

1 **Genomic epidemiology of SARS-CoV-2 within households in coastal** 2 **Kenya: a case ascertained cohort study**

3 4 **Authors**

5 Charles N. Agoti^{1,2,*}, Katherine E. Gallagher^{1,3}, Joyce Nyiro¹, Arnold W. Lambisia¹,
6 Nickson Murunga¹, Khadija Said Mohammed¹, Leonard Ndwiga¹, John M. Morobe¹,
7 Maureen W. Mburu¹, Edidah M. Ongera¹, Timothy O. Makori¹, My V.T. Phan⁴, Matthew
8 Cotten^{4,5}, Lynette Isabella Ochola-Oyier¹, Simon Dellicour^{6,7}, Philip Bejon^{1,8}, George
9 Githinji^{1,2}, and D. James Nokes^{1,9}

10 11 **Affiliations**

12 ¹ Kenya Medical Research Institute (KEMRI)-Wellcome Trust Research Programme
13 (KWTRP), Kilifi, Kenya

14 ² Pwani University, Kilifi, Kenya

15 ³ London School of Hygiene & Tropical Medicine, United Kingdom

16 ⁴ MRC/UVRI & London School of Hygiene and Tropical Medicine, Entebbe, Uganda.

17 ⁵ MRC-University of Glasgow, Centre for Virus Research Glasgow, United Kingdom

18 ⁶ Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, Bruxelles, Belgium

19 ⁷ Department of Microbiology, Immunology and Transplantation, Rega Institute,
20 Laboratory for Clinical and Epidemiological Virology, KU Leuven, University of Leuven,
21 Leuven, Belgium

22 ⁸ Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom

23 ⁹ School of Life Sciences and Zeeman Institute for Systems Biology and Infectious
24 Disease Epidemiology Research (SBIDER), University of Warwick, Coventry, United
25 Kingdom

26 27 *** Correspondence**

28 Dr Charles N. Agoti,
29 Epidemiology and Demography Department,
30 KEMRI-Wellcome Trust Research Programme,
31 P.O. Box 230, Kilifi – 80108,

32 Kenya.

33 Email: cnyaigoti@kemri-wellcome.org

34

35 **Article type**

36 Research Article

37

38 **Key words**

39 SARS-CoV-2, COVID-19, household, transmission, genomics, Kenya

40 **Abstract (221 words)**

41 Analysis of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomic
42 sequences from household infections may provide useful epidemiological information
43 for future control measures. Between December 2020 and July 2022, we conducted a
44 case-ascertained household cohort study whereby households were recruited if a
45 member was either a SARS-CoV-2 case or contact of a confirmed case. A total of 765
46 members of 214 households were prospectively monitored for SARS-CoV-2 infection
47 and transmission. Follow-up visits collected a nasopharyngeal/oropharyngeal (NP/OP)
48 swab on days 1, 4 and 7 for RT-PCR diagnosis. If any of these were positive, further
49 swabs were collected on days 10, 14, 21 and 28. Of 2,780 NP/OP swabs collected, 540
50 (19.4%) tested SARS-CoV-2 positive and viral genome sequences were recovered for
51 288 (53.3%) positive samples. The genomes belonged to 23 different Pango lineages.
52 Phylogenetic analysis including contemporaneous Coastal Kenya data estimated 233
53 putative transmission events involving 162 members of the 89 households, of which 60
54 (25%) were intra-household transmission events while 173 (75%) were infections that
55 likely occurred outside the households. In 34 (38%) households, multiple virus
56 introductions were observed (up to six) within the one-month follow-up period, in
57 contrast to high-income settings, where a single introduction seemed to occur during
58 epidemic waves. Our findings suggests that in this setting control of respiratory virus
59 spread by household member isolation will be ineffective.

60 **Introduction**

61 Households are a fundamental unit of social structure and the frequent scene of
62 respiratory pathogen transmission including for severe acute respiratory syndrome
63 coronavirus 2 (SARS-CoV-2), the aetiological agent of coronavirus disease 2019
64 (COVID-19) ^{1,2}. The household secondary attack rate (i.e., the probability of infection of
65 susceptible household members directly from the first case in the household) for SARS-
66 CoV-2 has been estimated to be about 21.1% (95%CI: 17.4-24.8%) with considerable
67 heterogeneity observed over geographic regions and time periods ³⁻⁶. An improved
68 understanding of SARS-CoV-2 household transmission, including the frequency of virus
69 transmission within a household compared to transmission from outside into the
70 household, may help refine local control measures.

71
72 As of August 2022, Kenya had experienced six waves of SARS-CoV-2 infections ⁷. In
73 this period, Kenya's COVID-19 countermeasures as measured by the Oxford Stringency
74 Index (SI), a measure based on nine key indicators rescaled from 0-100, fluctuated
75 between 36 and 89⁸ in response to the perceived threat. The different infection waves
76 were dominated by distinct virus variants: the first two by the early ancestral virus
77 lineage (mainly B.1), the third by the Alpha (B.1.1.7) variant of concern (VOC), the
78 fourth by Delta VOC (B.1.617.2), the fifth by Omicron VOC (BA.1), and the sixth by
79 Omicron sub-variants BA.4/5 ⁹. Local epidemiological studies suggest many SARS-
80 CoV-2 infections have been asymptomatic or mild ¹⁰, as evidenced by the high
81 seroprevalence (48.5% by March 2021 ¹¹) without high numbers of hospitalisations or
82 excess mortality ¹². However, confirmed active SARS-CoV-2 infections by the Kenyan
83 Ministry of Health (MoH) have been reported in less than 1% of the Kenyan population
84 indicating inadequate testing. Consequently, gaps exist in our quantitative
85 documentation of SARS-CoV-2 circulation within the Kenyan population.

86
87 To date, SARS-CoV-2 genomic analysis has played a key role in elucidating COVID-19
88 pandemic transmission patterns ¹³⁻¹⁷. For instance, genomic analysis has helped
89 uncover the nature of virus seeding into close living environments like, hospitals ^{18,19},
90 prisons²⁰, cruise ships ²¹, long-term care facilities ²², or learning institutions ²³, and has

91 also uncovered a number of superspreading events^{19,24}. Unlike many RNA viruses,
92 SARS-CoV-2 replication is believed to be under some level of proof-reading²⁵, limiting
93 its substitution rate (estimated at 9.90×10^{-4} nucleotide substitutions/site/year)²⁶, a
94 parameter critical in applying genomics to understand short-term epidemiological
95 dynamics.

96
97 Few studies have examined SARS-CoV-2 genomic diversity in putative within-
98 household transmission cases. In Ireland, Hare *et al*²⁷ found that most family members
99 testing positive had indistinguishable consensus genome sequences from other family
100 members and the early presumed index case had a divergent sub-lineage. In the
101 present study, we document SARS-CoV-2 transmission patterns within households in
102 coastal Kenya by analysing infections identified in a case-ascertained cohort (i.e.
103 households enrolled if a member was confirmed case of SARS-CoV-2 or a contact of
104 one) during successive local waves of infections²⁸. We undertook detailed genomic
105 analyses to identify patterns of SARS-CoV-2 introductions into households and to
106 document the frequency and patterns of infection spread within households in coastal
107 Kenya.

108

109 **Results**

110 **Baseline characteristics**

111 A total of 765 participants from 214 households were recruited between 10th December
112 2020 and 29th July 2022. From these, 2,780 nasopharyngeal/oropharyngeal (NP/OP)
113 swabs were collected, 540 (19.4%) of which tested SARS-CoV-2 positive by RT-PCR
114 (**Table 1**). The positive swabs were from 254 infected participants in 119 households.
115 The temporal distribution of the swab collections for all the 214 households and their
116 RT-PCR testing results are shown in **S1 Fig**. The participants with a positive swab had
117 a median age of 27 years (IQR: 12.0-46.0; **Table 1**), with 164 (64.6%) females.
118 Compared to participants who remained SARS-CoV-2 negative during the follow-up
119 period, positive cases were more likely to report at least one acute respiratory illness
120 (ARI) symptom (67.7% vs 32.5%; $p < 0.001$, chi-squared test). The household
121 recruitments were predominantly coincidental with the national COVID-19 waves 3, 4, 5

122 and 6 (**Fig. 1A and B**) with only one household recruited during wave 2. The changes in
123 the national countermeasures during the study period as estimated by the Oxford SI is
124 provided in **Fig. 1C**.

125

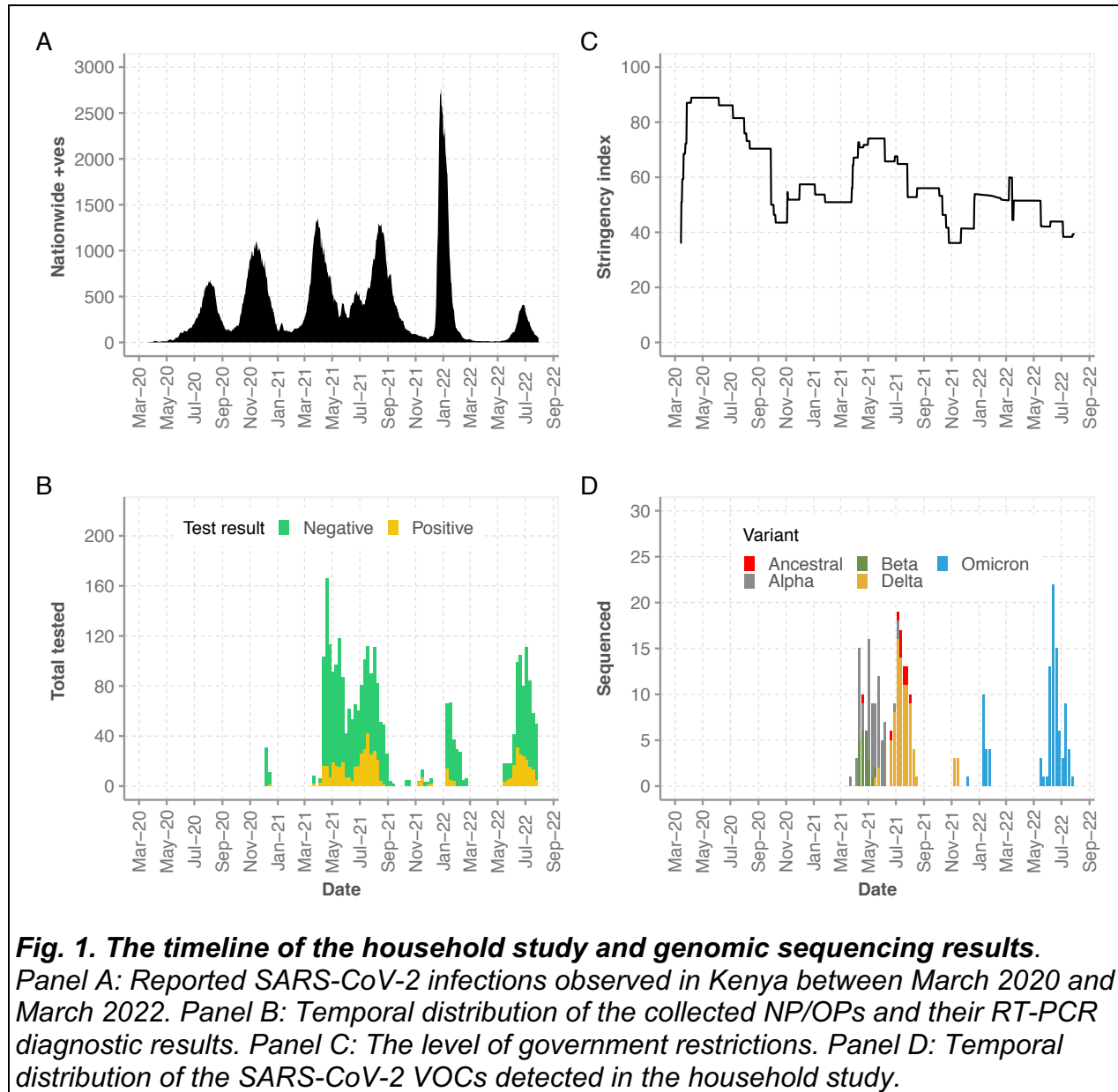
126 **Table 1. Demographic characteristics of study participants.**

Characteristic	Positive	Negative	Total
NP/OP swabs	540	2240	2780
Recruited households	119	95	214
Recruited participants	254	511	765
Sex			
female (%)	164 (64.6)	297 (58.1)	461 (60.3)
male (%)	90 (35.4)	214 (41.9)	304 (39.7)
Age distribution (years)			
Median (IQR)	27.0 (12.0-46.0)	20.0 (10.0-40.0)	24.0 (11.0-44.0)
Age categories			
0-4 y (%)	31 (12.2)	54 (10.6)	85 (11.1)
5-9 y (%)	23 (9.1)	54 (10.6)	77 (10.1)
10-19 y (%)	48 (18.9)	133 (26.0)	181 (23.7)
20-39 y (%)	73 (28.7)	113 (28.6)	219 (28.6)
40-64 y (%)	63 (24.8)	92 (18.0)	155 (20.3)
65+ y (%)	14 (5.5)	31 (6.1)	45 (5.9)
Unknown	2 (0.8)	1 (0.2)	3 (0.4)
Symptom status			
asymptomatic	82 (32.3)	345 (67.5)	427 (55.8)
symptomatic	172 (67.7)	166 (32.5)	338 (44.2)

127 IQR: Interquartile Range, y: years

128

129



130

131 **Fig. 1. The timeline of the household study and genomic sequencing results.**
132 *Panel A: Reported SARS-CoV-2 infections observed in Kenya between March 2020 and*
133 *March 2022. Panel B: Temporal distribution of the collected NP/OPs and their RT-PCR*
134 *diagnostic results. Panel C: The level of government restrictions. Panel D: Temporal*
135 *distribution of the SARS-CoV-2 VOCs detected in the household study.*

136

137

138

139 **Genomic sequencing and lineage/VOC classification**

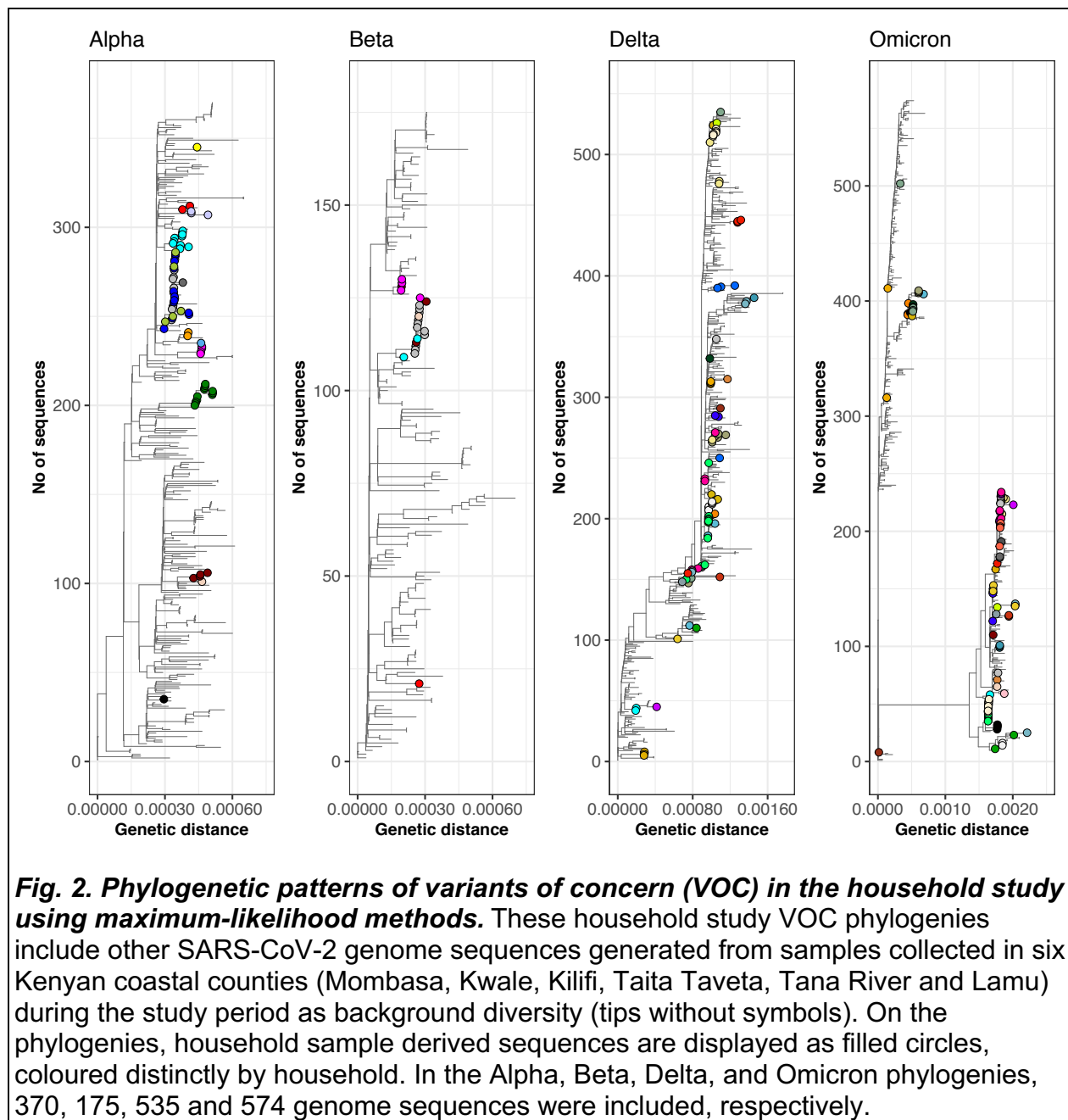
140 Near complete genomes (over 80% coverage) were obtained from 288 samples (53.3%
141 of positive samples) collected from 162 participants from 89 households. The samples
142 that failed sequencing (n = 252) had high Ct values (corresponding to low viral load;
143 >33.0; **S2. Fig.**) or yielded low quality PCR products during library preparation. The 288
144 genomes were classified into non-VOC (Pango lineage B.1; n = 11), Alpha VOC (n =
145 70), Beta VOC (n = 22), Delta VOC (n = 88) and Omicron VOC (n = 97; **Fig.1D**). All
146 Alpha and Beta sequences fell within lineage B.1.1.7 and B.1.351, respectively. Within
147 the Delta VOC, six Pango lineages were identified while within the Omicron VOC 14
148 Pango lineages were identified. The lineages within Omicron VOC included those that
149 classified under BA.1, BA.2, BA.4 and BA.5 sub-variants. A summary of the temporal
150 distribution of the 23 Pango lineages that were identified across the sequenced cases
151 and their history of global detection is presented in **S2 Fig. and S1 Table**.

152

153 **Phylogenetic clustering of the household study genomes**

154 To investigate the genetic diversity in the household study genome sequences, we
155 reconstructed a maximum likelihood (ML) phylogeny that included background coastal
156 Kenya co-circulating viruses (n = 2,555). As expected, the household genome
157 sequences clustered by VOC and Pango lineages with other Kenyan coastal sequences
158 (**S3 Fig.**). Notably, lineage B.1 sequences were found in multiple branches of the
159 phylogeny, including some at the base of branches leading to Beta and Delta VOCs. To
160 assess the genetic relatedness of the recovered genomes within and between
161 households, we reconstructed VOC-specific phylogenies with tips coloured by the
162 household of sampling (**Fig. 2**). Here we observed both intra- and inter-household
163 clustering (i.e., multiple sequences of a single clade identified in one household or
164 shared between two or more households). In a few households, sequences of a single
165 VOC but distinct clades were observed, indicating multiple distinct virus introductions
166 into the same household (**Fig. 2**).

167



168

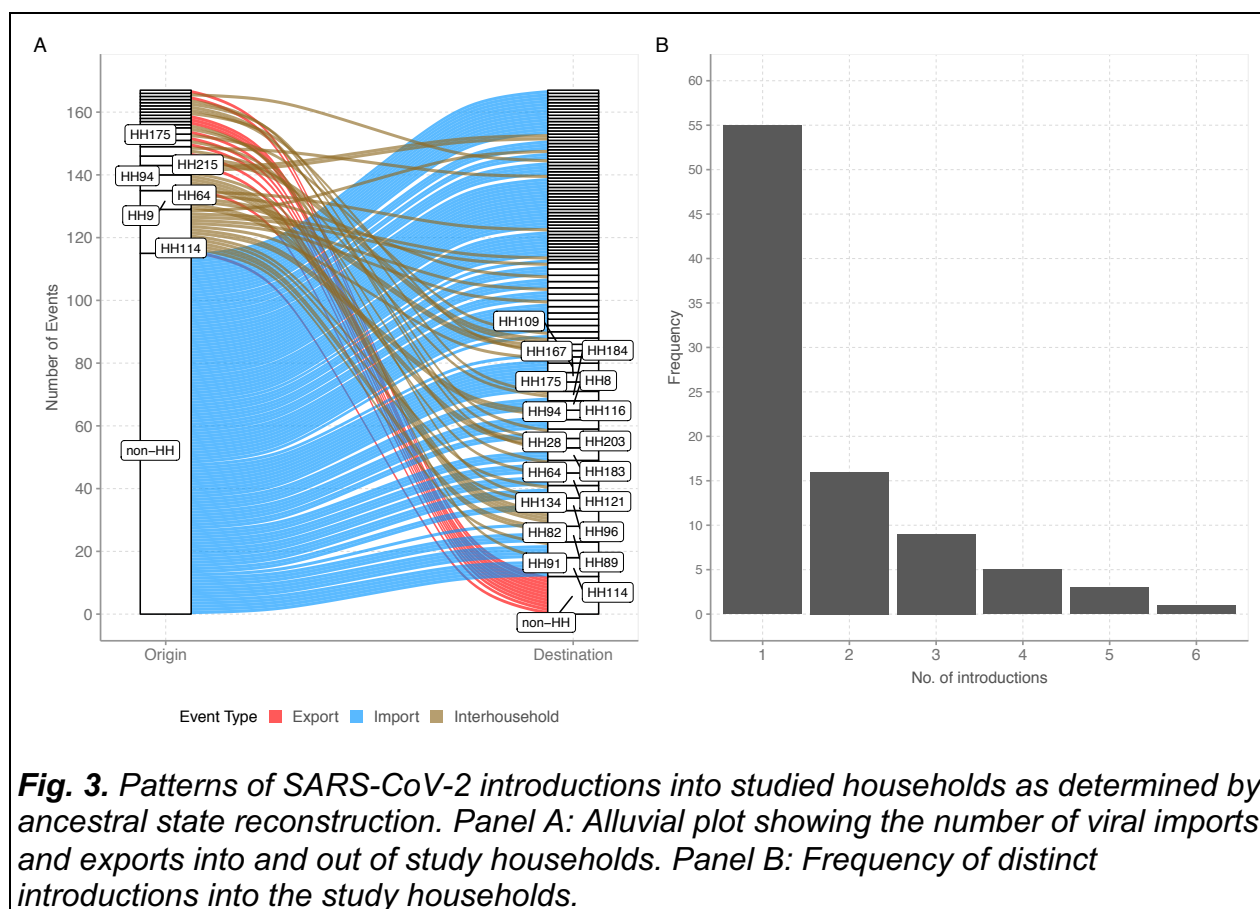
169 **Fig. 2. Phylogenetic patterns of variants of concern (VOC) in the household study**
170 **using maximum-likelihood methods.** These household study VOC phylogenies
171 include other SARS-CoV-2 genome sequences generated from samples collected in six
172 Kenyan coastal counties (Mombasa, Kwale, Kilifi, Taita Taveta, Tana River and Lamu)
173 during the study period as background diversity (tips without symbols). On the
174 phylogenies, household sample derived sequences are displayed as filled circles,
175 coloured distinctly by household. In the Alpha, Beta, Delta, and Omicron phylogenies,
176 370, 175, 535 and 574 genome sequences were included, respectively.

177

178 **Estimating the number of introductions into the households**

179 SARS-CoV-2 has been reported to have a genomic evolutionary rate of ~2 nucleotide
180 substitutions per genome per month²⁹. A heterogenous distribution of the pairwise
181 nucleotide differences of specimens identified in the same household was observed
182 (range 0-63; median 0.0, mean 2.8; **S4 Fig.**). More than two nucleotide differences
183 between genomes from the same household were observed in 25 households, implying

184 possible multiple introductions into these households. We further investigated the
185 number of virus introductions into each household using ancestral state reconstruction
186 (ASR) along the dated ML phylogeny^{16,30}. A total of 155 virus introductions were
187 predicted into the 89 households where we recovered SARS-CoV-2 sequence data. On
188 classifying the transition events by sequence origin ("non-household" events - those
189 individuals not from same household - and "household" events - those where both
190 involved individuals were members of same household), we found that most transitions
191 were "non-household" compared to "household" transition events (75% vs. 25%; **Fig.**
192 **3A**). Overall, we estimated that a single virus introduction occurred for 55 households
193 (61%), two introductions for 16 households (17%), three introductions for nine
194 households (10%), four introductions for five households (6%), five introductions for
195 three households (3%) and six introductions in one household (**Fig. 3B**).
196



197
198
199
200
201
202

203 Discussion

204 We provide evidence of frequent multiple introductions of SARS-CoV-2 into rural coastal
205 Kenyan households within a month, a finding that was unexpected. The conventional
206 view has been that households with concurrently infected members most likely acquire
207 the infection introduced through a single index case. This assumption has been
208 repeatedly supported by several genomic studies. For instance, a phylogenetic analysis
209 performed in a Dutch study following 85 households showed a single introduction into
210 all the studied households ⁴. By contrast, in the present study, about one-third of
211 infected households (38%) had two or more introduction events within a distinct
212 epidemic wave.

213
214 A variety of factors may explain the differences in household SARS-CoV-2 introduction
215 patterns identified in our study compared to previous observations. First, in our setting,
216 multiple families may live in one compound and eat from one kitchen and these were
217 defined as a single household in this study ³¹. Such living arrangement results in
218 relatively larger household sizes that may be associated with higher chances of multiple
219 virus introductions, especially at the peak of epidemic waves. Second, the dominance
220 of informal sector jobs in this setting may have made options like working from home
221 difficult to implement. As result the potential for acquiring infection outside the
222 household setting was high relative to high-income settings.

223
224 Our study followed-up participants for a period of up to 1 month with serial sample
225 collection and recovered genomes were analysed in the context of contemporaneous
226 locally circulating diversity in coastal Kenya ³². Despite observing minimal nucleotide
227 variation between samples from members of the same household infection clusters,
228 when we incorporated sampling dates through the ASR analysis, we were able to
229 partially reconstruct potential within-household transmission events. This allowed the
230 identification of multiple introductions into the households of closely related viruses,
231 observation of putative within-household transmission and short-interval reinfections.

232

233 Few studies have examined SARS-CoV-2 household transmission dynamics within
234 Africa³³⁻³⁵, and these have resulted in diverse findings. In rural Egypt, a 6-month study
235 reported a SAR of 89.8%³³, in South Africa a 13-month study reported a 25% infection
236 rate among vulnerable household contacts³⁴, and in Madagascar, a SAR of 38.8%
237 (CI:19.5-57.2)³⁶ was reported. None of these studies included a viral genomic analysis
238 to support any conclusions that the inferred household transmission clusters were
239 epidemiologically linked and arose from a single index case.

240
241 The Kenyan government countermeasures in place during the study period may have
242 had an impact on the spread of SARS-CoV-2 within the study households. In June
243 2020, the Kenyan government announced guidelines for home-based care for
244 asymptomatic or mildly symptomatic patients without co-morbidities³⁷. Kenya started
245 vaccinating its population in March 2021, but the coverage was low (<15%) during the
246 study period³⁸, and thus it is unlikely that it affected transmission during our study. The
247 stringency index in the country during the study period fluctuated from 35% to 75%.
248 However, we did not detect variation in the pattern of introductions over time, which
249 could suggest that the various restrictions had minimal impact at the household level.
250 However, concluding on this would likely require more advanced epidemiological
251 modelling.

252
253 This study had several limitations. First, several positive NP/OP samples (46.7%) failed
254 to yield viral genome sequences or had large genome sequencing gaps due to PCR
255 amplicon drop-off. With such a high level of genome data missingness, the overall
256 phylogenetic signal was reduced in trying to establish who infected whom or
257 directionality of transmission. Second, we cannot rule out that a few of the sequence
258 changes observed could be sequencing or assembly artifacts³⁹. Third, the case-
259 ascertained study design we used had the drawback that by the time the first sample
260 was collected, multiple positive cases had already occurred in most households. Most of
261 the index cases were recruited following presentation to a health facility with ARI. This
262 complicated our effort of fully determining who infected whom back in the household. To
263 overcome this challenge, future studies should observe members before entry of the

264 virus into households and genomic data co-analysed with other relevant epidemiological
265 data ⁴⁰. Fourth, our sampling interval, especially after week two, may have missed
266 persons who had been positive for less than the 7 days sample collection interval.
267 Higher density sampling has previously been associated with a higher attack rate⁴.
268 Finally, we did not bring in other datatypes like, contact patterns, symptoms and shared
269 intra-host variation ⁴¹, which would have provided further insights into potential
270 transmission linkages. An analysis of secondary attack rates is being conducted that
271 includes all data, including serological evidence of infection, which is not included here.

272
273 In conclusion, we identified an unusually high number of independent virus introductions
274 into households in coastal Kenya during household temporally clustered infections. Our
275 findings suggests that control of SARS-CoV-2 transmission by household member
276 isolation alone may not stop community transmission in this setting. Our study further
277 highlights the importance of examining genomic data for accurate estimation and
278 interpretation of SARS-CoV-2 household epidemiological parameters.

279

280 **Methods**

281 **Study design and recruitment**

282 We conducted a case-ascertained study in coastal Kenya, where new households were
283 recruited into the study via five local health facilities or the Kilifi County Department of
284 Health rapid response team (RRTs). Households were defined as dwellings or groups of
285 dwellings that share the same kitchen or cooking space (**S2 Table**). Many of the
286 recruited households were from within the Kilifi Health and Demographic Surveillance
287 System (KHDSS) area located in Kilifi County, Coastal Kenya ⁴². To get enrolled, a
288 household needed to have at least two occupants, to be accessible by road, and to
289 obtain permission from the household head. In the beginning, only households with a
290 member who was a contact of a confirmed case from a different household were
291 recruited, but due to slow enrolment, we revised the protocol to include households with
292 confirmed cases at first sampling. A household was exempted if at the time of
293 recruitment: two or more members had already developed COVID-19 symptoms (e.g.

294 fever, sore throat, cough etc), a member had been hospitalized due to COVID-19, or the
295 household had been enrolled in a trial of therapeutic COVID-19 product.

296

297 **Follow-up**

298 During each household visit, a NP/OP swab was obtained from all participants and
299 transported in cool boxes with ice packs to KEMRI-Wellcome Trust Research
300 Programme (KWTRP) laboratories within 48 hr for real-time RT-PCR testing. The study
301 had broadly two follow-up arms: "reduced follow-up" and "intense follow-up".

302 Households in the "reduced follow-up" arm were those where all the members tested
303 SARS-CoV-2 negative at day 1, 4, and 7. Therefore, in that case, follow-up was
304 discontinued henceforth with a few exceptions (**S1 Fig.**). The "intense follow-up arm"
305 was activated when a household member tested positive on day 1, 4, or 7. In that case,
306 the household was sampled again on day 10, 14, 21, and 28. Data on baseline
307 household and demographic characteristics were collected by the study team at
308 enrolment. During all households' visits, data on presence of ARI symptoms (e.g., fever,
309 cough, runny nose, sore throat, headache) were collected.

310

311 **Laboratory procedures**

312 *SARS-CoV-2 diagnosis*

313 SARS-CoV-2 testing of study samples was undertaken alongside samples collected in
314 six coastal counties of Kenya as part of the national COVID-19 testing as previously
315 described¹⁰. Four different viral RNA extraction kits (namely, QIAamp Viral RNA Mini
316 Kit, RNeasy® QIAcube® HT Kit, TIANamp Virus RNA Kit and Da An Gene Nucleic
317 acid Isolation and Purification Kit) were deployed in combination with five different RT-
318 PCR kits/protocol (namely, Da An Gene Co. detection Kit, European Virus Archive-
319 Global (EVAg) E gene protocol, Standard M Kit, Sansure Biotech Novel Coronavirus
320 (2019-nCoV) Nucleic Acid Diagnostic Real-time RT-PCR kit¹⁰). Positives were
321 determined using the kit/protocol-defined cycle thresholds (Ct). In kits where multiple
322 SARS-CoV-2 genomic regions were targeted, the average cycle threshold (Ct) was
323 calculated from the individual Cts.

324

325 *Genome sequencing*

326 We aimed to sequence whole genomes of all the RT-PCR positive samples with a cycle
327 threshold of < 33.0. Viral RNA was re-extracted from the specimens using QIAamp viral
328 RNA mini-Kit following the manufacturer's instructions and converted to cDNA using
329 Lunascript kit with ARTIC protocol primers⁴³. Genome amplification was conducted
330 using Q5 PCR kit and ARTIC protocol primers (initially v3 and then v4). Sequencing
331 libraries were prepared using Oxford Nanopore Technologies (ONT) ligation sequencing
332 kit SQK-LSK109 and the ONT Native Barcoding Expansion kit as described in the
333 ARTIC protocol⁴³. Sequencing was performed on Oxford Nanopore Technologies'
334 MinION or GridION devices using R9.4.1 flow cells.

335

336 **Bioinformatic analysis**

337 *Genome assembly and lineage assignment*

338 The ONT's raw sequencing reads (FAST5) were base-called and demultiplexed using
339 ONT's Guppy v3.5-4.2. The resultant files (FASTQ) were assembled into consensus
340 genomes using ARTIC bioinformatic pipeline reference-based approach
341 (<https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>; last accessed 2022-
342 09-17). Only nucleotides with a read depth of more than $\times 20$ were included into the
343 consensus sequence. Only sequences with >80% genome coverage were further
344 analysed. The genomes were assigned Pango lineages using the command-line
345 installation of pangolin v4.1.3, PUSHER-v1.3, scorpio v0.3.16 and constellation v0.1.6
346 ^{44,45}.

347

348 *Phylogenetic analysis*

349 Multiple sequence alignments were generated using Nextalign v.1.10.1 referenced-
350 based aligner within the Nextclade tool v0.14.2⁴⁶ using the command:

351

352 *nextalign -r NC_045512.fasta -r <input_file.fasta>*

353

354 Alignments were visualized using a custom Python script and "snipit" tool
355 (<https://github.com/aineniamh/snipit>; last accessed 2022-05-20). Pairwise distances

356 were calculated using pairsnp.py (<https://github.com/gtonkinhill/pairsnp/>; last accessed
357 2022-05-20) using the command:

358

```
359     pairsnp -f <input_file.fasta> --csv --output <output_file.fasta>
```

360

361 Phylogenetic relationships between all recovered genomes and between viruses
362 classified under the same VOC were obtained through the inference of maximum
363 likelihood (ML) trees performed with the program IQTREE v2.1.3 with a general time
364 reversible (GTR) substitution model using the command:

365

```
366     iqtree -s <input_alignment.fasta> -nt <no_of_cores> -m GTR
```

367

368 We included contemporaneous genomes from the six coastal Kenya counties
369 (Mombasa, Kilifi, Kwale, Taita Taveta, Tana River, and Lamu) that were sequenced as
370 part of the national SARS-CoV-2 genomic surveillance to provide phylogenetic context
371 to the household study genomes. Each ML tree was subsequently time-calibrated using
372 the program TreeTime, assuming a constant evolutionary rate of SARS-CoV-2 genome
373 of 8.0×10^{-4} using the command:

374

```
375     treetime -tre <tree_file> --aln <alignment_file> --clock-rate 0.00084 --dates  
376     <metadata_file.csv>
```

377

378 The phylogenetic trees were combined with metadata and visualized with the R
379 package "ggtree" v2.4.2⁴⁷.

380

381 *Virus introductions and putative transmissions*

382 The number of independent virus introductions into the households was investigated
383 using two approaches.

384

385 *i. Pairwise nucleotide substitution.*

386 Here, we compared observed nucleotide differences between pairs of household
387 genomes with the number of nucleotide mutations expected over the time interval
388 between the two sampling dates. The pairwise differences between the household study
389 genomes were computed using the program pairsnp
390 (<https://github.com/gtonkinhill/pairsnp>; accessed 06-Feb-2023).

391
392 *ii. Ancestral state reconstruction (ASR)*
393 ASR approach was used to identify the introduction events into a household and count
394 the transitions between household members¹⁶. This was performed along the time-
395 scaled phylogeny obtained for each VOC. To infer the number of introduction events
396 into each household, a variable "source" was generated for which the sequences were
397 assigned the household ID or noted as "non-household". Using the dated phylogeny,
398 the nucleotide sequence alignment, the "source" metadata file, a migration analysis
399 was run using TreeTime⁴⁸, using the command:

400
401 `treetime migration --tre <dated_tree_file> --states <metadata_file.csv> --attributes`
402 `<source_attribute>`

403
404 To infer within household transition events, a variable "participant" was generated for
405 which the sequences were assigned the participant ID or noted as "non-household".
406 The migration analysis was rerun using the dated phylogeny and transitions between
407 members of the same households noted using python scripts described in^{16,30}.

408 **Statistical analysis**

409
410 Summary statistics were computed for key demographic characteristics including mean,
411 median, standard deviation as appropriate. Infection prevalence was expressed using
412 proportions and comparison between groups included appropriate statistical tests (e.g.,
413 chi-square or Fisher's exact). All statistical analyses were performed in R version 4.0.5.

414

415 **Ethical consideration**

416 The study protocol was reviewed and approved by both the Scientific and Ethics
417 Research Unit (SERU) at Kenya Medical Research Institute (KEMRI), Nairobi, Kenya
418 (REF: SERU protocol # 4077) and the University of Warwick, Biomedical and Scientific
419 Research Ethics Committee, Coventry, United Kingdom (REF: BSREC 150/19-20
420 AM01). Prior to baseline data and biospecimen collection, written informed consent was
421 obtained from all participants aged 18 years or older, while for participants aged less
422 than 18 years consent was obtained from their parents or legal guardians. Assent was
423 also sought for adolescents (13-17 years of age).

424

425

426 **Data availability**

427 The consensus genome sequences obtained in this study that passed our quality
428 control filters have been submitted to both GISAID and GenBank database (accession
429 numbers available in appendix pages of the supplementary material). The raw data files
430 have been prepared for deposition in Harvard DataVerse. For more detailed information
431 beyond the metadata used in the paper, there is a process of managed access requiring
432 submission of a request form for consideration by our Data Governance Committee
433 (http://kemri-wellcome.org/about-us/#ChildVerticalTab_15).

434

435 **Code availability**

436 The code for the analyses presented in this manuscript is available from Github using
437 the link https://github.com/cnyaigoti/Household_study_2021-2022.

438

439 **Acknowledgements**

440 We thank (a) the members of the Kilifi County rapid response team who worked with our
441 field study team in collecting the samples analysed here; (b) the members of the
442 COVID-19 KWTRP Testing Team who undertook real-time RT-PCR processing of the
443 samples received at KWTRP to identify positives (see full list of members below).

444

445 This work was supported by the National Institute for Health and Care Research (NIHR)
446 (project reference 17/63/82) using UK aid from the UK Government to support global
447 health research, The UK Foreign, Commonwealth and Development Office and
448 Wellcome Trust (grant# 220985). Members of COVID-19 Testing Team at KWTRP are
449 supported by multiple funding sources including UNITAD (BOHEMIA study received by
450 Dr Marta Maia funded UNITAID), EDCTP (Senior Fellowship and Research and
451 Innovation Action (RIA) grants received by Dr Francis Ndungu), GAVI (PCIVS grant
452 received by Prof. Anthony Scott). Dr Simon Dellicour acknowledges support from the
453 *Fonds National de la Recherche Scientifique* (F.R.S.-FNRS, Belgium; grant
454 n°F.4515.22), from the Research Foundation - Flanders (*Fonds voor Wetenschappelijk*
455 *Onderzoek-Vlaanderen*, FWO, Belgium; grant n°G098321N), and from the European
456 Union Horizon 2020 project MOOD (grant agreement n°874850). MP and MC were
457 supported by the Wellcome Trust and FCDO - Wellcome Epidemic Preparedness –
458 Coronavirus (AFRICO19, grant agreement number 220977/Z/20/Z). The views
459 expressed in this publication are those of the author (s) and not necessarily those of
460 NIHR, the Department of Health and Social Care, Foreign Commonwealth and
461 Development Office, Wellcome Trust or the UK government.

462

463 **List of members of the COVID-19 Testing Team at KWTRP**

464 Agnes Mutiso, Alfred Mwanzu, Angela Karani, Bonface M. Gichuki, Boniface Kaaria,
465 Brian Bartilol, Brian Tawa, Calleb Odundo, Caroline Ngetsa, Clement Lewa, Daisy
466 Mugo, David Amadi, David Ileri, Debra Riako, Domtila Kimani, Donwilliams Omuoyo,
467 Edwin Machanja, Elijah Gicheru, Elisha Omer, Faith Gambo, Horace Gumba, Isaac
468 Musungu, James Chemweno, Janet Thoya, Jedida Mwacharo, Jennifer Musyoki, John
469 Gitonga, Johnstone Makale, Justine Getonto, Kelly Ominde, Kelvias Keter, Lydia
470 Nyamako, Margaret Nunah, Martin Mutunga, Metrine Tendwa, Moses Mosobo, Nelson
471 Ouma, Nicole Achieng, Patience Kiyuka, Perpetual Wanjiku, Peter Mwaura, Rita Warui,
472 Robinson Cheruiyot, Salim Mwarumba, Shaban Mwangi, Shadrack Mutua, Susan
473 Njuguna, Victor Osoi, Wesley Cheruiyot, Wilfred Nyamu, Wilson Gumbi and Yiakon
474 Sein.

475

476

477 **Role of funding source**

478 The funders of the study had no role in study design, data collection, data analysis ,
479 data interpretation or writing of the report.

480

481 **Author's contributions**

482 The project was conceived and designed by CNA, KEG, JUN, MC, and DJN; Laboratory
483 processing of specimens was conducted by JUN, LIO, NM, AWL, KSM, LN, JMM,
484 MWM, EMO, and TOM; Management and analysis of data were handled by AWL, CNA,
485 NM, SD and GG. CNA wrote the first draft; MP, MC, LIO, SD, PB, and DJN critically
486 reviewed the manuscript to produce the final draft.

487

488 References

- 489 1 Lee, E. C., Wada, N. I., Grabowski, M. K., Gurley, E. S. & Lessler, J. The engines of SARS-CoV-2
490 spread. *Science* **370**, 406-407 (2020). <https://doi.org/doi:10.1126/science.abd8755>
- 491 2 Park, Y. J. *et al.* Contact Tracing during Coronavirus Disease Outbreak, South Korea, 2020.
492 *Emerging Infectious Disease journal* **26**, 2465 (2020). <https://doi.org:10.3201/eid2610.201315>
- 493 3 Thompson, H. A. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) Setting-
494 specific Transmission Rates: A Systematic Review and Meta-analysis. *Clin Infect Dis* **73**, e754-
495 e764 (2021). <https://doi.org:10.1093/cid/ciab100>
- 496 4 Kolodziej, L. M. *et al.* High SARS-CoV-2 household transmission rates detected by dense saliva
497 sampling. *Clinical Infectious Diseases* (2022). <https://doi.org:10.1093/cid/ciac261>
- 498 5 Madewell, Z. J., Yang, Y., Longini, I. M., Jr, Halloran, M. E. & Dean, N. E. Household Secondary
499 Attack Rates of SARS-CoV-2 by Variant and Vaccination Status: An Updated Systematic Review
500 and Meta-analysis. *JAMA Network Open* **5**, e229317-e229317 (2022).
501 <https://doi.org:10.1001/jamanetworkopen.2022.9317>
- 502 6 Jørgensen, S. B., Nygård, K., Kacelnik, O. & Telle, K. Secondary Attack Rates for Omicron and
503 Delta Variants of SARS-CoV-2 in Norwegian Households. *JAMA* **327**, 1610-1611 (2022).
504 <https://doi.org:10.1001/jama.2022.3780>
- 505 7 Tegally, H. *et al.* The evolving SARS-CoV-2 epidemic in Africa: Insights from rapidly expanding
506 genomic surveillance. *Science*, eabq5358 (2022). <https://doi.org:10.1126/science.abq5358>
- 507 8 Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government
508 Response Tracker). *Nat Hum Behav* **5**, 529-538 (2021). <https://doi.org:10.1038/s41562-021-01079-8>
- 509 9 Nasimiyu, C. *et al.* Imported SARS-CoV-2 Variants of Concern Drove Spread of Infections across
510 Kenya during the Second Year of the Pandemic. *COVID* **2**, 586-598 (2022).
- 511 10 Nyagwange, J. *et al.* Epidemiology of COVID-19 infections on routine polymerase chain reaction
512 (PCR) and serology testing in Coastal Kenya. *Wellcome Open Res* **7**, 69 (2022).
513 <https://doi.org:10.12688/wellcomeopenres.17661.1>
- 514 11 Uyoga, S. *et al.* Prevalence of SARS-CoV-2 Antibodies From a National Serosurveillance of
515 Kenyan Blood Donors, January-March 2021. *Jama* **326**, 1436-1438 (2021).
516 <https://doi.org:10.1001/jama.2021.15265>
- 517 12 Otiende, M. *et al.* Impact of COVID-19 on mortality in coastal Kenya: a longitudinal open cohort
518 study. *medRxiv*, 2022.2010.2012.22281019 (2022).
519 <https://doi.org:10.1101/2022.10.12.22281019>
- 520 13 Li, J., Lai, S., Gao, G. F. & Shi, W. The emergence, genomic diversity and global spread of SARS-
521 CoV-2. *Nature* **600**, 408-418 (2021). <https://doi.org:10.1038/s41586-021-04188-6>
- 522 14 Bugembe, D. L. *et al.* Main Routes of Entry and Genomic Diversity of SARS-CoV-2, Uganda.
523 *Emerg Infect Dis* **26**, 2411-2415 (2020). <https://doi.org:10.3201/eid2610.202575>
- 524 15 Githinji, G. *et al.* Tracking the introduction and spread of SARS-CoV-2 in coastal Kenya. *Nat*
525 *Commun* **12**, 4809 (2021). <https://doi.org:10.1038/s41467-021-25137-x>
- 526 16 Wilkinson, E. *et al.* A year of genomic surveillance reveals how the SARS-CoV-2 pandemic
527 unfolded in Africa. *Science* **374**, 423-431 (2021). <https://doi.org:10.1126/science.abj4336>
- 528 17 Vöhringer, H. S. *et al.* Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature*
529 **600**, 506-511 (2021). <https://doi.org:10.1038/s41586-021-04069-y>
- 530 18 Ellingford, J. M. *et al.* Genomic and healthcare dynamics of nosocomial SARS-CoV-2
531 transmission. *Elife* **10** (2021). <https://doi.org:10.7554/eLife.65453>
- 532 19 Illingworth, C. J. *et al.* Superspreaders drive the largest outbreaks of hospital onset COVID-19
533 infections. *Elife* **10** (2021). <https://doi.org:10.7554/eLife.67308>
- 534

- 535 20 Hershov, R. B. *et al.* Rapid Spread of SARS-CoV-2 in a State Prison After Introduction by Newly
536 Transferred Incarcerated Persons - Wisconsin, August 14-October 22, 2020. *MMWR Morb*
537 *Mortal Wkly Rep* **70**, 478-482 (2021). [https://doi.org:10.15585/mmwr.mm7013a4](https://doi.org/10.15585/mmwr.mm7013a4)
538 21 Hoshino, K. *et al.* Transmission dynamics of SARS-CoV-2 on the Diamond Princess uncovered
539 using viral genome sequence analysis. *Gene* **779**, 145496 (2021).
540 [https://doi.org:10.1016/j.gene.2021.145496](https://doi.org/10.1016/j.gene.2021.145496)
541 22 Aggarwal, D. *et al.* The role of viral genomics in understanding COVID-19 outbreaks in long-term
542 care facilities. *Lancet Microbe* **3**, e151-e158 (2022). [https://doi.org:10.1016/s2666-](https://doi.org/10.1016/s2666-5247(21)00208-1)
543 [5247\(21\)00208-1](https://doi.org/10.1016/s2666-5247(21)00208-1)
544 23 Baumgarte, S. *et al.* Investigation of a Limited but Explosive COVID-19 Outbreak in a German
545 Secondary School. *Viruses* **14** (2022). [https://doi.org:10.3390/v14010087](https://doi.org/10.3390/v14010087)
546 24 Popa, A. *et al.* Genomic epidemiology of superspreading events in Austria reveals mutational
547 dynamics and transmission properties of SARS-CoV-2. *Sci Transl Med* **12** (2020).
548 [https://doi.org:10.1126/scitranslmed.abe2555](https://doi.org/10.1126/scitranslmed.abe2555)
549 25 Smith, E. C., Blanc, H., Surdel, M. C., Vignuzzi, M. & Denison, M. R. Coronaviruses lacking
550 exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and
551 potential therapeutics. *PLoS Pathog* **9**, e1003565 (2013).
552 [https://doi.org:10.1371/journal.ppat.1003565](https://doi.org/10.1371/journal.ppat.1003565)
553 26 Nie, Q. *et al.* Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Res* **287**, 198098
554 (2020). [https://doi.org:10.1016/j.virusres.2020.198098](https://doi.org/10.1016/j.virusres.2020.198098)
555 27 Hare, D. *et al.* Genomic epidemiological analysis of SARS-CoV-2 household transmission. *Access*
556 *Microbiol* **3**, 000252 (2021). [https://doi.org:10.1099/acmi.0.000252](https://doi.org/10.1099/acmi.0.000252)
557 28 Githinji, G. *et al.* The genomic epidemiology of SARS-CoV-2 variants of concern in Kenya.
558 *medRxiv*, 2022.2010.2026.22281446 (2022). [https://doi.org:10.1101/2022.10.26.22281446](https://doi.org/10.1101/2022.10.26.22281446)
559 29 Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus*
560 *Evolution* **6** (2020). [https://doi.org:10.1093/ve/veaa061](https://doi.org/10.1093/ve/veaa061)
561 30 Tegally, H. *et al.* Sixteen novel lineages of SARS-CoV-2 in South Africa. *Nat Med* **27**, 440-446
562 (2021). [https://doi.org:10.1038/s41591-021-01255-3](https://doi.org/10.1038/s41591-021-01255-3)
563 31 Munywoki, P. K. *et al.* The source of respiratory syncytial virus infection in infants: a household
564 cohort study in rural Kenya. *J Infect Dis* **209**, 1685-1692 (2014).
565 [https://doi.org:10.1093/infdis/jit828](https://doi.org/10.1093/infdis/jit828)
566 32 Agoti, C. N. *et al.* Transmission networks of SARS-CoV-2 in Coastal Kenya during the first two
567 waves: a retrospective genomic study. *eLife* **11**, e71703 (2022).
568 [https://doi.org:10.7554/eLife.71703](https://doi.org/10.7554/eLife.71703)
569 33 Gomaa, M. R. *et al.* Incidence, household transmission, and neutralizing antibody
570 seroprevalence of Coronavirus Disease 2019 in Egypt: Results of a community-based cohort.
571 *PLoS Pathog* **17**, e1009413 (2021). [https://doi.org:10.1371/journal.ppat.1009413](https://doi.org/10.1371/journal.ppat.1009413)
572 34 Cohen, C. *et al.* SARS-CoV-2 incidence, transmission, and reinfection in a rural and an urban
573 setting: results of the PHIRST-C cohort study, South Africa, 2020-21. *Lancet Infect Dis* (2022).
574 [https://doi.org:10.1016/s1473-3099\(22\)00069-x](https://doi.org/10.1016/s1473-3099(22)00069-x)
575 35 Semakula, M. *et al.* The secondary transmission pattern of COVID-19 based on contact tracing in
576 Rwanda. *BMJ Global Health* **6**, e004885 (2021). [https://doi.org:10.1136/bmjgh-2020-004885](https://doi.org/10.1136/bmjgh-2020-004885)
577 36 Ratovoson, R. *et al.* Household transmission of COVID-19 among the earliest cases in
578 Antananarivo, Madagascar. *Influenza Other Respir Viruses* **16**, 48-55 (2022).
579 [https://doi.org:10.1111/irv.12896](https://doi.org/10.1111/irv.12896)
580 37 Herman-Roloff, A. *et al.* Adapting Longstanding Public Health Collaborations between
581 Government of Kenya and CDC Kenya in Response to the COVID-19 Pandemic, 2020-2021.
582 *Emerg Infect Dis* **28**, S159-s167 (2022). [https://doi.org:10.3201/eid2813.211550](https://doi.org/10.3201/eid2813.211550)

- 583 38 Orangi, S. *et al.* Epidemiological impact and cost-effectiveness analysis of COVID-19 vaccination
584 in Kenya. *BMJ Glob Health* **7** (2022). <https://doi.org/10.1136/bmjgh-2022-009430>
- 585 39 Bendall, E. E. *et al.* SARS-CoV-2 Genomic Diversity in Households Highlights the Challenges of
586 Sequence-Based Transmission Inference. *mSphere* **7**, e00400-00422 (2022).
587 <https://doi.org/doi:10.1128/msphere.00400-22>
- 588 40 Agoti, C. N. *et al.* Genomic analysis of respiratory syncytial virus infections in households and
589 utility in inferring who infects the infant. *Sci Rep* **9**, 10076 (2019).
590 <https://doi.org/10.1038/s41598-019-46509-w>
- 591 41 Gallego-García, P. *et al.* Limited genomic reconstruction of SARS-CoV-2 transmission history
592 within local epidemiological clusters. *Virus Evolution* **8** (2022).
593 <https://doi.org/10.1093/ve/veac008>
- 594 42 Scott, J. A. *et al.* Profile: The Kilifi Health and Demographic Surveillance System (KHDSS). *Int J*
595 *Epidemiol* **41**, 650-657 (2012). <https://doi.org/dys062> [pii]
596 10.1093/ije/dys062
- 597 43 Tyson, J. R. *et al.* Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome
598 sequencing using nanopore. *bioRxiv* (2020). <https://doi.org/10.1101/2020.09.04.283077>
- 599 44 O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the
600 pangolin tool. *Virus Evol* **7**, veab064 (2021). <https://doi.org/10.1093/ve/veab064>
- 601 45 Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic
602 epidemiology. *Nat Microbiol* **5**, 1403-1407 (2020). <https://doi.org/10.1038/s41564-020-0770-5>
- 603 46 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-
604 4123 (2018). <https://doi.org/10.1093/bioinformatics/bty407>
- 605 47 Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and
606 annotation of phylogenetic trees with their covariates and other associated data. *Methods in*
607 *Ecology and Evolution* **8**, 28-36 (2017). [https://doi.org:https://doi.org/10.1111/2041-
608 210X.12628](https://doi.org/https://doi.org/10.1111/2041-210X.12628)
- 609 48 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis.
610 *Virus Evol* **4**, vex042 (2018). <https://doi.org/10.1093/ve/vex042>
- 611