Eindopdracht DAUR2 – V8

Introductie

Hersenziekten, zoals bijvoorbeeld de ziekte van Parkinson, worden veroorzaakt door verstoorde functie van hersencellen. Voor veel hersenziekten geldt dat we niet weten welke moleculaire mechanismen de ziekte veroorzaakt. Het zou daarom mooi zijn als we de hersencellen in het laboratorium kunnen onderzoeken. Het mooiste zou het zijn als we de hersencellen van gezonde individuen kunnen vergelijken met de hersencellen van patiënten.

Omdat het niet mogelijk is om hersenweefsel uit levende personen te halen, zijn er veel methoden ontwikkelt om hersencellen in het laboratorium te maken uit andere cellen. Eén mogelijkheid is om fibroblasten om te zetten naar geïnduceerde pluripotente stamcellen (zie formatieve opdracht) en die stamcellen vervolgens te differentiëren naar hersencellen. Een voordeel van deze route is dat je voor zowel patiënten als gezonde personen hersencellen kunt verkrijgen. Een nadeel is dat deze methode relatief veel tijd kost. Een alternatief is de directe omzetting van fibroblasten naar hersencellen. Dit noemen we **transdifferentiatie**. Hierbij wordt een transcriptiefactor in de fibroblasten tot overexpressie gebracht, waardoor de cellen van identiteit veranderen en om worden gezet in hersencellen.

In deze opdracht maken we gebruik van gepubliceerde data. In deze studie zijn fibroblasten uit één proefpersoon gehaald. Een deel van de fibroblasten is behandeld met een controle transcriptiefactor (BCLXL). Een ander deel van de fibroblasten is behandeld met een transcriptiefactor (ONECUT2) waarvoor wordt vermoed dat het fibroblasten om kan zetten naar hersencellen. De onderzoeksvraag van deze transdifferentiatie studie is: wat zijn de verschillen in genexpressie tussen BCLXL behandelde cellen en ONECUT2 behandelde cellen? Voor beide type fibroblasten zijn daarom RNA-seq datasets gemaakt. Voor elke conditie zijn twee datasets gegenereerd (duplo metingen). De samples in deze studie zijn de volgende:

GEO Sample id	Cell line	Treatment	Run id
GSM3393013	Fibroblast line 1 (CL1500023)	BCLXL	SRR7866699
GSM3393014	Fibroblast line 1 (CL1500023)	BCLXL	SRR7866700
GSM3393017	Fibroblast line 1 (CL1500023)	ONECUT2	SRR7866703
GSM3393018	Fibroblast line 1 (CL1500023)	ONECUT2	SRR7866704

Voor de transdifferentiatie studie gelden de volgende experimentele details:

- De RNA-seq datasets bestaan uit **paired reads**.
- Voor het genereren van de datasets is gebruik gemaakt van een **stranded protocol** (de reads kunnen dus gebruikt worden om te bepalen of de RNA moleculen afkomstig waren van de plus of min strand).

Verder zijn voor de transdifferentiatie studie verschillende databestanden beschikbaar:

- Sample informatie is beschikbaar in de file **/home/daur2/rnaseq/rnaseq_onecut/onecut_sampledata_OC2.csv**
- De fastqQC resultaten zijn al gegenereerd en zijn te vinden **/home/daur2/rnaseq/rnaseq_onecut/fastq_output/**

- De fastq files zijn al gedownload en zijn te vinden in **/home/daur2/rnaseq/rnaseq_onecut/fastq/**
- De alignment is al uitgevoerd. De bam files en alignment statistics zijn te vinden in **/home/daur2/rnaseq/rnaseq_onecut/bam/**
- De count table is al gegenereerd. De count table is te vinden in **/home/daur2/rnaseq/rnaseq_onecut/counts/read_counts_OC2.rds**

Opdrachten

1. Schrijf een korte introductie voor de Rmarkdown met daarin in eigen woorden het doel van jouw onderzoek/analyses.

2. Bekijk de FastQC html rapporten voor de verschillende fastq files. Wat zijn jullie conclusies? Vermeld deze conclusies in jullie Rmarkdown. Neem een of meerdere screenshot(s) op in jullie Rmarkdown om jullie conclusie te onderbouwen.

3. Schrijf code voor het genereren van de count table met de Rsubread package. Jullie hoeven deze code niet uit te voeren, omdat de count table al voor jullie is gemaakt.
   NB: de bam directory bevat meer bam files dan alleen jullie bam files. Zorg dus dat het script daar rekening mee houdt!

4. Gebruik de count table en het csv bestand met sample informatie om een DESeq2 object te maken.

5. Voer een PCA analyse uit. Gebruik de resultaten voor de volgende opdrachten:

   a. Maak een staafdiagram met daarin voor elke PC (PC1 t/m PC4) het *percentage* variatie dat wordt verklaard door die PC. Bijvoorbeeld: stel dat PC1 80% van de variatie verklaard, dan moet er voor PC1 een staaf in de grafiek komen met een hoogte van 80%; hetzelfde voor PC2 t/m PC4.

   b. Maak een grafiek voor PC1 versus PC2. Kleur de punten op basis van de behandeling.

   c. Noteer voor zowel de grafieken bij onderdeel (a) en (b) jullie observaties in jullie Rmarkdown bestand.

6. Voer de DGE analyse uit met behulp van DESeq2. Gebruik de resultaten voor de volgende opdrachten:

   a. Maak een Volcano plot waarin je alle genen met een adjusted p-value < 0.01 and |LFC| > 1 donkeroranje maakt. Geef de LFC en p-waarde thresholds in de plot weer als stippellijnen. Geef ook in de grafiek aan hoeveel genen er upgereguleerd zijn bij deze thresholds en hoeveel genen er downgereguleerd zijn; geef deze aantallen weer als tekst in de grafiek (op de juiste plaats!).

   b. Maak een heatmap met daarin de count values voor de 5 meest upgereguleerde genen en de 5 meest downgereguleerde genen. Selecteer alleen genen met een

adjusted p-waarde < 0.01. Zorg ervoor dat de rijen van de heatmap gelabeld worden met het gensymbool i.p.v. de Entrez identifier.

7. Schrijf een functie die voor gegeven Entrez identifiers (bijvoorbeeld “3175”, “9480” en “390874”) de Ensembl identifier, de Uniprot identifier én het gensymbool kan opzoeken. De functie moet aan de volgende voorwaarden voldoen:

    - De functie moet gebruik maken van de org.Hs.eg.db library.
    - De functie hoeft als output voor elke Entrez identifier slechts één Uniprot identifier, één Ensembl identifier en één gensymbool terug geven.
    - De functie moet overweg kunnen met meerdere Entrez identifiers in de vorm van een vector.
    - De output van de functie is een tibble.

    Bijvoorbeeld:

```
# Run function
entrezConverter(entrezid = c("3175","9480","390874"))
```

```
## # A tibble: 3 x 4
##   entrez symbol  ensembl         uniprot
##   <chr>  <chr>   <chr>           <chr>
## 1 3175   ONECUT1 ENSG00000169856 Q9UBC0
## 2 9480   ONECUT2 ENSG00000119547 O95948
## 3 390874 ONECUT3 ENSG00000205922 O60422
```

8. Voer een GO term enrichment analyse uit voor de upgereguleerde genen (gedefinieerd als adjusted p-value < 0.01 en LFC > 1) en voor de downgereguleerde genen (gedefinieerd als adjusted p-value < 0.01 en LFC < -1). Maak voor beide analyses een grafiek.

    Wat zijn jullie conclusies? Klopt het vermoeden van de onderzoekers dat ONECUT ervoor zorgt dat fibroblasten worden omgezet naar hersencellen? Vermeld jullie conclusie in jullie Rmarkdown.

9. Zorg ervoor dat de Rmarkdown voorzien is van informatieve headers en netjes is opgemaakt (zodanig dat de resulterende html file een mooi rapport is voor jullie resultaten). De Rmarkdown moet uiteraard geknit kunnen worden.