NGS Data Formats

Austin Herbert
1_25_2023

CLEMSON® UNIVERSITY
CENTER FOR
HUMAN GENETICS

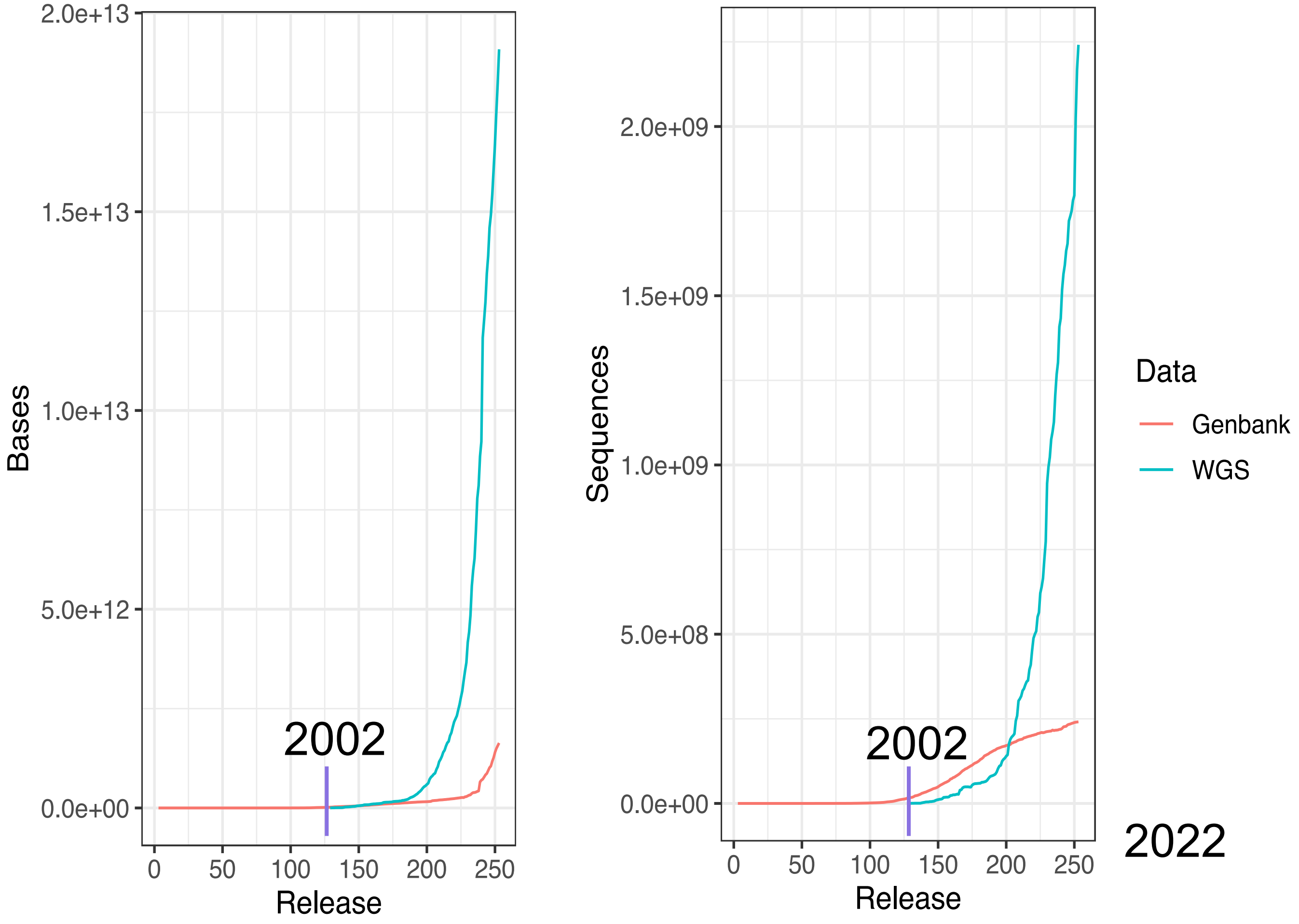
Learning objectives

- **Identify & describe different NGS file formats**
  - Point of origin
  - Contents

- **Analyze NGS data for sequencing & mapping quality**
  - Understand QC metrics
  - Conduct alignments

Relevance

- @ 2002
  - ~19.70 **B** as in **B**illion bases sequenced
- @ 2022
  - 20.70 **T** as in **T**rillion bases sequenced



Data from: https://www.ncbi.nlm.nih.gov/genbank/statistics/

Relevance

- @ 2002
  - ~19.70 **B** as in **B**illion bases sequenced
- @ 2022
  - 20.70 **T** as in **T**rillion bases sequenced
- 20,000 times the amount of sequencing data in 20 years

Here is the plan: we sequence another 20 TRILLION bases

Don't you think that's kind of a low number?

CLEMSON® UNIVERSITY
CENTER FOR
HUMAN GENETICS

Data from: https://www.ncbi.nlm.nih.gov/genbank/statistics/

Sequence file formats

- Fasta
- Nanopore sequencing – hdf5/fast5 format
- .fastq, .fq, .fastq.gz
- **S**equence **A**lignment **M**ap (.sam)
  - .sam file index == .sai
- **B**inary **A**lignment **M**ap
  - .bam file index == .bai

Fasta

- Standard text file
- Stores strings of nucleotide and amino acid sequences
- Used as reference sequences
- Denote new sequence with ‘>’

```
>[sequence name line]
ACTGCAGTCAGTGACTNNTCGA
```

Illumina versus Nanopore

- Nanopore sequencing versus conventional illumina technology
- What chemistry does illumina use?
- How does nanopore sequencing work?

Illumina versus Nanopore

illumina

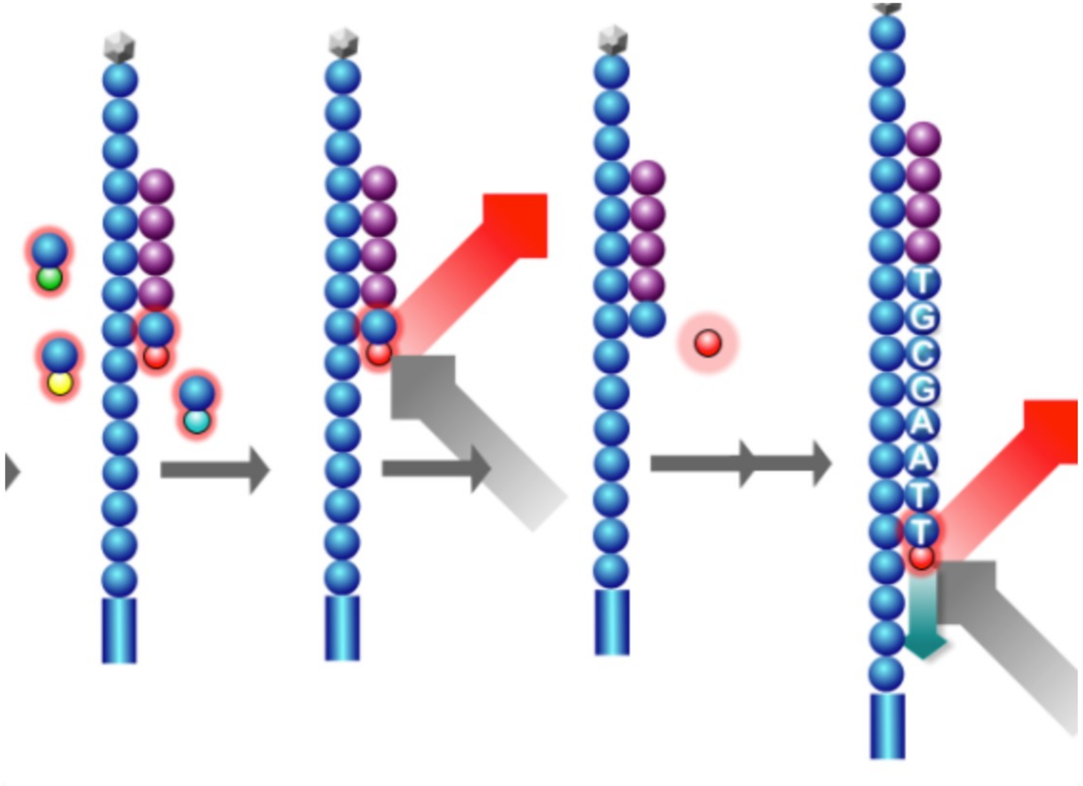
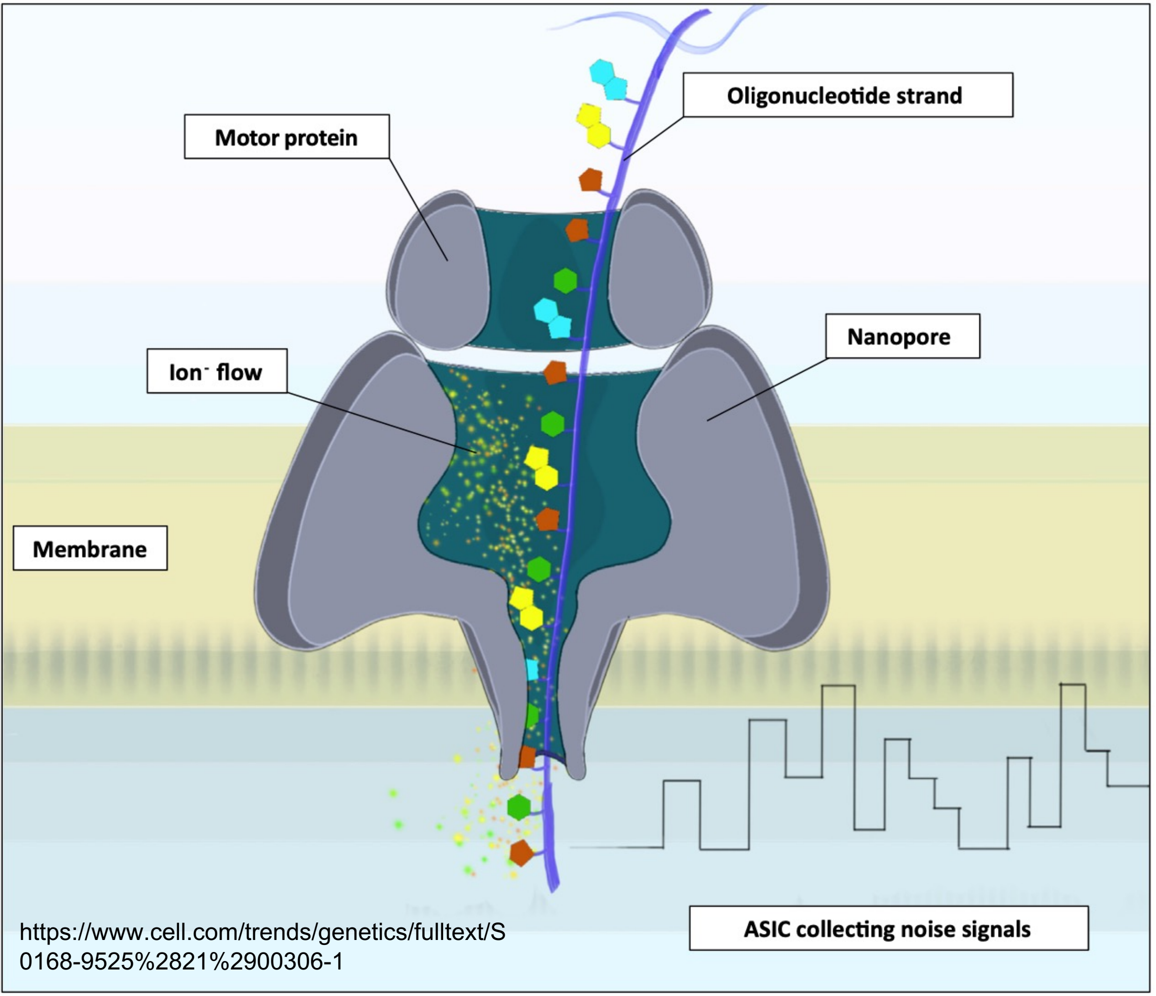


Figure 4: Sequencing-by-Synthesis

https://pediaa.com/how-does-illumina-sequencing-work/

Nanopore

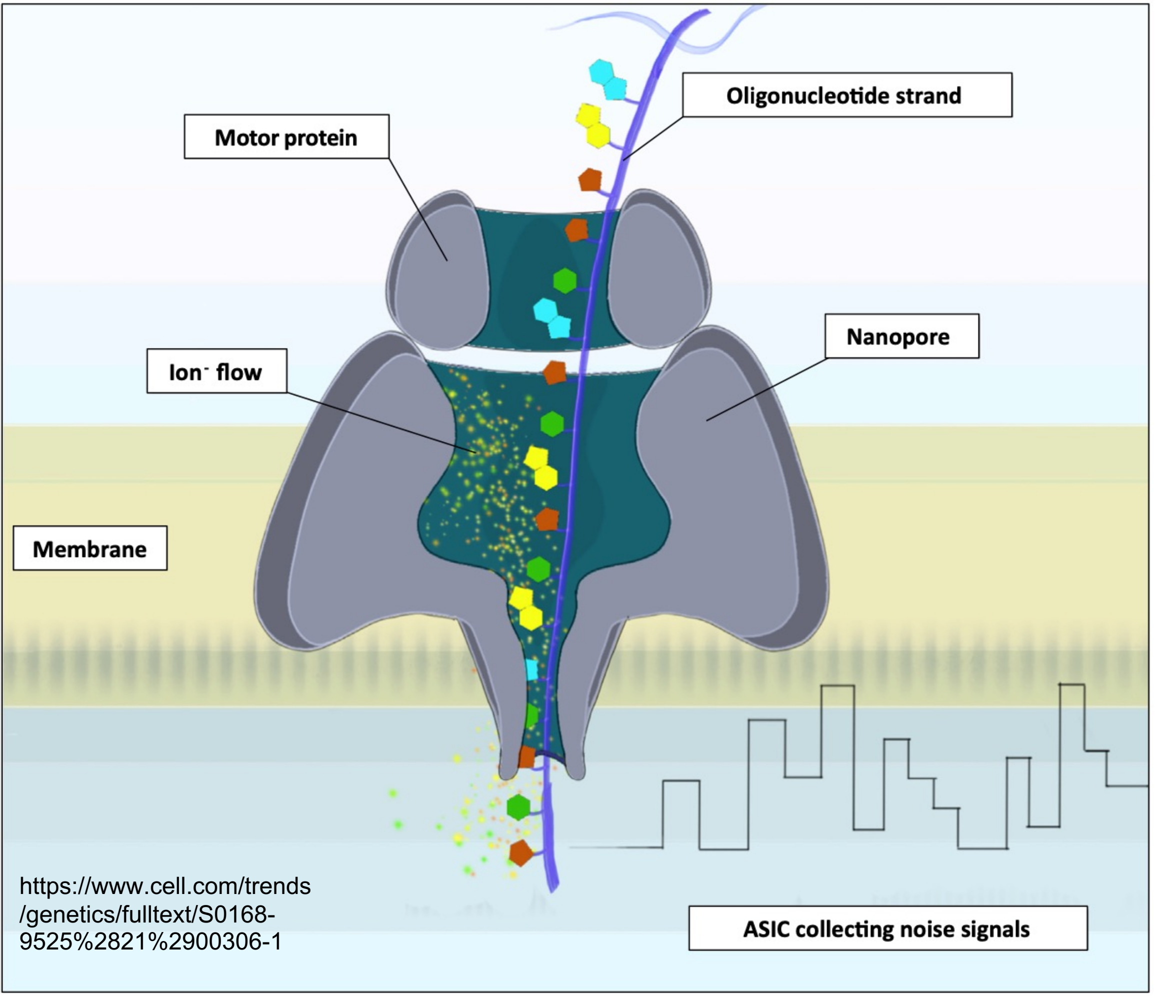


https://www.cell.com/trends/genetics/fulltext/S0168-9525%2821%2900306-1

Nanopore sequencing - HD5/fast5 file format

- Binary, compressed, hierarchical data scheme

- Holds
  - Fastq reads, quality scores
  - Raw nanopore signal (squiggles)

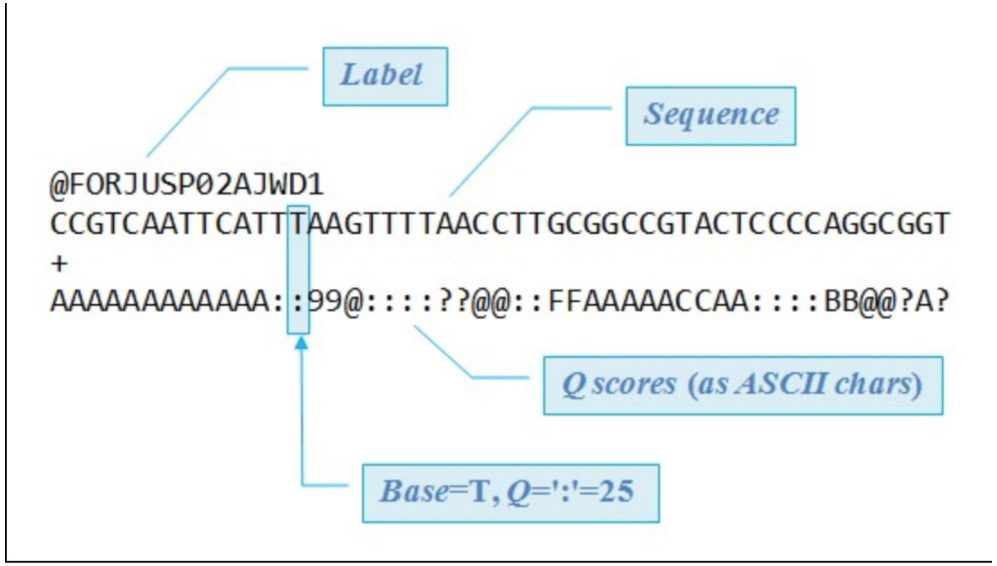


https://www.cell.com/trends/genetics/fulltext/S0168-9525%2821%2900306-1

Fastq - description

- Fastq files generated directly by the sequencer
  - Paired end sequencing generated two files
    - _R1.fastq.gz
    - _R2.fastq.gz
- Contains raw sequencing data & QScores

- Unzip and view
  - Gunzip {filename}_R1.fastq.gz
  - More {filename}_R1.fastq



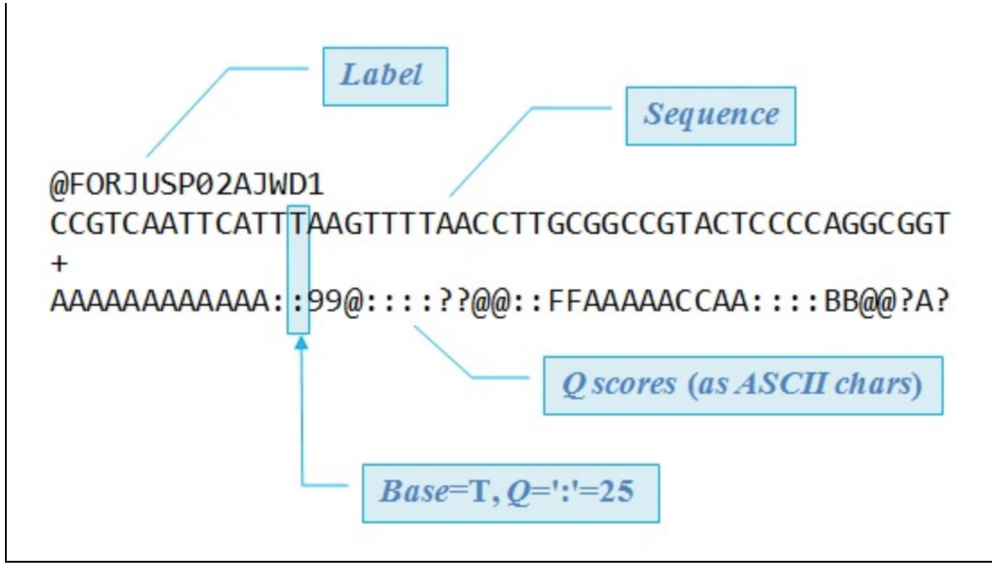
```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Photo from: https://www.drive5.com/usearch/manual/fastq_files.html

Fastq – quality control and metrics

- Fastq quality and statistics

- Purpose –
  - Ensure data quality meets usable standards
  - Make adjustments to library prep



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

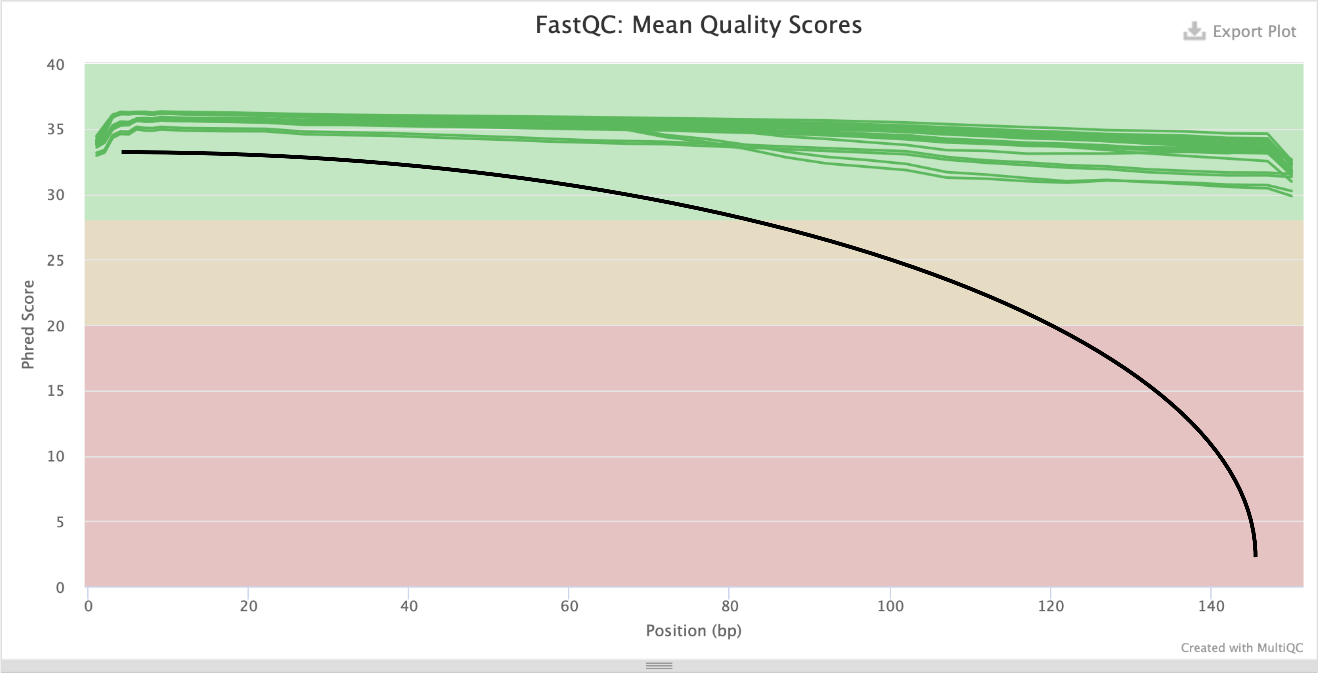
Photo from: https://www.drive5.com/usearch/manual/fastq_files.html

Fastq – Base quality scores

- Base quality scores

- Poor quality reads have high drop off towards ends
- High quality = low probability of error

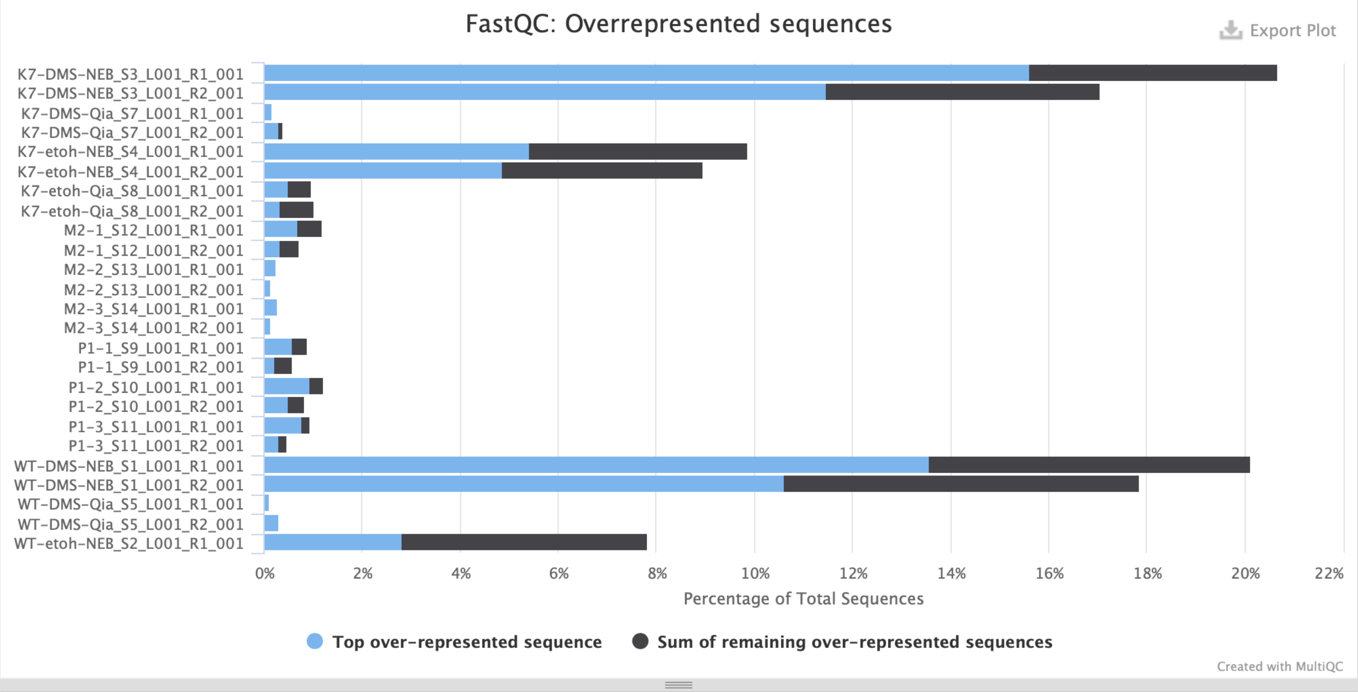
- Issue with actual sequencer
- Library diversity



CLEMSON® UNIVERSITY
CENTER FOR HUMAN GENETICS

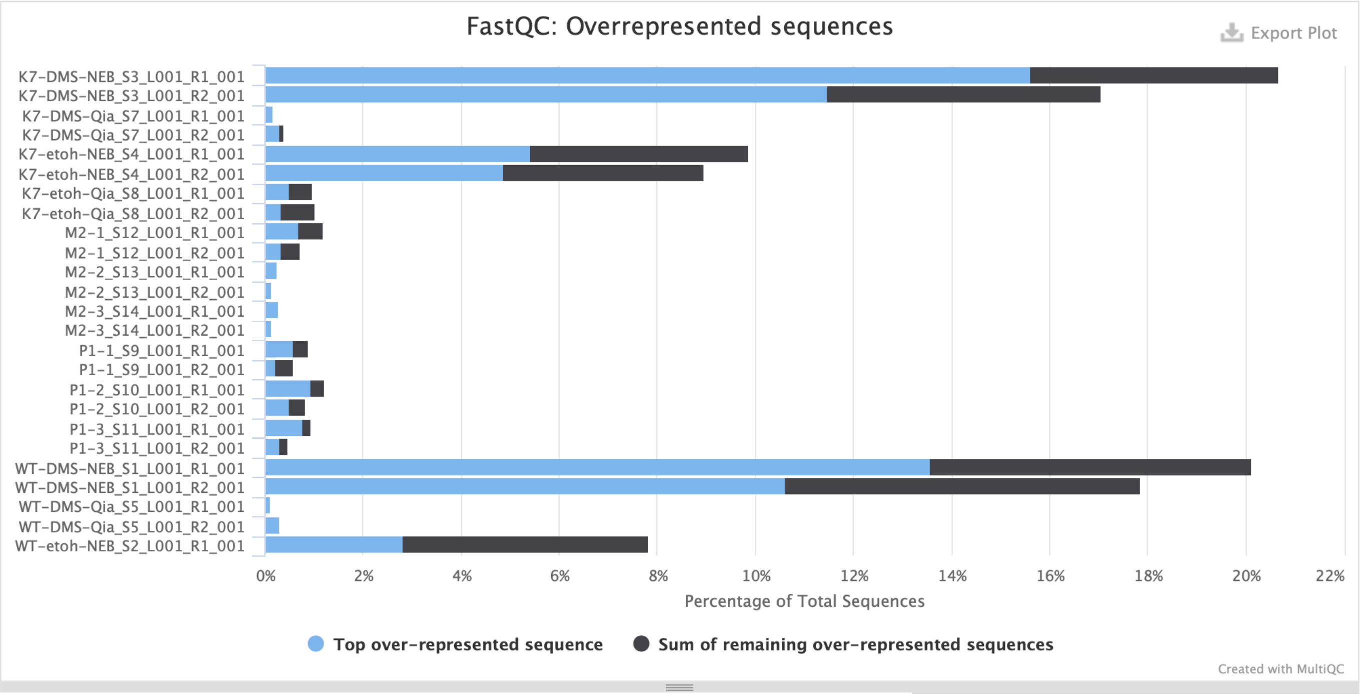
Fastq – overrepresented sequences

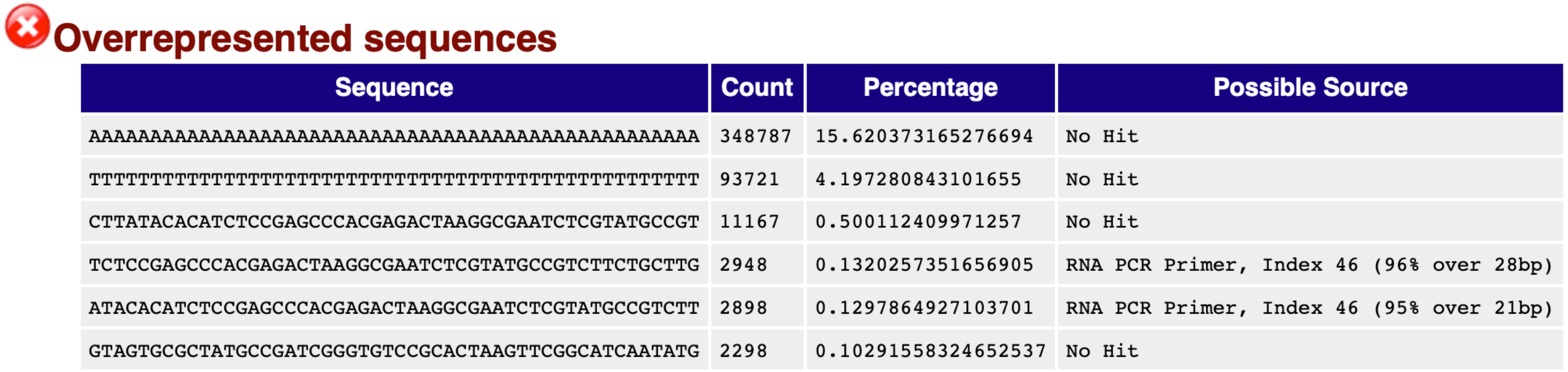
- Duplications, contamination, rRNA or adaptors/primers
- Overrepresented if > .1% of reads
- RNA sequencing – rRNA makes up about 80% of RNA
- Methods
  - polyA tail enrichment
  - rRNA depletion



Fastq – overrepresented sequences

- Duplications, contamination rRNA or adaptors/primers
- Overrepresented if > .1% of reads
- RNA sequencing – rRNA makes up about 80% of RNA



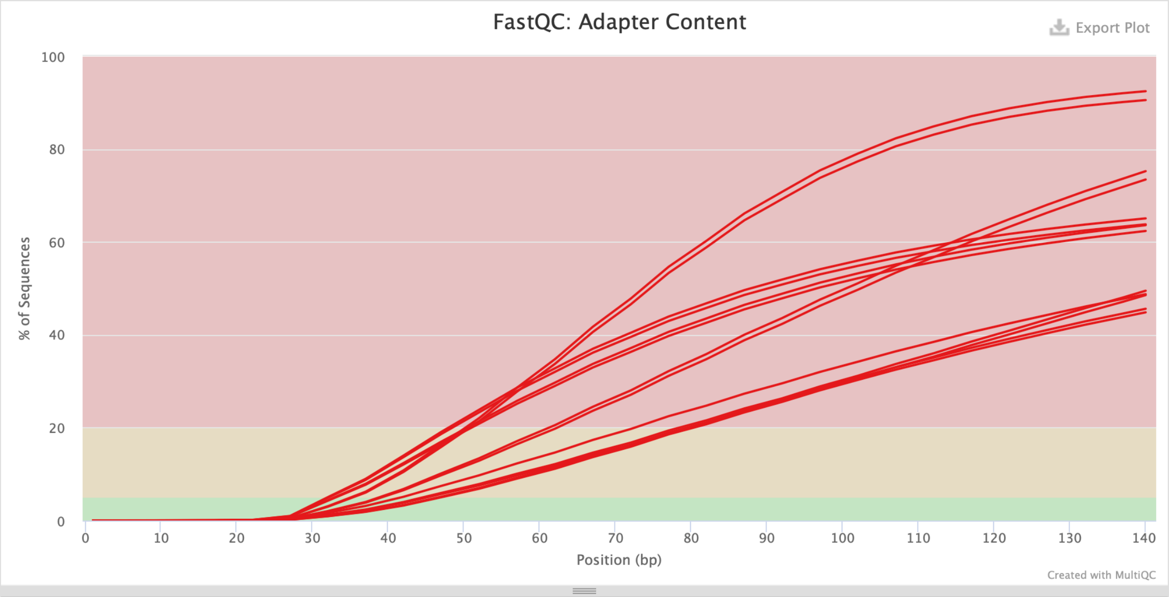


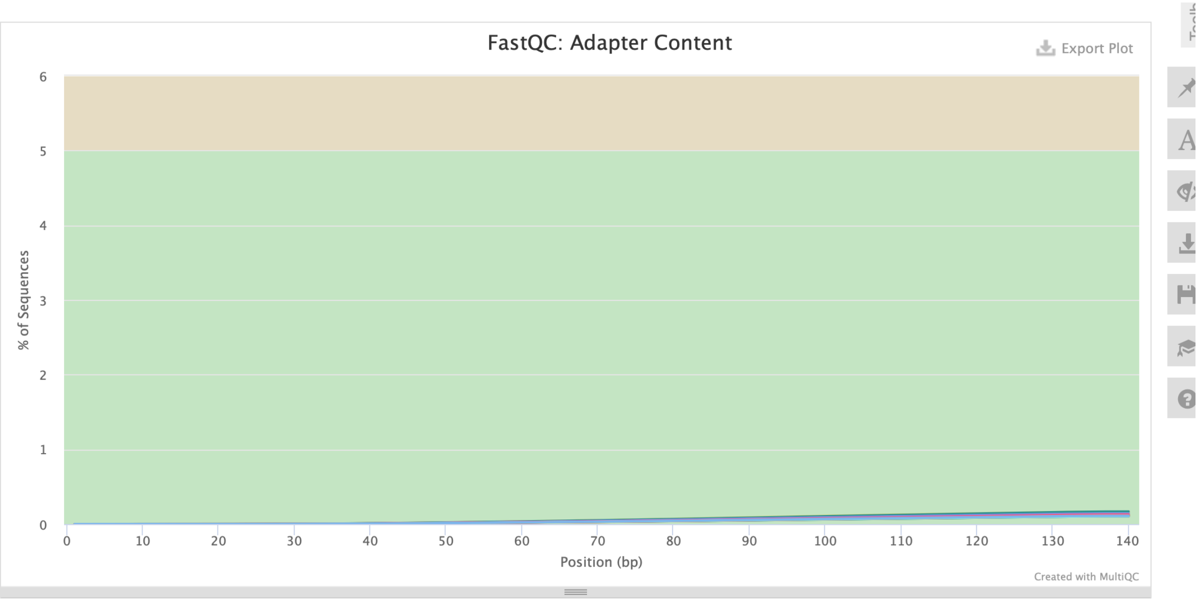
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	348787	15.620373165276694	No Hit
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	93721	4.197280843101655	No Hit
CTTATACACATCTCCGAGCCCACGAGACTAAGGCGAATCTCGTATGCCGT	11167	0.500112409971257	No Hit
TCTCCGAGCCCACGAGACTAAGGCGAATCTCGTATGCCGTCTTCTGCTTG	2948	0.1320257351656905	RNA PCR Primer, Index 46 (96% over 28bp)
ATACACATCTCCGAGCCCACGAGACTAAGGCGAATCTCGTATGCCGTCTT	2898	0.1297864927103701	RNA PCR Primer, Index 46 (95% over 21bp)
GTAGTGCGCTATGCCGATCGGGTGTCCGCACTAAGTTCGGCATCAATATG	2298	0.10291558324652537	No Hit

Fastq - Adapter content

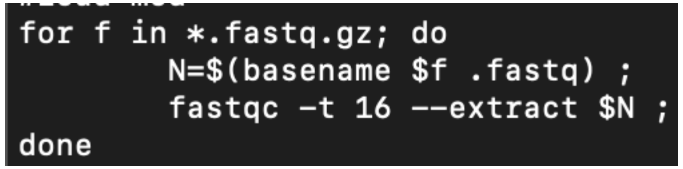
- Degraded/low quality RNA input
- Poor RT





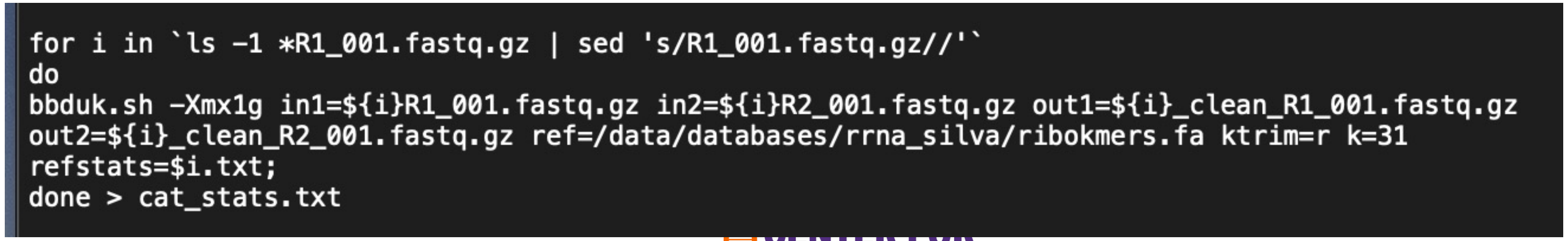
Fastq – quality metrics

- Fastqc loop



```
for f in *.fastq.gz; do
        N=$(basename $f .fastq) ;
        fastqc -t 16 --extract $N ;
done
```

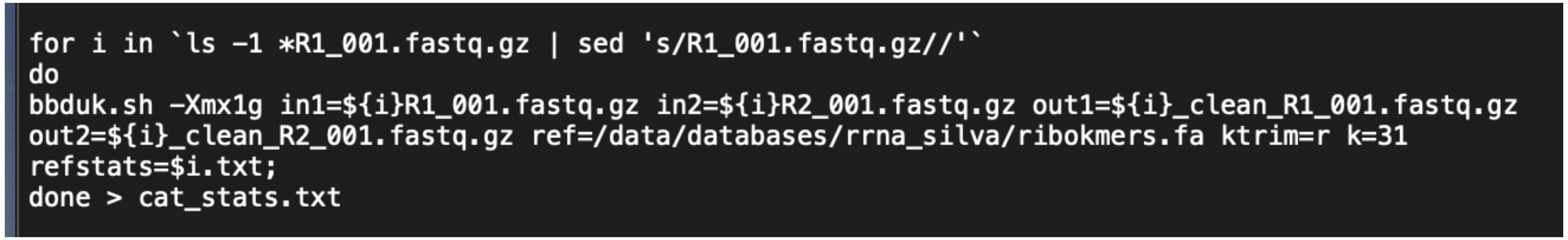
- Generates .html analysis file
- Html files to one folder
- Run: multiqc .



```
for i in `ls -1 *R1_001.fastq.gz | sed 's/R1_001.fastq.gz//'`
do
bbduk.sh -Xmx1g in1=${i}R1_001.fastq.gz in2=${i}R2_001.fastq.gz out1=${i}_clean_R1_001.fastq.gz
out2=${i}_clean_R2_001.fastq.gz ref=/data/databases/rrna_silva/ribokmers.fa ktrim=r k=31
refstats=$i.txt;
done > cat_stats.txt
```

Fastq – quality metrics

- Bbduk.sh
  - Read trimming
  - Accurate rRNA QC metrics



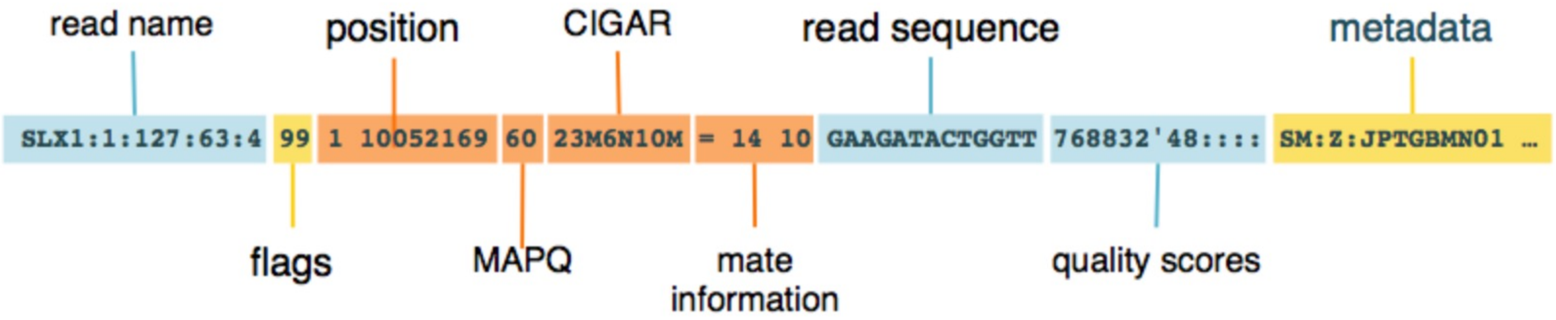
```
for i in `ls -1 *R1_001.fastq.gz | sed 's/R1_001.fastq.gz//'`
do
bbduk.sh -Xmx1g in1=${i}R1_001.fastq.gz in2=${i}R2_001.fastq.gz out1=${i}_clean_R1_001.fastq.gz
out2=${i}_clean_R2_001.fastq.gz ref=/data/databases/rrna_silva/ribokmers.fa ktrim=r k=31
refstats=$i.txt;
done > cat_stats.txt
```

- Cats rRNA stats to single text file

SAM – sequence alignment map

- Standard text file
- Generated by alignment/mapping reads to reference sequence
- Stores alignment information
  - Alignment coordinates, mapping quality,
- Hisat2, STAR, minimap2

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)

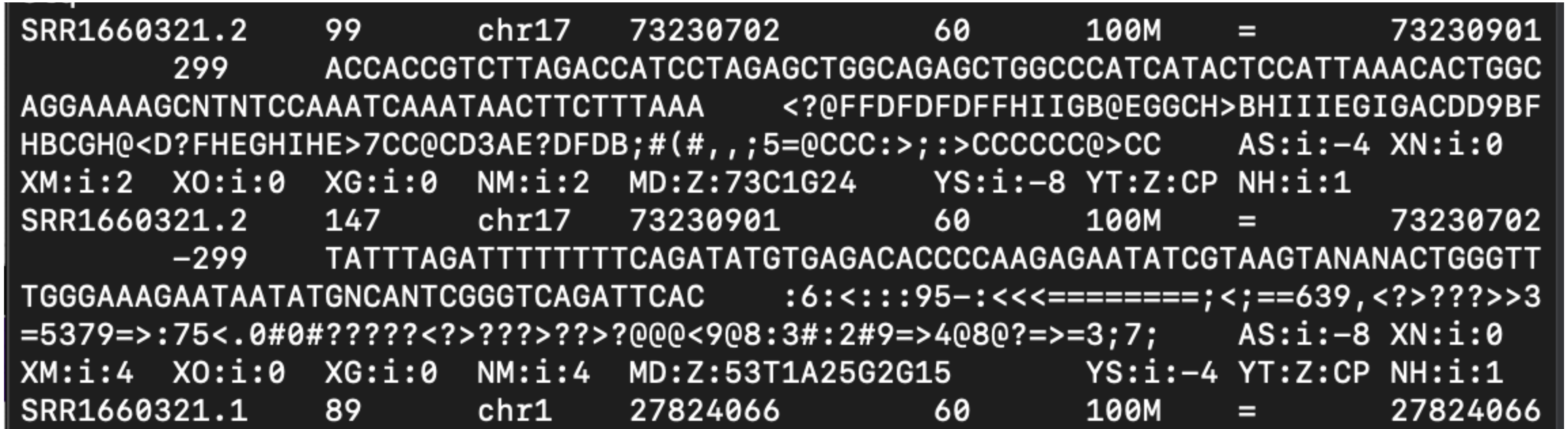


```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …
```

Photo from: https://sites.google.com/a/broadinstitute.org/legacy-gatk-forum-discussions/dictionary/11014-SAM-BAM-CRAM-Mapped-sequence-data-formats

SAM – sequence alignment map

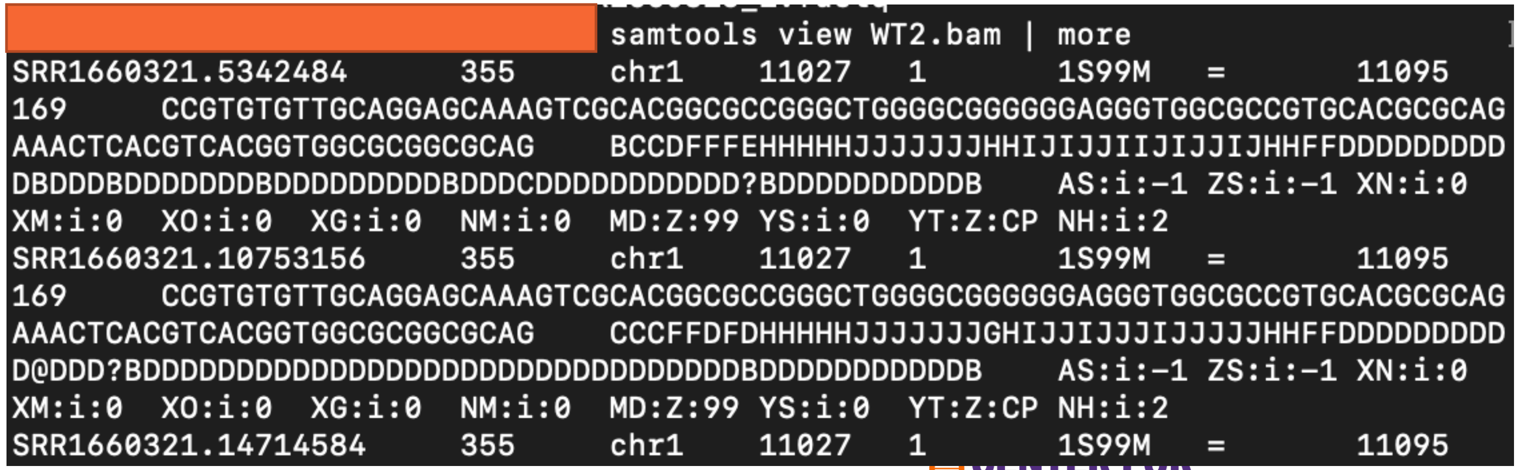
- Standard text file
- Generated by alignment/mapping reads to reference sequence
- Stores alignment information
  - Alignment coordinates, mapping quality,
- Hisat2, STAR, minimap2



```
SRR1660321.2    99      chr17   73230702        60      100M    =       73230901
        299     ACCACCGTCTTAGACCATCCTAGAGCTGGCAGAGCTGGCCCATCATACTCCATTAAACACTGGC
AGGAAAAGCNTNTCCAAATCAAATAACTTCTTTAAA    <?@FFDFDFDFFHIIGB@EGGCH>BHIIIEGIGACDD9BF
HBCGH@<D?FHEGHIHE>7CC@CD3AE?DFDB;#(#,,;5=@CCC:>;:>CCCCCC@>CC    AS:i:-4 XN:i:0
XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:73C1G24    YS:i:-8 YT:Z:CP NH:i:1
SRR1660321.2    147     chr17   73230901        60      100M    =       73230702
        -299    TATTTAGATTTTTTTTCAGATATGTGAGACACCCCAAGAGAATATCGTAAGTANANACTGGGTT
TGGGAAAGAATAATATGNCANTCGGGTCAGATTCAC    :6:<:::95-:<<<========;<;==639,<?>???>>3
=5379=>:75<.0#0#?????<?>???>??>?@@@<9@8:3#:2#9=>4@8@?=>=3;7;    AS:i:-8 XN:i:0
XM:i:4  XO:i:0  XG:i:0  NM:i:4  MD:Z:53T1A25G2G15       YS:i:-4 YT:Z:CP NH:i:1
SRR1660321.1    89      chr1    27824066        60      100M    =       27824066
```

BAM – binary alignment map

- Compressed SAM file into binary
- Decreases size of alignments, frees up space
- Usually work out of bam files
  - Can gzip fq files and delete intermediate sam files



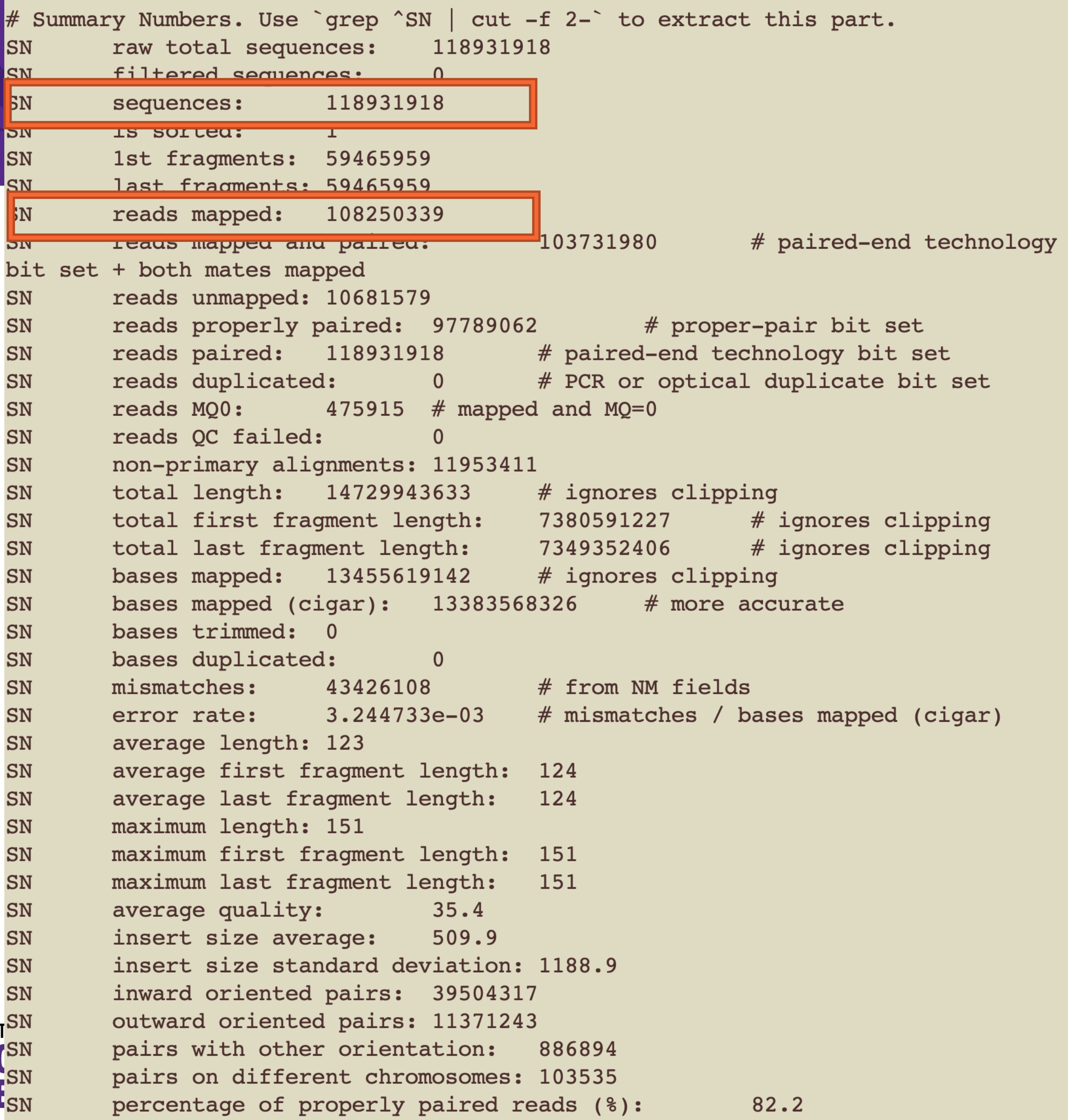
```
[redacted] samtools view WT2.bam | more
SRR1660321.5342484      355     chr1    11027   1       1S99M   =       11095
169     CCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGGGCTGGGGCGGGGGAGGGTGGCGCCGTGCACGCGCAG
AAACTCACGTCACGGTGGCGCGGCGCAG    BCCDFFFEHHHHHJJJJJJHHIJIJJIIJIJJIJHHFFDDDDDDDDD
DBDDDBDDDDDDDBDDDDDDDDDBDDDCDDDDDDDDDDD?BDDDDDDDDDDB    AS:i:-1 ZS:i:-1 XN:i:0
XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:99 YS:i:0  YT:Z:CP NH:i:2
SRR1660321.10753156     355     chr1    11027   1       1S99M   =       11095
169     CCGTGTGTTGCAGGAGCAAAGTCGCACGGCGCCGGGCTGGGGCGGGGGAGGGTGGCGCCGTGCACGCGCAG
AAACTCACGTCACGGTGGCGCGGCGCAG    CCCFFDFDHHHHHJJJJJJGHIJJIJJJIJJJJHHFFDDDDDDDDD
D@DDD?BDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDBDDDDDDDDDDDB    AS:i:-1 ZS:i:-1 XN:i:0
XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:99 YS:i:0  YT:Z:CP NH:i:2
SRR1660321.14714584     355     chr1    11027   1       1S99M   =       11095
```

Mapping quality metrics

- Generate mapping quality metrics via samtools stats

- Important metrics
  - High mapping %
  - Low MAPQ

- [Reads/Reads Mapped]*100 = percent mapped
  - 91% for this sample

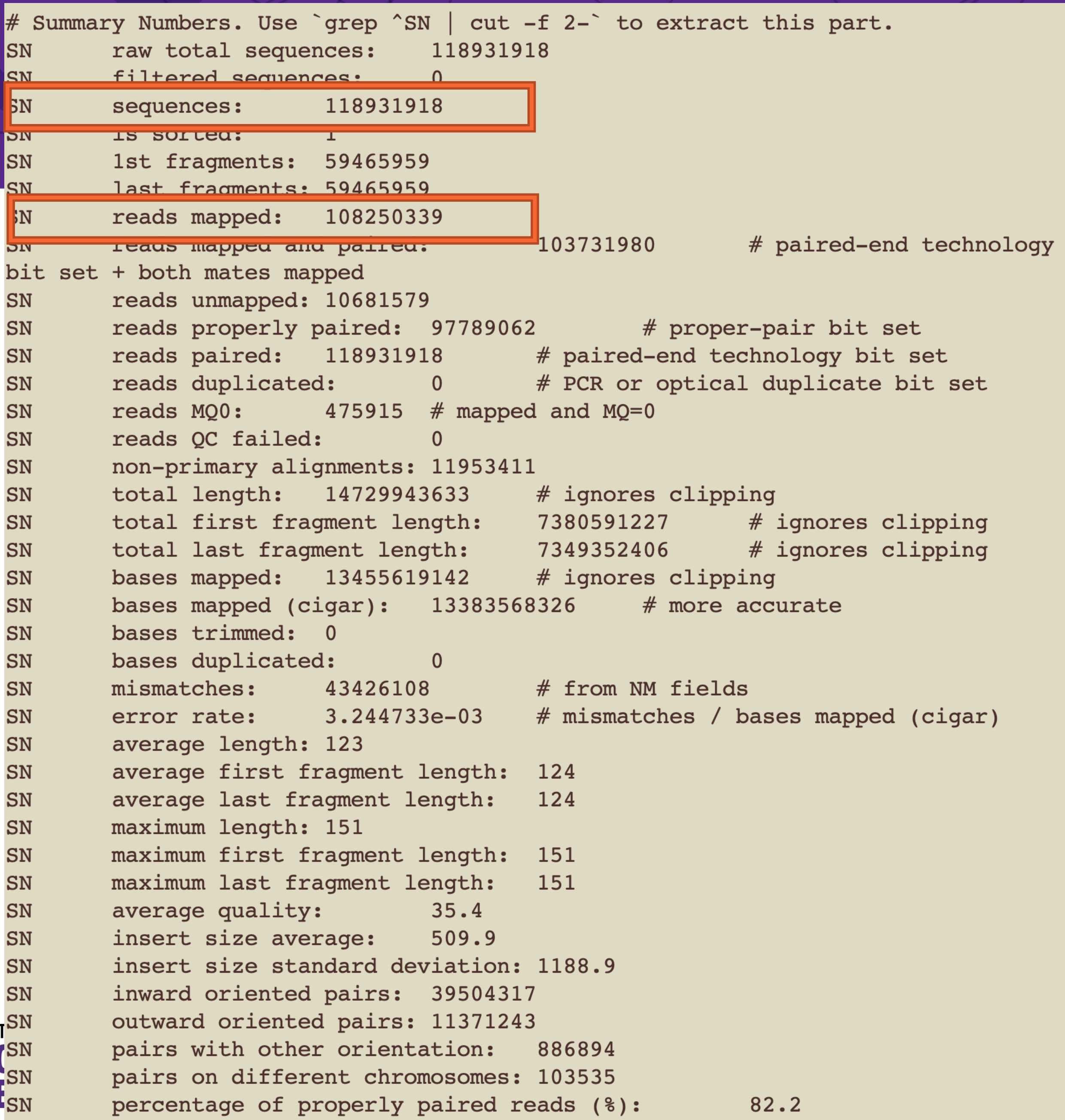


```
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN      raw total sequences:    118931918
SN      filtered sequences:     0
SN      sequences:      118931918
SN      is sorted:      1
SN      1st fragments:  59465959
SN      last fragments: 59465959
SN      reads mapped:   108250339
SN      reads mapped and paired:        103731980       # paired-end technology
bit set + both mates mapped
SN      reads unmapped: 10681579
SN      reads properly paired:  97789062        # proper-pair bit set
SN      reads paired:   118931918       # paired-end technology bit set
SN      reads duplicated:       0       # PCR or optical duplicate bit set
SN      reads MQ0:      475915  # mapped and MQ=0
SN      reads QC failed:        0
SN      non-primary alignments: 11953411
SN      total length:   14729943633     # ignores clipping
SN      total first fragment length:    7380591227      # ignores clipping
SN      total last fragment length:     7349352406      # ignores clipping
SN      bases mapped:   13455619142     # ignores clipping
SN      bases mapped (cigar):   13383568326     # more accurate
SN      bases trimmed:  0
SN      bases duplicated:       0
SN      mismatches:     43426108        # from NM fields
SN      error rate:     3.244733e-03    # mismatches / bases mapped (cigar)
SN      average length: 123
SN      average first fragment length:  124
SN      average last fragment length:   124
SN      maximum length: 151
SN      maximum first fragment length:  151
SN      maximum last fragment length:   151
SN      average quality:        35.4
SN      insert size average:    509.9
SN      insert size standard deviation: 1188.9
SN      inward oriented pairs:  39504317
SN      outward oriented pairs: 11371243
SN      pairs with other orientation:   886894
SN      pairs on different chromosomes: 103535
SN      percentage of properly paired reads (%):        82.2
```

Mapping quality metrics

- **Mapping rate**
- [Reads/Reads Mapped]*100 = percent mapped
  - 91% for this sample

- Low mapping < 70%
  - Library contamination
  - Low base quality/mutations
  - Strict mapping parameters



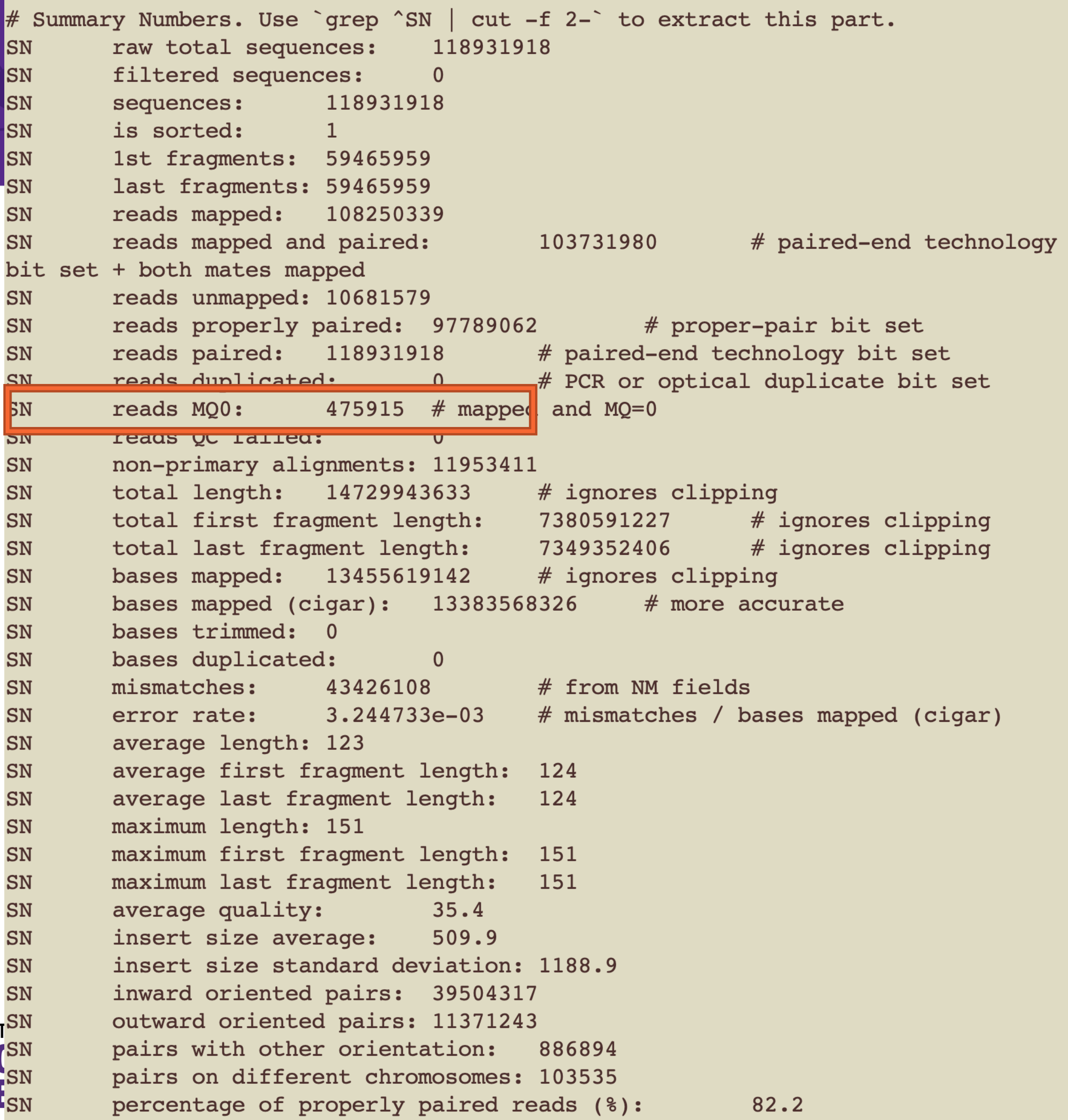
```
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN      raw total sequences:    118931918
SN      filtered sequences:     0
SN      sequences:      118931918
SN      is sorted:      1
SN      1st fragments:  59465959
SN      last fragments: 59465959
SN      reads mapped:   108250339
SN      reads mapped and paired:        103731980       # paired-end technology bit set + both mates mapped
SN      reads unmapped: 10681579
SN      reads properly paired:  97789062        # proper-pair bit set
SN      reads paired:   118931918       # paired-end technology bit set
SN      reads duplicated:       0       # PCR or optical duplicate bit set
SN      reads MQ0:      475915  # mapped and MQ=0
SN      reads QC failed:        0
SN      non-primary alignments: 11953411
SN      total length:   14729943633     # ignores clipping
SN      total first fragment length:    7380591227      # ignores clipping
SN      total last fragment length:     7349352406      # ignores clipping
SN      bases mapped:   13455619142     # ignores clipping
SN      bases mapped (cigar):   13383568326     # more accurate
SN      bases trimmed:  0
SN      bases duplicated:       0
SN      mismatches:     43426108        # from NM fields
SN      error rate:     3.244733e-03    # mismatches / bases mapped (cigar)
SN      average length: 123
SN      average first fragment length:  124
SN      average last fragment length:   124
SN      maximum length: 151
SN      maximum first fragment length:  151
SN      maximum last fragment length:   151
SN      average quality:        35.4
SN      insert size average:    509.9
SN      insert size standard deviation: 1188.9
SN      inward oriented pairs:  39504317
SN      outward oriented pairs: 11371243
SN      pairs with other orientation:   886894
SN      pairs on different chromosomes: 103535
SN      percentage of properly paired reads (%):        82.2
```

CLEMSON® UNIVERSIT
CENTER FC
HUMAN GE

Mapping quality metrics

- **Mapping quality**
  - Metrics **read multi-mapping**
  - **Also utilizes hit identity**
- Changes by aligner

- HISAT2
  - MAPQ 60 = uniquely mapped,
  - MAPQ 1 = multiple mapped, high hit identity
  - MAPQ 0 = unmapped, multiple mapped, low hit identity



```
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN      raw total sequences:    118931918
SN      filtered sequences:     0
SN      sequences:      118931918
SN      is sorted:      1
SN      1st fragments:  59465959
SN      last fragments: 59465959
SN      reads mapped:   108250339
SN      reads mapped and paired:        103731980       # paired-end technology
bit set + both mates mapped
SN      reads unmapped: 10681579
SN      reads properly paired:  97789062        # proper-pair bit set
SN      reads paired:   118931918       # paired-end technology bit set
SN      reads duplicated:       0       # PCR or optical duplicate bit set
SN      reads MQ0:      475915  # mapped and MQ=0
SN      reads QC failed:        0
SN      non-primary alignments: 11953411
SN      total length:   14729943633     # ignores clipping
SN      total first fragment length:    7380591227      # ignores clipping
SN      total last fragment length:     7349352406      # ignores clipping
SN      bases mapped:   13455619142     # ignores clipping
SN      bases mapped (cigar):   13383568326     # more accurate
SN      bases trimmed:  0
SN      bases duplicated:       0
SN      mismatches:     43426108        # from NM fields
SN      error rate:     3.244733e-03    # mismatches / bases mapped (cigar)
SN      average length: 123
SN      average first fragment length:  124
SN      average last fragment length:   124
SN      maximum length: 151
SN      maximum first fragment length:  151
SN      maximum last fragment length:   151
SN      average quality:        35.4
SN      insert size average:    509.9
SN      insert size standard deviation: 1188.9
SN      inward oriented pairs:  39504317
SN      outward oriented pairs: 11371243
SN      pairs with other orientation:   886894
SN      pairs on different chromosomes: 103535
SN      percentage of properly paired reads (%):        82.2
```

CLEMSON® UNIVERSIT
CENTER FC
HUMAN GE

Conclusions

- Fasta text files contain only sequence and no meta data
- Fastq formats come straight from sequencers
  - Read and base quality information

- Sam files store alignment information in text format
  - Alignments, quality metrics
- BAM files are store alignments as compressed, binary data
- Visit github page for more scripts and resources on today lecture: https://github.com/herber4/NGS_Formats_QC.git