| | | |
|---|---|---|
| **Code :**<br>BFVH3STA3 | **Voorbeeld tentamen:**<br>Statistiek 3 | |
| **Datum:**<br>2013 | **Tijd:** | **School:**<br>ILST |
| **Lokaal:** | **Klas:**<br>BFV2 / VWO-BFV2 | **Duur:** 90 min |
| **Docent :** M.E.F. Apol<br>**Tijdens het tentamen te bereiken onder nummer:** | | **Aantal pagina's:** 5<br>(incl. voorblad en bijlage) |

**Hulpmiddelen:**

PowerPoint presentaties Statistiek 3 (als pdf)

Extra R-functies

An Introduction to R (als pdf)

R Reference Card (als pdf)

R, RStudio en LibreOffice Writer software in een gesloten omgeving zonder internet verbinding

R packages: gplots, limma

**Overig hulpmiddelen:**

Kladpapier

**Opgave inleveren:** Ja

**Kladpapier inleveren:** Ja

***Bijzonderheden:***

*Typ je **antwoorden** in een **LibreOffice Writer** file, geef daarbij goed aan met welke vraag je bezig bent, en lever je antwoorden uiteindelijk in als **pdf file**. Kopieer in deze file ook relevante (statistische) uitvoer van R, die nodig is om de vraag te beantwoorden.*
*Lever elke gevraagde **grafiek** in als **pdf file** (los of samengevoegd).*
*Lever daarnaast je **R script** in die je hebt gebruikt om het tentamen te maken. Geef daarin duidelijk aan of je eventueel extra packages of functies hebt gebruikt. Geef ook de functiedefinitie van eigen functies (m.u.v. de bijgeleverde extra R-functies).*

*Heel veel succes!*

| | | |
|---|---|---|
| **Naam student:** | **Klas:** | **Studentnummer:** |

**Tentamencijfer = 1 + 9 x punten/160**

## Opgave 1. Kiezen van een statistische toets voor DEG's

Geef voor elk van de volgende microarray experimenten/datasets aan met welke statistische toets je kunt onderzoeken welke genen differentieel tot expressie komen. Je kunt daarbij kiezen uit:

- 1-sample *t*-toets
- 2-sample/Welch *t*-toets
- gepaarde *t*-toets
- 1-way ANOVA
- Wilcoxon's signed-rank toets
- Mann-Whitney toets = Wilcoxon's rank-sum toets
- Kruskal-Wallis toets

Soms zijn meerdere opties goed. Leg je keuze kort uit. De data zijn op de juiste manier voorbewerkt (d.w.z. background correctie, normalisatie, log2-transformatie).

**a.** [10 pt] Microarray data file met $M = {}^2\log(R/G)$ waarden van dual-channel ($R$ = rood, $G$ = groen) chips voor 2 biologische samples ($R$ = reuma, $G$ = controle). Er werden 6 replica's gemeten (op 6 verschillende chips). Het data format:

| gene | M.1 | M.2 | M.3 | M.4 | M.5 | M.6 |
|---|---|---|---|---|---|---|
| ABC_7 | 0.71 | 0.53 | 0.33 | 0.45 | 0.36 | 0.63 |
| ABC_8 | -1.31 | -1.27 | -1.14 | -1.22 | -1.39 | -0.99 |
| … | … | … | … | … | … | … |

**b.** [10 pt] Microarray data file met $E$-waarden (= log2-getransformeerde intensiteiten) van single-channel Affymetrix chips voor 3 biologische samples (EHEC, MRSA en ECOLI). Er werden 4, 5 en 3 replica's gemeten (op 12 verschillende chips). Het data format:

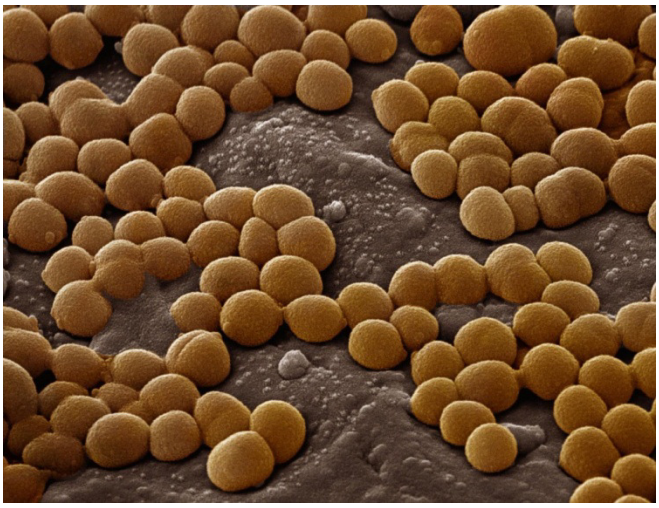
| gene | EHEC.1 | … | EHEC.4 | MRSA.1 | … | MRSA.5 | ECOLI.1 | … | ECOLI.3 |
|---|---|---|---|---|---|---|---|---|---|
| ZZP_04 | 14.40 | … | 14.21 | 12.41 | … | 12.39 | 13.55 | … | 13.62 |
| ZZP_76 | 9.44 | … | 9.50 | 8.88 | … | 9.23 | 9.45 | … | 9.37 |
| … | … | … | … | … | … | … | … | … | … |

**c.** [10 pt] Microarray data file met log-getransformeerde $R$- en $G$-waarden van dual-channel ($R$ = rood, $G$ = groen) chips voor 2 biologische samples ($R$ = met medicijn, $G$ = zonder medicijn). Er werden 3 replica's gemeten (op 3 verschillende chips). Het data format:

| gene | R.1 | G.1 | R.2 | G.2 | R.3 | G.3 |
|---|---|---|---|---|---|---|
| ATX_av | 4.41 | 3.51 | 4.38 | 3.49 | 4.31 | 3.37 |
| ATX_az | 8.52 | 5.19 | 8.47 | 5.33 | 8.39 | 5.27 |
| … | … | … | … | … | … | … |

**d.** [10 pt] Voer de test bij vraag **c.** uit op genen ATX_av en ATX_az. Komen ze differentieel tot expressie? Neem $\alpha = 0.05$.

## Opgave 2. *Bloomococcus* en antibiotica



In een microarray experiment werden de genexpressies van de bacterie *Bloomococcus scoloferi* (de veroorzaker van de ziekte van Bloom) onderzocht onder drie omstandigheden: met het antibioticum learicycline, met het antibioticum doxycyline en zonder antibioticum. Het platform is een single-channel Affymetrix genchip, en de gegevens in de data file `Opgave2.txt` zijn log2-getransformeerde intensiteiten (met juiste background correctie en normalisatie). Per rij één gen, per kolom een (replica van een) biologisch sample (LEAR, DOXY en CONTR). Er zijn 6 replica's met learicycline (LEAR.1 t/m LEAR.6), 6 replica's met doxycycline (DOXY.1 t/m DOXY.6) en 5 replica's zonder antibioticum (CONTR.1 t/m CONTR.5), in totaal 17 verschillende chips. Bij een Affymetrix platform geeft elke chip een kolom in de data file. Elke Affy chip bevatte 2 000 genen. In de laatste kolom van de data file (FUNCTION) staat of elk gen volgens de Gene Ontology betrokken is bij celdeling (= 1) of niet (= 0).

Data format:

gene LEAR.1 … LEAR.6 DOXY.1 … DOXY.6 CONTR.1 … CONTR.5 FUNCTION

De gen namen zijn rownames in de file, de kolommen zijn tab-separated.

**a.** [10 pt] Leg uit welke parametrische analyse geschikt is om te onderzoeken welke van de 2000 genen differentieel tot expressie komen. Neem hierbij aan dat de varianties binnen beide samples vergelijkbaar zijn.

**b.** [10 pt] Hoeveel vals positieve DEG's verwacht je met een significantieniveau van $\alpha = 0.05$?

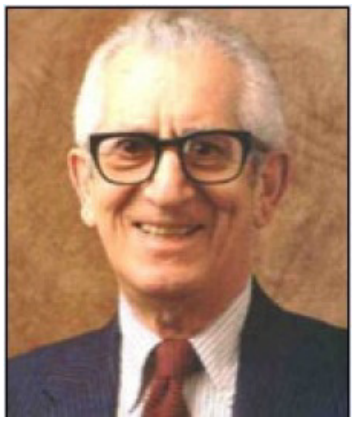
**c.** [10 pt] Onderzoek *hoeveel* genen differentieel tot expressie komen. Neem $\alpha = 0.05$, en gebruik nog geen multiple-testing correctie.

**d.** [20 pt] Pas drie verschillende multiple-testing correctie methodes toe: de Bonferroni methode, de Holm methode en de Benjamini-Hochberg methode. Welke genen zijn volgens elk van de drie methodes DEG's? Neem $\alpha = 0.05$.

**e.** [10 pt] Maak een Venn diagram van de drie verzamelingen DEG's die je bij **d.** hebt gevonden. Hoeveel genen worden door *alledrie* multiple-testing correctie methodes als DEG's bestempeld?

**f.** [10 pt] Bereken of er binnen de DEG's volgens de Holm methode (zie **d.**) sprake is van significante enrichment/depletion voor de functionaliteit "celdeling". Neem $\alpha = 0.05$.

## Opgave 3. Ziekte van Bloom



In de file `Opgave3.txt` staat een verzameling van 20 genen die volgens een statistische analyse differentieel tot expressie zijn gekomen. Voor elk gen (= rij) staan in de file $E$-waarden (= log2-getransformeerde intensiteiten met background correctie en juiste normalisatie) van een Affymetrix platform. Er zijn 3 verschillende biologische samples: drie verschillende stadia van de ziekte van Bloom (StageI, StageII en StageIII). Het aantal replica's per sample is 3, 4 en 3, respectievelijk.

Data format:

gene StageI.1 … StageI.3 StageII.1 … StageII.4 StageIII.1 … StageIII.3

waarbij gene een rowname is; de kolommen zijn tab-separated.

**a.** [10 pt] Voer op de *genen* een hiërarchische clustering uit (Euclidische afstanden, average linkage), en maak een goed dendrogram van het resultaat ($\rightarrow$ pdf).

**b.** [5 pt] Maak 4 subclusters. Uit welke DEG's bestaat elke subcluster? Geef de gen namen.

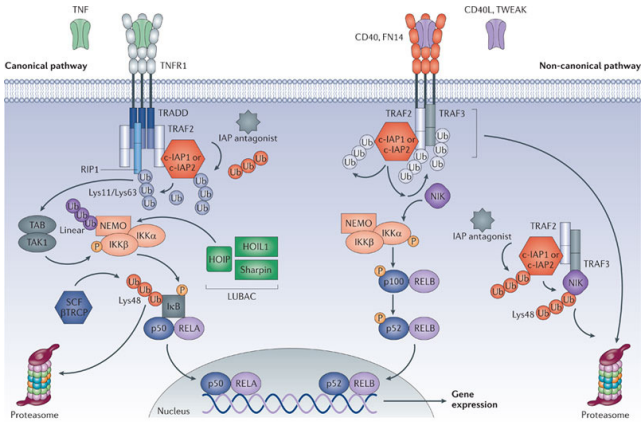
**c.** [10 pt] Voer op de *samples* een *k*-means clustering uit (met 3 clusters), en maak een goed dendrogram van het resultaat ($\rightarrow$ pdf).

**d.** [5 pt] Uit welke samples bestaat elke cluster? Geef de sample namen. Klopt dit ook met de biologische indeling van de samples (Stage I, Stage II en Controle)?

**e.** [10 pt] Maak een goede heatmap van deze data ($\rightarrow$ pdf), o.a. met een legenda, waarbij je zowel de genen als de samples clustert volgens een hiërarchische clustering (Euclidische afstanden, average linkage).

## Opgave 4. Gene Enrichment Analysis

Analyse van een microarray experiment met 20 000 genen resulteert in een cluster van 48 DEG's. Vanuit de Gene Ontology is informatie over de functies van alle 20 000 genen opgevraagd. We zijn o.a. geïnteresseerd of deze cluster van DEG's significant van doen heeft met metabolism en/of signalling.



Nature Reviews | Drug Discovery

**a.** [10 pt] Een gen analyse geeft voor de functionaliteit "metabolism" het volgende resultaat:

| aantal | | functionaliteit: metabolism ja | functionaliteit: metabolism nee | totaal |
|---|---|---|---|---|
| DEG | ja | 24 | 24 | 48 |
| | nee | 6 844 | 13 108 | 19 952 |
| totaal | | 6 868 | 13 132 | 20 000 |

d.w.z., van de 48 DEG's zijn er 24 betrokken bij celdeling, en van alle 20 000 genen op de chip zijn er 6868 betrokken bij celdeling. Is er sprake van een significante enrichment en/of depletion van deze functionaliteit binnen de cluster DEG's? Neem $\alpha = 0.05$. Zo ja, is er dan sprake van enrichment of van depletion ?

**b.** [10 pt] Een analyse voor de functionaliteit "signalling" geeft het volgende resultaat:

| aantal | | functionaliteit: signalling ja | functionaliteit: signalling nee | totaal |
|---|---|---|---|---|
| DEG | ja | 29 | 19 | 48 |
| | nee | 16 172 | 3 780 | 19 952 |
| totaal | | 16 201 | 3 799 | 20 000 |

d.w.z., van de 48 DEG's zijn er 21 betrokken bij ion transport, en van alle 20 000 genen op de chip zijn er 4982 betrokken bij ion transport. Is er sprake van een significante enrichment en/of depletion van deze functionaliteit binnen de cluster DEG's? Neem $\alpha = 0.05$. Zo ja, is er dan sprake van enrichment of van depletion ?