Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, WS 2021/22
Dr. Roland Wittler · M. Sc. Tizian Schulz
http://gi.cebitec.uni-bielefeld.de/teaching/2021winter/sequaprak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 09 vom 14./15.12.2021
Abgabe bis Sonntag bzw. Montag, 24:00 Uhr.

Die Firma GrowBoost möchte herausfinden, warum sich ihr neu entwickelter Dünger „GrowBlast 2.0“ auf das Wachstum der Zuckerrübe positiv auswirkt. Dazu lässt sie eine Population von Zuckerrüben mit Zugabe eines gebräuchlichen Düngers und eine Population mit Zugabe des neuen Düngers wachsen. Nach zwei Monaten werden Blätter von beiden Populationen geerntet, um eine differentielle Expressionsanalyse des *Minichloroplastens*[1] durchzuführen. Dazu werden jeweils drei technische Replikate hergestellt. Die Sequenzierung erfolgt mit einer Illumina-Sequenziermaschine, es werden einfache Reads mit einer Länge von 150 bp generiert.
Du findest die Reads der sechs Sequenzierungen in fastq-Dateien und die Referenzsequenz als Genbank- und als fasta-Datei im Ordner `/prj/seqan/Praktikum/DiffExp`.

Aufgabe 1 (Mapping)
Zuerst müssen die Reads der einzelnen Sequenzierungen auf das Referenzgenom gemappt werden. Dazu soll wieder `Bowtie 2` verwendet werden. Zur Erinnerung: Das Tool ist auf dem CeBiTec-System unter /vol/biotools/lib/bowtie2-2.2.7 installiert. Wie alle rechenintensive Befehle, sollte auch Bowtie nur auf einem Compute Cluster, also von einem `qxterm` aus aufgerufen werden.

1. Kopiere dir die oben beschriebenen fastq-Dateien und die beiden Referenzdateien in dein Heimatverzeichnis.
2. Erstelle einen Index der Referenzsequenz *chloroplast.fasta* mit `bowtie2-build`. Wie genau lautet dein Aufruf?
3. Mappe nun nacheinander die Reads der sechs Sequenzierungen gegen die von `Bowtie 2` erstellte Referenz und speichere die Ergebnisse in sechs verschiedenen sam-Dateien ab. Welchen Aufruf hast du hier verwendet und was bedeuten die Parameter?
4. Wie viele Reads konnten in etwa pro Durchlauf nicht gemappt werden? Woran könnte das liegen? Wie könnte man einen Teil der ungemappten Reads doch noch mappen?

Aufgabe 2 (Differenzielle Expressionsanalyse mit ReadXplorer)
Benutze `ReadXplorer`, um eine differenzielle Expressionsanalyse durchzuführen. Das Tool liegt auf dem CeBiTec-System unter /vol/readxplorer/bin/readxplorer. Es muss von einem `lqxterm` aus aufgerufen werden und benötigt `Java 8`. Ein ausführliches Manual findest du unter: `https://www.uni-giessen.de/fbz/fb08/Inst/bioinformatik/software/ReadXplorer/documentation/userManual`.

1. Erstelle zuerst eine neue Datenbank. Importiere dann die GenBank-Datei *chloroplast.gb* als Referenz und anschließend deine sechs Mapping-Dateien als einfache *Tracks*.
2. Öffne nun die Referenz und die sechs verschiedenen Tracks und mache dich mit der Anzeige vertraut. Was bedeuten die grünen und gelben Bereiche im Track Viewer?
3. Warum findest du Bereiche innerhalb von Genen, in die kein Read mappt?
4. Um eine differentielle Expressionsanalyse durchführen zu können, muss der `ReadXplorer` zunächst mit der lokalen Rserve-Instanz des CeBiTec verknüpft werden. Wie das funktioniert, ist unter `https://www.cebitec.uni-bielefeld.de/intranet/de/spezielle-anwendungen/readxplorer` dokumentiert. Die Konfiguration ist bei jedem Start von `ReadXplorer` zu wiederholen. Zumindest das Passwort muss neu eingegeben werden. Nach der Konfiguration erscheint ein Dialog, der nach einem Masterpasswort verlangt. Dieses ist jedoch nicht erforderlich. Der Dialog kann über die Schaltfläche *Cancel* geschlossen werden.

[1] Für diesen Zettel wurde das Chloroplastengenom verkleinert und verändert.

Führe nun eine differentielle Expressionsanalyse mit `DESeq` für unsere zwei Konditionen durch. Betrachte dabei nur die Genbereiche. Wähle zudem aus, dass die Reads von beiden Strängen kombiniert werden sollen. Welche Gene zeigen eine geänderte Expression bei einem kleinen p-Value (< 0.05)? Wie werden diese Gene im Experiment mit dem neuen Dünger im Vergleich zu dem Experiment mit dem gebräuchlichen Dünger exprimiert?

5. Erstelle mit `ReadXplorer` eine Grafik, die die logarithmierte Expressionsveränderung *(Log2 fold change)* der durchschnittlichen Basenabdeckung eines Gens *(base means)* gegenüber stellt. Bilde die Grafik in deinem Protokoll ab. Welche Gene werden von den roten Punkten unten und oben in der Grafik dargestellt?

6. Welcher Schritt muss nun theoretisch noch erfolgen, um herausfinden zu können, warum der neue Dünger das Wachstum der Zuckerrübe positiv beeinflusst?