Clinical Infectious Diseases

BRIEF REPORT

Phylogenetic Analyses of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) B.1.1.7 Lineage Suggest a Single Origin Followed by Multiple Exportation Events Versus Convergent Evolution

A. Chaillon and D. M. Smith

Division of Infectious Diseases and Global Public Health, University of California San Diego, La Jolla, California, USA

The emergence of new variants of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) herald a new phase of the pandemic. This study used state-of-the-art phylodynamic methods to ascertain that the rapid rise of B.1.1.7 "Variant of Concern" most likely occurred by global dispersal rather than convergent evolution from multiple sources.

Keywords. SARS-CoV-2; variant of concern; B.1.1.7; convergent evolution; phylogenetics.

Following phylogenetic and epidemiological investigations, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genetic lineage B.1.1.7 is suspected to be associated with an increase in human-to-human viral transmissibility [1, 2] and was classified as a "variant of concern" (VOC B.1.1.7) on 18 December 2020 [3]. The variant was first discovered in Kent, United Kingdom, on 21 September 2020 and has since been identified in over 40 countries across the world, including the United States [3–6]. We sought to evaluate whether the breadth of VOC B.1.1.7 identification represents convergent evolution [7] or rapid local and global dispersal after this lineage's genesis.

On 14 January 2021, we downloaded all B.1.1.7 lineage SARS-CoV-2 genomic sequences available on the GISAID (Global Initiative on Sharing All Influenza Data) public database [8] (17 118 full length genome sequences across 36 countries, Supplementary Table 1). The vast majority were from the United Kingdom (95%, n = 16 263, generated by the national COVID-19 [coronavirus disease 2019] Genomics UK [COG-UK] Consortium) [9], but 855 sequences were from other countries, including 80 from North America (74 from the US; Supplementary Figure 1).

We combined these B.1.1.7 sequences with a representative set of non-B.1.1.7 sequences (n = 4768) based on sequence homology. All sequences were aligned using MAFFT and highly homoplasic sites were masked [10]. To reduce the data set size while maintaining an appropriate set of epidemiologically relevant background sequences, we used BLAST [11, 12] to identify the 50 closest non-B.1.1.7 variants to each of the 17 118 B.1.1.7 genomic sequences in the data set [13, 14]. After keeping one copy of duplicated entries that ranked among the 50 best hits, a total of 4768 sequences out of the 316 075 non-B1.1.7 sequences available on GISAID were kept for further analyses and combined with the B.1.1.7 data set. The final set of 21 886 sequences was aligned with MAFFT [15], and a maximum likelihood phylogeny was inferred using IQ-TREE v2.1.2 [16]. The resulting phylogeny showed that all available B.1.1.7 samples clustered together with high support (0.99 Shimodaira Hasegawa [SH] support [17–19]). Non-UK VOC B.1.1.7 sequences intermix within those from the United Kingdom (Figure 1). As convergent evolution can induce incorrect clustering [20], the same approach was repeated after excluding variable positions that define the B.1.1.7. lineage (Supplementary Table 2), which yielded a similar picture. These patterns are in line with the view that this variant successfully spread around the world after it arose in the United Kingdom.

To estimate the timing of introduction of B.1.1.7 variants outside the United Kingdom, we applied a multistep analytic approach, as previously described by our group for human immunodeficiency virus (HIV) [21, 22] (see Supplementary Information). B.1.1.7 clusters of size ≥ 2 including only non-UK sequences were identified from the ML phylogeny in R [23]. For each non-UK clade, the phylogeny was rescaled into units of time with treedater [24], assuming a strict molecular clock with the rate of SARS-CoV-2 genome evolution drawn from an externally estimated distribution, as previously described [25], and the rate was a mean of 9.41×10^{-4} nucleotide substitutions per site per year with a standard deviation of 4.99×10^{-5}. To incorporate uncertainty in the estimated clock rate, molecular clock estimation was replicated 100 times for each non-UK B.1.1.7 clade. We identified a total of 90 clades of size ≥ 2 for a total of 513 sequences (ranging from 2 to 135) including only B.1.1.7 variants from outside the United Kingdom. The largest cluster of 135 sequences was identified in Denmark across 5 regions. One third (60/90) were European exclusive clusters (Supplementary Table 1), whereas 12 clusters included sequences from the United States (5 sampled in California).

Received 20 January 2021; editorial decision 22 March 2021; published online 26 March 2021.
Correspondence: D. M Smith, University of California San Diego, La Jolla, CA (d13smith@health.ucsd.edu).

Clinical Infectious Diseases® 2021;73(12):2314–7

DOI: 10.1093/cid/ciab265

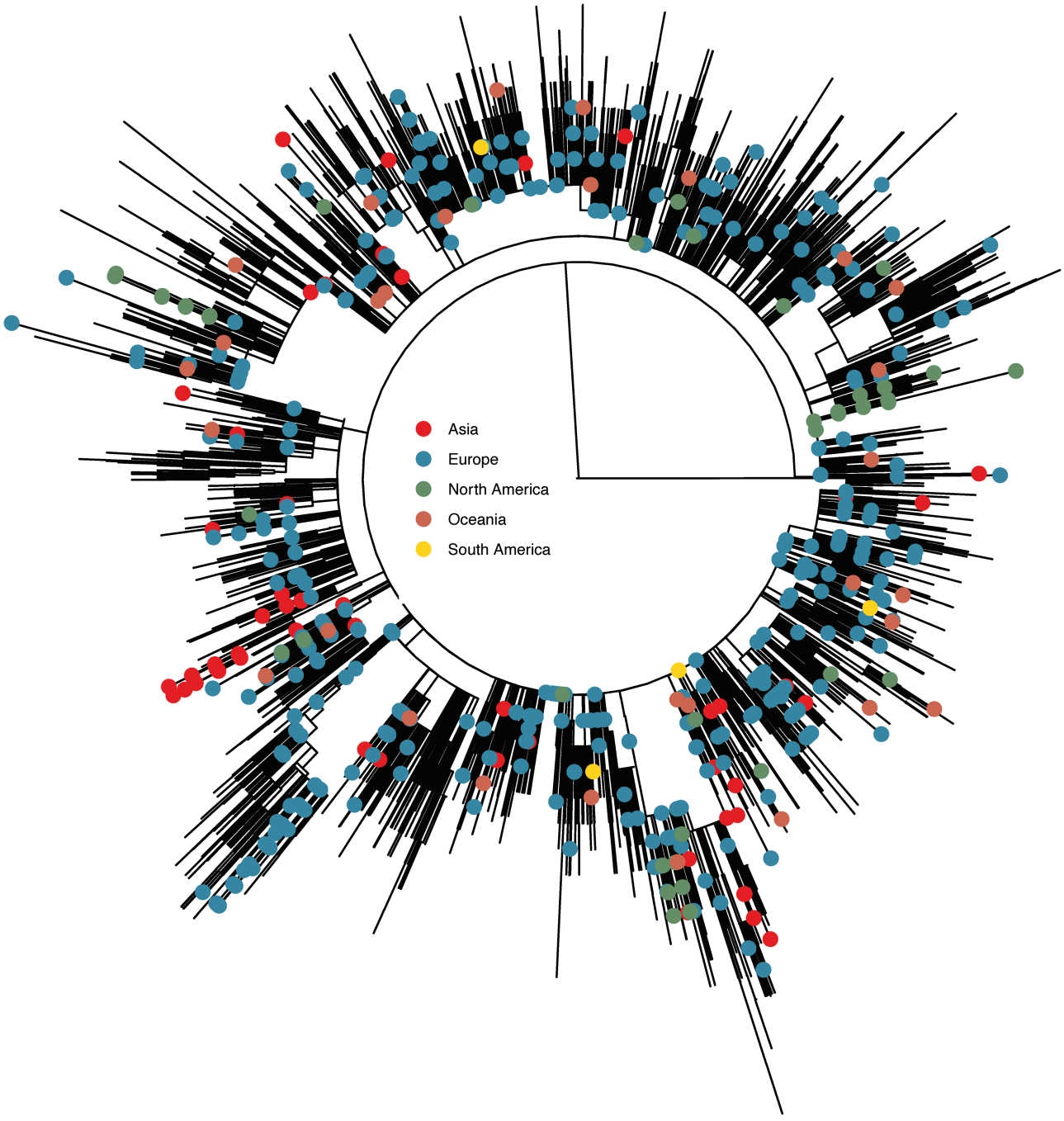

Figure 1. SARS-CoV-2 B.1.1.7 phylogenetic tree. Tips outside the UK ("non-UK B.1.1.7 tips") in the phylogeny are colored according to the continent of origin (red denotes taxa from Asia, blue denotes taxa from Europe, green denotes taxa from North America, maroon denotes taxa from Oceania, and yellow denotes taxa from South America; B.1.1.7 sequences from the UK are not colored). Abbreviations: SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; UK, United Kingdom.

The earliest estimated seeding of B.1.1.7 from the United Kingdom dates to 9 September 2020 in Denmark, and the most recent to 8 January 2021 in Spain (see Supplementary Table 3 and Supplementary Figure 2). The number of weekly introductions outside the United Kingdom peaked in mid-December (Figure 2). In the United States, the first introduction was estimated on 14 November in Florida. Five distinct introductions in California were also identified from 3 December to 26 December, including one cluster of 19 sequences. Of note, 6 international non-UK clusters including ≥2 countries were identified of whom 2 did not include European sequences (Supplementary Table 3).

In response to the rapid increase in viral infections and spread, UK officials announced a lockdown on 31 October that came into force on 5 November and ended on 5 December. Given time to the most recent common ancestor (TMRCA) estimates, we

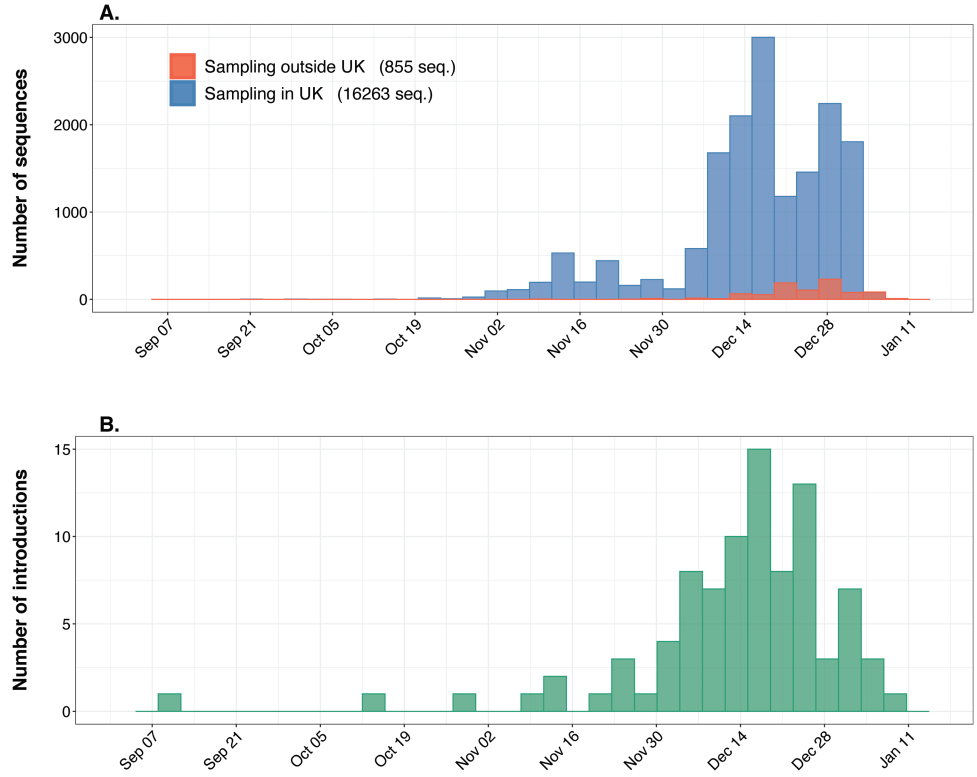

Figure 2. Number of sequences (*A*) and number of introductions of B.1.1.7 outside the UK (*B*). *A*, Biweekly number of B.1.17 VOC genome collected through time in the UK (*blue*) and in other countries (*red*) is presented. *B*, Vertical green bar represents the biweekly number of introductions. Abbreviations: UK, United Kingdom; VOC, variant of concern.

determined that 19% (17/90) of the exportation events that gave rise to detectable non-UK VOC B1.1.7 transmission lineages occurred during this period (the remaining 81% occurred before or after these dates). The emergence and rapid dispersal of this new VOC led to the implementation of a new national strict lockdown in the United Kingdom on 4 January 2021 [26].

As previously described by du Plessis et al[14], we next used the TMRCA of each non-UK clade to estimate the genomic "detection lag" for each cluster, which represents the duration that a transmission lineage went undetected before it was first sampled by genome sequencing. The mean detection lag was ~10.6 days (interquartile range [IQR] = 4–15). This largely agrees with detection lag-time estimates from SARS-CoV-2 importation into the United Kingdom in the first months of the pandemic [14], which was on average 8 days (IQR = 3–15, ~10 days for lineages comprising ≤10 genomes and <1 day for lineages of >100 genomes).

Of note, virus genome sequences have been determined for only a fraction of infections. Even in the United Kingdom, where the by far largest sequencing effort is done, only an estimated 4.3% (129 939 available sequences out of 3 039 797 cases reported on 14 January) [27] of infections have been sequenced. For this reason, and also because not all sequenced SARS-CoV-2 genomes are being deposited in the GISAID repository, many B.1.1.7 variants that successfully established transmission chains outside of the United Kingdom likely remain undetected (for now). Our estimated number of B.1.1.7 exportation events from the United Kingdom thus represents an underestimate. The sparse sampling and sequencing also poses limits to the accuracy with which introduction events can be dated (see du Plessis and colleagues [25] for a more detailed explanation).

Our results do not suggest that the canonical mutations of VOC B.1.1.7 evolved independently in different locations. Instead, our analyses point to an origin in and spread of the VOC B.1.1.7 from the United Kingdom. As for the virus' initial [28] and subsequent [29, 30] spread, global connectedness and high levels of human mobility undoubtedly facilitated VOC B.1.1.7 dissemination. The swift global spread of VOC B.1.1.7 illustrates that current restrictions are insufficient to prevent the spread of new and emerging variants [31–37]. Similar to Ebola [38], hepatitis C virus (HCV) [39, 40] and HIV [22], countermeasures to SARS-CoV-2 spread should be developed with a broader perspective than the national level. Otherwise, without population immunity, successful local reductions in SARS-CoV-2 burden will be counteracted by imported infections that set off new waves of viral spread, possibly exacerbated by novel phenotypic characteristics of the imported strains.

Supplementary Data

Supplementary materials are available at *Clinical Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Acknowledgments. The authors gratefully acknowledge the authors from the originating laboratories and the submitting laboratories who generated and shared via GISAID the data on which this research is based. In particular, the authors acknowledge the role of the COVID-19 Genomics UK (COG-UK) Consortium who generated the vast majority of sequences from the United Kingdom. See Supplementary Material for the acknowledgment table.

Financial support. This work was supported by grants from the National Institutes of Health (NIH) (San Diego Center for AIDS Research, CFAR, AI306214 and AI100665), the Department of Veterans Affairs, the John and Mary Tu Foundation, and the James B. Pendleton Charitable Trust. A. C. was also supported by NIH Grant AI131971 (R21). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Potential conflicts of interest. The authors: No reported conflicts of interest. Both authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

References

1. Rambaut A, Loman N, Pybus O, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. virological.org **2020**. Available at: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
2. Volz E, Mishra S, Chand M, et al. Transmission of SARS-CoV-2 lineage B.1.1.7 in England: insights from linking epidemiological and genetic data. medRxiv **2021**. doi:10.1101/2020.12.30.20249034.
3. Chand M, Hopkins S, Dabrera G, et al. Variant of Concern 202012/01. Public Health Engl **2020**. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959426/Variant_of_Concern_VOC_202012_01_Technical_Briefing_5.pdf.
4. Russell G, Woodhouse A, Dempsey H, Clarfelt H, Ralph O. Coronavirus: New York detects first case of UK variant, California reveals further occurrences—as it happened. Financial Times **2020**. Available at: https://www.ft.com/content/3abc2a29-3318-3001-9d1f-b5bead6c88f8.
5. Global report investigating novel coronavirus haplotypes (grinch). *B.1.1.7 report*. Available at: https://cov-lineages.org/global_report_B.1.1.7.html.
6. Rambaut A, Holmes EC, O'Toole Á, Hill V, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol **2020**. doi:10.1038/s41564-020-0770-5.
7. Volz E, Hill V, McCrone JT, et al; COG-UK Consortium. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. Cell **2021**; 184:64–75.e11.
8. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data: from vision to reality. Euro Surveill **2017**; 22. doi:10.2807/1560–7917.Es.2017.22.13.30494.
9. COVID-19 Genomics UK (COG-UK). An integrated national scale SARS-CoV-2 genomic surveillance network. Lancet Microbe **2020**; 1:e99–100. doi:10.1016/S2666-5247(20)30054–9.
10. De Maio N, Walker C, Borges R, et al. Issues with SARS-CoV-2 sequencing data. virological.org **2020**. Available at: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol **1990**; 215:403–10.
12. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics **2009**; 10:421.
13. Gräf T, Vrancken B, Maletich Junqueira D, et al. Contribution of epidemiological predictors in unraveling the phylogeographic history of HIV-1 subtype C in Brazil. J Virol **2015**; 89:12341–8.
14. Vrancken B, Adachi D, Benedet M, et al. The multi-faceted dynamics of HIV-1 transmission in Northern Alberta: a combined analysis of virus genetic and public health data. Infect Genet Evol **2017**; 52:100–5.
15. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **2013**; 30:772–80.
16. Minh BQ, Schmidt HA, Chernomor O, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol **2020**; 37:1530–4.
17. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **2010**; 59:307–21.
18. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **2003**; 52:696–704.
19. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol **1999**; 16:1114.
20. Lemey P, Derdelinckx I, Rambaut A, et al. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. J Virol **2005**; 79:11981–9.
21. Vrancken B, Zhao B, Li X, et al. Comparative circulation dynamics of the five main HIV types in China. J Virol **2020**; 94. doi:10.1128/jvi.00683-20.
22. Vrancken B, Mehta SR, Ávila-Ríos S, et al. Dynamics and dispersal of local HIV epidemics within San Diego and across the San Diego-Tijuana Border. Clin Infect Dis **2020**; doi:10.1093/cid/ciaa1588.
23. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics **2019**; 35:526–8.
24. Volz EM, Frost SDW. Scalable relaxed clock phylogenetic dating. Virus Evol **2017**; 3. doi:10.1093/ve/vex025.
25. du Plessis L, McCrone JT, Zarebski AE, et al. Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. medRxiv **2020**. doi:10.1101/2020.10.23.20218446.
26. Steed L, Cavanagh N. LOCKED IN When did lockdown start in the UK? Sun **2021**. Available at: https://www.the-sun.com/news/622129/when-lockdown-start/.
27. GOV.UK. Coronavirus (COVID-19) in the UK. **2020**. Available at: https://coronavirus.data.gov.uk/details/cases. Accessed 15 January 2021.
28. Worobey M, Pekar J, Larsen BB, et al. The emergence of SARS-CoV-2 in Europe and the US. bioRxiv **2020**. doi:10.1101/2020.05.21.109322: the preprint server for biology.
29. Badr HS, Gardner LM. Limitations of using mobile phone data to model COVID-19 transmission in the USA. Lancet Infect Dis **2020**; doi:10.1016/s1473-3099(20)30861–6.
30. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. Lancet Infect Dis **2020**; 20:1247–54.
31. Peiris JS, Yuen KY, Osterhaus AD, Stöhr K. The severe acute respiratory syndrome. N Engl J Med **2003**; 349:2431–41.
32. Johnson NP, Mueller J. Updating the accounts: global mortality of the 1918–1920 "Spanish" influenza pandemic. Bull Hist Med **2002**; 76:105–15.
33. Shortridge KF, Peiris JS, Guan Y. The next influenza pandemic: lessons from Hong Kong. J Appl Microbiol **2003**; 94 Suppl:70–9S.
34. Subbarao K, Katz J. Avian influenza viruses infecting humans. Cell Mol Life Sci **2000**; 57:1770–84.
35. Chua KB, Bellini WJ, Rota PA, et al. Nipah virus: a recently emergent deadly paramyxovirus. Science **2000**; 288:1432–5.
36. Nash D, Mostashari F, Fine A, et al; 1999 West Nile Outbreak Response Working Group. The outbreak of West Nile virus infection in the New York City area in 1999. N Engl J Med **2001**; 344:1807–14.
37. Reid AH, Taubenberger JK. The origin of the 1918 pandemic influenza virus: a continuing enigma. J Gen Virol **2003**; 84:2285–92.
38. Kamorudeen RT, Adedokun KA, Olarinmoye AO. Ebola outbreak in West Africa, 2014–2016: Epidemic timeline, differential diagnoses, determining factors, and lessons for future response. J Infect Public Health **2020**; 13:956–62.
39. Pérez AB, Vrancken B, Chueca N, et al. Increasing importance of European lineages in seeding the hepatitis C virus subtype 1a epidemic in Spain. Euro Surveill **2019**; 24. doi:10.2807/1560–7917.Es.2019.24.9.1800227.
40. Vrancken B, Cuypers L, Pérez AB, et al. Cross-country migration linked to people who inject drugs challenges the long-term impact of national HCV elimination programmes. J Hepatol **2019**; 71:1270–2.