GENETICS AND MOLECULAR BIOLOGY FOR ENVIRONMENTAL ANALYSIS

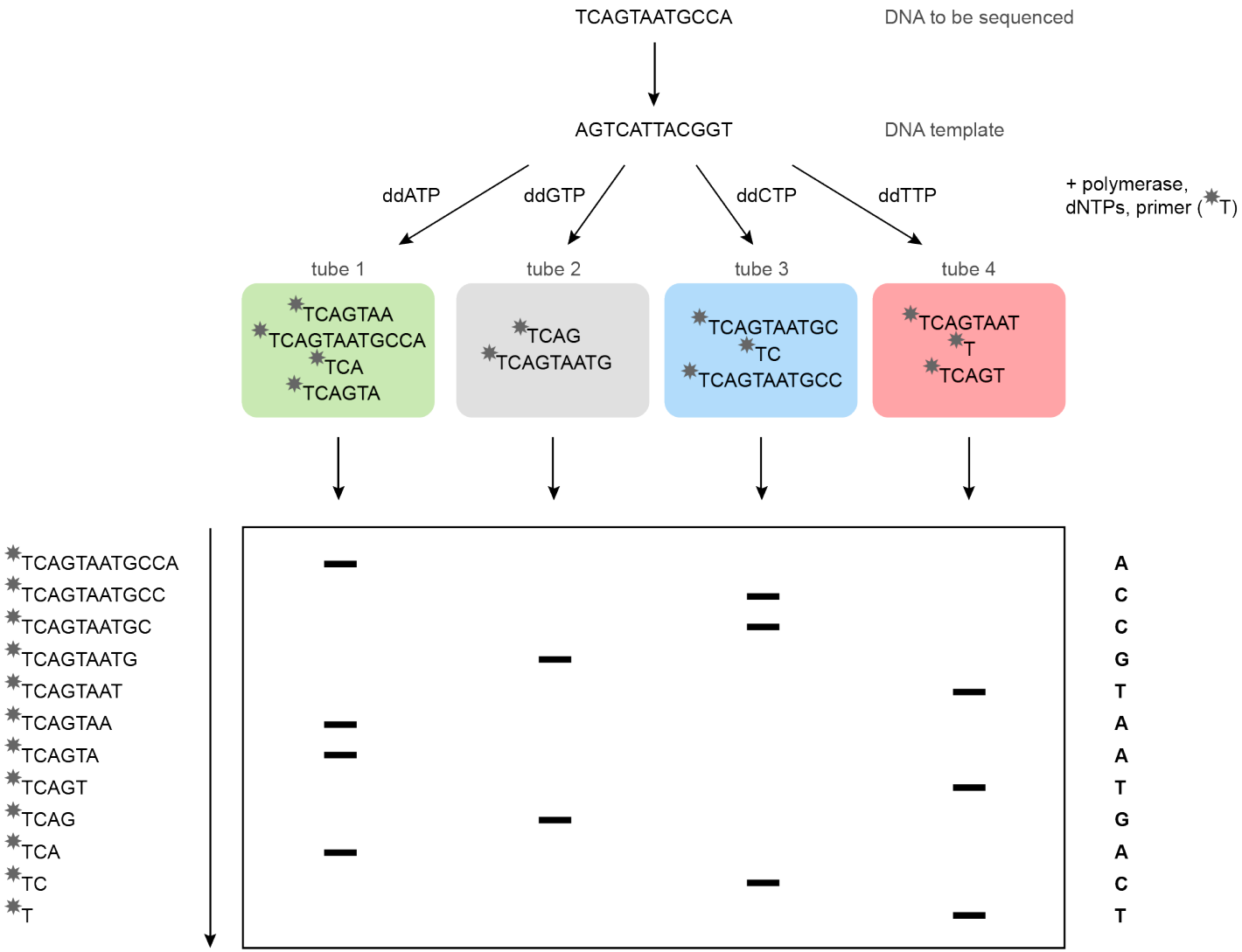
MOLECULAR ECOLOGY
LESSON 6: DNA SEQUENCING

Prof. Alberto Pallavicini
pallavic@units.it

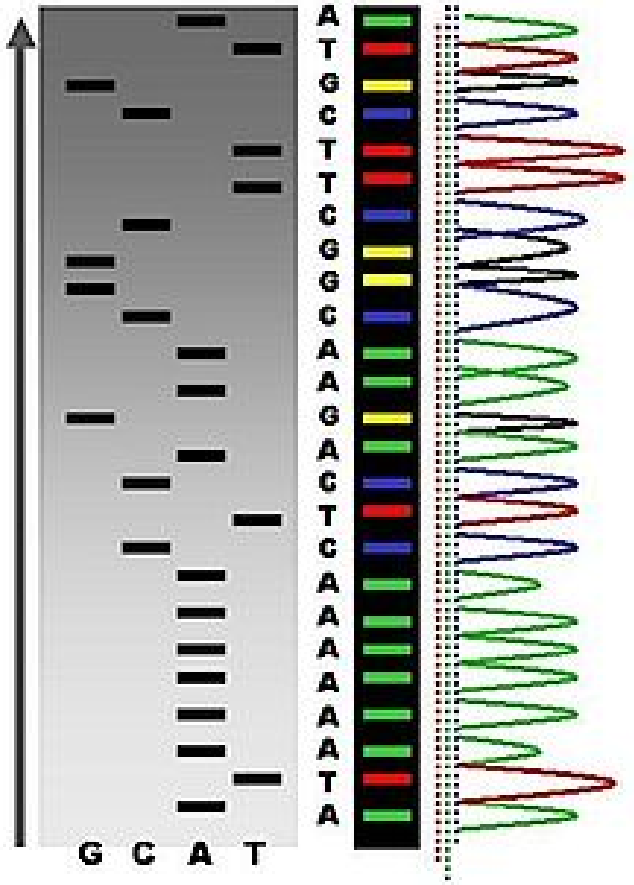
GENOMICS ANALYSIS

- Genomics is highly technology driven. The enormous impact of genomics research on medical and agro-technological sciences has inspired commercial life-science companies to develop innovative genomic tools at a tremendously high speed.

…BUT LET'S START FROM THE BEGINNIN: SANGER SEQUENCING



SANGER SEQUENCING



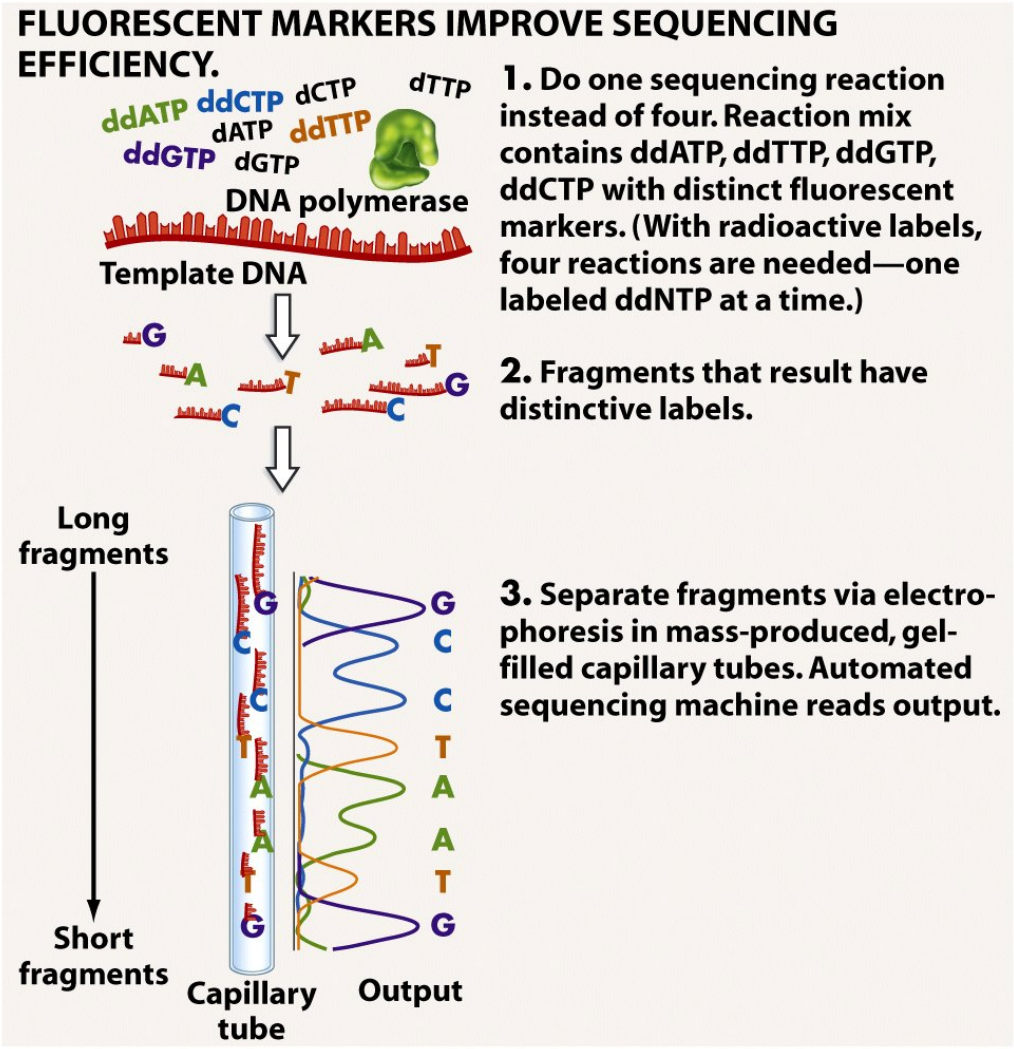


Figure 20-1 Biological Science, 2/e

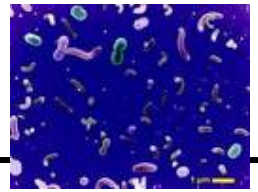
SANGER SEQUENCING

- Advantages
  - Long reads (~900bps)
  - Suitable for small projects
- Disadvantages
  - Low throughput
  - Expensive

SANGER SEQUENCING

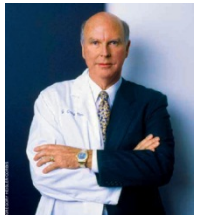
1994: *H. Influenzae*
1.8 Mbp
(Fleischmann et al.)

2007: *Global Ocean Sampling Expedition*
~3,000 organisms, 7Gbp (Venter et al.)



1980 — 1990 — 2000

1982: *lambda virus*
DNA stretches up to
30-40Kbp
(Sanger et al.)



2001: *H. Sapiens, D. Melanogaster*
3 Gbp
(Venter et al.)

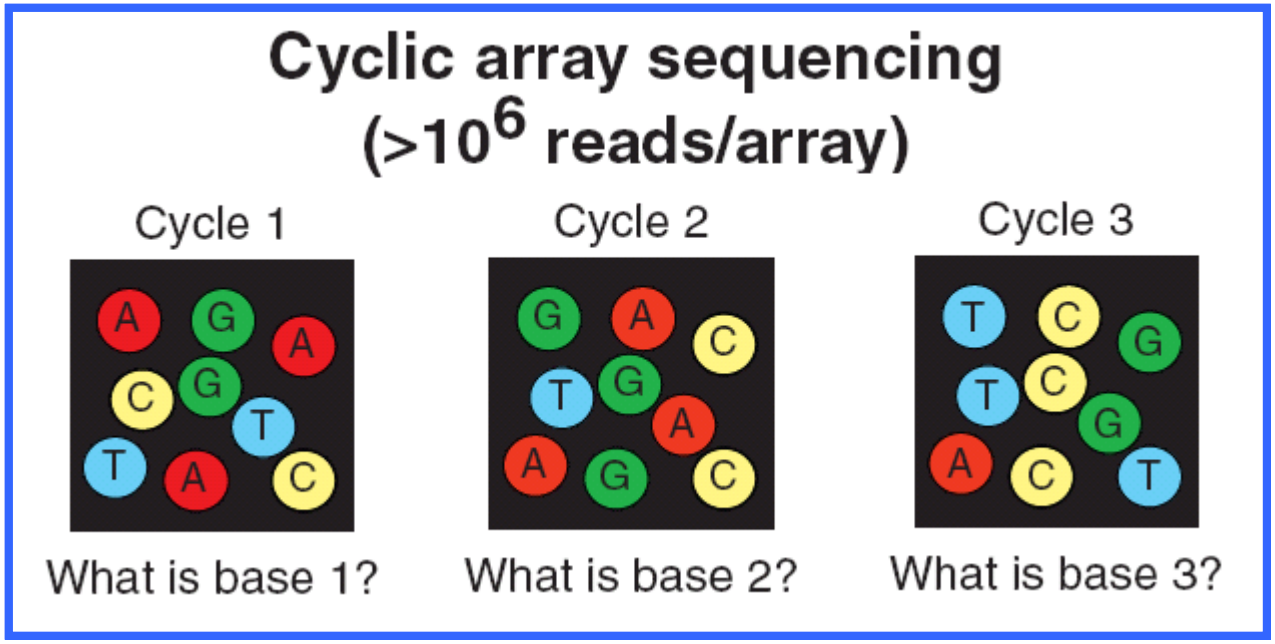
NEXT GENERATION SEQUENCING: WHY NOW?

- **Motivation:** HGP and its derivatives, personalized medicine
- **Short reads applications:** (re-)sequencing, other methods (e.g. gene expression)
- Advancements in technology

HIGH PARALLELISM IS ACHIEVED IN POLONY SEQUENCING

Sanger

Polony



454

Illumina

Ion torrent

MinION

TECHNOLOGY SUMMARY

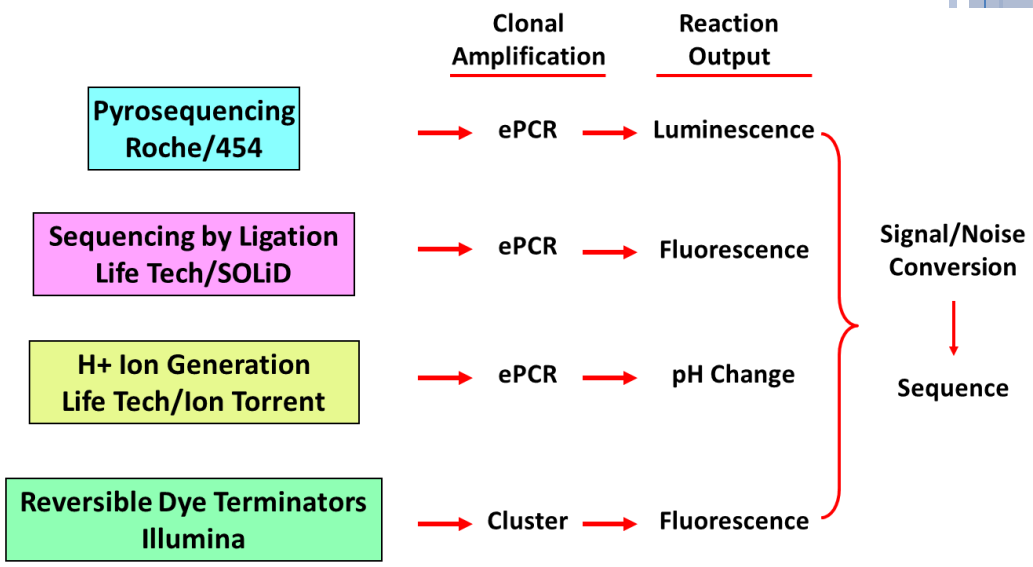
Summary of the five major next-generation sequencing platform families

Platform Family	*Clonal Amplification*	*Chemistry*	*Highest Average Read Length*
454	Emulsion PCR	Pyrosequencing (seq-by-synthesis)	700 bp (paired-end sequencing available)
Illumina	Bridge amplification	Reversible dye terminator (seq-by-synthesis)	300 bp (overlapping paired-end sequencing available)
SOLiD	Emulsion PCR	Oligonucleotide 8-mer chained ligation (seq-by-ligation)	75 bp (paired-end sequencing available)
Ion Torrent	Emulsion PCR	Proton detection (seq-by-synthesis)	400 bp (bidirectional sequencing available)
PacBio	N/A (single molecule)	Phospholinked fluorescent nucleotides (seq-by-synthesis)	8,500 bp

The average read length is given for the platform/chemistry combination in each family that provides the longest reads.

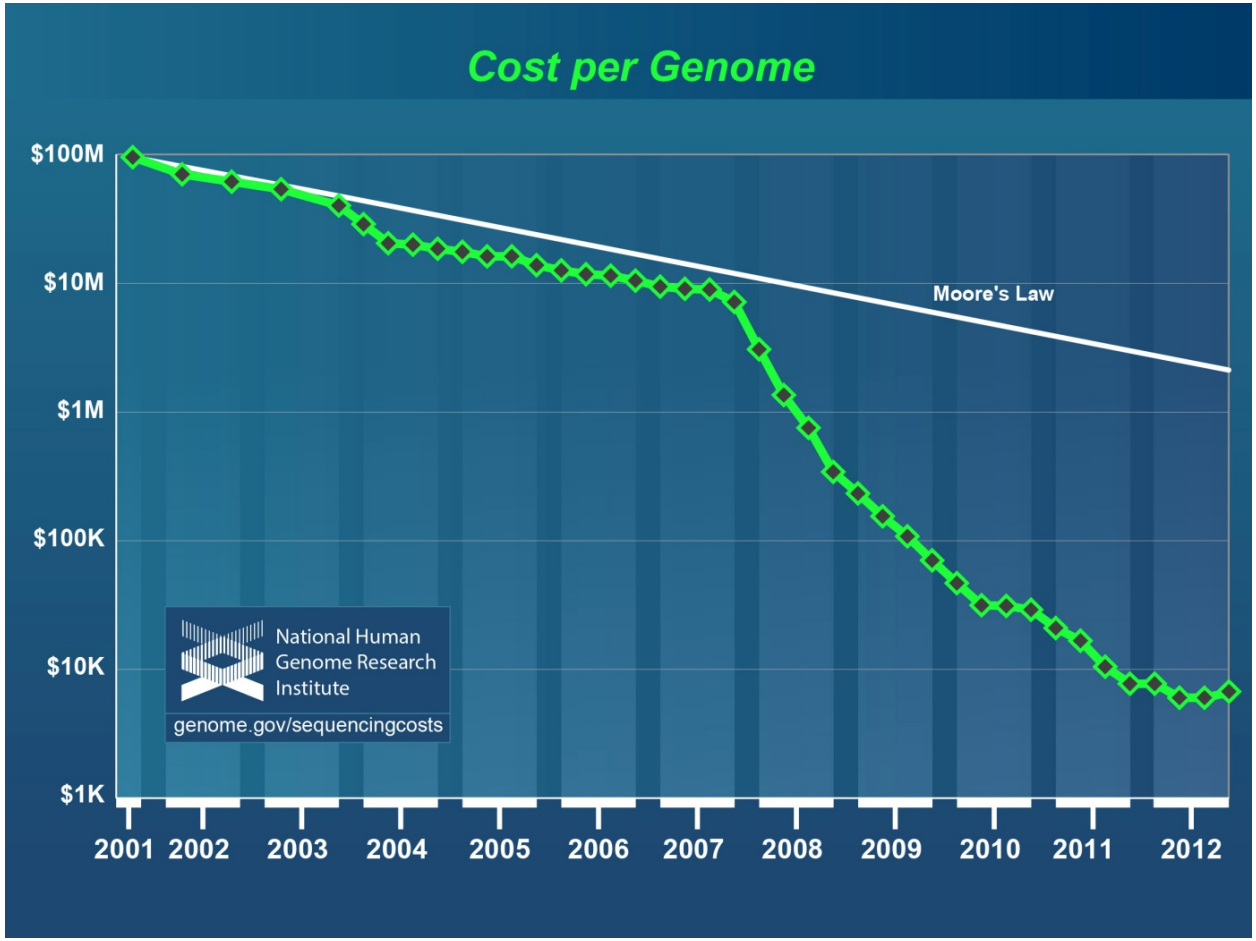
NGS PLATFORMS OVERVIEW

- Differ in design and chemistries
- Fundamentally related-sequencing of thousands to millions of clonally amplified molecules in a massively parallel manner
- Orders of magnitude more information-will continue to evolve

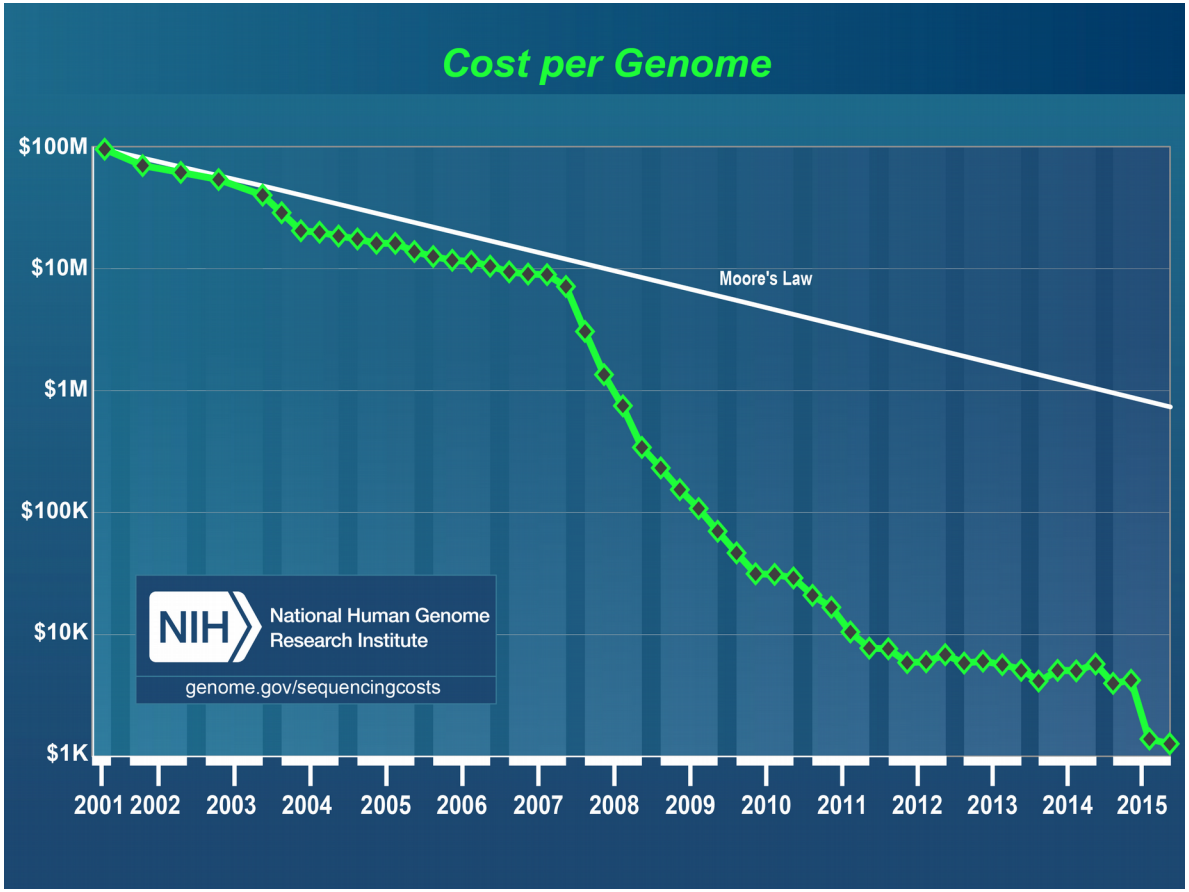


Pacific Biosciences
Helicos Biosciences
NABsys
VisiGen Biotechnologies
Complete Genomics
Oxford Nanophore Technologies

Sequencing costs have fallen

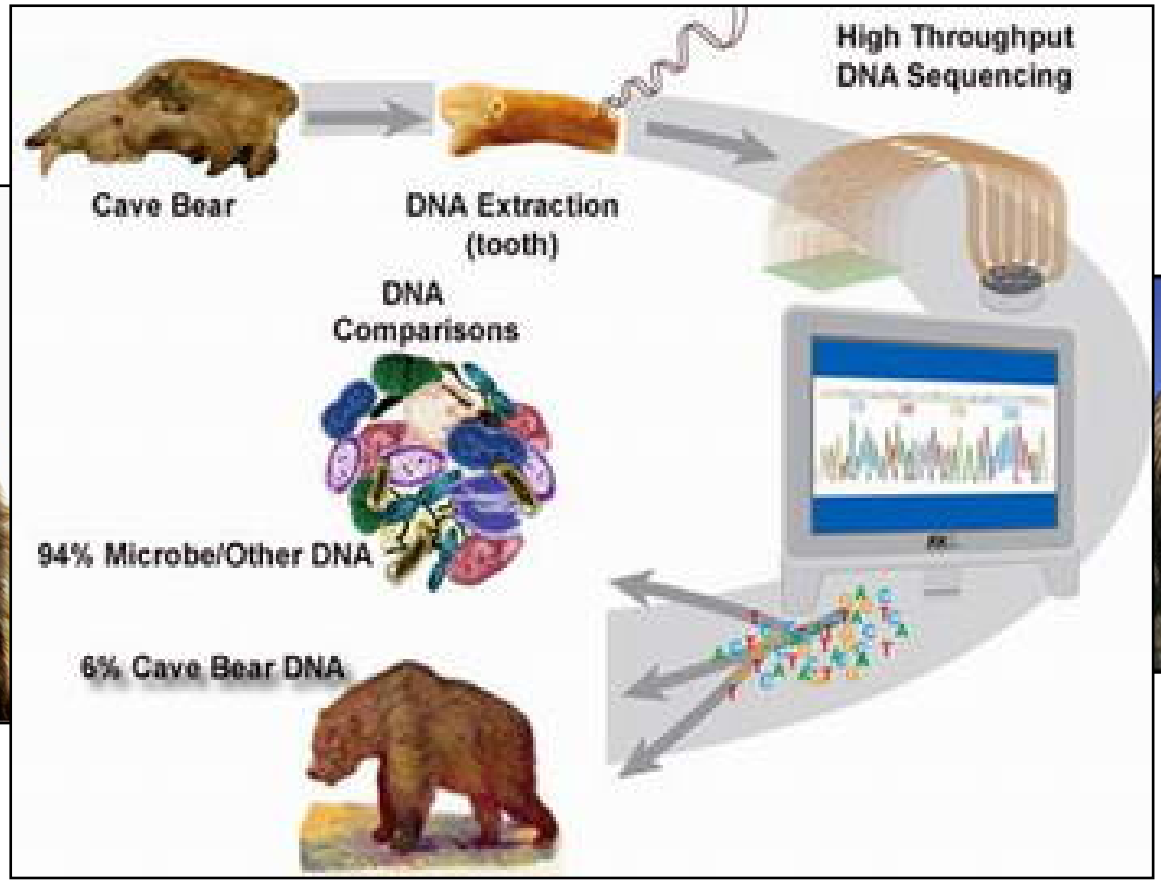


Sequencing costs have fallen



ANCIENT GENOMES RESURRECTED

- **Degraded state of the sample mitDNA sequencing**
- **Nuclear genomes of ancient remains: cave bear, mommoth, Neanderthal (10^6 bp)**



Problems: contamination modern humans and coisolation bacterial DNA

AB 3730 reduced

Megabace offline

454 Titanium

454 in production

Solexa in production

PacBio

Illumina HiSeq 2000

Ion Torrent

Illumina MiSeq

ONT

Solexa early access

SOLiD early access

Illumina GAIIx

454 1K

08/2007

12/2007

07/2005

01/2007

04/2007

10/2007

07/2008

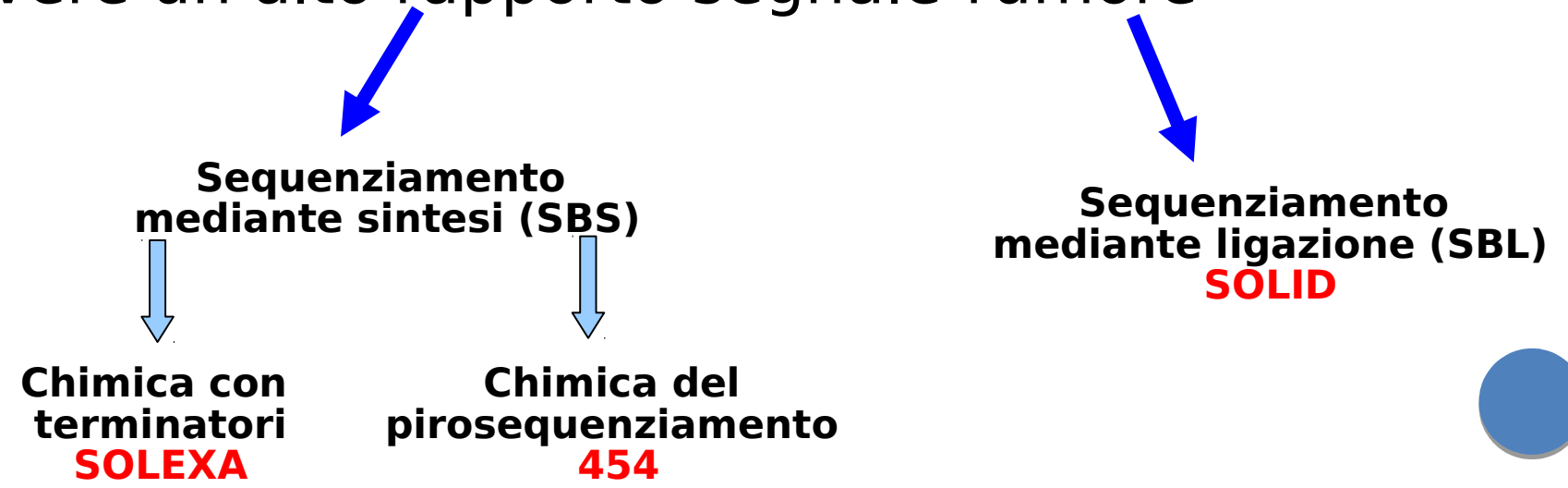
05/2009

12/2009

SEQUENZIAMENO DI NUOVA GENERAZIONE

Si basano sul principio del sequenziamento di '*cluster*' clonali

Il processo, che incomincia con una singola molecola target, prevede la creazione di targets clonali durante un processo intermedio di amplificazione. Copie multiple identiche sono infatti necessarie per avere un alto rapporto segnale-rumore



SEQUENZIAMENTO SANGER AD ALTA PROCESSIVITÀ

PREPARAZIONE DELLA LIBRERIA

Frammentazione casuale del DNA genomico
clonazione e trasformazione in batteri

↓

Raccolta delle colonie

↓

Purificazione del DNA dalle colonie
Sequenziamento Sanger
Elettroforesi capillare

↓

Whole genome *de novo* assembly or mapping to a reference (re-sequencing)

7-10 giorni
assumendo di possedere una piattaforma robotica per alta processività

Settimane-anni (!),
dipendentemente dalla dimensione del genoma (e copertura richiesta), dal numero di sequenziatori capillari

SEQUENZIAMENTO DI NUOVA GENERAZIONE

PREPARAZIONE DELLA LIBRERIA
Frammentazione casuale del DNA genomico
Ligazione degli adattatori

1 – 3 giorni

Amplificazione clonale dei frammenti

Sequenziamento mediante sintesi o ligazione

1 – 6 giorni

Processamento delle immagini

Mappatura delle reads su un genoma di riferimento
(o assemblaggio *de novo*)

Vantaggi delle piattaforme di nuova generazione

- Non sub-clonazione, non utilizzo di cellule batteriche *E. coli*
  - abolizione di *bias* di clonazione
  - rapidità nel preparare le librerie (non c'e' colony picking!)
- Ciascuna sequenza proviene da una molecola di DNA unica.
  - quantificazione attraverso 'conta' digitale
  - aumento del range dinamico
  - rilevazione di varianti rare
- Fornisce una eccezionale risoluzione per molti tipi di esperimenti (es. analisi di espressione, sequenziamento di DNA immunoprecipitato, di RNA piccoli, analisi di medie/grandi inserzioni-delezioni nei genomi….)
- Rivoluzionaria diminuzione del costo e del tempo per generare dati di sequenza (lavorano in multi-parallelo)
- Richiesta meno robotica nelle fasi precedenti al caricamento sul sequenziatore

Svantaggi delle piattaforme next-gen

- Sono prodotte sequenze più corte

  - relativamente alle sequenze da sequenziatori capillari (metodo Sanger)

  - è necessario ri-parametrizzare l'accuratezza della procedura di chiamata delle basi

  - enorme difficoltà nell'analisi dei dati; richiesto un grande sforzo di programmazione per costruire nuovi algoritmi.

- La mole enorme di dati 'traumatizza' le infrastrutture informatiche.

  - da 10 Gb a diversi Tb di dati grezzi prodotti per corsa (dipende dalla piattaforma)

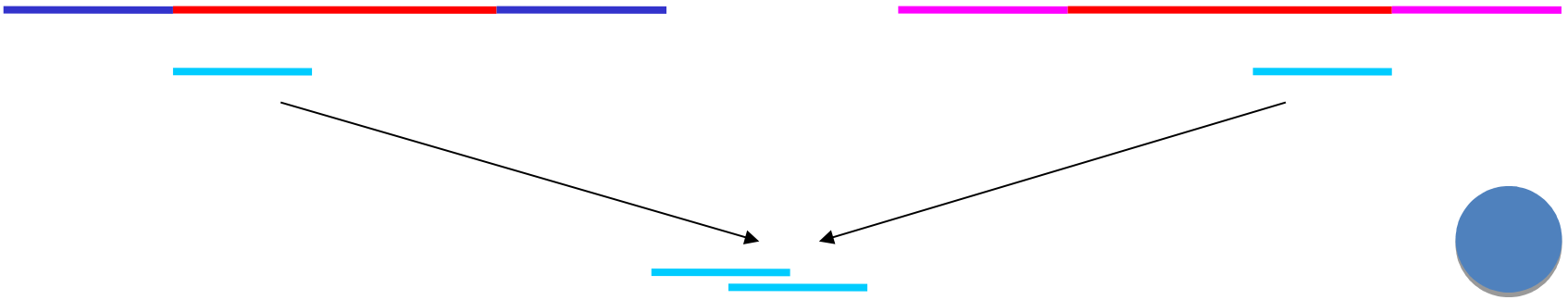
  - il processamento delle *read* tramite *pipeline* informatiche richiede molta capacità di calcolo (CPU)

  - è necessario prendere accurate decisioni su cosa salvare e cosa cancellare

SEQUENZE CORTE

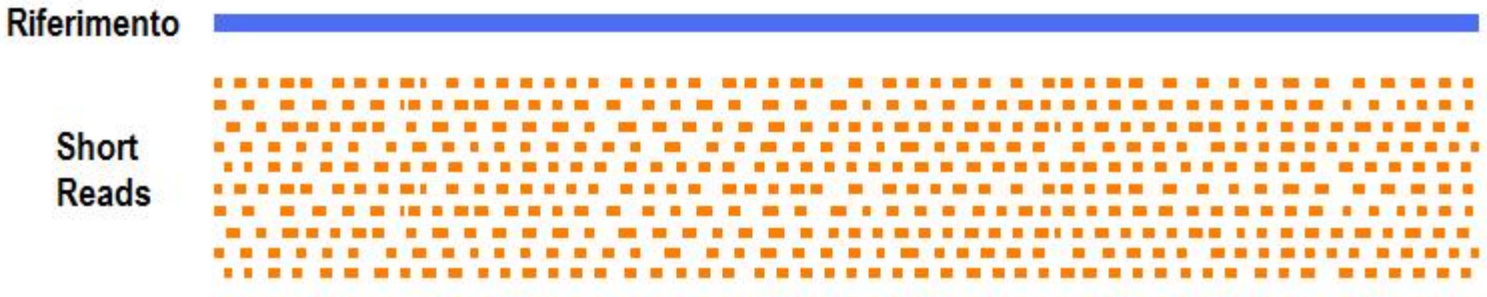
- Sequenze corte, ma tecnologia in continua evoluzione:
  - 454: 100 basi → 200 → **400-500** → ?
  - Solid: 25 basi → 35 → **50** → 100 → ?
  - PGM: 200 → 400
  - Illumina: 32 → 36 → 75-100 → **125 → 150** → 250?

- Difficoltà di assemblare sequenze corte *de novo*, soprattutto per il problema delle sequenze ripetute complicato ancora di più rispetto a Sanger (lunghezza media 700-900bp)

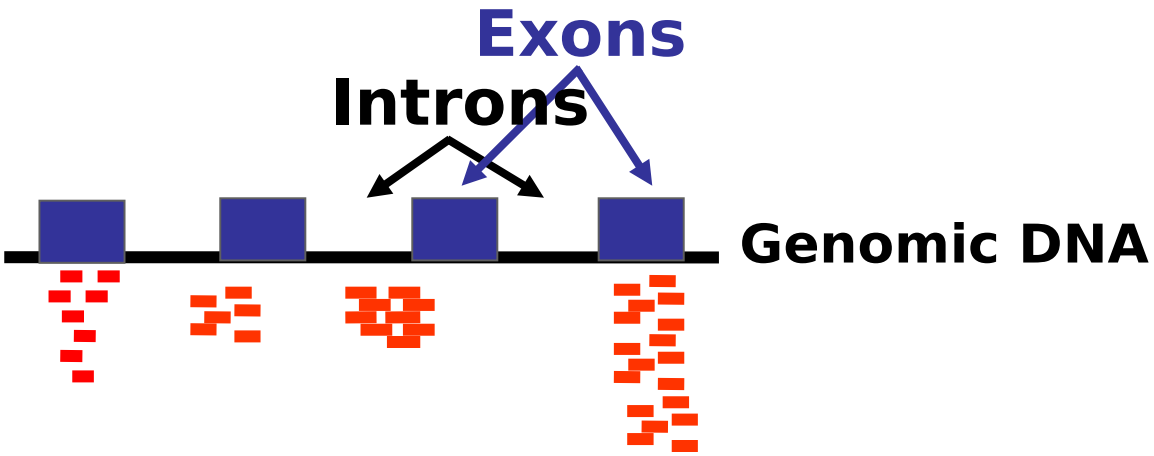


RISEQUENZIAMENTO

- In presenza di un genoma di riferimento di buona qualità posso effettuare un ri-sequenziamento e allineare tutte le reads ottenute:

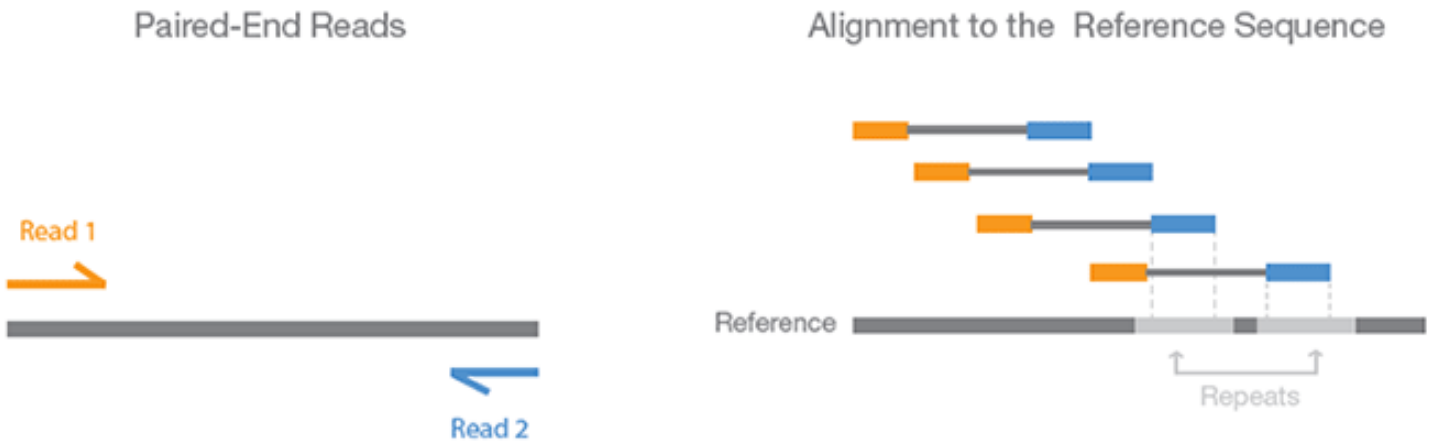


- Non solo del genoma, ma anche del trascrittoma



Paired-end (PE)

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

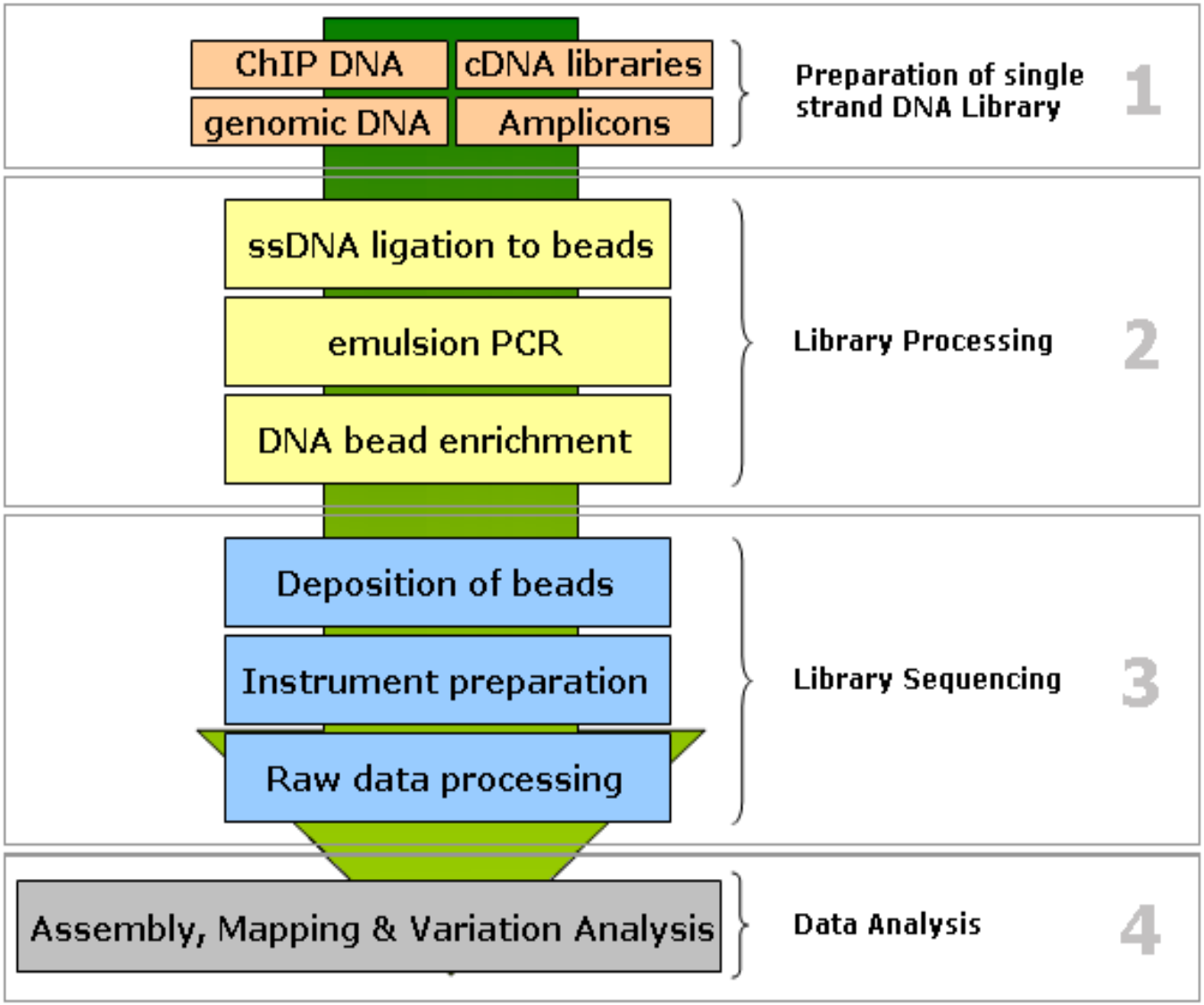
IL PROBLEMA (!) DELLA ENORME MOLE DI DATI PRODOTTA

- E’ un problema chiave che limita una più ampia adozione di questi strumenti da parte dei laboratori
- 1 ABI3730xl genera fino un max di 260 milioni di paia di basi di sequenza all’anno
- Quando nel 2004-2005 è stato lanciato il primo 454 produceva una quantità di dati in un anno superiore a quella prodotta da più di 50 ABI3730xl
- Il problema dell’ ‘indigestione’ di dati è dal 2005 ulteriormente peggiorato sia per il 454 che a causa della possibilità di scelta anche delle altre due piattaforme (Illumina/Solexa lanciata sul mercato nel 2006 e Solid nel 2007)
- Produzione una decina di gigabytes di dati per corsa per 454, 1-4 terabytes di dati per corsa per Illumina e Solid

SEQUENZIAMENTO CON LA TECNOLOGIA 454

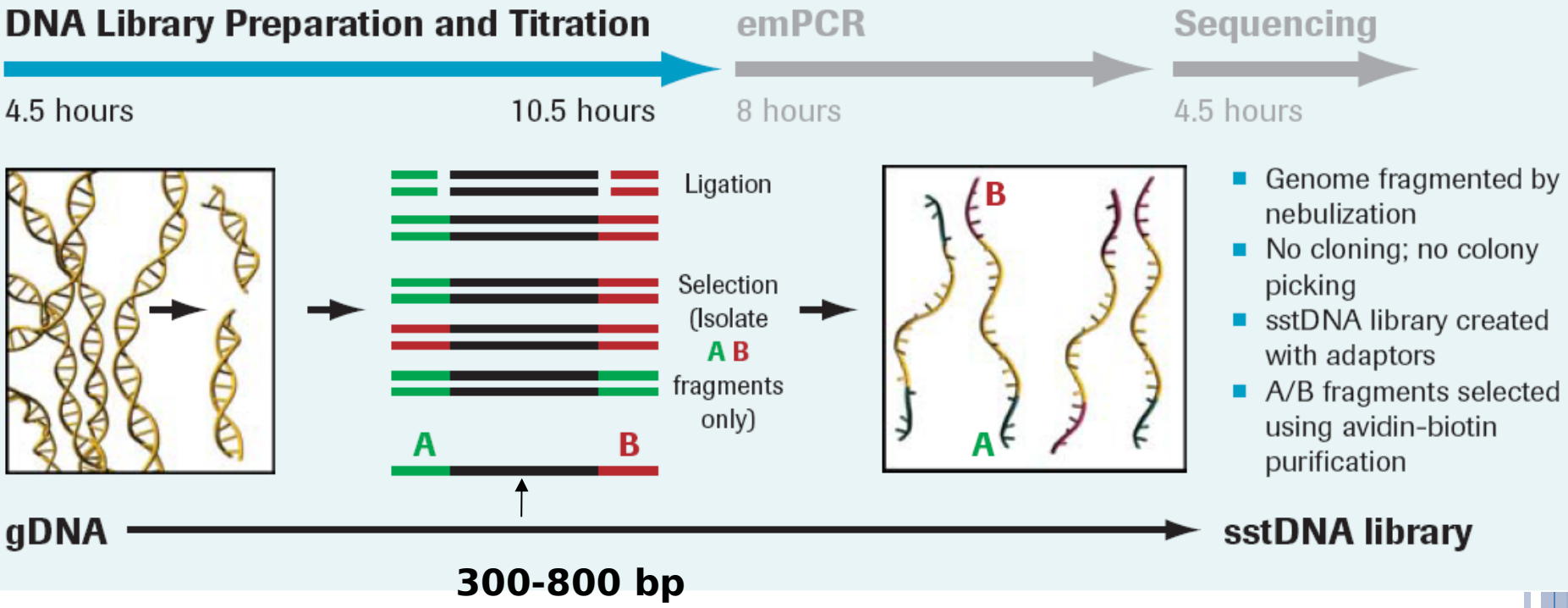
Workflow

(Click on the chart)

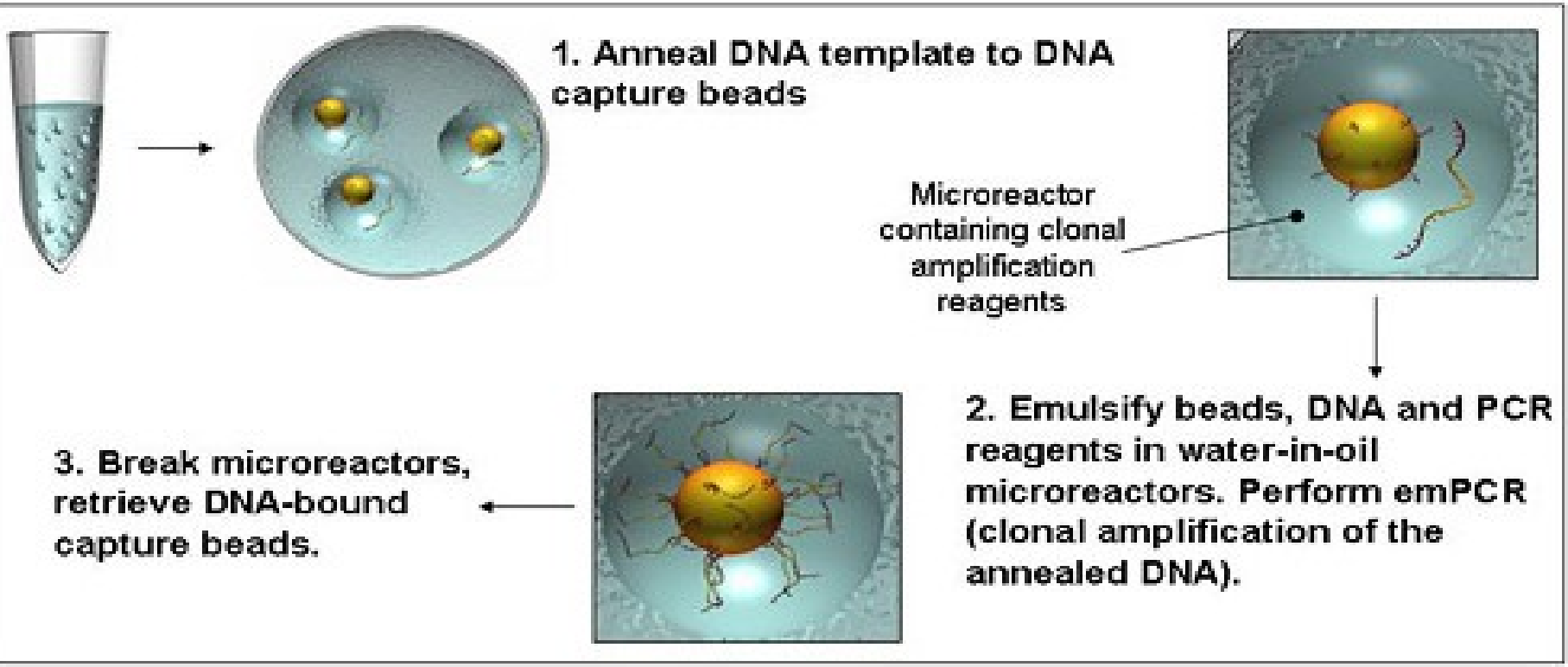


>>top

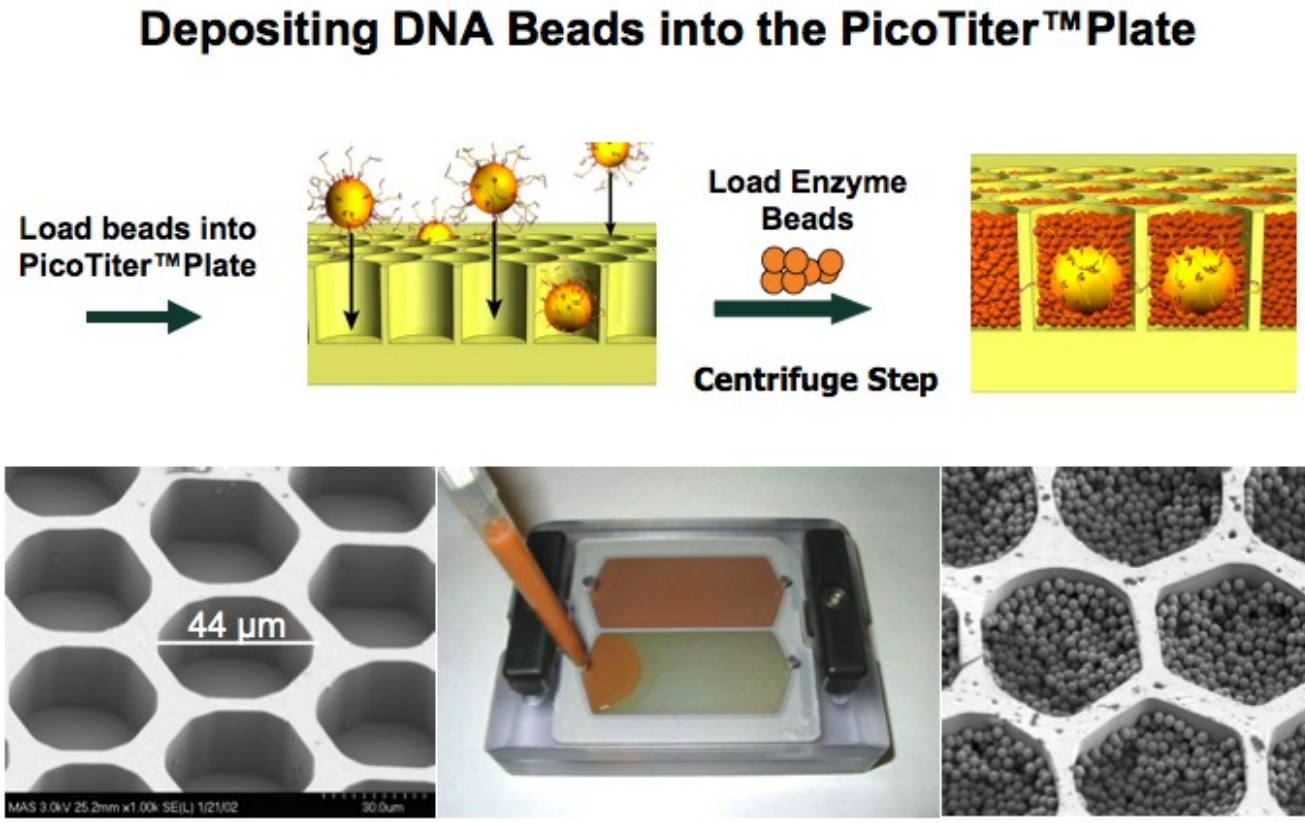
Tecnologia 454



emPCR™: Clonal Amplification of Annealed DNA fragments



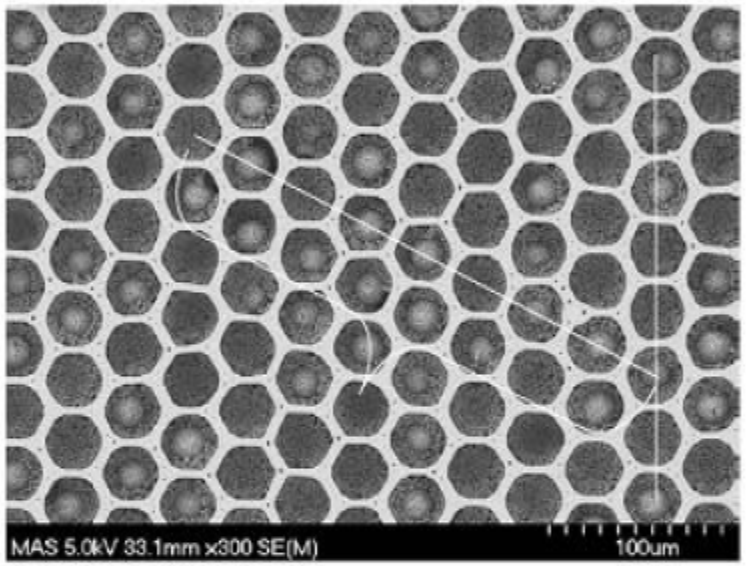
Tecnologia 454

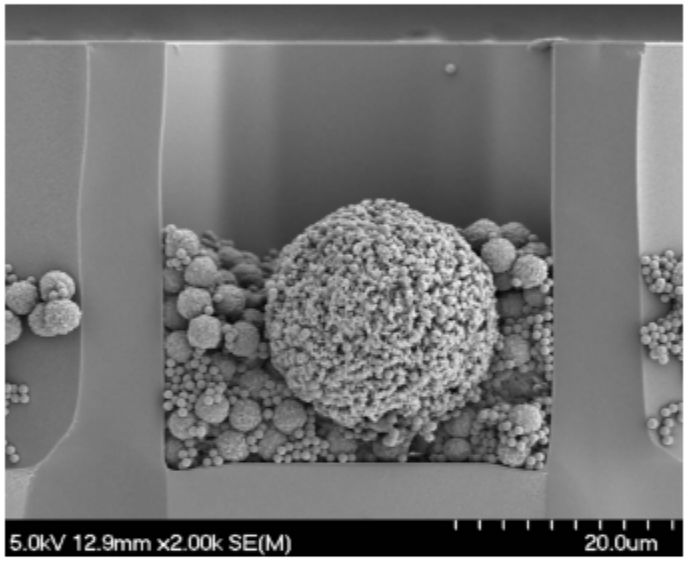


- **Il sequenziamento inizia con la preparazione della piastra PicoTiter. Durante questo passaggio una miscela di beads, enzimi per il sequenziamento e la libreria sstDNA vengono depositati nei pozzetti di 44um**
- **Il processo di deposizione delle beads massimizza il numero di pozzetti che contengono un frammento individuale della libreria sstDNA**
- **La piastra PicoTiter viene caricata sul sequenziatore**

TITANIUM

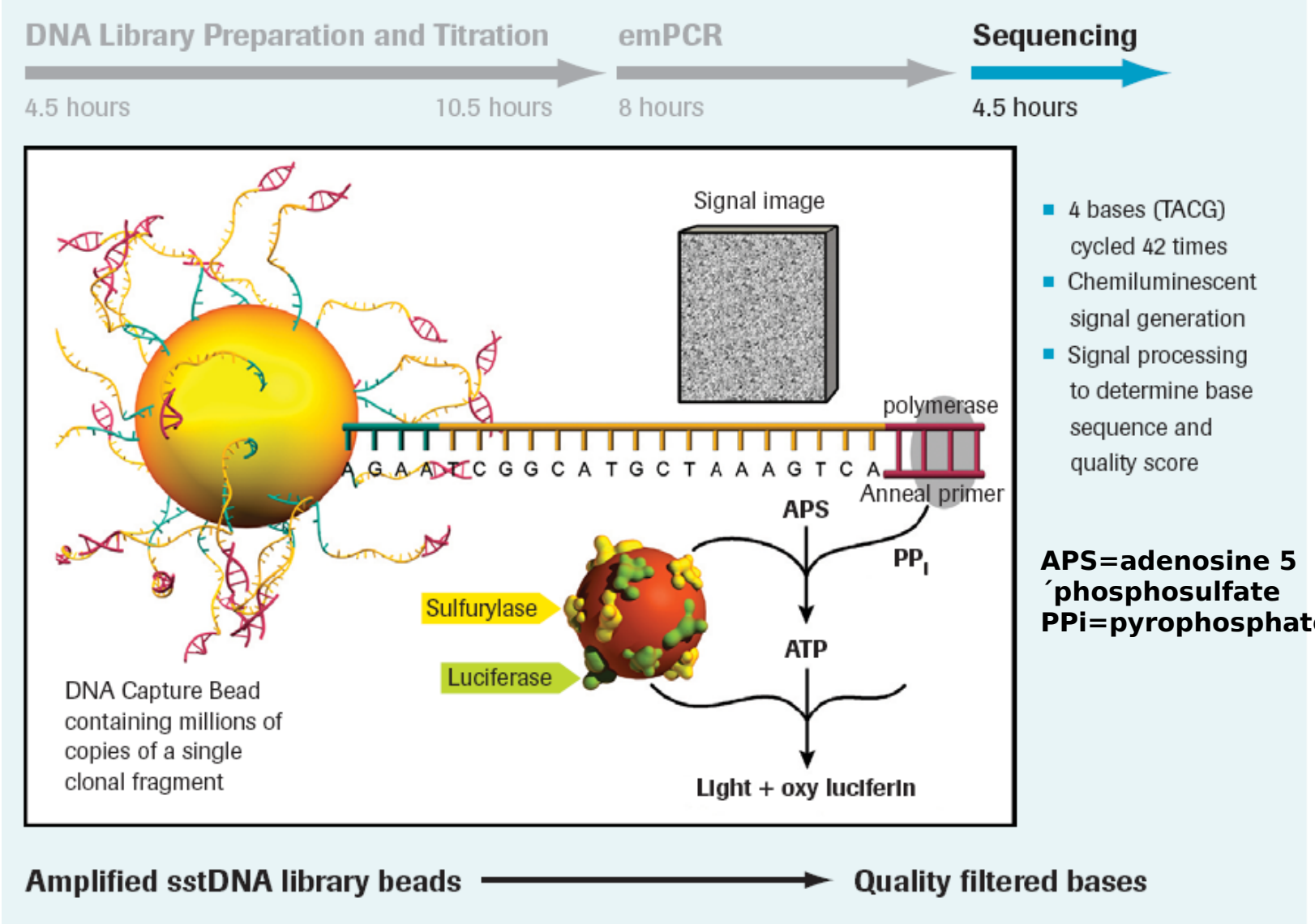
More wells: New PTP and Beads



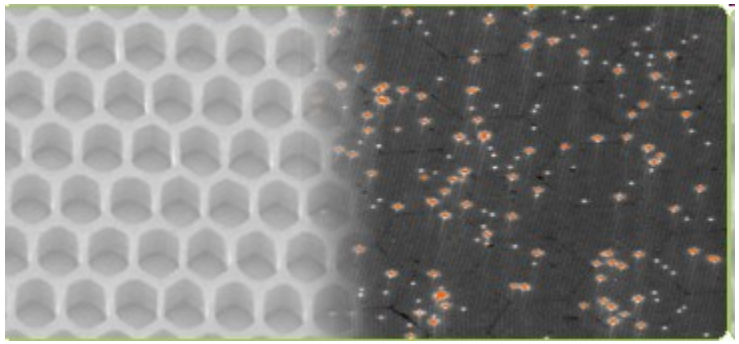


- Pitch changed from 50 micron to 34 micron
- Number of wells increased from 1.6M to 3.4M
- Cross talk significantly reduced to permit higher loading

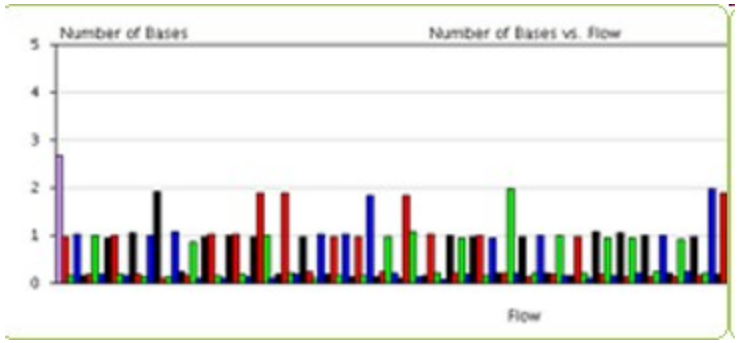
Tecnologia 454



For each cycle four pictures are captured (one picture per nucleotide);



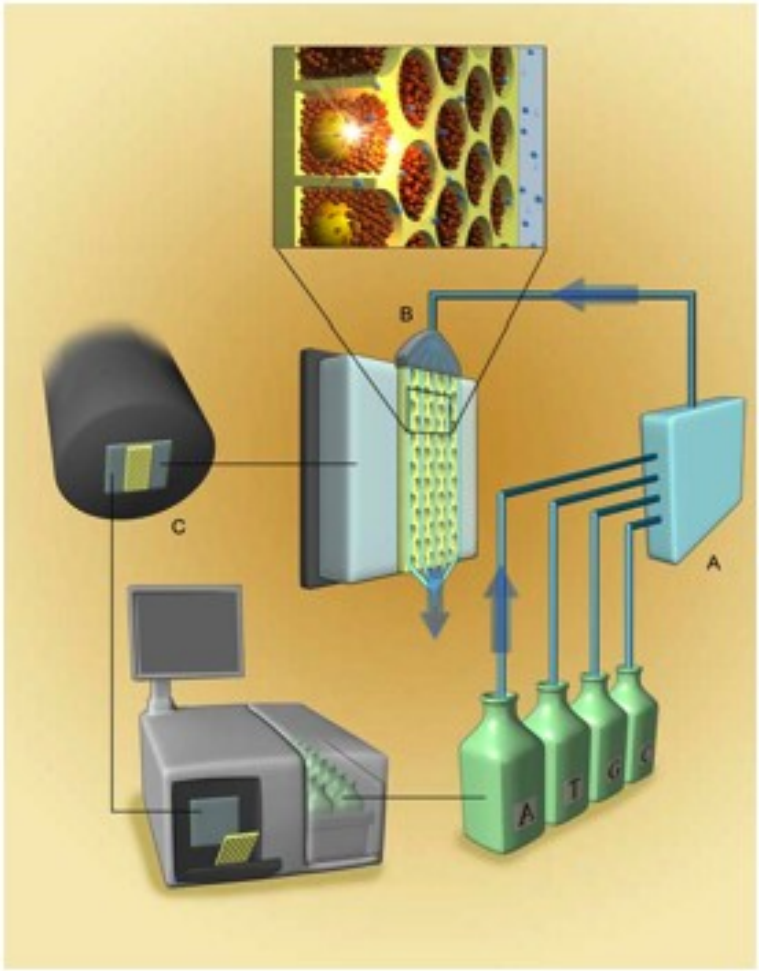
Extraction, Qualification/Quantification and Normalization of wells data

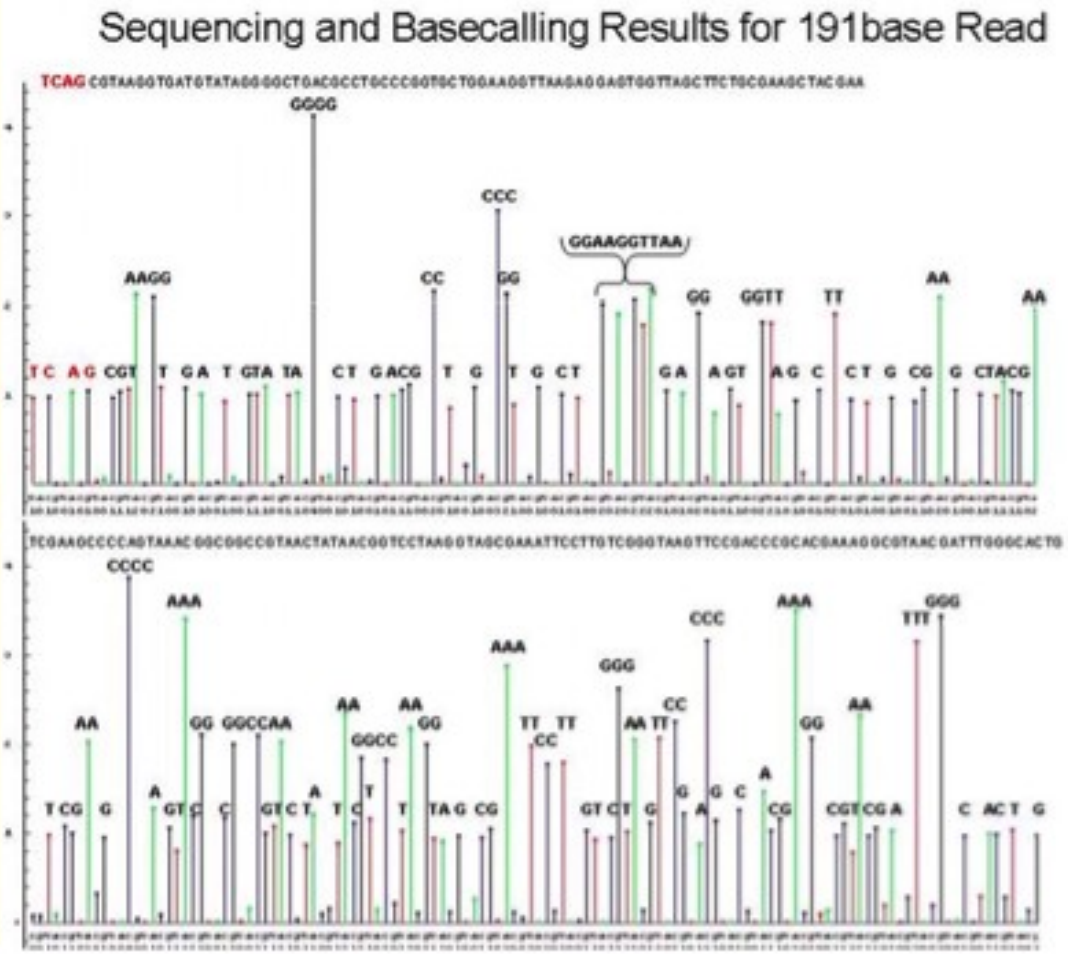


Read data are converted into "flowgrams".

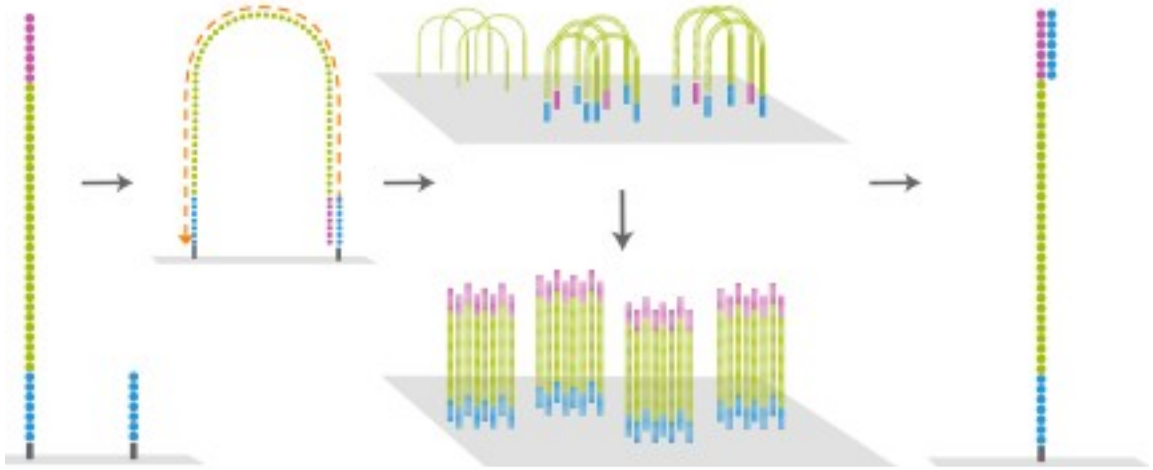
Tecnologia 454

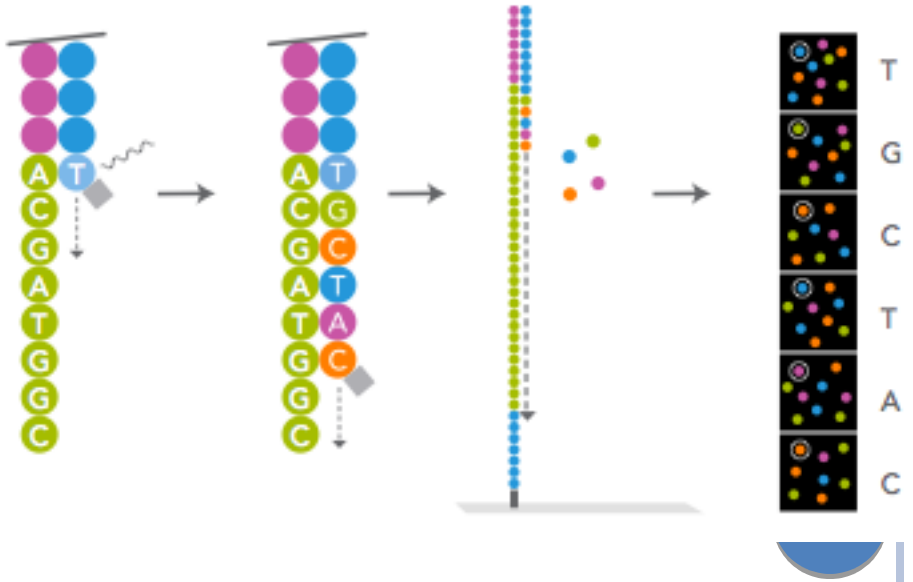
454 Technology - Sequencing Instrument





Illumina

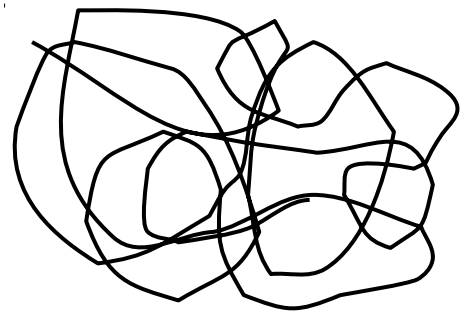




Library Construction

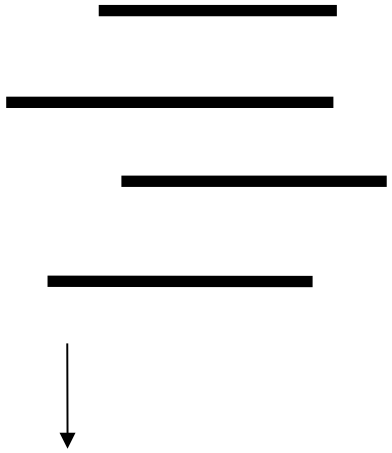
1. Fragmentation

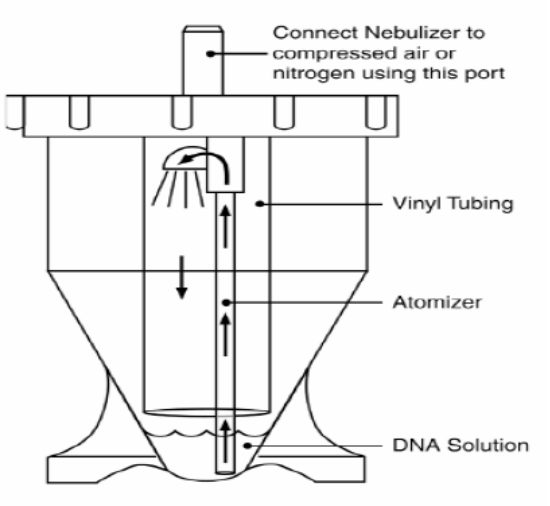
ds DNA (genomic, cDNA)

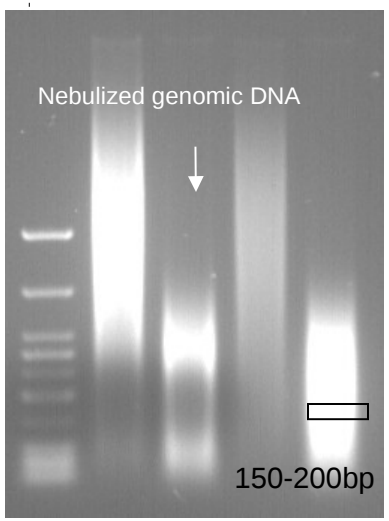


Nebulization

Covaris sonicator







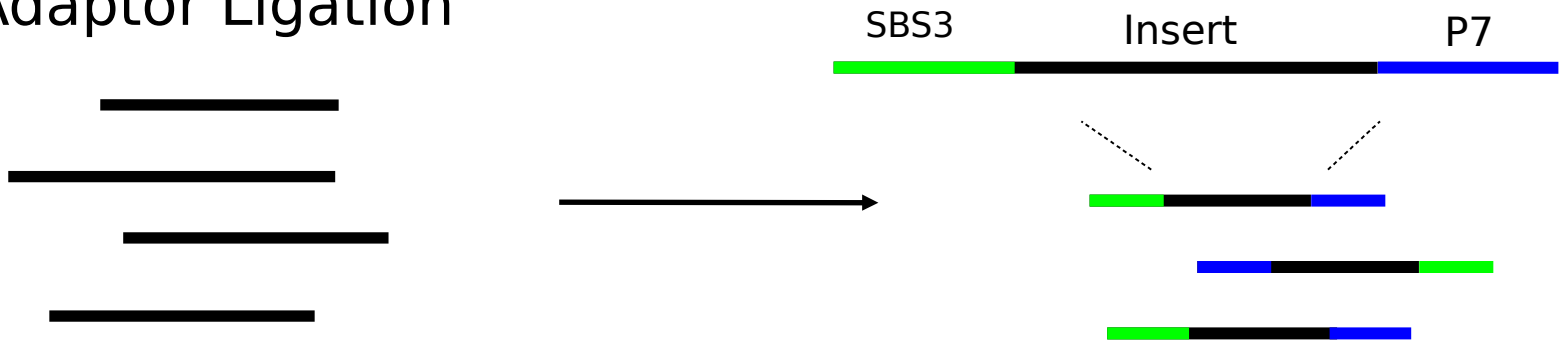
2. End repair and 'A' addition

3' overhangs removed by exonuclease (Klenow fragment)
5' overhangs extended by T4 DNA polymerase
Blunt ends phosphorylated by kinase

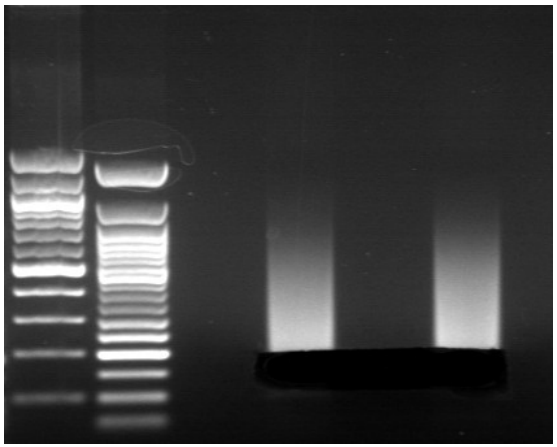
Added an 'A' base to the 3' end of the blunt phosphorylated fragment

Library Construction

3. Adaptor Ligation



4. Size selection by gel recovery



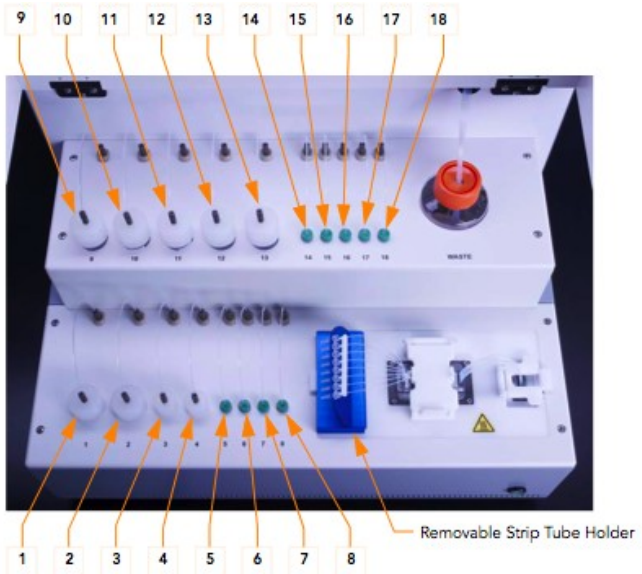
LIBRARY CONSTRUCTION

4. PCR Enrichment

5. Library validation (i.e. plasmid cloning/Sanger sequencing or quantitative PCR)

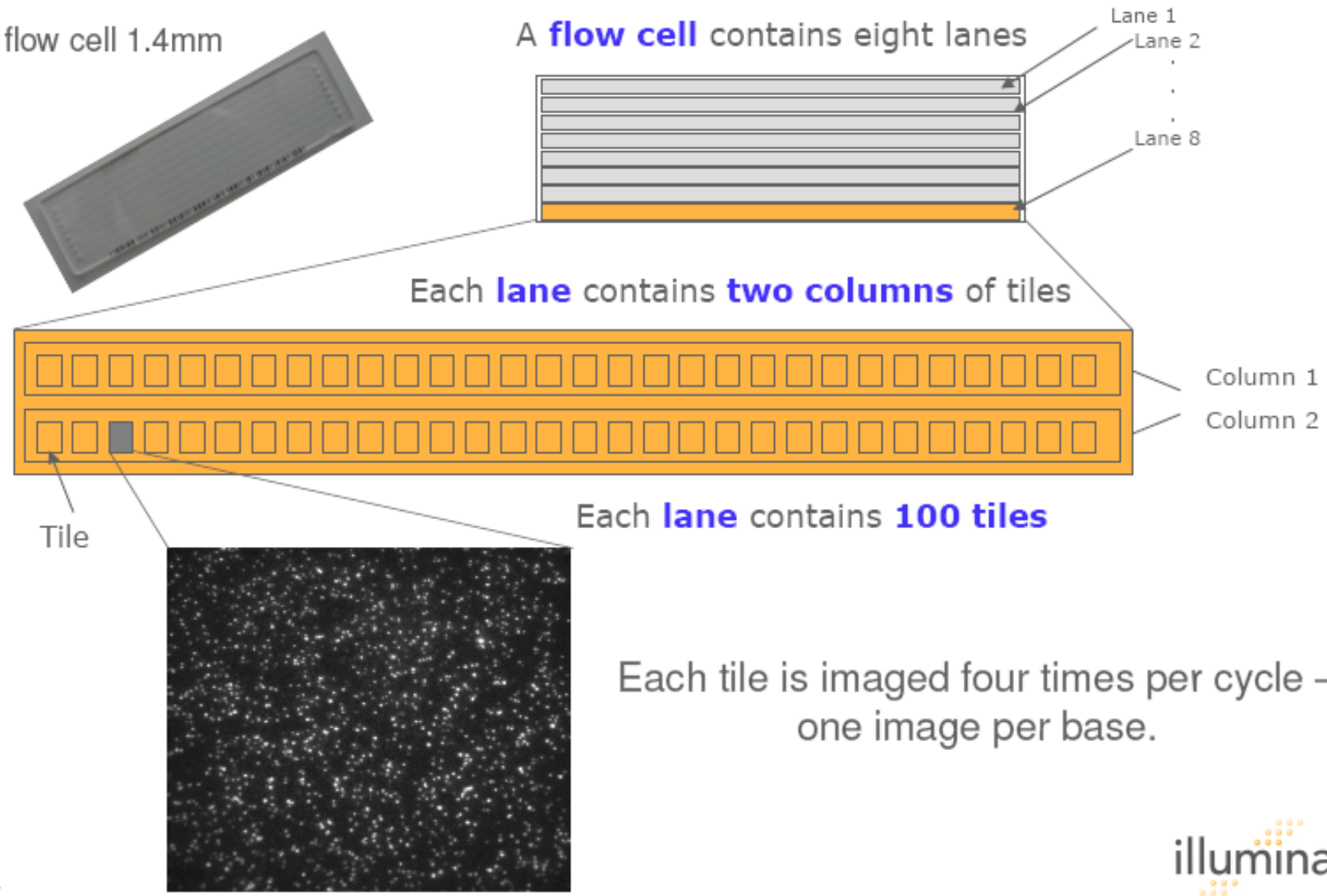
6. Library quantification (to have ~100K cluster I need 2-4pM library)

Cluster Station



Strumento che permette di preparare la *flow-cell* (=supporto di vetro su cui i frammenti della libreria verranno sequenziati in parallelo)

Illumina flow cell

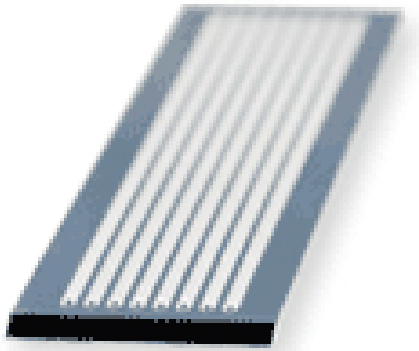


Each tile is imaged four times per cycle – one image per base.

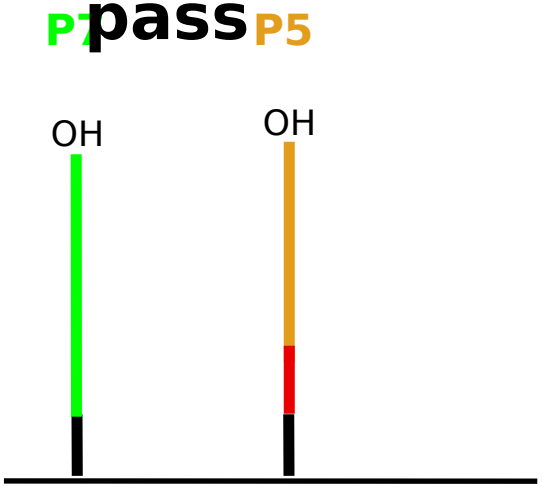
illumina

Cluster Generation

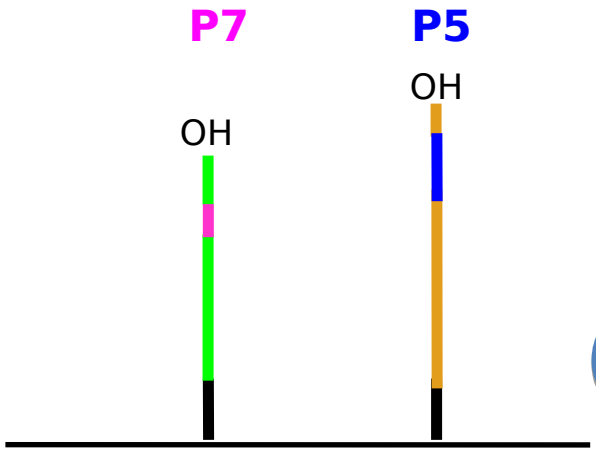
Grafted Flow Cells



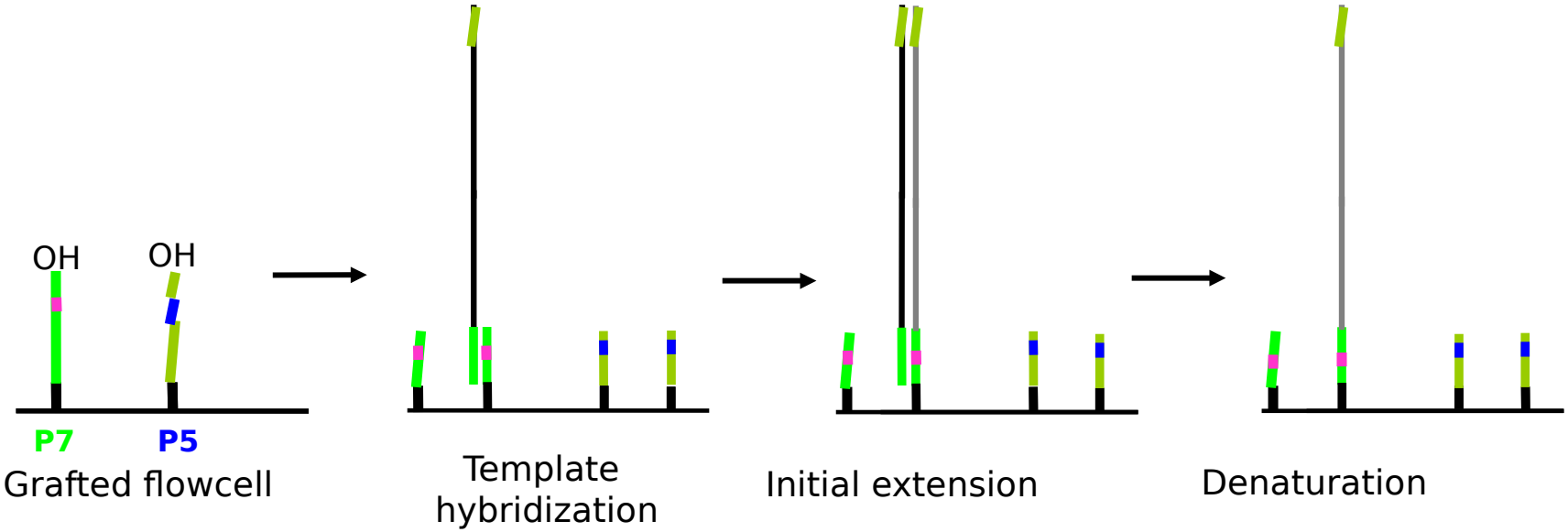
Single pass



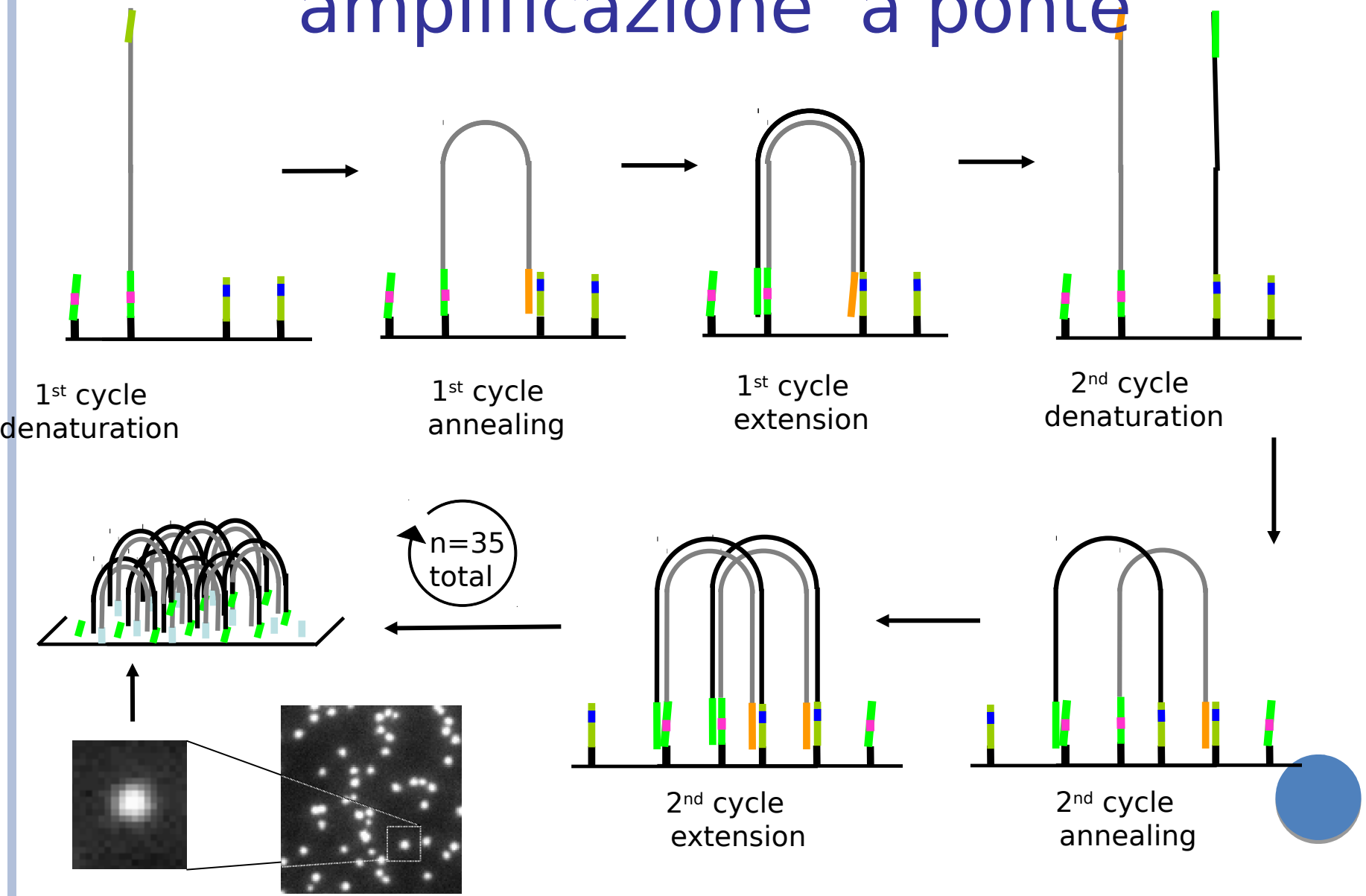
Paired end



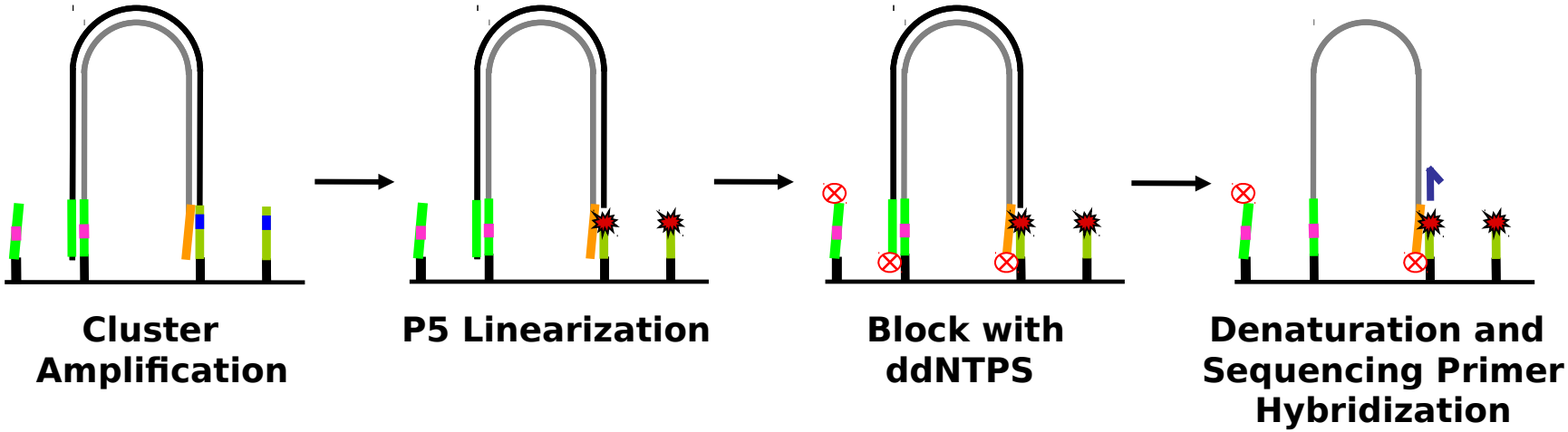
Cluster Generation



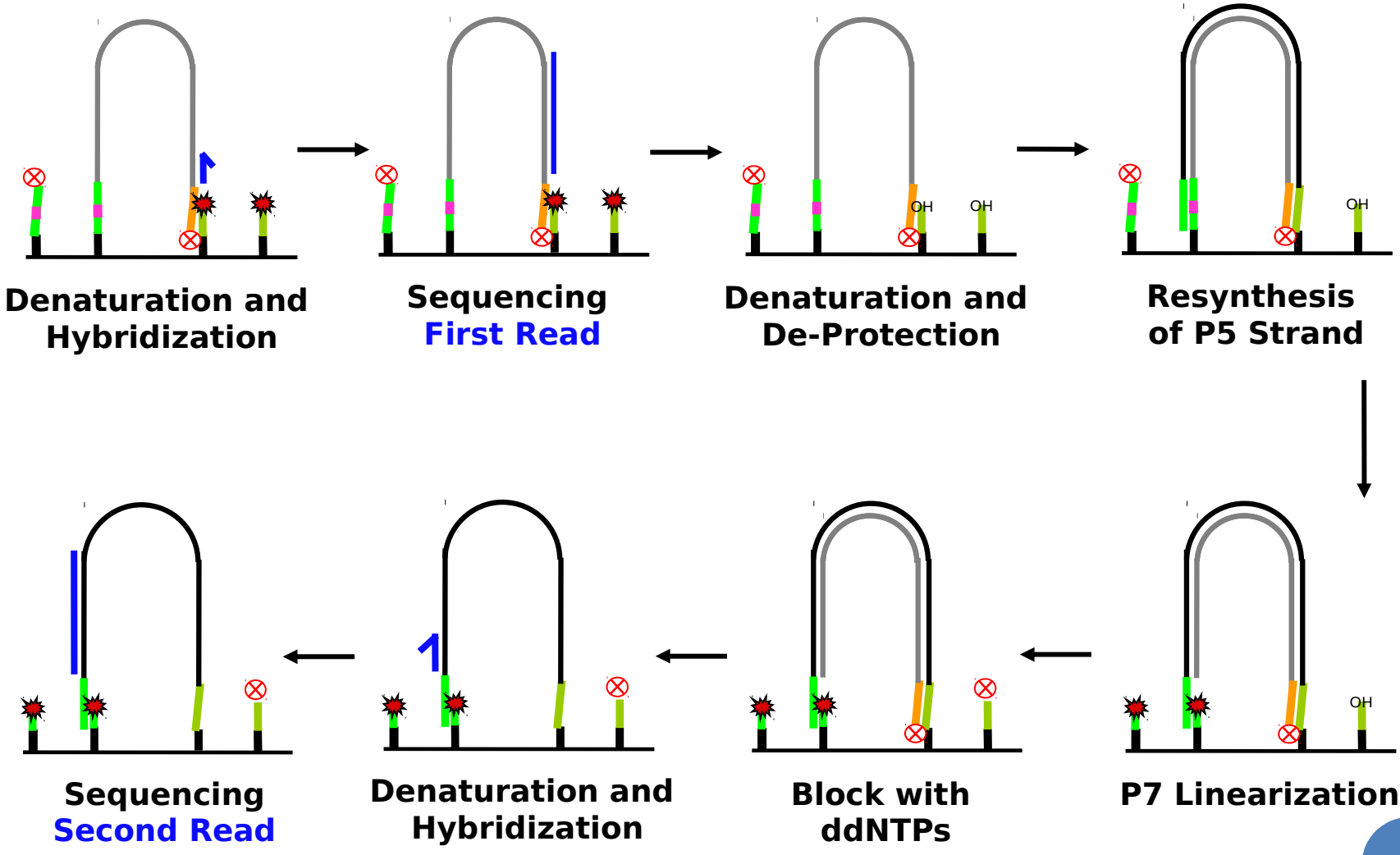
La generazione dei cluster: amplificazione a ponte



La generazione dei cluster

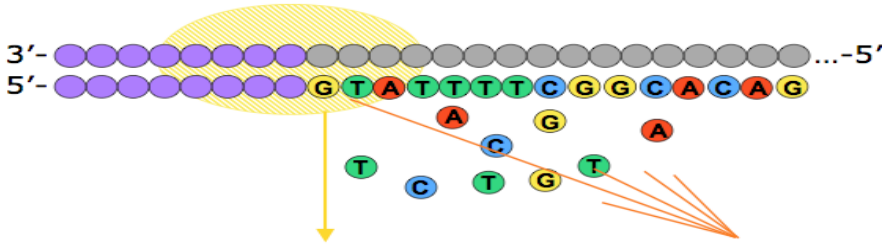


Sequenziamento pair-end



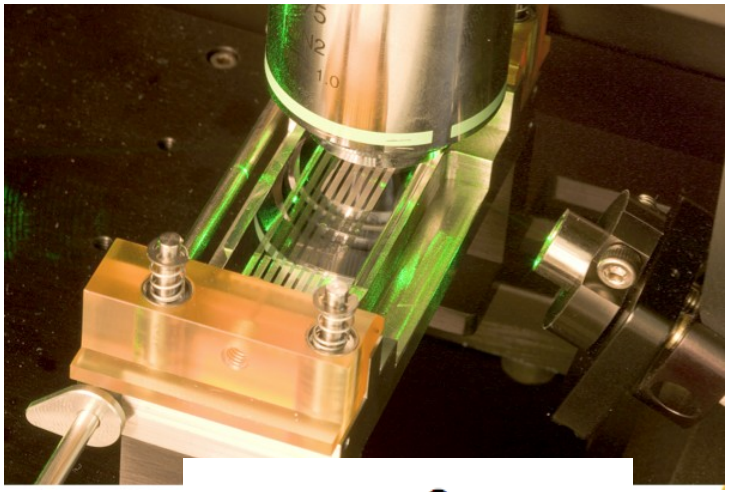
Sequencing Chemistry

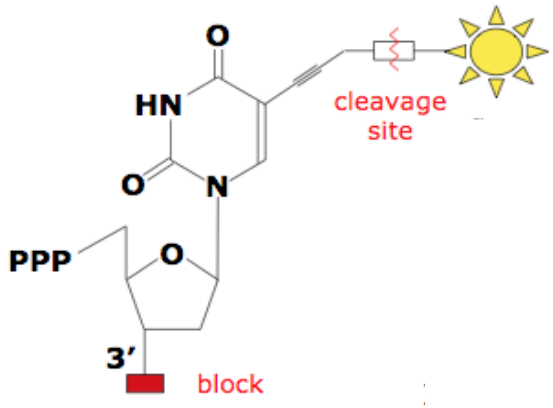
Sequencing by Synthesis



Cycle 1: Add sequencing reagents
First base incorporated
Remove unincorporated bases
Detect signal

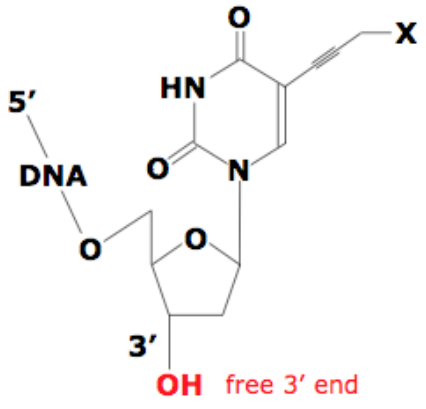
Cycle 2-n: Add sequencing reagents and repeat

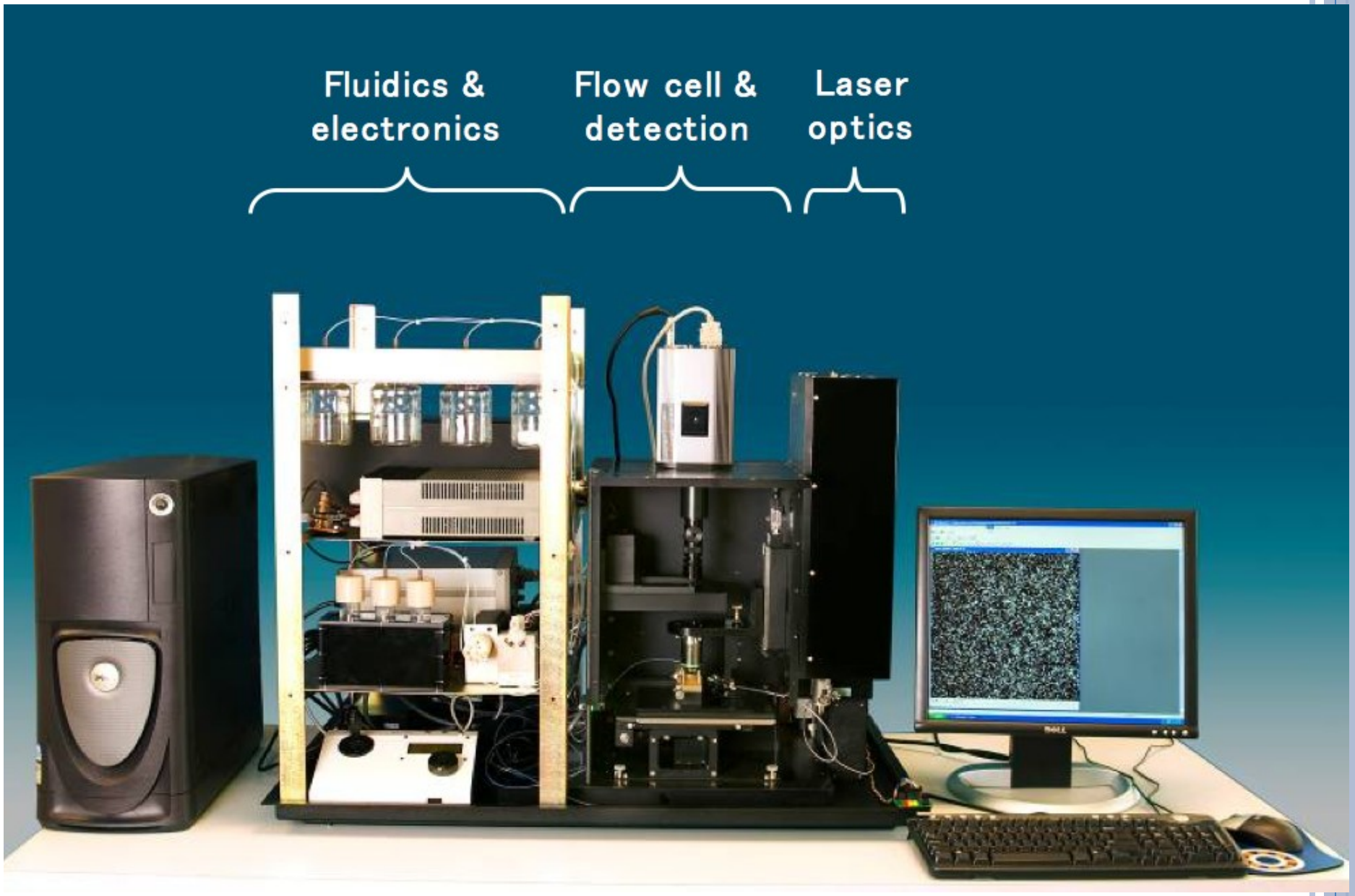




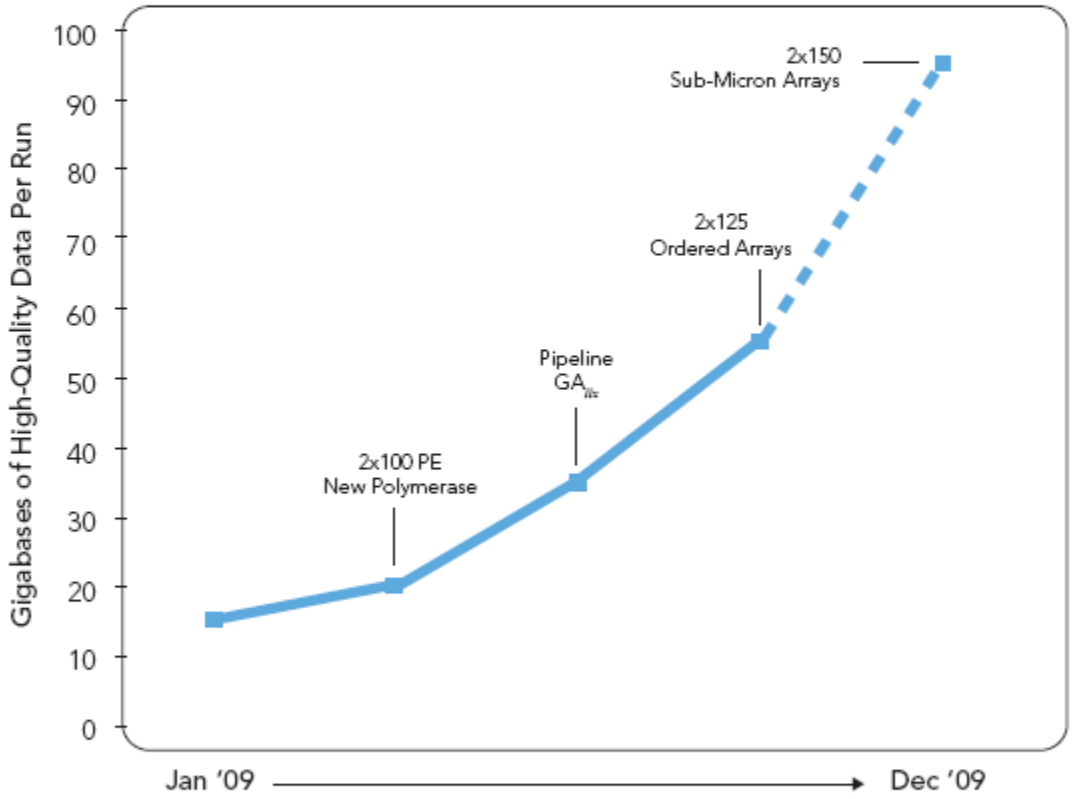
Cleave fluorophore

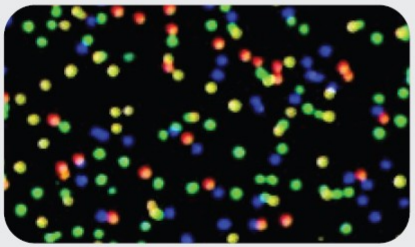
De-block 3′ terminus



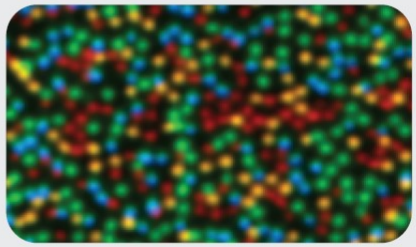

Fluidics & electronics
Flow cell & detection
Laser optics

umento progressivo della lunghezza delle reads e della processi





Random Array Clusters



Semi-ordered Array

CARATTERISTICHE DELLA TECNOLOGIA SOLEXA/ILLUMINA

- **Immobilizzazione della libreria su *flow-cell***
- **Amplificazione clonale tramite PCR a ponte (‘bridge amplification’)**
- **Sequenziamento mediante sintesi (SBS)=simile a Sanger ma la sintesi della catena continua**
- **Sequenze corte (75-100 bp, ma promesse 150bp entro l’anno)**
- **Accuratezza 99-99,50% (0.5-1.0% di errore)**
- **95 Gbp/corsa**

Illumina MiSeq (improvement)

Illumina HiSeq 2500

Ion Torrent PGM

Ion Torrent Proton

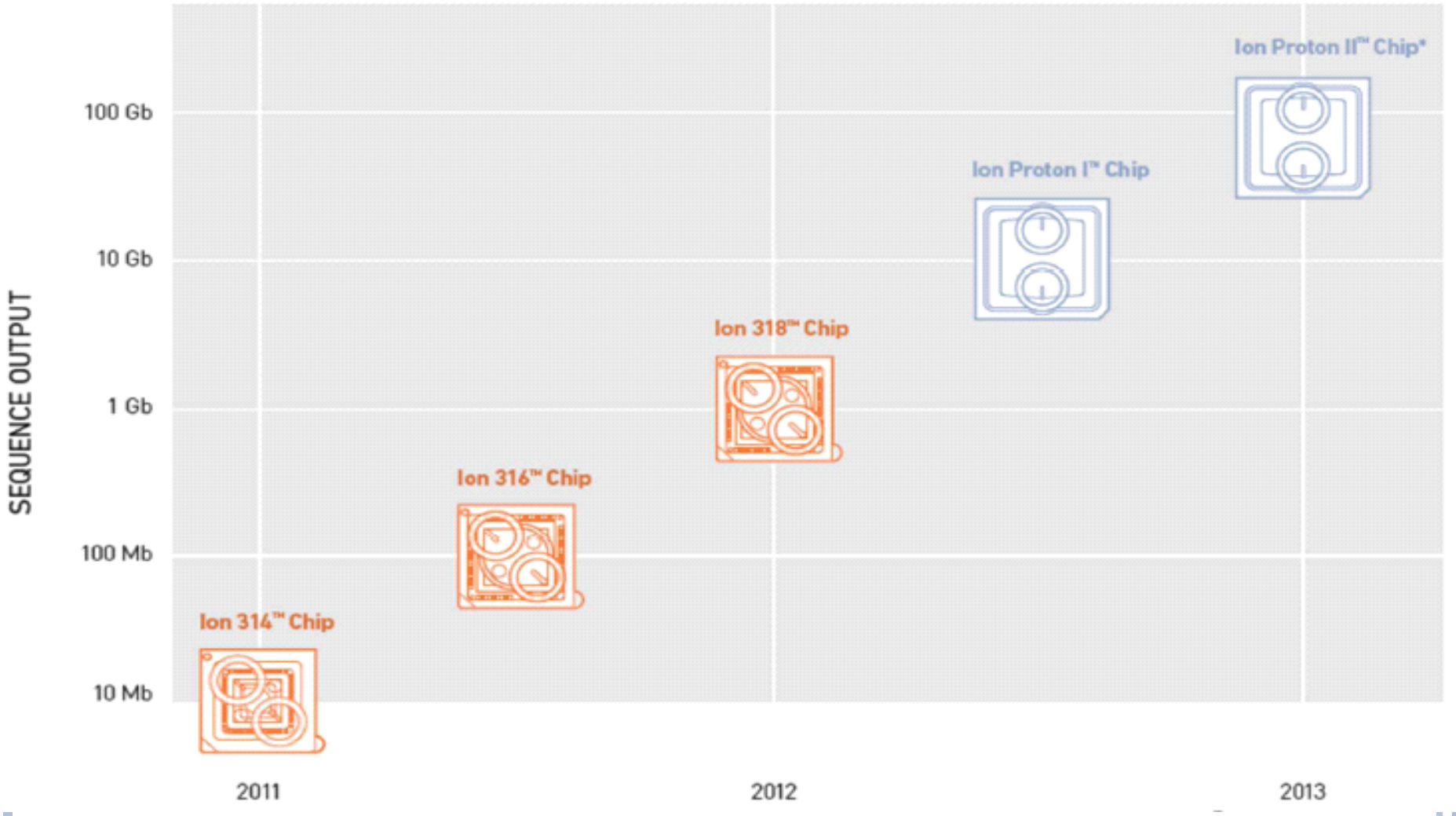
ILLUMINA IMPROVEMENT

- Longer read length (250 bp)
- 3-fold more reads (15 M)
- Higher throughput (5-7 Gb)
- Faster run time

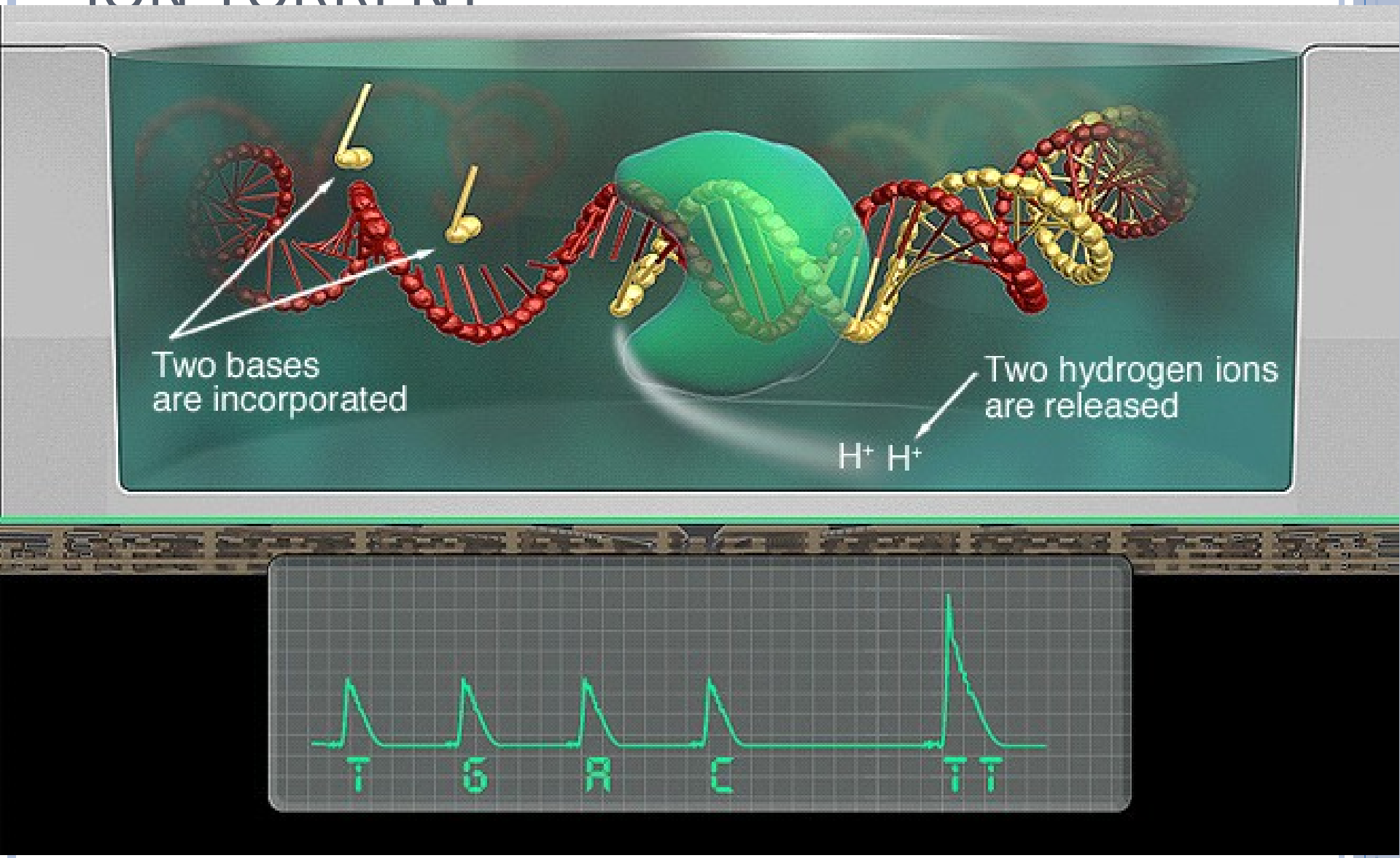
Two run configurations

- Fast run config can be done in 27 hours and produce 120 Gb
- Standard run config remains the same (600 Gb in 17 days)

PROMISES FROM ION TORRENT



ION TORRENT


Two bases
are incorporated
Two hydrogen ions
are released
H+ H+
T G A C TT

OVERALL PREFORMANCES

Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314'chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+[e]	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316'chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318'chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (4hq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i]*	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)

Field Guide to next-generation DNA Sequencers, Glenn (2011)

Costs and Error Rate

Table 3 Instrument purchase cost, additional instrument costs, service agreement costs, computational resources needed, size of data files, primary errors and error rates for commercially available DNA sequencing platforms in 2011. All costs are list price in thousands of US dollars

Instrument	Purchase cost	Additional instruments[a]	Service contract[b]	Computational resources[c]	Data file sizes (GB)[d]	Primary errors	Error rate (%)[e]
3730xl (capillary)	$376	–	$19.8	Desktop	0.03	Substitution	0.1–1
454 GS Jr. Titanium	$108	$16	$12.6	$5 (desktop)	<3 images, <1 sff	Indel	1
454 FLX Titanium	$500	$30	$50.0	$5 (desktop)	20 images, 4 sff	Indel	1
454 FLX+[f]	$29.5	$30	$50.0	$5 (desktop)	~40 images, 8 sff	Indel	1*
PacBio RS	$695	–	$85	$65 cluster	20 pulsed, 2 Fastq	CG deletions	16
Ion Torrent – 314 chip	$49.5	$18[g]	$7.5	Desktop – $35	0.1Fastq	Indel	~1
Ion Torrent – 316 chip	$49.5	$18[g]	$7.5	Desktop – $35	0.6Fastq	Indel	~1*
Ion Torrent – 318 chip	$49.5	$18[g]	$7.5	Desktop – $35	TBD	Indel	~1*
SOLiD – 4	$475	$54[h]	$38.4	$35 cluster[i]	680[j]	A-T bias	>0.06*
SOLiD – 5500	$349	$54[h]	$29.0	$35 cluster[i]	74[k*]	A-T bias	>0.01*
SOLiD – 5500xl	$595	$54[h]	$38.4	$35 cluster[i]	148[k*]	A-T bias	>0.01*
Illumina MiSeq	$125	–	$12.5	Desktop	1[k*]	~Substitution	>0.1*
Illumina HiScanSQ	$405	$55[l]	$41.5	$222 cluster[m]	50[k*]	Substitution	≥0.1
Illumina GAIIx	$250	$100[n]	$44.5	$222 cluster[m]	600	Substitution	≥0.1
Illumina HiSeq1000	$560[o]	$55[l]	$62.0	$222 cluster[m]	≤300[k*]	Substitution	≥0.1
Illumina HiSeq2000	$690	$55[l]	$75.9	$222 cluster[m]	≤600[k*]	Substitution	≥0.1

Field Guide to next-generation DNA Sequencers, Glenn (2011)

Jonathan M. Rothberg

- 1999: founded 454 Life Sciences
- 2007: founded Ion Torrent

Pacific Biosciences: il futuro, il sequenziamento massivo di singole molecole

Published online before print January 23, 2008, 10.1073/pnas.0710982105
PNAS | **January 29, 2008** | vol. 105 | no. 4 | **1176-1181**

◂ Previous Article | Table of Contents | Next Article ▸

BIOLOGICAL SCIENCES / BIOPHYSICS

Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures

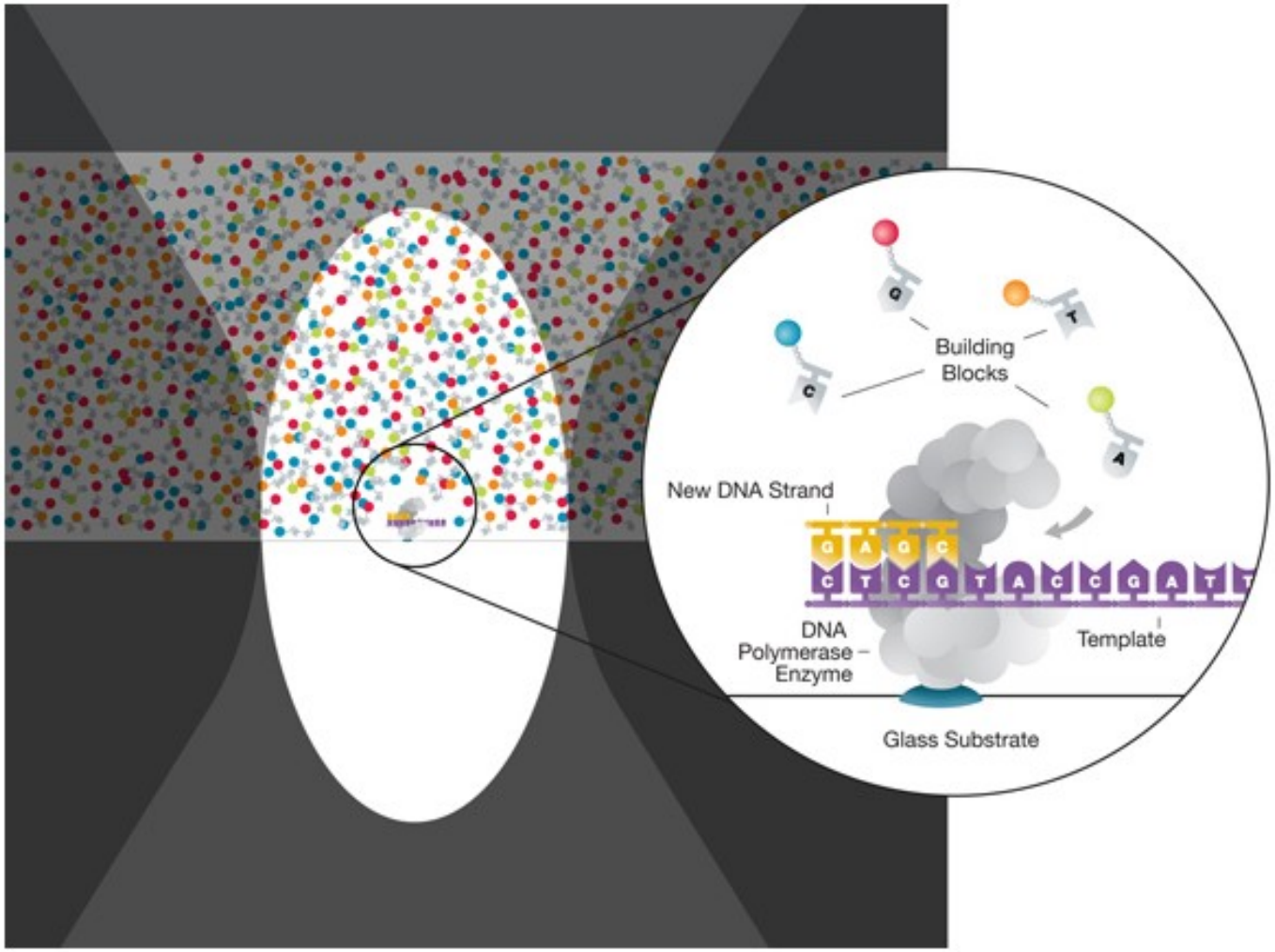
Jonas Korlach, Patrick J. Marks, Ronald L. Cicero, Jeremy J. Gray, Devon L. Murphy, Daniel B. Roitman, Thang T. Pham, Geoff A. Otto, Mathieu Foquet, and Stephen W. Turner*

Pacific Biosciences, 1505 Adams Drive, Menlo Park, CA 94025

Communicated by Watt W. Webb, Cornell University, Ithaca, NY, November 20, 2007 (received for review August 6, 2007)

Optical nanostructures have enabled the creation of subdiffraction detection volumes for single-molecule fluorescence microscopy. Their applicability is extended by the ability to place molecules in the confined observation volume without interfering with their biological function. Here, we demonstrate that processive DNA synthesis thousands of bases in length was carried out by individual DNA polymerase molecules immobilized in the observation volumes of zero-mode waveguides (ZMWs) in high-density arrays.

Pacific Biosciences ZMW



ZMW=‘Zero Mode Waveguide’

SMRT SEQUENCING

- **SMRT=Single Molecule Real-Time**
- **una ZMW è un foro, di diametro di qualche decina di nanometri, fabbricato in un film di metallo di 100nm depositato su un substrato di diossido di silicone**
- **Ciascun ZMW diventa una camera di visualizzazione nanofotonica che fornisce un volume di rilevazione di solo 20 zeptolitri (10^{-21} litri).**
- **In un tale volume, l'attività di una singola molecola può essere rilevata in un background di migliaia di nucleotidi marcati in fluorescenza**

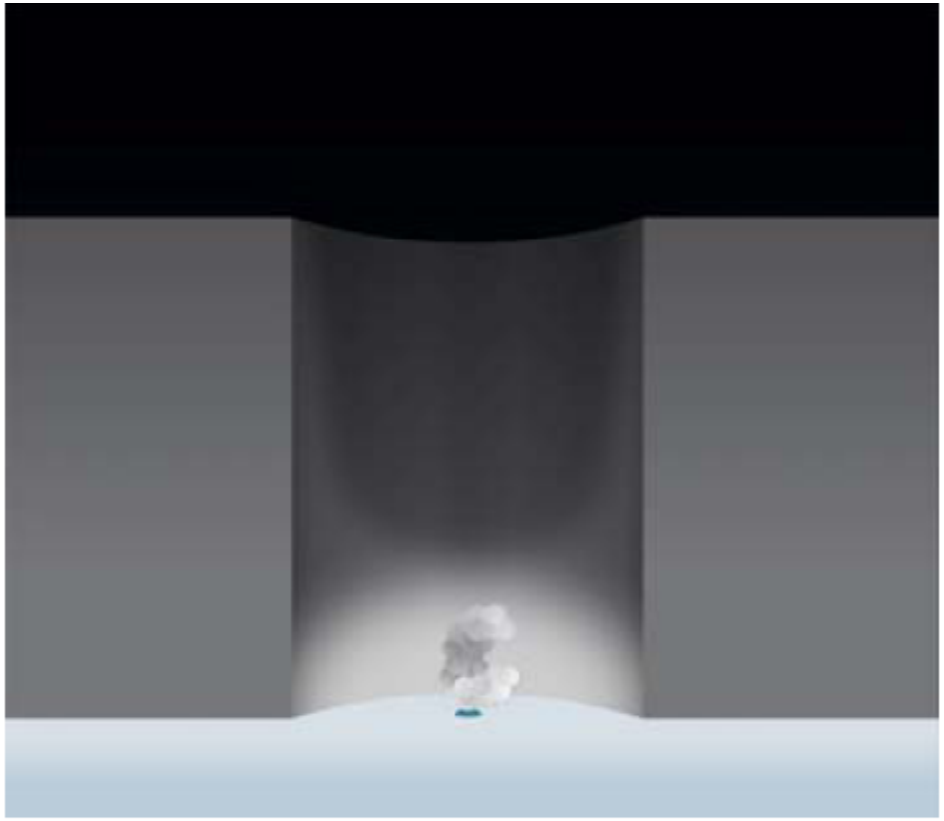
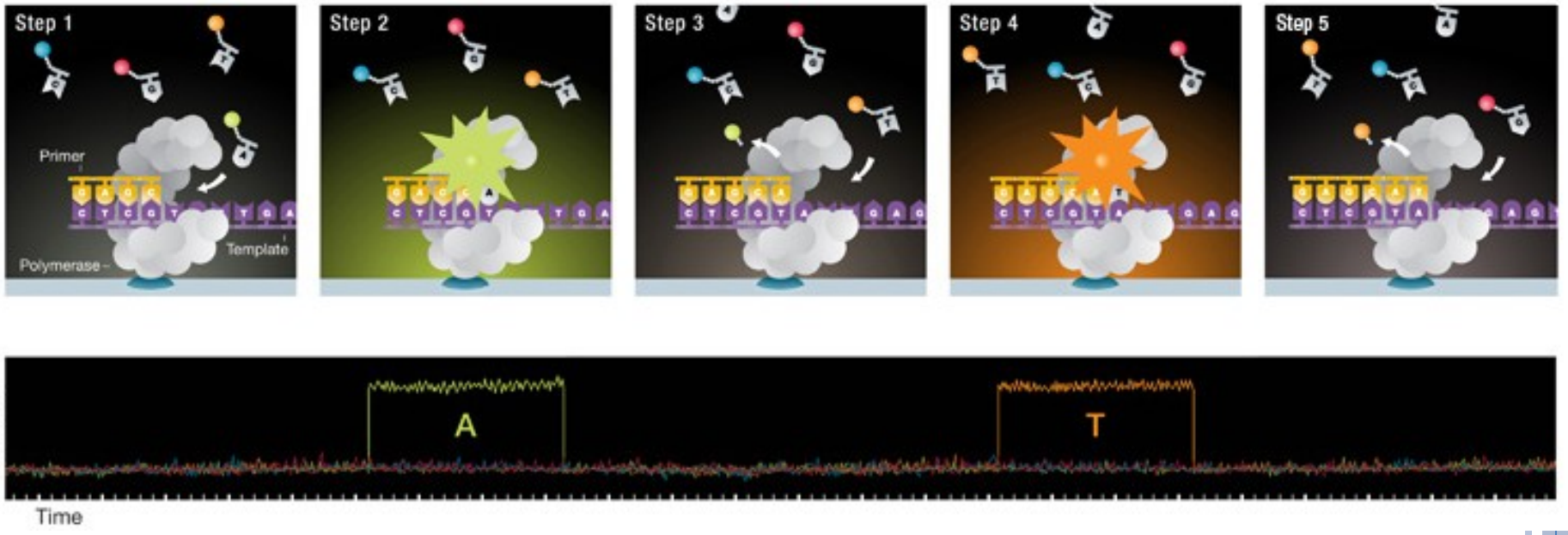


Figure 5. ***ZMW with DNA polymerase***

A single DNA polymerase molecule is attached to the bottom of the ZMW using a proprietary biased immobilization process.

Pacific Biosciences



- ✓ **Sequenziamento di singole molecole**
- ✓ **Incorporazione di molecole fluorescenti (sequenziamento mediante sintesi)**
  - ✓ **Monitoraggio in tempo reale dell'attività della polimerasi**
    - **(eccitazione/rilevazione)**

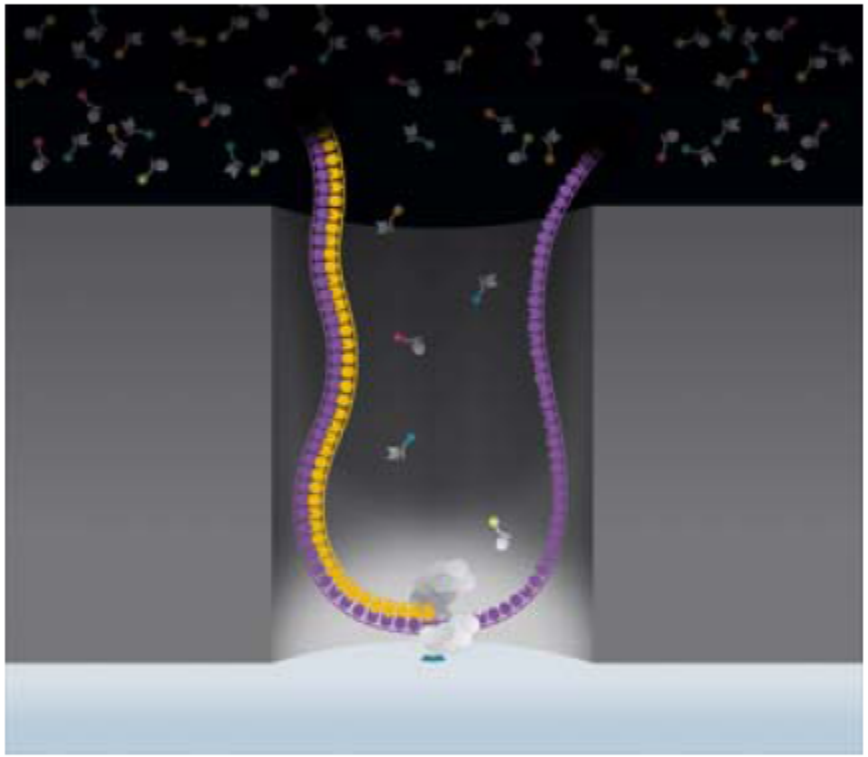


Figure 11. Synthesis of long DNA.

DNA polymerase processively incorporates nucleotides producing long, natural DNA.

OXFORD NANOPORES TECHNOLOGY

Long read length: > 50kb

High output: > 1gb/hr

"Run until…"

Cheap: ~$40/gb

Error rate: < 4%