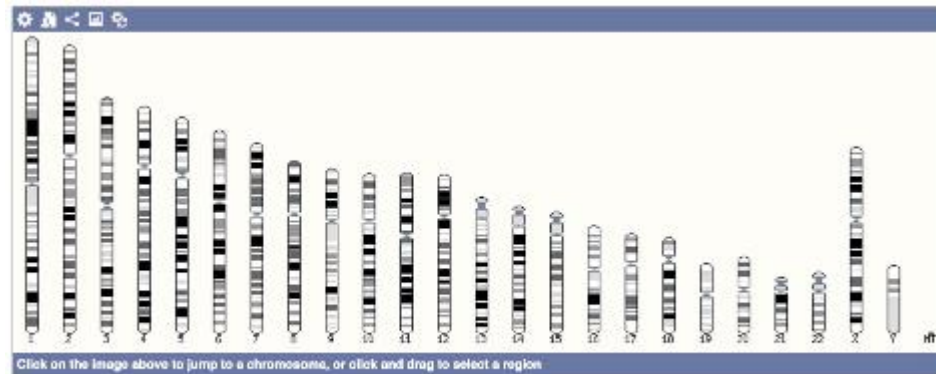


R을 활용한 TCGA 암유전체 데이터 분석

- 변이란 무엇인가?
 - 유전체란?
 - 생식세포 변이, 체세포 변이
 - 변이의 종류 (SNP, SNV, fusion, CNV..)
- 변이를 연구하는 이유는?
 - 유전질환
 - 암
 - Precision medicine
- 변이를 어떻게 검출할 것인가
 - NGS 이론
 - 변이 검출 파이프라인
 - 변이 들여다보기: IGV (실습 1)

What are genetic variants?

The human genome - basic stats



- 3.096 billion base pairs (haploid)
- 20,441 protein coding genes
- 198,002 coding transcripts (isoforms of a gene that each encode a distinct protein product)

Summary

Assembly	GRCh38.p7 (Genome Reference Consortium Human Build 38), R50C Assembly GCA_000001405.226, Dec 2013
Database version	87.86
Base Pairs	3,547,762,741
Golden Path Length	3,085,648,726
Genebuild by	Ensembl
Genebuild method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jun 2018
Gencode version	GENCODE 25

Gene counts (Primary assembly)

Coding genes	20,441 (incl 528 readthrough)
Non coding genes	82,218
Small non coding genes	5,052
Long non coding genes	14,757 (incl 214 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,806 (incl 5 readthrough)
Gene transcripts	198,002

http://uswest.ensembl.org/Homo_sapiens/Location/Genome

What is genetic variation?

- Differences in DNA content or structure among individuals
 - Any two individuals have ~99.5% identical DNA.
- But the human genome is big - each haploid set of 23 chromosomes has 3.1 billion nucleotides.
 - There are >100,000,000 known genetic variants in the human genome
- Effectively infinite combinations of alleles. The details matter.

~99.8% identical DNA (differ at 1/ 620 - 1/750 bp)



99% identical DNA



CGCAAATTTGCCGGATTTTCCTTTGCTGTTTCCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCT
TTTTCTTTGACCTCTTCTTTCTGTTTCATGTGTATTTGCTGTCTCTTAGCCAGACTTCCCGTGTCTTTCCACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTG
GTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCTGTGGAGAGGAGCCATGCCTAGAGTG
GGATGGGCCATTGTTTCATCTTCTGGCCCCGTGTGTCTGCATGTAACCTAATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTT
CTCTCACAACCTAGGCCAGTAAGTAGTGCTTGTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTGCTGCCA
/TCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGC/GCTCAGCAGGTCTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCA
GGCCATCGCTGCCACAGAACCCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCAGG
CACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCCACGGCCCCAGGCTCCTGTCTCCCCCAGGTGTGTGGTGATGCCAGGCAT
GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG/AAGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGAC
AGTGAGTGGGAGTGGCGTCGCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGGCTTCGATGCCCTCCACACCCTCTTGATCTTCCCTGTGATGT
CATCTGGAGCCCTGCTGCTTGCGGTGGCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCAGGGAAAGCTACA/TAGCAGCAAACAGTCTGC
ATGGGTTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCAAGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAG
CCTGGCTGCCTCCAGCTGGGTGACAGACAGGGGCTGGAGAAGGGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGG
ATGCACTGTTGGGGAGGCAGCTGTAACCAAAGCCTTAGCCTCTGTTCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACC
AGTGATACACCCGGCACCCCTGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTC
ACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCAGTTTCACCAGAAGTAGGCCTCTTCCCTGACAGGC/
TAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCA
GGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCTTGGCCATCCGTGA
GATCTTCCCAGGTGTGCCGTTTTCTCTGGAAGCCTCTTAAGAACACAGTGCGCAGGCTGGGTGGAGCCGTCCCCCATGGAGCACAGGCA/GGACAGAAGTCCCCG
CCCCAGCTGTGTGGCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCTCCCAAGGAAGTAGGTCTGAGCAGCTTGTCC
TGGCTGTGTCCATGTCAGAGCAACGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGATTCCCAGTCGTCCTCCTCTGCCTGTGGCT
GCTGCGGTGGCGGCAGAGGAGGATGGAGTCTGACACGCGGGCAAAGGCTCCTCCGGGCCCTCACCAGCCCCAGGTCCTTTCCCAGAGATGCCTGGAGGGAAAAGG
CTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTCCCCCAGCCCCGGG/A/GACTTAAATACAGGAAGAAAAAGGCAGGACAGAATTACAAGGTGCTGGCCCAGGGCG
GGCAGCGGCCCTGCCTCCTACCTTGCCTCATGACCGGAGCCATAGCCCAGGCAGGAGGGCTGAGGACCTCTGGTGGCGGCCAGGGCTTCCAGCATGTGCCCTA
GGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACCCGGTGGACTCAGGGCTGGAGGGGAGGAGGCGATCTTGCCCAAGGCCCTCCGACTGCAA
GCTCCAGGGCCCGCTCACCTTGCTCCTGCTCCTTCTGCTGCTGCTTCTCCAGCTTTCGCTCCTTCATGCTGCGCAGCTTGGCCTTGCCGATGCCCCAGCTTGGCGG
ATGGACTCTAGCAGAGTGGCCAGCCACCGGAGGGGTCAACCACTTCCC

Types of genetic variation

ctc**c**gag
ctc**t**gag

Single-nucleotide
polymorphisms
(**SNPs**)

“DNA spelling mistakes”

ctc--ag
ctc**tg**ag

Insertion-deletion
polymorphisms
(**INDELs**)

*“extra or missing
DNA”*

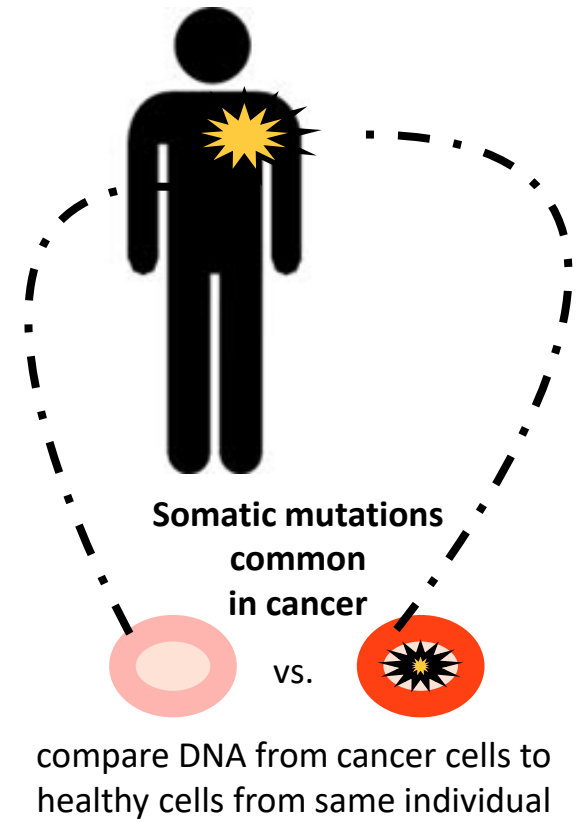
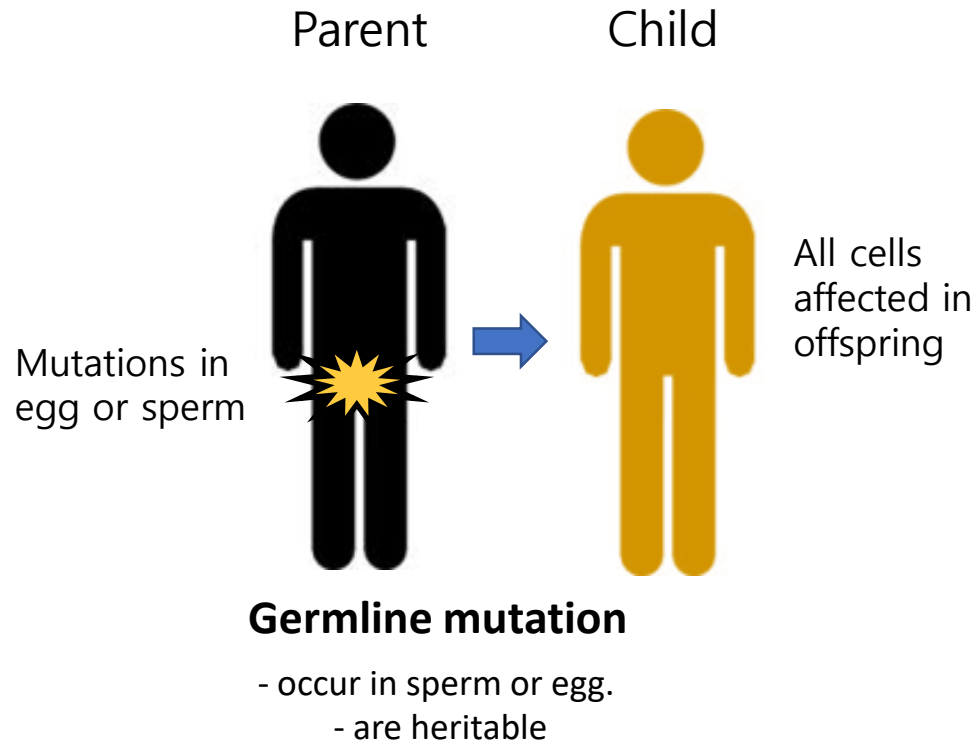
ctc ctcaag
  ag

Structural
variants
(**SVs**)

*“Large blocks of extra, missing
or rearranged
DNA”*

Mutation != Polymorphism (or SNP)

Somatic mutations



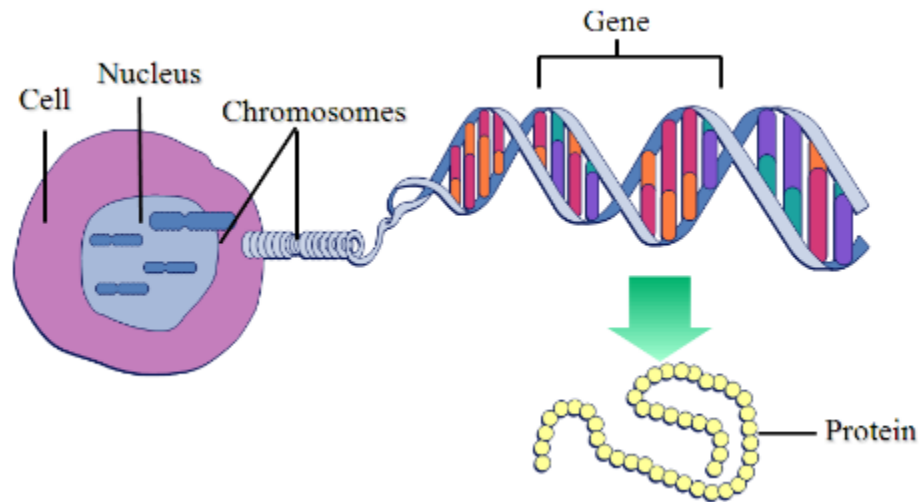
Types of Genetic Alterations in Cancer

- Subtle alterations
- Chromosome number changes
- Chromosomal translocation
- Amplifications
- Exogenous sequences

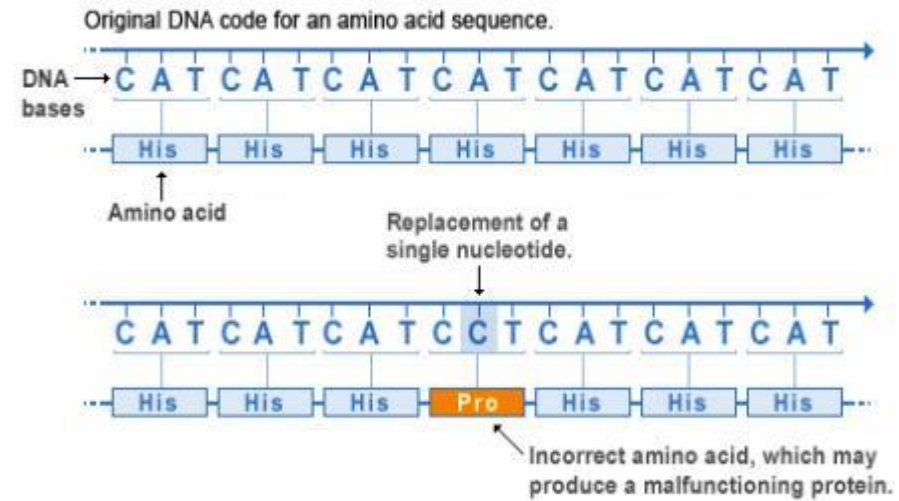
Subtle Alterations

- Small deletions
- Insertions
- Single base pair substitutions
 - (Point mutations)

Missense mutation



Missense mutation



Point Mutations

Normal THE BIG RED DOG RAN OUT.

Missense THE BIG RAD DOG RAN OUT.

Nonsense THE BIG RED.

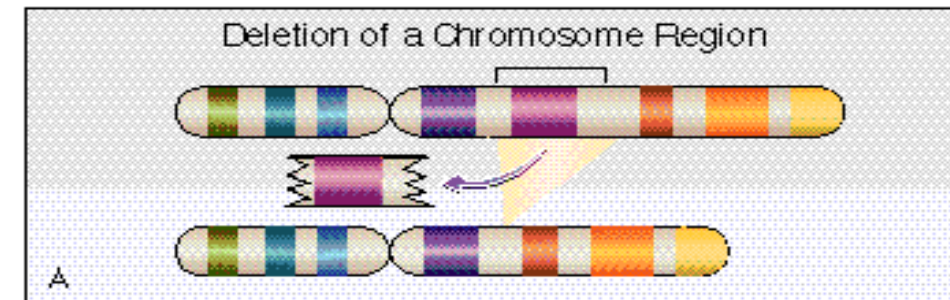
Frameshift (deletion) THE BRE DDO GRA.

Frameshift (insertion) THE BIG RED ZDO GRA.

Point mutation: a change in a single base pair

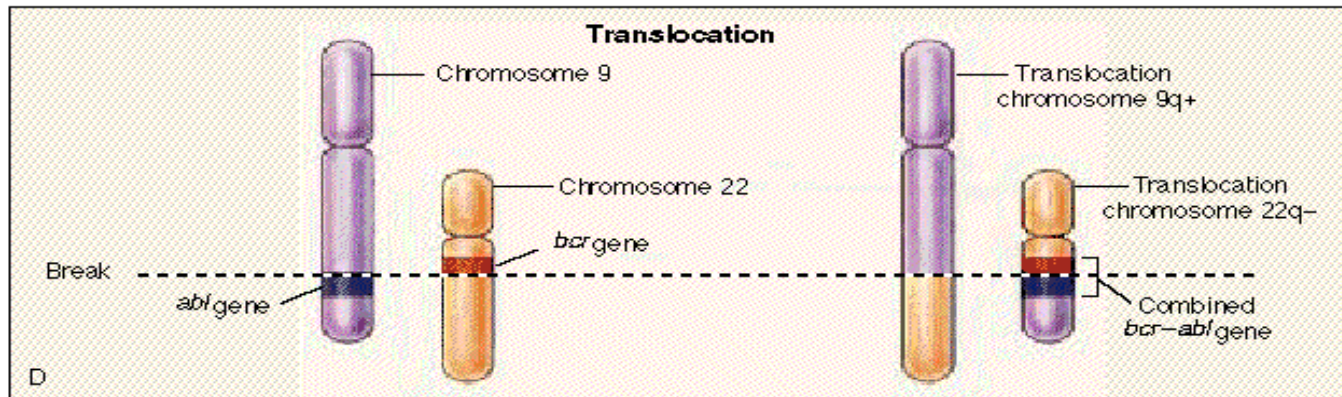
Chromosome Number Changes

- Aneuploidy
 - somatic losses or gains
- Whole chromosome losses often are associated with a duplication of the remaining chromosome.
- LOH
 - loss of heterozygosity



Chromosome Translocations

- Random translocations
 - breast, colon, prostate (common epithelial tumors)
- Non-random translocations
 - leukemia, lymphoma



Amplifications

- Seen only in cancer cells
 - 5 to 100-fold multiplication of a small region of a chromosome
- "Amplicons"
 - contain one or more genes that enhance proliferation
- Generally in advanced tumors

Why do we care?

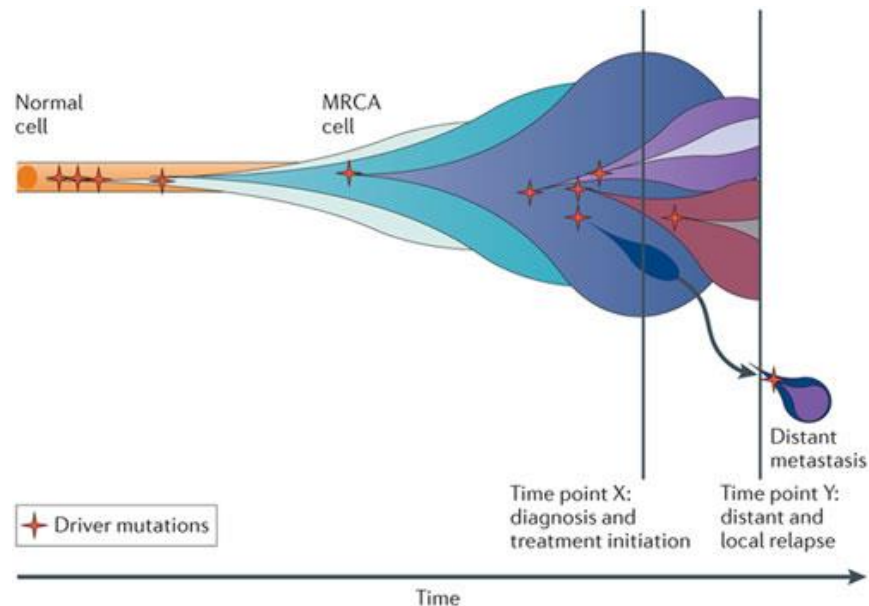
Driver vs. Passenger Mutation

- **Driver mutation**

- A mutation that gives a selective advantage to a clone in its microenvironment, through either increasing its survival or reproduction.
- Driver mutations tend to cause clonal expansions.

- **Passenger mutation**

- A mutation that has no effect on the fitness of a clone but may be associated with a clonal expansion because it occurs in the same genome with a driver mutation.
- This is known as a hitchhiker in evolutionary biology.



A driver and passengers

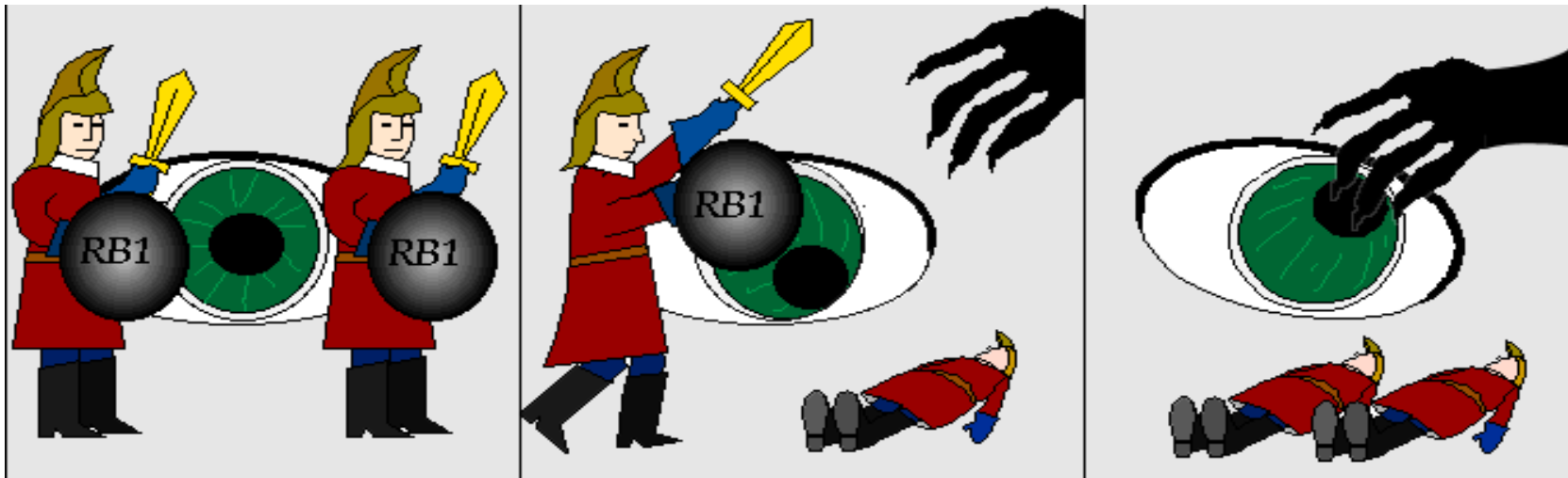


Features of Retinoblastoma



- 1 in 20,000 children
- Most common eye tumor in children
- Occurs in heritable and nonheritable forms
- Identifying at-risk infants substantially reduces morbidity and mortality

Knudson's "Two-Hit" Model for Retinoblastoma



Normal
2 intact copies

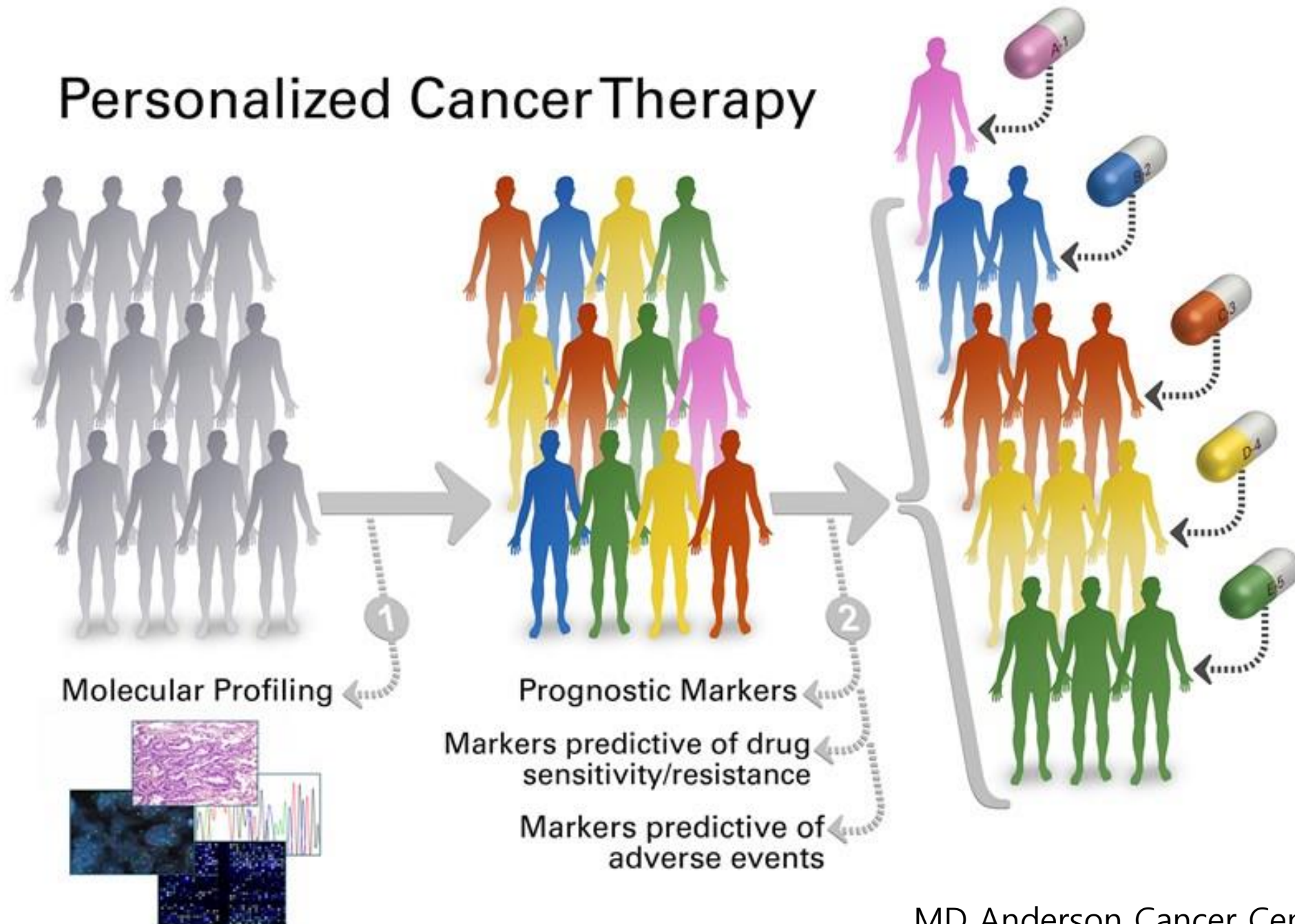
Modified from *Time*, Oct. 27, 1986

Predisposed
1 intact copy
1 mutation

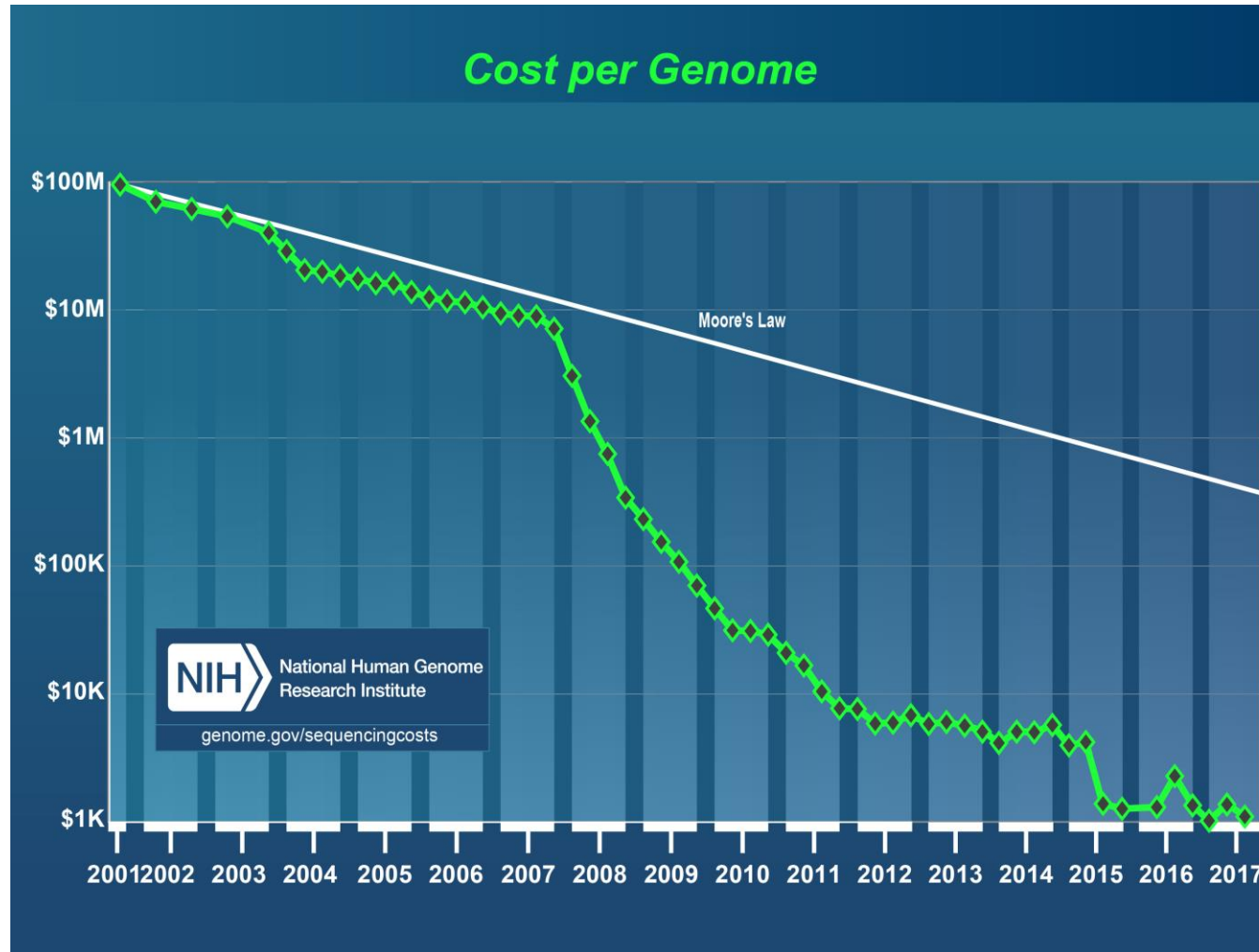
Affected
Loss of both copies

ASCO

Personalized Cancer Therapy



Cost of sequencing



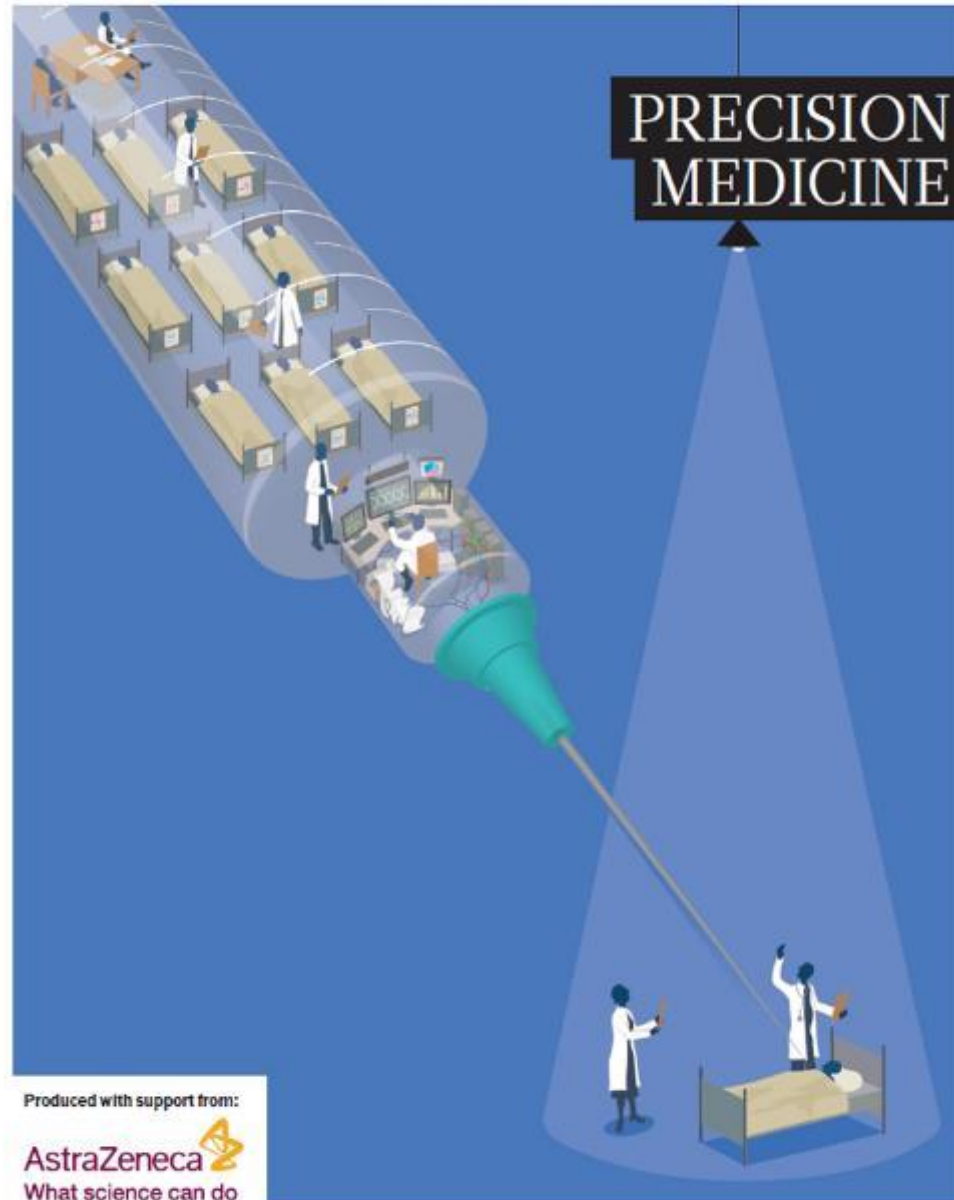
NovaSeq: \$100

NATIONAL CANCER INSTITUTE PRECISION MEDICINE IN CANCER TREATMENT

Discovering unique therapies that treat an individual's cancer based on the specific genetic abnormalities of that person's tumor.

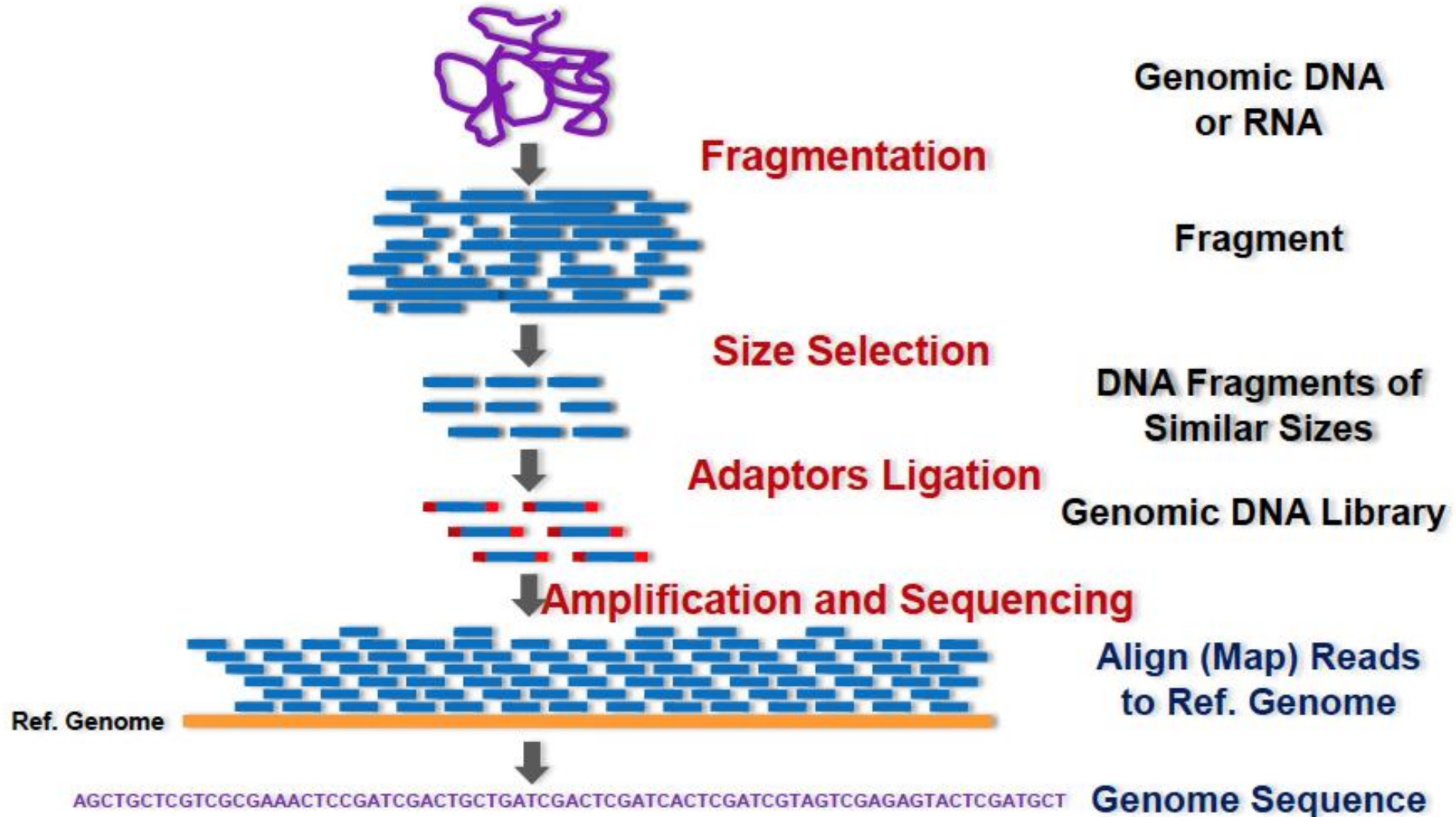


natureINSIGHT

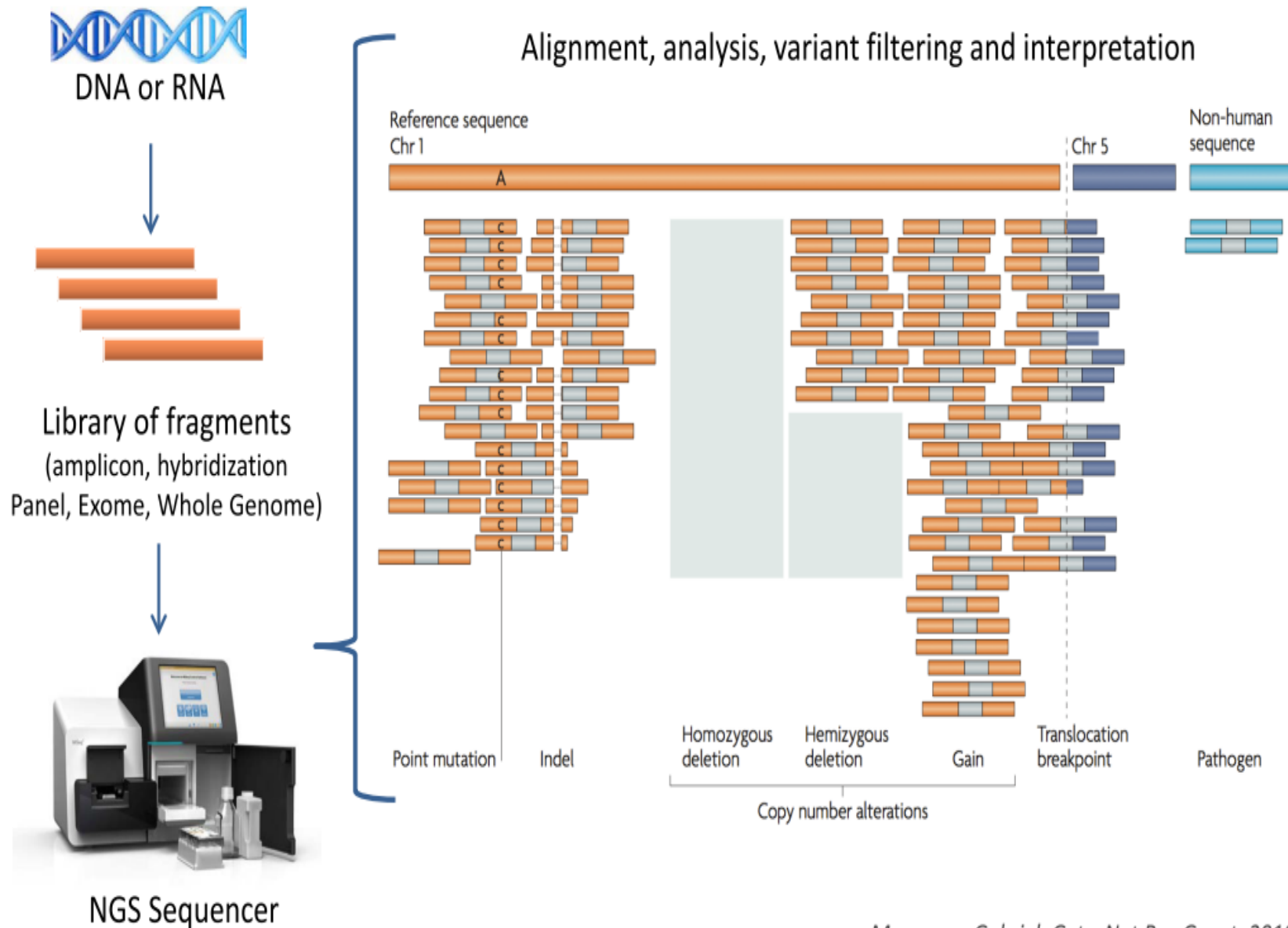


How to Detect Variants?

(Next-Generation Sequencing)



NGS – SNV, indel, CNV, SV



실습

TCGA data visualization using R

Load data

```
brca.cnv <- read.delim("TCGA_BRCA_CNV_processed.txt")  
brca.snv <- read.delim("TCGA_BRCA_SNV_processed.txt")  
brca.expr <- read.delim("TCGA_BRCA_Expr_processed.txt")
```

실습 문제

- 주어진 dataset에서 ERBB2의 CNV가 3보다 큰 tumor sample들의 ERBB2 expression의 평균값을 구하시오.
- 94개 이상의 tumor sample들에서 CNV가 2 이상인 유전자를 도출하시오.

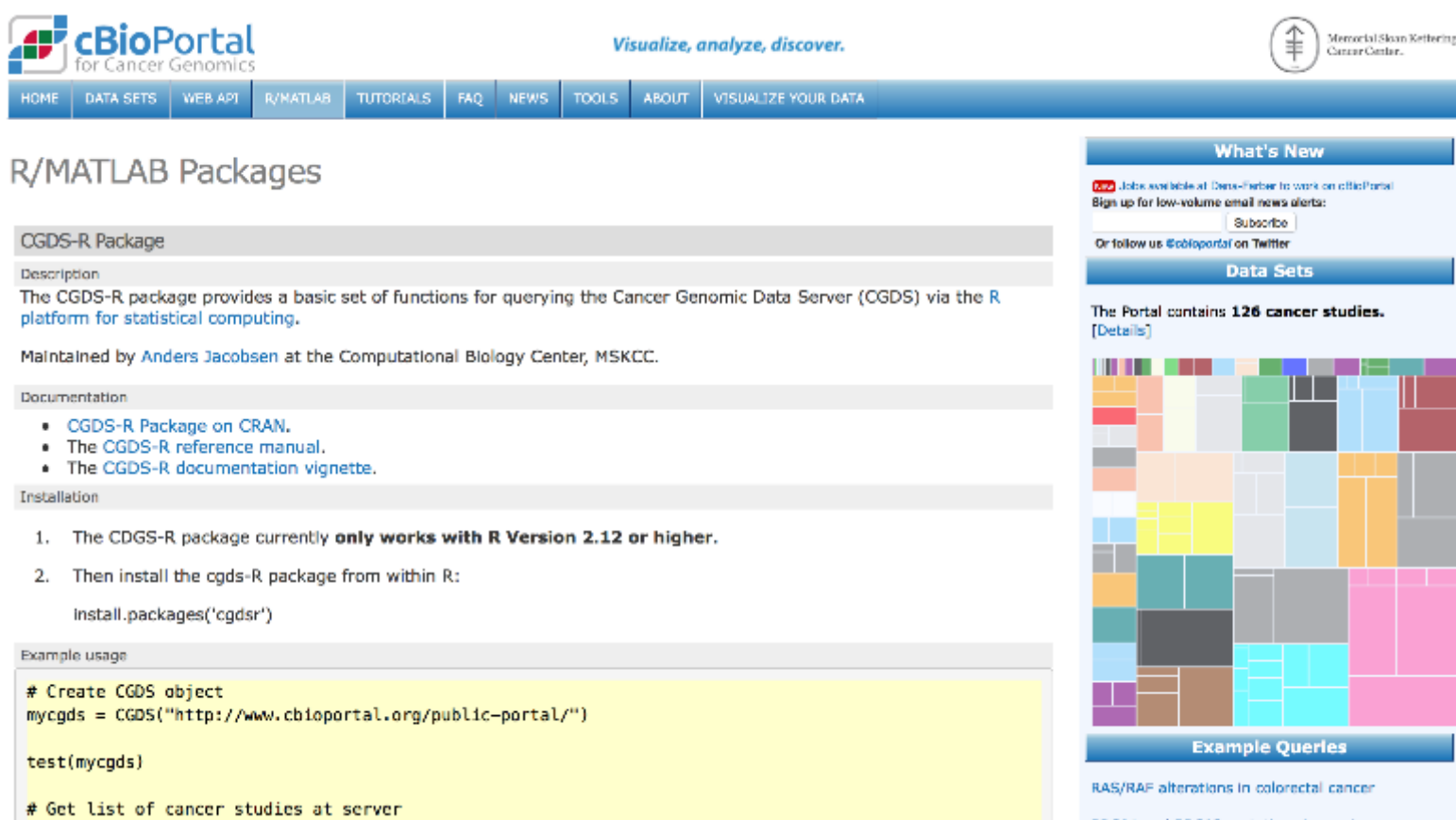
Question

- Is there a correlation between ERBB2 CNA and expression in breast cancer?

CGDS of cBioPortal

library install

install.packages("cgdsr")



The screenshot displays the cBioPortal website interface. At the top, the cBioPortal logo is on the left, the tagline "Visualize, analyze, discover." is in the center, and the Memorial Sloan Kettering Cancer Center logo is on the right. A navigation bar below the header contains links for HOME, DATA SETS, WEB API, R/MATLAB, TUTORIALS, FAQ, NEWS, TOOLS, ABOUT, and VISUALIZE YOUR DATA. The main content area is titled "R/MATLAB Packages" and features a section for the "CGDS-R Package". This section includes a description of the package's purpose for querying the Cancer Genomic Data Server (CGDS) via R, mentions its maintenance by Anders Jacobsen at MSKCC, and provides links to documentation and installation instructions. The installation instructions specify that the package works with R Version 2.12 or higher and provide the command to install it. An example usage section shows R code to create a CGDS object and test it. On the right side of the page, there are sidebars for "What's New" (with a job alert and a subscribe button), "Data Sets" (stating 126 cancer studies are available), and "Example Queries" (listing RAS/RAF alterations in colorectal cancer).

cBioPortal
for Cancer Genomics

Visualize, analyze, discover.

Memorial Sloan Kettering Cancer Center

HOME DATA SETS WEB API R/MATLAB TUTORIALS FAQ NEWS TOOLS ABOUT VISUALIZE YOUR DATA

R/MATLAB Packages

CGDS-R Package

Description

The CGDS-R package provides a basic set of functions for querying the Cancer Genomic Data Server (CGDS) via the [R platform for statistical computing](#).

Maintained by [Anders Jacobsen](#) at the Computational Biology Center, MSKCC.

Documentation

- [CGDS-R Package on CRAN](#).
- [The CGDS-R reference manual](#).
- [The CGDS-R documentation vignette](#).

Installation

1. The CGDS-R package currently **only works with R Version 2.12 or higher**.
2. Then install the cgds-R package from within R:

```
install.packages('cgdsr')
```

Example usage

```
# Create CGDS object
mycgds = CGDS("http://www.cbioportal.org/public-portal/")

test(mycgds)

# Get list of cancer studies at server
```


What's New

Job Alert Jobs available at [Data-Portal](#) to work on cBioPortal
Sign up for low-volume email news alerts:

Or follow us [@cbioportal](#) on Twitter

Data Sets

The Portal contains **126 cancer studies**.
[\[Details\]](#)



Example Queries

[RAS/RAF alterations in colorectal cancer](#)

[BRCA1 and BRCA2 mutations in breast cancer](#)

Getting conneted to cBioPortal

```
## Loading library
```

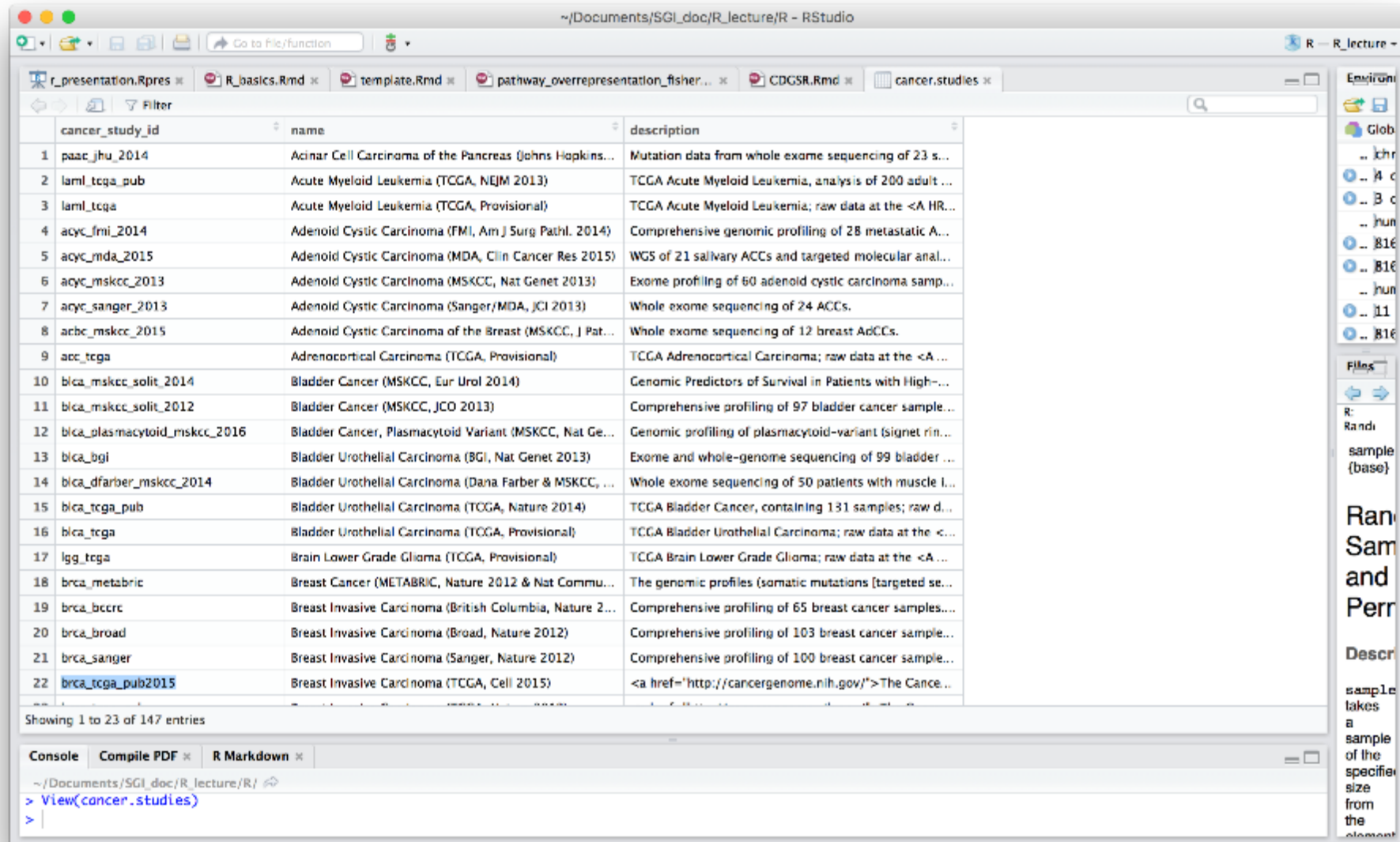
```
library(cgdsr)
```

```
## Create CGDS object
```

```
mycgds = CGDS("http://www.cbioportal.org/public-portal/")
```

```
cancer.studies = getCancerStudies(mycgds)
```

```
View(cancer.studies) #View table of studies list
```



The screenshot shows the RStudio interface with a table of cancer studies. The table has three columns: cancer_study_id, name, and description. The first 23 entries are visible, showing various cancer types and studies. The console at the bottom shows the command View(cancer.studies) being executed.

cancer_study_id	name	description
1 pasc_jhu_2014	Acinar Cell Carcinoma of the Pancreas (Johns Hopkins...)	Mutation data from whole exome sequencing of 23 s...
2 laml_tcga_pub	Acute Myeloid Leukemia (TCGA, NEJM 2013)	TCGA Acute Myeloid Leukemia, analysis of 200 adult ...
3 laml_tcga	Acute Myeloid Leukemia (TCGA, Provisional)	TCGA Acute Myeloid Leukemia; raw data at the <A HR...
4 acyc_fm1_2014	Adenoid Cystic Carcinoma (FMI, Am J Surg Pathl. 2014)	Comprehensive genomic profiling of 28 metastatic A...
5 acyc_mda_2015	Adenoid Cystic Carcinoma (MDA, Clin Cancer Res 2015)	WGS of 21 salivary ACCs and targeted molecular anal...
6 acyc_mskcc_2013	Adenoid Cystic Carcinoma (MSKCC, Nat Genet 2013)	Exome profiling of 60 adenoid cystic carcinoma samp...
7 acyc_sanger_2013	Adenoid Cystic Carcinoma (Sanger/MDA, JCI 2013)	Whole exome sequencing of 24 ACCs.
8 acbc_mskcc_2015	Adenoid Cystic Carcinoma of the Breast (MSKCC, J Pat...	Whole exome sequencing of 12 breast AdCCs.
9 acc_tcga	Adrenocortical Carcinoma (TCGA, Provisional)	TCGA Adrenocortical Carcinoma; raw data at the <A ...
10 blca_mskcc_solit_2014	Bladder Cancer (MSKCC, Eur Urol 2014)	Genomic Predictors of Survival in Patients with High-...
11 blca_mskcc_solit_2012	Bladder Cancer (MSKCC, JCO 2013)	Comprehensive profiling of 97 bladder cancer sample...
12 blca_plasmacytoid_mskcc_2016	Bladder Cancer, Plasmacytoid Variant (MSKCC, Nat Ge...	Genomic profiling of plasmacytoid-variant (signet rin...
13 blca_bgi	Bladder Urothelial Carcinoma (BGI, Nat Genet 2013)	Exome and whole-genome sequencing of 99 bladder ...
14 blca_dfarber_mskcc_2014	Bladder Urothelial Carcinoma (Dana Farber & MSKCC, ...	Whole exome sequencing of 50 patients with muscle l...
15 blca_tcga_pub	Bladder Urothelial Carcinoma (TCGA, Nature 2014)	TCGA Bladder Cancer, containing 131 samples; raw d...
16 blca_tcga	Bladder Urothelial Carcinoma (TCGA, Provisional)	TCGA Bladder Urothelial Carcinoma; raw data at the <...
17 lgg_tcga	Brain Lower Grade Glioma (TCGA, Provisional)	TCGA Brain Lower Grade Glioma; raw data at the <A ...
18 brca_metabric	Breast Cancer (METABRIC, Nature 2012 & Nat Commu...	The genomic profiles (somatic mutations [targeted se...
19 brca_bccrc	Breast Invasive Carcinoma (British Columbia, Nature 2...	Comprehensive profiling of 65 breast cancer samples...
20 brca_broad	Breast Invasive Carcinoma (Broad, Nature 2012)	Comprehensive profiling of 103 breast cancer sample...
21 brca_sanger	Breast Invasive Carcinoma (Sanger, Nature 2012)	Comprehensive profiling of 100 breast cancer sample...
22 brca_tcga_pub2015	Breast Invasive Carcinoma (TCGA, Cell 2015)	The Cance...

Showing 1 to 23 of 147 entries

Console: ~ / Documents / SGI_doc / R_lecture / R /
> View(cancer.studies)
>

Getting data

```
# Get data
```

```
mrnadata = getProfileData(mycgds, c("ERBB2"), "brca_tcga_pub2015_rna_s  
eq_v2_mrna", "brca_tcga_pub2015_3way_complete")
```

```
head(mrnadata)
```

	ERBB2
TCGA.LQ.A4E4.01	6846.946
TCGA.A2.A3KC.01	14814.131
TCGA.A2.A3KD.01	8941.431
TCGA.A7.A0D9.01	5291.478
TCGA.A7.A0DA.01	5035.810
TCGA.A7.A0CD.01	15139.034

Loading TCGA data

Get available case lists for a given cancer study

```
View(getCaseLists(mycgds,mycancerstudy))
```

```
mycaselist = getCaseLists(mycgds,mycancerstudy)[1,1] #All Complete Tumors
```

Get available genetic profiles

```
View(getGeneticProfiles(mycgds,mycancerstudy))
```

```
cna = getGeneticProfiles(mycgds,mycancerstudy)[5,1] #linear_CNA
```

```
mrna = getGeneticProfiles(mycgds,mycancerstudy)[3,1] #rna_seq_v2_mrna
```

Get data slices for a specified list of genes, genetic profile and case list

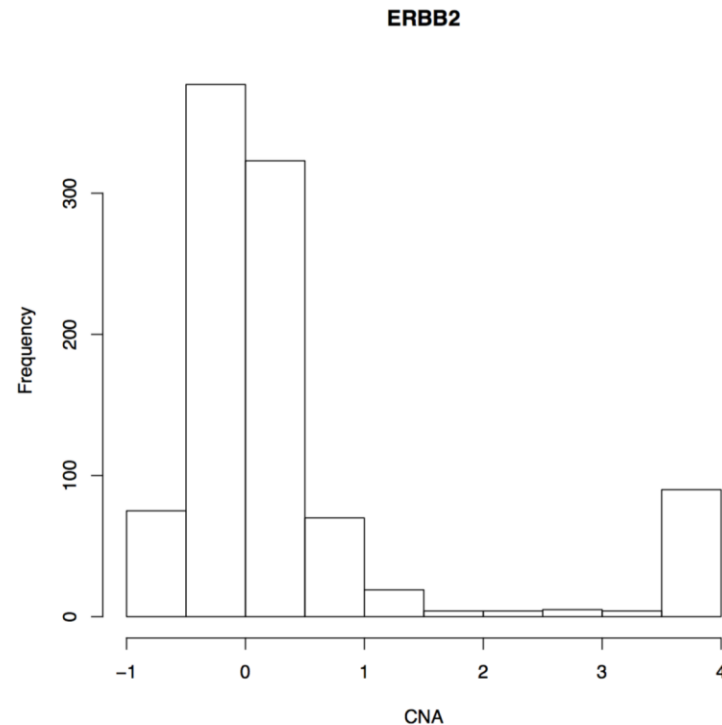
```
cnadata=getProfileData(mycgds,c("ERBB2"),cna,mycaselist) #CNA data of ERBB2
```

```
mrnadata=getProfileData(mycgds,c("ERBB2"),mrna,mycaselist) #mRNA data of ERBB2
```

Data visualization

##Histogram

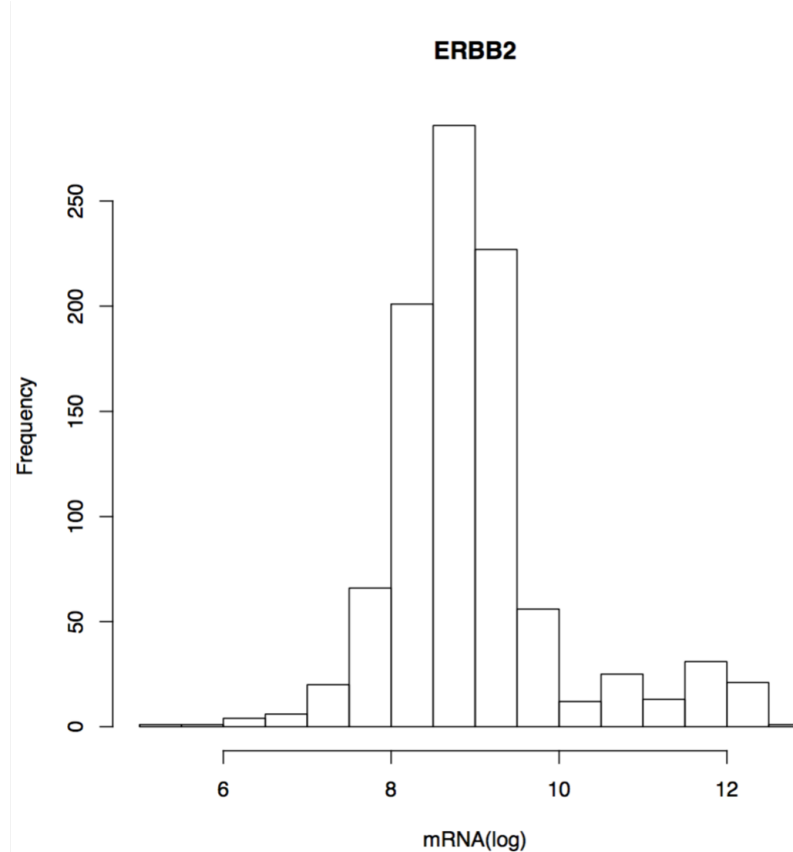
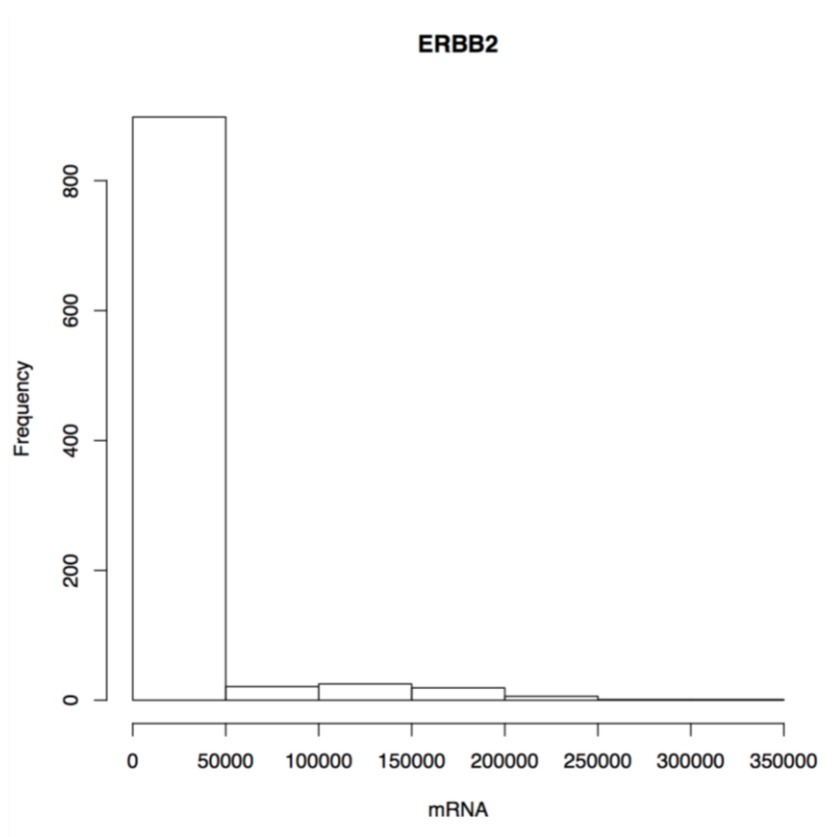
```
hist(cnadata$ERBB2,main="ERBB2",xlab="CNA  
A") # plot generation
```



Data visualization

```
hist(mrnadata$ERBB2,main="ERBB2",xlab="mRNA")
```

```
hist(log(mrnadata$ERBB2),main="ERBB2",xlab="mRNA") # log transfor  
mation
```

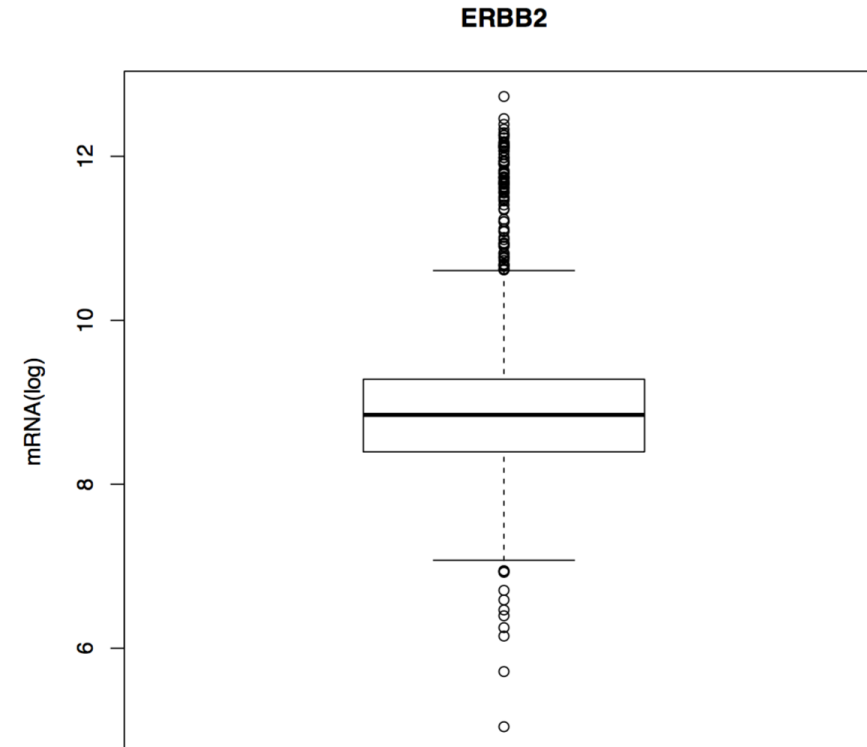
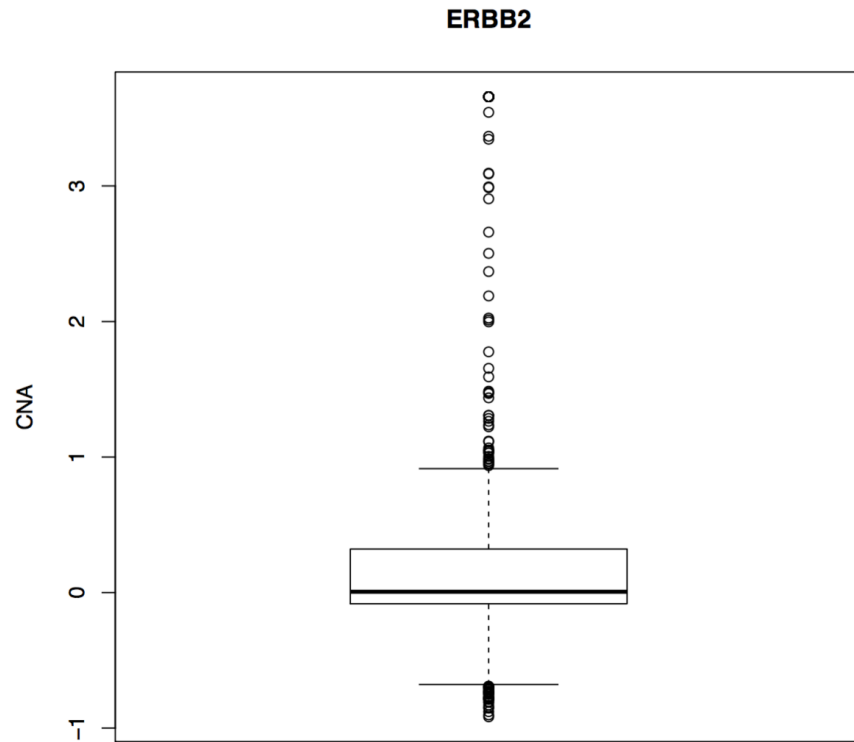


Data visualization

```
##box plot
```

```
boxplot(cnadata$ERBB2,main="ERBB2",ylab="CNA")
```

```
boxplot(log(mrnadata$ERBB2),main="ERBB2",ylab="mRNA(log)")
```



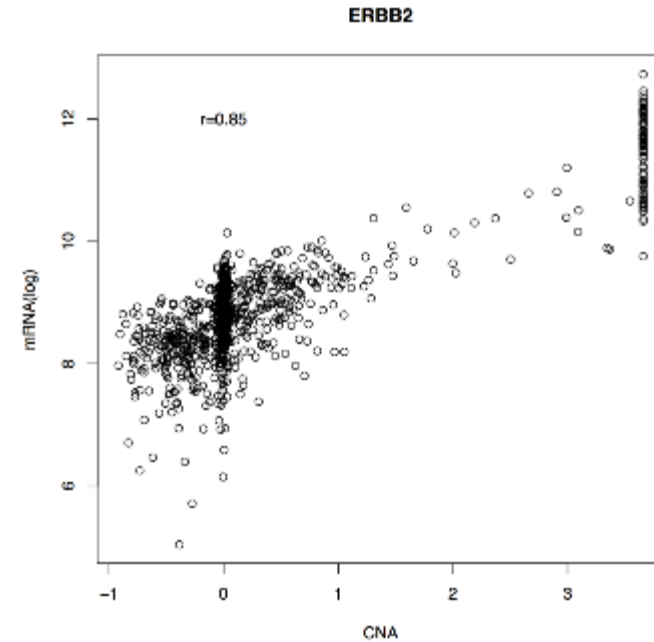
Data visualization

##Scatter plot

```
co=cor(cnadata$ERBB2,log(mrnadata$ERBB2)) #pearson correlation
```

```
plot(cnadata$ERBB2,log(mrnadata$ERBB2),main="ERBB2",xlab="CNA",ylab="mRNA(log)")
```

```
text(0,12,"r=0.85") # add text
```



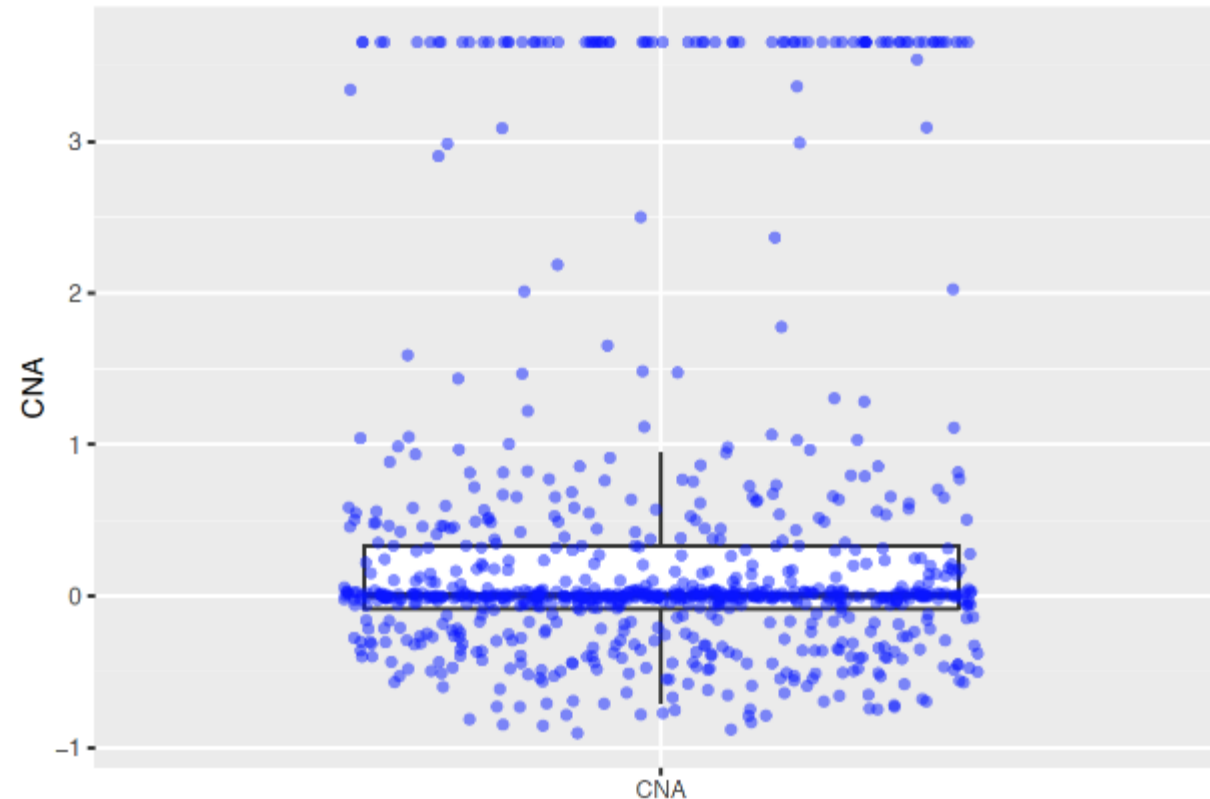
```
install.packages("ggplot2") ## library install
```

```
library(ggplot2) ## Loading library
```

```
input.cna = data.frame(cnadata, "CNA", stringsAsFactors = F)  
colnames(input.cna) = c("CNA", "Type")
```

```
plot1 = ggplot(input.cna, aes(x = Type, y = CNA)) + geom_boxplot(width = 0.7,  
  outlier.size = NA) + geom_jitter(width = 0.7, colour = "blue", alpha = 0.5)
```

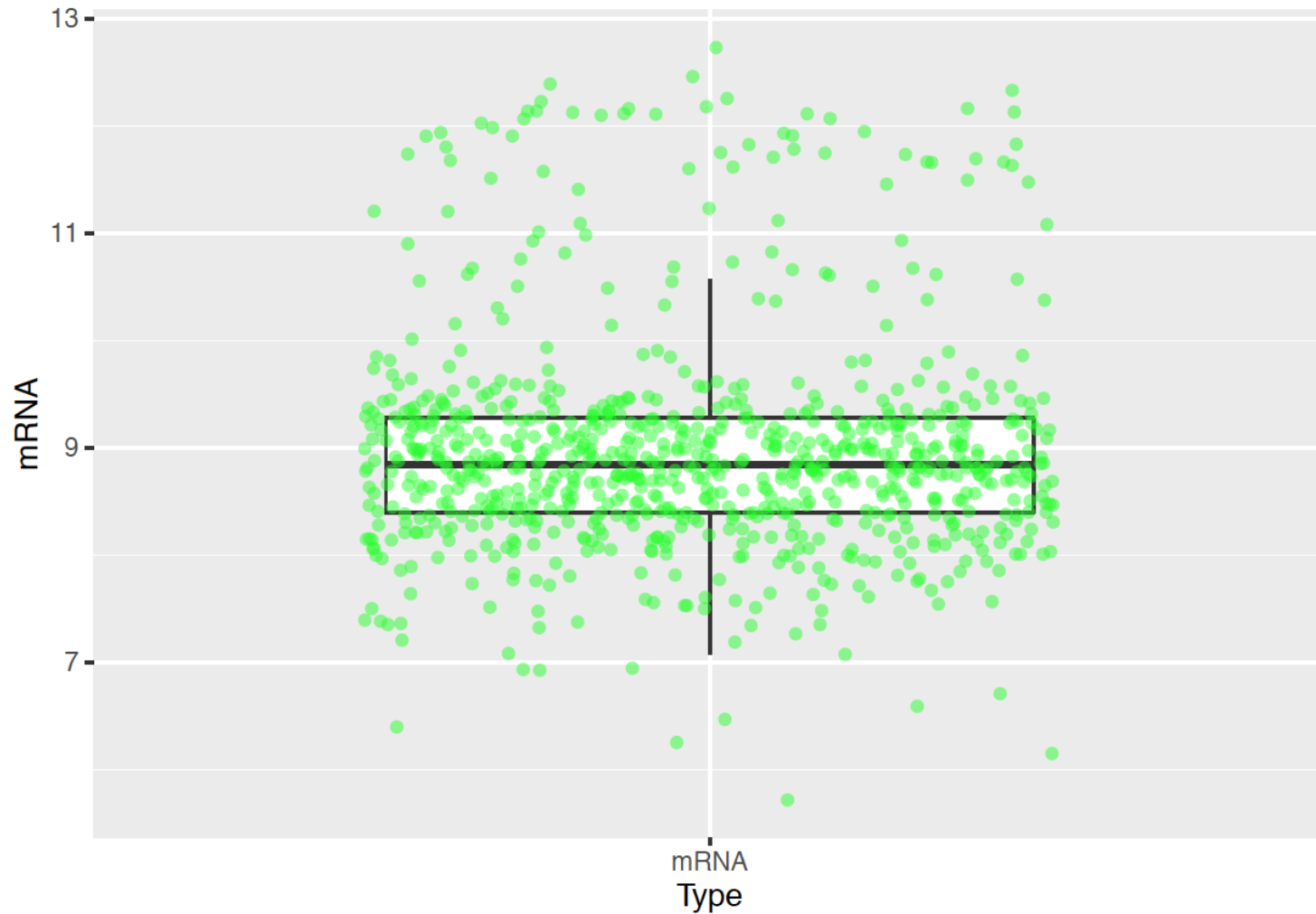
```
plot1
```



```
input.mrna = data.frame(log(mrnadata), "mRNA", stringsAsFactors = F)
colnames(input.mrna) = c("mRNA", "Type")

plot2 = ggplot(input.mrna, aes(x = Type, y = mRNA)) + geom_boxplot(width = 0.7,
  outlier.size = NA) + geom_jitter(width = 0.7, colour = "green", alpha = 0.5)

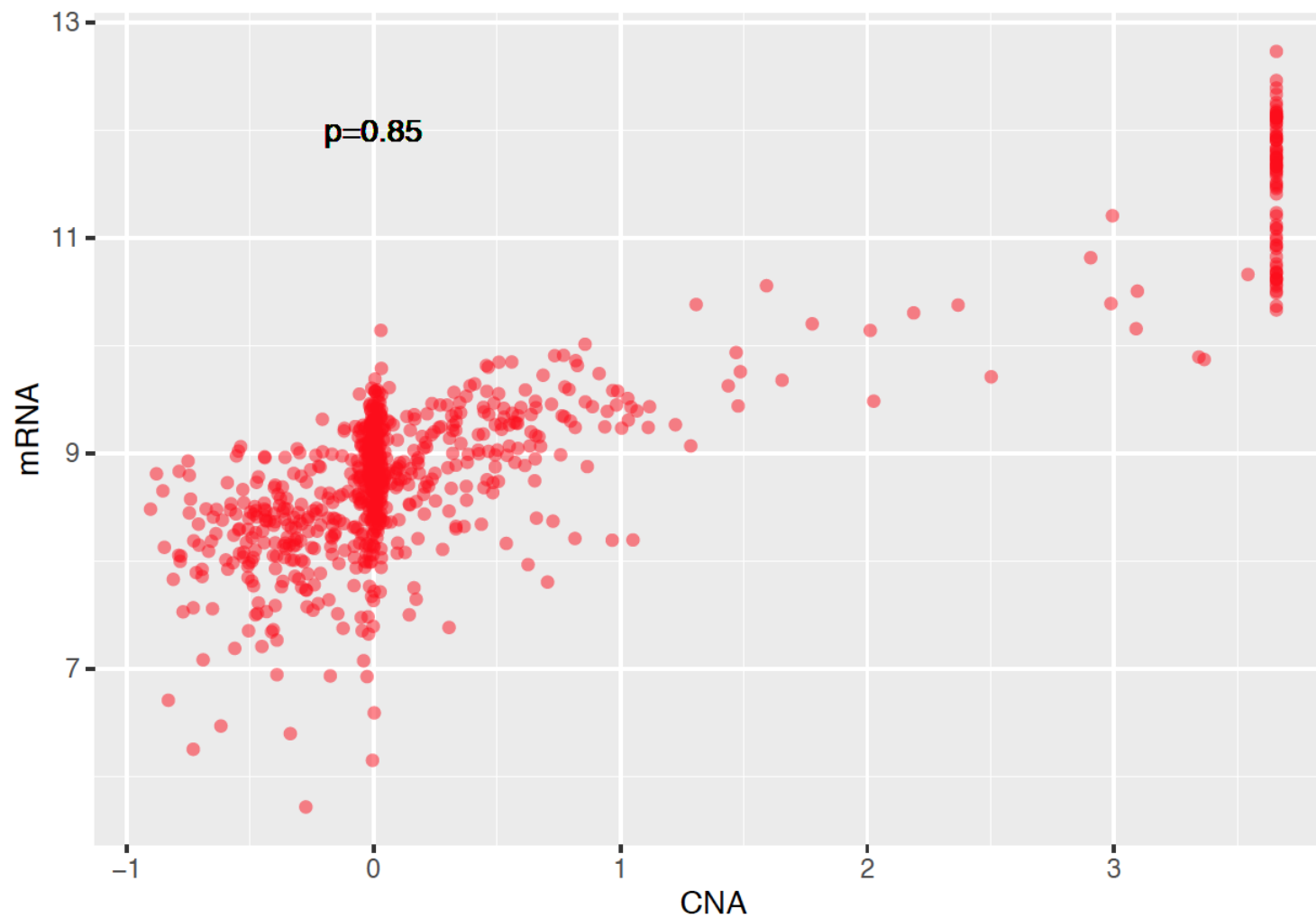
plot2
```



```
input.data = data.frame(cnadata, log(mrnadata), stringsAsFactors = F)
colnames(input.data) = c("CNA", "mRNA")

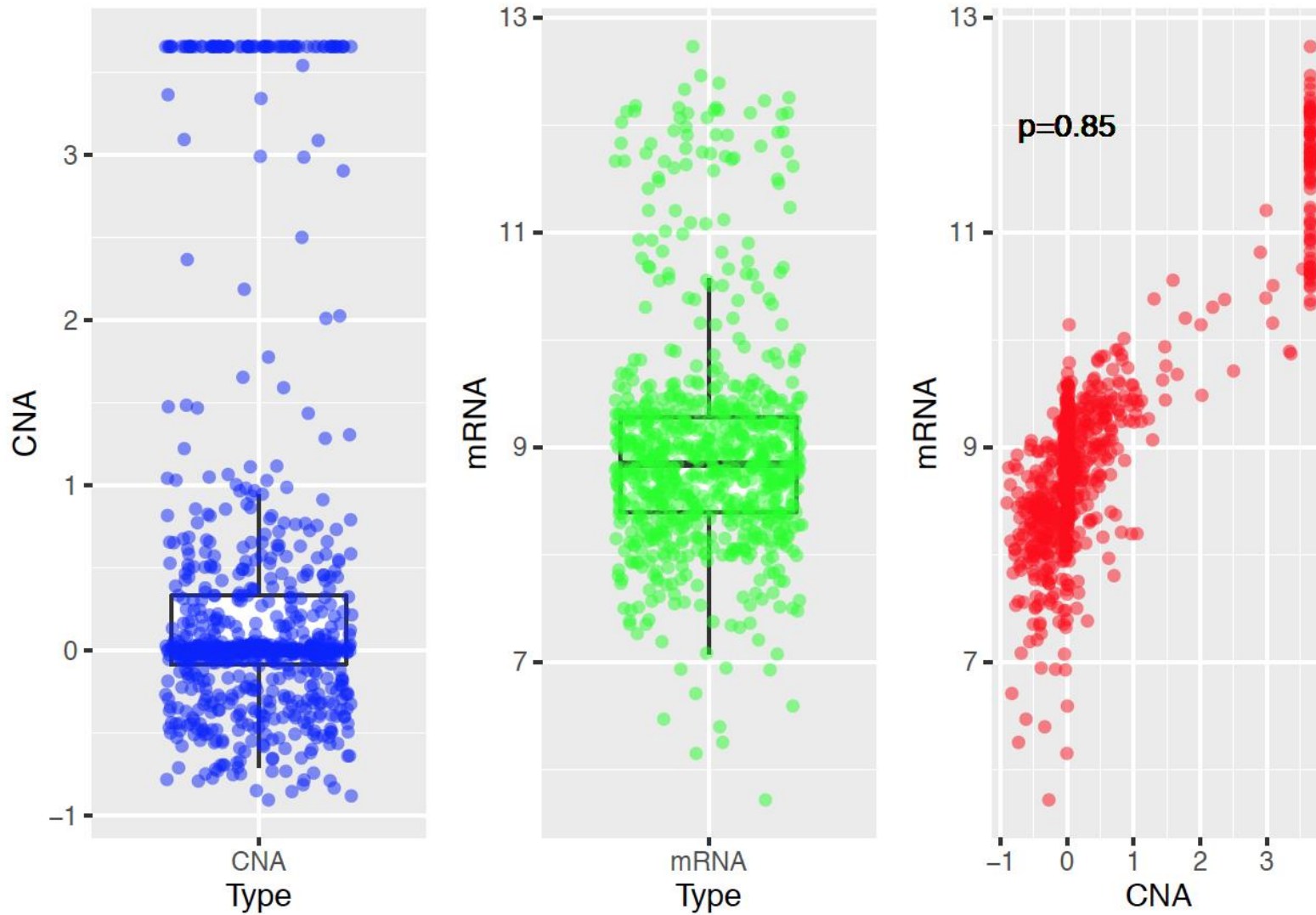
plot3 = ggplot(input.data, aes(x = CNA, y = mRNA)) + geom_point(color = "red",
  alpha = 0.5) + geom_text(aes(0, 12), label = "p=0.85")
```

plot3



```
install.packages("gridExtra")
```

```
library(gridExtra)  
grid.arrange(plot1, plot2, plot3, nrow = 1, ncol = 3)
```



Heatmap

Install NMF package

```
install.packages("NMF") #install package
```

```
library(NMF)
```

```
geneset = c("RUNX1", "PIK3CA", "TP53", "GATA3", "FOXA1", "SF3B1", "PTEN", "CBFB",  
            "CDH1", "TBX3", "MAP2K4", "MAP3K1", "ERBB2", "KMT2C", "NCOR1", "FAM86B2",  
            "CDKN1B", "HIST1H3B", "THEM5", "FAM86B1", "GPS2", "AQP12A", "PIK3R1", "ACTL6B",  
            "ZFP36L1", "RB1", "KRAS", "EPDR1", "C1QTNF5", "ZFP36L2", "CTCF", "ASB10",  
            "FBXW7", "RPGR", "MYB", "TBL1XR1", "CASP8", "TCP10", "WSCD2", "AARS", "FAM20C",  
            "HIST1H2BC", "ARID1A", "PTHLH")  
  
geneset_cndata = getProfileData(mycgds, geneset, cna, mycaselist) #cna data  
  
geneset_cndata = geneset_cndata[c(1:44), ] #sample selection  
  
geneset_cndata = t(geneset_cndata) # transpose data  
  
geneset_cndata[is.na(geneset_cndata)] <- 0  
# View(geneset_cndata)
```

```
ann_col = HeatmapAnnotation(Group = c(rep("A", 22), rep("B", 22)))
```

```
Heatmap(geneset_cndata, top_annotation = ann_col)
```

