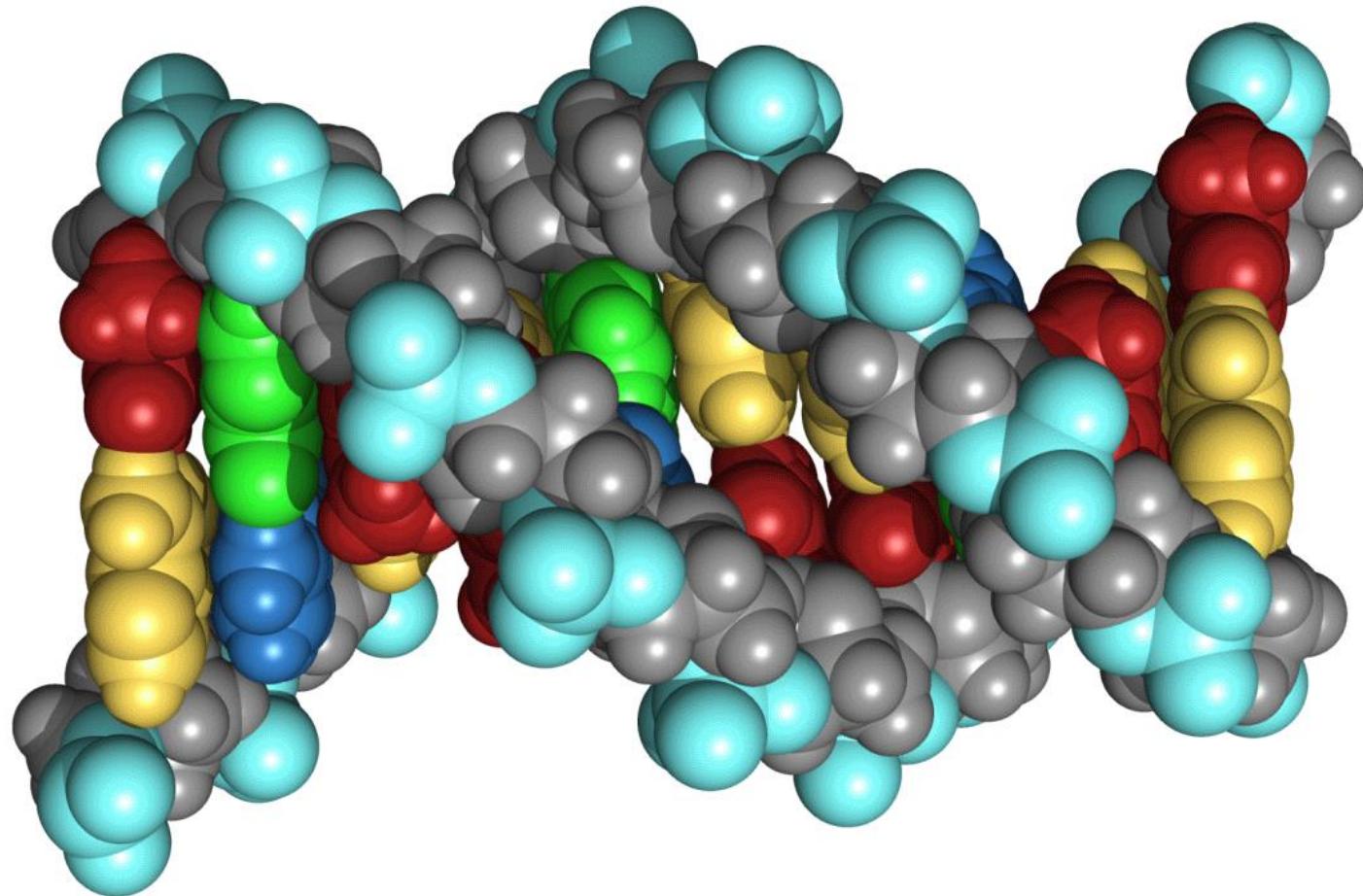
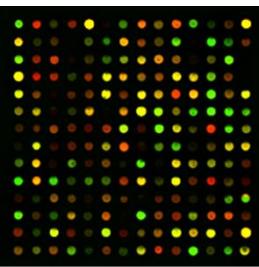


Statistiek 3 BIN

- Les 10



MICROARRAY ANALYSE: STAPPENPLAN



- Background correctie
- Log transformatie
- Normalisatie (bijv. loess)
- Toetsen op DEG's:
 - t -toets, 1-way ANOVA, ...
 - Wilcoxon's toets, Kruskall-Wallis toets, ...
- Aanpassen p -waarden voor multiple toetsing
- Clustering van DEG's:
 - Hiërarchisch clusteren
 - k -means
 - Principale Componenten Analyse (PCA)
- Grafische weergave: heatmaps, volcano plot, ...

MICROARRAY'S: GEN EXPRESSIONS (1-CHANNEL)

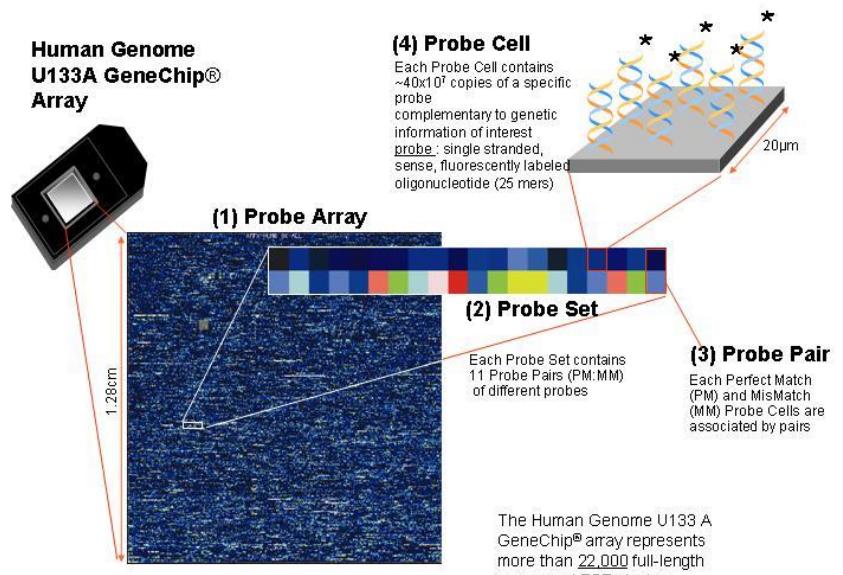
- Voorbeelden:

- Affymetrix: Gene Chip
- Illumina: Bead Chip
- Agilent single-channel arrays
- Applied Microarrays: CodeLink arrays
- Eppendorf: DualChip & Silverquant

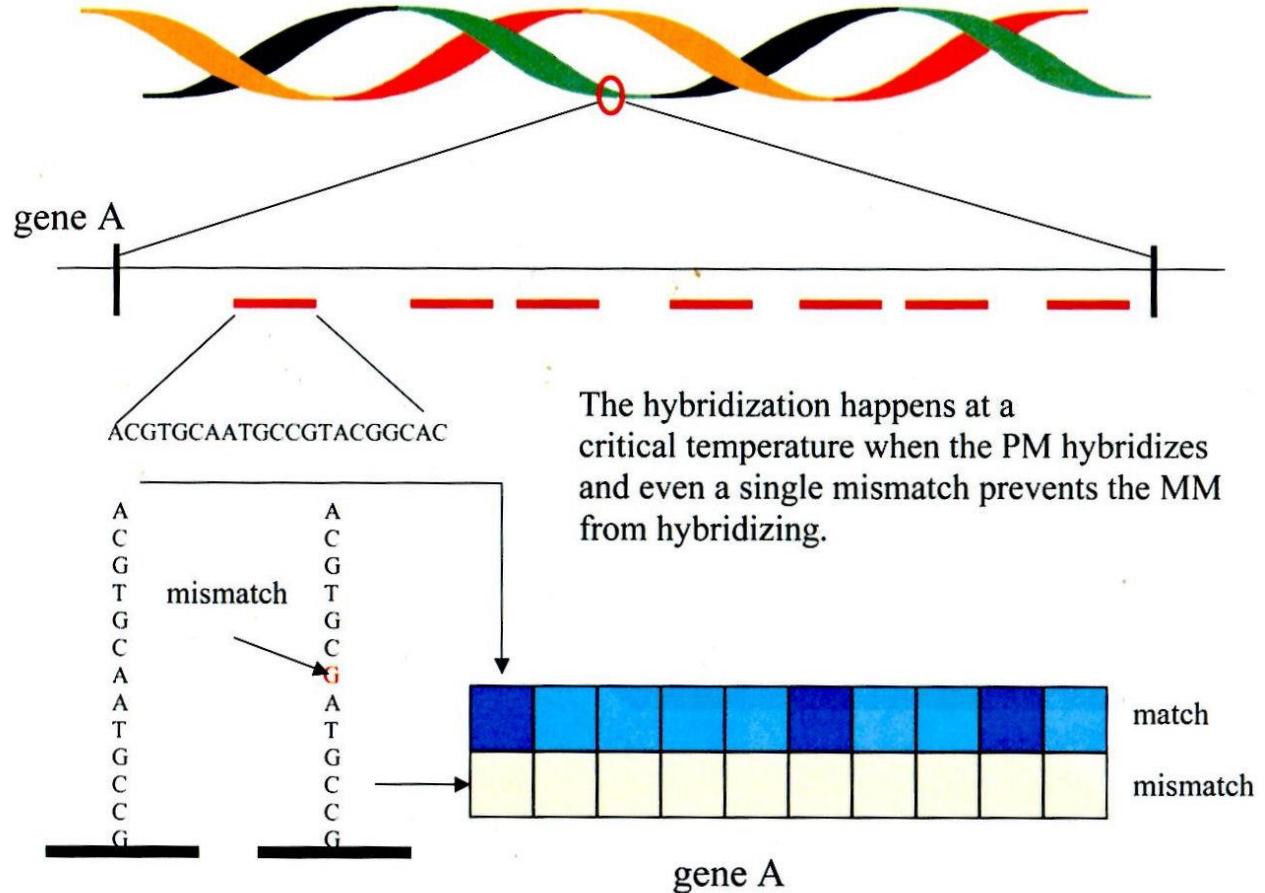


- Probe pair:

- PM = Perfect Match
- MM = Mismatch



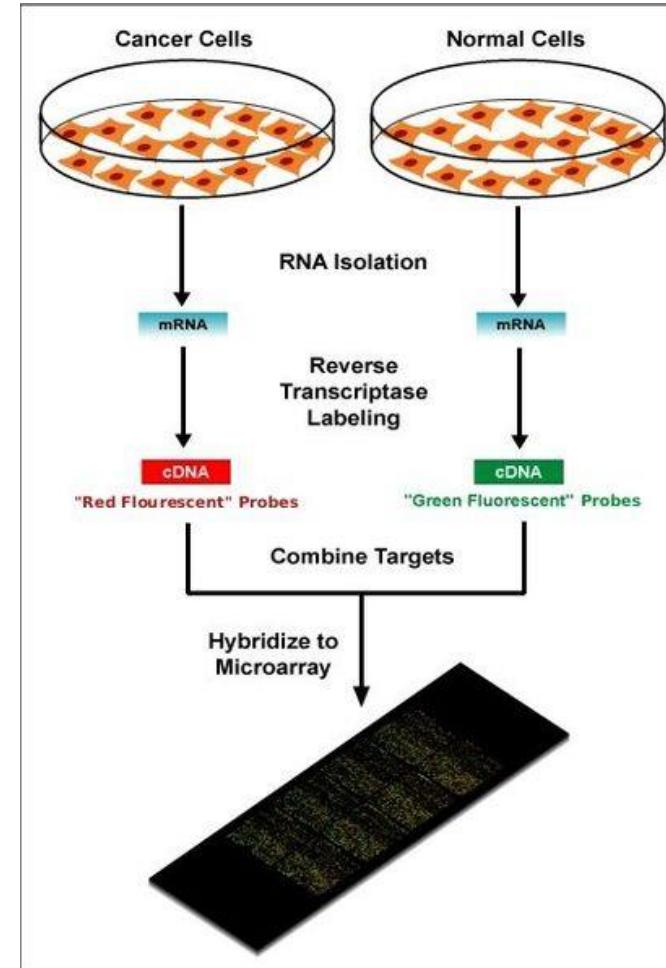
MICROARRAY'S: GEN EXPRESSIES (1-CHANNEL)



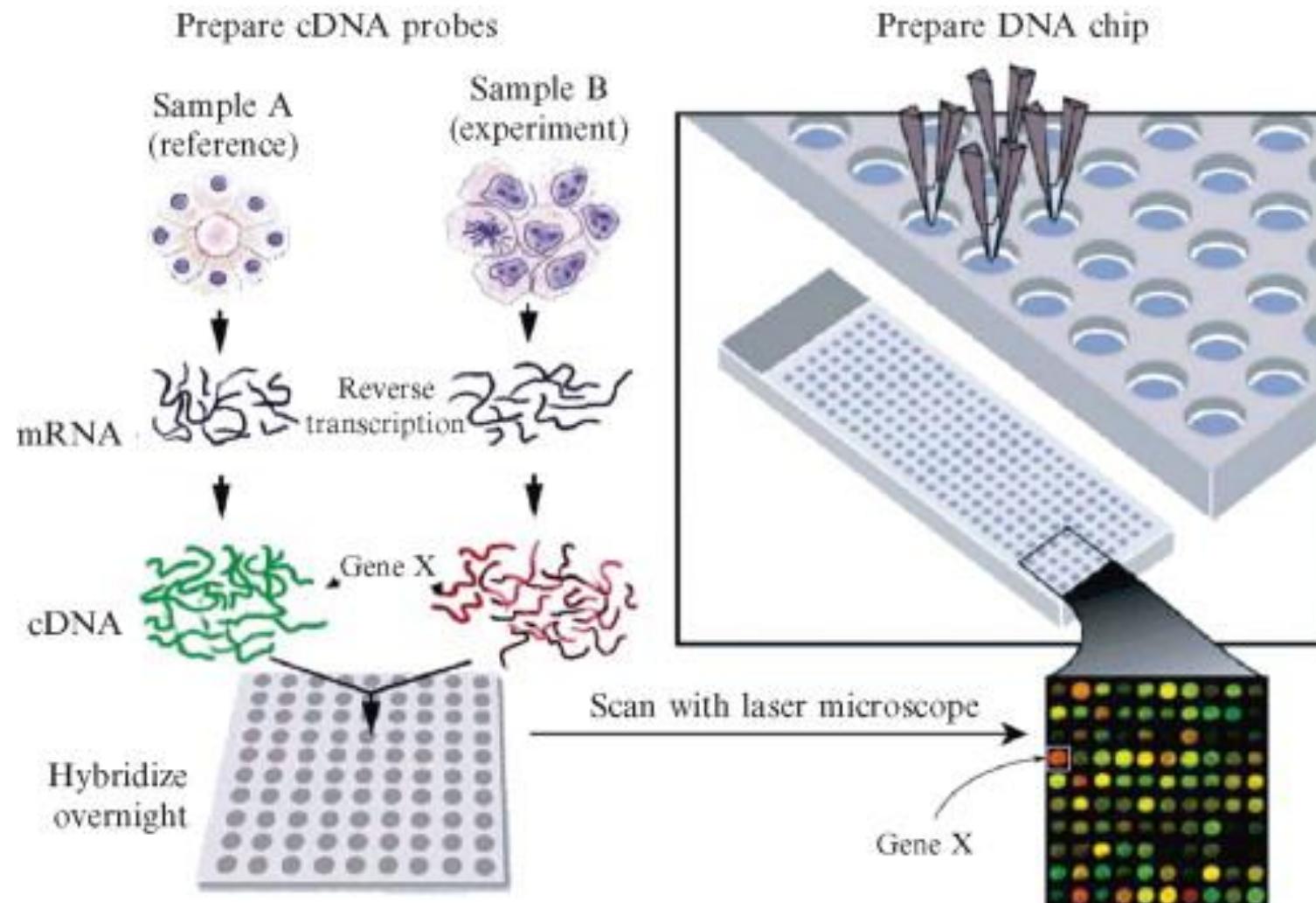
Draghici, 2012

MICROARRAY'S: GEN EXPRESSIES (2-CHANNEL)

- Voorbeelden:
 - Agilent: Dual-Mode platform
 - Eppendorf: DualChip platform for colorimetric Silverquant labeling
 - TeleChem International: Arrayit
- Twee fluorescente channels:
 - R(ed) Cy5 @ 670 nm
 - G(reen) Cy3 @ 570 nm
- Hybridisatie van 2 samples op zelfde spot in array



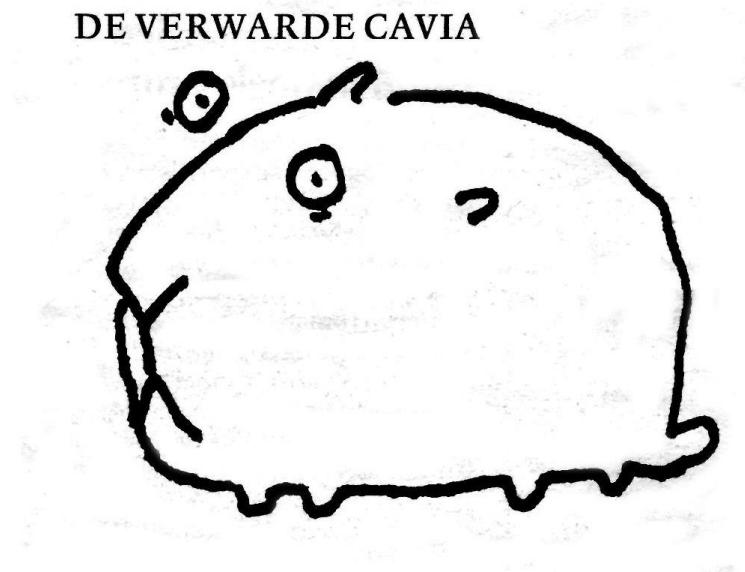
MICROARRAY'S: GEN EXPRESSIONS (2-CHANNELS)



Chang et al., 2006

DATA PRE-PROCESSING

- Verschillende data pre-processing stappen
- Niet altijd elke stap en niet altijd in zelfde volgorde
 - background correctie
 - log transformatie
 - normalisatie



DATA PRE-PROCESSING: BACKGROUND CORRECTIE

- Corrigeer de gemeten fluorescentie intensiteiten voor de background fluorescentie
- 1-channel methode (Affymetrix):
 - PM (perfect match) en MM (mismatch) spots opzelfde array voor elk gen
 - MM is background signaal: $E = PM - MM$
- 2-channel methode:
 - Bij uitlezen chip ook waarden van bg per spot per channel: $R' = R - R_{bg}$, $G' = G - G_{bg}$
- Soms wordt *niet* gecorrigeerd voor background!

DATA PRE-PROCESSING: EXPRESSIE RATIO

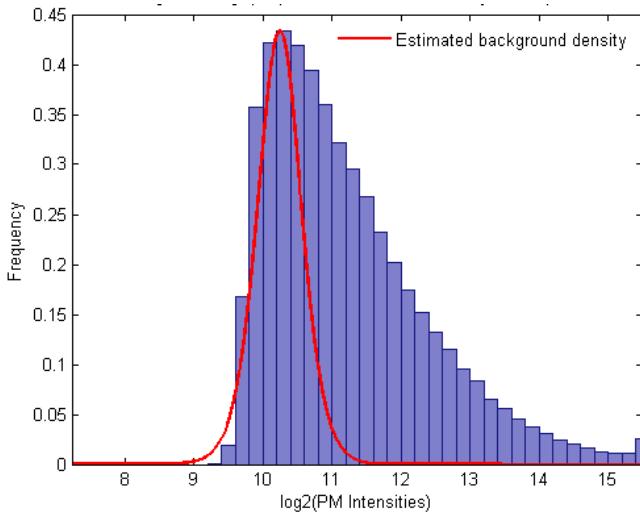
- Bereken de ratio van de gen expressie van het sample (R') en van de controle (G')

$$T' = \frac{R'}{G'}$$

In Engels ook wel de “fold change”

DATA PRE-PROCESSING: LOG TRANSFORMATIE

- Gen intensiteiten E hebben vaak **asymmetrische** verdeling:



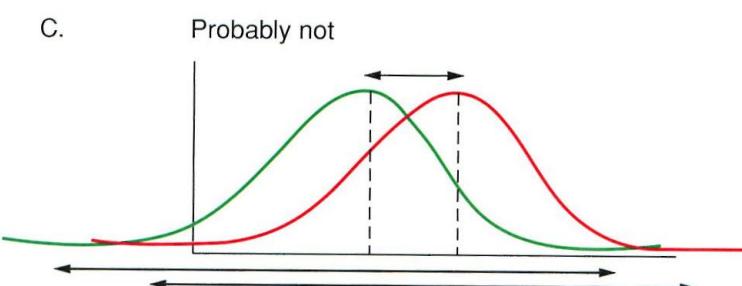
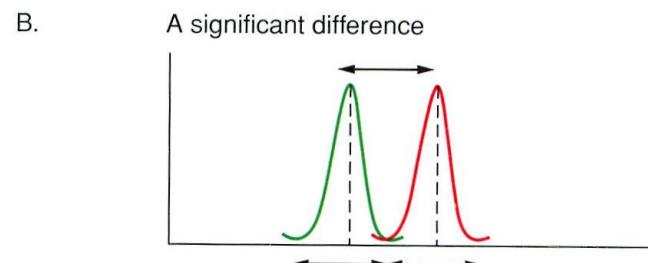
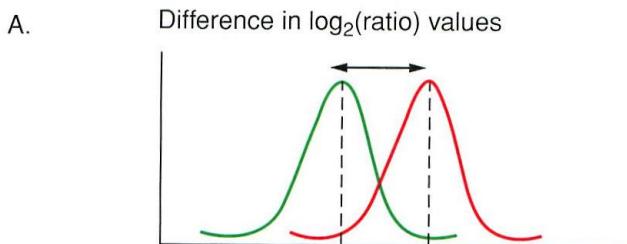
- Veel statistische analyses (t -toets, ANOVA) gaan uit van **normaal verdeelde** data
- Standaard truc in statistiek: $\log(E)$ is veel meer normaal verdeeld!
- Microarray data: meestal $^2\log$ of \log_2 gebruikt

R:

Voorbeeld 10.4 – Logtransformatie

DIFFERENTIALLY EXPRESSED GENES (DEGs)

- Genen met verschillende expressie tussen sample groep(en)
 - Bijv. gen X komt *meer tot expressie* bij gezonde mensen ten opzicht van mensen met kanker
 - Bijv. gen Y komt *minder tot expressie* bij vrouwen dan bij mannen



DEGs: VALKUILEN

- Transcriptie is geen translatie
- Translatie is geen garantie voor veel (functioneel) eiwit product

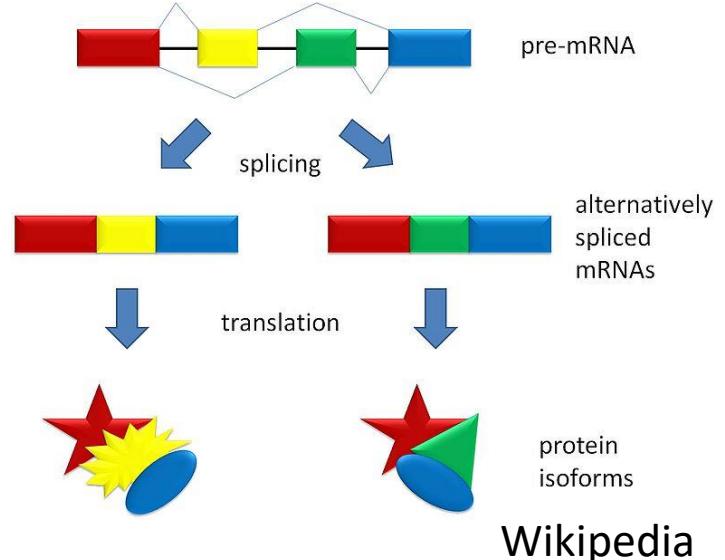
- Cross mapping

- Splice variants

- Weefsel / cel type niet relevant

- Confounders

- Biasses in microarray



- Interpretatie: een gen dat (significant) hoger tot expressie komt in kanker cellen hoeft niet per se de oorzaak te zijn van kanker

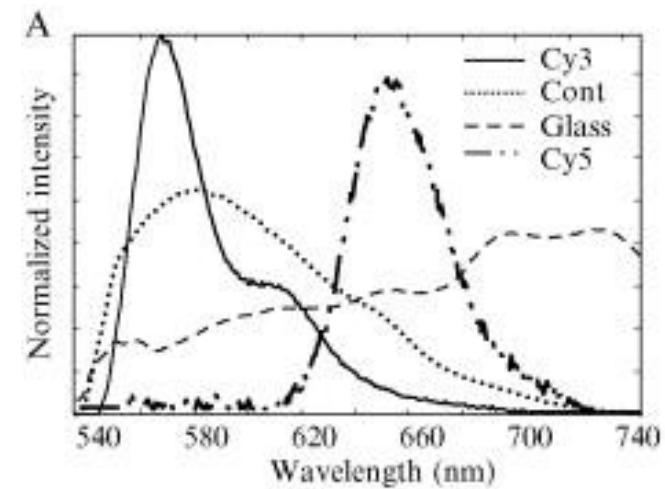
- Hoeft zelfs niet eens belangrijk te zijn

PROBLEMEN: BRONNEN VAN VARIATIE

- Verschil tussen array's
- Verschil in probe dichtheid tussen spots binnen array
- Delen array gemaakt door verschillende printer tips
- Background correctie
- Verschil hybridisatie tussen genen
- Effect van dye
- Interactie gen - dye
- Dag/tijdstip/analist/...

- Verschil tussen weefsel
- Verschil tussen individuen

- Differentieel verschil tussen genen



DATA PRE-PROCESSING: NORMALISATIE

- Corrigeer zoveel mogelijk voor array variatie, labeling variatie etc.: dus **technische variatie**
- Verschillende mogelijkheden:
 - Normalization to a reference RNA
 - Mean or median normalization
 - Scaling normalization
 - Lowess (Loess) normalization
 - Print-tip normalization

NORMALISATIE

- Dual channel ($R = Cy5$, $G = Cy3$) array
- Voer per channel background correcties uit per gen (spot) i

$$R_i \equiv R_{i,\text{raw}} - R_{i,\text{bg}} \quad G_i \equiv G_{i,\text{raw}} - G_{i,\text{bg}}$$

- Bereken per microarray per gen (spot) i de log expressie ratio (“log fold change”)

ook wel “ R ”

$$M_i \equiv {}^2\log\left(\frac{R_i}{G_i}\right) = {}^2\log(T_i) = [{}^2\log(R_i) - {}^2\log(G_i)]$$

- Bereken per microarray over alle genen (spots) i de log gemiddelde expressie (“log intensity”)

ook wel “ I ”

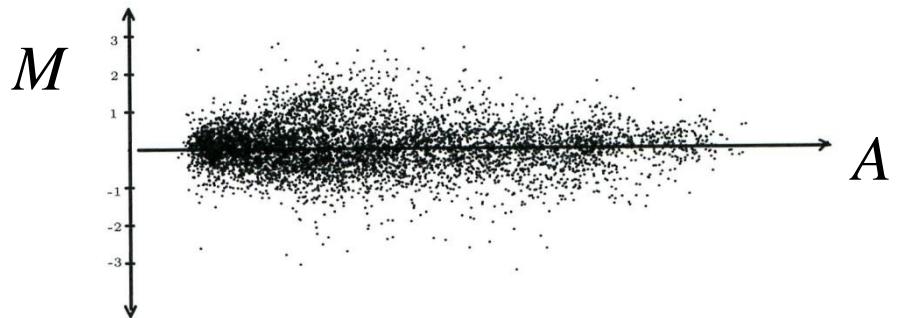
$$A_i \equiv {}^2\log\left(\sqrt{R_i \cdot G_i}\right) = \frac{1}{2} [{}^2\log(R_i) + {}^2\log(G_i)]$$

geometrisch gemiddelde

NORMALISATIE

- Verwachting:

- M vs A is een puntenwolk rondom $M = 0$, en ongeveer symmetrisch in A



- Zo niet: dan effect van (technische) bias

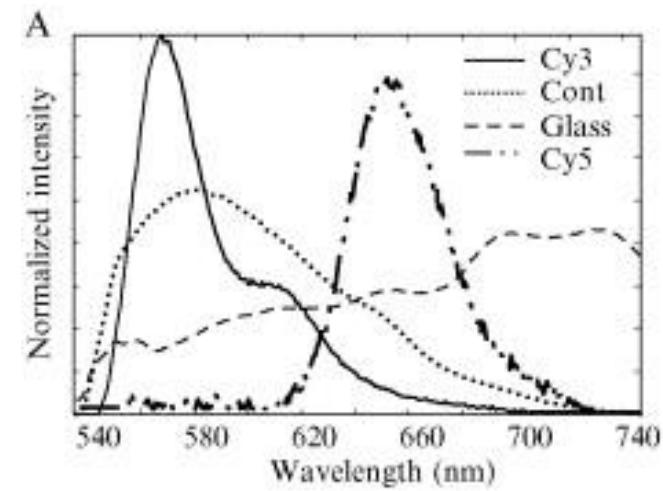
- ongelijke hoeveelheden mRNA sample
- verschil in labeling efficiëntie dyes
- verschil in detectie efficiëntie dyes
- ...

PROBLEMEN: BRONNEN VAN VARIATIE

- Verschil tussen array's
- Verschil in probe dichtheid tussen spots binnen array
- Delen array gemaakt door verschillende printer tips
- Background correctie
- Verschil hybridisatie tussen genen
- Effect van dye
- Interactie gen - dye
- Dag/tijdstip/analist/...

- Verschil tussen weefsel
- Verschil tussen individuen

- Differentieel verschil tussen genen



NORMALISATIE

- Per microarray meestal per gen (spot) i

- mean normalisatie

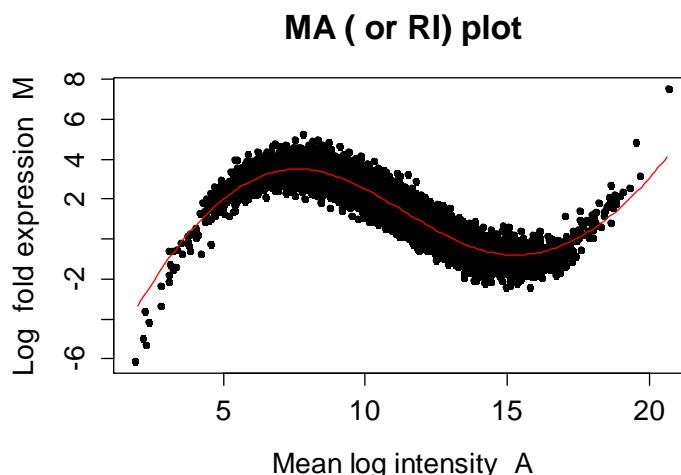
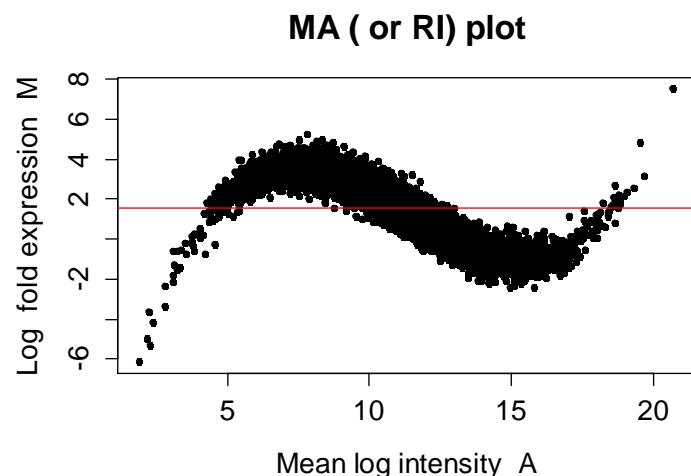
$$M'_i = M_i - \bar{M}$$

gemiddeld over hele array

- lowess (loess) normalisatie

$$M'_i = M_i - \text{lowess}(A_i)$$

locale fit door puntenwolk

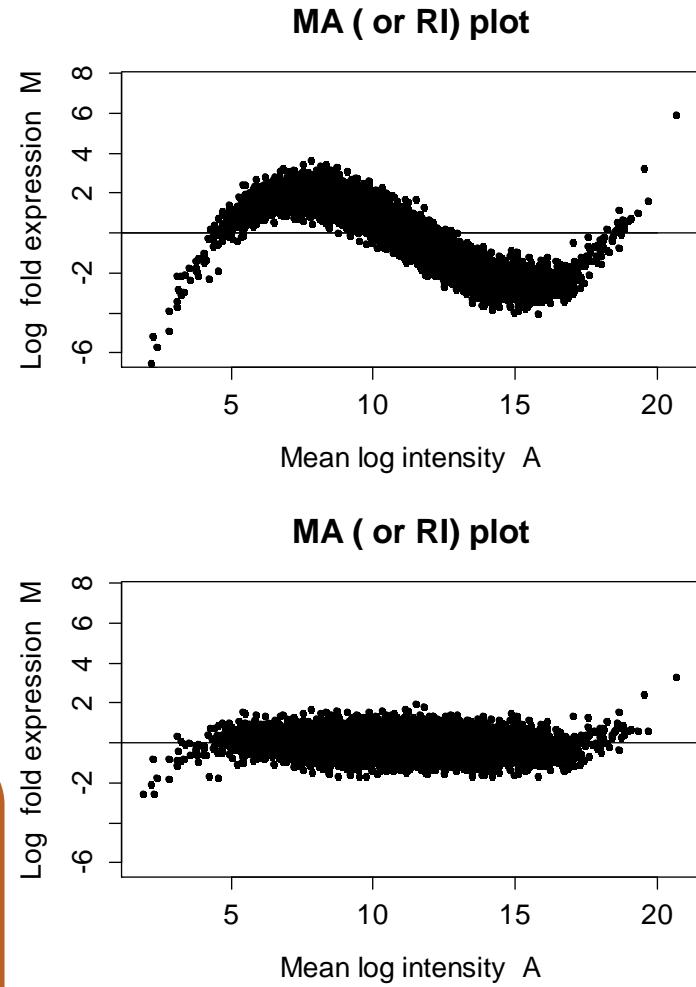


NORMALISATIE

- Resultaat:
 - mean normalisatie

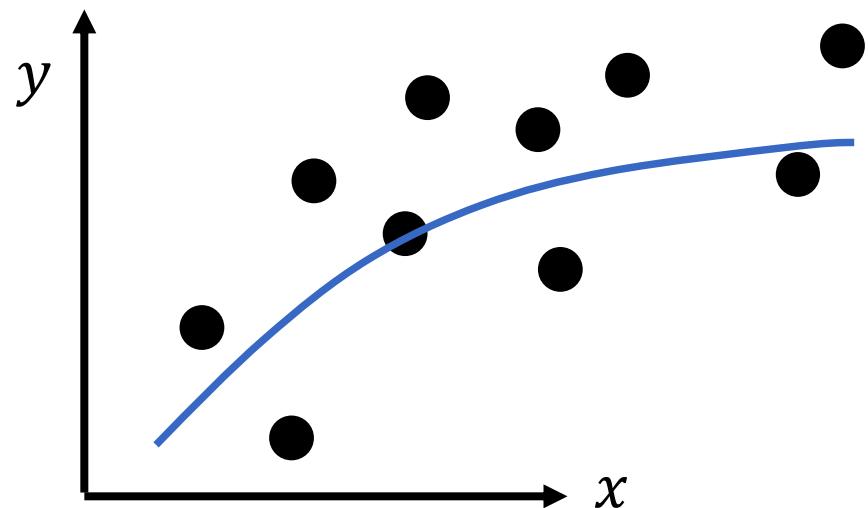
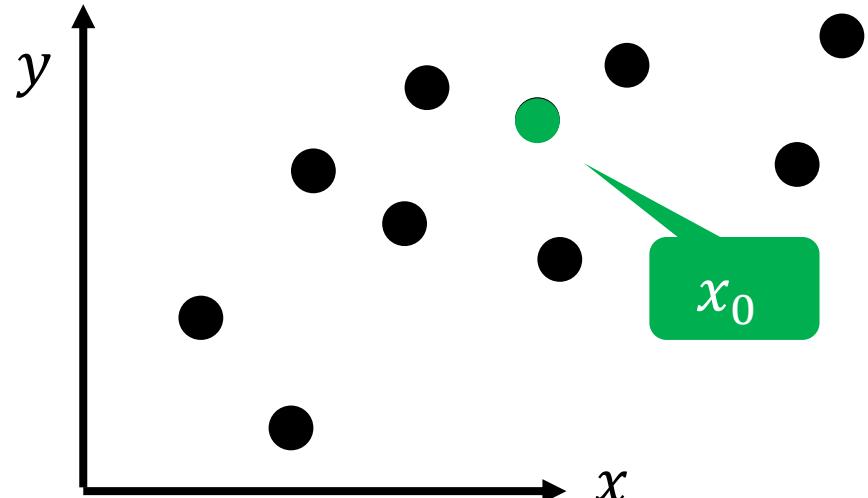
- lowess (loess) normalisatie

Je update dus ALLEEN de $M = \log(R/G)$ waarden, NIET de intensiteiten per channel!



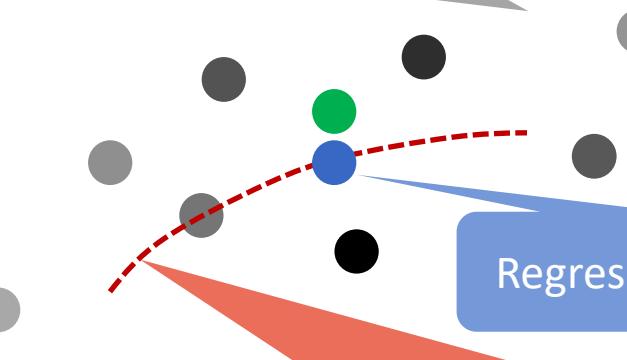
Lo(w)ESS REGRESSIE

Local Polynomial Regression:



Gewicht van elk punt x bij locale regressie afhankelijk van afstand tot x_0 :

$$w(x) = \begin{cases} \left(1 - \frac{|x - x_0|^3}{h}\right) & |x - x_0| < h \\ 0 & |x - x_0| \geq h \end{cases}$$



voor alle x_0

$$y = a_0 + a_1 \cdot (x - x_0) + a_2 \cdot (x - x_0)^2$$

REGRESSIE: LO(w)ESS

Model: lowess (locale regressie)

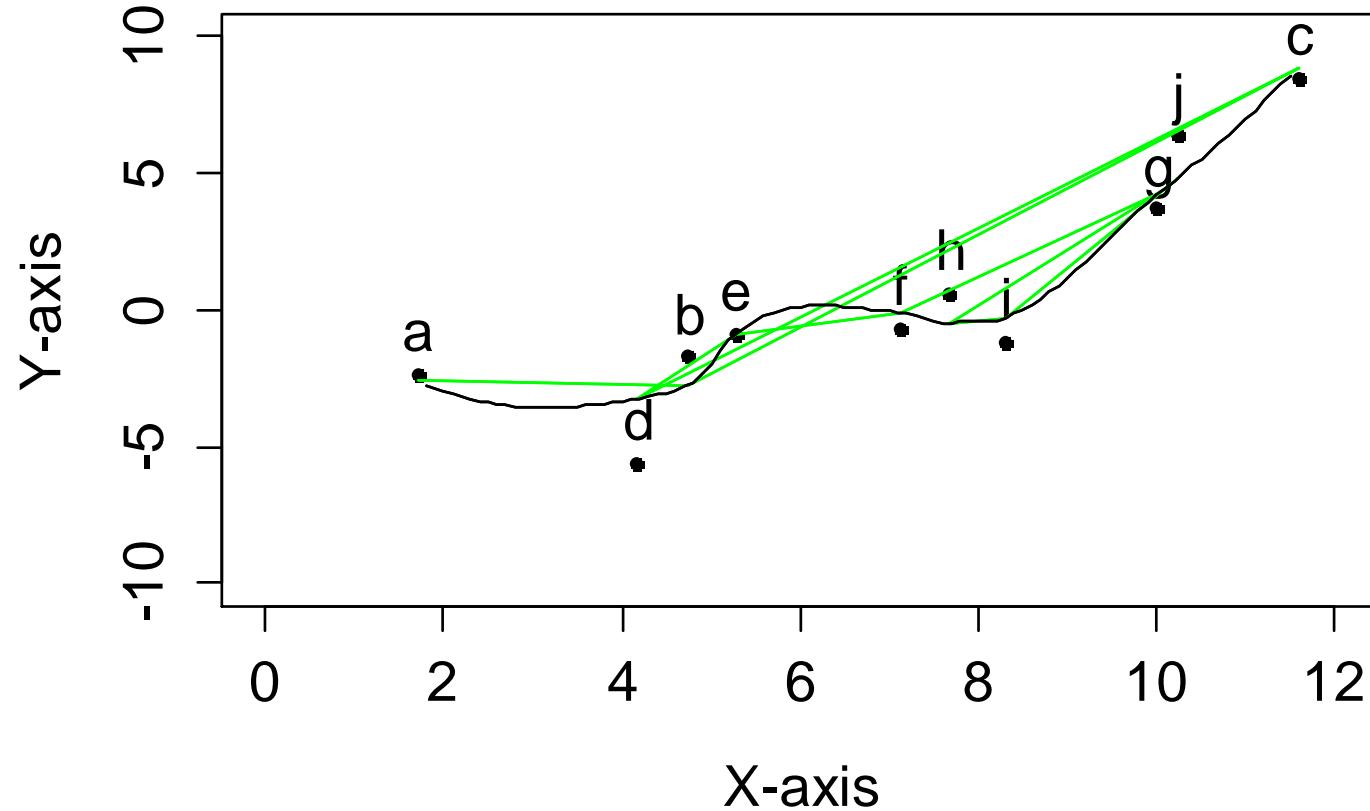
```
# fit and plot a loess function  
fit.loess <- loess(y ~ x)  
  
y.loess.1 <- fit.loess$fitted  
y.loess.1 <- predict(fit.loess)  
lines(x, y.loess.1, col="green")  
lines(y.loess.1 ~ x, col="green")  
  
xplot <- seq(0,12,0.1)  
y.loess.2 <- predict(fit.loess, newdata=data.frame(x=xplot))  
lines(xplot, y.loess.2, col="black")  
lines(y.loess.2 ~ xplot, col="black")  
  
# in 1 stap data en loess curve plotten:  
scatter.smooth(x, y, xlab="X-axis", ylab="Y-axis")  
scatter.smooth(y ~ x, xlab="X-axis", ylab="Y-axis")
```

Belangrijk: omdat in de (loess) fit "x" de verklarende variabele is, moet je de variabele bij newdata OOK "x" noemen!

REGRESSIE: LO(w)ESS

- Resultaat:

Test of plot options



Dus:

```
plot(M ~ A)
```

```
fit <- lowess(M ~ A)
```

```
# Loess correctie:
```

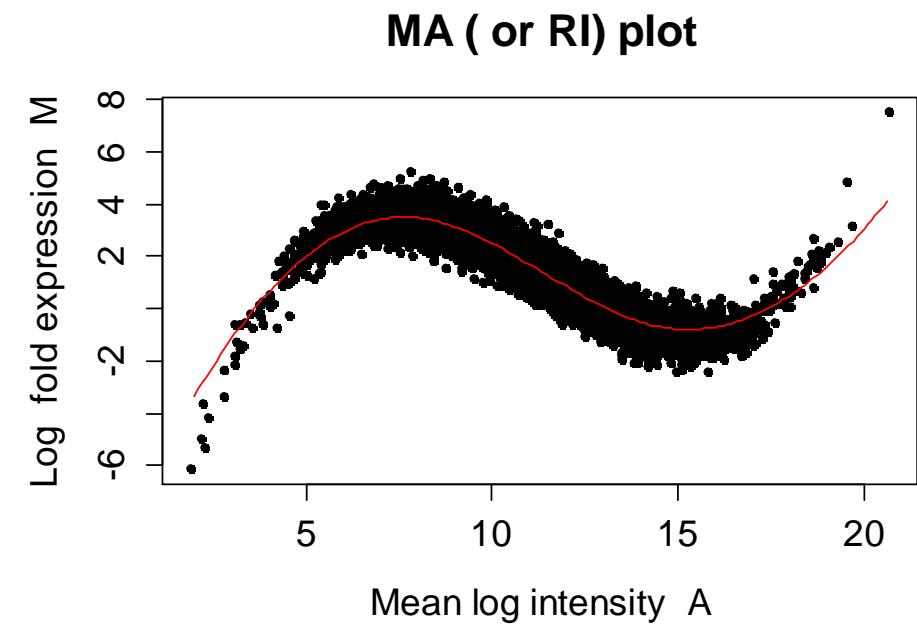
```
M.corr <- M - predict(fit)
```

```
# Grafisch weergeven in MA plot:
```

```
xplot <- seq(0, 12, 0.1)
```

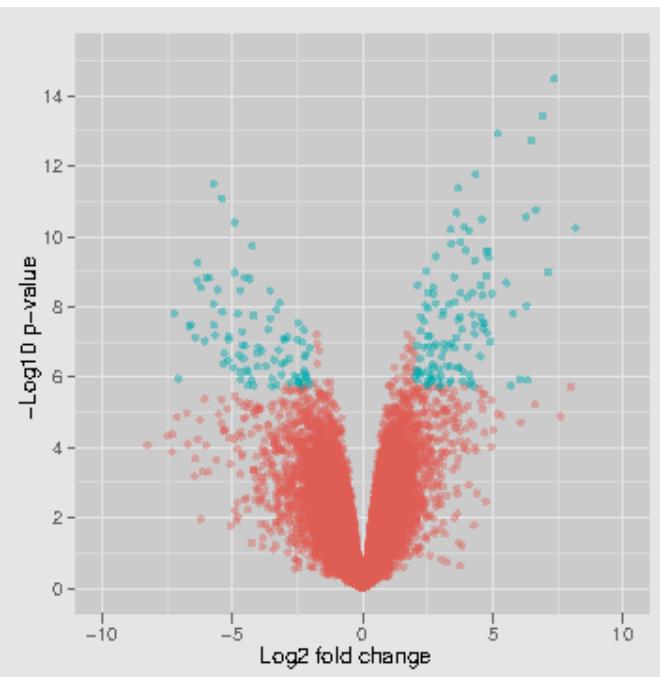
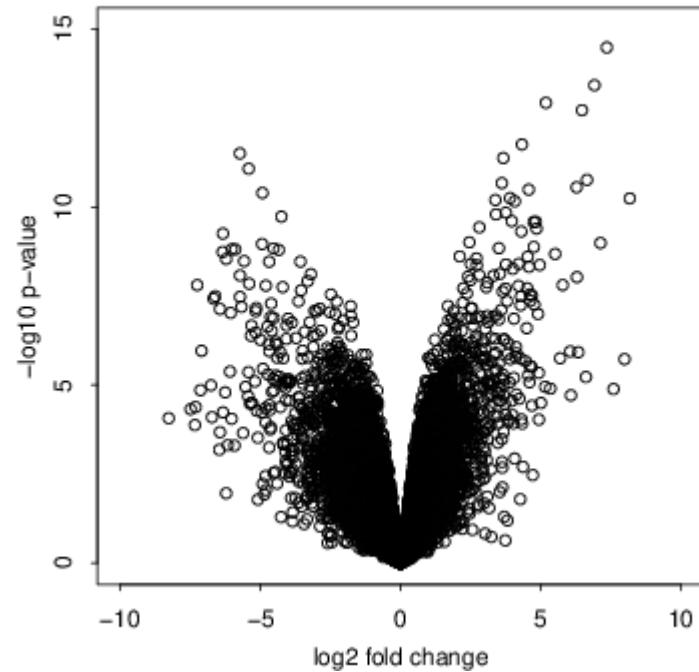
```
yplot <- predict(fit, newdata = data.frame(A=xplot))
```

```
lines(yplot ~ xplot, col="red")
```



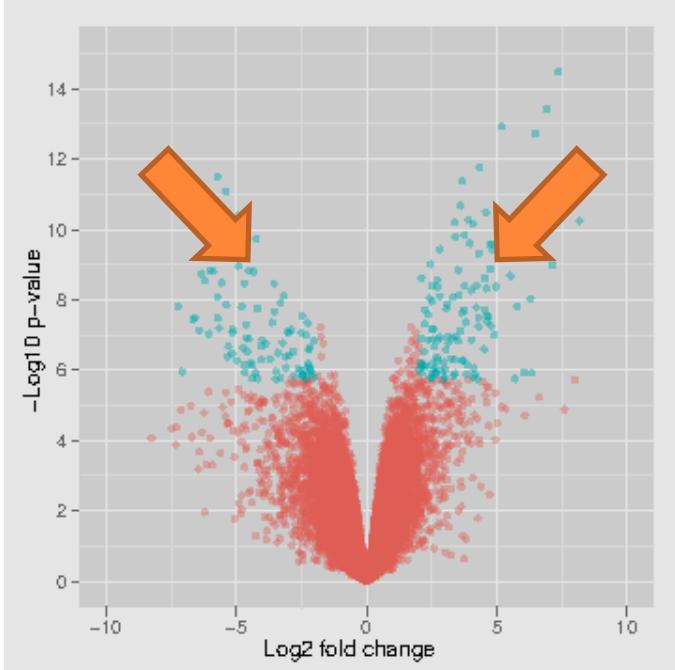
VOLCANO PLOT

- Wat zijn “biologisch relevante” DEG’s?
 - Kleine p -waarde
 - Grote (absolute) log fold change $|M|$



VOLCANO PLOT

- Plot $-10\log(p\text{-waarden})$, evt. met multiple toets correctie, als functie van M -waarden.
- “Biologisch interessante” genen:
 - bijv. $|M| > 2$
 - bijv. $-10\log(p_{adj}) > -10\log(0.05)$



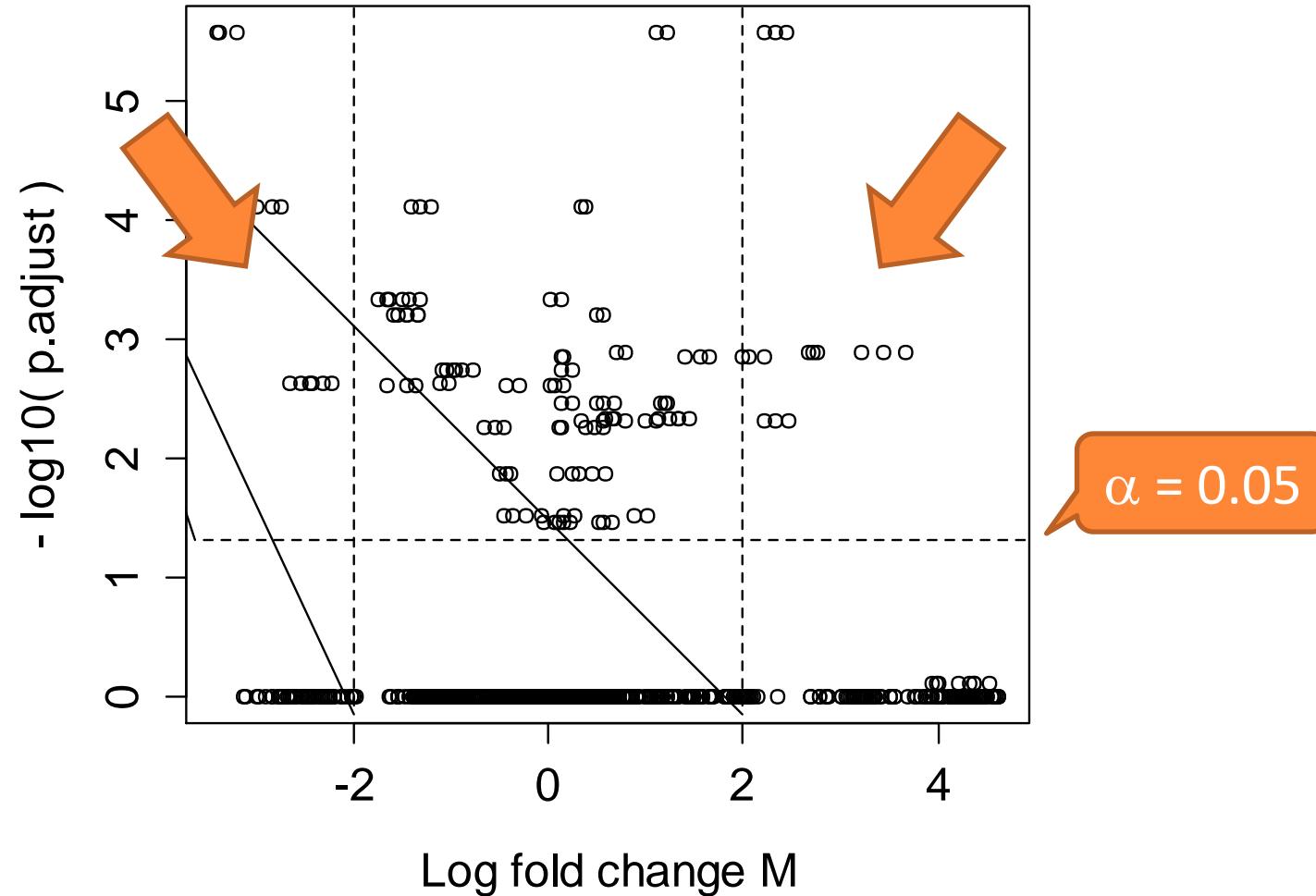
VOLCANO PLOT

- Snelle manier om volcano-plot te maken uit dataframe **M** (met enkel M -waarden, per regel één gen), en een vector **pVec** en/of **pVec.adjust** met (aangepaste) p -waarden uit t -toets/ANOVA/... per gen:

```
nGenes <- nrow(M)
nSamples <- ncol(M)
pVec.ALL <- rep(pVec, nSamples)
pVec.adjust.ALL <- rep(pVec.adjust, nSamples)
logFold.ALL <- as.vector(as.matrix(M))
plot(-log10(pVec.adjust.ALL) ~ logFold.ALL,
      xlab="Log fold change M",
      ylab="- log10( p.adjust )")
abline(h=-log10(0.05), lty=2)
abline(v=-2, lty=2)
abline(v=2, lty=2)
```

VOLCANO PLOT

- Resultaat:



VOLCANO PLOT

- Gewone punten zwart, maar **rood** als $p_{\text{adj}} < 0.05$ en $|M| > 2$:

```
# make colors red for special points
colors.ALL <- 1 + (pVec.adjust.ALL < 0.05 & abs(logFold.ALL) > 2)

plot(-log10(pVec.adjust.ALL) ~ logFold.ALL,
      xlab="Log fold change M",
      ylab="- log10( p.adjust )",
      col=colors.ALL)
abline(h=-log10(0.05), lty=2)
abline(v=-2, lty=2)
abline(v=2, lty=2)
```

