Sequence evidence that the D614G clade of SARS-CoV-2 was already circulating in northern Italy in the fall of 2019

Xuhua Xia[1]

1 University of Ottawa

Abstract

The D614G clade is characterized by TTTG at four nucleotide sites (sites 241, 3037, 14408 and 23403 following the reference genome NC_045512), in contrast to CCCA shared among early SARS-CoV-2 genomes sampled in China and those that can be traced to China. It was believed that the TTTG lineage descended from the early viral CCCA lineages. A set of SARS-CoV-2 sequences collected from Sept. 12 to Dec. 18, 2019, in Lombardy, Milan and Turin in Italy provided, for the first time, strong evidence that the D614G/TTTG lineage has already been circulating in Italy in 2019.

Xuhua Xia[1,2]

[1] *Department of Biology, University of Ottawa, Ottawa, Canada K1N 6N5xxia@uottawa.ca*

[2] *Ottawa Institute of Systems Biology, Ottawa, Canada K1H 8M5*

Keywords: SARS-CoV-2; COVID-19; viral evolution; viral origin; viral dating; Italy.

Two key questions are asked following a viral outbreak, i.e., when and where the zoonotic or lab-leak event occurred. The "When" question is typically addressed by estimating the viral evolutionary rate and dating the most recent common ancestor (MRCA) of representative viral strains [1][2][3][4][5][6]. The "Where" question is approximated by the location where the earliest viral lineage was sampled [5][6][7]. Obviously, such large-scale analyses are better performed with full-length high-quality viral genomes than short sequence fragments. For SARS-CoV-2, there are currently more than 13 million genomic sequences. Thus, there seems no shortage of data for any large-scale data analysis. Indeed, the MRCA of SARS-CoV-2 genomes have been dated to the summer of 2019 with a gigantic tree of 455,251 leaves [6].

One major shortcoming of using full-length high-quality SARS-CoV-2 genomes is that these genomes are sampled after the viral outbreak in Wuhan in late December 2019, when SARS-CoV-2 has already spread globally. SARS-CoV-2 samples sequenced before the viral outbreak are present in GenBank only as short sequence fragments. However, these sequence fragments provide critical insights into the origin and early evolution of the D614G lineage.

The D614G clade differs from the early Wuhan strains at four nucleotide sites (sites 241, 3037, 14408 and 23403 following the reference genome NC_045512) [8]. The D614G strain has TTTG at these four sites, in contrast to CCCA shared among early SARS-CoV-2 genomes sampled in Wuhan. It was believed that the D614G/TTTG lineage descends from the CCCA lineage. However, sequence data from early SARS-CoV-2 samples collected in northern Italy (Lombardy, Milan and Turin regions) from Sept. 12 to Dec. 18, 2019, provided strong evidence that the D614G/TTTG lineage was already circulating in northern Italy in the fall of 2019.

Thirteen of these Italy-derived samples of SARS-CoV-2 were partially sequenced and deposited in GenBank (Table 1). These sequence fragments vary in length from 209 nt to 778 nt and do not cover the entire SARS-CoV-2 genome. However, they do cover three of the four critical sites distinguishing the D614G/TTTG lineage from the early CCCA lineage (Table 1). Site 3037 is represented by two sequences with a T, site 14408 by four sequences with a T, and site 23403 by one sequence with a G. This "?TTG" configuration is compatible with that of D614G/TTTG lineage but not with the early CCCA lineage. The inescapable conclusion is that the D614G/TTTG lineage has been circulating in northern Italy in 2019, concurrently with or even earlier than the CCCA lineages represented in early SARS-CoV-2 samples in Wuhan.

The evidence from these SARS-CoV-2 sequence data from northern Italy is consistent with the dating of the MRCA to the summer of 2019 [5][6]. It is also consistent with the twin-beginnings hypothesis on the origin, spread and evolution of SARS-CoV-2 [9].

These sequence fragments have been overshadowed by the huge number of full-length high-quality SARS-CoV-2 genomes and forgotten for too long. In spite of their high relevance, these sequence fragments (Table 1) have not been used in any previous studies addressing the "When" and "Where" questions, including the recent study by Pekar et al.[10]. Not only did Pekar et al.[10] not analyze these early SARS-CoV-2 sequence fragments, but they also omitted full-length high-quality SARS-CoV-2 genomes from the D614G clade sampled in January 2020 in both GenBank and GISAID (Table 2). While there is no guarantee that collection dates found in GenBank and GISAID records are accurate, it is not appropriate to omit crucially relevant data because such omission will almost certainly bias the estimation of the root of the tree. If one includes early samples in Location A but omits early samples from all other locations, then Location A will necessarily become associated with the earliest SARS-CoV-2 lineage.

Fig. 1 is a phylogenetic tree including early SARS-CoV-2 genomes from Europe and USA, together with representatives of lineage A and lineage B samples from China. There were two early but now obsolete reasons for placing the root within the shaded clade (Fig. 1). First, the shaded clade represents the earliest SARS-CoV-2 samples from Wuhan and those that can be traced to China. Second, Huanan Seafood Market is likely where the zoonotic transmission occurred. Associated with these two reasons is the assumption that SARS-CoV-2 was sampled and sequenced right after its presumed zoonotic transmission [11]. The specific placement of the root at the red dot or within the red clade was tenuously based on two nucleotides at sites 8782 and 28144[11][12]. Lineage A (represented by the red clade in Fig. 1) has T and C at these two sites, and the close relatives of bat-derived viruses such as RaTG13 also have T and C at these two sites. This TC configuration is therefore assumed to be ancestral. In contrast, the CT configuration at these two sites observed in lineage B, which was sampled earlier than lineage A and originally represented by the non-red lineages within the shaded clade (Fig. 1), was deemed derived. Pekar et al.[10] did briefly mention an alternative

hypothesis that the TC configuration may not be ancestral because of the existence of SARS-CoV-2 with TT and CC configurations at these two sites indicating an evolutionary trajectory of CT (lineage B) (TT or CC) TC (lineage A).

All these interpretations fall apart given the evidence of early circulation of the D614G strain in Table 1. The root of SARS-CoV-2 is more likely at the black dot than at the red dot (Fig. 1). This new root immediately answers two key questions raised by Pekar et al.[10] at beginning of their paper: (i) why were lineage B viruses detected earlier than lineage A viruses and (ii) why did lineage B predominate early in the pandemic? The answers to these two questions are obvious given the root at the black dot, but become difficult if the root is at the red dot as assumed by Pekar et al.[10].

Sequence data currently available is inadequate for one to infer when and where the D614G/TTTG strain originated. Although D614G with the TTTG configuration was not found in China except for imported cases, viral genomes with TTCG configuration were later found in China through retrospective sequencing[13]. What is clear is that viral lineages with the TTTG/TTCG configuration coexisted with the CCCA strain at the time when the latter caused the viral outbreak in Wuhan. Both the TTTG clade and the CCCA clade are likely descendants of a much earlier common ancestor. If scientists around the world would respond to WHO's call to sequence archived samples, then there is a chance to find when and where SARS-CoV-2 originated, either through a natural zoonotic event or a lab-leak event. Limiting the search for a common ancestor within the early Wuhan clade is reminiscent of a mink government insisting that the common ancestor of SARS-CoV-2 is within the mink farm in the Netherlands where the first SARS-CoV-2 outbreak was recorded (Oude Munnink *et al.*, [14]).

Figures and Tables

Table 1. All SARS-CoV-2 sequences sampled in 2019 in Lombardy and nearby Turin in Italy and deposited in GenBank. The sequence ID is in the form of Accession_Collection date_Location. No sequence includes site 241 but some include sites 3037, 14408 and 23403.

Sequence ID	Start[1]	End[1]	3037	14408	23403
MZ223388_2019-10-22_Lombardy	2977	3185	T		
MZ223391_2019-11-22_Lombardy	2977	3185	T		
MZ223389_2019-10-17_Lombardy	14366	14655		T	
MZ223387_2019-10-19_Lombardy	14366	14655		T	
MZ223386_2019-10-23_Lombardy	14366	14655		T	
MZ223392_2019-12-15_Lombardy	14366	14655		T	
MZ223393_2019-12-15_Lombardy	22904	23681			G
MW303957_2019-12-05_Milan	22935	23343			
MZ223385_2019-09-12_Lombardy	22904	23330			
MZ223390_2019-10-12_Lombardy	22904	23330			
MT843234_2019-12-18_Milan	18484	18770			
MT843235_2019-12-18_Turin	18484	18770			
MT843236_2019-12-18_Turin	18484	18770			

[1] *Start and ending site of the sequence with site numbering according to the reference genome (NC_045512)*

Table 2. Early SARS-CoV-2 samples in the G614G lineage with collection date (Coll. Date), country, and Type (nucleotide configuration at sites 241, 3037, 14408 and 23403) that are not used in Pekar et al.[10].

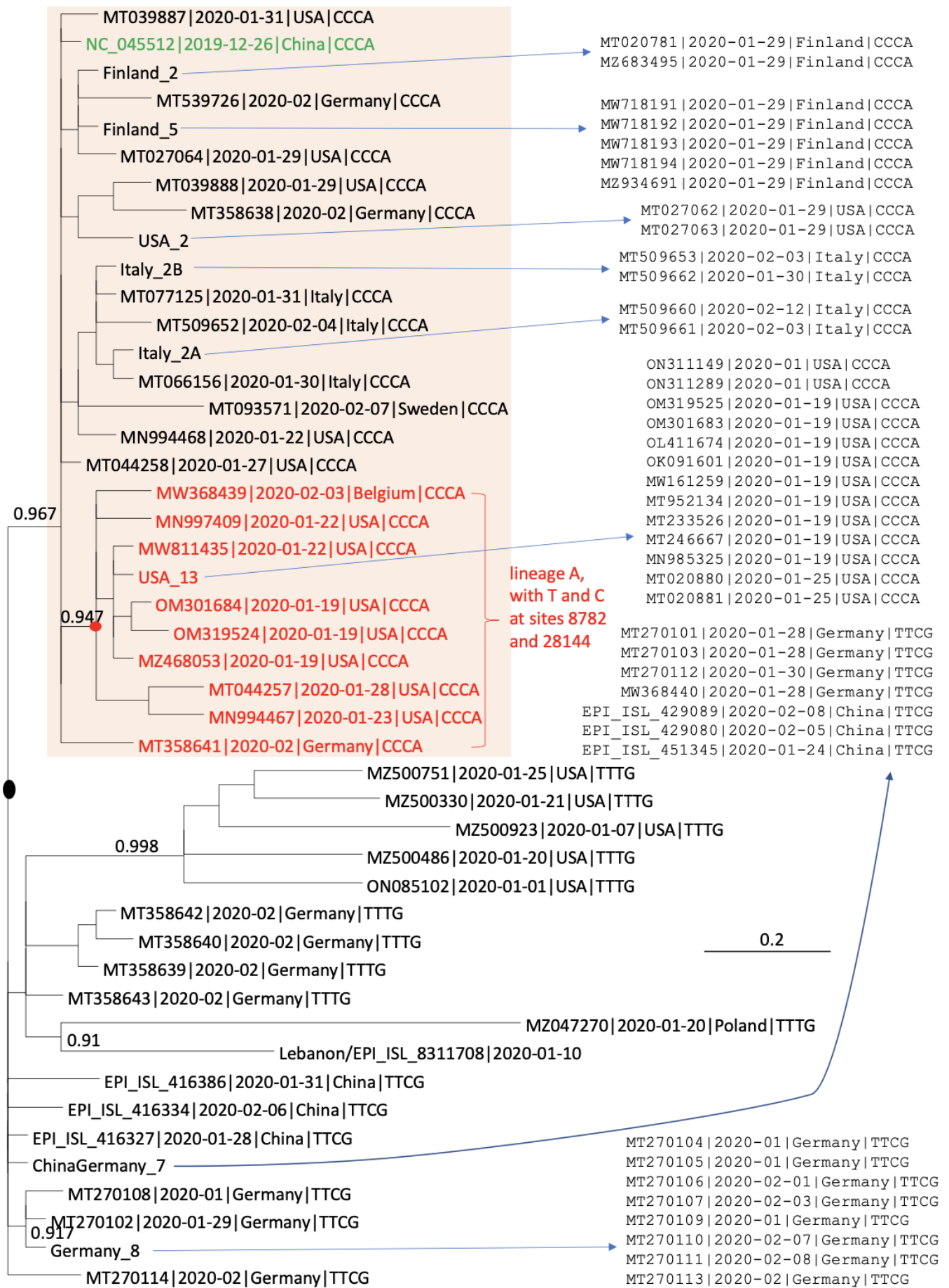| Accession	Coll. Date	Country	Type
ON085102	2020-01-01	USA	TTTG
MZ047270	2020-01-20	Poland	TTTG
MZ500923	2020-01-07	USA	TTTG
MZ500486	2020-01-20	USA	TTTG
MZ500330	2020-01-21	USA	TTTG
MZ500751	2020-01-25	USA	TTTG
EPI_ISL_8311708	2020-01-10	Lebanon	TTTG

Fig. 1. Phylogenetic tree of early lineages of SARS-CoV-2 isolated in Europe (excluding UK) and USA, together with one early D614G genome from Lebanon and seven key representative SARS-CoV-2 genomes from China, including the reference genome (NC_045512, colored green). All genomes have at least 29740 nt, and "human" as host. Collection dates were up to February 2020 for European sequences, and by Jan. 31, 2020 for USA sequences. OTU names are in the form of accession|collection date|country|type where "accession" is either GenBank or GISAID accession and "type" is defined by the four nucleotides at sites 241, 3037, 14408 and 23403 that separate the D614G (TTTG) lineage from those early strains isolated in China (CCCA). Viral genomes missing any of these four sites were excluded. Lineage A is characterized by C8782T and T28144C. The shaded area includes the original viral classification of lineage A and lineage B. Genomes were aligned by MAFFT[15] with the FFT-NS-2 option. The unrooted phylogenetic tree was reconstructed with PhyML [16], with a GTR model and optimization of topology, branch lengths and rates.

References

1. ^Oscar A. MacLean, Spyros Lytras, Steven Weaver, Joshua B. Singer, Maciej F. Boni et al. (2021). Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. PLOS Biology. 19(3):e3001115. doi:10.1371/journal.pbio.3001115.

2. ^Hongru Wang, Lenore Pipes, Rasmus Nielsen. (2021). Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. Virus evolution. 7:veaa098. doi:10.1093/ve/veaa098.

3. ^Maciej F. Boni, Philippe Lemey, Xiaowei Jiang, Tommy Tsan-Yuk Lam, Blair Perry et al. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat Microbiol. 5:1408–1417. doi:10.1101/2020.03.30.015008.

4. ^S. Lytras, J. Hughes, D. Martin, Klerk Arné de, Lourens Rentia et al. (2021). Exploring the natural origins of SARS-CoV-2 in the light of recombination. bioRxiv (accessed on Sept 1, 2021). doi:10.1101/2021.01.22.427830 ,.

5. a, b, c Xuhua Xia. (2021). Dating the Common Ancestor from an NCBI Tree of 83688 High-Quality and Full-Length SARS-CoV-2 Genomes. Viruses. 13(9):1790. doi:10.3390/v13091790.

6. a, b, c, d X. Xia. (2022). Improved method for rooting and tip-dating a viral phylogeny. In: H. H.-S. Lu, B. Scholkopf, M. T. Wells, H. Zhao, editors. Handbook of Computational Statistics, II. Berlin: Springer (In press). p.

7. ^M. T. Gilbert, A. Rambaut, G. Wlasiuk, T. J. Spira, A. E. Pitchenik et al. (2007). The emergence of HIV/AIDS in the Americas and beyond. Proceedings of the National Academy of Sciences of the United States of America. 104(47):18566-18570. PubMed PMID: 17978186.

8. ^B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler et al. (2020). Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 182(4):812-827 e819. doi:10.1016/j.cell.2020.06.043. PubMed PMID: 32697968; PubMed Central PMCID: PMC7332439.

9. ^Yongsen Ruan, Haijun Wen, Mei Hou, Ziwen He, Xuemei Lu et al. (2022). The twin-beginnings of COVID-19 in Asia and Europe—one prevails quickly. Natl Sci Rev. 9(4):nwab223. doi:10.1093/nsr/nwab223.

10. a, b, c, d, e, f Jonathan E. Pekar, Andrew Magee, Edyth Parker, Niema Moshiri, Katherine Izhikevich et al. (2022). The

molecular epidemiology of multiple zoonotic origins of SARS-CoV-2. Science. 377(6609):960-966. doi:doi:10.1126/science.abp8337.

11. [a, b]Andrew Rambaut, Edward C. Holmes, Verity Hill, Áine O'Toole, J. T. McCrone et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology. bioRxiv.2020.2004.2017.046086. doi:10.1101/2020.04.17.046086.

12. [^]Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao et al. (2020). On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 7(6):1012–1023. doi:10.1093/nsr/nwaa036.

13. [^]E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen et al. (2020). Evaluating the Effects of SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. Cell. 184:64-75. doi:10.1016/j.cell.2020.11.020. PubMed PMID: 33275900; PubMed Central PMCID: PMC7674007.

14. [^]Bas B. Oude Munnink, Reina S. Sikkema, David F. Nieuwenhuijse, Robert Jan Molenaar, Emmanuelle Munger et al. (2021). Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science. 371(6525):172. doi:10.1126/science.abe5901.

15. [^]K. Katoh, H. Toh. (2008). Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics. 9:212. doi:10.1186/1471-2105-9-212. PubMed PMID: 18439255; PubMed Central PMCID: PMC2387179.

16. [^]S. Guindon, O. Gascuel. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52(5):696-704.