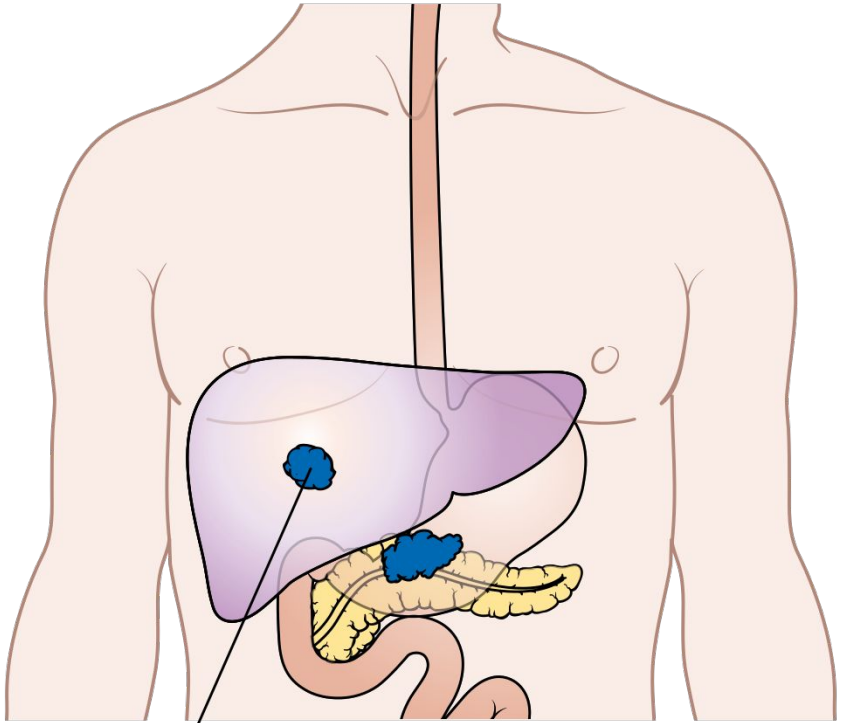
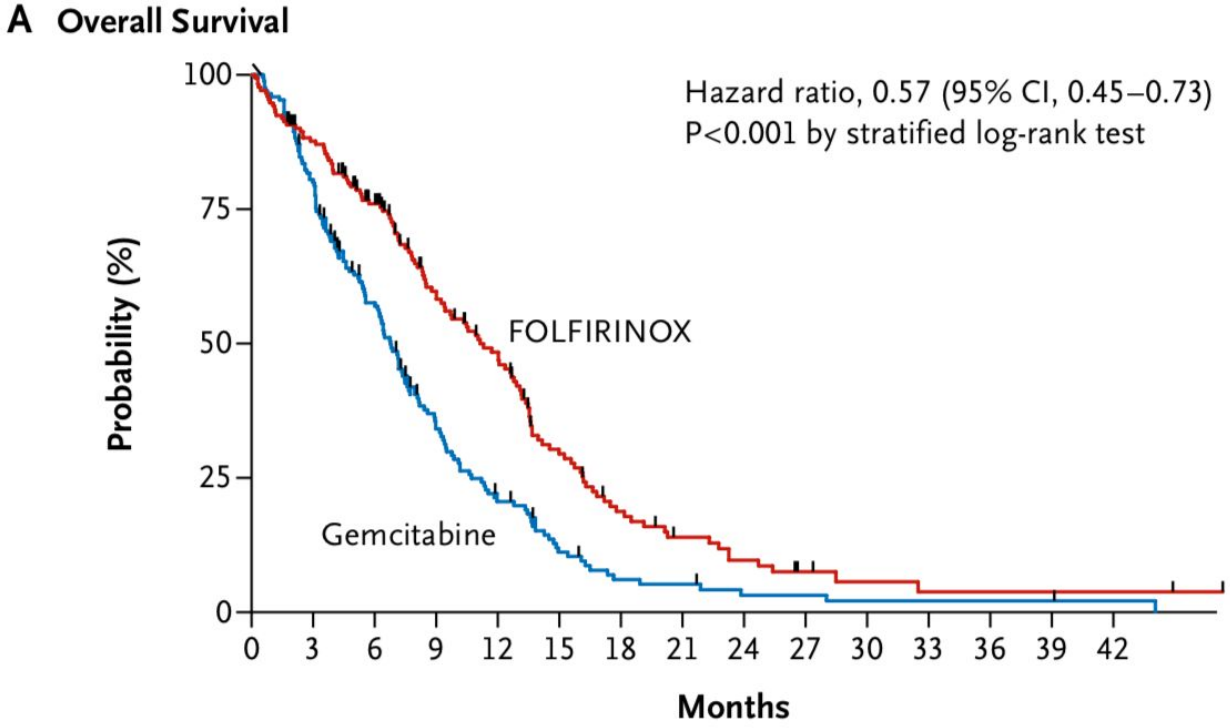
Transcriptomic Subtyping in Cancer

Rob Grant and Nathan Taback
30 January 2020

Pancreatic Cancer

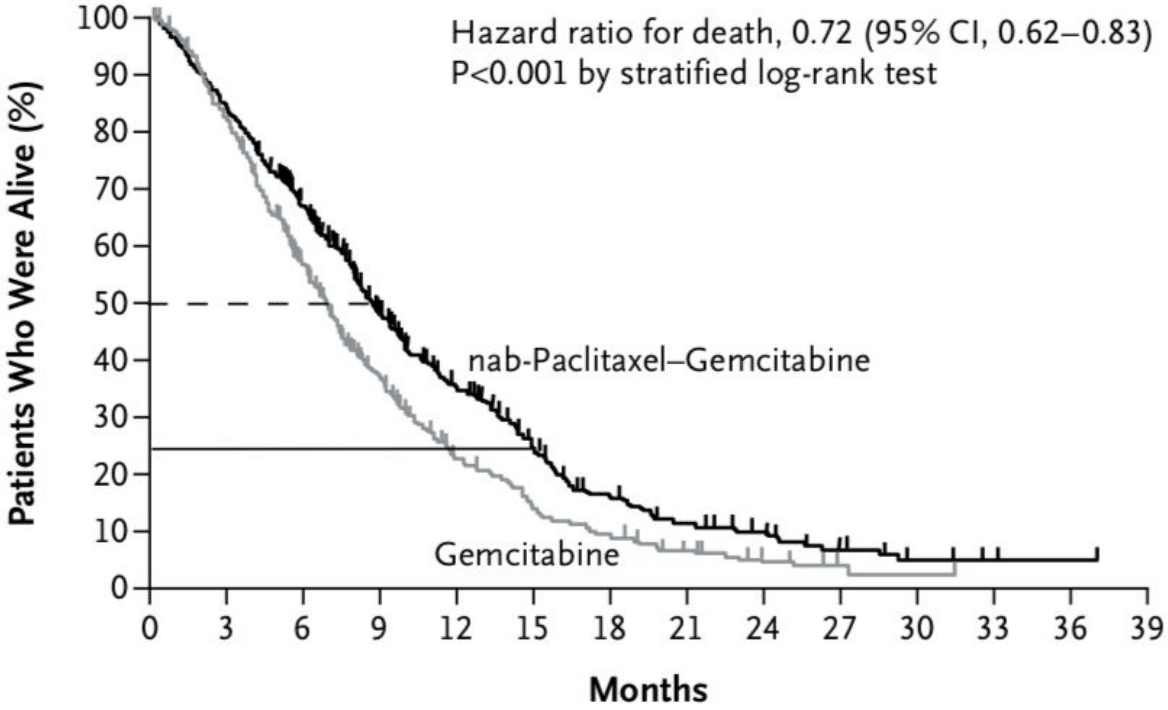


PRODIGE 4



Conroy *et al., NEJM,* 2011

MPACT

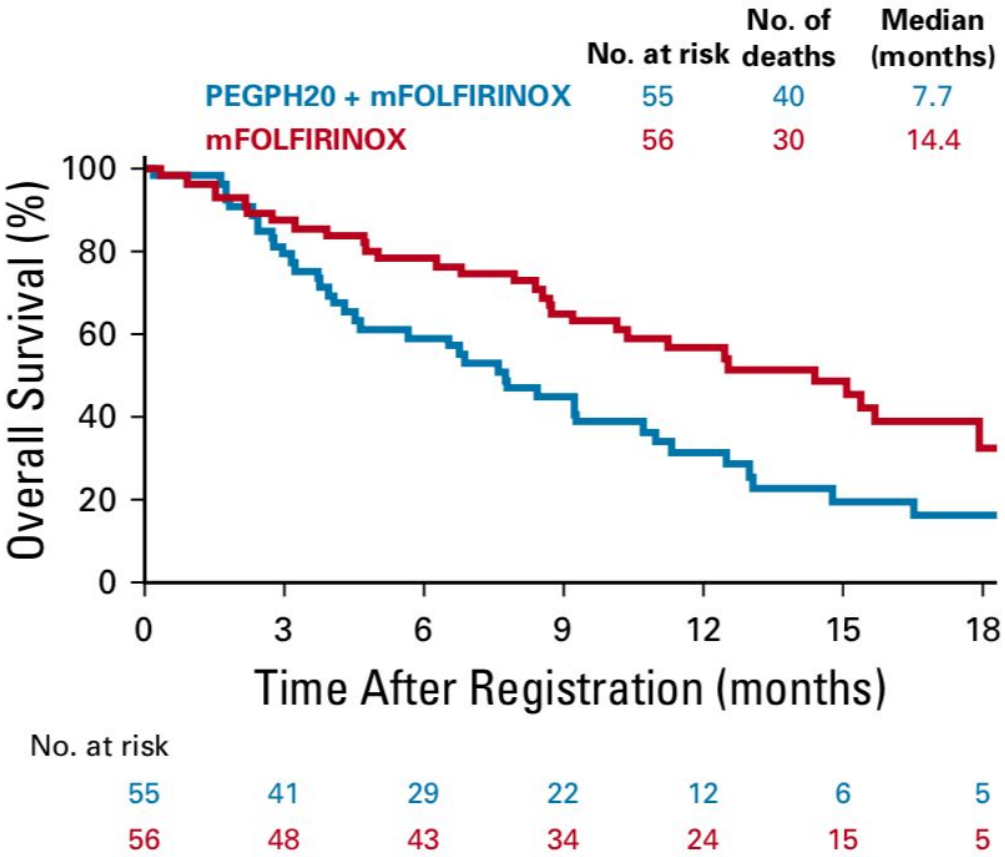


No. at Risk

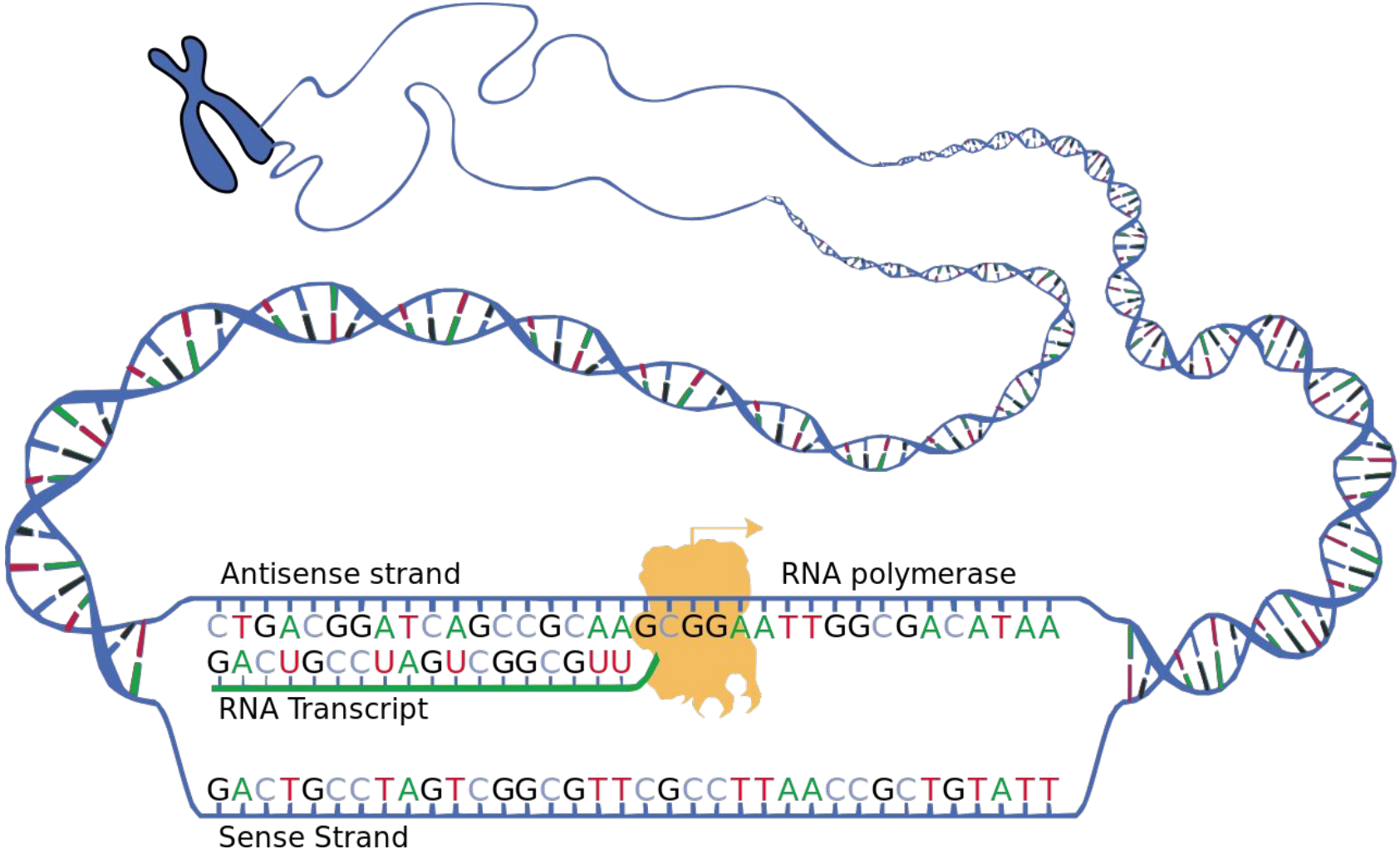
nab-Paclitaxel–Gemcitabine	431	357	269	169	108	67	40	27	16	9	4	1	1	0
Gemcitabine	430	340	220	124	69	40	26	15	7	3	1	0	0	0

Von Hoff *et al., NEJM,* 2011

PEGPH20

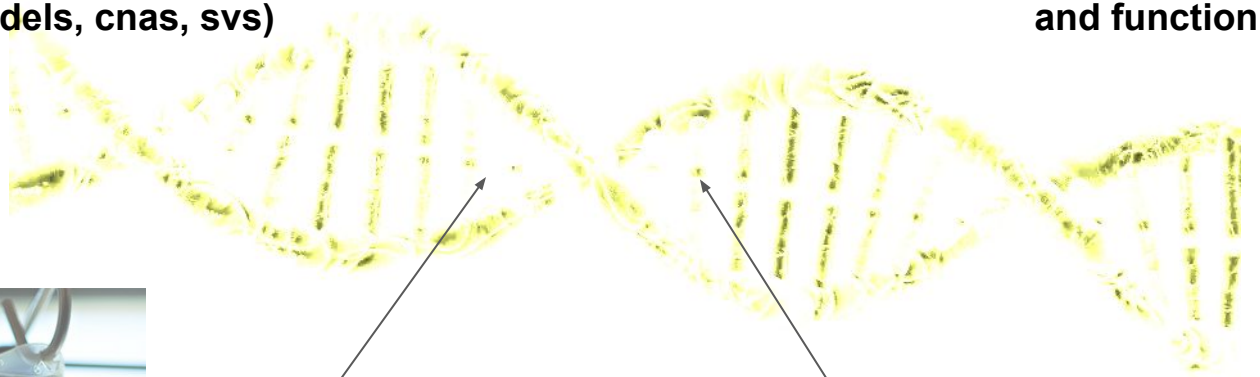


Ramanathan *et al., JCO,* 2019

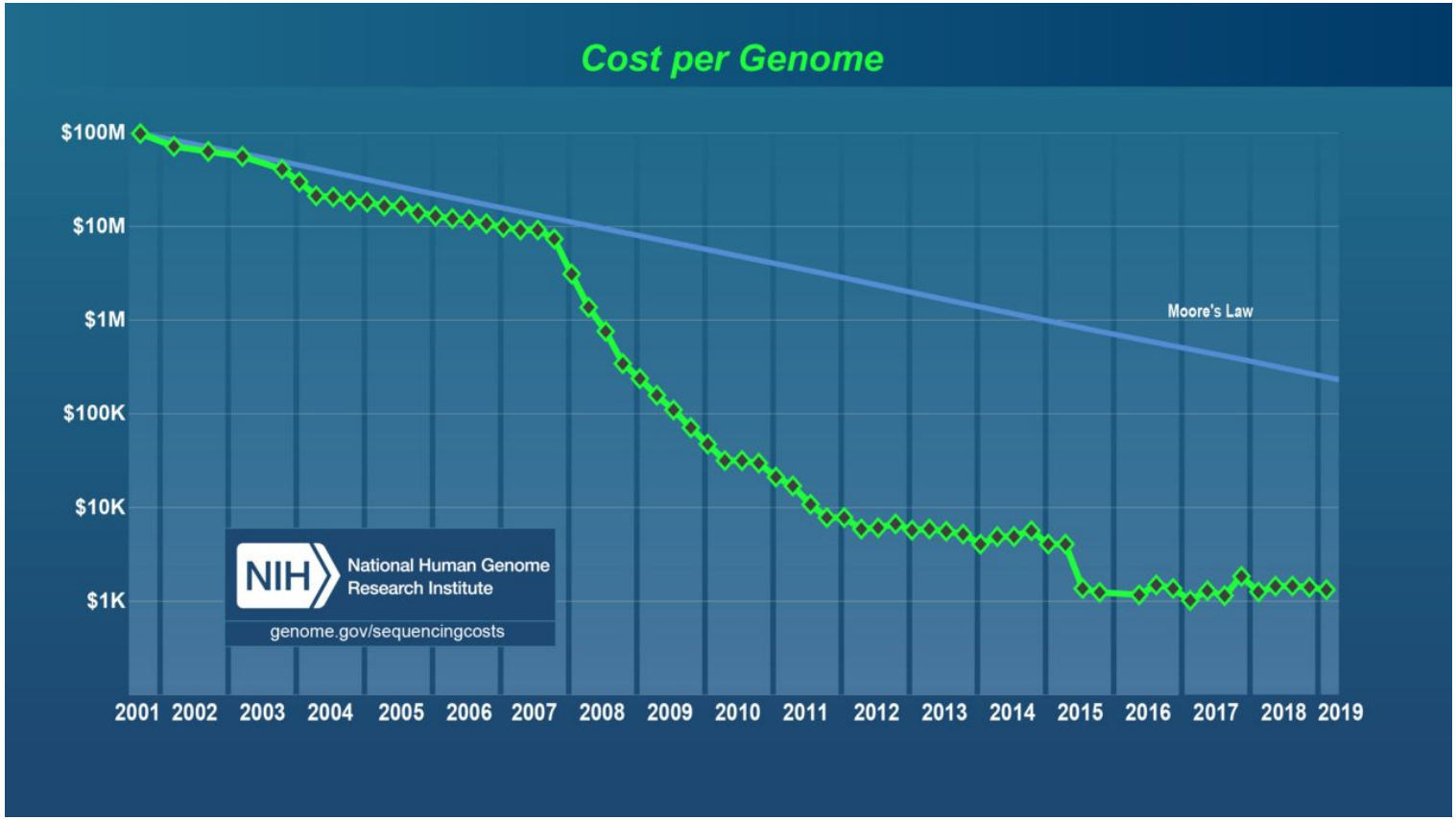

Antisense strand
RNA polymerase
CTGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
GACUGCCUAGUCGGCGUU
RNA Transcript
GACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT
Sense Strand

Cancer

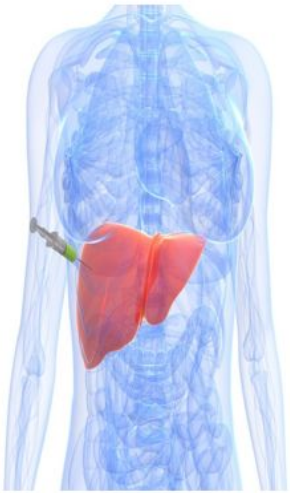
DNA mutations (snvs, indels, cnas, svs) ⟷ **RNA expression** ⟷ **Protein expression and function**

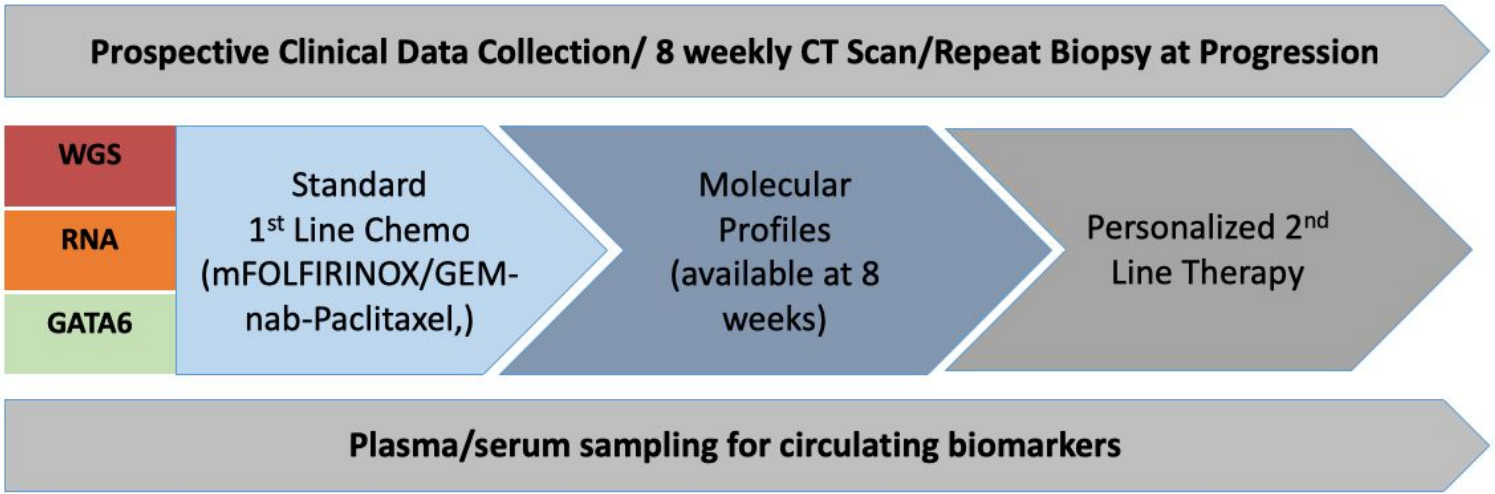


Treatments ⟷ **Patients**

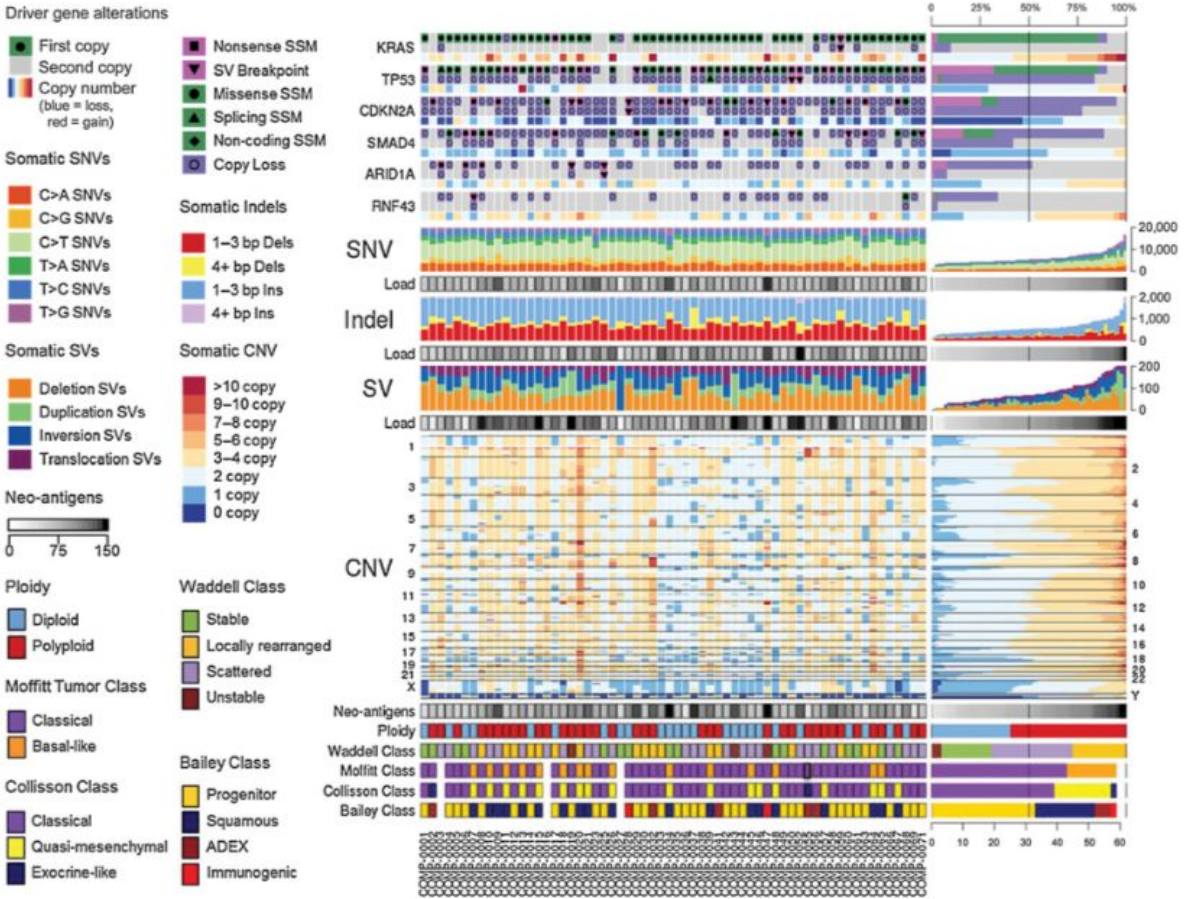

Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019

COMPASS Trial



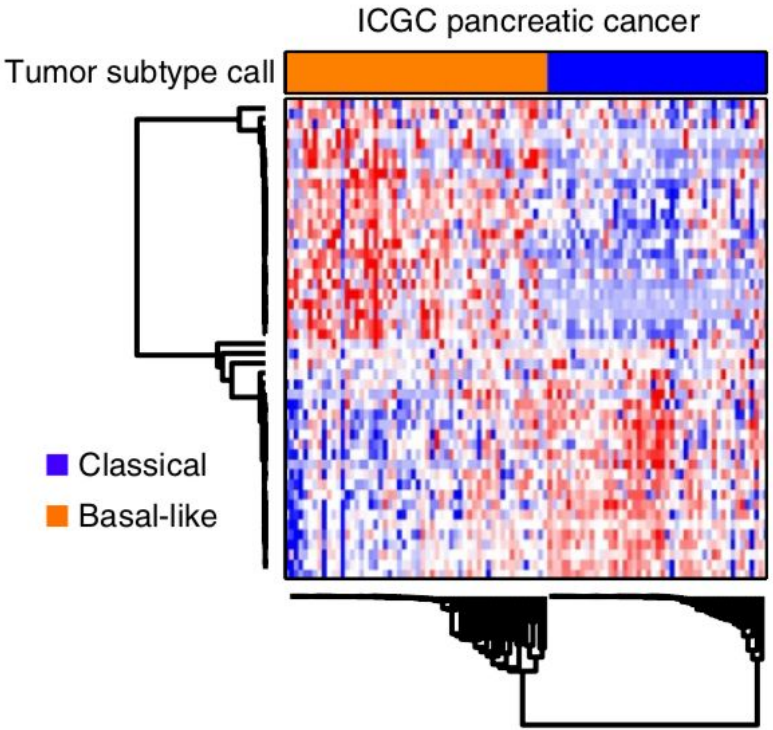


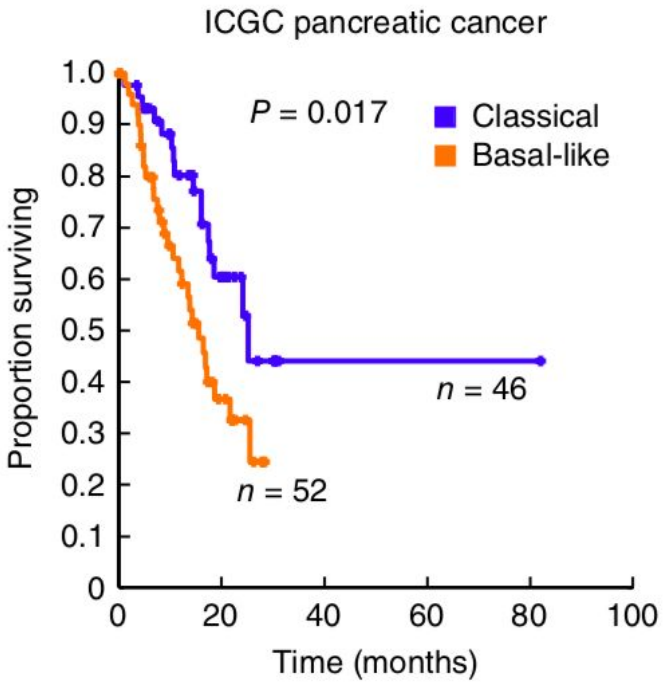
Aung *et al., Clinical Cancer Research,* 2017



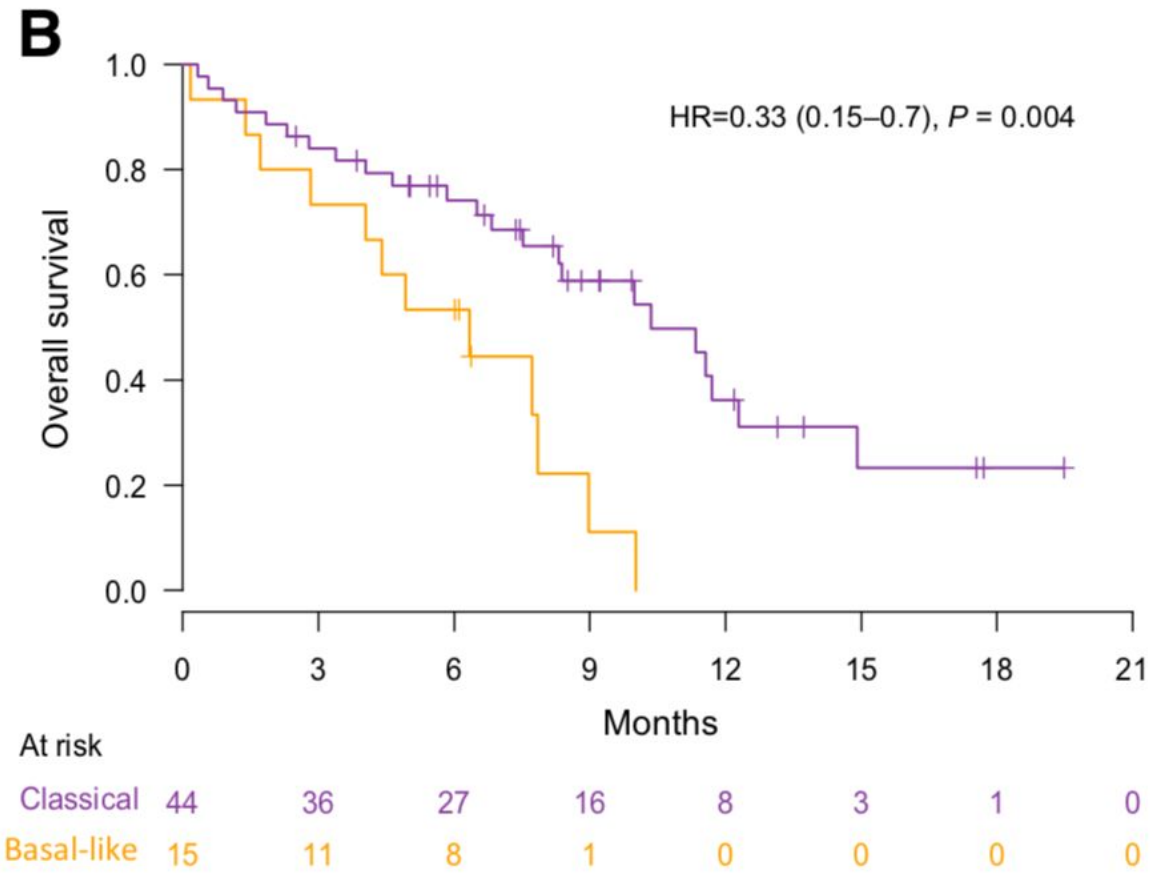
Aung *et al., Clinical Cancer Research,* 2017

Moffitt subtyping





Moffitt *et al., Nature Genetics,* 2015



Aung *et al., Clinical Cancer Research,* 2017

NATIONAL CANCER INSTITUTE

The Cancer Genome Atlas

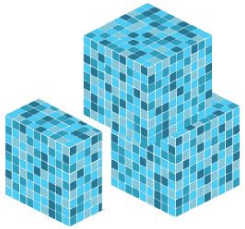
TCGA BY THE NUMBERS

TCGA produced over

2.5

PETABYTES

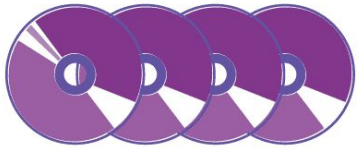
of data



To put this into perspective, **1 petabyte** of data is equal to

212,000

DVDs



TCGA data describes

33

DIFFERENT TUMOR TYPES

...including

10

RARE CANCERS

...based on paired tumor and normal tissue sets collected from

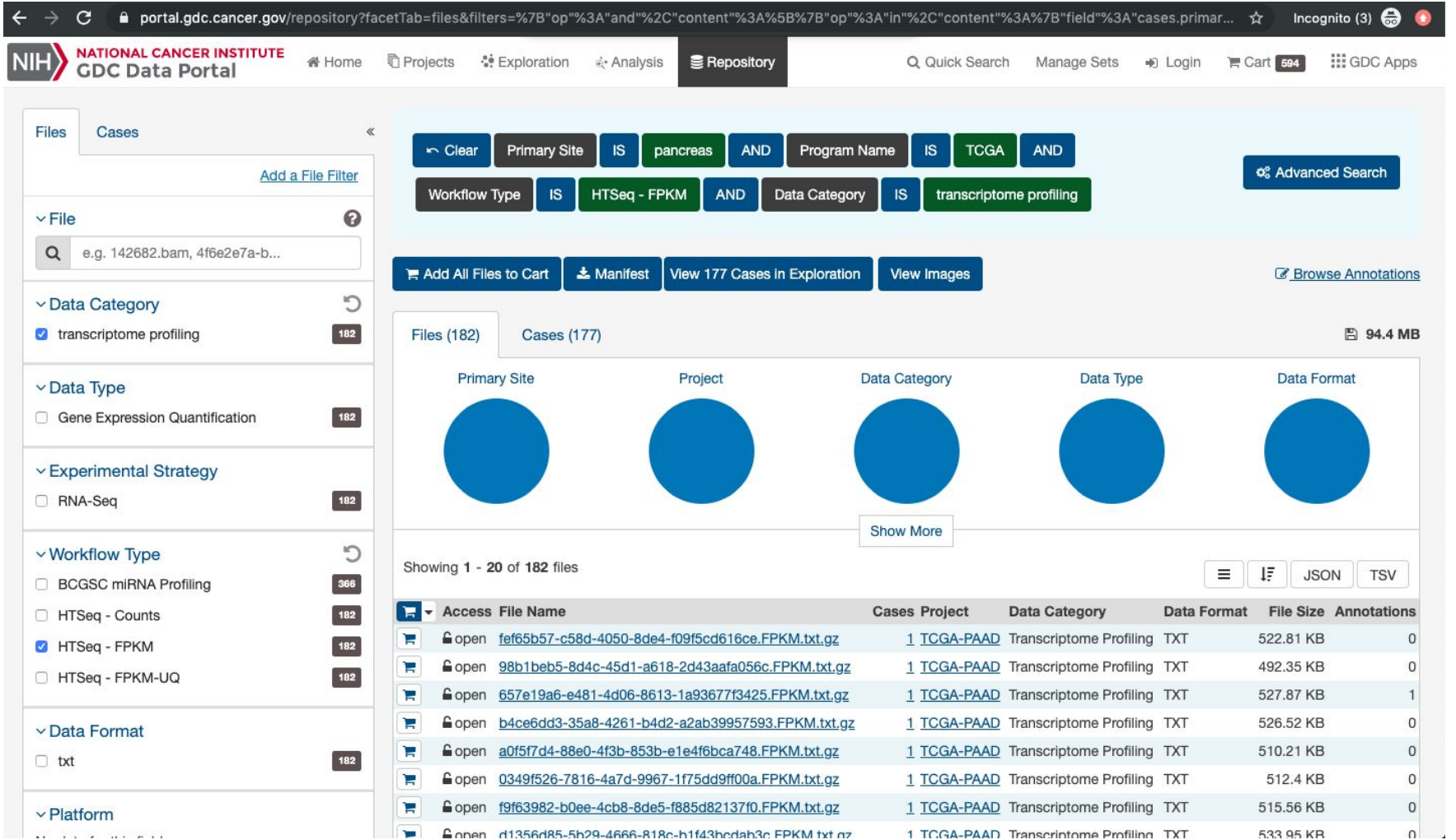
11,000

PATIENTS

...using

7

DIFFERENT DATA TYPES


portal.gdc.cancer.gov/repository?facetTab=files&filters=%7B"op"%3A"and"%2C"content"%3A%5B%7B"op"%3A"in"%2C"content"%3A%7B"field"%3A"cases.primar...
Incognito (3)
NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Quick Search
Manage Sets
Login
Cart 594
GDC Apps
Files
Cases
Add a File Filter
File
e.g. 142682.bam, 4f6e2e7a-b...
Data Category
transcriptome profiling 182
Data Type
Gene Expression Quantification 182
Experimental Strategy
RNA-Seq 182
Workflow Type
BCGSC miRNA Profiling 366
HTSeq - Counts 182
HTSeq - FPKM 182
HTSeq - FPKM-UQ 182
Data Format
txt 182
Platform
Clear
Primary Site IS pancreas AND Program Name IS TCGA AND
Workflow Type IS HTSeq - FPKM AND Data Category IS transcriptome profiling
Advanced Search
Add All Files to Cart
Manifest
View 177 Cases in Exploration
View Images
Browse Annotations
Files (182)
Cases (177)
94.4 MB
Primary Site
Project
Data Category
Data Type
Data Format
Show More
Showing 1 - 20 of 182 files
JSON
TSV
Access File Name Cases Project Data Category Data Format File Size Annotations
open fef65b57-c58d-4050-8de4-f09f5cd616ce.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 522.81 KB 0
open 98b1beb5-8d4c-45d1-a618-2d43aafa056c.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 492.35 KB 0
open 657e19a6-e481-4d06-8613-1a93677f3425.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 527.87 KB 1
open b4ce6dd3-35a8-4261-b4d2-a2ab39957593.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 526.52 KB 0
open a0f5f7d4-88e0-4f3b-853b-e1e4f6bca748.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 510.21 KB 0
open 0349f526-7816-4a7d-9967-1f75dd9ff00a.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 512.4 KB 0
open f9f63982-b0ee-4cb8-8de5-f885d82137f0.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 515.56 KB 0
open d1356d85-5b29-4666-818c-b1f43bcdab3c.FPKM.txt.gz 1 TCGA-PAAD Transcriptome Profiling TXT 533.95 KB 0

FILES
182

CASES
177

FILE SIZE
94.4 MB

File Counts by Project

Project	Cases (n=177)	Files (n=182)	File Size (∑=94.4 MB)
TCGA-PAAD	177	182	94.4 MB

File Counts by Authorization Level

Level	Files (n=182)	File Size (∑=94.4 MB)
Authorized	182	94.4 MB

How to download files in my Cart?

Download Manifest:
Download a manifest for use with the GDC Data Transfer Tool. The GDC Data Transfer Tool is recommended for transferring large volumes of data.

Download Cart:
Download Files in your Cart directly from the Web Browser.

Biospecimen | Clinical | Sample Sheet | Metadata | Download | Remove From Cart

Cart Items

Showing **1** - **20** of **182** files

TSV

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	fef65b57-c58d-4050-8de4-f09f5cd616ce.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	522.81 KB	0
open	98b1beb5-8d4c-45d1-a618-2d43aafa056c.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	492.35 KB	0
open	657e19a6-e481-4d06-8613-1a93677f3425.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	527.87 KB	1
open	b4ce6dd3-35a8-4261-b4d2-a2ab39957593.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	526.52 KB	0
open	a0f5f7d4-88e0-4f3b-853b-e1e4f6bca748.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	510.21 KB	0
open	0349f526-7816-4a7d-9967-1f75dd9ff00a.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	512.4 KB	0
open	f9f63982-b0ee-4cb8-8de5-f885d82137f0.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	515.56 KB	0
open	d1356d85-5b29-4666-818c-b1f43bcdab3c.FPKM.txt.gz	1	TCGA-PAAD	Transcriptome Profiling	TXT	533.95 KB	0

Data

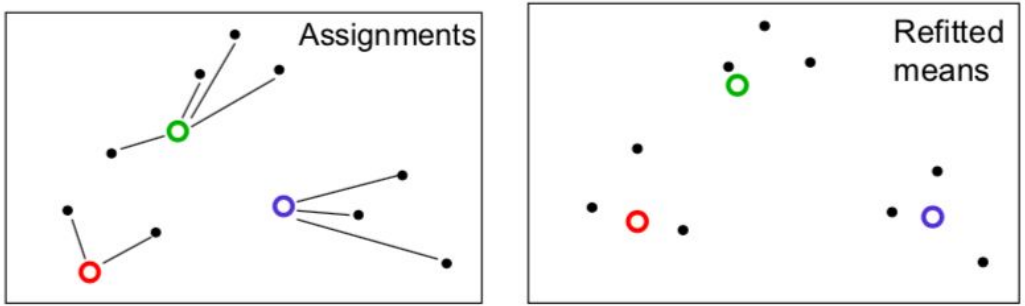
- Clinical data on cancer patient: age, sex, stage, treatment, *etc*
- Typically hundreds of patients for each cancer type
- Multiple types of genomic data
- We will focus on RNA sequencing
- 60,000 gene expression levels per patient
    - Fragments per Kilobase Million (FPKM): A measure of the amount of RNA fragments detected for a specific gene, adjusted for sequencing depth and gene length
    - Non-negative right-skewed variable, typically transformed for analysis *i.e.* log(FPKM + 1)
- Higher FPKM => the gene has higher expression

Question

- Are there distinct biological “subtypes” within a cancer type?
- May have different clinical behavior and benefit from different treatments

K-means

- Initialization: randomly initialize cluster centers
- The algorithm iteratively alternates between two steps:
  - Assignment step: Assign each data point to the closest cluster
  - Refitting step: Move each cluster center to the center of gravity of the data assigned to it



The K-means Algorithm

- Initialization: Set K cluster means $\mathbf{m}_1, \ldots, \mathbf{m}_K$ to random values
- Repeat until convergence (until assignments do not change):
  - Assignment: Each data point $\mathbf{x}^{(n)}$ assigned to nearest mean

$$\hat{k}^n = arg\min_k d(\mathbf{m}_k, \mathbf{x}^{(n)})$$

(with, for example, L2 norm: $\hat{k}^n = arg\min_k ||\mathbf{m}_k - \mathbf{x}^{(n)}||^2$)

and Responsibilities (1-hot encoding)

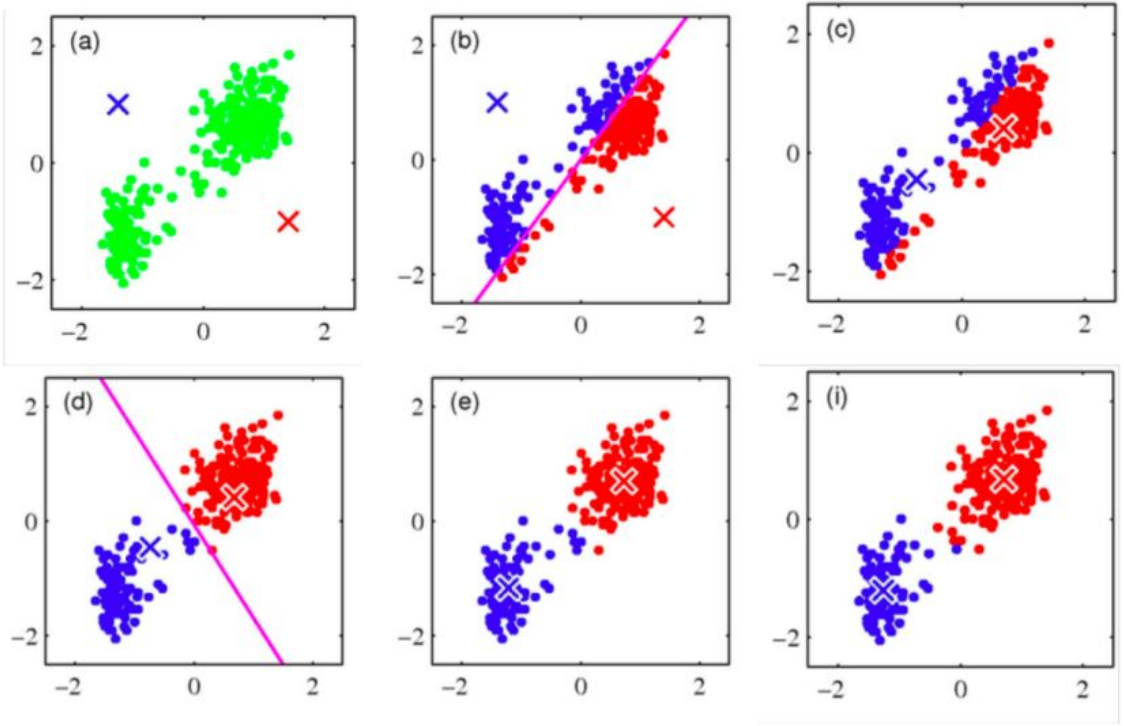
$$r_k^{(n)} = 1 \longleftrightarrow \hat{k}^{(n)} = k$$

  - Refitting: Model parameters, means are adjusted to match sample means of data points they are responsible for:

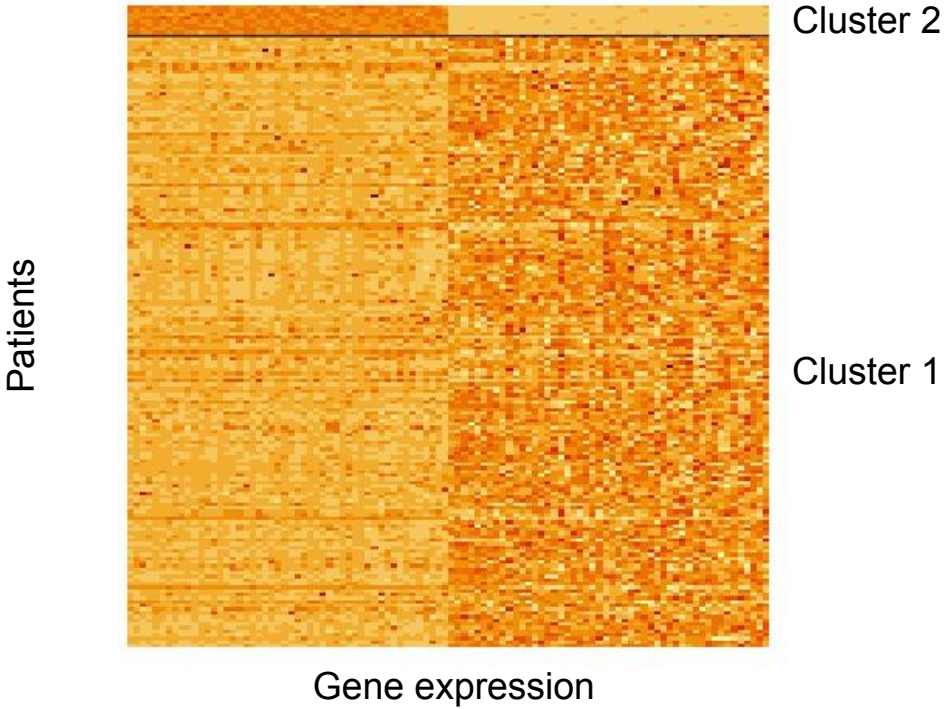
$$\mathbf{m}_k = \frac{\sum_n r_k^{(n)} \mathbf{x}^{(n)}}{\sum_n r_k^{(n)}}$$

Example

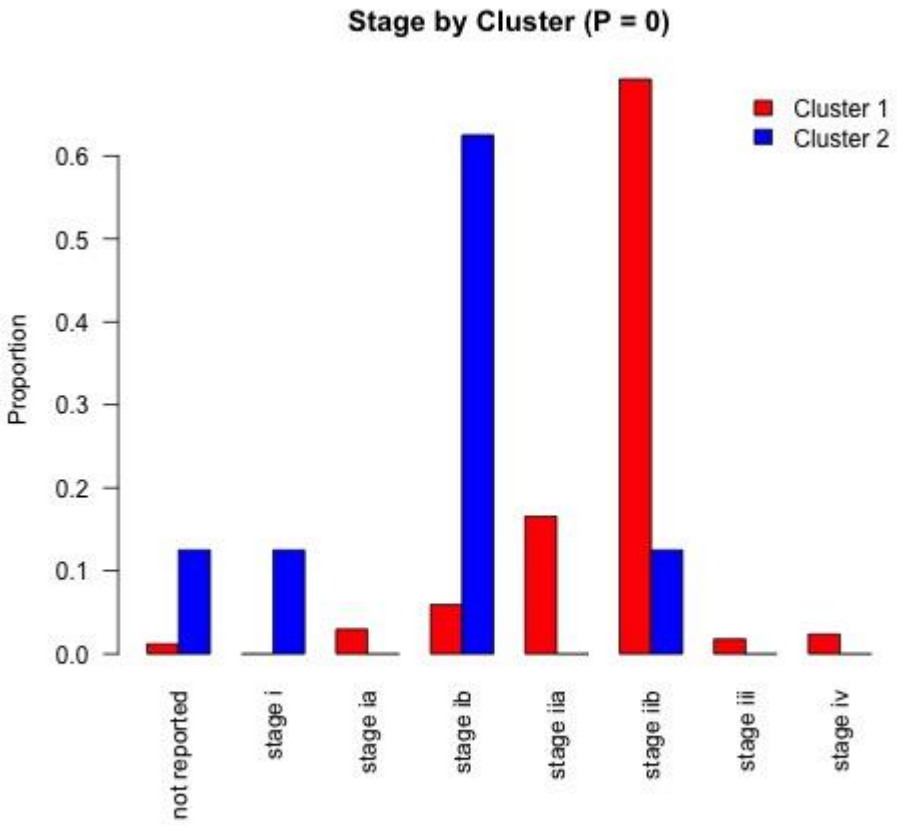
- Example of using K-means (K=2) on Old Faithful dataset.



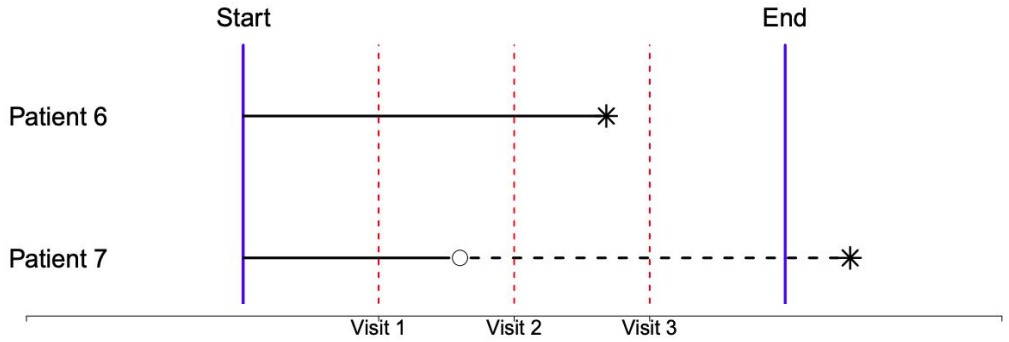
Heatmap of the Top 100 Genes with the Furthest Centres



Some of these are important cancer genes i.e. S100P and other are reasonable candidates (i.e. interleukins, trypases etc)


Stage by Cluster (P = 0)
Cluster 1
Cluster 2
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
not reported
stage i
stage ia
stage ib
stage iia
stage iib
stage iii
stage iv

Survival Analysis



- Notation (i denotes the patient)
  - ▷ T_i^* 'true' time-to-event
  - ▷ because of censoring we do not always observe T_i^*
  - ▷ C_i the censoring time

- Available data for each patient
  - ▷ observed event time: $T_i = \min(T_i^*, C_i)$
  - ▷ event indicator: $\delta_i = 1$ if event; $\delta_i = 0$ if censored

Kaplan-Meier Curves

$S(t) = \Pr(T > t)$

Let $t_1, t_2, \ldots, t_k$ denote the unique event times in the sample at hand

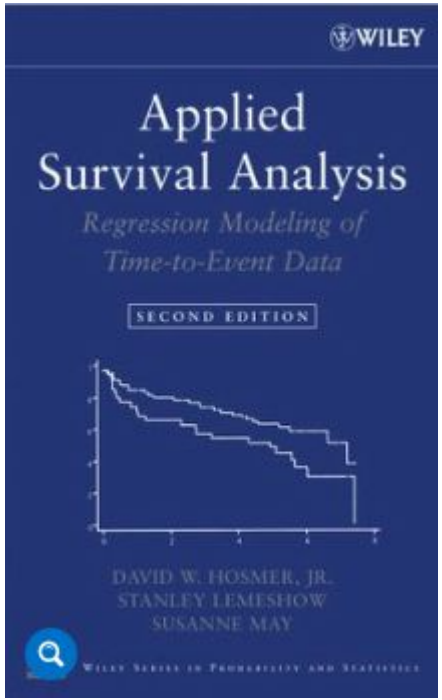
$$\hat{S}_{KM}(t) = \prod_{i:\, t_i \leq t} \frac{r_i - d_i}{r_i}$$

where d_i is the number of events at time t_i, and r_i the number of patients still at risk at time t_i

Resource

U of T library link

R code link



Lukemia Survival Times

Survival in patients with Acute Myelogenous Leukemia. The question at the time was whether the standard course of chemotherapy should be extended ('maintainance') for additional cycles.

```
##   time status           x
## 1    9      1 Maintained
## 2   13      1 Maintained
## 3   13      0 Maintained
```

`time`: survival or censoring time

`status`: censoring status

`x`: maintenance chemotherapy

Lukemia Survival Times

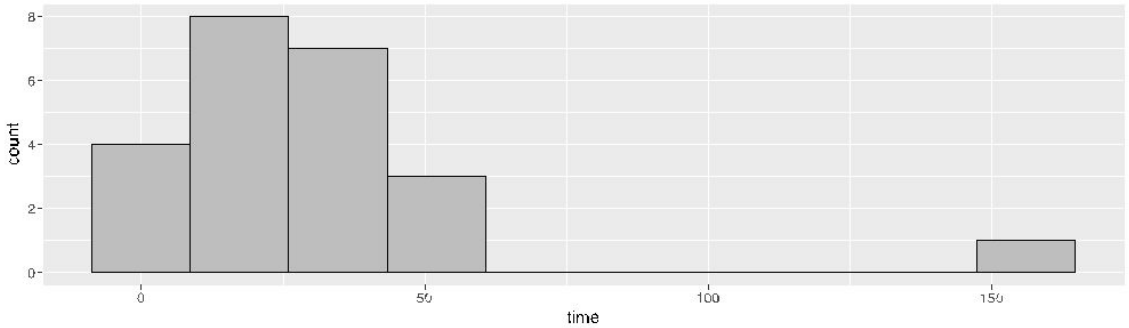
```
##   time status          x
## 1    9      1 Maintained
## 2   13      1 Maintained
## 3   13      0 Maintained
```

- The third observation has a status of 0.
- Person was followed for 13 months and after that was lost to follow up.
- So we only know that the patient survived AT LEAST 13 months, but we have no other information available about the patient's status.
- This type of censoring (also known as "right censoring") makes linear regression an inappropriate way to analyze the data due to censoring bias.

Lukemia Survival Times

```
library(survival)
library(tidyverse)

leukemia %>% ggplot(aes(time)) +
  geom_histogram(colour = "black",
                 fill = "grey", bins = 10)
```



Survival Analysis in R

```
Surv(time, status)
```

creates the dependent variable for a survival object in a survival model.

Survival Analysis in R

```
library(broom)
km <- survfit(Surv(time, status) ~ 1, data = leukemia)
tidy(km) %>% head(3) %>% rename(survival = estimate)
```

```
## # A tibble: 3 x 8
##    time n.risk n.event n.censor survival std.error conf.high conf.lo
##   <dbl>  <dbl>   <dbl>    <dbl>    <dbl>     <dbl>     <dbl>    <dbl
## 1     5     23       2        0    0.913    0.0643     1        0.80
## 2     8     21       2        0    0.826    0.0957     0.996    0.68
## 3     9     19       1        0    0.783    0.110      0.971    0.63
```

- At time 9 the probability of survival is:

```
((23 - 2)/23)*((21 - 2)/21)*((19 - 1)/19)
```

```
## [1] 0.7826087
```

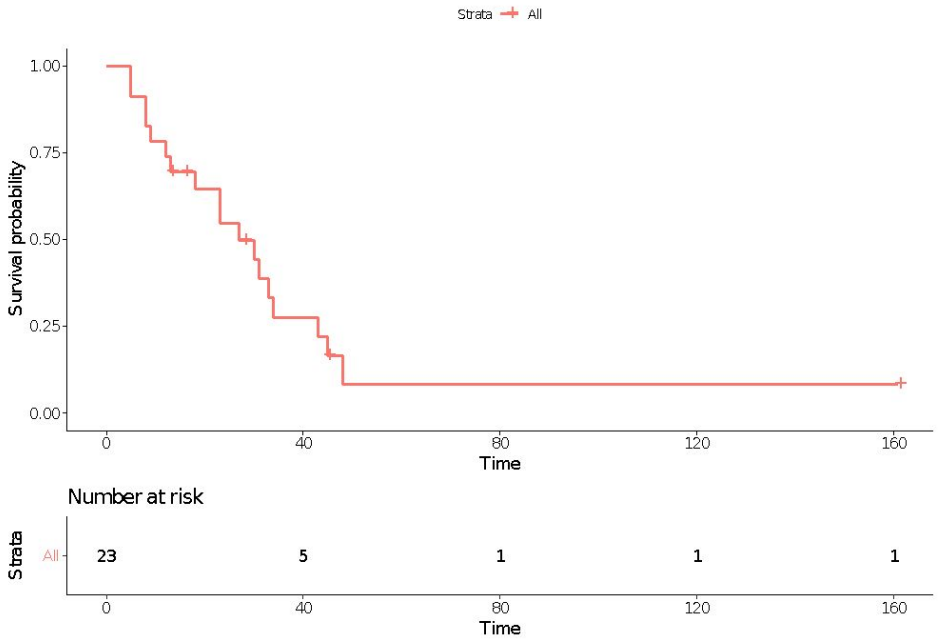
```
tidy(km) %>% head(6) %>% rename(survival = estimate)
```

```
## # A tibble: 6 x 8
##    time n.risk n.event n.censor survival std.error conf.high conf.lo
##   <dbl>  <dbl>   <dbl>    <dbl>    <dbl>     <dbl>     <dbl>    <dbl
## 1     5     23       2        0    0.913    0.0643         1    0.80
## 2     8     21       2        0    0.826    0.0957     0.996    0.68
## 3     9     19       1        0    0.783     0.110     0.971    0.63
## 4    12     18       1        0    0.739     0.124     0.942    0.58
## 5    13     17       1        1    0.696     0.138     0.912    0.53
## 6    16     15       0        1    0.696     0.138     0.912    0.53
```

- At time 16 there are 17 - 1 - 1 = 15 people at risk since one person was censored at time 13.

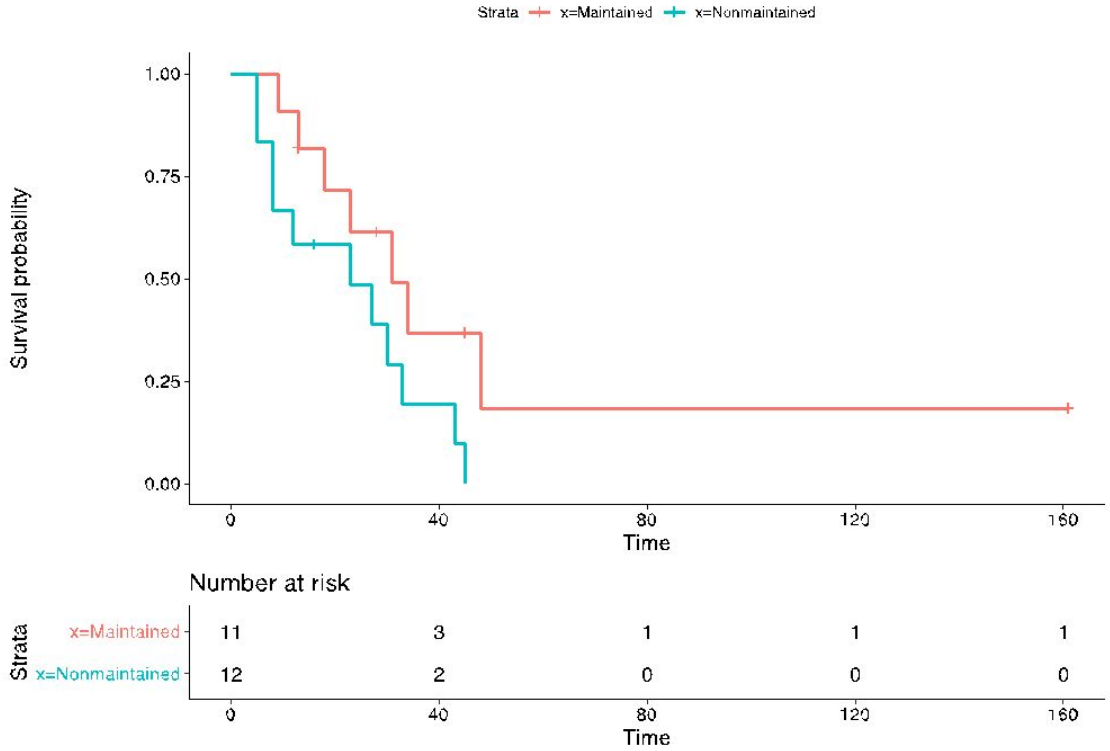
Kaplan - Meir Analysis

```
library(survminer)
ggsurvplot(km, data = leukemia, risk.table = T)
```



Comparing Survival Curves

```
kmx <- survfit(Surv(time, status) ~ x, data = leukemia)
ggsurvplot(kmx, data = leukemia, risk.table = T)
```



Comparing Survival Curves

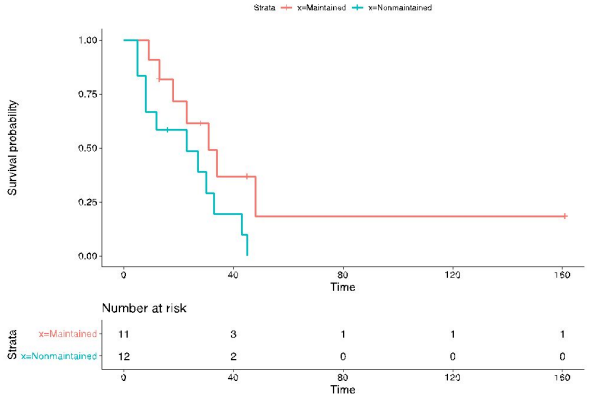
Log-Rank Test

H_0: There is no difference in the survival function between those who were on maintenance chemotherapy and those who weren't on maintenance chemotherapy.

H_a: There is a difference in the survival function between those who were on maintenance chemotherapy and those who weren't on maintenance chemotherapy.

Comparing Survival Curves

```
kmx <- survfit(Surv(time, status) ~ x, data = leukemia)
ggsurvplot(kmx, data = leukemia, risk.table = T)
```



Comparing Survival Curves

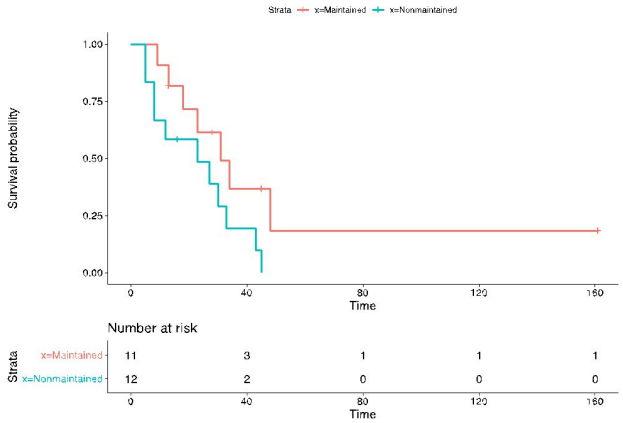
Log-Rank Test

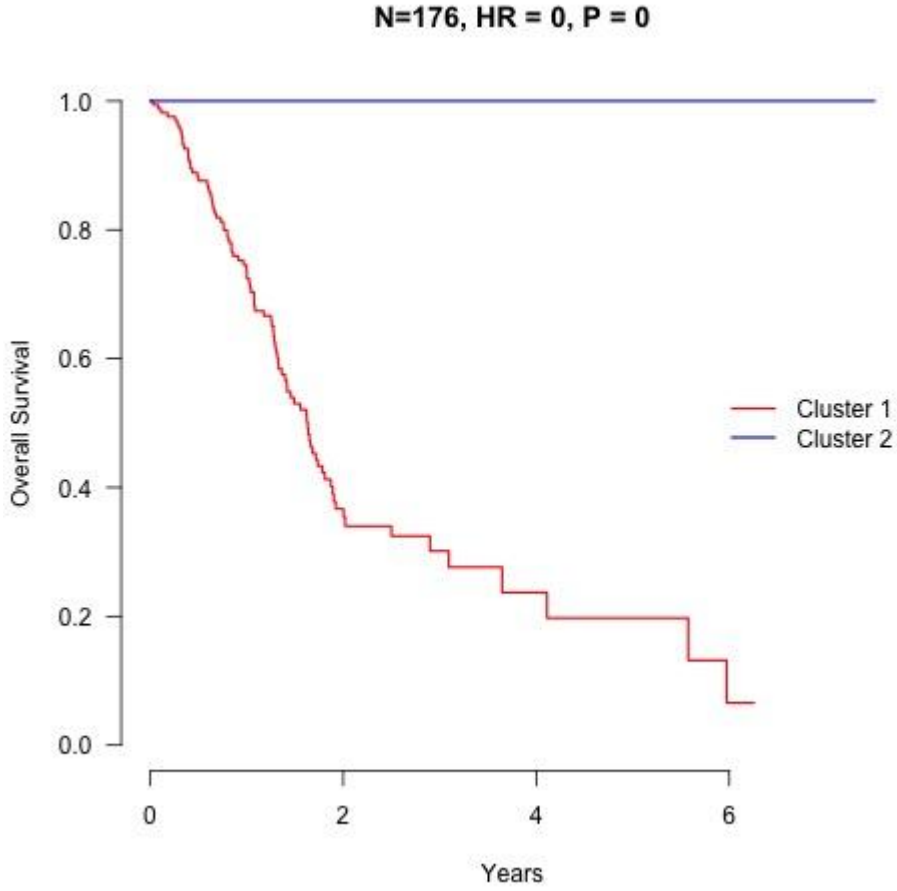
```
survdiff(Surv(time, status) ~ x, data = leukemia)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ x, data = leukemia)
##
##                 N Observed Expected (O-E)^2/E (O-E)^2/V
## x=Maintained    11        7    10.69      1.27       3.4
## x=Nonmaintained 12       11     7.31      1.86       3.4
##
##  Chisq= 3.4  on 1 degrees of freedom, p= 0.07
```

Comparing Survival Curves

```
kmx <- survfit(Surv(time, status) ~ x, data = leukemia)
ggsurvplot(kmx, data = leukemia, risk.table = T)
```




N=176, HR = 0, P = 0
1.0
0.8
0.6
0.4
0.2
0.0
Overall Survival
0
2
4
6
Years
Cluster 1
Cluster 2

SCIENTIFIC REPORTS
nature research

OPEN

Machine Learning and Network Analyses Reveal Disease Subtypes of Pancreatic Cancer and their Molecular Characteristics

Musalula Sinkala*, Nicola Mulder & Darren Martin

Questions

- How many clusters?
- Continuous latent features?
- Training and test sets?
- Supervised or semi-supervised approaches?
- Other cancers
  - TCGA-LIHC, TCGA-LUAD
  - All cancers together
- Tumour cellularity and sampling issues

Weill Cornell Medicine
Englander Institute for Precision Medicine
Weill Cornell Precision Medicine Program (USA, multiple cancers)

Swiss Oncology and Cancer Immunology Breakthrough Platform (Switzerland, multiple cancers)

SAMSUNG MEDICAL CENTER
Personalized Genomic Characterisatiopn of Korean Lung Cancers (Korea)

elixir
Precision Medicine for esophageal Cance (UK)

PBCP
Personalised Breast Cancer Program (United Kingdom)

SNUH
Korean Myeloma Project (Korea)

Pan Prostate Cancer Group (United Kingdom)

국립암센터
Korean Rare Cancers Project (Korea)

OICR
Enhanced Pancreatic Cancer Profiling for Individualised Care (Canada)

BC Cancer personalised OncoGenomics Program (Canada, multiple cancers)

Papillary Thyroid Cancer Project (Saudi Arabia)

CCGC
Chinese Cancer Genome Consortium (China, colorectal cancer)

CANCER RESEARCH UK | CANCER CHALLENGE
Mutographs Study (UK, France, multiple cancers)

PRECISION PANCREAS
Precision Panc (UK, pancreatic cancer)

P1000
1000 Polyethnic Study (USA, multiple cancers)

UMG G-CCC
European Peripheral T Cell Lymphoma Study (Germany)

China Diffuse Gastric Cancer Study (China)

TRACERx Study (UK, lung cancer)

Oesophageal Squamous Cell Carcinoma Study (China)

AMED
Genomic Medicine for Asia Prevalent Cancers (Japan, multiple cancers)

Profiling Orphan Neoplasms for Treatment (Italy, multiple cancers)

ARGO Accelerating Research in Genomic Oncology

International Cancer Genome Consortium