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

Edoardo Silva 816560 Davide Marchetti 815990

A.A.: 2019/2020

1 Abstract

La seconda parte del progetto prevede di elaborare i file prodotti in precedenza ricavando informazioni relative alle alterazioni rilevate e producendo in output una tabella riassuntiva contenente:

- il gene id del gene in cui cade la variazione con lo start e l'end della sua CDS rispetto alla reference
- il codone (o i codoni) alterato della reference, con posizione di inizio rispetto alla CDS, sequenza del codone e amminoacido codifcato
- il nuovo codone generato dalla variazione (o i nuovi codoni generati) specificando la sequenza del codone e il nuovo amminoacido codificato

2 Algoritmo

L'algoritmo inizia caricando tutti i file necessari per l'elaborazione, in particolare quelli prodotti in output nella parte precedente del progetto:

- 1. Caricamento della sequenza reference dal file corrispondente memorizzato in /project-1/input/reference.fasta.
- Caricamento di uno dei file di output prodotti nella prima parte di progetto. Nel nostro caso è stato utilizzata l'analisi dell'allineamento di ClustalW.
- 3. Lettura del file Genes-CDS.xlsx contenente le informazioni sui geni e le CDS della sequenza di reference. In particolare, per le CDS che derivano dalla join di due sequenze è possibile specificare il punto di unione della sequenza.

Dopo la lettura del materiale rilevante a questa fase di elaborazione, l'algoritmo itera le variazioni rilevate nell'allineamento e per ciascuna di esse esegue i seguenti step:

- 1. Trova le CDS nelle quali avviene l'alterazione rispetto alla reference.
- 2. Recupera le informazioni del gene associato alle CDS rilevate calcolando le posizioni globali e relative alla CDS dell'alterazione.
- 3. Identifica i codoni alterati e ne effettua la ritraduzione in amminoacidi grazie ad una look-up table (listato 1). Vengono ignorate le alterazioni che presentano sequenze di soli -, derivate probabilmente da un sequenziamento errato o un'alterazione posta ai capi dell'allineamento.
- 4. Memorizza tutte le informazioni ricavate in una struttura dati tramite cui derivare la tabella per l'output finale associando i valori a chiavi prestabilite (listato 2).

Al termine dell'elaborazione di tutte le alterazioni, viene costruito un oggetto di tipo DataFrame fornito dalla libreria pandas.

Le chiavi utilizzate nella costuzione della struttura dati a lista diventeranno le colonne del DataFrame. Questo sarà esportato in CSV nella cartella /project-2/output/alteration-table.csv per permettere una visualizzazione più semplice tramite programmi terzi (come riportato in fig. 1)

3 Informazioni memorizzate

Ad ogni variazione analizzata corrisponde un entrata nella struttura dati a lista contenente le seguenti informazioni:

- gene id: id del gene in cui cade la variazione
- gene_start: inizio del gene in cui cade la variazione (1-based)
- gene end: fine del gene in cui cade la variazione (1-based)
- cds_start: inizio della Coding DNA Sequence della porzione del gene in cui cade la variazione (1-based)
- cds_end: fine della Coding DNA Sequence della porzione del gene in cui cade la variazione (1-based)
- relative_start: inizio della variazione in rispetto all'inizio della cds (1-based)
- relative_end: fine della variazione in rispetto all'inizio della cds (1-based)
- alteration: sequenza della variazione
- original codone: codone della reference prima della modifica
- original aminoacid: amminoacido codificato da original codone
- altered codone: codone della reference modificati dalla variazione
- encoded aminoacid: amminoacido codificato da altered codone

4 Output

Come riportato in fig. 1 la maggior parte delle alterazioni coinvolgono un singolo codone e quelli ottenuti rimangono traducibili.

In alcuni casi, l'amminoacido risultante dalla traduzione dell'alterazione non viene modificato. La maggior parte delle variazioni si concentrano nel gene ORF1ab identificato da gene_id = 43740578.

Le ultime righe della tabella riportano delle alterazioni che determinano la cacellazione di alcune basi rispetto alla sequenza reference. Queste sono relative solo alla sequenza MT262993.1 e si pensa possano derivare da un errore in fase di sequenziamento.

15.5. 25.5. <th< th=""><th>gene_id 💌 gene_start 💌 gene_end</th><th>ne_start 🔽 ge</th><th>ne_end 🔽 cd:</th><th>s_start 🔽 co</th><th>s end r relat</th><th>ive_start 💌 rela</th><th>tive_end alteration</th><th>✓ original_codone</th><th>v original aminoa</th><th>cid valtered_codone</th><th><pre>encoded_aminoacid</pre></th></th<>	gene_id 💌 gene_start 💌 gene_end	ne_start 🔽 ge	ne_end 🔽 cd:	s_start 🔽 co	s end r relat	ive_start 💌 rela	tive_end alteration	✓ original_codone	v original aminoa	cid valtered_codone	<pre>encoded_aminoacid</pre>
267 21552 267 267 267 267 267 267 267 267 267 267 267 267 267	43740578	267	21555	267	13483	1131	1131 A	GTA	^	ATA	_
267 21555 267 21545 267 1348 10817 176 17 17 267 21555 267 21555 267 21555 18011 18017 17 17 17 17 267 21555 267 21555 267 21555 18011 18017 17 17 17 17 267 21555 21555 21555 2	43740578	267	21555	267	21555	1131	1131 A	GTA	>	ATA	_
287 2155 287 2155 287 2815 1855 1855 1887 1810 110	43740578	267	21555	267	13483	10817	10817 T	TTG	7	Ш	ш
207 21552 208 208 18811 1811 ACA T TCA 207 21553 2057 12583 2057 12583 2091 ACA T ACA CAA T ACA A	43740578	267	21555	267	21555	10817	10817 T	TTG		Ш	u.
2877 2853 2875 2853 2875 <th< td=""><td>43740578</td><td>267</td><td>21555</td><td>267</td><td>21555</td><td>18111</td><td>18111 T</td><td>ACA</td><td>_</td><td>TCA</td><td>S</td></th<>	43740578	267	21555	267	21555	18111	18111 T	ACA	_	TCA	S
267 21555 267 13483 793 793 T ACC 1 ACC	43740575	28275	29533	28275	29533	1100	1100 A	GAG	ш	GAA	Е
267 21555 266 12555 277 2771 TMC F TMT 267 21555 266 12555 277 2771 TMC F TMT 267 21555 267 13583 267 13584 2771 2771 TMC F TMT 267 21555 267 13583 363 3827 CCA P CTA 267 21555 267 13583 369 3827 CCA P CTA 267 13555 267 13583 369 368 9248 CA CA CA CA CA 267 13555 267 1358 354 3827 CCA CA C	43740578	267	21555	267	13483	793	793 T	ACC	_	ATC	. —
267 21555 269 11556 2771 2771 TTT T	43740578	267	21555	267	21555	793	793 T	ACC	⊢	ATC	_
267 21555 267 11555 267 11555 267 11555 1154 4492 117 1171	43740578	267	21555	267	13483	2771	2771 T	TTC	ш	Ш	ш
267 21555 267 13483 3467 3487 CCC P CTA 267 21555 267 21835 367 31827 CCC P CTA 267 21555 267 21585 367 31826 367 3187 CCC P CTA 267 21555 267 21585 267 21585 267 21586 1110 1110 1110 CCC T T CCC 267 21555 267 21585 267 21586 1110 1110 3110 1110 1110 CCC T T CCC CCC T T CCC CCC T T CCC	43740578	267	21555	267	21555	2771	2771 T	TTC	ш	Ш	ш
267 21555 267 13483 3857 3857 CCA P CTA 287 21555 267 13483 3857 3857 CCA C 287 21555 267 13483 3928 3948 TTA I TTA CTA 287 21555 267 13483 13110 13110 ACC T C CCA 287 21555 2155 13120 13100 ACC T C CC 287 21555 2155 1320 13200 GC AC T CC 287 21555 267 13483 1310 13100 GC AC T GC 287 21555 267 13483 1310 13100 GC AC T CC 287 21554 2348 414 417 CC CC AC CC 287 2252 2348 4	43740578	267	21555	267	21555	14142	14142 T	CCT	Ь	TCT	S
267 21555 367 3887 3887 CCA P CTA 279 21555 267 13483 9248 9748 17A 1 TTG 279 21555 267 13483 13110 13110 4CC 1 TTG 270 21555 267 13483 13110 13110 4CC 1 TTG 270 21555 267 13483 13210 13101 4CC 1 TTG 270 21556 21550 1320 13201 4CC 1 TTG 2877 21556 21520 13100 13101 4CC TTG TTG 2877 21556 2753 1320 13101 1400 GC CC TTG 2877 21556 2753 2441 441C TG CC CC TTG 2877 21554 2358 461 441C TG CC CC <	43740578	267	21555	267	13483	3637	3637 T	CCA	Ь	CTA	
267 21555 267 13483 9248 9248 TTA I TTG 267 21555 267 13483 1310 1310 G TTA I TTG 267 21555 267 13483 13110 1310 G T T CC 267 21555 267 13483 1310 1310 G T T CC 267 21555 267 1355 1320 1310 1310 GC AC T GC 267 21555 267 1355 1320 1310 1310 GC C T CC 267 21554 267 267 1360 1310 <td>43740578</td> <td>267</td> <td>21555</td> <td>267</td> <td>21555</td> <td>3637</td> <td>3637 T</td> <td>CCA</td> <td>Ь</td> <td>CTA</td> <td></td>	43740578	267	21555	267	21555	3637	3637 T	CCA	Ь	CTA	
267 2155 267 21555 267 21555 267 21555 267 21555 267 13150 34310 4CC 1 1 CCC 267 21555 267 21555 13150 13110 13110 13110 140 1 CC CC 1 CC CC 1 CC 1 CC CC 1 CC CC <t< td=""><td>43740578</td><td>267</td><td>21555</td><td>267</td><td>13483</td><td>9248</td><td>9248 G</td><td>ТТА</td><td>7</td><td>TTG</td><td>7</td></t<>	43740578	267	21555	267	13483	9248	9248 G	ТТА	7	TTG	7
867 11555 267 13483 13110 13110 ACC T CCC 267 21555 267 13483 13110 13110 GCG T T GC 267 21555 267 13483 13210 13210 GCG A GC GC 267 21555 267 21555 267 21553 1321 13210 GCG A GC	43740578	267	21555	267	21555	9248	9248 G	TTA	_	TTG	
267 21555 267 13555 267 13555 267 13435 13110 31310 GCG T C 267 12555 267 13436 13210 13210 16C C T C 267 12555 267 13555 1321 19218 197 GC C T TC 2624 27191 26534 27191 197 197 GC GC T TC 2625 27191 26534 27191 197 197 GC GC T TC 28275 28343 2877 29533 2877 2654 1714 441 C GG GC CCA 2867 21556 2834 2155 2841 441 C GG GG GC CCA 2877 21556 2834 2156 2334 417 417 CC CCA 2877 21556 2834 136 </td <td>43740578</td> <td>267</td> <td>21555</td> <td>267</td> <td>13483</td> <td>13110</td> <td>13110 G</td> <td>ACC</td> <td>F</td> <td>209</td> <td>A</td>	43740578	267	21555	267	13483	13110	13110 G	ACC	F	209	A
267 21555 267 13483 13210 13210 GGG A GIG 287 21555 268 2185 13250 13210 170 GGG C TITC 2867 21555 2875 2583 2875 19210 19210 1767 C TTC TTC 28675 2583 2875 2933 441 441 TG TG TC TC 2877 21555 2873 2875 2933 8441 441 TG TG TC TC 2877 21555 287 21558 2861 166 GG GG GG 2877 21556 23584 464 441 TG TG TG TG 2877 21556 23584 562 2621 2621 AA AA TG TG TG TG 2877 21564 23584 49 441 TG TG	43740578	267	21555	267	21555	13110	13110 G	ACC	-	209	A
267 21555 267 21555 267 21555 267 21555 2157 19210 19210 16C CT TTC 26274 21554 27594 21555 19238 19218 19718 16C 17C 17C 28275 2533 28275 25933 28275 25933 28275 2662 27191 17C 17C 17C 28275 25834 28275 29333 28275 2664 44 17C 17C 17C 17C 287 21556 2934 2844 844 17C 1	43740578	267	21555	267	13483	13210	13210 T	909	A	GTG	>
267 2155 267 2155 267 19218 19218 CCT A TCT 28273 28234 28234 2491 419C GTG Y GTC 28275 29533 28273 28533 4441 441C TCA CCA CCA 28275 21553 2877 21553 8441 8441 GGT GGT G GCA 267 21555 267 21556 21564 2384 641 AAA GGT G GGC GCA 2767 21556 21564 2384 641 2601 GGT G GCA 2867 21558 2876 21564 2384 641 AAA AAAA AAAA AAAA AAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	43740578	267	21555	267	21555	13210	13210 T	TGC	O	TTC	u.
26524 2791 26524 2791 197 137 610 7 610 7 610 7 <td>43740578</td> <td>267</td> <td>21555</td> <td>267</td> <td>21555</td> <td>19218</td> <td>19218 T</td> <td>901</td> <td>A</td> <td>TCT</td> <td>S</td>	43740578	267	21555	267	21555	19218	19218 T	901	A	TCT	S
28275 29333 28275 29533 414 414 C ITG CCA 28275 28333 28671 261 551 C CCA 2877 21555 267 21555 267 267 CCA 287 21555 267 13555 267 1355 8441 641 GCT GC GC 21564 21555 267 1355 2621 1252 661 GC GC GC 21564 23584 1164 1524 641 641 AAA GG GC GC GC 21564 23584 1164 1172 1172 TC T AA AA 28275 2463 267 1172 1172 TC C TC TC 28275 2464 267 147 TC C TC TC 28275 2468 256 127 TC TC TC	43740571	26524	27191	26524	27191	197	197 C	GTG	>	GTC	>
2877 29533 28275 29533 561 TCA TCA CCA 267 21555 267 13483 8441 8441 66T 66C 66C 267 21555 267 13584 8441 8441 66T 66T 66C 267 21555 267 21558 20621 20621 20621 67T 67T 66C 21564 25384 21564 25384 1172 1172 10C 1 1 1 1 1 1 2 66C	43740575	28275	29533	28275	29533	414	414 C	TTG		CTG	_
267 21555 267 134883 8441 8441 GGT GGT GG 267 21555 267 21555 8441 8441 GGT GGT GG 267 21555 267 21555 2661 AAA KA AA AAA 21564 25384 21564 25384 1172 1707 CC C AAA 21564 25384 21564 25384 1172 1707 CC C C AAA 2167 21555 267 21555 267 21556 25334 2170 CC C C CT C CT C	43740575	28275	29533	28275	29533	561	561 C	TCA	S	CCA	d
267 21555 267 21555 8441 66T 66T 6 6 267 21555 267 21555 2661 AAA AAA AAA AAA 21564 23584 21564 2584 6 64 ACT T ATT 21564 25384 21564 2584 1172 1172 T TCC C T ATT 28275 25333 2875 2563 556 T TCC S TC TC TC TC 28275 25533 2875 2584 47 47 T CC C TC	43740578	267	21555	267	13483	8441	8441 C	GGT	9	399	9
267 21555 267 21555 267 21554 2172 T CC C C C C T TC 28775 21555 267 21555 267 21554 47 47 T C C C TTC 2877 21555 267 21555 267 21554 47 47 T C C C TTC 2877 21555 267 21554 47 47 T C C C TTC 2877 21554 21554 21554 21547 47 T C C C TTC 2877 21664 260 400 400 400 AC	43740578	267	21555	267	21555	8441	8441 C	GGT	Ŋ	299	g
21564 25384 21564 25384 21564 25384 21564 25384 1172 1172 TCC T T ATT ACT T T ATT ACT T T ATT ACT T T T ATT ACT T T T T T T ATT ACT T	43740578	267	21555	267	21555	20621	20621 A	AAA	×	AAA	~
21564 25384 21564 25384 21564 25384 21564 25384 1172 1172 TGC C TGC C TGC LGC	43740568	21564	25384	21564	25384	64	64 T	ACT	_	ATT	_
28275 29533 28275 29533 556 7 TCC 1 267 21555 267 13483 47 471 CTC L CTT 21564 21554 25384 13483 47 471 CTC L CTT 21564 25384 21546 25384 1840 1840 604 AC CT L CTT CTT 21564 25384 21585 8516 8516 AGC S AACGAACA AACGAACA AACGAACA AACGAACA AACGAACA AACGAACA AACGAACA AACGAACA AACGAACAACA AACGAACAACA AACGAACAACAACAACAACAACAACAACAACAACAACAAC	43740568	21564	25384	21564	25384	1172	1172 T	T6C	U	TGT	C
267 21555 267 13483 47 47 T CTC CTC CTT CTT 287 21554 21554 21554 21554 1836 67	43740575	28275	29533	28275	29533	556	556 T	TCC	S	TTC	ш
267 21555 267 21555 47 47 T CTC CTC CTT 21354 25384 21564 25384 1840 1840 G 6AT D GGT 28775 25384 21554 25384 2156 25384 8516 860AC 6GG 8 267 21555 267 13483 8516 8516 T 4GC 5 AGT 21564 25384 21554 25384 3751 4GC 5 AGT 21564 25384 21564 25384 3751 4GC 5 AGT 21564 25384 3751 3751 4GG 6 GG GG GG 21864 25384 3751 3751 4GG 5 AGT	43740578	267	21555	267	13483	47	47 T	CTC		CTT	_
21564 25384 21564 25383 21564 25383 1840 GG ACC GAT AACGAA 28275 22633 28275 22633 609 AC AGC AACGAA 266 21555 267 21585 8516 AGC S AGT 21564 21555 267 11583 8516 AGC S AGT 21564 21555 267 11583 816 AGC S AGT 21564 21584 21564 5384 3751 AGG S AGT 21564 25384 21564 3751 AGG AGG AGT AGT 21565 2825 2583 2825 250 AG AGT AGT 2877 21555 267 13483 618 CG AGT AGT 2877 21555 267 13483 8893 1082 CG AGT AGT 2876<	43740578	267	21555	267	21555	47	47 T	CTC	7	CTT	
2877 2953 2877 29533 607 AGGGGA RGGGGA RG AAACGA 267 21555 267 11483 8516 8516 T AGC S AGT 267 21555 267 11555 8516 8516 T AGC S AGT 21564 22384 21564 25384 905 T AGC S AGT 21564 25384 21564 25384 3751 AGC S AGT 21564 25384 21569 2502 2507 170 AGT AGT 21564 25384 2155 2604 604 AGT C AGT 2877 21555 267 13483 618 108 C P C 267 21555 267 13483 618 108 C C C C 267 21555 267 13483 8893 108 C C <t< td=""><td>43740568</td><td>21564</td><td>25384</td><td>21564</td><td>25384</td><td>1840</td><td>1840 G</td><td>GAT</td><td>D</td><td>GGT</td><td>9</td></t<>	43740568	21564	25384	21564	25384	1840	1840 G	GAT	D	GGT	9
267 21555 267 13483 8516 8516 T AGC S AGT 268 21556 2255 2567 21555 8216 T AGC S AGT 21564 21554 21564 21564 25384 3751 3751 T AGC S AGT 21564 25384 21564 25384 3751 3751 T AGC G T AGT 27895 28259 27895 28239 2503 250 C 170 TAA G AGT T AGT 28275 28275 28239 28275 2623 AGT T AGT T AGT	43740575	28275	29533	28275	29533	209	609 AAC	AGGGGA	RG	AAACGA	KR
267 21555 267 21555 8516 8616 AGC S 21564 25384 905 905 AGG S AGT 21564 25384 31564 5384 905 905 AGG G 21564 25384 3154 25384 3751 GG G GT 21785 28259 2539 250 250 250 TA C GT 2877 28275 2953 604 604 AGT S AAT 2877 21555 267 13483 618 G1 R G 267 21555 267 13483 1082 1082 C C P C 267 21555 267 13483 8893 1082 C C C P C 267 21555 267 13483 8893 1 C C P C <td< td=""><td>43740578</td><td>267</td><td>21555</td><td>267</td><td>13483</td><td>8516</td><td>8516 T</td><td>AGC</td><td>S</td><td>AGT</td><td>S</td></td<>	43740578	267	21555	267	13483	8516	8516 T	AGC	S	AGT	S
21364 23384 21564 25384 905 905 7 ACG ACG 7 ACG ACT 21364 23384 21564 25384 3731 3751 6GA 6A 6A <td>43740578</td> <td>267</td> <td>21555</td> <td>797</td> <td>21555</td> <td>8516</td> <td>8516 T</td> <td>AGC</td> <td>S</td> <td>AGT</td> <td>S</td>	43740578	267	21555	797	21555	8516	8516 T	AGC	S	AGT	S
21364 2384 21564 25384 3751 3751 6GA GGA GG TAA 27785 28759 28259 2829 250 250 17A 17A 17A 28775 21553 28775 29533 604 A 4GT 5 AT 267 21555 267 13483 618 618 7 GGT R 1GT 267 21555 267 13483 1082 1082 CC P 1GT 267 21555 267 13483 8893 1082 CC P CC 267 21555 267 13483 8893 1082 CC P CC 267 21555 267 21555 287 2313 2313 CC CC CC CC 267 21554 2524 2313 2313 CC CC CC CC CC 2855 <	43740568	21564	25384	21564	25384	905	905 T	ACG	⊢	ACT	_
2785 2825 2785 2826 250 TTA ITA ICA 28275 2833 28275 2933 604 604 AGT 5 AAT 267 21555 267 13483 618 108 1 R 1GT 267 21555 267 13483 618 1082 1 CC 1 1GT 267 21555 267 13483 1082 1 CC	43740568	21564	25384	21564	25384	3751	3751 T	GGA	g	GTA	^
28275 29533 28275 29533 604 604 A 4GT S AAT 267 21555 267 13483 618 1 GGT R IGT 267 21555 267 1355 618 1 GGT R IGT 267 21555 267 13483 1082 1 CC P CT IGT 267 21555 267 13483 8893 1082 C C P C	43740577	27895	28259	27895	28259	250	250 C	TTA		TCA	S
267 21555 267 13483 618 618 CGT R TGT 267 21555 267 21555 267 21555 267 13483 1082 1082 CGT P CGT 267 21555 267 21555 267 13483 1082 1082 CGT P CGT 267 21555 267 21555 267 13483 8893 T CGT P CGT 267 21556 21554 21554 21554 21554 21313 2313 CGT P CGT 28559 29674 2558 29674 5 5 GGC CGT P CGT 2857 21554 2558 29674 5 5 GGC CGT P CGT 2867 21555 267 21348 2313 2313 CGT CGT P CGT 287 21555 267 <td>43740575</td> <td>28275</td> <td>29533</td> <td>28275</td> <td>29533</td> <td>604</td> <td>604 A</td> <td>AGT</td> <td>S</td> <td>AAT</td> <td>Z</td>	43740575	28275	29533	28275	29533	604	604 A	AGT	S	AAT	Z
267 21555 267 21555 618 618 CGT R TGT 267 21555 267 21555 267 21555 267 21555 267 21555 267 21555 267 21555 267 21555 267 21557 2188 2893 7 CCT	43740578	267	21555	267	13483	618	618 T	CGT	œ	TGT	C
267 21555 267 13483 1082 1082 CCC P CCT 267 21555 267 21555 267 21555 267 21555 267 21555 267 21555 267 21555 267 21555 267 21557 2834 2313 2313 CCT P CTT 21556 22554 22554 2354 2313 2313 CTT CT ATT 267 21555 2674 5 5 6GC GTT ATT 267 21555 2674 5 5 6GC GTT ATT 267 21555 2674 5 5 6GC GTT CTT 267 21555 267 13483 9469 CATTCCATGGTCTT HFWHE CATTCCATGGTCTT 268 21555 267 21555 2469 CATTCCATGGTCTTATGGTCTTT HFWHE CATTCCATGGTCTTATGGTCTT	43740578	267	21555	267	21555	618	618 T	CGT	W.	TGT	C
267 21555 267 21555 267 21555 267 21555 267 21548 8893 1082 T CCT P 267 21555 267 21555 8893 8893 T CCT P 21564 21555 267 21555 8893 8893 T CCT P 29559 29674 29559 29674 5 7 GCT P 267 21555 267 1483 9456	43740578	267	21555	267	13483	1082	1082 T	222	Ь	CCT	Ь
267 21555 267 13483 8893 8893 T CCT P 267 21555 267 21555 8893 8893 T CCT P 21564 21554 21554 21554 21534 2313 A GT P 29579 29674 25 5 A GGC G A 267 21555 2967 3455 9455 9466	43740578	267	21555	797	21555	1082	1082 T	222	Ь	CCT	Ь
267 21555 267 21555 8893 8893 T CCT P 21564 25384 25384 2313 2313 A 6TT V 29559 29674 5 5 6GC C C 267 21555 267 13483 945 9469	43740578	267	21555	267	13483	8893	8893 T	CCT	Ь	EJ.	_
21564 25384 2534 2313 2313 A 6TT V 29559 29674 29589 29674 5 7 6GC G 267 21555 267 13483 9455 9469 ————————————————————————————————————	43740578	267	21555	267	21555	8893	8893 T	ССТ	Ь	СТТ	٦
29559 29674 29559 29674 5 7 GGC G 267 21555 267 13483 9455 9456 ————————————————————————————————————	43740568	21564	25384	21564	25384	2313	2313 A	GTT	>	ATT	_
267 21555 267 13483 9455 9469	43740576	29559	29674	29559	29674	5	5 T	299	Ŋ	GGT	9
267 21555 267 21555 9455 9469	43740578	267	21555	267	13483	9455	9469	CATTICTATTGGTTCTTT	HFYWFF	CAT	
267 21555 267 21555 19250 19270	43740578	267	21555	267	21555	9455	9469	CATTTCTATTGGTTCTTT	HFYWFF	CAT	
	43740578	267	21555	267	21555	19250	19270	AGATTGTATCTCGATGCTTAT	AAC RLYLDAYN	AG	C

Figura 1: Tabella di output delle alterazioni

5 Listati di codice

Code Listing 1: Tabella per la traduzione in amminoacidi

```
aminoacids_lookup_table = {
 1
 2
      'F': ['TTT', 'TTC'],
      'L': ['TTA', 'TTG', 'CTT', 'CTA', 'CTC', 'CTG'],
'I': ['ATT', 'ATC', 'ATA'],
 3
 4
      'M': ['ATG'],
 6
      'V': ['GTT', 'GTA', 'GTC', 'GTG'],
      'S': ['TCT', 'TCA', 'TCC', 'TCG', 'AGT', 'AGC'],
 7
      'P': ['CCT', 'CCA', 'CCC', 'CCG'],
 8
      'T': ['ACT', 'ACA', 'ACC', 'ACG'],
 9
      'A': ['GCT', 'GCA', 'GCC', 'GCG'],
'Y': ['TAT', 'TAC'],
'H': ['CAT', 'CAC'],
10
11
12
      'Q': ['CAA', 'CAG'],
13
      'N': ['AAT', 'AAC'],
14
      'K': ['AAA', 'AAG'],
15
      'D': ['GAT', 'GAC'],
16
      'E': ['GAA', 'GAG'],
17
      'C': ['TGT', 'TGC'],
18
19
      'W': ['TGG'],
      'R': ['CGT', 'CGA', 'CGC', 'CGG', 'AGA', 'AGG'],
20
      'G': ['GGT', 'GGA', 'GGC', 'GGG'],
21
22
      'START': ['ATG'],
23
      'STOP': ['TAA', 'TAG', 'TGA']
24
```

Code Listing 2: Memorizzazione dei risultati nella struttura dati a lista

```
1
     for key, value in variations:
2
       for index, cds in affected_cdses.iterrows():
3
4
         variations_to_genes.append({
5
           'gene_id': gene_id,
6
           'gene_start': gene_start + 1, # 1-based position
7
           'gene_end': gene_end,
           'cds_start': cds_start + 1, # 1-based position
8
9
           'cds_end': cds_end,
10
           'original_codone': original_codone,
11
           'altered_codone': altered_codone,
12
           'relative_start': relative_start + 1, # 1-based position
13
           'relative_end': relative_end,
14
           'alteration': sequence,
15
           'original_aminoacid': original_aminoacid,
           'encoded_aminoacid': encoded_aminoacid
16
17
         })
```

6 Analisi dei risultati e conclusioni

Sulla base degli allineamenti prodotti ed analizzati in questa prima parte, è stato prodotto uni grafico riassuntivo per fornire un idea generale delle sequenze prese in esame.

6.1 Geni coinvolti

Variations per gene

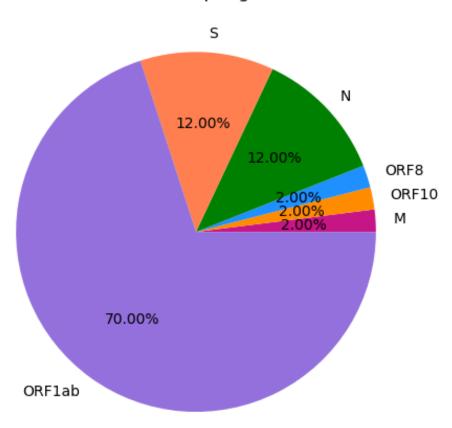


Figura 2: Tipologia di variazioni

Si può notare che quasi tutte le modifiche (70%) si concentrano sul primo gene della CDR: gene_name= ORF1ab; seguito dai geni gene_name= S e gene __name= N (12% ciascuno) e che gli altri siano quasi invariati. Un'analisi finale si trova nella terza e ultima parte del progetto.

7 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 e gestire l'output JSON del progetto1 e delle funzioni di supporto.
- Davide Marchetti si è occupato principalmente di generare i file di output e correggere le porzioni di codice relative alle letture delle reference.
- Entrambi hanno lavorato alla creazione ed elaborazione dei dati, alla matrice delle mutazioni e le traduzioni di quest'ultime.