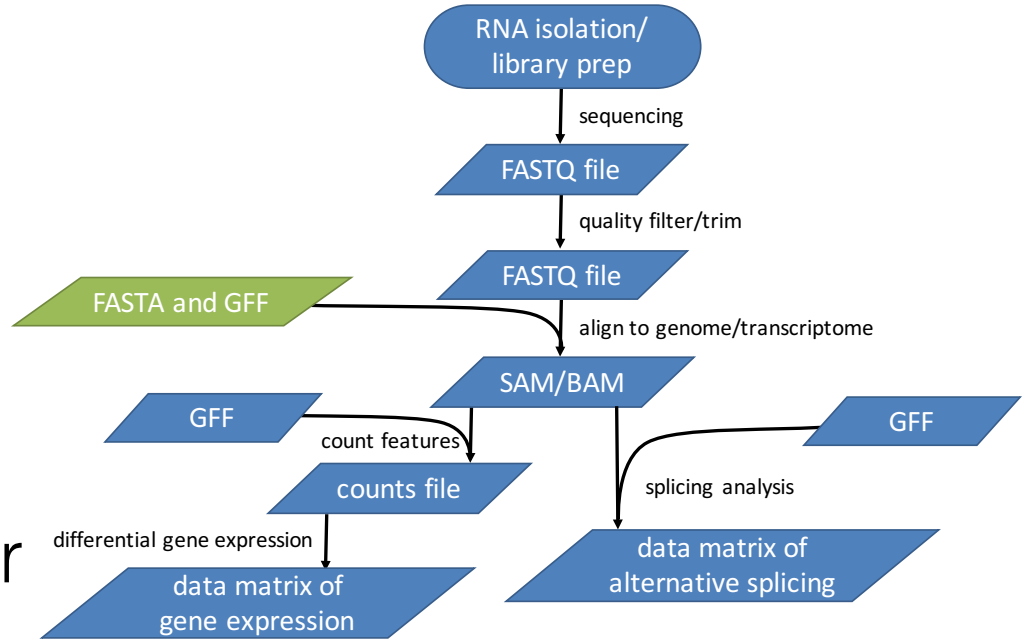
Common file formats, Part 1

FASTA/FASTQ

The FASTA file format

- Text-based format for representing
  - nucleotide sequence OR
  - peptide sequence (single letter codes)
- Each sequence in the file begins with a unique header line denoted with a “>”



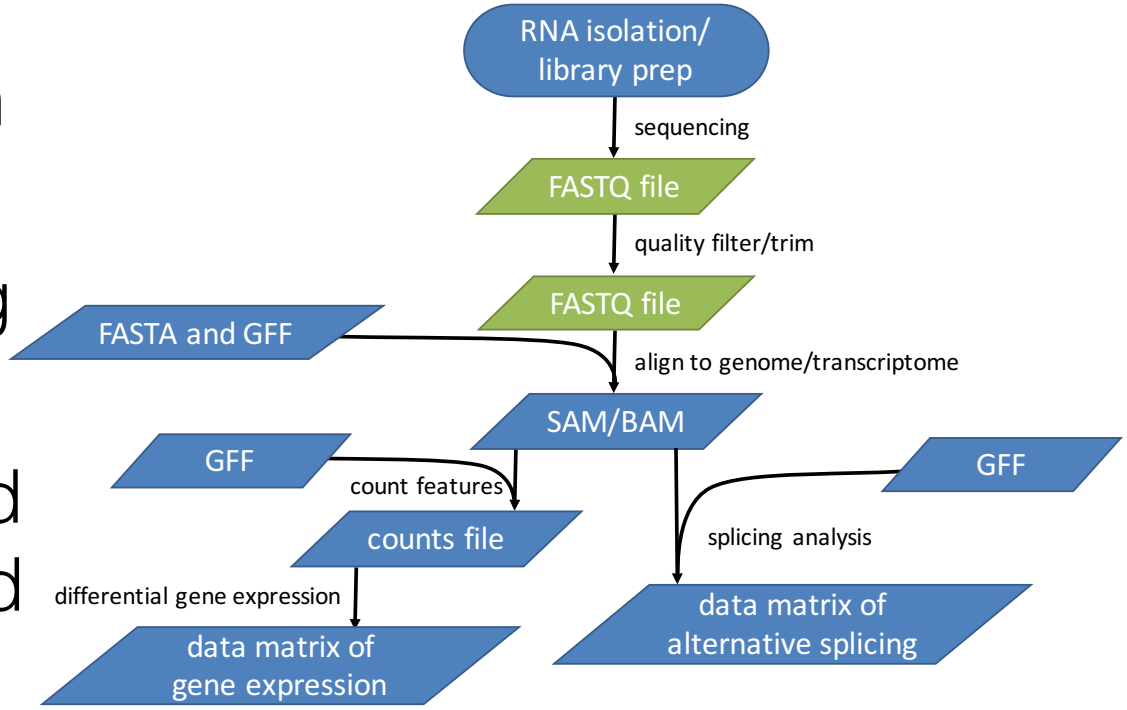
- The following line(s) are are sequence, typically between 60 and 130 characters in length (may include spaces and/or numbers)

What does a FASTA file look like?

```
>NODE_1140748_length_208_cov_4.298077
GTATATTAGAAGGGGCCGCGCGGATGAGATGGGTGACAGTACACTTTCCATGCAAGAACG
GGCGGGTTTGTAATATTCCTTAAATTATTGTCAGAAACTCTGTGATGGAGACATTGACCT
CAGTTAATTAGTCTGGGTATTGCTGTCTGATTTAATAAGACACCAAACCAGTCTCCCTGG
CGACAGGAAGTGCTGCATGCATGCCACTGCATTTGTTGCAGCGGTGTTCTAGAGCGGTCA
CCCTGAGGGATGCTGc
>NODE_1140775_length_818_cov_9.328851
TACGGAAAGAGGACACGCAGGCCGGTTCCTCTCCAGCAGTCCGGTGACTGCACACCGCAG
AAGGAACCGGACGCCACCTCGACCTTCCTTTCCTCTTCGCANNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAAACACATCATAAATCACTTAA
GTTAAGACTCCTCGGGGTTCACAGCTGTCTGATAACAGAGAAGACCGAGTGACCTCGCCT
CAGTTGGCAGTGGCCCCAAACTGATCACAGCAACAGCAAAGATCTGTAACAATGCCAACA
TATCGCCCCTTGGAGACCAGGAGAACGTGCTCCTTACAGGAGAACATGTTCCTTAGAGGA
GAACATGCTCCTTACAGGAGAACAGG
>NODE_1141553_length_172_cov_7.779070
ACAGGAAGTAGAACGATGCGAGGGAGGGGCTTCGGGGGACGCAGGGACCCTCCTCTCTTT
GGCCACCATTTTGGAAGTGACACGAGAGGGGGGAGCGTTTCTTTGTCAGCGACTAAAATG
TTAAAAAAAAACTTTCATCAAGTAAAGAACACAACGTGTTTTAAACTGGGTGGAGACCTG
TCAATCAGCGTGTAGCCCCGCCCCAAAGCATCTCCTGCTT
```

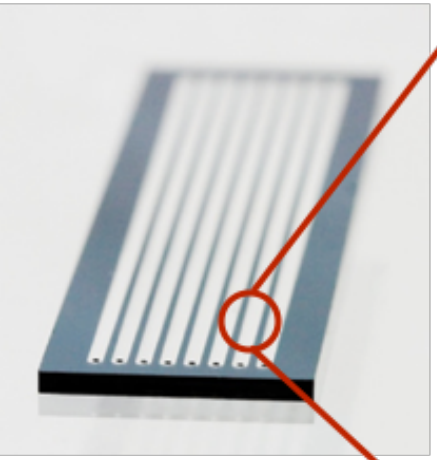
The FASTQ file format

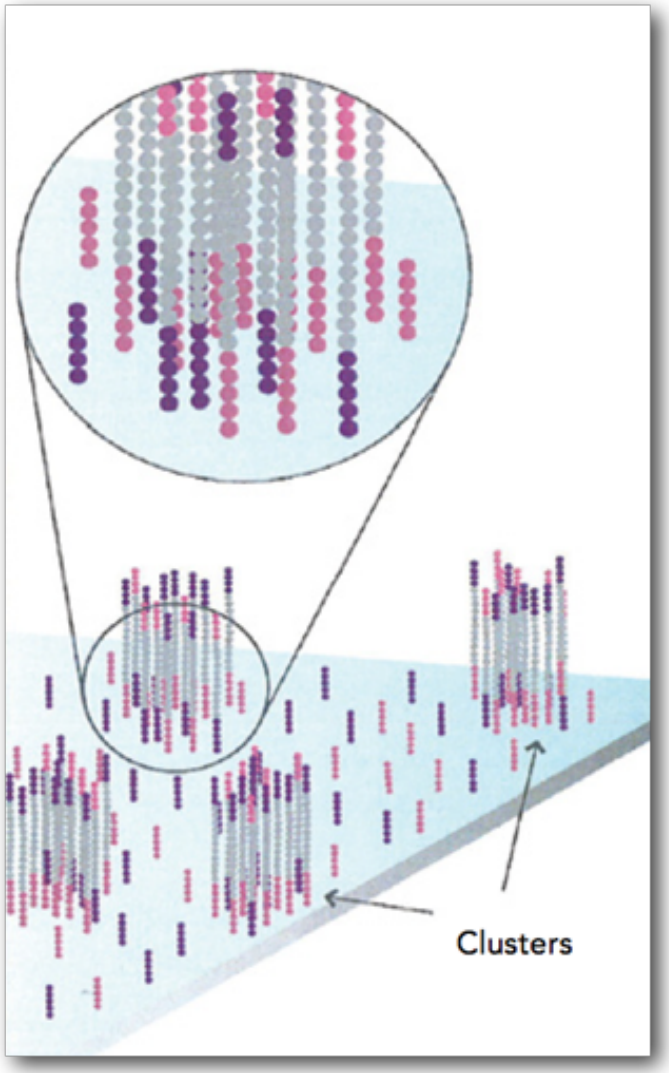
- FASTQ is a modified form of FASTA
- Goal: include sequencing quality scores
- Has become the standard for storing data generated by NGS data

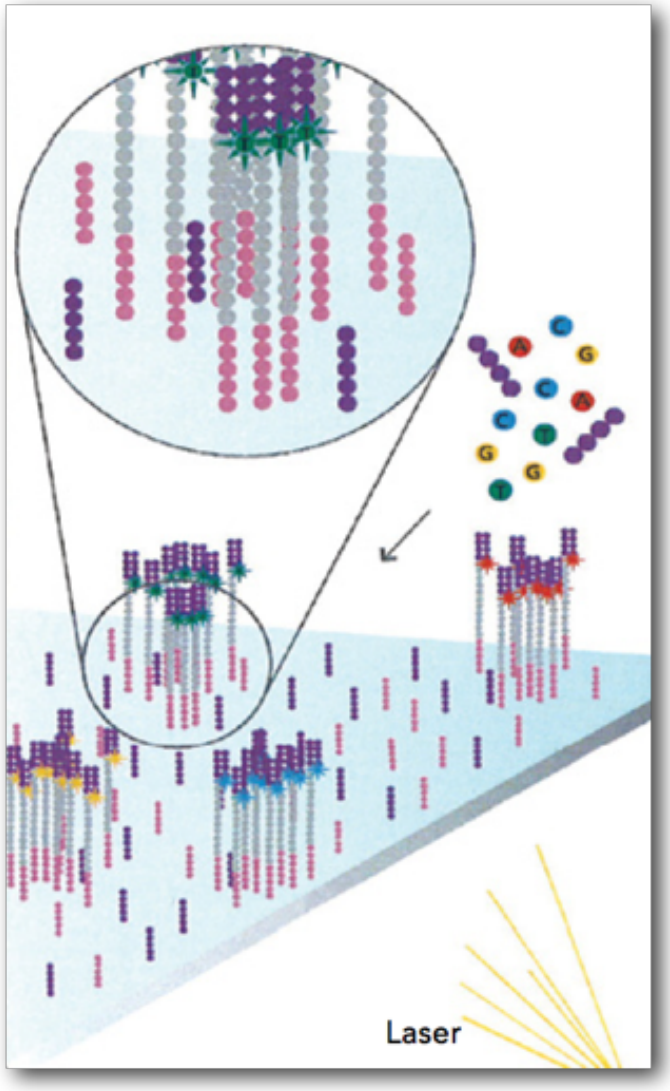


- To understand quality scores, it's helpful to understand how sequencing data are generated

Sequencing on Illumina's flow cell

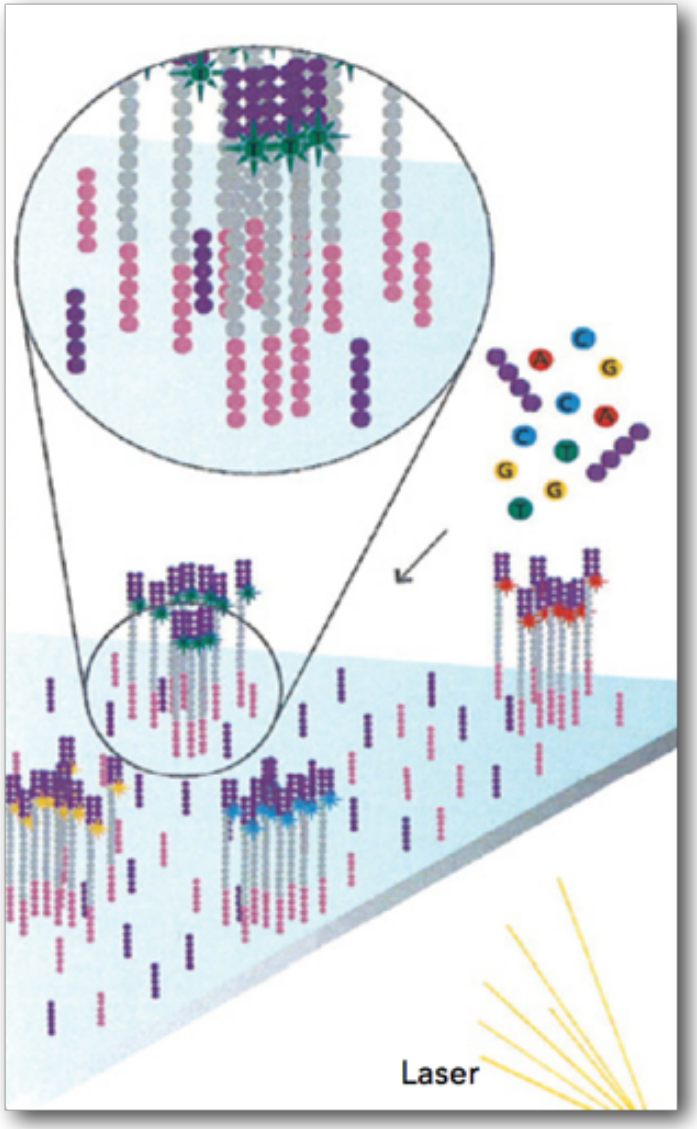


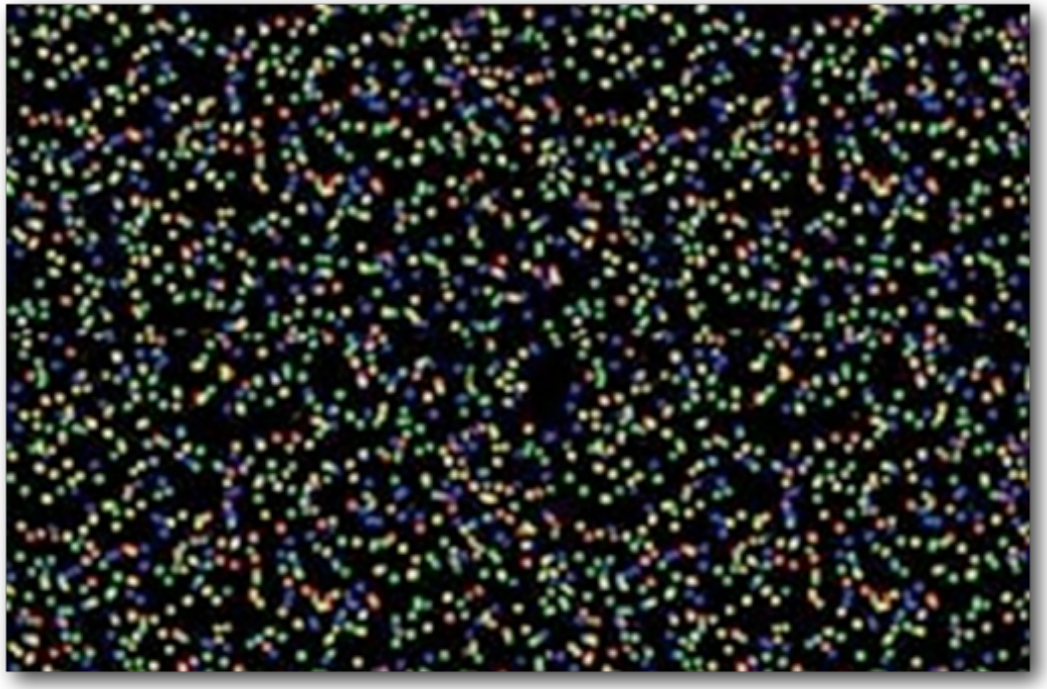




Source: Illumina Publication 770-2007-002 01May07

Sequencing on Illumina's flow cell



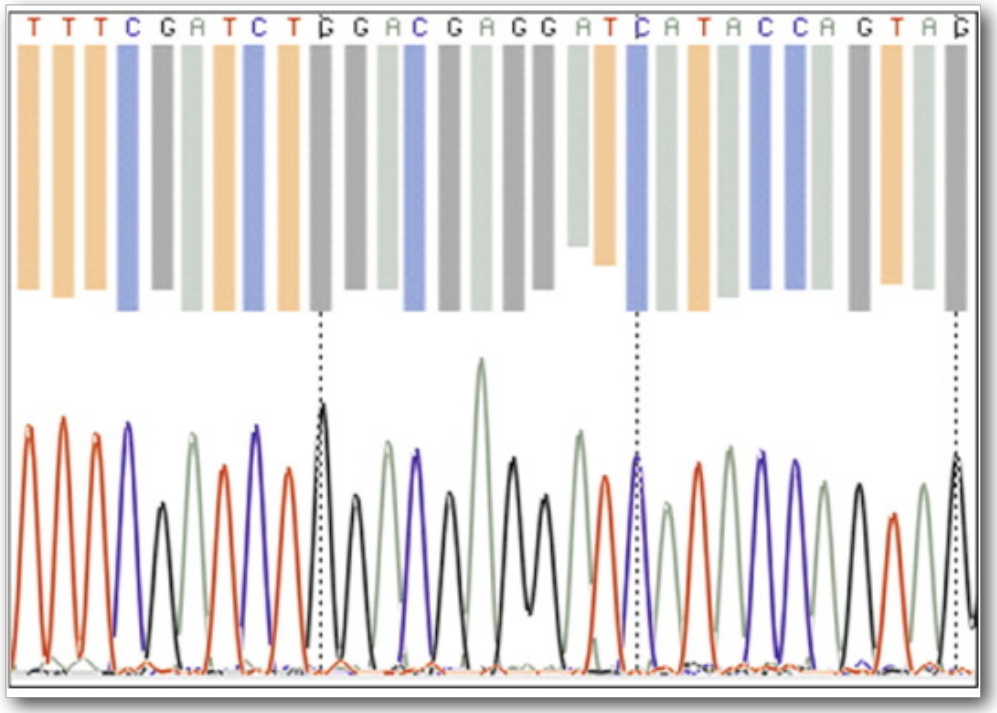


Source: Illumina Publication 770-2007-002 01May07

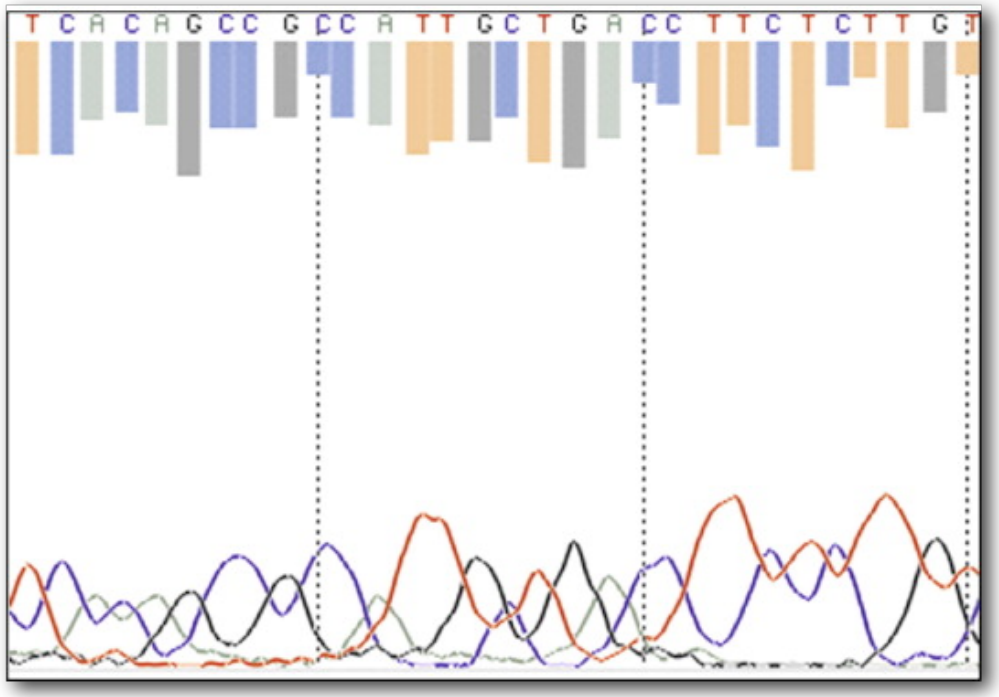
https://www.illumina.com/company/video-hub/fCd6B5HRaZ8.html

What does high/low quality data look like?

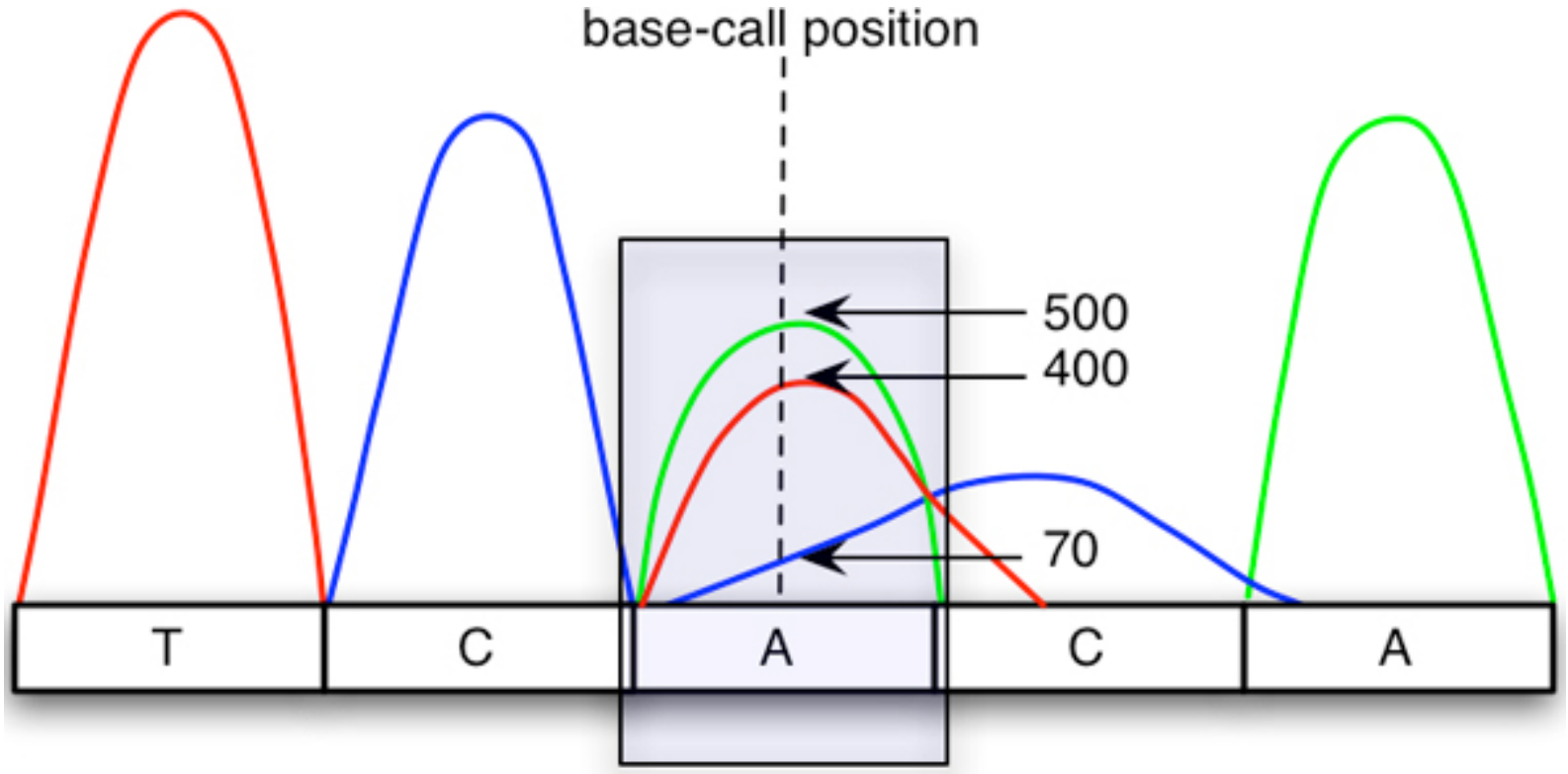
High quality data



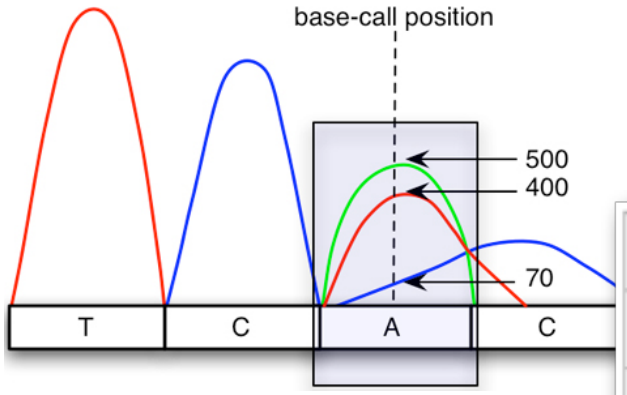
Lower quality data



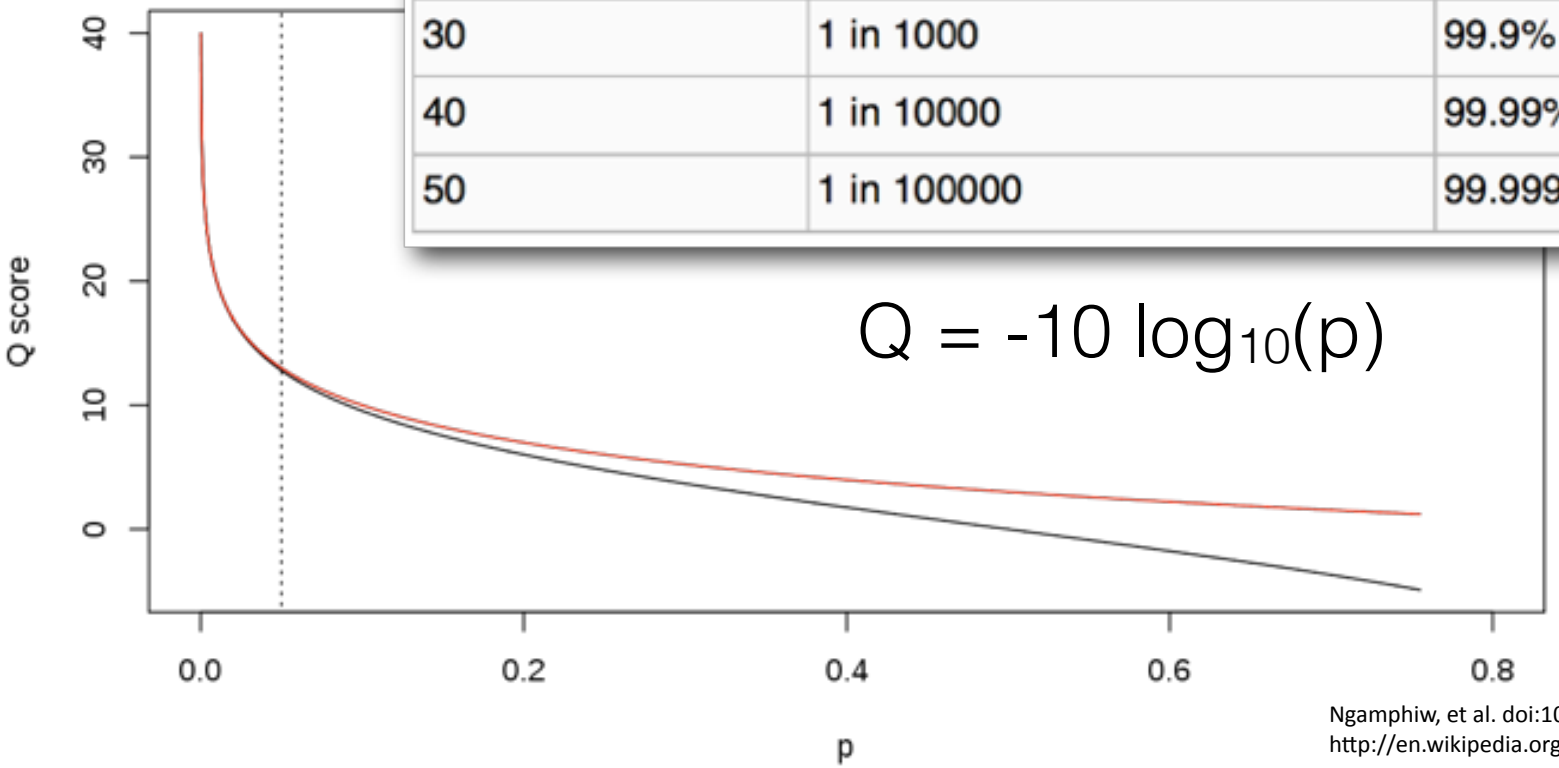
Each base call is given a score:
Phred quality score



Each base call is given a score:
Phred quality score



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%



$$Q = -10 \log_{10}(p)$$

Ngamphiw, et al. doi:10.1186/1471-2105-9-S12-S9
http://en.wikipedia.org/wiki/FASTQ_format

What does a FASTQ file look like?

```
@NS500451:154:HWKTMBGXX:1:11101:10065:1121 1:N:0:TAGAACAC
AGGTTGCTATGAAATTTTAGTTGTCGTAGTAGGCAAACAATAAGGAATGTTGATCCAATAATTACATGGAGTCCATGGAA
+
AAAAAEEEA6EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEE
```

File uses four lines per read

1. Begins with “@” followed by unique sequence identifier
2. Sequence of the read
3. Begins with “+” and is optionally followed by identifier again
4. Encodes the quality scores (must be same length as line 2)

Notice phred scores encoded by a single letter – how can we decode?

First, decode ASCII

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	¡	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	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	Ù
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	∂	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	∑	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	∏	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	∫	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	fl	255	ˇ

Each character has been assigned a numerical value that can be encoded by 8 bits

8 bits = 2^8 combinations = 256 possibilites

01111001

First, decode ASCII

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	¡	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	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	Ù
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	∂	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	∑	215	◊	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	∏	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	∫	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	fl	255	ˇ

Each character has been assigned a numerical value that can be encoded by 8 bits

8 bits = 2^8 combinations = 256 possibilites

01111001 $0 \times 2^7 + 1 \times 2^6 + 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 121 = y$

Then convert quality score

```
@NS500451:154:HWKTMBGXX:1:11101:10065:1121 1:N:0:TAGAACAC
AGGTTGCTATGAAATTTTAGTTGTCGTAGTAGGCAAACAATAAGGAATGTTGATCCAATAATTACATGGAGTCCATGGAA
+
AAAAAEEEA6EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEE
```

Quality scores

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                                  |                     |
33                        59   64       73                               104                   126

 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)
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

ASCII values 33 through 73 correspond to phred scores 0 through 40

'E' = 69 69 – 33 = **36** p_{error} = 0.025% Base call accuracy: 99.975%

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