Introduction to Sequencing and Quality Control

Dr. James Emmanuel San
CERI-KRISP, UKZN
Sanemmanueljames@gmail.com

What We're Going To Do

THIS AFTERNOON

- Introduction to Sequencing
- Checking the Quality of Sequencing Data

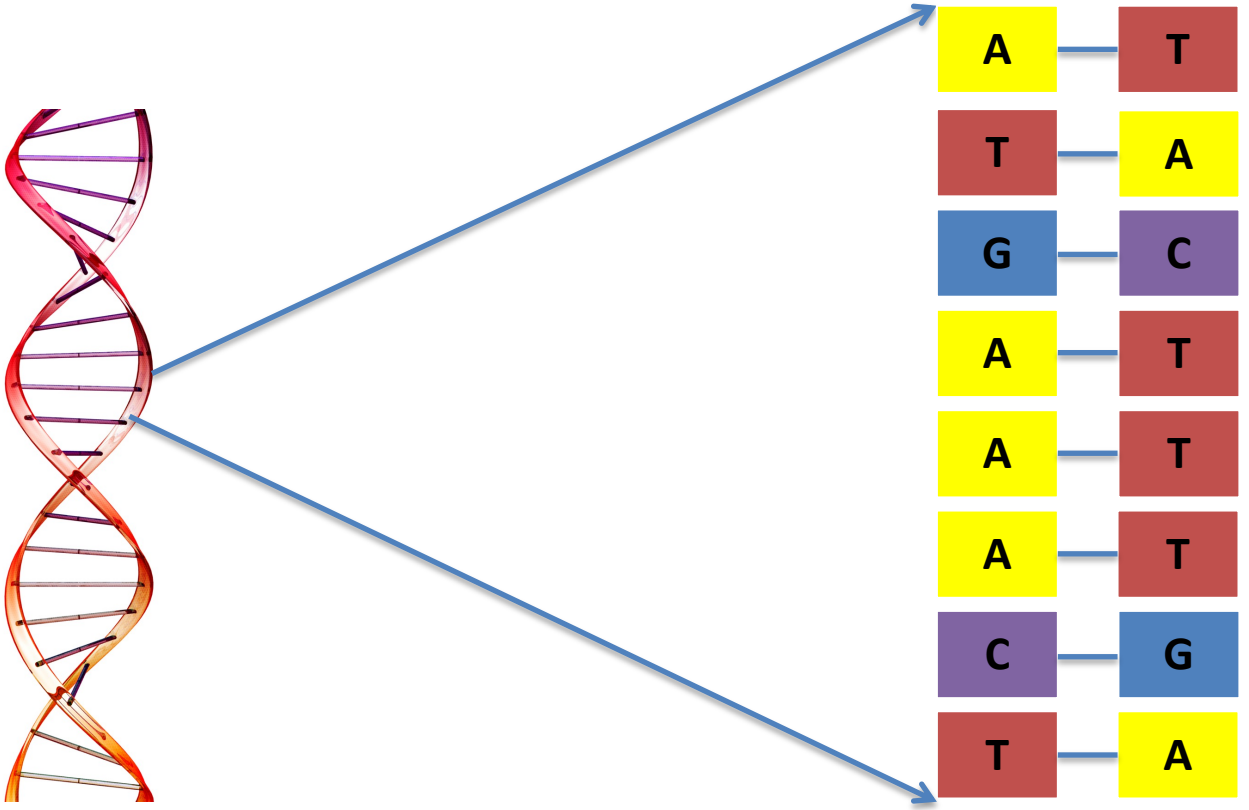
Watching vs Doing

Listen when you see this cat

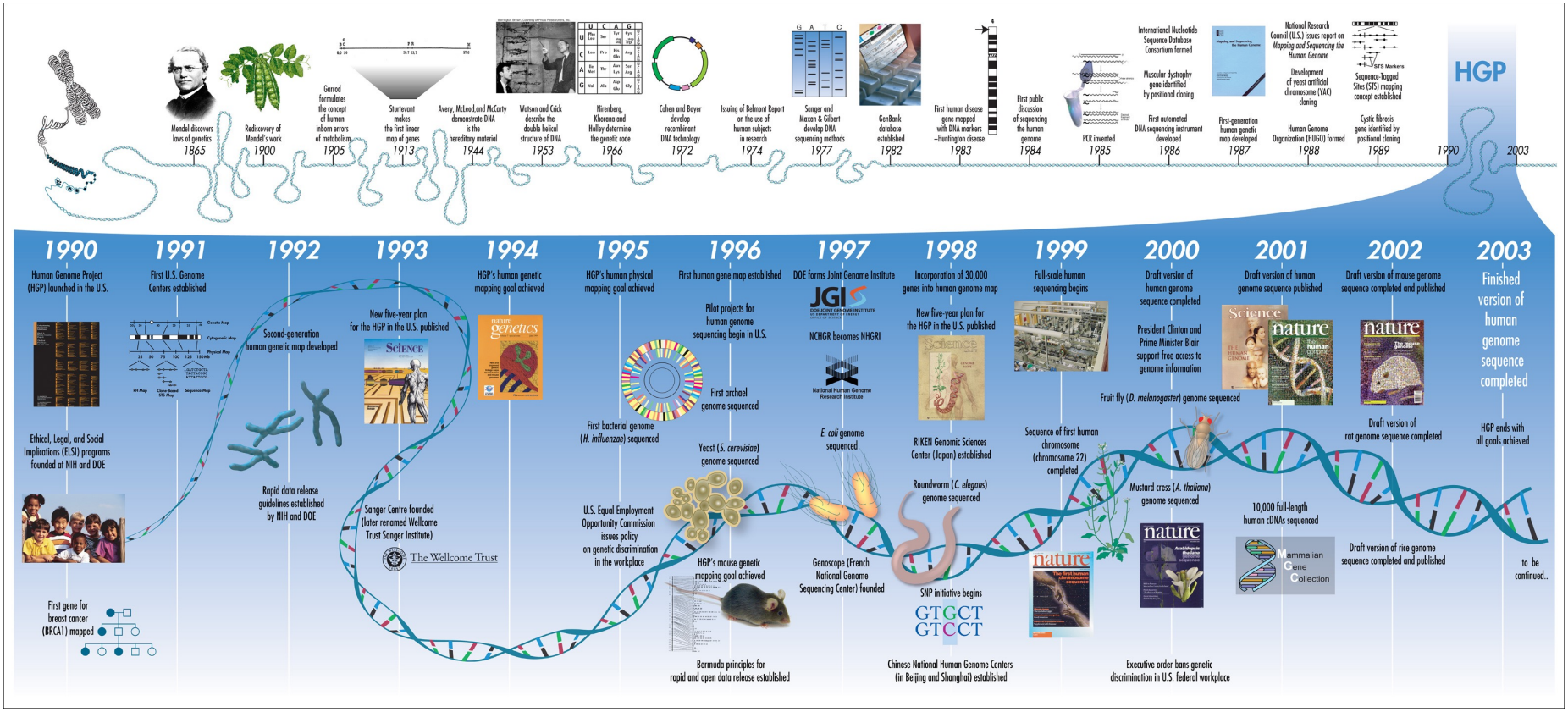
Do when you see this cat

How do we study genomes?

Sequencing – determining the order of nucleotide bases to produce a REFERENCE GENOME.



History of sequencing



History of sequencing

2004 Solexa

2007 Illumina Sequencing Technologies

2005-2013 Roche 454 Sequencer

2006-~2015 ABI SOLiD Sequencer

2010 Ion Torrent

2011 Pacific Biosciences

2014 Oxford Nanopore

"Second" Generation

"Third" Generation

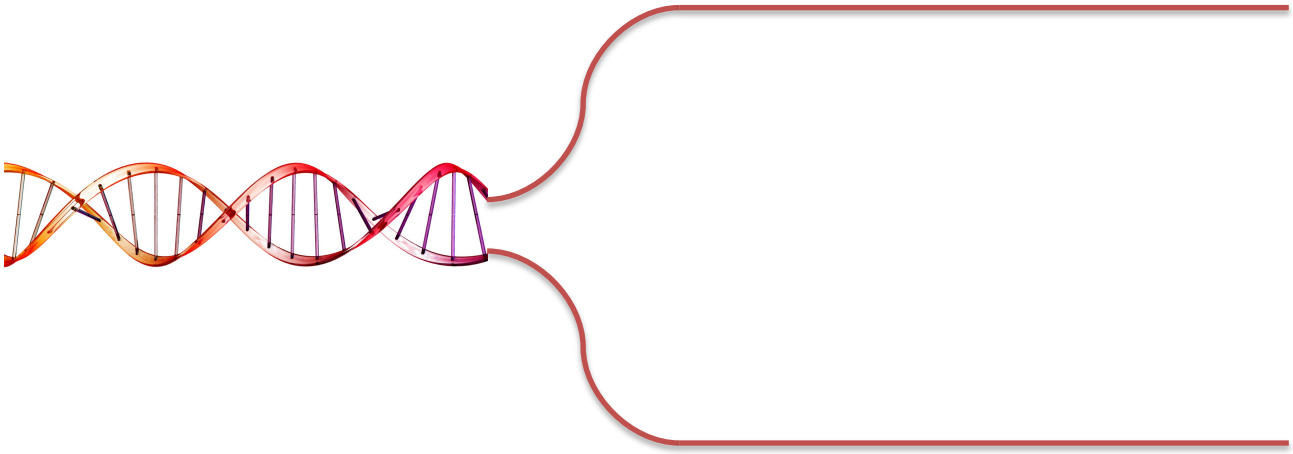
The First Sequencers

Sanger Sequencing

- Known as chain termination sequencing.
- Uses a mixture of dNTP and fluorescently labelled ddNTP.
- Polymerase is used to extend the DNA fragment, until a ddNTP is reached.

Sanger Sequencing

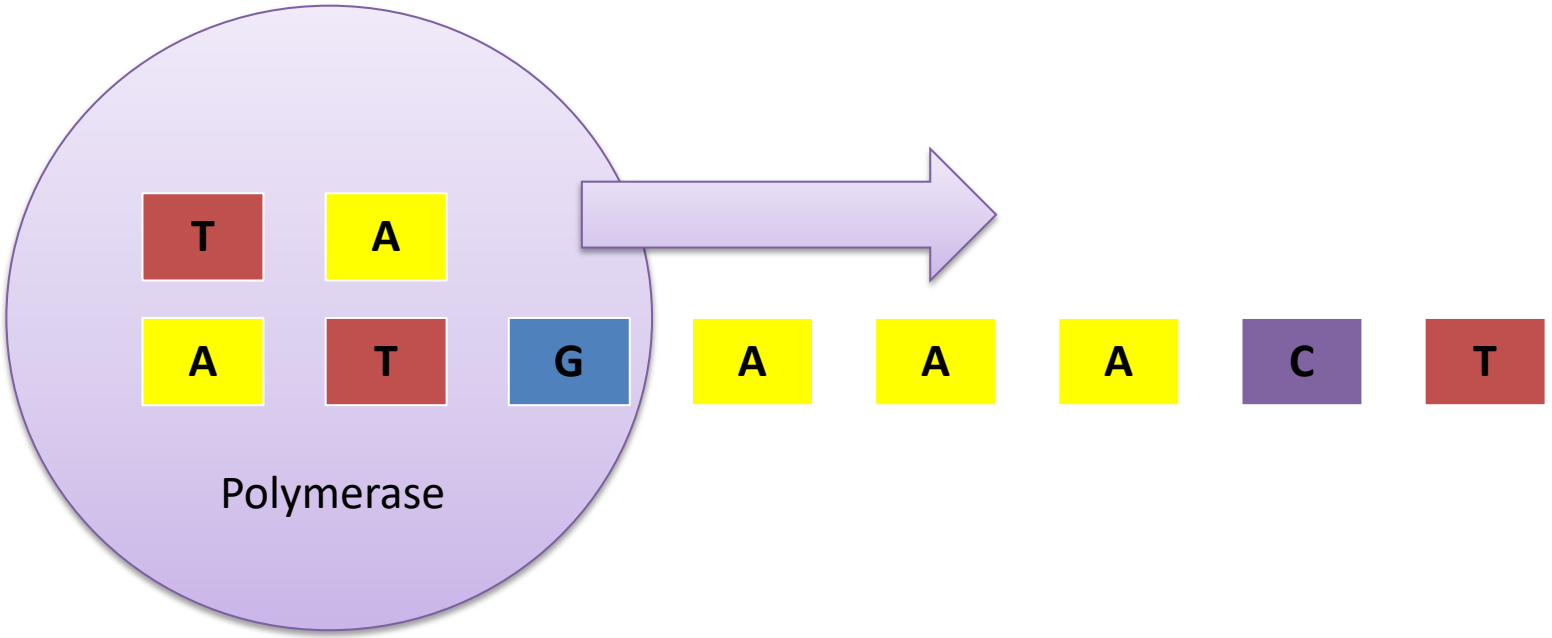
1. Unwind the DNA helix to get a single strand.



https://www.youtube.com/watch?v=KTstRrDTmWI

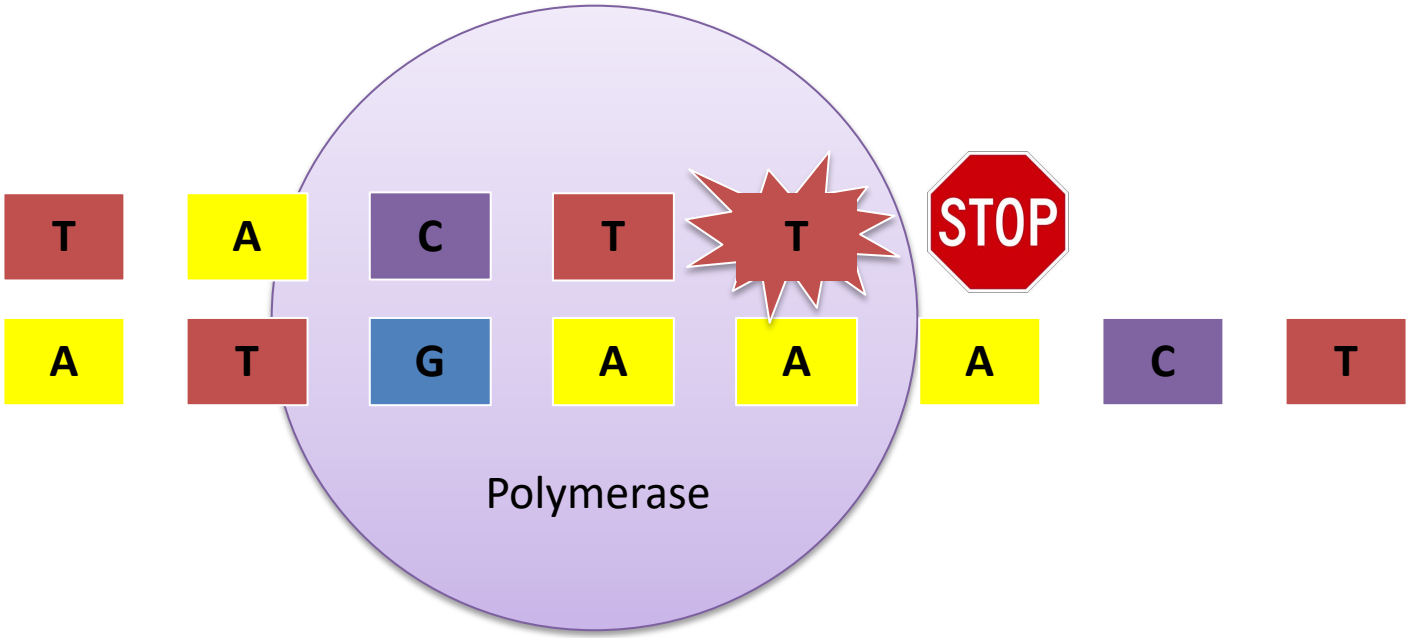
Sanger Sequencing

2. Polymerase makes a new strand based on complementary bases.



Sanger Sequencing

3. When a ddNTP is encountered, the polymerase stops.



Sanger Sequencing

4. Run the sequences on a gel

AT

ATA

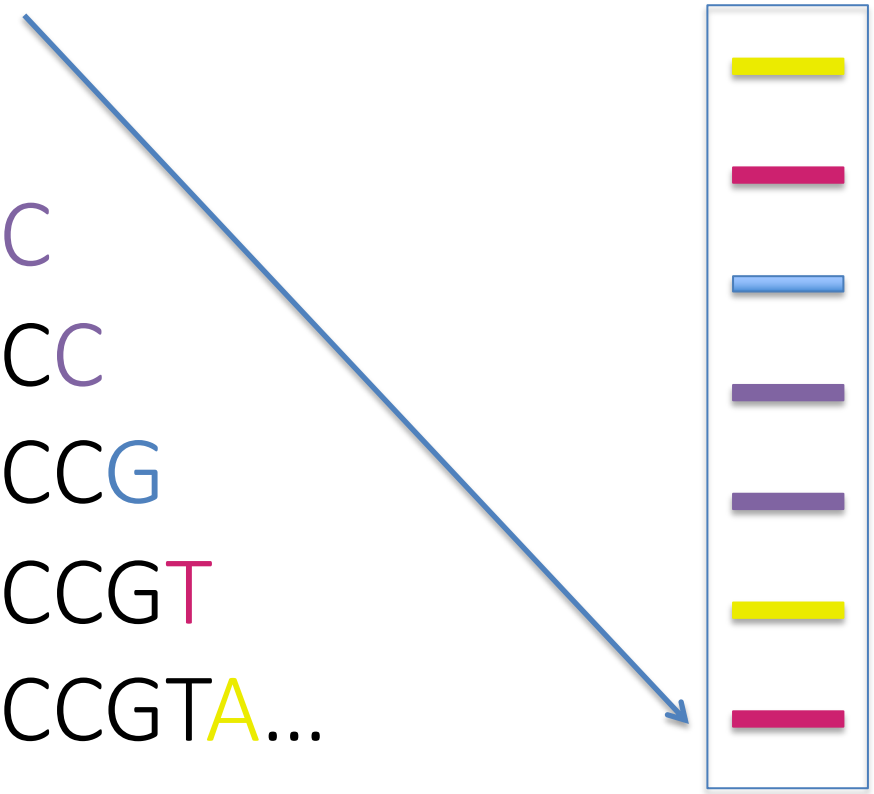
ATAC

ATACC

ATACCG

ATACCGT

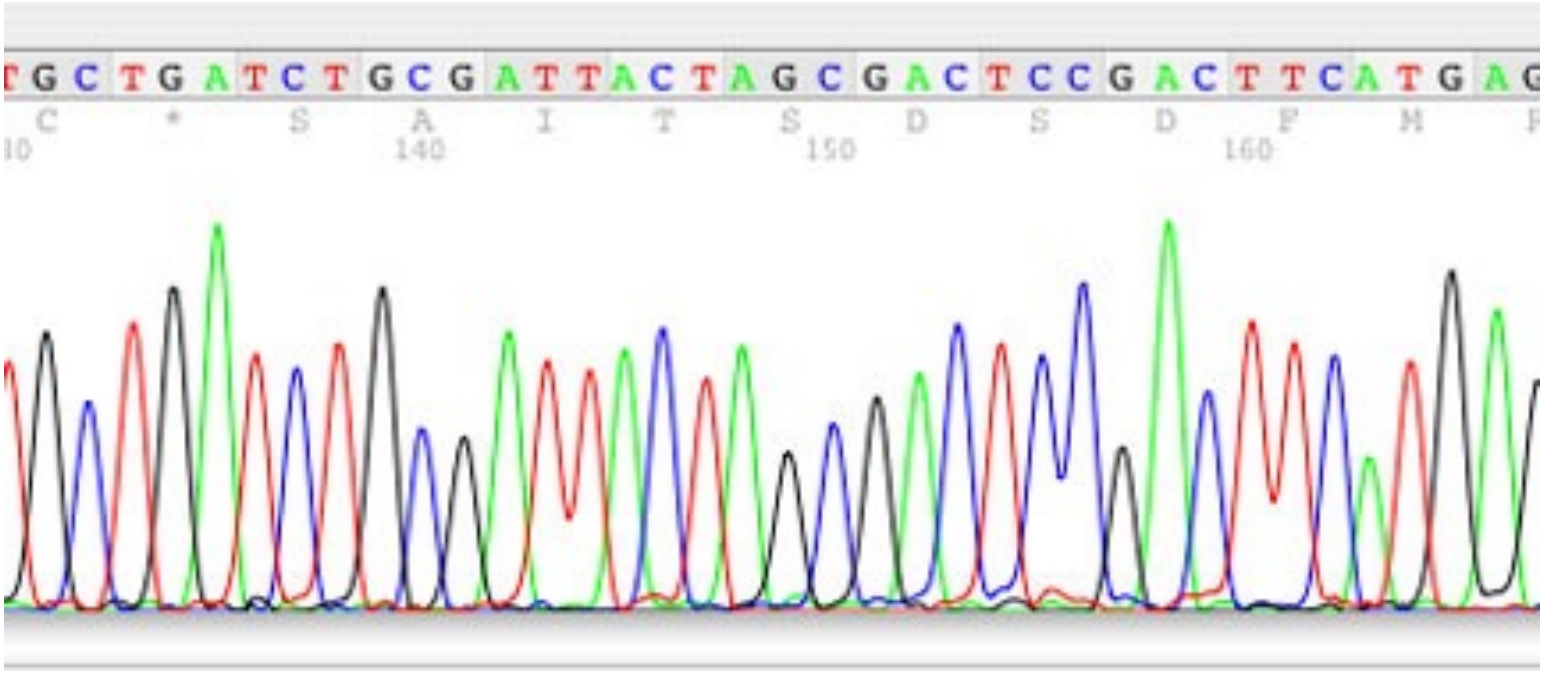
ATACCGTA…



Capillary Electrophoresis

The shorter the fragment, the further it travels

Sanger Sequencing



The First Sequencers

- Slow ✗
- Expensive ✗
- Prone to error – repetitive regions ✗
- Prone to bias – amplification ✗
- Not high throughput ✗

Next Generation Sequencers

Roche 454

Life Technologies SOLiD

Illumina

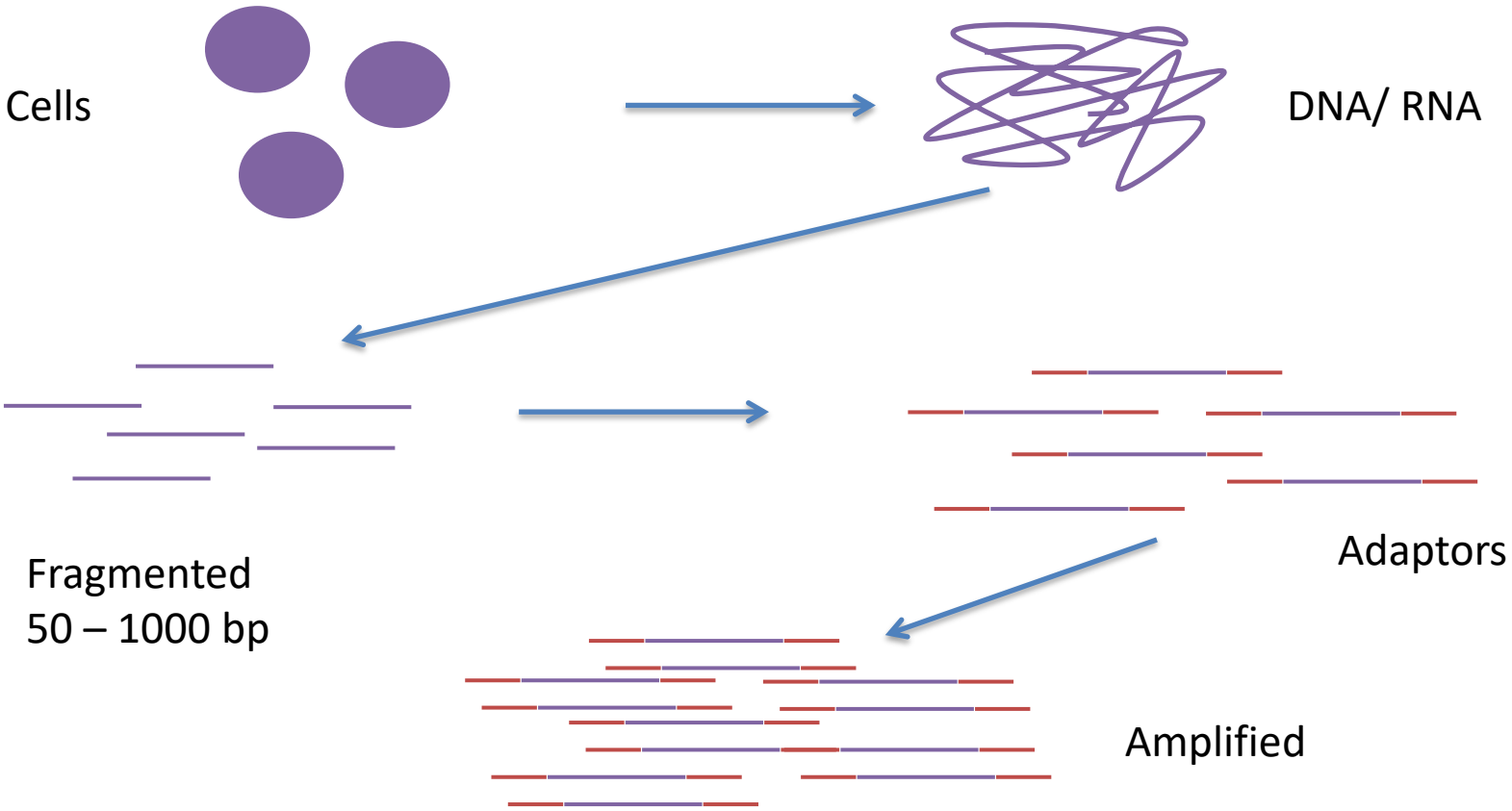
MiSeq

HiSeq

NovaSeq

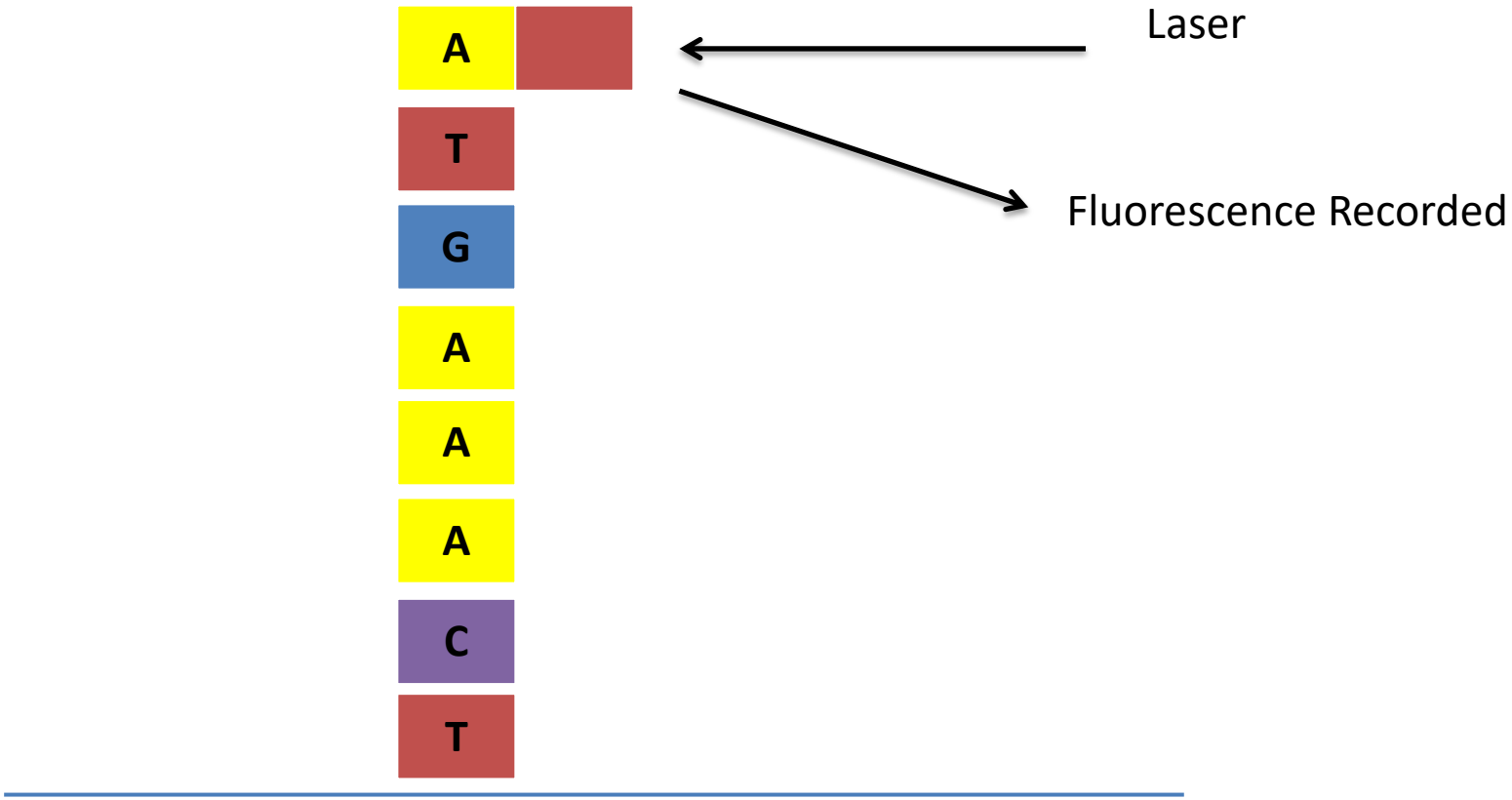
How it Works

Library Preparation



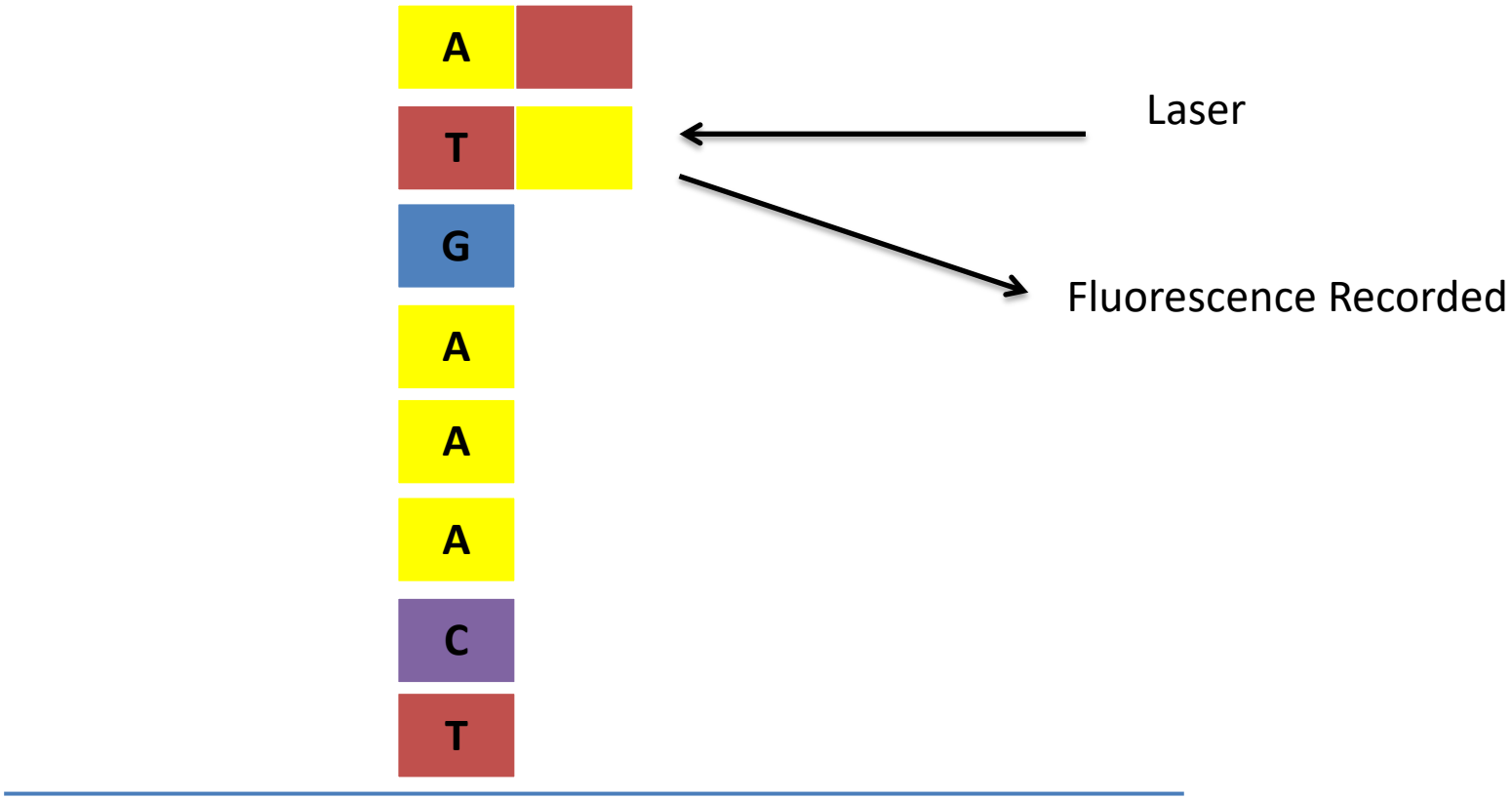
How it Works

Sequencing



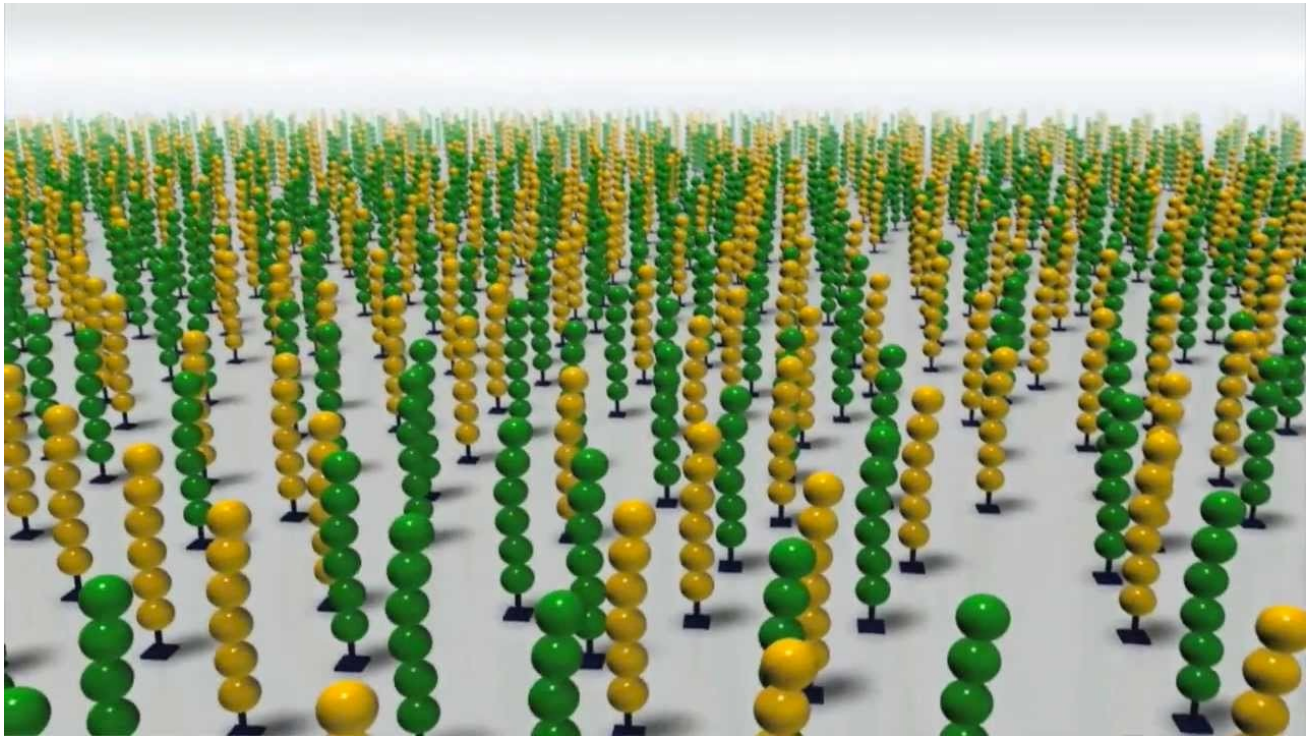
How it Works

Sequencing



How it Works

Sequencing



Millions of “clusters”

How it Works

Sequencing

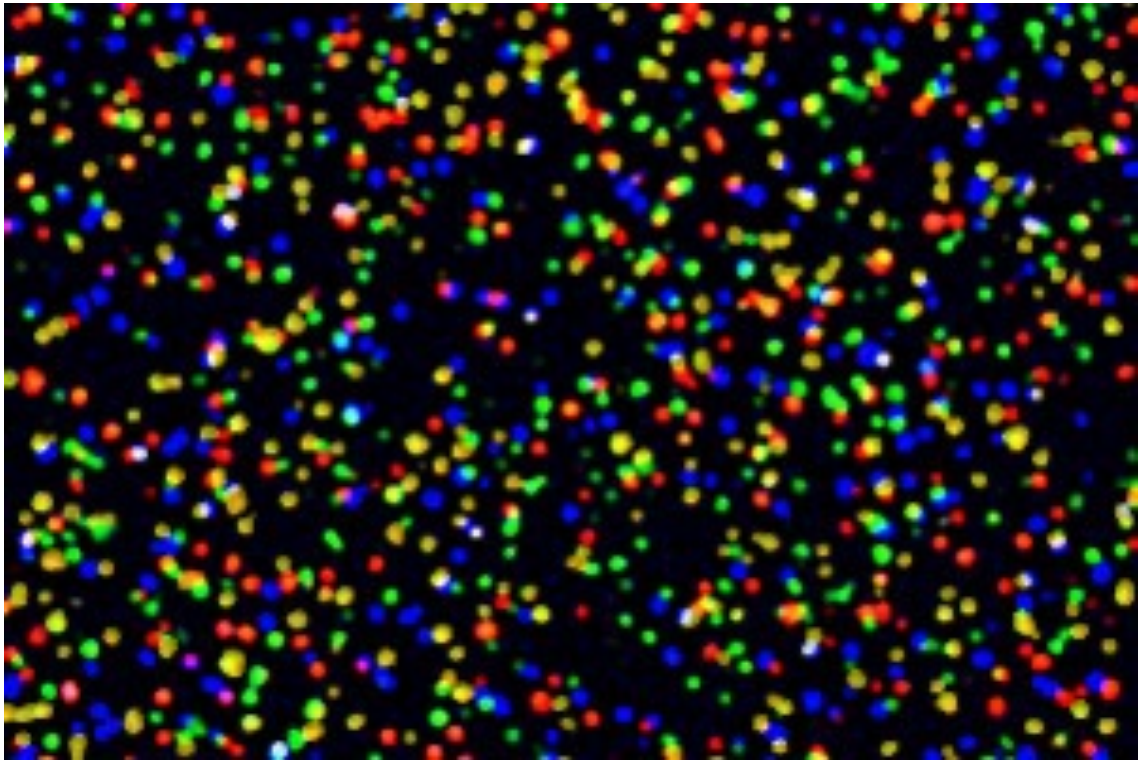


Image taken at every nucleotide incorporation

Next Generation Sequencers

- Fast ✓
- Cheap ✓
- High throughput ✓
- Prone to error – repetitive regions ✗
- Prone to bias – amplification ✗
- Restricted by length ✗

“Third” Generation Sequencers

PacBio SMRT

- Single Molecule, Real-Time Technology
- Reads are held in “Zero-mode waveguides”
- Signal read from DNA polymerase during DNA replication

“Third” Generation Sequencers

Oxford Nanopore MinION

“Third” Generation Sequencers

Oxford Nanopore MinION

“Third” Generation Sequencers

- Fast ✓
- Cheap(ish) ✓
- High throughput ✓
- Single strand ✓
- Long ✓
- Prone to error ✗

Types of Sequencing Reads

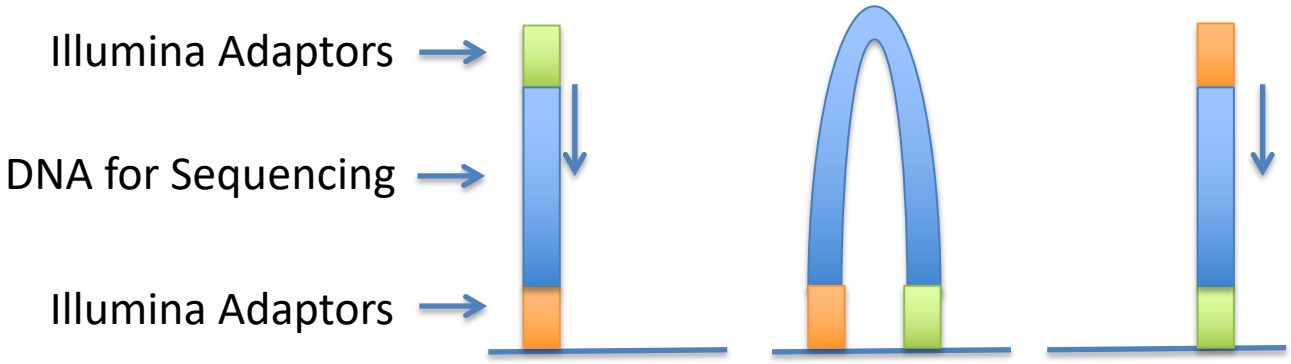
Single end (up to 300 bp)

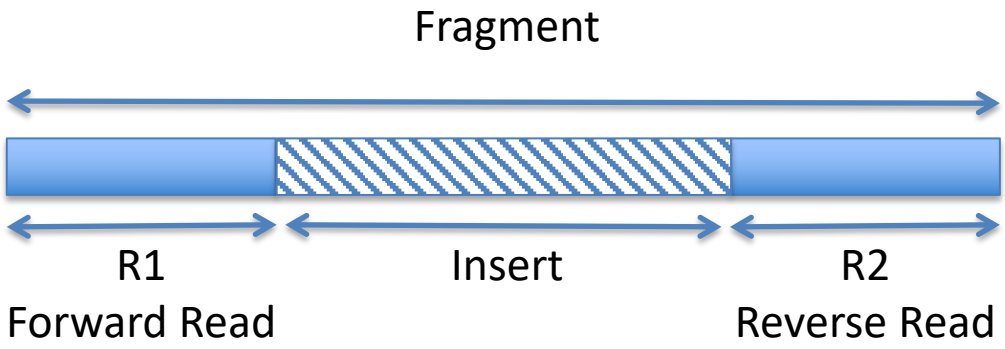
Paired end (up to 800 bp span)

Mate pair (up to 20 kbp span)

Long reads (up 1 Mbp)

Paired Reads





An example: 300 bp paired end reads with a 700 bp fragment size
R1 = 300 bp, R2 = 300 bp, Insert = 100bp

Any Questions So Far?

Looking at File Contents

Navigate back to the Sequence files that we unarchived and renamed this morning. They should be in this path:

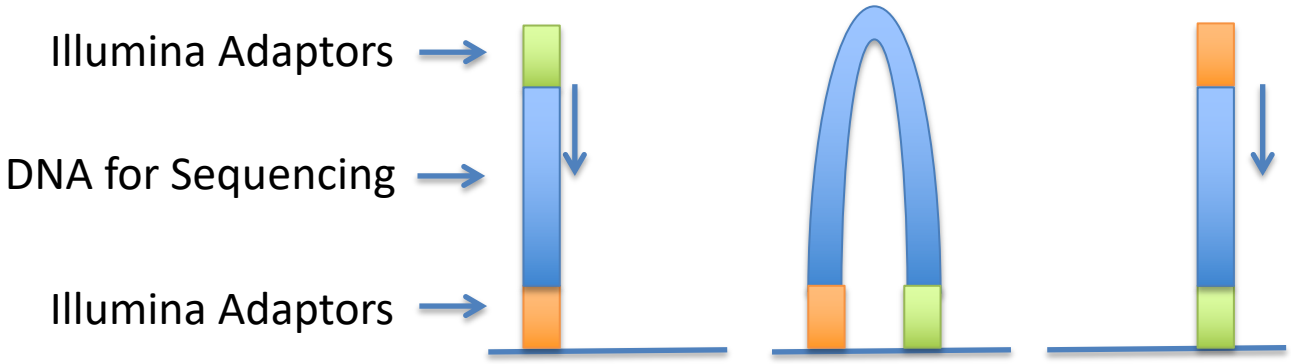
```
cd ~/workshop_materials/Unix/Sequences
```

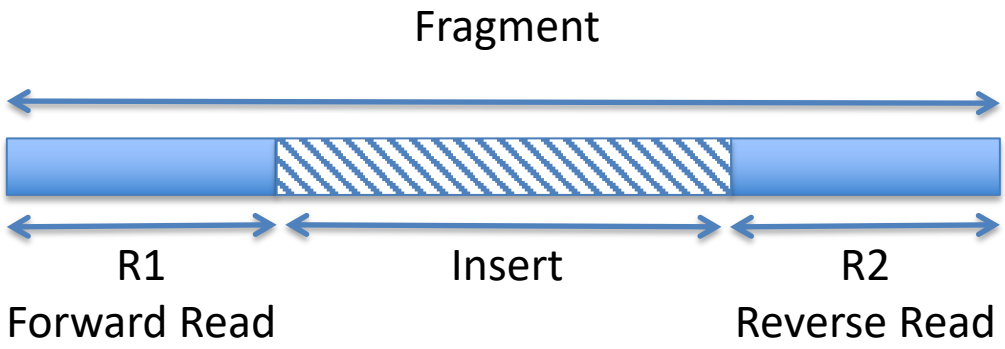
Don’t forget about tab complete whilst changing directories!

List the files to make sure that you have two sequencing files called sequence_1.fq and sequence_2.fq

If your file names are different, you will need to use these names in all of the following examples.

Paired Reads





An example: 300 bp paired end reads with a 700 bp fragment size
R1 = 300 bp, R2 = 300 bp, Insert = 100bp

Looking at File Contents

head	tail	more	less	cat
Shows the top lines of a file	Shows the bottom lines of a file	Shows the file one full screen at a time	Shows the file one full screen at a time	Shows an entire file all at once AND Used to combine files
-n specifies the number of lines (default 10)	-n specifies the number of lines (default 10)	Enter to scroll one line Space to scroll a page q to quit	Enter to scroll one line Space to scroll a page q to quit / to search n shows next	Ctrl + C to stop

Use these command line programs to look at the sequence files

```
head -n 10 sequence_1.fq
```

Let's put this to use

```
[s10ss5@login2(maxwell) Basic_Unix]$ head -n 4 sequence_1.fq
@E.-371320/1
GCTGGTCAGCCAGGATAAAACCACCACTGACCCGATGGCGGTTGTTGACTGGATCAACATGTTTGCACTGGCAGTGAACGAAGAGAACGCTGCTGGCGGTCGCGTGGTGACTGCG
CCGACTAACGGTGCGTGCGGGATTATCCCGGCAGTTCTGGCGTACTACGACAAGTTTATCCGCGAAGTGAACGCTAACTCACTGGCTCGTTACCTGCTGGTAGCCAGCGCCATTG
GTACTCTTTATAAGATGAAC
+
????,BBBDDD<BD?<FGFFGFCFFIIHIHIHIGIIDHGIIDIIIIIGIFHHHIIHHHIHIIII-HHDIIIHIIHIIFIHHHFIGIHHHH:HGI=GHHHHFGIBEGHHHHHBFH=
HHHFEEFGGEGHGDEBFG?FIFEG:8EBEEDG:GEEBGGGGGGGG(GEGGGGE?FECGFFCGFDFGFGEFEE??GG?GCGE*FG?:6E/FGCCEEHC:FGF-:G?6GCGGAAGGG
6G)EGEC:GGFE'G:GC?G8
[s10ss5@login2(maxwell) Basic_Unix]$
```

Fastq File Format:

- Header
- Sequence
- Second Header (often +)
- Phred Quality Score

Lot's of analysis software like paired reads to be in the same order

CHALLENGE!

Use head and tail to check that the five bottom and top headers are in the same order in sequence_1.fq and sequence_2.fq

(You may need to use the –n option to show more lines)

Challenge

```
[s10ss5@login1(maxwell) Basic_Unix]$ head -n 20 sequence_1.fq
@E.-371320/1
GCGCCGACTAACGGTGCGTGCGGGATTATCCCGGCAGTTCTGGCGTACTACGACAAGTTTATCCGCGAAGTGAA
CCATTGGTACTCTTTATAAGATGAAC
+
????,BBBDDD<BD?<FGFFGFCFFIIHIHIHIGIIDHGIIDIIIIIGIFHHHIIHHHIHIIII-HHDIIIHII
FH=HHHFEEFGGEGHGDEBFG?FIFEG:8EBEEDG:GEEBGGGGGGGG(GEGGGGE?FECGFFCGFDFGFGEFE
@E.-371318/1
GCGCGAAAGTGATCGCCAGCGCCTATTCCGATCTCGGTTTCGACGTAGATTTAAGCCCGATGTTCTCTACACCT
CGTAGTGGGCGCATCCTCACTGGCTG
+
??AA?BABDDD?BDDDCGFFFGIIHHIIIHIGIIDIHHIFEHFGIIHIHIFIHBICHIIHGHHBFIAIBIIFHF
@E.-371316/1
ATTGAAACTAAAGGCTCCGACGCCAGAAATAGCCTGCAACCGCTGTGCGGAGATTAACCGTTCAATAATGTCCA
GGATAGAGCAGATGGCCTTGCTTATC
+
?????BBADDDDBDDDG;GGCGEIFHHHIIGHHHCGGHHIHIFHIAIICHHEFHIH5HIICDFHHGIIIHI>EI
@E.-371314/1
TGACGTGTGACCCGGTCGCCGGACAGGTACAGGTGCCATGCATCGAGCGTAACGCCATTGCGGAAGTAAAAGCG
GCCGCGCGTCTGCCTCGATAAAGTTA
+
=????BBBDDBBD6<@FFBGFFFHIIHHFIHHHHIEIHIIGIIIIIIFFIIHH?IDIFHICGGIHAIHHIAIHI
@E.-371312/1
CACCGATTTTCGGCTGGTTCCACCACTGCAGGGCAGGGTAGAGGGAAACCGGTTCTTTGCACTCATATTGCAGA
GGGTAATTCATGCACCAGCGGAATAT
+
?A?A?BABDDDDDDDDAGFGFFIIC1H@IEHECI.IHGH=IGHHIGIII@HFFHIGHFHHHHHHGHFIIIIHIG
H.FHHFGGHGHH=GD:GEGEEG,GGEEGBGGGG*GEEEGCGGGGBAGGFG8GG:EEEGEGGGEEG;GEEGEAEE
?FGEGEE*CGG?FGGEG*GGE?GEGD
[s10ss5@login1(maxwell) Basic_Unix]$
```

```
[s10ss5@login1(maxwell) Basic_Unix]$ head -n 20 sequence_2.fq
@E.-371320/2
CGGCGCAGCGCCATACGTGCGGCGTTCACCACTATTACTGCCGCAATGGCGTTACGCTCGATGCATGGCACCTGTACCTGTCC
TGTGCTCCCTGGCGATTTCCGGCGCG
+
?A???BBBDDDDDD@DFFFDCCIEI-IFEIGHIEFHIEHHIFFGDHIH-FIDHHGHHFI-FEHBGIHGCB=7HGHH*FF-HF
HHHHGHHHGHEGG>@DF?@=C;BGF/EEDH*.G'GEGGEEGG;E(CGEE(GG8(EG/(;E(?EG?EC6CGE:EGEE:E?FE)
@E.-371318/2
TTTCTGCCAGCGTGGCTGCATTAATCATGATGCTGGCTTATCAGATTAAGTACGTCGCGCACACTGTCGAGCATAGGTGTACG
CTTGCAGGAAGGCGTAAACCTGCGGC
+
?????BAB+DDDDDEDGA;F8FHCHIF>CH-ICHIHCCICHIIEHI-G+>I5GHDDHHFHCFHF-HICHHFGHEHCAHHBIGB
EHHFEF@FFGHHHFEH?-FF9FFEEGG2FFE*EDEGGFDCC@ACGF@(FGEGE;A((?2FGEFC(?AGGF;?GEEG6GGGGCI
@E.-371316/2
CAGAACTACGGTCCGTTGTACAAGCGTTAATACATAACTCCCTAATGTAATGCCCGGTCGCTCCGGCTACCGGGCCAGAACAA
TATCCGGCTATATTGCAGGAGCGATT
+
=???ABBBBBE<DB5DFEGGGGIIIC>-HFHHHD9HHI@,F@@IFCHHFHHCFHIEI/FFF-=8DBHIAI*GI?FFECAHHII
E=2H)FDHC+D(HG*HH+?G0GGAG4F;EG:GH?G(F(GG*AG#(GGG(:G/(EAGF(.AG<=GC(G6(G**4(.8#..(>G/
@E.-371314/2
CTGATCGAGCGCATCCTGGCAGGTTTCGACACAGGGCGACCCCTGAACGGCATGAATGCGACGAATAAGATTCAGTGCATCGA
CCCTTAAAGATCAGGCGATAAGGGTA
+
?,???BBBDBDEDDDDFEGF8CCCAEEHCFHHGAF@H+IHEHH.FDHHGHAHIEH0H'HHFEHHFII5IHHAHCCIFIHIHHH
HHCGBHCFDG6HFF,EGC+G0GE4GGEGGGEEGF,,EBEGGEGE8GGG8GF=G6AEGGEFG(6FF;GAEA:CG*G0GF(.EG
@E.-371312/2
TACCCGCGCACGGACGATCTCAACGGTTATCAGCAGGAAGGTTTTGGTCCGATGGATCGCACCGTCACGCCGCAGGGCCGTCC
AGGGCAAATCGCGTCCTAACCTGACC
+
=??A?BBBEDDBD<@BFGEEEDF;IIFHH;AIGGH>IFHIIECGFFHH>8IFFDGDAH)HIAGEHCFFHICF=FIGHH?IH5
HHC?GFGCHHGEFEEEFGGFEFGEEDAG4GGGG;,EFGFEED*:GEEEAC2G*G?E?E?CC<A;CCE02CFFGGGE>E0EG;/
GAG(*AA86:G8CEE*/8F*C:GEE:
[s10ss5@login1(maxwell) Basic_Unix]$
```

Sequencing Stats

How many reads?
Count the number of lines

```
$ wc -l sequence_1.fq
```

This is the letter l

742640 lines THEREFORE 185660 reads

Are there the same number of reverse reads?

How about just counting the header lines?
Each header line starts @E

```
$ grep -c "@E" sequence_1.fq
```

219153

BUT the numbers from the two programs don't match?!

```
$ grep "@E" sequence_1.fq
```

How about with this

```
$ grep -c "^@E" sequence_1.fq
```

185660

^ matches this pattern at the start of the line

So You've Got Some Sequencing Data – Now What?

The first thing you should do is check the quality of the data…

Fastq Format

```
[s10ss5@login2(maxwell) Basic_Unix]$ head -n 4 sequence_1.fq
@E.-371320/1
GCTGGTCAGCCAGGATAAAACCACCACTGACCCGATGGCGGTTGTTGACTGGATCAACATGTTTGCACTGGCAGTGAACGAAGAGAACGCTGCTGGCGGTCGCGTGGTGACTGCG
CCGACTAACGGTGCGTGCGGGATTATCCCGGCAGTTCTGGCGTACTACGACAAGTTTATCCGCGAAGTGAACGCTAACTCACTGGCTCGTTACCTGCTGGTAGCCAGCGCCATTG
GTACTCTTTATAAGATGAAC
+
????,BBBDDD<BD?<FGFFGFCFFIIHIHIHIGIIDHGIIDIIIIIGIFHHHIIHHHIHIIII-HHDIIIHIIHIIFIHHHFIGIHHHH:HGI=GHHHHFGIBEGHHHHHBFH=
HHHFEEFGGEGHGDEBFG?FIFEG:8EBEEDG:GEEBGGGGGGGG(GEGGGGE?FECGFFCGFDFGFGEFEE??GG?GCGE*FG?:6E/FGCCEEHC:FGF-:G?6GCGGAAGGG
6G)EGEC:GGFE'G:GC?G8
[s10ss5@login2(maxwell) Basic_Unix]$
```

Fastq File Format:

Header

Sequence

Second Header (often +)

Phred Quality Score

This line contains the Phred Quality Score for every base in the sequence

Read Quality – Base Calling

Sequencing

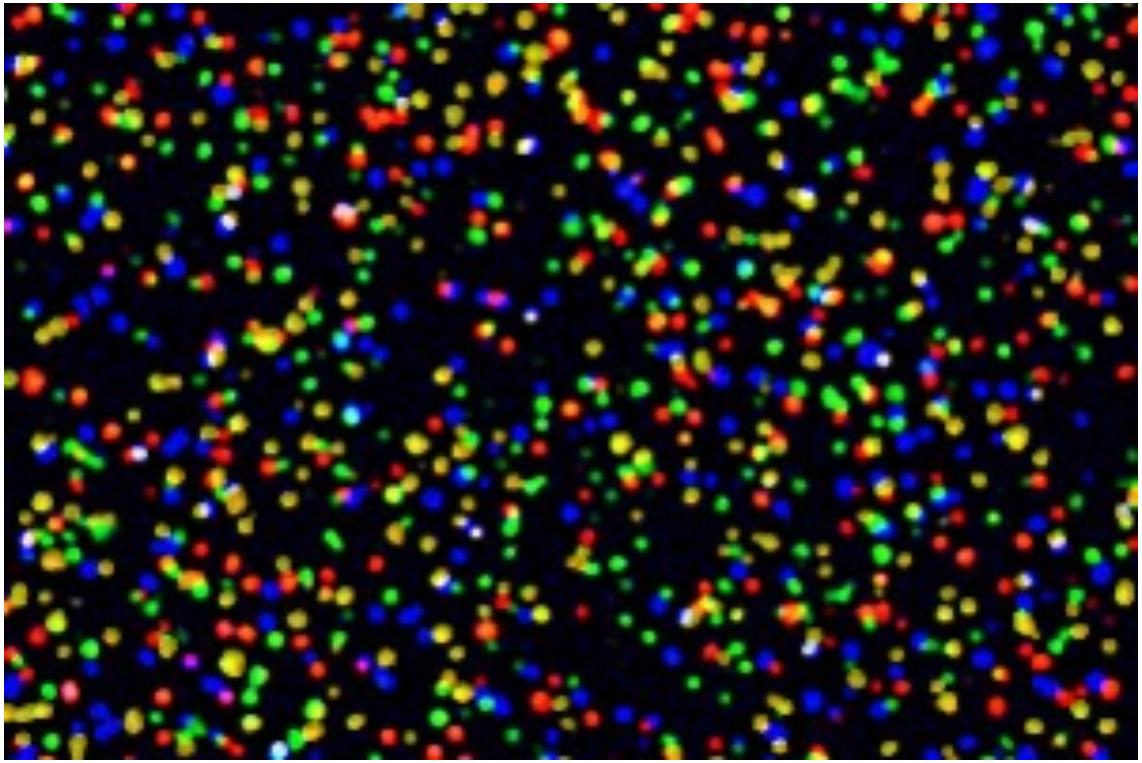
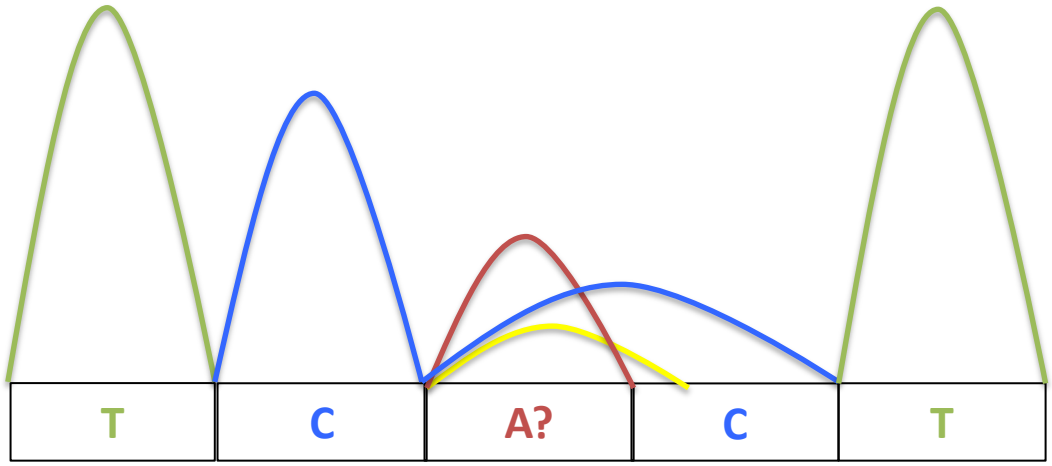


Image taken at every nucleotide incorporation

Read Quality – Base Calling



Base pair sequences are called based on the fluorescence signal.

Quality scores are associated with each base call depending on how well it can indiscriminately determine the base.

Why are some bases poorer quality?

The base caller is unable to determine the base when:

- Cluster fluorescence is unclear.
  - Dephasing of cluster.
  - Reagent degradation.
- Clusters overlap or are too close together.
- Lack of sequence diversity.

Read Quality – Phred Score

Phred 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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

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

ASCII Encoding

- Each number is converted to one symbol:

40:@	90:Z	141:a
41:A	91:[	142:b
42:B	92:\	143:c
43:C	93:]	144:d
44:D	94:^	145:e
45:E	95:_	146:f
… :…	… :…	… :…

Different Phred Scores

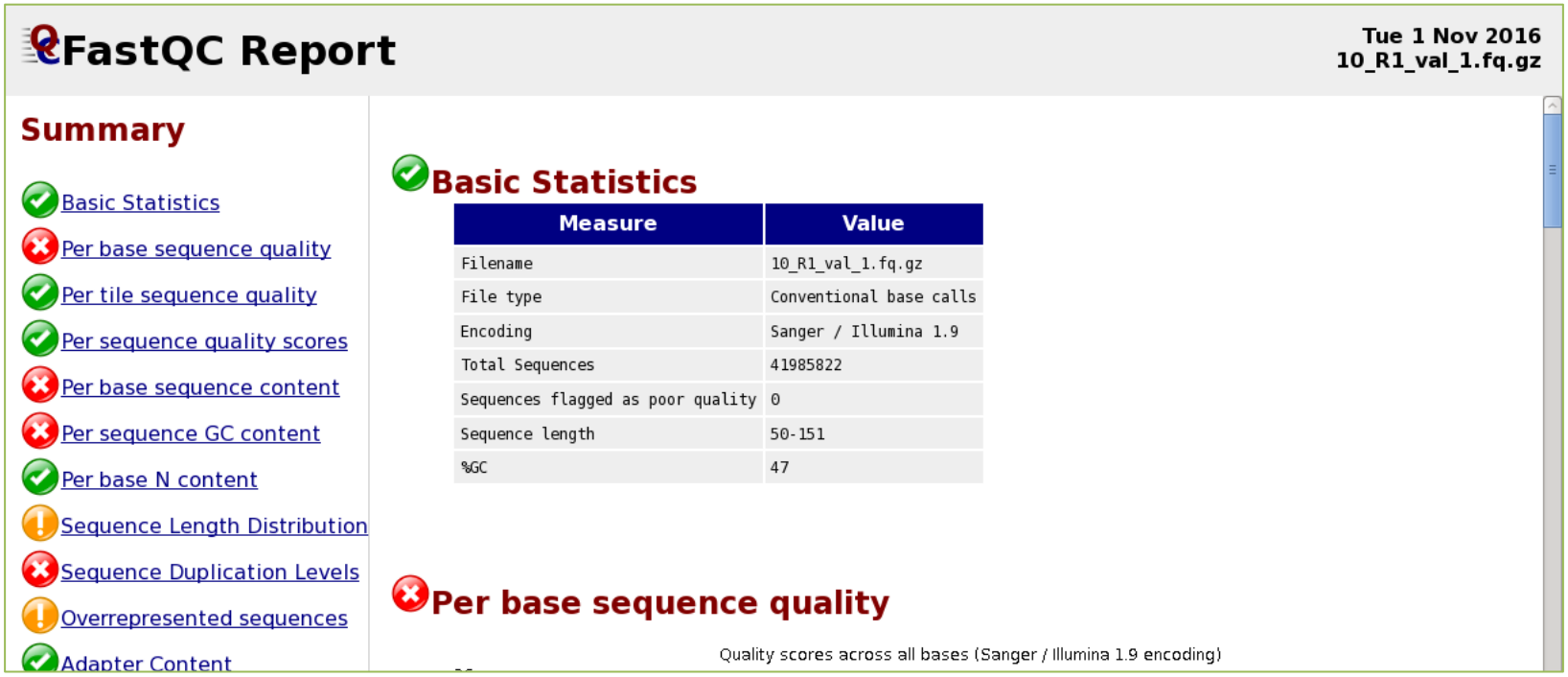
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |        |                                 |                     |
33                         59   64       73                               104                   126
 0.........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................41
 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, 41)
    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)
```

Why Do We Care?

- Low quality reads, contamination and adaptors introduce errors into data.
- Filtering and trimming these sequences may help to improve downstream analysis.
- Filtering isn't always needed. Some programs take quality into account.
- HOWEVER a visualisation of data quality should be carried out at the beginning of ANY analysis.

FastQC

- How do we easily look at the quality of >100000 sequences?
- FastQC is software that does this for us.



FastQC Report

Tue 1 Nov 2016
10_R1_val_1.fq.gz

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

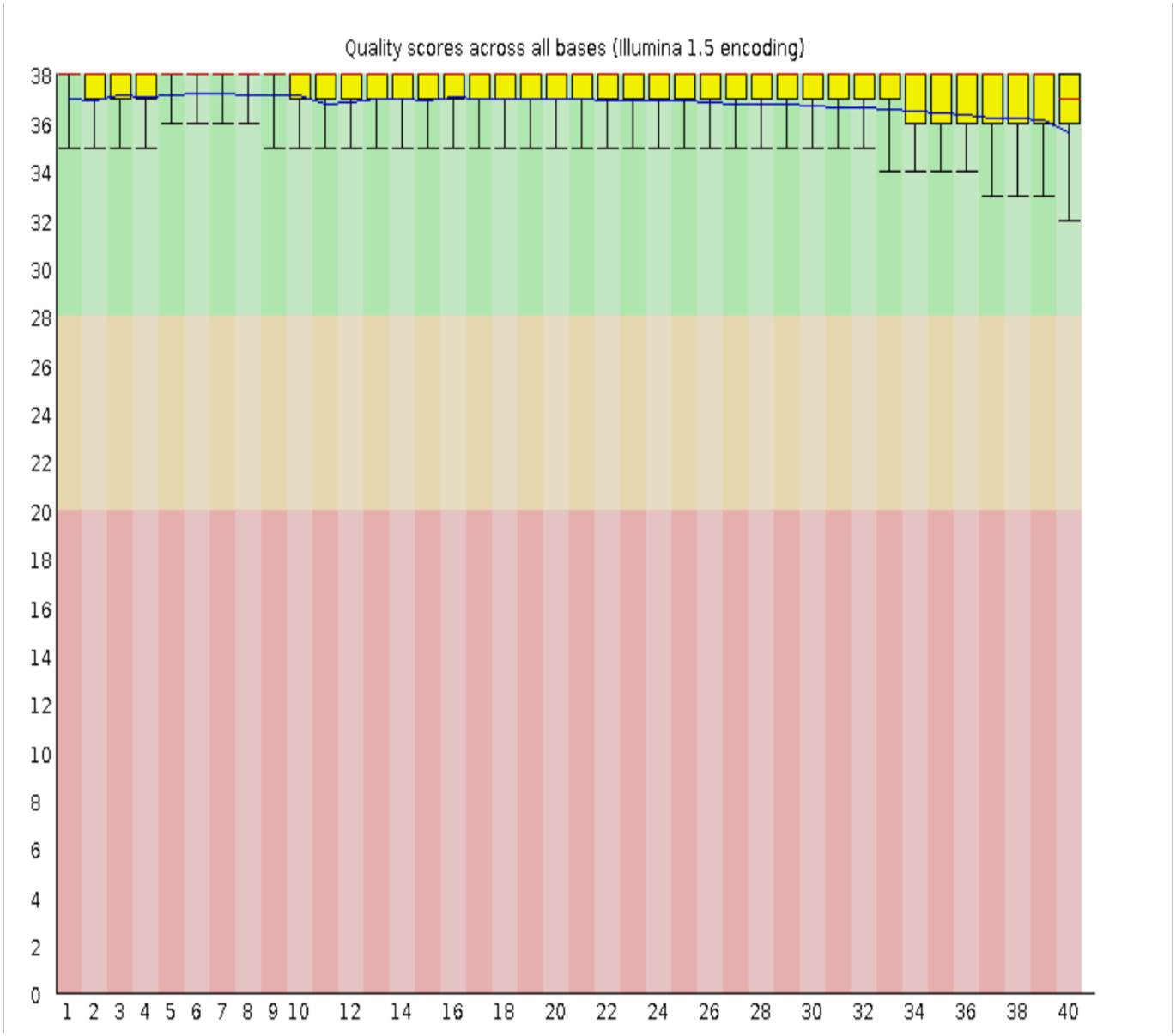
Basic Statistics

Measure	Value
Filename	10_R1_val_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	41985822
Sequences flagged as poor quality	0
Sequence length	50-151
%GC	47

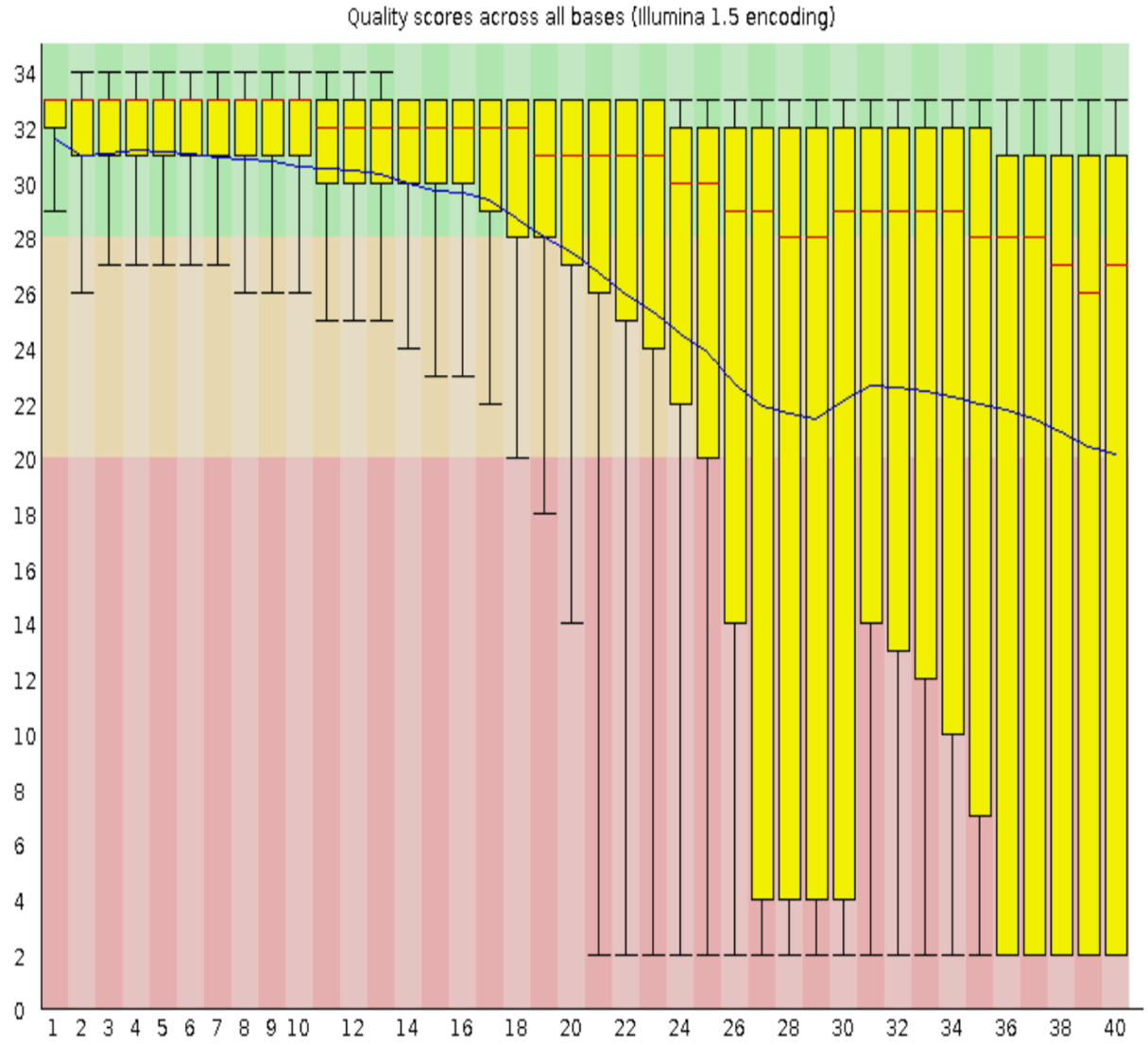
Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

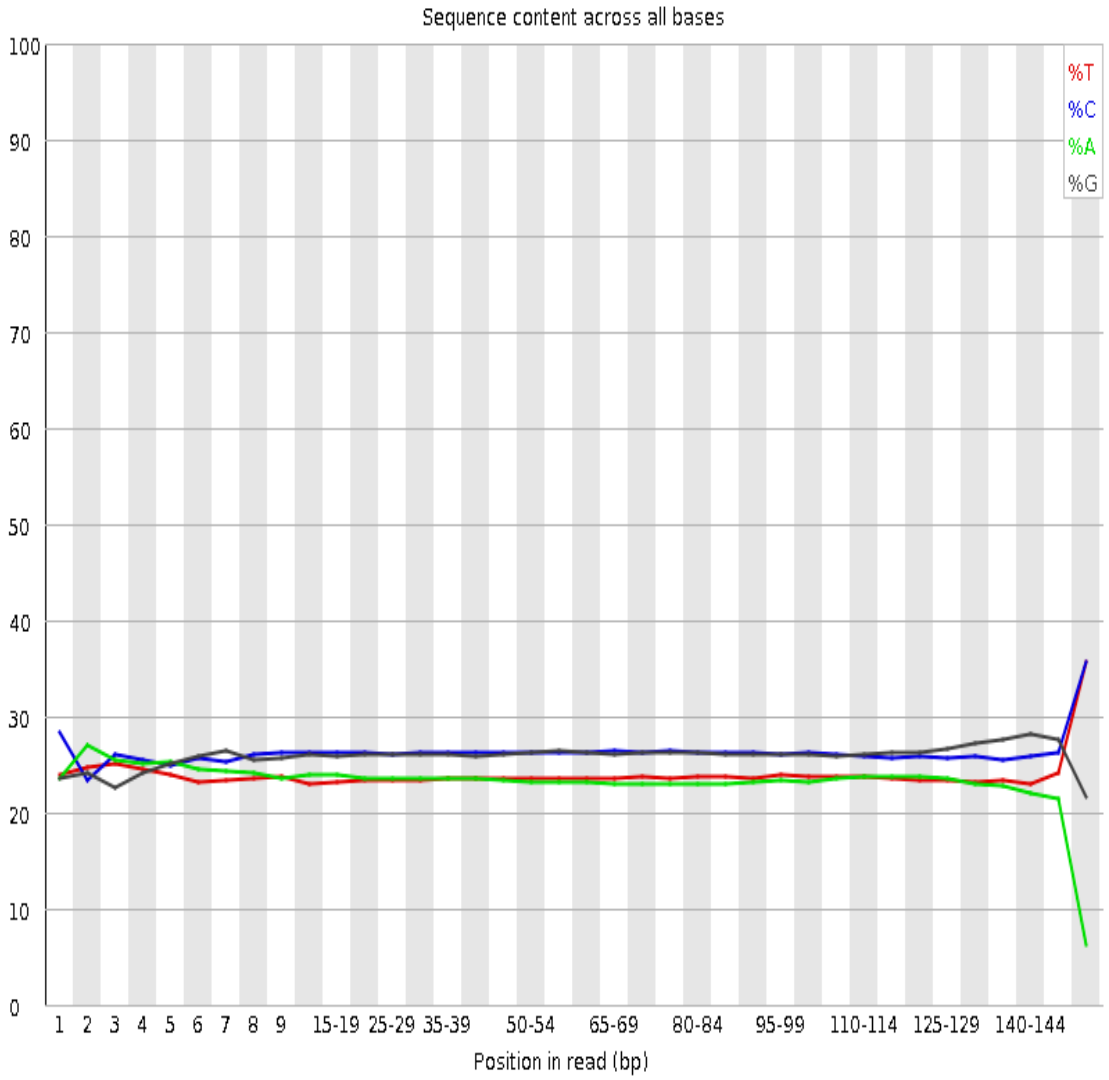
Per Base Sequence Quality



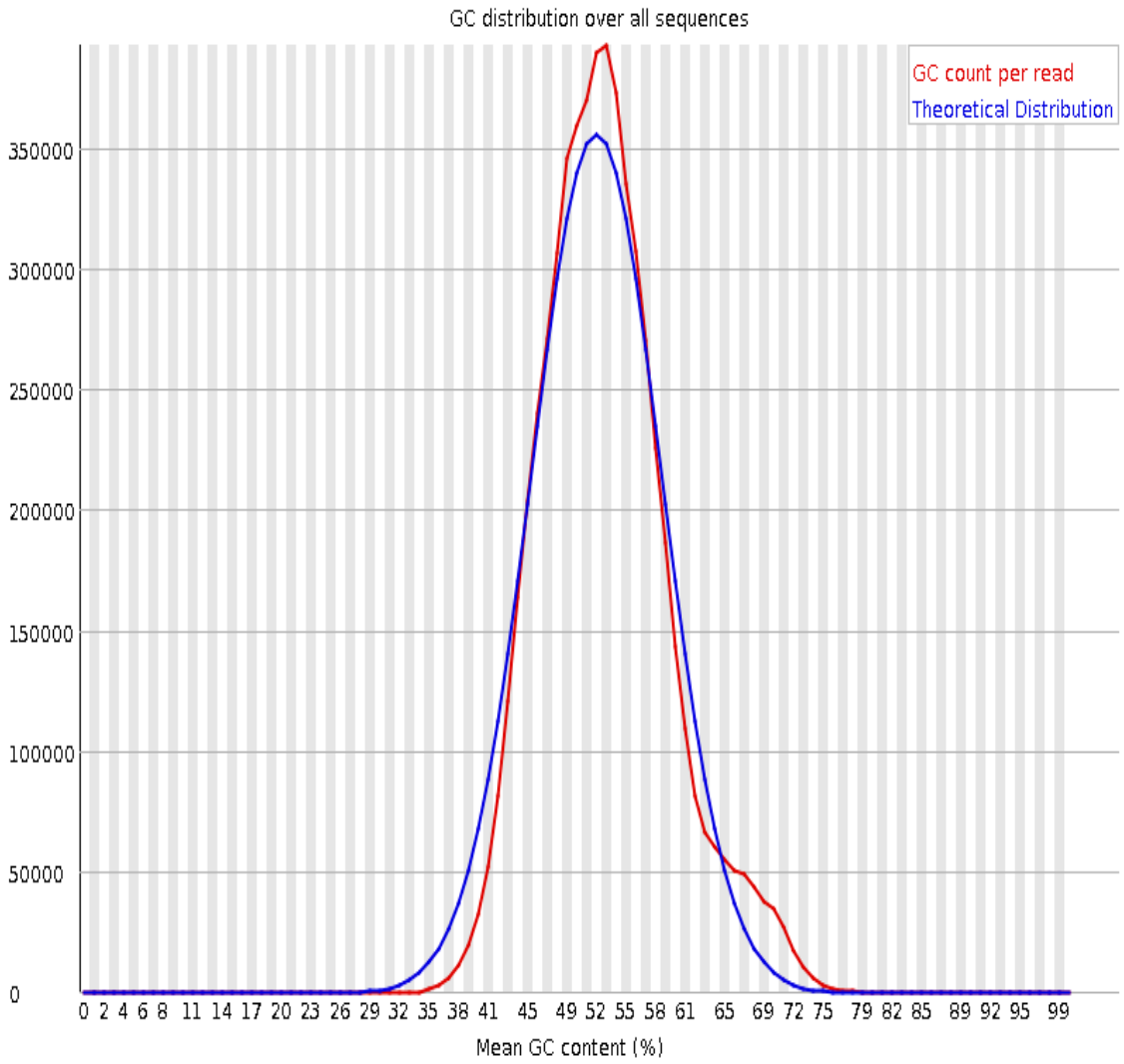
Per Base Sequence Quality



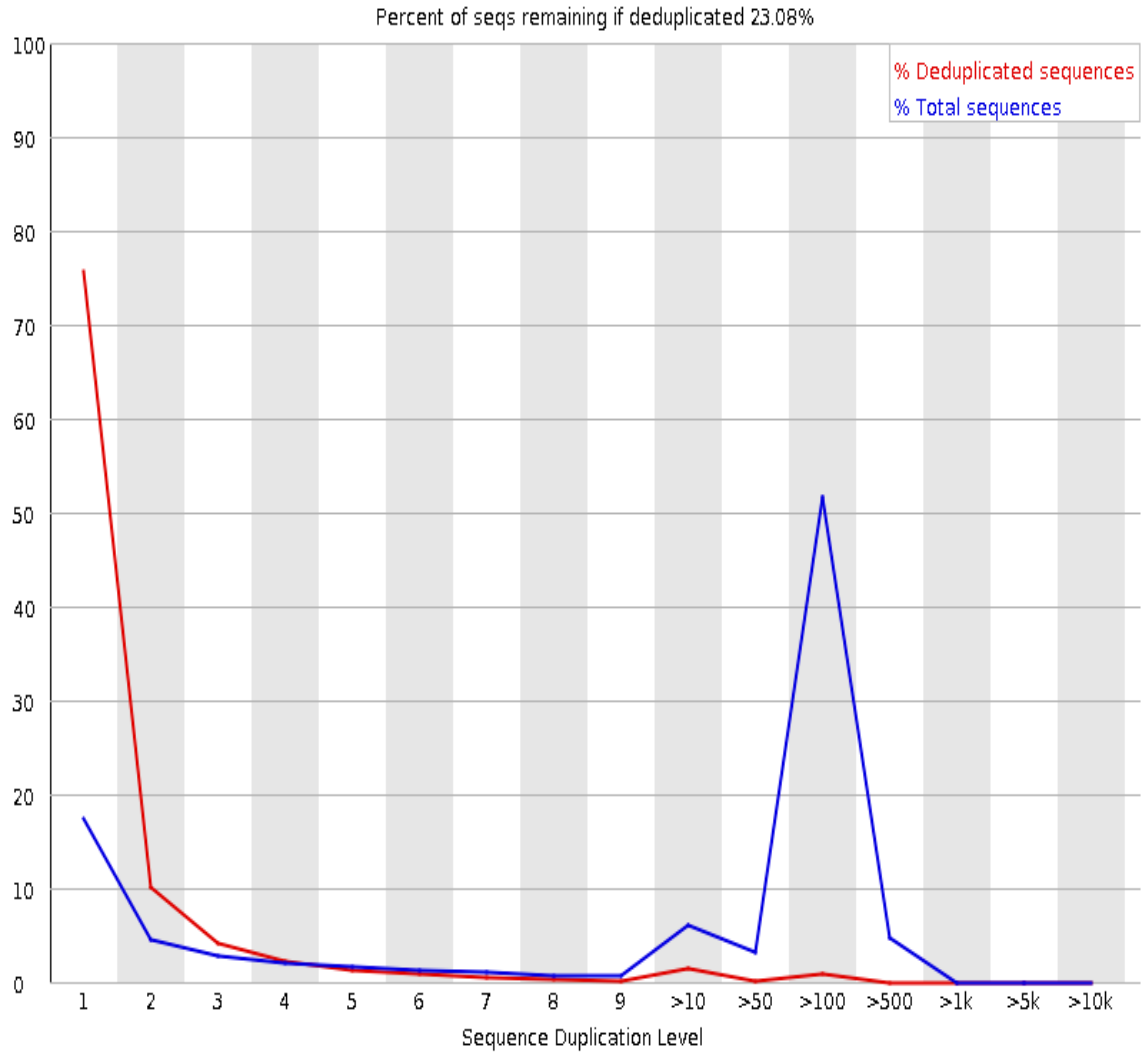
Per Base Sequence Content



Per Sequence GC Content



Sequence Duplication Levels

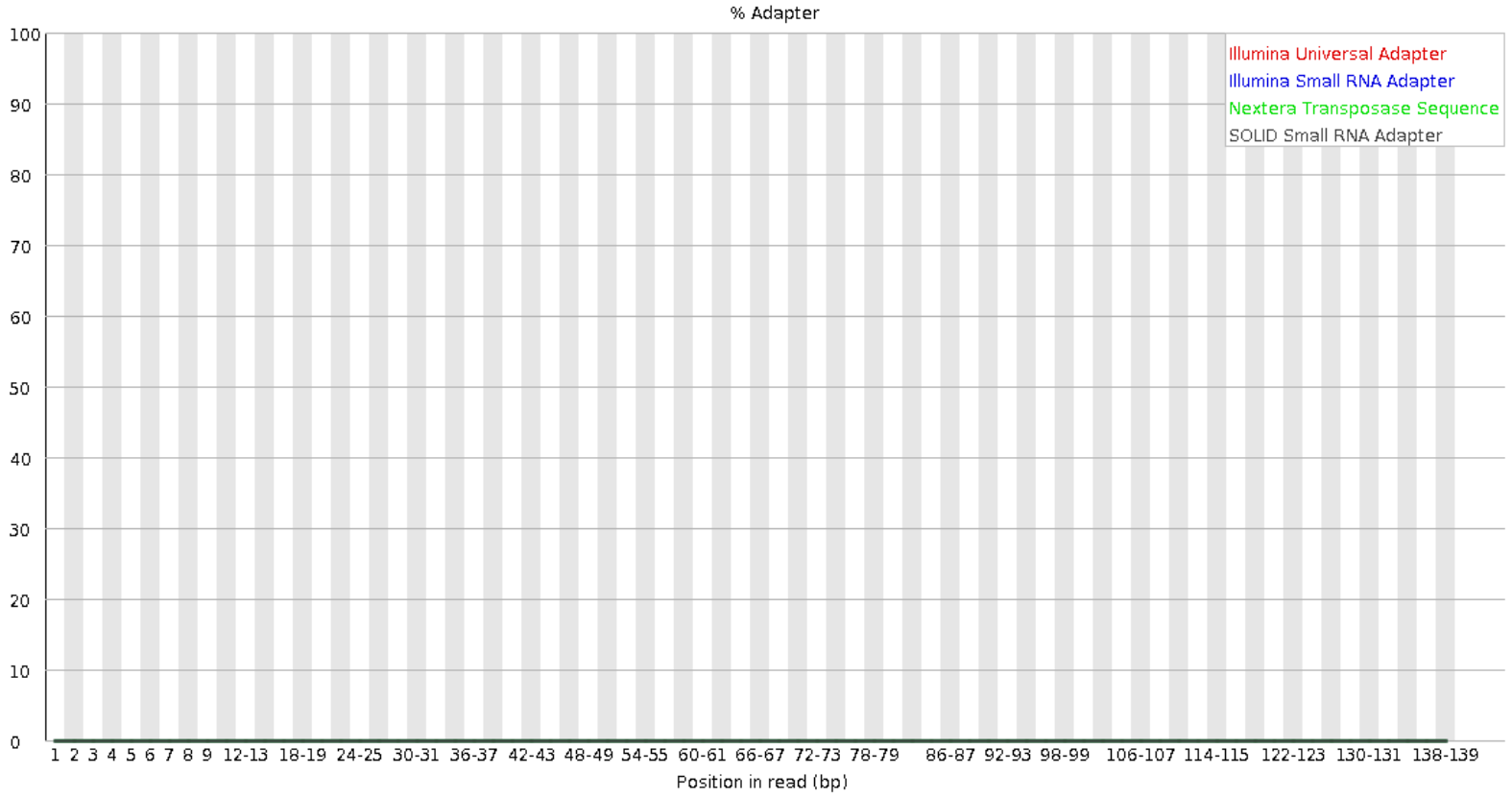


Over Represented Sequences and Adaptor Content

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGC	16515	0.14749543112472432	TruSeq Adapter, Index 6 (100% over 50bp)

Adapter Content



And the list goes on…

- Per sequence quality scores
- Per base N content
- And more! (this changes slightly depending on your FastQC version)

How Do We Improve the Quality?

- Software is available to filter reads based on quality and trim adaptor sequences.
- Many(!) different software available but all essentially do the same thing.
  - Today we're going to use TrimGalore!

Any Questions So Far?

So let's give this a go!

- Exercise 1: Genome sequencing of the bacteria *Bartonella*
- Exercise 2: Amplicon sequencing of 16S rRNA from the Earth Microbiome Project
- Exercise 3: RAD sequencing data
- Exercise 4: Amplicon sequencing of 16S rRNA from environmental samples
- Exercise 5: Filtering of amplicon sequencing data
- Exercise 6: Filtering of microRNA sequencing data

Amplicon Sequencing

Amplify Region via PCR
(16S rRNA gene)

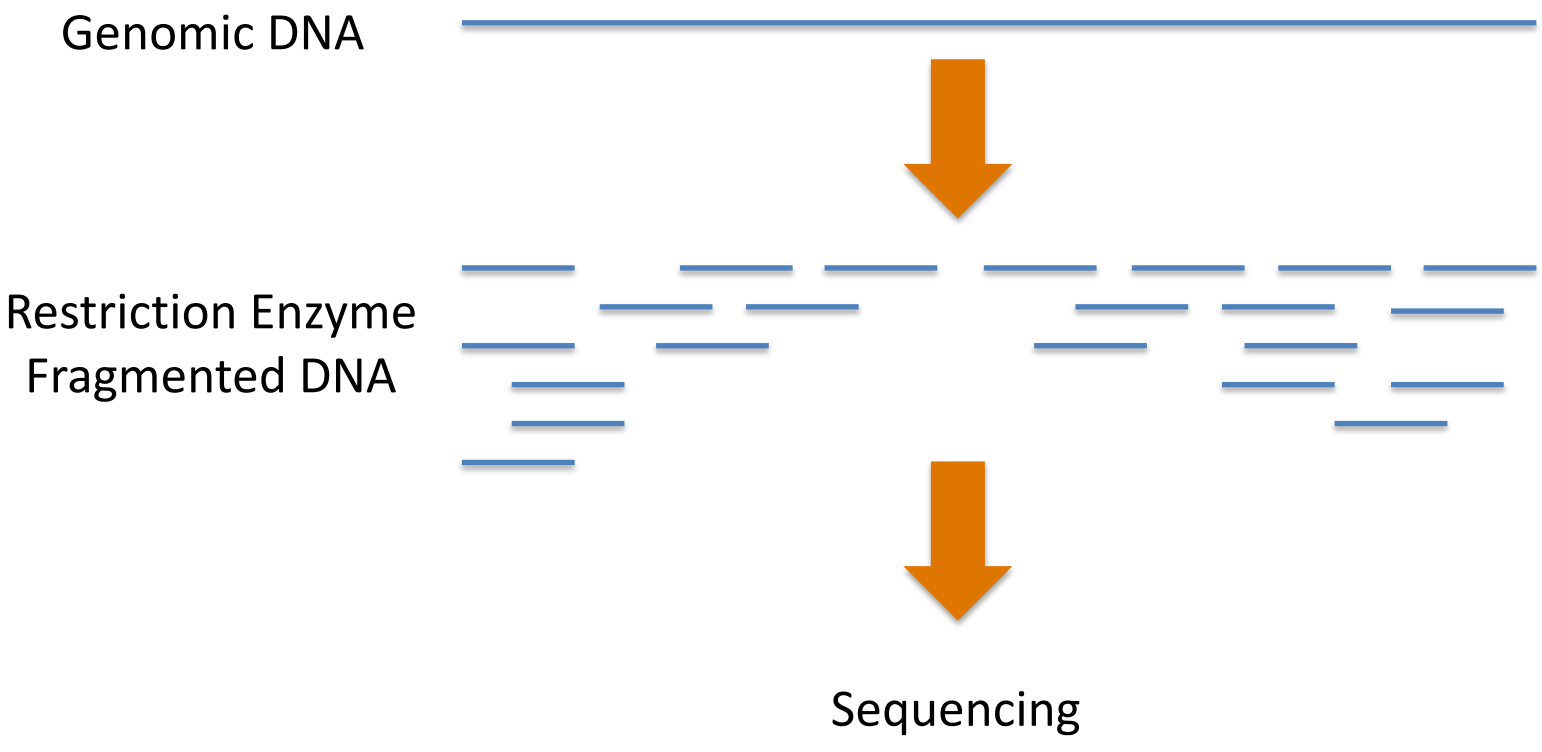
Sequence

Clustering:
Based on Sequence
Similarity

Scott will talk about this data type and analysis in more detail on Wednesday!

RAD Sequencing

RAD = Restriction-site Associated DNA



How to Run Software on the Command Line

Usage Statement:

```
$ software [options] <files>
```

For example:

```
$ fastqc [-o output] [-f fastq|bam|sam] <seqfile1..seqfileN>
```

Generally:

[ ] means optional

< > means mandatory files

| means or

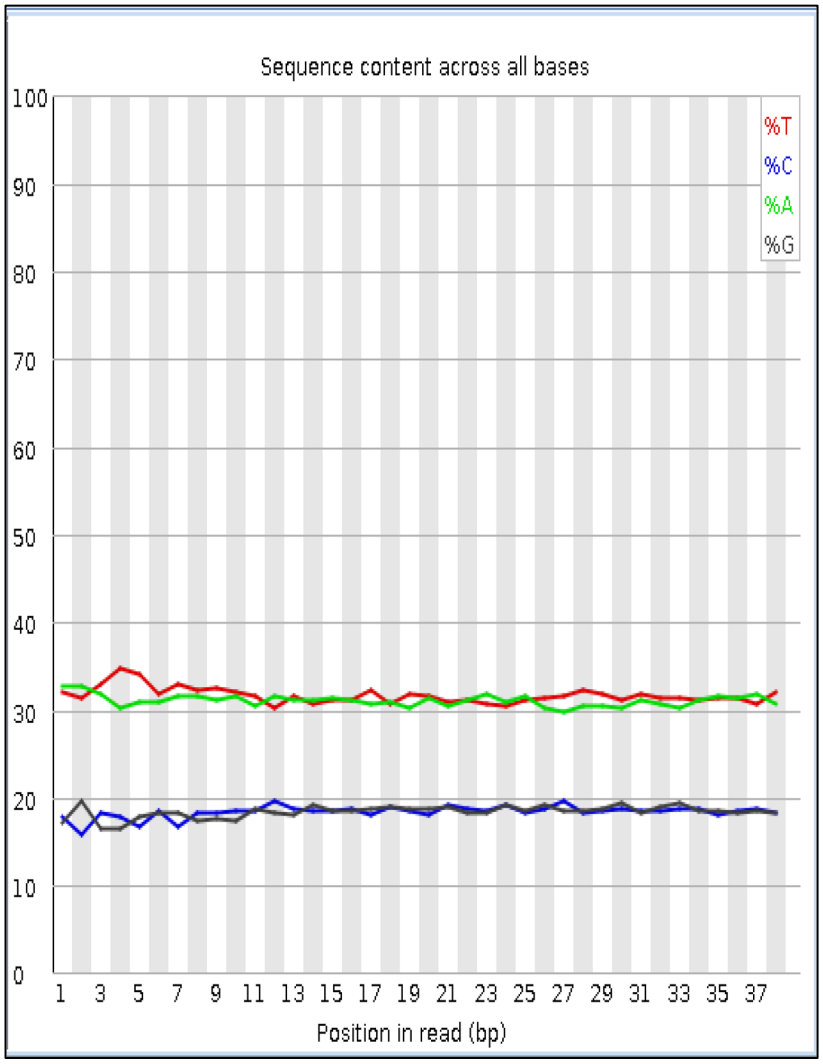
Now What?!

- Work through the Quality Control Exercises (link on the workshop web page).
- At the end of the session, we'll go over the answers!

Exercise One
Genome Sequencing Data

- There are 10,000 sequences of 38 bp in length
- The GC content is 37%
- Quality score is > Q30, which = 1 error in 1000 so base call quality is 99.9%
- Would you think this is good quality sequencing data?

Exercise One
Genome Sequencing Data



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCAC...	17	0.17	TruSeq Adapter, In...

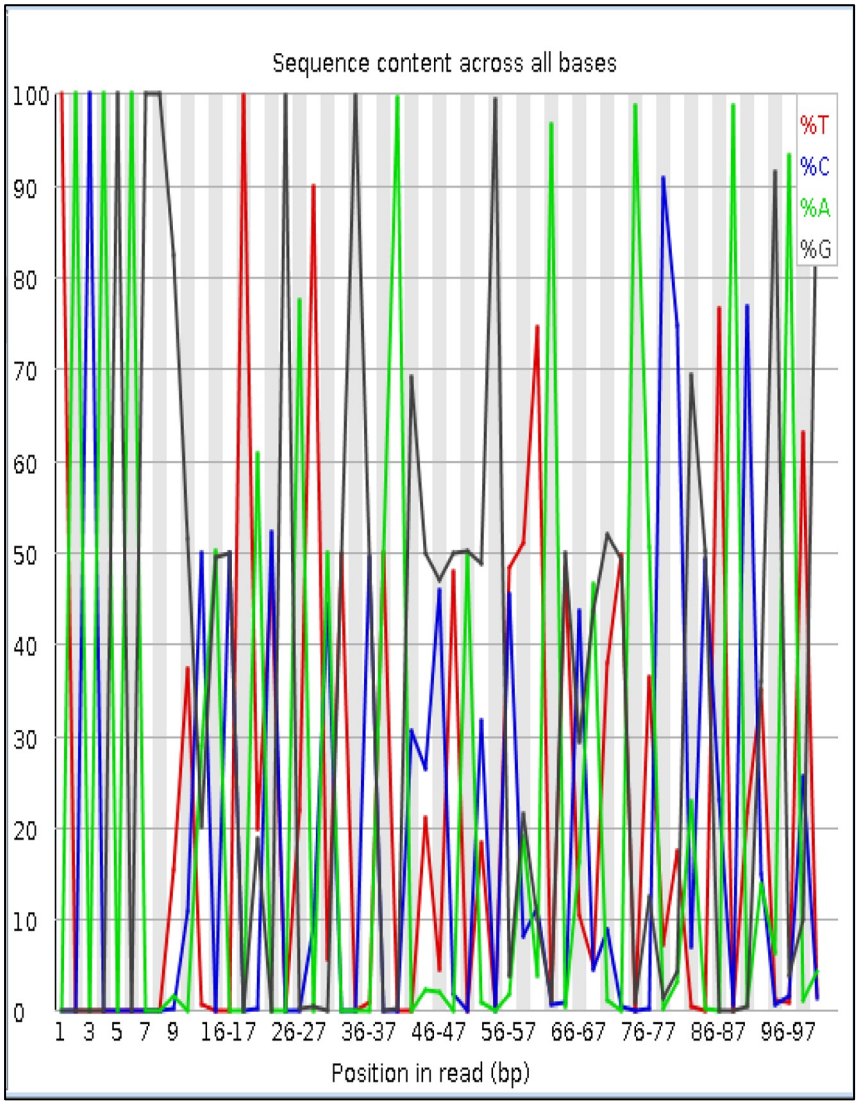
Are we worried about this data?!
No - This shows the GC content we expect and very few adaptors.

Exercise Two
Amplicon Sequencing Data

Conserved sequence at the beginning of the reads:

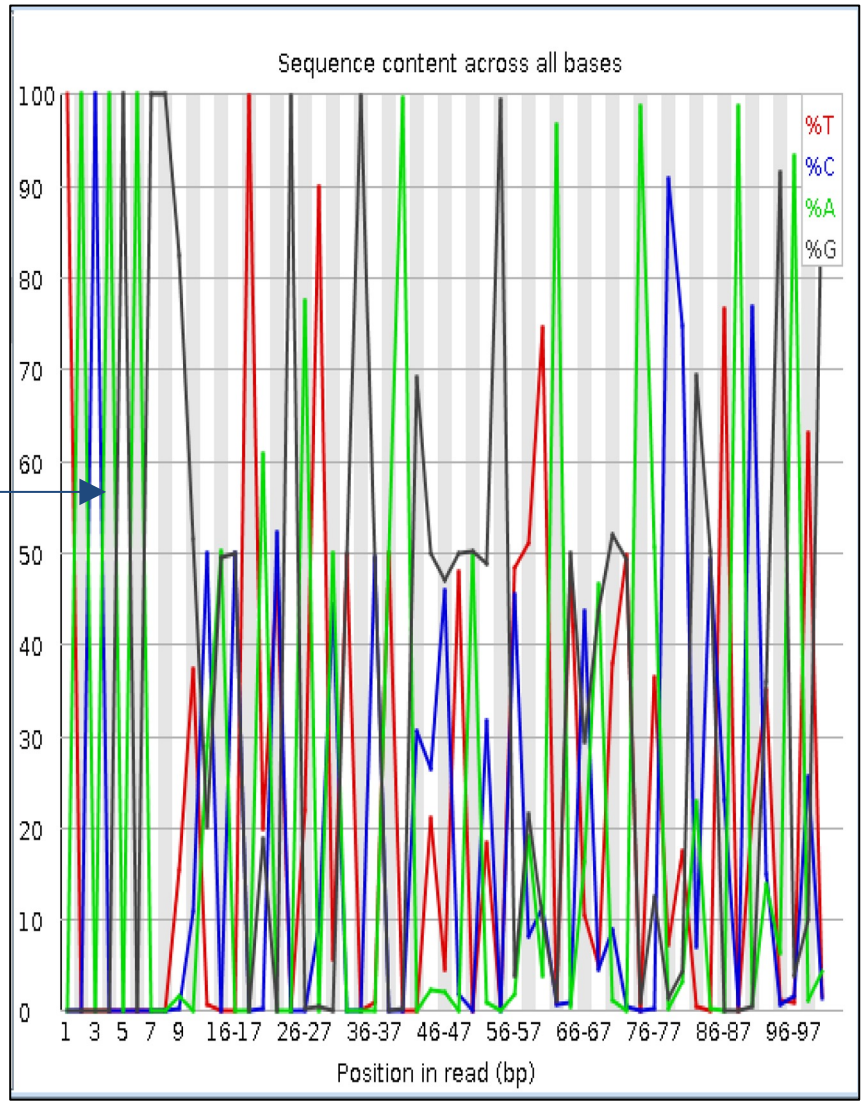
TACAGAGG

Lots of sequences with very similar sequence.



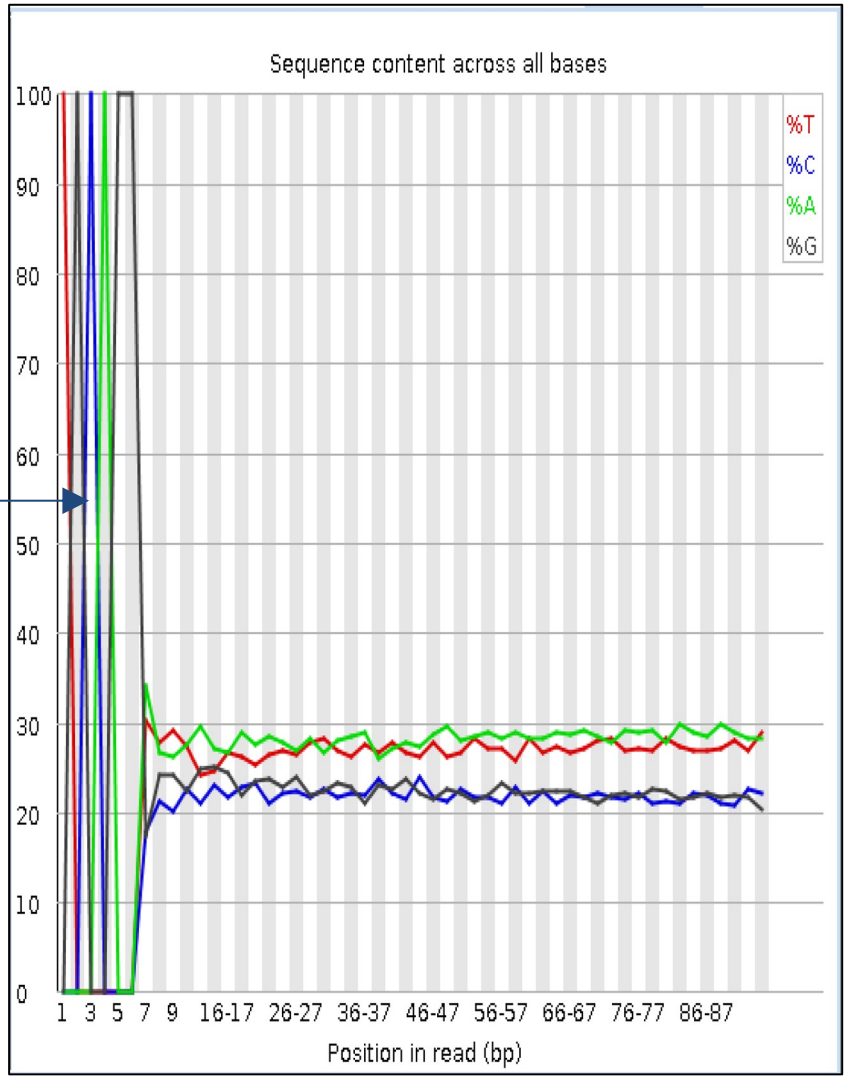
Exercise Two
Amplicon Sequencing Data

Filter Low Quality
&
Remove the Primer
Sequence

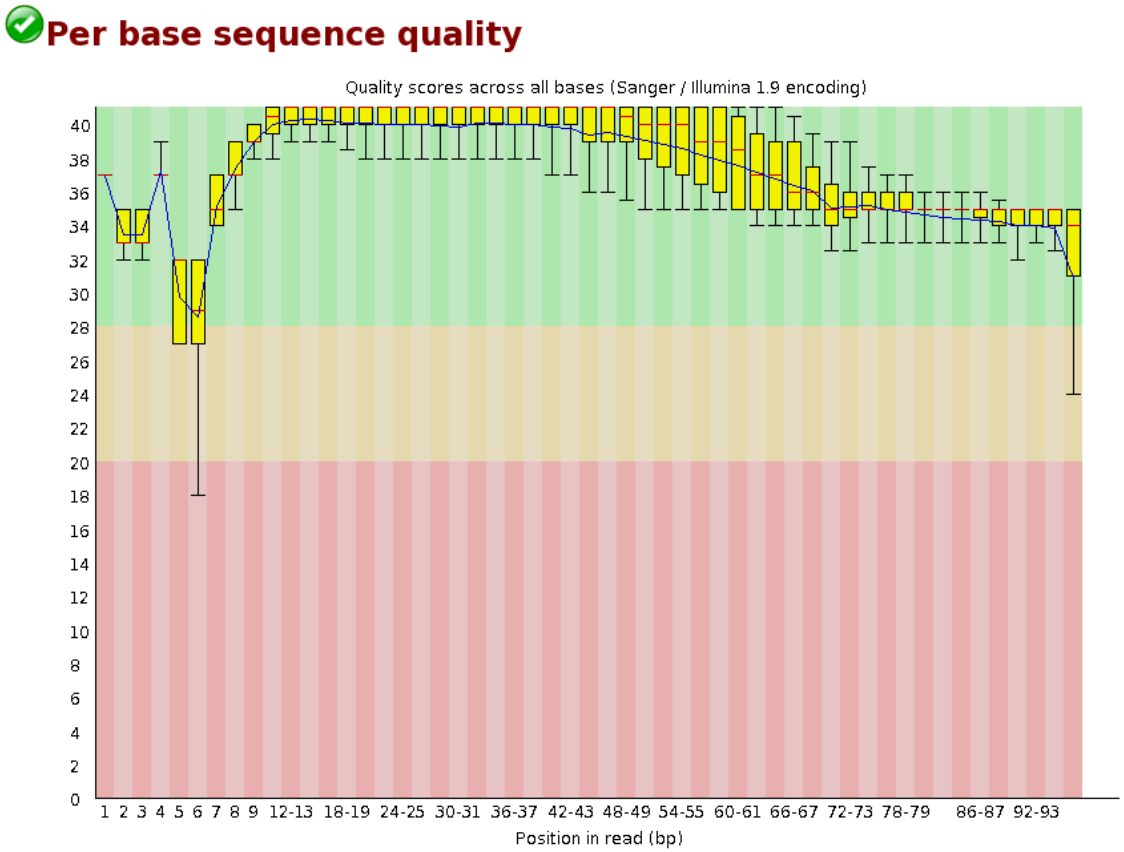


Exercise Three
RAD Sequencing Data

Restriction Enzyme
Digestion Site



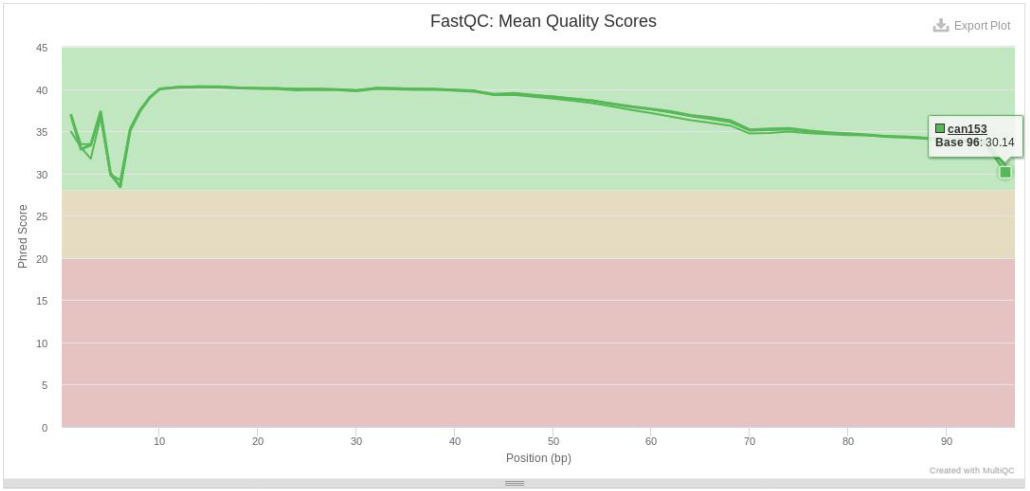
Exercise Three
RAD Sequencing Data

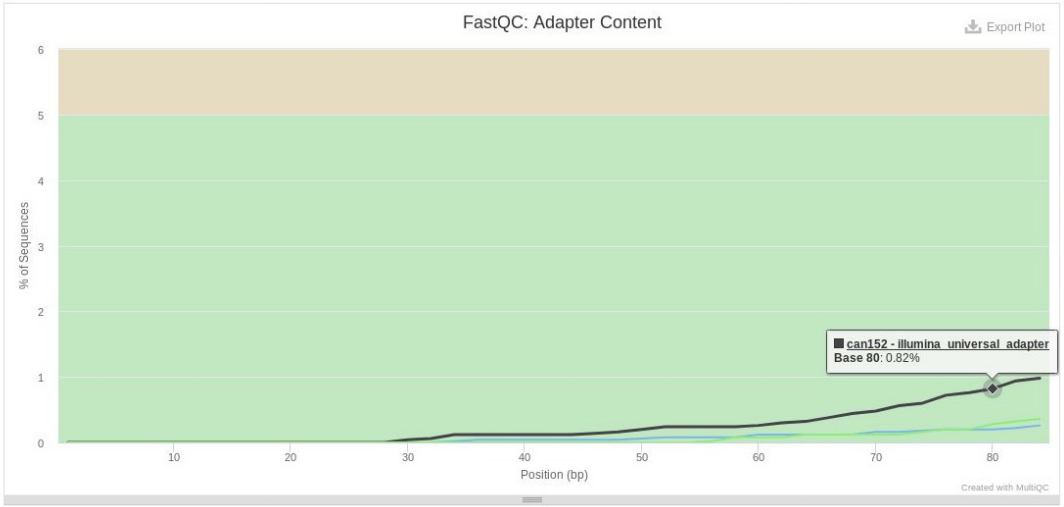


Lack of diversity between clusters confuses the sequencer.
All the dots are identical and therefore it can not decide where one cluster starts and one ends!

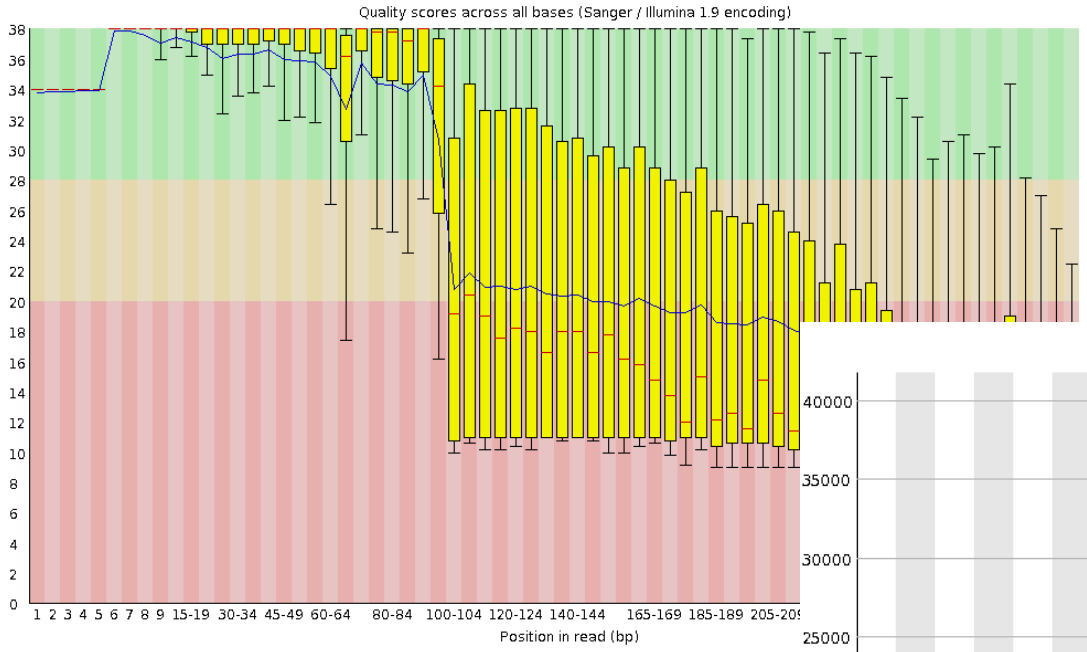
Exercise Three
RAD Sequencing Data

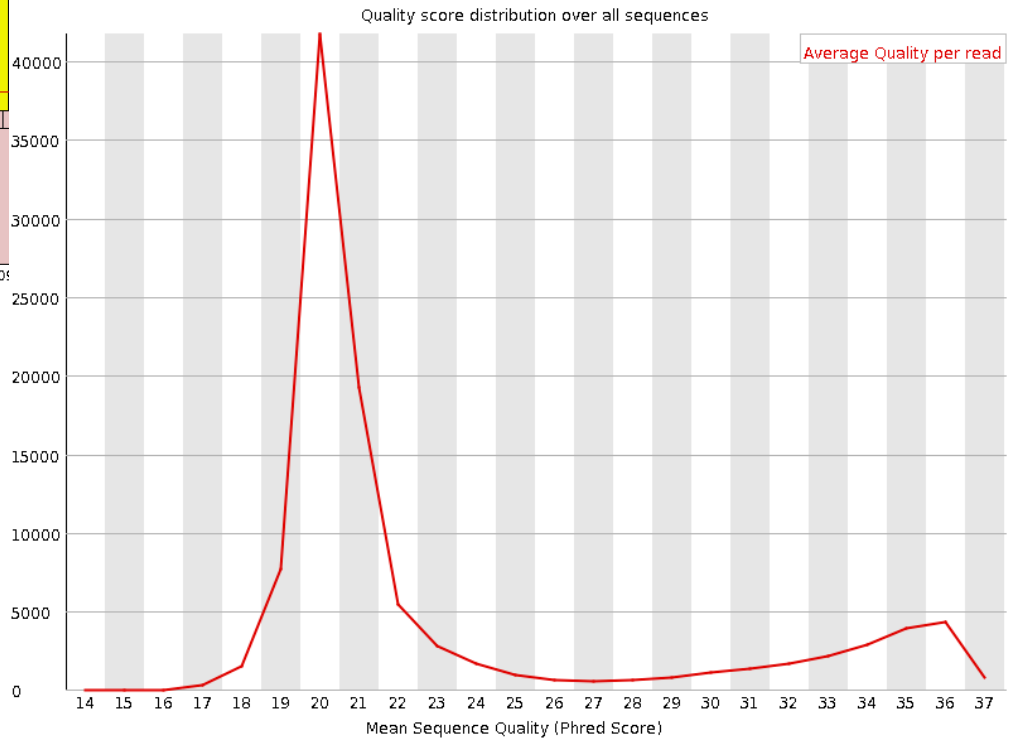
MultiQC makes it easier to compare multiple FastQC files at once.





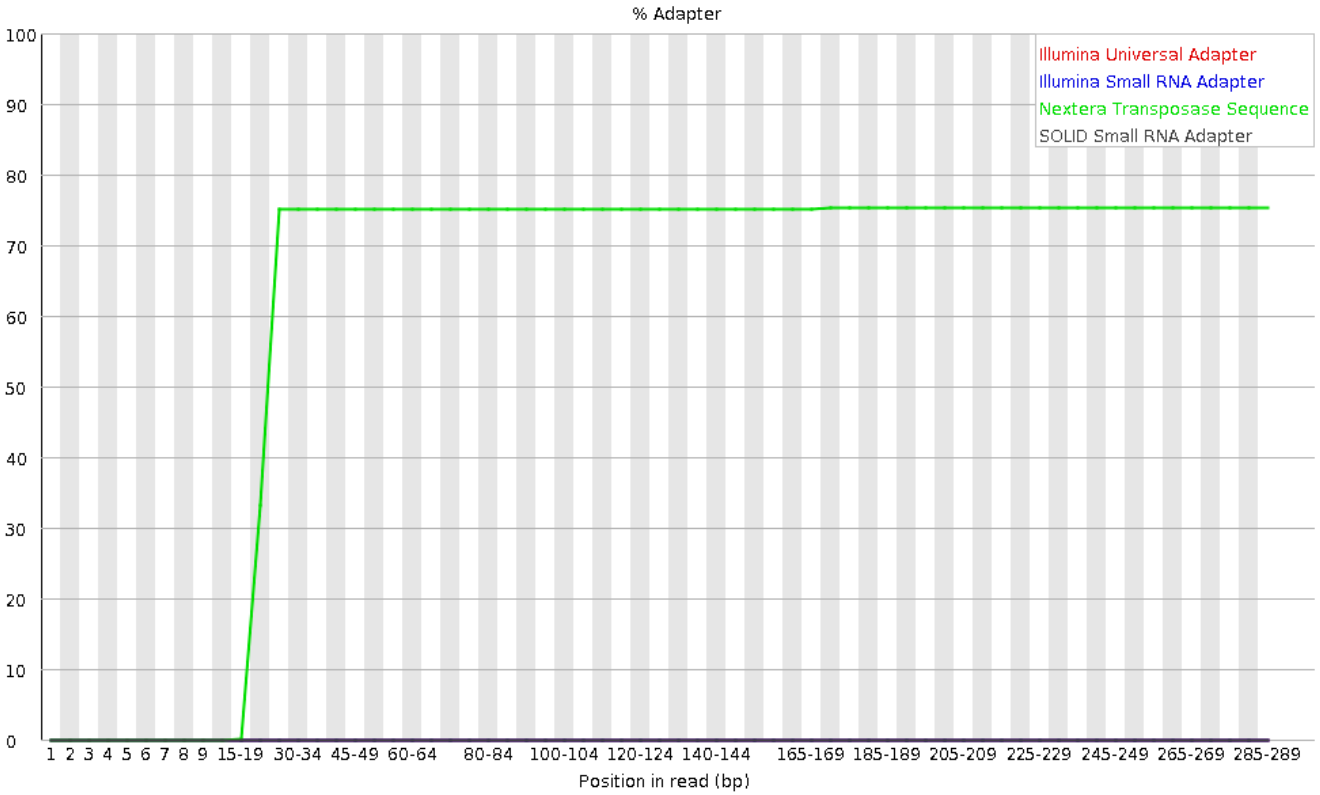
Exercise Four
Amplicon Sequencing Data





Very low average quality

Exercise Four
Amplicon Sequencing Data



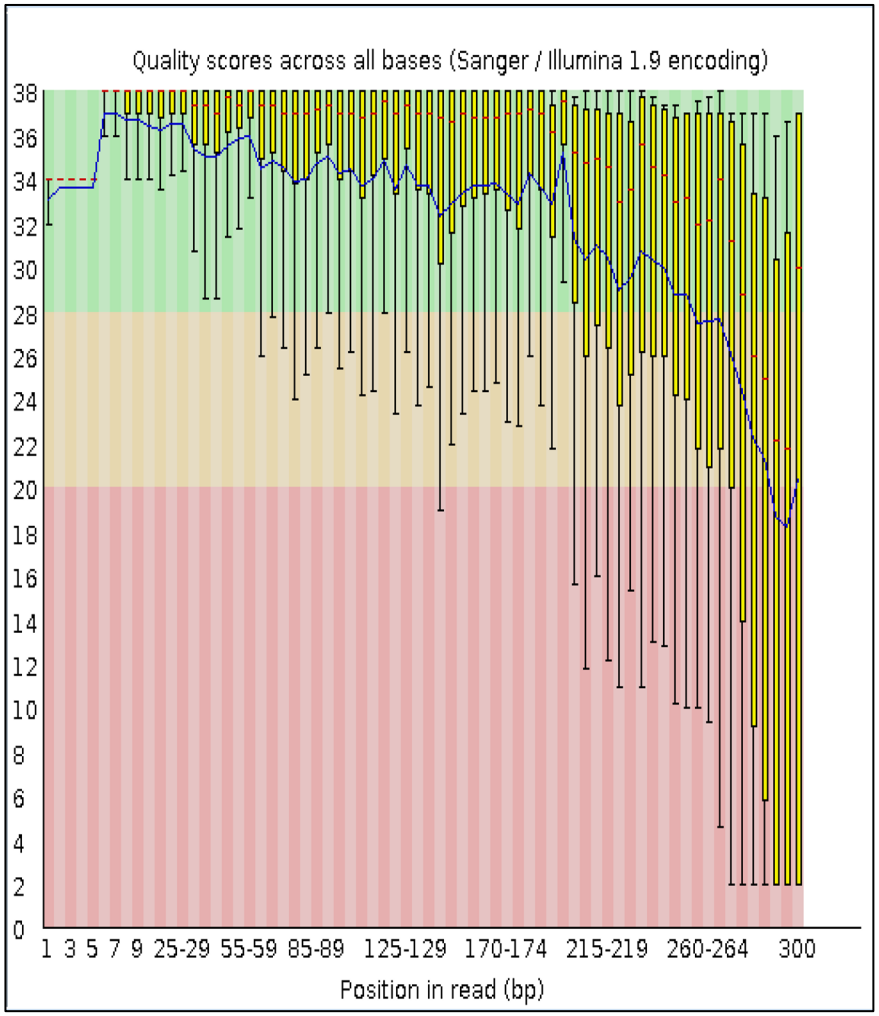
75% of bases are adapters after 15 bp

Majority adaptor dimers

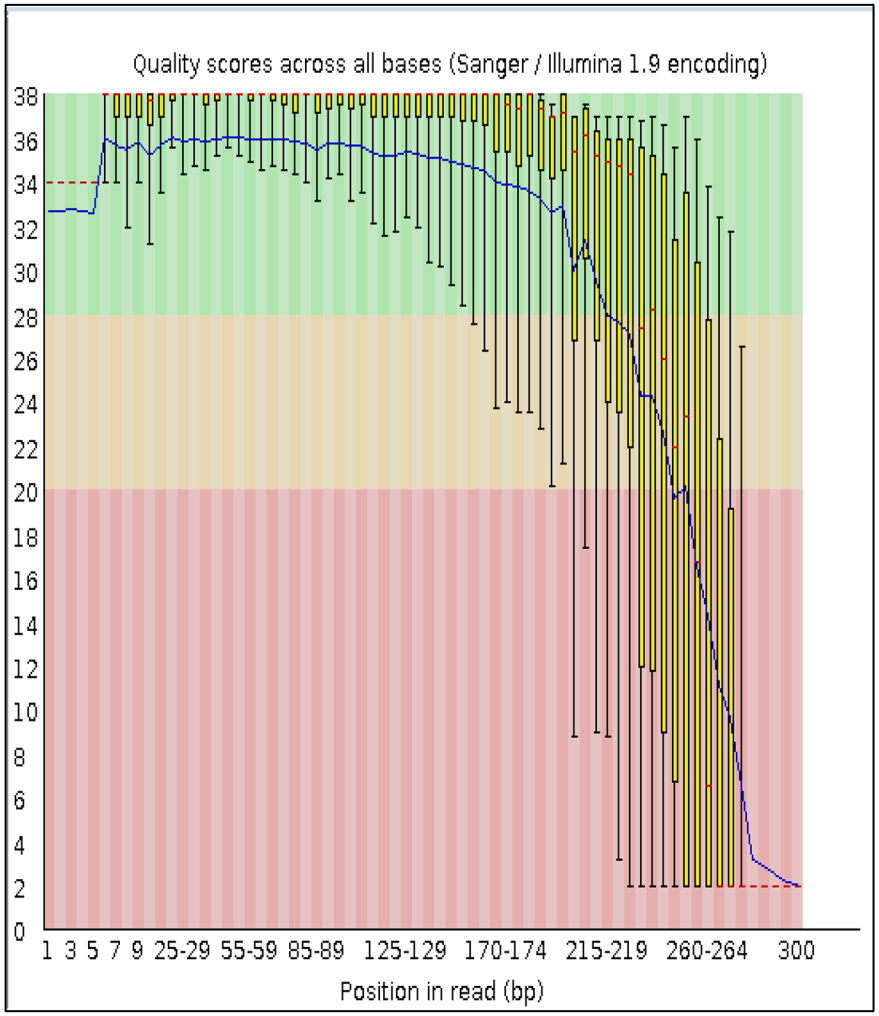
Unlikely to be able to save this data!

Exercise Five
Amplicon Sequencing Data

Read 1



Read 2



Exercise Five
Amplicon Sequencing Data

Your Trim Galore! Command
(all on one line):

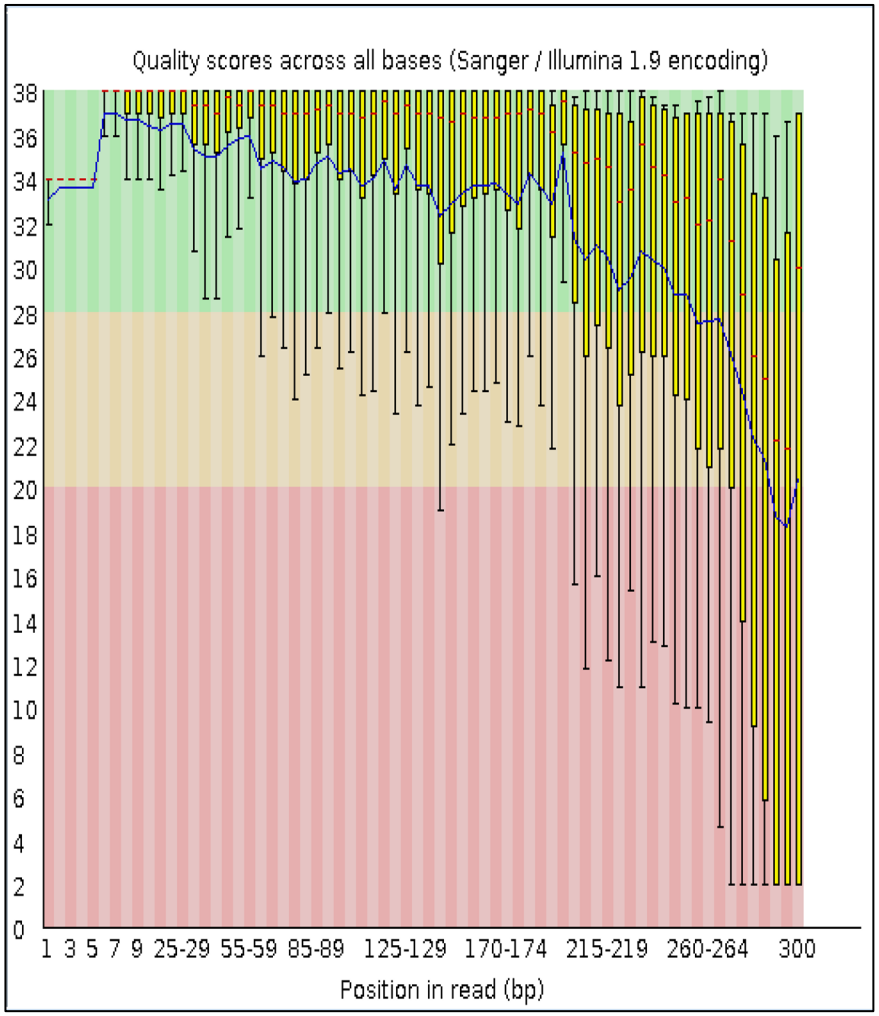
```
$ trim_galore -q 35 --paired
1_TAAGGCGA-TAGATCGC_L001_R1_001.fastq.gz
1_TAAGGCGA-TAGATCGC_L001_R2_001.fastq.gz
```

Read 1 – 58204 (9.4 %) of reads with adaptors
Read 2 – 233549 (37.9 %) of reads with adaptors

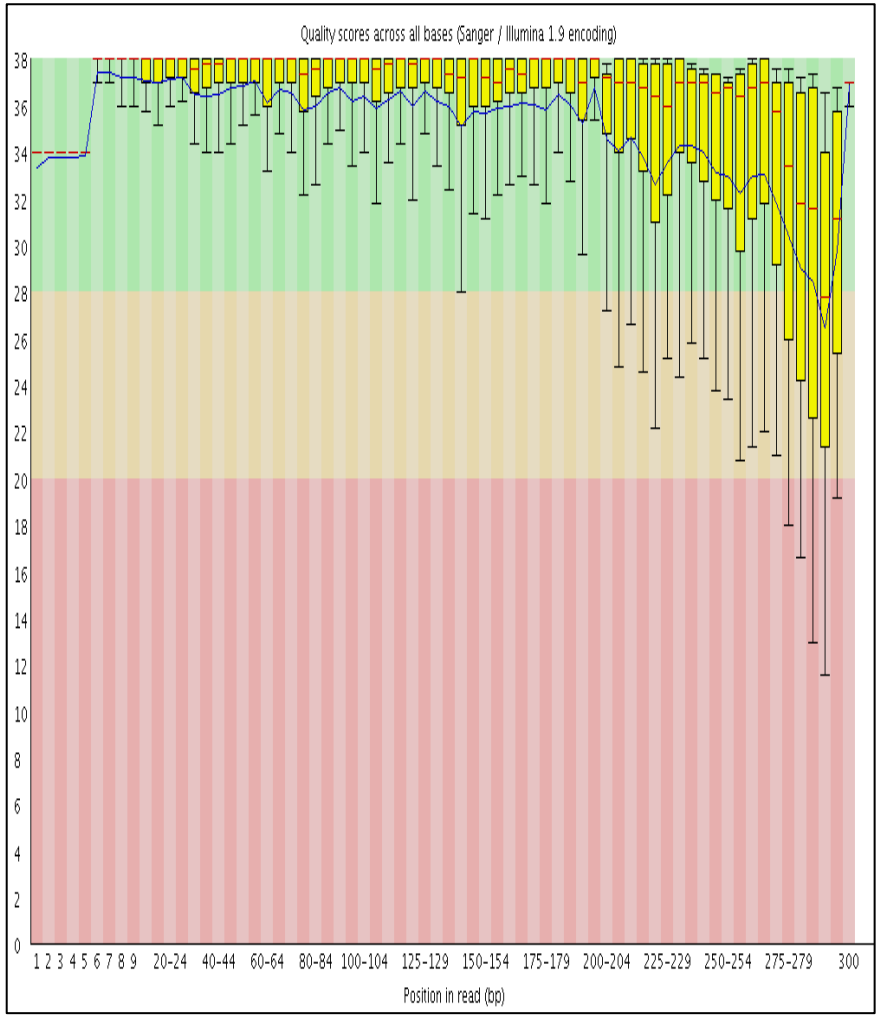
138277 read pairs removed (22.45 %)

Exercise Five
Amplicon Sequencing Data

Read 1 Before

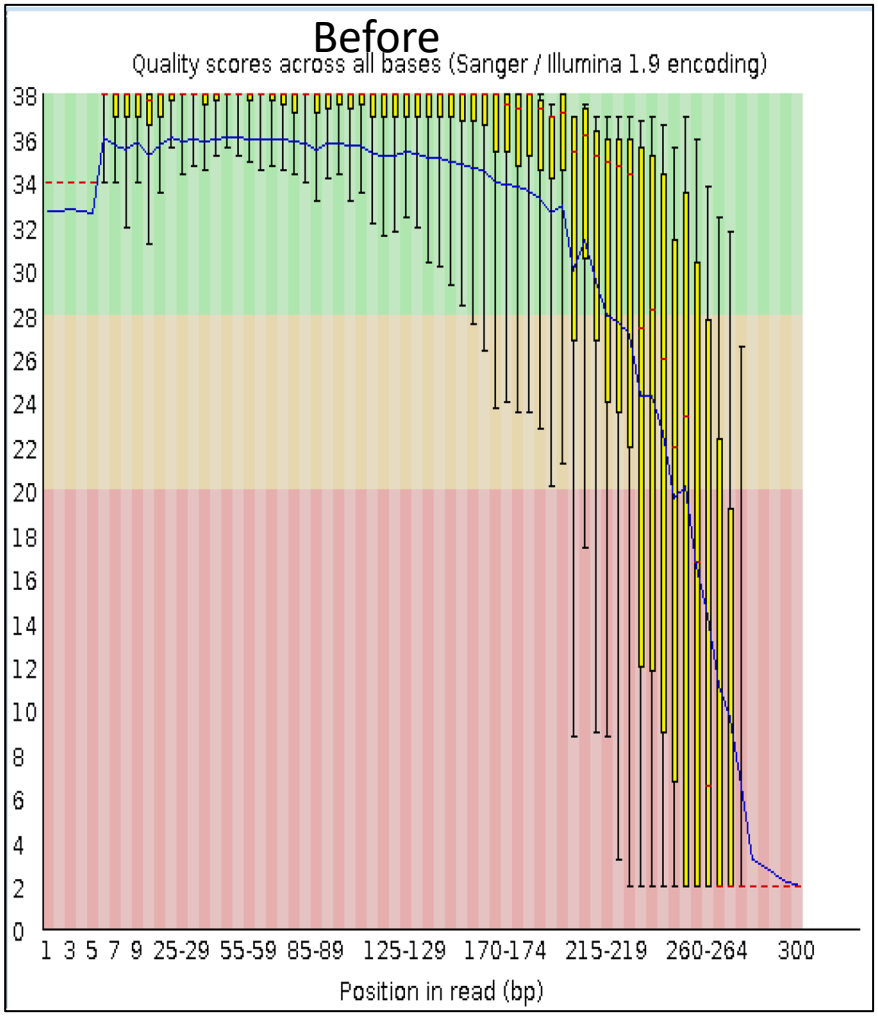


Read 1

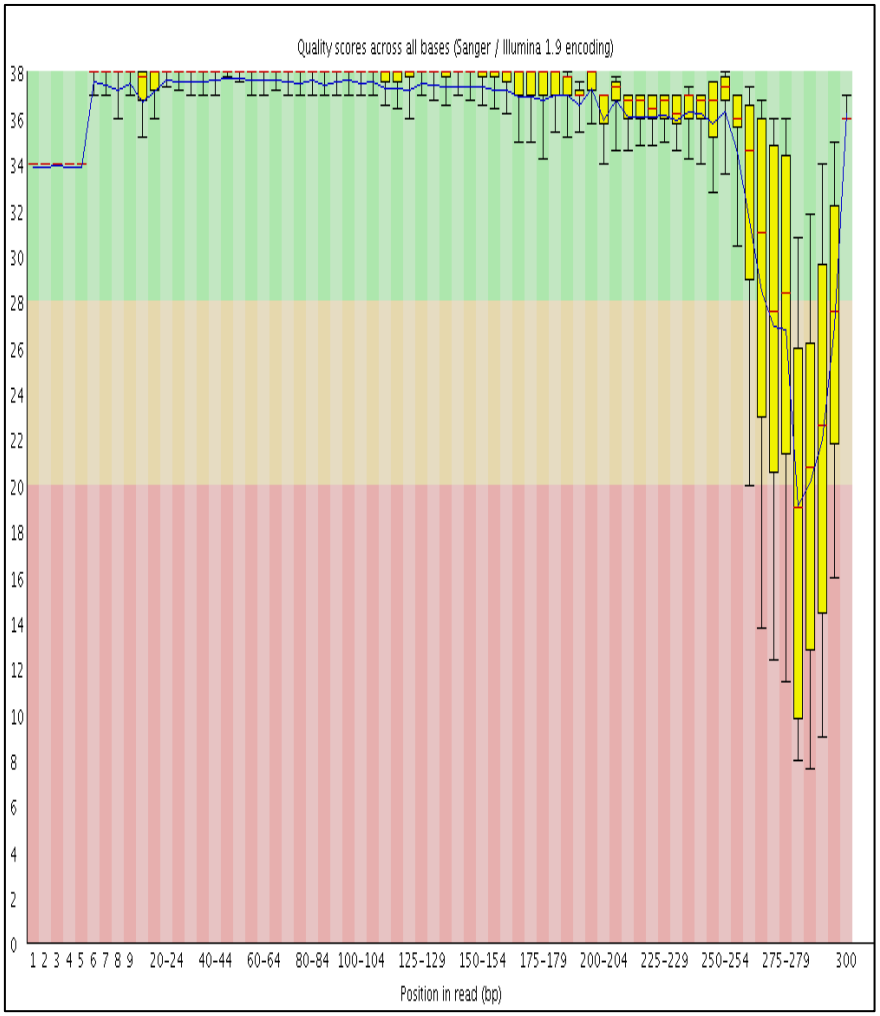


Exercise Five
Amplicon Sequencing Data

Read 2
Before



Read 2



Exercise Five
Amplicon Sequencing Data

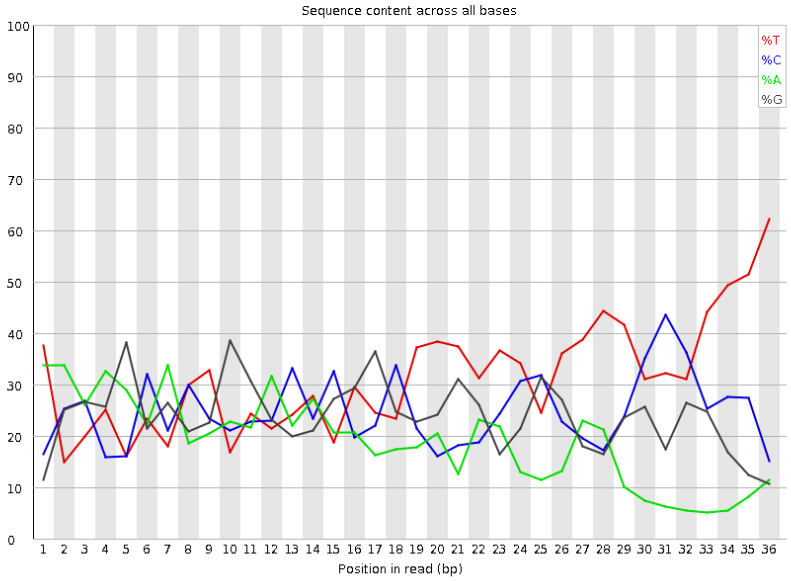
When would you be less strict?

- Quality vs. Quantity - Do you have enough data after filtering?
- SOME alignments - RNA sequencing vs. SNP calling
- Does your downstream software handle quality scores and filtering? E.G. DADA2, QIIME2, A5 assembler and Stacks

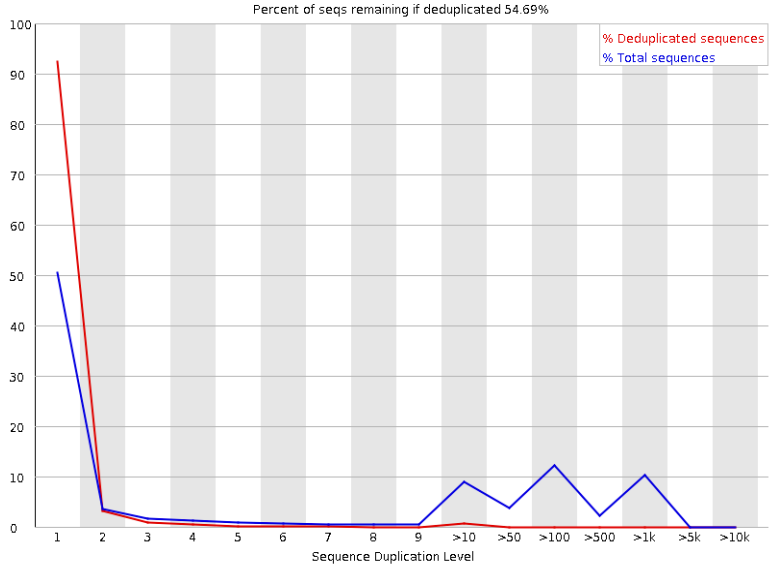
Exercise Six
microRNA Sequencing Data

Think about the data you have! microRNAs are likely to be similar.

Per base sequence content



Sequence Duplication Levels



Exercise Six
microRNA Sequencing Data

What is the source of the overrepresented sequences?

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGCAGCATTGTACA...	3398	3.398	No Hit
TACAGTCCGACGAT...	1814	1.814	Illumina PCR Prime...
TCTACAGTCCGACG...	1570	1.57	RNA PCR Primer, In...
TATTGCACTTGTCCC...	1421	1.421	No Hit
TTCTACAGTCCGAC...	1181	1.181	RNA PCR Primer, In...
CTACAGTCCGACGA...	1168	1.168	Illumina PCR Prime...
CATTGCACTTGTCTC...	839	0.839	No Hit
ACAGTCCGACGATC...	835	0.835	RNA PCR Primer, In...
AGTTCTACAGTCCG...	648	0.648	Illumina PCR Prime...
AAAGTGCTGCGACA...	491	0.491	No Hit
TCGTATGCCGTCTT...	465	0.465	Illumina Single En...
CAGTCCGACGATCT...	436	0.436	Illumina PCR Prime...
TNNNNNNNNNNNN...	392	0.392	No Hit
TAGCTTATCAGACT...	388	0.388	No Hit
TATTGCACTCGTCC...	366	0.366	TruSeq Adapter, I...
ACCGGGCGGAAAC...	357	0.357	No Hit
ANNNNNNNNNNNN...	355	0.355	No Hit
GTTCTACAGTCCGA...	353	0.353	Illumina PCR Prime...
AAGTGCTGCGACAT...	341	0.341	No Hit

Exercise Six
microRNA Sequencing Data

Your Trim Galore! Command
(all on one line):

```
$ trim_galore -a ATCTCGTATGCCGTCTTCTGCTTG
SRR026762-sample.fastq.gz
```

Read 1 – 37661 (37.7 %) of reads with adaptors

33738 reads removed (33.7 %)

Exercise Five
Amplicon Sequencing Data

Are we happy with the final data?

What about the primer sequences that remain?

- May need further rounds of removal or manual trimming.
- This is an alignment project and therefore complete removal less important.

What about the quality trimming?

- Trimmers work with a sliding window and calculates the average Q score in that window, if it's higher on average than the cut off, it stays.

Any Questions So Far?

Take Home Message!

- It is essential to QC your data before beginning analysis.
- What are you expecting? Think about your experimental design, your species etc…
- What are you doing with the data? Alignment, assembly etc…
- No two datasets are the same!

What We're Going To Do

THIS AFTERNOON

- Introduction to Sequencing ✓
- Checking the Quality of Sequencing Data ✓