

Multiple Sequence Alignment (MSA) di sequenze SARS-CoV-2

EDOARDO SILVA 816560
DAVIDE MARCHETTI 815990

A.A.: 2019/2020

Abstract

Il SARS-CoV-2 (dall'inglese *Severe Acute Respiratory Syndrome CoronaVirus 2*), è un ceppo virale della specie SARS-related coronavirus/SARS-CoV, facente parte del genere Betacoronavirus (ceppo di virus a RNA).

Il virus è stato sequenziato genomicamente dopo un test di acido nucleico effettuato su un campione prelevato da un paziente colpito da una polmonite, di cui non si conosceva la causa, ad inizio Dicembre 2019 a Wuhan, città continentale a est della Cina.

L'obiettivo del progetto consiste nell'analizzare, allineare ed identificare le differenze con la sequenza di riferimento prelevata su un campione di Wuhan, progettando un formato di output nel quale memorizzare i risultati ottenuti.

Usando i sequenziamenti genomici del virus denominato Covid-19, reperibili tramite NCBI¹ e GISAID² e sfruttando gli strumenti messi a disposizione dall'European Bioinformatics Institute³ abbiamo effettuato l'allineamento di un insieme di sequenze relative a paesi del medioriente.

¹<https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/>

²<https://www.gisaid.org/>

³<https://www.ebi.ac.uk/Tools/msa/>

1 Sequenze Analizzate

In aggiunta alla sequenza di riferimento di Wuhan, ne sono state selezionate alcune relative a paesi dell'area mediorientale.

Reference di Wuhan

NC_045512.2 pubblicata il 17/01/2020 (ultimo aggiornamento)

Iran

MT320891.2 pubblicata il 10/04/2020

MT281530.2 pubblicata il 04/04/2020

EPI_ISL_442523 sequenziata il 09/03/2020

EPI_ISL_437512 sequenziata il 26/03/2020

Israele

MT276598.1 sequenziata il 02/04/2020

MT276597.1 sequenziata il 02/04/2020

EPI_ISL_447469 sequenziata il 14/04/2020

Pakistan

MT262993.1 pubblicata il 25/03/2020

MT240479.1 pubblicata il 25/03/2020

EPI_ISL_417444 sequenziata il 04/03/2020

Turchia

MT327745.1 pubblicata il 13/04/2020,

EPI_ISL_437334 sequenziata il 24/03/2020

EPI_ISL_437317 sequenziata il 27/03/2020

2 Strumenti utilizzati

L'analisi è stata effettuata utilizzando Clustal Omega e MUSCLE, due strumenti per l'allineamento di sequenze multiple (MSA) accessibili attraverso un'interfaccia web^{4 5}.

3 Analisi preliminare

Una prima analisi delle sequenze è stata effettuata con l'ausilio di Jalview⁶, un software open source che offre la possibilità di generare una visualizzazione grafica degli allineamenti effettuati dai tool.

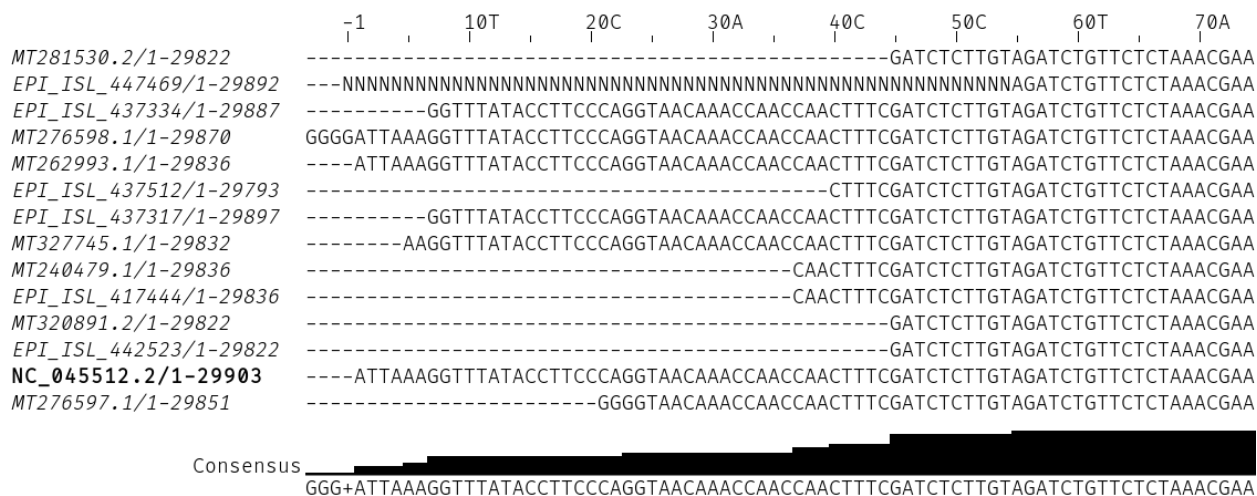


Figura 1: Differenze all'inizio dell'allineamento

Tramite questo strumento è possibile notare aspetti interessanti: esaminando l'allineamento la maggior parte delle differenze sono concentrate agli estremi dello stesso (fig. 1 e 2).

⁴<https://www.ebi.ac.uk/Tools/msa/clustalo/>

⁵<https://www.ebi.ac.uk/Tools/msa/muscle/>

⁶<https://www.jalview.org/>

4 Struttura del codice

L'esecuzione dell'analisi prevede, per ogni file di allineamento fornito in input, un corrispondente file di output che raggruppa le regioni di non match delle sequenze analizzate.

Lo script è composto da diversi file, ciascuno con la propria funzione specifica:

`alignment_model.py`: definisce le strutture dati che verranno utilizzate dal parser per memorizzare i contenuti letti.

`matcher.py`: contiene la classe che identifica le regioni di match/mismatch.

`main.py`: entrypoint dello script che coordina l'esecuzione dell'analisi.

`parsers.py`: definisce una classe per ogni tool di allineamento utilizzato per permettere di effettuare il parsing in una struttura dati omogenea.

`utils.py`: contiene utility generiche e di supporto per l'input/output.

5 Formato di Output

L'output dell'analisi prevede un file json per ogni allineamento elaborato ed un file json di confronto dei due tool di allineamento raggruppando gli allineamenti identici.

La struttura dell'output di analisi di un singolo allineamento comprende:

`tool`: tool di allineamento utilizzato.

`timestamp`: data e ora dell'analisi.

`reference`: identificativo della sequenza reference.

`analyzed_sequences`: identificativi delle sequenze allineate.

`mismatch`: vettore contenente i gruppi di mismatch identificati dall'hash della posizione di inizio e fine:

`from`: posizione di partenza rispetto alla sequenza di reference.

`to`: posizione finale del mismatch.

`reference`: basi della sequenza reference da `from` a `to`.

`alt`: basi delle sequenze non coincidenti con il reference da `from` a `to`.

`sequences`: sequenze nelle quali si verifica il mismatch.

Come riportato nel listato 1, la struttura dell'output di analisi del confronto tra due tool, prevede un'informazione aggiuntiva a ciascun gruppo di mismatch:

tools: contiene i tool che hanno rilevato quel particolare mismatch

Listing 1: Esempio di output

```
{
  "tools": ["clustal", "muscle"],
  "timestamp": "2020/05/14 12:05:18 UTC+0200",
  "reference": "NC_045512.2",
  "analyzed_sequences": ["MT281530.2", "EPI_ISL_447469", "
    ↪ EPI_ISL_437334", "EPI_ISL_437512", "MT276598.1", "
    ↪ EPI_ISL_437317", "MT240479.1", "EPI_ISL_417444", "MT327745
    ↪ .1", "MT320891.2", "EPI_ISL_442523", "MT276597.1", "
    ↪ MT262993.1"],
  "mismatches": {
    ...,
    "3715055078265645856": {
      "from": 3040,
      "to": 3041,
      "reference": "C",
      "alt": "T",
      "sequences": ["EPI_ISL_447469", "MT276598.1"],
      "tools": ["muscle", "clustal"]
    },
    ...
  }
}
```

6 Analisi dei risultati

In questa sezione veguiranno dei grafici dei risultati ottenuti e una loro breve spiegazione.

6.1 Variazioni

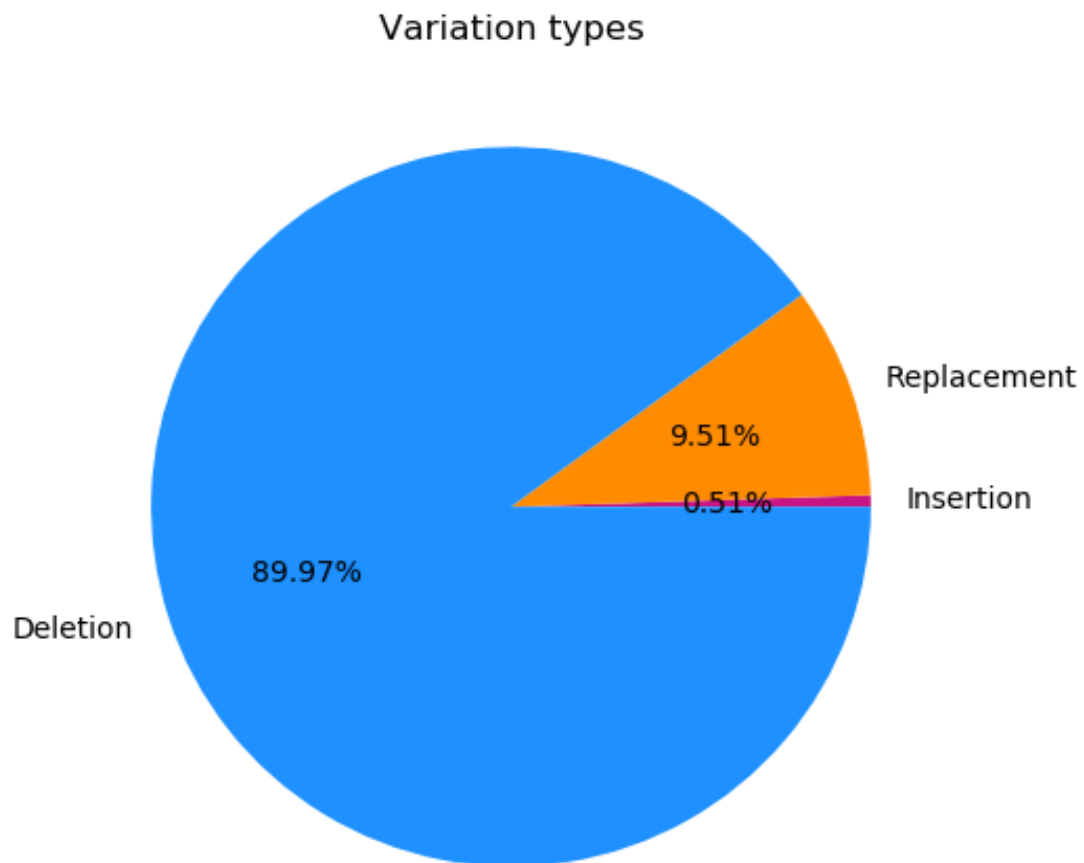


Figura 4: Tipologia di variazioni

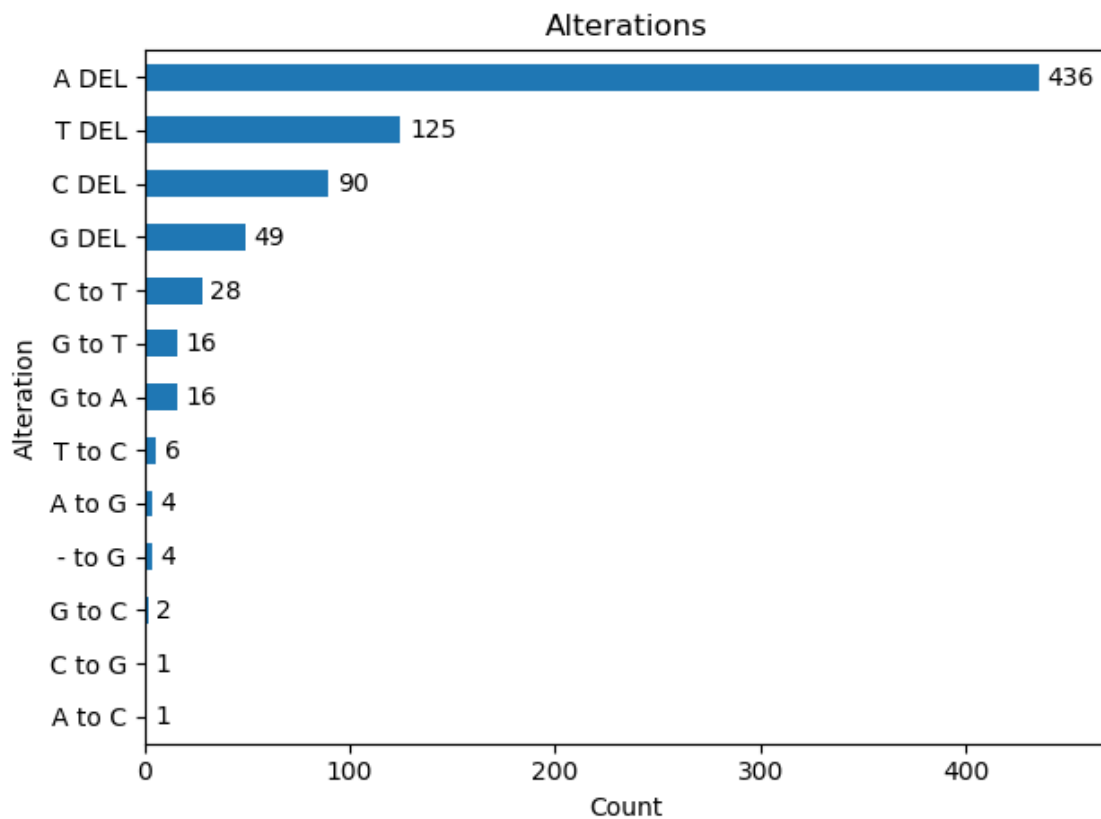


Figura 5: Tipologia di variazioni specifiche

La causa principale di mismatch pare essere una cancellazione (89,97% dei casi), principalmente molte **adenine** risultano mancanti rispetto alla reference, probabilmente compongono la maggior parte delle basi negli estremi non codificanti.

6.2 elementi mancanti

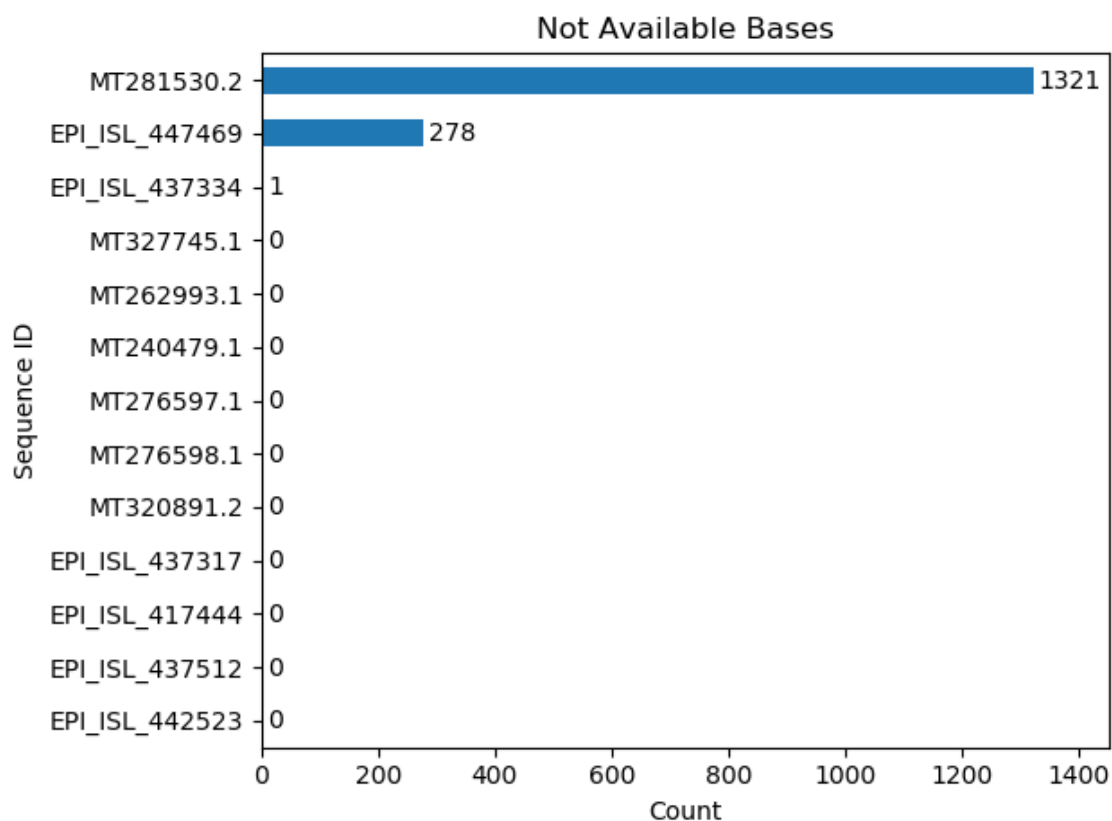


Figura 6: ??

Questo grafo mostra come ??? porta a supporre che 2 delle sequenze esaminate siano contengano molti errori di sequenziamento (MT281530.2 iraniano) è un particolare problema, l'errore possibile in EPI_ISL_437334 è trascurabile.

6.3 Mismatch

I prossimi 3 grafi mostrano la tipologia di mismatch, una ogni grafo, distribuita nelle varie sequenze.

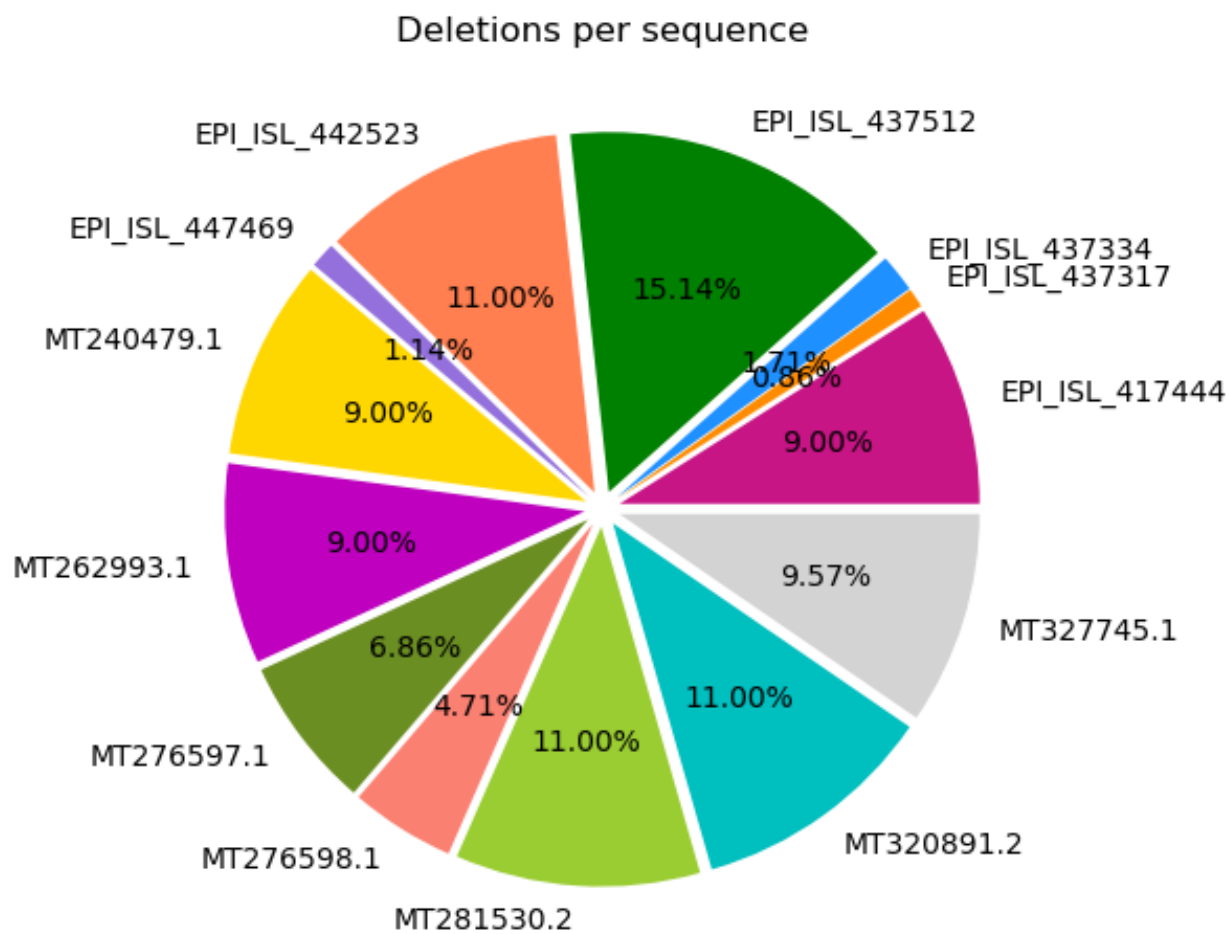


Figura 7: Cancellazioni

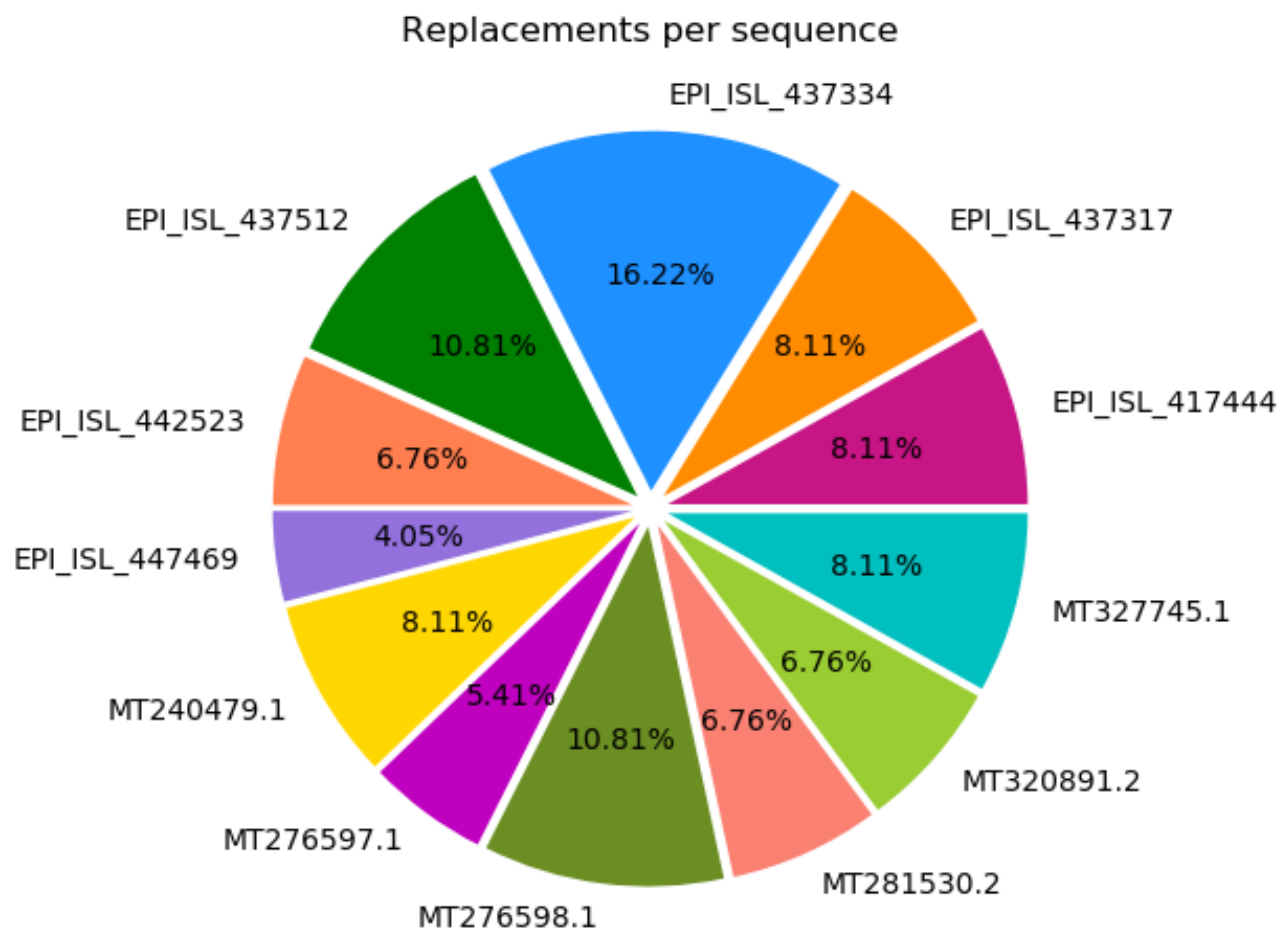


Figura 8: sostituzioni

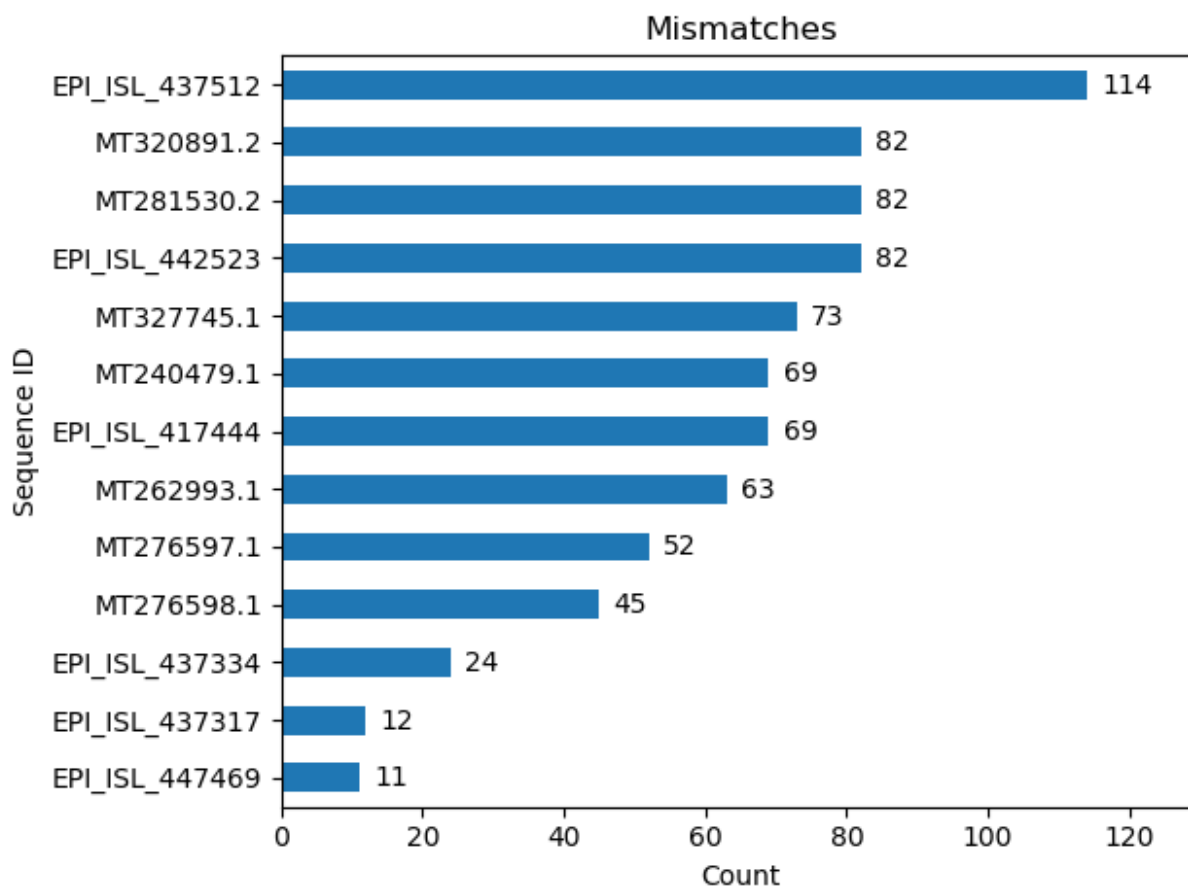


Figura 9: mismatches

7 Conclusioni

Dall'analisi preliminare si nota che le sequenze non hanno tutte lo stesso numero di basi, ma la maggior parte dei mismatch si manifesta agli estremi delle sequenze.

Attraverso il confronto degli allineamenti effettuati con Clustal Omega e MUSCLE, la sequenza MT262993.1 è quella che presenta evidenti cancellazioni di più basi contigue che non siano posizionate ai capi della sequenza stessa. In particolare, sono state rilevate due cancellazioni considerevoli rilevate da entrambi i tool: da 9724 a 9739; da 19519 a 19540. La sequenza

EPI_ISL_437334 pare essere sia quella con più sostituzioni rispetto al reference; mentre EPI_ISL_437512 è la sequenza con un maggior numero di mismatch complessivi.

Tuttavia, non possiamo determinare se l'incongruenza delle sequenze MT281530.2 e EPI_ISL_437334 rispetto alla reference sia determinata da un'effettiva mutazione piuttosto che da un errore in fase di sequenziamento della stessa. Le restanti variazioni sporadiche tra le sequenze risultano coinvolgere solamente singole basi.

7.1 Divisone del lavoro

Durante la realizzazione del progetto entrambi i componenti del gruppo hanno partecipato attivamente alla sua realizzazione. In particolare:

- **Edoardo Silva** si è occupato principalmente di recuperare le sequenze identificate su NCBI e GISAID e della generazione dei file di output.
- **Davide Marchetti** si è occupato principalmente di scrivere i parser per i tool di allineamento e dell'elaborazione dei gruppi di mismatch per generare il confronto tra i due tool.
- Entrambi hanno lavorato all'identificazione dei mismatch tra sequenze sui singoli tool e collaborato alla stesura di questo documento.