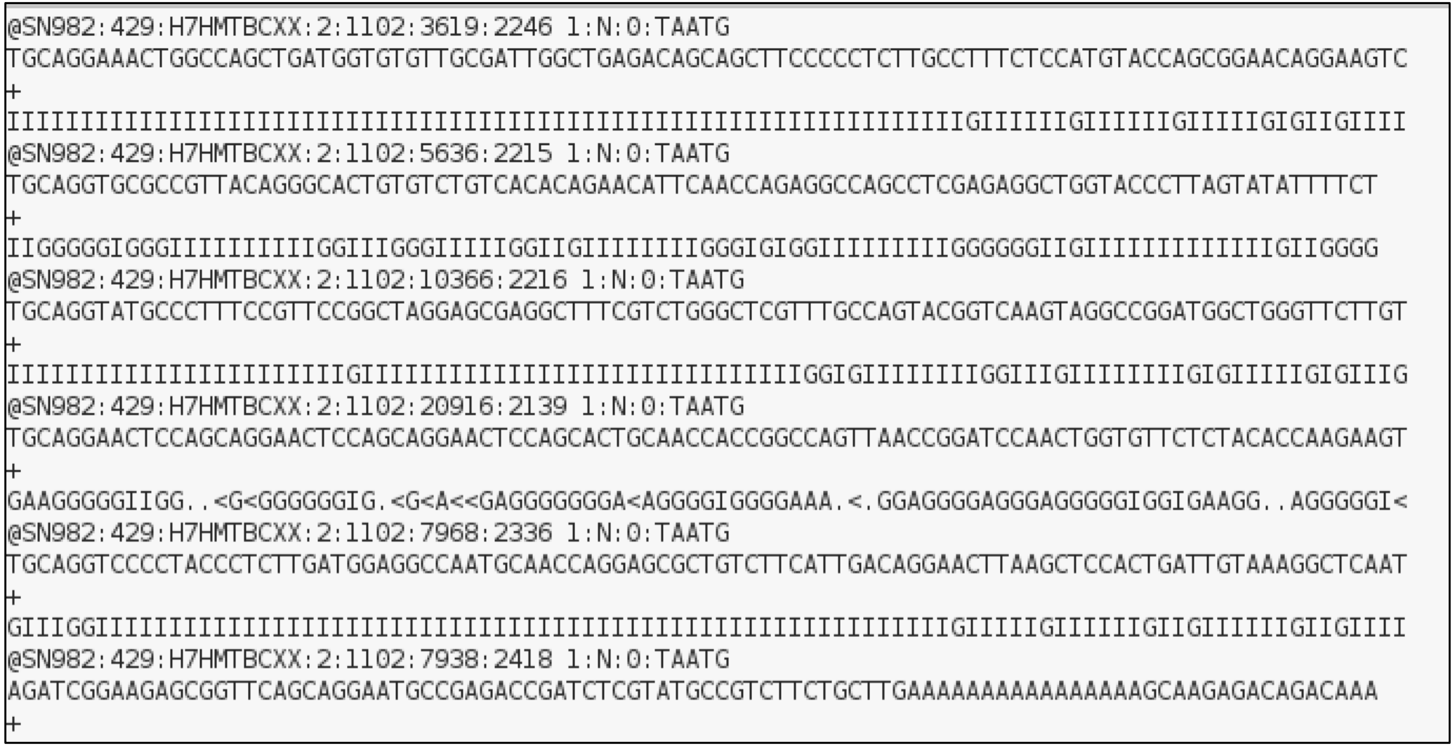
Quality Control Laboratory

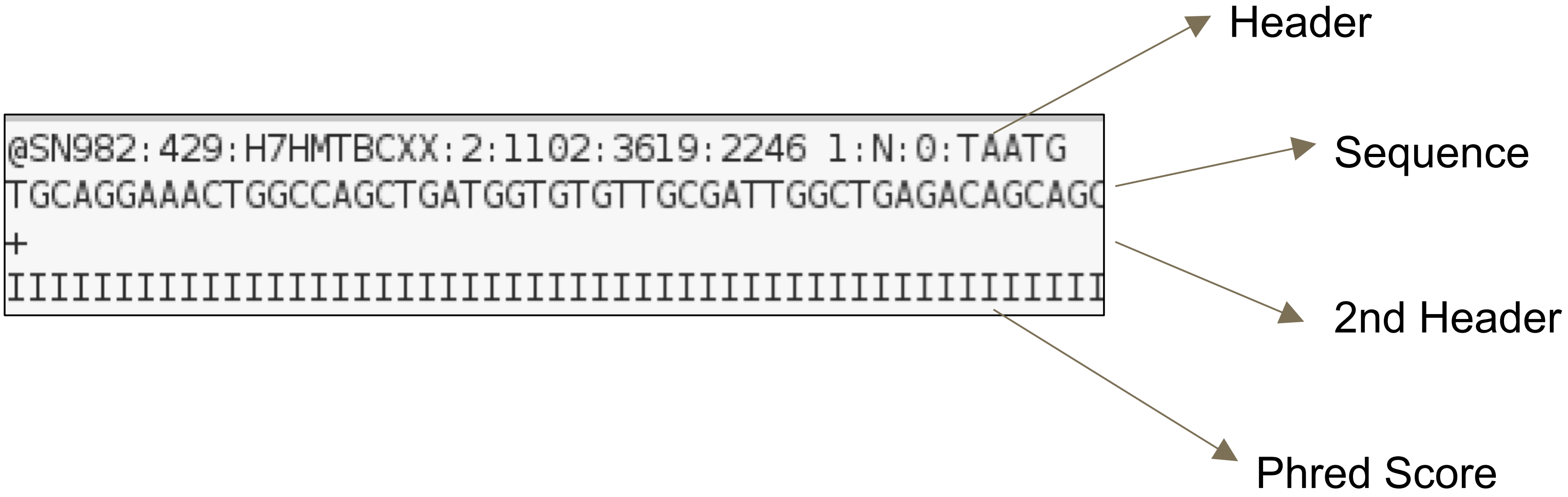
Josie Paris & Sophie Shaw
Workshop on Genomics 2016

Fastq Format



```
@SN982:429:H7HMTBCXX:2:1102:3619:2246 1:N:0:TAATG
TGCAGGAAACTGGCCAGCTGATGGTGTGTTGCGATTGGCTGAGACAGCAGCTTCCCCCTCTTGCCTTTCTCCATGTACCAGCGGAACAGGAAGTC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIGIIIIIIGIIIIIGIGIIGIIII
@SN982:429:H7HMTBCXX:2:1102:5636:2215 1:N:0:TAATG
TGCAGGTGCGCCGTTACAGGGCACTGTGTCTGTCACACAGAACATTCAACCAGAGGCCAGCCTCGAGAGGCTGGTACCCTTAGTATATTTTCT
+
IIGGGGGIGGGIIIIIIIIIIGGIIIGGGIIIIIGGIIGIIIIIIIIGGGIGIGGIIIIIIIIIIGGGGGGIIGIIIIIIIIIIIIIGIIGGGG
@SN982:429:H7HMTBCXX:2:1102:10366:2216 1:N:0:TAATG
TGCAGGTATGCCCTTTCCGTTCCGGCTAGGAGCGAGGCTTTCGTCTGGGCTCGTTTGCCAGTACGGTCAAGTAGGCCGGATGGCTGGGTTCTTGT
+
IIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGIGIIIIIIIIIGGIIIGIIIIIIIIIGIGIIIIIGIGIIIG
@SN982:429:H7HMTBCXX:2:1102:20916:2139 1:N:0:TAATG
TGCAGGAACTCCAGCAGGAACTCCAGCAGGAACTCCAGCACTGCAACCACCGGCCAGTTAACCGGATCCAACTGGTGTTCTCTACACCAAGAAGT
+
GAAGGGGGIIGG..<G<GGGGGGIG.<G<A<<GAGGGGGGGA<AGGGGIGGGGAAA.<.GGAGGGGAGGGAGGGGGIGGIGAAGG..AGGGGGI<
@SN982:429:H7HMTBCXX:2:1102:7968:2336 1:N:0:TAATG
TGCAGGTCCCCTACCCTCTTGATGGAGGCCAATGCAACCAGGAGCGCTGTCTTCATTGACAGGAACTTAAGCTCCACTGATTGTAAAGGCTCAAT
+
GIIIGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIGIIIIIIGIIGIIIIIIGIIGIIII
@SN982:429:H7HMTBCXX:2:1102:7938:2418 1:N:0:TAATG
AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAAAAGCAAGAGACAGACAAA
+
```

Fastq Format



```
@SN982:429:H7HMTBCXX:2:1102:3619:2246 1:N:0:TAATG
TGCAGGAAACTGGCCAGCTGATGGTGTGTTGCGATTGGCTGAGACAGCAGC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

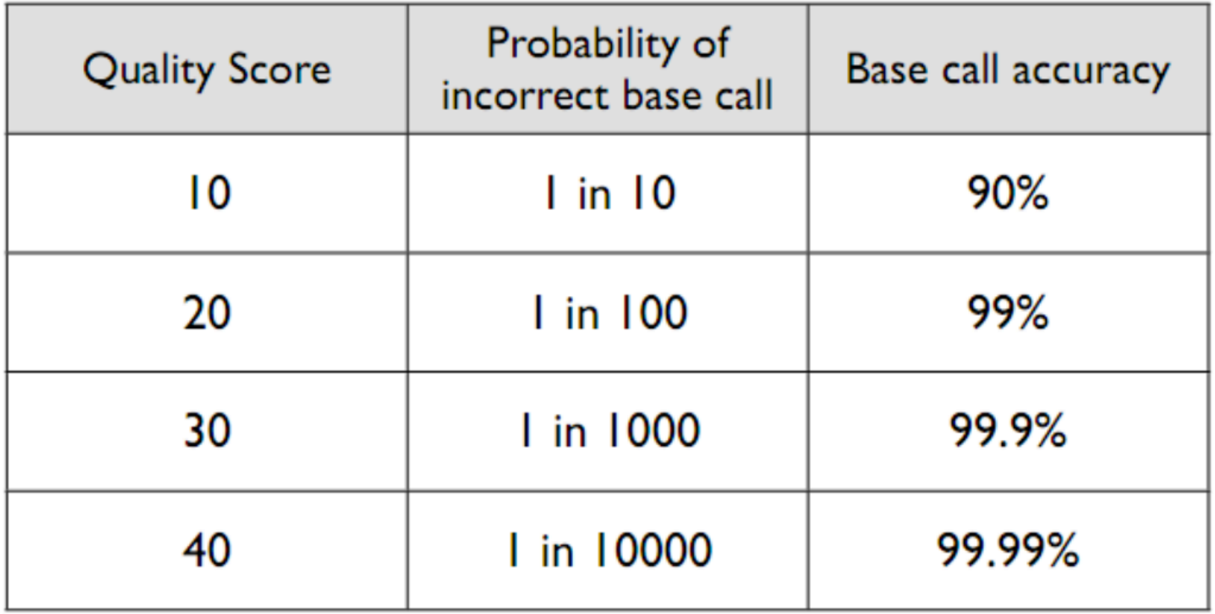
Header

Sequence

2nd Header

Phred Score

Phred Scores (Q)



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%

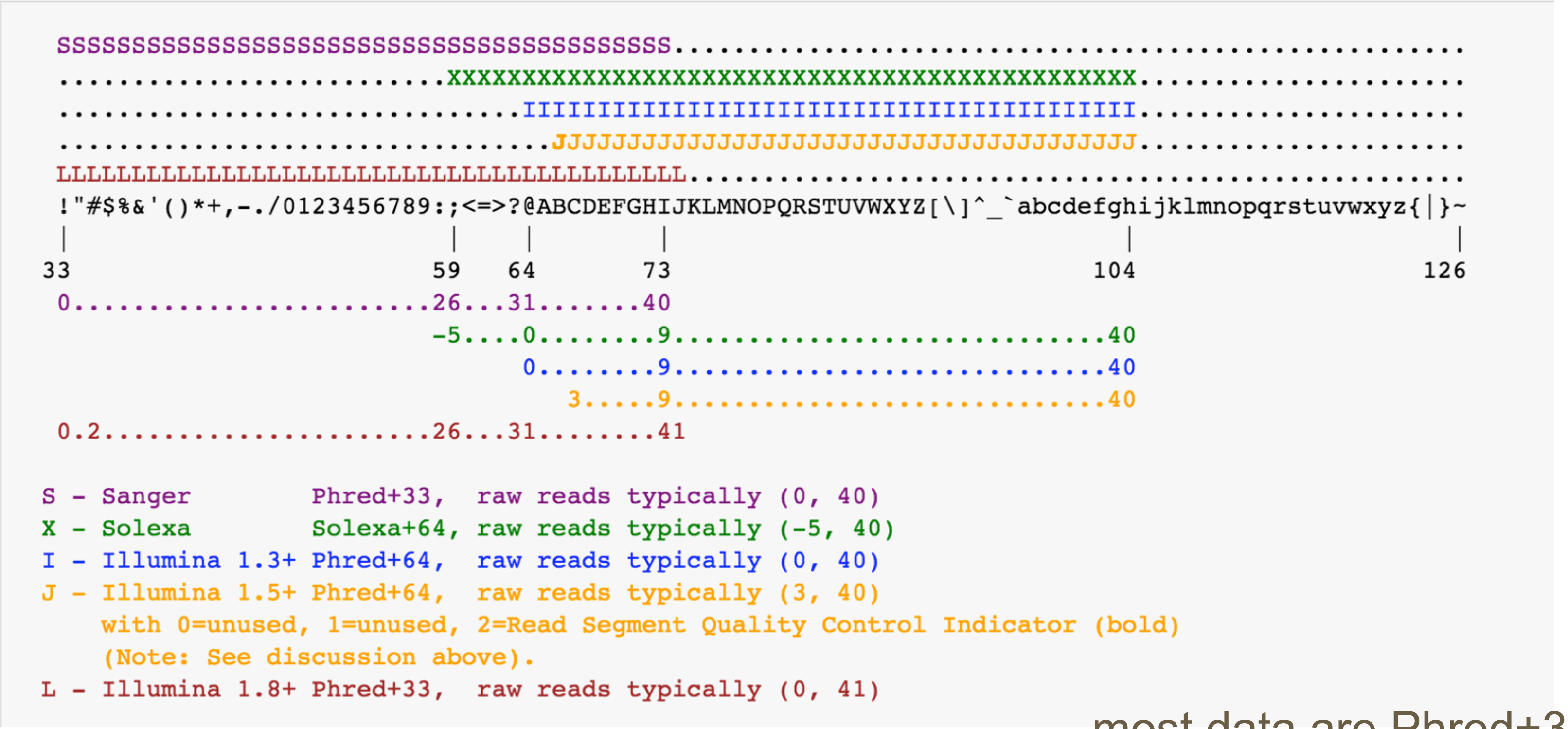
https://en.wikipedia.org/wiki/FASTQ_format

ASCII encoding of phred scores

one number : one letter

```
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



```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................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
  0.2......................26...31........41

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

...most data are Phred+33

Quality Control

Why? Low quality reads, contamination and adaptors introduce errors into data.

Filtering and trimming these sequences may help to improve downstream analysis.

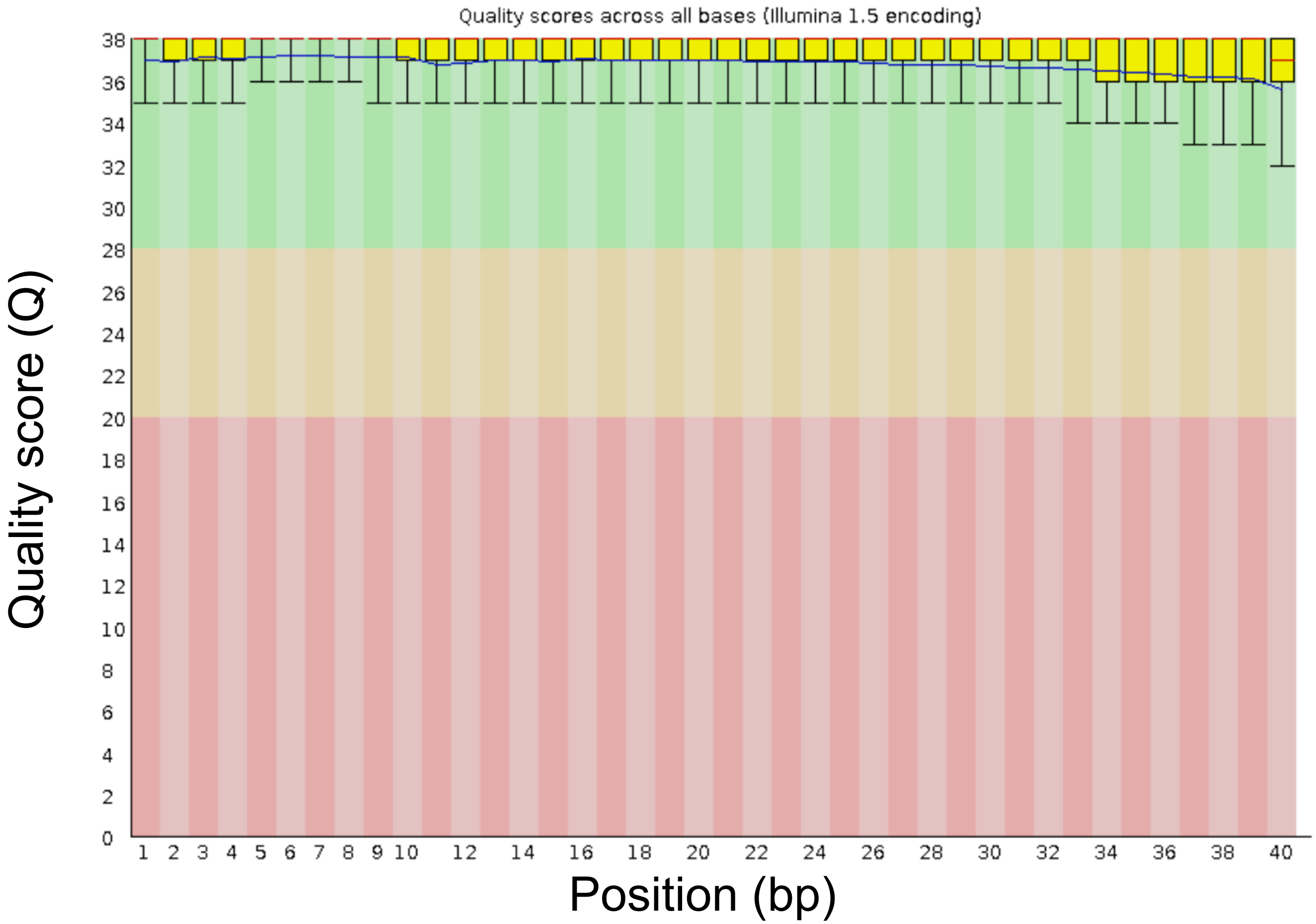
Filtering isn’t always needed. Some programs take quality into account.

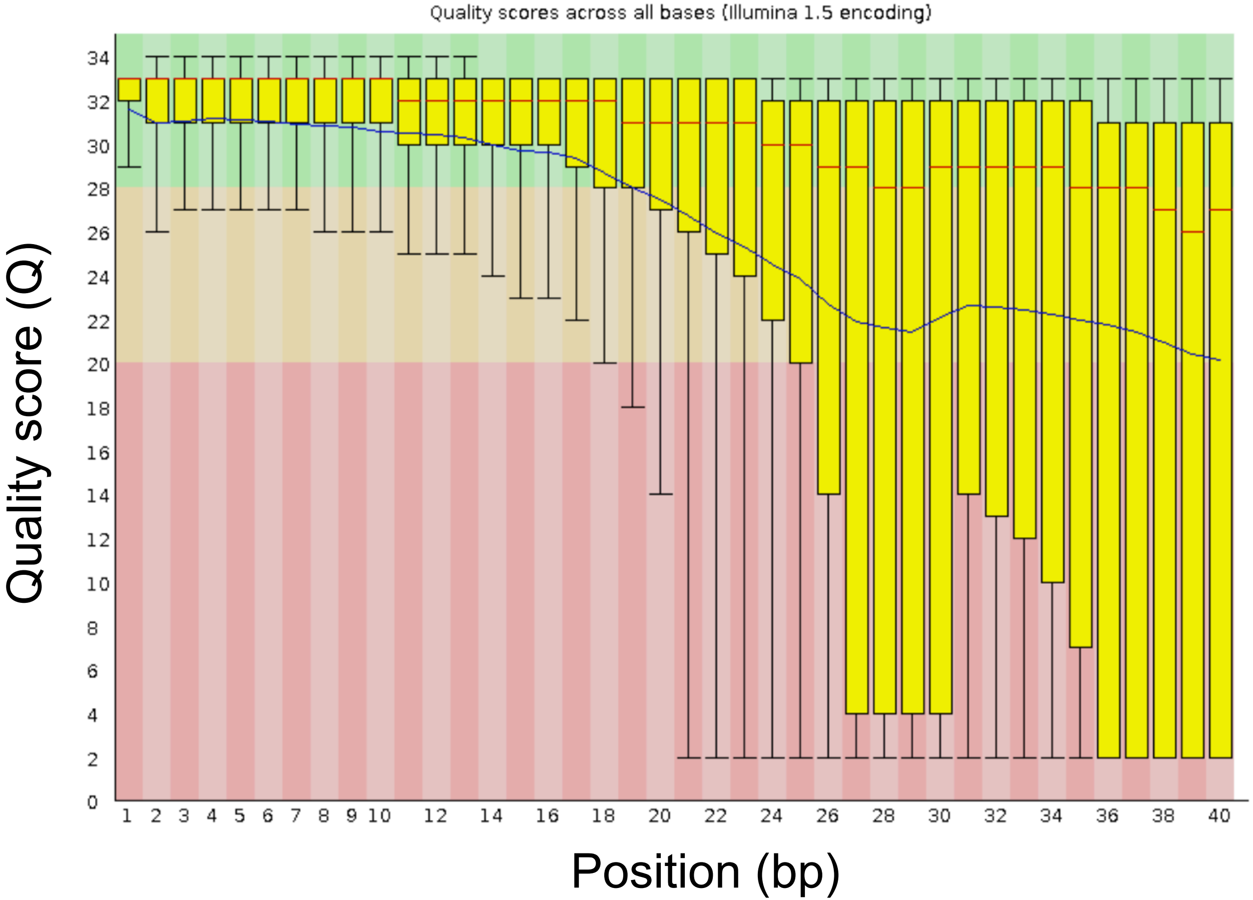
HOWEVER a visualisation of data quality should be carried out at the beginning of ANY analysis.

FastQC

Do you want to manually assess the quality of each read? All 1000000?! NO!

FastQC is a software programme that analyses quality and produces a report showing key information.


Quality scores across all bases (Illumina 1.5 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
Quality score (Q)
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
Position (bp)


Quality scores across all bases (Illumina 1.5 encoding)
Quality score (Q)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
Position (bp)

Kmer

A “string” of letters (sequence) that can be any length

For example:

ATGC

ATCGCTTGTGTGACCAGTGATTGACGATGGTCATTATGTC

Datasets

Exercise 1: Genome sequence of the bacteria *Bartonella*

Exercise 2: Amplicon sequencing of 16S rRNA

Exercise 3: RAD Sequencing data

Exercise 4: Amplicon sequencing of COI genes

Exercise 5: microRNA sequencing

Exercise 6: PacBio data from *Arabidopsis*

How to Run a Programme - Command Line

```
$ program_name [OPTIONS] <files>
```

For example:

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

Generally,

[ ]	optional
< >	mandatory, e.g. files
|	OR

Genome sequence of the bacteria *Bartonella*

Exercise One

There are 10,000 sequences of 38 nucleotide length

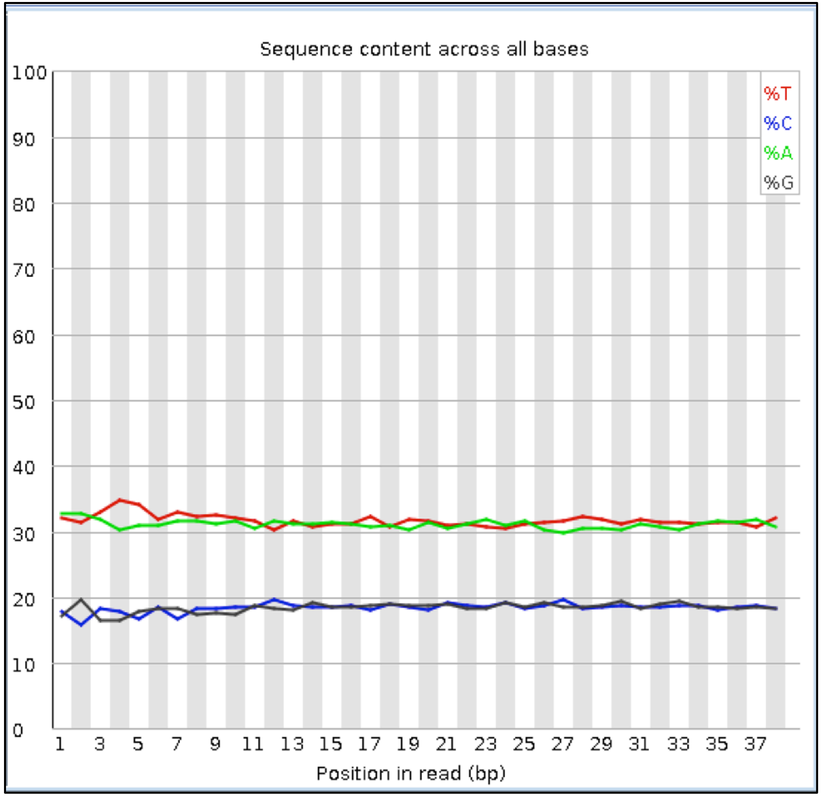
The GC content is 37%

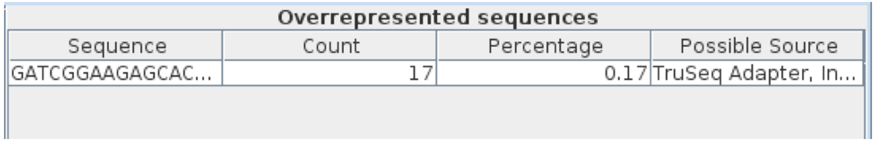
Quality score is > Q30, which = 1 error in 1000 so base call quality is 99.9%

Would you think this is good quality sequencing data?

Genome sequence of the bacteria *Bartonella*

Exercise One - Are we worried about this data?





Overrepresented sequences

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

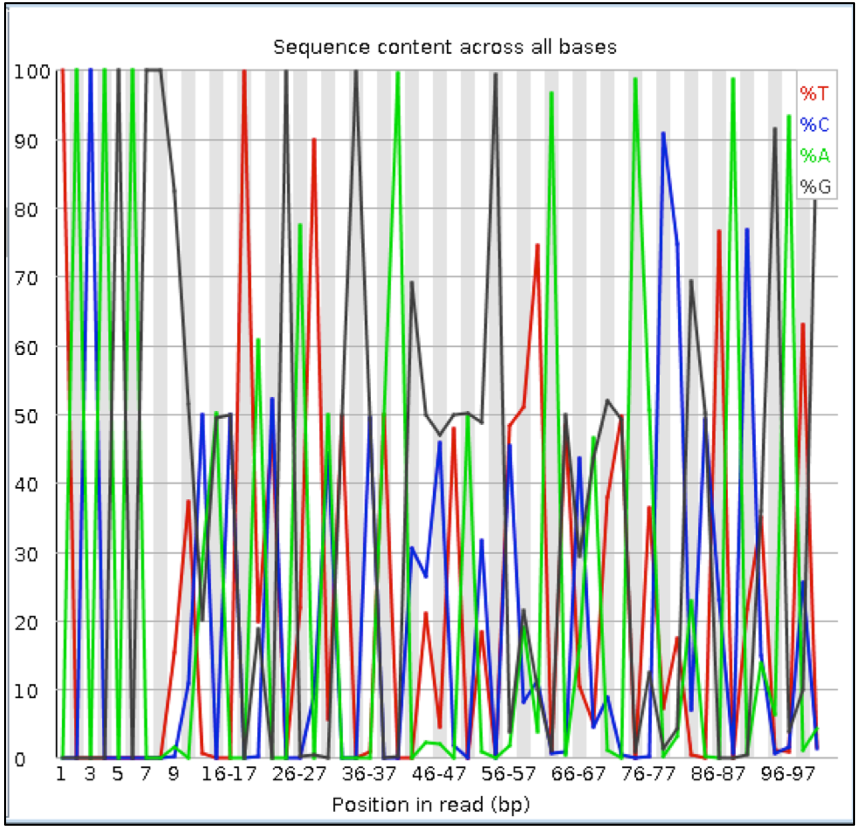
No - This shows the GC content we expect and very few adaptors

Exercise Two - Per Base Sequence Quality

Conserved sequence at the beginning of the reads:

TACAGAGG

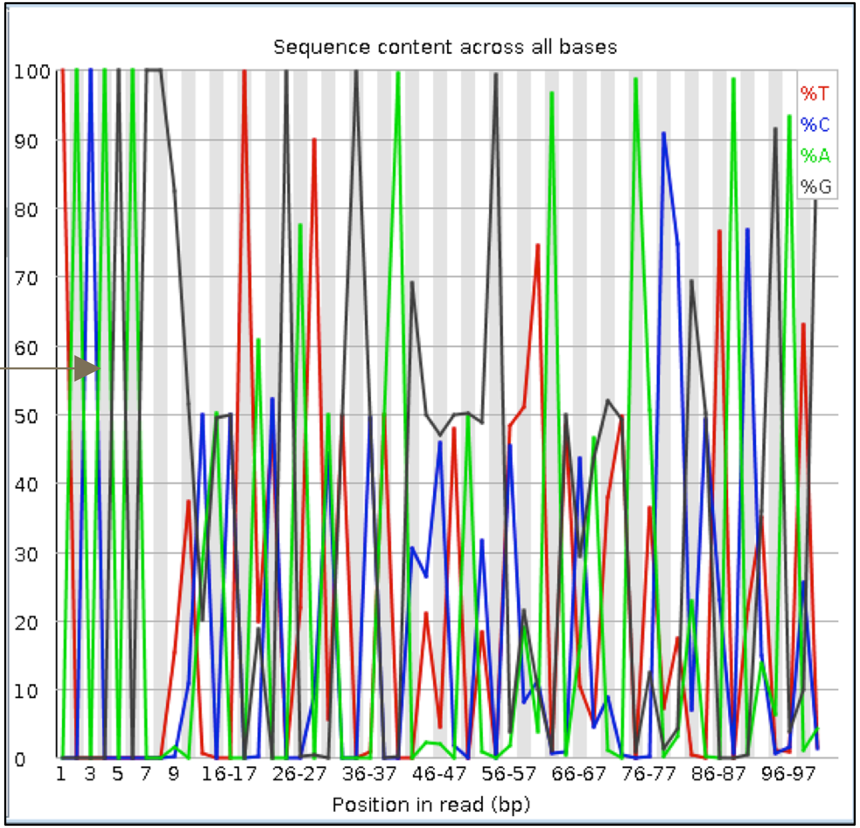
Lots of sequences with very similar sequence.



Exercise Two - Filtering/Trimming?

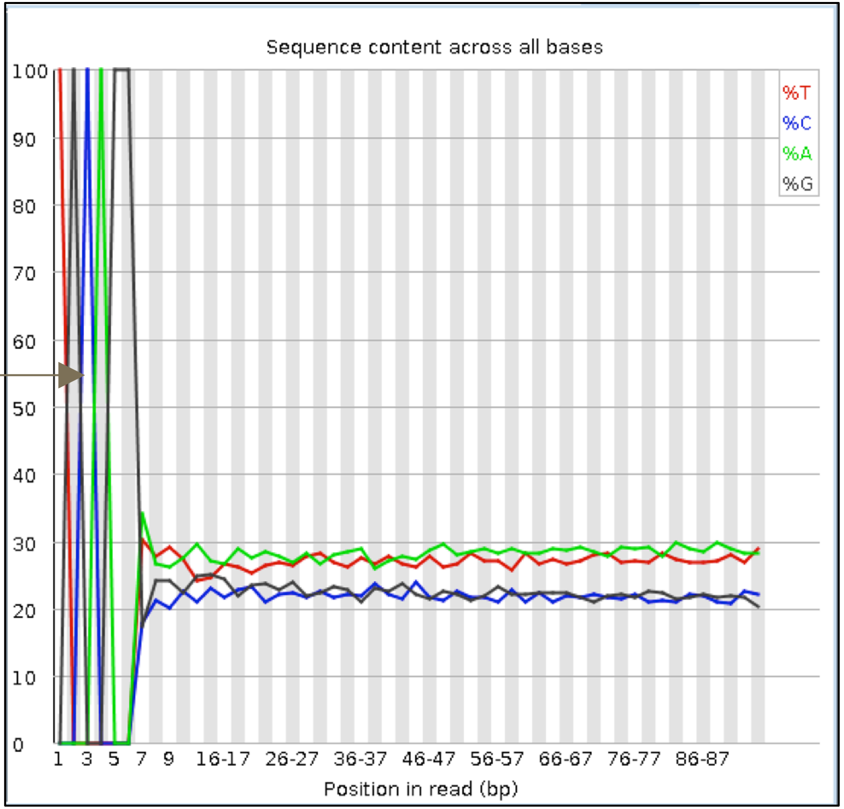
Amplicon sequencing of 16S rRNA

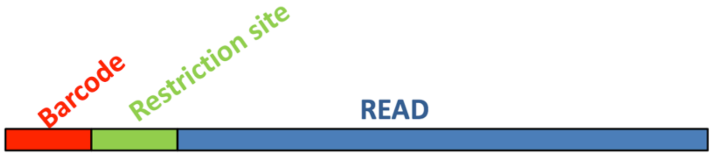
Filter Low Quality & Remove the Primer Sequence



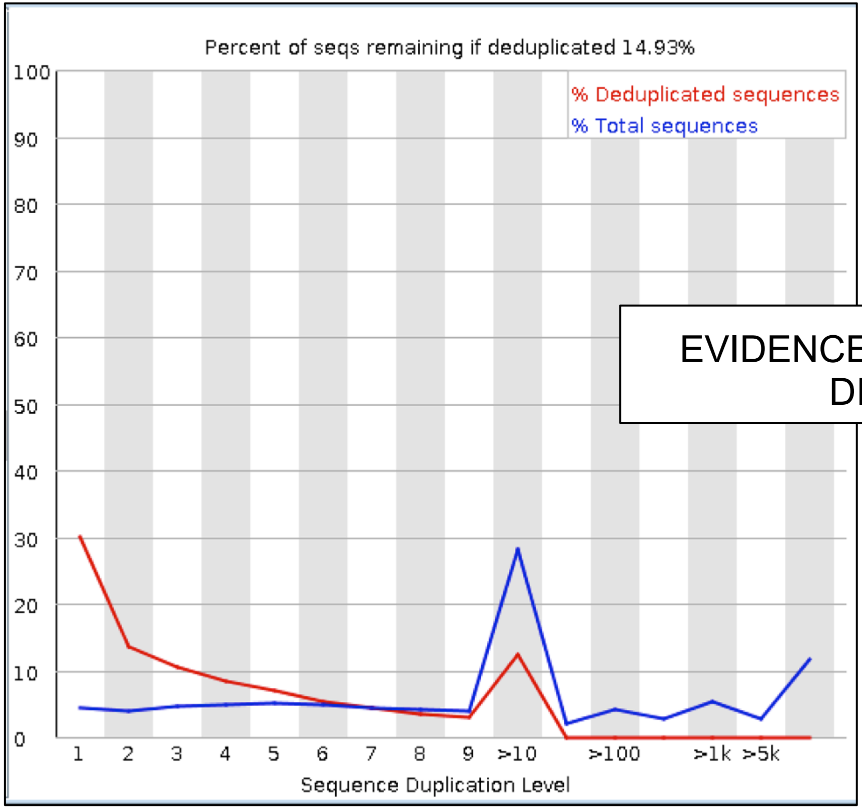
Exercise Three - Dataset 1: The First 6 Bases

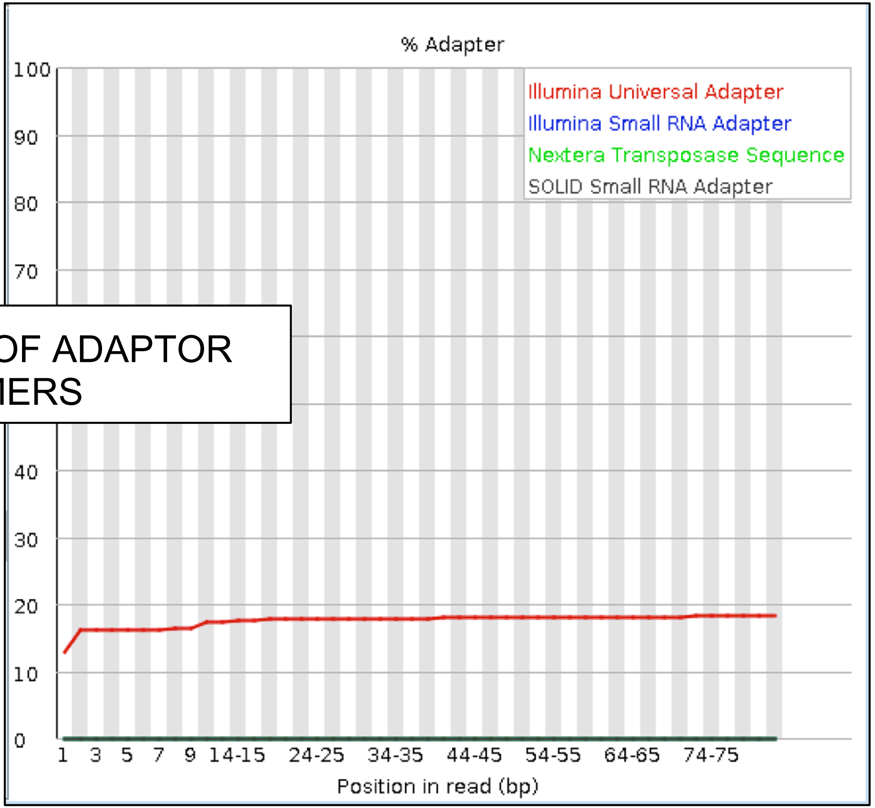
Restriction Enzyme Digestion Site





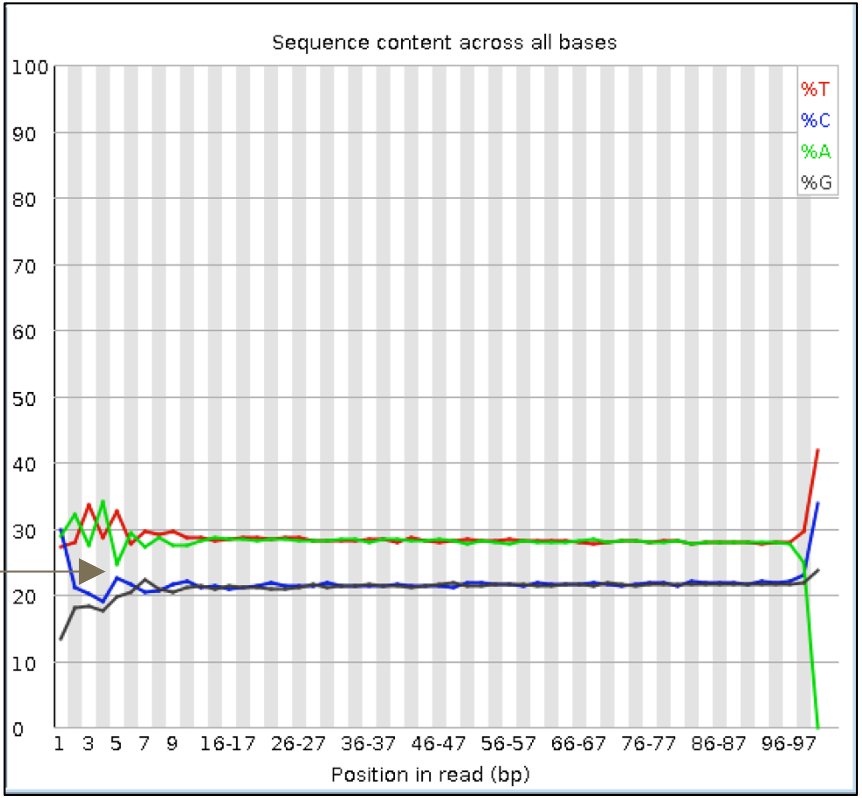
Exercise Three - Dataset 2: Read 1





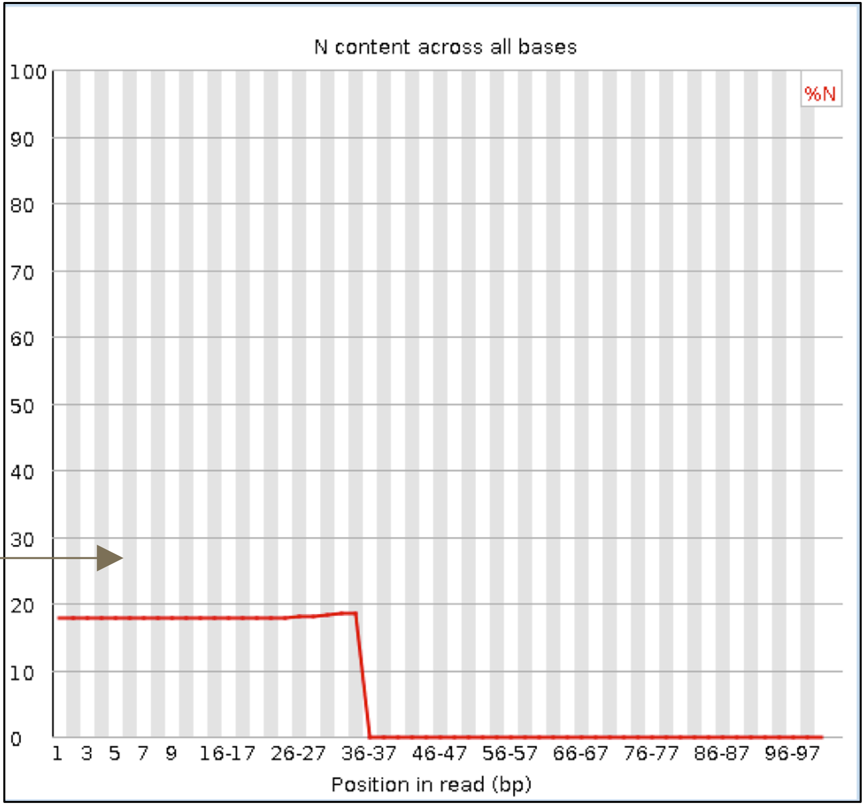
EVIDENCE OF ADAPTOR DIMERS

Exercise Three - Dataset 2: Read 2



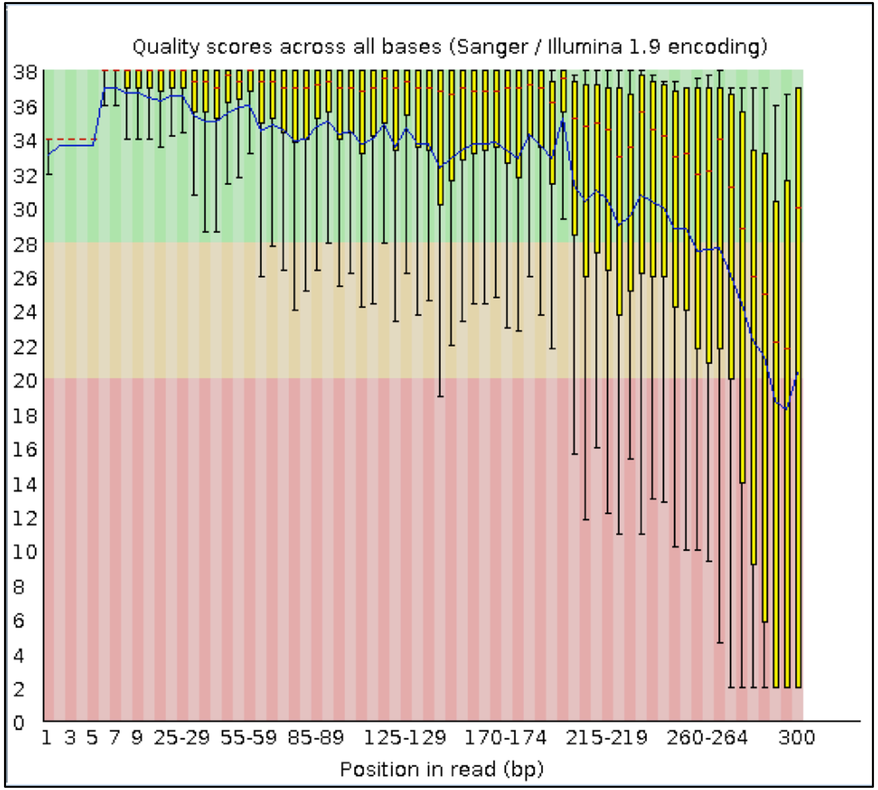
No restriction enzyme digestion site, therefore not double digested

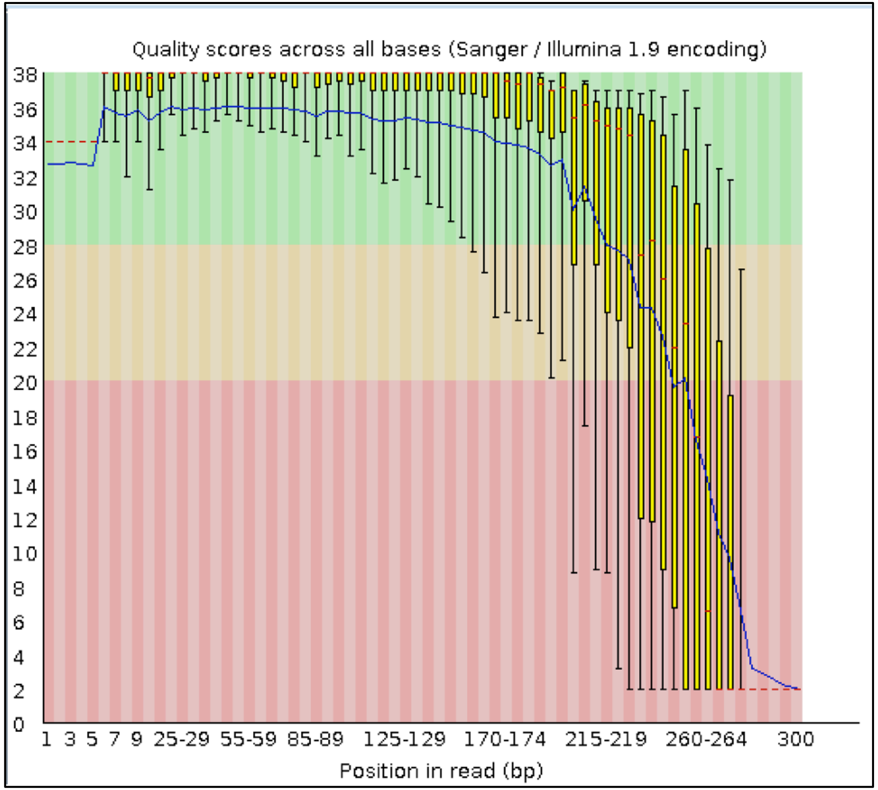
Exercise Three - Dataset 2: Read 2



Lots of Ns at the beginning of sequence

Exercise Four - Why is read 2 poorer quality?





Exercise Four - When would you be less strict?

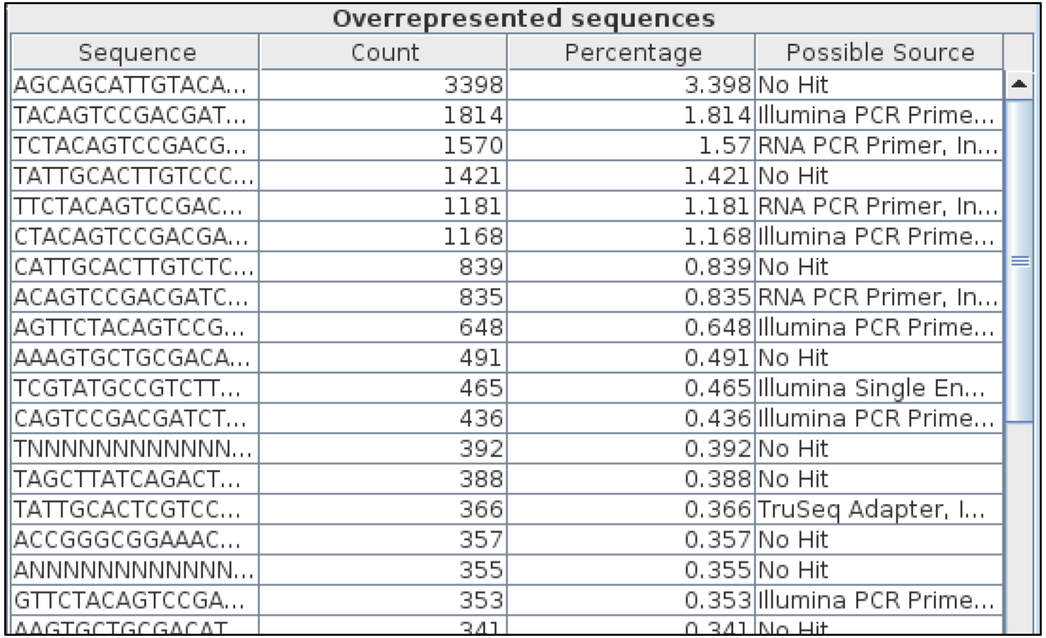
Quality vs. Quantity - Do you have enough data after filtering?

SOME alignments - RNA sequencing vs. SNP calling

Does your downstream software handle quality scores and filtering? E.G. QIIME and Stacks

Exercise Five - Adaptor Contamination

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 Five

Are we happy with the final data?

What about the adaptor sequences that remain?

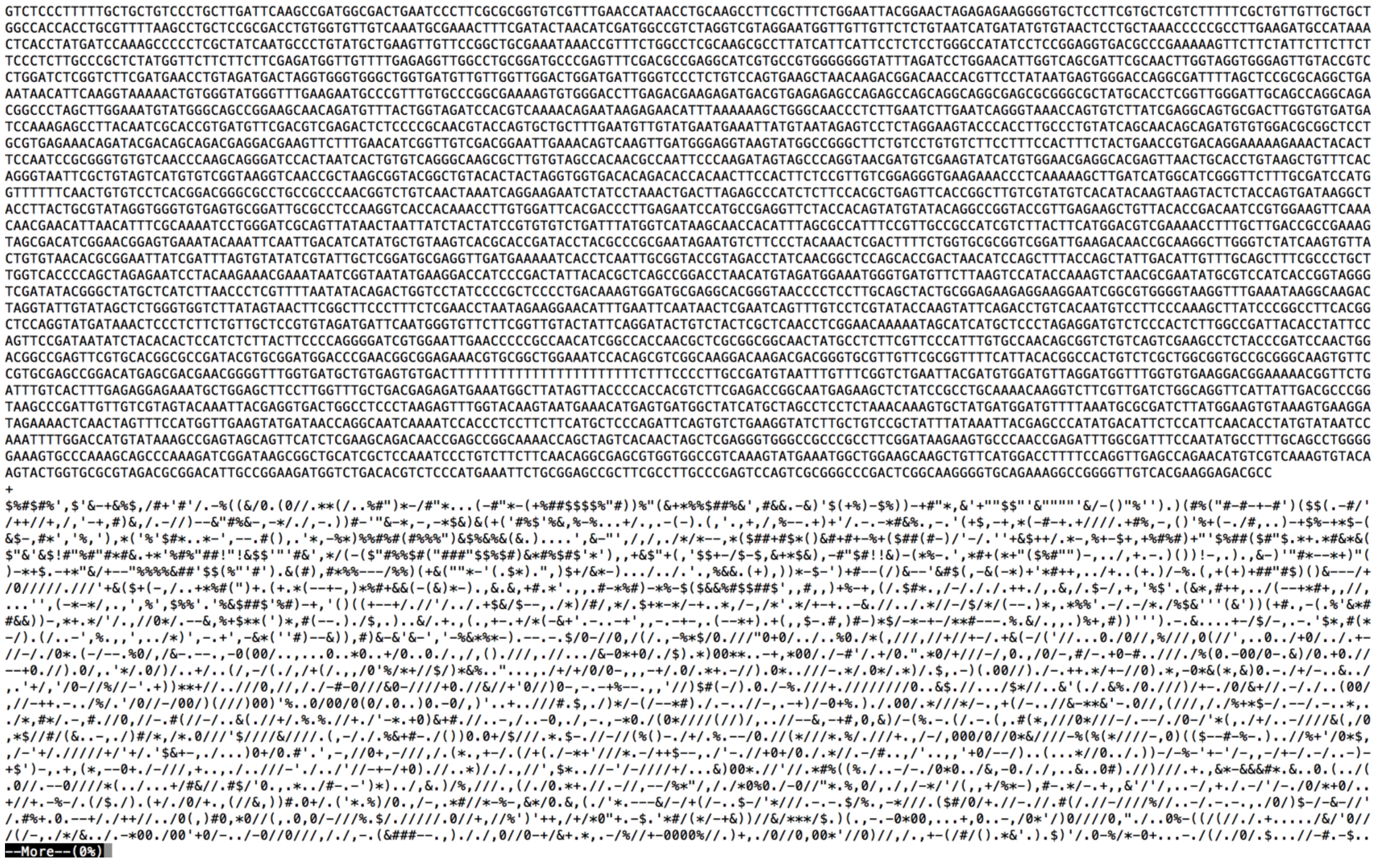
- Could be real adaptors that are too small to match based on software parameter OR could be real sequence

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.

Exercise Six

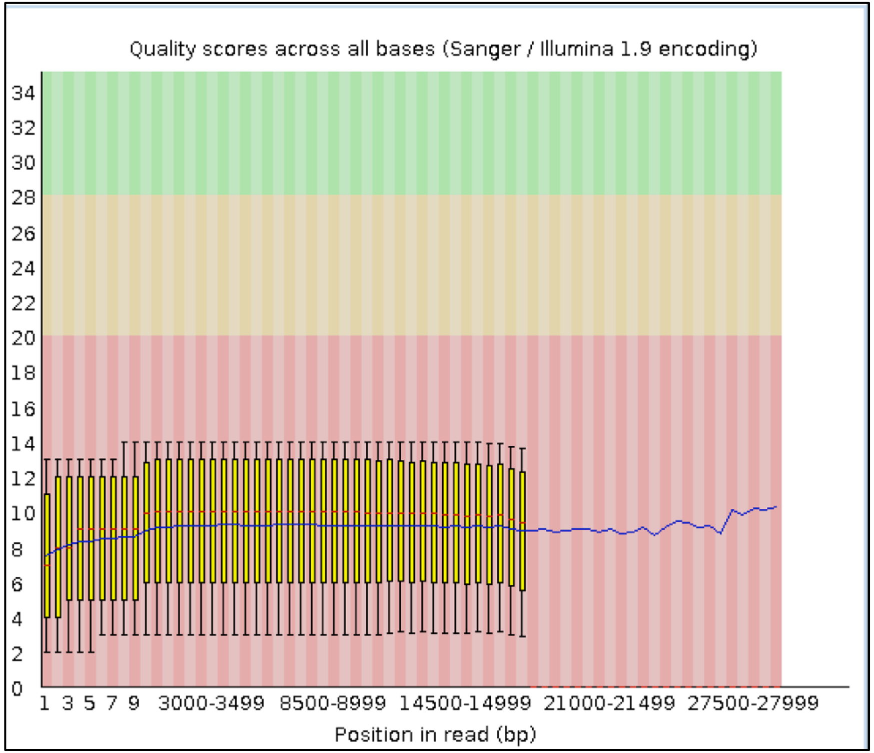
PacBio data

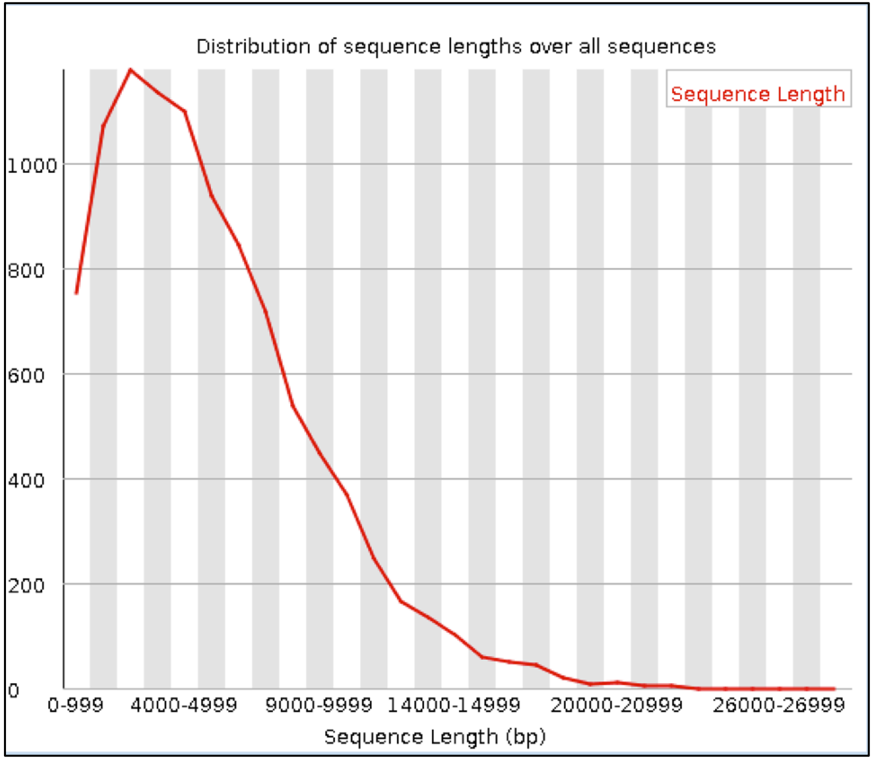


```
GTCTCCCTTTTTGCTGCTGTCCCTGCTTGATTCAAGCCGATGGCGACTGAATCCCTTCGCGCGGTGTCGTTTGAACCATAACCTGCAAGCCTTCGCTTTCTGGAATTACGGAACTAGAGAGAAGGGGTGCTCCTTCGTGCTCGTCTTTTTCGCTGTTGTTGCTGCT
GGCCACCACCTGCGTTTTAAGCCTGCTCCGCGACCTGTGGTGTTGTCAAATGCGAAACTTTCGATACTAACATCGATGGCCGTCTAGGTCGTAGGAATGGTTGTTGTTCTCTGTAATCATGATATGTGTAACTCCTGCTAAACCCCCGCCTTGAAGATGCCATAAA
CTCACCTATGATCCAAAGCCCCCTCGCTATCAATGCCCTGTATGCTGAAGTTGTTCCGGCTGCGAAATAAACCGTTTCTGGCCTCGCAAGCGCCTTATCATTCATTCCTCTCCTGGGCCATATCCTCCGGAGGTGACGCCCGAAAAAGTTCTTCTATTCTTCTTCT
TCCCTCTTGCCCGCTCTATGGTTCTTCTTCTTCGAGATGGTTGTTTTGAGAGGTTGGCCTGCGGATGCCCGAGTTTCGACGCCGAGGCATCGTGCCGTGGGGGGGTATTTAGATCCTGGAACATTGGTCAGCGATTCGCAACTTGGTAGGTGGGAGTTGTACCGTC
CTGGATCTCGGTCTTCGATGAACCTGTAGATGACTAGGTGGGTGGGCTGGTGATGTTGTTGGTTGGACTGGATGATTGGGTCCCTCTGTCCAGTGAAGCTAACAAGACGGACAACCACGTTCCTATAATGAGTGGGACCAGGCGATTTTAGCTCCGCGCAGGCTGA
AATAACATTCAAGGTAAAAACTGTGGGTATGGGTTTGAAGAATGCCCGTTTGTGCCCGGCGAAAAGTGTGGGACCTTGAGACGAAGAGATGACGTGAGAGAGCCAGAGCCAGCAGGCAGGCGAGCGCGGGCGCTATGCACCTCGGTTGGGATTGCAGCCAGGCAGA
CGGCCCTAGCTTGGAAATGTATGGGCAGCCGGAAGCAACAGATGTTTACTGGTAGATCCACGTCAAAACAGAATAAGAGAACATTTAAAAAAGCTGGGCAACCCTCTTGAATCTTGAATCAGGGTAAACCAGTGTCTTATCGAGGCAGTGCGACTTGGTGTGATGA
TCCAAAGAGCCTTACAATCGCACCGTGATGTTCGACGTCGAGACTCTCCCCGCAACGTACCAGTGCTGCTTTGAATGTTGTATGAATGAAATTATGTAATAGAGTCCTCTAGGAAGTACCCACCTTGCCCTGTATCAGCAACAGCAGATGTGTGGACGCGGCTCCT
GCGTGAGAAACAGATACGACAGCAGACGAGGACGAAGTTCTTTGAACATCGGTTGTCGACGGAATTGAAACAGTCAAGTTGATGGGAGGTAAGTATGGCCGGGCTTCTGTCCTGTGTCTTCCTTTCCACTTTCTACTGAACCGTGACAGGAAAAAGAAACTACACT
TCCAATCCGCGGGTGTGTCAACCCAAGCAGGGATCCACTAATCACTGTGTCAGGGCAAGCGCTTGTGTAGCCACAACGCCAATTCCCAAGATAGTAGCCCAGGTAACGATGTCGAAGTATCATGTGGAACGAGGCACGAGTTAACTGCACCTGTAAGCTGTTTCAC
AGGGTAATTCGCTGTAGTCATGTGTCGGTAAGGTCAACCGCTAAGCGGTACGGCTGTACACTACTAGGTGGTGACACAGACACCACAACTTCCACTTCTCCGTTGTCGGAGGGTGAAGAAACCCTCAAAAAGCTTGATCATGGCATCGGGTTCTTTGCGATCCATG
GTTTTTTCAACTGTGTCCTCACGGACGGGCGCCTGCCGCCCAACGGTCTGTCAACTAAATCAGGAAGAATCTATCCTAAACTGACTTAGAGCCCATCTCTTCCACGCTGAGTTCACCGGCTTGTCGTATGTCACATACAAGTAAGTACTCTACCAGTGATAAGGCT
ACCTTACTGCGTATAGGTGGGTGTGAGTGCGGATTGCGCCTCCAAGGTCACCACAAACCTTGTGGATTCACGACCCTTGAGAATCCATGCCGAGGTTCTACCACAGTATGTATACAGGCCGGTACCGTTGAGAAGCTGTTACACCGACAATCCGTGGAAGTTCAAA
CAACGAACATTAACATTTCGCAAAATCCTGGGATCGCAGTTATAACTAATTATCTACTATCCGTGTGTCTGATTTATGGTCATAAGCAACCACATTTAGCGCCATTTCCGTTGCCGCCATCGTCTTACTTCATGGACGTCGAAAACCTTTGCTTGACCGCCGAAAG
TAGCGACATCGGAACGGAGTGAAATACAAATTCAATTGACATCATATGCTGTAAGTCACGCACCGATACCTACGCCCGCGAATAGAATGTCTTCCCTACAAACTCGACTTTTCTGGTGCGCGGTCGGATTGAAGACAACCGCAAGGCTTGGGTCTATCAAGTGTTA
CTGTGTAACACGCGGAATTATCGATTTAGTGTATATCGTATTGCTCGGATGCGAGGTTGATGAAAAATCACCTCAATTGCGGTACCGTAGACCTATCAACGGCTCCAGCACCGACTAACATCCAGCTTTACCAGCTATTGACATTGTTTGCAGCTTTCGCCCTGCT
TGGTCACCCCAGCTAGAGAATCCTACAAGAAACGAAATAATCGGTAATATGAAGGACCATCCCGACTATTACACGCTCAGCCGGACCTAACATGTAGATGGAAATGGGTGATGTTCTTAAGTCCATACCAAAGTCTAACGCGAATATGCGTCCATCACCGGTAGGG
TCGATATACGGGCTATGCTCATCTTAACCCTCGTTTTAATATACAGACTGGTCCTATCCCCGCTCCCCTGACAAAGTGGATGCGAGGCACGGGTAACCCCTCCTTGCAGCTACTGCGGAGAAGAGGAAGGAATCGGCGTGGGGTAAGGTTTGAAATAAGGCAAGAC
TAGGTATTGTATAGCTCTGGGTGGTCTTATAGTAACTTCGGCTTCCCTTTCTCGAACCTAATAGAAGGAACATTTGAATTCAATAACTCGAATCAGTTTGTCCTCGTATACCAAGTATTCAGACCTGTCACAATGTCCTTCCCAAAGCTTATCCCGGCCTTCACGG
CTCCAGGTATGATAAACTCCCTCTTCTGTTGCTCCGTGTAGATGATTCAATGGGTGTTCTTCGGTTGTACTATTCAGGATACTGTCTACTCGCTCAACCTCGGAACAAAAATAGCATCATGCTCCCTAGAGGATGTCTCCCACTCTTGGCCGATTACACCTATTCC
AGTTCCGATAATATCTACACACTCCATCTCTTACTTCCCCAGGGGATCGTGGAATTGAACCCCCGCCAACATCGGCCACCAACGCTCGCGGCGGCAACTATGCCTCTTCGTTCCCATTTGTGCCAACAGCGGTCTGTCAGTCGAAGCCTCTACCCGATCCAACTGG
ACGGCCGAGTTCGTGCACGGCGCCGATACGTGCGGATGGACCCGAACGGCGGAGAAACGTGCGGCTGGAAATCCACAGCGTCGGCAAGGACAAGACGACGGGTGCGTTGTTCGCGGTTTTCATTACACGGCCACTGTCTCGCTGGCGGTGCCGCGGGCAAGTGTTC
CGTGCGAGCCGGACATGAGCGACGAACGGGGTTTGGTGATGCTGTGAGTGTGACTTTTTTTTTTTTTTTTTTTTTTTTTCTTTCCCCTTGCCGATGTAATTTGTTTCGGTCTGAATTACGATGTGGATGTTAGGATGGTTTGGTGTGAAGGACGGAAAAACGGTTCTG
ATTTGTCACTTTGAGAGGAGAAATGCTGGAGCTTCCTTGGTTTGCTGACGAGAGATGAAATGGCTTATAGTTACCCCACCACGTCTTCGAGACCGGCAATGAGAAGCTCTATCCGCCTGCAAAACAAGGTCTTCGTTGATCTGGCAGGTTCATTATTGACGCCCGG
TAAGCCCGATTGTTGTCGTAGTACAAATTACGAGGTGACTGGCCTCCCTAAGAGTTTGGTACAAGTAATGAAACATGAGTGATGGCTATCATGCTAGCCTCCTCTAAACAAAGTGCTATGATGGATGTTTTAAATGCGCGATCTTATGGAAGTGTAAAGTGAAGGA
TAGAAAACTCAACTAGTTTCCATGGTTGAAGTATGATAACCAGGCAATCAAAATCCACCCTCCTTCTTCATGCTCCCAGATTCAGTGTCTGAAGGTATCTTGCTGTCCGCTATTTATAAATTACGAGCCCATATGACATTCTCCATTCAACACCTATGTATAATCC
AAATTTTGGACCATGTATAAAGCCGAGTAGCAGTTCATCTCGAAGCAGACAACCGAGCCGGCAAAACCAGCTAGTCACAACTAGCTCGAGGGTGGGCCGCCCGCCTTCGGATAAGAAGTGCCCAACCGAGATTTGGCGATTTCCAATATGCCTTTGCAGCCTGGGG
GAAAGTGCCCAAAGCAGCCCAAAGATCGGATAAGCGGCTGCATCGCTCCAAATCCCTGTCTTCTTCAACAGGCGAGCGTGGTGGCCGTCAAAGTATGAAATGGCTGGAAGCAAGCTGTTCATGGACCTTTTCCAGGTTGAGCCAGAACATGTCGTCAAAGTGTACA
AGTACTGGTGCGCGTAGACGCGGACATTGCCGGAAGATGGTCTGACACGTCTCCCATGAAATTCTGCGGAGCCGCTTCGCCTTGCCCGAGTCCAGTCGCGGGCCCGACTCGGCAAGGGGTGCAGAAAGGCCGGGGTTGTCACGAAGGAGACGCC
+
$%#$#%',$'&-+&%$,/#+'#'/.-%((&/0.(0//.**(/..%#")*-/#"*...(-#"*-(+%##$$$$%"#))%"(&+*%%$##%&',#&&.-&)'$(+%)-$%))-+#"*,&'+""$$"'&"""""&/-()"%'').)(#%("#-#-+-#')($$(.-#/'
/++//+,/,'-+,#)&,/.-//)--&"#%&-,-*/./,-.))#-'"&-*,-,-*$&)&(+('#%$'%&,%-%...+/.,.-(-).(,'.,+,/,%--.+)+'/.-.-*#&%.,-.'(+$,-+,*(-#-+.+////.+#%,-,()'%+(-./#,..)-+$%-+*$-(
&$-,#*','%,'),*('%'$#*..*-',--.#(),.'*,-%*)%%#%#(#%%%")&$%&%&(&.)....',&-"',/,/,./*/*--,*($##+#$*()&#+#+-%+($##(#-)/'-/.''+&$++/.*-,%+-$+,+%#%#)+"'$%##($#"$.*+.*#&*&(
$"&'&$!#"%#"#*#&.+*'%#%"##!"!&$$'"'#&',*/(-($"#%%$#("###"$$%$#)&*#%$#$'*'),,+&$"+(,'$$+-/$-$,&+*$&),-#"$#!!&)-(*%-.',*#+(*+"($%#"")-,../,+.-.)())!-,.).,&-)'"#*--*+)"(
)-*+$.-+*"&/+--"%%%%&##'$$(%"'#').&(#),#*%%---/%%)(+&(""*-'(.$*).",)$+/&*-).../../.',%&&.(+),))*-$-')+#--(/)&--'&#$(,-&(-*)+'*#++,../+..(+.)/-%.(,+(+)+##"#$)()&---/+
/0/////.///'+&($+(-,/..+*%#(")+.(+.*(--+-,)*%#+&&(-(&)*-).,&.&,+#.*'.,,.#-*%#)-*%-$($&&%#$$##$',,#,,)+%-+,(/.$#*.,/-/././.++./,.&,/.$-/,+,'%$'.(&*,#++,../(--+*#+,,//,
...'',(-*-*/,.,',%',$%%'.'%&$##$'%#)-+,'()((+--+/.//'/../.+$&/$--,./*)/#/,*/.$+*-*/-+..*,/-,/*'.*/+-+..-&.//../.*//-/$/*/(--.)*,.*%%'.-/.-/*./%$&'''(&'))(+#.,-(.%'&*#
#&&))-,*+.*/'/.,//0*/.--&,%+$**(')*,#(--.)./$,.)..&/.+.,(.,+-.+/*(-&+'.-..-+',,-.-+-,.(--*+).+(,,$-.#,)#-)*$/-*-+-/**#---.%.&/.,,.)%+,#))''').-.&....+-/$/-,.-.'$*,#(*
-/).(/..-',%.,,',../*)',-.+',-&*(''#)--&)),#)&-&'&-','-%&*%*-).--.-.$/0-//0,/(/.,-%*$/0.///"0+0/../..%0./*(,///,//+//+-/.+&(-/('//...0./0//,%///,0(//',..0../+0/../.+-
//-/./0*.(-/--.%0/,/&-.--.,-0(00/..,...0..*0..+/0..0./.,/,().///,.//.../&-0*+0/./$).*)00**..-+,*00/./-#'/.+/0.".*0/+///-/,0.,/0/-,#/-.+0-#..///./%(0.-00/0-.&)/0.+0.//
--+0.//).0/,.'*/.0/)/..+/..(/,-/(./,/+(/.,,/0'%/*+//$/)*&%.."....,./+/+/0/0-,,,-+/.0/.*+.-//).0*..///-.*/.0*/.*)/.$,.-)(.00//)./-.++.*/+-//0).*,-0*&(*,&)0.-./+/-..&../
,.'+/,'/0-//%//-'.+))**+//..///0,//,/./-#-0///&0-////+0.//&//+'0//)0-,-.-+%--.,,'//)$#(-/).0./-%.///+.///////0..&$.//.../$*//..&'(./.&%./0.///)/+-./0/&+//.-/./..(00/
,//-++.-../%/.'/0//-/00/)(///)00)'%..0/00/0(0/.0..)0.-0/,)'..+..///#.$,./)*/-(/--*#)./.-..//-,.-+)/-0+%.)./.00/.*///*/-.,+(/-..//&-**&'-.0//,(///,/./%+*$-/.--/.-..*,.
./*,#*/.-,#.//0,//-.#(//-/..&(.//+/.%.%.//+./'-*.+0)&+#.//..-,/..-0,./,-.,-*0./(0*////(//)/,..//--&,-+#,0,&)/-(%.-.(/.-.(,.#(*,///0*///-/.--/./0-/'*(,./+/..-////&(,/0
,*$//#/(&..-,./)#/*,/*.0///'$////&////.(,-/./.%&+#-./())0.0+/$///.*.$-.//-//(%()-./+/.%.--/0.//(*///*.%/.///+.,/-/,000/0//0*&////-%(%(*////-,0)(($--#-%-.)..//%+'/0*$,
,/-'+/./////+/'+/.'$&+-,./...)0+/0.#'.',-,//0+,-///,/.(*.,+-/.(/+(./-*+'///*.-/++$--,./'-.//+0+/0./.*//.-/#..,/'..,,'+0/--/)..(...*//0../.))-/-%-'+-'/-,,-/+-/.-/..-)-
+$')-,.+,(*,--0+./-///,+..,./..///-'./../'//-+-/+0).//..*)/./.,//',$*..//-'/-////+/...&)00*.//'//.*#%((%./..-/-./0*0../&,-0././,..&..0#).//)///.+.,&*-&&&#*.&..0.(../(
.0//.--0////*(../...+/#&//.#$/'0.,.*../#-.-')*)../,&.)/%,///.,(/./0.*+.//.-//,--/%*"/,/./*0%0./-0//"*.%,0/,./,/-*/'/(,,+/%*-),#-.*/-.+,,&'/'/,..-/,+./.-/'/-./0/*+0/..
+//+.-%-/.(/$./).(+/./0/+.,(//&,))#.0+/.('*.%)/0.,/-,.*#//*-%-,&*/0.&,(./'*.---&/-/+(/-..$-/'*///.-.-.$/%.,-*///.($#/0/+.//-.//.#(/.//-////%//..-/.-.-.,./0/)$-/-&-//'
/.#%+.0.--+/./++//../0(,)#0,*0//(,.0,0/-///%.$/./////.0//+,//%')'++,/+/*0"+.-$.'*#/(*/-+&))//&/***/$.)(.,-.-0*00,...+,0..-,/0*'/)0////0,"./..0%-((/(//./.+...../&/'0//
/(/-,./*/&../.-*00./00'+0/-../-0//0///,/./,-.(&###--.,)././,0//0-+/&+.*,.-/%//+-0000%//.)+,./0//0,00*'//0)//,/.,+-(/#/().*&'.).$)'/.0-%/*-0+...-./(/./0/.$...//-#.-$..
--More--(0%)
```

Exercise Six

PacBio data





Exercise Six

Why can’t we filter PacBio data based on quality?

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…

No two datasets are the same!