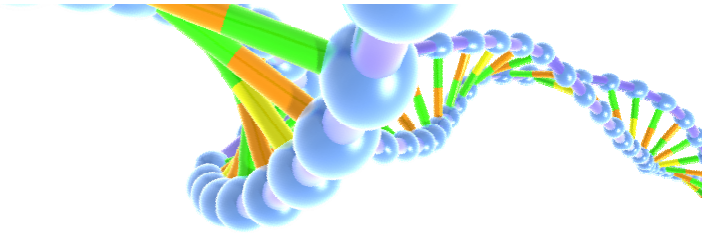
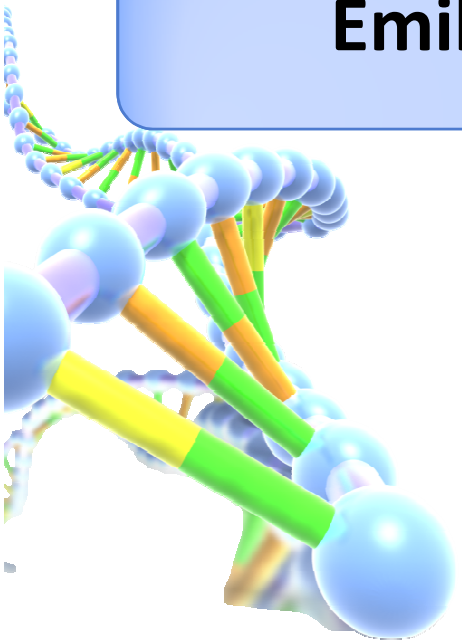


Les 7 - Microarray's en differentiële gen expressie (1)

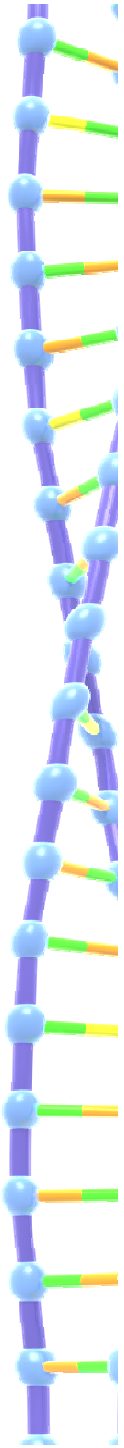
Emile Apol, Patrick Deelen, 2013-2014



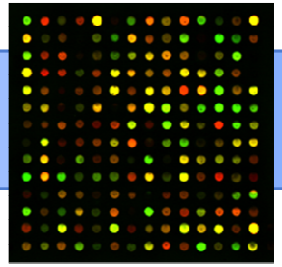
Hanze University Groningen
Applied Sciences
Institute for
Life Science & Technology

LES 7

- Microarray's
- Experimentele opzet
- Differentiële gen expressies (Differentially Expressed Genes)
- Statistische analyse per gen
  - *t*-toets
- Multiple testing correcties



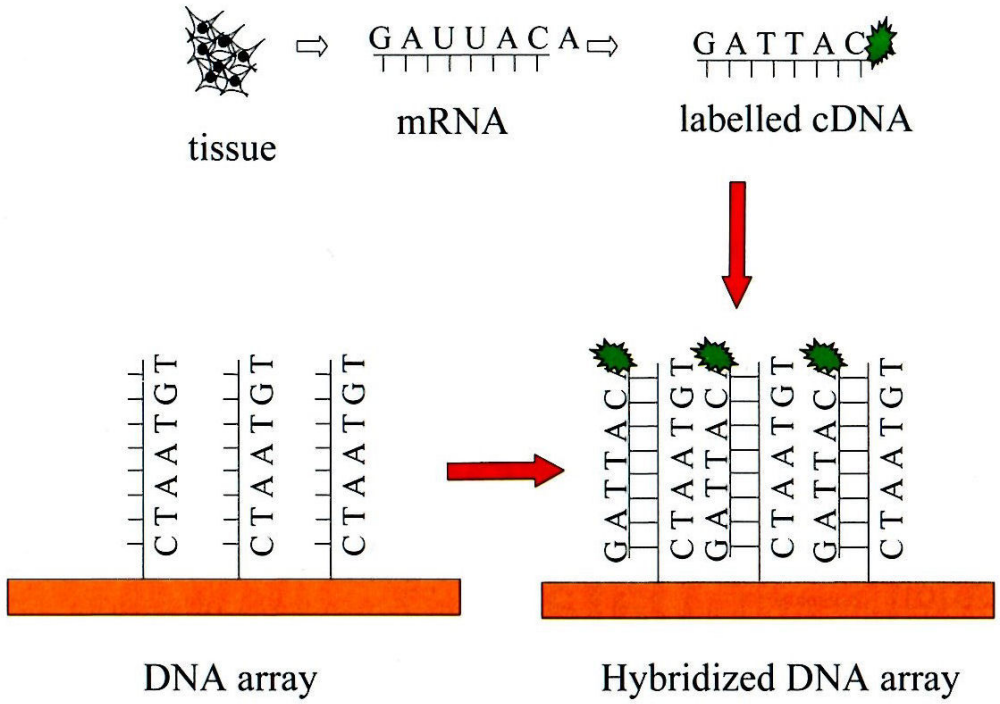
MICROARRAY ANALYSE: STAPPENPLAN

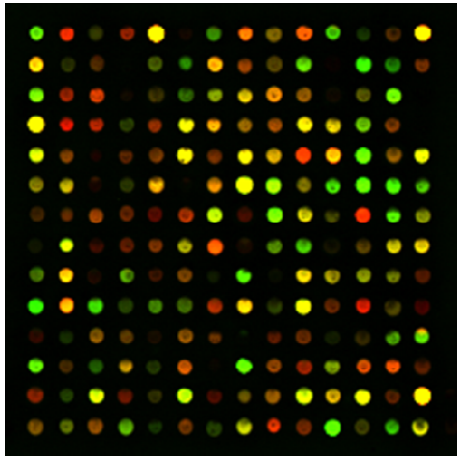


- Background correctie
- Log transformatie
- Normalisatie (bijv. loess)
- Toetsen op DEG's:
  - *t*-toets, 1-way ANOVA, …
  - Wilcoxon's toets, Kruskall-Wallis toets, …
- Aanpassen *p*-waarden voor multiple toetsing
- Clustering van DEG's:
  - Hiërarchisch clusteren
  - *k*-means
  - Principale Componenten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

MICROARRAY'S

- Schematische werking

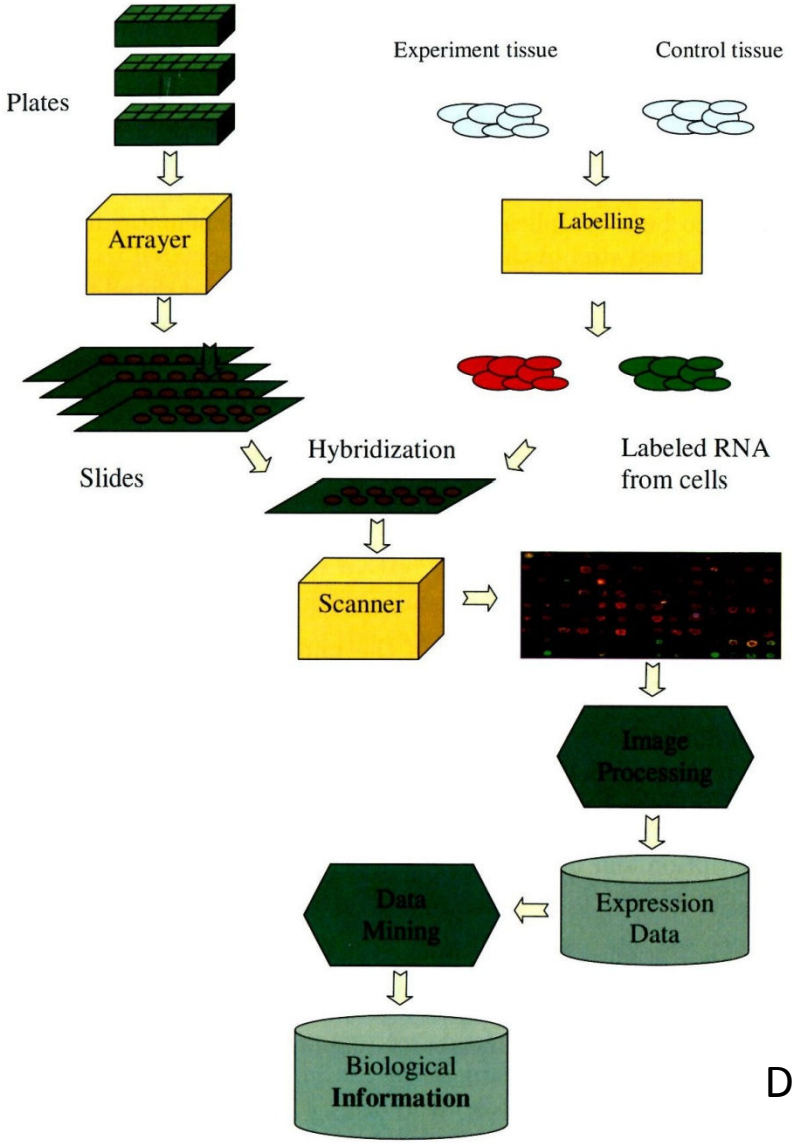




Draghici (2012)

FIGURE 3.1: A general overview of the DNA array used in gene expression studies. The mRNA extracted from tissue is transformed into complementary DNA (cDNA), which is hybridized with the DNA previously spotted on the array.

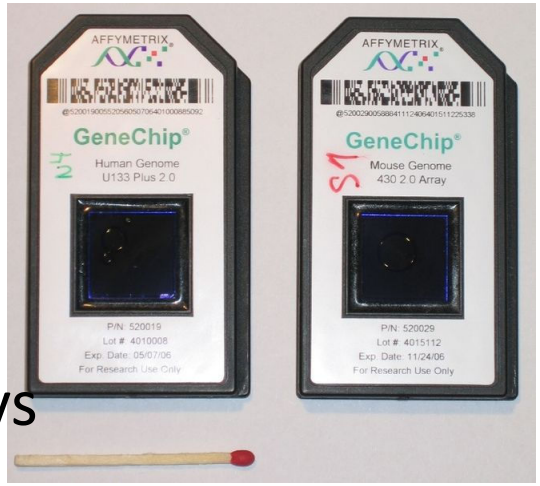
MICROARRAY'S: STAPPEN IN ANALYSE

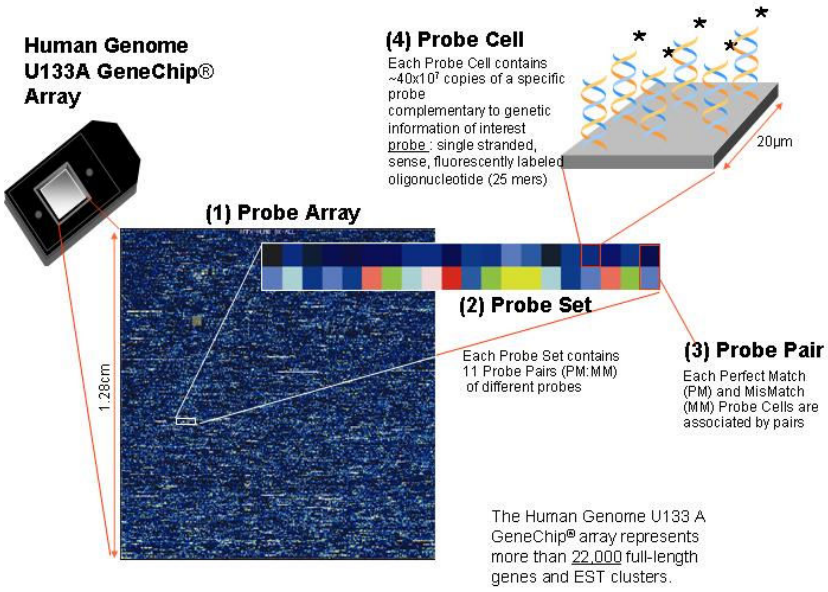


Draghici (2012)

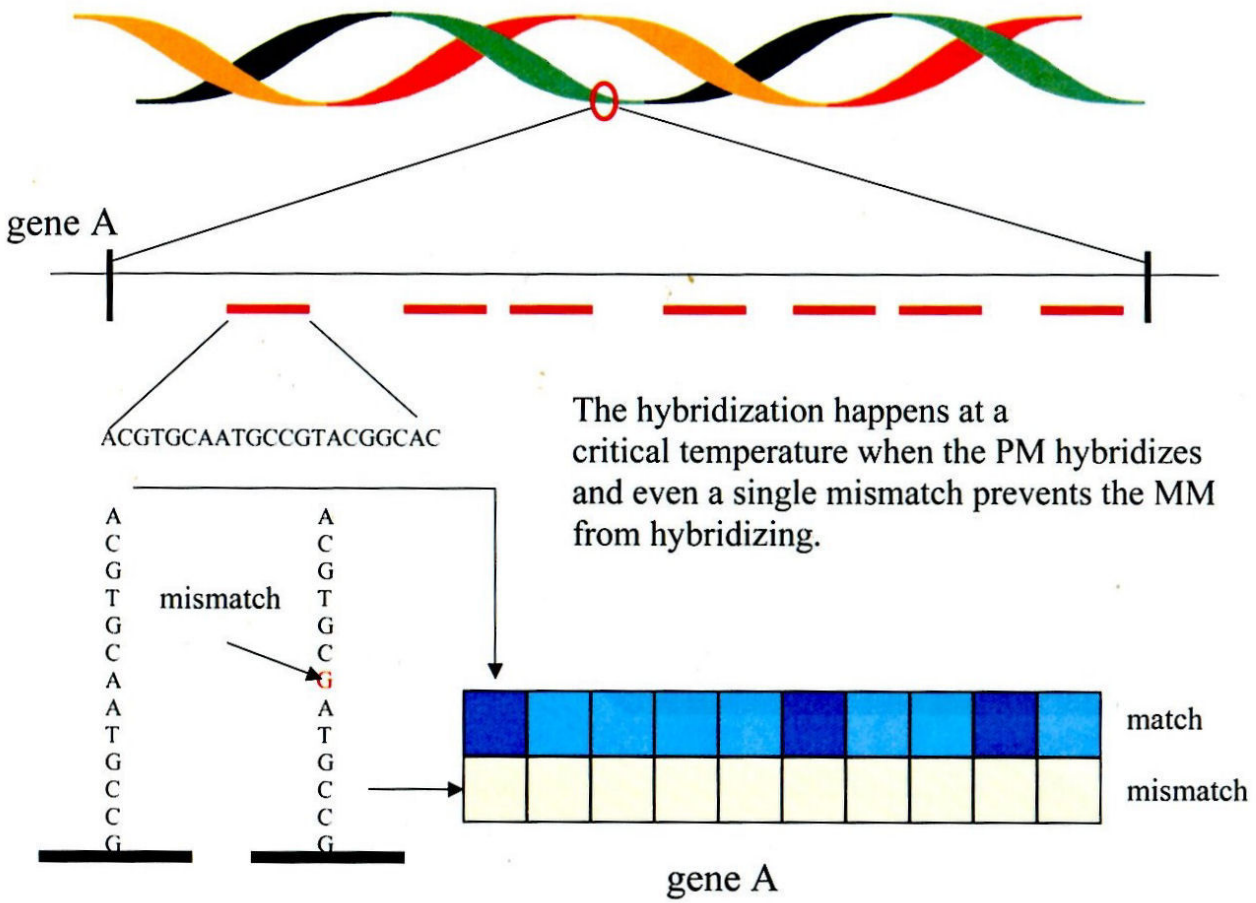
MICROARRAY'S: GEN EXPRESSIES (1-CHANNEL)

- 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





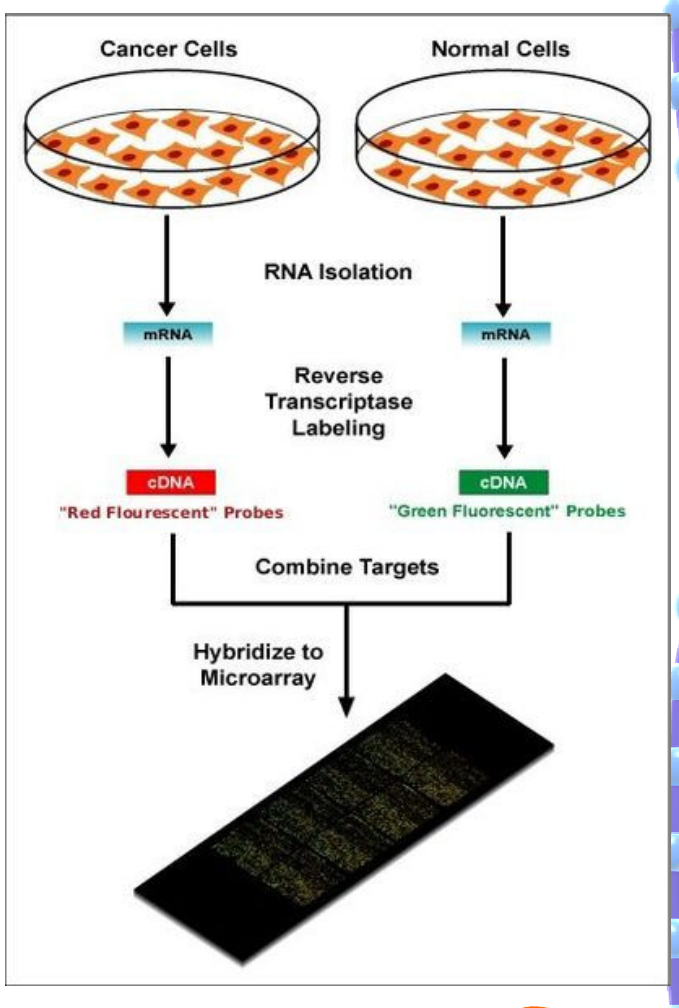
MICROARRAY'S: GEN EXPRESSIES (1-CHANNEL)



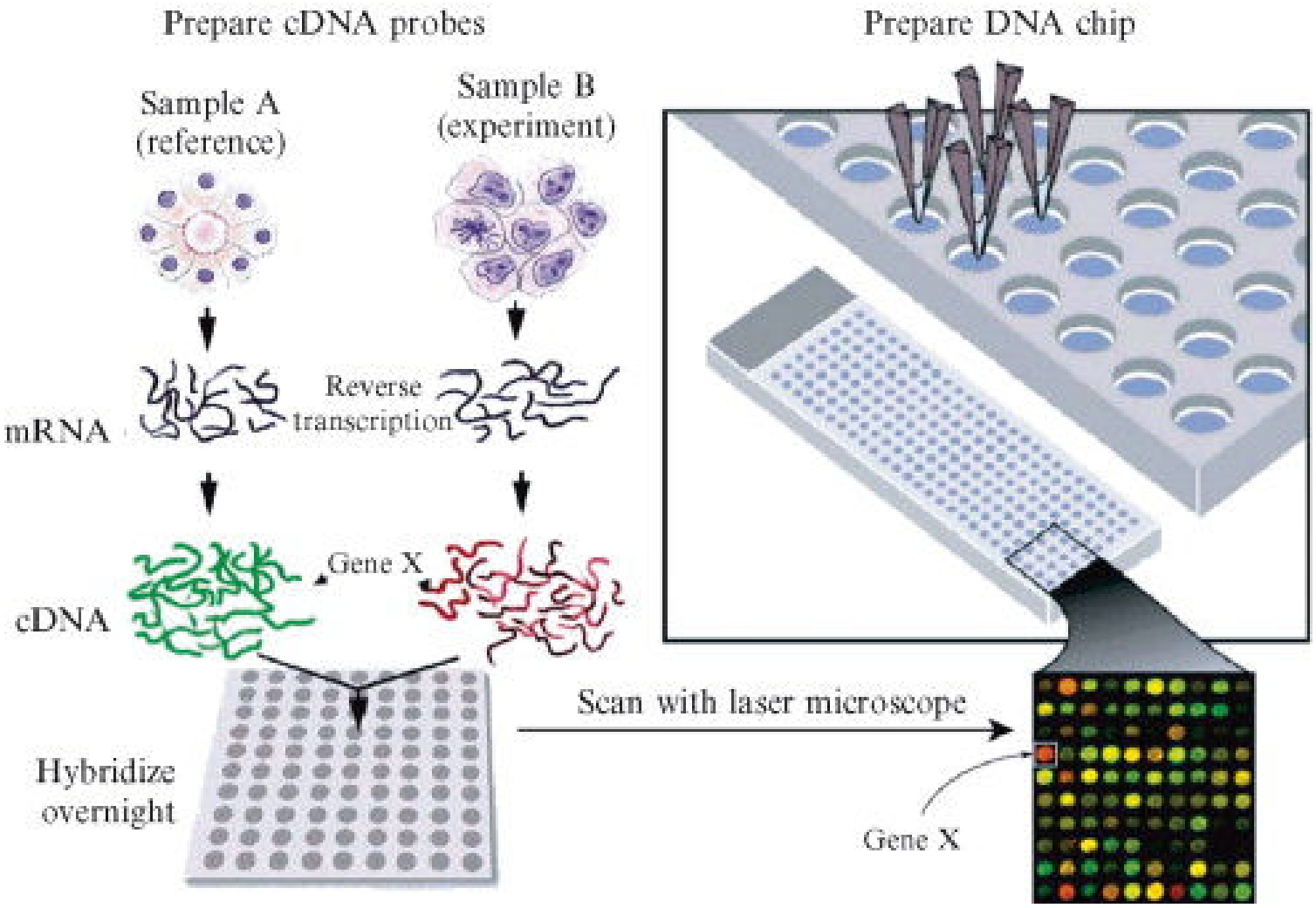
Draghici, 2012

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

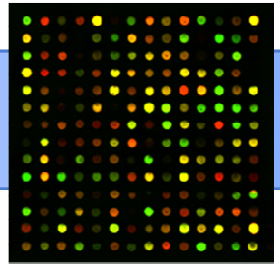


MICROARRAY'S: GEN EXPRESSIES (2-CHANNELS)



Chang et al., 2006

MICROARRAY ANALYSE: STAPPENPLAN

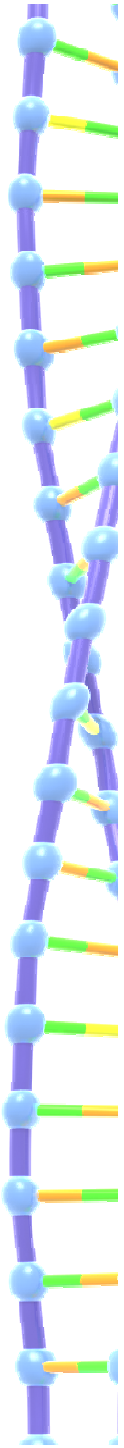


- Background correctie
- Log transformatie
- Normalisatie (bijv. loess)
- Toetsen op DEG's:
  - *t*-toets, 1-way ANOVA, …
  - Wilcoxon's toets, Kruskall-Wallis toets, …
- Aanpassen *p*-waarden voor multiple toetsing
- Clustering van DEG's:
  - Hiërarchisch clusteren
  - *k*-means
  - Principale Componenten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

DATA PROCESSING

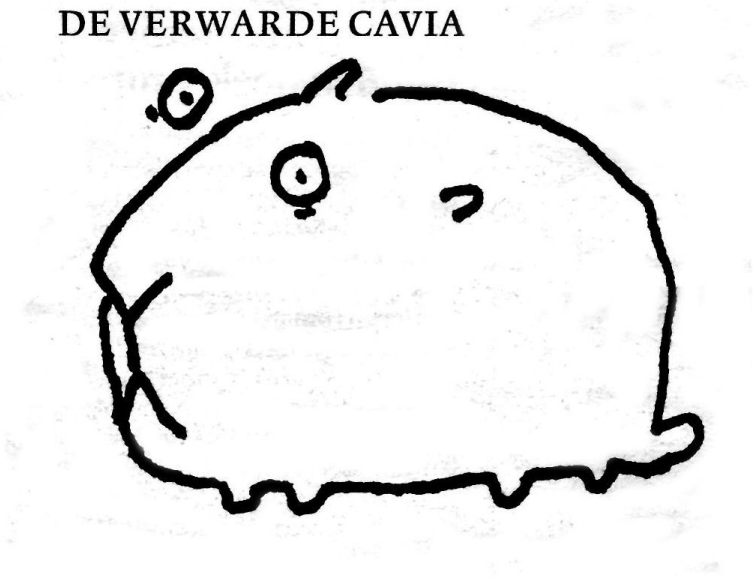
- Wikipedia (2013):

  "Microarray data is difficult to exchange due to the lack of standardization in platform fabrication, assay protocols, and analysis methods. This presents an interoperability problem in bioinformatics. Various grass-roots open-source projects are trying to ease the exchange and analysis of data produced with non-proprietary chips."



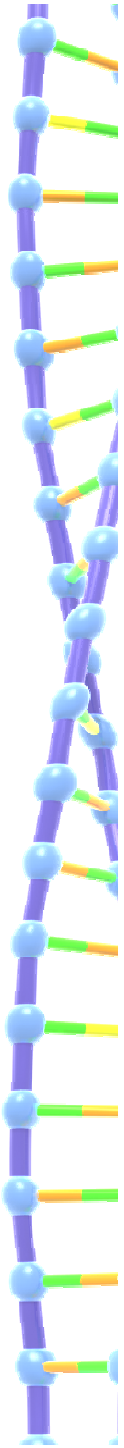
DATA PRE-PROCESSING

- 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 op zelfde array voor elk gen
  - MM is background signaal: $E = \mathrm{PM} - \mathrm{MM}$
- 2-channel methode:
  - Bij uitlezen chip ook waarden van bg per spot per channel: $R' = R - R_{\mathrm{bg}}$, $G' = G - G_{\mathrm{bg}}$
- Soms wordt *niet* gecorrigeerd voor background!

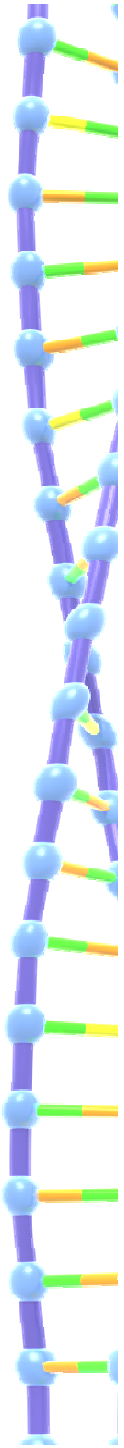


DATA PRE-PROCESSING: EXPRESSIE RATIO

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

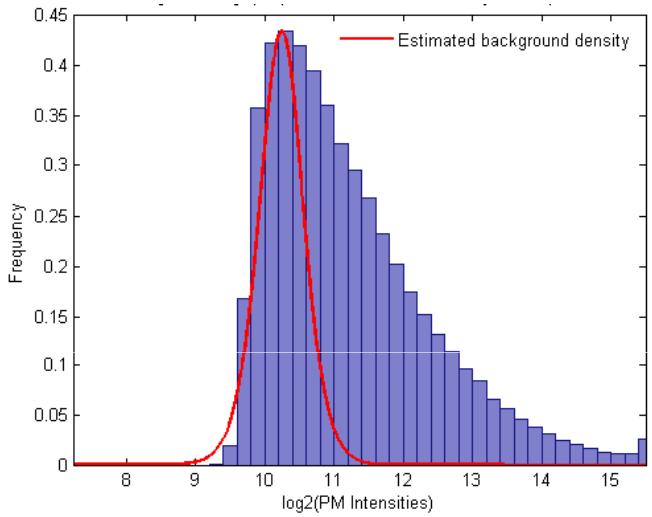
$$T' = \frac{R'}{G'}$$

In Engels ook wel de “fold change”



DATA PRE-PROCESSING: LOG TRANSFORMATIE

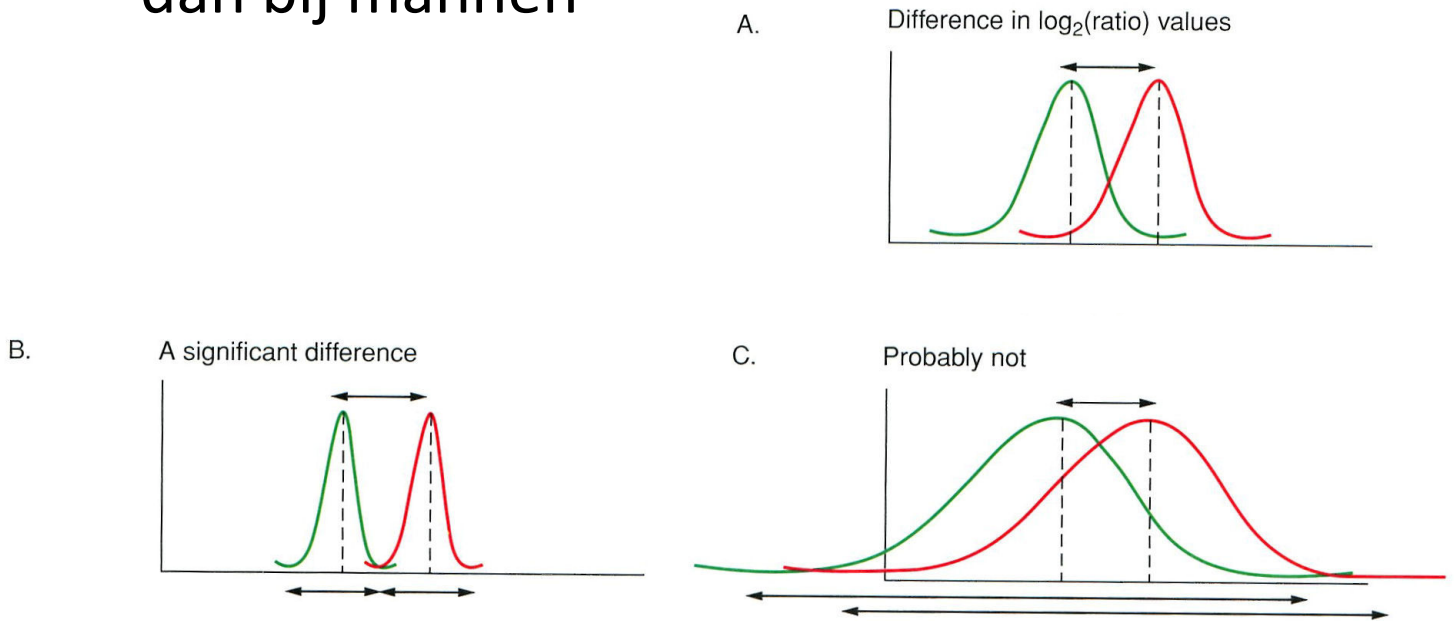
- Gen intensiteiten E hebben vaak asymmetrische verdeling:



- 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

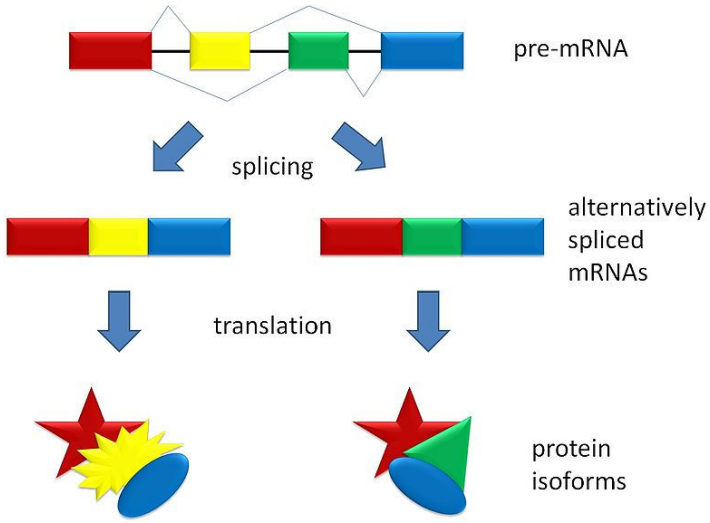
DIFFERENTIALLY EXPRESSED GENES (DEGS)

- Genen met verschillende expressie tussen sample groep(en)
  - Bijv. gen X komt *meer tot expressie* bij gezonde mensen ten opzichten 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



Wikipedia

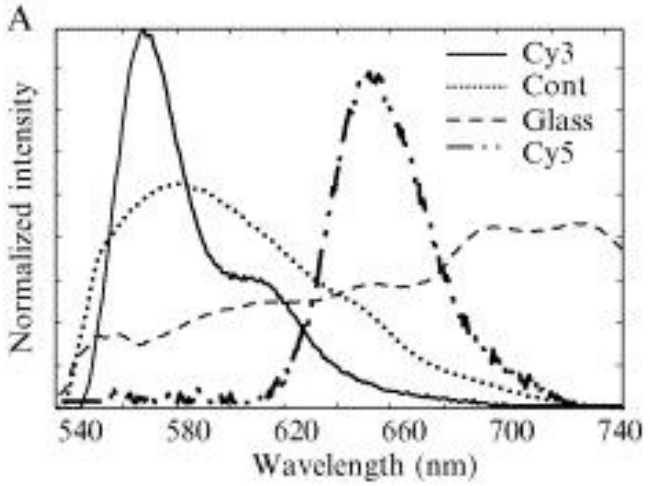
- Interpretatie: een gen dat (significant) hoger tot expressie komt in kanker cellen hoeft niet per se de oorzaak te zijn van kanker
  - Hoeft zelfs niet eens belangrijk te zijn

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/…

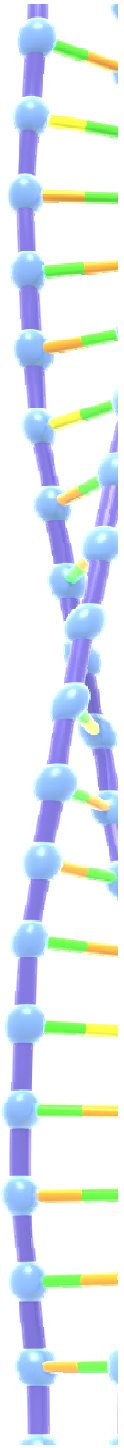
- Verschil tussen weefsel
- Verschil tussen individuen

- 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



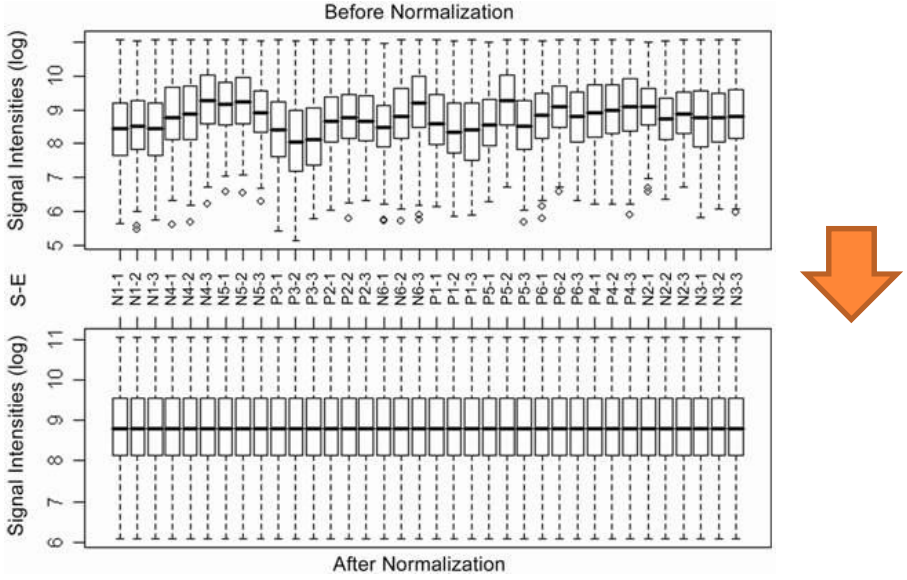
DATA PRE-PROCESSING: NORMALISATIE

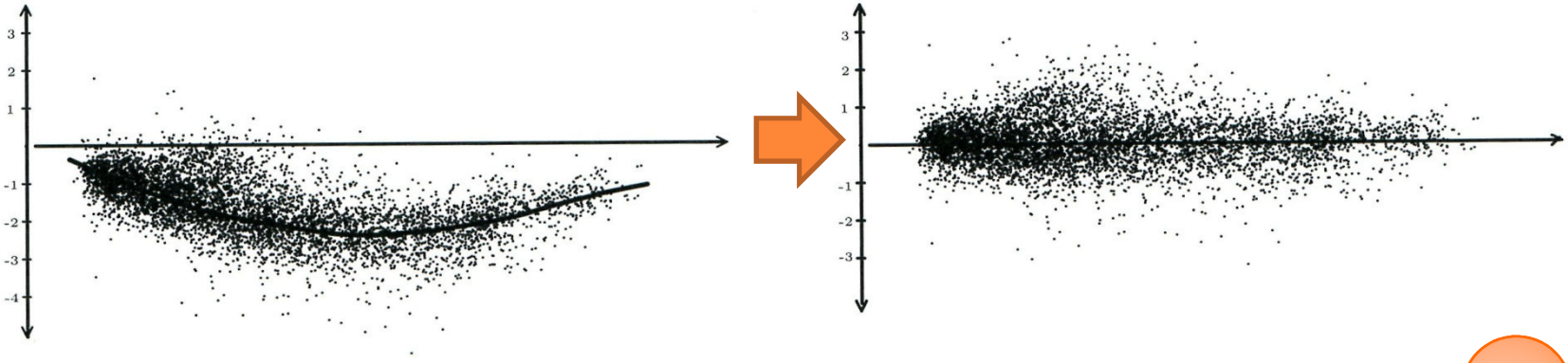
- Scaling normalization:

$$x = \frac{\log(T_i') - \overline{\log(T_i')_a}}{s_a}$$

$$x = \frac{\log(T_i') - \text{median}_a}{MAD_a}$$

- Loess normalization:



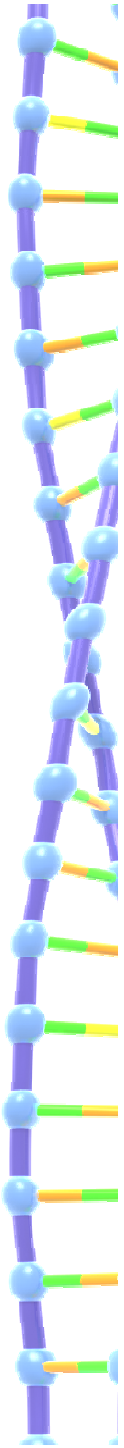


DATA PRE-PROCESSING: LOG TRANSFORMATIE

- Differentiële Gen Expressie:

$$^2\log T' = {}^2\log\left(\frac{R'}{G'}\right) = {}^2\log R' - {}^2\log G'$$

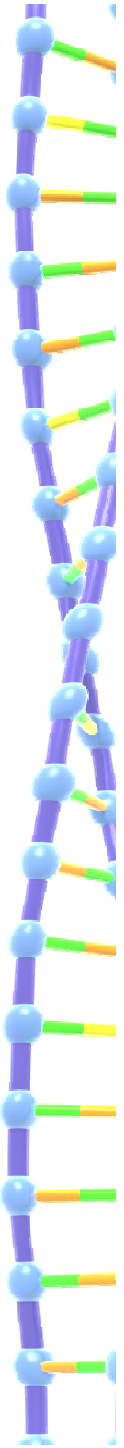
Gen regulatie	R', G'	T'	$^2\log T'$
upregulatie	$R' > G'$	$T' > 1$	$^2\log T' > 0$
niet	$R' = G'$	$T' = 1$	$^2\log T' = 0$
downregulatie	$R' < G'$	$T' < 1$	$^2\log T' < 0$



DATA PRE-PROCESSING: LOG TRANSFORMATIE

- Betekenis van log-waarden:

$^2\log T'$	T'		DE
-3	0.125	8 x kleiner	down regulatie
-2	0.25	4 x kleiner	
-1	0.5	2 x kleiner	
0	1	gelijk	
1	2	2 x groter	up regulatie
2	4	4 x groter	
3	8	8 x groter	

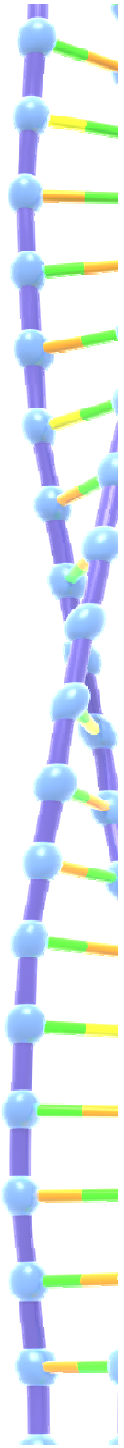


Differentiële Gen Expressies: 1 gen

- Toets of log(genexpressies) verschillen tussen (biologische) samples
- Parametrische toetsen (aanname: normaal verdeeld)
  - 2 samples: *t*-toets
  - > 2 samples: 1-way ANOVA

  → *p*-waarde per gen
- Niet-parametrische toetsen (alternatief: aanname dat verdelingen gelijk zijn)
  - 2 samples: Wilcoxon's toets
  - > 2 samples: Kruskal-Wallis toets

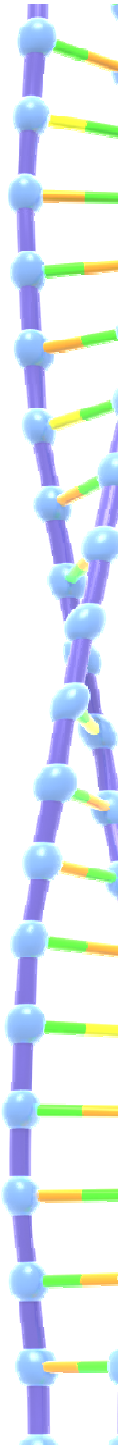
  → *p*-waarde per gen



DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- Data format 1: r replica's van $\log(T')$ waarden per gen:

Gene	$\log(T')_1$	$\log(T')_2$	...	$\log(T')_i$	...	$\log(T')_r$
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	0.45	...	0.66	...	0.91



DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- Data format 1: r replica's van $\log(T')$ waarden per gen:

Gene	$\log(T')_1$	$\log(T')_2$	…	$\log(T')_i$	…	$\log(T')_r$
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 2: r replica's van $\log(R')$ en $\log(G')$ waarden per gen:

Gene	$\log(R')_1$	$\log(R')_2$	...	$\log(R')_r$	$\log(G')_1$	$\log(G')_2$	...	$\log(G')_r$
gene 1	8.54	8.68	...	7.99	7.77	7.92	...	8.13
gene 2	8.11	8.31	...	8.25	8.35	8.46	...	8.27
...								
gene g	9.78	9.85	...	9.55	8.69	8.42	...	8.75
...								
gene G	11.15	11.45	...	10.98	10.66	10.54	...	10.91

DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- Data format 2: r replica's van $\log(R')$ en $\log(G')$ waarden per gen:

Gene	$\log(R')_1$	$\log(R')_2$	...	$\log(R')_r$	$\log(G')_1$	$\log(G')_2$	...	$\log(G')_r$
gene 1	8.54	8.68	...	7.99	7.77	7.92	...	8.13
gene 2	8.11	8.31	...	8.25	[illegible]	8.46	...	8.27
...								
gene g	9.78	9.85	...	9.55	8.69	8.42	...	8.75
...								
gene G	11.15	11	..	10.98	10.66	10.54	...	10.91

gepaarde *t*-toets (paring van *R'* en *G'*)

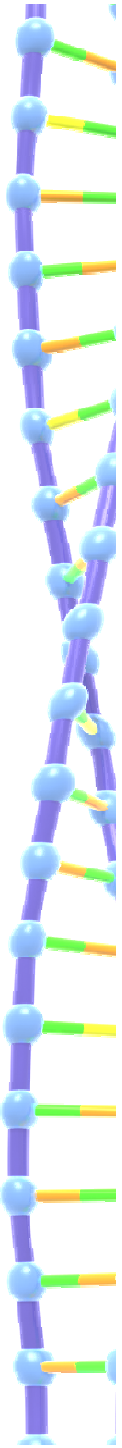
DIFFERENTIËLE GEN EXPRESSIES: 1 GEN, 2 SAMPLES

- 1-sample *t*-toets vs gepaarde *t*-toets: zelfde?

$$t_{1-\text{sample}} = \frac{\overline{\log(T')} - 0}{s_{\overline{\log(T')}}} = \frac{\overline{\log(R'/G')} - 0}{s_{\overline{\log(R'/G')}}}$$

$$= \frac{\overline{\log(R') - \log(G')} - 0}{s_{\overline{\log(R') - \log(G')}}} = t_{\text{gepaard}}$$

Ja! Beide data formaten geven zelfde resultaten (*p*-waarden)



MULTIPLE TOETSEN

- Per toets een kans α op vals positieve (FP) uitslag, d.w.z. H_0 waar maar toch verwerpen
- Bij G (= aantal genen) toetsen achter elkaar dus een verwacht (= gemiddeld) aantal vals positieven:

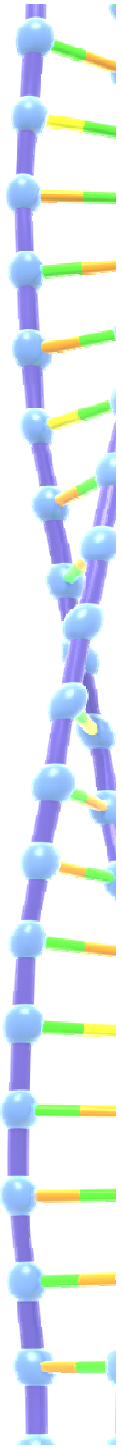
$$\overline{FP} = \alpha \cdot G$$

- Voorbeeld: 10 000 genen met α = 0.05 geeft

$$\overline{FP} = 0.05 \cdot 10\,000 = 500$$

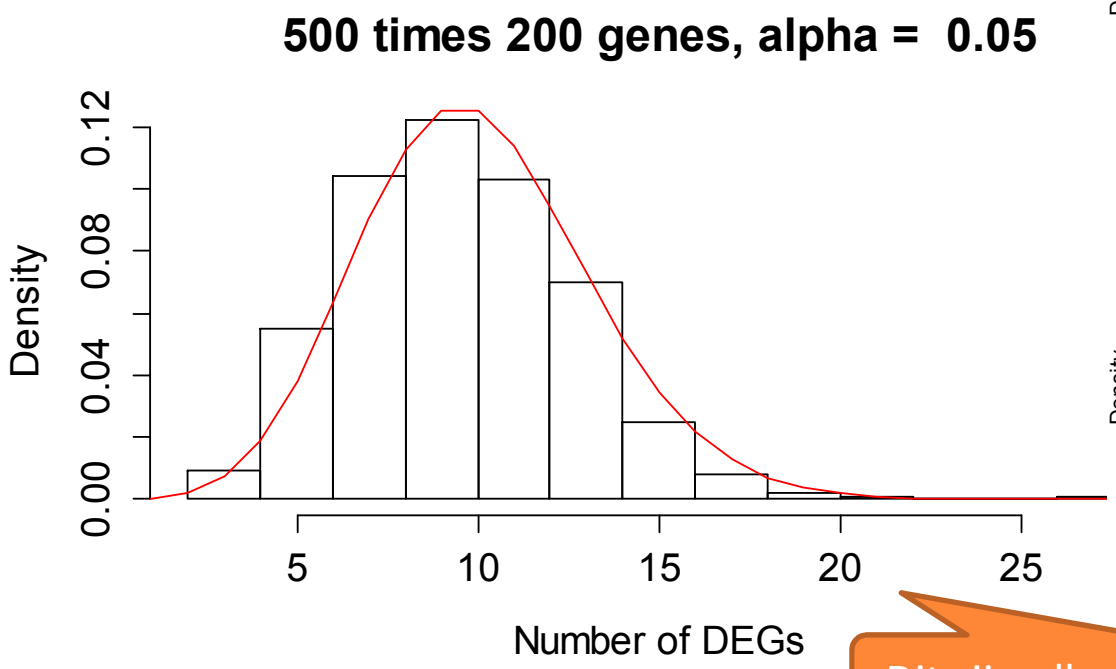
- De verdeling van FP is ongeveer Poisson verdeeld, met

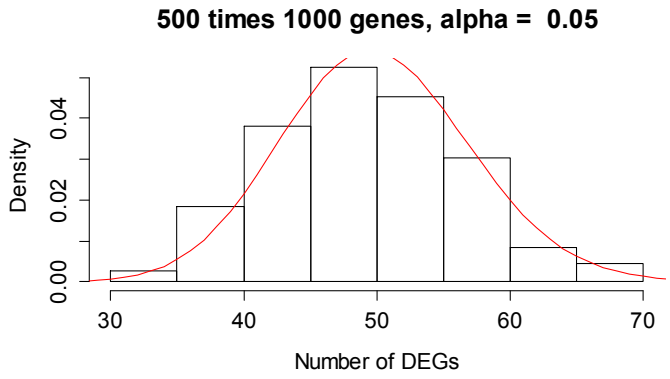
$$\lambda = \overline{FP} \qquad p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

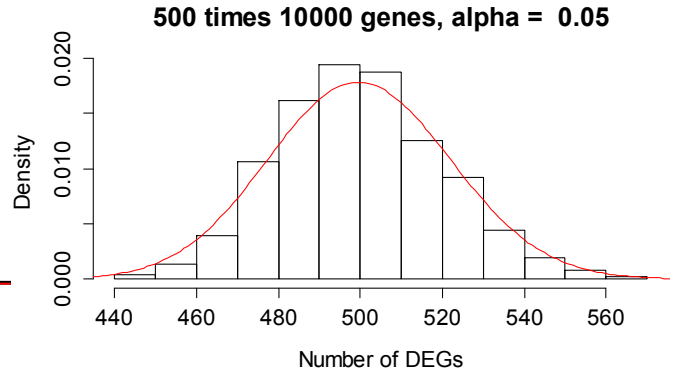


MULTIPLE TOETSEN

- Simulaties in R (G = 200, 1 000 of 10 000 genen, 500 x random MA met 5 replica's), met theoretische Poisson verdeling (in rood):







Dit zijn allemaal vals positieven!

MULTIPLE TOETSEN

- We voeren een series (= “familie”) van R (t-)toetsen uit, elk met een significantie van α_C :
- De Family-Wise Error Rate α_F voor R toetsen:

kans op minstens 1x vals positief

$$\alpha_F = 1-(1-\alpha_C)^R \quad \Rightarrow \quad \alpha_C = 1-(1-\alpha_F)^{1/R}$$

- Voorbeeld: 15 testen met α_C = 0.05:

$$\alpha_F = 1-(1-0.05)^{15} = 0.54 \quad \Rightarrow$$

$$\alpha_C = 1-(1-0.05)^{1/R} = 0.0034$$

aangepaste significantie!

- We kunnen
  - het significantie niveau per test α_C aanpassen (verkleinen)
  - de p-waarde per test aanpassen (vergroten)

MULTIPLE TOETSEN

- Aanpassen van *p*-waarden in R: **`p.adjust()`**
- Kan via verschillende methoden voor R toetsen
- Sidak methode:

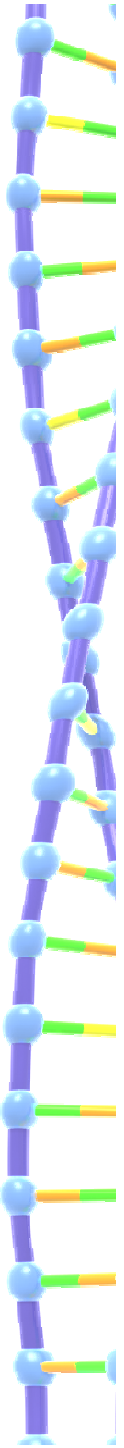
$$\alpha_F = 1-(1-\alpha_C)^R \quad \Rightarrow \quad p_{\text{adjust}} = 1-(1-p)^R$$

- Bonferroni methode:

$$\alpha_F = 1-(1-\alpha_C)^R \approx R\,\alpha_C \quad \Rightarrow \quad p_{\text{adjust}} = R\,p$$

- Holm (Holm-Bonferroni) methode:
  - Orden *p*-waarden oplopend: $i = 1$ (kleinste *p*-waarde)

$$p_{\text{adjust},i} = (R-i+1)\,p_i$$



MULTIPLE TOETSEN

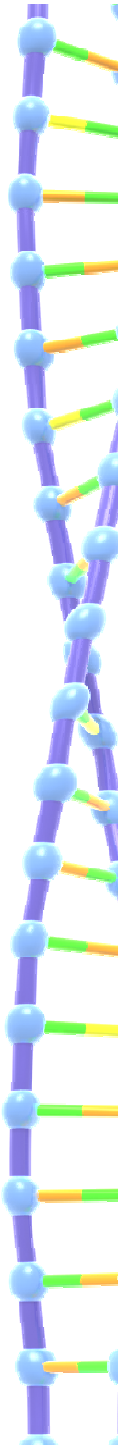
- "Klassieke" statistiek: 5 groepen, dus R = 10

$$\alpha_C = 1-(1-0.05)^{1/10} = 0.0051$$

- Microarray: 10 000 genen, dus R = 10 000

$$\alpha_C = 1-(1-0.05)^{1/10000} = 5.1\cdot 10^{-6}$$

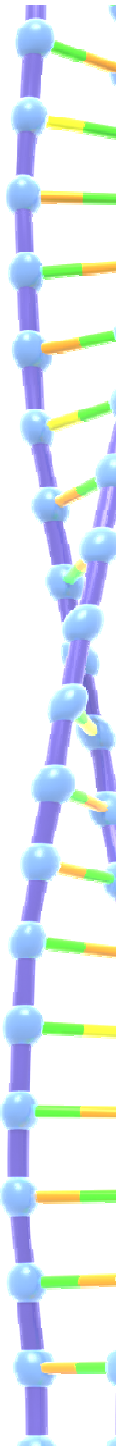
- De Sidak en Bonferroni correcties zijn erg conservatief: met de nauwkeurigheid van microarrays zijn er bijna geen significante differential gene expressions!
- Beter: Holm methode of resampling (software)



MULTIPLE TOETS CORRECTIE IN R

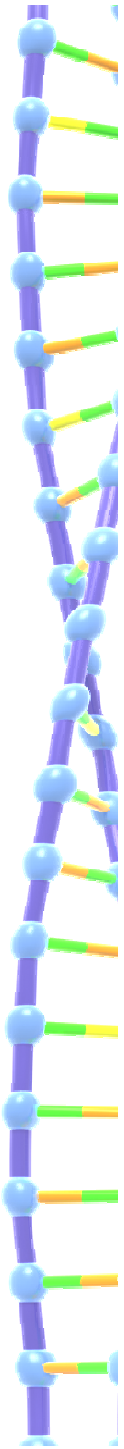
- **`pVec`** is vector met *p*-waarden uit multiple toetsen (bijv. *t*-toetsen)
  - **`pVec <- c(0.01,0.05,0.005,0.0015,0.2)`**
- Sidak methode: niet aanwezig
- Bonferroni methode:
  - **`p.adjust(pVec, method="bonferroni")`**
- Holm methode:
  - **`p.adjust(pVec, method="holm")`**

```
> pVec
[1] 0.0100 0.0500 0.0050 0.0015 0.2000
> p.adjust(pVec, method="bonferroni")
[1] 0.0500 0.2500 0.0250 0.0075 1.0000
> p.adjust(pVec, method="holm")
[1] 0.0300 0.1000 0.0200 0.0075 0.2000
```



MULTIPLE T-TOETS CORRECTIE IN R

- Geen correctie:
  - `pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="none")`
- Sidak methode: bijna als Bonferroni , en niet aanwezig (maar zie opgave!)
- Bonferroni methode:
  - `pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="bonferroni")`
- Holm methode:
  - `pairwise.t.test(y, sample, pool.sd=T, p.adjust.method="holm")`



MULTIPLE TOETSEN: FALSE DISCOVERY RATE

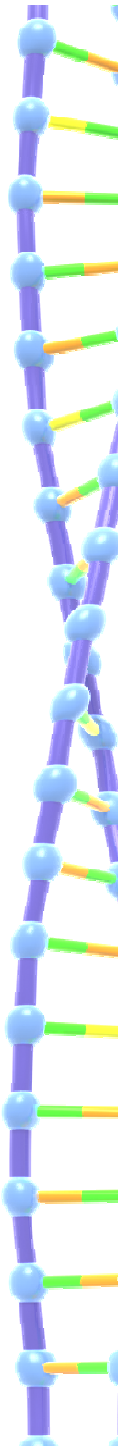
- Vorige methodes (Bonferroni, Sidak, Holm) corrigeren de Family-Wise Error Rate (FWER) α_F
- Andere aanpak: corrigeer de False Discovery Rate (FDR)

$$FDR = \frac{FP}{FP + TP}$$

		Waarheid	
		H_0 is waar	H_0 is niet waar
Uitkomst toets	H_0 niet verworpen	true negatives (goede beslissing) $1 - \alpha$	false negatives (Type II error) β
	H_0 verworpen	false positives (Type I error) α	true positives (goede beslissing) $1 - \beta$

MULTIPLE TOETSEN: FALSE DISCOVERY RATE

- Probleem:
- Methoden om de FWER (Family-Wise Error Rate α_F) te controleren zijn “te streng”
  - Gecorrigeerde α_C waarden zijn zó klein voor duizenden genen, dat kleiner dan technisch haalbare precisie van microarray’s!
  - Geen enkel gen als DEG gedetecteerd…
- Methoden die de FDR (False Discovery Rate) controleren zijn “minder streng”
  - Zo binnen de haalbare precisie van microarray’s toch DEGs detecteren

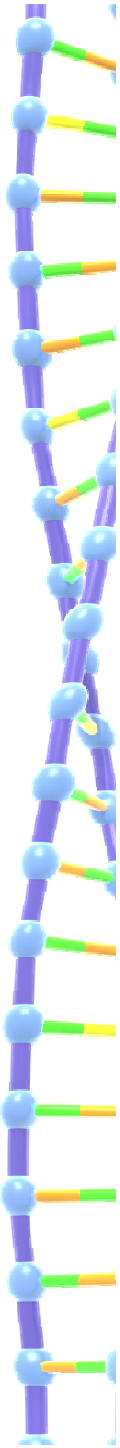


MULTIPLE TOETSEN: FALSE DISCOVERY RATE

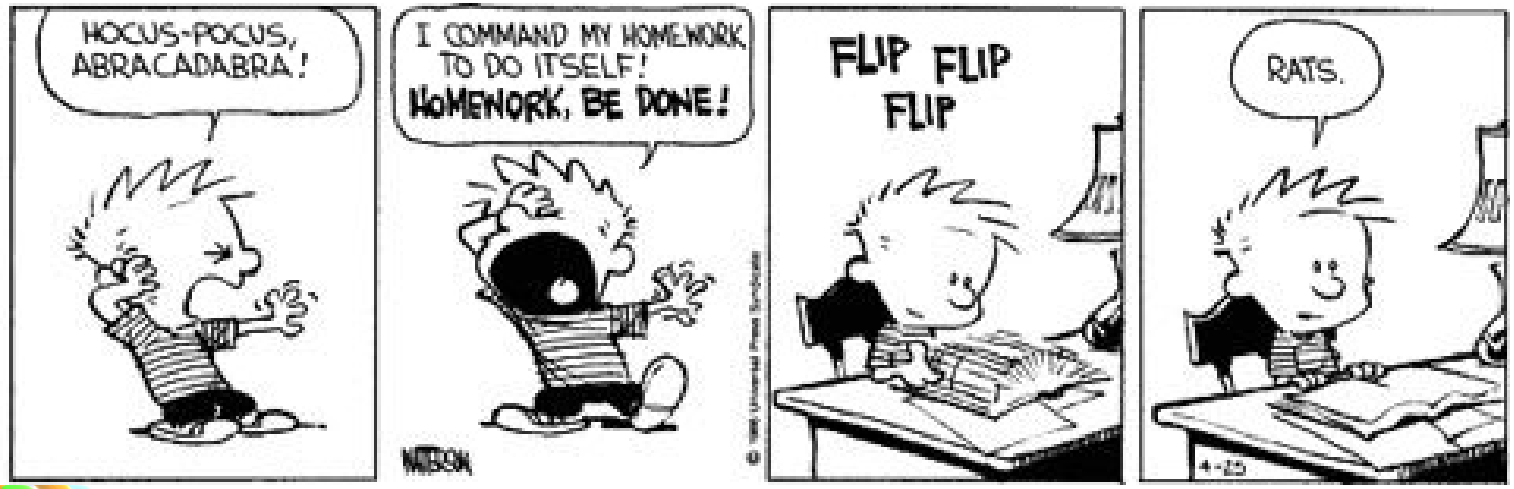
- De methoden van Benjamini & Hochberg (1995)
  - `p.adjust(pVec, method = "BH")`
  - `p.adjust(pVec, method = "fdr")`

  of Benjamini & Yekutieli (2001)
  - `p.adjust(pVec, method = "BY")`

  zijn in R geïmplementeerd
- Ook in `pairwise.t.test(…, p.adjust.method = …)`
- Meer functionaliteit in package **multtest** (Bioconductor)



Jullie kunnen nu de opdrachten van les 7 maken



Hanze University Groningen
Applied Sciences
Institute for
Life Science & Technology