Distanza mutazioni nei ceppi CoVid2 - 2019

Introduzione:

Le comparazioni eseguite ai fini della presente ricerca sono da intendersi effettuate con le seguenti premesse:

1 – Da un punto di vista matematico-informatico il bit in elettronica ha due stadi (acceso o spento), mentre il "bit" biologico del DNA (per fare un paragone di larghe intese) ha quattro stadi (le quattro basi azotate Adenina Citosina Guanina Timina).

2 – Da questo punto di partenza si svolgerà lo studio successivo di comparazione sequenze, applicando ad esse un modello matematico/statistico, coinvolgente frequenze e modalità di mutazione all'interno delle stesse.

3 – Il software di base usato è il MEGA-X per quanto riguarda la composizione grafica degli alberi filogenetici. Del sopracitato programma sono stati modificati alcuni algoritmi, eliminando i calcoli relativi alla sintesi proteica (ovvero la combinazione successiva delle basi azotate in quanto non inerente la ricerca primaria); sono stati altresì usati degli scripts programmati "ad hoc" per facilitare la formattazione dei file di testo con le basi in un formato leggibile dal software in oggetto.

4 – Le sequenze virali sono state acquisite dalla banca dati mondiale per la genetica:
https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/
SARS-CoV-2 (Severe acute respiratory syndrome coronavirus sequences).

Scopo della ricerca:

1 – Comparazione delle sequenze genetiche in ambito della stessa famiglia CoVid2.

2 – Comparazione dei ceppi individuati, volta all' individuazione di tipi e differenze in questi ultimi, al fine della costruzione di un albero filogenetico.

3 – Localizzazione di eventuali mutazioni in tale albero, raffrontato con localizzazione e datazione dei ceppi individuati.

Teorie e metodi di calcolo distanza nucleotidi:

Per comprendere come si svolga il calcolo per determinare le differenze fra due o più sequenze virali è necessario prima salire sul gradino immediatamente successivo a quello delle basi azotate, ovvero i nucleotidi:

I nucleotidi (derivati dai nucleosidi con l'aggiunta di uno o più gruppi fosfatici) sono in sostanza le unità che compongono gli acidi nucleici (DNA o RNA).

Ognuno di essi è composto da:

a – una base azotata

b – una molecola di zucchero (deossiribosio per essere specifici)

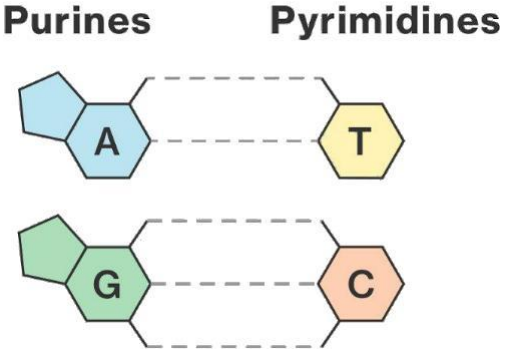
c – una molecola di acido fosforico (gruppo fosfatico)

I nucleotidi sono quanti sono le basi azotate da cui derivano, in dettaglio:

- Adenosina, derivata dalla Adenina
- Citidina, derivata dalla Citosina
- Guanosina, derivata dalla Guanina
- Timidina, derivata dalla Timina
- Uradina, derivata dall'Uracile

Nota: si presti attenzione al fatto che nelle sequenze genetiche da comparare l'uracile viene normalmente sostituito dalla timina, in quanto nei software si verificherebbe un errore di formato ed incompatibilità nel confrontare sequenze miste DNA - RNA) (*)

La cosa fondamentale da capire è che i nucleotidi possono legarsi perché è possibile formare un legame fra le loro basi azotate, ma non indistintamente: difatti Adenina lega con Timina e la Guanina lega con Citosina.
Le basi azotate si dividono poi, a seconda della configurazione molecolare, in puriniche (Adenina e Guanina) o pirimidiniche (Timina e Citosina).
Nell'immagine seguente un breve riepilogo dello schema AT-GC:



Altro concetto fondamentale da capire è la presenza/distanza dei cosiddetti Siti, ovvero delle regioni del DNA dove una citosina si trova vicino ad una guanina nella sequenza lineare di basi. Essi possono essere classificati in CpG o CG, ("CpG" è l'abbreviazione di " --C--phosphate--G-- ") dove i nucleosidi deossicitidina e deossiguanosina sono separati unicamente da un gruppo fosfato, che fa normalmente da ponte tra i nucleotidi nel DNA. La notazione "CpG" viene usata per distinguere questa sequenza lineare dall'appaiamento di basi complementari CG (citosina e guanina) su due diversi filamenti. (**)
È proprio sulle regole di legami sopracitate, proprie dei nucleotidi che si basa il calcolo per individuare eventuali mutazioni nel codice genetico, ovvero:
- Sostituzioni di nucleotidi compatibili con altri nella stessa sequenza genetica
- Sostituzioni di basi puriniche/ pirimidiniche con altre basi azotate compatibili
- Sostituzioni di Siti nelle Isole CpG ripetuti nella stessa sequenza

Per applicare questo ad un modello matematico i metodi sono ovviamente molteplici e dipendono da come vengano impostate le sostituzioni delle sequenze, tuttavia a "basso livello" (cioè non coinvolgente sintesi proteica) si usa frequentemente il metodo "Tajima D", che prende appunto il nome dal ricercatore Tajima Fumio: tale metodo si basa sui confronti del numero medio delle differenze di coppie di basi con i relativi Siti. Il relativo algoritmo è usato da molteplici programmi di comparazione ed evoluzione genetici.

Di seguito gli algoritmi usati:
1 – Calcolo della distanza per sostituzione di nucleotidi adiacenti:
Con p si indica la distanza di mutazione ovvero: la distanza (p) è la proporzione dei Siti dei nucleotidi nei quali le due sequenze comparate differiscono.
Questo si ottiene dividendo il numero della differenza dei nucleotidi (n_d) con il numero totale dei nucleotidi comparati (n), ovvero:
$p = n_d/n$
dove la variazione V rispetto a p è data da:
$V(p) = [p(1 - p)]/n$
In questo modo la sostituzione risulta consecutiva e lineare

Questo tipo di funzione è particolarmente efficace rispetto al numero di sostituzione di nucleotidi in un sito (*d*) solo quando la distanza è piccola ovvero quando la distanza *p* è approssimativamente uguale al numero delle sostituzioni nucleotidiche per Sito (*d*) e risulta ottimale quando quest'ultimo ha in linea di massima $p < 0.1$. Considerando tuttavia la semplicità di comparazione di questo algoritmo si possono ottenere risultati comunque apprezzabili ed affidabili anche per costruire alberi filogenetici più complessi dove la distanza è molto più lunga a patto che tutte le distanze delle coppie siano comunque piccole.
Di seguito il grafico del metodo Tajima:

A	-	β	γ	δ
T	α	-	γ	δ
C	α	β	-	δ
G	α	β	γ	-

Dove α, β, γ, δ sono i rapporti di sostituzione.

2 – Calcolo della distanza per Transizione e Transversione:
A quanto detto sopra può essere applicato in aggiunta il cosiddetto calcolo di transizione e transversione: La transizione è la sostituzione di una Purina con un'altra Purina e la transversione è la sostituzione di una Pirimidina con un'altra Pirimidina.
Si può dunque calcolare la proporzione delle differenze di transizione (*P*) e di transversione (*Q*) con le seguenti equazioni:
$P = n_s/n$
$Q = n_v/n$
dove rispettivamente n_s e n_v sono i numeri delle differenze transizionali e transversionali tra le due sequenze, con $n_s + n_v = n_d$. Le varianti di *P* e *Q* vengono calcolate con la stessa funzione usata per la variazione V della distanza detta prima. In aggiunta il rapporto delle differenze di transizione e transversione (R_d) è calcolato da:
$R_d = P/Q$
E la sua variazione è calcolata da
$V(R_d) = [c_1^2 P + c_2^2 Q - (c_1 P + c_2 Q)^2]/n$
Dove
$c_1 = 1/Q$
$c_2 = -P/Q^2$
Questo metodo, integrato al precedente, riesce a dare un prospetto più accurato di albero filogenetico.

3 – Calcolo della distanza Tajima-Nei
Poiché nella realtà può comunque spesso accadere un ulteriore distanziamento di frequenza di mutazione dei nucleotidi o comunque quando si vogliano calcolare dati che implicano mutazioni più distanti dalla semplice base vicina, vengono in aiuto le seguenti equazioni, che il dott. Fumio Tajima perfezionò con il dott. Masatoshi Nei. In questo metodo la distanza non viene più calcolata linearmente ma logaritmicamente ed è espressa da:

$d = -b \log_e(1 - p/b)$
e
$V(d) = p(1 - p)/[(1 - p/b)^2 n]$ dove

$$b = \frac{1}{2}(1 - \sum_{i=1}^{4} g_i^2 + p^2 / c)$$

$$c = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{x_{ij}}{2 g_i g_j}$$

In quest'ultima g_i e g_j rappresentano la frequenza con cui sono presenti gli ennesimi nucleotidi "i" e "j"
Questo metodo si mostra particolarmente efficace quando ci si trova di fronte a delle sequenze oggettivamente mutate che lasciano intendere un cambiamento profondo nel ceppo genetico. (***)

Risultati preliminari:

Come ceppo principale, usato come punto di partenza e riferimento per le sequenze successive confrontate, è stato scelto il ceppo originale di Wuhan. Quest'ultimo, oltre ad essere risultato il più diffuso sulla base delle sequenze analizzate a livello planetario (in quanto ogni continente ha almeno un luogo dove sia presente tale ceppo), risulta anche coinvolgere in particolare tutta la zona sud asiatica.
Di seguito il grafico del ceppo sudasiatico che coinvolge per similitudine lineare altri paesi:

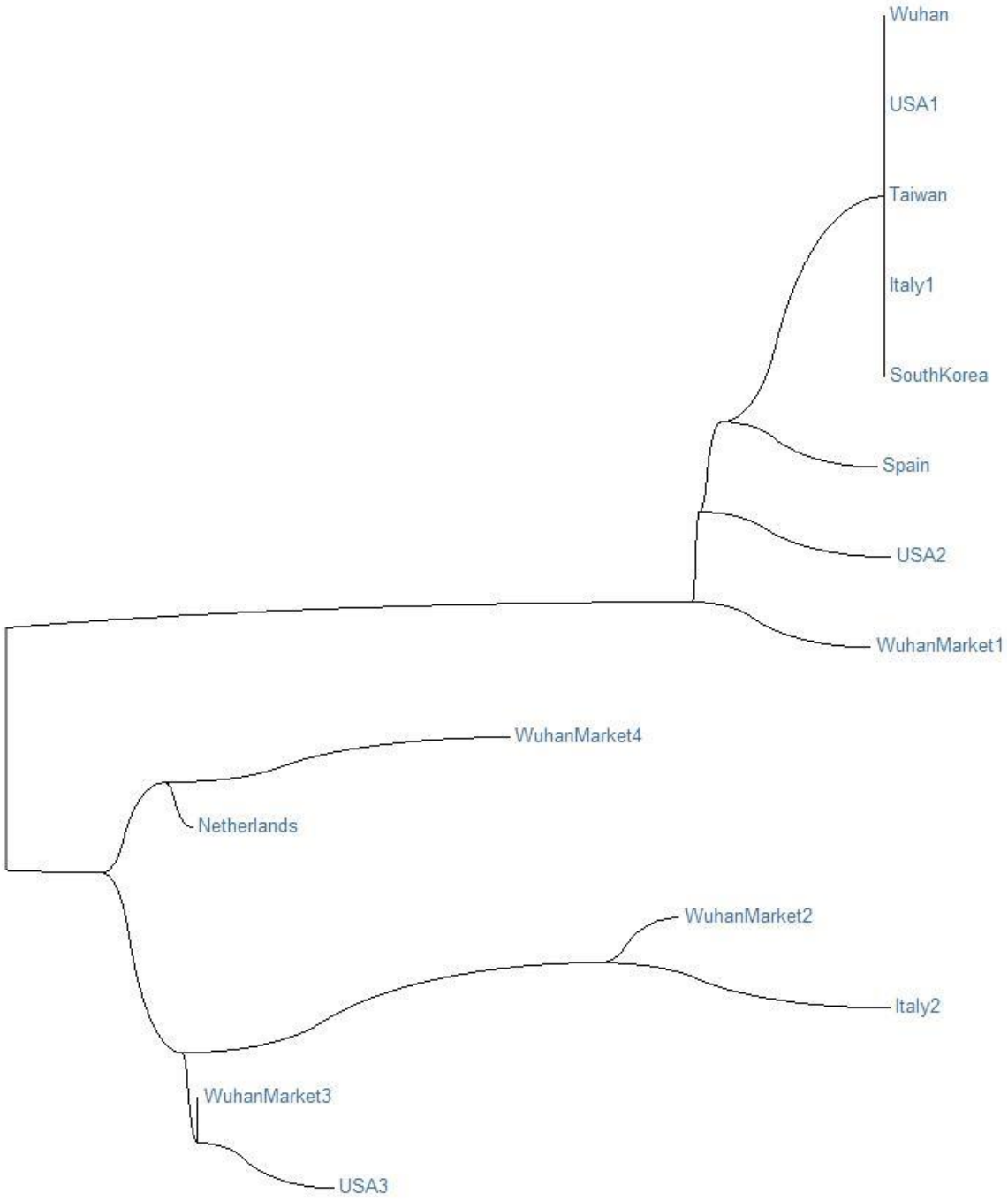
I primi risultati sempre basati su comparazione lineare indicano un virus con svariati ceppi (grafico circolare):

Già da questo primo grafico risulta che il ceppo dei Netherlands è il più lontano seguito da uno dei ceppi del cosiddetto “Mercato del pesce di Wuhan”, identificato con “Wuhan Mrkt”; molto più vicino alla sequenza originale c'è lo spagnolo, mentre le sequenze Wuhan, Taiwan, Korea del Sud praticamente identiche (insieme ad una delle tante che è arrivata negli USA, che verranno analizzate in seguito)

Anomalia dei ceppi del “Mercato di Wuhan”

Premettendo che a detta del governo cinese il campione del mercato del pesce di Wuhan dovrebbe essere il “protovirus” ovvero quello da cui poi è scoppiata la vera e propria pandemia e sempre premettendo che tutte le sequenze orientali esaminate sono più o meno uguali (anche quelle di Taiwan e Sud Korea, isolate da ricercatori indipendenti) e che tutti i dati finora esaminati a questo punto sono stati analizzati solo a livello comparativo - lineare, l’anomalia è la seguente: Il campione del 2019 del mercato cinese dove il governo di Pechino asserisce essersi diffuso questo virus ha una sequenza genetica che non corrisponde al ceppo sudasiatico.

Da prendere in considerazione il fatto che esistano solamente quattro sequenze del "Mercato del pesce" tutte inserite nella banca dati genetica (sequenze LR767995 – 96 – 97 – 98), e tutte pubblicate il 6 marzo 2020.

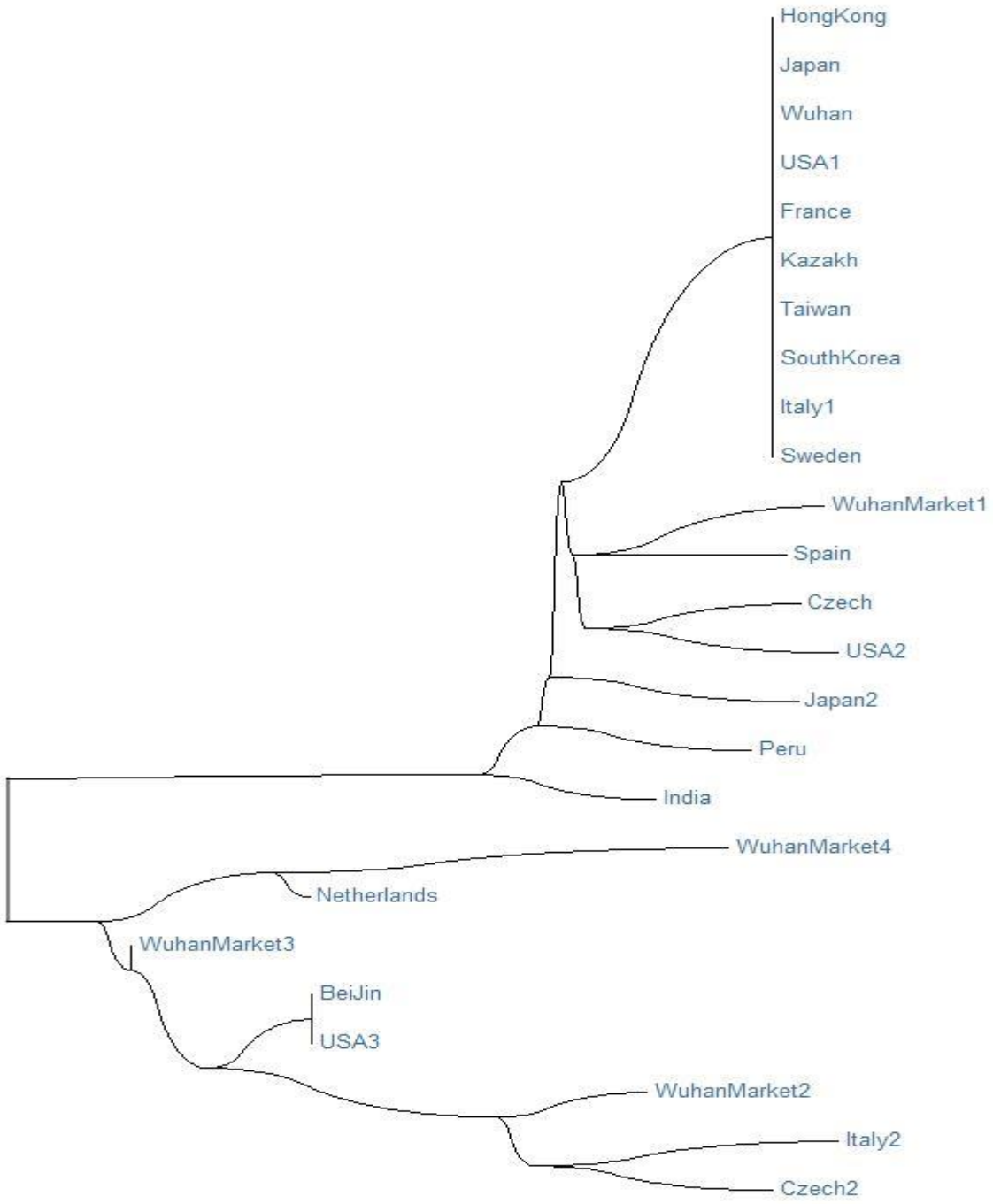


Da questo grafico emerge come i quattro ceppi del "Mercato del Pesce" (Wuhan Market da 1 a 4) non siano collegati direttamente con il ceppo principale sudasiatico.
Poiché la data di pubblicazione è identica e copre un periodo di un anno è difficile dedurre con i dati attuali se tali ceppi fossero mutati o pre-esistenti.

Ceppi Statunitensi:

Gli USA sembrano, sulle comparazioni attuali (effettuate con algoritmo di calcolo per similitudine di basi azotate), il paese più colpito dalla pandemia non solo a livello epidemico ma anche a livello di diversità di genere: difatti ogni ceppo analizzato presente a livello planetario è stato isolato anche negli Stati Uniti con caratteristiche uguali o comunque di similitudine filogenetica all'interno della stessa famiglia.

Di seguito il grafico dove si evince come Gli USA abbiano almeno un caso isolato in ogni ceppo a livello mondiale, inclusa la sequenza primaria sudasiatica ed un altro ceppo con caratteristiche familiari simili:



Ceppo Brasiliano e stabilità nella sequenza genetica:

A data di Maggio 2020, sulla base di 32 campioni differenti analizzati, si confermano i dati precedenti ed è stato riscontrato che il ceppo del Brasile più diffuso è stabilmente uguale al ceppo comune primario sudasiatico. Questo potrebbe comprovare la teoria che il virus si mantenga stabile all'interno di ogni ceppo, non presentando mutazioni significative ed avendo alla fine del genoma una solida sequenza di stop della codifica ben visibile e marcata, rimanendo su un numero oscillante fra 29890-29903 basi azotate inalterate, dalle quali derivano 29860 Siti.

Da tenere in considerazione che la percentuale di similitudine di tale ceppo applicando solamente il modello comparativo delle basi azotate. Di seguito il risultato della comparazione diretta (nell'immagine viene mostrata solamente la parte iniziale del codice, mentre nel calcolo è inclusa l'intera sequenza).

Di seguito i risultati ottenuti applicando alla sequenze virali di Wuhan, Italia e Brasile i metodi di comparazione Tajima-Nei, ovvero sostituzioni di nucleotidi e similitudine di siti adiacenti:

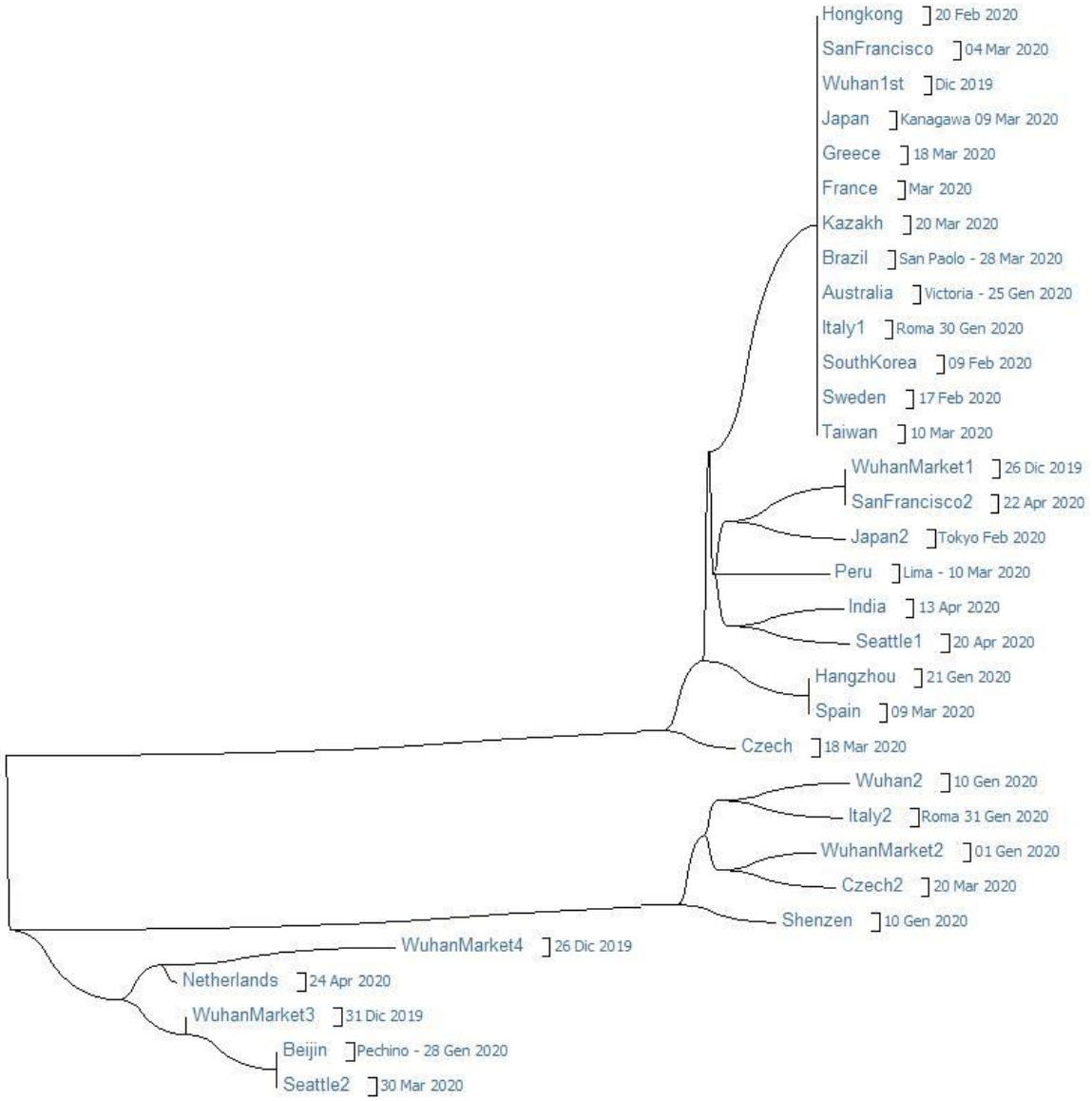
Table. Results from the Tajima's test for 3 Sequences

Configuration	Count
Identical sites in all three sequences	29860
Divergent sites in all three sequences	0
Unique differences in Sequence A	1
Unique differences in Sequence B	1
Unique differences in Sequence C	4

Tale risultato evidenzia che su 29866 siti le divergenze sono pari a 0, i siti identici sono 29860 con differenze univoche tra la sequenza di Wuhan e quella dell'Italia pari a 1, con differenze fra le sopracitate sequenze e quella del Brasile pari a 4.
Non risulta in questo ceppo alcuna mutazione significativa all'interno del suo genoma.

Albero filogenetico a 32 campioni:

Ciò che si è evidenziato dalla comparazione a 32 campioni è la progressiva attenuazione della sequenza poli(A)polimerasi finale, presente nella parte terminale del genoma RNA (codifica ACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA nella sequenza WuHan-1st – Dic 2019), di fatto abbreviata (codifica ACAAAAAAAAAAA nella sequenza France – Mar 2020) o non presente in alcuni ceppi cronologicamente successivi (codifica ATTTTAGTAGTGCTATCCCCATGT nella sequenza India – Apr 2020)
Di seguito il grafico che ricostruisce l'albero filogenetico a 32 campioni con le date di quando sono state isolate le sequenze virali. Tale grafico è basato sul modello Tajima-Nei:



Conclusioni:

Sulla base dei campioni analizzati, tenendo in considerazione anche i metodi di calcolo utilizzati non univoci, si presenta un modello virale comprendente più ceppi, all'interno dei quali sono presenti minime variazioni. Non si riscontrano fino ad ora significative mutazioni nel codice virale.

Note:

- (*) La banca dati mondiale per la genetica usa la Timina anche per sequenze RNA, ovvero sostituendola all'Uracile, per la standardizzazione con i programmi compatibili con la formattazione ACGT e che non riconoscerebbero quella ACGU.
- (**) I modelli di comparazione Tajima-Nei utilizzati in questo studio sono fondamentalmente intercambiabili ed applicabili sia per DNA che RNA, in quanto analizzano la sequenza, le basi ed i siti nella mappatura progressiva ad un livello più basso e basilare, che prescinde dal singolo o doppio filamento.
- (***) Si parla in questo caso di distanza logaritmica. E' doveroso specificare che nell'albero filogenetico generato la distanza fisica misurabile fra un ceppo ed un altro deve essere sempre riferita al tipo di algoritmo usato ed obbligatoriamente descritto (lineare, logaritmico, temporale, ecc.), per avere appunto una misura di riferimento

Ringraziamenti:

1 – Il mio Ryzen Octacore, rivelatosi fondamentale nella tempistica dei calcoli.
2 – Il software MEGA X, un grande strumento alla portata di tutti.
3 – Il dott. Fumio Tajima per gli eccellenti modelli evolutivi elaborati.
4 – Le banche dati, i ricercatori e gli ingegneri che progettano le macchine per estrapolare dati biologici in modelli matematici.
5 – L'Istituto Nazionale Malattie Infettive Lazzaro Spallanzani per la sequenza Italiana del codice virale isolato e pubblicato tempestivamente.

Bibliografia:

- Lehninger A.L., Nelson D.L., Cox M.M – Principi di biochimica
- Multiplication and permutation, Axioms independence – University of Illinois (Urbana-Champaign)
- Evolutionary distance between Nucleotide sequences – Tajima F., Nei M. - University of Texas (Huston)
- MEGA X: Molecular evolutionary genetics analysis across Computing platforms
- Kumar S., Tamura K., Jakobsen I.B., Nei M. - MEGA2: Molecular evolutionary genetics analysis software - Bioinformatics

Virologicamente vostro… Mike Yoshi