

1 **A2B-COVID: A method for evaluating potential SARS-CoV-2 transmission events**

2

3 **Authors**

4 Christopher J. R. Illingworth^{*1,2}, William L. Hamilton^{*3,4}, Chris Jackson^{*1}, Ashley Popay⁵, Luke
5 Meredith⁶, Charlotte J. Houldcroft³, Myra Hosmillo⁶, Aminu Jahun⁶, Matthew Routledge^{4,9}, Ben
6 Warne^{3,4}, Laura Caller⁷, Sarah Caddy⁸, Anna Yakovleva⁶, Grant Hall⁶, Fahad A. Khokhar⁶,
7 Theresa Feltwell³, Malte L. Pinckert⁶, Iliana Georgana⁶, Yasmin Chaudhry⁶, Martin Curran⁹,
8 Surendra Parmar⁹, Dominic Sparkes^{4,9}, Lucy Rivett^{4,9}, Nick K. Jones^{4,9}, Sushmita Sridhar^{3,8,10},
9 Sally Forrest⁸, Tom Dymond⁴, Kayleigh Grainger⁴, Chris Workman⁴, Effrossyni Gkrania-
10 Klotsas^{4,11,12}, Nicholas M. Brown^{4,9}, Michael P. Weekes^{3,8}, Stephen Baker^{3,8}, Sharon J. Peacock³,
11 Theodore Gouliouris^{4,9}, Ian Goodfellow⁶, Daniela De Angelis^{1,13}, M. Estée Török^{3,4}

12

13 * Contributed equally

14

15 **Affiliations**

- 16 1. MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site,
17 Robinson Way, Cambridge, CB2 0SR, United Kingdom
- 18 2. Institut für Biologische Physik, Universität zu Köln, Zùlpicherstr. 77, 50937, Köln, Germany
- 19 3. University of Cambridge, Department of Medicine, Cambridge Biomedical Campus, Hills
20 Road, Cambridge CB2 0QQ, United Kingdom
- 21 4. Cambridge University Hospitals NHS Foundation Trust, Cambridge Biomedical Campus,
22 Hills Road, Cambridge CB2 0QQ, United Kingdom
- 23 5. Public Health England Field Epidemiology Unit, Cambridge Institute of Public Health,
24 Forvie Site, Cambridge Biomedical Campus, Cambridge CB2 0SR, United Kingdom
- 25 6. University of Cambridge, Department of Pathology, Division of Virology, Cambridge
26 Biomedical Campus, Hills Road, Cambridge CB2 0QQ, United Kingdom
- 27 7. Francis Crick Institute, 1 Midland Rd, Somers Town, London NW1 1AT, United Kingdom
- 28 8. Cambridge Institute for Therapeutic Immunology and Infectious Disease, Jeffrey Cheah
29 Biomedical Centre, Puddicombe Way, Cambridge CB2 0AW, United Kingdom
- 30 9. Public Health England Clinical Microbiology and Public Health Laboratory, Cambridge
31 Biomedical Campus, Hills Road, Cambridge CB2 0QQ, United Kingdom
- 32 10. Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1RQ
- 33 11. MRC Epidemiology Unit, University of Cambridge, Level 3 Institute of Metabolic Science,
34 Cambridge CB2 0SL, United Kingdom
- 35 12. University of Cambridge, School of Clinical Medicine, Cambridge Biomedical Campus,
36 Hills Road, Cambridge CB2 0QQ, United Kingdom
- 37 13. Public Health England, National Infection Service, 61 Colindale Avenue, London NW9
38 5EQ, United Kingdom

39

40

41 **Correspondence**

42 Dr Chris Illingworth cjri2@cam.ac.uk

43

44

45 **Abstract**

46
47 Identifying linked cases of infection is a key part of the public health response to viral infectious
48 disease. Viral genome sequence data is of great value in this task, but requires careful analysis,
49 and may need to be complemented by additional types of data. The Covid-19 pandemic has
50 highlighted the urgent need for analytical methods which bring together sources of data to inform
51 epidemiological investigations. We here describe A2B-COVID, an approach for the rapid
52 identification of linked cases of coronavirus infection. Our method combines knowledge about
53 infection dynamics, data describing the movements of individuals, and novel approaches to
54 genome sequence data to assess whether or not cases of infection are consistent or inconsistent
55 with linkage via transmission. We apply our method to analyse and compare data collected from
56 two wards at Cambridge University Hospitals, showing qualitatively different patterns of linkage
57 between cases on designated Covid-19 and non-Covid-19 wards. Our method is suitable for the
58 rapid analysis of data from clinical or other potential outbreak settings.

59

60 **Introduction**

61
62 Having emerged via zoonotic transfer in late 2019, the COVID-19 pandemic remains an ongoing
63 public health priority [1,2]. Understanding the nature of viral transmission is a key factor in all
64 strategies to prevent and control disease spread. The earliest stages of the outbreak were
65 characterised by small, localised clusters of infection [3,4]. Identifying chains of linked cases
66 remains crucial for containing disease spread [5,6], particularly in healthcare settings. Outbreaks
67 in these settings, however, provide a particular challenge to the identification of linked cases as
68 regular new introductions from the community have to be distinguished from potential cases of
69 nosocomial transmission [7–9].

70

71 Viral genome sequencing provides one strategy for identifying clusters of transmission. The rapid
72 evolution of viral populations leads to the accumulation over time of genetic differences which
73 distinguish linked from unlinked cases [10]. A broad range of phylogenetic approaches for
74 identifying linked infection clusters have previously been described [11–15]

75

76 With this background, a recent study of hospital-based COVID-19 infection used genome
77 sequencing to identify potential clusters of infection. In this study, sets of individuals with identical
78 viral genome sequences were often verified by a process of epidemiological follow-up to
79 correspond to likely cases of nosocomial transmission [5]. Data from genome sequencing allowed
80 feedback in real time to clinical, infection control and hospital management teams to inform their
81 response.

82

83 Whilst genome sequence data is of great value for studying viral transmission, it has some
84 drawbacks in the context of SARS-CoV-2. Given the recent emergence of SARS-CoV-2, and the
85 low diversity of global viral sequences, finding identical sequences in different individuals does
86 not necessarily imply a connection between those people. Furthermore, sequence data is not the
87 only type of information that may be available. For example, studies of known SARS-CoV-2
88 transmission events have quantified the distribution of times between the onset of symptoms and
89 the transmission of the virus, and of the subsequent time between infection and the onset of
90 symptoms [16–19]. In addition the contribution of asymptomatic individuals to transmission is
91 difficult to ascertain as they may not be sampled [20,21]. Such information has implications for
92 the analysis of potential transmission events. Information about the location of individuals might
93 also contribute to the identification or ruling out of connections; people who never physically
94 interact cannot directly transmit the virus from one to another. Potential, therefore, exists for novel
95 approaches in the identification of transmission clusters.

96

97 New methodologies have the potential to gain insights into viral sequence data. For example,
98 noise in the process of viral sequencing can affect phylogenetic analyses which rely on finding
99 identical sequences. Multiple studies have considered the problem of noise in genome sequence
100 data, particularly with regard to identifying variant frequencies [22–25]. The potential for error in
101 variant frequencies means that a viral consensus sequence is also stochastic, since it is
102 generated from a potentially diverse viral population. Further, evolution would be expected to
103 change viral sequences over time. Metrics which account for measurement error and viral
104 evolution may be advantageous for identifying linked cases of infection.

105

106 A variety of studies have used short-read viral sequencing to evaluate the nature and likely
107 direction of viral transmission. The ability of such data to capture the diversity of within-host viral
108 populations has proved very valuable in the assessment of transmission events [26–29].
109 However, the low cost of data collection via nanopore-based methods has often outweighed the
110 additional precision provided by Illumina technology [30,31]. Analyses which can utilise the
111 consensus sequences provided by such methods have broader application, increasing their
112 value.

113

114 Here, we describe a tool which implements a combined statistical and evolutionary framework to
115 analyse genome sequence data with location data and knowledge of SARS-CoV-2 infection
116 dynamics to rapidly identify clusters of infection. Our approach provides a rapid data analysis,
117 outputting information in an easily interpretable format. We demonstrate its use with augmented
118 sequence data from outbreaks on hospital wards, illustrating the value of different kinds of data
119 in this context. While designed with clinical application in mind, the generality of the properties of
120 viral transmission make our method applicable to any setting in which appropriate data has been

121 collected from more than one individual. We hope that our approach will be of value in ongoing
122 public health efforts to combat the COVID-19 pandemic.

123

124 **Results**

125

126 Our method exploits data from genome sequencing alongside other information about individuals
127 and COVID-19 infection. Given short periods of time between samples, the extent of
128 measurement error in a viral sample may exceed the extent of evolutionary change in a population
129 [32]. As a preliminary step therefore, we evaluated the extent of measurement error in our
130 sequencing pipeline.

131

132 We examined cases among data collected from patients at the Cambridge University Hospitals
133 National Health Service Foundation Trust (CUH) for which more than one viral sample was
134 sequenced. A total of 136 such patients were identified, with between 2 and 9 (median 2) samples
135 collected from each individual and 336 samples in total. Intervals between pairs of samples varied
136 from 0 to 39 days. Each sample gave rise to a consensus sequence. We filtered the data to
137 remove sequences with less than 90% coverage of the genome. Combining these data through
138 regression, we inferred a mean error rate of approximately 0.207 nucleotide errors per sequence
139 (S1 Figure). While small, this rate is significant. In our model, the expected time between
140 symptoms being reported from individuals in a transmission pair is 5.7 days; the expected amount
141 of sequence evolution within this time is not greater than the expected difference between two
142 sequences resulting from noise (S2 Figure).

143

144 For each pair of individuals in a dataset we compared symptom onset, location, and genome
145 sequence data with an underlying model of transmission, identifying whether or not the data were
146 consistent with the null hypothesis that direct viral transmission occurred between the two

147 individuals. Our model produces a simple output, stating that the data are either 'consistent' with
148 the hypothesis of transmission (nominal p-value > 0.05), that transmission is 'unlikely' to have
149 occurred (p-value < 0.01) or that the case is 'borderline' (p-value between 0.05 and 0.01).

150
151 We applied our model to data from two wards within CUH, which we term X and Y. Ward X was
152 a 'green' ward, used for patients considered to be free from COVID-19 infection. By contrast,
153 ward Y was a 'red' ward within the hospital, designated for the treatment of patients with COVID-
154 19 infection, on which multiple cases of infection in healthcare workers (HCWs) were identified.
155 Information collected for these individuals included genome sequence data from viral swabs,
156 dates of symptom onset and dates on which individuals were present on the wards in question.

157
158 Our method combines multiple types of data, using the information available to identify potentially
159 linked cases of infection (Figure 1). We conducted two analyses of the data from wards X and Y,
160 the first looking at the complete data collected for individuals, including dates of symptom onset,
161 viral genomes and location data, and the second excluding location data, considering only times
162 of symptom onset and the viral sequence data.

163
164 Analysing the complete data from each ward identified potential transmission events in each case
165 (Figure 2). Outputs from our method are asymmetrical. For example, the data is consistent with
166 transmission from 7069 to 7074 having occurred, but transmission from 7074 to 7069 was ruled
167 unlikely. This is in part explained by 7069 reporting symptoms four days before 7074; the order
168 of reporting symptoms provides information on the likely direction of transmission. Examination
169 of the output from our code can be used to identify potential clusters of linked transmission events.
170 For example, data from ward X suggest that the ten infections for which data were available were
171 potentially mutually connected by transmission, though some underlying structure can be seen.
172 The bottom three individuals coded 7108, 7128, and 7074, could each have infected each other,

173 but have only borderline chances of having infected anyone else on the ward. These individuals
174 could have been infected by either 7069 or 7112 infecting 7074, leading to further transmission
175 events. The remainder of the individuals on the ward show multiple plausible transmission events,
176 though we note that 7129 has only a borderline probability of having infected any other person.
177 Examination of location data collected for the ward provides some insight into these findings
178 (Figure 2B). Individuals 7108 and 7128 were never present on ward X, but were mutually in
179 contact with 7074 throughout the duration of the outbreak via known household contacts. The
180 health care worker 7129 was present on ward X at the same time as many of the other individuals,
181 but became symptomatic at a later point; the later onset of symptoms suggests that 7129 was
182 infected by another individual in the outbreak without passing the virus on within the ward. The
183 overall picture we discover is what may be a single outbreak of linked cases on ward X, potentially
184 started by a patient, but largely involving HCWs both on and off the ward.

185
186 Analysis of data from ward Y revealed a less complex chain of events (Figure 2C). Precisely two
187 clusters of individuals were inferred from the complete dataset, with individuals 1773, 2914, and
188 4255 forming one mutually interconnected set of infections, and the individuals 4902, 4633, and
189 4493 forming a second. While location data was missing for the latter three individuals the mutual
190 connections between individuals can again be understood from the location data that do exist
191 (Figure 2D). We note that in ward Y, cases were spread over a longer period of time than in ward
192 X. In this case, we find two potential outbreaks between HCWs on the ward, with other cases
193 reported on the ward not being linked to these individuals. Our results reflect the 'green' and 'red'
194 natures of the two wards. Where ward X was designated for patients who were in theory free
195 from COVID-19 infection, a coherent cluster of infection, potentially indicating a single introduction
196 of the virus into the ward, was responsible for all of the observed cases. By contrast ward Y,
197 being designated for COVID-19 cases, had multiple introductions of the virus onto the ward,
198 perhaps two of those cases leading to ongoing nosocomial transmission.

199

200 Sensitivity analyses suggested that the measurement error has some effect on the outputs of our
201 model. Calculations performed with increased and decreased error parameters led to changes
202 in which events were identified as 'Consistent', 'Borderline' or 'Unlikely' (S3 Figure). Reducing
203 the input error parameter to zero has the potential to induce more significant changes in our model
204 output, as discussed further in the Methods section.

205

206 In order to assess the value of location data we repeated our estimation in their absence, using
207 only viral genome sequence data and the dates on which individuals first reported symptoms.
208 Location data generally reduces the inferred potential for viral transmission; under the logic of our
209 method, individuals who are not in the same place at the same time cannot infect one another.
210 The value of location data in increasing the precision with which networks of links between cases
211 of infection is shown by the results from ward X (Figure 3A). While the overall pattern of data
212 shows multiple potential connections between individuals, the independence of the transmissions
213 between 7108, 7128, and 7074 from the remainder of the network was lost in the absence of
214 spatial data. Spatial data in our model was defined in terms of presence or absence on a ward,
215 more refined information not being available. An analysis of data from ward Y showed an intriguing
216 result, with previously unseen potential links between individual 2019 and the first cluster, and
217 between 3327 and the second cluster (Figure 3B). These HCW displayed symptoms earlier than
218 the people in their respective clusters, consistent with being the original cases in each case.
219 Further, phylogenetic reconstruction showed that the sequences from these individuals were
220 consistent with their being linked (Figure 3C; similar data for ward X is shown in S4 Figure).
221 However, the location data do not show them as working shifts on ward Y at the same time as
222 any of the other linked individuals were present. We suggest that unrecorded contacts may exist
223 in this case.

224

225 The potential for missing location data to prevent the identification of linked individuals shows that
226 measuring location can be difficult, more so for HCWs than patients. While patients are unlikely
227 to be highly mobile, HCWs move around the hospital outside of their shifts. Unless explicitly
228 recorded, off-ward contacts between HCWs are unlikely to be noted, leading to the potential non-
229 inference of genuine links. While missing location data cannot be unambiguously diagnosed as
230 the cause of our result, this case highlights the limitations intrinsic to our approach. Our software
231 may provide valuable insights, but does not replace the need for full epidemiological investigation.

232

233 **Discussion**

234

235 We have here set out a method for identifying potential cases of direct transmission between pairs
236 of individuals, based upon the dynamics of SARS-CoV-2 infection, data describing times of co-
237 location between individuals, and genome sequence data collected during infection. In a first
238 application of our method, we analysed data from two hospital wards. In each, we identified cases
239 where the data were consistent with viral transmission occurring between either patients or HCWs
240 on the ward. Our method builds upon information that can be obtained from a phylogenetic
241 analysis, incorporating data from multiple sources to present an easily-interpretable map of
242 potential linked cases of infection. It is likely to be valuable in the initial assessment of potential
243 cases of nosocomial transmission, highlighting pairs or clusters of individuals for further
244 epidemiological assessment, and allowing for a more strategic deployment of resources for
245 outbreak investigation and targeted interventions.

246

247 Our method brings together a variety of data, combining an evolutionary model for the analysis of
248 sequence data with location information and details of the dynamics of viral infection. In contrast
249 to standard phylogenetic approaches to sequence data, our model explicitly accounts for noise in
250 the generation of a viral consensus sequence; using within-host data we identified a magnitude

251 of error of a fraction of one nucleotide per genome. In rapidly evolving viruses for which
252 transmissions are separated by longer periods of time, the within-host evolution of viral
253 populations is likely to overwhelm the effect of noise in the sequencing process. However, for
254 cases of acute infection, separated by only a few days, the extent of noise may be close to the
255 expected evolutionary change in the population, making it an important factor to consider.

256
257 One limitation of our method is that it deals with consensus viral sequences rather than deep
258 sequence data. Where available, detailed measurements of within-host viral diversity may lead to
259 an improved picture of relationships between cases of viral infection. We note further that our
260 tool analyses data in a pairwise manner; while distinguishing plausible from implausible links
261 between cases of infection, it does not attempt to infer a complete reconstruction of a transmission
262 network. Unobserved cases of infection are not considered. Our model used parameters which
263 in some cases have been derived from early studies into SARS-CoV-2 spread. To account for
264 the event that further research leads to a better understanding of viral transmission we provide
265 options to perform calculations with user-specified parameters. We finally note that a statistical
266 inference from our model does not describe the probability of transmission having occurred
267 between two individuals. Instead it describes how consistent the data are with transmission. Our
268 model is intended as a first step towards further epidemiological investigation.

269
270 Our model has a range of features specifically tailoring it to the real-time analysis of data in a
271 hospital context during an outbreak of a rapidly spreading viral disease. Our method is designed
272 for simplicity both in being easy to use and in rapidly producing an interpretable output. To this
273 extent our method is limited in what we try to infer, highlighting only pairs of individuals where the
274 data are consistent with transmission. For example, in a case where an individual A infects B and
275 C, it could be that our method highlights not only the real transmission events, but also reports
276 that the data from B and C is consistent with transmission occurring between them. This does

277 not comprise an error in our method, but does require our method to be understood. We note
278 that, in a hospital environment, a positive output from our method could be followed up by
279 investigative efforts and epidemiological follow up; such efforts have the potential to collect data
280 beyond that considered by our method.

281
282 We believe that the key application of our method will be in investigating nosocomial transmission
283 of SARS-CoV-2. Within a hospital, potential cases of transmission may be obscured by a large
284 number of cases of community-acquired infection. In a busy clinical setting, our tool has the ability
285 to rapidly separate potentially linked cases from those which are likely to be unlinked. In this way
286 we allow investigative efforts and epidemiological followup to be focused more precisely,
287 concentrating effort on cases where transmission is a real possibility.

288

289 **Methods**

290

291 **Model overview**

292

293 We here consider pairs of individuals, who for the purpose of notation, we describe as individuals
294 A and B. Given data on when the individuals became symptomatic for SARS-CoV-2 infection,
295 their locations, and their viral genome sequences, we generate a statistic to test whether the data
296 are consistent with viral transmission having occurred from A to B.

297

298 To outline this process, suppose that we have observed data y from this pair of individuals. The
299 null hypothesis of transmission is supported by the data if these data have high probability of
300 having arisen given transmission from A to B. More formally, the hypothesis is accepted at a
301 confidence level ψ if the probability of observing y , or data that are “less extreme” (i.e. data that

302 are even more consistent with transmission than y) is at least ψ under the hypothesis of direct
303 transmission from A to B. We outline our method in detail below.

304

305 **Available data**

306

307 *Notation*

308

309 An overview of the notation used in the description of our model is shown in Figure 4. The dates
310 of symptom onset and the dates when viral sequence data were collected are denoted S_A and S_B
311 and D_A and D_B , respectively. These dates are assumed to be known, or in the case of symptom
312 dates can be estimated from times at which individuals tested positive. Further data described
313 the locations of the individuals A and B on each day, with the binary indicator $C_A(L,T)$ denoting
314 whether individual A was present in location L on day T. The information describing the location
315 of individuals may be uncertain, so we represent it by $w_A(L,T)$, the probability that individual A is
316 present in location L on day T. For example, if A is known to be in location L on day T we have
317 $w_A(L,T)=1$, while if A is known not to be in location L on day T we have $w_A(L,T)=0$. If the location
318 of A at this time is unknown, $w_A(L,T)$ is defined as described below. Analogously to this, the binary
319 indicator $C_{AB}(T)$ denotes whether or not A and B were in contact on day T. Uncertainty in this
320 indicator is represented by the probability $w_{AB}(T)$ that A and B were present in the same location
321 on this day. In describing genomic data, H_A and H_B describe Hamming distances between the
322 viral sequences collected from A and B and their mutual consensus. The CT scores of the viral
323 samples are denoted V_A and V_B .

324

325 *Symptom data*

326

327 Due to extensive monitoring of individuals in hospital, we often had information on the dates of
328 symptom onset for individuals. When these were unknown we estimated them from positive test
329 dates. To perform this estimation we used symptom onset dates and positive test dates from 86
330 health care workers and 393 patients from Cambridge University Hospitals, fitting an offset
331 gamma distribution to these data (S5 Figure, S1 Table). Where symptom dates were missing,
332 the mean of this distribution was used to impute symptom onset dates from positive test dates.
333 We write \hat{S}_A to denote an estimate for S_A . Where positive test dates are used in place of symptom
334 onset dates, greater care is required in the interpretation of results.

335

336 *Location data*

337

338 In our study, time was measured in whole days. For example, if an individual was known to be in
339 location L for any part of day T , we set $w_A(L,T)=1$. Known location data were edited for health
340 care workers to account for their increased mobility, night shifts which span more than one day,
341 and uncertainties such as the potential for fomite transmission. If for a healthcare worker we had
342 that $w_A(L,T)=1$ for some L and T we set $w_A(L,T-1)$ and $w_A(L,T+1)$ to be equal to a minimum value
343 of 0.5.

344

345 Where location data were missing it was necessary to specify values $w_A(L,T)$. Data from our
346 study were centred on cases from a specific part of the hospital, usually a single ward; this location
347 was denoted L^* . Where location data were missing for a patient, we set $w_A(L^*,T)=1$ for all T ,
348 assuming that a patient was always on the most common ward. Where location data were missing
349 for health care workers, we set $w_A(L^*,T)=4/7$ for all T , reflecting shift patterns. We note that in
350 other circumstances (e.g. a dataset spanning an entire hospital), an alternative prior for the
351 location of individuals could be more appropriate.

352

353 Contact information was derived from the location data. For any two individuals we note that
354 there could be multiple locations in which they could be in contact on a single day. We combined
355 probabilities of contact across potential locations, calculating

356

$$357 \quad w_{AB}(t) = 1 - \prod_L (1 - w_A(L, t)w_B(L, t))$$

358

359 *Viral genome sequence data*

360

361 Consensus genome sequences were calculated from viral sequence data. Sequences were
362 subjected to two levels of quality control. The first considered the coverage of the genome. An
363 unambiguous nucleotide is here defined as an instance in which sequencing describes an A, C,
364 G, or T. We applied the criterion that sequences had to unambiguously describe nucleotides at
365 80% or more of the sites in the genome.

366

367 The second level of quality control counted ambiguous nucleotides that were found at sites in the
368 genome that were found to be polymorphic between the collected viral sequences. These sites
369 are more likely to be informative with regards to the number of genetic differences between two
370 sequences; a genome with high overall coverage but ambiguity at multiple of these positions
371 would in practice be quite uninformative. Having identified polymorphic sites, we required
372 sequences to have no more than one ambiguous nucleotide at these positions.

373

374 In some cases, multiple viral samples were collected from the same individual. Viral genomes
375 collected from the same individual were usually extremely similar to one another (S1 Figure). In
376 such a case, we identified the earliest sequence with sufficient coverage of the viral genome,
377 using this sequence for analysis. Where positions in this genome were ambiguous, and where

378 other sequences from the same individual had unambiguous nucleotides at these positions, the
379 other sequences were used to construct a more complete consensus sequence for the individual.

380
381 Given viral sequences from the pair of individuals A and B we calculated Hamming distances from
382 each sequence to a pairwise consensus sequence; we denote these distances as H_A and H_B .

383

384 **Assessing viral transmission**

385

386 We denote as X_T an indicator for the event that transmission took place at time T , and as X an
387 indicator for the event that transmission took place at all. To test the hypothesis of transmission,
388 we calculated the probability of observing the data y under the null hypothesis that transmission
389 occurred, $p(y|X) = \sum_T p(y|X_T)P(X_T|X)$, where $P(X_T|X)$ is the probability that transmission took place
390 at T given transmission, which we abbreviate as $P(T)$.

391
392 Let Y represent the observable data. Y consists of the symptom time S_B , the Hamming distances
393 H_A and H_B , and the set of $C_{AB}(T)$ for all T , denoted C_{AB} . We will write an expression for the
394 probability of the observable data as follows:

395

$$396 \quad p(Y|D, X, \theta) = \sum_T P(T|S_A, \theta)P(S_B|\theta, X_T)P(C_{AB}|X_T)P(H_A, H_B|\theta, D, E, X_T) \quad (1)$$

397

398 where $D=\{D_A, D_B\}$, E is the error in sequencing, and θ represents the set of parameters that are
399 assumed to be known. We note that we condition on S_A ; an alternative approach would be to
400 write the equation in terms of S_B-S_A . We consider the parts of this equation in turn.

401

402 **Assessing viral transmission: Symptom and location data**

403

404 In equation (1), $P(T|S_A, \theta)$ describes the probability that transmission is at time T , where time is
 405 measured relative to S_A , the time of onset of symptoms in A. This term describes the infectivity
 406 profile of the virus, that is, the time from symptom onset to transmission. We follow previously
 407 published work which has characterised this as an offset gamma distribution[16,17,35].

408
 409 The term $P(S_B|\theta, X_T)$ describes the probability that B becomes symptomatic at time S_B , given that
 410 transmission occurs at time T . Again we have information from the same literature characterising
 411 this as a lognormal distribution. We therefore write:

412
 413
$$P(T|S_A, \alpha, \beta, s) = \frac{e^{-(T-S_A+s)/\beta} (T-S_A+s)^{\alpha-1} \beta^{-\alpha}}{\Gamma(\alpha)}, \quad (2)$$

414
 415 where s is the offset and $\alpha=97.1875$, $\beta=0.2689$, and $s=25.625$, and

416
 417
$$P(S_B|\mu, \sigma, X_T) = \frac{e^{-\frac{(\log(S_B-T)-\mu)^2}{2\sigma^2}}}{(S_B-T)\sigma\sqrt{2\pi}}, \quad (3)$$

418
 419 where $\mu=1.434$, and $\sigma=0.6612$. Each of these expressions treat T as a continuous variable ; we
 420 use an approximation to discretise the formula to a resolution of single days, obtaining

421
 422
$$P(T|S_A, \theta)P(S_B|\theta, X_T) = \left[\int_{T-S_A-0.5}^{T-S_A+0.5} \frac{e^{-(x+s)/\beta} (x+s)^{\alpha-1} \beta^{-\alpha}}{\Gamma(\alpha)} dx \right] \left[\int_{S_B-T-0.5}^{S_B-T+0.5} \frac{e^{-\frac{(\log(x)-\mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}} dx \right]. \quad (4)$$

423
 424 We next consider the term $P(C_{AB}|X_T)$. Where $|C|$ is the length of the vector C_{AB} , we note that there
 425 are $2^{|C|}$ possible such vectors. If $C_{AB}(T)=0$, transmission cannot have occurred at time T , so that,
 426 if transmission occurred at time, then $C_{AB}(T)$, the T th element of C_{AB} , equals 1. Regarding other

427 elements of C_{AB} , we take a naive approach to contact patterns, assuming that $P(C_{AB}(t)=1|X_T)=0.5$
428 for each t not equal to T . This implies the result that $P(C_{AB}|X_T)=0.5^{|C_{AB}|-1}$ for any C_{AB} with $C_{AB}(T)=1$,
429 with $P(C_{AB}|X_T)=0$ otherwise.

430

431 **Assessing viral transmission: Viral sequence data**

432

433 Finally, we consider the term $P(H_A, H_B|\theta, D, X_T)$, which is derived from the viral genome sequence
434 data. In order to generate H_A and H_B , we first calculated a local consensus sequence across all
435 of the viral genomes in our data. Next, for each pair of sequences from individuals A and B, we
436 calculated a pairwise consensus, defined as the nucleotide shared by the two sequences where
437 the sequences agreed, and the nucleotide in the local consensus where the sequences differed.
438 H_A and H_B were then calculated as the Hamming distances from each of the two sequences to
439 the pairwise consensus sequence. These distances describe the number of substitutions
440 observed to have been gained by the viral population in each individual.

441

442 In our analysis we assumed an infinite sites model; among our sequences any given mutation
443 can be obtained only once, while the reversion of mutations back to the consensus never occurs.

444

445 We used a Poisson model to compare the number of observed substitutions in each sequence
446 with an expected rate of viral evolution. Our model includes a term accounting for errors in the
447 viral consensus sequences. In the notation of Figure 4, we note that if D_A is before T , any variants
448 observed in sequence data from A but not in the data from B can only arise from error. Under our
449 infinite sites assumption, such variants cannot revert in the time between D_A and D_B so must be
450 caused by error in the observation. However, if D_A is after T , such variants have the potential to
451 evolve in the time between D_A and T , in addition to being potentially caused by measurement
452 error.

453

454 Similarly, variants observed in data from B but not from A can arise either from error, or as a result
455 of evolution going back an assumed common ancestor at the earlier of the time of transmission
456 T and the previous time of sequencing D_A . We can thus describe the probability of observing the
457 data H_A and H_B under the assumption of transmission at time T:

458

459 $P(H_A, H_B | \theta, D, E, X_T)$

460
$$= \left(\frac{(E/2 + \gamma_G P_A)^{H_A} e^{-(E/2 + \gamma_G P_A)}}{H_A!} \right) \left(\frac{(E/2 + \gamma_G (D_B - Q_A))^{H_B} e^{-(E/2 + \gamma_G (D_B - Q_A))}}{H_B!} \right)$$

461

462 where $P_A = \max\{0, D_A - T\}$ and $Q_A = \min\{D_A, T\}$. The rate of evolution γ_G describes the expected
463 number of substitutions per genome per day, while the parameter E is the mean number of errors
464 in the Hamming distance between two viral sequences, estimated as described below.

465

466 Estimating noise in genome sequence data

467

468 In order to estimate the extent of measurement error in a consensus viral genome, we examined
469 cases among data collected at Cambridge University Hospitals (CUH) for which more than one
470 viral sample was sequenced. We identified 136 such patients, with between 2 and 9 samples
471 collected from each individual and 336 samples in total. Each sample gave rise to a consensus
472 sequence; we filtered the data to remove sequences with less than 90% coverage of the genome.
473 For each pair of samples i and j , collected from the same individual, we recorded H_{ij} , the Hamming
474 distance between them, ΔT_{ij} , the absolute difference in time between the dates on which the
475 samples were collected, measured in days, and the viral load of each sample, as represented by
476 the CT scores V_i and V_j .

477

478 Following in principle a previous approach to estimating noise and rates of evolution [32], we then
479 fitted a Poisson model to the data, deriving for each pair the log likelihood

480

$$481 \quad \log L^D(\varepsilon, \lambda, \gamma \mid H_{ij}, \Delta T_{ij}, V_i, V_j) = \log \left(\frac{\left(\frac{\varepsilon}{2} (V_i + V_j) + \lambda + \gamma \Delta T_{ij} \right)^{H_{ij}} e^{-\left(\frac{\varepsilon}{2} (V_i + V_j) + \lambda + \gamma \Delta T_{ij} \right)}}{H_{ij}!} \right)$$

482

483 and estimating the parameters ε , λ and γ so as to maximise the sum of the log likelihoods across
484 all pairs of sequences; we inferred the parameters $\hat{\varepsilon}=0.0200$, $\hat{\lambda}=-0.0693$ and $\hat{\gamma}=0.0453$. Here the
485 value $\hat{E}(V_i, V_j) = \hat{\lambda} + \hat{\varepsilon}(V_i + V_j)$ provides a simple estimate of the extent of measurement error in
486 a Hamming distance, expressed in terms of the CT scores of the two samples. For the purposes
487 of our model this function was evaluated at the mean CT score of 24.091. This provided an
488 estimate for the pairwise difference arising through measurement error, \hat{E} , of 0.414 nucleotides,
489 equivalent to 0.207 nucleotide errors per genome sequence. The estimate $\hat{\gamma}$ describes the mean
490 rate of within-host evolution calculated across the within-host sample. It is expressed as a number
491 of substitutions per genome per day, and is equivalent to a rate of 6.0×10^{-4} substitutions per
492 locus per year, close to the value of 8×10^{-4} that has been calculated from global sequence data
493 [33]. In so far as we require an estimated rate of evolution spanning both within-host and
494 between-host evolution, we used in our model a rate $\hat{\gamma}_G$ of 0.0655 nucleotides per day, equivalent
495 to this latter, globally estimated, rate of evolution.

496

497 To examine the effect of CT score upon our inference, a repeat calculation was performed in
498 which these data were ignored; this gave a worse fit to the data under the Bayesian Information
499 Criterion[34] (S2 Table).

500

501 In a case where no sequence data was observed for an individual, we excluded that individual
502 from our calculation. An option within our method allows for calculations to be performed between
503 individuals where no sequence data was collected; under this option we set $P(H_A, H_B | \theta, D, E,$
504 $X_T) = 1$ for all A and B.

505

506 **Assessing viral transmission: Hypothesis testing**

507

508 Having derived the expression (1) for $P(Y|D,X)$, we now derive the probability $P(y|D,X)$ of the
509 specific observed data y . The data y consist of the symptom time S_B , if it is known, the Hamming
510 distances H_A and H_B , the set of those $C_{AB}(T)$ that are known, and the information about potential
511 locations and contacts in cases where the $C_{AB}(T)$ are unknown, which are encapsulated in $w_{AB}(T)$.
512 $p(y|D,X)$ is defined by setting Y to equal the data y that are observed, and then integrating
513 $P(Y|D,X)$ over the potential values for any missing data.

514

515 Integration was required with respect to the unknown contact dates, applying to the term
516 $P(C_{AB}|X_T)$. We generalise the argument made for this term in the case of Y to show that in this
517 case we have

518

$$519 P(C_{AB} | X_T) = 0.5^{|C|-1} w_{AB}(T)$$

520

521 A full derivation is given in S1 Text.

522

523 We thus have the result

524

$$525 p(y|D, X) = \sum_T P(T | \widehat{S}_A, \theta) P(\widehat{S}_B | \theta, X_T) 0.5^{|C|-1} w_{AB}(T) P(H_A, H_B | \theta, D, X_T),$$

526

527 where $\theta = \{\alpha, \beta, s, \mu, \sigma, \hat{E}, \gamma_G\}$, and seek to compare this to potential values $p(Y|D, X)$. To achieve this,
528 for confidence level ψ , we define a threshold $p_\psi(D)$ by

529

530
$$\int_{Y \in \Omega_\psi} p(Y|D, X) = \psi$$

531

532 where $\Omega_\psi = \{Y: p(Y|D, X) \geq p_\psi(D)\}$. Note that the threshold $p_\psi(D)$ depends on the sample collection
533 times $D = \{D_A, D_B\}$. For values $\psi = 0.95$ and $\psi = 0.99$, the observed data y is deemed 'consistent' with
534 transmission if $p(y|D, X) \geq p_{95}(D)$, 'borderline' if $p_{95}(D) > p(y|D, X) \geq p_{99}(D)$, and unlikely if $p_{99}(D) >$
535 $p(y|D, X)$.

536

537 To identify these threshold values we calculated $p(Y|D, X)$ across large numbers of sets of data
538 Y , in which we assumed without loss of generality that $S_A = 0$. Calculations were performed for all
539 Y in which $S_B \in [-11, 87]$, and for all values H_A and H_B for which $H_A + H_B \in [0, 10]$ and $H_A \in [0,$
540 $H_A + H_B]$; these ranges were chosen to return values of at least 10^{-6} from each component of
541 $p(Y|D, X)$. In our code these statistics are calculated for $D_A \in [-10, 40]$, and $D_B \in [S_B - 10, S_B + 40]$;
542 values outside of these parameters are unlikely.

543

544 In the integral, we note that there are a large number of possible vectors C_{AB} that indicate all times
545 when a pair were in contact. We approximated the sum by generating 100 random vectors C_{AB}
546 for each set of other parameters, and calculating the sum over these vectors, altering the value
547 $0.5^{|C|-1}$ in $P(C_{AB}|X_T)$ so as to normalise the integral. Reflecting our approach to contact patterns,
548 we generated the C_{AB} as random vectors of draws from a Bernoulli distribution with mean 0.5.
549 Repeating this calculation with different sets of 100 vectors did not substantially change the
550 thresholds obtained. Our code allows for the generation of alternative thresholds with different

551 probabilities of an element of C_{AB} being equal to 1. We note that if this probability is higher, fewer
552 datasets will be judged consistent with transmission.

553

554 *Study setting, participants and data collection*

555

556 This study was conducted at Cambridge University Hospitals NHS Foundation Trust (CUH), a
557 secondary and tertiary referral centre in the East of England. SARS-CoV-2 positive cases tested
558 at the on-site Public Health England (PHE) Clinical Microbiology and Public Health Laboratory
559 (CMPHL) were identified prospectively from 26th February to 17th June 2020. The CMPHL tests
560 SARS-CoV-2 samples submitted from over thirty organisations across the East of England (EoE)
561 region and all samples from CUH. The majority of samples were tested using an in-house
562 validated qRT-PCR assay targeting the SARS-CoV-2 RdRp genes, as described in a previous
563 publication [5], with more recent samples tested using the Hologic Panther™ platform [36]. Patient
564 metadata were accessed via the electronic healthcare record system (Epic Systems, Verona, WI,
565 USA). Metadata collected included patient demographic information, duration of symptoms,
566 sample collection date and location (ward and hospital). Patients and samples were assigned
567 unique anonymised study codes. Metadata manipulations were performed using the R
568 programming language and the *tidyverse* packages installed on CUH Trust computers.

569

570 *Sample sequencing*

571

572 All samples collected at CUH and a randomised selection of samples from the EoE region were
573 selected for nanopore sequencing on-site in the Division of Virology, Department of Pathology,
574 University of Cambridge. This enabled us to rapidly investigate suspected hospital acquired
575 infections at CUH as previously described [5]. Briefly, a multiplex PCR based approach was used
576 according to the modified ARTIC version 2 protocol with version 3 primer set, and amplicon

577 libraries sequenced using MinION flow cells version 9.4.1 (Oxford Nanopore Technologies,
578 Oxford, UK). Sequences were made publicly available as part of COG-UK
579 (<https://www.cogconsortium.uk/>) via weekly uploads with linked metadata onto the MRC-CLIMB
580 server (<https://www.climb.ac.uk/>).

581
582 Samples collected via the CUH healthcare worker (HCW) screening programme were also
583 prioritised for on-site nanopore sequencing, as previously described [37]. This programme
584 entailed asymptomatic screening of selected wards, symptomatic testing of self-presenting HCW
585 and testing of symptomatic contacts of positive HCW. After a HCW tested positive, members of
586 the HCW screening team contacted the HCW and retrospectively collected data on symptom
587 onset date, symptomatology, household contacts, their job role, and which wards they had worked
588 in for the preceding two weeks. Most positive HCW could identify symptoms on retrospective
589 questioning, even if they were identified in the asymptomatic screening arm; however, a small
590 minority were genuinely asymptomatic and never went on to develop symptoms. HCW presenting
591 acutely to medical services at CUH were not part of the HCW screening programme, but were
592 identified as HCW from their medical records as part of hospital surveillance.

593
594 *Identifying hospital-associated outbreaks for investigation*

595
596 Patients tested at CUH were categorised on the basis of time between admission and first positive
597 swab into different groups reflecting the likelihood that their infection was community or hospital
598 acquired, as previously described (Meredith et al, LID 2020). The categories used were: 1)
599 Community onset, community associated (first positive sample <48 hours from admission and no
600 healthcare contact in the preceding 14 days); 2) Community onset, suspected healthcare
601 associated (first positive sample <48 hours from admission with healthcare contact in the
602 preceding 14 days); 3) Hospital onset, indeterminate healthcare associated (first positive sample

603 48 hours to 7 days post admission); 4) Hospital onset, suspected healthcare associated (first
604 positive sample 8 to 14 days post admission); 5) Hospital onset, healthcare associated (first
605 positive sample >14 days post admission); 6) HCW.

606

607 All CUH patients in categories 3, 4, and 5 (hospital onset with indeterminate, suspected or definite
608 healthcare associated COVID-19 infections) and 6 (HCW) were included for analysis and
609 integrated into the HCW screening dataset of positive HCW. The main wards the HCW had
610 worked in prior to testing positive and the ward where each patient had first tested positive were
611 used to identify ward clusters of hospital-associated infections. The ward clusters are named
612 anonymously here as Wards X and Y.

613

614 **Ethics statement**

615

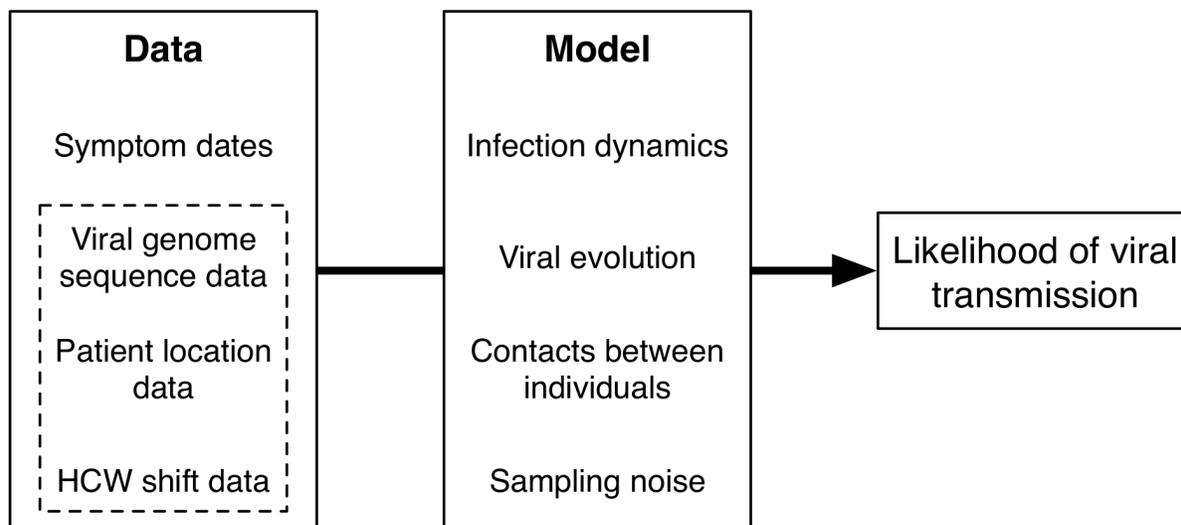
616 This study was conducted as part of surveillance for COVID-19 infections under the auspices of
617 Section 251 of the NHS Act 2006. It therefore did not require individual patient consent or ethical
618 approval. The COG-UK study protocol was approved by the Public Health England Research
619 Ethics Governance Group (reference: R&D NR0195).

620

621

622 **Figures**

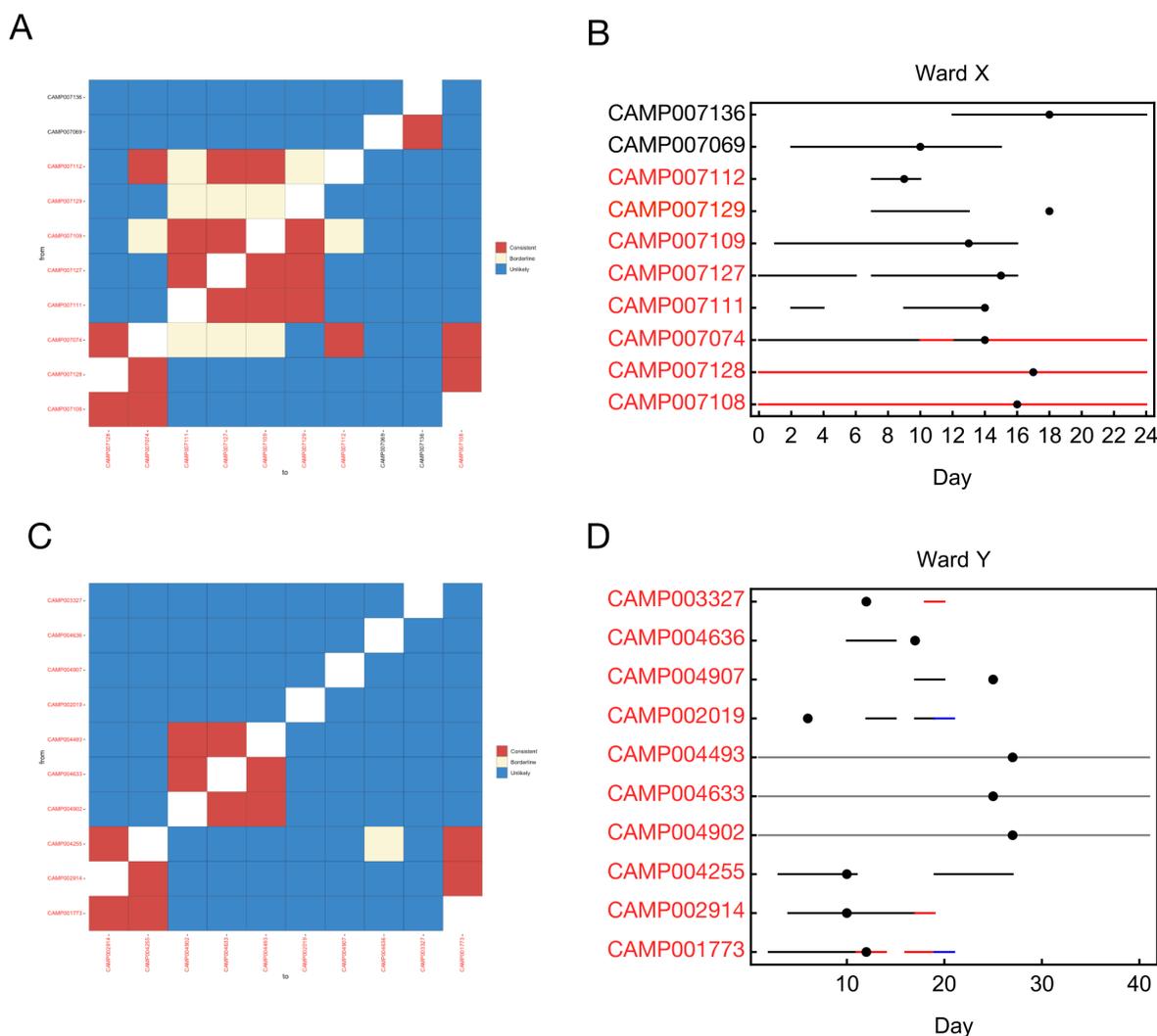
623



624

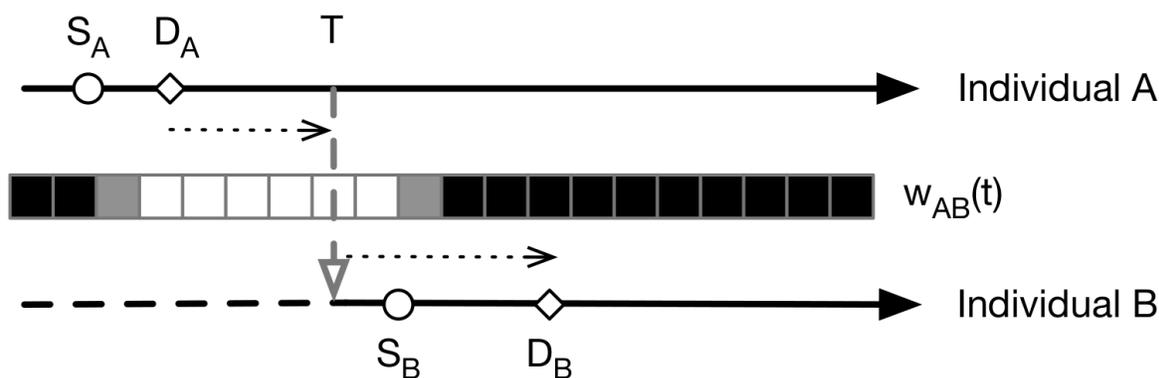
625

626 **Figure 1: Overview of our method.** Our approach estimates the likelihood that transmission
627 could have occurred between pairs of individuals. The model takes as input dates on which
628 individuals became symptomatic for COVID-19 infection. Further data which can be considered
629 includes viral genome sequence data, and time-resolved location data for each individual. Our
630 model combines details of COVID-19 infection dynamics with a model of viral evolution,
631 information about potential contacts between individuals, and measurement error in the sequence
632 data. Increasing amounts of data provide increasing amounts of resolution about the potential for
633 viral transmission.



634
 635 **Figure 2: Analysis of the full datasets collected from wards X and Y.** **A.** Output from the
 636 a2bcovid package given data from ward X. The plot shows potential links between cases,
 637 assessed in a pairwise fashion between potential donors (rows) and recipients (columns).
 638 Identifiers of individuals are coloured in either black (patients) or red (HCWs). Squares in the grid
 639 indicate that transmission from one individual to another is consistent with our model (red),
 640 borderline (yellow) or unlikely (blue). **B.** Locations of individuals linked to the ward X outbreak.
 641 Black lines indicate presence on ward X. Red lines indicate known household contacts between
 642 three individuals. Dots show times at which individuals first reported symptoms. **C.** Output from
 643 the a2bcovid package given data from ward Y. **D.** Locations of individuals linked to the ward Y

655 Y. The tree was constructed manually using a principle of maximum parsimony[38]. Red digits
656 indicate the number of substitutions between sequences from each individual.
657



S_i : Time of individual i becoming symptomatic

D_i : Time of collecting sequence from individual i

T : Time of transmission from i to j

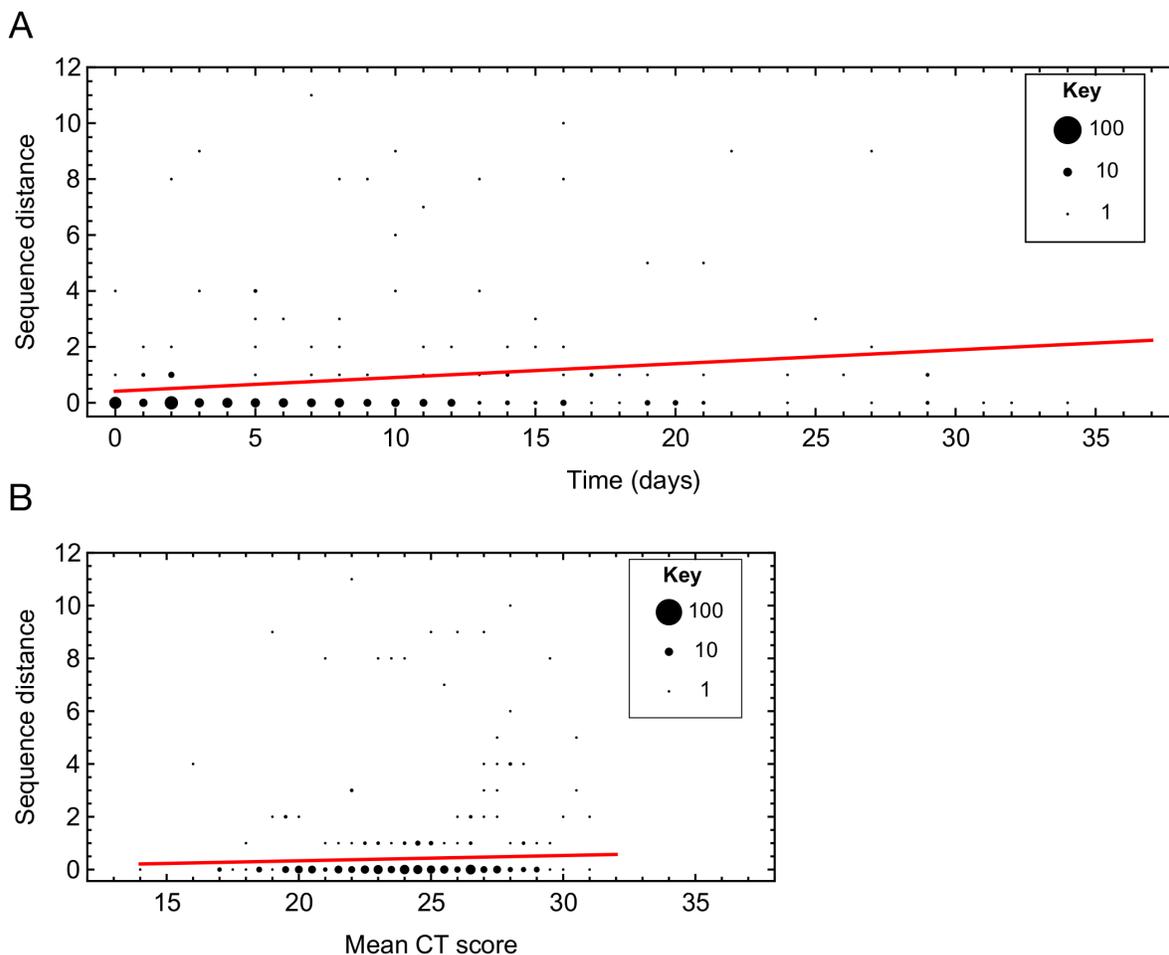
w_{ij} : Location weighting for individuals i and j

658
659 **Figure 4: Notation used in our method.** An overview of our model for transmission events is
660 shown in Figure 3. We divide time into discrete days. For the individual A, we denote by S_A the
661 date at which that individual became symptomatic, and by D_A the date at which a sample of viruses
662 were collected for genome sequencing. For each pair of individuals A and B we denote by $w_{AB}(t)$
663 the probability that A and B were co-located on day t . Within our model, we assume that dates of
664 sample collection are known, while times of symptom onset are known or estimated. Using these
665 statistics, in combination with viral sequence data, we calculate a statistic describing the potential
666 for individual A to infect individual B on any given day T . Summing this statistic across T , we
667 obtain an estimate of the consistency of our data with transmission having occurred between the
668 two individuals.

669

670 **Supporting Information**

671



672

673 **S1 Figure:** Analysis of Hamming distances between pairs of genome sequences collected from

674 viral samples in the same host. Figures show projections through a multi-linear model fit to the

675 data using a Poisson likelihood. **A.** Relationship between the Hamming distance and time

676 between samples. The line shows the fit to the data at the mean CT score. The size of a dot is

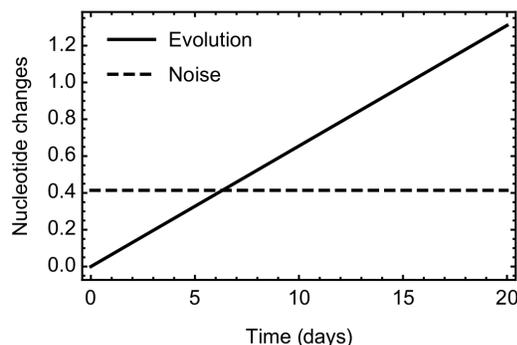
677 proportional to the number of pairs with given parameters. **B.** Relationship between the Hamming

678 distance and mean CT score of the two samples. The line shows the fit to the data calculated at

679 zero time between samples.

680

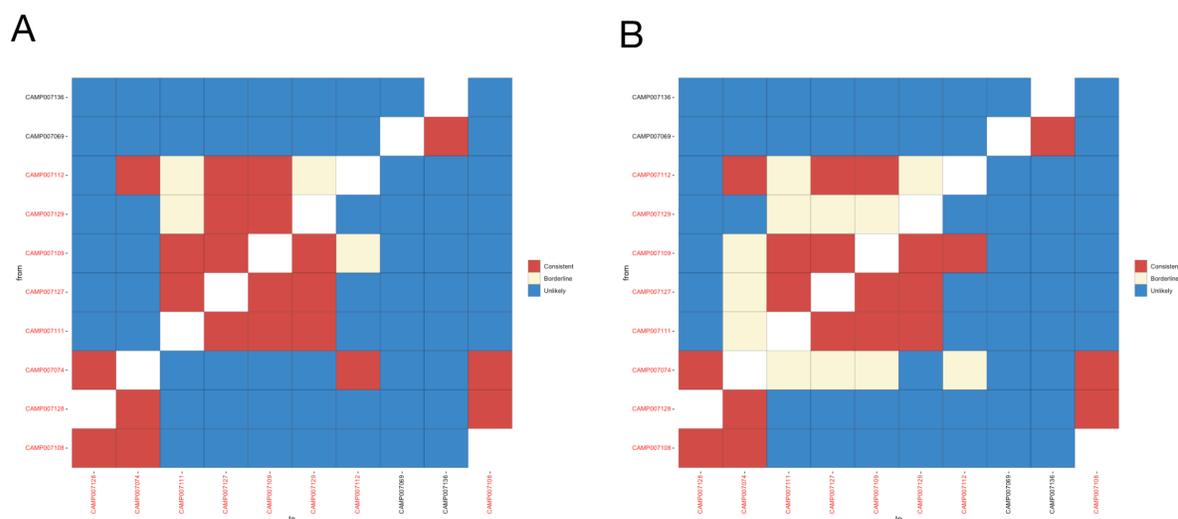
681



682

683 **S2 Figure:** Comparison between the expected rate of SARS-CoV-2 evolution within our model,
684 and the expected difference between two sequences caused by noise. The expected time
685 between symptoms being reported from individuals in a transmission pair is 5.7 days, in which
686 time the expected number of substitutions arising via evolution is 0.373. The expected number
687 of differences between two genome sequences resulting from error was estimated as 0.414.

688



689

690

691 **S3 Figure: Sensitivity of our results to changes in the noise parameter.** A. Inferences of
692 potential transmission events for ward X given a noise parameter of zero. B. Inferences of

706 equivalent symptom onset dates for individuals who were asymptomatic or for whom no data on
707 symptom onset date were available.

708

709 **S1 Table:** Parameters for the offset gamma distribution fitted to data describing intervals between
710 times of reporting symptoms and positive test results. Inferred values were generated using
711 maximum likelihood; the range describes a window of size two likelihood units from the maximum.

712

Model	Inferred distribution
Parameter	Value (Range)
α	2.593 (2.269, 2.976)
β	3.776 (3.292, 4.353)
Offset (σ)	3.112 (3.011, 3.431)

713

714 **S2 Table:** Error models fitted to Hamming distance data. A model incorporating dependence
715 upon the time between samples and upon viral load gave the best fit to the data. The error
716 parameter used was calculated at zero time between samples and at the mean viral load.

717

Model	Parameters			BIC
	Constant (λ)	Time dependence (γ)	CT dependence (α)	
Constant error	0.842			887.3

Time-dependent	0.385	0.0525		858.5
Time and CT dependent	-0.0692	0.0492	0.0200	857.8

718

719

720 **S1 Text: Further methodological details**

721

722 In the main text we stated that:

723

$$724 P(C_{AB} | X_T) = 0.5^{|C|-1} w_{AB}(T)$$

725

726 To derive this result we note that, if it is observed that $C_{AB}(T)=0$, transmission cannot occur at
727 time T , so that $P(C_{AB}|X_T)=0$. If it is observed that $C_{AB}(T)=1$, we next consider the element $C_{AB}(t)$

728 of C_{AB} for a time $t \neq T$. If $C_{AB}(t)$ is observed, we apply our approach to contact patterns, assuming

729 that $P(C_{AB}(t)=1|X_T) = P(C_{AB}(t)=0|X_T) = 0.5$, such that the probability of this observation is 0.5. If

730 $C_{AB}(t)$ is not observed, its probability is obtained by integration. We have that $P(C_{AB}(t) | X_T) =$

731 $w_{AB}(t) * 0.5 + (1 - w_{AB}(t)) * 0.5 = 0.5$. Hence if $C_{AB}(T)=1$, the probability $P(C_{AB} | X_T)$ of the whole

732 contact vector is equal to $0.5^{|C|-1}$. Finally, we consider the case in which $C_{AB}(T)$ is missing data.

733 Integrating over the missing value, we have that

734

$$735 P(C_{AB}(T) | X_T) = w_{AB}(T) * P(C_{AB} | X_T, C_{AB}(T)= 1) + (1 - w_{AB}(T)) * P(C_{AB} | X_T, C_{AB}(T)= 0) = w_{AB}(T)$$

736

737 Applying again the reasoning above, this gives us the result $P(C_{AB} | X_T) = 0.5^{|C|-1}w_{AB}(T)$. As we
738 defined $w_{AB}(T) = C_{AB}(T)$ when $C_{AB}(T)$ was observed, we thus have that $P(C_{AB} | X_T) = 0.5^{|C|-1}w_{AB}(T)$
739 in every case.

740

741 **S2 Text: GISAID identifiers for sequences used in this study**

742

743 **Ward X:** EPI_ISL_473479, EPI_ISL_473505, EPI_ISL_473478, EPI_ISL_473470,
744 EPI_ISL_473475, EPI_ISL_473464, EPI_ISL_473467, EPI_ISL_473466, EPI_ISL_473465,
745 EPI_ISL_473472, EPI_ISL_473471

746

747 **Ward Y:** EPI_ISL_425263, EPI_ISL_433686, EPI_ISL_444320, EPI_ISL_433740,
748 EPI_ISL_444407, EPI_ISL_433479, EPI_ISL_433779, EPI_ISL_433481, EPI_ISL_448052,
749 EPI_ISL_433492, EPI_ISL_433990

750

751 **Measurement error analysis:** EPI_ISL_444341, EPI_ISL_434058, EPI_ISL_438599,
752 EPI_ISL_425289, EPI_ISL_433900, EPI_ISL_425316, EPI_ISL_425424, EPI_ISL_433822,
753 EPI_ISL_433820, EPI_ISL_425314, EPI_ISL_447952, EPI_ISL_433796, EPI_ISL_433814,
754 EPI_ISL_433816, EPI_ISL_433666, EPI_ISL_425333, EPI_ISL_438723, EPI_ISL_433473,
755 EPI_ISL_434042, EPI_ISL_425334, EPI_ISL_433978, EPI_ISL_438648, EPI_ISL_434020,
756 EPI_ISL_444418, EPI_ISL_434059, EPI_ISL_438627, EPI_ISL_433671, EPI_ISL_433893,
757 EPI_ISL_433827, EPI_ISL_434004, EPI_ISL_433775, EPI_ISL_434060, EPI_ISL_433938,
758 EPI_ISL_444420, EPI_ISL_438714, EPI_ISL_448108, EPI_ISL_433895, EPI_ISL_438631,
759 EPI_ISL_438594, EPI_ISL_433911, EPI_ISL_444425, EPI_ISL_425309, EPI_ISL_433899,
760 EPI_ISL_433967, EPI_ISL_433681, EPI_ISL_425235, EPI_ISL_425259, EPI_ISL_425423,
761 EPI_ISL_425252, EPI_ISL_425251, EPI_ISL_438650, EPI_ISL_425274, EPI_ISL_425427,
762 EPI_ISL_425271, EPI_ISL_425270, EPI_ISL_425268, EPI_ISL_433673, EPI_ISL_433698,

763 EPI_ISL_433675, EPI_ISL_433697, EPI_ISL_425453, EPI_ISL_433679, EPI_ISL_433677,
764 EPI_ISL_433748, EPI_ISL_433784, EPI_ISL_438669, EPI_ISL_433672, EPI_ISL_434057,
765 EPI_ISL_433737, EPI_ISL_438580, EPI_ISL_438673, EPI_ISL_433477, EPI_ISL_433727,
766 EPI_ISL_433750, EPI_ISL_433752, EPI_ISL_433706, EPI_ISL_433707, EPI_ISL_433846,
767 EPI_ISL_433721, EPI_ISL_438706, EPI_ISL_433732, EPI_ISL_444416, EPI_ISL_433806,
768 EPI_ISL_434039, EPI_ISL_433966, EPI_ISL_433761, EPI_ISL_433768, EPI_ISL_433792,
769 EPI_ISL_433873, EPI_ISL_433881, EPI_ISL_433868, EPI_ISL_444329, EPI_ISL_444408,
770 EPI_ISL_433867, EPI_ISL_433878, EPI_ISL_433863, EPI_ISL_433467, EPI_ISL_433864,
771 EPI_ISL_433888, EPI_ISL_444403, EPI_ISL_434023, EPI_ISL_433892, EPI_ISL_433907,
772 EPI_ISL_433904, EPI_ISL_433906, EPI_ISL_438583, EPI_ISL_438647, EPI_ISL_444427,
773 EPI_ISL_433909, EPI_ISL_433920, EPI_ISL_438721, EPI_ISL_444413, EPI_ISL_433921,
774 EPI_ISL_433919, EPI_ISL_433980, EPI_ISL_433937, EPI_ISL_433993, EPI_ISL_433940,
775 EPI_ISL_434031, EPI_ISL_438586, EPI_ISL_433466, EPI_ISL_444419, EPI_ISL_433944,
776 EPI_ISL_433943, EPI_ISL_438651, EPI_ISL_433945, EPI_ISL_433969, EPI_ISL_438595,
777 EPI_ISL_434006, EPI_ISL_433972, EPI_ISL_434030, EPI_ISL_433986, EPI_ISL_434061,
778 EPI_ISL_433992, EPI_ISL_438668, EPI_ISL_438702, EPI_ISL_438649, EPI_ISL_444428,
779 EPI_ISL_434034, EPI_ISL_438578, EPI_ISL_438660, EPI_ISL_434036, EPI_ISL_438643,
780 EPI_ISL_444417, EPI_ISL_434015, EPI_ISL_433493, EPI_ISL_444316, EPI_ISL_434044,
781 EPI_ISL_438662, EPI_ISL_438622, EPI_ISL_438596, EPI_ISL_434045, EPI_ISL_444374,
782 EPI_ISL_434041, EPI_ISL_448052, EPI_ISL_433481, EPI_ISL_438632, EPI_ISL_433471,
783 EPI_ISL_444412, EPI_ISL_433468, EPI_ISL_444369, EPI_ISL_433478, EPI_ISL_433475,
784 EPI_ISL_438652, EPI_ISL_433476, EPI_ISL_438719, EPI_ISL_444326, EPI_ISL_447957,
785 EPI_ISL_444402, EPI_ISL_433474, EPI_ISL_444406, EPI_ISL_438670, EPI_ISL_444404,
786 EPI_ISL_444376, EPI_ISL_452887, EPI_ISL_438582, EPI_ISL_444429, EPI_ISL_444411,
787 EPI_ISL_447980, EPI_ISL_444422, EPI_ISL_444373, EPI_ISL_438672, EPI_ISL_444421,
788 EPI_ISL_477785, EPI_ISL_456720, EPI_ISL_448012, EPI_ISL_453003, EPI_ISL_456719,

789 EPI_ISL_438568, EPI_ISL_444324, EPI_ISL_444414, EPI_ISL_448110, EPI_ISL_438623,
790 EPI_ISL_438588, EPI_ISL_438592, EPI_ISL_444339, EPI_ISL_438609, EPI_ISL_438646,
791 EPI_ISL_438584, EPI_ISL_438728, EPI_ISL_438637, EPI_ISL_438591, EPI_ISL_438679,
792 EPI_ISL_438671, EPI_ISL_438593, EPI_ISL_438658, EPI_ISL_438656, EPI_ISL_447948,
793 EPI_ISL_438653, EPI_ISL_438645, EPI_ISL_438644, EPI_ISL_447974, EPI_ISL_438678,
794 EPI_ISL_438674, EPI_ISL_438676, EPI_ISL_438699, EPI_ISL_448016, EPI_ISL_438680,
795 EPI_ISL_447961, EPI_ISL_438701, EPI_ISL_438681, EPI_ISL_444371, EPI_ISL_447978,
796 EPI_ISL_448014, EPI_ISL_444323, EPI_ISL_438697, EPI_ISL_448087, EPI_ISL_438737,
797 EPI_ISL_452946, EPI_ISL_452962, EPI_ISL_444343, EPI_ISL_444361, EPI_ISL_456743,
798 EPI_ISL_447977, EPI_ISL_452997, EPI_ISL_447955, EPI_ISL_447990, EPI_ISL_452922,
799 EPI_ISL_453002, EPI_ISL_447994, EPI_ISL_456703, EPI_ISL_452973, EPI_ISL_452912,
800 EPI_ISL_448017, EPI_ISL_448036, EPI_ISL_456686, EPI_ISL_448092, EPI_ISL_473460,
801 EPI_ISL_448043, EPI_ISL_452911, EPI_ISL_452936, EPI_ISL_448050, EPI_ISL_456717,
802 EPI_ISL_448106, EPI_ISL_473469, EPI_ISL_456724, EPI_ISL_448105, EPI_ISL_452935,
803 EPI_ISL_456737, EPI_ISL_452868, EPI_ISL_448099, EPI_ISL_456741, EPI_ISL_452863,
804 EPI_ISL_452963, EPI_ISL_461562, EPI_ISL_452939, EPI_ISL_452938, EPI_ISL_452951,
805 EPI_ISL_456704, EPI_ISL_452994, EPI_ISL_456709, EPI_ISL_452989, EPI_ISL_452985,
806 EPI_ISL_461558, EPI_ISL_452984, EPI_ISL_461573, EPI_ISL_473451, EPI_ISL_456708,
807 EPI_ISL_456725, EPI_ISL_456728, EPI_ISL_456701, EPI_ISL_461585, EPI_ISL_456699,
808 EPI_ISL_461561, EPI_ISL_456749, EPI_ISL_473484, EPI_ISL_473485, EPI_ISL_456742,
809 EPI_ISL_461546, EPI_ISL_461587, EPI_ISL_461571, EPI_ISL_461554, EPI_ISL_473463,
810 EPI_ISL_461565, EPI_ISL_461568, EPI_ISL_473501, EPI_ISL_473473

811

812

813 **Author contributions**

814

Conceptualization	CI, WLH, BW, MR, AP, TG, DdA, MET
Data Curation	WLH, AP, LM, CJH, MH, AJ, MR, BW, LC, SC, AY, GH, FAK, TF, MP, IGe, YC, MC, SP, DS, LR, NJ, SS, SF, TD, KG, CW, EGK, NMB, MPW, SB, MET
Formal Analysis	CI, WLH, CJ, AP, BW, MR, MET
Funding Acquisition	SJP, IG, SB, MPW, MET, EGK
Investigation	WLH, AP, LM, CJH, MH, AJ, MR, BW, LC, SC, AY, GH, FAK, TF, MP, IGe, YC, MC, SP, DS, LR, NJ, SS, SF, TD, KG, CW, EGK, NMB, MPW, SB, MET
Methodology	CI, WLH, CJ
Project Administration	TG, IG, DdA, MET
Resources	MC, SP, NMB, MPW, SB, IG
Software	CI, WLH, CJ
Supervision	SJP, IG, TG, DdA, MET
Validation	CI, WLH, CJ, TG
Visualization	CI, WLH, CJ
Writing – Original Draft Preparation	CI, WLH, MET

Writing – Review & Editing	All authors
----------------------------	-------------

815

816

817 **Acknowledgements**

818

819 This work was funded by COG-UK, which is supported by funding from the Medical Research
820 Council (MRC) part of UK Research & Innovation (UKRI), the National Institute of Health
821 Research (NIHR) and Genome Research Limited, operating as the Wellcome Sanger Institute;
822 We also acknowledge the support from the Wellcome (Senior Clinical Fellowship to MPW (ref:
823 108070/Z/15/Z), Senior Research Fellowship to SB (ref: 215515/Z/19/Z), Senior Fellowship to IG
824 (ref: 097997/Z/11/Z); Collaborative Grant to CJH (ref: 204870/Z/16/Z); the Academy of Medical
825 Sciences & the Health Foundation (Clinician Scientist Fellowship to MET), the NIHR Cambridge
826 Biomedical Research Centre (to BW, MET) and the NIHR Clinical Research Network
827 Greenshoots award (to EGK). CJRI was supported by Deutsche Forschungsgemeinschaft (DFG)
828 Grant SFB 1310. We acknowledge MRC funding (ref: MC_UU_00002/11).

829

830 **Availability of code**

831

832 Our app is suitable for use with the R package and can be downloaded from
833 <http://github.com/chjackson/a2bcovid>.

834

835 **References**

836 1. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-
837 CoV-2. Nat Med. 2020;26: 450–452.

838 2. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real

- 839 time. *Lancet Infect Dis.* 2020;20: 533–534.
- 840 3. Danis K, Epaulard O, Bénét T, Gaymard A, Campoy S, Botelho-Nevers E, et al. Cluster of
841 Coronavirus Disease 2019 (COVID-19) in the French Alps, February 2020. *Clin Infect Dis.*
842 2020;71: 825–832.
- 843 4. Gao Y, Shi C, Chen Y, Shi P, Liu J, Xiao Y, et al. A cluster of the Corona Virus Disease
844 2019 caused by incubation period transmission in Wuxi, China. *J Infect.* 2020;80: 666–670.
- 845 5. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid
846 implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated
847 COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis.* 2020.
848 doi:10.1016/S1473-3099(20)30562-4
- 849 6. Kretzschmar ME, Rozhnova G, Bootsma MCJ, van Boven M, van de Wijgert JHHM, Bonten
850 MJM. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a
851 modelling study. *Lancet Public Health.* 2020;5: e452–e459.
- 852 7. Rickman HM, Rampling T, Shaw K, Martinez-Garcia G, Hail L, Coen P, et al. Nosocomial
853 Transmission of Coronavirus Disease 2019: A Retrospective Study of 66 Hospital-acquired
854 Cases in a London Teaching Hospital. *Clinical Infectious Diseases.* 2020.
855 doi:10.1093/cid/ciaa816
- 856 8. Wake RM, Morgan M, Choi J, Winn S. Reducing nosocomial transmission of COVID-19:
857 implementation of a COVID-19 triage system. *Clin Med .* 2020;20: e141–e145.
- 858 9. Lucey M, Macori G, Mullane N, Sutton-Fitzpatrick U, Gonzalez G, Coughlan S, et al.
859 Whole-genome sequencing to track SARS-CoV-2 transmission in nosocomial outbreaks.
860 *Clin Infect Dis.* 2020. doi:10.1093/cid/ciaa1433
- 861 10. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the
862 genomic era. *Trends Ecol Evol.* 2015;30: 306–313.
- 863 11. Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, et al. Transmission
864 clustering drives the onward spread of the HIV epidemic among men who have sex with
865 men in Quebec. *J Infect Dis.* 2011;204: 1115–1119.
- 866 12. Ragonnet-Cronin M, Hodcroft E, Hué S, Fearnhill E, Delpech V, Brown AJL, et al.
867 Automated analysis of phylogenetic clusters. *BMC Bioinformatics.* 2013;14: 317.
- 868 13. Jacka B, Applegate T, Krajdén M, Olmstead A, Harrigan PR, Marshall B, et al. Phylogenetic
869 clustering of hepatitis C virus among people who inject drugs in Vancouver, Canada.
870 *Hepatology.* 2014;60: 1571–1580.
- 871 14. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic
872 surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak.
873 *Science.* 2014;345: 1369–1372.
- 874 15. McCloskey RM, Poon AFY. A model-based clustering method to detect infectious disease
875 transmission outbreaks from sequence variation. *PLoS Comput Biol.* 2017;13: e1005868.
- 876 16. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in
877 Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med.* 2020;382: 1199–

- 878 1207.
- 879 17. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding
880 and transmissibility of COVID-19. *Nat Med.* 2020;26: 672–675.
- 881 18. Du Z, Xu X, Wu Y, Wang L, Cowling BJ, Meyers LA. Serial Interval of COVID-19 among
882 Publicly Reported Confirmed Cases. *Emerg Infect Dis.* 2020;26: 1341–1343.
- 883 19. Prete CA, Buss L, Dighe A, Porto VB, da Silva Candido D, Ghilardi F, et al. Serial Interval
884 Distribution of SARS-CoV-2 Infection in Brazil. *J Travel Med.* 2020.
885 doi:10.1093/jtm/taaa115
- 886 20. Lipsitch M, Donnelly CA, Fraser C, Blake IM, Cori A, Dorigatti I, et al. Potential Biases in
887 Estimating Absolute and Relative Case-Fatality Risks during Outbreaks. *PLoS Negl Trop*
888 *Dis.* 2015;9: e0003846.
- 889 21. Buitrago-Garcia D, Egli-Gany D, Counotte MJ, Hossmann S, Imeri H, Ipekci AM, et al.
890 Occurrence and transmission potential of asymptomatic and presymptomatic SARS-CoV-2
891 infections: A living systematic review and meta-analysis. *PLoS Med.* 2020;17: e1003346.
- 892 22. Beerenwinkel N, Zagordi O. Ultra-deep sequencing for the analysis of viral populations.
893 *Current Opinion in Virology.* 2011. pp. 413–418. doi:10.1016/j.coviro.2011.07.008
- 894 23. Laehnemann D, Borkhardt A, McHardy AC. Denoising DNA deep sequencing data-high-
895 throughput sequencing errors and their correction. *Brief Bioinform.* 2016;17: 154–179.
- 896 24. Sandmann S, de Graaf AO, Karimi M, van der Reijden BA, Hellström-Lindberg E, Jansen
897 JH, et al. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing
898 Data. *Sci Rep.* 2017;7: 43169.
- 899 25. Illingworth CJR, Roy S, Beale MA, Tutill H, Williams R, Breuer J. On the effective depth of
900 viral sequence data. *Virus Evol.* 2017;3: vex030.
- 901 26. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of
902 Transmission Routes Using Pathogen Deep-Sequence Data. *Am J Epidemiol.* 2017;186:
903 1209–1216.
- 904 27. De Maio N, Worby CJ, Wilson DJ, Stoesser N. Bayesian reconstruction of transmission
905 within outbreaks using genomic variants. *PLoS Comput Biol.* 2018;14: e1006117.
- 906 28. McCrone JT, Woods RJ, Martin ET, Malosh RE, Monto AS, Lauring AS. Stochastic
907 processes constrain the within and between host evolution of influenza virus. *Elife.* 2018;7.
908 doi:10.7554/eLife.35962
- 909 29. Lumby CK, Nene NR, Illingworth CJR. A novel framework for inferring parameters of
910 transmission from viral sequence data. *PLoS Genet.* 2018;14: e1007718.
- 911 30. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome
912 sequencing. *Nat Rev Microbiol.* 2017;15: 183–192.
- 913 31. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E,
914 et al. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease
915 Research and Public Health: Approaches, Applications, and Considerations for

- 916 Development of Laboratory Capacity. *J Infect Dis*. 2020;221: S292–S307.
- 917 32. Lumby CK, Zhao L, Breuer J, Illingworth CJ. A large effective population size for
918 established within-host influenza virus infection. *Elife*. 2020;9. doi:10.7554/eLife.56915
- 919 33. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-
920 time tracking of pathogen evolution. *Bioinformatics*. 2018;34: 4121–4123.
- 921 34. Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics*. 1978. pp. 461–
922 464. doi:10.1214/aos/1176344136
- 923 35. Ashcroft P, Huisman JS, Lehtinen S, Bouman JA, Althaus CL, Regoes RR, et al. COVID-19
924 infectivity profile correction. *Swiss medical weekly*. 2020. p. w20336.
- 925 36. Sridhar S, Forrest S, Kean I, Young J, Scott JB, Maes M, et al. A blueprint for the
926 implementation of a validated approach for the detection of SARS-Cov2 in clinical samples
927 in academic facilities. doi:10.1101/2020.04.14.041319
- 928 37. Rivett L, Sridhar S, Sparkes D, Routledge M, Jones NK, Forrest S, et al. Screening of
929 healthcare workers for SARS-CoV-2 highlights the role of asymptomatic carriage in COVID-
930 19 transmission. *Elife*. 2020;9. doi:10.7554/eLife.58728
- 931 38. Fitch WM. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree
932 Topology. *Systematic Zoology*. 1971. p. 406. doi:10.2307/2412116