bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introductions and early spread of SARS-CoV-2 in France

Fabiana Gámbaro[1,6]*, Sylvie Behillil[2,3]*, Artem Baidaliuk[1]*, Flora Donati[2,3], Mélanie Albert[2,3], Andreea Alexandru[4], Maud Vanpeene[4], Méline Bizard[4], Angela Brisebarre[2,3], Marion Barbet[2,3], Fawzi Derrar[5], Sylvie van der Werf[2,3] $, Vincent Enouf[2,3,4] $ and Etienne Simon-Loriere[1] $

*These authors contributed equally

$These authors co-supervised this work

Correspondence: sylvie.van-der-werf@pasteur.fr or etienne.simon-loriere@pasteur.fr

Affiliations

[1]Evolutionary genomics of RNA viruses, Institut Pasteur, Paris, France

[2]National Reference Center for Respiratory Viruses, Institut Pasteur, Paris, France

[3]Molecular Genetics of RNA Viruses, CNRS - UMR 3569, University of Paris, Institut Pasteur, Paris, France

[4]Mutualized Platform of Microbiology, Pasteur International Bioresources Network, Institut Pasteur, Paris, France

[5]National Influenza Centre, Viral Respiratory Laboratory, Algiers, Algeria

[6]Université de Paris, Paris, France

Abstract

Following the emergence of coronavirus disease (COVID-19) in Wuhan, China in December 2019, specific COVID-19 surveillance was launched in France on January 10, 2020. Two weeks later, the first three imported cases of COVID-19 into Europe were diagnosed in France. We sequenced 97 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genomes from samples collected between January 24 and March 24, 2020 from infected patients in France. Phylogenetic analysis identified several early independent SARS-CoV-2 introductions without local transmission, highlighting the efficacy of the measures taken to prevent virus spread from symptomatic cases. In parallel, our genomic data reveals the later predominant circulation of a major clade in many French regions, and implies local circulation of the virus in undocumented infections prior to the wave of COVID-19 cases. This study emphasizes the importance continuous and geographically broad genomic sequencing and calls for further efforts with inclusion of asymptomatic infections.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was identified as the cause of an outbreak of severe respiratory infections in Wuhan, China in December 2019 (Zhu, Zhang et al. 2020). Although Chinese authorities implemented strict quarantine measures in Wuhan and surrounding areas, this emerging virus has rapidly spread across the globe, and the World Health Organization (WHO) declared a pandemic of coronavirus disease 2019 (COVID-19) on March 11, 2020.

Strengthened surveillance of COVID-19 cases was implemented in France on January 10, 2020, with the objective of identifying imported cases early to prevent secondary transmission in the community, and the National Reference Center for Respiratory Viruses (NRC) hosted at Institut Pasteur identified the first cases in Europe. With the extension of the epidemic, identification of SARS-CoV-2 cases was shared with the NRC associated laboratory in Lyon and then extended to additional first line hospital laboratories, with the NRC at Institut Pasteur focusing on the Northern part of France, including the densely populated capital. Screening and sampling for SARS-CoV-2 was targeted towards patients who had symptoms (fever and/or respiratory problems) or had travel history to risk zones of infection. With the spread of the virus, it became clear that clinical characteristics of COVID-19 patients vary greatly (Guan, Ni et al. 2020, Onder, Rezza et al. 2020) with a proportion of asymptomatic infections or mild disease cases (Li, Pei et al. 2020).

Viral genomics coupled with modern surveillance systems is transforming the way we respond to emerging infectious diseases (Gardy and Loman 2018, Ladner, Grubaugh et al. 2019). Real-time genomic epidemiology data has proven to be useful to reconstruct outbreaks dynamics: from virus identification to understanding the factors contributing towards global spread (Grubaugh, Ladner et al. 2019). Here, we sequenced SARS-CoV-2 genomes from clinical cases sampled since the beginning of the syndromic surveillance in France. We used the newly generated genomes to investigate the origins of SARS-CoV-2 lineages circulating in Northern France and better understand its spread.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results/Discussion

We generated complete SARS-CoV-2 genome sequences from nasopharyngeal or sputum samples addressed to the National Reference Center for Respiratory Viruses at the Institut Pasteur in Paris as part of the ongoing surveillance (Figure 1A). We combined the 97 SARS-CoV-2 genomes from France and 3 from Algeria generated here with 338 sequences published or freely available from the GISAID database, and performed a phylogenetic analysis focusing on the initial introductions and spread of the virus in France.

Early introductions do not appear to have resulted in local transmission

Our analysis indicates that the quarantine imposed on the initial COVID-19 cases in France appears to have prevented local transmission. The first European cases sampled on January 24, 2020 (IDF0372 and IDF0373 from Île-de-France, described in (Lescure, Bouadma et al. 2020) were direct imports from Hubei, China, and the genomes fall accordingly near the base of the tree, within clade V, according to GISAID nomenclature (Figure 2, Figure 3A). These identical genomes both harbor a V367F (G22661T) mutation in the receptor binding domain of the Spike, not observed in other genomes. Similarly, IDF0515 corresponds to a traveler from Hubei, China. This basal genome falls outside of the three major GISAID proposed clades V, G, and S (Figure 2, Figure 3A), but carries the G11083T mutation associated with putative lineage V1 (Figure 3A, Figure S2), suggesting convergent evolution or more likely a reversion of the V-clade defining G26144T change. Subsequent early cases in the West or East of France (B2334/B2340, clade V and GE1583, clade S), all with recent history of travel to Italy, add to the genomic diversity of viruses from Northern Italy, but also do not appear to have seeded local transmission with the current sampling (Figure 2).

The current outbreak lineages

All other sequences from Northern France fall in clade G (defined by a single non-synonymous mutation, D614G (A23403G) in the Spike, Figure 2), and this includes sequences captured during the steep increase of reported cases in many strongly affected regions (Figure 1). While a more thorough sampling will be needed to confirm this observation, it suggests that, unlike what is observed for many other European countries (Gudbjartsson, Helgason et al. 2020, Zehender, Lai et al. 2020), the French outbreak has been mainly seeded by one or several variants of this clade. This clade can be further classified into lineages, albeit supported again by only 1 to 3 substitutions (putatively named G1, G2, G3, G3a, G3b), and the diversity of the sequence from Northern France is spread out, with most regions represented in the different

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lineages. Several genomes correspond to patients with recent history of travel in Europe (GE3067, N1620, IDF2792), United Arab Emirates (IDF2936), Madagascar (HF1993) or Egypt (B1623, B2330), and might represent additional introductions of the same clade. On the other hand, in lineage G3b, three sequences sampled in Algeria are closely related to French sequences and likely represent exported cases in light of recent history of travel to France.

The syndromic surveillance allowed to capture one of the earliest representatives of clade G (HF1463, sampled on February 19th) (Figure 2). Importantly, this sequence carries 2 additional mutations compared to the reconstructed ancestral sequence of this clade (Figure 3B). Other sequences sampled weeks later (IDF2849, GE1973) are more basal to the clade, highlighting the complexity and risk of inferences based on 1 or 2 nucleotide substitutions. Because of this, and the scarcity of early sequences in many countries in Europe, country and within-country level phylogeographic estimations for the origin of clade G are also unreliable with the current dataset.

Crucially, while all early symptomatic suspected COVID-19 cases were addressed to the NRC for testing, this was no longer the case as the epidemic developed (Figure 1A). In addition, pauci or asymptomatic cases are scarcely represented here. As the earliest representative of clade G (HF1463) had no history of travel or contact with returning travelers, we can infer that the virus was silently circulating in France in February, a scenario compatible with the large proportion of mild or asymptomatic diseases (Li, Pei et al. 2020), and observations in other European countries (Gudbjartsson, Helgason et al. 2020, Onder, Rezza et al. 2020). While this is also compatible with the time to the most recent common ancestor estimate for clade G (Figure 2), the current sampling clearly prevents reliable inference for the timing of introduction in France.

This study reveals areas for potential improvement of SARS-CoV-2 genomic surveillance in France. Several regions are poorly represented yet, likely due to the heavy burden on hospitals, which were quickly able to perform local testing as the molecular detection tools were rapidly shared by the NRC. Because of this, and of the syndromic-only based surveillance, we likely underestimate the genetic diversity of SARS-CoV-2 circulating in France.

In conclusion, our study sheds light on the origin and diversity of the COVID-19 outbreak in France with insights for Europe, and highlights the challenges of containment measures when a significant proportion of cases are asymptomatic.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Materials and Methods

Ethical statement

Samples used in this study were collected as part of approved ongoing surveillance conducted by the National Reference Center for Respiratory Viruses (NRC) at Institut Pasteur (WHO reference laboratory providing confirmatory testing for COVID-19). The investigations were carried out in accordance with the General Data Protection Regulation (Regulation (EU) 2016/679 and Directive 95/46/EC) and the French data protection law (Law 78–17 on 06/01/1978 and Décret 2019–536 on 29/05/2019).

Sample collection

After reports of severe pneumonia in late December 2019, enhanced surveillance was implemented in France to detect suspected infections. For each suspected case, respiratory samples from the upper respiratory tract (nasopharyngeal swabs or aspirates) and when possible from the lower respiratory tract, were sent to the NRC, to perform SARS-CoV-2-specific real-time RT-PCR. Demographic information, date of illness onset, and travel history, were obtained when possible. A subset of samples were selected according to the viral load and their sampling location in order to have a broad representation across different regions of France.

Molecular test

RNA extraction was performed with the Extraction NucleoSpin Dx Virus kit (Macherey Nagel). RNA was extracted from 100 µl of specimen, eluted in 100 µl of water and used as a template for RT-qPCR. Samples were tested with a one-step RT-qPCR using three sets of primers as described on the WHO website (https://www.who.int/docs/default-source/coronaviruse/real-time-rt-pcr-assays-for-the-detection-of-sars-cov-2-institut-pasteur-paris.pdf?sfvrsn=3662fcb6_2).

Virus Sequencing

Viral genome sequences were generated by two different approaches. The first consisted in direct metagenomic sequencing, which resulted in complete or near complete genome sequences for samples with viral load higher than 1.45×10^4 viral genome copies/ul, which corresponds to a Ct value of 25.6 with the IP4 primer set (Figure 1B and Supplementary material, Table S3). Briefly, extracted RNA was first treated with Turbo DNase (Ambion) followed by purification using SPRI beads RNA clean XP. RNA was converted to double

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

stranded cDNA. Libraries were then prepared using the Nextera XT DNA Library Prep Kit and sequenced on an Illumina NextSeq500 (2x150 cycles) on the Mutualized Platform for Microbiology (P2M) at Institut Pasteur.

For samples with lower viral load, we implemented a highly multiplexed PCR amplicon approach (Quick, Grubaugh et al. 2017) using the ARTIC Network multiplex PCR primers set v1 (https://artic.network/ncov-2019), with modification as suggested in (Kentaro, Tsuyoshi et al. 2020). Synthesized cDNA was used as template and amplicons were generated using two pooled primer mixtures for 35 rounds of amplification. We prepared sequencing libraries using the NEBNext Ultra II DNA Library Prep Kit for Illumina and barcoded with NEBNext Multiplex Oligos for Illumina (Dual Index Primers Set 1) (New England Biolabs). We sequenced prepared libraries on an Illumina MiSeq using MiSeq Reagent Kit v3 (2x300 cycles) at the biomics platform of Institut Pasteur.

Genome assembly

Raw reads were trimmed using Trimmomatic v0.36 (Bolger, Lohse et al. 2014) to remove Illumina adaptors and low quality reads, as well as primer sequences for samples sequenced with the amplicon-based approach. We assembled reads from all sequencing methods into genomes using Megahit, and also performed direct mapping against reference genome Wuhan/Hu-1/2019 (NCBI Nucleotide – NC_045512, GenBank –MN908947) using the CLC Genomics Suite v5.1.0 (QIAGEN). We then used SAMtools v1.3 to sort the aligned bam files and generate alignment statistics (WysokerA, RuanJ et al.). Aligned reads were manually inspected using Geneious prime (2020) (https://www.geneious.com/), and consensus sequences were generated using a minimum of 3X read-depth coverage to make a base call. No genomic deletions were detected in the genomes analyzed.

Phylogenetic analysis

A set of 100 SARS-CoV-2 sequences generated in this study (97 from France, 3 from Algeria) was complemented with 338 genomes published or freely available sequences on GenBank or the GISAID database. From the latter, only published sequences were chosen (Deng, Gu et al. 2020, Fauver, Petrone et al. 2020) (Supplementary material, Table S2). A total of 438 full genome sequences were analyzed with augur and auspice as implemented in the Nextstrain pipeline (Hadfield, Megill et al. 2018) version from March 20, 2020 (https://github.com/nextstrain/ncov). Within the pipeline, sequences were aligned to the reference Wuhan/Hu-1/2020 strain of SARS-CoV-2 (GenBank accession MN908947) using

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MAFFT (Katoh and Toh 2010). The alignment was visually inspected and French sequences were subset to analyze shared SNPs. No evidence of recombination was detected using RDP4 (Martin, Murrell et al. 2015). A Maximum likelihood phylogenetic tree was built using IQ-TREE with the GTR model (Kalyaanamoorthy, Minh et al. 2017), after masking 130 and 50 nucleotides from the 5' and 3' ends of the alignment, respectively, as well as single nucleotides at positions 18529, 29849, 29851, 29853 as normally set in Nextstrain implementation for SARS-CoV-2. We checked for temporal signal using Tempest v1.5.3 (Rambaut, Lam et al. 2016). The temporal phylogenetic analyses were performed with augur and TreeTime (Sagulenko, Puller et al. 2018), assuming clock rate of 0.0008±0.0004 (SD) substitutions/site/year (Rambaut 2020), coalescent skyline population growth model and the root set on the branch leading to the Wuhan/Hu-1/2020 sequence. The time and divergence trees were visualized with FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). Nucleotide substitutions from the reference sequence that define internal nodes of the tree were extracted from the final Nextstrain build file and annotated on the tree using a custom R script (https://www.R-project.org/) with packages *tidyverse* v1.3.0 (Wickham, Averick et al. 2019) and *jsonlite* v1.6.1 (Ooms 2014). Adobe Illustrator 2020 was used to prepare final tree figures. Sequence metadata (Supplementary material, Table S2), Nextstrain build, and R script is available at https://github.com/Simon-LoriereLab/SARS-CoV-2-France. The phylogeny can be visualized interactively at https://nextstrain.org/community/Simon-LoriereLab/SARS-CoV-2-France. In this study, we used the proposed nomenclature from GISAID to annotate three major clades V, G, and S according to specific single-nucleotide polymorphisms that are shared by all sequences in the clade. Clade defining variants according to GISAID nomenclature are included in Supplementary material, Table S1.

Data sharing

The assembled SARS-CoV-2 genomes generated in this study were deposited on the GISAID database (https://www.gisaid.org/) as soon as they were generated, accession numbers can be found in Supplementary material (Table S2).

Acknowledgements

We would like to thank all of the health care workers, public health employees, and scientists involved in the COVID-19 response. We acknowledge the hospital laboratories from the RENAL network in the north of France (list of names in Supplementary material, Table S4). We acknowledge the authors, originating and submitting laboratories of the sequences from GISAID

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and GenBank (Supplementary material, Table S2). We avoided any direct analysis of genomic data not submitted as part of this paper and used this genomic data only as background. This study has received funding from Institut Pasteur, CNRS, Université de Paris, Santé publique France, the French Government's Investissement d'Avenir program, Laboratoire d'Excellence "Integrative Biology of Emerging Infectious Diseases" (grant n°ANR-10-LABX-62-IBEID), REACTing (Research & Action Emerging Infectious Diseases), France Génomique (ANR-10-INBS-09-09), IBISA, and the EU grant Recover. ESL acknowledges funding from the INCEPTION program (Investissements d'Avenir grant ANR-16-CONV-0005). We thank Laurence Ma (Biomics Platform, C2RT, Institut Pasteur, Paris, France) for the MiSeq sequencing. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris. FG is part of the Pasteur-Paris University (PPU) International PhD program, BioSPC doctoral school.

Figures

Figure 1. SARS-CoV-2 genome sequencing effort in the Northern French regions in a rapidly growing pandemic.

A. The plot represents the numbers of daily sequenced genomes in this study (red filled or hollow circles) overlaid with the number of reported positive cases (grey circles) obtained from Santé Publique France (www.santepubliquefrance.fr). Hollow circles indicate samples obtained on dates with zero reported positive cases. The data are shown separately for each region of Northern France as indicated on the map on the right. **B.** Percentage of SARS-CoV-2 genome coverage obtained in relation to the Ct values for the 97 genomes reported here. Colors indicate sequencing approach: untargeted metagenomics (green) or amplicon-based sequencing (red).

Figure 2. Northern France sequences from early introductions and currently circulating lineages.

Time calibrated tree of 438 SARS-CoV-2 sequences including Northern France, Algeria and publicly available global sequences. The tree is rooted using the reference strain Wuhan/Hu-1/2019n (MN908947). The tips of the tree are shaped and colored according to sampling location. Branch lengths are proportional to the time span from the sampling date to the inferred date of the most recent common ancestor. The three major clades according to GISAID nomenclature are indicated. Strain names of the sequences discussed in this study are indicated next to the corresponding tips.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. Divergence of SARS-CoV-2 sequences from Northern France.

Divergence tree of SARS-CoV-2 genomes including sequences from Northern France with the three major clades (**A**) or only clade G (**B**). The tips of the tree are shaped and colored according to sampling location. Branch lengths are proportional to the number of nucleotide substitutions from the reference and tree root Wuhan/Hu-1/2019 (MN908947). GISAID clades and putative lineages are indicated on the right of each panel. Strain names of the sequences discussed in this study are indicated next to the corresponding tips in italic. Nucleotide substitutions shared among all the sequences of each clade or lineage are indicated next to the corresponding nodes. Some monophyletic lineages are collapsed for ease of representation. A complete tree is shown in Figure S1.

Supplementary figures

Figure S1**. Phylogenetic divergence tree of all labeled SARS-CoV-2 sequences used in this study.**

Maximum-likelihood tree including all sequences from Northern France, Algerian sequences and publicly available global SARS-CoV-2 sequences, corresponding to the collapsed tree shown in Figure 3 (same ordering as in Figure 2 and Figure 3). GISAID clades are indicated next to the corresponding nodes and branches are colored distinctly. Tips indicate strain names, colored in red for sequences from France, and are noted in bold if discussed in this study.

Figure S2**. Single-nucleotide polymorphisms representing the diversity among French sequences across the regions of Northern France.**

Multiple sequence alignment of all SARS-CoV-2 genomes sampled across the northern part of France from different clades and lineages. Single nucleotide variants with respect to the reference (MN908947) are shown as black vertical bars and shared substitutions among the sequences of each clade or lineage are annotated. A substitution only found in sequences from Normandie is noted in italic.

Supplemental material

A single Excel document with multiple sheets, representing tables below.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table S1. Clade defining SNPs according to GISAID nomenclature.

Table S2. Metadata associated with the sequences used in this study.

Table S3. Viral RNA load and genome recovery data.

Table S4. List of collaborators in the RENAL network in the north of France.

References

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics **30**(15): 2114-2120.
Deng, X., W. Gu, S. Federman, L. Du Plessis, O. Pybus, N. Faria, C. Wang, G. Yu, C.-Y. Pan and H. Guevara (2020). "A Genomic Survey of SARS-CoV-2 Reveals Multiple Introductions into Northern California without a Predominant Lineage." medRxiv.
Fauver, J. R., M. E. Petrone, E. B. Hodcroft, K. Shioda, H. Y. Ehrlich, A. G. Watts, C. B. Vogels, A. F. Brito, T. Alpert and A. Muyombwe (2020). "Coast-to-coast spread of SARS-CoV-2 in the United States revealed by genomic epidemiology." medRxiv.
Gardy, J. L. and N. J. Loman (2018). "Towards a genomics-informed, real-time, global pathogen surveillance system." Nature Reviews Genetics **19**(1): 9.
Grubaugh, N. D., J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes and K. G. Andersen (2019). "Tracking virus outbreaks in the twenty-first century." Nature microbiology **4**(1): 10-19.
Guan, W.-j., Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei and D. S. Hui (2020). "Clinical characteristics of coronavirus disease 2019 in China." New England Journal of Medicine.
Gudbjartsson, D. F., A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem and A. B. Agustsdottir (2020). "Spread of SARS-CoV-2 in the Icelandic Population." New England Journal of Medicine.
Gudbjartsson, D. F., A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem, A. B. Agustsdottir, B. Eiriksdottir, R. Fridriksdottir, E. E. Gardarsdottir, G. Georgsson, O. S. Gretarsdottir, K. R. Gudmundsson, T. R. Gunnarsdottir, A. Gylfason, H. Holm, B. O. Jensson, A. Jonasdottir, F. Jonsson, K. S. Josefsdottir, T. Kristjansson, D. N. Magnusdottir, L. le Roux, G. Sigmundsdottir, G. Sveinbjornsson, K. E. Sveinsdottir, M. Sveinsdottir, E. A. Thorarensen, B. Thorbjornsson, A. Love, G. Masson, I. Jonsdottir, A. D. Moller, T. Gudnason, K. G. Kristinsson, U. Thorsteinsdottir and K. Stefansson (2020). "Spread of SARS-CoV-2 in the Icelandic Population." N Engl J Med.
Hadfield, J., C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford and R. A. Neher (2018). "Nextstrain: real-time tracking of pathogen evolution." Bioinformatics **34**(23): 4121-4123.
Kalyaanamoorthy, S., B. Q. Minh, T. K. Wong, A. von Haeseler and L. S. Jermiin (2017). "ModelFinder: fast model selection for accurate phylogenetic estimates." Nature methods **14**(6): 587.
Katoh, K. and H. Toh (2010). "Parallelization of the MAFFT multiple sequence alignment program." Bioinformatics **26**(15): 1899-1900.
Kentaro, I., S. Tsuyoshi, H. Masanori, T. Rina and K. Makoto (2020). "Aproposal ofan alternative primer fortheARTIC Network's multiplex PCRto improve coverageof SARS-CoV-2 genomesequencing." bioRxiv
Ladner, J. T., N. D. Grubaugh, O. G. Pybus and K. G. Andersen (2019). "Precision epidemiology for infectious disease control." Nature medicine **25**(2): 206-211.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lescure, F.-X., L. Bouadma, D. Nguyen, M. Parisey, P.-H. Wicky, S. Behillil, A. Gaymard, M. Bouscambert-Duchamp, F. Donati and Q. Le Hingrat (2020). "Clinical and virological data of the first cases of COVID-19 in Europe: a case series." The Lancet Infectious Diseases.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang and J. Shaman (2020). "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)." Science.

Martin, D. P., B. Murrell, M. Golden, A. Khoosal and B. Muhire (2015). "RDP4: Detection and analysis of recombination patterns in virus genomes." Virus evolution **1**(1).

Onder, G., G. Rezza and S. Brusaferro (2020). "Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy." Jama.

Ooms, J. (2014). "The jsonlite package: A practical and consistent mapping between json data and r objects." arXiv preprint arXiv:1403.2805.

Quick, J., N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara, Jr., E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen and N. J. Loman (2017). "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples." Nat Protoc **12**(6): 1261-1276.

Rambaut, A. (2020). "Phylogenetic analysis of nCoV-2019 genomes." Retrieved 6-Mar-2020.

Rambaut, A., T. T. Lam, L. Max Carvalho and O. G. Pybus (2016). "Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)." Virus evolution **2**(1): vew007.

Sagulenko, P., V. Puller and R. A. Neher (2018). "TreeTime: Maximum-likelihood phylodynamic analysis." Virus evolution **4**(1): vex042.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry and J. Hester (2019). "Welcome to the Tidyverse." Journal of Open Source Software **4**(43): 1686.

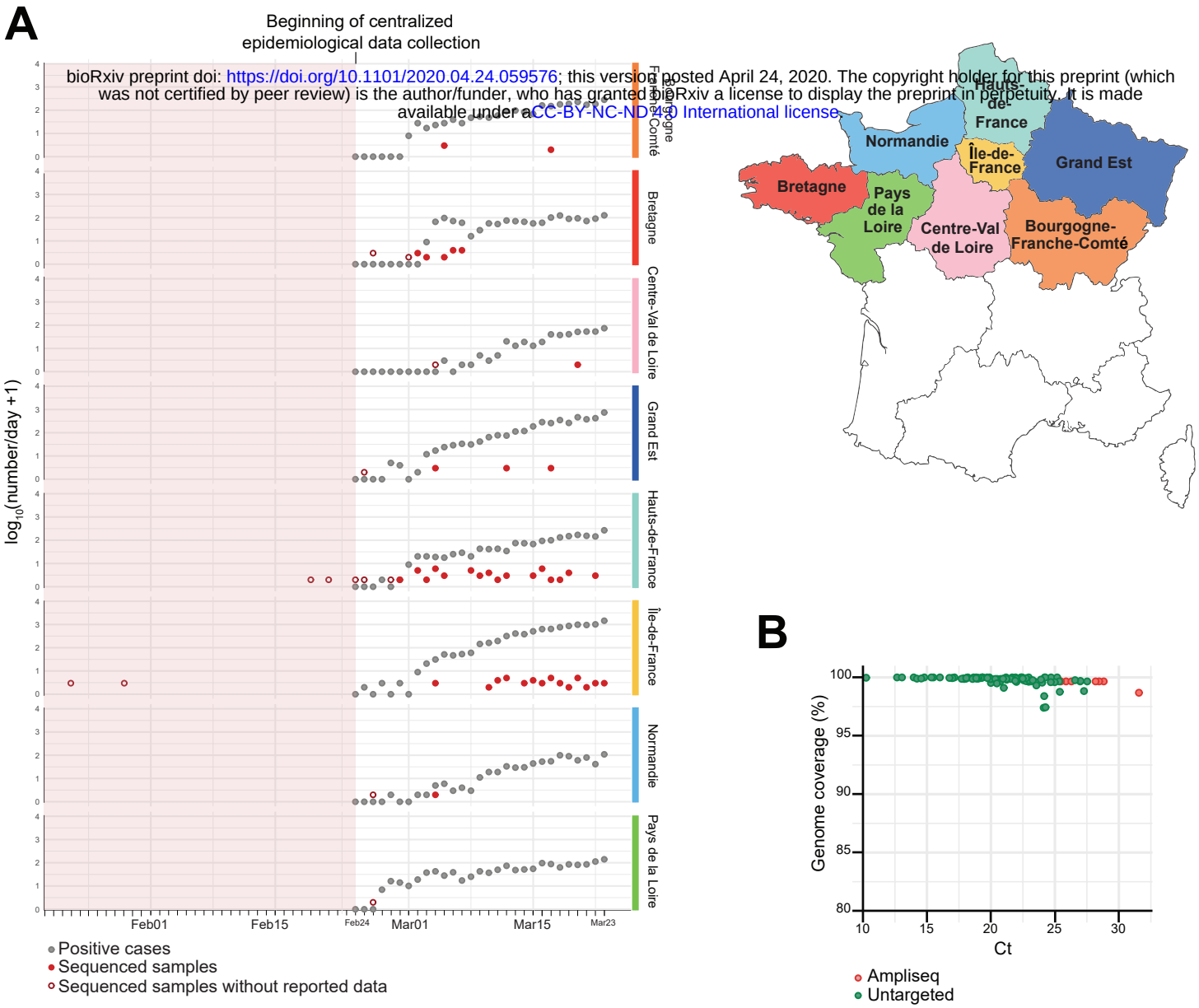
WysokerA, F., H. RuanJ and A. MarthG "DurbinR. 2009a. The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

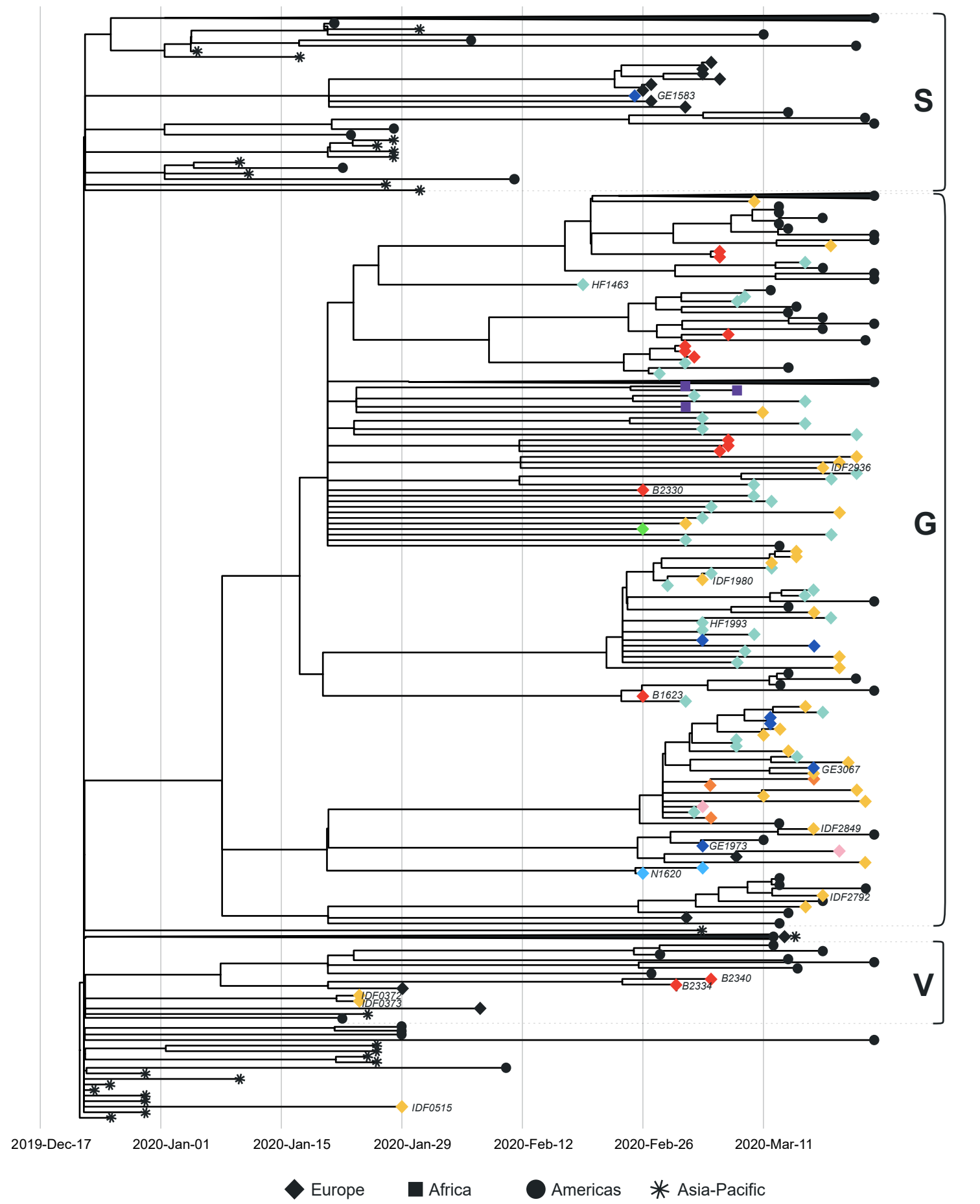
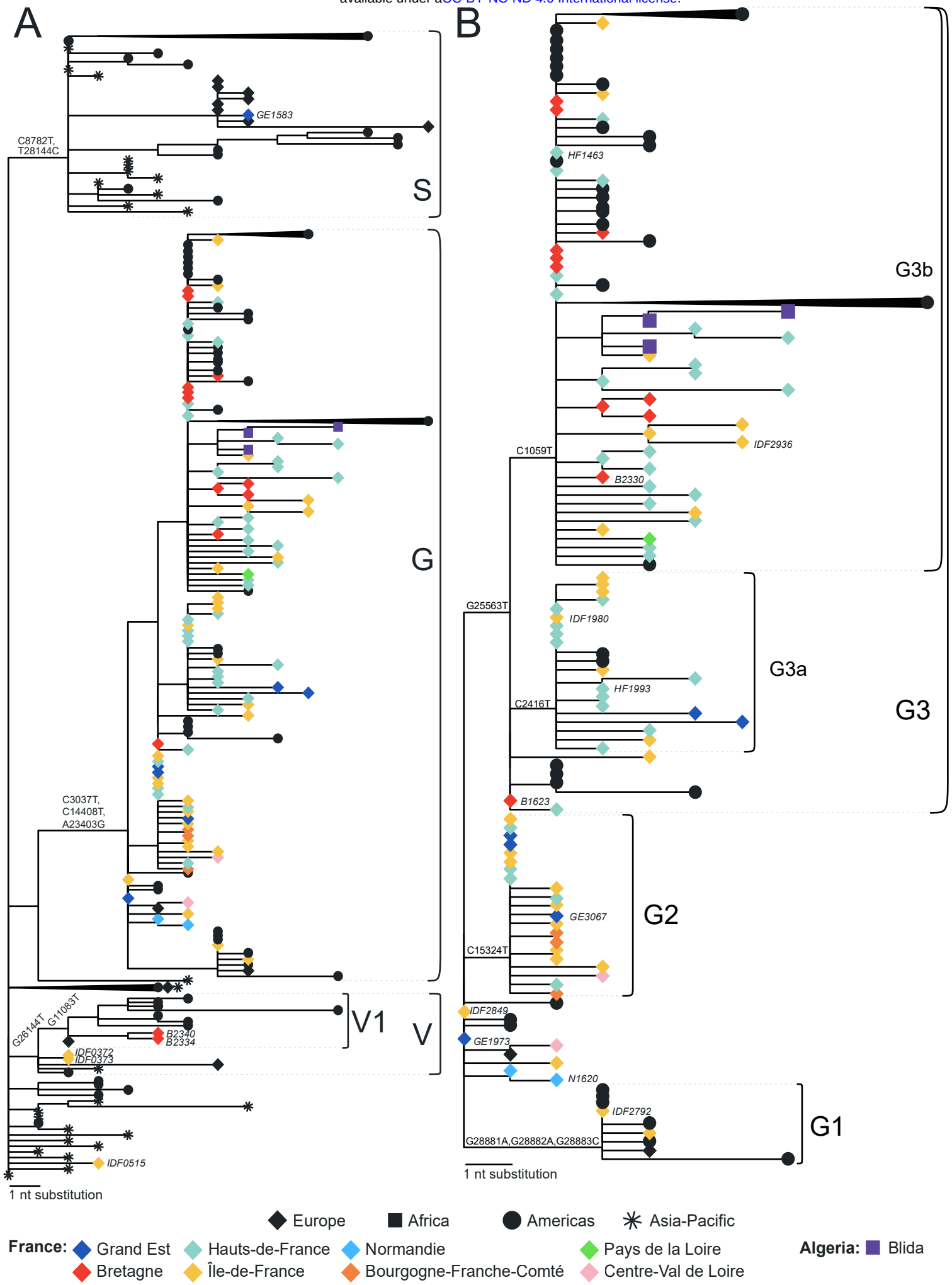
Zehender, G., A. Lai, A. Bergna, L. Meroni, A. Riva, C. Balotta, M. Tarkowski, A. Gabrieli, D. Bernacchia and S. Rusconi (2020). "Genomic Characterisation And Phylogenetic Analysis Of Sars-Cov-2 In Italy." Journal of Medical Virology.

Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi and R. Lu (2020). "A novel coronavirus from patients with pneumonia in China, 2019." New England Journal of Medicine.

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.24.059576; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.




S
G
V
GE1583
HF1463
IDF2936
B2330
IDF1980
HF1993
B1623
GE3067
IDF2849
GE1973
N1620
IDF2792
B2340
B2334
IDF0372
IDF0373
IDF0515
2019-Dec-17
2020-Jan-01
2020-Jan-15
2020-Jan-29
2020-Feb-12
2020-Feb-26
2020-Mar-11
Europe
Africa
Americas
Asia-Pacific
France:
Grand Est
Hauts-de-France
Normandie
Pays de la Loire
Bretagne
Île-de-France
Bourgogne-Franche-Comté
Centre-Val de Loire
Algeria:
Blida

available under aCC BY-NC-ND 4.0 International license.