CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/150513/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Attwood, Stephen W., Hill, Sarah C., Aanensen, David M., Connor, Thomas R. and Pybus, Oliver G. 2022. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nature Reviews Genetics 23 , pp. 547-562. 10.1038/s41576-022-00483-8

Publishers page: http://dx.doi.org/10.1038/s41576-022-00483-8

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

Phylogenetic and phylodynamic approaches to understanding and combating the SARS-CoV-2 pandemic

Stephen W. Attwood[1,2], Sarah C. Hill[3], David M. Aanensen[4,5], Thomas R. Connor[2,6], and Oliver G. Pybus[1,3]

[1]Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

[2]Pathogen Genomics Unit, Public Health Wales NHS Trust, Cardiff, UK

[3]Department of Pathobiology and Population Sciences, Royal Veterinary College, University of London, UK

[4]Centre for Genomic Pathogen Surveillance, Wellcome Genome Campus, Hinxton, UK

[5]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[6]School of Biosciences, Cardiff University, Cardiff, UK

Corresponding authors:
OGP oliver.pybus@zoo.ox.ac.uk
SWA stephen.attwood@zoo.ox.ac.uk

Key words: COVID-19; epidemiology; evolution; genomics; hCoV-19; phylodynamics; phylogenetics; SARS-CoV-2

ABSTRACT

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

Introduction

The COVID-19 pandemic has triggered an unprecedented global response in pathogen genome sequencing. Nearly 400,000 full or partial SARS-CoV-2 genomes were generated and shared publicly within the first year of transmission. Whilst phylogenetic tools have become increasingly important in the public health management of a range of viral epidemics[1–4], the COVID-19 crisis is the first global health emergency in which large-scale, real-time genomic sequencing and analysis have underpinned public health decision making. The first 12 months of the pandemic were characterised by continual change in the global epidemiological and virological situation, and the analysis of genome sequences has proven essential in tracking the pandemic. Phylogenetic and phylodynamic approaches (Box 1) can unlock the potential of sampled gene sequences, and are often analysed in conjunction with other data sources. Such analyses have been used to quantify international 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 interest, to discover and analyse variants of concern, and to investigate intra-host virus evolution.

This review focuses on how SARS-CoV-2 transmission, epidemiology, and spatial dispersal has been measured and investigated through the phylogenetic and phylodynamic analysis of SARS-CoV-2 genomes (Fig. 1). It is intended to be a retrospective overview that uses examples from the first year of the pandemic to demonstrate the varied contributions of phylogenetics, in the context of different phases of pandemic responses. We will examine how such analyses have informed global efforts to understand, control and predict the course of the pandemic, and outline arising new challenges and how they are 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 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 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 interventions implemented in different nations, often in an *ad hoc* and pragmatic manner. We further highlight research that was influential in contributing to epidemiological

understanding and public health decision making. The first year also best demonstrates the potential of these methods for risk assessment, prediction, and control of future emerging viruses. We mostly refer to the genetic diversity of SARS-CoV-2 using the Phylogenetic Assignment of Named Global Outbreak (Pango) dynamic nomenclature[5] (see Box 1), but also sometimes use the World Health Organization (WHO) "Greek letter" nomenclature scheme for particular variants of concern (VOCs) and variants of interest (VOIs).

Tracking the global pandemic

Revealing how SARS-CoV-2 spread globally in early 2020 was important in informing public health strategies world-wide. Phylodynamic and phylogeographic methods can be used to estimate the timing and location of ancestral nodes within a molecular phylogeny[6–8], allowing inference of the 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.

International travel restrictions

A range of studies have investigated the impact of international travel restrictions in a phylogeographic framework, quantifying the absolute number of lineage introductions from abroad and the relative contribution of local 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 likely reflects the rapid declines in mobility as many countries concurrently imposed restrictions on international travel[9].

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 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 initially by dispersal from China, however the number of exportations declined rapidly following the cessation of China's international flights in January 2020[10]. Endemic transmission began in Italy during mid-February 2020, with establishment in other European countries soon thereafter[11]. The shift in global drivers of dissemination to predominantly intercontinental exportations from Europe, became

associated with the expansion of a lineage bearing the D614G spike mutation[12] (later designated as Pango lineage B.1). Virus lineage migrations from Europe to North America increased until the declaration by WHO of a pandemic on March 11th 2020, suggesting that air travel restrictions slowed spread[13]. In South Africa, international introductions plummeted after travel restrictions began on March 26th 2020[14]. Similar observations were made in other nationally-focused studies, including Italy[15], New Zealand, Australia, Iceland, Taiwan[16], and the UK[17].

The impact of international travel restrictions appeared significant, especially when combined with domestic transmission control, or when restrictions were implemented before full establishment of local transmission. A study of 427 genomes from Brazil applied a discrete asymmetric phylogeographic model and estimated at least 104 international introductions during March and April 2020; these fell into three monophyletic clades (see Box 1) of apparently European origin, and a molecular clock approach 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 have had little impact[18]. In the US, an early study investigated the efficacy of international travel restrictions in Connecticut[19]. Seven of nine Connecticut genomes fell into a clade of mostly Washington-state genomes, whereas two clustered with genomes from China and Europe. As the former had no history of recent travel, their phylogenetic 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 if implemented earlier[19]. Similar patterns were observed in other countries, including Italy[11] and the UK[20]. During 2020, more SARS-CoV-2 genomes were generated in the UK than any other country, allowing quantitative assessments of interventions in controlling introductions and domestic transmission. One study estimated that 33% of UK lineages originated in Spain, 12% from Italy, and 26% from elsewhere[17].

Many countries strengthened travel restrictions later in 2020, aiming to slow the spread of variants 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 indicated two independent international introductions of lineage B.1.1.7 (the alpha VOC) from London, UK[21]. These introductions occurred despite the suspension of flights to and from the UK. Similarly, phylodynamics was used to evidence multiple

international introductions to the US and hidden transmission of B.1.1.7 since November 2020, and to infer that lineage B.1.1.7 expanded to 33 states by January 2021 with a doubling time of 9.8 days[22]. Investigations have also considered the factors that drove the resurgence of transmission in Europe in late summer 2020. A recent study using a Bayesian time-scaled phylogeographic model (Box 1) found that by mid-August a large fraction of the lineages then circulating in each country had been introduced after June 15th, the date when many countries in the Schengen area opened their borders[23]. The study also found that newly introduced lineages tended to expand faster when entering a region of low incidence, and that for most countries resurgence was driven by new introductions rather than persistence of lineages from the spring[23].

Local transmission and interventions

Non-pharmaceutical interventions (NPIs) include travel restrictions, person-to-person distancing, or mandatory mask wearing. Two main phylogenetic approaches have been adopted for investigation of NPI impact. First, the frequency of lineage movement among regions within a country can be assessed using phylogeographic analyses (as discussed above for international dissemination). Second, estimates of virus population size, epidemic doubling time, and R_t can be calculated from virus genome sequences using phylodynamic approaches.

Molecular clock dating of SARS-CoV-2 lineages indicated multiple introductions from Wuhan to Guangdong in early January 2020, with a fall in lineage diversity thereafter, suggesting that travel restrictions combined with comprehensive tracing and isolation in Guangdong were effective in controlling transmission[24]. A phylogenetic study of transmission in Boston, USA, also reported a drop in importations to Boston from other domestic locations after national restrictions began[13.] By contrast, a study of NPI for Italy[15] suggested that domestic travel restrictions failed to prevent community transmission. One global study of 29,000 SARS-CoV-2 genome sequences used a compartmental structured coalescent model to estimate the time of epidemic seeding in 57 different locations[25]. The authors found that locations with early implementation of strong NPIs experienced less severe morbidity and mortality during the study[26] and that stringent interventions two weeks earlier would have approximately halved cumulative deaths in the immediate post-intervention period.

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 2020show Such methods were used to from genetic sequence data. effective population sizes models can estimate Coalescent[27]. Similar approaches were used to show that R_t fell in New Zealand in March 2020, from 7.0 at the beginning of the month to 0.2 by the end, demonstrating the impact of NPIs aimed at national disease elimination[28]. Taiwan,In $R_{t\,\text{decreased}}$ throughout the early pandemic even in the absence of substantially decreased local human mobility or stay-at-home orders[16]. This suggested that interventions such as contact tracing and widespread face mask use could be sufficient for adequate outbreak control (at least before the evolution of VOCs). Phylodynamic studies have provided other parameter estimates that are useful for understanding virus biology and transmission, or for use as priors (Box 1) in further Bayesian modelling (Table 1).

Phylodynamic analyses have repeatedly demonstrated hidden circulation of SARS-CoV-2 for days to months prior to first-case detection. Such results are important in determining whether existing surveillance adequately captures ongoing community transmission[29]. A US study of 346 genomes, 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 genomes clustered in a clade containing WA1 (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 2020. This date was used to parameterise a stochastic epidemiological model that suggested 1600 active infections in the State by mid-March[30]. Similarly, a molecular clock analysis of genomes from Scotland estimated transmission began around February 19th 2020, predating first case detection by almost two weeks[20].

Outbreak phylogenetics

Evolutionary approaches can help refute or confirm suspected transmission routes, supplementing 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 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 among cases.

Nosocomial transmission

Studies of healthcare settings are used to determine whether personal protective equipment (PPE) guidelines are sufficient to prevent nosocomial transmission. In Durban, South Africa, routine phylogenomic surveillance identified a monophyletic clade of cases in co-workers at a city hospital, suggesting a nosocomial outbreak. The observation of community cases with additional mutations, implied community transmission beyond the hospital[31]. A lack of phylogenetic clustering of cases by ward among healthcare workers in a Netherlands hospital showed community transmission to be more likely than nosocomial transmission[32]. Furthermore, a study in Australia ruled out associations among 54 cases across four health services, where shared health care workers had been initially implicated in dissemination. Phylogenomics revealed that the cases instead actually clustered according to a common social event[27].

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 extensive sequencing resulted in timely revision of the hospital's infection control procedures[33]. In a second UK study, phylogenetic analysis of infections from 31 care home staff and 61 residents 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 settings[34]. Policy change was also called for in a Boston hospital study; virus genomes with shared substitutions suggested at least two patient-to-staff transmission events, despite a lack of aerosol-generating procedures and the staff wearing masks and face-shields[35]. 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 seeded by a patient transferred from the other[36]. Multiple introductions were inferred for an outbreak at a San Francisco nursing facility, with one worker, who had also worked in Washington State, apparently responsible for introduction of WA1-related virus genomes[37]. Other applications of phylogenetics in investigations of outbreaks in medical or care settings are found in reports from Chile[38], France[39], Minnesota[40], and the Netherlands[41]. Nevertheless, whilst phylogenetics has allowed confirmation of nosocomial transmission in some cases, it has also helped reveal the

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

Public gatherings and super-spreading

Epidemiological studies of SARS-CoV-2 have indicated a relatively high attack rate[42,43]. A high attack rate has been supported by phylogenetics; for example, an outbreak affecting 11 workers in a 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)[44]. In some cases, local bursts of 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 monophyletic clade. 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 genomes[13]. This finding showed that individual mass-infection events could facilitate transmission and virus dissemination.

The role of large celebrations in triggering super-spreading can be also explored using phylogenetics. A discrete-state phylogeographical model 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 there[45]. Resurgence of an early outbreak in Japan was hypothesised initially to be linked to increased travel to cherry blossom sites during the national holiday of March 20th to 22nd 2020, potentially causing resurgence and growth of persistent low-level community transmission. Clarification through sequencing later showed that the late March cases were not directly related to cases from the first epidemic "wave"[46]. In Germany, three events at a Berlin nightclub in early March 2020 led to a series of outbreaks. Phylogenetics confirmed the club as a potential focus of super-spreading, and supported the decision in Germany to prohibit such events from March 16th [47]. In the US, phylodynamics linked the establishment of B.1.1.7 to the Thanksgiving holiday travel surge in November 2020[22].

Travel and transport

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 flight-attendants and two related passengers were detected with

B.1 lineage infections, despite B.1 being unknown in Hong-Kong at that time[48]. A similar indication of in-flight transmission was reported for a flight between Dubai (United Arab Emirates) and Auckland (New Zealand)
[49].

The predominance of one major clade in the February 2020 Diamond Princess cruise ship outbreak, suggested that most passengers became infected whilst attending on-board events, with a single introduction prior to quarantine measures[50]. Similarly, a phylogenetic study involving samples from 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. 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 California. This suggested that the source(s) of infection on the cruise were more likely local (i.e. California), rather than either of the cruise destinations. The second cruise, immediately following the first 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 WA1 genomes in general. This suggested that the shared cohort of passengers seeded the outbreak on the second cruise[51]. The patterns of shared, derived, mutations in the Grand Princess outbreaks 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, rather than from any station of disembarkation. The implications are that extensive revision and intensification of infection management procedures and practices are essential to protect passengers should cruise travel be permitted during pandemic events.

Genomic analyses aided the tracing of transmission during a Chinese-German business meeting in 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 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, and revealed the order of transmissions in a subsequent household cluster[52].

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 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 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 and recorded reverse transcription PCR (RT-PCR) negative in mid-April 2020[53] (see also REF[54]).

Tracking lineages of interest

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 analysis has revealed the independent emergence of VOCs, some of which share identical mutations (evolutionary convergence), and has reconstructed the accumulation of substitutions in time and space, shedding light on virus evolutionary or adaptive strategies.

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 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 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 Δ69/Δ70)[55]. Lineage B.1.1.7 became dominant in the UK just a few months after its emergence, and phylodynamic studies have shown it to have an estimated growth rate 40-70% higher than previous lineages[56]. 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 the variant may have arisen through a discrete evolutionary event during which multiple mutations were acquired, possibly during protracted infection of a single patient[57]. Slightly before the emergence of 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 receptor-binding domain (RBD) of its spike protein[14].

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 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 months[58]. 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 that define B.1.351 and appeared in samples from the lineage in November 2020[59]. Nevertheless, it is sometimes not possible to resolve the order of evolutionary events, either because genome sampling through time is insufficiently frequent or several mutations occurred very quickly. For example, ΔH69/V70 has arisen independently in several lineages (Fig. 3a), and is thought to compensate for decreased infectivity due to antibody escape substitutions such as N501Y (Fig. 4a,c); however, it is currently not clear whether or not the deletion preceded the RBD substitution in B.1.1.7[60].

The E484K mutation in B.1.351 has been associated with antibody escape and potential resistance to convalescent plasma therapies[55,61]. *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 phenotype mostly attributed to E484K and K417N)[62,63] and shows increased transmissibility[62]. Whilst the B.1.1.7 lineage did not carry E484K when it first emerged, by February 1st 2021 this mutation had appeared in thirteen English and two Welsh B.1.1.7 genomes. The phylogenetic relationships among these suggested at least two independent acquisitions of E484K in the UK. Lentiviral and vesicular stomatitis virus (VSV) pseudotyping experiments indicate that the E484K mutation on the B.1.1.7 lineage backbone results in a reduction of neutralising activity by vaccine sera[63,64]. The P.1 lineage was first reported in international travellers from Brazil entering Japan[65], and showed 11 amino acid substitutions relative to its ancestral lineage B.1.1.28. Three of these 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 lineages[55]. P.1 appears to have originated in Brazil[65,66] and also shows signs of increased transmissibility relative to its parental lineage B.1.1.28[67].

Whilst the phenotypic effect of mutations carried by VOCs can be investigated *in vitro* (see[68–71] for examples), their epidemiological significance is harder to evaluate. Changes in 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, ascertainment bias, and uneven sampling among regions[29]. Studies with a phylogenetic or phylodynamic basis have the potential to ameliorate some of these 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 614 rose from 10% prevalence before March 1st 2020, to overall global predominance by April 2020[69]. Relative growth rates for D614G and other substitutions were estimated by phylogenetic diversification; this suggested that most variants were weakly deleterious, and not more transmissible[72]. Sequence data from repeated international introductions of SARS-CoV-2 to the UK were leveraged to provide replicate observations of the growth of 614D and 614G lineages[73]. Modelling and phylodynamic analyses of 307 independent introductions between January 29th and June 16th, 2020 suggested a genuine (i.e. not a sampling effect) replacement of D by G in the UK, with a growth effect of around 20% and phylogenetic estimates of R_o for D of 2.7–3.5 and G 3.1–4.8; however, indications of 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 614G[74], noting that the expansion of 614G coincided with a shift in the nexus of global dispersal from Asia to Europe. Other lineages of potential concern have been detected through phylogenetic analyses, such as the multiple, expanding B.1 sublineages in the US and Canada, including some that have acquired mutation E484K[75,76].

Homoplasy and recombination

Lineages bearing N501Y and E484K appeared independently in Brazil, South Africa, Canada, and the UK in late 2020. Evolutionary convergence was observed, with the same changes being acquired independently on several branches scattered across the virus phylogeny (Fig. 3a; homoplasy), 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 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 different places and on different genomic backgrounds suggests there were shared selective pressures acting on the virus[77], such as the need to increase intrinsic transmissibility, extend the duration of infection, or to evade host immune responses (whether elicited by natural infection or vaccination)[78]. The parallel emergence of constellations of functionally-relevant mutations[79] further 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 or compensatory mutation (Fig. 4). Deletions and other rearrangements tend to command less discussion in the literature than substitutions; however, as SARS-CoV-2 has exonuclease-based proofreading (which can correct nucleotide substitutions), a role for deletions is not unexpected. Indeed, deletions with demonstrated effects on the binding of some neutralising monoclonal antibodies have been found, concentrated in four recurrent deletion regions in the spike N-terminal domain (NTD)[80]. The potentially-important role of deletions also indicates that indels might best be coded as phylogenetic characters and not excluded from analyses.

The epidemiological context of these convergent changes indicates that they arose through independent, parallel mutation. However, it is known that such changes (homoplasies) can arise also through recombination, and evolutionary analyses suggest that recombination could be now relevant to SARS-CoV-2 evolution[81]. The level, scale and consequences of recombination during the pandemic are unclear; one earlier study of phylogenetic inconsistency found no clear signals of recombination[82], whereas a more recent analysis of UK sequence data discovered at least four groups of natural recombinants of B.1.1.7, and other parental lineages[83]. The increasing co-circulation in 2021 of genetically-diverse viruses increases the likelihood that further SARS-CoV-2 recombinants will be detected.

It is also possible that co-infections may complicate tracing of transmission networks[84]. An individual involved in one outbreak, may also participate in a second, heterochronous and phylogenetically distinct, transmission chain (Fig. 5d). Such confusion may occur where different lineages of virus dominate the intra-host

population at different times (perhaps due to antagonistic evolution with lineage-specific host immune responses[85]) in the co-infected index patient. Nevertheless, the problem is likely to be restricted to certain scenarios, such as a major nosocomial outbreak, where co-infection is particularly likely.

To date there is little evidence that specific intra-host single nucleotide substitutions (iSNSs) are associated with antiviral drug resistance or infection outcomes. Elevated C→U in SARS-CoV-2 relative to the four long established circulating coronaviruses of humans, and the low GC content of seasonal coronaviruses in general, has been attributed to targeting of SARS-CoV-2 RNA genomes by host apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like cytidine deaminases (APOBECs) (Fig. 5c) and to selection against CpG dinucleotides to avoid targeting by host zinc finger antiviral proteins[86]. The same hypermuation has been detected in SARS-CoV-2 from infections in mink, and appears to be an effect mediated by the virus rather than the host[87]. The association of C→U with particular APOBEC targets can generate highly recurrent homoplasies that mimic convergent evolution and lead to false inference of selection; this can potentially impact phylogenetic methods and bias molecular clock studies[88]. Nevertheless, many C→U changes occur outside APOBEC target motifs and phylodynamic studies of host-virus interactions would benefit from a greater understanding of the processes involved. The observation that U nucleotides were less common than expected at 4-fold degenerate sites[89], led to inference of ongoing selection for a reduced U content, as selection acted against the many deleterious mutations generated through the C→U bias. By contrast, other studies[90] compared low-frequency against high-frequency mutations within each mutational class and found that although rates of C→U and G→U were both high, no significant signals of selection against U remained after accounting for changes in mutation rates caused by a skewed mutational spectrum and selective pressures associated with a recent host shift.

Several studies have used deep-sequencing to estimate levels of virus genetic variation within infected individuals[91–93]. The majority of iSNSs do not appear to be effectively transmitted between patients[92,94,95]. Nevertheless, at least one study found that, despite the low proportion that are stably transmitted, some iSNSs were phylogenetically associated, suggesting preservation of diversity at some sites after transmission[94]. A study of 210

sequence-read sets suggested that, despite narrow transmission bottlenecks, non-synonymous iSNSs that were shared among hosts were three times more likely to occur than unshared iSNSs; again host defences (e.g. APOBEC3 enzymes) were suggested as shaping this pattern of recurrence[92].

The lack of onward transmission of iSNSs suggests that antibody escape mutations may carry some cost in terms of transmissibility. For example, spike D796H and ΔH69/V70 emerge during therapy with convalescent plasma[96]. Although observed in patients, D796H is not characteristic of any major lineage transmitting among humans—D796Y occurs in A.27 but this lineage showed low global prevalence (<0.1%) and range, and has been effectively unobserved since May 2020[97]. Deep-sequencing and phylogenomics are used to understand co-infections (Fig. 5a) and prolonged infections (Fig. 5b), and to distinguish chronic infections from re-infections. A study of within-host variation[84] used a linear model to estimate over 36 possible co-infections among 1179 cases. A later study[85] speculated that different lineages detected in a patient within 21 days were the result of changing dominance of two co-infecting lineages over time, although the authors could not fully exclude reinfection.Studies of within-host variation

Tackling sampling bias in epidemiology

Uneven sampling of genomes is a considerable problem for SARS-CoV-2 phylogenetics. Sampling was effectively absent during the first days and weeks, becoming more extensive as the pandemic progressed[98], and often concentrated toward particularly large outbreaks or the radiation of VOCs. Some countries sequence routinely, others only for outbreak investigation, and some not at all.

Ascertainment bias towards symptomatic cases potentially complicated attempts to determine any greater transmissibility conferred by spike 614G over 614D[73], and undersampling in a region of high incidence has been suggested as a cause of overestimation of size and duration of SARS-CoV-2 transmission chains[24]. Methods that can better accommodate known sampling biases are urgently required;

one current solution is the use of a structured (epoch based) model. Such models condition on the rate of genomic sampling relative to all PCR-confirmed SARS-CoV-2 cases, and reportedly improve molecular clock accuracy[99]. Methods have been developed that can accommodate changing rate of sequencing through time, for example the Bayesian Epoch Skyline Plot (ESP) approach[100.] (see also Box 1). The ESP infers virus outbreak size over time, whilst modelling the sampling strategies commonly adopted by epidemiologists. An alternative is to model sampling whilst linking sample location to regional variations in sampling effort; this has improved estimation of population size history for at least some datasets[101]Genetic variation and transmission patterns are interdependent, hence a clearer picture can be obtained by combining phylodynamic estimation with epidemiological data. There has been notable

· progress on such integrated approaches. One recent method allowed incorporation of non-genomic incidence data and epidemic dynamics models with a novel phylodynamic approach extended to handle sampling of both original and downstream members of transmission chains (i.e. phylogenies with extant internal nodes); this joint epidemiological and phylodynamic analysis is relatively less susceptible to bias due to undiagnosed cases, imported cases and changes in sampling levels, and so produces more reliable estimates of transmission rates than epidemiological data alone[102].

Undetected SARS-CoV-2 transmission has been incorporated in Bayesian phylodynamics using an epidemiological model that includes data on confirmed, but unsequenced, cases 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 in sampling strategy, such as a decision to begin testing asymptomatic individuals[103]. Both approaches have been successfully used to infer R_0 and cumulative case count trajectories for the Diamond Princess cruise ship outbreak, a closed system for which reliable epidemiological (non-genetic) data are available to validate corresponding phylodynamic estimates[103,104]. These, and similar efforts to unite phylogenetics and epidemiology, are promising tools for the study of viral epidemics, including SARS-CoV-2[105].

Further innovation is nevertheless required, particularly concerning the estimation of large phylogenies of rapidly evolving viruses. Many current solutions have inherent assumptions such as negligible variation within patients, and absence of superinfections — assumptions which 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 biological science disciplines, often in a highly collaborative manner. For example, approaches to quasispecies deconvolution were adopted from practices in oncology[106,107]. In addition, there is a vast literature devoted to the reduction of technical error and improvement of genome sequence quality[108,109].

Conclusions and the way forwards

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 over the past decade in virus genomics and phylodynamic analyses, whilst revealing technical and social challenges that remain to be addressed. The rapid, open sharing of protocols and data has been critically important, and more extensive for SARS-CoV-2 than ever before, yet hesitancy to share sequencing data prior to publication remains[110] because of concerns that data may be used elsewhere without appropriate credit being given to producers[111]. Greater insights into SARS-CoV-2 transmission could be gained through the incorporation of more and varied data (e.g., mobility data); however, this must be balanced with privacy and anonymisation concerns.

Nomenclature of lineages and variants was initially inconsistent; this complicates scientific discussion, and encourages the media to adopt simple but inappropriate naming of lineages based on the location of their first detection (e.g., “South Africa variant”)[112]. The problem of toponymic naming in the popular literature has been partly overcome by the adoption of Greek 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. 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 designation[113].

In many countries, current research recruitment, evaluation and funding frameworks disincentivise 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 evaluate career progress (e.g. research publications and grants)[114]. Consequently, new career pathways or evaluation systems are required to encourage greater embedding of evolutionary 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, where the capacity for computational analysis sometimes lags behind that for genetic sequencing[115]. 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 recombinants, and cross-species transmission events.

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 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 overview of virus epidemiology to be obtained even with samples taken in limited geographical areas or unevenly through time. Accordingly, phylogenomic 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.

Acknowledgements
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 the Wellcome Trust [Sir Henry Wellcome Postdoctoral Fellowship 220414/Z/20/Z].

Table 1 Pango lineages of interest or concern during the first year of the pandemic

Pango lineage	*Nextstrain Clade*	*Public Health Authority denotations*	*Territory of first reporting*
A.23.1	NA	VUI-21FEB-01* (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	VOC-20DEC-01 (WHO alpha)	UK
B.1.1.7	NA	VOC-21FEB-02* (VOC-202102/02)	UK
B.1.324.1	NA	VUI-21MAR-01* (VUI-202103/01)	UK (links with travel from Antigua)
B.1.351	20H/501Y.V2	VOC-20DEC-02 (WHO beta)	South Africa
B.1.525	20A/S:484K	VUI-21FEB-03 (VUI-202102/03)†	UK (associations with Angola)
B.1.617	NA	VUI-21APR-01 (617.2 WHO delta)	India
P.1	20J/501Y.V3	VOC 202101/02 (WHO gamma)	Japan (in arrivals from Brazil)
P.2	NA	VUI-21JAN-01 (VUI-202101/01)	
P.3	NA	VUI-21MAR-02	Philippines (Central Visayas)

Although alternative denotations may, to varying degrees, correspond to Phylogenetic Assignment of Named Global Outbreak (Pango) lineages, the 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 WHO or UK public health authorities[66,97,116].

* Refers only to variants within the respective lineage that show E484K.
† Briefly known as UK1188.

Table 2 SARS-CoV-2 epidemiological parameter estimation 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)[16]	6.91e-04 (6.00e-04, 7.78e-04)[16]	MTBD
Australia	Prior to 27/03 Post 27/03	R_t = 1.63 (1.45, 1.8)[27] R_t = 0.48 (0.27, 0.69)[27]	1.1e-03 1.1e-03[27]	BCP+SC
Iceland	18/03–29/04	R_t = 1.4 (1.2, 1.59)[16]	5.75e-04 (4.96e-04, 6.47e-04)[16]	MTBD
Italy	22/02–04/04	R_t = 2.25 (1.5, 3.1)[117]	1.16e-03 (1.01e-03, 1.32e-03)[118]	BCP, BCP+SC
New Zealand	26/03–29/04	R_t = 1.41 (1.07, 1.89)[16]	6.09e-04 (5.16e-04, 7.03e-04)[16]	MTBD
Russia (Vreden hospital)	27/03–08/04 08/04–23/04	R_t = 3.72 (2.48, 5.05)[119] R_t = 1.38 (0.48, 2.41)[119]	9.43e-04 (8.46e-04, 1.04e-03) 9.43e-04 (8.46e-04, 1.04e-03)[119]	BCP+SC BCP+SC
Taiwan, Tâi-oân pún-tó	27/03–29/04	R_t = 1.02 (0.825, 1.22)[16]	8.00e-04 (6.89e-04, 9.17e-04)[16]	MTBD
Weifang, Shandong	25/01–10/02	R_0 = 3.4 (2.1, 5.2)[120]	1.30e-03 (0.98e-03, 1.7e-03) [120]	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[71,121] or reproduction numbers[26,28,73,122,123]. Confidence intervals are provided where available (95% highest posterior density, (HPD)). Dates are dd/mm in 2020. Abbreviations: BCP, Bayesian Coalescent Phylodynamic; CFEM, Coalescent Fitted Epidemiological Model; MTBD, Multitype Birth-Death Model; SC, Structured Coalescent.

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-Infected-Recovered (SIR) 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 and interactions within the host. TMRCA, time to the most recent common ancestor. Panel header colours indicate related themes: colour 1, public health; colour 2, epidemiological parameters; colour 3, clinical parameters.

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 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 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 origins. The B.1.1.7 lineage coloured in red differs from the other B.1.1.7 viruses (and all other lineages here) in that it bears S494P rather than a substitution at E484. Lineage B.1.617 bears E484Q rather than E484K. Some 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 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 aimed at detecting any selective advantage of the VOCs, including the search for vaccine escape phenotypes. Lineages and variants are based on the following publications: A.23.1[124]; B.1.1.318, B.1.1.7+E484K, B.1.1.7+S494P, B.1.324.1[66]; B.1.351[14,63,66]; B.1.525[66]; B.1.617[125]; P.1[65]; P.2[126,127]; P.3[128]. Pin heights indicate time relative to detection of the first lineage, i.e. P.2 in Rio de Janeiro, October 13th, 2020 (not to scale, but ranked in time).

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 ΔH69/V70, indicated in blue, in genomes of the B.1.1.7 and B.1.258 lineages respectively — phylogeny from Nextstrain[129,130] (Europe ncov GISAID dataset), visualised in Figtree. . B.1.258 includes branches both with and without the deletionShaded region indicating the deletions inFor clarity, not all Pango lineages are shown. **b |** By the start of 2020 a number of commonly occurring spike substitutions and deletions had been recognised as shared among lineages. The illustrated substitutions are found in the exposed (i.e. outermost on the surface of the virion) subunit of spike, 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. "Mink" refers to the SARS-CoV-2 mink-human sublineage, termed ‘Cluster 5’, which exhibited ΔH69/V70 and N501T (and other spike substitutions)[131]; the second B.1.1.7 lineage (VOC-202102/02, the ellipse with broken-line border) is a cluster of B.1.1.7 that also bears E484K[132]. 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 original mink farm epidemic in Denmark. Nevertheless, N501T appeared to have emerged independently four times, and has been detected in ten human cases[133]. L18F is an NTD substitution found in B.1.351 and increasing in frequency in B.1.1.7[60]. As in Fig. 2 we see that the same substitutions appear in multiple lineages, implying that they arose 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 lineage; this suggests epistatic interactions, with perhaps compensatory changes following immune escape variants (see Fig. 4).

Fig. 4: Stylised interactions between Spike mutations and antibodies, and potential compensatory and epistatic interactions with other, co-occurring, mutations.

a | The receptor-binding domain (RBD) mutation K417T of the SARS-CoV-2 spike protein has been empirically shown to reduce affinity to angiotensin-converting enzyme 2 (ACE2; the primary host receptor protein for SARS-CoV-2), but confer immune escape properties[134]. By contrast, N501Y showed increased ACE2 binding *in vitro*[134] — of the several mutations in the P.1 (gamma) variant, N501Y is hypothesized to partially rescue the loss of ACE2 affinity caused by the co-occurring K417T mutation[135]. **b** | E484K and K417N inhibit neutralising antibody action against some important antibody classes *in vivo* [79]; however, the inhibitory effect of K417N is generally weaker than that of E484K, and K417N/T mutations tend to occur only following replacements E484K or N501Y, as these are hypothesised to compensate for the reduction in ACE2 affinity caused by K417N/T[134] (a reduction also demonstrated by *in vivo* binding experiments[136]). Nevertheless, E484K in the presence of K417N (and/or N501Y) appears to confer a major conformational change in spike, which leads to increased ACE2 contact in model simulations[137]. **c** | ΔH69/V70 often arises after antibody escape substitutions at the RBD, such as N439K and Y453F[60]. The double deletion may help compensate for lower infectivity following the RBD mutations, and shows 2-fold increase in cell infectivity *in vitro*[131]. **d** | The Δ242-244 deletion shows host antibody escape *in vitro*[138], with molecular modelling indicating that this is due to disruption of the N-terminal domain (NTD) hydrophobic pocket[139]. Relative protein stabilities suggested that the destabilising effect of Δ242-244 could be offset by co-occurring stabilising mutations such as D215G and K417N, which restore spike expression *in vitro*[140]. **e** | Hypothesis explaining antibody avoidance by ΔY144, as suggested by molecular modelling[131]. ΔY144 shows neutralising antibody escape *in vitro*[141,142], and commonly occurs with ΔH69/V70, which is hypothesised to cause near complete removal of a commonly targeted epitope on the front side of the NTD. ΔY144 is conformationally sensitive and likely to be affected by co-mutations outside the epitope region[143] — further work is required to asses these interactions. **f** | N439K (characteristic of lineage B.1.258)[144] is hypothesised to compensate for K417V (empirically shown to reduce ACE2 binding) through enhanced ACE2 affinity seen *in vitro*. N439K shows higher viral loads and antibody escape *in vivo*[145]. Molecules are simplified and subunits are not to scale. LC CDR, Light Chain Complementarity-Determining Region of antibody; MoAb, Monoclonal Antibody; S1 and S2 refer to Spike subunits one and two; Vh, variable region of antibody heavy chain; Vl, variable region of antibody light chain.

Fig. 5: 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 (i.e. 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, for example with APOBECs, C→U homoplasies (convergent evolution) and changes in virus genome CpG content as a response. 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.

Box 1. Phylogenetic terminology and concepts [Contains a figure]

Phylogenetics

Phylogenetics has made an invaluable contribution to our understanding of the first year of the pandemic, and continues to do so. 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) hypothesised 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 evolutionary steps (parsimony) or that maximises 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) as marginal probabilities for the events (e.g. a prior distribution for outbreak onset time) gives the approach an advantage over simple maximum likelihood estimation.

Phylogenies have also formed the basis of a system for the identification, definition and monitoring of outbreak clusters and VOCs. Although nomenclatures such as that currently adopted by the 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 phylogenetic clades. 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 clade A in the figure). 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. Pango also delimits lineage clades using defining constellations of mutations, which we here refer to as **variants**.

Phylodynamics

Phylodynamics been used to incorporate epidemiological data in phylogenetic studies of the pandemichasmolecular clocks, models of virus population growth, sampling models, and epidemiological models, and , such as methods with other models

phylogeneticcombines. 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 of a clade (TMRCA).

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 visualised 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. The coalescent is most commonly used in '**skyline plot**' methods to estimate historical changes in population size (e.g. virus demographics); these are plots of effective population size (N_e) with time. Various "skyline" methods exist, and these generally differ in the smoothing of the population size transition boundaries, where growth is modelled as stepwise between 'epochs' of constant size. The model can also be modified to allow for expected population structure (**structured coalescent**). Similarly, epidemiological models, such as the Susceptible-Infected-Recovered (SIR) 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.

Glossary

Attack rate

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

Convalescent plasma

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

Convergence (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.

Effective population size

(N_e). A population genetic parameter that is proportional to the true population size.

Founder effect

Patterns in gene (variant) frequencies resulting from chance colonisation 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.

Indel

An insertion or deletion of bases in a DNA or RNA sequence observed as a mutation within an individual or as a heritable polymorphism in a population; these may be used as phylogenetic or taxonomic characters.

Nocosomial transmission

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

Pseudotyping experiment

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).

Quasispecies

Genetically distinct virus populations coexisting within one individual host; these may exhibit turnover (see Fig. 5b).

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 of a nascent virus).

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 and the accumulation of resistant individuals in the population.

Substitution

Here substitution refers to a mutation that has persisted through viral generations (i.e. is transmissible), reaching sufficient population frequency so as to appear in consensus genomes, and therefore representing a polymorphism of non-trivial frequency.

Superinfection

A second infection, or subsequent infections of the same or a different organism, establishes 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.

TMRCA

(Time to the most recent common ancestor). The time back to 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.

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 US.

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. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).
6. Rambaut, A. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**, 395–399 (2000).
7. 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).
8. Dellicour, S. *et al.* Epidemiological hypothesis testing using a phylogeographic and phylodynamic framework. *Nat. Commun.* **11**, 5620 (2020).
9. 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.
10. Yang, J. *et al.* Uncovering two phases of early intercontinental COVID-19 transmission dynamics. *J. Travel Med.* **27**, taaa200 (2020).
11. 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).
12. Fountain-Jones, N. M. *et al.* Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *Virus Evol.* **6**, veaa082 (2020).
13. Lemieux, J. E. *et al.* Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. *Science* **371**, eabe3261 (2021).
14. Tegally, H. *et al.* Emergence of a SARS-CoV-2 variant of concern with mutations in spike glycoprotein. *Nature* **592**, 438–443 (2021).
15. Di Giallonardo, F. *et al.* Genomic Epidemiology of the First Wave of SARS-CoV-2 in Italy. *Viruses* **12**, (2020).
16. 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).
17. Plessis, L. du *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).
18. Candido, D. S. *et al.* Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science* **369**, 1255–1260 (2020).
19. 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).
20. 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).
21. 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).

22. 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.
23. 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.
24. Lu, J. *et al.* Genomic Epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997–1003.e9 (2020).
25. Siveroni, I. A. & Volz, E. M. PhyDyn: Epidemiological Modelling in BEAST. (2017).
26. 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).
27. Seemann, T. *et al.* Tracking the COVID-19 pandemic in Australia using genomics. *Nat. Commun.* **11**, 4376 (2020).
28. 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).
29. 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.
30. Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571–575 (2020).
31. 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).
32. 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).
33. 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).
34. 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).
35. 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.
36. 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).
37. 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).
38. Olmos, C. *et al.* SARS-CoV-2 infection in asymptomatic healthcare workers at a clinic in Chile. *PLOS ONE* **16**, e0245913 (2021).
39. 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).
40. 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).
41. 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).
42. 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).
43. Liu, Y., Eggo, R. M. & Kucharski, A. J. Secondary attack rate and superspreading events for SARS-CoV-2. *The Lancet* **395**, e47 (2020).

44. 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).
45. 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.
46. Sekizuka, T. *et al.* A Genome Epidemiological Study of SARS-CoV-2 Introduction into Japan. *mSphere* **5**, e00786-20 (2020).
47. Muller, N. *et al.* Severe Acute Respiratory Syndrome Coronavirus 2 Outbreak Related to a Nightclub, Germany, 2020. *Emerg. Infect. Dis.* **27**, 645–648 (2020).
48. Choi, E. M. *et al.* In-Flight Transmission of SARS-CoV-2. *Emerg. Infect. Dis. J. - CDC* **26**, eid2611.203254 (2020).
49. 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).
50. 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).
51. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582–587 (2020).
52. 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).
53. 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).
54. Tillett, R. L. *et al.* Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect. Dis.* **21**, 52–58 (2021).
55. ECDC. Risk related to the spread of new SARS-CoV-2 variants of concern in the EU/EEA, first update–21January 2021. (2021).
56. Lemey, P. *et al.* Accommodating individual travel history and unsampled diversity in Bayesian phylogeographic inference of SARS-CoV-2. *Nat. Commun.* **11**, 5110 (2020).
57. 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).
58. 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).
59. Tegally, H. *et al.* Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **592**, 438–443 (2021).
60. Gupta, R. K. Will SARS-CoV-2 variants of concern affect the promise of vaccines? *Nat. Rev. Immunol.* **21**, 340–341 (2021).
61. Andreano, E. *et al.* SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proc. Natl. Acad. Sci.* **118**, e2103154118 (2021).
62. 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).
63. Wang, P. *et al.* Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. *Nature* **593**, 130–135 (2021).
64. Collier, D. A. *et al.* Sensitivity of SARS-CoV-2 B.1.1.7 to mRNA vaccine-elicited antibodies. *Nature* **593**, 136–141 (2021).
65. Faria, N. R. *et al.* Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. (2021).
66. Variant Technical Group. SARS-CoV-2 variants of concern and variants under investigation in England: Technical briefing 7. (2021).

67. 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).
68. Hou, Y. J. *et al.* SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. *Science* **370**, 1464–1468 (2020).
69. 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).
70. 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.
71. Díez-Fuertes, F. *et al.* A Founder Effect Led Early SARS-CoV-2 Transmission in Spain. *J. Virol.* **95**, JVI.01583-20 (2021).
72. van Dorp, L. *et al.* No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat. Commun.* **11**, 5986 (2020).
73. Volz, E. *et al.* Evaluating the effects of SARS-CoV-2 Spike mutation D614G on transmissibility and pathogenicity. *Cell* **184**, 64-75.e11 (2020).
74. 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).
75. Pater, A. A. *et al.* Emergence and Evolution of a Prevalent New SARS-CoV-2 Variant in the United States. *bioRxiv* 2021.01.11.426287 (2021) doi:10.1101/2021.01.11.426287.
76. Tu, H. *et al.* Distinct Patterns of Emergence of SARS-CoV-2 Spike Variants including N501Y in Clinical Samples in Columbus Ohio. *bioRxiv* 2021.01.12.426407 (2021) doi:10.1101/2021.01.12.426407.
77. Gutierrez, B., Escalera-Zamudio, M. & Pybus, O. G. Parallel molecular evolution and adaptation in viruses. *Curr. Opin. Virol.* **34**, 90–96 (2019).
78. Day, T., Gandon, S., Lion, S. & Otto, S. P. On the evolutionary epidemiology of SARS-CoV-2. *Curr. Biol.* **30**, R849–R857 (2020).
79. 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).
80. McCarthy, K. R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139–1142 (2021).
81. 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.
82. 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.
83. 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).
84. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife* **10**, e66857 (2021).
85. 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).
86. Mourier, T. *et al.* Host-directed editing of the SARS-CoV-2 genome. *Biochem. Biophys. Res. Commun.* **538**, 35–39 (2021).
87. Forni, D., Cagliani, R., Pontremoli, C., Clerici, M. & Sironi, M. The substitution spectra of coronavirus genomes. *Brief. Bioinform.* (2021) doi:10.1093/bib/bbab382.

88. 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).
89. 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).
90. De Maio, N. *et al.* Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, (2021).
91. Avanzato, V. A. *et al.* Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer. *Cell* **183**, 1901-1912.e9 (2020).
92. Armero, A., Berthet, N. & Avarre, J.-C. Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* **13**, 133 (2021).
93. Tarhini, H. *et al.* Long term SARS-CoV-2 infectiousness among three immunocompromised patients: from prolonged viral shedding to SARS-CoV-2 superinfection. *J. Infect. Dis.* jiab075 (2021) doi:10.1093/infdis/jiab075.
94. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* abg0821 (2021) doi:10.1126/science.abg0821.
95. Wang, D. *et al.* Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2. *Front. Med.* **8**, 585358 (2021).
96. Kemp, S. A. *et al.* SARS-CoV-2 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).
97. O'Toole, A. *et al.* PANGO lineages. (2021).
98. 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).
99. 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).
100. 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).
101. Cappello, L. & Palacios, J. A. Adaptive preferential sampling in phylodynamics. *ArXiv200902307 Q-Bio Stat* (2020).
102. Vaughan, T. G. *et al.* Estimating Epidemic Incidence and Prevalence from Genomic Data. *Mol. Biol. Evol.* **36**, 1804–1816 (2019).
103. 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.
104. 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.
105. 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).
106. 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.
107. 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).
108. Charre, C. *et al.* Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* **6**, veaa075 (2020).
109. 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).

110. Kalia, K., Saberwal, G. & Sharma, G. The lag in SARS-CoV-2 genome submissions to GISAID. *Nat. Biotechnol.* **39**, 1058–1060 (2021).
111. Noorden, R. V. Scientists call for fully open sharing of coronavirus genome data. *Nature* **590**, 195–196 (2021).
112. Callaway, E. 'A bloody mess': Confusion reigns over naming of new COVID variants. *Nature* **589**, 339–339 (2021).
113. Ferguson, C. Don't let "delta plus" confuse you. The strain hasn't learned any new tricks. (2021).
114. Hodcroft, E. B. *et al.* Want to track pandemic variants faster? Fix the bioinformatics bottleneck. *Nature* **591**, 30–33 (2021).
115. WHO. *Genomic sequencing of SARS-CoV-2: A guide to implementation for maximum impact on public health*. (World Health Organization, 2021).
116. Public Health England. Variants: distribution of cases data. (2021).
117. Lai, A. *et al.* Molecular Tracing of SARS-CoV-2 in Italy in the First Three Months of the Epidemic. *Viruses* **12**, 798 (2020).
118. 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).
119. Komissarov, A. B. *et al.* Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia. *Nat. Commun.* **12**, 649 (2021).
120. Geidelberg, L. *et al.* Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evol.* **7**, veaa102 (2021).
121. Duchene, S. *et al.* Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* **6**, veaa061 (2020).
122. 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.
123. Moreno, G. K. *et al.* Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat. Commun.* **11**, 5558 (2020).
124. 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).
125. 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).
126. 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).
127. 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).
128. Public Health England. What do we know about the new COVID-19 variants? - Public health matters. (2021).
129. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
130. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
131. 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).
132. 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).
133. 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.
134. 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).

135. Dejnirattisai, W. *et al.* Antibody evasion by the P.1 strain of SARS-CoV-2. *Cell* **184**, 2939-2954.e9 (2021).
136. Tanaka, S. *et al.* An ACE2 Triple Decoy that neutralizes SARS-CoV-2 shows enhanced affinity for virus variants. *Sci. Rep.* **11**, 12740 (2021).
137. Nelson, G. *et al.* Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y.V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant. *bioRxiv* (2021).
138. Wibmer, C. K. *et al.* SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nat. Med.* **27**, 622–625 (2021).
139. Resende, P. C. *et al.* The Ongoing Evolution of Variants of Concern and Interest of SARS-CoV-2 in Brazil Revealed by Convergent Indels in the Amino (N)-Terminal Domain of the Spike Protein. *Virus Evol.* (2021) doi:10.1093/ve/veab069.
140. Javanmardi, K. *et al.* Rapid characterization of spike variants via mammalian cell surface display. *biorxiv* 2021.03.30.437622 (2021) doi:10.1101/2021.03.30.437622.
141. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332-2347.e16 (2021).
142. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).
143. Hastie, K. M. *et al.* Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study. *Science* **374**, 472–478 (2021).
144. Verkhivker, G. M., Agajanian, S., Oztas, D. Y. & Gupta, G. Comparative Perturbation-Based Modeling of the SARS-CoV-2 Spike Protein Binding with Host Receptor and Neutralizing Antibodies: Structurally Adaptable Allosteric Communication Hotspots Define Spike Sites Targeted by Global Circulating Mutations. *Biochemistry* **60**, 1459–1484 (2021).
145. Thomson, E. C. *et al.* Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell* **184**, 1171-1187.e20 (2021).