Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, SoSe 2015
Dr. Roland Wittler · M.Sc. Linda Sundermann
http://wiki.techfak.uni-bielefeld.de/gi/Teaching/2015summer/SequaPrak
praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 2 vom 20.04.2015
Abgabe bis Donnerstag, 24:00 Uhr.

Aufgabe 1 (Kompression mit `bzip2`)

Erstelle 6 Textdateien mit je 100.000 Zeichen, die folgendermaßen aufgebaut sein sollen:

1. alle Zeichen sind identisch,
2. Zufallszeichen aus einem 4-Buchstabenalphabet,
3. DNA-Sequenz,
4. Protein-Sequenzen,
5. natürlichsprachlicher deutscher Text,
6. natürlichsprachlicher Text einer anderen Sprache deiner Wahl.

Komprimiere diese Dateien mit `bzip2`. Beschreibe in deinem Protokoll kurz, wie du die Dateien erstellt hast. Kontrolliere, ob alle sechs Dateien vor der Komprimierung die gleiche Größe haben. Vergleiche dann deine Beobachtungen in einer Tabelle und versuche sie zu erklären.

Aufgabe 2 (Bowtie 2)

Installiere auf deinem System die aktuelle Version von Bowtie 2, die du hier findest: `http://bowtie-bio.sourceforge.net/bowtie2`. Verwende entweder eines der vorkompilierten Pakete oder übersetze die Quellversion.

1. Verschaffe dir im MANUAL einen Überblick über das Tool. Du musst nicht alles im Detail lesen.
2. Im Ordner examples/ findest du einige Beispieldateien. Führe die nachfolgenden Schritte mit der Referenzsequenz *lambda_virus.fa* aus.
   (a) Erstelle den Index für die Referenz mit dem Programm *bowtie2-build*.
   (b) Aligniere die Reads in der Datei *reads_1.fq* als ungepaarte Reads zur Referenzsequenz mit dem Programm *bowtie2* und speichere die Ergebnisse im SAM-Format ab (du musst hier nicht wissen, wie genau dieses Format aufgebaut ist). Schaue dir das Ergebnis an. Bowtie gibt zusätzlich eine kleine Statistik auf der Konsole aus. Wie viele Reads konnten erfolgreich aligniert werden?
   (c) Als nächstes kommt ein Beispiel für das paired-end Alignment. Die Readpaare sind aufgeteilt in die Dateien *reads_1.fq* und *reads_2.fq*. Rufe *bowtie2* mit beiden Dateien auf und speichere das Ergebnis wieder im SAM-Format. Vergleiche kurz die Statistik zum vorherigen Aufruf.
   (d) Als letztes wollen wir Variationen in den Reads im Vergleich zur Referenzsequenz finden (SNPs, kurze Insertionen und Deletionen). Dazu brauchst du auch Zugriff auf samtools und bcftools, die du hier findest: `http://sourceforge.net/projects/samtools`.
   Nutze *samtools view* und *samtools sort*, um das Ergebnis deines paired-end Alignments in ein sortiertes BAM-file zu konvertieren (BAM ist die komprimierte Version von SAM).
   Um die Varianten zu finden, führe *samtools mpileup -uf <referenz>.fa <sortiertes bam> | bcftools call -vc > <ausgabename>.raw.bcf* aus.
   Du kannst dir das Ergebnis mit *bcftools view* anschauen. Wie viele Varianten enthält deine Ausgabe?

Beschreibe in deinem Protokoll die einzelnen Programmaufrufe und was sie bewirken. Du musst deine Ergebnisdateien nicht ins Protokoll schreiben, solltest aber alle Fragen beantworten.