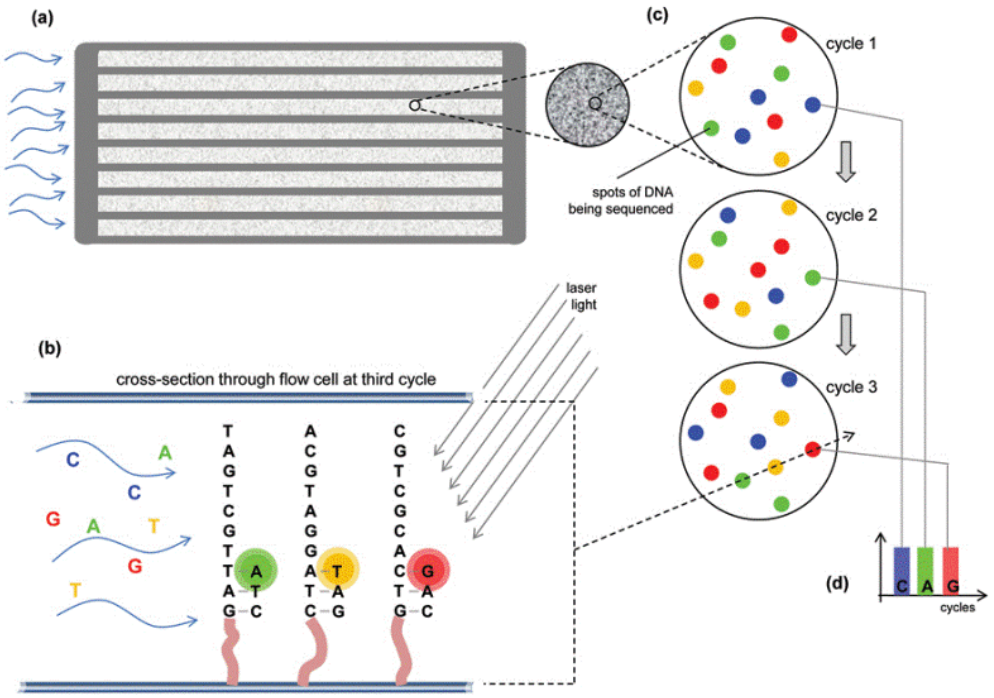
NGS Quality Control

Sandra Smit

RNAseq course, June 5-7

NBIC/EPS/WUR

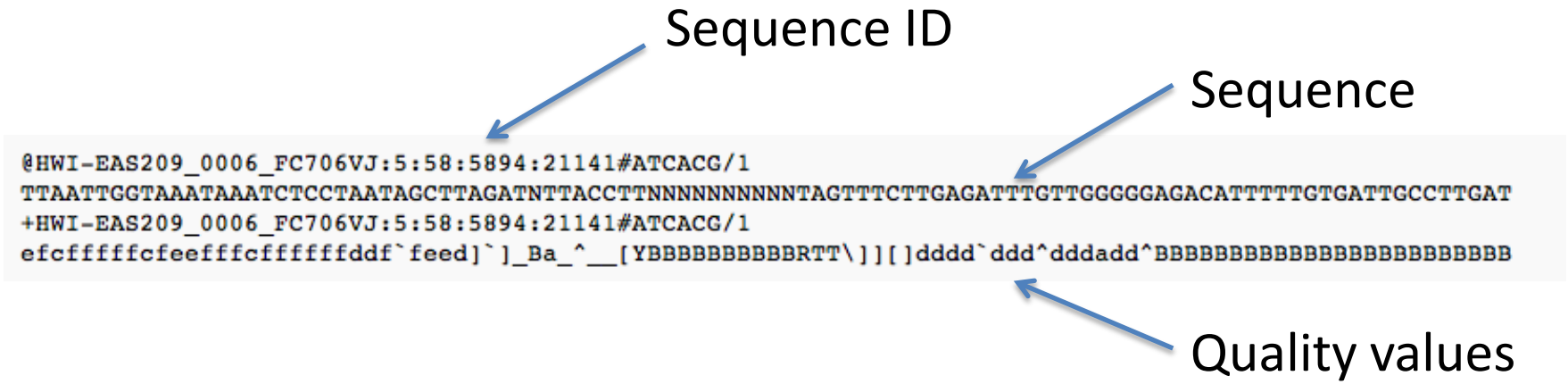
NGS data



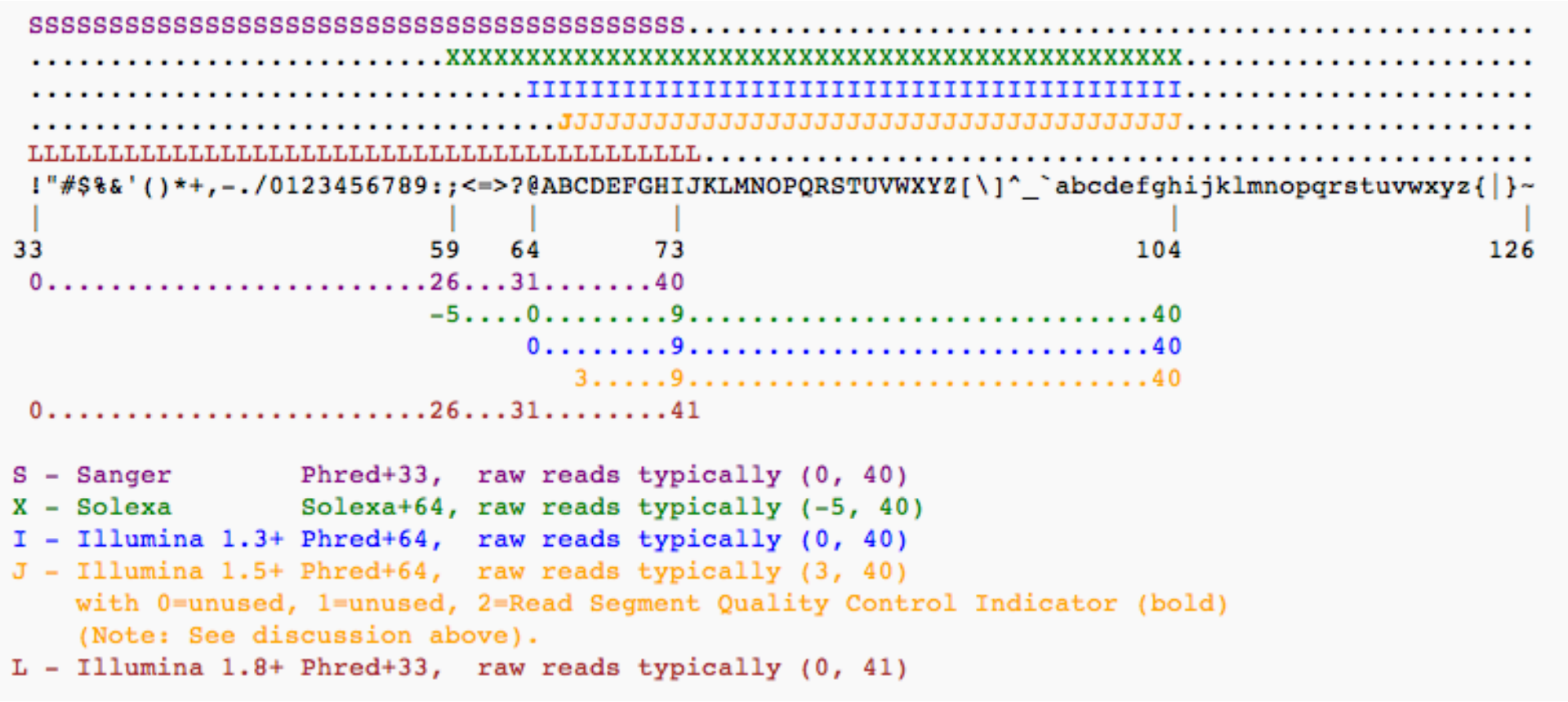
DATA: sequences (reads) and quality values

FASTQ

Many formats:
- FASTQ
- FASTA & QUAL
- SCARF
- SFF



FASTQ quality values



Quality Control (QC) of NGS data

- Quality assessment of data
- Improvement of quality (trimming and filtering)
  - selection: remove problematic reads
  - trimming: discard regions of low fidelity
  - correction: replace improbable basecalls
- Quality assessment again, see improvement?

What could be wrong?

- Base calling errors
- Uncalled bases
- GC bias
- Homopolymers
- Low quality bases (3’ end)
- Clonal duplicates
- Contamination (pathogens)
- Nuclear vs. organellar reads
- Sequencing artifacts (adaptors, vectors, clones, chimeric reads)

QC Tools

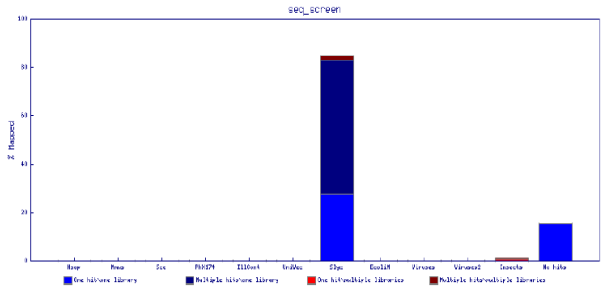
- FastQC
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- PINSEQ
  - http://prinseq.sourceforge.net/
- FASTX Toolkit
  - http://hannonlab.cshl.edu/fastx_toolkit/
- NGS QC Toolkit
  - http://www.nipgr.res.in/ngsqctoolkit.html
- NGS QC in Galaxy

QC and manipulation in Galaxy

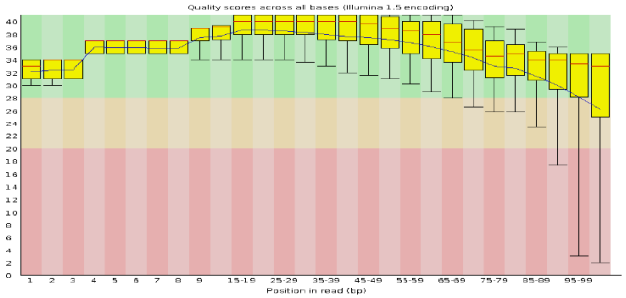
Manipulation of FASTQ data with Galaxy
Blankenberg et al.
Bioinformatics 2010

FastQC Quality Checking Tool

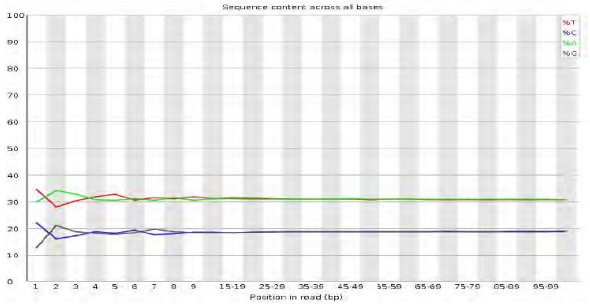
Contamination screen
fastq screen



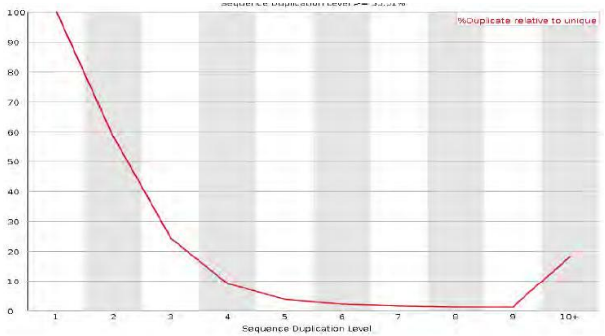
Per base sequence quality



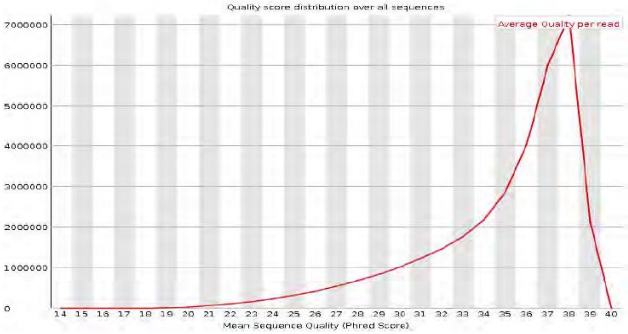
Per base sequence content



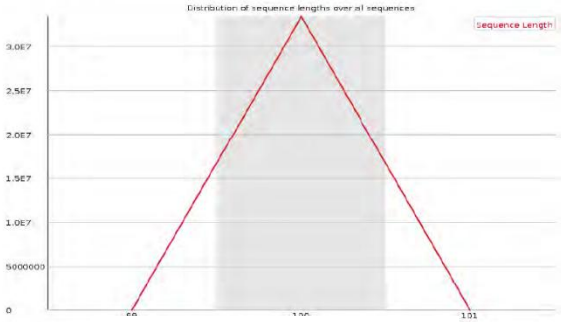
Sequence duplication



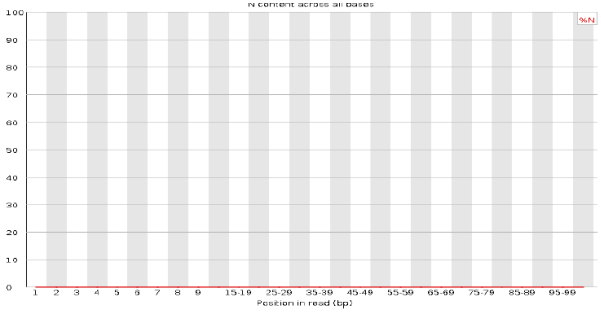
Per sequence quality



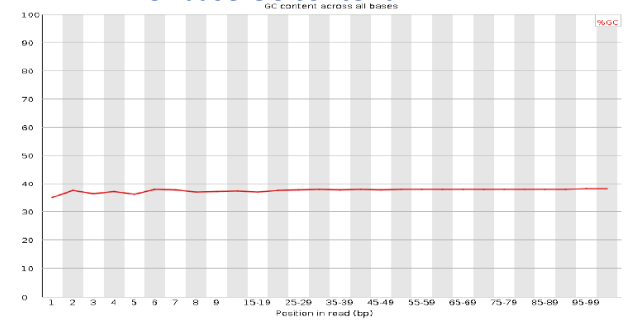
Sequence length distribution



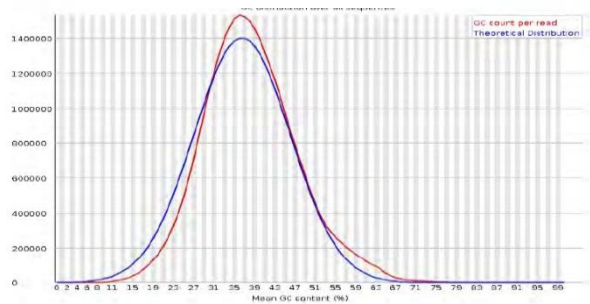
Per base N-content



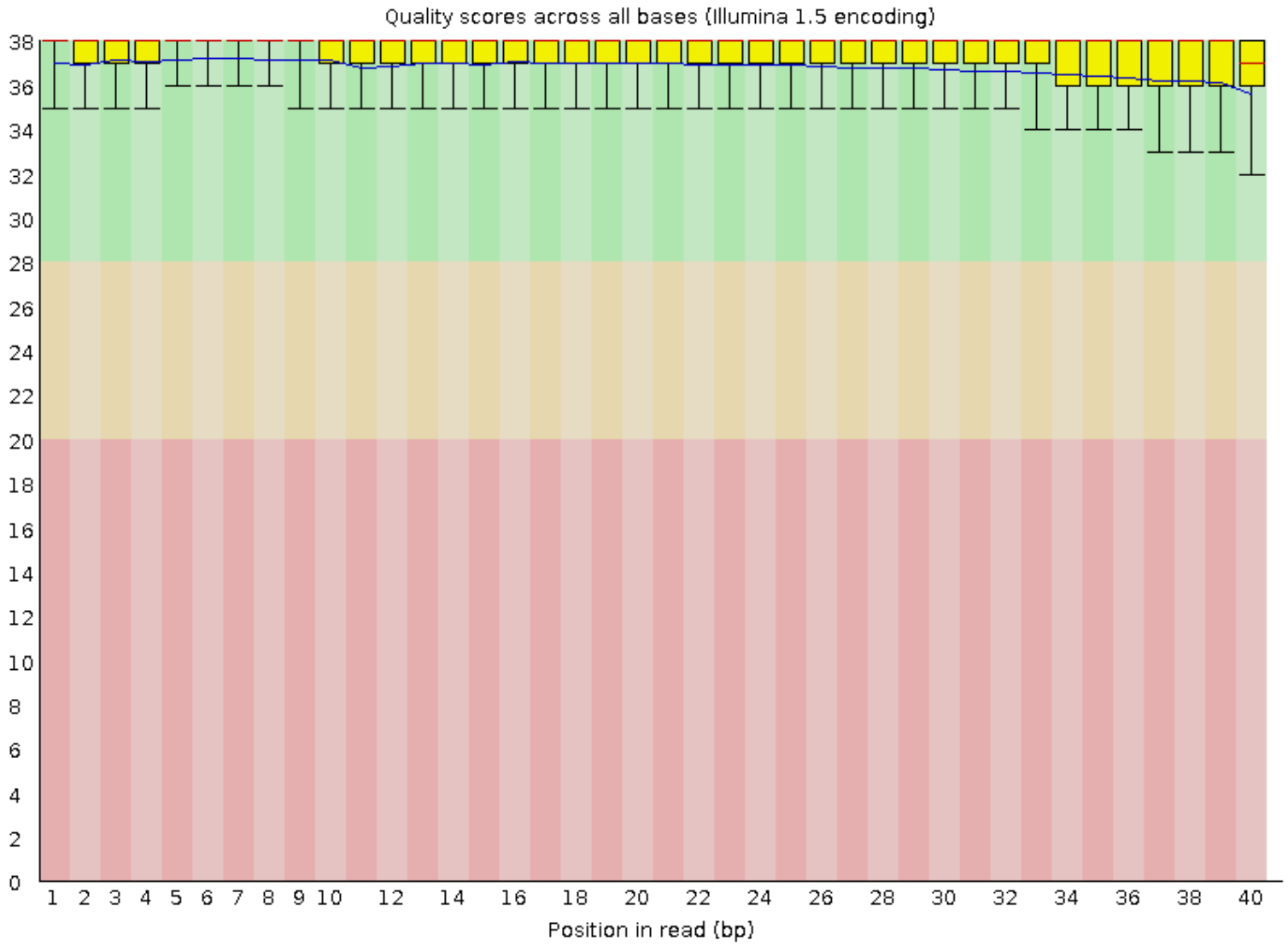
Per base GC content



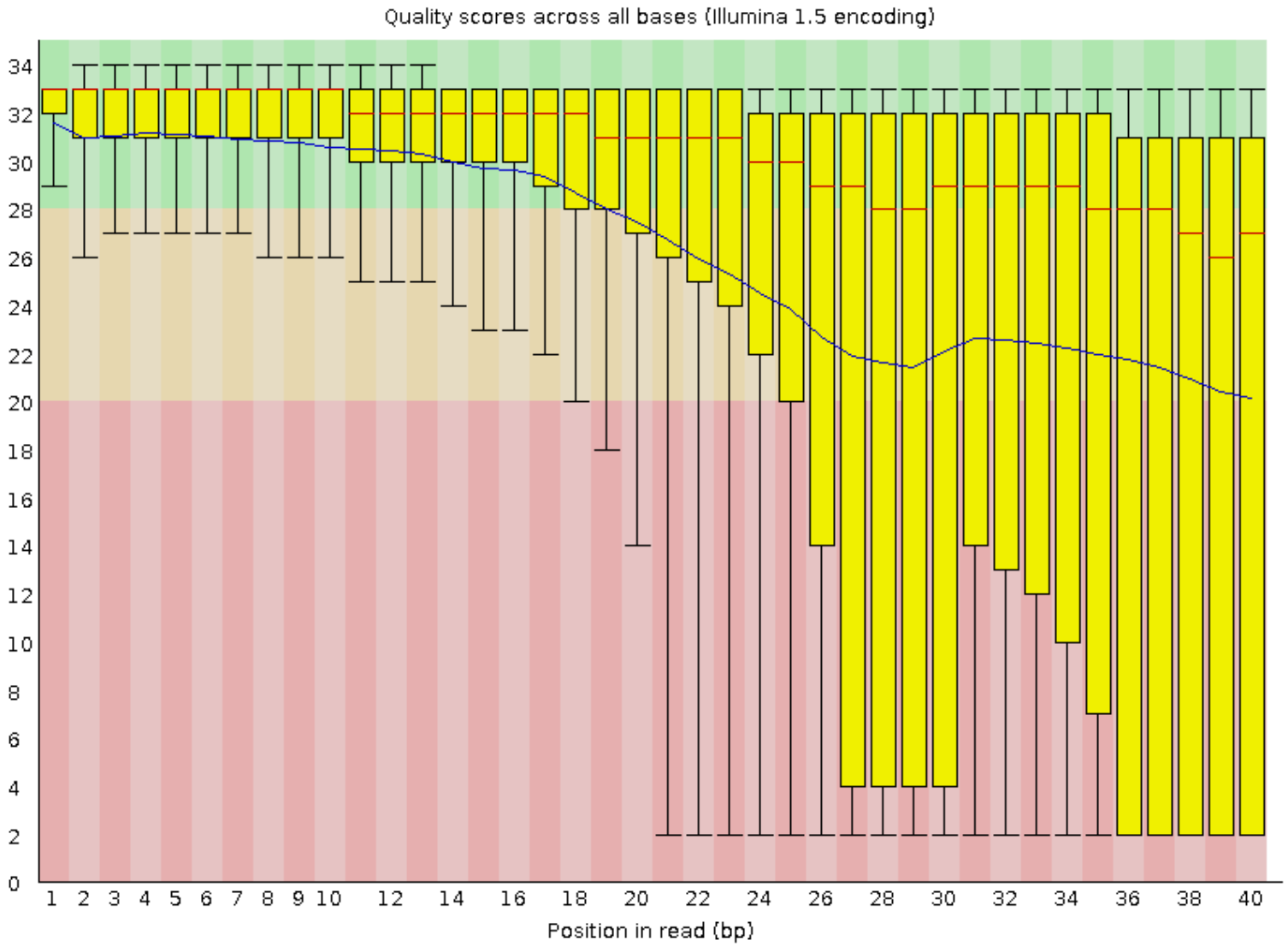
Per sequence GC content



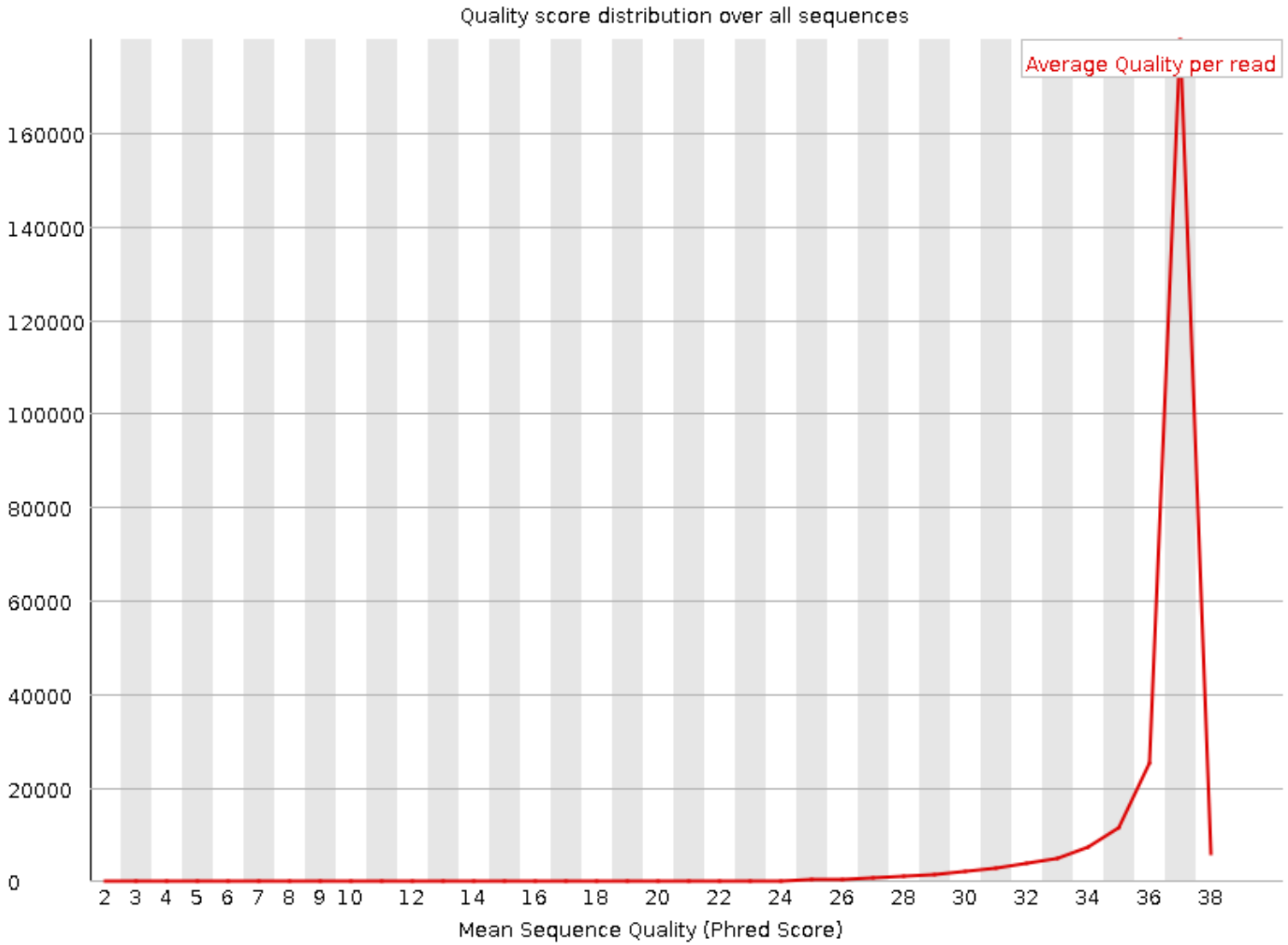
Good


Quality scores across all bases (Illumina 1.5 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

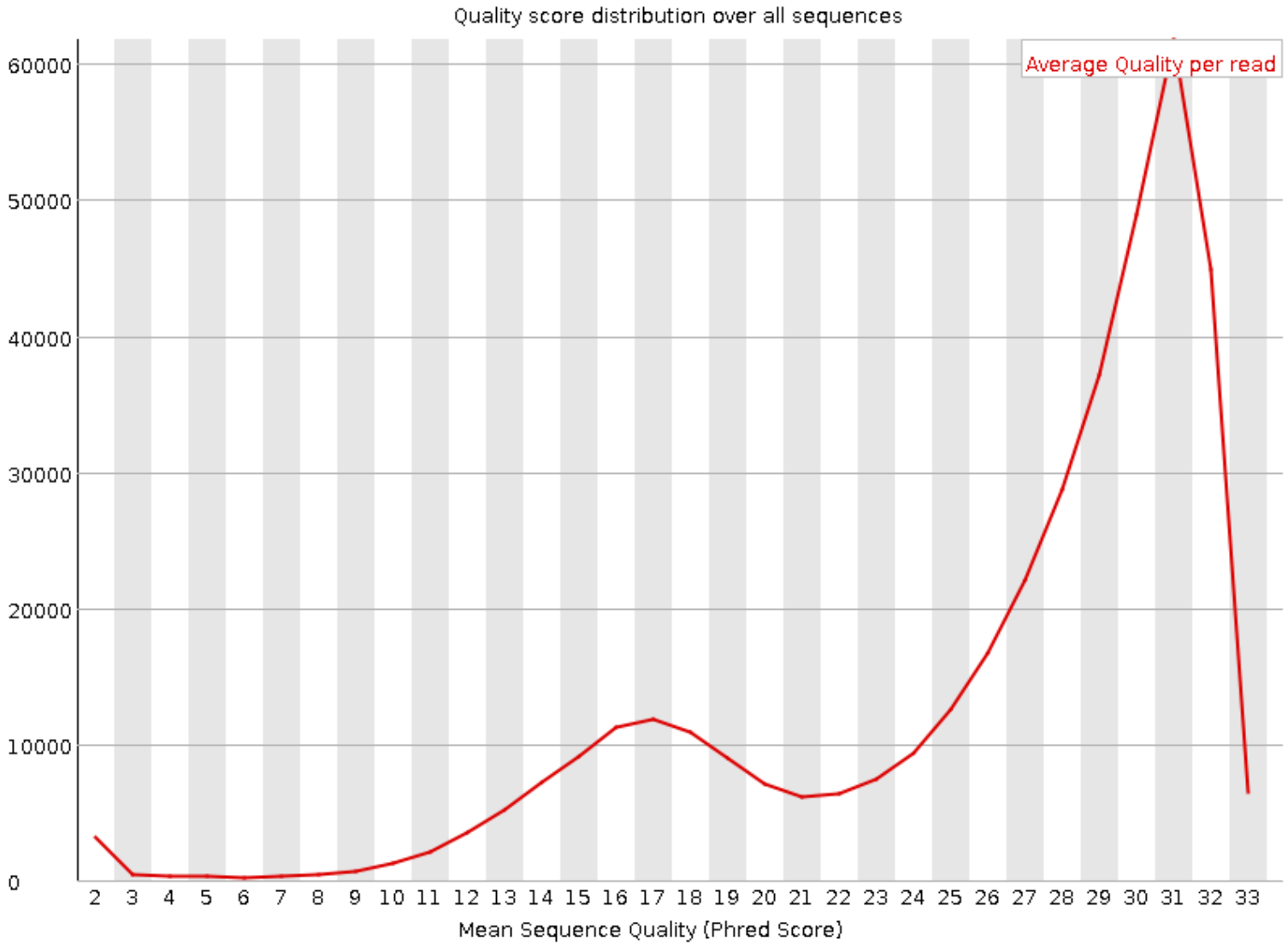
Bad


Quality scores across all bases (Illumina 1.5 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Position in read (bp)

Good


Quality score distribution over all sequences
Average Quality per read
160000
140000
120000
100000
80000
60000
40000
20000
0
2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
Mean Sequence Quality (Phred Score)

Bad


Quality score distribution over all sequences
Average Quality per read
60000
50000
40000
30000
20000
10000
0
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33
Mean Sequence Quality (Phred Score)

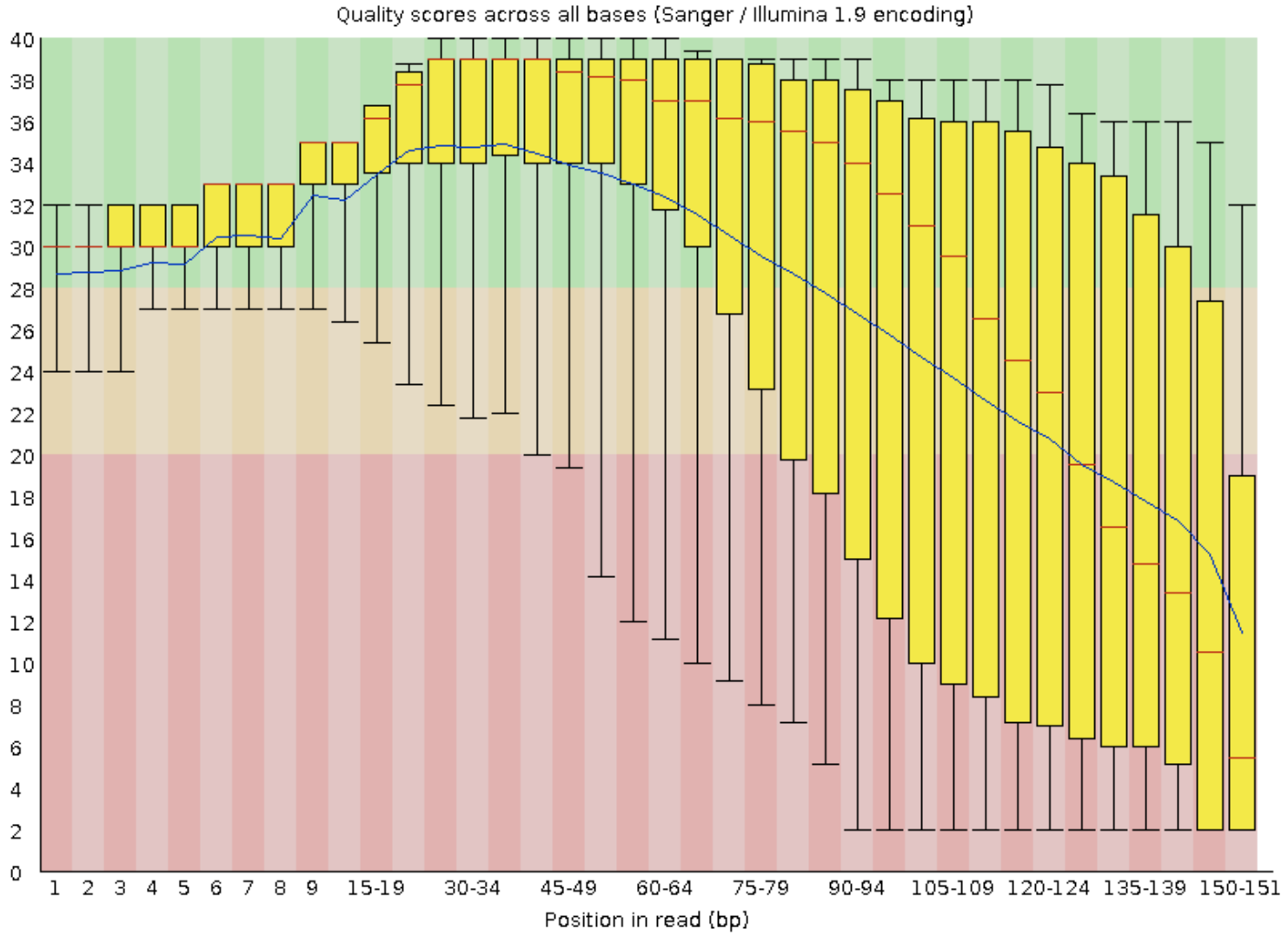
Tools in galaxy

- FASTQ groomer
  - Verify and convert between the known FASTQ variants
- Quality statistics
- Read Trimmer
- Quality filter
- Dealing with paired-end data

Tutorial

- Assess the quality in a good and bad data set
- Clean up the bad data set
- Examine the improvements

bad_MiSeq_r10000.fq (100000 reads)



Strategy

Tool: Filter by quality

Name:	Filter by quality on data 5
Created:	Jun 04, 2013
Filesize:	19.2 MB
Dbkey:	?
Format:	fastqsanger
Galaxy Tool Version:	1.0.0
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	5abf71dc00cff9f5

Input Parameter	Value	Note for rerun
Library to filter	5: FASTQ Groomer on data 2	
Quality cut-off value	20	
Percent of bases in sequence that must have quality equal to / higher than cut-off value	80	

Tool: FASTQ Quality Trimmer

Name:	FASTQ Quality Trimmer on data 7
Created:	Jun 04, 2013
Filesize:	18.9 MB
Dbkey:	?
Format:	fastqsanger
Galaxy Tool Version:	1.0.0
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	af7fa13876122ee5

Input Parameter	Value	Note for rerun
FASTQ File	7: Filter by quality on data 5	
Keep reads with zero length	False	
Trim ends	3' only	
Window size	1	
Step Size	1	
Maximum number of bases to exclude from the window during aggregation	0	
Aggregate action for window	min score	
Trim until aggregate score is	>=	
Quality Score	20.0	

After cleaning (50989 reads)