Sequence read quality

What it means

How it is determined

How it is represented in data files

The basics

- Sequence read quality is evaluated on a **base-by-base basis**
  - The quality of each base call is evaluated by a number
  - I.e. **quality scores**
- Whole reads, or sections of reads, can therefore be assessed on the quality properties of their bases
- E.g. each 'window' of *N* consecutive bases can easily have its properties determined, such as:
  - Average quality
  - Lowest quality base in the window, etc
- These properties can be used to decide where to 'trim' reads to remove poor quality segments
  - And which reads to discard entirely

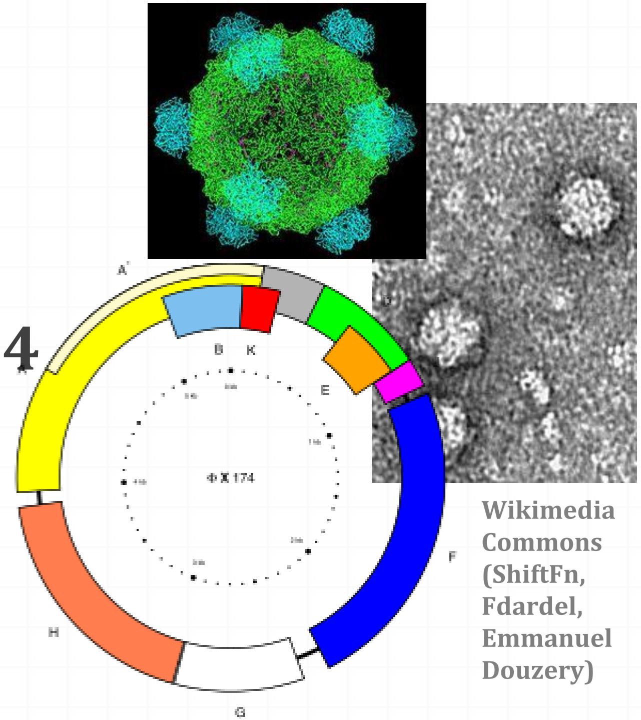
Dealing with sequence quality scores

How sequence quality is determined

- A gold-standard reference sequence
  - i.e. you know the sequence for certain
- An objective way of measuring properties of your sequencing readout
- As an example, we will consider –
  - Good, old fashioned, **Sanger sequencing**
- Why Sanger?
  - It's as good as any as understanding the principles which link sequencing readout metrics with the reliability of the sequence
- Similar principles apply to NGS platforms

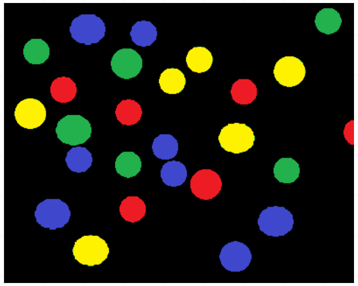
The reference sequence

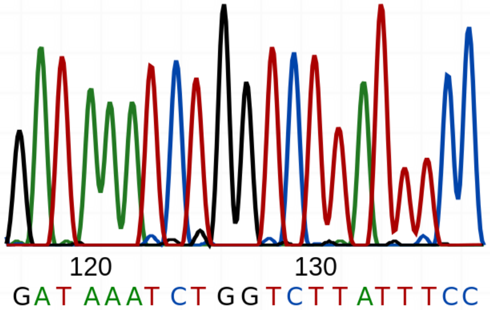
- You need to be certain of the sequence.
- Therefore, it needs to be a piece of DNA which has been sequenced many, many times
  - so that each base is in no doubt
- So ideally it won’t be huge (or tiny)
  - and it will be easily maintainable
- Step forward **bacteriophage ΦX174**
- It’s genome is only about 5Kb



Wikimedia Commons (ShiftFn, Fdardel, Emmanuel Douzery)

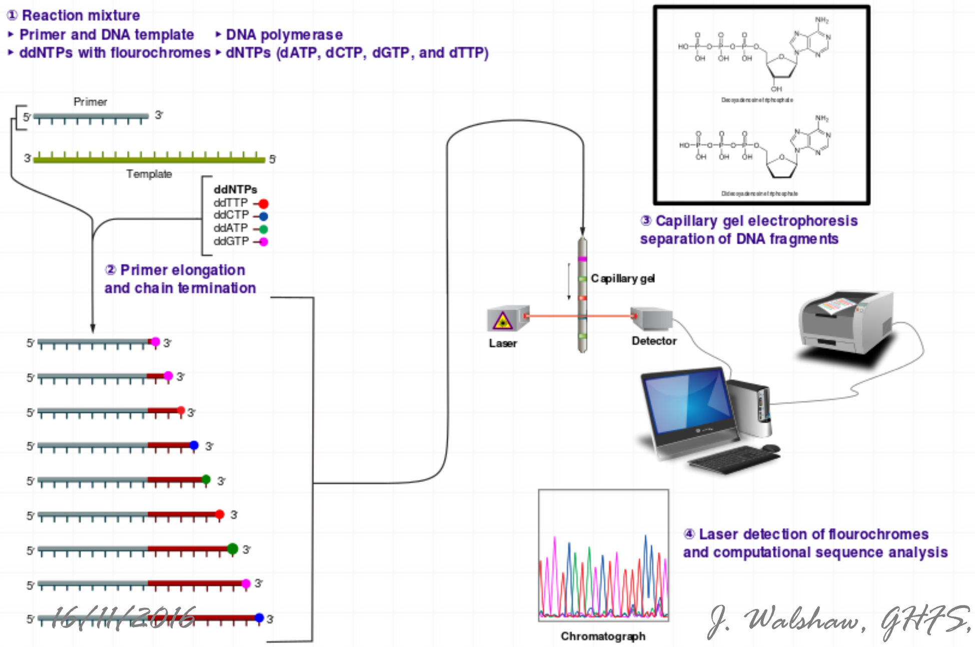
- Having a known sequence enables benchmarking
  - Basically, evaluate the characteristics of the sequencing readout
- Measurable properties of the readout
- what are these properties when a based is called:
  - Correctly
  - Incorrectly



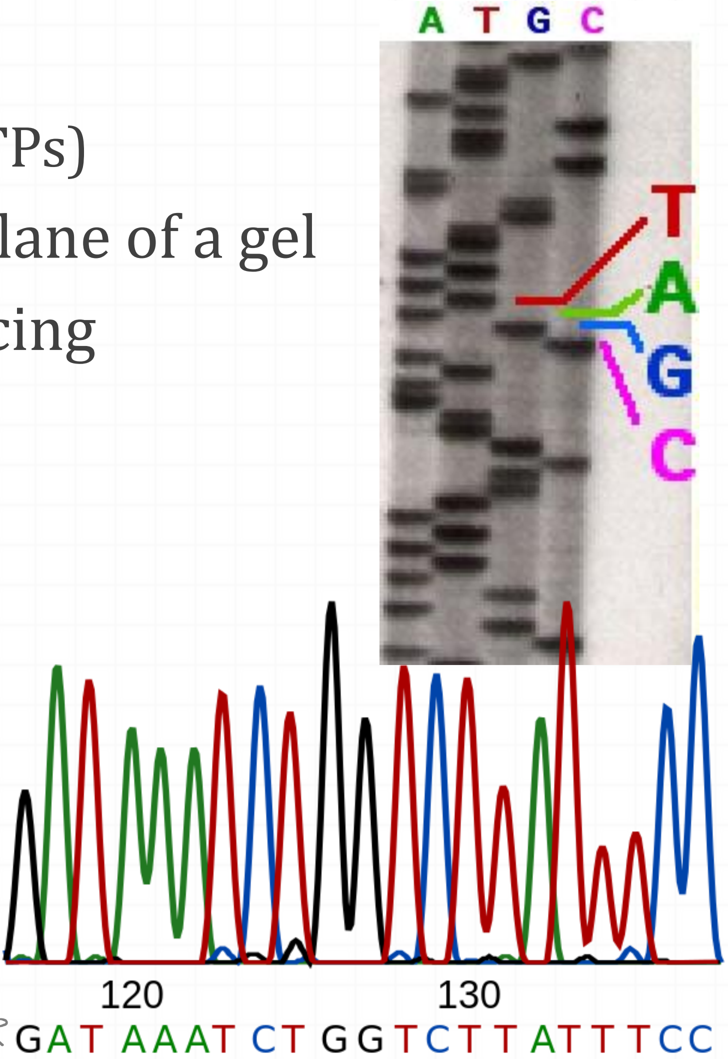


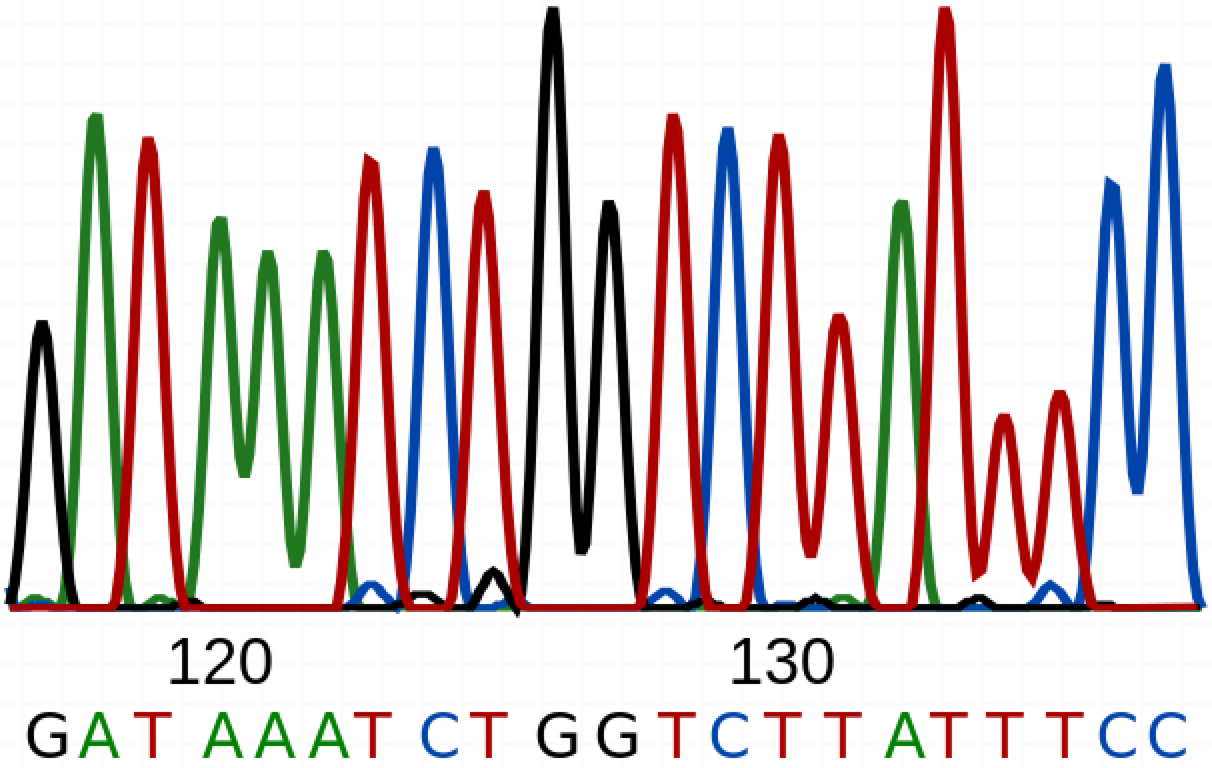
Example readout: Sanger

- One 'lane' for each of A, C, G, T
  - (due to labelled terminating dNTPs)
- Really old-fashioned – an actual lane of a gel
- Superceded by capillary sequencing

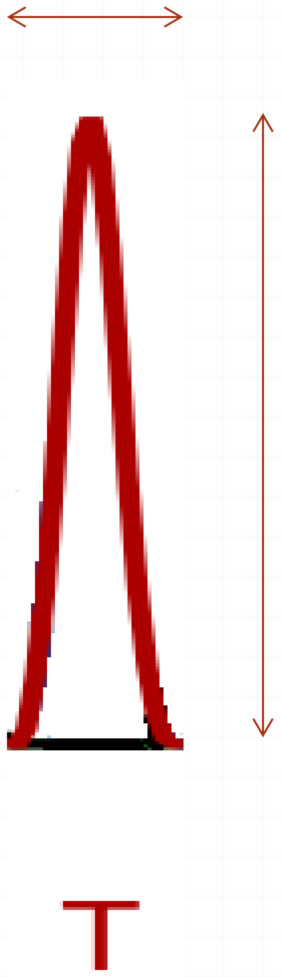


① Reaction mixture
‣ Primer and DNA template
‣ DNA polymerase
‣ ddNTPs with flourochromes
‣ dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
5'
3'
3'
5'
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
NH₂
HO
OH
Deoxyadenosine triphosphate
Dideoxyadenosine triphosphate
③ Capillary gel electrophoresis separation of DNA fragments
Capillary gel
Laser
Detector
④ Laser detection of flourochromes and computational sequence analysis
Chromatograph





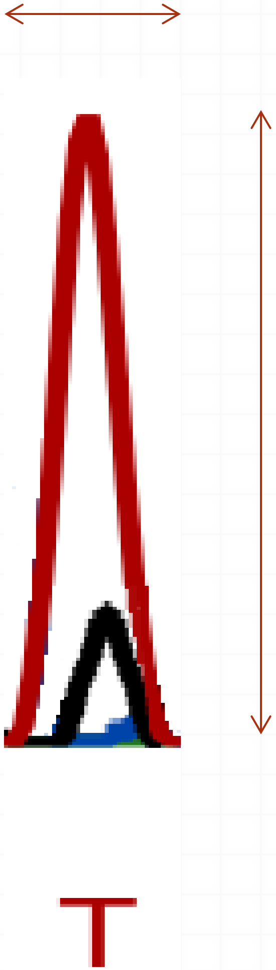
- A rather nice section of sequencing readout



Principles of this **benchmarking process**:

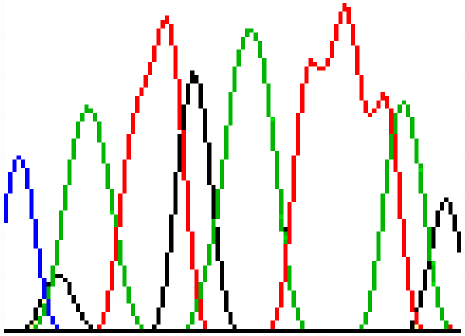
- This nice, sharp peak in the T lane has measurable properties
- So too do the readouts in the A, C, G lanes at the same point (they are all flat)
- The base call is obvious
- Using the **known reference sequence**, all of the peaks with identical properties can be evaluated
- How often does a peak with these properties identify the correct base in the reference?
- How often is it wrong?

- **Probabilities of error for each base call can therefore be calculated**
  - (strictly speaking, estimated)

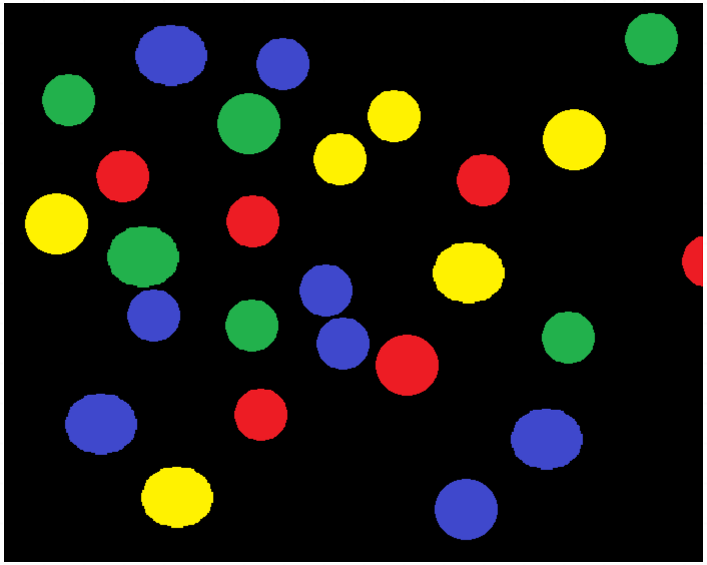


- What about this situation? A peak in two lanes simultaneously, albeit one is much smaller
- How often does a peak pair with these properties identify the correct base?
- How often is it wrong?

- What about these peaks:



- Similar principles can be applied to completely different sequencing platforms
- E.g. Illumina
- Again, it's a case of association readout characteristics with probabilities of error (error rates)



When is benchmarking done?

- In principle, it could be done as a one-off exercise
- In practice, it may make more sense for this to be done frequently
- i.e. as part of the **calibration** process prior to any sequencing run
- This is common in an Illumina sequencing operation
  - And uses sequences from the same ΦX genome
- So, need never concern you (unless you are actually operating the sequencing machine)

Using the benchmark error rates

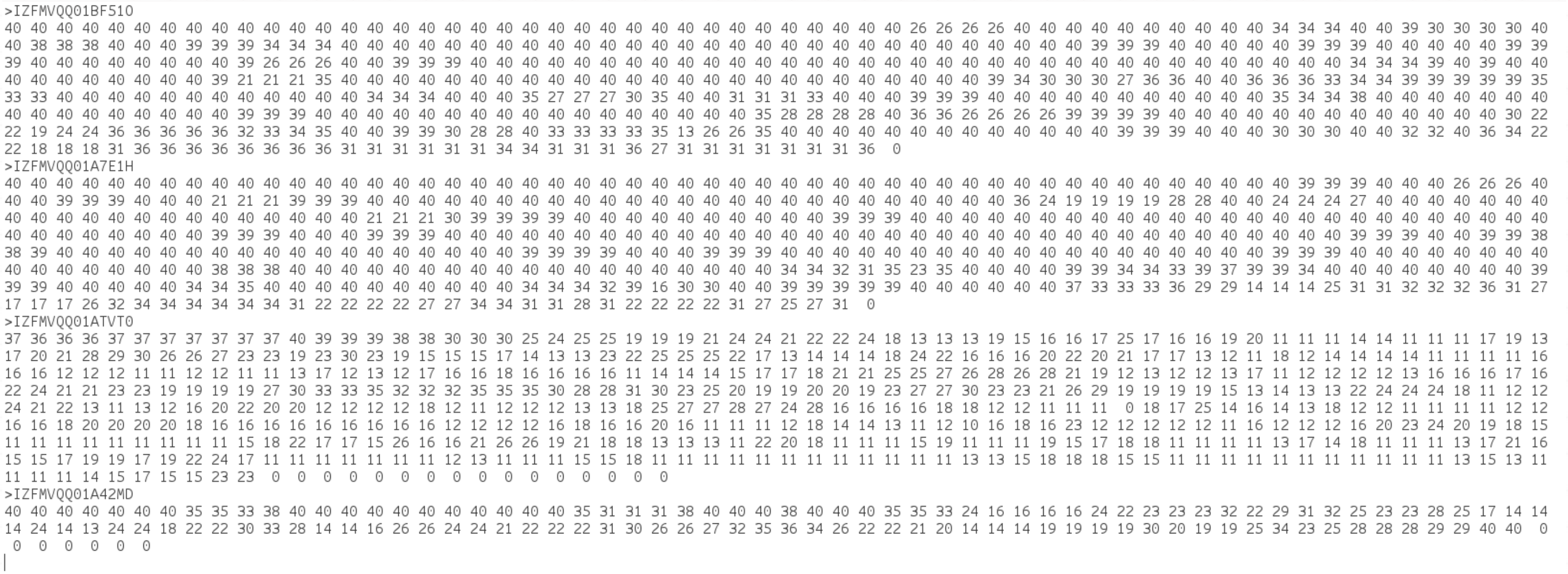
- Once the error rates are known, these can then be applied to new output, i.e. of new, unknown sequences, base-call by base-call
  - I.e. **one error probability per base**
- The software which outputs the data from the sequencer does all this for you
- So again, no need to worry about **how** it's done
- But important to know what these probabilities **mean –** and how they are **expressed**

Phred scores

- Usually, the error probability (***p)*** of each base call is expressed as a **Phred score**
- This value is simply: $-10 \times \log_{10} p$
  - **rounded to an integer**
- E.g. 1 in a thousand probability of being wrong
  - $p = 0.001$ (generally acceptable) → phred score = 30
  - 1 in 10 (awful) : $p = 0.1$ → phred score = 10
  - 1 in 4 (disastrous) : $p = 0.25$ → phred score = 6
- Picking a base completely at random, i.e. 3 / 4 chance that it is wrong:
  - 3 in 4: $p = 0.75$ → phred score = 1

Presenting the phred score for each base call

- QUAL file format: a PLAIN TEXT ('ASCII') file; each sequence has a header, then one digit per base



```
>IZFMVQQ01BF510
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 26 26 26 26 40 40 40 40 40 40 40 40 40 40 34 34 34 40 40 39 30 30 30 30 40
40 38 38 38 40 40 40 39 39 39 34 34 34 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 39 39 39 40 40 40 40 40 39 39
39 40 40 40 40 40 40 40 40 39 26 26 26 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 34 34 34 39 40 39 40 40
40 40 40 40 40 40 40 40 39 21 21 21 35 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 34 30 30 30 27 36 36 40 40 36 36 36 33 34 34 39 39 39 39 39 35
33 33 40 40 40 40 40 40 40 40 40 40 40 40 34 34 34 40 40 40 35 27 27 27 30 35 40 40 31 31 31 33 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 35 34 34 38 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 35 28 28 28 28 40 36 36 26 26 26 26 39 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 30 22
22 19 24 24 36 36 36 36 36 32 33 34 35 40 40 39 39 30 28 28 40 33 33 33 33 35 13 26 26 35 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 30 30 30 40 40 32 32 40 36 34 22
22 18 18 18 31 36 36 36 36 36 36 36 36 31 31 31 31 31 31 34 34 31 31 31 36 27 31 31 31 31 31 31 31 36  0
>IZFMVQQ01A7E1H
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 26 26 26 40
40 40 39 39 39 40 40 40 21 21 21 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 36 24 19 19 19 19 28 28 40 40 24 24 24 27 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 21 21 21 30 39 39 39 39 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 39 39 39 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 39 39 38
38 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 39 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 38 38 38 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 34 34 32 31 35 23 35 40 40 40 40 39 39 34 34 33 39 37 39 39 34 40 40 40 40 40 40 40 40 39
39 39 40 40 40 40 40 34 34 35 40 40 40 40 40 40 40 40 40 40 34 34 34 32 39 16 30 30 40 40 39 39 39 39 39 40 40 40 40 40 40 37 33 33 33 36 29 29 14 14 14 25 31 31 32 32 32 36 31 27
17 17 17 26 32 34 34 34 34 34 34 31 22 22 22 22 27 27 34 34 31 31 28 31 22 22 22 22 31 27 25 27 31  0
>IZFMVQQ01ATVT0
37 36 36 36 37 37 37 37 37 37 37 40 39 39 39 38 38 30 30 30 25 24 25 25 19 19 19 21 24 24 21 22 22 24 18 13 13 13 19 15 16 16 17 25 17 16 16 19 20 11 11 11 14 14 11 11 11 17 19 13
17 20 21 28 29 30 26 26 27 23 23 19 23 30 23 19 15 15 15 17 14 13 13 23 22 25 25 25 22 17 13 14 14 14 18 24 22 16 16 16 20 22 20 21 17 17 13 12 11 18 12 14 14 14 14 11 11 11 11 16
16 16 12 12 12 11 11 12 12 11 11 13 17 12 13 12 17 16 16 18 16 16 16 16 11 14 14 14 15 17 17 18 21 21 25 25 27 26 28 26 28 21 19 12 13 12 12 13 17 11 12 12 12 12 13 16 16 16 17 16
22 24 21 21 23 23 19 19 19 19 27 30 33 33 35 32 32 32 35 35 35 30 28 28 31 30 23 25 20 19 19 20 20 19 23 27 27 30 23 23 21 26 29 19 19 19 19 15 13 14 13 13 22 24 24 24 18 11 12 12
24 21 22 13 11 13 12 16 20 22 20 20 12 12 12 12 18 12 11 12 12 12 13 13 18 25 27 27 28 27 24 28 16 16 16 16 18 18 12 12 11 11 11  0 18 17 25 14 16 14 13 18 12 12 11 11 11 11 12 12
16 16 18 20 20 20 20 18 16 16 16 16 16 16 16 16 16 12 12 12 12 16 18 16 16 20 16 11 11 11 12 18 14 14 13 11 12 10 16 18 16 23 12 12 12 12 12 11 16 12 12 12 16 20 23 24 20 19 18 15
11 11 11 11 11 11 11 11 11 15 18 22 17 17 15 26 16 16 21 26 26 19 21 18 18 13 13 13 11 22 20 18 11 11 11 15 19 11 11 11 19 15 17 18 18 11 11 11 11 13 17 14 18 11 11 11 13 17 21 16
15 15 17 19 19 17 19 22 24 17 11 11 11 11 11 11 11 12 13 11 11 11 15 15 18 11 11 11 11 11 11 11 11 11 11 11 11 13 13 15 18 18 18 15 15 11 11 11 11 11 11 11 11 11 11 11 13 15 13 11
11 11 11 14 15 17 15 15 23 23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
>IZFMVQQ01A42MD
40 40 40 40 40 40 40 35 35 33 38 40 40 40 40 40 40 40 40 40 40 40 35 31 31 31 38 40 40 40 38 40 40 40 35 35 33 24 16 16 16 16 24 22 23 23 23 32 22 29 31 32 25 23 23 28 25 17 14 14
14 24 14 13 24 24 18 22 22 30 33 28 14 14 16 26 26 24 24 21 22 22 22 31 30 26 26 27 32 35 36 34 26 22 22 21 20 14 14 14 19 19 19 19 30 20 19 19 25 34 23 25 28 28 28 29 29 40 40  0
 0  0  0  0  0  0
|
```

Presenting the phred score for each base call

- QUAL file format: header, then one digit per base

```
>IZFMVQQ01BF51O
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 38 38 38 40 40 40 39 39 39 34 34 34 40 40 40 40 40 40 40 40 40
39 40 40 40 40 40 40 40 40 39 26 26 26 40 40 39 39 39 40 40 40 40
40 40 40 40 40 40 40 40 39 21 21 21 35 40 40 40 40 40 40 40 40 40
33 33 40 40 40 40 40 40 40 40 40 40 40 40 34 34 34 40 40 40 35 27
40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40
22 19 24 24 36 36 36 36 36 32 33 34 35 40 40 39 39 30 28 28 40 33
22 18 18 18 31 36 36 36 36 36 36 36 36 31 31 31 31 31 31 34 34 31
>IZFMVQQ01A7E1H
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
40 40 39 39 39 40 40 40 21 21 21 39 39 39 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 21 21 21 30 39 39 39 39
40 40 40 40 40 40 40 40 39 39 39 40 40 40 39 39 39 40 40 40 40 40
38 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39
40 40 40 40 40 40 40 40 38 38 38 40 40 40 40 40 40 40 40 40 40 40
39 39 40 40 40 40 40 34 34 35 40 40 40 40 40 40 40 40 40 40 34 34
17 17 17 26 32 34 34 34 34 34 34 31 22 22 22 22 27 27 34 34 31 31
```

QUAL and FASTA files

- Each QUAL file is used in conjunction with a FASTA file, which contains the corresponding base calls
  - I.e. the actual sequence
- The FASTA and QUAL files have corresponding headers

E.g. two sequence reads

```
>IZFMVQQ01BF51O
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
40 38 38 38 40 40 40 39 39 39 34 34 34 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
39 40 40 40 40 40 40 40 40 39 26 26 26 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 4
40 40 40 40 40 40 40 40 39 21 21 21 35 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
33 33 40 40 40 40 40 40 40 40 40 40 40 40 34 34 34 40 40 40 35 27 27 27 30 35 40 40 31 31 31 3
40 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 35 28 2
22 19 24 24 36 36 36 36 36 32 33 34 35 40 40 39 39 30 28 28 40 33 33 33 33 35 13 26 26 35 40 4
22 18 18 18 31 36 36 36 36 36 36 36 36 31 31 31 31 31 31 34 34 31 31 31 36 27 31 31 31 31 31 3
>IZFMVQQ01A7E1H
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
40 40 39 39 39 40 40 40 21 21 21 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
40 40 40 40 40 40 40 40 40 40 40 40 40 40 21 21 21 30 39 39 39 39 40 40 40 40 40 40 40 40 40 4
40 40 40 40 40 40 40 40 39 39 39 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 4
38 39 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39
```

QUAL

```
>IZFMVQQ01BF51O
TCTCTATGCGGTGTCAGCCGCCGCGGTAATACGTAGGGGCAAGCGTTATCCCGGATTTAC
TGGGTGTAAAGGGAGCGTAGACGGCAGCGCAAGTCTGAAGTGAAATGCCAGGGCTTAACC
CTGGAACTGCTTTGGAAACTGTGCAGCTAGAGTGCAGGAGAGGTAAGTGGAATTTCTAGT
GTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCAGTGGCGGAGGCGGCTTACTGGAC
GGTAACTGACGCTGAGGCTCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGT
CCACGCCGTAAACGATGAATACTAGGTACAGGGGCACAAAAGTGCTTCTGTGCCGCAGCT
AACGCAATAAGTATTCCACCTGGGGAGTACGTTCGCAAGAATGAAACTCAAAGGAATTGA
CGGGCTGAGACTGCCAAGGCACACAGGGGATAGGN
>IZFMVQQ01A7E1H
CGTGTCTCTAGTGCCAGCCGCCGCGGTAATACGTAGGTGGCAAGCGTTATCCCGGATTTA
CTGGGTGTAAAGGGCGTGTAGGCGGACGCTTAAGTCAGCGGTAAATTGCGGGGCTCAACC
TCGTCGAGCCGTTGAAACTGGGTGCCTTGAGTGGGCGAGAAGTACGCGGAATGCGTGGTG
TAGCGGTGAAATGCATAGATATCACGCAGAACTCCGATTGCGAAGGCAGCGTACCGGCGC
CCAACTGACGCTGAAGCACGAAGGCGTGGGTATCGAACAGGATTAGATACCCTGGTAGTC
CACGCAGTAAACGATGAATGCTAGTTGTCCGGGGCGATTGAGTTCTGGGTGACACAGCGA
AAGCGTTAAGCATTCCACCTGGGGAGTACGCCGGCAACGGTGAAACTTAAATGAATTGAC
GGGCTGAGACTGCCAAGGCACACAGGGGATAGGN
```

FASTA

FASTQ format

- A more commonly-used alternative to QUAL+FASTA
- Also a plain-text format
- Stores both the sequences and the quality (phred) scores in the same file
- Stores the quality scores in a more compact format than QUAL format
- Each possible Quality score is represented by a **single character** (letter, digit or symbol)
  - Thus, **encoding** of quality scores
  - As is sometimes the case with bioinformatics formats, things are not as simple as they might be:
  - There have been **different** encoding schemes employed

FASTQ format, and Example

- 4 lines per read:
  1. *@sequenceID additional-data*
  2. *sequence (base calls) – one letter per base (of course…)*
  3. *+sequenceID additional-data*
  4. *encoded quality scores – one letter per base*

```
@HWI-M01242:112:000000000-AM193:1:1101:15596:1678 1:N:0:ACGCTACTGGATATCT
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGACGCTTAAGTCAG
+HWI-M01242:112:000000000-AM193:1:1101:15596:1678 1:N:0:ACGCTACTGGATATCT
BABBBFFDBFFFGGFGGGGAGGHFEAEBEFGHGFHGGGFDGGHGHGGGGGGGHGGEGGGGGGGGHHHH
```

Note that lines 1 and 3 are necessarily identical (apart from the first character)

Example FASTQ

```
@HWI-M01242:112:000000000-AM193:1:1101:15596:1678 1:N:0:ACGCTACTGGATATCT
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGACGCTTAAGTCAGCGGTAAAATTGCGGGGCTCAACCTCG
+HWI-M01242:112:000000000-AM193:1:1101:15596:1678 1:N:0:ACGCTACTGGATATCT
BABBBFFDBFFFGGFGGGGAGGHFEAEBEFGHGFHGGGFDGGHGHGGGGGGGHGGEGGGGGGGGHHHHGGCEFFFGGGHFGGGGGGGGFCGHHG
@HWI-M01242:112:000000000-AM193:1:1101:16562:1707 1:N:0:ACGCTACTGGATATCT
TACGGAGGATACGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGAGTGTCAAGTCAGCGGTAAAATTTCGGGGCTCAACCCCG
+HWI-M01242:112:000000000-AM193:1:1101:16562:1707 1:N:0:ACGCTACTGGATATCT
BBBBBBBBAFFFFGGGGGGEGGHGGAEGHGGHHHHEGGHHHHHGGGGGFGGGHGGFGGHHHHHHHHHHHGGGGGHHGHGHFGGGFGEGHHHEGG
(….etc)
```

- (2 reads are shown above)

Example FASTQ

- In practice, the sequenceID and additional data is often omitted from line 3 to save space, with the 4th line being assumed to be the Q-scores of the most recent sequence
- (same 2 reads are shown below, with implicit headers prior to the Q-scores)

```
@HWI-M01242:112:000000000-AM193:1:1101:15596:1678 1:N:0:ACGCTACTGGATATCT
TACGGAGGATGCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGACGCTTAAGTCAGCGGTAAAATTGCGGGGCTCAACCTCG
+
BABBBFFDBFFFGGFGGGGAGGHFEAEBEFGHGFHGGGFDGGHGHGGGGGGGHGGEGGGGGGGGHHHHGGCEFFFGGGHFGGGGGGGGFCGHHG
@HWI-M01242:112:000000000-AM193:1:1101:16562:1707 1:N:0:ACGCTACTGGATATCT
TACGGAGGATACGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGAGTGTCAAGTCAGCGGTAAAATTTCGGGGCTCAACCCCG
+
BBBBBBBBAFFFFGGGGGGEGGHGGAEGHGGHHHHEGGHHHHHGGGGGFGGGHGGFGGHHHHHHHHHHHGGGGGHHGHGHFGGGFGEGHHHEGG
```

Encoding quality-scores in FASTQ

- `BABBBFFDBFFFGGFGGGAGGHFEAEBEFGHGFHGGGFDGGHG...`
- So what do those quality-score lines mean?
  - Depending on the coding scheme, quality scores of up to **93** can be stored
  - In a read output by a sequencer, even the best scores you see in practice will be **far lower**
  - E.g. best scores of around **40** ('I' in the current Illumina format)
- For more details on encoding of quality scores and their relation to **ASCII codes**, refer to:
  - MF's slides ("*It's all about the text*") of 19th Oct: http://ghfs1.ifr.ac.uk/ghfs/wp-content/uploads/2016/10/Bitesize_ngs_formats.pdf
  - JW's background ("*A brief guide to how computers encode data*"): http://ghfs1.ifr.ac.uk/ghfs/wp-content/uploads/2016/11/slides_file_storage_1.pdf

Format conversion

- Sequencing providers give Illumina-format data in the form of FASTQ
- Many tools will understand FASTQ
- So conversion may not be necessary
- Some conversion tools can convert between FASTA+QUAL and FASTQ
  - E.g. PRINSEQ (prinseq-lite.pl)
- EMBOSS seqret (a very versatile program)
  - FASTQ→FASTA (loses quality scores)

Other formats you may have come across

- SFF : Standard Flowgram File
  - The native format of the 454 sequencing platform
  - Comes in binary and plain-text varieties
  - Can be converted to other formats such as FASTA/QUAL with various tools, e.g. MOTHUR sffinfo
- BCL: earlier Illumina Base Call format
  - With current Illumina software, BCL files will have been converted to FASTQ before you are likely to have seen the data
- FAST5: used with Oxford Nanopore sequencers
  - Based on HDF5 (Hierarchical Data Format – a generic and versatile data format definition)