



Explore the data

Anja Bråthen Kristoffersen

[Biomedical Research Group](#)



UNIVERSITY
OF OSLO

Outline

- Basics:
 - Boxplots
 - Histograms
 - Scatter plots
 - Transformations
 - QQ-plot
- Applications to microarray data

Outline

- Mostly graphical
- Plotting the raw data (histograms, scatterplots, etc.)
- Plotting simple statistics such as means, standard deviations, medians, box plots, etc
- Positioning such plots so as to maximize our natural pattern-recognition abilities
- A **clear** picture is worth a thousand words!

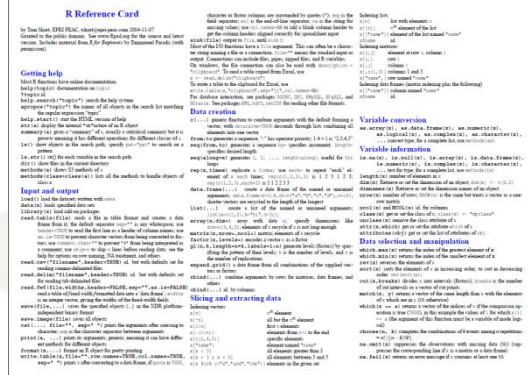
R

- r-project.org and bioconductor.org

- R reference card

<http://cran.r-project.org/doc/contrib/Short-refcard.pdf>

- *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* by Gentleman et al.



What is R?

- R (<http://www.r-project.org>). R is a language and environment for statistical computing and graphics
- Provide many statistical techniques
- R provides a great environment for plotting with great graphics capabilities
- Open source
- Highly extensible (e.g. CRAN, Bioconductor)

Probability distributions

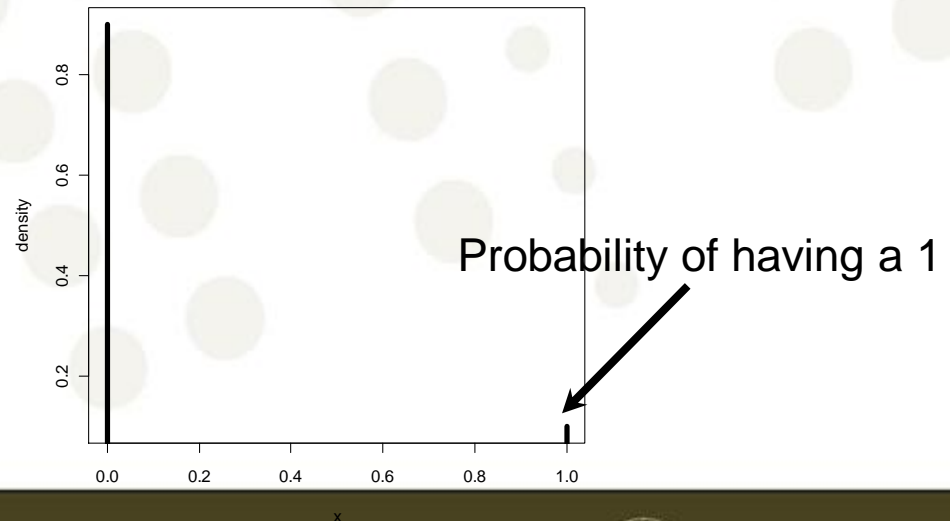
Can be either discrete or continuous (uniform, bernoulli, normal, etc)

Defined by a density function, $p(x)$ or $f(x)$

Bernoulli distribution $Be(p)$

Flip a coin (T=0, H=1). Probability of H is 0.1.

```
x<-0:1  
f<-dbinom(x, size=1, prob=0.1)  
plot(x,f,xlab="x",ylab="density",type="h",lwd=5)
```

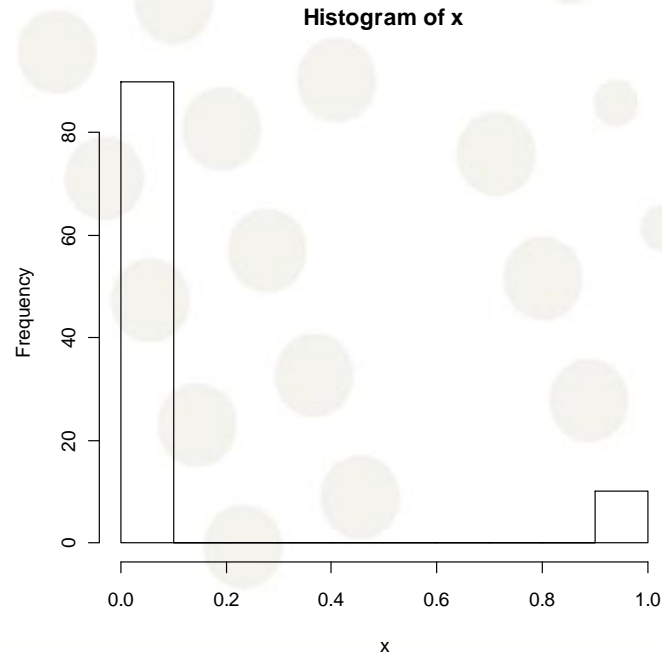


Probability distributions

Random sampling

Generate 100 observations from a $Be(0.1)$

```
x<-rbinom(100, size=1, prob=0.1)  
hist(x)
```

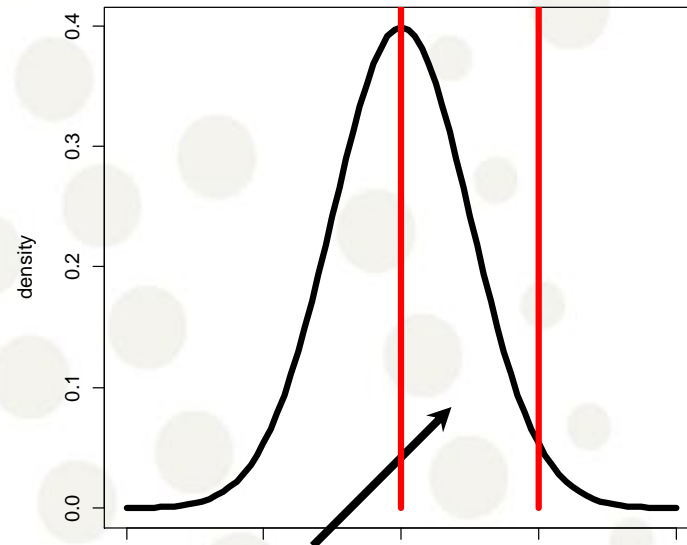


Probability distributions

Normal distribution $N(\mu, \sigma^2)$

μ is the mean and σ^2 is the variance

```
x<-seq(-4,4,0.1)
f<-dnorm(x, mean=0, sd=1)
plot(x,f,xlab="x",ylab="density",lwd=5,type="l")
lines(c(0,0),c(0,0.5), col=2, lwd=5)
lines(c(2,2),c(0,0.5), col=2, lwd=5)
```



Area under the curve is the probability of having an observation between 0 and 2.

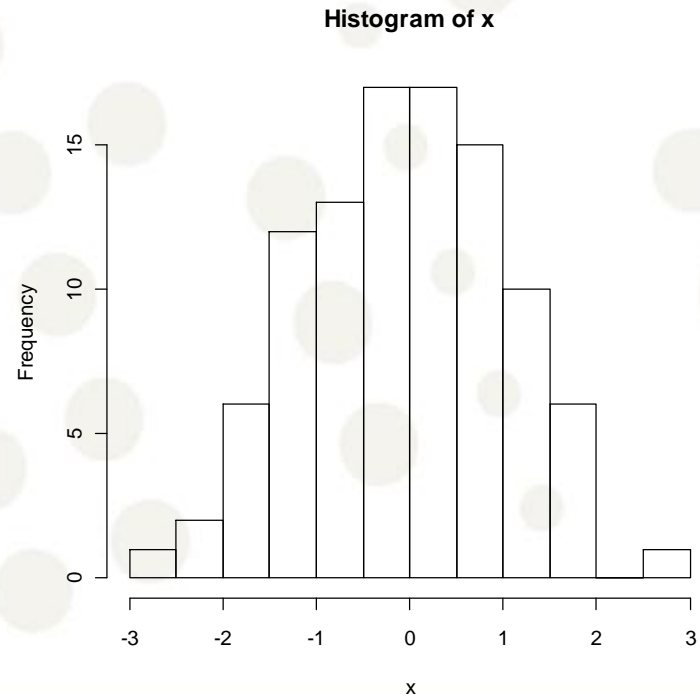
Probability distributions

Random sampling

Generate 100 observations from a $N(0,1)$

```
x<-rnorm(100, mean=0, sd=1)  
hist(x)
```

Histograms can be used
to estimate densities!

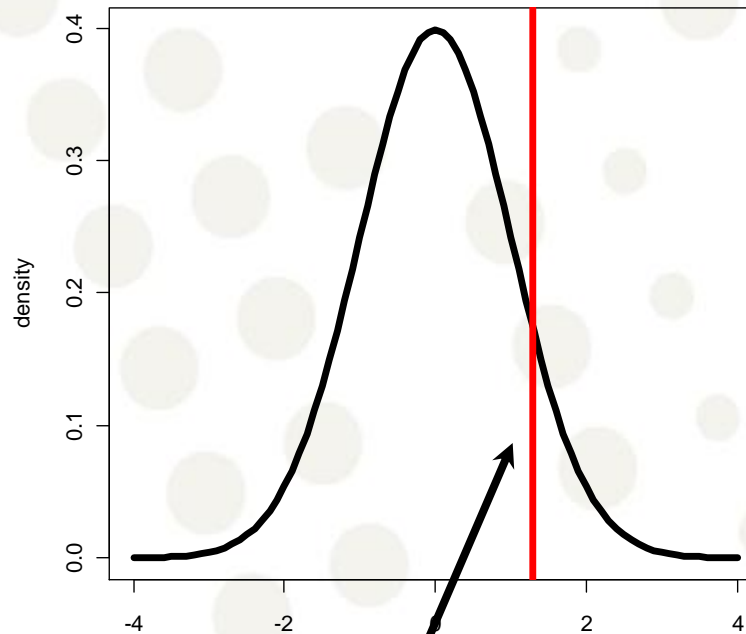


Quantiles

(Theoretical) Quantiles: The p -quantile is the value with the property that there is a probability p of getting a value less than or equal to it.

```
q90<-qnorm(0.90, mean = 0, sd = 1)
x<-seq(-4,4,.1)
f<-dnorm(x, mean=0, sd=1)
plot(x,f,xlab="x",ylab="density",type="l",lwd=5)
abline(v=q90,col=2,lwd=5)
```

The 50% quantile is called the median



90% of the prob. (area under the curve) is on the left of red vertical line.

Descriptive Statistics

Empirical Quantiles: The p -quantile is the value with the property that $p\%$ of the observations are less than or equal to it.

Empirical quantiles can easily be obtained in R.

```
x<-rnorm(100, mean=0, sd=1)  
quantile(x)
```

0%	25%	50%	75%	100%
-2.2719255	-0.6088466	-0.0594199	0.6558911	2.5819589

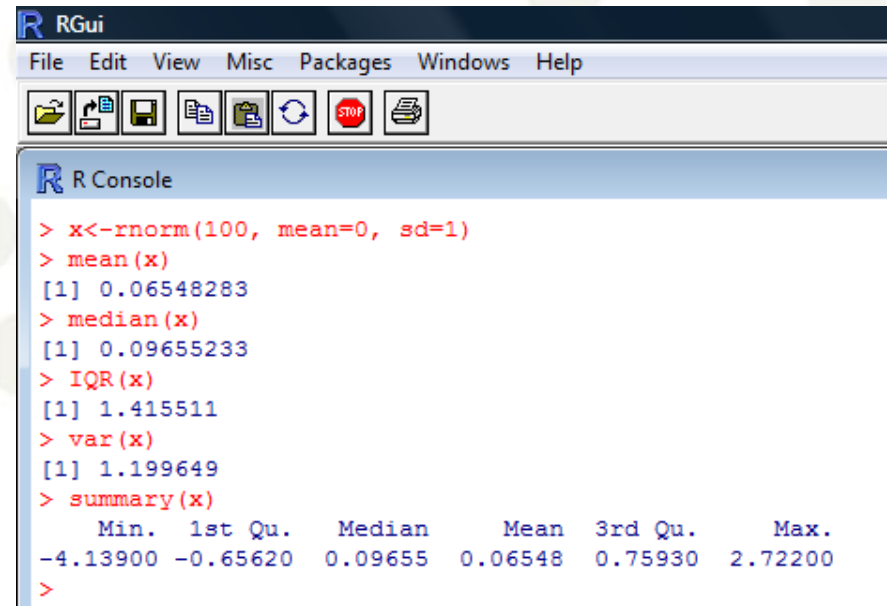
```
quantile(x, probs=c(0.1,0.2,0.9))
```

10%	20%	90%
-1.1744996	-0.8267067	1.3834892

Descriptive Statistics

We often need to quickly ‘quantify’ a data set. This can be done using a set of **summary statistics** (mean, median, variance, standard deviation)

```
x<-rnorm(100, mean=0, sd=1)
mean(x)
median(x)
IQR(x)
var(x)
sd(x)
summary(x)
```



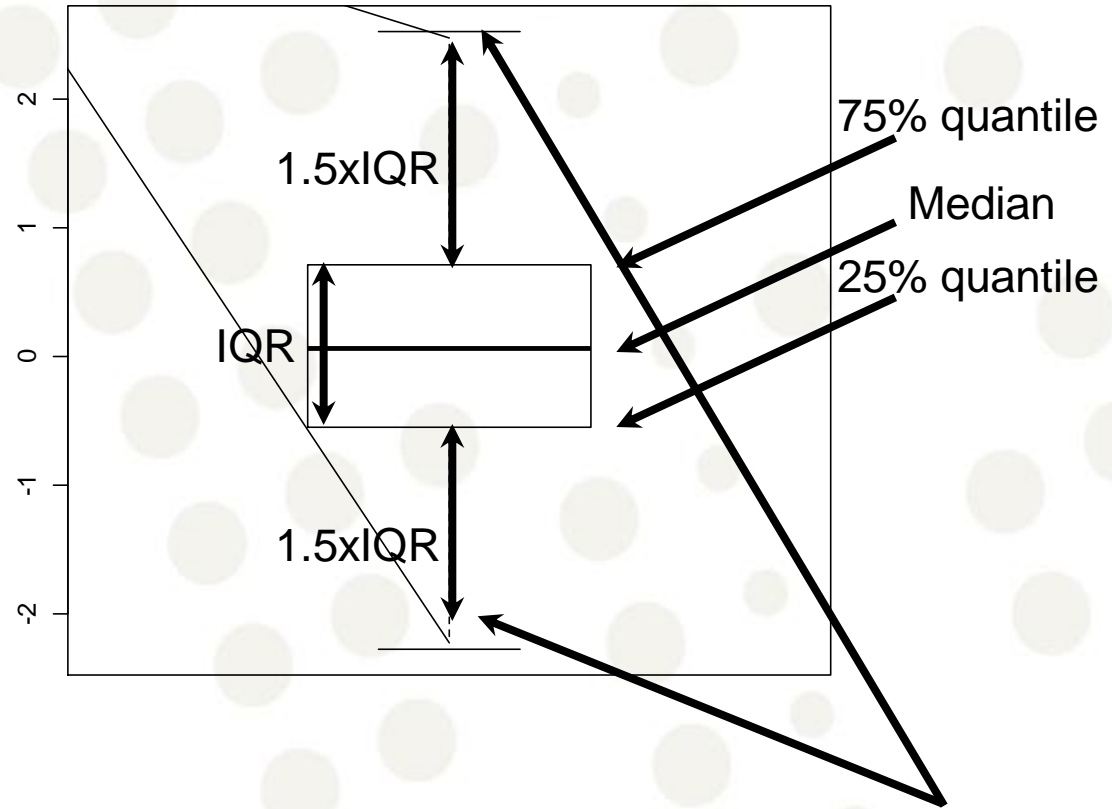
```
RGui
File Edit View Misc Packages Windows Help
R Console
> x<-rnorm(100, mean=0, sd=1)
> mean(x)
[1] 0.06548283
> median(x)
[1] 0.09655233
> IQR(x)
[1] 1.415511
> var(x)
[1] 1.199649
> summary(x)
   Min.   1st Qu.   Median     Mean   3rd Qu.    Max.
-4.13900 -0.65620  0.09655  0.06548  0.75930  2.72200
>
```

‘summary’ can be used for almost any R object!

R is object oriented (methods/classes).

Descriptive Statistics - Box-plot

```
x<-rnorm(100, mean=0, sd=1)  
boxplot(x)
```



$IQR = 75\% \text{ quantile} - 25\% \text{ quantile} = \text{Inter Quantile Range}$

Everything above or below are considered outliers

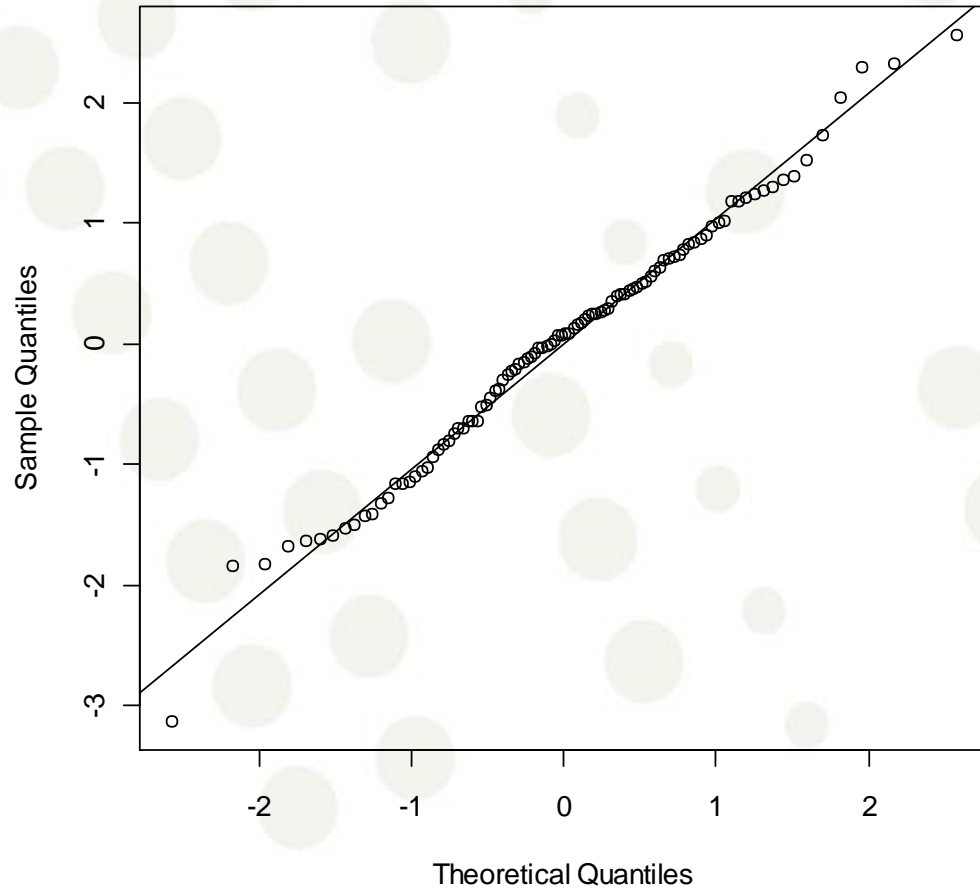
QQ-plot

- Many statistical methods make some assumption about the distribution of the data (e.g. Normal).
- The quantile-quantile plot provides a way to visually verify such assumptions.
- The QQ-plot shows the theoretical quantiles versus the empirical quantiles. If the distribution assumed (theoretical one) is indeed the correct one, we should observe a straight line.

QQ-plot

```
x<-rnorm(100, mean=0, sd=1)  
qqnorm(x)  
qqline(x)
```

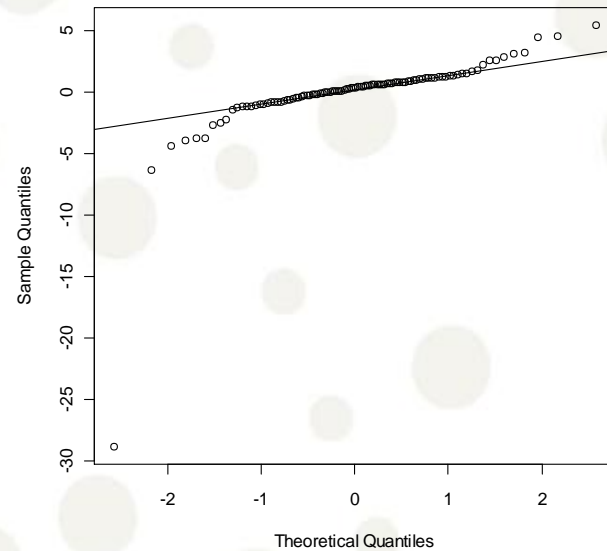
Normal Q-Q Plot



Only valid for the normal distribution!

QQ-plot

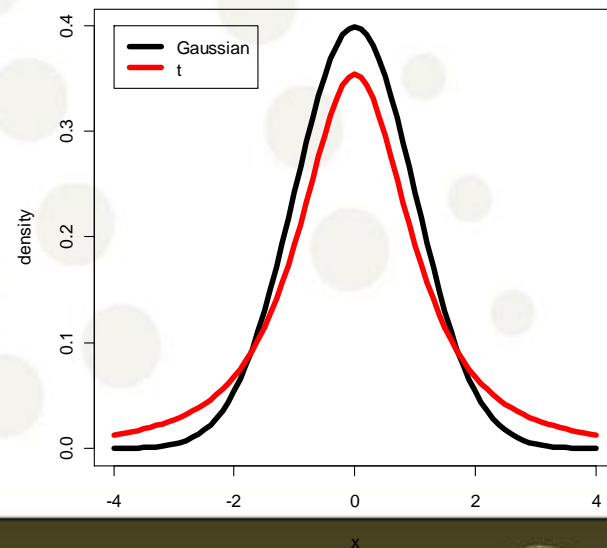
Normal Q-Q Plot



```
x<-rt(100,df=2)
qqnorm(x)
qqline(x)
```

Clearly the t distribution with two degrees of freedom is different from the Normal!

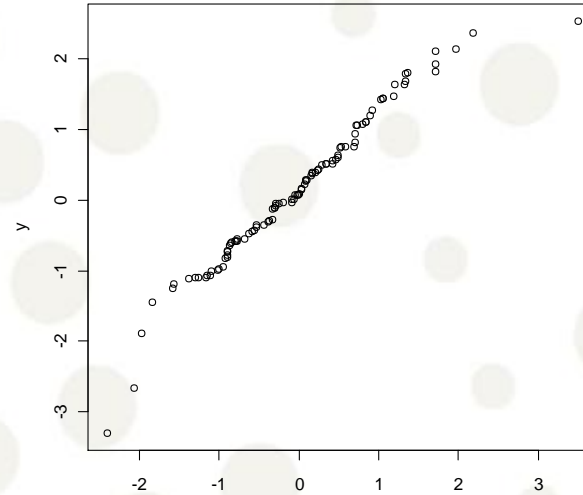
```
x<-seq(-4,4,.1)
f1<-dnorm(x, mean=0, sd=1)
f2<-dt(x, df=2)
plot(x,f1,xlab="x",ylab="density",lwd=5,type="l")
lines(x,f2,xlab="x",ylab="density",lwd=5,col=2)
legend(-4,0.4,c("Gaussian", "t"),col=c(1,2),lty=c(1,1),lwd=5)
```



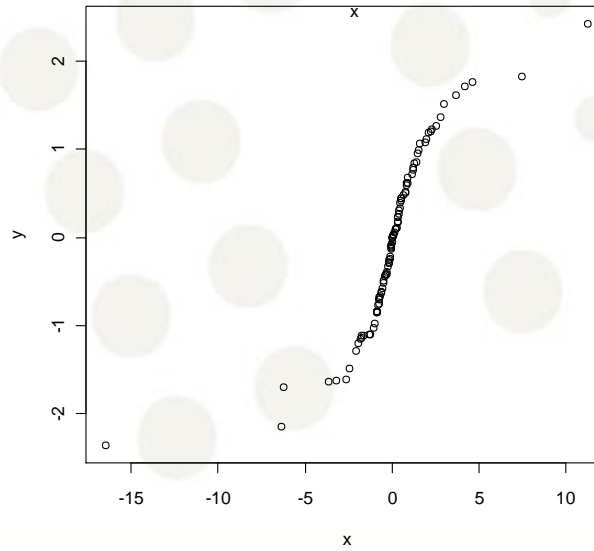
QQ-plot

Comparing two samples

```
x<-rnorm(100, mean=0, sd=1)
y<-rnorm(100, mean=0, sd=1)
qqplot(x,y)
```



```
x<-rt(100, df=2)
y<-rnorm(100, mean=0, sd=1)
qqplot(x,y)
```



Ex: Try with different values of df.

Main idea behind
quantile normalization

<http://Imppp.nih.gov/DLBCL/>

The New England
Journal of Medicine

Copyright © 2002 by the Massachusetts Medical Society

VOLUME 346

JUNE 20, 2002

NUMBER 25



THE USE OF MOLECULAR PROFILING TO PREDICT SURVIVAL
AFTER CHEMOTHERAPY FOR DIFFUSE LARGE-B-CELL LYMPHOMA

ANDREAS ROSENWALD, M.D., GEORGE WRIGHT, PH.D., WING C. CHAN, M.D., JOSEPH M. CONNORS, M.D.,
ELIAS CAMPO, M.D., RICHARD I. FISHER, M.D., RANDY D. GASCOYNE, M.D., H. KONRAD MULLER-HERMELINK, M.D.,
ERLEND B. SMELAND, M.D., PH.D., AND LOUIS M. STAUDT, M.D., PH.D.,
FOR THE LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT

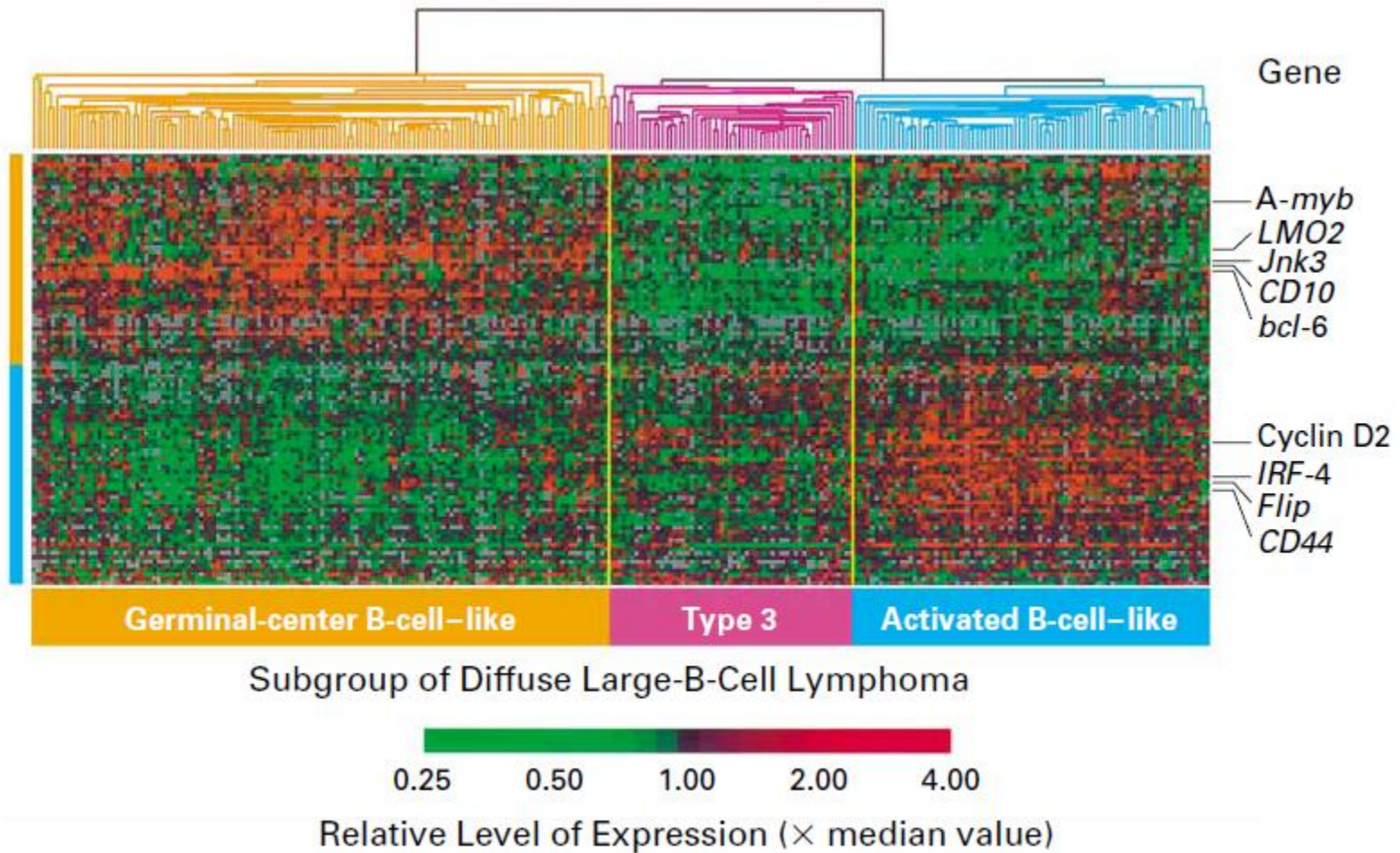
ABSTRACT

Background The survival of patients with diffuse large-B-cell lymphoma after chemotherapy is influenced by molecular features of the tumors. We used the gene-expression profiles of these lymphomas to develop a molecular predictor of survival.

Methods Biopsy samples of diffuse large-B-cell lymphoma

DIFFUSE large-B-cell lymphoma, the most common type of lymphoma in adults, can be cured by anthracycline-based chemotherapy in only 35 to 40 percent of patients.¹ The multiple unsuccessful attempts to increase this rate² suggest that diffuse large-B-cell lymphoma

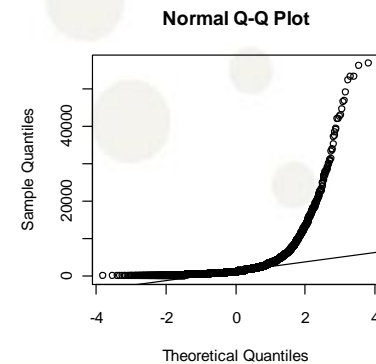
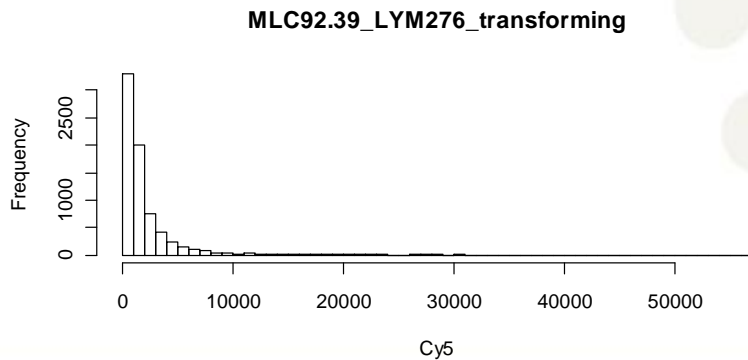
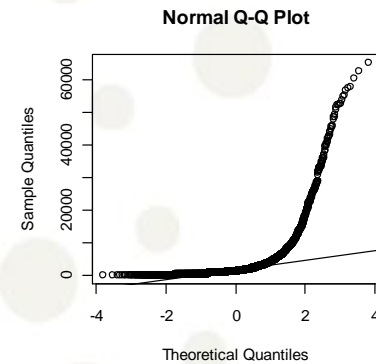
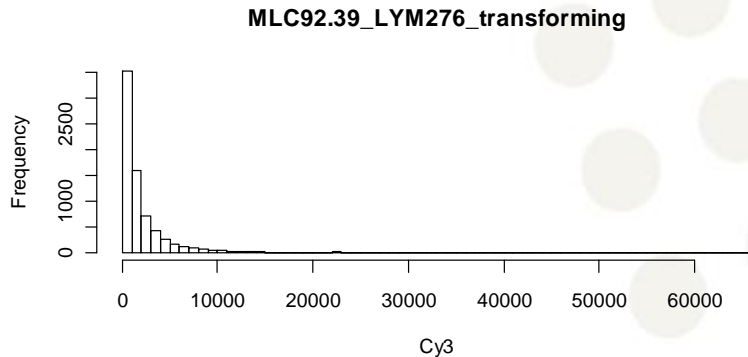
Microarray data



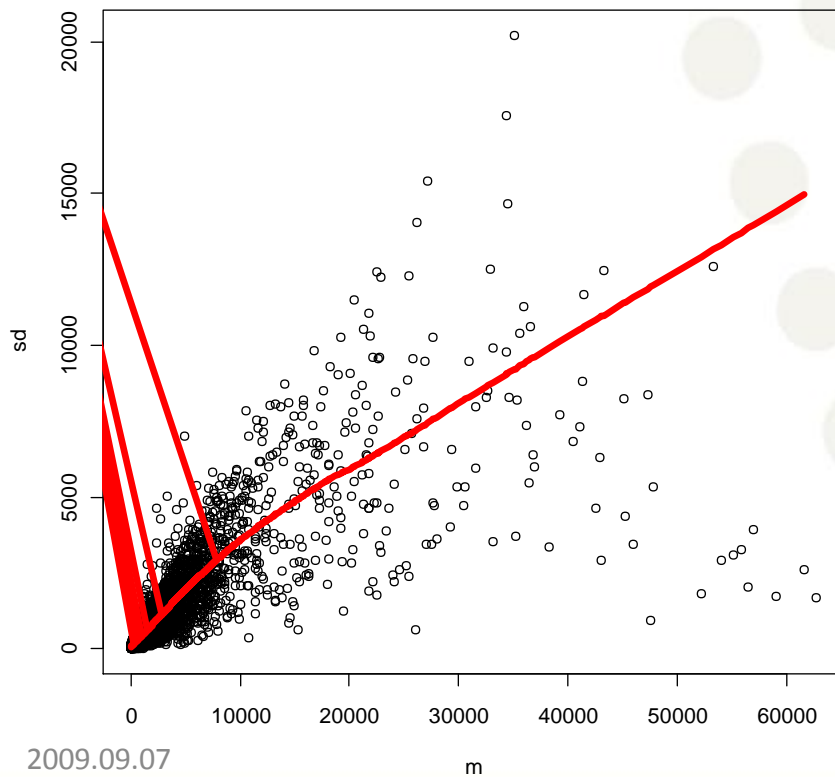
Microarray raw data

```
Cy3 <- read.table(file="NEJM_Web_Fig1data_CY3.txt", header=TRUE, sep="\t", dec=",")  
Cy5 <- read.table(file="NEJM_Web_Fig1data_CY5.txt", header=TRUE, sep="\t", dec=",")  
  
par(mfrow=c(2,1))  
hist(Cy3[,55], 50, main=names(Cy3)[55], xlab="Cy3")  
hist(Cy5[,55], 50, main=names(Cy3)[55], xlab="Cy5")
```

```
par(mfrow=c(2,1))  
qqnorm(Cy3[,55])  
qqline(Cy3[,55])  
qqnorm(Cy5[,55])  
qqline(Cy5[,55])
```



Standard deviation depends on signal



```
# 'apply' will apply the function to all rows of the data matrix  
m <- apply(Cy3[,55:58],1,mean,na.rm=TRUE)  
sd <- apply(Cy3[,55:58],1,sd,na.rm=TRUE)  
plot(m,sd)  
trend<-lowess(m,sd)  
lines(trend,col=2,lwd=5)
```

— lowess fit

LOcally WEighted Scatter plot Smoother
used to estimate the trend in a scatter plot

Non parametric!

Transformations

Observations:

The data are highly skewed.

The standard deviation is not constant as it increases with the mean.

Solution:

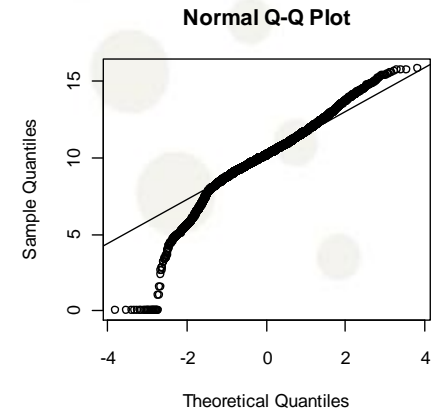
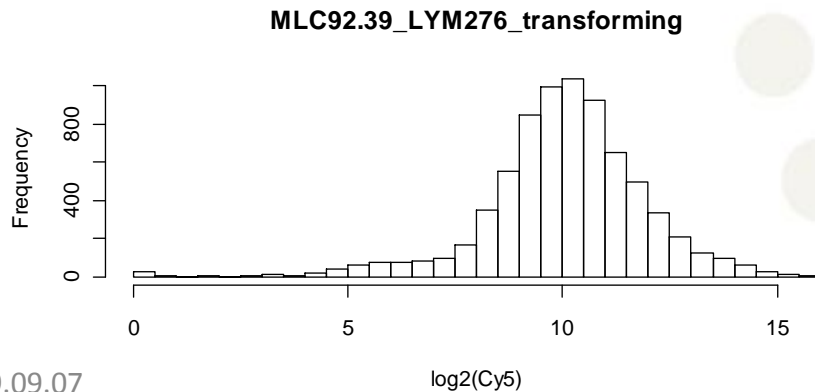
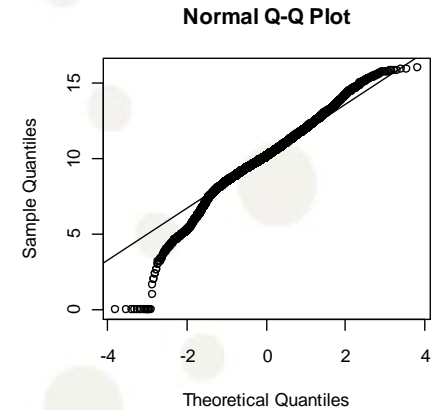
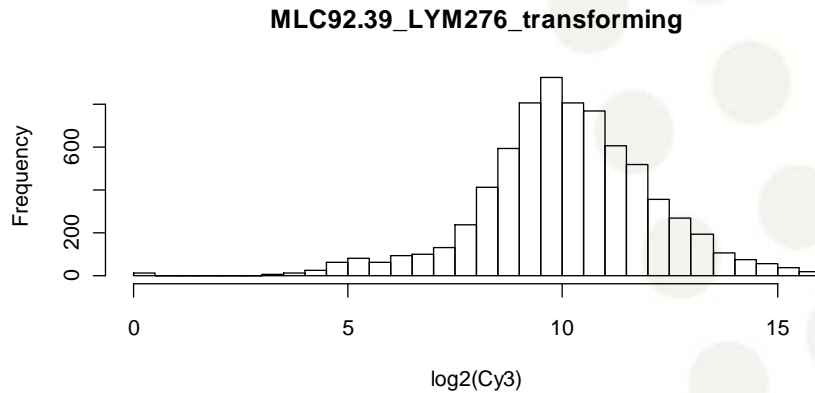
Look for a transformation that will make the data more symmetric and the variance more constant.

With positive data the log transformation is often appropriate.

Transformation

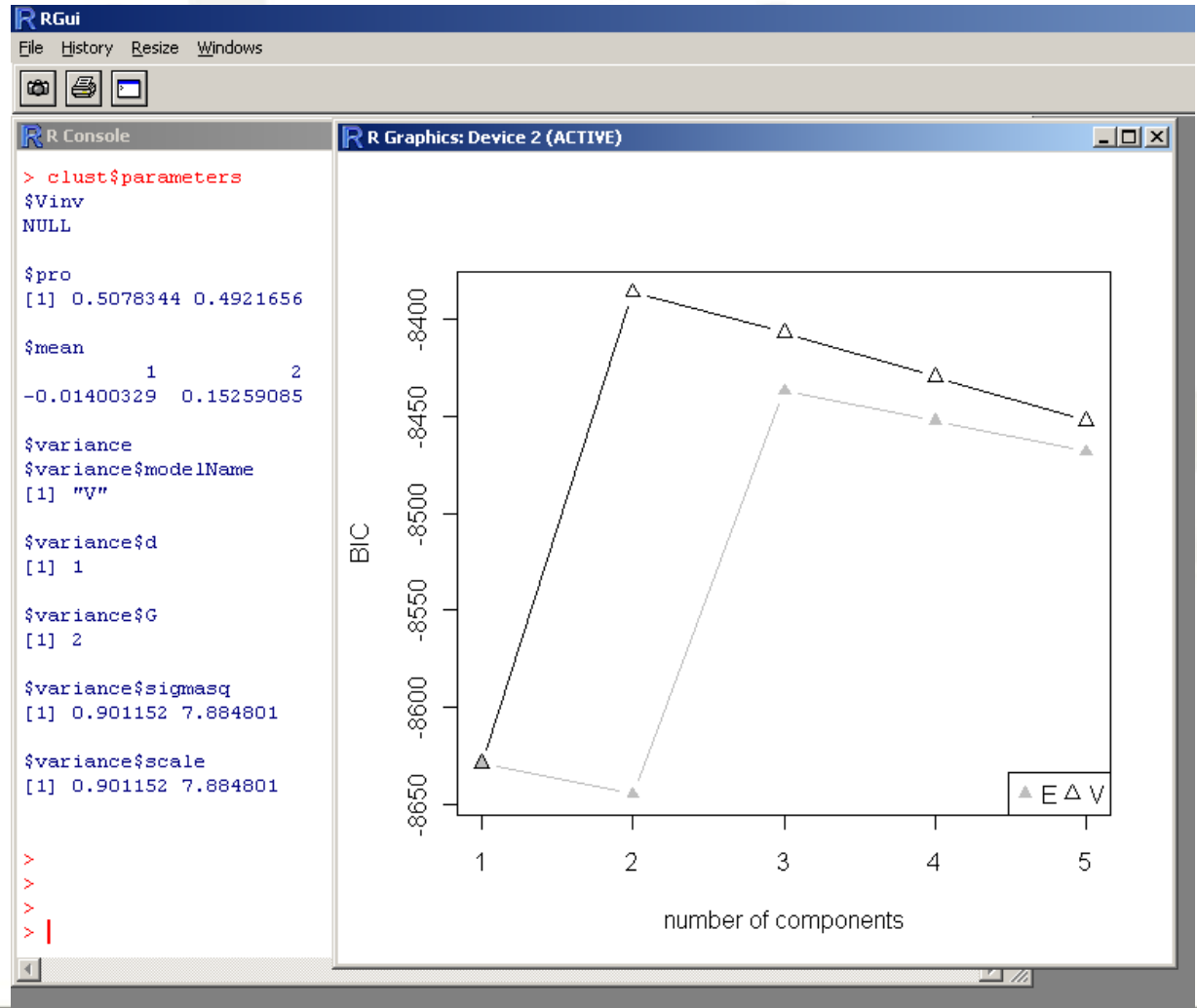
```
hist(log2(Cy3[,55]), 50, main=names(Cy3)[55], xlab="log2(Cy3)")  
hist(log2(Cy5[,55]), 50, main=names(Cy3)[55], xlab="log2(Cy5)")
```

```
par(mfrow=c(2,1))  
qqnorm(log2(Cy3[,55]))  
qqline(log2(Cy3[,55]))  
qqnorm(log2(Cy5[,55]))  
qqline(log2(Cy5[,55]))
```



One Gaussian distribution?

```
library(mclust)
y<-rnorm(1000,0,1)
x<-rnorm(1000,0,3)
clust <- Mclust(c(x,y), G=1:5)
plot(clust)
clust$parameters
```



R Console

```
> z<-which(is.na(Cy3[,55]))
> clust <- Mclust(log2(Cy3[-z,55]),G=1:5,na.rm=TRUE)
> clust$parameters
```

```
$Vinv
NULL
```

```
$pro
[1] 0.2828964 0.3778001 0.3393036
```

```
$mean
      1      2      3
9.102280 9.670364 11.328825
```

```
$variance
$variance$modelName
[1] "V"
```

```
$variance$d
[1] 1
```

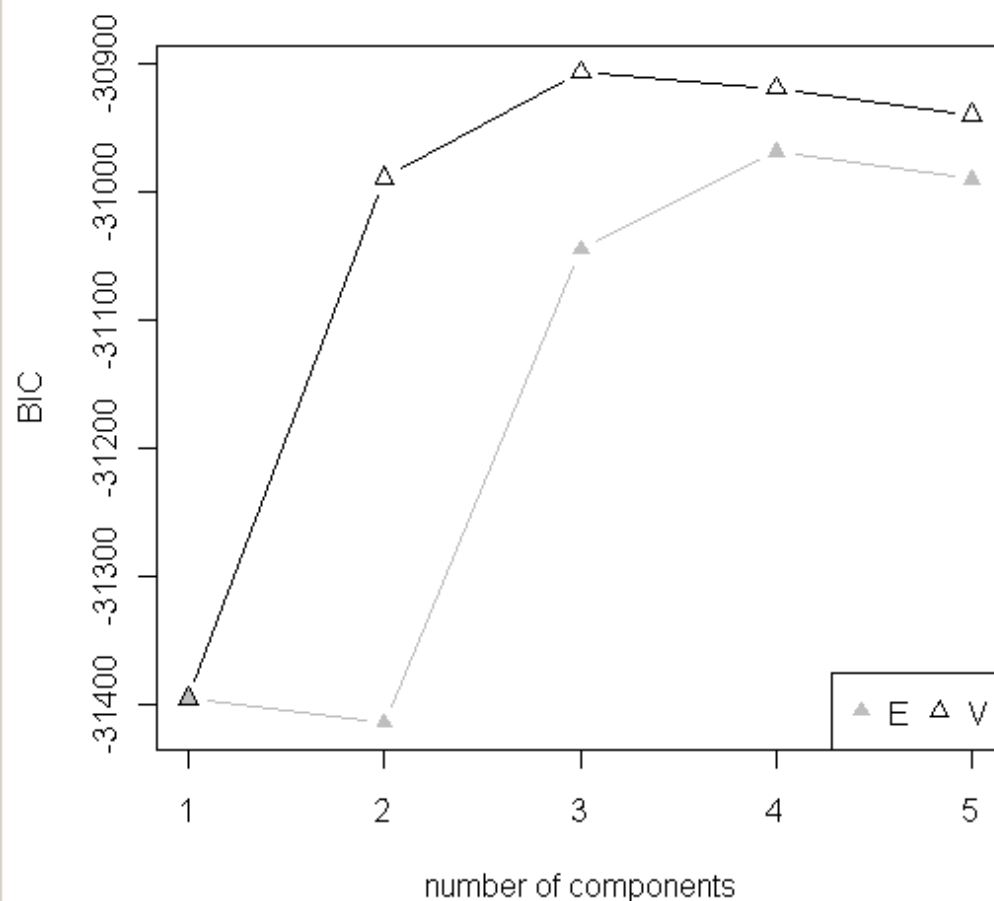
```
$variance$G
[1] 3
```

```
$variance$sigma$sq
[1] 7.4456537 0.9734005 2.2411960
```

```
$variance$scale
[1] 7.4456537 0.9734005 2.2411960
```

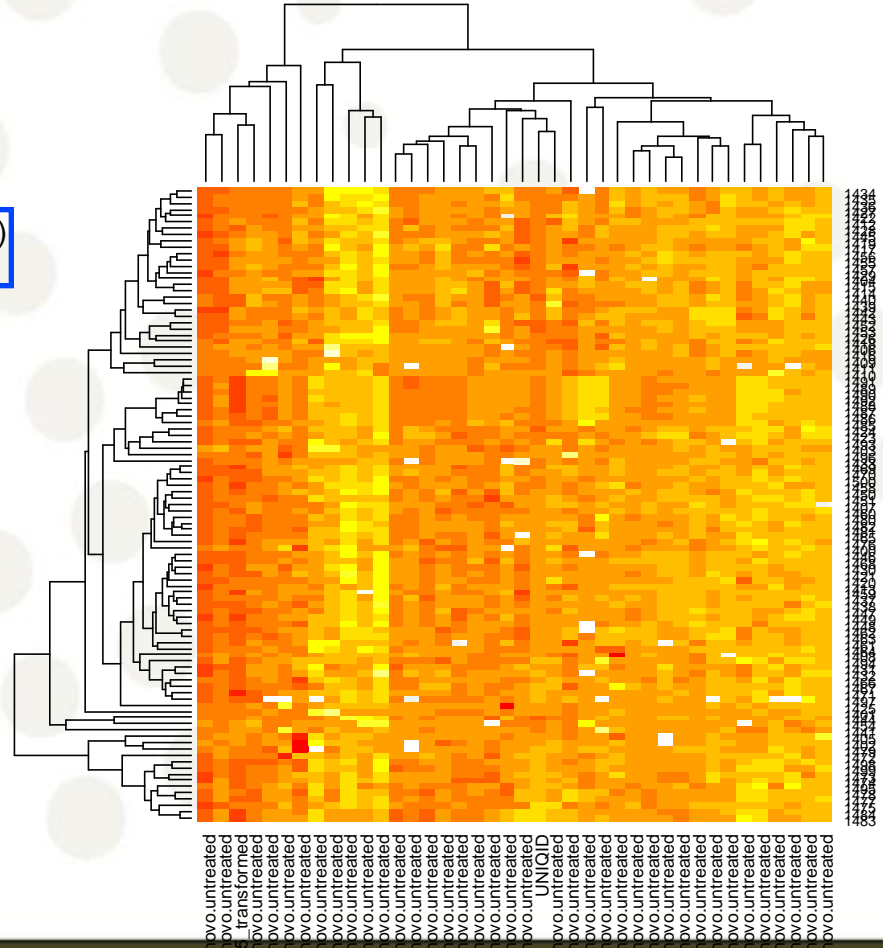
```
> plot(clust)
Waiting to confirm page change...
Warning message:
In plot.Mclust(clust) : data not supplied
> |
```

R Graphics: Device 2 (ACTIVE)

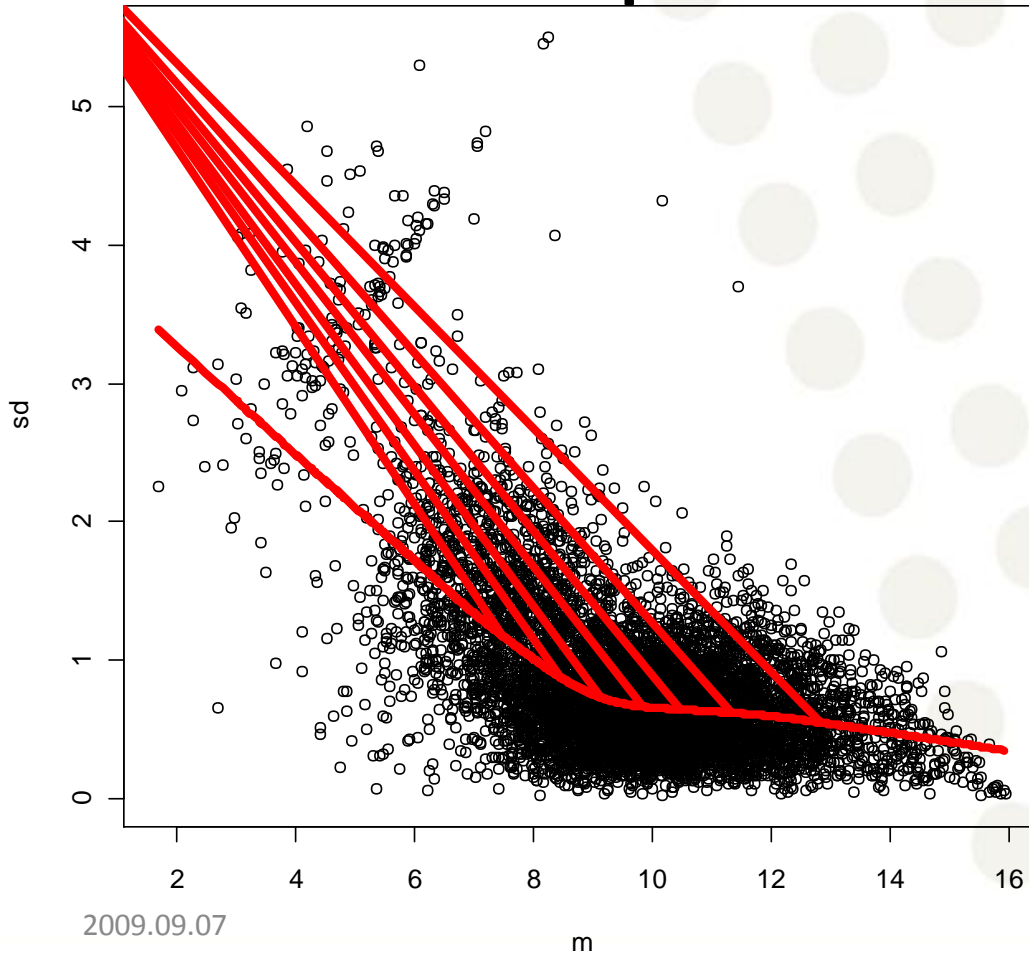


Heatmap

```
z<-as.matrix(log2(Cy3[1400:1500,1:40])-log2(Cy5[1400:1500,1:40]))  
heatmap(z)
```



Standard deviation depends on signal



```
# 'apply' will apply the function to all rows of the data matrix  
m <- apply(log2(Cy3[,55:58]),1,mean,na.rm=TRUE)  
sd <- apply(log2(Cy3[,55:58]),1,sd,na.rm=TRUE)  
plot(m,sd)  
trend<-lowess(m,sd)  
lines(trend,col=2,lwd=5)
```

But the dependency is weaker
Especially where most of the
data are located.

microarray: Always log

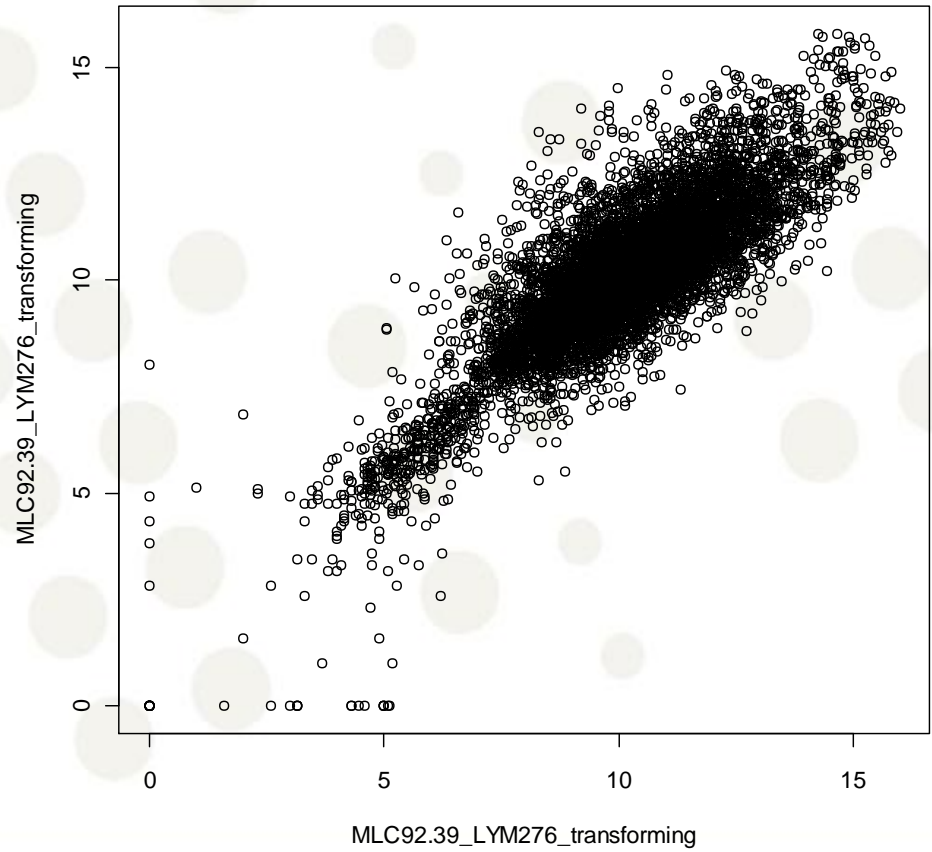
- Makes the data more symmetric, large observations are not as influential
- The variance is (more) constant
- Turns multiplication into addition ($\log(ab)=\log(a)+\log(b)$)
- In practice use log base 2, $\log_2(x)=\log(x)/\log(2)$

gene expression

```
plot(Cy3[,55],Cy5[,55], xlab=names(Cy3)[55], ylab=names(Cy5)[55])
```

What can you say?

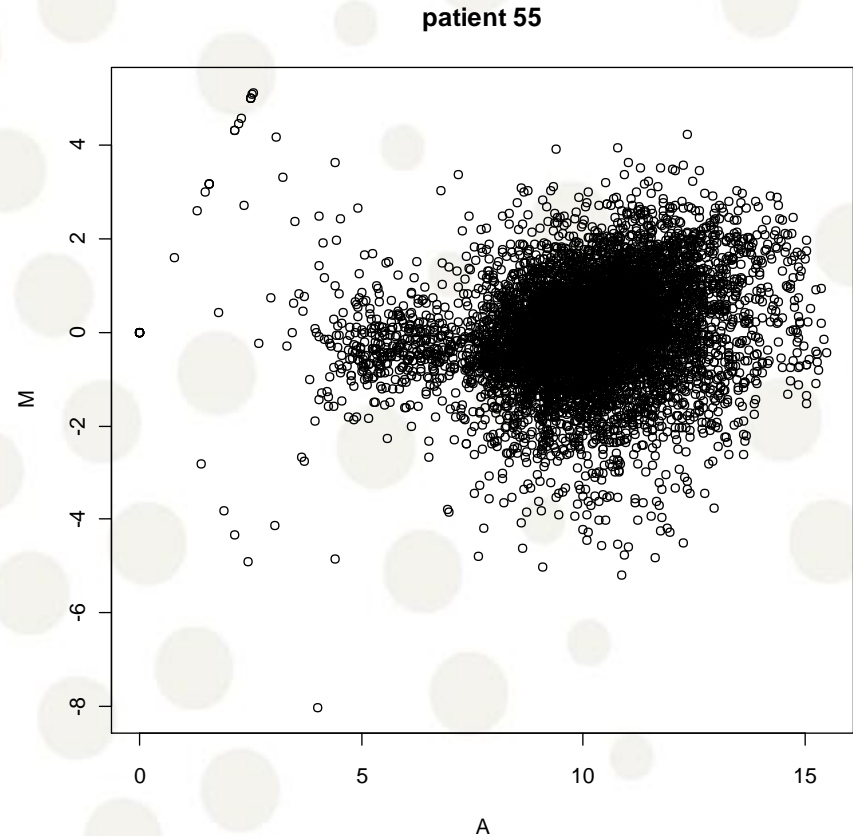
Is this the best way to look at the data?



MA plots

```
# MA plots per replicate  
A<-(log2(Cy3[,55])+log2(Cy5[,55]))/2  
M<-(log2(Cy3[,55])-log2(Cy5[,55]))  
plot(A,M,xlab="A",ylab="M",main="patient 55")
```

M (minus) is the log ratio
A (average) is overall intensity



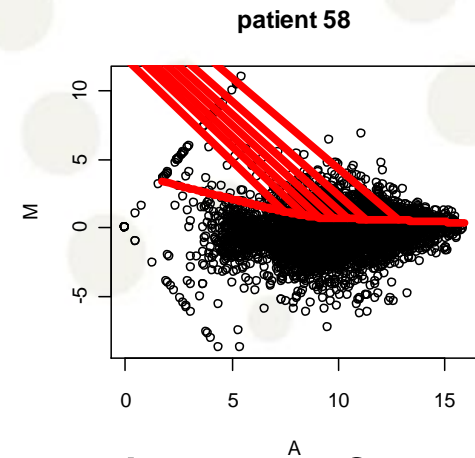
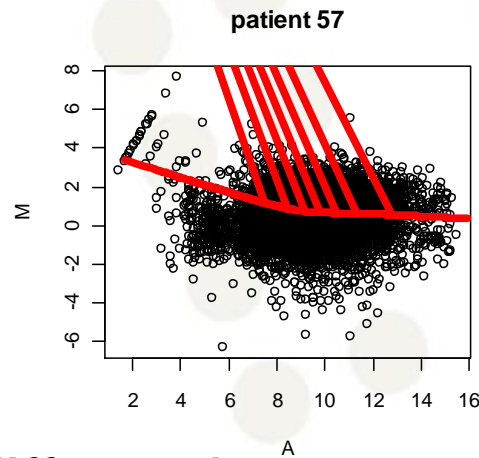
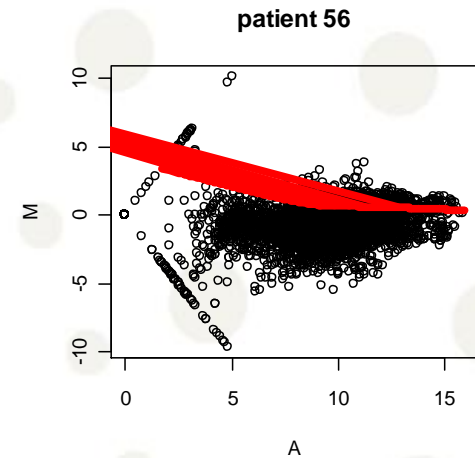
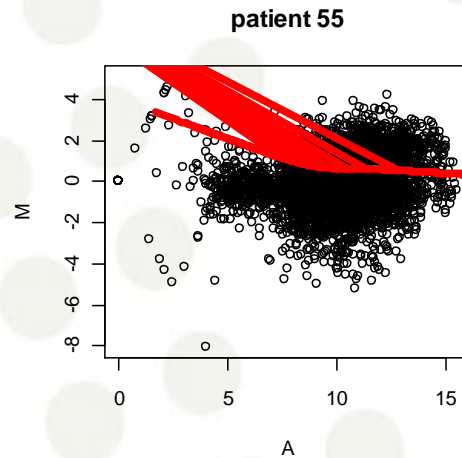
MA plots

```
par(mfrow=c(2,2))
A<-(log2(Cy3[,55])+log2(Cy5[,55]))/2
M<-(log2(Cy3[,55])-log2(Cy5[,55]))
plot(A,M,xlab="A",ylab="M",main="patient 55")
trend<-lowess(A,M)
lines(trend,col=2,lwd=5)

A<-(log2(Cy3[,56])+log2(Cy5[,56]))/2
M<-(log2(Cy3[,56])-log2(Cy5[,56]))
plot(A,M,xlab="A",ylab="M",main="patient 56")
trend<-lowess(A,M)
lines(trend,col=2,lwd=5)

A<-(log2(Cy3[,57])+log2(Cy5[,57]))/2
M<-(log2(Cy3[,57])-log2(Cy5[,57]))
plot(A,M,xlab="A",ylab="M",main="patient 57")
trend<-lowess(A,M)
lines(trend,col=2,lwd=5)

A<-(log2(Cy3[,58])+log2(Cy5[,58]))/2
M<-(log2(Cy3[,58])-log2(Cy5[,58]))
plot(A,M,xlab="A",ylab="M",main="patient 58")
trend<-lowess(A,M)
lines(trend,col=2,lwd=5)
```



How do we find differentially expressed genes?

Combining micro-array and survival data

- For each patient five signature are calculated from the micro-array as the mean of the signal from each of the group of genes:
 - Germinal.center.B.cell.signature
 - Lymph.node.signature
 - Proliferation.signature
 - BMP6
 - MHC.class.II.signature

head(dat)

```
> dat <- read.table(file = "M:/Undervisning/Statistikk/DLBCLpatientDataNEW.txt", header =TRUE, sep="\t")
> head(dat)
DLBCL.sample..LYM.number. Analysis.Set Follow.up..years. Status.at.follow.up Subgroup IPI.Group
1 2 Training 4.0 Alive GCB Low
2 4 Training 4.9 Alive GCB Medium
3 6 Training 5.6 Alive GCB Low
4 7 Training 12.1 Alive GCB Medium
5 8 Training 0.6 Dead ABC Medium
6 11 Training 0.3 Dead GCB High
Germinal.center.B.cell.signature Lymph.node.signature Proliferation.signature BMP6 MHC.class.II.signature
1 0.28 -0.07 -0.56 0.46 0.57
2 1.01 -1.15 -1.04 0.23 0.63
3 0.83 -2.11 0.52 -0.28 0.38
4 0.89 -1.33 0.01 -0.64 0.93
5 0.27 -1.56 1.56 -0.67 -2.50
6 -0.05 0.06 -0.68 -0.38 -2.32
Outcome.predictor.score
1 -0.23
2 -0.38
3 0.20
4 -0.41
5 1.25
6 0.44
```

summary(dat)

```
> summary(dat)
DLBCL.sample..LYM.number.      Analysis.Set Follow.up..years.
Min.   : 1.00                   Training   :160   Min.   : 0.000
1st Qu.: 91.75                 Validation: 80   1st Qu.: 0.900
Median :177.50                 Median    : 2.800
Mean   :190.29                 Mean      : 4.411
3rd Qu.:284.25                 3rd Qu.  : 7.100
Max.   :439.00                 Max.     :21.800

Status.at.follow.up   Subgroup   IPI.Group
Alive:102             ABC       : 73     High    : 32
Dead :138             GCB      :115    Low     : 82
                        Type III: 52    Medium  :108
                        missing: 1
                        NA's    : 17

Germinal.center.B.cell.signature Lymph.node.signature Proliferation.signature
Min.   :-2.61000          Min.   :-2.6500          Min.   :-1.700000
1st Qu.: -0.91000        1st Qu.: -0.8675        1st Qu.: -0.410000
Median : -0.16000        Median : 0.0600         Median : -0.010000
Mean   : -0.03062        Mean   : 0.0065         Mean   : 0.005958
3rd Qu.: 0.86000        3rd Qu.: 0.8675        3rd Qu.: 0.412500
Max.   : 2.48000        Max.   : 2.9800         Max.   : 2.180000

BMP6                   MHC.class.II.signature Outcome.predictor.score
Min.   :-1.87000        Min.   :-3.020000       Min.   :-1.700000
1st Qu.: -0.65250      1st Qu.: -0.537500     1st Qu.: -0.537500
Median : -0.13500      Median : 0.125000       Median : -0.085000
Mean   : -0.04362      Mean   : -0.006083      Mean   : -0.003208
3rd Qu.: 0.49250      3rd Qu.: 0.680000     3rd Qu.: 0.522500
Max.   : 2.69000      Max.   : 1.890000       Max.   : 2.360000
```

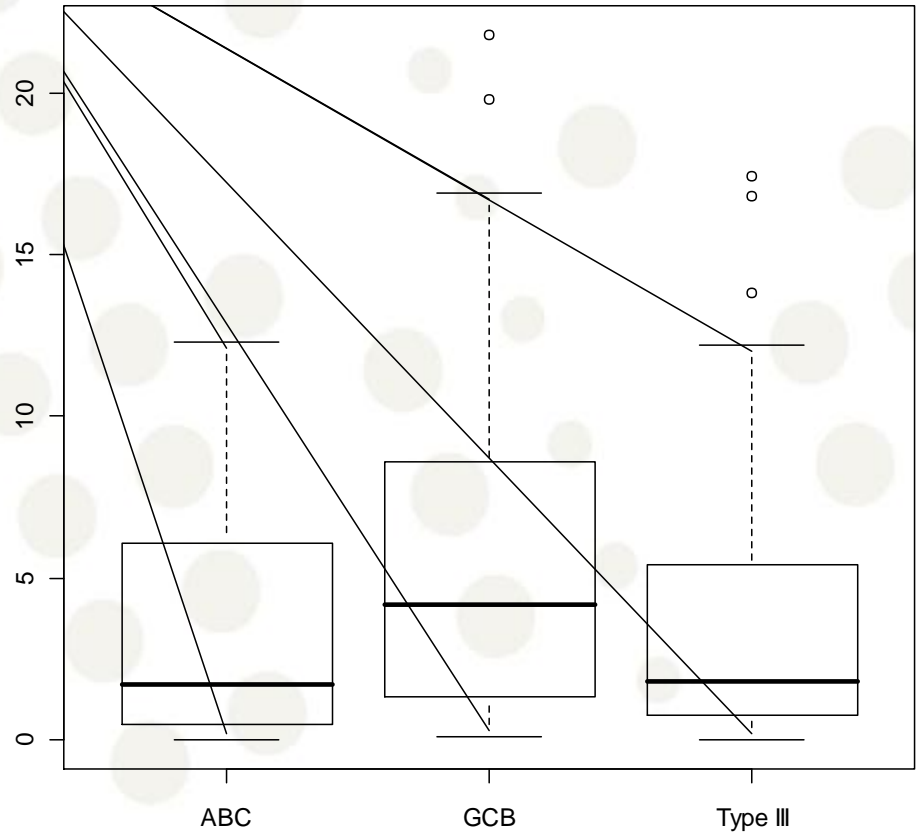
```
> |
```

Boxplot: follow up time for each subgroup

```
boxplot(Follow.up..years.~Subgroup, data = dat)
```

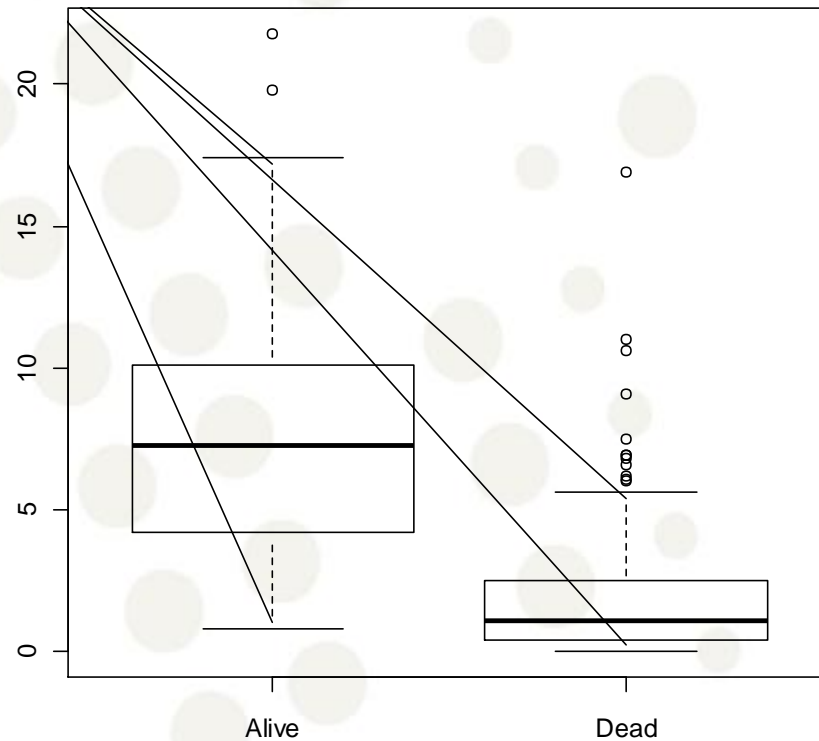
The boxplot function can be used to display several variables at a time!

What can you say here?



Boxplot: follow up time for each subgroup

```
boxplot(Follow.up..years.~Status.at.follow.up, data = dat)
```



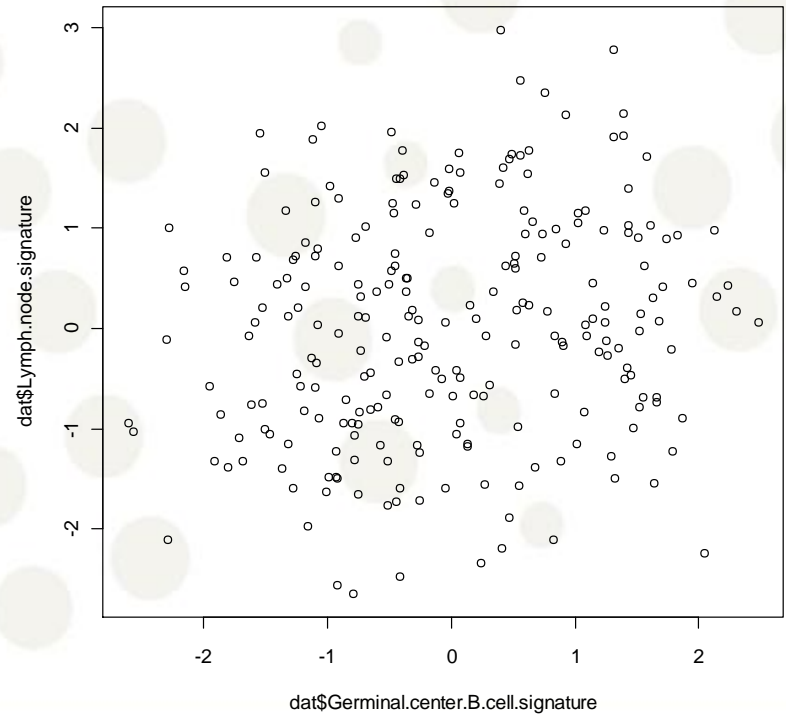
Scatter plots

Biological data sets often contain several variables
So they are **multivariate**.

Scatter plots allow us to look at two variables at a time.

```
plot(dat$Germinal.center.B.cell.signature,dat$Lymph.node.signature)  
cor(dat$Germinal.center.B.cell.signature,dat$Lymph.node.signature)  
#[1] 0.1633608
```

This can be used
to assess **independence!**

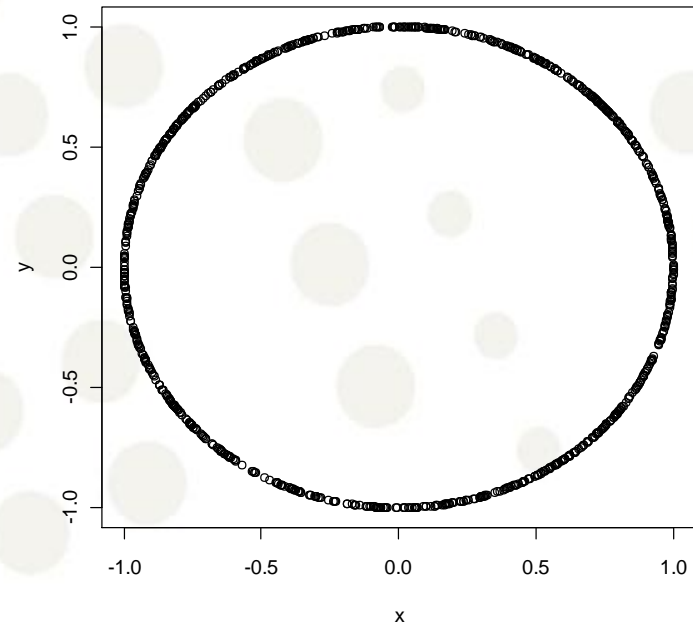


Scatter plots vs. correlations

Correlation is only good for **linear dependence**.

```
# Quick comment on correlation  
theta<-runif(1000,0,2*pi)  
x<-cos(theta)  
y<-sin(theta)  
plot(x,y)  
cor(x,y)
```

What is the correlation?



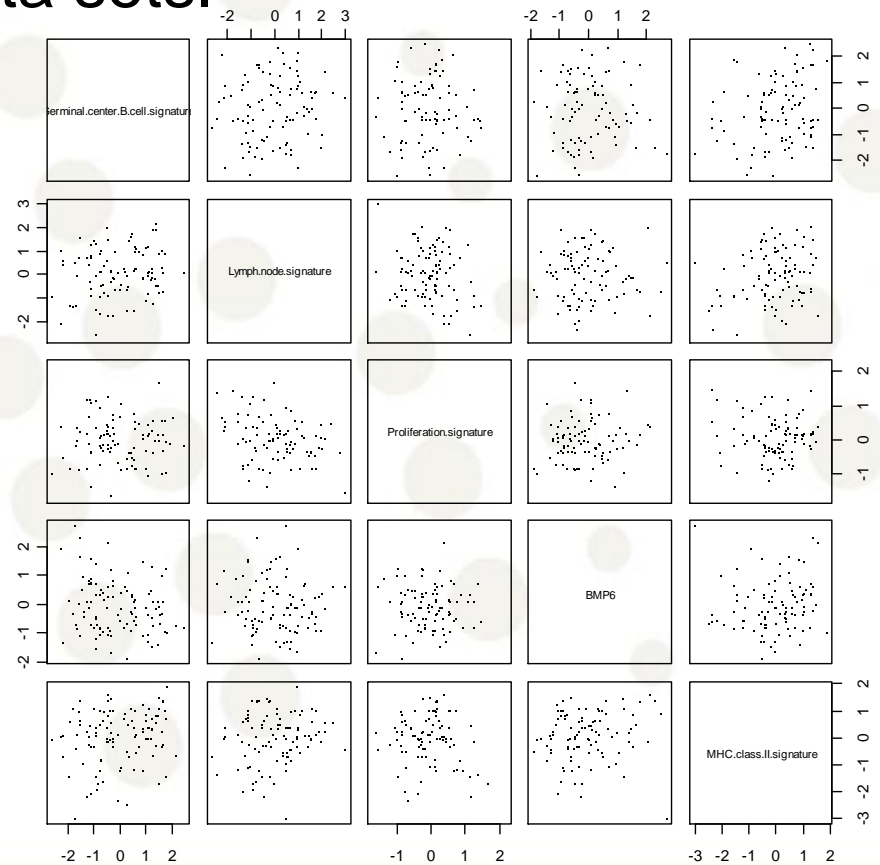
Trellis graphics

Trellis Graphics is a family of techniques for viewing complex, multi-variable data sets.

```
plot(dat[,7:11] pch=".")
```

Note that the plotting symbol is changed.

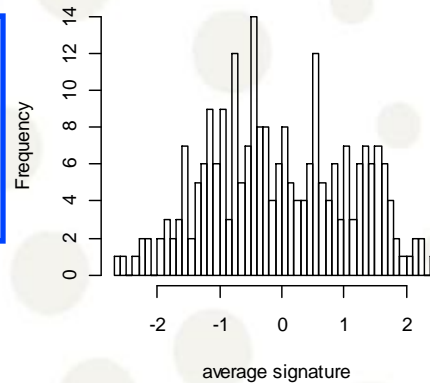
Many more possibilities in the 'lattice' package!



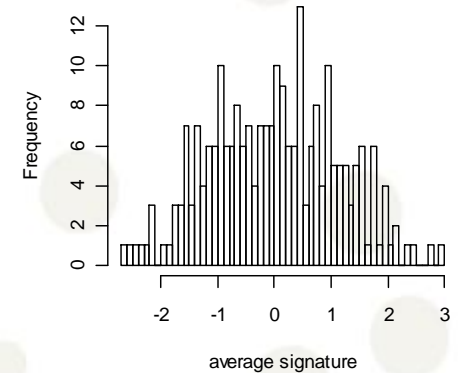
Histogram

```
par(mfrow=c(2,2))  
hist(dat[,7], 50, main = names(dat)[7], xlab="average signature")  
hist(dat[,8], 50, main = names(dat)[8], xlab="average signature")  
hist(dat[,9], 50, main = names(dat)[9], xlab="average signature")  
hist(dat[,10], 50, main = names(dat)[10], xlab="average signature")
```

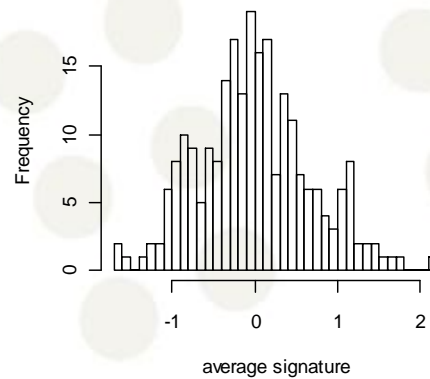
Germinal.center.B.cell.signature



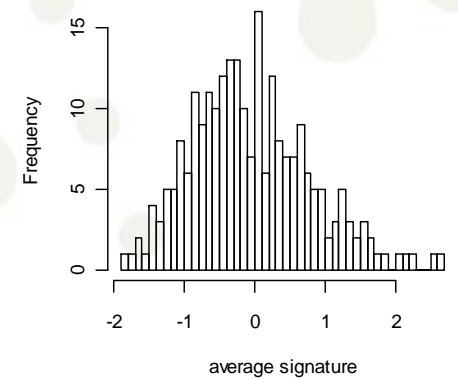
Lymph.node.signature



Proliferation.signature



BMP6



Summary

- Plotting should be the first step in any statistical analysis!
- **Extremely Important**
- Good modeling starts and ends with plotting
- R provides a great framework for plotting