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 embeddings5. Compare embeddings quality for DNN training

Results after implementing Step 2

Logic Used:

```
# Define the quality check function
def is_quality_good(quality_scores):
return np.min(quality_scores) >= 30

# Step 2: Filter by sequence length (>= 100 bp)
df_quality = df[df['sequence'].str.len() >= 100]

# Step 3: Filter by Phred quality score (all scores >= 30)
df_quality = df_quality[df_quality['quality'].apply(is_quality_good)]
```

Implementation on Forward Sequences

Almost half the sequences got dropped after implementing the condition.

	id	sequence	quality
0	SRR5177930.1	NTACCTTCAGGCCCCTGGACCCTTGCTCCCCAGCTGGTCCGTCCGG...	[2, 27, 27, 33, 33, 37, 37, 37, 37, 37, 37, 37...
1	SRR5177930.2	NTCCCCTCTGGGCACCTCATTCCCAGAGGCATGTAAGGCTGGAAGG...	[2, 27, 27, 33, 33, 37, 37, 37, 37, 37, 37, 37...
2	SRR5177930.3	NATGTGAACACCTGAATGAATGAGTGCCCTGAAAATATGACTGGCT...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
3	SRR5177930.4	NGCCTGTGGGCCAGGGCCAGAGCCTTCAGGGACCCTTGACTCCCCG...	[2, 27, 27, 27, 33, 37, 37, 37, 37, 37, 37, 37...

	id	sequence	quality
4	SRR5177930.5	NATTGAGACTGGCCCAACAAACATTCAATCCACTCCACCCATGGAC...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
5	SRR5177930.6	NACTCAGTTCTTTTCATGGCCAGACTCTGCCAGTCCCTGGGAGAGC...	[2, 27, 27, 27, 33, 37, 37, 37, 37, 37, 37, 37...
6	SRR5177930.7	NAAGTTCCGCACAATACTTTTCAGAAAGAGAAAAGCCATGCAGTTG...	[2, 27, 27, 33, 33, 37, 37, 37, 37, 37, 37, 37...
7	SRR5177930.8	NTCTGTTTCTATGTGGAAATAACCTCCTTCATTTCCTGATGCAAAT...	[2, 27, 27, 27, 27, 37, 37, 37, 37, 33, 14, 37...
8	SRR5177930.9	NGCCCCCTGTTCTCTAGTTGGCCTGTGCCCCTCTCCCATGTGGAGT...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
9	SRR5177930.10	NATTTCTCAAGACTTGCACATTTATATTATGCAAAACACAGCATGA...	[2, 27, 27, 33, 33, 37, 37, 37, 37, 37, 37, 37...
10	SRR5177930.11	NCTTTTTTCAGGAAACCATTGCCTACCTCAAGATTAAAAAAAAGTT...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
11	SRR5177930.12	NGCTGCACTTCAAAACTGTAAAATTAATGATCTTTGGATATTCAAT...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
12	SRR5177930.13	NACTGGATTTCAACAGGCTAAATGGCCTTTGGCGATTTCTTTCTTT...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...

	id	sequence	quality
13	SRR5177930.14	NCAGGCCAAGGTCCGCGTGCATGTGCAGGACACCAACGAGCCCCCC...	[2, 27, 27, 33, 33, 37, 37, 37, 37, 37, 37, 37...
14	SRR5177930.15	NTTACCACTGTATTAAAGATATCAGTGTCATGGTTTTCTAATTCTT...	[2, 27, 27, 33, 33, 37, 37, 37, 37, 37, 37, 37...
15	SRR5177930.16	NTCTGATGTGTGACTGATGCGGCATTCATTAATCCGATTATCAGAG...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
16	SRR5177930.17	NAGGCTCACAGCTACTTAGAGTACTAGGGTTATTCCCAGCAGAGGA...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...
17	SRR5177930.18	NCCATGGCCACCCTGCCCCCACCCCTCCAGGTTGCAGGAAGTGAAC...	[2, 27, 27, 27, 33, 37, 37, 37, 37, 37, 37, 37...
18	SRR5177930.19	GCCATAGCCATTGCCATTGCCACTTGGGGCAAAGCCATTTCCCCCA...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
19	SRR5177930.20	NGCCATCCCGCAGATCTTCATAAAGATCATTGATGTGCTTGCGGAC...	[2, 27, 33, 33, 33, 37, 37, 37, 37, 37, 37, 37...

> This is an example; in our original df (taken from the first batch of the original sequence), as it can be noticed the first good quality sequence is at ID 19 (as for the rest of the sequences, the quality is 2, 27, etc).

> And as a result to verify, the first sequence in our cleaned batch looks like this:

	id	sequence	quality
0	SRR5177930.19	GCCATAGCCATTGCCATTGCCACTTGGGGCAAAGCCATTTCCCCCA...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...

	id	sequence	quality
1	SRR5177930.28	ATGTGGGATTTTGATATTTATGGTACTGTGTCTATGTGCTGATTGT...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
2	SRR5177930.38	ACCTTTATAGGTGGGGATTAGGAGTCCCTTCTGGGCTGGGTGTGGT...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
3	SRR5177930.39	GCACAGGTAGCCAGACTCTGATCATGGCTCTGAGGAGGAGCCCTGG...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
4	SRR5177930.58	ATCCTGGGTTTTAATGCTAGGGTGGAAAGGTATTTCTGAAGCCTTG...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
...	...	...	...
51685	SRR5177930.99988	GAAAGATGTTGTTTTTGGTGAGTTTGACGCTTTTGGGCCTTGGGTG...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
51686	SRR5177930.99989	ATGCCGTGGGTTATTTCCTAAGGTTTCCTAGGTTATAGCCTAACCT...	[33, 33, 33, 33, 33, 37, 37, 37, 37,

	id	sequence	quality
			37, 37, 3...
51687	SRR5177930.99995	TAATCGTTTCATATATGATGGAATTGACAGCAACTTTGAACCTGAG...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
51688	SRR5177930.99996	ACCACAATTCCAGAAAATGACATAGAGAAGACTGACCCTTGGTTTG...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 37, 37, 3...
51689	SRR5177930.100000	CGCCCCCCTGCCCCTGCACCCTCACACCCATCTTCCTCTCTCAGCC...	[33, 33, 33, 33, 33, 37, 37, 37, 37, 33, 37, 3...

51690 rows × 3 columns

> i.e the code is working well.

Implementation on Backward Sequences

In backward sequences, very less records are of good quality. Mostly resampling shall be required during training.

```
Number of records with quality < 30: 82935
```

Almost 82K out of 100000 are of low quality.
The average number of records for backward sequence in each batch is ~15-20K.

Conclusion

The cleaned records have been stored as parquet batches in two folders:

```
dna_sequencing/clean_forward_reads
dna_sequencing/clean_backward_reads
```

The code for cleaning these records can be found here:

```
dna_sequencing/cleaning_sequences.ipynb
```

Steps 3 and 4 shall be implemented in the next report.

Updated folder structure as of 30 May, 2025, after working on the non-cancerous readings:

Code for Non-Cancerous readings handling: `dna_sequencing/Anushka/noncan_readings.ipynb`

- the only change is after quality cleansing, there were ~2 files that got dropped completely due to the records being completely empty. Also talking about the sequence, this sequence has a mix of 100 and 101 bp.
- So we have accordingly selected only those sequences that have 100bp.

The sequences can be found here:

`dna_sequencing/clean_forward_reads`

`dna_sequencing/clean_backward_reads`

`dna_sequencing/clean_forward_noncan`

`dna_sequencing/clean_backward_noncan`