Exercise 2

Questions

1. Look at the sequences below and answer the questions for the SECOND sequence -
   a. How many lines belong to each sequence (read)?
   b. Look at the quality value of the first base from the second sequence. Convert the character into a numeric value using the conversion table below.
   c. What is the Q value (Phred score) and the probability of an error in the first base of the second sequence?
   d. What is the lowest (worst) observed Phred score observed for the second sequence?

```
@HWI-ST808:87:C068VACXX:2:1101:1441:2180 1:N:0:ATCACG
GGGGGATTGTACTTATAGCAGTTGGAGAACATAAGCCGTACATCAGCAGCAAACTCCTGTGCATCCCGGTAATCACGGTTCTCCATCTTCCGCTTGACAG
+
@@@DDD6A:AC?D@EGGHCGHF9FG@F9EB?FADBHGH@GFFGGICGCECD9CGFG>CC7@>CADEEEB?D?DC@6=>=?5:ACCCC@;A>/)55?2@@C
@HWI-ST808:87:C068VACXX:2:1101:1483:2198 1:N:0:ATCACG
CACAGATATTCAGGACTTACTGAGAGGGGATGCTGCTTTGTTGGATGCTGCCAAGAAGGGCTGCCTGGCAAGAGTGCAGAAGCTCTGTACCCCAGAGAAT
+
CCCFFFFFHHGHHJJJJJJJJIJJJIJJJHIJIJJJJJJIJJGIGIJJIJJJJJJIJJJJIIJHHGGF@BDFCE;;ACCDDDDDDDDCDCDDD?BDDDDC
@HWI-ST808:87:C068VACXX:2:1101:1424:2226 1:N:0:ATCACG
TGAACAGTTTAGCATCATTAACAGCTCAGGCTGTTTCCTTTTCAGCTGATCAAGGAGATGTTCACGTGGTTCAGTTCTTACCAGTGTTGATGGGGATAAA
+
@@@DD?DBFBAFHEBGGFHGIDH>FCG:C>GHDGHBFCC@DC9:CG?FIEHGBHE?8BFGBC))8==FEIH;@@E7AEE?AEDB;7?;;CC>>B((35@A
```

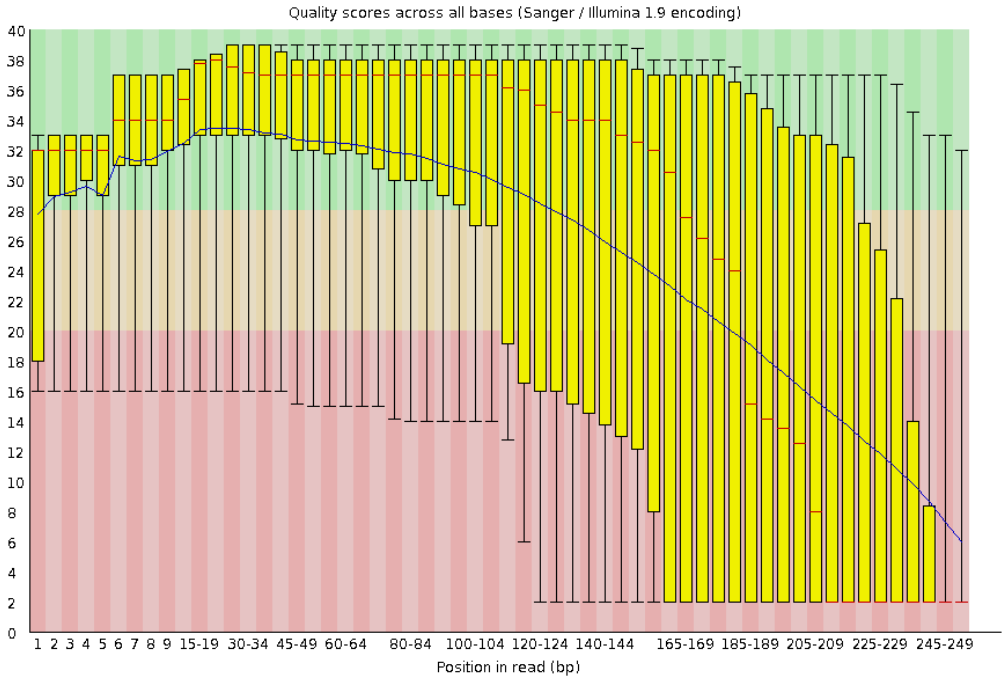
Illumina (L; 1.8+) Phred Score encoding

Symbol	Phred Score	Symbol	Phred Score	Symbol	Phred Score	Symbol	Phred Score
!	0	,	11	7	22	B	33
"	1	-	12	8	23	C	34
#	2	.	13	9	24	D	35
$	3	/	14	:	25	E	36
%	4	0	15	;	26	F	37
&	5	1	16	<	27	G	38
'	6	2	17	=	28	H	39
(	7	3	18	>	29	I	40
)	8	4	19	?	30	J	41
*	9	5	20	@	31		
+	10	6	21	A	32		

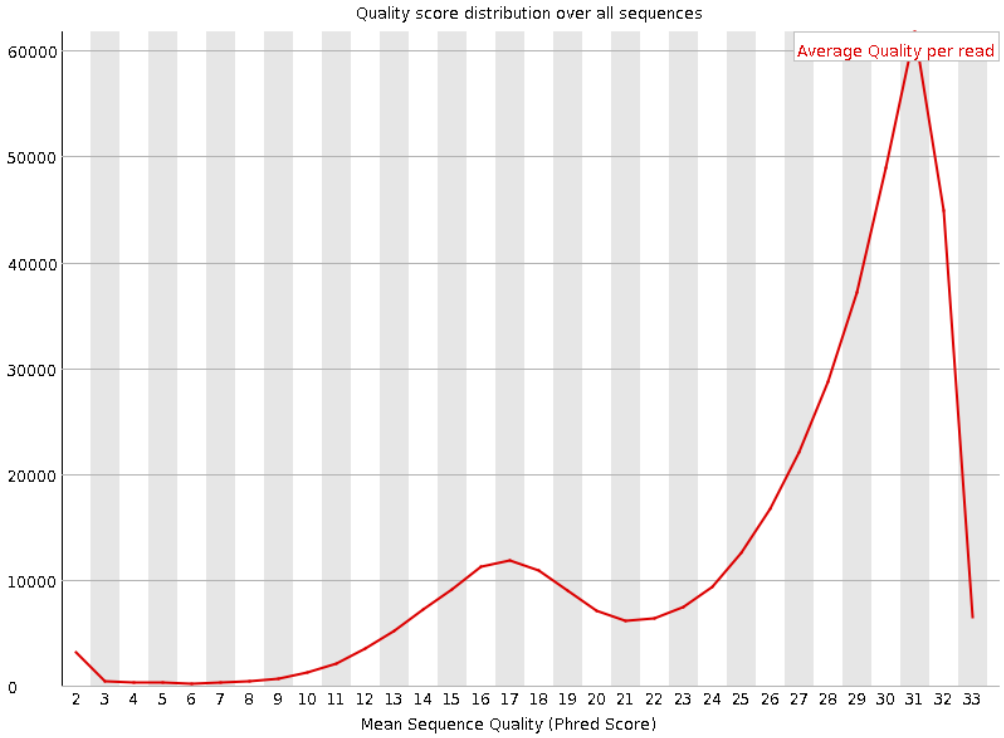
2. Look at the fastqc report for experiment 73 (exp_73) via the link on the course site and answer the following questions -
   a. How many sequences does the fastq file contain?

b. Is the base quality the same for all the cycles? Please elaborate.
c. Do all the cycles have an equal per base sequence content?

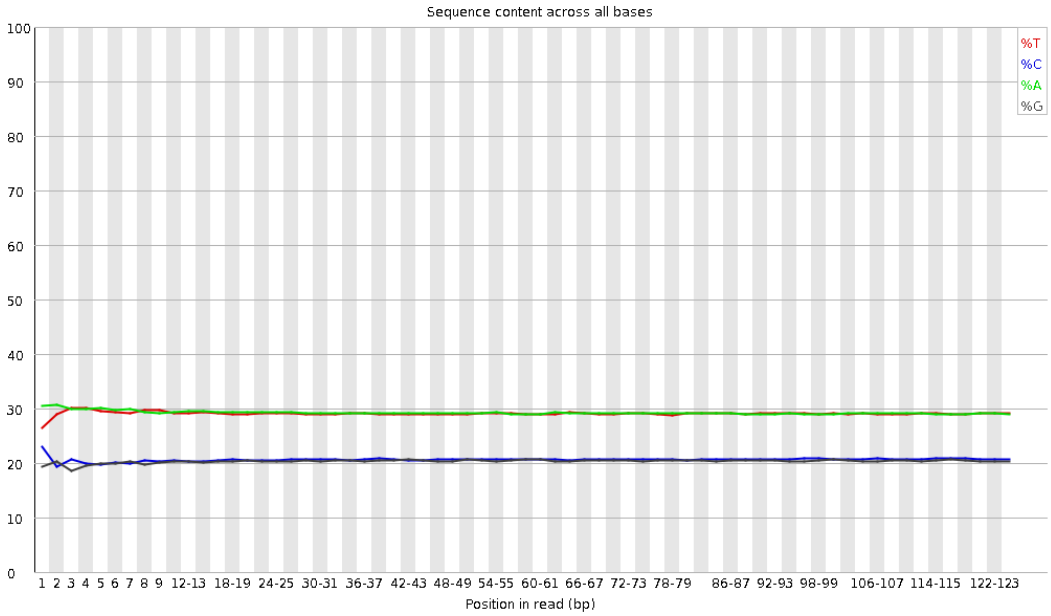
3. Do the sequences in the following plot indicate good quality? If not, what can be done?



4. What is the maximum average per sequence quality score in the following plot? Is there anything unexpected? If so, what does this indicate?



5. Does the plot below indicate a sequencing problem? Describe what could have brought about this data.



6. Look at the fastqc report for experiment 16 (exp_16) via the link on the course site and answer the following questions –
   a. Which sequence(s) are duplicated and make up almost 10% of the sequences \ reads?
   b. What is the origin of the overrepresented sequences? Are they similar? Are they different? Please detail.
7. When comparing a patterned flow cell to random one –
   a. Explain the difference between them?
   b. How does this influence the percentage of passed filtered reads (%PF)?
8. In which stages of Illumina sample manipulation and sequencing is there amplification (PCR)?
9. Why is amplification required on the flow cell?
10. What is the advantage of using 2 vs 4-channel chemistry in Illumina sequencing?