Softwarewerkzeuge der Bioinformatik

Prof. Dr. Volkhard Helms
Dr. Michael Hutter, Andreas Denger,
Marcial Josef Paszkiel, Markus Hollander

Saarland University
Department of Computational Biology

Wintersemester 2022/2023

Tutorial 8

15. Dezember 2022

Genexpression

In diesem Tutorium werden Sie zwei verschiede Methoden zur Microarray-Analyse auf zwei verschiedene Microarray-Datensätze anwenden. Alle Datensätze stammen von Patienten oder Zelllinien mit *akuter lymphatischer Leukämie* (T-ALL), sowohl vor als auch nach der Behandlung mit einem potentiellen Wirkstoff.

Exercise 8.1: Vorbereitung

Normalerweise wird diese Art von Analyse meistens mit den *Bioconductor*-Paketen für die Programmiersprache R durchgeführt. Glücklicherweise stellt der Webserver *CARMAweb* (https://carmaweb.genome.tugraz.at/) ein Frontend für diese Pakete bereit, sodass Sie selbst nichts programmieren müssen.

(a) Rufen sie die Website auf und erstellen sie einen Benutzer-Account. Eine Email-Adresse wird dafür nicht benötigt, nur ein Benutzername und Passwort. Benutzen sie diese anschließend um sich einzuloggen.

(b) Für diese Übung werden wir die Test-Daten benutzen die vom Webserver zur Verfügung gestellt werden. Klicken sie auf den Menüeintrag *Data directory* auf der linken Seite. In diesem Ordner finden Sie einen Knopf mit dem die Testdaten geladen werden können. Diese Dateien können dann für die nächsten Übungen genutzt werden.

Exercise 8.2: Fold change Analyse für zweifarbige Microarrays

Zuerst werden wir die *fold changes* zwischen den roten und grünen Signalen eines zweifarbigen Microarrays berechnen. Die grünen Intensitätswerte stehen für die Expression der Gene einer T-ALL Zelllinie vor der Zugabe eines Wirkstoffs, die roten Signale für die Genexpression die 6 Stunden nach der Zugabe gemessen wurde.

(a) Preprocessing

(1) Klicken sie auf *New Analysis*, und wählen sie dort *Perform a two color microarray analysis* aus. Wählen sie nun die Tabelle mit den Gen-Expressionsdaten aus. Fügen sie die *GenePix* Datei mit dem Namen *Nr026004.gpr* hinzu, und gehen sie zum nächsten Schritt.

(2) CARMAweb hat schon aus dem Dateinamen hergeleitet dass es sich um eine GenePix Datei handelt, und die korrekten Spalten für Rot und Grün ausgewählt. Die Test-Dateien enthalten außerdem eine .GAL Datei, welche die Punkte auf dem Microarray jeweils einem Gennamen und weiteren Annotationen zuordnet. Wählen sie in dem dropdown Menü die Datei *Batch08_modUG.GAL* aus.

(3) Als nächstes kommt das Preprocessing. Wählen sie **normexp** für die *background correction*, **printtiploess** für die *within-array-normalization*, und **quantile normalization** für die *between-array-normalization*. Klicken sie auf anschließend auf *next*.

(4) Das *replicate handling* können wir überspringen, da wir nur einen Array betrachten.

(b) Analyse

(1) Wählen sie **Fold change analysis to define differentially expressed genes** auf der nächsten Seite aus.

(2) Nun ist es an der Zeit, die Log Fold Change (LFC) Werte zu berechnen. Führen sie einen Vergleich (hier: *Comparison*) zwischen den roten und grünen Kanälen des Microarrays durch. Stellen sie sicher dass Red vs. Green ausgewählt ist.

(3) Weiter unten können wir einen LFC threshold angeben, um nur Gene als Resultat zu bekommen die höher oder niedriger als ein bestimmter Wert sind. LFC wird hier als *M (log ratio)* bezeichnet. Wir wollen uns nur Gene angucken die einen LFC-Wert größer als 1,5 oder kleiner als -1,5 haben. Wählen sie außerdem dass ein *MA plot* für diesen Vergleich erstellt werden soll.

(4) Da wir nur einen Vergleich durchführen, müssen wir keine Vergleiche kombinieren. Daher können sie die Analyse nun starten.

(c) Auswertung der Ergebnisse

(1) Öffnen sie die PDF-Datei. Wie viele hoch- bzw. runter-regulierte Gene wurde laut dem LFC-Cutoff von 1,5 bzw. -1,5 gefunden? Interpretieren sie den MA plot auf der letzten Seite.

(2) Die Leukämie-Zellen wurden mit *Glucocorticoiden* (GC), einer Wirkstoffklasse die oft für die Behandlung von ALL benutzt wird, behandelt. Ihren zytotoxischen Effekt erreichen sie durch das Binden an den Glucocorticoid-Rezeptor GR, welcher von dem Gen **NR3C1** kodiert wird. Wurde die Expression von GR durch die Präsenz von GC beeinflusst? Die .txt Datei enthält eine Tabelle mit LFC-Werten für die Gene.

Exercise 8.3: Differenzielle Genexpressions-Analyse für einfarbige Microarrays

Als nächstes werden wir differentiell exprimierte Gene mit einen t-test suchen, dieses mal in *hgu133plus2* Microarrays von Patienten mit T-ALL. Proben wurde vor einer Behandlung mit Glucocoricoiden, sowie 6-8 Stunden nach der Behandlung entnommen.

(a) Preprocessing

(1) Starten sie eine Affymetrix GeneChip analysis und fügen sie die sechs Dateien mit der Endung *.CEL.gz* hinzu.

(2) Die GeneChips sind vom Typ *hgu133plus2*, also wählen wir *conventional 3' array*. Wählen sie *robust multiarray average* (RMA) als Preprocessing-Methode. RMA lässt sich schneller berechnen als die Affymetrix Standard-Methode MAS5. Lassen sie das Programm vor und nach der Normalisierung jeweils ein Histogramm erstellen.

(3) *Replicate handling* wird hier auch nicht benötigt, da wir mehrere Replikate brauchen um einen t-test durchzuführen.

(b) Analyse

(1) Wählen sie **Test statistics to detect differentially expressed genes**.

(2) Definieren sie zwei Gruppen: Die Stichproben ohne Wirkstoff (0h) sind in Gruppe 0, die Stichproben mit Wirkstoff (6h oder 8h) sind in Gruppe 1.

(3) Als nächstes wählen sie die Testmethode. Da es für jeden Patienten zwei Stichproben gibt die zu verschiedenen Zeitpunkten entnommen wurden eignet sich der *paired t-test* am besten. Wählen sie *paired moderated t-statistic (limma)* als Test aus. Diese spezielle Variante des paarweisen t-tests eignet sich besonders für Datensätze mit wenigen Stichproben. Überprüfen sie ob die zwei Stichproben von einem Patienten jeweils als Paar eingetragen sind. Patient 2 ist Paar 1, Patient 20 ist Paar 2, Patient 25 ist Paar 3.

(4) Beim gleichzeitigen Testen von mehreren Hypothesen sollte *multiple testing correction* auf die p-values angewendet werden. Wählen sie hierfür Bonferroni und Benjamini-Hochberg (BH) als Methoden.

(5) Außerdem soll das Programm uns die 100 Gene mit den niedrigsten p-values geben. Um die Spots auf dem Microarray später Genen zuzuordnen, sollten die Ergebnisse mit Gen-identifiern annotiert werden, also klicken sie die entsprechende Option. Lassen sie CARMAweb außerdem noch einen Volcano-Plot der p-values erstellen, hierzu müssen die untersten beiden Optionen gewählt werden.

(c) Auswertung der Ergebnisse

(1) Die Histogramme, die vor und nach der Normalisierung erstellt wurden, sollten unter den Ergebnissen sein, als PDF-Dateien mit dem Namen *analysis....pdf*. Vergleichen sie die Plots miteinander. Hat die Normalisierung gut funktioniert?

(2) Interpretieren sie den Volcano Plot. Wofür stehen die x- und y-achse? Wo würde sich ein signifikant differenziell exprimiertes Gen mit einem hohen Fold Change auf dem Plot befinden?

(3) Öffnen sie die Datei mit den 100 Top-Genen, sortiert nach p-value. Welches Gen hat den höchsten durchschnittlichen LFC (meanM)?

(4) Schauen sie sich nun die p-value dieses Gens an, sowie die zwei korrigierten p-values. Warum ist Fold Change alleine nicht ausreichend um signifikant differenziell exprimierte Gene zu finden? Erklären sie den Unterschied zwischen den p-values die Bonferroni und Benjamini-Hochberg berechnet haben.

(5) Hatte die Behandlung mit GC einen signifikanten Effekt auf die Genexpression in Patienten mit T-ALL, laut dieser Analyse?

Have fun!