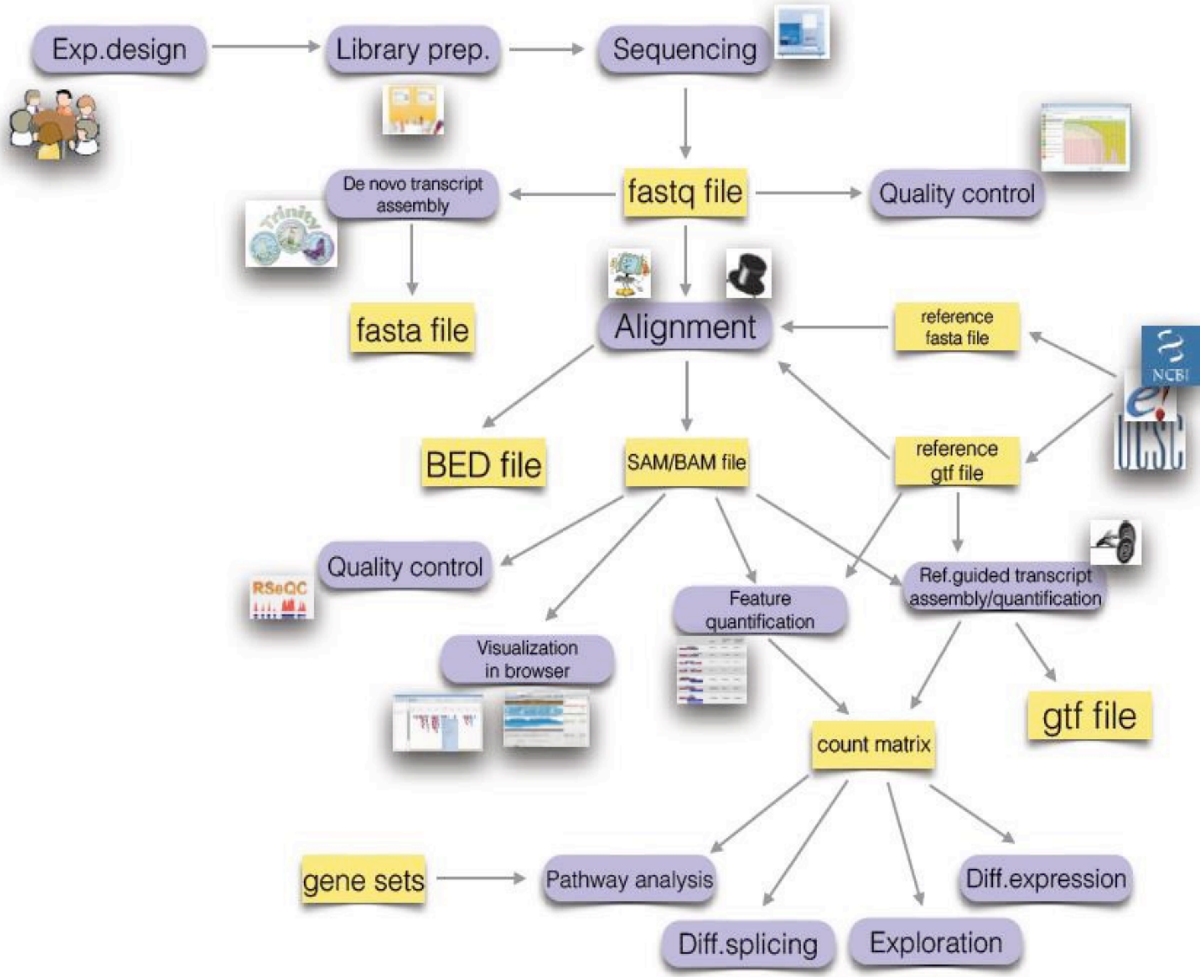
Bi5444 Analysis of sequencing data

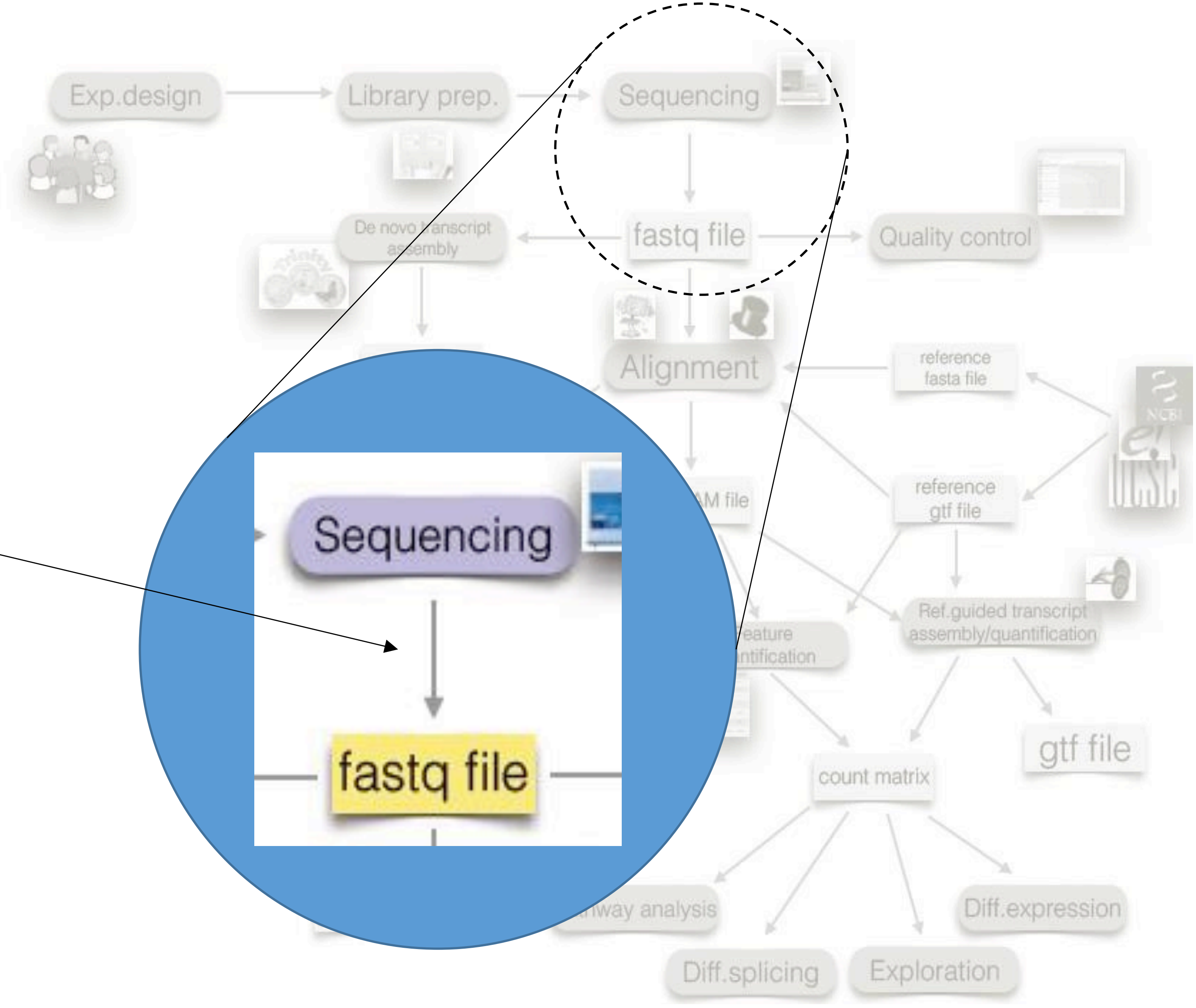
NGS quality control

Eva Budinska

The NGS analysis pipeline

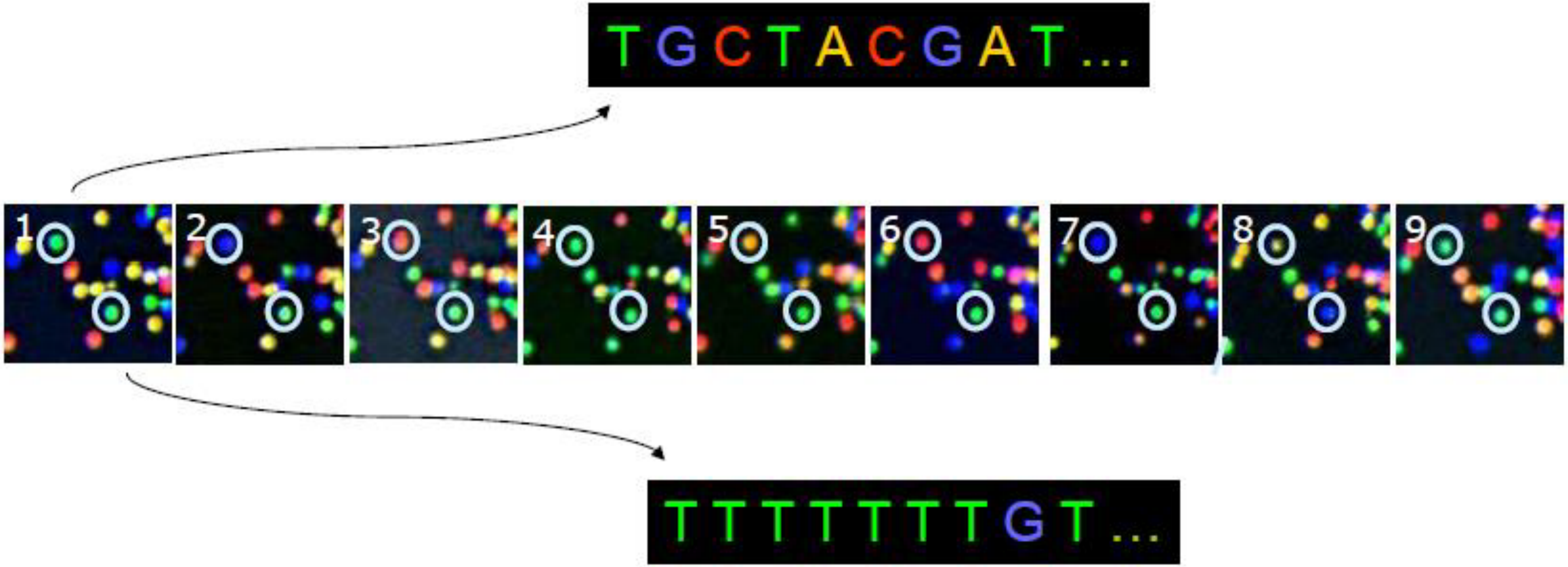


Step 0: base calling (image analysis) + base quality control



Step 0: base calling (image analysis)

- The identity of each base of a cluster is read off from **sequential images**
- One cycle -> one image



The PHRED score

$$Q_{phred} = - 10 \times \log_{10}P(\mathbf{error})$$

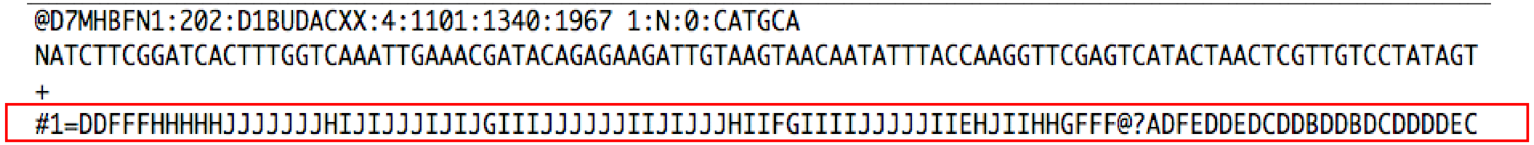
- The *Phred* quality score is the negative ratio of the error probability to the reference level of *P* = 1 expressed in Decibel (dB).
- The **error estimate** is based on **statistical model** providing measure of **certainty** of each base call in addition to the nucleotide itself

Phred score	Probability of incorrect base call	Base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10 000	99.99%
50	1 in 100 000	99.999%
60	1 in 1 000 000	99.9999%

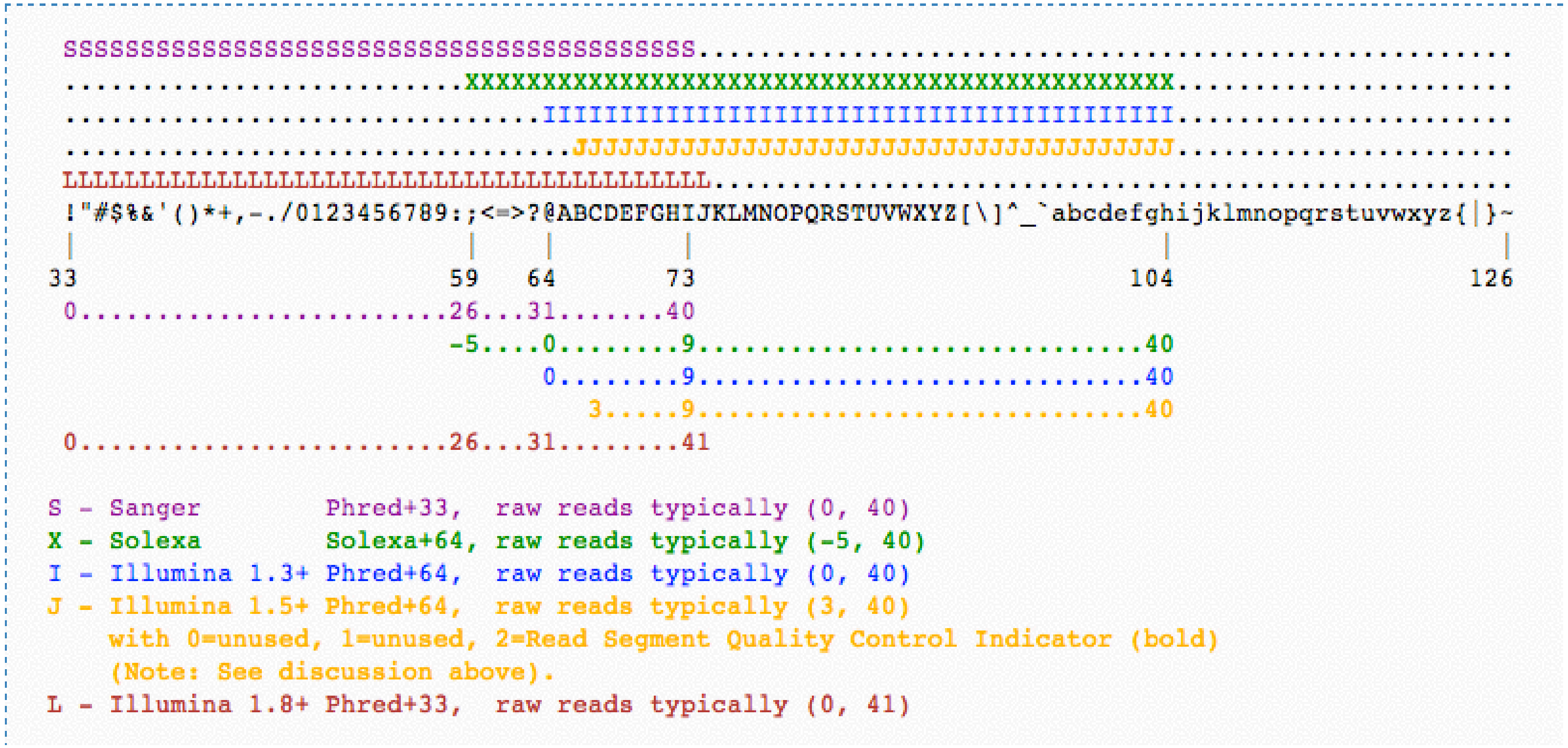
- These statistical models base their error estimate on:
  - Signal intensities from the recorded image
  - Number of the sequencing cycle
  - Distance to other sequence colonies
- *Phred* score is recoded using ASCII in fastq file

Phred score encoding in ASCII

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



```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHHJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

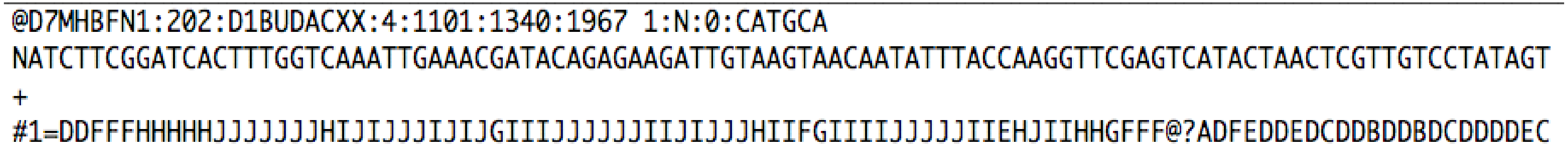


```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0........................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)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

FASTQ format

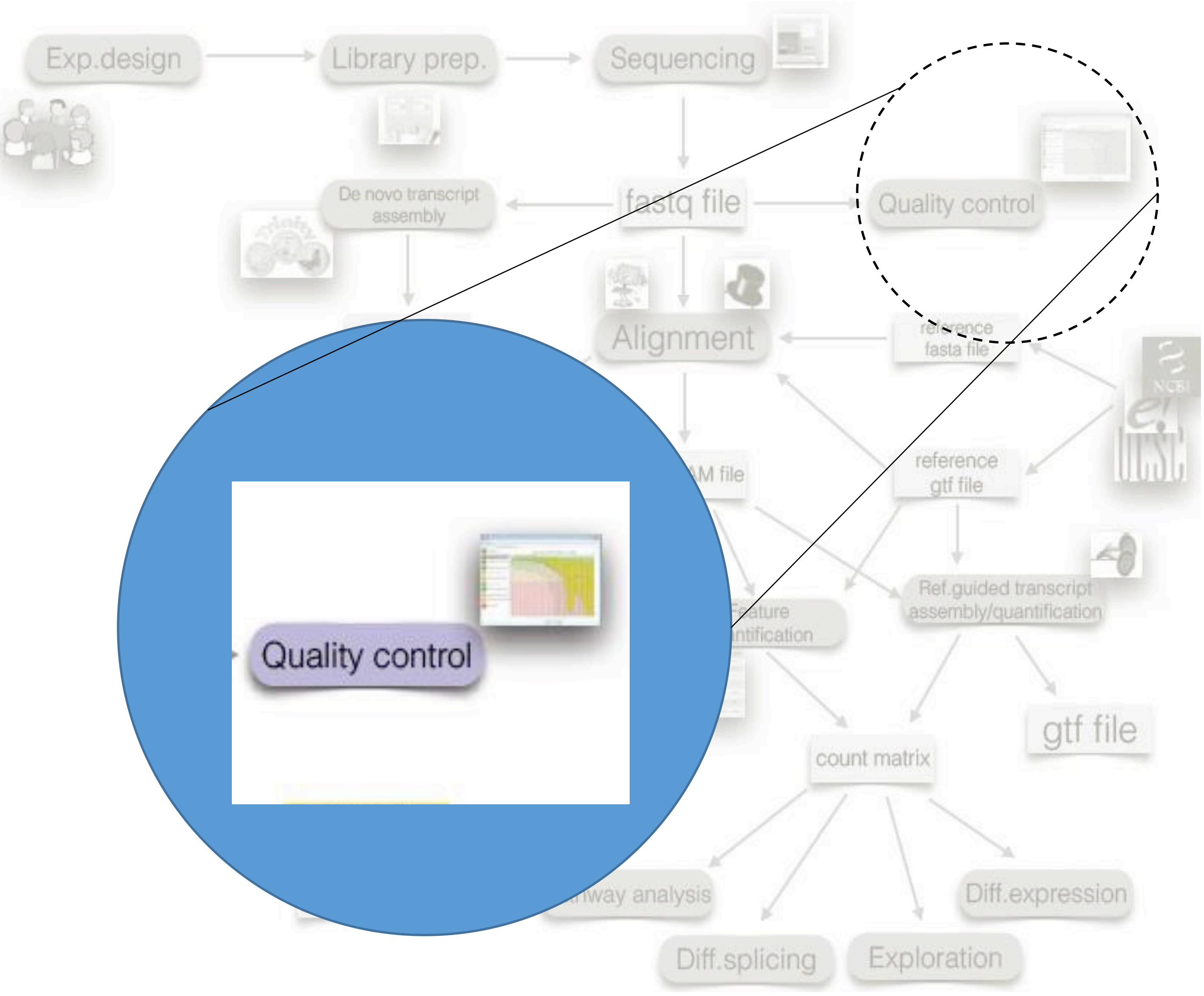
- Combines sequence and base call quality information.
- Typical file extension: `.fastq`



```
@D7MHBFN1:202:D1BUDACXX:4:1101:1340:1967 1:N:0:CATGCA
NATCTTCGGATCACTTTGGTCAAATTGAAACGATACAGAGAAGATTGTAAGTAACAATATTTACCAAGGTTCGAGTCATACTAACTCGTTGTCCTATAGT
+
#1=DDFFFHHHHJJJJJJHIJIJJJIJIJGIIIJJJJJJIIJIJJJHIIFGIIIIJJJJJIIEHJIIHHGFFF@?ADFEDDEDCDDBDDBDCDDDDEC
```

- Four lines per sequence (read):
  - ID (starting with @)
  - Sequence line
  - Another ID line (starting with +)
  - Base qualities (one for each letter in the sequence)

Step 1:
Read quality control and data filtering



Before we dive in…

… let’s review few concepts and expressions

The steps of Illumina sequencing

1. Fragment genomic DNA, e.g. with a sonicator.

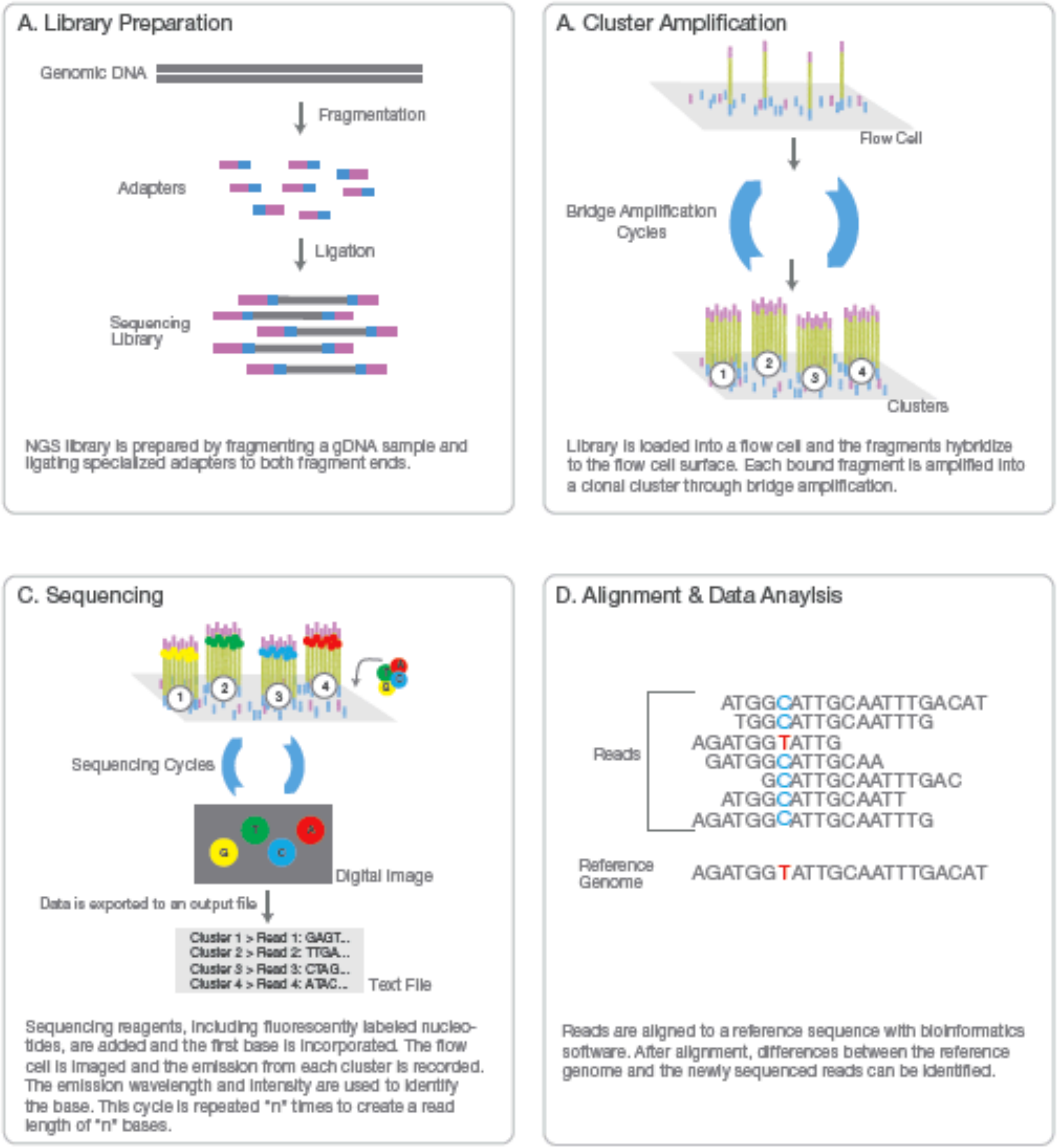
2. Ligate adapters to both ends of the fragments.

3. PCR amplify the fragments with adapters

4. Spread DNA molecules across flowcells. Goal is to get exactly **one DNA molecule** per flowcell lawn of primers. This depends purely on probability, based on the concentration of DNA.

5. Use bridge PCR to amplify the single molecule on each lawn so that you can get a strong enough signal to detect. Usually this requires several hundred or low thousands of molecules.

6. Sequence by synthesis of complementary strand: [reversible terminator chemistry](#).



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

Library is loaded into a flow cell and the fragments hybridize to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

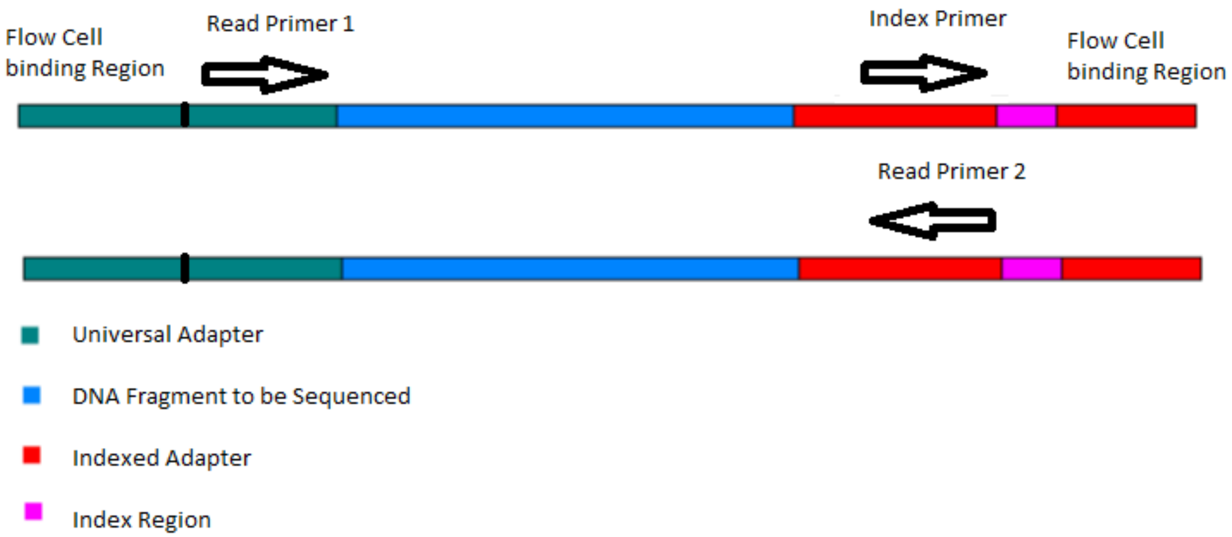
Cluster 1 > Read 1: GAGT..
Cluster 2 > Read 2: TTGA..
Cluster 3 > Read 3: CTAG..
Cluster 4 > Read 4: ATAC..

Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

Sources of errors: adapters

- In step 2, adapters are ligated to the end of the fragments



Sequencing random fragments of DNA is possible via the addition of short nucleotide sequences which allow any DNA fragment to:

● Bind to a flow cell for next generation sequencing

● Allow for PCR enrichment of adapter ligated DNA fragments only

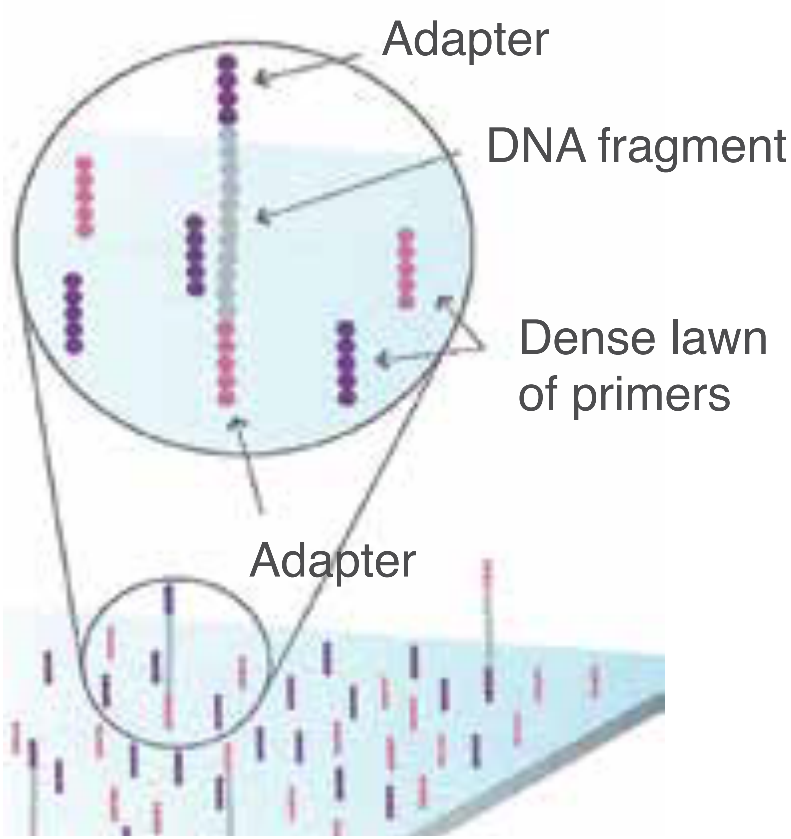
● Allow for indexing or 'barcoding' of samples so multiple DNA libraries can be mixed together into 1 sequencing lane (known as multiplexing)

From:
http://tucf-genomics.tufts.edu/documents/protocols/TUCF_Understanding_Illumina_TruSeq_Adapters.pdf

Sources of errors: PCR duplicates

- In step 3 we are *intentionally* creating multiple copies of each original genomic DNA molecule so that we have enough of them.
- PCR duplicates occur when **two copies of the same original molecule get onto different primer lawns in a flowcell**.
- In consequence we read the very same sequence twice!

Higher rates of PCR duplicates e.g. 30% arise when you have too little starting material such that greater amplification of the library is needed in step 3, or when you have too great a variance in fragment size, such that smaller fragments, which are easier to PCR amplify, end up over-represented.

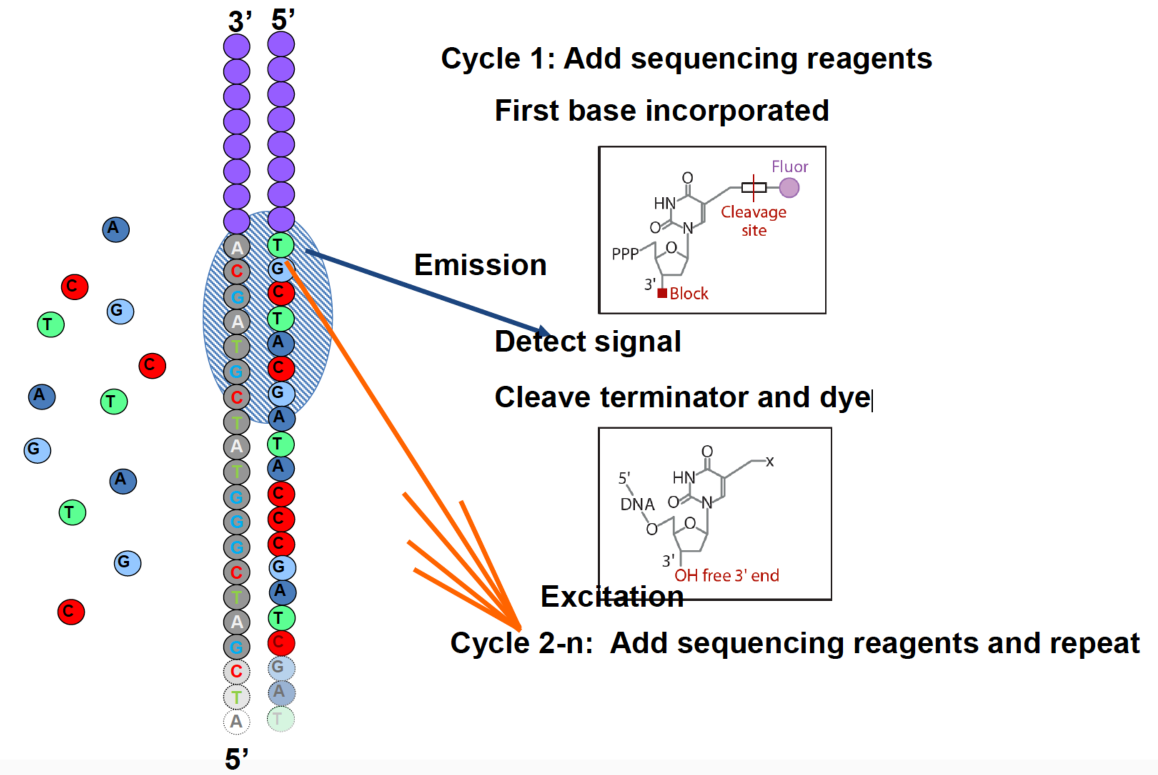


Find beautiful explanation of probabilities and much more at: https://www.cureffi.org/2012/12/11/how-pcr-duplicates-arise-in-next-generation-sequencing/

Sources of errors: sequencing by synthesis – the fluorescence

- In step 5 we amplify the signal and detect the fluorescence of each base
- The assumption is that in a cycle, every molecule on the flowcell is extended by one base

- The reality:
  - Some molecules are not extended or their base has no fluorescent dye
  - The previous fluorescent dye is not cleaved – the signal from the cluster after a few cycles is a mix of signals from previous bases



Cycle 1: Add sequencing reagents

First base incorporated

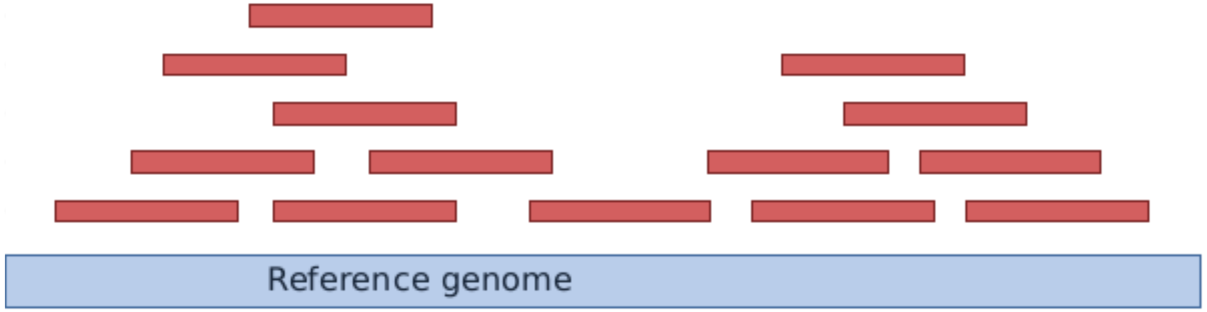
Detect signal

Cleave terminator and dye

Cycle 2-n: Add sequencing reagents and repeat

Sequencing coverage

Coverage in DNA sequencing **is the number of unique reads that include a given nucleotide** in the reconstructed sequence.



www.metagenomics.wiki

Depth of coverage
(coverage depth / mapping depth)

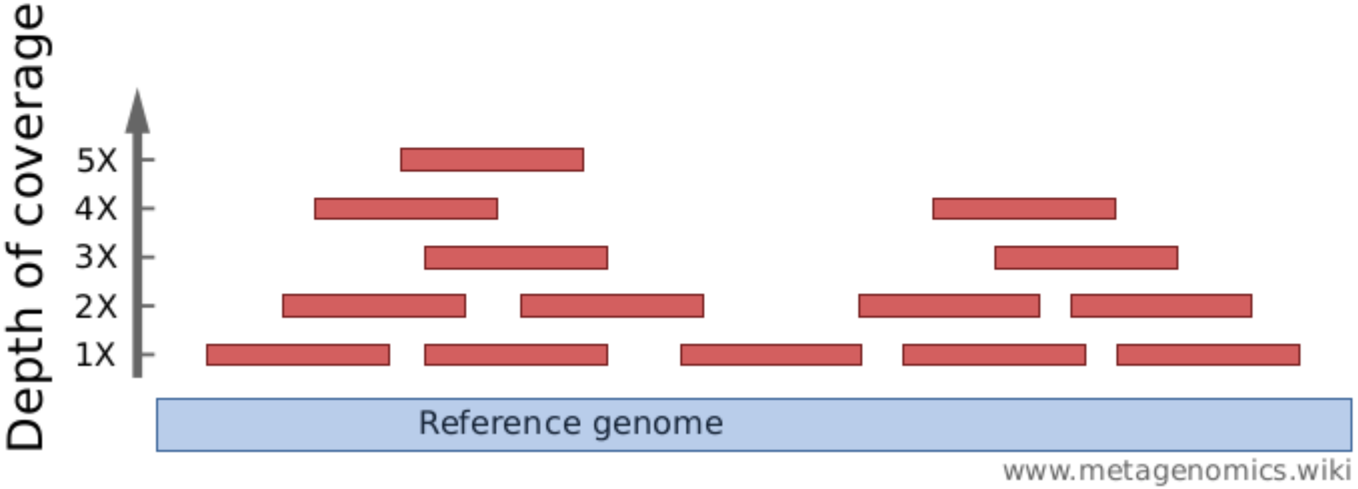
How strongly is the genome "covered" by sequenced fragments (short reads)?

Per-base coverage is the average number of times a base of a genome is sequenced (in other words, how many reads cover it).

Average coverage of the genome (Av)

Av = (NxL)/G

G - length of the original genome
N - number of reads
L - average read length

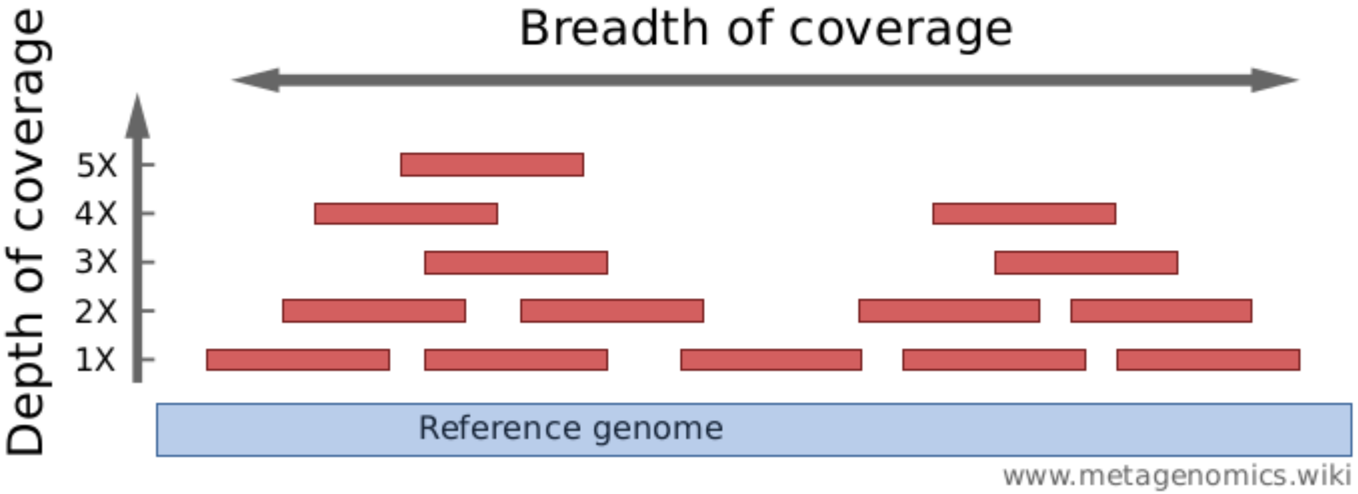


www.metagenomics.wiki

The coverage depth of a genome is calculated as **the number of bases of all short reads that match a genome divided by the length of this genome**. It is often expressed as 1X, 2X, 3X,... (1, 2, or, 3 times coverage).

Breadth of coverage
(covered length)

What proportion of the genome is "covered" by short reads? Are there regions that are not covered, even not by a single read?



<u>Breadth of coverage</u> is the percentage of bases of a reference genome that are covered with a certain depth.
For example: "90% of a genome is covered at 1X depth; and still 70% is covered at 5X depth."

Sequencing **coverage**

- **Deep sequencing** refers to the general concept of aiming for high number of unique reads of each region of a sequence.

Step 1: **Read quality control** and data filtering

- Uses the output file with information about the quality of base calls (.fastq)
- First step in the pipeline that **deals with actual sequencing data** in base or color space

- Several metrics are evaluated, not all of them use the Phred score information:
  - Distribution of quality scores at each sequence, Sequence composition, Per-sequence and per-read distribution of GC content, Library complexity, Overrepresented sequences

- Initial overview – already in base calling SW
- More quality overview – SW solutions `SolexaQA, FastQC`

Main quality control points

1. **Base quality**
2. **Sequence composition** – sequence content across bases should not change with cycle (exception are targeted sequencing SNP experiments)
3. **Per-sequence and per-read distribution of GC content** (shift from expected can indicate contamination by rRNA for instance)
4. **Library complexity** (too many duplicates?)
5. **Overrepresented sequences** –may represent highly expressed genes, or presence of adapters or rRNA contamination or PCR duplicates

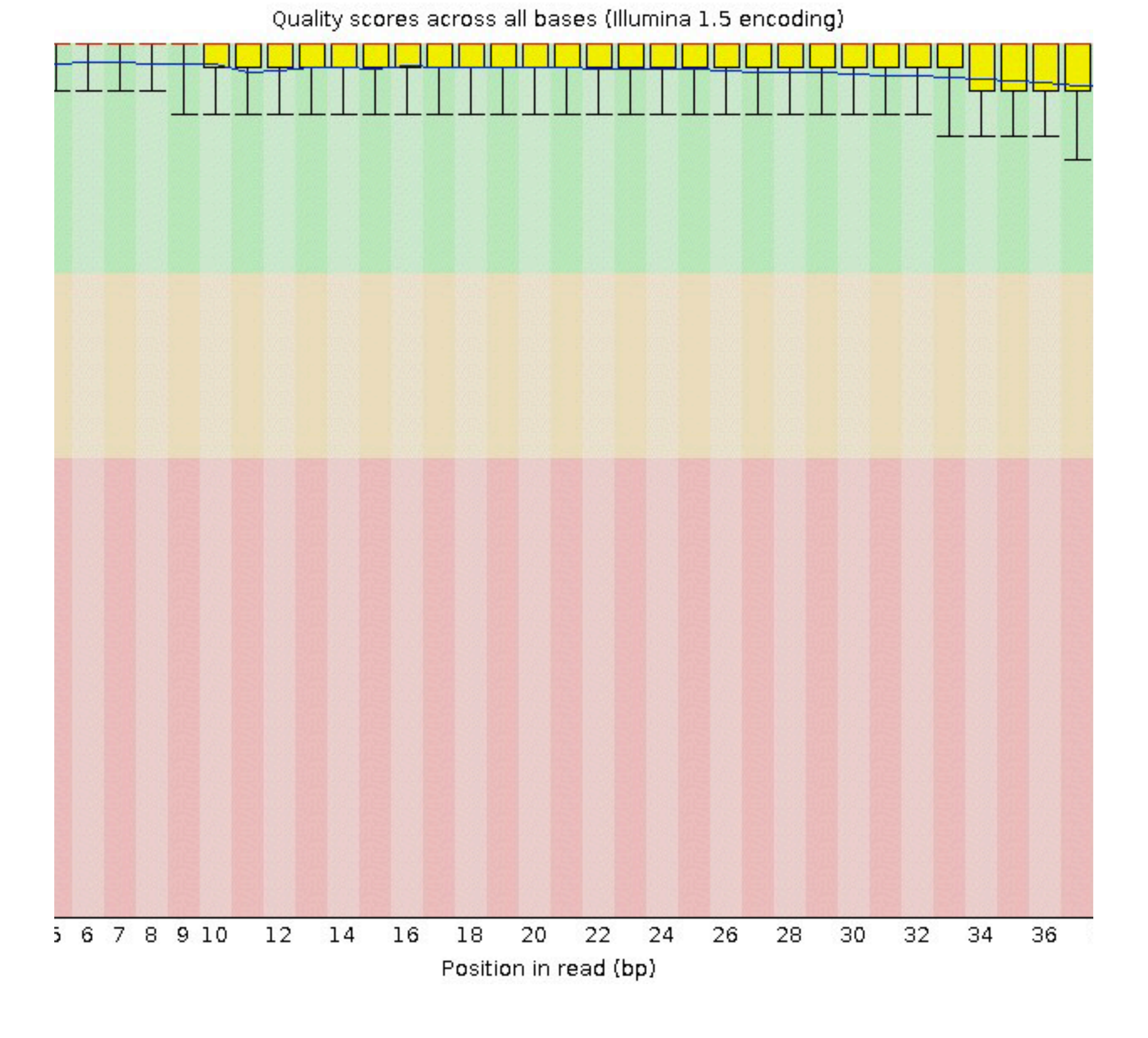
Base quality

- Quality of bases (Phred score) should be good across all cycles

(all the sequence)

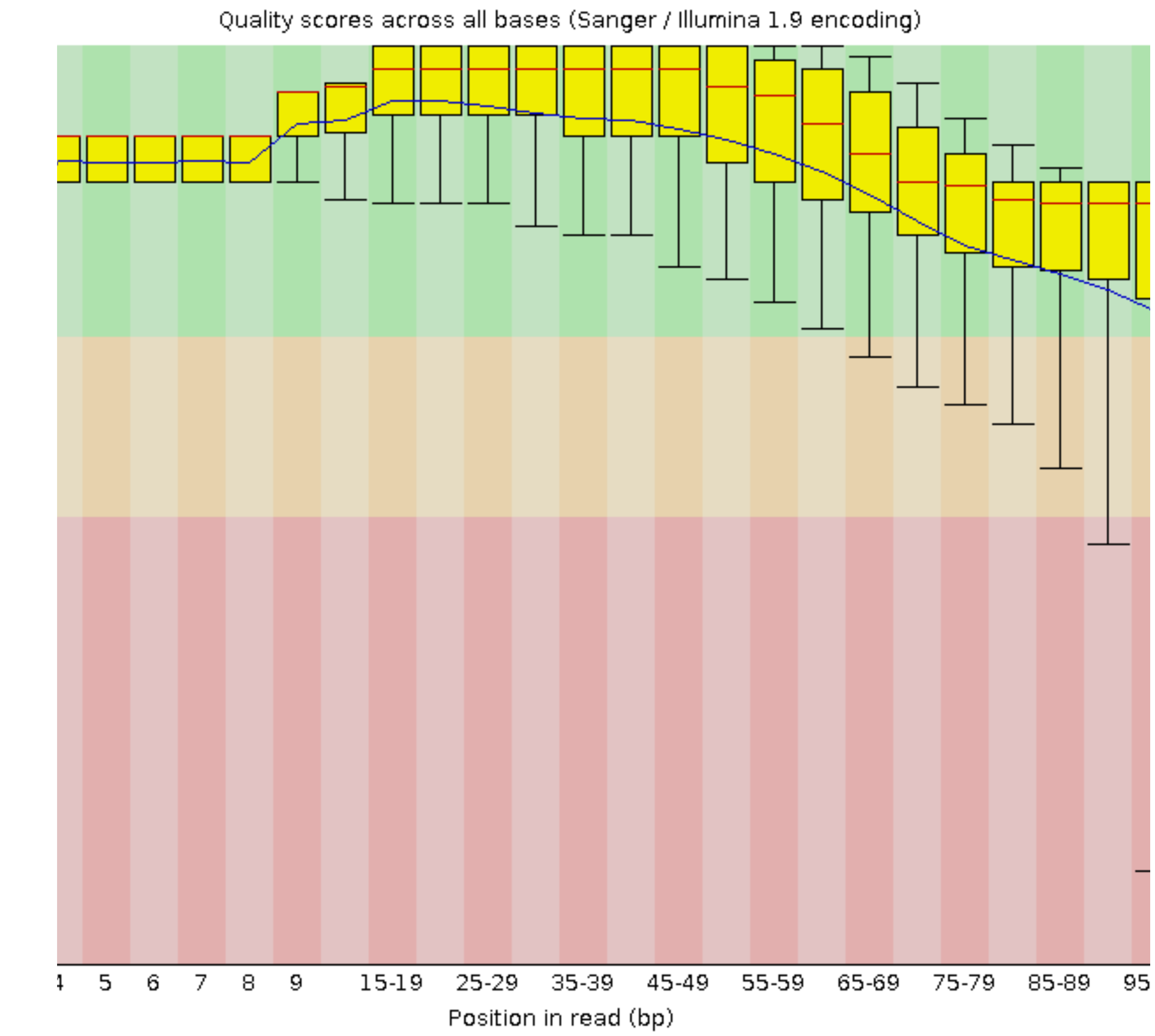
Base quality – an excellent example

- Shows distribution (boxplot) of quality of bases (Phred scores) across all reads in each cycle



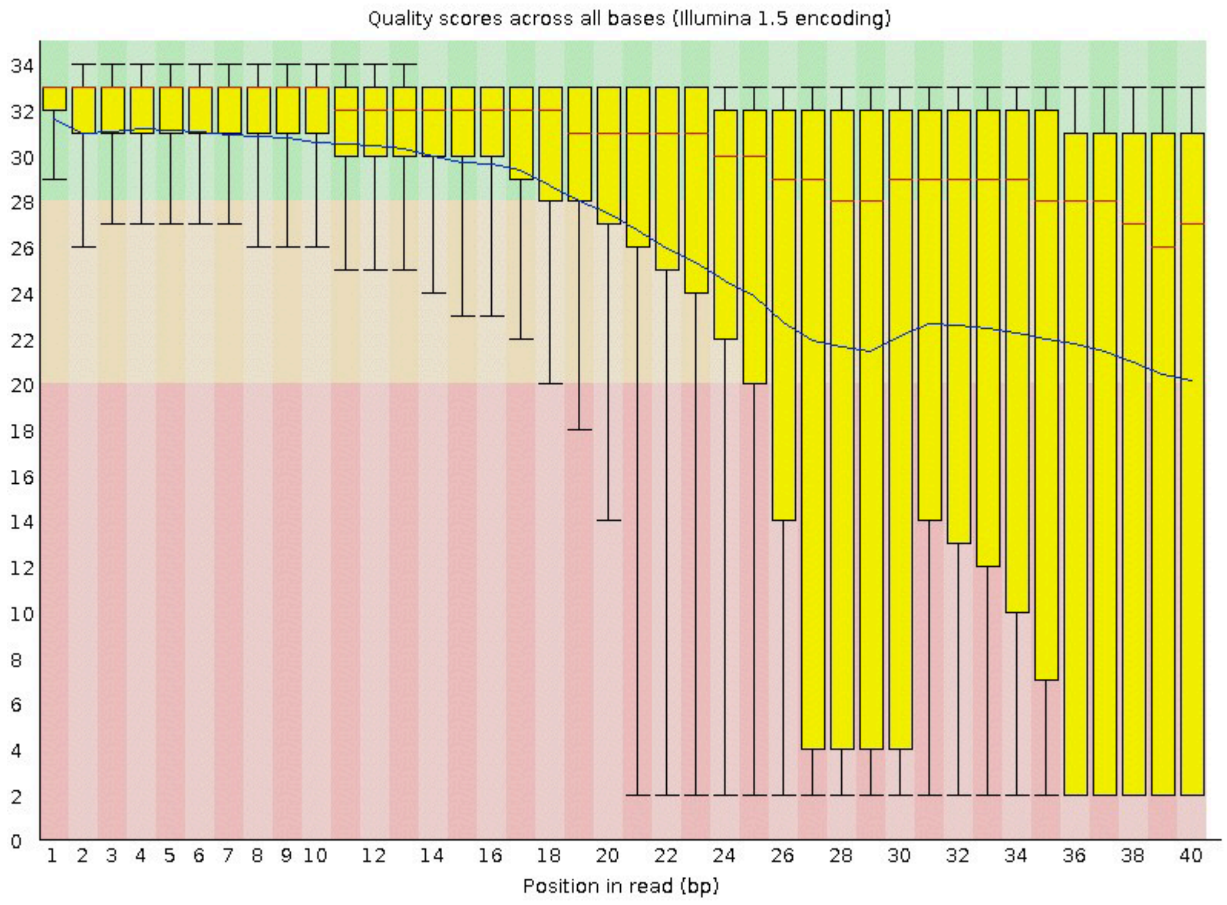
Base quality – a more common example

- Decrease of quality towards the end of reads (late cycles)



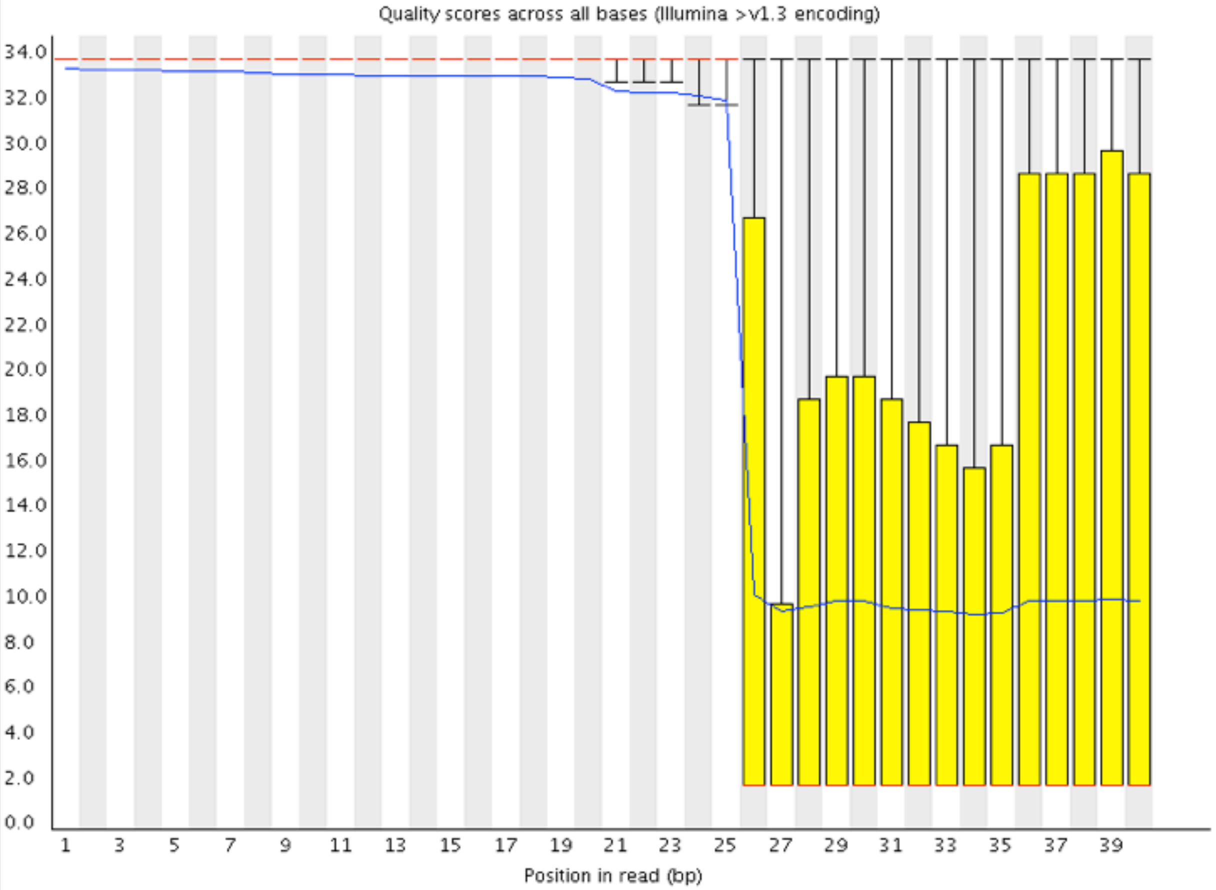
Base quality – bad example

- …



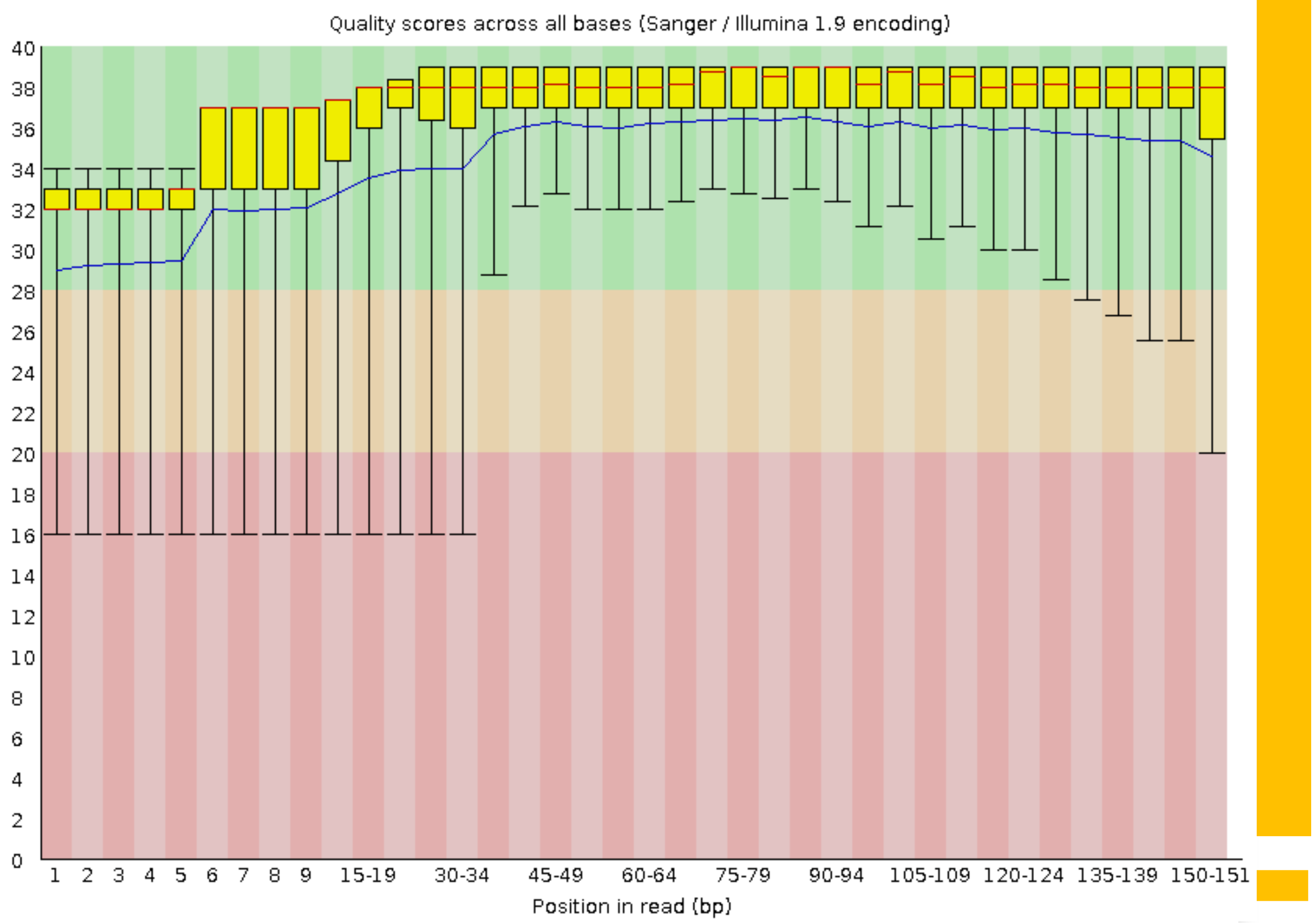
Base quality - sudden quality drop

- Indicates problems with flow cell, trimming needed



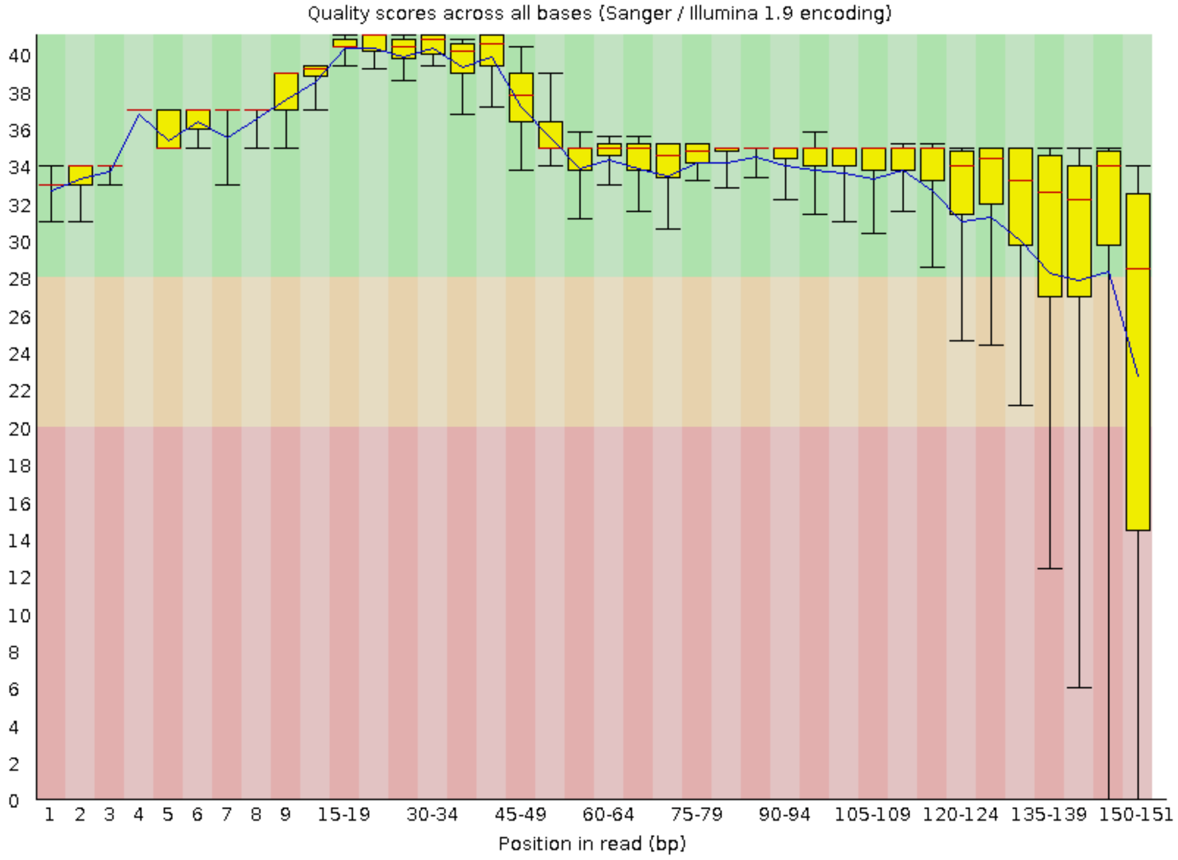
Base quality – targeted sequencing

- The low quality extremes suggest a problem in the beginning of the reads
- (primers?, NNNNN sequences...)



Base quality – microbiome

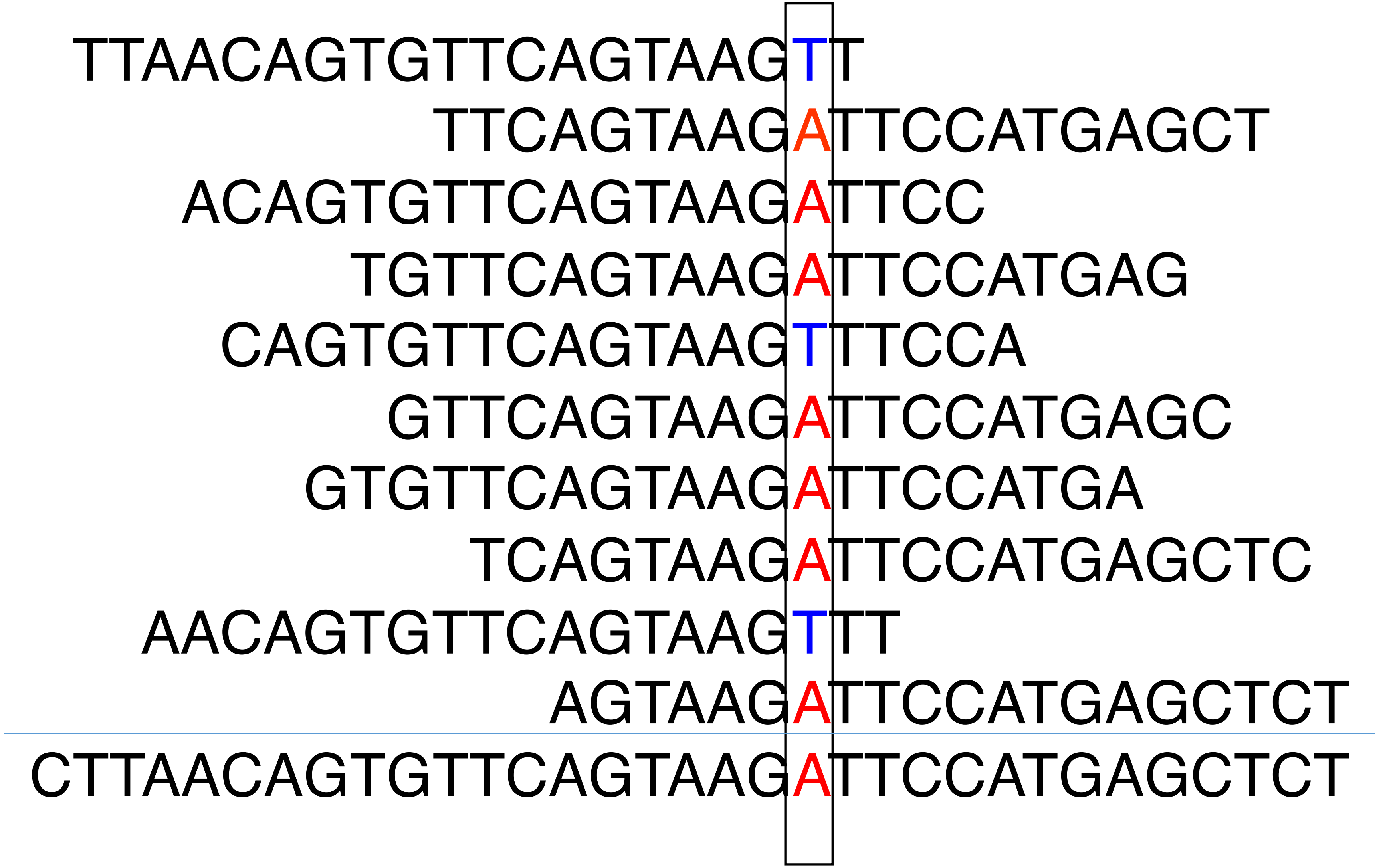
- This is what it can look like with very small sample and sequence size



Base call errors in last cycles

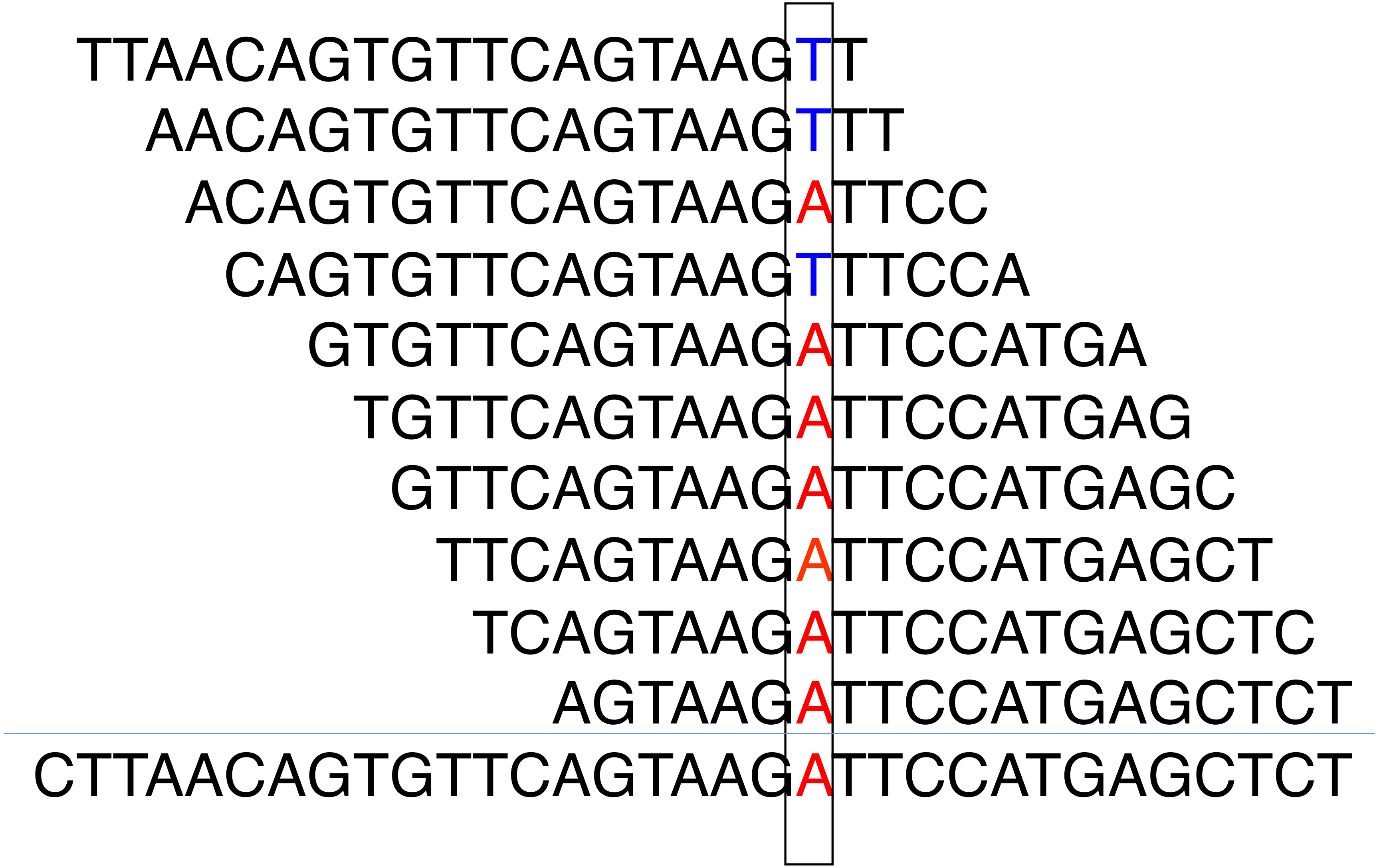
- Towards the end of sequencing, the quality drops, signal is worse
- We can see it for Illumina and SOLID
- Not very important for RNAseq, but crucial for variant calling

SNP calling



```
 TTAACAGTGTTCAGTAAGTT
          TTCAGTAAGATTCCATGAGCT
    ACAGTGTTCAGTAAGATTCC
        TGTTCAGTAAGATTCCATGAG
     CAGTGTTCAGTAAGTTTCCA
         GTTCAGTAAGATTCCATGAGC
       GTGTTCAGTAAGATTCCATGA
           TCAGTAAGATTCCATGAGCTC
   AACAGTGTTCAGTAAGTTT
             AGTAAGATTCCATGAGCTCT
CTTAACAGTGTTCAGTAAGATTCCATGAGCTCT
```

SNP calling



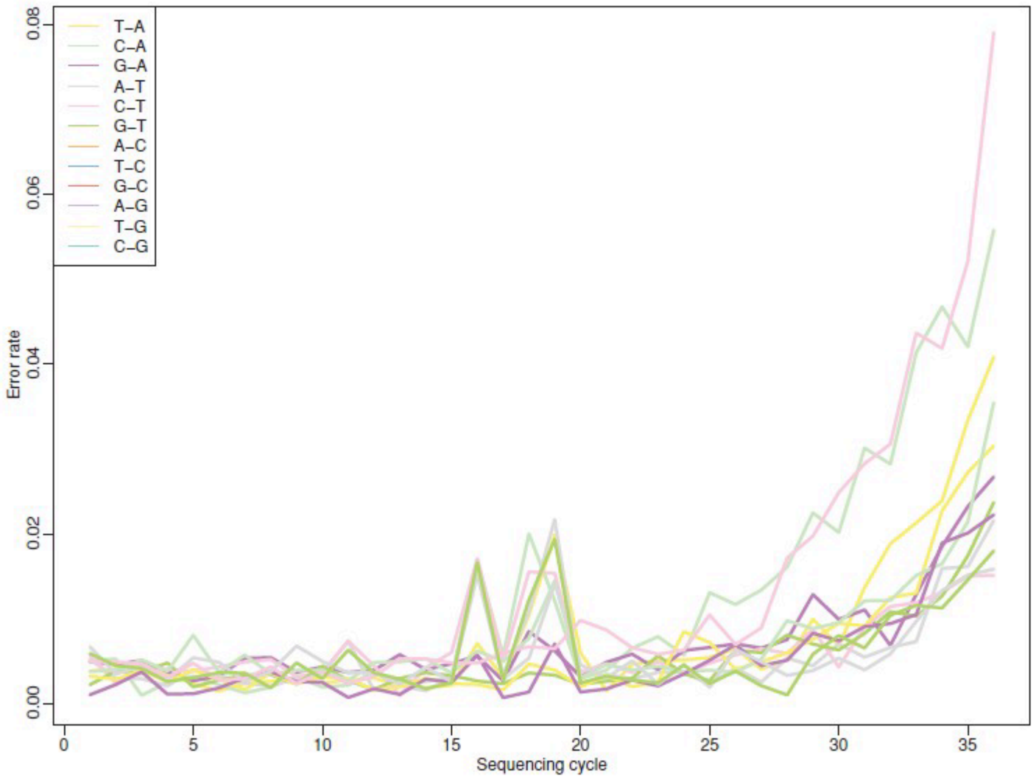
```
 TTAACAGTGTTCAGTAAGTT
   AACAGTGTTCAGTAAGTTT
    ACAGTGTTCAGTAAGATTCC
     CAGTGTTCAGTAAGTTTCCA
       GTGTTCAGTAAGATTCCATGA
        TGTTCAGTAAGATTCCATGAG
         GTTCAGTAAGATTCCATGAGC
          TTCAGTAAGATTCCATGAGCT
           TCAGTAAGATTCCATGAGCTC
             AGTAAGATTCCATGAGCTCT
CTTAACAGTGTTCAGTAAGATTCCATGAGCTCT
```

SNP error dependent on cycle

These errors are not random and look like SNPs (e.g. if there were randomly distributed T, C, G and A's, we would conclude it is error directly)

We want the SNPs to be distributed evenly across cycles

SNPs coming from towards end of the read are sign of false positive

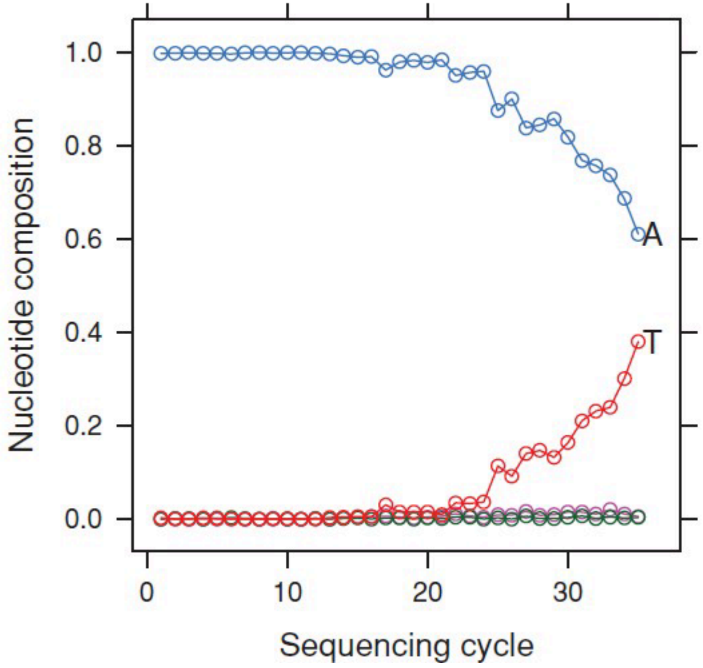


SNP error dependent on cycle

These errors are not random and look like SNPs (e.g. if there were randomly distributed T, C, G and A's, we would conclude it is error directly)

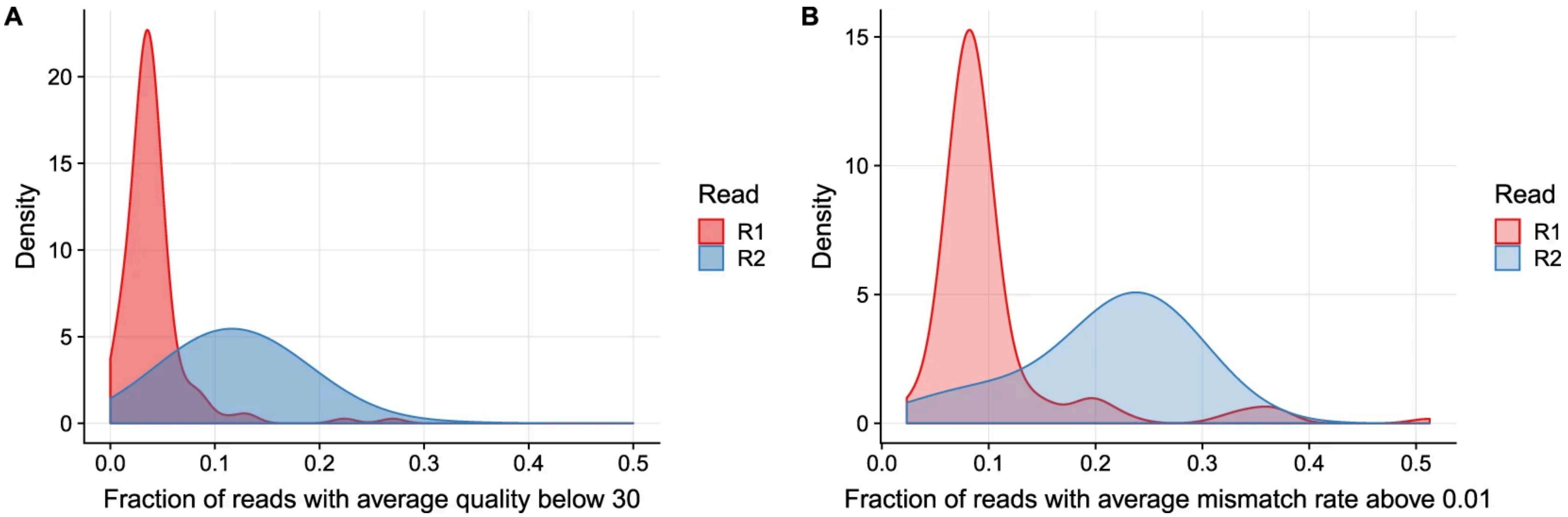
We want the SNPs to be distributed evenly across cycles

SNPs coming from towards end of the read are sign of false positive



Long fragments have lower base quality

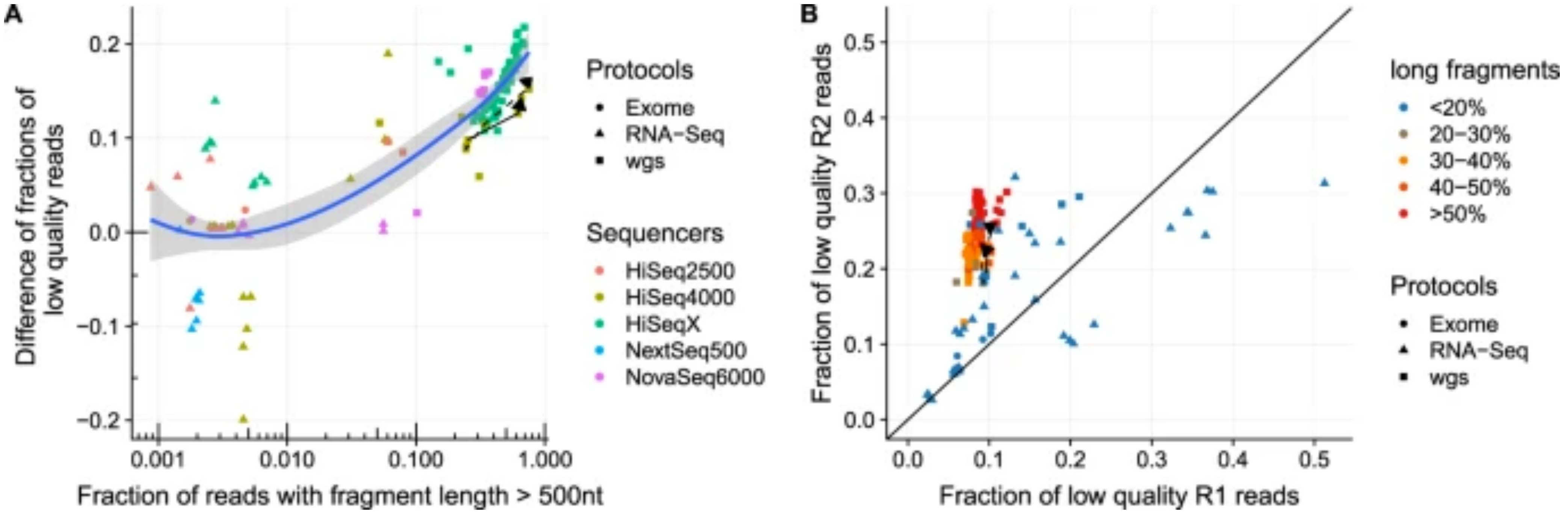
From: Long fragments achieve lower base quality in Illumina paired-end sequencing



We plot the fraction of low quality reads in the 138 samples analyzed in our study. Across all samples the R2 reads harbor more low quality reads than the R1 reads. We plot two alternative definitions of 'low quality'. Reads are called low quality if (**A**) the average Phred score is below 30, or (**B**) the average mismatch rate of the aligned bases is above 0.01. Both plots show that the R2 reads harbor more low quality reads and that the fraction of low quality reads is more variable across samples.

Increase of R2 low quality reads as a function of the content of long fragments

From: Long fragments achieve lower base quality in Illumina paired-end sequencing



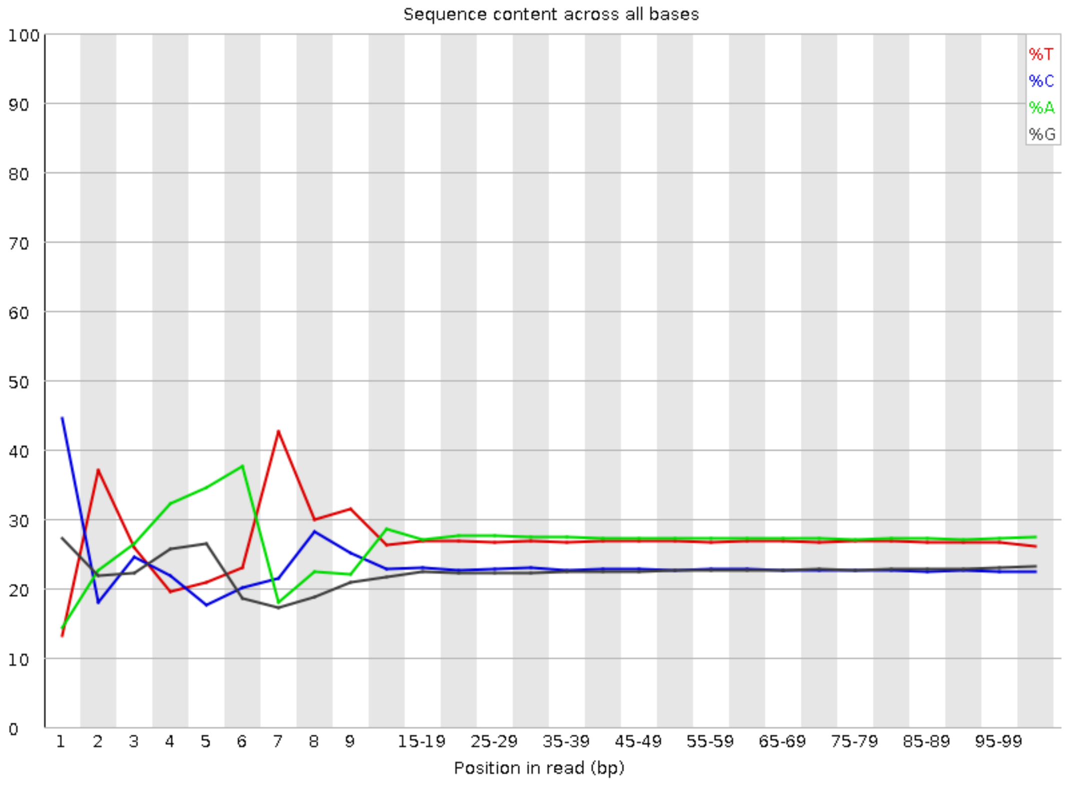
Increase of R2 low quality reads as a function of the content of long fragments. In (**A**) we plot for individual samples the difference in low quality read content among the R2 and the R1 reads versus the content of long fragments. The plot shows that the more long fragments a samples has the more prevalent are low quality reads among the R2 reads. In (**B**) we directly compare the fraction of low quality reads in R2 and R1 and color-code the content long fragments. Low quality reads are defined as reads having a mismatch rate above 0.01 in the bases after alignment. The plotted samples have been generated using various protocols on various sequencers in various labs. The dashed lines connect three samples each that have been processed identically except with an increasing targeted fragment length.

Per base sequence content

- Sequence content across bases should not change with cycle

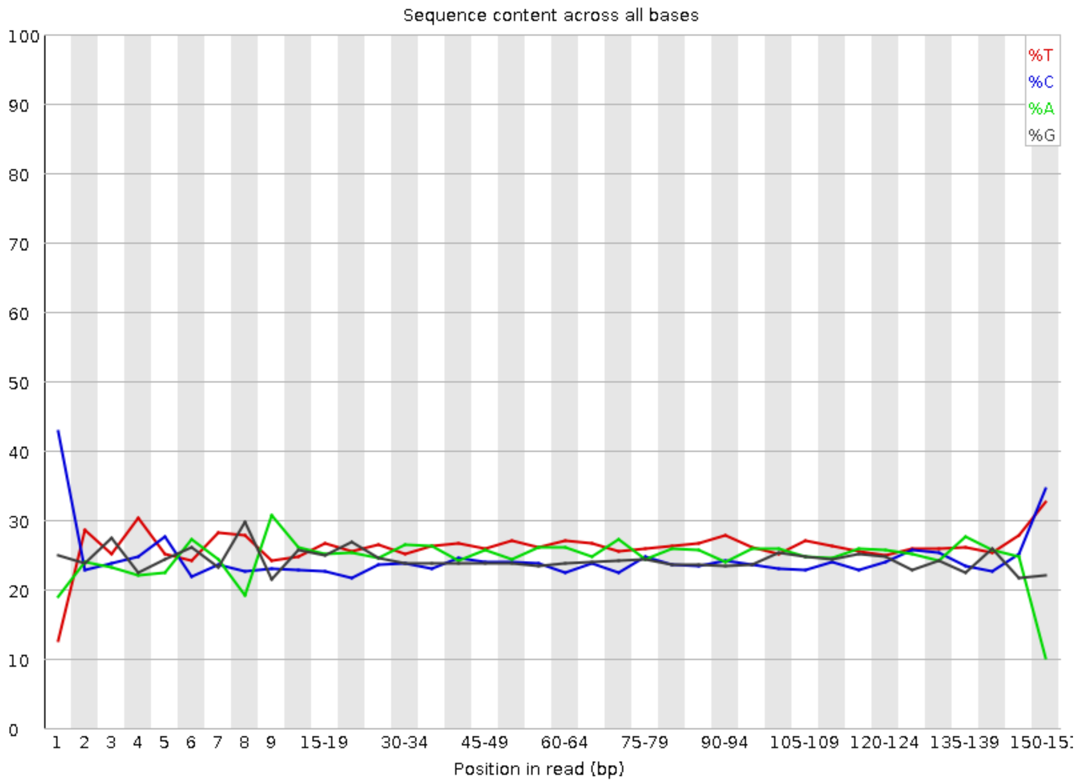
Per base sequence content – RNAseq – typical Illumina library

- The primers used in the library are typically not removed



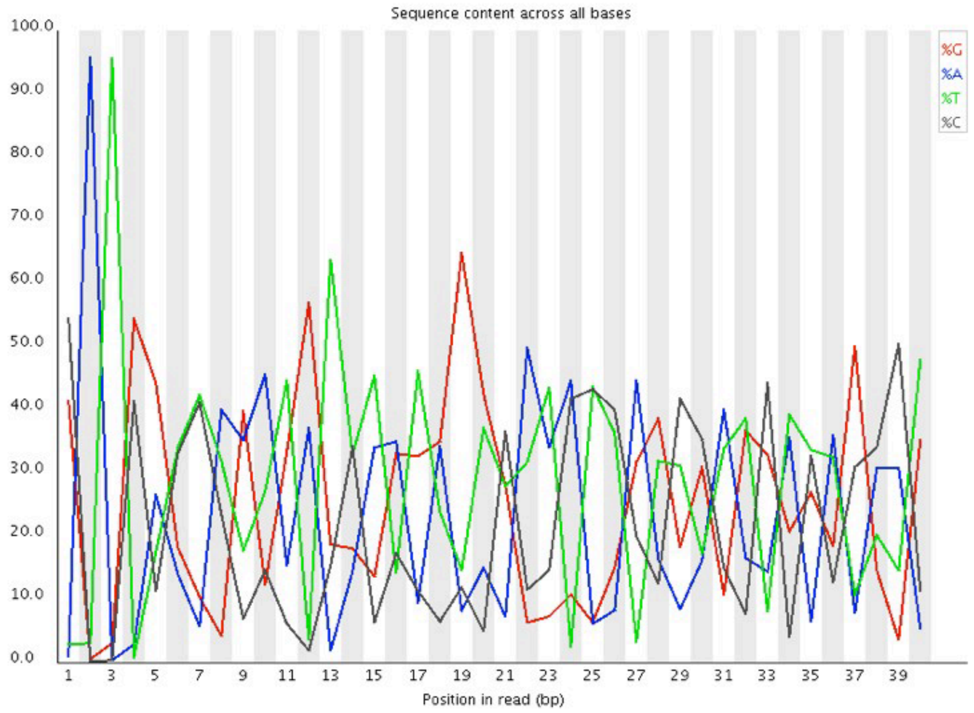
Per base sequence content – targeted sequencing

- In targeted sequencing there is much less genes being sequenced so the base composition of reads is non-random



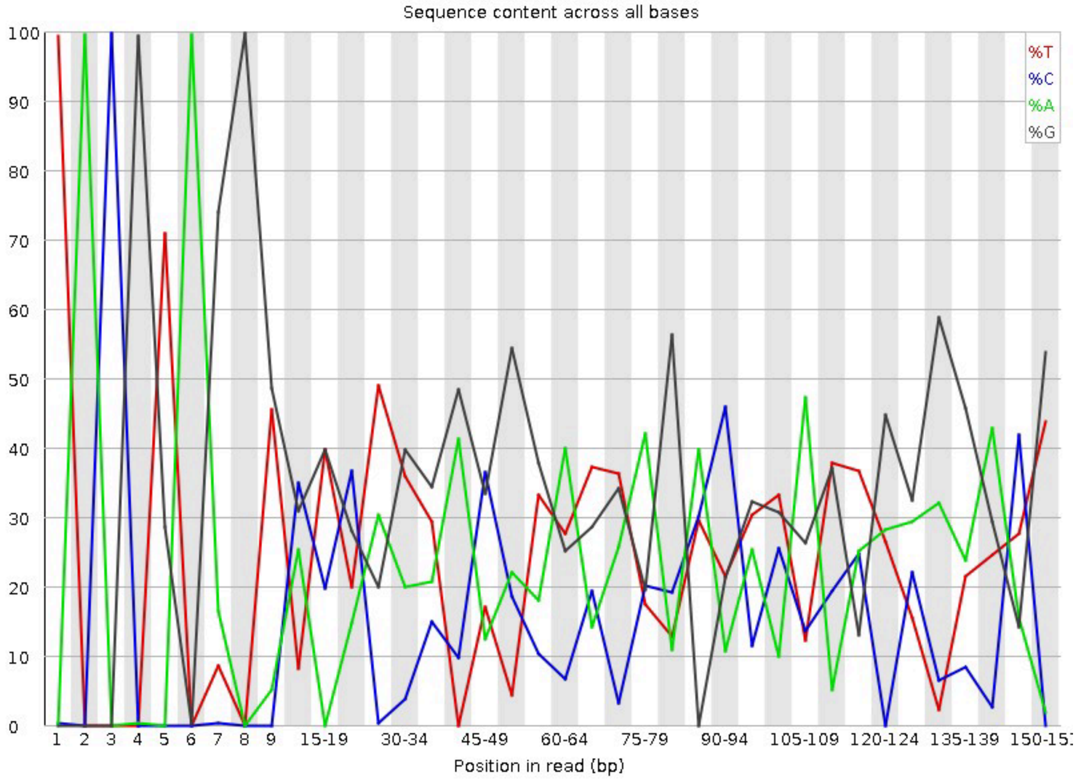
Per base sequence content – a bad example?

- This suggests that a single sequence makes up a large part of the library – this can mean rRNA contamination in RNAseq



Per base sequence content - microbiome

- … however, it is excepted if we sequence 16S rRNA of microbiome where one or few bacteria strains are dominating

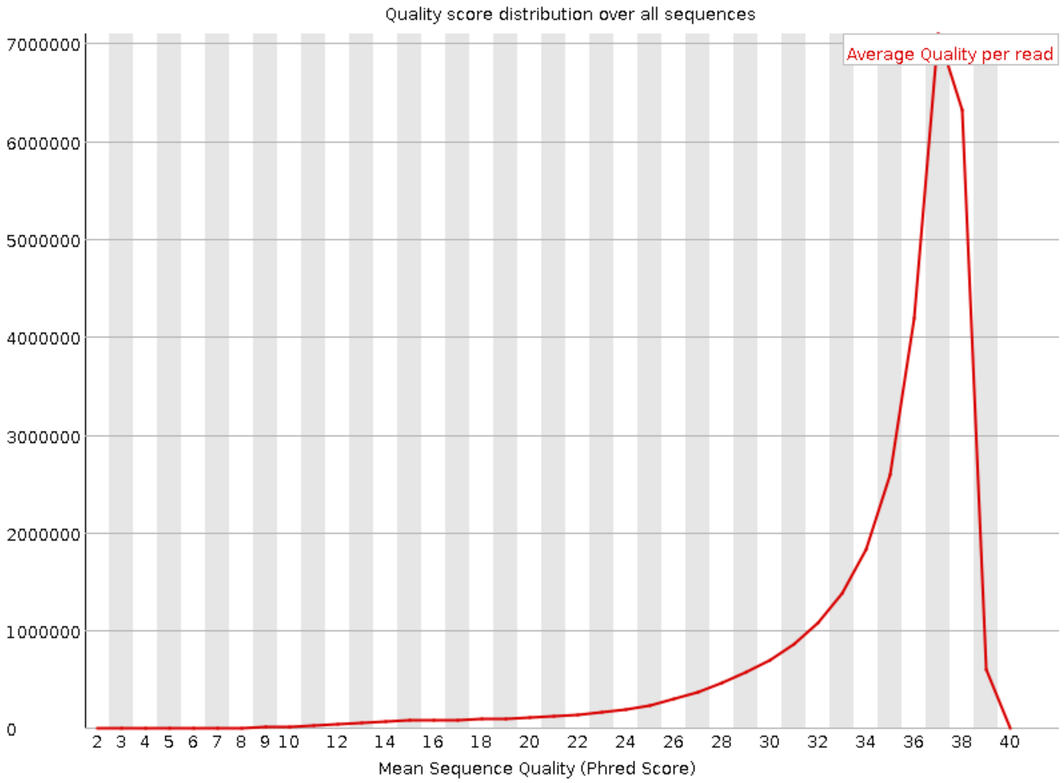


Per sequence quality

- All – or at least majority of the sequences should have good average quality (average Phred score across all read bases)

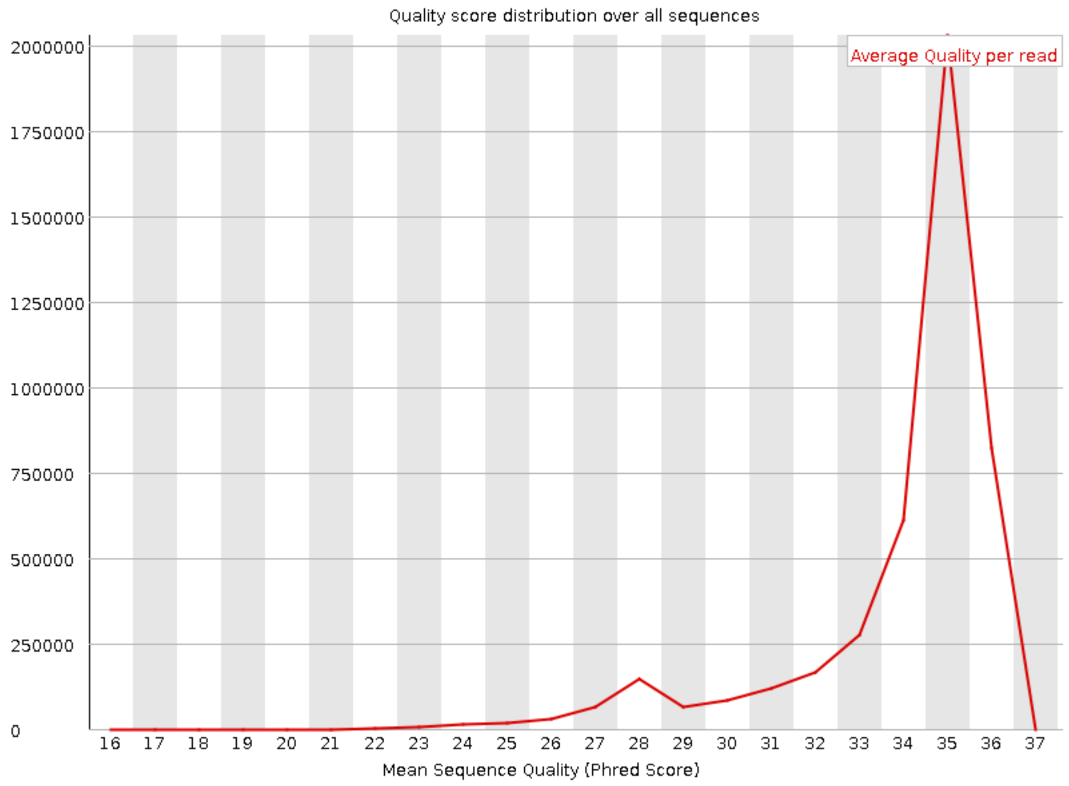
Per sequence quality - RNAseq

- majority of the sequences have good average quality



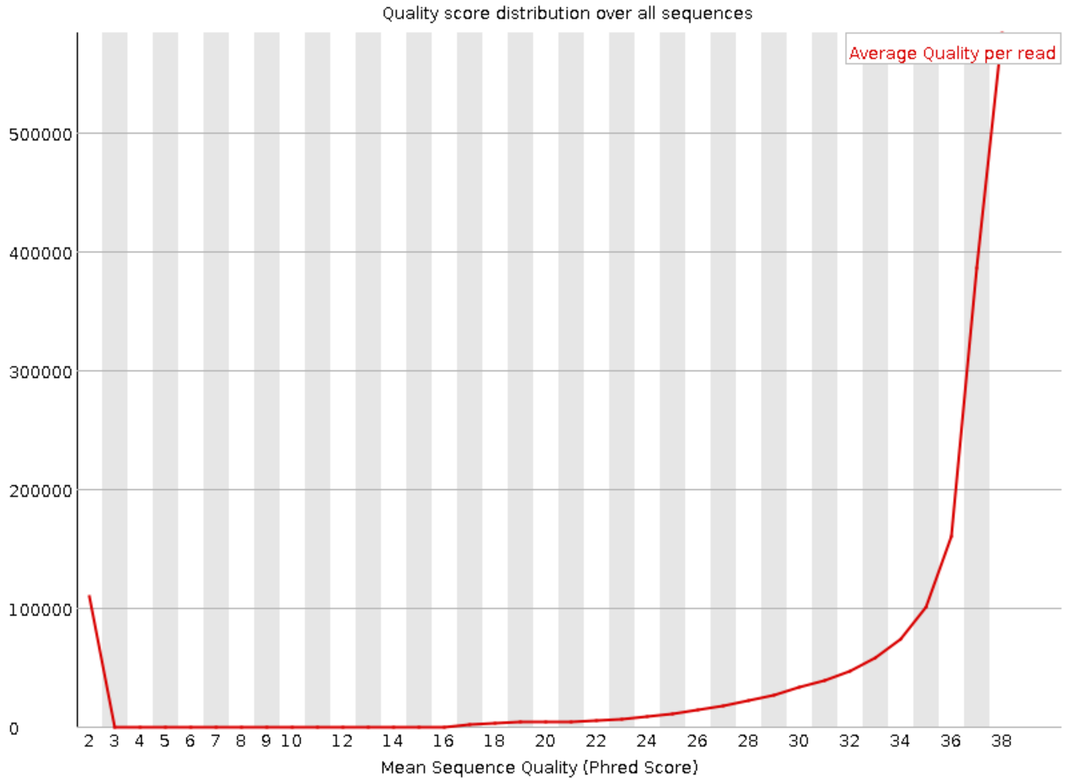
Per sequence quality - microbiome

- Small peaks in lower average quality can suggest low quality ends on part of sequences – attention, if small read diversity (e.g. microbiome), this can be due to highly duplicated reads due to too deep sequencing



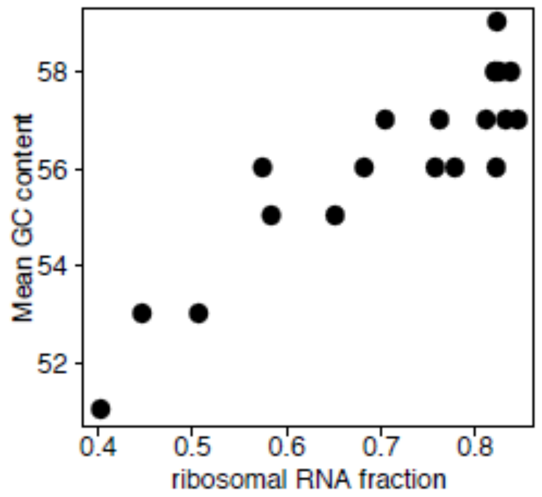
Per sequence quality – targeted sequencing

- Small peaks in lower average quality can suggest low quality ends on part of sequences – attention, if small read diversity (e.g. microbiome), this can be due to highly duplicated reads due to too deep sequencing



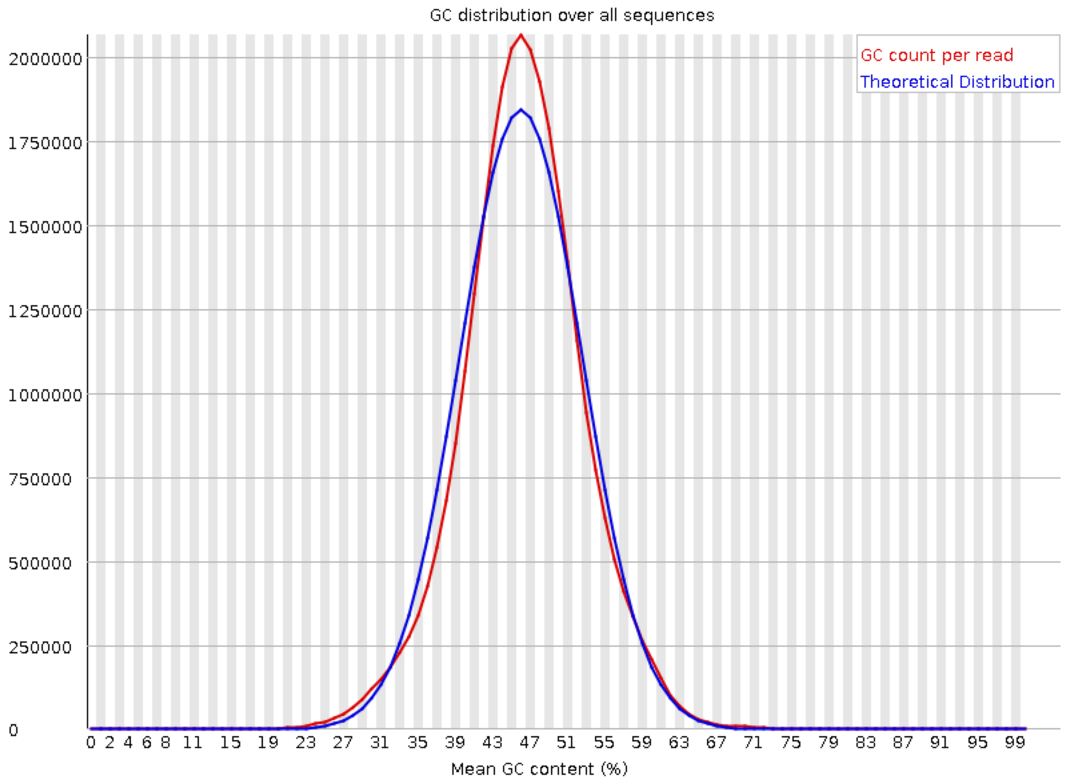
Per sequence and per read GC content

- Mean GC content across reads should correspond to the overall GC content of the genome
- Evan small shifts can indicate contamination with GC rich sequences (ribosomal RNA with high GC content for instance)



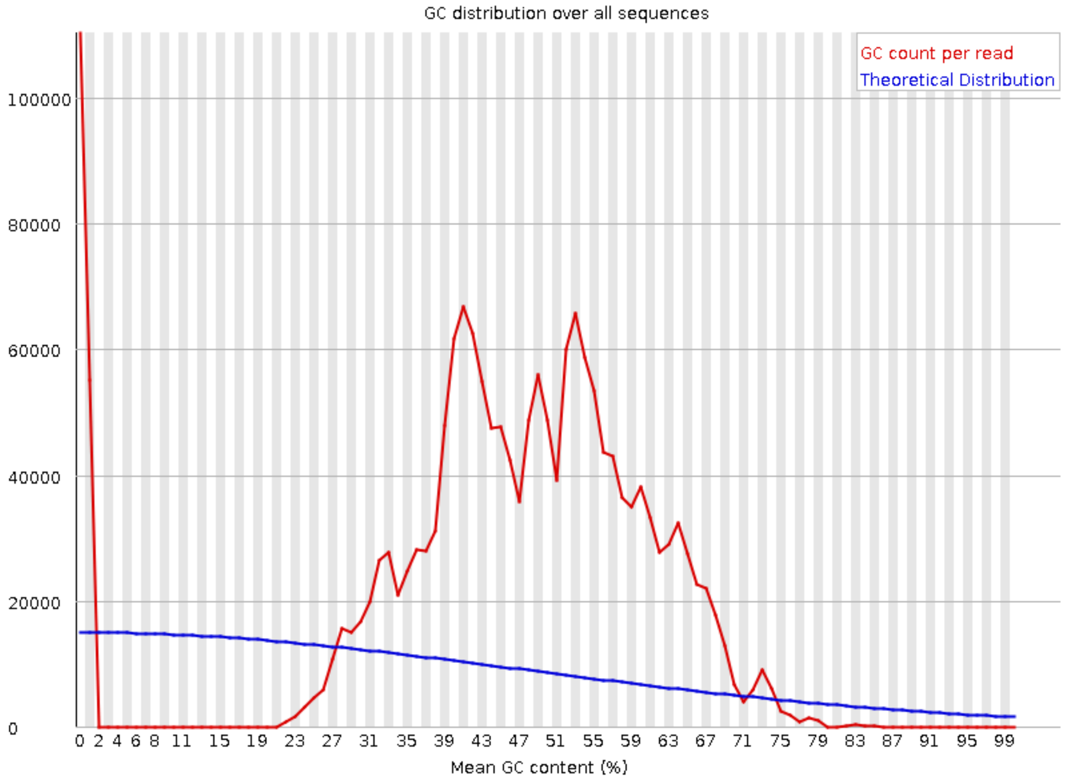
Per sequence GC content - RNAseq

- A relatively good example of GC content



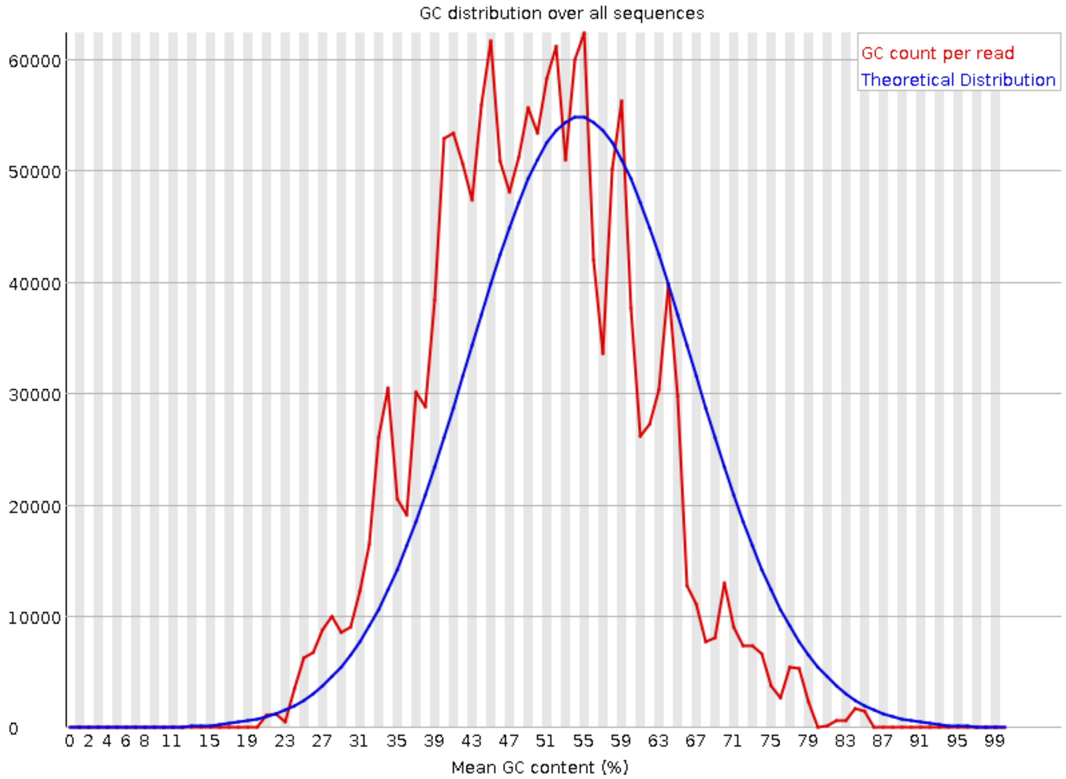
Per sequence GC content – targeted sequencing

- This strange theoretical distribution is due to high amount of NNNNNN sequences in the reads



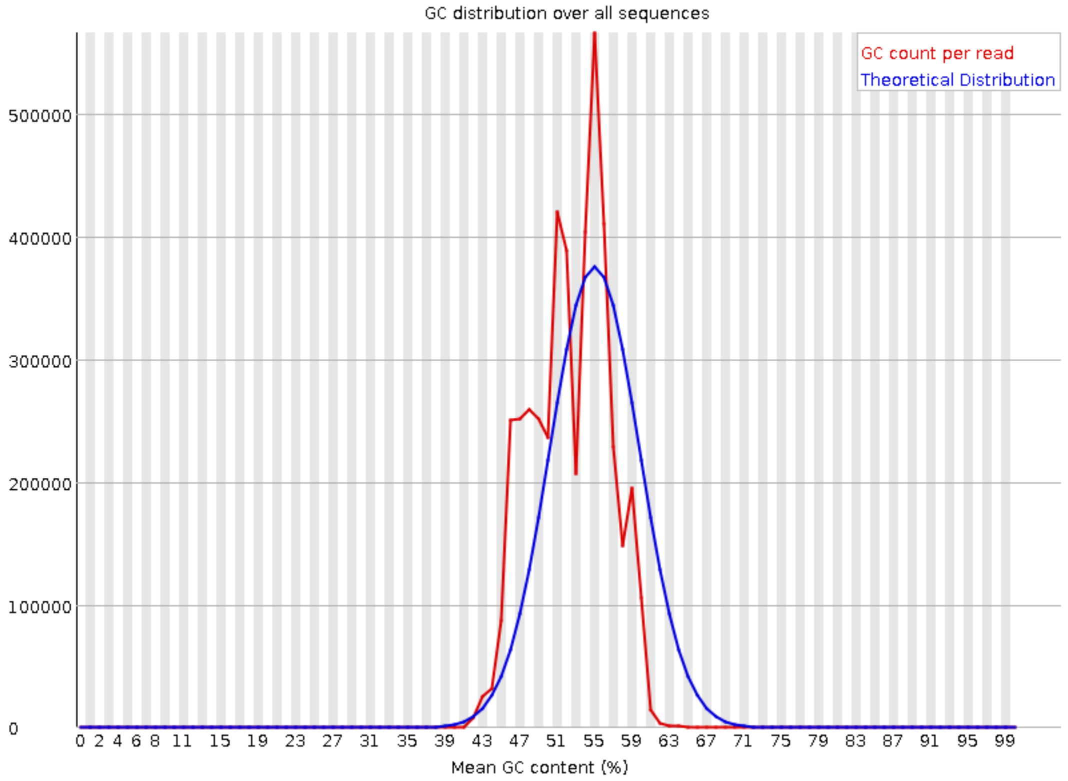
Per sequence GC content – targeted sequencing after trimming

- The GC count per read is disturbed because of small number of genes sequenced!



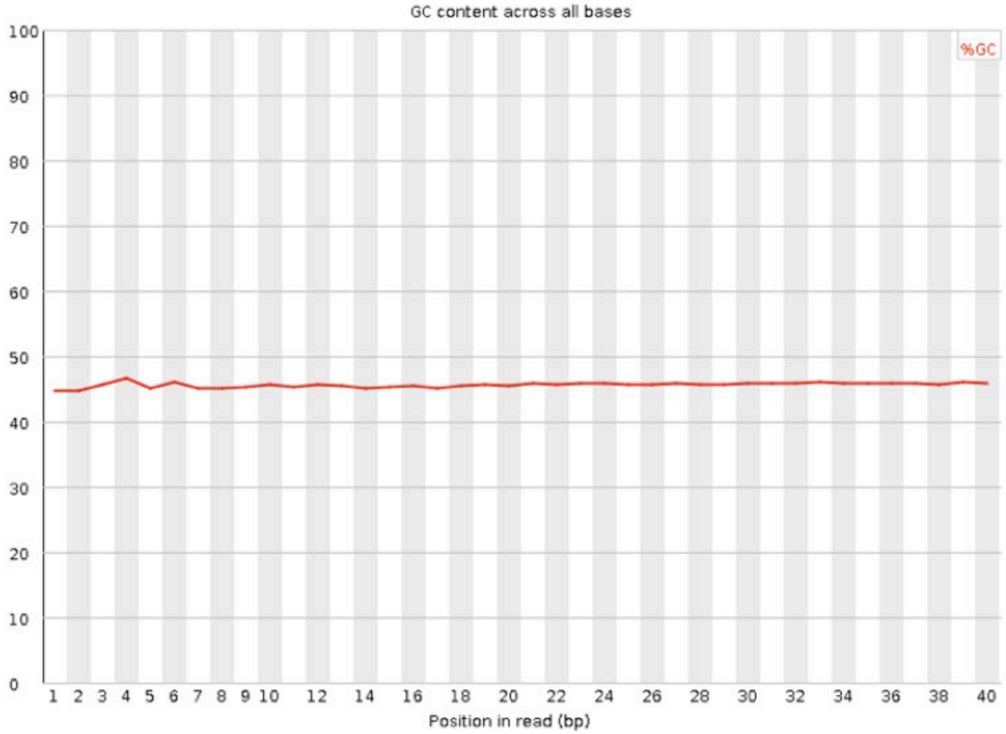
Per sequence GC content – microbiome

- The GC count per read is disturbed because of small diversity of sequences



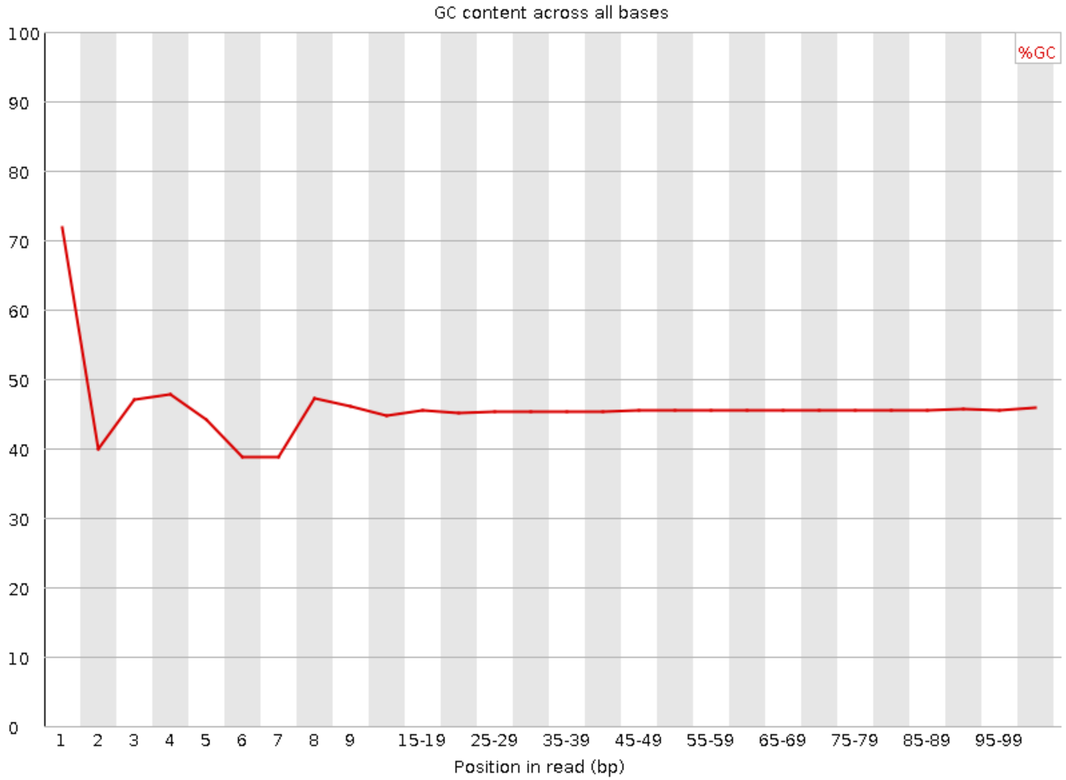
Per read GC content – good example

- The GC count per read is disturbed because of small diversity of sequences



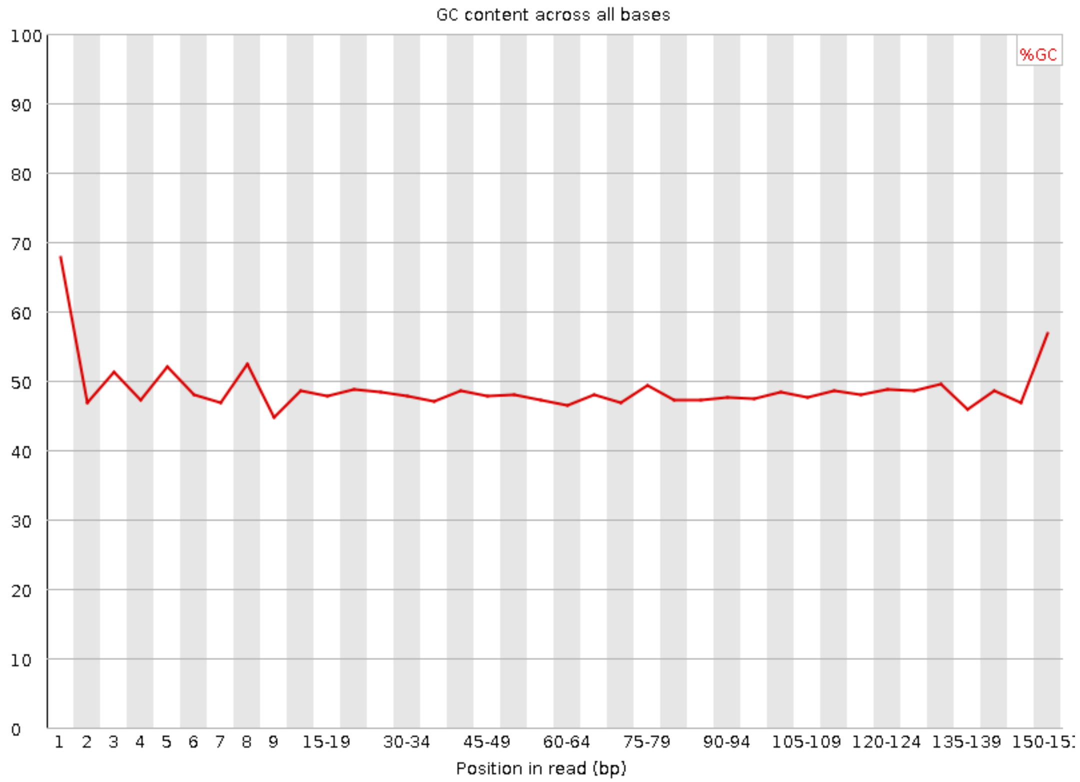
Per read GC content – typical RNAseq

- GC content different in first 8-10 bases, due to presence of primers



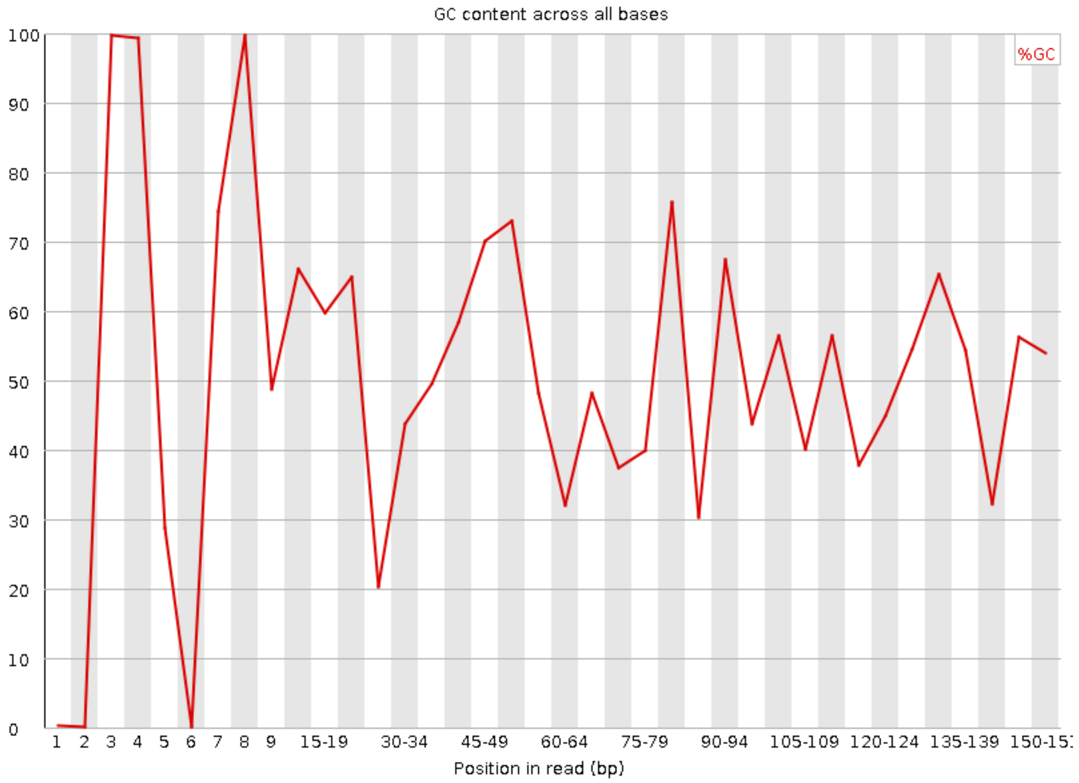
Per read GC content – targeted sequencing

- GC content across different base positions due to high duplication level of reads and small diversity



Per read GC content – microbiome

- GC content across different base positions due to high duplication level of reads and even smaller diversity.
- Zero GC in first two bases can be due to adapters.



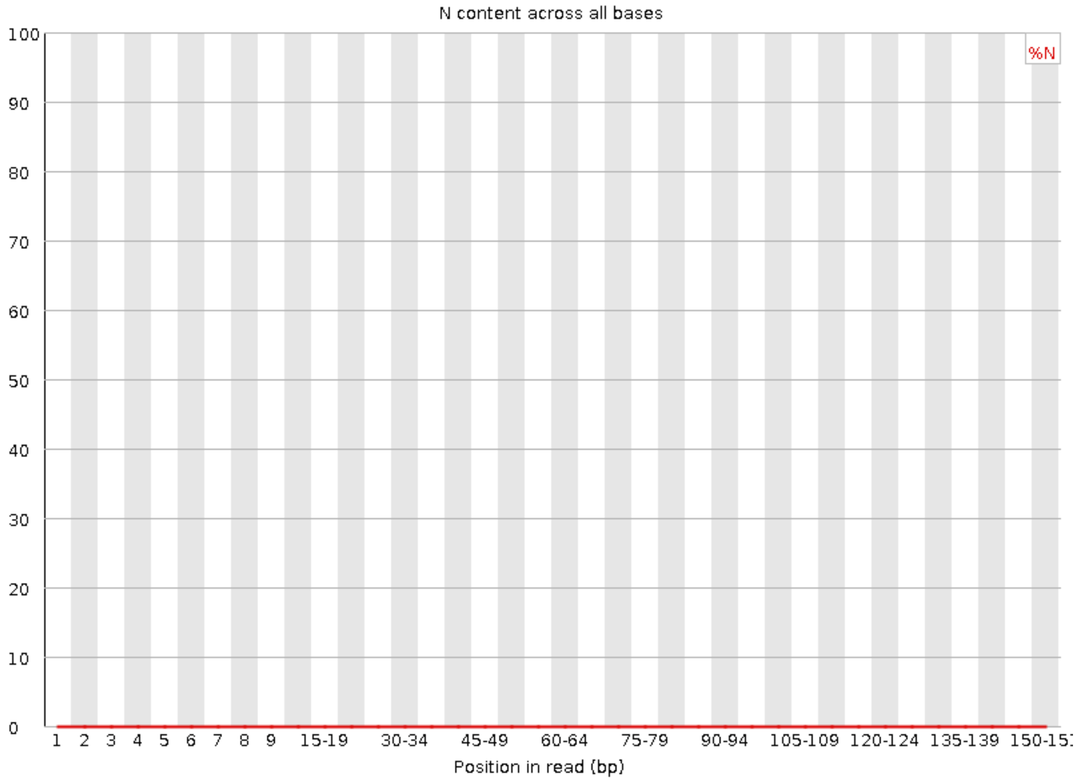
Per base N content

- In ideal case, there should be minimum of N calls in the reads
- “The HiSeq2000 produces very few Ns. It is very rare to see N content greater than 30%. When Ns are produced it is usually the result of some temporary instrument issue. For example a small bubble in the flow cell may cause focus problems at a certain cycle. Downstream processing of Ns depends on your analysis software and strategy.”
- Source:

https://www.biotech.wisc.edu/services/dnaseq/sequencing/Illumina_old/Illumina_QC_FAQs

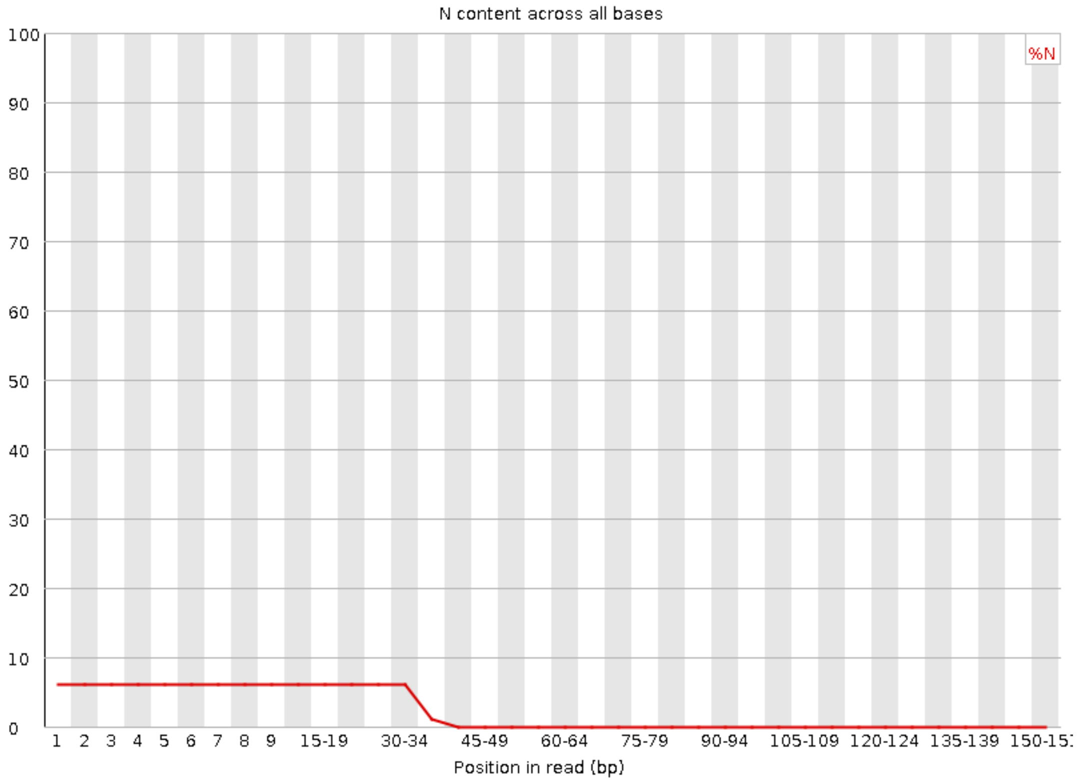
Per base N content – ideal case

- …



Per base N content – targeted sequencing

Enrichment for N calls at the beginning of the sequence

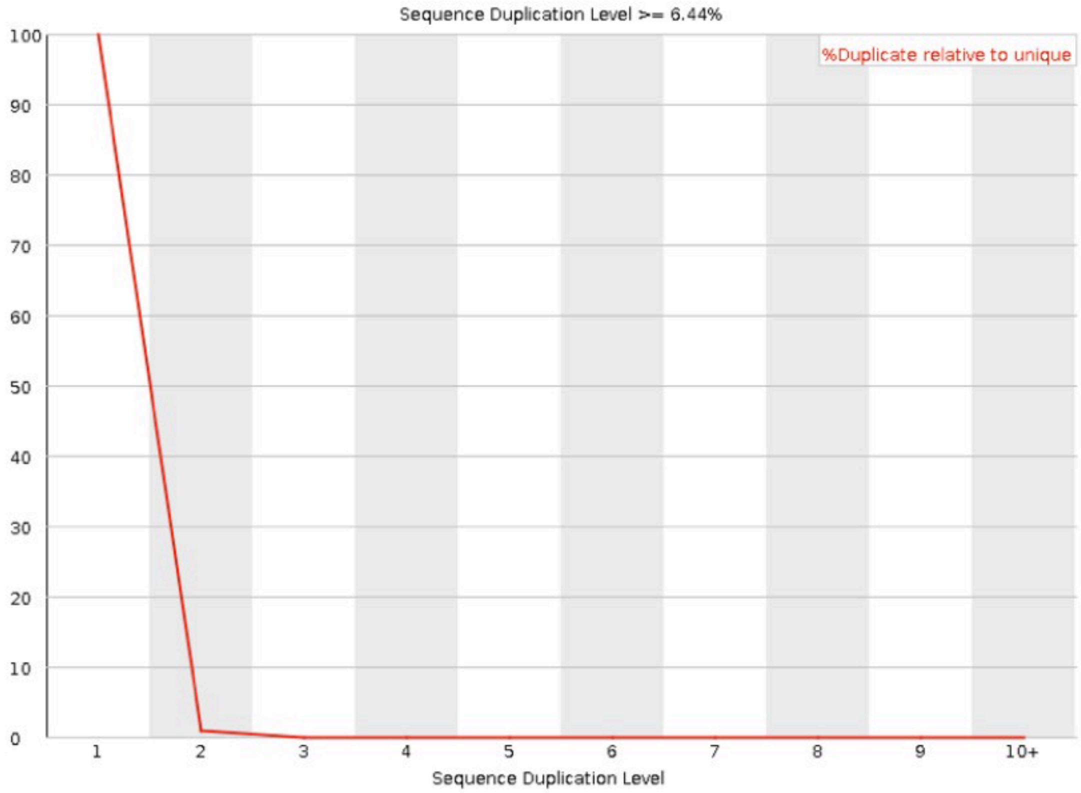


Sequence duplication level and overrepresented sequences

- Indicates the library complexity and possible contamination
  - (the less duplicates, the more complex)
- Too many duplicated sequences means we sequenced “too much”.
- Overrepresented sequences may indicate:
  - Presence of adapters, presence of contamination (rRNA), PCR problems
- This holds, however, mainly for WGS, WES or RNAseq

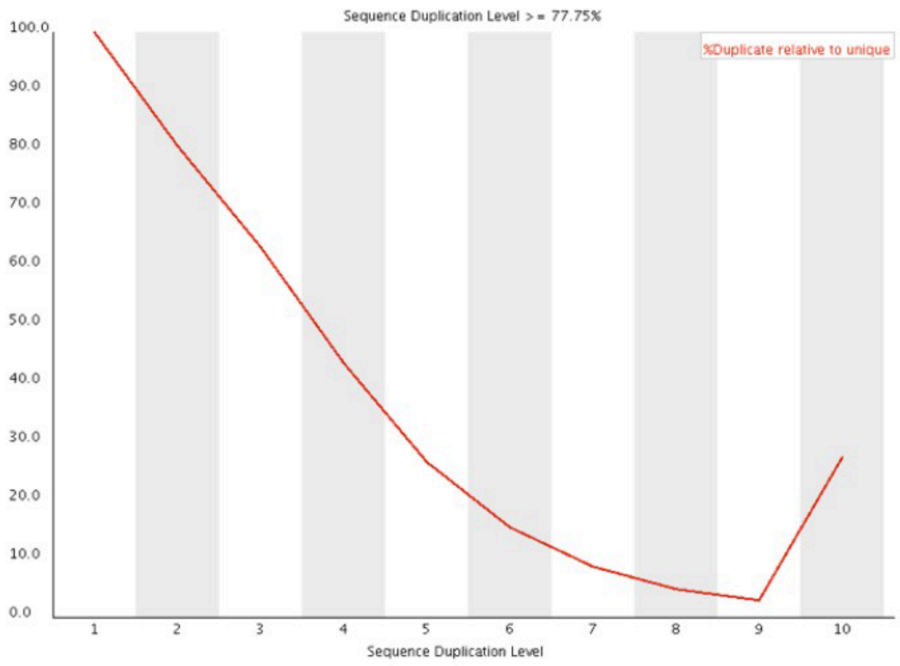
Sequence duplication level – good example (RNAseq)

- Most sequences occur only once



Sequence duplication level – bad example (RNAseq)?

- Over-amplification! May come from highly expressed transcripts.

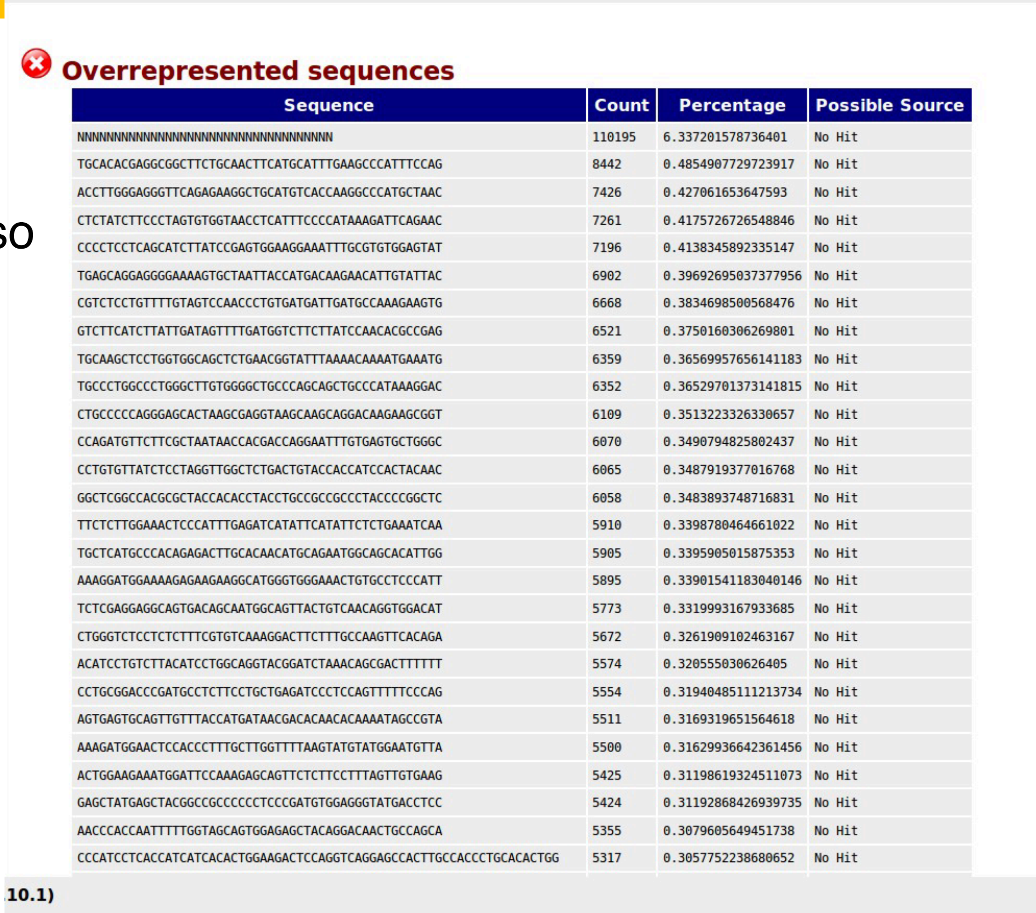


Overrepresented sequences

Overrepresented sequences

- Indicate remaining adapters, PCR duplicates, but also can be real sequences!

Always judge based on type of data and check before filtering!

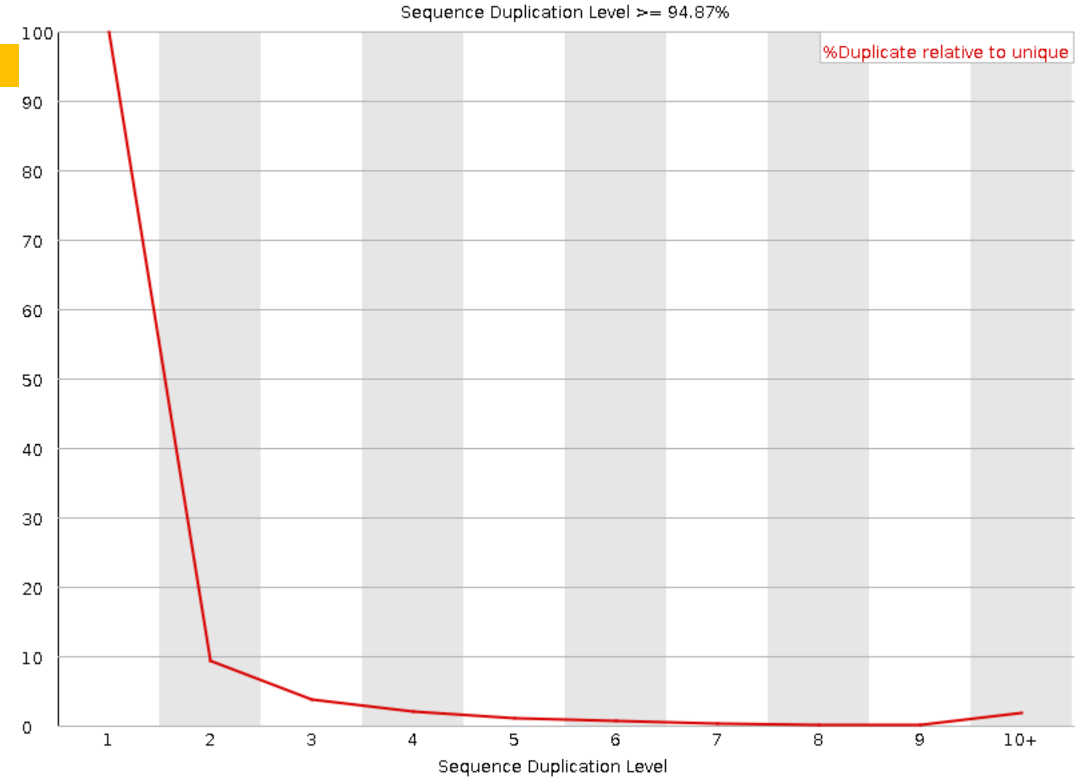


Overrepresented sequences

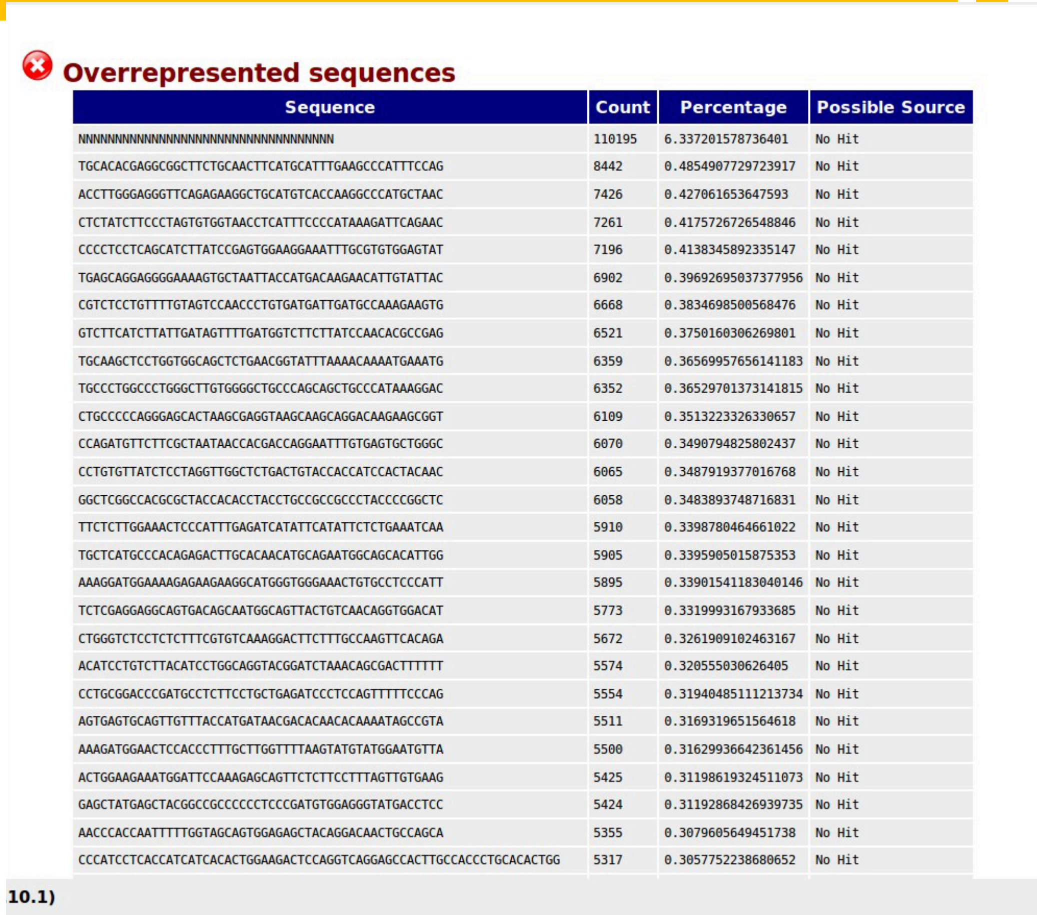
Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	110195	6.337201578736401	No Hit
TGCACACGAGGCGGCTTCTGCAACTTCATGCATTTGAAGCCCATTTCCAG	8442	0.4854907729723917	No Hit
ACCTTGGGAGGGTTCAGAGAAGGCTGCATGTCACCAAGGCCCATGCTAAC	7426	0.427061653647593	No Hit
CTCTATCTTCCCTAGTGTGGTAACCTCATTTCCCCATAAAGATTCAGAAC	7261	0.4175726726548846	No Hit
CCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTAT	7196	0.4138345892335147	No Hit
TGAGCAGGAGGGGAAAAGTGCTAATTACCATGACAAGAACATTGTATTAC	6902	0.39692695037377956	No Hit
CGTCTCCTGTTTTGTAGTCCAACCCTGTGATGATTGATGCCAAAGAAGTG	6668	0.3834698500568476	No Hit
GTCTTCATCTTATTGATAGTTTTGATGGTCTTCTTATCCAACACGCCGAG	6521	0.3750160306269801	No Hit
TGCAAGCTCCTGGTGGCAGCTCTGAACGGTATTTAAAACAAAATGAAATG	6359	0.36569957656141183	No Hit
TGCCCTGGCCCTGGGCTTGTGGGGCTGCCCAGCAGCTGCCCATAAAGGAC	6352	0.36529701373141815	No Hit
CTGCCCCCAGGGAGCACTAAGCGAGGTAAGCAAGCAGGACAAGAAGCGGT	6109	0.3513223326330657	No Hit
CCAGATGTTCTTCGCTAATAACCACGACCAGGAATTTGTGAGTGCTGGGC	6070	0.3490794825802437	No Hit
CCTGTGTTATCTCCTAGGTTGGCTCTGACTGTACCACCATCCACTACAAC	6065	0.3487919377016768	No Hit
GGCTCGGCCACGCGCTACCACACCTACCTGCCGCCGCCCTACCCCGGCTC	6058	0.3483893748716831	No Hit
TTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAA	5910	0.3398780464661022	No Hit
TGCTCATGCCCACAGAGACTTGCACAACATGCAGAATGGCAGCACATTGG	5905	0.3395905015875353	No Hit
AAAGGATGGAAAAGAGAAGAAGGCATGGGTGGGAAACTGTGCCTCCCATT	5895	0.33901541183040146	No Hit
TCTCGAGGAGGCAGTGACAGCAATGGCAGTTACTGTCAACAGGTGGACAT	5773	0.3319993167933685	No Hit
CTGGGTCTCCTCTCTTTCGTGTCAAAGGACTTCTTTGCCAAGTTCACAGA	5672	0.3261909102463167	No Hit
ACATCCTGTCTTACATCCTGGCAGGTACGGATCTAAACAGCGACTTTTTT	5574	0.320555030626405	No Hit
CCTGCGGACCCGATGCCTCTTCCTGCTGAGATCCCTCCAGTTTTTCCCAG	5554	0.31940485111213734	No Hit
AGTGAGTGCAGTTGTTTACCATGATAACGACACAACACAAAATAGCCGTA	5511	0.3169319651564618	No Hit
AAAGATGGAACTCCACCCTTTGCTTGGTTTTAAGTATGTATGGAATGTTA	5500	0.31629936642361456	No Hit
ACTGGAAGAAATGGATTCCAAAGAGCAGTTCTCTTCCTTTAGTTGTGAAG	5425	0.31198619324511073	No Hit
GAGCTATGAGCTACGGCCGCCCCCCTCCCGATGTGGAGGGTATGACCTCC	5424	0.31192868426939735	No Hit
AACCCACCAATTTTTGGTAGCAGTGGAGAGCTACAGGACAACTGCCAGCA	5355	0.3079605649451738	No Hit
CCCATCCTCACCATCATCACACTGGAAGACTCCAGGTCAGGAGCCACTTGCCACCCTGCACACTGG	5317	0.3057752238680652	No Hit

.10.1)

Sequence duplication level – targeted sequencing



Some reads are present more than 10 times. This is due to NNNN sequences and due to few genes sequenced (longer genes get more reads)



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	110195	6.337201578736401	No Hit
TGCACACGAGGCGGCTTCTGCAACTTCATGCATTTGAAGCCCATTTCCAG	8442	0.4854907729723917	No Hit
ACCTTGGGAGGGTTCAGAGAAGGCTGCATGTCACCAAGGCCCATGCTAAC	7426	0.427061653647593	No Hit
CTCTATCTTCCCTAGTGTGGTAACCTCATTTCCCCATAAAGATTCAGAAC	7261	0.4175726726548846	No Hit
CCCCTCCTCAGCATCTTATCCGAGTGGAAGGAAATTTGCGTGTGGAGTAT	7196	0.4138345892335147	No Hit
TGAGCAGGAGGGGAAAAGTGCTAATTACCATGACAAGAACATTGTATTAC	6902	0.39692695037377956	No Hit
CGTCTCCTGTTTTGTAGTCCAACCCTGTGATGATTGATGCCAAAGAAGTG	6668	0.3834698500568476	No Hit
GTCTTCATCTTATTGATAGTTTTGATGGTCTTCTTATCCAACACGCCGAG	6521	0.3750160306269801	No Hit
TGCAAGCTCCTGGTGGCAGCTCTGAACGGTATTTAAAACAAAATGAAATG	6359	0.36569957656141183	No Hit
TGCCCTGGCCCTGGGCTTGTGGGGCTGCCCAGCAGCTGCCCATAAAGGAC	6352	0.36529701373141815	No Hit
CTGCCCCCAGGGAGCACTAAGCGAGGTAAGCAAGCAGGACAAGAAGCGGT	6109	0.3513223326330657	No Hit
CCAGATGTTCTTCGCTAATAACCACGACCAGGAATTTGTGAGTGCTGGGC	6070	0.3490794825802437	No Hit
CCTGTGTTATCTCCTAGGTTGGCTCTGACTGTACCACCATCCACTACAAC	6065	0.3487919377016768	No Hit
GGCTCGGCCACGCGCTACCACACCTACCTGCCGCCGCCCTACCCCGGCTC	6058	0.3483893748716831	No Hit
TTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAA	5910	0.3398780464661022	No Hit
TGCTCATGCCCACAGAGACTTGCACAACATGCAGAATGGCAGCACATTGG	5905	0.3395905015875353	No Hit
AAAGGATGGAAAAGAGAAGAAGGCATGGGTGGGAAACTGTGCCTCCCATT	5895	0.33901541183040146	No Hit
TCTCGAGGAGGCAGTGACAGCAATGGCAGTTACTGTCAACAGGTGGACAT	5773	0.3319993167933685	No Hit
CTGGGTCTCCTCTCTTTCGTGTCAAAGGACTTCTTTGCCAAGTTCACAGA	5672	0.3261909102463167	No Hit
ACATCCTGTCTTACATCCTGGCAGGTACGGATCTAAACAGCGACTTTTTT	5574	0.320555030626405	No Hit
CCTGCGGACCCGATGCCTCTTCCTGCTGAGATCCCTCCAGTTTTTCCCAG	5554	0.31940485111213734	No Hit
AGTGAGTGCAGTTGTTTACCATGATAACGACACAACACAAAATAGCCGTA	5511	0.3169319651564618	No Hit
AAAGATGGAACTCCACCCTTTGCTTGGTTTTAAGTATGTATGGAATGTTA	5500	0.31629936642361456	No Hit
ACTGGAAGAAATGGATTCCAAAGAGCAGTTCTCTTCCTTTAGTTGTGAAG	5425	0.31198619324511073	No Hit
GAGCTATGAGCTACGGCCGCCCCCCTCCCGATGTGGAGGGTATGACCTCC	5424	0.31192868426939735	No Hit
AACCCACCAATTTTTGGTAGCAGTGGAGAGCTACAGGACAACTGCCAGCA	5355	0.3079605649451738	No Hit
CCCATCCTCACCATCATCACACTGGAAGACTCCAGGTCAGGAGCCACTTGCCACCCTGCACACTGG	5317	0.3057752238680652	No Hit

10.1)

Quality control exercise

- We continue in our exercise from 1_Preprocessing.sh

Step 1: Read quality control and **data filtering**

- Based on the quality measures, we decide to remove low quality bases and reads

- **Trimming** – removes low quality or unwanted bases from reads, thus shortening them. Is applied to increase the number of mappable reads.

- **Filtering –** removes whole reads that do not meet quality standards (e.g. too short etc)

Trimming reads

- Read trimming is applied to increase the number of mappable reads by:
    - Removing low quality bases at the end of the reads that are likely to contain sequencing errors
    - Removing adapter sequences

Removing adapters

- Important mainly for very short read sequences of interest (when the input DNA fragment is less than the read length)

e.g. for miRNA with 22nt length the adapter gets sequenced more often than for RNA sequences, which are much longer

What is the sequence of adapters?

Best option: ask which kit was used for preparing libraries

Programs: `cutadapt, trimmomatic`

TruSeq Universal Adapter: 5
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3
TruSeq Indexed Adapter 5
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC NNNNNN ATCTCGTATGCCGTCTTCTGCTTG 3

Here "N" is any nucleotide, and the 6 of them together are a unique sequence which can readily be identified as unique to 1 library.

Filtering reads

- We can remove whole reads based on:
- quality of its base calls
- its length (too short reads)
- level of duplication
- …

Trimming and filtering - exercise

Trimming and filtering - exercise

- We continue in our exercise from 1_Preprocessing.sh
- We will use `grep` command to find adapter sequences and `cutadapt` to remove them
- We will trim low quality bases
- Independent work: find specific QC problems in your project data and suggest solutions (what to trim, filter, etc)

Recommended literature

- Fuller et al. 2009: The challenges of sequencing by synthesis http://arep.med.harvard.edu/pdf/Fuller_09.pdf
- https://sequencing.qcfail.com/