Understanding & Manipulating Big Genomic Data - Getting Started with FASTQ files -

Biocomputing Bootcamp
Day 3 – Session 1
Instructor : Hyun Min Kang

What is FASTQ?

- A text file format for storing
  - Nucleotide sequences (A, C, G, T)
  - Their quality scores
- Developed by Sanger Institute
- Widely used after high-throughput sequencing technology

What does a FASTQ format look like?

```
@HWI-D00196:189:C6WU5ACXX:8:1215:17539:66708 1:N:0:
TGCTTTGGGCAGTGTCCTGACTGTAAGATCAAGTCCAAACCTGTTTTGGAA
+
@@@FFBDEHH?ACGFGE?HHIIFEHIGEEE@EH?>GA?HHGHFFGEHC6##
@HWI-D00196:189:C6WU5ACXX:8:1215:17690:66714 1:N:0:
TAGTGTGGGCCGGCGGCGGCGCCCCACGAGGCGGTGCCGAGTTCGGTCCCA
+
CCCFFFDDDHFGGIJIGIFDDDDDD?BB6B8BDD7<BDBDD@BDDDDDDDD
@HWI-D00196:189:C6WU5ACXX:8:1215:17723:66717 1:N:0:
TAGATGGGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCAGATCATCT
+
@@CBDBDACFFFHHGIGCHI@FFFGBDGGGHGDFHC>DGHJJIIDFFIJIJ
```

(4N+1)-th line : (Unique) Read Name

```
@HWI-D00196:189:C6WU5ACXX:8:1215:17539:66708 1:N:0:
TGCTTTGGGCAGTGTCCTGACTGTAAGATCAAGTCCAAACCTGTTTTGGAA
+
@@@FFBDEHH?ACGFGE?HHIIFEHIGEEE@EH?>GA?HHGHFFGEHC6##
@HWI-D00196:189:C6WU5ACXX:8:1215:17690:66714 1:N:0:
TAGTGTGGGCCGGCGGCGGCGCCCCACGAGGCGGTGCCGAGTTCGGTCCCA
+
CCCFFFDDDHFGGIJIGIFDDDDDD?BB6B8BDD7<BDBDD@BDDDDDDDD
@HWI-D00196:189:C6WU5ACXX:8:1215:17723:66717 1:N:0:
TAGATGGGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCAGATCATCT
+
@@CBDBDACFFFHHGIGCHI@FFFGBDGGGHGDFHC>DGHJJIIDFFIJIJ
```

(4N+2)-th line : Sequence Reads

```
@HWI-D00196:189:C6WU5ACXX:8:1215:17539:66708 1:N:0:
TGCTTTGGGCAGTGTCCTGACTGTAAGATCAAGTCCAAACCTGTTTTGGAA
+
@@@FFBDEHH?ACGFGE?HHIIFEHIGEEE@EH?>GA?HHGHFFGEHC6##
@HWI-D00196:189:C6WU5ACXX:8:1215:17690:66714 1:N:0:
TAGTGTGGGCCGGCGGCGGCGCCCCACGAGGCGGTGCCGAGTTCGGTCCCA
+
CCCFFFDDDHFGGIJIGIFDDDDDD?BB6B8BDD7<BDBDD@BDDDDDDDD
@HWI-D00196:189:C6WU5ACXX:8:1215:17723:66717 1:N:0:
TAGATGGGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCAGATCATCT
+
@@CBDBDACFFFHHGIGCHI@FFFGBDGGGHGDFHC>DGHJJIIDFFIJIJ
```

(4N+4)-th line : Quality Scores

```
@HWI-D00196:189:C6WU5ACXX:8:1215:17539:66708 1:N:0:
TGCTTTGGGCAGTGTCCTGACTGTAAGATCAAGTCCAAACCTGTTTTGGAA
+
@@@FFBDEHH?ACGFGE?HHIIFEHIGEEE@EH?>GA?HHGHFFGEHC6##
@HWI-D00196:189:C6WU5ACXX:8:1215:17690:66714 1:N:0:
TAGTGTGGGCCGGCGGCGGCGCCCACGAGGCGGTGCCGAGTTCGGTCCCA
+
CCCFFFDDDHFGGIJIGIFDDDDDD?BB6B8BDD7<BDBDD@BDDDDDDDD
@HWI-D00196:189:C6WU5ACXX:8:1215:17723:66717 1:N:0:
TAGATGGGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACCAGATCATCT
+
@@CBDBDACFFFHHGIGCHI@FFFGBDGGGHGDFHC>DGHJJIIDFFIJIJ
```

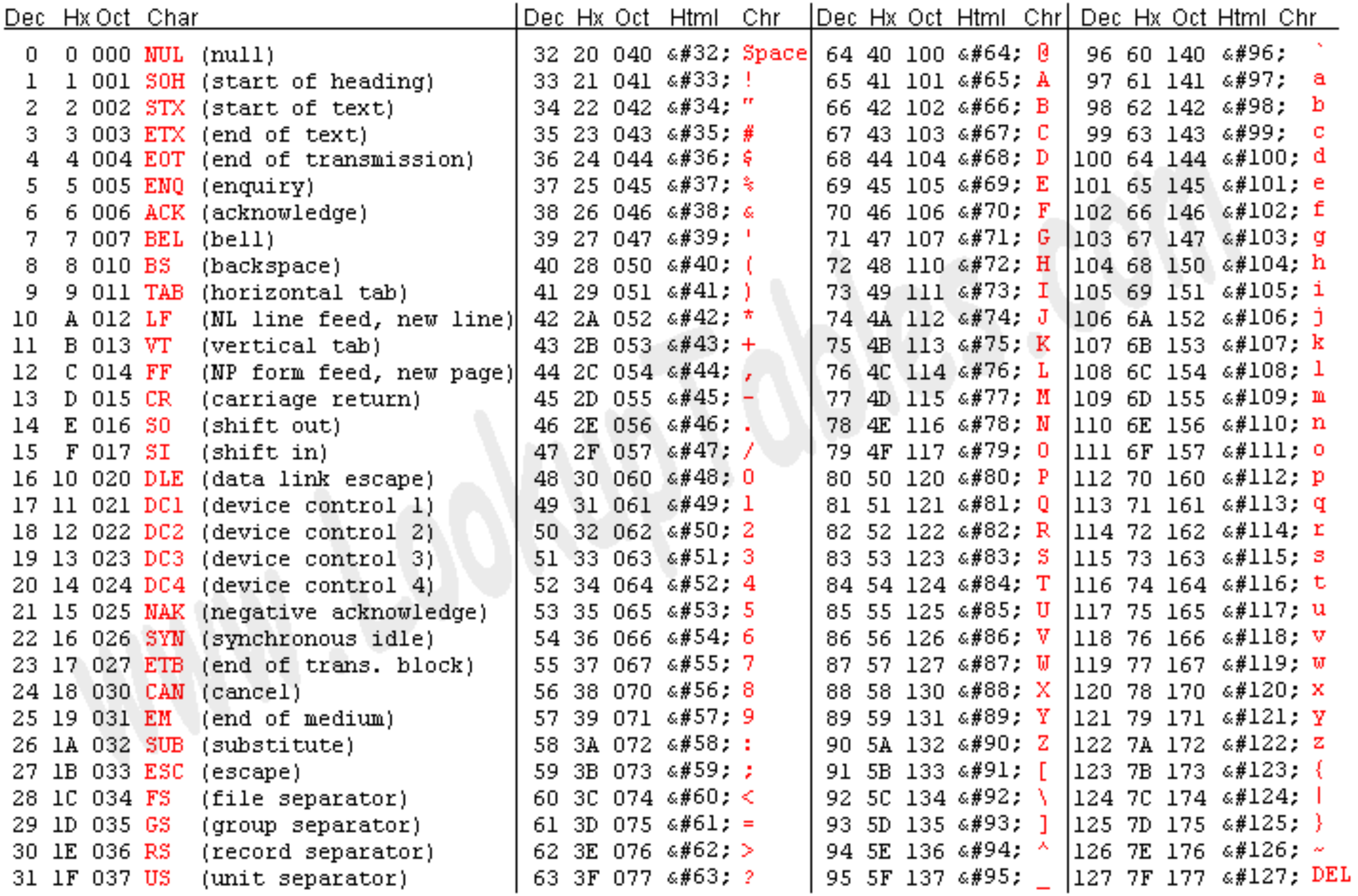
Quality Scores in FASTQ

```
@@@FFBDEHH?ACGFGE?HHIIFEHIGEEE@EH?>GA?HHGHFFGEHC6##
```

- Each character represent an integer
  - as [ASCII code of the character] – 33
  - Not human-friendly, but storage-friendly
    (requires one character rather than two characters)
- The integer represents the estimated error of sequence read
  - as translated by the equation: $\Pr(e|Q) = 10^{-\frac{Q}{10}}$

Character Quality	Integer Quality	Pr(error)	Pr(correct)
I	40	10^{-4} = 0.01%	99.99%
?	30	10^{-3} = 0.1%	99.9%
5	20	10^{-2} = 1%	99%
+	10	10^{-1} = 10%	90%
#	2	$10^{-0.2}$ = 63%	37%

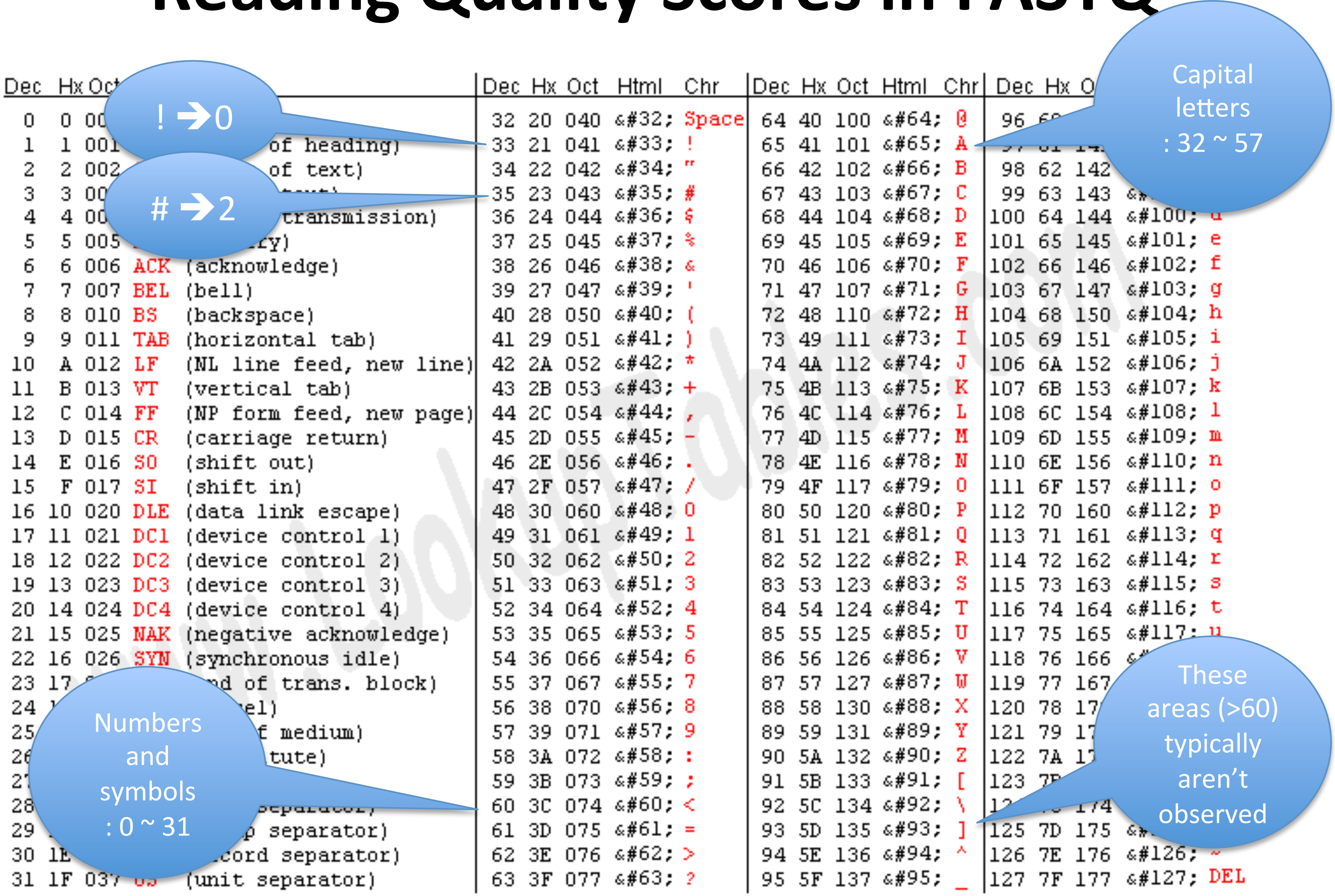
Reading Quality Scores in FASTQ



Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	60	140	&#96;	`
1	1	001	SOH (start of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	97	61	141	&#97;	a
2	2	002	STX (start of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	&#98;	b
3	3	003	ETX (end of text)	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#99;	c
4	4	004	EOT (end of transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	d
5	5	005	ENQ (enquiry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK (acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL (bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS (backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB (horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF (NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT (vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF (NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR (carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO (shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI (shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE (data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1 (device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2 (device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3 (device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4 (device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK (negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN (synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#118;	v
23	17	027	ETB (end of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	&#119;	w
24	18	030	CAN (cancel)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	170	&#120;	x
25	19	031	EM (end of medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	171	&#121;	y
26	1A	032	SUB (substitute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	172	&#122;	z
27	1B	033	ESC (escape)	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	173	&#123;	{
28	1C	034	FS (file separator)	60	3C	074	&#60;	<	92	5C	134	&#92;	\	124	7C	174	&#124;	|
29	1D	035	GS (group separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#125;	}
30	1E	036	RS (record separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US (unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

Reading Quality Scores in FASTQ



Dec	Hx	Oct			Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	O		
0	0	00	[illegible]	[illegible]	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	6	[illegible]	[illegible]	[illegible]
1	1	001	[illegible]	of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	[illegible]	[illegible]	14	[illegible]	[illegible]
2	2	002	[illegible]	of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	[illegible]	[illegible]
3	3	00	[illegible]	[illegible]	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#	[illegible]
4	4	00	[illegible]	transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	[illegible]
5	5	005	[illegible]	ry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK	(acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL	(bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS	(backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB	(horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF	(NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT	(vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF	(NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR	(carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO	(shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI	(shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE	(data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1	(device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2	(device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3	(device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4	(device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK	(negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN	(synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#	[illegible]
23	17	[illegible]	[illegible]	nd of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	[illegible]	[illegible]
24	[illegible]	[illegible]	[illegible]	el)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	17	[illegible]	[illegible]
25	[illegible]	[illegible]	[illegible]	f medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	17	[illegible]	[illegible]
26	[illegible]	[illegible]	[illegible]	tute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	17	[illegible]	[illegible]
27	[illegible]	[illegible]	[illegible]	[illegible]	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	[illegible]	[illegible]	[illegible]
28	[illegible]	[illegible]	[illegible]	[illegible]	60	3C	074	&#60;	<	92	5C	134	&#92;	\	12	[illegible]	174	[illegible]	[illegible]
29	[illegible]	[illegible]	[illegible]	p separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#	[illegible]
30	1E	[illegible]	[illegible]	cord separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US	(unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

! → 0

→ 2

Capital letters : 32 ~ 57

Numbers and symbols : 0 ~ 31

These areas (>60) typically aren't observed

Reading Quality Scores in FASTQ

```
@@@FFBDEHH?ACGFGE?HHIIFEHIGEEE@EH?>GA?HHGHFFGEHC6##
```

High qualities > 30

Low quality = 2

- If you become familiar with FASTQ format,
  you may be able to interpret the quality scores above as follows..

  "most of sequences are of high quality (>30),
  except for a few nucleotides at the end"

Paired FASTQ files

- Often a sequence read has one or more mates…
  - when both ends of a DNA fragment is sequenced.
  - when the multiple samples are barcoded and pooled into a single sequencing lane.

- The paired FASTQ files
  - have exactly the same number of lines (and reads)
  - the read name for each corresponding read is identical

Practice : Demultiplexing FASTQ files

- Given : You’re given a pair of FASTQ files
  - bioboot_2015a_R1.fastq.gz, bioboot_2015a_R2.fastq.gz
  - Read 1 is 51bp, Read 2 is 7bp

- This is a mixture of 5 samples, barcoded by Read 2
  - Sample1 : ACAGTGA
  - Sample2: CAGATCA
  - Sample3 : GCCAATA
  - Sample4: TGACCAA
  - Sample5 : TTAGGCA

- Want : Split the first FASTQ files into six parts
  - (1)-(5) for each sample, (6) for UNKNOWN classification