

NGS : reads quality control

Data used in this tutorials are available on <https://urgi.versailles.inra.fr/download/Tuto/NGS-reads-quality-control>.

Select [genome solexa.fasta](#), [illumina.fastq](#), [solexa.fastq](#) and import them into your current history. With the respectively type of data **fasta** , **fastqIllumina** and **fastqsolexa**.

Download from web or upload from disk

Regular Composite

You added 3 file(s) to the queue. Add more files or click 'Start' to proceed.

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

[https://urgi.versailles.inra.fr/download/tuto/NGS-reads-quality-control/genome_solexa.fasta](#)

New File **86 b**

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

[https://urgi.versailles.inra.fr/download/tuto/NGS-reads-quality-control/illumina.fastq](#)

New File **84 b**

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

[https://urgi.versailles.inra.fr/download/tuto/NGS-reads-quality-control/solexa.fastq](#)

Type (set all): **Genome (set all):**

FastQC	2
Webpage Sections :	
Summary Index	3
Basic Statistics	3
Box plot per base sequence quality	4
Quality per tile	5
Average sequence quality	6
Sequence content across bases	7
GC content (sequence)	8
N content (base)	9
Sequence length	10
Sequence duplication	11
Overrepresented sequences	12
Adapter content	13
Kmer content	14
Fastq Groomer	15
Trimming	17
Filter by quality	18


FastQC

Load the FastQC tool (section : **NGS: QC and manipulation**) -> **FastQC: Read Quality reports**




Choose illumina.fastq file as input and execute.

FastQC run several tests on a maximum subset of 200000 reads (first 200000 reads) of your fastq file. More information on www.bioinformatics.babraham.ac.uk fastqc on line help




Two files in output : rawdata is a txt file and web page that is a html file.

 **FastQC Read Quality reports (Galaxy Tool Version 0.65)**

Short read data from your current history




   2: illumina.fastq

Contaminant list

   Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Prime
CAAGCAGAAGACGGCATAACGA













Submodule and Limit specifying file

   Nothing selected


a file that specifies which submodules are to be executed (default=all) and also specifies the threshold submodules warning parameter

Execute

Webpage Sections : Summary Index

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

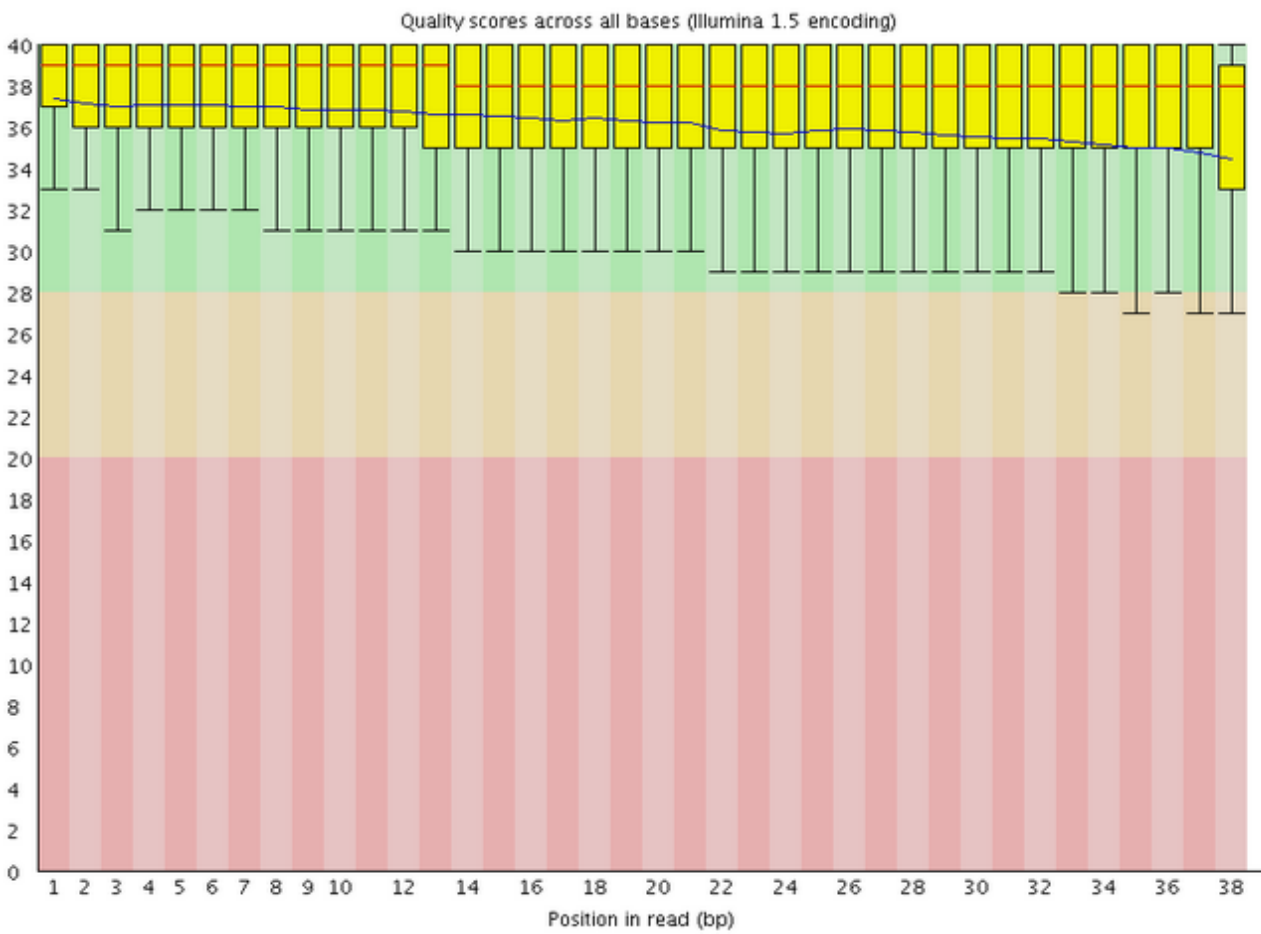
Basic Statistics

 Basic Statistics	
Measure	Value
Filename	illumina.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	1138429
Sequences flagged as poor quality	0
Sequence length	38
%GC	43

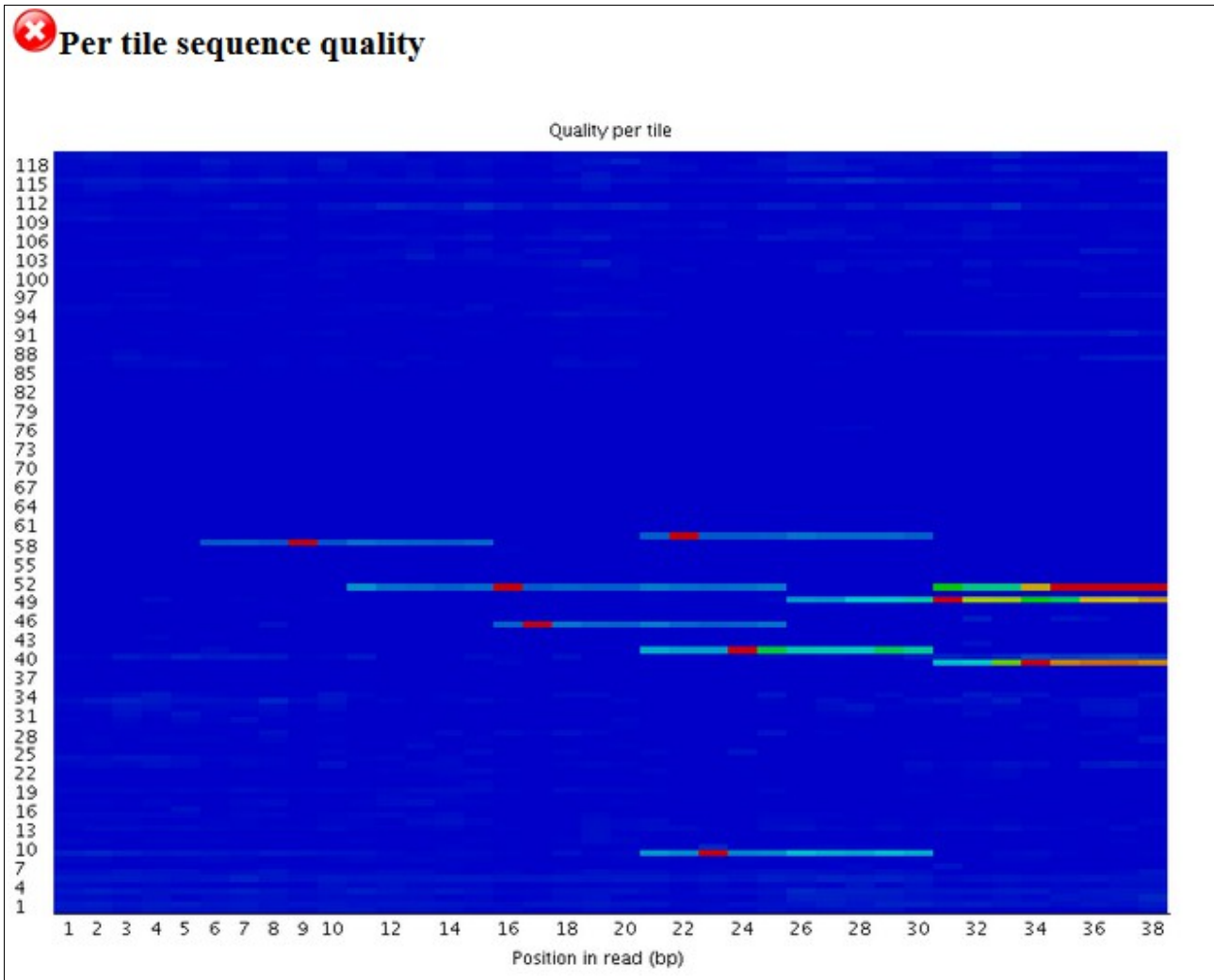
Box plot per base sequence quality



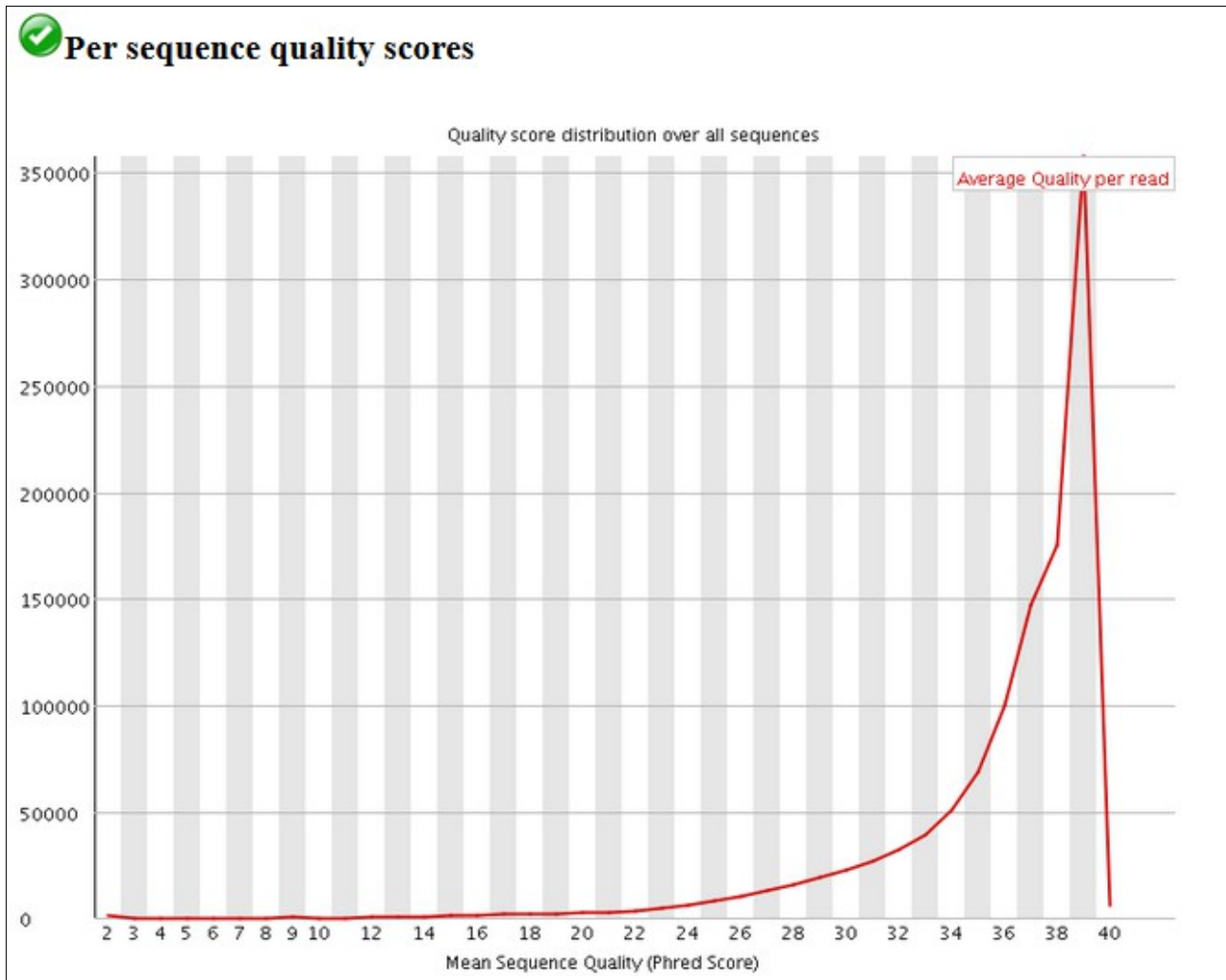
Per base sequence quality



Quality per tile

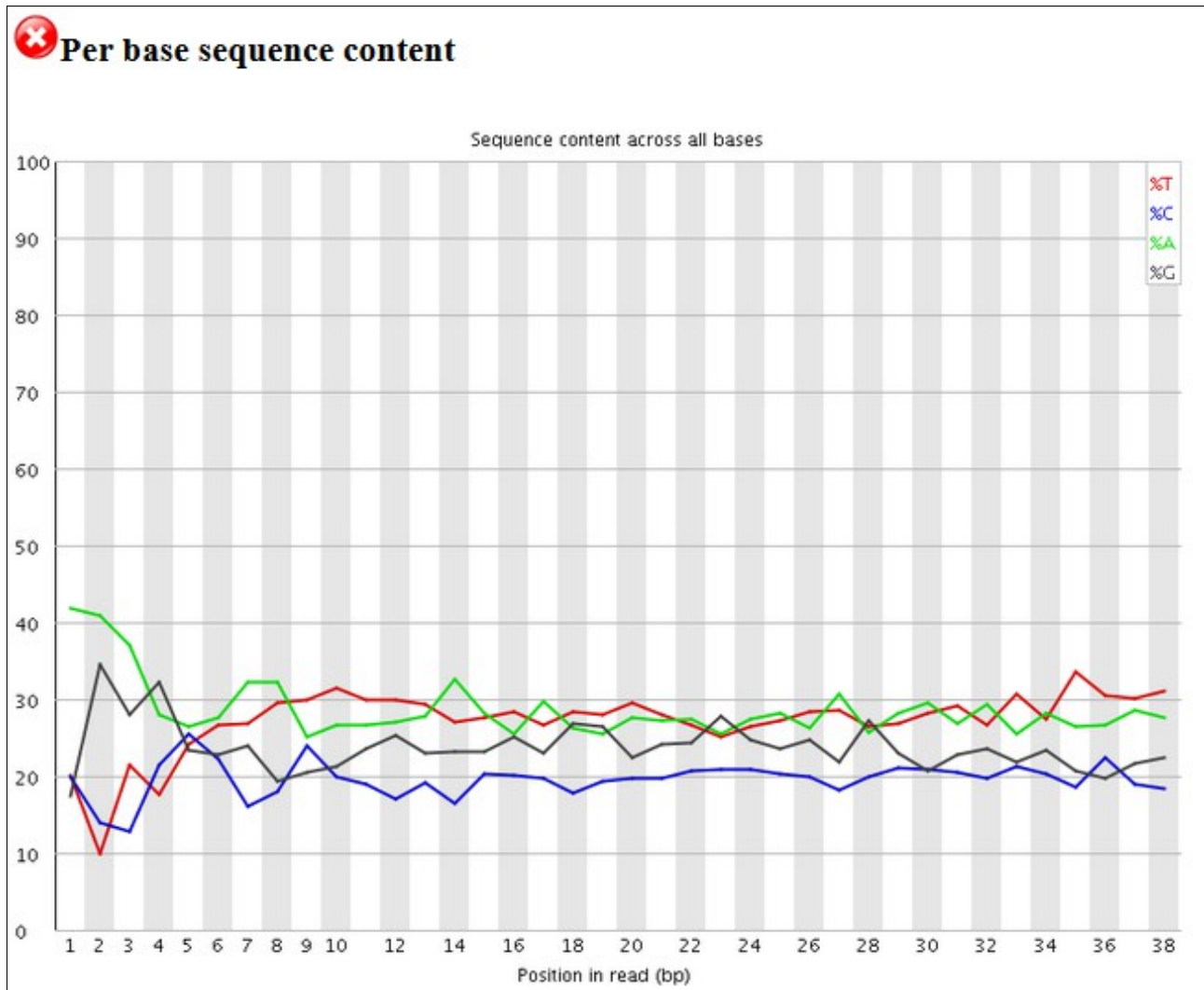


Average sequence quality



Sequence content across bases

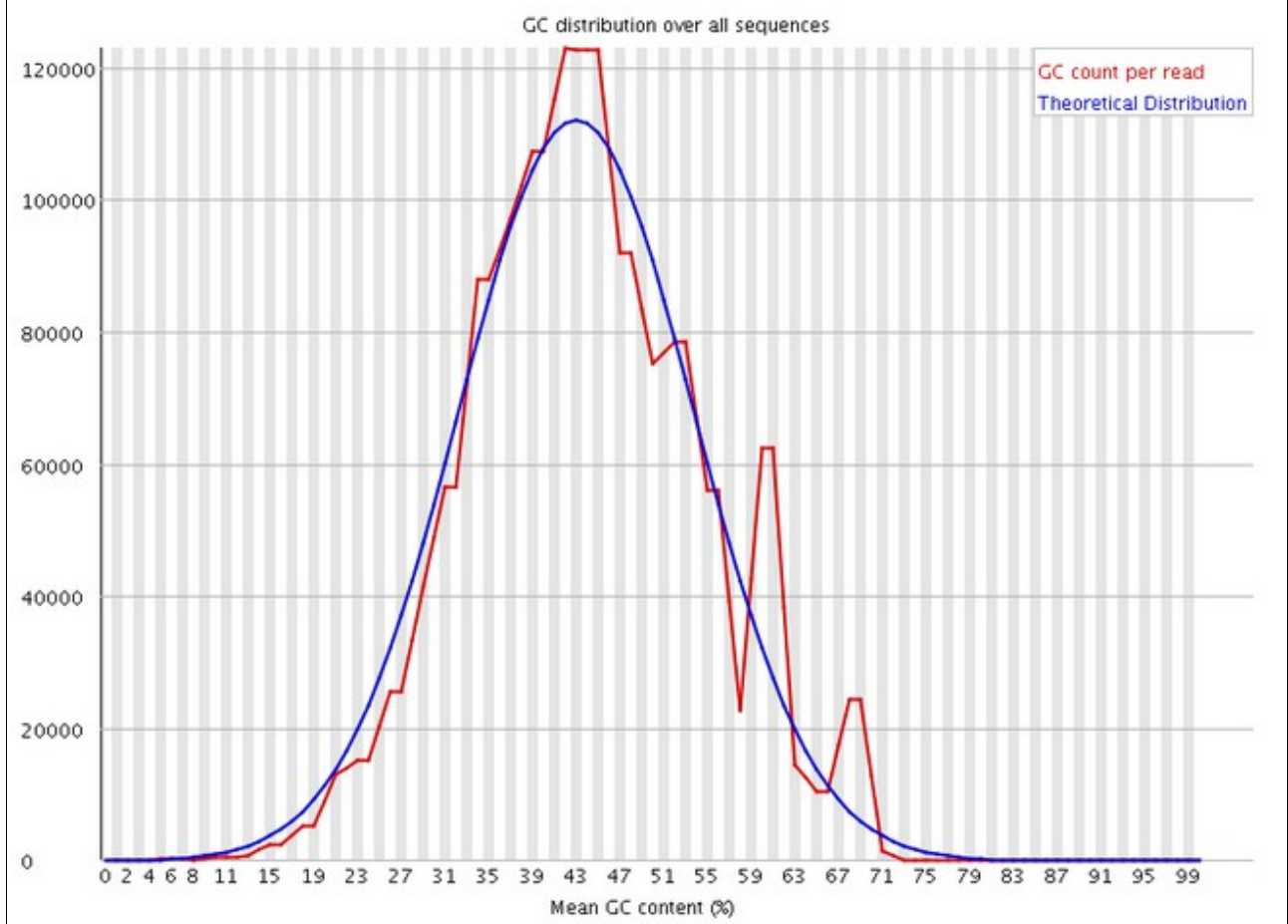
The bias in nucleotide composition at the start of illumina reads is explained by random primers (ref: Biases in Illumina transcriptome sequencing caused by random hexamer priming, Hansen et al, 2010, NAR)



GC content (sequence)



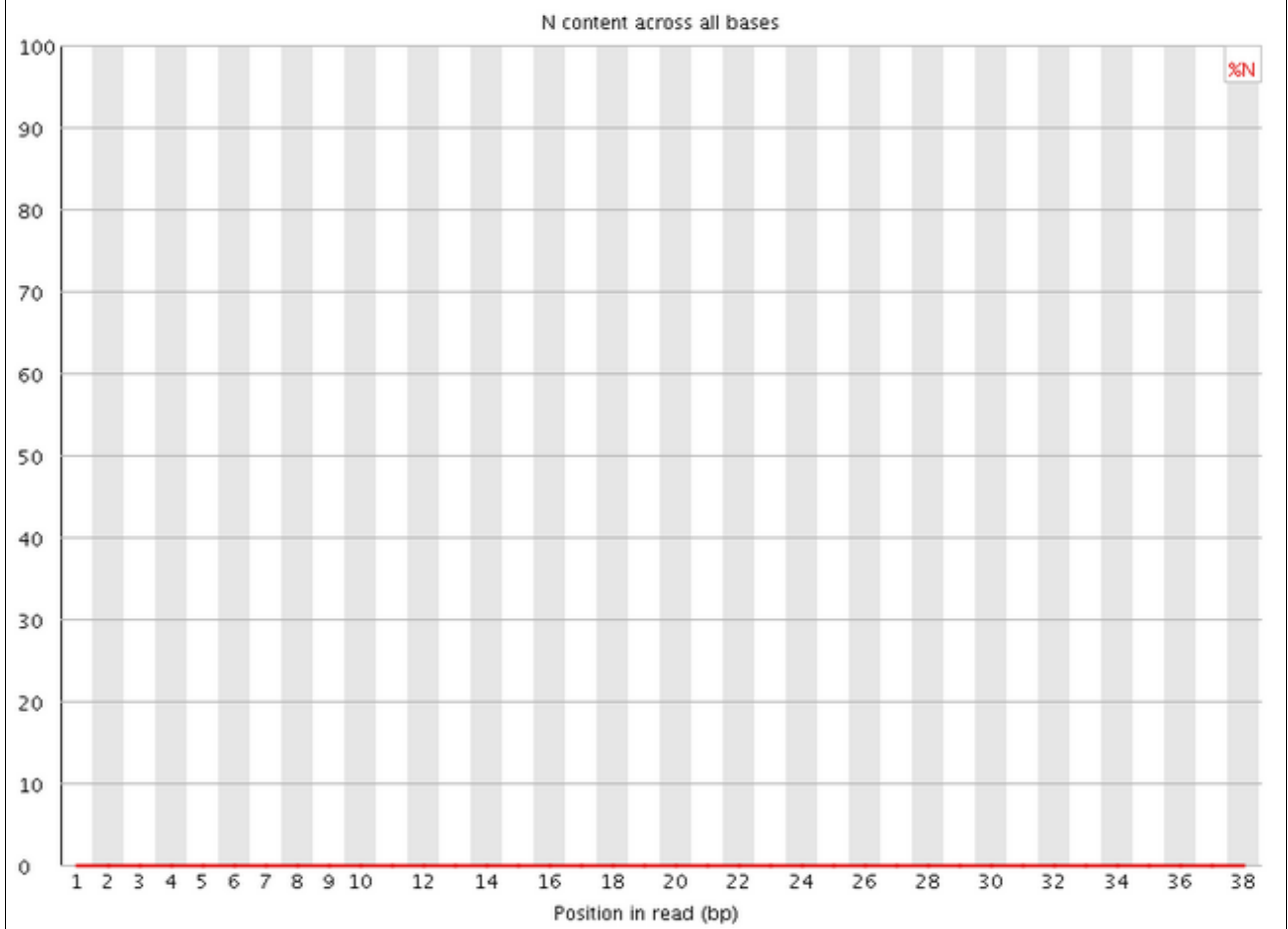
Per sequence GC content



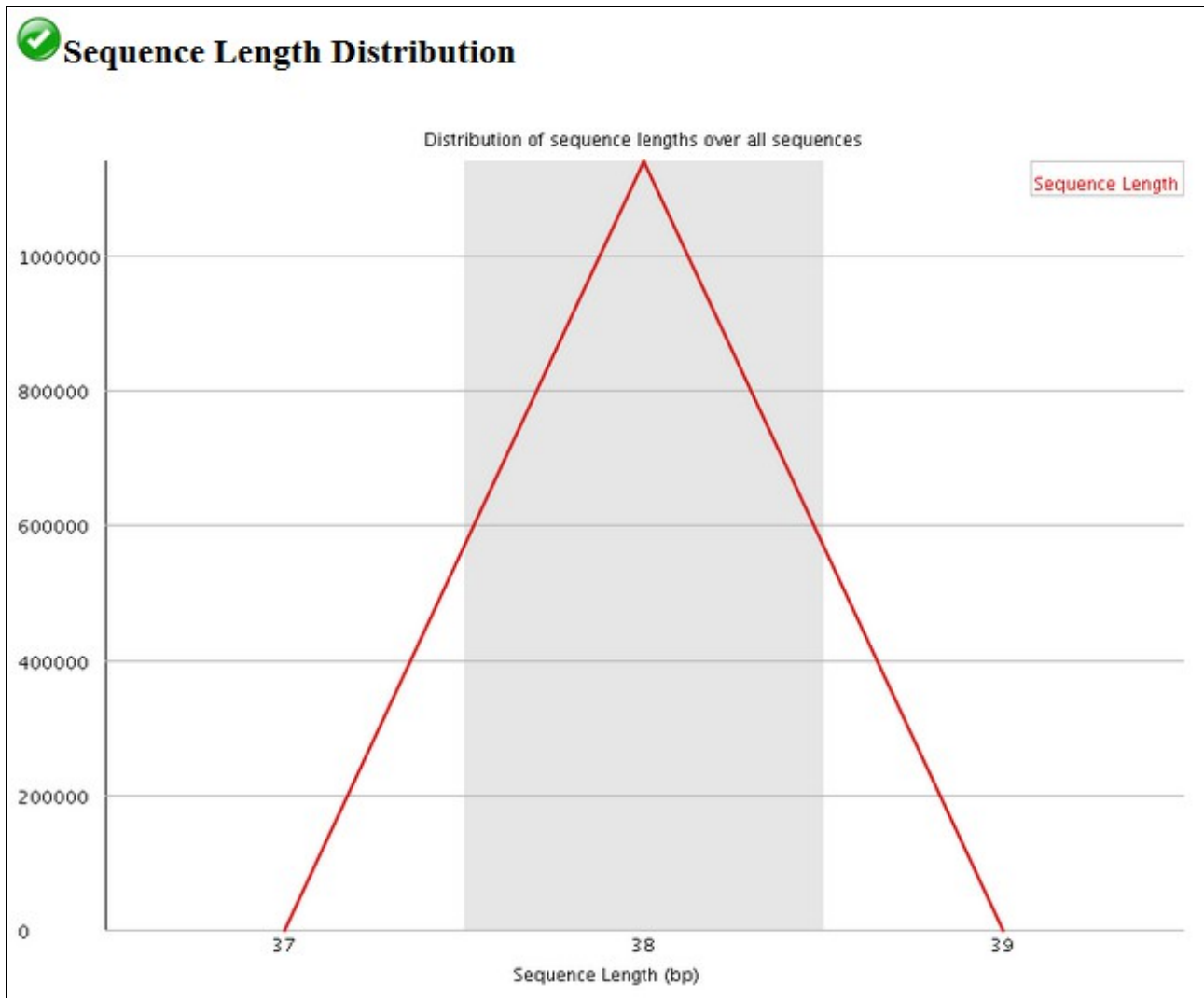
N content (base)



Per base N content



Sequence length

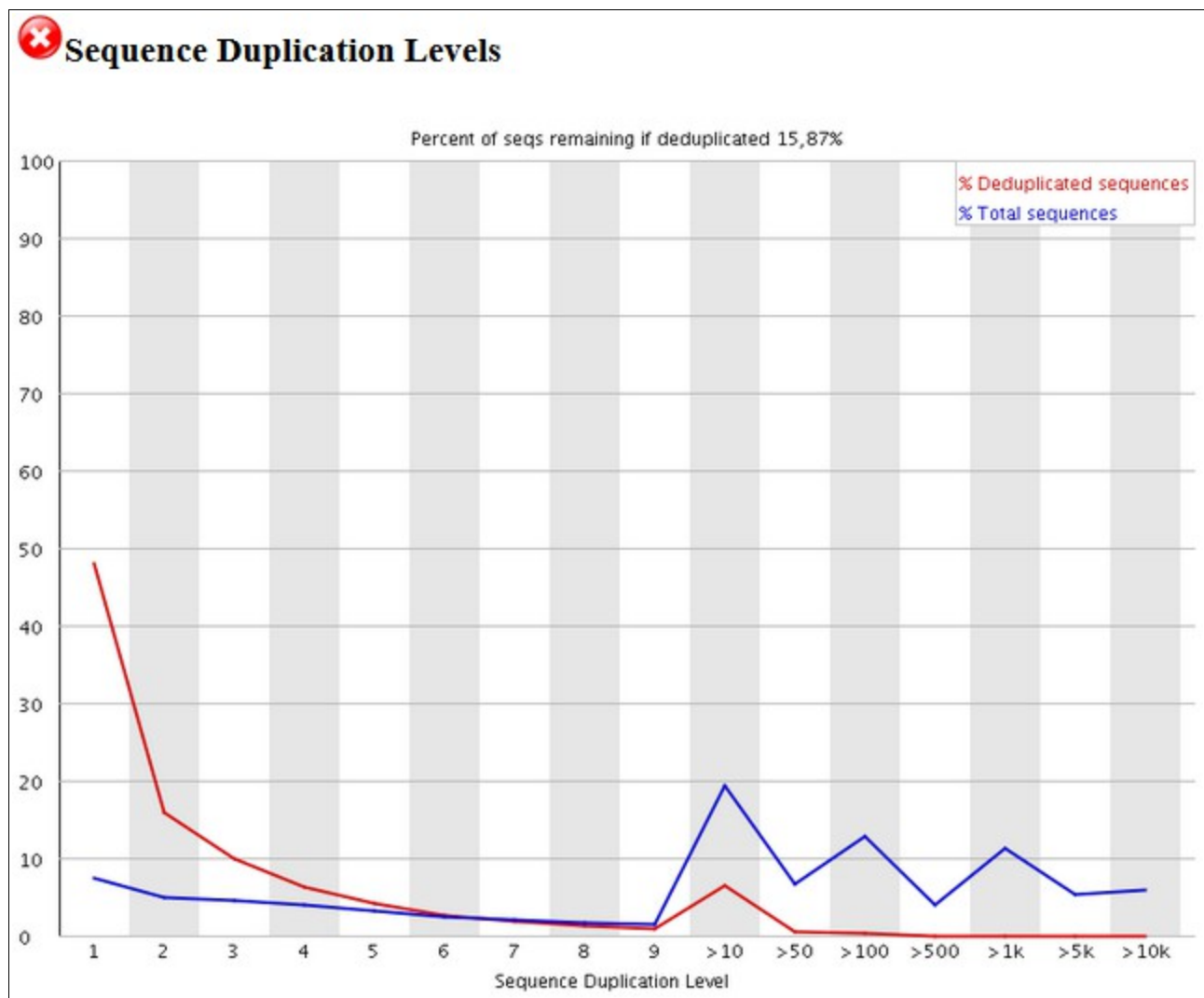


Sequence duplication


High sequence level duplication could be explained by several factors. In our case, the fastq file contains few sequences extracted from a RNA-Seq analysis. The duplicated sequences are the multiple counts for each gene.

(info: only the 50 first bases are taken into account. The percentage of duplicated sequences is on a relative scale, with the number of sequences occurring exactly once.)

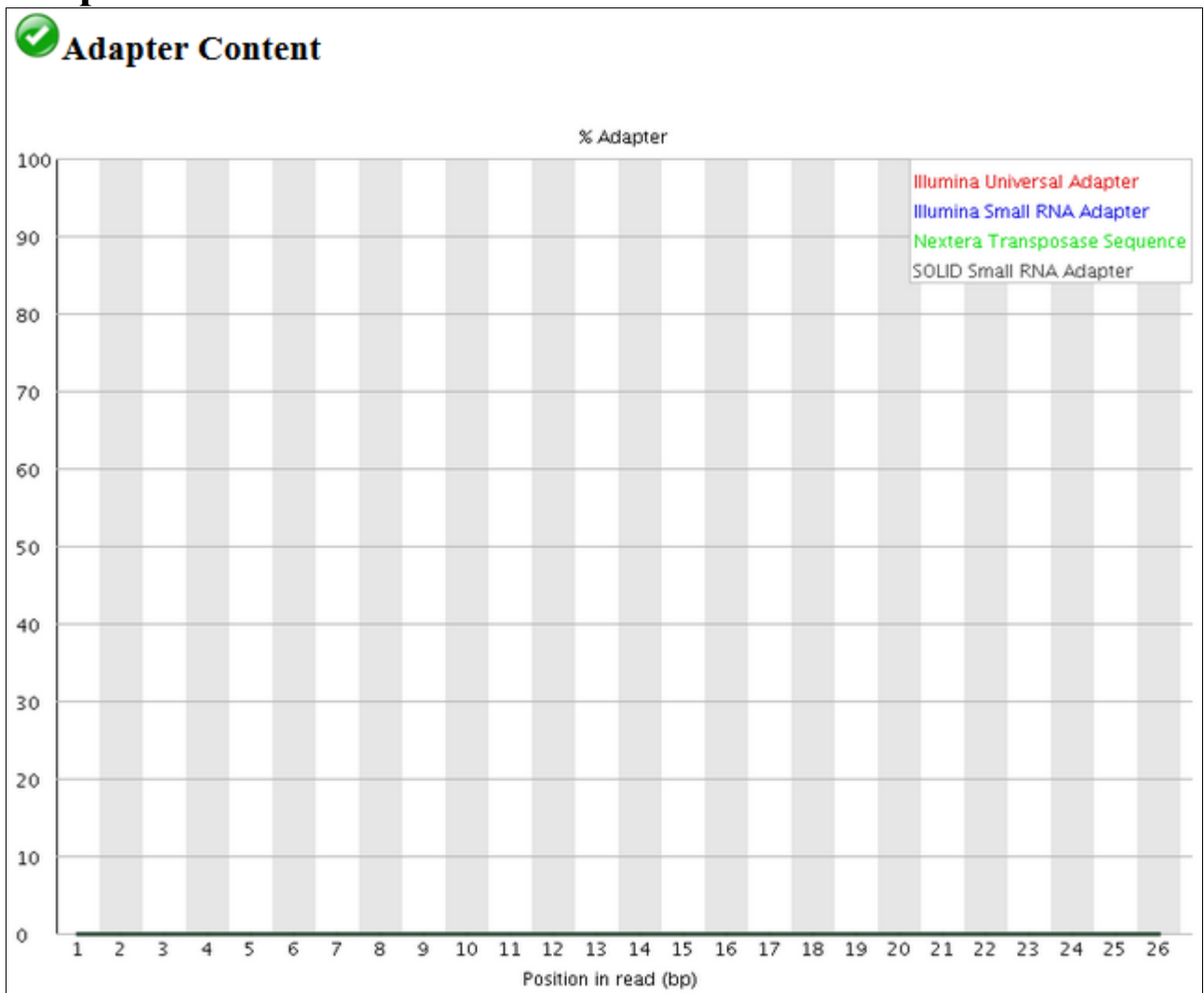
Blog post to read : [here](#)



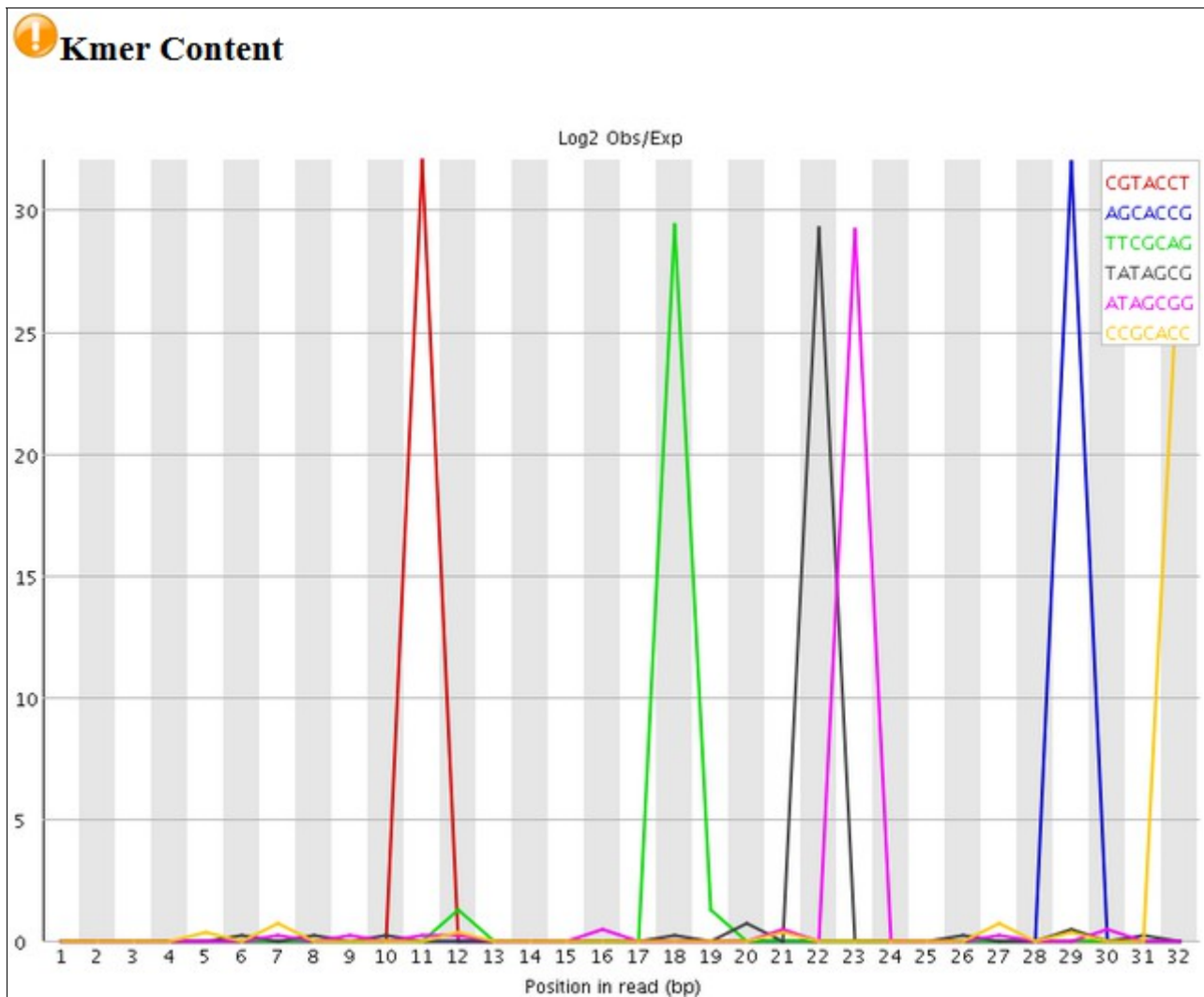
Overrepresented sequences

 Overrepresented sequences			
Sequence	Count	Percentage	Possible Source
TGACCCGTGATGGAGGCGGCCGGGATAGCACATATCAG	21255	1.8670466054536559	No Hit
CGGCTATAATAAATCGATCTTTGCGGGCAGCCC GTTG	12397	1.0889567992382485	No Hit
AAGGTGACCCGTGATGGAGGCGGCCGGGATAGCACATA	12187	1.0705103260721573	No Hit
CGGGCAGCCC GTTGCCAGGAGGCGTGAGGAATCCGTCT	11906	1.0458271881689591	No Hit
TACTATGATGAATGACATTAGCGTGAACAATCTCTGAT	10282	0.9031744623511875	No Hit
ACGTTTGACAACACTCTGGGACCAAGTTTCGATGGTTA	9689	0.8510851357440823	No Hit
AAAGGTGACCCGTGATGGAGGCGGCCGGGATAGCACAT	9327	0.8192869296196775	No Hit
GGGCAGCCC GTTGCCAGGAGGCGTGAGGAATCCGTCTC	8773	0.7706233766005609	No Hit
TGGAGGCGGCCGGGATAGCACATATCAGTCGGATAATT	8714	0.7654407960443734	No Hit
AGGATTTACTCGCACATTGTGGCCGTTCCCTCGGGGAT	7273	0.638862853985624	No Hit
GAGGCGGCCGGGATAGCACATATCAGTCGGATAATTGT	6750	0.5929223517672161	No Hit
CGTGGCCTTTTTACCACCTTTATAGCGGTGCTTTAAC	5983	0.5255488045367783	No Hit
GTTATAATGATGATAACCTTCTCAGCTCACTCAGATCTT	5536	0.4862841687975271	No Hit
AATCCGTCTCTCTGTCTGGTGCGGCAAGGTAGTTCTGG	4503	0.3955450888900406	No Hit
AGATAGATTGAACGTTGCTGGGCGCCTGGTGTTGATCA	4379	0.3846528856872058	No Hit
AGCTGAAAAGTTCTGGGTTAACCCAGATTGTGGTTTGA	3802	0.33396900465466006	No Hit
TTTGCGGGCAGCCC GTTGCCAGGAGGCGTGAGGAATCC	3717	0.3265025750398136	No Hit
GGAGGCGGGAGAGTCCGTTCTGAAGTGTCCCGGCTATA	3716	0.3264147346914037	No Hit
CAAACAAGATTAATTTAGGCGATTACTCACTAAGATAT	3652	0.32079295239316635	No Hit

Adapter content



Kmer content



Sequence	Count	PValue	Obs/Exp	Max Obs/Exp	Max Obs/Exp Position
CGTACCT	30	2.2343991E-5	32.024796	11	
AGCACCG	20	0.0037578726	31.992388	29	
TTCGCAG	120	0.0	29.384535	18	
TATAGCG	635	0.0	29.279364	22	
ATAGCGG	635	0.0	29.244553	23	
CCGCACC	425	0.0	29.00048	32	
GTGGCCT	660	0.0	28.790955	2	
AGCGGTG	660	0.0	28.556538	25	
CGCCTGG	555	0.0	28.556366	23	
ACGTTTG	1125	0.0	28.535923	1	
TAGGCGA	430	0.0	28.337175	16	
AACACTC	1140	0.0	28.335407	10	
ACCACCT	670	0.0	27.96441	14	
TACTCAC	435	0.0	27.942457	24	
TCACTAA	430	0.0	27.854992	27	
AGGCGAT	435	0.0	27.629488	17	
GGCCTTT	695	0.0	27.36749	4	
CACCACC	685	0.0	27.352053	13	
CTGGGAC	1175	0.0	27.29	16	
CTAAGAT	440	0.0	27.28665	30	

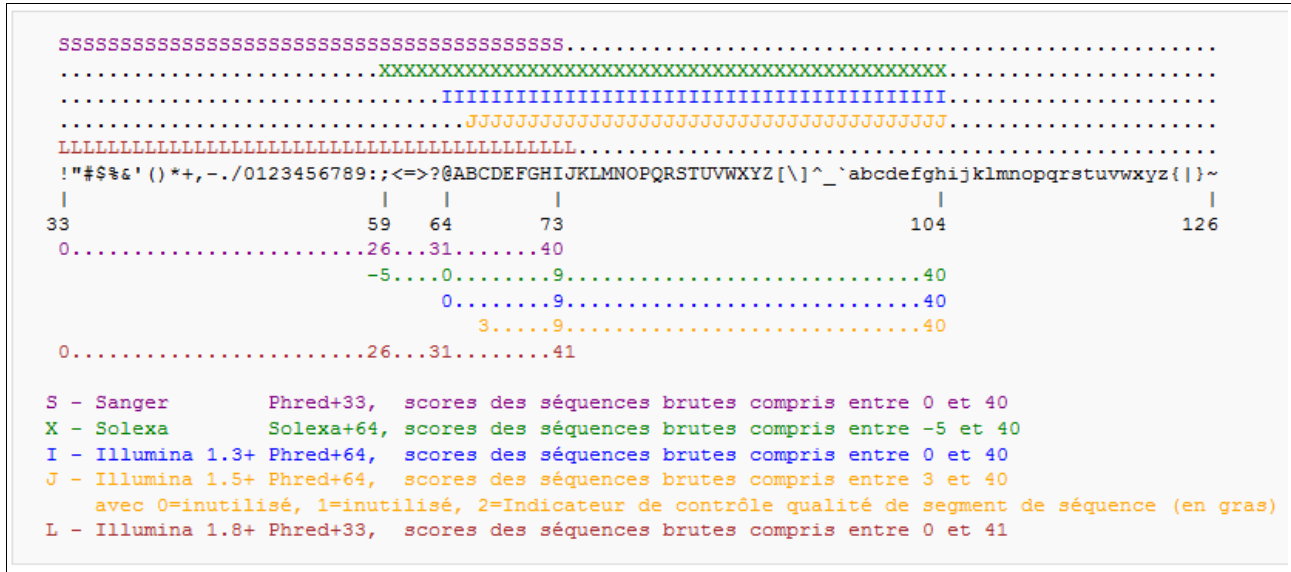
[Back](#)

ads-quality-control

Fastq Groomer

The aim of the fastq Groomer is to homogenize the quality of the different fastq formats. The usual quality is in Sanger encoding (phred+33) with expected characters from ! to J. The role of the fastq Groomer is to perform this transform. Moreover, it sets the metadata associated with the fastq format. So, the format is not just a fastq, it becomes a fastqsanger, fastqillumina or fastqsolexa. If you know that your data are encoded in phred+33, Sanger, you can avoid this step and edit the metadata associated with your sequences. To do that, clic on the pen in data box and edit attributes. Select "fastqsanger" as datatype instead of "fastq".

Figure: Explanation of data quality encoding and the relation between quality value and corresponding ASCII table elements.



Explore the uploaded file illumina.fastq content (clic on the eye).

```

! This dataset is large and only the first megabyte is shown below.
Show all | Save

@HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
ACGATATTTTGTCCGTGCTAGACTCCNACTTAATTCCA
+HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
gggggfggccccfffcffcfbcFcccdcdggg
@HWUSI-EAS1656_0009_FC:1:1:1150:15676#0
ATGGCTGATATTACTGATAAGACAGCTGAACAATTGGA
+HWUSI-EAS1656_0009_FC:1:1:1150:15676#0
cee_gffc[ff_fcf[_ffc[]ccc[dfcWfcW_ff
@HWUSI-EAS1656_0009_FC:1:1:1150:20431#0
CGGGCAGCCCGTTGGCAGGAGCGTGAGGAATCCGTCT
+HWUSI-EAS1656_0009_FC:1:1:1150:20431#0
cYfffffdff]ff_ae`Wacc2^Xa^J]Q``^acc
@HWUSI-EAS1656_0009_FC:1:1:1152:17358#0
TTGGCTACTTACTTCGGTACCCTTGTCCCTAACTTAGA
+HWUSI-EAS1656_0009_FC:1:1:1152:17358#0
fgggagggggaggg^ccReeeddea`eeddcfcf_fe_
@HWUSI-EAS1656_0009_FC:1:1:1155:17012#0

```

2: illumina.fastq [eye] [edit] [close]

174.1 MB
format: **fastqillumina**,
database: ?

Uploaded file from
<https://urgi.versailles.inra.fr/download/tuto/NGS-reads-quality-control/illumina.fastq>
uploaded fastqillumina file

[save] [info] [share]

```

@HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
ACGATATTTTGTCCGTGCTAGACTCCNACTTAATTCCA
+HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
gggggfggccccfffcffcfbcFcccdcdggg
@HWUSI-EAS1656_0009_FC:1:1:1150:15676#0
ATGGCTGATATTACTGATAAGACAGCTGAACAATTGGA



```

The quality line contains g,f,c,d characters. The quality is encoding in illumina 1.3-1.7.

Select the FASTQ Groomer in (NGS: QC and manipulation -> ILLUMINA FASTQ).
 Select the file, the incoming encoding is illumina 1.3-1.7. Let default option, the outgoing encoding is Sanger.

FASTQ Groomer convert between various FASTQ quality formats (Galaxy Tool Version 1.0.4) Options

File to groom

  2: illumina.fastq


Input FASTQ quality scores type

illumina 1.3-1.7




Advanced Options

Hide Advanced Options

The output is now in "fastqsanger" format.

 This dataset is large and only the first megabyte is shown below.
[Show all](#) | [Save](#)

```
@HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
ACGATATTTTGTCCGTGCTAGACTCCNACTTAATCCCA
+HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
gggggfgggcccccfffcfffcdfefebcFccccddcdggg
@HWUSI-EAS1656_0009_FC:1:1:1150:15676#0
ATGGCTGATATTACTGATAAGACAGCTGAACAATTGGA
+HWUSI-EAS1656_0009_FC:1:1:1150:15676#0
cce_gffc(ff_fcf[_ffc[]ccc[dfffcWfcW_ff
@HWUSI-EAS1656_0009_FC:1:1:1150:20431#0
CGGGCAGCCCGTTGGCAGGAGGCGTGAGGAATCCGTCT
+HWUSI-EAS1656_0009_FC:1:1:1150:20431#0
cYfffffdaff]ff_ae`WacccZ^Xa^J]Q``^acc
@HWUSI-EAS1656_0009_FC:1:1:1152:17358#0
TTGGCTACTTACTTCGGTACCGTTGTCCTAACTTAGA
+HWUSI-EAS1656_0009_FC:1:1:1152:17358#0
fgggagggggagg^ccReeeddea`eeddcfcf_fe_
@HWUSI-EAS1656_0009_FC:1:1:1155:17012#0
AAGAGCAAACGGTCTTTGTGATAGCTCAACGTCATTCCG
+HWUSI-EAS1656_0009_FC:1:1:1155:17012#0
c]ff_b^_ffeffcffgg_gcf^ffcf^ad^ac\\caa
@HWUSI-EAS1656_0009_FC:1:1:1155:20602#0
+HWUSI-EAS1656_0009_FC:1:1:1155:20602#0
```

6: FASTQ Groomer on data   






2 [View data](#)

174.1 MB

format: **fastqsanger**,
 database: ?

Groomed 1138429 sanger reads into sanger reads. Based upon quality and sequence, the input data is valid for: solexa, sanger, illumina

Input ASCII range: 'B'(66) - 'h'(104)
 Input decimal range: 33 - 71

```
@HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
ACGATATTTTGTCCGTGCTAGACTCCNACTTAATCCCA
+HWUSI-EAS1656_0009_FC:1:1:1145:8238#0
gggggfgggcccccfffcfffcdfefebcFccccddcdggg
@HWUSI-EAS1656_0009_FC:1:1:1150:15676#0
ATGGCTGATATTACTGATAAGACAGCTGAACAATTGGA
```


Trimming

Trimmomatic is a tool dedicated to reads cleaning in Single or Paired End mode. It performs adaptor removal, trimming by sliding window or cut off, minimum length, ...

Select **Trimmomatic** in (**NGS: QC and manipulation**). We will remove adaptors (used with GAI sequencer library type) and trim sequence with a sliding window of 4 bases and an average quality of 20.

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Tool Version 0.36.0)

Paired end data?

Input FASTQ file
 6: FASTQ Groomer on data 2

Perform initial ILLUMINACLIP step?

Cut adapter and other illumina-specific sequences from the read

Adapter sequences to use
TruSeq2 (single-ended, for Illumina GAI)

Maximum mismatch count which will still allow a full match to be performed
2

How accurate the match between the two 'adapter ligated' reads must be for PE palindrome read alignment
30

How accurate the match between any adapter etc. sequence must be against a read
10

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform
Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across
4

Average quality required
20

+ Insert Trimmomatic Operation

Important: There is no common way to clean data. You have to analyze your read contents to define adequate filtering process.

Instead of Trimmomatic, you can use simple task tools of the FastX Toolkit or Generic Fastq Manipulation.

FASTQ Quality Trimmer by sliding window

Select **FASTQ Quality Trimmer** in (**NGS: QC and manipulation** -> **GENERIC FASTQ MANIPULATION**). We will trim read end until the quality of the current base is higher than 20.

FASTQ Quality Trimmer by sliding window (Galaxy Tool Version 1.0.1)

FASTQ File
 6: FASTQ Groomer on data 2

Keep reads with zero length

Trim ends

Window size

Step Size

Maximum number of bases to exclude from the window during aggregation

Aggregate action for window

Trim until aggregate score is

Quality Score

Filter by quality

Filter by quality (Galaxy Tool Version 1.0.0)

Library to filter
 6: FASTQ Groomer on data 2

Quality cut-off value

Percent of bases in sequence that must have quality equal to / higher than cut-off value