High-throughput DNA sequencing

Robert Lyle

Department of Medical Genetics

Norwegian Sequencing Centre

Oslo University Hospital

Robert.Lyle@medisin.uio.no

FUGE Functional genomics

CEES Center for Ecological and Evolutionary Synthesis

UNIVERSITAS OSLOENSIS MDCCCXI

UNIVERSITY OF OSLO

Oslo universitetssykehus

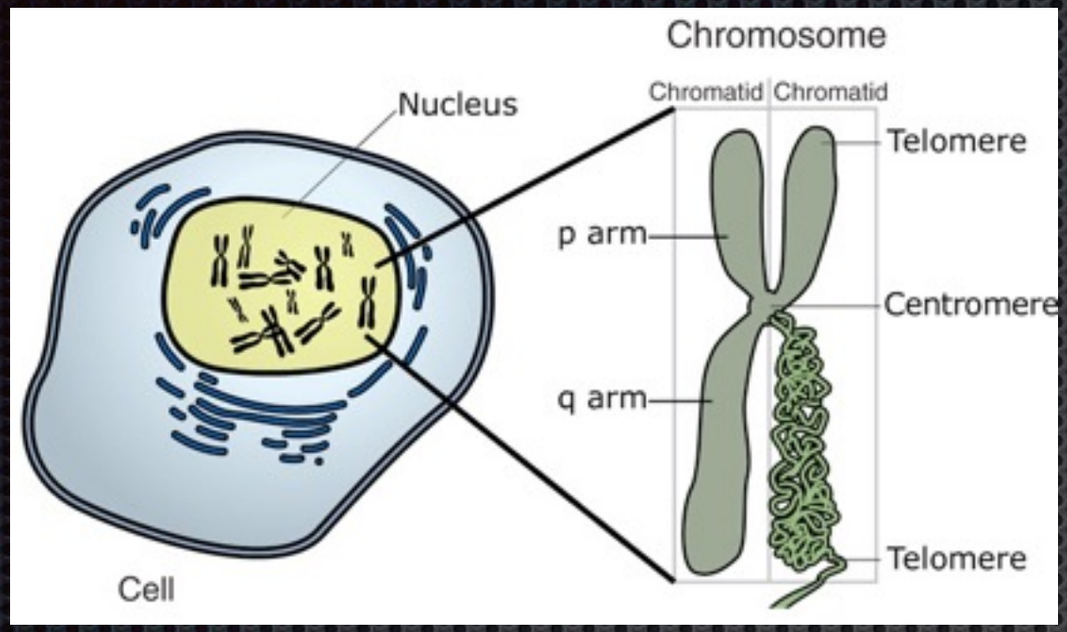
HELSE SØR-ØST

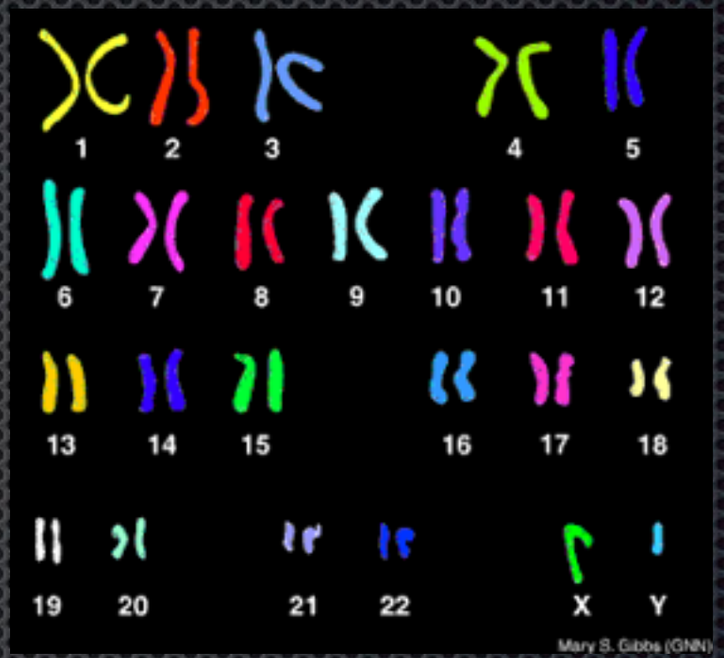
Overview

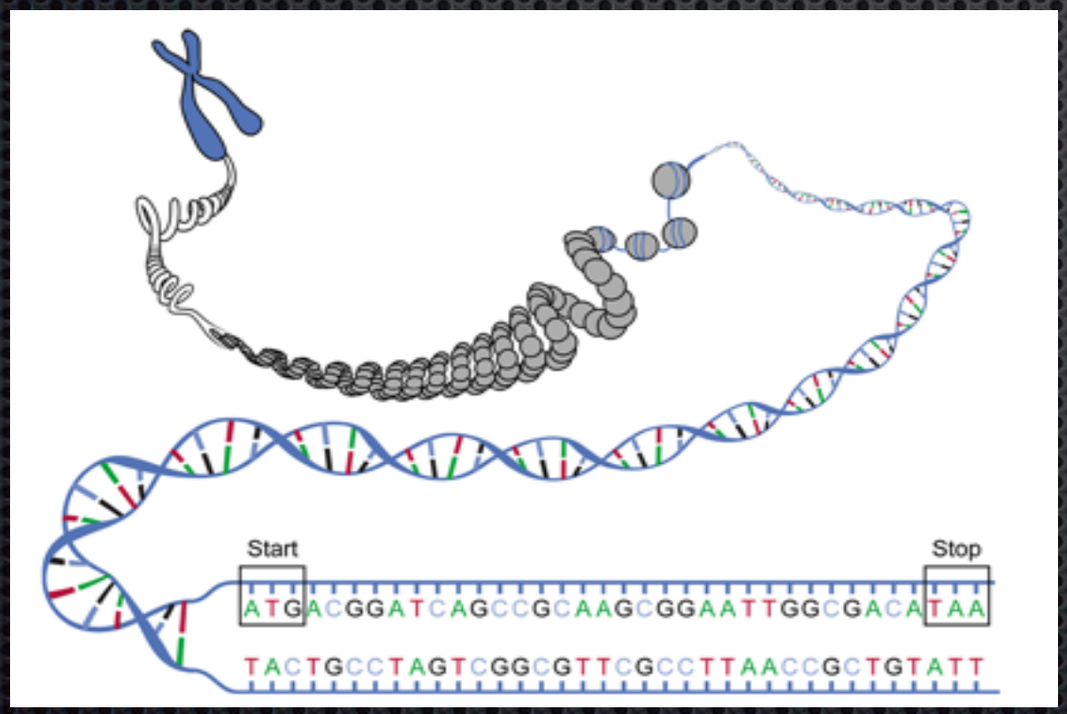
- DNA sequencing
- High-throughput sequencing
- Break?
- HTS and medical genetics
- Example

DNA sequencing

DNA







- 4 bases - A, G , C, T

G A T C
C T A G

- Human ***genome*** ~3 billion bases
- How can we read the sequence of bases?

How many bases?

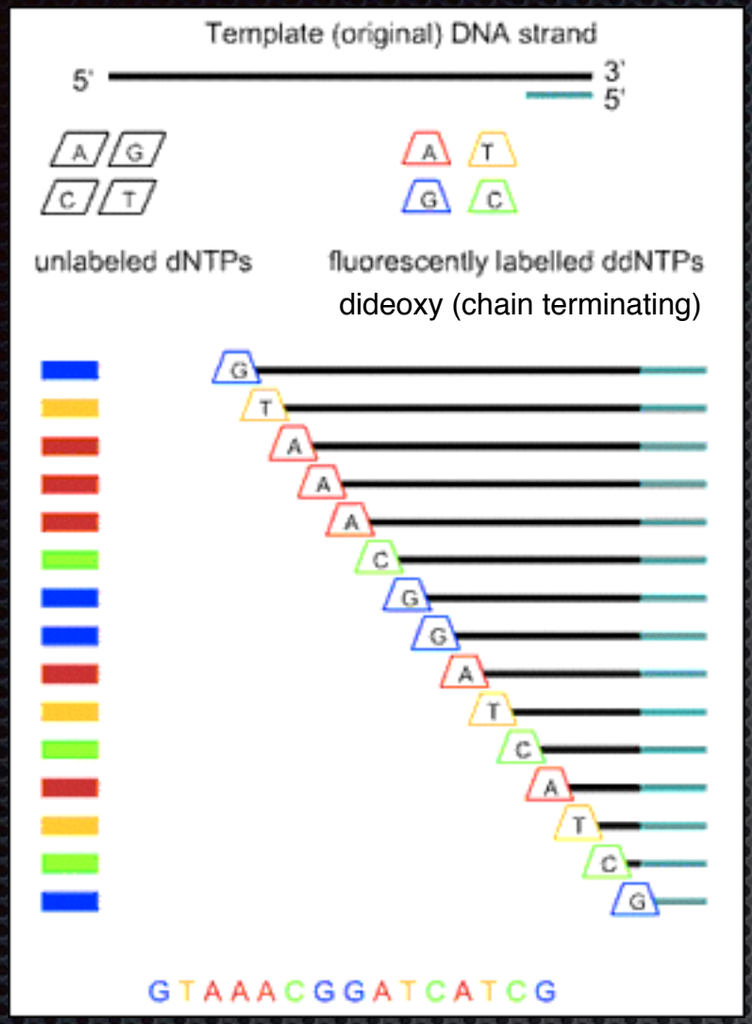
	bp	1
kilo	kb	1 000
mega	Mb	1 000 000
giga	Gb	1 000 000 000
tera	Tb	1 000 000 000 000

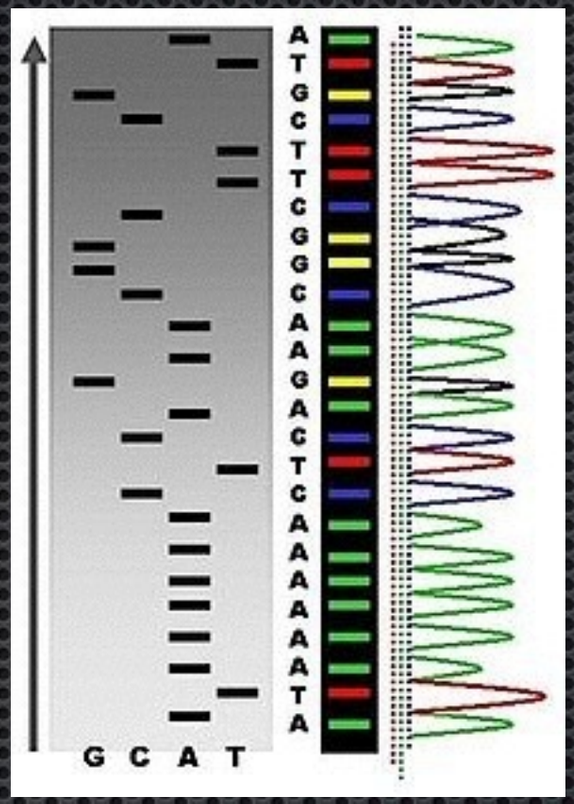
- Human ***genome*** ~3 Gb

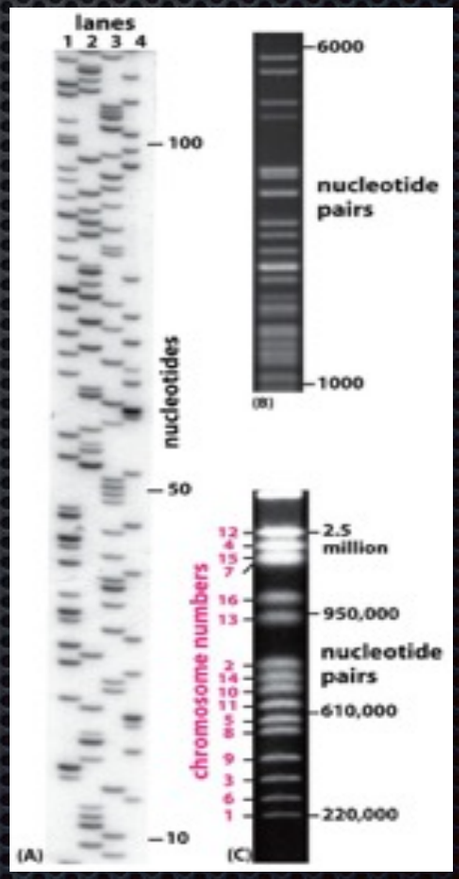
DNA replication

- Denature DNA
- Prime with short DNA (oligonucleotide)
- Polymerase extends base by base

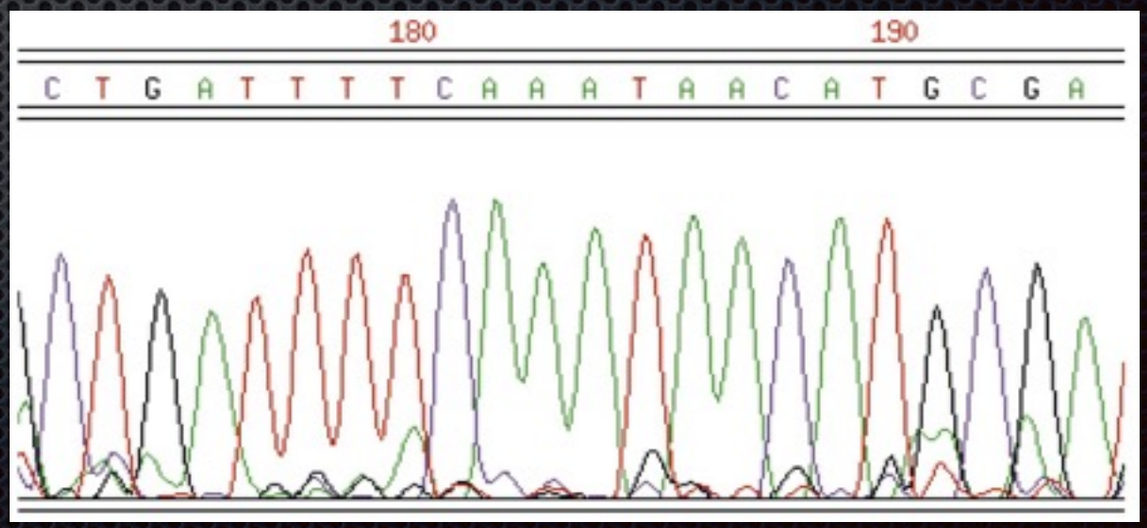
Sanger DNA sequencing







- Sanger sequencing
- Detect nucleotide extension with radioactivity or fluorescence
- Accurate but slow



High-throughput sequencing

Sequencing: old and next

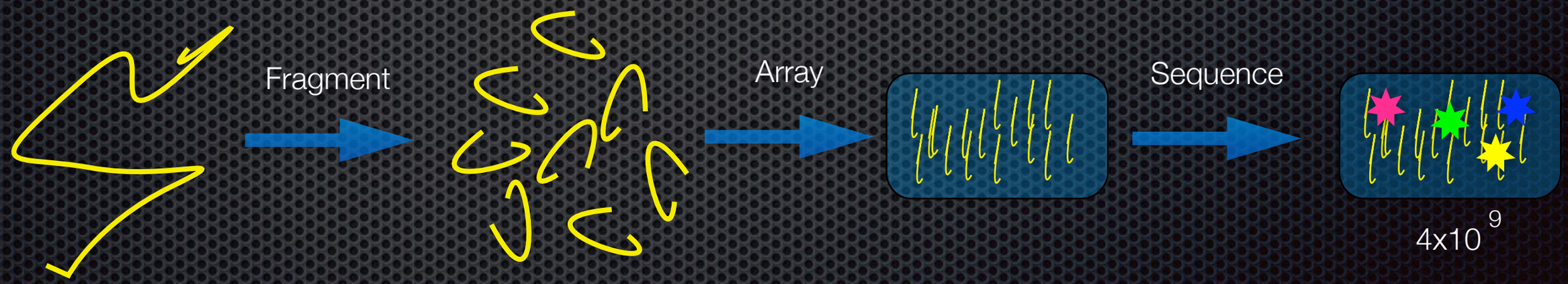
LTS (Sanger)

Molecules sequenced

1, 48, 96...

...unless you have a lot of machines

HTS (high-throughput sequencing)



...on one machine

Massively parallel

Sequencing platforms



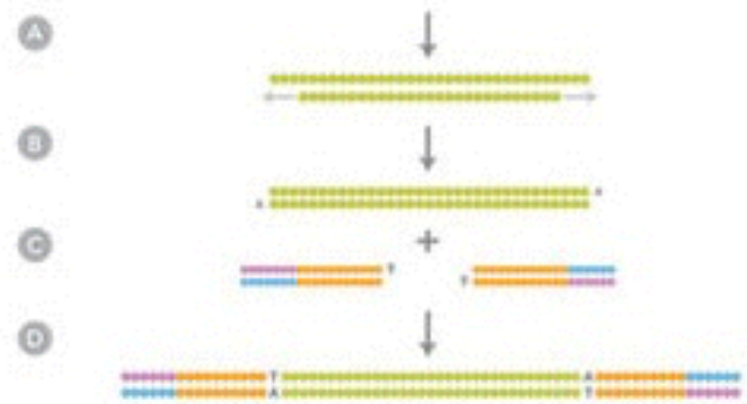


Platform	Illumina HiSeq	Illumina NextSeq	Illumina MiSeq	PacBio	Ion Torrent
@NSC	**4**	**1**	**3**	**1**	**1+1**
System cost	$750 k	$250k	$125k	---	+
Prep	++	++	++	+	+
Running cost	+	+	+	++	+
Run time	1-6 days	29 hours	65 hours	4 hours	2-4 hours
Read accuracy	98%	98%	98%	87%	98.8%
Read number	4 000 000 000	400 000 000	20 000 000	50 000	70 000 000
Read length	2x125 bp	2x150 bp	2x300 bp	10 (40) kb	1-200 bp
Output	1000 Gb	129 Gb	12 Gb	<1 Gb	10 Gb

Illumina sequencing technology

1. Library preparation

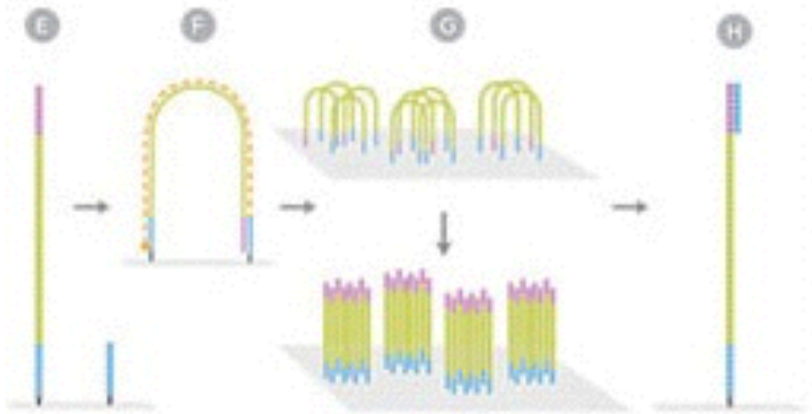
6 hours
3 hours hands-on time



A Fragment DNA
B Repair ends
Add A overhang
C Ligate adapters
D Select ligated DNA

2. Cluster generation

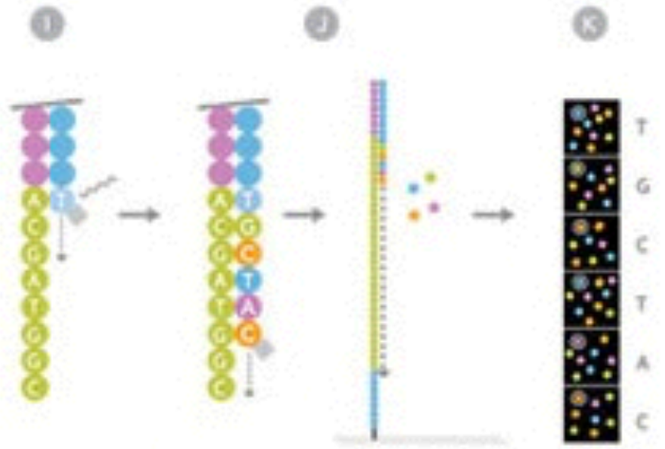
4 hours
30 minutes hands-on time
1–96 samples



E Attach DNA to
flow cell
F Perform bridge
amplification
G Generate clusters
H Anneal sequencing
primer

3. Sequencing

1–3 days single-read run
3–7 days paired-end run
30 minutes hands-on time
1–96 samples



I Extend first base,
read, and deblock
J Repeat step above
to extend strand
K Generate base calls

Illumina sequence data

- Random DNA library of short fragments ~300 bp
- ***4 billion DNA sequence reads***
- 75,...,300 bp long
- Single-end reads

- Paired-end reads

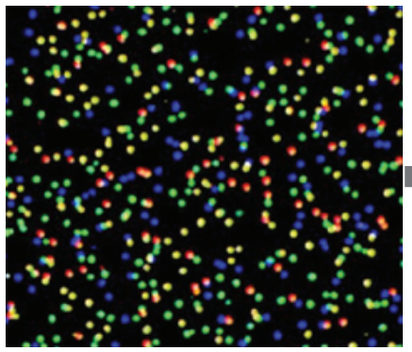
- Run time: 1-6 days
- Data volume: < 1 TB

Analysis pipeline

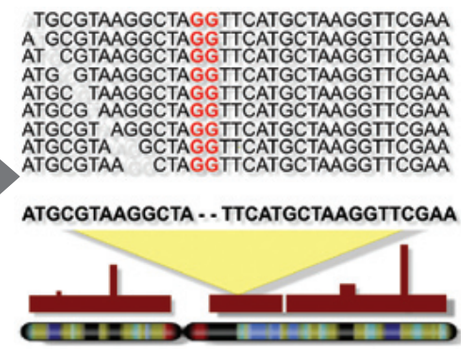
Illumina Pipeline

SCS | Firecrest | Bustard | GERALD

Images | Image Analysis | Base Calling | Aligned Reads



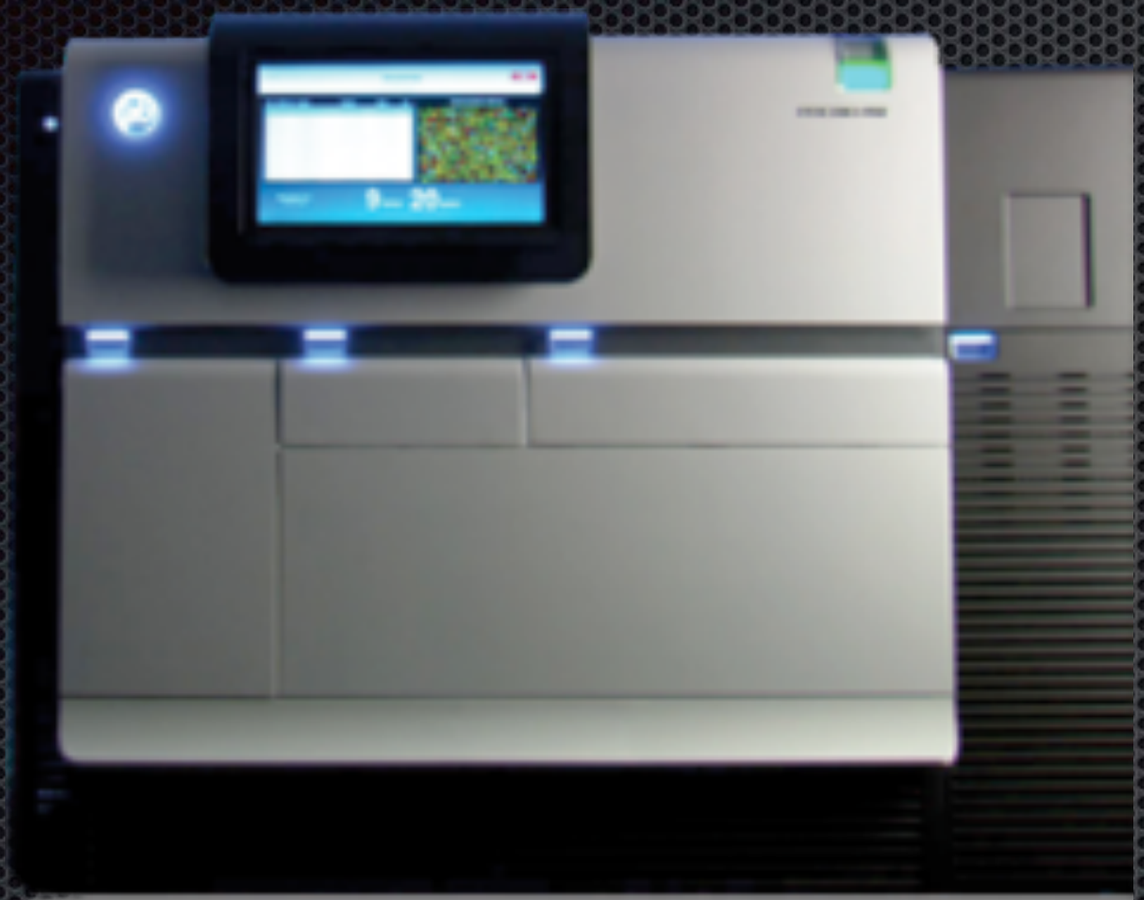
Lane	Tile	X	Y	Cycle 1 - A C G T	Cycle 2 - A C G T
5	12	924	1560	493.1 388.9 3626.7 2359.4	185.6 122.3 360.4 307.8
5	12	773	395	85.5 113.0 2327.5 1158.0	156.3 166.9 113.5 909.6
5	12	165	786	1243.8 741.1 45.8 67.4	318.4 692.6 48.3 41.7
5	12	598	690	1342.6 760.0 60.6 716.6	423.6 505.7 1919.1 959.3
5	12	1107	1207	59.9 63.0 957.5 818.2	98.6 230.5 815.1 512.1
5	12	1074	466	254.7 664.4 47.2 45.1	38.4 41.8 64.9 1102.9
5	12	887	356	743.1 486.4 42.2 305.0	230.3 603.6 -63.1 -20.1
5	12	642	1769	63.2 54.3 861.7 595.7	81.5 86.0 54.9 385.4
5	12	599	314	845.5 533.2 45.2 581.0	260.9 560.9 13.0 78.4
5	12	839	1103	372.0 812.6 16.7 70.5	59.4 69.4 35.4 1394.9
5	12	347	1792	343.8 706.9 108.4 638.5	73.2 43.9 121.6 1882.2
5	12	807	1114	63.9 63.8 828.3 1369.0	1074.4 714.3 -39.9 29.4




```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;39333
```

FASTQ format

Pacific BioSciences



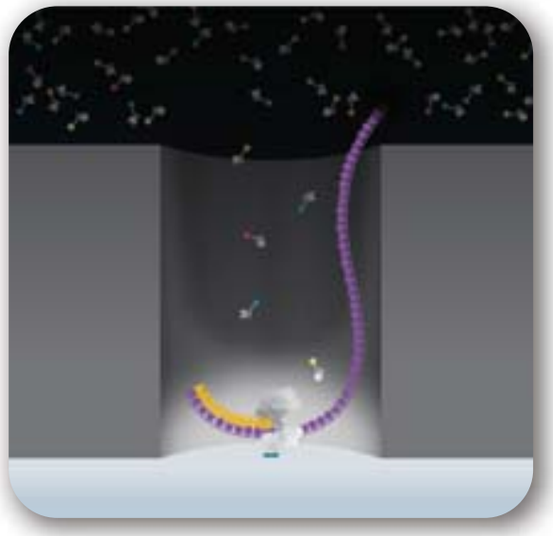
- Single molecule
- Real-time
- SMRT

http://www.pacificbiosciences.com

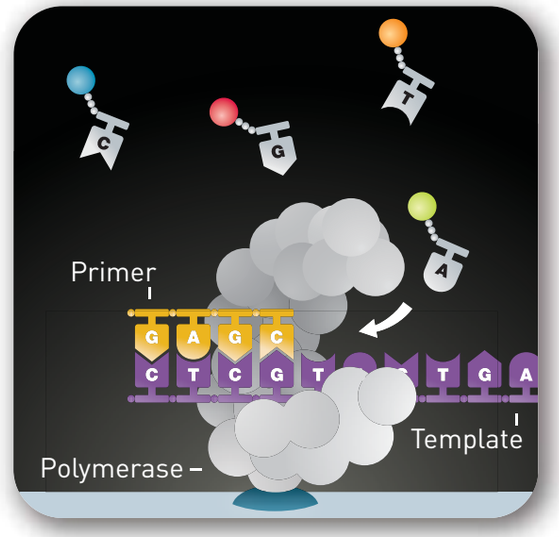
Pacific Biosciences RSII

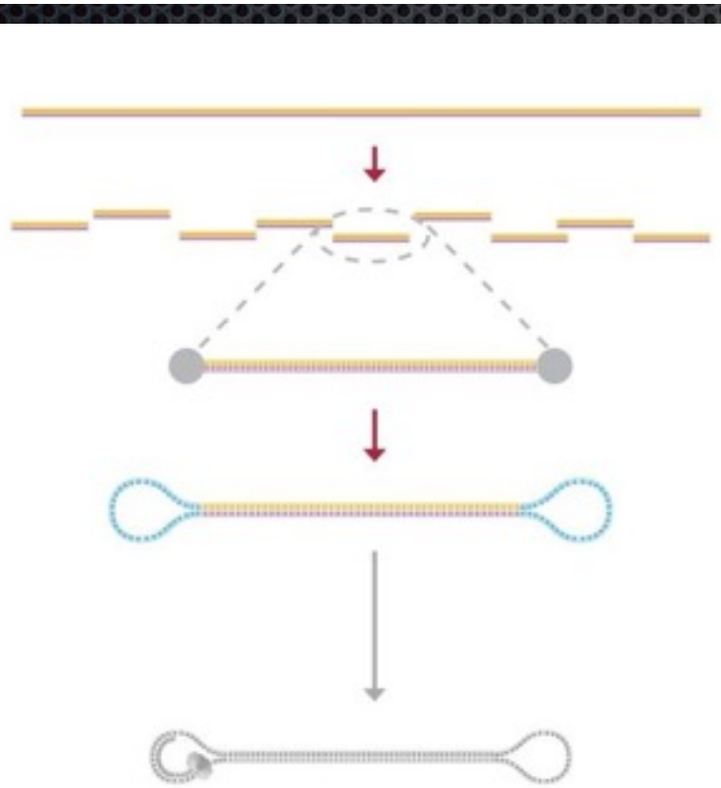
SMRT Cells

Zero-Mode Waveguides 150 000/SMRT cell



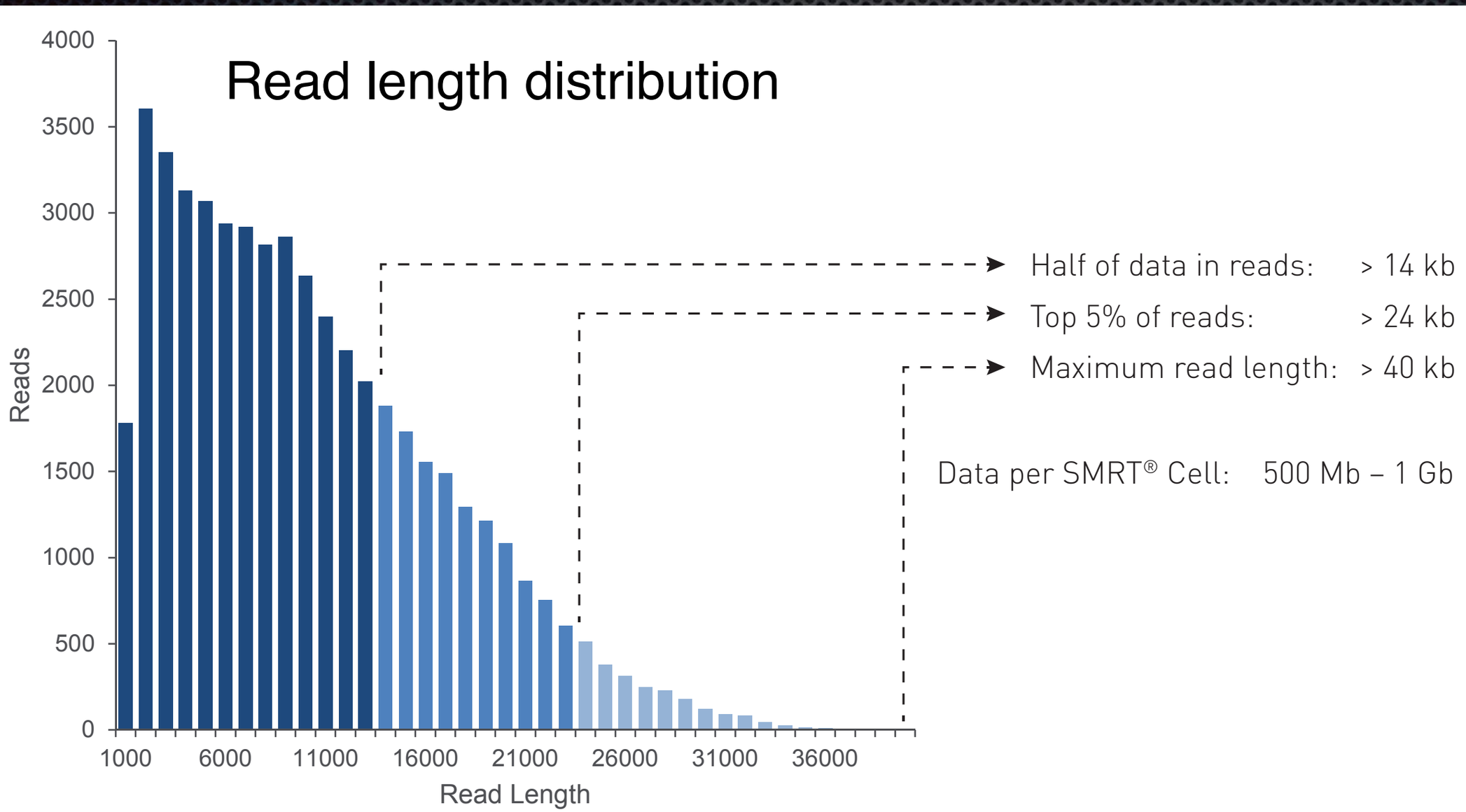
Phospholinked Nucleotides





- (Very) long reads
  - 50% > 10kb
  - Huge advantage for de novo sequencing
- Single molecule
  - No amplification bias

Pacific Biosciences RSII



Based on data from a 20 kb size-selected *E. coli* library using a 4-hour movie.
Each SMRT Cell yields ~ 50,000 reads.

What can I sequence?

Sequencing platforms





Platform	Illumina HiSeq	Illumina NextSeq	Illumina MiSeq	PacBio	Ion Torrent
@NSC	**4**	**1**	**3**	**1**	**1+1**
System cost	$750 k	$250k	$125k	---	+
Prep	++	++	++	+	+
Running cost	+	+	+	++	+
Run time	1-6 days	29 hours	65 hours	4 hours	2-4 hours
Read accuracy	98%	98%	98%	87%	98.8%
Read number	4 000 000 000	400 000 000	20 000 000	50 000	70 000 000
Read length	2x125 bp	2x150 bp	2x300 bp	10 (40) kb	1-200 bp
Output	1000 Gb	129 Gb	12 Gb	<1 Gb	10 Gb

Types of sequence projects

Project	Description
Resequencing	Align and compare to a reference sequence
de novo	Assemble new genome
metagenomics	Sequence DNA pool of multiple species
mRNA	Sequence cDNA for gene expression
miRNA	Sequence small RNAs for expression
ChIP	Study chromatin structure
DNA meth	Study DNA methylation

Sequencing platforms





Platform	Illumina HiSeq	Illumina NextSeq	Illumina MiSeq	PacBio	Ion Torrent
@NSC	4	1	3	1	1+1
System cost	$750 k	$250k	$125k	---	+
Prep	++	++	++	+	+
Running cost	+	+	+	++	+
Run time	1-6 days	29 hours	65 hours	4 hours	2-4 hours
Read accuracy	98%	98%	98%	87%	98.8%
Read number	4 000 000 000	400 000 000	20 000 000	50 000	70 000 000
Read length	2x125 bp	2x150 bp	2x300 bp	10 (40) kb	1-200 bp
Output	1000 Gb	129 Gb	12 Gb	<1 Gb	10 Gb