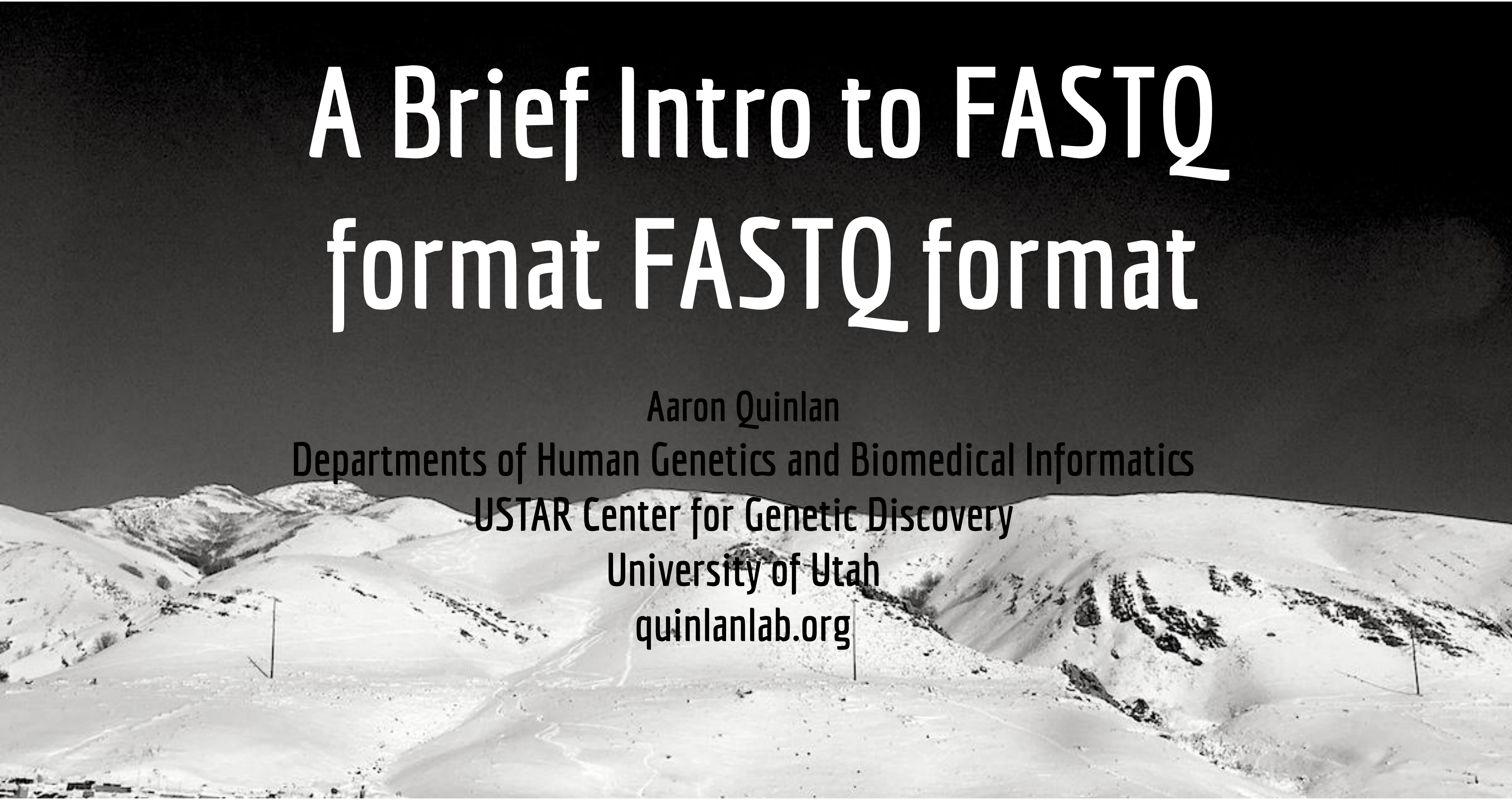
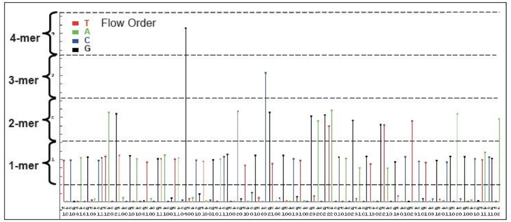
# A Brief Intro to FASTQ format FASTQ format



Aaron Quinlan
Departments of Human Genetics and Biomedical Informatics
USTAR Center for Genetic Discovery
University of Utah
quinlanlab.org

# Base calling: the conversion of signal to a nucleotide sequence



Raw signal
(e.g., 454 Life Sciences)

Base calling algorithms

Errors happen.
Hopefully infrequently

ACCTTCGAACGGCGGGGGGTTACAA

# (Mostly) all technologies yield DNA sequences in FASTQ format

DNA

```
@seq1
ACCTTCGAACGGCGGGGGGTTACAA
+
!''*((((***+))%%%++).1***
@seq2
TGGAACCGAACGGCCCCGGTTACAT
+
!''*!!!!***+))+++++).1***
And so on...
```

# The FASTQ format. Welcome to a minor hell.

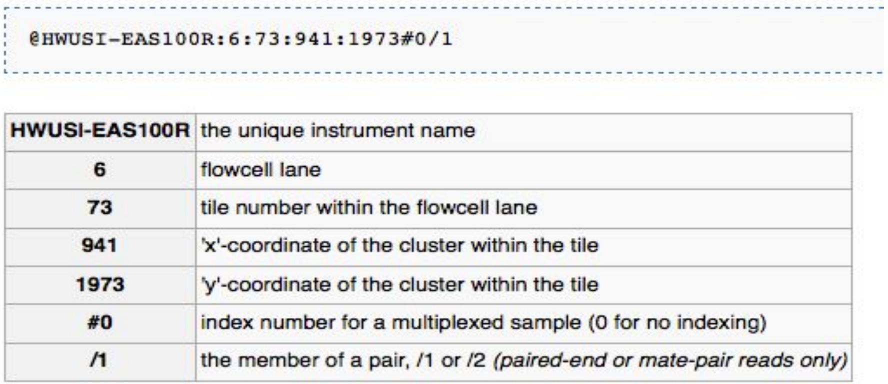
A “standard” format for storing and defining sequences from next-generation sequencing technologies.

| | |
|---|---|
| Sequence ID | @SEQ_ID |
| Sequence | GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT |
| ‹separator› | + |
| Quality scores | !''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65 |

http://en.wikipedia.org/wiki/FASTQ_format

# The FASTQ format's sequence identifier (first line of each record)

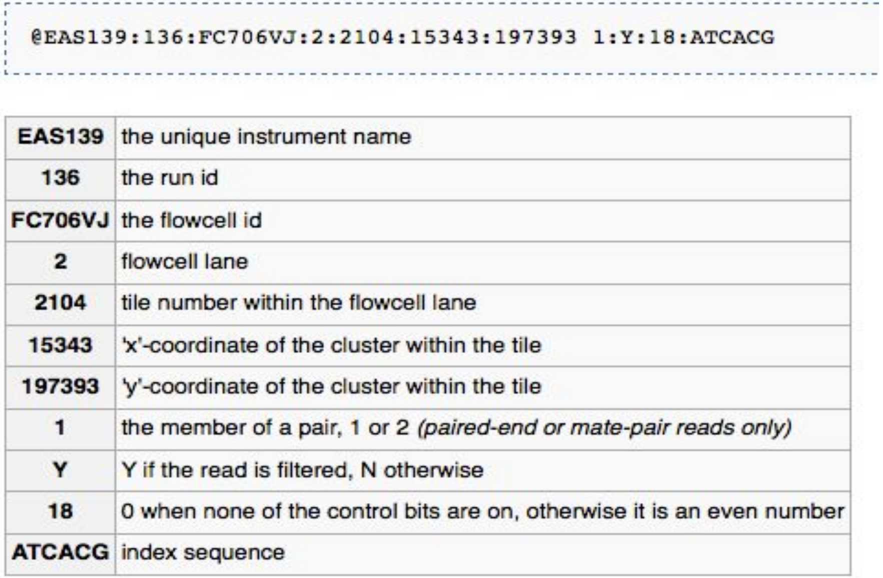
## Old format



```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

| | |
|---|---|
| **HWUSI-EAS100R** | the unique instrument name |
| **6** | flowcell lane |
| **73** | tile number within the flowcell lane |
| **941** | 'x'-coordinate of the cluster within the tile |
| **1973** | 'y'-coordinate of the cluster within the tile |
| **#0** | index number for a multiplexed sample (0 for no indexing) |
| **/1** | the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)* |

## New format



```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

| | |
|---|---|
| **EAS139** | the unique instrument name |
| **136** | the run id |
| **FC706VJ** | the flowcell id |
| **2** | flowcell lane |
| **2104** | tile number within the flowcell lane |
| **15343** | 'x'-coordinate of the cluster within the tile |
| **197393** | 'y'-coordinate of the cluster within the tile |
| **1** | the member of a pair, 1 or 2 *(paired-end or mate-pair reads only)* |
| **Y** | Y if the read is filtered, N otherwise |
| **18** | 0 when none of the control bits are on, otherwise it is an even number |
| **ATCACG** | index sequence |

# FASTQ quality scores: estimate of confidence in each base

## (sequencing technologies make errors!)

| | |
|---|---|
| **Sequence ID** | `@SEQ_ID` |
| **Sequence** | `GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT` |
| **‹separator›** | `+` |
| **Quality scores** | `!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65` |

↓

Qualities are based on the Phred scale and are *encoded*

$$Q = -10*\log_{10}(P_{err})$$

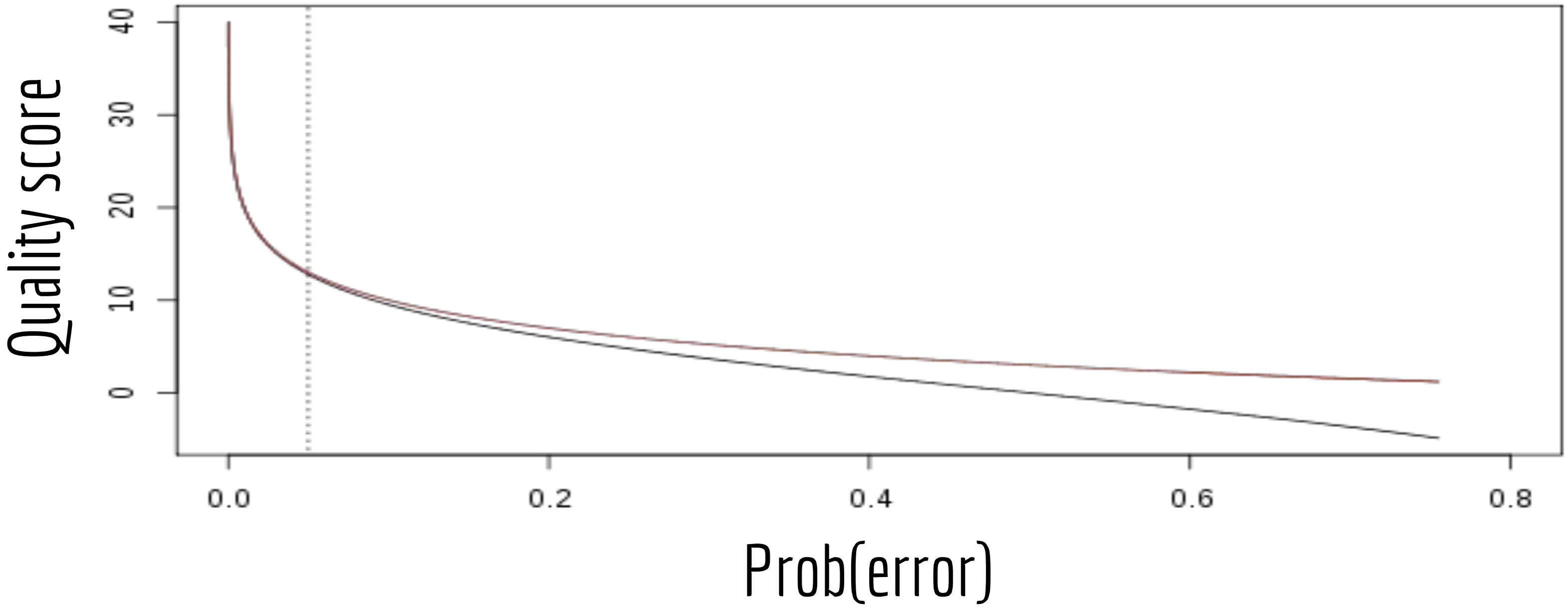
**Note**:
The Ph in Phred comes from Phil Green, the inventor of the encoding
http://www.gs.washington.edu/faculty/green.htm

# Phred quality score calculation

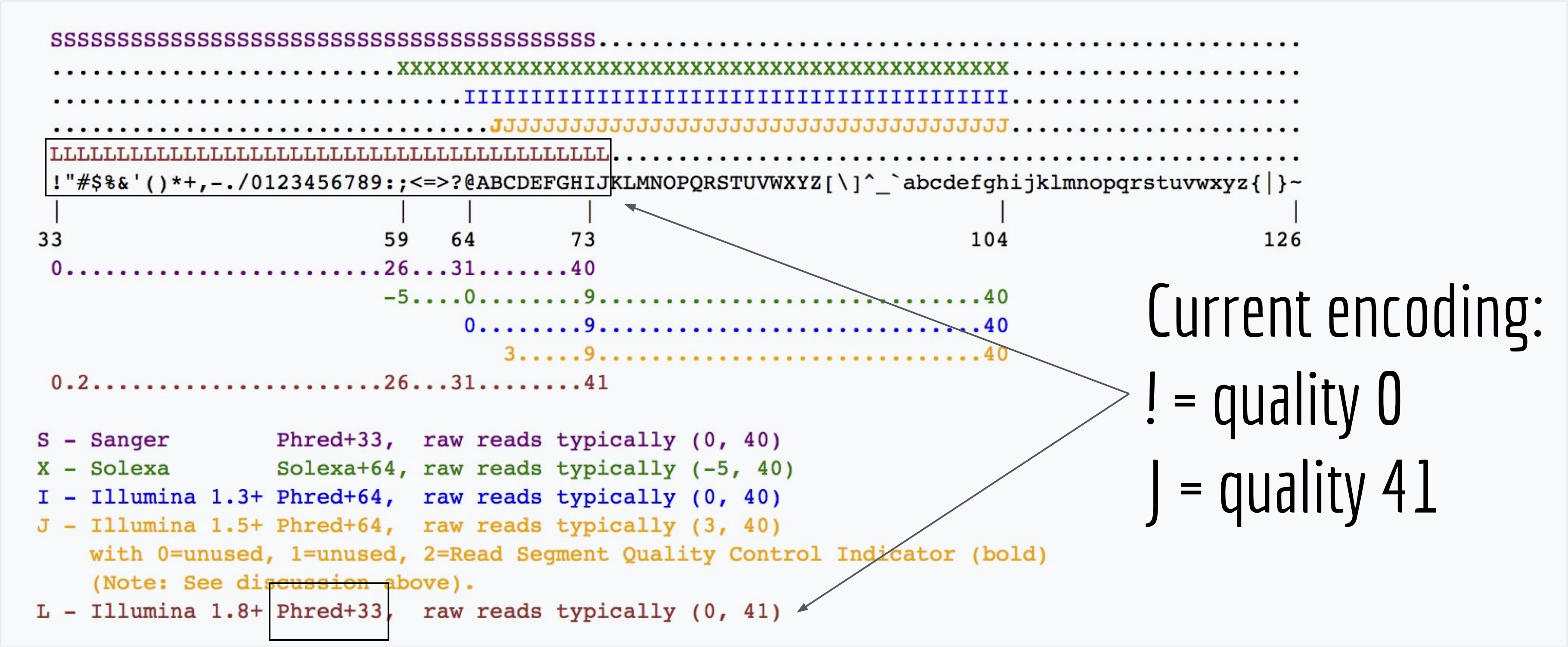
$$Q = -10*\log_{10}(P_{err})$$

| Error probability $(P_{err})$ | $\log_{10}(P_{err})$ | Phred quality score |
|---|---|---|
| 1 | 0 | 0 |
| 0.1 | -1 | 10 |
| 0.01 | -2 | 20 |
| 0.001 | -3 | 30 |
| 0.0001 | -4 | 40 |

# A higher quality score is better (>=20 is considered "good")



# Historically, FASTQ has had different encoding schemes for encoding PHRED quality scores. Ouch.



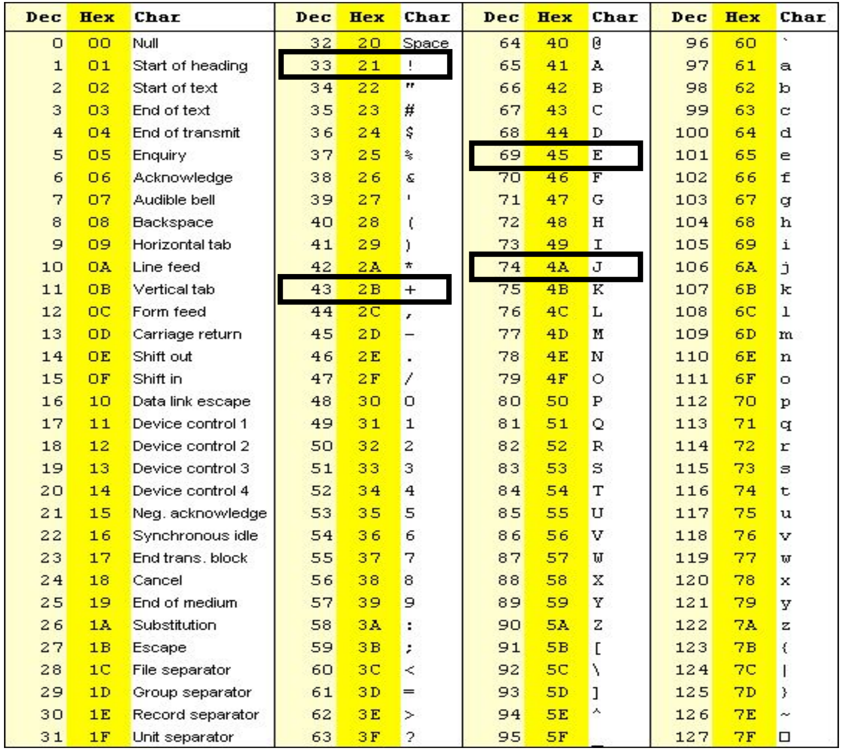
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
  !"#$%&'()*+,-./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)
```

Current encoding:

! = quality 0

J = quality 41

# Quality score encoding based on ASCII table. Geekery.

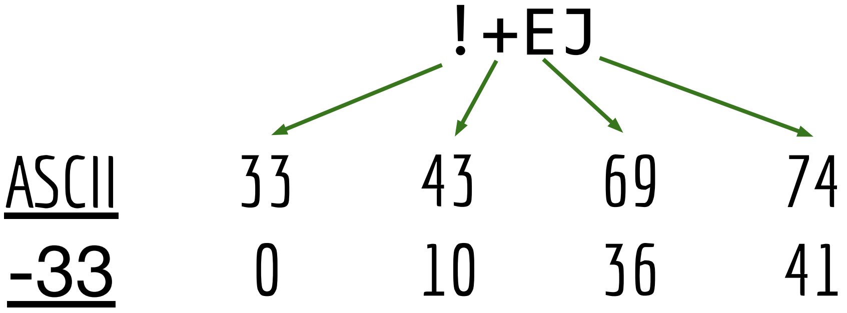


| Dec | Hex | Char | Dec | Hex | Char | Dec | Hex | Char | Dec | Hex | Char |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 00 | Null | 32 | 20 | Space | 64 | 40 | @ | 96 | 60 | ` |
| 1 | 01 | Start of heading | 33 | 21 | ! | 65 | 41 | A | 97 | 61 | a |
| 2 | 02 | Start of text | 34 | 22 | " | 66 | 42 | B | 98 | 62 | b |
| 3 | 03 | End of text | 35 | 23 | # | 67 | 43 | C | 99 | 63 | c |
| 4 | 04 | End of transmit | 36 | 24 | $ | 68 | 44 | D | 100 | 64 | d |
| 5 | 05 | Enquiry | 37 | 25 | % | 69 | 45 | E | 101 | 65 | e |
| 6 | 06 | Acknowledge | 38 | 26 | & | 70 | 46 | F | 102 | 66 | f |
| 7 | 07 | Audible bell | 39 | 27 | ' | 71 | 47 | G | 103 | 67 | g |
| 8 | 08 | Backspace | 40 | 28 | ( | 72 | 48 | H | 104 | 68 | h |
| 9 | 09 | Horizontal tab | 41 | 29 | ) | 73 | 49 | I | 105 | 69 | i |
| 10 | 0A | Line feed | 42 | 2A | * | 74 | 4A | J | 106 | 6A | j |
| 11 | 0B | Vertical tab | 43 | 2B | + | 75 | 4B | K | 107 | 6B | k |
| 12 | 0C | Form feed | 44 | 2C | , | 76 | 4C | L | 108 | 6C | l |
| 13 | 0D | Carriage return | 45 | 2D | – | 77 | 4D | M | 109 | 6D | m |
| 14 | 0E | Shift out | 46 | 2E | . | 78 | 4E | N | 110 | 6E | n |
| 15 | 0F | Shift in | 47 | 2F | / | 79 | 4F | O | 111 | 6F | o |
| 16 | 10 | Data link escape | 48 | 30 | 0 | 80 | 50 | P | 112 | 70 | p |
| 17 | 11 | Device control 1 | 49 | 31 | 1 | 81 | 51 | Q | 113 | 71 | q |
| 18 | 12 | Device control 2 | 50 | 32 | 2 | 82 | 52 | R | 114 | 72 | r |
| 19 | 13 | Device control 3 | 51 | 33 | 3 | 83 | 53 | S | 115 | 73 | s |
| 20 | 14 | Device control 4 | 52 | 34 | 4 | 84 | 54 | T | 116 | 74 | t |
| 21 | 15 | Neg. acknowledge | 53 | 35 | 5 | 85 | 55 | U | 117 | 75 | u |
| 22 | 16 | Synchronous idle | 54 | 36 | 6 | 86 | 56 | V | 118 | 76 | v |
| 23 | 17 | End trans. block | 55 | 37 | 7 | 87 | 57 | W | 119 | 77 | w |
| 24 | 18 | Cancel | 56 | 38 | 8 | 88 | 58 | X | 120 | 78 | x |
| 25 | 19 | End of medium | 57 | 39 | 9 | 89 | 59 | Y | 121 | 79 | y |
| 26 | 1A | Substitution | 58 | 3A | : | 90 | 5A | Z | 122 | 7A | z |
| 27 | 1B | Escape | 59 | 3B | ; | 91 | 5B | [ | 123 | 7B | { |
| 28 | 1C | File separator | 60 | 3C | < | 92 | 5C | \ | 124 | 7C | \| |
| 29 | 1D | Group separator | 61 | 3D | = | 93 | 5D | ] | 125 | 7D | } |
| 30 | 1E | Record separator | 62 | 3E | > | 94 | 5E | ^ | 126 | 7E | ~ |
| 31 | 1F | Unit separator | 63 | 3F | ? | 95 | 5F | _ | 127 | 7F | □ |

Formula for getting PHRED quality from encoded quality:

**Q = ascii(char) - 33**

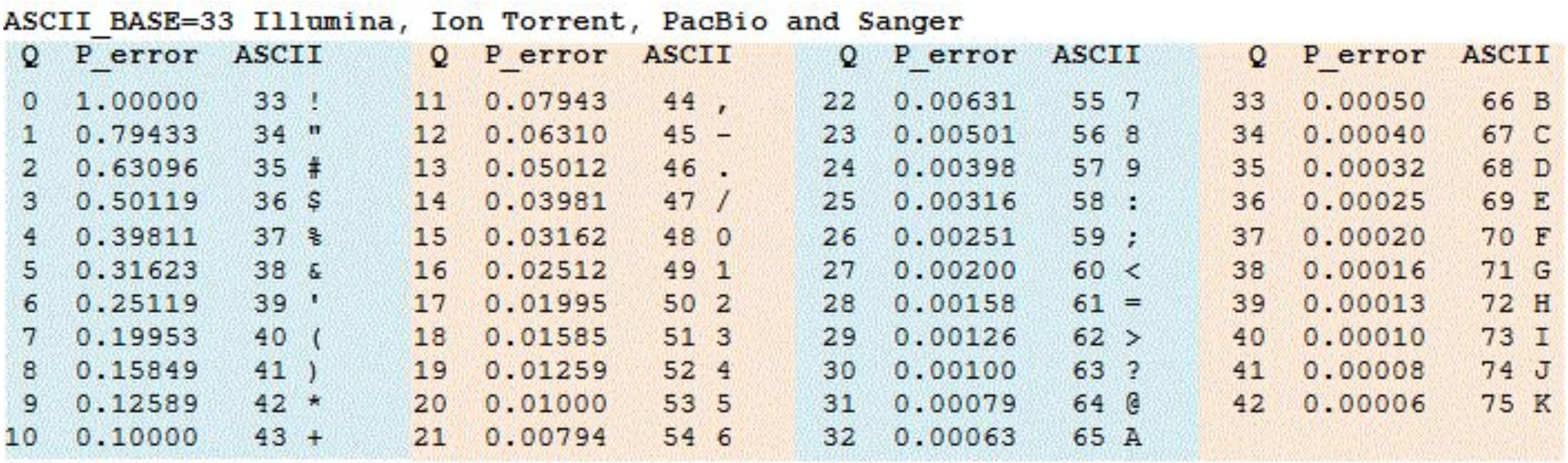
Example:



!+EJ

| ASCII | 33 | 43 | 69 | 74 |
|---|---|---|---|---|
| -33 | 0 | 10 | 36 | 41 |

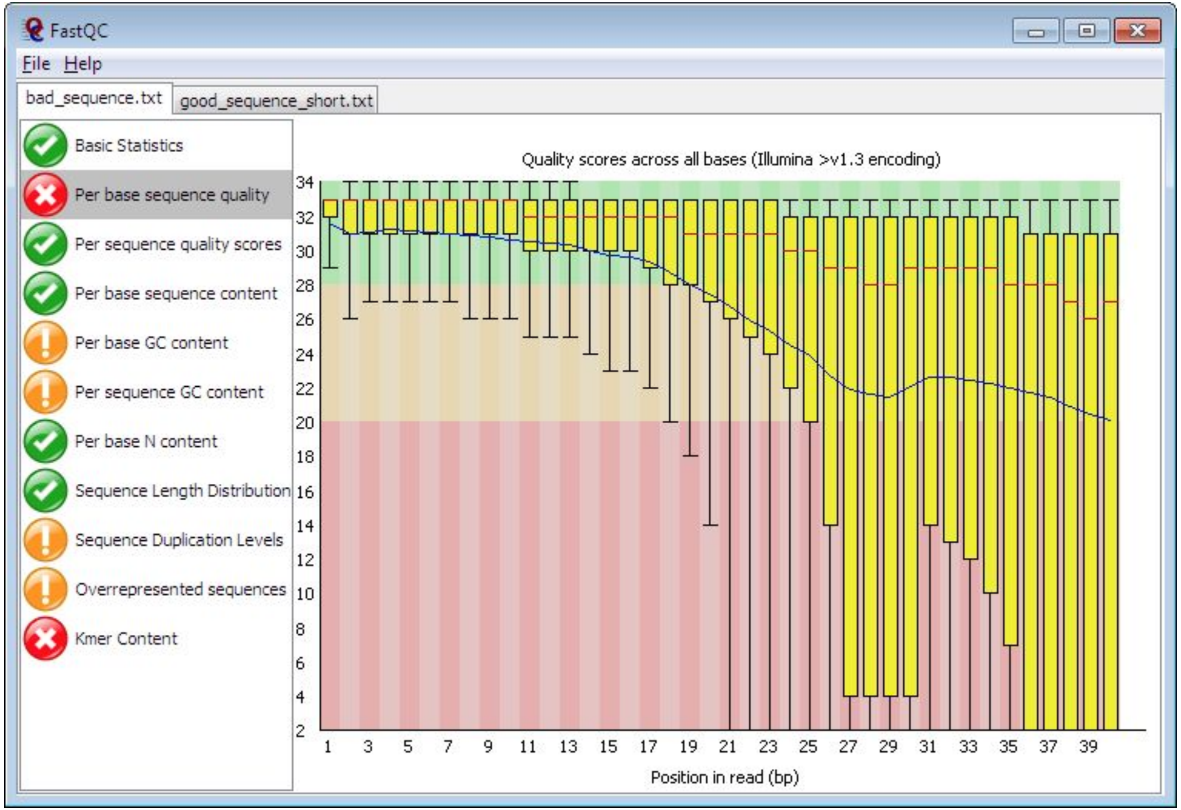
# Quality score encoding based on ASCII table. Geekery.



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

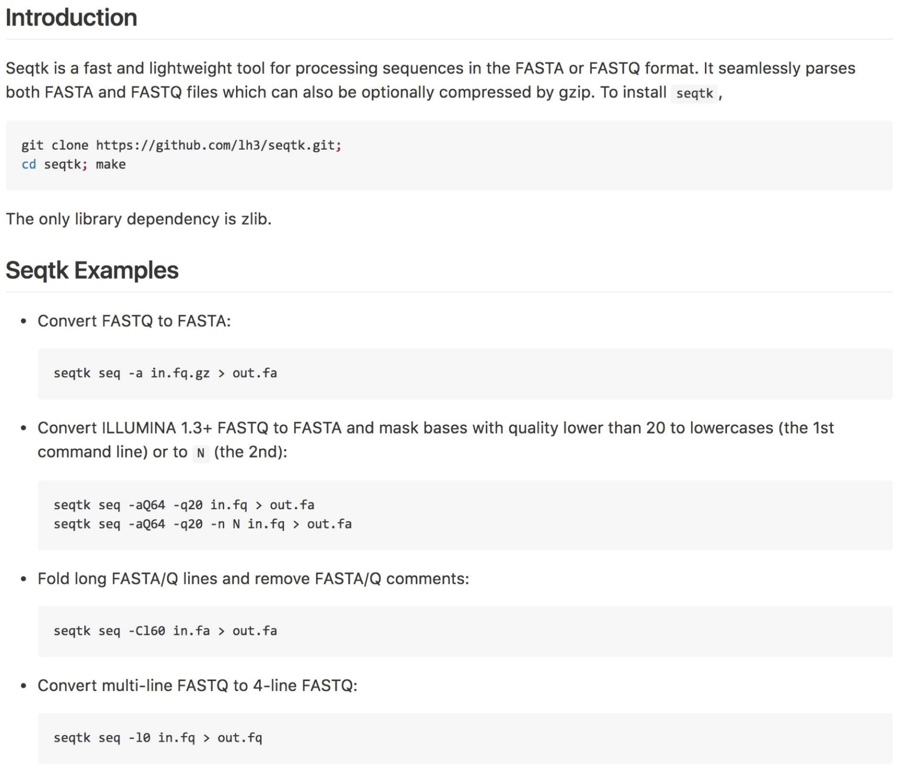
| Q | P_error | ASCII | Q | P_error | ASCII | Q | P_error | ASCII | Q | P_error | ASCII |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.00000 | 33 ! | 11 | 0.07943 | 44 , | 22 | 0.00631 | 55 7 | 33 | 0.00050 | 66 B |
| 1 | 0.79433 | 34 " | 12 | 0.06310 | 45 - | 23 | 0.00501 | 56 8 | 34 | 0.00040 | 67 C |
| 2 | 0.63096 | 35 # | 13 | 0.05012 | 46 . | 24 | 0.00398 | 57 9 | 35 | 0.00032 | 68 D |
| 3 | 0.50119 | 36 $ | 14 | 0.03981 | 47 / | 25 | 0.00316 | 58 : | 36 | 0.00025 | 69 E |
| 4 | 0.39811 | 37 % | 15 | 0.03162 | 48 0 | 26 | 0.00251 | 59 ; | 37 | 0.00020 | 70 F |
| 5 | 0.31623 | 38 & | 16 | 0.02512 | 49 1 | 27 | 0.00200 | 60 < | 38 | 0.00016 | 71 G |
| 6 | 0.25119 | 39 ' | 17 | 0.01995 | 50 2 | 28 | 0.00158 | 61 = | 39 | 0.00013 | 72 H |
| 7 | 0.19953 | 40 ( | 18 | 0.01585 | 51 3 | 29 | 0.00126 | 62 > | 40 | 0.00010 | 73 I |
| 8 | 0.15849 | 41 ) | 19 | 0.01259 | 52 4 | 30 | 0.00100 | 63 ? | 41 | 0.00008 | 74 J |
| 9 | 0.12589 | 42 * | 20 | 0.01000 | 53 5 | 31 | 0.00079 | 64 @ | 42 | 0.00006 | 75 K |
| 10 | 0.10000 | 43 + | 21 | 0.00794 | 54 6 | 32 | 0.00063 | 65 A | | | |

# FASTQC: Is my sequence data any good?


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34 32 30 28 26 24 22 20 18 16 14 12 10 8 6 4 2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

# seqtk: manipulating FASTQ and FASTA files



## Introduction

Seqtk is a fast and lightweight tool for processing sequences in the FASTA or FASTQ format. It seamlessly parses both FASTA and FASTQ files which can also be optionally compressed by gzip. To install `seqtk`,

```
git clone https://github.com/lh3/seqtk.git;
cd seqtk; make
```

The only library dependency is zlib.

## Seqtk Examples

- Convert FASTQ to FASTA:

  ```
  seqtk seq -a in.fq.gz > out.fa
  ```

- Convert ILLUMINA 1.3+ FASTQ to FASTA and mask bases with quality lower than 20 to lowercases (the 1st command line) or to `N` (the 2nd):

  ```
  seqtk seq -aQ64 -q20 in.fq > out.fa
  seqtk seq -aQ64 -q20 -n N in.fq > out.fa
  ```

- Fold long FASTA/Q lines and remove FASTA/Q comments:

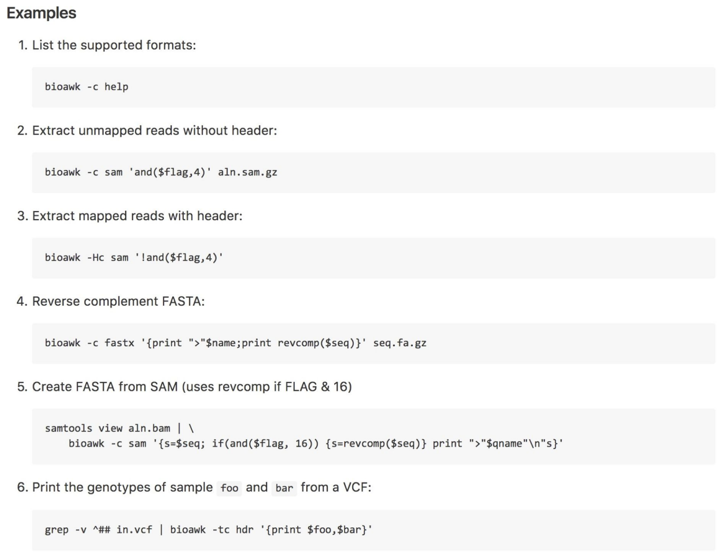
  ```
  seqtk seq -Cl60 in.fa > out.fa
  ```

- Convert multi-line FASTQ to 4-line FASTQ:

  ```
  seqtk seq -l0 in.fq > out.fq
  ```

https://github.com/lh3/seqtk

# bioawk: awk that is enhanced for genomics formats.



### Examples

1. List the supported formats:

```
bioawk -c help
```

2. Extract unmapped reads without header:

```
bioawk -c sam 'and($flag,4)' aln.sam.gz
```

3. Extract mapped reads with header:

```
bioawk -Hc sam '!and($flag,4)'
```

4. Reverse complement FASTA:

```
bioawk -c fastx '{print ">"$name;print revcomp($seq)}' seq.fa.gz
```

5. Create FASTA from SAM (uses revcomp if FLAG & 16)

```
samtools view aln.bam | \
    bioawk -c sam '{s=$seq; if(and($flag, 16)) {s=revcomp($seq)} print ">"$qname"\n"s}'
```

6. Print the genotypes of sample `foo` and `bar` from a VCF:

```
grep -v ^## in.vcf | bioawk -tc hdr '{print $foo,$bar}'
```

Print tab separated sequence ID and sequence from a FASTQ file

```
$ bioawk -c fastx '{print $name"\t"$seq}' test.fastq
SRR3750603.1 NTCGGAACATTTTTTCTTCAAAAATATGAAAAATCACCTAATTTATCTGAAAATGACATTTANNNCAGTNNNNNNNATTGGGAAAGTGCTCGATTTNCGGA
SRR3750603.2 TGTAATTTACTTTGTTCAGTTAGACTCTTAATTAGACTAAAAACGGTCTCAAAAAGTATAATTTCATAATGAGACACCTTTAAAAATTCTACGTTTTTATG
SRR3750603.3 AGTTTTCTCAAACACAGAAAACATATGGGAGTTTCTCAAACAATGGACAATGAGTGATCACCGATATTTGATACAAATCGACCAACTCGGCTCATATTCTC
```

https://github.com/lh3/bioawk