1. The output of my tool resulted in a mean quality score of 30.00 across all the bases in my FASTQ influenza A virus file SRX23723064. According to the PHRED quality score table, any quality score at 30 indicates a 0.001 error probability, meaning this sequencing data is extremely high quality as each base call has a 99.99% chance of being correct. Although encouraging data, I would be cautious in automatically assuming this data is correct as the quality score of each sequence in the FASTQ file is a "?." This could possibly indicate some questionable data reliability, which could further impact some downstream analysis.

```
In [2]: FastQC('SRR28073705.fastq')
```

Position	Mean
1	30.00
2	30.00
3	30.00
4	30.00
5	30.00
6	30.00
7	30.00
8	30.00
9	30.00
10	30.00
11	30.00
12	30.00
13	30.00
14	30.00
15	30.00
16	30.00
17	30.00
18	30.00
19	30.00
20	30.00
21	30.00
22	30.00
23	30.00
24	30.00
25	30.00
26	30.00
27	30.00
28	30.00
29	30.00
30	30.00
31	30.00
32	30.00
33	30.00
34	30.00
35	30.00
36	30.00
37	30.00
38	30.00
39	30.00
40	30.00
41	30.00
42	30.00
43	30.00
44	30.00
45	30.00
46	30.00
47	30.00
48	30.00
49	30.00
50	30.00
51	30.00
52	30.00
53	30.00
54	30.00
55	30.00
56	30.00
57	30.00
58	30.00
59	30.00
60	30.00
61	30.00
62	30.00
63	30.00
64	30.00
65	30.00
66	30.00
67	30.00
68	30.00
69	30.00
70	30.00
71	30.00
72	30.00
73	30.00
74	30.00
75	30.00
76	30.00
77	30.00
78	30.00
79	30.00
80	30.00
81	30.00
82	30.00
83	30.00
84	30.00
85	30.00
86	30.00
87	30.00
88	30.00
89	30.00
90	30.00

2. Unfortunately, I was not able to install Mascara into my machine. Despite many attempts, there seemed to be some sort of configuration issues that I could not troubleshoot. I was

also unable to obtain a contig.fa file from velvet despite many attempts and different flags with my run. Regardless, the following analysis will be based on my assemblies in megabit and unicycler. In terms of completeness, the megahit assembly had a larger assembled genome size of 13770, compared to that of the unicycler which only had 11126. This could indicate that the unicycler assembly process may not have captured all the regions of the genome when assembling, in turn leading to a smaller percentage of reference genome covered of 81.639%, while megahit had 99.053%. In terms of contiguity, unicycler did have a smaller largest contig compared to megahit, however. unicycler had a much better N50 and N90 score of 2243 and 1570 respectively, compared to megahit. This could suggest that overall unicycler may have a better distribution across contigs. In terms of correctness, unicycler had 0 assemblies, indicating a much higher correctness compared to megahit and while megahit only had 1 misassembly, it did span a contig length of 1066. Unicycler also have a lower mismatch score, in turn also having a lower indel score and a much smaller length of the indels than megahit, indicating a much higher form of correctness. The full html files are attached to my submission.

3. After running my alignment through RAxML, I generated the following phylogenetic tree shown below (the full tree can be found at: http://etetoolkit.org/treeview/?treeid=38e4956a646f65b88e50a5627094a739&algid=ce443ed1e53858bf4e11d1e069c7a927). Looking at the full tree, I can see that my alignments have been clustered into about 6 big clades, some more distinct than others. This could indicate that a few of the clades may be closer related than others, or just have a more recent common ancestor, such as the second, third and fourth clade. Overall, the branch length of the root node is 0.51, meaning that there is a smaller evolutionary divergence between the root and common ancestor. This would be due to genomic evolution of H3N2, caused by various mutations along the genome and re-assortment. It is because of this the influenza virus has been able to maintain its existence.

Support values are shown in red.
(Loading big trees/algs may take a few seconds)