Übungen zur Vorlesung Algorithmische Bioinformatik

Freie Universität Berlin, WS 2014/15
Martin Vingron · Juliane Perner · Annkatrin Bressin

Blatt 7 · Ausgabe am 24.11.2014
Abgabe am 01.12.2014 vor Beginn der Vorlesung

Name: Matrikelnummer: Übungsgruppe:

Aufgabe 1 (*10 Punkte; Theorie*)**.** Welche Verteilung würden Sie zur Modellierung folgender Messungen verwenden? Begründen Sie ihre Aussage.

1. Durchmesser von Hefezellen in einer Population
2. Die Anzahl an Kopf-Würfen bei 100 Münzwürfen.
3. Blitzeinschläge pro km^2
4. Die besten Sprintzeiten der Sportler bei einem Wettkampf

Aufgabe 2 (*20 Punkte; Theorie*)**.** Beschreiben Sie den Gibbs Sampler Algorithmus, welcher zur Motivsuche in DNA Sequenzen verwendet wird. Beschreiben Sie außerdem warum der Gibbs Sampler geeigneter ist als ein Greedy-Algorithmus.

Aufgabe 3 (*30 Punkte; Theorie*)**.** In der Vorlesung wurde die Burrows-Wheeler Transformation vorgestellt.

1. Beschreiben Sie kurz in ihren eigenen Worten das Prinzip der Burrows-Wheeler Transformation. Was ist die Anwendung der Transformation?
2. Führen Sie die Burrows-Wheeler Transformation auf folgendem String aus:

   ACCCGTGAA$

3. Berechnen Sie den Originalstring der folgenden Transformation:

1.Spalte:	$AAACCGGTTT
Letzte Spalte:	T$TGAACCTAG

Aufgabe 4 (*40 Punkte; Praxis*)**.** Wir möchten die auf der Vorlesungsseite verlinkte FastQ-Datei[1] analysieren. Sie enhält die bei der Sequenzierung eines menschlichen Chromosom 21 detektierten Reads.
Benutzen Sie für Ihre Analyse den GALAXY-Server[2], der verschiedene Programme zur einfachen Verarbeitung von Sequenzierdaten bereitstellt (*NGS toolbox*).
Beachten Sie folgende Hinweise: Je nach Auslastung des Servers kann die Analyse ggf. länger dauern. Außerdem müssen Sie sich um die Aufgabe vollständig bearbeiten zu können als User registrieren.

1. Laden Sie die FastQ-Datei (Typ *fastqillumina* und Genom *hg19*) auf den Server und schauen Sie sich zuerst die Qualität der sequenzierten Reads an. Nutzen Sie dazu die FASTQC-Option (unter *QC and manipulation*). Fassen Sie kurz die Statistiken zu den Reads in ihrer Library (z.B. Länge, Sequenzkomposition, etc.) zusammen. Was fällt Ihnen an den positionsspezifischen Quality-scores auf?

2. Mappen Sie die Reads mit *Bowtie for Illumina* auf das humane Genom (*hg19*). Nutzen Sie dazu den 'build-in'-Genomindex, erlauben Sie maximal zwei Alignmentfehler und unterdrücken Sie Reads, falls diese nicht eindeutig gemapped werden können. Geben Sie die Optionen, die Sie verändert haben, an.

3. Analysieren Sie die Mapping-Statistiken mit *SAM tools: flagstat*. Werden Reads während des Mappings gefiltert? Wenn ja, wodurch?

4. Berechnen Sie mit Hilfe von *Bedtools: Create a Bedgraph* die Genom-weite Coverage. Laden Sie diese als Custom-Track in den UCSC Genome Browser. Was fällt Ihnen anhand der Coverage auf, an welchen Elementen im Genom ist die Coverage besonders hoch?

5. Wie Sie im Genome Browser sehen können, gibt es einen bekannten SNP an der Position *chr21:30255074*. Ist diese Region durch unsere Library abgedeckt?

6. Schauen Sie, ob der SNP auch in unserer Library detektiert wurde. Generieren Sie dazu eine Zusammenfassung mit *SAM tools: pileup*, bei der für jede Position das Nukleotid im Referenzgenom und das Konsensus-Nukleotid in den Reads angegeben wird.

[1] Material 1: `https://ws.molgen.mpg.de/ws/318395/test_21.fastq`
[2] https://usegalaxy.org/