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 13 vom 13.07.2015
Abgabe bis Donnerstag, 24:00 Uhr.

Aufgabe 1 (IGV)

1. Lade die Datei `http://www.cebitec.uni-bielefeld.de/~roland/IGV.zip` herunter, extrahiere sie und informiere dich in der README-Datei, wie IGV mittels `java -jar` gestartet werden kann.

2. Im Universitätsklinikum Düsseldorf wurden drei Proben eines Patienten mit akuter lymphatischer Leukämie sequenziert:

   initial: vor der Behandlung,

   remission: nach der Behandlung (Remission = dauerhafte Nachlassen von Symptomen) und

   relapse: nach einem Rückfall.

   Am Institut für Medizinische Informatik der Universität Münster wurden diese Daten mittels `BWA` auf das Referenzgenom *hg19* gemapped. Im Ordner `data` ist ein Ausschnitt der Daten bereitgestellt (Chromosom 6: 43.382.800–43.389.200). Lade die Daten (`*.bam`) in IGV ein, wähle das entsprechende Referenzgenom aus und lass die entsprechende Region anzeigen.

3. Die drei Proben wurden leicht unterschiedlichen Fragmentlängen um etwa 300 bp (und variierenden Standardabweichungen) sequenziert:

   initial: 261 (45,5)

   remission: 296 (36,83)

   relapse: 305 (37.83)

   Stelle die erwartete Fragmentlänge für alle drei Datensätze so ein, dass Paired-end-Mappings, deren Länge um mehr als die dreifache Standardabweichung vom Mittelwert abweichen, farblich hervorgehoben werden. (Achtung: Man muss nicht nur den Grenzwert manuell einstellen, sondern auch die automatische Berechnung, die aufgrund des kleinen Ausschnitts der Daten nicht anwendbar ist, deaktivieren.) Stelle außerdem ein, dass die gemappten Reads als Paare angezeigt werden.

4. Erstelle ein Screenshot und füge es in dein Protokoll ein – als **einzige** Antwort auf **alle** bisherigen Fragen.

Aufgabe 2 (Strukturelle Variationen in IGV)

1. Welche Art von struktureller Variation ist in der dargestellten Region zu erkennen? Mache dir klar, inwiefern die *Coverage* und die Mappingdistanzen Hinweise auf die Variation geben. (Es muss nur die erste Frage im Protokoll beantwortet werden. Erläuterungen sind nicht notwendig.)

2. Im Datensatz "initial" scheinen auf den ersten Blick ausschließlich erwartete Mappingdistanzen aufzutreten. Sei bei den Filtereinstellungen der Mappingdistanzen/Fragmentlängen für diesen Datensatz nun etwas restriktiver, so dass die Variation nun auch im Datensatz "initial" aufgrund der Mappings zu erkennen ist.

3. Reads, welche im Randbereich der Variation gemapped wurden, konnten teilweise nicht vollständig aligniert werden, sondern wurden mittels *soft clipping* gekürzt. Dies im sog. *Cigar-String* z.B. wie folgt dargestellt. (Diesen String findet man auch in den Mapping-Informationen, die bei IGV mittles *Mouse over* angezeigt werden.)

   `51M` bedeutet: alle 51 Basen wurden gematched,
   `6S40M1D6M` bedeutet: 6 Basen *soft clipping*, 40 Matches, eine Base gelöscht, 6 Matches

   Im obigen Beispiel scheint der zweite Read also sechs Basen in die Variation "hereinzuragen". Finde in einem der Datensätze ein *geclipptes* Readmapping, welches auf die genauen Grenzen der Variation hinweisen könnte.

4. Als **einzige** Antwort auf die Fragen 2–4, füge dem Protokoll ein Screenshot hinzu, auf dem sowohl die nun hervorgehobenen Mappings im Datensatz "initial", als auch die Mapping-Informationen des gefundenen *soft-clipped* Reads zu erkennen sind.