Übungen „Angewandte Bioinformatik“
Sommersemester 2024

Jan Grau, Carolin Brune, Gwendolyn Döbel

Abgabe: Montag, 22.05.2024, 23:59 Uhr

Institut für Informatik
Universität Halle

Blatt 4

Im Rahmen dieser Übungsserie bearbeiten wir ein kleines Übungsprojekt. Für dessen Bearbeitung sind zwei Wochen Zeit, in der kommenden Woche gibt es daher keine zusätzliche Übungsserie.

In diesem Übungsprojekt betrachten wir die Auswertung von RNA-seq-Daten von den Rohdaten (FastQ) über die Quantifikation bis zur Erstellung von Plots zu differentiell exprimierten Transkripten.

Die untersuchten Beispieldaten stammen wiederum von Reispflanzen nach Infektion mit *Xanthomonas*-Bakterien, in diesem Fall zwei Stämmen (RS105 und BLS256) von *Xanthomonas oryzae* pv. *oryzicola*, bzw. von einer Kontroll-Behandlung (Mock). Die komplette Experimentserie (mit weiteren Stämmen) ist im Sequence Read Archive des NCBI unter der Accession PRJNA280380 verfügbar.

Auf der Seite zur Veranstaltung im Stud.IP finden Sie eine Datei `sample_table.tsv`, die in der ersten Spalte die Accessions der individuellen Experimente und in der zweiten Spalte die zugehörige Bedingung (BLS256, RS105, Mock) auflistet.

Aufgabe 4.1 (3 Punkte)

Laden Sie mit `fastq-dump` aus den `sra-tools` die Daten zum ersten Mock-Experiment (SRR1952808) als komprimiertes FastQ-File herunter. Der zugehörige Befehl lautet

```
fastq-dump --gzip SRR1952808
```

Diese Datei ist in komprimierter Form ca. 2,3 GB groß. Sollte entweder der Download oder die Speicherung der Datei Ihre technischen Möglichkeiten übersteigen, so ist die Datei auf den Pool-Rechnern im Verzeichnis `/lehre/agprbio/angewandte_bioinformatik/` hinterlegt. Diese und die folgende Aufgabe können Sie auch auf Basis dieser Datei lösen, indem Sie den Tools den absoluten Pfad `/lehre/agprbio/angewandte_bioinformatik/SRR1952808.fastq.gz` angeben.

Nutzen Sie FastQC, um eine Qualitätskontrolle auf dieser Datei vorzunehmen. Geben Sie für jede Kategorie in FastQC (“Basic Statistics”, “Per base sequence quality”, ...) eine kurze Zusammenfassung Ihrer Beobachtungen ab. Eine genauere Beschreibung der Bedeutung der einzelnen Kategorien finden Sie in der FastQC-Hilfe oder unter `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/` → “3 Analysis Modules”.

Aufgabe 4.2 (4 Punkte)

Für dieses erste Mock-Experiment soll nun die Transkript-Quantifikation mittels `kallisto` durchgeführt werden.

(a) Laden Sie das Reis-Transkriptom von

`http://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/all.cdna`

(trotz der ungewöhnlichen Dateiendung eine FastA-Datei) herunter.
Auch diese Datei finden Sie im Zweifelsfall unter `/lehre/agprbio/angewandte_bioinformatik/`.

(b) Erstellen Sie auf dem Reis-Transkriptom einen `kallisto`-Index.

(c) Quantifizieren Sie nun auf Basis dieses Index die Transkript-Abundanzen entsprechend der Daten aus dem ersten Mock-Experiment (`SRR1952808.fastq.gz`). Nehmen Sie dabei eine Fragmentlänge von 200 bp mit einer Standardabweichung von 40 an.

(d) Schreiben Sie ein bash-Skript, das die entsprechenden Befehle in geeigneter Weise kombiniert, so dass die Quantifikation auch für eine beliebige Anzahl von FastQ-Files möglich wäre.

*Geben Sie das bash-Skript aus Teilaufgabe (d) und die Datei **`abundance.tsv`** aus Teilaufgabe (c) bzw. (d) ab. Verpacken Sie beide Dateien in ein **komprimiertes** Archiv.*

Aufgabe 4.3 (7 Punkte)

Da die Analyse aller Datensätze mit `kallisto` hier zu zeitaufwändig und datenintensiv wäre, stellen wir die Quantifikationsergebnisse für alle Experimente auf der Seite zur Veranstaltung im Stud.IP als `kallisto.tgz` bereit. Nutzen Sie diese Daten für die folgenden Aufgaben.

(a) Lesen Sie alle Quantifikations-Ergebnisse (d.h. alle Dateien mit Namen `abundance.tsv` aus dem Archiv) gemeinsam über `tximport` auf Ebene der Transkripte (`txOut=TRUE`) in R ein. Lesen Sie außerdem die Experiment-Tabelle aus der Datei `sample_table.tsv` ein. Stellen Sie dabei sicher, dass die Experimente im `tximport`-Aufruf die gleiche Reihenfolge haben wie in der Experiment-Tabelle. Legen Sie nun auf Basis der importierten Abundanzen und der Experiment-Tabelle ein `DESeqDataSet` an.

(b) Lassen Sie durch `DESeq` die Zählwerte normieren und stellen Sie sicher, dass "Mock" die Basis-Stufe des Faktors "condition" ist.

(c) Vergleichen Sie die Verteilung der normierten Zählwerte in den einzelnen Experimenten auf geeignete Art grafisch miteinander (Histogramme, Density-Plots, Boxplots, ...). Wählen Sie dazu mindestens zwei Plot-Varianten. Es bietet sich dazu an, die Zählwerte zunächst zu logarithmieren ($\log(x+1)$) oder die rlog-Transformation anzuwenden. Halten Sie die Verteilung der Zählwerte für ähnlich?

(d) Vergleichen Sie außerdem mit zwei unterschiedlichen Plot-Varianten (Scatter-Plot, pairs, hclust) die nach Transkripten gepaarten (und logarithmierten) Zählwerte zwischen den Experimenten. Interpretieren Sie die so erstellten Plots.

Falls Sie `pairs` *benutzen möchten, empfiehlt es sich aufgrund der großen Anzahl Datenpunkte direkt in eine PNG-Datei zu plotten.*

Reichen Sie Ihr R-Skript für alle Teilaufgaben und die generierten Plots als komprimiertes Archiv ein.

Aufgabe 4.4 (6 Punkte)

(a) Bestimmen Sie korrigierte p-Werte und log2-fold-changes nach der Infektion mit jeweils einem *Xanthomonas*-Stamm relativ zur Kontrolle, also zwischen den Bedingungen BLS256 vs. Mock bzw. RS105 vs. Mock. Speichern Sie die jeweilige Ergebnis-Tabelle in einer eigenen Variablen.

Da wir anders als im Vorlesungsbeispiel hier drei unterschiedliche Bedingungen haben, müssen wir die gewünschten Vergleiche explizit spezifizieren. Dies erfolgt über den Parameter `name` *von* `results()`*. Die möglichen Werte erhalten Sie von der Funktion* `resultsNames()`*. Wenn Sie alle Daten (Spaltennamen etc.) der bereitgestellten Dateien unverändert gelassen haben, sollten dies "condition_BLS256_vs_Mock" bzw. "condition_RS105_vs_Mock" sein.*

(b) Bestimmen Sie jeweils (für BLS256 bzw. RS105) die IDs der differentiell exprimierten Gene nach folgenden Kriterien: korrigierter p-Wert kleiner 0.01 und log2-fold-change größer als 4 (also mindestens 16-fach hochreguliert). Nutzen Sie diese IDs, um jeweils eine Heatmap über die Expressionswerte (logarithmierte oder rlog-transformierte Zählwerte) aller Experimente zu plotten. Was beobachten Sie? Zwischen welchen Experiment-Gruppen sehen Sie jeweils deutliche Unterschiede?

(c) Vergleichen Sie die Mengen der differentiell exprimierten Transkripte für BLS256 bzw. RS105 über Venn-Diagramme oder Upset-Plots. Wie viele differentiell exprimierte Transkripte haben die Infektionen mit den beiden Stämmen gemeinsam? Wie viele sind jeweils spezifisch?

Reichen Sie Ihr R-Skript für alle Teilaufgaben und die generierten Plots als komprimiertes Archiv ein.