Haplotype inference with Long-read sequencers

Yosuke Tanigawa (ytanigaw@stanford.edu)
M. Rivas Lab (rotation student) | Biomedical Informatics Ph.D. Program

[Background]
Long read sequencer & Haplotype reference panel

- Long read sequencers (Oxford Nanopore, PacBio)
  - Read length >= 8 kb
- Haplotype reference
  - UK BioBank (152,729 imputed haplotypes)
  - Haplotype reference consortium (not available yet?)

[Rotation project]
Compressed representation of haplotypes

- How can we compress haplotype information ?

- Is it possible to infer haplotype on the fly ???

[Data]
cDNA data set & WGS data set (by Helio)

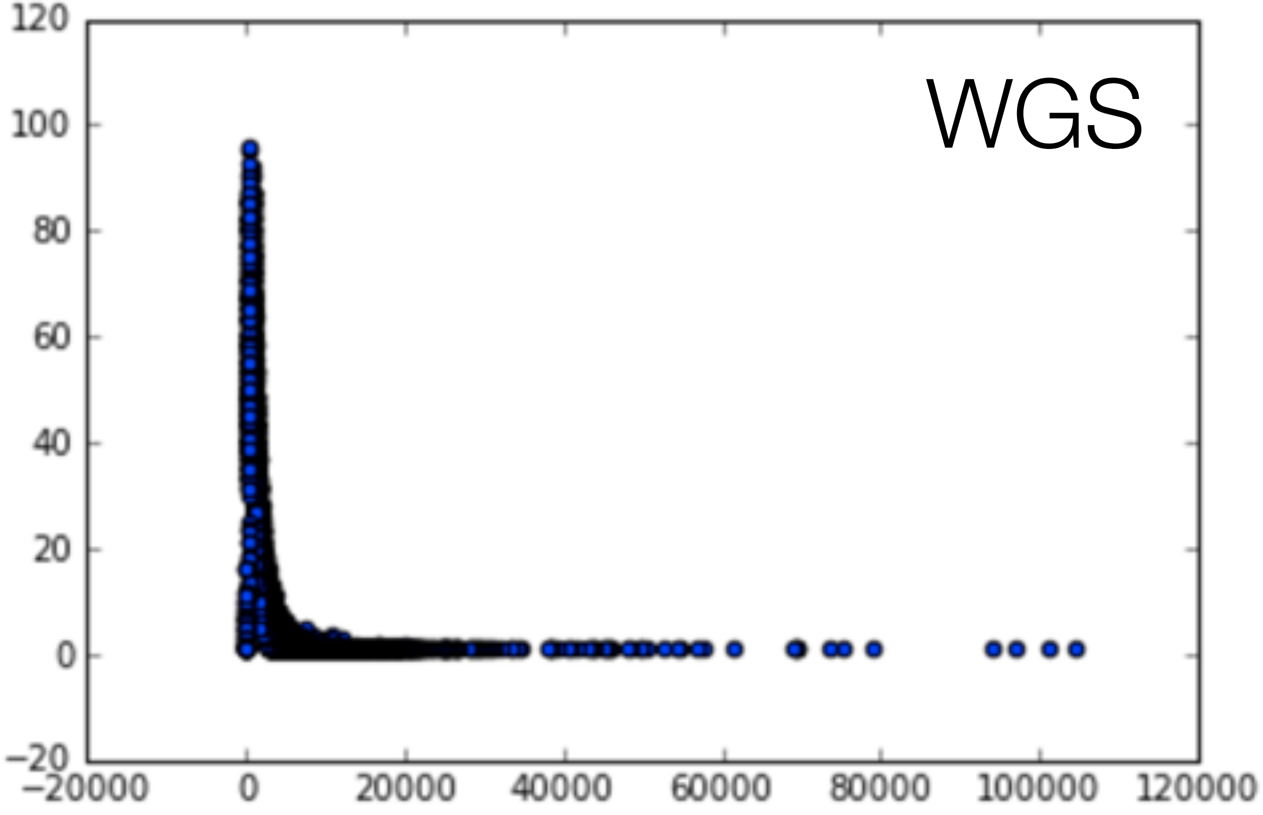
- Technology: Oxford Nanopore MinION sequencer
  R9.3 chemistry + minKnow v1.1.14
  (current version: R9.4 & v1.1.17)

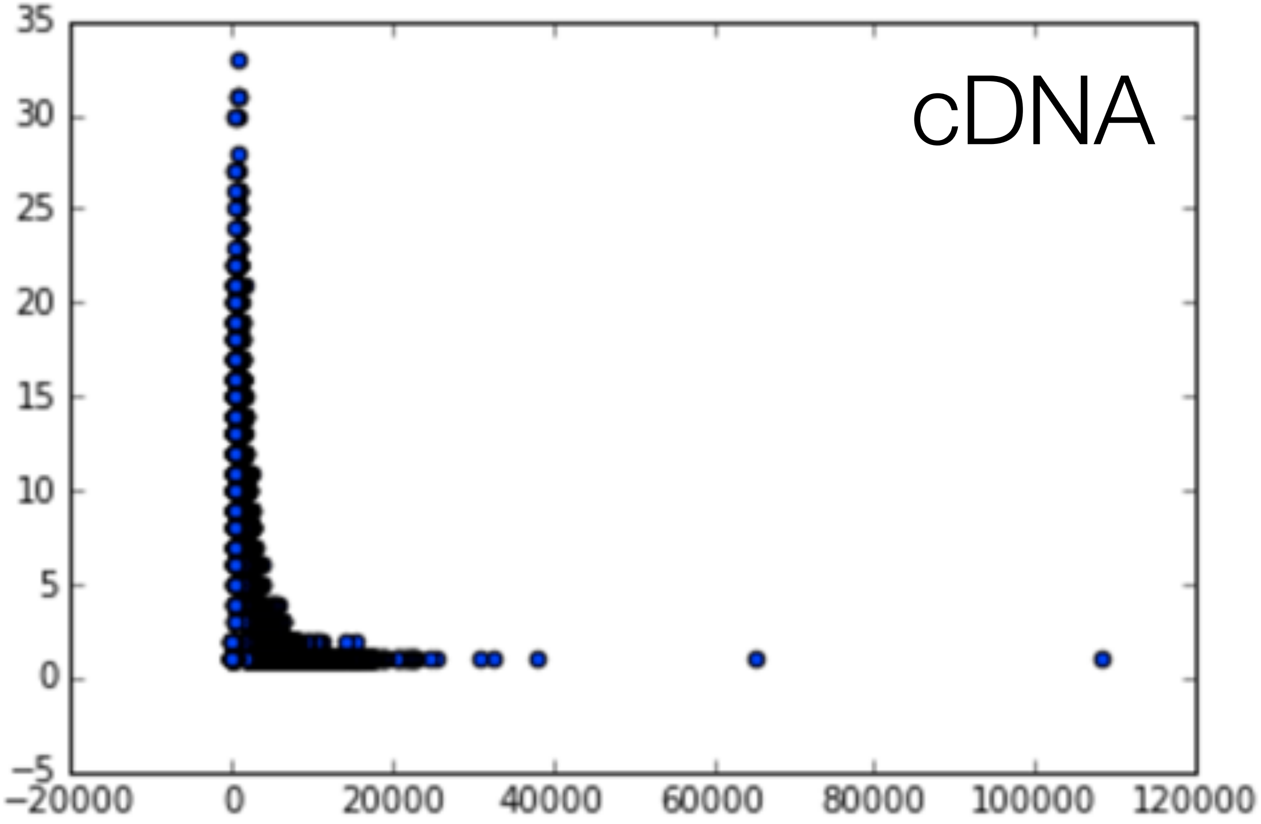
- “WGS” data set
  20161008_wgs_caucasian_48hr
  48hr(?) run

- “cDNA” data set
  20161006_minion_human_cDNA
  6hr run

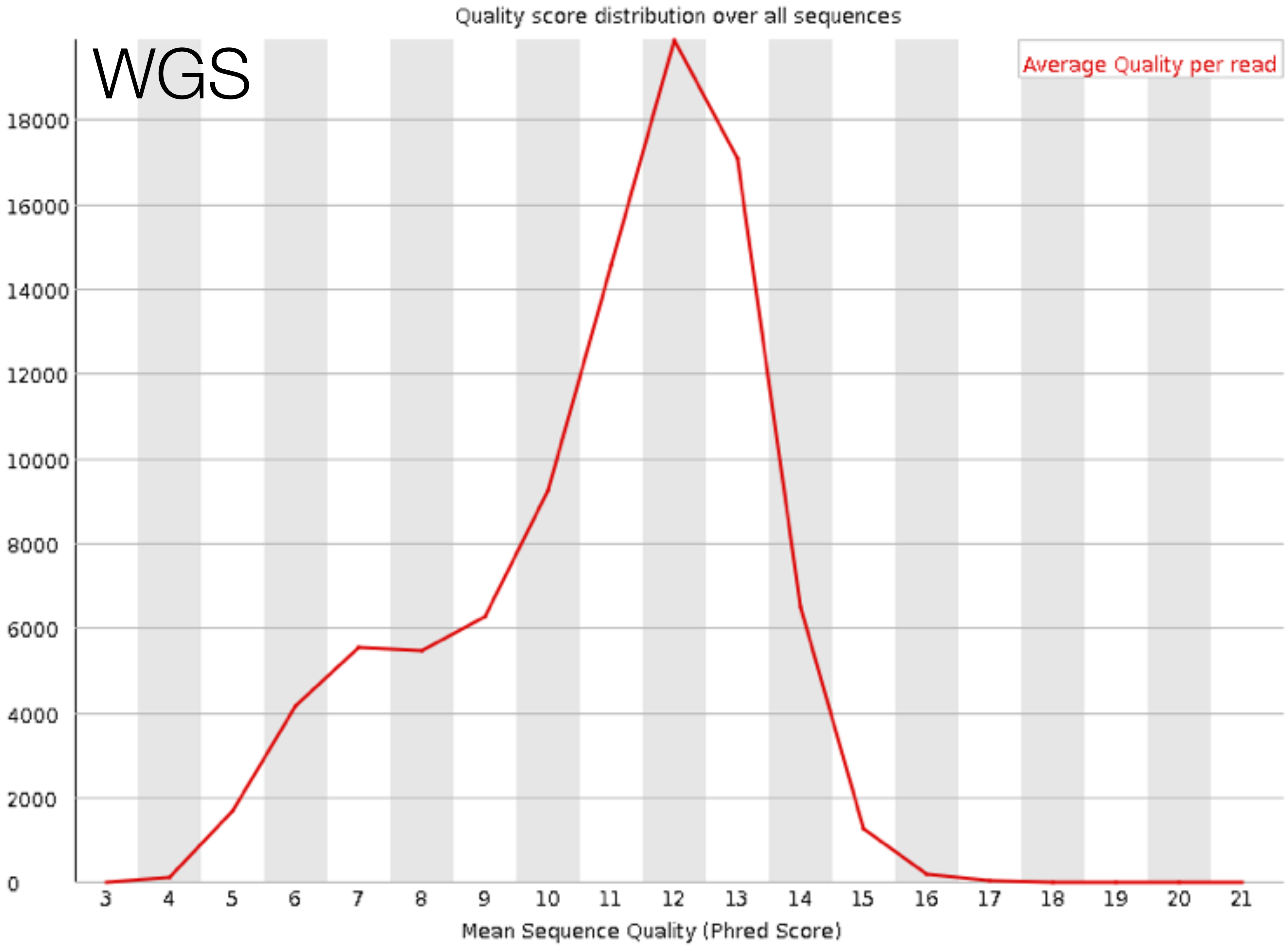
Read Length Distribution

	WGS	cDNA
total base pairs	44,839,915	46,314,462
N50	2,227	2,529
# of seq (total)	92,186	26,854
# of seq > 1kb	42,572	14,550
# of seq > 10kb	660	302
# of seq > 20kb	102	21

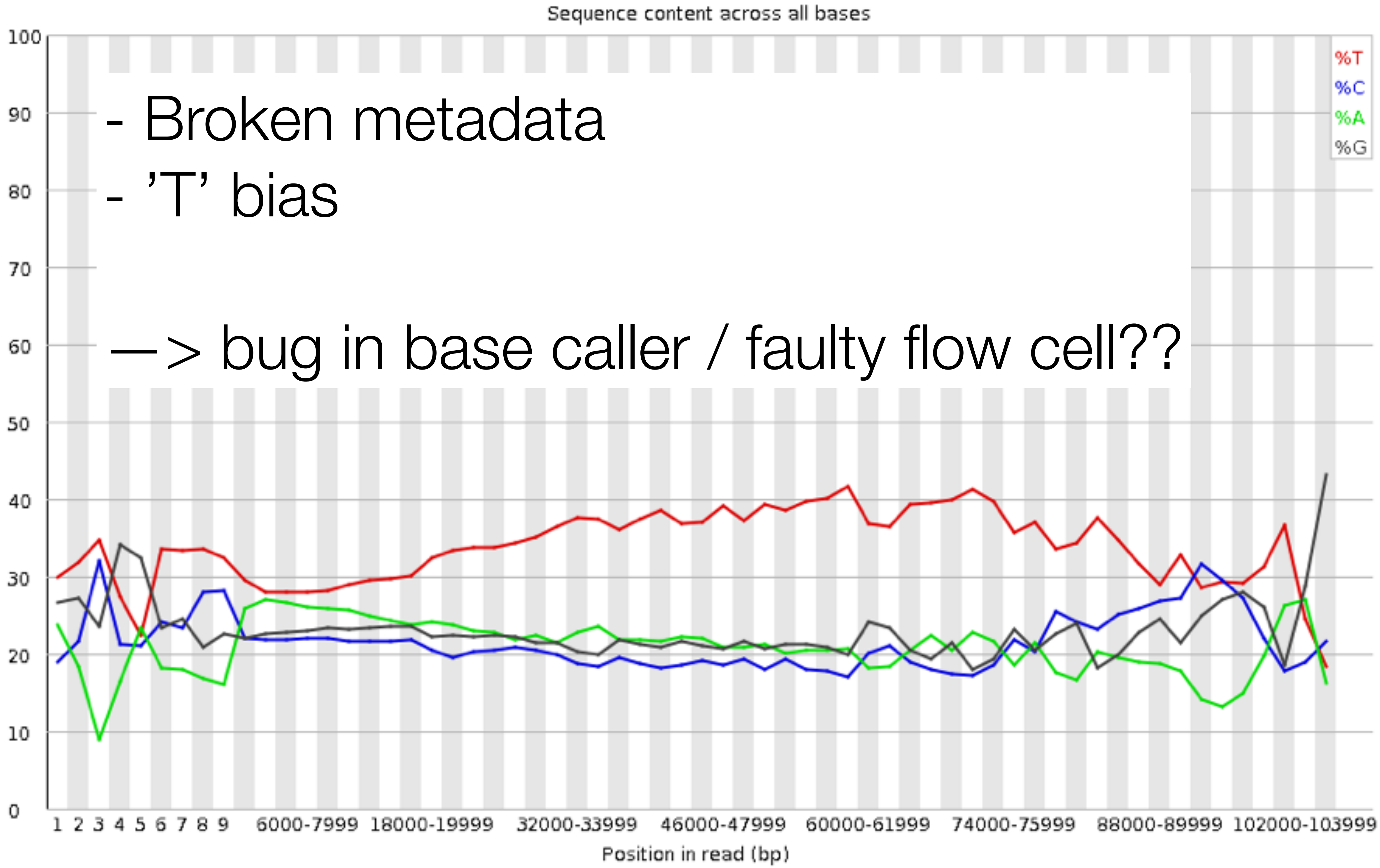




Error rate is 10-20%: Quality score freq. dist.



WGS data set looks strange



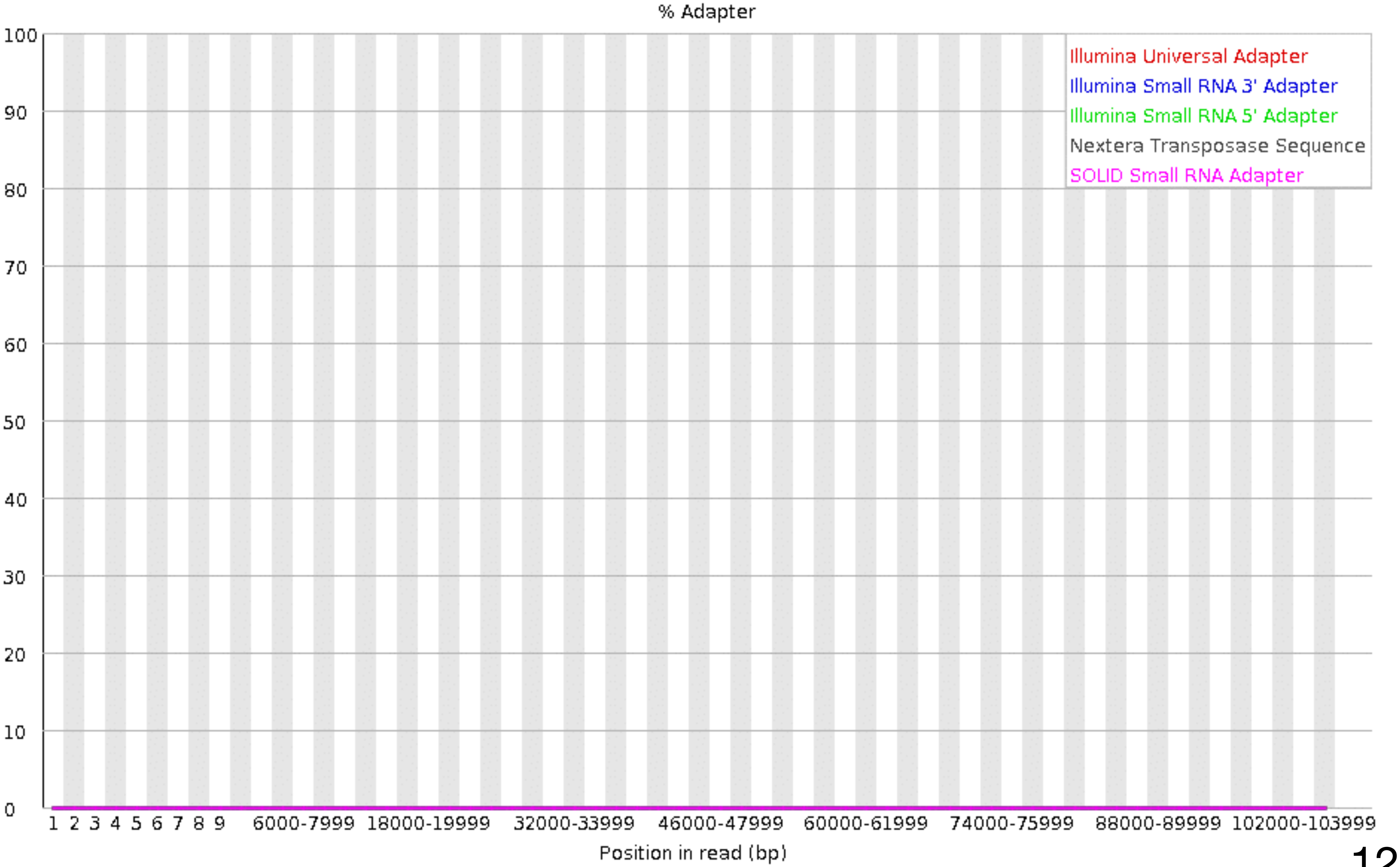
- Broken metadata
- ’T’ bias

—> bug in base caller / faulty flow cell??

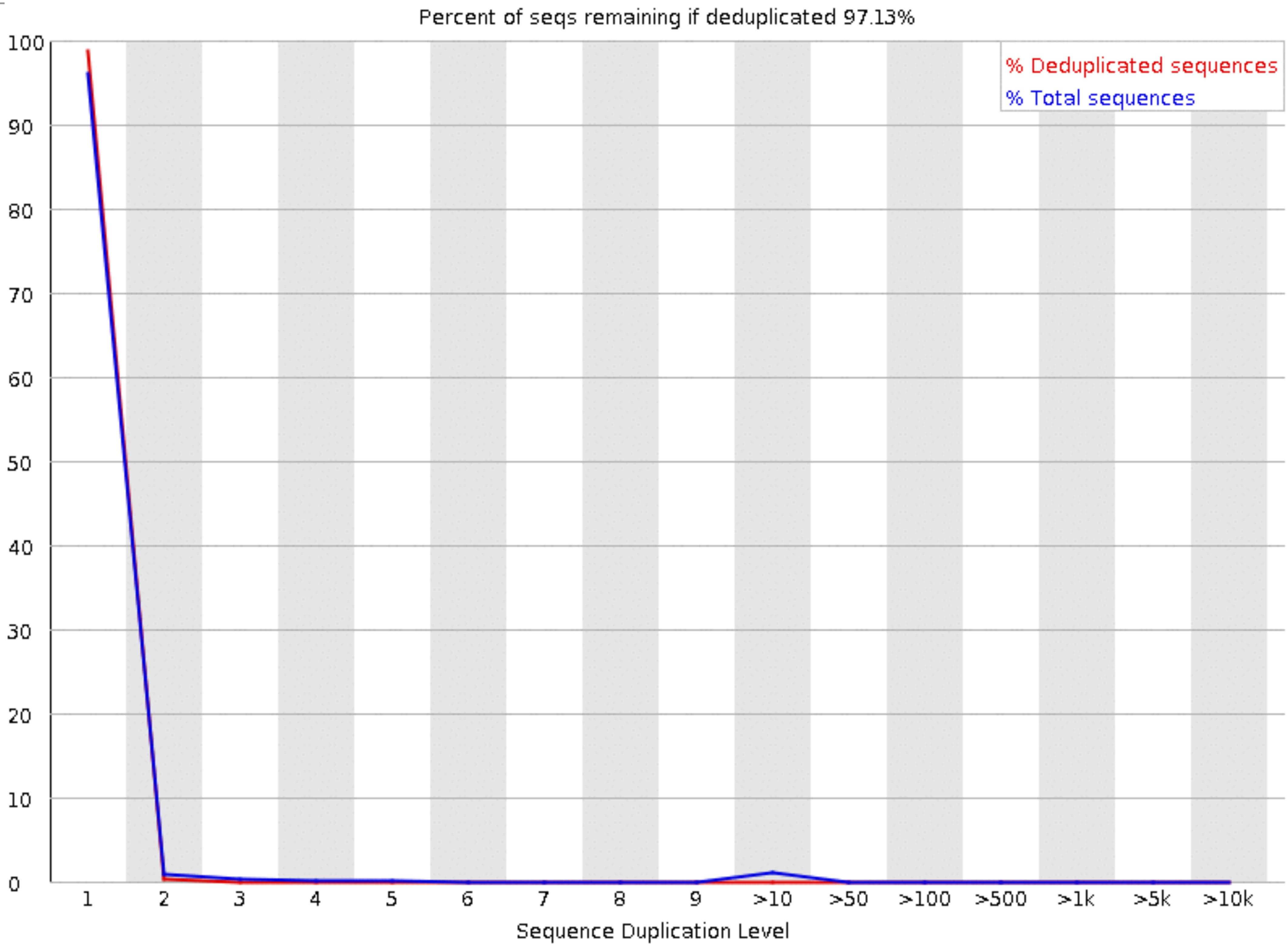
Stats.

	WGS	cDNA
Number of FAST5 files	184,911	48,280
total reads(*)	29,964	26,854
total base pairs	44,839,915	46,314,462
mean	1496.46	1724.68
median	925	1094
min	35	58
max	94024	108262
N25	4547	5201
N50	2227	2529
N75	1140	1327
# of seq (total)	92,186	26,854
# of seq > 1kb	42,572	14,550
# of seq > 10kb	660	302
# of seq > 20kb	102	21

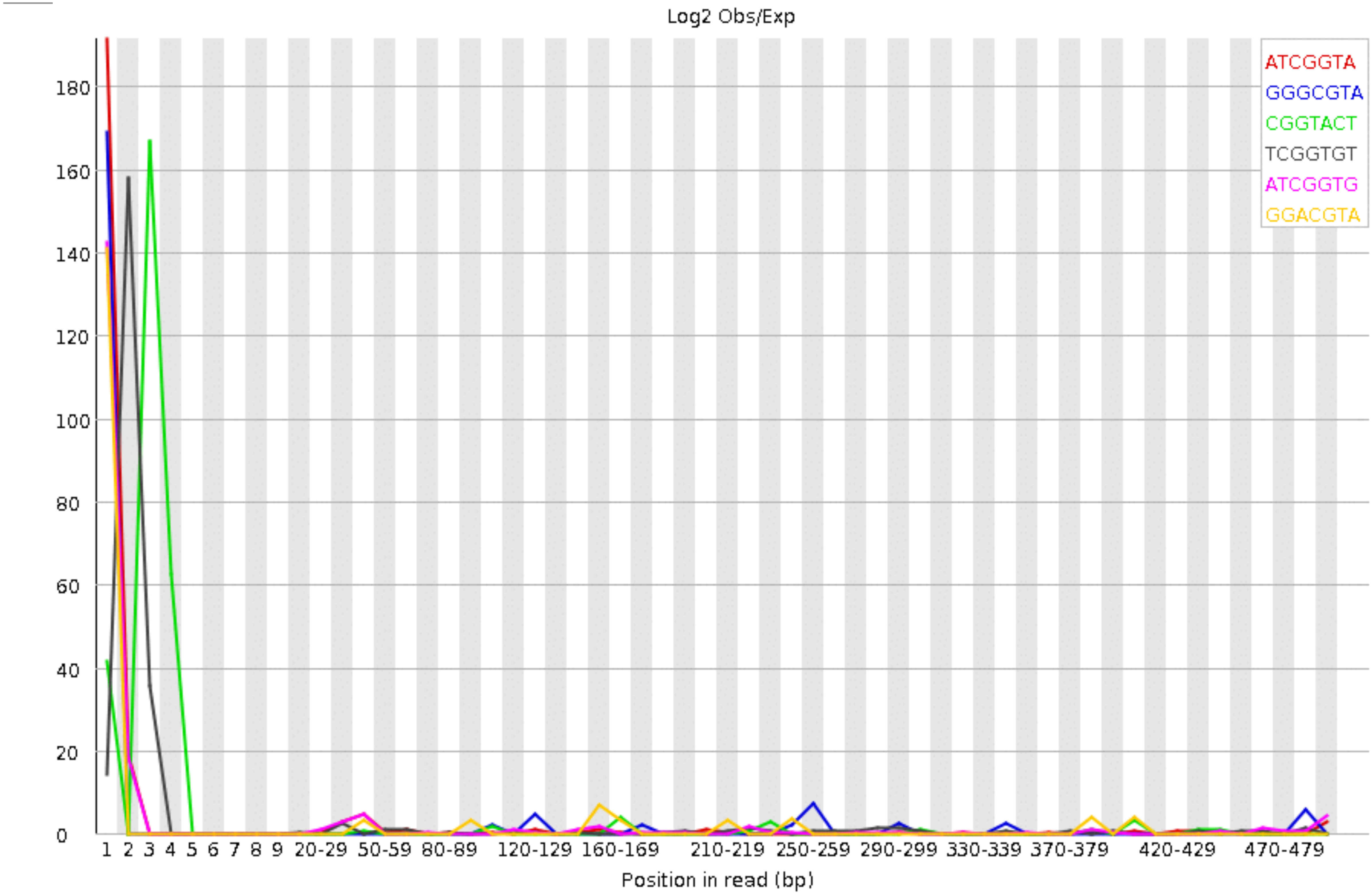
WGS — Adapter content



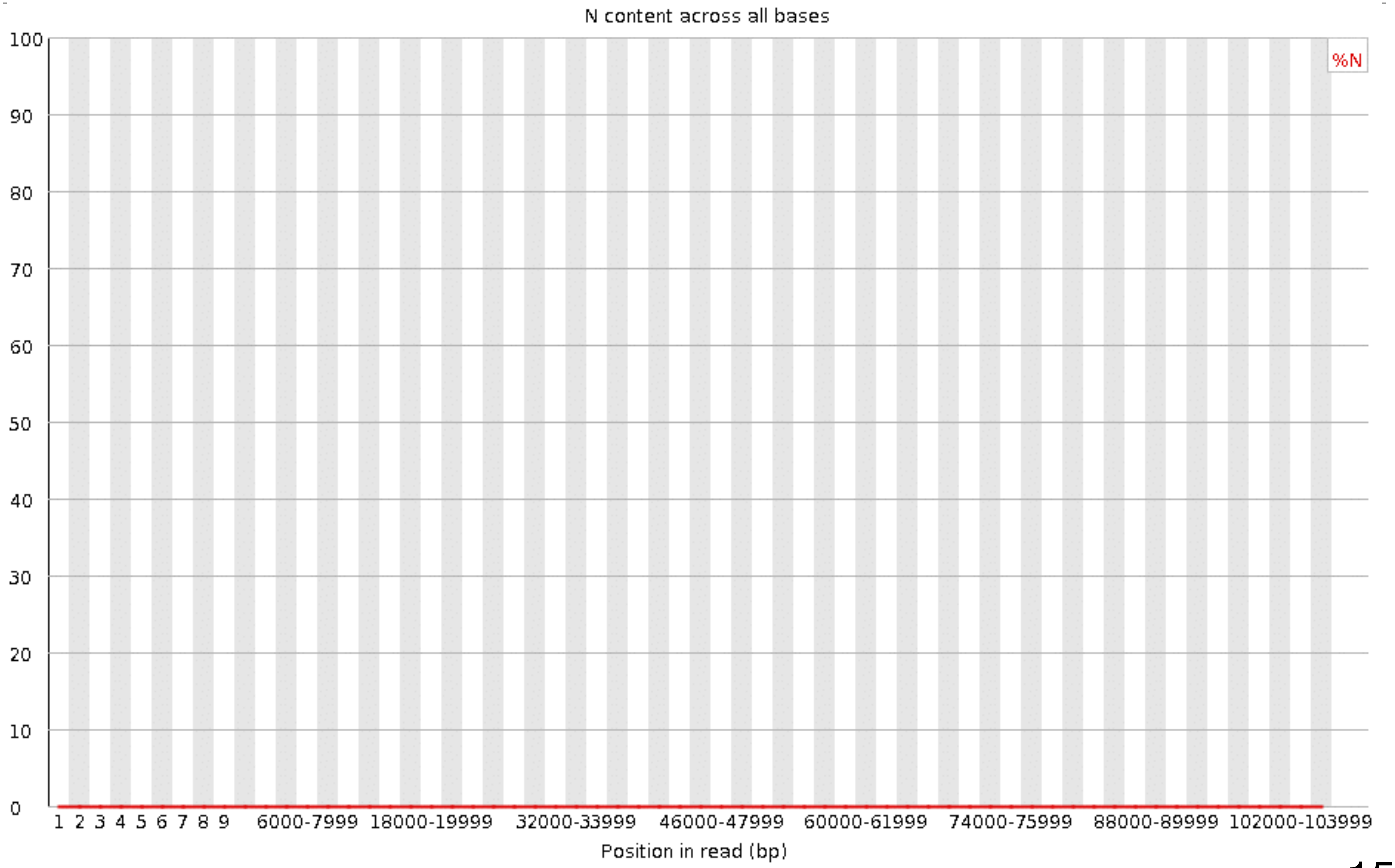
WGS — Duplicate levels



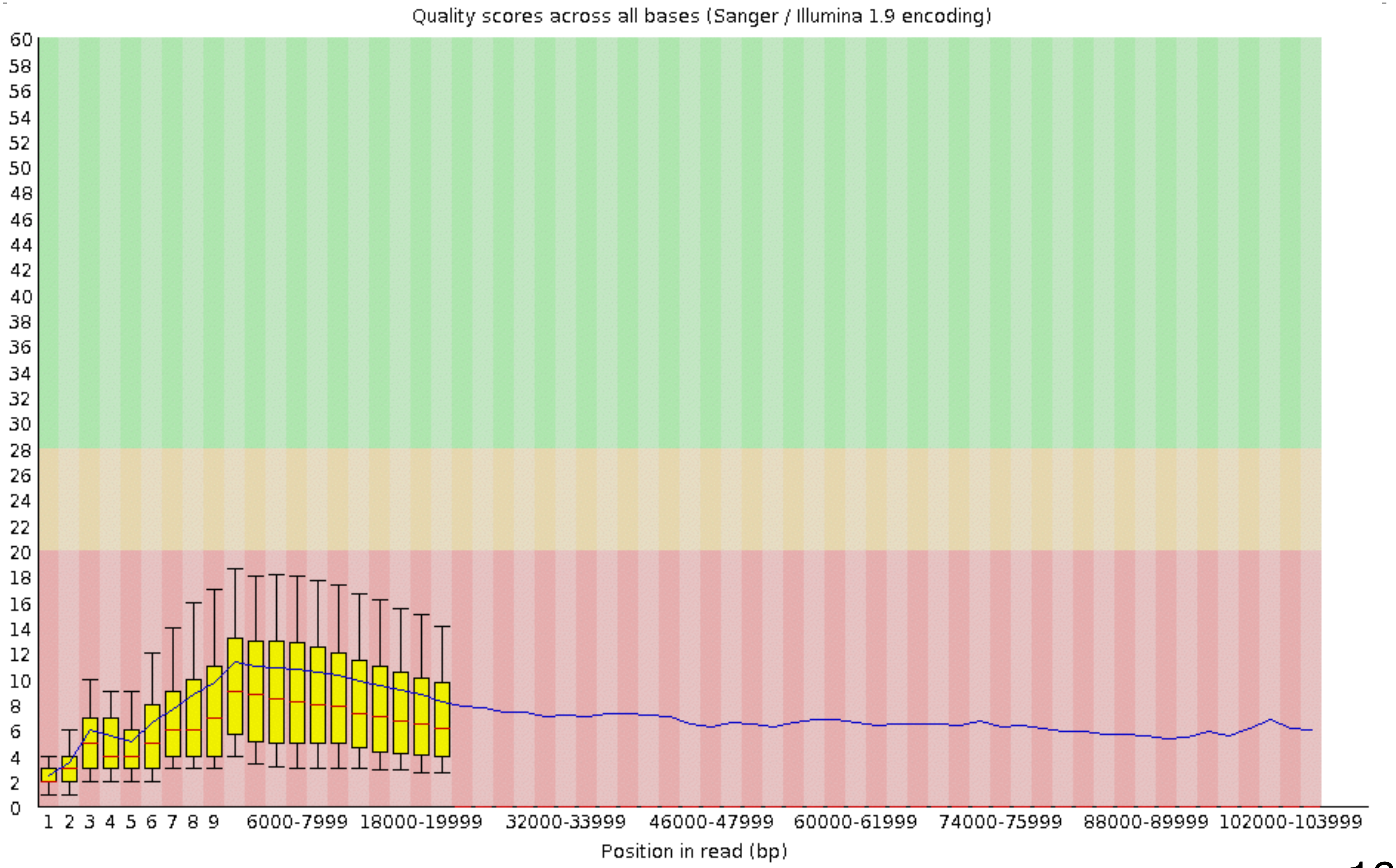
WGS — k-mer profiles



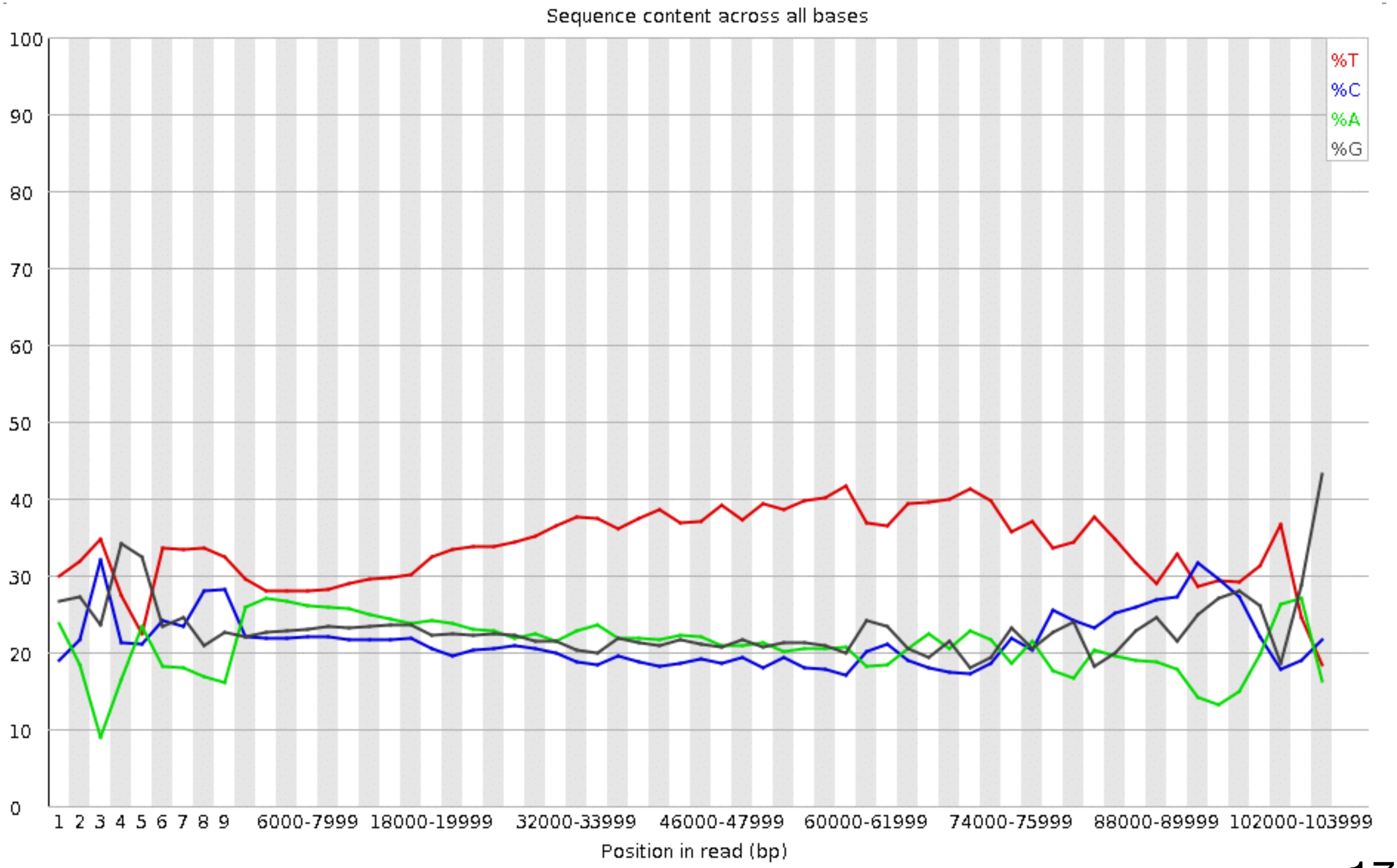
WGS — Per base N content



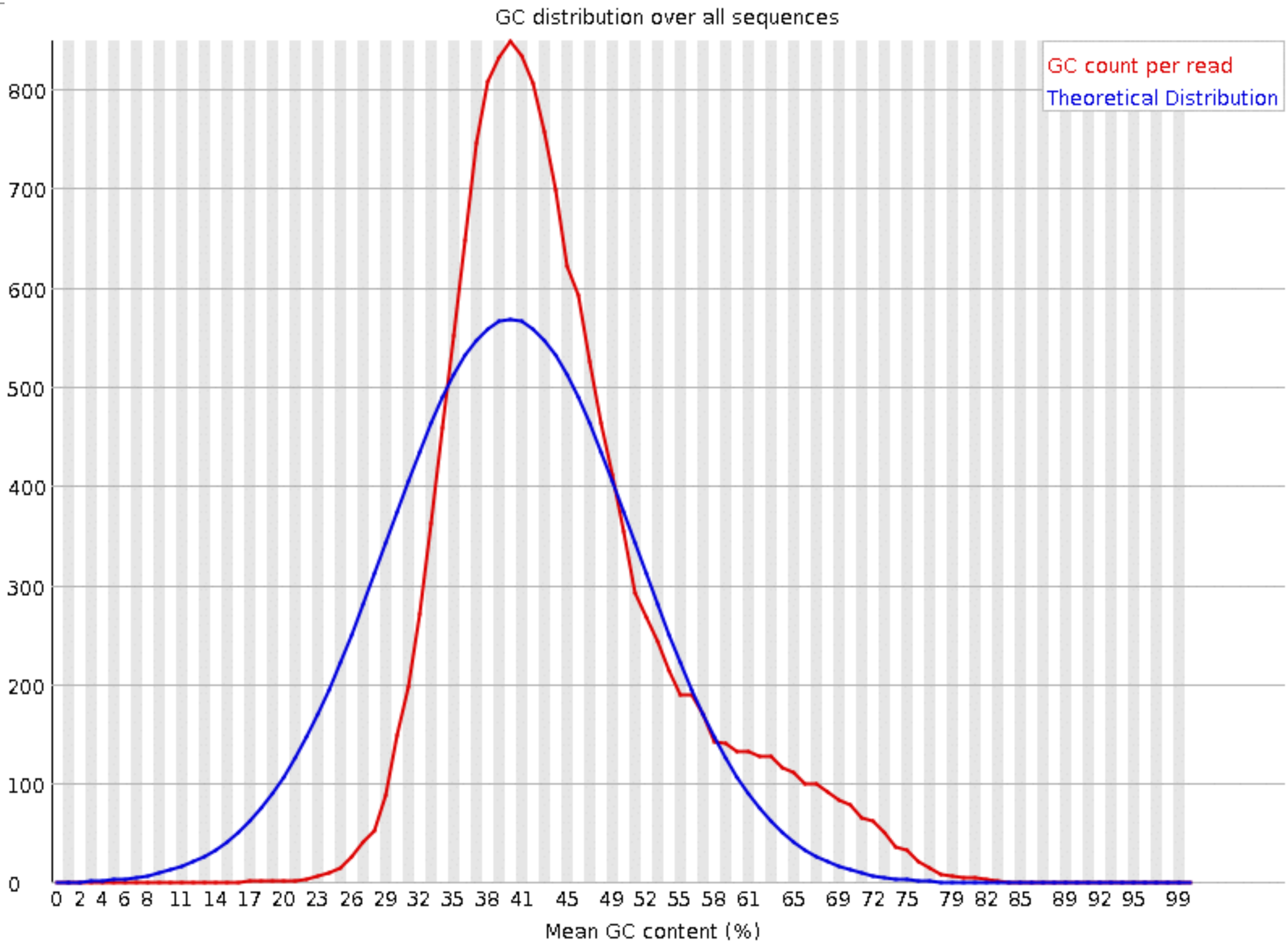
WGS — Per base quality



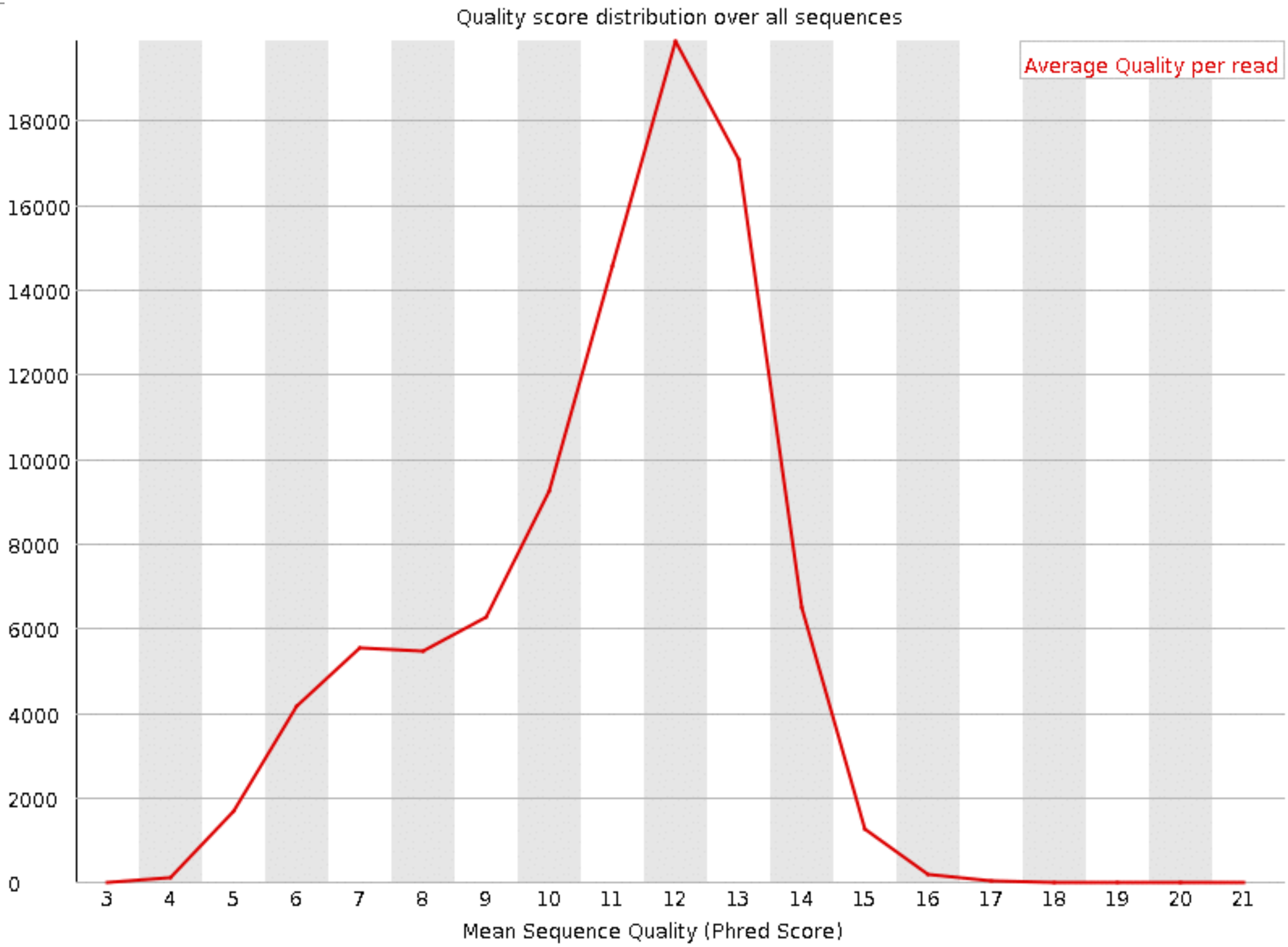
WGS — Per base sequence content



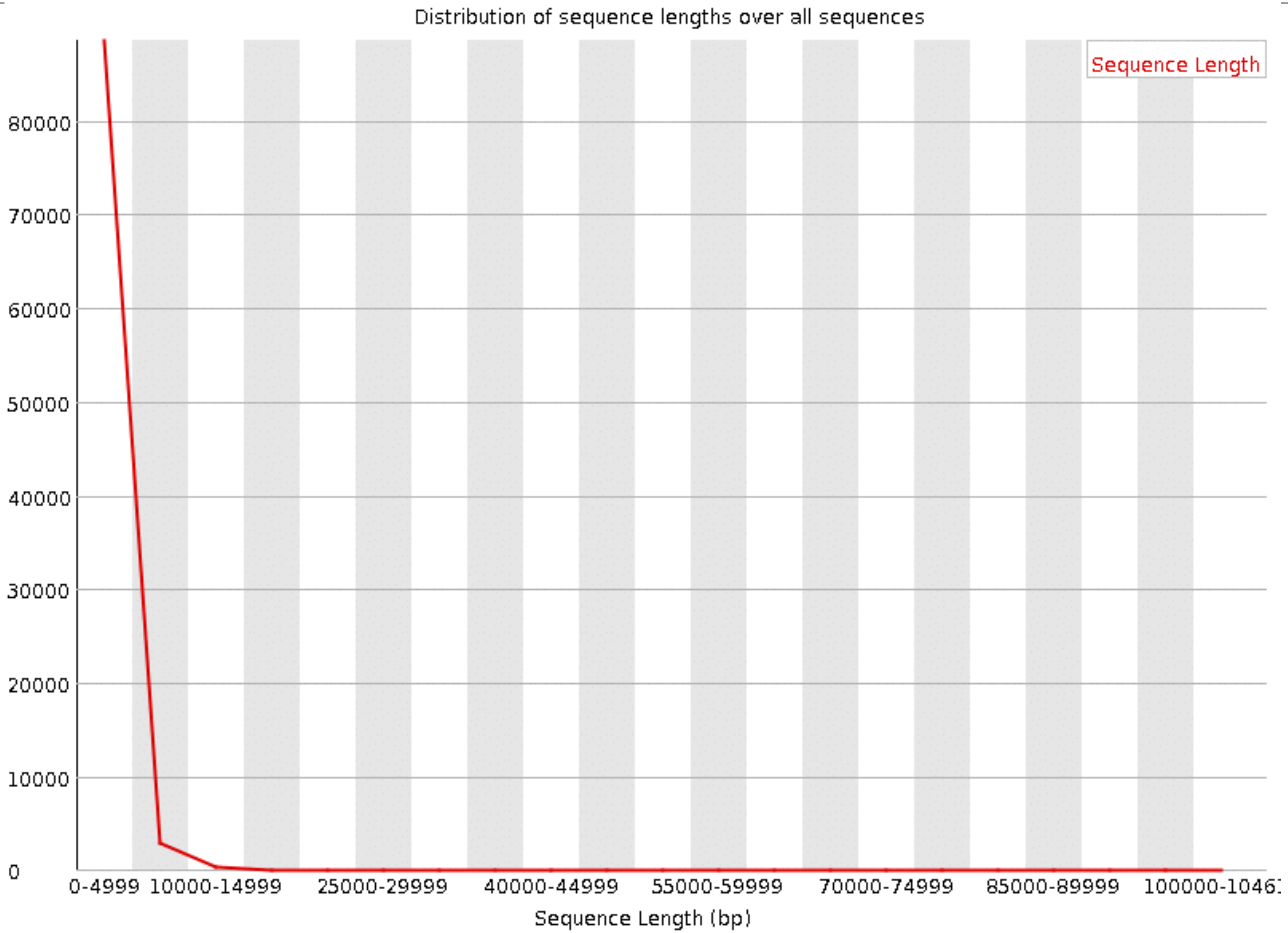
WGS — Per sequence GC content



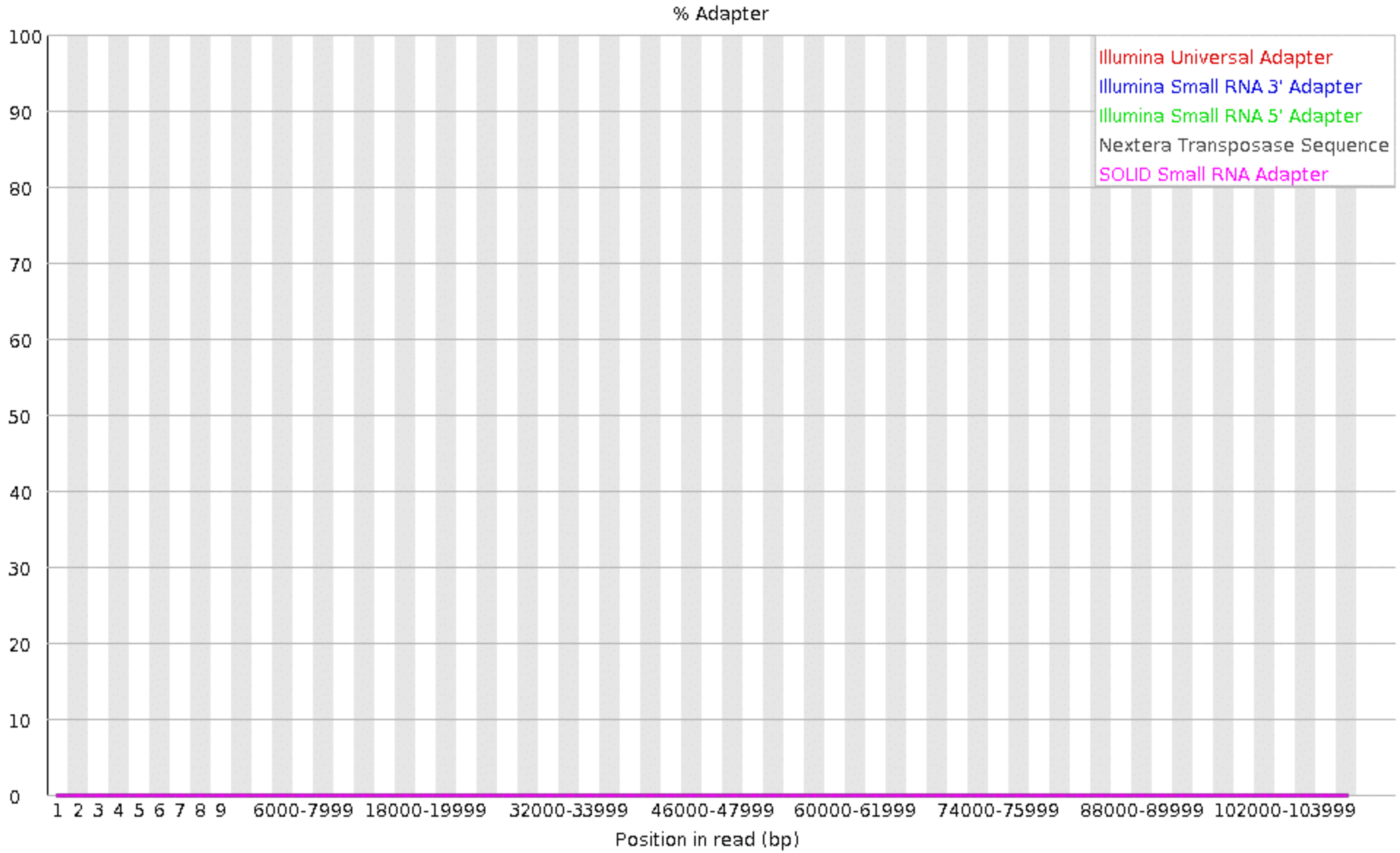
WGS — Per sequence quality



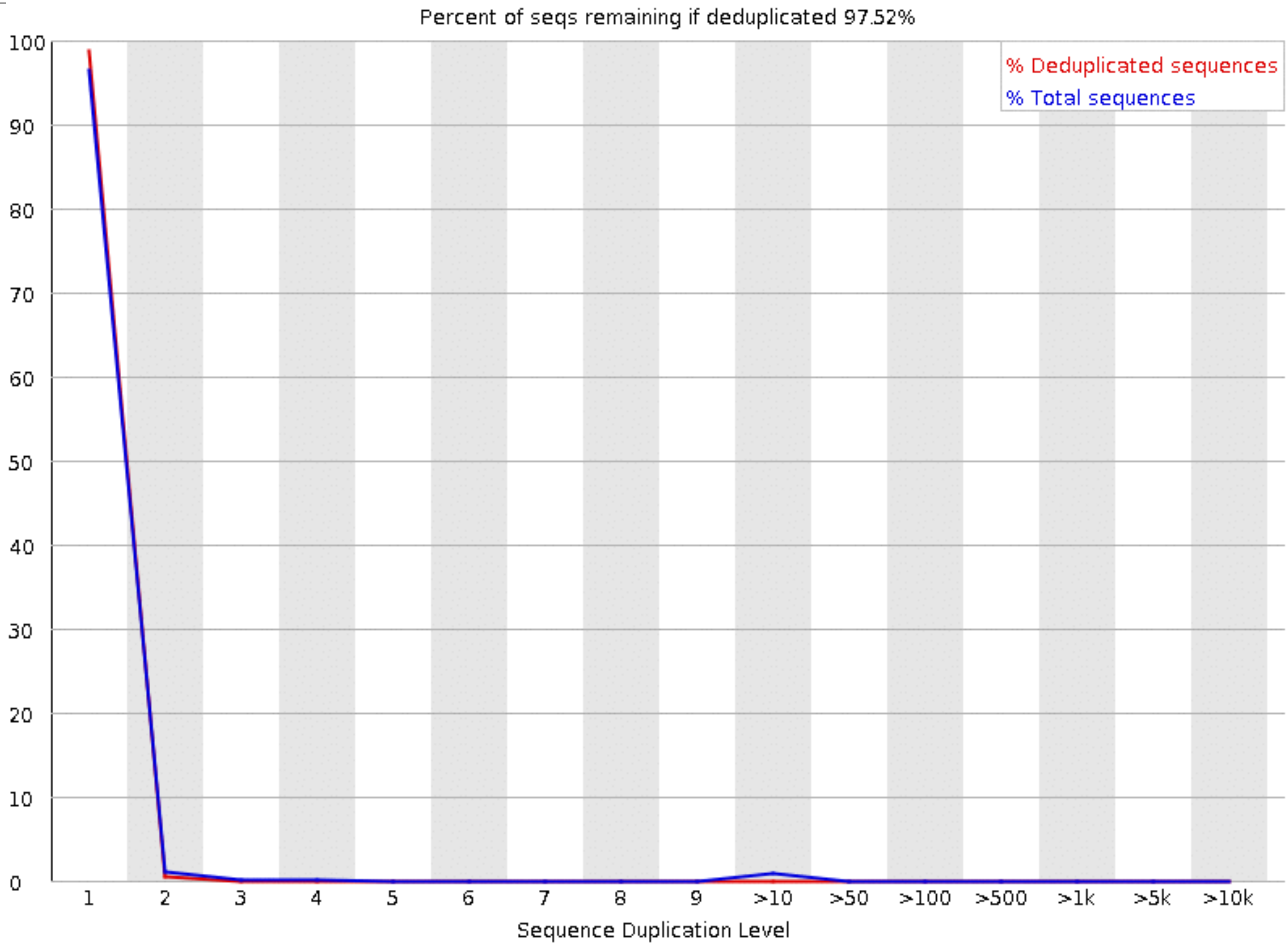
WGS — Sequence length distribution



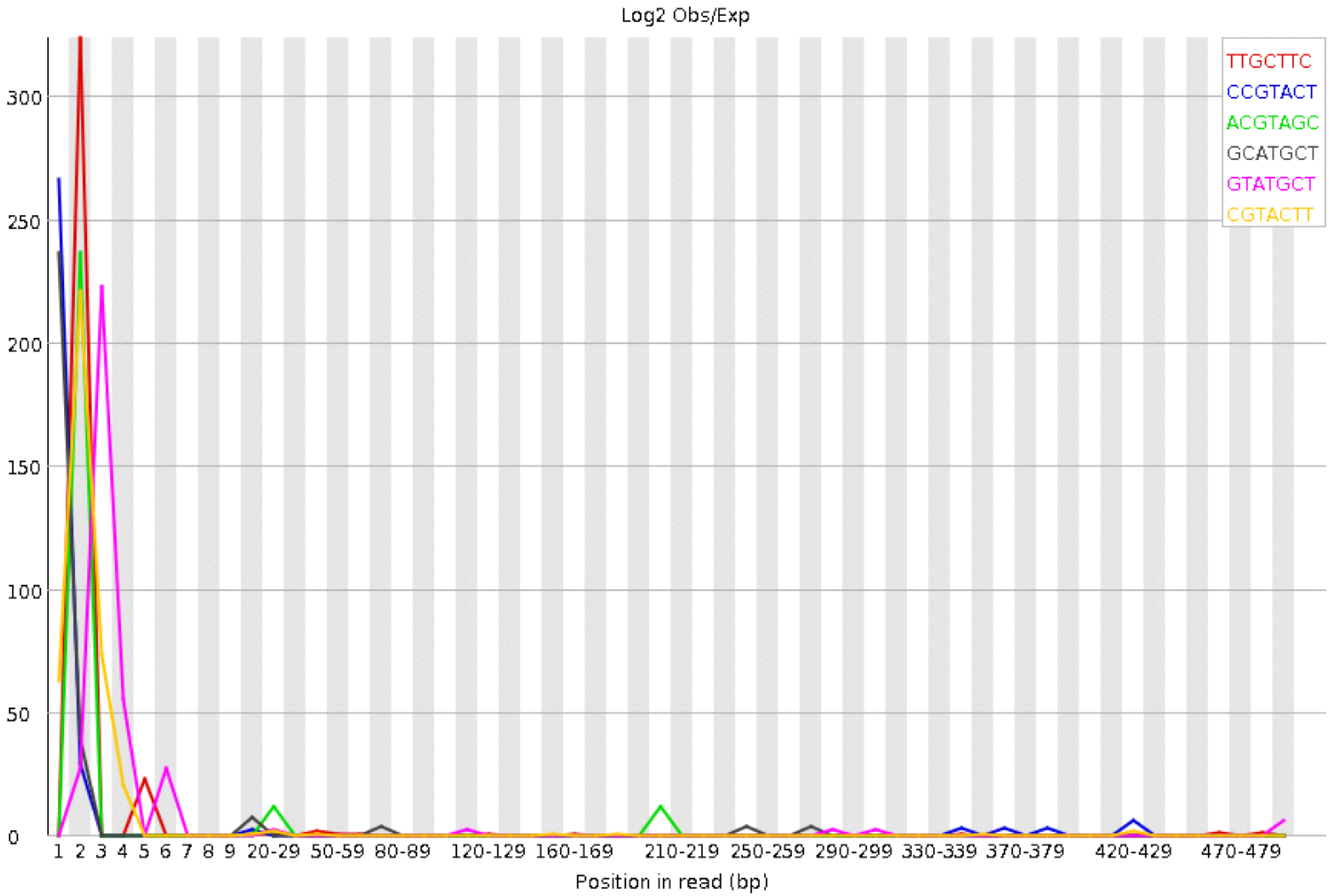
cDNA — Adapter content



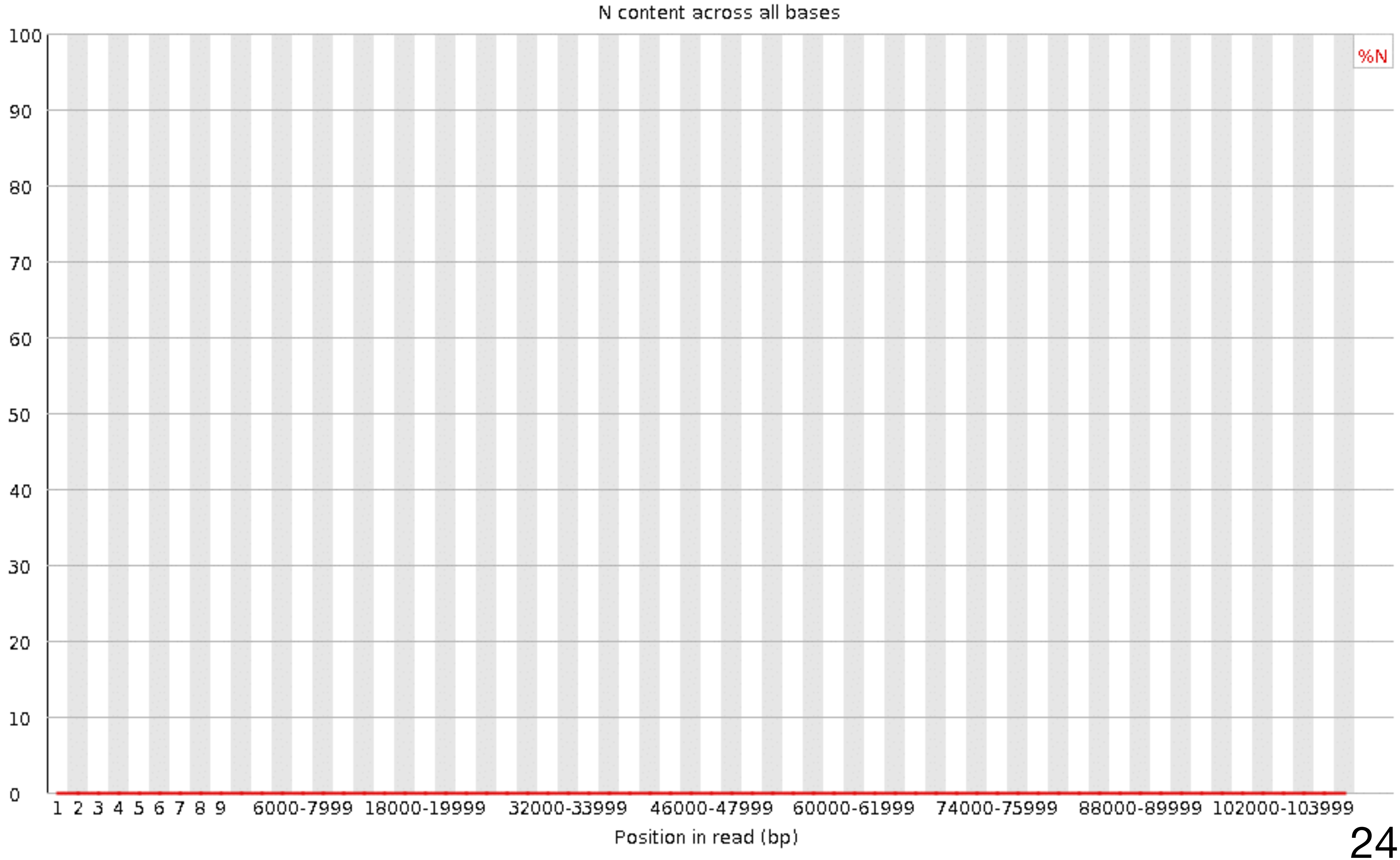
cDNA — Duplicate levels



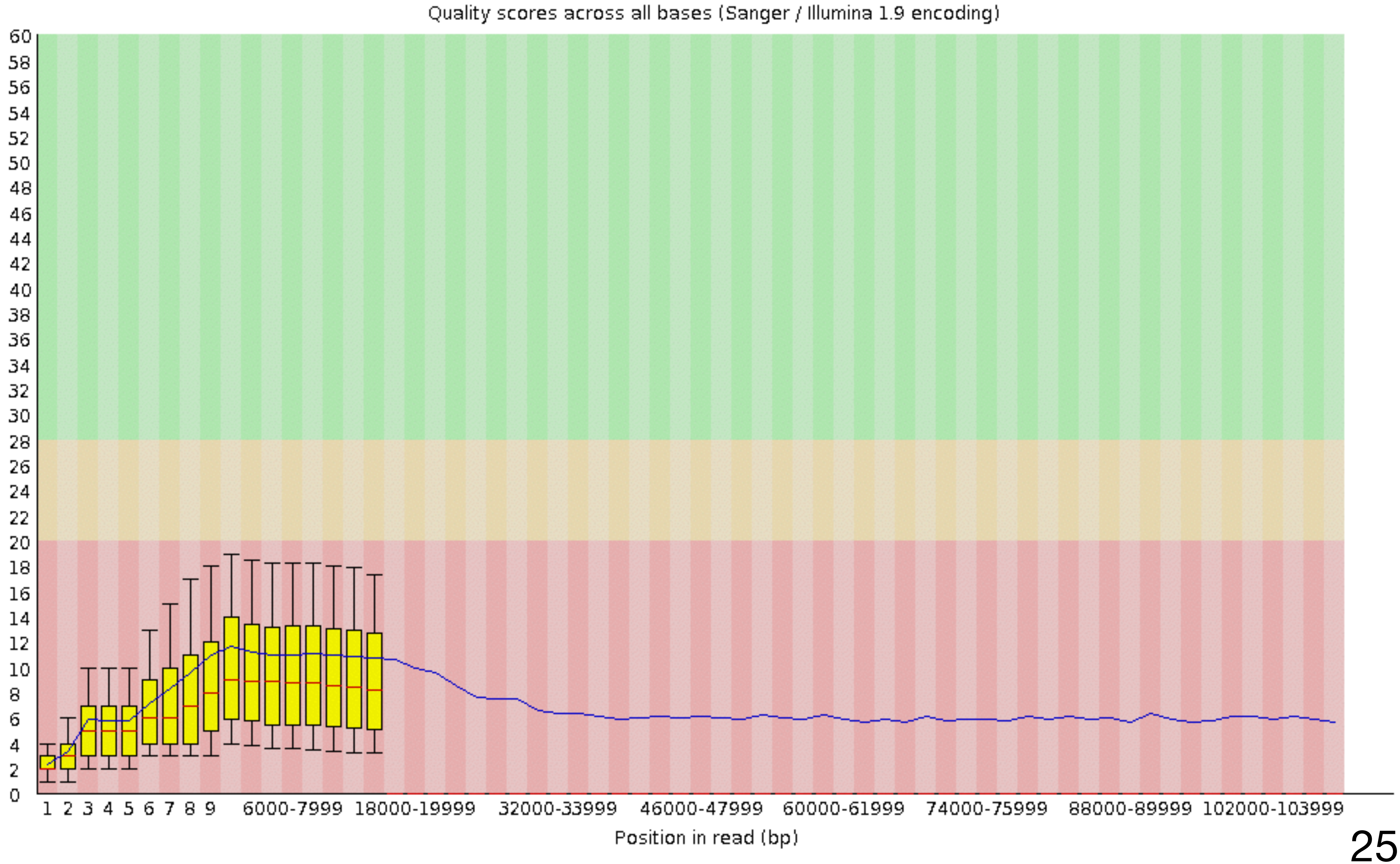
cDNA — k-mer profiles



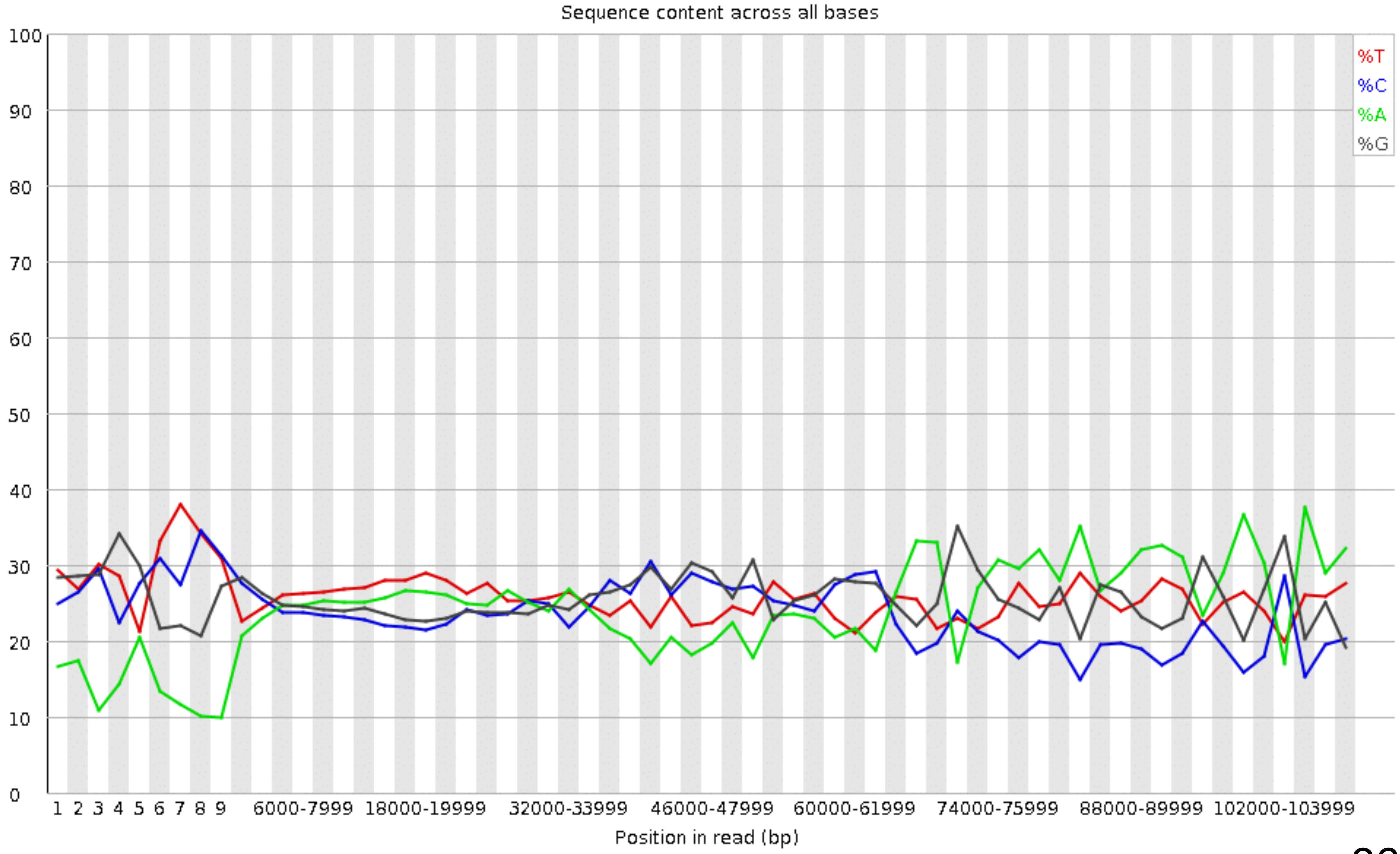
cDNA — Per base N content



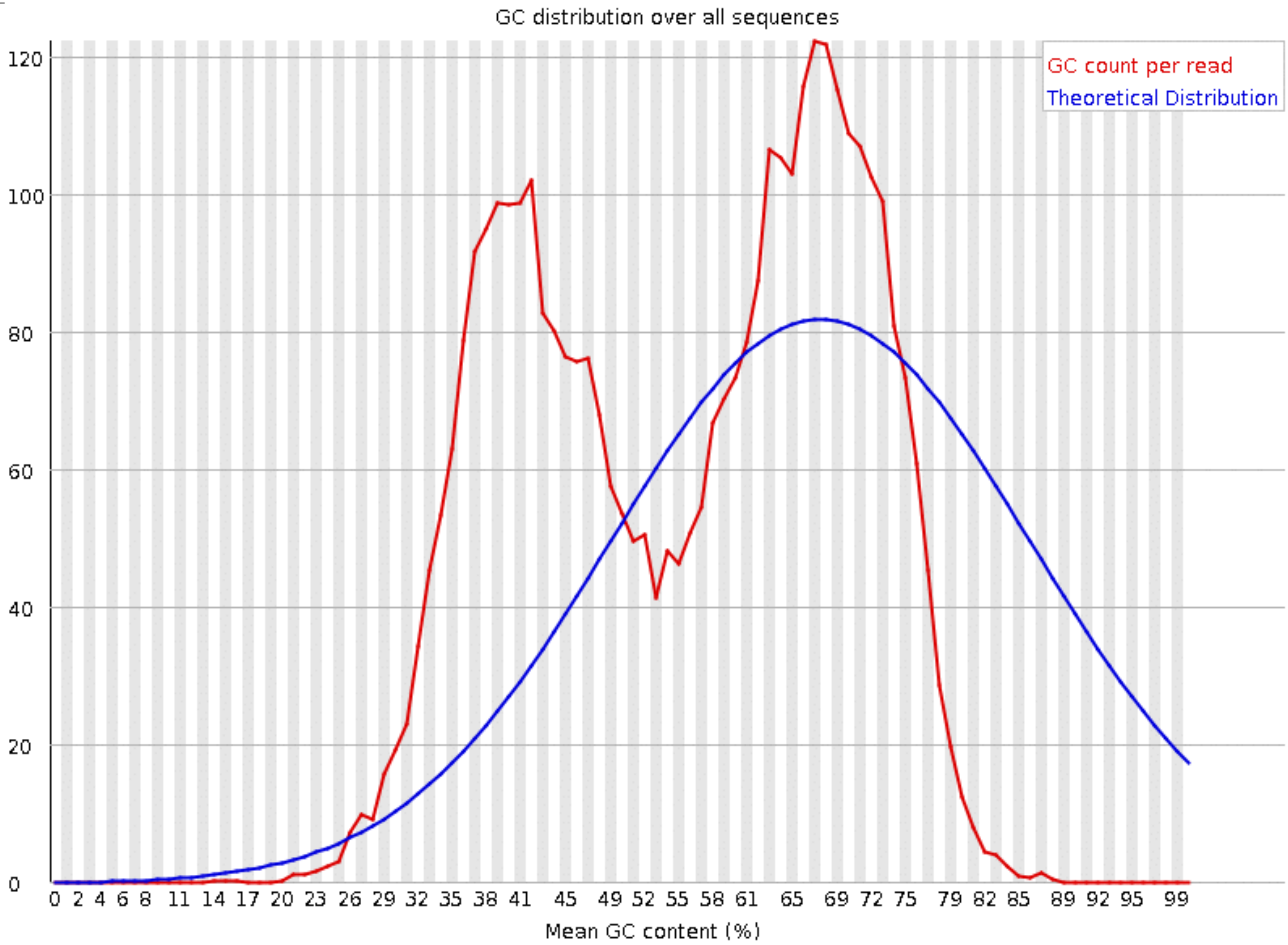
cDNA — Per base quality



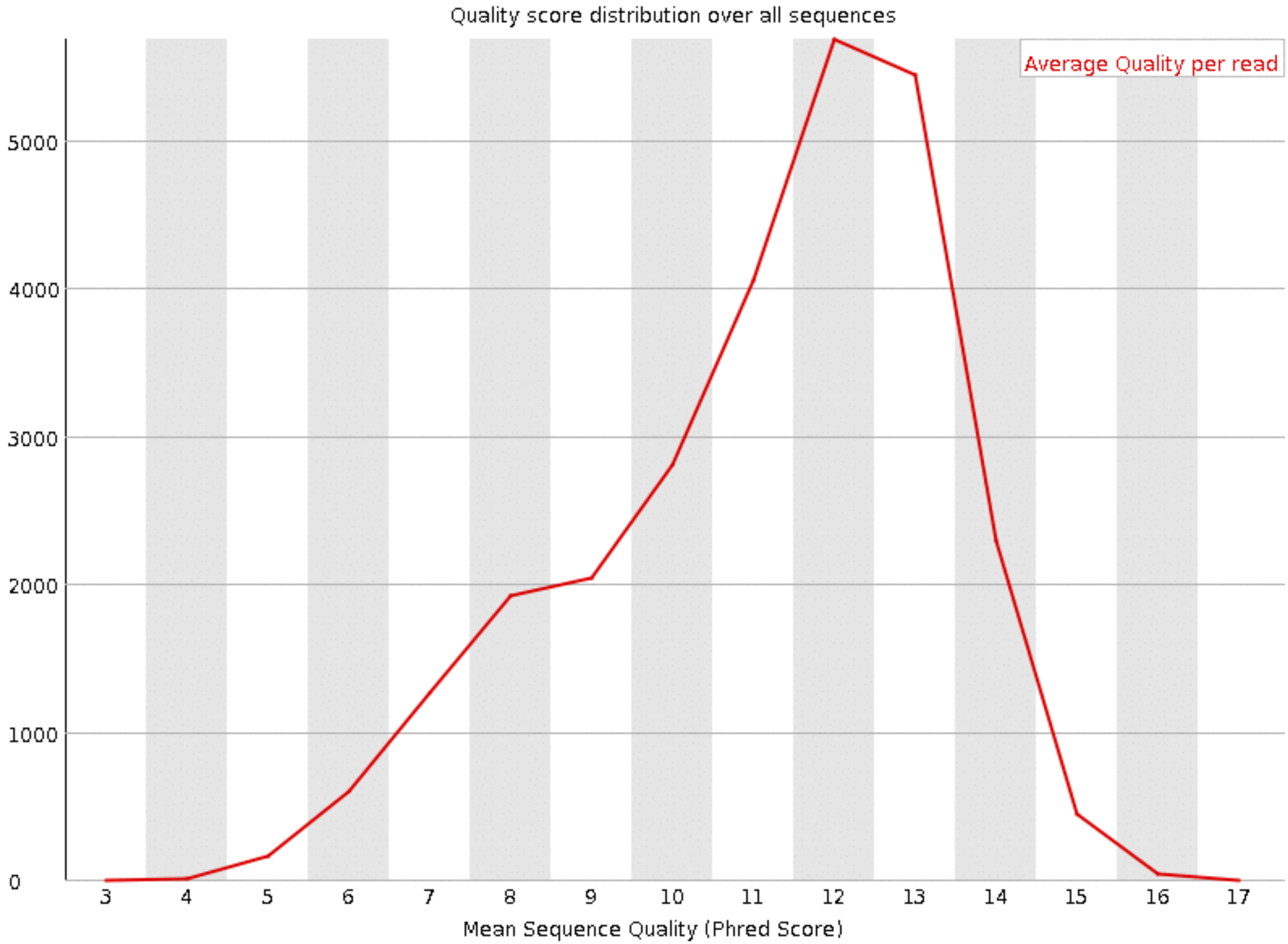
cDNA — Per base sequence content



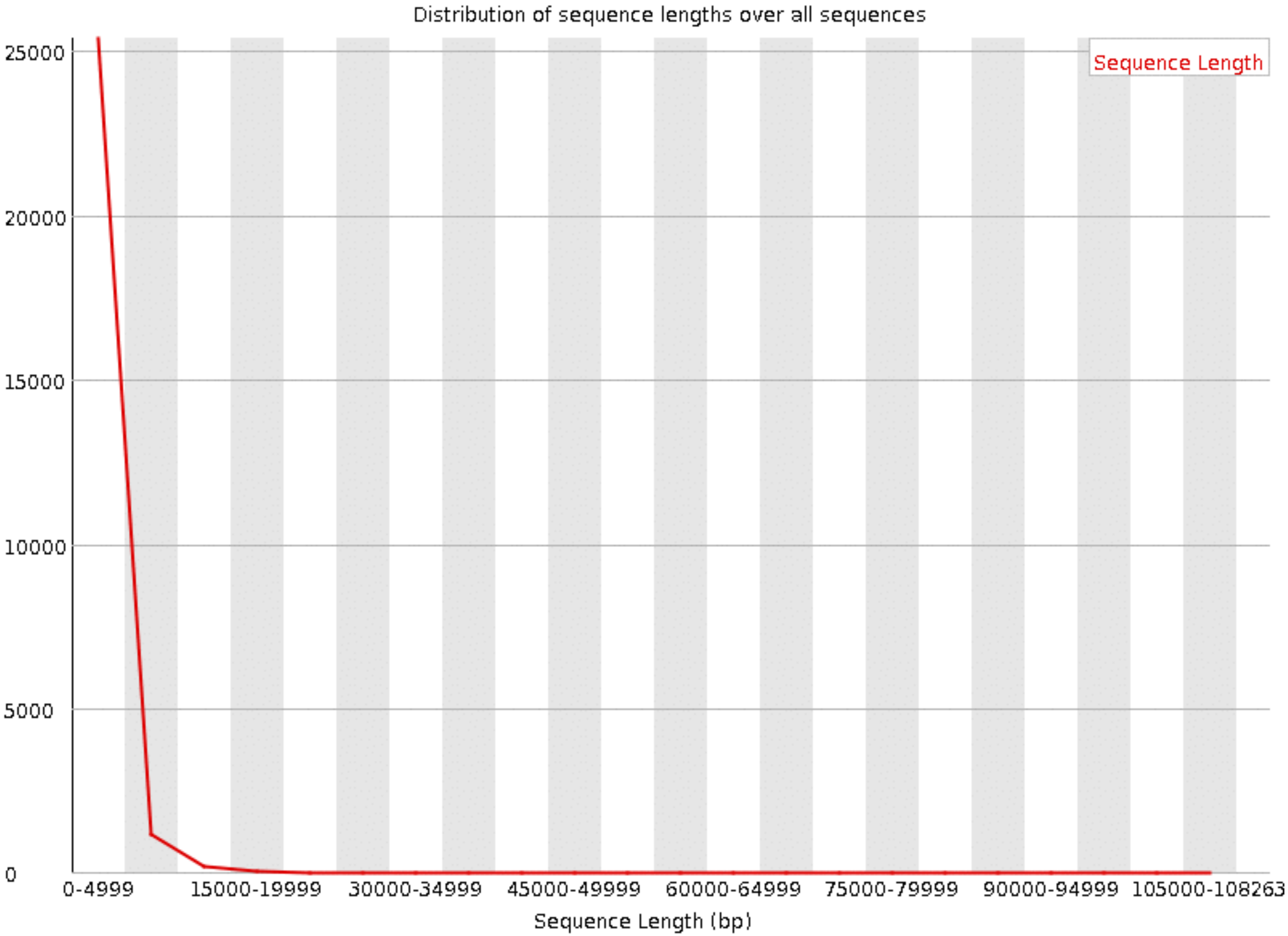
cDNA — Per sequence GC content



cDNA — Per sequence quality



cDNA — Sequence length distribution


Distribution of sequence lengths over all sequences
Sequence Length
Sequence Length (bp)