

Introduction to Next Generation Sequencing (NGS)

Hugh Patterton

hpatterton@sun.ac.za

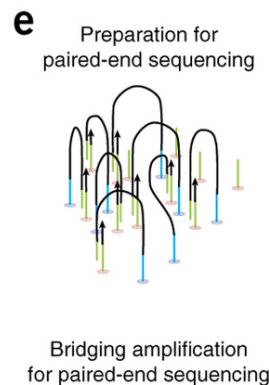
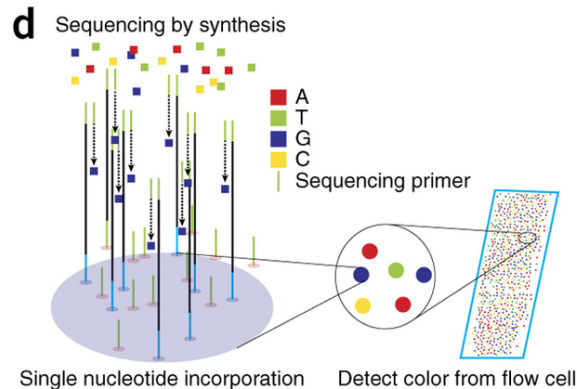
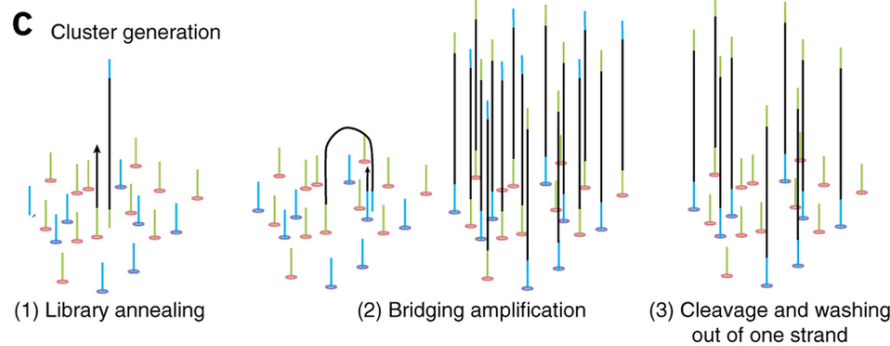
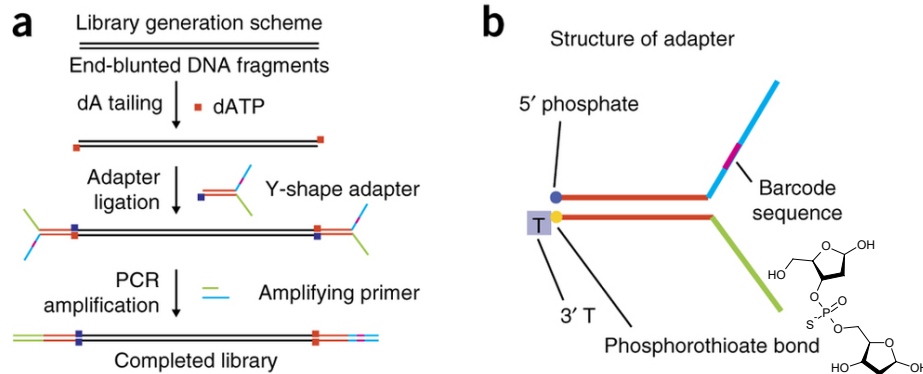
Room 2010 AI Perold Building

Many different NGS techniques

- Sanger sequencing sequenced one molecule per 4 reaction (= 4 lanes on a sequencing gel)
- NGS is highly parallelised, i.e., sequencing millions of molecules simultaneously
- One run takes hours to days
- Samples can be multiplexed
- Human Genome Project cost ~\$3B over 15 years
- Can obtain whole genome sequences today for ~\$1K

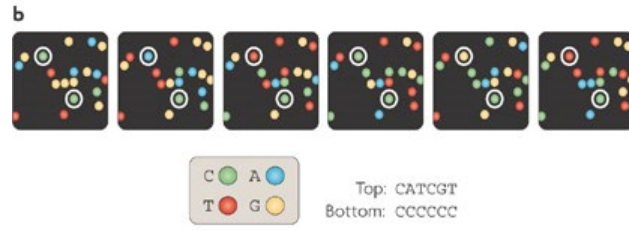
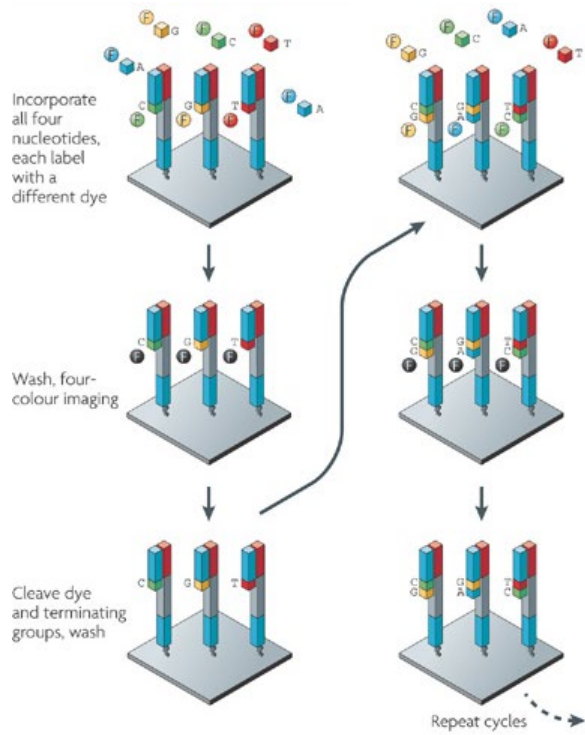
Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase (US\$)	Accuracy (%)
Roche 454	700	1 million	1 day	10	99.90
Illumina	50–250	<3 billion	1–10 days	~0.10	98
SOLiD	50	~1.4 billion	7–14 days	0.13	99.90
Ion Torrent	200	<5 million	2 hours	1	98
Pacific Biosciences	2900	<75,000	<2 hours	2	99
Sanger	400–900	N/A	<3 hours	2400	99.90

Illumina: Cyclic Reversible Termination

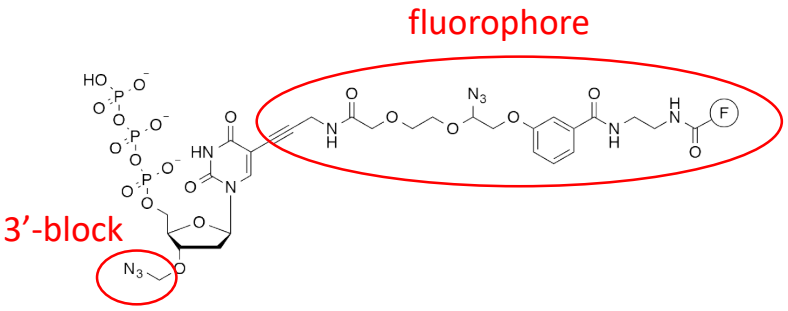


- Randomly fragment the DNA (genome) sample
- Ligate a Y-adapter to the DNA fragments
- Adapter can contain barcode sequence for multiplexing
- Amplify DNA
- Anneal single strands to flow-cell surface
- Bridge amplify the generate clusters
- Wash out one strand
- Perform sequencing with fluorescent nucleotides
- Perform bridge amplification to regenerate clusters, wash out "other" strand
- Sequence from other end of original fragment (paired-ends)

Reversible terminators



Metzker 2010



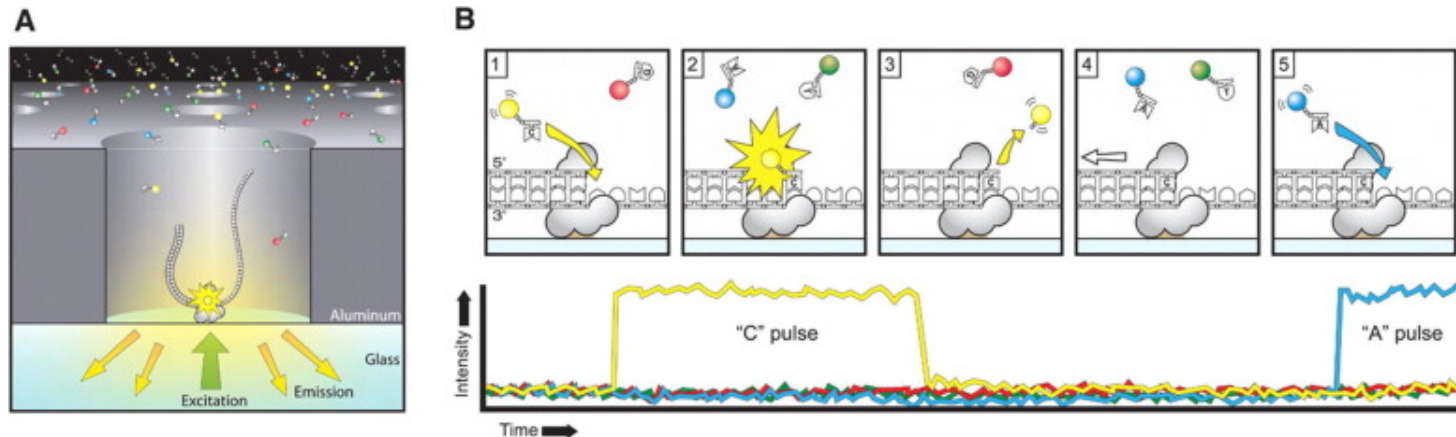
Structure of fluorescently labelled dTTP used in Illumina sequencing

1. Sequencing primer extended in the presence of fluorescent dNTPs
2. Remove unincorporated nucleotides
3. Record image of clusters
4. Cleave of fluorophore and 3'-block using tris(2-carboxyethyl)phosphine (TCEP)
5. Repeat from step 1
6. The series of colours for each cluster is the sequence of that cluster
7. Can re-synthesize strand and sequence from other end (paired ends)

PacBio Single Molecule Real Time



- Hairpin adaptors (green) are ligated to the ends of a double-stranded DNA molecule (yellow and purple), forming a closed circle (= SMRTbell)
- The polymerase (gray) is anchored to the bottom of a zero-mode waveguide (ZMW) and incorporates bases into the read strand (orange)
- Each single molecule real time (SMRT) cell contains 150,000 ZMWs
- Approximately 35,000–75,000 of these wells produce a read in a run lasting 0.5–4 h, resulting in 0.5–1 Gb of sequence



- A SMRTbell (gray) diffuses into a ZMW, and the adaptor binds to a polymerase immobilized at the bottom
- Incorporation by the polymerase of the fluorescent nucleotide places it at the bottom of the well, allowing detection of fluorescence
- Each of the four nucleotides is labeled with a different fluorescent dye (indicated in red, yellow, green, and blue)


There any many NGS technologies

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US\$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo- polymer repeats	Bacterial and insect genome <i>de novo</i> assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA _{II}	Frag, MP/ solid-phase	RTs	75 or 100	4 [‡] , 9 [§]	18 [‡] , 35 [§]	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7 [‡] , 14 [§]	30 [‡] , 50 [§]	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non- cleavable probe SBL	26	5 [§]	12 [§]	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8 [‡]	37 [‡]	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. [‡]Fragment run. [§]Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Typical NGS pipeline

- A pipeline is a series of programs or scripts, used consecutively, to perform a complete analysis

Stage	Examples/explanation	File formats
Laboratory work	Experimental design Library preparation Enrichment (capture)	
Next-generation sequencing	Platforms include Illumina, SOLiD, Pacific Biosciences, other	Output: FASTQ-Sanger, FASTQ-Illumina
Analysis pipeline	Quality assessment Trimming, filtering Software: FastQC	FASTQ  Typical starting data file
	Alignment to reference genome Software: BWA, Bowtie2	Reference: FASTA Output: SAM/BAM
	Variant identification Single nucleotide variants (SNVs), structural variants (e.g. indels) Software: GATK, SAMTools Realignment, recalibration	Variant Call Format (VCF/BCF)
	Annotation Comparison to public database (dbSNP, 1000 Genomes); functional consequence scores	
Visualization	Variant visualization; read depth; comparison to other samples Software: IGV, BEDTools, BigBED	
Prioritization	Discovery of relevant variants Software: PolyPhen-2, VEP, VAAST	VCF
Storage	Deposit data in ENA, SRA, dbGaP	BAM, VCF

FASTQ file format

```
@HWI-D00466:62:C6UETANXX:5:1101:1270:2163 1:N:0:CGATGT
TGTCAGTATAAAAAAATTTTCCGCAGGATATAGAAAAAAGAAATGAAATTATAGTAGCGGTTATTTCCGTGGGGTGCTTTTTTACACCTGTACATCTGTT
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00466:62:C6UETANXX:5:1101:1332:2180 1:N:0:CGATGT
GCCAATGAAGAAAATACGATGAAACCATGGCAGATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTGAAAAAAA
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00466:62:C6UETANXX:5:1101:4490:2229 1:N:0:CGATGT
GAATCTTATTATTTTCTTTATTTAAATTTATAAAAAATATAAAGTCCCCGCCCCCTTTTTATTTTATTTAATTAAGAAGGTATTTTAAAAAAGGAGTGAGGGA
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

⋮

The FASTQ file is a **text file** composed of blocks of text, with each block containing the following:

1. **@Sequence name**
2. **Sequence**
3. **+Sequence name (may be absent)**
4. **Phred quality score**

Phred score

Phred was an early base calling program written by Phil Green of University of Washington Genome Center, and was probably derived from “Phil’s Read Editor”

The **Phred score (Q)** is a quantity, where the **probability (p)** that the assignment is incorrect is given by:

$$p = 10^{\frac{-Q}{10}}$$

∴ if $Q = 30$, $p = 10^{\frac{-30}{10}} = 10^{-3} = 0.001$

∴ a 1 in 1000 chance that the assigned base is incorrect

Why are Phred scores given as **letters**?

The classic 7 bit (0-127) ASCII table

The first “real” character (!) has an ASCII value of 33

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

ASCII = American Standard Code for Information Interchange

Phred score table

Accuracy (p)	Phred score (Q)	Q+33	ASCII code
1.0	0	33	!
0.1	10	43	+
0.01	20	53	5
0.001	30	63	?
~0.0002	37	70	F
↓			
$10^{-9.3}$	93	126 (127 is "DEL")	~

```
@HWI-D00466:62:C6UETANXX:5:1101:1270:2163 1:N:0:CGATGT
TGTCAGTATAAAAAAAAAATTTCCGCAGGATATAGAAAAAAAAAGAAATGAAATTATAGTAGCGGTTATTTCCGTGGGGTGCTTTTTTACACCTGTACATCTGTT
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBF//77BBF<FFFFFFFFFFFFFFFFB/<F
```

- Each sequenced nucleotide is assigned a quality score
- A Phred score higher than 20 (99% accurate) is generally OK