Assignment for University of Chicago

Author: Snehal K. Talati

Assignment:

Background Information:
FastQ file used for analysis was derived from one of the gene panel runs with a tumor cell line. It is illumine sequence data with pred33 q-scores, with 151 base pair reads

Task:
To write a python script that goes through the FastQ file and output both the average of the quality scores, and the standard deviation of the quality scores for each sequencing cycle.

Sample Output:

Cycle	Avg	Stdev
1	30.456	5.531
2	31.992	3.529
3	32.096	3.774
4	32.452	3.150
5	32.296	3.609
6	35.344	3.575
.. ....		
.. ....		
151	(151 avg)	(151 stdev)

Program Usage Instructions:
To run the program first run the program as follows:

      python talatisk_UOC_assignment.py [input file] [output file]

NOTE: There are two versions of the script so if the first one does not work and gives error due to Scipy or Numpy not being installed please proceed with Version 2.0 of the program named talatisk_UOC_assignment_V2.py

Also if the R-Script is utilized to generate the plot please name the output file:

      R_Data_UOC.txt

Output of R-Script:
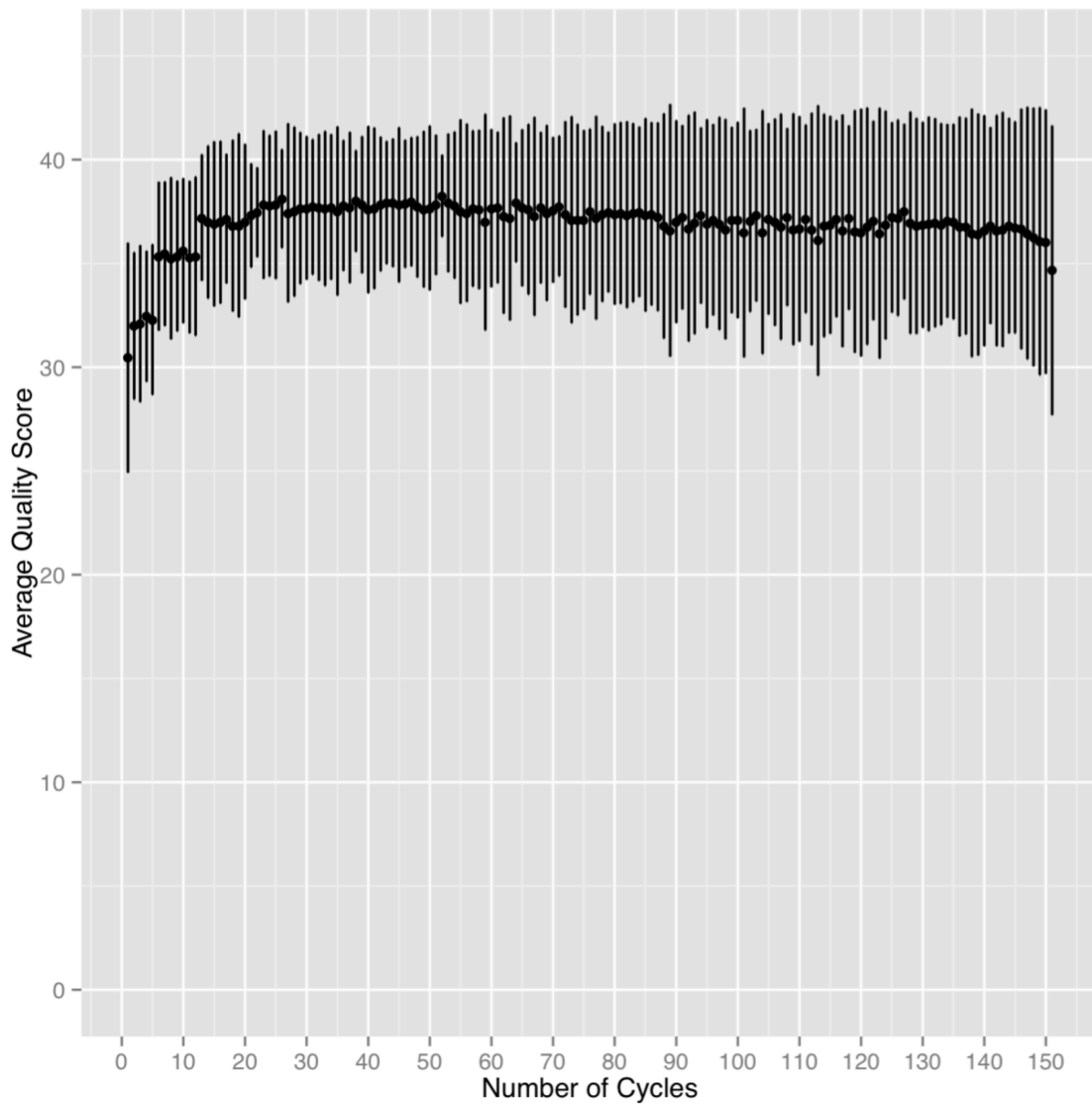
This R-Plot shows the dots, which are the average quality, score per cycle, and the bars represents the minus-plus distributions. A clear image is provided in the PDF.