What about amplicons?

For example, amplified variable regions of 16S

Goldilocks and the Three pyrosequenced 16S amplicon datatasets

Datasets from 3 different experiments
All are amplicons of the same V4-V5 region,
of ~ 375 b.p., not including linker/primers ;
Including calibration key, barcode, linker/primers:
expected length is ~ 430 b.p.

454 …?

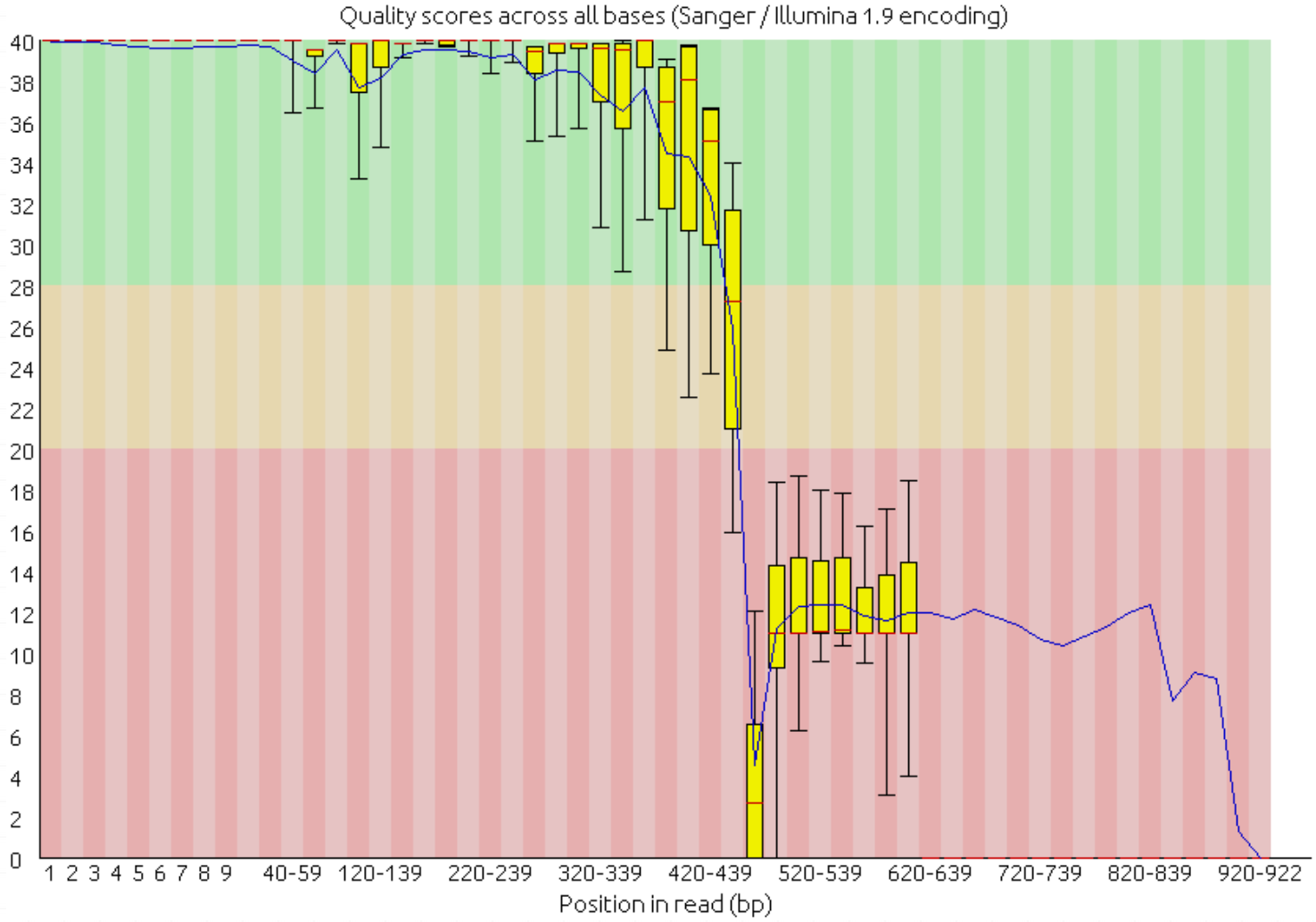
- So why are we bothering looking at 454 data? No-one uses that any more?
- Well, it's true that these days you are much more likely to be looking at Illumina data sets than 454.
- But it's quite possible you may need to reanalyse older datasets from your research group
  - And possibly even process 3rd-party, published 454 datasets
- Aside from that – 454 datasets are quite instructive
  - If you can understand the issues with those, then you'll probably be fine with any Illumina datasets you are likely to analyse

These are all real datasets

From different projects within IFR-GHFS

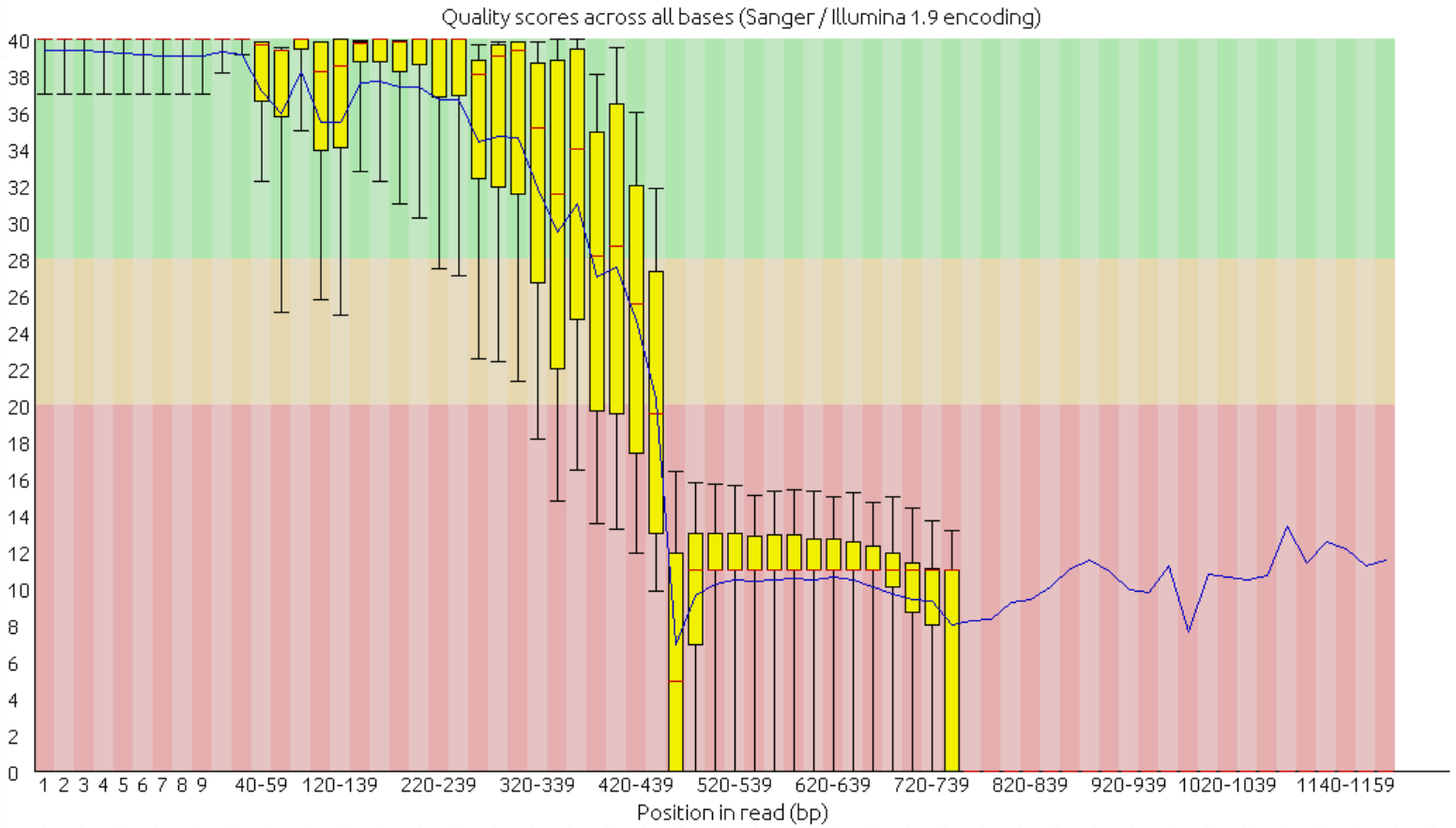
115,210 reads
Length range 69-922 b.p. (mode **462** b.p.)

A



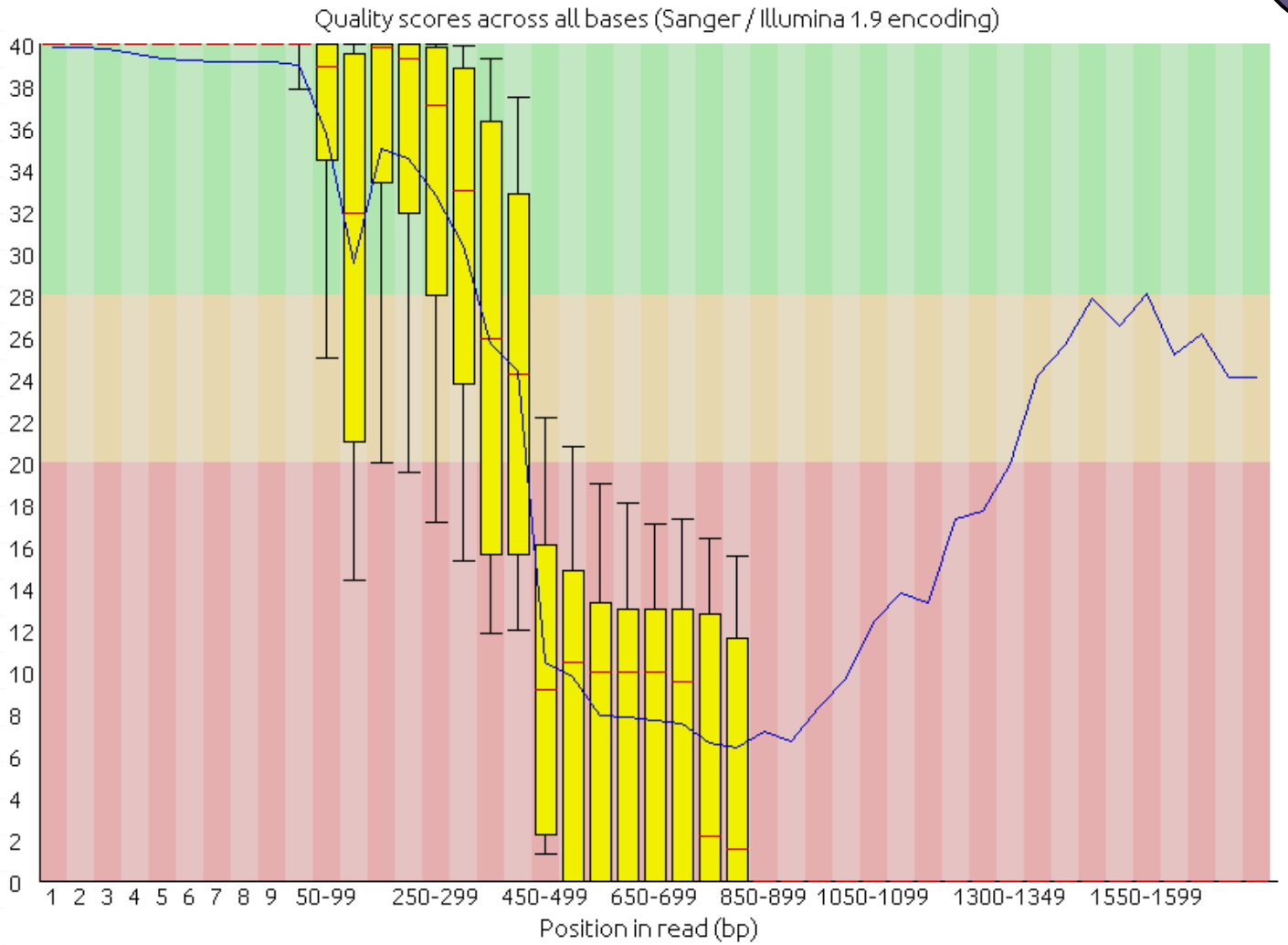
44,909 reads
Length range 82-1199 b.p. (mode **461** b.p.)

B

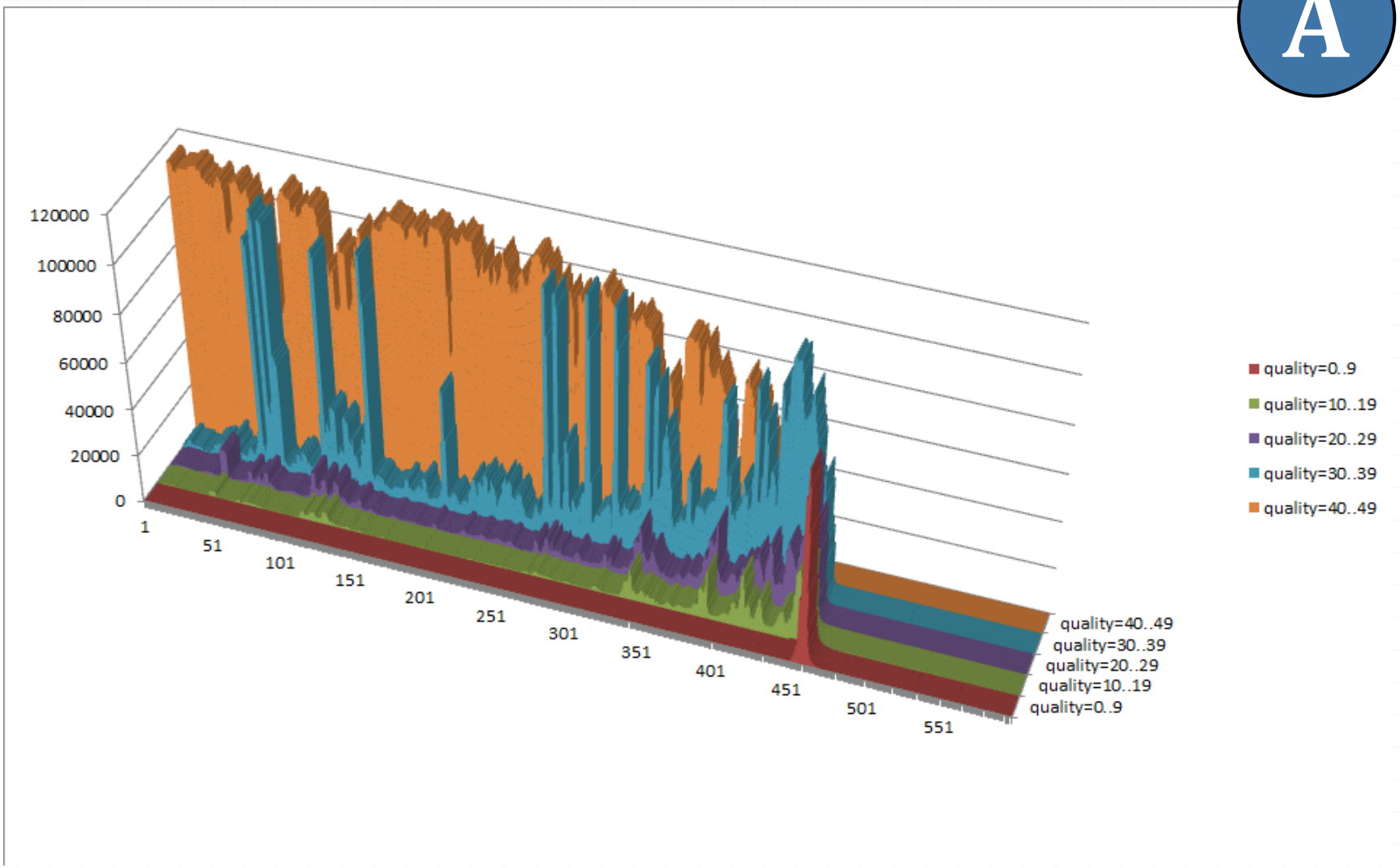


44,909 reads
Length range 105-1779 b.p. (mode **469** b.p.)

C

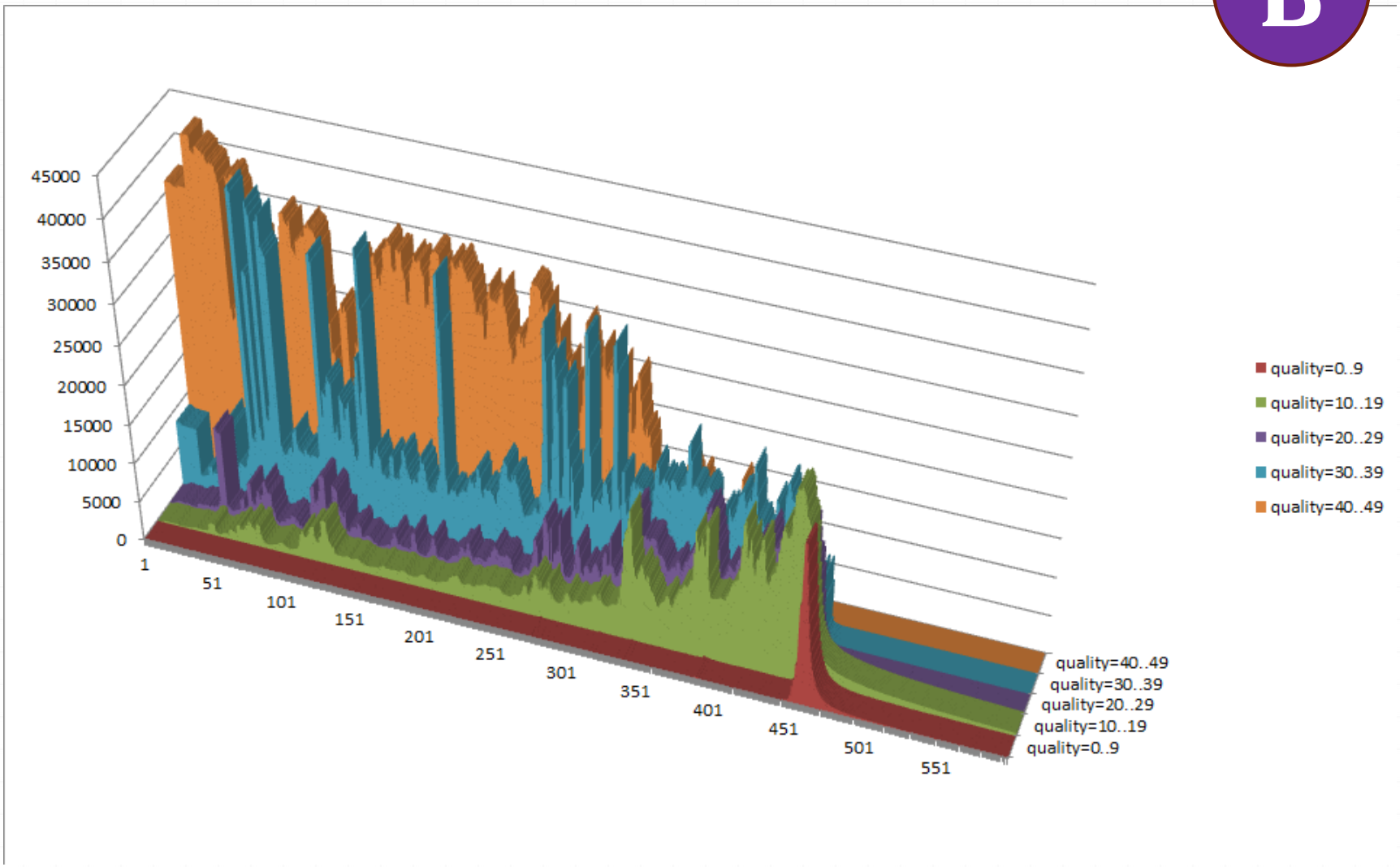


Frequencies of quality scores at each b.p. position



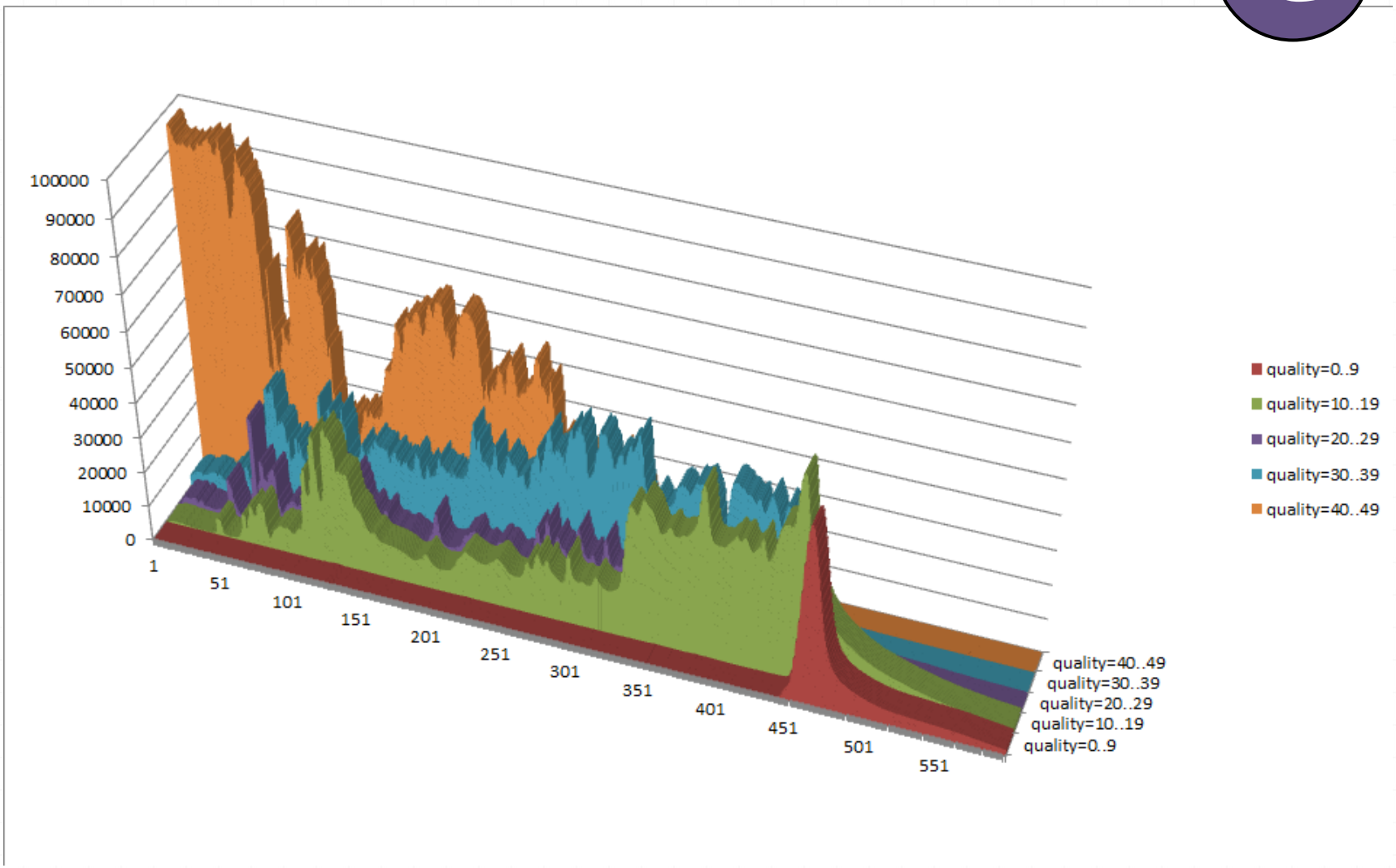
Frequencies of quality scores at each b.p. position

B



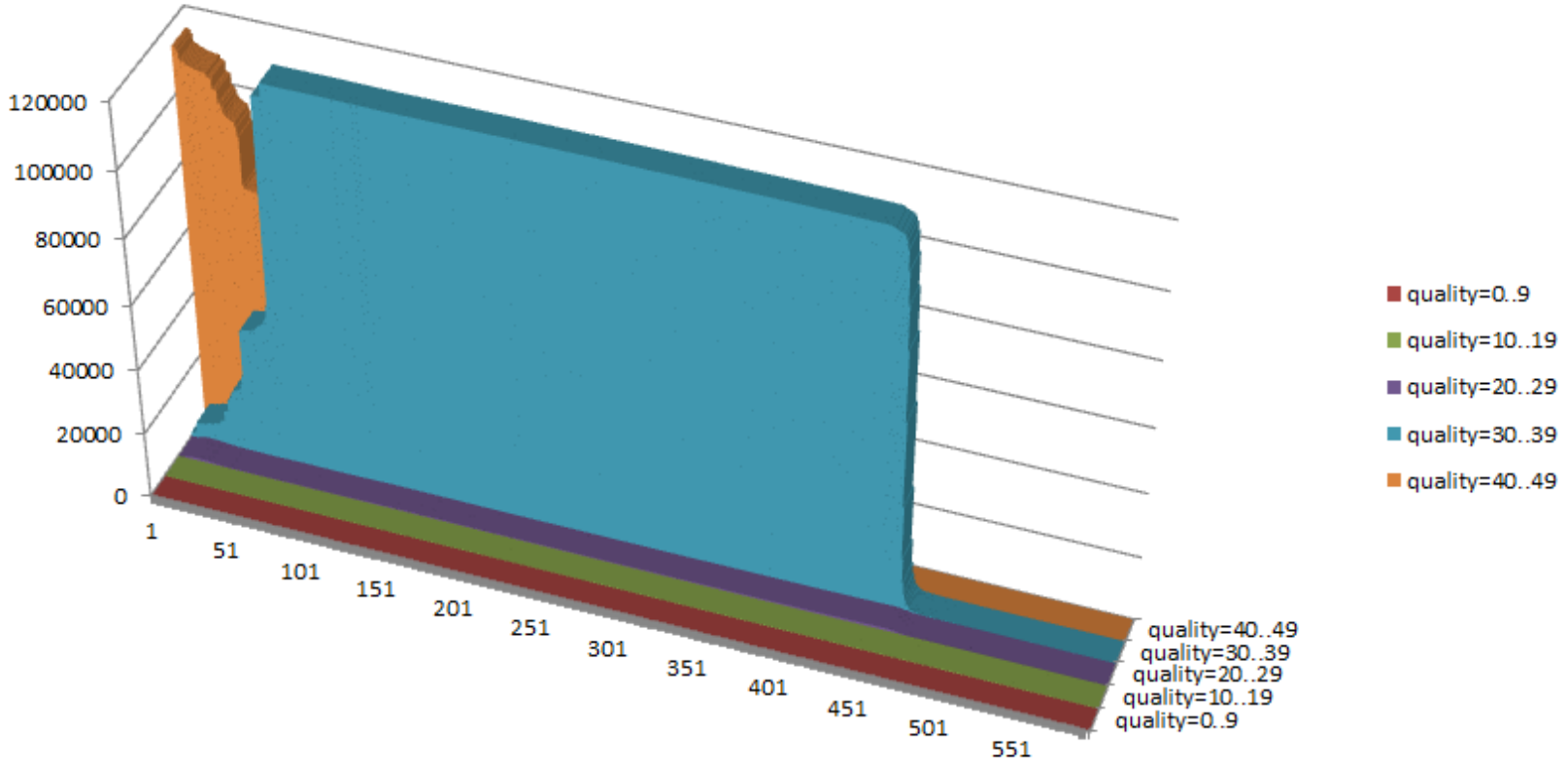
Frequencies of quality scores at each b.p. position

C



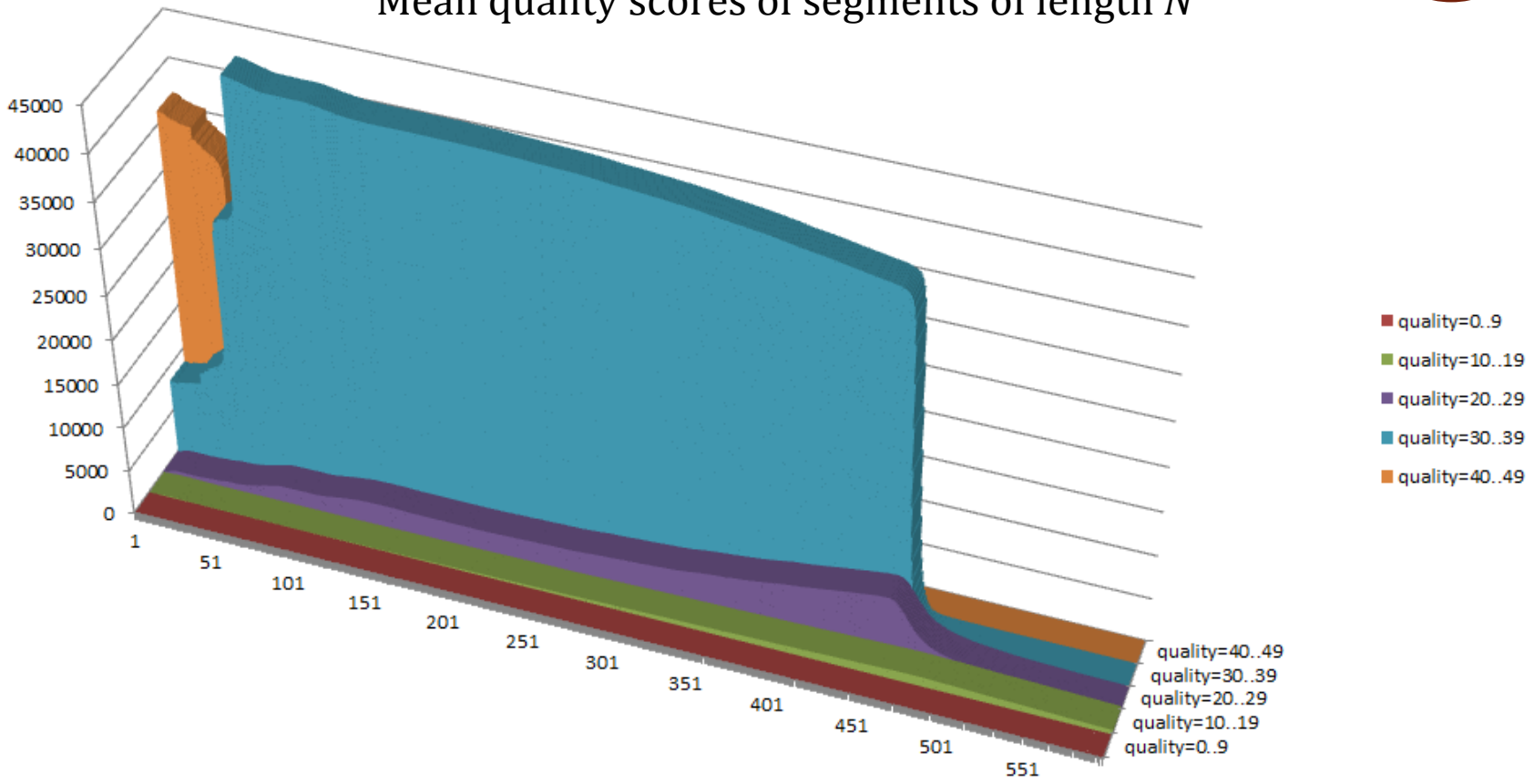
A

Frequencies of :
Mean quality scores of segments of length *N*



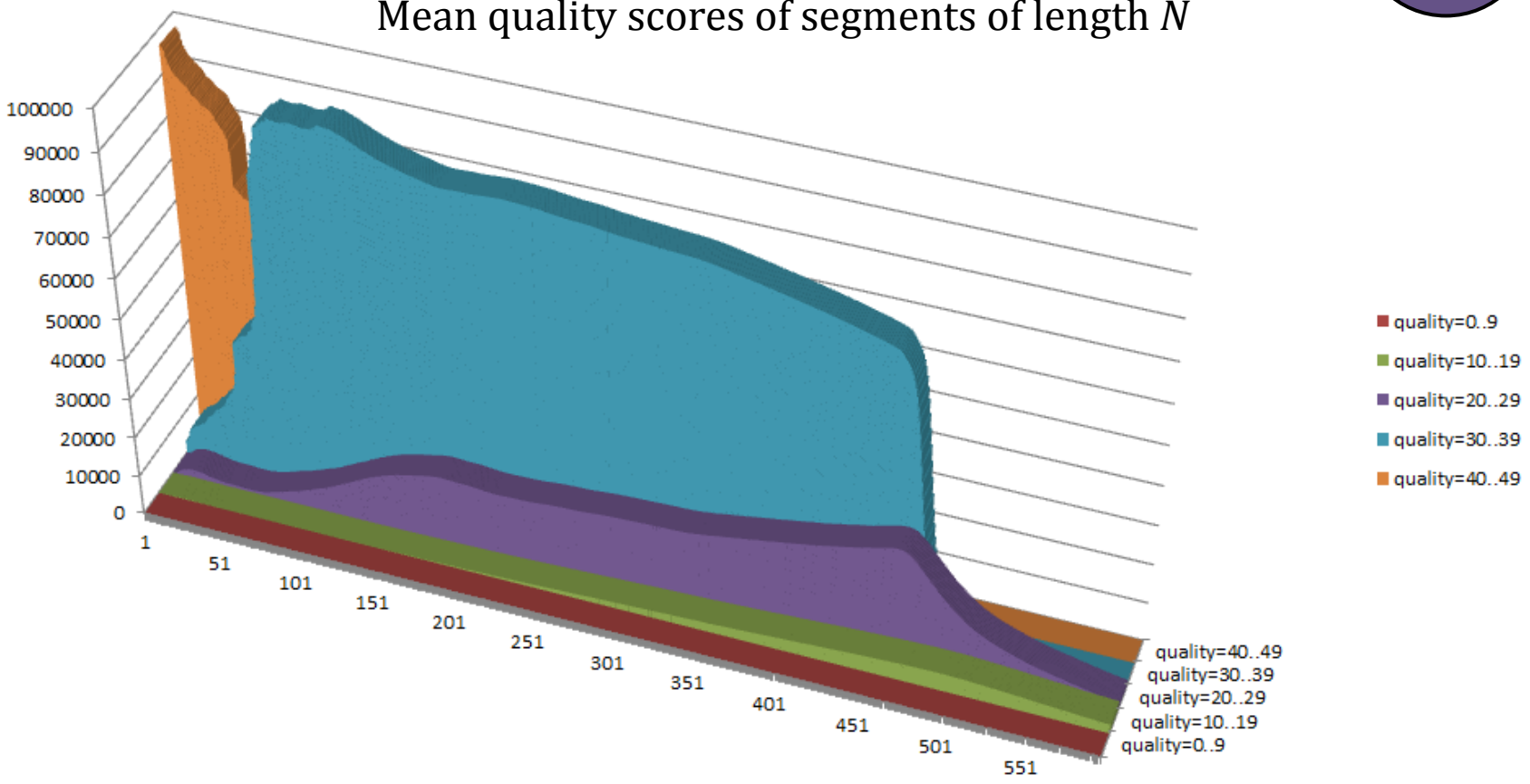
B

Frequencies of :
Mean quality scores of segments of length *N*



C

Frequencies of :
Mean quality scores of segments of length *N*



So are any of these datasets "just right"?

- Dataset 'A' is extremely good – about as good a set of 454 16S amplicon sequence data that you will see.
- 'B' is not of as good quality overall
- 'C' is poorer still
- However, 'B' and 'C's overall metrics are **very likely** the result of a significant number – *still a minority* – of very poor reads
  - ***Not*** by all or most of the reads being polluted by a significant number of poor-quality bases
- **This means it's likely a large number of reads are of <u>very good quality throughout the length of the amplified region</u>**
  - **10s of thousands of reads in both 'B' and 'C'**
  - (This is confirmed by examining post-trimming length distributions, etc. – not shown here)