Heliyon 10 (2024) e36162

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Epi-Clock: A sensitive platform to help understand pathogenic disease outbreaks and facilitate the response to future outbreaks of concern

Cong Ji [*], Junbin (Jack) Shao [**]

Liferiver Science and Technology Institute, Shanghai ZJ Bio-Tech Co., Ltd., Shanghai, China

ARTICLE INFO

Keywords:
Divergence
Host bias
Mutation type
Codon usage bias
ZHU algorithm

ABSTRACT

To predict potential epidemic outbreaks, we tested our strategy, Epi-Clock, which applies the novel ZHU algorithm to different SARS-CoV-2 datasets before outbreaks to search for significant mutational accumulation patterns correlated with outbreak events. Surprisingly, some interspecies genetic distances in Coronaviridae may represent intermediate states of different species or subspecies in the evolutionary history of Coronaviridae. The insertions and deletions in whole-genome sequences between different hosts were separately associated with important roles in host transmission and shifts in Coronaviridae. Furthermore, we believe that non-nucleosomal DNA may play a dominant role in the divergence of different lineages of SARS-CoV-2 in different regions of the world owing to the lack of nucleosome protection. We suggest that strong selective variation among different lineages of SARS-CoV-2 is required to produce strong codon usage bias, which appears in B.1.640.2 and B.1.617.2 (Delta). Notably, we found that an increasing number of other types of substitutions, such as those resulting from the hitchhiking effect, accumulated, especially in the pre-breakout phase, although some of the previous substitutions were replaced by other dominant genotypes. From most validations, we could accurately predict the potential pre-phase of outbreaks with a median interval of 5 days.

1. Introduction

Determining how viruses originate and diverge, and how a disease is transmitted, is extremely important for understanding viral disease outbreaks and facilitating responses to future outbreaks. Since viruses lack fossilisation, preventing the availability of any references or ancestors for inferring evolutionary processes, some theories about the origin of viruses have been reported, such as the degeneracy theory, DNA escape from plasmids or transposons, and viroid or satellite viruses [1]. Several methods have been proposed, including phylogenetic [2], neutral selection [3], shared ancestry inference [4], and divergence [5] approaches. Most highly pathogenic mutations become extinct in the population, whereas adaptive mutations are fixed in the population [3]. To answer the above crucial questions about outbreaks [6], sequencing of viral samples and supplementation of epidemiological methods could play an important role in providing nucleotide-level resolution data for outbreak-causing pathogens.

How does viral evolution occur in different hosts? Why do pathogens successfully jump between host species but not between

[*] Corresponding author.

[**] Corresponding author.

E-mail addresses: c_ji@liferiver.com.cn (C. Ji), junbin_shao@liferiver.com.cn (J.(J. Shao).

https://doi.org/10.1016/j.heliyon.2024.e36162
Received 29 October 2023; Received in revised form 8 August 2024; Accepted 11 August 2024
Available online 29 August 2024

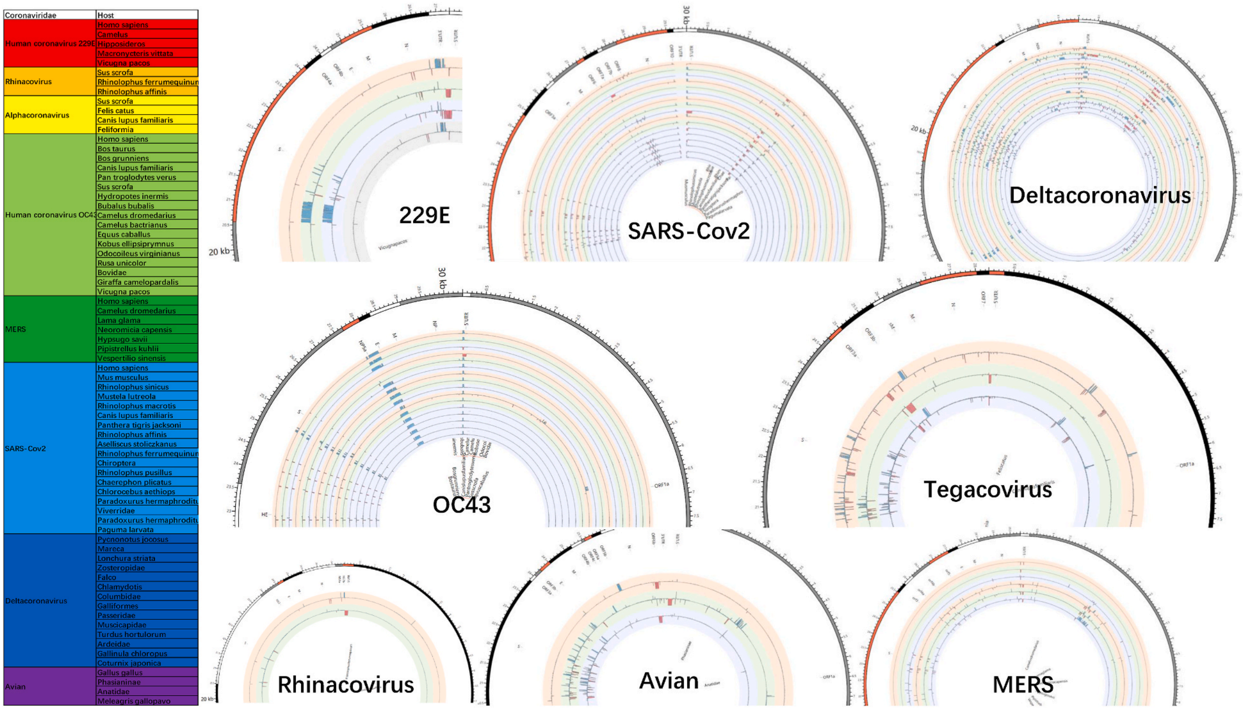
others? Occasionally, preventing host shifts may come at the cost of other aspects of the pathogen's fitness [7–10]. The evolution of EBOV [11] and the specific amino acid substitutions in the EBOV GP have increased the tropism for human cells and infectivity, enhancing the ability of transmission among humans [12]. The association between viral divergence and adaptation to host transfer is strongly selective. Nucleosome occupancy nearly eliminated cytosine deamination and suppressed, spontaneous mutations by nucleosomes in a base-specific manner in eukaryotes [13]. Viral doublet histones are essential for viral infectivity, localise to the cytoplasmic viral factories after viral infection, and are ultimately found in mature virions [14]. Giant viruses belonging to the nucleocytoplasmic large DNA virus (NCLDV) group possess histone-like genes in their genomes. Host−virus arm races are a powerful source of adaptation, and most genes gained by NCLDVs from their different hosts are likely linked to viral defences [15].

To compare viral genomes before and after epidemics, the Global Epidemic and Mobility Model [16] and Epi-Factors [17] were used to search for the accumulation pattern of pathogenic mutations contributing to the outbreak, to prevent disease recurrence worldwide [18]. As a result of the virus evolving under immune system selection pressure in infected individuals, lineage B.1.1.7 has presumably arisen [19]. Lambda, a new variant of interest, is now spreading in some South American countries and is attributed to the T76I and L452Q mutations [20]. With the first wave fixed at the start of the pandemic, D614G was the foundation for all subsequent waves of strains(C241T, C3037T, C14408T, and A23403G) [21]. Tracking the spread of infectious diseases to assist in their control has traditionally relied on analysing case data gathered as the outbreak proceeds [22]. Here, we demonstrate our strategy, to develop a sensitive process for predicting the potential trigger of outbreaks to facilitate the response to future outbreaks of concern.

2. Results

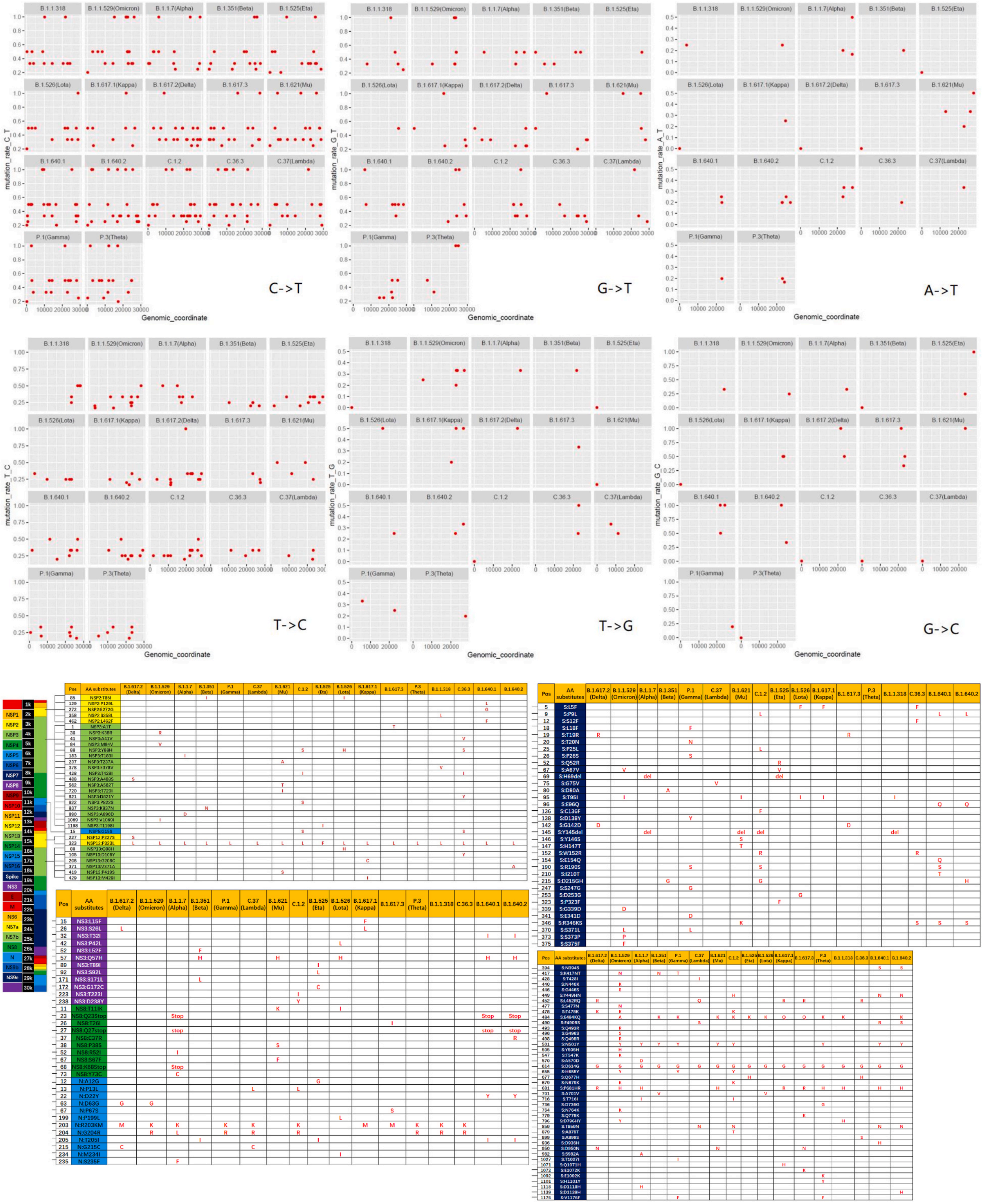
2.1. Divergence of the whole Coronaviridae family within different populations and adaptation to cross host species barriers

To comprehensively explore the evolution of the whole family of *Coronaviridae*, we performed a genomic comparison of



Coronaviridae	Host
Human coronavirus 229E	Homo sapiens
	Camelus
	Hipposideros
	Macronycteris vittata
	Vicugna pacos
Rhinacovirus	Sus scrofa
	Rhinolophus ferrumequinun
	Rhinolophus affinis
Alphacoronavirus	Sus scrofa
	Felis catus
	Canis lupus familiaris
	Feliformia
Human coronavirus OC43	Homo sapiens
	Bos taurus
	Bos grunniens
	Canis lupus familiaris
	Pan troglodytes verus
	Sus scrofa
	Hydropotes inermis
	Bubalus bubalis
	Camelus dromedarius
	Camelus bactrianus
	Equus caballus
	Kobus ellipsiprymnus
	Odocoileus virginianus
	Rusa unicolor
	Bovidae
	Giraffa camelopardalis
	Vicugna pacos
MERS	Homo sapiens
	Camelus dromedarius
	Lama glama
	Neoromicia capensis
	Hypsugo savii
	Pipistrellus kuhlii
	Vespertilio sinensis
SARS-Cov2	Homo sapiens
	Mus musculus
	Rhinolophus sinicus
	Mustela lutreola
	Rhinolophus macrotis
	Canis lupus familiaris
	Panthera tigris jacksoni
	Rhinolophus affinis
	Aselliscus stoliczkanus
	Rhinolophus ferrumequinun
	Chiroptera
	Rhinolophus pusillus
	Chaerephon plicatus
	Chlorocebus aethiops
	Paradoxurus hermaphroditu
	Viverridae
	Paradoxurus hermaphroditu
	Paguma larvata
Deltacoronavirus	Pycnonotus jocosus
	Mareca
	Lonchura striata
	Zosteropidae
	Falco
	Chlamydotis
	Columbidae
	Galliformes
	Passeridae
	Muscicapidae
	Turdus hortulorum
	Ardeidae
	Gallinula chloropus
	Coturnix japonica
Avian	Gallus gallus
	Phasianinae
	Anatidae
	Meleagris gallopavo

Fig. 1. Host shifts in the evolution of *Coronaviridae*. Here, we partly present the divergence of whole genome among different hosts in different species or subspecies of *Coronaviridae*, such as human coronavirus 229E in *Camelus*, *Hipposideros*, *Macronycterisvittata*, and *Vicugnapacos* compared with in *Homo sapiens*; SARS-CoV-2 in *Mus musculus, Rhinolophus sinicus, Mustela lutreola, Rhinolophus macrotis, Canis lupus familiaris, Panthera tigris jacksoni, Rhinolophus affinis, Aselliscus stoliczkanus, Rhinolophus ferrumequinum, Chiroptera, Rhinolophus pusillus, Chaerephon plicatus, Chlorocebus aethiops, Paradoxurus hermaphroditus, Viverridae, Paradoxurus hermaphroditus, Paguma larvata* compared with in *Homo sapiens*; Deltacoronavirus in *Mareca, Lonchura striata, Zosteropidae, Falco, Chlamydotis, Columbidae, Galliformes, Passeridae, Muscicapidae, Turdus hortulorum, Ardeidae, Gallinula chloropus, Coturnix japonica;* Human coronavirus OC43 in *Bos Taurus, Bos grunniens, Canis lupus familiaris, Pan troglodytes verus, Sus scrofa, Hydropotes inermis, Bubalus bubalis, Camelus dromedaries, Camelus bactrianus, Equus caballus, Kobus ellipsiprymnus, Odocoileus virginianus, Rusa unicolor, Bovidae, Giraffa Camelopardalis, Vicugna pacos;* Alphacoronavirus in *Felis catus, Canis lupus familiaris, Feliformia* compared with in *Sus scrofa*; Rhinacovirus in *Rhinolophus ferrumequinum, Rhinolophus affinis* compared with in *Sus scrofa*; Avian in *Phasianinae, Anatidae, Meleagris gallopavo* compared with in *Gallus gallus*; Middle East respiratory syndrome-related coronavirus in *Camelus dromedaries, Lama glama, Neoromicia capensis, Hypsugo savii, Pipistrellus kuhlii, Vespertilio sinensis* compared with in *Homo sapiens*.



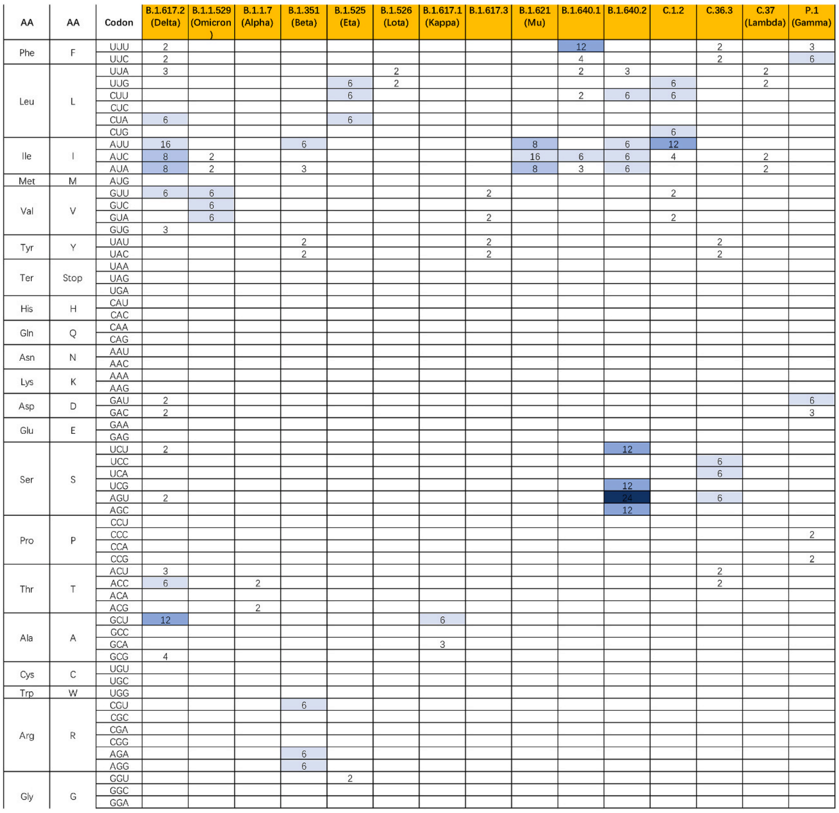
Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
85	NSP2:T85I				I						I							
129	NSP2:P129L																L	
272	NSP2:E272G																G	
358	NSP2:S358L														L			
462	NSP2:L462F																F	
1	NSP3:A1T												T					
38	NSP3:K38R		R															
41	NSP3:A41V															V		
84	NSP3:M84V		V															
88	NSP3:Y88H								S		H					S		
183	NSP3:T183I			I														
237	NSP3:T237A							A										
378	NSP3:E378V														V			
428	NSP3:T428I								I							I		
488	NSP3:A488S	S																
562	NSP3:A562T							T										
720	NSP3:T720I							I										
821	NSP3:D821Y															Y		
822	NSP3:P822S								S									
837	NSP3:K837N				N													
890	NSP3:A890D			D														
1069	NSP3:V1069I		I															
1198	NSP3:T1198I									I								
15	NSP5:G15S								S							S		
227	NSP12:P227S	S																
323	NSP12:P323L	L	L	L	L	L	L	L	L	F	L	L	L	L	L	L	L	L
88	NSP13:Q88H										H							
105	NSP13:D105Y															Y		
206	NSP13:G206C											C						
371	NSP13:V371A																	A
419	NSP13:P419S							S										
429	NSP13:M429I											I						

Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
15	NS3:L15F											F						
26	NS3:S26L	L										L						
32	NS3:T32I																I	I
42	NS3:P42L										L							
52	NS3:L52F				F													
57	NS3:Q57H				H			H			H						H	H
89	NS3:T89I									I								
92	NS3:S92L									L								
171	NS3:S171L				L													
172	NS3:G172C									C								
223	NS3:T223I								I									
238	NS3:D238Y								Y									
11	NS8:T11IK							K			I							
23	NS8:Q23Stop			Stop													Stop	Stop
26	NS8:T26I												I					
27	NS8:Q27stop			stop													stop	stop
37	NS8:C37R																	R
38	NS8:P38S							S										
52	NS8:R52I			I														
67	NS8:S67F							F										
68	NS8:K68Stop			Stop														
73	NS8:Y73C			C														
12	N:A12G									G								
13	N:P13L						L		L									
22	N:D22Y																Y	Y
63	N:D63G	G	G															
67	N:P67S												S					
199	N:P199L										L							
203	N:R203KM	M	K	K		K	K		K			M	M	K	K	K		
204	N:G204R		R	L		R	R		R					R	R	R		
205	N:T205I				I					I							I	I
215	N:G215C	C					C											
234	N:M234I										I							
235	N:S235F			F														

Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
5	S:L5F										F	F				F		
9	S:P9L								L								L	L
12	S:S12F															F		
18	S:L18F					F												
19	S:T19R	R											R					
20	S:T20N					N												
25	S:P25L								L									
26	S:P26S					S												
52	S:Q52R									R								
67	S:A67V		V							V								
69	S:H69del			del						del								
75	S:G75V						V											
80	S:D80A				A													
95	S:T95I		I					I			I	I			I			
96	S:E96Q																Q	Q
136	S:C136F								F									
138	S:D138Y					Y												
142	S:G142D	D											D					
145	S:Y145del			del				del	del						del			
146	S:Y146S							S										
147	S:H147T							T										
152	S:W152R								R							R		
154	S:E154Q																Q	
190	S:R190S					S			S								S	
210	S:I210T																T	
215	S:D215GH				G				G									H
247	S:S247G					G												
253	S:D253G										G							
323	S:P323F									F								
339	S:G339D		D															
341	S:E341D					D												
346	S:R346KS							K								S	S	S
370	S:S371L		L			L												
373	S:S373P		P															
375	S:S375F		F															

Pos	AA substitutes	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	P.1 (Gamma)	C.37 (Lambda)	B.1.621 (Mu)	C.1.2	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	P.3 (Theta)	B.1.1.318	C.36.3	B.1.640.1	B.1.640.2
394	S:N394S																S	S
417	S:K417NT		N		N	T												
428	S:T428I						I											
440	S:N440K		K															
446	S:G446S		S															
449	S:Y449HN								H								N	N
452	S:L452RQ	R					Q					R	R			R		
477	S:S477N		N															
478	S:T478K	K	K						K									
484	S:E484KQ		A		K	K		K	K	K	K	Q	Q	K	K			K
490	S:F490RS						S										R	S
493	S:Q493R		R															
496	S:G496S		S															
498	S:Q498R		R															
501	S:N501Y		Y	Y	Y	Y		Y	Y					Y			Y	Y
505	S:Y505H		H															
547	S:T547K		K															
570	S:A570D			D														
614	S:D614G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
655	S:H655Y		Y			Y			Y									
677	S:Q677H									H						H		
679	S:N679K		K						K									
681	S:P681HR	R	H	H				H				R	R	H	H		H	H
701	S:A701V				V						V							
716	S:T716I			I					I									
736	S:D736G													G				
764	S:N764K		K															
779	S:Q779K												K					
796	S:D796HY		Y												H			
859	S:T859N						N		N								N	N
879	S:A879T								T									
899	S:A899S															S		
936	S:D936H																H	
950	S:D950N	N						N					N					
982	S:S982A			A														
1027	S:T1027I					I												
1071	S:Q1071H											H						
1072	S:E1072K												K					
1092	S:E1092K													K				
1101	S:H1101Y													Y				
1118	S:D1118H			H														
1139	S:D1139H																	H
1176	S:V1176F					F								F				

Fig. 2. Distribution of mutation rates for different mutation types and codon usage bias of different lineages of SARS-CoV-2. 2a. Different lineages are represented by different mutation types, i.e., C- > T, G- > T, A- > T, T- > C, T- > G, G- > C. **2b.** The distribution of amino acid substitutions among different SARS-CoV-2 lineages. Here the rainbow bars on the left show represent all the coding genes in the whole genome of SARS-CoV-2. From 1k to 30kbp, there are separately *NSP1, NSP2, NSP3, NSP4, NSP5, NSP6, NSP7, NSP8, NSP9, NSP10, NSP11, NSP12, NSP13, NSP14, NSP15, NSP16, Spike, NS3, E, M, NS6, NS7a, NS7b, NS8, N, NS9b,* and *NS9c*. **2c.** RSCU values in different lineages of SARS-CoV-2.



AA	AA	Codon	B.1.617.2 (Delta)	B.1.1.529 (Omicron)	B.1.1.7 (Alpha)	B.1.351 (Beta)	B.1.525 (Eta)	B.1.526 (Lota)	B.1.617.1 (Kappa)	B.1.617.3	B.1.621 (Mu)	B.1.640.1	B.1.640.2	C.1.2	C.36.3	C.37 (Lambda)	P.1 (Gamma)
Phe	F	UUU	2									12			2		3
		UUC	2									4			2		6
Leu	L	UUA	3					2				2	3			2	
		UUG					6	2						6		2	
		CUU					6					2	6	6			
		CUC															
		CUA	6				6										
		CUG												6			
Ile	I	AUU	16			6					8		6	12			
		AUC	8	2							16	6	6	4		2	
		AUA	8	2		3					8	3	6			2	
Met	M	AUG															
Val	V	GUU	6	6						2				2			
		GUC		6													
		GUA		6						2				2			
		GUG	3														
Tyr	Y	UAU				2				2					2		
		UAC				2				2					2		
Ter	Stop	UAA															
		UAG															
		UGA															
His	H	CAU															
		CAC															
Gln	Q	CAA															
		CAG															
Asn	N	AAU															
		AAC															
Lys	K	AAA															
		AAG															
Asp	D	GAU	2														6
		GAC	2														3
Glu	E	GAA															
		GAG															
Ser	S	UCU	2										12				
		UCC													6		
		UCA													6		
		UCG											12				
		AGU	2										24		6		
		AGC											12				
Pro	P	CCU															
		CCC															2
		CCA															
		CCG															2
Thr	T	ACU	3												2		
		ACC	6		2										2		
		ACA															
		ACG			2												
Ala	A	GCU	12						6								
		GCC															
		GCA							3								
		GCG	4														
Cys	C	UGU															
		UGC															
Trp	W	UGG															
Arg	R	CGU				6											
		CGC															
		CGA															
		CGG															
		AGA				6											
		AGG				6											
Gly	G	GGU					2										
		GGC															
		GGA															

Fig. 2. (*continued*).

Coronaviridae to search for new age patterns related to virulence or transmissibility (Supplementary Fig. 1). In the analysis of intra- and inter-species genetic distances of whole genome sequences of *Coronaviridae*, it was evident that most inter-species genetic distances of *Coronaviridae* were longer than intra-species distances, that is SARS-CoV-2, SARS-CoV, SADS, NL63, MERS, London1, HKU5, HKU4, HKU3, HKU2, and BATS. For instance, the sequencing similarities of SARS-CoV-2 and bat-SL-CovZC45 and SARS-CoV-2 and bat-SL-CovZXC21 were nearly 88 %; however, the similarities of SARS-CoV-2, SARS-CoV, SARS-CoV-2, and MERS were 79 % and 50 %, respectively. Based on a comparison of the genetic distances of H1N1 and H3N2 with those of SARS-CoV-2 and SARS-CoV, we found that the p-distance of SARS-CoV-2 and Bat SARS-like CoV was nearly 0.13, while that of SARS-CoV-2 and SARS-CoV was 0.24 MEGA [23]. Both these values are lower than the p-distances of H1N1 and H3N2 (0.8) and H1N1 and H7N9 (0.73). Conversely, some interspecies genetic distances of *Coronaviridae* are shorter than the intraspecies genetic distances, such as those for OC43 and 229E, which may be the intermediate states of different species or subspecies in the entire evolutionary history of *Coronaviridae.*

Here, we explored the divergence time of species and subspecies of *Coronaviridae*, which presented the diversity of such ages within populations separately associated with new functions or important turning points. We explored the dated variants of different species or subspecies and generated a new age atlas using SARS-CoV-2 as the reference genome (Supplementary Fig. 2). It has been demonstrated that SARS-CoV, HKU3, and BATS are very close to SARS-CoV-2 and far away from OC43 and HKU1 on the S protein, although there are 5 kb insertions of BATS in Orf7a, Orf7b, and Orf8 and 15 kb deletions of OC43, HKU9, and HKU1 in Orf7a, Orf7b, Orf8, and N. Most strikingly, the insertions and deletions of whole genome sequences between different hosts play an important role in host transmission and the shift of *Coronaviridae* presented by Circos (Fig. 1 and Supplementary Table 1). For instance, human coronavirus 229E in *Camelus* and alpaca (*Vicugna pacos)* has S protein sequences similar to those observed in *Homo sapiens*, while having 500 bp deletions in striped leaf-nosed bats (*Hipposideros* and *Macronycteris vittata)*. When used as the reference genome, *Rhinacovirus* in piglets (*Sus scrofa)* was similar to that in wild greater horseshoe bats (*Rhinolophus ferrumequinum)* and divergent from that in intermediate horseshoe bats (*Rhinolophus affinis)* in that insertions and deletions all appeared in the S protein, which may be related to host bias. It is likely that the accumulation of insertions and deletions emerged in the S protein of alphacoronavirus, human coronavirus OC43, Middle East respiratory syndrome-related coronavirus, severe acute respiratory syndrome-related coronavirus, *Deltacoronavirus*, and avian coronavirus. The similarity of SARS-CoV-2 S proteins in the Chinese rufous horseshoe bat (*Rhinolophus sinicus)*, big-eared horseshoe bat (*Rhinolophus macrotis)*, dogs (*Canis lupus familiaris)*, and Malayan tiger *(Panthera tigris jacksoni)* is very high with the SARS-CoV-2 S protein in *Homo sapiens*, implying that the ancestral hosts of SARS-CoV-2 in *Homo sapiens* should be closely related to the species, as detailed in Supplementary Figs. 3–10. In summary, the more mutations viruses have, the greater the possibility of adaptation, allowing them to cross host species barriers. This indicates that increased host diversity of pathogens would accumulatively enlarge the population size of potential hosts with different species or subspecies and protect against the emergence of pathogens.

2.2. Distribution of rates of different mutation types and codon usage bias in different lineages of SARS-CoV-2

Across different lineages of SARS-CoV-2, 538 nucleotide substitutions such as C1876T(B.1.1.318), A2791G(B.1.1.529), C874T (B.1.1.7), C1055T(B.1.351), C1473T(B.1.525), C1034T(B.1.526), C2999T(B.1.617.1), C3012T(B.1.617.2), C787T(B.1.617.3), C3033T(B.1.621), C535T(B.1.640.1), C1154T(B.1.640.2), C515T(C.1.2), C569T(C.36.3), C3032T(C.37), T702C(P.1), and C1714T (P.3) were found in the regions of the NSP2, NSP3, NSP12, NSP13, S, N, NS8, and NS3 gene as presented in Supplementary Fig. 11 and the "AA substitutes" sheet of Supplementary Table 1. Then, we demonstrated the richness distribution of different mutation types across the whole genomes, in which mutation rates of C- > T, G- > T and T- > C were dominant in driving the divergence of different lineages of SARS-CoV-2 (Fig. 2a and Supplementary Figs. 12–17). In particular, the mutation rate distributions of C- > T are highly enriched in all lineages of SARS-CoV-2, and the mutation rate distributions of G- > T are relatively enriched in B.1.1.318, B.1.1.529 (Omicron), B.1.617.1 (Kappa), B.1.617.3, B.1.621 (Mu), B.1.640.1, B.1.640.2, C.1.2, C.37 (Lambda), and P.3 (Theta). Conversely, the patterns of A- > T, T- > G, G- > C are relatively rare, and it was clear that some points with high mutation rates in the A- > T distribution derive from B.1.1.7 (Alpha) and B.1.621 (Mu); several points with high mutation rates in the T- > G distribution from B.1.526 (Lota), B.1.617.1 (Kappa), B.1.617.2 (Delta), and C.36.3; and some points in the G- > C distribution from B.1.525 (Eta), B.1.617.2 (Delta), B.1.617.3, B.1.621 (Mu), B.1.640.1 and B.1.640.2. Therefore, we believe that nucleosomal hereditary material (DNA/RNA) undergoes fewer C- > T mutations in SARS-CoV-2, protected by viral doublet histones essential for viral infectivity and viral defences leading to multiple host adaptation/transmission. In contrast, non-nucleosomal deoxyribonucleic acids may play a dominant role in the divergence of various SARS-CoV-2 lineages in different regions of the world without nucleosome protection.

Furthermore, we present the recent amino acid substitutions among different SARS-CoV-2 lineages, including B.1.617.2 (Delta), B.1.1.529 (Omicron), B.1.1.7 (Alpha), B.1.351 (Beta), P.1 (Gamma), C.37 (Lambda), B.1.621 (Mu), C.1.2, B.1.525 (Eta), B.1.526 (Lota), B.1.617.1 (Kappa), B.1.617.3, P.3 (Theta), B.1.1.318, C.36.3, B.1.640.1, and B.1.640.2, with a focus on NSP2, NSP3, NSP5, NSP12, NSP13, Spike, NS3, NS8, and N proteins in Fig. 2b and Supplementary Table 1, which have strong effects on the divergence and evolutionary history of different lineages. This is especially true for the spike protein, which contains a receptor-binding domain (RBD), a fusion domain, a transmembrane domain, and the NSP3 protein, which is the N-terminus of the coronavirus SARS-CoV non-structural protein 3 (Nsp3) and related proteins. In addition to the prevalent amino acid substitutions such as NSP12:P323L and S:D614G, distinct and specific substitutions were observed in one lineage compared to another. Examples include the NSP2:P129L, NSP2:E272G, NSP2:L462F, S:E154Q, S:I210T, and S:D936H substitutions in the B.1.640.1 lineage; NSP13:V371A, S:P129L, S:D1139H, and NS8:C37R substituted in the B.1.640.2 lineage; NSP2:S358L, and NSP3:E378V substitutions in the B.1.1.318 lineage; NSP3:A41V, NSP3:D821Y, NSP13:D105Y, S:S12F, S:G212V, and S:A899S substitutions in the C.36.3 lineage; S:K2Q, S:L280F, S:G313S, S:A368V, S:D736G, S:E1092K, and S:H1101Y substitutions in the P.3 (Theta) lineage; NSP3:A1T, S:Q779K, S:E1072K, NS8:T26I, and N:P67S substitutions in the B.1.617.3 lineage; NSP13:G206C, NSP13:M429I, S:Q1071H, and NS3:L15F in the B.1.617.1 (Kappa) lineage; NSP13:Q88H, S:D253G, NS3:P42L, N:P199L, N:M234I, S:D253G, NS3:P42L, N:P199L, and N:M234I substitutions in the B.1.526 (Lota) lineage; NSP3:T1198I, S:Q52R, S:P323F, S:Q677H, NS3:T89I, NS3:S92L, NS3:G172C, and N:A12G substitutions in the B.1.525 (Eta) lineage; NSP3:P822S, S:P25L, S:C136F, S:A879T, NS3:T223I, and NS3:D238Y substitutions in the C.1.2 lineage; NSP3:T237A, NSP3:A562T, NSP3:T720I, NSP13:P419S, S:Y146S, S:H147T, S:T205I, S:R346K, NS8:P38S, and NS8:S67F substitutions in the B.1.621 (Mu) lineage; S:P13L, S:G75V, S:V76I, and S:T428I substitutions in the C.37 (Lambda) lineage; S:L18F, S:T20N, S:P26S, S:D138Y, S:S247G, S:E341D, S:S371L, S:K977Q, and S:T1027I substitutions in the P.1 (Gamma) lineage; NSP3:K837N, S:D80A, NS3:L52F, and NS3:S171L substitutions in the B.1.351 (Beta) lineage; NSP3:T183I, NSP3:A890D, S:A570D, S:S982A, S:D1118H, NS8:R52I, NS8:K68Stop, NS8:Y73C, and N:S235F substitutions in the B.1.1.7 (Alpha) lineage; NSP3:K38R, NSP3:M84V, NSP3:V1069I, S:G339D, S:S371L, S:S373P, S:S375F, S:N440K, S:G446S, S:S477N, S:Q493R, S:G496S, S:Q498R, S:Y505H, S:T547K, and S:N764K substitutions in the B.1.1.529 (Omicron) lineage; and NSP3:A488S, NSP12:P227S, and S:P77L substitutions in the B.1.617.2 (Delta) lineage.

The amino acid composition of proteomes reflects the action of natural selection to enhance metabolic efficiency, synonymous codon usage bias as a measure of translation rates, and increases in the abundance of less energetically costly amino acids in highly expressed proteins [24]. The codon usage data compiled for different lineages of SARS-CoV-2 are presented in Supplementary Table 1. We summarised the synonymous codon usage of all different lineages of SARS-CoV-2 in Fig. 2c and Supplementary Table 1. In the B.1.640.2 lineage, codon AGU was biased toward Ser, codon UUU was stronger toward Phe in the B.1.640.1 and P.3 (theta) lineages, and codon GCU was biased toward Ala in the B.1.617.2 (delta) lineage. Here, we summarise the mutation rates across closely related species over short time-scales in a large, effective population. Thus, a large selective difference between the lineages of SARS-CoV-2 is required for a strong codon usage bias.

2.3. Clock-like prediction of focal outbreak points worldwide to provide warnings

To explain the epidemic outbreak points related to key mutations, we set up an EpiClock device to predict potential epidemics and assist in the presentation of detailed mutation information for severely affected areas. Therefore, we analysed the entire evolutionary pathway in the timeline of lineages of SARS-CoV-2 presented in Supplementary Table 2, where each lineage shows information for the earliest publicly collected samples in different regions of the world. Simultaneously, we summarise the smoothed distribution of new cases per million cases of SARS-CoV-2 in different severely affected areas, such as Africa, Asia, Europe, North America, Oceania, and South America, along the timeline from the OWID [25,26] in Supplementary Figs. 18–24. We defined amino acid substitutes according to the latest data reported around Jun 9th, 2022.Notably, we found an increasing number of other types of substitutions (indicated by the red asterisk) as the hitchhiking effect progressed; that is, a few other types increased dominantly, especially in the pre-breakout phase, even though some substitutions were replaced by other dominant genotypes (shown with the red box) in Fig. 3a and. b, Supplementary Table 2, and Supplementary Figs. 25–32.

We hypothesised that specific amino acid substitutions triggered these outbreaks. To predict potential epidemic outbreaks, we proposed the ZHU algorithm presented in Fig. 4a and tested it on different true sets of SARS-CoV-2 data before outbreaks to search for significant mutational accumulation patterns related to outbreak events. We found 171 statistically significant substitutions (significance level, $p < 0.05$) as potential epifactors in 55 countries and regions. We proposed the ZHU prediction which was similar to “China’s abacus”, to perform the cycling of N generations of training by GLM and reordering according to AIC (The Akaike Information Criterion). Finally, we performed ZHU prediction based on the weighted intercept estimates provided by the supporting information of the true sets and accessed 42 validation sets with positive precision, sensitivity, and accuracy as follows.

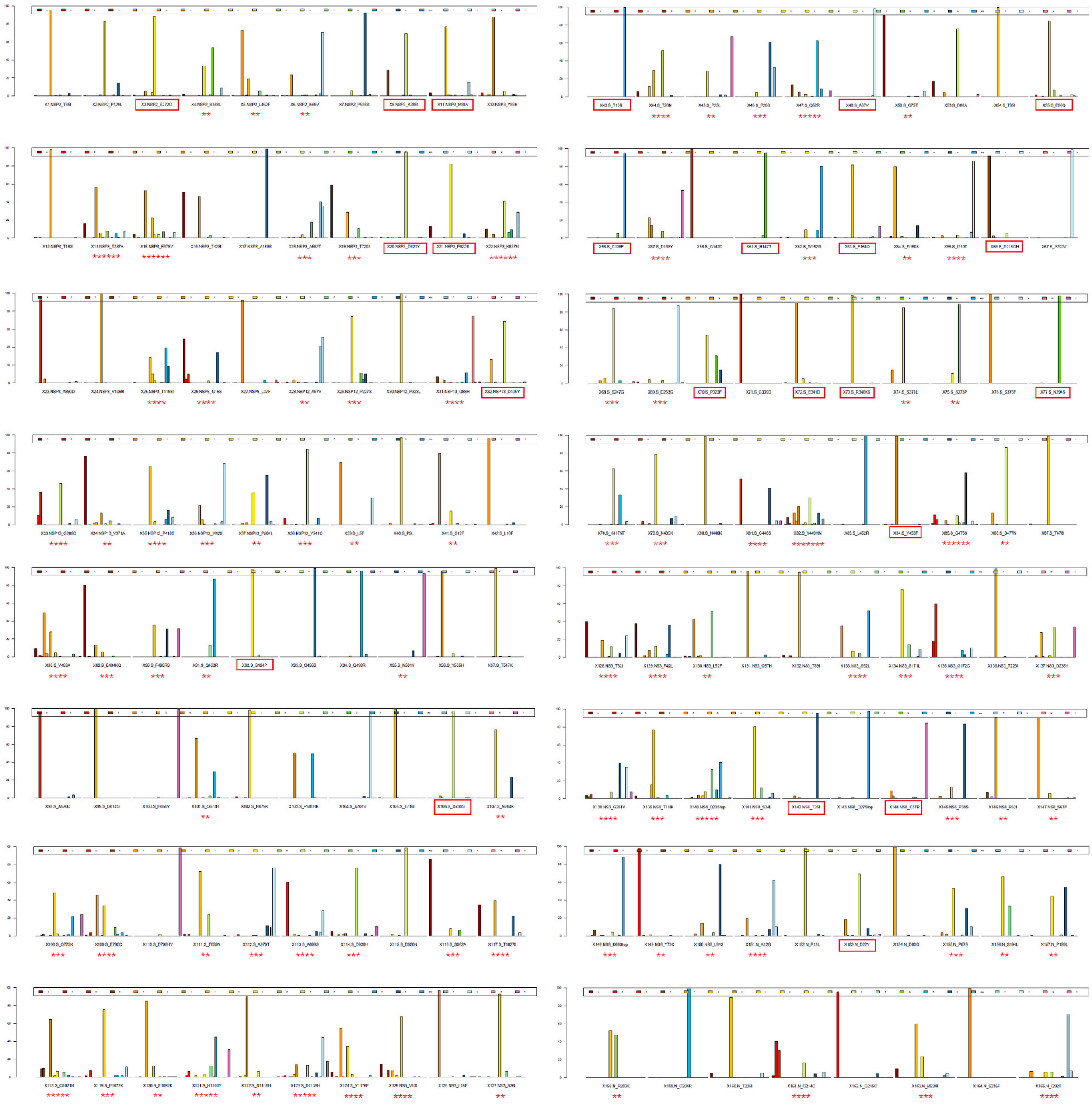
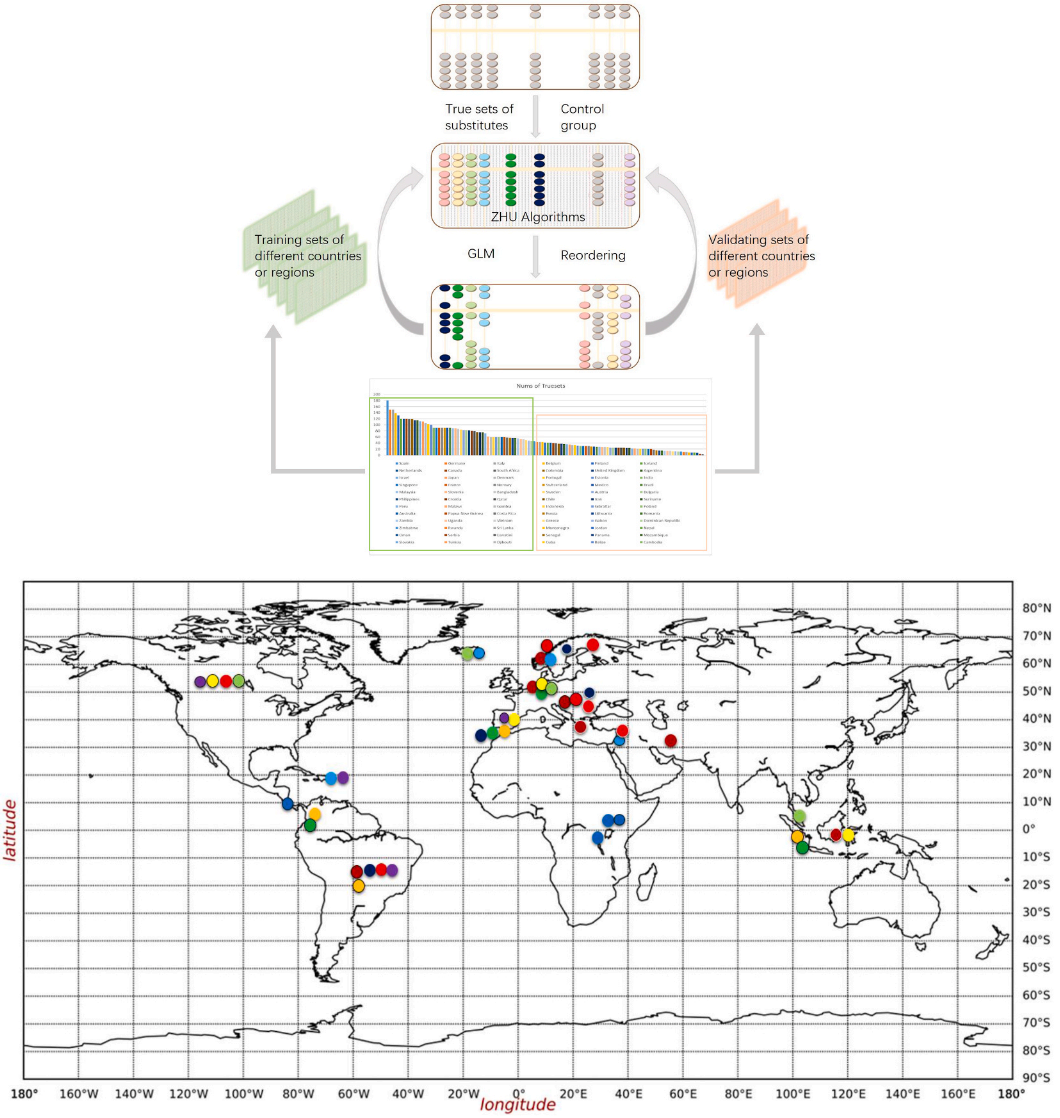


Fig. 3. Illustration of mutation information for different lineages of SARS-CoV-2 in severely affected areas. 3a. Frequency distribution of types of amino acid substitutions 1–87 in the pre-breakout phase. **3b.** Frequency distribution of types of amino acid substitutions 88–165 in the pre-breakout phase. Asterisks represent the pattern of an increasing number of other types of substitutions as the hitchhiking effect progressed, and red boxes show that some substitutions were replaced by other dominant genotypes recently.



●	X13.NSP3_T183I.I	2020/03 - 2021/08 Iran: 352.54	2020/03 - 2021/11 Netherlands: 817.422	2020/10 - 2021/09 Norway: 41.525	●	X69.S_D253G.G	2020/10 - 2021/09 Norway: 67.733	2021/08 - 2022/02 Slovenia: 387
●	X74.S_S371L.L	2020/09 - 2022/01 Finland: 65.9662	2020/03 - 2022/01 Canada: 206.464	2020/06 - 2022/01 Brazil: 221.244	●	X89.S_E484KQ.V	2020/03 - 2022/02 Singapore: 361.997	2020/06 - 2022/01 Brazil: 1104.313
●	X11.NSP3_M84V.I	2020/12 - 2021/12 Gibraltar: 5590.54	2020/12 - 2022/01 Colombia: 63.5		●	X101.S_Q677H.H	2020/03 - 2021/11 Netherlands: 1051.28	2020/03 - 2022/01 Canada: 21.4259
●	X29.NSP12_P227S.P	2021/01 - 2021/07 Indonesia: 2.9861	2020/03 - 2022/01 Spain: 250.599		●	X101.S_Q677H.Q	2020/03 - 2021/11 Netherlands: 1099.87	2020/03 - 2022/01 Canada: 24.4729
●	X31.NSP13_Q88H.Q	2021/01 - 2021/08 Malaysia: 1998.7116	2020/03 - 2021/08 Iceland: 205.8784		●	X110.S_D796HY.D	2020/03 - 2022/02 Singapore: 940.076	2020/12 - 2022/01 Colombia: 41.1
●	X42.S_L18F.F	2020/12 - 2021/12 Gibraltar: 546.44	2020/03 - 2021/11 Netherlands: 89.456		●	X129.NS3_P42L.P	2020/03 - 2022/01 Israel: 1200	2020/03 - 2021/08 Iceland: 1293.6818
●	X53.S_D80A.D	2020/10 - 2021/09 Norway: 5.148	2020/12 - 2022/01 Dominican Republic: 10.58		●	X56.S_C136F.C	2021/05 - 2022/01 Uganda: 56.381	2021/04 - 2022/01 Costa Rica: 701.2643
●	X56.S_C136F.stop	2021/06 - 2022/01 Rwanda: 30.347	2021/05 - 2022/01 Uganda: 56.452		●	X76.S_S375F.S	2020/11 - 2021/11 Slovakia: 1857.245	2020/05 - 2022/01 Sweden: 1290
●	X62.S_W152R.W	2020/12 - 2021/12 Gibraltar: 1272.88	2020/06 - 2022/01 Brazil: 276.801		●	X80.S_N440K.K	2020/03 - 2022/01 Spain: 292.031	2020/03 - 2022/01 Canada: 117.2041
●	X68.S_S247G.S	2020/12 - 2022/01 Dominican Republic: 113.174	2020/06 - 2022/01 Brazil: 306.985		●	X81.S_G446S.S	2021/01 - 2021/07 Indonesia: 113.0921	2020/10 - 2022/01 Greece: 3060.4
●	X69.S_D253G.D	2021/08 - 2022/02 Slovenia: 245	2020/06 - 2022/01 Brazil: 28.109		●	X86.S_S477N.N	2020/10 - 2021/12 Jordan: 140.499703	2020/07 - 2020/12 Serbia: 316.05

Fig. 4. Algorithm and performance of Epi-Clock. 4a. ZHU algorithm for inferring the timing of an epidemic of a new lineage related to increasing confirmed cases. **4b.** Prediction of 22 significant substitutions as the potential trigger for the outbreak from 75 training sets. **4c.** Performance of ZHU prediction measured by the relationship between observed confirmed numbers of new cases per million and normalised predicted numbers in different regions. **4d.** Performance of Epi-Clock evaluated using 42 validation sets.

However, our prediction of an outbreak was based only on significant amino acid substitutes and excluded other epi-factors. Notably, amino acid substitution type X13.NSP3_T183I.I, is a potential epifactor in Iran, the Netherlands, and Norway. X74. S_S371L.L significantly correlated with the number of new cases in Finland, Canada, and Brazil (Fig. 4b). Across N generations of training using GLM and reordering, we identified 171 significant substitutions as potential epi-factors within 55 different countries and regions (Supplementary Table 3). Finally, we successfully demonstrate the performance of the ZHU algorithm on 42 validation sets (Fig. 4c). From most validations, we accurately predicted the potential pre-phase of the outbreak using the ZHU prediction. By counting the number of true instances, we summarised the positive precision, sensitivity, and accuracy of the Epi-Clock with the true validation sets, as presented in Fig. 4d and Supplementary Table 3, where the median interval before the outbreaks was five days.

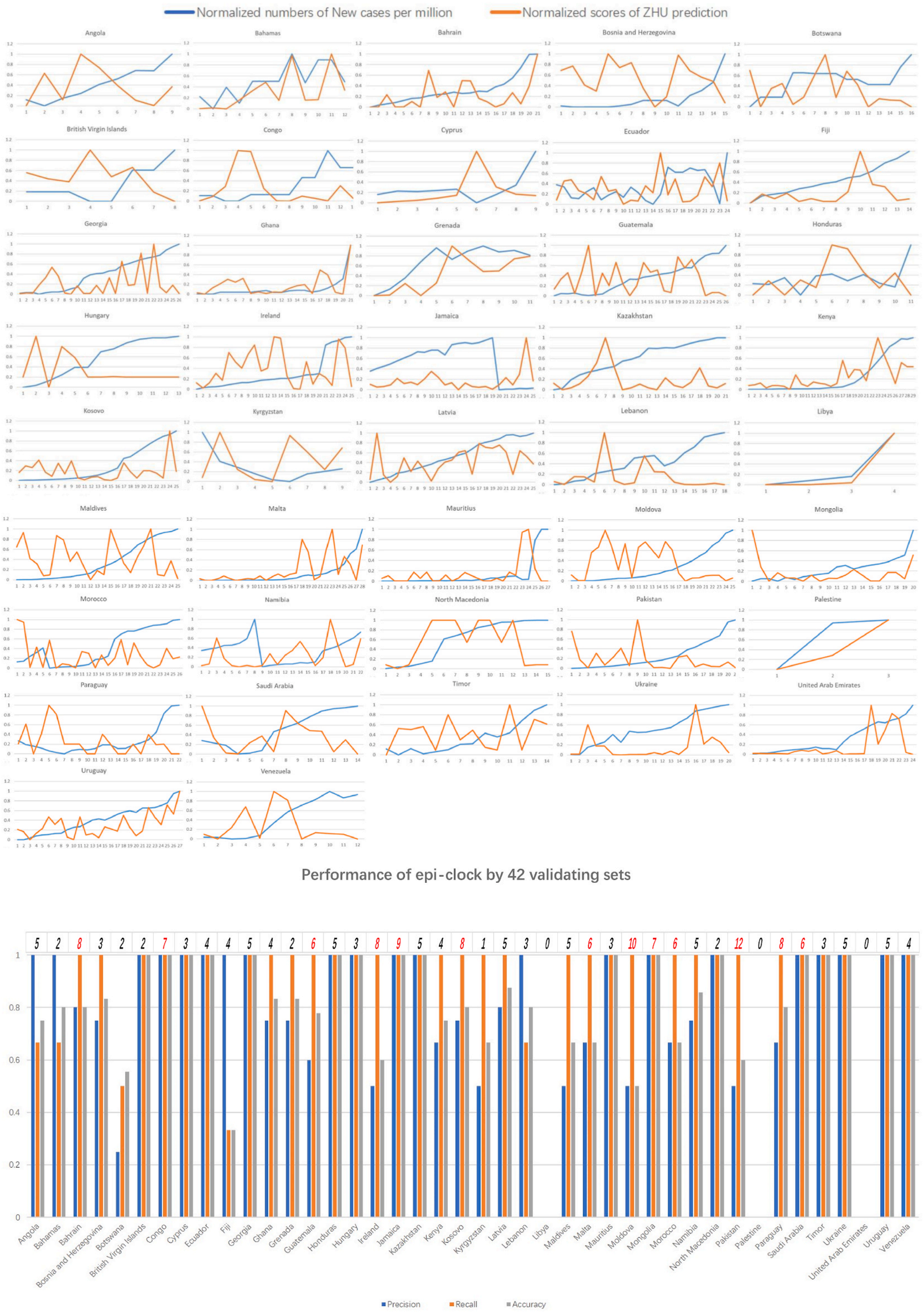


Fig. 4. (*continued*).

3. Discussion

We found significant mutational accumulation based on the frequency distribution of amino acid substitution types among different lineages of SARS-CoV-2 in severely affected areas. A large epidemic is a heterogeneous and spatially dissociated collection of transmission clusters of varying sizes, durations and connectivities [27]. Several studies have proposed the epidemiological principles of pathogens, such as two co-circulating lineages of influenza B virus [28], the West African Ebola outbreak [22], and repeatable/predictable parallel adaptation, including cross-species transmission, drug resistance, and host immune escape [29,30]. In particular, omicron strains are the turning point of the SARS-CoV-2 epidemic with strong infectivity, which provides a great opportunity for viral adaptive evolution within large scales of individual populations. It represents diverse amino acid substitutes at different sites, achieved by different ages, species, or physiological environments. Except for mutational shifts among different hosts, the pathogenicity and fatality rates gradually decreased. Here, we could only accurately predict the potential pre-phase of the outbreak by ZHU prediction, where the median interval before the outbreaks was 5 days. Nevertheless, it is difficult to ascertain the significance of this short prediction timeframe to take necessary precautionary measures to prevent future outbreaks. In only four countries, Jamaica, Kyrgyzstan, Libya, and Palestine, our model could not be fitted because of unavailable sequencing samples in the outbreak phase. As distinguished by EpiFactors [18], which include 815 proteins with 95 histones and protamines involved in epigenetic regulation, we extracted 171 significant amino acid substitutions as potential epifactors within 55 different countries and regions (Supplementary Table 3). We believe that it will never repeat severe epidemic clinical events in a few years, because SARS-CoV-2 has been completely adaptive in the human body. The strict and comprehensive pandemic control strategies implemented in Shanghai were able to reduce the number of people infected so that the case fatality rate could be minimised and buy time for full vaccination coverage [31].

Here, we summarised the amino acid substitutions among different lineages to determine the codon usage bias according to relative synonymous codon usage, involved with the dated variants of different species or subspecies, and the unique and specific amino acid substitutions in one lineage. The complexity of many host interactions broadens the definition of a pathogen's immunological niche [32]. From the host transmission and shift of Coronaviridae shown in Supplementary Figs. 3–10 and Supplementary Table 1, we illustrated the insertions and deletions of whole genome sequences between different hosts and codon usage bias among different lineages. It has been proposed that the richness of wildlife host species is an important predictor of disease emergence. Similarly, host populations with low biodiversity might have an increased risk of emergence. Conversely, high host biodiversity has been linked to a 'dilution effect' with a decrease in disease risk [8]. How do influenza viruses evolve in their human hosts? This could be because of factors such as antigenic selection, antiviral treatment, tissue specificity, spatial structure, and multiplicity of infection [33]. Host switching often leads to viral emergence to overcome barriers to infection in a new host [10]. Host gene editing, the major source of existing SARS-CoV-2 mutations [34], results in a higher rate of severe outcomes and considerable mortality in unvaccinated individuals, especially in older adults. How does natural selection shape immunity and host defence genes [35]? Natural selection causes microevolutionary changes that increase fitness, whereas random gene drift is strongly linked to the gene size fixed by the population [36,37]. Long-term co-speciation [38], host range properties [8], and single-cell technologies [39,40] could explore the properties of host cells harbouring infection, the host pathogen-specific immune responses, and the mechanisms by which pathogens have evolved to escape host control.

Our study has some limitations in exploring future work. Pathogens have always imposed strong selection pressure on the human genome [35]. This might be expected to follow viral emergence in a new host species because positive selection would entail a major boost in the number of susceptible hosts and a concomitant increase in fitness [41]. We demonstrated the richness distribution of different mutation types across the whole genomes, in which mutation rates of C- > T, G- > T and T- > C were dominant in driving the divergence of different lineages. Similar to the reported vRNP structures, phosphorylation of the N protein in its disordered serine/arginine region weakens these interactions, generating less compact vRNPs to support other N protein functions in viral transcription [34]. All populations showed evidence of insertions, deletions, or substitutions that have driven the divergence of the entire family of *Coronaviridae* in response to natural selection, random genetic drift, host gene editing, and viral proofreading [37]. It suggested that these probably represent another independent acquisition of new "functional sequences" through either specific horizontal gene transfer or recombination events [42–44]. The value of using age information can be used to interpret variants of functional and selective importance, such as allele age estimates, to infer ancestry shared between individual genomes [4]. However, we could not demonstrate whether codon usage bias contributed to natural selection or genetic drift. In particular, closely related species do not usually exhibit major shifts in codon preferences; however, changes in mutation rates over short timescales are common in large effective population sizes [45]. Transgenes are commonly designed to increase gene expression levels through codon optimisation. Only in large effective population sizes is the selection of codon usage strong among species. With a reduction in the effective population size, the codon bias declines. This long-term reduction has led to a major shift in genome evolution. Until the effective population size is small, genetic drift becomes dominant over natural selection [45].

4. Conclusions

We developed a platform for predicting future pathogenic disease outbreaks. With the accumulation of sequencing datasets, the performance of our strategy will improve and the approach will become more sensitive in its ability to predict the potential triggers of epidemic outbreaks and describe the spatiotemporal and geographical mutational landscape of the pathogens, with special emphasis on SARS-CoV-2, which will ultimately facilitate responses to future outbreaks of concern. We believe that the host's non-nucleosomal DNA played a key role in the evolutionary divergence and emergence of the new strains responsible for the outbreak.

5. Materials and methods

To explore the divergence of the whole family of *Coronaviridae*, we computed the intra- and inter-species p-distances of whole genome sequences using MEGA11 and plotted the distances using the boxplot function of R [46]. To illustrate the atlas of new ages within different populations, we applied a genome alignment-based pipeline to infer the origin time of a given genomic region using a 6 bp sliding window with the numbers of dated variants, including insertions in red and deletions in blue. We scanned the whole genome sequences with sliding windows and summarised the mean values of the mutation rates for different mutation types. Based on the reference genomes of different hosts, we demonstrated host shifts within sliding windows in the evolution of *Coronaviridae* and presented them using Circos.

Similarly, we identified nucleotide and amino acid substitutions in different SARS-CoV-2 lineages. We computed mutation rates using sliding windows for different mutation types. We presented the 144 amino acid substitution distributions of the different lineages using GeneWise [47]. The codon usage numbers were converted into relative synonymous-codon usage (RSCU) values [28], which were simply the observed frequency of a codon divided by the expected frequency under the assumption of equal usage of synonymous codons for an amino acid [28,48], where i is the specific number of amino acids, j is the specific number of codons, xij is the observed number of the j-th codon type for the i-th amino acid, and ni is the number of alternative codon types for the i-th amino acid.

$$\mathrm{RSCU_{ij}} = \frac{\mathrm{Xij}}{1/\mathrm{ni}} \sum_{j=1}^{ni} Xij$$

A neighbour-joining phylogenetic tree of separate lineages of SARS-CoV-2 was constructed using MEGA v11.0.11. To predict potential epidemic mutation patterns in severely affected areas, we summarised testing data from Our World in Data (OWID) until Feb 7th, 2022 that were continually collated from official government sources worldwide. We plotted the frequency distribution of confirmed cases of SARS-CoV-2 in different severely affected areas, including Africa, Asia, Europe, North America, Oceania, and South America, along a timeline. Simultaneously, we set the baseline of new cases per million substitutions as the internal control group and widespread types of substitutions as the external control group to exclude other effects of epidemic outbreak events. We excluded areas without any outbreak time, as labelled by the red box, and searched for mutation patterns that appeared to be related to the outbreak until the appearance of the Omicron variant because this corresponded with the time of universal immunisation [49,50]. To avoid system errors, that is, errors caused by regions varying in other epi-factors, such as population density, geographical environment, vaccination coverage rates, and national/social rules and norms, we extracted the true sets by the control group in different individual countries or regions as individual-related values of new cases per million. To ensure the true values of the outbreak of new SARS-CoV-2 cases, we carefully detected peaks following strict standards; that is, the values of new cases per million had to be above 30. Then, we separately split the sets of population samples from 1 to 30 days before the outbreaks based on the same location according to 144 amino acid substitutions or deletions (Fig. 2b and Supplementary Table 2). The dataset contained 13,740,300 observations, including 6300 true sets and 2181 features, which amounted to 117 different countries or regions on all five continents, divided into the 75 training sets and 42 validation sets, as listed in Supplementary Table 3. We then separately performed a generalised linear model (GLM) analysis on individual countries/regions to determine the optimal mutational patterns, as in the training phase presented in Supplementary Table 3.

```
glm(formula, family = gaussian, data, weights, subset,
na.action, start = NULL, etastart, mustart, offset,
control = list( …), model = TRUE, method = "glm.fit",
x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, …)
```

We found 171 significant substitutions ($p < 0.05$) as potential epifactors in 55 countries and regions. We proposed the ZHU prediction, similar to “China’s abacus”, to perform the cycling of N generations of training by GLM and reordering according to The Akaike Information Criterion (AIC). Finally, we performed ZHU prediction based on the weighted intercept estimates provided by the supporting information of the true sets and accessed 42 validation sets with positive precision, sensitivity, and accuracy as follows.

$$\text{Positive precision} = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

$$\text{Sensitivity} = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

$$\text{Accuracy} = \frac{TP + TN}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Ethics approval and consent to participate

All protocols were approved by the Liferiver Science and Technology Institute, Shanghai ZJ Bio-Tech Co., Ltd. and the Use Committee (Shanghai, China).

Data availability statement

All shared mutations and substitutes from different hosts, lineages or regions are available in the Supplementary Materials and https://bioinfo.liferiver.com.cn/#/home. There we also provided the source data and technological method descriptions on this manuscript.

Funding statement

Thanks for the financial support provided by the Liferiver Science and Technology Institute of Shanghai ZJ Bio-Tech Co, Ltd for this research.

CRediT authorship contribution statement

Cong Ji: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Junbin (Jack) Shao:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Cong Ji reports financial support was provided by Liferiver Science and Technology Institute. Junbin Jack Shao reports a relationship with Liferiver Science and Technology Institute that includes: board membership. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Han Chen and Guangzhong Wang for their valuable suggestions and comments. Thank you for your discussion of Si Chen, Yan Liu, and Hanyan Zhang. Many facets of the user-interface design benefited from the help of Xingsheng Niu, Yajie Pan, and Wang Lu. We thank Elizabeth Sung and Elsevier Language Editing Services for the flow and language corrections. All other data supporting the findings of this study and the computational code used in this study are available from the corresponding author upon request. The authors declare no conflicts of interest.

List of abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
vRNP	Viral ribonucleoprotein
RBD	Receptor-binding domain
NSP3	non-structural protein 3
EBOV	Ebola virus
GP	Glycoprotein
NCLDV	Nucleocytoplasmic large DNA virus
MVs	*Marseilleviridae*
DG	D614G substitute
GLM	Generalised linear model
AIC	The Akaike Information Criterion
MEGA11	Molecular Evolutionary Genetics Analysis
RSCU	Relative synonymous-codon usage
OWID	Our World in Data
GLM	Generalised linear model
TP	True positive
FP	False positive
FN	False negative
TN	True negative
SARS-CoV	Severe acute respiratory syndrome-related coronavirus
SADS	Swine acute diarrhoea syndrome coronavirus
NL63	Human coronavirus NL63
MERS	Middle East respiratory syndrome-related coronavirus

London1	Betacoronavirus England 1
HKU5	Bat coronavirus HKU5
HKU4	Bat coronavirus HKU4
HKU3	Bat coronavirus HKU3
HKU2	Bat coronavirus HKU2
BATS	Bat SARS-like coronavirus WIV1
OC43	Human coronavirus OC43
HKU9	Rousettus bat coronavirus HKU9
HKU1	Human coronavirus HKU1
229E	Human coronavirus 229E
ORF	Open reading frame
NSP	Non-structural protein
S	Surface glycoprotein
E	Envelope protein
M	Membrane glycoprotein
N	Nucleocapsid phosphoprotein
HIV	Human immunodeficiency virus
H1N1pdm09	2009H1N1 Pandemic

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e36162.

References

[1] P. Forterre, D. Prangishvili, The origin of viruses, Res. Microbiol. 160 (7) (2009) 466–472.

[2] A. Kitchen, L.A. Shackelton, E.C. Holmes, Family level phylogenies reveal modes of macroevolution in RNA viruses, Proceedings of the National Academy of Sciences of the United States of America 108 (1) (2011) 238–243.

[3] M.C. Vieira, D. Zinder, S. Cobey, Selection and neutral mutations drive pervasive mutability losses in long-lived anti-HIV B-cell lineages, Mol. Biol. Evol. 35 (5) (2018) 1135–1146.

[4] P.K. Albers, G. McVean, Dating genomic variants and shared ancestry in population-scale sequencing data, PLoS Biol. 18 (1) (2020) e3000586.

[5] P. Simmonds, P. Aiewsakun, Virus classification - where do you draw the line? Arch. Virol. 163 (8) (2018) 2037–2046.

[6] S. Wohl, S.F. Schaffner, P.C. Sabeti, Genomic analysis of viral outbreaks, Annual review of virology 3 (1) (2016) 173–195.

[7] B. Longdon, et al., The evolution and genetics of virus host shifts, PLoS Pathog. 10 (11) (2014) e1004395.

[8] J.L. Geoghegan, E.C. Holmes, Predicting virus emergence amid evolutionary noise, Open biology 7 (10) (2017) 170189.

[9] G.L. Kaján, et al., Virus–host coevolution with a focus on animal and human DNA viruses, J. Mol. Evol. 88 (1) (2020) 41–56.

[10] C.R. Parrish, et al., Cross-species virus transmission and the emergence of new epidemic diseases, Microbiol. Mol. Biol. Rev. 72 (3) (2008) 457–470.

[11] D.J. Park, et al., Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone, Cell 161 (7) (2015) 1516–1526.

[12] R.A. Urbanowicz, et al., Human adaptation of Ebola virus during the West African outbreak, Cell 167 (4) (2016) 1079–1087.e5.

[13] X. Chen, et al., Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes, Science 335 (6073) (2012) 1235–1238.

[14] Y. Liu, et al., Virus-encoded histone doublets are essential and form nucleosome-like structures, Cell 184 (16) (2021) 4237–4250 e19.

[15] A. Vannini, I. Marazzi, A small nucleosome from a weird virus with a fat genome, Mol Cell 81 (17) (2021) 3447–3448.

[16] M. Tizzoni, et al., Real-time numerical forecast of global epidemic spreading: case study of 2009 A/H1N1pdm, BMC Med. 10 (2012), pp. 165–165.

[17] Y.A. Medvedeva, et al., EpiFactors: a comprehensive database of human epigenetic factors and complexes, Database (bav067) (2015).

[18] P. Forster, et al., Phylogenetic network analysis of SARS-CoV-2 genomes, Proc. Natl. Acad. Sci. USA 117 (17) (2020) 9241.

[19] B. Meng, et al., Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7, Cell Rep. 35 (13) (2021).

[20] I. Kimura, et al., The SARS-CoV-2 Lambda variant exhibits enhanced infectivity and immune resistance, Cell Rep. 38 (2) (2022).

[21] Y. Ruan, et al., The twin-beginnings of COVID-19 in Asia and Europe-one prevails quickly, Natl. Sci. Rev. 9 (4) (2022) nwab223.

[22] N.D. Grubaugh, et al., Tracking virus outbreaks in the twenty-first century, Nature Microbiology 4 (2019) 10–19.

[23] S. Kumar, , et al.X. Mega, Molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol. 35 (6) (2018) 1547–1549.

[24] Z. Yang, et al., Codon-substitution models for heterogeneous selection pressure at amino acid sites, Genetics 155 (1) (2000) 431–449.

[25] J. Hasell, et al., A cross-country database of COVID-19 testing, Sci. Data 7 (1) (2020) 345.

[26] E. Mathieu, et al., A global database of COVID-19 vaccinations, Nat. Human Behav. 5 (7) (2021) 947–953.

[27] G. Dudas, et al., Virus genomes reveal factors that spread and sustained the Ebola epidemic, Nature 544 (7650) (2017) 309–315.

[28] D. Vijaykrishna, et al., The contrasting phylodynamics of human influenza B viruses, Elife 4 (2015) e05055.

[29] B. Gutierrez, M. Escalera-Zamudio, O.G. Pybus, Parallel molecular evolution and adaptation in viruses, Curr Opin Virol 34 (2019) 90–96.

[30] X. Han, et al., SARS-CoV-2 nucleic acid testing is China's key pillar of COVID-19 containment, Lancet 399 (10336) (2022) 1690–1691.

[31] X. Zhang, W. Zhang, S. Chen, Shanghai's life-saving efforts against the current omicron wave of the COVID-19 pandemic, Lancet 399 (10340) (2022) 2011–2012.

[32] S. Cobey, Pathogen evolution and the immunological niche, Ann. N. Y. Acad. Sci. 1320 (1) (2014) 1–15.

[33] K.S. Xue, et al., Within-host evolution of human influenza virus, Trends Microbiol. 26 (9) (2018) 781–793.

[34] C.R. Carlson, et al., Reconstitution of the SARS-CoV-2 ribonucleosome provides insights into genomic RNA packaging and regulation by phosphorylation, J. Biol. Chem. 298 (11) (2022) 102560.

[35] L.B. Barreiro, L. Quintana-Murci, From evolutionary genetics to human immunology: how selection shapes host defence genes, Nat. Rev. Genet. 11 (1) (2010) 17–30.

[36] A.L. Vargas-Aguilar, et al., Genomic and molecular evolutionary dynamics of transcriptional response regulator genes in bacterial species of the Harveyi clade of Vibrio, Gene 783 (2021) 145577.

[37] R. Wang, et al., Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries, Genomics 113 (4) (2021) 2158–2170.

[38] D.M. de Vienne, et al., Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution, New Phytol. 198 (2) (2013) 347–385.
[39] P.K. Chattopadhyay, M. Roederer, D.L. Bolton, A deadly dance: the choreography of host–pathogen interactions, as revealed by single-cell technologies, Nat. Commun. 9 (1) (2018) 4638.
[40] S.A.-O. Kazer, et al., Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection, Nat. Med. 26 (4) (2020) 511–518.
[41] E.C. Holmes, The evolution of viral emergence, Proceedings of the National Academy of Sciences of the United States of America 103 (13) (2006) 4803.
[42] D. Forni, et al., Homology-based classification of accessory proteins in coronavirus genomes uncovers extremely dynamic evolution of gene content, Mol. Ecol. 31 (13) (2022) 3672–3692.
[43] V. Makarenkov, et al., Horizontal gene transfer and recombination analysis of SARS-CoV-2 genes helps discover its close relatives and shed light on its origin, BMC Ecol Evol 21 (1) (2021) 5.
[44] A. Shukla, R. Hilgenfeld, Acquisition of new protein domains by coronaviruses: analysis of overlapping genes coding for proteins N and 9b in SARS coronavirus, Virus Gene. 50 (1) (2015) 29–38.
[45] S.T. Parvathy, V. Udayasuriyan, V. Bhadana, Codon usage bias, Mol. Biol. Rep. 49 (1) (2022) 539–565.
[46] R.C. Team, R: a language and environment for statistical computing, MSOR connections 1 (2014).
[47] E. Birney, M. Clamp, R. Durbin, GeneWise and genomewise, Genome Res. 14 (5) (2004) 988–995.
[48] W.H. Sharp Pm Fau - Li, W.H. Li, An evolutionary perspective on synonymous codon usage in unicellular organisms, J. Mol. Evol. 24 (1–2) (1986) 28–38.
[49] S. Chowdhury, et al., Omicron variant of SARS-CoV-2 infection elicits cross-protective immunity in people who received boosters or infected with variant strains, Int. J. Immunopathol. Pharmacol. 36 (2022) 3946320221133001.
[50] D. Zarębska-Michaluk, et al., COVID-19 vaccine booster strategies for omicron SARS-CoV-2 variant: effectiveness and future prospects, Vaccines (Basel) 10 (8) (2022).