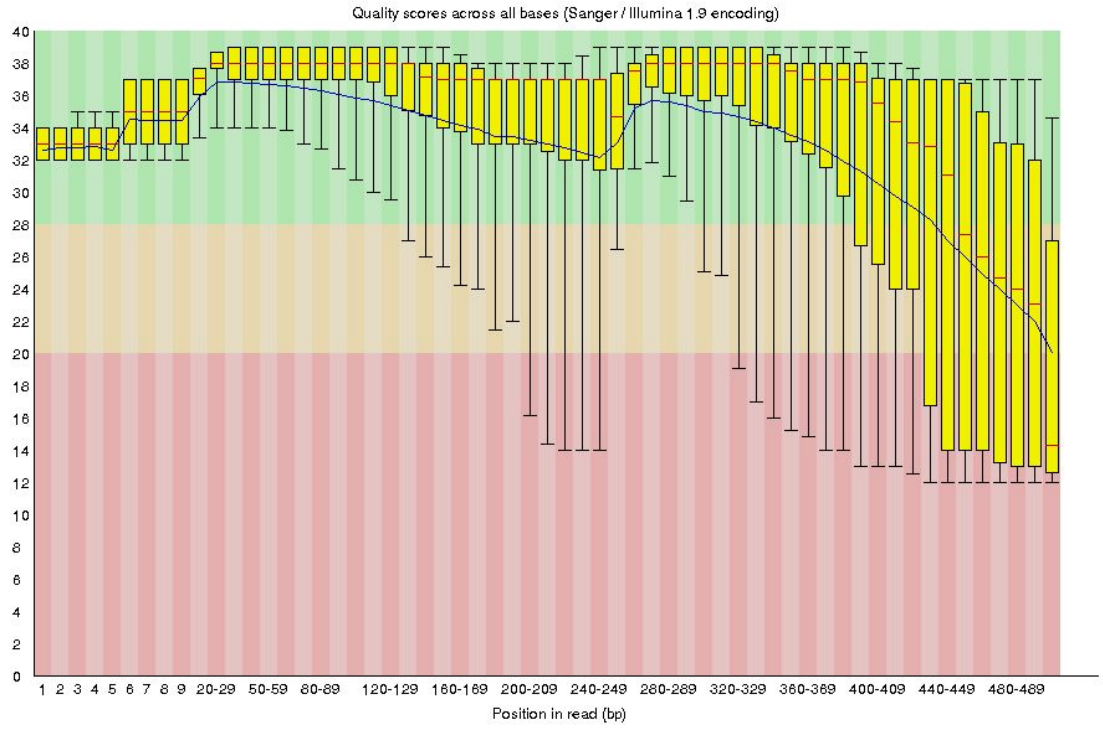
Genome Assembly
Team-II
Homework Assignment (50 pts)

- Answers should not exceed 4 sentences
- Please do not run anything related to this homework assignment on the server. None of the questions require you to run any commands.
- This homework is due Tuesday, March 6th at 1:25PM.

Preassembly (10 pts)

1. The following is a per base sequence quality graph for a Illumina MiSeq paired end sequenced sample generated by FastQC. The sample has not been subject to trimming and adapter removal.
    a. Are these forward reads, reverse reads or both? Explain your answer. (Hint: consider the average length of a single read) (3pts)
    b. Why is there a "bump" in quality at 240-249bp? (2pts)
    c. Why do untrimmed reads always have low quality scores towards the end? (2pts)
    d. Will there be a significant difference in quality between forward reads and reverse reads from paired end sequencing? Why/why not? (2pts)



2. What are adapters and why do we want to remove them during trimming? (1pts)

Reference Assembly (15 pts)

3. Upload sample SRR5666547_1_27.f (found in /projects/data/homework/hw1/) to the webtool version of StrainSeeker through the following link:

   http://bioinfo.ut.ee/strainseeker/index.php?r=site/webtool

   Based on the results you get:
   a. What reference genome would you use to assemble this sample? (2pts)
   b. Why is it important to have a highly related species as your reference? (2pts)

4. Describe a command line sequence for reference assembly using SMALT. If you were to assemble this sample (with the reference genome chosen in the last question), what options would you use and why? (5pts)

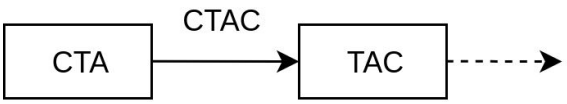
5. In order to process these assembled reads via Quast, they must be converted to Fasta files. However, a simple conversion will not work. Look at the SampleSystemCalls.txt file found in /projects/data/homework/hw1/. Describe the general steps that had to be taken in order to convert our files from SAM to FASTA, and explain why this had to be done. (6pts)

***De Novo* Assembly (15 pts)**

6. Draw the De Bruijn graph for the following sequence. Use k-mer length of 4 for edges (meaning k-mer length of 3 for nodes). (7pts)

   CTACGTACTTACC

   (Hint: The graphs starts as:)



7. Does the graph lead to one unambiguous assembly? If not, report contigs. (3pts)

8. Draw the De Bruijn graph for the above sequence with a k-mer length of 7 for edges. Is the graph unambiguous? What impact does k-mer length have on graph traversal? (5pts)

Assembly Evaluation (10 pts)

9. The following is a Quast report comparing the assemblies from 4 different *de novo* tools.
    a. Which aligner has the best L50 value? How is the L50 different from N50? (2pts)
    b. Why would 'N's be introduced into the assemblies? (3pts)
    c. Which aligner has the largest assembly? Is this an indicator of assembly quality? Why/why not? (3pts)
    d. What is the significance of GC%? (2pts)

Statistics without reference	skesa	masurca	spades	unicycler
# contigs	94	150	109	91
# contigs (>= 0 bp)	97	153	344	123
# contigs (>= 1000 bp)	91	140	81	87
# contigs (>= 5000 bp)	67	122	59	63
# contigs (>= 10000 bp)	57	104	52	55
# contigs (>= 25000 bp)	41	69	38	39
# contigs (>= 50000 bp)	30	40	28	28
Largest contig	545 362	155 124	619 026	432 941
Total length	5 309 452	5 263 663	5 313 635	5 283 100
Total length (>= 0 bp)	5 310 742	5 265 026	5 374 870	5 290 099
Total length (>= 1000 bp)	5 307 576	5 256 252	5 295 449	5 280 222
Total length (>= 5000 bp)	5 258 777	5 209 897	5 240 425	5 224 685
Total length (>= 10000 bp)	5 195 434	5 077 432	5 194 434	5 173 809
Total length (>= 25000 bp)	4 898 118	4 475 069	4 938 645	4 904 645
Total length (>= 50000 bp)	4 516 625	3 453 247	4 618 950	4 546 316
N50	143 821	67 231	198 148	177 707
N75	94 572	35 799	98 154	98 000
L50	11	26	10	11
L75	22	52	19	21
GC (%)	57.31	57.39	57.3	57.33
Mismatches				
# N's	11	0	212	0
# N's per 100 kbp	0.21	0	3.99	0