

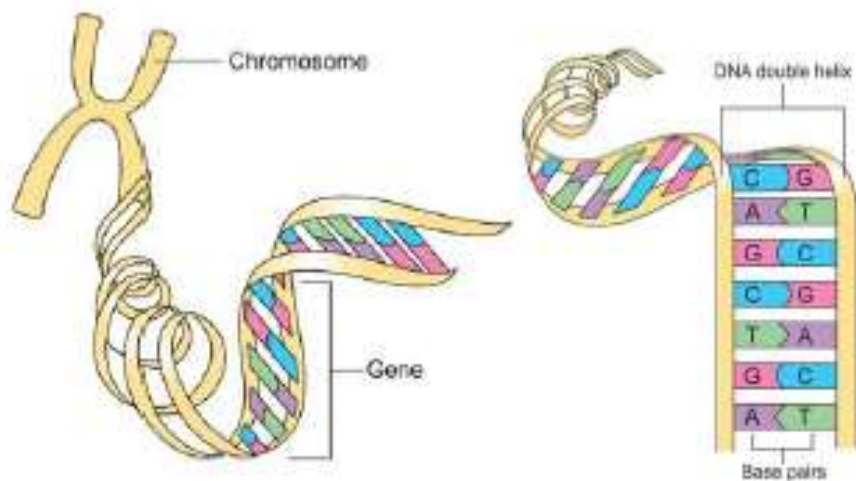
# Principi di sequenziamento e piattaforme NGS

ROMA, 26 OTTOBRE 2021

KATIA SPINELLA



# Cosa significa sequenziare il DNA?



DNA: acido desossiribonucleico

Unità di base del DNA sono i  
nucleotidi:

Adenina (A)  
Timina (T)  
Guanina (G)  
Citosina (C)

Nell'uomo ~ 20000 geni diversi

GCA AGA GAT AAT TGT

Sequenza nucleotidica (DNA)

Ala Arg Asp Asn Cys

Sequenza amminoacidica (proteina)

GCA AGA GAT AAT TGT

Sequenziamento del DNA

GCA AGA GAT AAT TGT  
CGT GCA AGA GAT AAT TGT  
\*\*\* GCA AGA GAT AAT TGT  
\*\*\* GCA AGA GAT AAT TGT



# Sequenziamento dei genomi: obiettivi della genomica

- ☐ Identificazione di tutti i geni e delle altre sequenze significative
- ☐ Costruire mappe genetiche e fisiche
- ☐ Confronto tra genomi di specie diverse (evoluzione)
- ☐ Produzione di un database per l'accesso alle informazioni



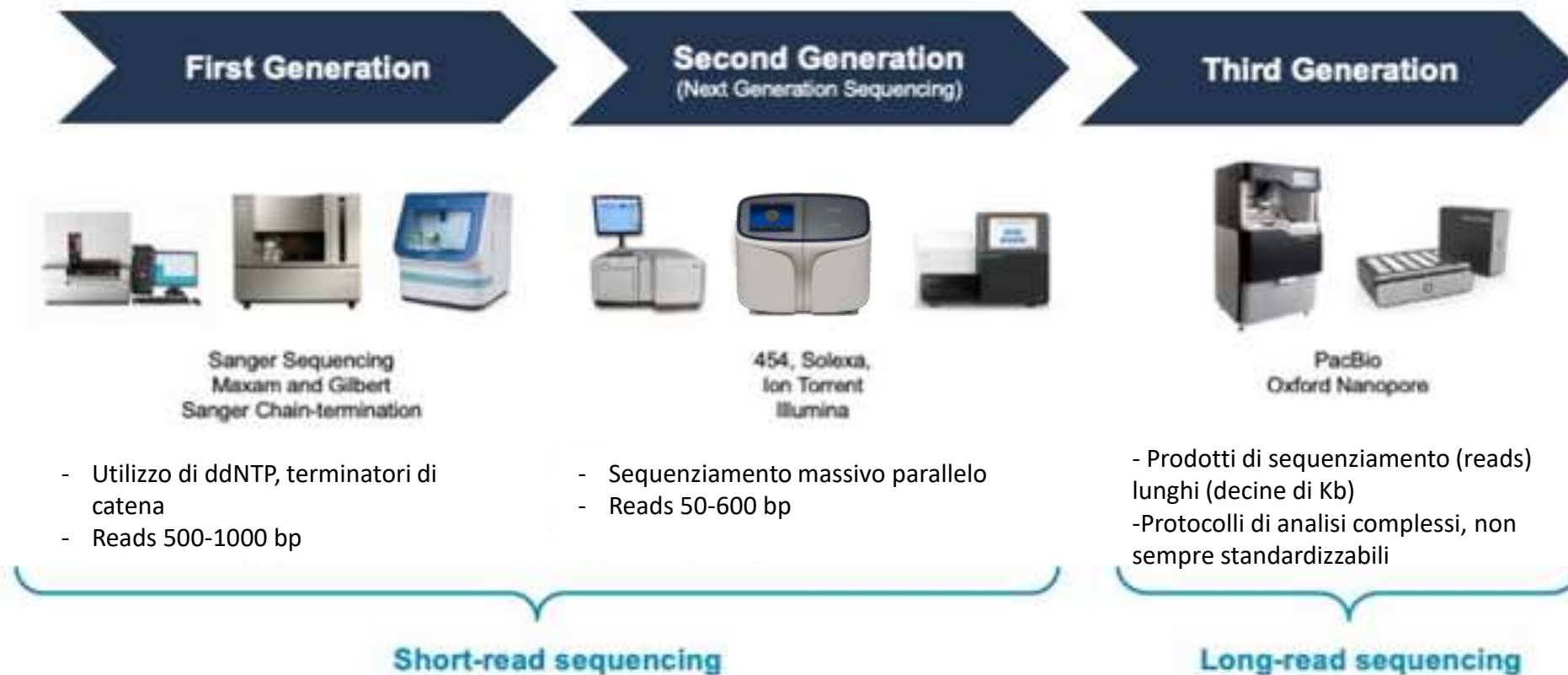


# Tecnologie di sequenziamento

- ❖ Il sequenziamento automatico secondo il metodo Sanger ha dominato nella scienza e nell'industria per almeno 20 anni e ha consentito il sequenziamento del genoma umano e la scoperta di almeno 2.500 malattie genetiche
- ❖ Il sequenziamento automatico secondo il metodo Sanger è considerato la tecnologia di “prima generazione”, mentre i nuovi metodi di “deep sequencing” sono denominati “next-generation sequencing (NGS)”
- ❖ **NGS** una serie di metodi che permettono di analizzare sequenze di DNA con strategie che permettono di ottenere milioni di sequenze in parallelo (sequenziamento massivo parallelo).



# Generazioni a confronto



# Piattaforme di 2° generazione

- ❖ Piattaforme di sequenziamento basate su diverse tecnologie
- ❖ Elevata accuratezza del dato di sequenza
- ❖ Standardizzazione protocolli di sequenziamento e analisi
- ❖ Prodotti di sequenziamento (**reads**) corti (max 600 bp) (Short-read sequencing )
- ❖ Diversi modelli di strumento in grado di generare **output** diversi (in Gb)



# NGS

## Wet-lab

1. preparazione della library
2. sequenziamento e imaging

## Dry-lab

3. analisi bioinformatica dei dati





# FORMAZIONE:

*Tecnici di laboratorio* per la **wet part**: estrazione del DNA, preparazione del campione e sequenziamento

✓ Esperienza pregressa (biologia molecolare)

Formazione su:

- Basi e principi di sequenziamento
- Flussi di laboratorio e protocolli
- Identificazione e risoluzione dei problemi

*Bioinformatici* per la **dry part**: analisi post-sequenziamento e gestione degli stessi dati

✓ Esperienza pregressa (biologia molecolare e/o bioinformatica)

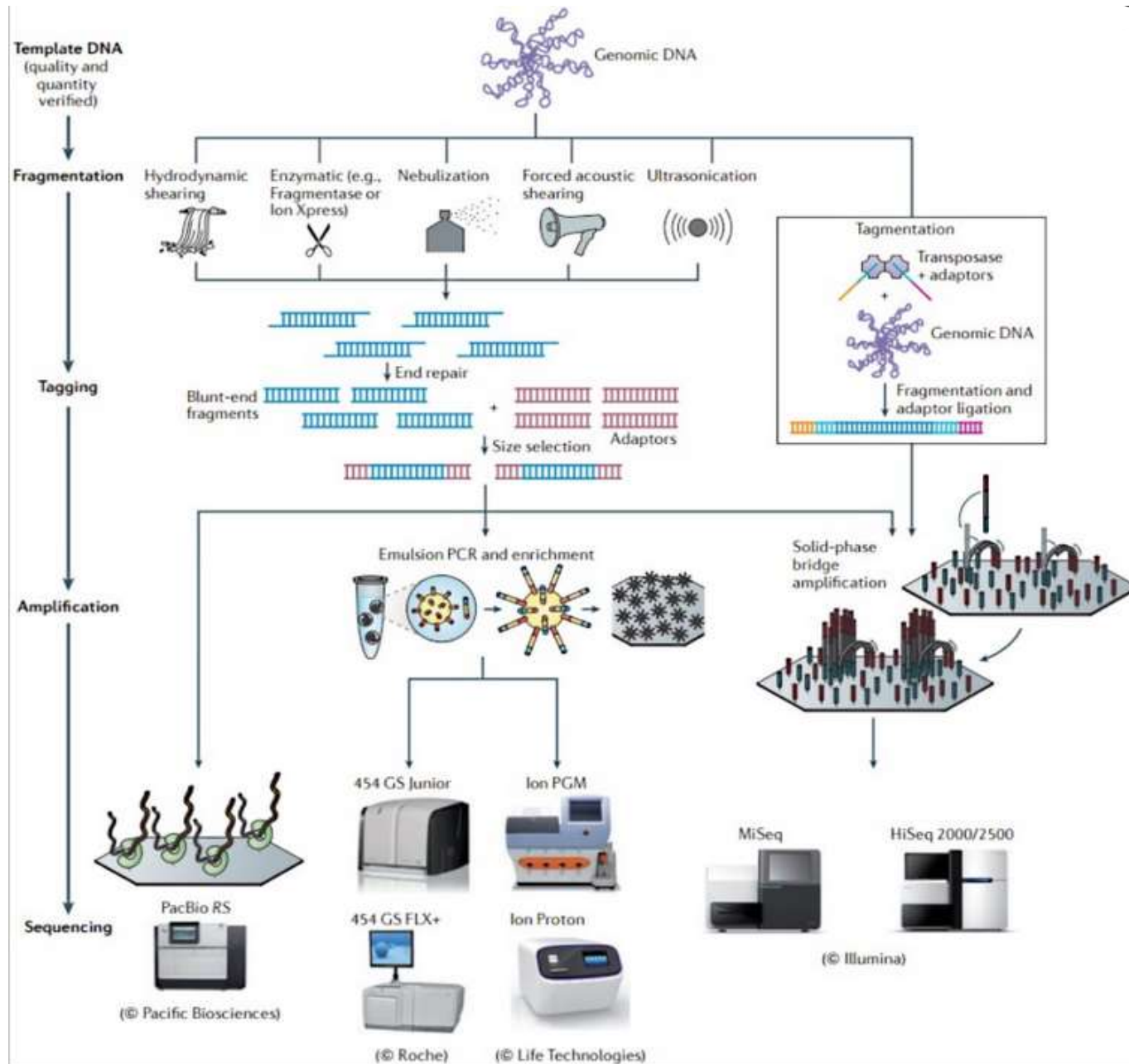
Formazione su:

- Sistemi operative UNIX e Linux
- Utilizzo di programmi ad interfaccia grafica (Galaxy, Geneious)
- Utilizzo dei comandi in riga e analisi di pipeline; uso di software commerciali o disponibili online
- Archiviazione dei dati di sequenza
- Identificazione e risoluzione dei problemi





# NGS WORKFLOW



## 1. Preparazione della library;

- Frammentazione (fisica o enzimatica)
- Le estremità vengono riparate
- Alle estremità vengono legati degli adattatori

## 2. Amplificazione dei frammenti;

- Sfere in emulsione
- Bridge amplification
- Amplificazione avviene sulla flow cell

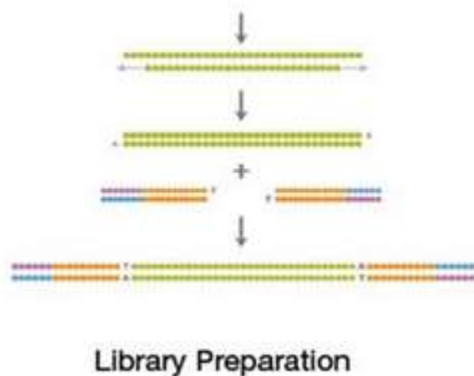
## 3. Sequenziamento;

- Sequenziamento per sintesi
- Sequenziamento per ligazione

# Illumina (tecnologia sequencing by synthesis (SBS))

Il flusso di lavoro NGS su piattaforma Illumina prevede 4 passaggi fondamentali:

A. Preparazione della  
library



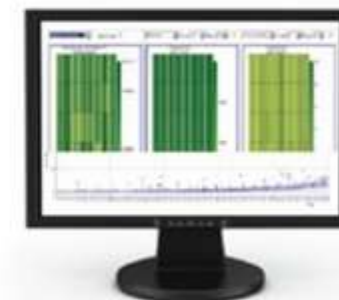
B. Generazione dei  
cluster



C. Sequenziamento per  
sintesi



D. Analisi dei dati  
(dry-lab)



Laboratorio è bello



# A. Preparazione della library

**LIBRARY:** Miscela di frammenti di DNA da sequenziare modificati per essere compatibili con la piattaforma di sequenziamento e con la strategia da applicare.



## Per la generazione di cluster

Le librerie devono avere le regioni di legame P5 e P7, che interagiscono con gli oligonucleotidi sulla superficie della cella a flusso.

## Per il multiplex

Le librerie devono avere un indice univoco o una sequenza marcata per miscelare i campioni.

## Per il sequenziamento

Le librerie devono avere regioni di legame per far sì che i primer di sequenziamento attivino Read 1 (Lettura 1) e Read 2 (Lettura 2).





## B. Generazione dei cluster

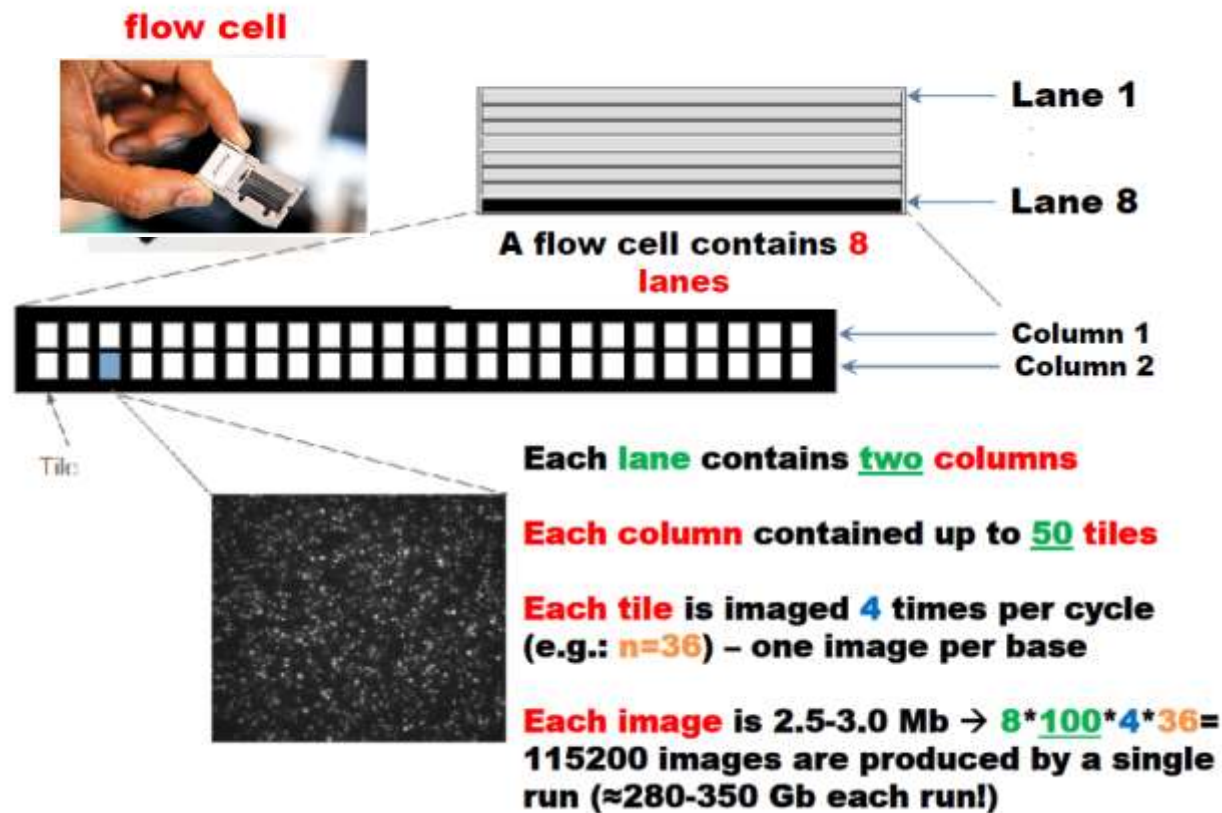
La generazione dei cluster avviene sulla flow cell

### FLOW CELL

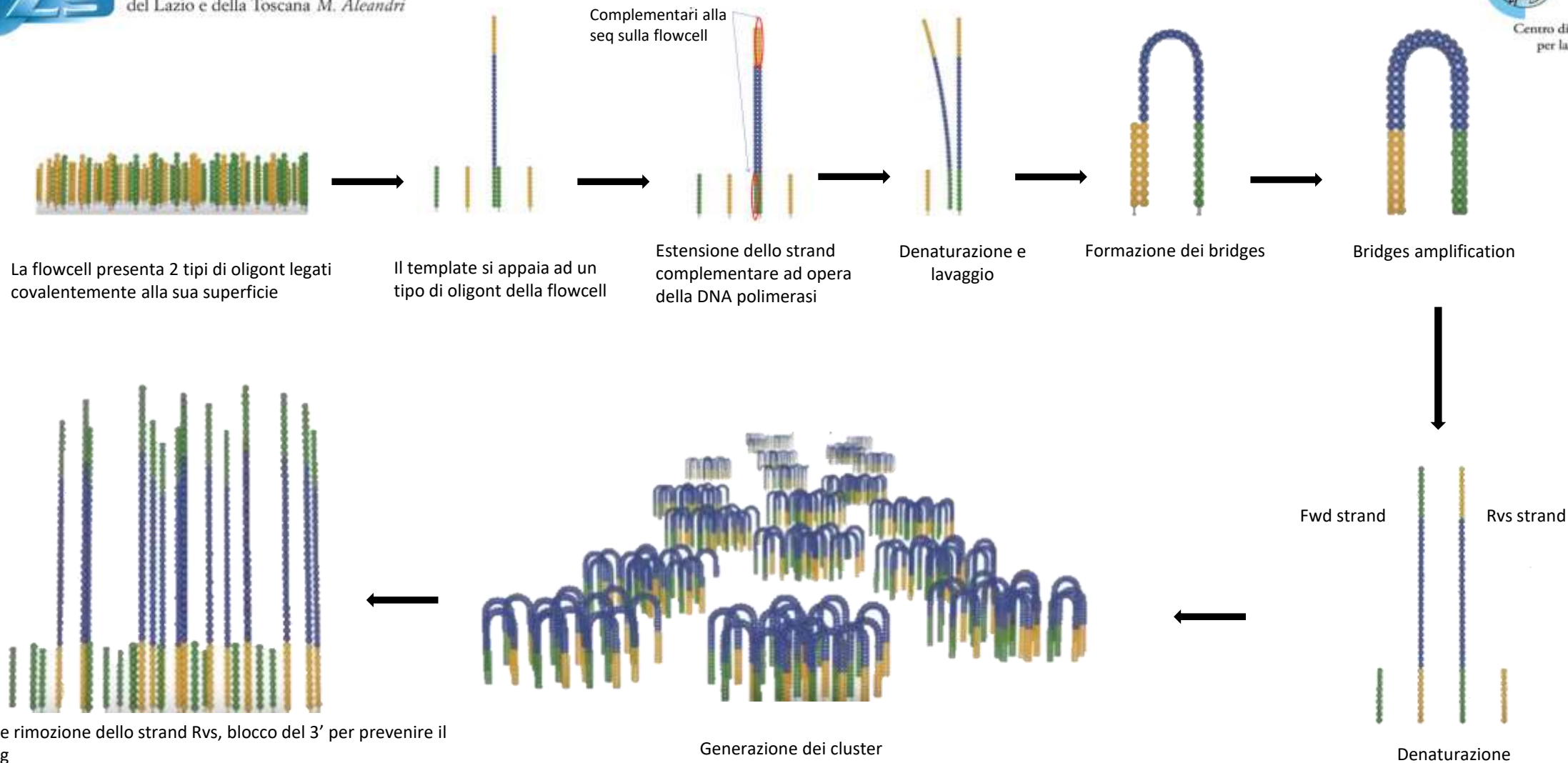
Supporto solido su cui viene immobilizzato il DNA da sequenziare (template). I reagenti liquidi possono fluire all'interno dei nano-pozzetti ed essere ciclicamente rimossi tramite lavaggi.

Una flow-cell è un vetrino suddiviso in corsie composte da milioni di nano-celle dove avviene fisicamente la reazione di sequenziamento.

Ogni strumento di sequenziamento utilizza una determinata flow cell.



## B. Generazione dei cluster

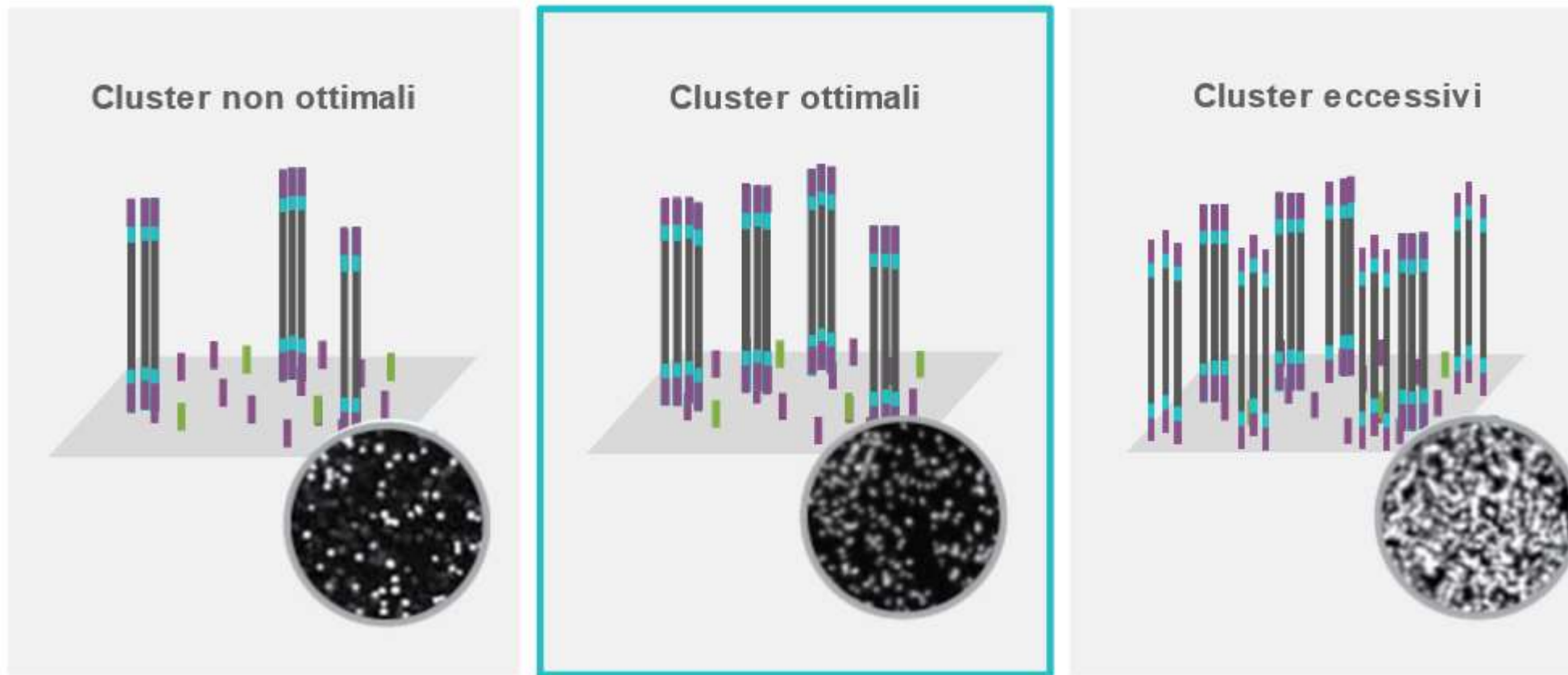


Ciascun frammento della libreria viene clonato in migliaia di copie identiche



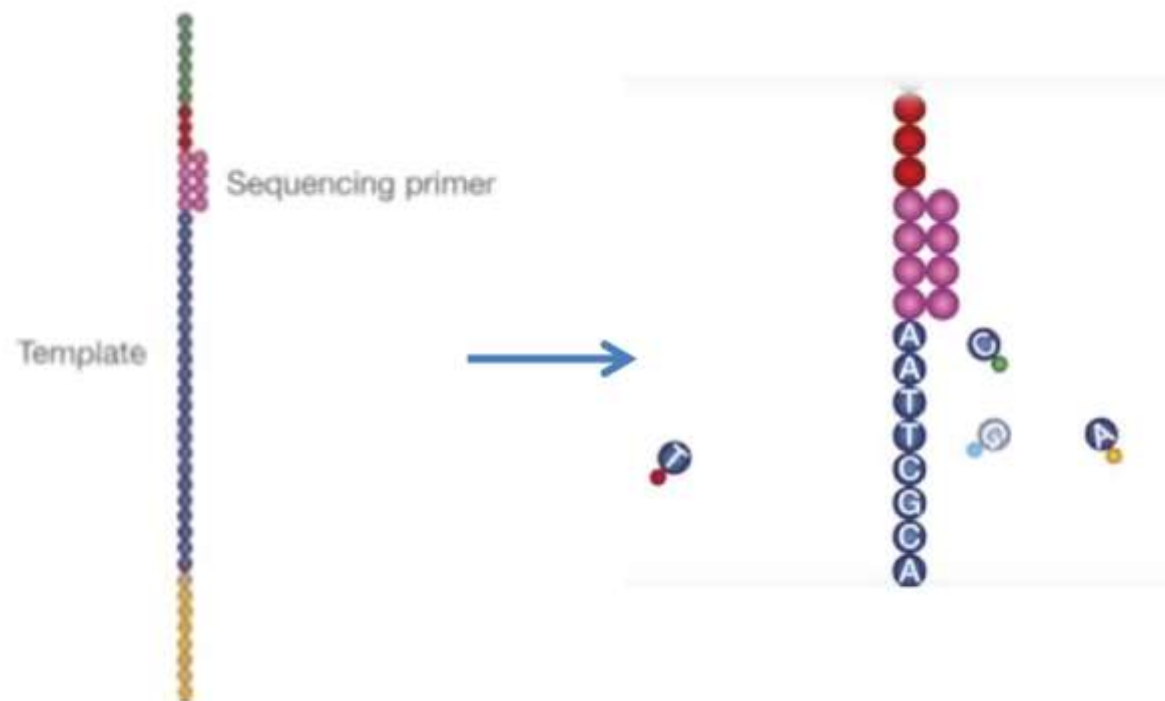
# Cluster density

La generazione di cluster sulla flow cell determina la qualità e la resa dei dati. E' importante quantificare la library prima del sequenziamento.





## C. Sequenziamento mediante sintesi (SBS)



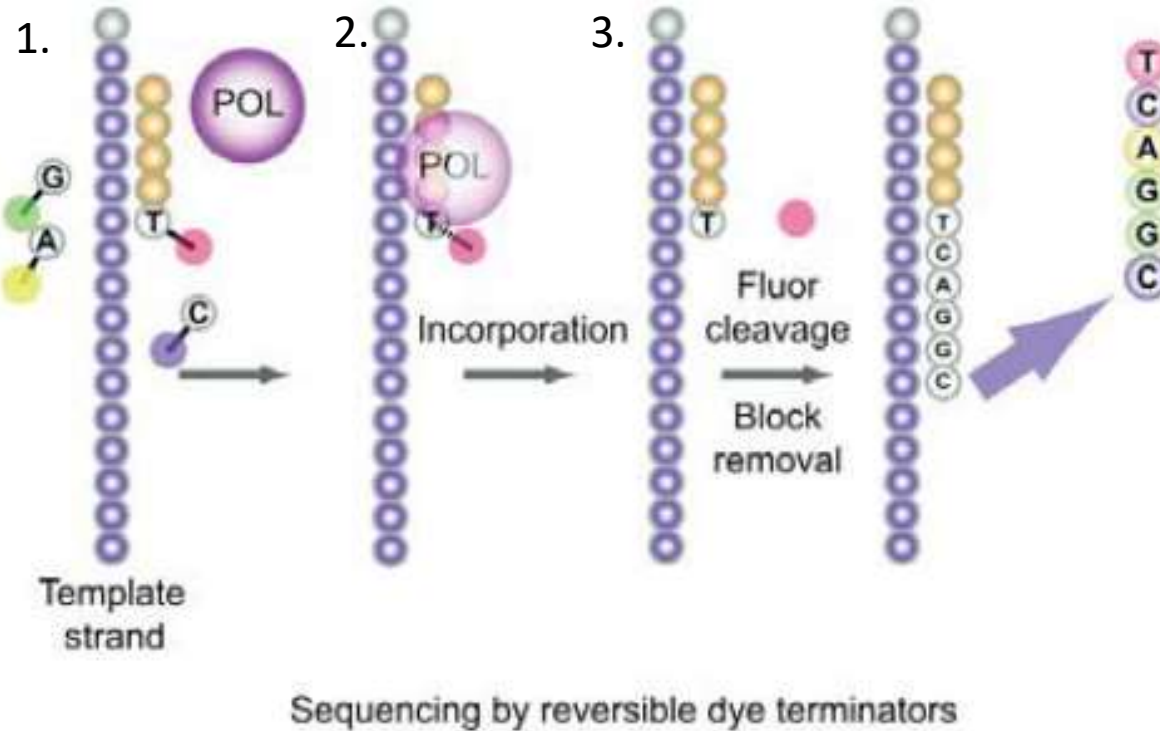
I primer di sequenziamento sono aggiunti per  
sequenziare la prima read

Sequenziamento per sintesi  
(ogni ciclo legge una base per ogni cluster)



# C. Sequenziamento mediante sintesi (SBS)

Un metodo basato su terminatori reversibili che consente il sequenziamento massivo in parallelo di miliardi di frammenti di DNA, rilevando singole basi mentre vengono incorporate in filamenti di DNA in crescita.



## 1. Incorporazione del nucleotide

Dei nucleotidi marcati con fluorofori legano la base complementare sul template. Si utilizzano 4 fluorofori differenti.

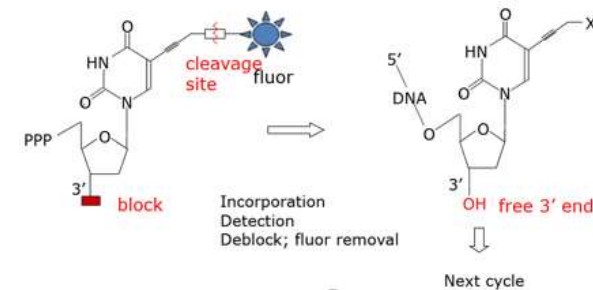
Il fluoroforo blocca il sequenziamento. Ogni cluster può incorporare una base diversa.

## 2. Acquisizione della fluorescenza

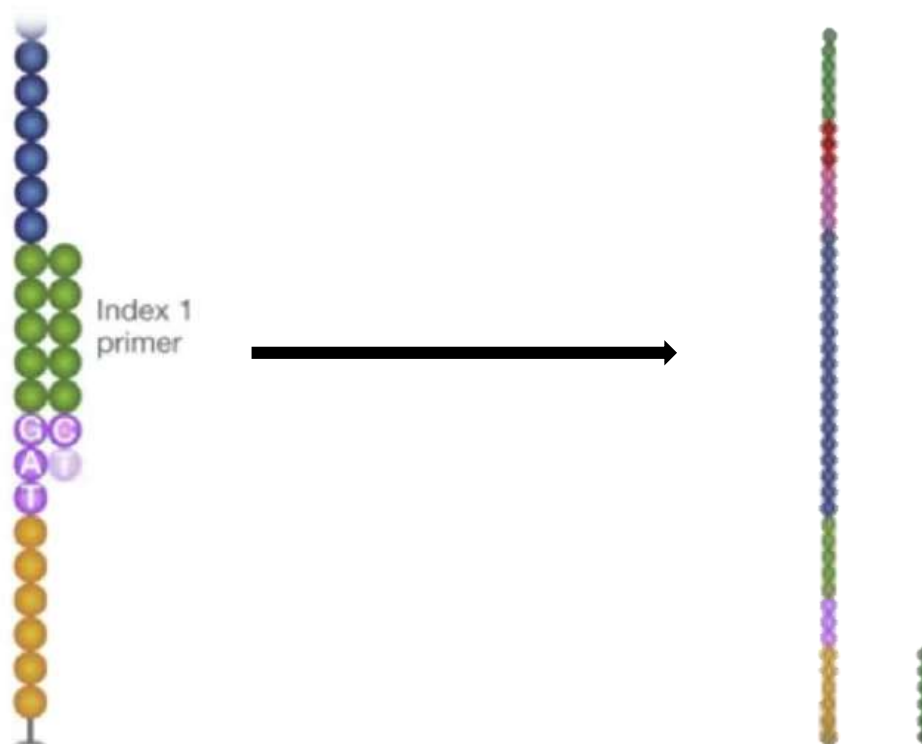
Un laser colpisce il cluster. Ogni cluster emette un colore corrispondente al nucleotide incorporato durante questo ciclo.

## 3. Taglio (cleavage)

I fluorofori vengono tagliati e lavati via dalla flow cell rigenerando il 3'OH. Inizia un nuovo ciclo con l'aggiunta di nuovi nucleotidi marcati.



## C. Sequenziamento mediante sintesi (SBS)



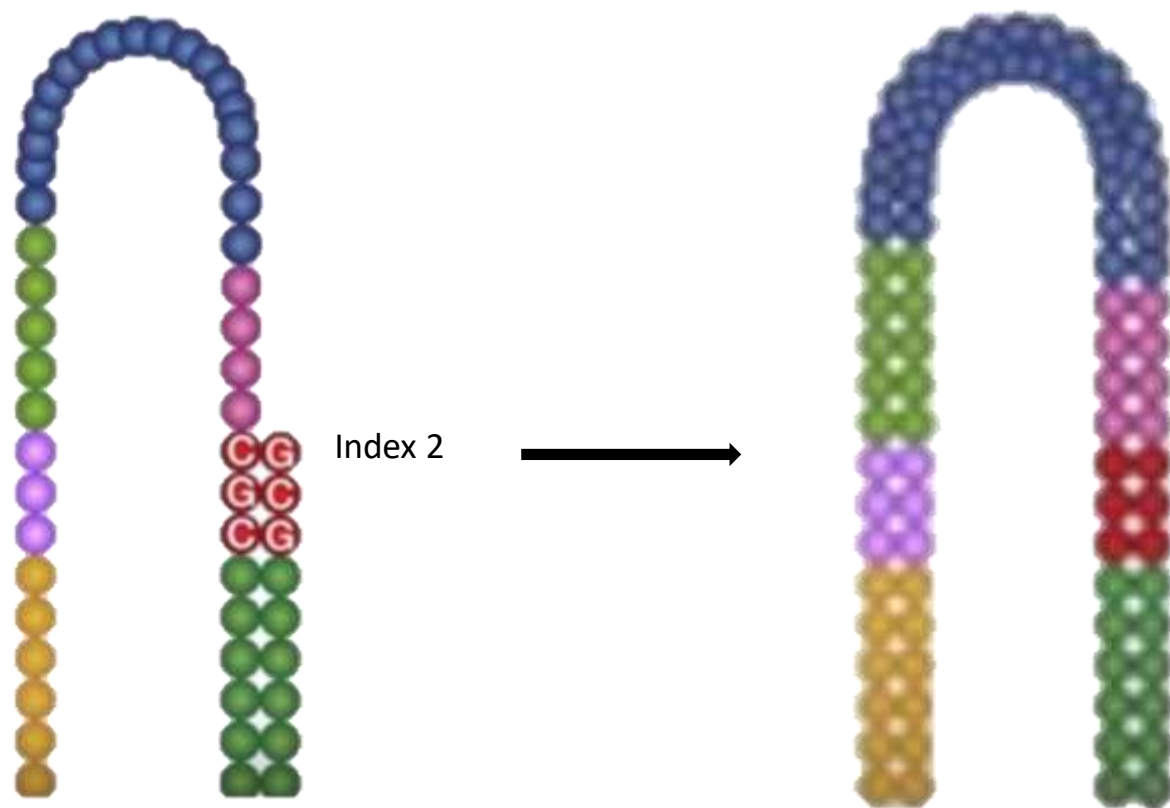
Il primer per l'index è aggiunto  
alla reazione e l'index viene letto

Rimozione del prodotto e  
deprotezione del 3' del template





## C. Sequenziamento mediante sintesi (SBS)

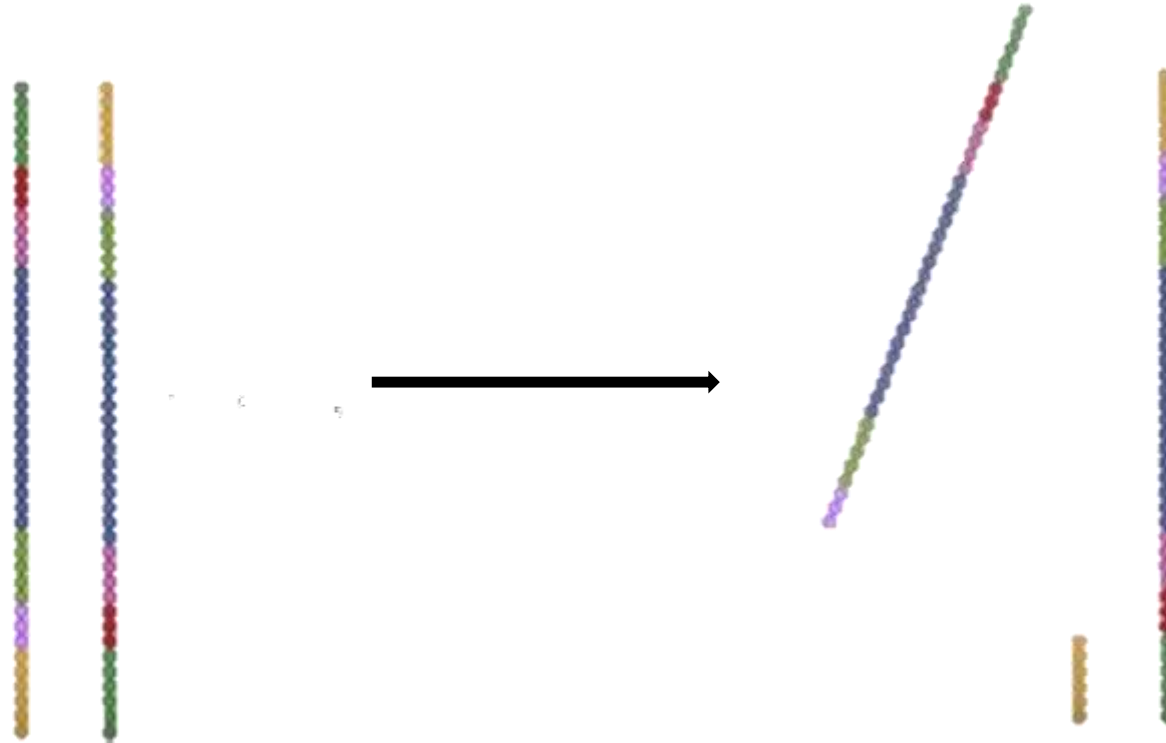


Il template si appaia al 2 oligont della  
flowcell e l'index 2 è letto

Il prodotto di index è lavato via e la  
polimerasi forma il double strand bridges



## C. Sequenziamento mediante sintesi (SBS)

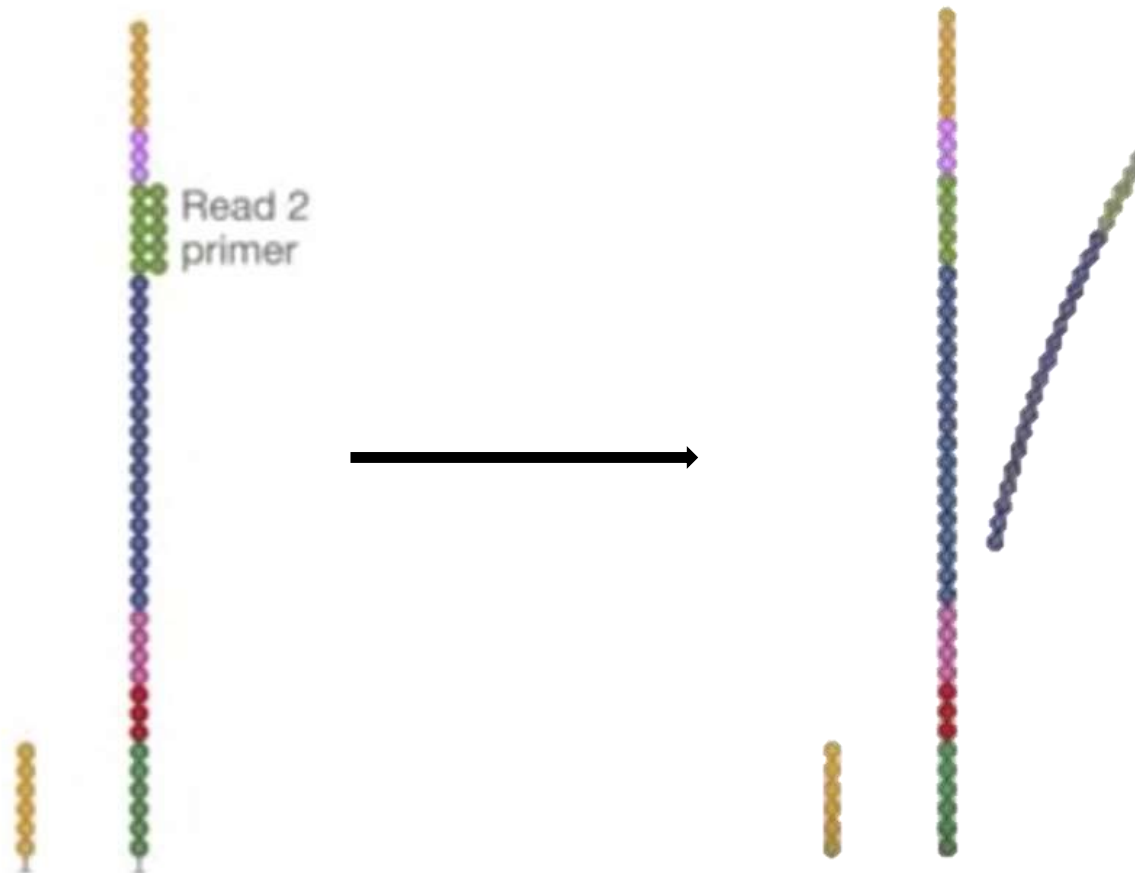


Linearizzazione del double strand bridges

L'estremità 3' è bloccata e il Fwd è tagliato e rimosso



## C. Sequenziamento mediante sintesi (SBS)



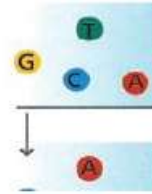
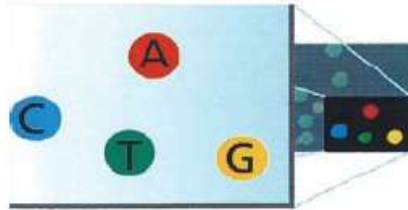
Il primer 2 è aggiunto e il  
frammento è sequenziato

Rimozione della read



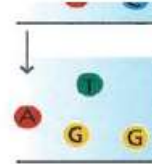


# Data collection



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

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



GCTGA...

Each cycle produces 4 images

Sequence of n bases of each fragment is  
read by inspecting signals at each  
location of the flowcell



	Reagent Kit v2			Reagent Kit v3	
Lunghezza delle reads	2 x 25 bp	2 x 150 bp	2 x 250 bp	2 x 75 bp	2 x 300 bp
kit size (cicli)	50	300	500	150	600
Tempo per il sequenziamento	5.5 ore	24 ore	39 ore	21 ore	56 ore
Output	850 Mb	4.5 Gb	7.5 Gb	3.3 Gb	13.2 Gb
n. di reads		15 M		25 M	
	Reagent Kit v2 Micro			Reagent Kit v2 Nano	
Lunghezza delle reads	2 x 150 bp			2 x 250 bp	2 x 150 bp
kit size (cicli)	300			500	300
Tempo per il sequenziamento	19			28 ore	17 ore
Output	1.2 Gb			500 Mb	300 Mb
n. di reads	4 M			1 M	



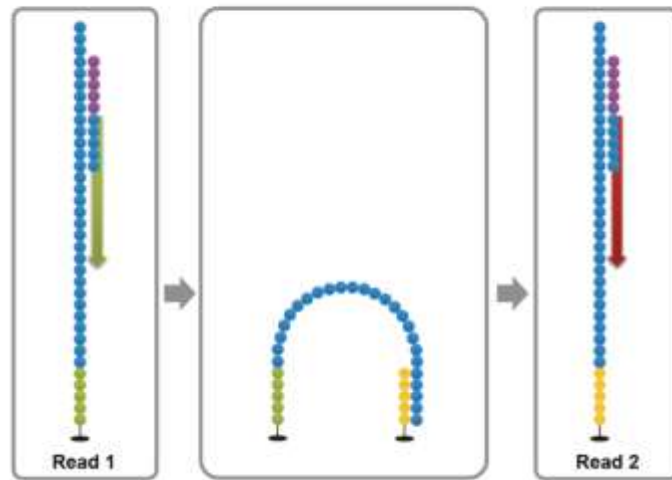
- Flow cell ; Standard, Micro o Nano
- Numero di cicli
- Output
- Coverage



## Come calcolare la lunghezza delle reads?

Tutti i reagenti di sequenziamento Illumina presentano un certo numero di cicli di sequenziamento. Questi cicli sono direttamente correlati alla lunghezza della sequenza. Poiché una base è sequenziata per ciclo, il numero totale di cicli indica il numero massimo di basi che possono essere sequenziate. È possibile utilizzare reagenti di sequenziamento per generare letture continue singole (single end) o per sequenziamento paired-end in entrambe le direzioni. (Ad esempio, un kit da 300 cicli può essere utilizzato per una corsa in lettura singola  $1 \times 300$  bp o una corsa in paired-end  $2 \times 150$  bp.)

### Illumina Paired-End Sequencing



### DNA Sequencing Applications

Application	Recommended Read Length
Whole-genome sequencing	$2 \times 150$ bp
Whole-exome sequencing	$2 \times 150$ bp
Targeted enrichment sequencing	$2 \times 150$ bp
Amplicon sequencing	Length of the entire amplicon insert
De novo sequencing	Ranges from $2 \times 150$ to $2 \times 300$ bp

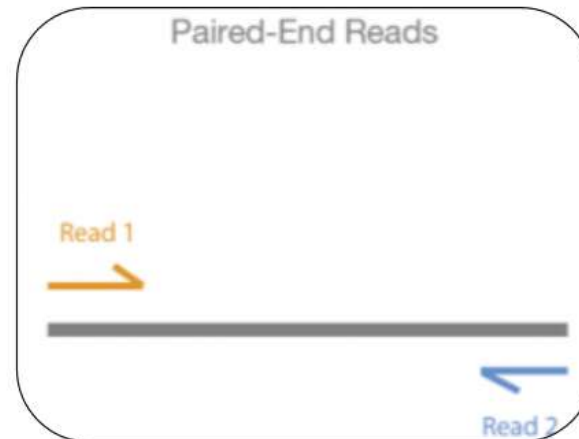
Il numero di cicli determina la lunghezza delle reads ed è uno dei parametri da considerare quando si seleziona il kit di sequenziamento (**KIT SIZE**)





- **Read** (lettura): si riferisci ad una stringa di dati che corrisponde ad una data sequenza
- **Numero di reads**: numero di letture effettuate per singolo campione, espresso in M (milioni) di reads
- **Single end** sequencing: sequenziamento di una sola estremità del DNA
- **Paired End** sequencing: vengono sequenziate entrambe le estremità del frammento di DNA, per migliorarne l'accuratezza e l'allineamento

Il sequenziamento PE permette all'algoritmo di mappare meglio le regioni ripetute.



L'output è il numero di basi totali lette dal sequenziatore, espresso in Mb (mega basi) o Gb (giga basi).

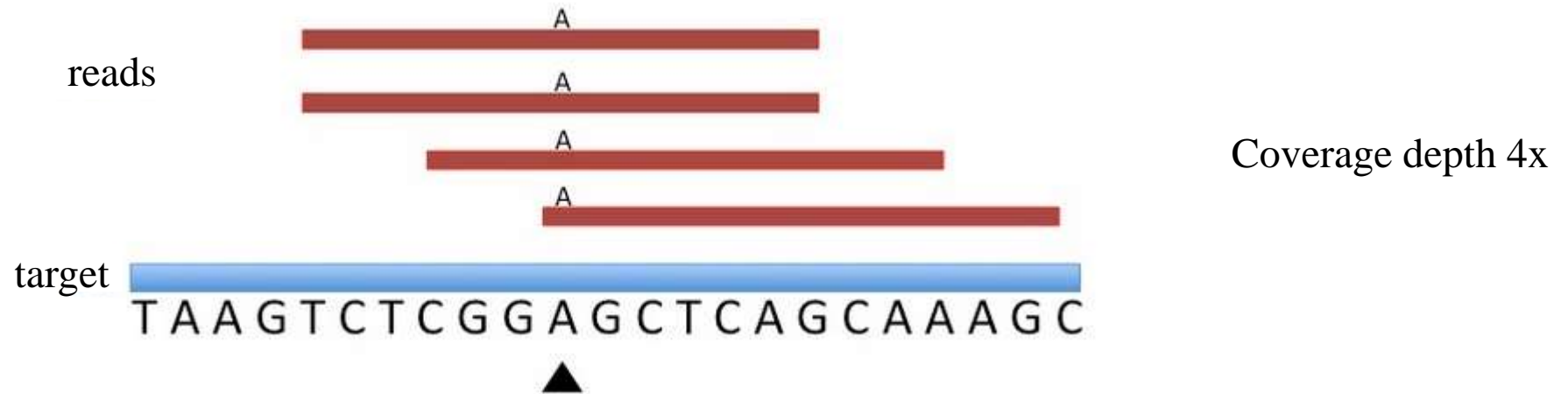
		Reagent Kit v2		Reagent Kit v3	
Lunghezza delle reads	2 x 25 bp	2 x 150 bp	2 x 250 bp	2 x 75 bp	2 x 300 bp
kit size (cicli)	50	300	500	150	600
Tempo per il sequenziamento	5.5 ore	24 ore	39 ore	21 ore	56 ore
Output	850 Mb	4.5 Gb	7.5 Gb	3.3 Gb	13.2 Gb
n. di reads		15 M		25 M	
		Reagent Kit v2 Micro		Reagent Kit v2 Nano	
Lunghezza delle reads		2 x 150 bp		2 x 250 bp	2 x 150 bp
kit size (cicli)		300		500	300
Tempo per il sequenziamento		19		28 ore	17 ore
Output		1.2 Gb		500 Mb	300 Mb
n. di reads		4 M		1 M	

Es: Standard Flow Cell con kit da 300 cicli (reads paired- end lunghe 150) ha un output di 4,5Gb pari a 15 milioni di reads.  $150 \text{ bp} \times 2 \times 15.000.000 = 4,5 \text{ Gb}$



# Coverage depth (profondità di lettura)

Il numero medio di letture (reads) ottenute dal sequenziamento per ciascuna base nucleotidica costituente il target. Maggiore è la profondità di lettura più è accurato il dato di sequenza.





Il coverage è un parametro deciso dall' operatore in base; - all' applicazione e all'obiettivo dello studio

- dimensione del genoma da sequenziare o del target
- livello di espressione dei geni
- riferimenti in letteratura di studi simili

## Come scegliere il coverage?

- Un coverage di 20-30X permette di sequenziare in maniera poco profonda il genoma : si può scegliere questo coverage per un resequencing di genomi già noti e per la ricerca di SNP già annotati
- Un coverage di 50-100X permette di sequenziare in profondità i genomi ed analizzare tutte le varianti presenti,

$$\text{Coverage} = \frac{(\text{n. reads}) \times (\text{lunghezza delle reads})}{\text{dimensione del target}}$$



# Quanti campioni posso caricare in una corsa?

Dipende da:

- Dimensione del genoma
- Coverage richiesto
- Output del sequenziatore ( Gb)

$$\text{n. campioni per run} = \frac{\text{output del sequenziatore}}{(\text{coverage} \times \text{dimensione del genoma})}$$



Per un esperimento di targeted sequencing:

- Numero di ampliconi
- Coverage richiesto
- Output del sequenziatore ( # reads)

$$\text{n. campioni per run} = \frac{\text{output del sequenziatore (n. di reads)}}{(\text{coverage} \times \text{n.ampliconi})}$$





## Esempio

Voglio sequenziare *Bacillus Subtilis*

- Dimensione del genoma: circa 4 Mbp
- Coverage: 50X
- Output del sequenziatore: 13 Gb

$$\text{n. campioni per run} = \frac{\text{output del sequenziatore}}{(\text{coverage} \times \text{dimensione del genoma})}$$

# Quanti campioni posso caricare in una corsa?

65 campioni di *B. Subtilis*



## D. Analisi dei dati

L'analisi dei dati ottenuti dal sequenziamento, rappresenta un complesso “montaggio di dati” che avviene con l'ausilio di algoritmi bioinformatici ben strutturati e con l'impiego di software particolari. Questo passaggio è specifico a seconda dell'obiettivo da raggiungere.

### Analisi bioinformatica



### D. Alignment and Data Analysis

Reads

```

ATGGCATTGCAATTGACAT
TGGCATTGCAATTG
AGATGGTATTG
GATGGCATTGCAA
GCATTGCAATTGAC
ATGGCATTGCAATT
AGATGGCATTGCAATTG
  
```

Reference Genome

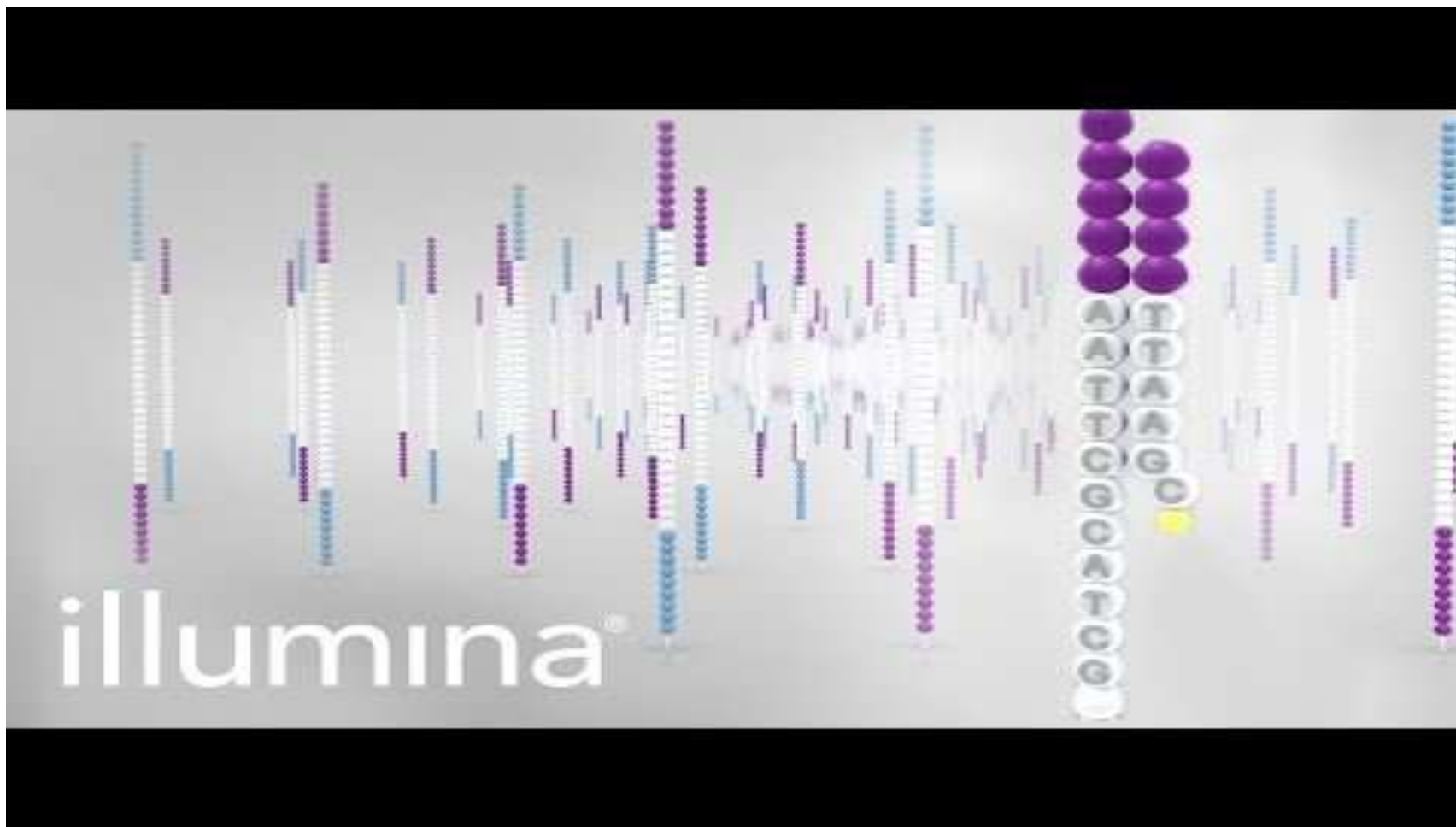
```

AGATGGTATTGCAATTGACAT
  
```

Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

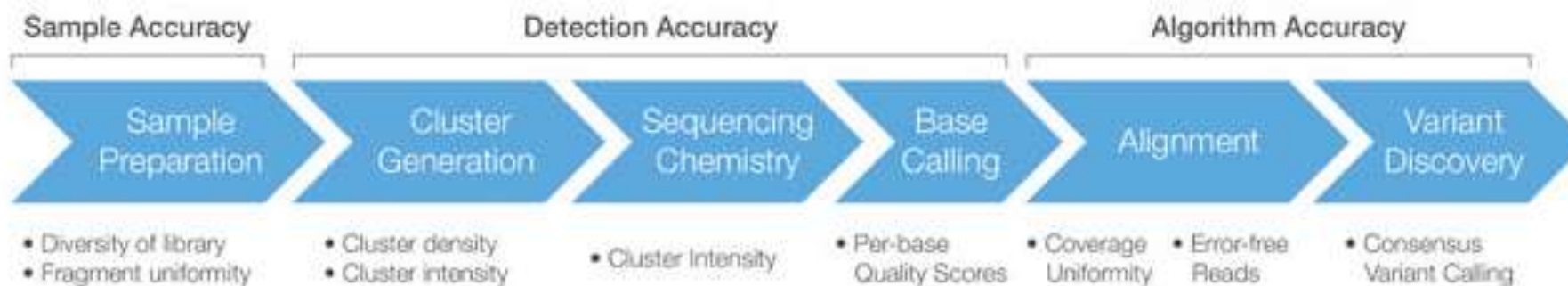


# Video; Tecnologia Illumina





# Aspetti critici che influenzano la qualità dei dati



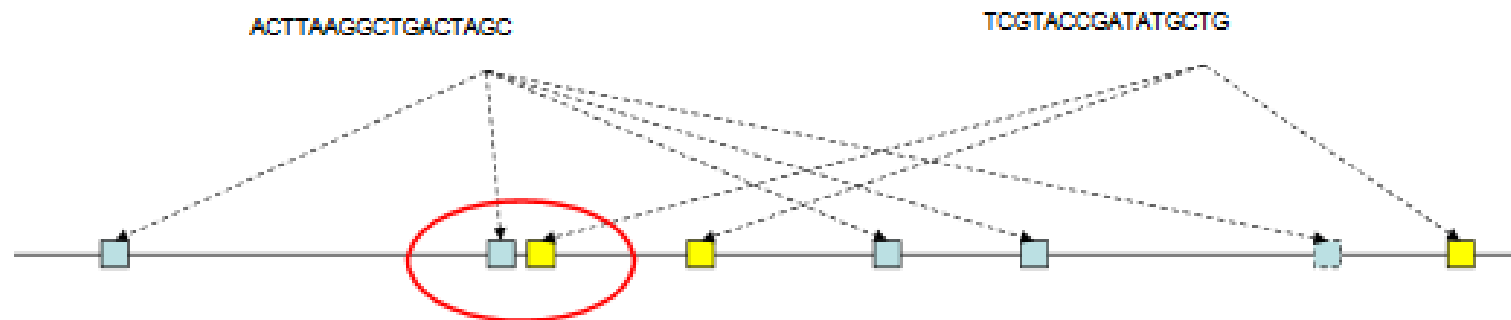
Factors that Contribute to Platform Accuracy



- Versatilità dello strumento
- Efficienza dello strumento;
  - Costo per corsa
  - Dati per corsa
- Risultati confrontabili perché largamente diffuso
- Sequenze pulite
- Sequenze corte



# Shorts reads



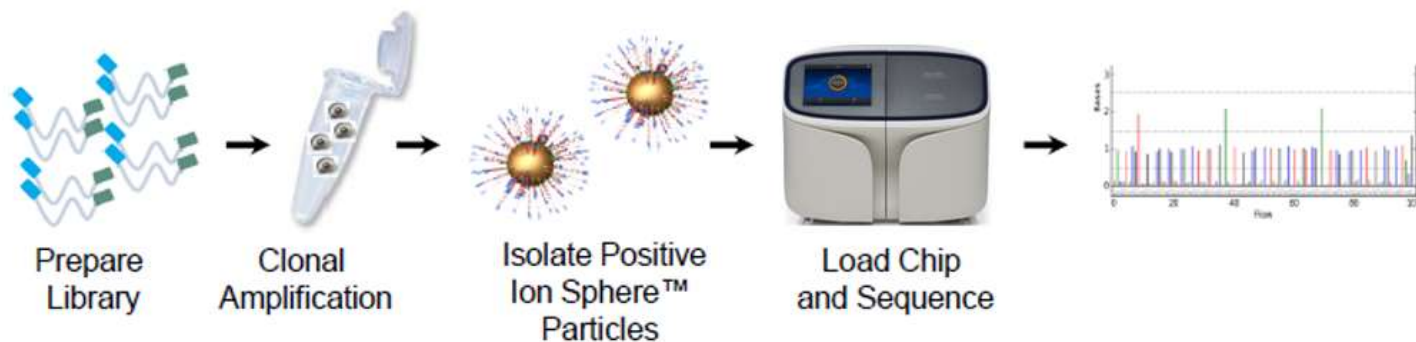
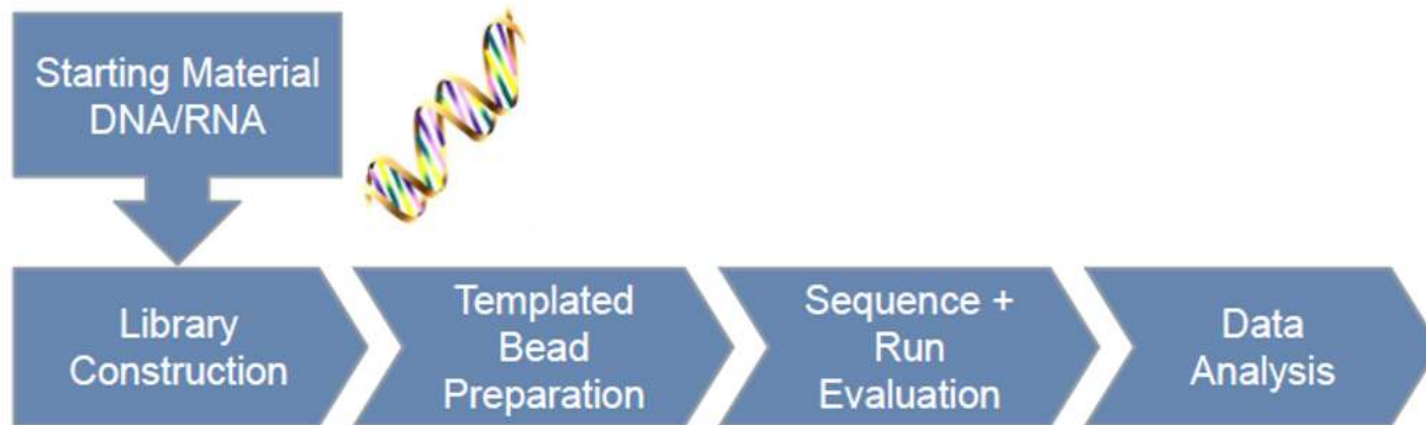
Le reads corte sono problematiche, perchè corte sequenze possono appaiare in vari punti del genoma

Soluzione #1: ottenere lunghe reads

Solution #2: sequenziare in entrambe le direzioni (paired-end)



# Ion Torrent

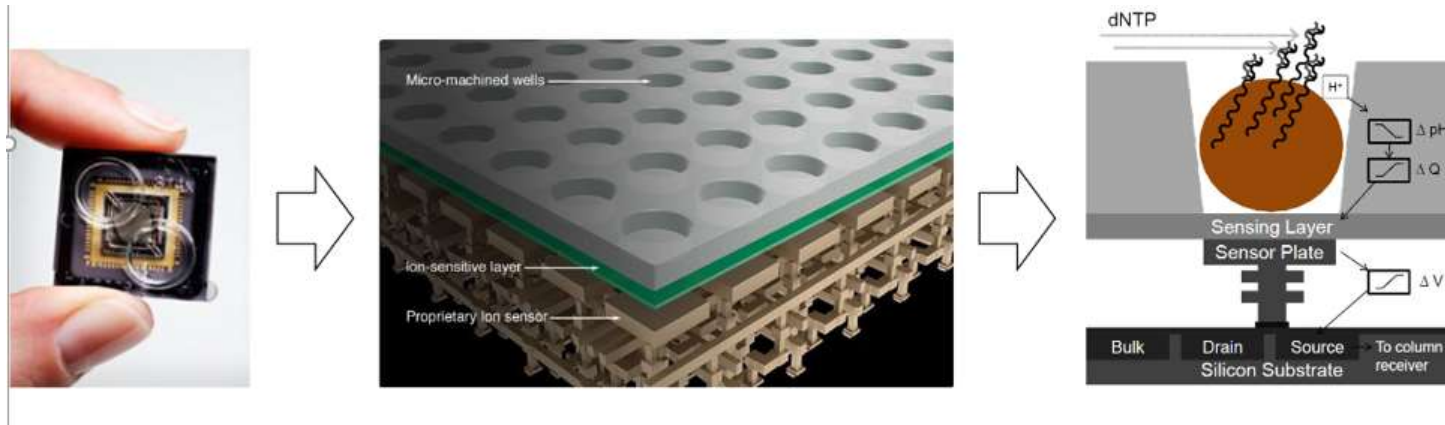




# Dove avviene la reazione di sequenziamento?

□ DNA → Ioni → Sequenza

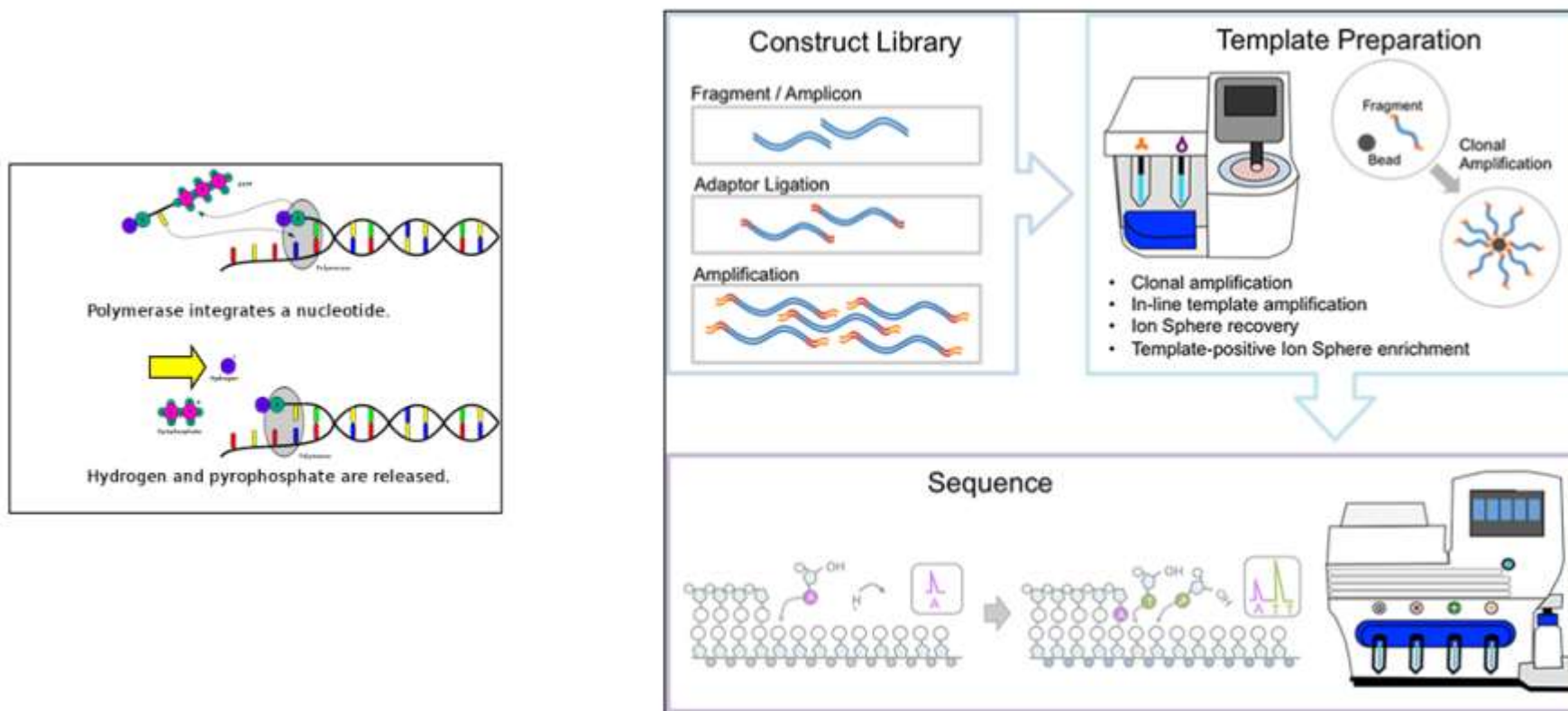
- IonTorrent utilizza una tecnologia basata su semiconduttori che consente di sequenziare il DNA senza utilizzare sistemi ottici di lettura.
- Flusso sequenziale dei nt sul chip (ogni 4 sec)
- Un sensore per well per reazione di sequenziamento
- Milioni di reazioni di sequenziamento per chip
- Analisi in real time



# Ion Torrent

L'idrolisi del nucleotide trifosfato causa la liberazione di un singolo protone per ogni nucleotide incorporato.

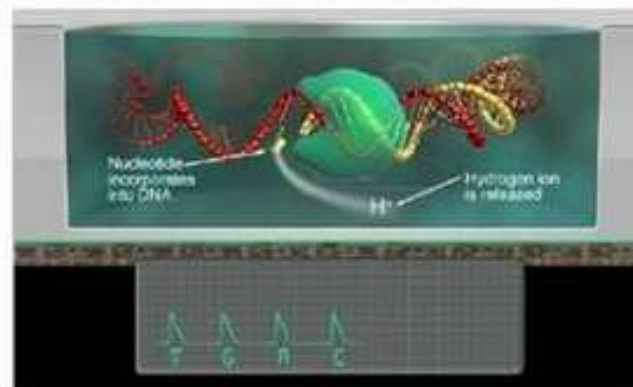
Una variazione di pH localizzata nella soluzione circostante causa una variazione di pH pari a 0.02 unità per nucleotide incorporato



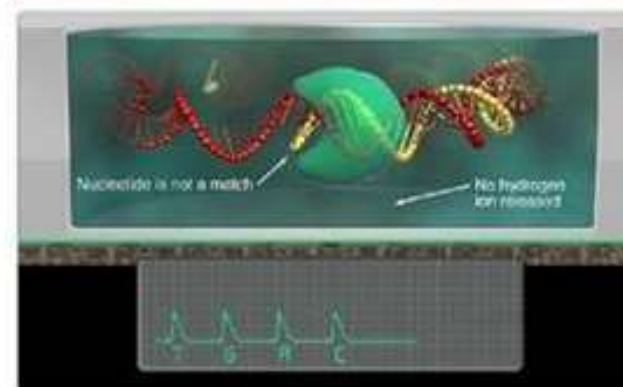
# quindi...

...sequenzialmente flusseranno in continuo i 4 deossinucleotidi...

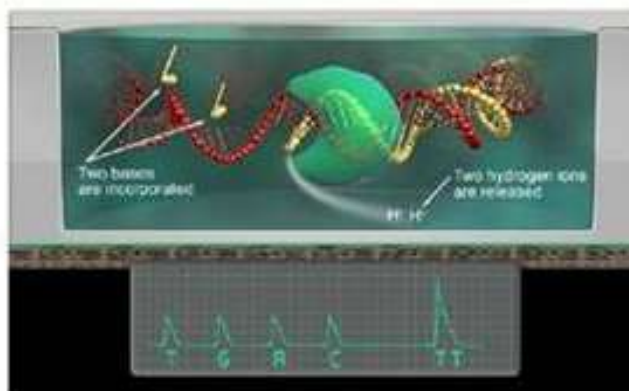
...e in seguito all'incorporazione verranno rilasciati Ioni idrogeno ( $H^+$ )



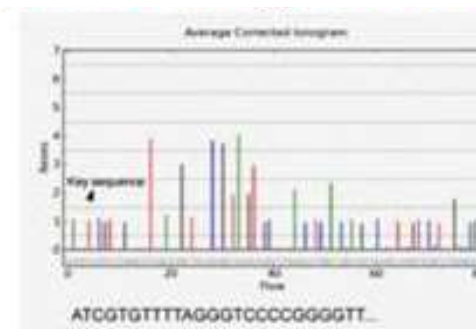
Non si avrà rilascio se non avrà luogo l'incorporazione!!



...in seguito all'incorporazione di due nucleotidi, si avrà il rilascio di due ioni  $H^+$

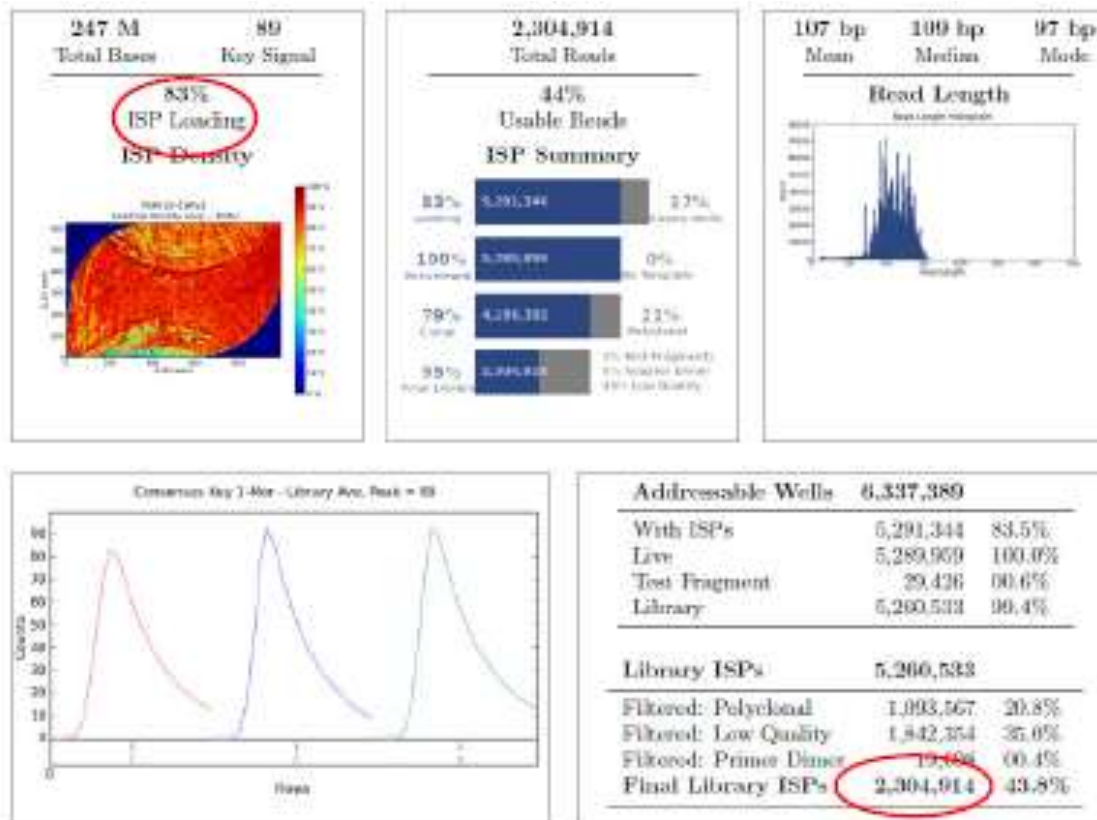


## Risultato...un ionogramma





# Report corsa ION TORRENT



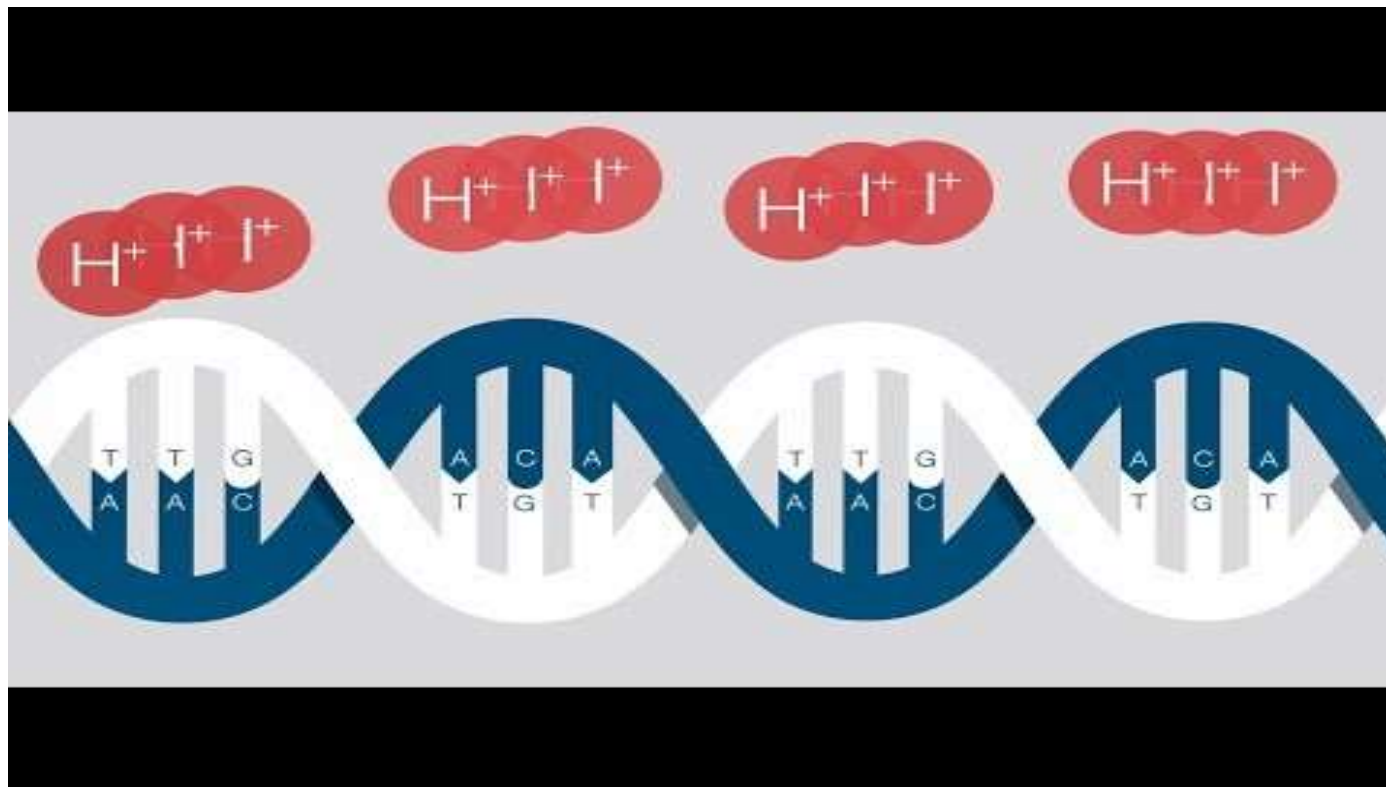
N° delle sfere con un frammento sequenziato

Barcode Name	Sample	Bases	≥ Q20	Reads	Mean Read Length
No barcode	None	1,210,465	1,094,732	10,932	110 bp
IonXpress_001	SI-23T	128,036,443	118,066,520	1,155,558	110 bp
IonXpress_002	SI-24T	118,727,478	110,212,069	1,138,347	104 bp





# Video; Tecnologia Ion Torrent



# Sequenziamento di 3° generazione

- ❖ Piattaforme di sequenziamento basate su diverse tecnologie dai costi altamente variabili
- ❖ Accuratezza del dato di sequenza generalmente più bassa
- ❖ Prodotti di sequenziamento (**reads**) lunghi (decine di Kb)
- ❖ Protocolli di analisi complessi, non sempre standardizzabili



PacBio

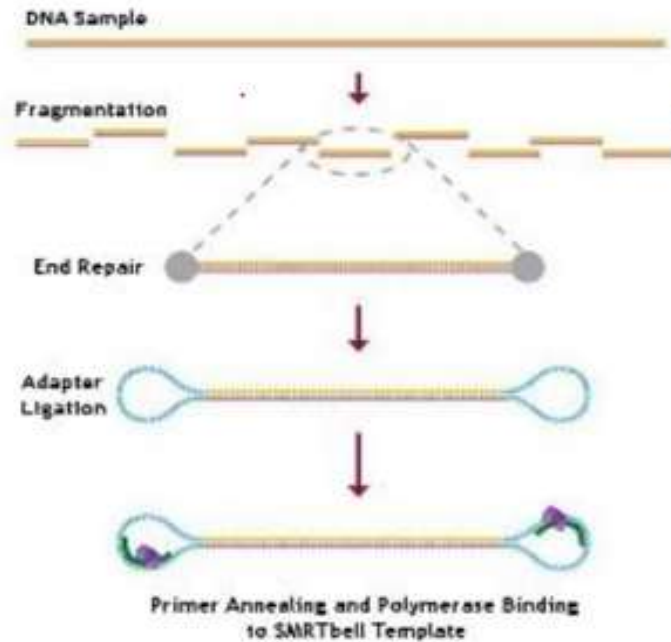


Nanopore

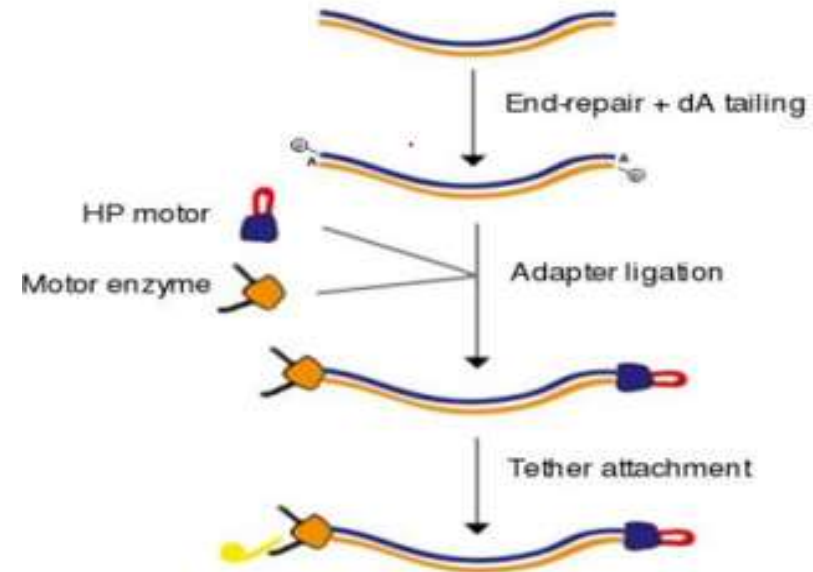


# Preparazione delle libraries

## ❖ Frammentazione del DNA, end-repair e ligazione



PacBio



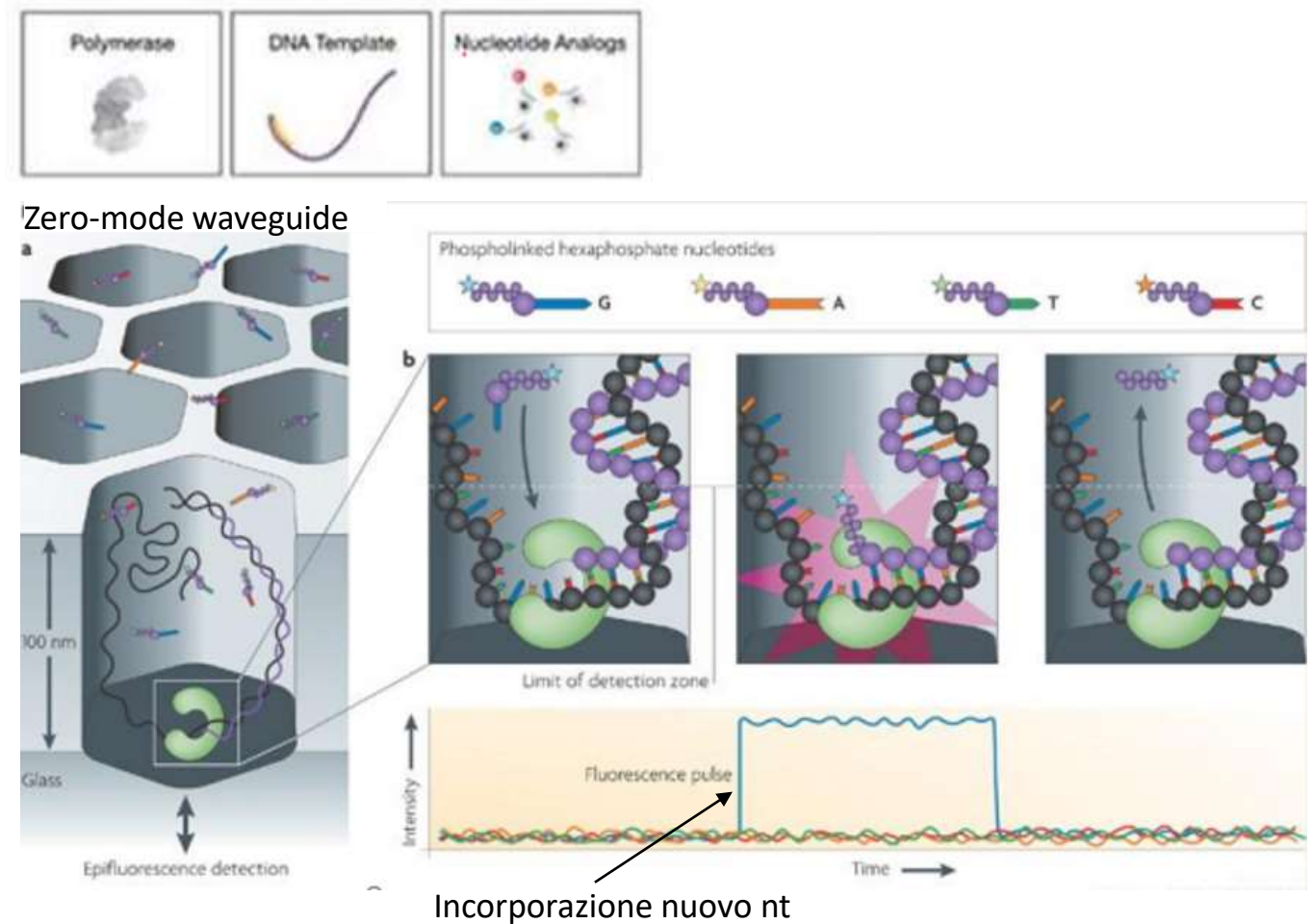
Nanopore





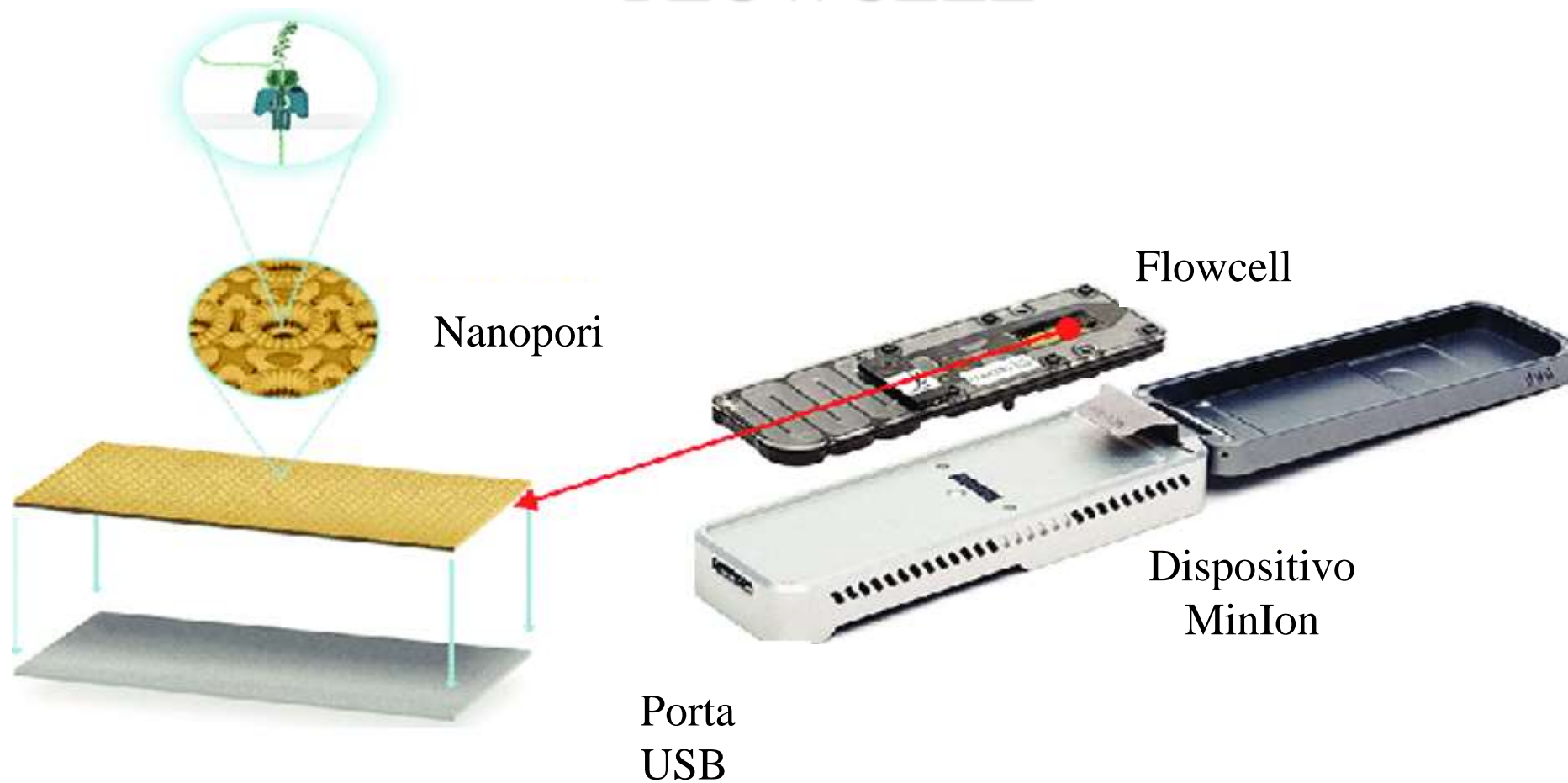
## Single Molecule Real Time Sequencing (SMRT)

- ❖ Template ancorato insieme alla polimerasi sul fondo di una ZMW
- ❖ 4 nt marcati ognuno con un fluorocromo diverso
- ❖ La polimerasi sintetizza il nuovo filamento, incorpora nuovi nt, il fluorocromo viene rilasciato
- ❖ Il sistema di rilevazione registra l'evento luminoso associato





# PIATTAFORMA NANOPORE FLOWCELL



# Preparazione della library: Adattatori



- **Leader- adaptor;** 2 oligont semi complementari che appaiandosi formano una struttura ad Y
- **Hairpin- adaptor;** 2 oligont con una regione complementare interna che formano una struttura ad hairpin

Entrambi gli adattatori presentano un complesso enzimatico (**motor protein**) che consente il passaggio del DNA attraverso il nanoporo, Gli adattatori guidano il DNA in prossimità del nanoporo attraverso oligont **tethering** che hanno affinità con i polimeri di membrana.



## Nanoporo;

proteine di membrana (porine, es  $\alpha$ -emolisina))

## Membrana;

Una differenza di potenziale applicata ai lati della membrana genera il passaggio di corrente attraverso il poro

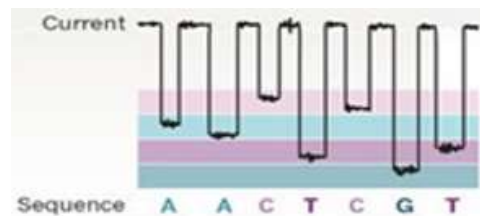
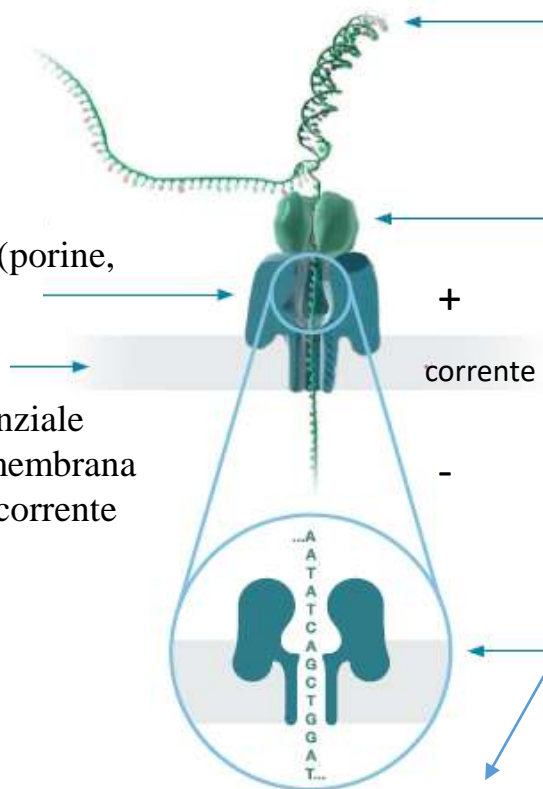
## Template;

dsDNA o cDNA con gli adattatori  
che consentono l'ancoraggio ai  
nanopori.

## Complesso enzimatico;

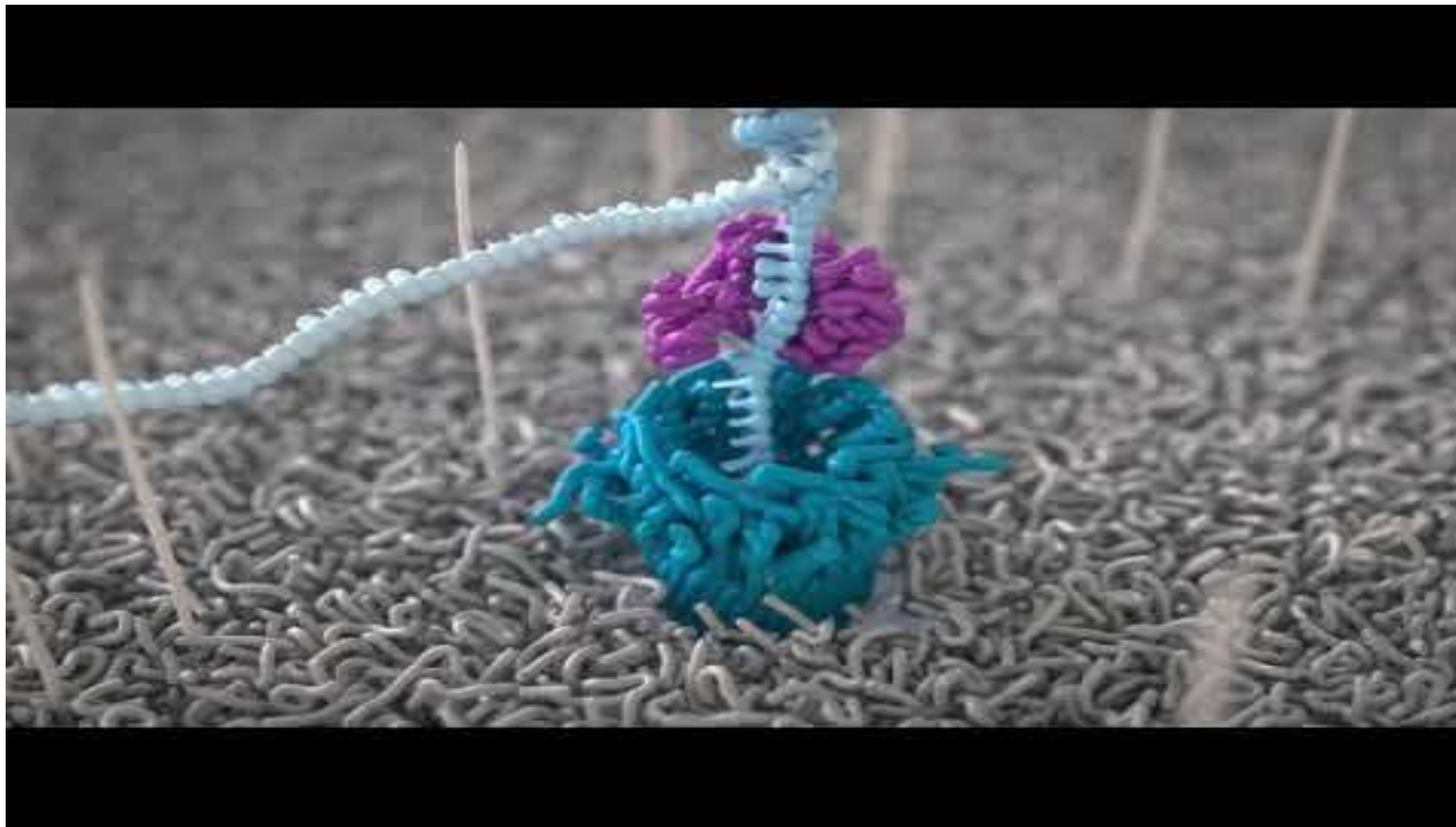
contatta il nanoporo, promuove lo svolgimento della doppia elica di DNA e catalizza il passaggio di uno dei due filamenti (template) attraverso il nanoporo (**single molecule strand sequencing**).

Differenti basi causano una diminuzione più o meno marcata della corrente misurata a seconda dell'ingombro.



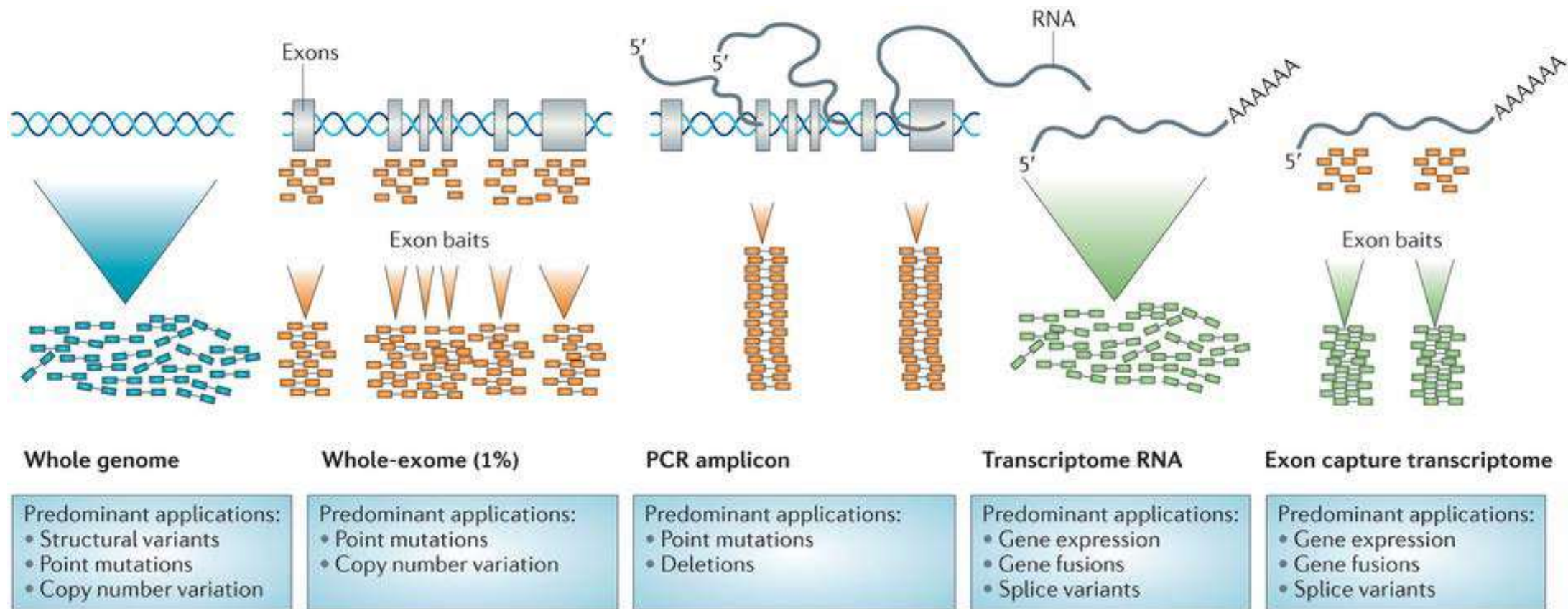


# Video; Tecnologia Nanopore





# Approcci



# Ambiti applicativi

## DNA

- Resequencing Geni/ Genomi Noti
- Sequenziamento De Novo
- Sequenziamento di regioni target
- Identificazione SNP/mutazioni
- Identificazione riarrangiamenti strutturali

## WGS

- Target seq
- Exome seq

## Transcriptome seq

## RNA

- RNA-seq
- Studi di espressione
- Small-RNA

## Regolazione

- Metilazione
- Analisi interazioni DNA-proteine

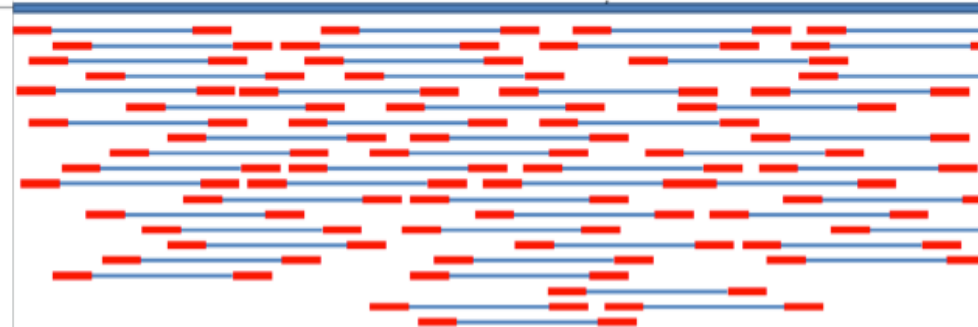


# Ricapitolando...

## Sequenziamento classico (Sanger)



## Sequenziamento Nuova Generazione



**READS:** Dati ottenuti dal sequenziamento, nella forma di una stringa di basi nucleotidiche.

### Sequenziamento Sanger

- Sequenze lunghe
- Poche sequenze
- **Costo più alto**
- Basso tasso di errore

### Sequenziamento NGS

- Sequenze corte
- Milioni di sequenze
- **Minor costo**
- Ridondanza per compensare errori di lettura





GRAZIE PER L'ATTENZIONE!!!

