BIO310 HW-2 / Beril Kiyanfer

History of "Marburg marburgvirus Jobs”:

https://usegalaxy.eu/u/beril.kiyanfer/h/marburg-marburgvirus-jobs

Exercise 1: Search for Marburg marburgvirus online. How and from what is it usually transmitted? What are the signs and the symptoms? How is it diagnosed? Is there any possible treatment? Is there any current/recent outbreak going on? How can people prevent themselves from getting ill or infected? Do not forget to include references for the information you have gathered for the sake of academic integrity, which is more important than getting a good grade.

- Marburg virus disease (MVD) is a severe, often fatal illness caused by the Marburg virus, like Ebola, and transmitted from bats or through direct contact with infected people's bodily fluids. Symptoms start suddenly with fever, headache, muscle aches, followed by diarrhea, abdominal pain, and hemorrhagic signs. The incubation period is 2 to 21 days, with death typically occurring between 8 and 9 days after symptom onset in fatal cases. Diagnosis is challenging, using tests like ELISA and RT-PCR among others. There's no specific vaccine or treatment, though supportive care can improve survival. Prevention involves avoiding contact with infected fluids and animals, and safe burial practices.

(references are at the end)

Exercise 2: Look at the fastq file for forward reads. Are the lengths of the reads same? Why do you think if not? Now, look at the first read, do you think it has good quality scores, or there are some nucleotides with bad scores?

- The lengths of the reads in our FASTQ snippet are not all the same.
- There are couple of reasons when we want to understand if the lengths of the reads same or not. First of all, it can be because of a sequencing technology. Different sequencing platforms and technologies have varying capabilities and limitations that can result in a range of read lengths. Another one can be quality trimming. During the preprocessing of sequencing data, low-quality ends of reads are often trimmed to improve data quality. This process can result in variable lengths of final reads.
- The first line contains a mix of ASCII characters representing different quality scores. Characters such as '!' (ASCII 33) would indicate very low-quality scores (error probability of 1), while characters such as 'I' (ASCII 73) would represent high-quality scores (error probability of about 0.0005).

- In this read, several scores fall below the threshold of 20, indicating potentially unreliable nucleotide calls in those positions. However, there are also many scores above 20, which suggests a mix of high and low confidence in the read's base calls.

Exercise 3: In what ways can Phred scores be beneficial when using the de Bruijn graph method for genome assembly?

There are couple of ways to show that Phred scores can be beneficial when using the Brujin graph method for genome assembly. Here are the two reasons I found through my research:

- Error Filtering: High-quality reads are essential for accurate assembly. Phred scores allow for the identification and removal of low-quality reads that could introduce errors into the assembly process.
- Read Trimming: The end of sequencing reads often have lower quality than the middle. Phred scores can be used to trim reads at points where the quality drops below a certain threshold, ensuring that only high-confidence nucleotides are used in the assembly.

By leveraging high-quality Phred scores, the accuracy and reliability of de Bruijn graph-based genome assembly can be significantly improved, leading to better assembly outcomes and more reliable downstream analysis.

Exercise 4: What is Illumina's rationale behind determining the quality score for a specific base?

Illumina uses quality scores to provide a probabilistic estimate of the accuracy of each base call made during sequencing. The quality score, often referred to as a Phred score, represents the likelihood that a particular base has been correctly identified. The higher the quality score, the lower the probability of an error in base calling.

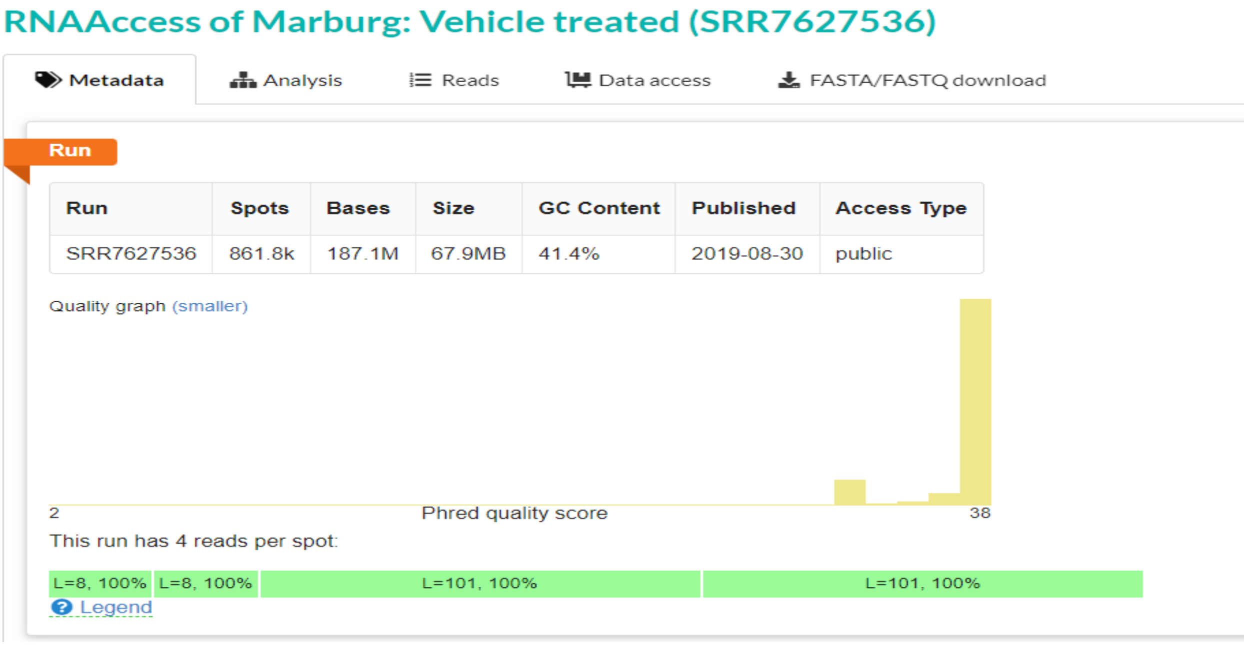
Error Probability: The Phred quality score Q is logarithmically linked to the base-calling error probability P as follows: $Q = -10 \log_{10} P$. So, a quality score of 20 would correspond to an error probability of 10^{-2} or 1 in 100.

Statistical Modeling: The quality score is derived from a statistical model that considers various aspects of the sequencing process. Factors such as the intensity of the fluorescent signal, the ratio of signal-to-noise for each base, the position of the base within the read, and any systematic errors known to be associated with the sequencing platform are used to calculate the score.

Quality Control and Filtering: These scores allow for quality control of the sequencing data. Low-quality bases can be filtered out or trimmed during the data processing steps, improving the overall reliability of the sequence data.

Exercise 5: Please navigate back to the SRA page for our dataset and go to the "Metadata" tab. Look for the section that says, "Quality graph" and click the "bigger" option if necessary to enlarge the histogram. Analyze the chart and determine if the quality scores are satisfactory.

- The highest peak of the histogram is at the far-right end, suggesting that the highest proportion of base calls have a Phred quality score of 38. A Phred score of 38 indicates a very low error probability, specifically a 0.000126 (or about 1 in 7943) chance of an incorrect base call, which is quite high-quality. The green bars labeled "L=8, 100%" and "L=101, 100%" suggest that all reads at lengths 8 and 101 have high-quality scores (although the length of 8 is likely to be a tag or index rather than a sequencing read).



RNAAccess of Marburg: Vehicle treated (SRR7627536)

Run	Spots	Bases	Size	GC Content	Published	Access Type
SRR7627536	861.8k	187.1M	67.9MB	41.4%	2019-08-30	public

Quality graph (smaller)

This run has 4 reads per spot:

L=8, 100%	L=8, 100%	L=101, 100%	L=101, 100%

Exercise 6: Check the .html report of fastp. How many reads are discarded? Were there any sequencing adapters detected? Why do you think if not? Was the average read length changed after fastp?

There were sequencing adapters detected in both read1 and read2, albeit in very small percentages (approximately 0.372308% for read1 and 0.376639% for read2). This low level of adapter content suggests that the majority of sequencing adapters were likely removed prior to the use of fastp. It’s possible the reads had undergone an initial processing step designed to trim adapters, or the library preparation kit used was effective in preventing adapter-dimer formation. The average read length did not change after fastp, remaining at 100bp for both reads before and after filtering. This indicates that the fastp tool did not need to trim the reads to remove low-quality bases or adapter sequences significantly. It also reflects the high quality of the reads, where only a minimal amount of filtering was required to maintain the quality thresholds set for the dataset.

Exercise 7: Report the following for the alignment. How much has it changed?

- Length: 24583
- Identity: 10630/24583 (43.2%)
- Similarity: 10630/24583 (43.2%)
- Gaps: 10942/24583 (44.5%)
- Score: 17131.0

NODE_1 is the longest ---> 0 to 18,000

Exercise 8: Why do you think there are gene(s) labeled as “hypothetical proteins”?

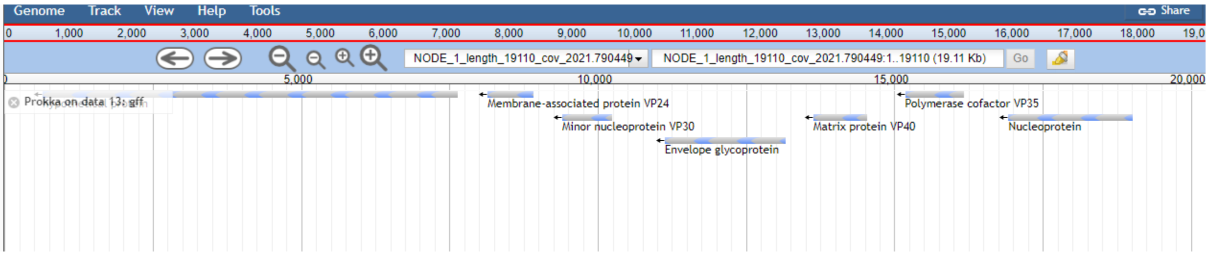
- Genes labeled as “hypothetical proteins” are sequences for which a coding function is predicted, but no direct evidence of the protein product exists (such as protein sequences in databases). These predictions are based on open reading frames that have characteristics of genes, including length, codon usage, and proximity to promoter-like sequences, but lack any known function or similarity to other characterized proteins. Essentially, they're called "hypothetical" because their function is not yet known or validated through experimental data.

Exercise 9: In addition to the “hypothetical” one(s), which genes did you find? Search for them online and report their viral functions.

- **Membrane-associated protein VP24:** Prevents the establishment of cellular antiviral state by blocking the interferon-alpha/beta (IFN-alpha/beta) and IFN-gamma signaling pathways. Blocks the IFN-induced nuclear accumulation of host phosphorylated STAT1 by interacting with the STAT1-binding region of host importins. Alternatively interacts also directly with host STAT1 and may additionally inhibit its non-phosphorylated form. Plays a role in assembly of viral nucleocapsid and virion budding. May act as a minor matrix protein that plays a role in assembly of viral nucleocapsid and virion budding.
- **Minor nucleoprotein VP30:** Acts as a transcription anti-termination factor immediately after transcription initiation but does not affect transcription elongation. This function has been found to be dependent on the formation of an RNA secondary structure at the transcription start site of the first gene.
- **Envelope glycoprotein:** GP1 is responsible for binding to the receptor(s) on target cells. Interacts with CD209/DC-SIGN and CLEC4M/DC-SIGNR which act as cofactors for virus entry into the host cell. Binding to CD209 and CLEC4M, which are respectively found on dendritic cells (DCs), and on endothelial cells of liver sinusoids and lymph node sinuses, facilitate infection of macrophages and endothelial cells. These interactions not only facilitate virus cell entry, but also allow capture of viral particles by DCs and subsequent transmission to susceptible cells without DCs infection (trans infection) (By similarity)
- **Matrix protein VP40:** Plays an essential role virus particle assembly and budding.
- **Polymerase cofactor VP35:** Plays an essential role in viral RNA synthesis and also a role in suppressing innate immune signaling. Acts as a polymerase cofactor in the RNA polymerase transcription and replication complexes
- **Nucleoprotein:** Encapsidates the genome, protecting it from nucleases. The encapsidated genomic RNA is termed the nucleocapsid and serves as template for transcription and replication. During replication, encapsidation by NP is coupled to RNA synthesis and all replicative products are resistant to nucleases.

Exercise 10: Are all the annotated genes on the same strand? If so, why?

- In the Marburg virus, all genes are indeed on the same strand of its RNA genome. They are on the same strand because Marburgvirus is a single-strand, negative-sense RNA. This setup simplifies transcription and ensures efficient gene expression, critical for the virus's replication. This streamlined genome organization is common in many RNA viruses, optimizing their lifecycle.



NODE_1_length_19110_cov_2021.790449

NODE_1_length_19110_cov_2021.790449:1..19110 (19.11 Kb)

REFERENCES:

Centers for Disease Control and Prevention. (2023, April 19). Diagnosis | Marburg (Marburg Virus Disease). CDC. Retrieved March 25, 2024, from https://www.cdc.gov/vhf/marburg/diagnosis/index.html

National Center for Biotechnology Information (NCBI). (n.d.). *NCBI*. Retrieved March

25, 2024, from https://www.ncbi.nlm.nih.gov/

Nucleic Acids Research, The Galaxy Community 2022 - The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update, gkac247. doi:10.1093/nar/gkac247

The Galaxy server that was used for some calculations is in part funded by Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grants 031 A538A/A538C RBC, 031L0101B/031L0101C de.NBI-epi, 031L0106 de.STAIR (de.NBI)). https://usegalaxy.eu/u/beril.kiyanfer/h/marburg-marburgvirus-jobs

UniProt Consortium. (n.d.). *UniProt: a worldwide hub of protein knowledge*. Retrieved March 25, 2024, from https://www.uniprot.org/

World Health Organization. (n.d.). Marburg virus disease. WHO. Retrieved March 25, 2024, from https://www.who.int/news-room/fact-sheets/detail/marburg-virus-disease