

Best practices in the analysis of RNA-seq and ChIP-seq data

27th – 31st, July 2015

University of Cambridge, Cambridge, UK

Quality assessment of NGS data

Ines de Santiago

CRUK Cambridge Research Institute

Ines.desantiago@cruk.cam.ac.uk



**UNIVERSITY OF
CAMBRIDGE**



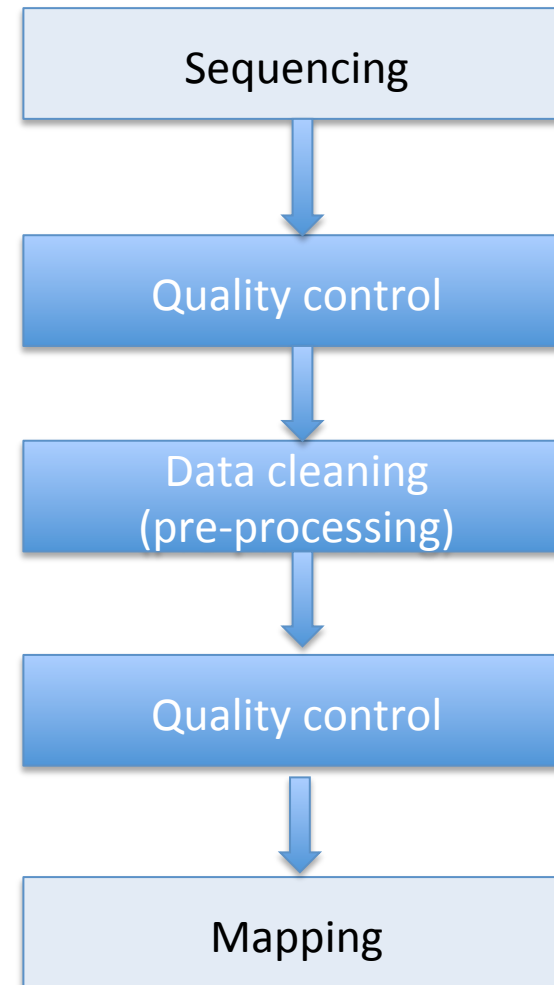
Quality control analysis

All sequencing platform have errors



Quality control

- It is important to check the quality of your sequenced reads!
- FASTQC: free program that reports quality profile of reads
- Pre-processing
 - Trim reads
 - exclude low quality reads
 - contaminations



Checking read quality with FASTQC

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

1. Run FASQC

fastqc sample.fastq

2. Open output file

sample_fastq.html

Summary



[Basic Statistics](#)



[Per base sequence quality](#)



[Per tile sequence quality](#)



[Per sequence quality scores](#)



[Per base sequence content](#)



[Per sequence GC content](#)



[Per base N content](#)



[Sequence Length Distribution](#)



[Sequence Duplication Levels](#)



[Overrepresented sequences](#)



[Adapter Content](#)



[Kmer Content](#)

FASTQC: Report

- 1) Basic statistics
- 2) Per base sequence quality
- 3) Per tile sequence quality
- 4) Per sequence quality scores
- 5) Per base sequence content
- 6) Per sequence GC content
- 7) Per base N content
- 8) Sequence Length Distribution
- 9) Sequence duplication levels
- 10) Over-represented sequences
- 11) Adapter/Kmer content



Basic Statistics

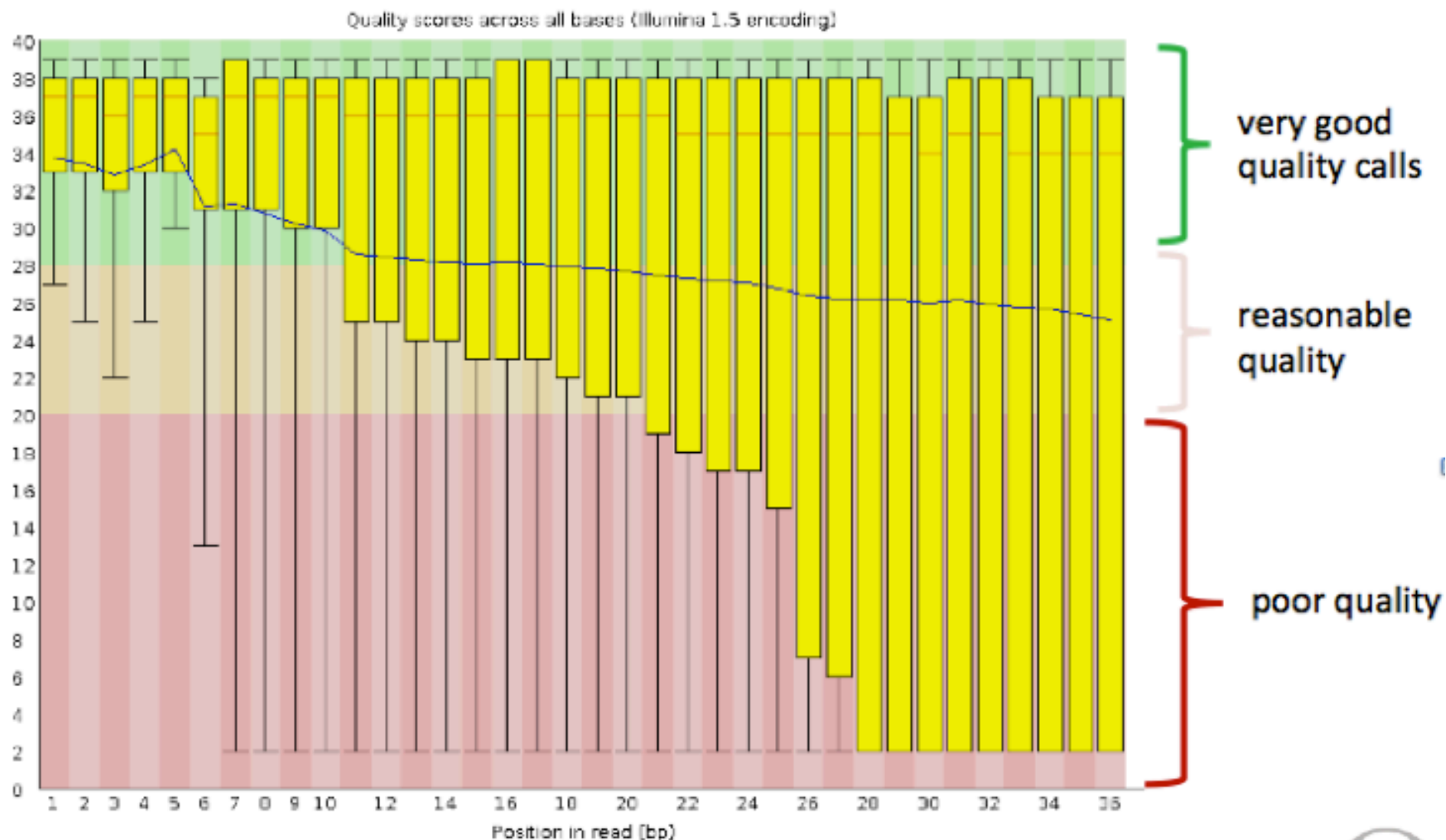
Measure	Value
Filename	sample.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	9053
Sequences flagged as poor quality	0
Sequence length	36
%GC	50

(2) FASTQC: Per base sequence content

- Poor quality at the end of reads



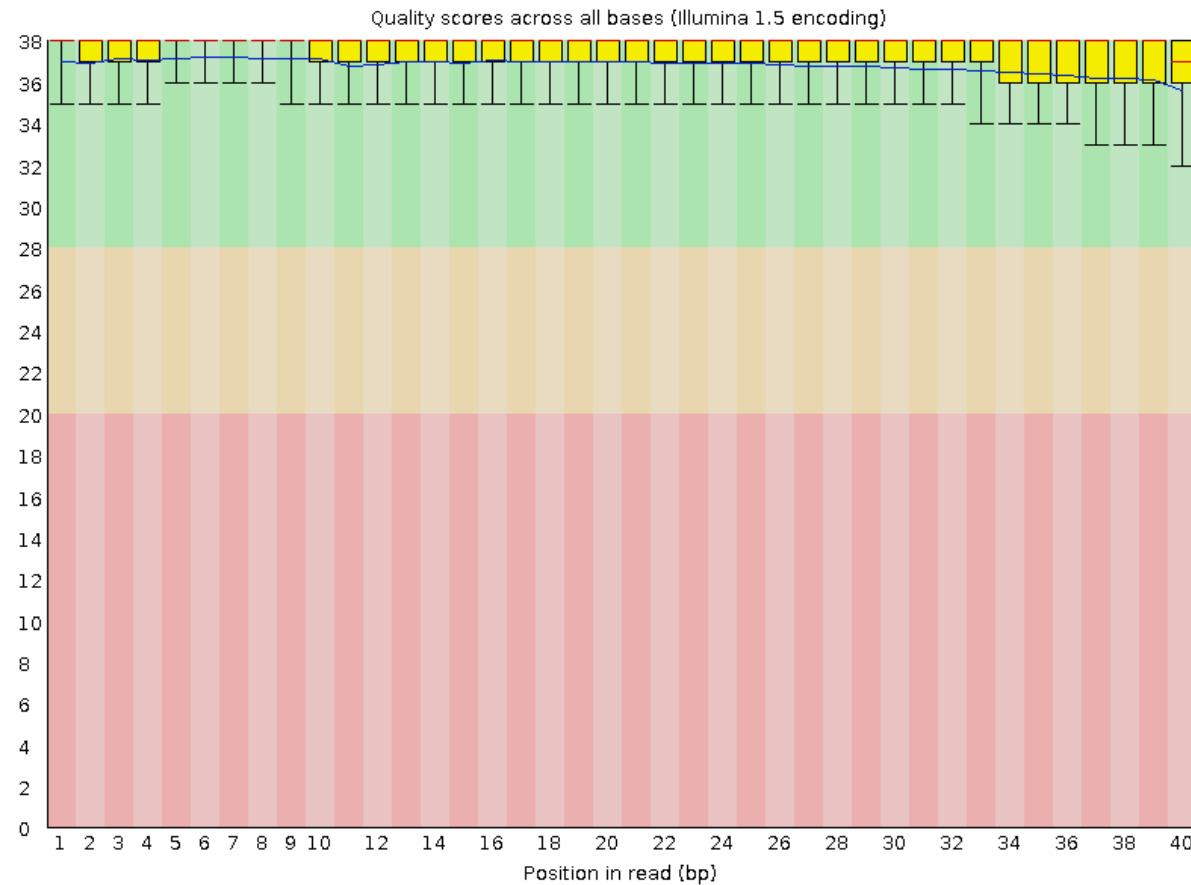
(2) FASTQC: Per base sequence content



(2) FASTQC: Per base sequence content

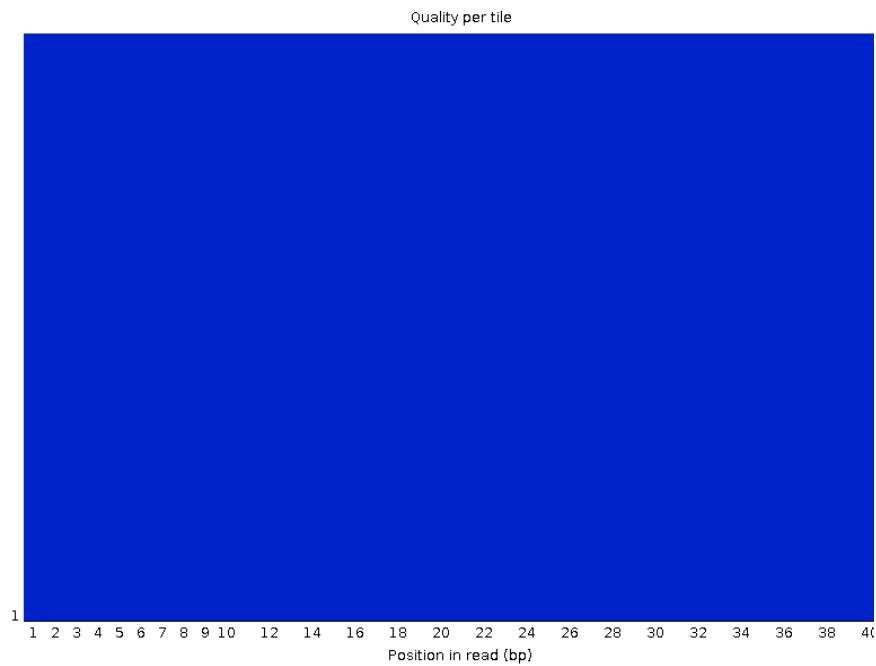
Good Illumina data:

✅ **Per base sequence quality**

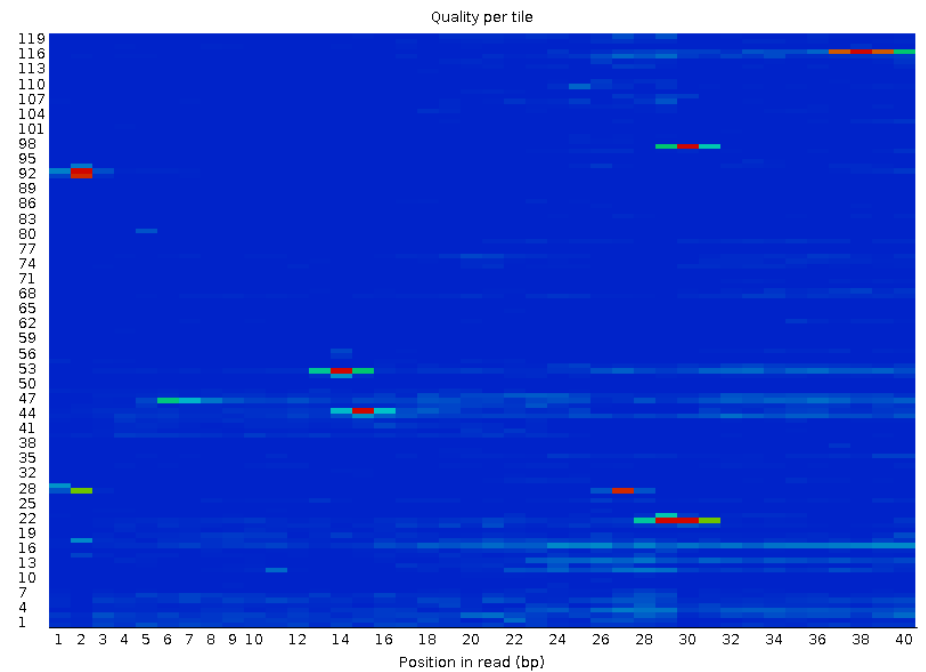


(3) FASTQC: Per tile sequence quality

✓ Per tile sequence quality

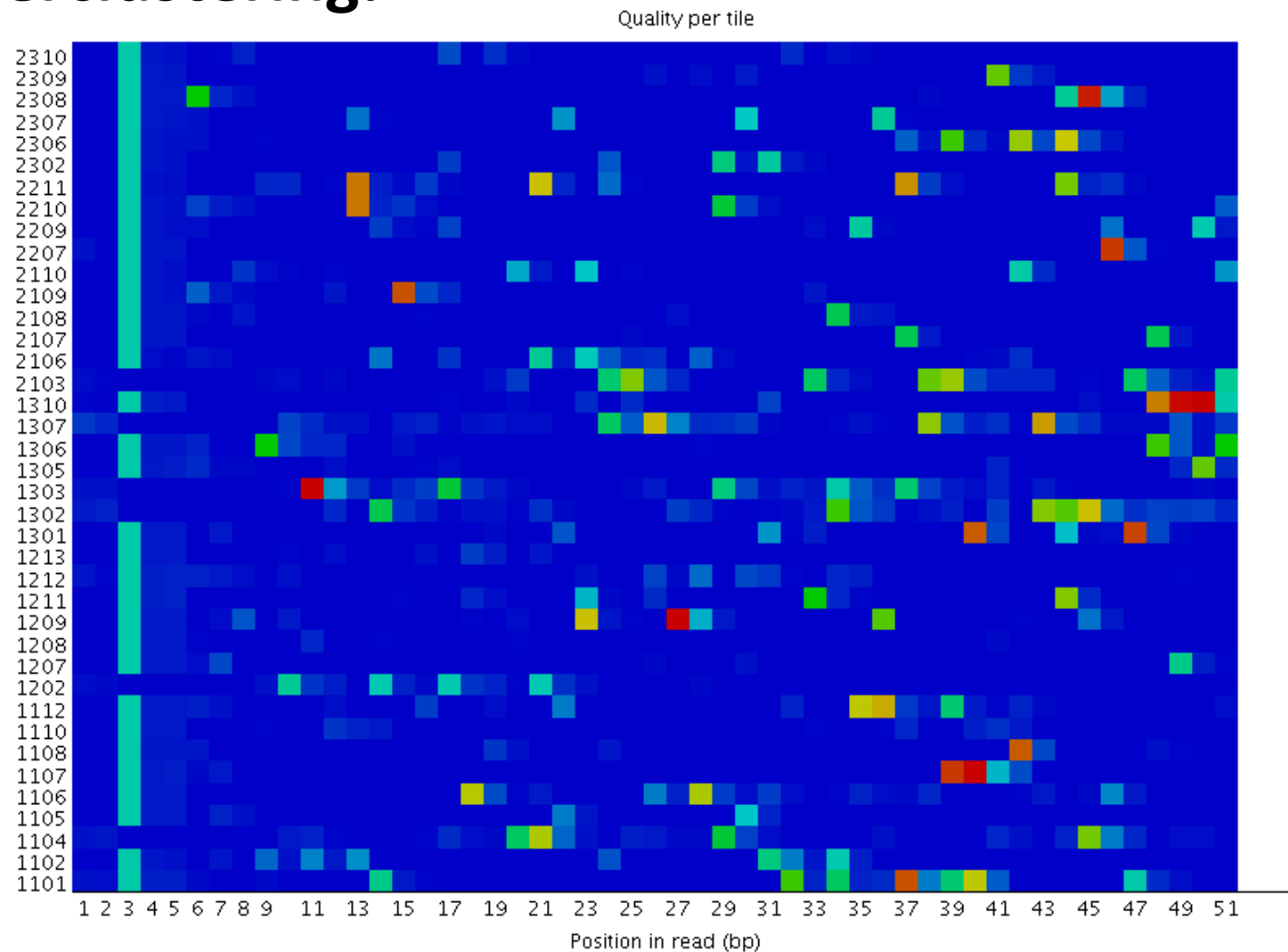


✗ Per tile sequence quality



(3) FASTQC: Per tile sequence quality

Overclustering:

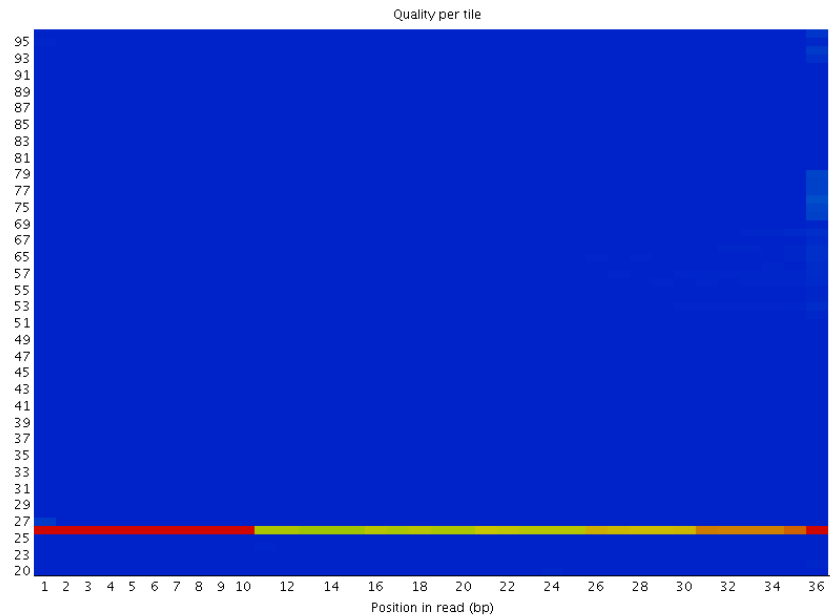


(3) FASTQC: Per tile sequence quality

Tile fail:

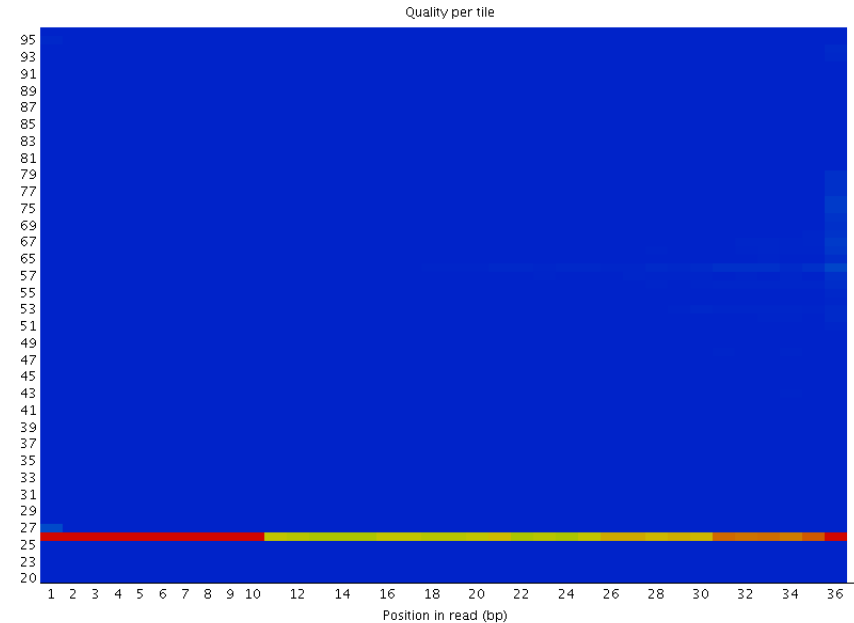
SRR576938
anaerobic INPUT DNA

✖ Per tile sequence quality



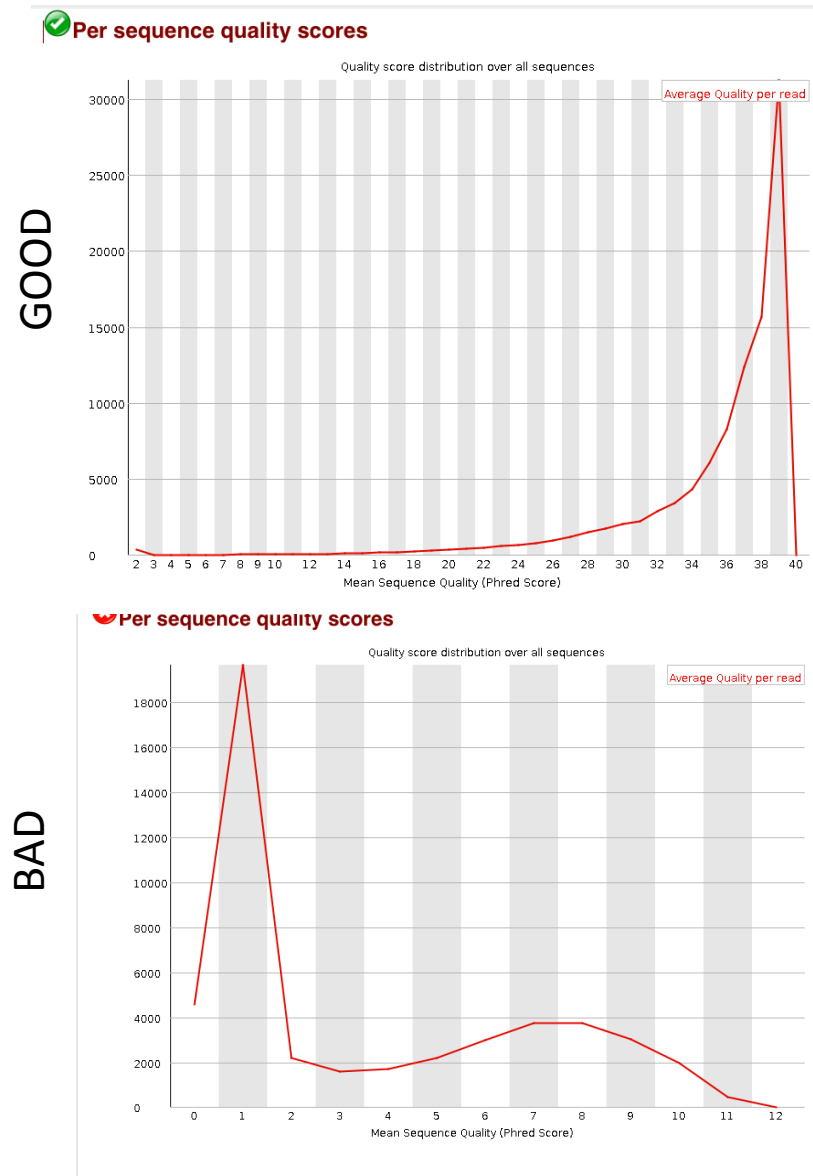
SRR576933
FNR IP ChIP-seq Anaerobic A

✖ Per tile sequence quality



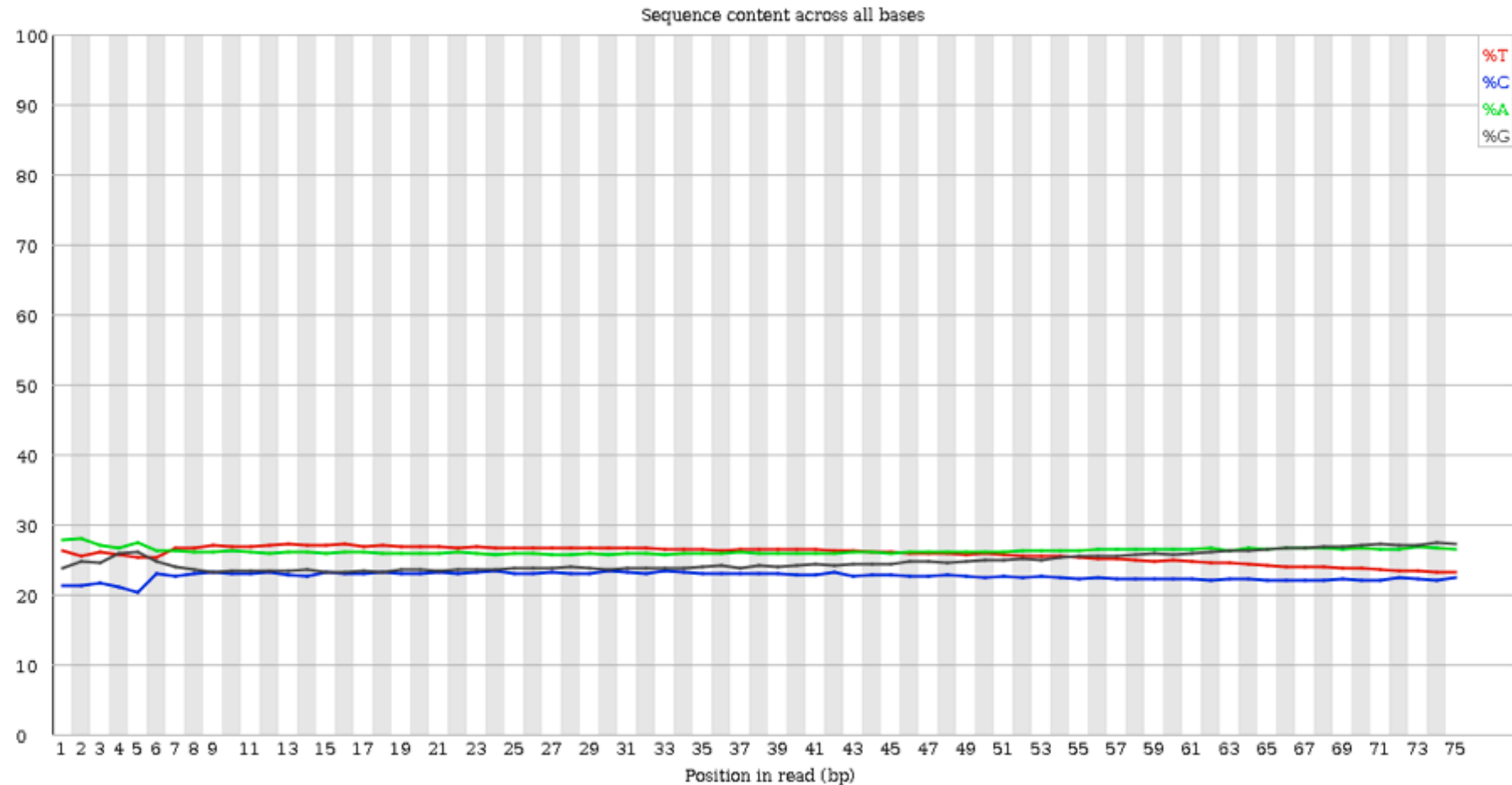
GSE41187: Genome-wide analysis of FNR and s70 in E. coli under aerobic and anaerobic growth conditions: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41187>

(4) FASTQC: Per sequence quality scores



(5) FASTQC: Per base sequence content

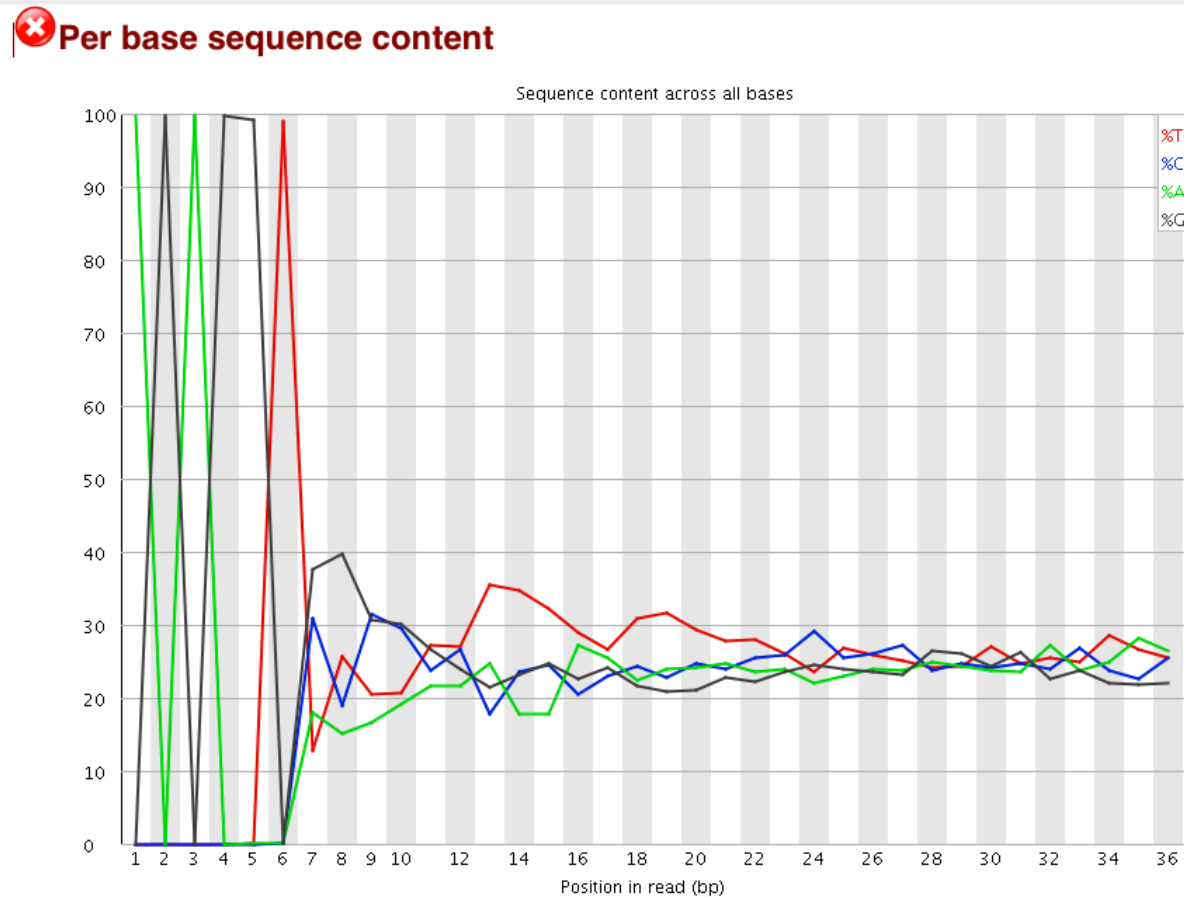
✔ Per base sequence content



http://bio-hpc.kisti.re.kr/MDS_03_normal_chr21.1.fq_fastqc/fastqc_report.html#M3

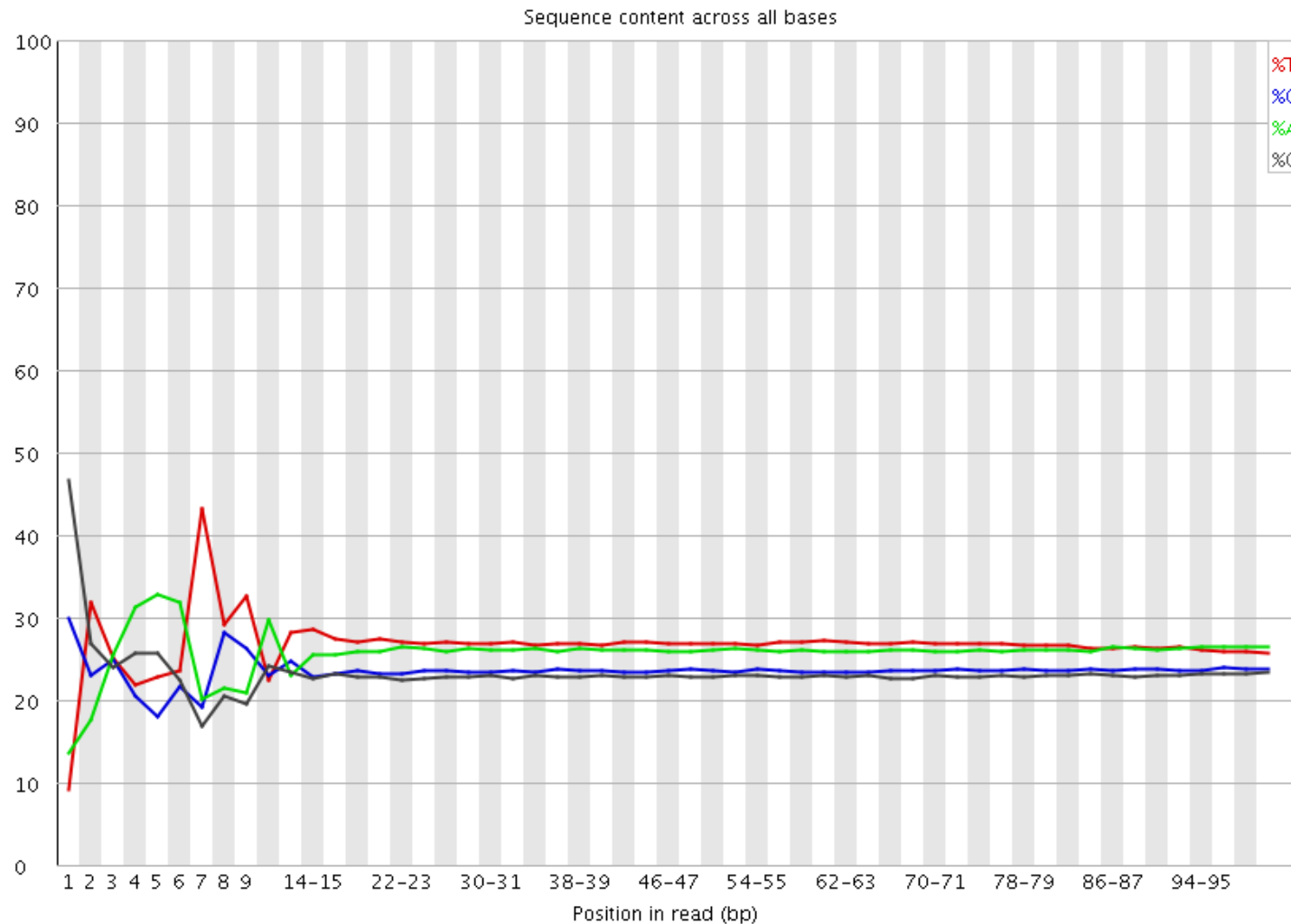
(5) FASTQC: Per base sequence content

Biased sequence composition (adapters?)



(5) FASTQC: Per base sequence content

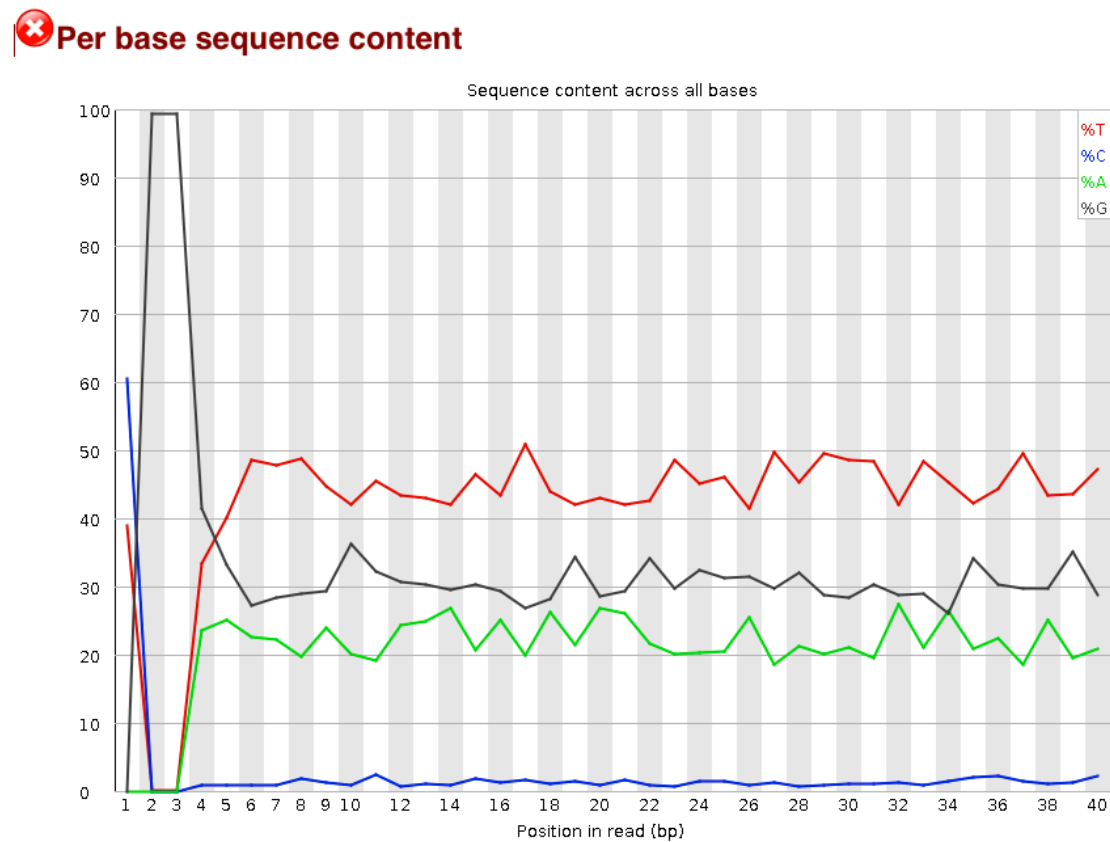
Unavoidable – RNA-Seq



(5) FASTQC: Per base sequence content

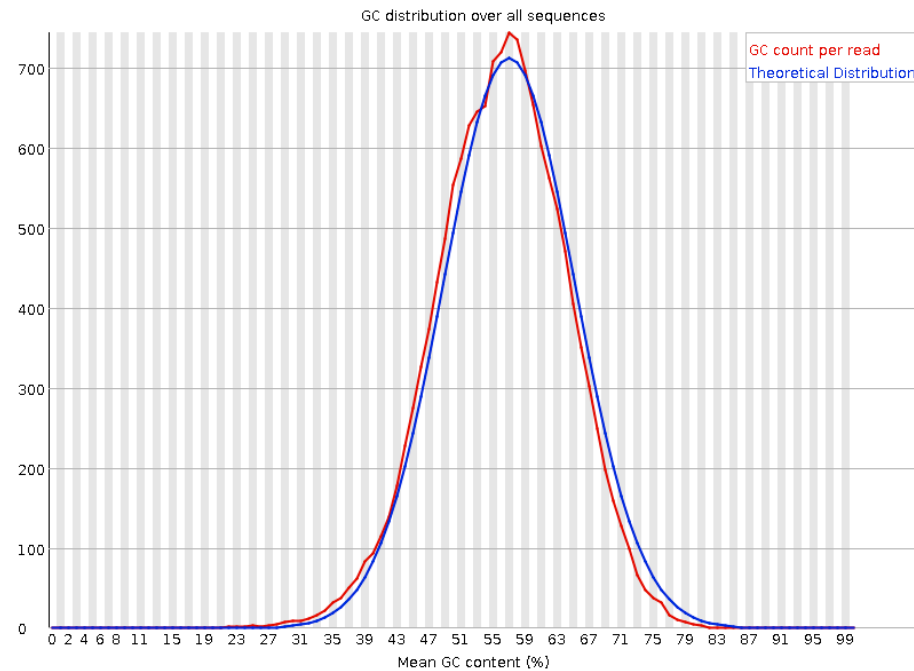
Unavoidable – RRBS

Devoided of cytosines because the library was treated with sodium bisulphite (which will have converted most of the C to T)

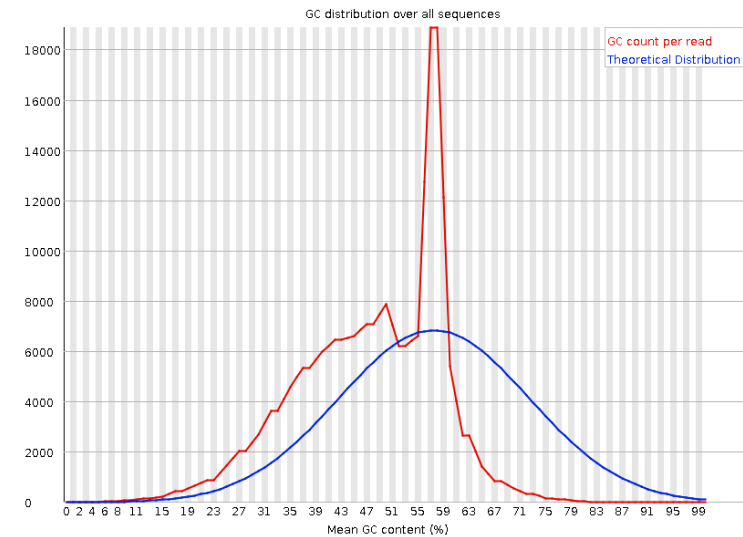


(6) FASTQC: Per sequence GC content

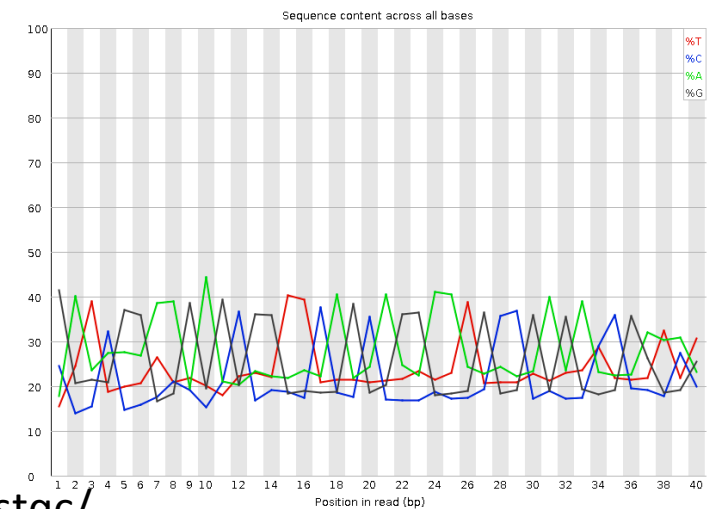
✓ Per sequence GC content



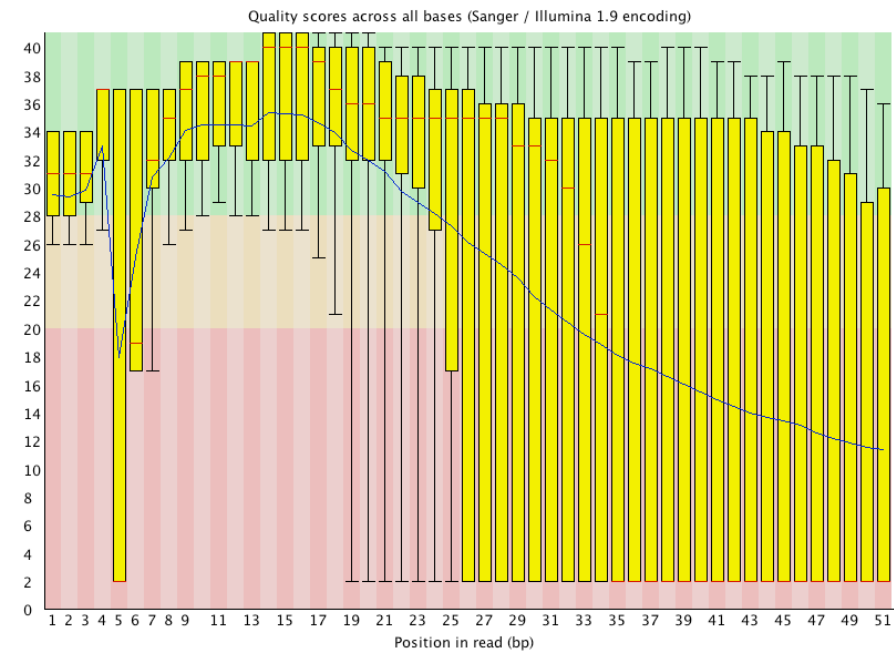
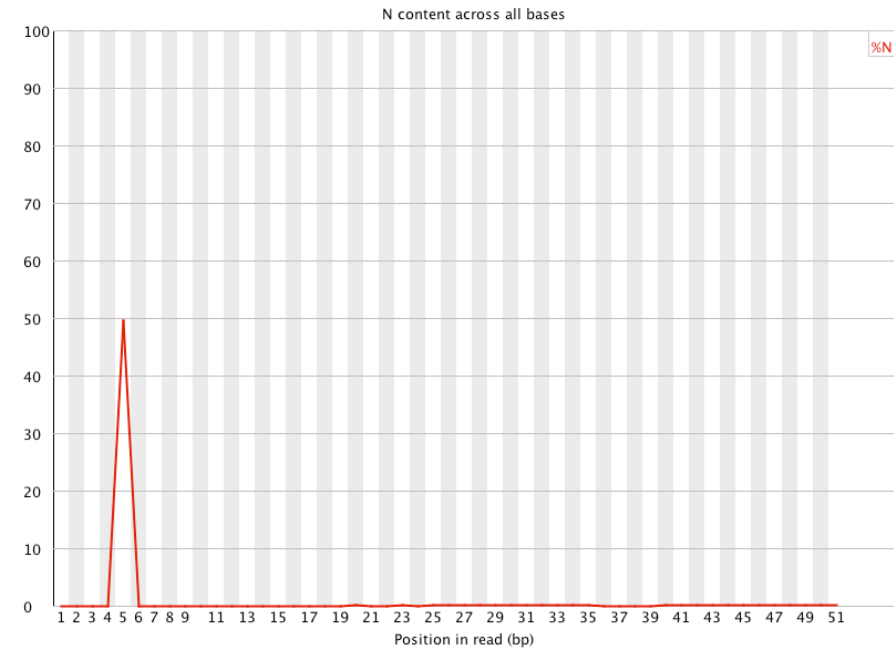
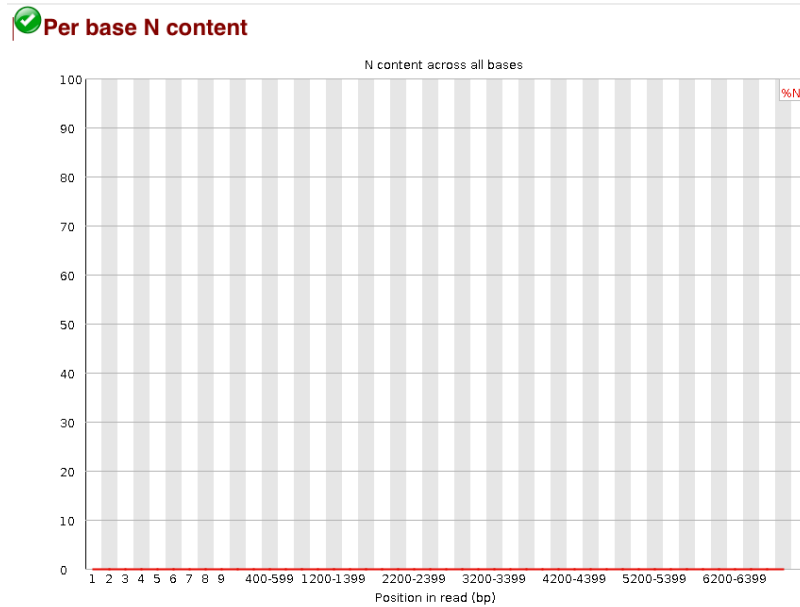
✗ Per sequence GC content



✗ Per base sequence content



(7) FASTQC: Per base N content



<http://cbio.mskcc.org/~lianos/files/scott/2011-11-21/qc/>

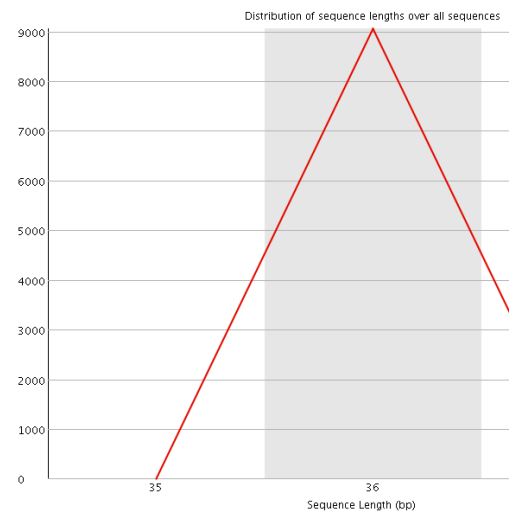
(8) FASTQC: Sequence Length Distribution

Summary

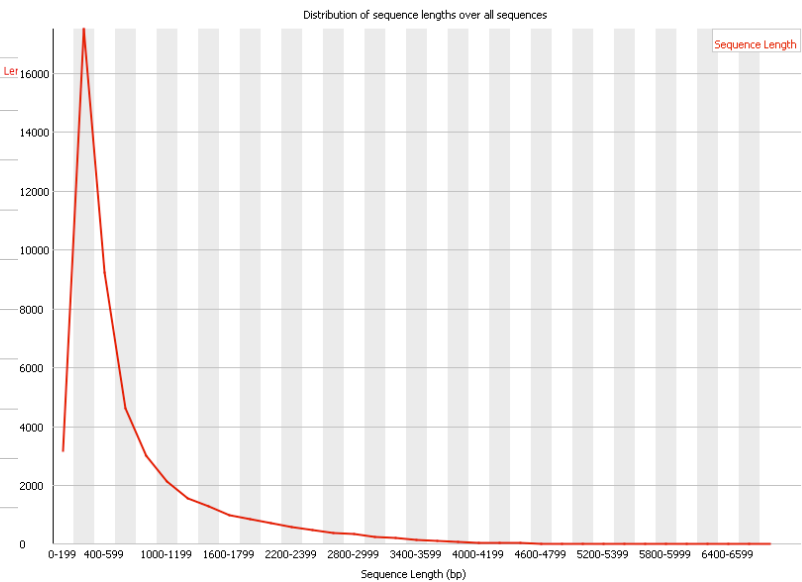
- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✗ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✗ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Kmer Content

Sequence fragments of uniform length (36bp)

Sequence Length Distribution



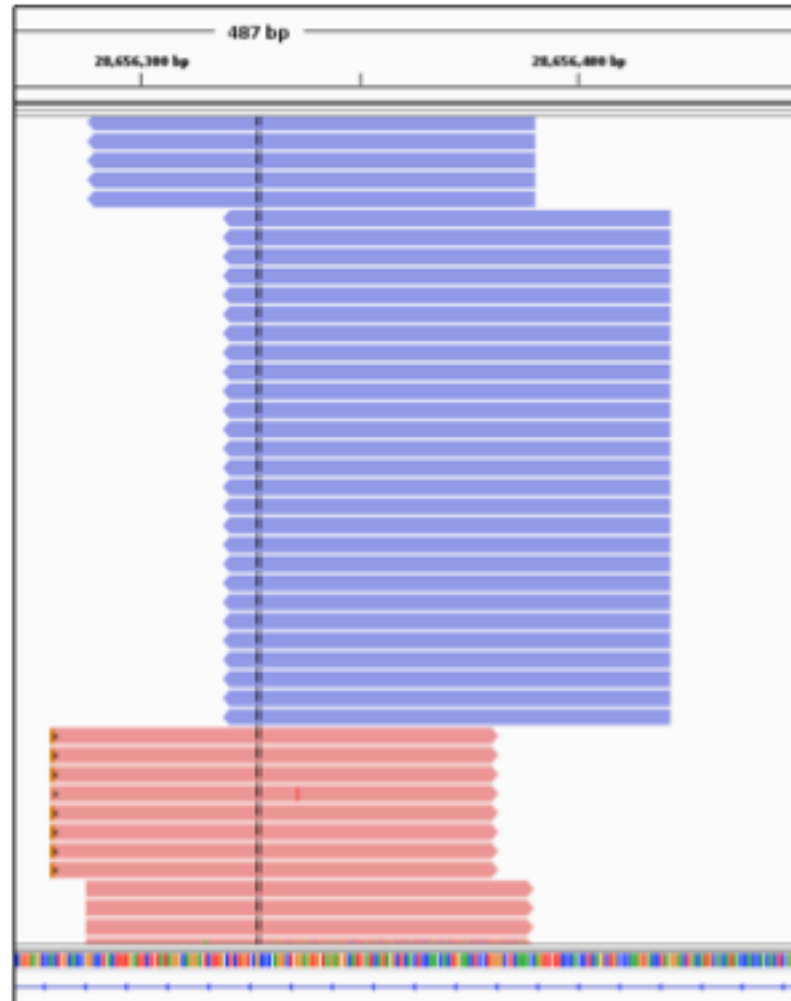
Reads of variable length:



http://cbio.mskcc.org/~lianos/files/scott/2011-11-21/qc/Bcnc2_ATCACG_L001_R1_001_fastqc/fastqc_report.html#M2

(9) FASTQC: Sequence duplication levels

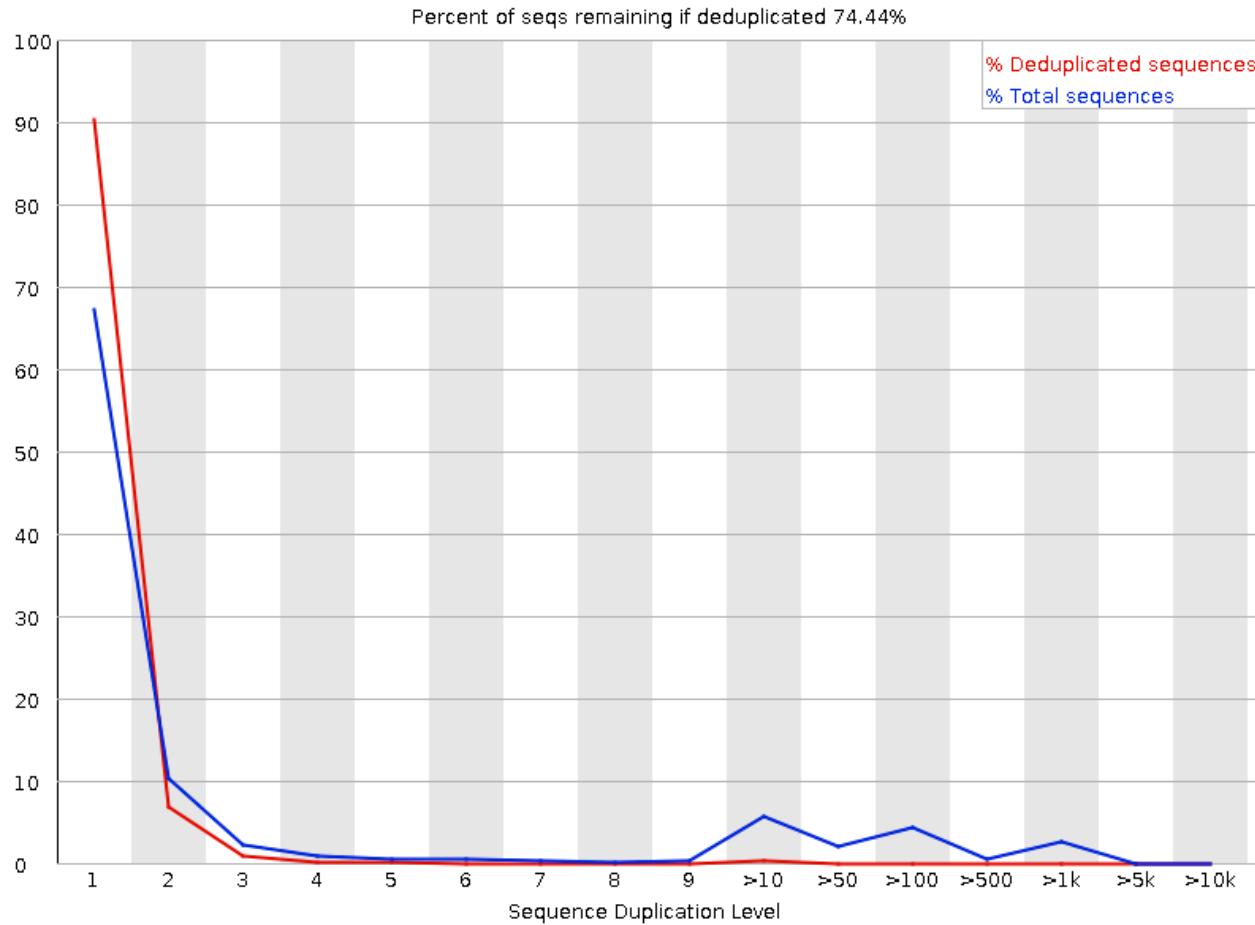
- PCR duplicates during sample preparation
- Optical duplicates: read the same cluster twice in the sequencer
- High duplication can lead to problems in downstream analysis (e.g. skew allele frequencies)



(9) FASTQC: Sequence duplication levels

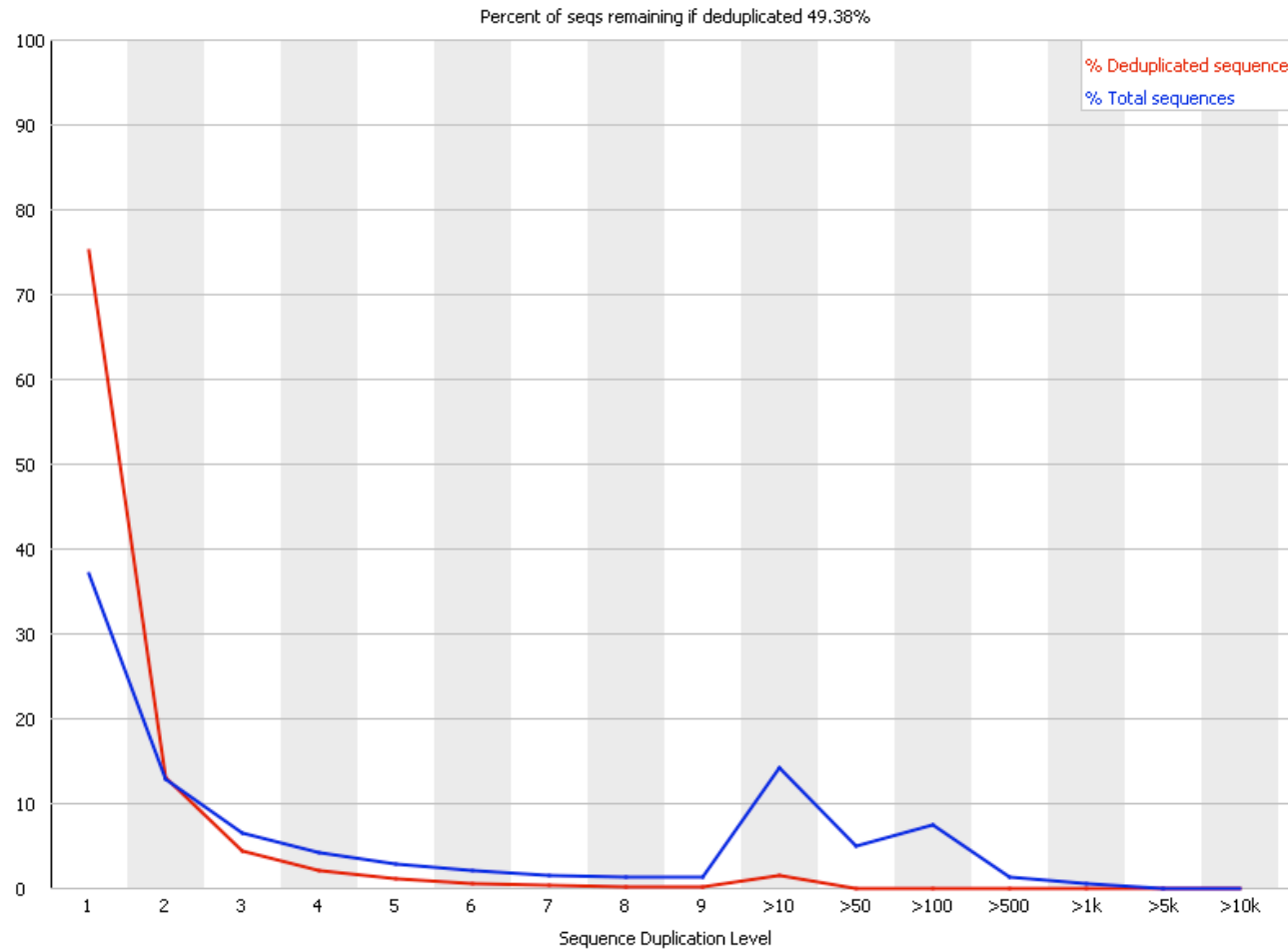
Very diverse library

✔ Sequence Duplication Levels



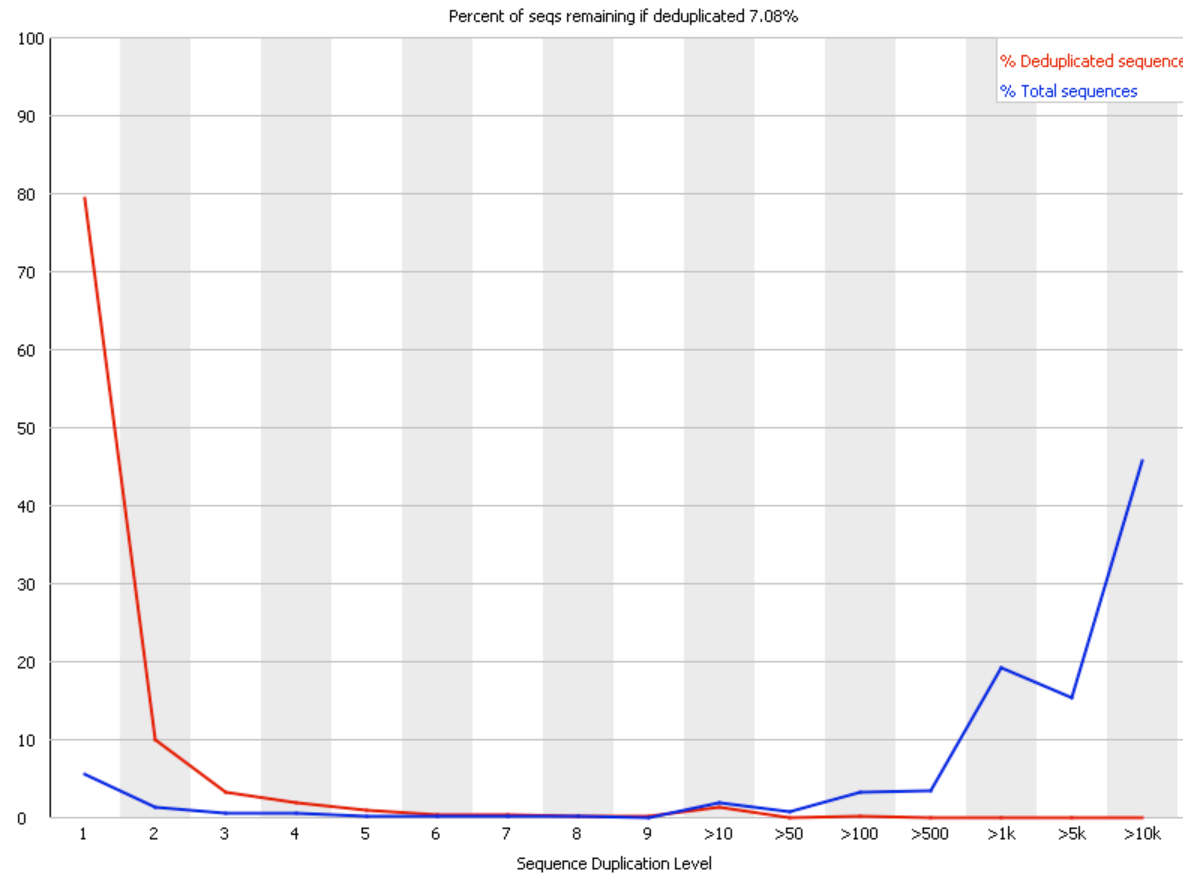
(9) FASTQC: Sequence duplication levels

A good RNA-Seq library (although dup levels > 50%)



(9) FASTQC: Sequence duplication levels

PCR duplication



<http://proteo.me.uk/2013/09/a-new-way-to-look-at-duplication-in-fastqc-v0-11/>

(10) FASTQC: Over-represented sequences

Good dataset



Overrepresented sequences

No overrepresented sequences

Bad datasets:



Overrepresented sequences

[illegible]

[Back to summary](#)

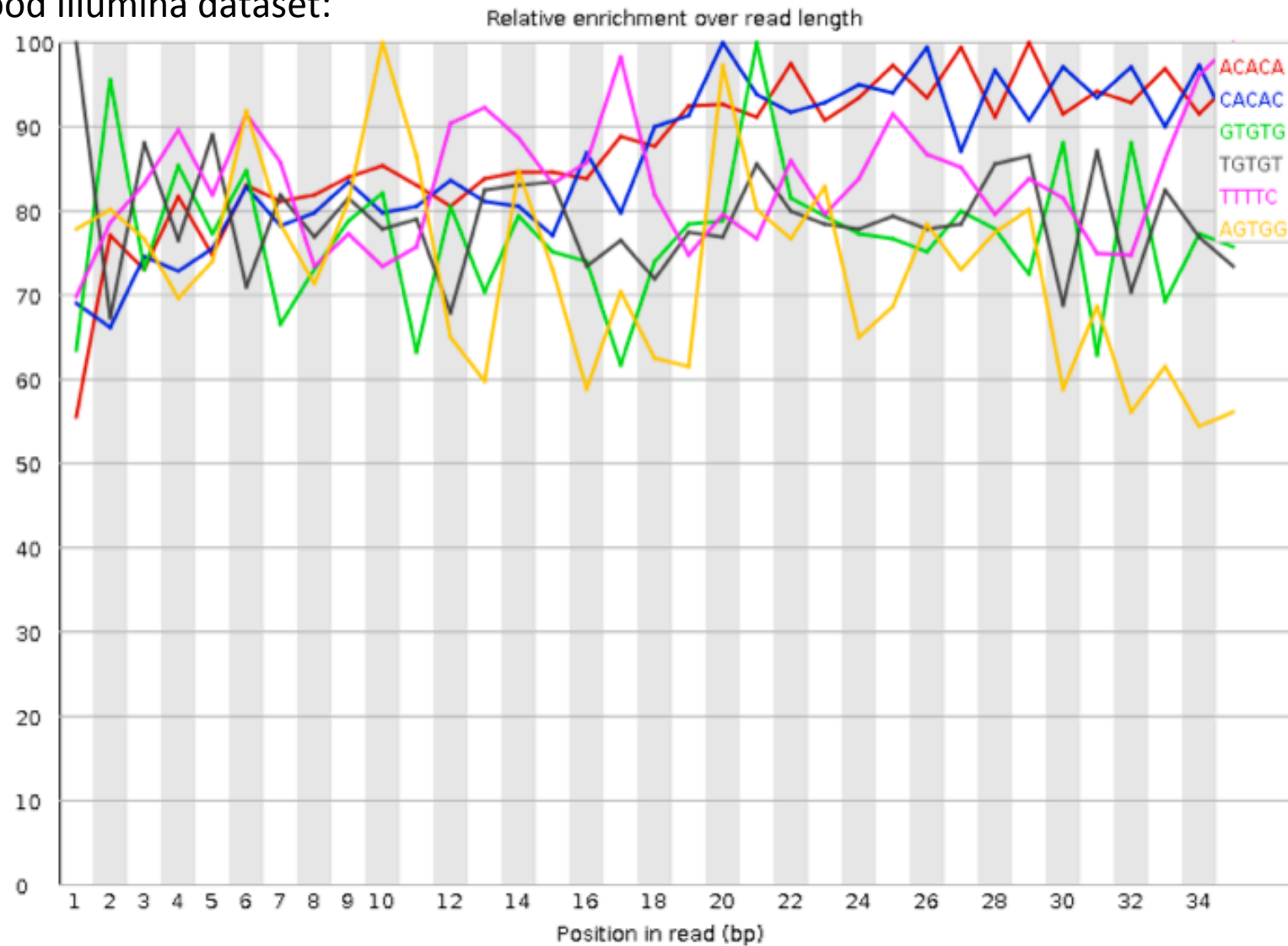


Overrepresented sequences

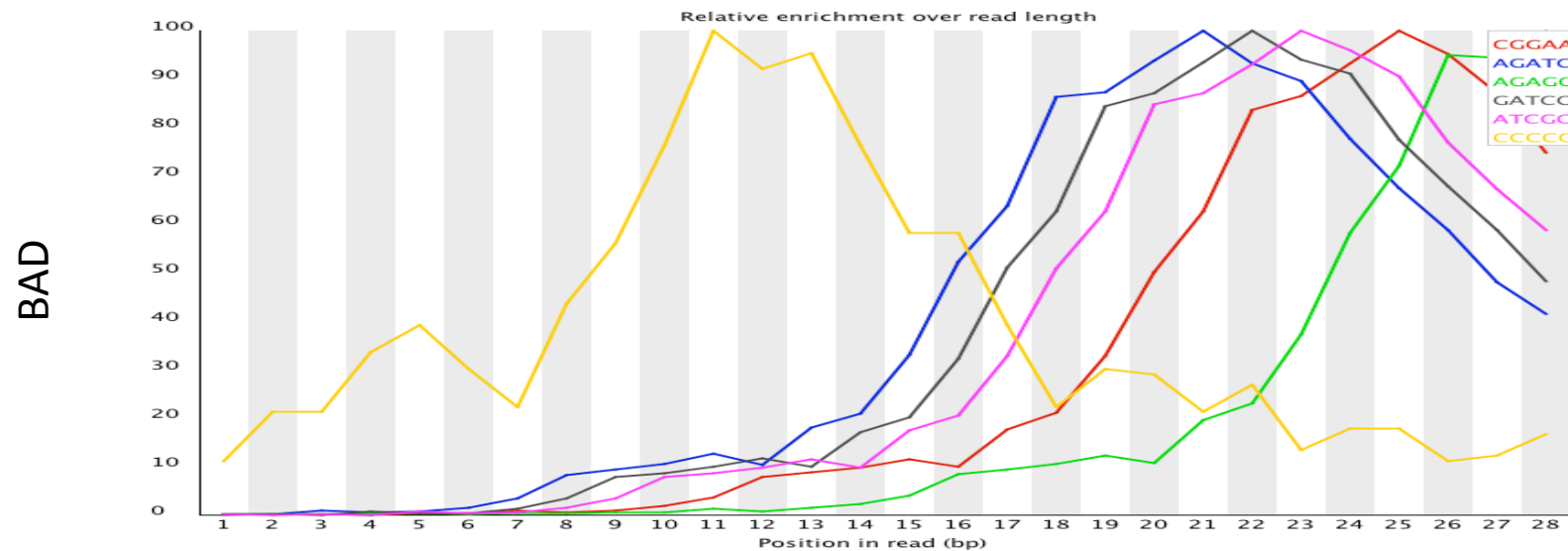
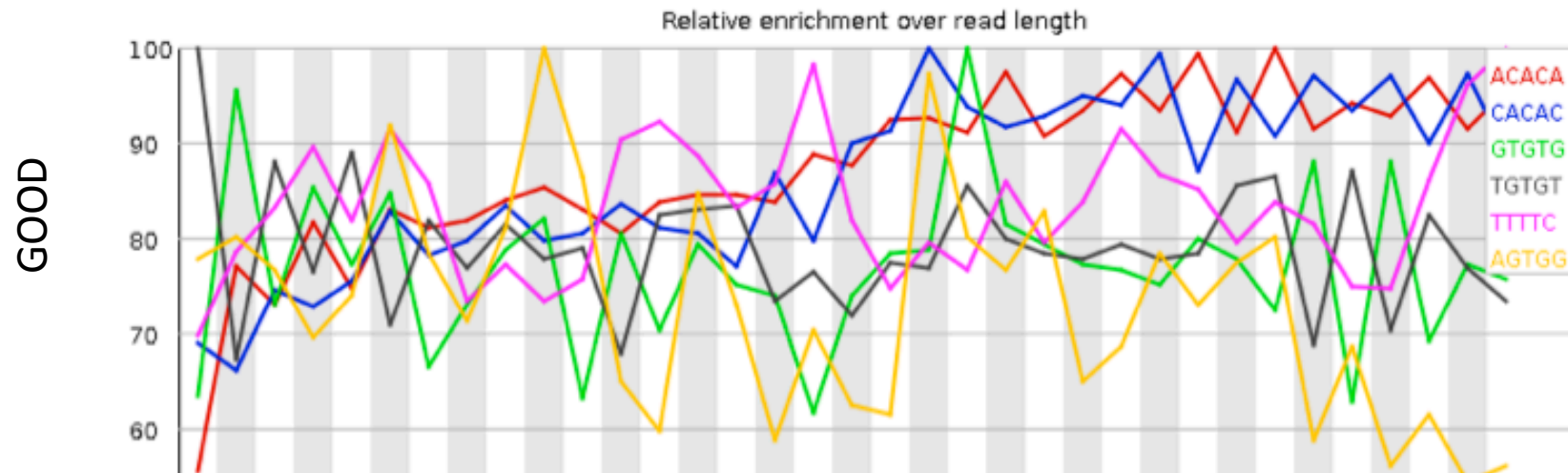
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	28971	28.971000000000004	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAATACCCGACTAAATCAGTCAAGTAAATA	392	0.392	No Hit
GTTAGCTATTTACTTGACTGATTTAGTCGGGTATTT	356	0.356	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	108	0.108	TruSeq Adapter, Index 1 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACG	107	0.107	TruSeq Adapter, Index 15 (97% over 36bp)

(11) FASTQC: Kmer content

Good Illumina dataset:



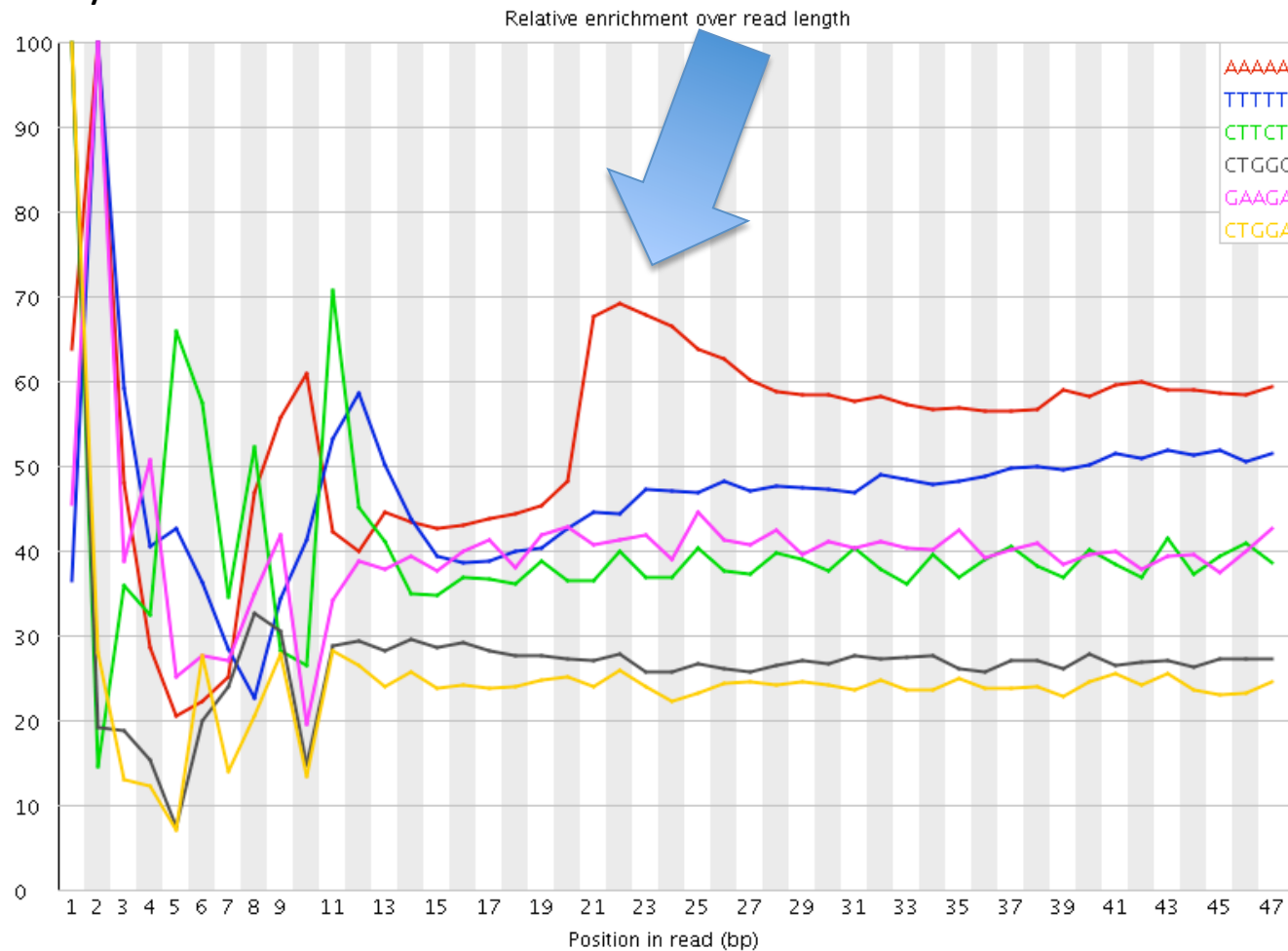
(11) FASTQC: Kmer content



<http://www.slideshare.net/suryasaha/sequencing-quality-filtering?related=1>

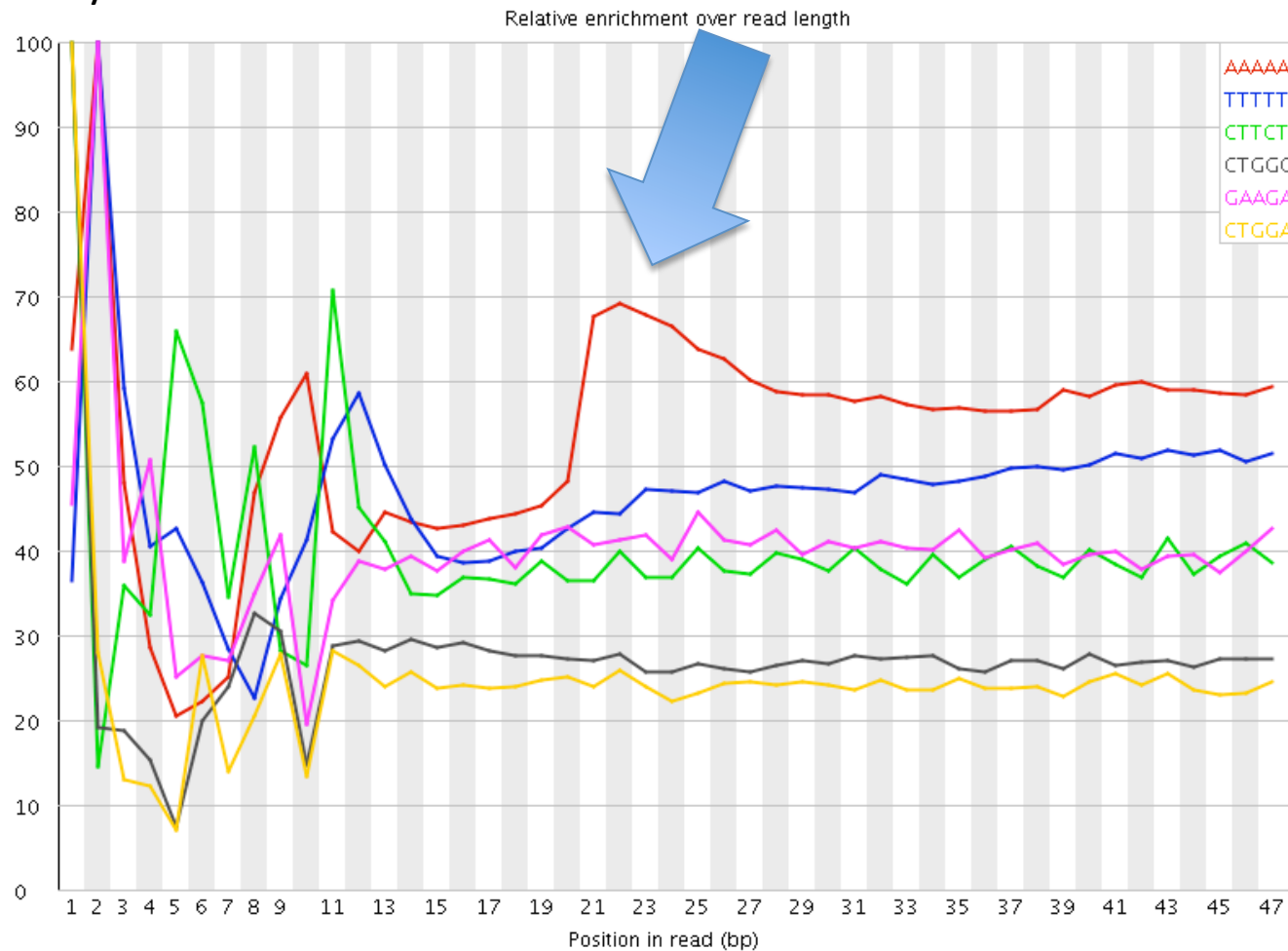
(11) FASTQC: Kmer content

AAAA k-mer that you're seeing at around 21 base pairs are arrested transcripts caused by cyclohexamide treatment.



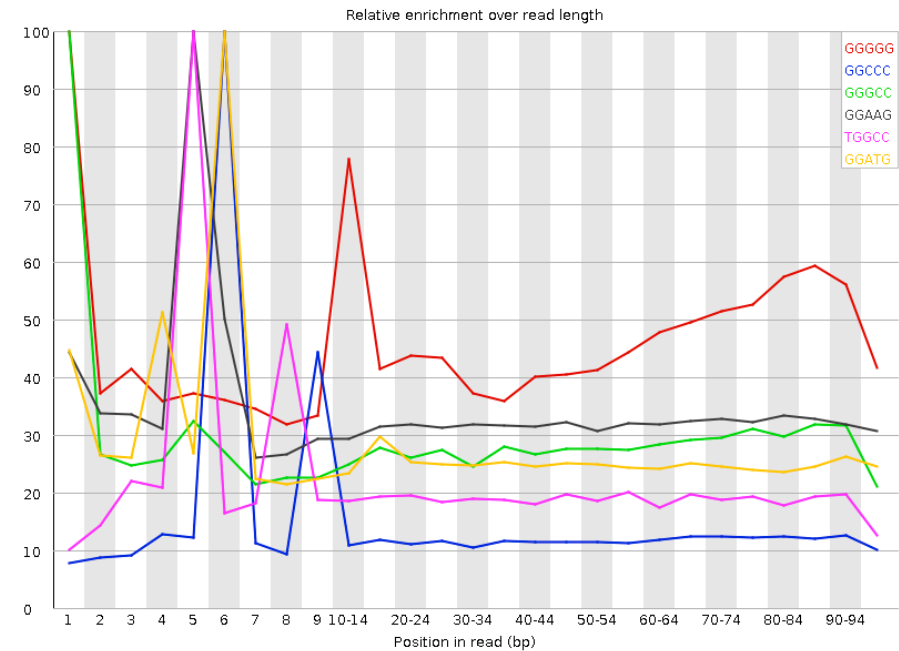
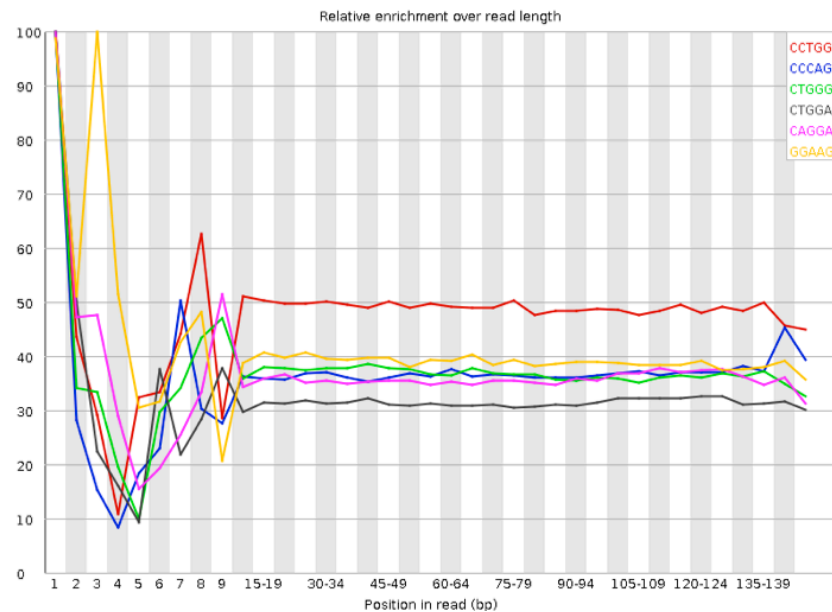
(11) FASTQC: Kmer content

AAAA k-mer that you're seeing at around 21 base pairs are arrested transcripts caused by cyclohexamide treatment.



(11) FASTQC: Kmer content

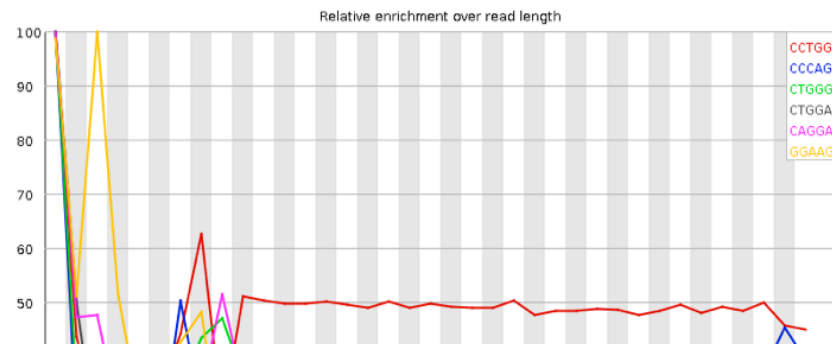
“Random” hexamer primer in RNA-seq libraries
(not that random after all)



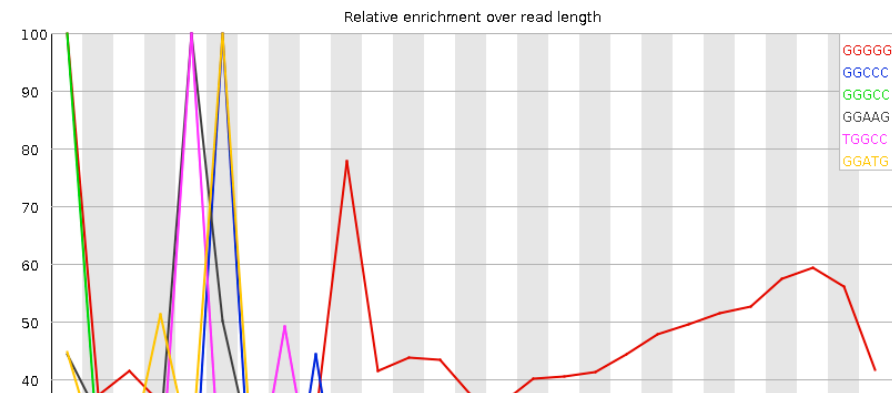
<http://seqanswers.com/forums/showthread.php?t=44770&highlight=kmer+fastq>
<http://seqanswers.com/forums/showthread.php?t=16669>

(11) FASTQC: Kmer content

“Random” hexamer primer in RNA-seq libraries
(not that random afterall)



Published online 14 April 2010



Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen^{1,*}, Steven E. Brenner² and Sandrine Dudoit^{1,3}

¹Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358,

²Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and

³Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Received December 1, 2009; Revised March 16, 2010; Accepted March 17, 2010

Hands on exercise:

Fastqc_sweave.pdf

Examples of FASTQC runs and preprocessing