BF528 - Sequence Analysis Fundamentals

Short Read Sequencing Review

- Original molecules are DNA fragments
- Fragments ~300-400 bases long
- Millions to billions of **short (<150nt) reads**
- All reads in dataset are the **same length**
- Each read base has a **quality score**
- Reads may be **single** or **paired end**
- Probably in **fastq** formatted files

Typical Analysis Workflow

1. Assess sequence quality
2. Trim adapters and low-quality sequence
3. Assess trimmed sequence quality
4. Align or quantify reads against a reference
5. Assess alignment quality
6. Analyze alignments

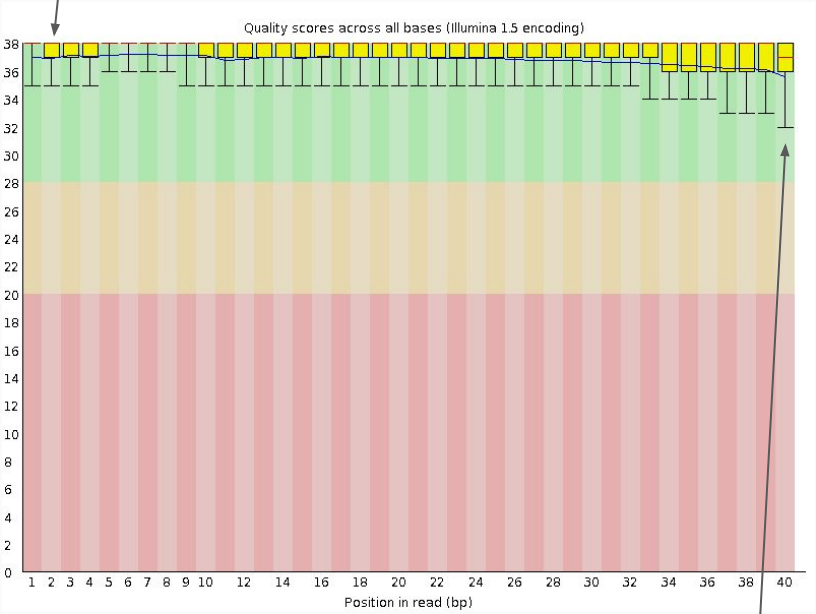
Quality check with FastQC

- FastQC: A quality control tool for high throughput sequence data
- Command line and graphical interfaces
- Produces HTML report
- Several metrics calculated on single set of sequences (e.g. fastq file)
- Useful to identify bad or outlier samples

FastQC: Read Quality

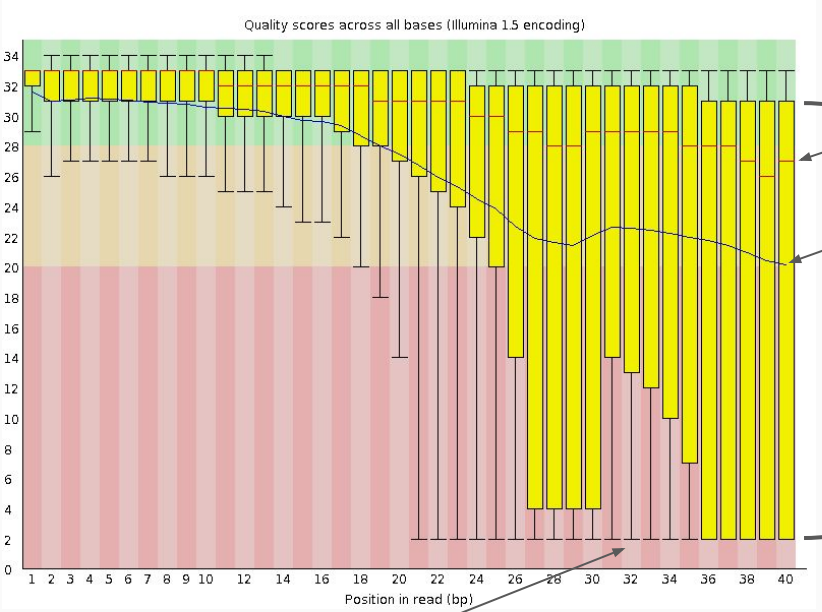
- Untrimmed reads all have same length
- Each position in each read has a Phred score
- Quantify the distribution of Phred scores in each position

Beginning of most reads is of high quality



Quality can degrade towards the end of read

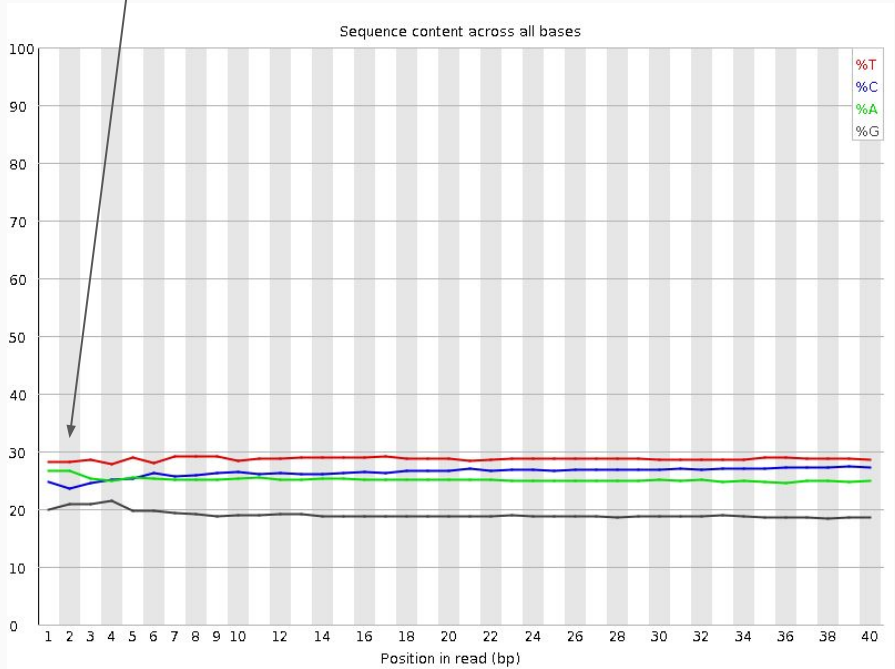
FastQC: Poor Read Quality Example



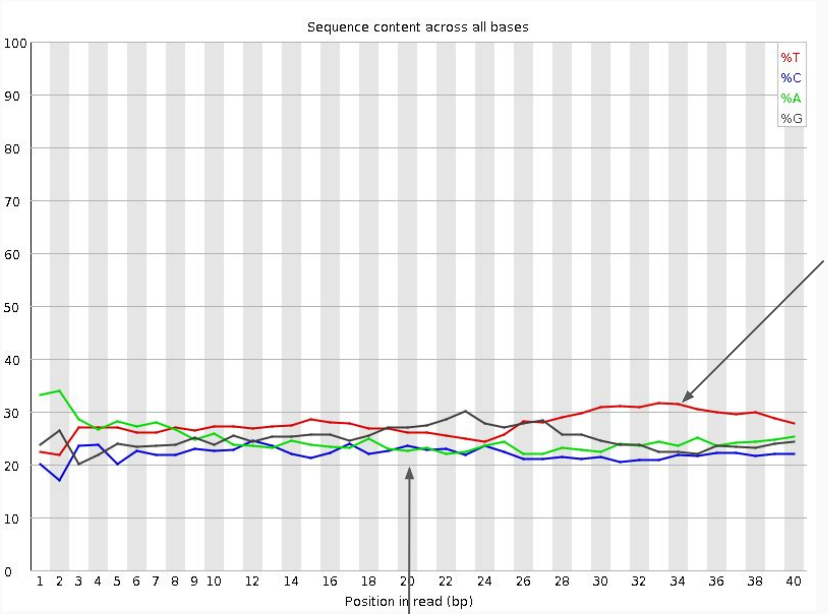
FastQC: Nucleotide distribution

- Each read position may be an A, C, G, or T (or N)
- Each position has a distribution across reads
- Distribution should match originating molecule distribution

Ligation biases cause first 8-10 bases to have non-uniform distribution



FastQC: Non-uniform nucleotide dist.



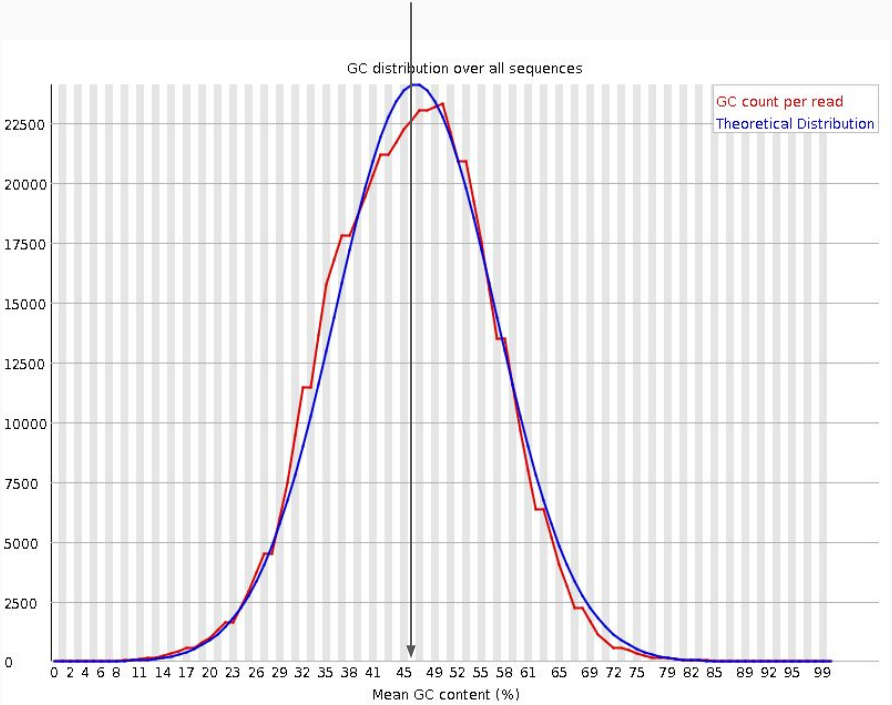
Higher than expected %T on end of reads for unknown reason

Nucleotide distribution non-uniform across read positions

FastQC: GC Content Distribution

- GC content is the fraction of C and G nucleotides
- Each read has a GC content (% GC)
- Most reads should have %GC equal to original molecules
- Should follow normal distribution

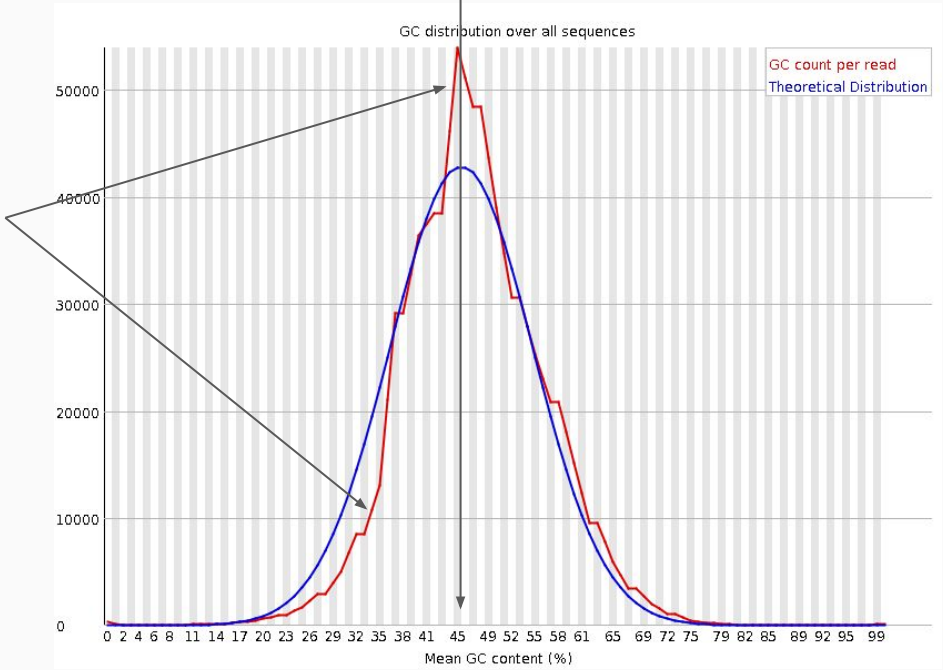
Mean GC content is not always 50%



FastQC: GC Content Mismatch

Mean GC content may be different from expected target molecules

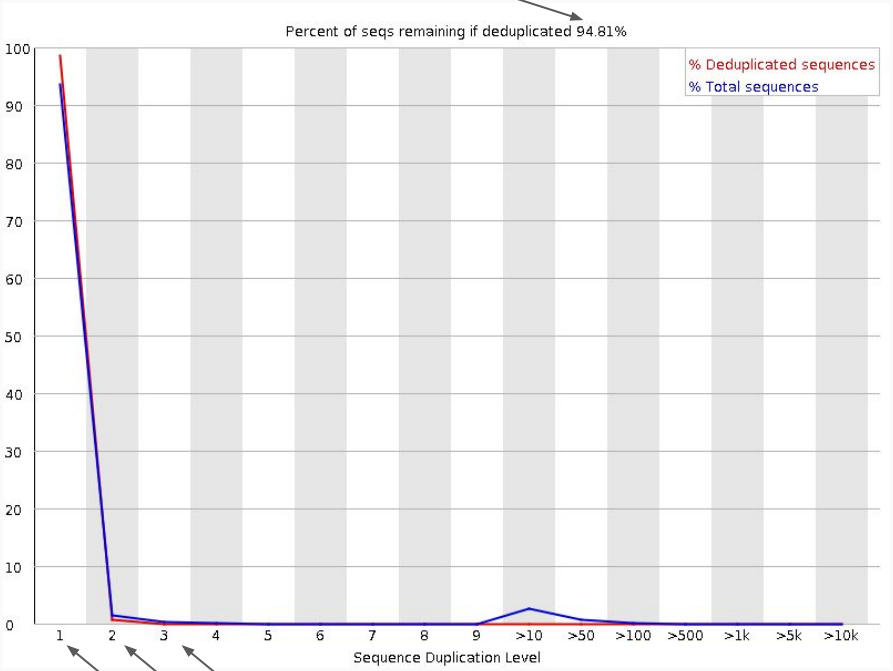
Empirical distribution deviates from expected normal distribution



FastQC: Read Duplication Levels

- PCR amplification may bias certain molecules
- Reads from these events have identical sequence
- For genome sequencing, observing two identical DNA fragments is unlikely (why?)

94.81% of sequences are unique

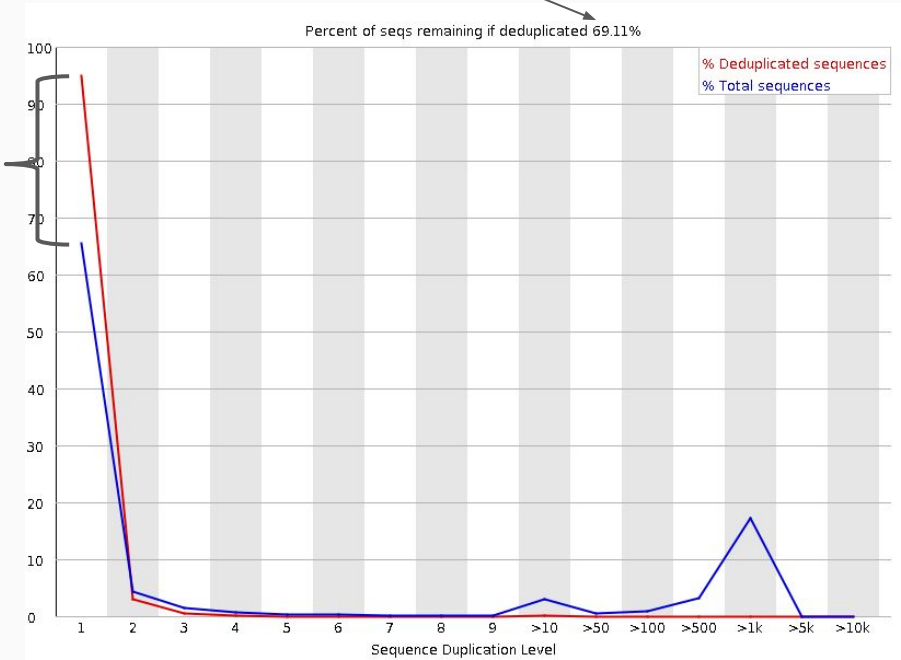


Read sequences that have exactly N copies in dataset

FastQC: Read Duplication Problem

69.11% of read sequences are unique, indicating possible contamination or PCR amp bias

~30% of reads have one or more duplicates



FastQC: Over-represented Sequences

- Over-represented sequences may come from:
  - Contamination
  - Sequencing adapter
  - Low library complexity
  - PCR amp bias
  - Inefficient rRNA depletion
- Some low-complexity libraries, e.g. miRNASeq, will naturally have over-represented sequences

Library prep adapter is sometimes sequenced

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

Sequence unrecognized by FastQC
(it's a barcoded Single End PCR Primer 1)

Running FastQC on SCC

```
$ module load fastqc
$ fastqc --help
$ fastqc SRR1919605_1.fastq.gz
```

Run as batch job!

Trimming

- Reads may have adapters and low quality 3’
- Want to **trim** reads to remove these effects
- Adapters/quality can be trimmed separately
- Many programs available, most common:
  - trimmomatic
  - cutadapt
- Trimmed reads are not same length

Trimming Example: Adapter

Untrimmed

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAAATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Trimmed

```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAA
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Trimming Example: Quality

```
Phred ASCII encoding:
Character: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
           |                          |    |        |
Score:     0                          28   31       40
```

Untrimmed

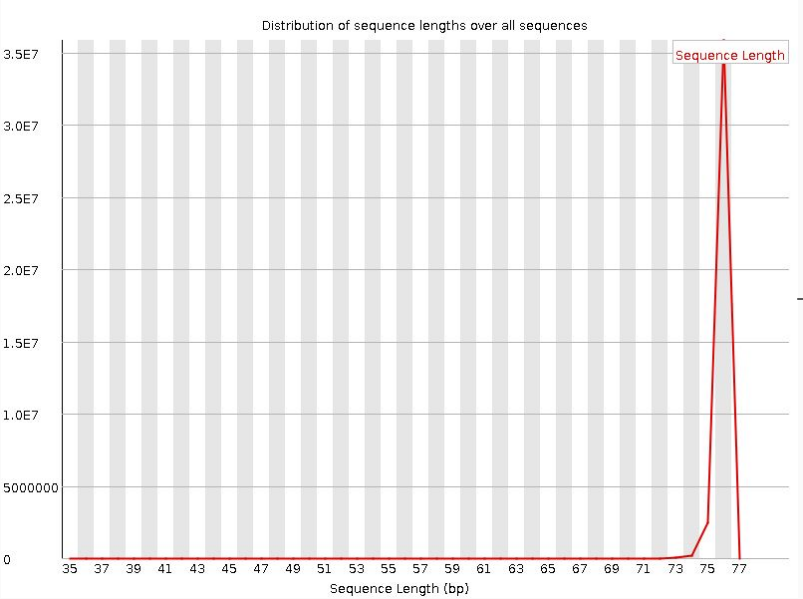
```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAAATGATACGGCGACCANCGAGATCTANANTCTTTNNNTNN
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFEC@@B<;9529.910,#(,-50.&%0#2#*('&%###$##
```

Trimmed

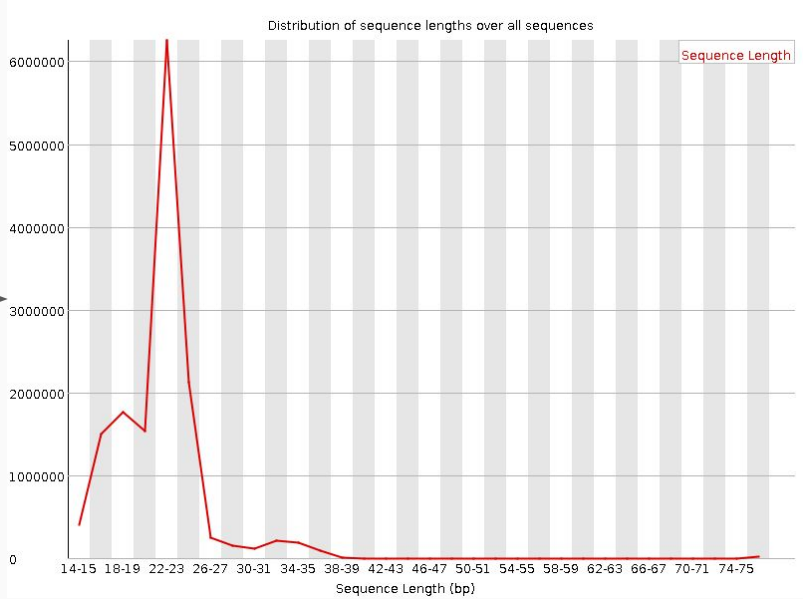
```
@SRR1997412.1 1 length=125
NTTGTAGCTGAGGAAACTGAGGCTCAGGAGGACAAGTGGCCTGCCAAAAATG
+SRR1997412.1 1 length=125
#<<BBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFEC@@
```

FastQC: Length Distribution

Untrimmed

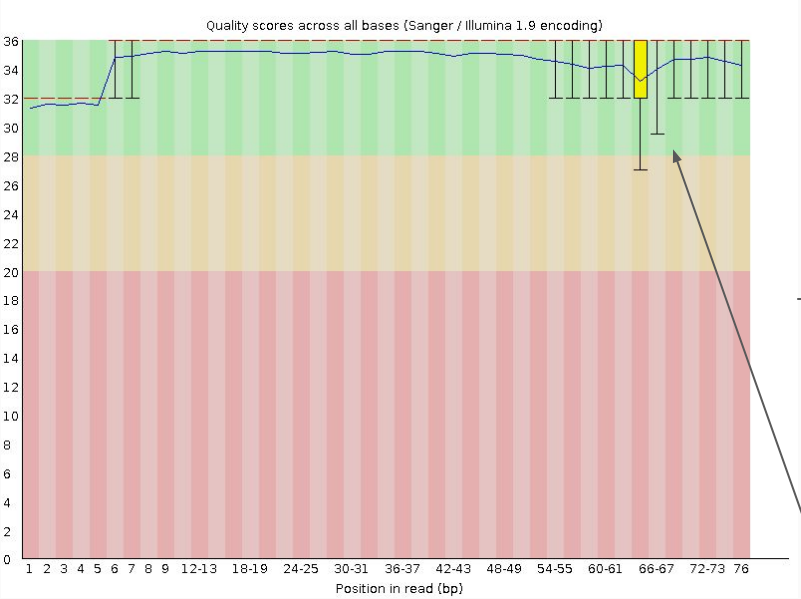


Trimmed

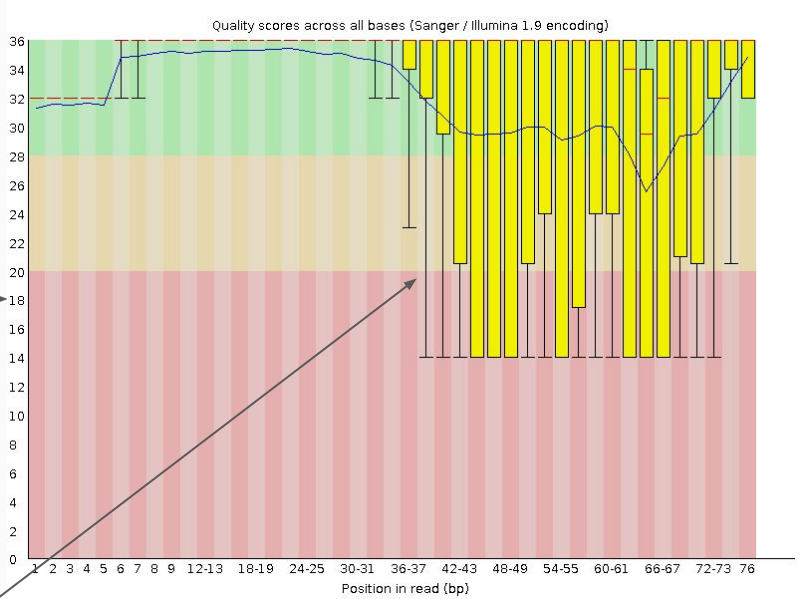


FastQC: Sequence Quality

Untrimmed



Trimmed



Why do we now see poor 3’ quality?

Alignment of short reads

- Trimmed reads used for all downstream analysis
- Alignment matches reads to some reference
- Usually one of two approaches:
  - Align - high resolution, explicit alignment, slow
  - Quasi-align - ‘good enough’ resolution, heuristic alignment, fast

Alignment tools

- Many tools to align reads
- General aligners
  - bwa
  - bowtie and bowtie2
  - SNAP → forces reads to align close to each other
- Purpose specific (e.g. RNA-Seq)
  - STAR
  - tophat

Alignment reference

- The reference is a haploid representation of the consensus of multiple individuals.
- Human genome:
  - GRCh37 (aka hg19)
  - GRCh38 (hg38)
- Mouse:
  - GRCm38 (mm10)
  - NCBI37 (mm9)
- It is in fasta format (.fa or .fasta)

Alignment Example: bwa

- Each aligner requires the reference to be prepared (indexed) in a certain way

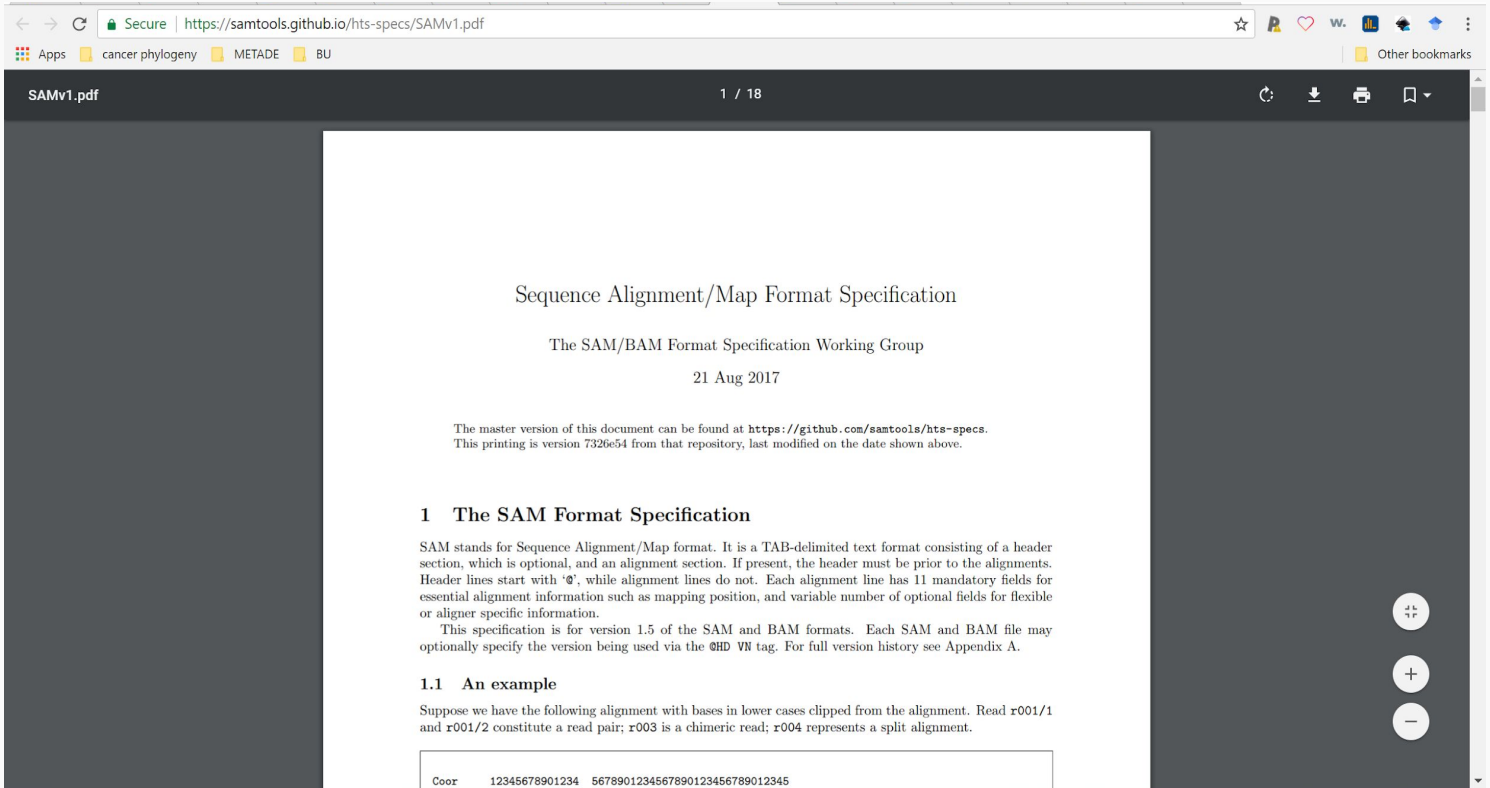
```
$ module load bwa
$ bwa index chr22.fa
```

- Once reference is indexed, align reads

```
$ bwa mem chr22.fa SRR1919605_1.fq.gz > SRR1919605.sam
```

SAM format

https://samtools.github.io/hts-specs/SAMv1.pdf



SAM format

- @ header line
- Each line has at least 11 fields
- 1 line per each read

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	`[0,2^16-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,2^31-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,2^8-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	`[0,2^31-1]`	Position of the mate/next read
9	TLEN	Int	`[-2^31+1,2^31-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SAM Format: Header

header lines start with @

```
@HD    VN:1.5  GO:none SO:coordinate
@SQ    SN:1    LN:248956422
@SQ    SN:10   LN:133797422
@SQ    SN:11   LN:135086622
@SQ    SN:12   LN:133275309
@SQ    SN:13   LN:114364328
@SQ    SN:14   LN:107043718
...
@RG    ID:50   LB:SRR1514950   SM:CHM1
...
@PG    ID:MarkDuplicates      VN:2.8.0-SNAPSHOT      CL:picard.sam.markduplicates.MarkDuplicates
@PG    ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa.real mem -t 16 GRCh38.fa SRR1919605_1.tar.gz
```

version

The reference names (i.e. chromosomes) and their lengths

read groups

Programs you ran

SAM Format: Aligned Read

- Each line has at least 11 fields
- 1 line per each alignment

```
SRR1514952.11241320     147     1       10000   27      41S60M  =       10034   -26
GAACCCTAGCCCTACCCCAACCCCGAACCCTACCCCGAACCATAACCCTAACCCTAACCCTAACCCTATCCCTAACCCTAGCCCTA
#############################################################################A2A+;H;F
XA:Z:X,+156030660,76M25S,5;Y,+57217180,76M25S,5;22,+50808410,60M41S,2;6,-147845,18S23M1D38M
1D22M,7; MC:Z:75M1I25M   MD:Z:27A11A20   NM:i:2  MQ:i:27 AS:i:50 XS:i:51 RG:Z:52
PG:Z:MarkDuplicates-24B543D9
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

samtools

- **BAM** is a compressed sam.
- **CRAM** is a better compressed SAM.
- You can read all with **samtools**

```
module load samtools
samtools view SRR1919605.sam
# unmapped reads
samtools view -f 4 SRR1919605.sam
# pair unmapped
samtools view -f 8 SRR1919605.sam
# only mapped ones
samtools view -F 4 SRR1919605.sam
# quality at least 30
samtools view -q 30 SRR1919605.sam
```

Important Alignment Properties

- **Mapping quality** - how good alignment is
- **Alignment rate** - % of reads that align
- **Multimapped reads** - align to multiple loci
- **Mate distance** - distance between paired read alignments (paired end only)
- **Coverage**, **depth**

Quasi-alignment tools

- Heuristic approach to alignment
  - i.e. fast, but not as sensitive as full alignment
- Typically against a reference of relatively short sequences (e.g. transcriptome)
- Include GC content, multimap adjustment
- Used for **quantification estimation**
  - i.e.what is the estimated expression of a gene?

Identifying Outlier Samples

- Flag sample if:
  - Deviate from expected (e.g. GC content)
  - Deviate from most other samples
  - Contain suspect sequences (e.g. contamination)
  - Aligns poorly
- MultiQC - tool for combining output from many tools run on many samples