GoSeqIt

Ex. 3 Recap

Q1. How many reads (sequences) do each of the two files contain?

Answer: 500.000

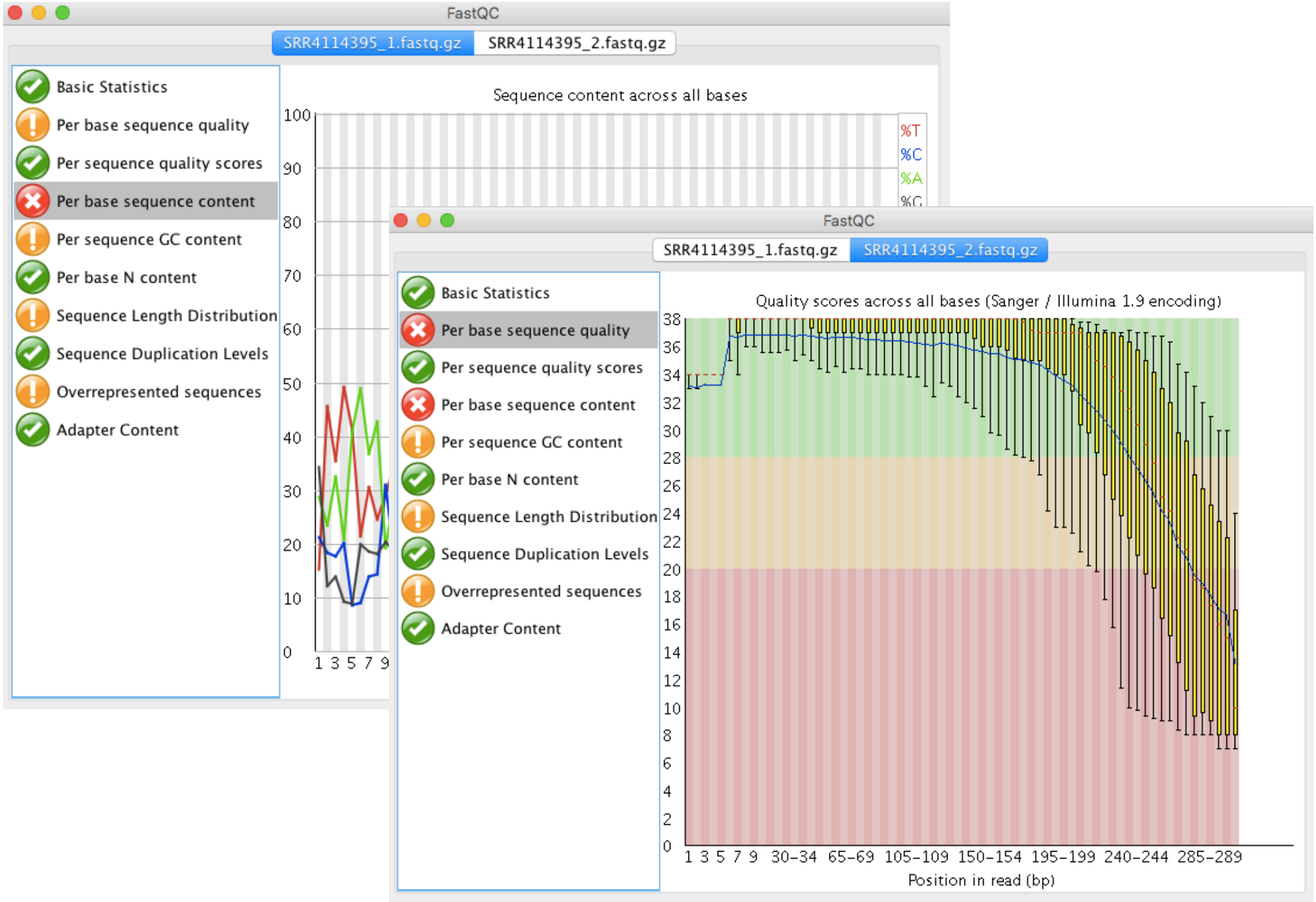
Q2. Which FastQC metrics should you be most concerned about? (Which metrics are flagged by FastQC as highly problematic via a red circle with a white “x” in it).

Answer:

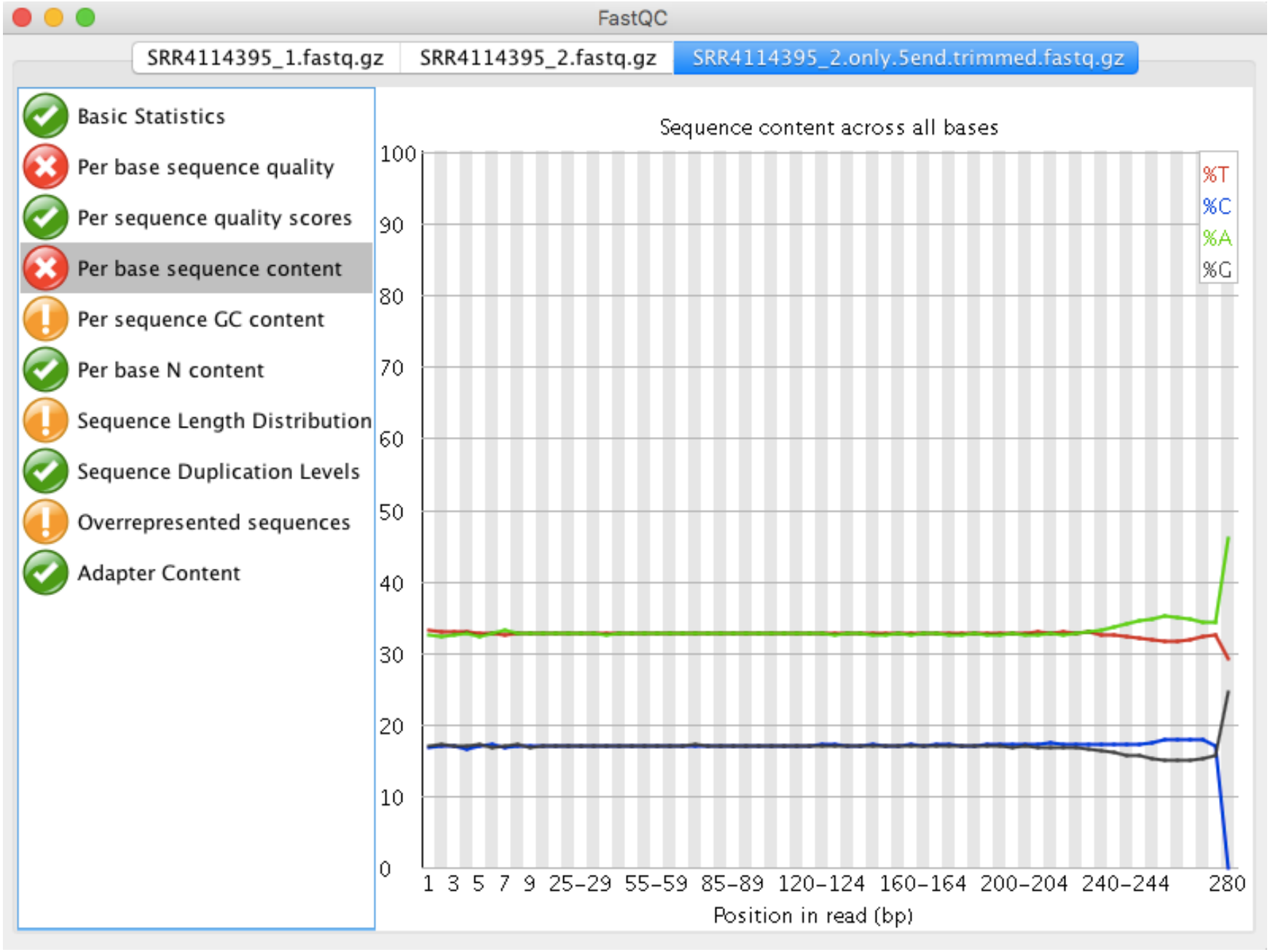
SRR4114395_1.fastq.gz: only “Per base sequence content”

SRR4114395_2.fastq.gz: “Per base sequence quality” and “Per base sequence content”

Before trimming



Q3: Use FastQC to validate that the reads in SRR4114395_2.only.5end.trimmed.fastq.gz are now 20 nucleotides shorter. Do the 5'ends according to the "Per base sequence content" look better?



Q4. How many reads are there in each of the two files and what is the mean read length after trimming?

Reads = 474123 (in both files). Mean length R1: 179.59 bp, mean length R2: 158.28 bp

Q5. Make a rough calculation of the depth of coverage based on the read statistics:

Depth = (no. of reads * av. read length) / genome size

Use an approximate genome size of 3.000.000 bp for *S. aureus*.

Answer: Depth = ((2*474123) * ((179,6+158,3)/2)) / 3.000.000 = 53

SQ1

The figure below shows a section of a fastq file. Has Phred+33 or Phred+64 been used to encode the quality scores?

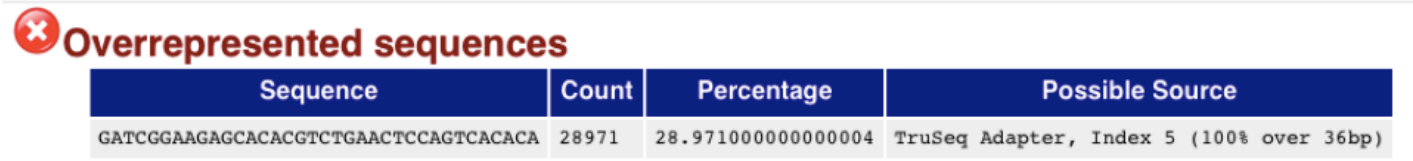
Answer: Phred+33 which can be seen by, for instance, the characters “?”, ”<“, “+”.

Cutadapt’s default behaviour is to assume your files use Phred+33, but if they use Phred+64, add the following to your command:

--quality-base=64

SQ2

SQ2. Now, let's pretend that FastQC found an adapter sequence in you reads:



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	28971	28.971000000000004	TruSeq Adapter, Index 5 (100% over 36bp)

Look into the documentation of Cutadapt's documentation (http://cutadapt.readthedocs.io/en/stable/guide.html#removing-adapters) to figure out what you should additionally add to the command to remove the adapter sequence if it is found in either the 5' or 3' end.

Answer: -b GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA