

Übungen zum Sequenzanalyse-Praktikum

Universität Bielefeld, WS 2019/20

Dr. Roland Wittler · M.Sc. Tizian Schulz

<http://gi.cebitec.uni-bielefeld.de/teaching/2019winter/sequaparak>

praktikum-seqan@CeBiTec.Uni-Bielefeld.DE

Übungsblatt 11 vom 10./11.12.2019

Abgabe bis Sonntag bzw. Montag, 24:00 Uhr.

Aufgabe 1 (IGV)

1. Kopiere die Datei `/prj/seqan/Praktikum/IGV.zip` in dein Home-Verzeichnis, 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 mit 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 die Grenzwerte 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 und dass die Mappings eng zusammen gezeigt werden sollen (*collapsed*).

4. Beschreibe kurz den allgemeinen Aufbau eines Tracks. Welche Elemente sind zu erkennen? Wie sind sie dargestellt?

Aufgabe 2 (Strukturelle Variationen in IGV)

1. Was bedeuten blau und was rot markierte Mappings? Auf was für strukturelle Variationen können sie hinweisen und warum?
2. Wie unterscheiden sich die Mappings der drei Tracks voneinander? Was ist ähnlich (auch in Bezug auf die Coverage)?
3. Welche Art von struktureller Variation ist in der dargestellten Region zu erkennen? Erkläre, inwiefern die *Coverage* die Hinweise durch die Mappingdistanzen unterstützt.
4. 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. Welche Werte hast du als minimale und als maximale Grenze gewählt, und wie viele rot markierte Mappings findest du nun im Bereich der strukturellen Variation?
5. Reads, welche im Randbereich der Variation gemapped wurden, konnten teilweise nicht vollständig aligniert werden, sondern wurden mittels *soft clipping* gekürzt. Dies wird im sog. *Cigar-String* z.B. wie folgt dargestellt. (Diesen String findet man auch in den Mapping-Informationen, die bei IGV mittels *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 im Datensatz „remission“ ein *geclipptes* Readmapping, welches auf die genauen Grenzen der Variation hinweisen könnte. Gib den *CIGAR-String* und den Mappingstart an. Wie viele Basen mussten „geklippt“ werden? Sagt der Read etwas über den linken oder rechten Rand der Variation aus? Wo genau liegt diesem Mapping zufolge der rechte bzw. linke Rand der Variation?