Sequence data format & quality measurements

Torsten Struck – t.h.struck@nhm.uio.no
NHM UiO

FASTA

```
>Stygocapitella_subterranea COX1 Population 1x415
IYRWLFSTNHKDIGTLYFILGIWAGLMGTSLSLLIRTELGQPGSLLGSDQLYNTIVTAHA
MLMIFFLVMPIMIGGFGNWLIPLMIGCPDMAFPRMNNMSFWLLPPALLLMLSSAAVEQGA
GTGWTVYPPLASNMAHSGASVDLVIFSLHLAGVSSILGSANFITTIINMRSTNLSLERIP
LFIWSVKITAILLLLSLPVLAGAITMLLTDRNLNTSFFDPAGGGDPILFQHLFWFFGHPE
VYILILPGFGMISHIVSFYGGKPTSFGTLGMIYAMAGIAILGFIVWAHHMFTVGMDVDTR
AYFTAATMIIAVPTGIKVFSWLTTLSGANLKLETPLLWAMGFIFLFTMGGLTGIILANSS
IDISLHDTYYVVAHFHYVLSMGAIFAIFAGFNFWFPLLSGMTLNPKWTQAHFGLMFIGVN
LTFFPQHFLGLAGMPRRYSDYPDSYMTWNIVSSVGSVISLVSLGLFILILWESFQSQRMI
ISTYHLPSMMEWQDIQLPVDWHTSHEPPLI
```

FASTA

Descriptor (one per sequence; always starts with >)

```
>Stygocapitella_subterranea COX1 Population 1x415
IYRWLFSTNHKDIGTLYFILGIWAGLMGTSLSLLIRTELGQPGSLLGSDQLYNTIVTAHA
MLMIFFLVMPIMIGGFGNWLIPLMIGCPDMAFPRMNNMSFWLLPPALLLMLSSAAVEQGA
GTGWTVYPPLASNMAHSGASVDLVIFSLHLAGVSSILGSANFITTIINMRSTNLSLERIP
LFIWSVKITAILLLLSLPVLAGAITMLLTDRNLNTSFFDPAGGGDPILFQHLFWFFGHPE
VYILILPGFGMISHIVSFYGGKPTSFGTLGMIYAMAGIAILGFIVWAHHMFTVGMDVDTR
AYFTAATMIIAVPTGIKVFSWLTTLSGANLKLETPLLWAMGFIFLFTMGGLTGIILANSS
IDISLHDTYYVVAHFHYVLSMGAIFAIFAGFNFWFPLLSGMTLNPKWTQAHFGLMFIGVN
LTFFPQHFLGLAGMPRRYSDYPDSYMTWNIVSSVGSVISLVSLGLFILILWESFQSQRMI
ISTYHLPSMMEWQDIQLPVDWHTSHEPPLI
```

FASTA

```
>Stygocapitella_subterranea COX1 Population 1x415
IYRWLFSTNHKDIGTLYFILGIWAGLMGTSLSLLIRTELGQPGSLLGSDQLYNTIVTAHA
MLMIFFLVMPIMIGGFGNWLIPLMIGCPDMAFPRMNNMSFWLLPPALLLMLSSAAVEQGA
GTGWTVYPPLASNMAHSGASVDLVIFSLHLAGVSSILGSANFITTIINMRSTNLSLERIP
LFIWSVKITAILLLLSLPVLAGAITMLLTDRNLNTSFFDPAGGGDPILFQHLFWFFGHPE
VYILILPGFGMISHIVSFYGGKPTSFGTLGMIYAMAGIAILGFIVWAHHMFTVGMDVDTR
AYFTAATMIIAVPTGIKVFSWLTTLSGANLKLETPLLWAMGFIFLFTMGGLTGIILANSS
IDISLHDTYYVVAHFHYVLSMGAIFAIFAGFNFWFPLLSGMTLNPKWTQAHFGLMFIGVN
LTFFPQHFLGLAGMPRRYSDYPDSYMTWNIVSSVGSVISLVSLGLFILILWESFQSQRMI
ISTYHLPSMMEWQDIQLPVDWHTSHEPPLI
```

The sequence itself

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

Includes quality score per base

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

Includes quality score per base

These are paired end sequence data

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

Includes quality score per base

These are paired end sequence data

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

Includes quality score per base

These are paired end sequence data

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATTGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFF<FBFFFFFFFFFF<FFFFFFBFFFFFFFFFFFF
```

These are paired end sequence data

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

Opening descriptor (one per sequence)

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>

FASTQ

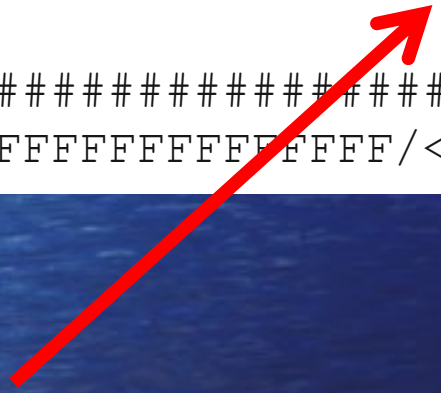
```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

Closing descriptor (one per sequence)

<read>:<is filtered>:<control number>:<index sequence>

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```



The sequences themselves

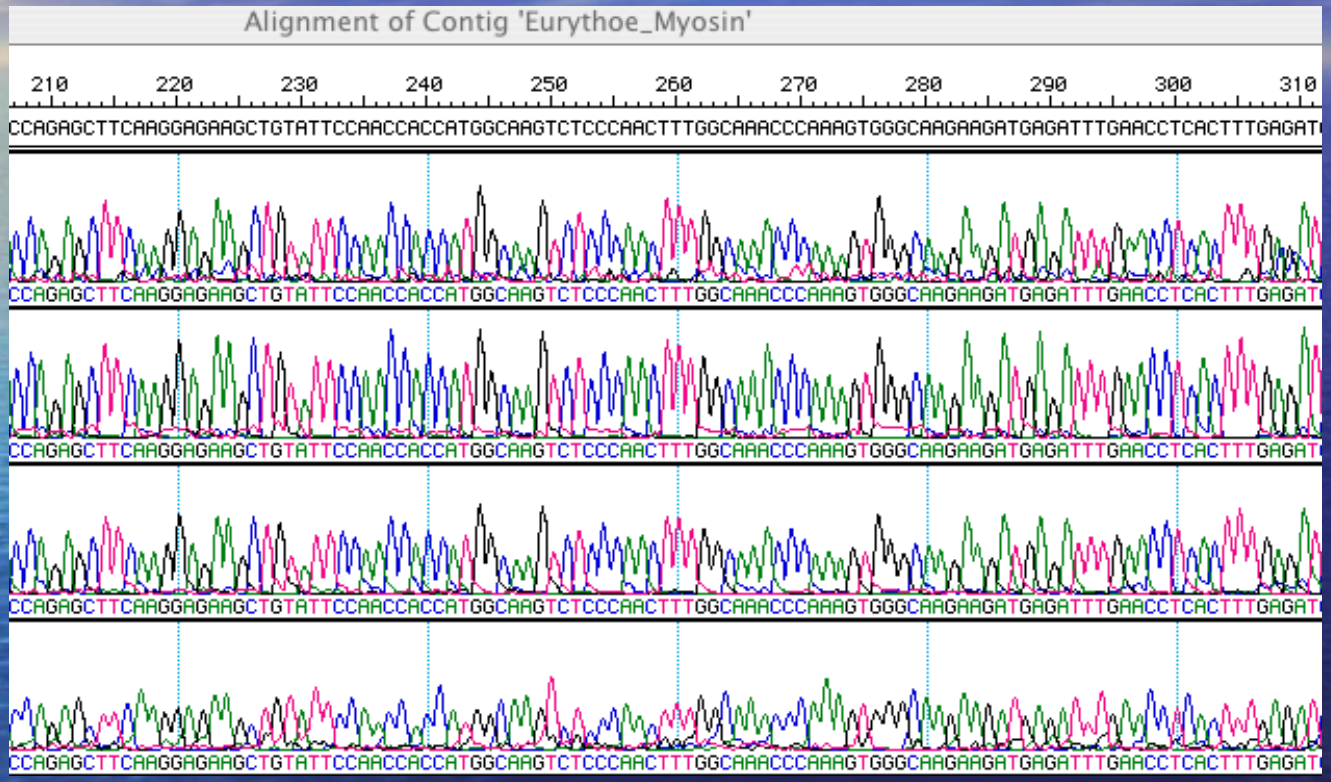
FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

The quality scores themselves

PHRED Quality Score

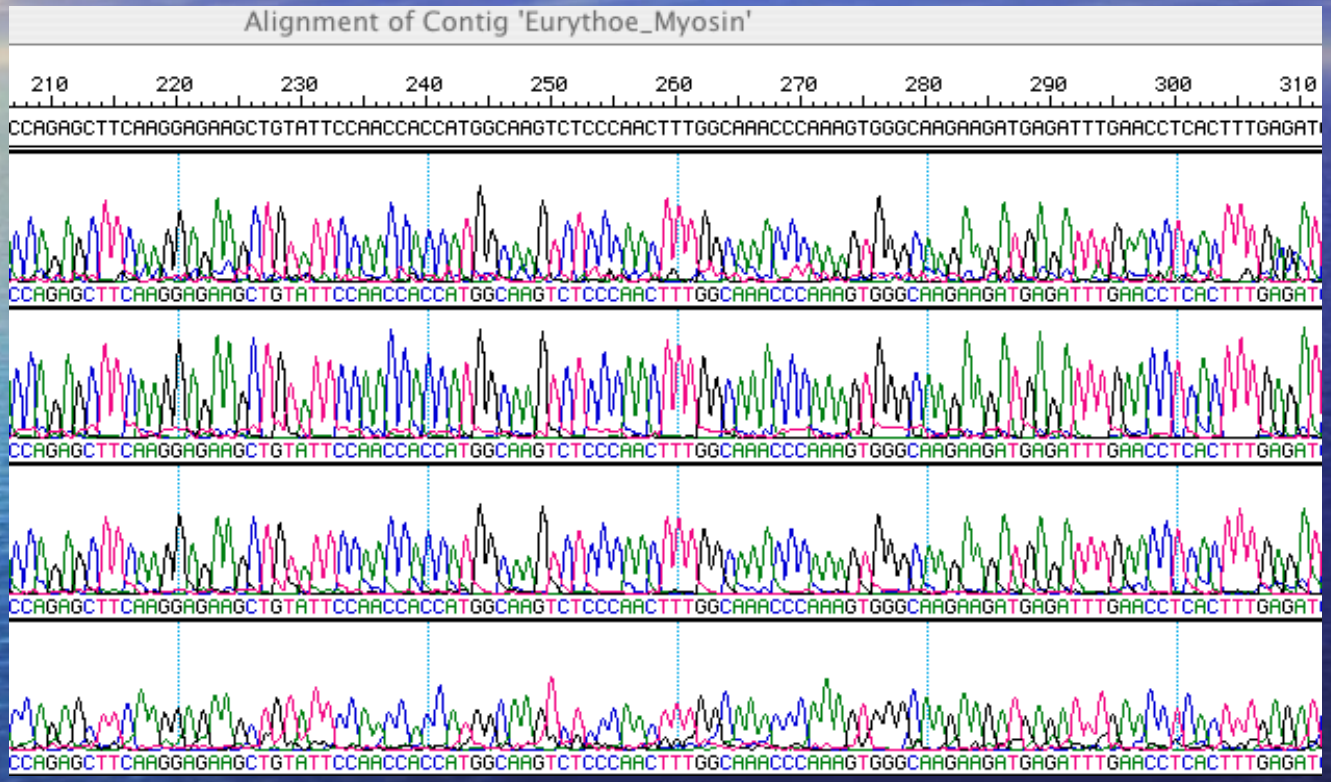
Developed by Ewing et al. (1998) Genome Research



1.) Determine idealized predicted peak locations for a given trace based upon peaks that appear to have regular spacing.

PHRED Quality Score

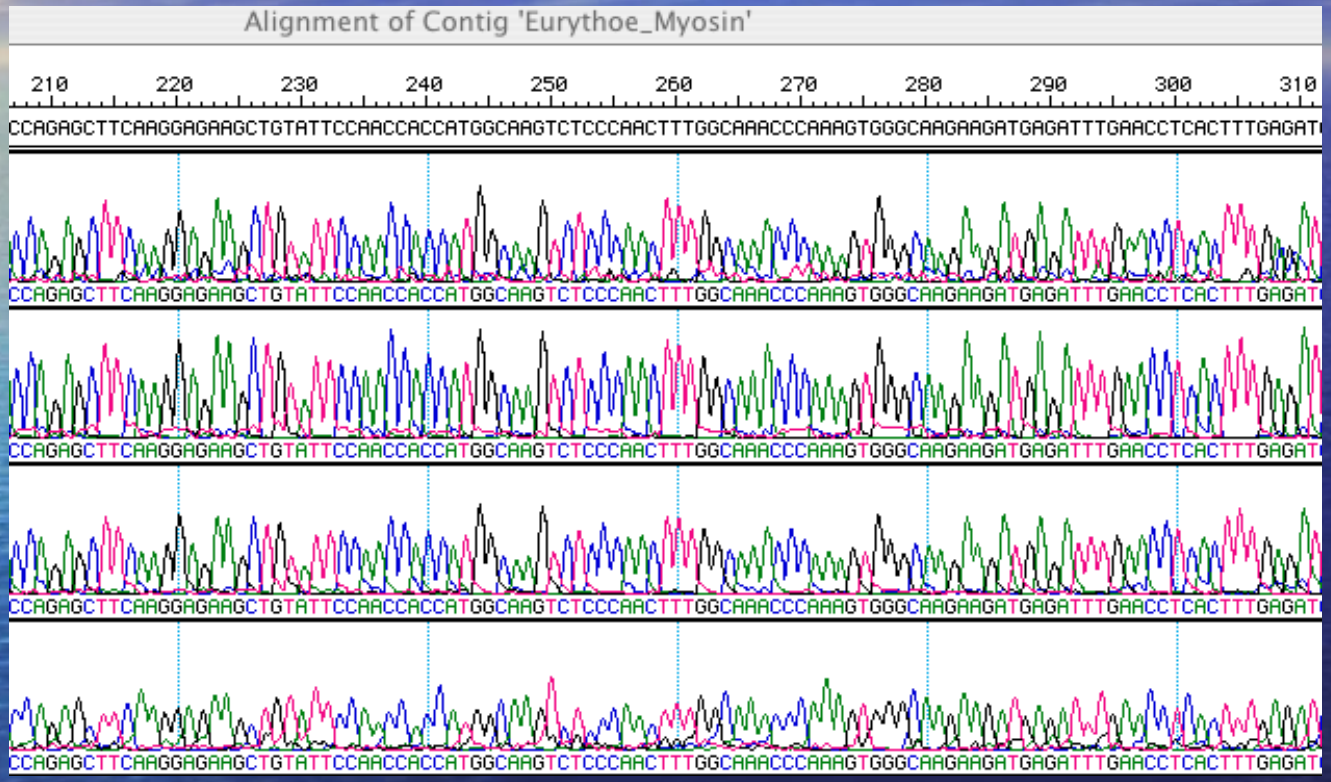
Developed by Ewing et al. (1998) Genome Research



2.) Identify observed peaks as those in the trace that exceed a minimum threshold peak area.

PHRED Quality Score

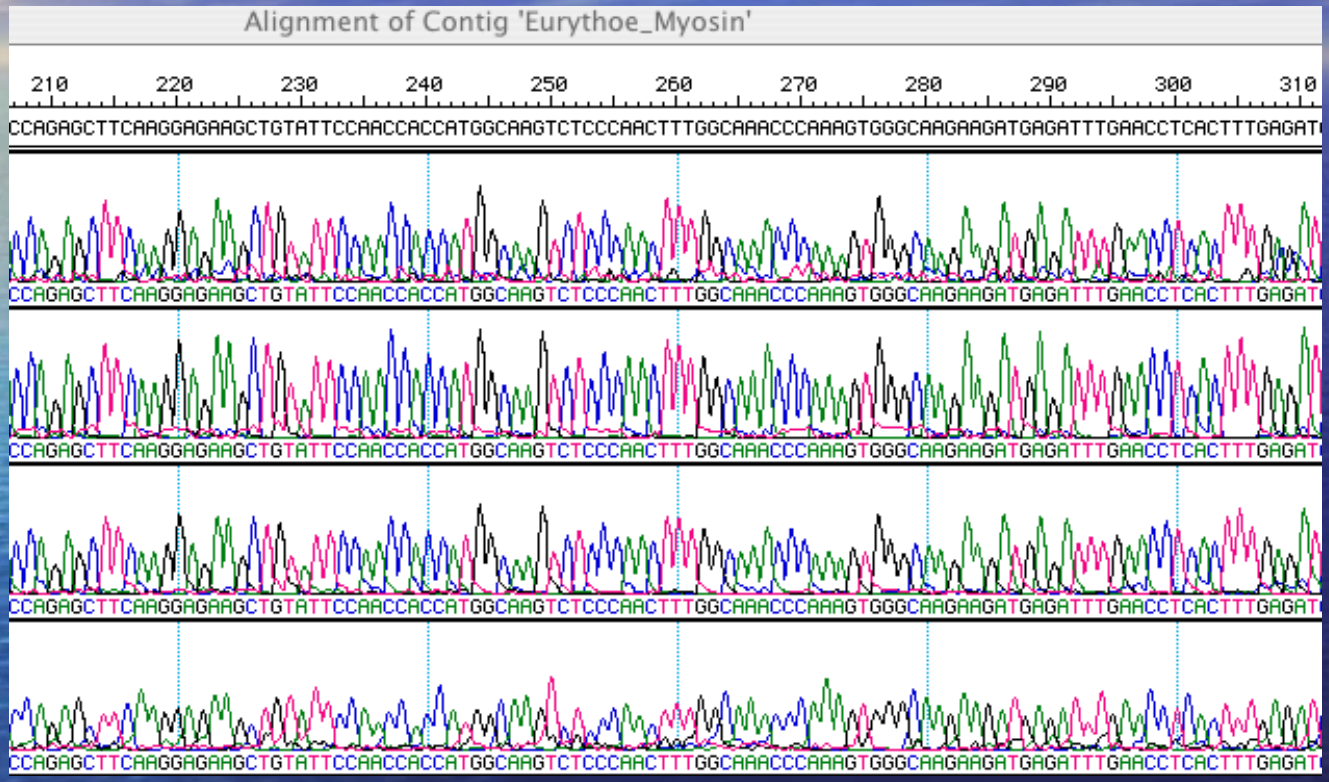
Developed by Ewing et al. (1998) Genome Research



3.) Observed peaks are matched predicted locations. Aberrantly large areas in comparison to neighbors are split in two or more peaks.

PHRED Quality Score

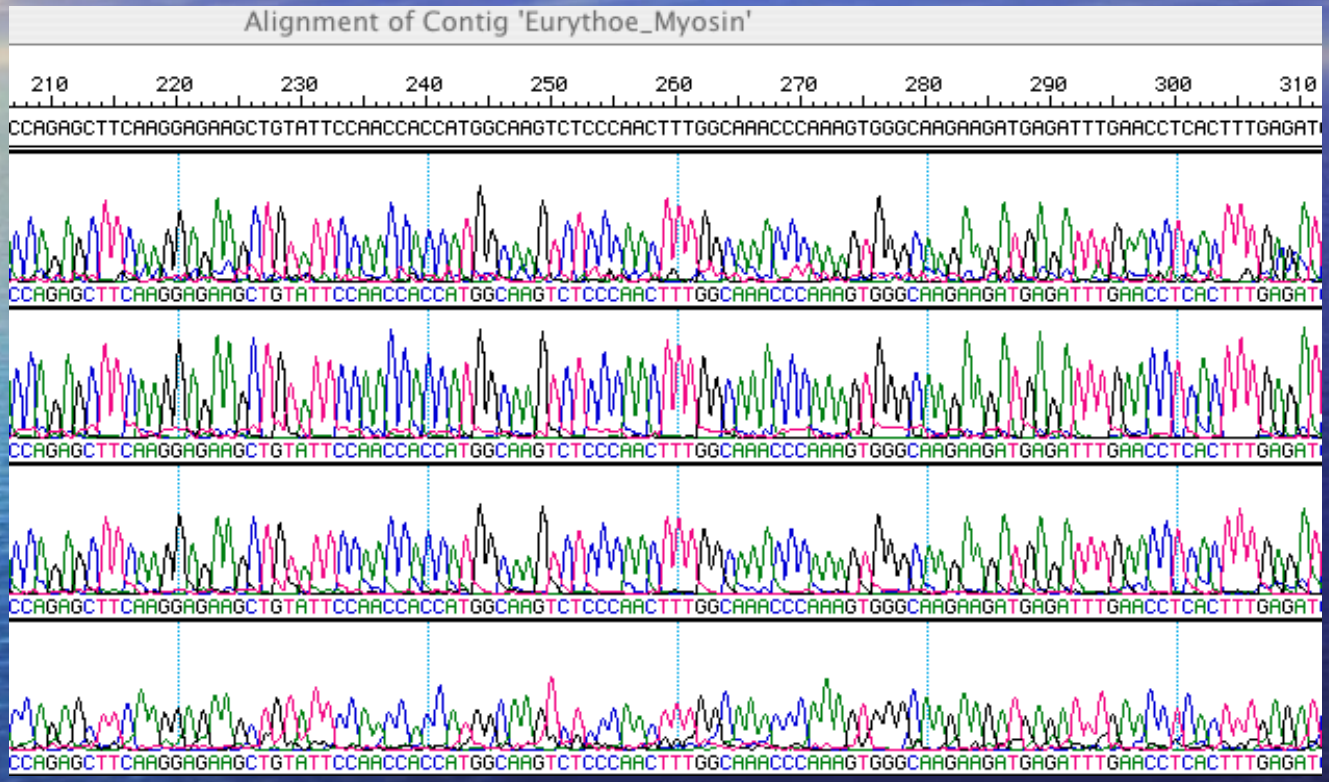
Developed by Ewing et al. (1998) Genome Research



4.) Missing peaks are accounted for from previously uncalled peaks.

PHRED Quality Score

Developed by Ewing et al. (1998) Genome Research



5.) Assign a probability from these measures (e.g., peak area, spacing, peak height, other traces) that the base call is an error.

PHRED Quality Score

5.) Probability p that the base call is an error.

$$\text{Quality score } Q = -10 * \log_{10} p$$

PHRED Quality Score

5.) Probability p that the base call is an error.

Quality score $Q = -10*\log_{10} p$

Phred quality scores are logarithmically linked to error probabilities

Phred 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%
50	1 in 100000	99.999%

PHRED Quality Score

5.) Probability p that the base call is an error.

Quality score $Q = -10*\log_{10} p$

Phred quality scores are logarithmically linked to error probabilities

Phred 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%
50	1 in 100000	99.999%

Per base, the higher the Phred score, the higher confidence it really is that base

FASTQ

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCCA
TACGCTGTTCCAGCAACATTTTTCAGTGAAATATTTGCATAGAAAACCCCGGCCGAAAGGCT
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGTA
TGGCTTCATTTTATTATGATCGATTTGCGACTTTCACCGTAATGTTGGTAGCTCTAAAGGCT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFFF
FFFFFFFFFFFFFFFFFFFFFFFFFF/<FBFFFFFFFFF<FFFFFFBFFFFFFFFFFF
```

The quality scores themselves

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCC
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFF
```

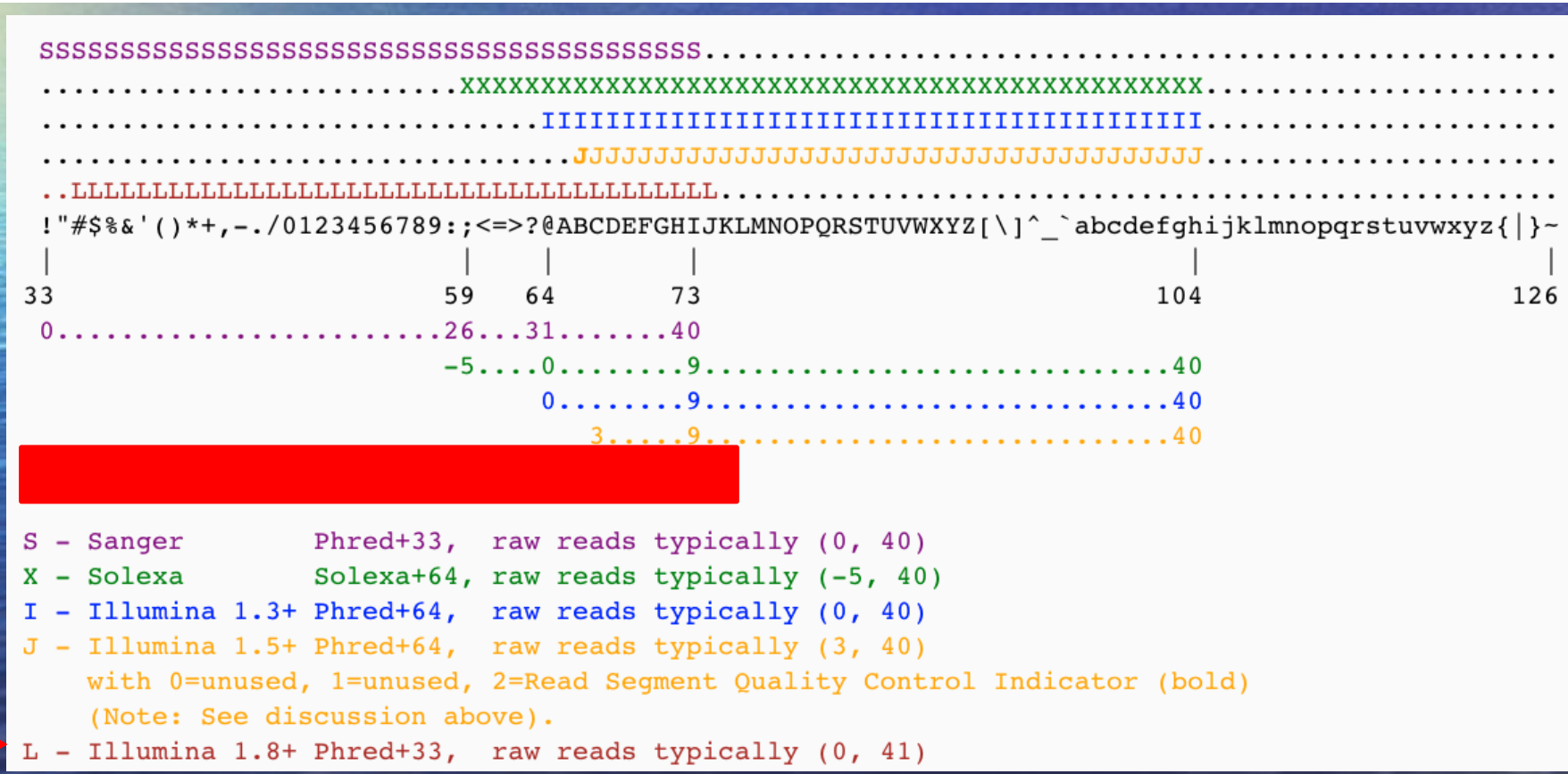
```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCC
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFF
```

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./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)
```

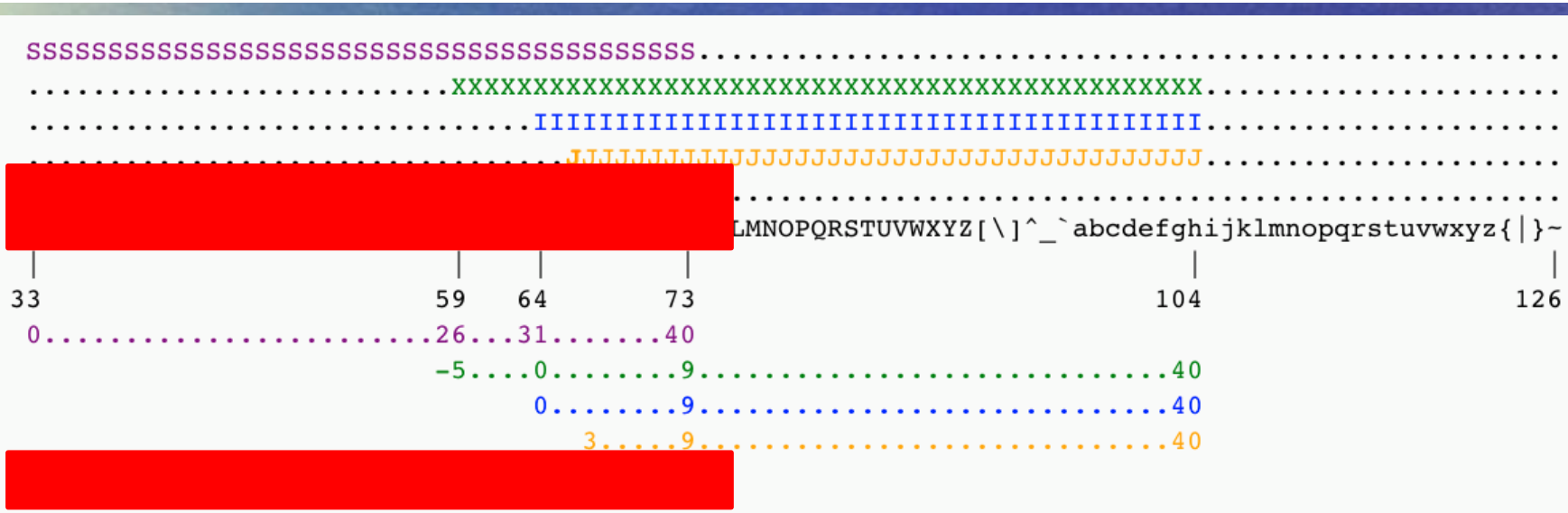
```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCC
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFF
```

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./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)
```

```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCC
+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGT
+
##############################/<<BBFBFFFFF#<<B#<<FFFFFFFFF
```



```
@HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 1:N:0:CCTAAGC
NAGCCTTTAGAGCTACCAACATTACGGTGAAAGTCGCAAATCGATCATAATAAAATGAAGCC+
#<<<<FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF@
HWI-D00351:144:C4R36ANXX:1:1101:2099:1964 2:N:0:CCTAAGC
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTCACTGAAAAANGTTNCTGGAACAGCGT+
##################################/<<BBFBFFFFF#<<B#<<FFFFFFFFF
```



Programs available

Assessing sequence quality scores:

– fastQValidator
(http://genome.sph.umich.edu/wiki/FastQValidator)
– fastqc
(http://www.bioinforma&cs.babraham.ac.uk/projects/fastqc/)
– fastx-toolkit
(http://hannonlab.cshl.edu/fastx_toolkit/download.html)

Trimming sequences based on quality scores (among others):

– fastx-toolkit
– trimmomatic, sickle, condetri, and many more

What to consider

How aggressive should I be?

What cut-offs (quality score and/or length do I apply?

Do I treat 5’ and/or 3’ ends differently?

Throw out or keep really “short” reads?