

Sequenziamento

dal dato alle strutture dati (e algoritmi)

How Do Biologists Assemble Genomic Puzzles from Millions of Pieces?

Fragment assembly and Graph Algorithms

Phillip Compeau and Pavel Pevzner

Bioinformatics Algorithms: an Active Learning Approach

cap. 3 rivisitato

Sequenziamento del DNA

- **Il dato di sequenziamento**
 - **Tecnologie di sequenziamento:**
 - SANGER sequencing
 - Next-Generation Sequencing (NGS)
 - **Qualità del dato di sequenziamento**
 - Formato Standard FASTQ
- **Fragment Assembly e approcci graph-based:**
 - Overlap Graph (OG) e de Bruijn Graph (dBG)

dato
(DNA) ottenuto il
lab e usato in silico
per assemblare un
frammento del genoma

Sequenziamento del DNA

- **Il dato di sequenziamento**
 - **Tecnologie di sequenziamento:**
 - SANGER sequencing
 - Next-Generation Sequencing (NGS)
 - **Qualità del dato di sequenziamento**
 - Formato Standard FASTQ
- **Fragment Assembly e approcci graph-based:**
 - Overlap Graph (OG) e de Bruijn Graph (dBG)

- Strutture di indicizzazione (BWT, SA, LCP)
- Bloom Filter per conteggio dei k-mer [& tecniche alignment-free]
- Problemi di spazio nella rappresentazione e calcolo overlap
- BWT+LCP+SA per costruzione lineare di OG
- ~~Pan-genome graph~~
- Fattorizzazione di Lyndon e varianti per: SA, BWT, signature per overlap

TO DO

IL DATO DI SEQUENZIAMENTO

Who Are These People?



Euler
1707-1783

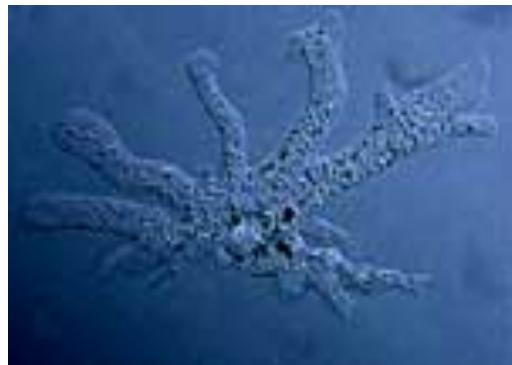
Hamilton
1805-1865

De Bruijn
1918-2012

The human genome is a three billion nucleotide long “book” written in A, C, G, T alphabet.

Some genomes are 100 X larger than the human genome:

Amoeba dubia



Paris japonica

Few Mutations Can Make a Big Difference...

- Different people have slightly different genomes: on average, roughly 1 mutation in 1000 nucleotides.
- The 1 in 1000 nucleotides difference accounts for height, high cholesterol susceptibility, and 1000s of genetic diseases.



```
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGAT  
CAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCG  
ATCGATCGATCGATTATCTACGATCGATCGATCGATCACTA  
TACGAGCTACTACGTACGTACGATCGCGGACTATTATCGA  
CTACAGATAAAACATGCTAGTACAACAGTATAACATAGCTGC  
GGGATACGATTAGCTAATAGCTGACGATATCCGAT
```

```
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGAT  
CAGCTACACACATCGTAGCTACGATGCATTAGCAAGCTATCG  
ATCGATCGATCGATTATCTACGATCGATCGATCGATCACTA  
TACGAGCTACTACGTACGTACGATCGCGTGACTATTATCGA  
CTACAGATGAAACATGCTAGTACAACAGTATAACATAGCTGC  
GGGATACGATTAGCTAATAGCTGACGATATCCGAT
```

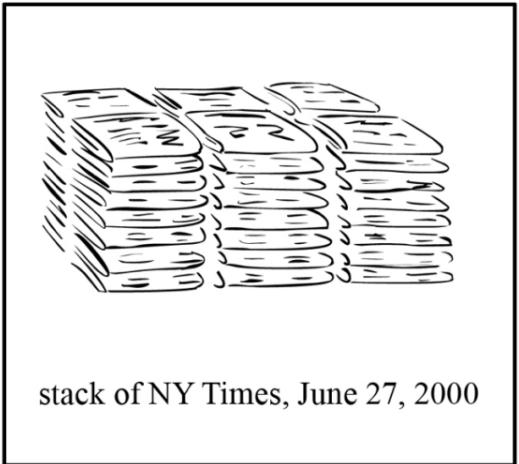


Why Do We Sequence Personal Genomes?

- **2010:** Nicholas Volker became first human being to be saved by genome sequencing.
 - Doctors could not diagnose his condition; he went through dozens of surgeries.
 - Sequencing revealed a rare mutation in a *XIAP* gene linked to a defect in his immune system.
 - This led doctors to use immunotherapy, which saved the child.



The Newspaper Problem

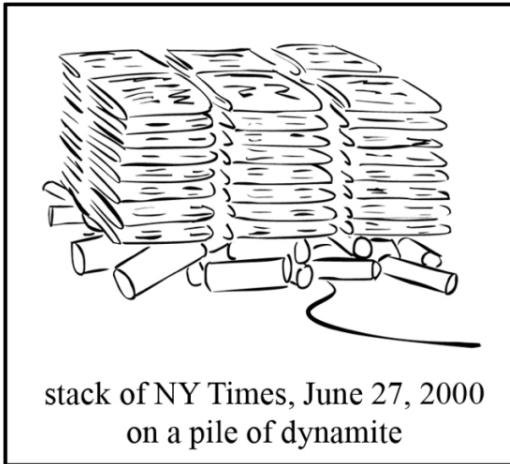


stack of NY Times, June 27, 2000

The Newspaper Problem



stack of NY Times, June 27, 2000

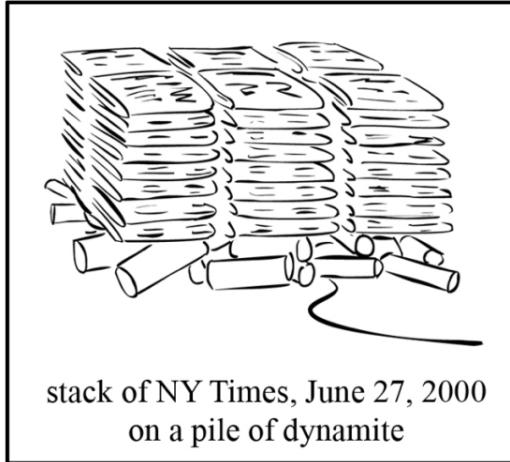


stack of NY Times, June 27, 2000
on a pile of dynamite

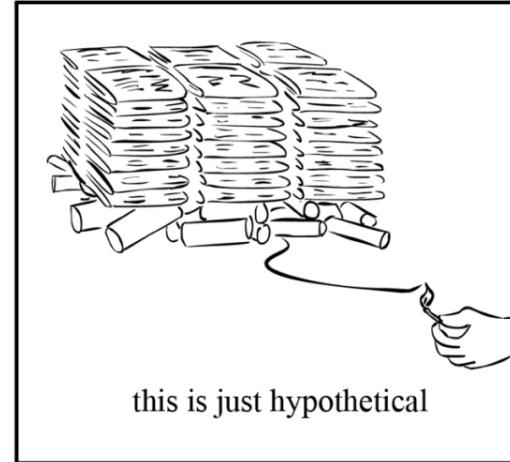
The Newspaper Problem



stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000
on a pile of dynamite

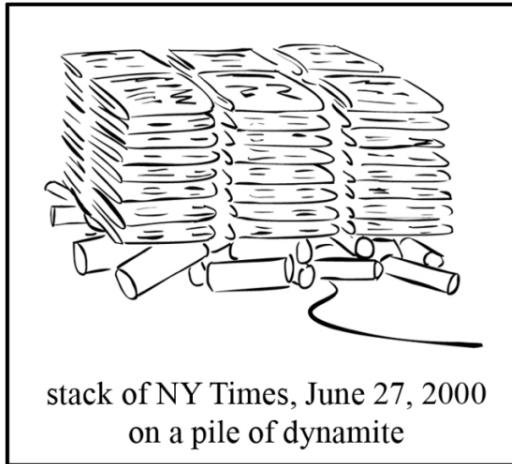


this is just hypothetical

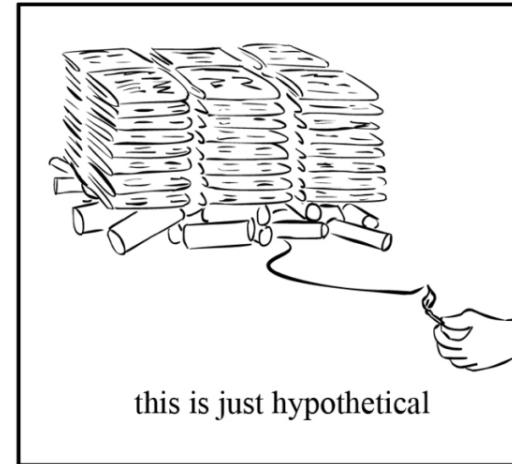
The Newspaper Problem



stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000
on a pile of dynamite



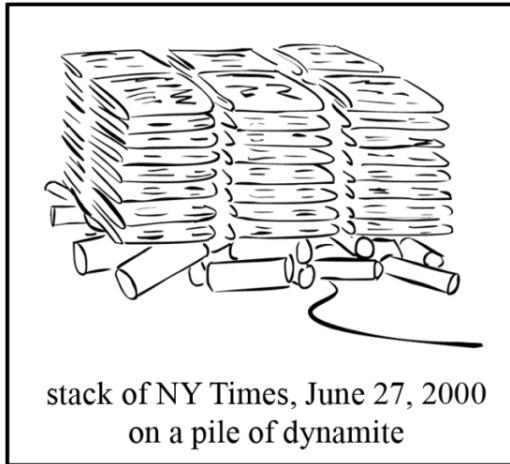
this is just hypothetical



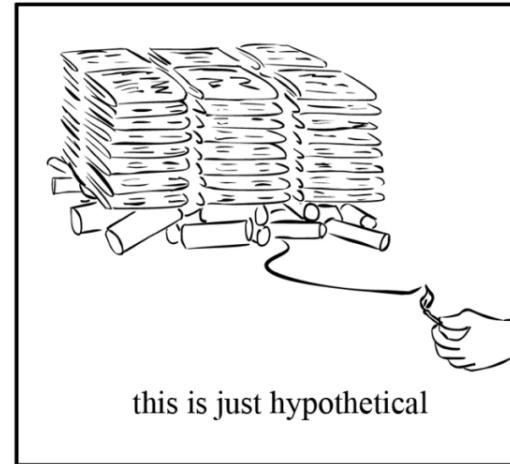
The Newspaper Problem



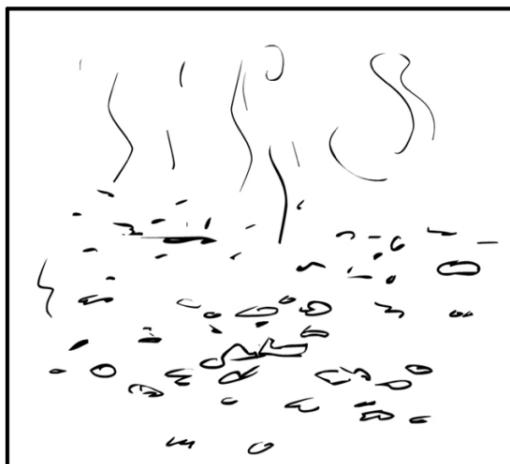
stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000
on a pile of dynamite



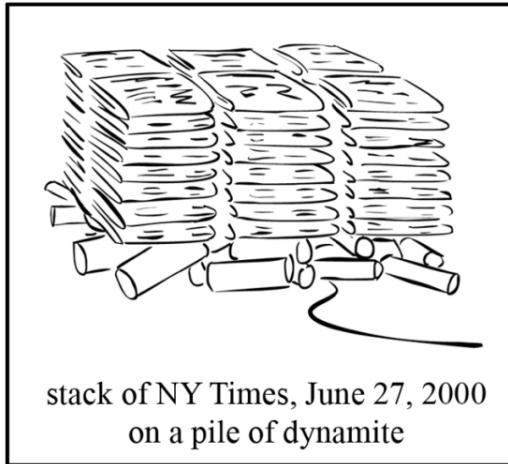
this is just hypothetical



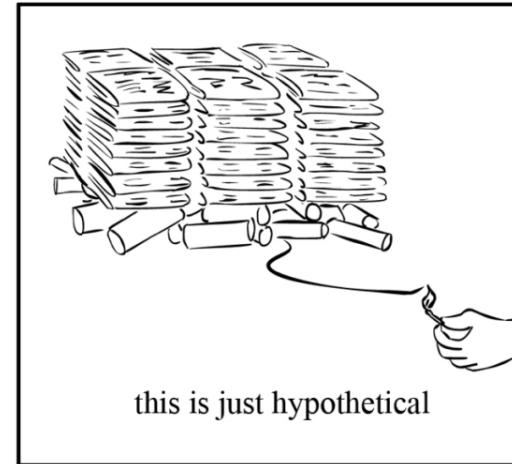
The Newspaper Problem



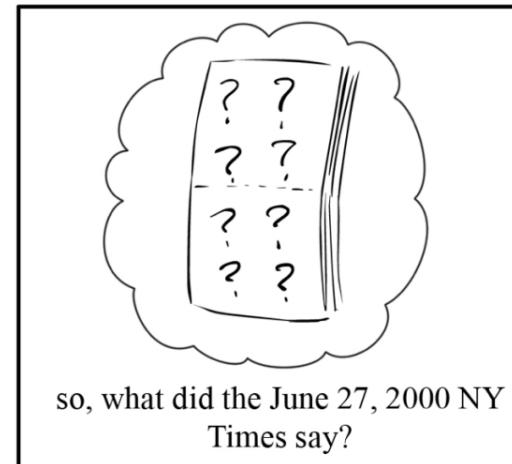
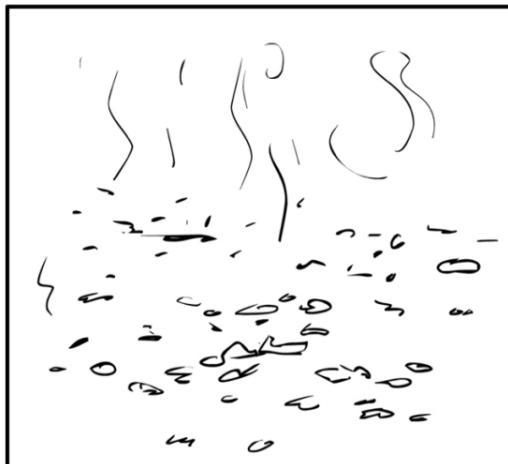
stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000
on a pile of dynamite

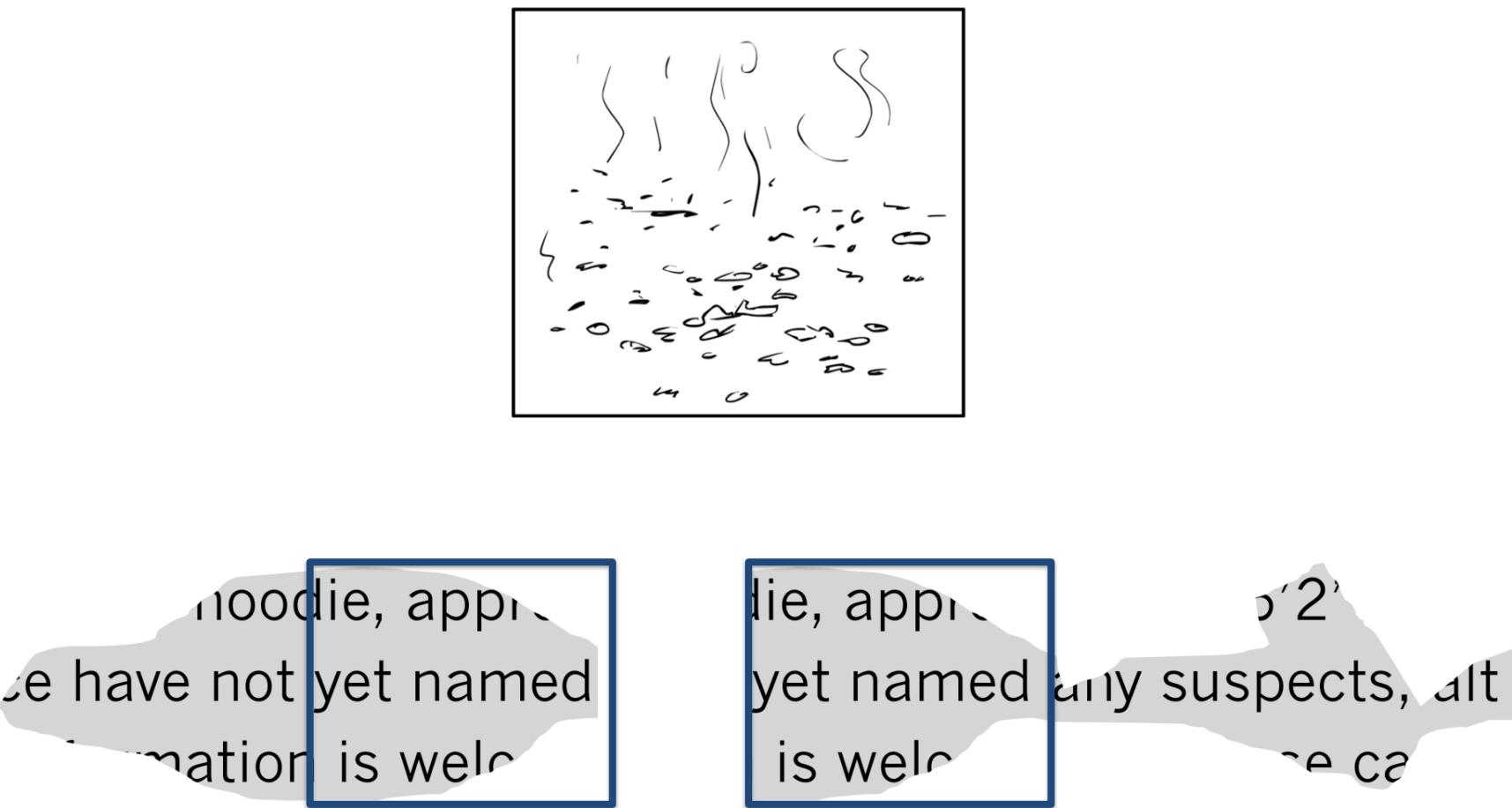


this is just hypothetical



so, what did the June 27, 2000 NY
Times say?

The Newspaper Problem as Overlapping Puzzle



The Newspaper Problem as an Overlapping Puzzle



Multiple Copies of a Genome (Millions of them)



CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

Breaking the Genomes at Random Positions



CTGATGA^{*}GGACTACGC^{*}ACTACTGCT^{*}GCTGTATT^{*}GATCAGCTACC^{*}CATCGTAGCTA^{*}GATGCATTAG^{*}AGCTATCC^{*}ATCAGCTAC^{*}ACATCGTAGC
CTGA^{*}GATGGACT^{*}ZGCTACTACT^{*}ZTAGCTGTAT^{*}ACGATCAGC^{*}ACCACATCGT^{*}GCTACGGATGC^{*}TAGCAAGC^{*}ATCGGATCA^{*}CTACCACAT^{*}GTAGC
CTGATGA^{*}GGACTACGC^{*}ACTACTGCTA^{*}ZTGATTAC^{*}ATCAGCTA^{*}CACATCGTAGC^{*}ACGATGCATT^{*}GCAAGCTA^{*}GGATCAGCT^{*}CCACATCGTAGC
CTGATGATGG^{*}CTACGCTAC^{*}ACTGCTAGCT^{*}TATTACGAT^{*}AGCTACCACA^{*}CGTAGCTACG^{*}TGCATTAGCA^{*}GCTATCGG^{*}TCAGCTACCA^{*}ATCGTAGC

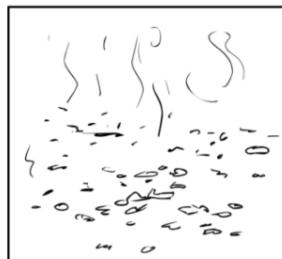
Generating “Reads”

CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC
CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC
CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

this is the last step in lab

—> file to be analyzed in silico

“Burning” Some Reads



CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC
CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC
CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

No Idea What Position Every Read Comes From

ATCAGCTACCA
TACTGCTAG
CTGATGA
ATGCATTAGCA
CTGATGATG
ACGCTACT
ACATCGTAGCT
TACTGCTAGCT
GCTAGCTGTAT
ATCGTAGCTACG
GGATCAGCTAC
ATCGGATCA
ACTACTGCTA
GCAAGCTAC
GACTACGCT
TACTGCTAGCT
ATCGTATTACG
CATCGTAGC
ACTACTGCTAC
GCAAGCTAC
ACTACGCTAC
TACGATCAGC
AGCTACCAC
AGCTACCAC
ACGATGCATTA
CTGATGATGG
TCGTAGCTACG
CTGATGATGG
TCGTAGCTACG
ATGCATTAGCAA
CACATCGTAGC
TACCAACATCGT
CTGATGATGG
ATCGTAGCTACG
AGCTACGATGCA
ATGCATTAGCA
CATCGTAGC
TCAGCTACCA
ATGCATTAGCA
ATGCATTAGCA
AGCTATCGG
AGCTATCGG
ATGCATTAGCA
ATGCATTAGCA

No Idea What Position Every Read Comes From

A collection of DNA sequence reads arranged in a grid, rotated diagonally. Some reads are highlighted with yellow boxes.

Highlighted reads (in yellow boxes):

- GCTATCGGA
- GCAAGCTATC

Other visible reads include:

- ATCAGCTACCA
- TACTGCTAG
- CTGATGATGGACT
- ATCAGCTACC
- GCTGTATTACG
- TGGACTACGCTAC
- TAGCAAGCT
- AGCTATCGG
- AGCTACGATGCA
- ATGCATTAGCA
- CTGATGAA
- TACTGCTAGCT
- ATGCATTAGCA
- CTGATGATG
- ACGCTACTACT
- ACATCGTAGCT
- GACTACGCT
- CTGTATTACG
- CATCGTAGC
- GCTACCACATC
- ATCAGCTACACA
- TACGATCAGC
- AGCTACCAC
- ACGATGCATTA
- CACATCGTAGC
- TACCAACATCGT
- CTGATGATGG
- ATCGTAGCTACG
- TACTGCTAGCT
- ATCGTAGCTACG
- GGATCAGCTAC
- ACTACTGCTA
- GCAAGCTATC
- ACTACGCTAC
- GCTAGCTGTAT
- GTATTACGATC

No Idea What Position Every Read Comes From

A collection of DNA sequence reads shown as overlapping diagonal lines:

- ATCAGCTACCA
- TACTGCTAG
- CTGATGATGGACT
- ATCAGCTACC
- GCTGTATTACG
- TGGACTACGCTAC
- TAGCAAGCT
- AGCTATCGG
- AGCTACGATGCA
- ATGCATTAGCA
- CTGATGA
- TACTGCTAGCT
- ATGCATTAGCA
- GCTATCGGA
- CTGATGATG
- ATGCATTAGCA
- ACGCTACTACT
- ACATCGTAGCT
- GCAAGCTATC
- GACTACGCT
- CTGTATTACG
- CATCGTAGC
- GCTACCACATC
- ATCAGTACACACA
- TACGATCAGC
- AGCTACCAC
- ACGATGCATTA
- CACATCGTAGC
- TACACATCGT
- CTGATGATGG
- ATCGTAGCTACG
- TACTGCTAGCT
- ATCGGATCA
- GGATCAGCTAC
- ACTACTGCTA
- ATCGTAGCTACG
- GCAAGCTATC
- ACTACGCTAC
- GCTAGCTGTAT
- GTATTACGATC

From Experimental to Computational Challenges

Multiple (unsequenced) genome copies **(by restriction enzymes)**



↓ ↓ ↓ ↓ ↓ **Read generation**

Reads **(by sequencing, in lab, I have the primary structure of each fragment)**



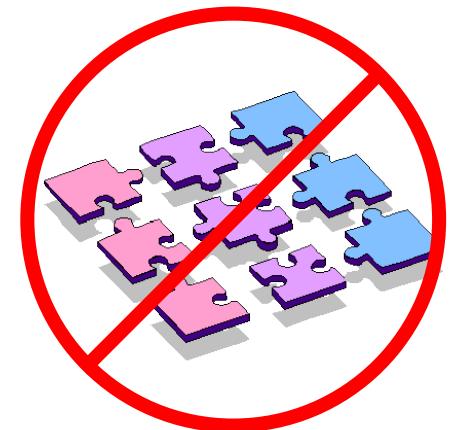
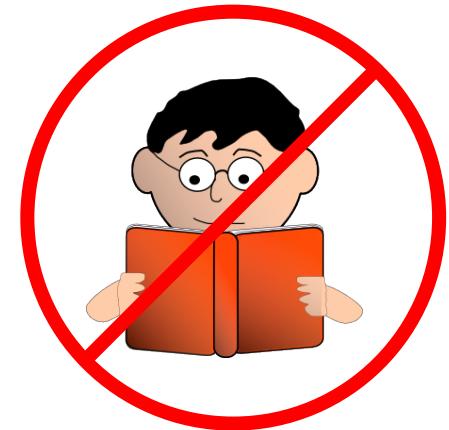
↓ ↓ ↓ ↓ ↓ **Genome assembly**

Assembled genome

...GGCATGCGTCAGAAACTATCATAGCTAGATCGTACGTAGCC...

What Makes Genome Sequencing Difficult?

- Modern sequencing machines cannot read an entire genome one nucleotide at a time from beginning to end (like we read a book)
- They can only shred the genome and generate short **reads**.
- The genome assembly is not the same as a puzzle: we must use *overlapping* reads to reconstruct the genome, a giant **overlap puzzle!**



Sequenziamento

Sequenziare → determinare la sequenza primaria di una molecola biologica

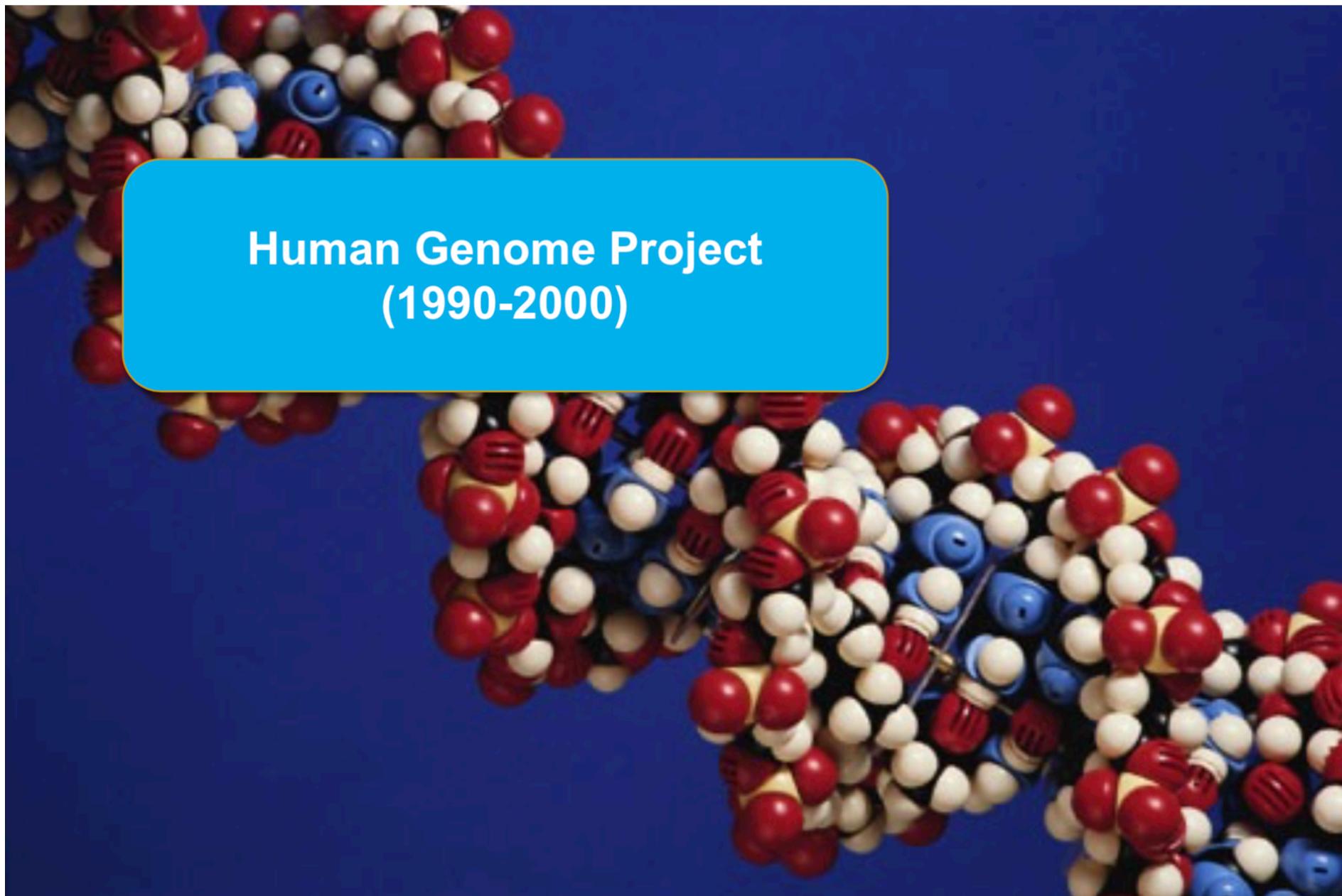
- sequenza delle basi {A,C,G,T|U} per DNA e RNA
- sequenza degli amminoacidi per le proteine

Due step:

1. Shotgun sequencing (in laboratorio)
2. Fragment assembly (*in silico*)

Shotgun Sequencing

Human Genome Project
(1990-2000)



Shotgun Sequencing → reads

each fragment is individually sequenced —> primary structure



Fragment Assembly

by using overlaps...



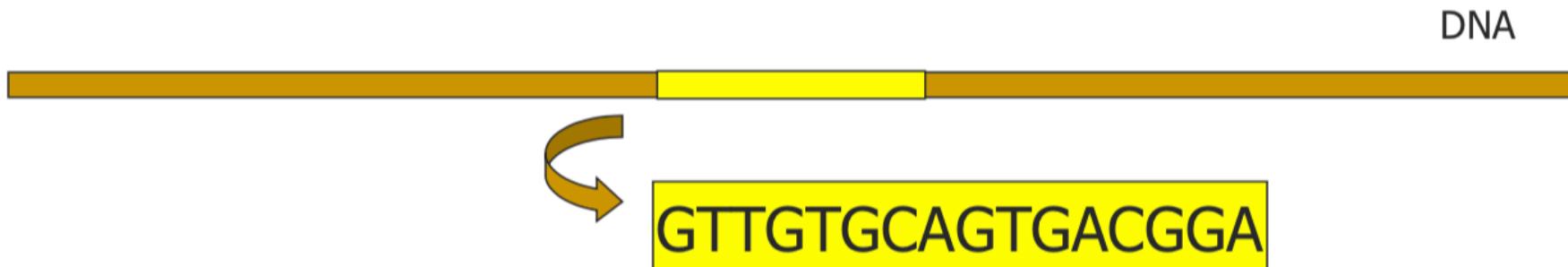
ACATGGCAAAATCCCATCTCTACAAAAAAATACAAAAAAA
TAAAAACTAGCCAGGTGTGGTGGCACATGCCTGTAATCGC
AGCTACTTGGGAGGGCTGAGGCAGAAGAACATTGAATC
TGGGAGGCAGAAGTTGCAGTGAGTTAAGATCATGCCACC
GCACTCCAGCCTGGGCAACAGAGCAAGACTTCTCAAAA
AATAAAAAATAAATAAAACATAAAAAAAATCAGCCACAG
GACTTGGTCTTGGACCCAAGTTAGAGCTAGGCCATGCTT
GCTTAAGGAGTGGCTGTAATTAAACAAGGCTAGTGGG
AAAGTTCCAGGCCATCTAACATTGTAGGTGCATTTT
TCTCTTCTTCACAGCTGACAACAGATGCCCTAATTGTT
TCACCATTAGCAGTTGACCATCTCATCACTTTACCT
CTCTTCTTTAGAAGAATGGAAAGACAGAAAATGCAG
AAAATTGATCAATTACAGAGAAAAA

Primo tipo di dato
ottenuto dal
sequenziamento

Single-end read

Cosa si ottiene da un esperimento di sequenziamento di una molecola di DNA (o RNA)?

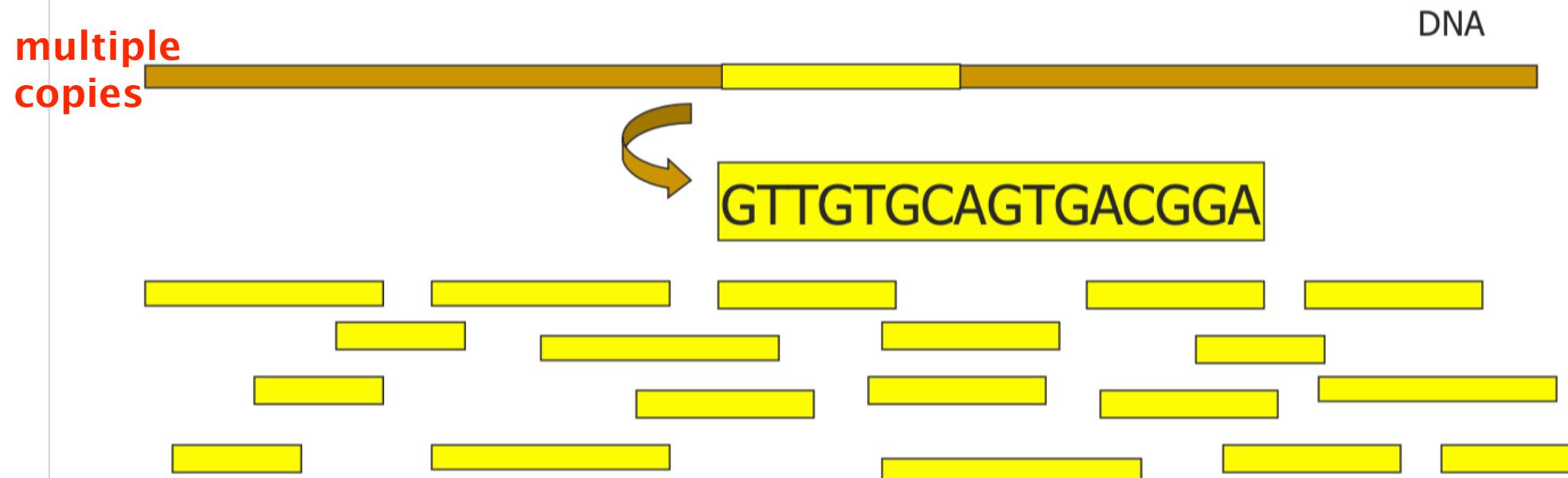
→ single-end *read*



Single-end read

Cosa si ottiene da un esperimento di sequenziamento di una molecola di DNA (o RNA)?

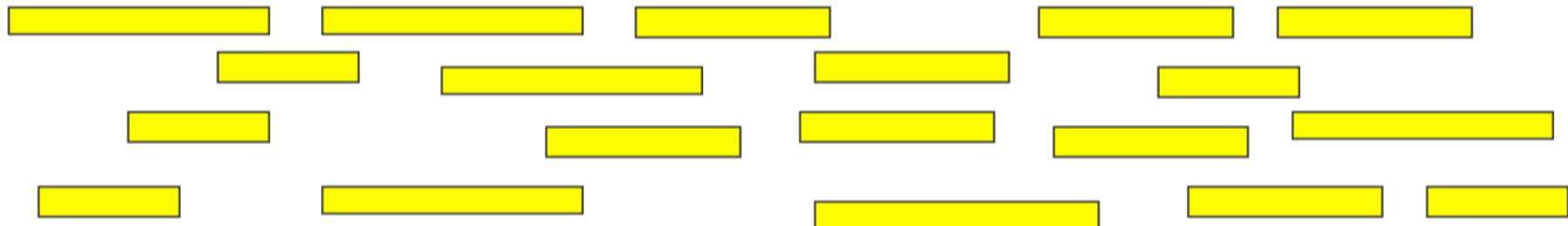
→ single-end *read*



Single-end read

Cosa si ottiene da un esperimento di sequenziamento di una molecola di DNA (o RNA)?

- Il fragment assembly ha il compito di assemblare i reads sequenziata nella sequenza primaria originale



Single-end read

Cosa si ottiene da un esperimento di sequenziamento di una molecola di DNA (o RNA)?

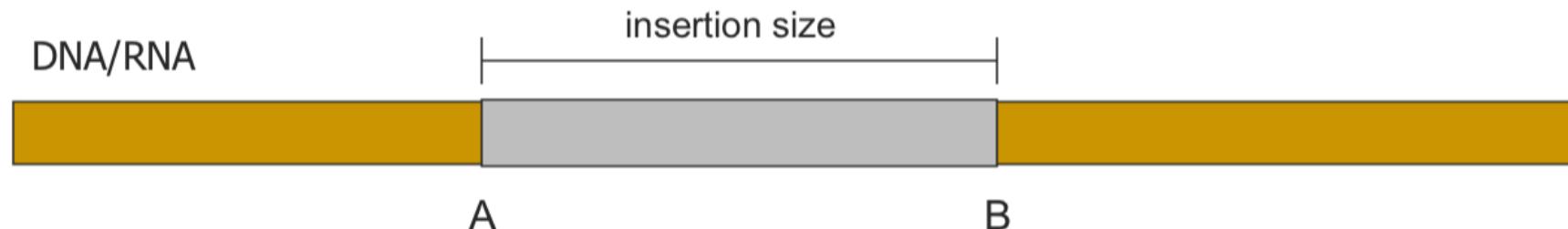
- Il fragment assembly ha il compito di assemblare i reads sequenziata nella sequenza primaria originale



Altro tipo di dato
ottenuto dal
sequenziamento

Pair-end / Mate-pair end

Paired-end/Mate-pair reads



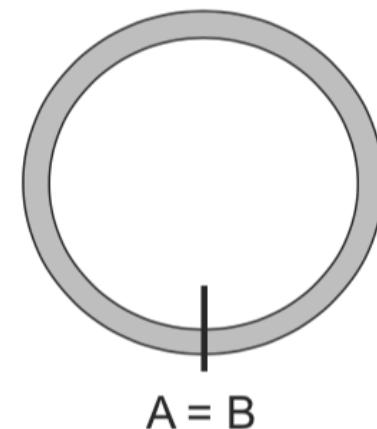
Considero la sottostringa tra A e B
(questo sarà sequenziato)

Pair-end / Mate-pair end

Paired-end/Mate-pair reads

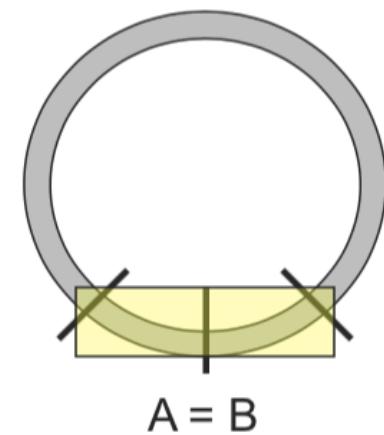
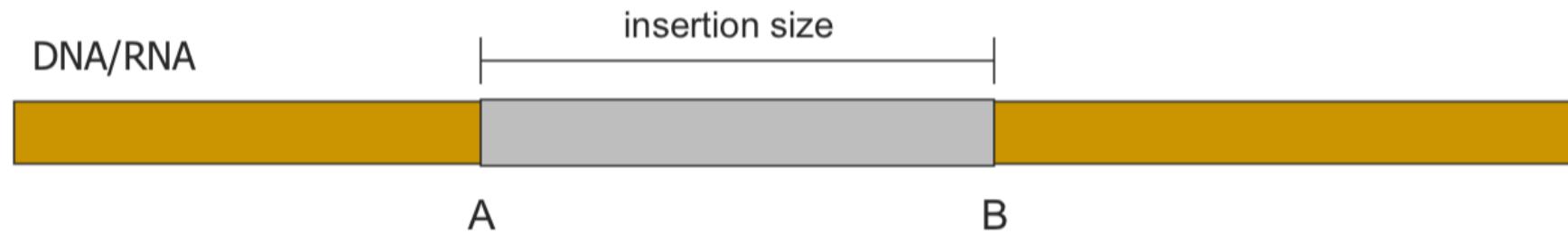


circolarizzo



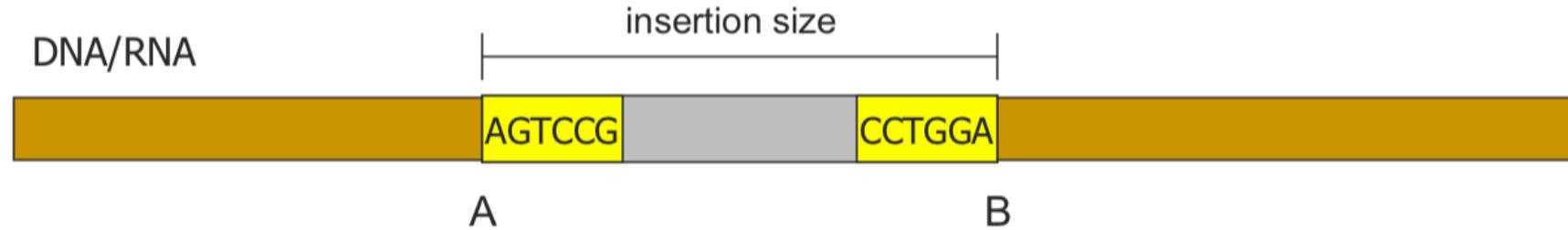
Pair-end / Mate-pair end

Paired-end/Mate-pair reads

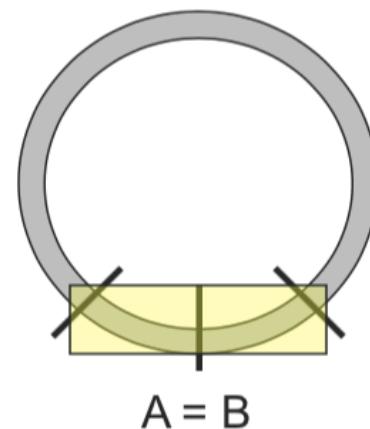


Pair-end / Mate-pair end

Paired-end/Mate-pair reads

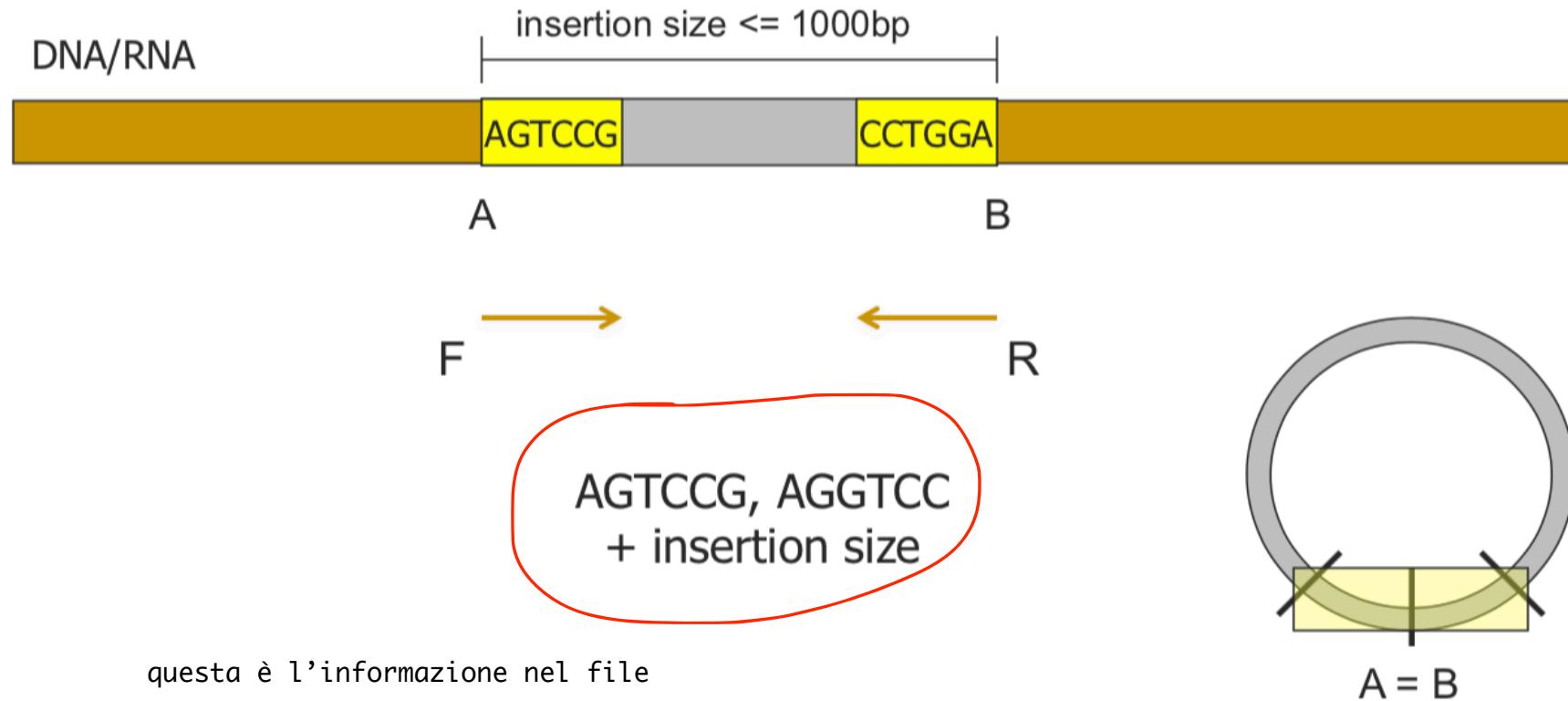


i due frammenti sono logicamente accoppiati
dall'insertion site (conosco la lunghezza che li
separa)



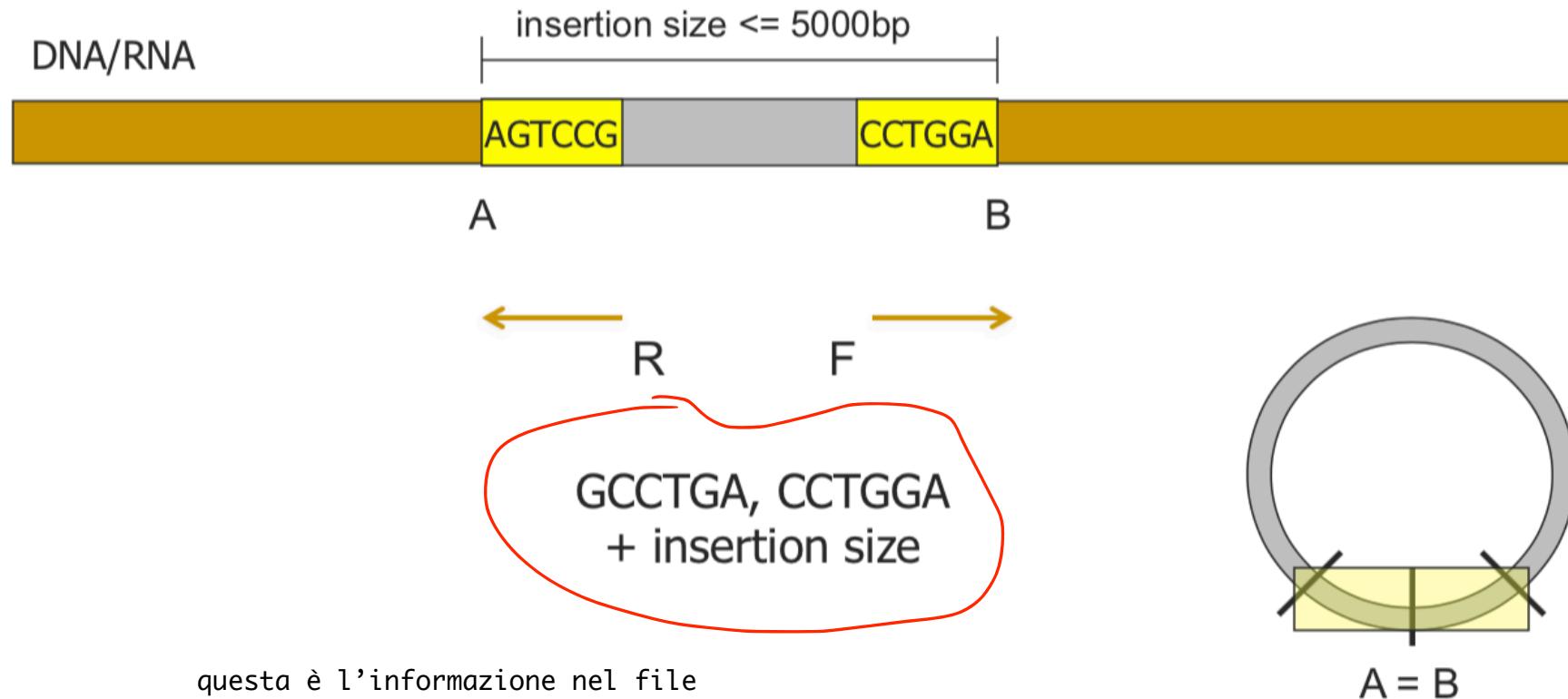
Pair-end / Mate-pair end

Paired-end/Mate-pair reads



Pair-end / Mate-pair end

Paired-end/Mate-pair reads



TECNICHE DI SEQUENZIAMENTO

Tecniche di sequenziamento

- ✓ 1977: Metodo tradizionale Sanger (first generation)
- ✓ 2000: Next-Generation Sequencing (NGS) (second generation)
- ✓ 2005: Next-Next-Generation Sequencing (third generation)

[Metodo Sanger]

- ✓ Ideato nel 1977 da Frederik Sanger
- ✓ Metodo di sequenziamento utilizzato per ottenere il primo genoma umano (Human Genome Project, HGP) completato nel 2001

Metodo Sanger

- ✓ Ideato nel 1977 da Frederik Sanger
- ✓ Metodo di sequenziamento utilizzato per ottenere il primo genoma umano (Human Genome Project, HGP) completato nel 2001



Collaborazione tra i gruppi di ricerca a livello internazionale che hanno portato a sviluppo di:

- ✓ banche dati genomiche
es: GenBank di NCBI
- ✓ software
es: BLAST (Basic Local Alignment Software Tool)

[Metodo Sanger]

Sanger → Chain-termination method

Componenti:

- ✓ DNA Template sequenza da sequenziare
(non la conosco) tante copie
- ✓ Primer
- ✓ DNA polimerasi
- ✓ Deossinucleosidi (dNTP → dATP, dCTP, dGTP, dTTP)
- ✓ Dideoxynucleosidi (ddNTP → ddATP, ddCTP, ddGTP, ddTTP)

[Metodo Sanger]

Sanger → Chain-termination method

Componenti:

T G C

- ✓ DNA Template
- ✓ Primer **tante copie**
- ✓ DNA polimerasi
- ✓ Deossinucleosidi (dNTP → dATP, dCTP, dGTP, dTTP)
- ✓ Dideoxynucleosidi (ddNTP → ddATP, ddCTP, ddGTP, ddTTP)

[Metodo Sanger]

Sanger → Chain-termination method

Componenti:

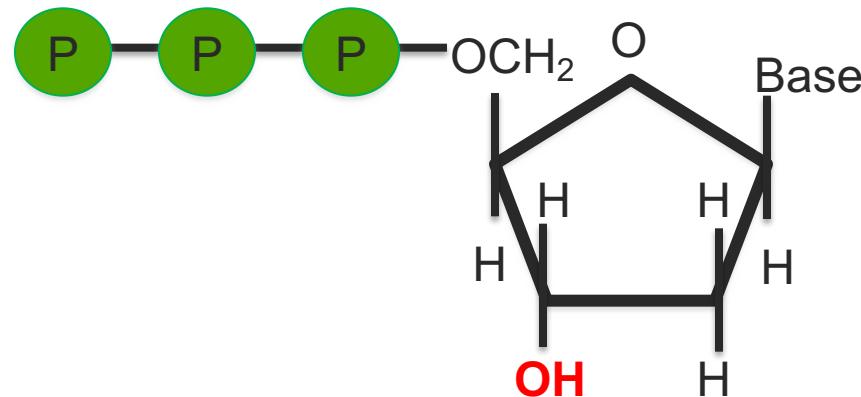
- ✓ DNA Template
- ✓ Primer
- ✓ DNA polimerasi enzima che “copia”
- ✓ Deossinucleosidi (dNTP → dATP, dCTP, dGTP, dTTP)
- ✓ Dideoxynucleosidi (ddNTP → ddATP, ddCTP, ddGTP, ddTTP)

[Metodo Sanger]

Sanger → Chain-termination method

Componenti:

- ✓ DNA Template
- ✓ Primer
- ✓ DNA polimerasi
- ✓ Deossinucleosidi (dNTP → dATP, dCTP, dGTP, dTTP) **tanti**
- ✓ Dideoxynucleosidi (ddNTP → ddATP, ddCTP, ddGTP, ddTTP)

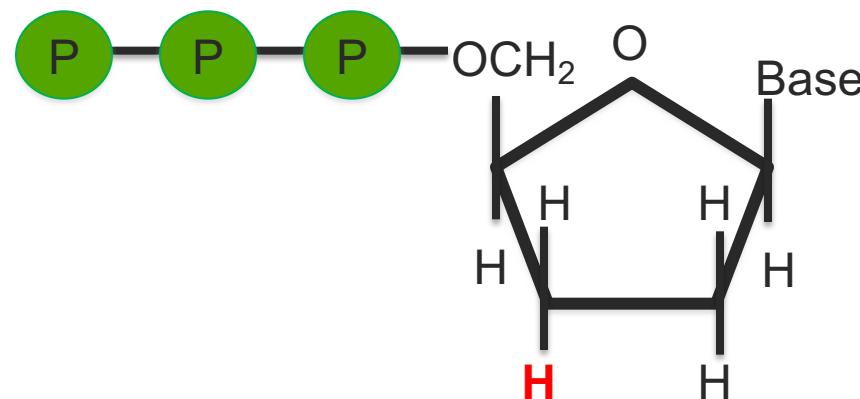


[Metodo Sanger]

Sanger → Chain-termination method

Componenti:

- ✓ DNA Template
- ✓ Primer
- ✓ DNA polimerasi
- ✓ Deossinucleosidi (dNTP → dATP, dCTP, dGTP, dTTP)
- ✓ Dideoxynucleosidi (ddNTP → ddATP, ddCTP, ddGTP, ddTTP)

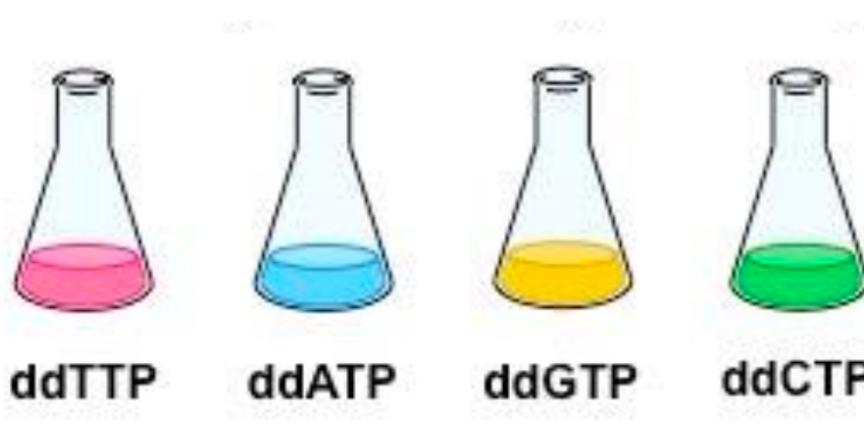


quantità minore

[Metodo Sanger]

Sanger → Chain-termination method

Componenti:

- ✓ DNA Template
 - ✓ Primer
 - ✓ DNA polimerasi
 - ✓ Deossinucleosidi (dNTP → dATP, dCTP, dGTP, dTTP)
 - ✓ Dideoxynucleosidi (ddNTP → ddATP, ddCTP, ddGTP, ddTTP)
- 

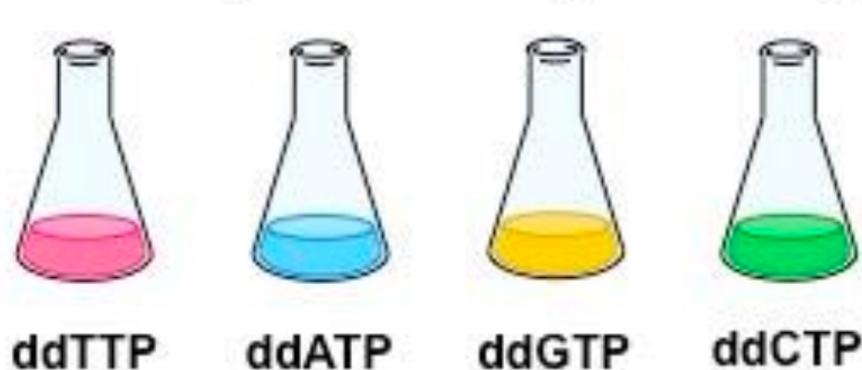
Metodo Sanger

Sanger → Chain-termination method

Procedimento:

- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

T G C A G G C A T C T G A
| | | | | | | | | | | | |
A C G T C C G T A G A C T



Metodo Sanger

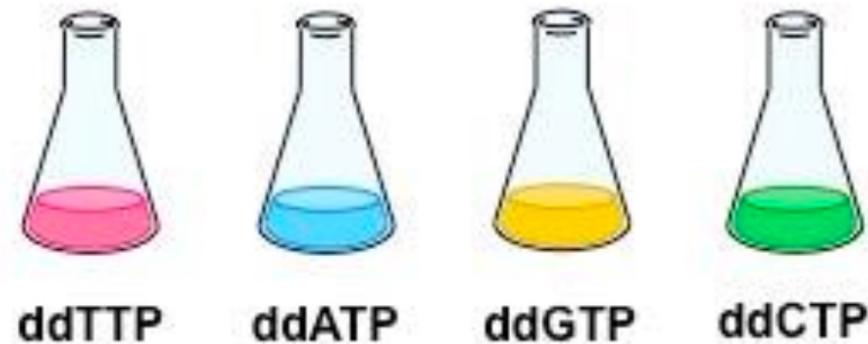
Sanger → Chain-termination method

Procedimento:

- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

T G C A G G C A T C T G A

| | | | | | | | | | | | |
A C G T C C G T A G A C T



[Metodo Sanger]

Sanger → Chain-termination method

Procedimento:

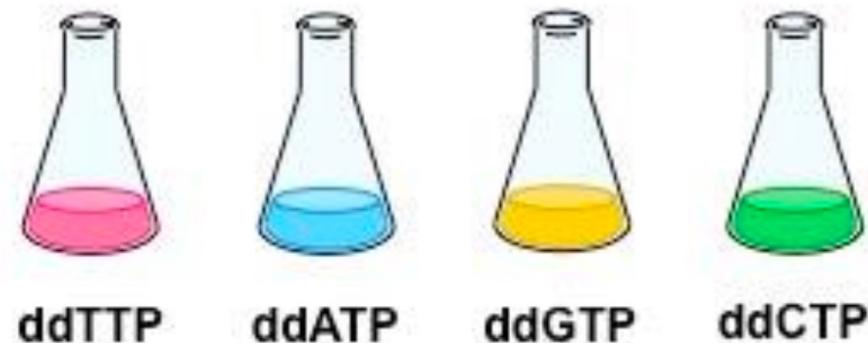
- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

T G C A G G C A T C T G A

T G C

| | | | | | | | | | | | | |

A C G T C C G T A G A C T



Metodo Sanger

Sanger → Chain-termination method

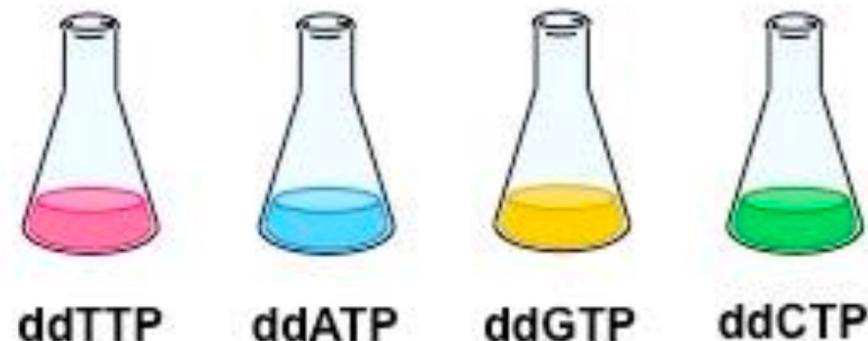
Procedimento:

- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

T G C A G G C A T C T G A

I G C A G G C

| | | | | | | | | | | | |
A C G T C C G T A G A C T



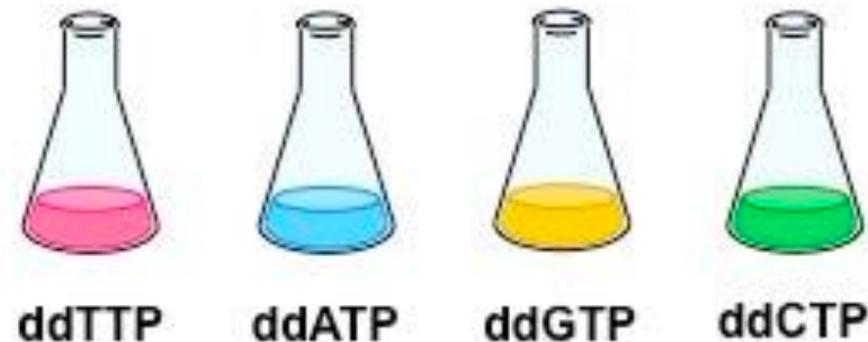
Metodo Sanger

Sanger → Chain-termination method

Procedimento:

- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

T G C A G G C A T C T G A
T G C A G G C A
| | | | | | | | | | | |
A C G T C C G T A G A C T



[Metodo Sanger]

Sanger → Chain-termination method

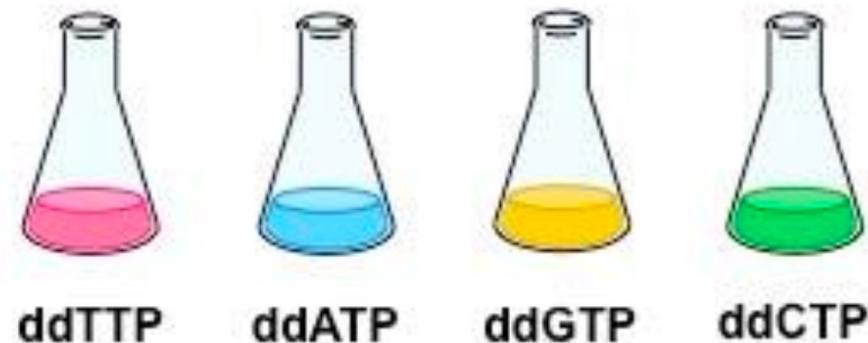
Procedimento:

- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

T G C A G G C A T C T G A

I G C A G G C A

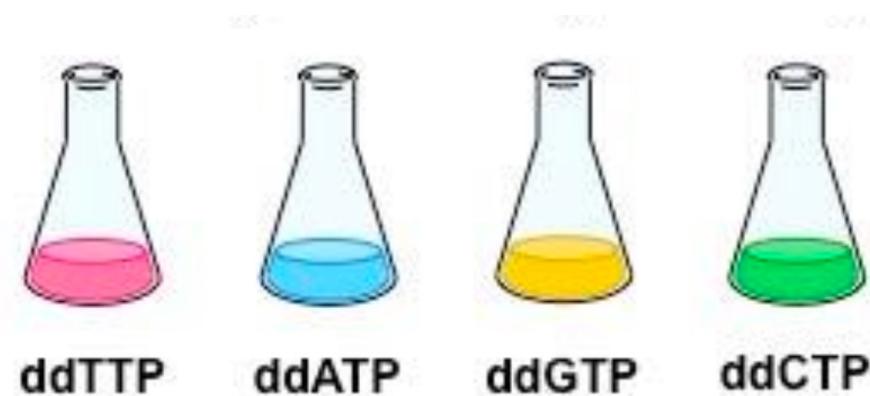
| | | | | | | | | | | | |
A C G T C C G T A G A C T



[Metodo Sanger]

Sanger → Chain-termination method

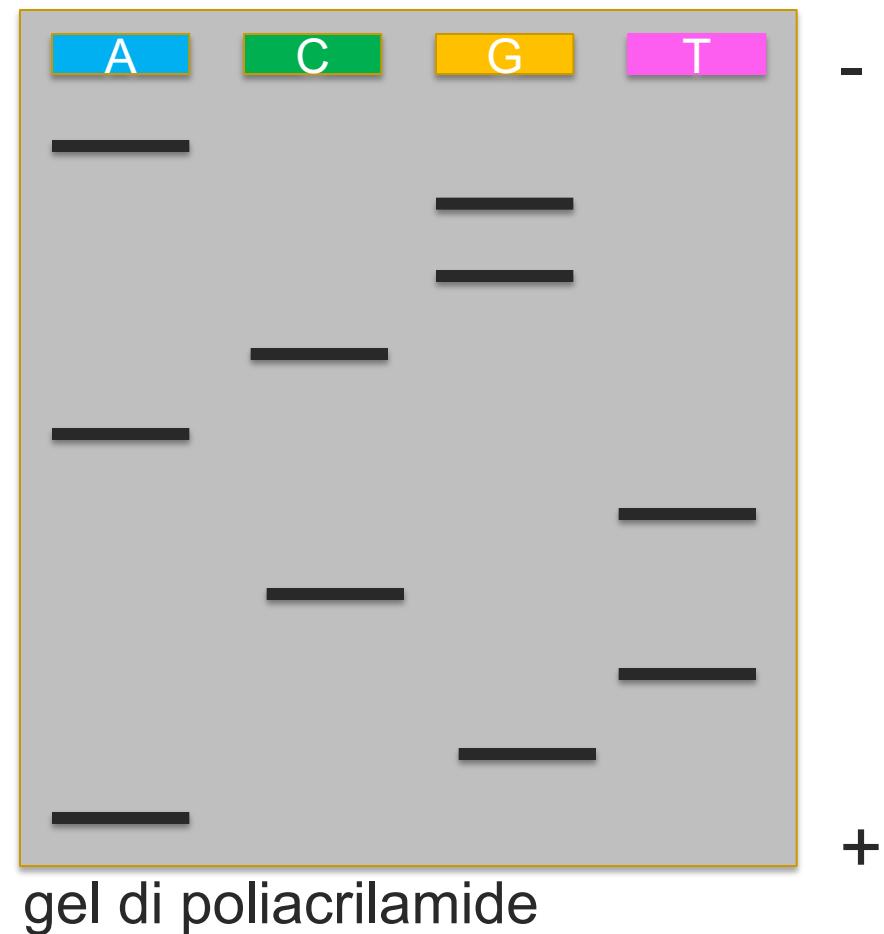
TGCA
TGCA~~G~~**A**
TGCA~~G~~**G**CATCT**G**
TGCA~~G~~**G**
TGCA~~G~~**G**
TGCA~~G~~**G**CATCT**G**
TGCA~~G~~**G**CAT**C**
TGCA~~G~~**G**
TGCA~~G~~**G**CAT**T**
TGCA~~G~~**G**CATCT**T**



[Metodo Sanger]

Sanger → Chain-termination method

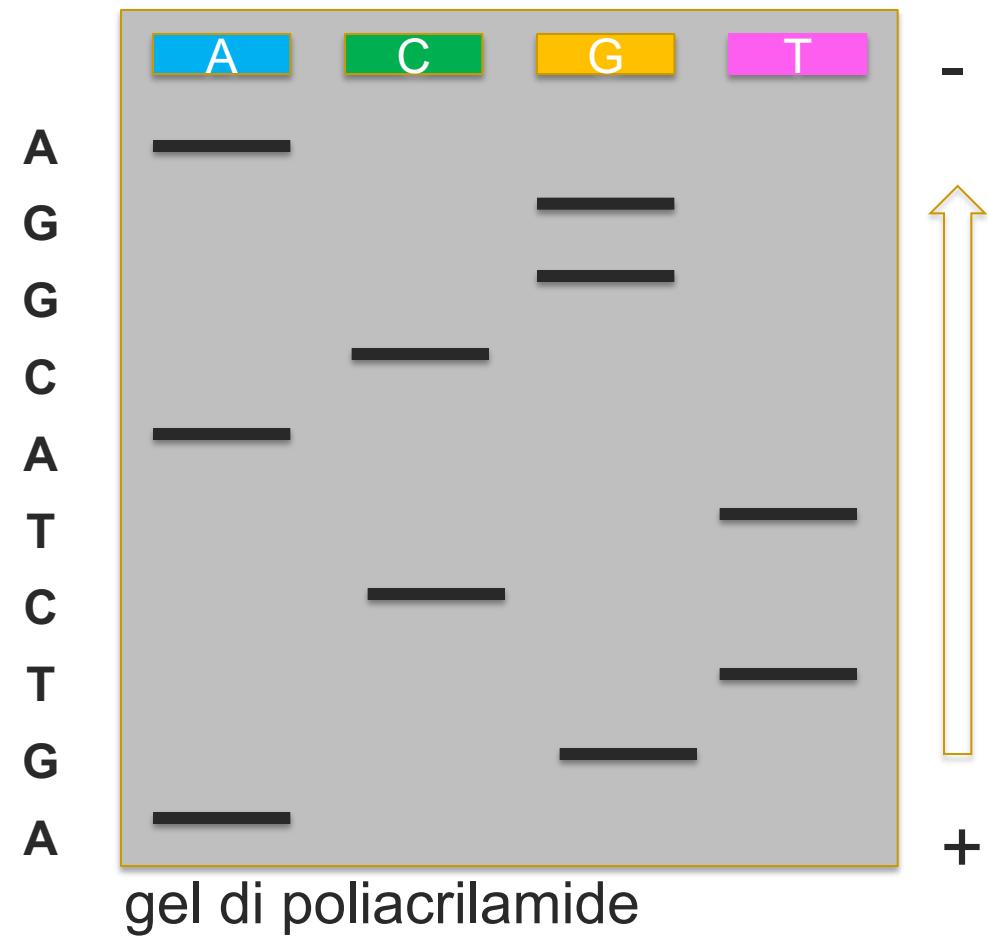
TGCA
TGCA~~G~~**G**
TGCA~~G~~**G**CATCT**A**
TGCA~~G~~**G**
TGCA~~G~~**G**
TGCA~~G~~**G**CATCT**G**
TGCA~~G~~**G**CAT**C**
TGCA~~G~~**G**C
TGCA~~G~~**G**CAT**T**
TGCA~~G~~**G**CATCT**T**



[Metodo Sanger]

Sanger → Chain-termination method

TGCA
TGCA**G**
TGCA**G**CATCT**G**
TGCA**G**
TGCA**G**
TGCA**G**CATCT**G**
TGCA**G**CAT**C**
TGCA**G**C
TGCA**G**CAT**T**
TGCA**G**CATCT**T**



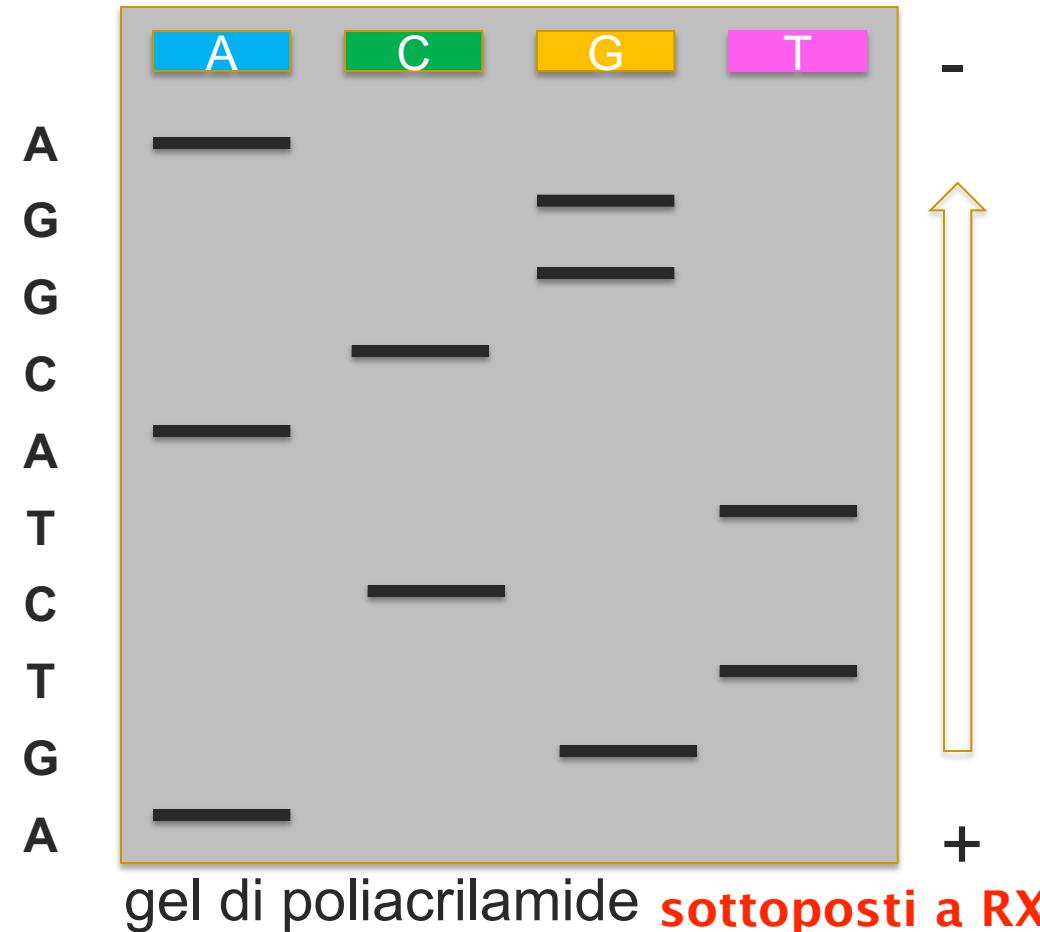


Metodo Sanger

T G C A G G C A T C T G A
| | | | | | | | | | | | | |
A C G T C C G T A G A C T

Sanger → Chain-termination method

TGCA
TGCA**GG**C
TGCA**GGC**A**T**CTGA
TGCA**G**
TGCA**GG**
TGCA**GGC**A**T**CT**G**
TGCA**GGC**A**T**C
TGCA**GGC**
TGCA**GGC**A**T**
TGCA**GGC**A**T**T



Metodo Sanger

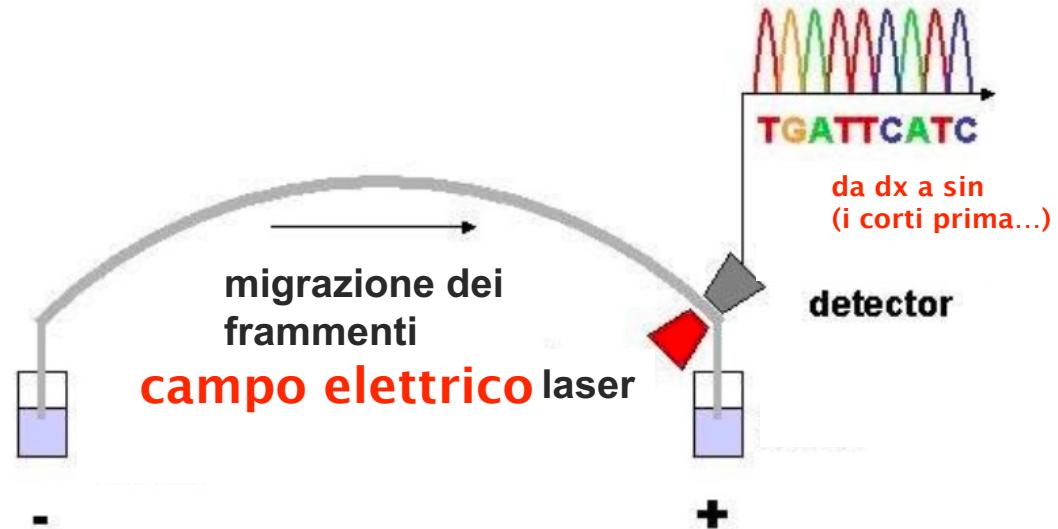
Sanger → Chain-termination method

metodo alternativo: ogni terminatore ha una sostanza fluorescente diversa

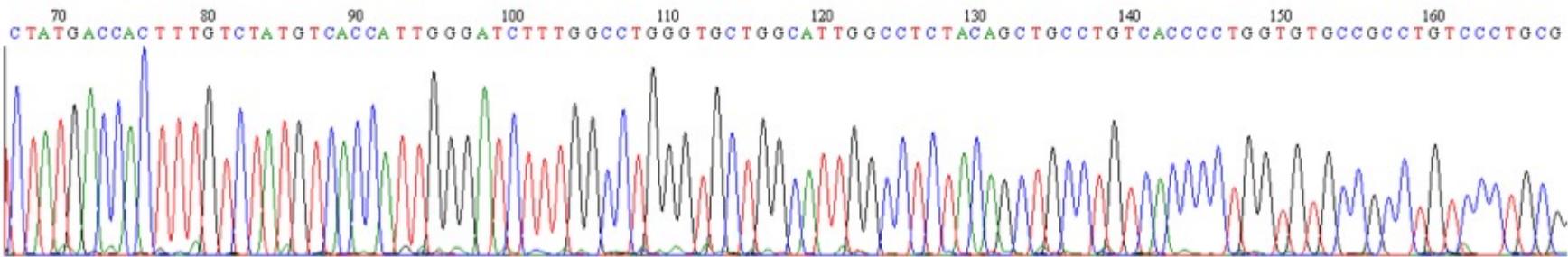
TGCA
TGCA~~G~~
TGCA~~G~~CATCTGA
TGCA~~G~~
TGCA~~G~~
TGCA~~G~~CATCTG
TGCA~~G~~CATC
TGCA~~G~~
TGCA~~G~~CAT
TGCA~~G~~CATCT

Elettroforesi capillare

1 provetta



[Metodo Sanger (automatici)]



cromatografia
(prodotta dal
sequenziatore
automaticamente)



Base Caller

software che legge la chromatografia



Sequenza primaria
del DNA template

[Metodo Sanger]

Il metodo Sanger produce un dato con:

- ✓ qualità elevata
- ✓ coverage bassa (5x-8x)
- ✓ lunghezza fino a 1000bp

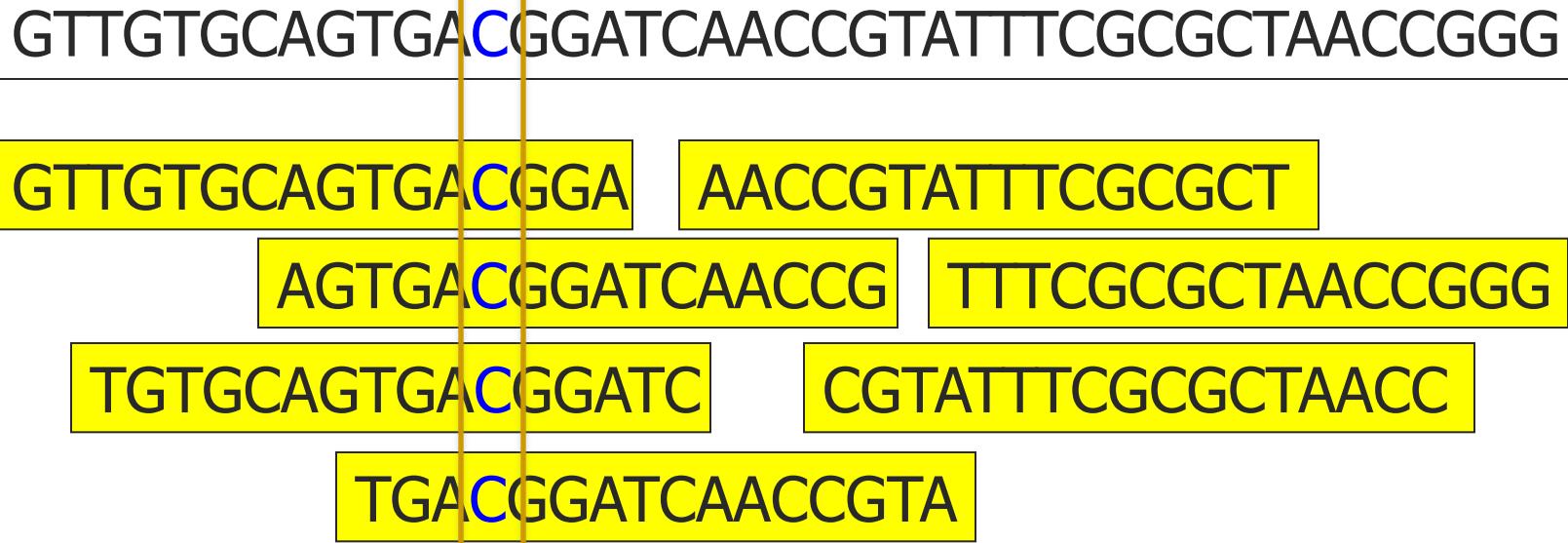


Il metodo Sanger è molto costoso (3 miliardi di dollari per HGP)

1\$ per nucleotide!

Coverage

numero medio di volte che una base viene analizzata, quindi corrisponde al numero di reads che mappano sulla base.



Rappresenta il numero di read che coprono quella base: da 5 a 8

Coverage

GTTGTGCAGTGACGGATCAACCGTATTTCGCGCTAACCGGG

GTTGTGCAGTGACGGAA AACCGTATTTCGCGCT
AGTGACGGATCAACCG TTTCGCGCTAACCGGG
TGTGCAGTGACGGATC CGTATTTCGCGCTAACCC
TGACGGATCAACCGTA

$$\text{coverage} = n / L$$

$n \rightarrow \# \text{ reads}$

$L \rightarrow \text{lunghezza dei reads}$ (fissata)

$L \rightarrow \text{lunghezza della molecola}$

[Qualità]

GTTGTGCAGTGACGGATCAACCGTATTTCGCGCTAACCGGG

sostituita

cancellazione

GTTGTGCATGTGACGGA

AACCGTATTTCGG-GCT

AGTGACGGATCAACCG

TTTCGCGCTAACCGGG

TGTGCAGTGACTGGATC

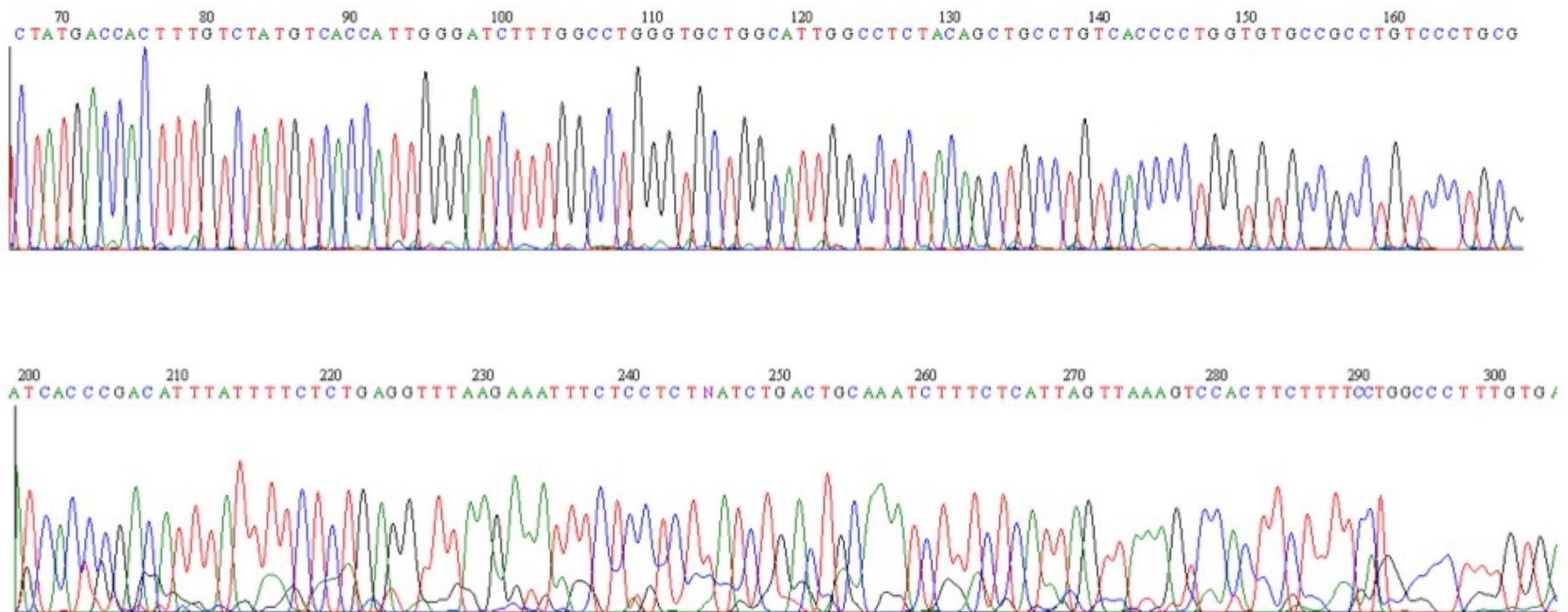
CGTATTTCGCGCTAACCC

T inserita

TGACGGATCAACGGTA

se ho una coverage alta, posso smorzare gli effetti di una bassa qualità

Qualità



[Tecnologie NGS]

- ✓ Tecnologie ideate a partire dagli anni 2000
- ✓ Tecnologie altamente parallele (high-throughput)
- ✓ Produce un dato con:
 - ✓ qualità variabile
 - ✓ coverage elevatissima
 - ✓ lunghezza fino a 300bp
- ✓ Veloci e poco costose (1500\$ per sequenziare un genoma) In un paio di giorni...
milioni di frammenti
(anche che coprono
33 genomi umani...)

Come sono sequenziate le reads?



<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

[Tecnologie NGS]

Illumina (Solexa)

- HiSeq System
- Genome analyzer ix
- MySeq

Ion Torrent - Life technologies

- Personal Genome Machine (PGM)
- Proton

[Tecnologie NGS]

Illumina (Solexa) → coverage: 50x-80x

- HiSeq System
- Genome analyzer Iix
- MySeq → 300bp

Ion Torrent - Life technologies:

- Personal Genome Machine (PGM) → 200 bp
- Proton

[Tecnologie NGS]

Illumina (Solexa) → coverage: 50x-80x

- HiSeq System
- Genome analyzer Iix
- MySeq → 25M reads/run

simil Sanger

Ion Torrent - Life technologies:

- Personal Genome Machine (PGM) → 11M reads/run
- Proton

altra tecnica

[Tecnologie Next-NGS]

> 1000 bp

Long Reads

Pacific Biosciences → lunghezza: 10-15 Kb

(quasi tutto
un trascritto)

- PacBio RS

Oxford Nanopore Technologies → lunghezza: 5-10 Kb

- GridION System
- MinION

[Tecnologie Next-NGS]

> 1000 bp

Pacific Biosciences → 10-15% di errore

- PacBio RS

Oxford Nanopore Technologies → 10-30% di errore

- GridION System
- MinION

[Tecnologie NGS e Next-NGS]

Conseguenze importanti:

- ✓ esplosione dei dati → Sequence Read Archive (SRA)
- ✓ nuovi problemi computazionali:
 - quantificazione dei trascritti
 - predizione di nuovi eventi di splicing
 - filogenesi tumorale
 - aplotipizzazione
 - variant calling
 - rappresentazione di più genomi → pan-genoma

→ nuovi algoritmi e tools