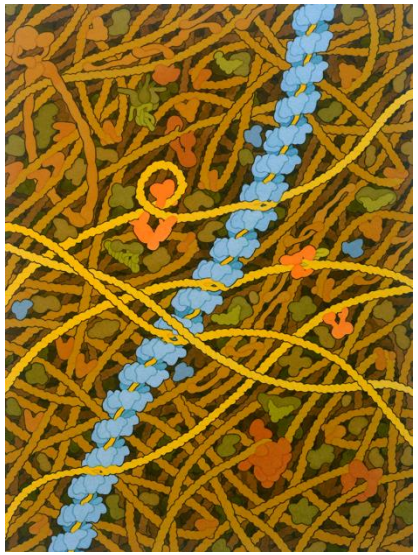
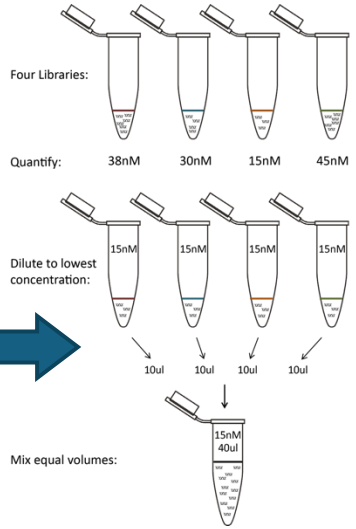
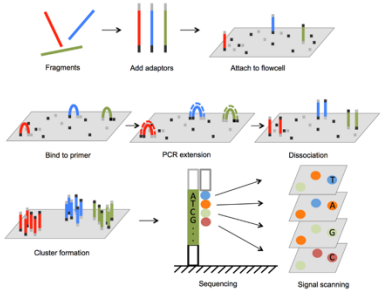
NICHD RNAseq Course

Week 2

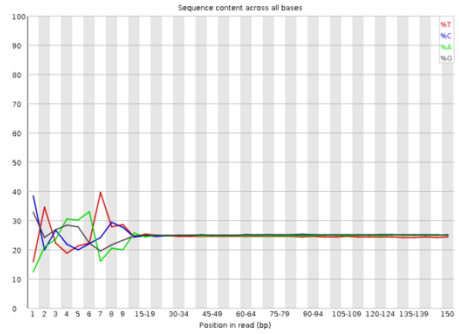
February 2025





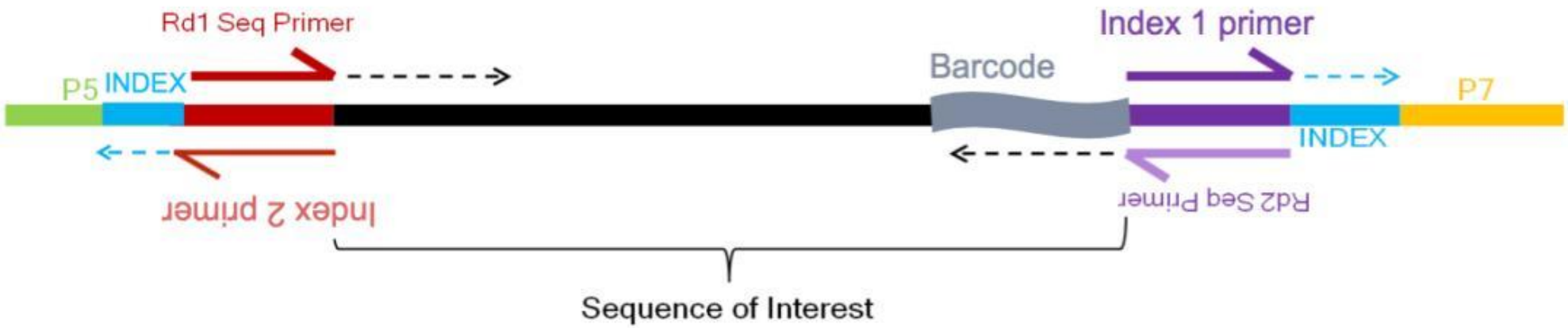



```
@SRR948304.5141 UNC14-SN744:253:D135LACXX:5:1101:4054:4287
AGGACTTTGGCTGTCCCCAACCGTACAGGTGGGTCTCTCCCTCATGGA
+
CCCFFFFFHHHHHGHIIJIJJJIEHGIJJCGGJAFHEIJDHGIFGGE<
@SRR948304.5160 UNC14-SN744:253:D135LACXX:5:1101:4625:4251
CCACGAAGTCAAGATCGCCGACAAGGCCTTCCTGATGAAGCAGAAGTT
+
BCCFFFFFHGHHHJJJJJJJJIJGJJJJJJJJJJFIJJGIJJIJJJIGI
@SRR948304.5196 UNC14-SN744:253:D135LACXX:5:1101:5864:4425
GGAACGCCTACTGGTACTATCTGAACATGGACTACTCCTTCTTCCTGG
+
CCCFFFFFFHHHHJGHIIJJJJJJJJJIIIIIIIIIIIIIIIIIIIII
@SRR948304.5200 UNC14-SN74
CAGGAAGAAGGAGTAGTCCATGTTCA
+
@@CFFBDDHHHBHIHIFGHHIGFFHI
@SRR948304.5215 UNC14-SN74
CCTTCTTCAACGACTACTACACCAA
+
@@@FFFFFHGDHFIEHGIJIFEHIHH
@SRR948304.5247 UNC14-SN74
CTACAGCTTCCGCAAGAACTACTAC
+
BB@FFFFFHHHHFIJIJJJJIJIJJ
@SRR948304.5250 UNC14-SN74
CCTGGTTGAACTCGTAGATATTCTCGCGCAGAATGAACTTGTCCTCGG
+
CCCFFFFFHHHHHJHGHIJIJJJJJJJJJJJIIIJJIJJJJJJIJJJE
@SRR948304.5259 UNC14-SN744:253:D135LACXX:5:1101:8714:4295
GTAGTCGAAGTGGCCGTAGTTGGAGTAGCTCTTGTACTTGTAGTAGGG
+
@@@FFFDFFBBFAHGIHIGGGHHIJFFGHIJIICFCHIIJEGIIEGCC
@SRR948304.5261 UNC14-SN744:253:D135LACXX:5:1101:8536:4428
TGCTCAACGACAAGATGATCTTCCAGGACGGCAAGTTCGTGTGGGACA
+
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJIIJJIIIIJJJHJJJIJJJJ
@SRR948304.5286 UNC14-SN744:253:D135LACXX:5:1101:9565:4401
CACCTTACTAGCTAGGTATCAGACACTCTAAAAACATCCGCGCTCATT
+
:?;ADD?B?,AA?<CEEEEEHF?AFDFDHAHEF9EHEGGIIEDFHHCC
@SRR948304.5309 UNC14-SN744:253:D135LACXX:5:1101:10403:4285
CGCACGGATTCGGCGAGATCCCCGAGTTCTCCTGGTACCAGCAGATCG
+
@@@DDDDDFHHFDEBHGFCHGIGBGHHCEH@FHAHAFHEIIIGEHFH9
@SRR948304.5310 UNC14-SN744:253:D135LACXX:5:1101:10382:4379
GGCGTTCCAGCCCAGATCCTCGGTGAAGTAGGACAGCTTGGACTCCTT
+
CCCFFFFFHHHHHJJJJJJJJJJHIJJJIIJJJJIJJJJJJJJJJJJJ
@SRR948304.5325 UNC14-SN744:253:D135LACXX:5:1101:10777:4369
CGACATGTCCACCAAAGCCGAGCTCGCCTGCGTCTACGCCTCCCTCAT
+
BCCFFFFFHHGHGIJJJJIJJJJIJGIIJIJGIJJ=FHEHGIJJJGHC
```



Sequencing Reads: The Raw Material of Bioinformatics

- A read is an inferred sequence (or base pair probabilities) corresponding to all or part of a single DNA/cDNA fragment.
- Can be accompanied by information about confidence
- Read length depends on the technology
- Often come with “accessories” like adaptor sequence, which can be trimmed



You may have seen a FASTA file

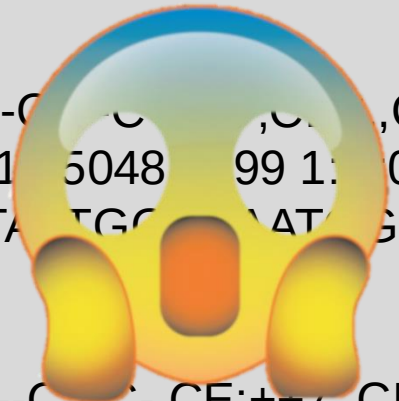
Header
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

https://www.researchgate.net/figure/A-sample-of-the-Multi-FASTA-file_fig1_309134977

Storing Reads: The Fastq Format

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCCAAGTAGC
GTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C[illegible],CD,CEFC,@E9<FCFCF?9
@M02286:19:000000000-AA549:1:1101[illegible]5048[illegible]99 1:[illegible]0:23
CCTACGGGTGGCTGCAGTGAGGAATA[illegible]TGC[illegible]AT[illegible]GTCGGAAGACTGATCCAGCCATGCCGC
GTGCAG
+
ABC@CC77CFCEG;F9<F89<9--C,CE,--C-[illegible]C-,CE:++7.,CF<,CEF,CFGGD8FFCFCFEGCF
@M02286:19:000000000-AA549:1:1101:11116:1322 1:N:0:23
CCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGACGGAAGTCTGACCGAGCAACGCCGC
GTGAGT
+
AAC<CCF+@@>CC,C9,F9C9@9-CFFFE@7@:+CC8-C@:7,@EFE,6CF:+8F7EFEEF@EGGGEEE
```



Storing Reads: The Fastq Format

Each sequencing read is represented by 4 lines

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCCAAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
@M02286:19:000000000-AA549:1:1101:15048:1299 1:N:0:23
CCTACGGGTGGCTGCAGTGAGGAATATTGGACAATGGTCGGAAGACTGATCCAGCCATGCCGCGTGCAG
+
ABC@CC77CFCEG;F9<F89<9--C,CE,--C-6C-,CE:++7:,CF<,CEF,CFGGD8FFCFCFEGCF
@M02286:19:000000000-AA549:1:1101:11116:1322 1:N:0:23
CCTACGGGAGGCAGCAGTAGGGAATCTTCGGCAATGGACGGAAGTCTGACCGAGCAACGCCGCGTGAGT
+
AAC<CCF+@@>CC,C9,F9C9@9-CFFFE@7@:+CC8-C@:7,@EFE,6CF:+8F7EFEEF@EGGGEEE
```

Read 1

Read 2

Read 3

Parts of A Single Read

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCC
AAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
```

1. @ followed by read ID and optional information about sequencing run (i.e. sample ID, sequencer)
2. Sequenced bases
3. + (optionally followed by the read ID and some additional info)
4. Quality scores for each base of the sequence encoded as ASCII Symbols

Parts of A Single Read

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCC
AAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
```

1. @ followed by read ID and optional information about sequencing run (i.e. sample ID, sequencer)
2. Sequenced bases
3. + (optionally followed by the read ID and some additional info)
4. Quality scores for each base of the sequence encoded as ASCII Symbols

Parts of A Single Read

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCC
AAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
```

1. @ followed by read ID and optional information about sequencing run (i.e. sample ID, sequencer)
2. Sequenced bases
3. + (optionally followed by the read ID and some additional info)
4. Quality scores for each base of the sequence encoded as ASCII Symbols

Parts of A Single Read

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCC
AAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
```

1. @ followed by read ID and optional information about sequencing run (i.e. sample ID, sequencer)
2. Sequenced bases
3. + (optionally followed by the read ID and some additional info)
4. Quality scores for each base of the sequence encoded as ASCII Symbols

Phred Scores

- Characters in last line of sequencing read encode for Phred Scores in ASCII
- One character (i.e. score) for every base in a read
- Measure of the quality of the identification (“base call”) for that particular base in a read
- Specifically, representation of the probability that the base call is incorrect – we want this probability to be low!

Quality score *characters* correspond to numeric *scores*

```
@SEQ_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|         |         |         |         |
0........10.......20........30........40
```

Quality score

A **GATC 1222430** *2-digit scores don't line up with sequence. Is that first one "1" or "12"?*

B **GATC 1,22,24,30** *Possible, but lots of extra characters*

C **GATC "79?** *Unambiguous & succinct*

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

Phred Scores

ASCII Character	Phred Quality Score	Probability of Error	Base Call Accuracy
+	10	1 in 10	90%
5	20	1 in 100	99%
?	30	1 in 1000	99.9%
I (the letter)	40	1 in 10,000	99.99%
2	50	1 in 100,000	99.999%

- We want high Phred scores, which mean a low probability of error
- Many reference tables for Phred scores exist online, but you will almost never be working with this data by hand!

Phred Scores

```
@M02286:19:000000000-AA549:1:1101:12677:1273 1:N:0:23
CCTACGGGTGGCAGCAGTGAGGAATATTGGTCAATGGACGGAAGTCTGAACCAGCC
AAGTAGCGTGCAG
+
ABC8C,:@F:CE8,B-,C,-6-9-C,CE9-CC--C-<-C++,,+;CE<,,CD,CEFC,@E9<FCFCF?9
```

Phred Score: 32
Base Call Accuracy: >> 99.999%

Phred Score: 12
Base Call Accuracy: 94.6%

Phred Score: 10
Base Call Accuracy: 90%

We can use CLI tools for some FASTQ manipulations

```
$ grep -B 1 -A 2 --no-group-separator NNNNNNNNNN Mov10_oe_1.subset.fq > bad_reads.fq
```

Quick review: What do the different arguments for grep do?

But we also need to have tools to help us make sense of our FASTQ data as a whole

- How long do you think it takes to scroll through a FASTQ of 200 million lines?

But we also need to have tools to help us make sense of our FASTQ data as a whole

- How long do you think it takes to scroll through a FASTQ of 200 million lines?

About 69 days!

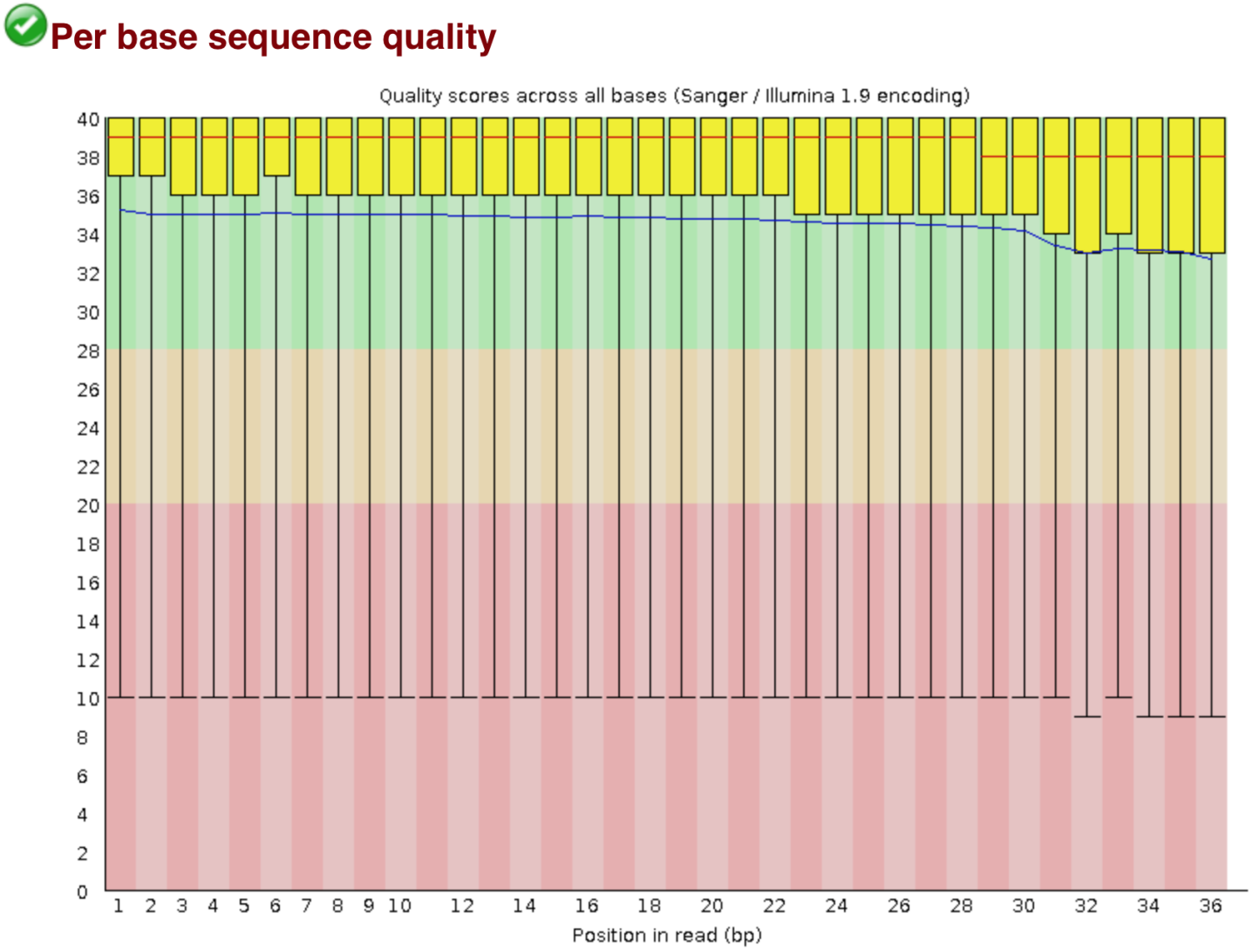
In fact…these files are so unwieldy we keep them **compressed** most of the time or else Biowulf staff start sending you e-mails!

Assessing Data Quality

- One of the first steps of sequence analysis is to see whether the raw data is any good (high confidence)
- Tools like FASTQC summarize overall quality trends in your data set over the millions of reads in your input data
- You use this information to decide whether to trim low-quality sections of the reads, or discard some reads altogether

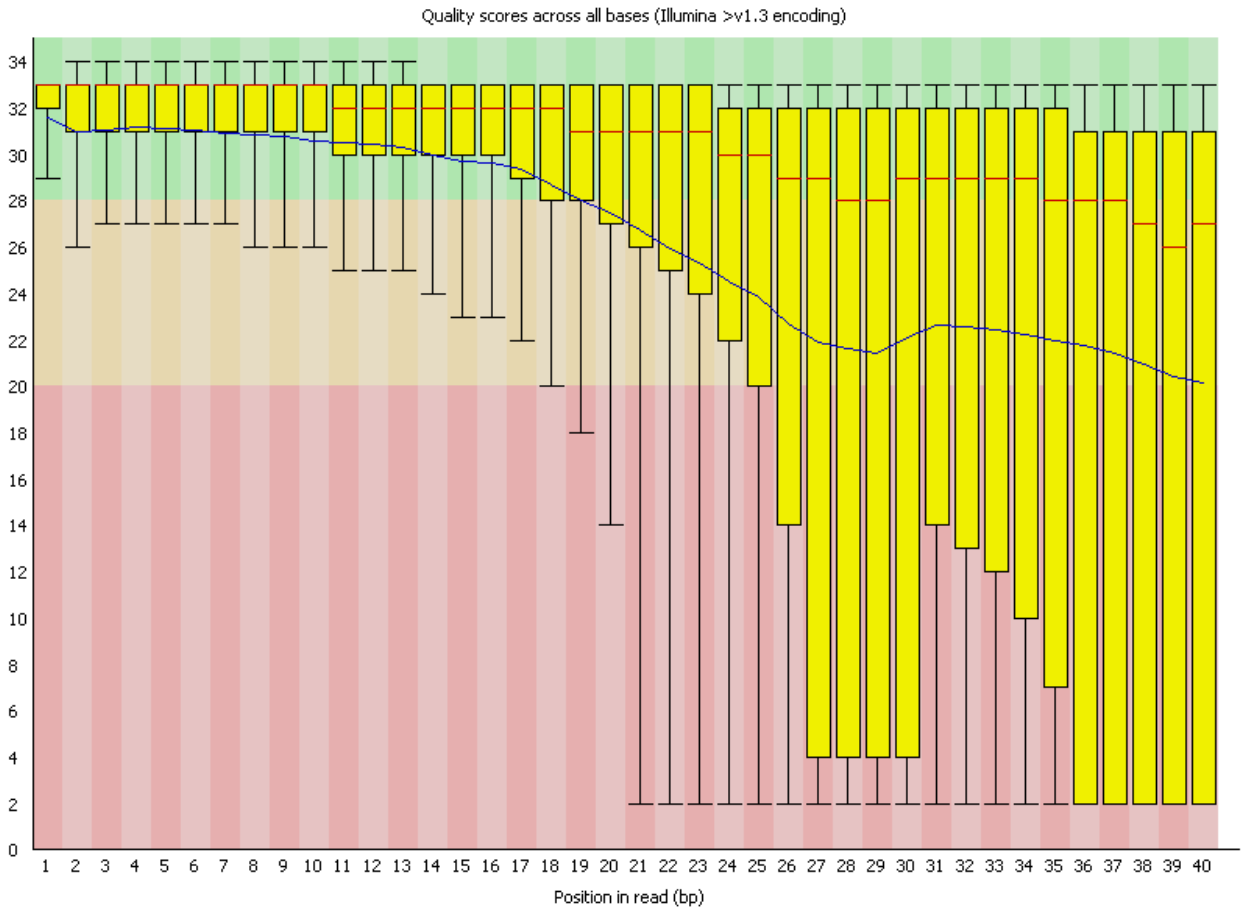
Example: Per-base sequence quality (good)

- Each of the columns is a boxplot of read quality at each basepair in the read, over all reads in your data
- We can see these are 36 basepair reads
- Data already very clean: Phred scores consistently high across length of reads



Per Base Sequence Quality: The Bad (but fixable)

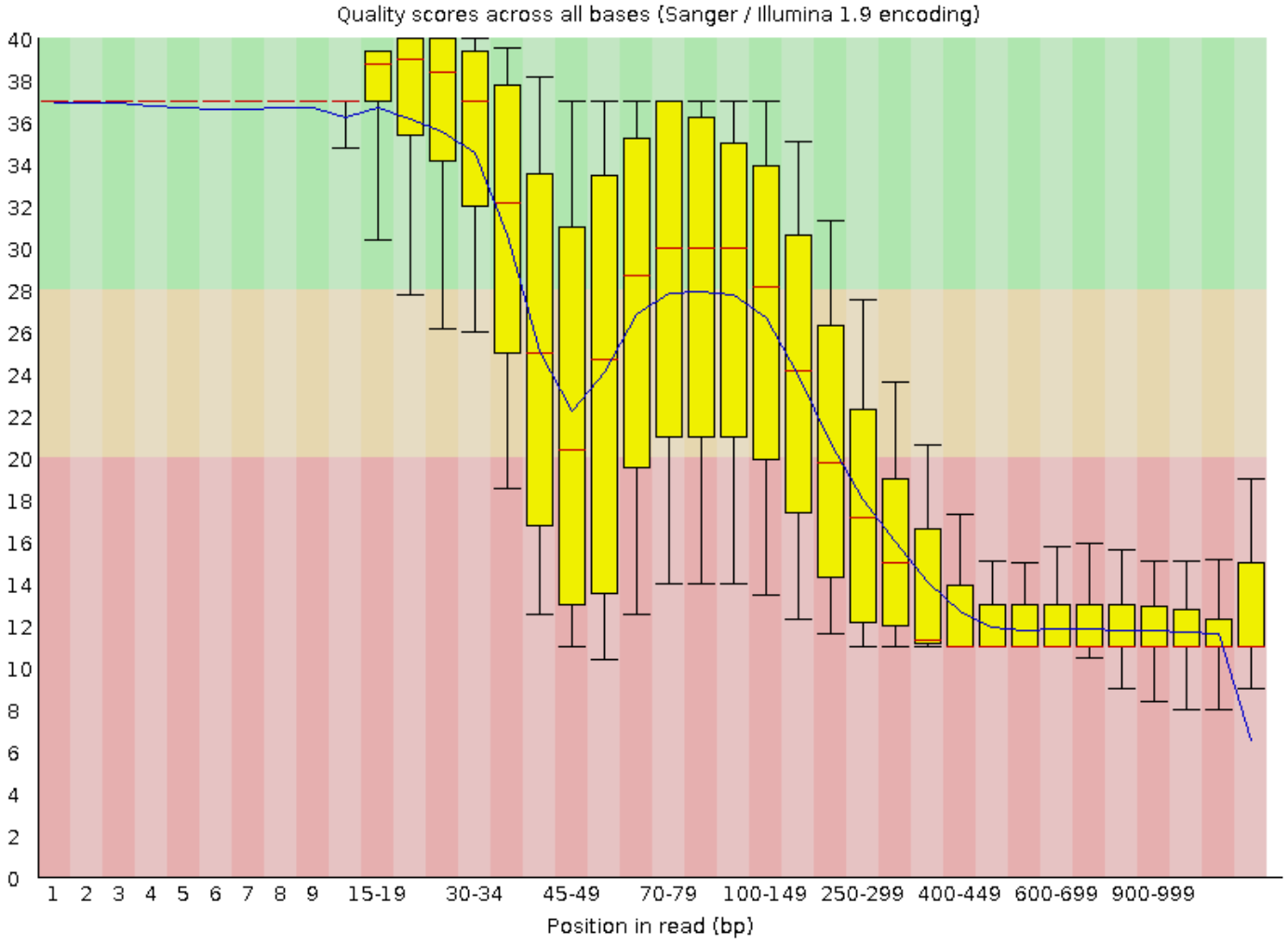
- Sequencing reads often decline in quality along the length of the reads
- As a result, pipelines trim bases off the end or software accounts for this quality drop-off



Per-Base Sequence Quality: The Ugly

More complex patterns might indicate issues with the actual sequencing itself!

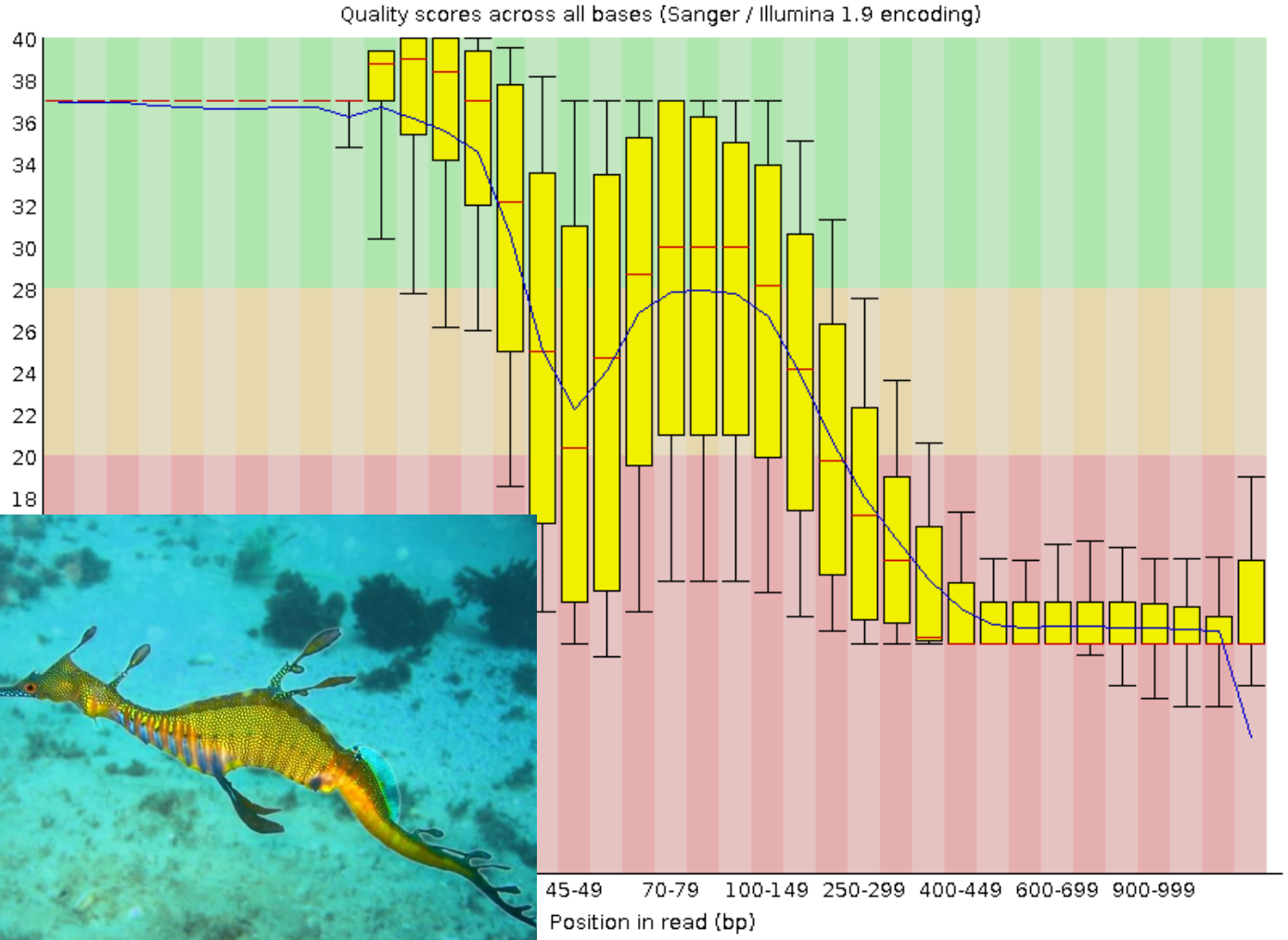
Learn more about some of these issues in Lesson 04 this week.



Per-Base Sequence Quality: The Ugly

More complex patterns might indicate issues with the actual sequencing itself!

Learn more about some of these issues in Lesson 04 this week.



Summary

- Sequencing reads are the basic unit of a sequencing project
- Raw data is often stored in FASTQ files, which contain the sequence and quality information about the sequence
- You can use tools to assess, and then decide how to treat, your input data before any other analyses
- A FASTQ file on its own doesn’t have an inherent order or relationship to a reference – we will map next week!

To run FASTQC we will use some Biowulf features:

- LMOD system for loading software
- Running software interactively
- Running software as a job submitted to Biowulf
- Parallelization

Week 2 Materials

https://nichd-bspc.github.io/intro-rnaseq-hpc/schedule/links-to-lessons.html#week-2