Read Quality Assessment

DECRYPT Workshop - Hands on session

0. Prior to the Quality Check...

We will use a small set of paired-end reads.

- Create a new folder in your home directory

```
mkdir ~/QC_session
cd ~/QC_session
```

- Copy this set of reads in it. Use the following command lines:

```
gunzip -c /netscratch/common/MPIPZ_SPP_workshop/RNA-Seq/RNA-Seq_smallraw_data/Cc
```

1. Fastq files

Example of a .fastq file

```
@J00137:166:H5LGGBBXY:5:1101:7770:1666 1:N:0:TCTCTTCA+TCGAAGTG
CCCGATTCGGGCCCTGAGGCTCAGCACTCGGTCCCCAGCCTCCAAATCCTCCTGCGGCAGCTGCTCCTCCCAGGCCGCTC
+
AAFFFJJFFFAA<FJJJJAJJJJJJ-<JAAJ7JJJFJ<FJ7JJFJFAA<7FJJJFJ-F<AFJ<AJJJJJJJJJJJJAJFJ
```

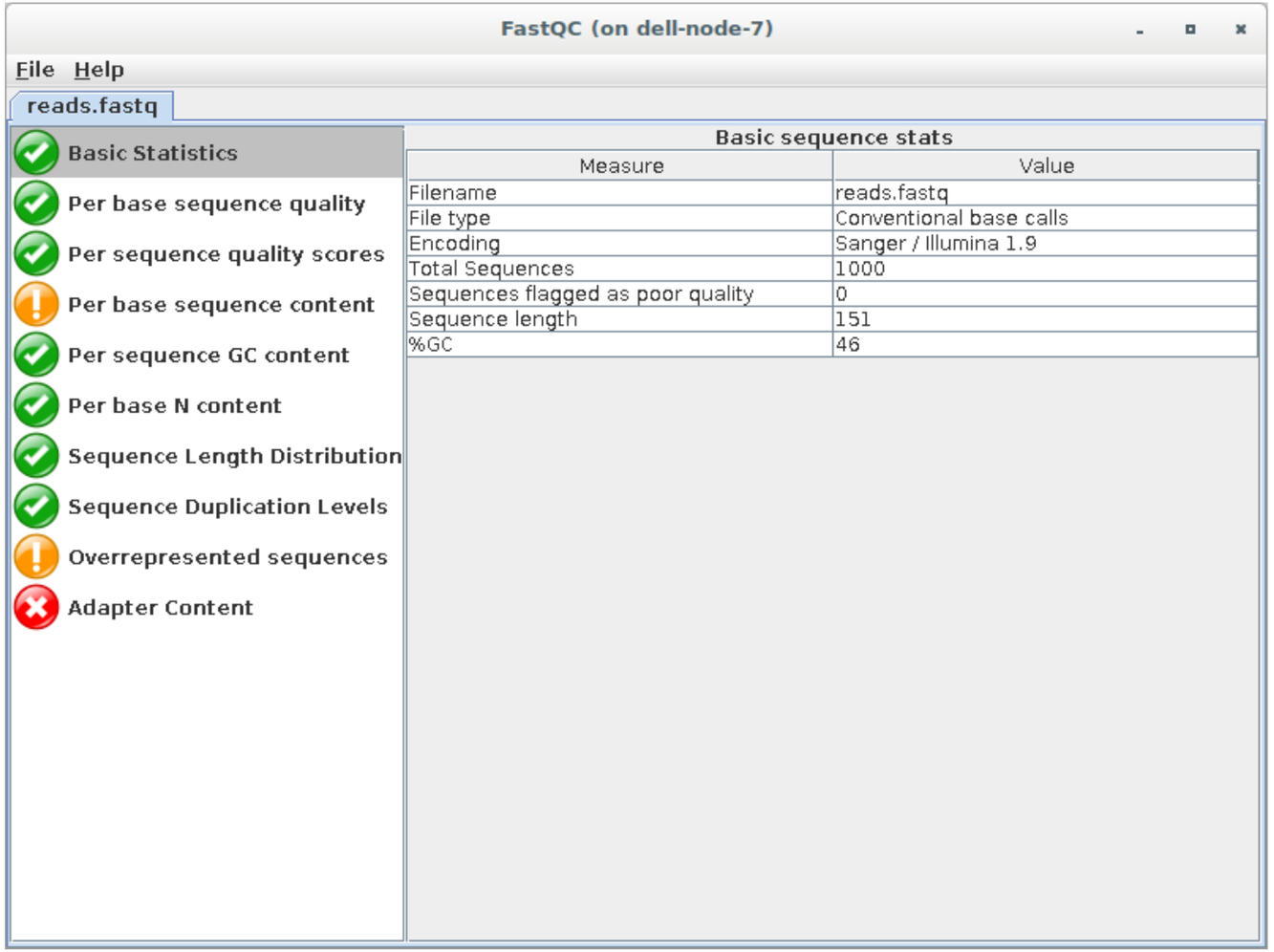
- Line 1: header
- Line 2: read sequence
- Line 3: may contain the same sequence identifier or a description
- Line 4: ascii-encoded quality values for each single base of the sequence

2. Assessing the global quality of a RNAseq dataset

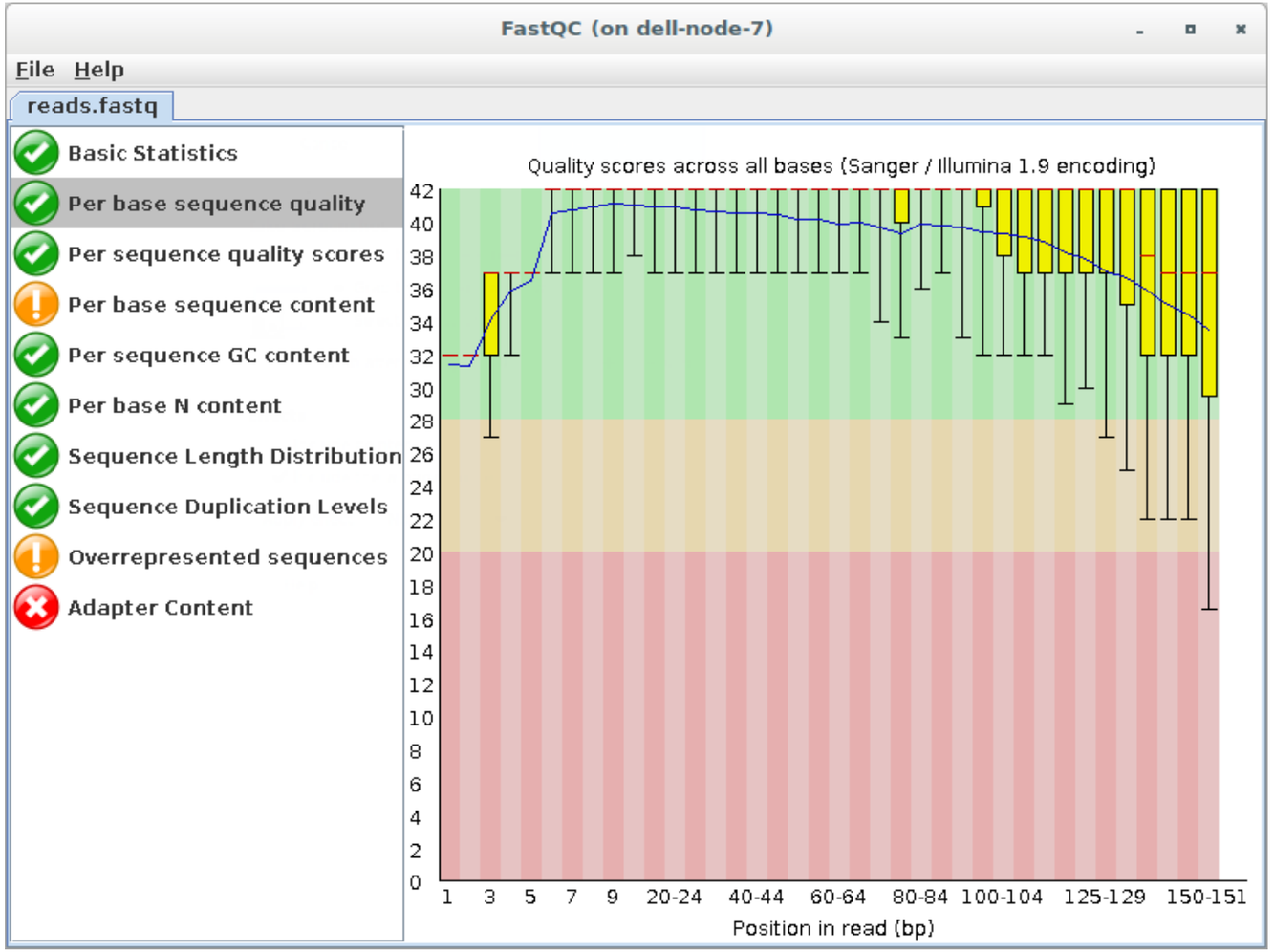
Display statistics about the reads included in a FASTQ file

```
fastqc
```

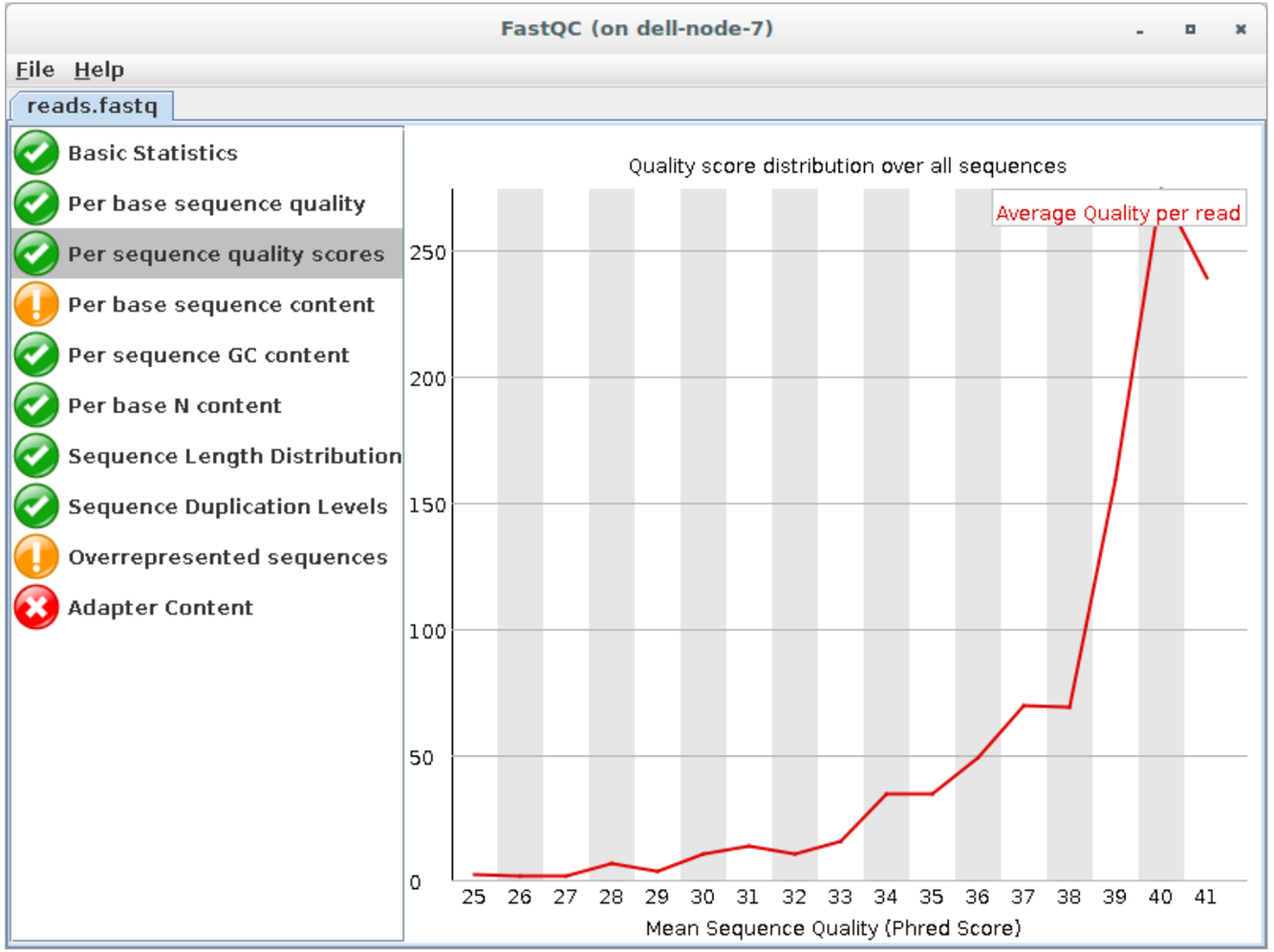
Basic statistics



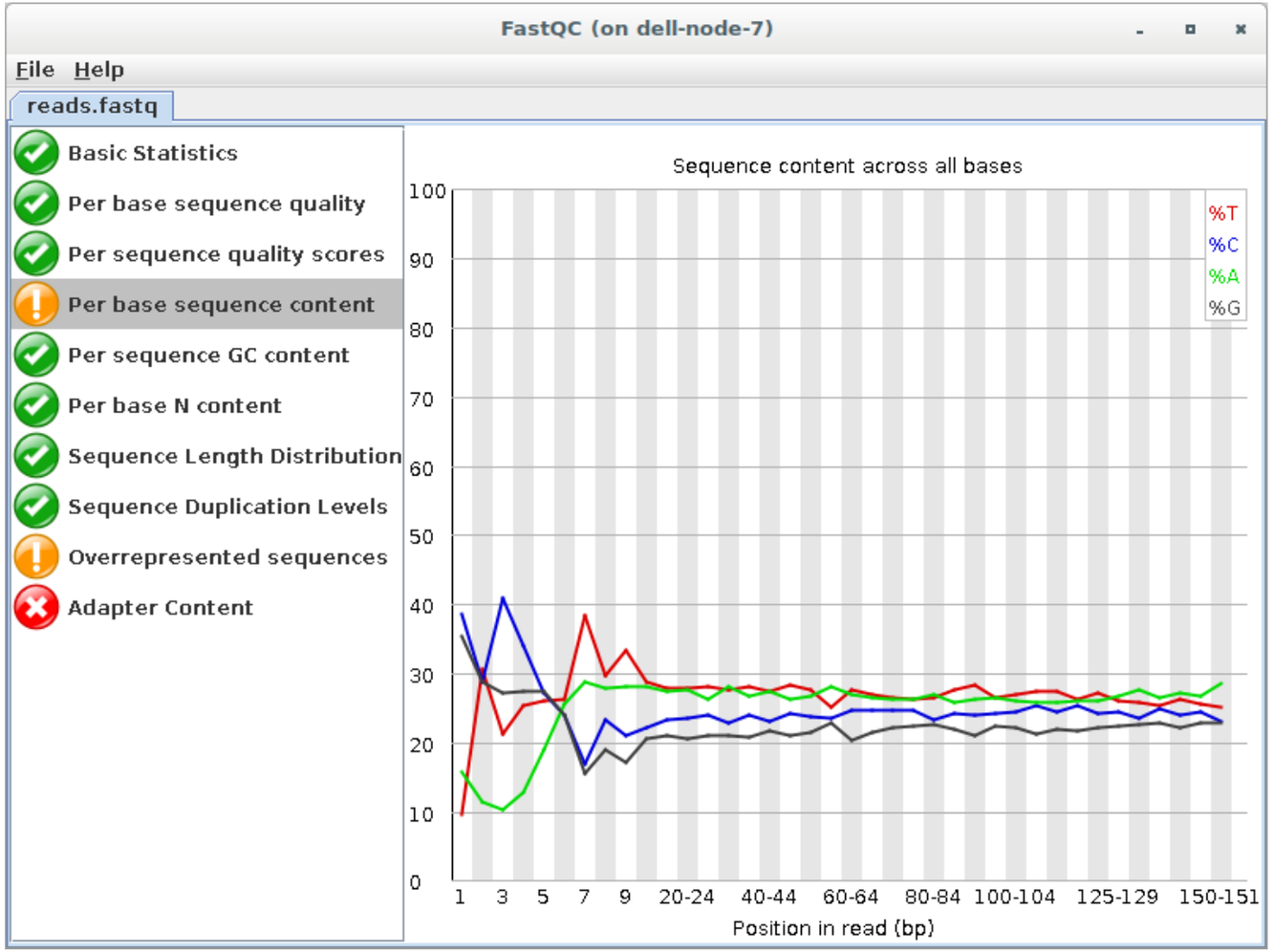
Per Base Sequence Quality



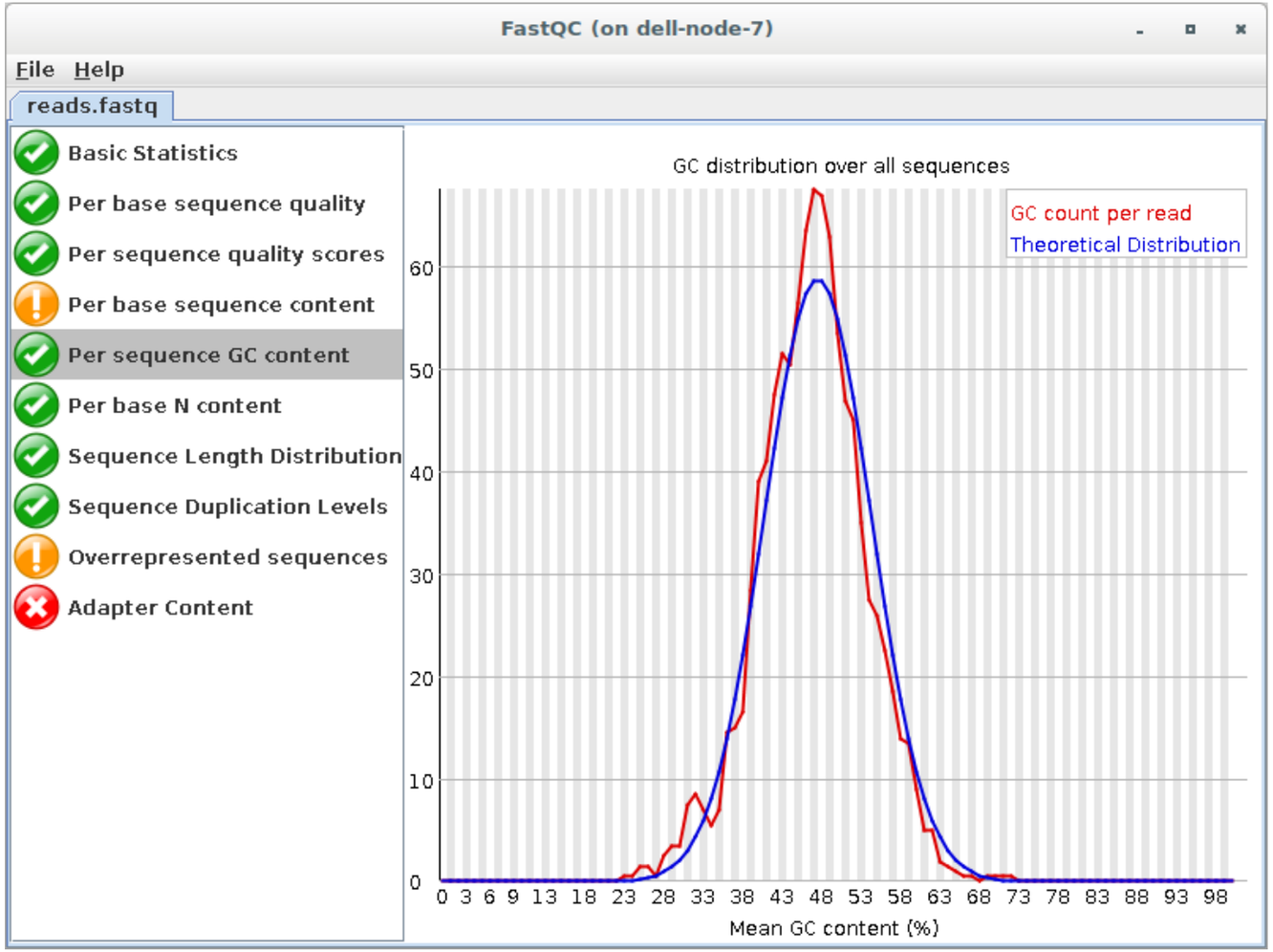
Per Sequence Quality Scores



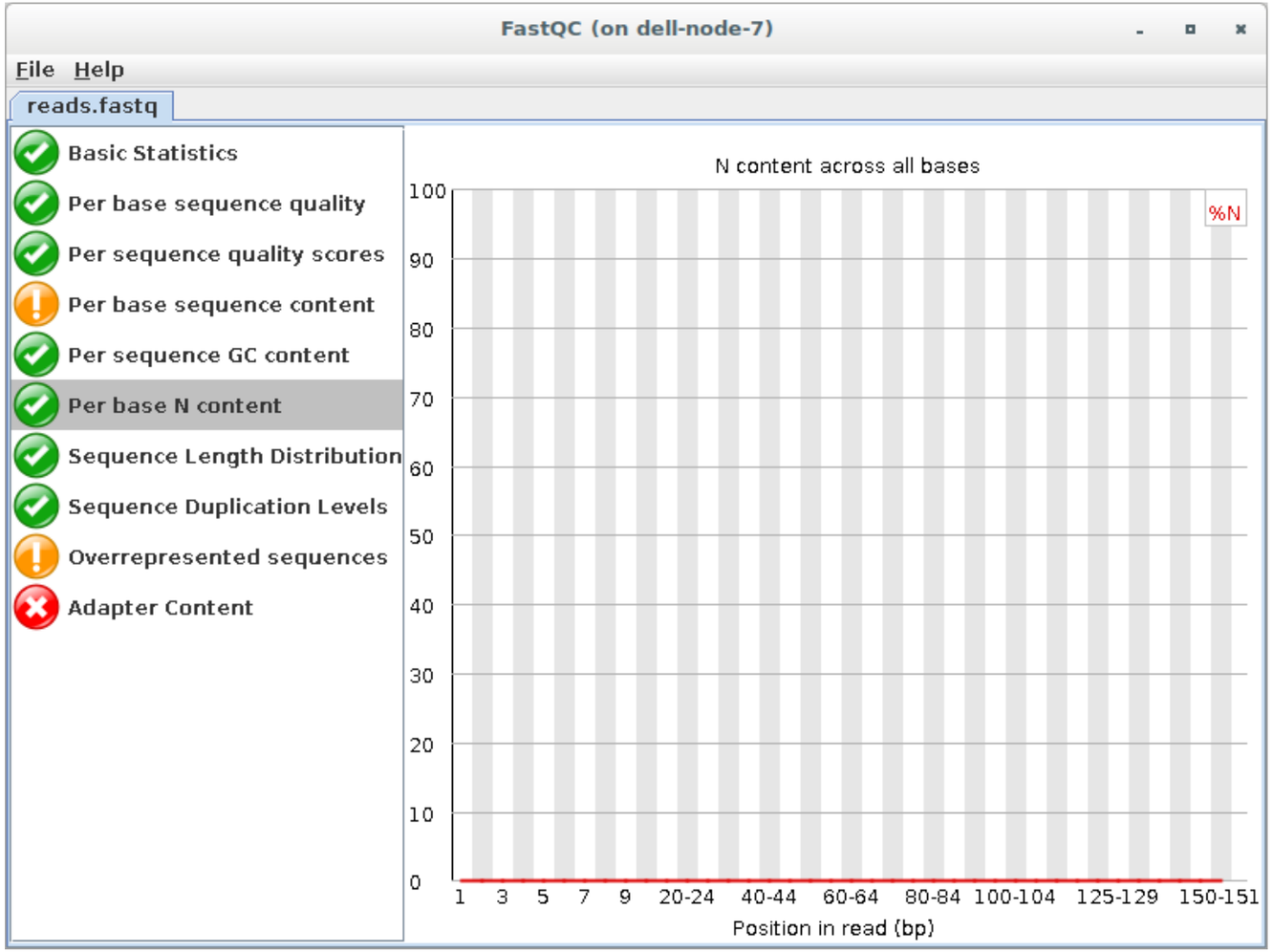
Per Base Sequence Content



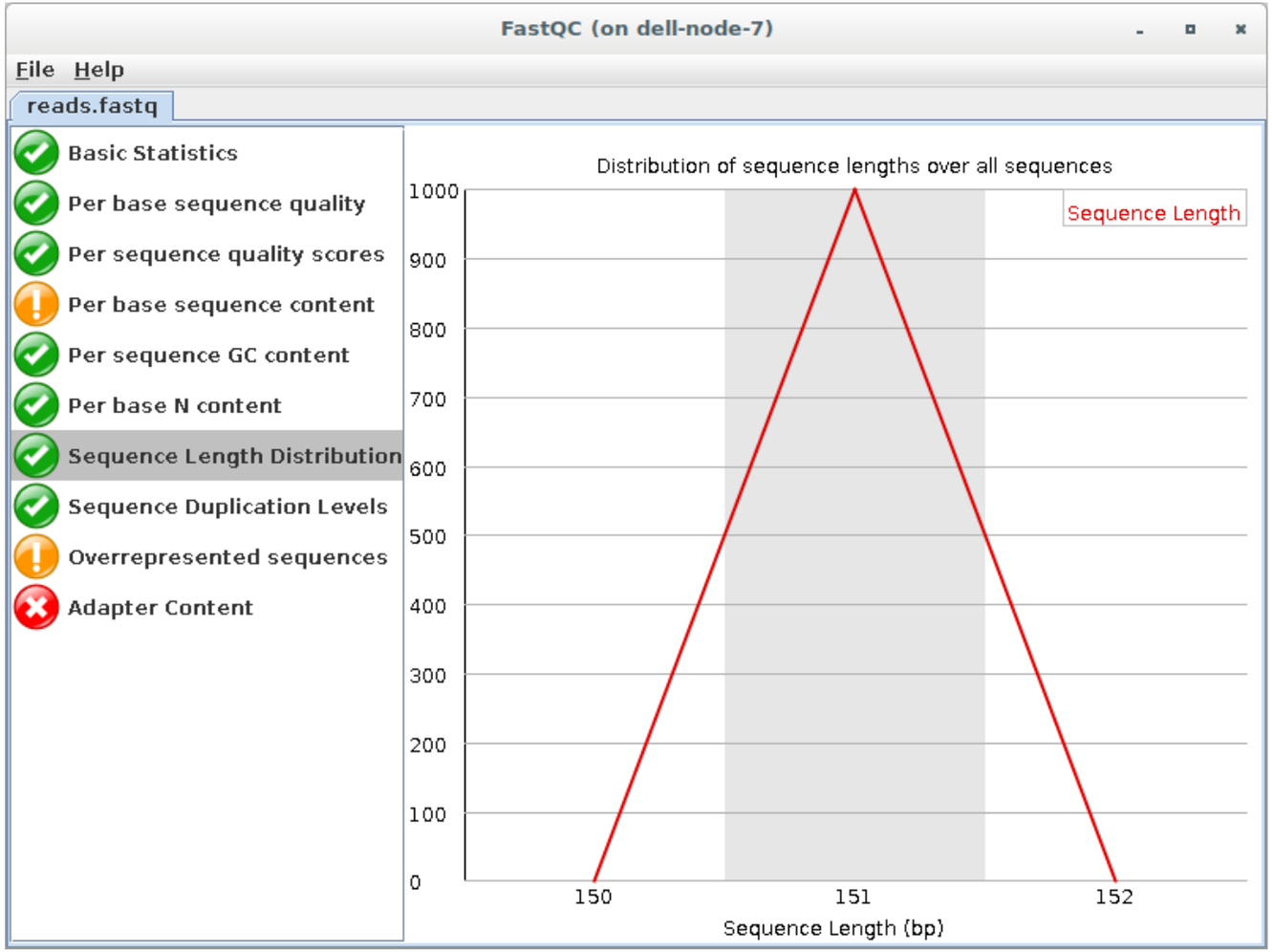
Per Sequence GC Content



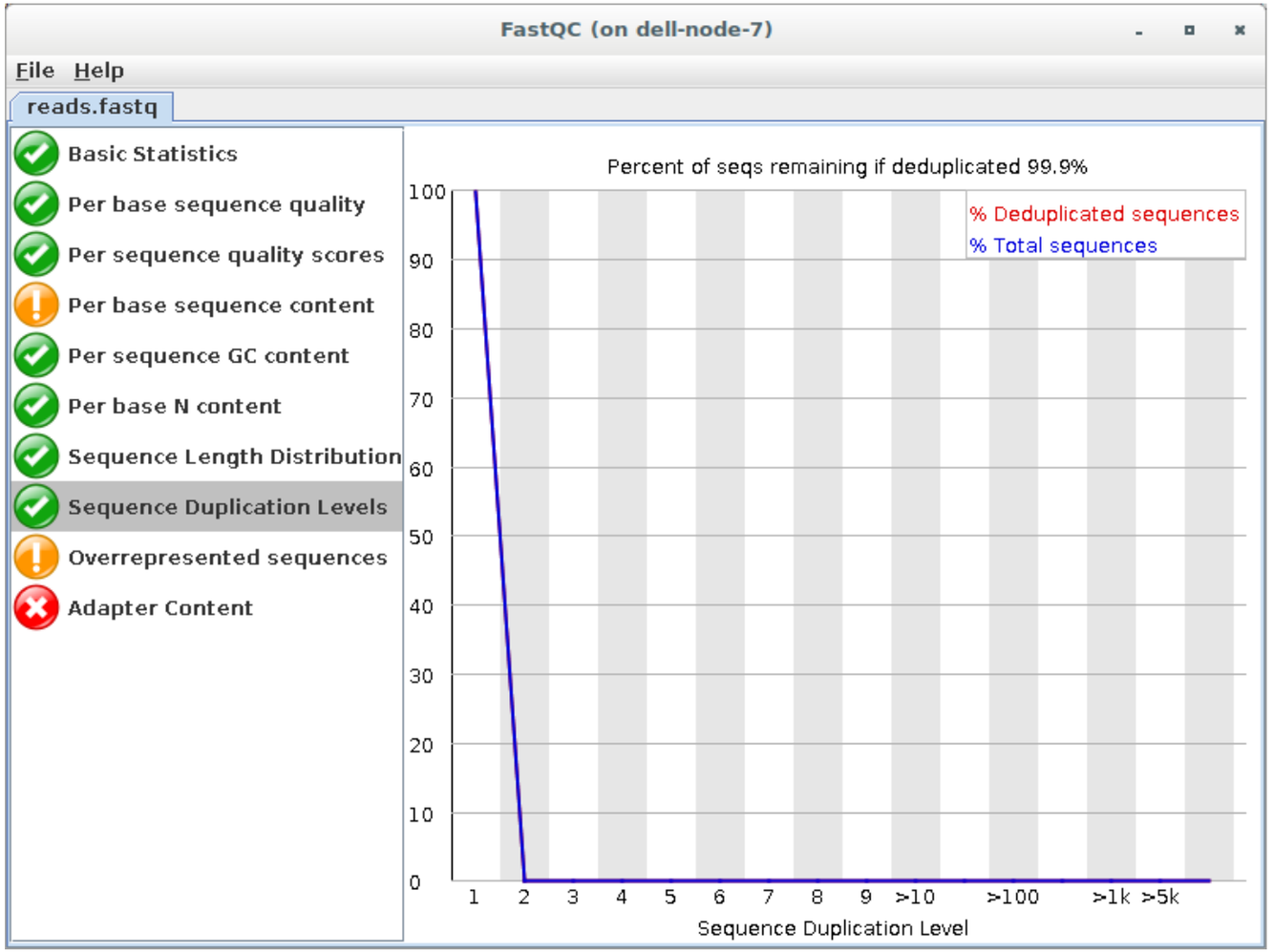
Per Base N Content



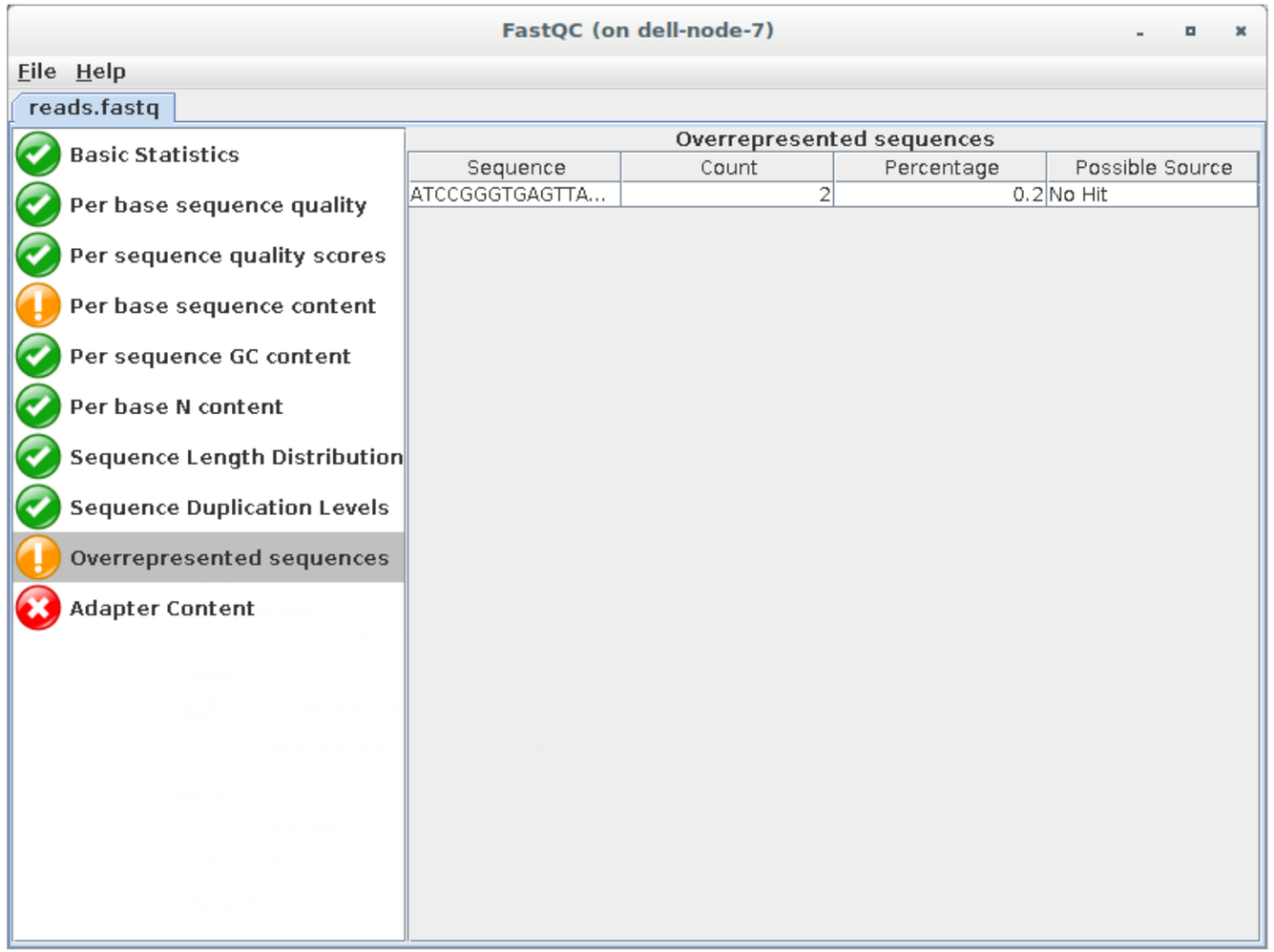
Sequence Distribution Lengths



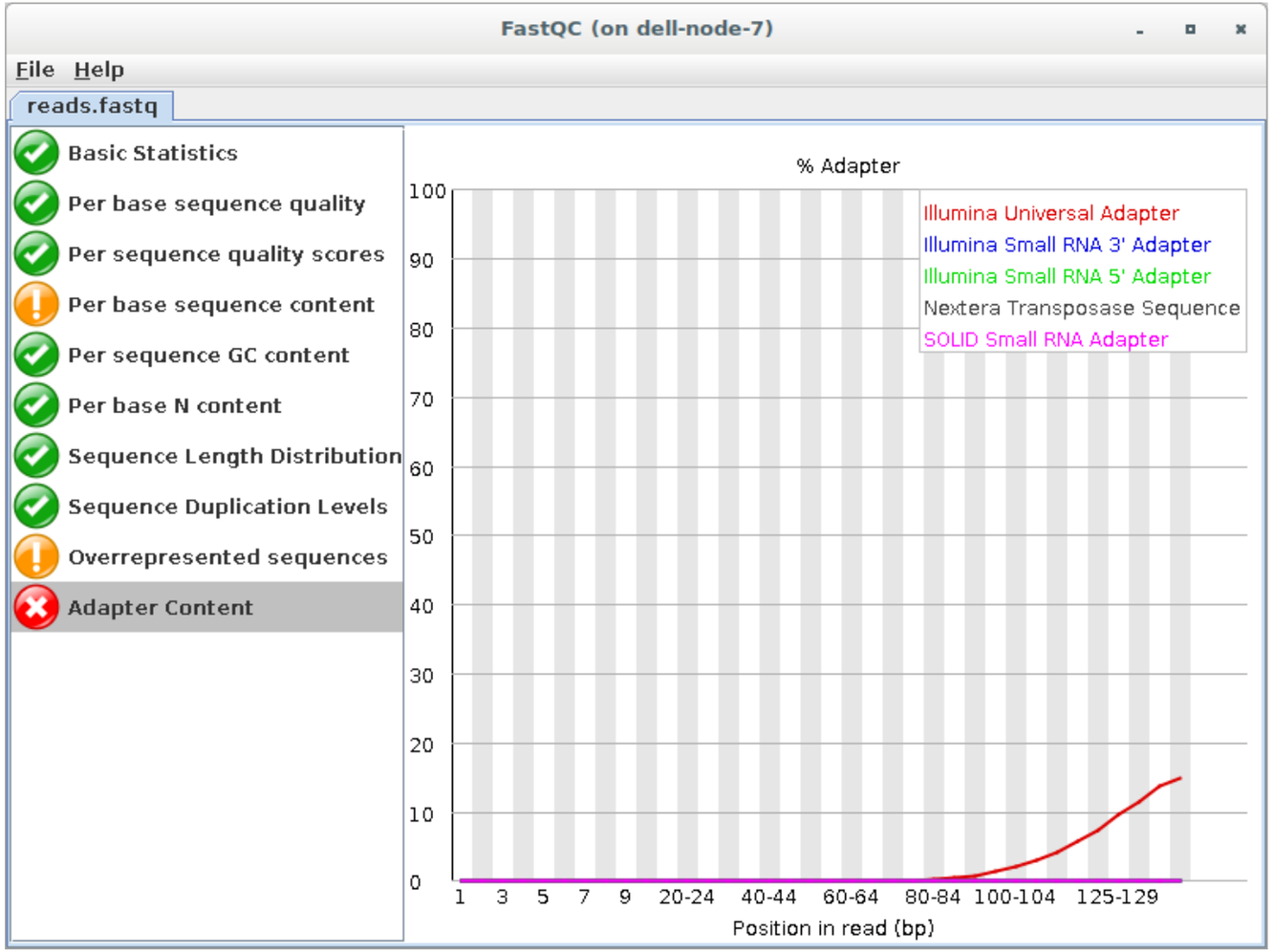
Sequence Duplication Levels



Overrepresented Sequences



Adapter Content



3. Read trimming

Trimmomatic parameters

```
- ILLUMINACLIP: Cut illumina-specific sequences from the read.
- SLIDINGWINDOW: Sliding window cutting when quality falls below a threshold
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
```

Example of a Trimmomatic command line:

```
trimmomatic SE reads.fastq reads_trimmed.fastq
        LEADING:20 TRAILING:20 AVGQUAL:20 HEADCROP:10 MINLEN:100
```

Use Trimmomatic to improve the reads quality, then observe the improvements in FastQC.