



BIOCOSE

Progetto di gruppo

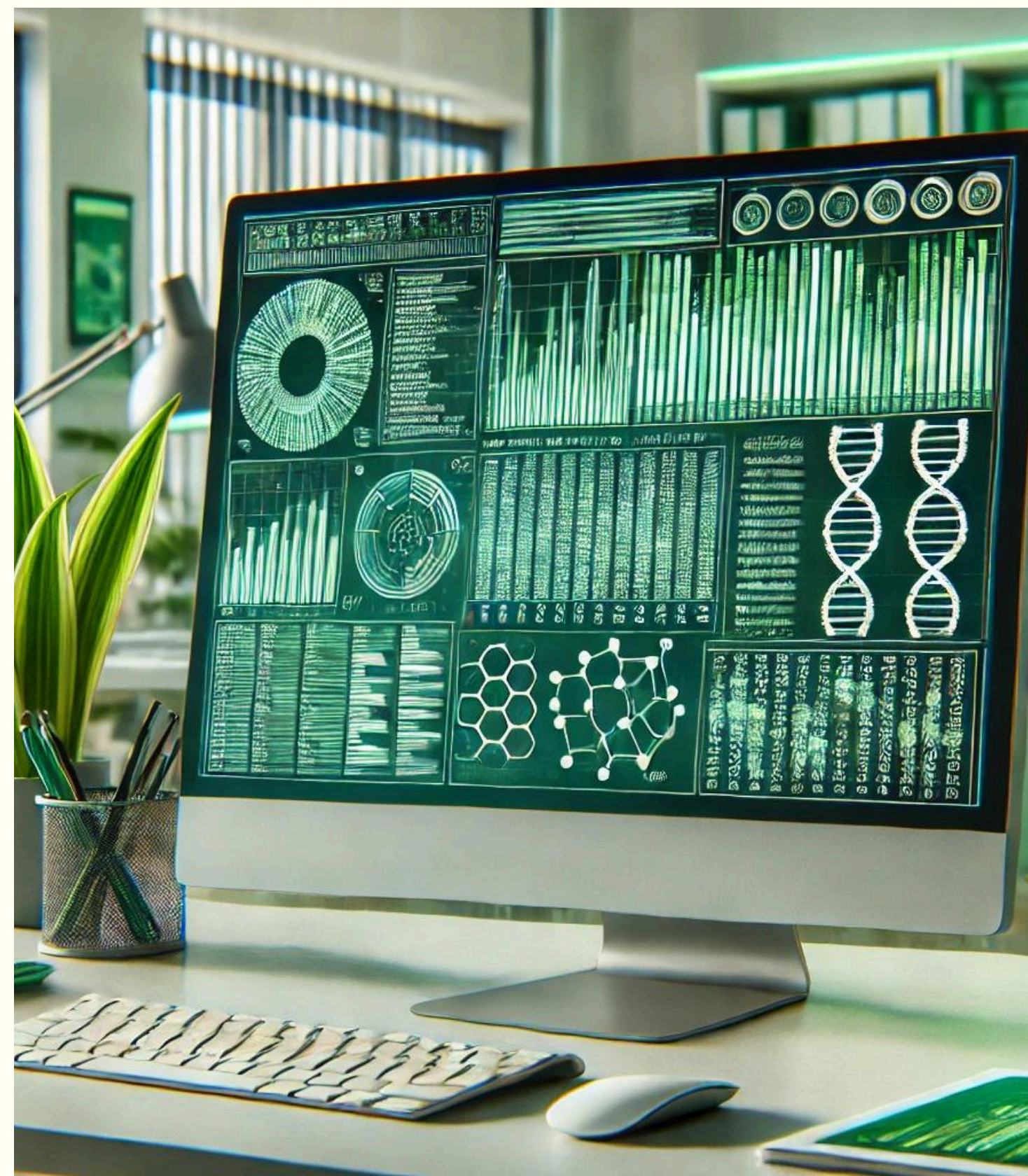
ANALISI DELLE VARIAZIONI

Il progetto mira a identificare e analizzare le variazioni puntuali nei genomi di riferimento. Utilizzando MAFFT per l'allineamento delle sequenze e uno script Python per l'analisi, vogliamo cercare di individuare e riportare le sostituzioni, inserimenti e cancellazioni rispetto al riferimento.

Giorgio Luigi Maria Bernasconi 885948

Alessio Farioli 879217

Silvia Cambiago 879382



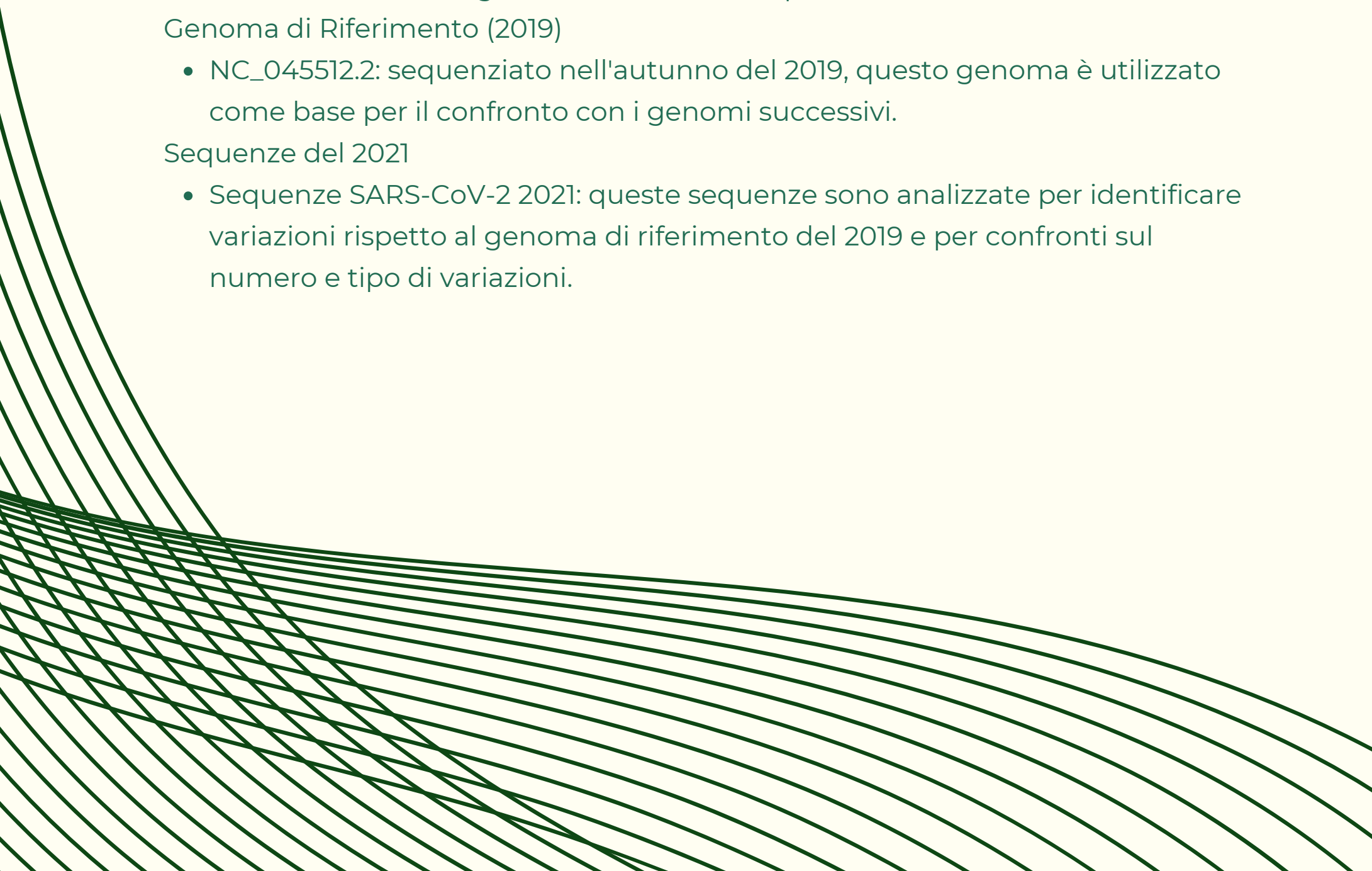


A PROPOSITO DEL PROGETTO

Ci sono stati forniti vari genomi derivati dal sequenziamento di SARS-CoV-2
Genoma di Riferimento (2019)

- NC_045512.2: sequenziato nell'autunno del 2019, questo genoma è utilizzato come base per il confronto con i genomi successivi.

Sequenze del 2021

- Sequenze SARS-CoV-2 2021: queste sequenze sono analizzate per identificare variazioni rispetto al genoma di riferimento del 2019 e per confronti sul numero e tipo di variazioni.
- 

Quali risultati cerchiamo?

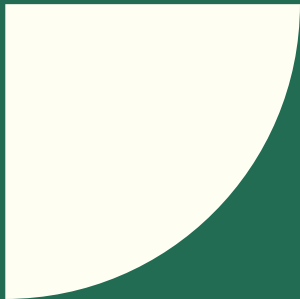
NUMERO DI VARIAZIONI

- Identificare il genoma con il maggior numero di variazioni rispetto al riferimento.
- Identificare il genoma con il minor numero di variazioni rispetto al riferimento.

POSIZIONE CON VARIAZIONE COSTANTE

- Elencare le posizioni del genoma di riferimento rispetto a cui tutti gli altri genomi variano.

POSIZIONI CON VARIAZIONI COMUNI

- Elencare le posizioni del genoma di riferimento rispetto a cui tutti gli altri genomi variano allo stesso modo.
- 

ABBIAMO DECISO DI DIVIDERE IL PROBLEMA

LETTURA FILE FASTA

Viene letto un file di tipo FASTA
prodotto da MAFFT

PREPARAZIONE DELLA MATRICE

I genomi puliti vengono inseriti in
una matrice, per confrontarli in
maniera più agevole.

IDENTIFICAZIONE DELLE VARIAZIONI

Mediante una matrice booleana
vengono segnalate tutte le
variazioni (senza distinguerle)

CLASSIFICAZIONE DELLE VARIAZIONI

Si itera sulla matrice stabilendo
quale variazione si è presentata

OUTPUT

Viene prodotto un output dettagliato
che elenca il tipo di variazione, le basi
coinvolte, la posizione e il numero di
genomi che la presentano.

CASI SPECIFICI DA RICERCARE

La ricerca dei casi specifici avviene
mediante l'utilizzo di appositi
contatori e confronti.



PANORAMICA DEI METODI

1

READ FASTA

Legge un file in formato FASTA ed estrae le sequenze e i relativi nomi.

2

PREPARE_MATRIX_FROM_FASTA_ALIGNED

Legge il contenuto di un file FASTA già allineato e lo struttura in una matrice per il confronto delle sequenze.

3

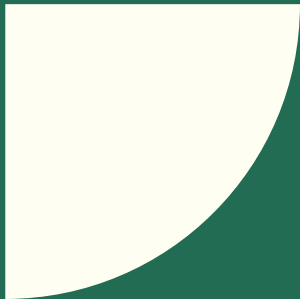
IDENTIFY_VARIATIONS

Identifica le variazioni e crea una matrice booleana che indica in una data posizione la presenza o l'assenza di variazione rispetto alla sequenza di riferimento.

4

PRINT_VARIATIONS

Individua la tipologia e stampa le variazioni rispetto alla sequenza di riferimento, producendo inoltre il file di report.



PANORAMICA DELL'OUTPUT

POINT 1

Per ogni genoma vengono restituite le variazioni (inserimento, cancellazione, sostituzione), mostrando la posizione e le basi coinvolte

POINT 2

Vengono mostrati i genomi con più e meno variazioni, mostrando il riferimento al nome e il numero di variazioni

POINT 3

Infine vengono mostrate tutte le posizioni per cui tutti i genomi variano rispetto al reference e le posizioni che variano rispetto al reference allo stesso modo.

```
output_variations.txt
Il genoma con più variazioni è OL700538.1, che presenta 60 variazioni
Il genoma con meno variazioni è OL700521.1, che presenta 49 variazioni

alla posizione 186 c'è una sostituzione C-->T in 1 genoma
alla posizione 210 c'è una sostituzione G-->T in 13 genomi
alla posizione 241 c'è una sostituzione C-->T in 13 genomi
alla posizione 521 c'è un inserimento di G in 1 genoma
alla posizione 522 c'è un inserimento di T in 1 genoma
alla posizione 523 c'è un inserimento di T in 1 genoma
alla posizione 1048 c'è una sostituzione G-->T in 1 genoma
alla posizione 1244 c'è una sostituzione G-->A in 1 genoma
alla posizione 1371 c'è una sostituzione A-->G in 1 genoma
alla posizione 1616 c'è una sostituzione C-->A in 1 genoma
alla posizione 1684 c'è una sostituzione C-->T in 2 genomi
alla posizione 1843 c'è una sostituzione G-->T in 1 genoma
alla posizione 1889 c'è una sostituzione C-->T in 1 genoma
alla posizione 2462 c'è una sostituzione C-->T in 1 genoma
alla posizione 2929 c'è una sostituzione A-->G in 1 genoma
alla posizione 3037 c'è una sostituzione C-->T in 13 genomi
alla posizione 3096 c'è una sostituzione C-->T in 1 genoma
alla posizione 3259 c'è una sostituzione G-->T in 1 genoma
alla posizione 3792 c'è una sostituzione C-->T in 1 genoma
alla posizione 3923 c'è una sostituzione C-->T in 1 genoma
alla posizione 3948 c'è una sostituzione A-->G in 1 genoma
alla posizione 4181 c'è una sostituzione G-->T in 11 genomi
alla posizione 4201 c'è una sostituzione G-->T in 1 genoma
alla posizione 4414 c'è una sostituzione A-->G in 1 genoma
alla posizione 5164 c'è una sostituzione G-->T in 2 genomi
alla posizione 5184 c'è una sostituzione C-->T in 2 genomi
alla posizione 5192 c'è una sostituzione C-->T in 1 genoma
alla posizione 5213 c'è una sostituzione T-->C in 1 genoma
alla posizione 5284 c'è una sostituzione C-->T in 1 genoma
alla posizione 5584 c'è una sostituzione A-->G in 2 genomi
alla posizione 6013 c'è una sostituzione A-->G in 1 genoma
alla posizione 6040 c'è una sostituzione C-->T in 1 genoma
alla posizione 6402 c'è una sostituzione C-->T in 11 genomi
alla posizione 6408 c'è una sostituzione C-->T in 1 genoma
alla posizione 6616 c'è una sostituzione A-->G in 1 genoma
alla posizione 6865 c'è una sostituzione G-->T in 1 genoma
alla posizione 7124 c'è una sostituzione C-->T in 11 genomi
alla posizione 7393 c'è una sostituzione G-->T in 1 genoma
alla posizione 7926 c'è una sostituzione C-->T in 1 genoma
alla posizione 8131 c'è una sostituzione G-->T in 1 genoma
alla posizione 8174 c'è una sostituzione G-->A in 1 genoma
alla posizione 8349 c'è una sostituzione G-->A in 1 genoma
alla posizione 8642 c'è una sostituzione G-->A in 1 genoma
alla posizione 8829 c'è una sostituzione C-->T in 2 genomi
alla posizione 8956 c'è una sostituzione C-->T in 1 genoma
alla posizione 8964 c'è una sostituzione C-->T in 1 genoma
alla posizione 8986 c'è una sostituzione C-->T in 11 genomi
alla posizione 9053 c'è una sostituzione G-->T in 11 genomi
alla posizione 9072 c'è una sostituzione C-->T in 1 genoma
```


VARIAZIONI PIÙ IMPORTANTI

OL700538.1

OL799538.1 è il genoma che ha mostrato il maggior numero di variazioni: 60

OL700521.1

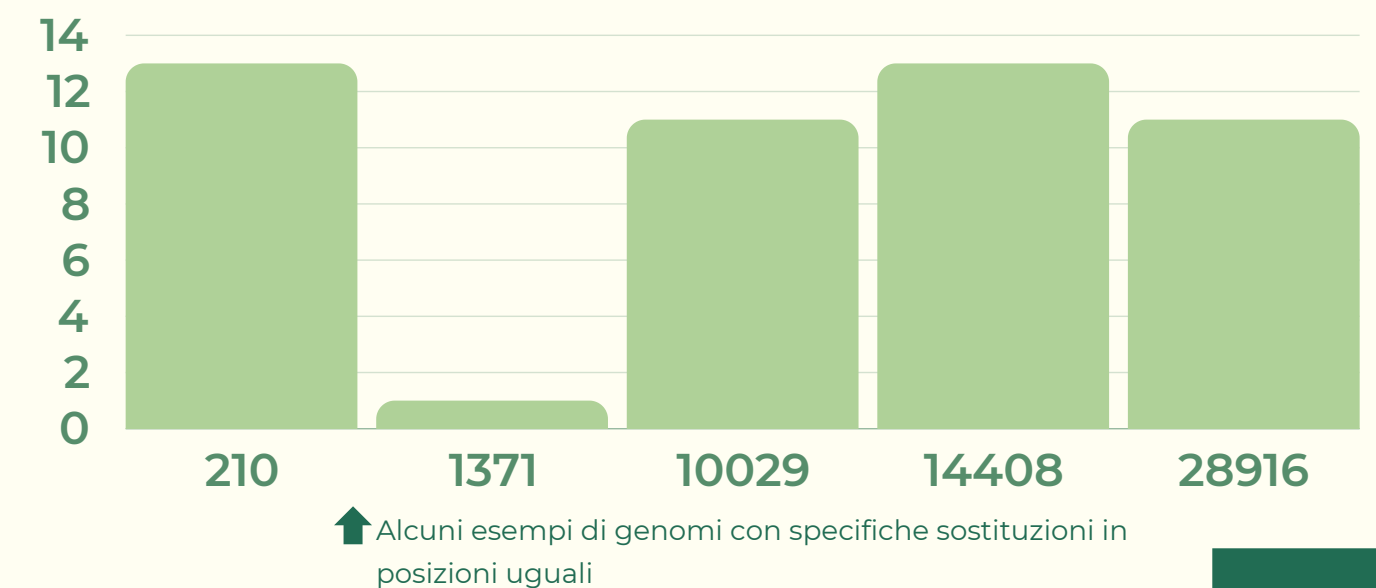
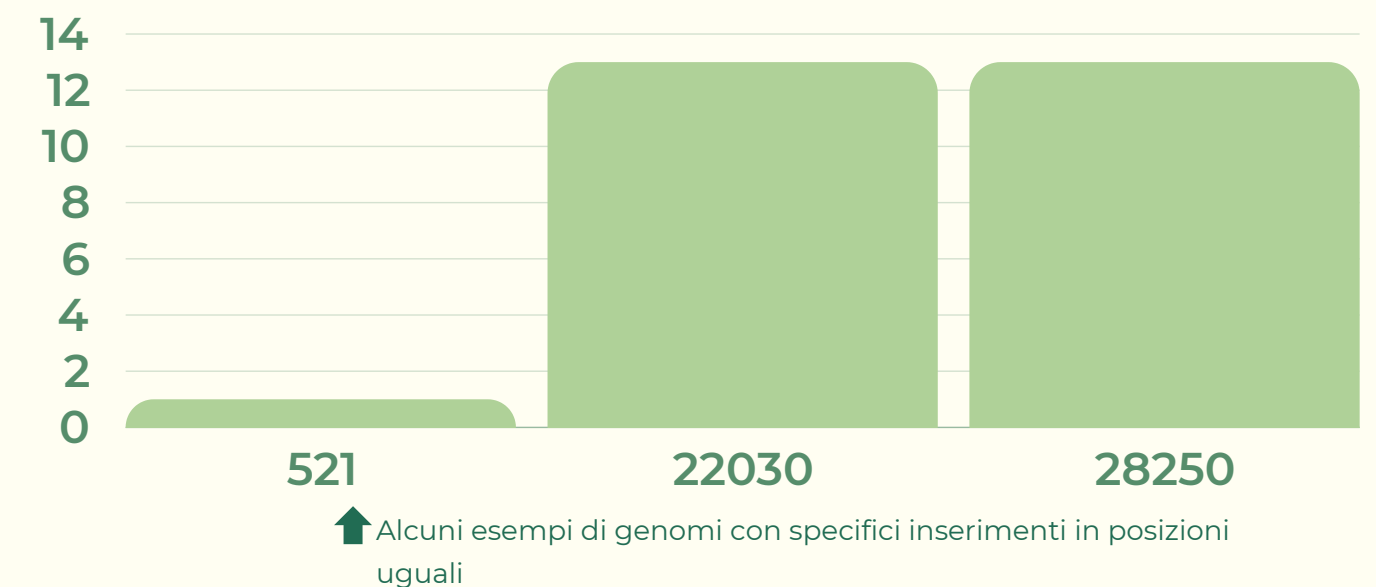
OL700521.1 è il genoma che ha mostrato il minor numero di variazioni: 49

CANCELLAZIONI

Il nostro script ha individuato la presenza di 0 cancellazioni

SOSTITUZIONI

Sono presenti mediamente più sostituzioni che inserimenti



PANORAMICA DEI DATI

Abbiamo raggiunto un risultato che ci ha portato a ottenere molteplici dati.

I dati più rilevanti:

- Abbiamo individuato il genoma di SARS-CoV-2 2021 che presenta il maggior numero di variazioni rispetto ai dati raccolti nel 2019: OL700538.1, con le sue 60 variazioni
- Abbiamo anche individuato il genoma di SARS-CoV-2 2021 che presenta il minor numero di variazioni rispetto ai dati raccolti nel 2019: OL700521.1, con le sue 49 variazioni

Da ciò si evince che il numero di variazioni nei 13 genomi si attesta in una soglia tra le 49 e le 60 variazioni

Osservando i risultati ottenuti è saltato subito all'occhio la totale assenza di cancellazioni. Esaminando la sequenza di riferimento abbiamo capito il motivo, non essendo presenti “-” all'interno della sequenza non possono verificarsi delezioni.



COSE CHE NON SONO STATE VISTE



Sono presenti alcune funzionalità gestite dal nostro codice ma che non compaiono visibili nel codice:

- Individua cancellazioni: come detto precedentemente non sono presenti cancellazioni, ma se fossero presenti il codice le individuerebbe
- Variazioni differenti stessa posizione: il codice segnalerebbe se, per una data posizione del reference, tutti i genomi variassero, anche in maniera differente
 - Con i nostri dati non accade: se tutti i genomi variano in una posizione, la mutazione è la stessa