Lecture 4

- Genome data analysis 1: Introduction of high-throughput sequencing data analysis

New challenges

- Dealing with millions of small reads
- Interpreting the sequencing results

First giant steps

Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene
W. FIERS et al, Nature April 1976

Nucleotide sequence of bacteriophage X174 DNA
F. Sanger et al, Nature December 1976.

More big steps

1984: The entire sequence of the HIV-1 genome was determined by Chiron Corp.

1995:The first genome of a free living organism (*H. influenzae*) was sequenced.

1996: The complete genome of the E. coli bacteria was sequenced.
The sequencing of the *S. cerevisiae* genome was completed.

1998: The first complete genome sequence of a multicellular organism (roundworm *C. elegans*) was published.

1999: The complete genome of the fruit fly (*D. melanogaster*) was sequenced.

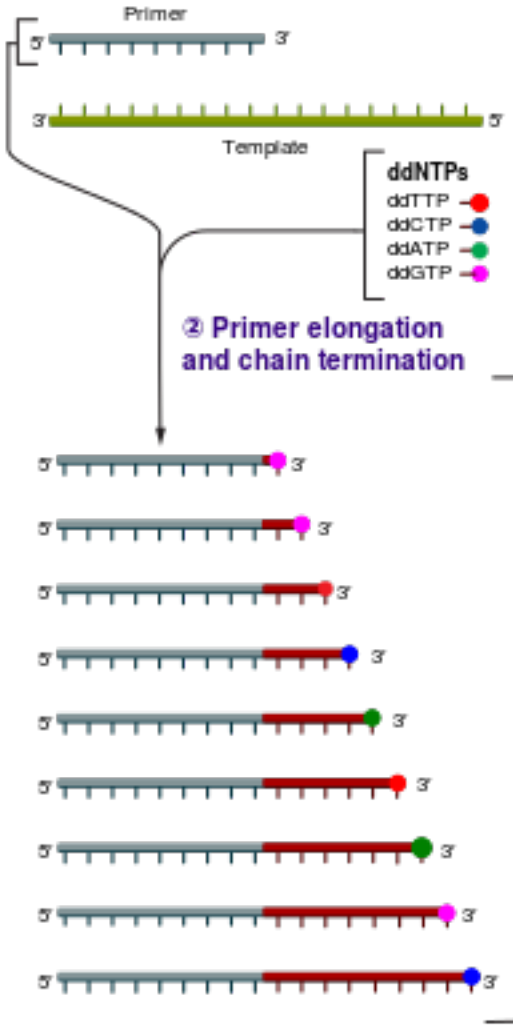
2000: The first plant genome (*A. thaliana*) was published.

2001: The first draft sequence of the entire human genome was published.
Celera Genomics announced the completion of a draft mouse genome sequence.

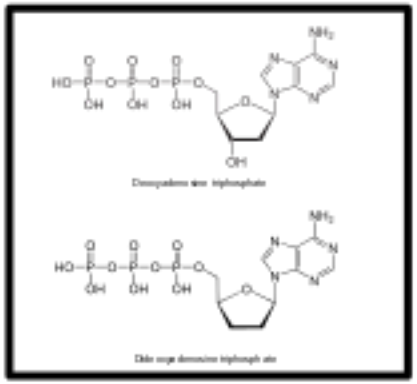
2003: The finished Human Genome was published.

Sanger sequencing

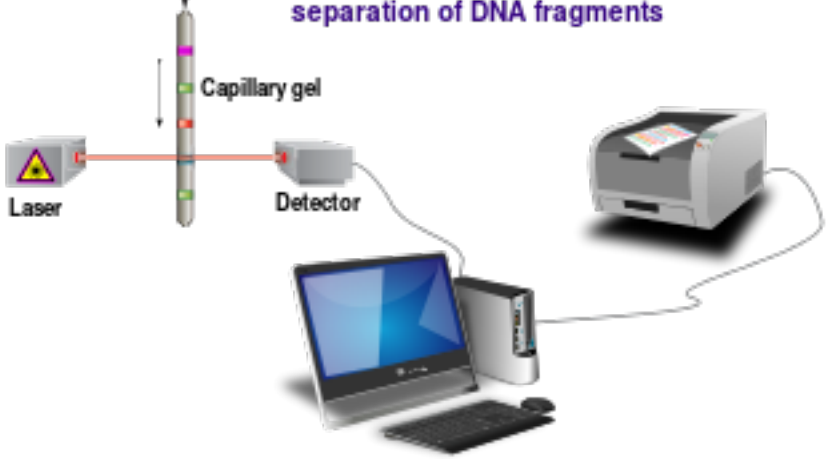
① Reaction mixture
- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

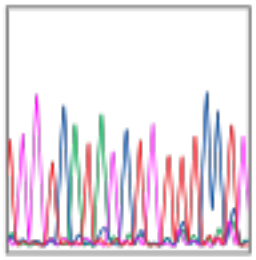


② Primer elongation and chain termination



③ Capillary gel electrophoresis separation of DNA fragments

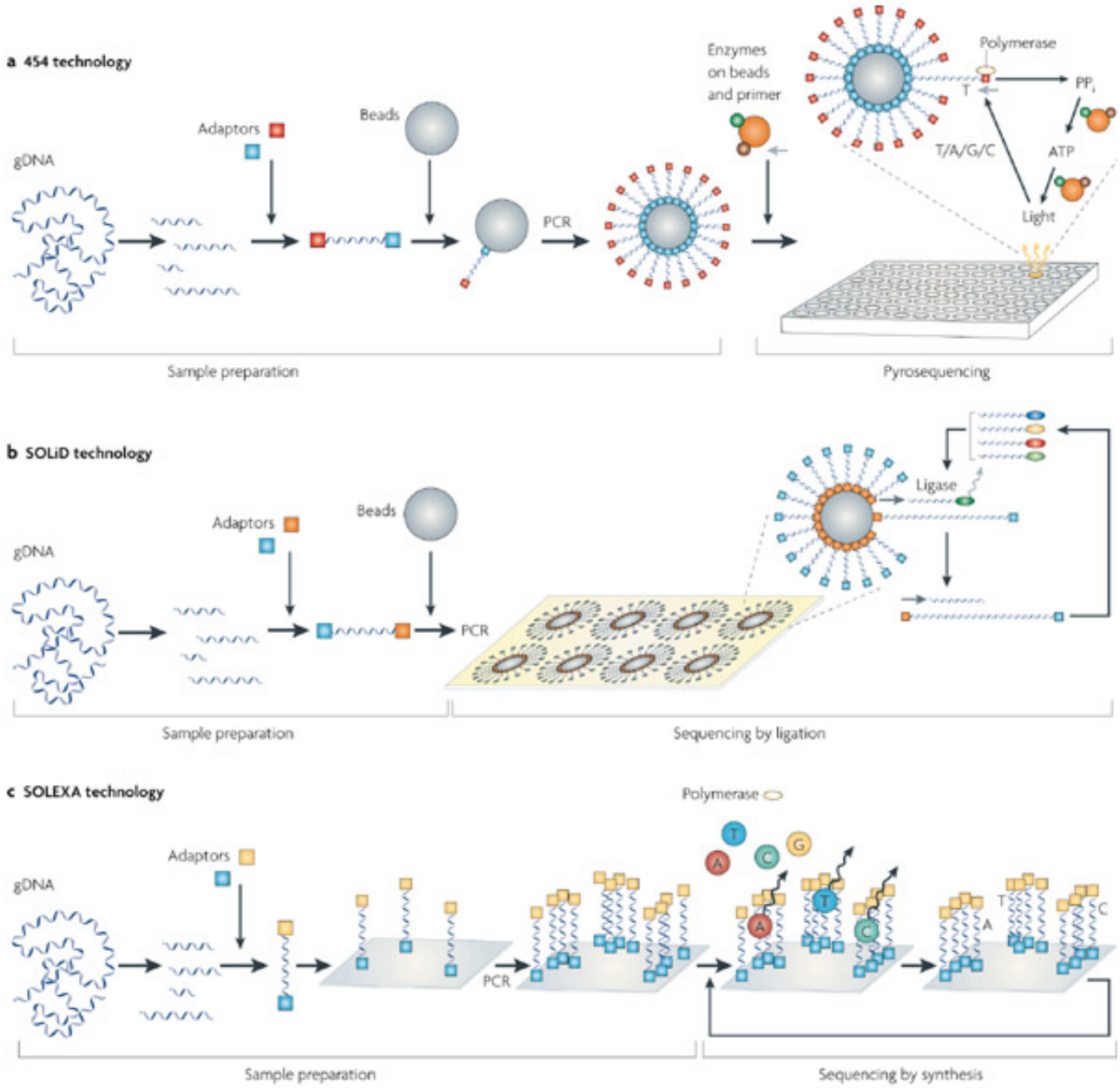




④ Laser detection of flourochromes and computational sequence analysis

Next generation Sequencing

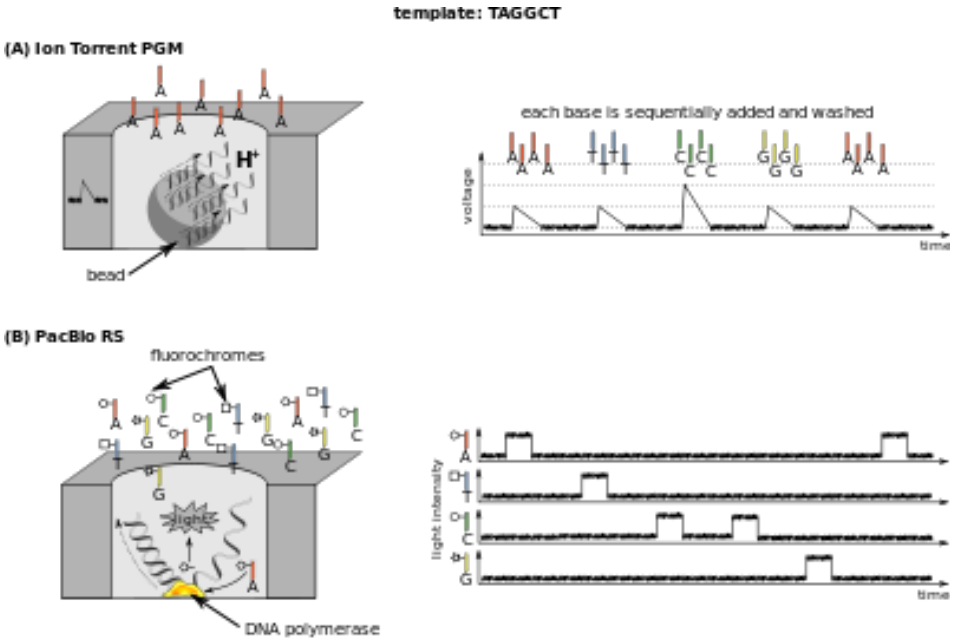
Pyrosequencing (454)
Sequencing by synthesis (illumina)
Sequencing by ligation (SOLID)


a 454 technology
gDNA
Adaptors
Beads
PCR
Enzymes on beads and primer
Polymerase
T
PPi
T/A/G/C
ATP
Light
Sample preparation
Pyrosequencing
b SOLiD technology
gDNA
Adaptors
Beads
PCR
Ligase
Sample preparation
Sequencing by ligation
c SOLEXA technology
gDNA
Adaptors
Polymerase
PCR
T
A
C
G
Sample preparation
Sequencing by synthesis

Next generation Sequencing
Single molecule real time platforms

Ion Torrent PGM. Life Technologies

PacBio RS. Pacific Biosciences



“Very small PH-meter”

“Very small microscope”

HTG platforms

Method	Single-molecule real-time sequencing (Pacific Bio)	Ion semiconductor (Ion Torrent sequencing)	Pyrosequencing (454)	Sequencing by synthesis (Illumina)	Sequencing by ligation (SOLiD sequencing)	Chain termination (Sanger sequencing)
Read length	2900 bp average	200 bp	700 bp	50 to 250 bp	50 to 100 bp	400 to 900 bp
Accuracy	87% (read length mode), 99% (accuracy mode)	98%	99.90%	98%	99.90%	99.90%
Reads per run	35-75 thousand	up to 5 million	1 million	up to 3 billion	1.2 to 1.4 billion	N/A
Time per run	30 minutes to 2 hours	2 hours	24 hours	1 to 10 days, depending upon sequencer and specified read length	1 to 2 weeks	20 minutes to 3 hours
Advantages	Longest read length. Fast.	Less expensive equipment. Fast.	Long read size. Fast.	Potential for high sequence yield, depending upon sequencer model and desired application.	Low cost per base.	Long individual reads. Useful for many applications.
Disadvantages	Low yield at high accuracy. Equipment can be very expensive.	Homopolymer errors.	Runs are expensive. Homopolymer errors.	Equipment can be very expensive.	Slower than other methods.	More expensive and impractical for larger sequencing projects.

The Fastq Format

A FASTQ file normally uses four lines per sequence.
•Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
•Line 2 is the raw sequence letters.
•Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
•Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Illumina Sequence ID formats

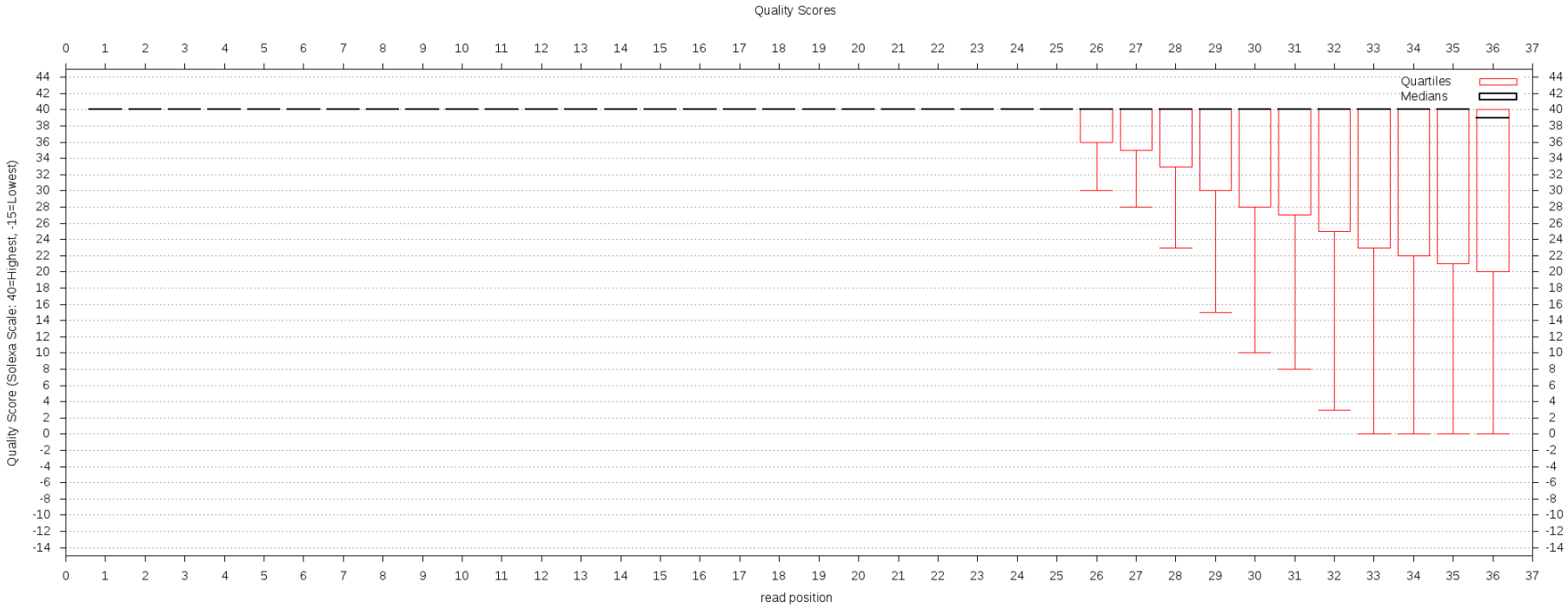
@HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	Instrument name
6	Flow cell lane
73	Tile number within the flow cell lane
941	X-coordinate of the cluster within the tile
1973	Y-coordinate of the cluster within the tile
#0	Index number for a multiplex sample (0 for no index)
/1	The member of a pair /1 or /2

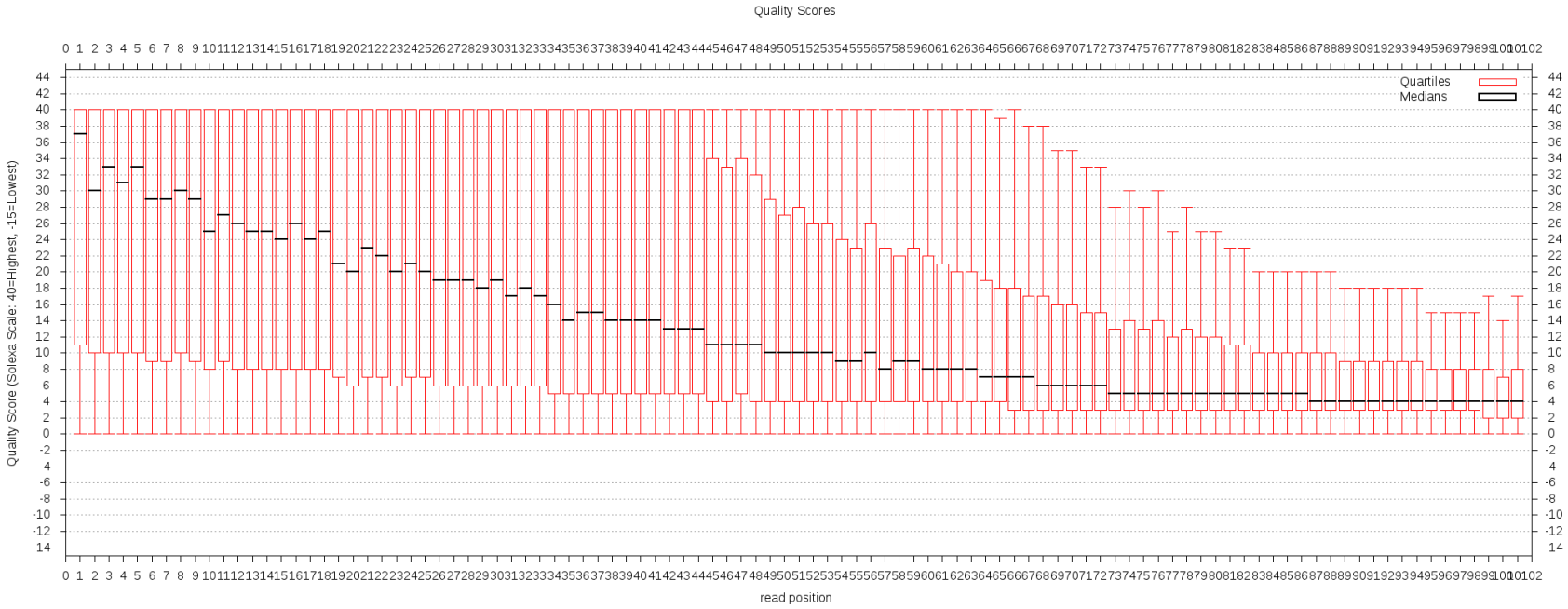
Quality Scores

- S - Sanger Phred+33, raw reads typically (0, 40)
- X - Solexa Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
- L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Good reads


Quality Scores
Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

Bad quality!


Quality Scores
Quartiles
Medians
Quality Score (Solexa Scale: 40=Highest, -15=Lowest)
read position

The fastx tool kit

- http://hannonlab.cshl.edu/fastx_toolkit/

FASTQ-to-FASTA

FASTX STATISTICS

FASTQ QC

FASTA/Q Clipper : clips adapters

FASTA/Q Renamer : Changes ID names

FASTA/Q Trimmer: cuts nucleotides

FASTA/Q Artifact Filter: Filter for artifacts in your reads

FASTQ Quality filter: Filter for reads with specific quality scores

FASTAQ/A Reverse Complement

FASTX Barcode Splitter

Exercise

Move the files cassava1.fq and cassava2.fq TO YOUR USER FOLDER AND WORK THERE =)

Lets start a session with the hipergator (interactive)

module load ufrc

srundev –t 120 –m=2g

module load fastqc

module load fastx_toolkit

- Use FastQC program and describe the quality and length of the reads
- type module load fastqc (to load the program)
- fastqc
- Use fastx Tool kit to:
- Cut cassava1.fq for the last 10 nucleotides of your reads and save in a new file1_trimmed.fastq
- Use fastqc program to load file1_trimmed.fastq and describe your new reads.
- Transform file1_trimmed.fastq in fasta format