Dropping Low Phred Quality Score Sequences

So in DNA Sequences, the Phred quality score is an important measure.

Phred Quality Score

Base calling accuracy, measured by the Phred quality score (Q score), is the most common metric used to assess the accuracy of a sequencing platform.
It indicates the probability that a given base is called *incorrectly* by the sequencer.

For example:

- If Phred assigns a Q score of 30 (Q30) to a base, this is equivalent to the probability of an incorrect base call 1 in 1000 times
  (Table 1). This means that the base call accuracy (i.e., the probability of a correct base call) is 99.9%.
- A lower base call accuracy of 99% (Q20) will have an incorrect base call probability of 1 in 100, meaning that every 100 bp sequencing read will likely contain an error.
- When sequencing quality reaches Q30, virtually all of the reads will be perfect, having zero errors and ambiguities. This is why **Q30 is considered a benchmark for quality in next-generation sequencing.**

Table 1: Quality Scores and Base Calling Accuracy

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Reference: Illumina ~ Q-Scores

Since the phred quality score is for each base, even if a single base has a score < 30, we have to drop that part of the sequence. (part of a parquet batch file, 60.3 M reads were split into batches, so we got 603 batches of 1,00,000 sequences)

Results from the code:

```
[14]
len(df_quality)

51690
```

Almost half the sequences got dropped

Next Steps

1. Verify if this is the right approach
2. Loop this over all the batches to have the best quality sequences only (i.e 100 bp only and phred >= 30)
3. Go to the embeddings codes (DNABERT and DNABERT 2, done in the report prior to this)
4. Loop that over all the batches to have the embeddings
5. Compare embeddings quality for DNN training