Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

1	2	3	4	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Blatt 7

(Abgabe am 23.06.2022)

Theoretical Assignments

Task 1: Assembly of the largest eukaryotic genome (3)

Identify the tool used for the reconstruction of the largest sequenced eukaryotic genome until date. Do not forget to correctly cite the tool! Furthermore, answer the following questions:

The largest sequenced eukaryotic genome that was reconstructed belongs to the *giant lungfish* [6]. It was assembled using the MARVEL assembler [9].

a) **Besides its size, what is the most striking feature of this genome?**
The most striking feature of this genome is that is still expanding in size [6]. Additionally (because the previous aspect still relates to size), it bears some similarities to the human genome, including the genes that control embryonic development of lungs [4].

b) **What is the assembly approach followed by the tool used to reconstruct this genome?**
According to another paper detailing the genome of the axolotl gene [7], the assembly was accomplished with an algorithm with two phases, the first phase keeping long PacBio-reads, while the second one correcting those reads with sequences obtained with Illumina, increasing accuracy and contingency.The corretion of short-read based error is done with Pilon. The scaffold was obtained using "de novo optical maps using the Bionano Saphyr system" (Figure1) [7].

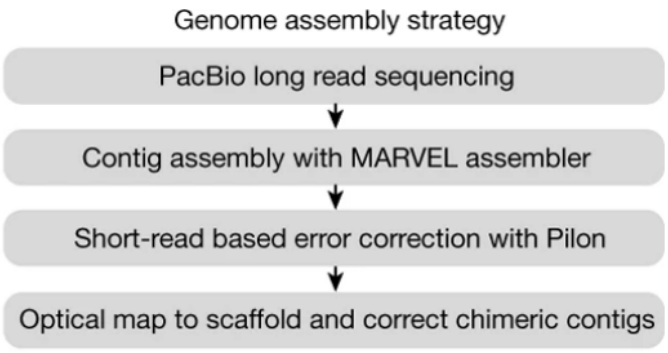

Figure 1: The assembly strategy consists out of four steps: long-read sequencing, a novel assembler (MARVEL), error correction and scaffolding. Modified from S.Nowoshilow et al. (2018)[7]

The assembly process of MARVEL[9] can be summarized as seven steps: overlap, patch reads, overlap (second time), scrubbing, construction of assembly graph and touring, optional read correction and the last step fasta file creation [9].

c) **What other large eukaryotic genome was assembled using this tool?**
The other large eukaryotic genome that was assembled using this tool is the genome of the axolotl salamander (*Ambystoma mexicanum*) [7].

Task 2: Greedy assembly with OLC(5)

a) **The overlap graph: Draw the overlap graph OG, labeling all edges as discussed in the lecture.**

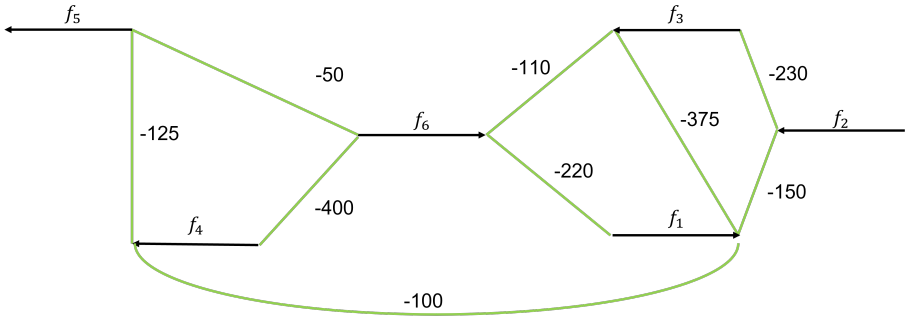

Figure 2: Our overlap graph (OG)

b) **A minimal spanning tree: Draw a minimal spanning tree in OG, containing all read edges**

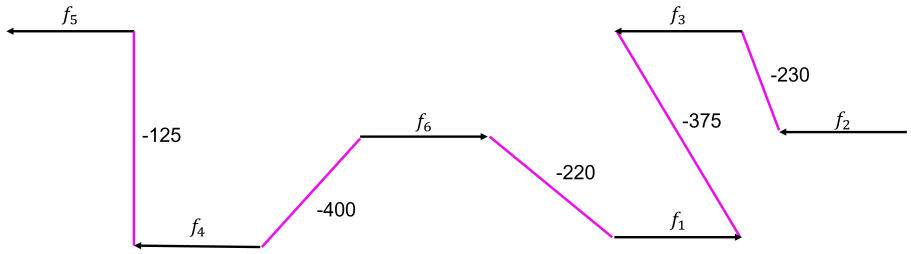

Figure 3: The minimal spanning tree for our OG in 2

c) **The layout: Draw the layout of the reads as given by the minimal spanning tree, indicating the approximate coordinates of the start and end of each read. What is the length of the final assembly?**

The length of the final assembly is 1650 (see Figure 4) which is equal to the six times the approximate length minus the length of the path of the spanning tree $(6 * 500 - 1350)$.

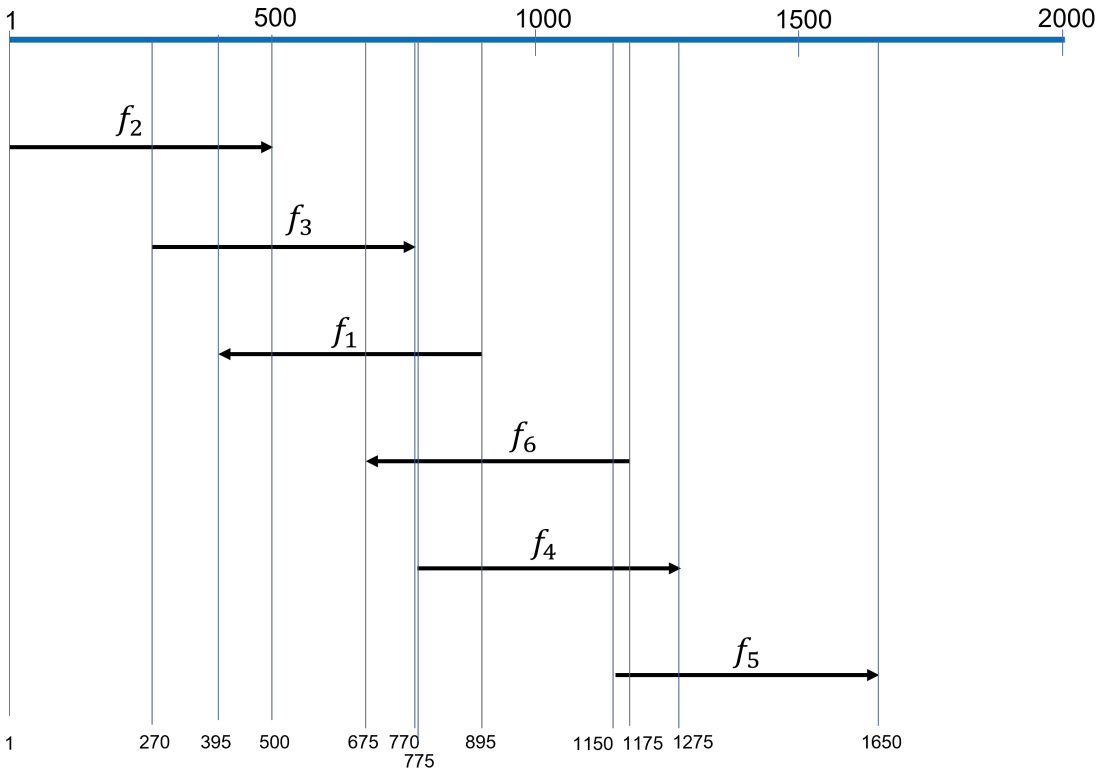

Figure 4: Layout of our minimal spanning tree (figure3)

d) **Consistent overlaps: Are all overlaps consistent with the computed layout? If not, which overlaps are not consistent with the layout, and why?**

All overlaps from the minimum spanning tree are consistent, but the others which are additional in the overlap graph are not (Table1). The layout is based on the minimum spanning tree and does not take the whole overlap graph into account. In our case not all overlaps are consistent, this is due to random overlaps (e.g., reads f_4 and f_1). There are also some inconsistencies in length, which could be due to the fact that the sequences only have an approximate, not an exact length of 500. The use of the layout approach contains a main problem, we are not able to distinguish between true overlaps, random overlaps and repeat-induced overlaps.

Overlap Graph	**Layout (minimum spanning tree)**	**Consistent (Yes or No)**
$o(5'f_5 - 5'f_6) = 50$	$o(5'f_5 - 5'f_6) = 25$	No
$o(3'f_4 - 3'f_1) = 100$	$o(5'f_4 - 5'f_1) = 80$	No
$o(3'f_1 - 3'f_2) = 150$	$o(3'f_1 - 3'f_2) = 105$	No
$o(3'f_6 - 3'f_3) = 110$	$o(3'f_6 - 3'f_3) = 95$	No

Table 1: Consistent overlaps: Comparing the overlap graph with the layout of the minimum spanning tree

Practical Assignments

Task 3: Contigs versus coverage(5)

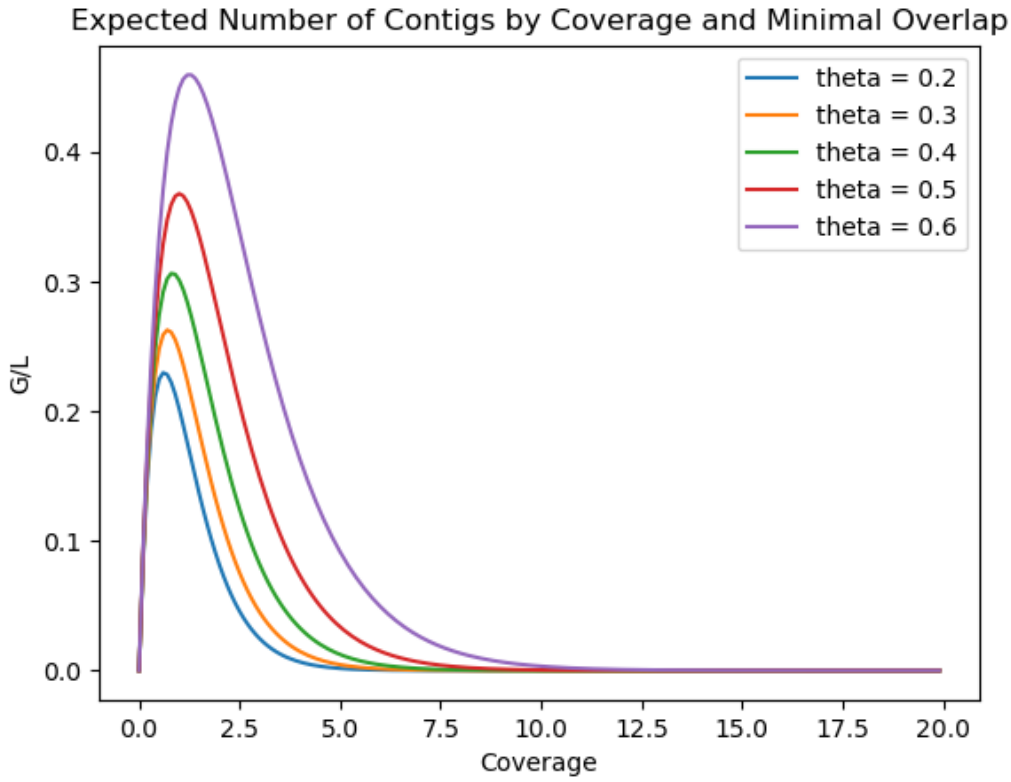

Figure 5: Number of contigs dependent on coverage values. The amount of contigs is generally higher with higher θ, while the number of contigs reaches a maximum at values below 2.5 and then drops exponentially, eventually approaching zero.

We can see that with increasing coverage, the number of contigs first approaches its maximum very fast at values below 2.5 for every θ but different for each θ. We can see that the expected number of contigs is generally higher if the minimal overlap θ between reads is higher. The course of each individual curve makes intuitive sense, because the number of contigs first increases quickly, but with increasing depth of coverage each base pair is read a higher number of times and therefore, the probability of overlap and coalescence between contigs is greater.

Task 4: Sequencing Read Quality Control (7)

a) **Shortly explain the term paired-end sequencing and its difference to a single-end sequencing approach.**

In paired-end sequencing, sequences can be read from both ends, which increases the read quality [5]. In single-paired sequence reads however, the sequence is only read from one end [2]. Paired-end sequencing, because of reading the same sequences twice and in reverse order, allows for easier assembly of the genome[2].

b) **Run FastQC[2] on the six compressed FastQ files. Hint: You do not need to gunzip the files to run FastQC.**

c) **Summarize all FastQC results into one common HTML report using MultiQC[3]. Hand in the created report as an additional file**

d) **Find out the length of the input genome and compute the mean coverage for each sequencing run. Remember: we have reads from paired-end sequencing approach!**

The total amount of base pairs is 3268203 base pairs [8]. The coverage is computed as $C = N \cdot L/G$, where N is the number of reads, G is the length of the original genome and L is the average read length. Because we have sequences from a paired-end approach and each run is sequenced twice, we multiply the read length with 2. Following this formula, we get the resulting coverages as:

$$Run_1 : C_1 = \frac{2 \cdot 108940 \cdot 150}{3268203} = 10 \tag{1}$$

$$Run_2 : C_2 = \frac{2 \cdot 217880 \cdot 150}{3268203} = 20 \tag{2}$$

$$Run_3 : C_3 = \frac{2 \cdot 326820 \cdot 150}{3268203} = 30 \tag{3}$$

e) **Discuss the quality of the three different paired-end sequencing experiments. Include high-quality and meaningful figures from the MultiQC report in your discussion. In your text, make sure that you correctly refer to the included figures. How certain would be the results of an assembly for each of the different sequencing runs?**

The MultiQC [3] report gives several measures that provide insight into the quality of the different sequencing runs. To illustrate the differences in quality, we focused on quality scores given by the Phred Score, figures concerning the GC content of each sequence run and base percentage by position in reads. Other figures were fairly similar between the different runs. As we will illustrate in the following text, sequencing runs 1 and 2 seemed to have good read quality, while sequencing run 3 was flagged as having inferior quality. In figure 8 one can see that while the trajectory of the curve is fairly similar for each sequencing run - first increasing slightly and then slowly dropping- the location on the y-scale representing the Phred score ranges from

ca. 5 for the lowest-scoring sequencing run (3.2) to ca. 55 for the highest-scoring sequencing run (2.1). Furthermore, runs 3.1 and 3.2 have the lowest quality with Phred scores never exceeding 20 and remaining in the graph area that is shaded red. On the other hand, runs 1.1., 1.2, 2.1 and 2.2 have high quality throughout and do not fall below the area that is shaded green in the plot, indicating good quality. We can see that the quality for the sequencing runs 2 is highest, while sequencing runs 1 still have good, albeit lower quality. This is also reflected in Figure 8. Here we can see the count of sequences per quality score for each sequencing run. We can see very narrow peaks with maxima located at the Phred-scores that were elaborated on above. this indicates that while the read quality differs greatly between sequencing runs, the Phred score within a run is fairly stable. While these two quality measures were the only ones that had the sequencing runs 3 flagged for inferior quality, there are also some noticeable differences in other measures. For example, the GC content of nearly all sequencing runs is very close to the theoretical normal distribution with peak values at 57%, but the GC content of run 3.2 is slightly lower at 51 %, indicating "a contaminated library or some other kinds of biased subset" [1]. Similarly, the base-content of the bases A, T, C and G is very stable across reads (see Figure 9), with A and T ranging at around 21% and G and C ranging at around 29% for all read positions, but in run 3.2, these proportions vary and even overlap, indicating inferior quality for this read (see Figure 10). The certaintly of an assembly for each of the different sequencing runs can be assessed using the Phred score, which is computed with the following formula:

$$Q = -10 \cdot log_{10}(P) \tag{4}$$

With Q being the score and P being the probability of an incorrectly classified base (see lecture slides chapter 8, slide 50). Therefore, with mean Phred scores of 2 and 16 for the third run, this probability would be 63.1% and 2.5% respectively, averaging on a false read probability of 32.8%. For the first run with Phred scores of 33 and 34, the probabilities would be 0.05% and 0.04%, averaging on 0.045%. For the second run with the highest mean Phred scores of 41 and 43, the probabilities of an incorrectly called base would be 0.008% and 0.005%, averaging on 0.0065%. We see that these probabilities reflect what we have outlined before in terms off read quality.

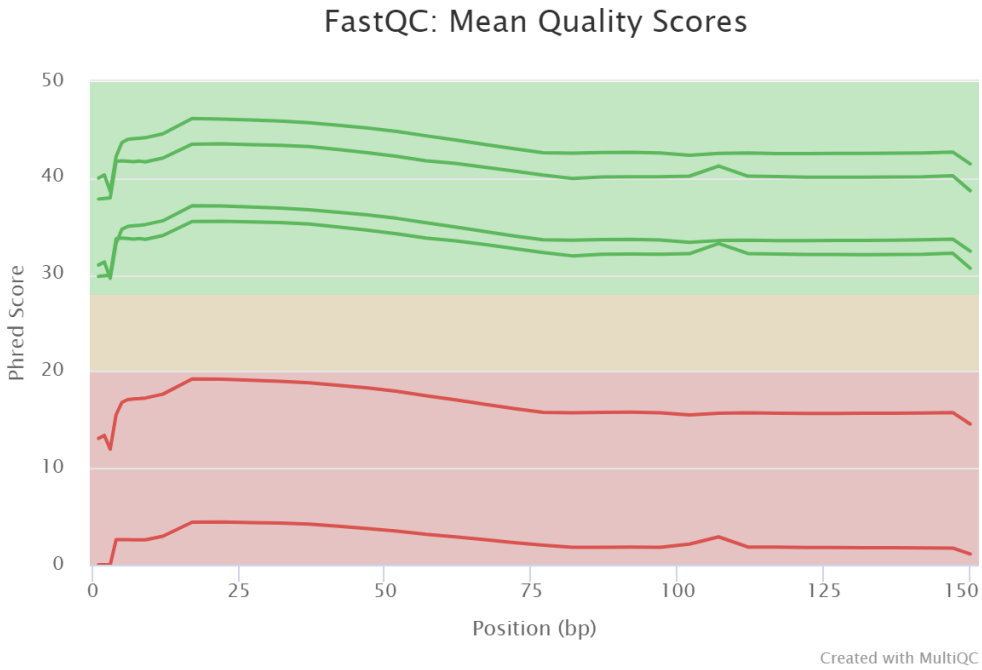

Figure 6: FastQC mean read quality dependent on position in sequence. From the top you can see the runs in the order 2_1, 2_2, 1_1, 1_2, 3_1 and 3_2

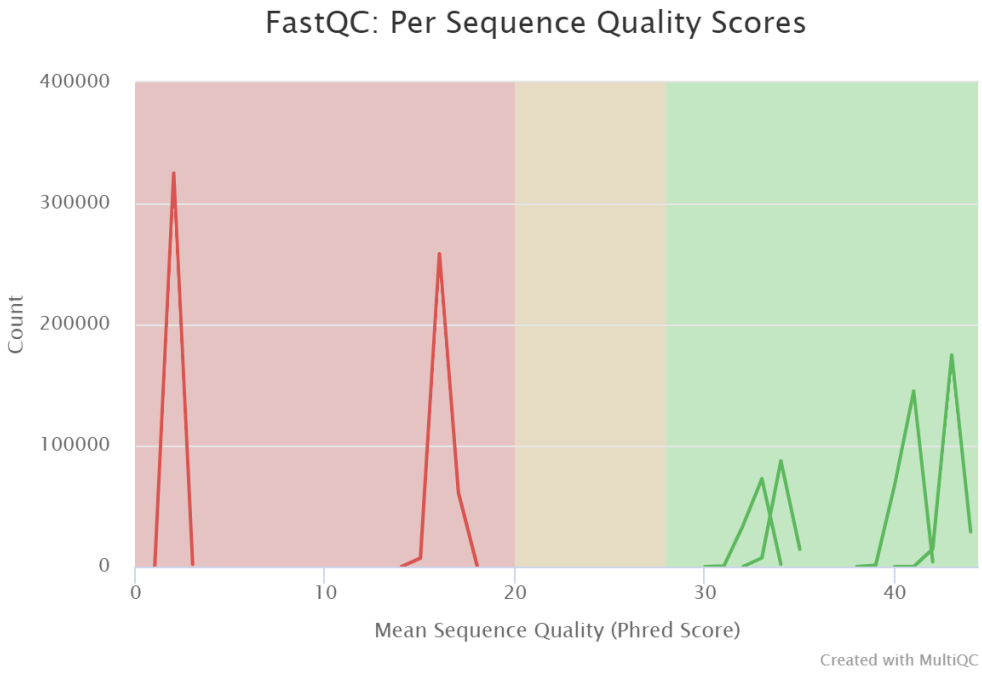

Figure 7: FastQC mean read quality per sequence. There are stark differences in quality, the lowest quality reads coming from the the runs 3_1 and 3_2, reads with the best quality from runs 2_1 and 2_2 and reads with still good quality from runs 1_1 and 1_2. From the left you can see the runs in the order 3_2, 3_1, 1_2, 1_1, 2_2 and 2_1

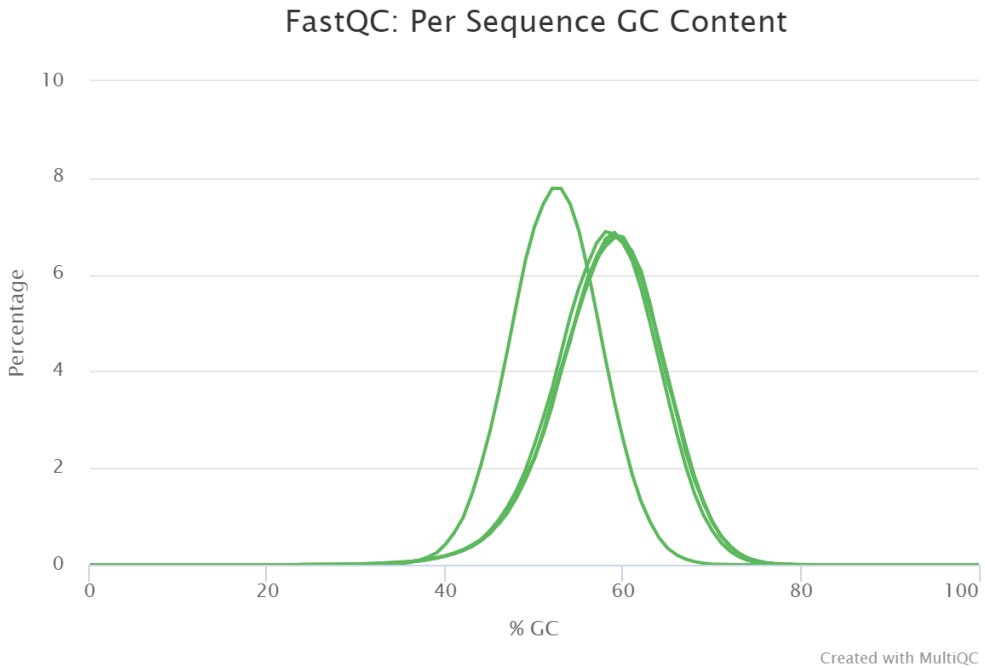

Figure 8: GC content of the reads . You can see that all the runs except for run 3_2 fit very closely to the same theoretical distribution. However, run 3_2 differs slightly from the theoretical distribution. This also indicates a drop in read quality for this run.

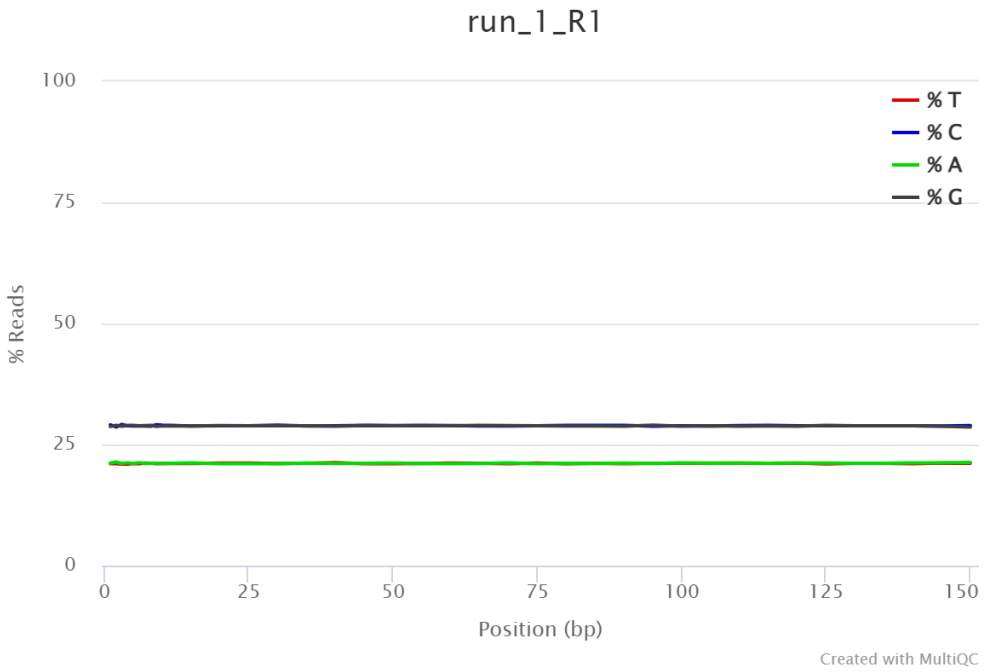

Figure 9: Example of a stable base content by read position, here in run 1.1

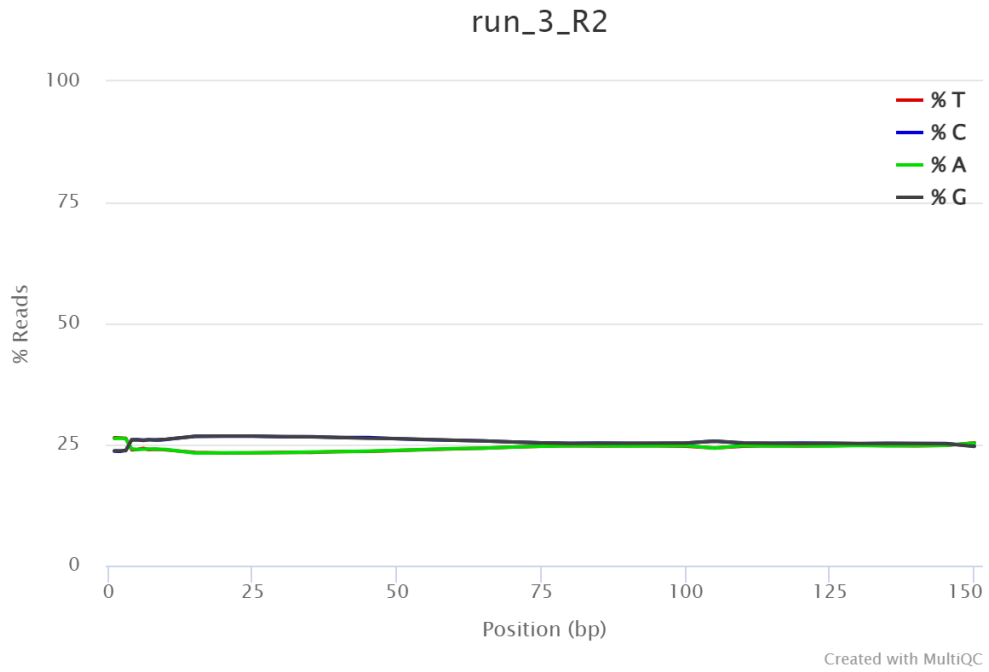

Figure 10: In Run 3.2, the base content dependent on read position was not stable, but fluctuated.

References

[1] Babraham Bioinformatics. Per sequence gc content. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html, 2022.

[2] Columbia Systems Biology. Genome sequencing: Defining your experiment. https://systemsbiology.columbia.edu/genome-sequencing-defining-your-experiment, 2019.

[3] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19):3047–3048, 2016.

[4] Alex Fox. Australian lungfish has biggest genome ever sequenced. https://www.smithsonianmag.com/smart-news/australian-lungfish-has-biggest-genome-ever-sequenced-180976837/#:~:text=This%20primitive%20looking%20fish%2C%20with,of%2032%20billion%20base%20pairs., 2021.

[5] Inc. Illumina. Paired-end vs. single-read sequencing technology. https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html, 2022.

[6] Axel Meyer, Siegfried Schloissnig, Paolo Franchini, Kang Du, Joost M Woltering, Iker Irisarri, Wai Yee Wong, Sergej Nowoshilow, Susanne Kneitz, Akane Kawaguchi, et al. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature, 590(7845):284–289, 2021.

[7] Sergej Nowoshilow, Siegfried Schloissnig, Ji-Feng Fei, Andreas Dahl, Andy WC Pang, Martin Pippel, Sylke Winkler, Alex R Hastie, George Young, Juliana G Roscito, et al. The axolotl genome and the evolution of key tissue formation regulators. Nature, 554(7690):50–55, 2018.

[8] J. Parkhill. Mycobacterium leprae tn, complete sequence. `https://www.ncbi.nlm.nih.gov/nuccore/15826865/`, 2021.

[9] Siegfried Schloissnig. The marvel assembler. `https://github.com/schloi/MARVEL`, 2020.