Informatics on High-throughput Sequencing Data

(Summer Course 2020)

Day 12

Notes

REMEMBER

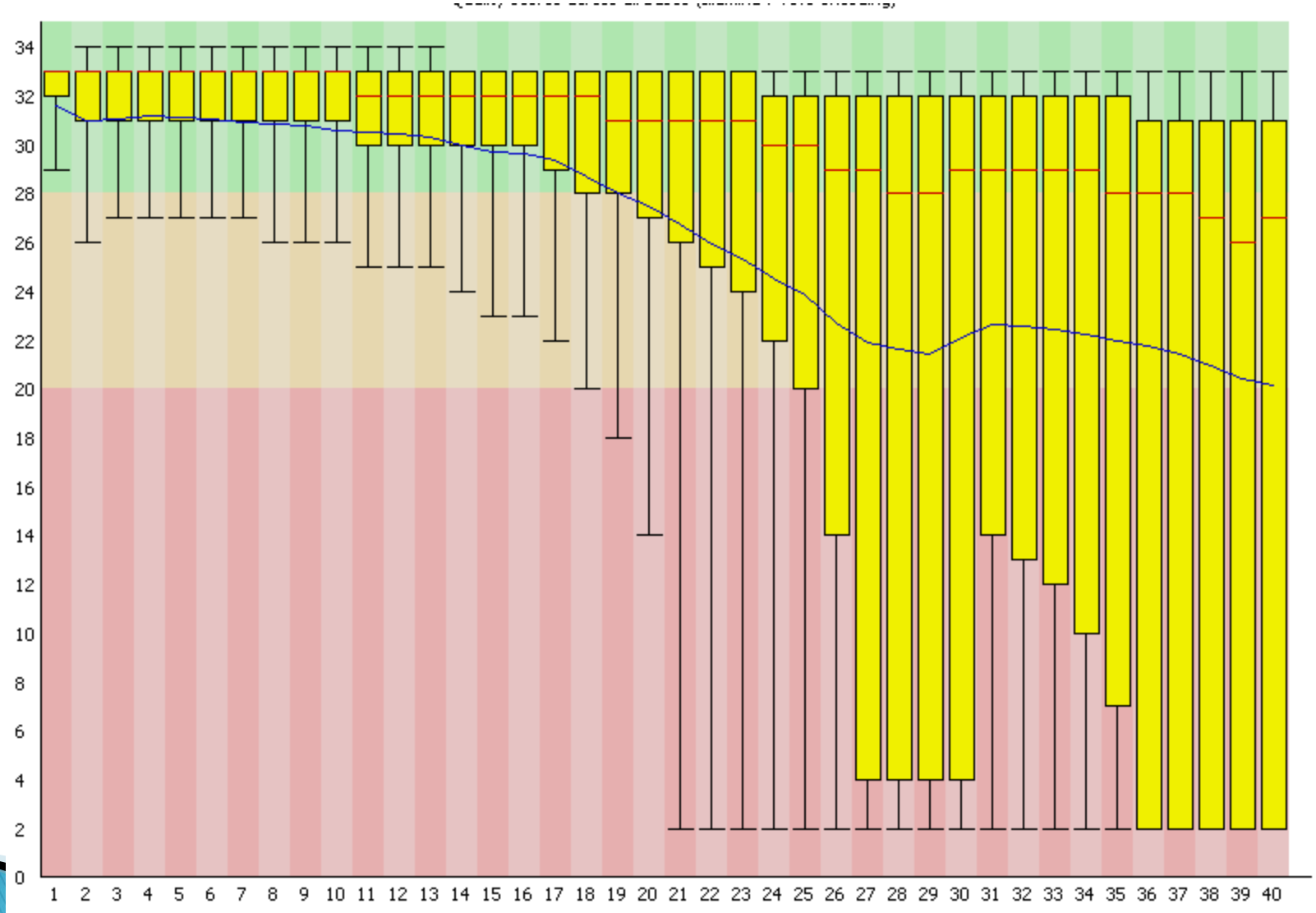
If one side of the box is longer than the other, it does not mean that side contains more data. In fact, you can't tell the sample size by looking at a boxplot; it's based on percentages of the sample size, not the sample size itself. Each section of the boxplot (the minimum to Q_1, Q_1 to the median, the median to Q_3, and Q_3 to the maximum) contains 25% of the data no matter what. If one of the sections is longer than another, it indicates a wider range in the values of data in that section (meaning the data are more spread out). A smaller section of the boxplot indicates the data are more condensed (closer together).

REMEMBER

Notice that the *IQR* ignores data below the 25th percentile or above the 75th, which may contain outliers that could inflate the measure of variability of the entire data set. So if data is skewed, the *IQR* is a more appropriate measure of variability than the standard deviation.

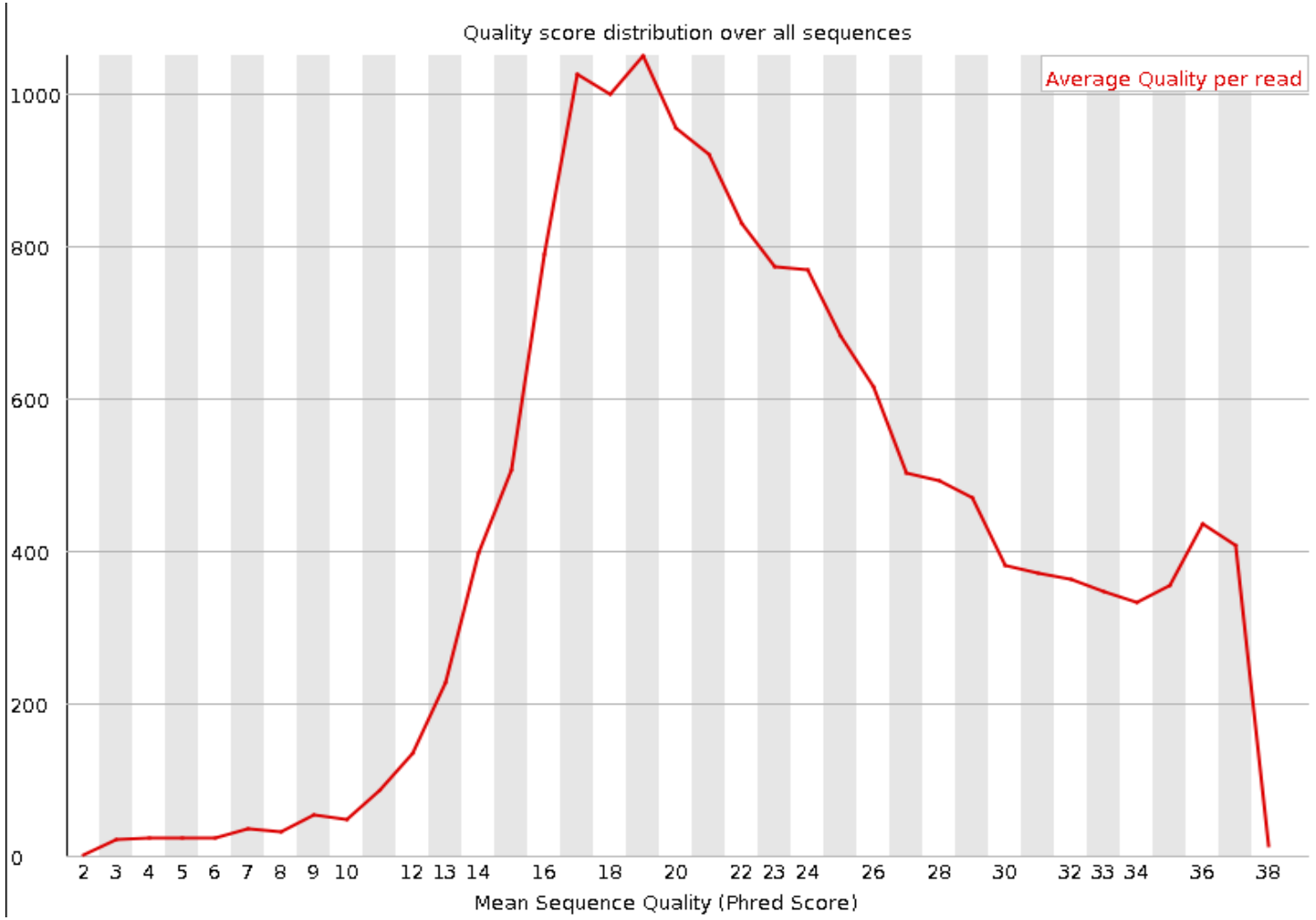
https://www.dummies.com/education/math/statistics/what-a-boxplot-can-tell-you-about-a-statistical-data-set/

FastQC



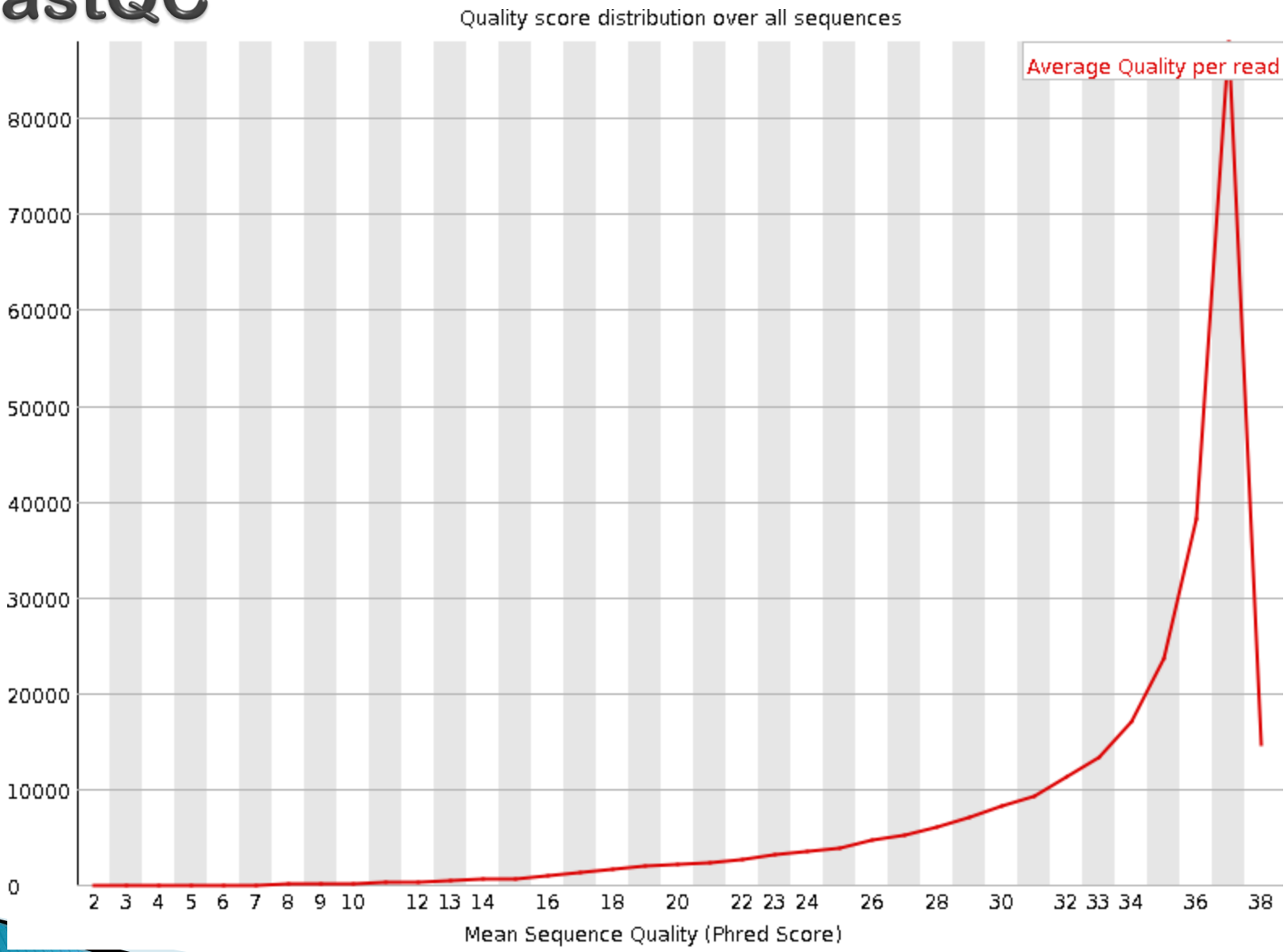
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html

FastQC



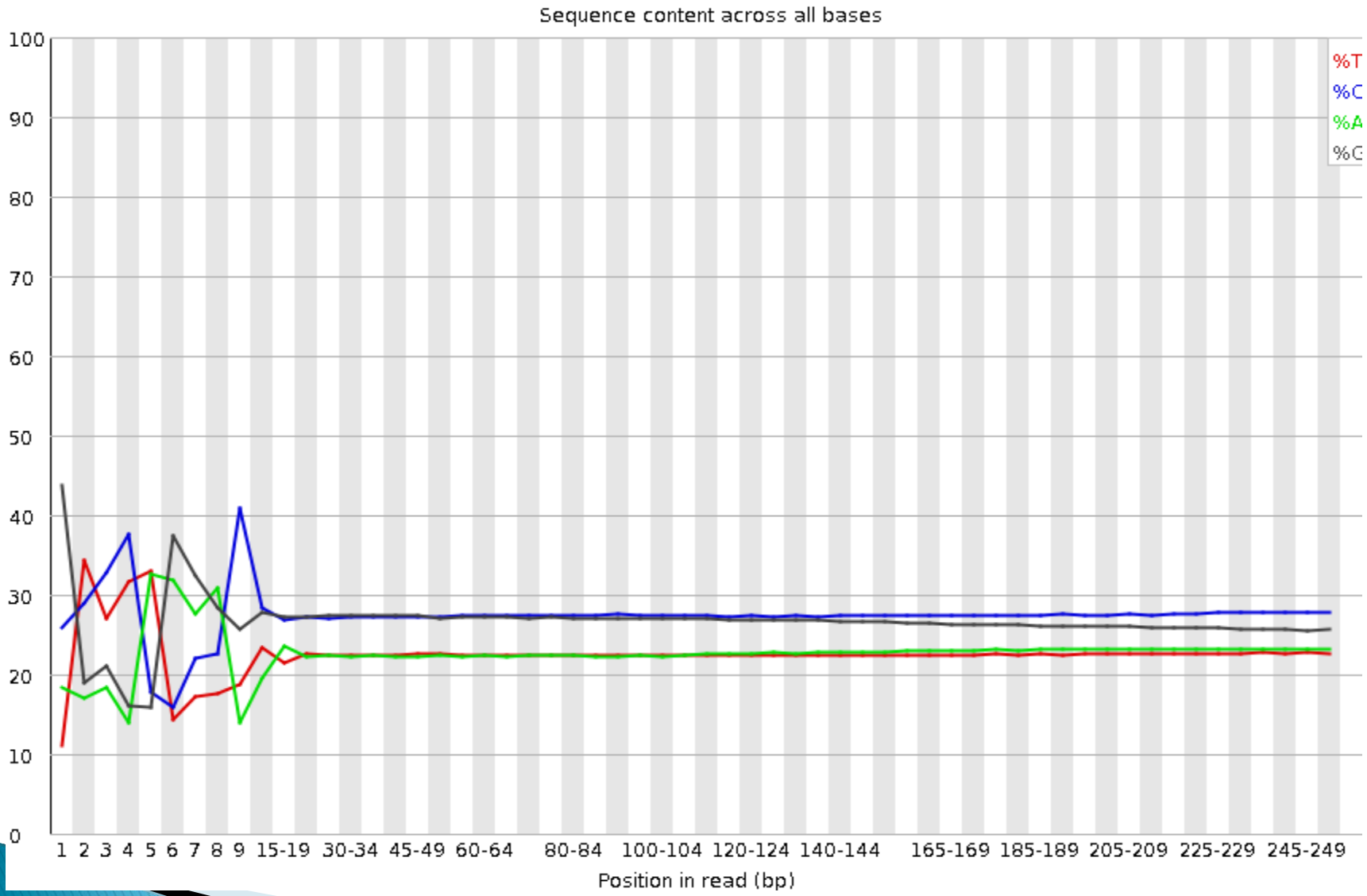
https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC



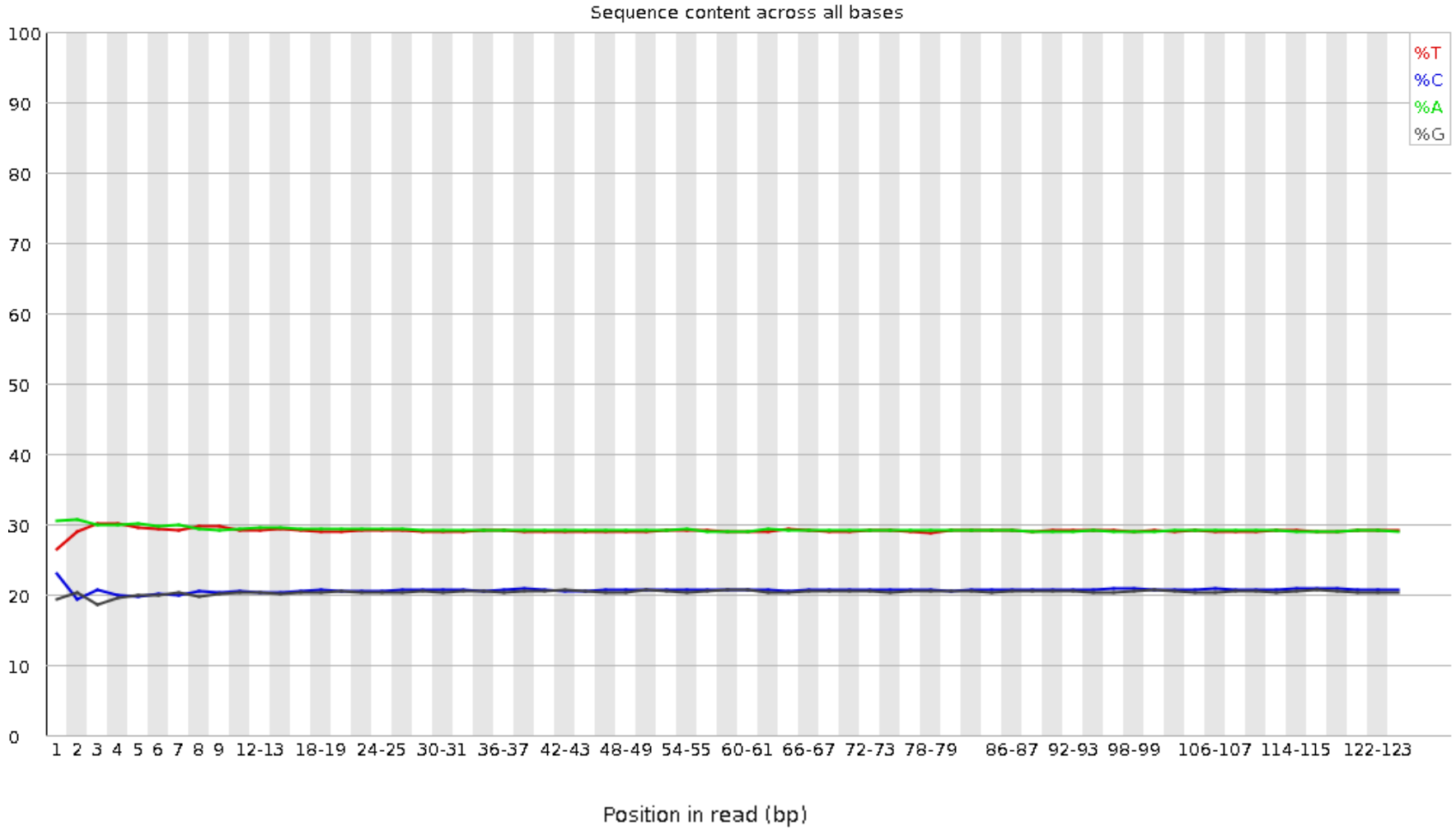
https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC



https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC



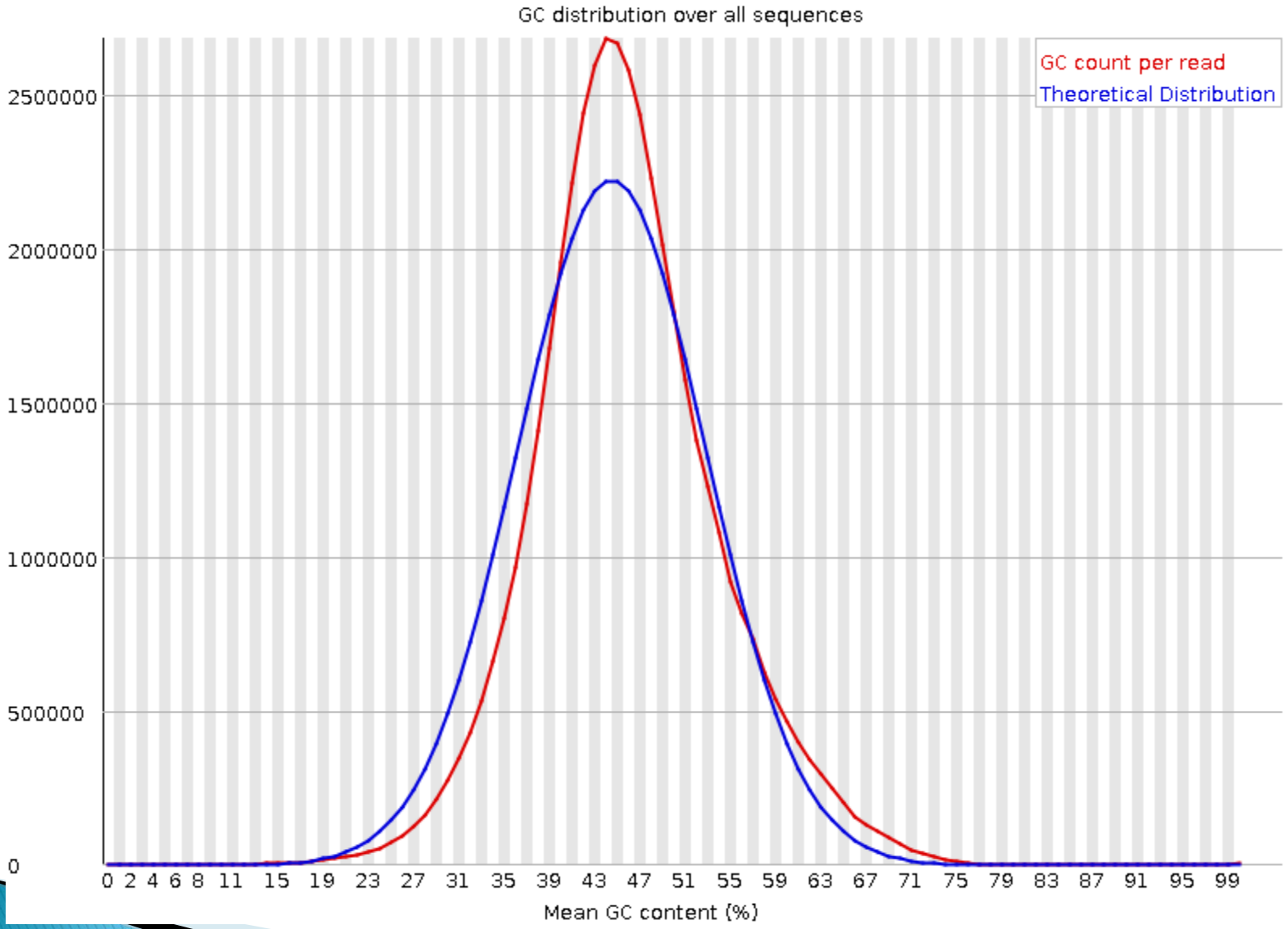
https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

Notes

- For whole genome shotgun DNA sequencing the proportion of each of the four bases should remain relatively constant over the length of the read with %A=%T and %G=%C.
- With most RNA-Seq library preparation protocols there is clear non-uniform distribution of bases for the first 10-15 nucleotides; this is normal and expected depending on the type of library kit used (e.g. TruSeq RNA Library Preparation)
- RNA-Seq data showing this non-uniform base composition will always be classified as Failed by FastQC for this module even though the sequence is perfectly good.

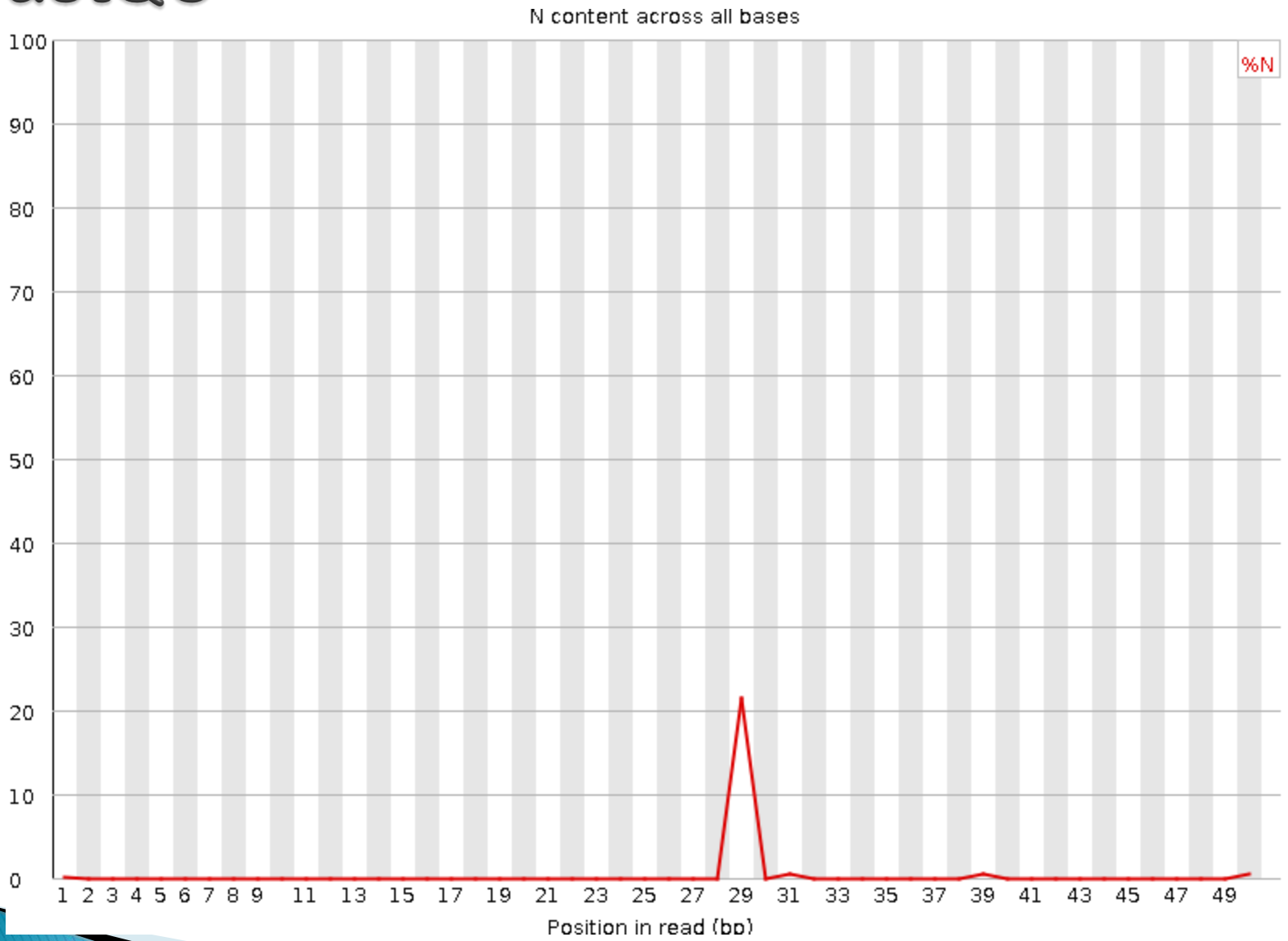
https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC



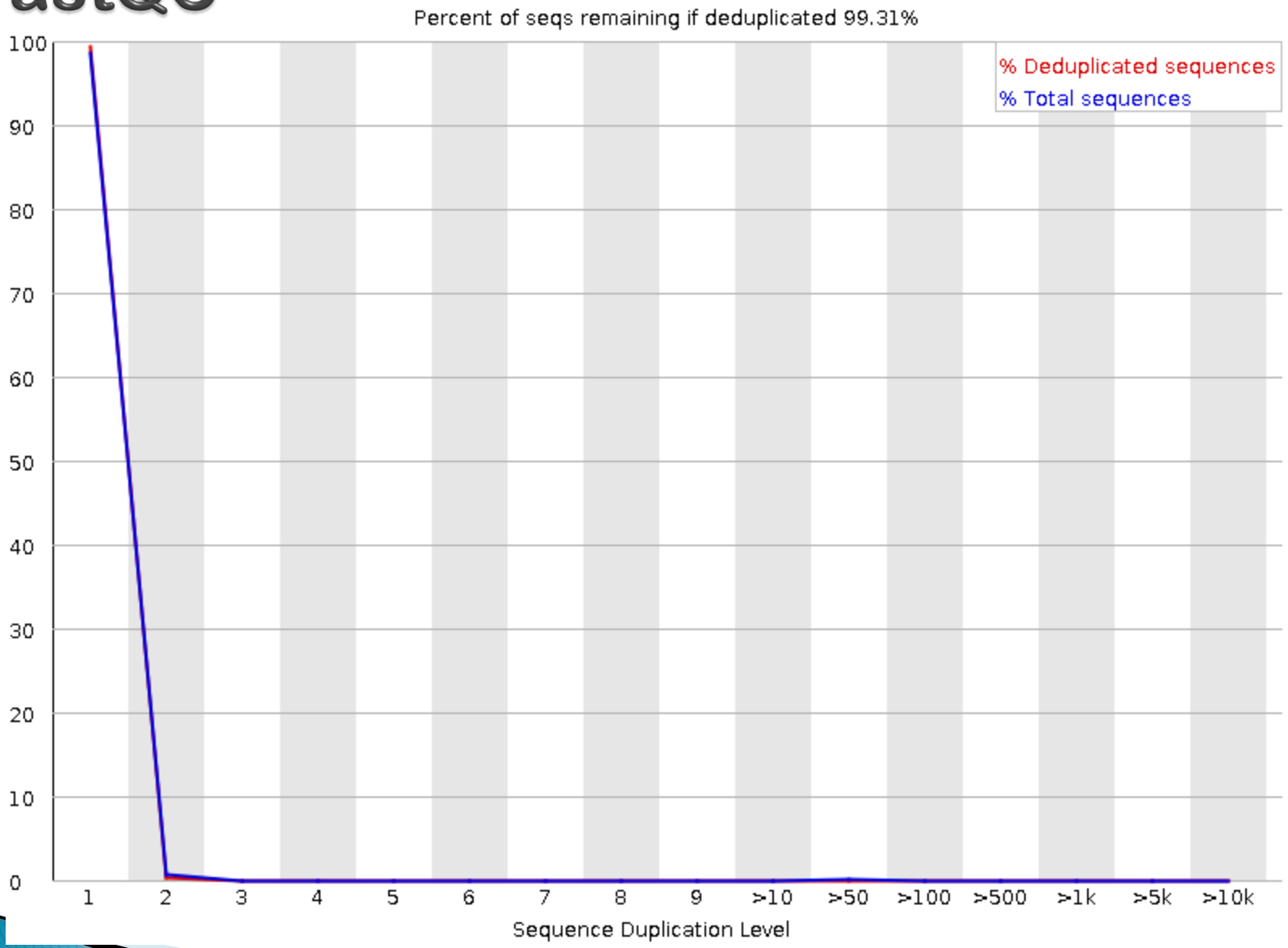
https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC



https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC



https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

FastQC

- There are generally two sources of duplicate reads:
  - PCR duplication in which library fragments have been over represented due to biased PCR enrichment or
  - truly over represented sequences such as very abundant transcripts in an RNA-Seq library.

https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

Run FastQC in non-interactively mode !!

Trimmomatic

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Downloading Trimmomatic

Version 0.39: binary, source and manual

Version 0.36: binary and source

Trimmomatic is a fast, multithreaded command line tool that can be used to trim and crop Illumina (FASTQ) data as well as to remove adapters. Trimmomatic works with FASTQ files (compressed vs. decompressed) (paired vs. single) (using phred + 33 or phred + 64 quality scores, depending on the Illumina pipeline used).

http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic

- What the sliding window trimming does is scan along the read in the 5'->3' direction calculating the average quality over the window size you have defined.
- Once it encounters a window with an average quality below the threshold you have specified it removes all bases in that window and beyond to the 3' end of the read, effectively trimming the read back to the point just before the sliding window fell below your threshold.

- `java -jar trimmomatic-0.39.jar SE SRR* output.fastq SLIDINGWINDOW:4:20 MINLEN:36`

http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic

- `java -jar trimmomatic-0.39.jar SE SRR* output.fastq SLIDINGWINDOW:4:20 MINLEN:36`

ACCCTTTGGTTTAATTACCTTTTATTTTTTATTTTTTTATTT

Avg Q < 20

ACCCTTTGGTTTAATTACCTTTTATTTTTA

Len < 36

http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic

- `java -jar trimmomatic-0.39.jar SE SRR* output.fastq ILLUMINACLIP:ad.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:36`

ACCCTTTGGTTTAATT
AGTCTTTGGT

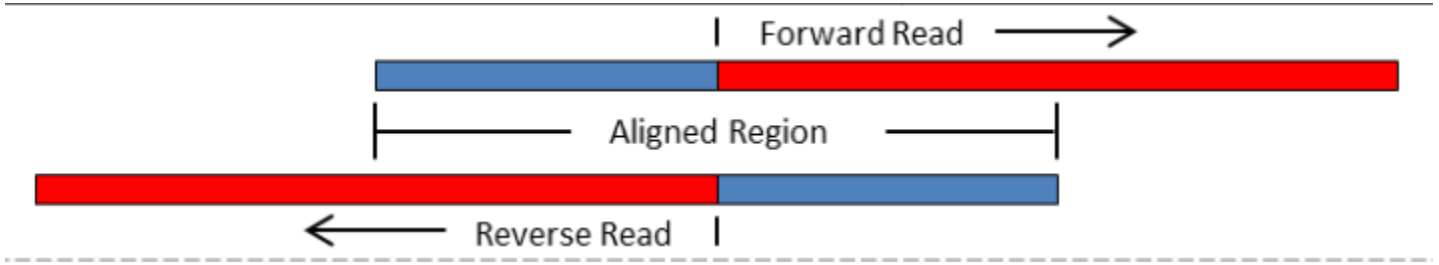
Allow for maximum 2 matches

ACCCTTTGGTTTAATT
AGTCTTTGGT

Alignment Score
match= +0.6
Mismatch = -Q/10

http://www.usadellab.org/cms/?page=trimmomatic

Trimmomatic



http://www.usadellab.org/cms/?page=trimmomatic

Thanks!

// | ?