2 - Sequencing Technologies

Tuesday afternoon

Bernardo J. Clavijo
Richard Smith-Unna
Gonzalo Garcia

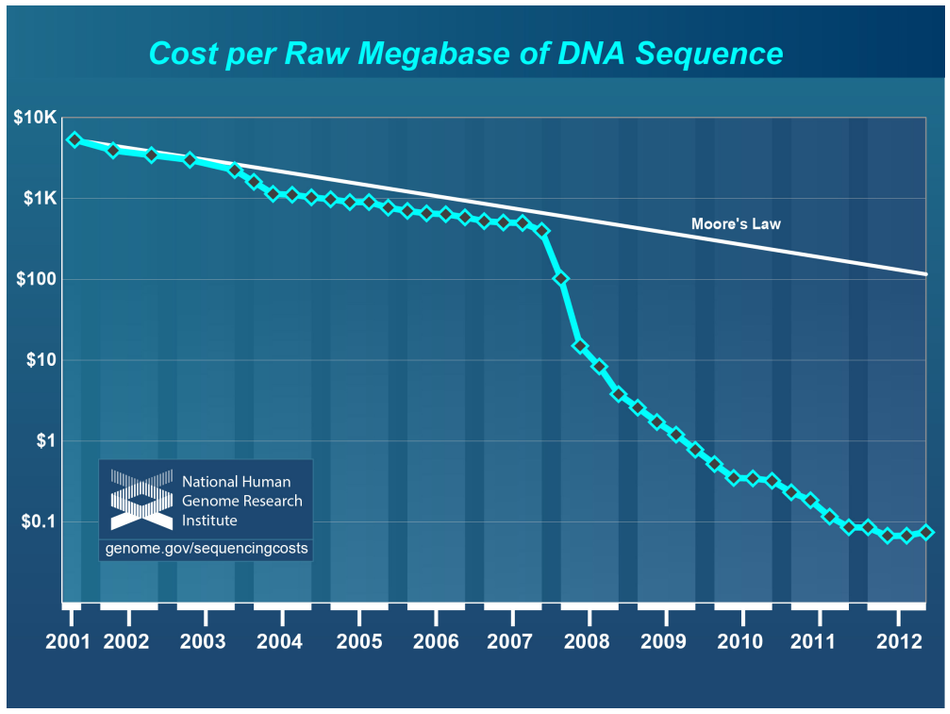




TGAC

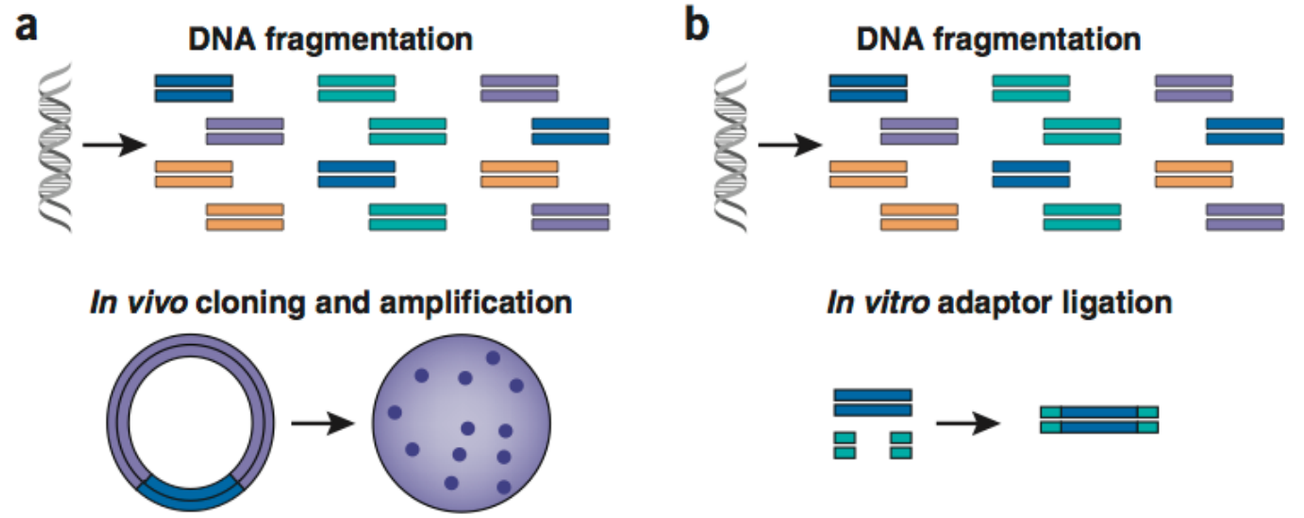
A brief history of DNA sequencing

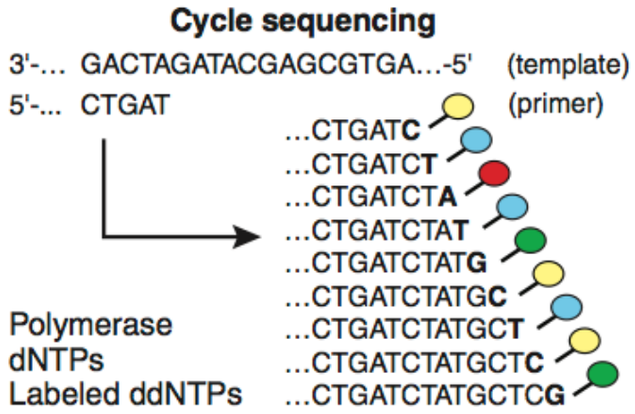
- 1953 – double helix structure, Watson & Crick
- 1977 – rapid DNA sequencing, Sanger
- 1977 – first full (5k) genome – bacteriophage Phi X
- Late 80s – first production ‘Sanger’ sequencers
- Mid 90s – DNA microarrays
- 2001 – draft human genome
- 2004 – first 454 pyrosequencing machine
- 2006 – first Solexa/Illumina sequencer
- 2011 – PacBio RS
- 2014 – Nanopore

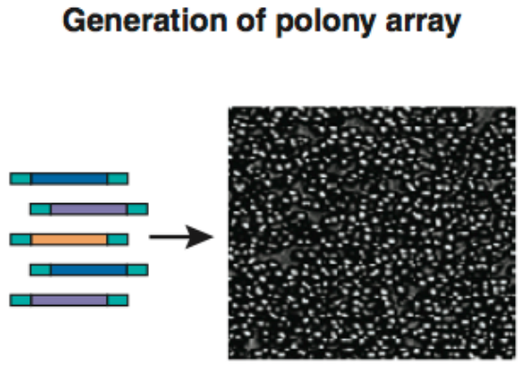


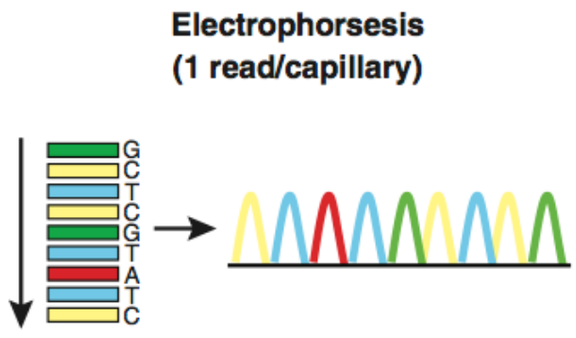
TGAC

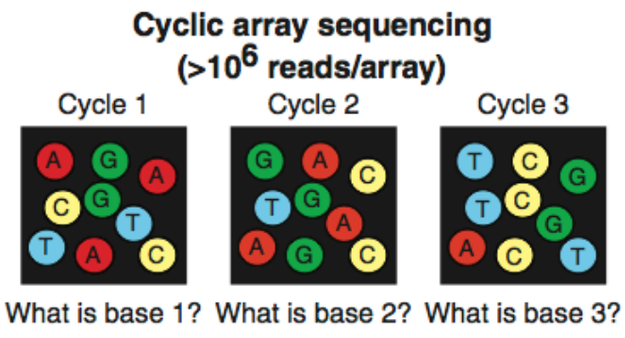
Next Generation Sequencing





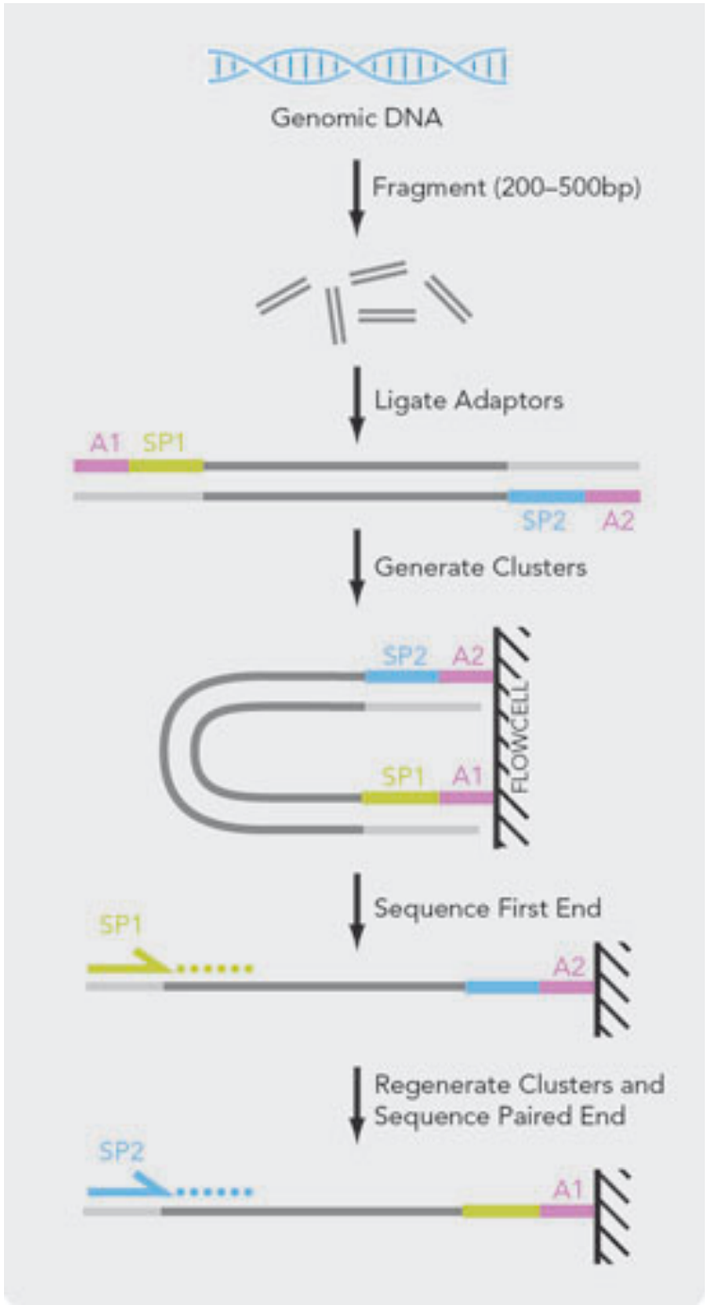


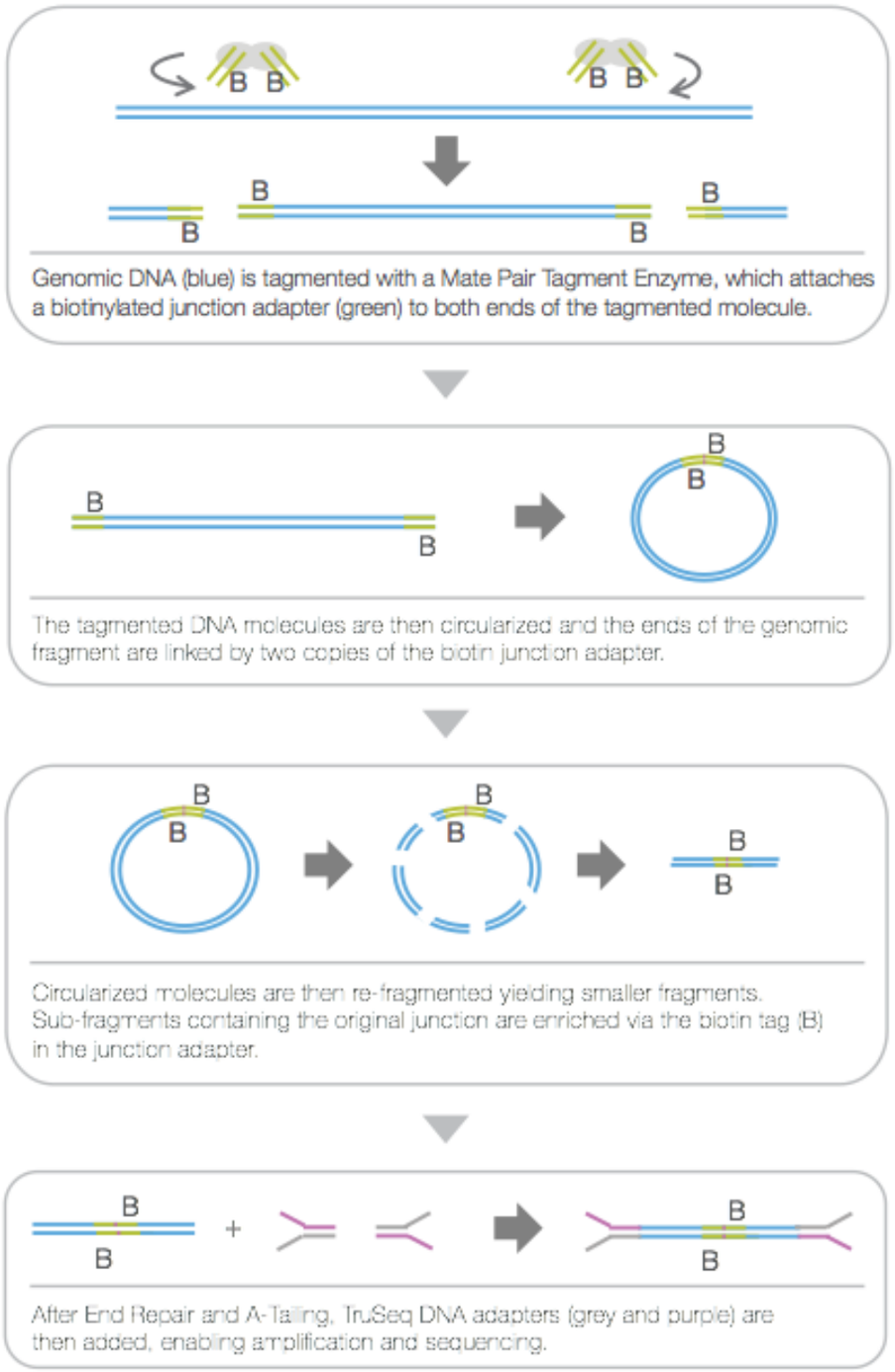




TGAC

Creating and Sequencing Paired Libraries





TGAC

TGAC Sequencing Platforms

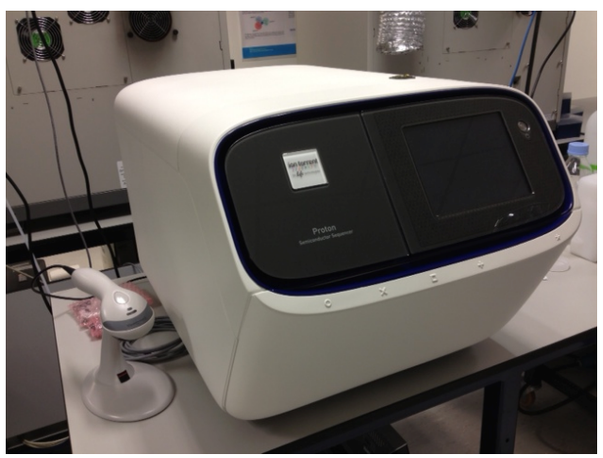
Illumina GAII x 1

Illumina HiSeq x 3

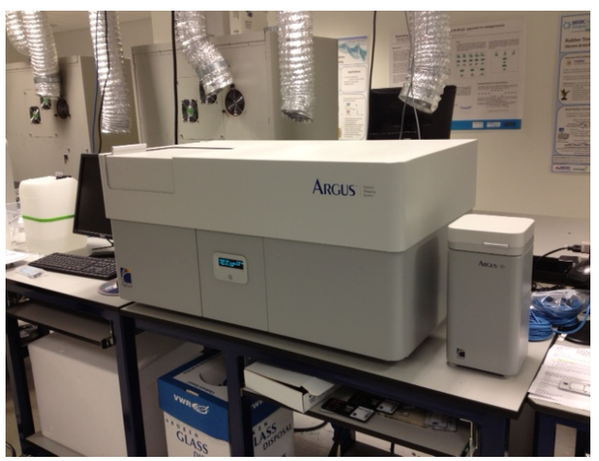
Illumina MiSeq x 3

Roche 454FLX x 2

PacBio RS x 1



Proton x 1



Opgen Argus x 1

TGAC

Platforms compared

	METHOD	READ LENGTH	NUMBER OF READS	THROUGHPUT	RUN TIME	ACCURACY	APPROX. COST
ILLUMINA HiSeq 2500 *High Output*	Sequencing by synthesis	Up to 100bp PE	1.5 billion per flowcell	300 Gb	11 days	99.9%	£14,000
ILLUMINA HiSeq 2500 *Rapid*	Sequencing by synthesis	Up to 150bp P.E	300 million per flowcell	90 Gb	40hours	99.9%	£4,400
ILLUMINA MiSeq	Sequencing by synthesis	Up to 250bp P.E	15 million per flowcell	8.5 Gb	39hours	99.9%	£1,400
454	Pyrosequencing	Up to 400 bp	1 million per plate	400 Mb	10 hours	99.9%	£6,000
PACBIO *Standard Run*	Real time sequencing	3Kb Upper 5% >6kb	50 000 per SMRT cell	100 Mb	2x55mins	86%	£300
PACBIO *Long Run*	Real time sequencing	3.5kb Upper 5% >10kb	25 000 per SMRT cell	60 Mb	1 x 120mins	86%	£300
OpGen Argus	Optical Map	150kb -> 2Mb	~2 000 per Map Card	3Gb	120mins	N/A	£500-£1000

TGAC

The FASTQ file

- 4 lines per read
- Stores sequence and quality

Read ID →
Sequence →
Quality →

```
@HWI-ST790:234:D0W8BACXX:1:1101:1792:2000 1:N:0:GCCAA
ACNATTAACAACCTTGGTGTTCAGCATGAGAACTTATCTGCAGCTGAGTCTCGTATCCGTGACG
+
CC#4ADDFHHHHHIIIIEGHIIIIIIIIIIGIIIIIIIIIIIIIIIIIDGHHIDHHIII6@FGI
@HWI-ST790:234:D0W8BACXX:1:1101:2592:1999 1:N:0:GCCAA
CTNGAATGCAGGTAGAATACATCTCCCGGATAAGCCTCGCGGCCCCCGGGGCGGGGGGGGAGAG
+
:=#44AA?:<DFFE>FED?3A<EHH>FIF?ADGCGBA?D########################
@HWI-ST790:234:D0W8BACXX:1:1101:4221:1999 1:N:0:GCCAA
GGNAAATACGAAAGATAAGCTACGCAAGAAACGAAGGATTACTGCGAAAGGCTGCGATGCGGCA
+
@@#4=BDDFDFHDIIBGIHHHIGGIIIBHHIF=ABB@?B<DE@BF<FHH@@EHACD<B3=8@:B
```


TGAC

Different Data for different Information

- Illumina paired end: a good and cheap way to get the motifs
- Long mate pairs: a hint at order and distances
- PacBio:
  - Long reads: longer, not very precise, motifs
  - Circular consensus reads: long, expensive, precise motifs
- Others include:
  - Optical maps (good positional information)
  - RNA-seq
  - Fosmid ends
  - Known deletion bins / markers / ESTs

Questions?