A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages

Simon Dellicour[1,2,*], Keith Durkin[3], Samuel L. Hong[2], Bert Vanmechelen[2], Joan Martí-Carreras[2], Mandev S. Gill[2], Cécile Meex[4], Sébastien Bontems[4], Emmanuel André[2], Marius Gilbert[1], Conor Walker[5], Nicola De Maio[5], James Hadfield[6], Marie-Pierre Hayette[4], Vincent Bours[3], Tony Wawina-Bokalanga[2], Maria Artesi[3], Guy Baele[2], and Piet Maes[2]

[1] Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles, CP160/12 50, av. FD Roosevelt, 1050 Bruxelles, Belgium.

[2] Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

[3] Department of Human Genetics, CHU Liège, and Medical Genomics, GIGA Research Center, University of Liège, Belgium.

[4] Department of Clinical Microbiology, University of Liège, 4000, Liège, Belgium.

[5] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.

[6] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.

∗ Corresponding author (simon.dellicour@ulb.ac.be)

Since the start of the COVID-19 pandemic, an unprecedented number of genomic sequences of the causative virus (SARS-CoV-2) have been publicly released. The resulting volume of available genetic data presents a unique opportunity to gain real-time insights into the pandemic, but also a daunting computational hurdle if analysed with gold-standard phylogeographic methods. We here describe and apply an analytical pipeline that is a compromise between fast and rigorous analytical steps. As a proof of concept, we focus on Belgium, one of the countries with the highest spatial density of sequenced SARS-CoV-2 genomes. At the global scale, our analyses confirm the importance of external introduction events in establishing transmission chains in the country. At the country scale, our spatially-explicit phylogeographic analyses highlight an impact of the national lockdown of mid-March on the dispersal velocity of viral lineages. Our pipeline has the potential to be quickly applied to other countries or regions, with key benefits in complementing epidemiological analyses in assessing the impact of intervention measures or their progressive easement.

Keywords: COVID-19, SARS-CoV-2, phylodynamic, phylogeography, Nextstrain, phylogenetic clusters

First reported in early December 2019 in the province of Hubei (China), COVID-19 (coronavirus disease 2019) is caused by a new coronavirus (SARS-CoV-2) and has since rapidly spread around the world[1,2], causing enormous public health, social and economic impacts[3,4]. Since the early days of the pandemic, there has been an important mobilisation of the scientific community to understand its epidemiology and help providing a real-time response. To this end, research teams around the world have massively sequenced and publicly released viral genome sequences to study the origin of the virus[5,6], as well as to allow evolutionary analyses to link worldwide infectious cases and retrace the dispersal history of the virus[6,7]. In this context, a platform like Nexstrain, already widely used and recognised by the academic community and beyond, has quickly become a reference to follow the travel history of SARS-CoV-2 lineages[8].

In the context of the COVID-19 pandemic, the volume of genomic data available presents a unique opportunity to gain valuable real-time insights into the dispersal dynamics of the virus. However, the number of available viral genomes is increasing every day, leading to substantial computational challenges. While Bayesian phylogeographic inference represents the gold standard for inferring the dispersal history of viral lineages[9], these methods are computationally intensive and will fail to provide useful results in an acceptable amount of time. To tackle this practical limitation, we here describe and apply an analytical pipeline that is a compromise between fast and rigorous analytical steps.

In practice, we propose to take advantage of the time-scaled phylogenetic tree available and kept up-to-date on the online Nextstrain platform[8]. Specifically, we propose to use this maximum-likelihood tree as a fixed empirical tree along which we infer the ancestral locations with the discrete[10] or spatially-explicit[11] phylogeographic models implemented in the software package BEAST 1.10[12].

In Belgium, there are mainly two different laboratories (from the University of Leuven and of Liège) involved in sequencing SARS-CoV-2 genomes extracted from confirmed COVID-19 positive patients. To date, a few genomes (20) have also been sequenced at the University of Ghent, but for which metadata about the geographic origin are unavailable. As of the 22nd of April 2020, 391 genomes had been sequenced by these research teams and deposited on the GISAID database (Global Initiative on Sharing All Influenza Data[13]), making Belgium, at that time, the third largest public contributor of SARS-CoV-2 genomes. In the present study, we exploit this comprehensive data set to unravel the dispersal history and dynamics of SARS-CoV-2 viral lineages in Belgium. In particular, our objective is to investigate the evolution of the circulation dynamics through time and assess the impact of lockdown measures on spatial transmission. Specifically, we aim to use phylogeographic approaches to look at the Belgian epidemic at three different levels: (i) the importance of introduction events into the country, (ii) viral lineages circulation at the nationwide level, and (iii) viral lineages circulation at a more localised level, i.e. a province for which we have a particularly dense sampling.

RESULTS

Importance of introduction events into the country

On the 20th of April 2020, we downloaded the time-scaled maximum-likelihood tree available on Nextstrain, which was based on 4,623 SARS-CoV-2 sequences originating from 63 countries, including the 391 Belgian genomes available at that time. We then ran a preliminary discrete phylogeographic analysis along this tree to identify internal nodes and descending clades that likely correspond to distinct introductions into the Belgian territory (Fig. 1, S2). We inferred a minimum number of 166 introduction events (95% HPD interval = [161-171]). When compared to the number of sequences sampled in Belgium (391), this number illustrates the relative importance of external introductions in establishing transmission chains in the country. Introduction events resulted in distinct clades (or "clusters") linking varying numbers of sampled sequences (Fig. 1). However, most clusters (127 out of 165) only consisted of one sampled sequence. According to the time-scaled phylogenetic tree and discrete phylogeographic reconstruction (Fig. S1), some of these introduction events could have occurred before the return of carnival holidays (around the 1st of

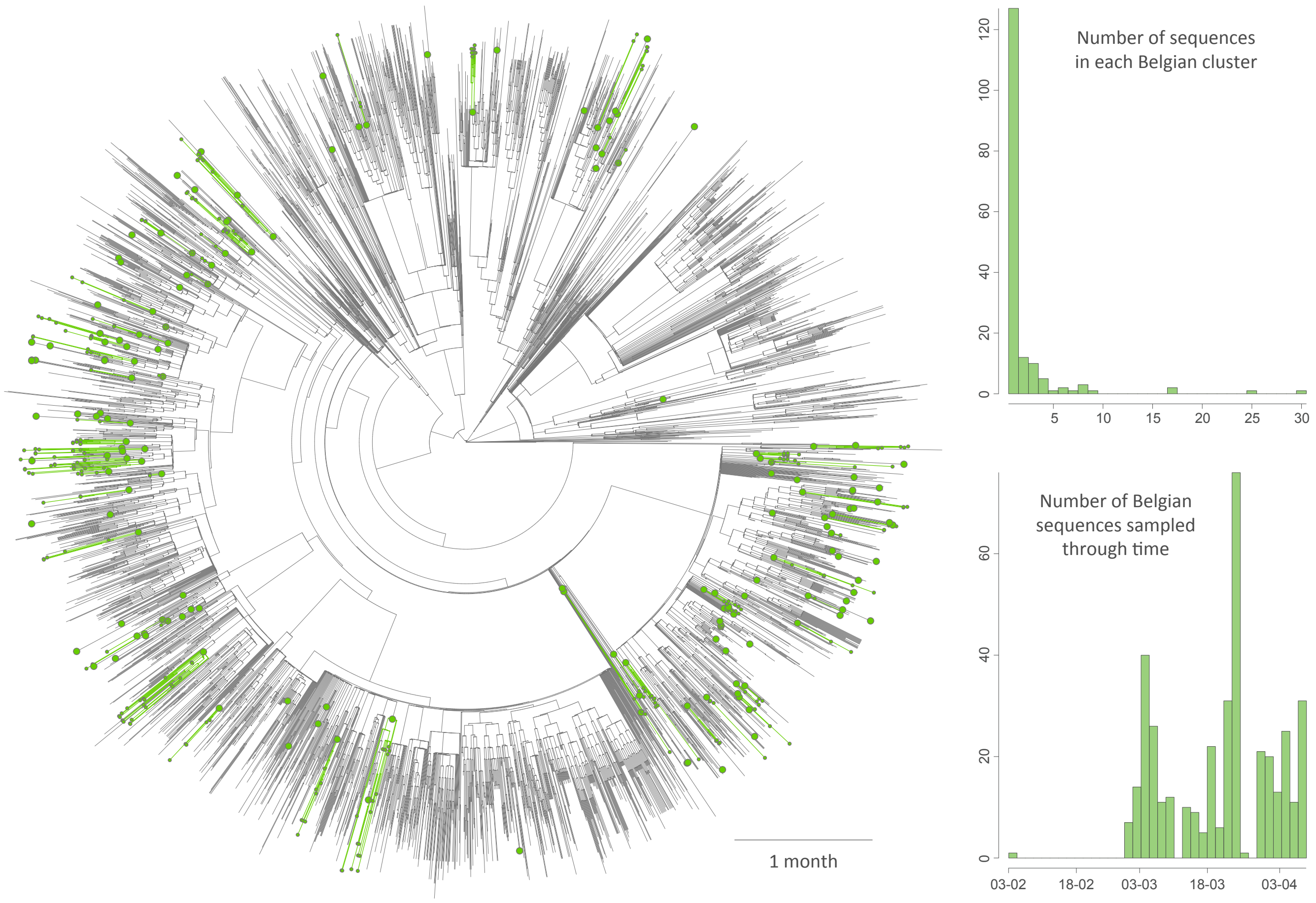

Figure 1. Time-scaled phylogenetic tree downloaded from Nextstrain (on the 22th of April 2020), in which we identified Belgian clusters. A cluster is here defined as a phylogenetic clade likely corresponding to a distinct introduction into the study area (Belgium). We delineated these clusters by performing a simplistic discrete phylogeographic reconstruction along the Nextstrain tree while only considering two potential ancestral locations: *Belgium* and *outside Belgium*. We identified a minimum number of 165 lineage introductions (95% HPD interval = [155-177]), which gives the relative importance of external introduction considering the number of sequences currently sampled in Belgium (391). On the tree, lineages circulating in Belgium are highlighted in green, and green nodes correspond to the most ancestral node of each Belgian cluster. See Figure S1 for a non-circular visualisation of the same tree, and Figure S2 for a first exploration of the spatial distribution of Belgian clusters. Besides the tree, we also report the distribution of cluster sizes (number of sampled sequences in each cluster) as well as the number of sequences sampled through time.

March), which was considered as the major entry point of transmission chains in Belgium.

Impact of lockdown measures at the country level

To analyse the circulation dynamics of viral lineages within the country, we performed spatially-explicit phylogeographic inference along Belgian clades, as identified by discrete phylogeographic analysis and corresponding to distinct introduction events. By placing phylogenetic branches in a geographical context, spatially-explicit phylogeographic inference allows treating those branches as conditionally independent movement vectors[14]. Here, we looked at these movement vectors to assess if the dispersal velocity of lineages was impacted by the national lockdown, of which the main measures were implemented on March 18th, 2020. When inspecting the duration of and geographic distance covered by lineages, our analysis revealed that fast long-distance dispersal events mainly occurred before the lockdown (Fig. 2). This overall pattern could, at least in part, result from the relatively lower sampling effort currently available at the country-scale for the lockdown time period. However, we further confirmed this trend by looking at the spatially-explicit phylogeographic reconstruction in itself (Fig. 3A) and estimates of the lineage dispersal velocity through time (Fig. 3B): while the lineage dispersal velocity was globally higher at the early phase of the Belgian epidemic, which corresponds to the week following the returns from carnival holidays, it then seemed to reach a plateau before further decreasing around the end of April. Furthermore, we recorded more long-distance lineage dispersal events before (Fig. 3C) than during (Fig. 3D) the lockdown.

Impact of lockdown measures at the province level

Taking advantage of the more intense sampling in the province of Liège, we also focused on this area to measure the lockdown impact at a more localised scale. The analysis of this subset of phylogenetic

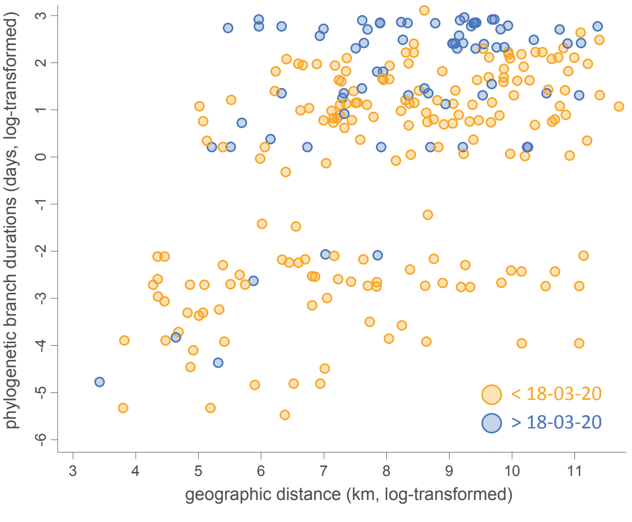

Figure 2. Comparison between phylogenetic branches occurring in Belgium before or after the national lockdown (18th March 2020). Geographic distances covered by phylogenetic branches were estimated by continuous phylogeographic inference.

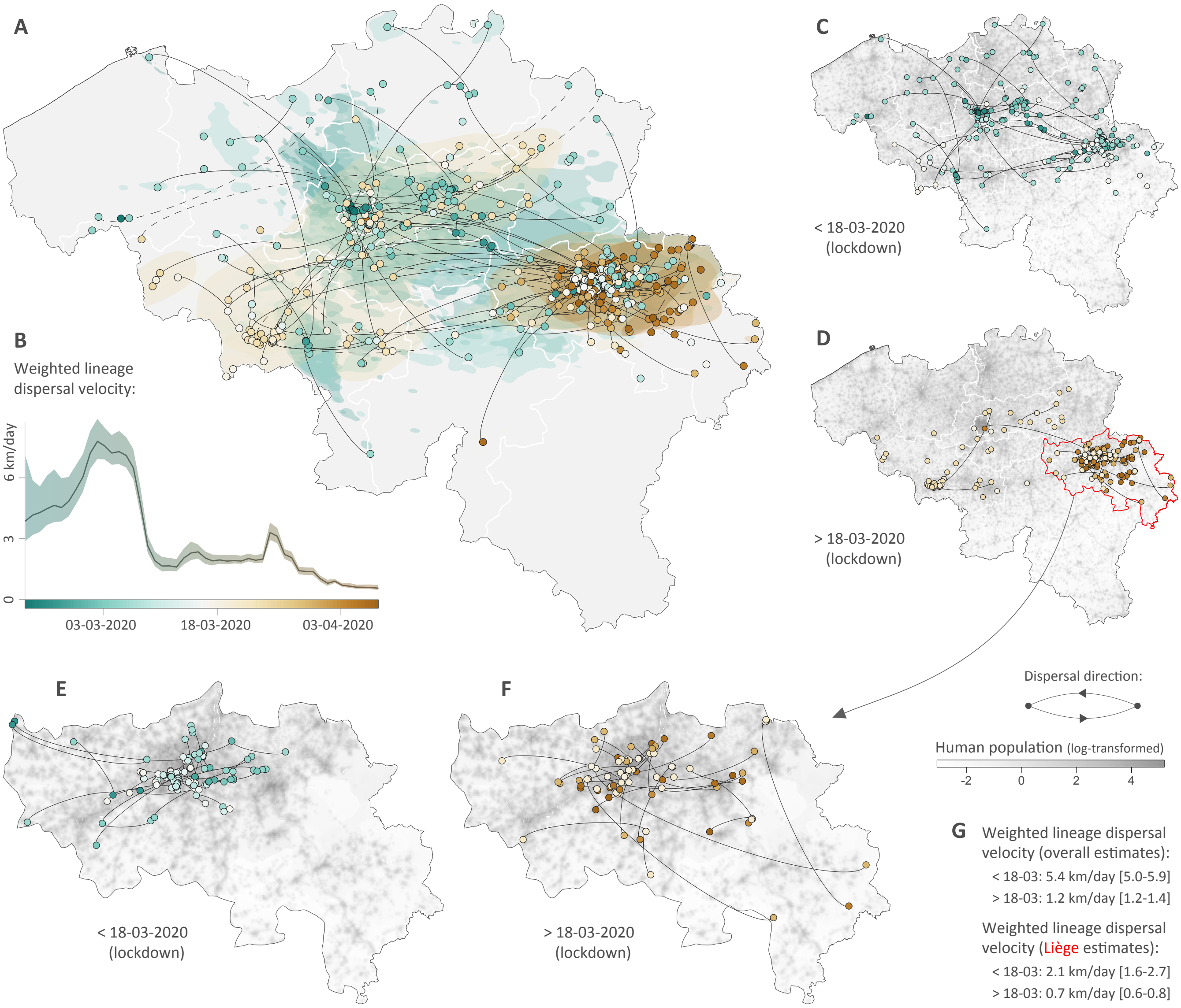
A
B
Weighted lineage dispersal velocity:
6 km/day
3
0
03-03-2020
18-03-2020
03-04-2020
C
< 18-03-2020 (lockdown)
D
> 18-03-2020 (lockdown)
Dispersal direction:
Human population (log-transformed)
-2
0
2
4
E
< 18-03-2020 (lockdown)
F
> 18-03-2020 (lockdown)
G
Weighted lineage dispersal velocity (overall estimates):
< 18-03: 5.4 km/day [5.0-5.9]
> 18-03: 1.2 km/day [1.2-1.4]
Weighted lineage dispersal velocity (Liège estimates):
< 18-03: 2.1 km/day [1.6-2.7]
> 18-03: 0.7 km/day [0.6-0.8]

Figure 3. Spatially-explicit phylogeographic reconstruction of the dispersal history of SARS-CoV-2 lineages in Belgium. (**A**) Continuous phylogeographic reconstruction performed along each Belgian clade (cluster) identified by the initial discrete phylogeographic analysis. For each clade, we mapped the maximum clade credibility (MCC) tree and overall 80% highest posterior density (HPD) regions reflecting the uncertainty related to the phylogeographic inference. MCC trees and 80% HPD regions are based on 1,000 trees subsampled from each post burn-in posterior distribution. MCC tree nodes were coloured according to their time of occurrence, and 80% HPD regions were computed for successive time layers and then superimposed using the same colour scale reflecting time. Continuous phylogeographic reconstructions were only performed along Belgian clades linking at least three sampled sequences for which the geographic origin was known (see the Methods section for further details). Besides the phylogenetic branches of MCC trees obtained by continuous phylogeographic inference, we also mapped sampled sequences belonging to clades linking less than three geo-referenced sequences. Furthermore, when a clade only gathers two geo-referenced sequences, we highlighted the phylogenetic link between these two sequences with a dashed curve connecting them. Sub-national province borders are represented by white lines. (**B**) Evolution of the weighted lineage dispersal velocity through time, with the surrounding polygon representing the 95% credible interval. (**C**) MCC tree branches occurring before the 18th March 2020 (beginning of the Belgian lockdown). (**D**) MCC tree branches occurring after the 18th March 2020. (**E**) MCC tree branches occurring before the 18th March 2020 in the province of Liège. (**F**) MCC tree branches occurring after the 18th March 2020 in the province of Liège. (**G**) Estimates of the weighted lineage dispersal velocity for different subsets of phylogenetic branches [and corresponding 95% HPD intervals].

branches did not highlight any apparent restriction in terms of long-distance dispersal events of viral lineages (Figs. 3E-F). Indeed, we still observed that viral lineages were able to travel non-negligible geographic distances despite the lockdown context (Fig. 3E). However, the estimation and comparison of lineage dispersal velocity before and during the lockdown confirmed the general decrease observed at the country-scale (Fig. 3G).

DISCUSSION

Our preliminary phylogeographic investigation reveals the important contribution of external introduction events. This overall result confirms that transmission chains circulating in Belgium were not established by a relatively restricted number of isolated infectious cases, e.g. people returning from skiing holidays in northern Italy. On the contrary, we identify a large number of distinct clades given the number of analysed sequences sampled in Belgium. This overall observation is in line with other reports, e.g. in California where no predominant lineage was identified[15] either, or with other country-specific patterns that can be visually explored on Nextstrain[8].

Our spatially-explicit phylogeographic analyses uncover the spatiotemporal distribution of Belgian SARS-CoV-2 clusters, indicating a potential impact of the national lockdown on long-distance dispersal events at the national scale, as well as on the lineage dispersal velocity at both the national and a more restricted scale. While the country will now enter in a phase of progressive easement of lockdown measures, it will be informative to keep monitoring lineage dispersal across the country. Increased dispersal velocity or more frequent long-distance dispersal events could point to an uptick in virus circulation.

Applying the present phylodynamic pipeline in a real-time perspective does not come without risk as new sequences can sometimes be associated with suspicious substitutions. For instance, several SARS-CoV-2 genome sequences are suspected to include unlikely homoplasies that could potentially result from systematic sequencing errors (`http://virological.org/t/issues-with-sars-cov-2-sequencing-data/473`). To assess the effect of such potential sequencing errors, we have repeated all the analyses after having discarded 21 tip branches associated with Belgian sequences carrying such suspicious homoplasies (C3130T, A4050C, T8022G, T13402G, C27046T, T28785G). Our results are however fully consistent with the analyses performed on the entire data set. Directly starting from inference results kept up to date by a platform like Nextstrain allows for fast analytical processing but also relies on newly deposited data that could sometimes carry potential errors, as we have investigated here for specific sequences that were generated in Belgium. To remedy such potentially challenging situations, our proposed pipeline could be extended with a sequence data resource component that makes uses of expert knowledge regarding a particular virus. The GLUE[16] software package allows new sequences to be systematically checked for potential issues, and could hence be an efficient tool to safely work with frequently updated SARS-CoV-2 sequencing data. Such a “CoV-GLUE” resource is currently being developed (`http://cov-glue.cvr.gla.ac.uk/#/home`).

While we acknowledge that a fully integrated analysis (i.e. an analysis where the phylogenetic tree and ancestral locations are jointly inferred) would be preferable, fixing an empirical tree represents a good compromise to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. Indeed, the number of genomes available, as well as the number of different sampling locations to consider in the analysis, would lead to a joint analysis requiring weeks of run-time in a Bayesian phylogenetic software package like BEAST. To illustrate the computational demands of such an approach, we ran a classic Bayesian phylogenetic analysis on a smaller SARS-CoV-2 data set (2,795 genomic sequences) using BEAST 1.10 (data not shown). This analysis required over 150 hours to obtain enough samples from the joint posterior, while using the latest GPU accelerated implementations[17] on 15 parallel runs. With a combined chain length of over 2.2x10^9 states, and an average runtime of 0.9 hours per million states, the significant computational demands required make this approach impractical when speed is critical. On the other hand, Nextstrain uses a maximum likelihood method implemented in the program TreeTime[18] to infer a time-scaled phylogenetic tree in a short amount of time (~3 hours for the data set analysed here). Given the present urgent situation, we have deliberately employed the Nextstrain tree as a fair estimate of the true time-scaled phylogenetic tree.

Our analytical workflow has the potential to be rapidly applied to study the dispersal history and dynamics of SARS-CoV-2 lineages in other restricted or even much broader study areas. We believe that spatially-explicit reconstruction can be a valuable tool for highlighting specific patterns related to the circulation of the virus or assessing the impact of intervention measures. While new viral genomes are sequenced and released daily, a limitation could paradoxically arise from the non-accessibility of associated metadata. Indeed, without sufficiently precise data about the geographic origin of each genome, it is not feasible to perform a spatially-explicit phylogeographic inference. In the same way that viral genomes are deposited in databases like GISAID, metadata should also be made available to enable comprehensive epidemiological investigations with a similar approach as we presented here.

METHODS

SARS-CoV-2 sequencing in Belgium. At the University of Leuven, RNA extracts from SARS-CoV-2 infected patients were selected anonymously and based on postcode city of Belgium. These RNA extracts were provided by the National Reference Center for Coronaviruses and UZ Leuven. Reverse transcription was carried out via SuperScript IV and cDNA was posteriorly amplified using Q5® High-Fidelity DNA Polymerase (NEB) with the ARTIC nCov-2019 primers and following the recommendations in the sequencing protocol of the ARTIC Network (`https://artic.network/ncov-2019`). Samples were multiplexed following the manufacturer’s recommendations using the Oxford Nanopore Native Barcoding Expansion kits NBD104 (1-12) and NBD114 (13-24), in conjunction with Ligation Sequencing Kit 109 (Oxford Nanopore). Sequencing was carried out on MinION sequencer using R9.4.1 flow cells and MinKNOW 2.0 software.

At the University of Liège, RNA was extracted from clinical samples (300μl) via a Maxwell 48 device using the Maxwell RSC Viral TNA kit (Promega) with a viral inactivation step using Proteinase K, following the manufacturer’s instructions. RNA elution occurred in 50μl of RNAse free water. Reverse transcription was carried out via SuperScript IV VILOTM Master Mix, and 3.3μl of the eluted RNA was combined with 1.2μl of master mix and 1.5μl of H_2O. This was incubated at 25°C for 10 min, 50°C for 10 min and 85°C for 5 min. PCR used Q5® High-Fidelity DNA Polymerase (NEB), the primers and conditions followed the recommendations in the sequencing protocol of the ARTIC Network. Samples were multiplexed following the manufacturer’s recommendations using the Oxford Nanopore Native Barcoding Expansion kits 1-12 and 13-24, in conjunction with Ligation Sequencing Kit 109 (Oxford Nanopore). Sequencing was carried out on a Minion using R9.4.1 flow cells. Data analysis followed the SARS-CoV-2 bioinformatics protocol of the ARTIC Network.

Preliminary discrete phylogeographic analysis. We performed a preliminary phylogeographic analysis using the discrete diffusion model[10] implemented in the software package BEAST 1.10[12]. The objective of this first analysis was to identify independent introduction events of SARS-CoV-2 lineages into Belgium. To this end, we used the Nextstrain tree as a fixed empirical tree and only considered two possible ancestral locations: *Belgium* and *outside Belgium*. Bayesian inference through Markov chain Monte Carlo (MCMC) was run on this empirical tree for 10^6 generations and sampled every 1,000 generations. MCMC convergence and mixing properties were inspected using the program Tracer 1.7[19] to ensure that effective sample size (ESS) values associated with estimated parameters were all >200. After having discarded 10% of sampled trees as burn-in, a maximum clade credibility (MCC) tree was obtained using TreeAnnotator 1.10[12]. We used the resulting MCC tree to delineate Belgian clusters here defined as phylogenetic clades corresponding to independent introduction events in Belgium.

Continuous and post hoc phylogeographic analyses. We used the continuous diffusion model[11] available in BEAST 1.10[12] to perform a spatially-explicit (or “continuous”) phylogeographic reconstruction of the dispersal history of SARS-CoV-2 lineages in Belgium. We employed a relaxed random walk (RRW) diffusion model to generate a posterior distribution of trees whose internal nodes are associated with geographic coordinates[11]. Specifically, we used a Cauchy distribution to model the among-branch heterogeneity in diffusion velocity. We performed a distinct continuous phylogeographic reconstruction for each Belgian clade identified by the initial discrete phylogeographic inference, again fixing a Nextstrain subtree as an empirical tree. As phylogeographic inference under the continuous diffusion model does not allow identical sampling coordinates assigned to the tips of the tree, we avoided assigning sampling coordinates using the centroid point of each administrative area of origin. For a given sampled sequence, we instead retrieved geographic coordinates from a point randomly sampled within the administrative area of origin, which is the maximal level of spatial precision in available metadata. This approach avoids using the common “jitter” option that adds a restricted amount of noise to duplicated sampling coordinates. Using such a jitter could be problematic because it can move sampling coordinates to administrative areas neighbouring their actual administrative area of origin[20]. Furthermore, the administrative areas considered here are municipalities and are rather small (there are currently 581 municipalities in Belgium). The clade-specific continuous phylogeographic reconstructions were only based on Belgian tip nodes for which the municipality of origin was known, i.e. 380 out of 391 genomic sequences. Furthermore, we only performed a continuous phylogeographic inference for Belgian clades linking a minimum of three tip nodes with a known sampling location (municipality).

Each Markov chain was run for 10^6 generations and sampled every 1,000 generations. As with the discrete phylogeographic inference, MCMC convergence/mixing properties were assessed with Tracer, and MCC trees (one per clade) were obtained with TreeAnnotator after discarding 10% of sampled trees as burn-in. We then used functions available in the R package “seraphim”[21,22] to extract spatiotemporal information embedded within the same 1,000 posterior trees and visualise the continuous phylogeographic reconstructions. We also used “seraphim” to estimate the following weighted lineage dispersal velocity, where d_i and t_i are the geographic distance travelled (great-circle distance in km) and the time elapsed (in days) on each phylogeny branch, respectively:

$$v_{\text{weighted}} = \sum_{i=1}^{n} d_i / \sum_{i=1}^{n} t_i.$$

1. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
2. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
3. McKee, M. & Stuckler, D. If the world fails to protect the economy, COVID-19 will damage health not just now but also in the future. *Nat. Med.* 1–3 (2020).
4. Holmes, E. A. *et al.* Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. *The Lancet Psychiat.* **0** (2020).
5. Andersen, K. G. *et al.* The proximal origin of SARS-CoV-2. *Nat. Med.* 1–3 (2020).
6. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).
7. Eden, J.-S. *et al.* An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6** (2020).
8. Hadfield, J. *et al.* Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
9. Baele, G. *et al.* Recent advances in computational phylodynamics. *Curr. Opin. Virol.* **31**, 24–32 (2018).
10. Lemey, P. *et al.* Bayesian phylogeography finds its roots. *PLoS Comp. Biol.* **5**, e1000520 (2009).
11. Lemey, P. *et al.* Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).
12. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
13. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
14. Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15066–15071 (2012).
15. Deng, X. *et al.* A genomic survey of SARS-CoV-2 reveals multiple introductions into northern California without a predominant lineage. *medRxiv* 2020.03.27.20044925 (2020).
16. Singer, J. B. *et al.* GLUE: A flexible software system for virus sequence data. *BMC Bioinf.* **19**, 532 (2018).
17. Ayres, D. L. *et al.* BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst. Biol.* **68**, 1052–1061 (2019).
18. Sagulenko, P. *et al.* TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
19. Rambaut, A. *et al.* Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
20. Dellicour, S. *et al.* Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat. Commun.* **9**, 2222 (2018).
21. Dellicour, S. *et al.* Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinf.* **17**, 1–12 (2016).
22. Dellicour, S. *et al.* SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics* **32**, 3204–3206 (2016).

Data accessibility The new sequences have been deposited in GISAID with accession IDs EPI_ISL_407976, EPI_ISL_417422, EPI_ISL_418800, EPI_ISL_418805-806, EPI_ISL_418863, EPI_ISL_418989, EPI_ISL_420314-454, EPI_ISL_415153-59, EPI_ISL_416469-76, EPI_ISL_417424-30, EPI_ISL_418981-87, EPI_ISL_418792-98, EPI_ISL_416997, EPI_ISL_417004-006, EPI_ISL_417008-009, EPI_ISL_417011-25, EPI_ISL_418623-666, EPI_ISL_421182-214, EPI_ISL_424628-665, EPI_ISL_427340-390. All data (sequence metadata, BEAST input and output files files) and R scripts for our analyses are available in the following GitHub repository: `https://github.com/sdellicour/covid19_spell`.

Acknowledgments We are grateful to Sébastien Kozlowskyj for his assistance in using the Dragon2 cluster of the University of Mons, and to Ine Boonen for her assistance during SARS-CoV-2 sequencing. SD and MG are supported by the Fonds National de la Recherche Scientifique (FNRS, Belgium). KD and MA, and the ULiège sequencing effort, are supported by the grant WALGEMED from the Walloon Region (convention n°1710180). BV is supported by a FWO SB grant for strategic basic research of the "Fonds Wetenschappelijk Onderzoek"/Research foundation Flanders (1S28617N). JMC is supported by a doctoral grant from HONOURs (Host switching pathogens, infectious outbreaks and zoonosis) Marie-Sklodowska-Curie training network (721367). GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094, and the Research Foundation Flanders ("Fonds voor Wetenschappelijk Onderzoek Vlaanderen", G0E1420N). TWB is supported by the Special Research Fund, KU Leuven (Bijzonder Onderzoeksfonds, KU Leuven, 3M170314 C14/17/100). Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (FNRS) under Grant No. 2.5020.11 and by the Walloon Region. We are grateful to all of the research teams that have deposited SARS-CoV-2 genome data on GISAID.