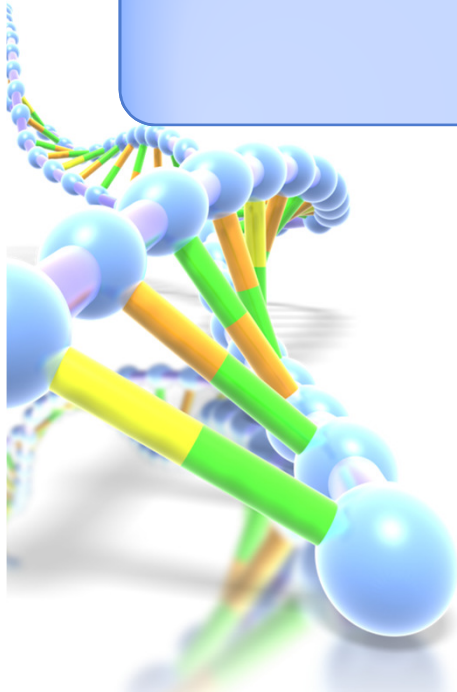


## Les 09 – Microarray's en differentiële gen expressie (2)

**Emile Apol**

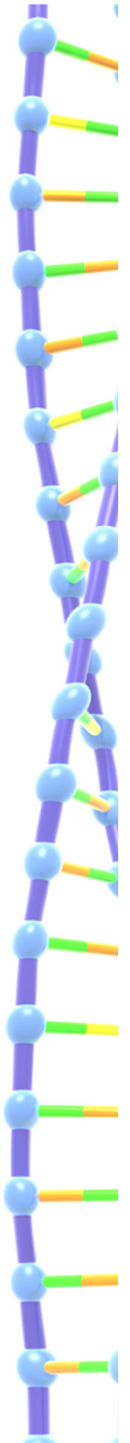


**Hanze University Groningen**  
APPLIED SCIENCES

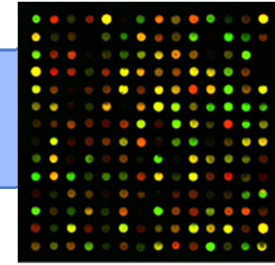
*Institute for  
Life Science & Technology*

## LES 09

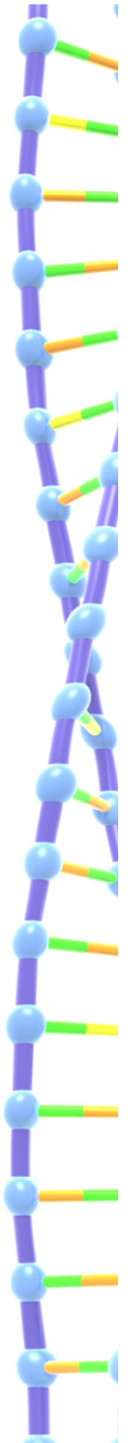
- Doe meer met DEG's:
  - Volcano plot
  - “Effect size ( $\eta^2$ ) plot”
- Algemene aanpak MA analyse: ANOVA
  - Meerdere factoren
  - Confounding
  - Experimentele opzet (DOE)



## MICROARRAY ANALYSE: STAPPENPLAN

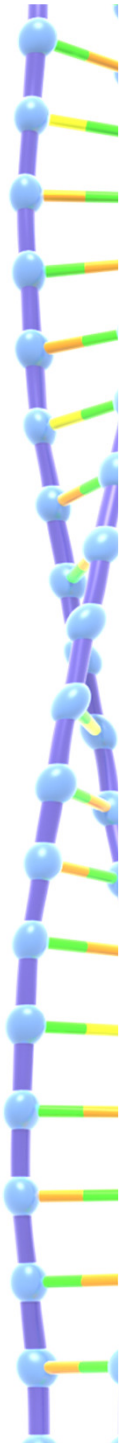
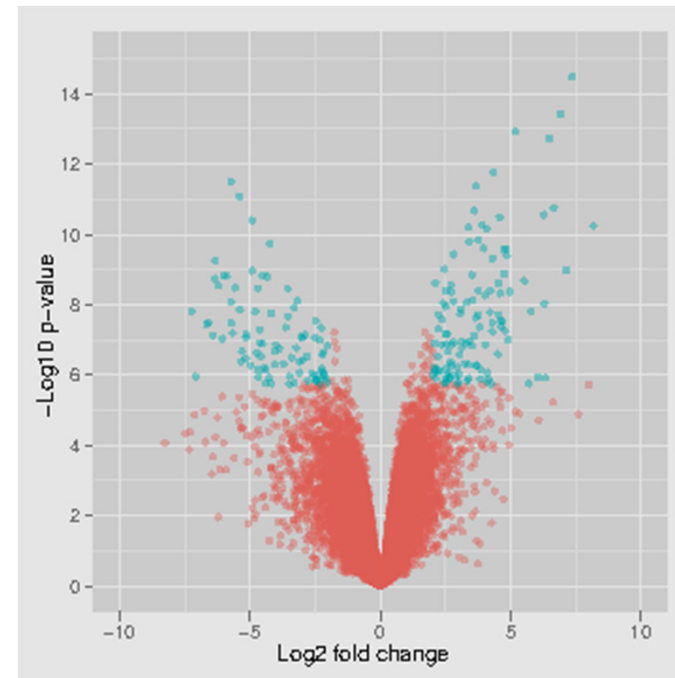
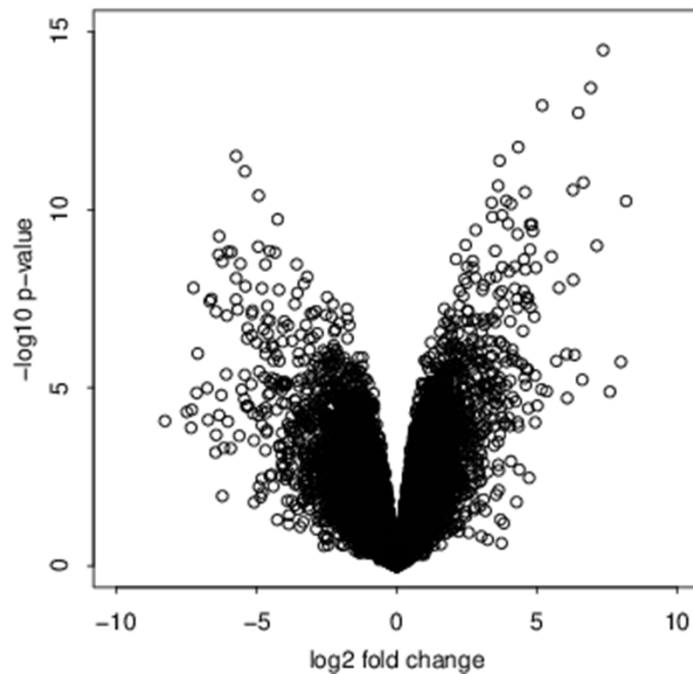
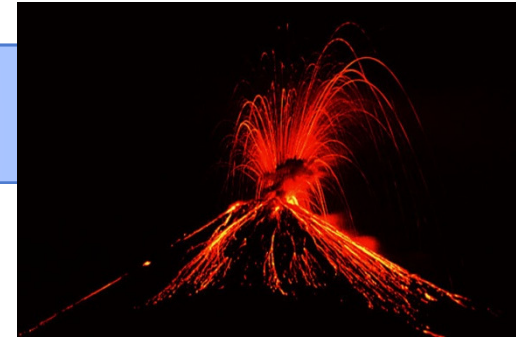


- Background correctie
- Log transformatie
- Normalisatie (bijv. loess)
- Toetsen op DEG's:
  - $t$ -toets, 1-way ANOVA, ...
  - Wilcoxon's toets, Kruskal-Wallis toets, ...
- Aanpassen  $p$ -waarden voor multiple toetsing
- Clustering van DEG's:
  - Hiërarchisch clusteren
  - $k$ -means
  - Principale Componenten Analyse (PCA)
- Toetsen op functionaliteit genen binnen clusters



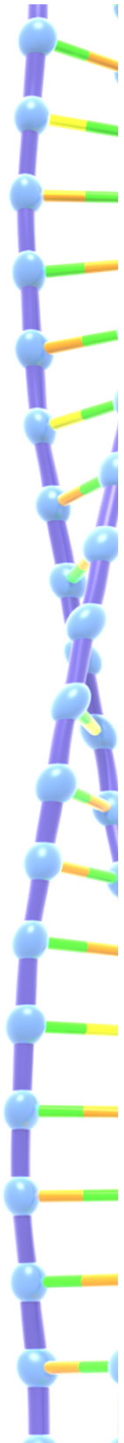
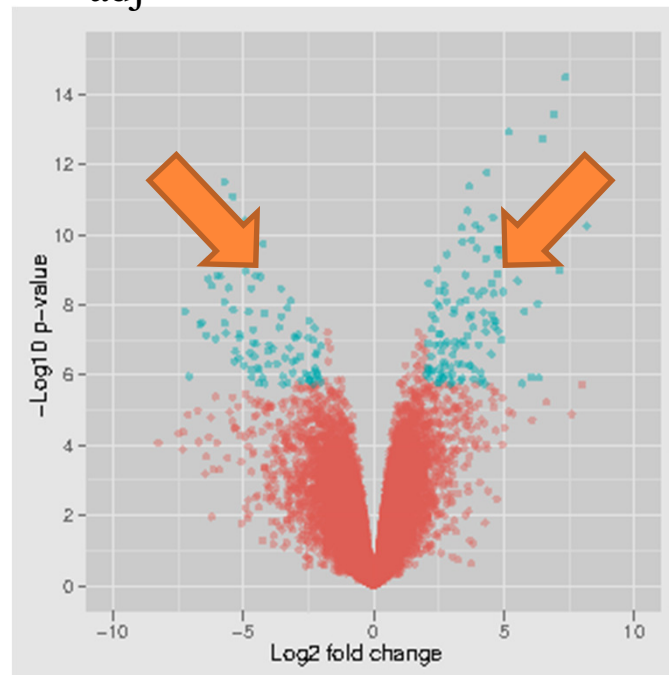
# VOLCANO PLOT

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



## VOLCANO PLOT

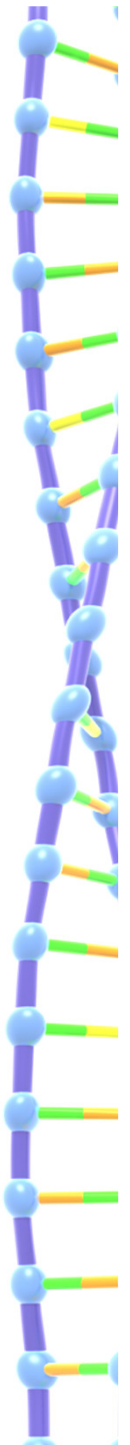
- Plot –  $^{-10}\log(p\text{-waarden})$ , evt. met multiple toets correctie, als functie van  $M$ -waarden.
- “Biologisch interessante” genen:
  - bijv.  $|M| > 2$
  - bijv.  $^{-10}\log(p_{\text{adj}}) > -^{-10}\log(0.05)$



## VOLCANO PLOT

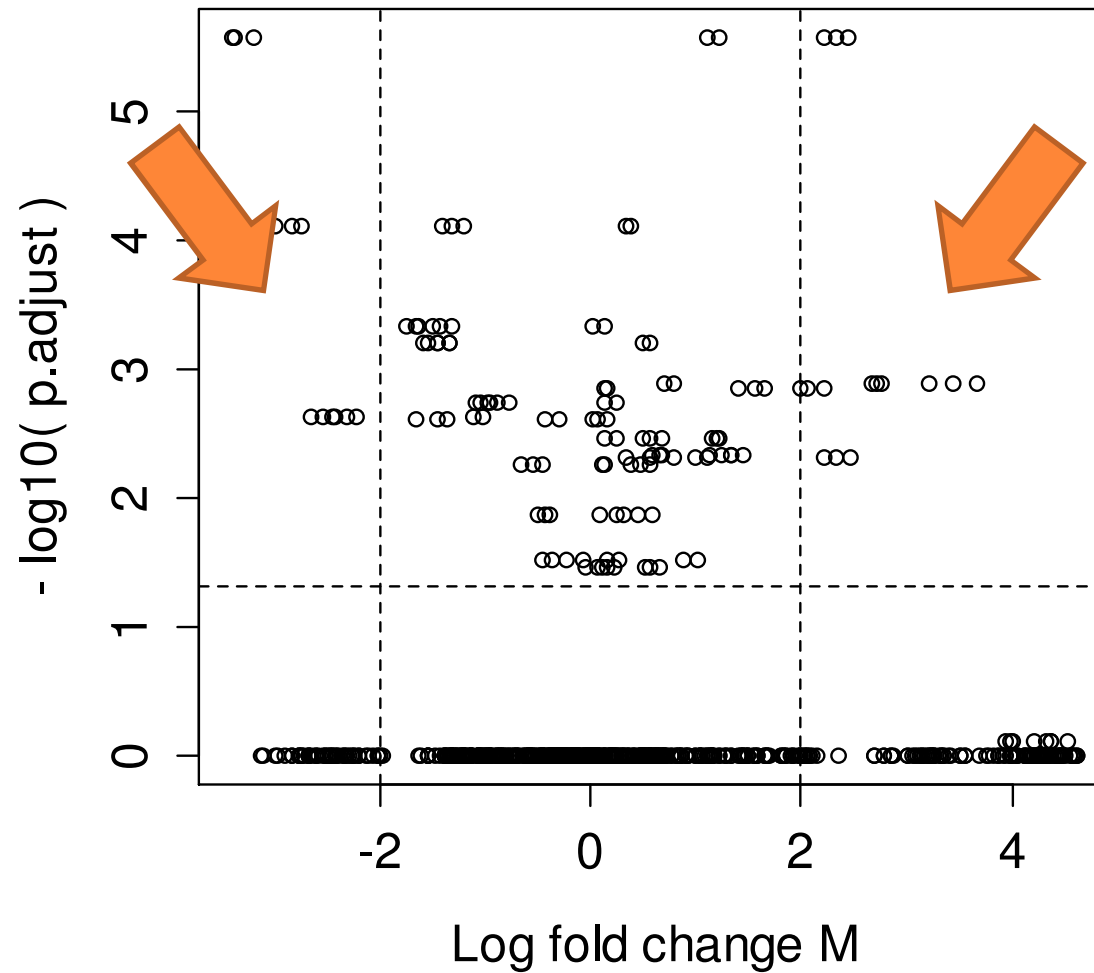
- Snelle manier om volcano-plot te maken uit dataframe **M** (met enkel *M*-waarden, per regel één gen), en een vector **pVec** en/of **pVec.adjust** met (aangepaste) *p*-waarden uit *t*-toets/ANOVA/... per gen:

```
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)
```

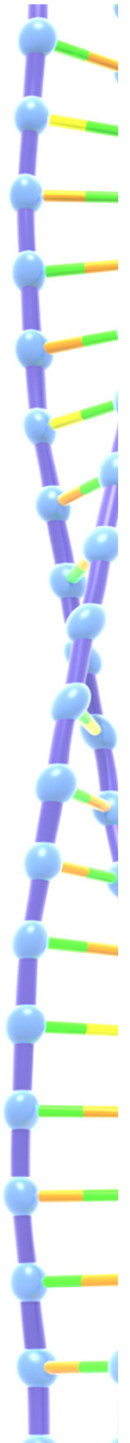


# VOLCANO PLOT

○ Resultaat:



$\alpha = 0.05$

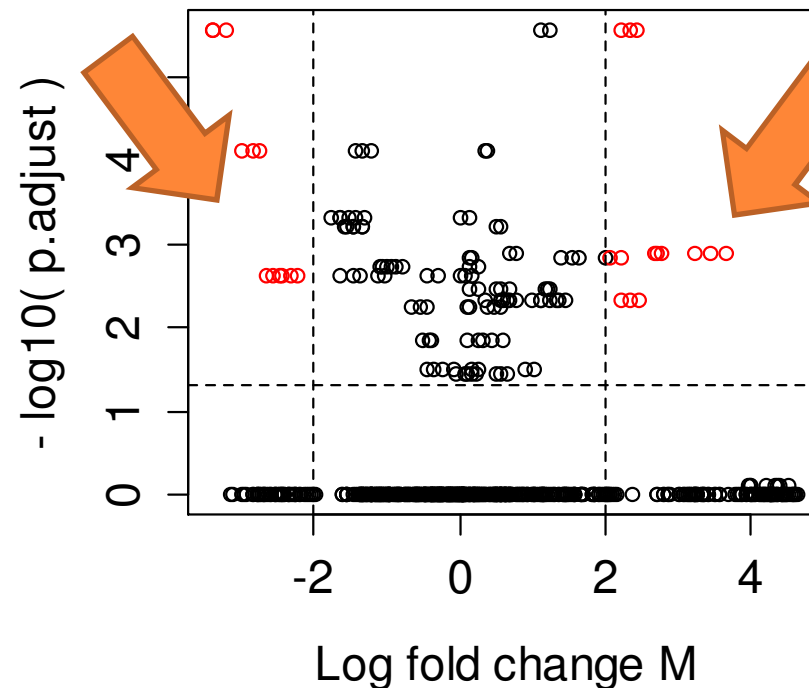


# VOLCANO PLOT

- 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)
```



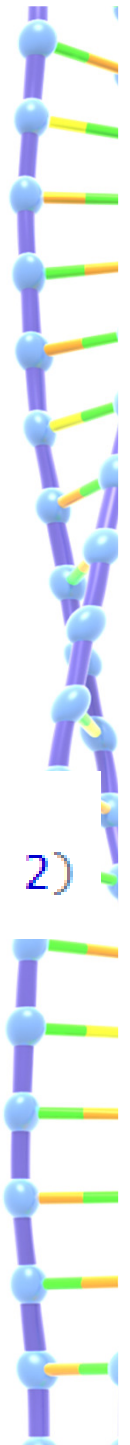


## KLEUREN IN R

- Verschillende manieren om in R **kleur** aan te geven:
  - `col = c("black", "red")`
  - `col = c(1, 2)`
  - `col = rainbow.colors(12)`
- Logicals: T = TRUE = 1, F = FALSE = 0, dus

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

levert normaal 1 (= “**black**”) maar als  $p_{\text{adj}} < 0.05$  en  $|M| > 2$ , dus  $1 + (T \& T) = 1 + T = 2$  (= “**red**”); dit maakt een vector van 1’en en 2’en...



## EFFECT STERKTE: $\eta^2$

- $p$ -waarde geeft **statistische significantie** (> toeval?)
- $|M|$ -waarde geeft “biologische” significantie: effect sterkte, aribraire grens  $|M| > 2$
- Bij  $t$ -toetsen en 1-way ANOVA ook *statistische* definitie van **effect sterkte**:  $\eta^2$  (eta kwadraat) = praktische significantie

$$\eta^2 \equiv \frac{SS_{\text{tussen}}}{SS_{\text{tot}}} = \frac{SS_{\text{tussen}}}{SS_{\text{tot}} + SS_{\text{binnen}}}$$

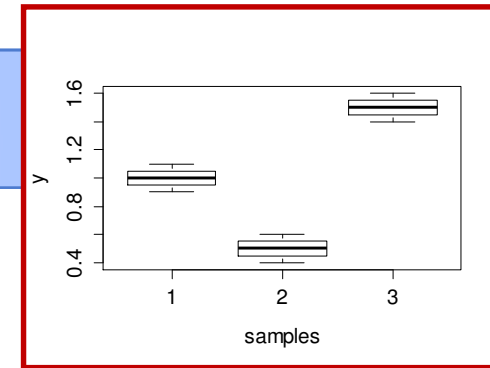
1-way ANOVA

$$\eta^2 \equiv \frac{SS_{\text{tussen}}}{SS_{\text{tot}}} = \frac{t^2}{t^2 + \text{df}}$$

$t$ -toets

## EFFECT STERKTE: $\eta^2$

- Voorbeeld: 1-way ANOVA



```
> summary(aov(y ~ samples))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
samples	2	1.50	0.75	75	5.69e-05
Residuals	6	0.06	0.01		

$$\eta^2 = \frac{SS_{\text{tussen}}}{SS_{\text{tot}} + SS_{\text{binnen}}} = \frac{1.50}{1.50 + 0.06} = 0.96$$

Dit betekent dat 96% van alle variatie in de data komt door verschillen tussen de groepen (de andere 4% is “ruis”)

## EFFECT STERKTE: $\eta^2$

### ○ Voorbeeld: 2-sample *t*-toets

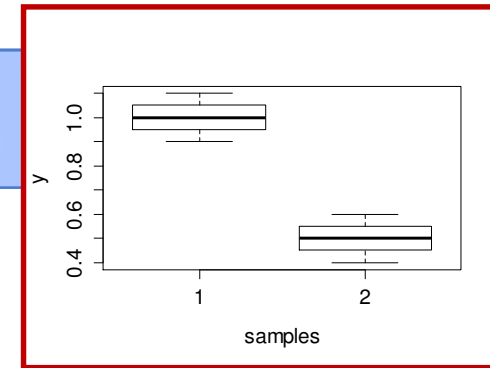
```
> t.test(y ~ samples, var.equal=T)
```

Two Sample t-test

data: y by samples

$t = 6.1237$ ,  $df = 4$ ,  $p\text{-value} = 0.003602$

alternative hypothesis: true difference in means is not equal to 0



$$\eta^2 = \frac{t^2}{t^2 + df} = \frac{6.1237^2}{6.1237^2 + 4} = 0.90$$

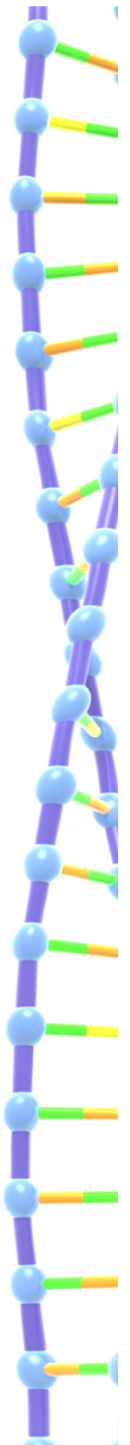
Dit betekent dat 90% van alle variatie in de data komt door verschillen tussen de groepen (de andere 10% is “ruis”)

## EFFECT STERKTE: $\eta^2$

- Wanneer is een effect (verschil tussen biologische samples) groot?
- Vuistregel (Cohen, 1988):

$\eta^2$ waarde	Effect sterkte
0.01	zwak effect
0.10	matig effect
> 0.25	sterk effect
1	perfecte relatie

NB. Vaak gaan statistische significantie ( $p$ -waarde < 0.05) en praktische significantie (effect sterkte) samen, soms niet!



## EFFECT STERKTE: $\eta^2$

- Voorbeeld van R functie om  $p$ -waarde en  $\eta^2$  per gen te berekenen: *t*-toets (2-sample, Welch, gepaard)
- Input:
  - $M$ -waarden per gen ( $\mathbf{x}$  = rij matrix)
  - Vector  $\mathbf{g}$  (factor) met groepslevels

```
matrixTTest <- function(x, g, ...){
```

```
  g <- as.factor(g)
```

```
  Q <- t.test(x ~ g, ...)
```

```
  p.value <- Q$p.value
```

```
  eta2 <- Q$statistic^2/(Q$statistic^2 + Q$parameter)
```

```
  a <- c()
```

```
  a[1] <- p.value
```

```
  a[2] <- eta2
```

```
  names(a) <- c("p-value", "eta2")
```

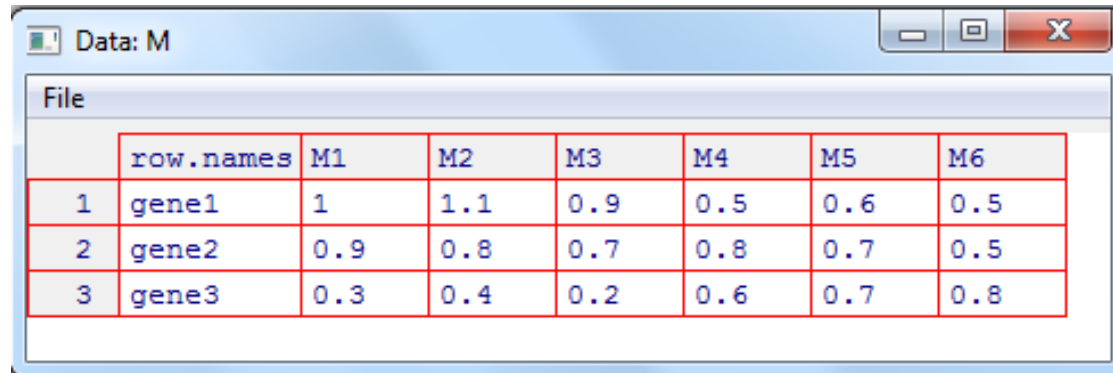
```
  return(a)
```

```
}
```

anonieme argumenten

## EFFECT STERKTE: $\eta^2$

### ○ Data:



	row.names	M1	M2	M3	M4	M5	M6
1	gene1	1	1.1	0.9	0.5	0.6	0.5
2	gene2	0.9	0.8	0.7	0.8	0.7	0.5
3	gene3	0.3	0.4	0.2	0.6	0.7	0.8

### ○ Resultaat:

```
samples <- as.factor(c(1,1,1,2,2,2))
```

```
> apply(M, 1, matrixTTest, g=samples, var.equal=T)
```

```
          gene1      gene2      gene3  
p-value 0.00219213 0.2745766 0.008049893  
eta2    0.92452830 0.2857143 0.857142857
```

anonieme argumenten

```
> apply(M, 1, matrixTTest, g=samples, var.equal=F)
```

```
          gene1      gene2      gene3  
p-value 0.004833894 0.2846272 0.008049893  
eta2    0.938697318 0.3169399 0.857142857
```

anonieme argumenten

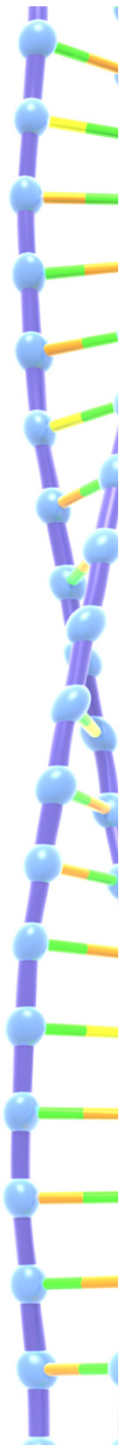
```
> apply(M, 1, matrixTTest, g=samples, paired=T)
```

```
          gene1      gene2      gene3  
p-value 0.005063324 0.05719096 0.05719096  
eta2    0.989898990 0.888888889 0.888888889
```

## EFFECT STERKTE: $\eta^2$

- Voorbeeld van R functie om  $p$ -waarde en  $\eta^2$  per gen te berekenen: **1-way ANOVA**
- Input:
  - $M$ -waarden per gen ( $\mathbf{x}$  = rij matrix)
  - Vector  $\mathbf{g}$  (factor) met groepslevels

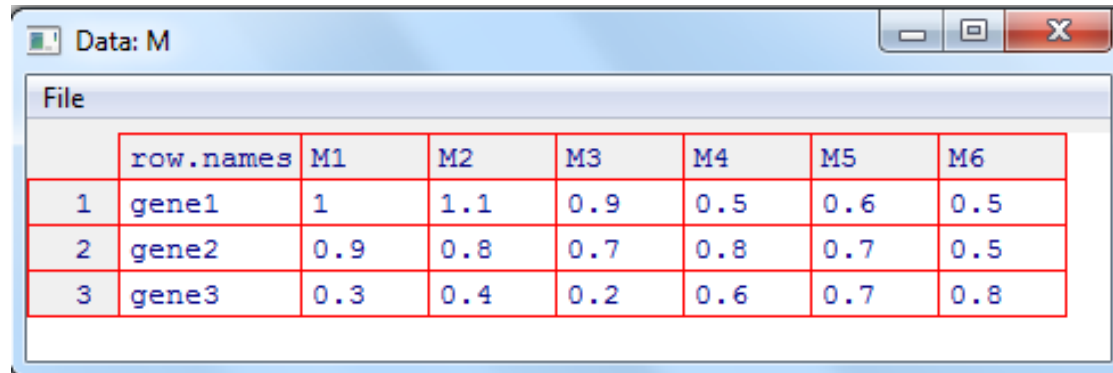
```
matrix1WayANOVATest <- function(x, g){  
  
  g <- as.factor(g)  
  Q <- summary(aov(x ~ g))[[1]]  
  p.value <- Q$Pr[1]  
  SS.g <- Q$Sum[1]; SS.tot <- sum(Q$Sum); eta2 <- SS.g/SS.tot  
  a <- c()  
  a[1] <- p.value  
  a[2] <- eta2  
  names(a) <- c("p-value", "eta2")  
  return(a)  
}
```





## EFFECT STERKTE: $\eta^2$

- Data:



	row.names	M1	M2	M3	M4	M5	M6
1	gene1	1	1.1	0.9	0.5	0.6	0.5
2	gene2	0.9	0.8	0.7	0.8	0.7	0.5
3	gene3	0.3	0.4	0.2	0.6	0.7	0.8

- Resultaat:

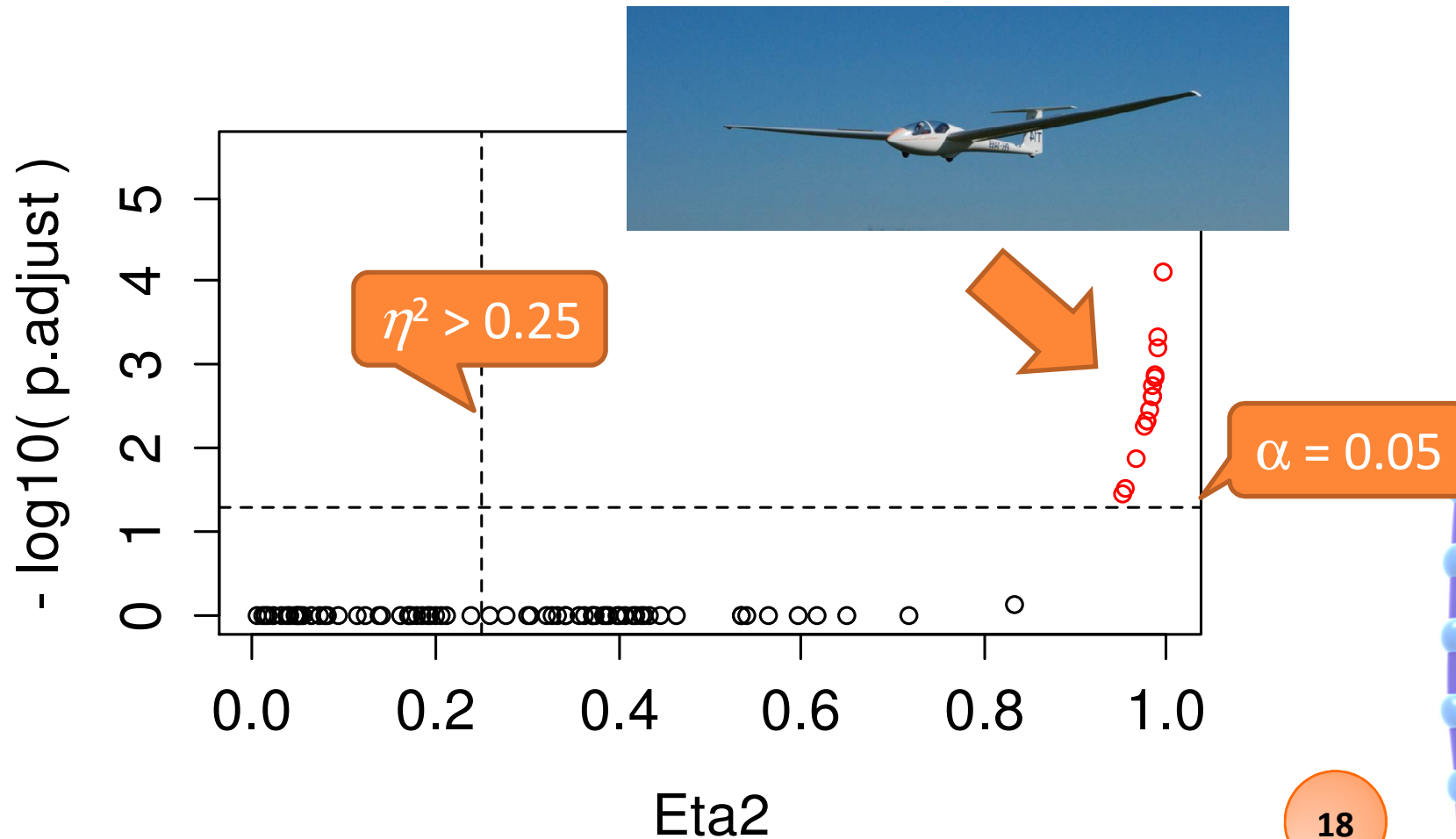
```
samples <- as.factor(c(1,1,2,2,3,3))
```

```
> apply(M, 1, matrix1WayANOVATest, g=samples)
```

	gene1	gene2	gene3
p-value	0.00219213	0.2745766	0.008049893
eta2	0.92452830	0.2857143	0.857142857

## ALTERNATIEF VOLCANO: GLIDER PLOT?

- Plot  $-\log(p_{\text{adj}})$  vs.  $\eta^2$  voor alle genen:



## ALGEMENE ANALYSE MA's: ANOVA

- Veel “ad hoc” preprocessings stappen, o.a. normalisatie, kunnen door systematische aanpak automatisch worden gedaan:
- ANOVA analyse (Jackson-groep: Churchill, Kerr, Cui)



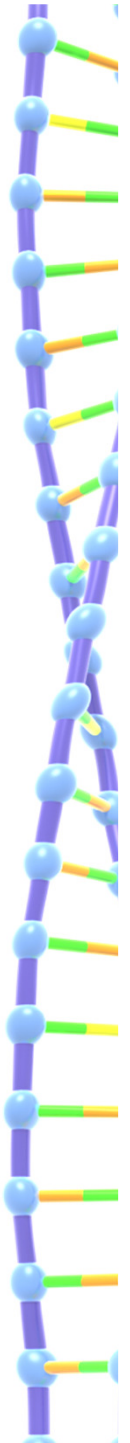
Gary Churchill



Kathleen Kerr



Xiangqui Cui

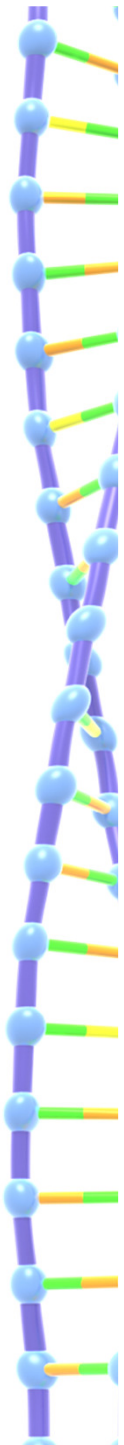


## DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- Data format 1:  $r$  replica's van  $M'$  waarden per gen:

Gene	$M'_1$	$M'_2$	...	$M'_i$	...	$M'_r$
gene 1	0.51	0.34	...	0.55	...	0.44
gene 2	-0.14	-0.31	...	0.11	...	-0.27
...						
gene $g$	0.78	0.85	...	0.69	...	0.75
...						
gene $G$	1.15		...	0.66	...	0.91

1-sample  $t$ -toets voor  $\mu = 0$



## DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, >2 SAMPLES

- **Data format 1:**  $k$  samples,  $r$  replica's  $j$  van  $M_{ij} = \log(T_{ij})$  waarden per gen per sample  $i$  :

Gene	sample 1				sample $k$		
	$M'_{11}$	...	$M'_{1r}$	...	$M'_{k1}$	...	$M'_{kr}$
gene 1	0.51	...	0.34	...	0.55	...	0.44
gene 2	-0.14	...	-0.31	...	0.11	...	-0.27
...							
gene $g$	0.78	...	0.85	...	0.69	...	0.75
...							
gene $G$	0.45	...	0.45	...	0.66	...	0.91

1-way ANOVA

## ANOVA: PER GEN OF HELE MA?

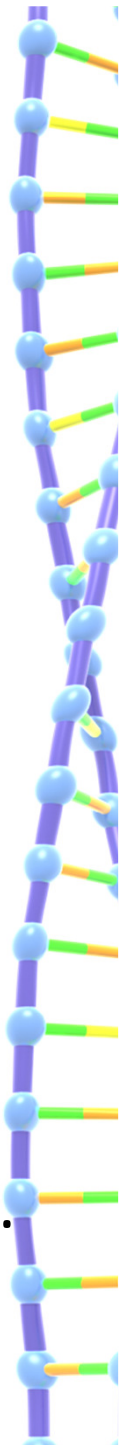
- Zoeken naar DEG's: *t*-toets of 1-way ANOVA *per gen*
- Model *per gen* (sample *s*, replica *k*):

$$M_{s,k} = \mu + S_s + \varepsilon_{s,k}$$

- **aov (x ~ S)** # S = factor met sample nrs.
- Alternatief: 2-way ANOVA op *hele microarray*:
- Model (gen *g*, sample *s*, replica *k*):

$$M_{g,s,k} = \mu + S_s + G_g + (SG)_{s,g} + \varepsilon_{g,s,k}$$

- **aov (x ~ S + G + S : G)** # G = factor met gen nrs.



## DIFFERENTIËLE GEN EXPRESSIES: 2-WAY ANOVA

- **Data format:**  $k$  samples,  $r$  replica's  $j$  van  $M_{ij} = \log(T_{ij})$  waarden per gen per sample  $i$  :

Gene	sample 1			...	sample $k$		
	$M'_{11}$	...	$M'_{1r}$		$M'_{k1}$	...	$M'_{kr}$
gene 1	0.51	...	0.34	...	0.55	...	0.44
gene 2	-0.14	...	-0.31	...	0.11	...	-0.27
...							
gene $g$	0.78	...	0.85	...	0.69	...	0.75
...							
gene $G$	1.15	...	0.45	...	0.66	...	0.91

## 2-WAY ANOVA

- 2-way ANOVA model:

$$M_{g,s,k} = \mu + S_s + G_g + (SG)_{s,g} + \epsilon_{g,s,k}$$

Gemiddelde log fold  
change: "normalisatie"

Gemiddelde effect van  
een biologisch sample  $s$

Gemiddelde effect van  
een gen  $g$

**Interactie**  
(versterkend/verzwakkend effect)  
tussen sample  $s$  en gen  $g$

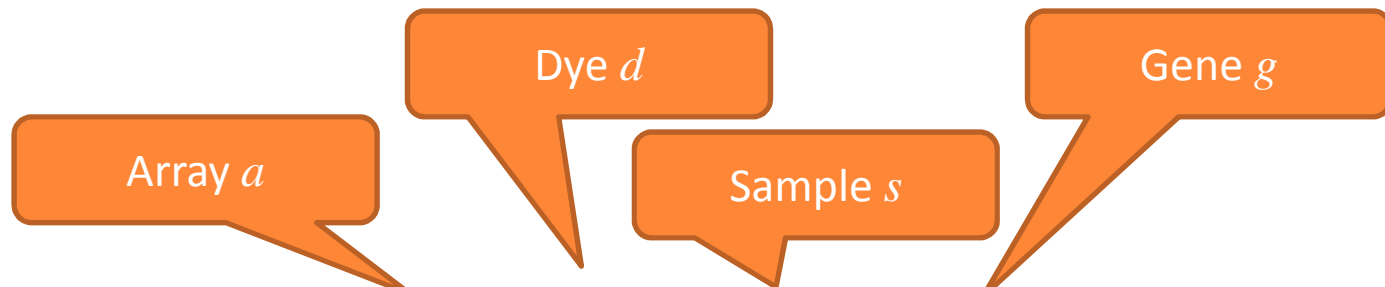

"Ruis"

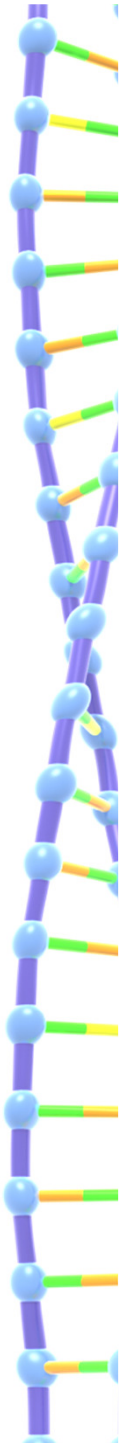
- DEG's zijn genen  $g$  waarvoor de interactie term  $(SG)_{s,g}$  in het model significant is!



## MULTI-WAY ANOVA

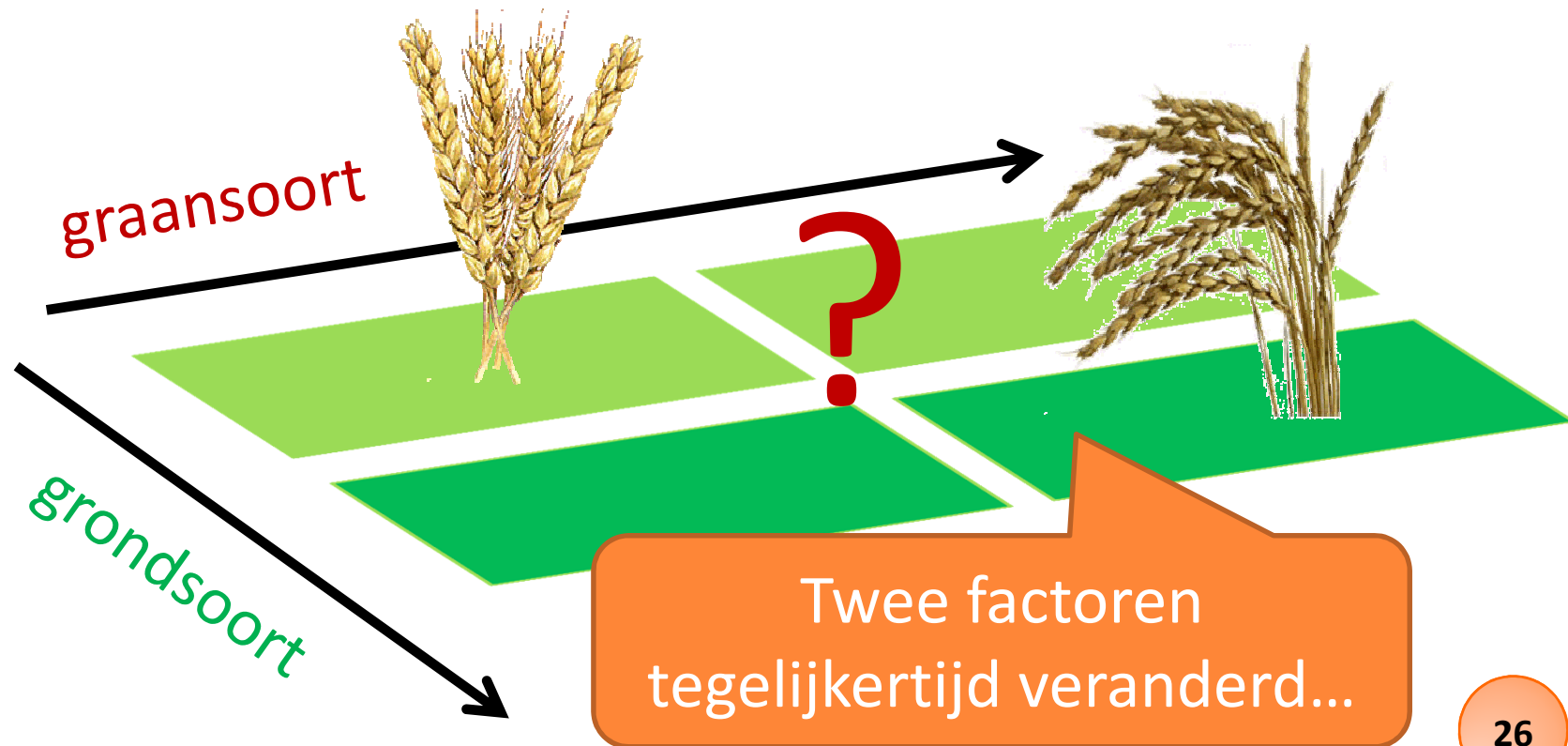
- 2-way ANOVA model is ook verder uit te breiden naar meer factoren (= verklaringen voor verschillen in log expressie waarden):


$$\begin{aligned} E_{g,s,k} = & \mu + A_a + D_d + S_s + G_g \\ & + (AG)_{a,g} + (DG)_{d,g} + (SG)_{s,g} + (DS)_{d,s} \\ & + \dots + \varepsilon_{g,s,k} \end{aligned}$$




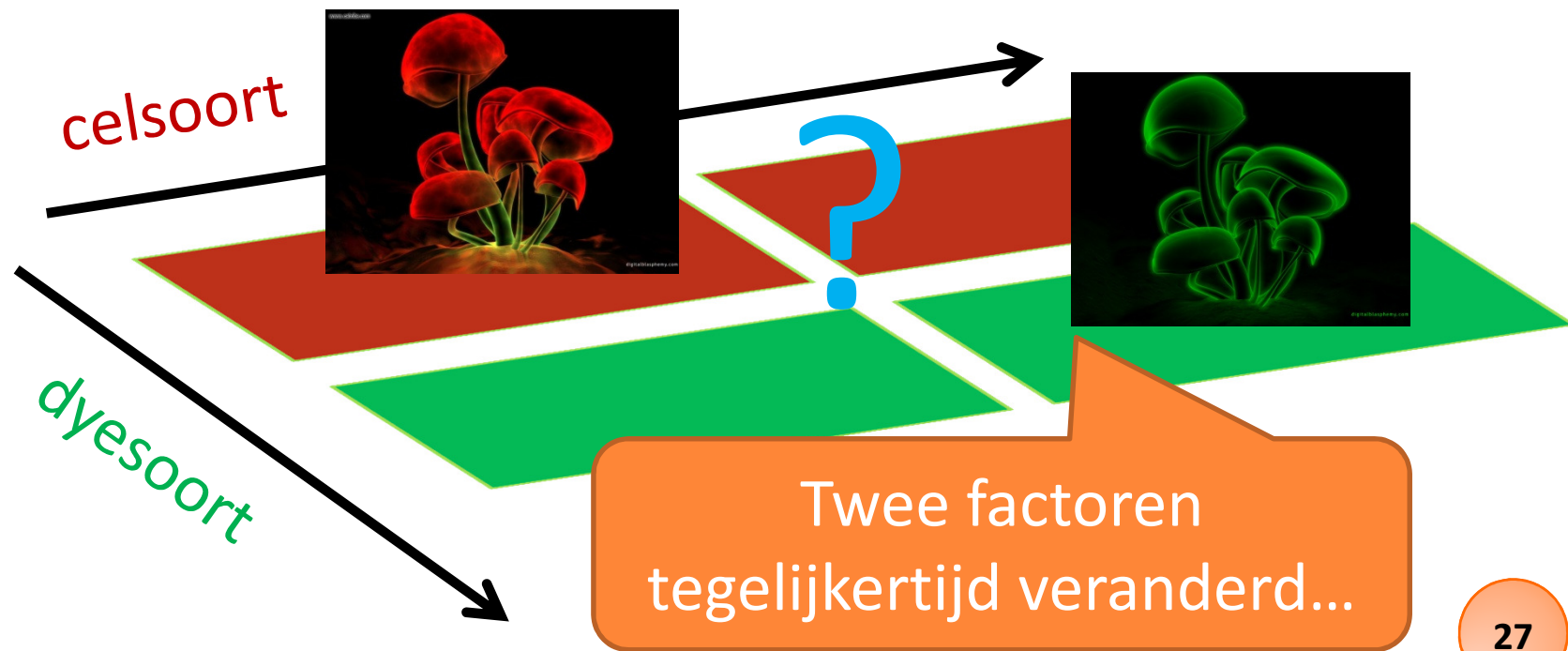
## CONFOUNDING

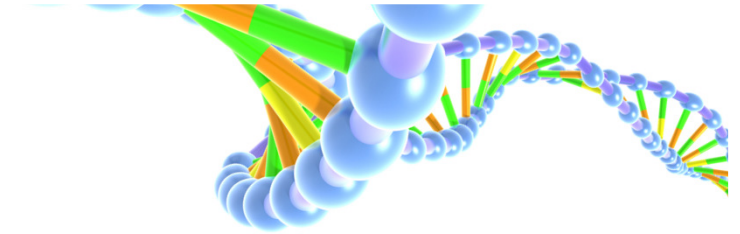
- Plantenveredelingsexperiment: 2 soorten graan op 2 soorten grond. Wat is het effect van graansoort op opbrengst?



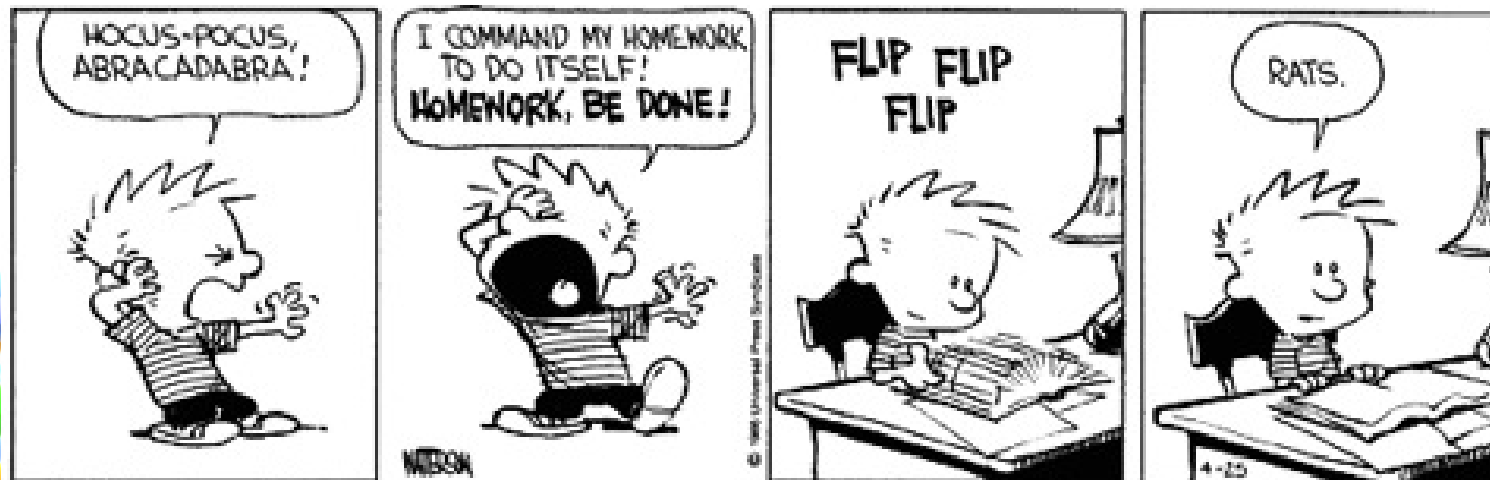
## CONFOUNDING

- Microarray experiment: 2 soorten cellen met 2 soorten dye. Wat is het effect van celsoort op fluorescentie intensiteit?





Jullie kunnen nu de opdrachten van les 13 maken



**Hanze University Groningen**  
APPLIED SCIENCES

*Institute for  
Life Science & Technology*