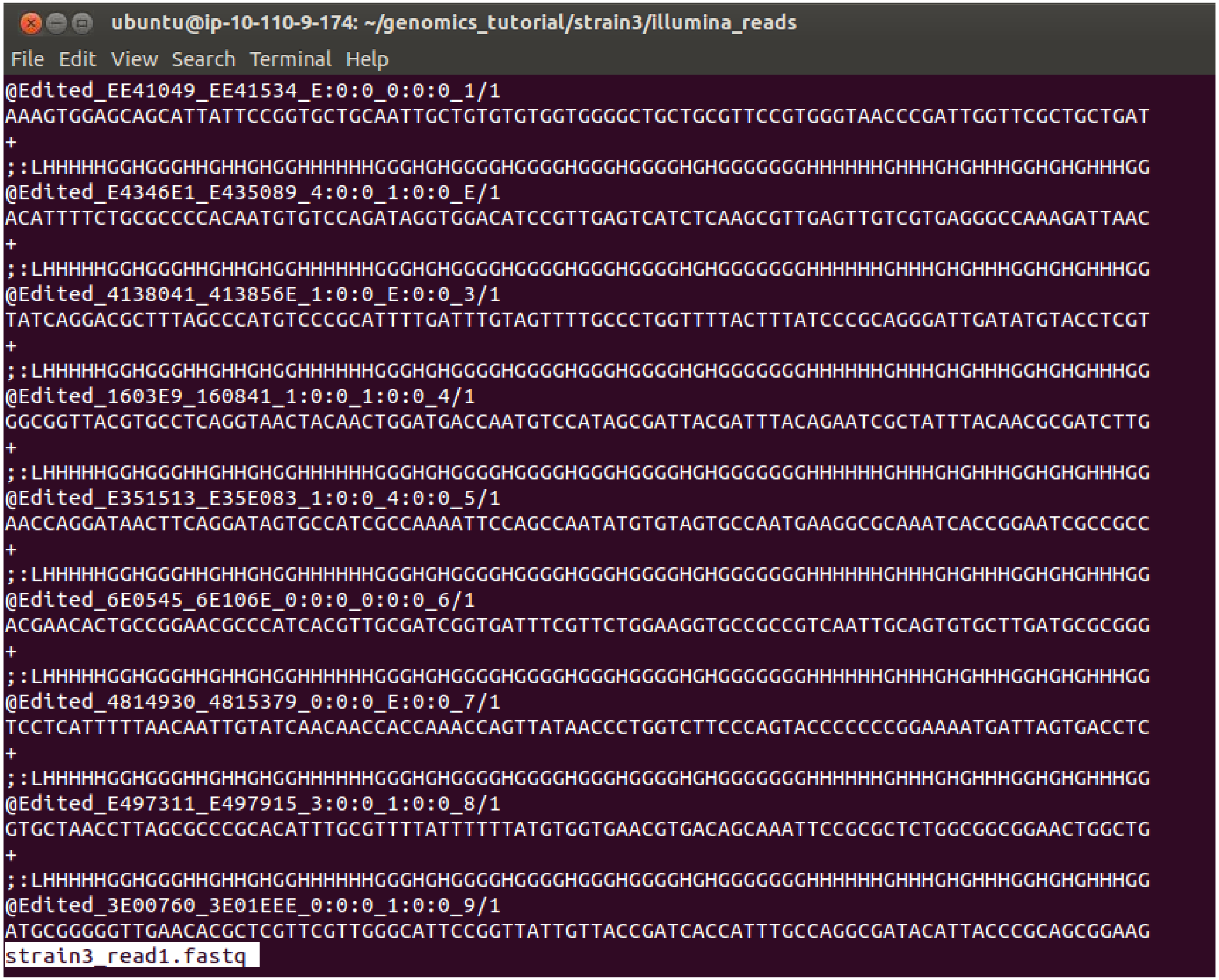
Quality assessment and control of sequence data

Naiara Rodríguez-Ezpeleta

Workshop on Genomics 2015
Cesky Krumlov

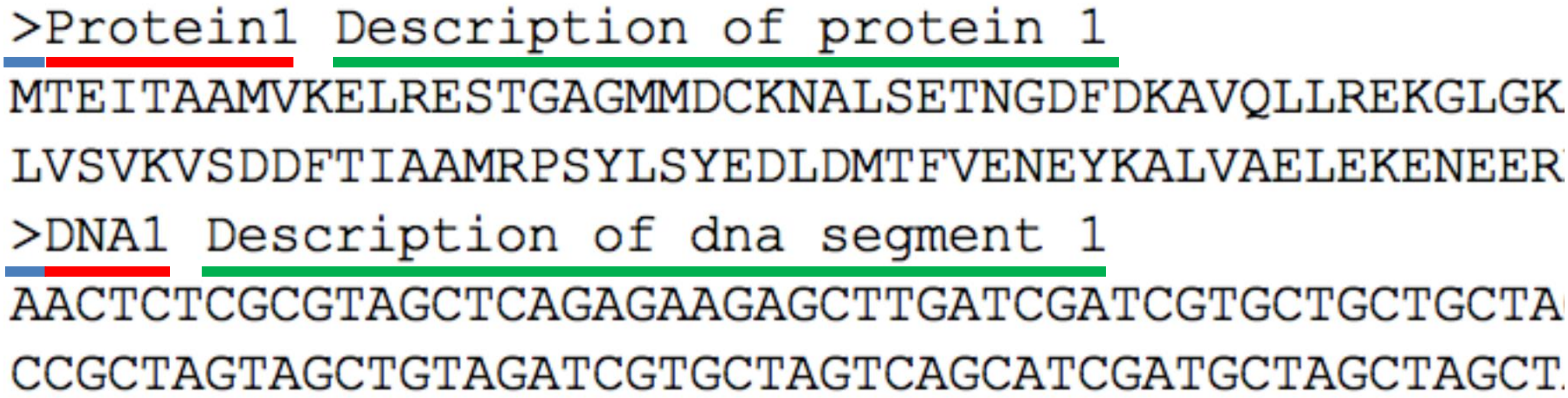
fastq format


ubuntu@ip-10-110-9-174: ~/genomics_tutorial/strain3/illumina_reads
File Edit View Search Terminal Help

```
@Edited_EE41049_EE41534_E:0:0_0:0:0_1/1
AAAGTGGAGCAGCATTATTCCGGTGCTGCAATTGCTGTGTGTGGTGGGGCTGCTGCGTTCCGTGGGTAACCCGATTGGTTCGCTGCTGAT
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_E4346E1_E435089_4:0:0_1:0:0_E/1
ACATTTTCTGCGCCCCACAATGTGTCCAGATAGGTGGACATCCGTTGAGTCATCTCAAGCGTTGAGTTGTCGTGAGGGCCAAAGATTAAC
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_4138041_413856E_1:0:0_E:0:0_3/1
TATCAGGACGCTTTAGCCCATGTCCCGCATTTTGATTTGTAGTTTTGCCCTGGTTTTACTTTATCCCGCAGGGATTGATATGTACCTCGT
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_1603E9_160841_1:0:0_1:0:0_4/1
GGCGGTTACGTGCCTCAGGTAACTACAACTGGATGACCAATGTCCATAGCGATTACGATTTACAGAATCGCTATTTACAACGCGATCTTG
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_E351513_E35E083_1:0:0_4:0:0_5/1
AACCAGGATAACTTCAGGATAGTGCCATCGCCAAAATTCCAGCCAATATGTGTAGTGCCAATGAAGGCGCAAATCACCGGAATCGCCGCC
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_6E0545_6E106E_0:0:0_0:0:0_6/1
ACGAACACTGCCGGAACGCCCATCACGTTGCGATCGGTGATTTCGTTCTGGAAGGTGCCGCCGTCAATTGCAGTGTGCTTGATGCGCGGG
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_4814930_4815379_0:0:0_E:0:0_7/1
TCCTCATTTTTAACAATTGTATCAACAACCACCAAACCAGTTATAACCCTGGTCTTCCCAGTACCCCCCCGGAAAATGATTAGTGACCTC
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_E497311_E497915_3:0:0_1:0:0_8/1
GTGCTAACCTTAGCGCCCGCACATTTGCGTTTTATTTTTTATGTGGTGAACGTGACAGCAAATTCCGCGCTCTGGCGGCGGAACTGGCTG
+
;:LHHHHHGGHGGGHHGHHGHGGHHHHHHGGGHGHGGGGHGGGGHGGGHGGGGHGHGGGGGGGHHHHHHGHHHGHGHHHGGHGHGHHHGG
@Edited_3E00760_3E01EEE_0:0:0_1:0:0_9/1
ATGCGGGGGTTGAACACGCTCGTTCGTTGGGCATTCCGGTTATTGTTACCGATCACCATTTGCCAGGCGATACATTACCCGCAGCGGAAG
strain3_read1.fastq
```

fasta

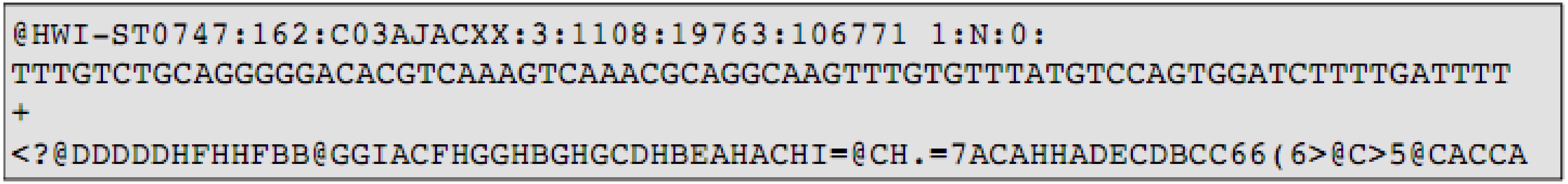
- Most basic file format to represent nucleotide or amino-acid sequences
- Each sequence is represented by:
  - A single description line (shouldn't exceed 80 characters):
    - Starts with "**>**"
    - Followed by the **sequence ID**, and a space, then
    - More information (**description**)
  - The sequence, over one or several lines (the number of characters per line is generally 70 or 80, but it does not matter)



```
>Protein1 Description of protein 1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGK
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEER
>DNA1 Description of dna segment 1
AACTCTCGCGTAGCTCAGAGAAGAGCTTGATCGATCGTGCTGCTGCTA
CCGCTAGTAGCTGTAGATCGTGCTAGTCAGCATCGATGCTAGCTAGCT
```

fastq

- same as fasta file but including quality scores
- contains 4 lines:
  - "@" and the sequence ID
  - the sequence
  - "+" (and the sequence ID)
  - the quality score



```
@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+
<?@DDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

ASCII encoding of phred scores

- one number : one leter

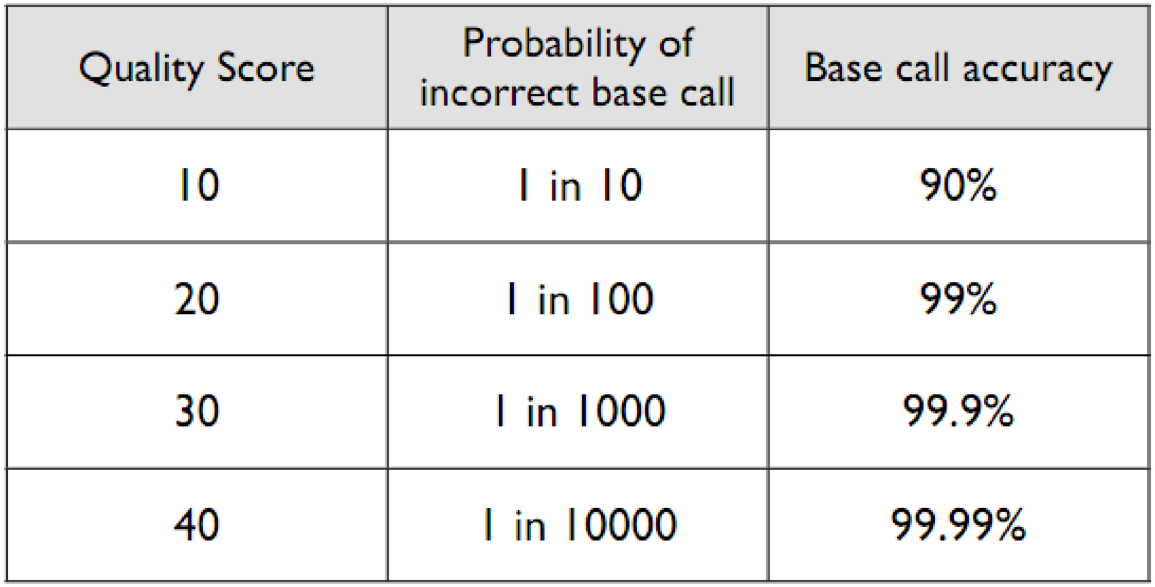
```
40:@
41:A
42:B
43:C
44:D
45:E
…  :…

90:Z
91:[
92:\
93:]
94:^
95:_
…  :…

141:a
142:b
143:c
144:d
145:e
146:f
…  :…
```

quality – Phred scores (Q)

- Most comonly used representation of qualities
- Related to the probability of errors (P) in a particular base

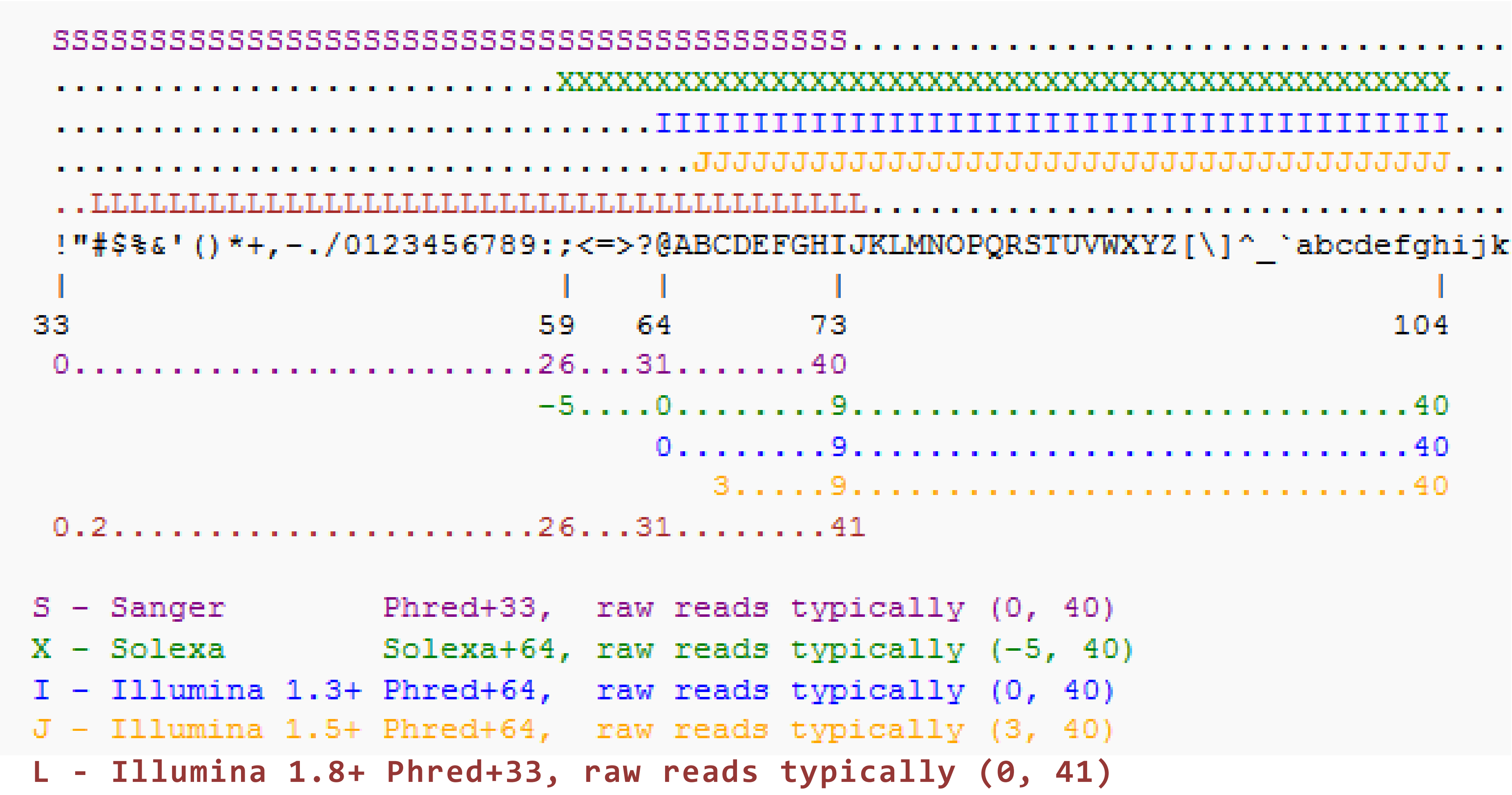
$$Q = -10\ \log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$



Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

Different scoring systems



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS..................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII...
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ...
  ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijk
  |                         |    |        |                              |
 33                        59   64       73                            104
  0........................26...31.......40
                           -5....0........9.............................40
                                 0........9.............................40
                                    3.....9.............................40
  0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)
```

http://en.wikipedia.org/wiki/FASTQ_format

You need to know the quality score encoding

STACKS:

- E: specify how quality scores are encoded, 'phred33' (Illumina 1.8+, Sanger, default) or 'phred64' (Illumina 1.3 - 1.5)

BOWTIE:

```
--phred33-quals    input quals are Phred+33 (default)
--phred64-quals    input quals are Phred+64 (same as --solexa1.3-quals)
--solexa-quals     input quals are from GA Pipeline ver. < 1.3
--solexa1.3-quals  input quals are from GA Pipeline ver. >= 1.3
--integer-quals    qualities are given as space-separated integers (not ASCII)
```

Quality control is important

- Some of the artefacts/problems that can be detected with QC
  - Sequencing
    - Sequence quality
  - Library preparation problems
    - Contaminations
    - Overrepresented sequences
    - Adaptor sequence presence
    - …

Sequence quality control

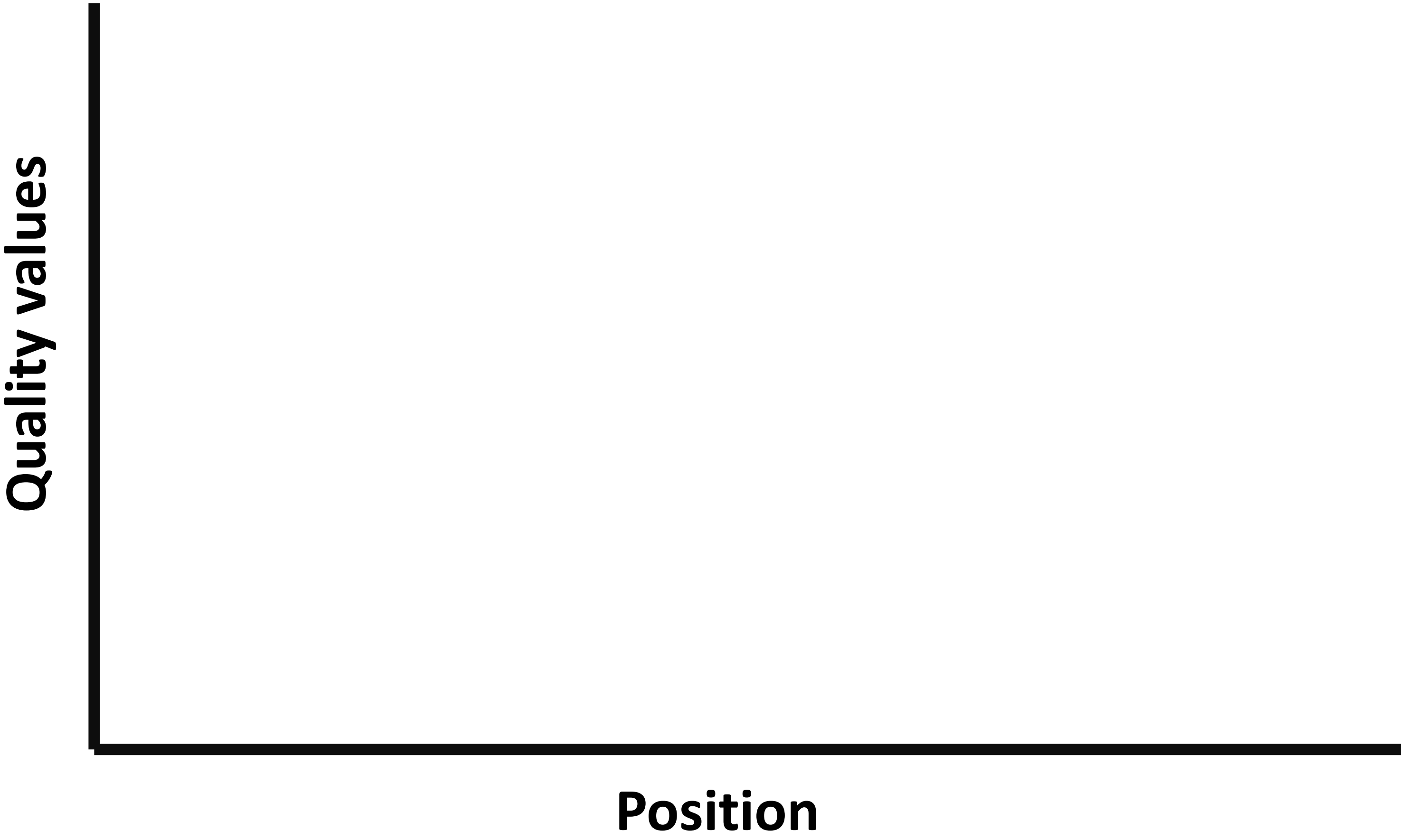
1. Look at how the data looks like

   Impossible to look at a >10 GB file to check if quality scores are adequate!

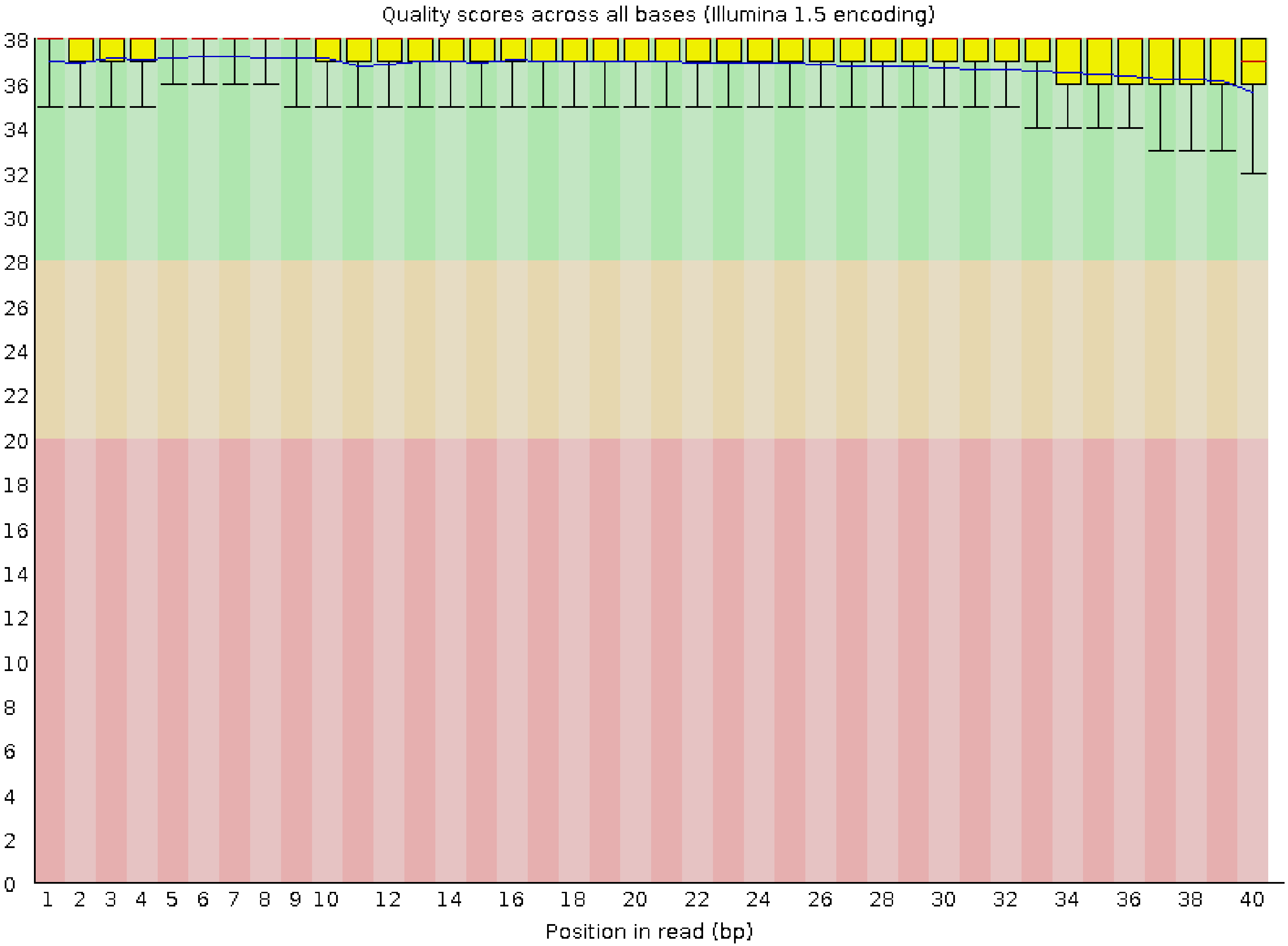
Sequence quality

```
@BENM_2011_0526_022:2:1:299:100/1
ATGGTGAACGCACATTGCCGTAATGGTCTGCGACTGTTCGCCAAAAGGCAATTCAATTAATTCTCCGCACTTGAACTCTTTTA
+
FEFFFEEEFEFEFFEEFFEEFFFEEFEFFEFFFFFEFEEFEFEEEFEFEEFFEFEFEFFEEFEEFFEFFEEFEEEFFFFFEEF
@BENM_2011_0526_022:8:2:303:100/1
GGCATTACGTTTTGCTACCATCGTGCCGAGGCGCTTATTGCTGGCATCATGCTCATGATAATGAATTTGAATCAGGGGTGATA
+
EDEDDEEEEEDEDDEEDEDEDDDEDEEEEEDDEEEEDDDDDDDDDDDDEEEDDEEDDEEEDEEEEDDDEEDDEDEEEDDEDD
@BENM_2011_0526_022:4:3:307:100/1
ATTCATAATGGCGCTGCAGCAATTGCTGGTTGAGGTTGTAACTGCGCACCAACATACCGATTTCATCGTCCTGATGCAGACGC
+
GGHGHHGHHGGHHHHGHGGHGHHHGGHHGHGHGGGGHGHGHHHGGHGHGGHGHHHGGHHGGHGGHHGHGHHHGGHGHHGGGHH
@BENM_2011_0526_022:7:4:292:100/1
TTCGACCGCATCCCACGGCACCCCACTGATCCCGGAAACACCAGCGGTTTCAATGCGGGTCAGCCTATGGTGAAGGCCGTGAC
+
IIHIHHIIHHHHIHIIHHHIHHIHHIHIHIHIIHHHIIHIIHIIHHIHHIIHIIIIIIHHHHIIHHIHHIIIHIHHHHHIHHI
@BENM_2011_0526_022:4:5:317:100/1
TTTTCACGCTGGGGCGTGACGGCATTCATTAACCCGCTTTCTTTGGCGATCTTCTCGATCTTCGCTTTCTCGCTTTCCGGCAC
+
BBBBCCCBCBBBBCBBCCCCCCBBCBCCCCBBBCBBBBBBCCCBCCBCCCCCCCCBCCBBBCBCCCCBCCBBBBCBBCBCCCB
@BENM_2011_0526_022:2:6:309:100/1
GATGGGAAAAGTAAAGTGACAGTTCGCGCATCCACCGCGACAACGGTTCCAAGTCCCAATTCGCTTTCTGTATCACTGATCCA
+
IHHIHIIHHIHHHHHHIIIIIIHHHIIHIIIIIHHIHHIHIIIHIIIIHIIIHHIIHHIIIIIHHHHHHIIIIIHIIHIIHHI
```

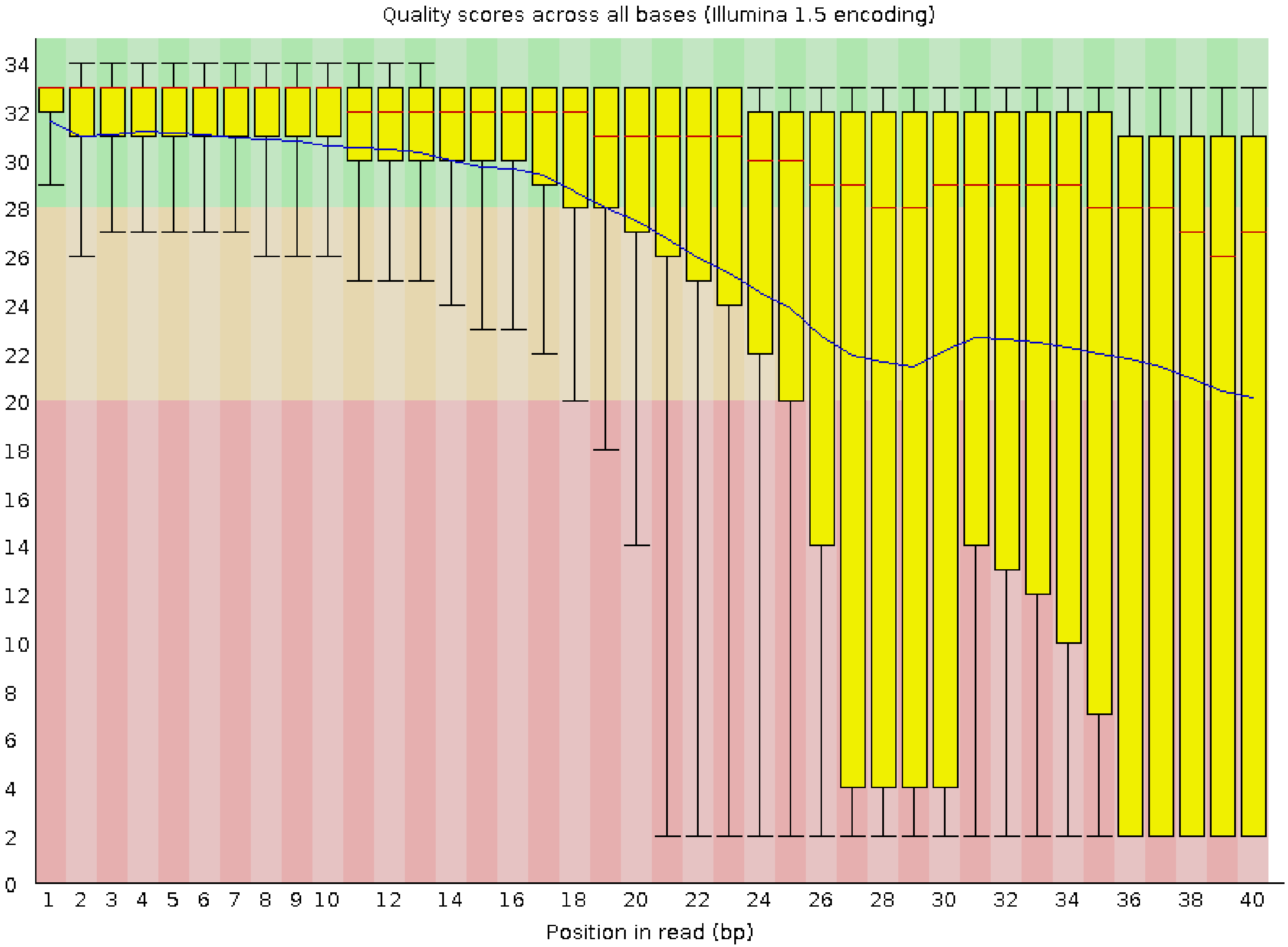
Quality plots



Sequence quality



Sequence quality



Sequence quality control

1. Look at how the data looks like

   Impossible to look at a >10 GB file to check if quality scores are adequate!

2. Decide what to do:

   Nothing (some programs take quality into account)

   Clean:

   - Trim all reads to a certain length
   - Trim bad quality bases
   - Discard bad quality reads

Data cleaning

Trimmomatic:

http://www.usadellab.org/cms/?page=trimmomatic

Fastx-toolkit:

http://hannonlab.cshl.edu/fastx_toolkit/index.html

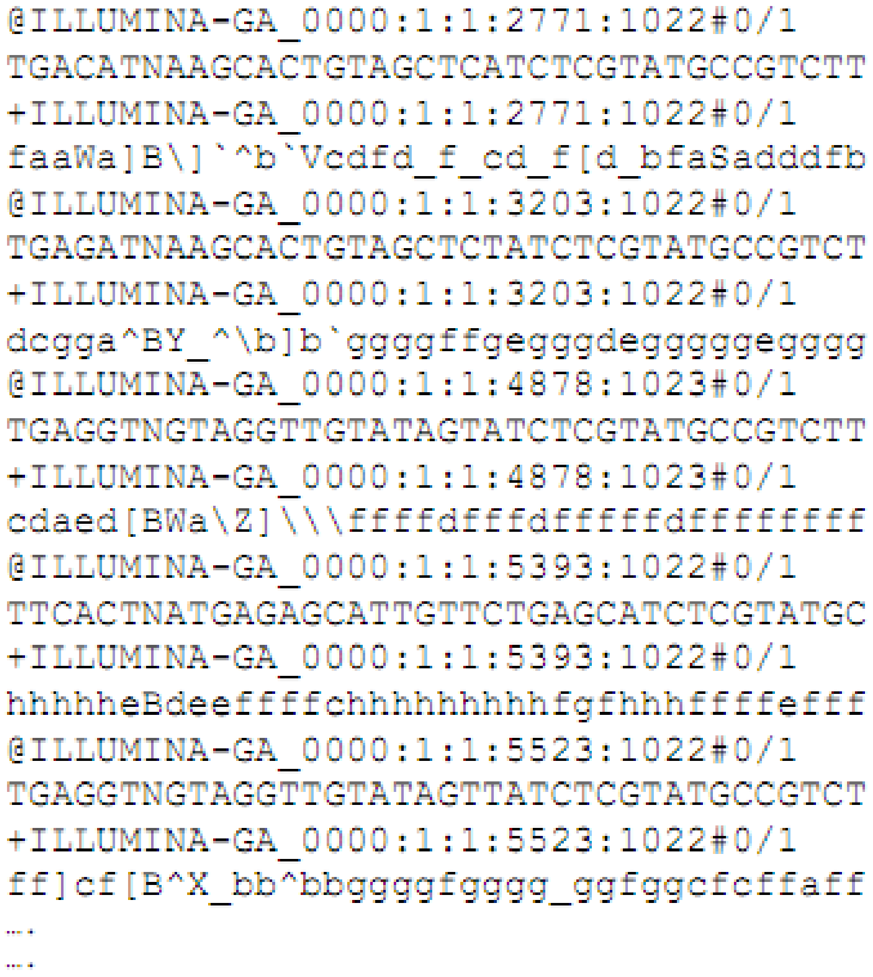
FastqMcf:

http://code.google.com/p/ea-utils/wiki/FastqMcf

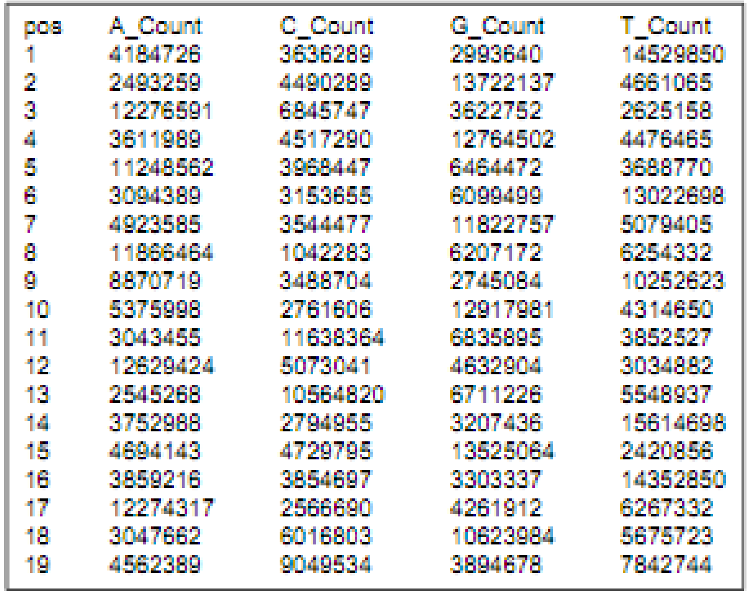
Quality control is important

- Some of the artefacts/problems that can be detected with QC
  - Sequencing
    - Sequence quality
  - Library preparation problems
    - Contaminations
    - Overrepresented sequences
    - Adaptor sequence presence
    - …

Nucleotide composition

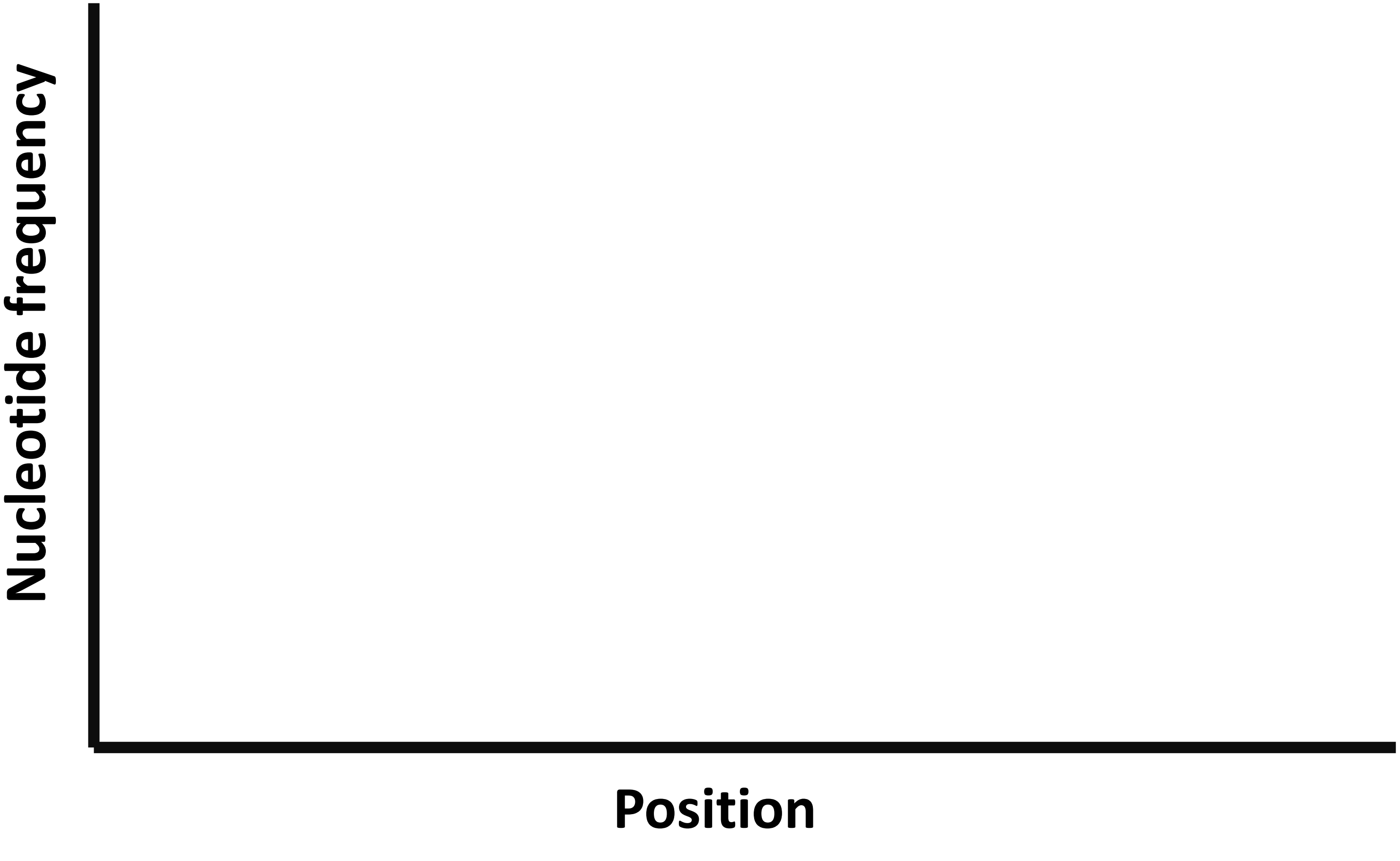


```
@ILLUMINA-GA_0000:1:1:2771:1022#0/1
TGACATNAAGCACTGTAGCTCATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:2771:1022#0/1
faaWa]B\]`^b`Vcdfd_f_cd_f[d_bfaSadddfb
@ILLUMINA-GA_0000:1:1:3203:1022#0/1
TGAGATNAAGCACTGTAGCTCTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:3203:1022#0/1
dcgga^BY_^\b]b`ggggffgegggdegggggegggg
@ILLUMINA-GA_0000:1:1:4878:1023#0/1
TGAGGTNGTAGGTTGTATAGTATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:4878:1023#0/1
cdaed[BWa\Z]\\\ffffdfffdffffdffffffff
@ILLUMINA-GA_0000:1:1:5393:1022#0/1
TTCACTNATGAGAGCATTGTTCTGAGCATCTCGTATGC
+ILLUMINA-GA_0000:1:1:5393:1022#0/1
hhhhheBdeeffffchhhhhhhhhfgfhhhffffefff
@ILLUMINA-GA_0000:1:1:5523:1022#0/1
TGAGGTNGTAGGTTGTATAGTTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:5523:1022#0/1
ff]cf[B^X_bb^bbggggfgggg_ggfggcfcffaff
….
….
```



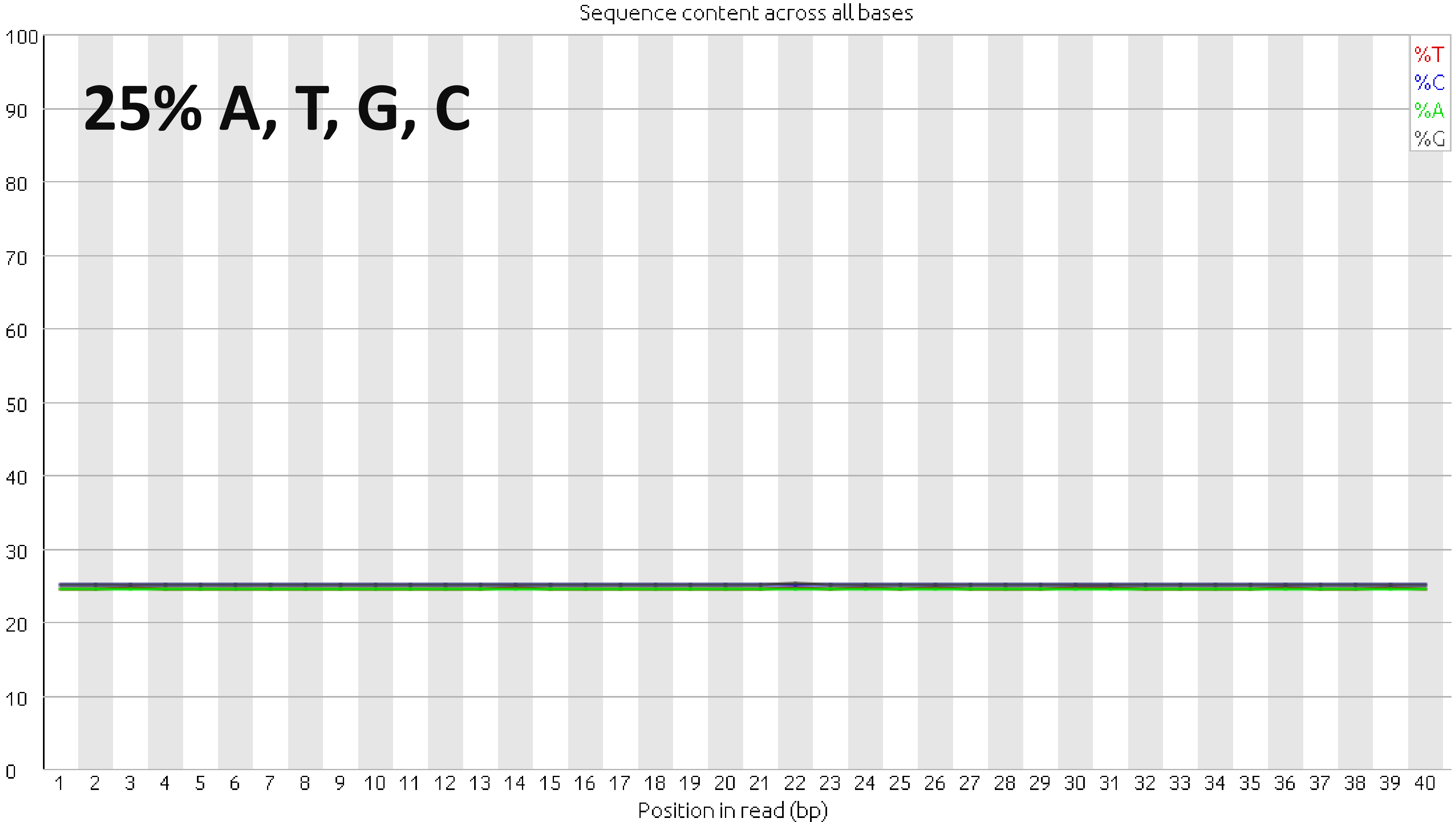
pos	A_Count	C_Count	G_Count	T_Count
1	4184726	3636289	2993640	14529850
2	2493259	4490289	13722137	4661065
3	12276591	6845747	3622752	2625158
4	3611989	4517290	12764502	4476465
5	11248562	3968447	6464472	3688770
6	3094389	3153655	6099499	13022698
7	4923585	3544477	11822757	5079405
8	11866464	1042283	6207172	6254332
9	8870719	3488704	2745084	10252623
10	5375998	2761606	12917981	4314650
11	3043455	11638364	6835895	3852527
12	12629424	5073041	4632904	3034882
13	2545268	10564820	6711226	5548937
14	3752988	2794955	3207436	15614698
15	4694143	4729795	13525064	2420856
16	3859216	3854697	3303337	14352850
17	12274317	2566690	4261912	6267332
18	3047662	6016803	10623984	5675723
19	4562389	9049534	3894678	7842744

Nucleotide composition graphs



Nucleotide composition graphs

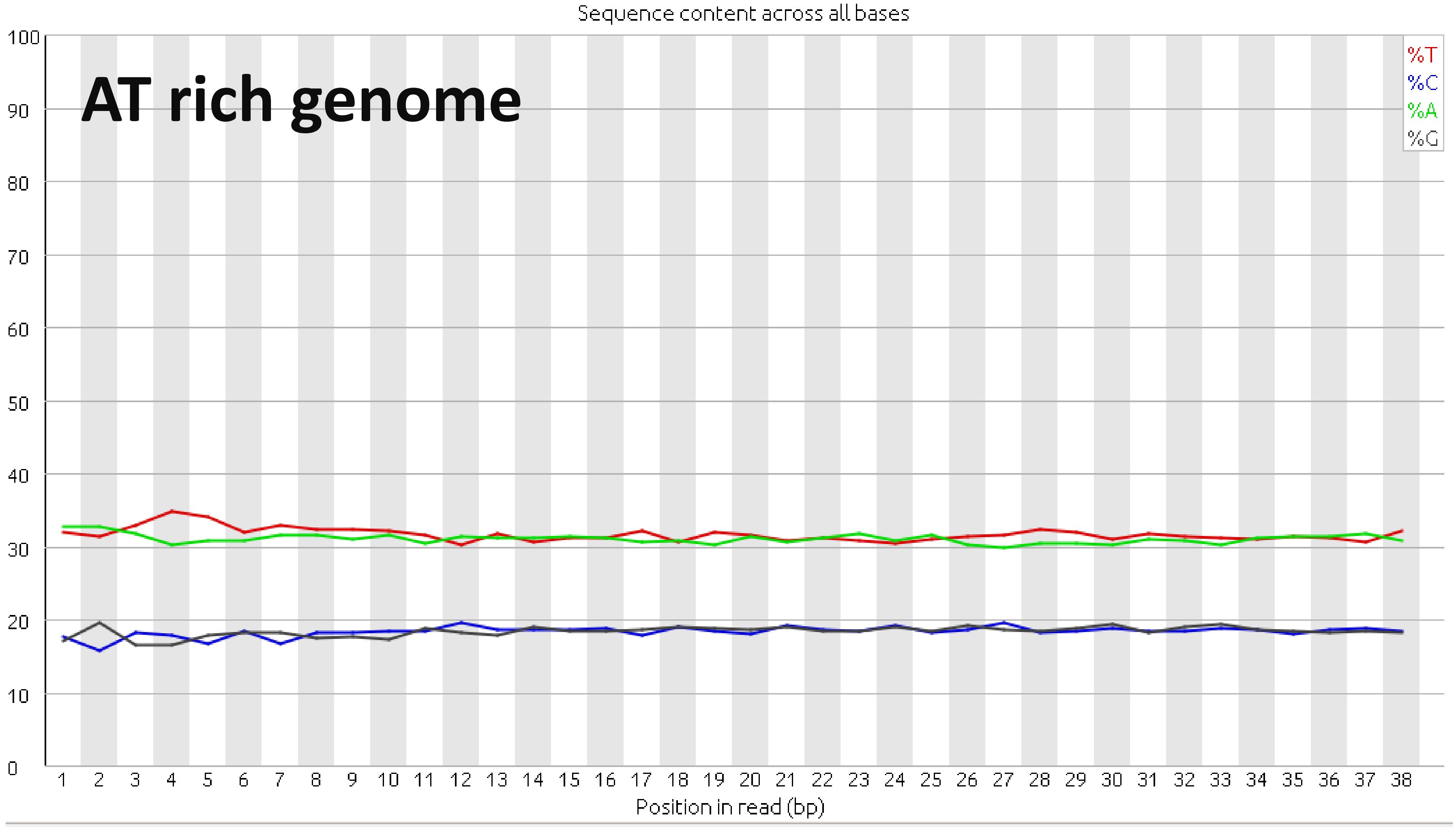
Are they what you expect?



25% A, T, G, C

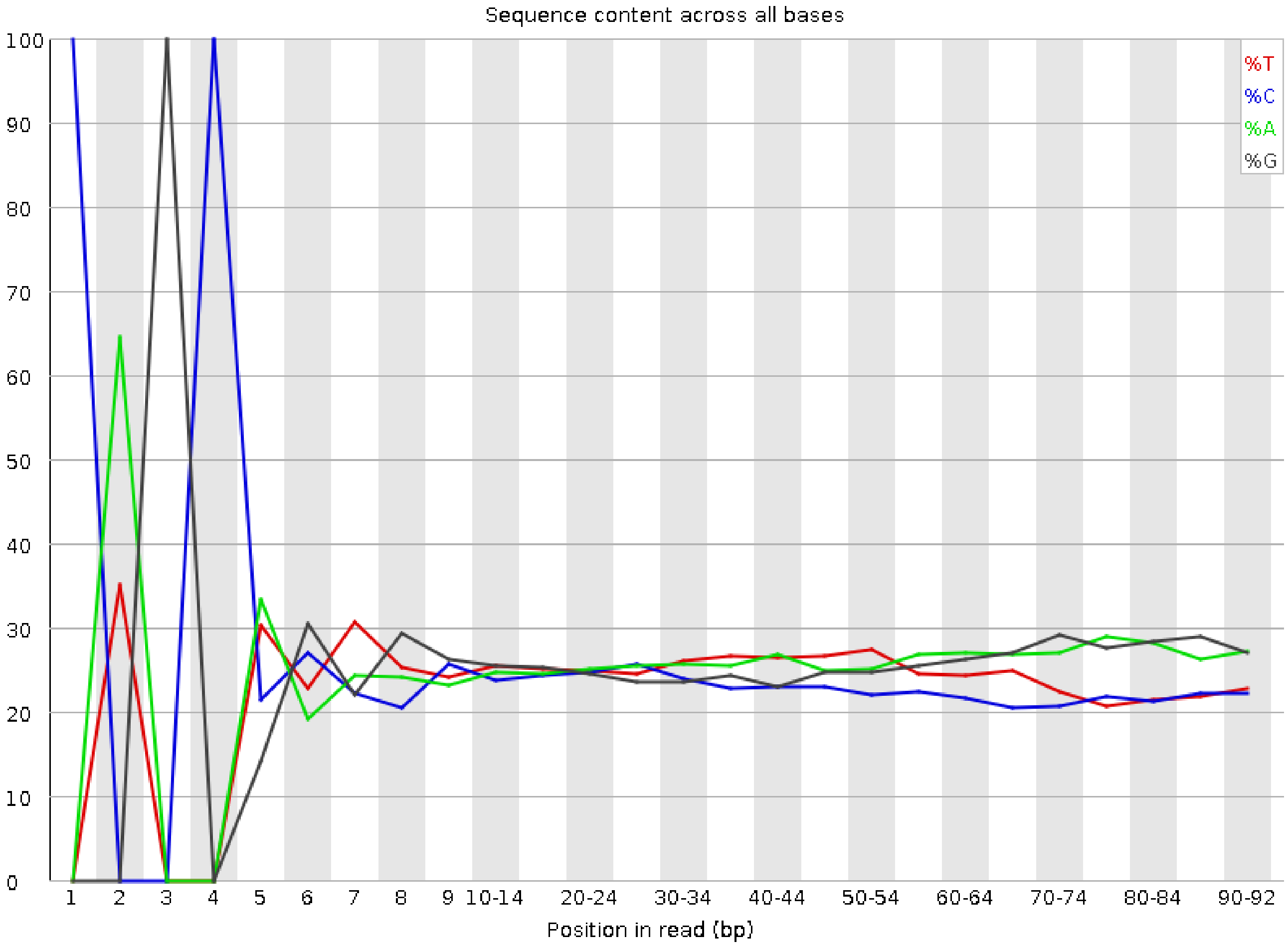
Nucleotide composition graphs

Are they what you expect?



Nucleotide composition graphs

Are they what you expect?



Data cleaning

Trimmomatic:

http://www.usadellab.org/cms/?page=trimmomatic

Fastx-toolkit:

http://hannonlab.cshl.edu/fastx_toolkit/index.html

FastqMcf:

http://code.google.com/p/ea-utils/wiki/FastqMcf

DATASETS

- **DATASET 1:** Genome sequencing of *Bartonella*
- **DATASET 2:** Amplicon sequencing of 16S rRNA
- **DATASET 3:** RAD-sequencing data of 12 samples
- **DATASET 4:** Shotgun metagenomics sequencing
- **DATASET 5:** microRNA sequencing

PROGRAMS

- **fastqc:**
  - ***User interface***
  - ***Command line***

- **Trimmomatic:**
  - **Command line**

For command line, check synopsis!

Synopsis

How you call a program: `program_name [SOMETHING]`

```
fastqc [-o output dir] [--(no)extract] [-f fastq|bam|sam]
           [-c contaminant file] seqfile1 .. seqfileN
```

```
trimmomatic-0.32.jar PE [-threads <threads>] [-phred33|-
phred64] [-trimlog <trimLogFile>] [-basein <inputBase> |
<inputFile1> <inputFile2>] [-baseout <outputBase> |
(<outputFile1P> <outputFile1U> <outputFile2P> <outputFile2U>]
<trimmer1>...
   or:
trimmomatic-0.32.jar SE [-threads <threads>] [-phred33|-
phred64] [-trimlog <trimLogFile>] <inputFile> <outputFile>
<trimmer1>...
```

Exercise

http://evomics.org/learning/quality-assessment-and-control-of-sequence-data/

Exercise 1: DATASET 1

- There are 10000 sequences of 38 nucleotides length. The total GC content is 37%.
- Quality score is ~37 (Q=37 → P = 1/5011)

$$P = 10^{\frac{-Q}{10}}$$

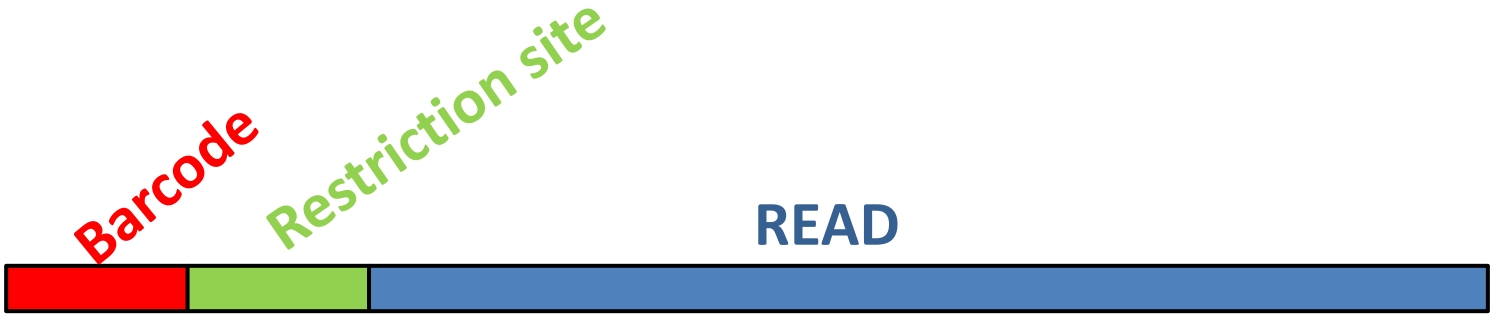
- The sequences are average good quality
- Sequences are AT rich – this is expected in Bartonella
- There is an adaptor contamination that is recognized by the program

Exercise 1: DATASET 2

- Conserved sequence at the beginning of the reads:
  - TACAGAGG
- Sequences from a conserved region of the 16S rRNA
- Some sequences are more frequent than others
  - Frequencies of the different bacteria in the sample are different

Exercise 1: DATASET 3

- fastqc –h
- RAD-tag



Exercise 2: Dataset 4

```
trimmomatic-0.32.jar  PE -phred33 -trimlog sample1.log sample_1_R1.fastq \
sample_1_R2.fastq  sample_1_P1.fastq sample_1_U1.fastq sample_1_P2.fastq \
sample_1_U2.fastq \
SLIDINGWINDOW:size:score \
LEADING:3 \
TRAILING:3 \
MINLEN:80
```

	FR keep	F keep	R keep	FR drop
4:35	0	0	0	100
4:32	83.6	8.9	6.7	0.7
10:35	43.5	20.8	23.9	11.7
10:32	96.6	1.9	1.4	0

Forward	Reverse
⬜	⬜
🟩	🟥
⬜	⬜
🟩	🟩
⬜	⬜
🟩	🟩
🟥	🟩
🟩	🟩
🟩	🟩
🟩	🟩
⬜	⬜

Exercise 2: DATASET 5

- The quality of some sequences drops down towards the end of the read
- The per base sequence content plot show that there are sequences that are more frequent than others
- The sources of the overrepresented sequences are:
  - Illumina adaptor /sequencing primer sequences
  - microRNAs that are more frequent than others

Exercise 2: DATASET 5

```
trimmomatic-0.32.jar  SE -phred33 -trimlog SRR026762.log \
SRR026762-sample.fastq  SRR026762-sample_trim.fastq \
ILLUMINACLIP:adapters/microRNA.fa:2:30:10
```

ADAPTER:

Surviving: 98966 (98.97%) Dropped: 1034 (1.03%)

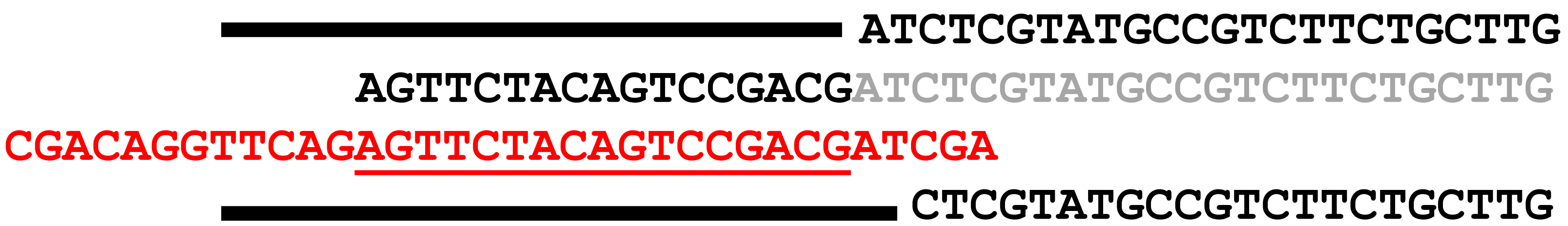
ADAPTER+SEQUENCING PRIMER

Surviving: 93163 (93.16%) Dropped: 6837 (6.84%)

ADAPTER+SEQUENCING PRIMER+POTENTIAL DIMER

Surviving: 70934 (70.93%) Dropped: 29066 (29.07%)

Exercise 2: DATASET 5



ATCTCGTATGCCGTCTTCTGCTTG

AGTTCTACAGTCCGACGATCTCGTATGCCGTCTTCTGCTTG

CGACAGGTTCAGAGTTCTACAGTCCGACGATCGA

CTCGTATGCCGTCTTCTGCTTG