Wintersemester 2006/2007 Abgabe am 18. Januar 2007
Prof. Dr. Stefan Posch, Dipl.-Bioinf. Jan Grau

12. Übung „Angewandte Bioinformatik mit Perl und R“

1. Auf der Seite zur Übung ist das Paper *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring* von *Golub et al., 1999* verlinkt. Lesen Sie dieses Paper und dabei insbesondere den Ergebnisteil.

2. Die Daten zur dem Paper aus Aufgabe 1 sind in `Bioconductor` im Package `multtest` als `golub` verfügbar.

   (a) Laden Sie das Package `multtest` und die `golub`-Daten und schauen Sie sie sich an.

   (b) Bestimmen Sie aus diesen Daten die differentiell exprimierten Gene nach folgenden Kriterien:

   - Fold-change ≥ 2 und Fold-change ≥ 3,
   - t-Test mit $\alpha = 0.05$ und $\alpha = 0.01$ (`ttest` aus `genefilter`),
   - t-Test mit Bonferroni-Korrektur und $\alpha = 0.05, \alpha = 0.01$ (`mt.rawp2adjp` aus `multtest` mit `t.test` oder `ttest` aus `genefilter` mit entsprechender Korrektur)

   (c) Vergleichen Sie die Ergebnisse der einzelnen Kriterien.

   (d) Bestimmen Sie aus den differentiell exprimierten Genen nach t-Test mit Bonferroni-Korrektur die 5 mit dem kleinsten p-Wert und plotten Sie die Ergebnisse, so dass eventuelle Unterschiede zwischen den Klassen (ALL, AML) deutlich werden.

   (e) Für jedes Kriterium haben Sie eine Auswahl von differentiell exprimierten Genen erhalten. Wenden Sie nun auf diese Auswahl jeweils die anderen Kriterien an und vergleichen Sie die daraus resultierenden Gene mit der ursprünglichen Selektion aus Aufgabe 2 (a). Welche Gene kommen hinzu, welche fallen weg? Versuchen Sie, eine Begründung für dieses Verhalten zu geben.

3. Unter den Genen im betrachteten Datensatz gibt es viele „uninteressante“ und einige „interessante“ Gene. Wir wollen im folgenden einige Kriterien anwenden, um die interessanten Gene vorzuselektieren.

   (a) Selektieren Sie aus allen Genen diejenigen, deren Expressionswerte grundsätzlich über einem Wert von 1.5 liegen. Schreiben Sie dazu eine Funktion, so dass sie später Gene für beliebige Schwellwerte selektieren können.

   (b) Selektieren Sie aus allen Genen diejenigen, für die mindestens 75% der Expressionswerte über einem Wert von 2.5 liegen. Schreiben Sie auch hierfür eine Funktion. (`pOverA` aus `genefilter`)

   (c) Selektieren Sie aus allen Genen diejenigen, deren *interquartile range* größer 1.5 ist. Dies ist ein Kriterium für die Varianz der Daten. Schreiben Sie auch hierfür eine Funktion. (`IQR` aus `genefilter`)

   Wenden Sie nun die Kriterien aus Aufgbe 2 (a) auf die jeweils selektierten Gene an. Gibt es Gene, die vorher als differentiell exprimiert eingestuft wurden, aber durch die Vorselektion ausgeschlossen wurden? Kommen Gene hinzu?

4. Vergleichen Sie die Ergebnisse aus Aufgabe 2 (a) mit denen von *Golub et al.*. Stellen Sie fest, wieviele und welche Gene sowohl bei Ihnen als auch bei Golub als differentiell exprimiert eingestuft wurden. Verwenden Sie das `annotate`-Package (`pm.getabst` mit `basename="hu6800"`), um etwas über diese Gene herauszufinden.