

**Sequence Formats:**

**FASTA, FASTQ**

**PHRED Scores coded as ASCII Characters**

# FASTA Sequence format

White Space (Space or Tab)

> Sequence\_Name\_1 Sequence Annotation ... ..

MDSKGSSQKGSRLLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLS  
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNPPLYHL  
VTEVRGMKGAPDAILSR AIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED  
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC

> Sequence\_Name\_2 Sequence Annotation ... ..

ADQLTEEQIAEFKEAFSLFDKDGDTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID  
FPEFLTMMARKMKD TDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA  
DIDGDGQVNYEEFVQMMTAK

> Sequence\_Name\_3 Sequence Annotation ... ..

LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPY  
IGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIP  
FHPYYTIKDFLGLLILILLLLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRS  
VPNKLGGVLALFLSIVILYGLMPFLHTSKHRSMMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYP  
YTIIGQMASILYFSIILAFLPIAGXIENY

# FASTA Sequence format

> Sequence\_Name\_1 Sequence Annotation ... ..

MDSKGSSQKGSRLLLLVSNNLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLS  
EMFNEFDKRYAQGKGFI TMAINSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNPYHL  
VTEVRGMKGAPDAILSRATIEIEEENKRLLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED  
ARYSAFYNNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC

> Sequence\_Name\_2 Sequence Annotation ... ..

ADQLTEEQIAEFKEAFSLFDKKGDTETTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID  
FPEFLTMMARKMKD TDSEEEI REAFRVFDKENGYSAAELRHVMTNLGEKLTDEEVDEMIREA  
DIDGDGQVNYEEFVQMMTAK

> Sequence\_Name\_3 Sequence Annotation ... ..

LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPY  
IGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIP  
FHPYYTIKDFLGLLILILLLLLLLLALLSPDMLGDPDNHMPADPLNTEFLHIKPEWYFLFAYAILRS  
VPNKLGGVLALFLSIVILYGLMPFLHTSKHRSMLRPLSQALFWTLTMDLLTLTWIGSQPVEYP  
YTIIGQMASILYFSIILAFLPIAGXIENY

# FASTQ Sequence format

> **Sequence\_Name\_1** **Sequence Annotation** ... ..

```
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGC
AGGGCATGCAGGGAATTCGACGTTTCAGGCGATTTACTTCGGCATGCATTGCGGCATTATATCGA
CTGTAGCTGACTGACTGGAGTATTAGCGCGAAAGGTCTATTTATTCTGGAGGGGCAGGTGTTCC
ACGGCATATCGGGACTATCTAACCCTCCTAGAAGTTC
```

**FASTA Format** is designed to store **DNA, RNA, Protein** Sequences with minimal **Annotation**.

Allowing **DNA** & **Protein** **Ambiguity Codes**, can make it difficult to ascertain **Sequence Type**.

**Sequence Type** is not specified in the **FASTA** format. Distinction is left to the software.

**FASTQ Format** is an adaptation designed to store **Sequencing Reads** with minimal **Annotation**.

Each sequence in a **FASTQ** file is always the that of a single **Sequencing Read**. Thus it can only ever be **DNA** sequence.

# FASTQ Sequence format

```
@ Sequence_Name_1 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGGCATCGAG ...
```

```
+ MORE Sequence Annotation ... ..
```

The transition from FASTA to FASTQ Format includes a couple of cosmetic “enhancements”:

The initial “>” becomes an “@”

An extra **Annotation** line beginning with a “+” is added.

The opportunity to include **Annotation** in this line is rarely used.

Currently, as the length of an individual read is modest, sequences occupy just a single line.

# FASTQ Sequence format

```
@ Sequence_Name_1 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGGCATCGAG ...  
+  
! ' ' * ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) * * 5 5 C C F > > > > > C C C C C C C 6 5 5 6 7 6 C C G L N 6 9 5 K P 4 ...
```

The **Q** in **FASTQ** stands for **Quality**. A definitive purpose of the **FASTQ** Format is to record the **Base Call Quality** of each element of a **Sequencing Read**.

**Base Call Qualities** are recorded in a fourth line of the **FASTQ** file. Each **Base Call Quality** is recorded as a single printable character. Each **Quality** character corresponds to one **Called Base**.

# FASTQ Sequence format

@ Sequence\_Name\_1 Sequence Annotation ... ..

CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGCA  
TCGAGCGATCAGCTGCATGACTGCATCAGCTGCGCATATGCGATCGGATTCGAGGGCTATGCATGCATGA  
TTTCAGCATCATTCAGAGCGGACTTTACGCATCGAGCTTACACACATGCATGCATTGGTGACCTGACTGG  
CGCGGCGGCGGCATTGGACTTACTGCATCGATGCATCAGTTGTGATCCAGCATGCATGCATGCGATCGA

+

Spurious Start of new read

! ' ' \* ( ( ( \* \* \* + ) ) % % % + + ) ( % % % ) . 1 \* \* \* - + \* ' ' ) ) \* \* 5 5 C C F > > > > > C C C C C C C 6 5 5 6 7 6 C C G L N 6  
@ 5 K P 4 5 5 C C F > > % % % % + + ) 5 6 7 6 C C G L ' \* ( ( ( \* \* \* + ) 1 1 8 4 7 7 G A A A D V V + + B B D A G G H 7 G G C 9 % % % )  
A C G G T A ( H G 4 8 \* \* ( @ @ H I A 9 0 % % + + ) ( % > > > > > C C J J A S ! ! ! ! ) ( ^ + + K ' ' ` ` ' ' ' \* ( ' ' \* ( L K \* & T T  
+ F 9 R S A ; ; : : A C V 6 ^ 7 & & & 3 £ 2 2 " ( 1 ! D F G H ( ( ( ( < < < < % + + ) ( % % % ) . 1 \* H I A 9 0 % % + + ) ( % > > >

Spurious Annotation line

Spurious Read Sequence

Most printable characters, *including* “@” and “+” are allowable **Base Call Quality** codes.

**Multiple line sequences** could introduce **Base Call Quality** lines beginning with “@” or “+”.

Such lines could be misinterpreted as: “@” - The start of a new **Sequencing Read**,

“+” - An additional **Annotation** line.

# FASTQ Sequence format

```
@ Sequence_Name_1 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGGCATCGAG ...  
+  
! ' ' * ( ( ( ***+ ) % % % ++ ) ( % % % ) . 1 *** - + * ' ' ) ) **55CCF>>>>>CCCCCCC655676CCGLN695KP4 ...  
  
@ Sequence_Name_2 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGGCATCGAG ...  
+  
( ** - + ( ( ***+ ) ) % >> CCCC ! CCC655676CCGL ! ' ' * ( N695KP4 % % ++ ) ( % % % ) . 1 * * ' ' ) ) **55CCF>>>> ...  
  
@ Sequence_Name_3 Sequence Annotation ... ..  
CATGAGCGTCAGCATGCATGGTCGGAGCAGGGGCATCGAGTCTTAATTCTCAGGGACCCGCTAGCTGGGGTATCAT ...  
+  
*** - + * ' ' ) ) **55CCF>>FF>> ! ' ' * ( ( ( ***+ ) ) % 1 CCACC % % ++ ) ( % % % ) . CCC655676CCGLN695KP4 ...
```

To avoid this, even longer **Read Sequences** are confined to one line.



# Base Quality ( **PHRED** ) Scores

The **Base Quality Scores** in **FASTQ** files are **PHRED Scores**.

The **PHRED Score** for a given **Base Call** is derived from an estimate of the **Probability** of that **Base** being **Called** incorrectly

Where **Q** is the **PHRED Score** for a given **Base Call** and **P** is the estimated **Probability of Error** for that **Base Call**, the following formula applies:

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

**P** is a **Probability** and so has **Range**  $0 \rightarrow 1$

**Q** is a function of **P** that has **Range**  $0 \rightarrow \text{infinity}$ .

# Base Quality ( **PHRED** ) Scores

Base Call Error Rate	Probability of Incorrect Call (P)	$\text{Log}_{10}(P)$	PHRED Score $Q = -10\text{Log}_{10}(P)$
1 in 1	1	0	00
1 in 10	0.1	-1	10
1 in 100	0.01	-2	20
1 in 1000	0.001	-3	30
1 in 10,000	0.0001	-4	40
1 in 100,000	0.00001	-5	50
1 in 1,000,000	0.000001	-6	60
1 in 10,000,000	0.0000001	-7	70
...	...	...	...
0	0.00	-infinity	infinity

**Q** is a function of **P** that has **Range** **0 → infinity**. and easily represents a useful subset of **P Values**, to an adequate accuracy, as a **Two Digit Integer**.

# Base Quality ( **PHRED** ) Scores

Base Call Error Rate	PHRED Score $Q = -10\log_{10}(P)$
1 in 1	00
1 in 10	10
1 in 100	20
1 in 1000	30
1 in 10,000	40
1 in 100,000	50
1 in 1,000,000	60
1 in 10,000,000	70
...	...
0	infinity

**PHRED Scores** can be intuitively thought of as directly representing various **Error Rates**.

An **Error Rate** worse than **1 in 100** (**PHRED = 20**) is usually considered a disaster!

Anything better than **1 in 10,000** (**PHRED = 40**) is usually considered as near perfect as makes no difference.

Achieving a **Consensus PHRED Score** of **30** (**1 in 1000**) is a common target for an **Assembly of Reads**.

# Base Quality ( PHRED ) Scores

```
@ Sequence_Name_1 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGGCATCGAG ...  
+  
! ' ' * ( ( ( ( * * * + ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) * * 5 5 C C F > > > > > C C C C C C C 6 5 5 6 7 6 C C G L N 6 9 5 K P 4 ...
```

## The Problem

Ideally at least, each base in a **FASTQ** file should correspond with one, *two digit* **PHRED Score**, represented as a **Single Character**.

# Base Quality ( PHRED ) Scores

@ Sequence\_Name\_1 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGCATCGAG ...  
+  
! ' ' \* ( ( ( ( \* \* \* + ) ) % % % + + ) ( % % % % ) . 1 \* \* \* - + \* ' ' ) \* \* 5 5 C C F > > > > > C C C C C C C 6 5 5 6 7 6 C C G L N 6 9 5 K P 4 ...

## The Solution

Is to map the *two digit* Phred Scores onto a single element of the **ASCII Character Set**.

# Base Quality ( **PHRED** ) Scores

The ASCII Character Codes are Stable and Universally Accepted.

They provide a mapping of the integers **000 → 255** to a set of **Individual Characters**.

Not all of these characters are visibly printable. Specifically, **00 → 31** are not printable.

**32** is printable, but not visible (it is a **Space**).

**33 → 126** are all printable however, and could be used to represent the **PHRED Scores 00 → 93** nicely. More than sufficient for practical purposes.

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	ì	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	í	225	·
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	,
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
4	<EOT>	36	\$	68	D	100	d	132	Ñ	164	§	196	ƒ	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Ö	165	•	197	≈	229	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	Δ	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	...	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	'	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	¨	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Ö	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	-	240	🍏
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	"	242	Ú
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	"	243	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	`	244	Ü
21	<NAK>	53	5	85	U	117	u	149	ï	181	μ	213	'	245	ı
22	<SYN>	54	6	86	V	118	v	150	ñ	182	ð	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	Σ	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	Π	216	ÿ	248	—
25	<EM>	57	9	89	Y	121	y	153	ô	185	π	217	ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	ƒ	218	/	250	·
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	°
28	<FS>	60	<	92	\	124		156	ú	188	º	220	<	252	¸
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	>	253	”
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	fi	254	˘
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	fi	255	˘

# Base Quality ( **PHRED** ) Scores

```
@ Sequence_Name_1 Sequence Annotation ... ..  
CCGCTAGCTGGGGTATCATCAGCATGCATGGCATGAGCGTTCTTAATTCTCAGGGACTCGGAGCAGGGGCATCGAG ...  
+  
! ' ' * ( ( ( ***+ ) %%%++ ) ( %%% ) . 1 ***-+* ' ' ) **55CCF>>>>>CCCCCCC655676CCGLN695KP4 ...
```

ASCII Codes →

42

70

62

53

78

PHRED Scores→

10

38

30

21

46

Thus it is **ASCII Codes** that represent the **Base Call Qualities (PHRED Scores)** in **FASTQ** files.

To compute the **PHRED Score** from the **ASCII Character**, simply look up the **ASCII Code** and **SUBTRACT 32**.

# Base Quality ( **PHRED** ) Scores

## A Warning

Long ago ... stupid people at the Sanger Centre coded **PHRED Scores** starting from **ASCII Code 64 (@)**.

Should you across any of this older data that has not been updated to reflect the current standards, you will need to instruct the software to subtract **63** instead of **32** in order to compute the **PHRED Score** from the **ASCII Code**.

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	ì	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	í	225	·
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	,
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
4	<EOT>	36	\$	68	D	100	d	132	Ñ	164	§	196	f	228	‰
5	<ENQ>	37	%	69	E	101	e	133	Ö	165	•	197	≈	229	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	Δ	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	...	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	'	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	..	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Õ	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	-	240	🍏
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
18	<DC2>	50	2	82	R	114	r	146	í	178	≤	210	"	242	Ú
19	<DC3>	51	3	83	S	115	s	147	ì	179	≥	211	"	243	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	`	244	Ü
21	<NAK>	53	5	85	U	117	u	149	ï	181	μ	213	'	245	ı
22	<SYN>	54	6	86	V	118	v	150	ñ	182	ð	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	Σ	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	Π	216	ÿ	248	—
25	<EM>	57	9	89	Y	121	y	153	ô	185	π	217	ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	ƒ	218	/	250	˙
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	˚
28	<FS>	60	<	92	\	124		156	ú	188	º	220	<	252	¸
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	>	253	”
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	fi	254	˘
31	<US>	63	?	95	_	127	<DEL>	159	ü	191	ø	223	fi	255	˘



TO DO:

Add Notes

Make video