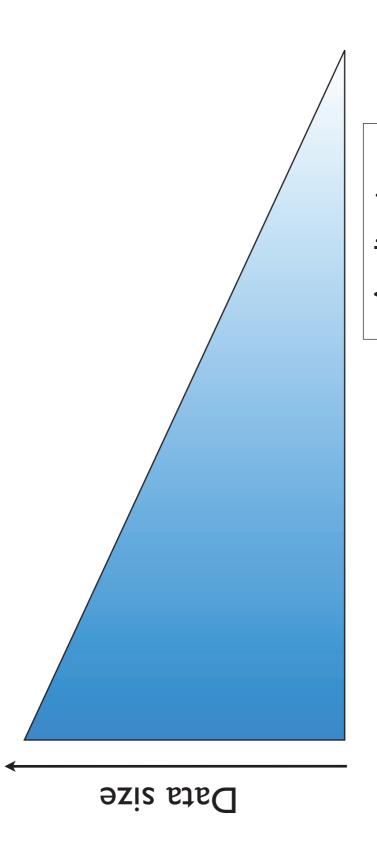
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Generalized NGS analysis



Application specific:

Assembly:

Alignment,

de novo

processing

reads

Question

Pre-

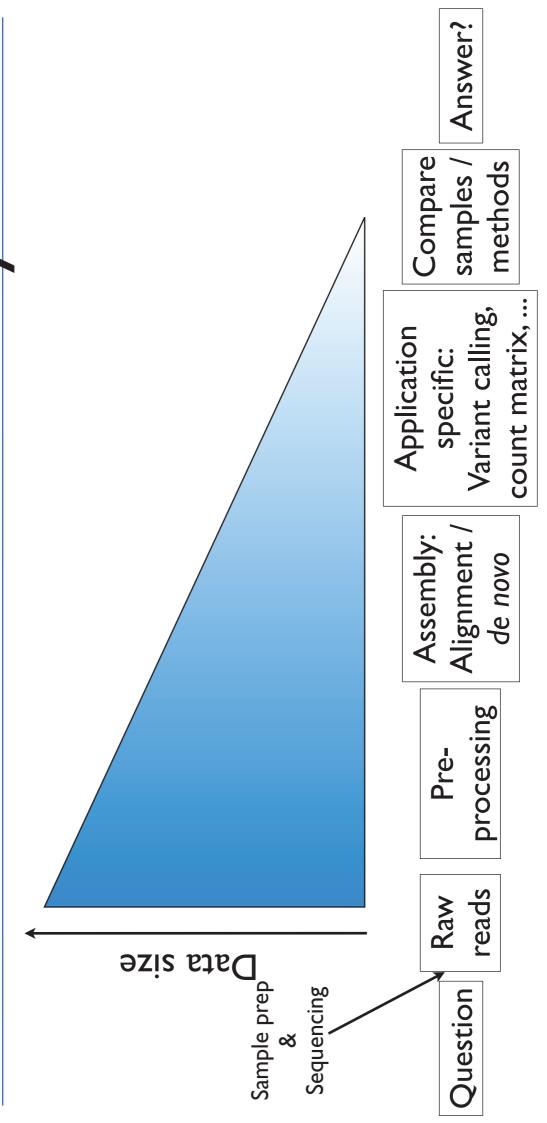
Raw

Variant calling, count matrix, ...

Answer? samples / Compare methods



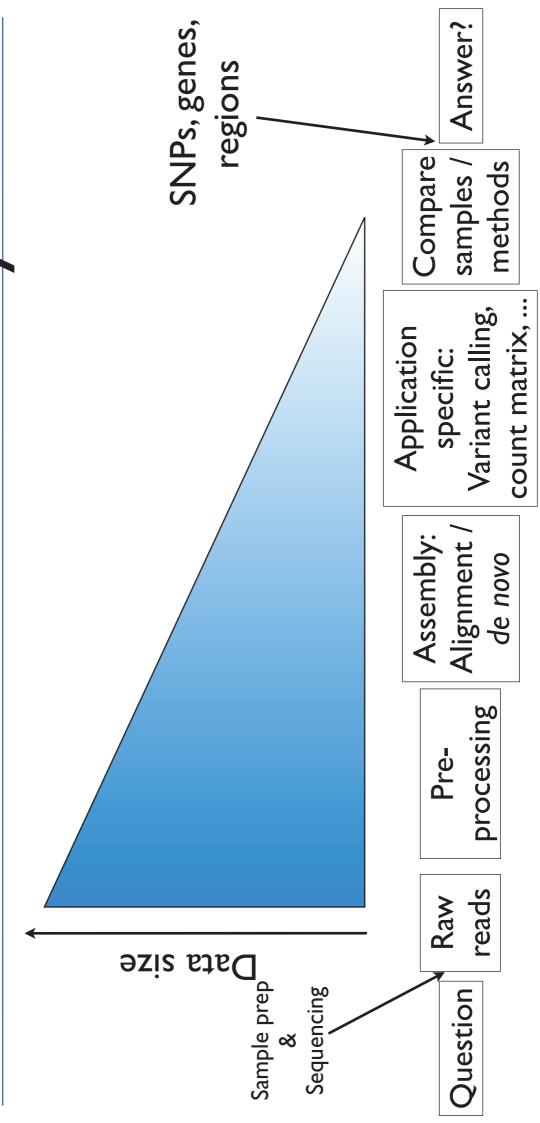
Generalized NGS analysis



27626 - Next Generation Sequencing Analysis



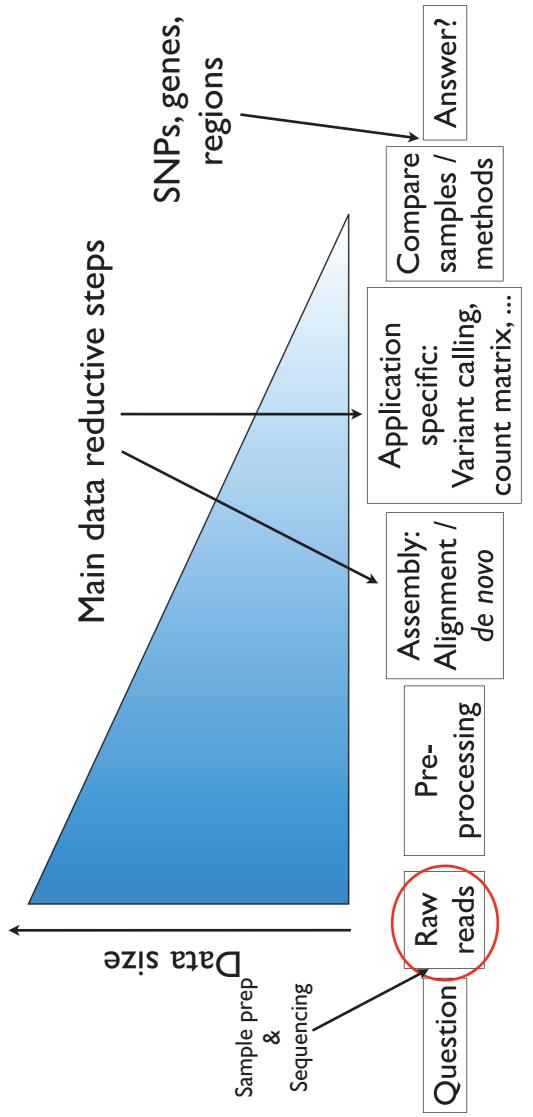
Generalized NGS analysis



27626 - Next Generation Sequencing Analysis



Generalized NGS analysis



27626 - Next Generation Sequencing Analysis



What is sequence data?

Sequences are stored in fasta-files

>gi|218693476|ref|NC_011748.1| Escherichia coli 55989 chromosome, complete genome ATACTTTAACCAATATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACACATCCATGAAACG GCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTG GTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTTAAATTTAAAATTTTATTGACTTAGGTCACTAA CATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAA GATCAGGAATTTGCCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGGCAGTGCCCGGATAGCA TGGCGATGATTGAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTT TGCCGAACTTTTGACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTC GTAAGTATTTTTCAGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGT Header

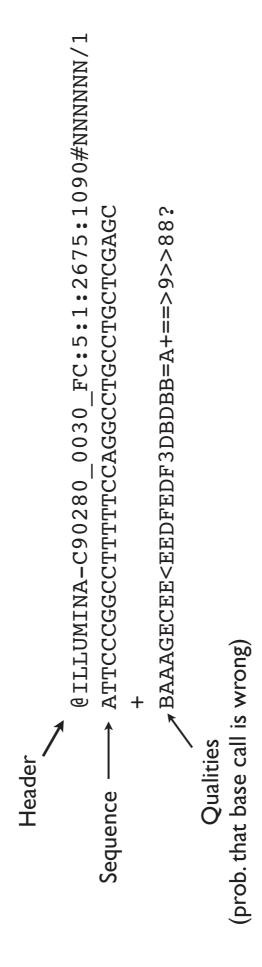
E.coli ~ 4.5 - 6 Mbases

Human ~ 3.2 Gbases



Then what is NGS data?

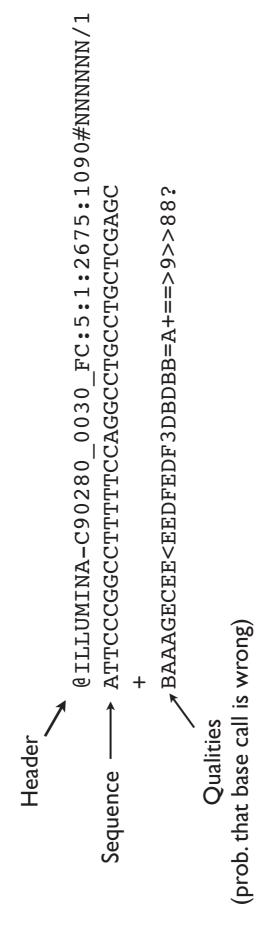
Fastg





Then what is NGS data?

Fastg



Millions to billions of these



A closer look at the qualities

@ILLUMINA-C90280 0030 FC:5:1:2675:1090#NNNN/1 ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC #BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88? (prob. that base call is wrong) Qualities Header Sequence

One character encodes a number using ascii table (0-255)

This number (Q) can be converted to P

Phred-scale

Q = -10 * log 10 P

 $P = 10^{\Lambda}(-Q/10)$



ATTIMINA_C90280 00 Dec	Dec	Hx Oct Char	Chai		Dec Hx C	Oct Html	Chr	Dec Hx	HX Oct H	Html Chr		Dec Hx Oct Html		Chr
	0	0000 0	NUL	(m11)	20	040 6#32;	Space	64 40	100 6#	6#64; B	96	60 140	:96#9	11
ATTCCCGGCCTTTTCCAG	-	1 001	SOH	(Start of heading)		041 6#33;	-	65 41	101 6#	6#65; A	97	61 141	£46#3	0
	7	2 002	STX	(start of text)	22	042 c#34;			102 &#	@#66; B	86	62 142	£86#3	A
+	n	3 003	MA	(end of text)	23	043 6#35;	*	67 43	103 6#	6#67; C	66	63 143	:66#3	0
	ঝ	4 004	EOL	(end of transmission)	24	044 6#36;	•	68 44	104 €#	c#68; D	100	64 144	«#100)	U
BAAAGECEE <fedfedf3d< td=""><td>r)</td><td>5 005</td><td>Z NO</td><td>(enquiry)</td><td>25</td><td>045 6#37;</td><td>*</td><td>69 45</td><td>105 6#</td><td>₹69 × ×</td><td>101</td><td>65 145</td><td>«#101;</td><td></td></fedfedf3d<>	r)	5 005	Z NO	(enquiry)	25	045 6#37;	*	69 45	105 6#	₹69 × ×	101	65 145	«#101;	
	9	900 9	ACK	(acknowledge)		046 6#38;	4	70 46	106 6#	4.470; F	102	66 146	6#102;	est.
	~	7 007	BEL	(bell)	27	047 6#39;	-	71 47	107 6#	6#71; 6	103	67 147	6#103;	Ch
	00	8 010	828	(backspace)		050 c#40;	_	72 48	110 6#	#72; H	104	68 150	c#104;	P
	0	9 011	TAB	(horizontal tab)		051 c#41;	_	73 49	111 6#	6#73; I	105	69 151	«#105;	ri.
	10	A 012	17	(NL line feed, new line)	42 2A 0	052 c#42;		74 4A	112 6#	6#74;	106	6A 152		
	11	B 013	×	(vertical tab)	43 2B 0	053 c#43;	+	75 4B	113 6#	6#75; K	107	6B 153	c#107;	м
	12	C 014	E	(NP form feed, new page)	44 2C 0	054 c#44;		76 4C	114 6#	6#76; L	108	6C 154	«#108;	-
	13	D 015	8	H)	45 2D 0	055 c#45;		77 4D	115 6#	6#77; M	109	6D 155	£#109;	
	14	E 016	20	(shift out)	46 2E 0	056 c#46;		78 4E	116 6#	6#78; M	110	6E 156	6#110;	B
	15	F 017	SI	(shift in)	47 2F 0	057 6#47;		79 4F	117 6#	0 162#9	111	6F 157	6#111;	0
	16	10 020	DIE	(data link escape)	30	060 c#48;	0		120 €#	C#80; 1	112	70 160	c#112;	a.
	17	11 021	100	(device control 1)	49 31 0	061 c#49;	-	81 51	121 6#	6#81; O	113	71 161	6#113;	
	18	12 022	200		50 32 0	062 c#50;	63	82 52	122 €#	6#82; R	114	72 162		м
	19	13 023	EUG M			063 c#51;	00	83 53	123 €#	£83;	115	73 163	6#115;	m
	20 7	14 024	DC4	(device control 4)	52 34 0	064 c#52;	7	84 54	124 €#	c#84; T	116	74 164	6#116;	4
	21	15 025	MAX	(negative acknowledge)	35	065 c#53;	us.	85 55	125 €#	6#85; U	117	75 165	6#117;	D
	22	16 026	SYN	(synchronous idle)	4 36	066 c#54;	٥	86 56	126 6#	₹ 198#°	118	991 94		P
	23	17 027	E	(end of trans. block)	37	067 c#55;	-	87 57	127 €#	@#87; W	119	77 167	c#119;	9
	24	18 030	CE	(cancel)	56 38 0	070 6#56;	0	88 58	130 c#	× 188#3	120	78 170	x	×
	25	180 61	H	(end of medium)	39	071 6#57;	o.	89 59	131 €#	¥89;	121	79 171	6#121;	24
	26	A 032	Sus	(substitute)	58 3A 0	072 :	**	90 SA	132 €#	2 :06#3	122	7A 172	6#122;	14
	27	LB 033	ESC	(escape)	59 3B 0	073 c#59;	**	91 SB	133 c#	161;	123	7B 173	{	-
	28	IC 034	73	(file separator)	30	074 6#60;	v	92 SC	134 €#	6#92; \	124	70 174		-
	29	ID 035	63	(group separator)	30	075 =		93 SD	135 €#	c#93;	125	7D 175		
	30	1E 036	22	(record separator)	62 3E 0	076 6#62;	٨	94 SE	136 €#	× 176#3	126	7E 176	c#126;	
	31	LF 037	ns	(unit separator)	3F	077 6#63;	**	95 SF	137 6#	:36	127	7F 177	6,#127;	DEL
:6 - Next Generation Sequencing Analysis										Sour	Source: w	ww.Lool	www.LookupTables.com	S.com



ATTITMIND COODSO ON Dec Hx Oct Char	Dec Hx Oct Ch	ar	Dec Hx Oct Html	Chr	Dec Hx Oct		Html	Chr D	Dec Hx Oct Html Chr	ct Htr	ol Chr	ř
	UN 000 0 0	(null)	20 040	Space	64 40	100	6#64;	9	60 1	40 6#96	. :96	
ATTCCCGGCCTTTTTCCAG	1 1 001 50	(start of heading)	33 21 041 6#33;	-	65 41	101	6#65;	A 97	61 1	41 6#97	97; 8	
	2 2 002 ST	(start of text)			66 42	102	£99#3	B 88	62 1	42 6#98	98; b	
+	3 3 003 ET	(end of text)	23 043 6	*	67 43	103	£467;	66	63 1	43 6#	0 166#3	
	4 4 004 20	(end of transmission)	24 044 6	**	68 44	104	:#68;	001 0	64 1	44 6#	#100; d	
BAAAGECEE <fedfedf3d< td=""><td>5 5 005 EW</td><td>(enquiry)</td><td>37 25 045 6#37;</td><td></td><td>69 45</td><td>105</td><td>£69#9</td><td>E 101</td><td>65 1</td><td>45 6#</td><td>#101; =</td><td></td></fedfedf3d<>	5 5 005 EW	(enquiry)	37 25 045 6#37;		69 45	105	£69#9	E 101	65 1	45 6#	#101; =	
	6 6 006 AC	(acknowledge)			70 46	106	. 0L#3	F 102	66 1	46 6#	6#102; E	
	7 7 007 BE	(bell)	39 27 047 6#39;	_	71 47	107	6#71;	6 103	67 1	47 6#	6#103; g	
	8 8 010 B3	(backspace)	40 28 050 6#40;		72 48	110	6#72;	H 104	68 1	\$0 c#	6#104; h	
	9 9 011 TM	(horizontal tab)	41 29 051 6#41;	_	73 49	111	6#73;	I 105	69 1	51 6#	6#105; 1	
	10 A 012 LF	(NL line feed, new line)	42 2A 052 6#42;	*	74 4A	112	6#74;	1 106	64	152 6#	c#106; 3	
	11 B 013 VT	(vertical tab)	43 2B 053 6#43;	+	75 4B	113	6#75;	K 107	68		6#107; k	
	12 C 014 FF	(NP form feed, new page)	44 2C 054		76 4C	114	c#76;	1 108	29	154 6#	c#108; 1	
99	13 D 015 CR	(carriage return)	45 2D 055		77 4	115	£177;	M 109	6D 1	55 c#	«#109; m	
	14 E 016 SO	(shift out)	46 2E 056 .		78 年	116	£82#3	M 110	SE.	156 6#	#110; m	
	15 F 017 SI	(shift in)	47 2F 057 6#47;		79 4F	117	662#3	0 111	6F	157 6#	6#111; o	
	16 10 020 DL	(data link escape)	48 30 060 6#48;	0	80 50	120	£80;	P 112	70	160 c#	6#112; p	
	17 11 021 DC	(device control 1)	49 31 061 c#49;	7	81 51	121	Q	0 113	71	161 6#	6#113; d	
	18 12 022 DC	2 (device control 2)	50 32 062 c#50;	63	82 52	122	6#82;	R 114	72	162 6#	6#114; E	
	19 13 023 DC	(device control 3)	51 33 063 c#51;	69	83 53	123	6#83;	\$ 115	73	163 €#	«#115; s	
	20 14 024 DC	(device control 4)	34 064	4	84 54	124	c#84;	T 116	74	164 c#	6#116; t	
	21 15 025 MA	(negative acknowledge)	53 35 065 c#53;	us.	85 55	125	£882	U 117	75	165 c#	6#117; u	
	22 16 026 ST	(synchronous idle)	36 066 €	9	86 56	126	¢#86;	V 118	94	166 c#	c#118; v	
	23 17 027 ET	8 (end of trans. block)	37 067	-	87 57	127	£81;	W 119	77	167 €#	#119; "	
	24 18 030 CM	(cancel)	38 070	60	88 58	130	188#3	X 120	78 1	₩» 02	6#120; ×	
	25 19 031 EM	(end of medium)	57 39 071 6#57;	04	89 59	131	£68#3	¥ 121	79	171 6#	6#121; ¥	
	26 1A 032 sur	(substitute)	58 3A 072 :		90 SA	132	106#3	2 122	7.14	172 6#	6#122; =	
	27 1B 033 ES	(escape)	59 3B 073 ;	**	91 SB	133	6#91;	123	7B 1	73 c#	6#123; 4	
	28 IC 034 FS	(file separator)	60 3C 074 <	v	92 SC	134	c#92;	124	20	174 6#	6#124;	
		(group separator)	3D 075		93 SD	135	¢#63;] 125	5		} 1	
	30 IE 036 PS	(record separator)	62 3E 076 >	٨		136	¢#64%	A 126	7E	176 6#	c#126; -	
	31 1F 037 US	(unit separator)		e.	95 SF	137	£36#3	12	7F	177 6#	6#127; D	-
- Next Generation Sequencing Analysis							So	Source:	www.Le	Take	www.LookupTables.com	E



GTT.T.IIMTNA_C90280 00 Dec Hx Oct Char	Dec	Hx Oct C	har		Dec	HX Oct	Httml	Chr	Dec	Hx Oct	Html	Ç	Dec	o Ť	Dec Hx Oct Html Chr	ਝੁੱ	X
	0	0 000 M		(null)		0 040	-8	Space	64 4	40 100				60 140			
ATTCCCGGCCTTTTCCAG(-	1 001 50	S) HO	start of heading)	33 2	21 041	L 6#33;	+	65 41	1 101		4	97 6	61 141	6#97	0	
	7	2 000 \$	TX (S	start of text)		22 042	2 6#34;	t	66 4	42 102				62 142		A	
+	n	3 003 2	E) XI	end of text)		23 043	3 c#35;	*	67 4	43 103		0	9 66	63 143		0	
	4	4 004 E	OT (e)	end of transmission)		24 044	4 6#36;	w	68 4	44 104		0		64 144		p : 0	
BAAAGECEE <eedfedf3d]< td=""><td>50</td><td>5 005 21</td><td>(e) (m</td><td>enquiry)</td><td></td><td>25 045</td><td>5 6#37;</td><td>*</td><td>69 4</td><td>45 105</td><td></td><td>64</td><td>101 6</td><td>65 145</td><td>5 e</td><td>11; =</td><td></td></eedfedf3d]<>	50	5 005 21	(e) (m	enquiry)		25 045	5 6#37;	*	69 4	45 105		64	101 6	65 145	5 e	11; =	
	9	€ 006 4	CK (a	acknowledge)		26 046	5 6#38;	4		46 106	. 6.#70;	-	102 6	66 146	5 6#102	12; €	
	-	7 007 BB	EL (b)	(be11)	39 2	27 047	7 6#39;	-		47 107	6#71	9 .	103 6	67 147	c#103	13; 0	
	00	8 010 B	(b)	(backspace)	40 2	28 050	0 6#40;	_	72 4	48 110		H	104 6	68 150	(#104°)	4; h	
	6	9 011 T	AB (h	(horizontal tab)	41 2	29 051	L)	_		49 111	6#73	-	105 6	69 151	c#105;	15; 1	
	10	A 012 L	N)	(NL line feed, new line)	42 2	2A 052	2 6#42;	+		4A 112		-	106 6	6A 152	, «#106;	16; 3	
	11	B OL3 V	T (V	(vertical tab)	43 2	2B 053	3 c#43;	+		4B 113		14	107 6	6B 153	8 k	17; 1	
	12	C 014 F	F (N	NP form feed, new page)	44 2	2C 054	1 c#44;		76 4	4C 114		1	108 6	6C 154		18; 1	
66 65	13	D 015 C	R (C	(carriage return)	45 2	2D 05	5 6#45	,	77 4	4D 115		N .	109 6	6D 155	%#109;	m :61	
	14	E 016 St	0 (3)	shift out)	46 2	2E 056	5 6#46;		78 4	4E 116	6.#78	N.	110 6	6E 156	6 6#110	u :0	
	15	F 017 S	I (S)	shift in)	47 2	2F 057	7 6#47;		79 4	4F 117	62#3	0	111 6	6F 157	0 8#111	11; 0	
	16	10 020 DI	LE (d	data link escape)	48 3	30 060	0 6#48	0	80 5	50 120	08#9			70 160	(#112;	2; p	
	17	11 021 DC	C1 (d	device control 1)	49 3	31 061	67#3 1	-	81 5	51 121	6#81;	0		71 161	6#113	3; 4	
	18	12 022 De	C2 (d)	device control 2)		32 062	2 6#50;	63		52 122	. Z;	OK.	114 7	72 162		4; =	
	13	13 023 D	(d)	device control 3)		33 063		00		53 123		572	115 7	3 163		5; =	
	20	14 024 D	C4 (d)	device control 4)		34 064	4	4		54 124		1	116 7	74 164		6; t	
	21	15 025 M	AX (no	negative acknowledge)		35 065	0	us;		55 125		D	117 7	75 165		7; u	
	22	16 026 51	TI (S)	synchronous idle)	54 3	36 066	W.	٥	86 5	6 126		1	118 7	991 94	811#3 S	8: 4	
	23	17 027 E	(e)	end of trans. block)		37 067	W.	-	87 5	7 127			119 7	77 167		a :6:	
	24	18 030 C	(C)	cancel)		38 070	W	60		58 130		×	120 7	78 170		× 10	
	25	19 031	m (e)	end of medium)		39 071	4	o.	89 5	59 131		1	121 7	171 67	. «#121;	11: 3	
	56	1A 032 ST	(S)	substitute)	58 3	3A 072	2 6#58;		90 5	5A 132	06#9	24	122 7	7A 172	. «#122;	2; =	
	27	1B 033 E	3C (e:	escape)	59 3	3B 073	8 c#59;	**	91 5	SB 133	16#91	-	123 7	7B 173		3;	
	28	IC 034 F	S (E	file separator)	60 3	30 074	09#9 t	v		5C 134	c#92	-	124 7	C 174		4;	
	53	1D 035 G	(g)	group separator)	250	3D 075				5D 135						:2:	
	30	1E 036 P	3 (1	record separator)	62 3	3E 076	9#3	٨	4	SE 136	3	*	126 7	7E 176		- 19	
	31	1F 037 U	m) s	unit separator)		3F 077	7 6#63;	64	LO.	5F 137	\$6#9			7F 177	0 	77, 10	-3
6 - Next Generation Sequencing Analysis												Source:		w.Loo	www.LookupTables.com	les .ce	E



GTT.T.IIMTNA_C90280 00 Dec Hx Oct Char	Dec Hx Oct	Char		Dec Hx Oct		Html Chr	Dec Hx Oct		Html (Chr	Dec Hx Oct Html Chr	Oct II	ml Chr	Ť
	000 0 0	NUL	(null)	20	040 6#	#32; Space	64 40	100	6#64;	9	6 60 1	40 6#	.#96 #°	
ATTCCCGGCCTTTTTCCAG	1 1 001	SOH	(start of heading)	33 21 (041 6#	#330 ·	65 41	101	6#65;	A 97	7 61 1	41 6#	£#97;	
	2 2 002	STX	(start of text)	22	042 6#	#34; **	66 42	102	£99#3	B6 B	8 62 1	42 6#	** ** 86#3	
+	3 3 003	ETX	(end of text)	23	043 6#	#32%	67 43	103	£#67;	66	9 63 1	43 6#	166#3	
	4 4 004	EOT	(end of transmission)	24	044 6#	#36,	68 44	1 104	£89#3	D 100		44 6#	«#100; e	
BAAAGECEE <eedfedf3d< td=""><td>5 5 005</td><td>EMO.</td><td>(enquiry)</td><td>25</td><td>045 6#</td><td>#372 *</td><td>69 45</td><td>105</td><td>£69#3</td><td>E 101</td><td>1 65 1</td><td>45 &#</td><td>6#101; e</td><td>20</td></eedfedf3d<>	5 5 005	EMO.	(enquiry)	25	045 6#	#372 *	69 45	105	£69#3	E 101	1 65 1	45 &#	6#101; e	20
	900 9 9	ACK	(acknowledge)	26	046 6#	#38; e	70 46	901	. 0L#3	F 102	99	146 6#	c#102;	300
	7 7 007	BEL	(bell)		047 6#	#36°	71 47	107	6#71;	6 103	3 67 1	47 6#	6#103; g	_
	8 8 010	18.50	(backspace)		050 6#	#40;	72 48	110	6#72;	H 104	89	150 €#	6#104; I	-
- -	9 9 011	TAB	(horizontal tab)	41 29 (051 6#	6#41;	73 49	111	6#73;	I 105	69	151 6#	6#105; a	
	10 A 012	E.	(NL line feed, new line)	42 2A (052 6#	c#42; *	74 4A	1112	6.#74;	106	64	152 €#	c#106;	
	11 B 013	1	(vertical tab)	43 2B (053 c#	#43:+	75 48	1113	c#75;	K 107	68	153 €#	6#107; k	
	12 C 014	1	(NP form feed, new page)	44 2C (054 c#	#444;	76 40	1114	c#76;	1 108	29	154 €#	c#108; 1	25
66 65 65	13 D 015	8		45 2D (055 c#	#457 -	77 40	1115	£177;	M 109	09	155 &#</td><td>#109;</td><td></td></tr><tr><td></td><td>14 E 016</td><td>20</td><td>(shift out)</td><td>46 2E (</td><td>056 c#</td><td>#46;</td><td>78 年</td><td>1116</td><td>£ 18;</td><td>M 110</td><td>6E</td><td>156 €#</td><td>c#110; "</td><td>_</td></tr><tr><td></td><td>15 F 017</td><td>SI</td><td>(shift in)</td><td>47 2F (</td><td>057 6#</td><td>.#47; /</td><td>79 4F</td><td>7117</td><td>166#3</td><td>0 111</td><td>6F</td><td>157 6#</td><td>6#111; c</td><td></td></tr><tr><td></td><td>16 10 020</td><td>DIE</td><td>(data link escape)</td><td>48 30 (</td><td>#9 090</td><td>#48; 0</td><td>80 50</td><td>120</td><td>\$ 80°</td><td>P 112</td><td>70</td><td>160 €#</td><td>6#112; r</td><td>-</td></tr><tr><td></td><td>17 11 021</td><td>DCI</td><td>(device control 1)</td><td>49 31 (</td><td>061 c#</td><td>6#49; 1</td><td>81 51</td><td>121</td><td>6#81;</td><td>0 113</td><td>71</td><td>161 &#</td><td>6#113; c</td><td>-</td></tr><tr><td></td><td>18 12 022</td><td>DCZ</td><td>(device control 2)</td><td>50 32 (</td><td>062 c#</td><td>c#50; 2</td><td>82 52</td><td>122</td><td>6#82;</td><td>R 114</td><td>72</td><td>162 €#</td><td>6#114; s</td><td></td></tr><tr><td></td><td>19 13 023</td><td>DC3</td><td>(device control 3)</td><td>51 33 (</td><td>063 c#</td><td>6#51; 3</td><td>83 53</td><td>123</td><td>£83;</td><td>\$ 115</td><td>73</td><td>163 €#</td><td>c#115;</td><td></td></tr><tr><td></td><td>20 14 024</td><td>DC4</td><td>(device control 4)</td><td>34</td><td>064 c#</td><td>6#52; 4</td><td>84 54</td><td>1 124</td><td>c#84;</td><td>1116</td><td>74</td><td>164 cm</td><td>6#116; t</td><td></td></tr><tr><td></td><td>21 15 025</td><td>MAN</td><td>(negative acknowledge)</td><td>53 35 (</td><td>99 S90</td><td>#53° S</td><td>85 5</td><td>5 125</td><td>£882</td><td>U 117</td><td>75</td><td>165 &#</td><td>6#117; "</td><td>-</td></tr><tr><td></td><td>22 16 026</td><td>SYN</td><td>(synchronous idle)</td><td>36</td><td>#3 990</td><td>#54, 6</td><td>10</td><td>126</td><td>¢#86;</td><td>V 118</td><td>94</td><td>166 c#</td><td>c#118;</td><td></td></tr><tr><td></td><td>23 17 027</td><td>2</td><td>(end of trans. block)</td><td>37</td><td>#≈ C90</td><td>#SS*</td><td></td><td>127</td><td>£81;</td><td>W 119</td><td>9 77 1</td><td>67 €#</td><td>6#119; "</td><td></td></tr><tr><td></td><td>24 18 030</td><td>CAN</td><td>(cancel)</td><td>38</td><td>%° 020</td><td>#562 8</td><td>88 58</td><td>130</td><td>188#3</td><td>X 120</td><td>78 1</td><td>70 €#</td><td>x *</td><td>-</td></tr><tr><td></td><td>25 19 031</td><td>EM</td><td>(end of medium)</td><td>39</td><td>071 6#</td><td>#572 9</td><td>89 59</td><td>131</td><td>£68#3</td><td>¥ 121</td><td>79</td><td>171 6#</td><td>6#121; y</td><td>-</td></tr><tr><td></td><td>26 1A 032</td><td>SUB</td><td>(substitute)</td><td>58 3A (</td><td>072 c#</td><td>#58; :</td><td>90 SA</td><td>132</td><td>106#3</td><td>2 122</td><td>7.4</td><td>172 €#</td><td>6#122;</td><td></td></tr><tr><td></td><td>27 1B 033</td><td>ESC</td><td>(escape)</td><td>59 3B (</td><td>073 c#</td><td>#59;</td><td>91 SB</td><td>133</td><td>6#91;</td><td>123</td><td>7.8</td><td>173 €#</td><td>6#123;</td><td></td></tr><tr><td></td><td>28 IC 034</td><td>FS</td><td>(file separator)</td><td>) DE 09</td><td>074 c#</td><td>> :09#9</td><td>92 50</td><td>134</td><td>c#92;</td><td>124</td><td>20</td><td>174 €#</td><td>6#124;</td><td></td></tr><tr><td></td><td>29 ID 035</td><td>65</td><td>(group separator)</td><td>30</td><td>075 6#</td><td>c#61; =</td><td>93 SD</td><td>135</td><td>£86#3</td><td>1125</td><td>5</td><td>175 cm</td><td>6#125;</td><td></td></tr><tr><td></td><td>30 IE 036</td><td>PS</td><td>(record separator)</td><td>62 3E (</td><td>#3 9L0</td><td>€#62;></td><td>94 SE</td><td>136</td><td>¢#84;</td><td>A 126</td><td>7E</td><td>176 c#</td><td>c#126; -</td><td></td></tr><tr><td></td><td>31 1F 037</td><td>ns</td><td>(unit separator)</td><td>3F</td><td>#2 220</td><td>63; 2</td><td>95 SF</td><td>137</td><td>456#°</td><td>_ 12</td><td>7F</td><td>177 6#</td><td>6#127; I</td><td>EL</td></tr><tr><td>6 - Next Generation Sequencing Analysis</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Ŝ</td><td>Source:</td><td>www.L</td><td>ookap</td><td>www.LookupTables.com</td><td>MO:</td></tr></tbody></table>		



ATTIMINA_C90280 00 Dec Hx Oct Char	Dec Hx Oct Cha		Dec Hx Oct		Html Chr	Dec Hx Oct		Html C	Chr De	Dec Hx Oct Html	oct H	nl Chr	i
	0 0 000 MUL	(nu11)	20	040 6#	#32; Space	64 40	100 6	#64;	96	60 1	40 6#9	:96	
ATTCCCGGCCTTTTTCCAG	1 1 001 SOH	(start of heading)	33 21 0	041 6#	#33°	65 41	101 6	6#65;	97	61 1	41 6#	6,497;	
	2 2 002 STX	(start of text)	22	042 c#	#34;	66 42	102 €	c#66;	96	62 1	42 6#	1 :86#3	30
+	3 3 003 ETX	(end of text)	23	043 6#	#32°	67 43		#67;	66	63 1	43 6#	66#3	
	4 4 004 EOT	(end of transmission)	24	044 c#	#36;	68 44	104 €	#68;	100	64 1	44 6#	c#100;	
BAAAGECEE <eedfedf3di< td=""><td>5 5 005 ENQ</td><td>(enquiry)</td><td>25</td><td>045 6#</td><td>#37; *</td><td>69 45</td><td>105 €</td><td>£ 69 * 9</td><td>101</td><td>65 1</td><td>45 6#</td><td>#101;</td><td>200</td></eedfedf3di<>	5 5 005 ENQ	(enquiry)	25	045 6#	#37; *	69 45	105 €	£ 69 * 9	101	65 1	45 6#	#101;	200
	6 6 006 ACK	(acknowledge)	26	046 c#	#38; 6	70 46	106 €	%#70;	102	66 1	46 6#	c#102;	
	7 7 007 BEL	(bell)		047 c#	#30;	71 47	107 €	6#71;	5 103	67	147 6#	6#103; a	
	8 8 010 BS	(backspace)		050 c#	#40;	72 48	110 6	6#72;	104	89	150 6#	6#104; I	
	9 9 011 TAB	(horizontal tab)	41 29 0	051 6#	#41;)	73 49	111 6	6#73;	105	69	151 6#	c#105;	
	10 A 012 LF	(ML line feed, new line)	42 2A 0	052 €#	#42; #	74 4A	112 6	6#74;	106	64	152 6#	«#106;	
	11 B 013 VT	(vertical tab)	43 2B 0	053 c#	6#43;+	75 48	113 6	6#75;	C 107	68	153 c#	6#107; k	
L	12 C 014 FF	(NP form feed, new page)	44 2C 0	054 c#	#444;	76 4C	114 6	c#76;	108	29	154 c#	c#108;	33
66 65 65	13 D 015 CR	(carriage return)	45 2D 0	055 c#	#45; -	77 4D	115 6	6#77;	109	09	155 c#	* 1001#°	
	14 E 016 SO	(shift out)	46 2E 0	056 c#	#46;	78 年	116 6	£#78;	110	6E	156 6#	c#110; "	
	15 F 017 SI	(shift in)	47 2F 0	057 c#	#47;	79 4F	117 6	162#9	111	6F	157 6#	6#111; c	
- (16 10 020 DLE	(data link escape)	48 30 0	#º 090	6#48; 0	80 50	120 €	£ 80 %	112	20	160 c#	6#112; r	
C ~ Prob	17 11 021 DC1	(device control 1)	49 31 0	061 c#	#49; 1	81 51	121 6	c#81;	113	71	161 6#	c#113;	_
	18 12 022 DCZ	(device control 2)	32	062 €#	c#50; 2			6#82;	114	72		c#114;	
	19 13 023 DC3	(device control 3)	33	063 c#	c#51; 3		123 €	¢#83;	1115	73	163 6#	c#115;	
	20 14 024 DC4	(device control 4)	34	064 c#	6#52; 4	84 54		6#84;	116	74	164 6#	c#116;	
	21 15 025 MAN.	(negative acknowledge)	35		6#53; 5	5	125 €	£882°	1 117	75	165 c#	6#117; "	_
	22 16 026 SYN	(synchronous idle)	36	# 990	#54; 6	w	126 €	798#7	1118	94	166 6#	c#118;	
	23 17 027 ETB	(end of trans. block)	37	#9 L90	#55%	87 57	127 6	c#87;	1119	77 1	67 c#	c#119;	
	24 18 030 CAN	(cancel)	56 38 0	\$ 040 040	#56; 8	88 58	130 €	188#3°	120	78 1		#120; *	
2	25 19 031 EM	(end of medium)	39	071 c#	#57, 9	89 59	131 €	*************	121	79	171 6#	6#121; y	
	26 1A 032 sub	(substitute)	34	072 €#	#58; :	90 SA	132 €	:06#°	122	7.4	172 €#	6#122; =	-
40 ~ 0 0000	27 1B 033 ESC	(escape)	59 3B 0	073 c#	#59;	91 SB	133 €	6#91;	123	7.8	173 c#	6#123; I	
	28 IC 034 FS	(file separator)	0 36 09	074 c#	> :09#9	92 50	134 €	6#92;	124	20	174 6#	6#124;	
	29 ID 035 GS	(group separator)	30	075 c#	#61; =	93 SD	135 €	¢#83;	125	5		6#125;	
	30 IE 036 PS	(record separator)	62 3E 0	#3 9L0	€#62;>	94 SE	136 €	¢#6#9	126	7E	176 6#	c#126; -	
	31 1F 037 US	(unit separator)	3F	#3 LLO	#63; 2	95 SF	137 6	#882	127	7F	177 6#	6#127; I	
.7626 - Next Generation Sequencing Analysis								Sou	Source:	www.L	Tokap	www.LookupTables.com	HO



GTT TIMTNA_CG0280 00 Dec Hx Oct Char	Dec Hx Oct Cha		Dec H	Hx Oct	Html	Chr	Dec Hx Oct	x Oct	Html	Chr	Dec Hx Oct Html	x Oct	Html C	Chr
	JUN 000 0 0	(mull)		040	6#32;	Space	64 40	001 0	6#64	8	96 60	140	:96#3	10
ATTCCCGGCCTTTTTCCAG	HO 1 1 001 SOH	(start of heading)	33 21	041	6#33	-	65 41	101	6#65	4	97 61	141	£46#3	0
	2 2 002 STX	(start of text)		2 042	c#34;	ŧ	66 42	2 102	c#66;	0	98 62	142	:86#3	A
+	3 3 003 ETX	(end of text)		3 043	£#32	*	67 43	N.		0	99 63	143	:66#3	0
	4 4 004 EOT	(end of transmission)		1 044	6#36;	w	68 44	1 104		0	100 64	144	«#100°	U
BAAAGECEE <fedfedf3d< td=""><td>5 5 005 ZWO</td><td>(enquiry)</td><td>37 25</td><td>5 045</td><td>6#37;</td><td>de</td><td>69 45</td><td>5 105</td><td>69#9</td><td>64</td><td>101 65</td><td>145</td><td>«#101;</td><td>60</td></fedfedf3d<>	5 5 005 ZWO	(enquiry)	37 25	5 045	6#37;	de	69 45	5 105	69#9	64	101 65	145	«#101;	60
	6 6 006 ACK	(acknowledge)		5 046	6#38;	9	70 46	901 9	6.#70	ja,	102 66	146	6#102	u
	7 7 007 BEL	(bell)		7 047	68#3	-		7 107	6#71	9	103 67	147	6#103,	D
	8 8 010 BS	(backspace)		3 050	6#40;	_	72 48	3 110	6#72	H	104 68	150	6#104,	4
- -	9 9 011 TAB	(horizontal tab)	41 29	9 051	6#41;		73 49	9 111	6#73	+	105 69	151	i	ri.
	10 A 012 LF	(NI line feed, new line)	42 2A	A 052	6#42;		74 4A	1112	6.#74	-	106 6A	152	%#10e	-
	11 B 013 VT	(vertical tab)	43 2B	8 053	6#43;	+	75 4B	8 113	6#75	14	107 6B	153	c#107;	м
	12 C 014 FF	(NP form feed, new page)	44 2C	054	6#44;		76 4C	114	6#76;	1	108 6C	154	6#108	-
66 65 65 ~ le-6	13 D 015 CR		45 2D	055	6#45;	,	77 4D	1115	6#77;	M	109 6D	155	£#109;	
	14 E 016 S0	(shift out)	46 2E	0.056	6#46;		78 4	2 116	6.#78		110 GE		6#110	F
	15 F 017 SI	(shift in)	47 2F	7 057	6#47;		79 4F	F 117	66#3	0	.11 6F	157	6#111;	0
- (16 10 020 DLE	(data link escape)	48 30	090 0	6#48;	0	80 50	0 120	6#80	di .	112 70	160	6#112,	0
○ ~ Prob	17 11 021 DC1	(device control 1)	49 31	190 1	6#49;	-	81 51	121	6#81;	0	113 71	161	6#113,	5
	18 12 022 DC2	(device control 2)	50 32	2 062	c#20;	63	82 52	2 122	6#82	06	114 72	162	6#1143	ш
	19 13 023 DC3	(device control 3)	51 33	3 063	c#21;	69		3 123	6#83	50	15 73	163	c#115;	m
	20 14 024 DC4	(device control 4)	52 34	1 064	6#52;	4	84 54	1 124	6#84	1	116 74	164	«#116,	4
	21 15 025 MAK	(negative acknowledge)	53 35	5 065	6#53	S	5	5 125	£#82	D	117 75	165	6#117;	2
	22 16 026 STM	(synchronous idle)		990 9	c#245	ي	w	5 126	98#9	Δ.	118 76	166	6#118	P
	23 17 027 ETB	(end of trans. block)	55 37	7 067	£ 822	-	87 57	7 127	£87;	. W .	77 61.	167	c#119;	2
	24 18 030 CAN	(cancel)	56 38	3 070	c#26;	60	88 58	3 130	88#3	×	120 78	170	£#120,	×
100.0 \(\text{OC} \)	25 19 031 EM	(end of medium)	57 39	9 071	6#57;	o.	89 5	9 131	68#3	Y	121 79	171	6#121;	21
	26 1A 032 sub	(substitute)	58 3A	\$ 072	¢#28;	**	90 SA	132	06#3	143	122 7A	172	6#122;	14
40 ~ ()()()()	27 1B 033 ESC	(escape)	59 3B	3 073	6,820	**	91 SB	3 133	6#91		123 7B	173	6#123;	_
	28 IC 034 FS	(file separator)	96 30	074	°#90	v	92 SC	134	6#92	-	124 70	174	6#124;	_
	29 ID 035 GS	(group separator)	61 3D	075	c#61;		93 SD	135	6#3		125 7D		6#125;	
	30 IE 036 PS	(record separator)	62 3E	076	c#62;	٨	94 SE	136	6#94	٧.	126 7E	176	6#126	
	31 1F 037 US	(unit separator)	63 3F	7 077	c#63;	64	95 SF	F 137	6#95	_	27 7F	177	6#127	DEL
626 - Next Generation Sequencing Analysis									41	Source:		Jook	www.LookupTables.com	S.com



Base qualities, read mapping qualities, variant qualities, ...

Straight-forward, except for when they are used in reads!

Offset: Sanger = 33, Illumina = 64

@ILLUMINA-C90280 0030 FC:5:1:2675:1090#NNNNN/1 ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC

Phred: 666565 ~ le-6

BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?

O ~ Prob 10 ~ 0.1 20 ~ 0.01 30 ~ 0.001 40 ~ 0.0001



Base qualities, read mapping qualities, variant qualities, ...

Straight-forward, except for when they are used in reads!

Offset: Sanger = 33, Illumina = 64

@ILLUMINA-C90280 0030 FC:5:1:2675:1090#NNNNN/1

ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC

+

BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?

Phred: 666565 ~ le-6

Sanger: 33 32 32 ~0.001

 $20 \sim 0.01$ $30 \sim 0.001$ $40 \sim 0.0001$

|0 ~ 0:|



Base qualities, read mapping qualities, variant qualities, ...

Straight-forward, except for when they are used in reads!

Offset: Sanger = 33, Illumina = 64

@ILLUMINA-C90280 0030 FC:5:1:2675:1090#NNNNN/1

ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC

+

BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?

Phred: 66 65 65 ~ le-6

Sanger: 33 32 32 ~0.001

Illumina: 2 | | ~

 $10 \sim 0.1$ $20 \sim 0.01$ $30 \sim 0.001$ $40 \sim 0.0001$



Base qualities, read mapping qualities, variant qualities, ...

Straight-forward, except for when they are used in reads!

Offset: Sanger = 33, Illumina = 64

@ILLUMINA-C90280 0030 FC:5:1:2675:1090#NNNNN/1

ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC

BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?

Phred: 666565 ~ le-6

Sanger: 33 32 32 ~0.001

Illumina: 2 | |

HUGE difference!

I.0 ~ 0.

 $20 \sim 0.01$ $30 \sim 0.001$

40 ~ 0.000 I



Base qualities, read mapping qualities, variant qualities, ...

Straight-forward, except for when they are used in reads!

Offset: Sanger = 33, Illumina = 64

@ILLUMINA-C90280_0030_FC:5:1:2675:1090#NNNN/1

ATTCCCGGCCTTTTTCCAGGCCTGCCTGCTCGAGC

BAAAGECEE<EEDFEDF3DBDBB=A+==>9>>88?

Phred: 666565 ~ le-6

Sanger: 33 32 32 ~0.001

Illumina: 2 | |

HUGE difference!

Exercise today

I.0 ~ 0.

 $20 \sim 0.01$ $30 \sim 0.001$

40 ~ 0.000 I



Sanger vs. Illumina vs. Solexa

- 454, Ion Torrent, Pac Bio, Sanger: "Sanger" encoding
- Illumina reads: "Illumina" or "Sanger" encoding. New reads are all "Sanger"
- Solexa data: Solexa encoding (bought by Illumina)
- All data from SRA: "Sanger"

Fragment DNA: ____



Single end

Paired end Ins: 200-800 bp

Mate pair Ins: 2kb - 40kb (~5kb)

Fragment DNA:

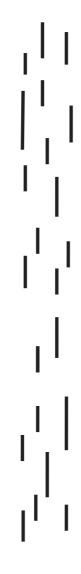


Single end

Paired end Ins: 200-800 bp

Mate pair Ins: 2kb - 40kb (~5kb)

Fragment DNA:





Mate pair Ins: 2kb - 40kb (~5kb)

Ins: 200-800 bp

Paired end

Single end

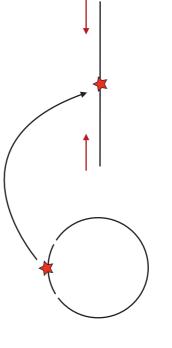
Fragment DNA:





Paired end Ins: 200-800 bp

Single end

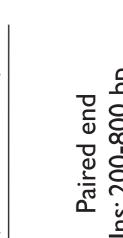


Mate pair Ins: 2kb - 40kb (~5kb)

Fragment DNA:





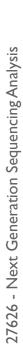


Mate pair Ins: 2kb - 40kb (~5kb)

Protocol/technology dependent









Read orientation

Single end

Forward

Paired end

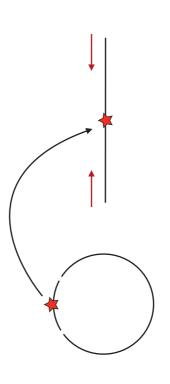


Illumina: Forward - Reverse





Mate pair



Illumina: Reverse - Forward



Different for other technologies!



Special applications

- Single end reads:
- Sometimes the only possibility (small DNA fragments / ancient DNA)
- Paired end reads:
- More precise mapping/alignment/variation calls
- Medium/Large indels (insertion/deletion)
- Structural variations
- Scaffolding in de novo assembly
- Mate pairs:
- Scaffolding in de novo assembly
- Structural variations



he reads comes in ?

Illumina: fastq

454: sff, fasta, fasta+qual, fastq

lon torrent: sff, fastq

Pac Bio: fastq

Solid: csfasta, qual, xseq



Never keep data in fasta+qual/csfasta+qual - the qual format is horrible (and large)

sff is large, convert to fastq unless you use flowgram-capable software

Convert csfasta+qual to csfastq (at least if you are doing alignment)



Question

What does it mean to have paired end reads?

Discuss with neighbor for 2-3 mins, we discuss