25th - 29th March 2019

Sequence Data and File Formats

THE UNIVERSITY OF MELBOURNE

Melbourne Bioinformatics

Sequence Readsets

What you get:

Millions to billions of **reads** - in one big file (or two!!)

```
ATGCTTCTCCGCCTTTAATTAAAATTCCATTTCGTGCACCAACACCCGTTCCTACCATAATAGCTGTTGGAGTCGCTAAACCTAATGCACATGGACACGC      <- 1st read
CTAAGATACTGCCATCTTCTTCCAACGTAAATTGTACGTGATTTTCGATCCATTTTCTTCGAGGTTCTACTTTGTCACCCATTAGTGTGGTTACTCGACG      <- 2nd read
GAATATGCGTGGACAGATGACGAATTGGCAGCAATGATTAAAAAAGTCGGCAAAGGATATATGCTACAGCGATATAAAGGACTTGGAGAGATGAATGCGG
ATCAATGCAAATACAAGATGTGACAATGCGCGCAATGCAATGATAACTGGTGTTGTCAAAAAGAAACCGAATGTCGTACCTAGTGCAACAGCCACTGCAA
GGAAAAAATGAGAAAAAAATTCAGTTCGAAAACTAACGATTTCTGCTTTATTGATTGGGATGGGGGTCATTATCCCAATGGTTATGCCTAAAATCATGATC
GATGAAACAATCCAACAAATACCATTCAATAATTTCACAGGGGAAAATGAGACNCTAAGTTTCCCCGTATCAGAAGCAACAGAAAGAATGGTGTTTCGCT
...
...                                         <--- 100 bp --->
...
AGGCATCTTGAAAAAACAAGTGTGTGCCTCTGCGATAATCAATGCCACAGAGGTGCATAAAATTAGTTGTCGAAAAAATAATCGCTACCGTTGAGACTTC
AAAGGAGCATTCTTCGCACGCGGCAAAAAAAGAATACAAACGCATGTCTATAAAAGAGACAACCCAAATTACCAGACAGTTAAACGCGATTTATAAGGCT
GTGACAAAAATCGTGTCACAGCTTCTTTTATATCCTGTCTTTTTTTAGTTATTTATTTTTCAACCTTATCAATATGACTTGATAGCCTTTTCTTTTTCGA
AACTTGTTAAAAAAGACGTCAATGCCTTAACTGTACGTGATTCTTCTGCAGTTAGGGGATGACCTTTGACTACTAAAACAGATGCCATATGCTTACCTTC
ACAAAGCATATTTGTAGGAACGATTGAAAGCATCACTCAAGTAGAAGCGGAAGAAGAAACGATTCAACTGAAACTCGTCGATGTCATGGCCAAAGAAGAT
AATTGGACTTTGTCACCGATTTTCAGTTCATCTATGTCCACGCTTATTTTTTCAGCAGTAGCATTCAAAATCACTCCGTCATTGCTGAATGATGTCCCCA
CTCCTGTTTCTTTATCTATAATTGAACTGTAAACATGAGGAATCACTTTTTTTACACCTGCATCGATTGCAATTTTCAGAATTTCTTCAAAGTTTGAAAG
AAACTGCCATTCAAATGCTGCAAGACATGGGAGGTACTTCAATCAAGTATTTCCCGATGAAAGGCTTAGCACATAGGGAAGAATTTAAAGCAGTTGCGGA
ATCATTCCTACGCCAGTCATTTCGCGTAGTTCTTTTACCATTTTAGCTGTAACGTCTGCCATGTTTAACTCCTCCTGTGTGTGTTCTTTTTAAAAAAAGC      <- last read
```

What you get:

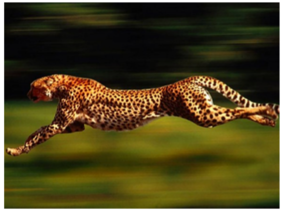
Millions to billions of **reads** - in one big file (or two!!)

```
ATGCTTCTCCGCCTTTAATTAAAATTCCATTTCGTGCACCAACACCCGTTCCTACCATAATAGCTGTTGGAGTCGCTAAACCTAATGCACATGGACACGC      <- 1st read

CTAAGATACTGCCATCTTCTTCCAACGTAAATTGTACGTGATTTTCGATCCATTTTCTTCGAGGTTCTACTTTGTCACCCATTAGTGTGGTTACTCGACG      <- 2nd read

GAATATGCGTGGACAGATGACGAATTGGCAGCAATGATTAAAAAAGTCGGCAAAGGATATATGCTACAGCGATATAAAGGACTTGGAGAGATGAATGCGG

ATCAATGCAAATACAAGATGTGACAATGCGCGCAATGCAATCATAACTGGTGTTGTCAAAAAGAAACCGAATGTCGTACCTAGTGCAACAGCCACTGCAA

GGAAAAAATGAGAAAAAAATTCAGTTCGAAAACTAACGATTTCTGCTTTATTGATTGGGATGGGGGTCATTATCCCAATGGTTATGCCTAAAATCATGATC

GATGAAACAATCCAACAAATACCATTCAATAATTTCACAGGGGAAAATGAGACTCTAAGTTTCCCCGTATCAGAAGCAACAGAAAGAATGGTGTTTCGCT

...

...                                        <--- 100 bp --->

...

AGGCATCTTGAAAAAACAAGTGTGTGCCTCTGCGATAATCAATGCCACAGAGGTGCATAAAATTAGTTGTCGAAAAAATAATCGCTACCGTTGAGACTTC

AAAGGAGCATTCTTCGCACGCGGCAAAAAAAGAATACAAACGCATGTCTATAAAAGAGACAACCCAAATTACCAGACAGTTAAACGCGATTTATAAGGCT

GTGACAAAAATCGTGTCACAGCTTCTTTTATATCCTGTCTTTTTTTAGTTATTTATTTTTCAACCTTATCAATATGACTTGATAGCCTTTTCTTTTTCGA

AACTTGTTAAAAAAGACGTCAATGCCTTAACTGTACGTGATTCTTCTGCAGTTAGGGGATGACCTTTGACTACTAAAACAGATGCCATATGCTTACCTTC

ACAAAGCATATTTGTAGGAACGATTGAAAGCATCACTCAAGTAGAAGCGGAAGAAGAAACGATTCAACTGAAACTCGTCGATGTCATGGCCAAAGAAGAT

AATTGGACTTTGTCACCGATTTTCAGTTCATCTATGTCCACGCTTATTTTTTCAGCAGTAGCATTCAAAATCACTCCGTCATTGCTGAATGATGTCCCCA

CTCCTGTTTCTTTATCTATAATTGAACTGTAAACATGAGGAATCACTTTTTTTACACCTGCATCGATTGCAATTTTCAGAATTTCTTCAAAGTTTGAAAG

AAACTGCCATTCAAATGCTGCAAGACATGGGAGGTACTTCAATCAAGTATTTCCCGATGAAAGGCTTAGCACATAGGGAAGAATTTAAAGCAGTTGCGGA

ATCATTCCTACGCCAGTCATTTCGCGTAGTTCTTTTACCATTTTAGCTGTAACGTCTGCCATGTTTAACTCCTCCTGTGTGTGTTCTTTTTAAAAAAAGC     <- last read
```

not in this format!

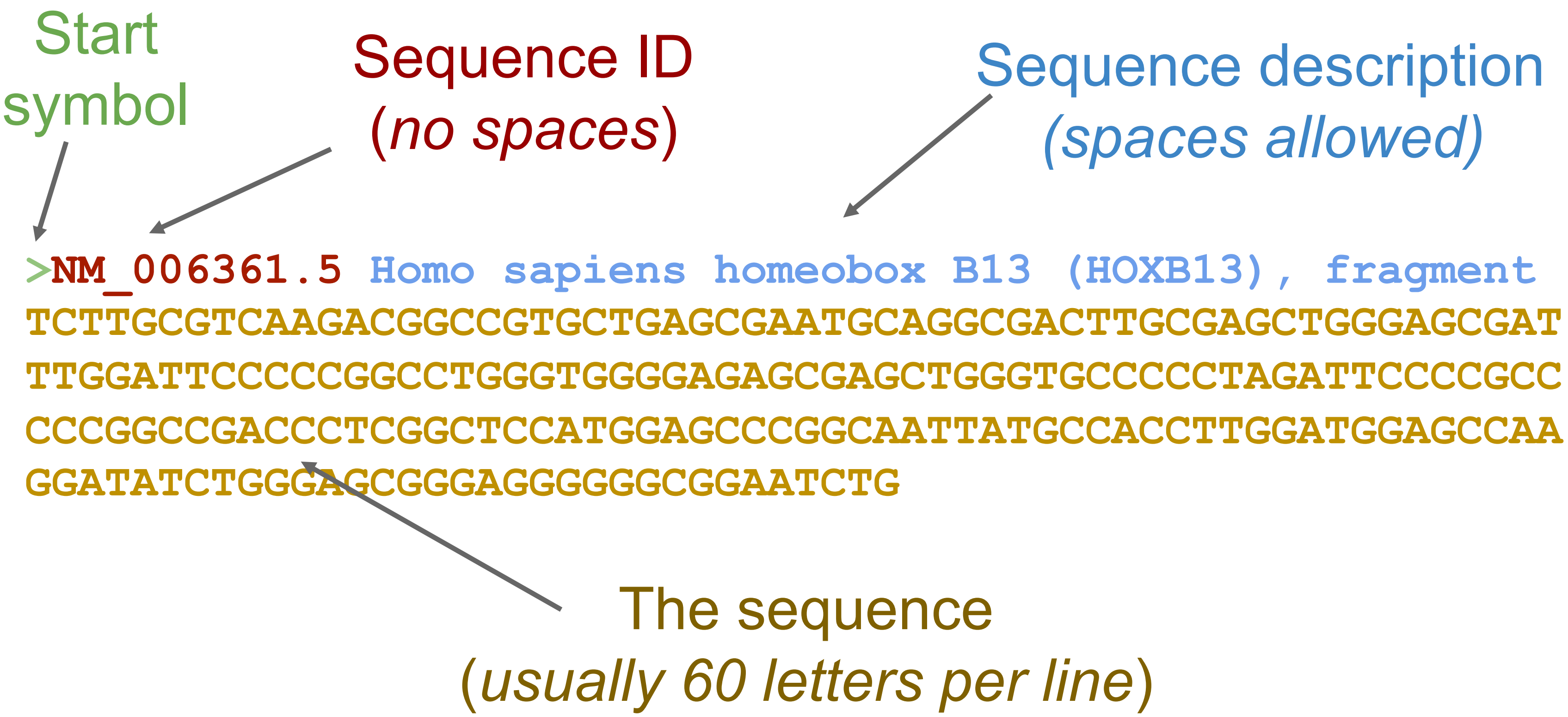
FASTA format

FASTA



```
>NM_006361.5 Homo sapiens homeobox B13 (HOXB13), fragment
TCTTGCGTCAAGACGGCCGTGCTGAGCGAATGCAGGCGACTTGCGAGCTGGGAGCGAT
TTGGATTCCCCCGGCCTGGGTGGGGAGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCC
CCCGGCCGACCCTCGGCTCCATGGAGCCCGGCAATTATGCCACCTTGGATGGAGCCAA
GGATATCTGGGAGCGGGAGGGGGGCGGAATCTG
```

FASTA components



Start symbol

Sequence ID *(no spaces)*

Sequence description *(spaces allowed)*

```
>NM_006361.5 Homo sapiens homeobox B13 (HOXB13), fragment
TCTTGCGTCAAGACGGCCGTGCTGAGCGAATGCAGGCGACTTGCGAGCTGGAGCGAT
TTGGATTCCCCGGCCTGGGTGGGAGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCC
CCCGGCCGACCCTCGGCTCCATGGAGCCCGGCAATTATGCCACCTTGGATGGAGCCAA
GGATATCTGGGAGCGGAGGGGGCGGAATCTG
```

The sequence *(usually 60 letters per line)*

Multi-FASTA

Concatenation of individual FASTA entries, using ">" as an entry separator

```
>read00001

TCTTGCGTCAAGACGGCCGTGCTGAGCGAATGCAGGCGACTTGCGAGCTGGGAGCGA

>read00002

TGGATTCCCCCGGCCTGGGTGGGGAGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCC

>read00003

GGCCGACCCTCGGCTCCATGGAGCCCGGCAATTATGCCACCTTGGATGGAGCCAAGG

>read00004

TCTGGGAGCGGGAGGGGGGCGGAATCTGGAGCGAGCTGGGTGCCCCCTAGATTCCCC

>read00004

GCGGAATCTGGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCATCGTAGATTAGATAT
```

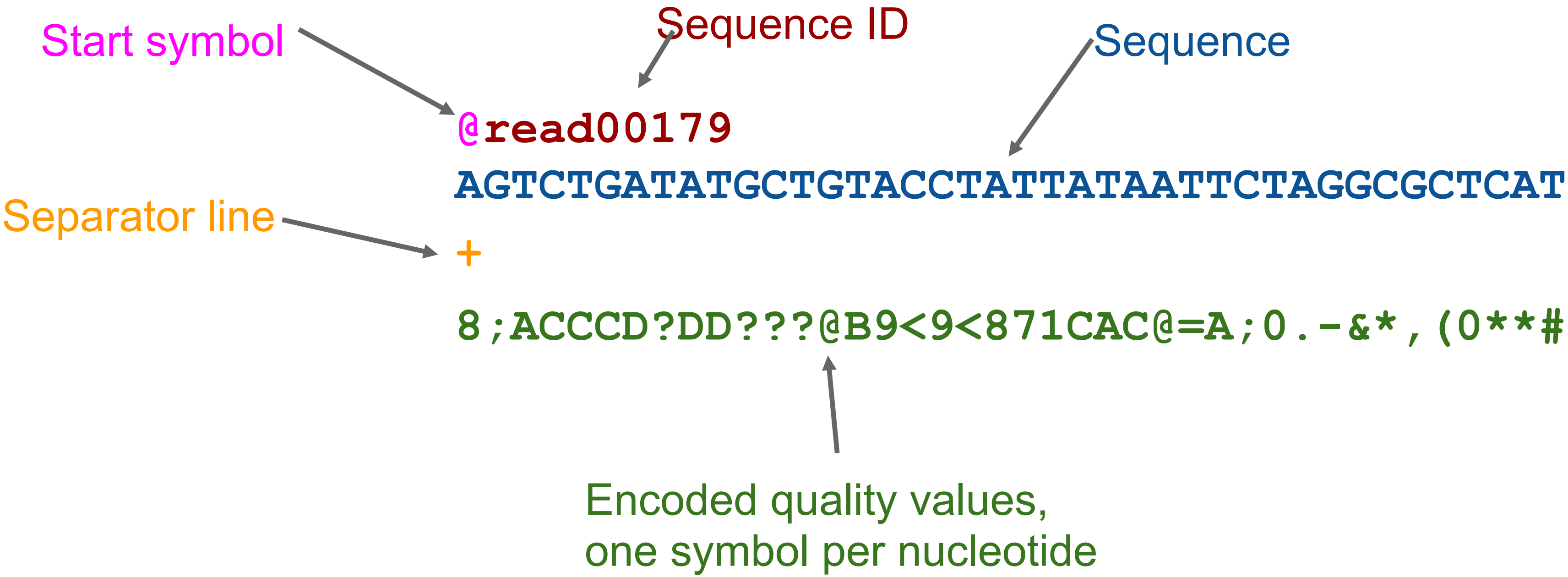
FASTQ files

FASTQ

FASTQ sequence entry looks like this:

```
@read00179
AGTCTGATATGCTGTACCTATTATAATTCTAGGCGCTCAT
+
8;ACCCD?DD???@B9<9<871CAC@=A;0.-&*,(0**#
```

FASTQ components



Start symbol

Sequence ID

Sequence

Separator line

```
@read00179
AGTCTGATATGCTGTACCTATTATAATTCTAGGCGCTCAT
+
8;ACCCD?DD???@B9<9<871CAC@=A;0.-&*,(0**#
```

Encoded quality values,
one symbol per nucleotide

Sequence Quality
“base calls”

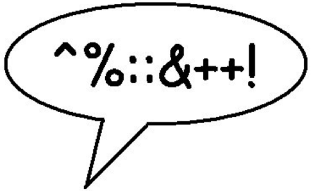
Phred Quality Scores

Quality Score	Chance of being wrong	Accuracy	Description
10	1 in 10	90%	Maybe
20	1 in 100	99%	OK
30	1 in 1000	99.9%	Good
40	1 in 10,000	99.99%	Excellent

$$Q = -10 \log_{10} P \quad \Leftrightarrow \quad P = 10^{-Q/10}$$

Q = Phred quality score P = probability of base call being incorrect

FASTQ quality encoding



Phred scores (1-40) each represented by a symbol/letter:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
|         |         |         |         |
Q0        Q10       Q20       Q30       Q40
bad       maybe     ok        good      excellent
```

What’s in a Readset?

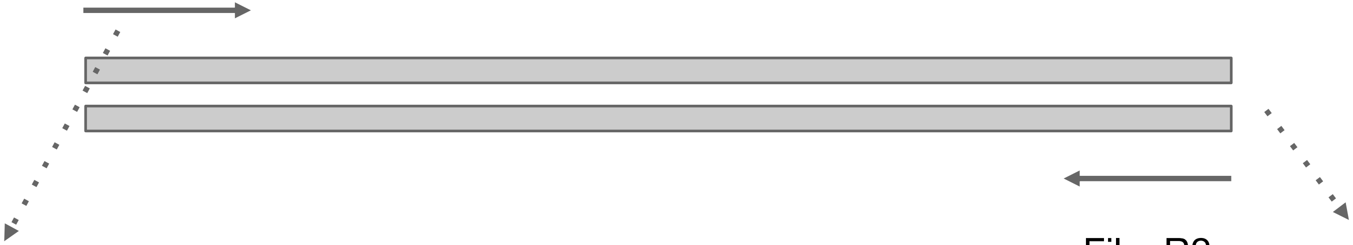
Multi-FASTQ

Same as with multi-FASTA: concatenated Fastq

```
@M00267:3:15997:1501
CTCGTGCTCTACTTTAGAAGCTAATGATTCTGTTTGTAGAACATTTTCTACCACTACATCTTTTTCTTGCTTCGCATCTT
+
:=?DD:BDDF>FFHI>E>B9AE>4C<4CCAE+AEG3?EAGEHCGIIIIIIIIIIIIGIIIEIIIIGGIDGIID/;4C<EE
@M00267:3:15997:1505
GCCTATAGTAGAAGAAAAAGAAGTGGCTCAAGAAATGAGTGCACCGCAGGAAGTTCCAGCGGCTGAATTACTTCATGAAA
+
<@@FFF?DHFHGHIIIFGIIGIGICDGEGCHIIIIIIIIIGIHIIFG<DA7=BHHGGIEHDBEBA@CECDD@CC>CCCAC
@M00267:3:14073:1508
GTCTTGCTAAATTTAAATAATCTGAAATAATTTGTTCTGCCCGGTCCAATTCAGCTAATACGAGACGCATATAATCCTTA
+
:?DDDDD?84CFHC><F>9EEH>B>+A4+CEH4FFEHFHIIIIIIIIIIIIIGGIIIIIIIIG>B7BBEBBB@CDDCFC
```

Reading from the End of the DNA

Paired End



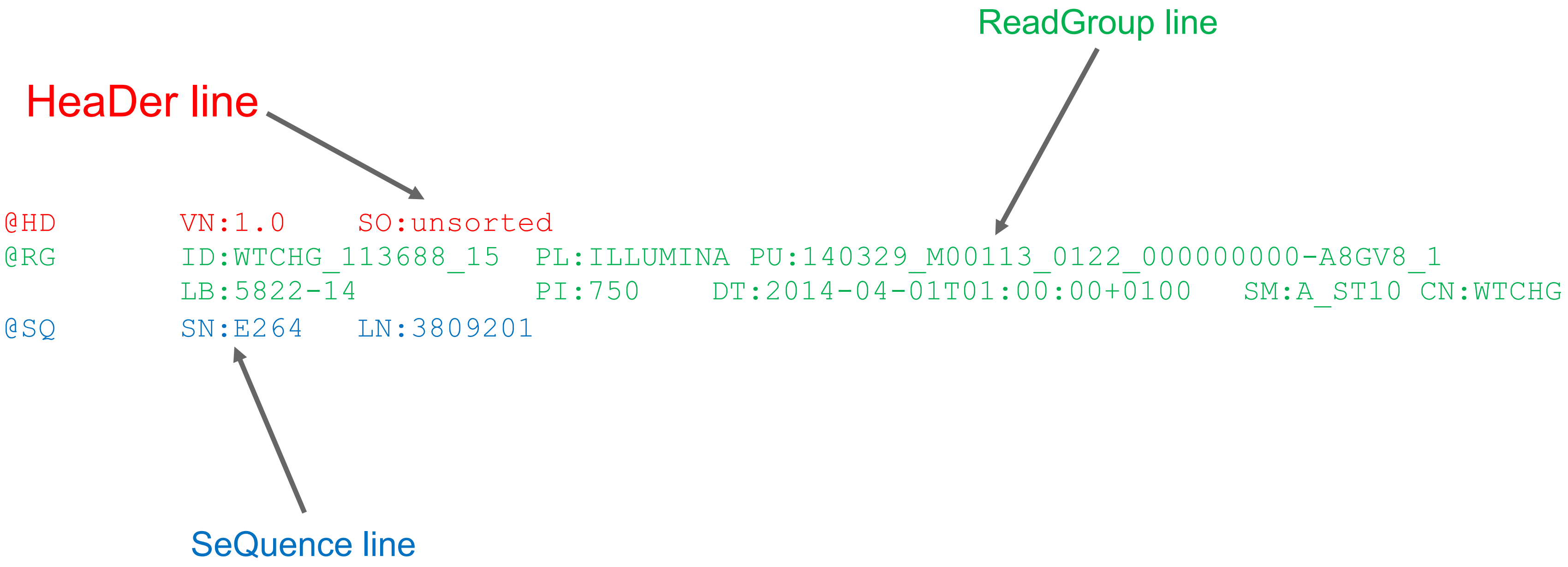
File: R1
>seq-template1
aaagtagctga
>seq-template2
aaagtagctga
…

File: R2
>seq-template1 (other strand)
ggaattccttaacc
>seq-template2 (other strand)
cgatcgtgtgagc
…

SAM/BAM file format

- Gives the reads plus alignment to a reference plus metadata
- BAM is a compressed form of SAM (smaller files)

SAM format



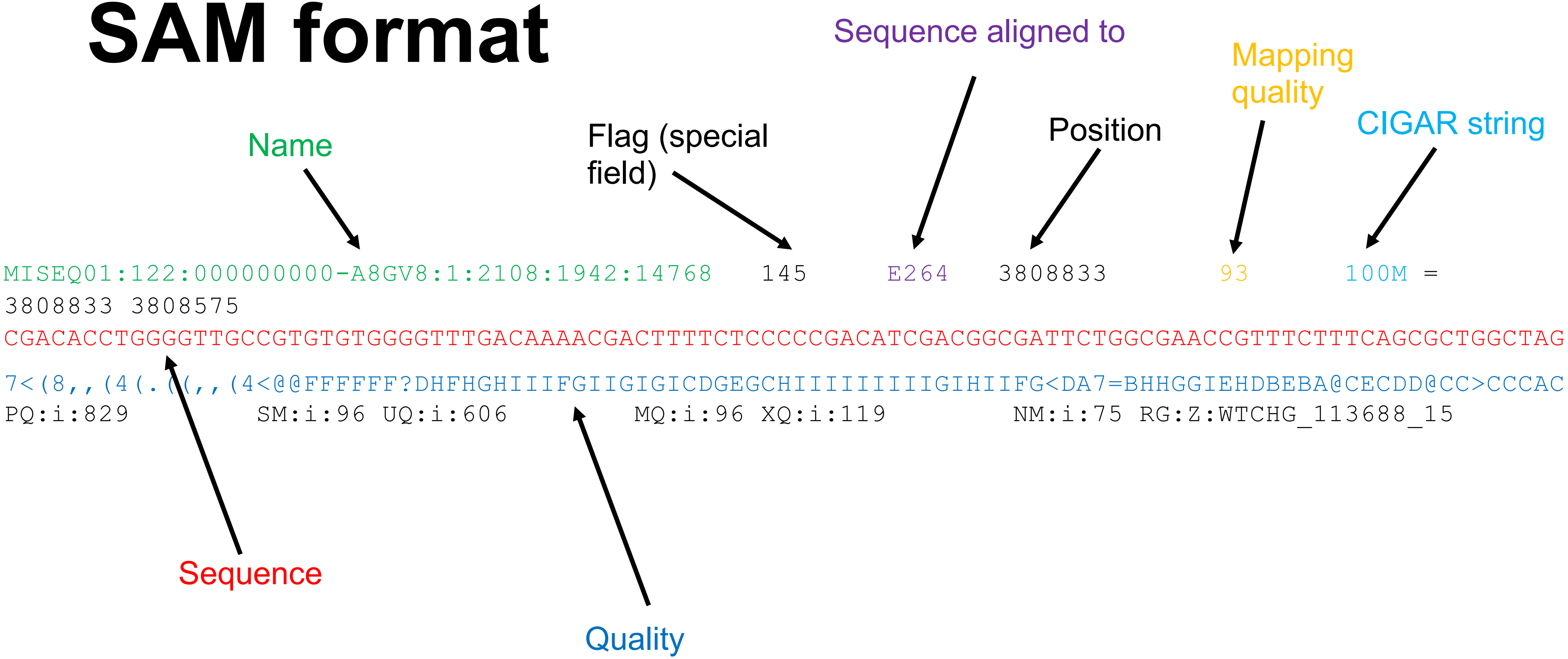
ReadGroup line

HeaDer line

```
@HD       VN:1.0    SO:unsorted
@RG       ID:WTCHG_113688_15  PL:ILLUMINA PU:140329_M00113_0122_000000000-A8GV8_1
          LB:5822-14          PI:750    DT:2014-04-01T01:00:00+0100   SM:A_ST10 CN:WTCHG
@SQ       SN:E264   LN:3809201
```

SeQuence line

SAM format



Name

Flag (special field)

Sequence aligned to

Position

Mapping quality

CIGAR string

```
MISEQ01:122:000000000-A8GV8:1:2108:1942:14768   145    E264   3808833       93      100M =
3808833 3808575
CGACACCTGGGGTTGCCGTGTGTGGGGTTTGACAAAACGACTTTTCTCCCCCGACATCGACGGCGATTCTGGCGAACCGTTTCTTTCAGCGCTGGCTAG

7<(8,,(4(.((,,(4<@@FFFFFF?DHFHGHIIIFGIIGIGICDGEGCHIIIIIIIIIGIHIIFG<DA7=BHHGGIEHDBEBA@CECDD@CC>CCCAC
PQ:i:829        SM:i:96 UQ:i:606       MQ:i:96 XQ:i:119        NM:i:75 RG:Z:WTCHG_113688_15
```

Sequence

Quality