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**

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

- **Fragment Assembly e approcci graph-based:**
  - **Overlap Graph (OG) e de Brujin Graph (dBG)**

Sequenziamento del DNA

- **Il dato di sequenziamento**
  - **Tecnologie di sequenziamento:**
    - **SANGER sequencing**
    - **Next-Generation Sequencing (NGS)**
  - **Qualità del dato di sequenziamento**
    - **Formato Standard FASTQ**

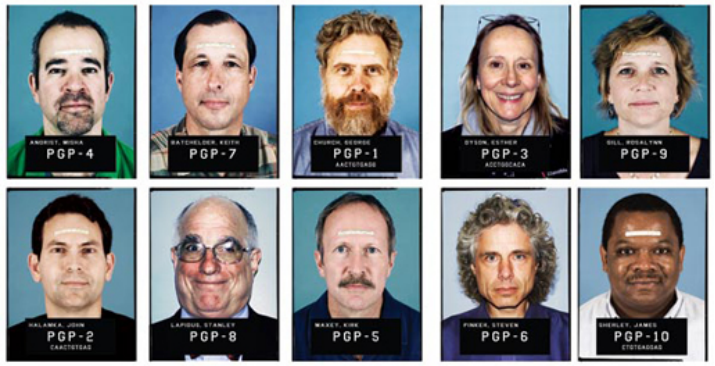
- **Fragment Assembly e approcci graph-based:**
  - **Overlap Graph (OG) e de Brujin Graph (dBG)**

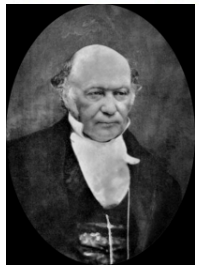
- Strutture di indicizzazione (BWT, SA, LCP)
- Bloom Filter per conteggio dei k-mer [& tecniche alignment-free]
- Problemi di spazio nella rappresentazione e calcolo overlap
- BWT+LCP+SA per costruzione lineare di OG
- ~~Pan-genome graph~~
- Fattorizzazione di Lyndon e varianti per: SA, BWT, signature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

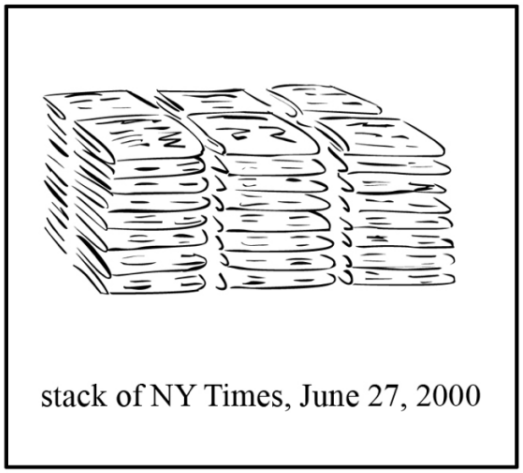
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGAT
CAGCTACAACATCGTAGCTACGATGCATTAGCAAGCTATCG
ATCGATCGATCGATTATCTACGATCGATCGATCGATCACTA
TACGAGCTACTACGTACGTACGATCGCGTGACTATTATCGA
CTACAGATGAAACATGCTAGTACAACAGTATACATAGCTGC
GGGATACGATTAGCTAATAGCTGACGATATCCGAT

Why Do We Sequence Personal Genomes?

- **2010:** Nicholas Volker became first human being to be saved by genome sequencing.

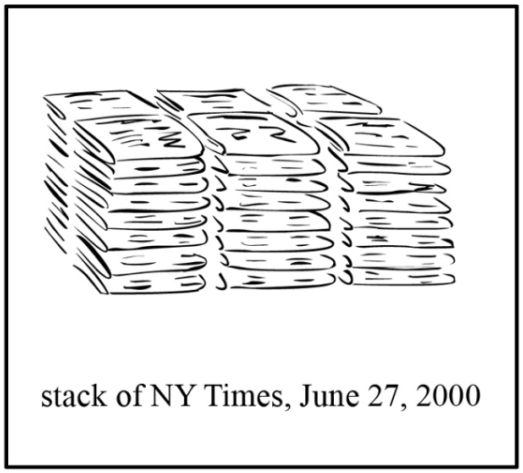
  - Doctors could not diagnose his condition; he went through dozens of surgeries.
  - Sequencing revealed a rare mutation in a *XIAP* gene linked to a defect in his immune system.
  - This led doctors to use immunotherapy, which saved the child.

The Newspaper Problem

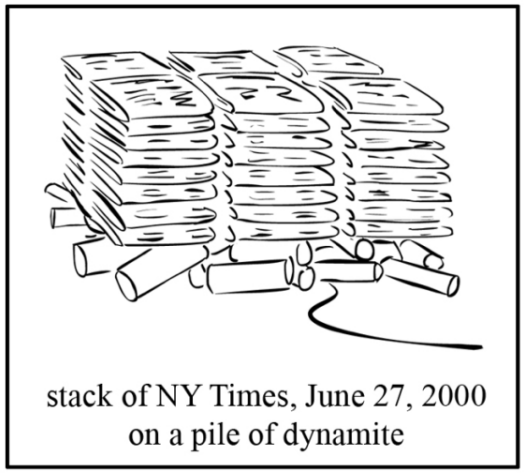


stack of NY Times, June 27, 2000

The Newspaper Problem

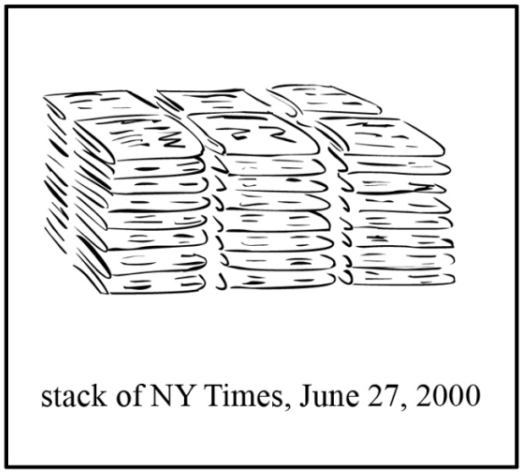


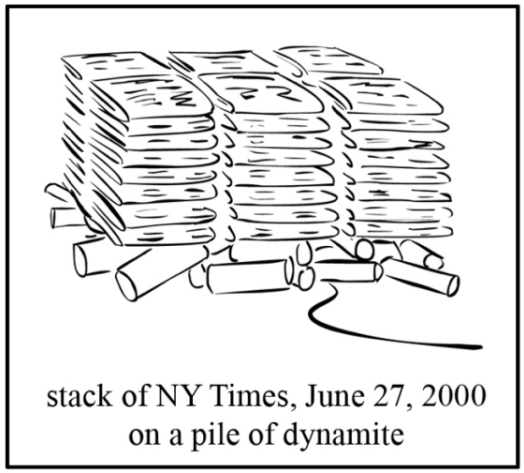
stack of NY Times, June 27, 2000

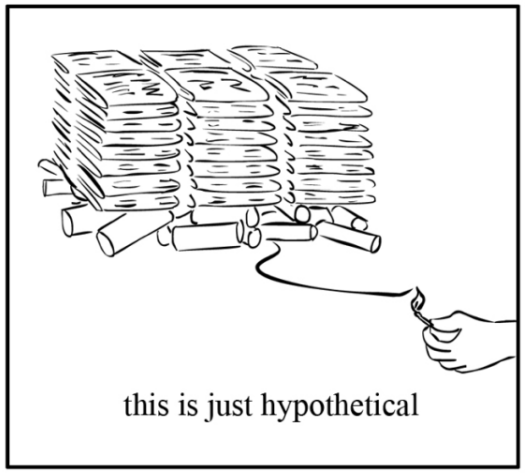


stack of NY Times, June 27, 2000
on a pile of dynamite

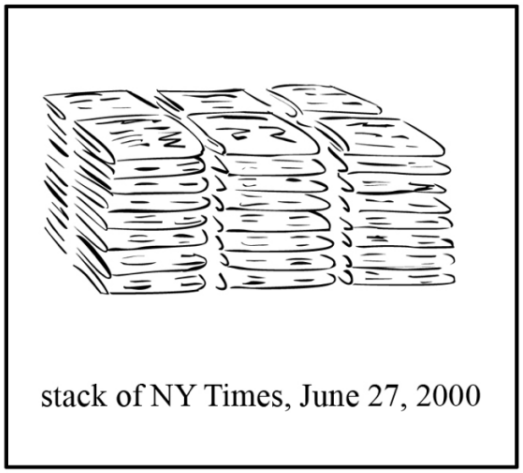
The Newspaper Problem

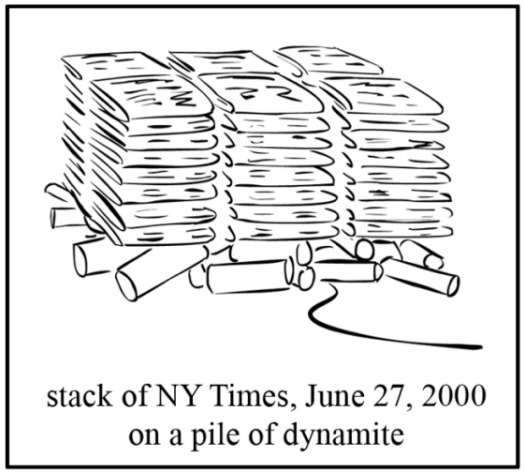


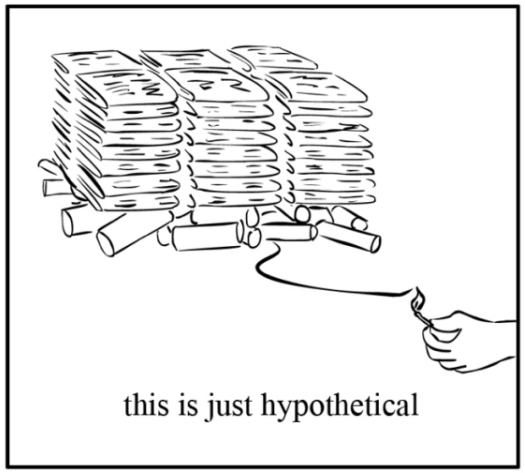


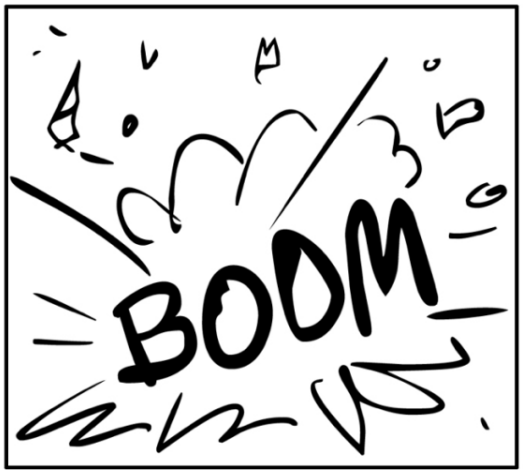


The Newspaper Problem

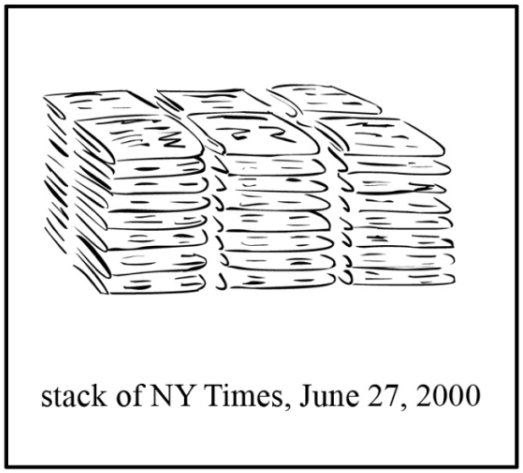


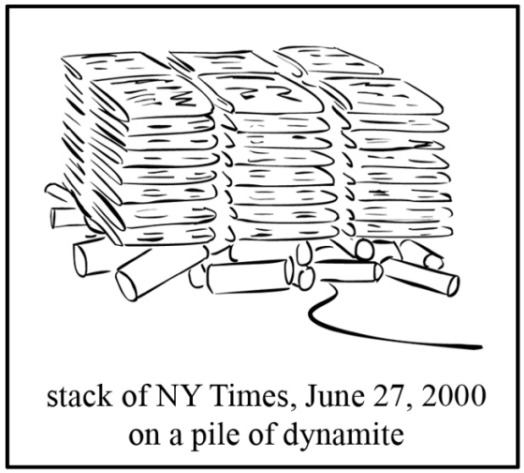


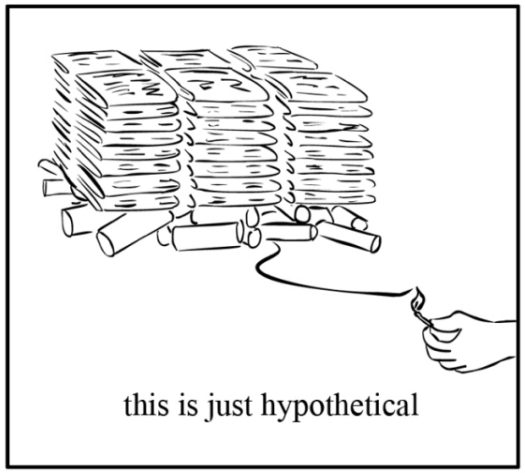


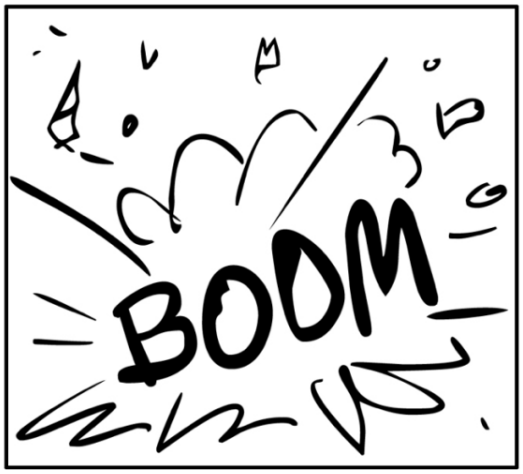


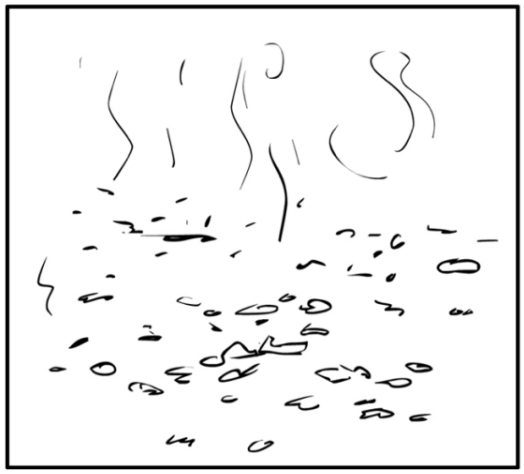
The Newspaper Problem



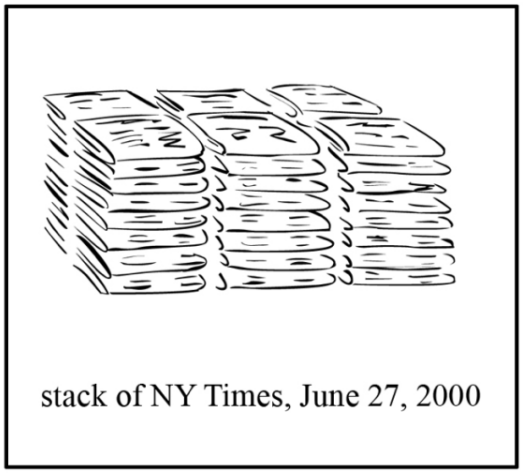


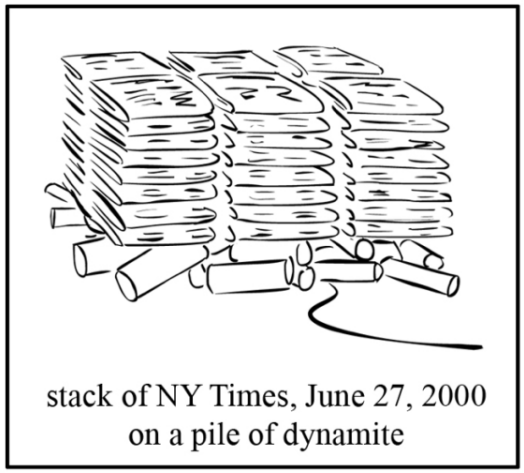


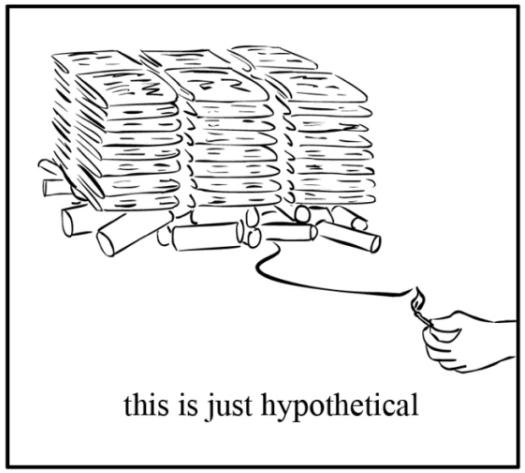


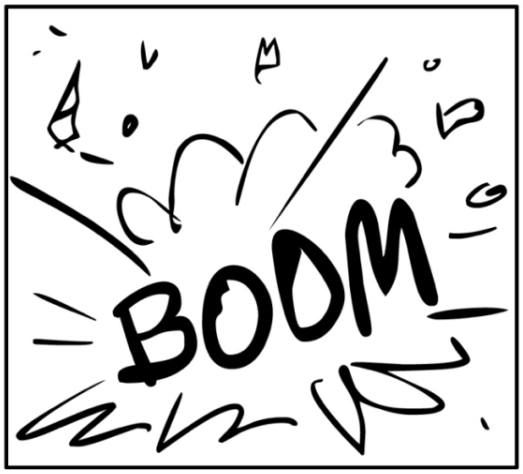


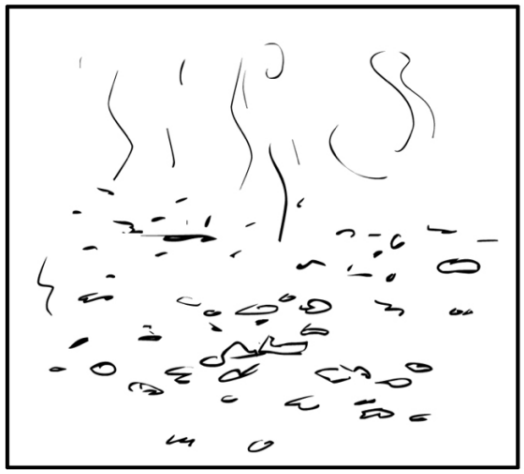
The Newspaper Problem

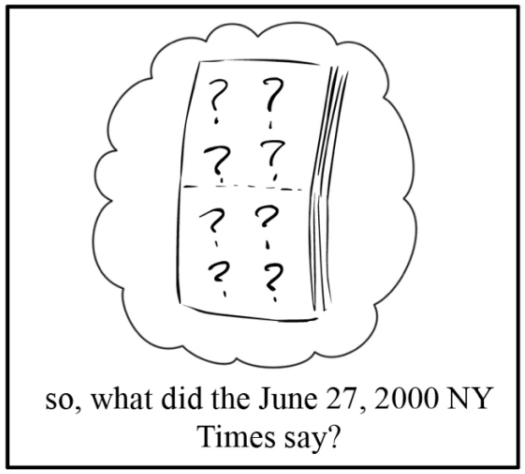




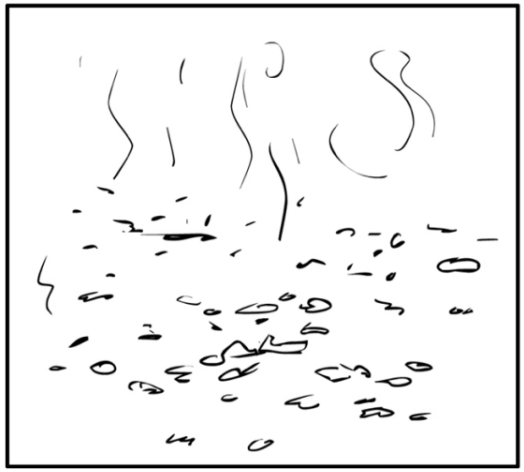


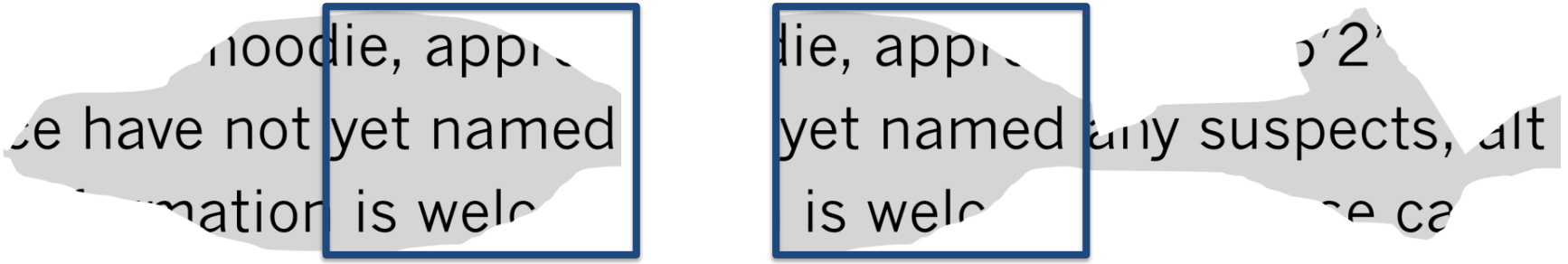






The Newspaper Problem as an Overlapping Puzzle





The Newspaper Problem as an Overlapping Puzzle

hoodie, appr
e have not yet name
mation is welc
2'
spects, alt
e ca

Multiple Copies of a Genome (Millions of them)

stack of NY Times, June 27, 2000

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

Breaking the Genomes at Random Positions

CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC
CTGATGATGGACTACGCTACTACTGCTAGCTGTATTACGATCAGCTACCACATCGTAGCTACGATGCATTAGCAAGCTATCGGATCAGCTACCACATCGTAGC

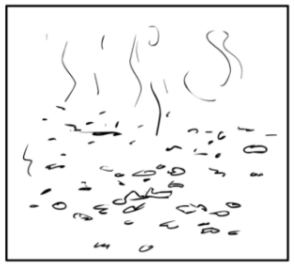
Generating “Reads”

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 anayzed in silico

“Burning” Some Reads



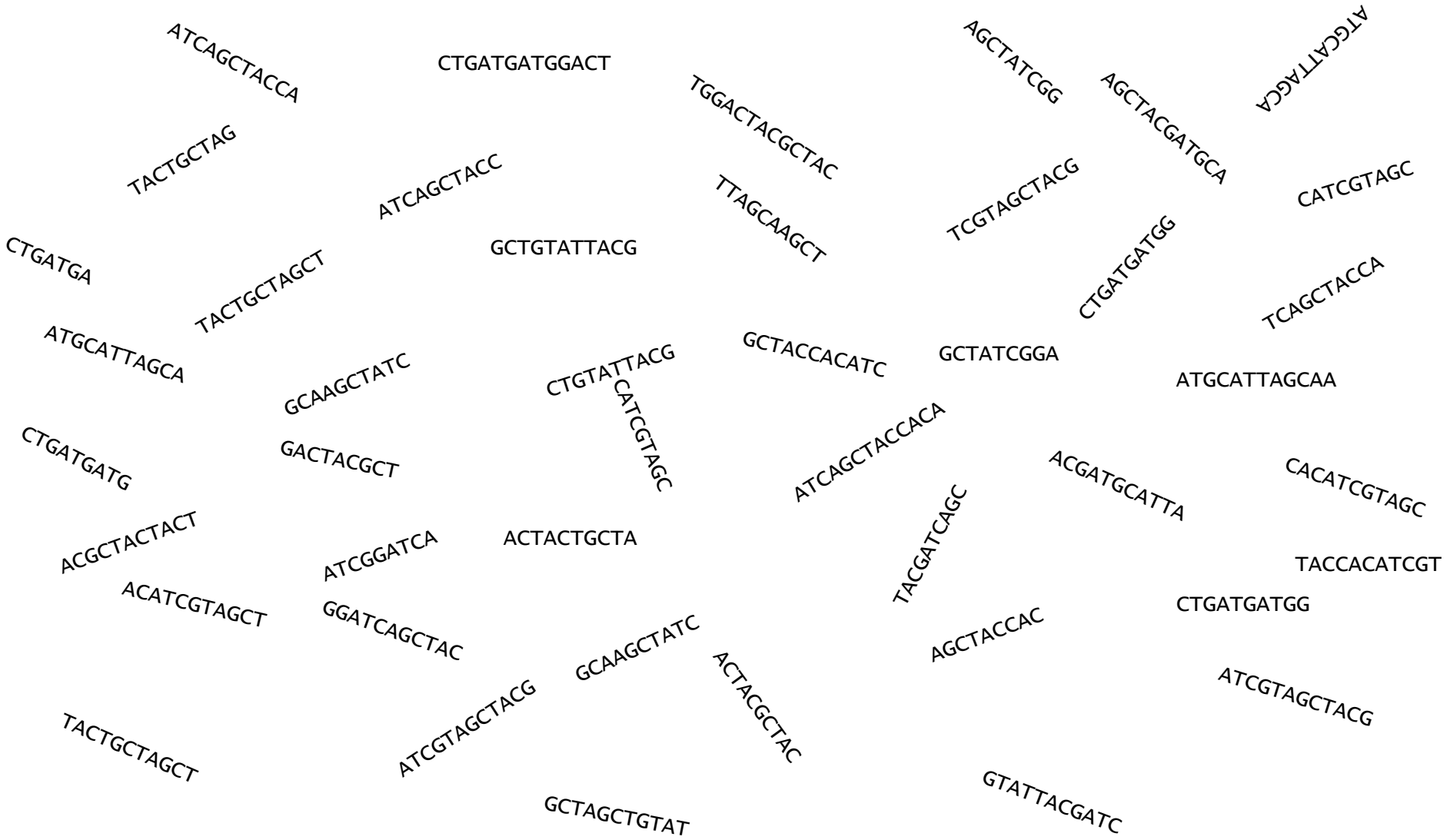
CTGATGA TGGACTACGCTAC TACTGCTAG CTGTATTACG ATCAGCTACCACA TCGTAGCTACG ATGCATTAGCAA GCTATCGGA TCAGCTACCA CATCGTAGC

CTGATGATG GACTACGCT ACTACTGCTA GCTGTATTACG ATCAGCTACC ACATCGTAGCT ACGATGCATTA GCAAGCTATC GGATCAGCTAC CACATCGTAGC

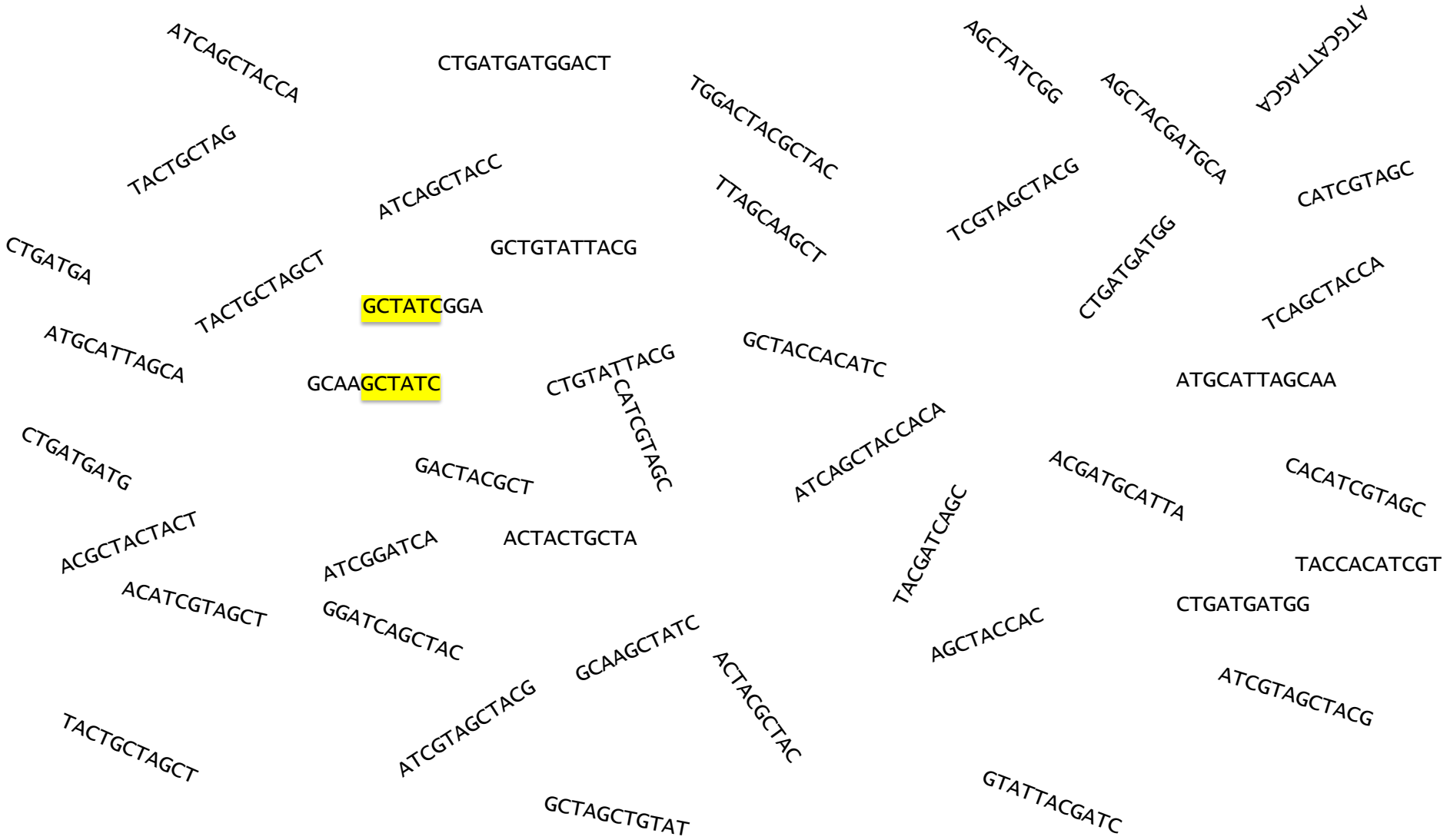
CTGATGATGG ACTACGCTAC TACTGCTAGCT GTATTACGATC AGCTACCAC ATCGTAGCTACG ATGCATTAGCA AGCTATCGG A TCAGCTACCA CATCGTAGC

CTGATGATGGACT ACGCTACTACT GCTAGCTGTAT TACGATCAGC TACCACATCGT AGCTACGATGCA TTAGCAAGCT ATCGGATCA GCTACCACATC GTAGC

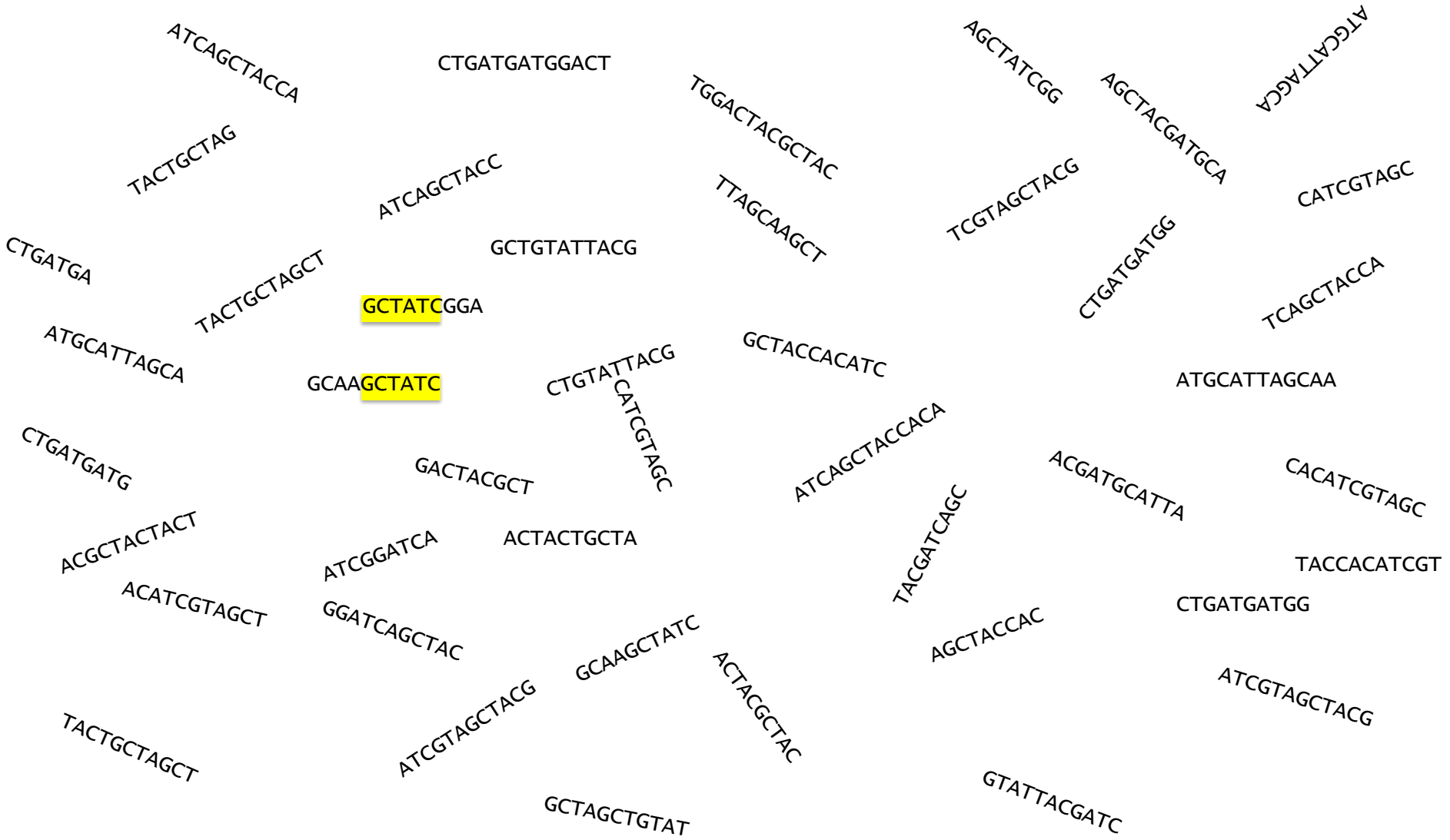
No Idea What Position Every Read Comes From



No Idea What Position Every Read Comes From

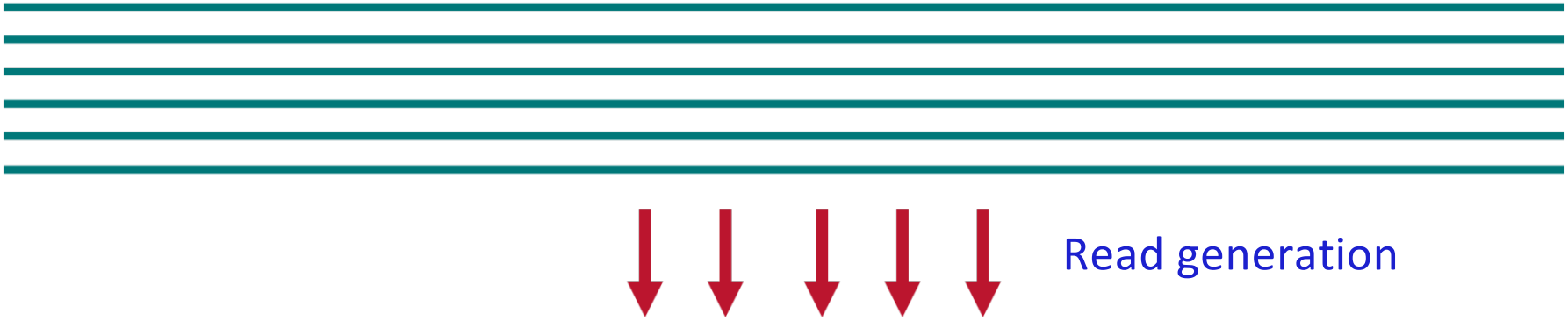


No Idea What Position Every Read Comes From

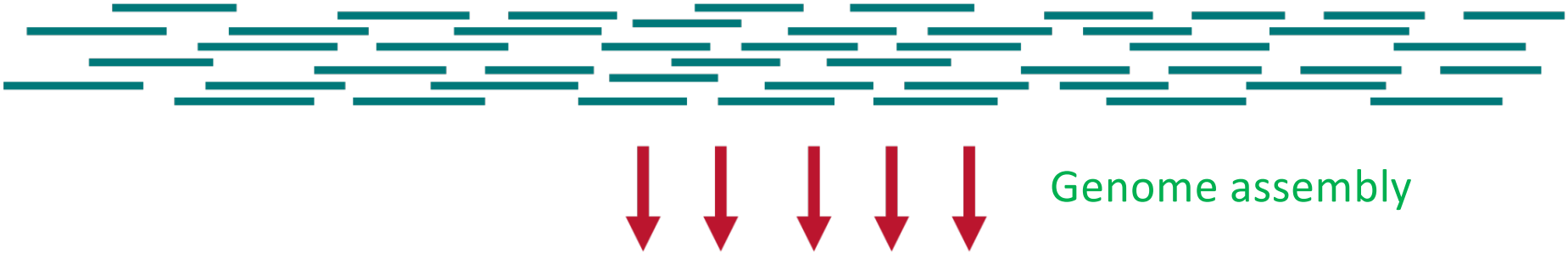


From Experimental to Computational Challenges

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



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

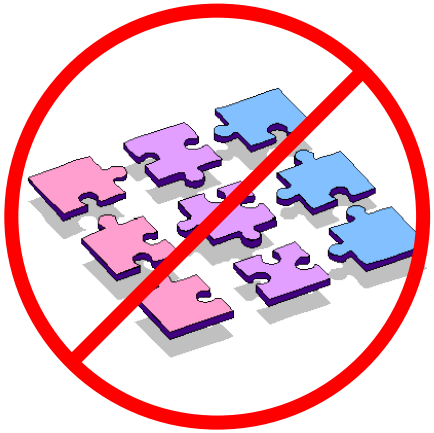


Assembled genome

...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

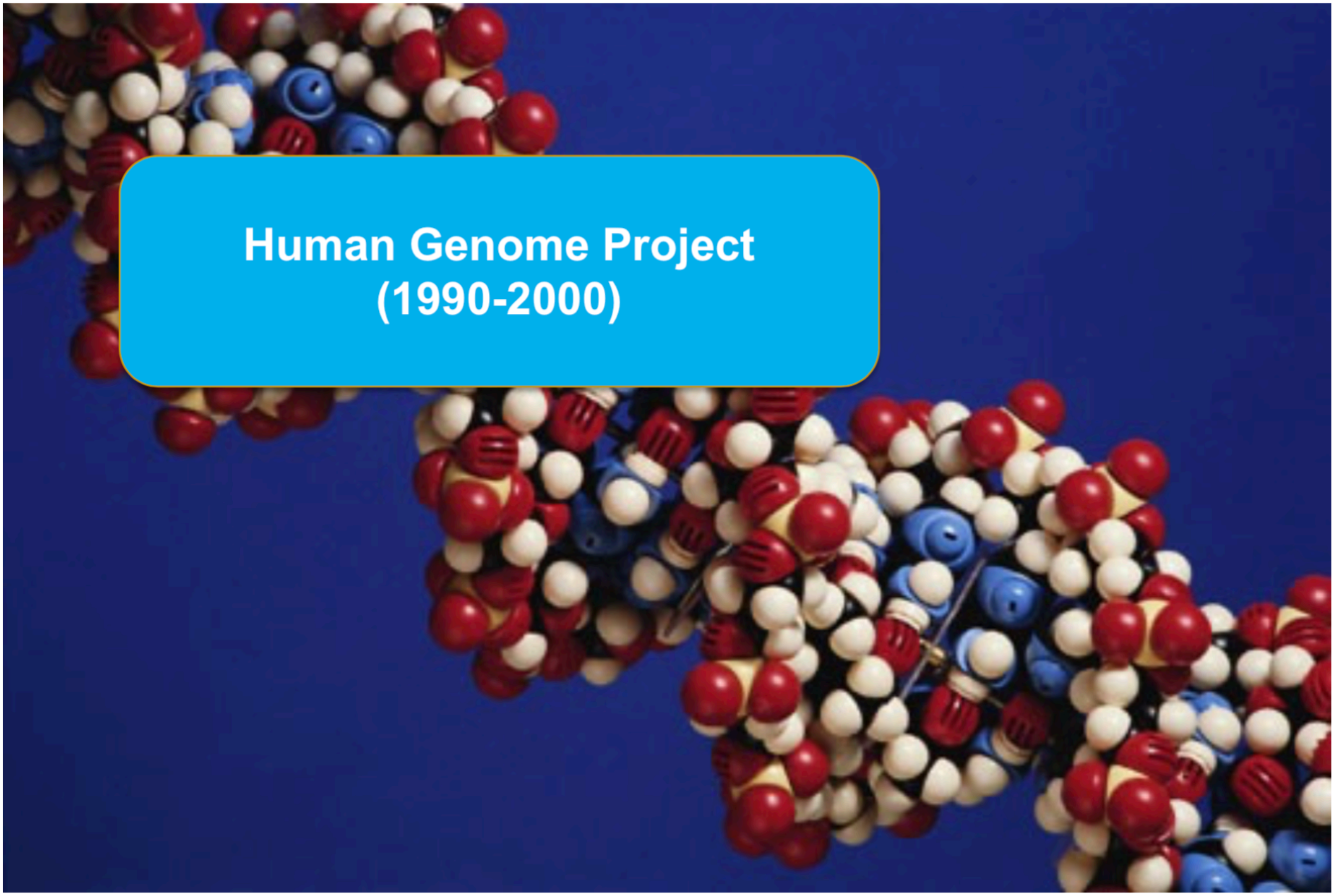
➔ sequenza delle basi {A,C,G,T|U} per DNA e RNA

➔ sequenza degli amminoacidi per le proteine

Due step:

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

Shotgun Sequencing



Shotgun Sequencing→ reads

each fragment is individually sequenced —> primary structure

Fragment Assembly

by using overlaps…

Single-end read

Primo tipo di dato ottenuto dal sequenziamento

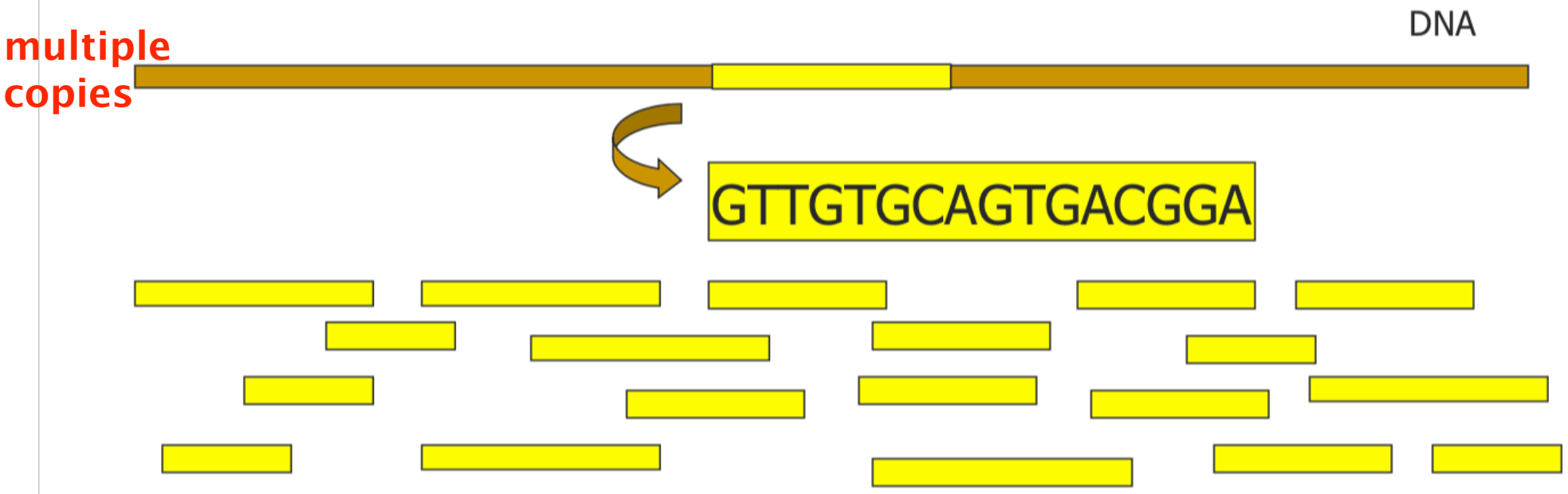
Cosa si ottiene da un esperimento di sequenziamento di una molecola di DNA (o RNA)?

➔ single-end *read*

Single-end read

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

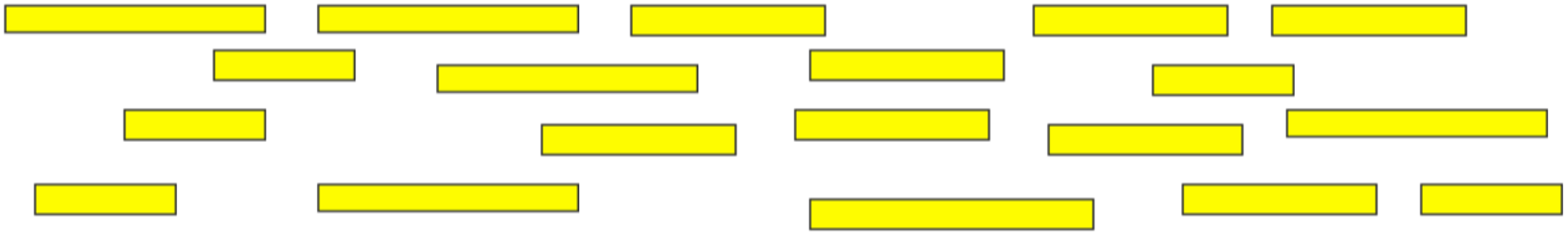
➔ single-end *read*



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



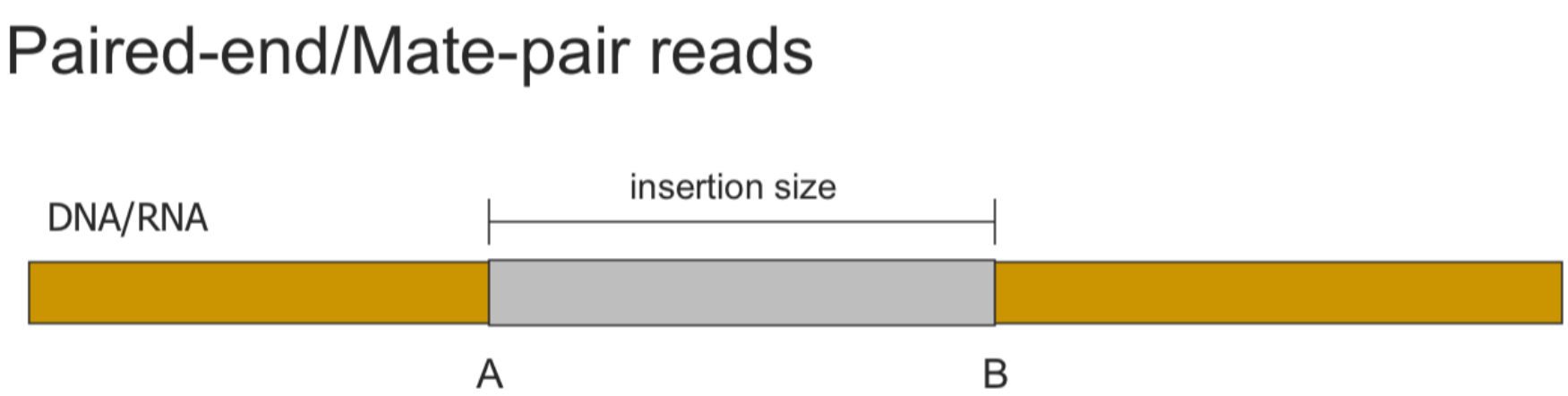
Single-end read

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

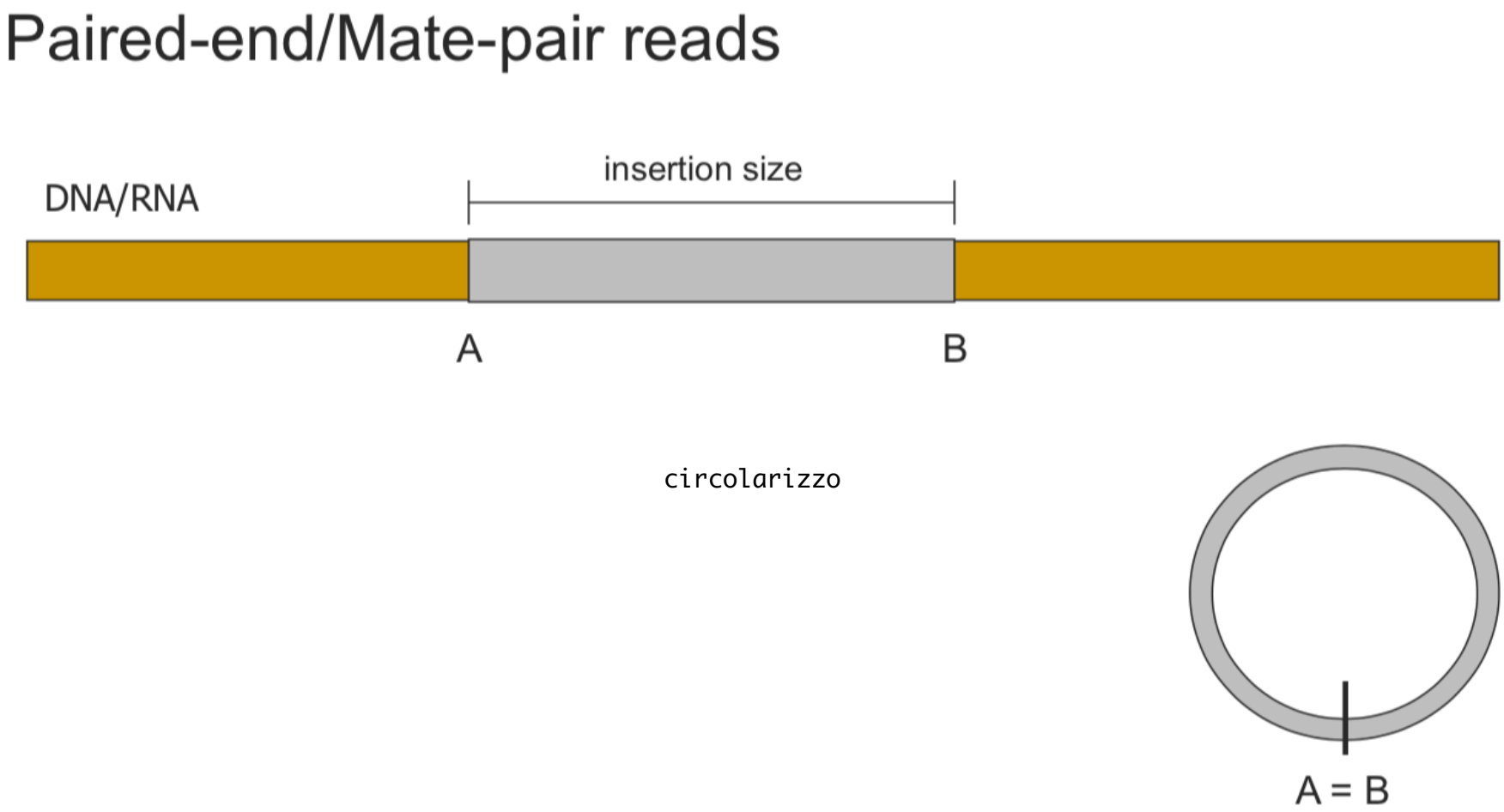
Pair-end / Mate-pair end

Altro tipo di dato ottenuto dal sequenziamento

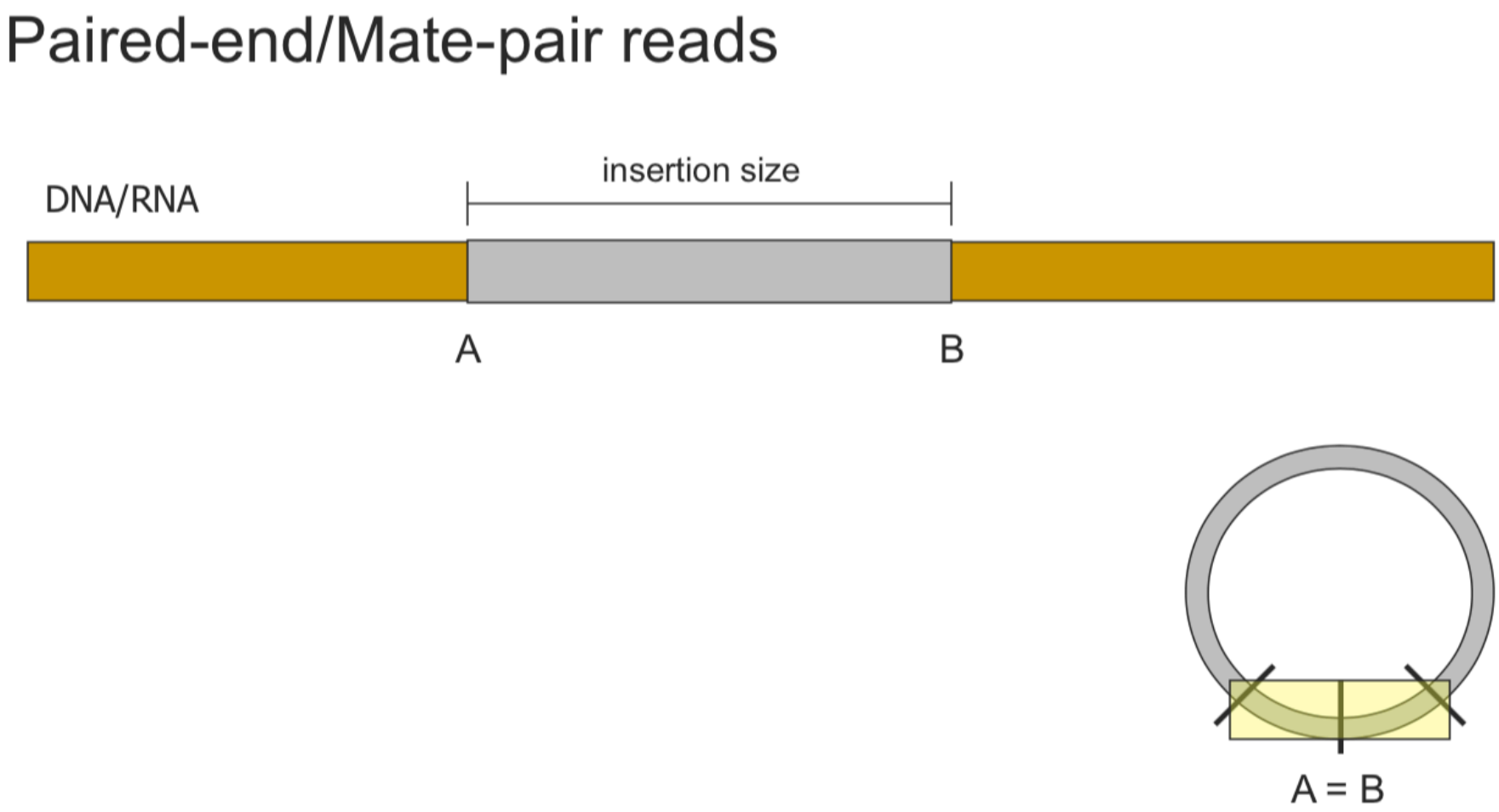


Considero la sottostringa tra A e B
(questo sarà sequenziato)

Pair-end / Mate-pair end

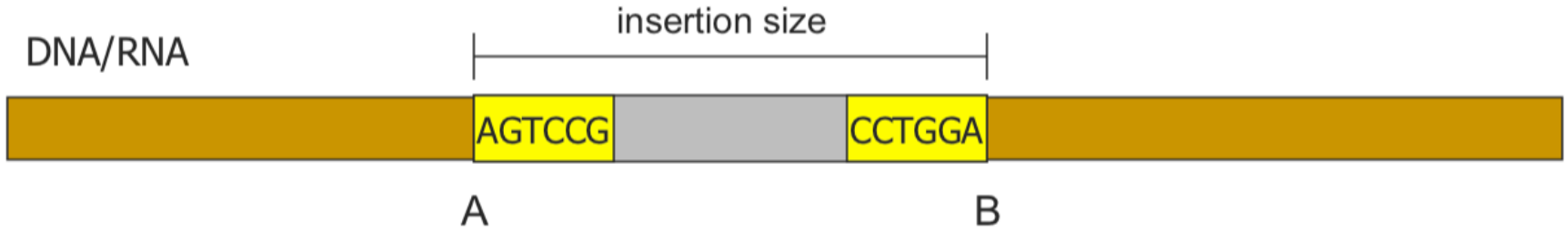


Pair-end / Mate-pair end

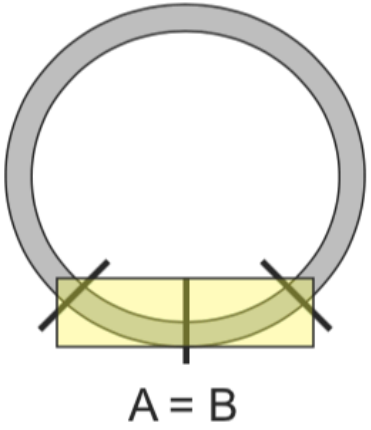


Pair-end / Mate-pair end

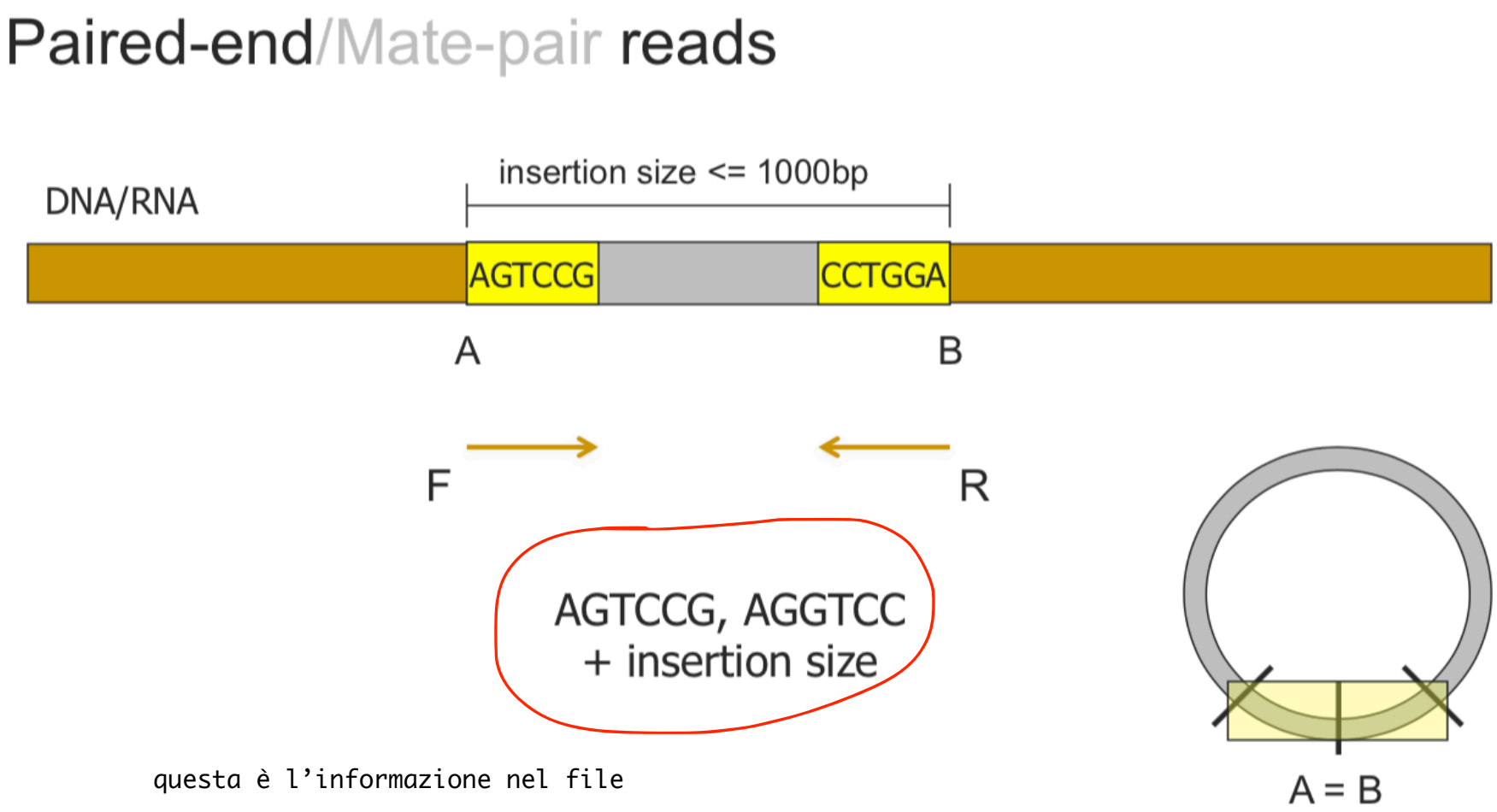
Paired-end/Mate-pair reads



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

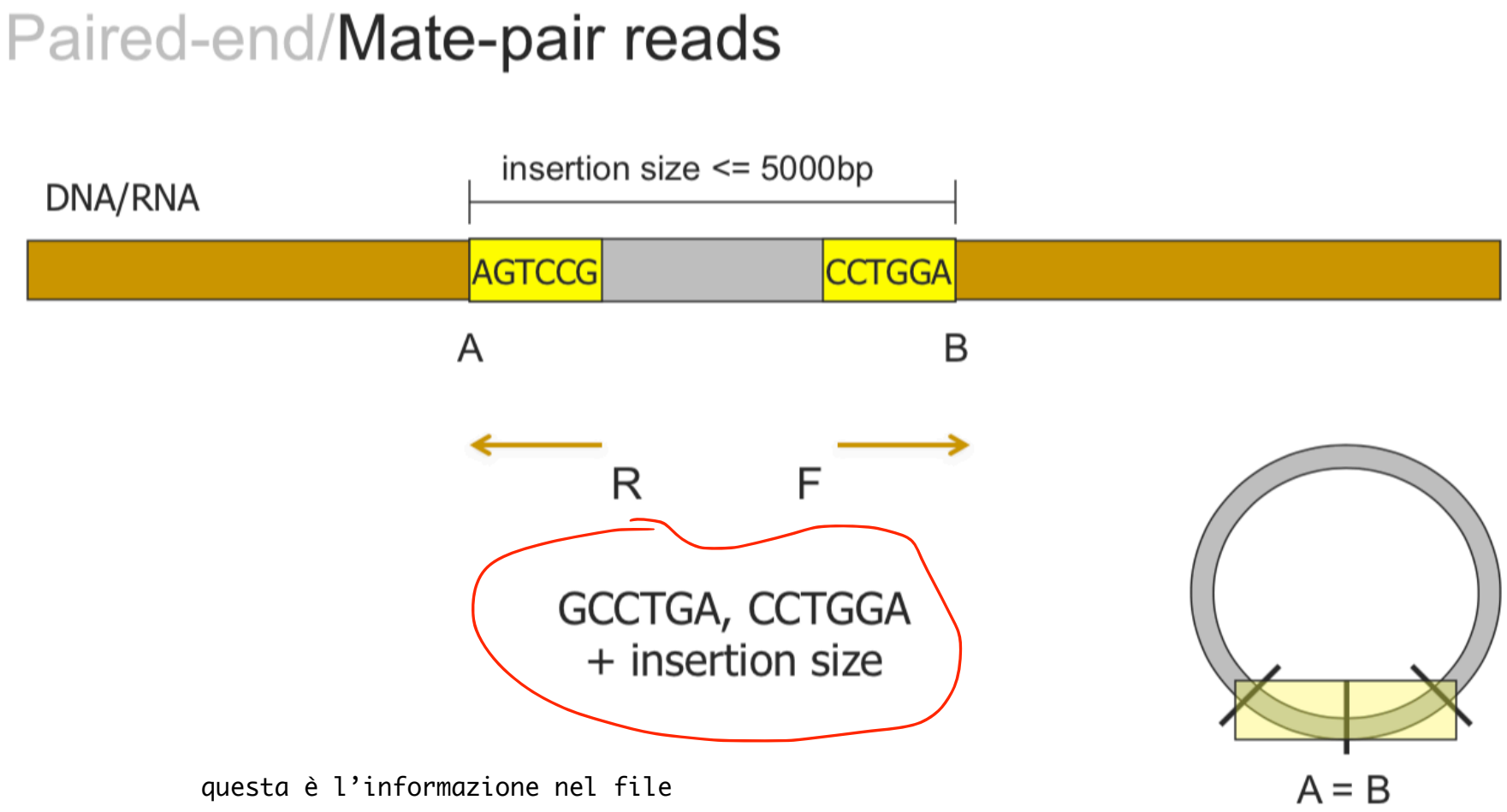


Pair-end / Mate-pair end



questa è l'informazione nel file

Pair-end / Mate-pair end



questa è l'informazione nel file

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

sequenza da sequenziare
(non la conosco)

tante copie

Metodo Sanger

Sanger → Chain-termination method

Componenti:

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

tante copie

Metodo Sanger

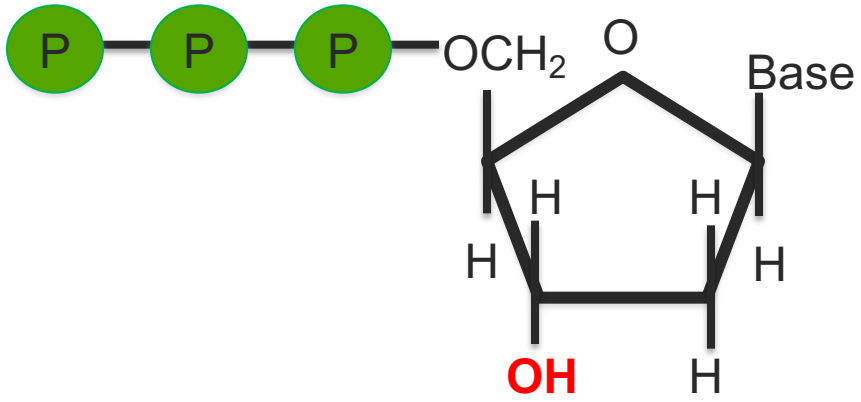
Sanger → Chain-termination method

Componenti:

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

Metodo Sanger

Sanger → Chain-termination method

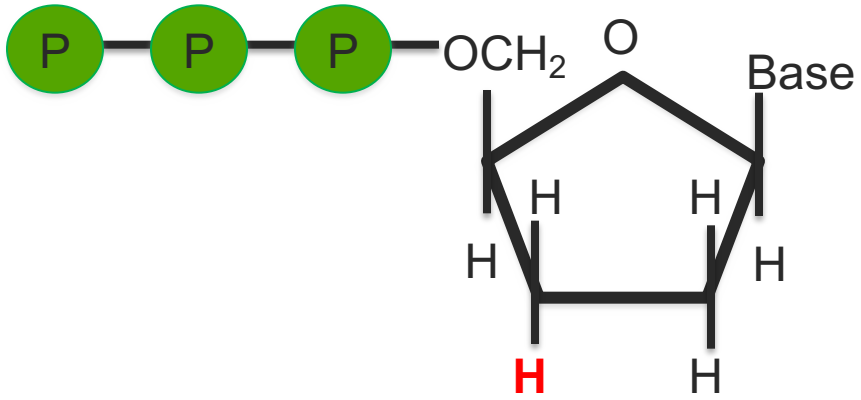


Componenti:

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

Metodo Sanger

Sanger → Chain-termination method



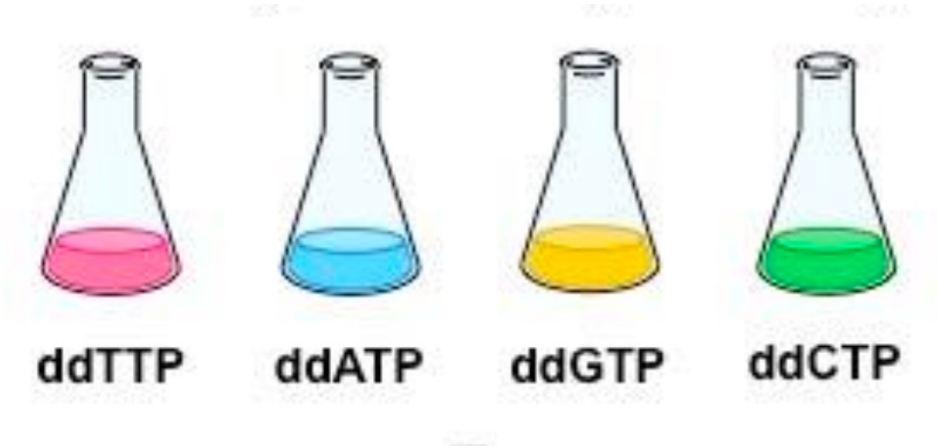
Componenti:

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

Metodo Sanger

Sanger → Chain-termination method

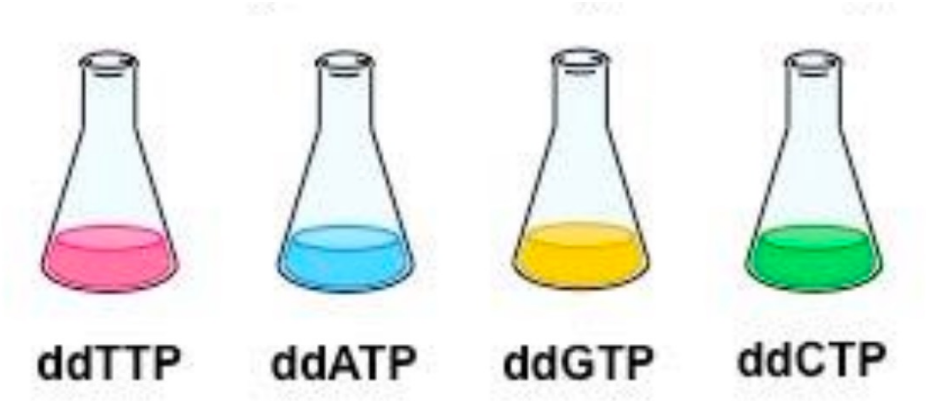
<u>Procedimento</u>:

- ✓ 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

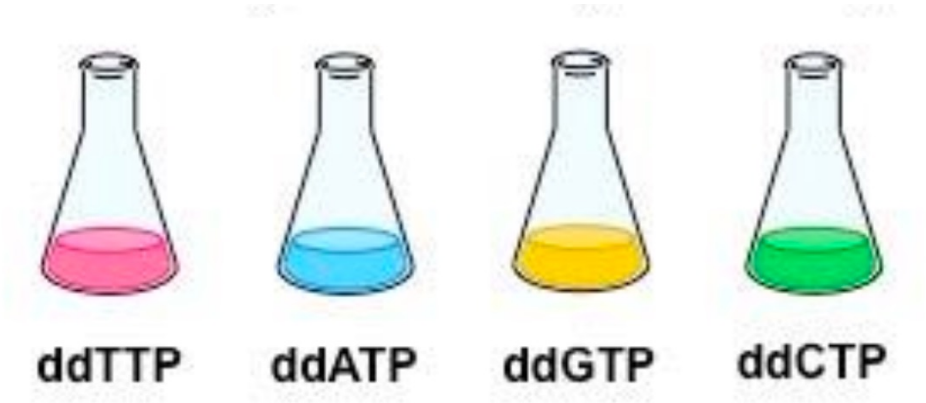
<u>Procedimento</u>:

- ✓ 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:

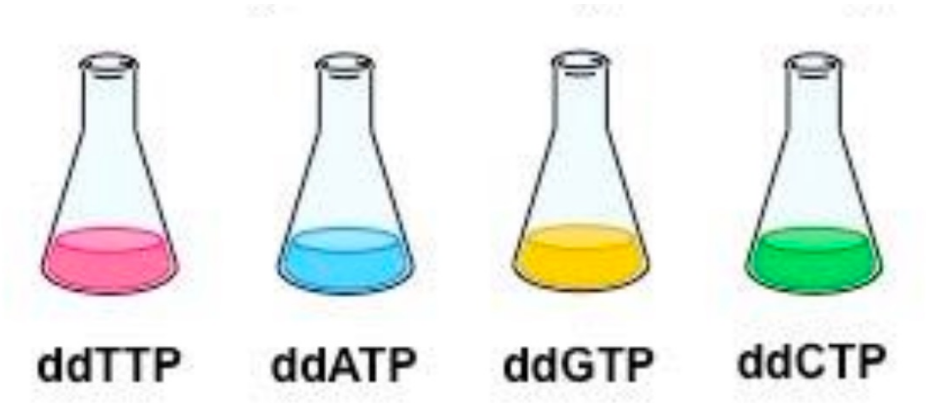
- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

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:

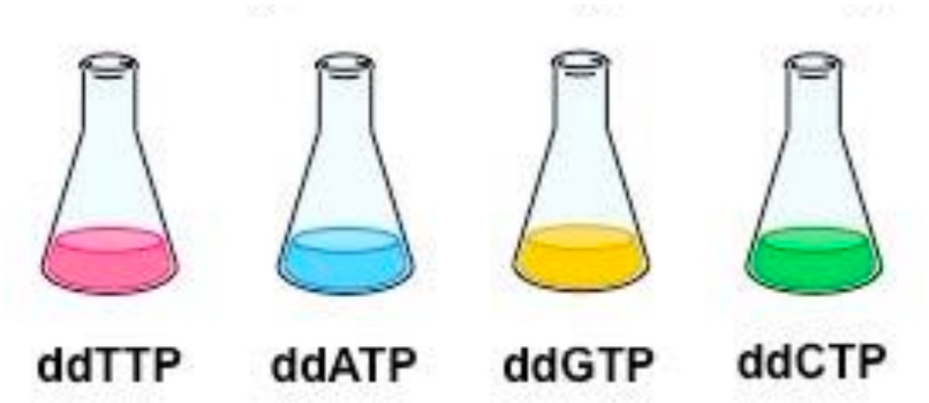
- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

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

T 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:

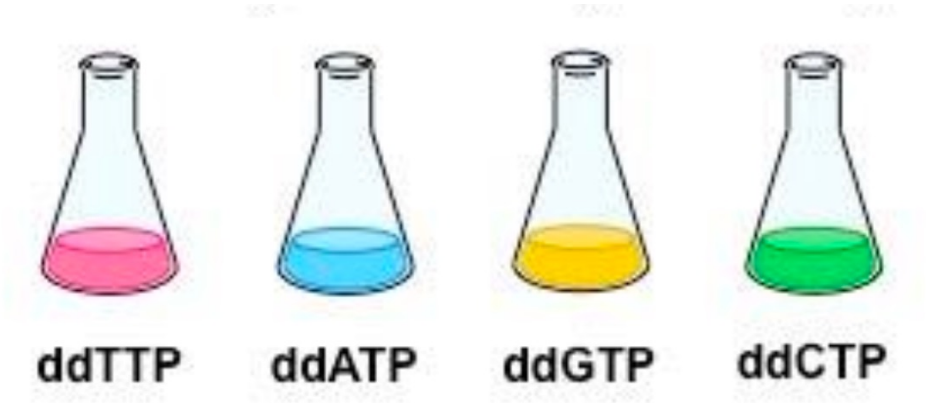
- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

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:

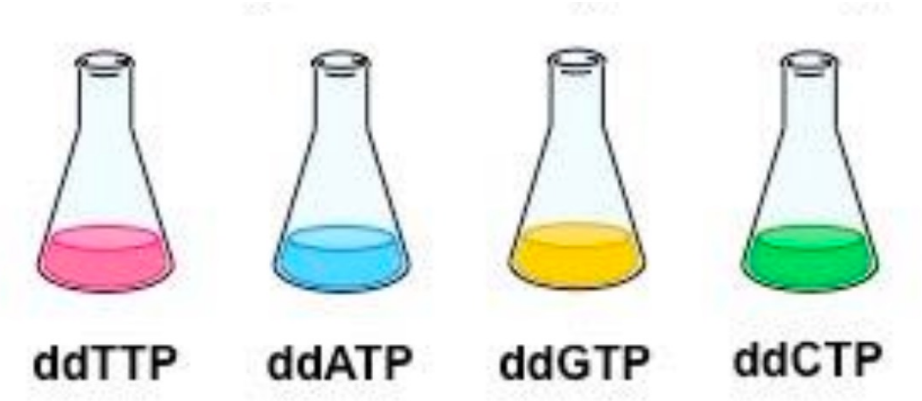
- ✓ Denaturazione
- ✓ Annealing
- ✓ Elongazione
- ✓ Terminazione (con ddNTP)

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

T G C A G G C A

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

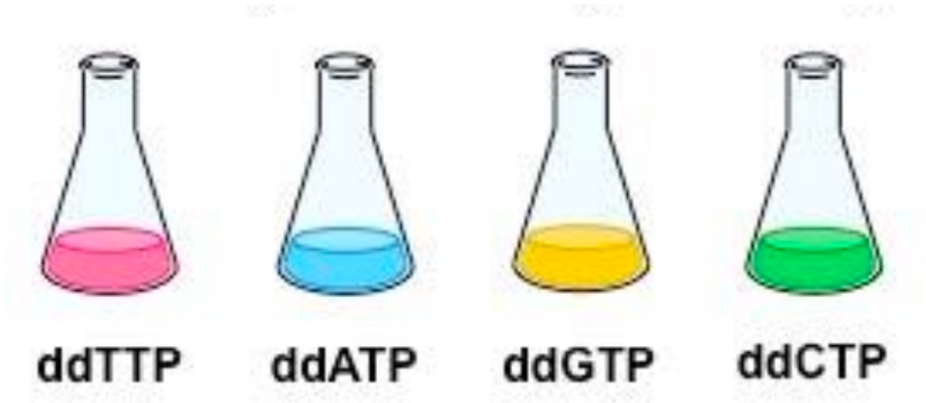
A C G T C C G T A G A C T



Metodo Sanger

Sanger → Chain-termination method

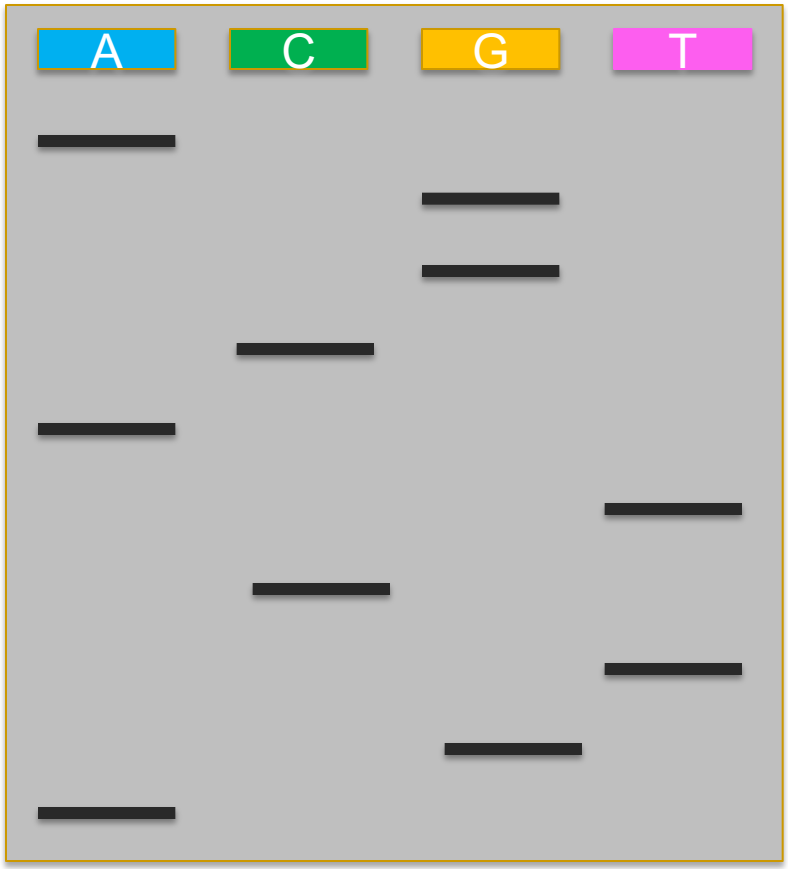
TGCA
TGCAGGCA
TGCAGGCATCTGA
TGCAG
TGCAGG
TGCAGGCATCTG
TGCAGGCATC
TGCAGGC
TGCAGGCAT
TGCAGGCATCT



Metodo Sanger

Sanger → Chain-termination method

TGCA
TGCAGGCA
TGCAGGCATCTGA
TGCAG
TGCAGG
TGCAGGCATCTG
TGCAGGCATC
TGCAGGC
TGCAGGCAT
TGCAGGCATCT

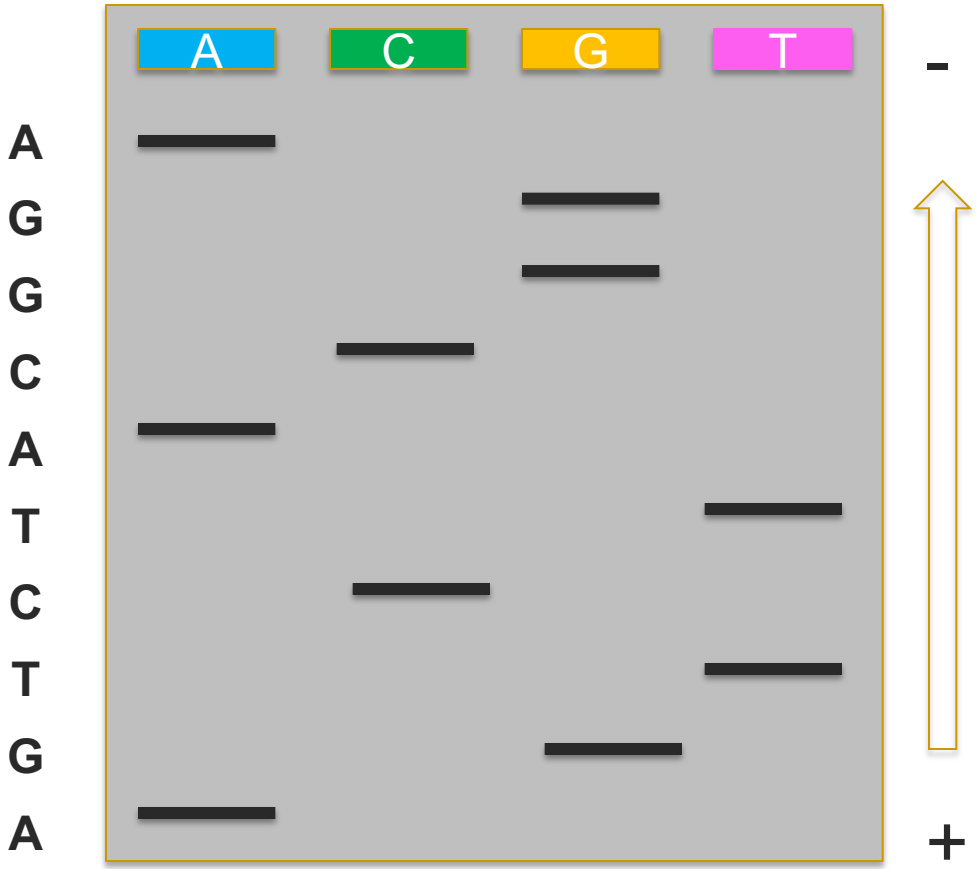


gel di poliacrilamide

Metodo Sanger

Sanger → Chain-termination method

TGCA
TGCAGGCA
TGCAGGCATCTGA
TGCAG
TGCAGG
TGCAGGCATCTG
TGCAGGCATC
TGCAGGC
TGCAGGCAT
TGCAGGCATCT



gel di poliacrilamide

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

T G C A

T G C A G G C A

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

T G C A G

T G C A G G

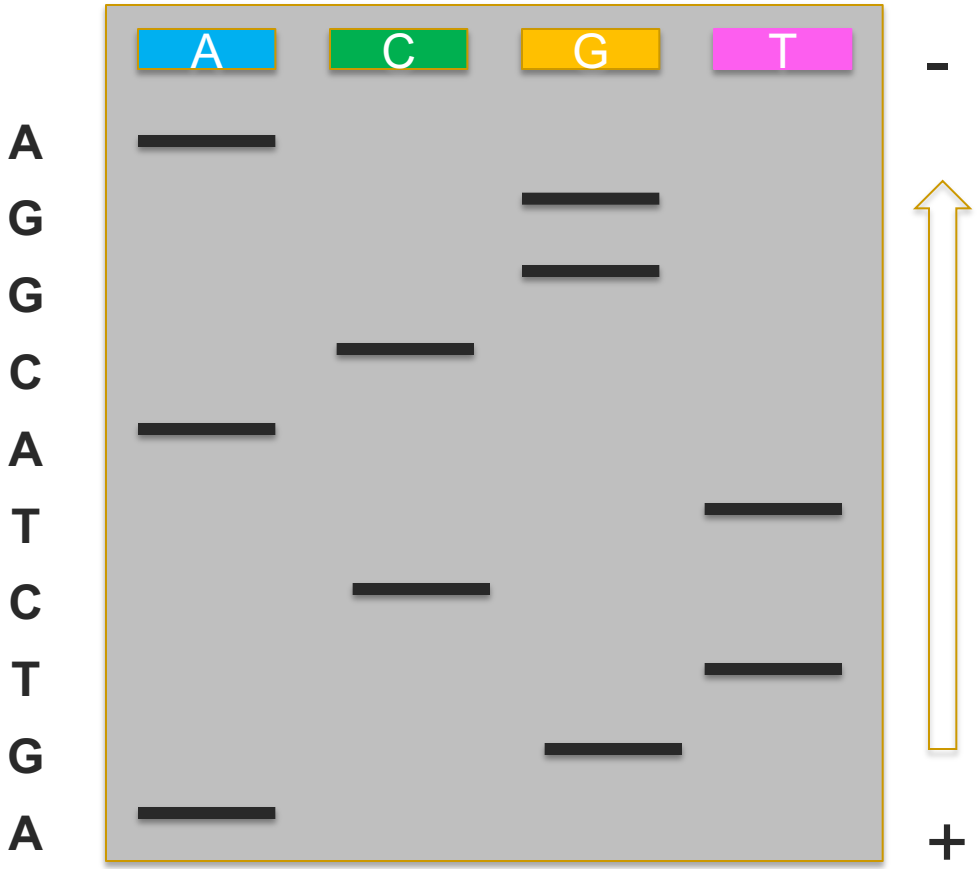
T G C A G G C A T C T G

T G C A G G C A T C

T G C A G G C

T G C A G G C A T

T G C A G G C A T C T



gel di poliacrilamide **sottoposti a RX**

Metodo Sanger

Sanger → Chain-termination method

metodo alternativo: ogni terminatore ha una sostanza fluorescente diversa

T G C A

T G C A G G C A

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

T G C A G

T G C A G G

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

T G C A G G C A T C

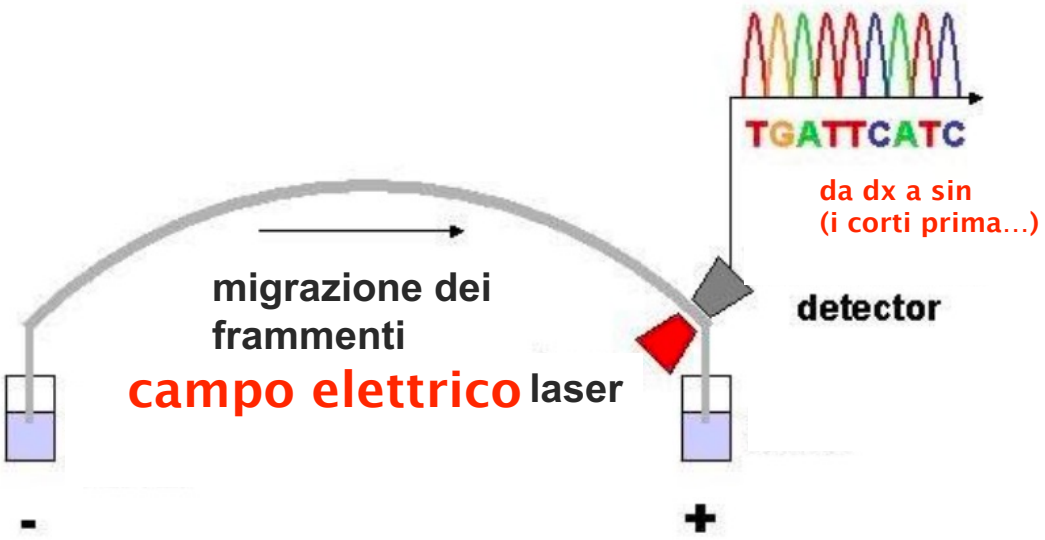
T G C A G G C

T G C A G G C A T

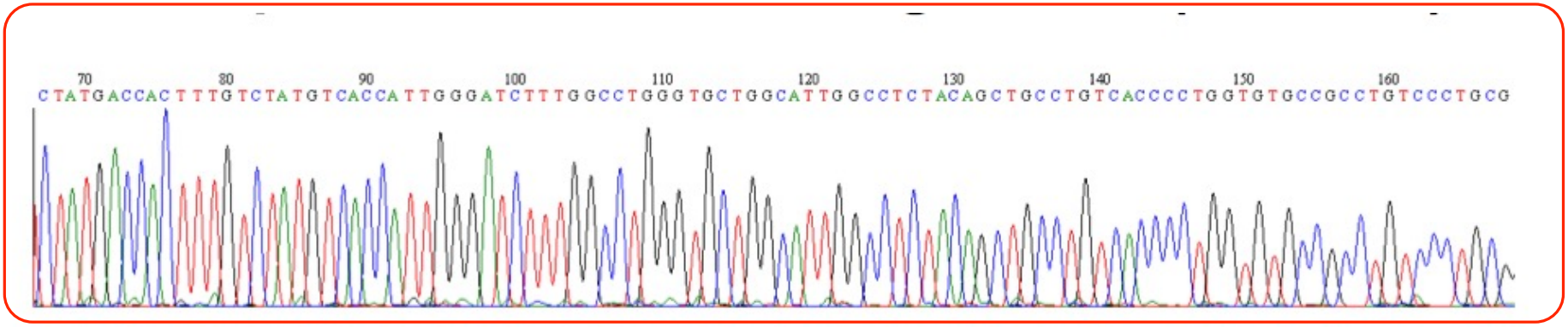
T G C A G G C A T C T

1 provetta

Elettroforesi capillare



Metodo Sanger (automatici)



cromatografia (prodotta dal sequenziatore automaticamente)

Base Caller

software che legge la cromatografia

Sequenza primaria del DNA template

Metodo Sanger

Il metodo Sanger produce un dato con:

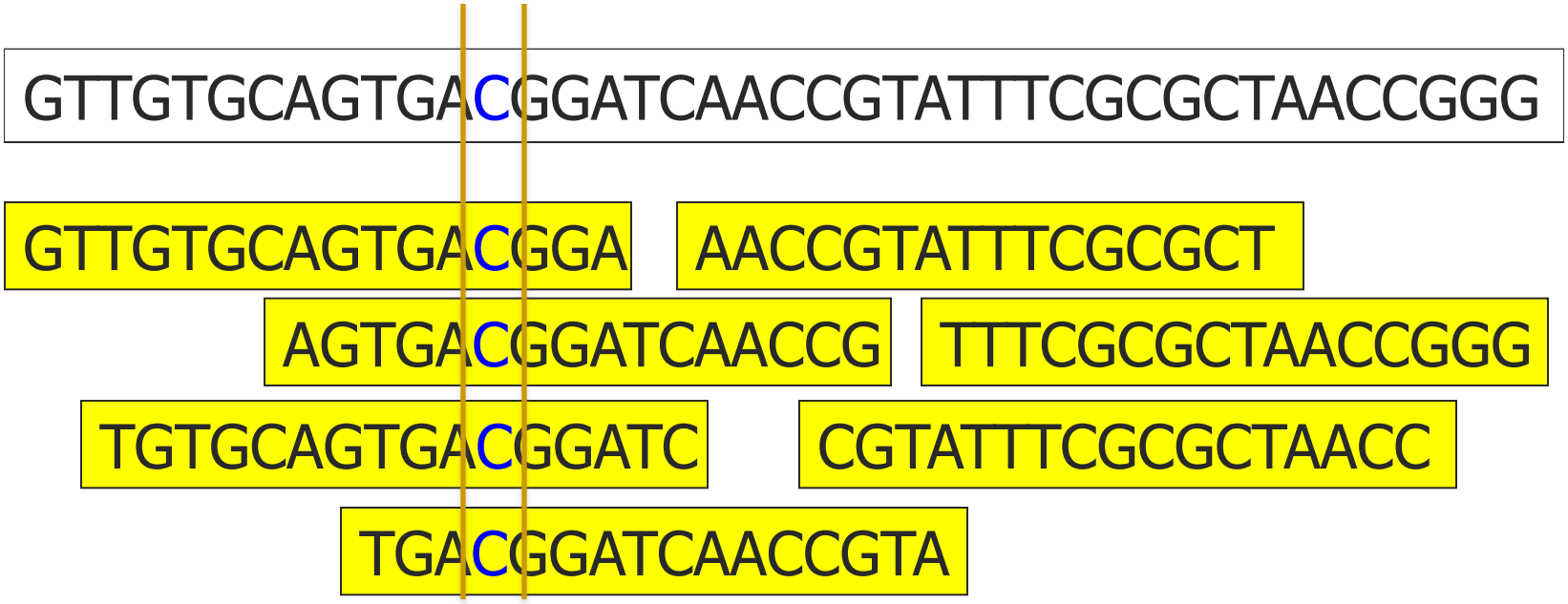
- ✓ qualità elevata
- ✓ *coverage* bassa (5x-8x)
- ✓ lunghezza fino a 1000bp

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

1$ per nucleotide!

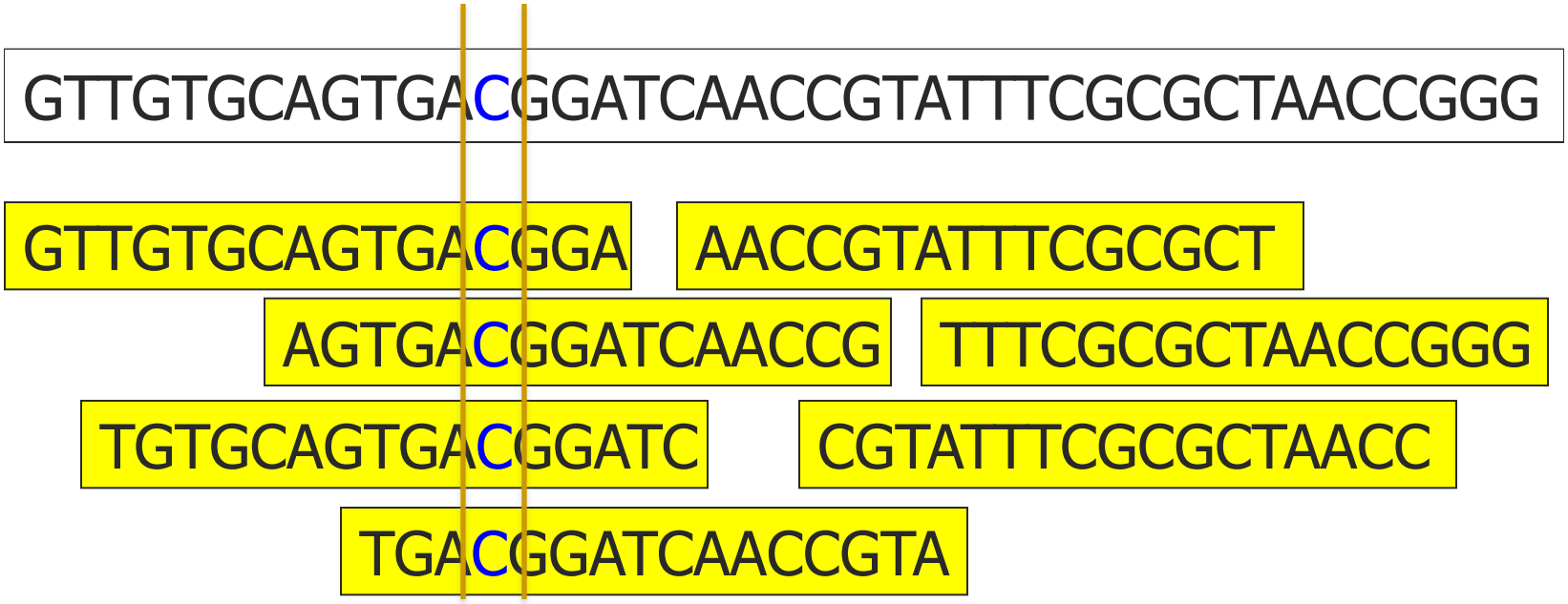
Coverage

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

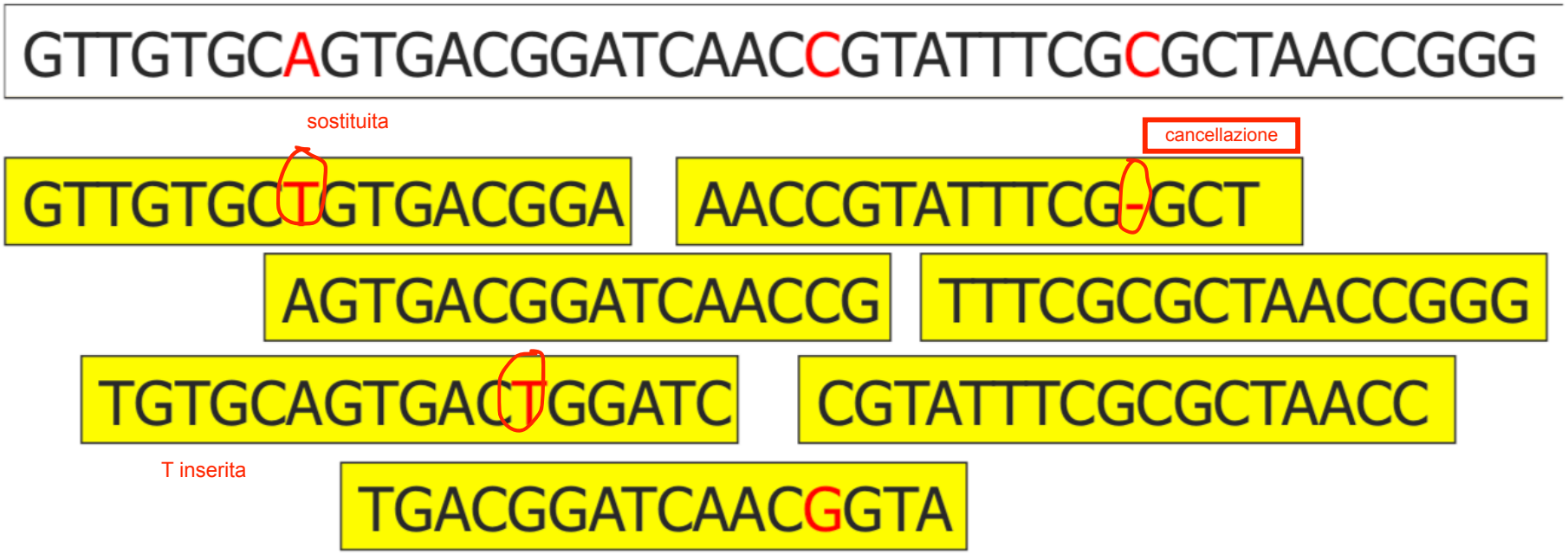
Coverage



coverage = n l / L

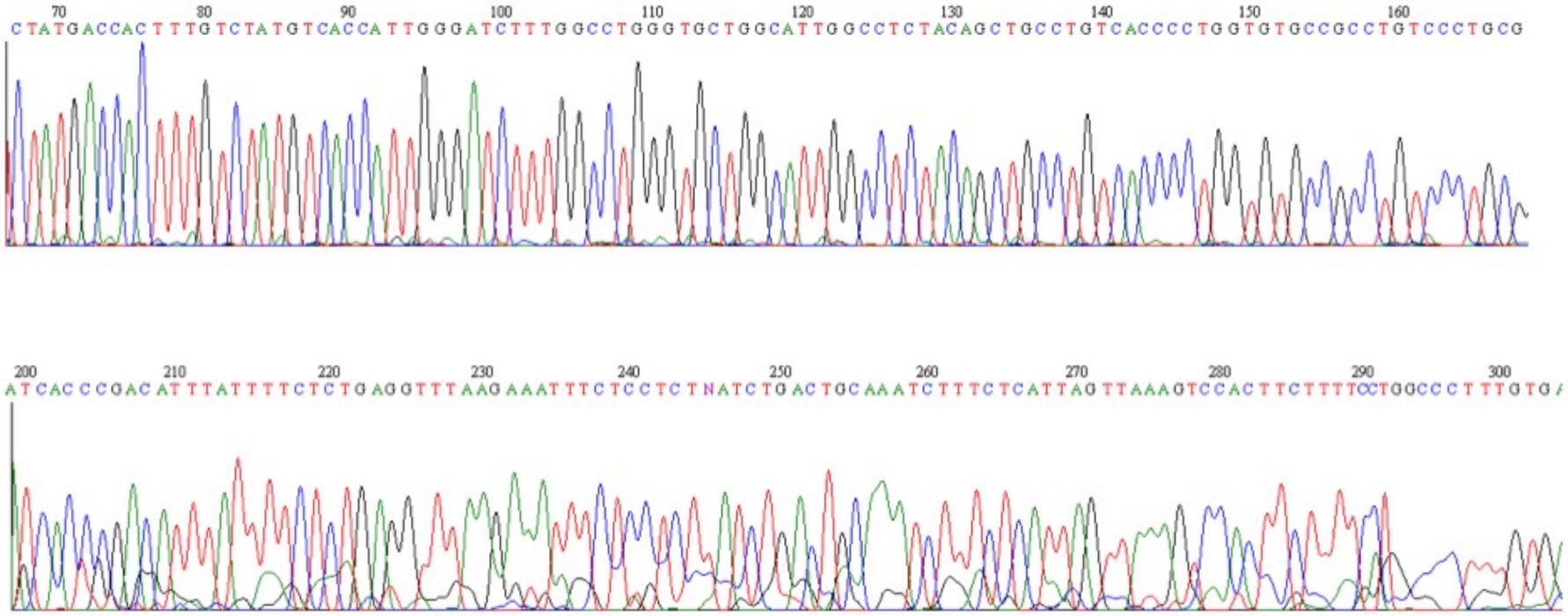
n → #reads
l → lunghezza dei reads (fissata)
L → lunghezza della molecola

Qualità



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

Qualità

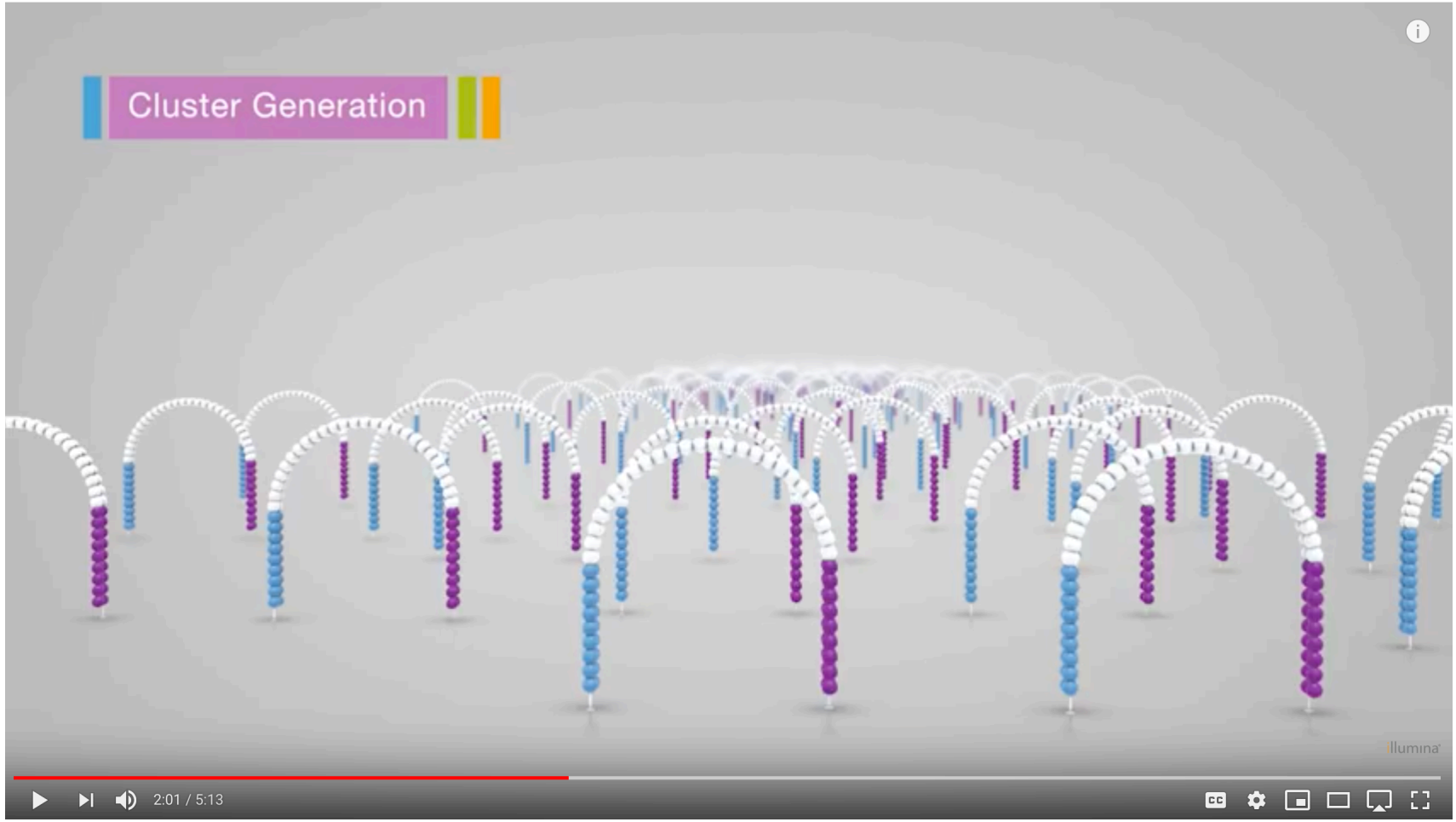

70 80 90 100 110 120 130 140 150 160
CTATGACCACTTTGTCTATGTCACCATTGGGATCTTTGGCCTGGGTGCTGGCATTGGCCTCTACAGCTGCCTGTCACCCCTGGTGTGCCGCCTGTCCCTGCG
200 210 220 230 240 250 260 270 280 290 300
ATCACCCGACATTTATTTTCTCTGAGGTTTAAGAAATTTCTCCTCTNATCTGACTGCAAATCTTTCTCATTAGTTAAAGTCCACTTCTTTTCCTGGCCCTTTGTG

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 Iix
- 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