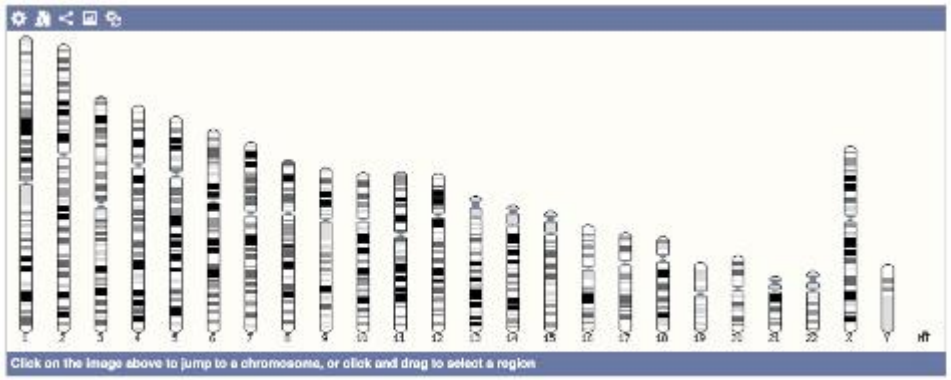
R을 활용한 TCGA 암유전체 데이터 분석

- 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 (isofor ms of a gene that each encode a d istinct protein product)

Summary

Assembly	GRCh38.p7 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.22 [illegible], Dec 2013
Database version	87.38
Base Pairs	3,547,782,741
Golden Path Length	3,095,649,726
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 526 readthrough)
Non coding genes	22,218
Small non coding genes	5,052
Long non coding genes	14,757 (incl 314 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,606 (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 know 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

CGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCT
TTTCTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAGCCAGACTTCCCGTGTCCTTTCCACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTG
GTCTTCATCTGCAGGTGTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGGGTGCAAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGCCATGCCTAGAGTG
GGATGGGCCATTGTTCATCTTCTGGCCCCTGTTGTCTGCATGTAACTTAATACCACAACCAGGCATAGGGGAAAGATTGGAGGAAAGATGAGTGAGAGCATCAACTT
CTCTCACAACCTAGGCCAGTAAGTAGTGCTTGTGCTCATCTCCTTGGCTGTGATACGTGGCCGGCCCTCGCTCCAGCAGCTGGACCCCTACCTGCCGTCTGCTGCCA
/TCGGAGCCCAAAGCCGGGCTGTGACTGCTCAGACCAGCCGGCTGGAGGGAGGGCC/GCTCAGCAGGTCTGGCTTTGGCCCTGGGAGAGCAGGTGGAAGATCAGGCA
GGCCATCGCTGCCACAGAACCCAGTGGATTGGCCTAGGTGGGATCTCTGAGCTCAACAAGCCCTCTCTGGGTGGTAGGTGCAGAGACGGGAGGGGCAGAGCCGCAGG
CACAGCCAAGAGGGCTGAAGAAATGGTAGAACGGAGCAGCTGGTGATGTGTGGGCCCACCGGCCCCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCAT
GCCCTTCCCCAGCATCAGGTCTCCAGAGCTGCAGAAGACGACGGCCGACTTGGATCACACTCTTGTG/AAGTGTCCCCAGTGTTGCAGAGGTGAGAGGAGAGTAGAC
AGTGAGTGGGAGTGGCGTCGCCCCTAGGGCTCTACGGGGCCGGCGTCTCCTGTCTCCTGGAGAGGCTTCGATGCCCCTCCACACCCTCTTGATCTTCCCTGTGATGT
CATCTGGAGCCCTGCTGCTTGCGGTGGCCTATAAAGCCTCCTAGTCTGGCTCCAAGGCCTGGCAGAGTCTTTCCCAGGGAAAGCTACA/TAGCAGCAAACAGTCTGC
ATGGGTCATCCCCTTCACTCCCAGCTCAGAGCCCAGGCCAGGGGCCCCCAAGAAAGGCTCTGGTGGAGAACCTGTGCATGAAGGCTGTCAACCAGTCCATAGGCAAG
CCTGGCTGCCTCCAGCTGGGTCGACAGACAGGGGCTGGAGAAGGGGAGAAGAGGAAAGTGAGGTTGCCTGCCCTGTCTCCTACCTGAGGCTGAGGAAGGAGAAGGGG
ATGCACTGTTGGGGAGGCAGCTGTAACTCAAAGCCTTAGCCTCTGTTCCCACGAAGGCAGGGCCATCAGGCACCAAAGGGATTCTGCCAGCATAGTGCTCCTGGACC
AGTGATACACCCGGCACCCTGTCCTGGACACGCTGTTGGCCTGGATCTGAGCCCTGGTGGAGGTCAAAGCCACCTTTGGTTCTGCCATTGCTGCTGTGTGGAAGTTC
ACTCCTGCCTTTTCCTTTCCCTAGAGCCTCCACCACCCCGAGATCACATTTCTCACTGCCTTTTGTCTGCCCAGTTTCACCAGAAGTAGGCCTCTTCCTGACAGGC/
TAGCTGCACCACTGCCTGGCGCTGTGCCCTTCCTTTGCTCTGCCCGCTGGAGACGGTGTTTGTCATGGGCCTGGTCTGCAGGGATCCTGCTACAAAGGTGAAACCCA
GGAGAGTGTGGAGTCCAGAGTGTTGCCAGGACCCAGGCACAGGCATTAGTGCCCGTTGGAGAAAACAGGGGAATCCCGAAGAAATGGTGGGTCCTGGCCATCCGTGA
GATCTTCCCAGGTGTGCCGTTTTCTCTGGAAGCCTCTTAAGAACACAGTGGCGCAGGCTGGGTGGAGCCGTCCCCCCATGGAGCACAGGCA/GGACAGAAGTCCCCG
CCCCAGCTGTGTGGCCTCAAGCCAGCCTTCCGCTCCTTGAAGCTGGTCTCCACACAGTGCTGGTTCCGTCACCCCCTCCCAAGGAAGTAGGTCTGAGCAGCTTGTCC
TGGCTGTGTCCATGTCAGAGCAACGGCCCAAGTCTGGGTCTGGGGGGGAAGGTGTCATGGAGCCCCCTACGATTCCCAGTCGTCCTCGTCCTCCTCTGCCTGTGGCT
GCTGCGGTGGCGGCAGAGGAGGGATGGAGTCTGACACGCGGGCAAAGGCTCCTCCGGGCCCCTCACCAGCCCCAGGTCCTTTCCCAGAGATGCCTGGAGGGAAAAGG
CTGAGTGAGGGTGGTTGGTGGGAAACCCTGGTTCCCCCAGCCCCCGGA/CGACTTAAATACAGGAAGAAAAAGGCAGGACAGAATTACAAGGTGCTGGCCCAGGGCG
GGCAGCGGCCCTGCCTCCTACCCTTGCGCCTCATGACCGGAGCCATAGCCCAGGCAGGAGGGCTGAGGACCTCTGGTGGCGGCCCAGGGCTTCCAGCATGTGCCCTA
GGGGAAGCAGGGGCCAGCTGGCAAGAGCAGGGGGTGGGCAGAAAGCACCCGGTGGACTCAGGGCTGGAGGGGAGGAGGCGATCTTGCCCAAGGCCCTCCGACTGCAA
GCTCCAGGGCCCGCTCACCTTGCTCCTGCTCCTTCTGCTGCTGCTTCTCCAGCTTTCGCTCCTTCATGCTGCGCAGCTTGGCCTTGCCGATGCCCCCAGCTTGGCGG
ATGGACTCTAGCAGAGTGGCCAGCCACCGGAGGGGTCAACCACTTCCC

Types of genetic variation

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

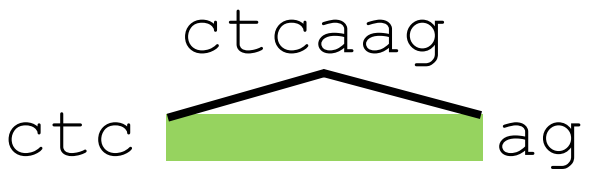
Single-nucleotide polymorphisms (**SNPs**)

"DNA spelling mistakes"

ctc--ag
ctc**tg**ag

Insertion-deletion polymorphisms (**INDELs**)

"extra or missing DNA"

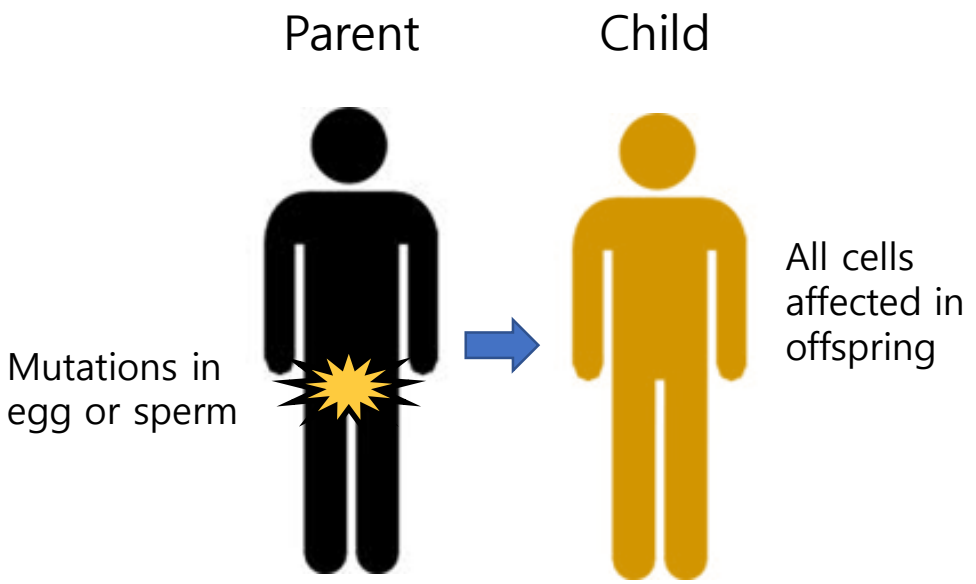


Structural variants (**SVs**)

"Large blocks of extra, missing or rearranged DNA"

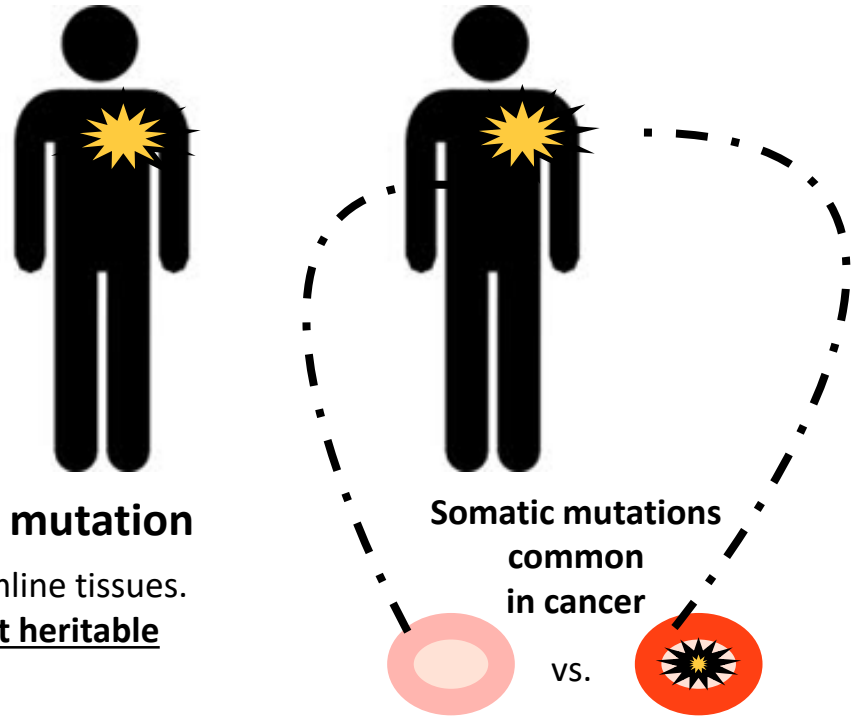
Mutation != Polymorphism (or SNP)

Somatic mutations



Germline mutation

- occur in sperm or egg.
- are heritable



Somatic mutation

- non-germline tissues.
- **<u>are not heritable</u>**

compare DNA from cancer cells to healthy cells from same individual

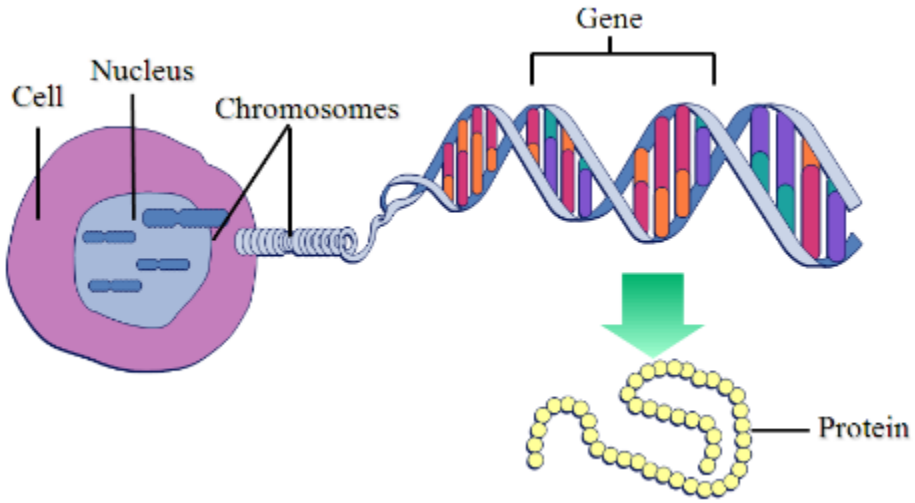
Types of Genetic Alterations in Cancer

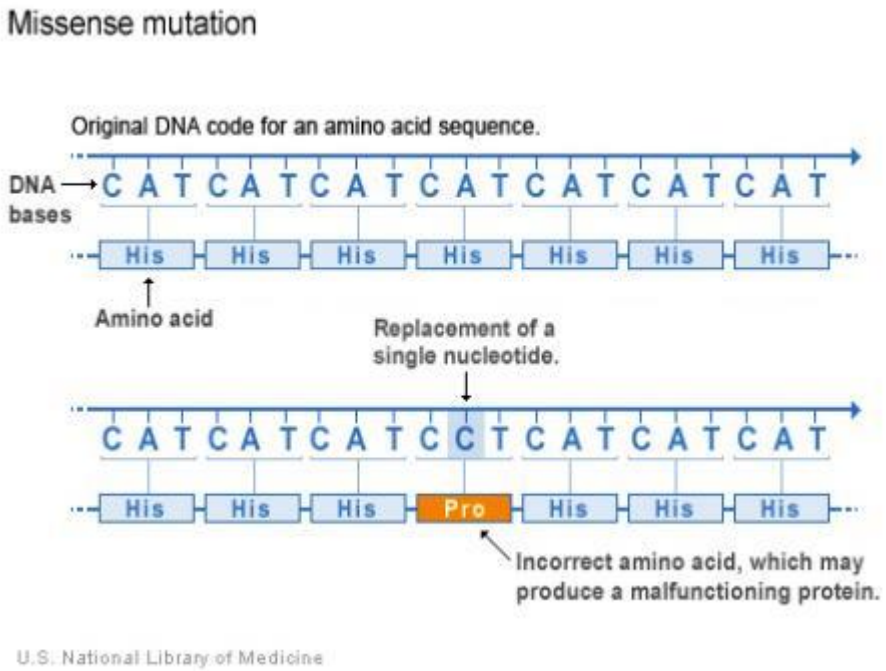
- Subtle alterations
- Chromosome number changes
- Chromosomal translocation
- Amplifications
- Exogenous sequences

Subtle Alterations

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

Missense mutation





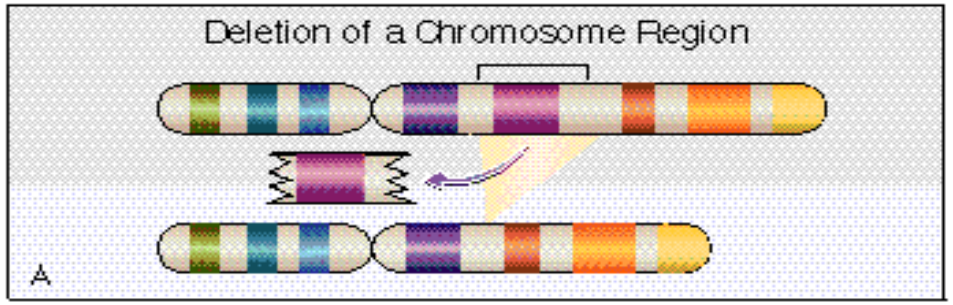
Point Mutations

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

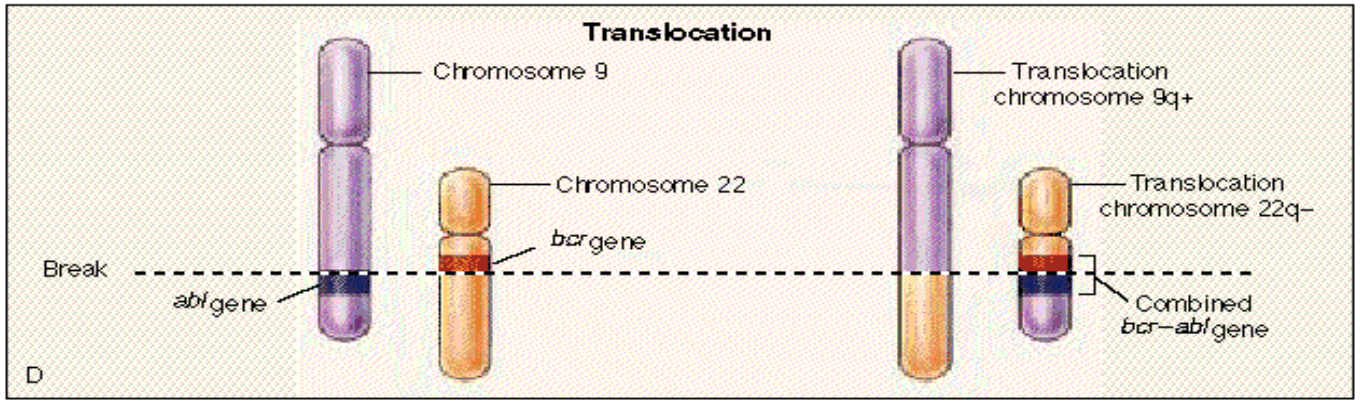
Chromosome Number Changes

- 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



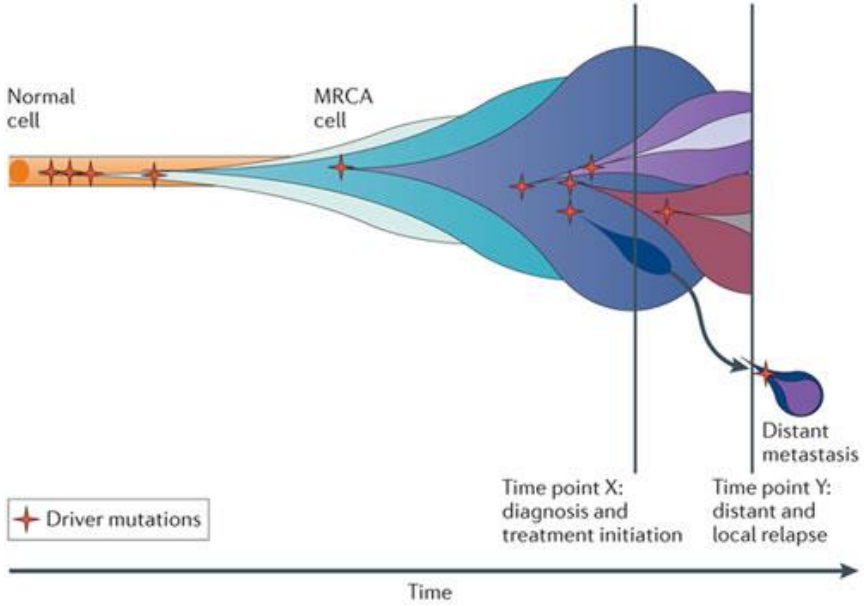
Amplifications

- 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

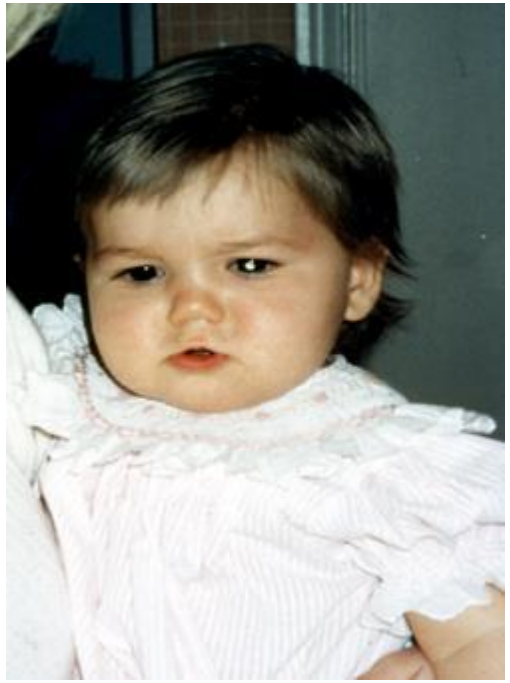
- **Driver mutation**
  - A mutation that gives a selective advantage to a clone in its microenvironment, through either increasing its survival or reproduction.
  - Driver mutations tend to cause clonal expansions.
- **Passenger mutation**
  - A mutation that has no effect on the fitness of a clone 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

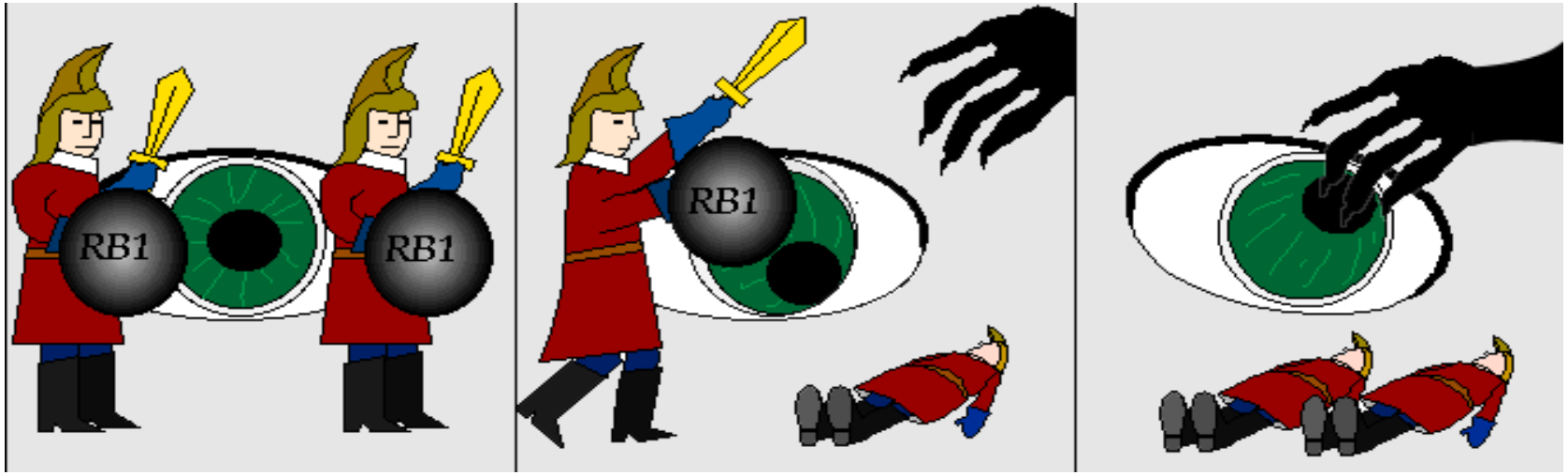
thedarwincancerblog.com

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



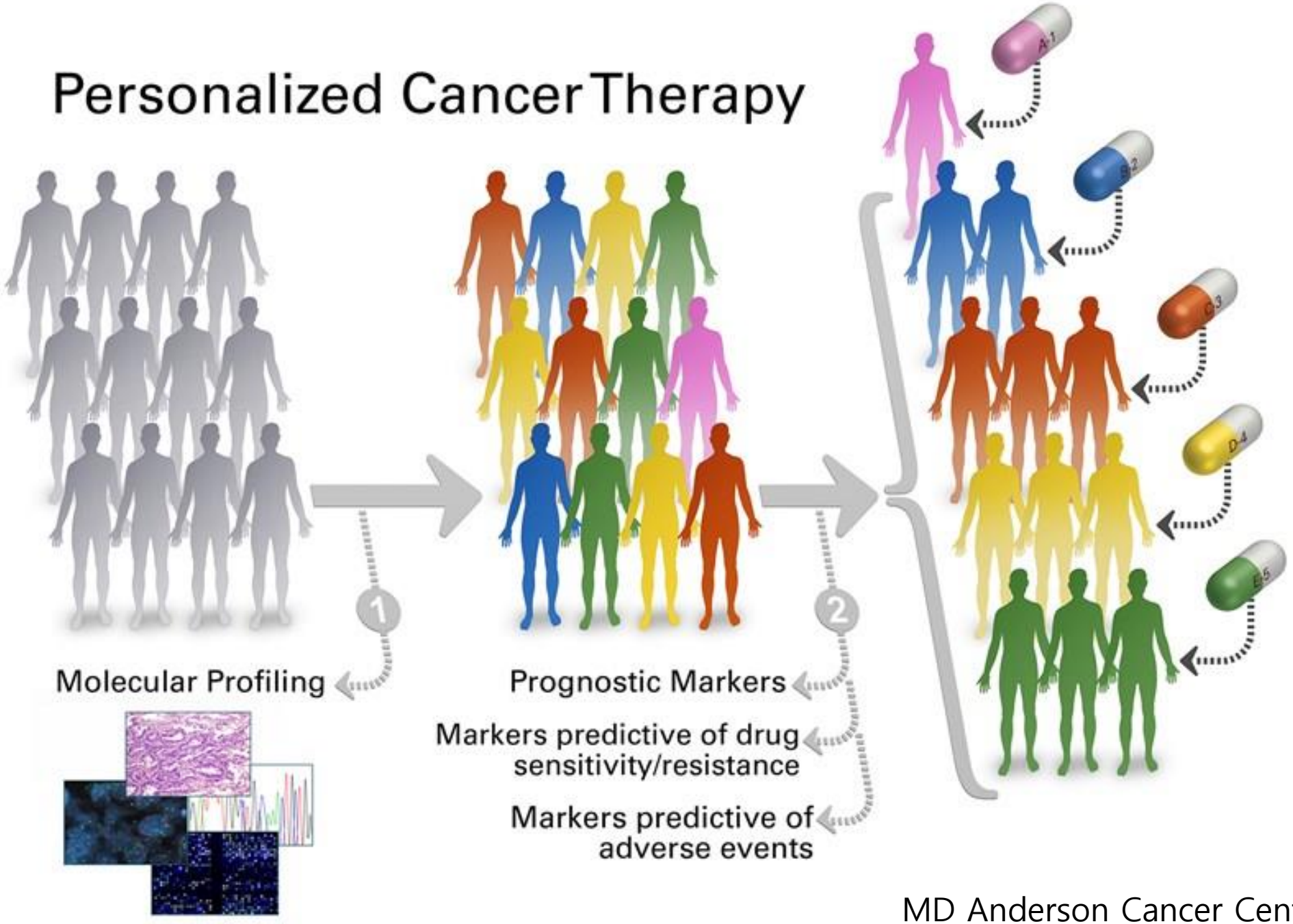
Normal
2 intact copies

Predisposed
1 intact copy
1 mutation

Affected
Loss of both copies

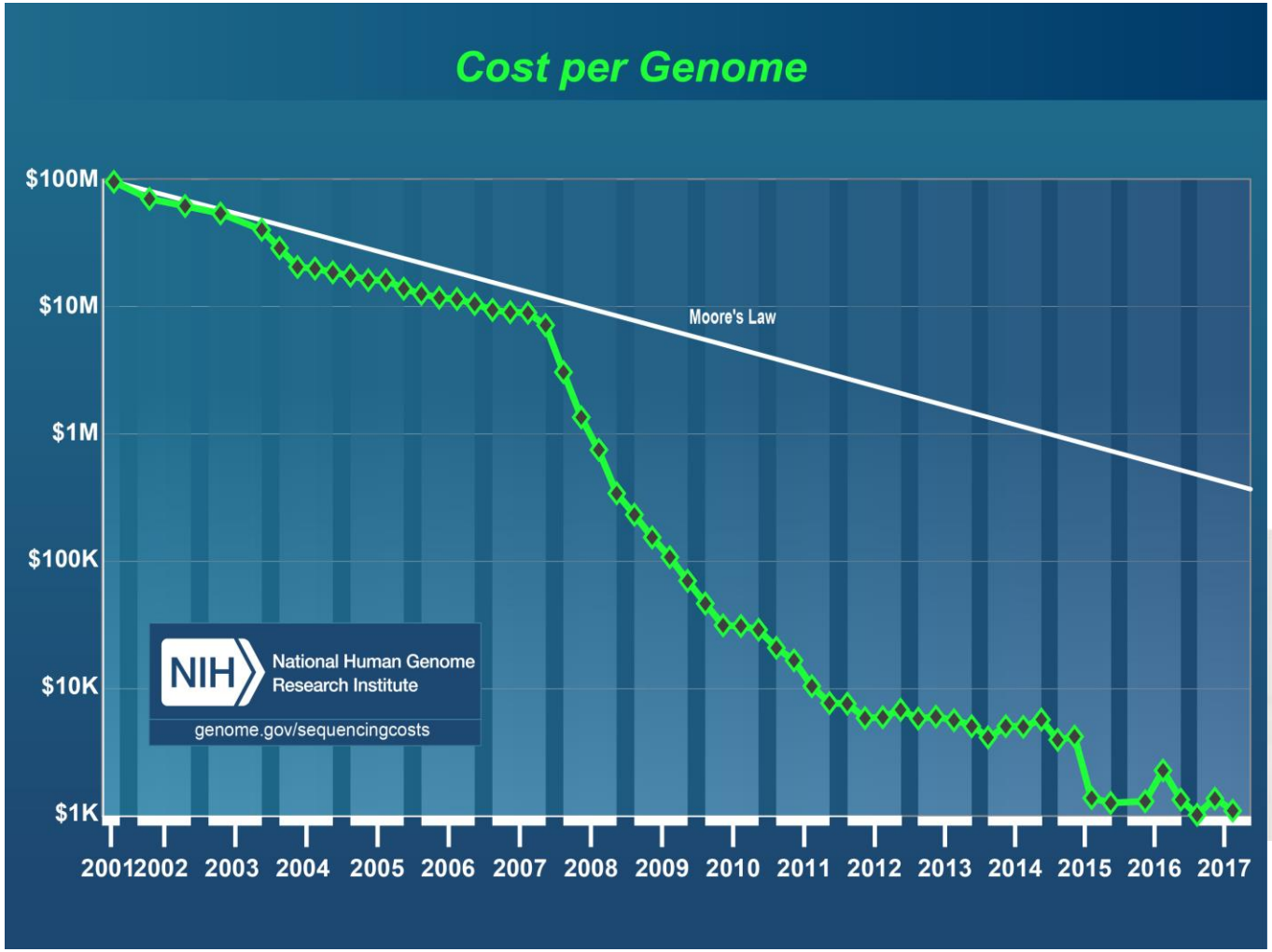
Modified from *Time*, Oct. 27, 1986

ASCO



MD Anderson Cancer Center

Cost of sequencing



NovaSeq: $100

https://www.genome.gov/images/content/costg_2017.jpg

NATIONAL CANCER INSTITUTE

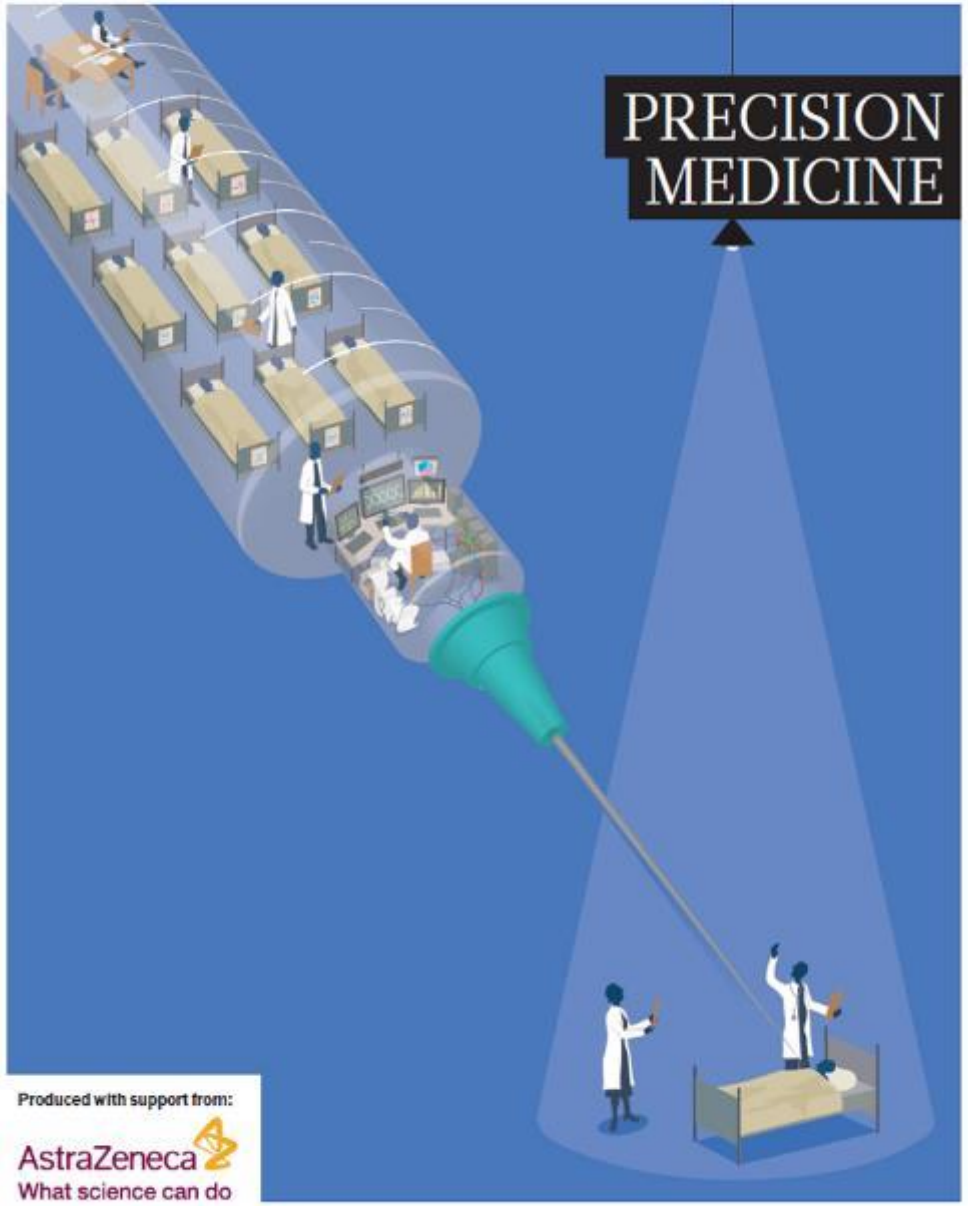
PRECISION MEDICINE IN CANCER TREATMENT

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

www.cancer.gov

nature INSIGHT

PRECISION MEDICINE

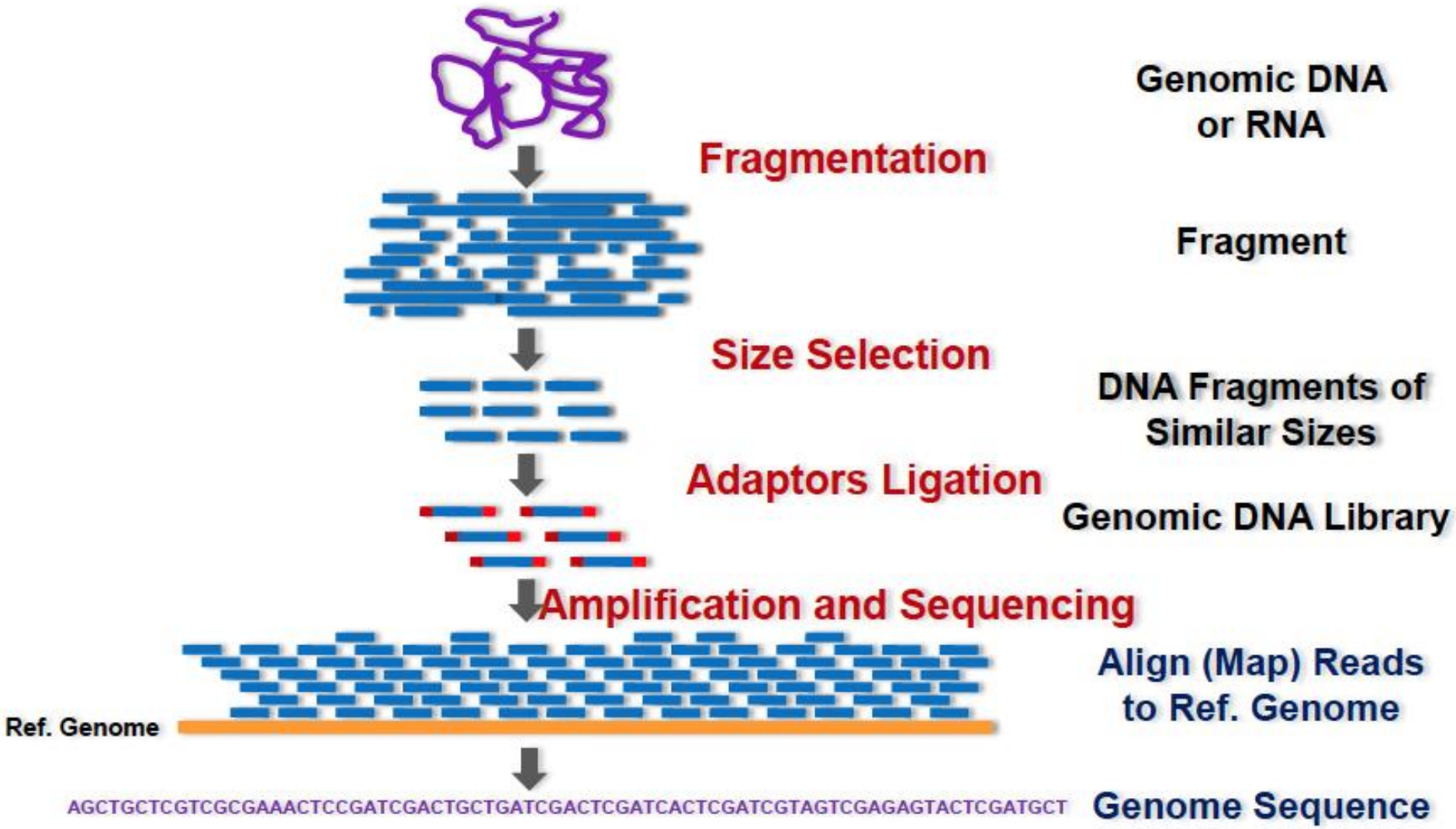


Produced with support from:

AstraZeneca
What science can do

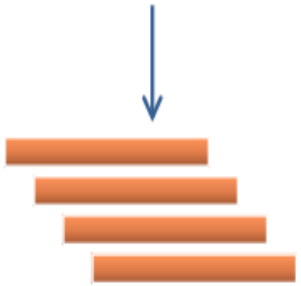
How to Detect Variants?

(Next-Generation Sequencing)



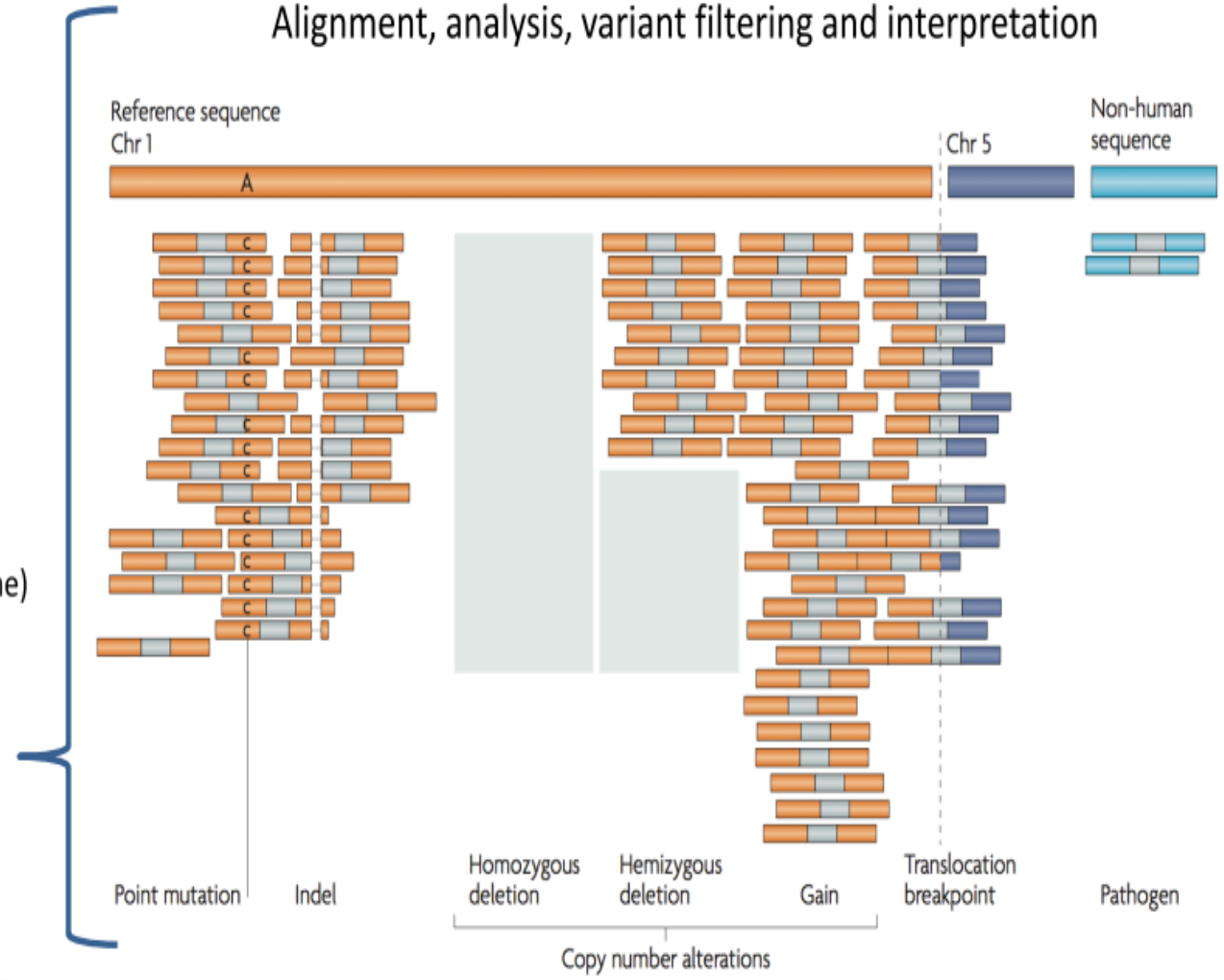
NGS – SNV, indel, CNV, SV

DNA or RNA



Library of fragments
(amplicon, hybridization
Panel, Exome, Whole Genome)

NGS Sequencer



Meyerson, Gabriel, Getz. Nat Rev Genet. 2010

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 sample 들의 ERBB2 expression의 평균값을 구하시오.
- 94개 이상의 tumor sample들에서 CNV가 2 이상인 유전자를 도출하시오.

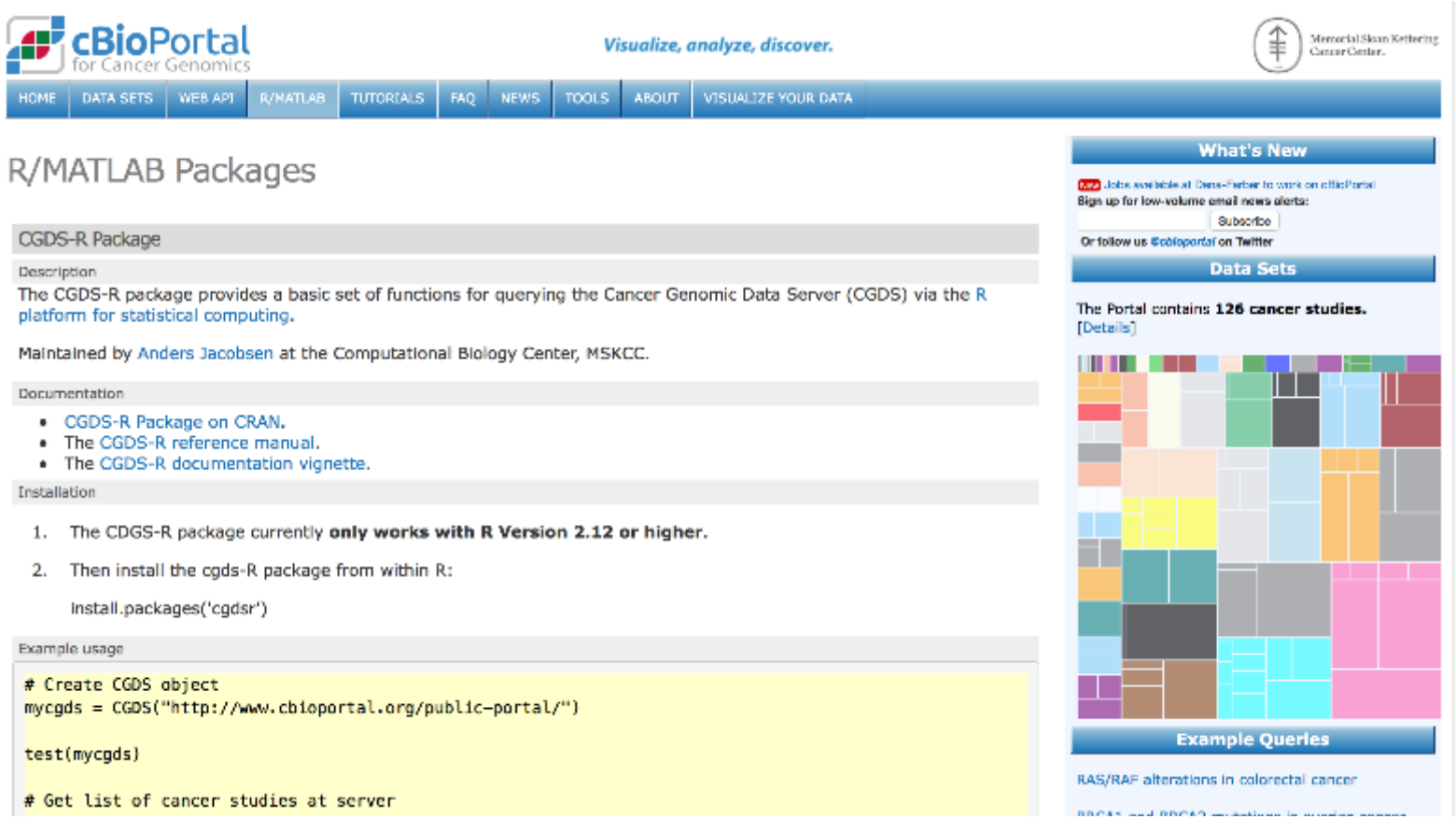
Question

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

CGDS of cBioPortal

library install

install.packages("cgdsr")

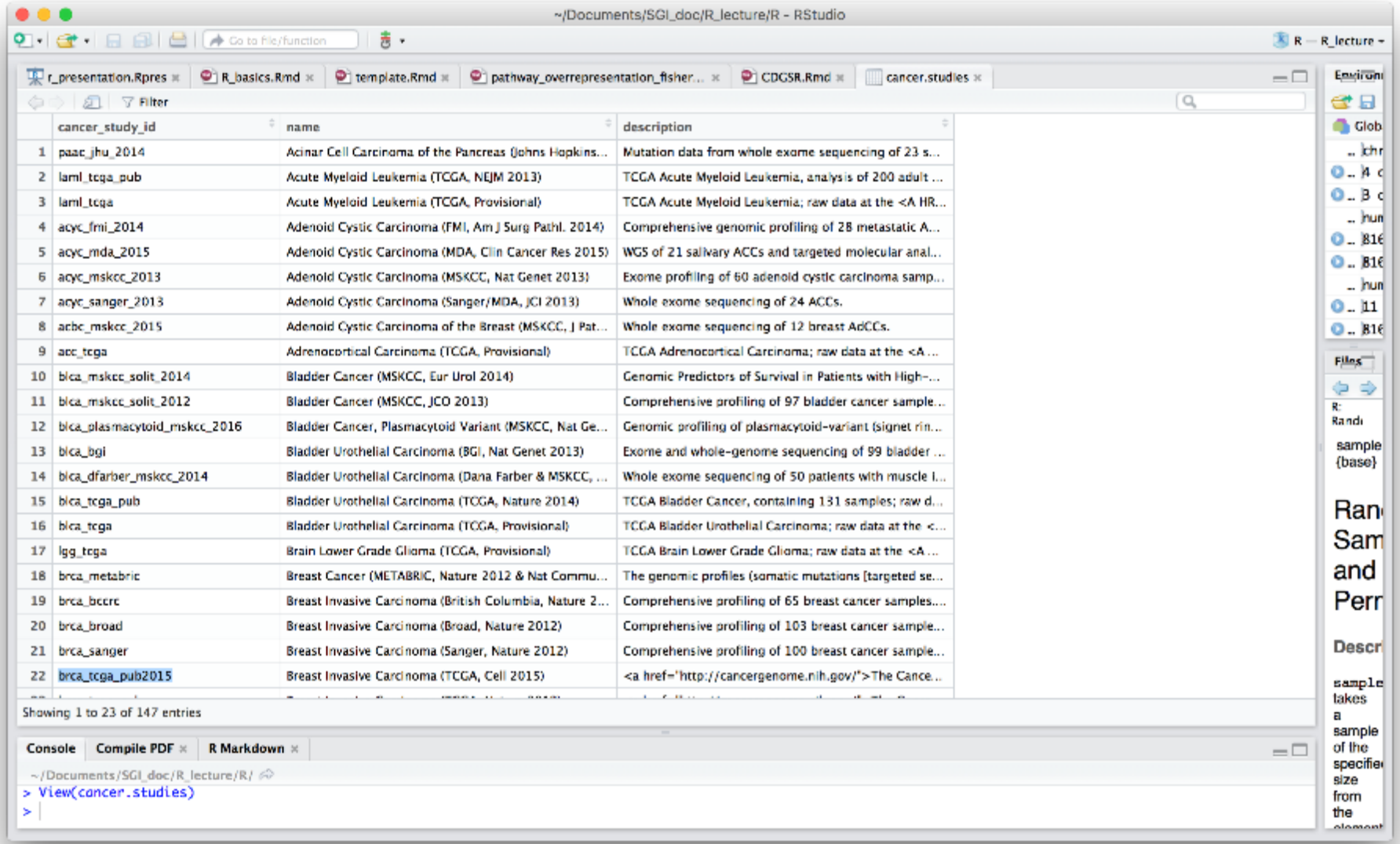


Getting conneted to cBioPortal

```
## 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
```



Getting data

```
# Get data

mrnadata = getProfileData(mycgds, c("ERBB2"), "brca_tcga_pub2015_rna_seq_v2_mrna", "brca_tcga_pub2015_3way_complete")

head(mrnadata)
```

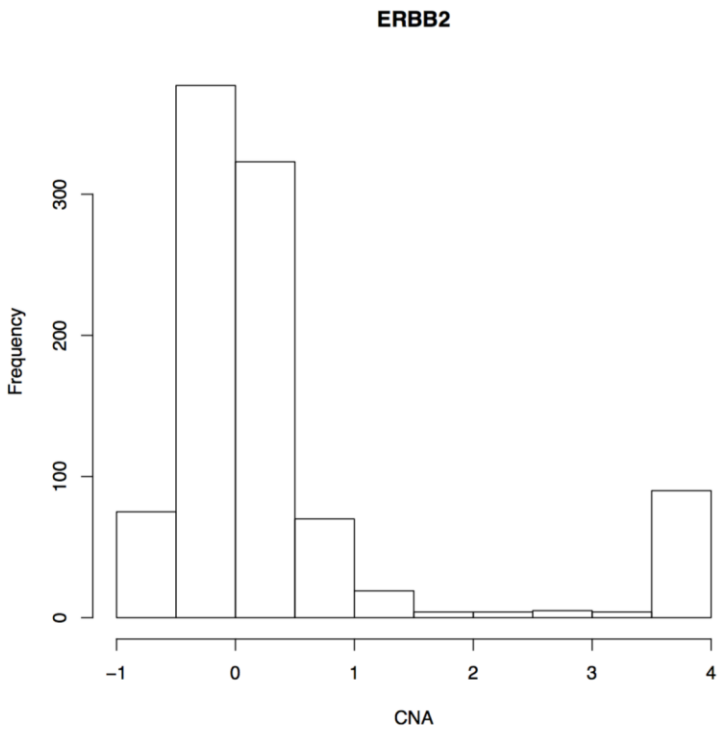
```
                   ERBB2
TCGA.LQ.A4E4.01  6846.946
TCGA.A2.A3KC.01 14814.131
TCGA.A2.A3KD.01  8941.431
TCGA.A7.A0D9.01  5291.478
TCGA.A7.A0DA.01  5035.810
TCGA.A7.A0CD.01 15139.034
```

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
```

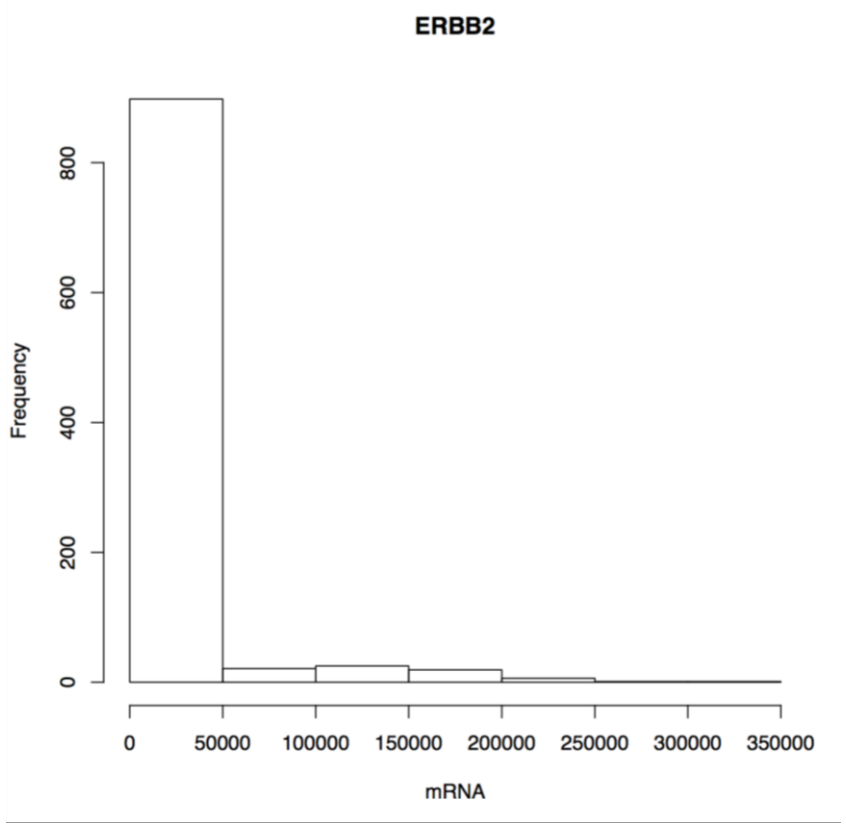
Data visualization

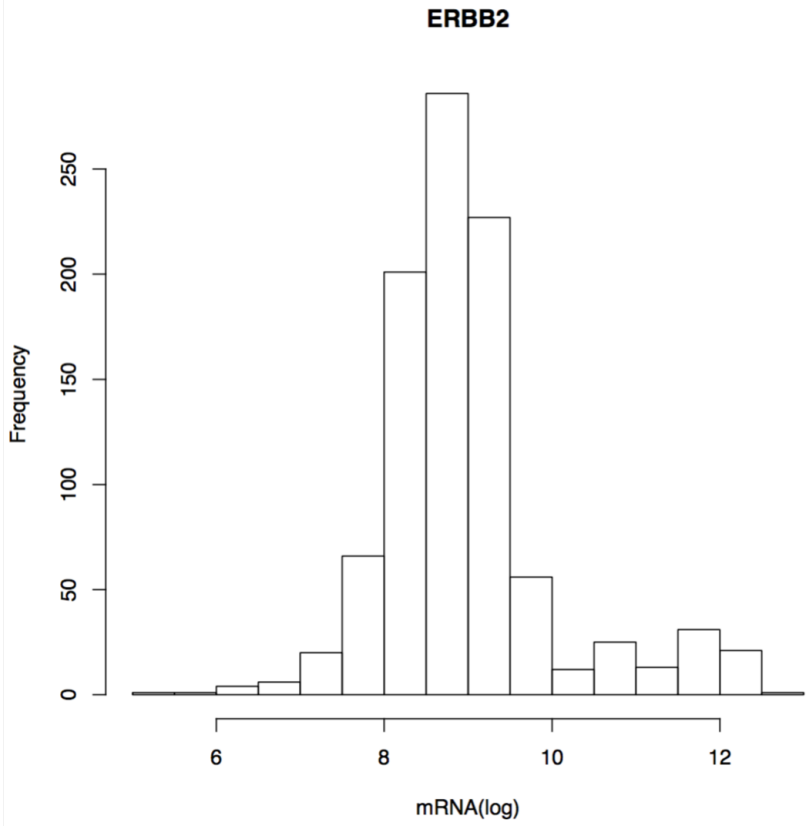
```
##Histogram

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



Data visualization

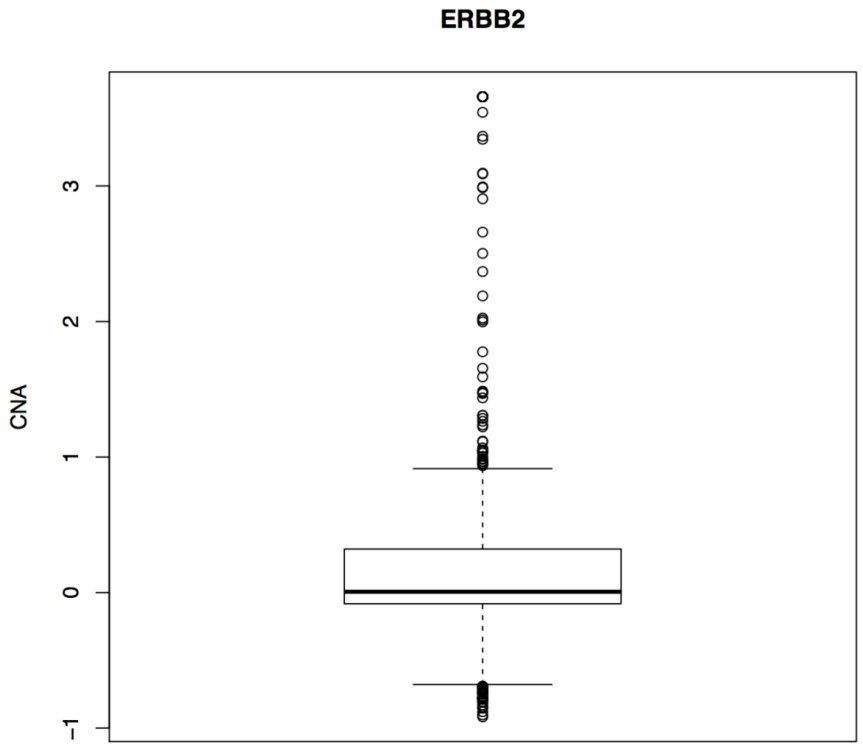
```
hist(mrnadata$ERBB2,main="ERBB2",xlab="mRNA")

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

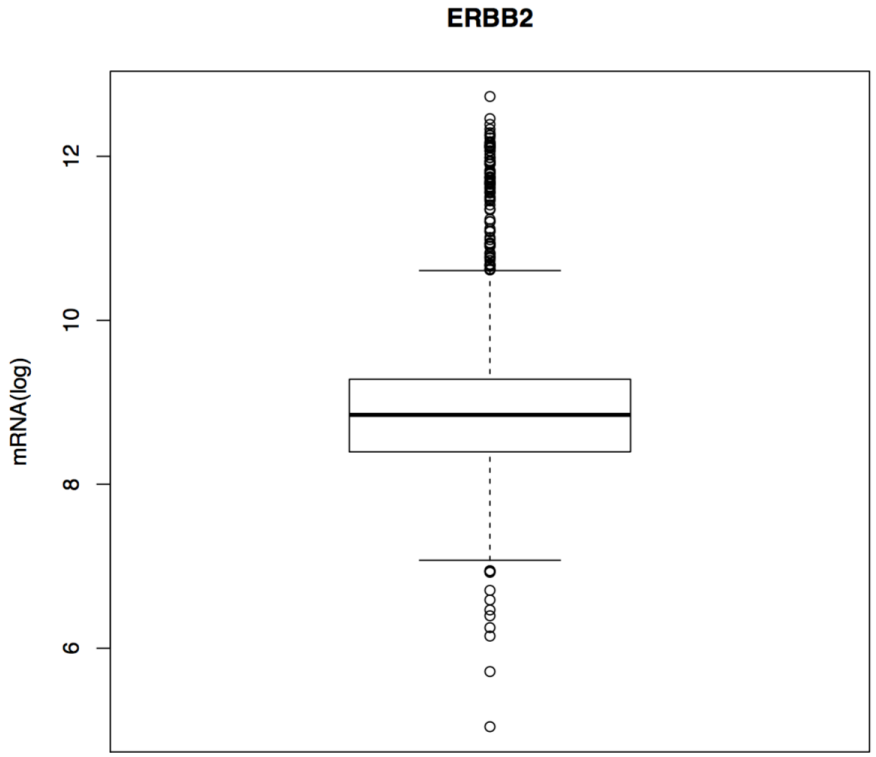




Data visualization

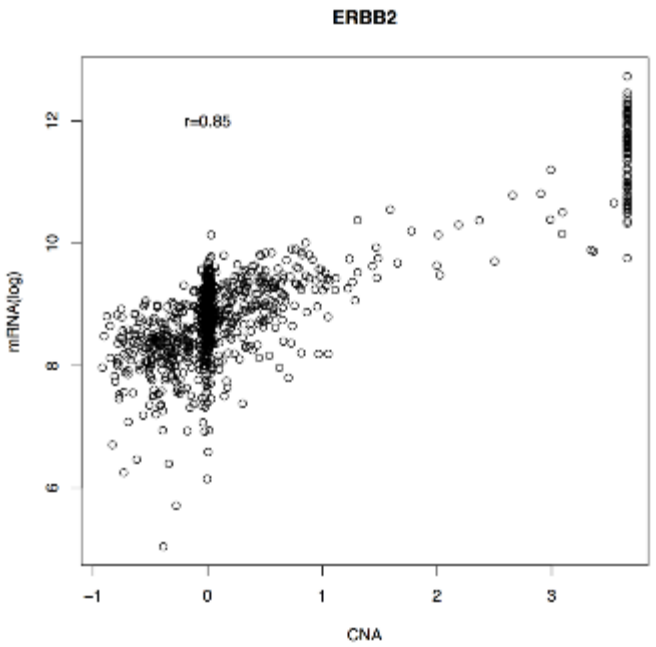
```
##box plot
boxplot(cnadata$ERBB2,main="ERBB2",ylab="CNA")
boxplot(log(mrnadata$ERBB2),main="ERBB2",ylab="mRNA(log)")
```





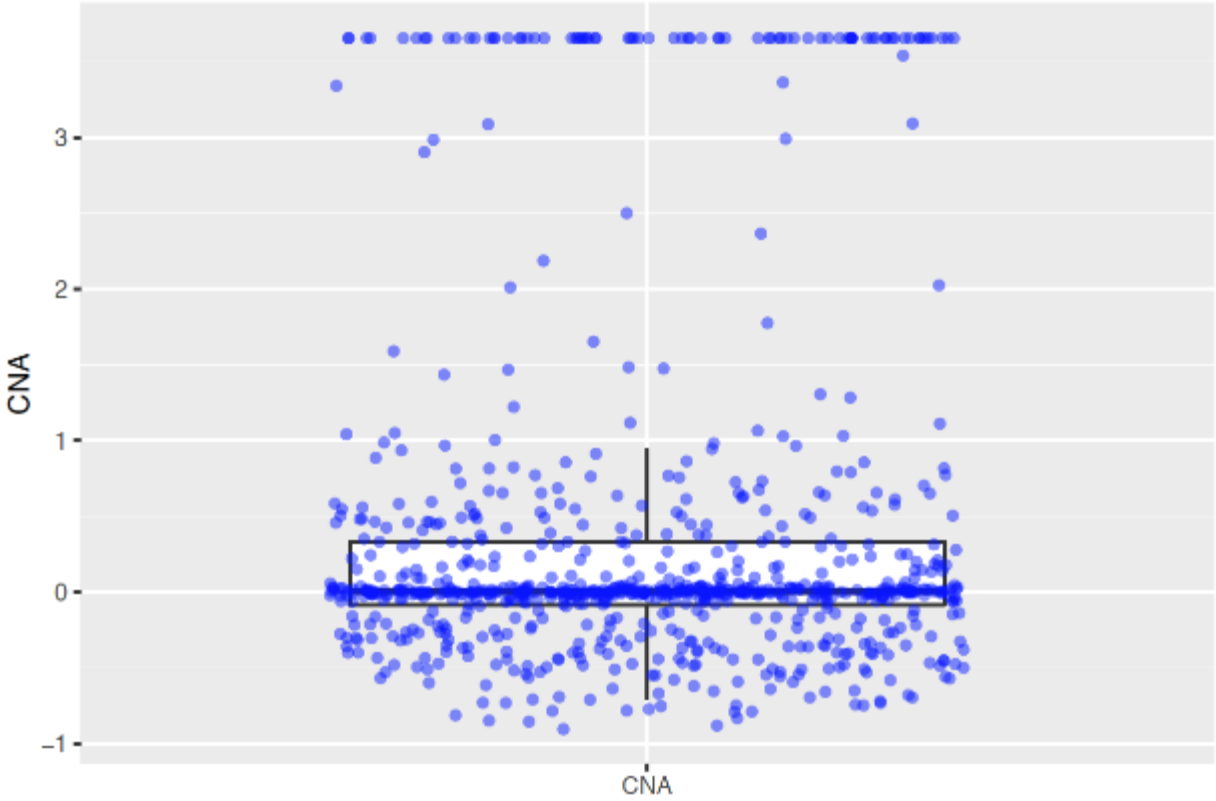
Data visualization

```
##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
```



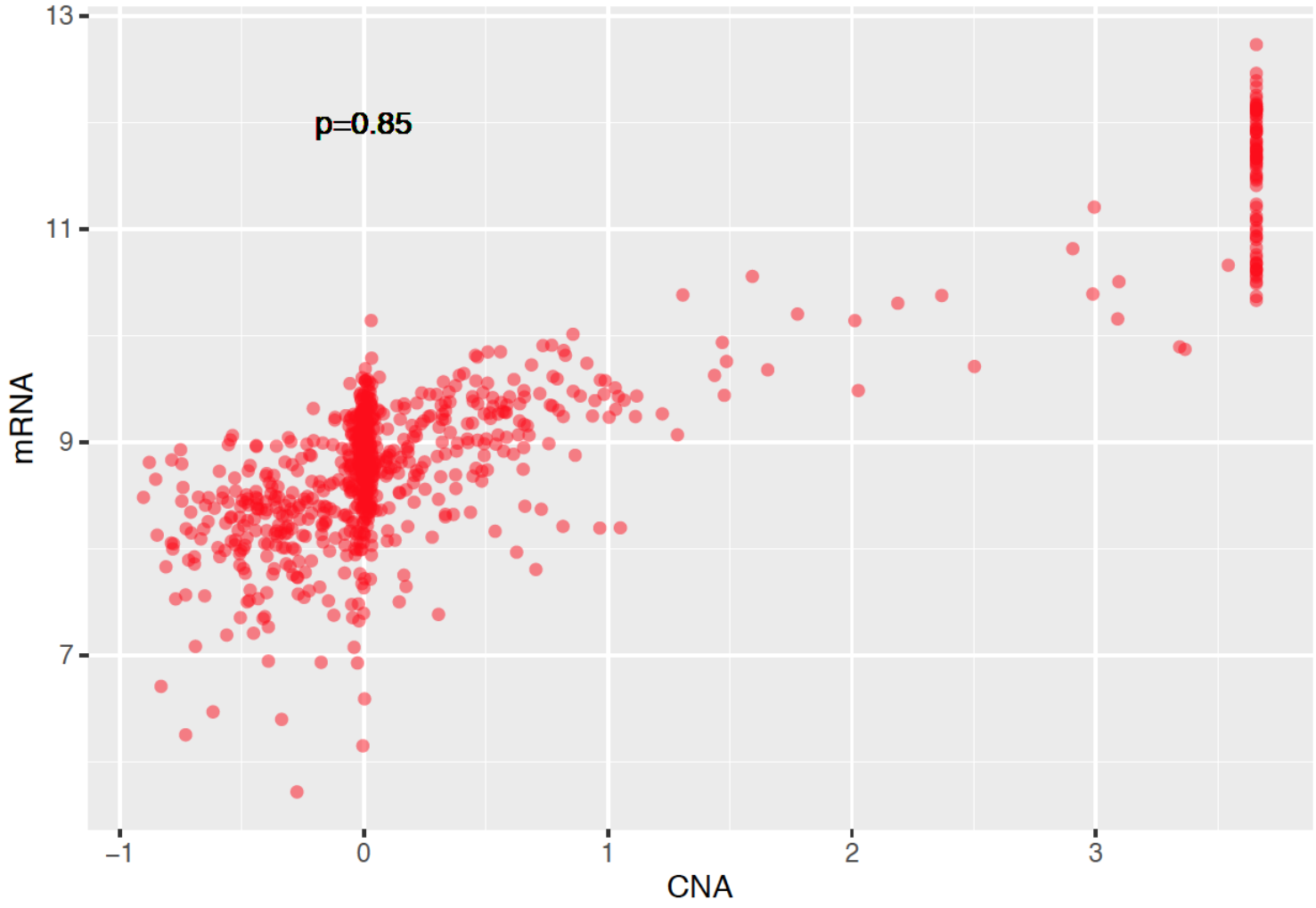
```r
install.packages("ggplot2")  ## library install
```

```r
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
```



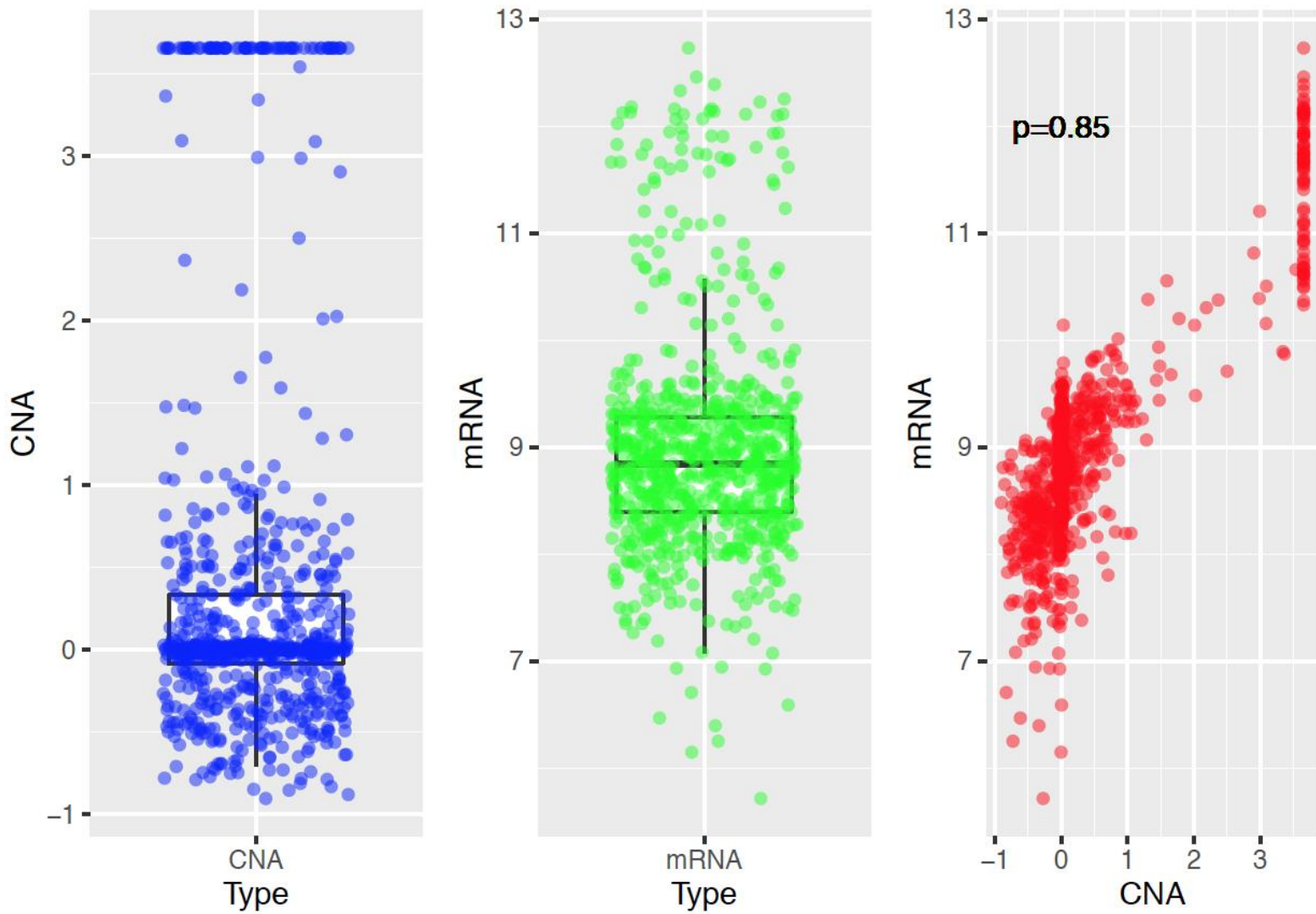
```r
input.mrna = data.frame(log(mrnadata), "mRNA", stringsAsFactors = F)
colnames(input.mrna) = c("mRNA", "Type")

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
```

```r
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
```



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

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



Heatmap

Install NMF package

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

```
library(NMF)
```

```r
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_cnadata = getProfileData(mycgds, geneset, cna, mycaselist)  #cna data

geneset_cnadata = geneset_cnadata[c(1:44), ]  #sample selection

geneset_cnadata = t(geneset_cnadata)  # transpose data

geneset_cnadata[is.na(geneset_cnadata)] <- 0
# View(geneset_cnadata)
```

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

Heatmap(geneset_cnadata, top_annotation = ann_col)
```