Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020

Maureen Smith  
Robert Koch Institute

Maria Trofimova  
Robert Koch Institute

Ariane Weber  
Max-Planck Institute

Yannick Duport  
Robert Koch Institute

Denise Kühnert  
Department of Archaeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany  
https://orcid.org/0000-0002-5657-018X

Max von Kleist  
kleistm@rki.de  
MF1 Bioinformatics, Robert Koch-Institute  
https://orcid.org/0000-0001-6587-6394

Article

Keywords: SARS-CoV-2, epidemiology, genomes

Posted Date: May 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-558667/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Version of Record: A version of this preprint was published at Nature Communications on October 14th, 2021. See the published version at https://doi.org/10.1038/s41467-021-26267-y.
Rapid incidence estimation from SARS-CoV-2 genomes reveals decreased case detection in Europe during summer 2020

Maureen Rebecca Smith¹, ², *, Maria Trofimova¹, ², *, Ariane Weber³, Yannick Duport¹, ², Denise Kühnert³, ⁴, and Max von Kleist¹, ², *, +

¹ Systems Medicine of Infectious Disease (P5), Robert Koch Institute, Berlin, Germany
² Bioinformatics (MF1), Robert Koch Institute, Berlin, Germany
³ Transmission, Infection, Diversification and Evolution Group, Max-Planck Institute for the Science of Human History, Jena, Germany
⁴ German COVID Omics Initiative (deCOI)
* these authors contributed equally to this work
+ smithm@rki.de
+ kleistm@rki.de

ABSTRACT

By May 2021, over 160 million SARS-CoV-2 diagnoses have been reported worldwide. Yet, the true number of infections is unknown and believed to exceed the reported numbers by several fold. National testing policies, in particular, can strongly affect the proportion of undetected cases.

Here, we propose a novel method (GInPipe) that reconstructs SARS-CoV-2 incidence profiles within minutes, solely from publicly available, time-stamped viral genomes. We validated GInPipe against in silico generated outbreak data and elaborate phylodynamic analyses. We apply the method to reconstruct incidence histories from sequence data for Denmark, Scotland, Switzerland, and Victoria (Australia). GInPipe reconstructs the different pandemic waves robustly and remarkably accurate. We demonstrate how the method can be used to investigate the effects of changing testing policies on the probability to diagnose and report infected individuals. Specifically, we find that under-reporting was highest in mid 2020 in parts of Europe, coinciding with changes towards more liberal testing policies at times of low testing capacities.

Due to the increased use of real-time sequencing, it is envisaged that GInPipe can complement established surveillance tools to monitor the SARS-CoV-2 pandemic. We anticipate that the method is particularly useful in settings where diagnostic and reporting infrastructures are insufficient. In ‘post-pandemic’ times, when diagnostic efforts are decreased, GInPipe may facilitate the detection of hidden infection dynamics.

Introduction

As of May 2021, the global SARS-CoV-2 pandemic is still ongoing in most parts of the world, with 160 million reported cases worldwide. Novel vaccines of high efficacy have been developed within a year of the outbreak [2, 46]. At the time of writing, approximately 8.2% of the world’s population had already received at least one vaccination. However, distribution of vaccines is uneven and achieving global herd immunity may pose an extremely difficult, long-term task [63, 36]. At the same time, novel variants of concern (VOC) have emerged in high prevalence regions [6, 34], which may be able to reinfect individuals [21, 37] and escape vaccine elicited immune responses [33, 66, 45]. For example, Manaus, Brazil, witnessed a massive second wave of infections [51], despite the fact that approx. 80% had already experienced an infection at the onset of the second wave [6].

Because of the evolutionary versatility of SARS-CoV-2 and difficulties in global vaccine distribution, some experts expect that the virus may not be eliminated globally [44]. Even without adaptation to vaccines in the future, it has been postulated that SARS-CoV-2 may surge [24, 50] and surveillance may have to be maintained into the mid 2020s to monitor virus spread and evolution [24].

Currently, the gold standard of SARS-CoV-2 surveillance is diagnostic testing via polymerase chain reaction (PCR) or antigen-based rapid diagnostic testing (RDT). Diagnostic test results currently define infection case reports, which are used to survey
epidemiological dynamics and to define thresholds for travel bans and non-pharmaceutical measures. Inevitably, case reporting data is affected by test coverage, which changes when testing policies are adapted. While RDT enables point-of-care diagnosis and is less costly than PCR testing [13, 12], gathering and reporting of test results still requires a sophisticated infrastructure, which is difficult to establish and maintain in many developing countries [35]. Independent and complimentary sources of information, such as social media reports [31, 53] or waste water analysis [9, 43] have been used early on to complement our knowledge of the pandemic dynamics. In addition, many regions of the world sequence SARS-CoV-2 genomes to track virus evolution and the emergence of variants of concern. The gathered viral sequences are regularly provided to public databases, such as GISAID [14, 54]. We hypothesize that the genetic data alone holds information about the pandemic trajectory. More specifically, we presume that the speed at which SARS-CoV-2 evolves on the population level contains information about the number of individuals who are actively infected.

In the vast majority of cases, SARS-CoV-2 is transmitted within a very short period, only days after infection [30, 17]. The consequence is a well-defined duration of intra-patient evolutionary time before transmission. Thus, the number of infected individuals is correlated to the rate of divergence of the viral population, implicating an ‘evolutionary signal’.

In this article, we introduce the computational pipeline GInPipe, which only uses time-stamped sequencing data, extracts the ‘evolutionary signal’ and reconstructs SARS-CoV-2 incidence histories. The approach builds on recent work by Khatri and Burt [23], who derived a simple function that relates the mean number of mutant origins to the current allele frequency and the mutational input, which is proportional to the effective population size. Herein, due to the short window of transmission, we anticipate that the effective population size may strongly correlate with the incidence of SARS-CoV-2. We adapt the function derived in [23] and embed it into an automatic computational pipeline (GInPipe) that reconstructs the time course of an incidence correlate $\phi$ merely from SARS-CoV-2 genetic data. GInPipe is validated threefold and performs robustly: (i) against \textit{in silico} generated outbreak data, (ii) against phylodynamic analysis and (iii) in comparison with case reporting data. We applied the method to SARS-CoV-2 sequencing data from Denmark, Scotland, Switzerland, and the Australian state Victoria to reconstruct their respective incidence histories. Lastly, we utilize the inferred epidemic trajectories to compute changes in the probability that an infected individual is reported and highlight how this probability is affected by changes in testing policies.

\section*{Results}

\subsection*{Incidence reconstruction}

An outline of GInPipe for SARS-CoV-2 incidence reconstruction is shown in Figure 1A-C. After compiling a set of time-stamped, full-length SARS-CoV-2 genomes, the sequences are placed into temporal bins $b$ (Fig. 1A). For each bin, we compute the number of mutant sequences $m_b$, as well as the number of haplotypes $h_b$. These two inputs are used to infer the incidence correlate $\phi_b$ (Fig. 1B). We then smooth over all $\phi_b$ point estimates and derive a reconstructed incidence history along the time axis (Fig. 1C). The reconstructed incidence histories can then be used as a basis to estimate the effective reproduction number $R_e$, as well as the relative case detection rate as outlined below.

\subsection*{Method validation: \textit{in silico} experiment}

To test whether GInPipe correctly reconstructs incidence histories, we first performed an \textit{in silico} experiment. We considered a population of $N(t)$ infected individuals at time $t$ that stochastically generate $N(t+1)$ infected individuals in the next time step $t+1$. Each individual is associated with a virus sequence, which can mutate randomly. Individuals can be removed (the associated sequence is removed), or they transmit their virus (the associated virus is copied over). We record the number of infected individuals per generation, as well as all sequences of the currently circulating viruses. We then use the simulated viral sequences to infer $\phi(t)$ and reconstruct the incidence history, as presented in Figure 1D-E.

In Figure 1D, we compare one trajectory of simulated population sizes with the reconstructed incidence histories. The simulated outbreak (red line, right axis) consists of two waves of increasing magnitude. GInPipe reconstructs these dynamics (blue lines and dots, left axis) quite accurately, although the incidence correlate $\phi(t)$ is on a different scale, implying a linear correlation to the number of infected individuals. To assess this correlation, we performed 10 stochastic simulations and compared the $\phi(t)$ point estimates with the corresponding number of infected individuals (Fig. 1E). We observed a strong ($r = 0.96$) and highly significant ($p < 10^{-16}$) linear relationship between the number of infected individuals $N(t)$ and the method’s incidence correlate $\phi(t)$.

While these simulations represent idealized scenarios, we evaluated the robustness of GInPipe with regards to incomplete, and sparse data sets, thoroughly elaborated in Supplementary Note 1.

Our analyses showed, that the method can still accurately reconstruct incidence histories over time, when data is missing or when data sampling is unbalanced. In scenarios of extreme under-sampling, the $\phi$ point estimates are prone to slight underestimation. However, through the smoothing step the reconstructed incidence trajectories still follow the overall population dynamics (Suppl. Note 1, section SN.1.7). Finally, we evaluated whether introductions of foreign sequences affect the reconstruction of incidence histories. Even for extreme and unrealistic cases, a stable reconstruction of the underlying dynamic is possible, but
**A** Re-sampling of sequence sets  

**B** Determination of $\phi_b$  

**C** Smoothing  

**D** Reconstruction of a simulated outbreak  

**E** Simulated incidence vs. reconstruction

---

**Figure 1. Reconstruction of incidence histories using the proposed method.**  
A–C Schematic of the incidence reconstruction method.  
A The sequences are chronologically ordered by collection date. The line shows the cumulative sum of sequences over time. The sequences are allocated into temporal bins, spanning either the same time frame $\Delta t_b$ (yellow and purple bins) or containing the same amount of sequences (green bins).  
B For each bin, the number of distinct variants $h_b$, as well as the total amount of mutant sequences $m_b$ are used to infer the incidence correlate $\phi_b$.  
C The point estimates for all bins $\phi_b$ (dots) are smoothed with a convolution filter. For uncertainty estimation, the point estimates are sub-sampled and interpolated.  
D–E Reconstruction of a simulated outbreak with GlnPipe.  
D $\phi$ estimates resemble the underlying population dynamics over time. The blue line shows the smoothed median of the sub-sampled $\phi$ estimates (dots) for a simulated outbreak. The red line indicates true incidence per generation.  
E Dotplot showing the true outbreak size from the simulation $N_{\text{true}}$ versus the $\phi_b$ point estimates for 10 stochastic simulations. The red line depicts the linear fit.

we do observe a slight tendency of overestimation in these extreme cases (Suppl. Note 1, section SN.1.8).

**Method validation: phylodynamics**

Phylogenetic methods combine phylogeny reconstruction with epidemic models. For example, the piecewise constant birth-death sampling process [55] implemented in BEAST2 [5], allows the reconstruction of the effective reproduction numbers $R_e(\tau)$ for given time periods $\tau$. However, these methods are computationally expensive, so that only moderately sized sequence sets can be used, and advanced knowledge is required to apply them properly to larger data sets. We conducted phylogenetic analyses of SARS-CoV-2 sequence data from Denmark, Scotland, Switzerland, and the Australian state Victoria. In analyzing the data we assumed that $R_e^{\text{BEAST}}(\tau)$ was piecewise constant in between major changes in SARS-CoV-2 non-pharmaceutical interventions (intervals stated in Supplementary Note 2). We then used BEAST2 to estimate
We used GlnPipe to reconstruct complete incidence histories for Denmark, Scotland, Switzerland, and Victoria (Australia) when comparing the piecewise constant \( R_e^\text{median} \) which may explain why the waves deduced from the reporting data, although there are quantitative differences between the reconstructed and reported \( R_e \) sequences) and in 7 minutes on the Victoria data set (n = 10.710 sequences) on a single notebook (2.3 Ghz, 2 cores). (Supplementary Note 2, Figure SN.12 therein, tree B.1). In the daily estimates, we see a transition from \( R_e \) value differs in magnitude between the two methods, with BEAST2 estimating a much lower value. The lower estimate from the BEAST2-analysis in the first interval may be explained by the approximation of transmission clusters, which results in the reconstruction of a relatively high number of transmission events many of which may have occurred outside Switzerland (Supplementary Note 2, Figure SN.12 therein, tree B.1). In the daily estimates, we see a transition from \( R_e^\text{BEAST} > 1 \) to \( R_e < 1 \), which may explain why the median prediction with GlnPipe is close to one for the entire interval. The estimates are qualitatively different for the second interval (mid May – mid June), where GlnPipe estimates \( R_e^\text{BEAST} < 1 \), while BEAST2 estimates \( R_e^\text{BEAST} \approx 1 \). Again, GlnPipe estimates a peak in summer (mid June-mid August \( R_e^\phi > 1 \)). While BEAST2 predicts the onset of transmission in the second wave to already start in mid August \( R_e > 1 \), GlnPipe estimates the first major rise in infections at the end of September.

For Switzerland, the estimates disagree slightly, particularly in the first interval (mid March to mid May), which spans both sides of the peak number of infections during the first wave. Although both methods predict a median \( R_e^\text{BEAST} < 1 \), the absolute value differs in magnitude between the two methods, with BEAST2 estimating a much lower value. The lower estimate from the BEAST2-analysis in the first interval may be explained by the approximation of transmission clusters, which results in the reconstruction of a relatively high number of transmission events many of which may have occurred outside Switzerland (Supplementary Note 2, Figure SN.12 therein, tree B.1). In the daily estimates, we see a transition from \( R_e^\text{BEAST} > 1 \) to \( R_e < 1 \), which may explain why the median prediction with GlnPipe is close to one for the entire interval. The estimates are qualitatively different for the second interval (mid May – mid June), where GlnPipe estimates \( R_e^\text{BEAST} < 1 \), while BEAST2 estimates \( R_e^\text{BEAST} \approx 1 \). Again, GlnPipe estimates a peak in summer (mid June-mid August \( R_e^\phi > 1 \)). While BEAST2 predicts the onset of transmission in the second wave to already start in mid August \( R_e > 1 \), GlnPipe estimates the first major rise in infections at the end of September.

For Victoria we observe an \( R_e^\text{BEAST} > 1 \) until mid March in the daily estimates. Overall, \( R_e \) is less than 1 for the first interval between mid March and May, versus \( R_e > 1 \) between June and August. Again, we see various peaks around June and July in the daily \( R_e \) estimates with the proposed method. For the final interval, both methods slightly disagree, with \( R_e^\text{BEAST} < 1 \) and \( R_e^\text{BEAST} > 1 \), though the daily \( R_e^\phi \) are decreasing towards the end of the final interval.

In terms of computational time, the entire GlnPipe analysis pipeline runs in 20 minutes on the full Denmark data set (n = 40.575 sequences) and in 7 minutes on the Victoria data set (n = 10.710 sequences) on a single notebook (2.3 Ghz, 2 cores). Furthermore, GlnPipe does not require to pre-assign any intervals, to exclude particular strains, construct a phylogenetic tree, or cluster sequences based on a their phylogenetic relationship. The BEAST2 analysis alone required about 15 hours on an Intel Xeon E5-2687W (3.1 Ghz, 2 x 12 cores) on a sub-sampled data set (n≈ 2500 sequences) with additional computation time needed to construct a multiple sequence alignment and approximate transmission clusters.

**Reconstructed incidence histories**

We used GlnPipe to reconstruct complete incidence histories for Denmark, Scotland, Switzerland, and Victoria (Australia) from publicly available full length SARS-CoV-2 sequencing data provided through GISAID [14, 54] (Supplementary Note 4). In Figure 3, we compare the reconstructed incidence histories (blue lines and dots, left axis) to the 7-day rolling average of officially reported new cases (red line, right axis). Overall, the reconstructed incidence estimates reflect the different pandemic waves deduced from the reporting data, although there are quantitative differences between the reconstructed and reported incidence trajectories over time. In particular, during the first wave in Scotland, and Victoria (Fig. 3B.D) our method estimates higher incidences than reported, whereas the curves align at later points for the second and third wave. It is worth mentioning that testing capacities were particularly low in Scotland in April (during the first wave), suggesting extensive under-reporting in the initial phase of the pandemic. This is also supported by test positive rates of almost 40% during April 2020 in Scotland (Supplementary Fig. 1). In Victoria, sufficient testing capacities were not available until May, but test positive rates were already declining from April to May (Supplementary Fig. 1). This indicates that the first wave may have been under-reported in magnitude, but had vanished by May.

Interestingly, the proposed incidence reconstruction method predicts small summer waves in August in the three European countries (Fig. 3A–C) that are not visible in the reporting data. In the incidence reconstruction method these ‘summer waves’
Figure 2. Effective reproduction number $R_e$ estimates using the proposed method ($\phi$) and phylodynamics (BEAST2). Piecewise constant $R_e^{BEAST}(\tau)$ estimates (green solid lines) where calculated using the BDSKY model for the indicated intervals, as described in the Methods section. Daily estimates $R_e^\phi(t)$ (blue dots) were directly calculated from the incidence correlates $\phi$ using the Wallinga-Teunis method [61]. The median of these values for the indicated intervals $R_e^\phi(\tau)$ is shown as solid blue lines. The 95% confidence interval is specified by the shaded areas. Justifications of the intervals are found in Supplementary Note 2.

are immediately followed by the second SARS-CoV-2 wave. For the second wave, reconstructed incidence histories correspond to the reported cases, particularly in Denmark, Scotland, and Victoria. (Fig. 3A-B & D). For Scotland, our method predicts a more long-lasting third wave with rising incidence rates until February 2021 and a moderate decline with several smaller peaks until May, whereas the reporting data indicates a peak in January 2021 with a subsequent fast regression. The argument, that ongoing vaccination in Great Britain could explain the immediate decline of reported infected cases, can be objected with the fact, that by March 2021 (end of the prediction horizon) only about 2% of the Scottish population were fully vaccinated. For Switzerland, we predict a larger wave around January-February 2021 (third wave) that is not reflected in the reporting data. Towards the end of the prediction horizon, from March 2021 onwards, the reported cases and the incidence estimation both indicate a rise in numbers (fourth wave).

Relative case detection rate
We investigated whether the proposed incidence reconstruction method may be used to learn about the proportion of infected cases that are actually tested, detected and reported, $P_I(\text{tested|infected})$. The proportion of SARS-CoV-2 infected who are actually reported can be calculated using Bayes' formula (see Methods section). In order to perform the calculation, the proportion of actively infected individuals in the population $P_I(\text{infected})$ needs to be known. We have shown that the incidence correlates $\phi$ from our method are proportional to the number of infected individuals, $c \cdot \phi = N_{\text{eff}}$ (Fig. 1D–E, Fig. 3), and hence to the probability of being infected $P_I(\text{infected})$. Consequently, we may use the reconstructed incidence profiles, together with the test sensitivity and specificity, the respective information about the proportion of positive tests, as well as the testing capacities for each country or region to calculate changes in the case detection rate, scaled by unknown factor $c$.

In Figure 4, we show the log$_2$ scaled detection probabilities for Denmark, Scotland, Switzerland, and Victoria (Australia).
Figure 3. Incidence reconstruction based on sequencing data. The graphic depicts the genome-based incidence reconstruction (in blue) using the proposed method (left axis) vs. the 7 days rolling average of newly reported cases in red (right axis). Blue dots depict φ̂ point estimates of the incidence correlate, where the size of the dot is related to the number of sequences used to infer φ̂. The solid and dashed blue lines denote the median smoothed trajectories and their 5th and 95th percentiles. The black markers on the x-axis depict the collected sequences at the given dates. A Denmark (n = 40,575 sequences) B Scotland (n = 30,258 sequences) C Switzerland (n = 25,779 sequences) D Victoria (n = 10,710 sequences)

The log scaling allows us to easily gauge the relative change in (under-)detection of the infected population over time (e.g. 2-fold, 4-fold increase or decrease in case detection rate). The dashed vertical lines in the graphics indicate major changes in testing policies in the respective countries. Individual parameters used in the inference procedure, \( P(\text{tested}) \), \( P(\text{inf} | \text{tested}) \), and \( c \cdot P(\text{infected}) \) are shown in Supplementary Figure 1.

For Denmark, we observe an initial period of massive SARS-CoV-2 under-detection in the beginning of March 2020, Fig. 4A (upper panel), which coincides with very low testing capacities at the beginning of the pandemic (Fig. 4A, lower panel). From mid March, case detection stabilizes at a 6-fold higher level, compared to the first week of March. The second interval begins around mid May with an important policy change, allowing every citizen to get tested without medical referral. Interestingly, compared to the fairly stable case detection levels from mid March to mid May, this policy change leads to a 2-3 fold drop in case detection in the summer months from July-September. Of note, while everybody is granted the possibility to test for SARS-CoV-2, testing capacities remained fairly unchanged (Fig. 4A, lower panel). According to our calculations, the largest proportion of infections remained undetected in July. From end of August, testing capacities were steadily increased in Denmark (Fig. 4A, lower panel), particularly in Copenhagen and at the airports, followed by prioritized testing. From September on, this leads to a nearly 8-fold increase of the case detection rate, with a peak in December. From end of December the detection rate drops more than 4-fold, despite continuous testing.

For Scotland (Fig. 4B), the earliest test data is available only from the end of March. Therefore, the data captures only the second part of the first wave, compare Fig. 3B. In the beginning of May, testing capacities were more than doubled (Fig. 3B, lower panel) and outbreak investigation intensified. This led to a doubling of the relative case detection rate from May, compared to the first phase. On 18 May, SARS-CoV-2 testing was opened for everyone with symptoms. However, only in July testing capacities were increased. This may have led to a drop in case detection from mid May to July, after which case detection increased and remained during August at roughly the levels achieved in May. After 25 August, testing capacities...
and accessibility of testing steadily increased. Accordingly, case detection increased about 6-fold until winter 2020/21. From 25 November, testing capacities were further expanded, especially in the health sector, including hospital patients, health and social care staff, with fairly stable case detection rates. Further increase of testing capacities in the end of December allowed to double the probability to detect infected individuals. From beginning of the year 2021, the Scottish government pushed community testing in areas with high SARS-CoV-2 prevalence. At the same time, the proportion of positive tests start to decline (Suppl. Fig. 1), and consequently the case detection rate collapses until April by 9-fold.

Similar to Denmark, Switzerland shows an initial period of massive SARS-CoV-2 under-detection in the beginning of March 2020 (Fig. 4C, upper panel), which coincides with very low testing capacities at the beginning of the pandemic (Fig. 4C, lower panel). When testing capacities increase by mid March, case detection rates grow 8-fold. However, from beginning of April, we observe drop in the probability to detect infections that lasts until mid May (overall 10-fold drop). This trend coincides with a drop of positivity rates (Supplementary Figure 1), as well as the extension of testing criteria on 22nd April: From this date, anybody with symptoms was allowed to get tested, despite the fact that the availability of tests was not increased (Fig. 4C, lower panel). From 18th May, tests were partly prioritized for hospitalized and vulnerable individuals. At the same time, testing capacities steadily increased and incidences dropped. As a net effect, the probability of detecting infected people increases steadily to a maximum at the end of October with a relative difference of nearly 20-fold compared to the low point in mid May. On 2 November, Switzerland begins to supply antigen-based rapid diagnostic tests (RDT) for self-testing as part of their COVID containment strategy. Interestingly, our model predicts that this led to a sharp decline in case detection, again corresponding with the decline in positivity rates (Supplementary Figure 1). From 21st February 2021, further precautionary actions were taken, and the government recommended repeated testing. This is associated with a stable, but relatively low detection rate for infected people until end of April 2021.
For the Australian state Victoria, the earliest data were available from end of March 2020, Fig. 4D. Detection probabilities in the first interval, until 14th April were changed proportionally to the test capacities during that interval Fig. 4D (upper and lower panel). On 14th April 2020, the testing criteria were expanded, allowing anyone with COVID-like symptoms to be tested. Unlike the situation in Switzerland, where we observed a downward trend in case detection after expanding the testing criteria (Fig. 4C), the detection probability in Victoria remains stable until end of April. In contrast to Switzerland, testing capacities were increased when testing criteria were expanded. On 30th April, the government initiated a two-week ‘testing blitz’, a large, coordinated testing campaign to locate viral spread. The ‘testing blitz’ was accompanied by mass sewerage testing and matched with a massive increase of testing capacities, which led, according to our simulations, to a 4-fold increase in the probability to detect infected individuals. At the end of the ‘testing blitz’, testing capacities steadily decreased and the proportion of detected infections decreased drastically (by roughly 9-fold). At the beginning of June, testing capacities rose again, matched by a rise in the proportion of detected cases. From 1st July onwards, several ‘testing blitzes’ were conducted in outbreak regions, which seemed to have stabilized case detection rates during the second wave of infections. After the second wave (end of August–September, Fig. 3D), case detection rates drop. From October 2020 onwards, our predictions become highly unreliable, as the incidence estimates credibility interval includes zero (compare Fig. 3D), which concludes that the case detection rate cannot be determined anymore.

In general, we make two striking observations: Firstly, and quite intuitively, whenever more tests were conducted, the proportion of detected SARS-CoV-2 cases increases. Secondly, and unexpectedly, whenever testing criteria were relaxed, this led to a drop in the probability of case detection. We see this drop in mid May in Denmark and Scotland and in mid April in Switzerland. Importantly, the expansions of testing criteria were not-, or insufficiently matched by increased testing capacities. Quite surprisingly, our simulations for Switzerland suggested a drop in case detection when antigen-based RDT self-testing became part of the national diagnostic strategies.

**Discussion**

SARS-CoV-2 continues to spread around the world, making epidemiological and molecular surveillance indispensable for the evaluation and guidance of public health interventions. Many national and international sequencing efforts are underway that closely monitor the dynamics and evolution of the virus. In the global fight against SARS-CoV-2, the vast majority of reconstructed sequence data has been made broadly available through public databases, such as GISAID [14, 54] and the COVID data portal. In this work, we introduce GInPipe, a pipeline that utilizes this data to reconstruct SARS-CoV-2 incidence histories. Viral infections are often characterized by a transmission bottleneck [32], where only a very small number of viruses initiate the infection and subsequently replicate within the host. A sufficient number of viruses (viral load) is required for further transmission. Hence, the temporal window of infectiousness begins with the intra-host viral population reaching a sufficiently large abundance and ends with the virus becoming eliminated by the immune system (or drugs). In contrast to HIV or HCV, SARS-CoV-2 is almost always transmitted within days after infection [30, 17]. If neutral or favourable mutations occur during this time, they may become abundant enough to be passed on to other hosts [32]. The consequence is a well-defined duration of intra-patient evolutionary time in which the virus can randomly mutate and become transmitted subsequently. In SARS-CoV-2, this intra-patient evolutionary time appears to be short and the analysis of outbreak clusters indicates that the virus genomes from linked cases were separated by either none, or very few mutations [4, 18, 52]. Taken together, these lines of evidence suggest that evolutionary change of SARS-CoV-2, the effective viral population size, and the number of infected people are correlated.

In the past, numerous approaches have been published, with the aim to estimate the effective population size from genetic properties (reviewed in [62, 38]). A variety of methods utilize the information of temporal changes in allele frequency (reviewed in [62]), while others build on population genetic theory and phylodynamic reconstruction [16, 59, 27]. GInPipe is rather related to the first class of methods as it adapts recent works of Khatri and Burt, 2019 [23]. Essentially, GlnPipe considers snap-shots of inter-patient evolution to estimate a mutational input parameter φ(t). The latter is proportional to the effective population size, which correlates with incidence. Taken together, GlnPipe uses time-stamped SARS-CoV-2 sequences and divides them into bins of inter-patient virus evolution to estimate time-dependent incidence correlates φb. From the set of φb estimates, the entire incidence history φ(t) can be reconstructed.

We assessed the suitability of GlnPipe using in silico simulated outbreaks, in comparison with phylodynamics and by comparing to reported case statistics. Using simulated data, the method accurately reconstructed incidence histories (Fig. 1). It also performed robustly with incomplete data, and when foreign sequence variants were introduced (Supplementary Note 1). The method even worked when the introduced variants made up a considerable fraction of the population and did not contribute to the mutational input of the outbreak.

We also compared the method with epidemiological estimates from phylodynamic reconstruction using BDSKY [55] in BEAST2 [5], shown in Figure 2. Bayesian phylodynamic methods use Monte Carlo Markov Chain (MCMC) or similar
techniques to allow for a Bayesian estimation of phylogenetic relatedness of genomes, by both estimating evolutionary parameters, as well as parameters governing an underlying epidemiological model \[64, 60\]. The MCMC sampling procedure makes phylodynamic inference computationally demanding and often requires to ‘down-sample’ data sets. When the epidemiological model entails time-varying parameters, changes in the effective reproduction numbers \(R_e (\tau)\) can be computed. However, to enable their estimation (practical parameter identifiability), parameters of the underlying epidemiological model are typically considered to be piecewise constant or to change smoothly. In Figure 2, we show the phylodynamic estimates of the effective reproduction numbers \(R^\text{BEAST} (\tau)\). Corresponding reproductive numbers \(R^\phi (\tau)\) were computed with GlnPipe by applying the method of Wallinga-Teunis \[61\] to the estimated incidence correlates \(\phi\). We compared the medians over the temporal windows used in the phylodynamic analysis. Overall, this methodological comparison yielded highly congruent predictions, with the exception of Switzerland in the first- (mid March – May 2020) and final intervals (mid September 2020 – January 2021). The ETH Zurich provides a visualization\(^1\) for the daily \(R_e\) estimates, based on reporting data. The ETH data, similarly to our daily \(R^\phi\) estimates with GlnPipe, shows a peak, followed by a decline in the daily \(R_e\) for the first interval. This could explain why the median \(R^\phi\) is only slightly smaller than 1 in this first interval, unlike the BEAST2 estimate, which is \(\approx 0.6\). For the final intervals (mid September 2020 – January 2021) \(R^\phi\) estimates fluctuate around- or slightly above \(R_e (\tau) = 1\), in line with the predictions of the ETH, and slightly below the BEAST2 estimate that resulted in a median \(R_e\) around 1.2. For the sake of this comparison, a relatively crude transmission cluster detection method was employed for the phylodynamic analyses, which may be causing a slight bias in the estimated effective reproduction numbers. Overall, it appears that both methods yield similar results with respect to inferring the pandemic trajectories in the majority of cases. The power of GlnPipe lies in the swift reconstruction of incidence histories with a fine temporal resolution, without requiring phylogenetic inference, construction of a multiple sequence alignment, down-sampling, clustering by e.g. lineages, or masking of problematic sites in the virus genomes. Moreover, GlnPipe performs robustly, even in case of large proportions of introduced variants, which would also include lab-specific errors (Supplementary Note 1). However, \(R_e\) estimation is obviously only a side-product of phylodynamic inference, which has many more applications such as the identification and analysis of transmission clusters \[19, 47\], which GlnPipe is not suited for. Hence, the two approaches could complement one another. To simplify the use of GlnPipe, we provide an automatic workflow that can be directly applied to data downloads from GISAID or the COVID Data Portal. The execution time appears to scale linearly with the number of sequences to be analyzed (\(\approx 1,500\) sequences per minute on a 2.3 Ghz computer with 2 cores). When we applied GlnPipe to available GISAID data from Denmark, Scotland, Switzerland, and Victoria (Australia), we observed that the reconstructed incidence histories agree well with the daily numbers of new reported infections (Fig. 3). Particularly for Denmark, reconstructed incidence histories match the reporting data quite well. Of the analyzed countries, Denmark conducted the largest number of SARS-CoV-2 tests per capita (see also \(P(\text{tested})\) in Supplementary Figure 1). This could imply that the pandemic was relatively well tracked, as also suggested by relatively small changes in the diagnostic rate (Fig. 4). Moreover, a large fraction of the diagnosed cases were sequenced, providing a comprehensive genomic profile of the virus population. For the first wave in Scotland and Victoria, we determined a much higher incidence than reported. Notably, the number of SARS-CoV-2 tests per capita was very low in Scotland, as well as in Victoria until May 2020 (\(P(\text{tested})\) in Supplementary Figure 1). Thus, a large proportion of infected individuals may not have been diagnosed during this time. In Victoria and Scotland, testing capacities were increased in May, i.e. after the peak of the first wave. Another striking difference of our predictions in comparison to the reported cases is that GlnPipe indicates a rise of infections in August 2020 in all European countries. Notably, this increase in infections coincides with the introduction and community spread of B.1.177 (the ‘Spanish’ variant, 20E (EU)) in most Western European countries as suggested by phylodynamic analyses \[28, 20\]. Our results, when compared with the reported cases, therefore imply an under-reporting of cases during the onset of community transmission of B.1.177. Quantifying case detection is usually not feasible without knowing, or approximating the proportion of infected individuals (compare Eq. (2)). In order to do so, others have used mathematical models to predict the proportion of infected individuals \[6, 1\] and with this, to estimate the level of under-reporting of SARS-CoV-2. However, these mathematical models cannot be fitted to the reported cases under the presumption of an unknown trajectory of under-reporting. It therefore remains extremely difficult to parameterize suitable models for the task of assessing under-reporting, in particular for non-monotonic pandemic trajectories. Random testing may inform the number of incidents, as well as asymptomatic infections \[41\]. Yet, usually only snap-shots of the incidence may be derived, which are insufficient to parameterize the aforementioned models. Moreover, it is not clear, whether the samples in the random testing scheme were representative. Sero-prevalence studies remain the gold-standard to estimate the cumulative number of infections \[6, 1\], as well as cumulative under-detection. Nevertheless, these studies only

---

\(^1\)https://ibz-shiny.ethz.ch/covid-19-re-international/
provide very coarse time resolution (if any) and require large sample sizes for robust analysis. A methodologically related approach uses a semi-Bayesian approach to assess under-detection in the US [65]. To enable estimation, the probability of case detection is constrained by the assumption of particular prior distributions.

With regards to the aforementioned approaches, our method to quantify case detecting profiles has the advantage that no complex mathematical modelling is needed, and no constraints are necessary. Instead, we use information about the conducted tests and the test positive rate, in combination with the incidence correlate $\phi$. This makes the proposed approach simple, interpretable and independent of additional assumptions.

Using this method, we observed that broad testing with little, or no suspicion of SARS-CoV-2 infection coincides with apparent under-reporting of infections from the second quarter of 2020. This coincides with a drastic decrease in the proportion of positive test results. From the latter, it is possible to compute the conditional probability that a tested person is actually infected ($P(\text{inf} | \text{tested})$, Supplementary Figure 1). A drop in $P(\text{inf} | \text{tested})$ coinciding with a steady amount of tests can negatively affect the probability to detect infected individuals $P(\text{tested} | \text{inf})$, which may have happened in the European summer of 2020.

In other words, the scarce testing resources available during that time, may not have been employed in the most effective way. This suggests that it may be advisable to focus on testing symptomatic individuals when testing capacity is low.

Nevertheless, the apparent under-reporting was overcome relatively quickly by either increasing testing capacities (Denmark, Scotland, Victoria) or re-focusing capacities or both (Switzerland), Fig. 4. Interestingly, our method predicts a decline in case detection in Switzerland after the broad introduction of antigen self-testing in November 2020. A potential explanation for this observation is that only a fraction of positive antigen self-tests is confirmed by PCR and hence enters the Swiss reporting system. At the time of writing, the final interpretation of this observation is still unclear and will require further analysis.

In summary, we have developed a method that allows to reconstruct incidence histories solely based on time-stamped genetic sequences of SARS-CoV-2. We implemented the method in a fully automated workflow that can be applied to publicly available data. Moreover, this method can be used to assess the impact of testing strategies on case reporting. Finally, we envision that the method will be particularly useful to estimate the extent of the SARS-CoV-2 pandemic in regions where diagnostic surveillance is insufficient for monitoring, but may still yield a few samples for sequencing. In some of these regions pandemic control may be impossible or cause more harm than benefit [56] and hence these regions may constitute reservoirs for the emergence of novel SARS-CoV-2 variants. Gaining insight in the pandemic dynamics in these regions through alternative methods, such as GInPipe, could yield valuable information that helps to direct global SARS-CoV-2 control efforts.

Methods

Data and data pre-processing

Sequences and meta data for Denmark, Scotland, Switzerland, and Victoria (Australia) were downloaded from the GISAID EpiCoV database [14, 54] (Supplementary Note 4). Sequences, where only the year of collection was provided were omitted. If year and month are specified, the 15th day of the month was added to the meta data. The retained sequences were individually mapped to the reference (NCBI Wuhan Reference Sequence: NC_045512.2 [39]) with minimap2 version 2.17 (r941) [29]. From the mapping files (SAM), we deduced the nucleotide substitutions for each sequence. Point mutations appearing less than three times in the whole data set were filtered out, as they may occur due to sequencing errors [58].

Construction of temporal sequence bins

SARS-CoV-2 sequences were sorted chronologically by collection date and assigned to temporal bins $b$ in a redundant manner. We subdivided the sequence set into bins of

- equal size (proportions of 2%, 5%, 7% of all samples)
- spanning an equal amount of days (10, 15, and 20, and one calendar week).

Bins that contain a proportion of sequences should however span at least 3 days and maximally 21 days, and bins that span a predefined time period should contain at least 15 sequences. The date assigned to a bin is the mean collection date of the comprised sequences.

The redundant binning (‘re-sampling’) allows to evaluate cases where there is insufficient data along the time line (Figure 1A), and makes the proposed method statistically more robust to outliers.

Incidence correlate $\phi_b$

The proposed method is inspired by the work of Khatri and Burt, 2019 [23], who derived a simple relation between the mean number of independent origins of soft selective sweeps in a population sample $T_f$, the current number of an allele $m$ and mutational input, i.e. the scaled (haploid) effective population size $\theta = 2N_{\text{eff}}\mu$, with $\mu$ denoting the mutation rate:
Unlike Khatri and Burt, who aim at estimating the recent effective population size utilizing the recurrent mutations which have been fixated in the population, we seek to reconstruct the history of incidences of a population over time. We adapted the equation accordingly, also under the presumption that the de novo occurrence of mutations is driven by random chance events, whose likelihood may increase with the number of infected individuals [25, 10]. Seeking to estimate the incidence correlate $\phi = c \cdot N_{\text{eff}}$, with the incidence being equivalent to the effective population size $N_{\text{eff}}$, scaled by a constant factor $c$, we parameterize the equation as follows: For each temporal bin $b$ we estimate incidence correlate $\phi_b$ at time $t_b$. From the sequences comprised in bin $b$, i.e. dated within a certain time frame $\Delta t_b$ (Fig 1A), we infer the number of haplotypes $h_b$ and the total number of mutant sequences $m_b$ in the bin (Fig 1B). The mutations are determined with respect to a given reference sequence. In the original equation, we replace the mean number of origins $\eta$ with the number of distinct variants $h_b$. In each temporal bin, however, haplotypes and mutants are accumulated over the time span $\Delta t_b$. To correct for biases that result from this accumulation, especially for large time spans, we normalize the inputs $h_b$ and $m_b$ using a logistic function $w_b = (\log(\sqrt{\Delta t_b}) + 1)^{-1}$.

The parameter $\phi_b$ is derived by numerically solving

$$\phi_b^* = \arg \min_{\phi_b} h_b \cdot w_b - \phi_b \log \left( 1 + \frac{m_b \cdot w_b}{\phi_b} \right).$$

Reconstructing the incidence history

Incidence point estimates $\phi_b$ are assigned to the mean collection date $t_b$ of the sequences contained in the bin. We applied a convolution filter with window size 7 days to derive a continuous, smoothed trajectory (Fig. 1C). For uncertainty estimation, we sub-sampled $\phi$ trajectories 1000 times, by randomly leaving out 50% of the point estimates and reconstructed the trajectory by smoothing and linear interpolation between the remaining point estimates.

Implementation and availability

All methods were implemented in Python version 3.9 and R version 4.0. A fully automated workflow has been generated using Snakemake [26] and is available from https://github.com/KleistLab/GInPipe.

Simulation study

To test the proposed incidence reconstruction method, we stochastically simulated the evolutionary dynamics of a viral outbreak using a Poisson process formalism. We started with $N(t_0) = 50$ copies of a random sequence of length $L = 200$ nt, that evolved in 120 discrete time steps, depending on a population dynamic. A succeeding generation was modelled to consist of $N(t+1) \sim \text{Poiss}(N(t) \cdot \rho(t))$ sequences (= effective population size), where we chose a sinusodial rate $\rho(t) = \frac{\sin(t \cdot 0.11)}{15} + 1.03$. Thus, $N(t+1)$ sequences from the actual generation were randomly chosen with replacement and copied over to the next generation. We then introduced $n_{\text{mut}} \sim \text{Poiss}(\mu \cdot N(t+1) \cdot L)$ random mutations into these sequences with per site mutation rate $\mu = 0.0001$.

For each generation, a fasta file with all sequences was stored and used as input for the incidence reconstruction pipeline. We ran 10 stochastic simulations with the settings stated above to compare the ground truth effective population sizes $N(t)$ from our simulations with the corresponding inferred incidence trajectories $\phi$.

In Supplementary Note 1, we evaluated scenarios where only a fraction of the sequences were sampled (10-90%) or, to rule out sampling biases, we sub-sampled equal amounts of sequences at each time point, independent of $N(t)$. Moreover, we assessed whether our predictions were affected by the introduction of unrelated sequence variants into the population.

Effective reproduction number $R_e$

Based on the reconstructed incidence histories, the effective reproduction number $R_e(t)$ was computed using the established method by Wallinga and Teunis [61] (R package R0 [40]). Daily estimates of $\phi$ were assigned a pseudocount of one and rounded to the nearest integer. For the generation time distribution $g(t)$ of SARS-CoV-2, we chose the Gamma distribution with a mean of 5 days and a standard deviation of 1 day [15, 7].

Phylogenetic analyses

Phylogenetic analyses were performed on subsampled sets of the data described above (Data and data pre-processing) using a birth-death-sampling process as implemented in the BDSKY [55] model in BEAST2 [5]. Here the precise collection day of sequence samples with only information on year and month was inferred during the analysis and not a priori set to the 15th. The full data sets were first grouped by Pango lineage [48, 8] and then subsampled by randomly selecting a specific percentage of sequences per week (Victoria: 10% for lineage D.2, 50% for other lineages; Switzerland: 50% for all lineages; Scotland: 20% for all lineages; Denmark: 5% for all lineages). In addition, sequences were excluded if they belonged to a lineage with...
less than two representatives in the analyzed set and lineages with periods longer than 75 days without any sample were split into parts. Retained sequences were aligned to the reference genome (Genbank-ID MN908947.3 [3]) in MAFFT [22] using the --keeplength option and problematic sites were masked by replacing the them with ‘N’ in the alignment [11].

For each approximate cluster a separate phylogeny was reconstructed. A strict clock model with a fixed rate of $8 \cdot 10^{-4}$ substitutions per site per year and an HKY substitution model were used. In the embedded transmission model, transmission ($\lambda$), recovery ($\mu$) and sampling ($\psi$) rates were assumed to be piecewise constant with changes allowed either when intervention measures changed, or in a uniform manner (Supplementary Note 2). The reproductive number $R_e(t) = \lambda(t)/(\mu(t) + \psi(t))$ was drawn from a log-normal distribution $R_e(t) \sim \log \mathcal{N}(0, 4)$, the rate to become non-infectious $\delta(t) = \mu(t) + \psi(t)$ from a narrow normal distribution with $\delta(t) \sim \mathcal{N}(27.11, 1)$ which is changed to $\mathcal{N}(48.8, 1)$ after first control measures are implemented in the respective area. The sampling proportion $s(t) = \psi(t)/(\psi(t) + \mu(t))$ was a priori assumed to arise from a uniform distribution with a lower limit of zero and the upper limit determined by the ratio of analyzed sequences over diagnosed cases $s \sim U(0, q_i/d_i)$ where $d_i$ is the number of diagnoses and $q_i$ the number of sequences included in the analysis in interval $i$. To account for the lineage specific subsampling, a separate sampling proportion for lineage D.2, $s_{D.2}$, was modelled in the analysis of the Victoria data. A uniform distribution with an upper limit corresponding to the subsampling percentage was thus used as prior distribution of the D.2 specific-, as well as general sampling proportion $s_g$, i.e. $s_{D.2} \sim U(0, 0.1)$ and $s_g \sim U(0, 0.5)$.

Setup files for all four analyses can be found as Supplementary Files.

MCMC chains were run until all parameters converged, which took about 300 million steps for analyses of data from Denmark, Scotland and Switzerland. Because of the large D.2 cluster consisting of more than 900 sequences, about 750 million steps were needed for convergence using data from Victoria. On an Intel Xeon CPU E5-2687W (3.1 Ghz; 2 x 12 cores), this corresponded to about 15 hours to run one analysis for at least 300 million MCMC steps (about 3 min/Msample). Log files were assessed using Tracer [49] and are included as Supplementary Files. TreeAnnotator was used to summarize the posterior sample of phylogenetic trees to a maximum clade credibility tree using median node heights. Lineage through time plots of all summary trees were calculated using the R package ape [42] and are shown in Supplementary Note 2.

Relative case detection rate

We used GlnPipe to detect changes in SARS-CoV-2 case detection. Let us denote by $P_{A.W.}$, $Y.D$. Writing - Original Draft, M.R.S., M.T., A.W. and M.v.K.; Writing-Review and Editing, M.R.S., M.T., A.W., D.K. and M.v.K.; Funding Acquisition, A.W., D.K. and M.v.K.; Supervision, D.K. and M.v.K.; Funding Acquisition, A.W., D.K. and M.v.K.; Supervision, D.K. and M.v.K.; Funding Acquisition, A.W., D.K. and M.v.K.; Supervision, D.K. and M.v.K.;
Funding
M.R.S., M.T., Y.D. and MvK acknowledge funding from the Germany ministry for science and education (BMBF; grant numbers 01KI2016 and 031L0176A). D.K. and A.W. acknowledge funding from the Max Planck Society. A.W. acknowledges financial support through a scholarship (Landesgraduiertenstipendium), funded by the State of Thuringia, Germany. The funders had no role in designing the research or the decision to publish.

Conflicts of interest
The authors declare that no conflicts of interest exist.

References
[1] Frederick J Angulo, Lyn Finelli, and David L Swerdlow. “Estimation of US SARS-CoV-2 infections, symptomatic infections, hospitalizations, and deaths using seroprevalence surveys”. In: JAMA network open 4.1 (2021), e2033706–e2033706.
[2] Lindsey R Baden et al. “Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine”. In: New England Journal of Medicine 384.5 (2021), pp. 403–416.
[3] Dennis A Benson et al. “GenBank”. In: Nucleic acids research 41.D1 (2012), pp. D36–D42.
[4] Merle M Böhmer et al. “Investigation of a COVID-19 outbreak in Germany resulting from a single travel-associated primary case: a case series”. In: The Lancet Infectious Diseases 20.8 (2020), pp. 920–928.
[5] Remco Bouckaert et al. “BEAST 2: a software platform for Bayesian evolutionary analysis”. In: PLoS Comput Biol 10.4 (2014), e1003537.
[6] Lewis F Buss et al. “Three-quarters attack rate of SARS-CoV-2 in the Brazilian Amazon during a largely unmitigated epidemic”. In: Science 371.6526 (2021), pp. 288–292.
[7] Robert Challen et al. “Meta-analysis of the SARS-CoV-2 serial interval and the impact of parameter uncertainty on the COVID-19 reproduction number”. In: medRxiv (2020). DOI: 10.1101/2020.11.17.20231548.
[8] cov-lineages/pangolin. Oct. 2020. URL: https://github.com/cov-lineages/pangolin.
[9] Christian G Daughton. “Wastewater surveillance for population-wide Covid-19: the present and future”. In: Science of the Total Environment 736 (2020), p. 139631.
[10] Troy Day et al. “On the evolutionary epidemiology of SARS-CoV-2”. In: Current Biology 30.15 (2020), R849–R857.
[11] Nicola De Maio, Conor Walker, and Rui Borges. Issues with SARS-CoV-2 Sequencing Data - SARS-CoV-2 Coronavirus / nCoV-2019 Genomic Epidemiology. Virological. 2020. URL: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (visited on 05/11/2021).
[12] Andreas Deckert et al. “Effectiveness and cost-effectiveness of four different strategies for SARS-CoV-2 surveillance in the general population (CoV-Surv Study): a structured summary of a study protocol for a cluster-randomised, two-factorial controlled trial”. In: Trials 22.1 (2021), pp. 1–4.
[13] Jacqueline Dinnes et al. “Rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection”. In: Cochrane Database of Systematic Reviews 3 (2021), p. CD013705.
[14] Stefan Elbe and Gemma Buckland-Merrett. “Data, disease and diplomacy: GISAID’s innovative contribution to global health”. In: Global Challenges 1.1 (2017), pp. 33–46.
[15] Luca Ferretti et al. “Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing”. In: Science 368.6491 (2020). DOI: 10.1126/science.abb6936.
[16] Simon DW Frost and Erik M Volz. “Viral phylodynamics and the search for an ‘effective number of infections’”. In: Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365.1548 (2010), pp. 1879–1890.
[17] John Griffin et al. “Rapid review of available evidence on the serial interval and generation time of COVID-19”. In: BMJ open 10.11 (2020), e040263.
[18] Thomas Günther et al. “SARS-CoV-2 outbreak investigation in a German meat processing plant”. In: EMBO molecular medicine 12.12 (2020), e13296.
[19] Kirsten Hanke et al. “Reconstruction of the genetic history and the current spread of HIV-1 subtype A in Germany”. In: Journal of virology 93.12 (2019).
[20] Emma B Hodcroft et al. “Emergence and Spread of a SARS-CoV-2 Variant through Europe in the Summer of 2020”. In: medRxiv (2020). DOI: 10.1101/2020.10.25.20219063.

[21] Markus Hoffmann et al. “SARS-CoV-2 Variants B.1.351 and P.1 Escape from Neutralizing Antibodies”. In: Cell 184.9 (2021), 2384–2393.e12.

[22] Kazutaka Katoh and Daron M Standley. “MAFFT multiple sequence alignment software version 7: improvements in performance and usability”. In: Molecular biology and evolution 30.4 (2013), pp. 772–780.

[23] Bhavin S Khatri and Austin Burt. “Robust estimation of recent effective population size from number of independent origins in soft sweeps”. In: Molecular biology and evolution 36.9 (2019), pp. 2040–2052.

[24] Stephen M Kissler et al. “Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period”. In: Science 368.6493 (2020), pp. 860–868.

[25] Max von Kleist et al. “HIV quasispecies dynamics during pro-active treatment switching: impact on multi-drug resistance and resistance archiving in latent reservoirs”. In: PloS one 6.3 (2011), e18204.

[26] Johannes Köster and Sven Rahmann. “Snakemake—a scalable bioinformatics workflow engine”. In: Bioinformatics 28.19 (2012), pp. 2520–2522.

[27] Denise Kühnert, Chieh-Hsi Wu, and Alexei J Drummond. “Phylogenetic and epidemic modeling of rapidly evolving infectious diseases”. In: Infection, genetics and evolution 11.8 (2011), pp. 1825–1841.

[28] Philippe Lemey et al. “SARS-CoV-2 European resurgence foretold: interplay of introductions and persistence by leveraging genomic and mobility data”. In: Research Square (2021). DOI: 10.21203/rs.3.rs-208849/v1.

[29] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. In: Bioinformatics 34.18 (2018), pp. 3094–3100.

[30] Qun Li et al. “Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia”. In: New England journal of medicine (2020). DOI: 10.1056/NEJMoa2001316.

[31] Milena Lopreite et al. “Early warnings of COVID-19 outbreaks across Europe from social media”. In: Scientific reports 11.1 (2021), pp. 1–7.

[32] Katrina A Lythgoe et al. “SARS-CoV-2 within-host diversity and transmission”. In: Science 372.6539 (2021).

[33] Shabir A. Madhi et al. “Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant”. In: The New England Journal of Medicine (2021). DOI: 10.1056/NEJMoa2102214.

[34] Anup Malani et al. “Seroprevalence of SARS-CoV-2 in slums versus non-slums in Mumbai, India”. In: The Lancet Global Health 9.2 (2021), e110–e111.

[35] Emmanuel Margolin et al. “Prospects for SARS-CoV-2 diagnostics, therapeutics and vaccines in Africa”. In: Nature Reviews Microbiology 18.12 (2020), pp. 690–704.

[36] Asher Mullard. “How COVID Vaccines Are Being Divvied up around the World”. In: Nature (2020). DOI: 10.1038/d41586-020-03370-6.

[37] Firzan Nainu et al. “SARS-CoV-2 reinfection and implications for vaccine development”. In: Human Vaccines & Immunotherapeutics 16.12 (2020), pp. 3061–3073.

[38] Thomas Obadia, Romana Haneef, and Pierre-Yves Boëlle. “The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks”. In: BMC medical informatics and decision making 12.1 (2012), pp. 1–9.

[39] Daniel P Oran and Eric J Tophel. “Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review”. In: Annals of internal medicine 173.5 (2020), pp. 362–367.

[40] Emmanuel Paradis and Klaus Schliep. “Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R”. In: Bioinformatics (Oxford, England) 35.3 (2019), pp. 526–528.

[41] Masatoshi Nei and Fumio Tajima. “Genetic drift and estimation of effective population size”. In: Genetics 98.3 (1981), pp. 625–640.

[42] Nuala A. O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: Nucleic Acids Research 44.D1 (2015), pp. D733–D745.

[43] Thomas Obadia, Romana Haneef, and Pierre-Yves Boëlle. “The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks”. In: BMC medical informatics and decision making 12.1 (2012), pp. 1–9.

[44] Nicky Phillips. “The Coronavirus Is Here to Stay - Here’s What That Means”. In: Nature 590.7846 (2021), pp. 382–384.
[45] Delphine Planas et al. “Sensitivity of infectious SARS-CoV-2 B. 1.1. 7 and B. 1.351 variants to neutralizing antibodies”. In: *Nature medicine* (2021). DOI: 10.1038/s41591-021-01318-5.

[46] Fernando P Polack et al. “Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine”. In: *New England Journal of Medicine* 383.27 (2020), pp. 2603–2615.

[47] Kaveh Pouran Yousef et al. “Inferring HIV-1 transmission dynamics in Germany from recently transmitted viruses”. In: *AIDS Journal of Acquired Immune Deficiency Syndromes* 73.3 (2016), pp. 356–363.

[48] Andrew Rambaut et al. “A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology”. In: *Nature Microbiology* 5.11 (2020), pp. 1403–1407.

[49] Andrew Rambaut et al. “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7”. In: *Systematic Biology* 67.5 (2018), pp. 901–904.

[50] Chadi M Saad-Roy et al. “Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years”. In: *Science* 370.6518 (2020), pp. 811–818.

[51] Ester C Sabino et al. “Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence”. In: *The Lancet* 397.10273 (2021), pp. 452–455.

[52] Sinaye Ngcapu San Emmanuel James et al. “High Resolution Analysis of Transmission Dynamics of Sars-Cov-2 in Two Major Hospital Outbreaks in South Africa Leveraging Intrahost Diversity”. In: *medRxiv: The Preprint Server for Health Sciences* (2020). DOI: 10.1101/2020.11.15.20231993.

[53] Cuihua Shen et al. “Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: Observational infoveillance study”. In: *Journal of medical Internet research* 22.5 (2020), e19421.

[54] Yuelong Shu and John McCauley. “GISAID: Global initiative on sharing all influenza data–from vision to reality”. In: *Eurosurveillance* 22.13 (2017), p. 30494.

[55] Tanja Stadler et al. “Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)”. In: *Proc Natl Acad Sci U S A* 110.1 (2013), pp. 228–33.

[56] John Stover et al. “The Risks and Benefits of Providing HIV Services during the COVID-19 Pandemic”. In: *medRxiv* (2021), p. 2021.03.01.21252663. DOI: 10.1101/2021.03.01.21252663.

[57] Yatish Turakhia et al. “Stability of SARS-CoV-2 phylogenies”. In: *PLOS Genetics* 16.11 (2020), pp. 1–34.

[58] Erik M Volz, Katia Koelle, and Trevor Bedford. “Viral phylodynamics”. In: *PLoS Comput Biol* 9.3 (2013), e1002947.

[59] Erik M Volz et al. “Phylodynamics of infectious disease epidemics”. In: *Genetics* 183.4 (2009), pp. 1421–1430.

[60] Jacco Wallinga and Peter Teunis. “Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures”. In: *American Journal of epidemiology* 160.6 (2004), pp. 509–516.

[61] J Wang, E Santiago, and Armando Caballero. “Prediction and estimation of effective population size”. In: *Heredity* 117.4 (2016), pp. 193–206.

[62] Olivier J Wouters et al. “Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment”. In: *The Lancet* 397.10278 (2021), pp. 1023–1034.

[63] Chieh-Hsi Wu and Alexei J Drummond. “Joint inference of microsatellite mutation models, population history and genealogies using transdimensional Markov Chain Monte Carlo”. In: *Genetics* 188.1 (2011), pp. 151–164.

[64] Sean L Wu et al. “Substantial underestimation of SARS-CoV-2 infection in the United States”. In: *Nature communications* 11.1 (2020), pp. 1–10.

[65] Daming Zhou et al. “Evidence of Escape of SARS-CoV-2 Variant B.1.351 from Natural and Vaccine-Induced Sera”. In: *Cell* 184.9 (2021), 2348–2361.e6.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- BEAST2ConfigurationFiles.zip
- nCovPopDynAppendix.pdf