Data Cleanup: Quality Filtering

Learning objectives:

1. Identify the need for quality filtering in raw .fastq files.
2. Become familiar with the Phred Quality scoring system used to assign base call accuracies in .fastq files.
3. Recognize the principles underlying software performance of the quality filtering step by performing a manual filter on a simplified .fastq data set.

Level of Difficulty: 3/5

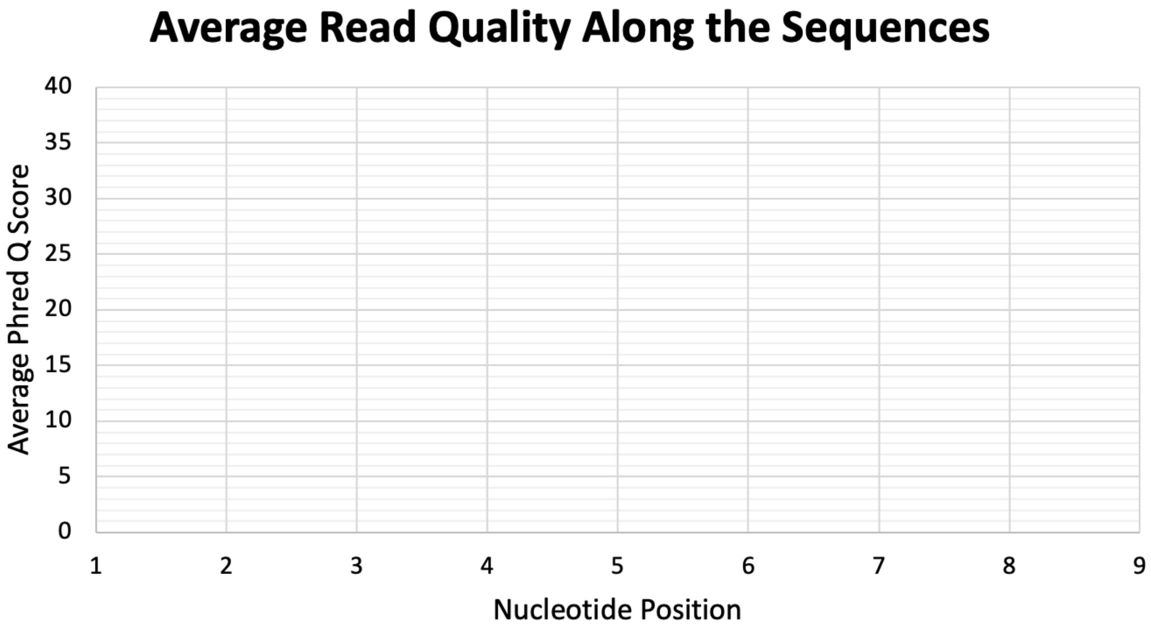
After removing contaminants, you'll need to get rid of the data that Illumina wasn't very sure about in a step called “quality trimming”. When an Illumina sequencer makes a base call (decides what nucleotide is present at a certain location in a read sequence), it also assigns a probability of error to the decision (the inverse of which would be the level of confidence). This can be thought of as sort of like Illumina's report card grade for that particular base call. When this confidence level is recorded, instead of writing a number Illumina uses an ASCII (pronounced “ask-ee”) character, based on the conversion key shown below. The data selected for removal from the read set are determined based on the data's Phred Quality score (aka Q Score) associated with the ASCII character. Low Phred scores correspond to low confidence base calls by Illumina. The minimum allowable quality score may vary based on the intended application and other context factors. Note: Quality scores are only present alongside the base calls in .fastq files, while .fasta files only contain the base calls.

ASCII character	Phred Q-Score	Probability of correct base assignment	Illumina Report Card Grade	ASCII character	Phred Q-Score	Probability of correct base assignment	Illumina Report Card Grade
"	1	21%	F	6	21	99.2%	A
#	2	37%	F	7	22	99.4%	A
$	3	50%	F	8	23	99.5%	A
%	4	60%	F	9	24	99.6%	A
&	5	68%	F	:	25	99.7%	A
'	6	75%	D	;	26	99.75%	A
(	7	80%	D	<	27	99.80%	A
)	8	84%	D	=	28	99.84%	A
*	9	87%	D	>	28	99.84%	A
+	10	90%	D	?	30	99.90%	A
,	11	92%	C	@	31	99.92%	A
-	12	94%	C	A	32	99.94%	A
.	13	95%	C	B	33	99.95%	A
/	14	96%	C	C	34	99.96%	A
0	15	96.8%	B	D	35	99.97%	A
1	16	97.5%	B	E	36	99.975%	A
2	17	98.0%	B	F	37	99.980%	A
3	18	98.4%	B	G	38	99.984%	A
4	19	98.7%	B	H	39	99.987%	A
5	20	99.0%	A	I	40	99.990%	A

To do:
1) Using the conversion key provided, convert the ASCII characters for sequences 1-4 below into Phred quality score values in the grey boxes. Then plot the average Phred quality score for each base pair position on the provided plot.

2) Now “trim” any quality scores below 20 by crossing out that portion of the sequence to ensure only high quality data remains. Recalculate and replot the average Q-score values for each position.

	T	A	T	T	A	A	T	G	T	Sequence 1
	H	H	F	C	A	<	9	7	6	ASCII characters 1
Quality Score 1										
	A	A	G	C	A	A	G	C	G	Sequence 2
	H	H	I	A	@	:	7	-	,	ASCII characters 2
Quality Score 2										
	A	C	G	T	G	G	C	T	A	Sequence 3
	G	G	F	E	=	6	4	0	+	ASCII characters 3
Quality Score 3										
	A	T	C	A	T	G	C	A	G	Sequence 4
	G	F	F	@	>	3	/	)	"	ASCII characters 4
Quality Score 4										
1) Average Quality Score pre-trimming										
2) Average Quality Score post-trimming										
	1	2	3	4	5	6	7	8	9	Basepair Position


Average Read Quality Along the Sequences
Average Phred Q Score
Nucleotide Position