

Quality assessment and control of sequence data

Genomic filetypes

<http://evomics.org/resources/file-formats/>

fasta + qual

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
```

qual

- Fasta-like quality format
- Always paired with a fasta file (sequences with same ids, same order)
- Description line as in fasta format
- Qualities: a number for each base in the corresponding fasta, separated by spaces

```
>DNA1 quality of dna segment 1  
23 23 22 24 23 12 3 34 23 23 22 21 23 34 32 9 14 25  
22 24 23 12 3 34 23 23 22 21 23 34 32 9 14 25 21 21
```

quality – Phred scores (Q)

- Most commonly 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%

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  
+  
<?@DDDDDDHFHHFBB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

ASCII encoding of phred scores

- one number : one leter

64 : @

65 : A

66 : B

67 : C

68 : D

69 : E

... : ...

90 : Z

91 : [

92 : \

93 :]

94 : ^

95 : _

... : ...

97 : a

98 : b

99 : c

100 : d

101 : e

102 : f

... : ...

Different scoring systems

[illegible]

Phred scores (Q) go from 0 to 40

Sanger and Illumina 1.8:

ASCII = 33 $\rightarrow$ Q = 0; Q = ASCII - 33, BUT
scores end in 40 and 41 respectively

Solexa, Illumina 1.3 and Illumina 1.5:

ASCII = 64 $\rightarrow$ Q = 0; Q = ASCII - 64, BUT
scores start in -5, 0 and 3 respectively

Quiz 1

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

- What are the phred scores (and probability of error) of the first five bases of this sequence?

```
@phredscoretest
```

```
AATTCTGGTACTCTAATGTGTTGTATATGGTCCTATC
```

```
+
```

```
FEGIFDCBAGFJDGFABBD9837.9##(. , +) ..+**
```

Quiz 2

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

- Can you guess the sequencing platform/base caller that was used to generate this data?

```
@phredscoretest
```

```
AATTCTGGTACTCTAATGTGTTGTATATGGTCCTATC
```

```
+
```

```
FEGIFDCBAGFJDGFABBD9837.9##(. , +) . . +**
```

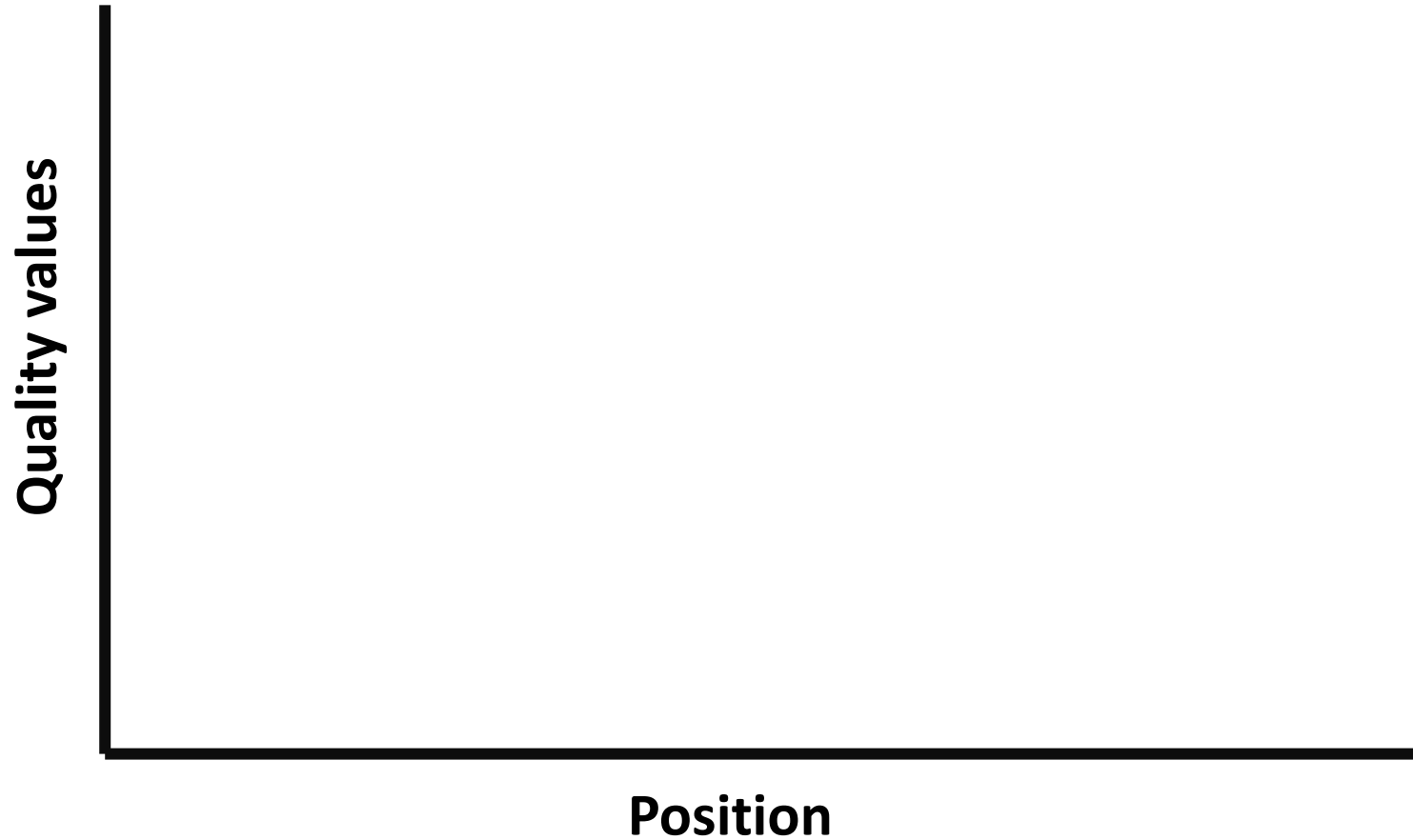
Quality control is important

- Some of the artefacts/problems that can be detected with QC
 - Sequence quality
 - Nucleotide composition
 - Technical issues
 - Overrepresented sequences
 - Adaptor sequence presence

Sequence quality

	pos	mean	Q1	med	Q3
@ILLUMINA-GA_0000:1:1:2771:1022#0/1	1	38.01	38	39	40
TGACATNAAGCACTGTAGCTCATCTCGTATGCCGTCTT	2	37.73	38	39	40
+ILLUMINA-GA_0000:1:1:2771:1022#0/1	3	37.53	38	39	40
faaWa]B\]^`b`Vcdfd_f_cd_f[d_bfaSadddfb	4	37.69	38	39	40
@ILLUMINA-GA_0000:1:1:3203:1022#0/1	5	37.41	37	39	40
TGAGATNAAGCACTGTAGCTCTATCTCGTATGCCGTCT	6	37.63	38	39	40
+ILLUMINA-GA_0000:1:1:3203:1022#0/1	7	37.52	38	39	40
dcgga^BY_`^b]b`ggggffgegdegggggggggg	8	36.95	36	38	40
@ILLUMINA-GA_0000:1:1:4878:1023#0/1	9	37.31	37	39	40
TGAGGTNGTAGGTTGTATAGTATCTCGTATGCCGTCTT	10	37.59	38	39	40
+ILLUMINA-GA_0000:1:1:4878:1023#0/1	11	37.61	38	39	40
cdaed[BWa\Z]\`\\ffffdffffdffffdffffdffff	12	37.28	36	39	40
@ILLUMINA-GA_0000:1:1:5393:1022#0/1	13	37.71	38	39	40
TTCACATNATGAGAGCATTGTTCTGAGCATCTCGTATGC	14	37.62	38	39	40
+ILLUMINA-GA_0000:1:1:5393:1022#0/1	15	37.54	37	39	40
hhhhheBdeeffffchhhhhhhhhfghhhffffefff	16	37.97	38	39	40
@ILLUMINA-GA_0000:1:1:5523:1022#0/1	17	37.44	37	39	40
TGAGGTNGTAGGTTGTATAGTTATCTCGTATGCCGTCT	18	37.49	38	39	40
+ILLUMINA-GA_0000:1:1:5523:1022#0/1	19	37.36	37	39	40
ff]cf[B^X_bb^bbggggfgggg_ggfggcfcffaff					

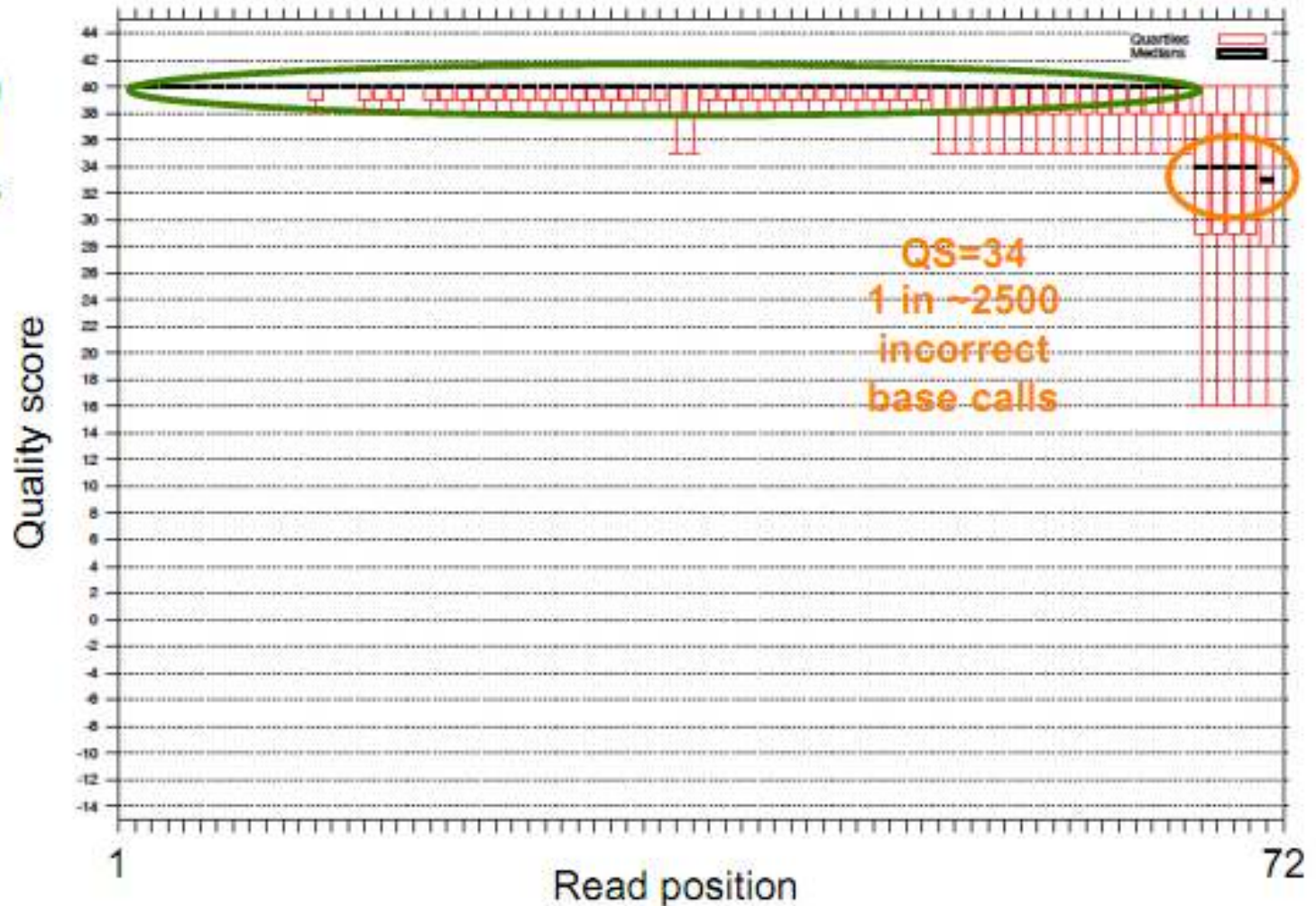
Quality plots



Sequence quality

GOOD QUALITY RUN

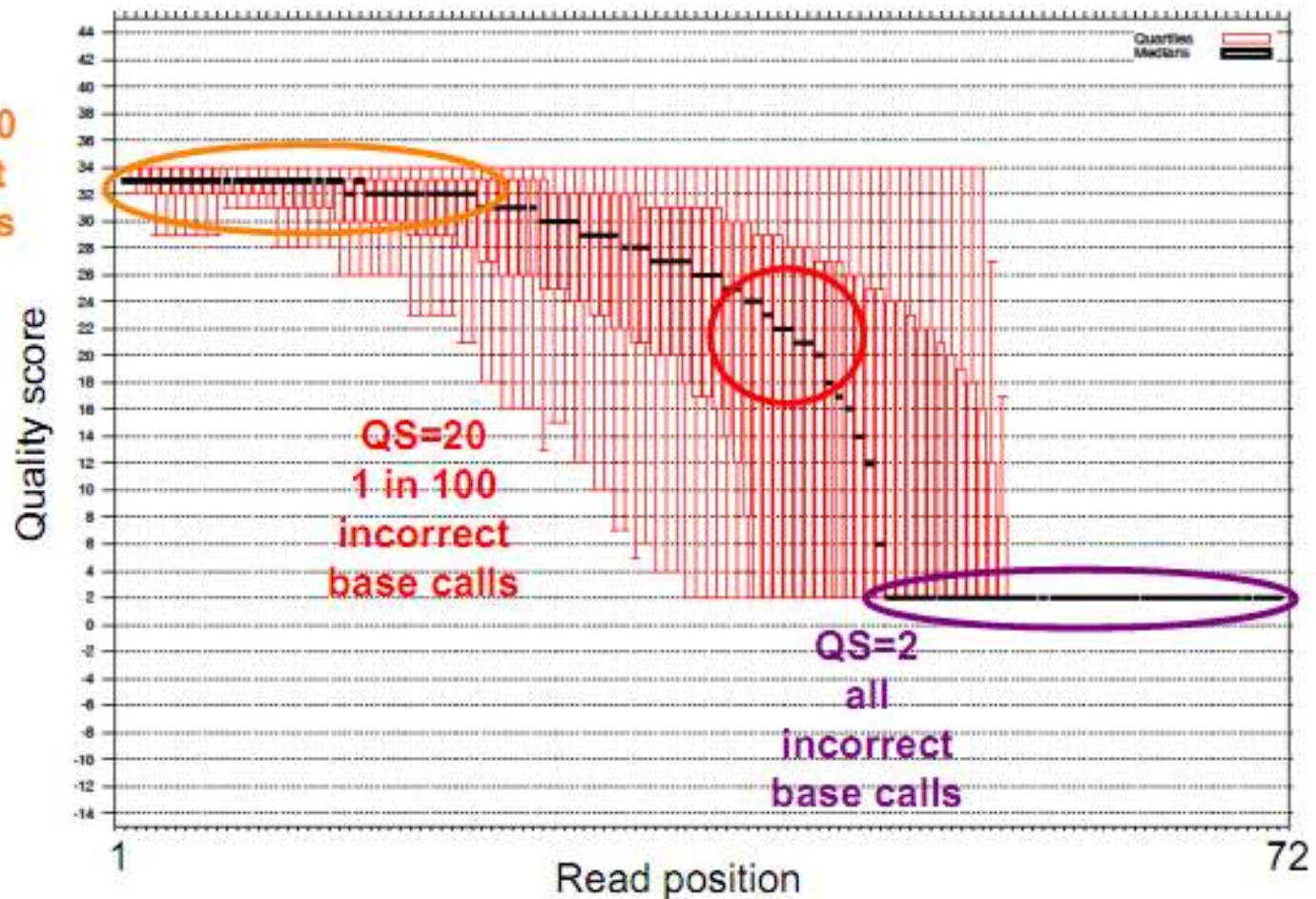
QS=40
1 in 10000
incorrect
base calls



Sequence quality

BAD QUALITY RUN

QS=32
1 in ~1500
incorrect
base calls



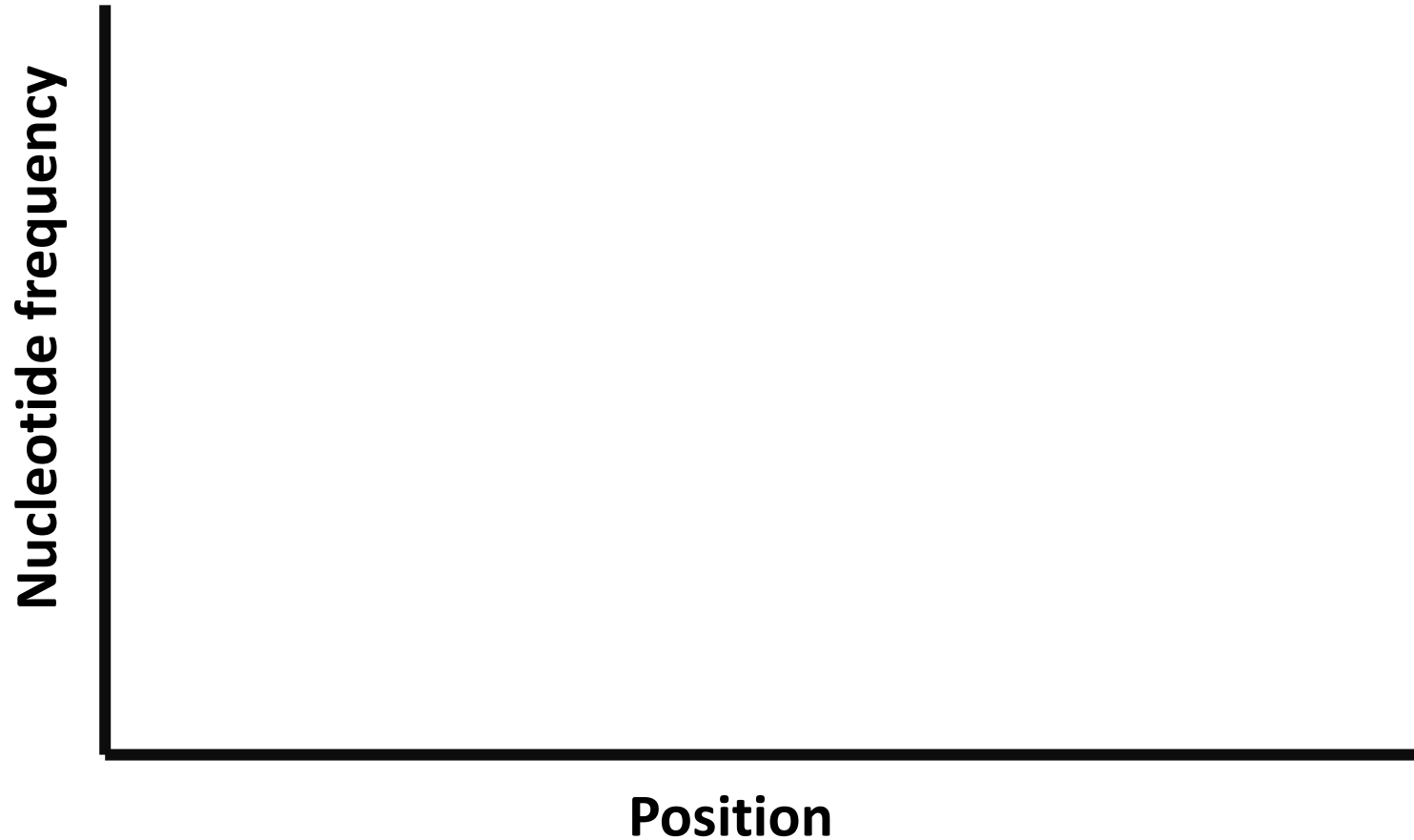
Nucleotide composition

```
@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`ggggffgegdeggggggeg
@ILLUMINA-GA_0000:1:1:4878:1023#0/1
TGAGGTNGTAGGTTGTATAGTATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:4878:1023#0/1
cdaed[BWa\Z]\\\ffffdffffdffffdffffffffff
@ILLUMINA-GA_0000:1:1:5393:1022#0/1
TTCACATNATGAGAGCATTTGTTCTGAGCATCTCGTATGC
+ILLUMINA-GA_0000:1:1:5393:1022#0/1
hhhhhheBdeeffffchhhhhhhhhfghhhfffefff
@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
```

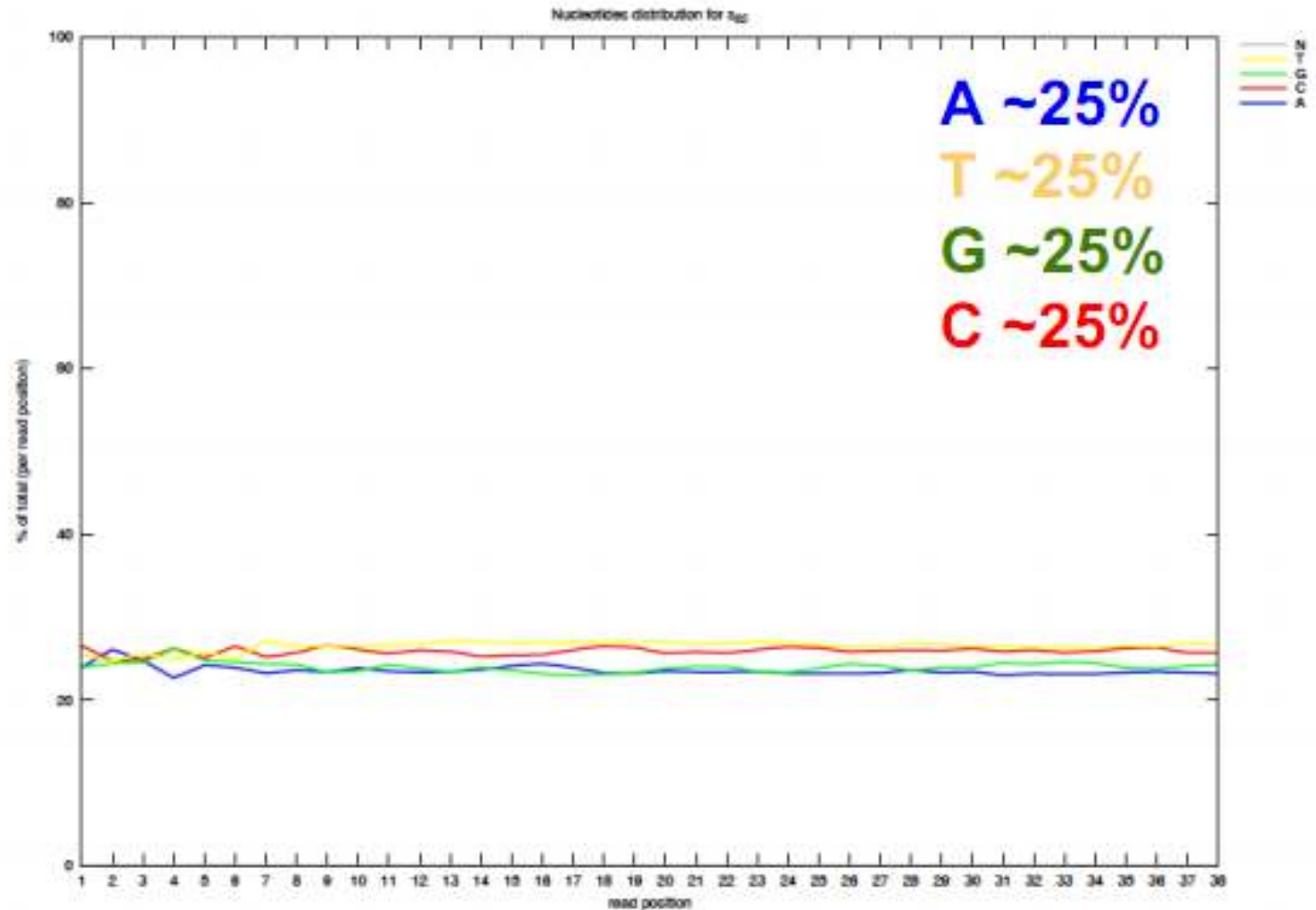
```
...
...
```

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

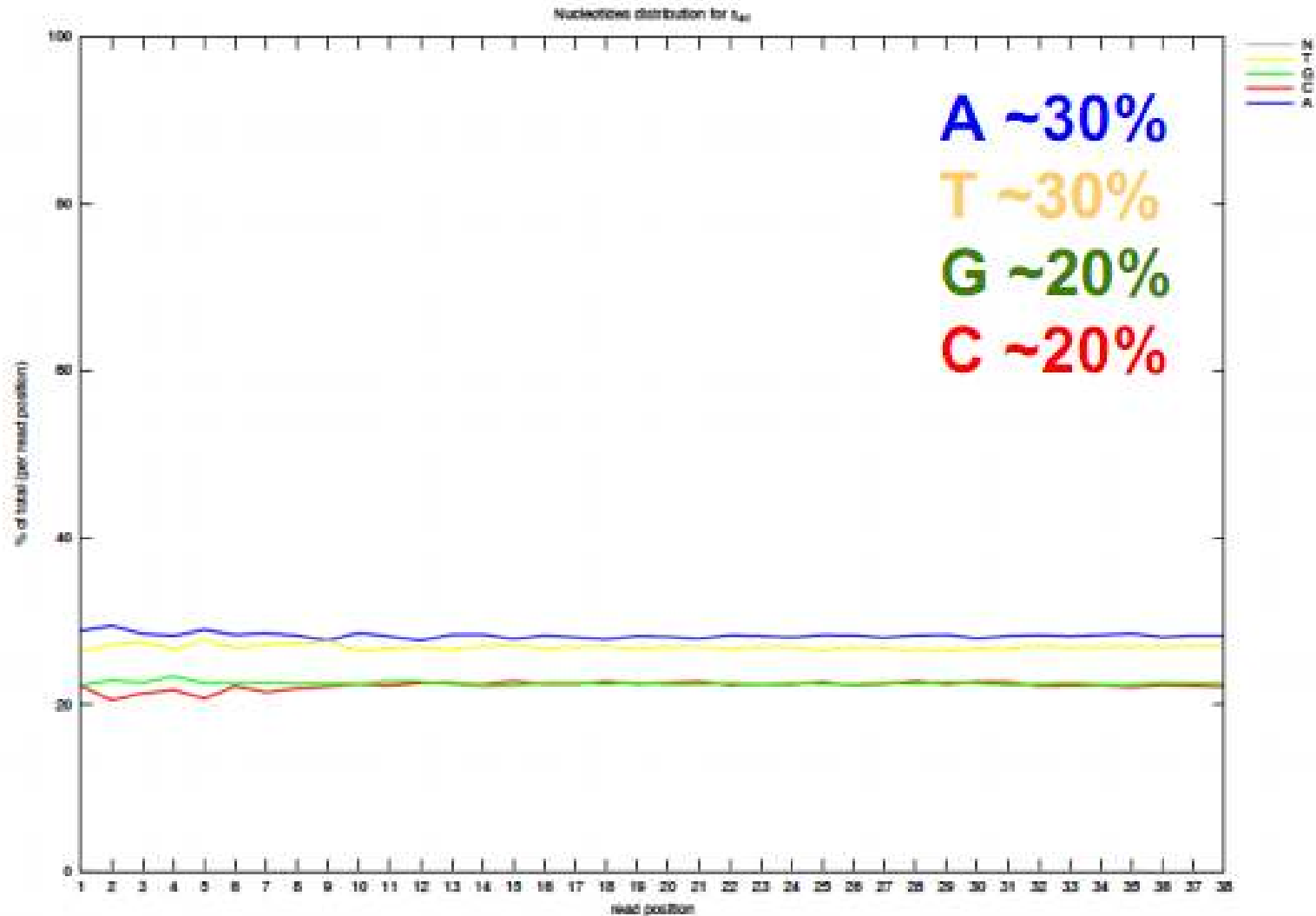
Nucleotide composition graphs



Nucleotide composition



Nucleotide composition



Exercise

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