Sara Kemmler 5760949
Robin Bonkaß 5769588

1	2	3	4	Σ

Übungsblatt Nr. 07

(Abgabetermin 23.06.22)

Aufgabe 1

Tool: Long- und ultra-long-read Nanopore Technologie
Genom: Australischer Lungenfisch (Neoceratodus forsteri) Meyer u. a. 2021

a)

Besonderheiten des Genoms:

1. Schnittstelle (Theoretisch sind die Merkmale für das Leben an Land, wie z.B. Füße im Genom angelegt, der Lungenfisch lebt aber im Wasser.
2. Lebende Fossilien, da die Morphologie sich in Millionen von Jahren kaum geändert hat.

b)

Assembly approach: Batches mit MARVEL assembler zusammengeführt.

c)

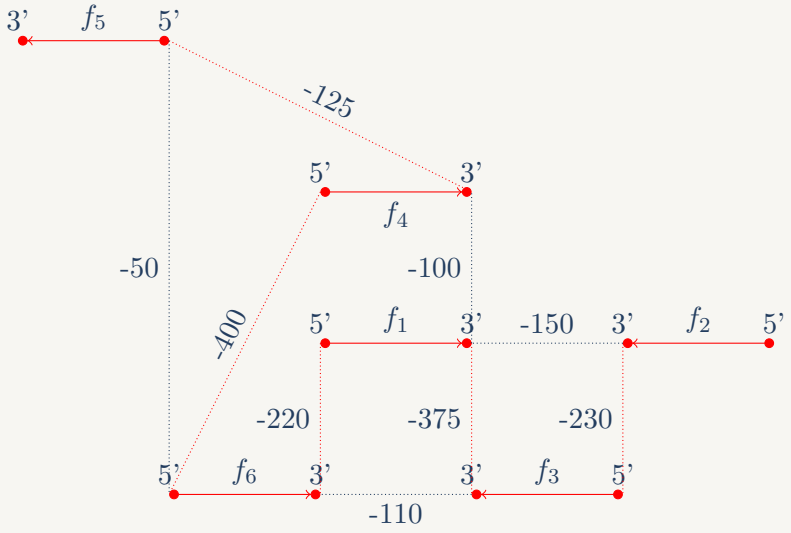
Andere eukaryotische Genome, die mit diesem Tool assembliert wurden: Axolotl Salamander (Ambystoma mexicanum)

3

Aufgabe 2

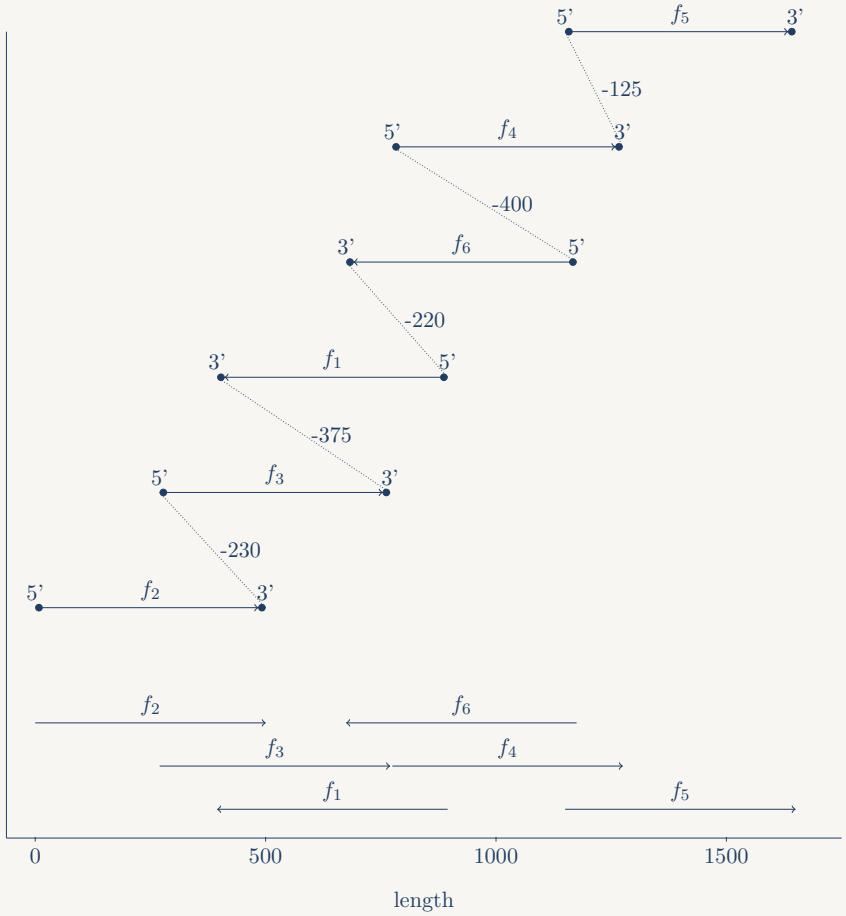
a) und b)

Overlap Graph und Minimaler Spannbaum (in rot):



c)

Layout der Reads und consensus sequence:



Die Länge der Assemblierung ergibt sich aus:

$$\text{Länge aller reads} - \text{Länge aller overlaps} = 6 \cdot 500 - 1350 = 3000 - 1350 = 1650$$

d)

Die Overlaps, welche in dem Layout dargestellt sind, sind alle konsistent, da sie Teil des minimalen Spannbaums waren. Drei Overlaps hingegen, sind nicht in dem minimalen Spannbaum, weshalb diese nicht unbedingt konsistent sein müssen:

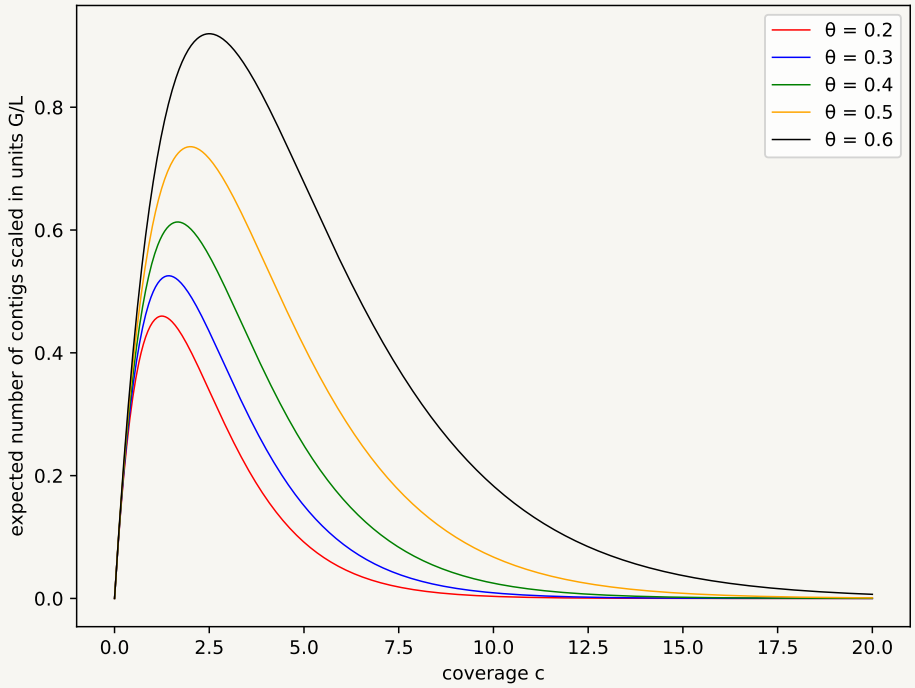
$$\begin{aligned}
&\text{Overlap } 5'f_5 - 5'f_6 = 25 \text{ er sollte aber } 50 \text{ sein} \\
&\text{Overlap } 3'f_6 - 3'f_3 = 95 \text{ er sollte aber } 100 \text{ sein} \\
&\text{Overlap } 3'f_1 - 3'f_2 = 105 \text{ er sollte aber } 150 \text{ sein}
\end{aligned}$$

Alle Overlaps, welche nicht in dem minimalen Spannbaum sind, sind somit nicht konsistent. Dies kann daran liegen, dass sich Teile der Sequenz wiederholen und somit bestimmte Teile an verschiedenen Stellen überlappen.

5

Aufgabe 3

Folgender Graph wurde mit dem Code in der Datei `Robin_Bonkass_Sara_Kemmler_A7.py` generiert. Dazu den Befehl `python Robin_Bonkass_Sara_Kemmler_A7.py` eingeben. Der generierte Plot wird zudem in der Datei `Robin_Bonkass_Sara_Kemmler_A7_plot.pdf` gespeichert.



Bei einem kleinen θ, muss ein kleinerer Teil überlappen, sodass die reads zusammengefasst werden. Deshalb ist die erwartete Abdeckung c bei kleinen θ ebenfalls geringer. Zudem ist auch die erwartete Anzahl an contigs geringer. Je größer der erforderliche überlappende Teil ist, desto größer ist die erwartete coverage und die erwartete Anzahl an contigs.

5

Aufgabe 4

a)

Paired-end Sequencing: Hierbei werden beide Enden eines Fragments sequenziert, wodurch hochwertige und alignierbare Sequenzdaten entstehen. Durch die paired-end Sequenzierung kann der Nachweis von genomischen Umlagerungen, repetitiven Sequenzelementen, sowie Genfusionen und neuartigen Transkripten ermöglicht werden. Zudem kann eine genauere Leserichtung und die Erkennung von Insertion-Deletion-Varianten ermöglicht werden.

Single-end Sequencing: Hierbei wird nur ein Ende des Fragments sequenziert. Es werden hierdurch große Mengen an qualitativ hochwertigen Daten generiert. Wobei es kostengünstig und schnell abläuft. Im Gegensatz zu paired-end kann single-end keine genaue Leserichtung oder die Erkennung von Insertion-Deletion-Varianten ermöglichen.

b) und c)

Der HTML Report, welcher durch MultiQC generiert wurde, befindet sich in der Datei `Robin_Bonkass_Sara_Kemmler_A7_multiqc_report.html`

d)

Länge des Input Genoms: 3268203 bp (NCBI accession code: NC_002677)

Mean Coverage:

$$c = \frac{p \cdot R}{G}$$

Wobei p der Anzahl an Reads entspricht, R entspricht der durchschnittlichen Länge der Reads, wobei dies mal 2 gerechnet wird, da es sich um reads von paired-end sequencing handelt und G entspricht der Größe des Genoms.

Berechnungen:

$$\begin{aligned}
c_{run1-R1} &= \frac{108940 \cdot 2 \cdot 150}{3268203} \\
&= 9,99 \approx 10 \\
c_{run1-R2} &= \frac{108940 \cdot 2 \cdot 150}{3268203} \\
&= 9,99 \approx 10 \\
c_{run2-R1} &= \frac{217880 \cdot 2 \cdot 150}{3268203} \\
&= 19,99 \approx 20 \\
c_{run2-R2} &= \frac{217880 \cdot 2 \cdot 150}{3268203} \\
&= 19,99 \approx 20 \\
c_{run3-R1} &= \frac{326820 \cdot 2 \cdot 150}{3268203} \\
&= 29,99 \approx 30 \\
c_{run3-R1} &= \frac{326820 \cdot 2 \cdot 150}{3268203} \\
&= 29,99 \approx 30
\end{aligned}$$

e)

In Abbildung 2 ist ein schlechter durchschnittlicher Qualitätsscore der Reads von run 3 zu sehen, da sie im roten Bereich liegen. Daraus lässt sich auf eine schlechte Qualität der Reads des runs schließen. Der durchschnittliche Qualitätsscore der Reads der ersten beiden runs liegen im grünen Bereich, weshalb sie eine gute Qualität aufweisen.
Die in Abbildung 2 beobachtete Qualität kann in Abbildung 1 an jeder Basenposition der Reads ebenfalls beobachtet werden.
Durch die gute Qualität der Reads der ersten beiden runs ist ein sehr genaues Ergebnis des Assemblies zu erwarten. Durch die oben begründete schlechte Qualität des runs 3 wird hier ein Ergebnis des Assemblies erwartet, welches anzuzweifeln ist.

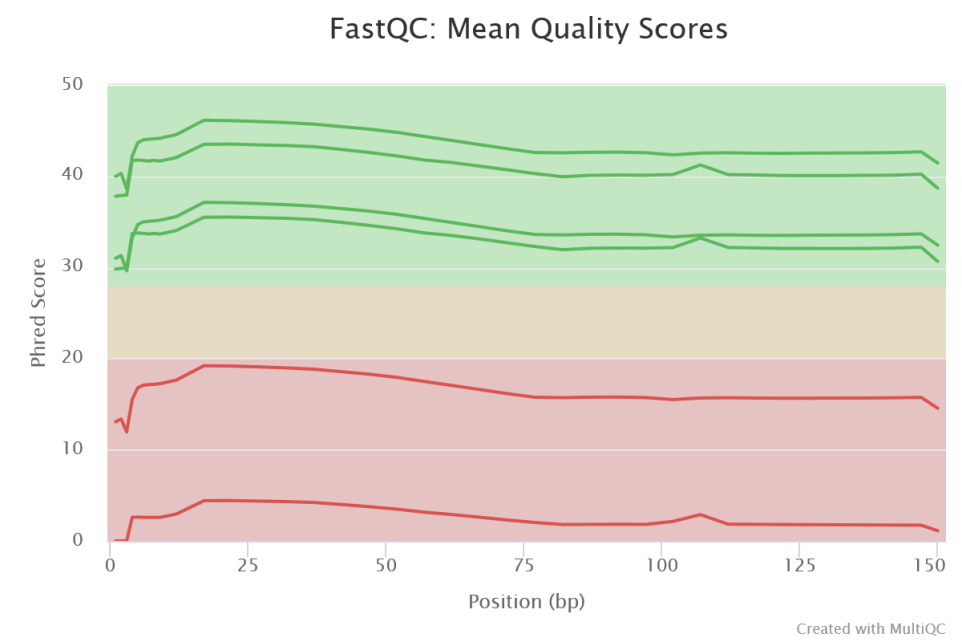


Abbildung 1: Mean Quality Scores

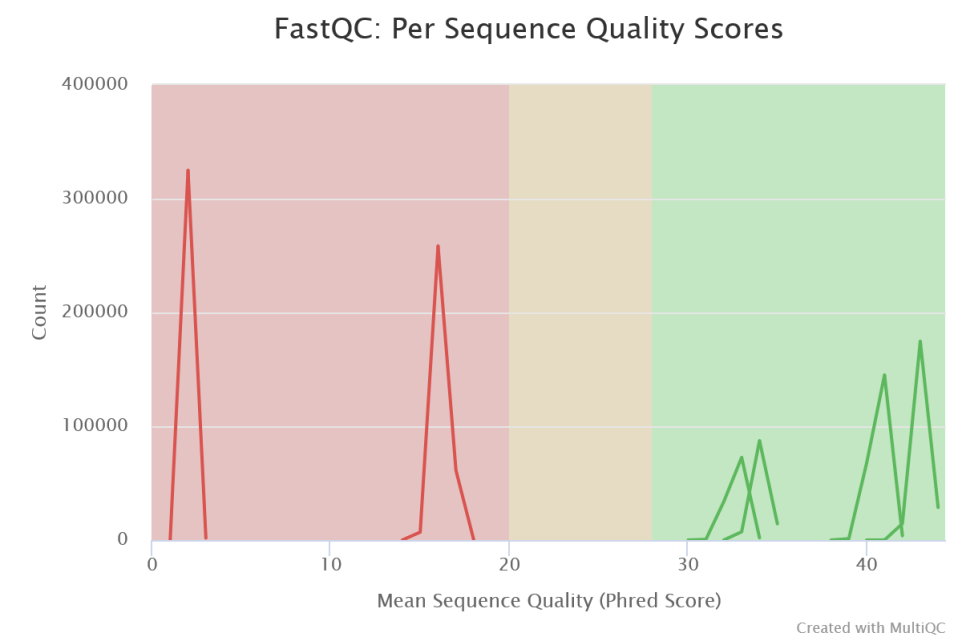


Abbildung 2: Per Sequence Quality Scores

7

Literatur

[Mey+21] Axel Meyer u. a. “Giant lungfish genome elucidates the conquest of land by vertebrates”. In: nature (2021).