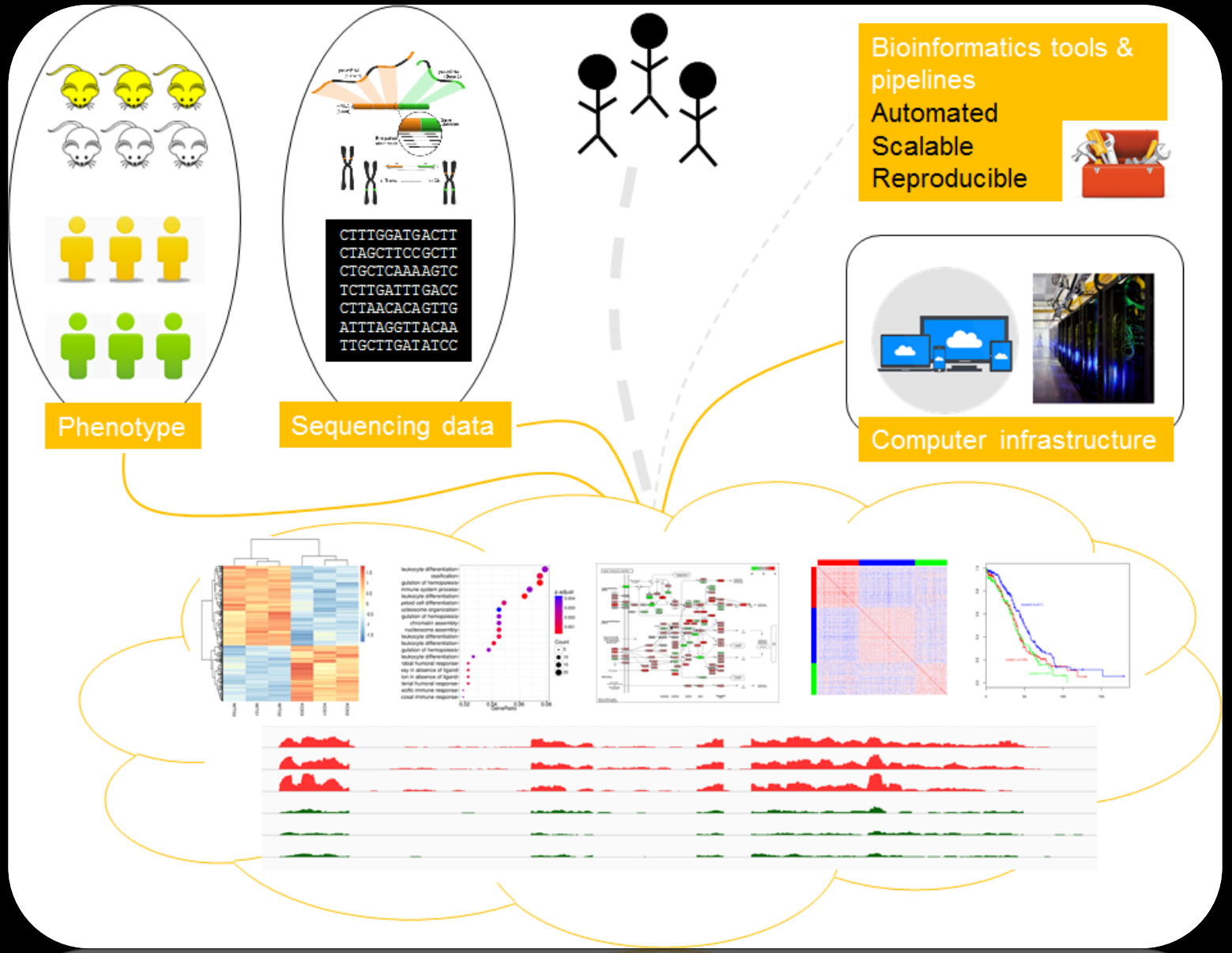
Class materials

- 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**

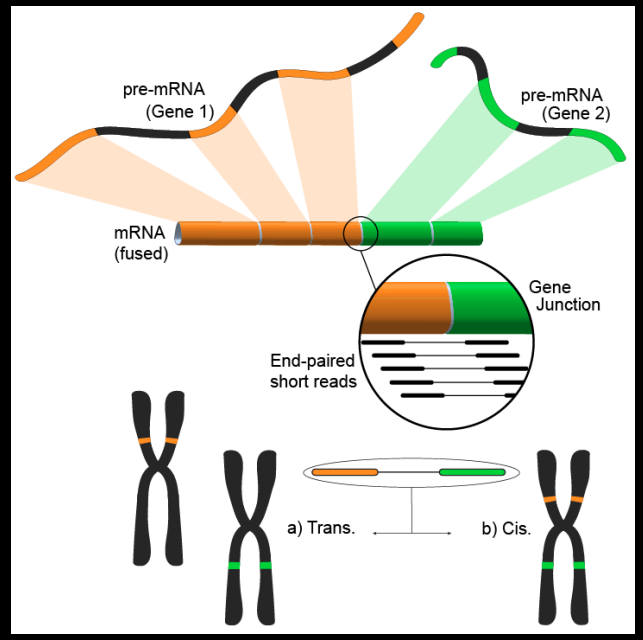
- Rstudio (or R console) on personal computers (hands on practice)

Objective

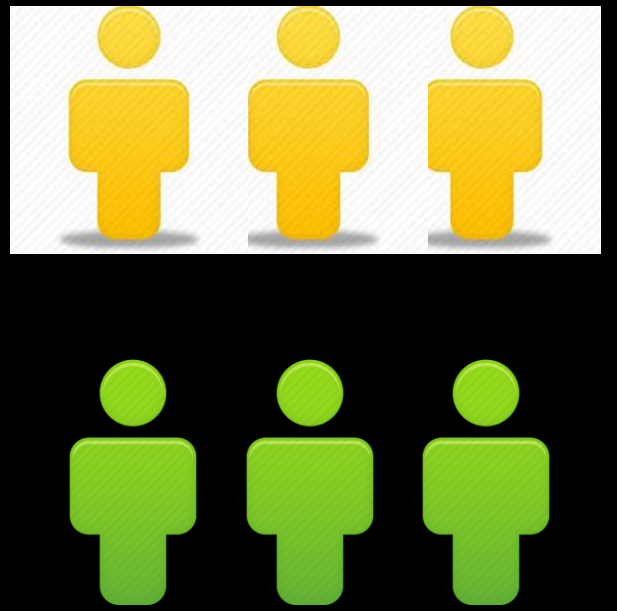
- Learn the background and application of The Cancer Genome Atlas (TCGA)
- Learn the structure and access of Genomics Data Commons (GDC)
- Explore datasets hosted on GDC
- Practice how to associate gene expression with clinical data
  - Use gene expression to identify tumor subtype
  - Detect survival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


Phenotype
CTTTGGATGACTT
CTAGCTTCCGCTT
CTGCTCAAAAGTC
TCTTGATTTGACC
CTTAACACAGTTG
ATTTAGGTTACAA
TTGCTTGATATCC
Sequencing data
Bioinformatics tools & pipelines
Automated
Scalable
Reproducible
Computer infrastructure

Background

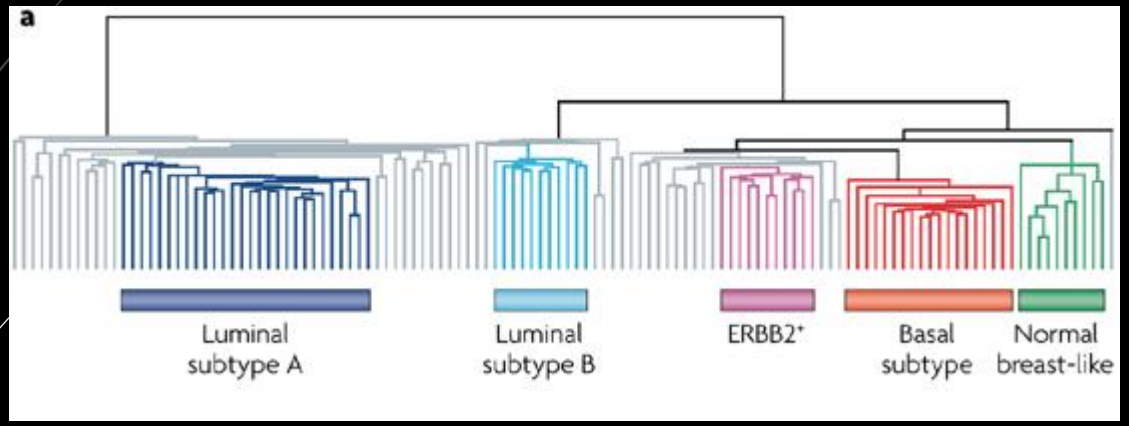


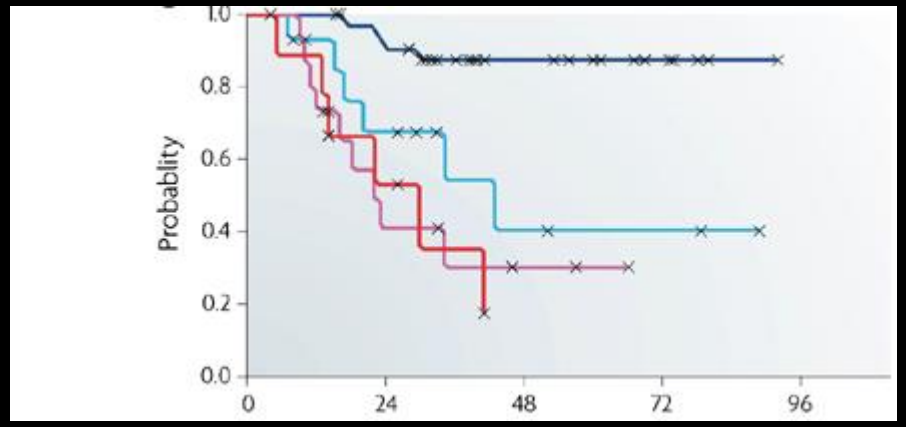
Gene Expression



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*, *RB1* 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

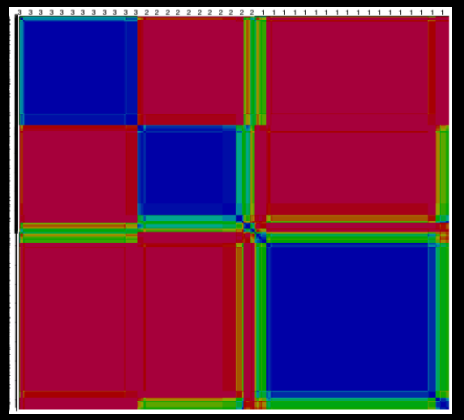
ARTICLE

doi:10.1038/nature10166

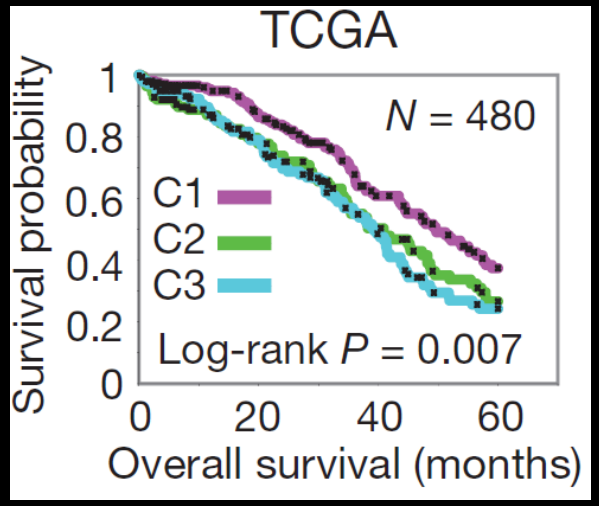
Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

150 most variable miRNAs for sample clustering



Cluster 1 has significantly better survival



https://portal.gdc.cancer.gov

? **Can't find your data?** Click here for more information. Dismiss ✖

NIH NATIONAL CANCER INSTITUTE GDC Data Portal | Home | Projects | Data | Analysis | Quick Search | Login | Cart 0 | GDC Apps

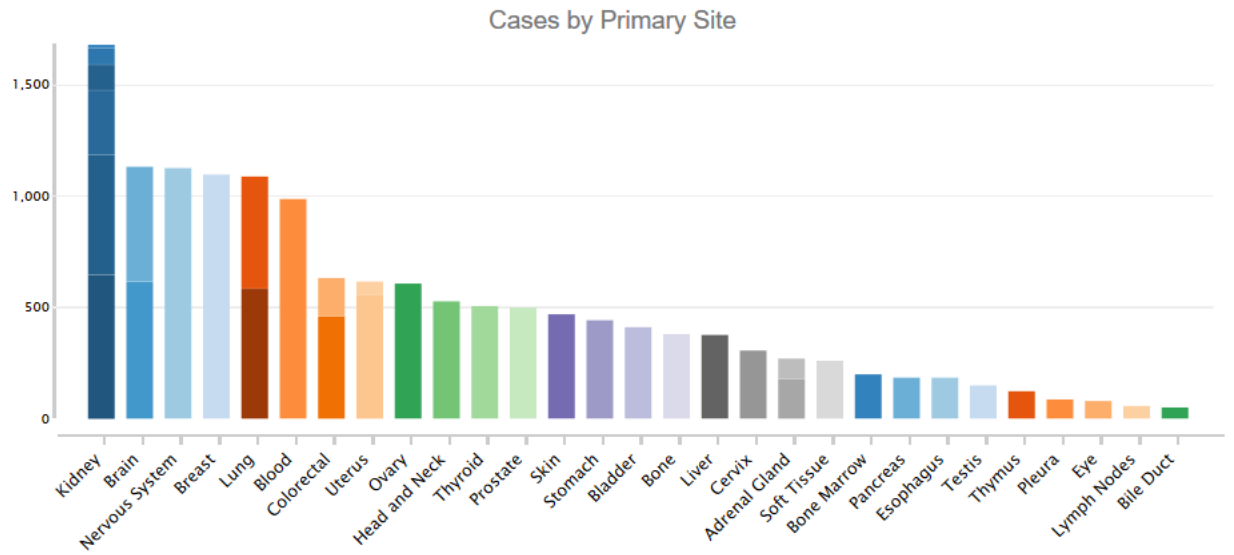
Harmonized Cancer Datasets
Genomic Data Commons Data Portal

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	PRIMARY SITE	CASES	FILES
39	29	14,551	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

Documentation

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

Browse Data using Facet Search

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:

https://portal.gdc.cancer.gov

? Can't find your data? Click here for more information. Dismiss ✖

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 1,681
- Brain 1,133
- Nervous System 1,127
- Breast 1,098
- Lung 1,089

24 More...

Cancer Program

- TCGA 11,315
- TARGET 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

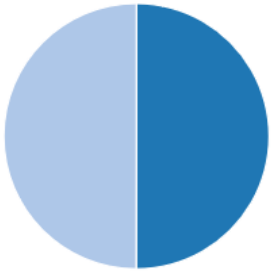
Add all files to the Cart | Download Manifest

FILES	CASES	FILE SIZE
274,724	14,551	470.58 TB

File Counts by Project

39 Projects

File Counts by Access Level



2 Access Levels

File Counts by Data Format

7 Data Formats

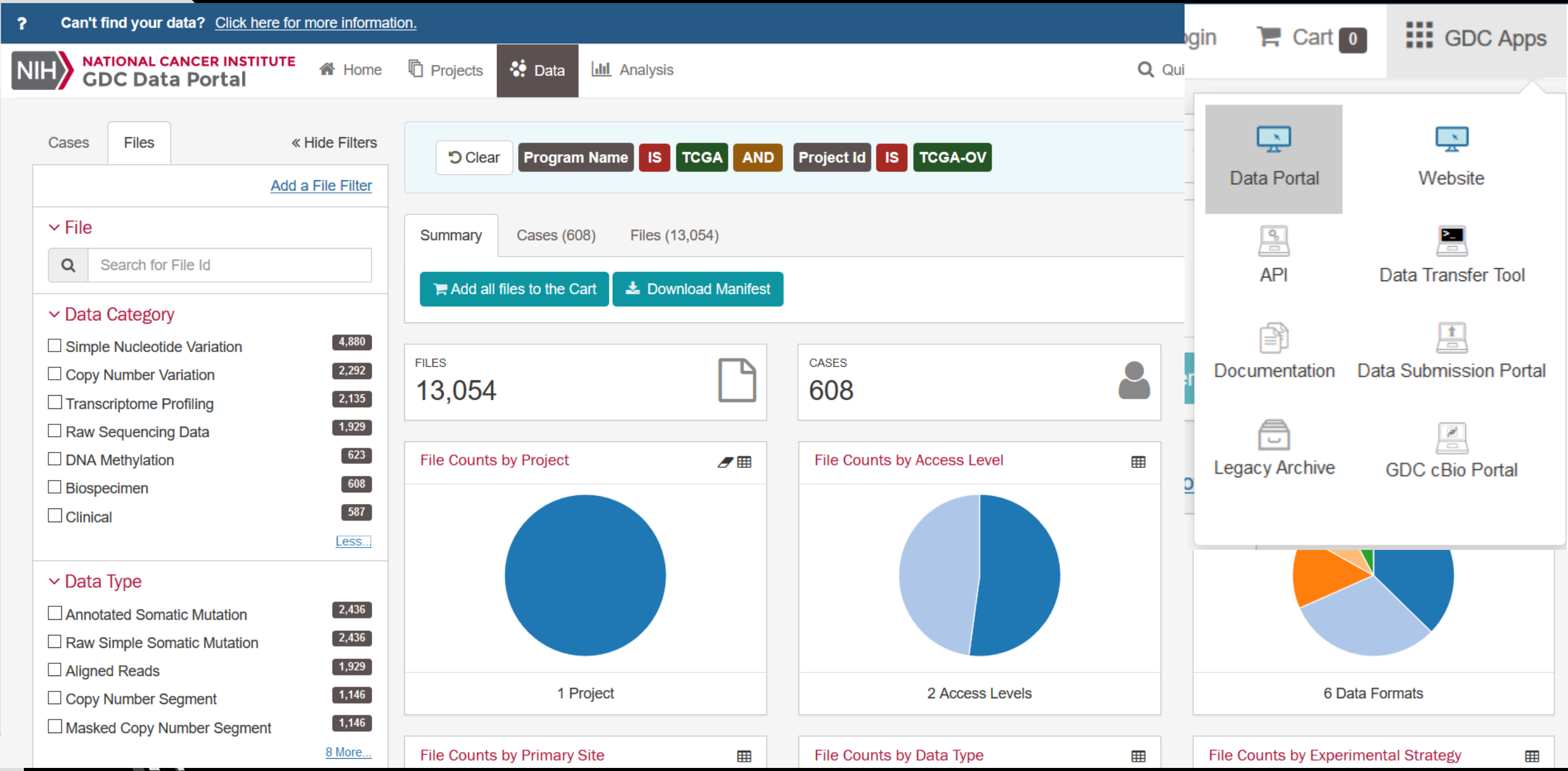
File Counts by Primary Site

File Counts by Data Type

File Counts by Experimental Strategy

The Cancer Genome Atlas (TCGA) Ovarian cancer

https://portal.gdc.cancer.gov



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!!

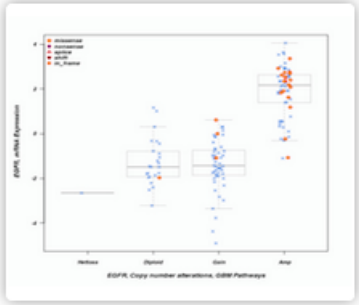
NIH NATIONAL CANCER INSTITUTE
GDC cBioPortal BETA

cBioPortal for Cancer Genomics

HOME DATA SETS

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.



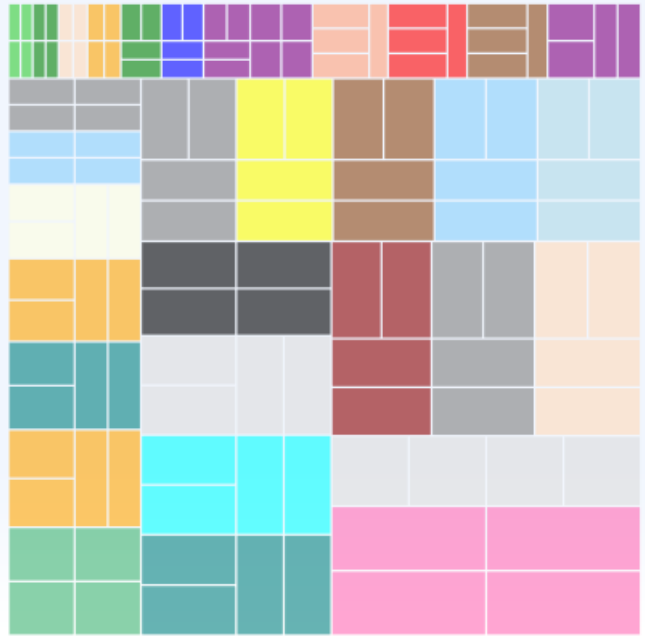
The GDC cBioPortal is an implementation of cBioPortal that supports the visualization of mutation data. For information on cBioPortal, please refer to cBioPortal.org.

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

```
[rbao@cri16in002 ~]$ 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

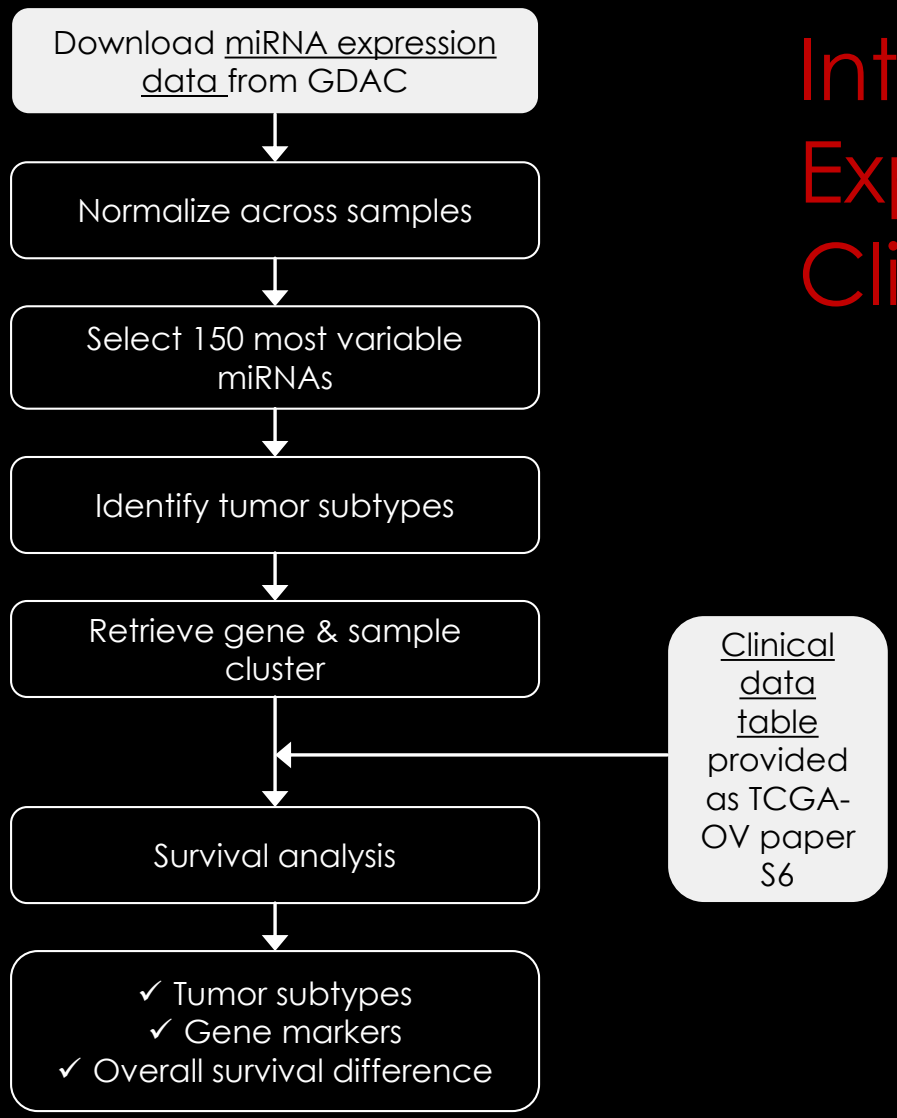
- library(TCGAbiolinks)
  - https://goo.gl/ytPe07
- library(GenomicDataCommons)
  - https://goo.gl/cbUfp3

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



Download miRNA expression data from GDAC

Normalize across samples

Select 150 most variable miRNAs

Identify tumor subtypes

Retrieve gene & sample cluster

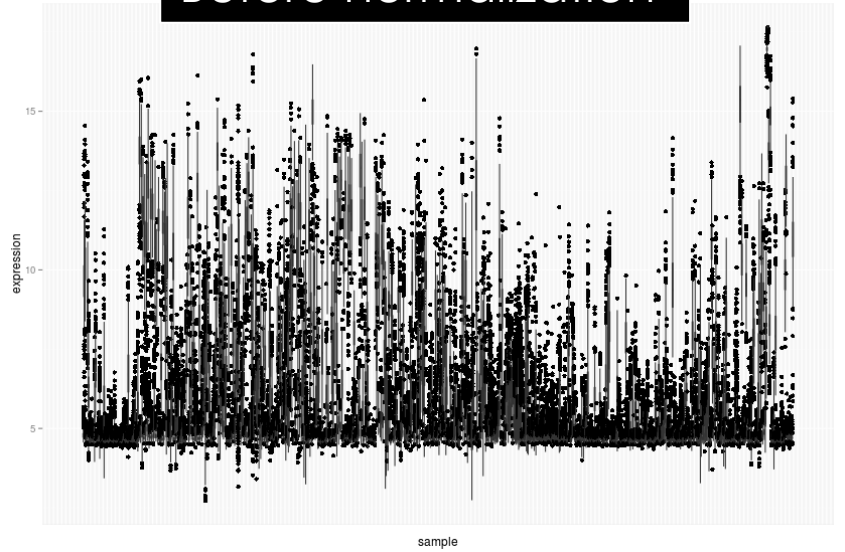
Clinical data table provided as TCGA-OV paper S6

Survival analysis

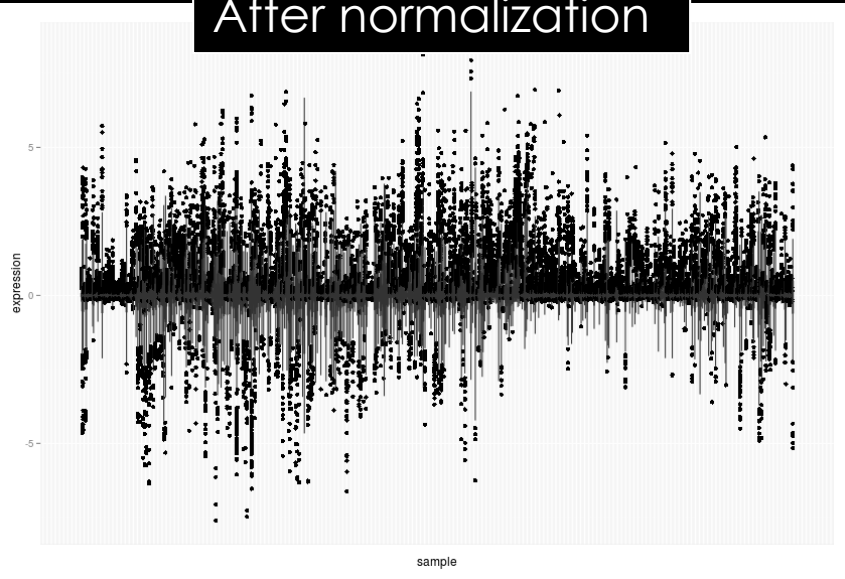
✓ Tumor subtypes
✓ Gene markers
✓ Overall survival difference

miRNA	Sample Total	Sample in TCGA-OV paper
821	594	486

Before normalization



After normalization



Download miRNA expression data from GDAC

Normalize across samples

Select 150 most variable miRNAs

Identify tumor subtypes

Retrieve gene & sample cluster

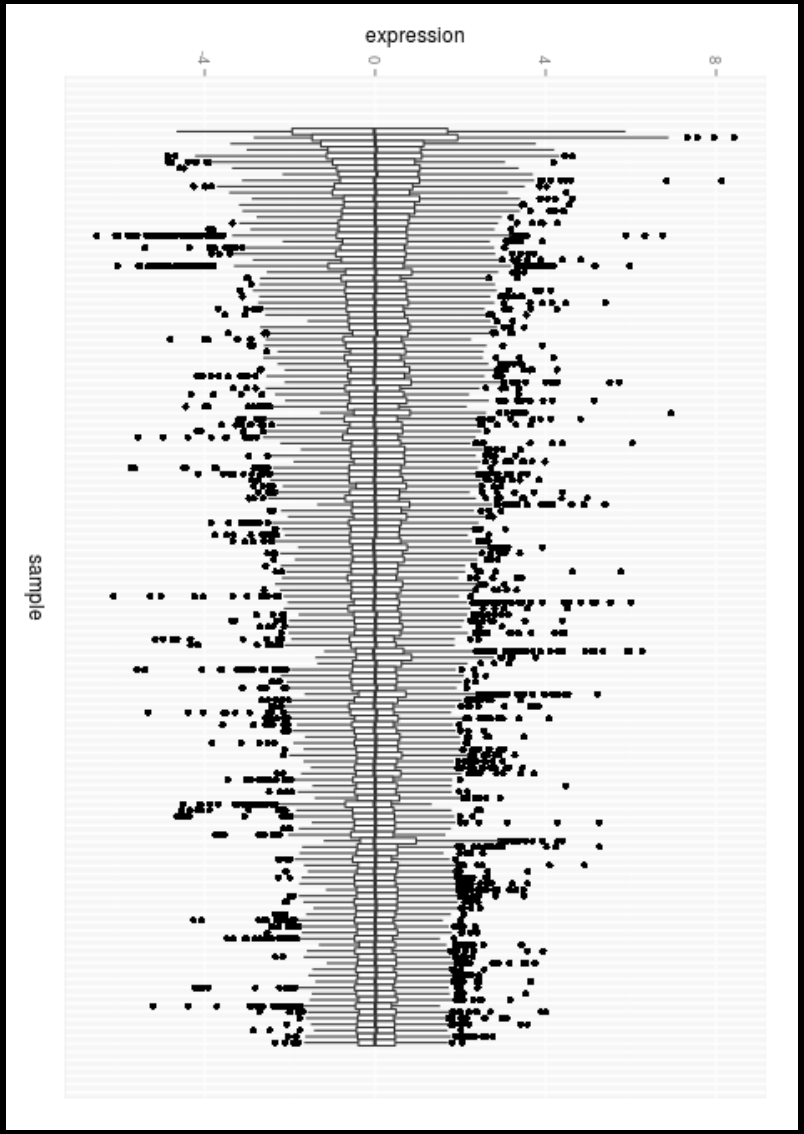
Clinical data table provided as TCGA-OV paper S6

Survival analysis

✓ Tumor subtypes
✓ Gene markers
✓ Overall survival difference

	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)

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

Download miRNA expression data from GDAC

Normalize across samples

Select 150 most variable miRNAs

Identify tumor subtypes

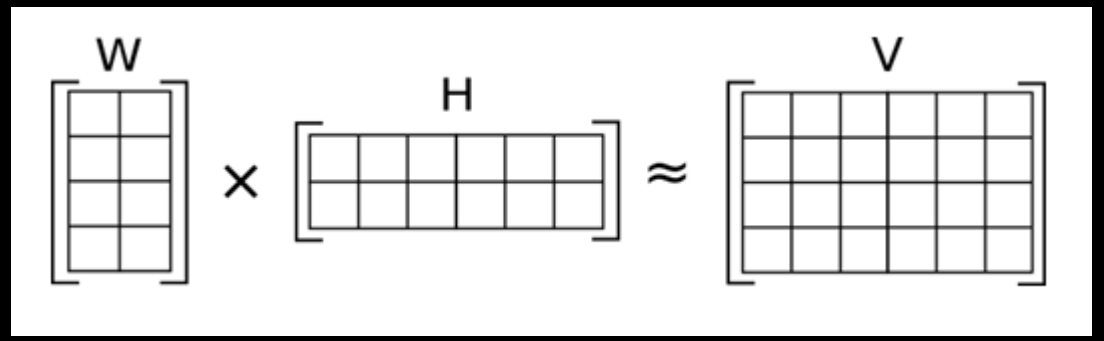
Retrieve gene & sample cluster

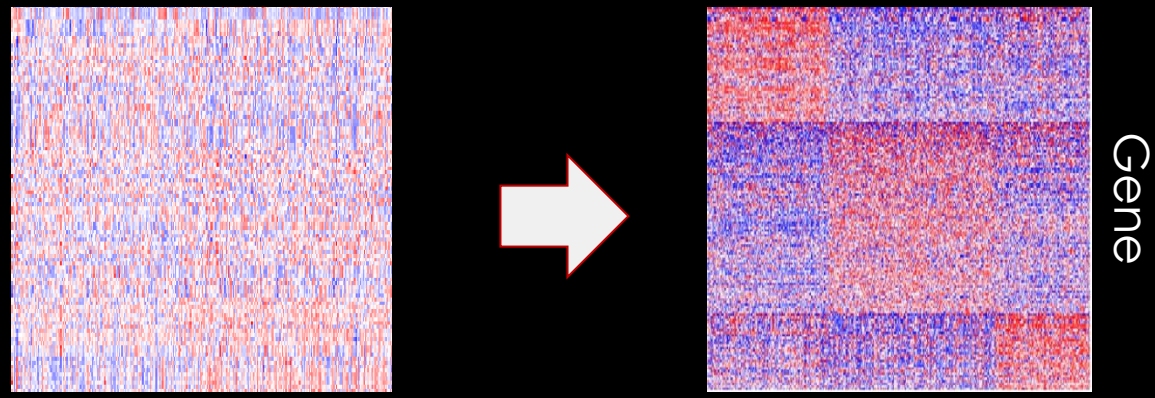
Clinical data table provided as TCGA-OV paper S6

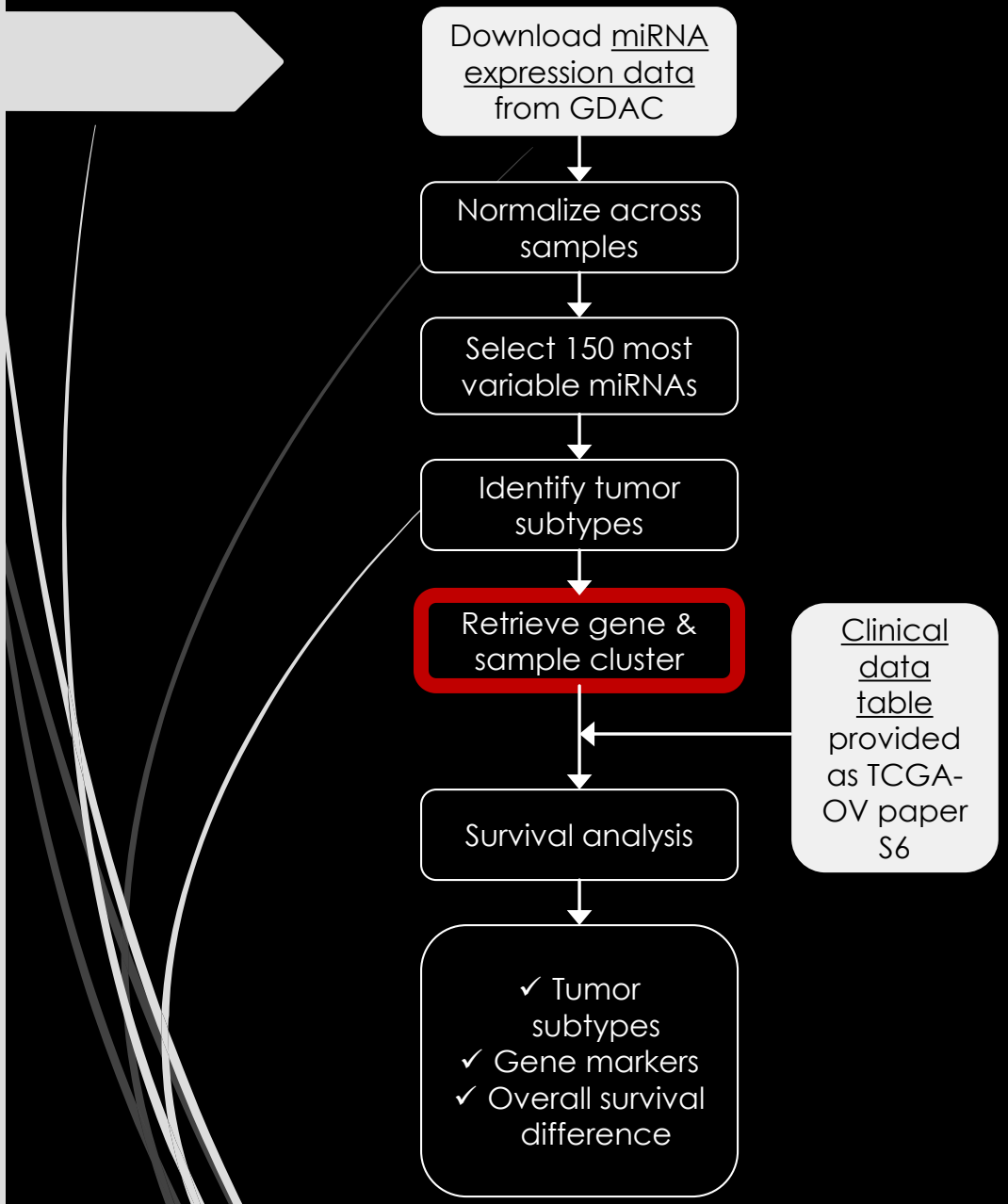
Survival analysis

✓ Tumor subtypes
✓ Gene markers
✓ Overall survival difference

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.





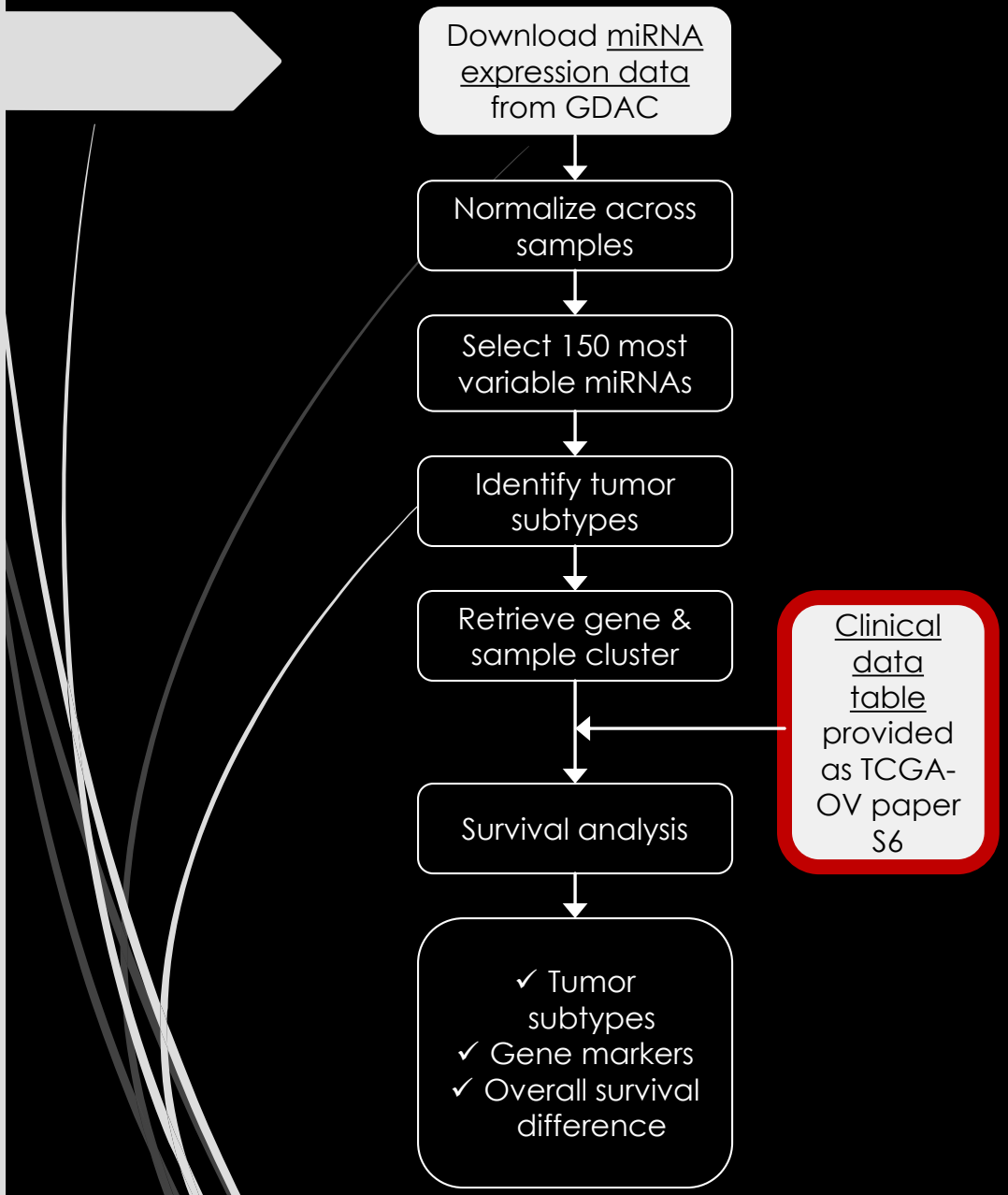


sample	cluster
TCGA.04.1331	1
TCGA.04.1338	1
TCGA.04.1341	1
TCGA.04.1343	1
TCGA.04.1348	1
TCGA.04.1362	1
TCGA.04.1365	1
TCGA.04.1530	1
TCGA.04.1542	1
TCGA.04.1648	1
TCGA.04.1649	1
TCGA.04.1651	1

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

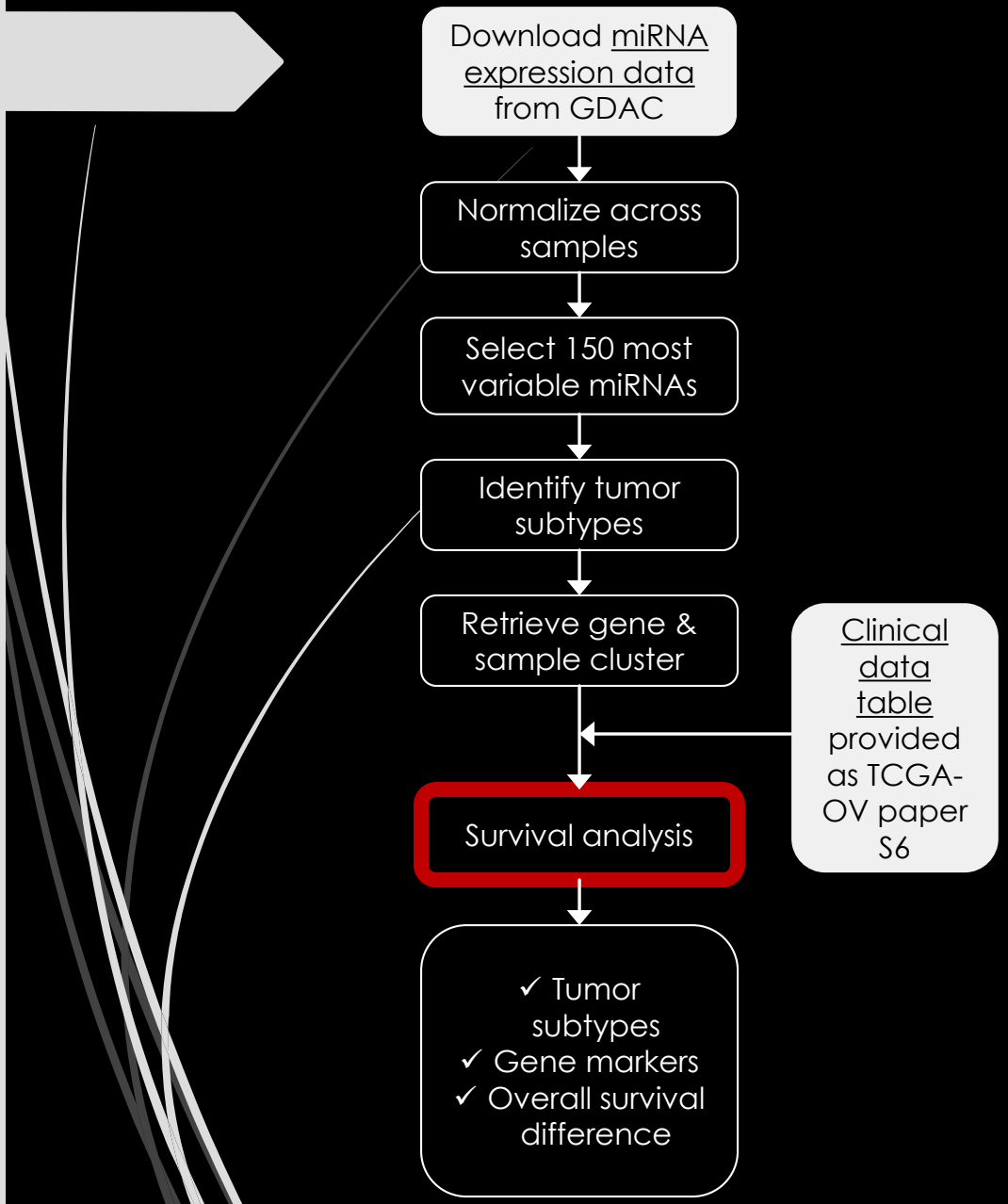
Gene clusters



	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	IIC	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)

Download miRNA expression data from GDAC

Normalize across samples

Select 150 most variable miRNAs

Identify tumor subtypes

Retrieve gene & sample cluster

Clinical data table provided as TCGA-OV paper S6

Survival analysis

✓ Tumor subtypes
✓ Gene markers
✓ Overall survival difference

Right Censoring

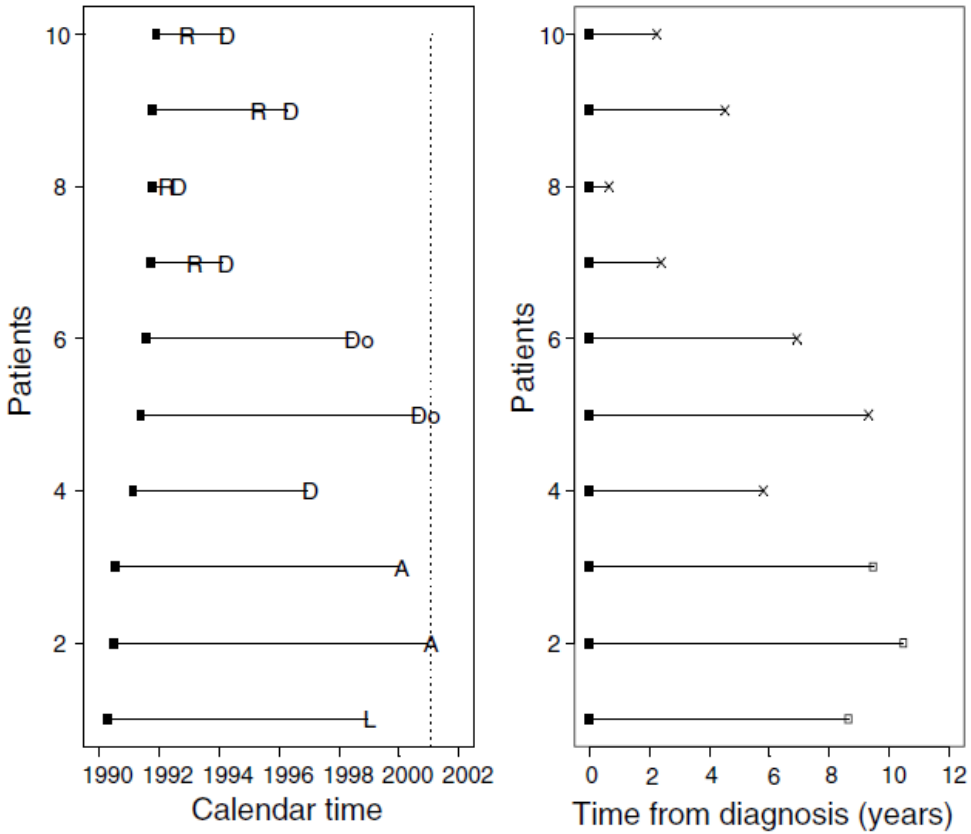
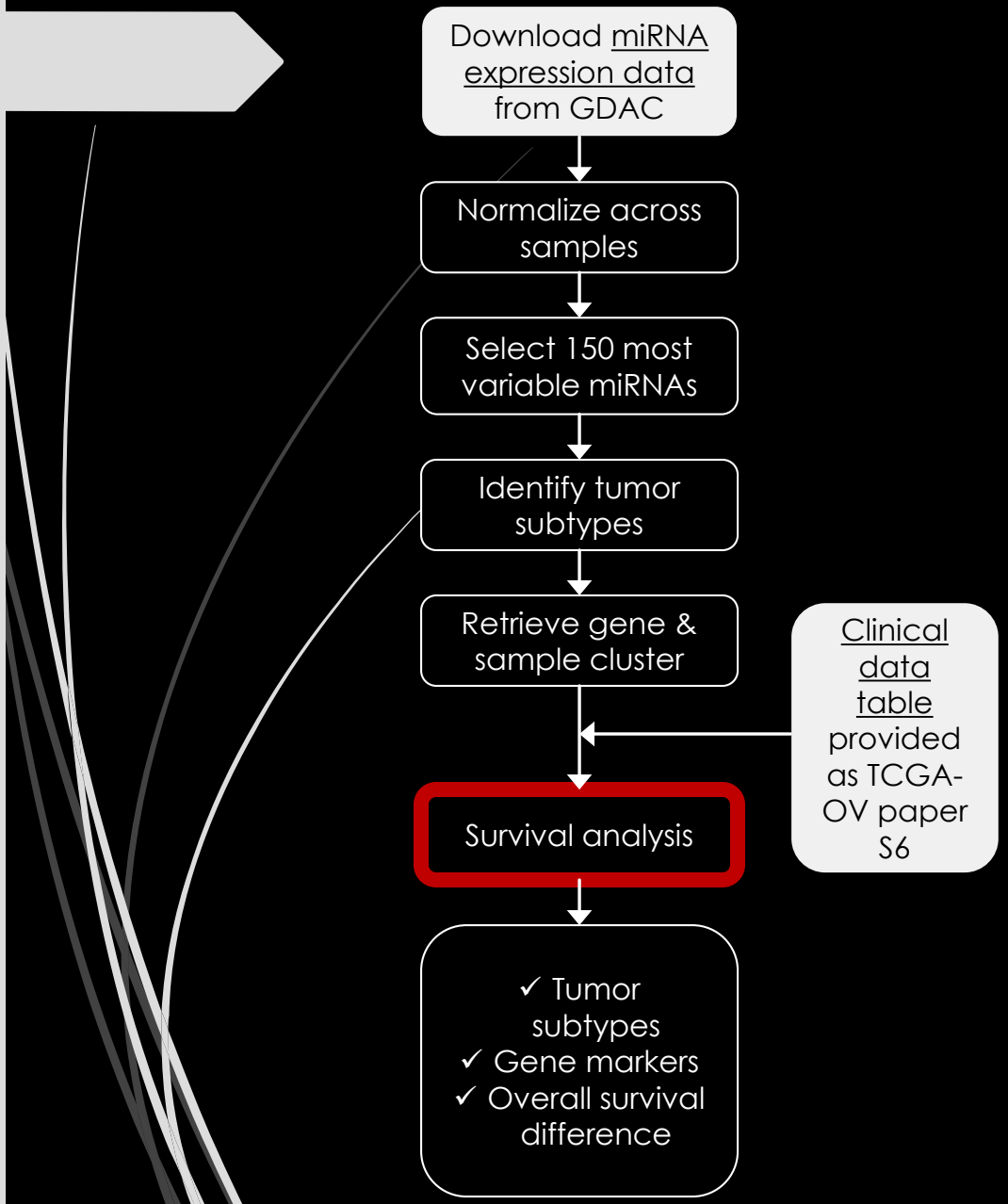
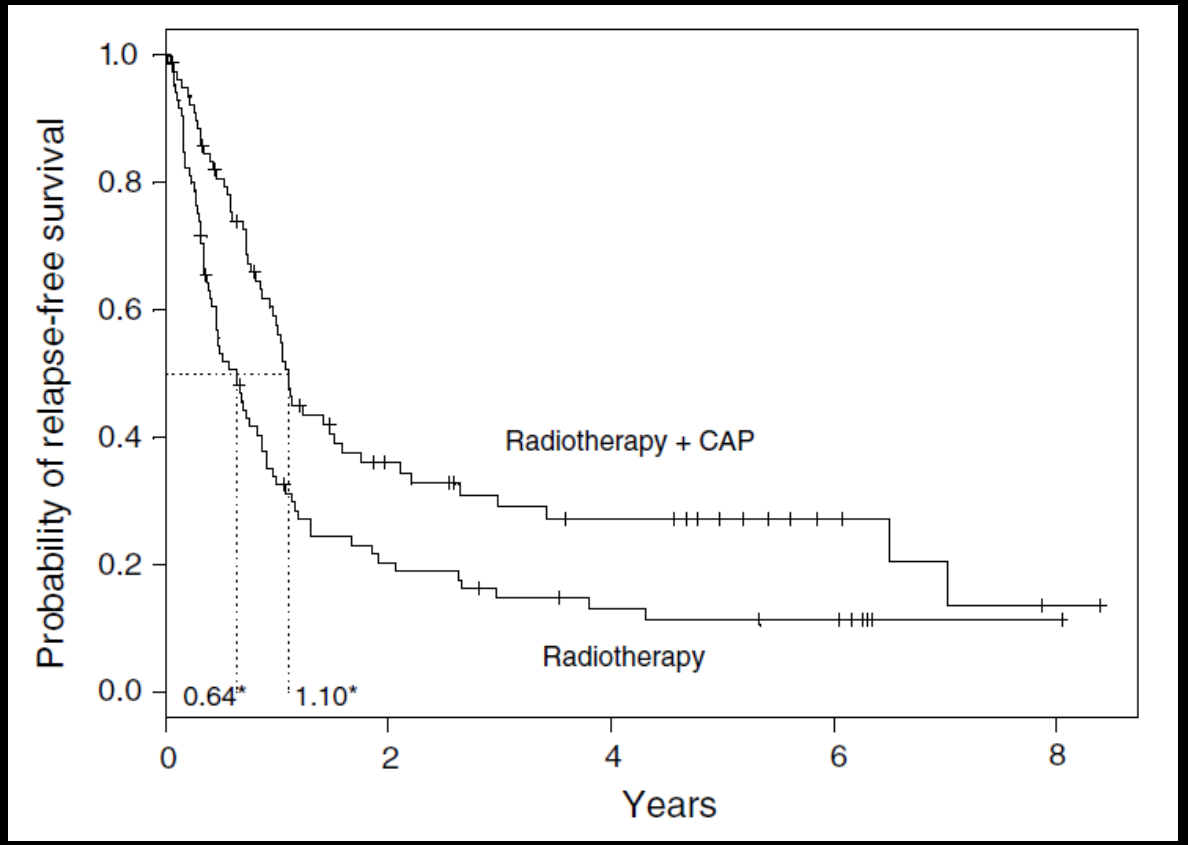


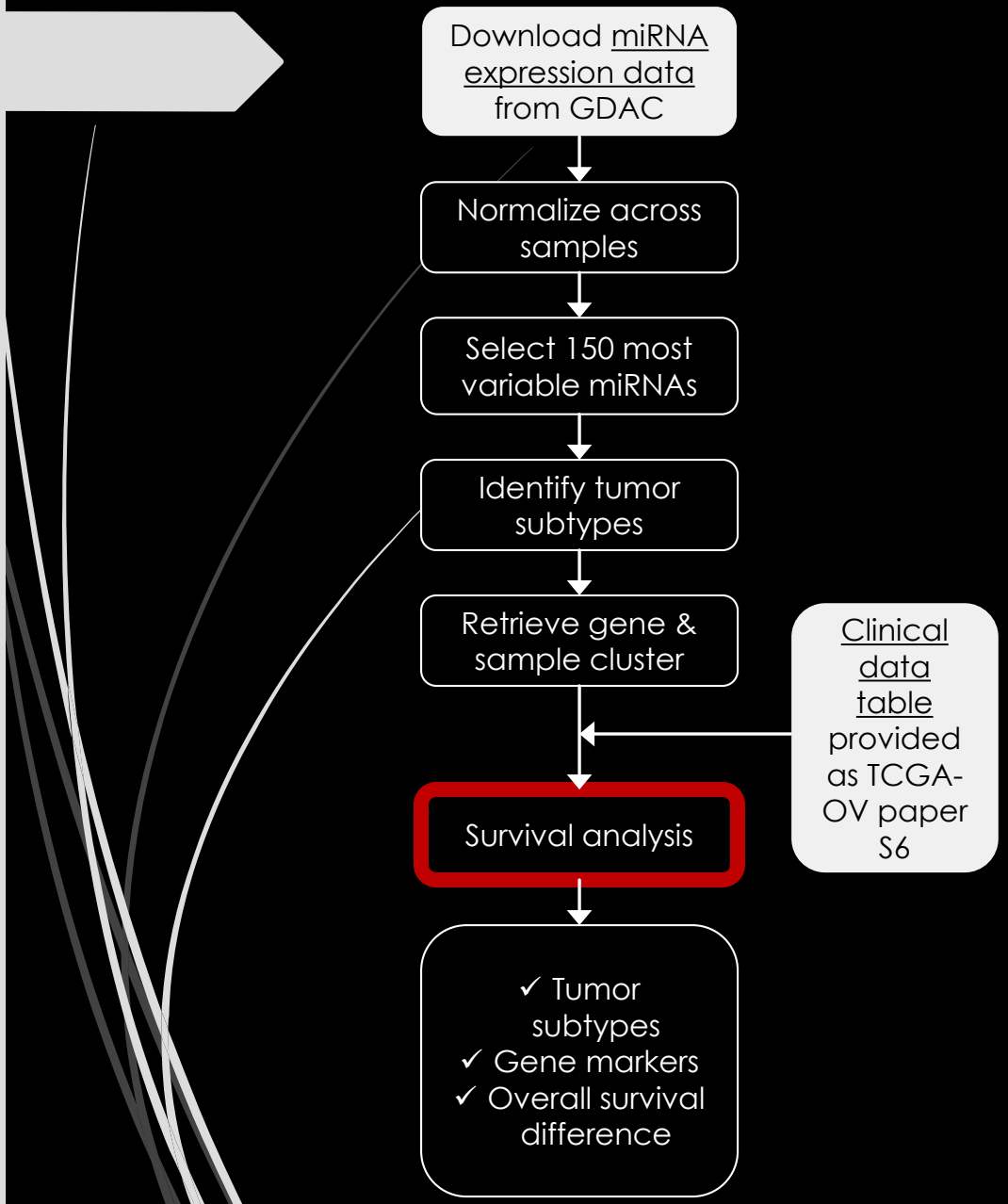
Figure 1 Converting calendar time in the ovarian cancer study to a survival analysis format. Dashed vertical line is the date of the last follow-up, R = relapse, D = death from ovarian cancer, Do = death from other cause, A = attended last clinic visit (alive), L = loss to follow-up, X = death, □ = censored.

Clark et al. Survival Analysis



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



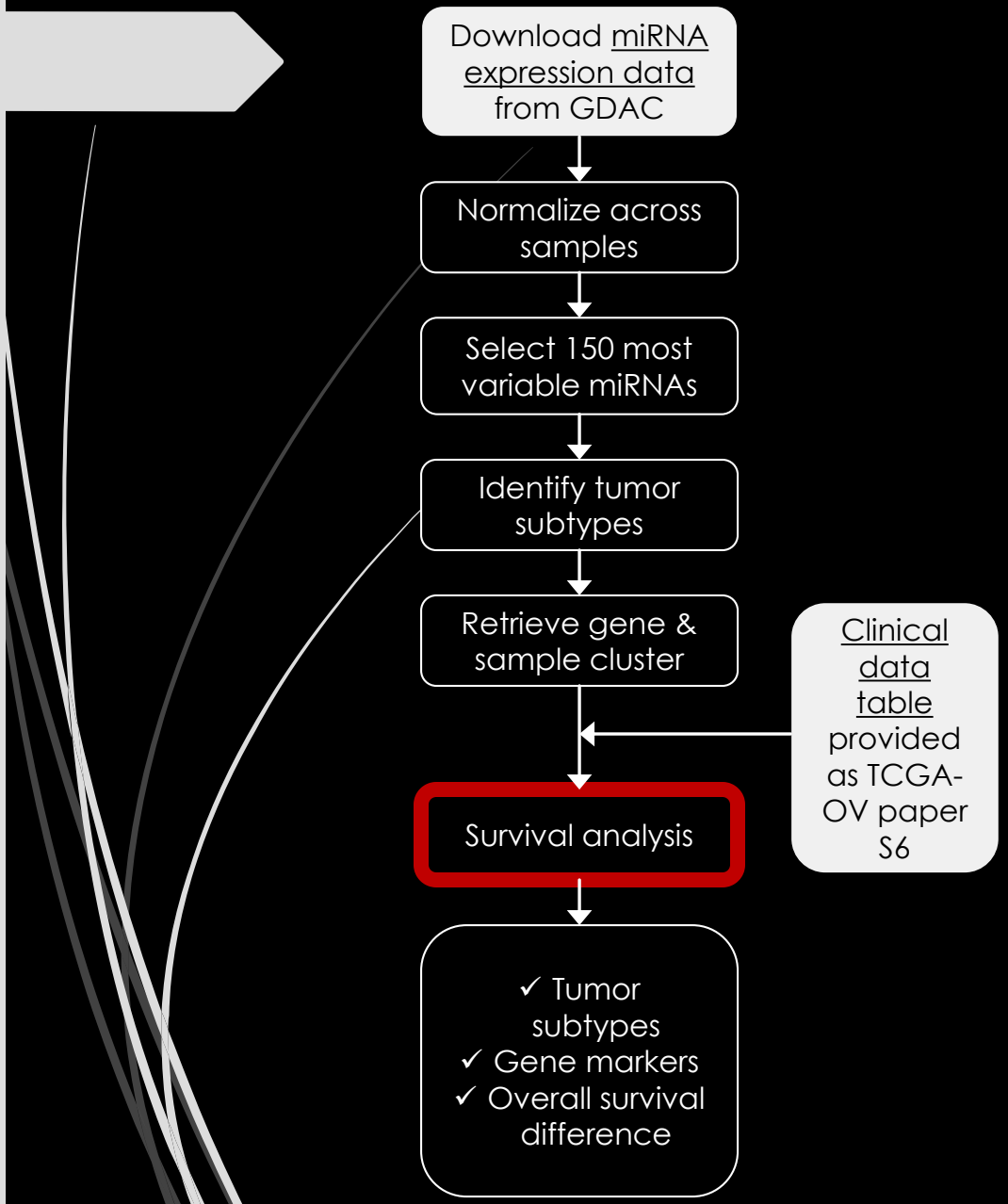


- Log-rank test
  - Chi-square test
  - Efficient in comparing groups differed by categorical variables, but not continuous ones
  - Univariate test

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

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

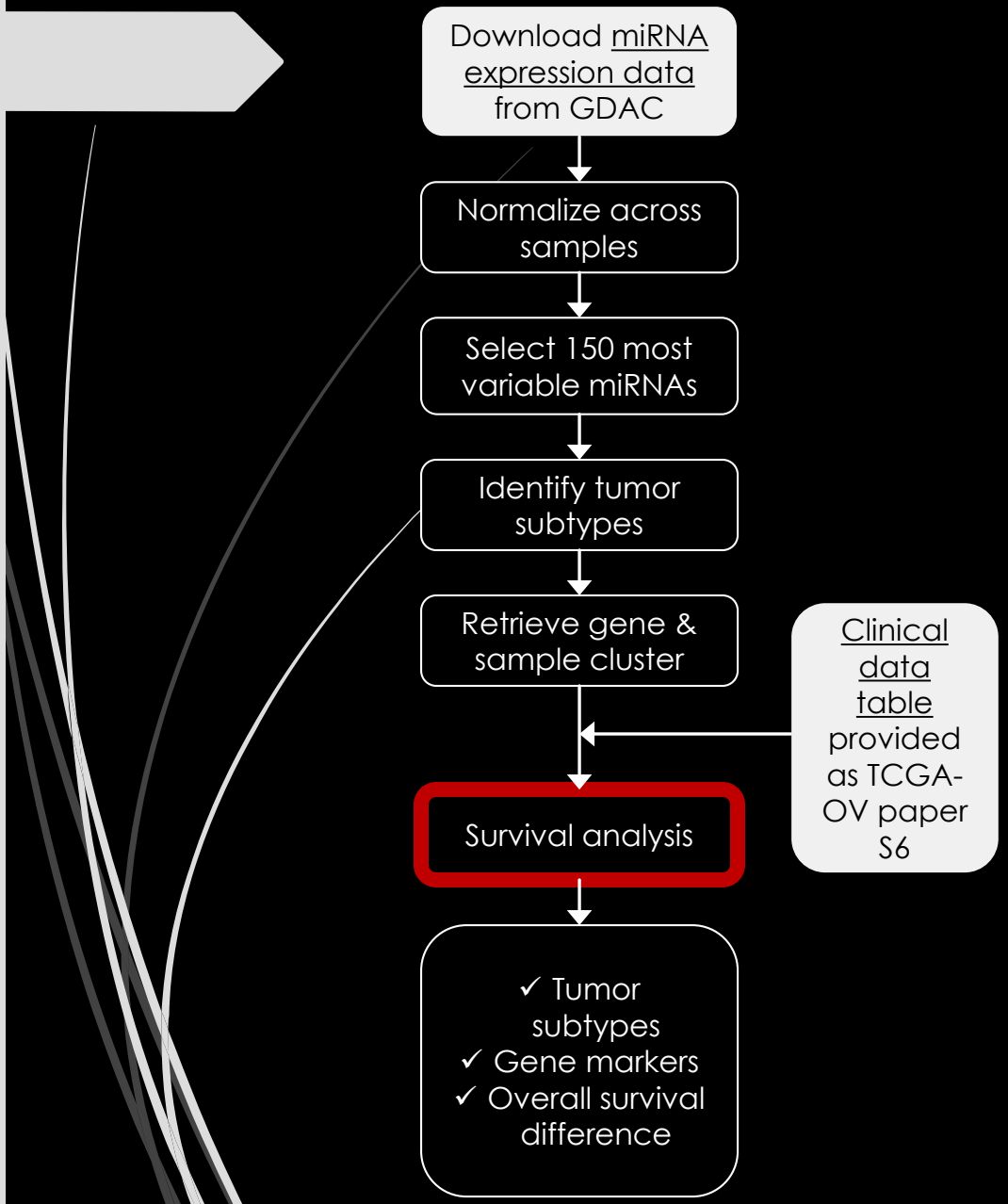
O: Observed
E: Expected (if no difference between group 1 and 2)



- 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
```

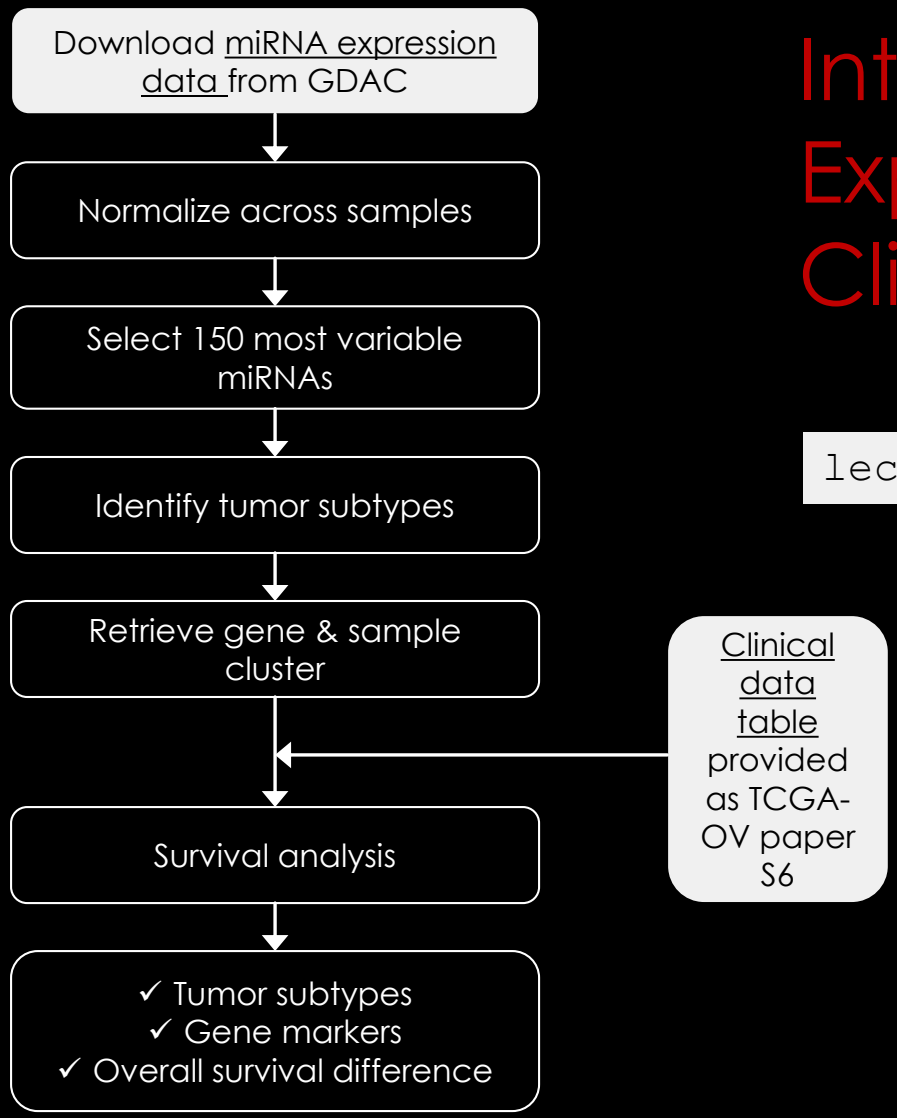
```
Call:
coxph(formula = surv ~ (clinical.sub$cluster + clinical.sub$age.at.diagnosis.year +
    clinical.sub$tumor.grade))


                                      coef exp(coef) se(coef)    z      p
clinical.sub$cluster                0.42658   1.53202  0.16031 2.66 0.0078
clinical.sub$age.at.diagnosis.year 0.02042   1.02063  0.00764 2.67 0.0075
clinical.sub$tumor.gradeG3         0.47041   1.60065  0.22140 2.12 0.0336

Likelihood ratio test=18.2  on 3 df, p=0.000403
n= 316, number of events= 169
   (15 observations deleted due to missingness)
```



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

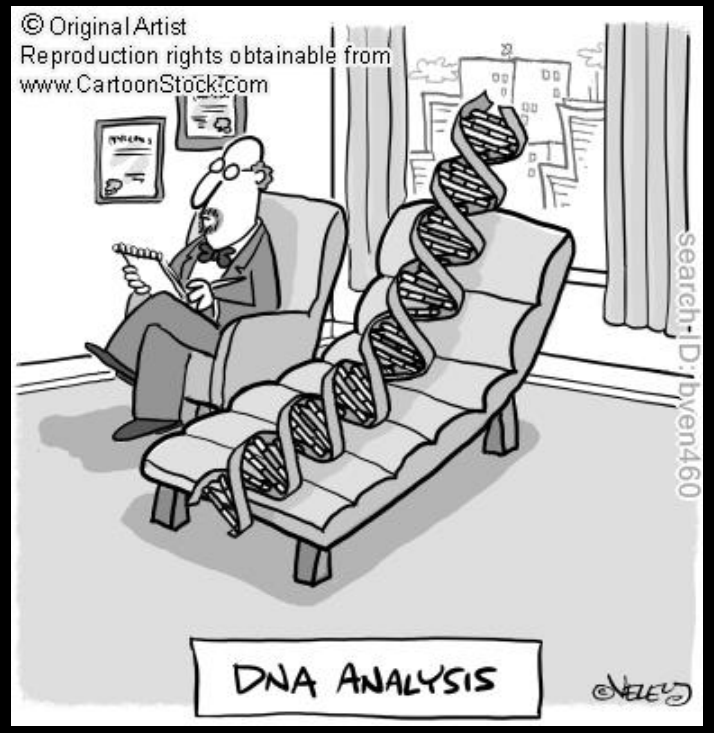
```
Library(survival)
survfit
survdiff
coxph
```

Integrate Expression with Clinical Data



`lecture10.Rmd`

Thank you!



Questions

INDEX
EXTRA QUALITY
90