

# **ASSEMBLAGGIO E ANNOTAZIONE DEL GENOMA DEL SARS-COV-2**

**Marianna Gambardella**

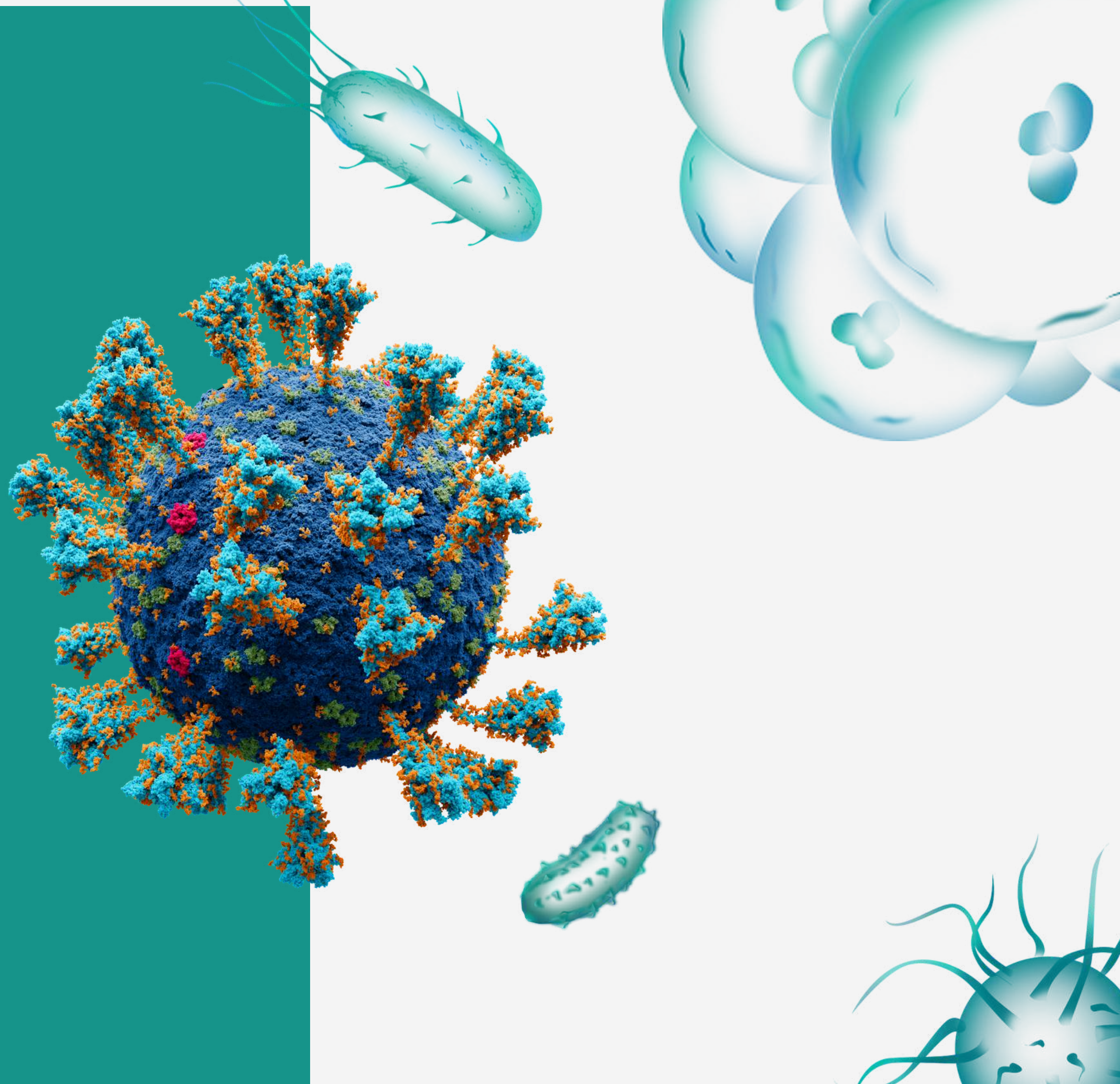


# **INTRODUZIONE**



# Sars-CoV-2

- Sindrome Respiratoria Acuta Grave
- Genoma a singolo filamento di RNA
- Malattia causata: Covid-19
- Periodo di incubazione che va da 2 a 14 giorni
- Sintomi che possono essere lievi (febbre, tosse) o gravi (polmonite, sindrome respiratoria acuta).

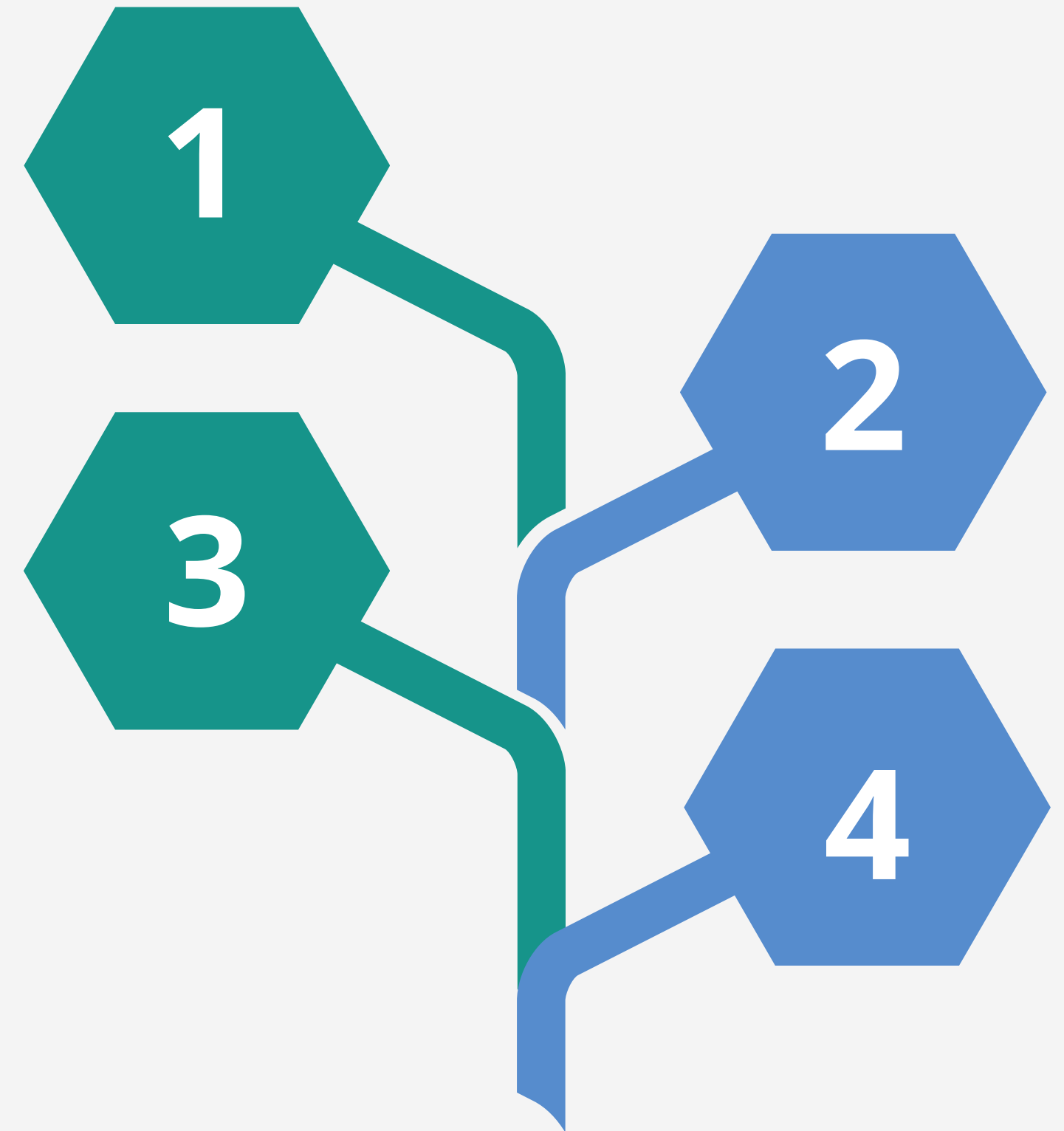
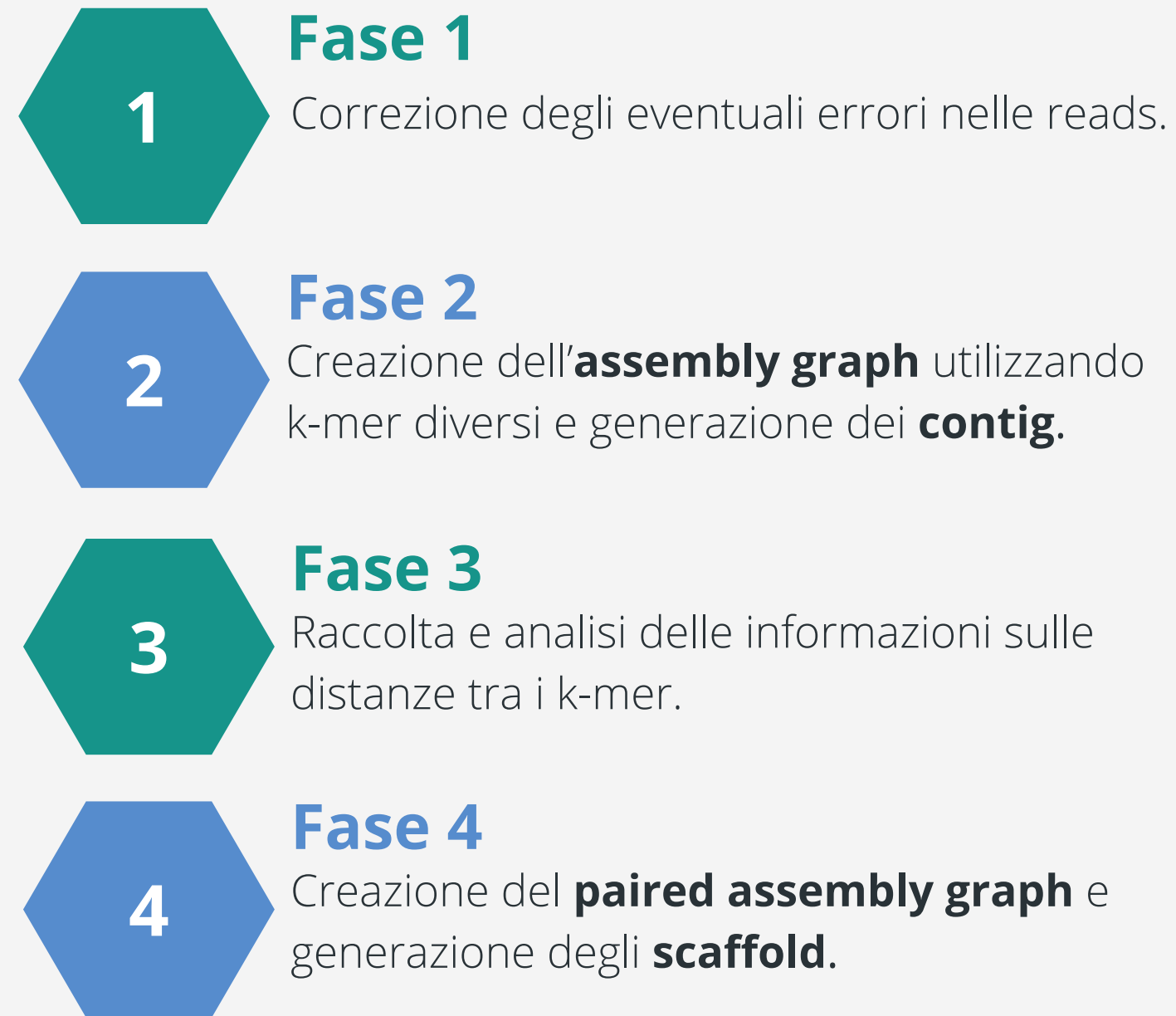






**TECNOLOGIE  
UTILIZZATE**

# SPAdes Assembler → Tool per l'assemblaggio di genomi



# Bandage

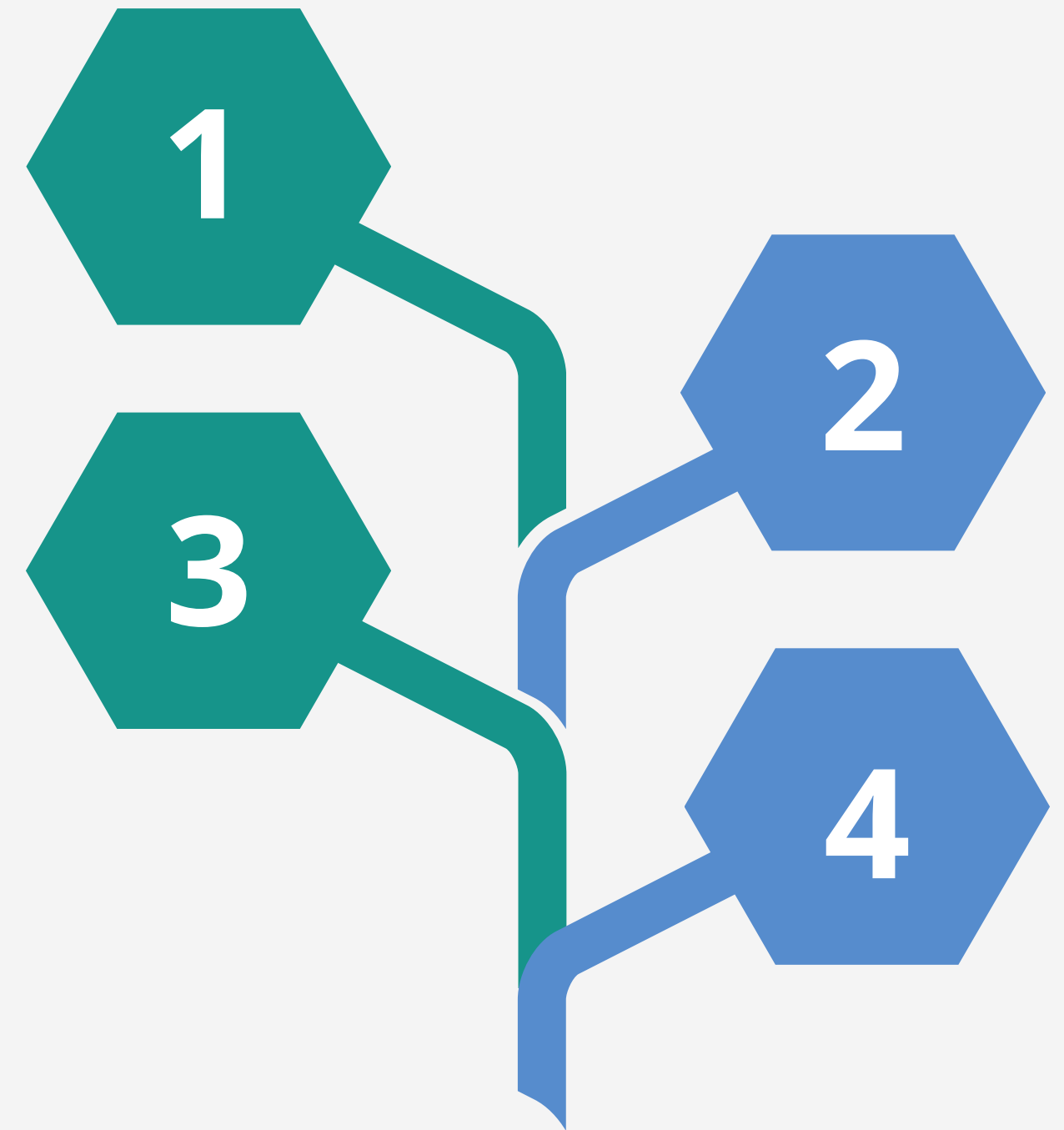
- GUI per la visualizzazione degli assembly graph.
- Posizione automatica ed efficiente dei nodi.
- Utilizzo di colori e forme diverse.
- Ciascun nodo rappresenta un contig o un k-mer e può essere etichettato in base a ID, lunghezza o coverage.
- Permette di valutare e confrontare assemblaggi ottenuti da tool differenti.
- Permette di identificare eventuali regioni critiche.



**Needle** → Tool per l'allineamento globale di coppie di sequenze

# Needleman-Wunsch Algorithm

- Fase 1**  
Creazione di una matrice  $(m+1) \times (n+1)$ , dove  $m$  e  $n$  sono le rispettive lunghezze delle due sequenze da allineare.
- Fase 2**  
Definizione dei punteggi corrispondenti ai **match** e ai **mismatch** e delle penalità da applicare ai **gap**.
- Fase 3**  
Ciascuna cella  $(i,j)$  viene riempita considerando il massimo tra:  
1. la somma tra il punteggio della cella  $(i-1, j-1)$  e quello della coppia di caratteri nelle posizioni  $i$  e  $j$ ;  
2. la somma tra il punteggio della cella  $(i-1, j)$  o  $(i, j-1)$  e il costo del gap inserito;
- Fase 4**  
La casella  $D(m,n)$  contiene il punteggio dell'allineamento ottimale, ricostruibile seguendo i punteggi massimi nella matrice.





# Annotazione di genomi

- Identificazione dei **geni putativi**.
- Mapping dei geni predetti con geni noti, utilizzando database esistenti.
- L'identificazione coinvolge la valutazione delle **ORF (Open Reading Frames)**.
- Le ORF sono sequenze sufficientemente lunghe che iniziano per **ATG (start codon)**, terminano per **TAA, TAG, TGA (stop codon)** e non contengono stop codon intermedi.
- Non tutte le ORF codificano necessariamente proteine.
- Sono identificati anche altri elementi funzionali, come regioni non codificanti e sequenze ripetute.
- Essenziale per comprendere la struttura e la funzione dei geni di un organismo





# Prokka

- Tool per l'annotazione di genomi procariotici.
- Prende in input l'assemblaggio del genoma.
- Coordina l'esecuzione di altri tool in base al tipo di genoma da annotare.
- Utilizza l'algoritmo **Prodigal** per la predizione dei geni.
- Utilizza l'algoritmo **BLAST** per il confronto tra i geni predetti e database di proteine note.



# Prodigal Algorithm

1

## Fase 1

Processo di training basato sulla presenza di **bias** in relazione alle basi **G** e **C**, in ogni posizione dei codoni per effettuare una prima scrematura tra le ORF.

2

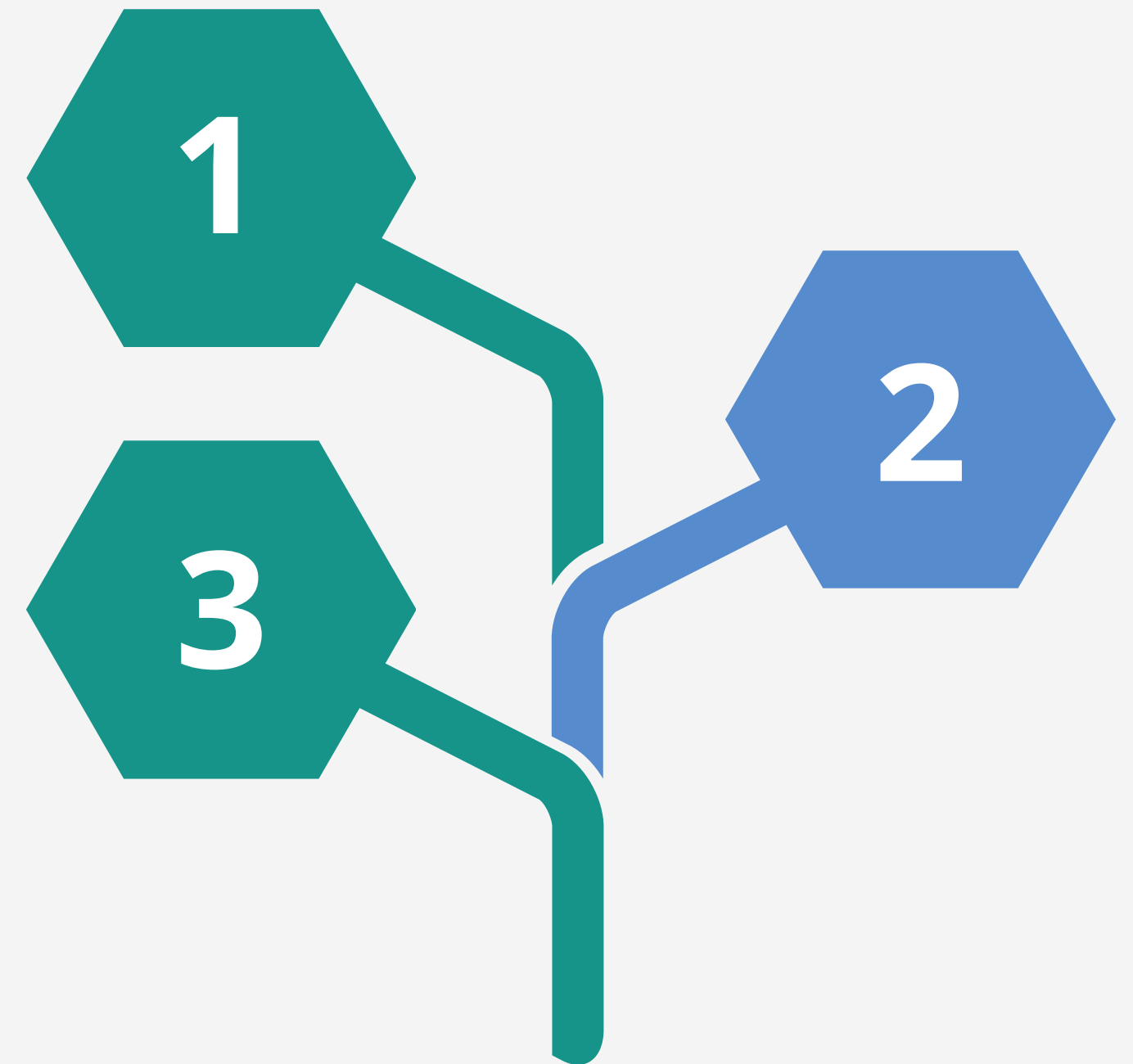
## Fase 2

Uso della tecnica di programmazione dinamica per costruire modelli di ORF potenzialmente codificanti, attraverso la valutazione dei **dicodoni** e assegnando punteggi e penalità fino a quando l'insieme delle ORF non resta costante.

3

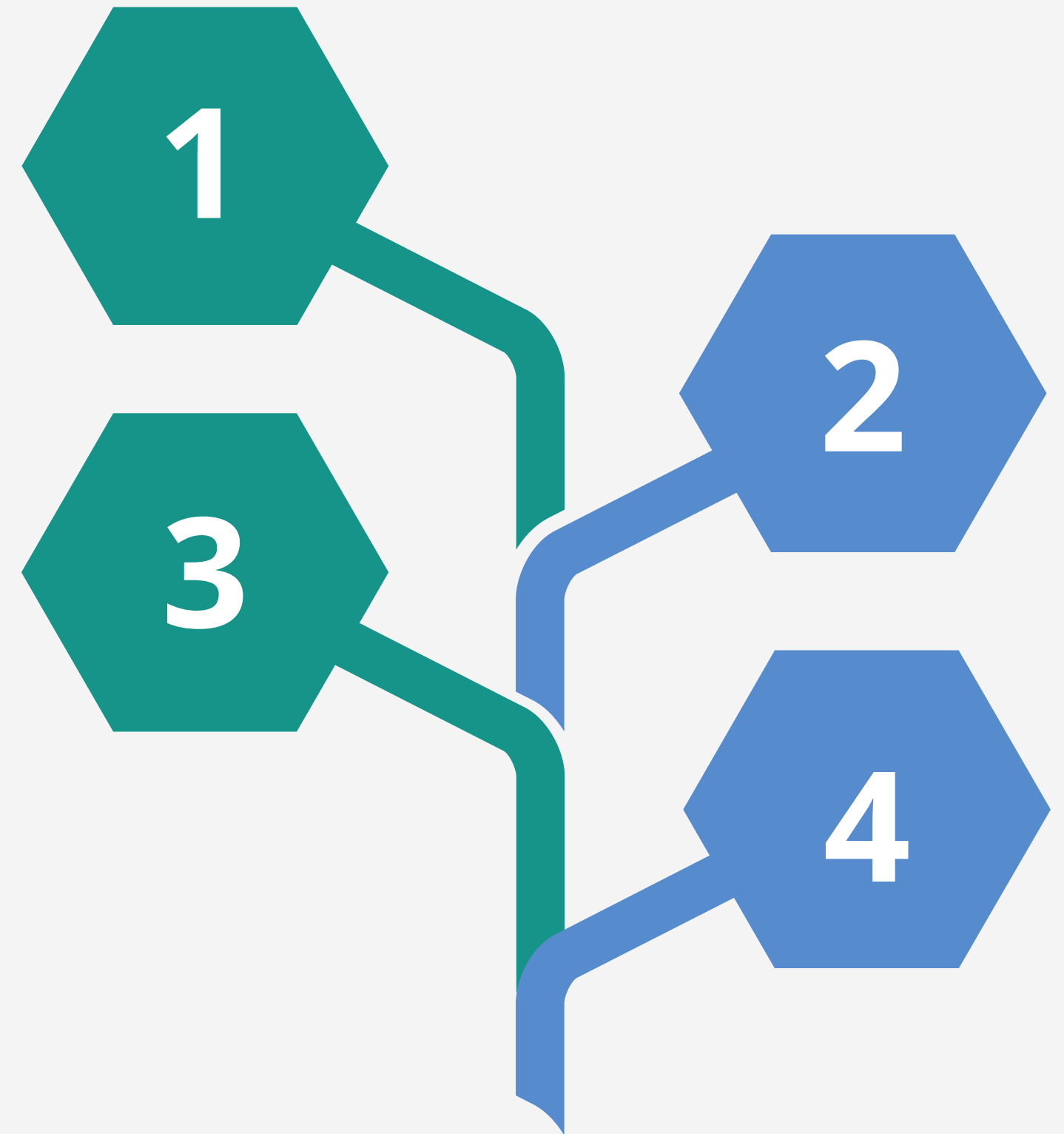
## Fase 3

Sono restituite le regioni codificanti.



# BLAST Algorithm

- Fase 1**  
Scomposizione della query in parole di lunghezza  $W$  (**W-mer**).
- Fase 2**  
Valutazione della somiglianza tra ogni W-mer e una lista di parole simili, assegnando punteggi e definendo una soglia al di sotto della quale le parole poco somiglianti vengono rimosse.
- Fase 3**  
Ogni parola nella lista viene cercata nei database di riferimento, e quando si verifica un match la ricerca viene estesa in entrambe le direzioni per individuare gli **High Scoring Pairs**.
- Fase 4**  
Sono restituiti i match con i punteggi totali più alti.







**ASSEMBLAGGIO**

# Dati di sequenziamento

Genome sequencing of SARS-CoV-2 isolates after 1 passage in Vero E6 cells (SB2) (SRR11528307)

Metadata Analysis Reads Data access FASTA/FASTQ download

Run

| Run         | Spots  | Bases  | Size   | GC Content | Published  | Access Type |
|-------------|--------|--------|--------|------------|------------|-------------|
| SRR11528307 | 719.8k | 186.5M | 76.9MB | 38.2%      | 2020-04-15 | public      |

Quality graph [\(bigger\)](#)

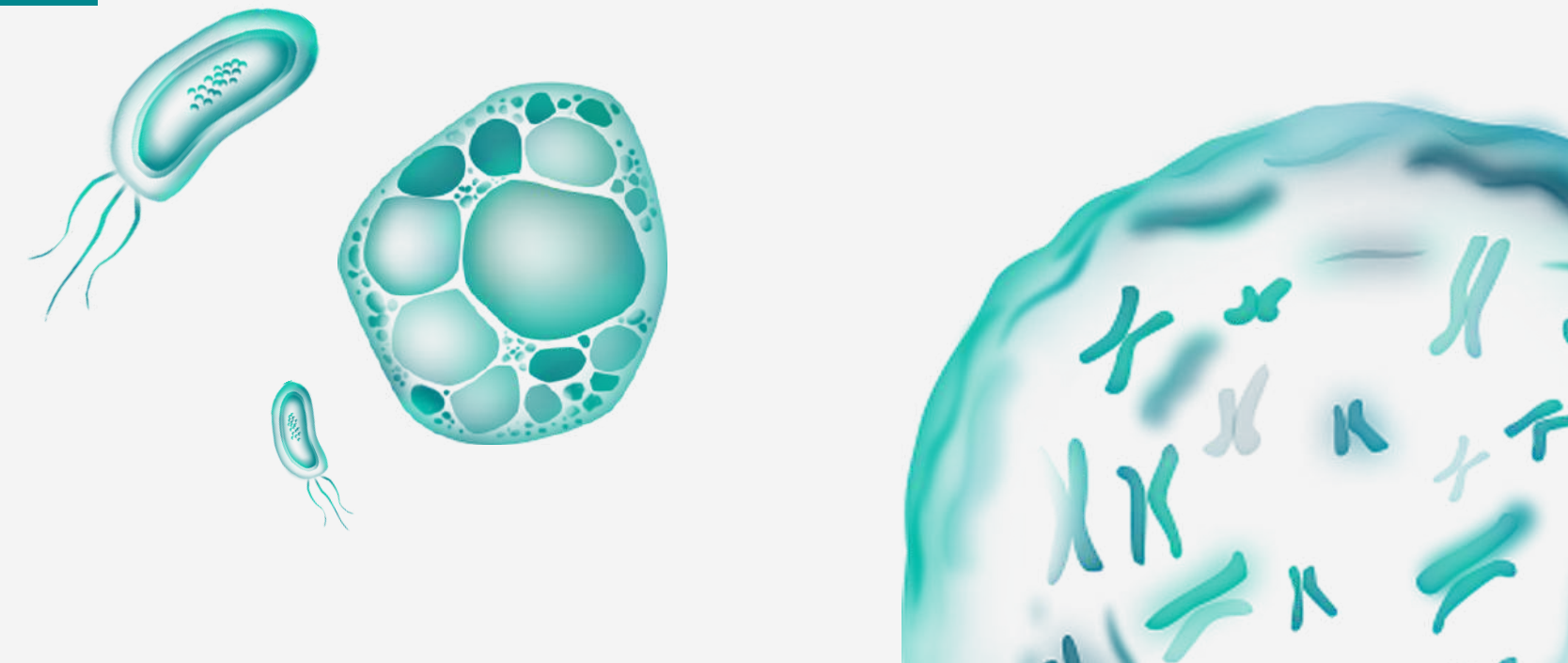
This run has 2 reads per spot:

$\bar{L}=131$ ,  $\sigma=38.0$ , 100%

$\bar{L}=128$ ,  $\sigma=41.3$ , 100%

[Legend](#)

Quante basi nucleotidiche si trovano all'interno di questo file? Qual è la percentuale di reads che corrispondono alle basi G o C?





# Dati di sequenziamento

Genome sequencing of SARS-CoV-2 isolates after 1 passage in Vero E6 cells (SB2) (SRR11528307)

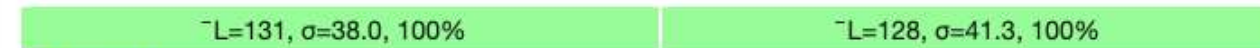
Metadata Analysis Reads Data access FASTA/FASTQ download

Run

| Run         | Spots  | Bases  | Size   | GC Content | Published  | Access Type |
|-------------|--------|--------|--------|------------|------------|-------------|
| SRR11528307 | 719.8k | 186.5M | 76.9MB | 38.2%      | 2020-04-15 | public      |

Quality graph [\(bigger\)](#)

This run has 2 reads per spot:



[Legend](#)

Se il genoma del Sars-CoV-2 ha una lunghezza pari a 30,000 nucleotidi, qual è la coverage delle reads di questo dataset?

$$\text{Coverage} = \frac{nl}{L} = \frac{719,800(2 \cdot 129.5)}{30,000} = 6214.3$$



# Dati importati in Galaxy:

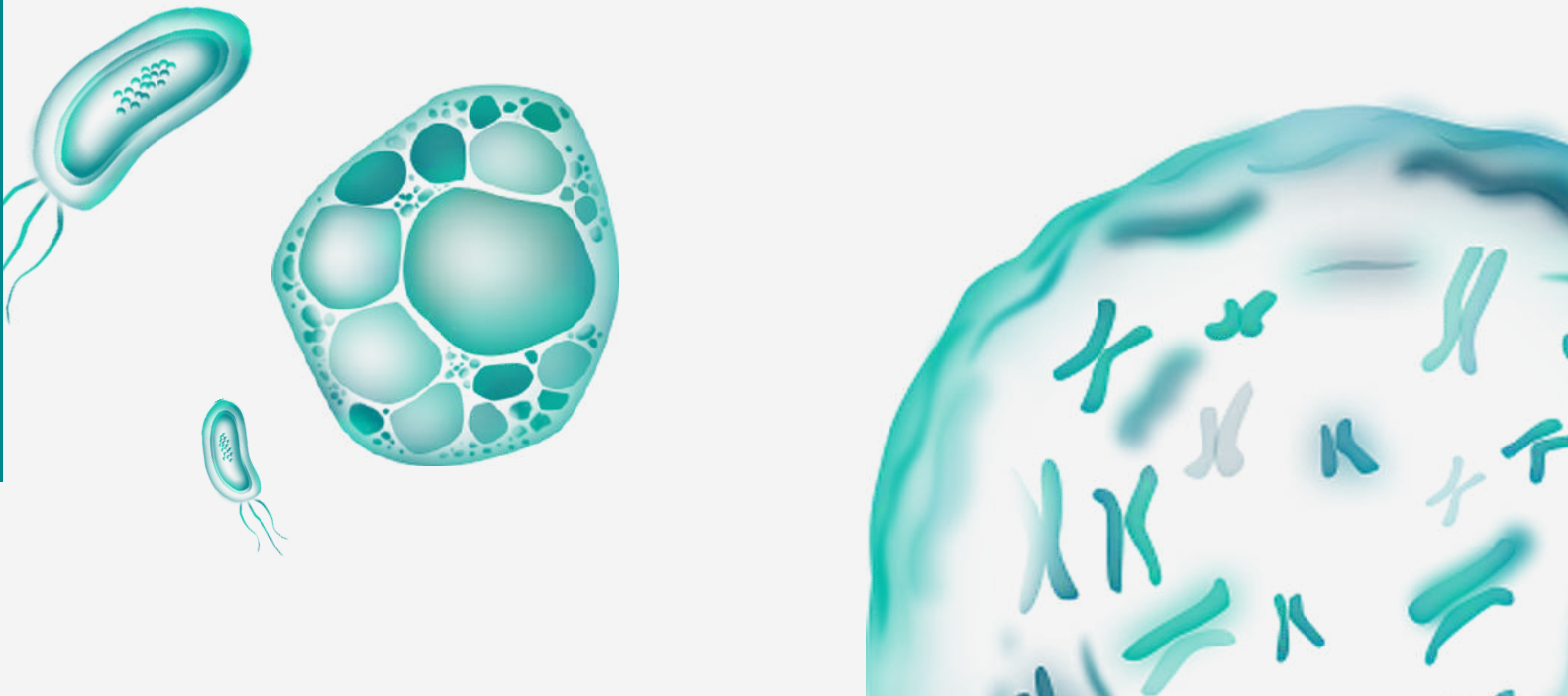
- Formato FastQ
- Phred Quality Scores

```
@MN01288:4:000H32WJK:1:11101:10036:8595/1
TTTATATACTGCTCATACTTTCCAAGTTCTTGGAGATCGATGAGAGATTCAATTTAAATTCTTGGCAACCTCATTGAGGCGGTCAATTTCTTTTGAATG
+
F/FFFFFFFFFFFFFFFF/FFFF=FFFFAFAAFFFFFFFFAAAF/FFFFFFFFFFFFFFFFFAFAF//FFFFFFFFF/F/AFFFF//A
```

Quali sono i punteggi di qualità della read "8595/1"? Sono buoni? Ci sono dei nucleotidi che ti preoccupano?

La maggior parte dei punteggi è pari a 37, pertanto si può affermare che sono **buoni** (37 > 30). Tuttavia ci sono nucleotidi che risultano preoccupanti: uno con punteggio pari a 28 e nove con punteggio pari a 14.

| ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger |         |       |    |         |       |    |         |       |    |         |       |
|--|---------|-------|----|---------|-------|----|---------|-------|----|---------|-------|
| Q  | P_error | ASCII | Q  | P_error | ASCII | Q  | P_error | ASCII | Q  | P_error | ASCII |
| 0  | 1.00000 | 33 !  | 11 | 0.07943 | 44 ,  | 22 | 0.00631 | 55 7  | 33 | 0.00050 | 66 B  |
| 1  | 0.79433 | 34 "  | 12 | 0.06310 | 45 -  | 23 | 0.00501 | 56 8  | 34 | 0.00040 | 67 C  |
| 2  | 0.63096 | 35 #  | 13 | 0.05012 | 46 .  | 24 | 0.00398 | 57 9  | 35 | 0.00032 | 68 D  |
| 3  | 0.50119 | 36 \$ | 14 | 0.03981 | 47 /  | 25 | 0.00316 | 58 :  | 36 | 0.00025 | 69 E  |
| 4  | 0.39811 | 37 %  | 15 | 0.03162 | 48 0  | 26 | 0.00251 | 59 ;  | 37 | 0.00020 | 70 F  |
| 5  | 0.31623 | 38 &  | 16 | 0.02512 | 49 1  | 27 | 0.00200 | 60 <  | 38 | 0.00016 | 71 G  |
| 6  | 0.25119 | 39 '  | 17 | 0.01995 | 50 2  | 28 | 0.00158 | 61 =  | 39 | 0.00013 | 72 H  |
| 7  | 0.19953 | 40 (  | 18 | 0.01585 | 51 3  | 29 | 0.00126 | 62 >  | 40 | 0.00010 | 73 I  |
| 8  | 0.15849 | 41 )  | 19 | 0.01259 | 52 4  | 30 | 0.00100 | 63 ?  | 41 | 0.00008 | 74 J  |
| 9  | 0.12589 | 42 *  | 20 | 0.01000 | 53 5  | 31 | 0.00079 | 64 @  | 42 | 0.00006 | 75 K  |
| 10   | 0.10000 | 43 +  | 21 | 0.00794 | 54 6  | 32 | 0.00063 | 65 A  |    |         |       |



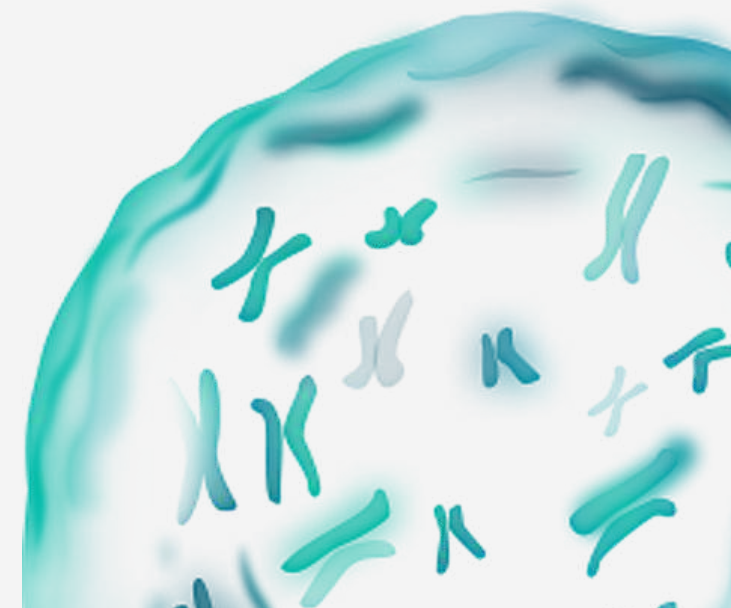
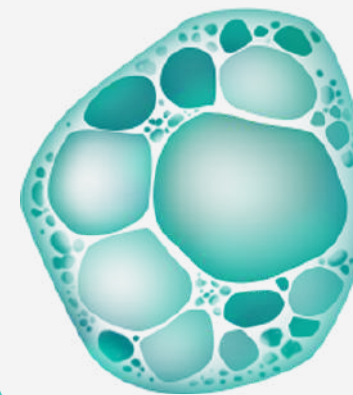
# Istogramma dei Phred Quality Scores

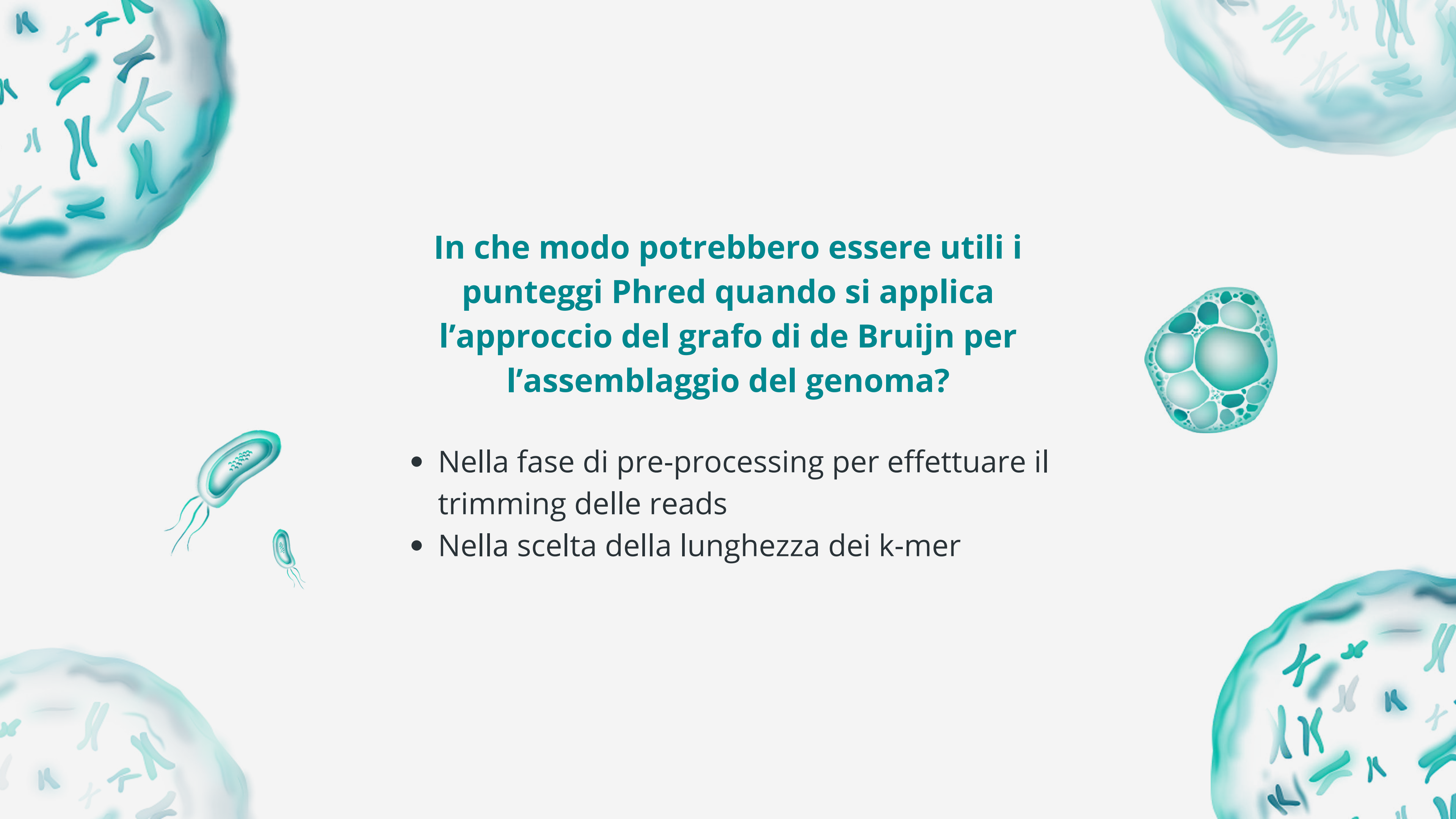
Quality graph (smaller)



Interpreta l'istogramma. I punteggi di qualità sono buoni? Cosa te lo fa pensare?

In generale, la più alta concentrazione dei punteggi rientra nel valore pari a 37, quindi si può affermare che la qualità delle reads è **buona**.



The background features several stylized, light blue, translucent cell-like shapes. Inside these shapes are various representations of chromosomes, some as simple lines and others as more complex, X-shaped structures. The overall aesthetic is clean and scientific.

## **In che modo potrebbero essere utili i punteggi Phred quando si applica l'approccio del grafo di de Bruijn per l'assemblaggio del genoma?**

- Nella fase di pre-processing per effettuare il trimming delle reads
- Nella scelta della lunghezza dei k-mer





**DEMO ASSEMBLAGGIO**

# Contigs ottenuti:

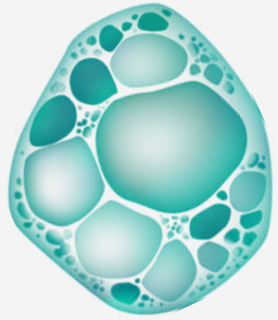
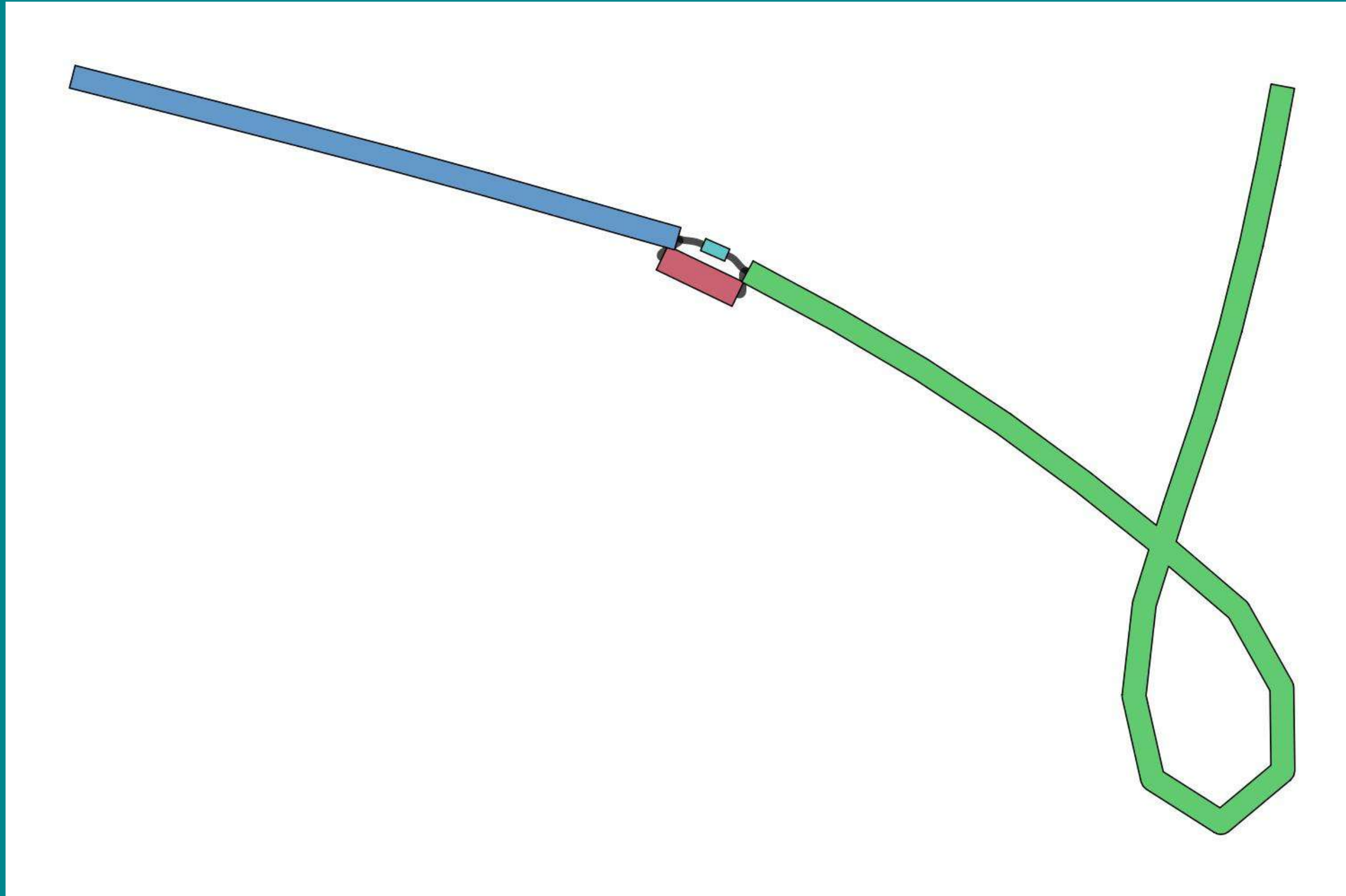
Quanti contigs ha prodotto l'assemblaggio? Che lunghezza hanno?

| name   | length | coverage    |
|--------|--------|-------------|
| #name  | length | coverage    |
| NODE_1 | 29600  | 2807.600820 |
| NODE_2 | 147    | 27.500000   |

Cosa pensi si intenda per *coverage* in questo contesto?

Si intende la coverage relativa ai k-mer usati da SPAdes per l'assemblaggio, che indica il numero di occorrenze di ogni k-mer all'interno delle reads.

# Assembly Graph







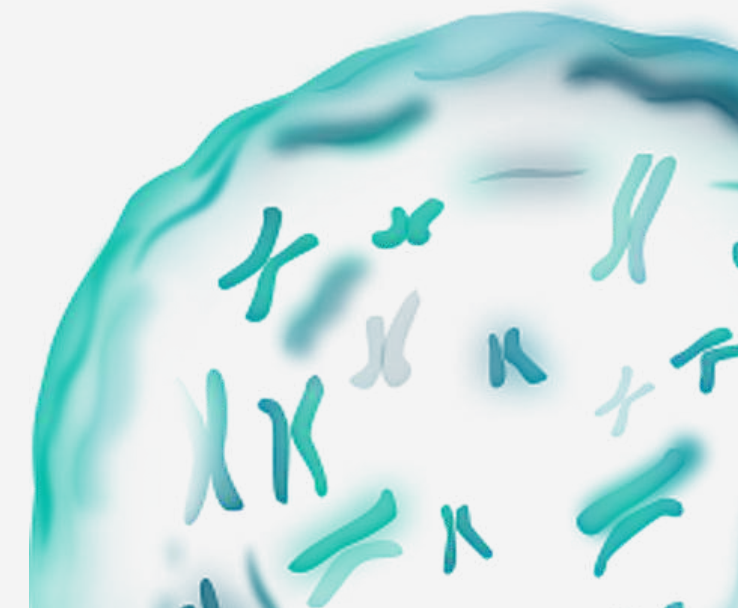
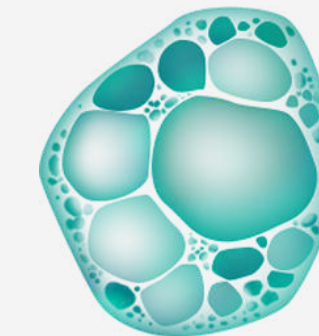
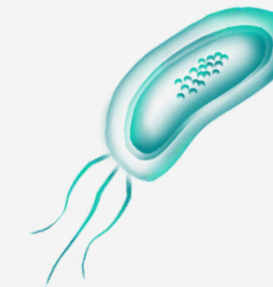
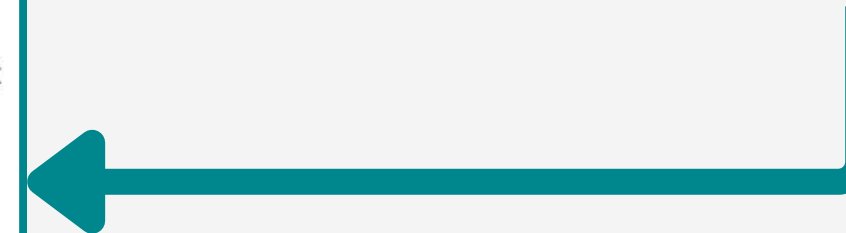
**DEMO ALLINEAMENTO**

# Allineamento

```
#####
# Program: needle
# Rundate: Fri 23 Feb 2024 16:45:50
# Commandline: needle
#   -asequence /mnt/user-data-volA/data11/f/6/0/dataset_f6038d4e-2b2f-42cd-bf4f-925f42a7e093.dat
#   -bsequence /mnt/user-data-volA/data11/4/5/0/dataset_45077f5a-5107-456a-abf8-040178ed3ff5.dat
#   -outfile /mnt/tmp/job_working_directory/008/265/8265809/outputs/dataset_28641a84-0826-4528-9720-8c7d7be04faa.dat
#   -gapopen 10.0
#   -gapextend 0.5
#   -brief yes
#   -aformat3 srspair
#   -auto
# Align_format: srspair
# Report_file: /mnt/tmp/job_working_directory/008/265/8265809/outputs/dataset_28641a84-0826-4528-9720-8c7d7be04faa.dat
#####

#=====
#
# Aligned_sequences: 2
# 1: NODE_1_length_29600_cov_2807.600820
# 2: NC_004718.3
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 30448
# Identity: 23875/30448 (78.4%)
# Similarity: 23875/30448 (78.4%)
# Gaps: 1545/30448 ( 5.1%)
# Score: 94510.0
#
#=====
```

Quanti simboli risultano allineati  
dall'allineamento? Quanti gap ci sono?

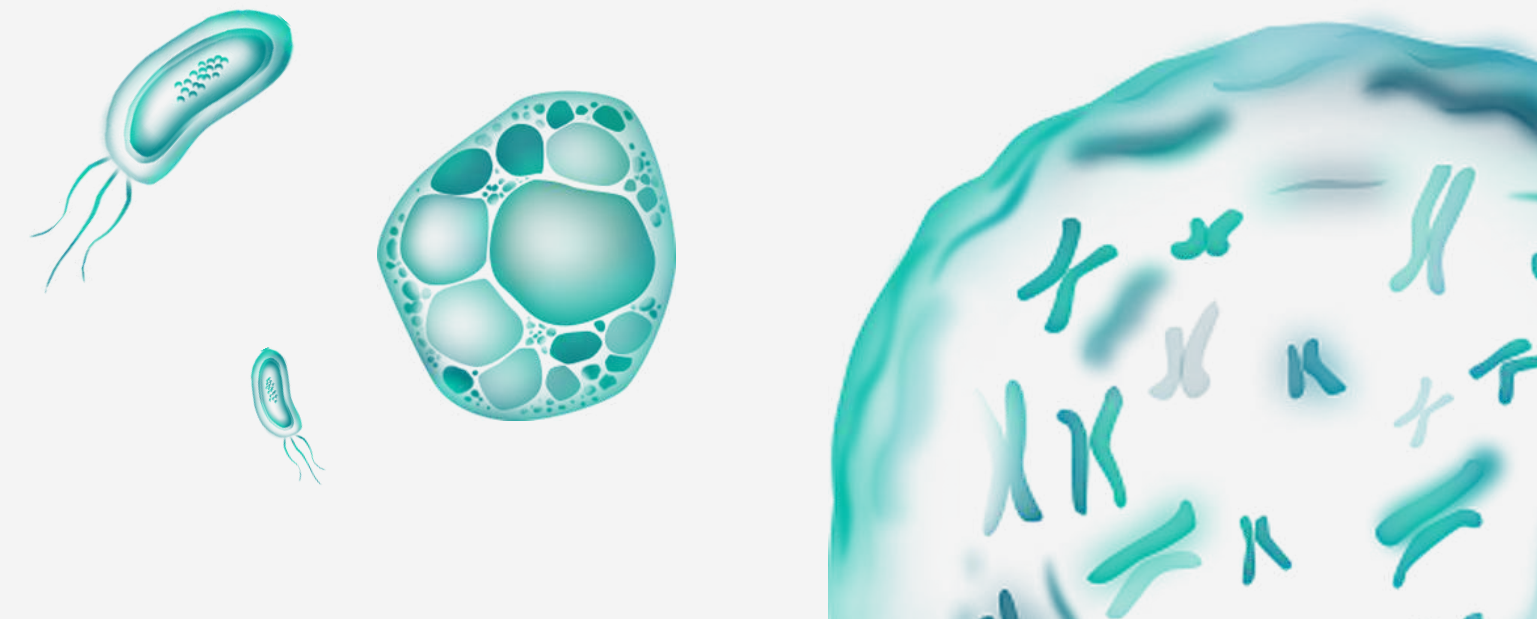




|               |       |   |       |
|---------------|-------|---|-------|
| NODE_1_length | 21743 | TAGATTTCGAAGACCCAGTCCCTACTTATTGTTAATAACGCTACTAATGTT | 21792 |
| NC_004718.3   | 21811 | TGAACAACAAGTCACAGTCGGTGATTATTATTAACAATTCTACTAATGTT  | 21860 |
| NODE_1_length | 21793 | GTTATTAAAGTCTGTGAATTTCAATTTTGTAAATGATCCATTTTTGGGTGT | 21842 |
| NC_004718.3   | 21861 | GTTATACGAGCATGTAACCTTTGAATTGTGTGACAACCCTTTCTTTGCTGT | 21910 |
| NODE_1_length | 21843 | TTAT-----TACCACAAAAACA-ACAAAAGTTGGATGGAAAGTG        | 21880 |
| NC_004718.3   | 21911 | TTCTAAACCCATGGGTA-CACAGACACATACTA-----TG            | 21944 |
| NODE_1_length | 21881 | AGTTCAGAGTTTATTC--TAGTGCGAATAATTGCACTTTTGAATATGTCT  | 21928 |
| NC_004718.3   | 21945 | A-----TATTCGATAATGCATTTAATTGCACTTTGAGTACATAT        | 21984 |
| NODE_1_length | 21929 | CTCA-GCCTTTTCTTATGGACCTTGAAG-----GAAAA--CAGGGTAATT  | 21970 |
| NC_004718.3   | 21985 | CTGATGCCTTTTC-----GCTTGATGTTTCAGAAAAGTCA-GGTAATT    | 22026 |
| NODE_1_length | 21971 | TCAAAAATCTTA-GGGAATTTGTGTTTAAGAATATTGATGGTTAT-----  | 22014 |
| NC_004718.3   | 22027 | TTAAACA-CTTACGAGAGTTTGTGTTTAAAAATAAAGATGGGTTTCTCTA  | 22075 |
| NODE_1_length | 22015 | --TTTAAAA-----TATATTCTAAGCACACGCCTATTAATTTAGTGCGTGA | 22058 |
| NC_004718.3   | 22076 | TGTTTATAAGGGCTAT-----CA-ACCTATAGATGTAGTTCGTGA       | 22114 |
| NODE_1_length | 22059 | TCTCCCTCAGGGTTTTTCGGCTTTAGAACC-ATTGGTAGATTTGCCAATA  | 22107 |
| NC_004718.3   | 22115 | TCTACCTTCTGGTTTTAACACTTTGAAACCTATTTTAA-AGTTGCCTCTT  | 22163 |
| NODE_1_length | 22108 | GGTATTAACATCACTAGGTTTCAAACCTTTA----CTTGCTTTACATAGAA | 22153 |
| NC_004718.3   | 22164 | GGTATTAACA-----TTACAAATTTTAGAGCCATTCTTAC-----A      | 22199 |
| NODE_1_length | 22154 | GTTATTTGACTCCTGGTGATTCTTCTTCAGG----TTGGACAGCTGGT--  | 22197 |
| NC_004718.3   | 22200 | GCCTTTT--CACCTG-----CTCAAGACATTTGG--GGCACGTCA       | 22235 |

**Scorrendo l'allineamento, noti qualche regione che appare più variabile di altre?**

Sì, sono presenti alcune regioni più variabili, ovvero caratterizzate da un numero maggiore di **mismatch** e **indel**.







**DEMO ANNOTAZIONE**

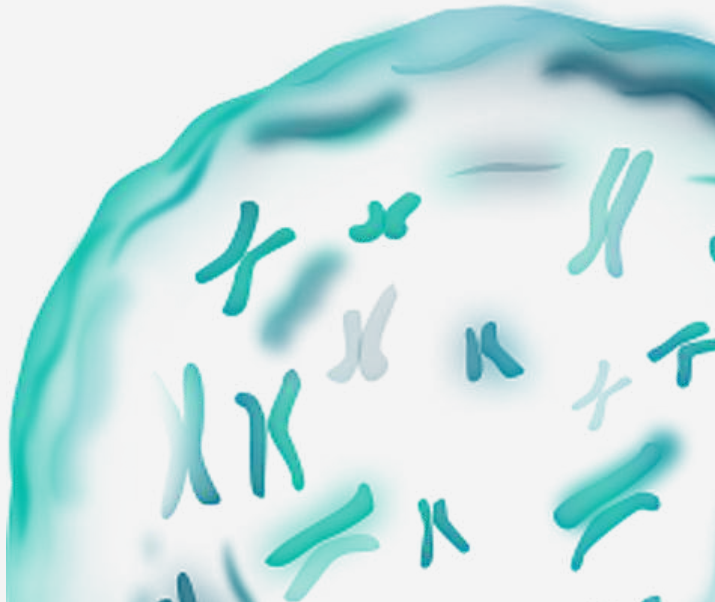
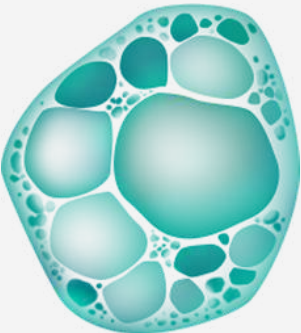
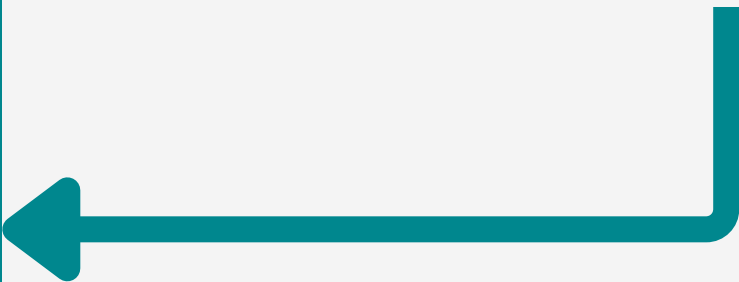
# Putative Genes:

Quanti geni sono stati identificati come putativi?

Sono stati identificati 9 geni putativi.

Qual è il più lungo e quale il più corto?

| Column 1       | Column 2 | Column 3  | Column 4 | Column 5  | Column 6 | Column 7                  |
|----------------|----------|-----------|----------|-----------|----------|---------------------------|
| locus_tag      | ftype    | length_bp | gene     | EC_number | COG      | product                   |
| AMFCHINL_00001 | CDS      | 13218     | 1a       |           |          | Replicase polyprotein 1a  |
| AMFCHINL_00002 | CDS      | 7788      | rep      |           |          | Replicase polyprotein 1ab |
| AMFCHINL_00003 | CDS      | 3822      | S        |           |          | Spike glycoprotein        |
| AMFCHINL_00004 | CDS      | 828       | 3a       |           |          | Protein 3a                |
| AMFCHINL_00005 | CDS      | 669       | M        |           |          | Membrane protein          |
| AMFCHINL_00006 | CDS      | 186       |          |           |          | hypothetical protein      |
| AMFCHINL_00007 | CDS      | 366       | 7a       |           |          | Protein 7a                |
| AMFCHINL_00008 | CDS      | 366       |          |           |          | hypothetical protein      |
| AMFCHINL_00009 | CDS      | 1260      | N        |           |          | Nucleoprotein             |



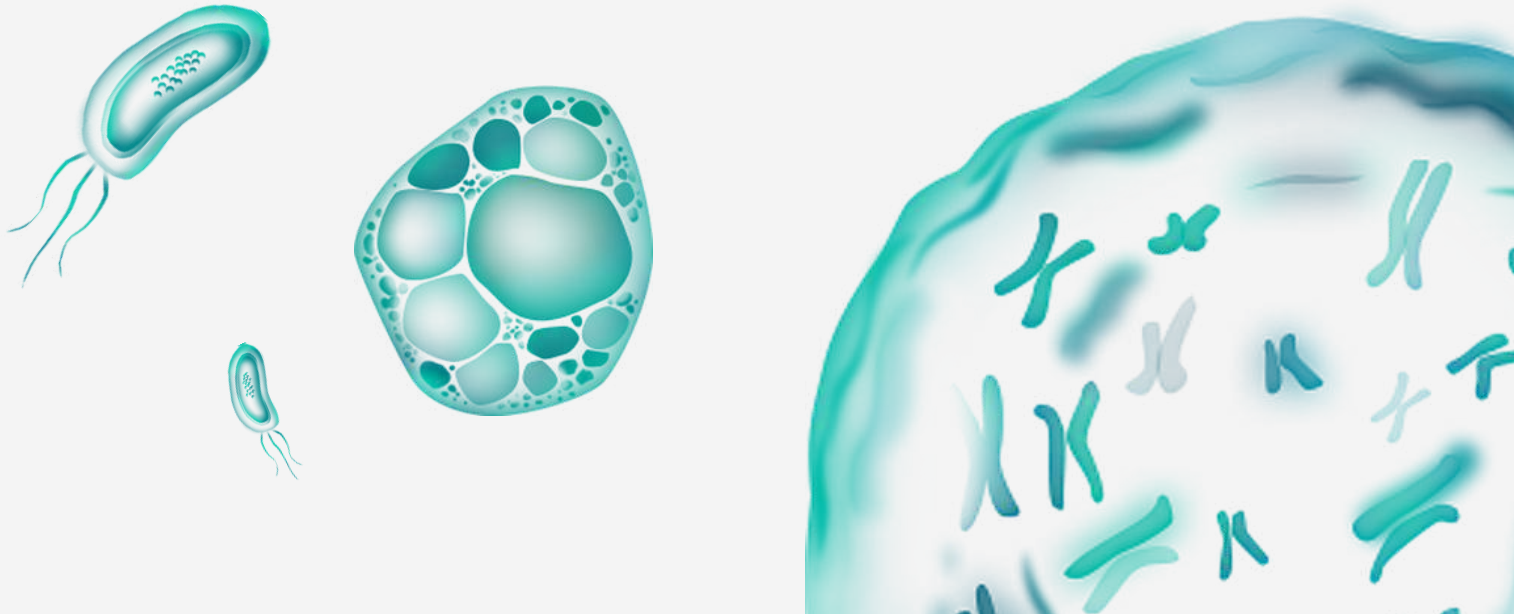


# Putative Genes:

| Column 1       | Column 2 | Column 3  | Column 4 | Column 5  | Column 6 | Column 7                  |
|----------------|----------|-----------|----------|-----------|----------|---------------------------|
| locus_tag      | ftype    | length_bp | gene     | EC_number | COG      | product                   |
| AMFCHINL_00001 | CDS      | 13218     | 1a       |           |          | Replicase polyprotein 1a  |
| AMFCHINL_00002 | CDS      | 7788      | rep      |           |          | Replicase polyprotein 1ab |
| AMFCHINL_00003 | CDS      | 3822      | S        |           |          | Spike glycoprotein        |
| AMFCHINL_00004 | CDS      | 828       | 3a       |           |          | Protein 3a                |
| AMFCHINL_00005 | CDS      | 669       | M        |           |          | Membrane protein          |
| AMFCHINL_00006 | CDS      | 186       |          |           |          | hypothetical protein      |
| AMFCHINL_00007 | CDS      | 366       | 7a       |           |          | Protein 7a                |
| AMFCHINL_00008 | CDS      | 366       |          |           |          | hypothetical protein      |
| AMFCHINL_00009 | CDS      | 1260      | N        |           |          | Nucleoprotein             |

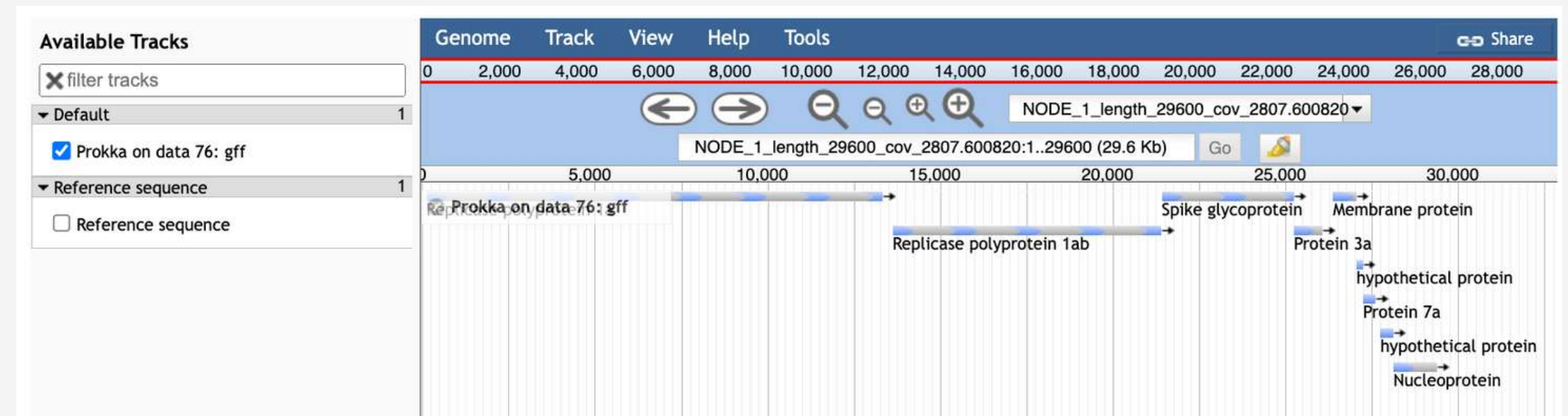
Perchè pensi che due geni siano stati etichettati come “ipotetiche proteine”?

Tale etichetta suggerisce che non sia stata riscontrata una significativa somiglianza con alcuna sequenza proteica nel database considerato dal tool Prokka.





# Conclusioni



Tutte le frecce puntano nella stessa direzione e ciò indica che i geni si trovano tutti sullo stesso filamento del genoma.

**GRAZIE PER  
L'ATTENZIONE!**

