

The background features four stylized DNA double helix structures in the corners. The top-left and bottom-right structures are rendered with dark grey spheres and lines, while the top-right and bottom-left structures are rendered with light grey spheres and lines. The central text is positioned between these decorative elements.

BIOINFORMATICA

II

SEQUENZIAMENTO

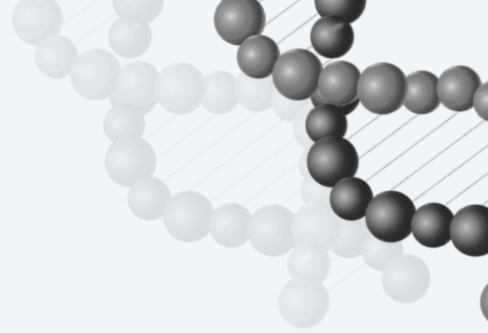
ARGOMENTI

01 SEQUENZIAMENTO

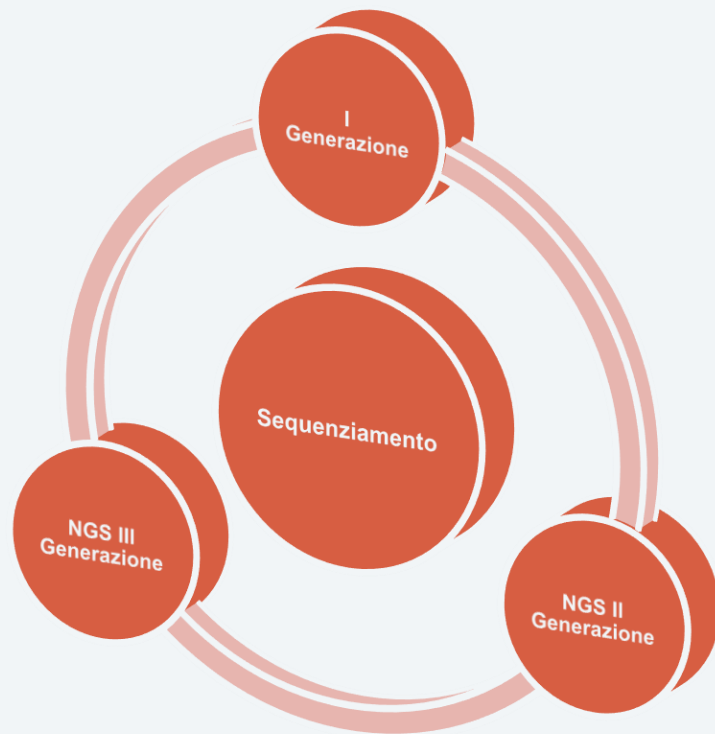
03 TECNICHE DI II
GENERAZIONE

02 TECNICHE DI I
GENERAZIONE

04 TECNICHE DI III
GENERAZIONE



SEQUENZIAMENTO



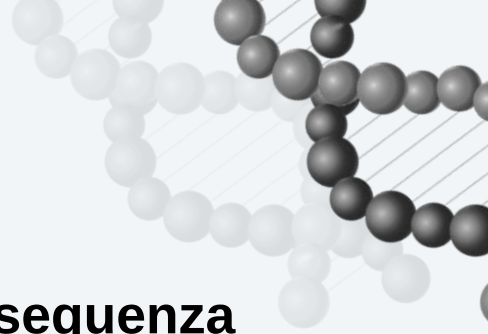


01

SEQUENZIAMENTO

SEQUENZIAMENTO: Definizione

Il sequenziamento è la tecnica per determinare la sequenza dei nucleotidi (tipica di una molecola di acido nucleico) o la sequenza di aminoacidi di una proteina, attraverso appropriate tecniche

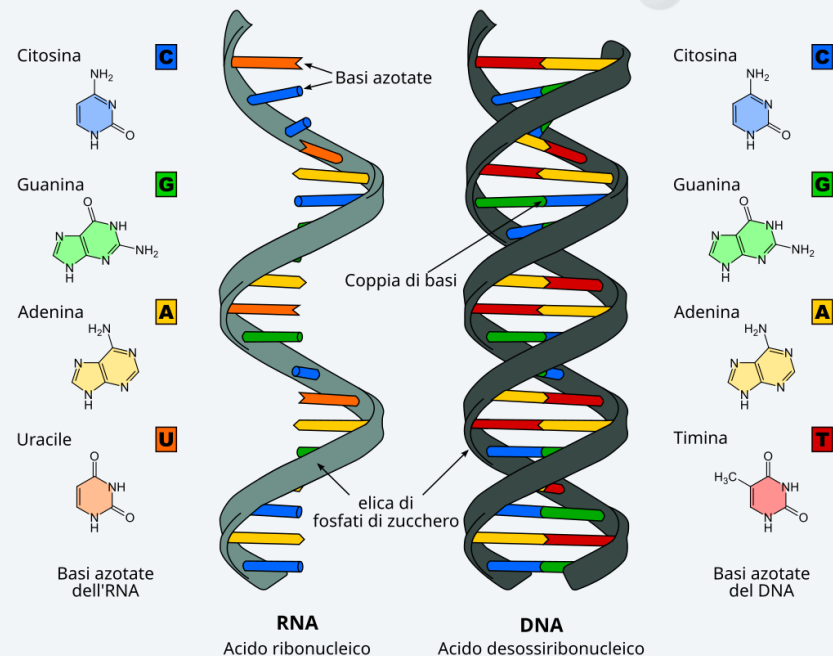


SEQUENZIAMENTO: Nucleotidi

I nucleotidi sono gli elementi che costituiscono gli acidi nucleici (il DNA o l'RNA) e sono formati da 3 porzioni: due non variano e sono sempre uno zucchero (il deossiribosio nel DNA e il ribosio nell'RNA) e un gruppo fosfato, mentre la terza, una base azotata.

Le **basi azotate del DNA** sono l'adenina (A), la citosina (C), la guanina (G) e la timina (T). Quest'ultima è sostituita dall'uracile (U) nell'RNA.

Il DNA è fatto di due filamenti avvolti tra loro, i cui legami si formano proprio tra le basi azotate, e in particolare tra A e T e tra C e G. Anche per questo si parla di **complementarità delle basi**. Sequenziando il DNA si conosce l'ordine delle sue basi



SEQUENZIAMENTO: Amminoacidi

Gli **aminoacidi** (o amminoacidi) sono l'unità strutturale primaria delle proteine. Gli aminoacidi sono paragonabili a mattoncini che, uniti da un collante chiamato legame peptidico, formano una lunga sequenza che dà origine ad una proteina

Aminoacido	Abbreviazione (3 lettere)	Abbreviazione (1 lettera)
Alanina	Ala	A
Arginina	Arg	R
Asparagina	Asn	N
Acido aspartico	Asp	D
Cisteina	Cys	C
Acido glutammico	Glu	E
Glutamina	Gln	Q
Glicina	Gly	G
Istidina	His	H
Isoleucina	Ile	I
Leucina	Leu	L
Lisina	Lys	K
Metionina	Met	M
Fenilalanina	Phe	F
Prolina	Pro	P
Serina	Ser	S
Treonina	Thr	T
Triptofano	Trp	W
Tirosina	Tyr	Y
Valina	Val	V



02

SEQUENZIAMENTO I GENERAZIONE

SEQUENZIAMENTO I GENERAZIONE

Nella seconda metà degli anni Settanta furono sviluppati **due metodi di sequenziamento**, uno nel 1973 a cura di Maxam e Gilbert e uno nel 1975 più veloce ed efficace, ideato dal chimico britannico Frederick Sanger all'Università di Cambridge. Per i loro contributi, Gilbert e Sanger hanno ricevuto il **premio Nobel** per la chimica nel 1980

La prima sequenza pubblicata, lunga 24 basi, è del **1973**, a opera degli scienziati statunitensi Maxam e Gilbert della Harvard University a Cambridge (Massachusetts, USA) dopo due anni di lavoro

5'--TGGAATTGTGAGCGGATAACAATT3'
3'--ACCTTAACACTCGCCTATTGTTAA 5'

I GENERAZIONE: Generalità

Il metodo di **sequenziamento del DNA con dideossinucleoterminalori** o **metodo di Sanger** consiste di tre fasi:

1. Preparazione del campione
2. Reazione di sequenziamento
3. Elettroforesi



Frederick Sanger

I GENERAZIONE: Preparazione campione

Nella preparazione del campione, il filamento di DNA che si vuole sequenziare è copiato artificialmente.

Sono due le tecniche che permettono questo processo di duplicazione:

1. **DNA ricombinante**
2. **Reazione a catena della polimerasi** (Polymerase Chain Reaction, PCR)

I GENERAZIONE: DNA ricombinate

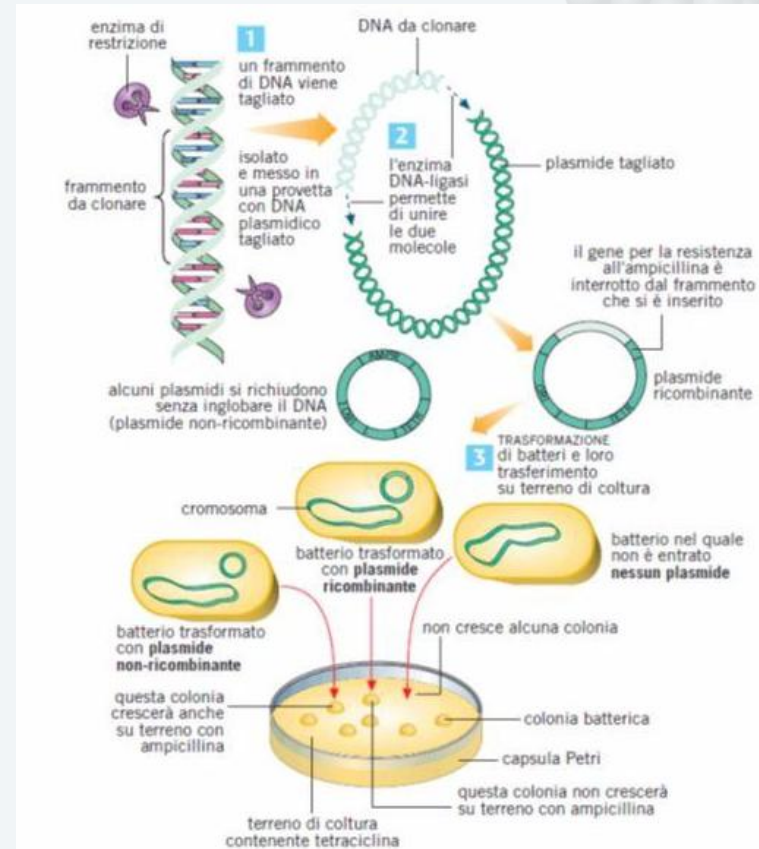
La replica mediante **DNA ricombinante** prevede l'inserimento di un frammento di DNA in un vettore (es.: plasmide, virus, cosmide, cromosoma artificiale) formando il DNA ricombinante e l'introduzione di questo in una cellula ospite mediante endonucleasi di restrizione (cioè enzimi, prodotti da batteri, in grado di riconoscere delle sequenze di DNA e di tagliare le molecole di acido nucleico in parti specifiche).

Questa tecnica è una di quelle definibili **clonaggio**

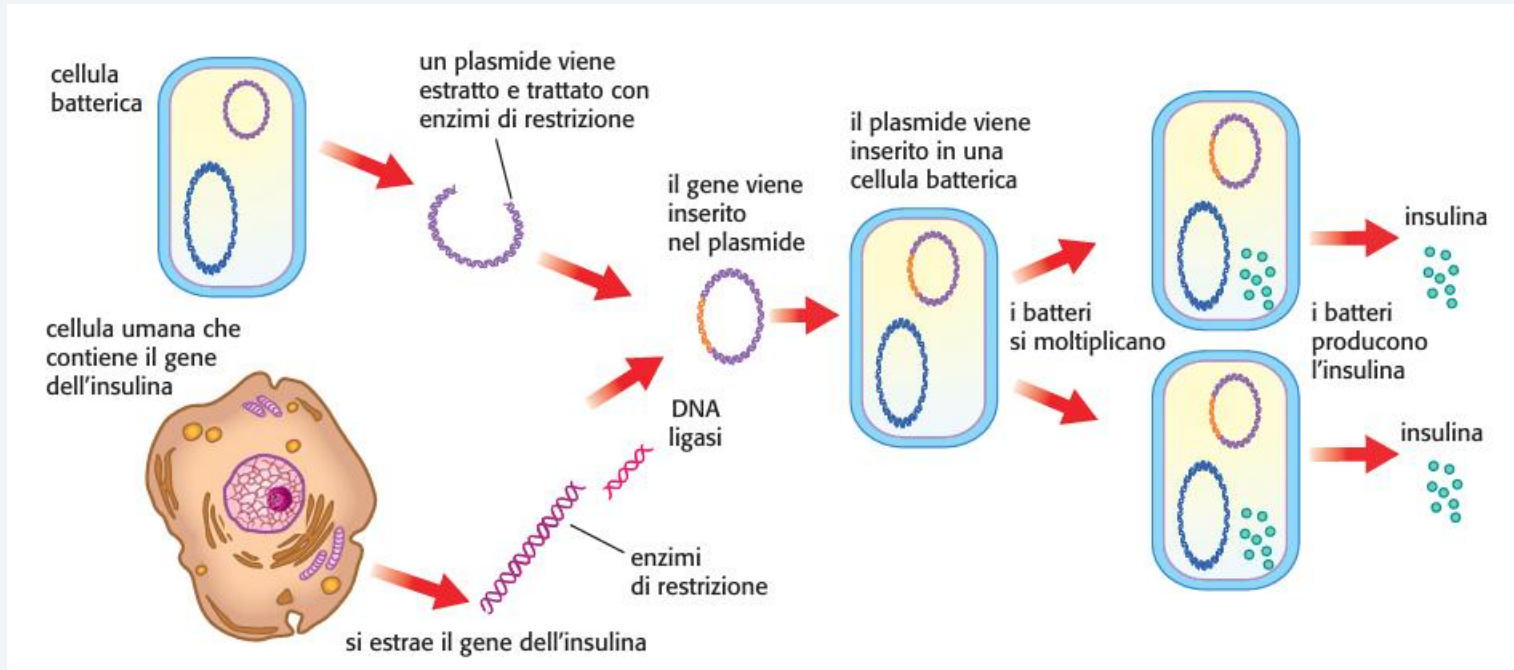
I GENERAZIONE: DNA ricombinate

Il metodo di replica usato nel **DNA ricombinate** può essere schematizzato nelle seguenti fasi:

1. isolamento di un gene;
2. inserimento del gene in un vettore;
3. inserimento del costrutto gene-vettore in una cellula ospite;
4. isolamento del clone desiderato;
5. purificazione del clone;
6. produzione di diverse cellule contenenti il clone



I GENERAZIONE: DNA ricombinate (esempio)

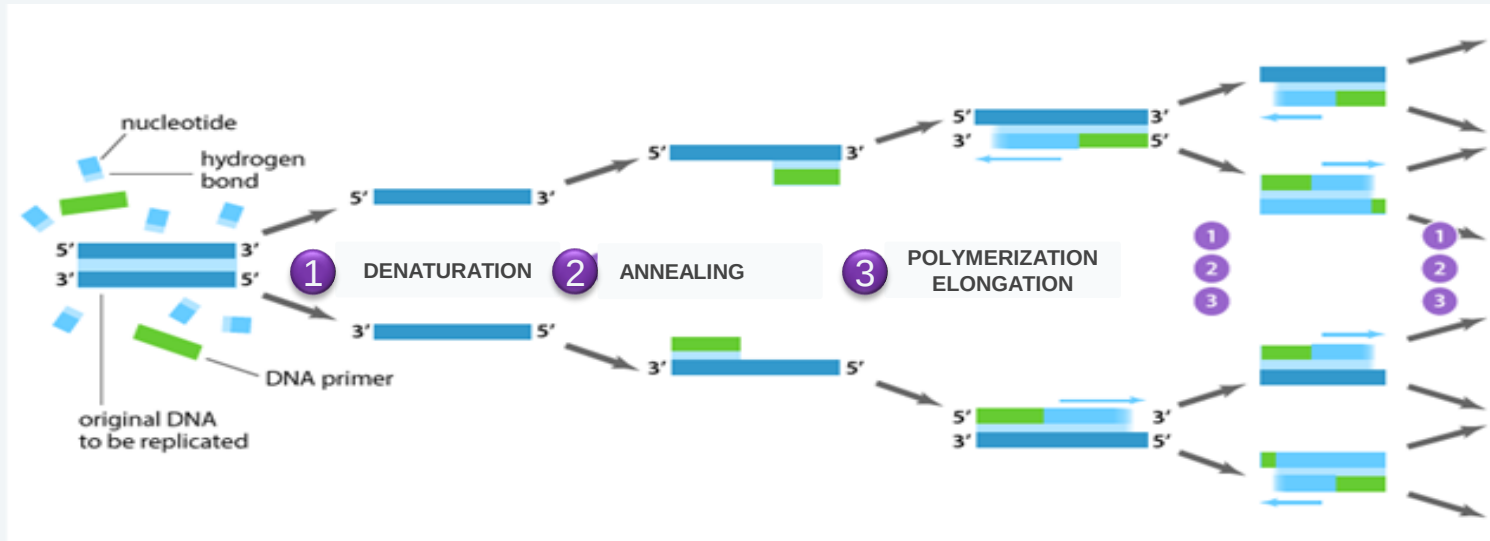


Esempio di DNA ricombinante per ottenere batteri modificati che producono insulina

I GENERAZIONE: PCR

Il metodo **PCR** si esplica in tre fasi:

- Denaturation (apertura del DNA)
- Annealing (appaiamento dei primer)
- Polymerization (copia del filamento) e Elongation (completamento copiatura ssDNA)

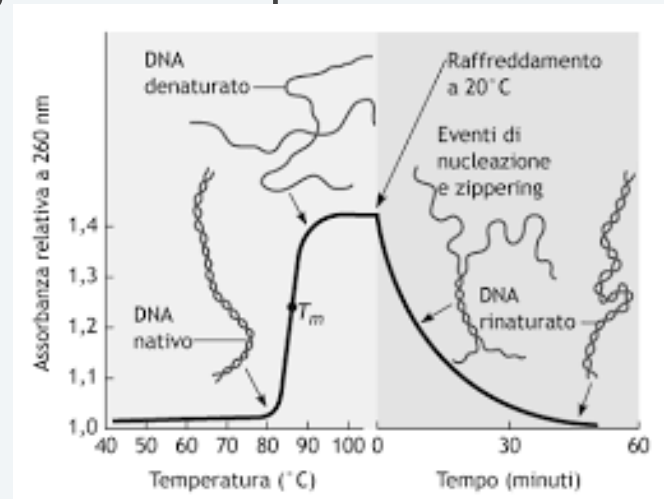
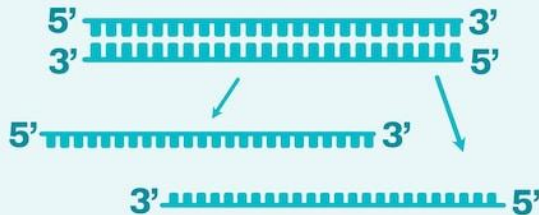


I GENERAZIONE: PCR

Nella fase di **denaturazione** (*denaturation*) si porta la temperatura fino a circa 90°C permettendo al doppio filamento di DNA di separarsi in due, aprendosi come fosse una cerniera lampo. Questo avviene perché si rompono i legami idrogeno che tengono uniti i filamenti complementari del DNA contenenti la regione da amplificare

DENATURAZIONE

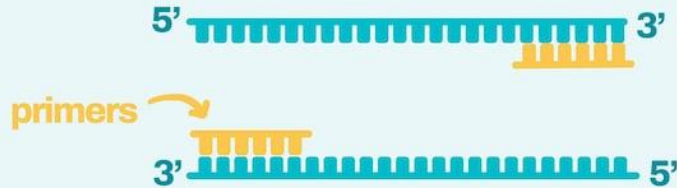
90-94°C circa



I GENERAZIONE: PCR

Abbassando la temperatura tra i 50 e i 60° C circa, si favorisce l'**appaiamento** (*alleaning*) dei **primers**, due segmenti formati da oligonucleotidi che definiranno gli estremi della regione da sequenziare

APPAIAMENTO 50-60°C circa



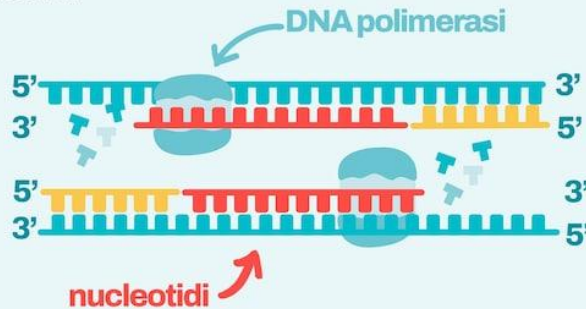
I primers hanno una lunghezza di circa 15-30 basi e si uniscono in maniera complementare al DNA stampo nella porzione di interesse, formando delle zone che serviranno da innesci. In una stessa reazione di PCR si utilizzano più primer differenti.

I GENERAZIONE: PCR

Nella estensione la temperatura è portata attorno ai 72°C in modo che la **polimerasi** sintetizzi i nuovi filamenti complementari nella direzione corretta, aggiungendo uno dopo l'altro i nucleotidi complementari al filamento stampo. In questo modo si ricrea il doppio filamento che è quindi formato da una stringa "originale" e una complementare nuova

ESTENSIONE

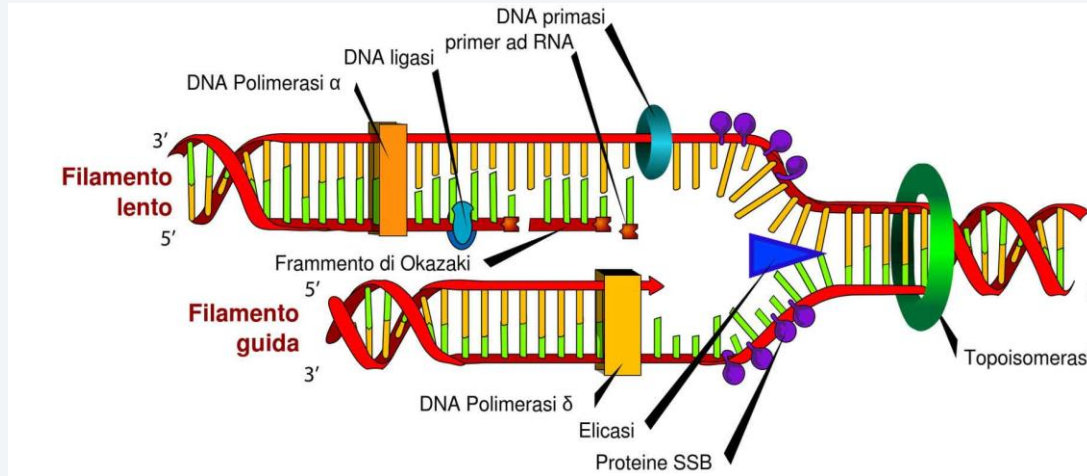
72°C circa



Ipotizzando di aver trattato un filamento di DNA a doppia elica alla fine del primo ciclo si ottengono due filamenti a doppia elica perfettamente uguali

I GENERAZIONE: Osservazione

La **DNA polimerasi** catalizza la produzione di nuovi filamenti complementari a quelli da copiare



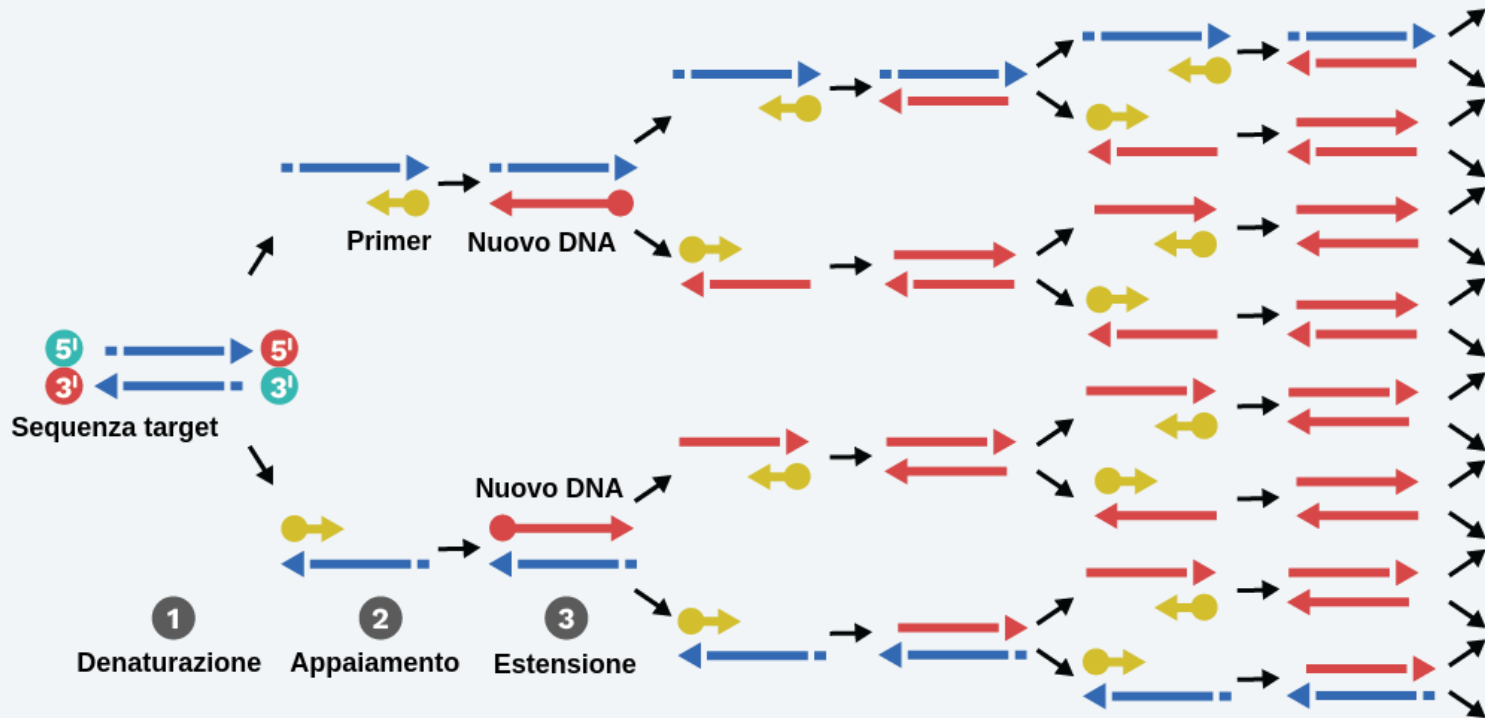
<https://www.dnaexpress.it/glossario/dna-polimerasi/>

La DNA polimerasi è un enzima che viene in gran parte distrutto dalle alte temperature necessarie durante la fase di denaturazione

Per ovviare a tal problema è utilizzata la DNA polimerasi del batterio *Thermus Aquaticus* che, vivendo in acque caldissime, dispone di un intero meccanismo termoresistente, compresa la DNA polimerasi



I GENERAZIONE: PCR (duplicazione)



I GENERAZIONE: Reazione di sequenziamento

Nella fase di **reazione di sequenziamento** sono allestite quattro miscele in quattro provette, una per ogni base. In ogni provetta è aggiunta la DNA polimerasi, i quattro nucleotidi (dATP, dCTP, dGTP, dTTP), e una piccola quantità di un dideossinucleoside trifosfato.

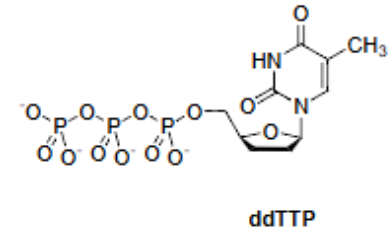
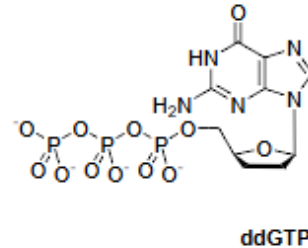
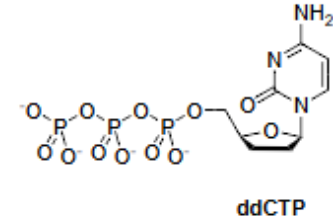
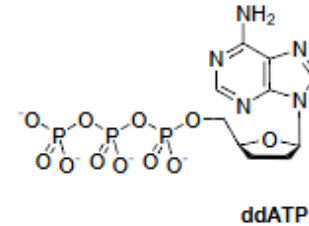
Un dideossinucleoside trifosfato (ddNTP), è un nucleotide che non ha il gruppo $-OH$ in posizione 3' dello zucchero.

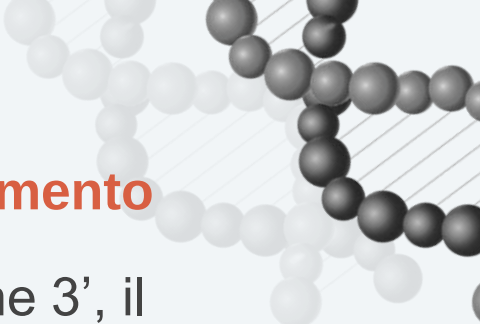


I GENERAZIONE: Reazione di sequenziamento

I quattro deossinucleosidi trifosfato (dNTPs), a destra la loro composizione chimica, modificati sono indicati dalle stringhe e dai rispettivi simboli:

- ddCTP 
- ddTTP 
- ddATP 
- ddGTP 





I GENERAZIONE: Reazione di sequenziamento

Poiché i ddNTPs mancano del gruppo -OH in posizione 3', il nucleotide successivo non può esser legato come avviene nella replica naturale del DNA

Per questo motivo la sintesi si arresta alla posizione in cui un ddNTP è stato incorporato all'estremità in accrescimento di un filamento di DNA (**fase di terminazione**)



I GENERAZIONE: Reazione di sequenziamento

La replicazione del DNA procede e nella provetta si viene a trovare una miscela di filamenti stampo del DNA insieme ad una varietà di filamenti neosintetizzati più brevi. I filamenti nuovi, ognuno dei quali termina con un ddNTP, avranno lunghezze differenti.

Ad esempio una volta che la DNA polimerasi incontra una base T sul filamento stampo, essa può aggiungere o un dATP o un ddATP. Se è aggiunto un dATP la crescita del filamento continua, mentre se è aggiunto un ddATP, la crescita del filamento si arresta.

Questo processo è ripetuto in altre provette rispettivamente per il ddGTP, ddTTP, ddCTP.

I GENERAZIONE: Reazione di sequenziamento

Sequenza:

3' GAGCAAATTCCGATACATTATTGT... 5'

Primer:

5' CTCGTTTAAG 3'

CTCGTTTAAGG **G**
CTCGTTTAAGGC **C**
CTCGTTTAAGGGCT **T**
CTCGTTTAAGGGCTA **A**
CTCGTTTAAGGGCTAT **T**
CTCGTTTAAGGGCTATG **G**

OSSERVAZIONE IMPORTANTE: Nel sequenziamento automatico si inserisce ai quattro ddNTPs un composto fluorescente, diverso per ogni base.

I GENERAZIONE: Elettroforesi

L'**elettroforesi** è un processo elettrocinetico nel quale molecole e particelle cariche, sotto l'influenza di un campo elettrico, migrano in direzione del polo che ha carica opposta.

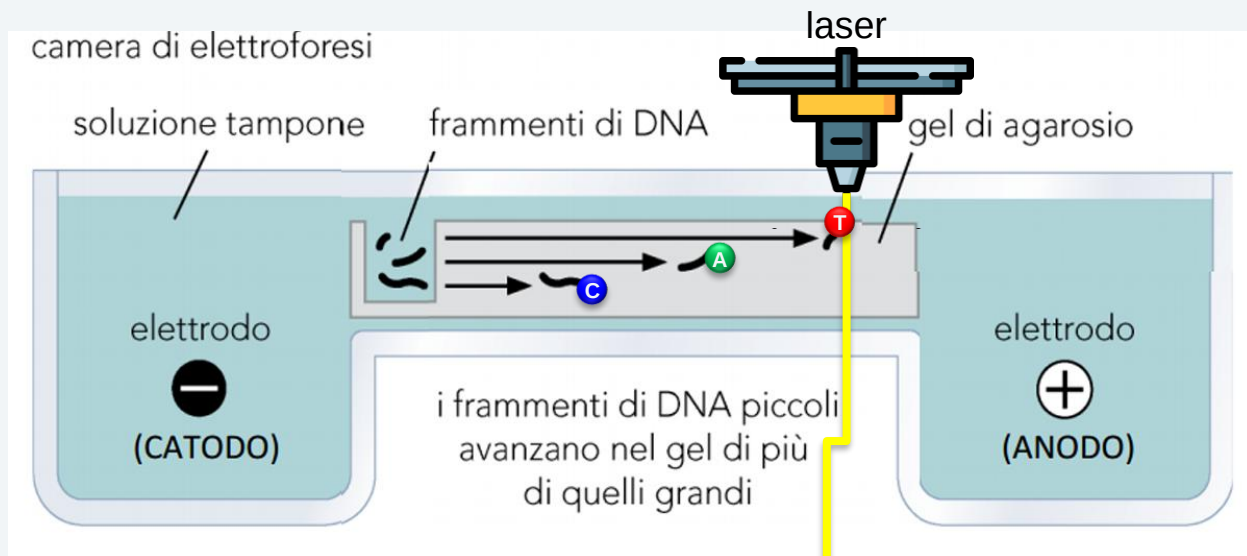
Grazie alla presenza dei gruppi fosfato, le molecole di DNA sono cariche negativamente e quindi migrano verso il polo positivo (anodo) se sottoposte a un campo elettrico, con velocità che dipende dalla loro lunghezza e dall'intensità della corrente.

I GENERAZIONE: Elettroforesi

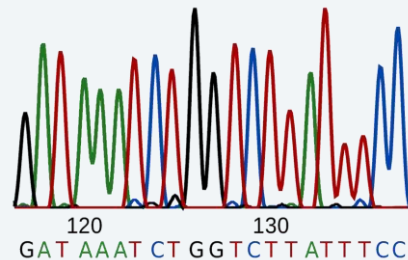
Durante la corsa elettroforetica, i frammenti sono letti in ordine di lunghezza crescente da un fascio laser che eccita i marcatori fluorescenti uniti alle ddNTP.

Quindi l'intensità della luce emessa è misurata e tale informazione, cioè quale colore di fluorescenza e dunque il tipo di ddNTP è presente all'estremità di ogni filamento di differente lunghezza, è inviata ad un elaboratore per, infine, essere ricostruita.

I GENERAZIONE: Elettroforesi



Elettroferogramma



I GENERAZIONE: Reazione di sequenziamento

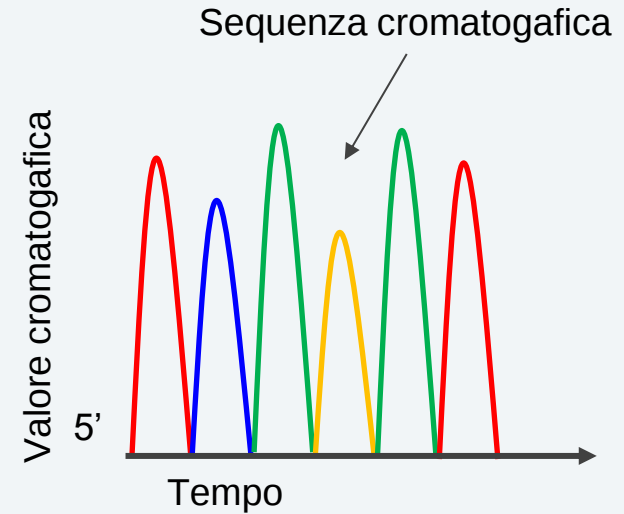
Sequenza:

3' GAGCAAATTCCGATACATTATTGT... 5'

Primer:

5' CTCGTTTAAG 3'

CTCGTTTAAGG **G**
CTCGTTTAAGGC **C**
CTCGTTTAAGGGCT **T**
CTCGTTTAAGGGCTA **A**
CTCGTTTAAGGGCTAT **T**
CTCGTTTAAGGGCTATG **G**



I GENERAZIONE

L'introduzione dell'elettroforesi capillare ha consentito un notevole miglioramento temporale e qualitativo del sequenziamento.

Sono stati inoltre sviluppati modelli di sequenziatori automatici che sono in grado di eseguire corse elettroforetiche multiple su apparecchi multi-capillari.

Ci sono comunque dei limiti:

- Necessità di clonare i frammenti da sequenziare (tramite la PCR)
- Limitazione massima di sequenziamento: per ogni corsa elettroforetica è di massimo 1000 coppie di basi (unità di misura, base pair – bp) per ciascuna elettroforesi capillare
- Costi elevati

FOTODIODO

Il fotodiiodo è un componente elettronico semiconduttore progettato per convertire la luce in una corrente elettrica.
È ampiamente utilizzato in dispositivi come sensori ottici, fotometri, telecomunicazioni e in strumenti scientifici come i sequenziatori.

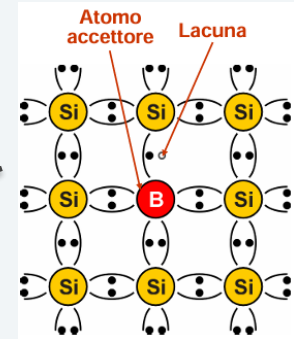
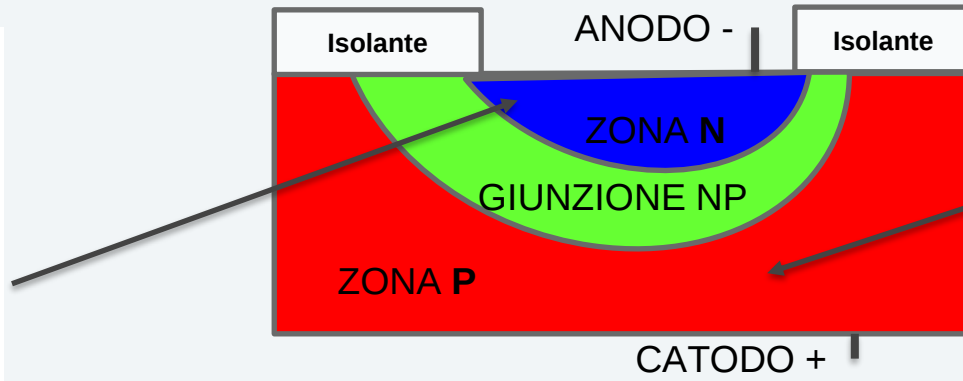
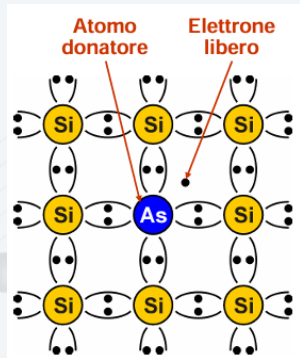
FOTODIODO

Un fotodiodo è costruito principalmente come una giunzione PN con le seguenti parti principali:

ZONA P: Materiale drogato positivamente (ha "lacune" come portatori di carica)

ZONA N: Materiale drogato negativamente (ha elettroni come portatori di carica)

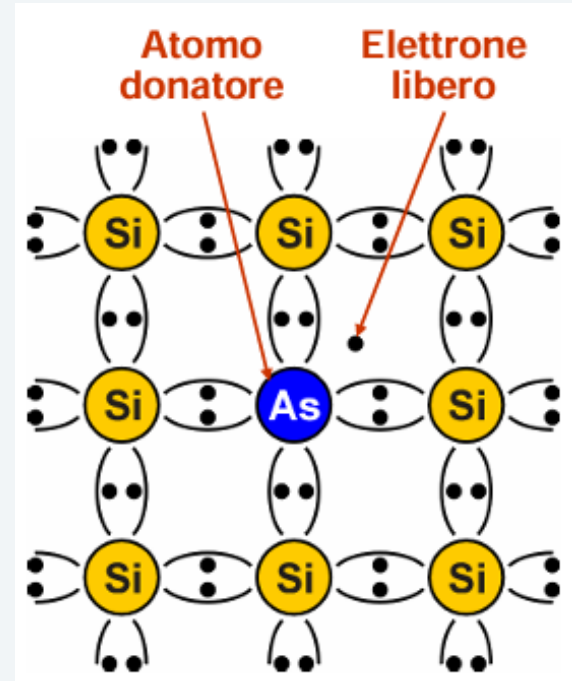
GIUNZIONE NP: sono presenti atomi donatori che hanno ceduto un elettrone e atomi accettori che hanno acquistato un elettrone



FOTODIODO: Drogaggio N

Se all'interno di un reticolo cristallino di silicio (elemento tetravalente) si sostituisce un atomo di questo elemento con uno di arsenico, che è un elemento pentavalente, allora quattro elettroni dell'arsenico si legano con altrettanti quattro atomi di silicio mentre il quinto resta libero e aumenta la conducibilità del silicio e nel reticolo cristallino resta uno ione fisso positivo.

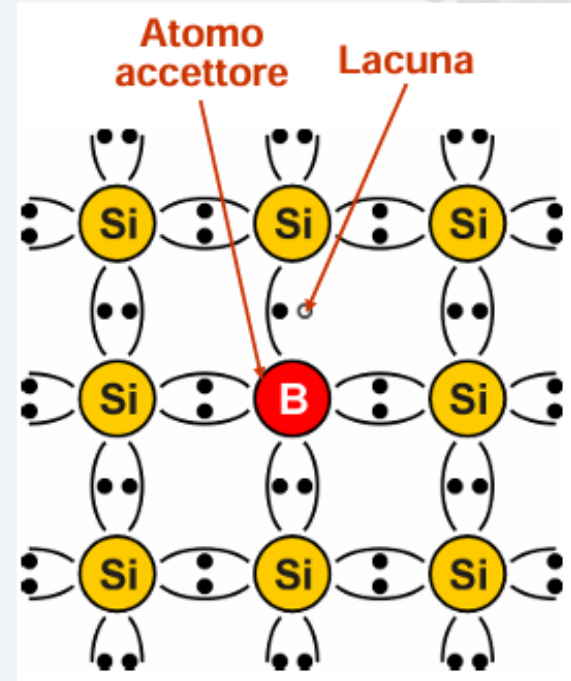
Il semiconduttore si dice **drogato di tipo N** perché i portatori di carica maggioritari sono gli elettroni negativi mentre le lacune sono i portatori minoritari. Gli atomi di arsenico si chiamano donatori.



FOTODIODO: Drogaggio P

Se all'interno di un reticolo cristallino di silicio si sostituisce un atomo di questo elemento con uno di boro, che è un elemento trivalente, questo non è in grado di saturare tutti e quattro i legami degli atomi vicini. Si forma, quindi, nel cristallo un posto libero che può essere occupato da un elettrone proveniente da un altro atomo generando così una lacuna. Questa aumenta la conducibilità del silicio e nel reticolo cristallino resta uno ione negativo fisso.

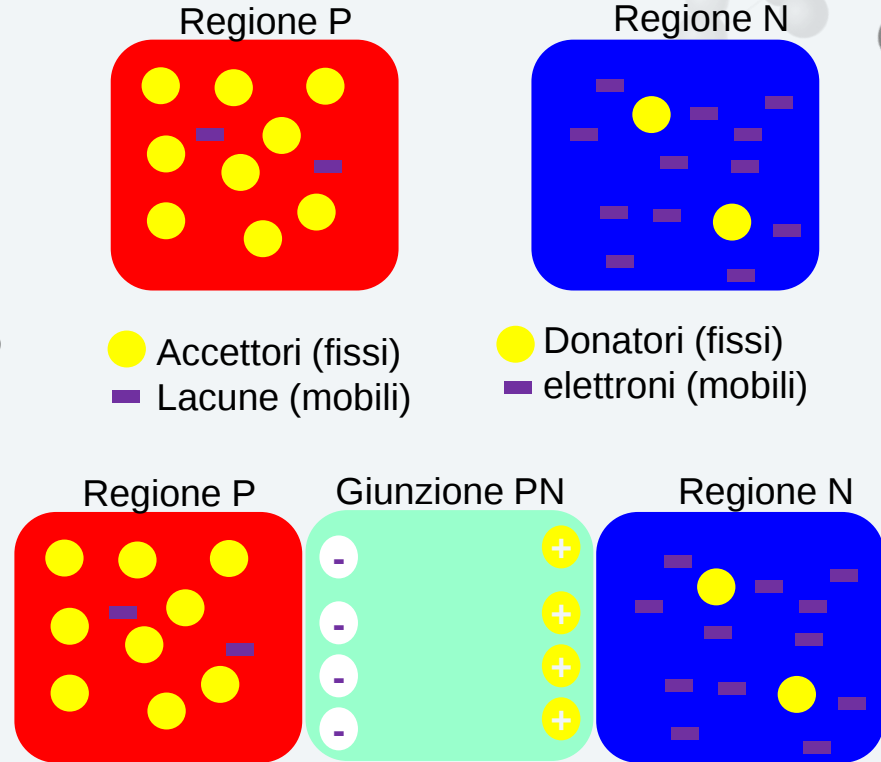
Il semiconduttore si dice **drogato di tipo P** perché le lacune positive sono i portatori di carica maggioritari mentre gli elettroni sono i portatori di carica minoritari. Gli atomi di boro si chiamano accettori.



FOTODIODO: Giunzione

Sia il semiconduttore drogato di tipo N che quello di tipo P sono elettricamente neutri, in quanto, le cariche degli elettroni e delle lacune in eccesso presenti rispettivamente nei due semiconduttori sono neutralizzate da quelle degli ioni fissi

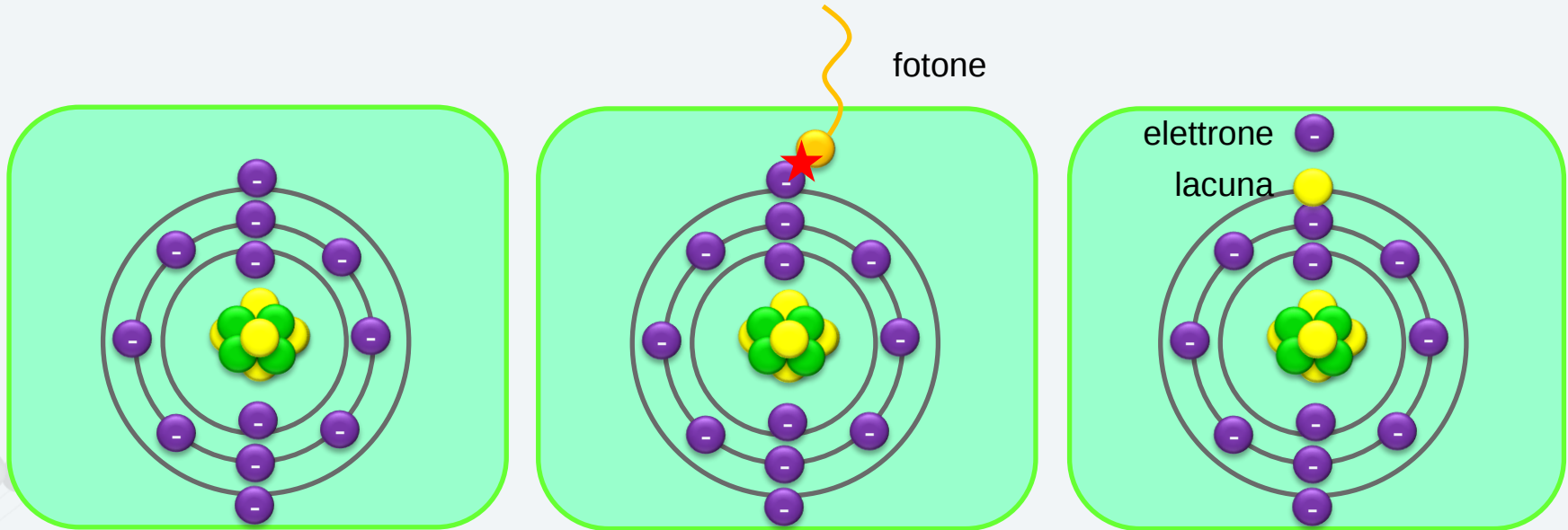
Quando si mettono a contatto i due blocchi (l'uno N, l'altro P) gli elettroni in eccesso, rispetto alle lacune libere, da N vanno in P neutralizzando le lacune in eccesso ivi presenti: le cariche fisse rimaste nei due semiconduttori formano una barriera di potenziale cioè una **giunzione P-N**. In questo modo nella giunzione PN a ridosso di N ci saranno elettroni che ostacoleranno gli elettroni di N a passare verso P e viceversa con le lacune



FOTODIODO

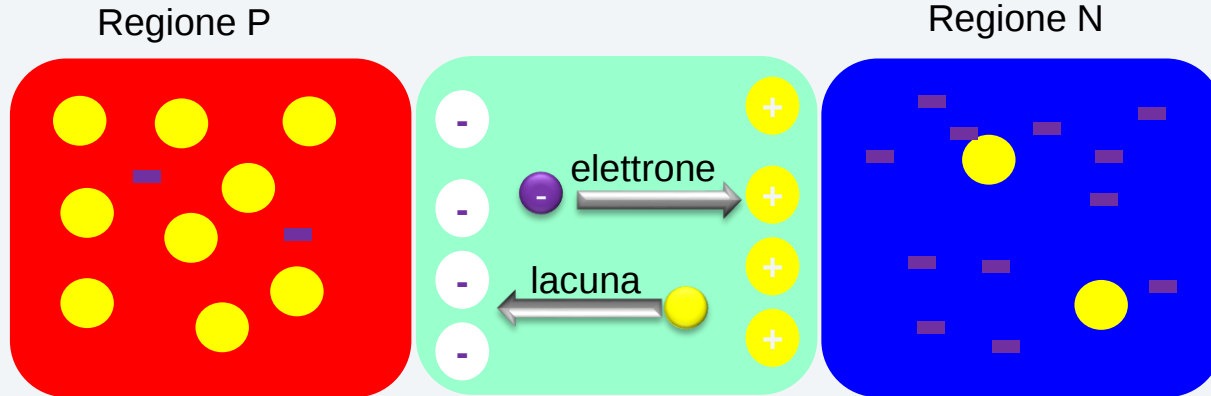
Se, nell'area di giunzione, un fotone colpisce un elettrone e la sua energia è sufficiente crea una coppia elettrone-lacuna.

I fotoni funzionano come “palle da biliardo”: un fotone colpendo un atomo scalza dall'orbita un elettrone e nell'orbita in cui si trovava si forma dunque una lacuna. Si dice che un fotone genera una coppia elettrone-lacuna



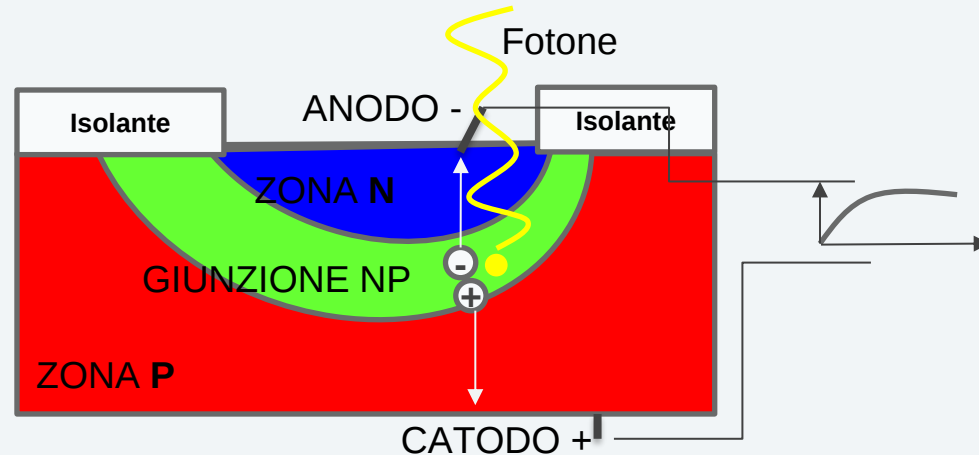
FOTODIODO

L'elettrone è spinto verso il lato N mentre la lacuna va verso il lato P.



FOTODIODO

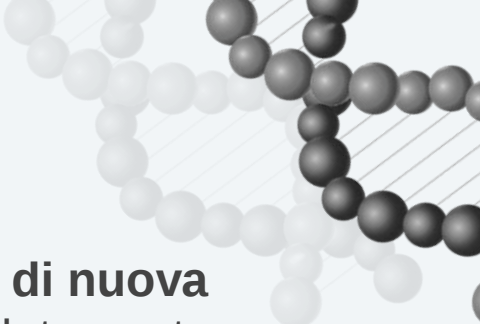
Si genera, quindi, passaggio di corrente: la cella diventa elemento attivo che trasforma l'energia elettromagnetica (energia fotonica) in energia elettrica





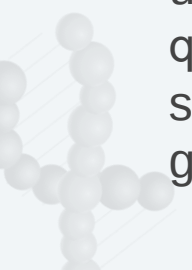
03

SEQUENZIAMENTO II GENERAZIONE



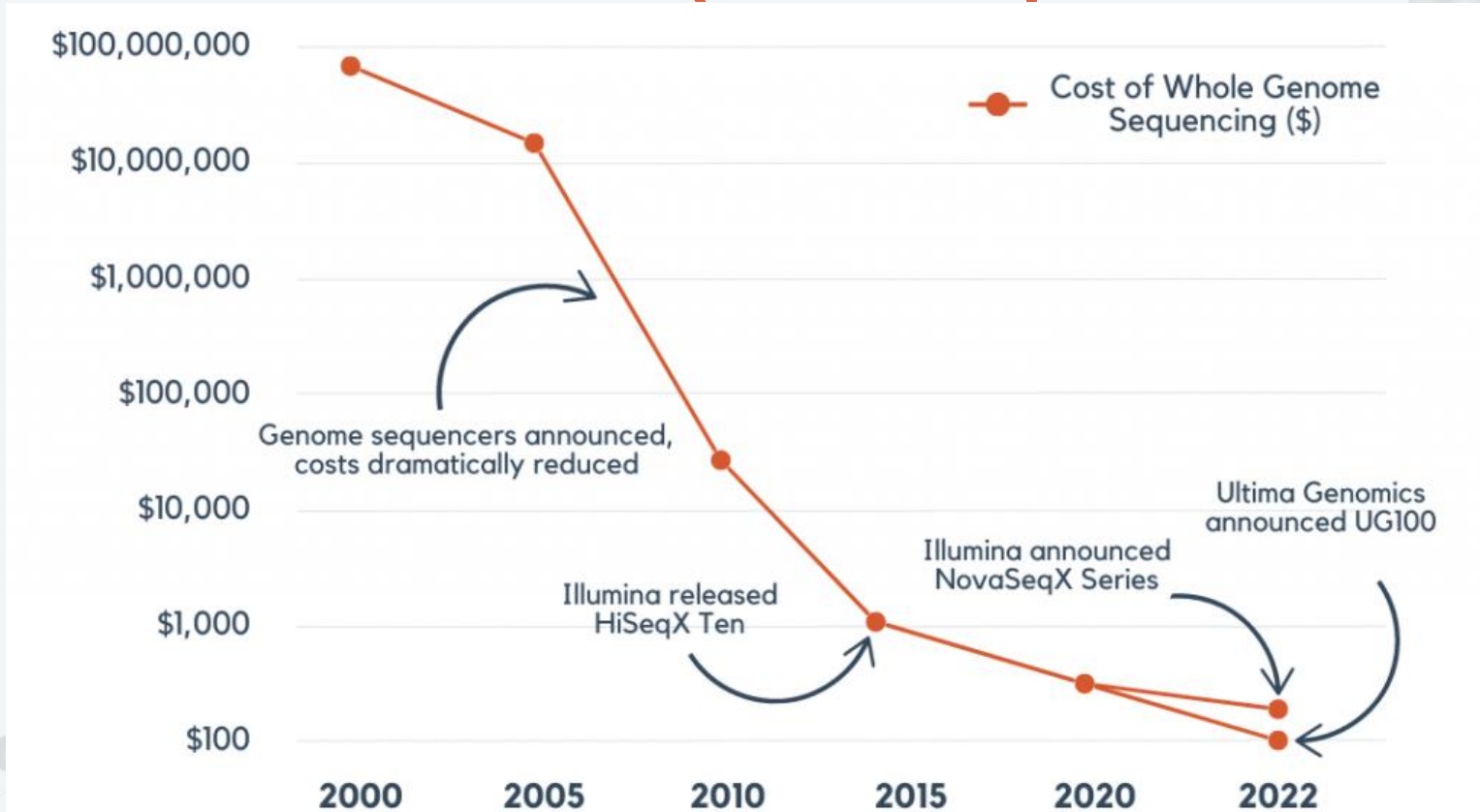
II GENERAZIONE

Nel 2007, l'introduzione delle **tecniche di sequenziamento di nuova generazione** (Next-Generation Sequencing, NGS) ha completamente rivoluzionato l'approccio al sequenziamento del DNA, dopo anni nei quali il metodo Sanger rappresentava ormai un punto fermo per gli scienziati e i ricercatori di tutto il mondo

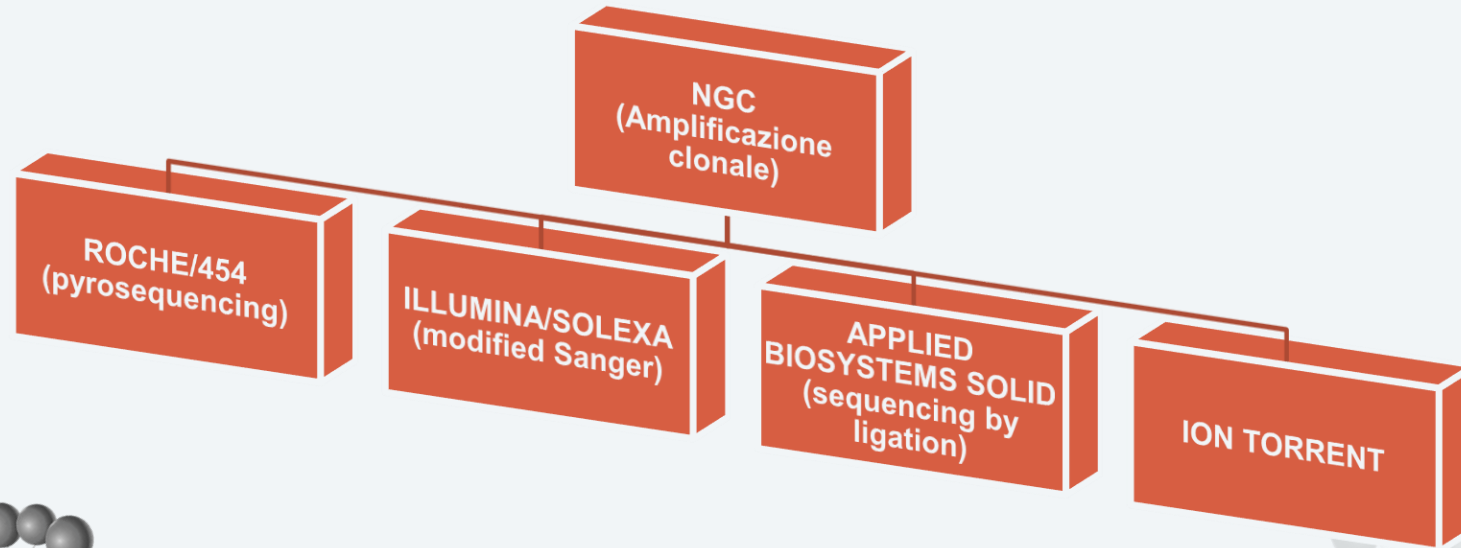


Per comprendere la portata dell'avanzamento tecnologico in termini di costi, basti pensare che il Progetto Genoma Umano impiegò 10 anni e costò \$3.000.000.000, mentre oggi un macchinario NGS è in grado di sequenziare un singolo genoma umano in un giorno e per meno di \$5.000. In termini quantitativi invece, nei primi anni '80 una persona avrebbe potuto sequenziare un genoma di 30000 paia di basi di DNA (circa le dimensioni del genoma del SARS-CoV-2) in circa **60 giorni**

II GENERAZIONE (costi sequenziamento)

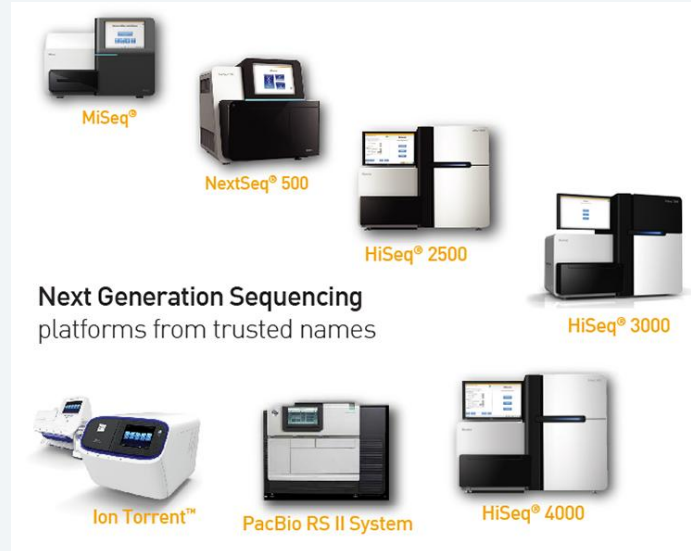


II GENERAZIONE: Classificazione



II GENERAZIONE: NGS - Generalità

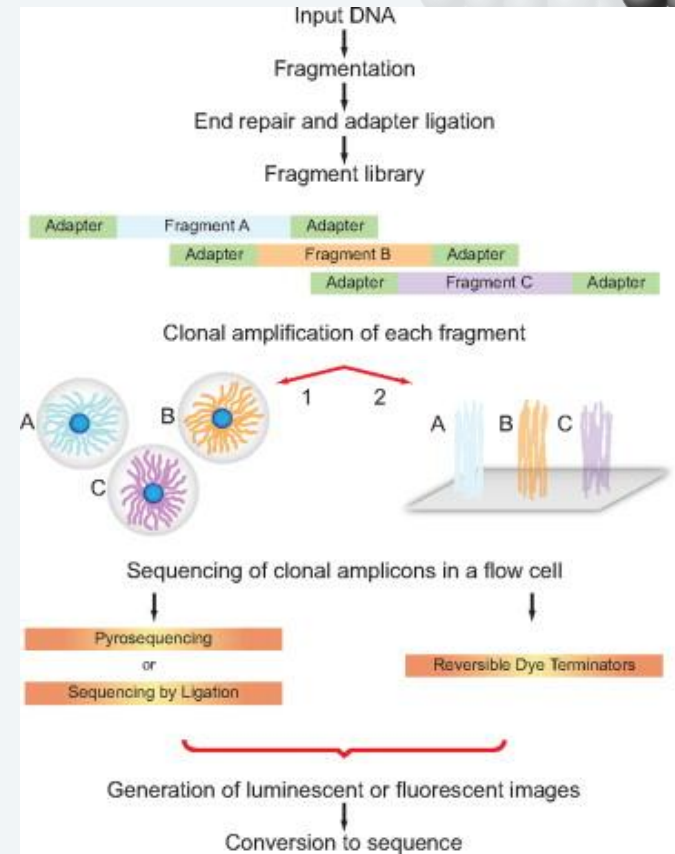
Il **sequenziamento genetico di nuova generazione** (Next generation sequencing, NGS) è l'insieme delle tecnologie di sequenziamento degli acidi nucleici che hanno in comune la capacità di sequenziare, in parallelo, milioni di frammenti di DNA



II GENERAZIONE: NGS - Generalità

Ci sono diverse tecniche NGS, ma in generale sono previste delle fasi comuni:

- Estrazione acidi nucleici (DNA o RNA)
- Frammentazione acidi nucleici e selezione dei frammenti di dimensione omogenea
- Aggiunta di adattatori alle estremità
- Clonaggio o immobilizzazione su un supporto per l'amplificazione
- Determinazione delle posizione dei nucleotidi



II GENERAZIONE: ROCHE 454

Il sistema 454 fu il primo processo di sequenziamento NGS lanciato sul mercato dalla azienda 454 (poi acquisita da Roche)

Il sistema Roche 454 è basato sulla tecnologia della **pirosequenziamento** (*pyrosequencing*), così chiamato per via dei pirofosfati che sono sfruttati nella fase di sequenziamento e successivamente in quella di individuazione dei nucleotidi

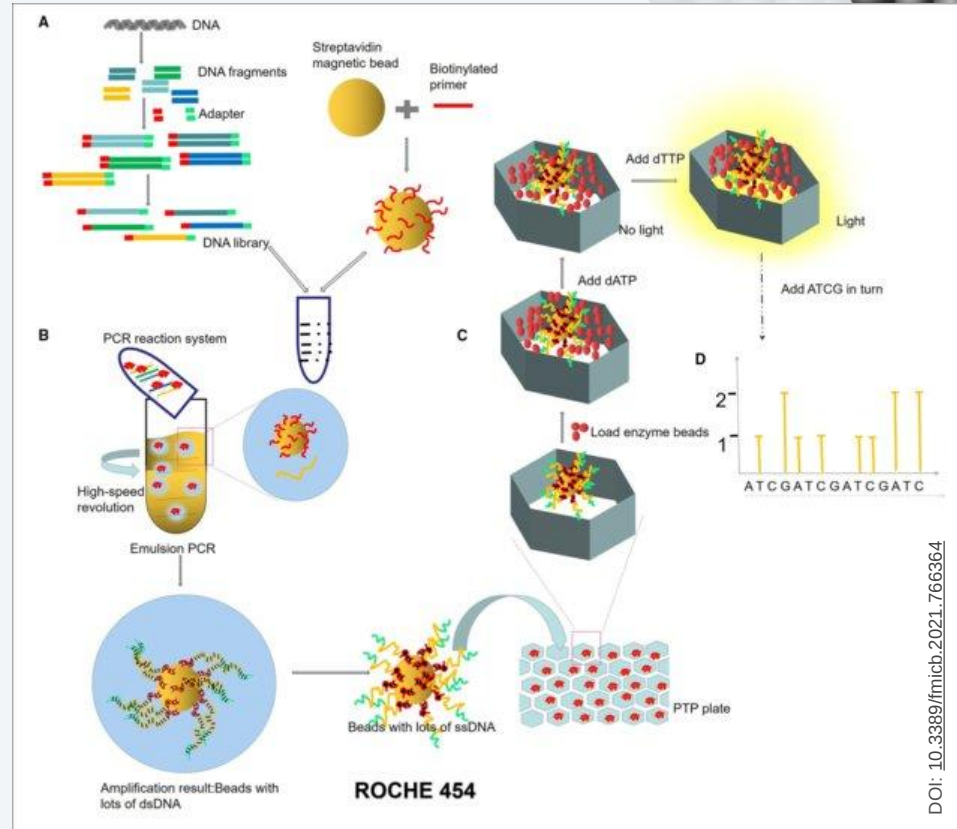
Il Roche 454 riconosce la sequenza grazie al susseguirsi di segnali luminosi causati dal distacco di un pirofosfato inserito ogniqualevolta un nucleotide è incorporato nel filamento in crescita.

II GENERAZIONE: ROCHE 454

Sequenziamento ROCHE 454:

A. Costruzione del sistema di microreazione. Il DNA è frammentato e a ciascuna sequenza si aggiunge un adattatore/primer alle estremità (DNA library). Le sfere magnetiche (*magnetic bead*) su cui sono innestati dei primer sono miscelate nel sistema di reazione PCR (frammenti di DNA, enzimi, dNTP)

B. PCR in emulsione. I filamenti replicati si legano alla sfera. Ogni molecola è depositata in un sistema ad alveare (PTP plate). Ciascuna cella contiene solo una sfera

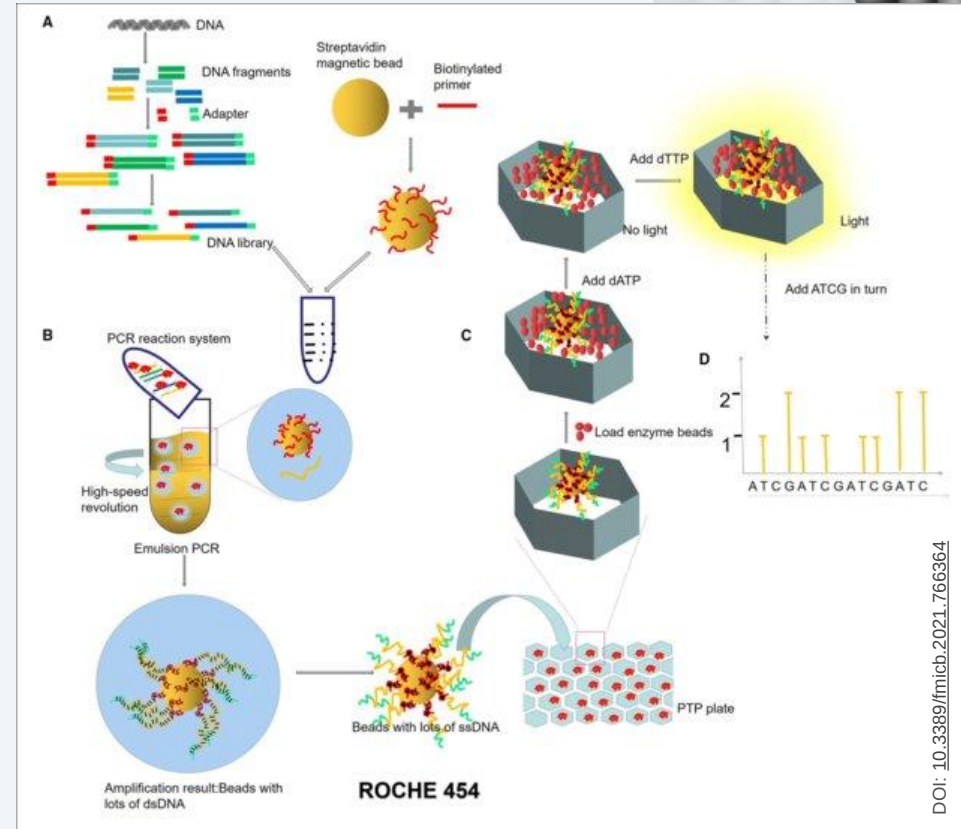


II GENERAZIONE: ROCHE 454

Sequenziamento ROCHE 454:

C. Sequenziamento del pirofosfato. In ciascuna cella è aggiunto un solo dNTP alla volta. Se corrisponde solo alla base successiva del modello di DNA, è aggiunto all'estremità 3' del primer di sequenziamento e allo stesso tempo rilascia un PPI molecolare altrimenti il processo di inserimento termina. Sotto l'azione dell'ATP solforilasi, luciferasi e apirasi, il PPI innesca una reazione di chemiluminescenza a cascata di enzimi

D. Ricezione del segnale luminoso



II GENERAZIONE: ROCHE 454

Frammentazione: un singolo filamento di DNA da sequenziare è usato come stampo su cui è sintetizzato un nuovo filamento. Si aggiunge un primer (un breve pezzo di DNA) che si lega allo stampo in una posizione specifica, marcando l'inizio del sequenziamento.

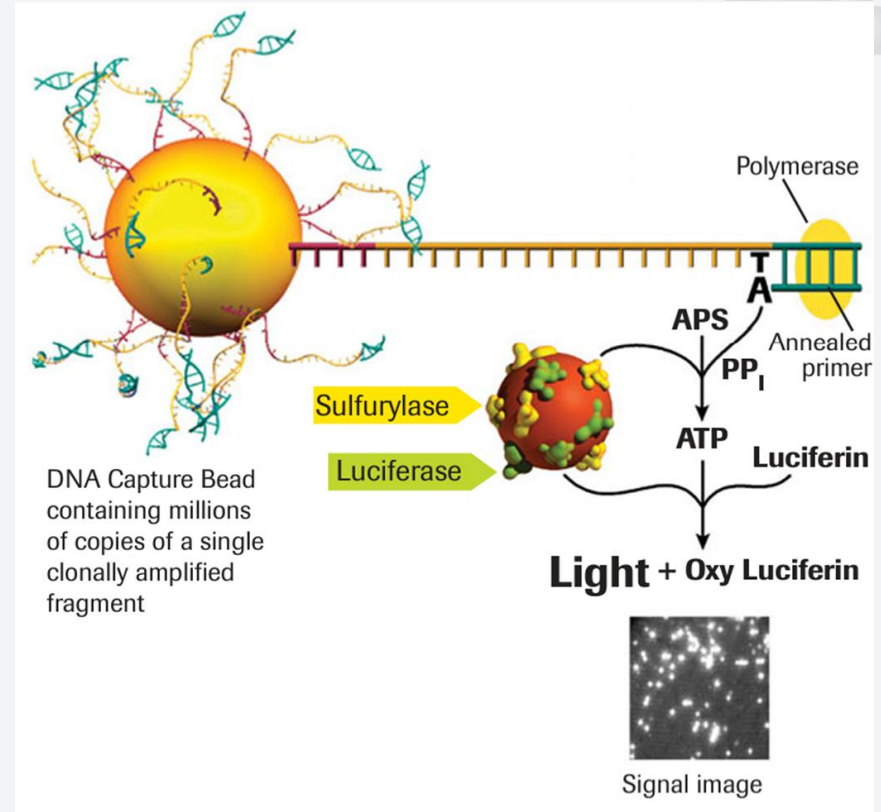
Sintesi del DNA e Rilevazione: si aggiungono nucleotidi (A, T, C o G) uno alla volta. Quando un nucleotide si lega al filamento in crescita, è rilasciato un pirofosfato (PPi).

Generazione di Luce: la presenza di PPi innesca una serie di reazioni enzimatiche che risultano nella produzione di luce visibile. Questa luce è rilevata da un sensore (un fotorecettore) e indica che il nucleotide è stato incorporato nel filamento in crescita. L'intensità della luce è proporzionale al numero di nucleotidi consecutivi incorporati (per esempio, se ci sono quattro adenine (A) consecutive nella sequenza del DNA e si aggiunge l'adenina come nucleotide, si otterrà una luce quattro volte più intensa di quella che si avrebbe con una singola adenina)

II GENERAZIONE: ROCHE 454

Lettura Sequenziale: Dopo ogni aggiunta di nucleotide, gli eventuali nucleotidi non incorporati sono lavati via ed è aggiunto il successivo nucleotide. La sequenza è determinata osservando l'ordine in cui la luce viene prodotta durante l'aggiunta di ciascun nucleotide

Analisi dei Dati: La sequenza di luce rilevata è poi tradotta quantizzata e convertita nella corrispondente sequenza di nucleotidi, che viene gestita da un elaboratore elettronico



II GENERAZIONE: ROCHE 454 PRO vs CONTRO

Read Lunghe: Uno dei principali vantaggi del Roche 454 è stata (fino al 2013) la sua capacità di produrre letture relativamente lunghe rispetto ad altre piattaforme NGS dell'epoca. Questo lo ha reso ideale per applicazioni come il sequenziamento di nuovi genomi e la risoluzione di regioni genomiche complesse.

Velocità: Rispetto ai metodi tradizionali, come il sequenziamento Sanger, il 454 è molto più veloce, consentendo il sequenziamento di interi genomi in tempi molto più brevi.

Flessibilità: Il sistema può essere utilizzato per una serie di applicazioni, dal sequenziamento del genoma completo, alla metagenomica, al sequenziamento mirato.

Senza Clonazione: A differenza del metodo Sanger, il 454 non richiede la clonazione di DNA in vettori batterici, il che riduce il tempo e la complessità del workflow.

Errori in Omopolimeri: Il metodo 454 ha difficoltà con le sequenze di omopolimeri (lungi filamenti di un unico tipo di monomero). Questo può portare a errori nel determinare il numero di basi ripetute in sequenza.

Costo per reads: il costo per singola reads è più alto rispetto ad altre piattaforme NGS, come Illumina.

Bassa Capacità: Sebbene il 454 fosse rivoluzionario al suo debutto, la sua capacità è stata superata da altre tecnologie NGS che possono produrre molti più dati per ciascuna lettura.

Fine della Produzione: Roche ha interrotto il sistema 454 nel 2013

Lunghezza Variabile delle Reads: Sebbene le reads siano generalmente lunghe, possono variare in lunghezza, il che può rendere alcune applicazioni, come l'assemblaggio del genoma, più complesse



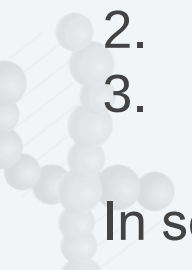
II GENERAZIONE: ILLUMINA

La tecnologia di **sequenziamento Illumina** è attualmente una delle più utilizzate al mondo per il sequenziamento di DNA ad alto rendimento (*throughput*) cioè la quantità di molecole di DNA lette contemporaneamente

La piattaforma Illumina si basa su un processo chiamato **sequenziamento di sintesi** (*sequencing-by-synthesis*).

Di seguito descrivo le fasi principali:

1. Preparazione della libreria dei frammenti di nucleotidi
2. Amplificazione bridge
3. Sequenziamento

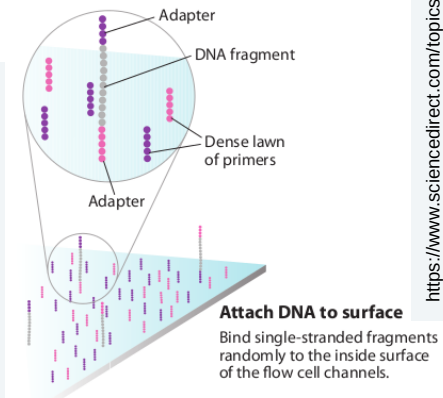
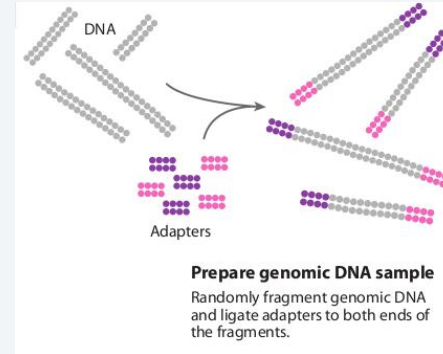


In seguito si svolge l'analisi dei dati

II GENERAZIONE: ILLUMINA

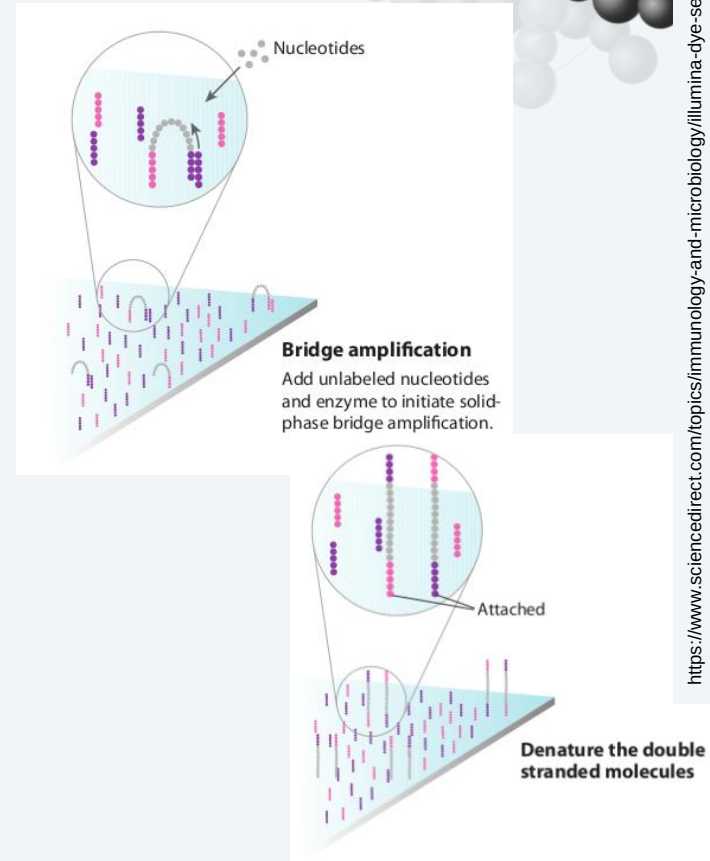
Il sequenziamento Illumina si articola in più passi:

1. Preparazione della Libreria: Il DNA di interesse è frammentato in piccoli pezzi e ad ogni frammento sono legati degli adattatori specifici. Questi adattatori servono per immobilizzare i frammenti su una superficie solida e per fornire siti di legame per i primer utilizzati nel sequenziamento



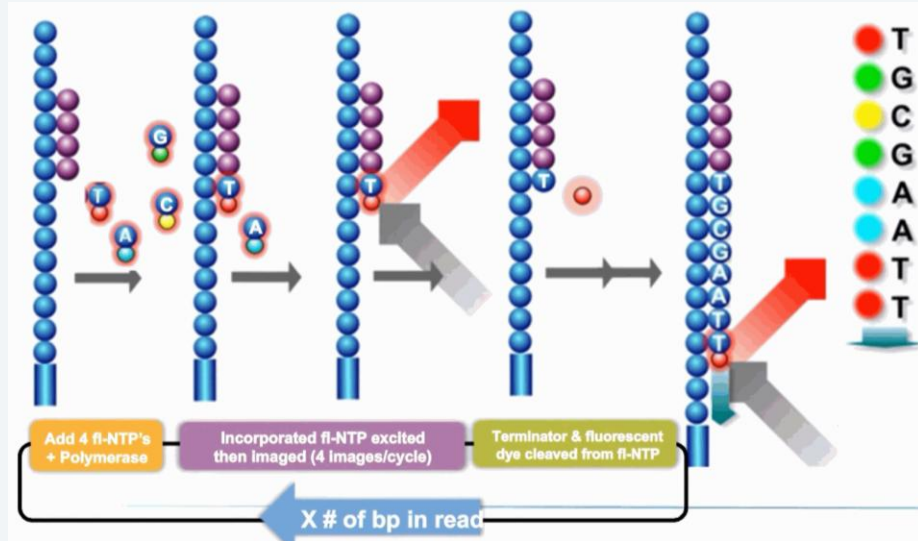
II GENERAZIONE: ILLUMINA

2. **Amplificazione di ponte:** I frammenti di DNA preparati sono caricati su una cella di flusso e poi riversati su una superficie rivestita con degli oligonucleotidi complementari agli adattatori. Una volta che i frammenti si sono legati alla superficie, iniziano a formare "ponti", piegandosi e legandosi ai siti vicini, creando molteplici copie del medesimo frammento. Questo processo è chiamato **amplificazione di ponte** (*bridge amplification*) e genera, dopo un **processo di dissociazione**, cluster di copie identiche del frammento di DNA che sono in seguito denaturate/separate



II GENERAZIONE: ILLUMINA

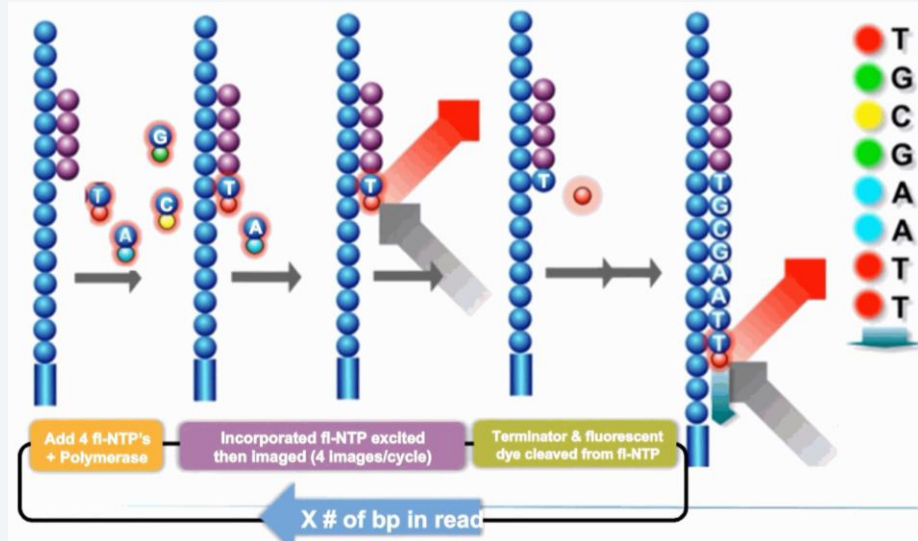
3. **Sequenziamento:** Una volta formati i cluster, inizia il sequenziamento vero e proprio. Si tratta di un sequenziamento "sintetico" basato sull'aggiunta sequenziale di nucleotidi fluorescentemente marcati. Ad ogni ciclo, è aggiunto un tipo di nucleotide marcato. Se il nucleotide è complementare al prossimo nucleotide del filamento stampo, è incorporato dal polimerasi altrimenti è espulso



II GENERAZIONE: ILLUMINA

Una volta legato il nucleotide con marcatore fluorescente, il filamento è sottoposta a scansione da un sensore ottico (simile ad un CMOS presente nelle videocamere digitali) che registra la fluorescenza emessa da ogni cluster.

Dopo ogni ciclo di lettura, la fluorescenza del nucleotide marcato è rimosso, e il ciclo ripete con il nucleotide successivo.





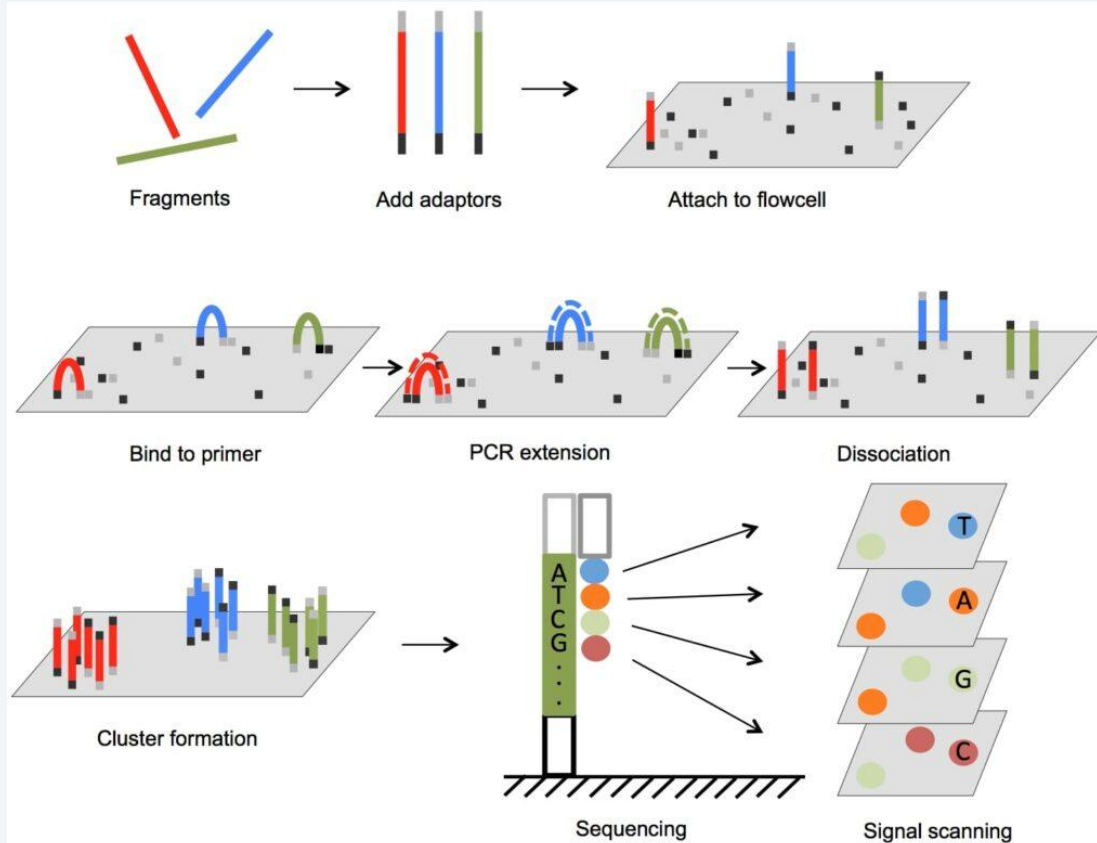
II GENERAZIONE: ILLUMINA

Alla fine del processo di sequenziamento, si ottiene una serie di "read" o letture, che rappresentano le **sequenze di DNA dei frammenti originali**

Utilizzando software specializzati, queste letture possono essere allineate, assemblate e analizzate per ottenere informazioni sul DNA di partenza.

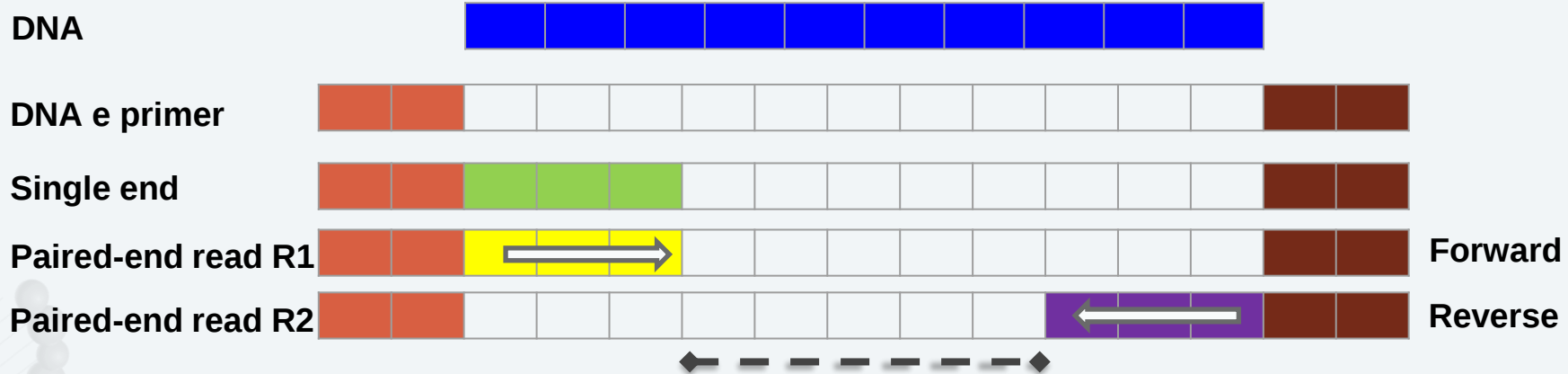


II GENERAZIONE: ILLUMINA



II GENERAZIONE: ILLUMINA

Il sequenziamento può essere fatto a partire da una sola estremità del frammento (sequenziamento con **single-end reads**) o partendo da entrambe le estremità e proseguendo in direzioni opposte (sequenziamento con **paired-end reads**).



L'inner mate distance è la lunghezza della sequenza interposta tra R1 ed R2. (Ad esempio, in un sistema nel quale si utilizzi una sequencing library con frammenti da 500bp e nel quale si producano reads di 100bp, la distanza intermedia fra R1 e R2 sarà di circa 300bp.) Grazie alla inner mate distance è possibile individuare mutazione genomiche. Quali?

II GENERAZIONE: ILLUMINA

Alto throughput: può sequenziare miliardi di basi in un singolo passaggio di lettura

Precisione: ha un tasso di errore molto basso

Flessibilità: può essere utilizzato per varie applicazioni, dal sequenziamento del genoma intero a panel di geni mirati

La lunghezza dei read è tipicamente più corta rispetto ad altre tecniche come il sequenziamento Sanger

Il sequenziamento Illumina ha rivoluzionato la genetica e la genomica, rendendo possibile sequenziare interi genomi in modo rapido e conveniente e contribuendo significativamente all'era della genomica personale e della medicina personalizzata.



II GENERAZIONE: SOLID

Il **sequenziamento SOLiD** (*Sequencing by Oligonucleotide Ligation and Detection*) è una tecnologia di sequenziamento sviluppata da Applied Biosystems. Il processo si basa sulla ligazione di oligonucleotidi marcati con fluorocromi per ottenere un'elevata accuratezza nel sequenziamento.

Una reazione di ligazione è un processo biochimico in cui due frammenti di DNA sono uniti insieme. Questo processo è mediato da un enzima chiamato ligasi, che facilita la formazione di un legame covalente tra i gruppi fosfato e i gruppi ossidrilici dei frammenti di DNA, risultando in un singolo pezzo di DNA.

Nel **sequenziamento SOLiD** la **reazione di ligazione è utilizzata per attaccare sonde fluorescenti specifiche al DNA** durante il processo di sequenziamento.

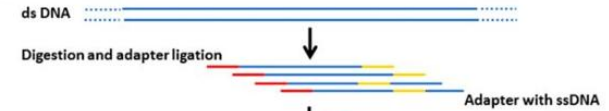


II GENERAZIONE: SOLID

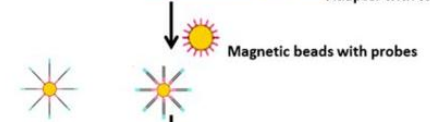
Il processo SOLID è costituito da cinque fasi fondamentali:

1 Preparazione della libreria: La sequenza di DNA da analizzare è frammentata in piccoli pezzi e legata a dei primer specifici.

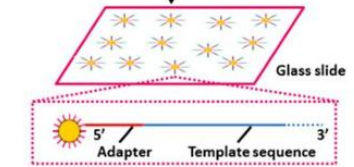
(1) Library preparation



(2) Emulsion PCR



(3) Bead deposition



(4) Sequence by ligation

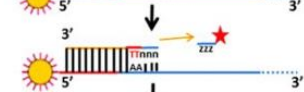
(i) Universal primer (n) hybridization



(ii) Specific de-base probe ligation and fluorescence detection



(iii) Cleave off fluor



(iv) Repeat steps (i) – (iv)



(v) Probe reset and universal primer (n - 1) hybridization



(vi) Repeat steps (i) – (iv) with new probe

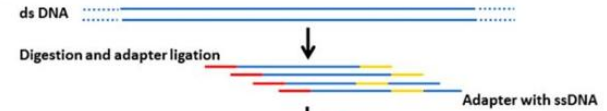
(vii) Repeat reset with, n - 2, n - 3, n - 4 universal primers



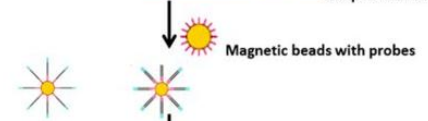
II GENERAZIONE: SOLID

2 PCR in emulsione: Ogni frammento è legato a delle **sfere magnetiche** (*beads*), su cui avviene un'amplificazione clonale tramite emPCR (emulsione PCR). Questo permette di ottenere una serie di copie identiche di ciascun frammento su ciascuna sfera.

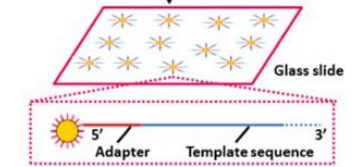
(1) Library preparation



(2) Emulsion PCR

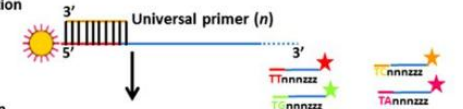


(3) Bead deposition



(4) Sequence by ligation

(i) Universal primer (n) hybridization



(ii) Specific de-base probe ligation and fluorescence detection



(iii) Cleave off fluor



(iv) Repeat steps (i) – (iv)



(v) Probe reset and universal primer ($n - 1$) hybridization



(vi) Repeat steps (i) – (iv) with new probe

(vii) Repeat reset with, $n - 2$, $n - 3$, $n - 4$ universal primers



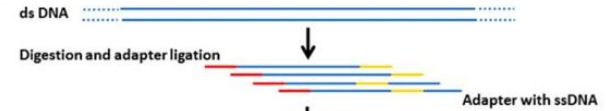
II GENERAZIONE: SOLID

3 Deposizione delle beads:

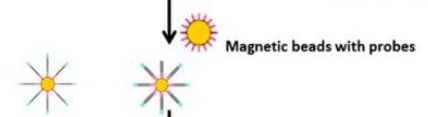
Le beads, ciascuna con molteplici copie dello stesso frammento di DNA, sono depositate su un supporto solido (come un vetrino).

A questo punto, ogni bead rappresenta una sequenza specifica che può essere letta simultaneamente insieme alle altre, rendendo il processo molto efficiente per il sequenziamento parallelo.

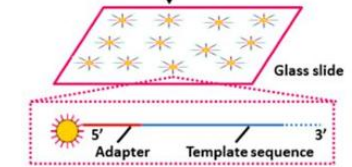
(1) Library preparation



(2) Emulsion PCR



(3) Bead deposition



(4) Sequence by ligation

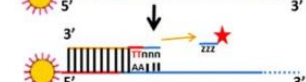
(i) Universal primer (n) hybridization



(ii) Specific de-base probe ligation and fluorescence detection



(iii) Cleave off fluor



(iv) Repeat steps (i) – (iv)



(v) Probe reset and universal primer (n - 1) hybridization



(vi) Repeat steps (i) – (iv) with new probe

(vii) Repeat reset with, n - 2, n - 3, n - 4 universal primers



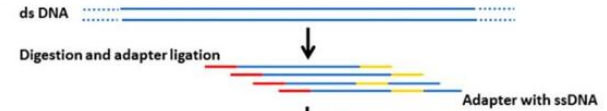
II GENERAZIONE: SOLID

4 Sequenziamento tramite ligazione:

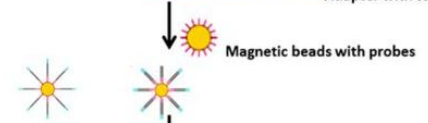
La sequenza di DNA è letta utilizzando una serie di oligonucleotidi marcati con fluorofori.

La ligazione avviene tra oligonucleotidi complementari alla sequenza di DNA bersaglio e un fluoroforo associato segnala la presenza del legame.

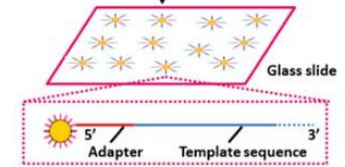
(1) Library preparation



(2) Emulsion PCR

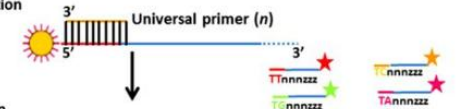


(3) Bead deposition



(4) Sequence by ligation

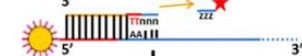
(i) Universal primer (n) hybridization



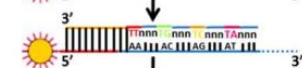
(ii) Specific de-base probe ligation and fluorescence detection



(iii) Cleave off fluor



(iv) Repeat steps (i) – (iv)



(v) Probe reset and universal primer (n - 1) hybridization



(vi) Repeat steps (i) – (iv) with new probe

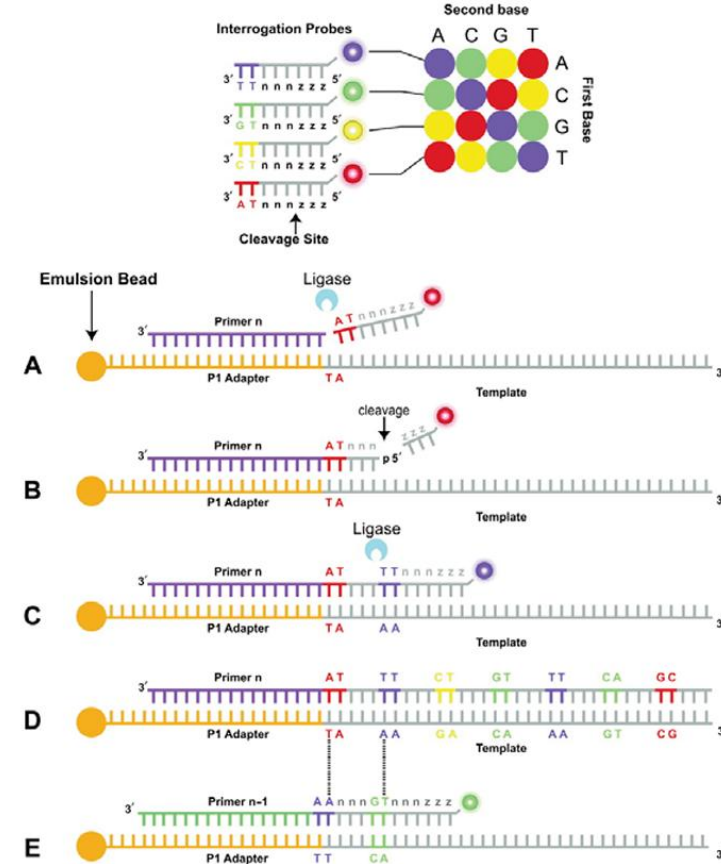
(vii) Repeat reset with, n - 2, n - 3, n - 4 universal primers



II GENERAZIONE: SOLID

4 Sequenziamento tramite ligazione:

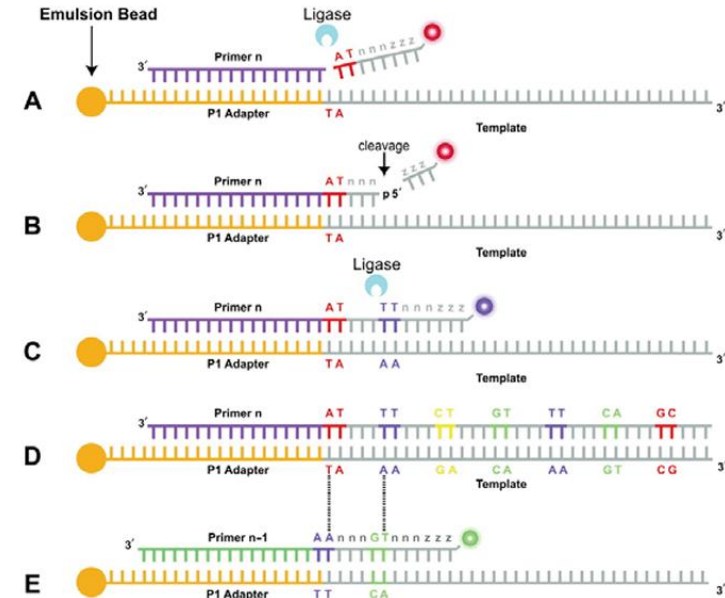
Ogni sonda oligonucleotidica è progettata per riconoscere una coppia specifica di basi e possiede un colorante fluorescente unico. Durante il processo di sequenziamento, le sonde competono per legarsi al frammento di DNA target. Quando una sonda si lega, emette una fluorescenza che è rilevata e registrata. Poiché ogni coppia di basi ha un colore unico, la sequenza del DNA è determinata in relazione alla serie di colori rilevati.



II GENERAZIONE: SOLiD

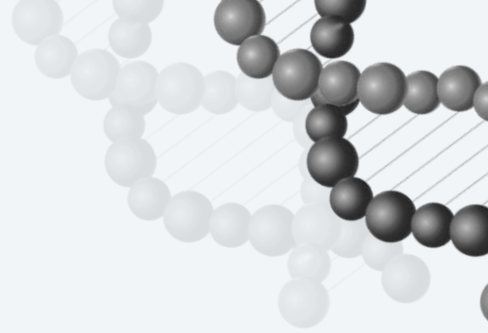
Nel sequenziamento tradizionale, una base è letta alla volta, e ciascuna (A, T, C o G) è identificata da una specifica fluorescenza o un segnale. Nel sistema SOLiD, invece, sono identificate coppie di basi consecutive

La codifica a 2 basi (2 *Base Encoding*) implica una lettura doppia della base nel frammento di DNA target (una volta in ogni coppia). Ciò aumenta l'accuratezza del sequenziamento poiché eventuali errori in una lettura possono essere corretti o confermati dalla lettura successiva



II GENERAZIONE: SOLID

Poiché ci sono quattro basi nel DNA (A, T, C, G), ci sono 16 possibili combinazioni di coppie di basi (per esempio, AA, AT, AC, AG, TA, TT, ecc.). Ogni combinazione ha un colore associato in relazione alla sonda che la riconosce.

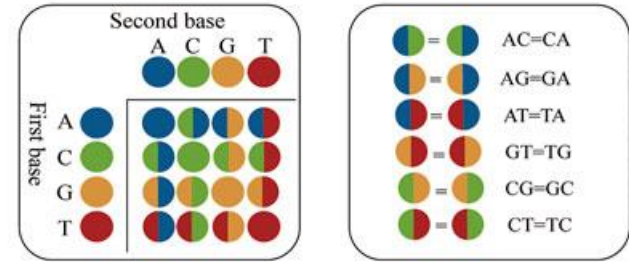


II GENERAZIONE: SOLID

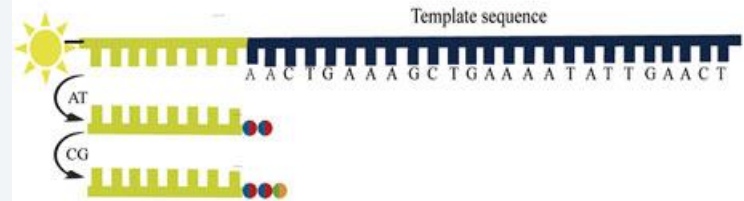
5 Lettura dati: I fluorofori emettono segnali che vengono rilevati da un sistema ottico. I colori identificano combinazioni specifiche di basi adiacenti. Questi segnali sono poi tradotti in sequenze nucleotidiche. La lettura è ripetuta più volte con uno spostamento della posizione di inizio, permettendo di "decodificare" ogni singola base con maggiore accuratezza.

Infine, l'algoritmo bioinformatico ricostruisce la sequenza completa basandosi sulle combinazioni lette, fornendo la sequenza finale del frammento.

Dual-mononucleotide scheme



Sequencing



Decoding



II GENERAZIONE: SOLID

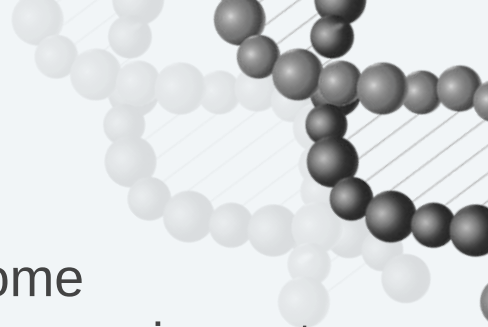
Questa tecnologia di nuova generazione produce 10^8 - 10^9 reads di breve lunghezza contemporaneamente. Quindi è capace di generare milioni di sequenze brevi in un singolo passaggio, rendendo la tecnica molto efficiente e ad alto rendimento.



II GENERAZIONE: ION TORRENT

Il **sequenziamento Ion Torrent**, conosciuto anche come sequenziamento a semiconduttore, è un metodo di sequenziamento del DNA che non fa uso di marcatori fluorescenti, a differenza di molte altre tecniche di sequenziamento di nuova generazione.

Questa tecnica di amplificazione si basa sulla **rilevazione di ioni idrogeno rilasciati durante la sintesi del DNA**



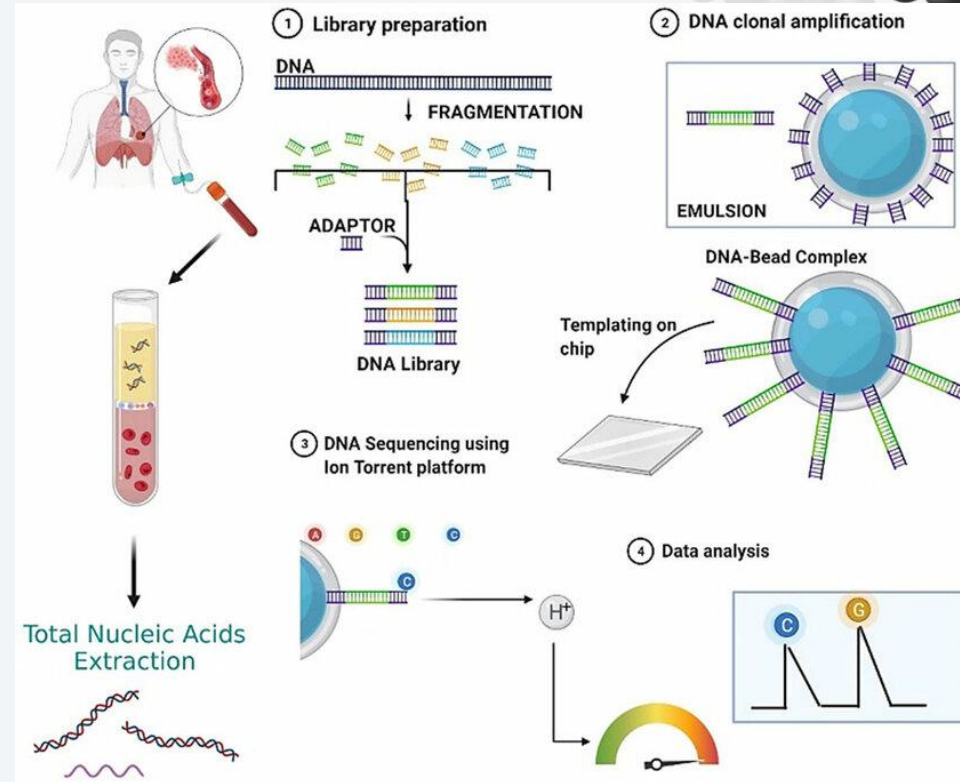
II GENERAZIONE: ION TORRENT

1. Preparazione della Libreria:

si prepara con frammenti di DNA a cui si uniscono adattatori. Ogni frammento è legato a una perlina (bead) in un processo noto come **emulsione PCR**.

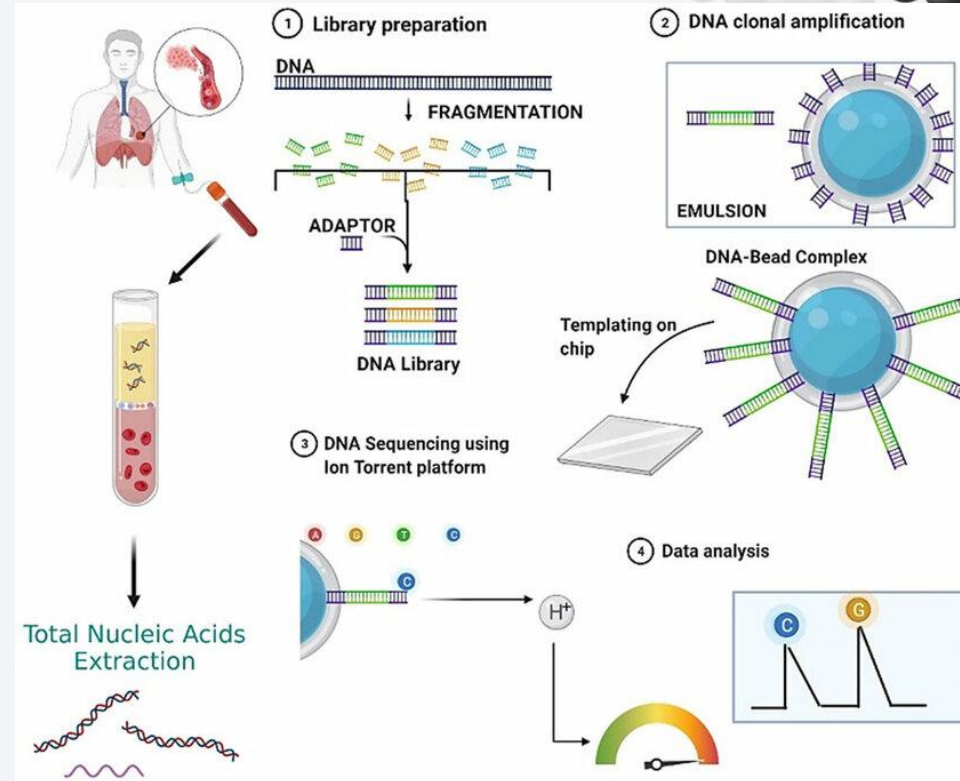
2. Deposito delle perline:

Ogni perlina avente una singola molecola di DNA replicata più volte è depositata su una piastra a pozzetti (*chip*)



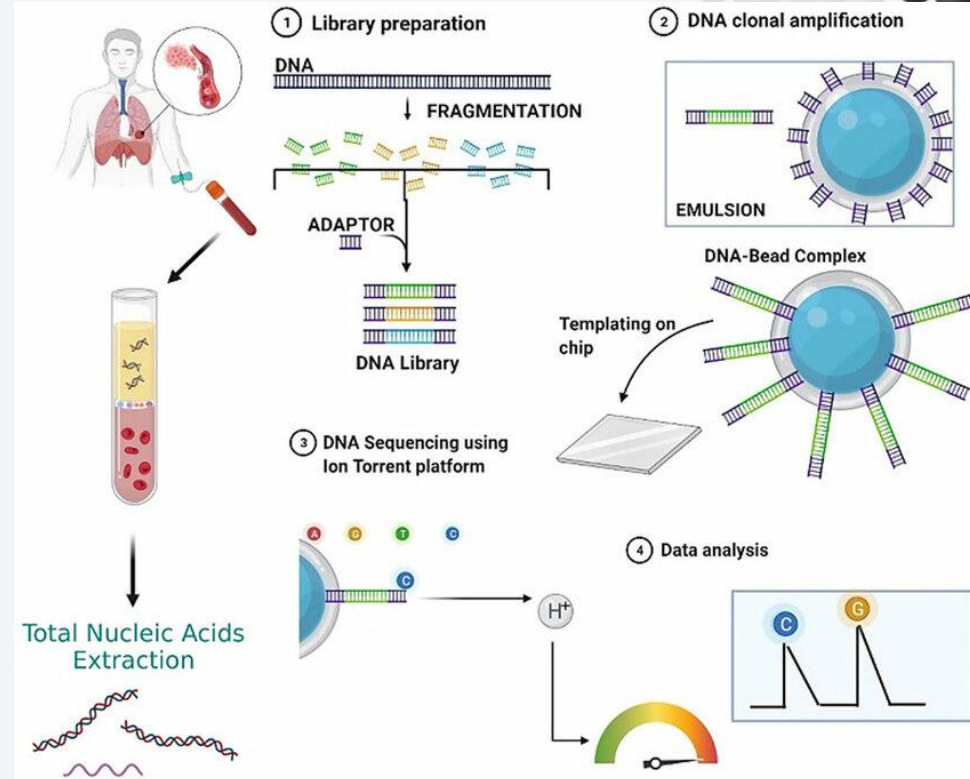
II GENERAZIONE: ION TORRENT

3. Sintesi del DNA e Rilevazione: Si aggiungono nucleotidi uno alla volta in ogni cella del chip. Quando un nucleotide si incorpora in una catena di DNA in crescita, rilascia un protone (ione idrogeno). Questo rilascio provoca una piccola variazione del pH nell'ambiente circostante



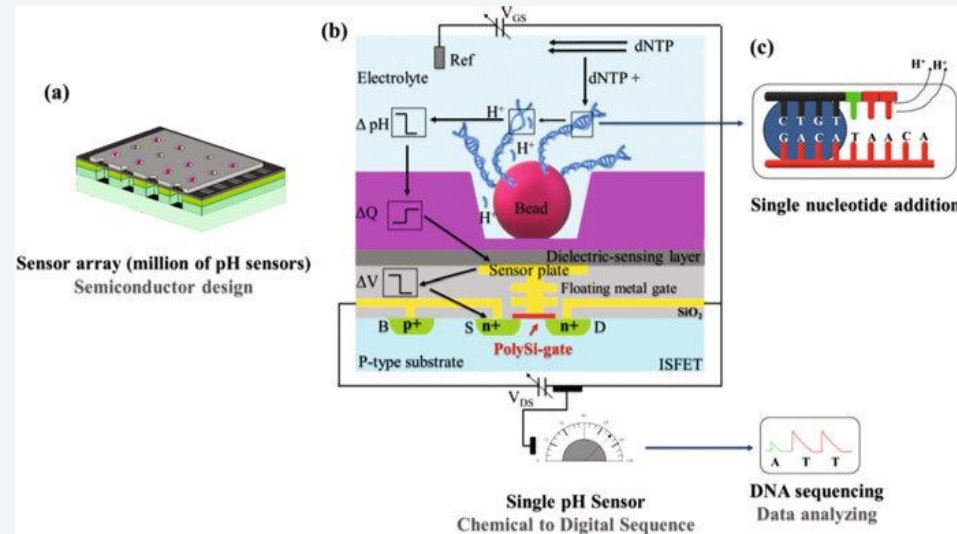
II GENERAZIONE: ION TORRENT

4. Rilevazione con Chip a Semiconduttore: Il chip (un semiconduttore CMOS) rileva i cambiamenti di pH (che corrispondono al rilascio di ioni idrogeno) e li traduce in un segnale elettrico. Questo segnale indica che un particolare nucleotide è stato incorporato nel frammento di DNA in quel pozzetto.



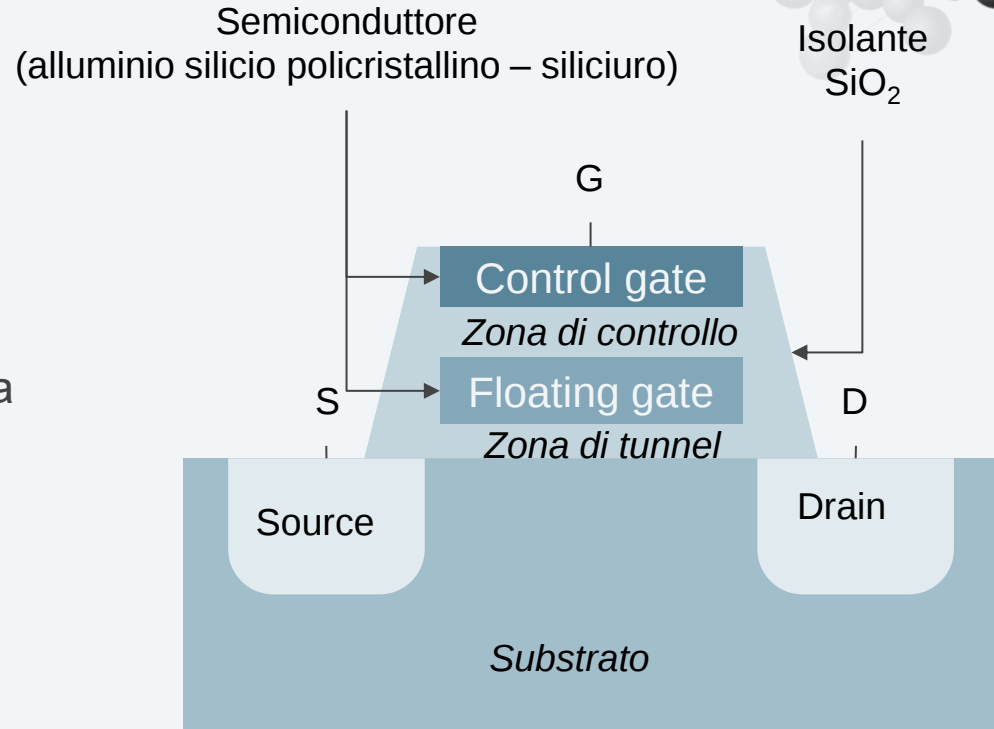
II GENERAZIONE: ION TORRENT

5. Traduzione in Sequenza: Ripetendo il processo di aggiunta di nucleotidi e rilevazione del segnale elettrico, si ottiene una sequenza di segnali che può essere tradotta in una sequenza di nucleotidi, fornendo la sequenza del frammento di DNA



L'informazione è custodita in una zona (la griglia flottante, *floating gate*) racchiusa entro due strati coperti da materiale isolante

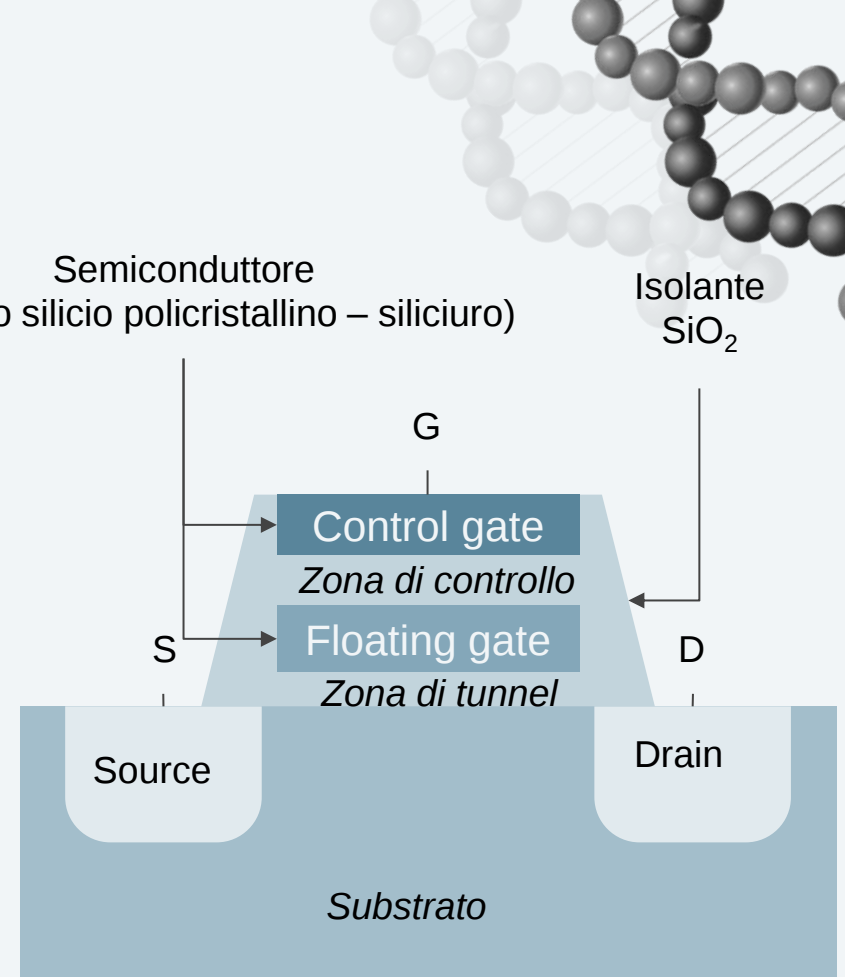
Tra la
zona
cunic
parte



Floating Gate MOSFET

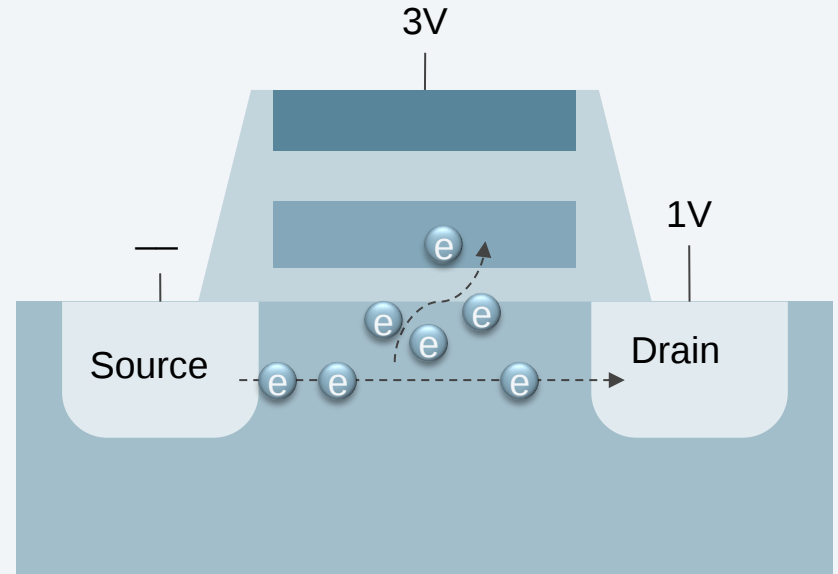
Il chip è formato da una matrice di **transistori a griglia fluttuante** (*floating-gate MOSFET*) (alluminio silicio policristallino – siliciuro)

La presenza di ioni crea un differenziale di potenziale sull'elettrodo denominato CONTROL GATE e grazie a quello si riesce a stabilire, analizzando la corrente risultante, il tipo di nucleotide individuato



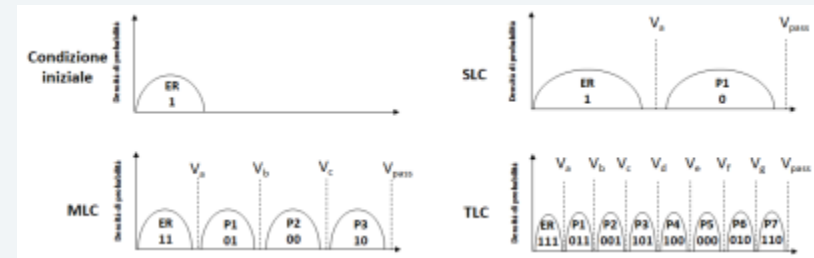
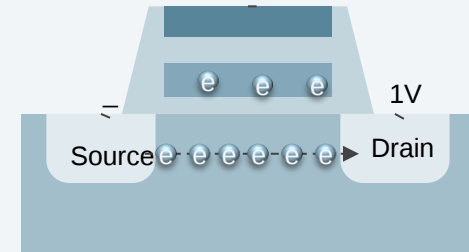
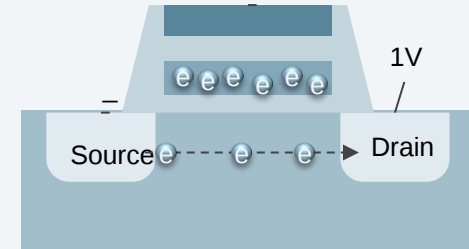
Floating Gate MOSFET: Scrittura

- ❑ In scrittura se si applica un voltaggio positivo al drain (es: 1V) si inducono gli elettroni nel settore della sorgente a muoversi verso la sua direzione
- ❑ Nel contempo se si applica un voltaggio più alto (es: 3V) al control gate che consente il passaggio degli elettroni in movimento a rimanere intrappolati nell'area flottante



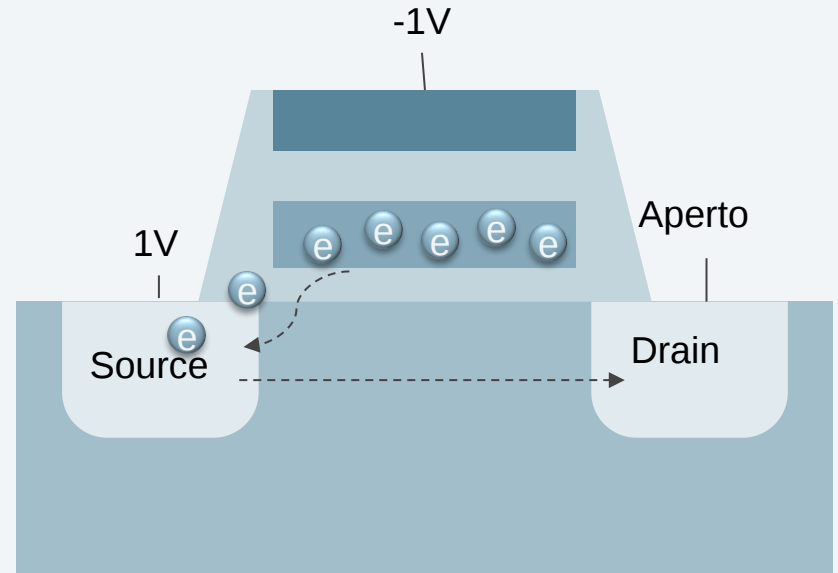
Floating Gate MOSFET: Lettura

- ❑ In lettura si applica un voltaggio di 1V al drain che induce gli elettroni nel settore della sorgente a muoversi verso la sua direzione
- ❑ Il passaggio degli elettroni in movimento risulta più o meno fluido in relazione al numero di elettroni intrappolati nell'area flottante. Maggiori sono gli elettroni imprigionati superiore sarà l'ostacolo degli elettroni mobili ad attraversare la base (stato logico 0), viceversa il flusso di elettroni individua un 1
- ❑ In base al numero di elettroni presenti nel floating gate si può determinare una variazione di tensione di uscita consentendo, per una cella di memoria, di archiviare più valori (es: 4 per A,T,G e C)



Floating Gate MOSFET: Cancellazione

- ❑ In cancellazione si applica un voltaggio di $-1V$ al control gate e uno di $1V$ alla sorgente. Questo induce gli elettroni imprigionati a rientrare nella sorgente
- ❑ La liberazione della zona flottante riporta la cella allo stato primordiale

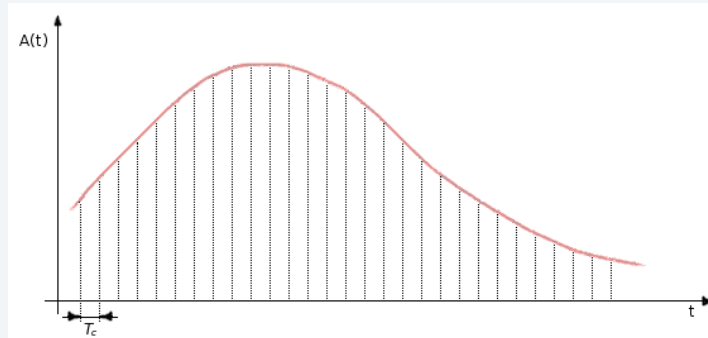


Dal segnale analogico al digitale

- ❑ Una volta che il segnale elettrico è stato ricavato bisogna riconoscere le basi
- ❑ Il processo di conversione di un segnale analogico in uno digitale avviene tramite due fasi di discretizzazione: il **campionamento** (discretizzazione temporale), in cui si stabiliscono e si individuano gli intervalli temporali in cui dovrà essere rilevato il segnale analogico; e la **quantizzazione** (discretizzazione dei livelli) cioè la trasformazione in valori numerici del segnale analogico nell'istante di rilevamento. Infine ci sarà una fase di codifica che permette di ottenere il segnale nella forma numerica finale (ovviamente nel caso di riproduzione del segnale si seguirà il procedimento inverso, trasformando il valore numerico discreto in segnale analogico).
- ❑ A questo ultimo aspetto provvedono due circuiti che possono avere caratteristiche architetture più o meno raffinate con un diverso tipo di accuratezza: lo ADC (Convertitore Analogico Digitale - *Analogic to Digital Converter*) che trasforma i segnali analogici in digitali e il DAC (Convertitore Digitale Analogico - *Digital to Analogic Converter*) che muta i segnali digitali in analogici

Passo di campionamento

- ❑ Con il **processo di campionamento** si considerano solamente i valori che il segnale ha in istanti di tempo prefissati. L'andamento del segnale nei rimanenti punti può essere ricostruito a partire dai valori conosciuti, a patto che questi ultimi corrispondano ad un numero finito di istanti sufficientemente vicini tra loro. Altrettanto intuitivamente si può capire che se il segnale varia rapidamente i valori rilevati devono essere più vicini tra loro, così che tra un campione all'altro non vi sia una insufficiente descrizione delle variazioni imprevedibili.
- ❑ La distanza tra due successivi valori noti del segnale è solitamente detta **passo o periodo di campionamento** (T_c) ed i valori assunti dal segnale negli istanti noti sono i campioni del segnale



Teoria del campionamento

La rapidità con cui un segnale varia è legata al suo contenuto in frequenza; per questo, la scelta del passo di campionamento deve essere legata alla banda di frequenze del segnale.

Il valore del periodo di campionamento, T_c , viene scelto in relazione al **Teorema del Campionamento (o Teorema di Shannon)**. Il teorema afferma che il numero dei campioni nell'unità di tempo deve essere almeno pari al doppio della frequenza massima del segnale che si vuol produrre.

$$T_c = \frac{1}{2f_{max}}$$

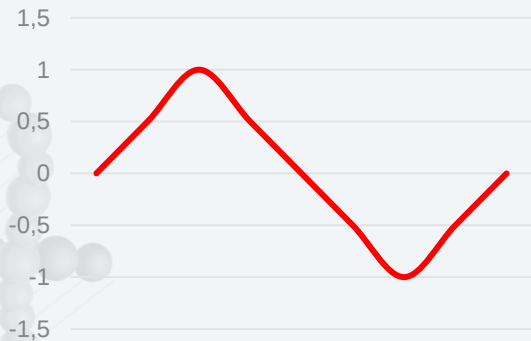
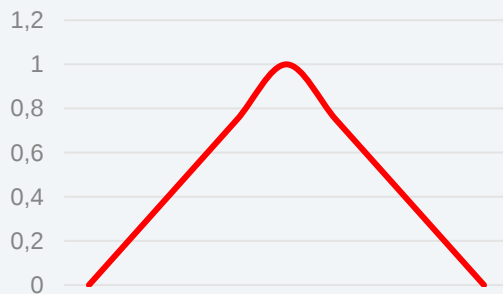
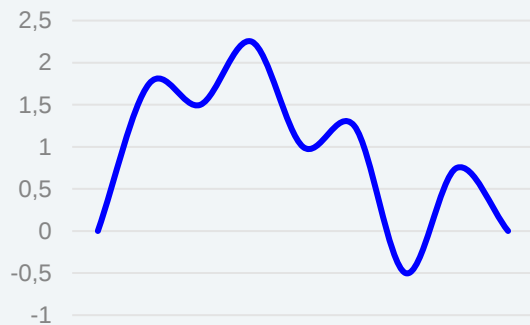
Teoria del campionamento

Per trovare la frequenza massima si usa la trasformata di Fourier
In seguito se ne descriverà l'uso con un esempio

$$f(x) = \frac{a_0}{2} + \sum_{k=1} [a_k \cos(2\pi \frac{k}{\tau} x) + b_k \sin(2\pi \frac{k}{\tau} x)]$$

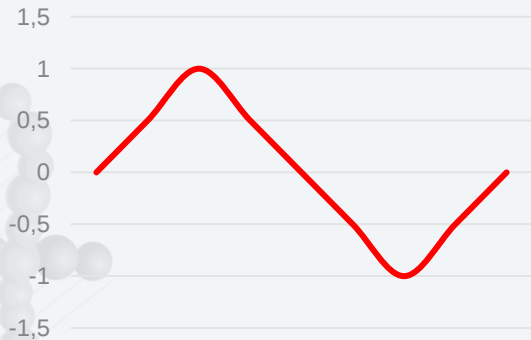
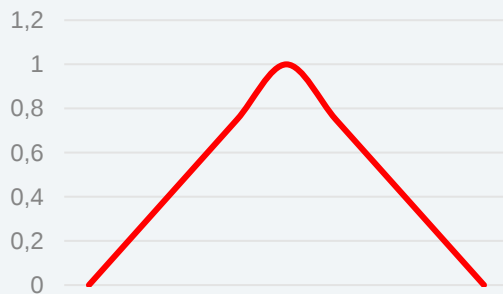
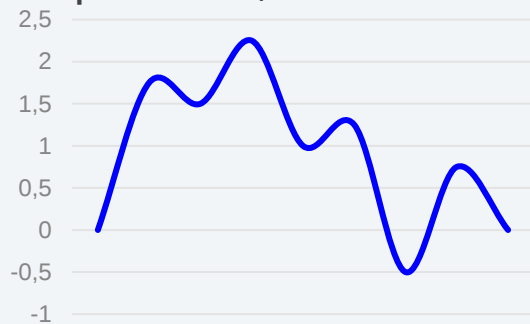
Teoria del campionamento

La curva può essere scomposta grazie al teorema di Fourier nelle successive (cioè la curva iniziale è data dalla somma delle curve derivate)



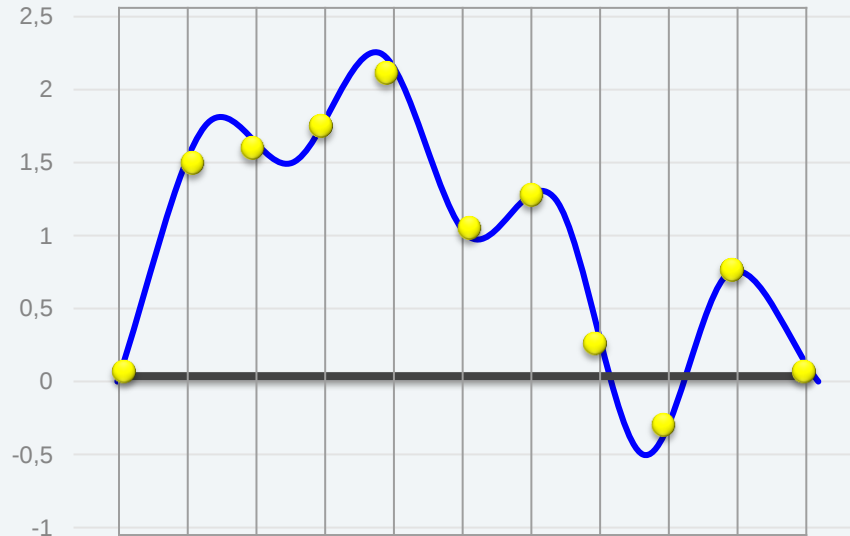
Teoria del campionamento

Supponendo che il campione sia stato analizzato nell'intervallo di tempo $[0,1]$ le sue curve derivate hanno frequenza 1, 1 e 4.



Teoria del campionamento

Quindi per il Teorema del campionamento si può pensare ad un $T_c > (1/2 \cdot 4)$.
Nell'esempio si preleva un campione ogni decimo di secondo (anche $1/9$ sarebbe andato bene)



Es: 0; 1.5;1.6;1.75; 2.2;1;1.6;0.5;-0.4;0.75;0

Teoria del campionamento⁸³

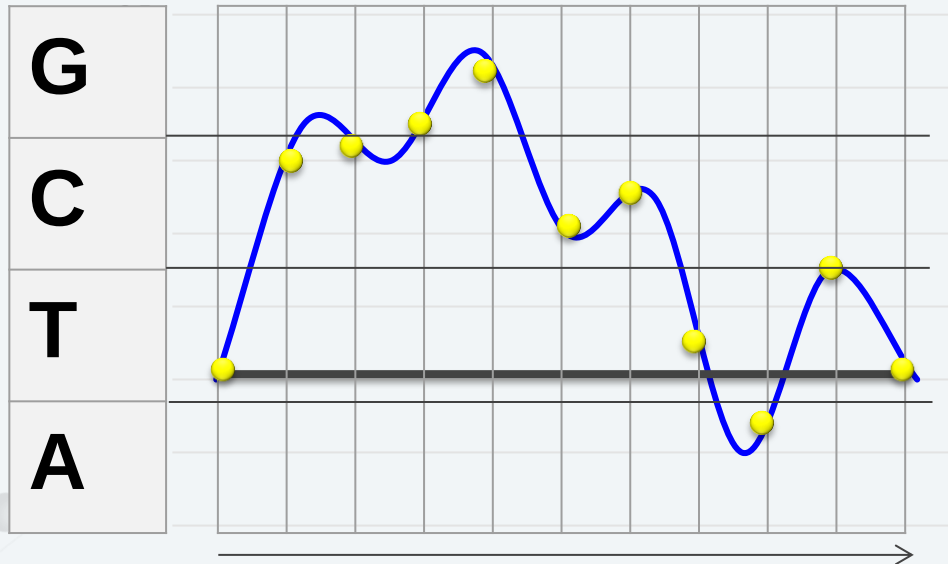
Il campionamento genera un segnale tempo discreto che può assumere valori in un insieme continuo (dei numeri reali)

Affinché il segnale possa essere elaborato con un sistema numerico è necessario allora discretizzare tali valori.

A tal fine essi sono raggruppati in un determinato numero di insiemi, chiamati celle o intervalli di quantizzazione, ognuno dei quali è caratterizzato da un singolo valore, detto **livello di restituzione**. Una volta individuato per ogni valore del segnale di ingresso la cella di appartenenza, si può associare ad esso il livello di restituzione relativo

Teoria del campionamento

Avendo 4 caratteri (A,T,C,G) i può stabilire un intervallo dell'ampiezza diviso per quattro (nel caso si sia a cavallo della linea di demarcazione si sceglie la lettera in basso)



0	1.5	1.6	1.75	2.2	1	1.6	0.5	-0.4	0.75	0
T	C	C	G	G	C	C	T	A	T	T

II GENERAZIONE: ION TORRENT

Rapidità: Una delle principali attrattive del sequenziamento Ion Torrent è la sua velocità. Può produrre risultati in poche ore, a differenza di altri metodi che possono richiedere giorni.

Semplicità del Rilevamento: Poiché il sistema rileva ioni idrogeno (cambiamenti di pH) invece di segnali fluorescenti, non sono necessari laser o camere specializzate, il che semplifica la strumentazione.

Costi: L'assenza di reagenti fluorescenti e la semplificazione della strumentazione possono portare a una riduzione dei costi operativi.

Flessibilità: Il sistema Ion Torrent offre flessibilità nella dimensione dei chip, consentendo agli utenti di scegliere la capacità di sequenziamento più appropriata per il loro progetto.

Lunghezza delle Reads: Sebbene la lunghezza delle reads dello Ion Torrent sia superiore a quella di alcune piattaforme, come la tecnologia Illumina, è ancora inferiore rispetto ad altre tecnologie, come il sequenziamento di terza generazione offerto da PacBio.

Errori di Omissione: Una delle principali limitazioni del sequenziamento Ion Torrent è la tendenza agli errori in regioni di omopolimeri (cioè sequenze in cui uno stesso nucleotide si ripete molte volte consecutivamente).

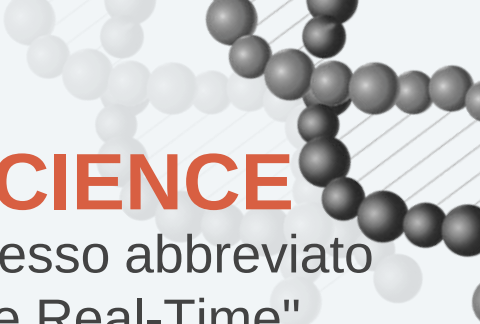
Capacità di Sequenziamento: Anche se i chip più grandi possono sequenziare grandi quantità di DNA, la capacità complessiva è generalmente inferiore rispetto a piattaforme come Illumina.

Affidabilità: si ha rumore tra celle del chip adiacenti (fenomeno noto anche in campo elettrico).



04

SEQUENZIAMENTO III GENERAZIONE



III GENERAZIONE: PACIFIC BIOSCIENCE

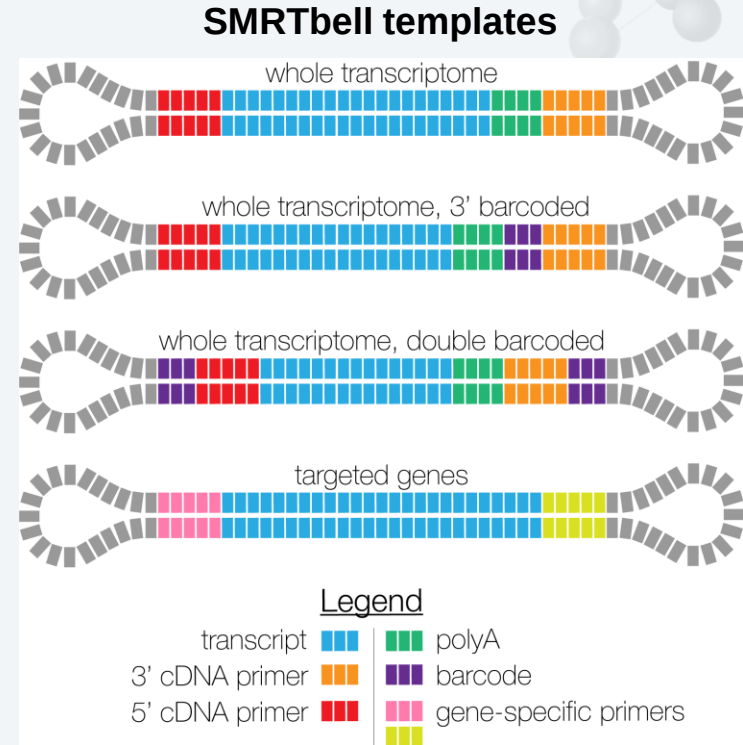
La piattaforma di sequencing **Pacific Biosciences** (spesso abbreviato come PacBio) si basa sulla tecnologia "Single Molecule Real-Time" (SMRT) sequencing.

E' il metodo di sequenziamento di terza generazione più utilizzato, principalmente per il fatto che le letture ottenute con essa sono molto lunghe, anche superiori a 20.000 bp (con una sola lettura (o corsa di sequenziamento, run) è possibile sequenziare un intero genoma batterico).



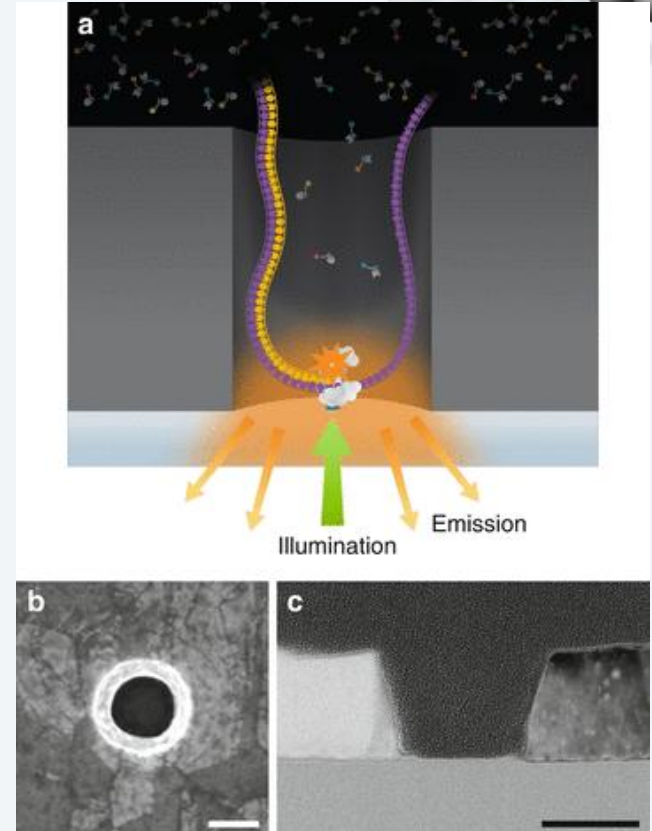
III GENERAZIONE: PACIFIC BIOSCIENCE

1. **Estrazione del DNA:** Il DNA è estratto dalla cellula e purificato per garantire che non vi siano contaminanti.
2. **Preparazione della libreria:** Le molecole di DNA sono legate a molecole adattatrici che permettono di creare dei SMRTbell templates. Questi sono costituiti da frammenti di DNA a doppio filamento con estremità circolari. Le estremità circolari permettono che il DNA sia letto ripetutamente, aumentando l'accuratezza attraverso il cosiddetto approccio di "consensus sequencing".



III GENERAZIONE: PACIFIC BIOSCIENCE

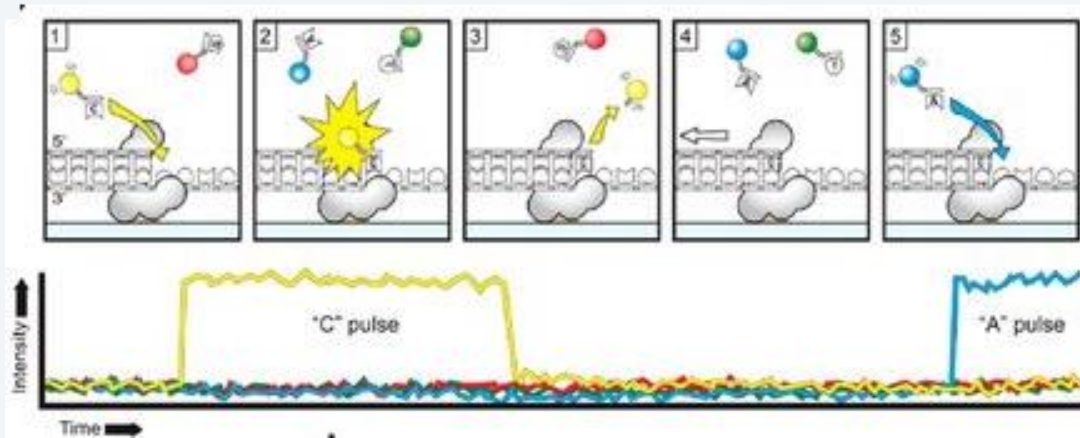
3. **Sequenziamento nel SMRT Cell.** Il cuore della tecnologia PacBio è il chip SMRT Cell, che contiene migliaia di piccoli pozzi chiamati Zero-Mode Waveguides (ZMWs). Questi pozzi consentono di osservare l'attività di una singola molecola in tempo reale
4. **Legame del complesso DNA-polimerasi:** Ogni SMRTbell template è associato a una DNA polimerasi. Questo complesso è introdotto in un ZMW



III GENERAZIONE: PACIFIC BIOSCIENCE

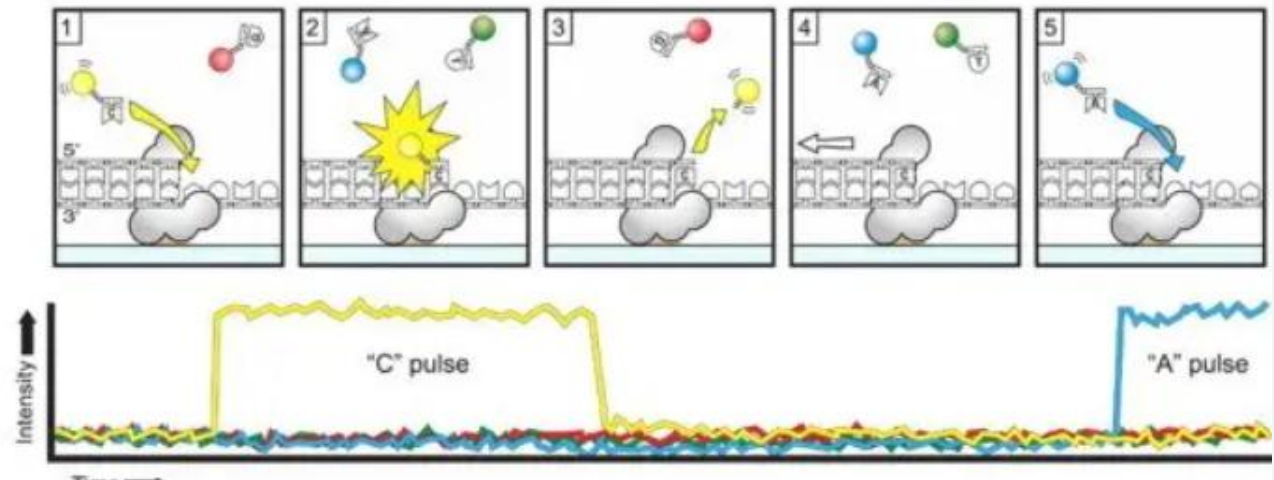
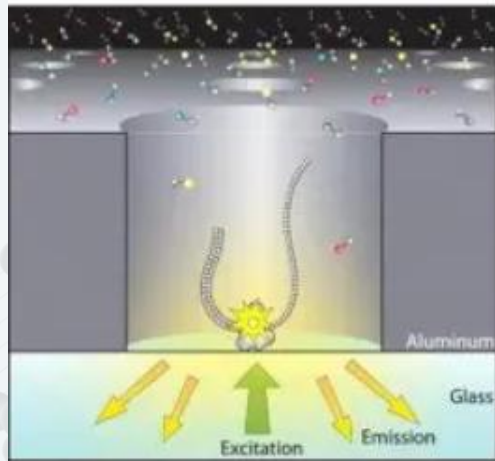
5. **Incorporazione di nucleotidi fluorescenti** Durante il sequenziamento, la polimerasi sintetizza il DNA complementare utilizzando nucleotidi marcati con fluorofori specifici per ogni base (A, T, G, C).

Quando un nucleotide viene incorporato, il fluoroforo emette un segnale luminoso (fluorescenza) che è rilevato dalla macchina

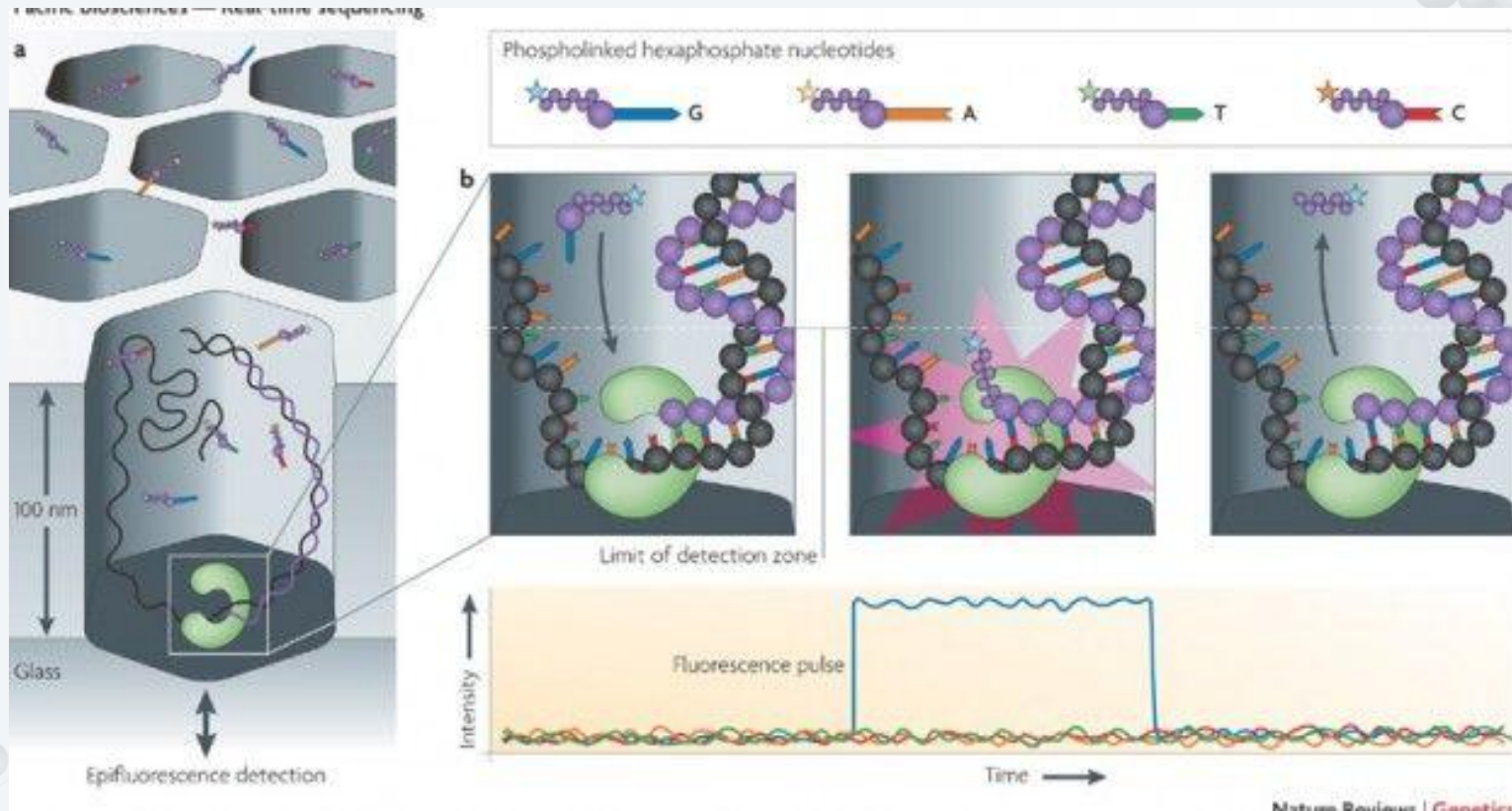


III GENERAZIONE: PACIFIC BIOSCIENCE

6. **Rilevazione in tempo reale.** Gli ZMW isolano ogni reazione di sequenziamento, consentendo di osservare l'incorporazione dei nucleotidi in tempo reale. La sequenza di segnali luminosi viene registrata, corrispondendo alla sequenza di basi del DNA.



III GENERAZIONE: PACIFIC BIOSCIENCE



III GENERAZIONE: PACIFIC BIOSCIENCE

Ogni SMRT cell contiene all'incirca 150.000 ZMW

Si tratta di un sequenziatore portatile, di piccole dimensioni, alimentato tramite una semplice porta USB e caratterizzato da un'elevatissima processività



III GENERAZIONE: NANOPORE

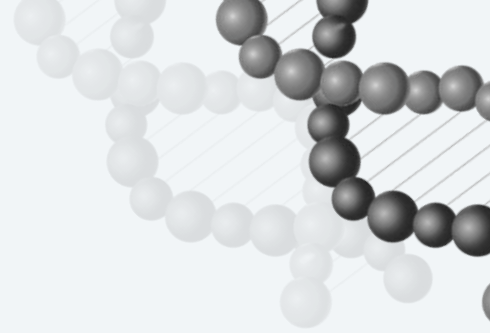
La tecnologia **Nanopore MinION** è uno strumento avanzato per il sequenziamento del DNA e dell'RNA, sviluppato da Oxford Nanopore Technologies.

Si applica con un dispositivo portatile e compatto, progettato per fornire un'analisi genetica flessibile, rapida e relativamente economica.

III GENERAZIONE: NANOPORE

1. Frammentazione (opzionale):

Sebbene il MinION possa leggere sequenze lunghe, talvolta si frammenta intenzionalmente il DNA in pezzi più piccoli per migliorare l'efficienza.



III GENERAZIONE: NANOPORE

2. **Preparazione della libreria.** Al DNA o allo RNA sono legati adattatori molecolari specifici, che includono:

- Tag di riconoscimento: aiutano il sequenziatore a identificare le estremità del frammento.
- Motore molecolare: una proteina che "spinge" il filamento di DNA o RNA attraverso il poro nanometrico.



III GENERAZIONE: NANOPORE

3. Caricamento sul dispositivo.

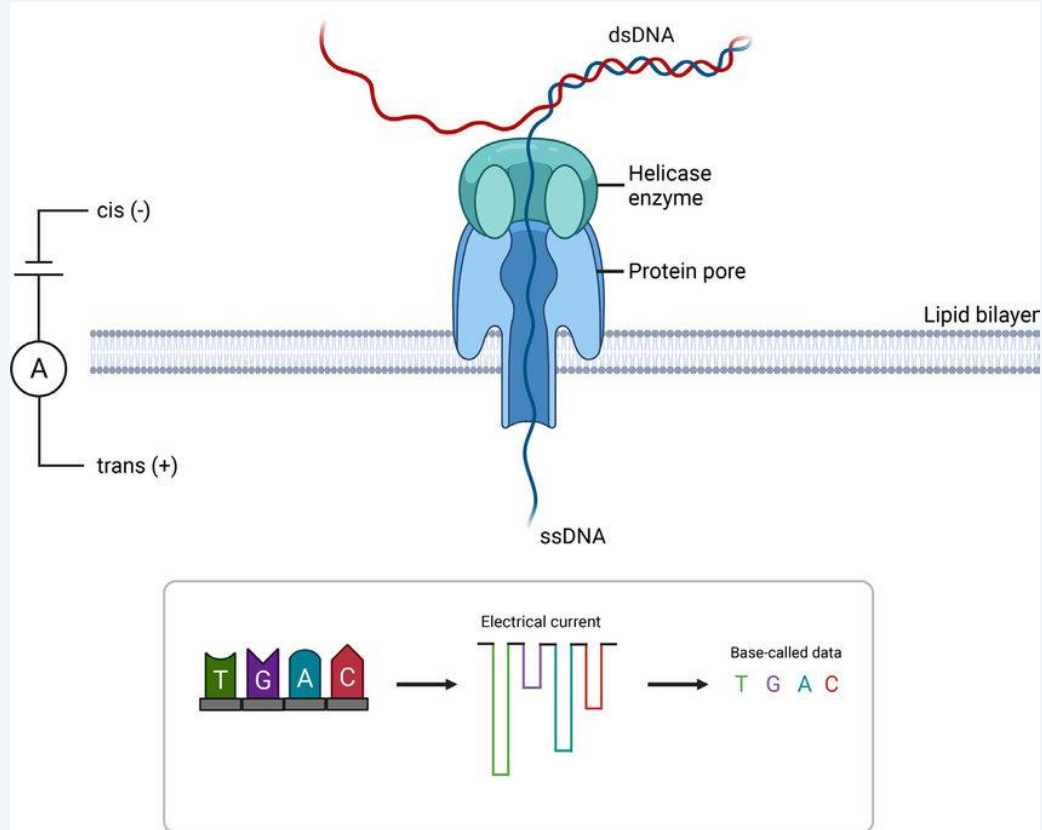
Il MinION utilizza una flow cell, una cartuccia contenente i pori nanometrici integrati in una membrana lipidica. La flow è calibrata e riempita con una soluzione per mantenere condizioni ottimali.

I pori sono proteine biologiche o sintetiche incorporate in una membrana lipidica. Una corrente ionica è generata attraverso il poro, e questa corrente muta (modulazione) quando una molecola (ad esempio, DNA o RNA) passa attraverso il poro



Caricamento della libreria: La libreria preparata viene immessa direttamente nella flow cell. Le molecole di DNA o RNA iniziano a entrare nei pori nanometrici.

III GENERAZIONE: NANOPORE



III GENERAZIONE: NANOPORE

4. Sequenziamento

Passaggio del DNA/RNA attraverso i pori: Una corrente elettrica viene applicata attraverso la membrana lipidica. Quando il filamento di DNA o RNA passa attraverso il poro, modifica la corrente in relazione alla sequenza delle basi nucleotidiche.

Rilevazione del segnale: Le variazioni nella corrente sono registrate in tempo reale come segnali elettrici.

Ogni base (A, T, C, G per il DNA; A, U, C, G per l'RNA) produce un segnale risultante elettrico univoco.

Decodifica in tempo reale: Un sensore dedicato converte i segnali elettrici in sequenze nucleotidiche leggibili. I dati vengono visualizzati e possono essere analizzati immediatamente.

III GENERAZIONE: NANOPORE

Nessuna Amplificazione: la tecnologia nanopore può sequenziare direttamente molecole di DNA o RNA senza la necessità di un'ampia fase di amplificazione precedente, il che riduce il bias (distorsioni sistematiche nella rappresentazione dei dati sequenziati rispetto al genoma o al campione originale) e gli errori associati all'amplificazione.

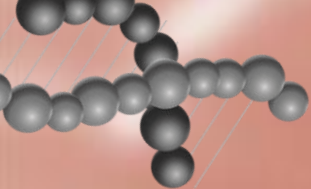
Lunghe reads: Il sequenziamento nanopore è noto per produrre letture estremamente lunghe, che possono estendersi per decine di migliaia di basi o addirittura oltre.

Portabilità: Un grande vantaggio del sequenziamento nanopore è la sua portabilità (ha le dimensioni di una chiavetta USB e può essere collegato a un laptop per il sequenziamento sul campo)

Applicazioni Versatili: Oltre al sequenziamento del DNA, la piattaforma nanopore può essere utilizzata per sequenziare l'RNA, identificare modifiche epigenetiche come la metilazione del DNA e analizzare proteine.

Errori: Come molte tecnologie di sequenziamento di terza generazione, il sequenziamento nanopore ha un tasso di errore intrinseco più elevato per lettura singola rispetto alle piattaforme di sequenziamento di seconda generazione. Tuttavia, poiché gli errori sono in gran parte casuali, possono essere corretti attraverso il sequenziamento profondo e l'uso di algoritmi appropriati





BIOINFORMATICA

FORMATO POST-NGS: FASTQ

Il formato standard di output dopo il sequenziamento NGS è il **FASTQ format**. Si tratta di un formato di testo (codice ASCII) che include sia la sequenza (in genere nucleotidica) sia la qualità di ogni base (score).

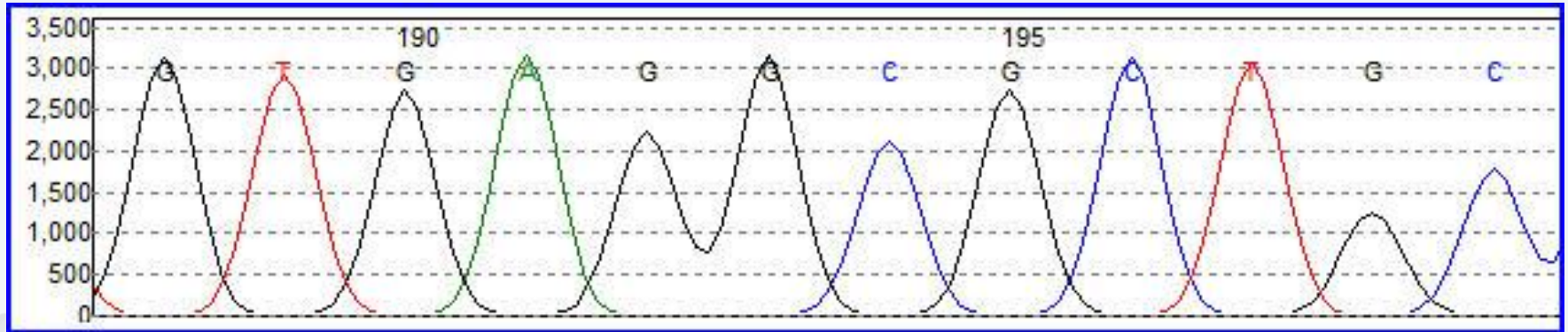
Header	Sequence	Quality
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979	GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT	+
	FFFFHHHHHHHJIJJJJJJJJJJIGIGIGIGGIJJJIJJJJJJJJII	
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893	GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA	+
	12BAFB?A:3<AE1@<FF;1*@EG*)?0?DBD>9BF9B*?#####	
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624	AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT	+
	BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?	

- ## Quality

@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHHJIIJJJJJJJIIJJJIGIGIGGIJJIIJJJJJJIII
@HWI-ST227:389:C4WA2ACXX:7:1205:15214:42893
GAGGATCCCAGGGAGGAAGGTCCTCGCTCCTCTTTCATCTAAGGGA
+
12BAFB?A:3<AE1@<FF;1*@(EG*)?0?DBD>9BF9B*?#####
@HWI-ST227:389:C4WA2ACXX:8:2208:2467:44624
AAAGAGGAGAGAGGACCATCCTCCCTGGGATCCTCAGAAGTCTACT
+
BDDA:DB?2AA@FC>F?EEGC<FED>GFD;?GBB?<?F99*/9?9?

FORMATO POST-NGS: FASTQ

Un esperimento di sequenziamento produce in generale un cromatogramma da cui viene derivata la sequenza delle basi tramite il cosiddetto processo (software) di **base calling** (chiamata del nucleotide).



La lettura di un cromatogramma permette l'associazione di un indice di qualità ad ogni base chiamata.

L'indice di qualità più usato attualmente è il **Phred Quality Score**.

FORMATO POST-NGS: PHRED

Phred è un algoritmo che prende le informazioni del cromatogramma durante il sequenziamento e rivaluta i picchi per produrre una "identificazione di base" che di solito è significativamente più accurata della chiamata originale. Oltre a ricalcolare i residui, Phred aggiunge anche informazioni sul punteggio di qualità a ciascun residuo.

Questo è un valore logaritmico da 0 a 99 dove un valore di 10 indica che c'è 1 possibilità su 10 che la chiamata sia in errore, un punteggio di 20 indica una possibilità su 1 su 100 che la chiamata sia in errore, un punteggio di 30 indica 1 possibilità su 1.000 di errore, ecc

$$Q = -10 \log_{10}(p)$$

Q = misura della qualità della base b in una sequenza

p = probabilità che la misurazione sia stata svolta correttamente $0 \leq p \leq 1$

FORMATO POST-NGS: PHRED

Calcolare la quality della sequenza A T G G C A con campionamenti aventi probabilità di errore di:

A,98.9%; T,67.5%; G,78.4%; G,12.3%; C,23.5%; A,88.9%

$$Q = -10 \log_{10}(p)$$

$$Q = -10 \cdot \log_{10} 98.9 \sim -10 \cdot -1.99 \sim +20$$

$$Q = -10 \cdot \log_{10} 67.5 \sim -10 \cdot -1.71 \sim +17$$

$$Q = -10 \cdot \log_{10} 78.4 \sim -10 \cdot -1.89 \sim +19$$

$$Q = -10 \cdot \log_{10} 12.3 \sim -10 \cdot -1.09 \sim +11$$

$$Q = -10 \cdot \log_{10} 23.5 \sim -10 \cdot -1.37 \sim +14$$

$$Q = -10 \cdot \log_{10} 88.9 \sim -10 \cdot -1.95 \sim +20$$

FORMATO POST-NGS: PHRED

La relazione del Phred score può anche essere scritta così:

$$p = 10^{\frac{-Q}{10}}$$

Ad esempio, se Phred assegna un punteggio di qualità pari a 30 a una base, le probabilità che questa base sia definita in modo errato sono 1 su 1000

FORMATO POST-NGS: PHRED

Calcolare la probabilità di errore supponendo che la sequenza A T G G C A abbia qualità:

A,33; T,22; G,30; G,7; C,15; A,41

$$p = 10^{\frac{-Q}{10}}$$

$$p = 10^{\frac{-33}{10}} = 10^{-3.3} = 0.0005$$

$$p = 10^{\frac{-22}{10}} = 10^{-2.2} = 0,0063$$

$$p = 10^{\frac{-30}{10}} = 10^{-3} = 0,001$$

$$p = 10^{\frac{-7}{10}} = 10^{-0.7} = 0,1995$$

$$p = 10^{\frac{-15}{10}} = 10^{-1.5} = 31,622$$

$$p = 10^{\frac{-15}{10}} = 10^{-1.5} = 0,0316$$

$$p = 10^{\frac{-41}{10}} = 10^{-4.1} = 0,000079$$

FORMATO POST-NGS: PHRED

L'intervallo del punteggio di qualità (Q) può iniziare da 0 (che indica una probabilità di errore del 100% o una certezza di 0% per la base sequenziata) a infinito (che indica una probabilità di errore di 0% o una certezza del 100% per la base sequenziata)

Nella pratica, i valori Phred seguono due rappresentazioni

PHRED33

PHRED64

FORMATO POST-NGS: PHRED

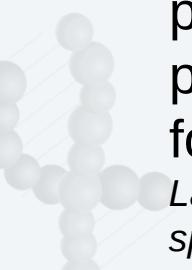
Phred+33

La codifica phred+33 prevede la somma del valore 33 al punteggio di qualità phred.

Ad esempio, utilizzando la codifica phred+33, un punteggio di qualità di 30 verrebbe rappresentato con il carattere ASCII con il valore 63 ($30 + 33$), che è '!'.

Phred+64

La codifica phred+64 funziona allo stesso modo della codifica phred+33, tranne per il fatto che si deve aggiungere 64 al punteggio phred per determinare il codice ascii del carattere di qualità (è un formato vetusto).

La scelta del Phred+33 è dovuta al fatto che i valori da 0 a 32 includono, nella codifica ASCII, spazi bianchi e caratteri non stampabili, che non possono essere identificati a occhio nudo

FORMATO POST-NGS: PRIMI 30 CARATTERI ASCII

Apreno un file testuale i caratteri dallo 0 al 31 non sono percepibili perché sono simboli di controllo (esempio il tasto di cancellazione, l'andata a capo) o di gestione dei file (la chiusura di un file)

Numero decimale	Numero binario	Codice	Significato
1	00000001	SOH	Start of Heading
2	00000010	STX	Start of Text
3	00000011	ETX	End of Text
4	00000100	EOT	End of Transmission
5	00000101	ENQ	Enquiry
6	00000110	ACK	Acknowledgment
7	00000111	BEL	Bell
8	00001000	BS	Back Space
9	00001001	HT	Horizontal Tab
10	00001010	LF	Line Feed
11	00001011	VT	Vertical Tab
12	00001100	FF	Form Feed
13	00001101	CR	Carriage Return
14	00001110	SO	Shift Out / X-On
15	00001111	SI	Shift In / X-Off
16	00010000	DLE	Data Line Escape
17	00010001	DC1	Device Control 1 (oft. XON)
18	00010010	DC2	Device Control 2
19	00010011	DC3	Device Control 3 (oft. XOFF)
20	00010100	DC4	Device Control 4
21	00010101	NAK	Negative Acknowledgement
22	00010110	SYN	Synchronous Idle
23	00010111	ETB	End of Transmit Block
24	00011000	CAN	Cancel
25	00011001	EM	End of Medium
26	00011010	SUB	Substitute
27	00011011	ESC	Escape
28	00011100	FS	File Separator
29	00011101	GS	Group Separator
30	00011110	RS	Record Separator
31	00011111	US	Unit Separator

FORMATO POST NGS: PHRED

Symbol	Phred Quality Score	Probability of Incorrect Base Call
!	0	1.000
"	1	0.794
#	2	0.631
\$	3	0.501
%	4	0.398
&	5	0.316
'	6	0.251
(	7	0.199
)	8	0.158
*	9	0.126
+	10	0.100
,	11	0.079
-	12	0.063
.	13	0.050
/	14	0.040
0	15	0.032
1	16	0.025
2	17	0.020
3	18	0.016
4	19	0.013
5	20	0.010

Symbol	Phred Quality Score	Probability of Incorrect Base Call
6	21	0.008
7	22	0.006
8	23	0.005
9	24	0.004
:	25	0.003
;	26	0.002
<	27	0.002
=	28	0.001
>	29	0.001
?	30	0.001
@	31	0.0008
A	32	0.0006
B	33	0.0005
C	34	0.0004
D	35	0.0003
E	36	0.0002
F	37	0.0002
G	38	0.0002
H	39	0.0001
I	40	0.0001



Grazie!

Domande?

franco.liberati@unitus.it

deb.scienceontheweb.com

