

SARS-CoV-2: filogenesi e filogeografia

Gianguglielmo Zehender
Università di Milano





CONVEGNO INTERNAZIONALE
GIORNATE
INFETTIVOLOGICHE
“LUIGI SACCO” 2020

Covid-19:
la sfida di una pandemia



 UNIVERSITÀ DEGLI STUDI
DI MILANO

 Regione
Lombardia

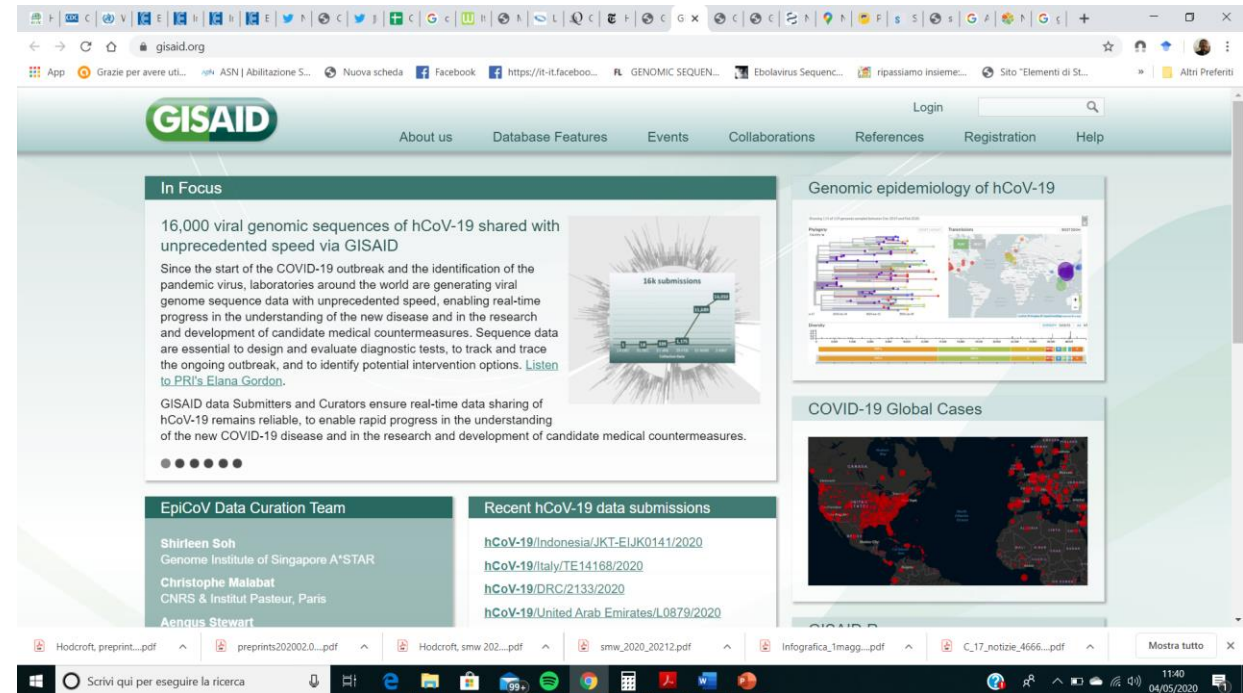
ASST Fatebenefratelli Sacco

**VIRTUAL
EDITION**

4-5 Febbraio 2021

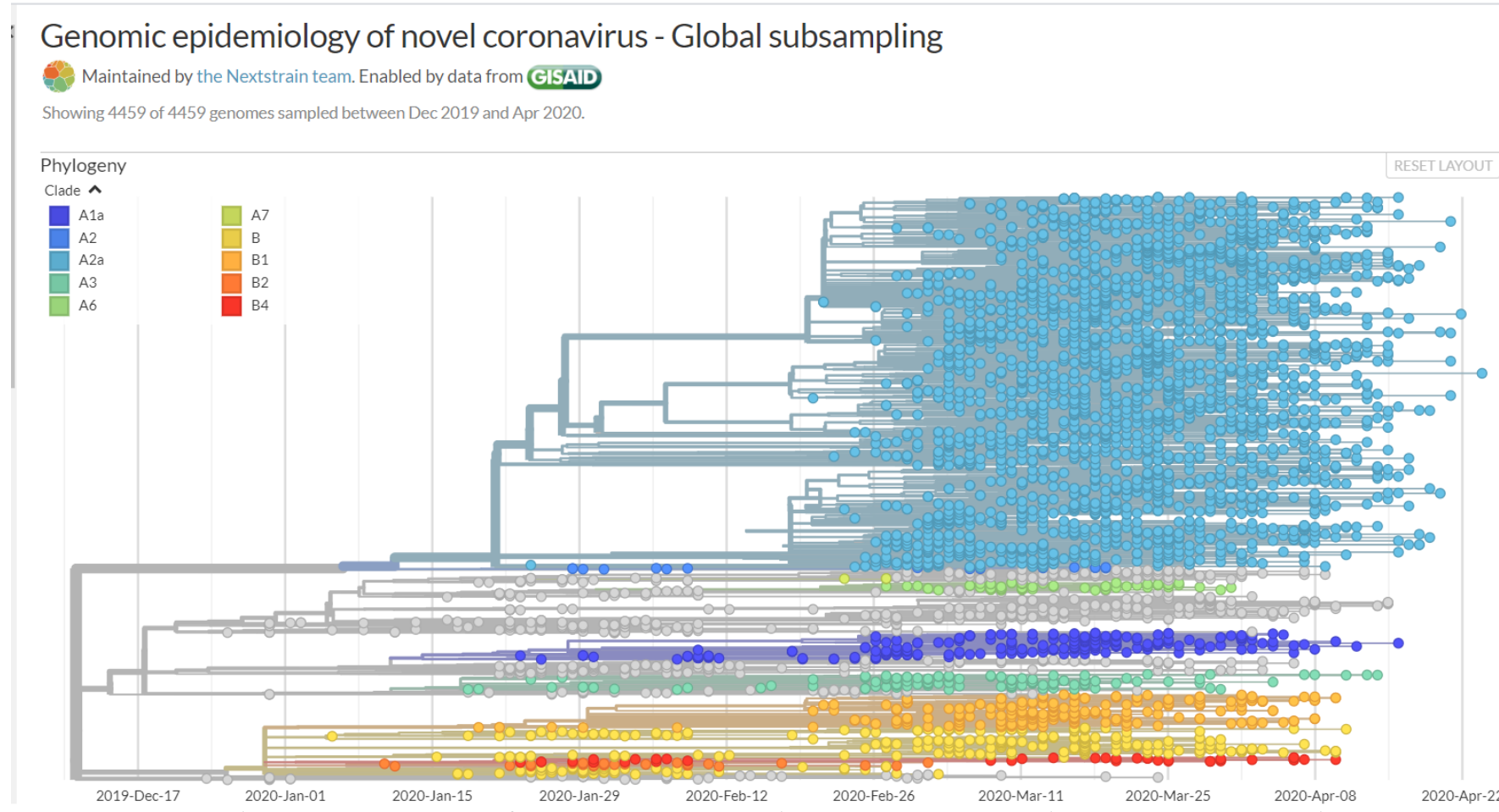
SARS-CoV-2 Filogenesi

- Virus ad RNA: alta variabilità genomica
- Stimata intorno a 0.8 sostituzioni ogni 1000 siti all'anno
- Per un genoma di quasi 30000 nucleotidi significa una frequenza di 24 mutazioni all'anno (2/mese)
- Sono più di 400,000 i genomi completi depositati sulla piattaforma **Global Initiative on Sharing All Influenza Data (GISAID)**
- dal 10/01/2020 (prima sequenza di Wuhan depositata) ad oggi



Nextstrain

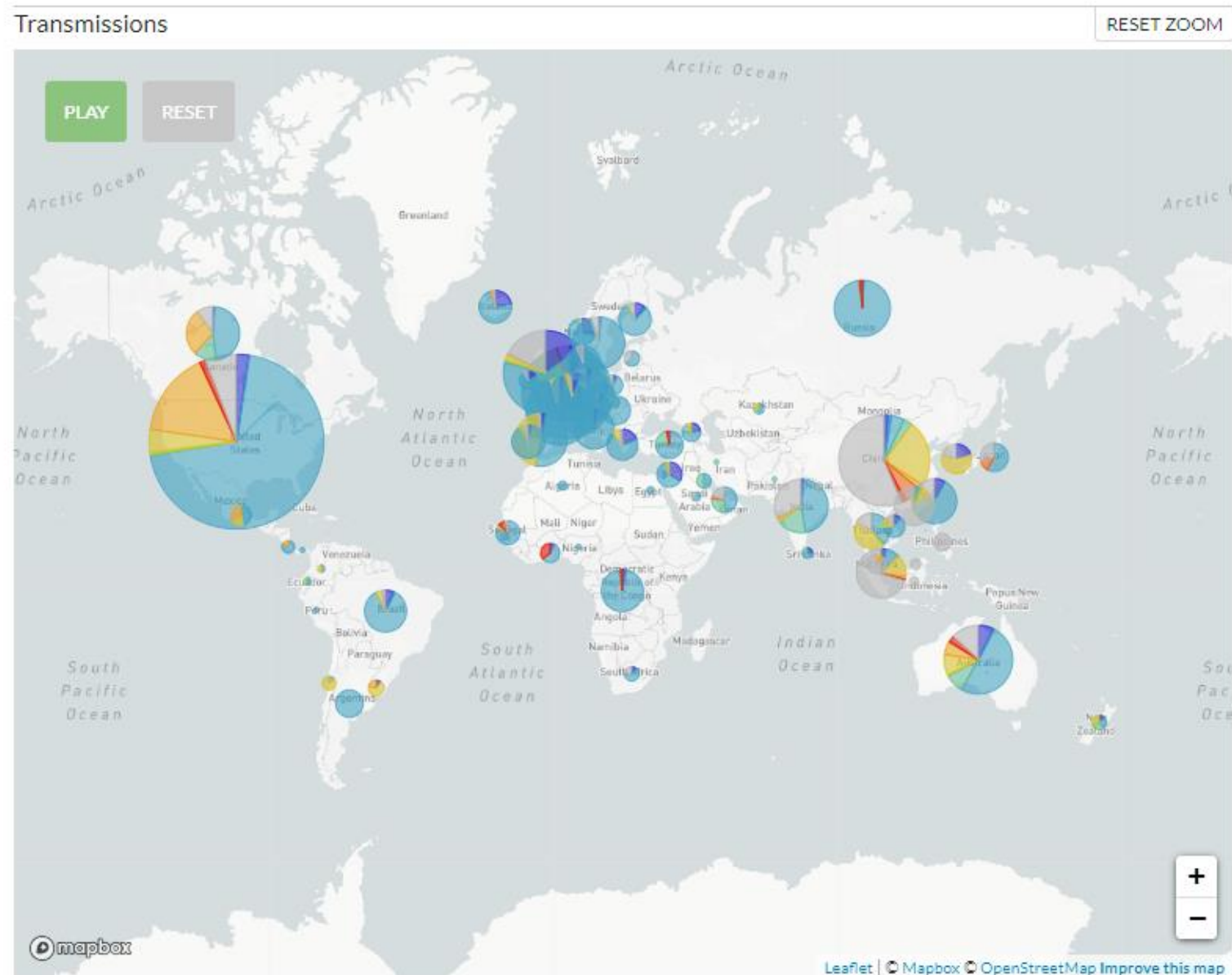
- Nextstrain è un progetto open source per sfruttare la disponibilità di genomi dei patogeni a beneficio della ricerca scientifica e della sanità pubblica
- Attinge i dati genomici da database pubblici, tra cui anche GISAID
- Fornisce una vista costantemente aggiornata dei dati disponibili al pubblico insieme a potenti strumenti di analisi e visualizzazione.
- L'obiettivo è aiutare la comprensione epidemiologica/evolutiva e contribuire alla risposta contro le epidemie.



Individuati 2 clades principali (A e B) ulteriormente suddivisi in subclades (esempio: A2-A2a, A1a, ecc)

Distribuzione geografica clades

- Mostra una mappa che riporta:
 - La frequenza dei diversi ceppi virali nelle diverse aree in tempi diversi
 - Gli spostamenti da un'area all'altra



I primi studi

SHORT COMMUNICATION

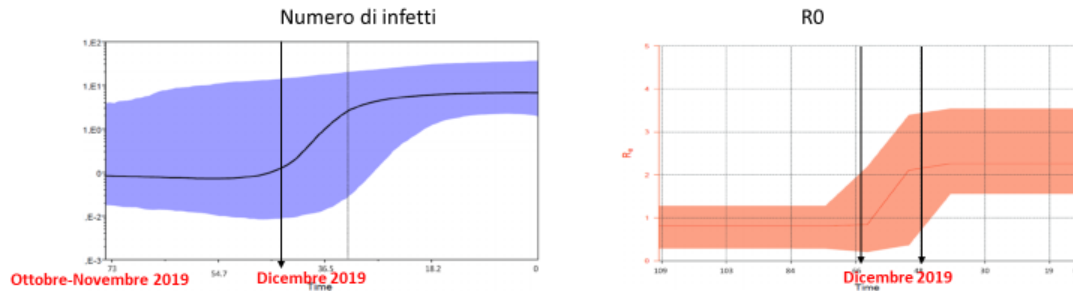
WILEY

Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2

Alessia Lai^{1,2} | Annalisa Bergna¹ | Carla Acciarri¹ | Massimo Galli^{1,2} |
Gianguglielmo Zehender^{1,2,3}

52 SARS-CoV-2 genomi
disponibili al 4 Febbraio
2020

- Più probabile origine dell'epidemia Novembre 2019 (tra Settembre e Dicembre 2019)
- Stima di R_0 2.6 (range, 2.1-5.1),
- Osservato un incremento sia del numero di infezioni che del valore R (da 0.8 a 2.4) nel mese di dicembre 2019.
- Tempo di raddoppiamento dell'epidemia tra 3.6 e 4.1 giorni
- Aumento della trasmissibilità interumana di SARS-CoV-2 nel mese di Dicembre 2019



SHORT COMMUNICATION

WILEY

Genomic characterization and phylogenetic analysis of SARS-CoV-2 in Italy

Gianguglielmo Zehender^{1,2,3} | Alessia Lai^{1,2} | Annalisa Bergna¹ | Luca Meroni⁴ |
Agostino Riva⁵ | Claudia Balotta¹ | Maciej Tarkowski¹ | Arianna Gabrielli¹ |
Dario Bernacchia⁶ | Stefano Rusconi^{1,4} | Giuliano Rizzardini⁷ | Spinello Antinori^{1,4} |
Massimo Galli^{1,2,4}

Isolation, molecular characterization, and phylogenetic analysis of the first **three complete genomes** of SARS-CoV-2 isolated from three patients involved in the first outbreak of COVID-19 in Lombardy, Italy.

Albero filogenetico datato di 161 genomi di SARS-CoV-2

Le sequenze dei primi tre pazienti di Codogno si raggruppano in un unico clade (A) insieme a

- Altre 2 sequenze Italiane già depositate
- Sequenze del sud-Centro America
- Una sequenza del primo paziente del cluster Bavarese (Germania) si mette all'outgroup del clade
- Sulla base della datazione del clade probabile ingresso in Italia alla fine di gennaio-inizi febbraio
- Identificazione della mutazione nel gene S della mutazione D614G

25/01/2020 (18/01-30/01)



31/01/2020 (23/01-07/02)

- Il virus era probabilmente presente in Italia dagli inizi di Febbraio
- Probabile legame con il cluster Bavarese

SARS-CoV-2: classificazione-clades

- Al momento sono stati descritti 11 clades
- 19A e 19B gli originari di Wuhan
- 20A emerso in Europa in Marzo da 19A
- 20B e 20C emersi successivamente nel 2020 dal 20A
- 20D-20I emersi dopo l'estate 2020

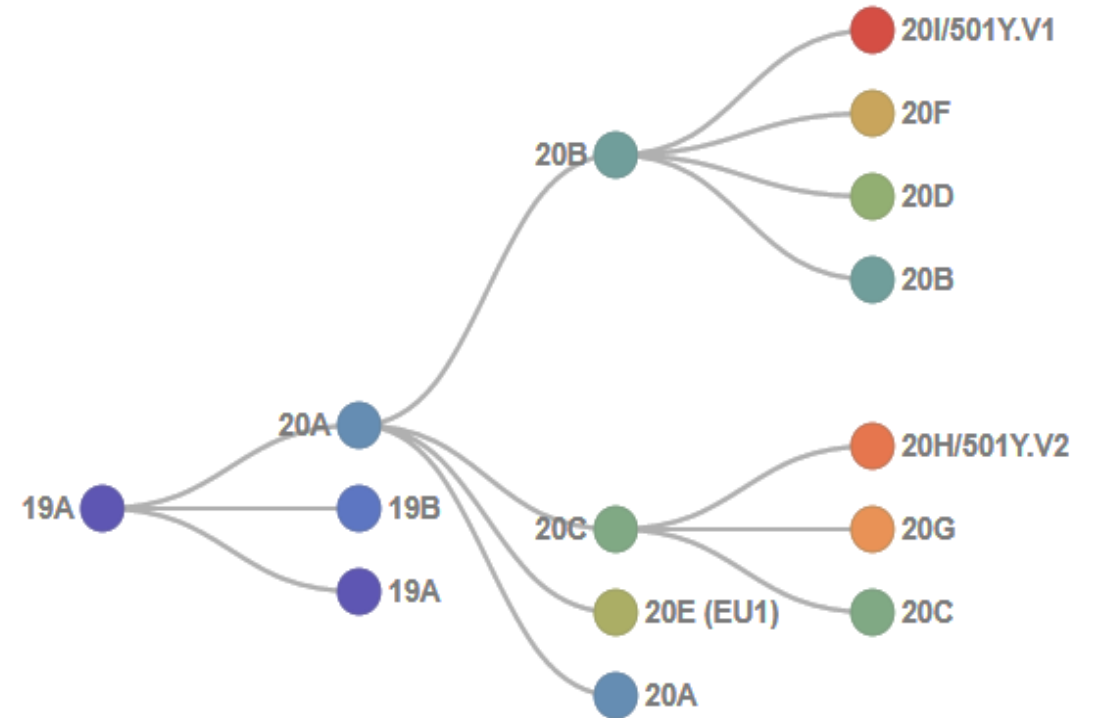
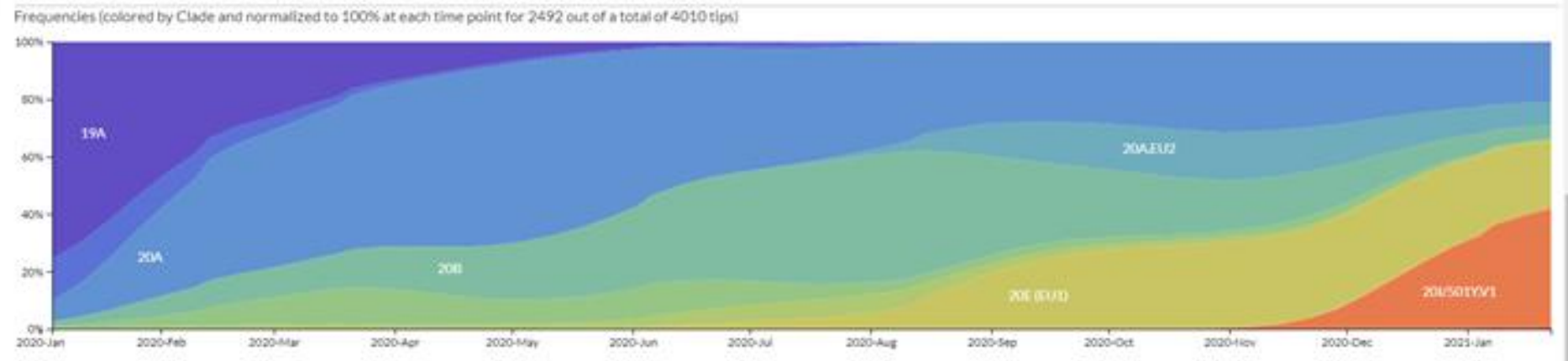
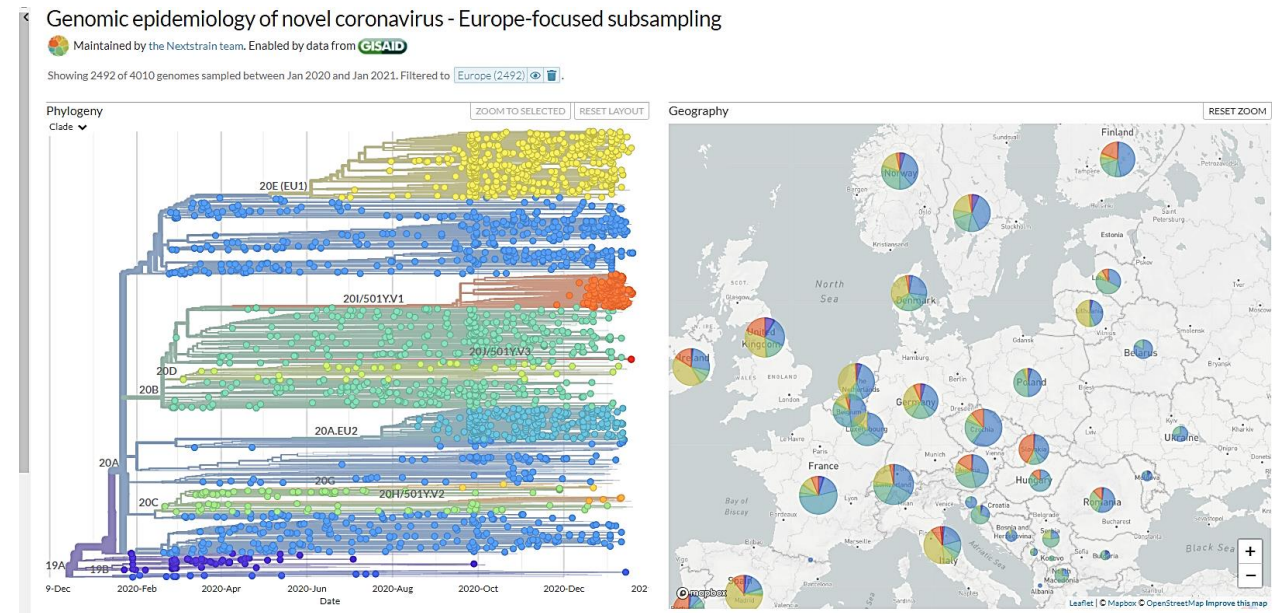


Fig.1. Illustration of phylogenetic relationship of clades, as defined by Nextstrain

Clades, la situazione in Europa ad oggi (feb 2021)

- In Europa sono virtualmente presenti tutti i clade (A e B)
- Quello numericamente più rappresentato è il clade 20A



SARS-CoV-2 lineages (Pangolin)

- Alla radice dell'albero 2 lineage denominati A (EPI_ISL_406801-il più vicino al pipistrello) e B (Wuhan-Hu-1 AN MN908947)
- Lineages discendenti dai primi numerati (A.1, B.1, B.2, ecc)
 - Almeno 1 nucleotide di differenza
- Questi poi possono essere progenitori di altri lineage che emergono in altre popolazioni (B.1.1, A.1.1....)
- Questo fino al 3° livello, quindi cambia lettera (es. A.1.1.1 corrisponde a C.1)
- Classificazione dinamica: la classificazione viene rivista ad intervalli regolari e i lineage etichettati come nuovi o non più evidenziati

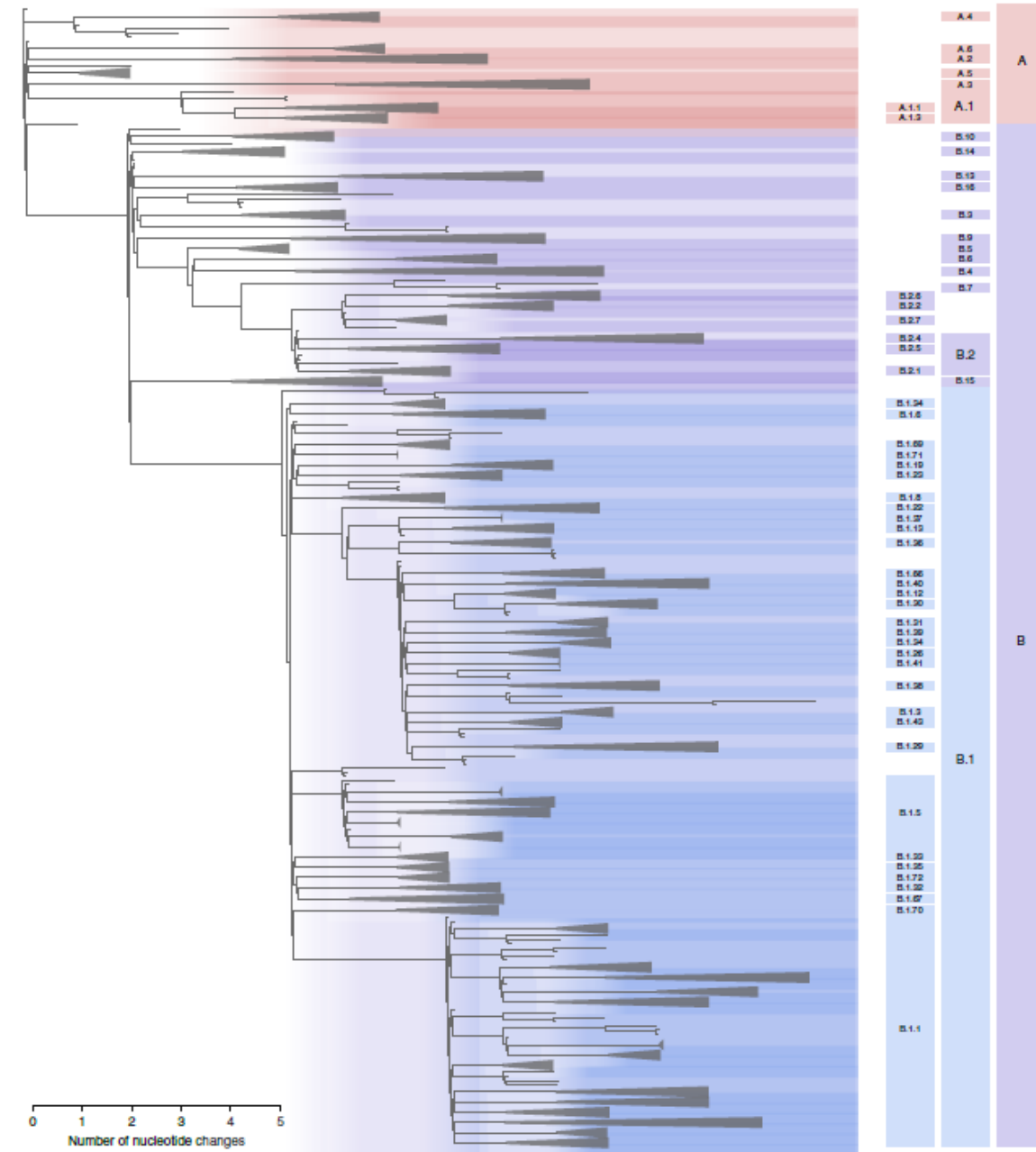


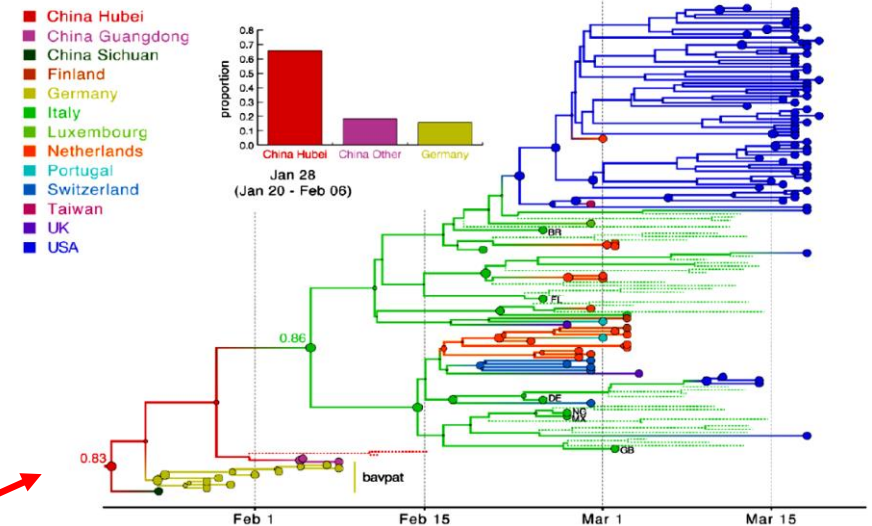
Fig. 1 | Maximum likelihood phylogeny of globally sampled sequences of SARS-CoV-2 downloaded from the GISAID database on 18 May 2020. Five

SARS-CoV-2 problemi aperti

- SARS-CoV-2 ha un grande genoma (29 kb) ed evolve rapidamente, ma meno degli altri RNA virus
- Evoluzione stimata intorno a 1×10^{-3} s/s/y
 - Che significa circa 2 sostituzioni per genoma per mese
- A Marzo 2020 l'intera popolazione di genomi differiva per un massimo di 12 mutazioni
- Cluster di trasmissione sono definiti già da 1-3 differenze nucleotidiche
- Carenza di genomi rispetto alla diffusione e soprattutto in alcune aree geografiche
- **Il tasso di trasmissione di SARS-CoV-2 è più elevato del suo tasso evolutivo**, quindi molti genomi identici si diffondono rapidamente e possono essere presenti ovunque!

Worobey, Science 2020

Tra la fine di gennaio e febbraio
SARS-CoV-2 è già presente in tutto il mondo

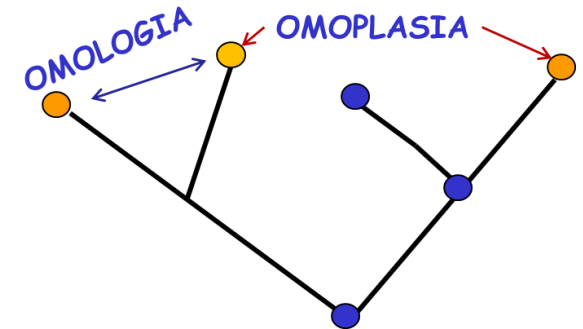


SARS-CoV-2 problemi aperti

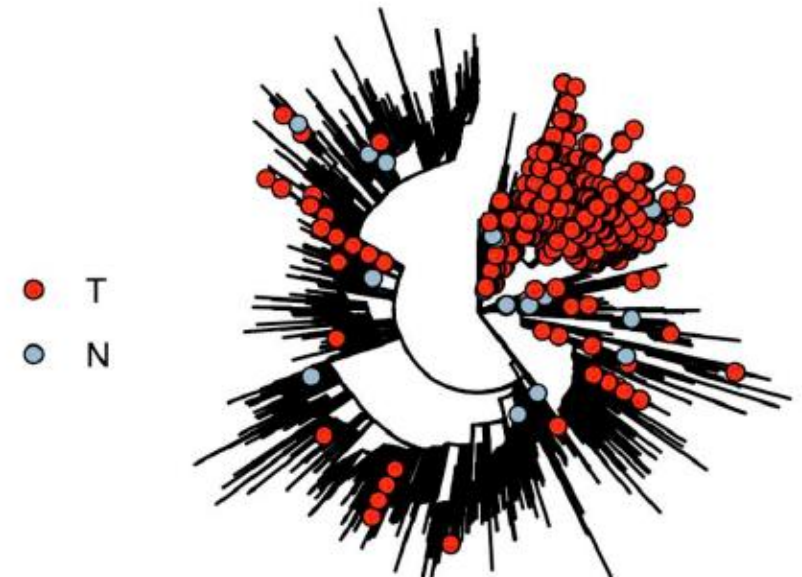
- Frequente omoplasia nel genoma
- Mutazioni ricorrenti in più lineage
- Individuati finora tra i 100 e i 200 siti soggetti a mutazioni ricorrenti
- Frequente sostituzione C->T (APOBEC?)
- O. può dipendere da 4 cause
 - Ricombinazione (rara)
 - Selezione positiva (come nel caso della G11083T, Orf1a, epitopo della risposta T CD4 e CD8)
 - Ipermutazione in alcuni geni
 - Dalla piattaforma utilizzata (dal lab)- artefatti (mutazioni Illumina o nanopore, Lab, country specifiche)-errori sistematici

Turakhia, Plos genetics, 2020

Van Dorp, IGE, 2020



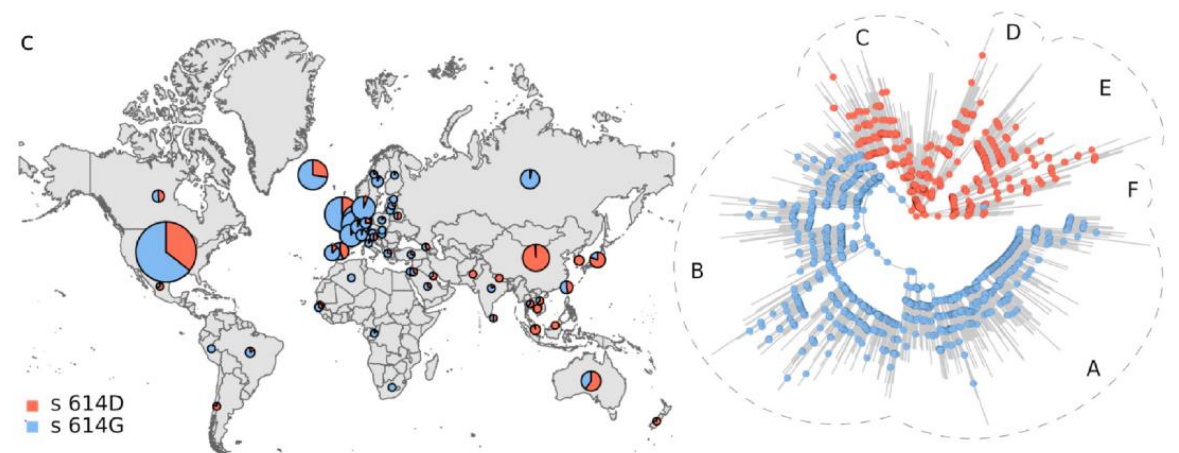
b)



Omoplasia esemplare in Orf1a (sito 11083)-van Dorp

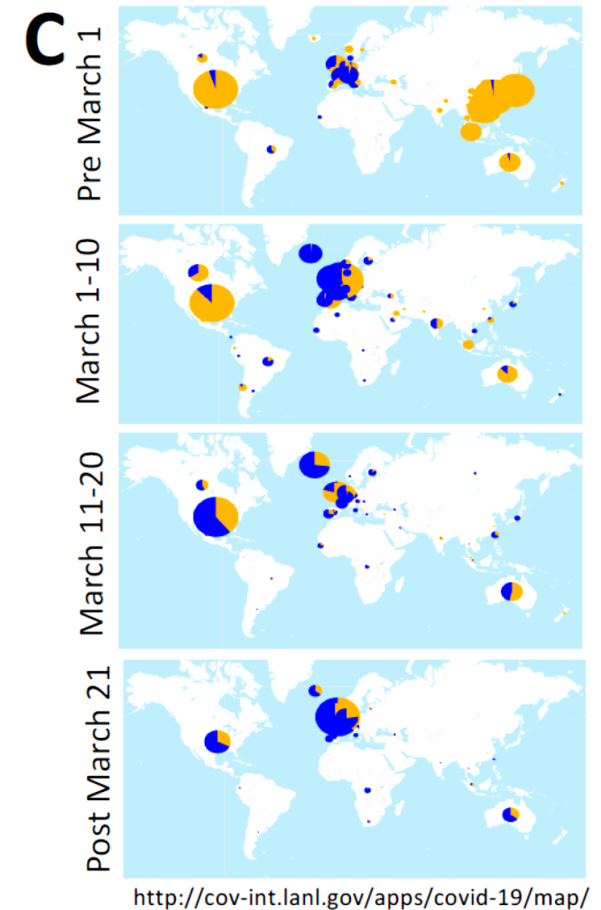
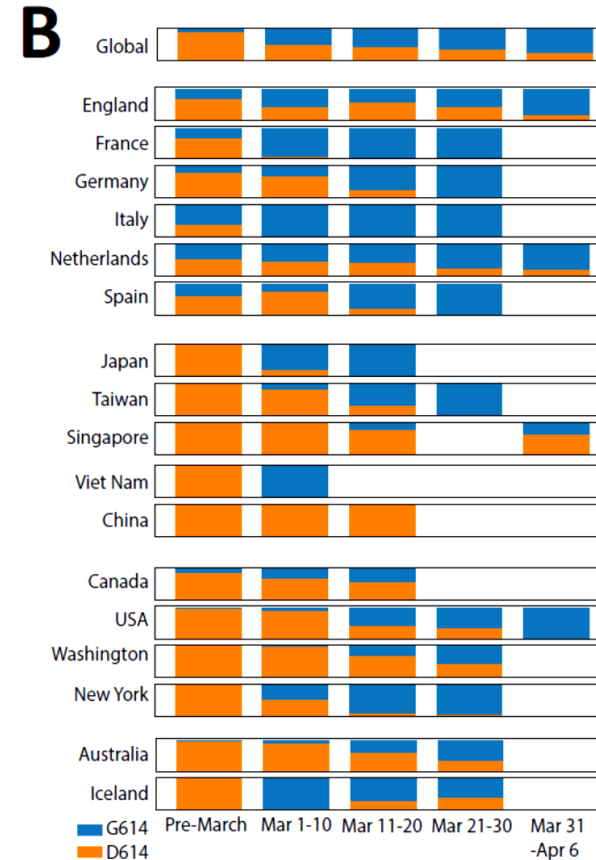
SARS-CoV-2 problemi aperti

- **Effetto fondatore** prevalente
 - La principale via di dispersione geografica del virus al mondo: i viaggi che uniscono regioni molto lontane tra loro
 - In questo caso solo una piccola frazione della diversità del virus viene (casualmente) trasportata in una nuova popolazione
 - Quando l'infezione si espande nella nuova popolazione si ha:
- Perdita di diversità progressiva (genetic drift)
 - Affermazione di mutanti nella popolazione originaria minoritari
 - Alcuni di questi mutanti possono portare a nuovi fenotipi virali.



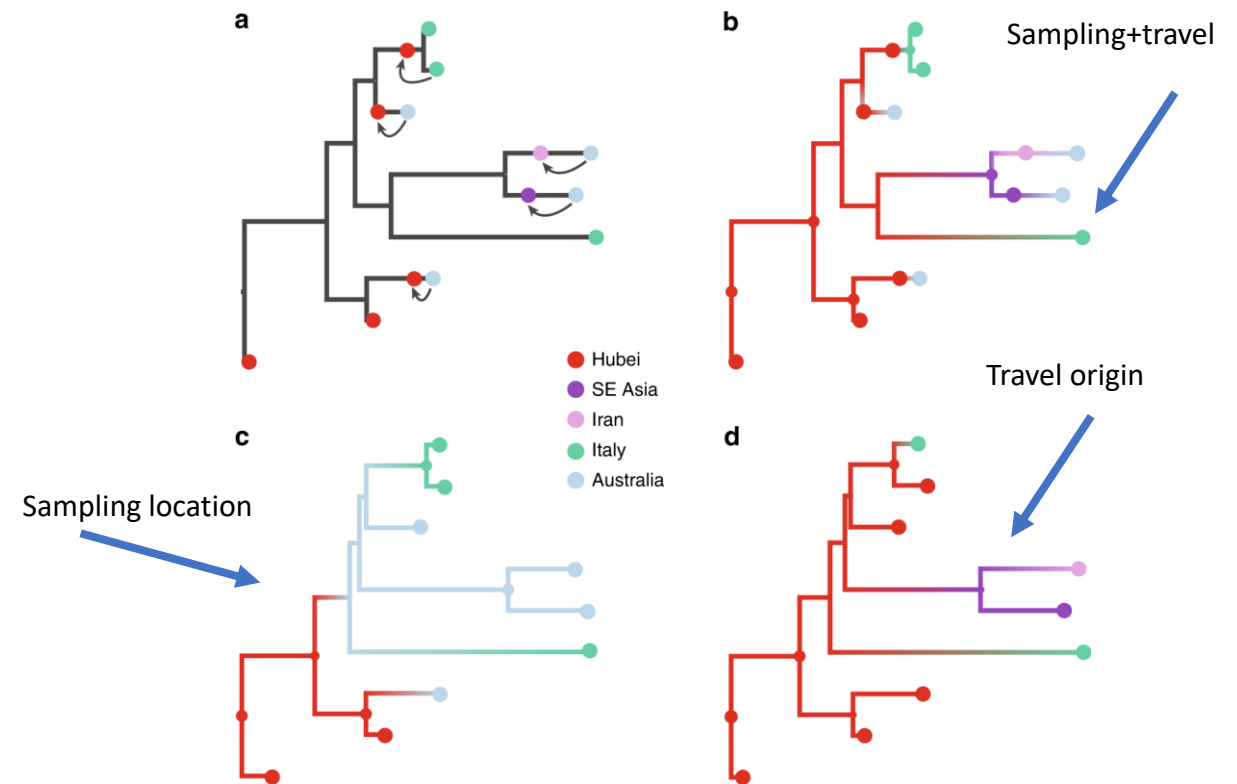
Il caso della variante S D614G

- Identificata in Cina, per la prima volta nella seconda metà di Gennaio 2020
- Infetta una donna-manager di una multinazionale che si reca in Germania per una serie di seminari qualche giorno prima di sviluppare i sintomi
- Trasmette l'infezione a un dipendente della stessa Azienda in Germania causando un cluster di 14 soggetti
- Prima di marzo rappresenta il 10% di tutti i genomi
- Alla fine di marzo rappresenta già il 67% di tutti i genomi globali
- Tra aprile e maggio il 78%...



SARS-CoV-2 diversità geografica e viaggi

- La distanza tra i genomi non è in relazione solo con la distanza geografica, quanto con i collegamenti aerei tra regioni anche distanti
- Necessario incorporare nell'analisi anche i dati relativi ai viaggi (origine e destinazione)

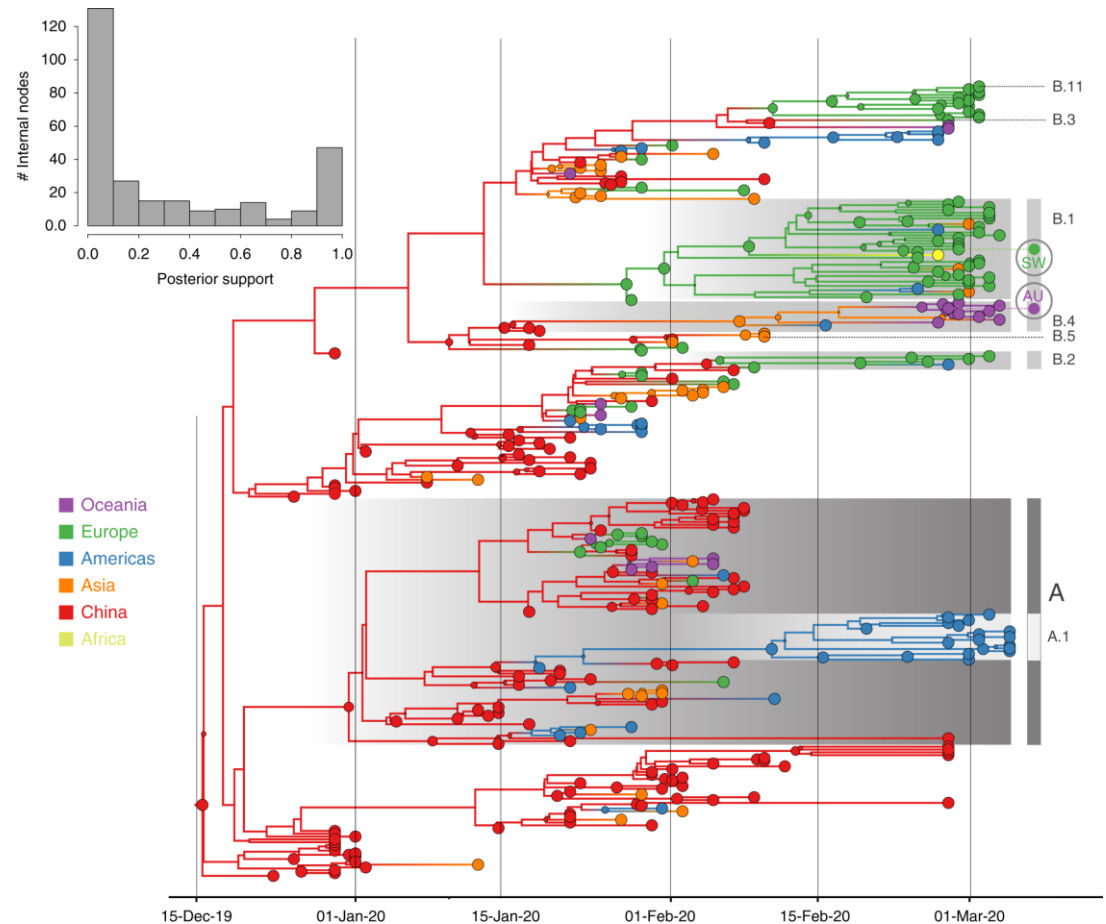


Il colore dei rami e dei nodi interni cambia (e quindi la storia filogeografica)
A seconda che si usi la sampling location, la travel origin o tutt'e due

Lemey, nat com 2020

Filogeografia di SARS-CoV-2 problemi

- Elevato grado di incertezza filogenetica
- Dovuta a scarso supporto dei nodi interni
- Un singolo albero filogenetico è quindi inadeguato a descrivere la storia filogeografica

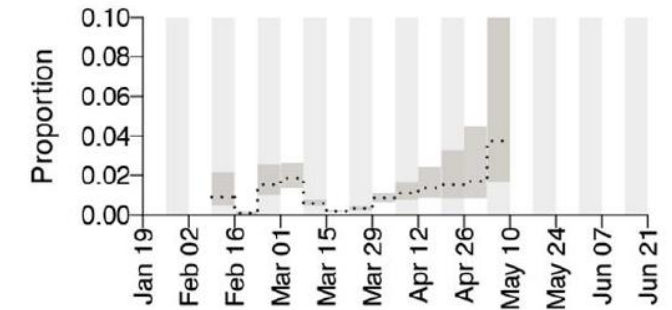


Lemey, nat com 2020

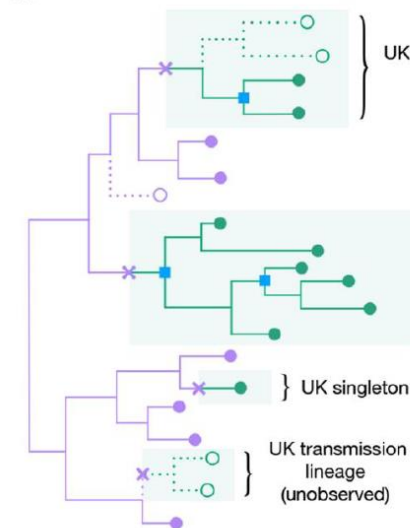
Filogeografia di SARS-CoV-2 in UK

- Studio su più di 26,000 genomi UK tra feb-mag 2020
 - circa l'1% dei casi confermati nel periodo
- Individuati più di 1000 lineage di trasmissione UK contenenti anche sequenze estere
- 8 i più grandi
- Sono i più precoci (tra 27 febbraio e 31 marzo)
- Sono anche i più persistenti (difficili da eliminare)
- Provenienza i più precoci da Cina (rari)
- Successivamente shift verso ceppi Europei
 - Spagna 33%
 - Francia 29%
 - Italia 12%
 - Altri EU 26%
- Variabilità nei clusters tra regioni diverse: più elevata nelle grandi città (Londra) e città con aeroporti. Meno variabili in Scozia.

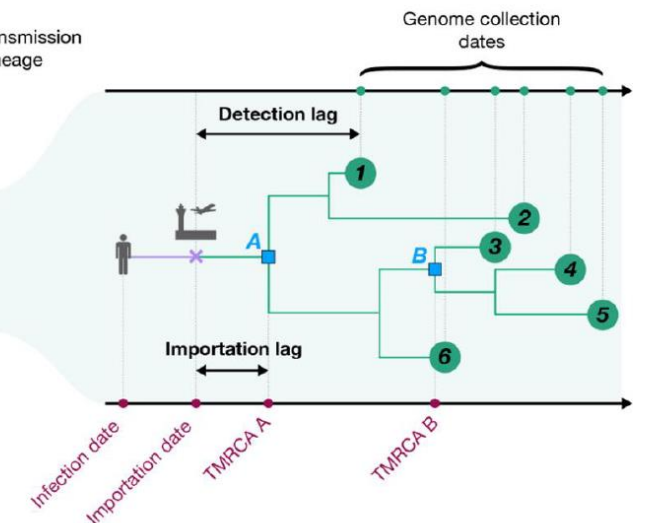
B



A

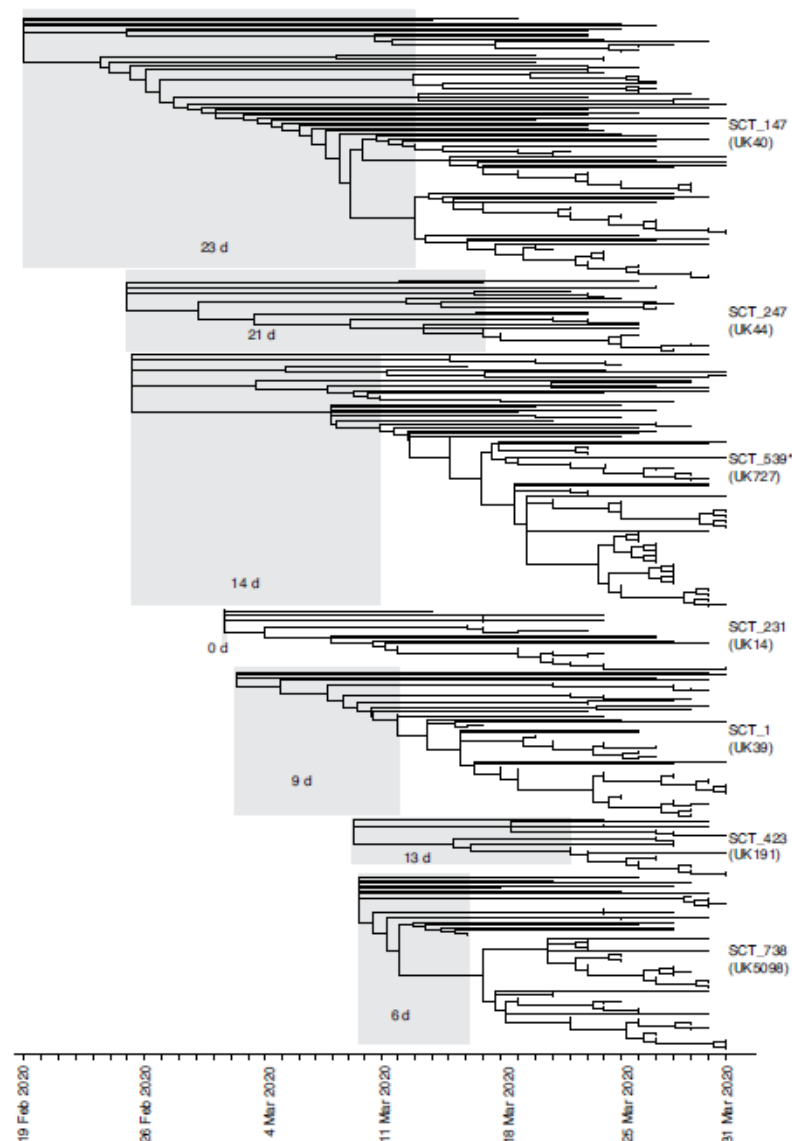


B



Filogeografia di SARS-CoV-2 in Scozia

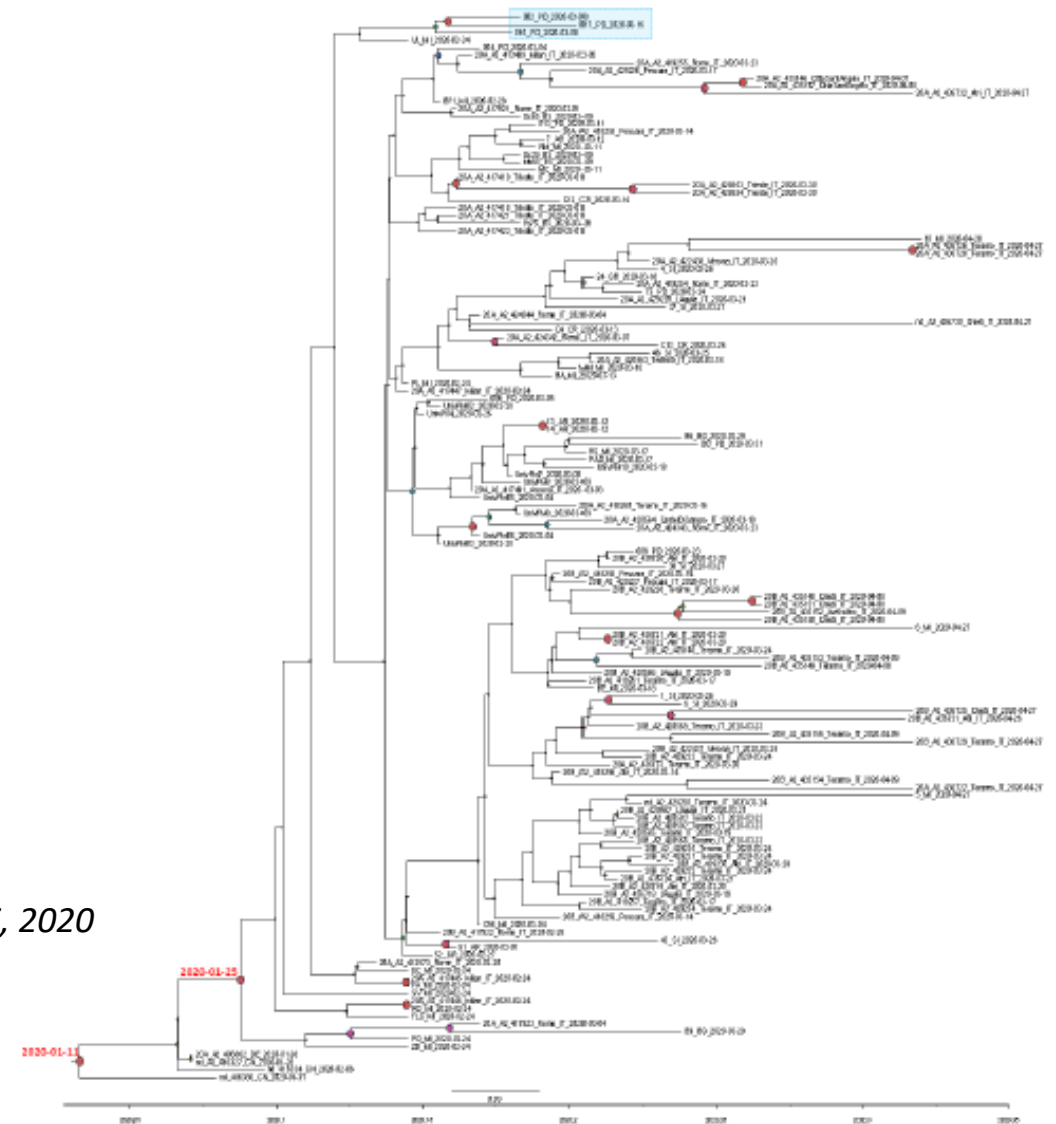
- Sequenziati 1314 genomi (il 49% dei casi confermati in Scozia)
- Combinati i dati genomici con quelli relativi ai viaggi e data di campionamento
- Maggioranza lineage B (B.1)
- Lineage locali 283 con genomi esteri
 - Il 40% contengono più di 1 seq scozzese
 - Il 60% erano singleton
- Dei cluster con + di 1 seq scozzese, circa il 30% coinvolgono persone con evidenze di viaggi internazionali
 - Maggioranza viaggi in Europa tra la fine di feb e primi di marzo (Italia, Spagna, Austria, Svizzera, ecc)
 - Cluster associati a viaggi in Asia rari
- Uno shift da cluster associati ai viaggi a cluster di trasmissione di comunità si è osservato dalla seconda metà di marzo
- Progenitore più precoce di cluster di comunità risale al 19 febbraio
 - Prima del primo caso individuato (1 marzo-viaggiatore)



Epidemiologia genomica di SARS-CoV-2 in Italia nella prima fase

- Ingresso stimato in Italia
 - 25/01/20 (23/01-10/02) *Lai et al*
 - 20/01/20 (16/01-23/01) *Altieri et al*
 - 15/01-30/01 *Giallonardo et al*
- Lineage prevalenti
 - B.1 (20A), 50-75%
 - B.1.1 (20B), 20-40%
 - B.1.5 (attualmente incluso nel B.1), 3-10%
- Il B.1 è andato diminuendo da febbraio a marzo/aprile sostituito da altri lineage (B.1.1 in particolare)
- Alcuni lineage si sono evoluti in Italia (B.1)
- Meno chiara è l'origine di B.1.1

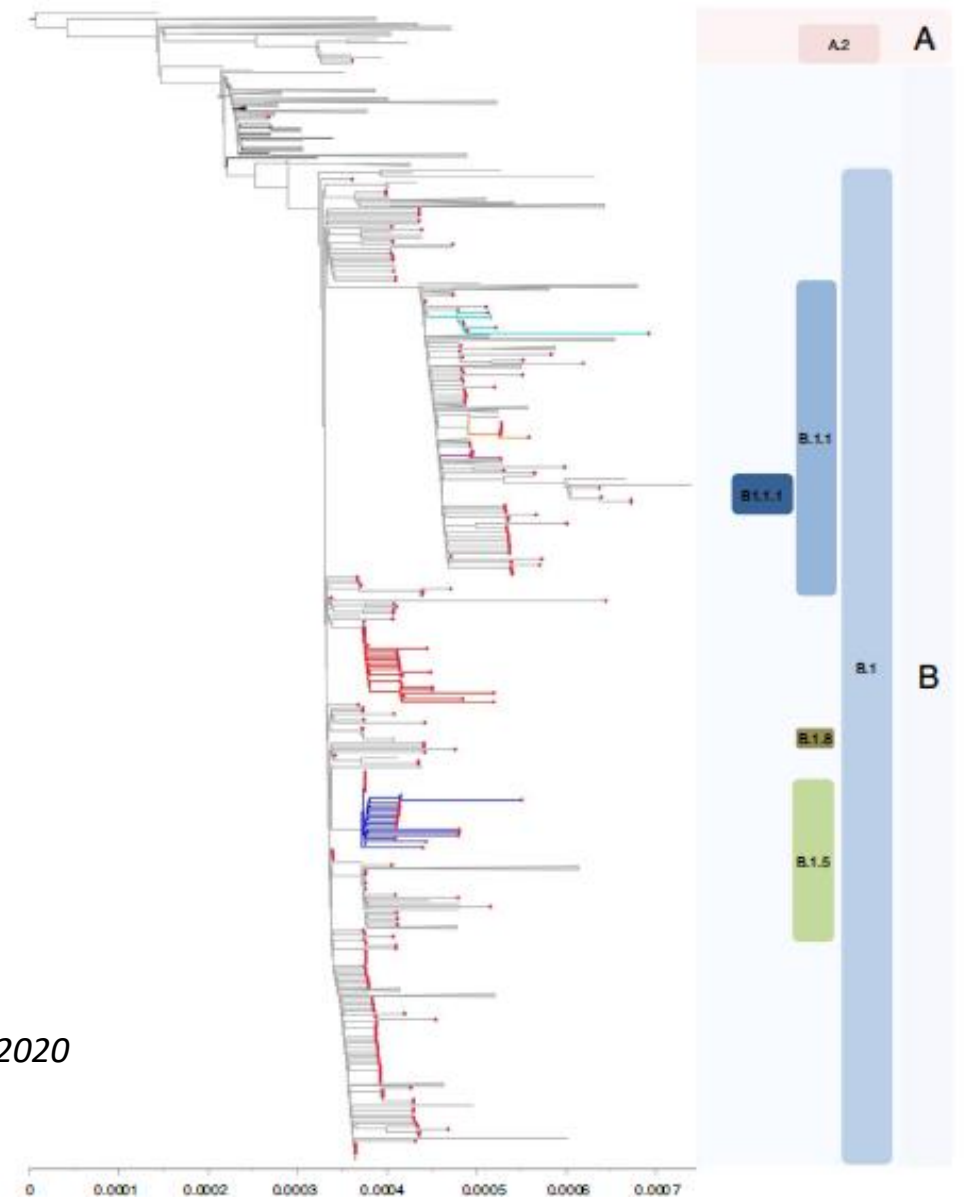
Lai, 2020



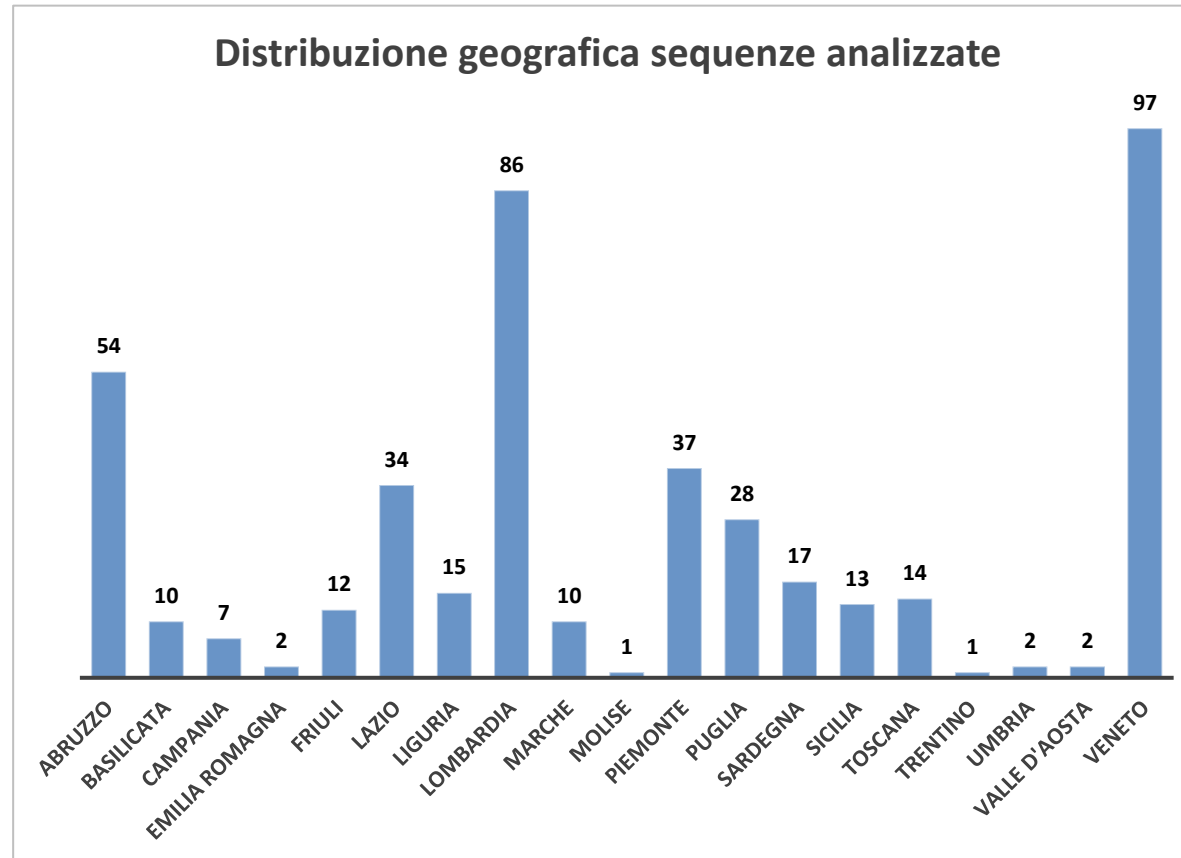
Epidemiologia genomica di SARS-CoV-2 in Italia nella prima fase

- Cluster di trasmissione:
 - Pochi nodi significativi
 - Relativamente pochi cluster (circa 10)
 - Tendenzialmente piccoli
 - Prevalentemente Italiani
 - Datazione precedente il lockdown

Altieri, 2020



Filogeografia



Totale sequenze italiane: 442

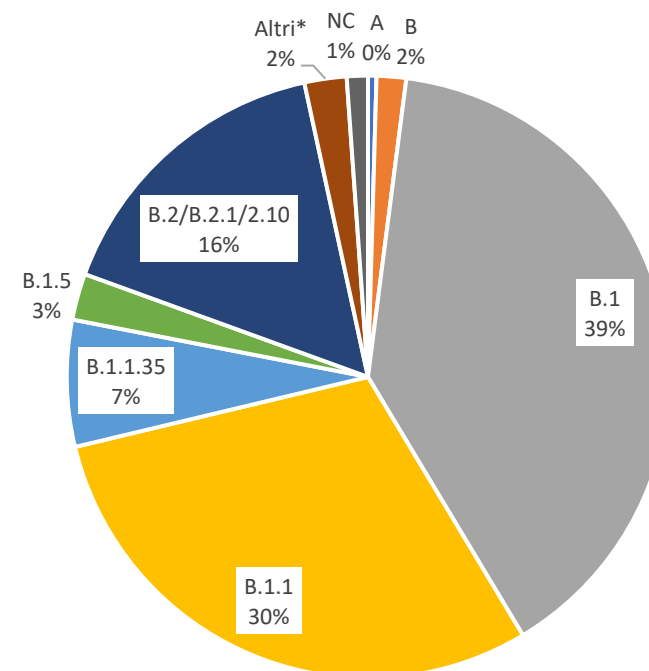
Totale sequenze nostre: 235

Lineages per località

Regione	Totale	A%	B%	B.1%	B.1.1%	B.1.1.35%	B.1.5%	B.2/2.1/2.	Altri%
Abruzzo	54			24.07	72.22		1.85		1.85
Basilicata	10			40.00	40.00	20.00			
Campania	7		14.29	28.57	14.29		28.57		14.29
Emilia Romagna	2			100.00					
Friuli	12			100.00					
Lazio	34	2.94	2.94	61.76	11.76			8.82	5.88
Liguria	15			40.00	33.33	6.67	13.33	6.67	
Lombardia	86		3.49	72.09	19.77		3.49		1.16
Marche	10			90.00			10.00		
Molise	1				100.00				
Piemonte	37			18.92	2.70	72.97	2.70	2.70	
Puglia	28	3.57	3.57	32.14	57.14				3.57
Sardegna	17		5.88		76.47				
Sicilia	13			7.69	76.92				15.38
Toscana	14			50.00	50.00				
Trentino	1			100.00					
Umbria	2			50.00	50.00				
Valle D'Aosta	2				100.00				
Veneto	97			17.53	11.34		1.03	68.04	2.06
Tot	437*	0.45	1.58	39.37	29.86	6.79	2.49	16.06	2.26

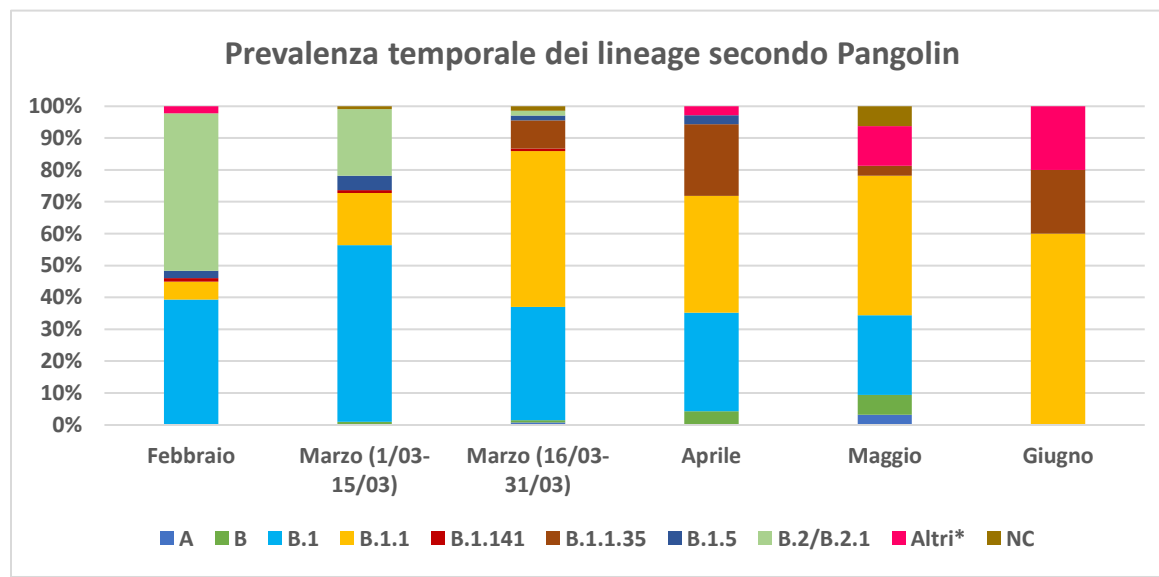
*5 non classificati

Prevalenza dei lineage secondo Pangolin



Lineages tempo

Data	Totale	A%	B%	B.1%	B.1.1%	B.1.1.35%	B.1.5%	B.2%
Febbraio	89			39.33	5.62	0.00	2.25	49.44
Marzo (1/03-15/03)	110		0.91	55.45	16.36	0.00	4.55	20.91
Marzo (16/03-31/03)	135	0.74	0.74	35.56	48.89	8.89	1.48	1.48
Aprile	71		4.23	30.99	36.62	22.54	2.82	0.00
Maggio	32	3.13	6.25	25.00	43.75	3.13	0.00	0.00
Giugno	5				60.00	20.00	0.00	0.00
Tot	442	0.45	1.58	39.37	29.86	6.79	2.49	15.61



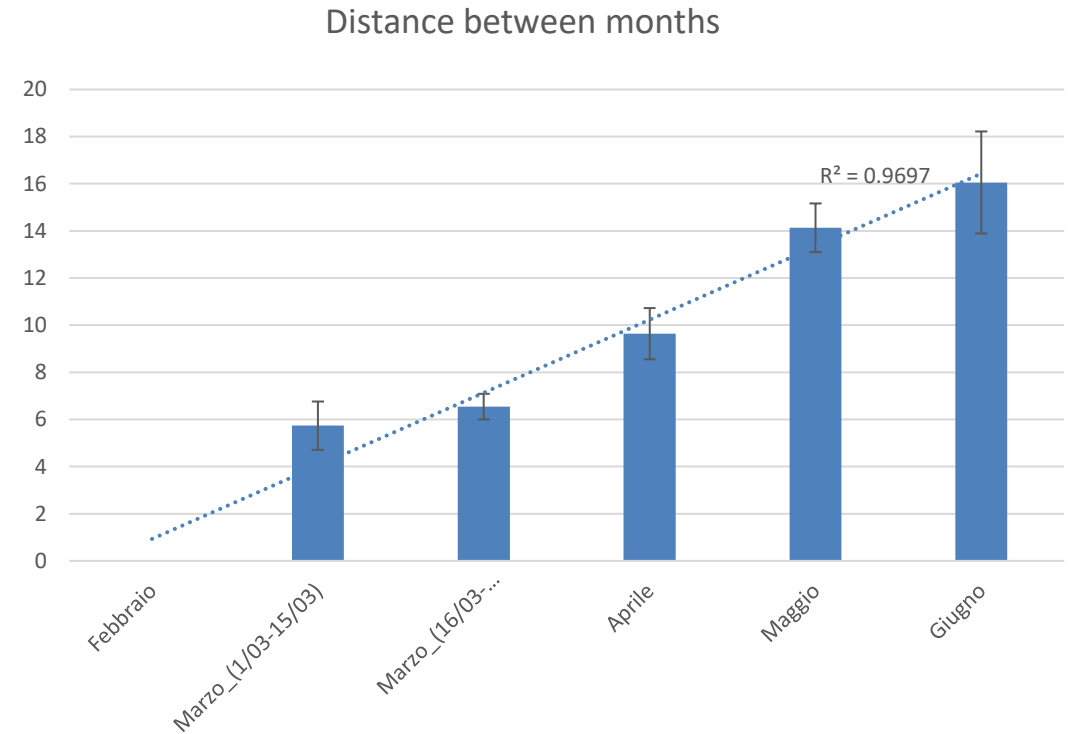
Differenze nucleotidiche within

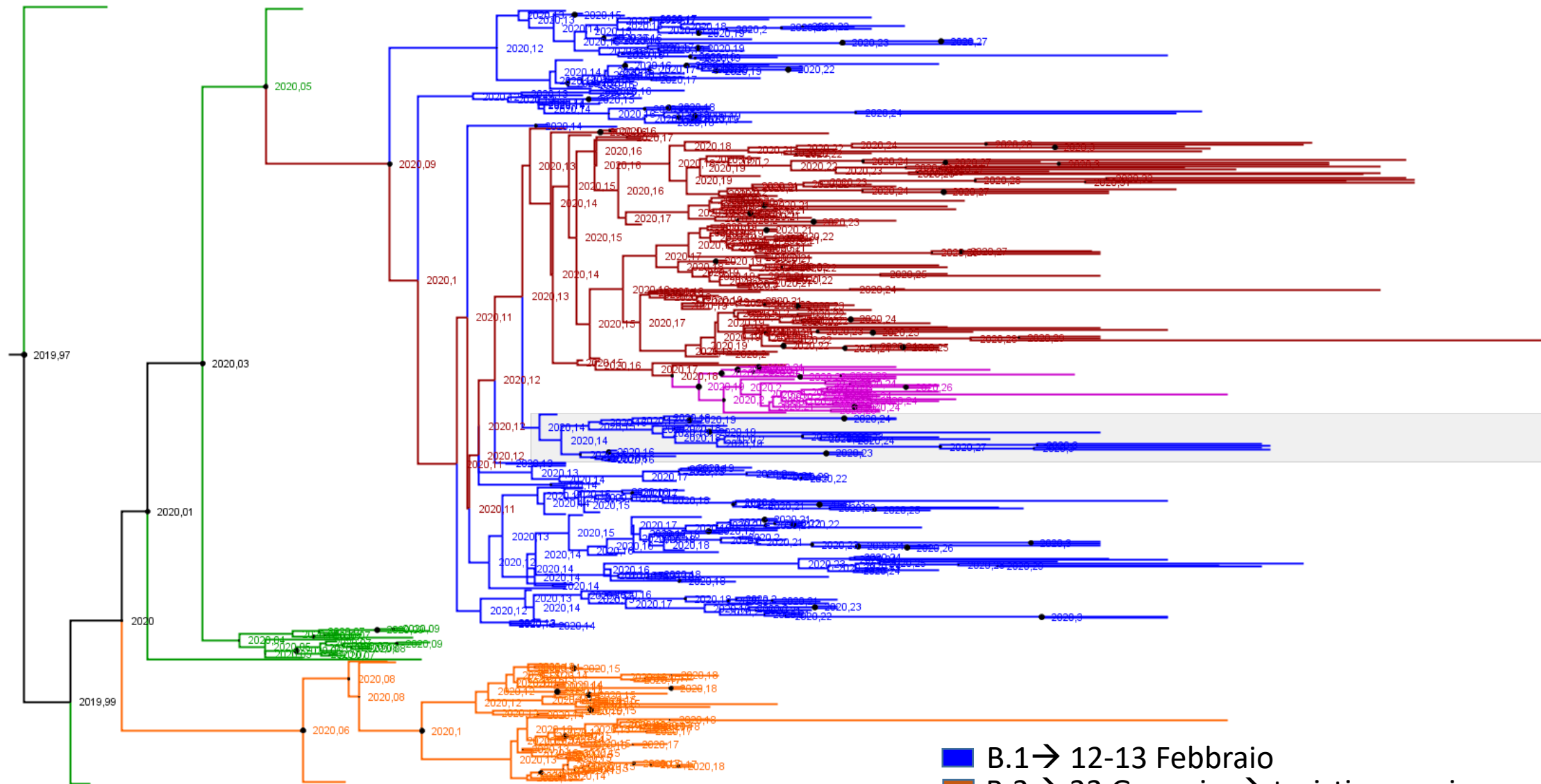
Mese	Numero differenze
Febbraio	5.8 (S.E. 0.834)
Marzo (1/03 -15/03)	7.2 (S.E. 0.856)
Marzo (16/03 -31/03)	8.7 (S.E. 1.044)
Aprile	13.0 (S.E. 1.226)
Maggio	20.0 (S.E. 1.323)

- Overall B1:
 - Nt =8.9 (0.85)
 - $pd (x10000)=3.6 (\pm 0.39)$
 - $ds (x10000)=3.8 (0.44)$
 - $dn (x10000)=3.0 (0.48)$
 - AA=4.8 (0.5)

Differenze per tempo

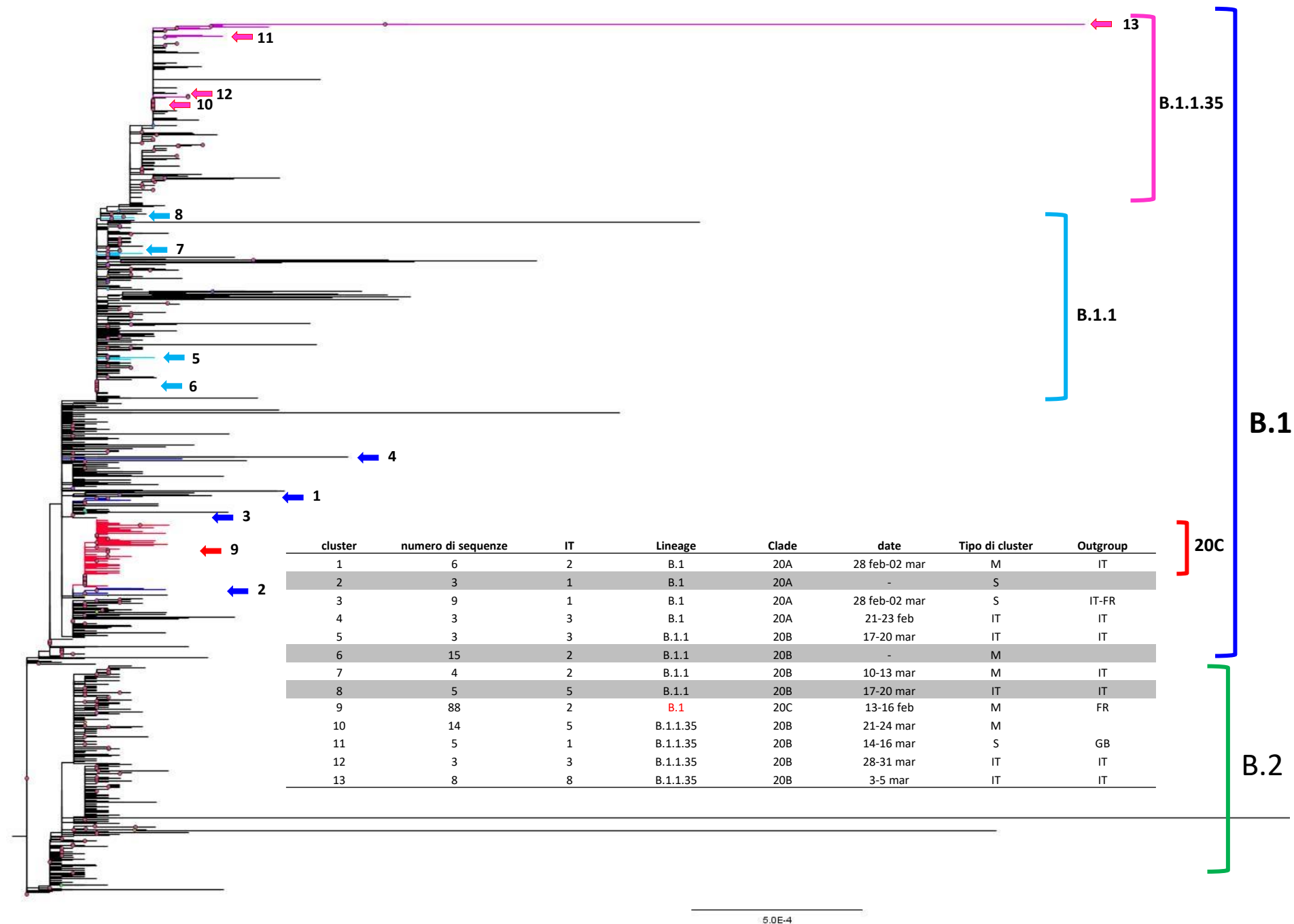
- Between





- B.1 → 12-13 Febbraio
- B.2 → 22 Gennaio → turisti romani
- B.1.1 → 16-17 Febbraio
- B.1.1.35 → 9-10 Marzo

5-6 Febbraio → 2 clade



Conclusioni

- Nella prima fase dell'infezione in Italia
- Introduzioni multiple tra febbraio e marzo 2020
 - B.1 in Lombardia
 - B2 in Veneto (Vo Euganeo)
 - B.1.1.35 in Piemonte
 - 20C (B.1) nel Lazio
- Diversi tempi di ingresso (inizialmente B.1 e B.2)
- La differenza genomica (scarsa-meno di 10 mutazioni) aumenta col tempo (in particolare da aprile in poi)-> evoluzione del virus? Aumento degli ingressi?
- Almeno 13 cluster individuati contengono circa 8% di sequenze Italiane (38/442)
 - 5 Italiani (trasmissione comunità)
 - 5 Misti (presenti in più Paesi EU)
 - 3 Singoli (introduzioni sporadiche?)
- Maggior parte dei genomi mescolati nell'albero intero

Grazie

Ai collaboratori

Al network (SCIRE)

Agli organizzatori

A Galli

E a tutti voi...