

28 **ABSTRACT**

30 Determining the transmissibility, prevalence, and patterns of movement of SARS-CoV-2 infections
is central to our understanding of the impact of the pandemic, and to the design of effective control
32 strategies. Phylogenies (evolutionary trees), have provided key insights into the the international
spread of SARS-CoV-2 and enabled investigation of individual outbreaks and transmission chains
34 in specific settings. Phylodynamic approaches combine evolutionary, demographic and
epidemiological concepts and have helped track virus genetic changes, identify emerging variants,
36 and inform public health strategy. Here we review and synthesise studies that illustrate how
phylogenetic and phylodynamic techniques were applied during the first year of the pandemic, and
38 summarise their contributions to our understanding of SARS-CoV-2 transmission and control.

[H1] Introduction

40 The COVID-19 pandemic has triggered an unprecedented global response in pathogen genome
sequencing, and nearly 400,000 full or partial SARS-CoV-2 genomes were generated and shared
42 publicly within its first year. Although phylogenetic tools have become increasingly relevant to the
public health management of a range of viral epidemics¹⁻⁴, the COVID-19 crisis is the first global
44 health emergency during which large-scale, real-time genomic sequencing and analysis have
underpinned public health decisions. The first 12 months of the pandemic were characterized by
46 continual change in the global epidemiological and virological situation, and the analysis of genome
sequences was essential in tracking the changing situation. Phylogenetic and phylodynamic
48 approaches (Box 1) can unlock information contained in sampled genomes, and are often analysed
in conjunction with other data sources⁵. Such analyses have been used to quantify international
50 virus spread, identify outbreaks and transmission chains in specific settings, estimate growth rates
and reproduction numbers, account for surveillance gaps and lags, identify and track mutations of
52 interest, discover and analyse variants of concern, and investigate intra-host virus evolution.

54 This Review focuses on how SARS-CoV-2 transmission, epidemiology, and spatial dispersal has
been measured and investigated through phylogenetic and phylodynamic analyses of SARS-CoV-2
56 genomes (Fig. 1). It is intended to be a retrospective overview that uses examples from the first year
of the pandemic to demonstrate the contributions of phylogenetics in the context of different phases
58 of pandemic responses. We will examine how such analyses have informed global efforts to
understand, control and predict the pandemic, and outline arising new challenges and how they are
60 being addressed. We do not review events that precede the widespread emergence of SARS-CoV-2
(such as the evolutionary origins of the pandemic in non-human host species), nor its functional
62 genomics (i.e. how virus mutations contribute to phenotypes such as transmissibility). Given the
scale of the field and the size of the literature on SARS-CoV-2 genomic epidemiology, we do not
64 attempt to provide a systematic review. Instead we focus on studies that represent the first year of
the pandemic, which saw evolutionary approaches applied to a wide variety of public health
66 interventions worldwide, often in an *ad hoc* or pragmatic manner. We further highlight research that
was influential in contributing to epidemiological understanding and public health decision making.
68 The pandemic's first year also best illustrates the potential of these methods for urgent risk
assessment, prediction, and control of future emerging viruses. We mostly refer to the genetic
70 diversity of SARS-CoV-2 using the Pango dynamic nomenclature⁶ (see Box 1), but also sometimes
use the World Health Organization (WHO) "Greek letter" nomenclature scheme for particular
72 variants of concern (VOCs) and variants of interest (VOIs).

74 **[H1] Tracking the global pandemic**

Revealing how SARS-CoV-2 spread globally in early 2020 was important in informing public
76 health strategies. Phylodynamic, particularly phylogeographic, methods can be used to estimate the
timing and location of ancestral nodes within a molecular phylogeny⁷⁻⁹, allowing inference of the
78 route and rate of spread of pandemic lineages, from the site of its initial detection in Wuhan, China,
to the location of each sampled patient from which a virus genome was obtained.

80

[H2] International travel restrictions.

82 Phylogeographic studies have investigated the impact of international travel restrictions,
quantifying the number of lineage introductions from abroad and the relative contribution of local
84 transmission. For example, a global phylogeny of the pandemic showed that earlier lineages were
highly cosmopolitan whereas later lineages tended to be continent-specific, which may reflect the
86 rapid declines in mobility as many countries concurrently imposed restrictions on international
travel¹⁰, although early sampling in some countries may have been biased towards cases in
88 international travellers.

90 At the national scale, studies have typically observed reduced numbers of introductions along
international routes covered by travel restrictions; however, the overall effects of this on controlling
92 national transmission depended on the extent to which lineages were already locally well
established. During the global expansion of SARS-CoV-2, international exportations were driven
94 initially by dispersal from China, however the number of exports declined rapidly following the
cessation of China's international flights in January 2020¹¹. Endemic transmission began in Italy
96 during mid-February 2020, with establishment in other European countries soon thereafter¹². The
shift in global dissemination towards greater inter-continental exportation from Europe, was
98 associated with the expansion of a lineage bearing the D614G spike mutation¹³. Virus lineage
migrations from Europe to North America increased until the declaration by WHO of a pandemic
100 on March 11th 2020, suggesting that air travel restrictions subsequently slowed international
spread¹⁴. In South Africa, international introductions plummeted after travel restrictions began on
102 March 26th 2020¹⁵. Similar observations were made in other nationally-focused studies, including
those from Italy¹⁶, New Zealand, Australia, Iceland, Taiwan¹⁷, and the UK¹⁸.

104

The impact of international travel restrictions depended on the level of domestic transmission
106 control, and whether restrictions were implemented before full establishment of local transmission.
A study of 427 genomes from Brazil applied a **discrete asymmetric phylogeographic model [G]** and
108 estimated at least 104 international introductions during March and April 2020; these fell into three

monophyletic clusters (see Box 1) of apparently European origin, and a molecular clock approach
110 indicated that they arrived in late-February 2020. Domestic transmission in Brazil was already well
established by early March, suggesting that international restrictions implemented thereafter may
112 have had little impact¹⁹. In the US, an early study investigated the efficacy of international travel
restrictions in Connecticut²⁰. Seven of nine Connecticut genomes fell into a clade of mostly
114 Washington-state genomes, whereas two clustered with genomes from China and Europe. As the
Connecticut genomes were derived from people with no history of recent travel, their phylogenetic
116 placement in a cluster of genetically similar genomes indicated community transmission of
recently-imported lineages; again, flight restrictions may have been more effective in reducing cases
118 if implemented earlier²⁰. Similar patterns were observed in other countries, including Italy¹² and the
UK²¹.

120

Many countries strengthened travel restrictions later in 2020, aiming to slow the spread of variants
122 associated with changes in transmissibility (see the section “Tracking lineages of interest”). In
Brazil, a phylogeny of SARS-CoV-2 from cases detected in São Paulo in late December 2020
124 indicated two independent international introductions of lineage B.1.1.7 (the alpha VOC) from
London, UK²². These introductions occurred despite the suspension of flights to and from the UK.
126 Similarly, phylodynamics suggested multiple international introductions to the US and hidden
transmission of B.1.1.7 since November 2020, and that lineage B.1.1.7 expanded to 33 US states by
128 January 2021 with a doubling time of 9.8 days²³. Investigations have also considered the factors that
drove the resurgence of transmission in Europe in late summer 2020. A recent study using a
130 Bayesian time-scaled phylogeographic model (Box 1) found that by mid-August a large fraction of
the lineages then circulating in European countries had been introduced after June 15th, the date
132 when many countries in the Schengen area opened their borders²⁴. The study also found that newly
introduced lineages tended to expand faster when entering a region of low incidence, and that for
134 most countries resurgence was driven by new introductions rather than persistence of lineages from
the spring²⁴.

Phylogeographic inference of SARS-CoV-2 migration patterns have typically used either a discrete
trait analysis (DTA; for example, analyses with travel and/or mobility data)^{18,19,25} or structured
event-based birth-death (BD) models^{12,17,26}. The approaches differ in that DTA assigns discrete states
(locations) to nodes on a phylogeny, whereas structured models explicitly model migration events
and rates at a population level. The advantage of DTA is its relatively low computational demand²⁶
and its ability to incorporate discrete metadata, such as travel histories (e.g. [REF]²⁵), in a
straightforward manner. However, DTA does not accommodate the interdependency of tree shape

and migration rate or population size, and it is more difficult to interpret DTA model parameters²⁶. Structured BD approaches are more computationally costly, but can model variable sampling among regions (DTA is less robust to sampling patterns²⁷), and they infer parameters that can be more readily compared with those obtained from epidemiological or mobility data sets¹⁷.

136

[H2] Local transmission and interventions.

138 Non-pharmaceutical interventions (NPIs) include travel restrictions, person-to-person distancing, and mandatory mask wearing. Two phylogenetic approaches were typically adopted for
140 investigating the impact of NPIs. First, the frequencies of lineage movement among regions within a country were assessed using phylogeographic analyses (as discussed above for international
142 dissemination). Second, estimates of virus population size, epidemic doubling time, and R_t [G] were calculated from virus genome sequences using phylodynamic approaches.

144

Molecular clock dating of SARS-CoV-2 lineages indicated multiple introductions from Wuhan to
146 Guangdong in early January 2020, with a fall in lineage diversity thereafter, suggesting that within-country travel restrictions combined with comprehensive tracing and isolation in Guangdong were
148 effective in controlling transmission²⁸. A phylogenetic study of transmission in Boston, USA, also reported a drop in importations to Boston from other domestic locations after national restrictions
150 began¹⁴, and phylodynamic methods estimated a reduction in R_t in Israel of at least two-thirds, coincident with the imposition of quarantine measures²⁹. By contrast, a study of NPIs in Italy¹⁶
152 suggested that domestic travel restrictions failed to prevent community transmission, although there was evidence that transmission was inhibited and the relatively low genome sampling density
154 means that NPIs could have greatly restricted the transmission of lower frequency lineages. One global study of 29,000 SARS-CoV-2 genome sequences used a compartmental structured coalescent
156 model to estimate the time of epidemic seeding in 57 different locations³⁰. The authors found that locations with early implementation of strong NPIs experienced less severe morbidity and mortality
158 during the study³⁰ and that stringent interventions two weeks earlier would have approximately halved cumulative deaths in the immediate post-intervention period.

160

Phylodynamic BD models can estimate growth rates and other population parameters from genetic
162 sequence data. A BD-skyline approach estimated a national fall in R_t from 1.63 to 0.48 in Australia after the introduction of travel restrictions and social distancing on 27th March 2020³¹. Similar
164 approaches were used to show that R_t for a New Zealand transmission cluster fell from 7.0 to 0.2 during March 2020, demonstrating the impact of NPIs targeting this cluster³². A multitype BD
166 model applied to data from Taiwan showed a decrease in R_t throughout the early pandemic even in

the absence of substantially decreased local human mobility or stay-at-home orders¹⁷, suggesting
168 that interventions such as effective contact tracing and widespread face mask use may be sufficient
for adequate outbreak control. Phylodynamic studies have provided other parameter estimates that
170 are useful for understanding virus biology and transmission, or for use as statistical priors (Box 1)
in further Bayesian modelling (Table 1).

172

Phylodynamic analyses have repeatedly demonstrated hidden circulation of SARS-CoV-2 for days
174 to months prior to first-case detection. Such results are important in determining whether existing
surveillance adequately captures ongoing community transmission³³. A US study of 346 genomes,
176 covering January to mid-March 2020, examined the establishment of community transmission in
Washington State. A phylogeny consistent with community transmission was reported, with most
178 genomes clustered in a clade containing **WA1 [G]** (USA-WA1-2020, the genome of the first
detected US case). The estimated date of origin for the major clade was January 18th to February 9th
180 2020. This date was used to parameterize a stochastic epidemiological model that suggested 1600
active infections in Washington State by mid-March³⁴. Similarly, a molecular clock analysis of
182 genomes from Scotland estimated transmission began around February 19th 2020, predating first
case detection by almost two weeks²¹.

184

Phylogenetics and phylodynamics have also contributed near real-time insights that are suitable for
guiding the responses of public health authorities. Many investigations of hospital or event-
associated outbreaks during the pandemic employed phylogenetic methods and rapidly provided
actionable information. For example, phylogenetics supported public health examinations of
numerous outbreaks in New Zealand³⁵, and in the Netherlands excluded a church service, initially
implicated, as a source of an outbreak in a care home³⁶. Other studies influenced policy changes; a
phylogeographic study of the impact of travel restrictions on lineage imports and transmission³⁷ was
used to support the reintroduction of restrictions in Wales in October 2020. A phylogenetic
investigation of the June 2020 re-emergence of epidemic transmission in Australia implicated the
national mandatory hotel quarantine system, and the findings led to reform of the quarantine
programme. The study³⁸ also used phylodynamics to show the initial growth rate of the second
wave to be similar to that of B.1.1.7 emerging in the UK. Furthermore, phylodynamic detection of
the increased transmissibility of B.1.1.7 in England³⁹ contributed to the evidence base that informed
responses to the first VOC.

186 **[H1] Outbreak phylogenetics**

Evolutionary approaches can help refute or confirm suspected transmission routes, supplementing
188 our understanding from contact tracing of cases. Phylogenetic insights can reveal factors associated
with transmission, help to establish the polarity of transmission among individuals, and estimate
190 outbreak parameters. Genetic analyses have reconstructed events in travel-associated outbreaks and
can be used to cross-validate epidemiological records, helping to rule out spurious connections
192 among cases.

194 **[H2] Nosocomial transmission.**

Studies of healthcare settings are used to determine whether personal protective equipment (PPE)
196 guidelines are sufficient to prevent **nosocomial transmission [G]**. In Durban, South Africa, routine
surveillance identified a clade of cases in co-workers at a city hospital, suggesting a nosocomial
198 outbreak. The observation of community cases with additional mutations, implied community
transmission beyond the hospital⁴⁰. A lack of phylogenetic clustering of cases by ward among
200 healthcare workers in a Netherlands hospital showed community transmission to be more likely
than nosocomial transmission⁴¹. Furthermore, a study in Australia ruled out associations among 54
202 cases across four health services, where shared health care workers had been initially implicated in
dissemination. Phylogenetics revealed that the cases instead actually clustered according to a
204 common social event³¹.

206 At a UK renal unit, virus genomes were used to assign responsibility for an outbreak to a shared bus
service used to transport outpatients, rather than to transmission from in-patients. Rapid and
208 extensive sequencing resulted in timely revision of the hospital's infection control procedures⁴². In a
second UK study, phylogenetic analysis of infections from 31 care home staff and 61 residents
210 indicated transmission within, and possibly between, care homes, as well as from staff to staff —
the study supported the case against the use of locum staff in such settings⁴³. Policy change was also
212 called for in a Boston hospital study; virus genomes with shared **substitutions [G]** suggested at least
two patient-to-staff transmission events, despite an apparent lack of aerosol-generating procedures
214 and the staff wearing masks and face-shields⁴⁴. In Korea, the observation of eight near-identical
B.2.1 lineage genomes across two Seoul hospitals suggested that the outbreak in one hospital was
216 seeded by a patient transferred from the other⁴⁵. Multiple introductions were inferred for an
outbreak at a San Francisco nursing facility, with one worker, who had also worked in Washington
218 State, implicated in the introduction of WA1-related virus genomes⁴⁶. Other applications of
phylogenetics in investigations of outbreaks in medical or care settings are found in reports from
220 Chile⁴⁷, France⁴⁸, Minnesota⁴⁹, and the Netherlands³⁶. Nevertheless, although phylogenetics has
supported the confirmation of nosocomial transmission in some cases, it has also helped reveal the

222 contributions of wider social contact, outside of hospitals and care homes, in the maintenance of
transmission networks that span nosocomial settings.

224

[H2] Public gatherings and super-spreading.

226 Epidemiological studies of SARS-CoV-2 have indicated a relatively high **attack rate [G]**^{50,51} and
phylogenetics has corroborated this finding. For example, an outbreak affecting 11 workers in a
228 large open-plan office in Sweden was supported by a phylogenetic clade of virus genomes from
eight workers (six genomes were identical, and two near-identical)⁵². In some cases, local bursts of
230 transmission appear to precede national-scale transmission. Phylogenetic analysis of the early
epidemic in Boston identified 28 cases from an international business conference that formed a
232 cluster. All cases shared a novel C2416T non-synonymous substitution and by November 2020
genomes containing this substitution appeared to underlie 35% of Boston's cases, and 1.9% of USA
234 genomes¹⁴. This finding showed that individual mass-infection events could facilitate transmission
and virus dissemination.

236

The role of large celebrations in triggering super-spreading has also been explored. Discrete-state
238 phylogeography was used to suggest that a *Mardi Gras*-associated super-spreading event led to
outward (inter-State) dissemination in the southern US, and the acceleration of the early epidemic
240 there⁵³. Resurgence of an early outbreak in Japan was hypothesized initially to be linked to
increased travel to cherry blossom sites during the national holiday of March 20th to 22nd 2020.
242 Clarification through sequencing later showed that the late March cases were not directly related to
cases from the first epidemic 'wave'⁵⁴. In Germany, three events at a Berlin nightclub in early
244 March 2020 led to a series of outbreaks. Phylogenetics confirmed the club as a potential focus of
super-spreading; Germany decided to prohibit such events from March 16th⁵⁵. In the US,
246 phylodynamics linked the establishment of B.1.1.7 to the Thanksgiving holiday travel surge in
November 2020²³.

248

[H2] Travel and transport.

250 The contribution of transport settings to SARS-CoV-2 transmission has been keenly debated. Virus
genomes supported the case for in-flight transmission on a Massachusetts to Hong Kong flight; two
252 flight-attendants and two related passengers were detected with B.1 lineage infections, despite B.1
being unknown in Hong Kong at that time⁵⁶. A similar indication of in-flight transmission was
254 reported for a flight between Dubai (United Arab Emirates) and Auckland (New Zealand)

⁵⁷.

256

258 The predominance of one major clade in the February 2020 Diamond Princess cruise ship outbreak,
suggested that most passengers became infected while attending on-board events, with a single
introduction prior to quarantine measures⁵⁸. Similarly, a phylogenetic study involving samples from
260 northern California and outbreaks on two consecutive cruises of the Grand Princess ship, with a
common crew, found that infected passengers carried three substitutions characteristic of WA1.
262 WA1 at that time was dominant in Washington State, and all cases sampled from the Grand Princess
also shared two substitutions that were common in WA1 viruses then circulating in Washington and
264 California. This suggested that the source(s) of infection on the cruise were more likely local (i.e.
California), than either of the cruise destinations. The second cruise, immediately following the first
266 outbreak, shared a subset of passengers with the first cruise. The outbreak phylogeny indicated that
one of the first-cruise genomes was ancestral to the second-cruise genomes, and also to Californian
268 WA1 genomes in general. This suggested that the shared cohort of passengers seeded the outbreak
on the second cruise⁵⁹. The patterns of shared, derived, mutations in the Grand Princess outbreaks
270 imply large numbers of infections from probably a single infected passenger or crew member (or
related transmission cluster), and that the source of infection was local, for example a crew member,
272 rather than from any station of disembarkation. The implications are that revision of infection
management procedures and practices were essential to protect passengers early during the
274 pandemic.

Genomic analyses aided the tracing of transmission during a Chinese-German business meeting in
276 greater Munich (January 19th to 22nd 2020), which began an outbreak in Bavaria and involved 16
cases (detected from January 27th to February 11th 2020). Genomes indicated that transmission may
278 have occurred in the pre-symptomatic phase of infection between two individuals who sat briefly
back-to-back in a canteen. Sequencing helped refine estimates of incubation periods and attack rate,
280 and revealed the order of transmissions in a subsequent household cluster⁶⁰.

282 The ability of virus genomes to distinguish prolonged infection from cases of reinfection clarifies
the reconstruction of transmission chains, and is crucial to understanding why some people
284 repeatedly test virus-positive. Similarly, co-infection with more than one virus phylogenetic lineage
in a host at the same time could mask an international lineage introduction. Sequencing supported
286 reinfection of an air traveller to Hong Kong (from Spain, via the UK) who had a high viral load and
a B.1.79 lineage infection in August 2020; the same passenger had a B.2 lineage infection in March
288 2020 and was reverse transcription PCR (RT-PCR) negative in mid-April 2020⁶¹ (see also a similar
case from the US⁶²).

290

[H1] Tracking lineages of interest

292 VOCs are genetic variants of SARS-CoV-2 that carry mutations that are known or suspected to
affect key virus phenotypes such as increased transmissibility or immune escape. Phylogenetic
294 analysis has revealed the independent emergence of VOCs, some of which share identical mutations
(**evolutionary convergence [G]**), and has reconstructed the accumulation of substitutions in time
296 and space, shedding light on virus evolutionary or adaptive strategies.

298 The end of 2020 saw the discovery of the first VOCs, with multiple instances of convergent
molecular evolution among them (Fig. 2; see the next section). For example, lineage B.1.1.7 (first
300 labelled VOC 202012/01 and now termed VOC alpha) (Table 2) was determined by Public Health
England to be a VOC on December 21st 2020 because its increase in frequency appeared to be
302 related to the presence of particular genetic changes in the virus' spike protein that had already been
implicated in greater transmissibility (e.g. N501Y and P681H) and antibody escape (e.g. deletion
304 $\Delta 69/\Delta 70$)⁶³. Lineage B.1.1.7 became dominant in the UK just a few months after its emergence, and
phylogenetic studies showed it to have an estimated growth rate 40–70% higher than previous
306 lineages²⁵. In the global SARS-CoV-2 phylogeny, B.1.1.7 descends from the B.1.1 parental lineage
via a long branch, suggesting that either the immediate ancestors of B.1.1.7 were unsampled, or that
308 the variant arose through a discrete evolutionary event during which multiple mutations were
acquired, possibly during protracted infection of a single patient⁶⁴. Slightly before the emergence of
310 B.1.1.7, the N501Y spike mutation was detected in an independent lineage in South Africa. This
lineage, B.1.351 (VOC 501Y.V2, now named VOC beta) also carried mutation E484K in the
312 receptor-binding domain (RBD) of its spike protein¹⁵.

314 Phylogenetics can help reveal the order in which variants accrue substitutions, which could provide
clues to the functional advantages of convergent variants. For example, a phylogeny for the then
316 emerging P.1 VOC (now named VOC gamma) indicated that the lineage's characteristic mutations
were gained in two phases, with a molecular clock analysis suggesting an intervening gap of several
318 months⁶⁵. Similarly, the nascent lineage B.1.351 detected in samples taken in South Africa during
October 2020, lacked L18F, R246I and K417N; the latter substitution is among the nine changes
320 that define B.1.351 and appeared in samples from the lineage in November 2020⁶⁶. Nevertheless, it
is sometimes impossible to resolve the order of evolutionary events, either because genome
322 sampling through time is insufficiently frequent or several mutations occurred very quickly. For
example, $\Delta H69/V70$ has arisen independently in several lineages (Fig. 3a), and is thought to
324 compensate for decreased infectivity due to antibody escape substitutions such as N501Y; however,
it is currently not clear whether or not the deletion preceded the RBD substitutions in B.1.1.7⁶⁷. The

326 sudden appearance of lineages with constellations of 30 or so key substitutions relative to ancestral
genomes, is unlikely *a priori* given the low long-term substitution rate of SARS-CoV-2. The recent
328 emergence of BA.1 (VOC omicron) has reignited interest in this; evolution during a prolonged
infection of an immunocompromised patient, or isolation within and then reintroduction from an
330 unsampled human or animal population are being considered as hypotheses for the origins of
omicron⁶⁸.

332

The E484K mutation in B.1.351 has been associated with antibody escape and potential resistance
334 to **convalescent plasma [G]** therapies^{63,69}. *In vitro*, B.1.351 exhibits improved ability to escape
antibody responses targeted at VOCs arising earlier in the pandemic, such as B.1.1.7, (an escape
336 phenotype mostly attributed to E484K and K417N)^{70,71} and shows increased transmissibility⁷⁰.
Although the B.1.1.7 lineage did not carry E484K when it first emerged, by February 1st 2021 this
338 mutation had appeared in thirteen English and two Welsh B.1.1.7 genomes. The phylogenetic
relationships among these genomes suggested at least two independent acquisitions of E484K in the
340 UK. Lentiviral and vesicular stomatitis virus (VSV) **pseudotyping experiments [G]** indicate that the
E484K mutation on the B.1.1.7 lineage backbone results in a reduction of neutralizing activity by
342 vaccine sera^{71,72}. The P.1 lineage was first reported in international travellers from Brazil entering
Japan⁷³, and showed 11 amino acid substitutions relative to its ancestral lineage B.1.1.28. Three of
344 these substitutions fall within the RBD (K417T, E484K, and N501Y), and all three sites are also
modified in B.1.351 and some B.1.1.7 lineages⁶³. P.1 appears to have originated in Brazil^{73,74} and
346 also shows signs of increased transmissibility relative to its parental lineage B.1.1.28⁷⁵.

348 Although the phenotypic effect of mutations carried by VOCs can be investigated *in vitro* (see
REFS⁷⁶⁻⁷⁹ for examples), their epidemiological significance is harder to evaluate. Changes in
350 mutation frequency during an emerging epidemic may not always directly reflect transmission
potential or selective advantage, because they can be also influenced by **founder effects [G]**,
352 genetic linkage to other mutations, ascertainment bias, and uneven sampling among regions³³.
Studies with a phylogenetic or phylodynamic basis have the potential to ameliorate some of these
354 issues. The first amino acid replacement substitution to show a marked change in prevalence was
D614G. Globally, SARS-CoV-2 with glycine (G) at spike position 614 rose from 10% prevalence
356 before March 1st 2020, to overall global predominance by April 2020⁷⁷. Relative growth rates for
D614G and other substitutions were estimated by phylogenetic diversification; this suggested that
358 most variants were weakly deleterious, and not more transmissible⁸⁰. Sequence data from repeated
international introductions of SARS-CoV-2 to the UK were leveraged to provide replicate
360 observations of the growth of 614D and 614G lineages⁸¹. Modelling and phylodynamic analyses of

307 independent introductions between January 29th and June 16th, 2020 suggested a genuine (i.e.
362 not a sampling effect) replacement of D by G in the UK, with a growth effect of around 20% and
phylogenetic estimates of R_0 of 2.7–3.5 for 614D and 3.1–4.8 for 614G; however, indications of
364 positive selection for 614G were not significant in all analyses. A separate analysis suggested that
founder effects were responsible for the apparent selective advantage of 614G⁸², noting that the
366 expansion of 614G coincided with a shift in the nexus of global dispersal from Asia to Europe.

Deep-sequencing and phylogenetics have also been used to track virus evolution during co-
368 infections⁸³ (Fig. 4a) and prolonged infections⁸⁴ (Fig. 4b), and to distinguish chronic infections from
re-infections. In addition, phylogenetic assumptions regarding the distribution and independence of
370 mutations can be violated by virus mutational patterns related to host antiviral defences⁸⁵ (Fig. 4c).
Within-host variation also has implications for SARS-CoV-2 phylogenetics, as co-infections may
372 complicate tracing of transmission networks⁸³ (Fig. 4d).

374 **[H1] Homoplasmy and recombination**

Lineages bearing N501Y and E484K appeared independently in Brazil, South Africa, Canada, and
376 the UK in late 2020. Evolutionary convergence was observed, with the same changes being
acquired independently on several branches scattered across the virus phylogeny (homoplasmy) (Fig.
378 3a), and several lineages may share one or more substitutions (Fig. 3b). For example, both B.1.351
and P.1 (VOCs beta and gamma) showed escape-associated RBD substitutions at sites 417, 484, and
380 501 (Fig. 3b), as well as at positions 614 and 701 in the spike protein, but these two lineages do not
share immediate common ancestry. The concurrent emergence and spread of the same mutations in
382 different places and on different genomic backgrounds suggests there were shared selective
pressures acting on the virus⁸⁶, such as the need to increase intrinsic transmissibility, extend the
384 duration of infection, or to evade host immune responses (whether elicited by natural infection or
vaccination)⁸⁷. The parallel emergence of constellations of functionally-relevant mutations⁸⁸ further
386 suggests the existence of fitness interactions (epistasis) among them. Some mutations may only
grow to a detectable population frequency if preceded, or closely followed by, a second permissive
388 or compensatory mutation— several such mutations have been suggested in SARS-CoV-2^{67,89–91}.

390 The epidemiological and phylogenetic context of these convergent changes indicates that they arose
through independent, parallel mutation. However, it is known that such changes (homoplasies) can
392 arise also through recombination, and evolutionary analyses suggest that recombination could be
now relevant to SARS-CoV-2 evolution⁹². The level, scale and consequences of recombination
394 during the pandemic are unclear; one earlier study of phylogenetic inconsistency found no clear

signals of recombination⁹³, whereas a more recent analysis of UK sequence data discovered at least
396 four groups of natural recombinants of B.1.1.7, and other parental lineages⁹⁴. The increasing co-
circulation in 2021 of genetically-diverse viruses increases the likelihood that further SARS-CoV-2
398 recombinants will be detected⁹⁵.

400 **[H1] Tackling sampling bias in genomic epidemiology**

Uneven sampling of genomes has emerged as an issue for SARS-CoV-2 phylogenetics. Sampling
402 was effectively absent during the first days and weeks, becoming more extensive as the pandemic
progressed⁹⁶, and often concentrated toward particularly large outbreaks or the radiation of VOCs.
404 Some countries sequence routinely, others only for outbreak investigation, and some not at all. The
UK and Danish virus sequencing programmes are examples of large-scale, sustained sampling
406 intensity, and they have allowed quantitative assessments of virus properties and public health
interventions.

408

Although these and other countries have generated and openly shared many virus genomes⁹⁷, even
large-scale genomic programmes sometimes achieve only moderate sampling densities. For
example, models of total UK infections to 4th May, 2020 suggested 3.4 million people were
infected, of these, around 0.3% had their viral genomes sequenced⁹⁸. The impact of low and uneven
sampling intensity on SARS-CoV-2 phylogeography has been recognized^{25,99}, and is commonly
addressed through downsampling or a bespoke sampling regimen⁹⁹. Some phylodynamic models
allow for explicit modelling of sampling bias, and this has enhanced several SARS-CoV-2
studies^{12,25,100}. In some cases, a single uniquely-shared variant may be sufficient to determine the
origin of a transmission chain^{60,101}, and in studies in which one or several regions or periods were
poorly sampled, unsampled individuals have been modelled and added to analyses¹⁰², or unobserved
ancestral locations jointly inferred using phylogenies¹⁰³. Delays between case detection¹⁰⁴ and
genome sequence availability, as well as insufficiently dense sampling can hinder outbreak analysis.
Further study of sampling effects is required and there is a need for standard definitions of sampling
schemes to minimize bias in large-scale analyses while maintaining practicality⁹⁹.

Ascertainment bias towards symptomatic cases potentially complicated attempts to determine any
410 greater transmissibility conferred by spike 614G over 614D⁸¹, and undersampling in a region of
high incidence has been suggested as a cause of overestimation of size and duration of SARS-CoV-
412 2 transmission chains²⁸. A generalized approach to assessing the impact of incomplete sampling on
outbreak analyses has been proposed, based on the probability of correctly detecting linked
414 transmission pairs given a certain level of sampling effort¹⁰⁵. Genetic detection of viruses in

wastewater has the potential to augment genome sampling before first case detection, and to
416 estimate the impacts of NPIs on local and regional prevalence¹⁰⁶. Recently, SARS-CoV-2 sequences
from wastewater samples have indicated genomes with novel constellations of mutations not seen in
418 any known circulating lineage or VOC, including ‘cryptic lineages’ with multiple immune escape
mutations¹⁰⁷, with some studies using phylogenetics to place cryptic wastewater lineages in relation
420 to known lineages¹⁰⁸.

422 Uneven sampling through time can be addressed by adding an explicit sampling model to
phylodynamic inference. One current solution uses structured (epoch based) models to condition on
424 the rate of genomic sampling relative to all PCR-confirmed SARS-CoV-2 cases, and reportedly
improve molecular clock accuracy¹⁰⁰. Methods that can accommodate changing rate of sequencing
426 through time have been developed, for example the coalescent-based Bayesian Epoch Skyline Plot
(ESP)¹⁰⁹ (see also Box 1), an approach analogous to the classic birth–death skyline¹¹⁰. An alternative
428 is to model sampling while linking sample location to regional variations in sampling effort; this
has improved estimation of population size history for at least some datasets¹¹¹, and the BD-skyline
430 with variable sampling rate has also been applied to SARS-CoV-2^{100,112}. The relationship between
genetic variation and transmission patterns are interdependent, and therefore combining
432 phylodynamic estimation with epidemiological data should generate stronger inferences. There has
been notable progress on such integrated approaches. One recent method¹¹³ allowed the
434 incorporation of non-genomic incidence data and epidemic dynamics models with a novel
phylodynamic approach that represented both original and downstream members of transmission
436 chains (i.e. phylogenies with extant internal nodes). This joint epidemiological and phylodynamic
approach is reportedly less susceptible to bias arising through undiagnosed cases, imported cases
438 and changes in sampling levels, and so produces more reliable estimates of transmission rates than
epidemiological data alone¹¹³. Analytical methods for *a priori* estimation of appropriate sampling
440 intensity, sizes and strategy for analyses in virus genomic epidemiology are urgently required but
not well developed. Such methods would reduce sequencing costs for longer-term initiatives as well
442 as help ameliorate sampling bias¹⁰⁵ Guidance is being developed to ensure that project objectives
are considered and addressed using economically efficient genome sampling and sequencing
444 approaches¹¹⁴.

446 Undetected SARS-CoV-2 transmission has been incorporated into standard Bayesian phylogenetic
inference using an epidemiological model that includes data on confirmed, but unsequenced, cases
448 and combines molecular sequence, case count data and temporal information. The model can infer
undetected transmission or track changes following interventions, and can also incorporate changes

450 in sampling strategy, such as a decision to begin testing asymptomatic individuals¹¹². The approach
has been successfully used to infer R_0 [G] and cumulative case count trajectories for the Diamond
452 Princess cruise ship outbreak, a closed system for which reliable epidemiological (non-genetic) data
are available to validate corresponding phylodynamic estimates^{112,115}. These, and similar efforts to
454 unite phylogenetics and epidemiology, are promising tools for the study of virus epidemics,
including SARS-CoV-2¹¹⁶.

456 Some phylodynamics approaches, especially those based on Bayesian inference, require substantial
computation time, and population genetics could provide simpler evaluations of estimated
458 population size¹¹⁷. One such method applied to SARS-CoV-2 used the relationship between number
of haplotypes and number of sequences deviating from a reference to estimate **effective population**
460 **size [G]** (N_e) from sequence data¹¹⁸. The size trajectories estimated were similar to those inferred
using phylodynamic methods. Such methods may be useful for rapid assessment, as they can be
462 computed within high-throughput pipelines or in simulations. Subsampling of large data sets in
virus genomic epidemiology has been a popular solution to reduce computational cost and to
464 ameliorate differences in sampling intensity among countries. Although a few studies have explored
how to optimize downsampling for parameter estimation (e.g. REF¹¹⁹), there is a current lack of
466 formal methods for downsampling model selection and implementation, and alternative strategies
have been discussed⁹⁹.

468

Further innovation is nevertheless required, particularly concerning the estimation of large virus
470 phylogenies. Several studies have noted that the comparatively low substitution rate of SARS-CoV-
2 reduces phylogenetic signal, potentially hampering studies of events early in the pandemic or of
472 local epidemics¹²⁰, hindering estimation of substitution rates¹²¹, and lowering phylogenetic
resolution^{122,123}. The effect was seen in the accumulation of sequence data for Washington State that
474 overturned the initial suggestion of prolonged cryptic transmission in the state, by linking the first
and second outbreaks there³⁴, to favour independent international introductions as founding the later
476 outbreaks¹⁰². Solutions to these problems include the use of sets of plausible trees, rather than a
single tree¹²⁰, testing of alternative root-placements¹²³, and randomisation tests for phylogenetic
478 signal¹²⁰. Consideration is also being given to rapid generation of maximally stable topologies, from
multiple studies based on different data, methods and assumptions, implemented using entropy-
480 weighted tree distances to highlight the least stable clades¹²⁴. Some methods in virus genomic
epidemiology have inherent assumptions such as negligible variation within patients, and absence
482 of **superinfections [G]** — that may not hold for SARS-CoV-2. The analyses being applied to the
virus matured throughout the first year of the pandemic, and solutions arose from across diverse

484 biological science disciplines, often in a highly collaborative manner. For example, approaches to
quasispecies [G] deconvolution were adopted from practices in oncology^{125,126}. In addition, there is a
486 vast literature devoted to the reduction of technical error and improvement of genome sequence
quality^{127,128}.

488

[H1] Conclusions and the way forwards

490 The contributions of evolutionary analyses to the global pandemic response are substantial and
varied. The first year of the SARS-CoV-2 pandemic highlighted the progress that has been made
492 over the past decade in virus genomics and phylodynamic analyses, while revealing technical and
social challenges that remain to be addressed. The rapid, open sharing of protocols and data has
494 been critically important, and more extensive for SARS-CoV-2 than ever before, yet hesitancy to
share sequencing data prior to publication remains¹⁰⁴ because of concerns that data may be used
496 elsewhere without appropriate credit being given to producers¹²⁹. Greater insights into SARS-CoV-2
transmission could be gained through the incorporation of more and varied data (e.g., mobility
498 data); however, this must be balanced with privacy and anonymization concerns. Flexible and
robust methods for incorporation of diverse metadata into phylodynamic analyses are also required,
500 as are standards for their collection and availability⁹⁹.

502 In addition, the nomenclature of lineages and variants was initially inconsistent; this complicated
scientific discussion, and encouraged the media to adopt simple but inappropriate naming of
504 lineages based on the location of their first detection (e.g., “South Africa variant”)¹³⁰. The problem
of toponymic naming in the popular literature has been partly overcome by the adoption of Greek
506 letter designation for VOCs and VOIs by the WHO, with the Pango nomenclature adopted by
researchers requiring a systematic nomenclature or for epidemiologically-relevant lineages.
508 Nevertheless, some confusion can still arise between the possible naming of recurring constellations
of variants by the WHO, and their phylogenetic context, as indicated by a Pango designation¹³¹.

510

In many countries, current research recruitment, evaluation and funding frameworks disincentivize
512 the long-term participation of researchers with phylodynamic analysis skills in public health
surveillance and control, because such participation diverts from those activities that are used to
514 evaluate career progress (e.g. research publications and grants)¹³². Consequently, new career
pathways or evaluation systems are required to encourage greater embedding of evolutionary
516 genomic approaches in public health. Investment in the training and retention of those with
bioinformatic and phylogenetic expertise is required in many low and middle income countries,
518 where the capacity for computational analysis sometimes lags behind that for genetic sequencing¹¹⁴.

520 Further investigation into these ethical and technical challenges is needed to prepare for future
pandemics, and to sustain our tracking of SARS-CoV-2, transmission, new VOCs, new
522 recombinants, and cross-species transmission events.

524 Phylodynamics has demonstrated the impact of interventions and highlighted cases where they
could have been applied more effectively or their use better timed. Phylogenetics has distinguished
526 local onward transmission from new introductions, and thereby informed infection control and
planning. The history of pandemic transmission is recorded in virus genomes, allowing a global
528 overview of virus epidemiology to be obtained even with samples taken in limited geographical
areas or unevenly through time. Accordingly, phylogenetic concepts are likely to continue to play an
important role in efforts to combat SARS-CoV-2 and in the prediction of the virus' next move.

530

Acknowledgements

532 OGP acknowledges the support of the Oxford Martin School. SWA was supported by the COVID-
19 Genomics UK (COG-UK) Consortium (<https://www.cogconsortium.uk>). SCH was supported by
534 the Wellcome Trust [Sir Henry Wellcome Postdoctoral Fellowship 220414/Z/20/Z].

Author contributions

536 S.W.A. researched data for the article. S.W.A., S.C.H. and O.G.P. wrote the article. S.W.A., D.M.A., T.R.C. and O.G.P. reviewed/edited the manuscript before
submission. All authors substantially contributed to the discussion of content.

Competing interests

538 The authors declare no competing interests.

Peer review information

540 *Nature Reviews Genetics* thanks Sebastián Duchêne, Damien Tully and David Welch for their contribution to the peer review of this work.

542

Table 1 SARS-CoV-2 epidemiological parameter estimates using phylodynamic approaches

Region	Period	Reproduction number	Substitution Rate (changes / site / yr)	Method
Australia	24/03–29/04	$R_t = 1.08 (0.99, 1.16)^{17}$	$6.91e-04 (6.00e-04, 7.78e-04)^{17}$	MTBD
Australia	Prior to 27/03 Post 27/03	$R_t = 1.63 (1.45, 1.8)^{31}$ $R_t = 0.48 (0.27, 0.69)^{31}$	$1.1e-03$ $1.1e-03^{31}$	BCP+SC
Iceland	18/03–29/04	$R_t = 1.4 (1.2, 1.59)^{17}$	$5.75e-04 (4.96e-04, 6.47e-04)^{17}$	MTBD
Italy	22/02–04/04	$R_t = 2.25 (1.5, 3.1)^{133}$	$1.16e-03 (1.01e-03, 1.32e-03)^{134}$	BCP, BCP+SC
New Zealand	26/03–29/04	$R_t = 1.41 (1.07, 1.89)^{17}$	$6.09e-04 (5.16e-04, 7.03e-04)^{17}$	MTBD
Russia (Vreden hospital)	27/03–08/04 08/04–23/04	$R_t = 3.72 (2.48, 5.05)^{135}$ $R_t = 1.38 (0.48, 2.41)^{135}$	$9.43e-04 (8.46e-04, 1.04e-03)$ $9.43e-04 (8.46e-04, 1.04e-03)^{135}$	BCP+SC BCP+SC
Taiwan, Tâi-oân pún-tó	27/03–29/04	$R_t = 1.02 (0.825, 1.22)^{17}$	$8.00e-04 (6.89e-04, 9.17e-04)^{17}$	MTBD
Weifang, Shandong	25/01–10/02	$R_0 = 3.4 (2.1, 5.2)^{136}$	$1.30e-03 (0.98e-03, 1.7e-03)^{136}$	BCP+CFEM

A selection of studies providing phylodynamic estimates of both growth and clock rates are listed; other studies have published estimates of clock rates^{79,122} or reproduction numbers^{30,32,81,137,138}. Confidence intervals are provided where available (95% highest posterior density, (HPD)). Dates are dd/mm in 2020. BCP, Bayesian coalescent phylodynamic; CFEM, coalescent fitted epidemiological model; MTBD, multitype birth-death model; SC, structured coalescent.

Table 2 Pango lineages of interest or concern during the first year of the pandemic

<i>Pango lineage</i>	<i>Nextstrain clade</i>	<i>WHO and Public Health England denotations</i>	<i>Territory of first reporting</i>
A.23.1	NA	VUI-21FEB-01 ^{a*} (VUI-202102/01)	UK (associations with Uganda)
B.1.1.318	NA	VUI-21FEB-04 (VUI-202102/04)	UK (TBC)
B.1.1.7	20I/501Y.V1	WHO alpha (VOC-20DEC-01)	UK
B.1.1.7	NA	VOC-21FEB-02 ^{a*} (VOC-202102/02)	UK
B.1.324.1	NA	VUI-21MAR-01 ^{a*} (VUI-202103/01)	UK (links with travel from Antigua)
B.1.351	20H/501Y.V2	WHO beta (VOC-20DEC-02)	South Africa
B.1.525	20A/S:484K	VUI-21FEB-03 (VUI-202102/03) ^b	UK (associations with Angola)
B.1.617.2	NA	WHO delta	India
P.1	20J/501Y.V3	WHO gamma (VOC 202101/02)	Japan (in arrivals from Brazil)
P.2	NA	VUI-21JAN-01 (VUI-202101/01)	Brazil
P.3	NA	VUI-21MAR-02	Philippines (Central Visayas)

Although alternative denotations may, to varying degrees, correspond to Pango lineages, Pango lineage designations are based on clades, whereas alternative denotations may refer to constellations of substitutions rather than to phylogenetic ancestry. For example, VOC 202102/02 (B.1.1.7 with E484K) refers to several independent origins of variants that all carry the definitive mutations. The majority of alternative designations in the table, arise from the World Health Organization (WHO) or UK public health authorities^{74,139,140}.^a

^aRefers only to variants within the respective lineage that show E484K.

^bBriefly known as UK1188. NA, not applicable; TBC, to be confirmed.

Fig. 1: Phylodynamic approaches to the investigation of SARS-CoV-2 transmission

Relevant clinical and public health questions are defined (top row), phylodynamic and epidemiological data and models are then combined (middle row), and used in combined or joint analyses to provide actionable insight into virus transmission (bottom row). **a** | Phylogenetic approaches estimate the rate of international lineage introductions, and distinguish introductions from community transmission. **b** | Genome sequences and phylogenetics support outbreak analyses by identifying or refuting links between local cases, this can lead to identification of outbreak sources and drivers or assessment of nosocomial transmission. **c** | Phylodynamic techniques using epidemiological demographic models, such as the susceptible-exposed-infected-recovered (SEIR) model, allow us to compare transmission rates among lineages bearing different key genotypes (e.g. variants of concern (VOCs) and pre-existing lineages). **d** | Relative timing of variant and lineage emergence from the global (or regional) phylogeny, and scattering of case genomes among clades can distinguish persistent from repeat infections in some scenarios. Phylogenetics is also useful in studies of **lineage turnover** [G] and interactions within the host. **TMRCA** [G], time to the most

recent common ancestor. Panel header colours indicate related themes: blue, public health; green, epidemiological parameters; pink, clinical parameters.

544 **Fig. 2: The emergence of E484 bearing lineages from late 2020 to March 2021**

Spike amino acid mutations and deletions are shown as symbols on the pins marking the
546 approximate locations of first detection. The symbols include only those mutations that were
implicated in possible immune escape or as suspected drivers of lineage growth, and were shared by
548 two or more lineages. The locality of first detection may not be that of the lineage's origin;
however, the intercontinental spread of first detections is consistent with multiple independent
550 origins. The B.1.1.7 lineage coloured in red differs from the other B.1.1.7 viruses in that it bears
S494P rather than a substitution at E484. Lineage B.1.617 bears E484Q rather than E484K. Some
552 lineages (B.1.1.7 and A.23.1) also have members that lack E484K, and some virus genotypes may
have arisen multiple times (e.g. B.1.1.7 with E484K). The near coincidental first detection of the
554 same variants in genomes of phylogenetically distant lineages in countries worldwide, in early
2020, is a clear sign of convergent evolution and was a major factor leading to numerous studies
556 aimed at detecting any selective advantage of the variants of concern (VOCs), including the search
for vaccine escape phenotypes. Lineages and variants are based on the following publications:
558 A.23.1¹⁴¹; B.1.1.318, B.1.1.7+E484K, B.1.1.7+S494P, B.1.324.1⁷⁴; B.1.351^{15,71,74}; B.1.525⁷⁴;
B.1.617¹⁴²; P.1⁷³; P.2^{143,144}; P.3¹⁴⁵. Note: B.1.324.1 was not designated as a sublineage of B.1.324, and
560 reference here is to the variant described as B.1.324.1 in the Technical briefing Table 17 of REF⁷⁴.
Pin heights indicate time relative to detection of the first lineage, i.e. P.2 in Rio de Janeiro, October
562 13th, 2020 (not to scale, but ranked in time, with days since detection of P.2 marked on each pin).

564 **Fig. 3: Convergent evolution of SARS-CoV-2 spike protein.**

a | Phylogenies for the first year of the pandemic show the independent emergence of spike
566 ΔH69/V70, indicated in red, in genomes of the B.1.1.7 and B.1.258 lineages respectively — note,
the B.1.258 clade in red includes some branches without the deletion. Phylogeny from
568 Nextstrain^{146,147} (which used data from the Europe ncov GISAID dataset¹⁴⁸), visualised in Figtree.
For clarity, not all Pango lineages are shown. **b** | By the start of 2020 several commonly occurring
570 spike substitutions and deletions had been recognized as shared among lineages. The illustrated
substitutions are found in the exposed (i.e. outermost on the surface of the virion) subunit of spike,
572 termed S1, or in the spike N-terminal domain (NTD), and are those shared by variants of interest or
concern, excluding those shared sporadically or in minor sublineages. B.1.351 and P.1 share K417T
574 N and (in some B.1.351 sublineages) L18F, as well as two other recurrent substitutions, this is
indicated by the overlap of their extended shading. 'Mink' refers to the SARS-CoV-2 mink-human

576 sublineage, termed ‘Cluster 5’, which exhibited Δ H69/V70 and N501T (and other spike
substitutions)⁹¹; the second B.1.1.7 lineage (VOC-202102/02, the ellipse with broken-line border) is
578 a cluster of B.1.1.7 that also bears E484K¹⁴⁹. N501T is a homoplasy that emerged in mink and may
have transferred to humans; it is relatively uncommon, as it was found in only five mink in the
580 original mink farm epidemic in Denmark. Nevertheless, N501T appeared to have emerged
independently four times, and has been detected in ten human cases¹⁵⁰. L18F is an NTD substitution
582 found in some B.1.351, and several of its sublineages, and increasing in frequency in B.1.1.7⁶⁷. As
in Fig. 2 we see that the same substitutions appear in multiple lineages, implying that they arose
584 independently at different times and places. Here, we also see that not only are individual
substitutions shared, but constellations of several changes also appear to co-occur in more than one
586 lineage; this suggests epistatic interactions, with perhaps compensatory changes following immune
escape variants.

588

Fig. 4: Effects of within-host evolution and dynamics on epidemiological observations.

Phylogenetic and phylodynamic approaches help detect and understand complex infections, measure within-patient lineage turnover, and explore how host-induced mutation affects outbreak investigations. **a** | Co-infections may confound transmissibility and aetiological studies, but they can be detected using phylogenetics. Specifically, co-infections are identified when viral genomes sequenced from multiple isolates from the same patient are not monophyletic. **b** | Lineage turnover can occur if within-host lineages share a recent common ancestor and arise from evolution within the host itself. Lineage turnover may complicate patient treatment, as a lineage with lesser susceptibility to host immune responses may give way to a more transmissible lineage after apparently successful completion of a course of therapy. Nevertheless, phylogenetic features such as longitudinal samples falling into different sister lineages, and relative branch lengths can help detect and account for lineage turnover. **c** | The antiviral activities of host APOBEC cytidine deaminases, which promote C→U hypermutation, adenosine deaminases that act on RNA (ADARs), and similar host systems, can lead to biases such as C→U homoplasies (convergent evolution) in the case of APOBECs, and changes in virus genome CpG content as a response^{151–153}. Phylogenetics can highlight such convergent changes, which will be seen arising in lineages that are not closely related, and phylogenetic and phylodynamic approaches can be adjusted to account for the elevated rate of particular transitions. **d** | Co-infections and superinfections can complicate attempts to trace transmission chains, either through lineage turnover or sampling bias (e.g. differential PCR amplification or through effects of organotropy). The result can be failure to connect two related transmission chains. A superinfected individual could also cryptically contribute to more than one heterochronous outbreak. The schema shows potential transmission events within households, or

similar units (e.g. work places), in a simplified transmission scenario. The broken lines indicate transmission events among households. Circles represent individuals, with empty circles indicating infection chains involving lineage 1, and filled circles those involving lineage 2. The red asterisk indicates a co-infected individual carrying both lineages. The phylogeny shows that the true relationship between individuals X and Y may be unclear if lineage 1 dominates the co-infection at the time of sampling.

590

Box 1. Phylogenetic terminology and concepts

[bH1] Phylogenetics

Phylogenetics provides a method for the generation of hypotheses about ancestor-descendant relationships using character-state data. The resulting phylogeny attempts to explain the observed character states in the sequences that we have sampled, as having evolved from a single common ancestor in the past, via a sequence of usually unobserved (unsampled or extinct) hypothesized intermediate ancestors represented by internal nodes or branch points on a bifurcating tree (see the figure). Phylogenetic methods typically search for the solution with the minimum number of evolutionary steps (parsimony) or that maximizing the likelihood of the data given the tree. A third alternative is a Bayesian approach, which applies Bayes theorem to estimate a probability distribution for population parameters of interest. The ability to incorporate prior information (priors) for the events (e.g. a prior distribution for outbreak onset time) gives the approach an advantage over maximum likelihood estimation.

Phylogenies have also formed the basis of a system for the identification, definition and monitoring of outbreak clusters and variants of concern (VOCs). Although nomenclatures such as that currently adopted by the World Health Organization (WHO), assign names to definitive constellations of substitutions that commonly occur together (e.g. VOC delta), most other current nomenclatures are lineage based (e.g. Pango and Nextstrain). In the case of the Pango nomenclature, lineages correspond either loosely or exactly to clades estimated on a reference phylogeny. A clade is a monophyletic subtree on a phylogeny, such subtrees include all descendants of their most recent common ancestor represented by the node joining them to the global phylogeny, and no others (see lineage A in the figure, which forms a clade). Nevertheless, Pango lineages can include any fairly cohesive, and exclusive (or nearly so) clustering of sequences on the global SARS-CoV-2 phylogeny, particularly where that cluster associates with an outbreak, epidemiologically significant

phenotype (e.g. greater transmissibility) or any noteworthy characteristic, whether proven or awaiting investigation.

[bH1] Phylodynamics

Phylodynamics focuses on the estimation of population dynamic parameters from genetic sequences and molecular phylogenies, such as epidemic growth rates, generation times, and reproductive numbers. Phylodynamic analysis relies on other statistical models such as nucleotide substitution models and molecular clock models. Phylodynamics has been used to incorporate epidemiological data in phylogenetic studies of the pandemic. Such models allow estimation of demographic or epidemic parameters over time; these often include changes in relative population size (including reproductive number and growth rate), and selection coefficients. Phylodynamics can help date the first cases in a region, and can provide public health officials with an estimate of the lag between importation and first case detection by estimating the time to the most recent common ancestor (TMRCA) of a clade. Phylogeography has been used during the pandemic to estimate rates of virus (lineage) movement among regions, and may be considered as a form of phylodynamics that uses phylogenetic methods to understand the spatial dissemination of lineages.

The coalescent model is central to a large class of phylodynamic methods. The coalescent considers mutation-drift (i.e. evolution without selection) backwards in time, with pairs of lineages coalescing rather than diverging. The model can be visualized as a genealogy, is computationally efficient, and deviations from the expected distribution of coalescence intervals (in time) can be used to infer processes such as selection, and migration. Phylodynamic applications of the coalescent often involve the use of ‘skyline plots’ or related methods to estimate historical changes in population size (e.g. virus demographics); these are estimates of effective population size (N_e) change through time. Various skyline methods exist, and these generally differ in the parameterization and smoothing of the population size changes. The coalescent model can also be modified to allow for expected population structure (structured coalescent). Similarly, epidemiological models, such as the susceptible-exposed-infected-recovered (SEIR) model, are incorporated into the phylodynamic framework as compartmental models in order to model disease transmission and prevalence. Compartmental models involve the partitioning of the individuals (e.g. hosts) within a population into mutually exclusive groups according to their properties, with their progression between the groups permitted according to the rules underlying the model. Due to computational advances, more mathematically complex birth-death (BD) models, including multitype BD models (see Table 1), are also being used in phylodynamic analysis⁵ in addition to coalescent approaches. These BD

models explicitly represent lineage diversification events and can estimate growth rates for distinct ‘types’ or sub-populations.

Glossary

592

Discrete asymmetric phylogeographic model

A phylogeographic model (Box 1) in which the transition rates between pairs of discrete location states can differ depending on the direction of movement between them. This asymmetry is an important property for some models in spatial epidemiology (e.g. source/sink models).

R_t

The general reproduction number (R_t or R_e) applicable to any stage in an epidemic or pandemic (it is R_0 at $t=0$). R_t is affected by public health interventions, host behaviour, and the accumulation of immune individuals in the population.

WA1

USA-WA1-2020, the curated genome sequence of SARS-CoV-2 found in a sample taken from the first officially reported case of COVID-19 in the United States.

594 **Nosocomial transmission**

Transmission chains initiated in, or driven by, activities undertaken in a hospital setting, particularly those related to patient treatment and care.

Substitutions

Here substitutions refer to mutations that have persisted through viral generations (i.e. they are transmissible), reaching sufficient population frequency so as to appear in consensus genomes, and therefore representing a polymorphism of non-trivial frequency or a fixation in the population.

Attack rate

The proportion of a potentially exposed susceptible population subsequently judged positive for infection, according to some approved criteria or test, over a specified time period.

Evolutionary convergence

(Also known as convergent evolution).

The independent emergence of the same character state (e.g. a nucleotide substitution such as N501Y) in distinct phylogenetic lineages (e.g. in different Pango lineages); this is a form of homoplasy.

Convalescent plasma

Passive transfer of antibodies in a therapeutic manner, from previously infected but recovered patients, through transfusion of plasma from donated blood.

Pseudotyping experiments

Experiments using a virus with a viral envelope from another virus, for example a SARS-CoV-2 core within a lentivirus envelope. This allows the convenient use of cell lines of a cell type that SARS-CoV-2 could not naturally infect. Safety is also enhanced as the construct lacks the genes coding for a functional autologous envelope.

Founder effects

Patterns in gene (variant) frequencies resulting from chance colonization events rather than selection; these events may be timed (e.g. coincident with a super-spreading event) or located (e.g. in a naive population) so as to give the impression that one lineage has a growth advantage over others.

R_0

The basic reproduction number (R_0) represents the average number of new infections arising as a result of contact with an infected individual in a naive population (this usually applies at the start of an epidemic).

Effective population size

(N_e). A population genetic parameter that affects the amount of genetic diversity in a population and is related to the true population size. Depending on circumstances, the relationship can be linear or more complex.

Superinfection

A second infection, or subsequent infections of the same or a different organism, established in a host already infected at some earlier time. This is in contrast to co-infection, where both infections are acquired at the same time.

Quasispecies

A population of genetically distinct viruses coexisting within one individual host; these may exhibit turnover (see Fig. 4b).

Lineage turnover

In the context of a pathogen evolving within an infected host, lineage turnover refers to the balance between new lineage divergence and growth on the one hand and decline and extinction on the other; this balance being affected by changes in selection pressures, which can accelerate turnover so that a previously dominant or sole lineage is replaced by a new or previously minor one.

TMRC

(Time to the most recent common ancestor). The time of the splitting of a clade into two sub-clades, when the sub-clades shared a common ancestor, or equivalently the date on the root of a clade.

Table of contents blurb

In this Review, the authors describe how phylogenetic and phylodynamic methods provide insight into viral evolution, focusing on the SARS-CoV-2 pandemic. The approaches reveal routes and timings of transmission events, and they can assess the effectiveness of various intervention measures aimed at controlling the virus.

References

1. Eickmann, M. *et al.* Phylogeny of the SARS Coronavirus. *Science* **302**, 1504–1505 (2003).
2. Arias, A. *et al.* Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, (2016).
3. Dudas, G. *et al.* Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315 (2017).
4. Grubaugh, N. D. *et al.* Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature* **546**, 401–405 (2017).
5. Ingle, D. J., Howden, B. P. & Duchene, S. Development of Phylodynamic Methods for Bacterial Pathogens. *Trends Microbiol.* **29**, 788–797 (2021).
6. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).

7. Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399 (2000).
8. Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307–1320 (2002).
9. Dellicour, S. *et al.* Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nat. Commun.* **11**, 5620 (2020).
10. Arévalo, S. J. *et al.* Analysis of the Dynamics and Distribution of SARS-CoV-2 Mutations and its Possible Structural and Functional Implications. *bioRxiv* 2020.11.13.381228 (2020) doi:10.1101/2020.11.13.381228.
11. Yang, J. *et al.* Uncovering two phases of early intercontinental COVID-19 transmission dynamics. *J. Travel Med.* **27**, taaa200 (2020).
12. Nadeau, S. A., Vaughan, T. G., Scire, J., Huisman, J. S. & Stadler, T. The origin and early spread of SARS-CoV-2 in Europe. *Proc. Natl. Acad. Sci.* **118**, e2012008118 (2021).
13. Fountain-Jones, N. M. *et al.* Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evol.* **6**, veaa082 (2020).
14. Lemieux, J. E. *et al.* Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).
15. Tegally, H. *et al.* Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature* **592**, 438–443 (2021).
16. Di Giallonardo, F. *et al.* Genomic Epidemiology of the First Wave of SARS-CoV-2 in Italy. *Viruses* **12**, (2020).
17. Douglas, J. *et al.* Phylodynamics reveals the role of human travel and contact tracing in controlling the first wave of COVID-19 in four island nations. *Virus Evol.* **7**, veab052 (2021).
18. Plessis, L. du *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
19. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
20. Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996.e5 (2020).

21. da Silva Filipe, A. *et al.* Genomic epidemiology reveals multiple introductions of SARS-CoV-2 from mainland Europe into Scotland. *Nat. Microbiol.* **6**, 112–122 (2021).
22. Claro, I. M. *et al.* Local Transmission of SARS-CoV-2 Lineage B.1.1.7, Brazil, December 2020. *Emerg. Infect. Dis. J. - CDC* **27**, eid2703.210038 (2021).
23. Washington, N. L. *et al.* Genomic epidemiology identifies emergence and rapid transmission of SARS-CoV-2 B.1.1.7 in the United States. *medRxiv* 2021.02.06.21251159 (2021) doi:10.1101/2021.02.06.21251159.
24. Lemey, P. *et al.* Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* 1–8 (2021) doi:10.1038/s41586-021-03754-2.
25. Lemey, P. *et al.* Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
26. De Maio, N., Wu, C.-H., O'Reilly, K. M. & Wilson, D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet.* **11**, e1005421 (2015).
27. Faria, N. R. *et al.* Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *J. Gen. Virol.* **93**, 889–899 (2012).
28. Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997-1003.e9 (2020).
29. Miller, D. *et al.* Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. Commun.* **11**, 5518 (2020).
30. Ragonnet-Cronin, M. *et al.* Genetic evidence for the association between COVID-19 epidemic severity and timing of non-pharmaceutical interventions. *Nat. Commun.* **12**, 2188 (2021).
31. Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *Nat. Commun.* **11**, 4376 (2020).
32. Geoghegan, J. L. *et al.* Genomic epidemiology reveals transmission patterns and dynamics of SARS-CoV-2 in Aotearoa New Zealand. *Nat. Commun.* **11**, 6351 (2020).
33. Brito, A. F. *et al.* Global disparities in SARS-CoV-2 genomic surveillance. *MedRxiv Prepr. Serv. Health Sci.* 2021.08.21.21262393 (2021) doi:10.1101/2021.08.21.21262393.
34. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).
35. Douglas, J. *et al.* Real-Time Genomics for Tracking Severe Acute Respiratory Syndrome

- Coronavirus 2 Border Incursions after Virus Elimination, New Zealand. *Emerg. Infect. Dis.* **27**, 2361–2368 (2021).
36. Voeten, H. A. C. M. *et al.* Unravelling the modes of transmission of SARS-CoV-2 during a nursing home outbreak: looking beyond the church super-spread event. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **73(Suppl 2)**, S163–S169 (2021).
37. Connor, T. *et al.* SARS-Cov-2 Genomic Insights with cover statement. *The Welsh Government: Genomic analysis of Covid-19 lineages in Wales* <https://gov.wales/sars-cov-2-genomic-insights-cover-statement-html> (2020).
38. Lane, C. R. *et al.* Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study. *Lancet Public Health* **6**, e547–e556 (2021).
39. Volz, E. *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* **593**, 266–269 (2021).
40. Giandhari, J. *et al.* Early transmission of SARS-CoV-2 in South Africa: An epidemiological and phylogenetic report. *Int. J. Infect. Dis.* **103**, 234–241 (2021).
41. Sikkema, R. S. *et al.* COVID-19 in health-care workers in three hospitals in the south of the Netherlands: a cross-sectional study. *Lancet Infect. Dis.* **20**, 1273–1280 (2020).
42. Meredith, L. W. *et al.* Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* **20**, 1263–1271 (2020).
43. Ladhani, S. N. *et al.* Increased risk of SARS-CoV-2 infection in staff working across different care homes: enhanced CoVID-19 outbreak investigations in London care Homes. *J. Infect.* **81**, 621–624 (2020).
44. Klompas, M. *et al.* A SARS-CoV-2 Cluster in an Acute Care Hospital. *Ann. Intern. Med.* M20-7567 (2021) doi:10.7326/M20-7567.
45. Park, K. *et al.* Epidemiologic Linkage of COVID-19 Outbreaks at Two University-affiliated Hospitals in the Seoul Metropolitan Area in March 2020. *J. Korean Med. Sci.* **36**, e38 (2021).
46. Karmarkar, E. *et al.* Timely Intervention and Control of a Novel Coronavirus (COVID-19) Outbreak at a Large Skilled Nursing Facility - San Francisco, California, 2020. *Infect. Control Hosp. Epidemiol.* **1–8**, doi:10.1017/ice.2020.1375 (2020).

47. Olmos, C. *et al.* SARS-CoV-2 infection in asymptomatic healthcare workers at a clinic in Chile. *PLOS ONE* **16**, e0245913 (2021).
48. Quéromès, G. *et al.* Characterization of SARS-CoV-2 ORF6 deletion variants detected in a nosocomial cluster during routine genomic surveillance, Lyon, France. *Emerg. Microbes Infect.* **10**, 167–177 (2021).
49. Taylor, J. *et al.* Serial Testing for SARS-CoV-2 and Virus Whole Genome Sequencing Inform Infection Risk at Two Skilled Nursing Facilities with COVID-19 Outbreaks — Minnesota, April–June 2020. *Morb. Mortal. Wkly. Rep.* **69**, 1288–1295 (2020).
50. Hamner, L. High SARS-CoV-2 Attack Rate Following Exposure at a Choir Practice — Skagit County, Washington, March 2020. *MMWR Morb. Mortal. Wkly. Rep.* **69**, (2020).
51. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *The Lancet* **395**, e47 (2020).
52. Weissberg, D. *et al.* Does respiratory co-infection facilitate dispersal of SARS-CoV-2? investigation of a super-spreading event in an open-space office. *Antimicrob. Resist. Infect. Control* **9**, 191 (2020).
53. Zeller, M. *et al.* Emergence of an early SARS-CoV-2 epidemic in the United States. *Cell* doi:10.1016/j.cell.2021.07.030 (2021) doi:10.1016/j.cell.2021.07.030.
54. Sekizuka, T. *et al.* A Genome Epidemiological Study of SARS-CoV-2 Introduction into Japan. *mSphere* **5**, e00786-20 (2020).
55. Muller, N. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 Outbreak Related to a Nightclub, Germany, 2020. *Emerg. Infect. Dis.* **27**, 645–648 (2020).
56. Choi, E. M. *et al.* In-Flight Transmission of SARS-CoV-2. *Emerg. Infect. Dis. J. - CDC* **26**, eid2611.203254 (2020).
57. Swadi, T. *et al.* Genomic Evidence of In-Flight Transmission of SARS-CoV-2 Despite Predeparture Testing. *Emerg. Infect. Dis. J. - CDC* **27**, 687–693 (2021).
58. Sekizuka, T. *et al.* Haplotype networks of SARS-CoV-2 infections in the Diamond Princess cruise ship outbreak. *Proc. Natl. Acad. Sci.* **117**, 20198–20201 (2020).
59. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582–587 (2020).
60. Böhmer, M. M. *et al.* Investigation of a COVID-19 outbreak in Germany resulting from a

- single travel-associated primary case: a case series. *Lancet Infect. Dis.* **20**, 920–928 (2020).
61. To, K. K.-W. *et al.* COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **25**, ciaa1275 (2020).
 62. Tillett, R. L. *et al.* Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect. Dis.* **21**, 52–58 (2021).
 63. ECDC. Risk related to the spread of new SARS-CoV-2 variants of concern in the EU/EEA, first update–21January 2021. (2021).
 64. Rambaut, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. (2020).
 65. Gräf, T. *et al.* Identification of SARS-CoV-2 P.1-related lineages in Brazil provides new insights about the mechanisms of emergence of Variants of Concern - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. (2021).
 66. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
 67. Gupta, R. K. Will SARS-CoV-2 variants of concern affect the promise of vaccines? *Nat. Rev. Immunol.* **21**, 340–341 (2021).
 68. Lennerstrand, J., Svensson, L. & Åke, L. Hur har omikron uppstått och varför sprider den sig så snabbt? *Läkartidningen* vol. 119 21242 (2022).
 69. Andreano, E. *et al.* SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proc. Natl. Acad. Sci.* **118**, e2103154118 (2021).
 70. Hu, J. *et al.* Emerging SARS-CoV-2 variants reduce neutralization sensitivity to convalescent sera and monoclonal antibodies. *Cell. Mol. Immunol.* **18**, 1061–1063 (2021).
 71. Wang, P. *et al.* Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).
 72. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021).
 73. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. *Virological.org* <https://virological.org/t/genomic-characterisation-of-an>

- emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586 (2021).
74. Variant Technical Group. SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 7. (2021).
 75. Naveca, F. G. *et al.* COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. *Nat. Med.* **27**, 1230–1238 (2021).
 76. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication *ex vivo* and transmission *in vivo*. *Science* **370**, 1464–1468 (2020).
 77. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
 78. Michaud, W. A., Boland, G. M. & Rabi, S. A. The SARS-CoV-2 Spike mutation D614G increases entry fitness across a range of ACE2 levels, directly outcompetes the wild type, and is preferentially incorporated into trimers. *bioRxiv* 2020.08.25.267500 (2020)
doi:10.1101/2020.08.25.267500.
 79. Díez-Fuertes, F. *et al.* A Founder Effect Led Early SARS-CoV-2 Transmission in Spain. *J. Virol.* **95**, JVI.01583-20 (2021).
 80. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
 81. Volz, E. *et al.* Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64-75.e11 (2020).
 82. Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. Making Sense of Mutation: What D614G Means for the COVID-19 Pandemic Remains Unclear. *Cell* **182**, 794–795 (2020).
 83. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10**, e66857 (2021).
 84. Pedro, N. *et al.* Dynamics of a Dual SARS-CoV-2 Lineage Co-Infection on a Prolonged Viral Shedding COVID-19 Case: Insights into Clinical Severity and Disease Duration. *Microorganisms* **9**, 300 (2021).
 85. Forni, D., Cagliani, R., Pontremoli, C., Clerici, M. & Sironi, M. The substitution spectra of coronavirus genomes. *Brief. Bioinform.* (2021) doi:10.1093/bib/bbab382.
 86. Gutierrez, B., Escalera-Zamudio, M. & Pybus, O. G. Parallel molecular evolution and adaptation in viruses. *Curr. Opin. Virol.* **34**, 90–96 (2019).

87. Day, T., Gandon, S., Lion, S. & Otto, S. P. On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* **30**, R849–R857 (2020).
88. Zhou, D. *et al.* Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* **184**, 2348-2361.e6 (2021).
89. Barton, M. I. *et al.* Effects of common mutations in the SARS-CoV-2 Spike RBD and its ligand the human ACE2 receptor on binding affinity and kinetics. *eLife* **10**, e70658 (2021).
90. Starr, T. N. *et al.* Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295-1310.e20 (2020).
91. Meng, B. *et al.* Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep.* **35**, 109292 (2021).
92. Giorgi, E. E. *et al.* Recombination and low-diversity confound homoplasy-based methods to detect the effect of SARS-CoV-2 mutations on viral transmissibility. *bioRxiv* 2021.01.29.428535 (2021) doi:10.1101/2021.01.29.428535.
93. Richard, D., Owen, C. J., Dorp, L. van & Balloux, F. No detectable signal for ongoing genetic recombination in SARS-CoV-2. *bioRxiv* 2020.12.15.422866 (2020) doi:10.1101/2020.12.15.422866.
94. Jackson, B. *et al.* Recombinant SARS-CoV-2 genomes involving lineage B.1.1.7 in the UK. *Virological* <https://virological.org/t/recombinant-sars-cov-2-genomes-involving-lineage-b-1-1-7-in-the-uk/658> (2021).
95. Gutierrez, B. *et al.* Emergence and widespread circulation of a recombinant SARS-CoV-2 lineage in North America. *medRxiv* 2021.11.19.21266601 (2021) doi:10.1101/2021.11.19.21266601.
96. Alm, E. *et al.* Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. *Eurosurveillance* **25**, 2001410 (2020).
97. Furuse, Y. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. *Int. J. Infect. Dis.* **103**, 305–307 (2021).
98. Flaxman, S. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584**, 257–261 (2020).
99. Hill, V., Ruis, C., Bajaj, S., Pybus, O. G. & Kraemer, M. U. G. Progress and challenges in virus genomic epidemiology. *Trends Parasitol.* **37**, 1038–1049 (2021).

100. Featherstone, L. A., Giallonardo, F. D., Holmes, E. C., Vaughan, T. G. & Duchêne, S. Infectious disease phylodynamics with occurrence data. *Methods Ecol. Evol.* **12**, 1498–1507 (2021).
101. Günther, T. *et al.* SARS-CoV-2 outbreak investigation in a German meat processing plant. *EMBO Mol. Med.* **12**, e13296 (2020).
102. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and North America. *Science* **370**, 564–570 (2020).
103. Dellicour, S. *et al.* Dispersal dynamics of SARS-CoV-2 lineages during the first epidemic wave in New York City. *PLOS Pathog.* **17**, e1009571 (2021).
104. Kalia, K., Saberwal, G. & Sharma, G. The lag in SARS-CoV-2 genome submissions to GISAID. *Nat. Biotechnol.* **39**, 1058–1060 (2021).
105. Wohl, S., Giles, J. R. & Lessler, J. Sample size calculation for phylogenetic case linkage. *PLOS Comput. Biol.* **17**, e1009182 (2021).
106. Larsen, D. A. & Wigginton, K. R. Tracking COVID-19 with wastewater. *Nat. Biotechnol.* **38**, 1151–1153 (2020).
107. Smyth, D. S. *et al.* Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.* **13**, 635 (2022).
108. Nemudryi, A. *et al.* Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in Municipal Wastewater. *Cell Rep. Med.* **1**, 100098 (2020).
109. Parag, K. V., du Plessis, L. & Pybus, O. G. Jointly Inferring the Dynamics of Population Size and Sampling Intensity from Molecular Sequences. *Mol. Biol. Evol.* **37**, 2414–2429 (2020).
110. Stadler, T., Kühnert, D., Bonhoeffer, S. & Drummond, A. J. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci.* **110**, 228–233 (2013).
111. Cappello, L. & Palacios, J. A. Adaptive preferential sampling in phylodynamics. *ArXiv200902307 Q-Bio Stat* (2020).
112. Andréoletti, J. *et al.* A skyline birth-death process for inferring the population size from a reconstructed tree with occurrences. *bioRxiv* 2020.10.27.356758 (2020)
doi:10.1101/2020.10.27.356758.
113. Vaughan, T. G. *et al.* Estimating Epidemic Incidence and Prevalence from Genomic Data.

- Mol. Biol. Evol.* **36**, 1804–1816 (2019).
114. WHO. *Genomic sequencing of SARS-CoV-2: A guide to implementation for maximum impact on public health*. (World Health Organization, 2021).
115. Vaughan, T. G., Sciré, J., Nadeau, S. A. & Stadler, T. Estimates of outbreak-specific SARS-CoV-2 epidemiological parameters from genomic data. *medRxiv* 2020.09.12.20193284 (2020) doi:10.1101/2020.09.12.20193284.
116. Hufsky, F. *et al.* Computational strategies to combat COVID-19: useful tools to accelerate SARS-CoV-2 and coronavirus research. *Brief. Bioinform.* **22**, 642–663 (2021).
117. Gattepaille, L. M., Jakobsson, M. & Blum, M. G. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity* **110**, 409–419 (2013).
118. Smith, M. R. *et al.* Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020. *Nat. Commun.* **12**, 6009 (2021).
119. Dearlove, B. *et al.* A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23652–23662 (2020).
120. Pereson, M. J. *et al.* Phylogenetic analysis of SARS-CoV-2 in the first few months since its emergence. *J. Med. Virol.* **93**, 1722–1731 (2021).
121. Villabona-Arenas, C. J., Hanage, W. P. & Tully, D. C. Phylogenetic interpretation during outbreaks requires caution. *Nat. Microbiol.* **5**, 876–877 (2020).
122. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, veaa061 (2020).
123. Morel, B. *et al.* Phylogenetic analysis of SARS-CoV-2 data is difficult. *Mol. Biol. Evol.* **38**, 1777–1791 (2021).
124. Turakhia, Y. *et al.* Stability of SARS-CoV-2 phylogenies. *PLOS Genet.* **16**, e1009175 (2020).
125. Utro, F., Levovitz, C., Rhrissorrakrai, K. & Parida, L. A Common Methodological Phylogenomics Framework for intra-patient heteroplasmies to infer SARS-CoV-2 sublineages and tumor clones. *bioRxiv* 2020.10.14.339986 (2020) doi:10.1101/2020.10.14.339986.
126. Ramazzotti, D. *et al.* VERSO: A comprehensive framework for the inference of robust phylogenies and the quantification of intra-host genomic diversity of viral samples. *Patterns* **2**, 100212 (2021).
127. Charre, C. *et al.* Evaluation of NGS-based approaches for SARS-CoV-2 whole genome

- characterisation. *Virus Evol.* **6**, veaa075 (2020).
128. Nasir, J. A. *et al.* A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture. *Viruses* **12**, 895 (2020).
129. Noorden, R. V. Scientists call for fully open sharing of coronavirus genome data. *Nature* **590**, 195–196 (2021).
130. Callaway, E. ‘A bloody mess’: Confusion reigns over naming of new COVID variants. *Nature* **589**, 339–339 (2021).
131. Ferguson, C. Don’t let “delta plus” confuse you. The strain hasn’t learned any new tricks. *MIT Technology Review* <https://www.technologyreview.com/2021/08/13/1031844/delta-plus-covid-variant-naming/> (2021).
132. Hodcroft, E. B. *et al.* Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).
133. Lai, A. *et al.* Molecular Tracing of SARS-CoV-2 in Italy in the First Three Months of the Epidemic. *Viruses* **12**, 798 (2020).
134. Alteri, C. *et al.* Genomic epidemiology of SARS-CoV-2 reveals multiple lineages and early spread of SARS-CoV-2 infections in Lombardy, Italy. *Nat. Commun.* **12**, 434 (2021).
135. Komissarov, A. B. *et al.* Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. *Nat. Commun.* **12**, 649 (2021).
136. Geidelberg, L. *et al.* Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evol.* **7**, veaa102 (2021).
137. Kepler, L., Hamins-Puertolas, M. & Rasmussen, D. A. Decomposing the sources of SARS-CoV-2 fitness variation in the United States. *bioRxiv* 2020.12.14.422739 (2020)
doi:10.1101/2020.12.14.422739.
138. Moreno, G. K. *et al.* Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat. Commun.* **11**, 5558 (2020).
139. Public Health England. Variants: distribution of cases data. *Gov.uk* <https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data> (2021).
140. O’Toole, A. *et al.* PANGO lineages. *Cov-lineages* <https://cov-lineages.org/index.html> (2021).
141. Bugembe, D. L. *et al.* Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1)

- with altered spike protein in Uganda. *Nat. Microbiol.* **6**, 1094–1101 (2021).
142. Cherian, S. *et al.* SARS-CoV-2 Spike Mutations, L452R, T478K, E484Q and P681R, in the Second Wave of COVID-19 in Maharashtra, India. *Microorganisms* **9**, 1542 (2021).
143. Nonaka, C. K. V. *et al.* Genomic Evidence of SARS-CoV-2 Reinfection Involving E484K Spike Mutation, Brazil. *Emerg. Infect. Dis.* **27**, 1522–1524 (2021).
144. Voloch, C. M. *et al.* Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* **95**, JVI.00119-21 (2021).
145. Public Health England. What do we know about the new COVID-19 variants? - Public health matters. *The official blog of Public Health England*
<https://publichealthmatters.blog.gov.uk/2021/02/05/what-do-we-know-about-the-new-covid-19-variants/> (2021).
146. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
147. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
148. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
149. Ho, D. *et al.* Increased Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7 to Antibody Neutralization. *Res. Sq.* **3**, rs-155394/v1 (2021).
150. van Dorp, L. *et al.* Recurrent mutations in SARS-CoV-2 genomes isolated from mink point to rapid host-adaptation. *bioRxiv* 2020.11.16.384743 (2020) doi:10.1101/2020.11.16.384743.
151. Simmonds, P. Rampant C → U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* **5**, e00408-20 (2020).
152. Rice, A. M. *et al.* Evidence for Strong Mutation Bias toward, and Selection against, U Content in SARS-CoV-2: Implications for Vaccine Design. *Mol. Biol. Evol.* **38**, 67–83 (2021).
153. Mourier, T. *et al.* Host-directed editing of the SARS-CoV-2 genome. *Biochem. Biophys. Res. Commun.* **538**, 35–39 (2021).

fig. 1

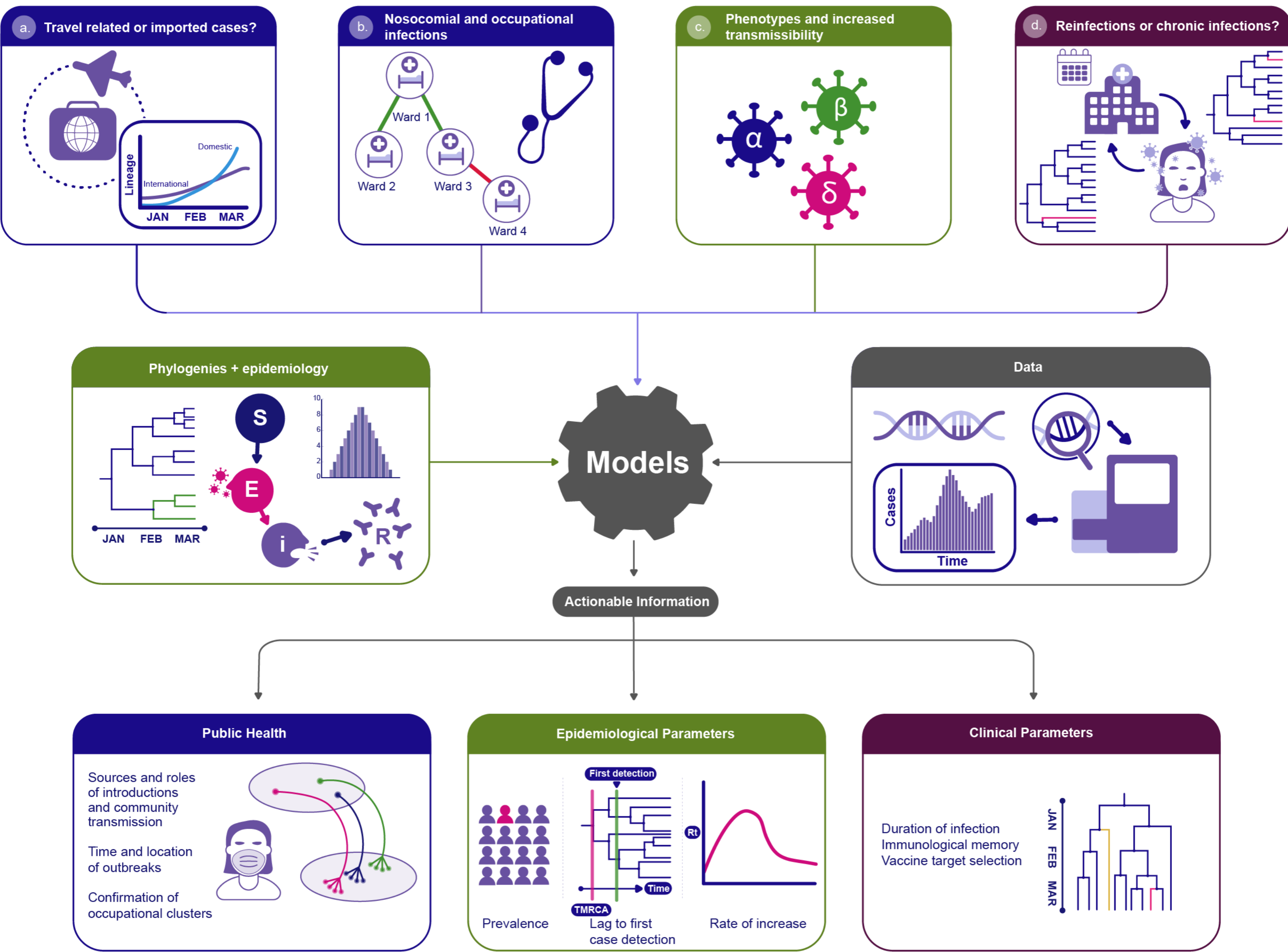


fig. 2

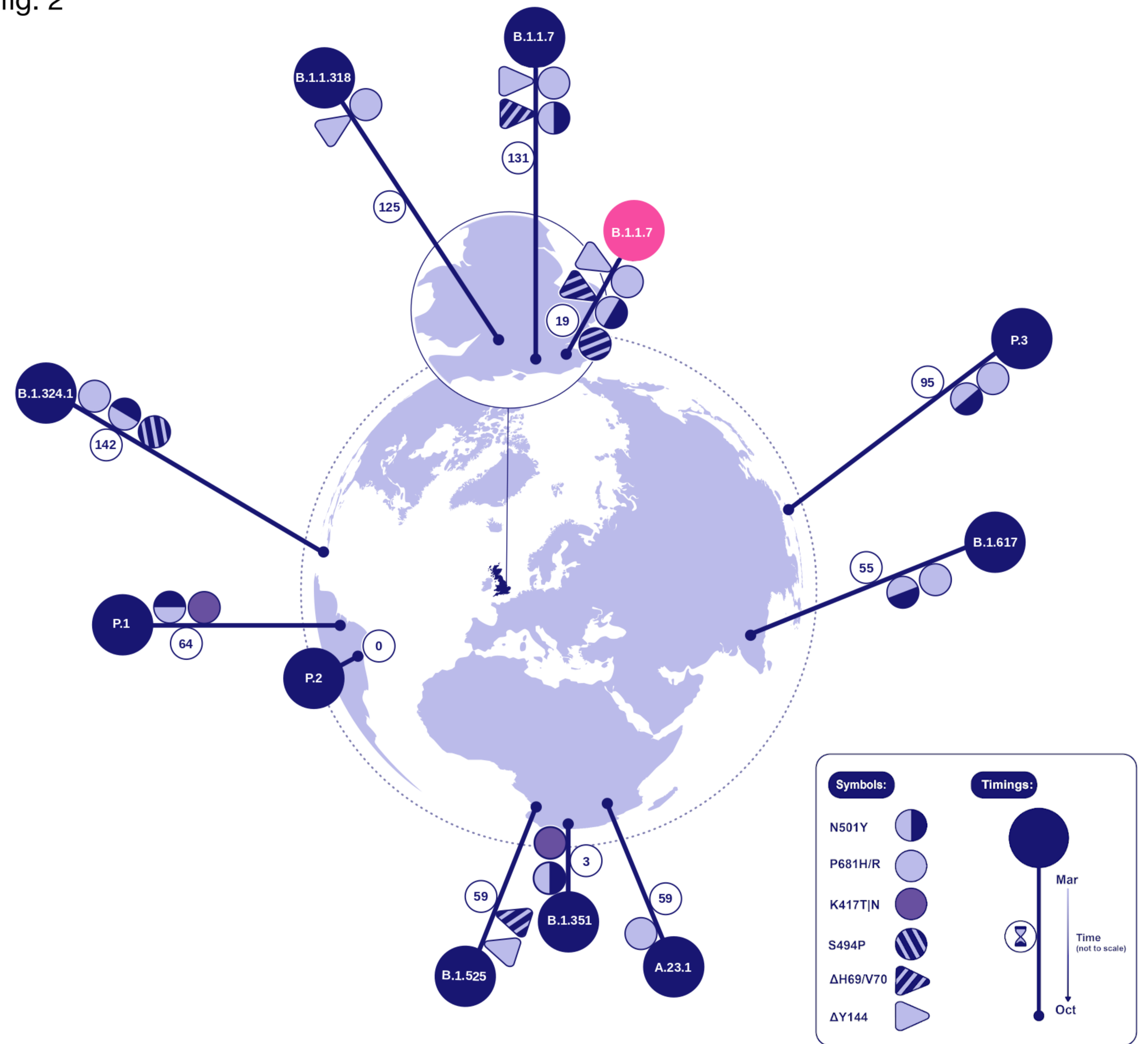


Fig. 3a

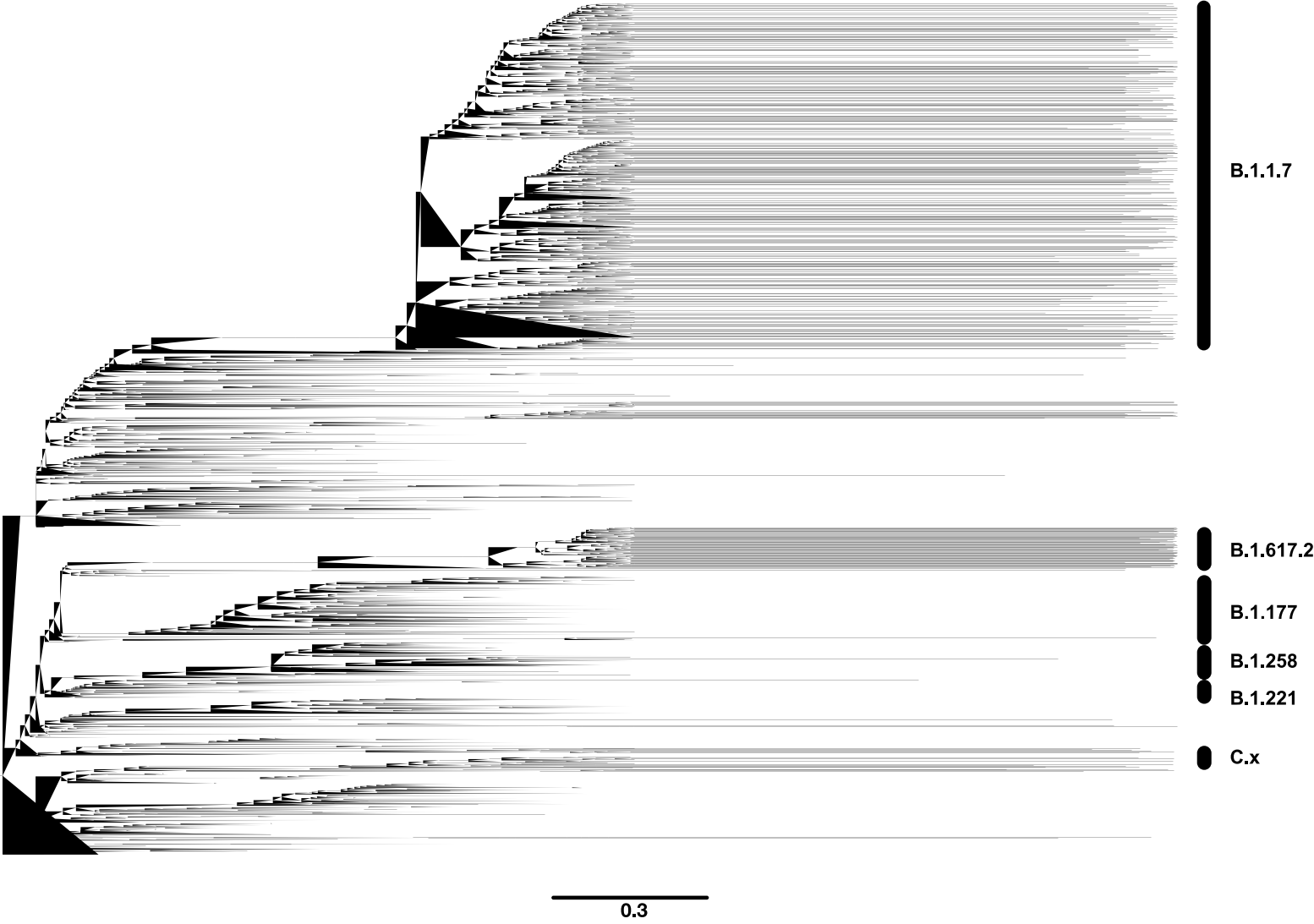


Fig. 3b

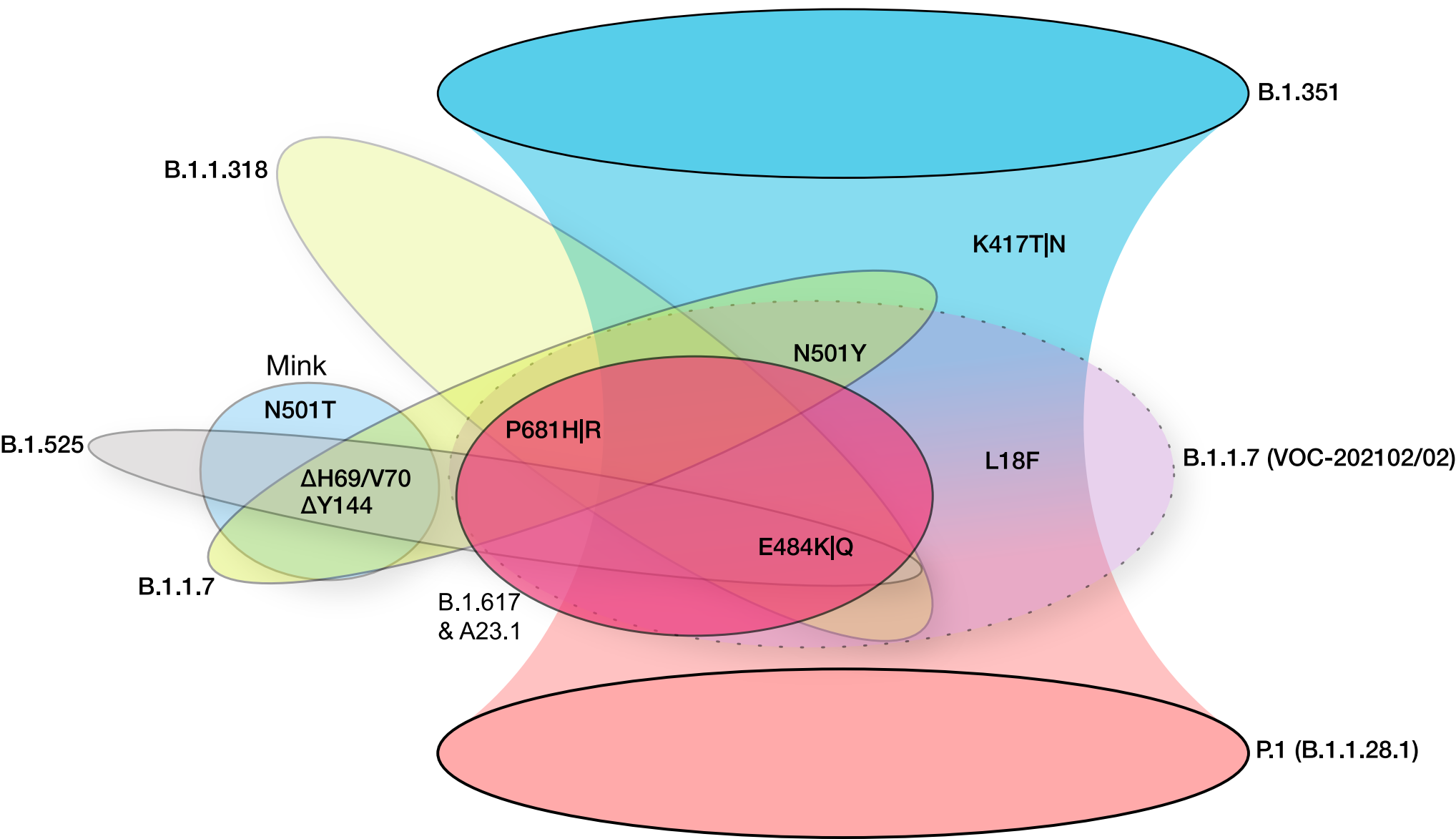
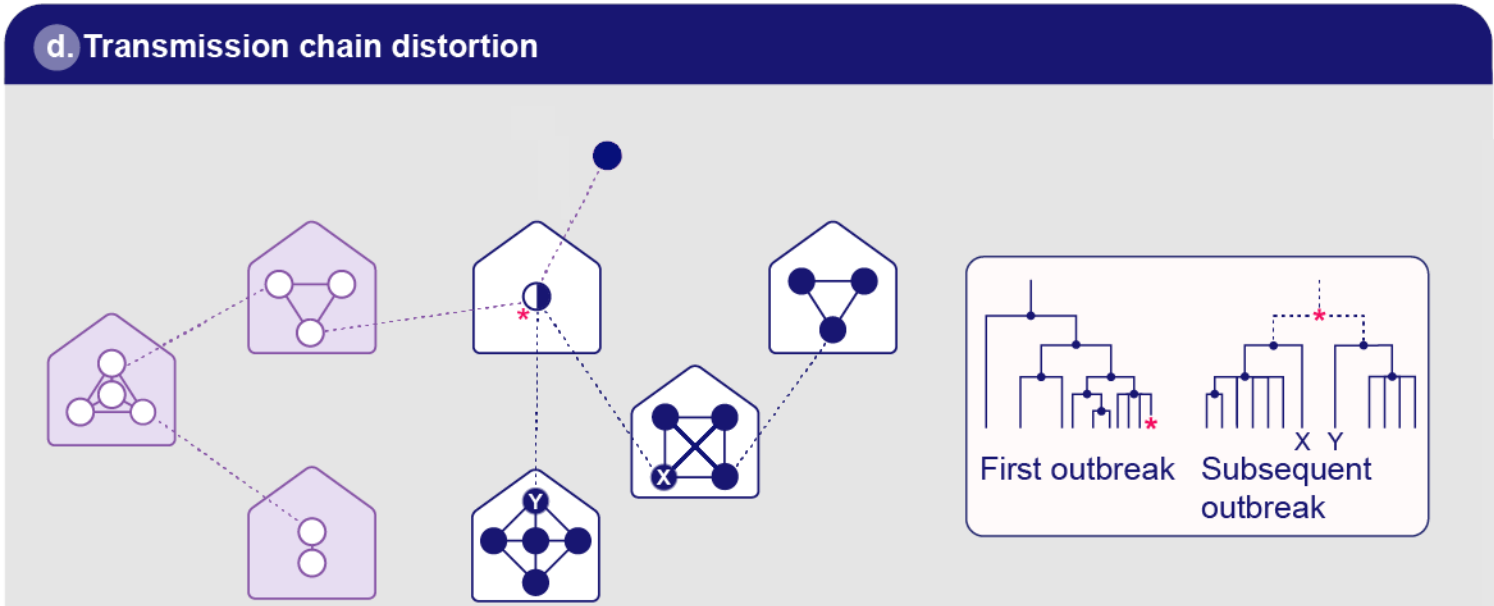
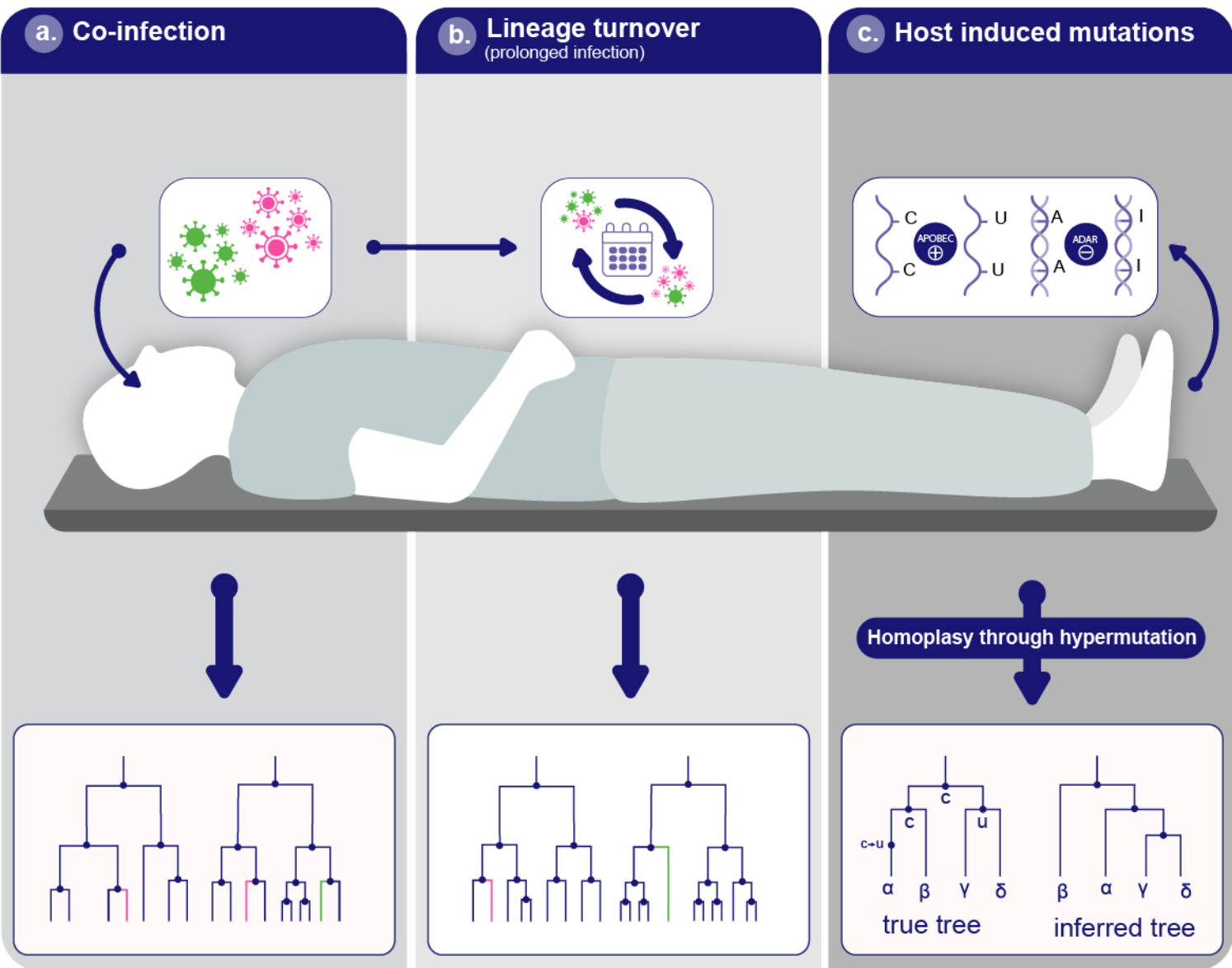
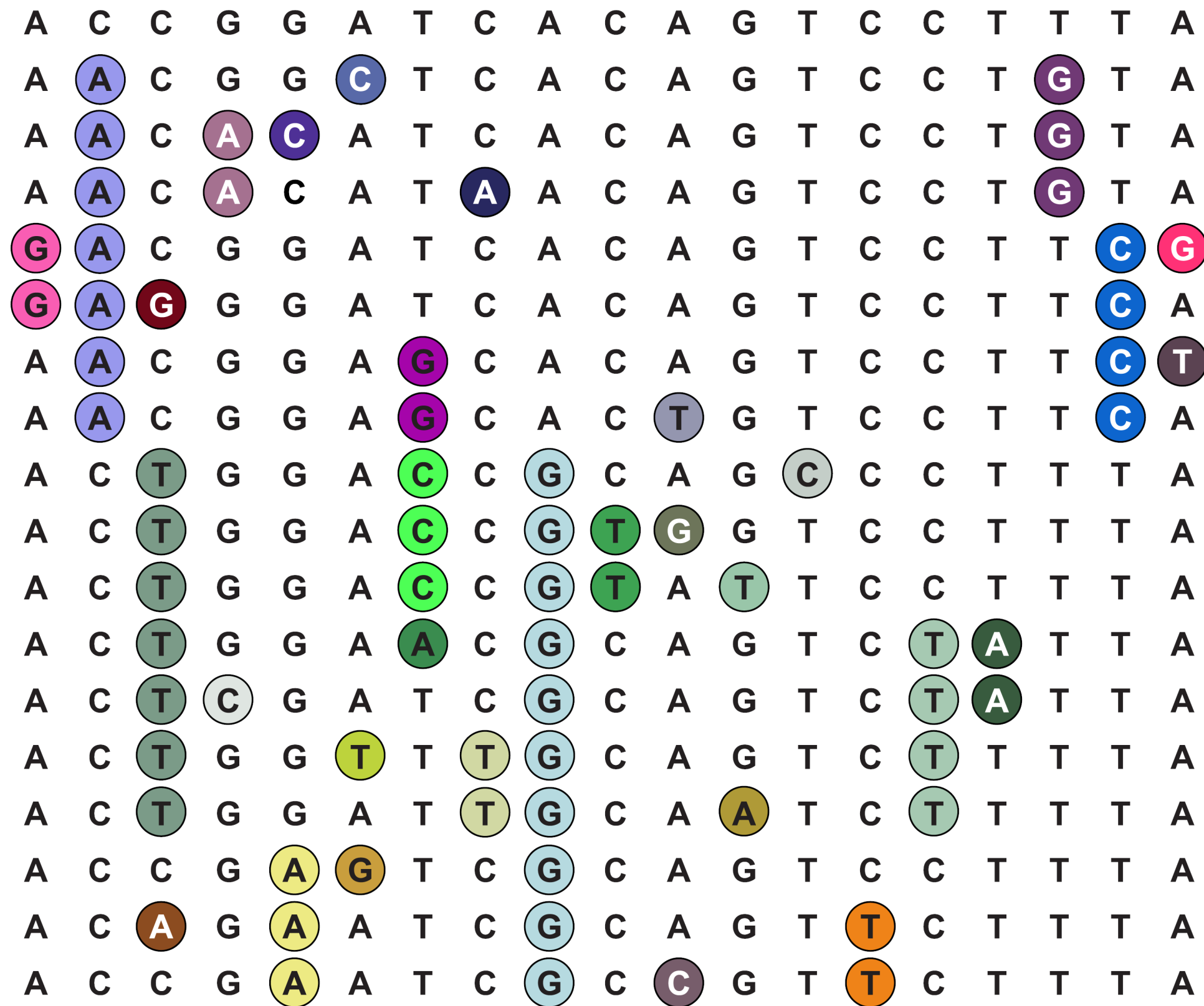


Fig. 4



Box 1. Fig.**Multiple Sequence Alignment****Phylogeny**