

[illegible]

Outline

Ways to generate large amounts of sequence

Understanding the contents of large sequence files

- Fasta format

- Fastq format

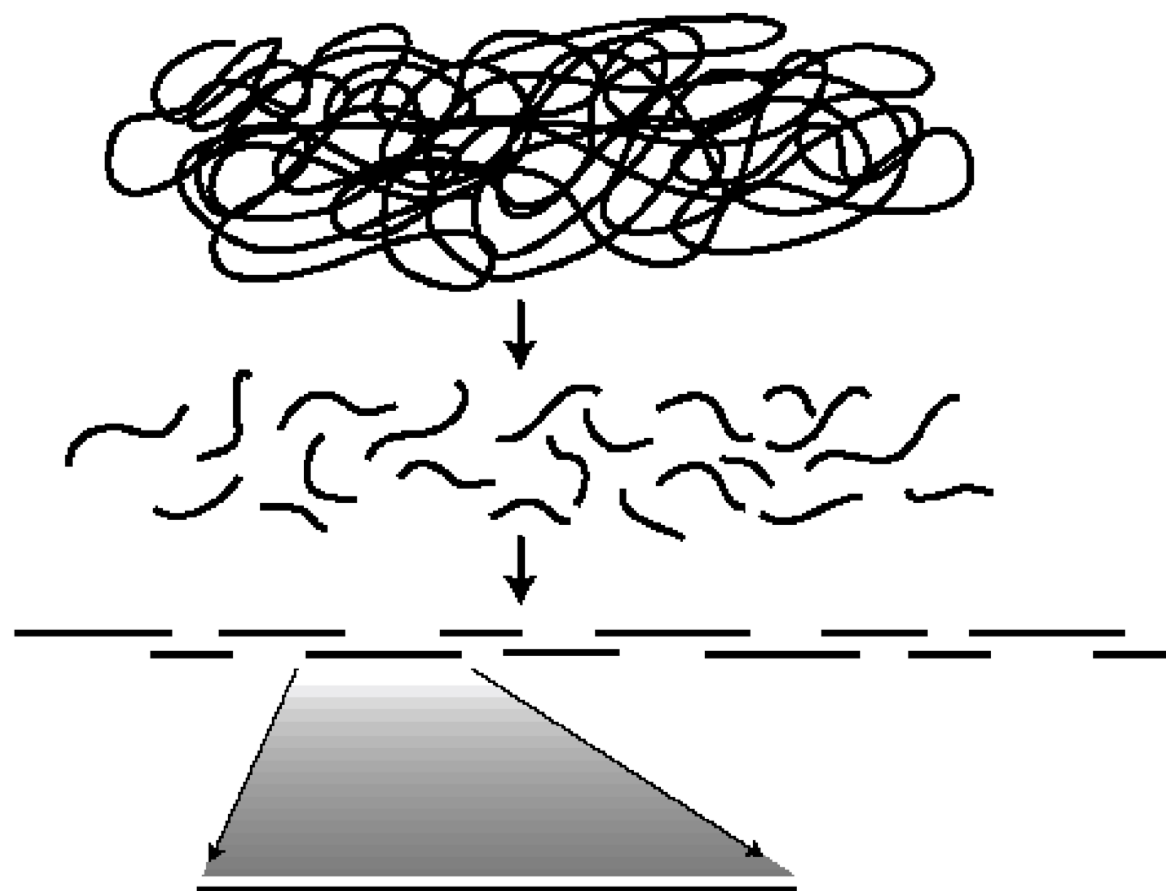
- Sequence quality metrics

- Summarizing sequence data quality/quantity

Using Unix to look at large files

Manipulating large files in Unix





DNA / RNA sequencing

- Sanger

- Long reads (800 bp), high quality
- targeted (primers), slow, expensive, hard to automate

- 454

- Long reads (600-800bp), fairly high quality
- Insertions/deletions, library prep is expensive, not cheap

- Illumina

- Many, many reads, high quality
- Short(ish) – 100bp-250bp

- Ion torrent – error rates, throughput

- PacBio – high error rates (10-15% errors) but very long reads
 - (up to 100kb)

- Oxford nanopore

DNA / RNA sequencing

- Sanger

- Long reads (800 bp), high quality
- targeted (primers), slow, expensive, hard to automate

- 454

- Long reads (600-800bp), fairly high quality
- Insertions/deletions, library prep is expensive, not cheap

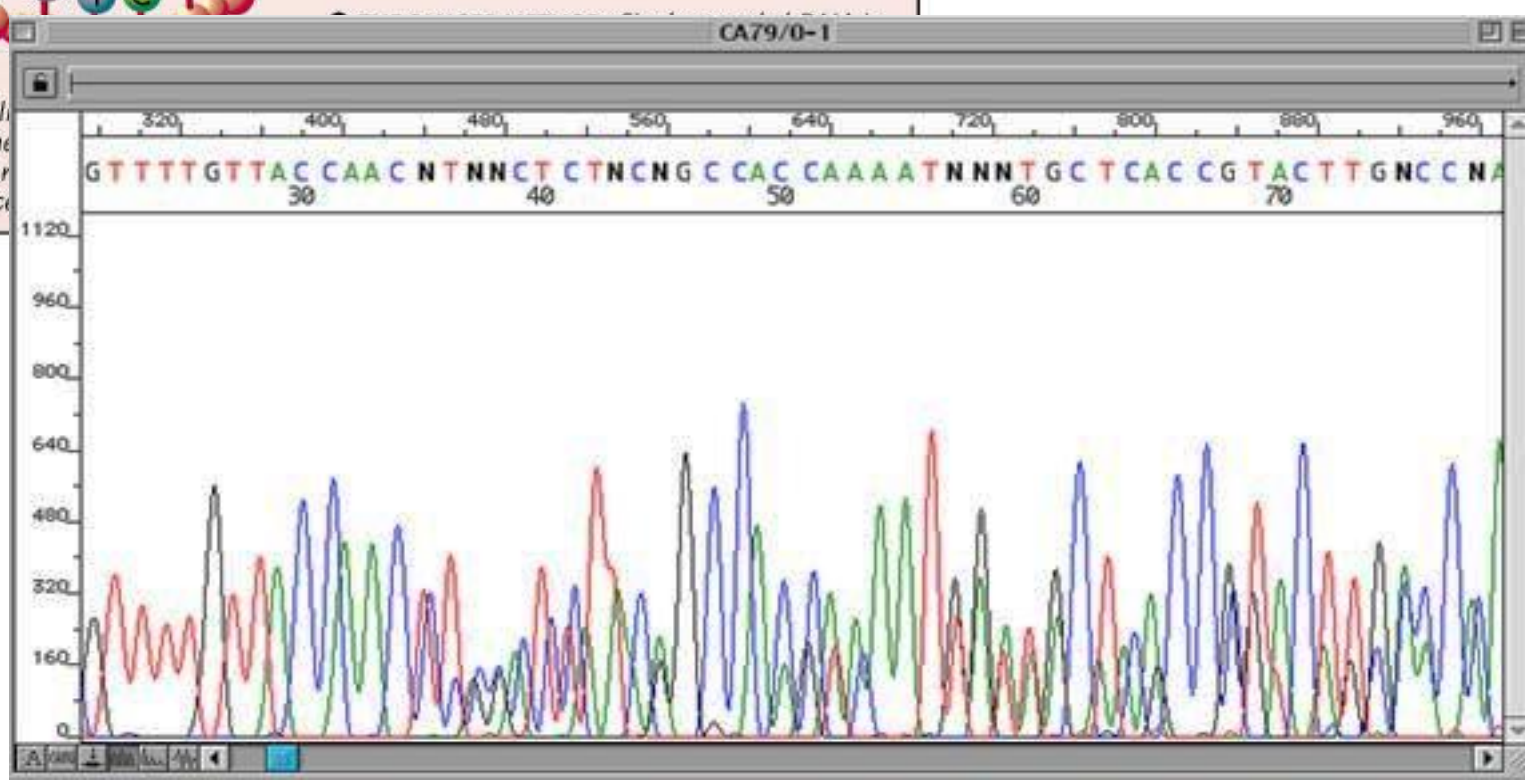
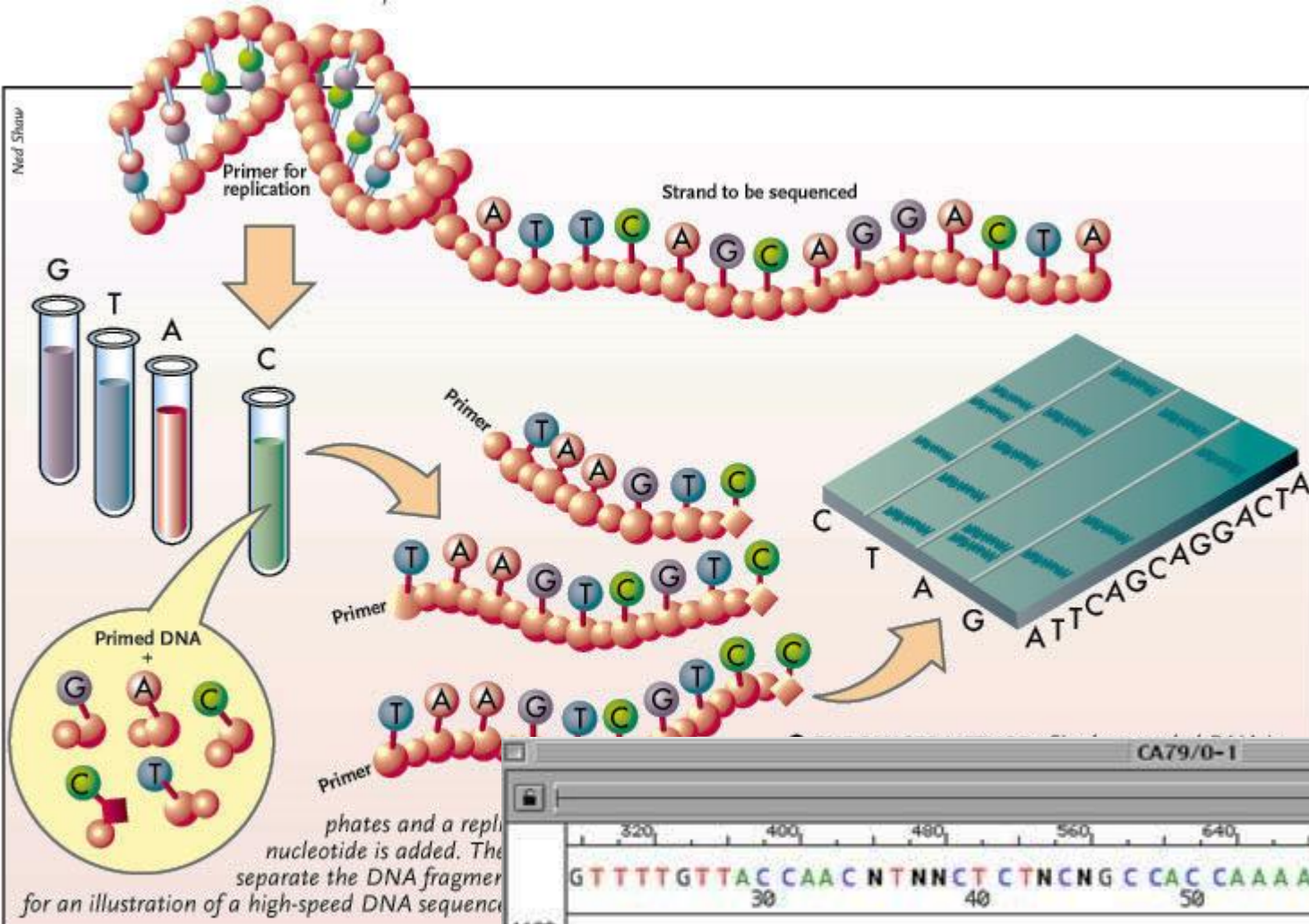
- Illumina

- Many, many reads, high quality
- Short(ish) – 100bp-250bp

- Ion torrent – error rates, throughput

- PacBio – high error rates (10-15% errors) but very long reads
 - (up to 100kb)

- Oxford nanopore



DNA / RNA sequencing

- Sanger

- Long reads (800 bp), high quality
- targeted (primers), slow, expensive, hard to automate

- 454**

- Long reads (600-800bp), fairly high quality**
- Insertions/deletions, library prep is expensive, not cheap**

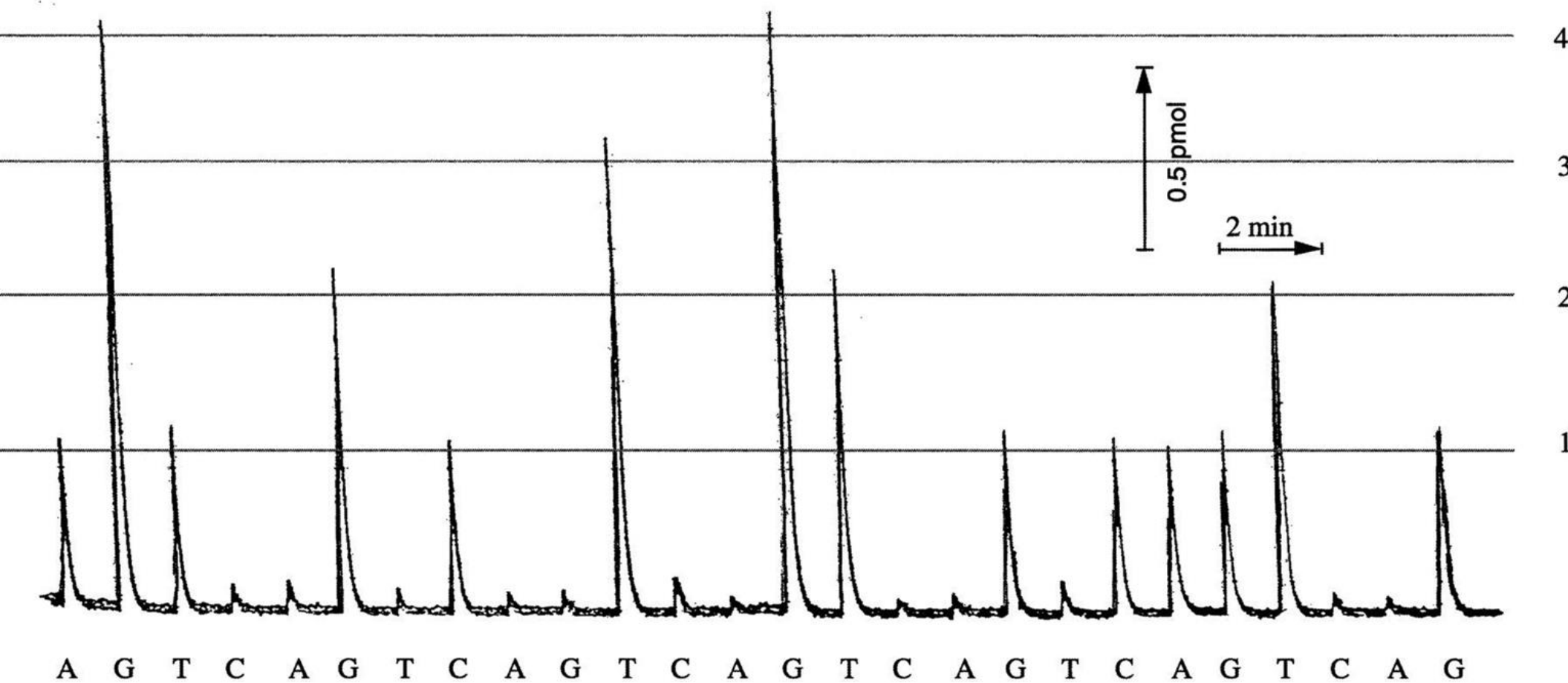
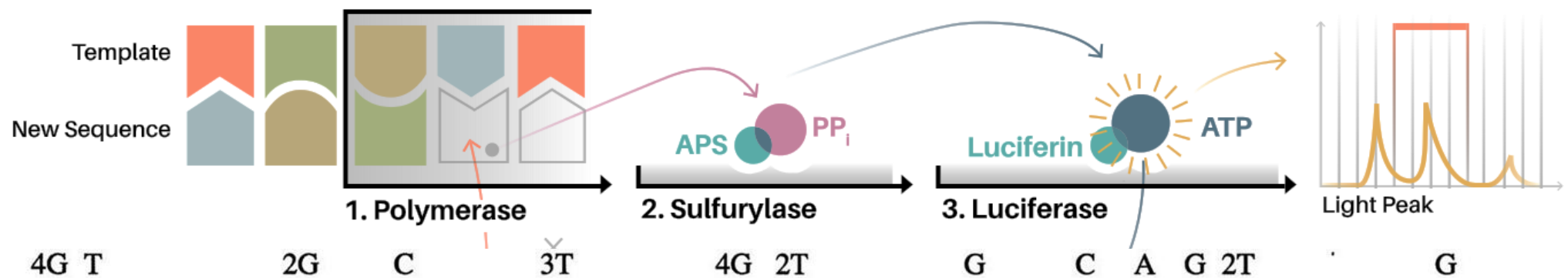
- Illumina

- Many, many reads, high quality
- Short(ish) – 100bp-250bp

- Ion torrent – high error rates, throughput

- PacBio – high error rates (10-15% errors) but very long reads
 - (up to 100kb)

- Oxford nanopore



DNA / RNA sequencing

- Sanger

- Long reads (800 bp), high quality
- targeted (primers), slow, expensive, hard to automate

- 454

- Long reads (600-800bp), fairly high quality
- Insertions/deletions, library prep is expensive, not cheap

- Illumina

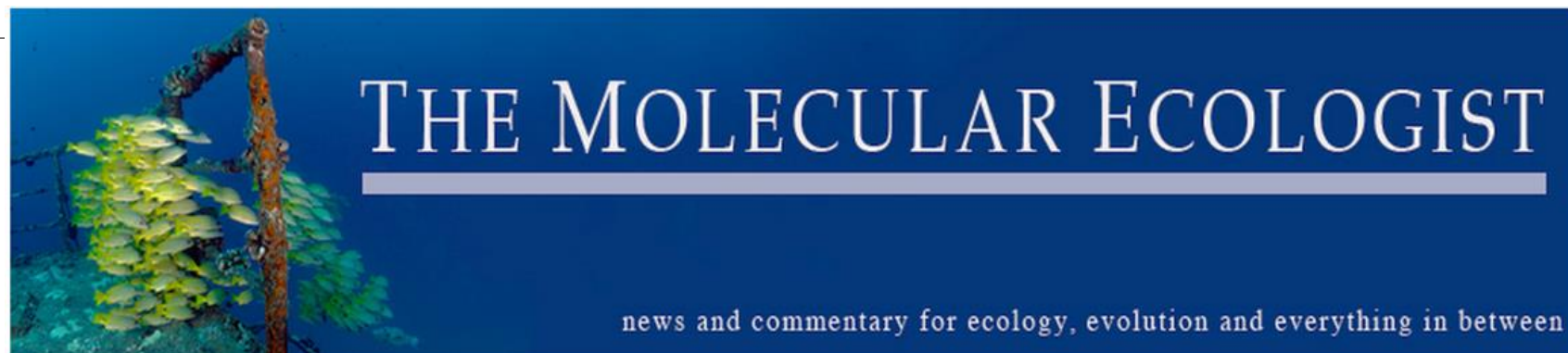
- Many, many reads, high quality
- Short(ish) – 100bp-250bp

- Ion torrent – error rates, throughput

- PacBio – high error rates (10-15% errors) but very long reads
 - (up to 100kb)

- Oxford nanopore

http://www.molecular ecologist.com /next-gen-fieldguide-2014/



[Home](#) [How to ...](#) [NGS Field Guide 2014](#) [About](#) [Legal Info](#) [News](#)


2014 NGS Field Guide: Overview

These pages update the tables presented in [Travis Glenn's \(2011\)](#) "Field Guide to Next Generation DNA Sequencers" for 2014 values. Previous years' tables have been archived: [2011](#), [2012](#), and [2013](#).

Please note that the contents of this guide are the opinion of Travis Glenn, and do not necessarily represent those of any other organisation or person with which he is associated. Neither the other authors of this blog nor John Wiley and Sons are responsible for the accuracy of any of the information supplied by Travis.

- [Table 1a-c](#). "Grades" for common applications on various NGS instruments. Other information from the original table 1 is relatively static.
- [Table 2](#). Run time, Millions of reads/run, Bases/read, and Yield/run for all common commercial NGS platforms (formerly 2a); and reagent costs/run, reagent costs/Mb, and minimum commercially available units for all common commercial

Subscribe in a reader

 [Subscribe to our RSS feed.](#)

Subscribe by e-mail

Enter your email address:

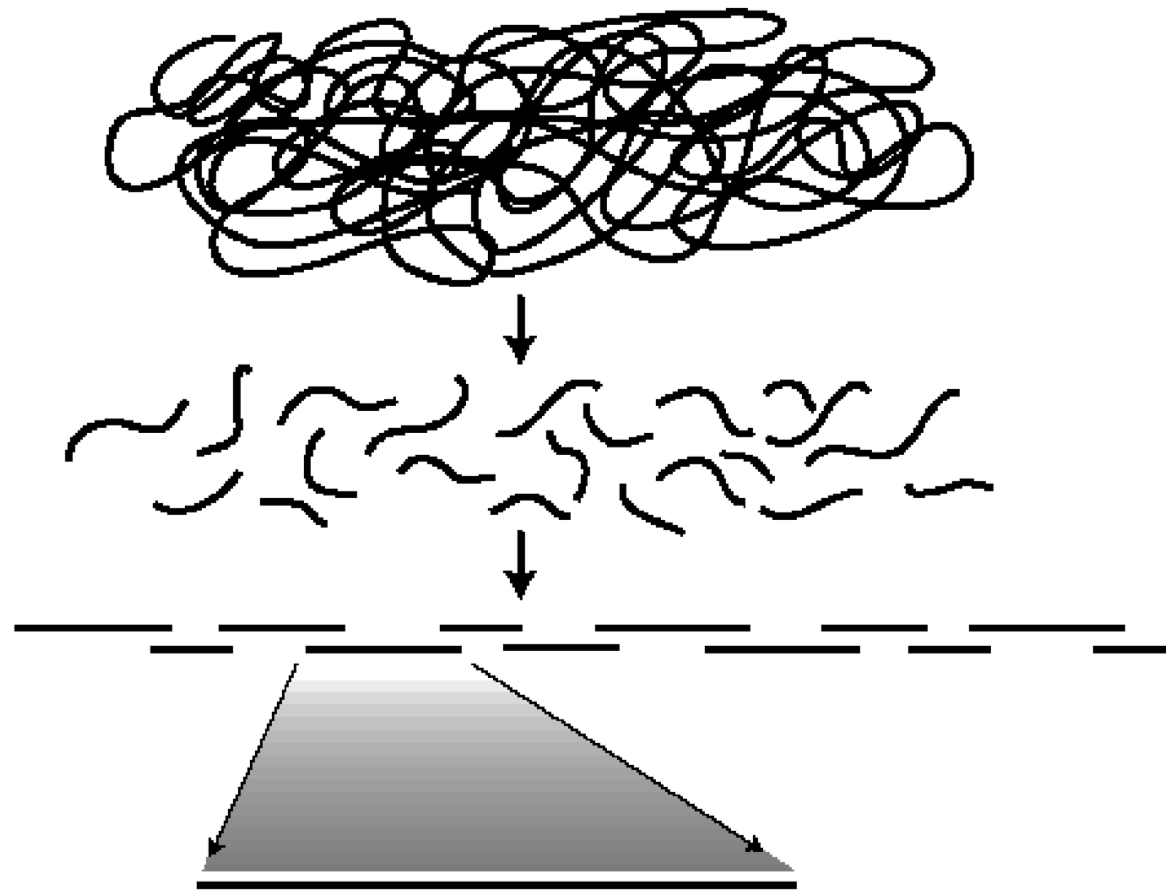
Delivered by [FeedBurner](#)

Latest comments



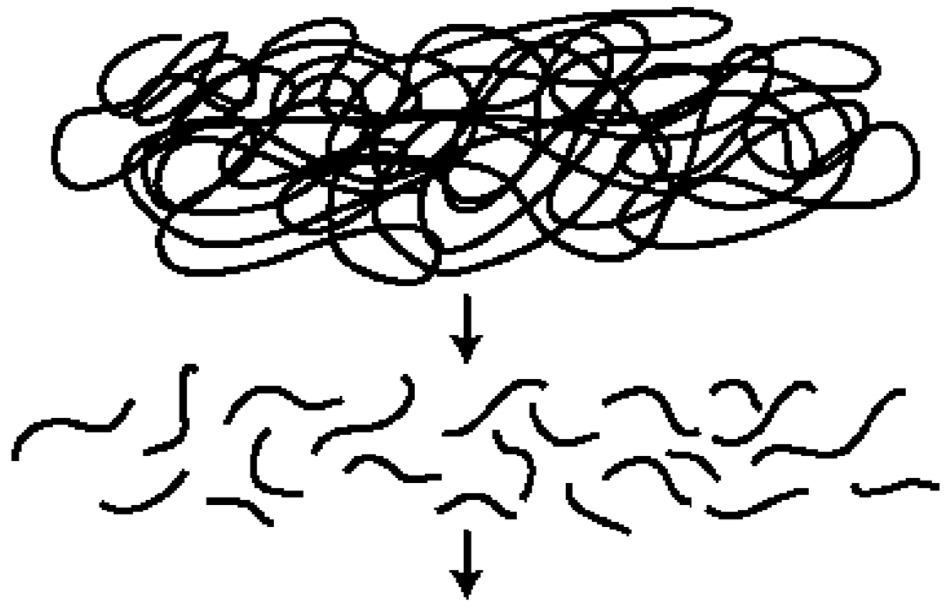
Markku Sorry for a late reply. Didn't see this blog post published for a long time, and drifted to other...

[Random drift and phenotypic](#)



Illumina

- Shear the DNA to specific fragment length
- Ligate on adaptors and barcodes



Shear Genomic DNA or begin with cDNA



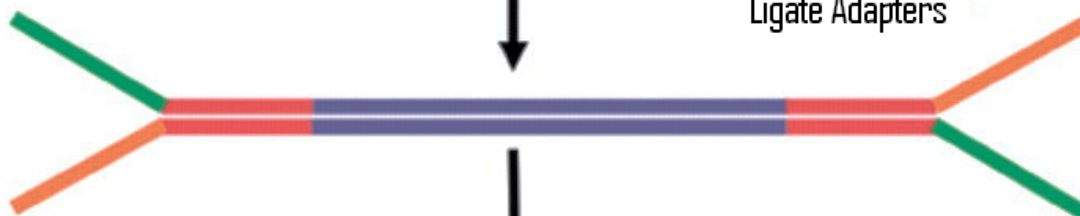
End Repair (Blunt ends)



Add 3' A Tail



Ligate Adapters



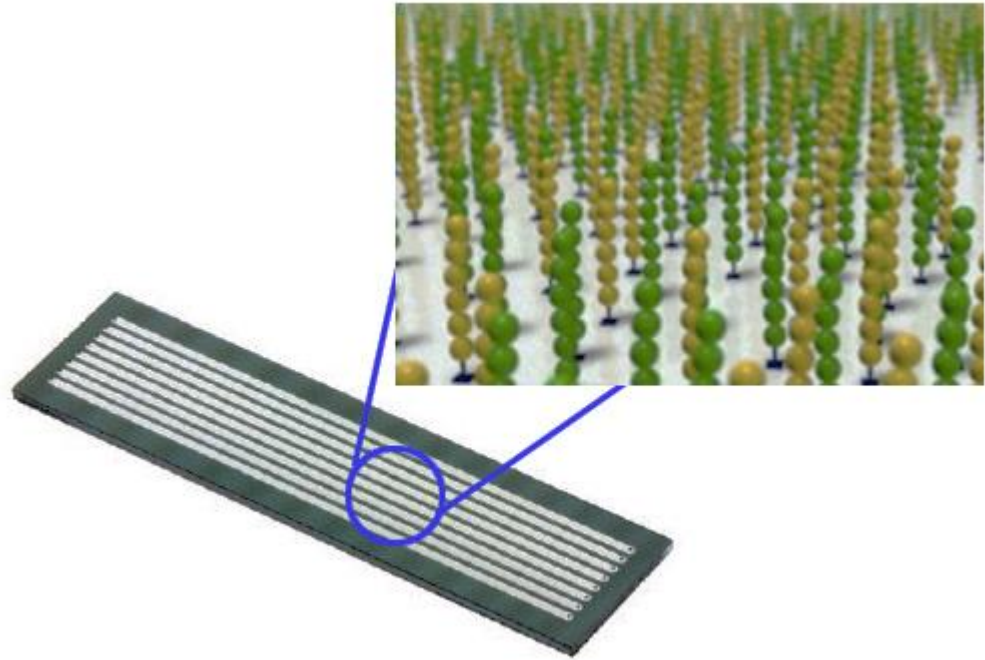
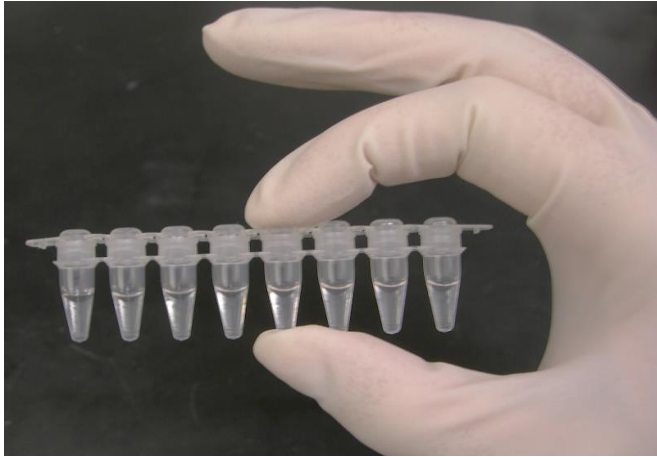
Enrich/Linearize with PCR



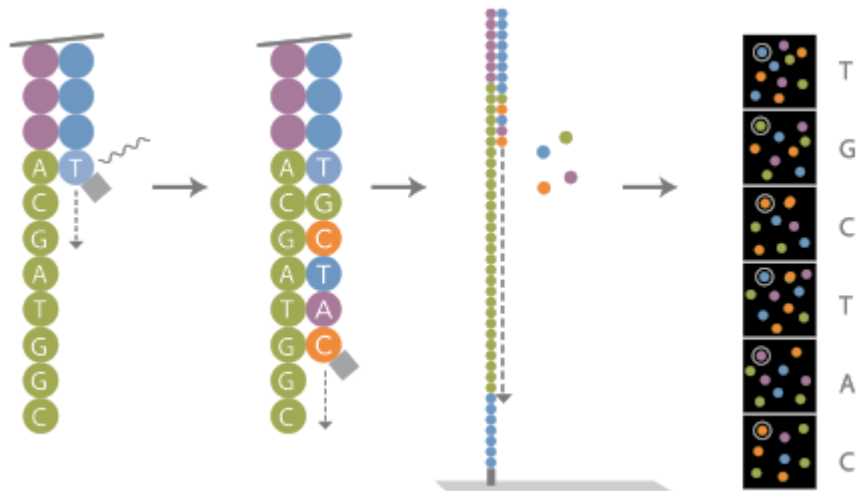
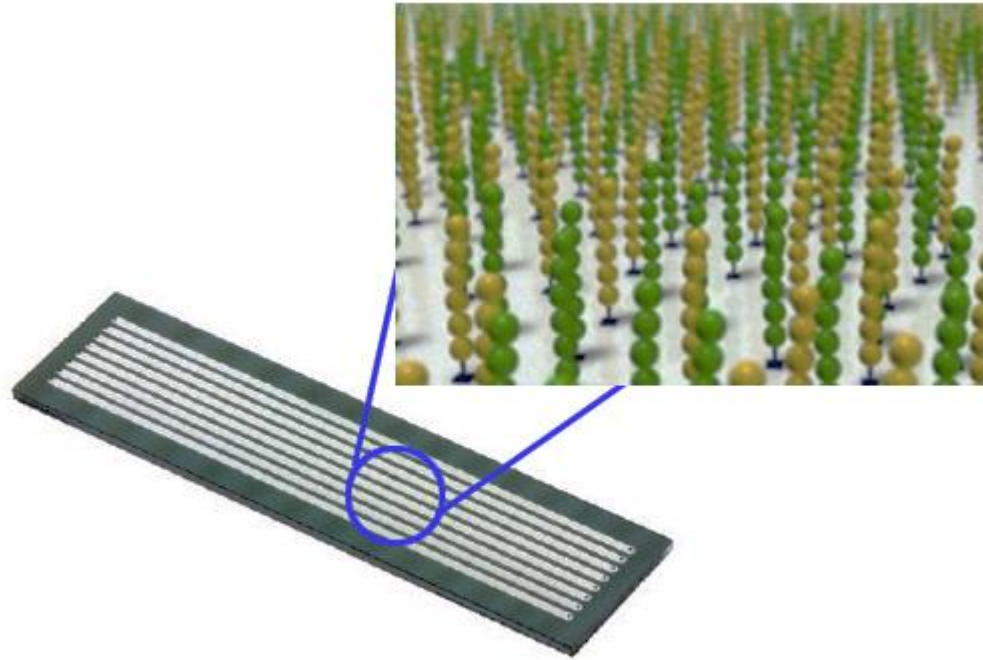
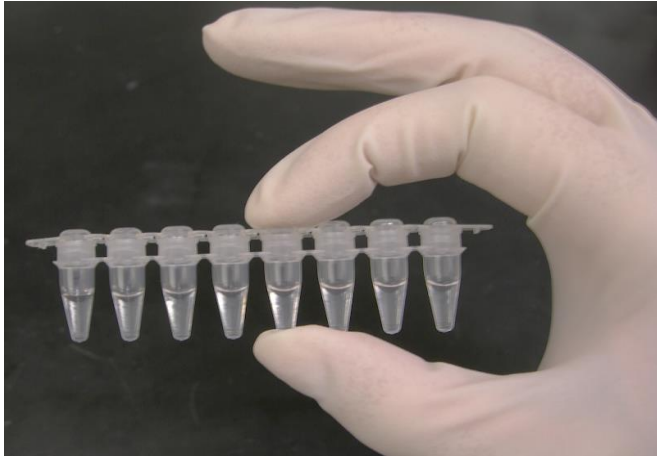
Sequencing

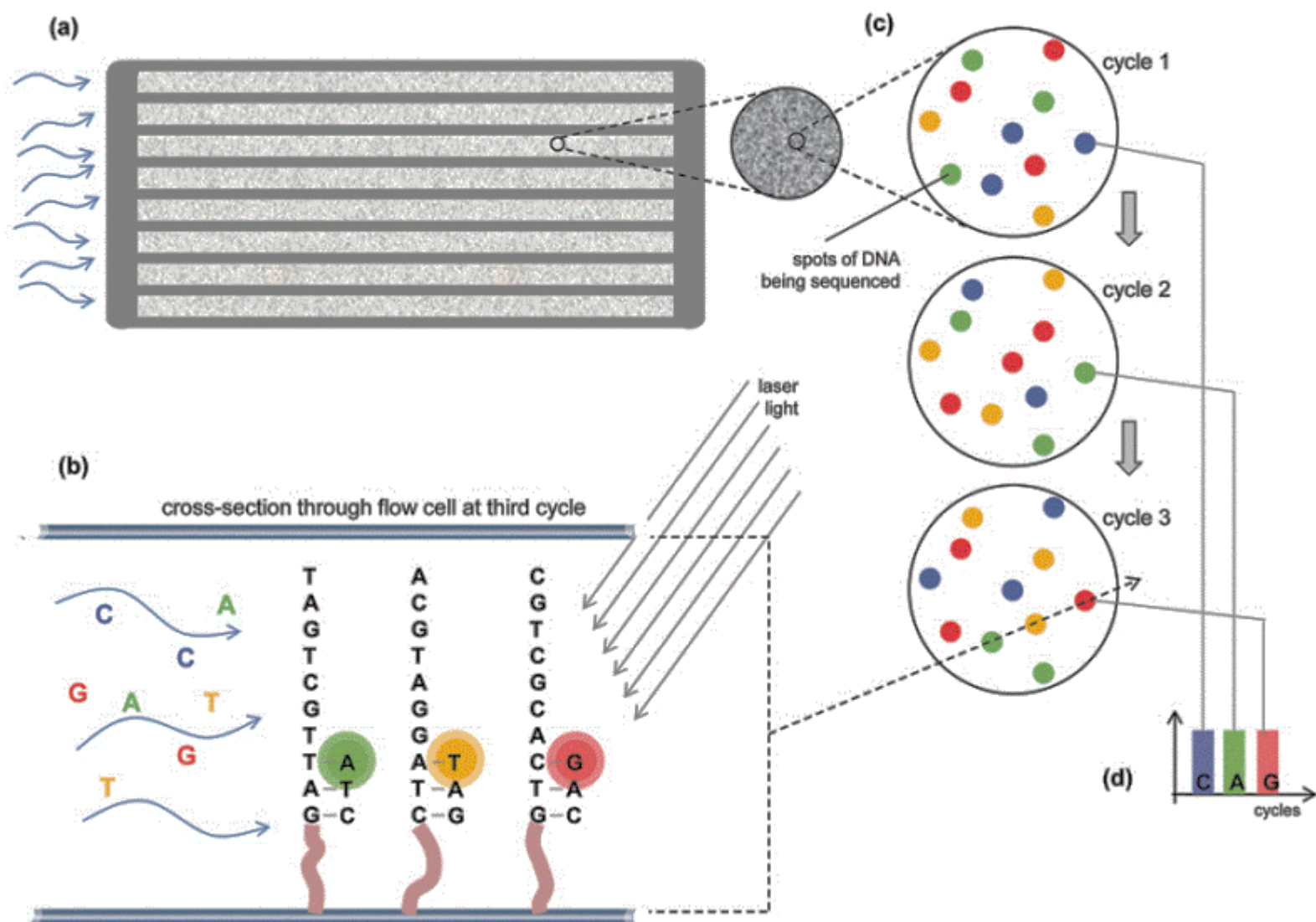


Illumina sequencing

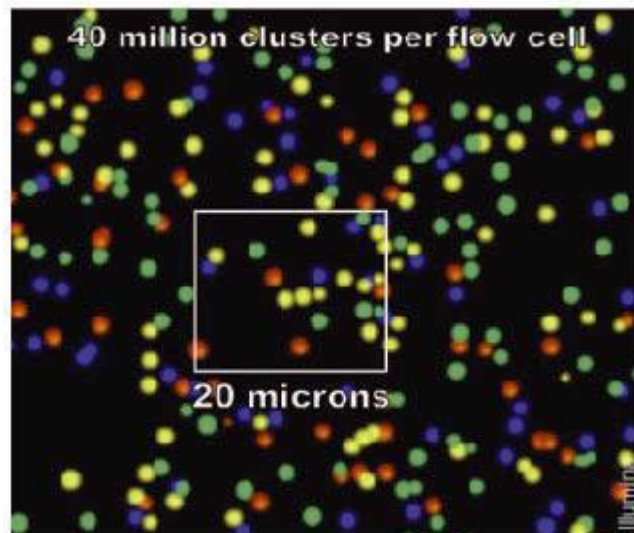
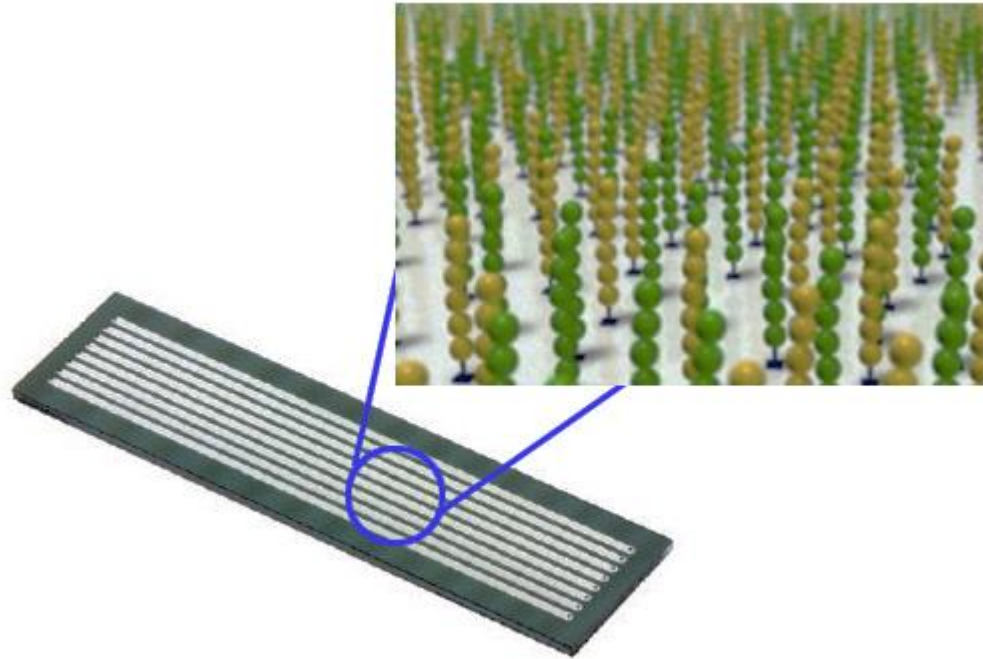
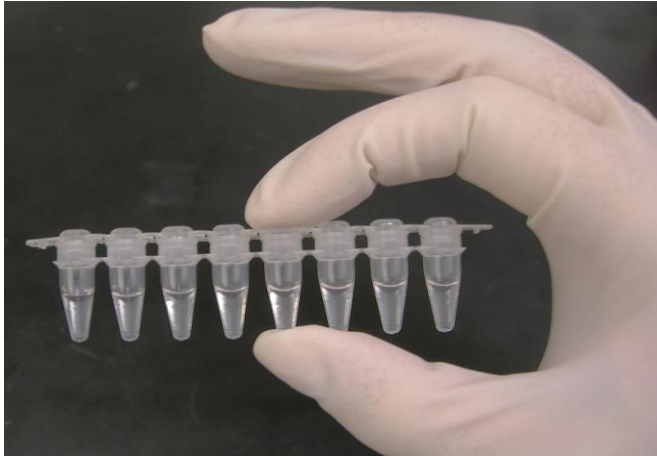


Illumina sequencing





Illumina sequencing



```
@SRR006511.105 8_1_663_27 length=36
ATACCGCACTGTGGTTCGCTTGTCTTGTGATC
+
IIII7II-9/0;+8I<03.+%-,&"+'($,#"'&"
@SRR006511.112 8_1_829_108 length=36
AGAAATTTTATGTATCTGGATGCAATAAAAAATGATG
+
II@IIIIIIIIIIIIIIIIIDII>0<I?>869;64(+%
@SRR006511.490 8_1_351_672 length=36
AGCACCGCGCGTGTGTCCCATGCTCCACACCTCT
+
IO>0A,I2H):$)6)#4$.>'&.$)"%7"1%)&&
@SRR006511.632 8_1_79_187 length=36
ATGCCGAAAGGTATCGGTAAACGTTGAAATTCTTC
+
IIIIII<I;II57G;II.I0**32.--)$32++9),
@SRR006511.726 8_1_300_437 length=36
ACCACTGGACTTCCAGGACCATGAGGCCAAATTGG
+
I1B>:IIII)3,I&0-;,$(%&%1$+1"&($%"&#"
```

@SRR006511.105 8_1_663_27 length=36
ATAGCGGCACTGTTGGTTCGCTTGTTCTTTGAGTC
+
IIII7II-9/0;+8I<03.+%-,&"+'(\$,#"'&"
@SRR006511.112 8_1_829_108 length=36
AGAATTTTATGTATCTGGATGCAATAAAAAATGATG
+
II@IIIIIIIIIIIIIIIDII>0<I?>869;64(+%
@SRR006511.490 8_1_351_672 length=36
AGCACCCGCGTGTGTCCCCCATGCTCCACACCTCT
+
IO>0A,I2H):\$)6)#4\$.>'&.\$)"%7"1%)&&
@SRR006511.632 8_1_79_187 length=36
ATGCCGAAAGGTATCGGTAAACGTTGAAATTCTTC
+
IIIIII<I;II57G;II.I0**32.--)\$32++9),
@SRR006511.726 8_1_300_437 length=36
ACCA CGTGGACTTCCAGGACCATGAGGCCAAATTGG
+
I1B>:IIII)3,I&0-,:\$(%&%1\$+1"&(\$%"&#"

Data formats

- Fasta
- Fastq
- .fastq
- .fq
- .fq.txt
- .fastq.txt
- SAM
- BAM

Basic Unix

- Unix tutorial

- <http://www.ee.surrey.ac.uk/Teaching/Unix/>

- Standard commands:

pwd	print working directory
-----	-------------------------

ls	list contents of working directory
----	------------------------------------

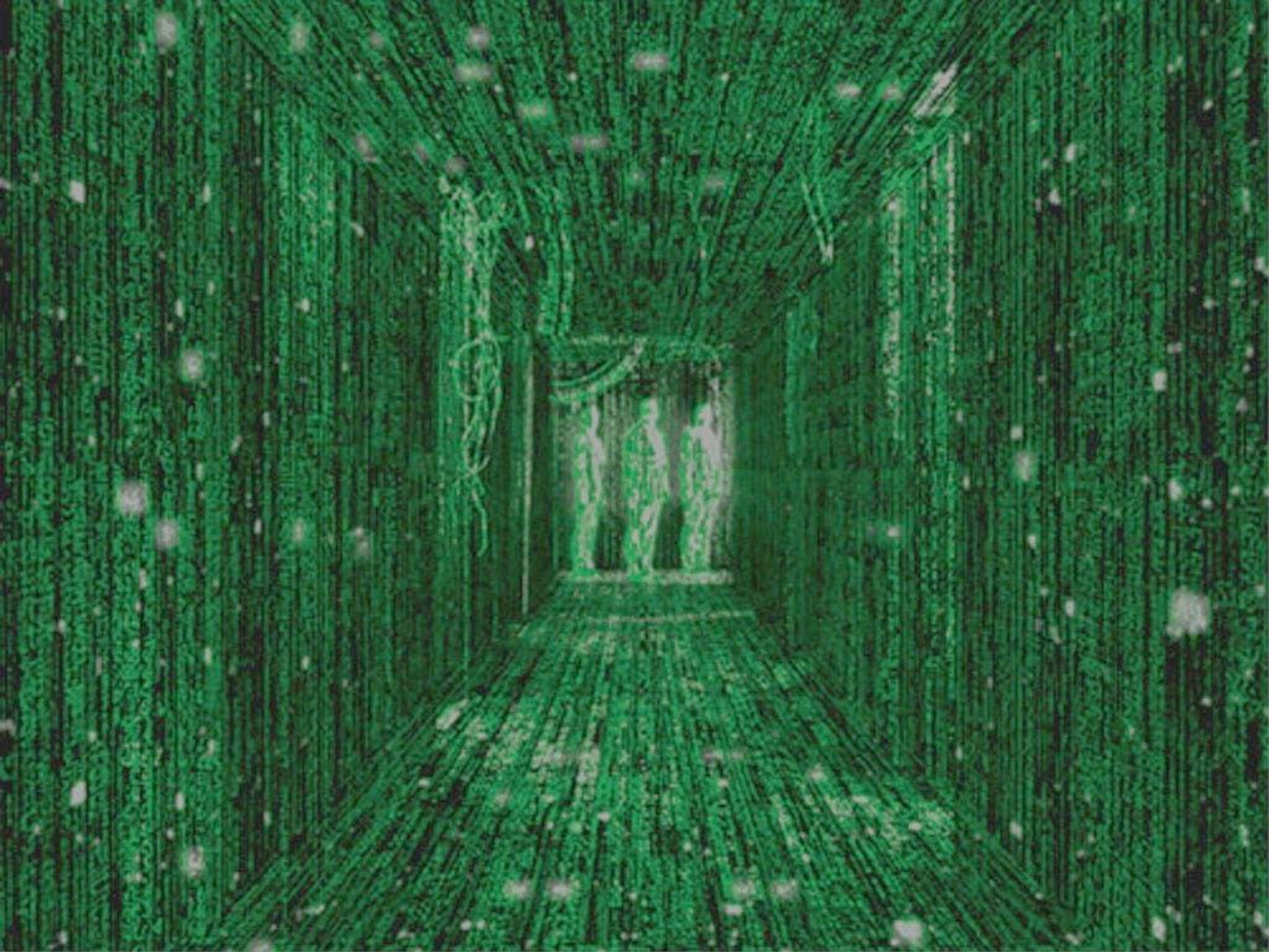
cd	change working directory
----	--------------------------

less	look at a text file
------	---------------------

man	read the manual – how to use a command
-----	---

wget	get a file from another machine
------	---------------------------------

[illegible]



21121 21131 21141 21151 21161 21171 21181 21191 21201 21211 21221 21231 21241 21251 21261 21271 21281
TCAATAGATCTATCTGGTCTGGATACGGTACAGTACAATACGAGACGATGGAATGCTATGGGATGGATGGTAGAGGGATGCCAGCGCCCAAAAGCGATGATTCACTTGTCCCTTGTCCATAGGGACCTCGTGGCATAACAACGAAACGACTCCCGCTAGATAGCGCCCTATCTCT

An example dataset

- These files are already on

- Reference genome – *Arabidopsis* mitochondrion

`wget ftp://ftp.arabidopsis.org/home/tair/Sequences/mitochondrial/mitochondrial_genomic_sequence`

- Illumina sequence for another genotype

`http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=dload&run_list=SRR307232&format=fastq`

(you may have to uncompress this file)

Data formats

- **Fasta**

- Fastq

- .fastq

- .fq

- .fq.txt

- .fastq.txt

- SAM

- BAM

Fasta format

Fasta format

First line: a “>” symbol, and a sequence name

After that – 1 or more lines of sequence data

May have another header and other sequence after that – or many headers and sequences

>Cannabis sativa CBDAS mRNA for cannabidiolic acid synthase, complete cds

```
ATGAAGTGCTCAACATTCTCCTTTTGGTTTGTTTGCAAGATAATATTTTTCTTTTTCTCATTCAATATCC
AAACTTCCATTGCTAATCCTCGAGAAAACCTTCCTTAAATGCTTCTCGCAATATATTCCCAATAATGCAAC
AAATCTAAAACCTCGTATACACTCAAAACAACCCATTGTATATGTCTGTCCTAAATTTCGACAATACACAAT
CTTAGATTCACCTCTGACACAACCCCAAACCACTTGTTATCGTCACTCCTTCACATGTCTCTCATATCC
AAGGCACTATTCTATGCTCCAAGAAAGTTGGCTTGCAGATTCGAACTCGAAGTGGTGGTCATGATTCTGA
GGGCATGTCCTACATATCTCAAGTCCCATTTGTTATAGTAGACTTGAGAAACATGCGTTCAATCAAAATA
GATGTTTCATAGCCAAACTGCATGGGTGAAGCCGGAGCTACCCTTGGAGAAGTTTATTATTGGGTTAATG
AGAAAAATGAGAATCTTAGTTTGGCGGCTGGGTATTGCCCTACTGTTTGCGCAGGTGGACACTTTGGTGG
AGGAGGCTATGGACCATTGATGAGAAACTATGGCCTCGCGGCTGATAATATCATTGATGCACACTTAGTC
```


Now, look at the file mt.fa

How do we look at this sequence?

What do we know about this sequence?

How do you know?

Data formats

- Fasta

- Fastq**

- .fastq

- .fq

- .fq.txt

- .fastq.txt

- SAM

- BAM

Fastq format

4 line repeating pattern:

1. Header line, starting with @
2. DNA sequence (ATGCN)
3. spacer line, starting with +
4. Sequence quality scores

Looking at a fastq file using less

```
@HWI-ST765:7:1101:1318:2091#0/1
GGCCACCTATGACCGGCTCGCGCCGCTCGTCGGGGAGCGGCTGCTCGTCAACGGGGGCGCGCCC GCGGACGCCGTCCGCGGCCCGCTCCGCGCGCCCC
+
_____ccccggggghhhhhh^b^^c__UZFLZWacdBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST765:7:1101:1628:2156#0/1
TCTTCGCGAGTATGTCTGTTGATGGCGCTGTGTCTATCTGCTCAAGGAAAGCAGCCCAACTCAATGTGTTACGCATTAGCGGCATTTGCTACATAATCCG
+
_____eeeeefgggfgf_bddgeafgihdgehghgfeghhhfibgfhhhhhihhhihdhigggfeede`d`]bbdbcccccccccccccccbcb`b`bdbcbce
@HWI-ST765:7:1101:2627:2192#0/1
ATTATGAAGACTGGAGAAGCCCTATATTTATTGTATTTCTTTCTGGATCACAAAATCCTCCCCCTCGAAACAAAAGATGTAGTTGGAATAAATAAAAGG
+
bbbeeeeegfggegghfffeffghiiihiihhhhfghhicégihihhihfhiiiiihfihiifhihhihifdggceeece_bdddbccbcbddbcbcb_
@HWI-ST765:7:1101:3236:2246#0/1
GCGGAAAGAGGGCTTGAGGATGACTTCCTCATAGACTGGGACCCCCACTTTGAGGTGGCTGACGTAGCCTTTAACGGAGTCCCCGCATTCCCGGTATCT
+
bbbeeeeegfggggiihiihiiiiiiihihiiiiiiihihiiiiiiiighhgaggfeeec`cdcccc`bcccc^bccac]aacdcc[_ccd
@HWI-ST765:7:1101:3400:2241#0/1
GCGGACAGCTAATGCGTTCCACTTATTGAACAGGGTTCTATGGTCGGTCCGTGACCCCCGGATGCCGAAGGCGTCCTTGGGGTAATCTCGTAGTTCCTACG
+
_____cacc_eeaegfffZa`e]]de`egdfig[cgffcgZf]e^aX^G[Ze_agfffdgc`bXZ^[]_aaa_GTTTTW_SX`aTX)`_bbaa_aacY`bbRO
@HWI-ST765:7:1101:4139:2060#0/1
NCTTCTCTCTTCATCAGAGAGTAGAGTTGGGGCAATTGTGGGATCACGACGGGGACAGGGGCAGGTGCGGGCGGCGTCTCCGGTTGAGGAAGAGGCTGCC
+
BS\cceeeeggsggggiiiiihifgliiiffhihiiiiiiighiihiiiiiiiggecccccccccccT___acX_c]][)acc_cT[_`bcbaa``caaa^^
@HWI-ST765:7:1101:4188:2089#0/1
ACAAGATATATTTGATATACTAAGATGATAGCTAGAGACTAGAGATGAGAGTGCAGGATCTAGATTGTAAACAAATATTGCACTTTGCTTATGCAAACGTG
+
bbbeeeeegggggiiiiiiiiiiiiiiiiiiiiihfghiiiiihiiiiifghiiiiiiiiihiiiiihiihhihiihgggggeeeeeedddddddcc
@HWI-ST765:7:1101:4440:2112#0/1
GACCTGCTCTGAGCTTCTGGAATGGGTATTAACAAAGGACATAGTCGGATAGGTAAAACCTCTTTTTTCGAGTGAAAGGCCTTATGTTATGAGGGTAA
+
bbbeeeeegffgghchfihiihiifiibbgghiiiiiihhihhifcghiihiihaeggfghhihhhedeeebdd`bbbbccccccccdddddccba^bc
@HWI-ST765:7:1101:5159:2138#0/1
CCCCGAGATGGCCTTCTCGCCGGCGGGTGTTGGGCACGGGCAGCGGTGCGAACACCTGGCCTTCGCTGCTGCGCGGCGCTTCGTTGGTGCGATAGAAGTTG
+
_____eeeeegfeeghfhihhihiihiieRZR[^aabaccc_aaaacTY_aacaccabccacccc_acca_]XX_]([)]_aR^ba^abaX`[_YbbYY]
@HWI-ST765:7:1101:5364:2245#0/1
CCCTTTCCGCCTAACCATTTGTTTAGAAAGTAAAAAAAGAATTCTTAAATTGTTAGACTAACTTTGTTCTTCGACTTCACTTTGTCTTCGTTTAGTCCA
+
bbbeeeeeggggfiyiiiiiiiihiiiiifgiyiiiiiiiiiiiihiiiiigliifgggggeeeeeeddddb^acccccccccccccccbccaccccc
@HWI-ST765:7:1101:5707:2110#0/1
CTGCGCGCCCATCTCGGCTTGATCTTCTTGGCCATCGCGCGGCGCAGCAGGTGGCGCGCGCCAGCGAGTAACCCGCCAGGATCTGGCGATCTGCATCA
+
bbbeeeeefggggiiiiiihiiiiiihiihiihibgf`ggecaccaccc^bccccaccccccc_]aT`bbccaacaaa^bcbcbcaaX]acccbcc`
@HWI-ST765:7:1101:6179:2187#0/1
TTCTAGTAGTGCAAAACACATATGTTTCTAGAGGTGGCAAAACATGAATTTGGGTCAAAGGTCCCTTTTGGTGCGGACCCAAAATACITTTTGGCTTGGGA
```

Fastq ASCII quality scores

<http://www.asciitable.com/>

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

Source: www.LookupTables.com

Illumina quality scores

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

- **Sanger format**

- 0 to 93 using ASCII 33 to 126

- **Solexa/Illumina 1.0 format**

- 5 to 62 using ASCII 59 to 126

- **Illumina 1.3+ format**

- 0 to 62 using ASCII 64 to 126

- **Illumina 1.5+**

- 0 and 1 are no longer used and the value 2, encoded by ASCII 66 "B", is used also at the end of reads as a Read Segment Quality Control Indicator [6].

Fastq ASCII quality scores

<http://www.asciitable.com/>

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

Source: www.LookupTables.com

Looking at a fastq file using less

```
@HWI-ST765:7:1101:1318:2091#0/1
GGCCACCTATGACC GGCTCGCGCCGCTCGTCGGGGAGCGGCTGCTCGTCAACCGGGGCGCGCCCCGCGGACGCCGTCCGCGGCCCGCTCCGCGCGCCCC
+
_____ccccggggghhhhhh^b^^c__UZFLZWacdBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST765:7:1101:1628:2156#0/1
TCTTCGCGAGTATGTCTGTTGATGGCGCTGTGTCCTATCTGCTCAAGGAAAGCAGCCC AACTCAATGTGTACGCATTAGCGGCATTTGCTACATAATCCG
+
_____eeeeefgggfgf_bddgeafgihdgehghgefeghhhfibgfhhhhhhhihhdhigggfeede`d`]bbdbcccccccccccccccbcb`b`bdbcbce
@HWI-ST765:7:1101:2627:2192#0/1
ATTATGAAGACTGGAGAAAGCCCTATATTTATTGTATTTCTTTCTGGATCACAAAATCCTCCCCCTCGAAACAAAAGATGTAGTTGGAATAAATAAAAGG
+
bbbeeeegfggggfhffffefghiiiiihiihhffghhicégihiihhihfhiiiiifhiifihihihihfdggeceecce_bdddbccbcbddbcbcb_
@HWI-ST765:7:1101:3236:2246#0/1
GCGGAAAGAGGGCTTGAGGATGACTTCCTCATAGACTGGGACCCCCACTTTGAGGTGGCTGACGTAGCCTTTAACCGAGTCCCCGCATTCCCGGTATCT
+
bbbeeeegfggggiiihiihiiiiiiihiiiiiiiiihiiiiiiiiihghhgaggfeeeeec`cdcccc`bcccc^bccac]aacdccc[_ccd
@HWI-ST765:7:1101:3400:2241#0/1
GCGGACAGCTAATGCGTTCCACTTATTGAACAGGGTTCTATGGTCGGTCCGTGACCCCCGGATGCCGAAGGCGTCCTTGGGGTAATCTCGTAGTTCCTACG
+
_____cacc_eeaegfffZa`e]]de`egdfig[cgffcgZf]e^aX^G[Ze_agffddgc`bXZ^[]_aaa_GTTTTW_SX`aTX)`_bbaa_aacY`bbRO
@HWI-ST765:7:1101:4139:2060#0/1
NCTTCTCTCTTCATCAGAGAGTAGAGTTGGGGCAATTGTGGGATCACGACGGGGACAGGGGCAGGTGCGGGCGGCGTCTCCGGTTGAGGAAGAGGCTGCC
+
BS\cceeeeggsgggiiiiihifgliiiffhiiiiiiighiiihiiiiiiiggeccccccccccT___acX_c]][[acc_cT[_`bcbaa` `caaa^^
@HWI-ST765:7:1101:4188:2089#0/1
ACAAGATATATTTGATATACTAAGATGATAGCTAGAGACTAGAGATGAGAGTGACAGGATCTAGATTGTAAACAAATATTCGACTTTGCTTATGCAAAGTGT
+
bbbeeeegggggiiiiiiihifgliiiffhiiiiiiighiiihiiiiifghiiiiiiihiiiiihiiihhihghgggeeeeeeddddddc
@HWI-ST765:7:1101:4440:2112#0/1
GACCTGCTCTGAGCTTCCTGGAATGGGTATTAACAAAGGACATAGTCGGATAGGTAAAACCTCTTTTTTCGAGTGAAAGGCCTTATGTTATGAGGGTAA
+
bbbeeeegffgghchfihihiifiibbgghiiiiihhihhihfchgiihihhaeggfghhihhhedeeebdd`bccccccccccdddddccba^bc
@HWI-ST765:7:1101:5159:2138#0/1
CCCCGAGATGGCCTTCCTCGCCGGCGGGTG TGGGCACGGGCAGCGGTGCGAACACCTGGCCTTCGCTGCTGCGCGGCGCTTCGTTGGTGCGATAGAAGTTG
+
_____eeeeegfeeghfhihhhiihiiieRZR[^aabaccc_aaaacTY_aacaccabccacccc_acca_]XX_] [[]_]aR^ba^abaX`[_YbbYY]
@HWI-ST765:7:1101:5364:2245#0/1
CCCTTTCGCGCTAACCATTTGTTTAGAAAAGTAAAAAAGAATTCTTAAATTGTTAGACTAACTTTGTTCTTCGACTTCACTTTGTCTTCGTTTAGTCCA
+
bbbeeeeggggfiyiiiiihiiiiifgiyiiiiihiiiiihiiiiigliifgggggeeeeeeddddb^acccccccccccccccbccaccccc
@HWI-ST765:7:1101:5707:2110#0/1
CTGCGCGCCCATCTCGCCTTGATCTTCTT GCCATCGCGCGGCGCAGCAGGTGCGCGCGGCCAGCGAGTAACCGCCAGGATCTGGCGATCTGCATCA
+
bbbeeeefggggiiiiiiihiiiiihiiiiihibi b g f`ggegaccaccc^bccccaccccccc_]aT`bbccaacaaa^bcbcbcaaX]acccbcc`
@HWI-ST765:7:1101:6179:2187#0/1
TTCTAGTAGTGCAAAACACATATGTTTCTAGAGGTGGCAAAACATGAATTTGGGTCAAAGGTCCCTTTTGGTGCGGACCCAAAATACITTTTGGCTTGGGA
```

Examining the data

Look at the fastq file -

Less SRR307232.fastq

Please work in groups of 2-4 again, and figure out – which quality score type is this?

Illumina quality scores

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

- **Sanger format**

- 0 to 93 using ASCII 33 to 126

- **Solexa/Illumina 1.0 format**

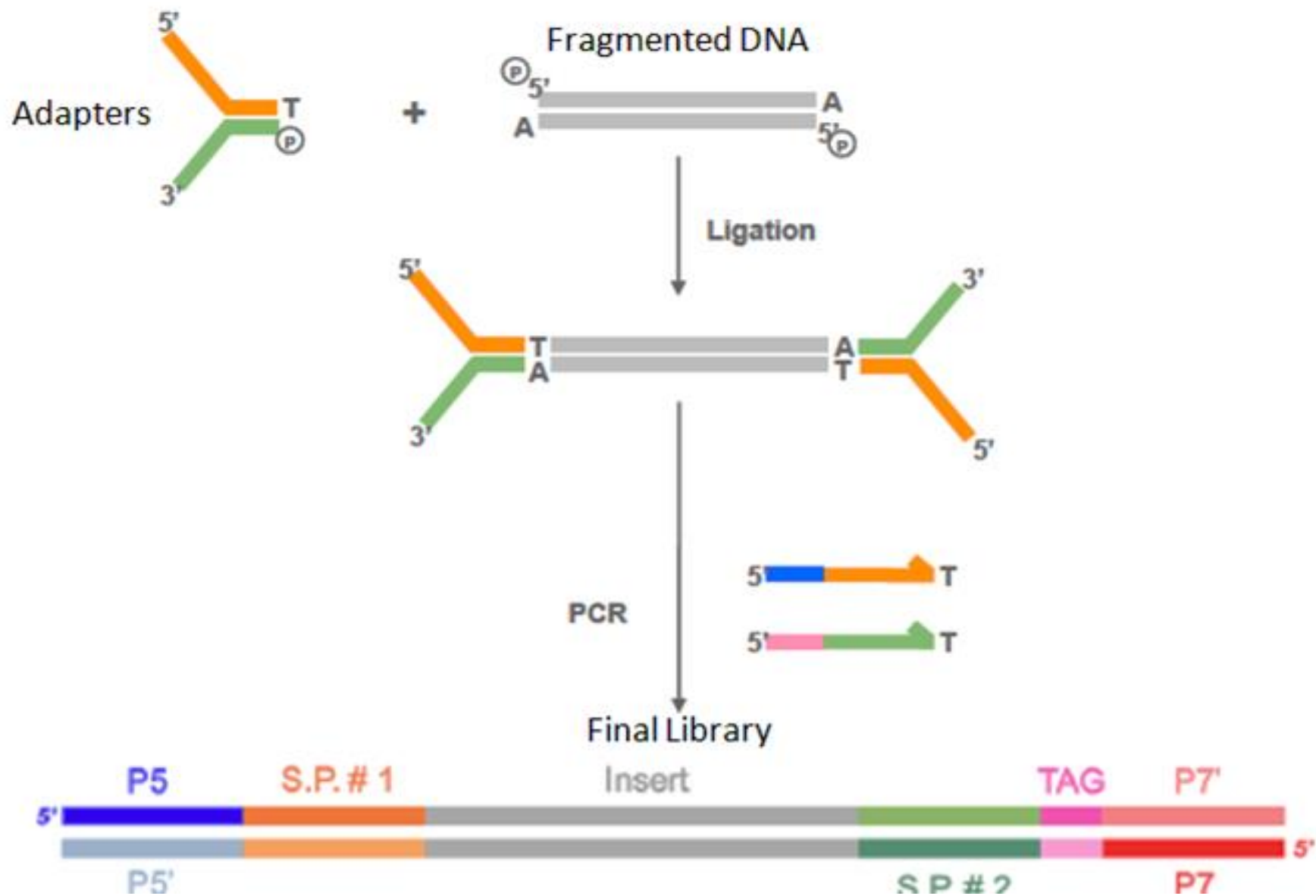
- 5 to 62 using ASCII 59 to 126

- **Illumina 1.3+ format**

- 0 to 62 using ASCII 64 to 126

- **Illumina 1.5+**

- 0 and 1 are no longer used and the value 2, encoded by ASCII 66 "B", is used also at the end of reads as a Read Segment Quality Control Indicator [6].



Evaluating quality

Fastqc

- a good program for quality metrics

```
fastqc sra_data.fastq
```

For Thursday

Do modules 3-4 in the tutorial

Read up on fastq files:

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

Summary

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

Basic Statistics

Measure	Value
Filename	sra_data.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	8507009
Filtered Sequences	0
Sequence length	35
%GC	55

Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

Position in read (bp)

FastQC Report

Tue 1 Sep 2015
sra_data.fastq

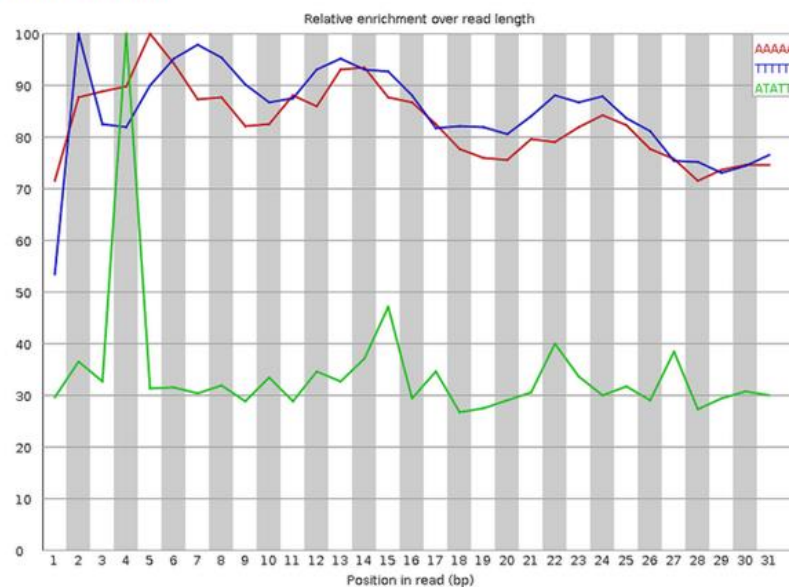
Summary

- ✓ 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

Overrepresented sequences

No overrepresented sequences

Kmer Content



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
AAAAA	500185	3.4980087	4.208535	5
TTTTT	463060	3.2985418	3.8755615	2
ATATT	259480	1.8348091	5.3278956	4