Übung Genexpression in R

1. December, 2008

Protokolle der Lösungen dieser Übung senden Sie an `anne.kupczok@univie.ac.at` bis spätestens **18. Januar 2009**. Erstellte plots können direkt ins Protokoll eingebunden werden oder mit einem bezeichnenden Namen auch als pdf mitgeschickt werden. Die Datensätze für die Übungen finden Sie in `http://www.cibiv.at/~akupczok/pubs/genexpr/`.

Die Übung wird in R und Bioconductor `http://www.bioconductor.org/` durchgeführt. Bioconductor stellt ein Skript bereit, welches das installieren der Pakete erleichtert. Weiterhin gibt es zu jedem Paket Hilfe üeber die Vignetten. Installieren, Laden und Referenz am Beispiel des Paketes `limma` (Hinweis: Jedes Paket muss nur einmal installiert werden, mit `library` kann überprüft werden, ob ein Paket verfügbar ist).

```
source("http://bioconductor.org/biocLite.R") #Laden des Skripts
biocLite("limma") #Installieren des Paketes
library("limma")  #Laden des Paketes
openVignette("limma") #Laden der Referenz
```

1. Erstellen Sie den folgenden Plot: Wahrscheinlichkeit eines seltenen Transkripts nicht detektiert zu werden gegen seine Anzahl in der Zelle, Größe der SAGE-Library: 50000, Totale Anzahl von Transkripten pro Zelle: 300000. Ab welcher Transkriptanzahl wird ein Transkript mit einer Wahrscheinlichkeit von mindestens 95 % mindestens einmal gezogen? (3+1 Punkte)

2. `sage.txt` enthält zwei SAGE-Libraries von dem gleichen Gewebe eines Krebs- und eines gesunden Patienten (Spalten `cancer` und `normal`). Wir wollen Kandidatengene finden, die in beiden Libraries unterschiedlich exprimiert sind. Lesen Sie die Datei in einen `data.frame` ein.

   (a) Entfernen Sie alle Tags, die nur einmal in mindestens einem der Samples gesehen wurden. Wie viele Tags bleiben erhalten? Wie groß ist jetzt die library für `cancer` und `normal`? (2 Punkte)

   (b) Testen Sie für jeden Tag ob die Häufigkeiten dieses Tags in den beiden Libraries unterschiedlich sind mittels des Chi-square-tests. Sortieren Sie den `data.frame` aufsteigend nach den dazugehörigen p-Werten. (Hinweis: Eine Möglichkeit ist den `data.frame` zeilenweise durchlaufen und das Ergebnis jeweils an einen Vektor zu hängen. Die Sortierung des Vektors erhält man mit der Funktion `order`, diese Indizes kann man auf den `data.frame` anwenden; 5 Punkte)

   (c) Schauen Sie bei SageGenie nach, ob es ein Gensymbol und Namen für die fünf Tags mit dem kleinsten P-Wert gibt (Zeile "Best Gene for Tag"). Welcher Tag scheint von Bedeutung zu sein? (3 Punkte)

   (d) Angenommen, Sie bekommen 3 weitere Samples von normalem Gewebe und 5 weitere Samples von Tumorgewebe. Welchen Test würden Sie vorschlagen um Gene zu finden, die zwischen den Geweben unterschiedlich exprimiert sind? Beschreiben sie ein mögliches Vorgehen kurz. (2 Punkte)

3. Microarray-Analyse: Es sollen sechs two-color Microarrays ausgewertet werden. Die Daten stammen von Rattenleberzellen, wobei drei Zelllinien mit einer mutmaßlich krebserregenden Substanz behandelt wurden und drei als Kontrolle dienen. Die Daten wurden vorverarbeitet und stehen als R-Objekt `dataset.robj` bereit. R-Objekte können mit `load` eingelesen werden, daraufhin können die Variablen mit `ls` abgefragt werden.

   (a) In `Mvals` und `Avals` sind die vorverarbeiteten M-Werte und A-Werte bereitgestellt, wobei jede Spalte das Experiment eines Microarrays ist. Überzeugen Sie sich von der Vorverarbeitung indem Sie einen M/A-Plot für das erste Experiment anfertigen. (1 Punkt)

   (b) `design` gibt die Art jedes Microarrayexperimentes an (1 falls die Kontrolle grün und die Transkripte behandelter Zellen rot markiert wurden und -1, falls umgekehrt), diese Information wird benötigt um die Experimente auszuwerten. In `limma` wird dafür ein lineares Modell für jedes Gen gefittet. Dies geschieht mit dem Befehl `fit = lmFit(Mvals, design)`, daraufhin kann man die "moderated T-Statistic" anwenden: `fit=eBayes(fit)`. Die Annotation in der Datei `feature` kann man einfach mit dem gefitteten Datensatz verwenden: `fit$genes=feature`. Eine Auswertung der Gene mit den kleinsten p-Wert findet man mit `topTable(fit)`. Fertigen Sie einen `volcanoplot` der Statistik an, in der die Gennamen der 100 besten Gene erscheinen (Hinweis: Es gibt dafür eine einfache Option in der entsprechenden Funktion in `limma`, die Gennamen stehen in `fit$genes$Name`; 2 Punkte)

   (c) Speichern Sie die Ergebnisse mithilfe von `write.fit` in einer Datei. Fügen Sie die ersten 10 Zeilen in ihr Protokoll ein. (1 Punkt)

   (d) Ordnen Sie `Mvals` absteigend nach den p-Werten der moderated T-Statistik (Hinweis: mit `names(fit)` erhalten sie die Komponenten von `fit`, dort sollten Sie auch die p-Werte finden.) Erstellen Sie ein Heatmap der M-Werte der 100 Gene mit den kleinsten p-Werten. Vergleichen Sie das Bild mit einem Heatmap von 100 beliebigen Genen. (3 Punkte)

   (e) Nennen Sie je ein Paket von Bioconductor, welches den Fisher-Test nutzt um Kategorien auf Überrepräsentation zu testen und welches die Gene Set Enrichment-Methode implementiert hat (2 Punkte).