PROPOSED TITLE: **Early evolution of the BA.2.86 variant sheds light on the origins of highly divergent SARS-CoV-2 lineages**

AUTHORS:
Marina Escalera-Zamudio[1,*], Cedric C S Tan[1,*], Lucy van Dorp[1,*], François Balloux[1,*]

[1]ORCID: 0000-0002-4773-2773
m.zamudio@ucl.ac.uk
[2]ORCID: 0000-0003-3536-8465
cedric.tan.18@ucl.ac.uk
[3]ORCID: 0000-0002-6211-2310
lucy.dorp.12@ucl.ac.uk
[4]ORCID: 0000-0003-1978-7715
f.balloux@ucl.ac.uk

AFFILIATIONS:
[1]UCL Genetics Institute, Department of Genetics, Evolution & Environment, University College London, London, UK, WC1E 6BT
*END-VOC Consortium

ABSTRACT (253)
Throughout the evolutionary history of SARS-CoV-2, divergent variants with an unusually high number of novel mutations have repeatedly emerged and displaced other lineages in co-circulation. The latest example is the BA.2.86 Variant Of Interest (VOI), which rapidly outcompeted all other lineages following its detection in late July 2023. Descending sublineages dominated globally by July 2024, at the time of writing this study. The mechanisms by which highly-divergent viral lineages tend to emerge remain unclear, with the most strongly supported hypothesis being the acquisition of multiple mutations through accelerated evolution in chronically infected individuals, with spread back into the community facilitated by prolonged shedding. We used BA.2.86 as a case study to investigate the evolutionary process behind the emergence of a highly divergent SARS-CoV-2 variant. Through fine-scale evolutionary analyses applied to four independently subsampled datasets, we identify over 100 genomes representing evolutionary intermediates occupying the branch in the phylogenetic tree leading to the BA.2.86*/JN.1* clade. Additionally, we detected more than 2,600 'hybrid' genomes across the evolutionary history of the Omicron-descending BA.2* sub-lineages, displaying a sequential acquisition of BA.2.86 lineage-defining mutation constellations through time. We observe that hybrid strains circulated up to 19 months prior to BA.2.86, denoting cryptic circulation in various countries, with further evidence of early transmission chains. We also found signals of recombination between pre-existing parental lineages, suggesting that this process contributed to the emergence of BA.2.86. Our results suggest that the BA.2.86 lineage emerged sequentially through a complex evolutionary process, rather than via accelerated evolution associated with a single emergence event.

MAIN TEXT WORD COUNT: 5692

INTRODUCTION (word count: 832)

The evolution of SARS-CoV-2 has been characterized by the periodic emergence and rapid spread of highly divergent viral lineages or variants. Such variants often replace previously dominant lineage(s) in circulation, likely due to the acquisition of several mutations that conferred an enhanced fitness either through more transmissible and/or immune evasive phenotypes [1]. Numerous schemes have been devised to classify such variants, including novel virus lineage designations under the 'Pango' nomenclature system [2]. The World Health Organisation WHO also devised a criteria for classifying highly divergent lineages with potential epidemiological relevance as Variants of Interest (VOI), Variants under Monitoring (VUM), and/or Variants of Concern (VOC) [3]. Elucidating the evolutionary processes which led to the emergence of highly divergent lineages (regardless of the 'variant monitoring status') is not only relevant for a retrospective understanding of the development of the SARS-CoV-2 pandemic, but can also inform decisions on future surveillance and control measures.

Several hypotheses have been put forward to explain the evolutionary origin of highly divergent SARS-CoV-2 lineages, deriving from a single or a combination of these scenarios: 1) chronic infections in immunocompromised individuals, where the virus linked to a single infection case is more likely to acquire multiple changes in a short period of time due to dampened host immune responses, 2) sustained circulation in areas with poor genomic surveillance, where the virus gradually accumulates changes that go undetected until reaching a region with stronger surveillance, and 3) zoonotic circulation in animal reservoirs, where changes occur in undetected virus subpopulations circulating in animals before spilling back into humans (extensively reviewed in [4]). The most widely accepted hypothesis is accelerated evolution in immunocompromised individuals [5,6]. However, the underlying evolutionary processes compatible with these hypotheses, such as a sequential accumulation of mutations and/or recombination, are often suggested as potential contributors but are typically not formally investigated. Thus, the evolutionary mechanisms driving the emergence of highly divergent SARS-CoV-2 lineages remains only partially addressed.

With over 50 mutations relative to its closest genetic background, the B.1.1.529 lineage ('Omicron' VOC) [3] was first detected in January 2022 [7]and rapidly dominated across the world. Following a period of gradual evolution, the detection of the directly descending BA.2.86 sublineage in late July 2023, marked the first recorded instance of apparent accelerated evolution since the detection of B.1.1.529 [8]. BA.2.86 was first identified in Denmark, UK, USA, and Israel [8,9], and rapidly spread globally to be recognized as a VOI in November 2023 [10].

Relative to its direct parental lineage (BA.2, also B.1.1.529-descending), BA.2.86 carries multiple non-synonymous mutations, with over 30 of these occurring in the Spike protein alone [9]. As with other highly divergent lineages, the BA.2.86 lineage forms a monophyletic clade separated by a long branch, diverging from its most closely related ancestral lineages within the global SARS-CoV-2 reference phylogeny [11,12].

Since its emergence, lineage BA.2.86 has further diversified giving rise to the JN.1 sublineage, defined by an additional single non-synonymous mutation L455S within the Spike protein (and three additional synonymous mutations in other genes) [13]. The ongoing rapid evolution of the JN.1* sublineages [13] is now evidenced by the recent emergence of the so-called "FLiRT variants" (designated as KP.*, direct descendants of JN.1), displaying convergent mutations at positions 456, 346, and 572 of the Spike protein. With increasing prevalence, FLiRT variants have dominated across the U.S. in July 2024, with KP.3 causing around 25% of newly sequenced cases over the last two weeks of June 2024, which may have in part been responsible for the infection wave in Summer 2024 [14].

Using the BA.2.86 VOI as a case study, we tested the hypothesis of accelerated change linked to a single emergence event driving the evolution of this highly divergent virus lineage. Such hypothesis would be falsified by detection of genomes denoting evolutionary intermediates collected at different timepoints and geographic locations, predating the detection of the BA.2.86 lineage. We applied multiple fine-scale evolutionary analyses to four datasets sequentially subsampled over time, comprising the BA.2.86*/JN.1* clade and the most closely related/parental lineages (BA.2, BA.2.75, and XBB.1.5). We recover multiple basally-positioned BA.2 genomes shortening the branch leading to the BA.2.86*/JN.1* clade, representing hybrid genomes that display some, but not all, BA.2.86-specific lineage defining mutations (LDMs). Some hybrid genomes correspond to precursors collected up to 19 months prior to the detection of BA.2.86 across multiple countries. Additionally, we found evidence for recombination between lineages BA.2 and BA.2.75 in the branch leading to the emergence of BA.2.86. Our results are incompatible with a scenario of accelerated evolution linked to a single emergence event, as we find evidence of multiple evolutionary processes (*i.e.*, gradual evolution during cryptic circulation and recombination between ancestral lineages) jointly contributing to the emergence of the BA.2.86 VOI. Our findings shed light on the overall evolutionary mechanisms that may drive the emergence of (at least some) highly divergent SARS-CoV-2 lineages, and raises questions over current best practice strategies in SARS-CoV-2 genomic surveillance, and in the initial interpretation of the phylogenetic placement of emerging SARS-CoV-2 lineages.

METHODS (word count: 1000)

In order to capture the evolution of the BA.2.86 lineage from its emergence following its diversification process as of May 2024, five different datasets (named here D1, D2, D3, D4 and D4.4) comprising complete SARS-CoV-2 genomes (low-coverage excluded, Ns ≤ 5%) were compiled from assemblies retrieved from GISAID [15-18]. Genomes within these datasets were collated at different time points and sequentially subsampled under a phylogenetically-informed approach. Further analyses included iterative phylogenetic inference, time-scaled analyses involving ancestral state reconstructions and node dating, as well as testing for recombination. Full details on the methodological approach and the tools used are available in the Supplementary Information (*Supplementary Text 1. Methods*).

Dataset acquisition and curation

D1 (n=100, compiled as of August 21st 2023), corresponds to the original phylogenetic characterisation of the BA.2.86 available from the Nextstrain platform as of July 30th 2023 [8,19]. D1 includes the eight earliest BA.2.86 genomes, and a subsample of 92 genomes assigned to the parental/most closely related virus lineages: BA.2, BA.2.75, and XBB.1.5 (with descending sublineages indicated with a *) (*Supplementary Data 1-Supplementary Table 1*). D1 was used to confirm the initial phylogenetic placement of BA.2.86 relative to its parental/most closely related virus lineages, and was further used to test for recombination (see Methods section *'Detection of Recombination'*).

D2 (n=9134, Including Wuhan-Hu-1 NC_045512 [Wu-1] used for rooting purposes) was compiled as of January 9th 2024 to consider a more representative sample reflecting the full diversity of the BA.2*, BA.2.75, and XBB.1.5* lineages (2,282 genomes, referred here as 'Background', labelled as BG on the tree). It also includes genomes representative of the expanding diversity of BA.2.86*/JN.1* within the first six months of circulation (referred here as 'Foreground', labelled as FG on the tree). To reduce computational costs for subsequent analyses, foreground genomes were subsampled from 35,902 to 6847 genomes (*Supplementary Data 1-Supplementary Table 2*) using a phylogenetically-informed approach.

D3 (n=7760) was compiled as of March 15th 2024, comprising an updated subsample of the foreground dataset, while balancing the number of sequences relative to the background genomes to avoid topological distortions. To further capture the unsampled diversity of virus genomes potentially representing evolutionary intermediates related to the foreground, we added a subsample of 2774 BA.2*, BA.2.75 and XBB.1.5* genomes subsampled from GISAID to include at least one BA.2.86-specific mutation at a nucleotide level (referred to here

112 'Mutated', labelled as MT on the tree) (*Supplementary Data 1-Supplementary Table 3*). All
113 genomes within D3 were re-assigned to their corresponding virus lineage prior to subsequent
114 analyses, ensuring a consistent and correct lineage assignation according to the 'Pango'
115 nomenclature (*Supplementary Data 3*). Sequences identified as 'misdated' outliers when
116 assessing root-to-tip divergence were also removed.
117
118 D4 (n=947, compiled as of May 22nd 2024) further reduces the size of D3, filtering out
119 sequences with incomplete collection dates and with an ambiguity content >0.1 (applicable to
120 'mutated' genomes only). This resulted in a streamlined dataset enabling subsequent
121 computationally costly analyses. Derived as a subsample of D4, D4.1 (n=100) comprises a
122 concatenated "Orf" only alignment, used for confirmatory testing applied to detection of
123 recombination (*Supplementary Data 1-Supplementary Table 4 and 5*).
124
125 *Alignment construction and iterative subsampling and phylogenetic inference*
126 All datasets were aligned and subsequently considered for phylogenetic inference under a
127 Maximum Likelihood (ML) framework, followed by phylogenetically-informed subsampling to
128 reduce phylogenetic redundancy while maintaining a representative diversity [20]. An iterative
129 subsampling/phylogenetic inference approach was employed until reaching computationally
130 manageable dataset sizes resulting in consistent phylogenies. Topological congruency (*i.e.,*
131 phylogenetic placement of the virus lineages of interest relative to the root of the tree) was
132 verified for all output trees through comparison with the global SARS-CoV-2 phylogeny
133 publicly available on the Nextstrain platform, updated between August 2023 and May 2024
134 [11,12].
135
136 *Reconstruction of Ancestral States and Time-Calibration*
137 The D3, D4 and D4.1 trees were time-calibrated using TreeTime [21], while BA.2.86-specific
138 LDMs defined as of August 2023 [9] were extracted at an amino acid level from the D2 and D3
139 alignments, coded as discrete traits, and mapped under reconstruction of ancestral states
140 onto nodes of the corresponding trees. A ML framework was employed to balance
141 computational costs relative to dataset size, and to facilitate further visualization. To limit
142 homoplasy (mutation co-occurrence across lineages linked to independent or convergent
143 origins) as a confounding factor impacting the identification of hybrid genomes, LDM co-
144 occurrence within hybrid genomes across the background was defined as at least (any) three
145 of the BA.2.86-specific LDMs. We further reconstructed the full mutational pathway derived
146 from the *in-frame* Orf1ab, Spike, M, and N alignments, (where LDMs occur) on the D4 tree.
147 For this, we extracted and mapped all amino acid states across alignment sites onto the
148 deeper nodes of the D4.1 tree. The definition of deep nodes within the tree considers all of

149 those occurring at the backbone of the phylogeny, including the most recent common
150 ancestors (MRCAs) for the virus lineages of interest and major clades.
151
152 *Time-scaled analysis*
153 We conducted time-scaled analysis using BEAST [22] to estimate the age of the node
154 representing the tMRCA at the root of the tree, and of ages of nodes representing the tMRCAs
155 of clusters of hybrid genomes identified to denote early transmission chains (C1, C2, and C3)
156 (*Supplementary Text 1. Methods*). The D4 time-calibrated ML tree was inputted as a starting
157 tree, customizing operators to keep the topology constant while re-estimating the root and
158 internal node heights. Drawing from existing literature [23,24] we used an informative strict clock
159 prior under a log-normal distribution centred on a mean of 8.0E-4 with a standard deviation of
160 1.0E-4. We ran two independent MCMC chains with length of 10E9 (sampling every 1E5
161 states) assessing for an effective sample size (ESS) > 200 for all parameters. Chains were
162 combined to generate summarized distributions by specifying a burn-in of 10%.
163
164 *Detection of Recombination*
165 To search for recombination breakpoints within the D1 and D4.1 alignments resulting in
166 topological incongruency, we used the GARD (Genetic Algorithm Approach for Detection of
167 Recombination) tool [25,26]. For further detection of recombinant sequences within D4, we used
168 the RDP4 suite (Recombination Detection Program) [27].
169
170 **RESULTS (word count: 2478)**
171
172 *A consistent phylogenetic placement is supported by an independent subsampling strategy*
173 The overall phylogenetic placement for all virus lineages observed across the D1-D4.1 trees
174 is consistent with the reference global SARS-CoV-2 phylogeny, indicating that our
175 independently subsampled datasets accurately capture the most recent evolutionary history
176 of the virus, particularly focusing on the evolution of the Omicron-descending BA.2* sub-
177 lineages (*Supplementary Figure 1*). The D1 tree shows a distal clustering of the BA.2, BA.2.75
178 and XBB.1.5* virus lineages, with the BA.2.86 lineage separated from the background by a
179 long branch suggesting accelerated evolution, compatible with observations raised as of
180 August 2023 [8].
181
182 After six months of circulation, the BA.2.86* lineage(s) diverged to give rise to JN.1*, first
183 detected in January 2024 [13]. Capturing the extended BA.2.86*/JN.1* diversity sampled as of
184 January 2024, the D2 tree shows a comparable pattern of phylogenetic placement, despite
185 the topological distortion driven by the over-representation of lineages BA.2.86*/JN.1* and

186 XBB.1.5*. The D2 tree confirms that BA.2.86*/JN.1* (foreground) forms a monophyletic clade
187 descending from a group of basally positioned BA.2 genomes. Nonetheless, basal BA.2 were
188 observed to shorten the branch leading to the foreground clade (*Supplementary Figure 1*).
189 These basal BA.2 genomes have collection dates ranging from March 2023 to June 2023 and
190 were sampled from multiple distinct geographic regions (Europe, Asia, North America, and
191 Australia). These observations support the presence of putative evolutionary intermediates
192 undetected during the initial characterisation of BA.2.86.
193
194 *Evolutionary intermediates that shorten the branch leading to BA.2.86 comprise hybrid*
195 *genomes*
196 To address potential sampling biases observed in D2 (*i.e.*, the over-representation of the
197 foreground relative to the background genomes) that could contribute to a phylogenetic
198 misplacement, a broader background diversity was included in the D3 tree. Additionally, to
199 increase the likelihood of capturing evolutionary intermediates, a subset of BA.2* 'mutated'
200 genomes displaying at least one BA.2.86-specific lineage defining mutation (LDMs) was also
201 included. We further identified ~130 basally-positioned BA.2 genomes that shorten the branch
202 to the foreground, supporting previous observations derived from the D2 tree (*Fig. 1a*)
203 (*Supplementary Data 1-Supplementary Table 6*). We subsequently investigated if putative
204 evolutionary intermediates also corresponded to hybrid genomes displaying (at least some)
205 BA.2.86-specific LDMs. We were able to identify 23 LDMs (5 in Orf1ab, 14 Spike, 3 in the M
206 and 1 in the N genes) with a consistent pattern of presence on nodes corresponding to BA.2
207 basal genomes (*Fig. 1a*) (*Supplementary Data 2*). Around 17 (13%) of these basal hybrid
208 genomes were collected up to 19 months prior to the initial detection of BA.2.86 (earliest
209 isolate: March 2022) from diverse geographic regions, mainly from North America, Europe
210 and Asia (*Fig. 1b*). Our results show multiple BA.2 genomes representing evolutionary
211 intermediates *i*) shorten the branch leading to the Foreground clade, and that also *ii*)
212 correspond to hybrid genomes displaying some, but not all, BA.2.86-specific LDMs.
213 Synchronous circulation of these hybrid isolates across multiple geographic regions, with
214 some collection dates preceding the initial detection of the BA.2.86, is indicative of cryptic
215 circulation.
216
217 Given that some 'mutated' genomes were retrieved under less stringent quality control, we
218 investigated if sequence quality could potentially impact the number of mutations observed,
219 particularly affecting the LDMs frequencies observed. Under the assumption of a positive
220 correlation between variables if sequence quality would increase mutation frequency, we
221 plotted normalised mutation counts per genome relative to ambiguous content. We observe a
222 negative correlation between mutation frequency counts (including LDMs) and genome

223 ambiguous content, indicating that an increased ambiguous content is related to the loss of
224 SNPs, rather than with the accumulation of non-synonymous mutations (*Supplementary*
225 *Figure 2*). Importantly, all hybrids are interspersed with other sampled genomes, with no
226 evident correlation. Thus, LDMs observed in hybrid genomes are unlikely to be an artefact
227 associated with reduced sequence quality. Moreover, after further addressing 'problematic'
228 genomes (filtering out potentially misdated sequences denoting outliers in the root-to-tip
229 divergence plots, lower quality sequences, as well as relabelling 'Pango' missassigned
230 genomes), the pattern of basally-positioned BA.2 genomes leading to the foreground is still
231 supported (*Supplementary Figure 1*), while LDM frequency is maintained. Consistently, only
232 a small proportion (<1.5%, 110/7760) of genomes within D3 showed conflict during 'Pango'
233 lineage reassignment, from which only 7 sequences were identified as BA.2 genomes
234 representing evolutionary intermediates, none of them with earlier collection dates
235 (*Supplementary Data 1-Supplementary Table 7, Supplementary Data 3*).
236
237 *Clusters of hybrid genomes denote early transmission chains*
238 We then assessed the temporal scale of the emergence of hybrid genomes in the background
239 along the full Omicron-descending BA.2* sub-lineages phylogeny, mapping these onto the D3
240 tree. We identify over 2600 hybrids genomes, most of which were placed amongst the
241 BA.2.75* and the XBB.1.5* lineage clades, and a minor proportion within the earlier BA.2 clade
242 (distal in the tree) (*Supplementary Data 1-Supplementary Table 6, Supplementary Figure 3*).
243 Despite most hybrids representing singletons in the tree, we identified three well-supported
244 clusters (named here C1, C2 and C3) that denote early transmission chains preceding the
245 emergence of BA.2.86. C1 locates within the distal early BA.2 clade, C2 within the XBB.1.5,
246 and C3 falls within the BA.2 genomes basal to the foreground (*Fig. 2a, Supplementary Figure*
247 *3*). Node ages corresponding to the tMRCAs of the clusters identified indicate that these
248 circulated up to 17 months prior to the initial detection of BA.2.86: C1 is estimated to have
249 emerged around February 2023 [95% HPD: 2022.89, 2023.26] in Mexico, C2 around April
250 2023 [95% HPD: 2023.15, 2023.34] in Japan, and C3 around February 2022 [95%
251 HPD:2022.05, 2022.12] in Malaysia (*Fig. 2b*). Consistent with previous estimates and
252 congruent with the known epidemiology of the virus [23,24], we inferred an overall evolutionary
253 rate of 7.3798E-4 substitutions/site/year [95% HPD: 7.0608E-4, 7.6924E-4], resulting in a
254 mean root age of the Omicron-descending BA.2* sub-lineages tree (excluding Wu-1) centred
255 around April 2021 [95% HPD: 2021.16, 2021.43] (*Fig. 2b*). Once more, hybrid background
256 genomes were ascribed collection dates ranging from January 2022 to February 2024, while
257 a synchronous co-circulation in multiple geographic regions (America, Europe, Africa, Asia,
258 and Australia) supports cryptic circulation dating back to the emergence of the BA.2 (*Fig. 2c*).

259

260 The clusters identified as early transmission chains are small (n<15) and exhibit some degree
261 of clonality regarding their collection region/date. C3 comprises genomes exclusively from
262 Malaysia all collected on the same date, while C1 and C2 comprise genomes obtained from
263 the same geographic locations with a reduced number of sequences (particularly evident in
264 the case of C1). Nonetheless, a high support value for the branch leading to all clusters [100%
265 bootstrap support in all three cases] (*Supplementary Figure 3*), coupled with a wide range of
266 collection dates observed in C1 and C2 (*Fig. 2c*), points out to these likely being true
267 representatives of undersampled localized transmission chains. On the other hand, despite
268 being assigned the same collection date from the same geographic region, assemblies within
269 C3 were not clonal, showing variability regarding the presence or absence of LDMs (*Table 1,*
270 *Supplementary Figure 3*) and carry other SNPs (reflected by the difference in branch lengths).
271 Importantly, no genomes sequences in any of the clusters identified were flagged as outliers
272 in the root-to-tip divergence plots (*Fig. 2b*). Even if C3 would not considered in our results,
273 node dates estimated for the C1 and C2 still support the early emergence of BA.2.86-like
274 genomes capable of successful transmission, months before the emergence and full
275 establishment of BA.2.86.
276

277 *Hybrid genomes show gradual evolution and a single rapid acquisition associated with*
278 *linkage/recombination*

279 We then looked into the presence/absence pattern of BA.2.86 LDMs in *i*) a subset of >100
280 hybrid genomes (both hybrid background genomes, those BA.2 basal to the foreground)
281 homogeneously subsampled according to collection date over a year (January 2023 to
282 January 2024), and *ii*) across genomes within C1-C3. We identify a pattern suggestive of a
283 gradual accumulation of LDMs over time, with at least 4 LDMs being already present in a
284 substantial number of genomes assigned to all background lineages (early BA.2, BA.2.75 and
285 XBB.1.5) as early as of January 2023 (*Supplementary Figure 4*). Genomes within C1-C3 also
286 display a distinctive subset of LDMs, suggestive of a progressive accumulation of LDMs over
287 time (*Table 1*). Derived from individual Orf alignments where LDMs occur (Orf1a, S, M and
288 N), we then reconstructed the full mutational trajectory at an amino acid level onto each node
289 of the time-calibrated D4 tree. We confirm that a proportion of LDMs (4 in Orf1a, 17 in Spike,
290 3 in M and 1 in N) map onto deeper nodes of the tree, while it is apparent that most LDMs
291 were sequentially fixed in the BA.2 genomes basal to the foreground (*Fig. 3a*). The mutational
292 trajectory of S revealed that some LDMs map onto earlier nodes of the tree corresponding to
293 the distally positioned BA.2, dating back to December 2021 (*i.e.,* D339H). Conversely, some
294 LDMs (*i.e.,* G446S, N460K, N481K, P621S, R158G S455L K481N, S50L, V127F) co-occur

295 with other non-LDMs, while some sites (*i.e.,* 484 and 445) show residue toggling and/or
296 reversions (*Fig. 3a*).
297
298 When plotting LDM frequency relative to sample collection date, we find a basal number of
299 LDMs (~5) already present in a substantial number of BA.2 and BA.2.75 genomes sampled
300 after January 2022. These results suggest that a reduced constellation of LDMs was already
301 fixed in these earlier lineages, with mutation frequency gradually increasing over time. Thus,
302 this subset of LMDs represent synapomorphic traits in a proportion of BA.2, BA.2.75 and XBB*
303 genomes, denoting pre-existing standing genetic variation. After July 2023, we observe that
304 the number of observed mutations explosively increased to over 20 LDMs, suggestive of a
305 single rapid acquisition of a secondary constellation of LDMs. Increase is mostly observed in
306 genomes assigned as BA.2.86*/JN.1*, but also in a considerable number of BA.2 and some
307 XBB* genomes (*Fig. 3b*). When plotting the frequency of all LDMs in the BA.2.86*/JN.1* clade
308 (beyond the 23 detected in hybrid genomes here) over the full timescale of the pandemic, it is
309 evident that four LDMs these were first observed as early as 2020-2021, displaying fluctuating
310 frequencies across time, prior to the detection of the BA.2.86 (*e.g.,* A1892T in NSP3 and
311 P681R, E484K and G446S in Spike). Importantly, four independent LDMs in Spike (*i.e.,*
312 G339H, N460K, G446S, F486P in Spike) show a pattern of a sequential emergence between
313 June and December 2022, with an increase in frequency reaching fixation in >90% of all
314 genomes sampled by June 2023. This is suggestive of multiple mutations having an additive
315 effect enhancing fitness, consistent with a soft selective sweep (*Supplementary Figure 5*).
316 Again, we observe another subset of LDMs displaying a synchronous increase in frequency
317 after June 2023 (*e.g.,* P681R, E484K in Spike), denoting a hard selective sweep following
318 acquisition through linkage/recombination. The temporal scale observed for the emergence of
319 LDMs is consistent with the timeframe of emergence of hybrid genomes, suggesting that many
320 of these mutations may have spread through a process of gene flow and/or recombination.
321
322 Of interest, over a broader temporal scale of SARS-CoV-2 evolution, tracking LDMs across
323 the global dataset available in Nextstrain (accessed as of June 26th 2024, with 3821 genomes
324 collected between Dec 2019 and June 2024) revealed that at least 7 LDMs are homoplasic
325 across different viral lineages not directly related to BA.2.86, with initial emergence in virus
326 lineages preceding the divergence of the Omicron complex. Specifically, Orf1ab/T2710 is
327 homoplasic between BA.2.86 and BA.1 (Nextstrain_2710) S/G158 and S/R681 are
328 homoplasic between BA.2.86 and Delta (Nextstrain_158, Nextstrain_681), S/H339 and
329 S/P486 are homoplasic between BA.2.86 and XBB* (Nextstrain_339, Nextstrain_486), and
330 S/S446 and S/K460 are homoplasic between BA.2.86 and XBB* and some BA.* lineages
331 (Nextstrain_446, Nextstrain_460).

332

333 *Evidence for recombination*

334 As early as August 2023, it had been suggested that recombination between the most closely
335 related (BA.2.75, XBB.1.5 and BA.2) virus lineages may have contributed to the emergence
336 of the BA.2.86 lineage [8], but this hypothesis was never formally tested. Through an initial
337 analysis using GARD applied to the earliest D1 dataset, we detected early evidence for
338 recombination, in which four breakpoints were inferred across the alignment (*Supplementary*
339 *Data 4)*. A single best breakpoint at codon site 21,314 was supported by four different models
340 (Δ c-AIC vs the null model: 2306.91, Δ c-AIC vs the single tree multiple partition: 2392.67),
341 yielding incongruent topologies corresponding to two partitions. Consistent with the reference
342 phylogeny, the Partition 1 tree shows that the BA.2.86 lineage directly descends from BA.2.
343 In the Partition 2 tree, the BA.2.86 is a sister lineage to the BA.2.75 (*Supplementary Figure*
344 *6*). Recombination analyses applied to the independently and later subsampled D4.1 dataset
345 (including a broader diversity representing the background and BA.2.86*/JN.1*) recapitulates
346 these results, with the same recombination breakpoint being inferred (codon site 21,314)
347 supported under five different models (Δ c-AIC vs the null model: 2363.56, Δ c-AIC vs the
348 single tree multiple partition: 1.6E-7) (*Supplementary Data 5)*. Again, in Partition 1 tree,
349 BA.2.86* directly descends from BA.2, while in Partition 2 tree, BA.2.86* is a sister lineage to
350 BA.2.75 (*Supplementary Figure 6*).

351

352 Topological incongruence in the branch leading to the BA.2.86*/JN.1* clade, corresponds to
353 a complete clade misplacement, rather than the misplacement of a subset of internal branches
354 within the clade. This indicates that recombination is likely to have significantly contributed to
355 the emergence of the BA.2.86 lineage, in line with our observation made on the subset of
356 LDMs denoting a hard selective sweep. Codon site 21,314 (A**T**A) locates to residue I6994
357 within Orf1ab (motif PKEQ**I**DGY), residing in a loop that connects a beta-sheet and an alpha-
358 helix of nsp16. This residue sits distally from where inhibitors of nsp16 bind, with no expected
359 interference to biological function, rendering it as a plausible location for recombination to
360 occur without having a significant impact on fitness linked to a disruptive function at a protein
361 level.

362

363 Using a combination of recombination detection methods available in the RDP4 suite [27], we
364 were able to further identify nine putative recombination events resulting in multiple candidate
365 recombinant sequences (*Supplementary Data 6)*. From these, five recombination events
366 supported by at least one method involved hybrid background (BA.2) and foreground
367 genomes. However, the number of synapomorphic SNPs (including LDMs) shared between
368 the background and foreground genomes poses a strong limitation to triplet-based

369 approaches, increasing the likelihood of recombination events detected as false positives [28].
370 Detecting recombination in large-scale data is computationally intensive, while identifying
371 recombination-associated patterns in recently diverged SARS-CoV-2 lineages proves
372 challenges around distinguishing homoplasies from synapomorphic changes [29]. Nonetheless,
373 a sudden increase in frequency in a subset of LDMs just before the emergence of BA.2.86
374 further supports that recombination contributed to the emergence of this highly divergent virus
375 lineage, providing a plausible explanation for how a basal number of LDMs between 5-10
376 increased to over 30 after July 2023. Our inability to detect clear LDM acquisition patterns may
377 be linked to the distribution of some LDMs across distinct background lineages, possibly linked
378 to the deep origins of some BA.2* lineages deriving from prior recombination events (such as
379 XBB.1.5 being a recombinant between BA.2.75 and BA.2.10) [30].
380
381 **DISCUSSION (word count: 1390)**
382
383 The emergence of mutations in SARS-CoV-2 (and RNA viruses in general) is driven by the
384 inherent rate at which substitutions occur, shaped by diverse biological processes such as a
385 low-fidelity viral genome replication, host-mediated genome editing, and recombination,
386 amongst others (extensively reviewed in [4]). Mutations that subsequently reach fixation as high
387 frequency alleles are likely to be adaptive Thus, an important question that remains to be
388 addressed is what drives the fixation of LDMs, especially in highly divergent virus lineages
389 (such as VOCs, and the BA.2.86 VOI considered here). Elucidating the evolutionary forces
390 shaping the emergence of LDMs is challenging due to changing selective landscapes, where
391 viral genetic and population dynamics are influenced by varying host immune responses
392 (reviewed in [1,31]). The possibility of recombination further adds to this complexity, representing
393 evolutionary events that may drive change as shifts, but can also inflate dN/dS estimates,
394 leading to false positives when inferring signatures of positive selection [32].
395
396 Here we formally assess the evolutionary history of the BA.2.86 VOI and related lineages,
397 analysing mutation distributions over time, with a focus on LDMs. Along the complete
398 Omicron-descending BA.2* sub-lineage phylogeny, we observe that an early constellation of
399 LDMs corresponding to synapomorphic SNPs was acquired through time sequentially in the
400 background and subsequently in the foreground. On the other hand, evidence for
401 recombination points to a secondary constellation of LDMs rapidly acquired in a single event
402 through linkage. Moreover, some LDMs were observed to co-occur with other non-LDMs,
403 suggestive of putative epistatic interactions (such as compensatory mutations). We can
404 speculate that this behaviour may be linked to functional optimization, or an antigenic drift
405 process shaped by positive selection linked to immune evasion [33,34]. Conversely, other LDM-

406 specific sites show residue toggling and/or reversions, while fluctuating frequencies for some
407 LDMs across the complete pandemic timescale putatively denote maladaptation within
408 specific genetic backgrounds, and/or the exploration of evolutionary sub-optima prior to
409 fixation in BA.2.86*/JN.1*. These observations are further supported by those LDMs denoting
410 homoplasies across unrelated virus lineages (*e.g.* BA.2.86 and Delta). Thus, our results
411 overall indicate that considerable proportion of LDMs 'unique' to BA.2.86 were fixed in various
412 SARS-CoV-2 subpopulations prior to the emergence of this lineage, following multiple
413 selective regimes and evolutionary processes.
414
415 In agreement with this, this subset of LDMs sequentially fixed over a short period time prior to
416 the emergence of BA.2.86, display a pattern consistent with a soft selective sweep derived
417 from pre-existing standing genetic variation [35]. In this scenario, rapid evolution is also plausible
418 since co-occurring mutations may provide additive adaptive benefits. This could also
419 potentially explain why not all LDMs may enhance fitness equally, for example, when linked
420 to immune evasive phenotypes. Supporting this observation, we recently found that a
421 considerable number (~15) of BA.2.86-specific LDMs may marginally contribute to diminished
422 T-cell responses relative to the parental Omicron VOC. Nonetheless, LDMs are expected to
423 have an additive effect (with some contributing positively and some negatively to MHC
424 binding), so the functional impact is mitigated. Moreover, LDMs are not expected to be driven
425 by selective forces exerted by T-cell immunity (*manuscript under review*).
426
427 For VOCs and other highly divergence lineages, such as the BA.2.86* VOI, it has been
428 hypothesised that these may arise through accelerated evolution linked to single emergence
429 events associated with individual chronic infection cases, with spread in the community
430 facilitated under a prolonged shedding scenario [4,6]. However, in the case of BA.2.86, we
431 observe the presence of early LDM constellations in evolutionary intermediates and hybrid
432 genomes along the full evolutionary history of the Omicron descending BA.2* sub-lineages.
433 Thus, emergence of BA.2.86/JN.1*during long-term infection is an unlikely scenario. In turn,
434 our results favour an alternative hypothesis of gradual evolution and cryptic circulation linked
435 to multiple emergence events. Consistent with this, it has been suggested that under a
436 constant evolutionary rate estimated for different SARS-CoV-2 lineages (including VOCs), the
437 emergence of highly divergent lineages with at least 12 amino acid changes would require the
438 variant to remain undetected for nearly six months before reporting [4]. Congruently, we observe
439 that up to 17 months elapsed before the emergence of the earliest recorded BA.2.86-like
440 genomes (with already a considerable number of LDMs), and the subsequent emergence and
441 establishment of the BA.2.86 lineage.
442

443 Long-term infections may explain a rapid accumulation of mutations, but does account for
444 evolutionary intermediates observed in the evolutionary history of highly divergent SARS-CoV-
445 2 lineages. We show a vast geographical and temporal extent of evolutionary intermediates
446 corresponding to hybrid genomes preceding the emergence of BA.2.86*. Nonetheless, such
447 observations are not novel, in the sense that evolutionary intermediates have been previously
448 described (to a lesser extent) in the evolutionary history of other VOCs. Despite controversy
449 in the detection of evolutionary intermediates in the emergence of highly divergent VOCs like
450 Omicron [36], there have been a handful of studies providing limited but credible evidence for
451 true evolutionary intermediates leading to highly divergent clades. For example, at least for
452 the Alpha and Gamma VOCs, ancestral intermediate genomes evidenced the acquisition of
453 LDM constellations through multiple interhost transmissions, with the persistence of endemic
454 lineages circulating cryptically [37-39].
455
456 Failing to detect evolutionary intermediates is inherently linked to our inability to search for
457 and detect *unknown* mutations that will become important in virus lineages of increasing
458 epidemiological relevance. Once highly divergent virus lineages have emerged, evolutionary
459 intermediates may go undetected, in part, due to the decrease in genomic surveillance as the
460 pandemic progressed, and given uneven surveillance efforts across geographic regions,
461 reflecting disparities in sequencing and analytical capacities. For example, during the initial
462 discovery of Omicron in South Africa and Botswana, where genomic surveillance was more
463 robust than in many other parts of Africa, it was suggested that the variant could have been
464 spreading unnoticed elsewhere. Incomplete sampling also challenges inference of the
465 geographic origins of lineages as well as reconstruction of viral diffusion patterns across time
466 and space [40]. Nonetheless, although genomic surveillance of SARS-CoV-2 has decreased
467 significantly (with a ten-fold reduction in the number of genomes uploaded to GISAID in August
468 2022 compared to July 2023, GISAID[16] accessed as of June 26, 2024), with over 1,600,000
469 virus genomes assigned to the Omicron VOC and descending sublineages (B.1.1.529+BA.*)
470 deposited between January 2023 and 2024 (GISAID[16] accessed as of June 26, 2024), we
471 were still able to detect a substantial number of evolutionary intermediates and hybrid
472 genomes sampled across multiple geographic regions over time.
473
474 Detecting evolutionary intermediates post emergence may be further limited by current efforts
475 to computationally assess and visualise the global SARS-CoV-2 evolutionary history, requiring
476 extensive subsampling to support tractability. Heavily down-sampled datasets, with genomes
477 selected under hierarchical strategy based on geographic and temporal distribution [12] can lead
478 to potential subsampling biases with underrepresented genome sequences [41]. More than
479 15,000 'mutated' genomes were retrieved under our search criteria, and only around 2,000 of

480 these were sampled for our analyses. Thus, it is likely that the proportion of evolutionary
481 intermediates and hybrid genomes we detect is underestimated. Related to global SARS-CoV-
482 2 phylogenetic interactive visualization platforms such as Nextstrain, and with over 16 million
483 genomes now available in GISAID, the likelihood of sampling some evolutionary intermediates
484 and hybrid genomes in a reduced dataset of less than 4,000 genomes is extremely low.
485 Therefore, we propose that for an improved periodic assessment of the phylogenetic
486 placement of emerging highly divergent virus lineages, were expanded subsampling and
487 iterative phylogenetic inference should be favoured. This could help identify intermediate
488 sequences denoting gradual evolution. Moreover, looking for LDMs beyond directly parental
489 virus lineages could help identify the acquisition of early constellations of mutations.
490
491 Here, we elucidate the evolutionary drivers of the emergence of BA.2.86, a highly divergent
492 lineage that led to currently dominating virus subpopulations globally. We implicate the
493 contribution of recombination and cryptic circulation over a single emergence scenario. Our
494 analyses highlight how uninformed subsampling strategies of SARS-CoV-2 datasets can lead
495 to overlooking complex evolutionary trajectories. While not scalable to understanding the
496 emergence of all highly divergent SARS-CoV-2 lineages, our work supports preliminary
497 observations made for other VOCs like Alpha and Gamma. Importantly, we test the general
498 evolutionary mechanisms underlying the emergence of this highly divergent SARS-CoV-2
499 variant. We suggest how emerging highly divergence variants should be routinely
500 investigated, providing a framework for future studies aimed at elucidating the mechanisms
501 behind viral diversification.

DATA AVAILABILITY

EPI_SET IDs for D1-D4 are as follow:

D1 EPI_SET ID:	EPI_SET_240717hn	doi:	10.55876/gis8.240717hn
D2 EPI_SET ID:	EPI_SET_240717ps	doi:	10.55876/gis8.240717ps
D3 EPI_SET ID:	EPI_SET_240717xg	doi:	10.55876/gis8.240717xg
D4 EPI_SET ID:	EPI_SET_240717xr	doi:	10.55876/gis8.240717xr

Unique identifiers are used for both the acknowledgment of data contributors and to permit registered users to retrieve the dataset of all the records encompassed in the EPI_SET ID. The resulting GARD outputs (*Supplementary Data 4 and 5*) can be visualized using Alignments corresponding to the D1-D4.1 datasets are available in Zenodo (accessible DOI 10.5281/zenodo.12764160 https://zenodo.org/records/12764160).

ACKNOWLEDGMENTS

The project and M.E.Z are Funded by the European Union under grant HORIZON-HLTH-2021-CORONA-01-02 agreement no. 101046314. LvD is additionally funded by a UKRI Future Leads Fellowship MR/X034828/1.Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We would like to acknowledge Pooja Swali and Alexei Yavlinsky for their support in dataframe construction and visualization.

FIGURE LEGENDS

Figure 1. Basal BA.2 genomes that shorten the branch to the BA.2.86/JN.1 lineage represent evolutionary intermediates.

(a) Zoom-in of the BA.2.86*/JN.1* clade within the D3 ML tree, showing over 100 basal BA.2 genomes that represent evolutionary intermediates corresponding to hybrid strains displaying some BA.2.86-specific LDMs. The presence/absence heatmap for a subset of evolutionary intermediates shows twenty-three LDMs as 'derived' states present on nodes corresponding to these basal BA.2 genomes (in teal), with a gradual acquisition pattern over time relative to the 'ancestral' states (in dark blue) present in the parental BA.2 (conserved/other states are indicated in ochre). (b) A proportion of basal BA.2 genomes were collected up to 17 months prior to the detection of BA.2.86 (earliest: March 2022, latest: February 2024) from diverse geographic regions, mainly North America, Europe, and Asia. The density of genomes collected per region is indicated by an increasing circle size on the map, with circles coloured according to the sampling locations. Collection dates per genome sequence are indicated per region, with the dashed line marking the detection of BA.2.86. Synchronous circulation across geographic regions, with some collection dates preceding the detection of BA.2.86, is indicative of cryptic circulation in multiple locations.

Figure 2. Three clusters identified across the complete Omicron-descending BA.2 sub-lineage tree denote early transmission chains of hybrid genomes.

(a) The D3 ML time-scaled tree with branches coloured according to Pango lineage. Grey circles indicating the placement of three clusters (C1-C3) representing early transmission chains (clusters of hybrid genomes) placed within the genetic background of the Omicron-descending BA.2* sub-lineages. (b) Root-to-tip divergence plot following tree time calibration shows the inferred evolutionary rate and mean root age of the Omicron-descending BA.2* sub-lineage tree with the corresponding 95% HDP interval indicated. Density plots showing the 95% HDP intervals for the tMRCA node age estimates for C1-C3, with C1 in ochre, C2 in red, and C3 in blue. (c) Hybrid genomes assigned to the background (early BA.2 in purple, BA.2.75 in red, and XBB.1.5 in yellow) were globally sampled up to 19 months prior to the detection of BA.2.86 (earliest: January 2022, latest: February 2024). Tip collection dates per genome sequence are indicated per virus lineage, indicating median sampling date per

cluster. The dashed line marks the detection of BA.2.86. Synchronous circulation across geographic regions, with collection dates preceding the detection of BA.2.86, is indicative of a cryptic circulation of hybrid genomes.

Figure 3. The full mutational trajectory at an amino acid level of individual Orf alignments where LDMs occur.

(a) The Orf1a, S, M, and N full mutational trajectory estimated under the ML framework, mapped onto the time-calibrated D4 tree. We confirm that a proportion of LDMs (4 in Orf1a, 17 in Spike, 3 in M, and 1 in N) (indicated with red circles) locate onto deeper nodes of the tree. For the S mutational trajectory, mutations indicated in black correspond to those co-occurring with known LDMs (R346T, L454S, F157S, S939F, P1143L), or sites denoting toggling residues (484 and 445). Branches are coloured according to Pango lineage, highlighting only the BA.2.86*/JN.1* clade (in purple and teal, respectively). Node ages corresponding to the tMRCA for the earliest emergence of a given LDM per Orf are indicated with an *. Most LDMs were sequentially fixed in the BA.2 genomes basal to the foreground, consistent with the notion of gradual evolution. (b) Scatter plot representing the frequency counts of LDMs relative to sample collection date, with dots coloured according to Pango lineage. The dashed line marks the detection of BA.2.86. A basal number of LDMs (~5) was already present in numerous BA.2 and BA.2.75 genomes as of July 2022. After July 2023, LDM frequency explosively increased to over 30, mostly in genomes assigned to the BA.2.86*/JN.1*, but also in a considerable number of XBB.1.5 and BA.2 genomes. LDM enrichment is particularly evident in BA.2 genomes representing evolutionary intermediates

Table 1. LDMs in presence/absence in sequences belonging to C1-C3 *

		FLIP		ORF1a					S																							M			N
		L455S	L456F	A211D	V1056L	N2526S	A2710T	V3593F	S50L	V127F	R158G	L216F	H245N	A264D	I332V	D339H	K356T	R403K	V445H	G446S	N450D	L452W	N460K	N481K	A484K	F486P	R493Q	E554K	A570V	P621S	H681R	D3H	T30A	A104V	Q229K
Earliest BA.2.86*	FG|BA.2.86.1|EPI_ISL_18097345|2023-07-24|Denmark	L	F	D	L	S	T	F	L	F	G	F	N	D	V	H	T	K	H	S	D	W	K	K	K	P	Q	K	V	S	R	H	A	V	K
Earliest BA.2†	BG|BA.2|EPI_ISL_17592056|2023-04-08|India	L	F	A	V	N	A	V	S	V	R	L	H	A	I	G	K	R	V	S	N	L	N	N	E	F	Q	E	A	P	H	D	T	A	Q
C3 (BA.2 basal to the FG)	Earliest collection date: 2022-03-11																																		
	MT|BA.2|EPI_ISL_18821487|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821492|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821480|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821490|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821494|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821486|2022-03-11|Malaysia	S	F	A	V	N	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821491|2022-03-11|Malaysia	S	F	A	V	N	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821483|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	D	T	A	Q
	MT|BA.2|EPI_ISL_18821484|2022-03-11|Malaysia	S	F	A	V	N	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	H	A	A	Q
	MT|BA.2|EPI_ISL_18821485|2022-03-11|Malaysia	S	F	A	V	S	A	V	S	V	R	L	H	A	I	D	T	K	H	S	D	W	K	K	X	P	Q	K	V	S	R	H	A	A	Q
C2 (XBB.1.5)	Earliest collection date: 2023-07-03																																		
	BG|XBB.1.5.90|EPI_ISL_17994831|2023-07-04|Finland	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	R	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18006690|2023-07-06|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	X	X	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18033115|2023-07-03|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	R	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18060516|2023-07-14|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	R	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18116248|2023-07-28|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	R	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18097793|2023-07-12|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	R	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18040349|2023-07-05|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	X	X	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18097862|2023-07-05|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	X	X	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18101027|2023-07-20|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	X	X	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18101036|2023-07-21|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	X	X	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18097643|2023-07-22|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	R	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
	BG|XBB.1.5.90|EPI_ISL_18106303|2023-08-03|Japan	L	F	A	V	N	A	V	S	V	R	L	H	A	I	H	K	X	P	S	N	L	K	N	A	P	Q	E	A	S	H	D	T	A	Q
C1 (distal BA.2)	Earliest collection date: 2023-07-13																																		
	MT|BA.2|EPI_ISL_18415854|2023-08-04|Mexico	L	F	A	V	N	A	V	S	V	K	L	H	A	I	D	K	R	V	S	N	R	N	N	E	L	Q	E	A	P	H	Y	T	A	Q
	MT|BA.2|EPI_ISL_18798234|2024-01-10|Mexico	L	F	A	V	N	A	V	S	V	K	L	H	A	I	D	K	K	V	N	N	R	N	N	K	L	Q	E	A	P	H	Y	T	A	Q
	MT|BA.2|EPI_ISL_18798204|2023-11-30|Mexico	L	F	A	V	N	A	V	S	V	K	L	H	A	I	D	K	K	V	N	N	R	N	N	K	L	Q	E	A	S	Y	Y	T	A	Q
	MT|BA.2|EPI_ISL_18415832|2023-07-13|Mexico	L	F	A	V	N	A	V	S	V	K	L	H	A	I	G	K	R	V	S	N	R	N	N	K	L	Q	E	A	S	Y	Y	T	A	Q

CONSERVED/OTHER
ANCESTRAL (BA.2)
DERIVED (BA.2.86)
ABSENT (X)

* Identified in D3 and D4

‡ Sampled in D3

Figure 1

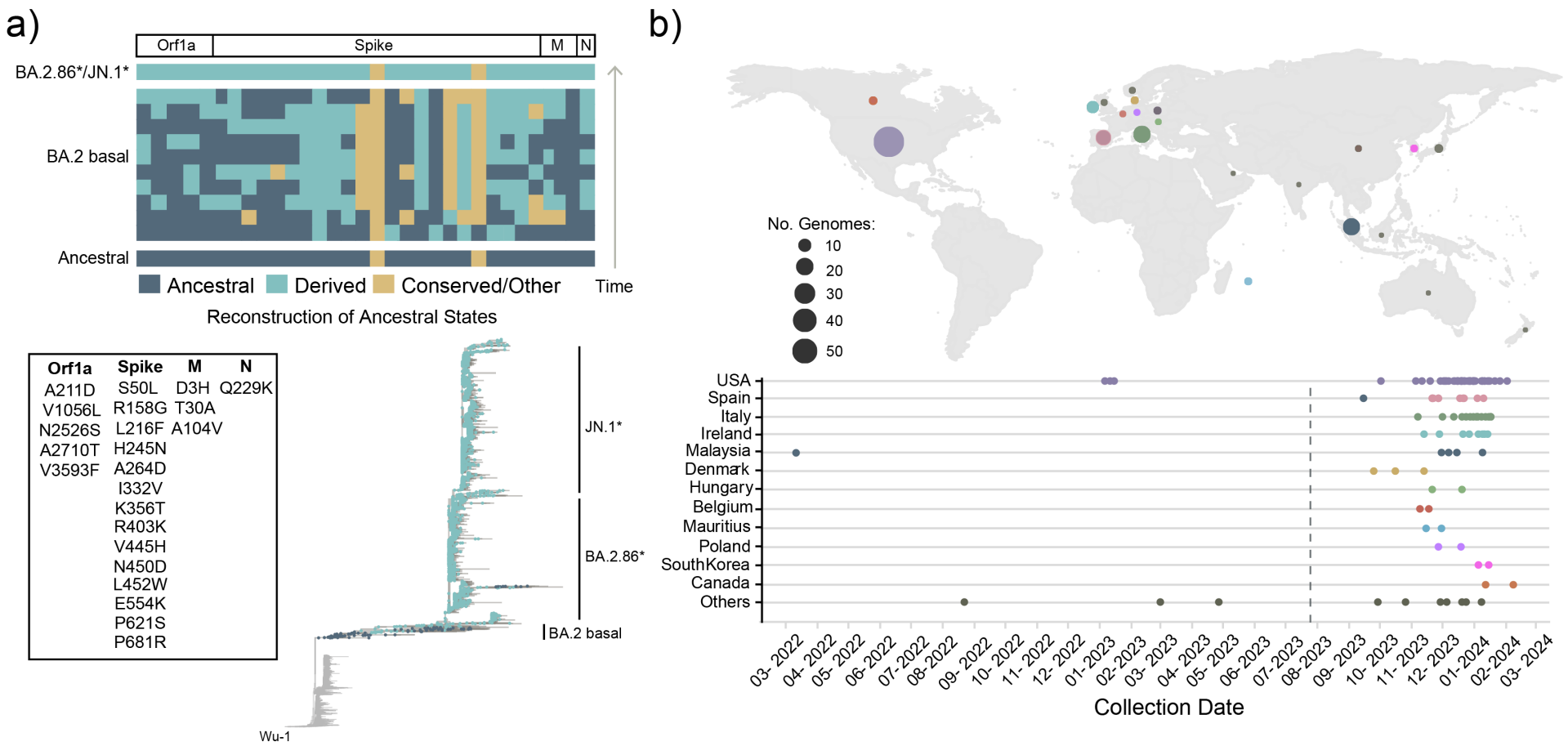

a)
Orf1a
Spike
M
N
BA.2.86*/JN.1*
BA.2 basal
Ancestral
Ancestral
Derived
Conserved/Other
Time
Reconstruction of Ancestral States
Orf1a Spike M N
A211D S50L D3H Q229K
V1056L R158G T30A
N2526S L216F A104V
A2710T H245N
V3593F A264D
I332V
K356T
R403K
V445H
N450D
L452W
E554K
P621S
P681R
JN.1*
BA.2.86*
BA.2 basal
Wu-1
b)
No. Genomes:
10
20
30
40
50
USA
Spain
Italy
Ireland
Malaysia
Denmark
Hungary
Belgium
Mauritius
Poland
SouthKorea
Canada
Others
03-2022
04-2022
05-2022
06-2022
07-2022
08-2022
09-2022
10-2022
11-2022
12-2022
01-2023
02-2023
03-2023
04-2023
05-2023
06-2023
07-2023
08-2023
09-2023
10-2023
11-2023
12-2023
01-2024
02-2024
03-2024
Collection Date

Figure 2

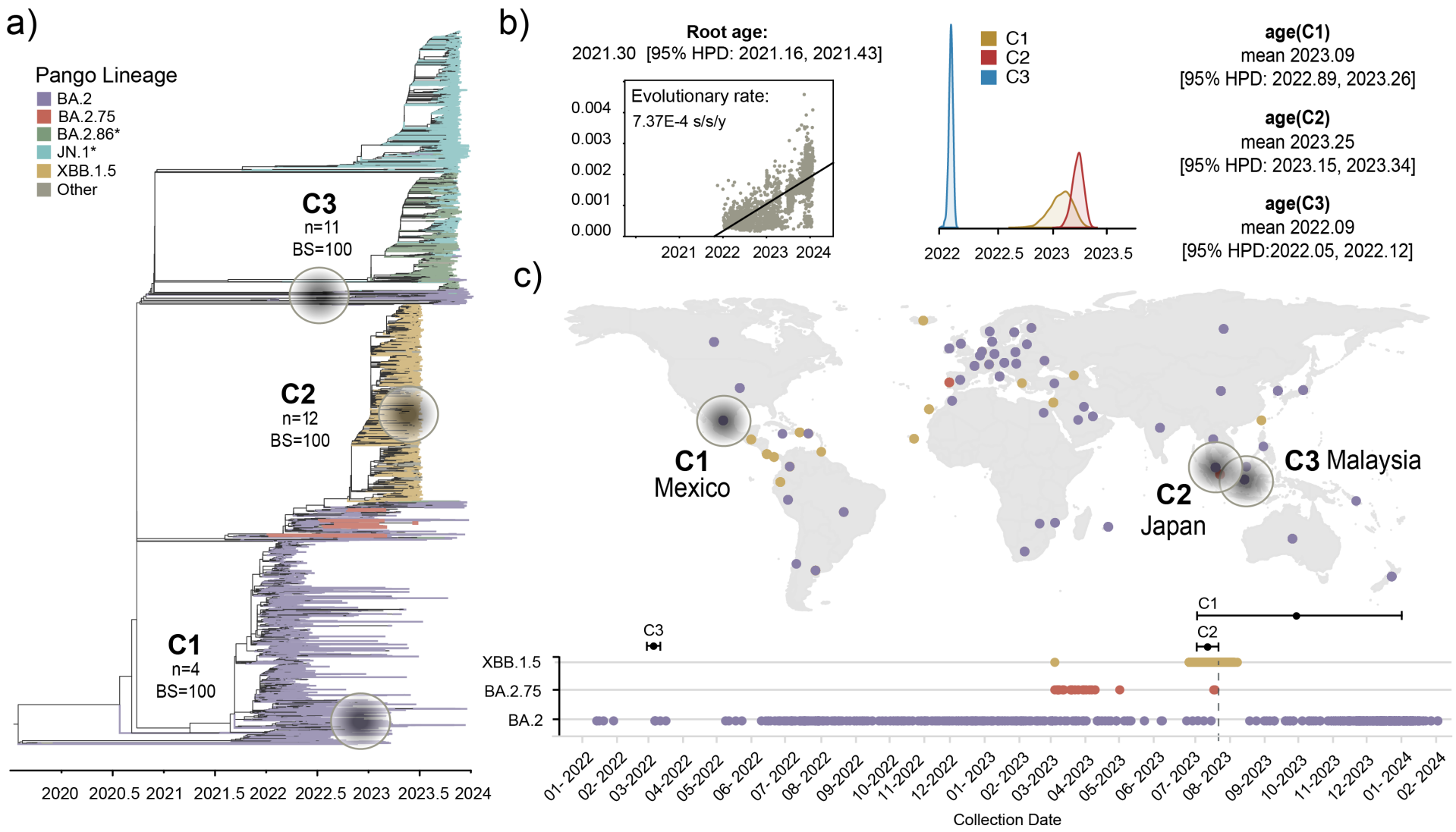
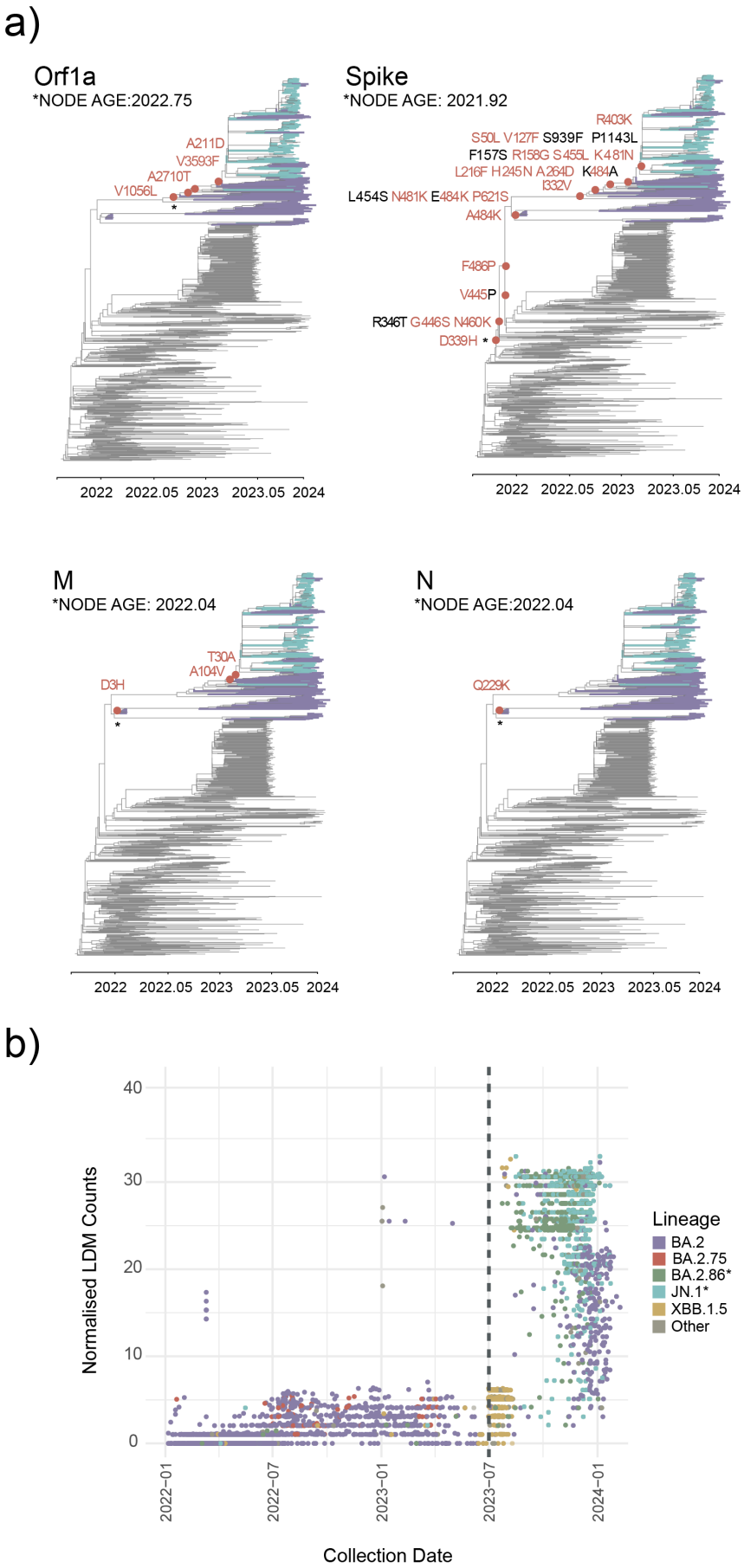

a)
Pango Lineage
BA.2
BA.2.75
BA.2.86*
JN.1*
XBB.1.5
Other
C3
n=11
BS=100
C2
n=12
BS=100
C1
n=4
BS=100
2020 2020.5 2021 2021.5 2022 2022.5 2023 2023.5 2024
b)
Root age:
2021.30 [95% HPD: 2021.16, 2021.43]
Evolutionary rate:
7.37E-4 s/s/y
0.004
0.003
0.002
0.001
0.000
2021 2022 2023 2024
C1
C2
C3
2022 2022.5 2023 2023.5
age(C1)
mean 2023.09
[95% HPD: 2022.89, 2023.26]
age(C2)
mean 2023.25
[95% HPD: 2023.15, 2023.34]
age(C3)
mean 2022.09
[95% HPD:2022.05, 2022.12]
c)
C1
Mexico
C2
Japan
C3 Malaysia
C1
C2
C3
XBB.1.5
BA.2.75
BA.2
01-2022 02-2022 03-2022 04-2022 05-2022 06-2022 07-2022 08-2022 09-2022 10-2022 11-2022 12-2022 01-2023 02-2023 03-2023 04-2023 05-2023 06-2023 07-2023 08-2023 09-2023 10-2023 11-2023 12-2023 01-2024 02-2024
Collection Date

Figure 3


a)
Orf1a
*NODE AGE:2022.75
A211D
V3593F
A2710T
V1056L
*
2022 2022.05 2023 2023.05 2024
Spike
*NODE AGE: 2021.92
R403K
S50L V127F S939F P1143L
F157S R158G S455L K481N
L216F H245N A264D K484A
I332V
L454S N481K E484K P621S
A484K
F486P
V445P
R346T G446S N460K
D339H *
2022 2022.05 2023 2023.05 2024
M
*NODE AGE: 2022.04
T30A
A104V
D3H
*
2022 2022.05 2023 2023.05 2024
N
*NODE AGE:2022.04
Q229K
*
2022 2022.05 2023 2023.05 2024
b)
Normalised LDM Counts
40
30
20
10
0
2022−01
2022−07
2023−01
2023−07
2024−01
Collection Date
Lineage
BA.2
BA.2.75
BA.2.86*
JN.1*
XBB.1.5
Other

REFERENCES

1 Balloux, F. *et al.* The past, current and future epidemiological dynamic of SARS-CoV-2. *Oxf Open Immunol* **3**, iqac003 (2022). https://doi.org:10.1093/oxfimm/iqac003

2 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403-1407 (2020). https://doi.org:10.1038/s41564-020-0770-5

3 WHO. *Tracking SARS-CoV-2 variants*, <https://www.who.int/activities/tracking-SARS-CoV-2-variants> (

4 Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat Rev Microbiol* **21**, 361-379 (2023). https://doi.org:10.1038/s41579-023-00878-2

5 Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277-282 (2021). https://doi.org:10.1038/s41586-021-03291-y

6 Ghafari, M., Liu, Q., Dhillon, A., Katzourakis, A. & Weissman, D. B. Ghafari, M., Liu, Q., Dhillon, A., Katzourakis, A., Weissman, D.B., . *Frontiers in Virology* **2** (2022). https://doi.org:https://doi.org/10.3389/fviro.2022.942555

7 de Oliveira, T. *et al.* A novel variant of interest of SARS-CoV-2 with multiple spike mutations detected through travel surveillance in Africa. *medRxiv* (2021). https://doi.org:https://doi.org/10.1101/2021.03.30.21254323

8 GitHub. *cov-lineages/pango-designation #2183* https://github.com/cov-lineages/pango-designation/issues/2183

9 outbreak.info. *BA.2.86 Lineage Report*.

10 WHO. *Initial Risk Evaluation of BA.2.86 and its sublineages 21 November 2023*, <https://www.who.int/docs/default-source/coronaviruse/21112023_ba.2.86_ire.pdf?sfvrsn=8876def1_3> (2023).

11 Nextstrain. *Nextstrain*, <https://nextstrain.org/ncov/gisaid/global/all-time?c=pango_lineage> (

12 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018). https://doi.org:10.1093/bioinformatics/bty407

13 outbreak.info. *JN.1 Lineage Report*, <https://outbreak.info/situation-reports?xmin=2024-01-18&xmax=2024-07-18&pango=JN.1> (

14 CDC. United States COVID-19 Deaths, Emergency Department (ED) Visits, and Test Positivity by Geographic Area.

15 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46 (2017). https://doi.org:10.1002/gch2.1018

16 GISAID. GISAID. https://gisaid.org/

17 Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly* **3**, 1049-1051 (2021). https://doi.org:10.46234/ccdcw2021.255

18 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill* **22** (2017). https://doi.org:10.2807/1560-7917.ES.2017.22.13.30494

19 Nextstrain. Phylogenetic analysis of BA.2.86 and BA.2.

20 Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018). https://doi.org:10.1186/s12859-018-2164-8

21 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018). https://doi.org:10.1093/ve/vex042

22 Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* **4**, vey016 (2018). https://doi.org:10.1093/ve/vey016

23 Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol* **6**, veaa061 (2020). https://doi.org:10.1093/ve/veaa061

24 Pekar, J., Worobey, M., Moshiri, N., Scheffler, K. & Wertheim, J. O. Timing the SARS-CoV-2 index case in Hubei province. *Science* **372**, 412-417 (2021). https://doi.org:10.1126/science.abf8003

25 Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. GARD: a genetic algorithm for recombination detection. *Bioinformatics* **22**, 3096-3098 (2006). https://doi.org:10.1093/bioinformatics/btl474

26 Kosakovsky Pond, S. L., Posada, D., Gravenor, M. B., Woelk, C. H. & Frost, S. D. Automated phylogenetic detection of recombination using a genetic algorithm. *Mol Biol Evol* **23**, 1891-1901 (2006). https://doi.org:10.1093/molbev/msl051

27 Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol* **1**, vev003 (2015). https://doi.org:10.1093/ve/vev003

28 Jaya, F. R., Brito, B. P. & Darling, A. E. Evaluation of recombination detection methods for viral sequencing. *Virus Evol* **9**, vead066 (2023). https://doi.org:10.1093/ve/vead066

29 Gutierrez, B. *et al.* Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. *Cell Host Microbe* **30**, 1112-1123 e1113 (2022). https://doi.org:10.1016/j.chom.2022.06.010

30 Tamura, T. *et al.* Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nat Commun* **14**, 2800 (2023). https://doi.org:10.1038/s41467-023-38435-3
31 Carabelli, A. M. *et al.* SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* **21**, 162-177 (2023). https://doi.org:10.1038/s41579-022-00841-7
32 Fares, M. in *Natural Selection: Methods and Applications* (ed CRC Press) (2014).
33 Uriu, K. *et al.* Transmissibility, infectivity, and immune evasion of the SARS-CoV-2 BA.2.86 variant. *Lancet Infect Dis* **23**, e460-e461 (2023). https://doi.org:10.1016/S1473-3099(23)00575-3
34 Yang, S. *et al.* Fast evolution of SARS-CoV-2 BA.2.86 to JN.1 under heavy immune pressure. *Lancet Infect Dis* **24**, e70-e72 (2024). https://doi.org:10.1016/S1473-3099(23)00744-2
35 Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* **28**, 659-669 (2013). https://doi.org:10.1016/j.tree.2013.08.003
36 Kupferschmidt, K. *'We made a mistake.' Omicron origin study retracted after widespread criticism*, <https://www.science.org/content/article/we-made-mistake-omicron-origin-study-retracted-after-widespread-criticism#:~:text=Omicron%20was%20first%20discovered%20in,the%20variants%20circulating%20before%20it> (2022).
37 Graf, T. *et al.* Identification of a novel SARS-CoV-2 P.1 sub-lineage in Brazil provides new insights about the mechanisms of emergence of variants of concern. *Virus Evol* **7**, veab091 (2021). https://doi.org:10.1093/ve/veab091
38 Naveca, F. G. *et al.* COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nat Med* **27**, 1230-1238 (2021). https://doi.org:10.1038/s41591-021-01378-7
39 Hill, V. *et al.* The origins and molecular evolution of SARS-CoV-2 lineage B.1.1.7 in the UK. *Virus Evol* **8**, veac080 (2022). https://doi.org:10.1093/ve/veac080
40 Layan, M. *et al.* Impact and mitigation of sampling bias to determine viral spread: Evaluating discrete phylogeography through CTMC modeling and structured coalescent model approximations. *Virus Evol* **9**, vead010 (2023). https://doi.org:10.1093/ve/vead010
41 Cheng, Y. *et al.* covSampler: A subsampling method with balanced genetic diversity for large-scale SARS-CoV-2 genome data sets. *Virus Evol* **8**, veac071 (2022). https://doi.org:10.1093/ve/veac071