SARS-COV-2 TRANSMISSION CHAINS FROM GENETIC DATA: A DANISH CASE STUDY

ANDREAS BLUHM, MATTHIAS CHRISTANDL, FULVIO GESMUNDO, FREDERIK RAVN KLAUSEN, LAURA MANČINSKA, VINCENT STEFFAN, DANIEL STILCK FRANÇA, ALBERT H. WERNER

Abstract. *Background:* The Covid-19 pandemic caused by the SARS-CoV-2 virus started in China in December and has since spread globally. Many countries have instated measures to slow the spread of the virus. Information about the introduction of the virus to a country and its further spread can inform the gradual opening of a country and the avoidance of a second wave of infections. Denmark has seen first cases in late February and is currently opening up.

Methods: Sequenced virus genome can be used to reconstruct transmission events. We perform a phylogenetic analysis of 742 publicly available Danish SARS-CoV-2 sequences and put them into context using sequences from other countries.

Results: Our findings are consistent with several introductions of the virus to Denmark from independent sources, the majority from a ski area in Austria. We identify several chains of mutations that occurred in Denmark and in at least one case find evidence that it spread from Denmark to other countries. There is a number of the Danish mutations that are non-synonymous, and in general there is a considerable variety of strains circulating in Denmark.

Conclusion: The introduction of the virus from the Austrian ski area happened after Iceland had declared this area high-risk, thereby giving an independent assessment of the effectiveness of the early identification of high-risk areas. The likely further spread of the virus from Denmark challenges the common narrative that Denmark only got infected from abroad, highlighting that Denmark was part of a network of countries among which the virus was being transmitted. Furthermore, our broad analysis of the mutations does not indicate that the virus underwent a systematic change in virulence in Denmark. We believe that the methods used can be a valuable tool to identifying and validating transmission chains during the reopening.

1. Introduction

The first cases of the Covid-19 pandemic were reported in the city of Wuhan (China) in December 2019 and a new virus, named SARS-CoV-2, was later identified as its origin [HWL+20]. At the time of writing, the pandemic is ongoing and has spread to more than 180 countries [WHO20b].

The first case in Europe was reported from France on January 24, 2020 [BSRS+20]. Italy confirmed the first two cases on January 31 [WHO20a]. Austria reported the first cases on February 25 [WHO20c]. In March, Europe was the center of the global pandemic with many European countries introducing lockdown measures and travel restrictions. Early on, the ski area of Ischgl in Tyrol, Austria, was identified as a

Date: May 29, 2020.

The authors are listed in alphabetical order.

transmission hot-spot, so Iceland already identified it as a risk area on March 5 [Emb]. Quarantine measures in Ischgl, however, were only imposed on March 13 [Amt].

The first case in Denmark was confirmed on February 27, after having returned from skiing holidays in Northern Italy on February 24 [TV2]. The second case was confirmed on February 28 after having entered Denmark from Italy on February 15 [Sunb]. The number of cases kept increasing, and, on March 8, when two cases at Rysensteen Gymnasium were confirmed [Dan], there was increased suspicion of community transmission within Denmark.

During this first phase of the pandemic in Denmark, there existed travel warnings to high risk areas. On March 2, Denmark advised against all travel to Northern Italy [Suna]. On March 10, Denmark additionally advised against travel to the Austrian state of Tyrol, as the country had seen many infected skiers returning from the ski village of Ischgl [Stad].

On March 11, the prime minister Mette Frederiksen announced a lockdown, which happened in several stages and included closures of borders and schools [Stae]. Overall, the measures were not as severe as in some other European countries. On April 6, the prime minister announced that a first phase of reopening would start after Easter from April 14 [Staf]. The country has opened up further since.

As of May 26, there were 11,511 confirmed infections and 567 deaths in Denmark in connection with the disease [Staa]. From March 12, only people with serious symptoms and people in risk groups were tested. Since April 1, the number of tests has been increased [Stab].

In this work, we study all the publicly available genome sequences of the SARS-CoV-2 from Denmark [HMB+18] as of May 26, and compare it sequences from abroad [HMB+18]. We use the mutations in the genomic data to identify transmission chains. We focus on chains highlighting the introduction of the virus to Denmark, its transmission within Denmark, and its spread to other countries.

2. Methods

We use publicly available sequenced genome of the Sars-Cov-2 virus in order to identify mutations for the purpose of finding transmission chains. In the following, we describe how we obtained and analysed the sequences. For a flow chart of this process, see Figure 1.

2.1. **Acquisition of samples:** Sequences were downloaded from the GISAID EpiCoV database [EBM17, SM17] on May 26, 2020, including 742 Danish sequences. See the supplementary file for a full list of sequences including their origins. From the sequences, we selected those which we deemed of high quality and used them for our analysis. Specifically, this means that we only consider sequences with at least 29,000 nucleotides and at most 300 unidentified nucleotides (N's). Moreover, we only consider sequences that originate from a human host. After these steps, we were left with 583 Danish sequences, which we focus on in our analysis.

2.2. **Sequence alignment and data preparation.** We perform multiple sequence alignment (MSA) using the R-package DECIPHER [Wri15, Wri16]. We apply the function AlignSeqs from this package to our data until convergence is achieved.

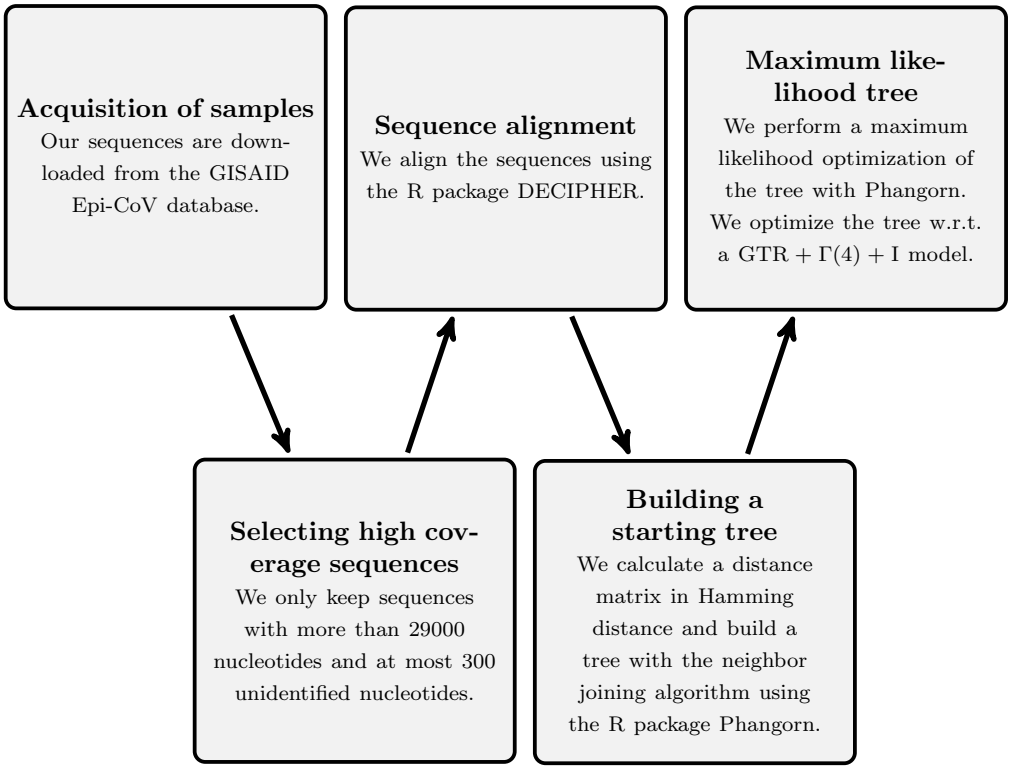


FIGURE 1. The tree building process as a flow chart.

For some of the larger data sets, we cluster identical sequences. For that, we build a distance matrix from the aligned data using the DECIPHER function DistanceMatrix. Subsequently, we use the distance matrix to identify clusters employing the IdClusters function, again from DECIPHER. We cluster identical sequences retaining only one representative sequence per cluster. We will remark whenever we use this procedure for clustering.

2.3. **Tree inference.** The tree topology was inferred using the R-package *phangorn* [Sch11]. From the aligned data, we build a distance matrix in Hamming distance using the function dist.hamming. In the next step, we build a phylogenetic tree from this distance matrix using the neighbor joining algorithm.

This tree is then used as the starting tree for a maximum likelihood optimization.

We run the modelTest function to estimate different models for the given tree and data. We optimize the parameters w.r.t. to a $\mathrm{GTR} + \Gamma(4) + \mathrm{I}$ model using the optim.pml function.

Tree visualization and annotation were done with the R-package *ggtree* [Yu20]. We plot the trees ignoring branch lengths to focus on the tree topology.

2.4. **Haplotypes and rooting conventions.** Our tree was rooted with respect to the reference sequence NC-045512.2 (SARS-CoV-2 isolate Wuhan-Hu-1). Haplotypes were subsequently inferred from the phylogenetic tree. We chose naming conventions in accordance with [GHJ+20] and Nextstrain [HMB+18].

To get an overview of haplotypes and mutations prevalent in Denmark, we identify positions where sufficiently many[1] of the analyzed sequences exhibit a substitution or a deletion as compared to the other sequences in the data set in question. We analyse the coincidence of the new mutations with previously identified haplotypes from [GHJ+20] and each other in the entire worldwide dataset.

2.5. **Other R-packages.** Most of our analysis uses the software R [R C17]. Except for the R-packages mentioned in the previous sections, we also use *stringr*, *dplyr*, *ggplot2* and *ape* [PS19].

3. Results

In this section, we review our results. We study three different types of mutations: First, we consider mutations which were present in some region of the world and appeared in Denmark at some point. Second, we look at chains of mutations which only appear in Denmark. Finally, we look at mutations, the majority of which occurs in Denmark. In the Discussion section, we analyse the spread of the virus to, within and from Denmark in light of these mutations.

3.1. **Mutations from other regions appearing in Denmark.** In the following we first discuss the most common haplotypes appearing in Denmark in relation to their appearance in other countries. We subsequently discuss specific examples of less prevalent mutations in Denmark that are present to a larger degree in other countries. For an overview we refer to Figure 6.

3.1.1. *A2a2a: A common mutation in Denmark and Ischgl.* Approximately 70% of the available Danish sequences have haplotype A2a2a. This makes it the most common haplotype in our Danish data with 405 out of 583 see Table 1. This haplotype was already reported in [GHJ+20, Figure 3B] where the authors point out that travelers from Austria had the haplotype A2a2 together with the mutation C1059T which is the definition of A2a2a. The haplotypes A2a2 and A2a2a correspond to amino acid changes Q57H and T265I in Orf3a and Orf1a respectively and have already been studied in [IMP+20].

The Austrian data presented in Figure 7 shows that the haplotype A2a2a is mostly present in sequences from the ski village of Ischgl in the region of Tyrol, Austria, and the adjacent region of Vorarlberg.

The Norwegian sequences have travel metadata and we see that four out of five sequences with the haplotype A2a2a also have recent travel history to Austria, whereas none of the other sequences is associated to these two mutations.

However, we see in our complete data set that 4343 out of 20254 sequences have the haplotype A2a2a (see Table 1). Hence, this haplotype is abundant also in other countries than the ones mentioned above (e.g. France and the UK).

On the other hand, some of the Danish sequences with A2a2a have additional mutations: Worldwide, 9 sequences have A2a2a and the mutation A6825C which

[1]To get a better overview and readability we choose different thresholds depending on the context.

corresponds to the amino acid change N2187T in Orf1a. Among those are the 6 Danish ones and the others are from Austria, Norway and Scotland. The Norwegian sequence can be traced with metadata to Austria and the Austrian one is from Ischgl. Similarly, the mutation G15380T (corresponding to S5039L in Orf1a) appears with haplotype A2a2a in 25 sequences in the world. Among those 25 are the 12 Danish ones. Of the remaining ones with A2a2a, there are 8 Austrian sequences. All of them stem from the region Tyrol and in particular 6 are from Ischgl.

In summary, the most common Danish haplotype A2a2a is abundant in several areas of Europe, including the area of Ischgl in Austria. Travel metadata from other countries shows that it has spread from Ischgl.

3.1.2. *A2a1: A common mutation in Denmark and Italy.* We see that the haplotype A2a1 appears 38 times in Denmark. Further, we also see this haplotype in the early targeted testing group (January 31-March 15) in the Icelandic study [GHJ$^+$20, Table] where 36 sequences with haplotype A2a1 were found, out of those 29 had a travel history from Italy and 3 from Austria. Further, the earliest dated Danish sequence is dated February 26 where there was only one confirmed case in Denmark, as reported in the news, has travel history in an Italian ski-area. This case also has haplotype A2a1.

3.1.3. *The triple deletion ATGA1605A with coincident mutation T514C.* Both in the Danish and the world data sets we observe sequences with a triple deletion at sites 1606–1608 (ATGA1605A)[2] which is sometimes coincident with a substitution T514C (identified as A6 from [GHJ$^+$20, Table]). Most of the sequences from our data set having the triple deletion ATGA1605A but not the substitution T514C are from the UK (293 out of 346). In contrast, most of the sequences with both the deletion ATGA1605A and the substitution T514C are from the Netherlands (98 out of 138). In addition, the earliest of the sequences with both ATGA1605A and T514C are from the Netherlands as well. Therefore, we conclude it to be likely that the triple deletion originated in the UK, spread to the Netherlands where it picked up the mutation T514C. In Denmark, we observe 9 sequences with ATGA1605A, two of which additionally have T514C. These latter two Danish sequences are likely of Dutch origin. See Figure 5 for an illustration.

3.2. **Chains of mutations in Denmark.** Now, we turn to chains of mutations which occured inside Denmark. From the Figure 6 one identifies several chains of mutations. Here we report two of the most interesting.

3.2.1. *Chain of mutations starting at C15842A.* The phylogenetic tree with an overview of the associated haplotypes for this mutation can be found in Figure 3. There are 20 sequences with the mutation C15842A and the haplotype A2a2a in the world and they are all Danish.

From the 20 (all Danish) sequences with A2a2a and C15842A, there are 17 which also have the mutation C12781T. Furthermore, of the sequences that have both the mutations C15842A (T5193N in Orf1a) and C12781T (synonomous), there are 8 which in addition have the non-synonymous mutation G22103C (G181R in the spike protein). Another 4 sequences have the mutation A23975G instead and finally, there are 2 which have C25499T. Some of the previously mentioned sequences have

[2]This presumably corresponds to the triple deletion ATGA1604A identified as haplotype A9 in [GHJ$^+$20, Table]. Note that [GHJ$^+$20, Table] places it at position 1604 rather than 1605.

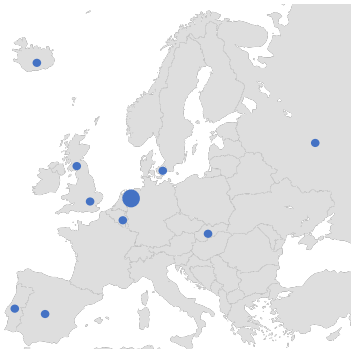
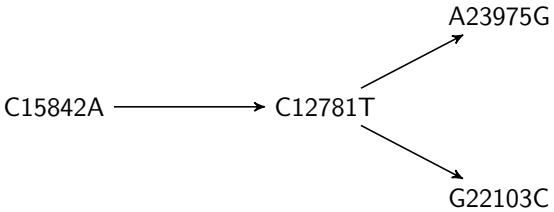


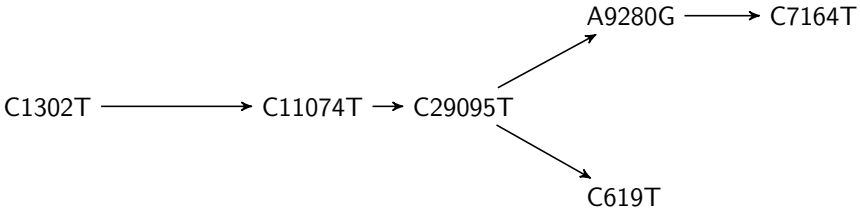
FIGURE 2. The graph shows the appearence of the strain with haplotype ATGA1605A and T514C. According to a maximum likelihood analysis, it likely originated from the Netherlands and spread to other European countries, including Denmark.

additional mutations. The highest number of mutations appearing at least twice in addition to having the haplotype A2a2a is 3.

We have thus identified the chains of mutations:



3.2.2. *Chains of mutations including C11074T.* In total, there are 115 sequences with C1302T in our data set and 103 of them are Danish. Of these, 11 sequences in Denmark and 48 worldwide have C11074T. Ten of the Danish have in addition C1302T (see Figure 4). Of these, 6 have the mutation C29095T; 3 of them moreover have the mutation A9280G. Further, 2 other sequences have the three mutations mentioned initially and C619T. Of the ones with mutation A9280G, 2 have a mutation at C7164T. Some of the sequences have additional single mutations. We have thus identified the chains of mutations:



3.3. **Mutation appearing mainly in Denmark.** The mutation C1302T, which was at the start of the Danish transmission chain discussed above, is also appearing outside Denmark. It is non-synonymous and corresponds to amino acid change T346L in Orf1a. Worldwide there are 115 sequences with the mutation C1302T

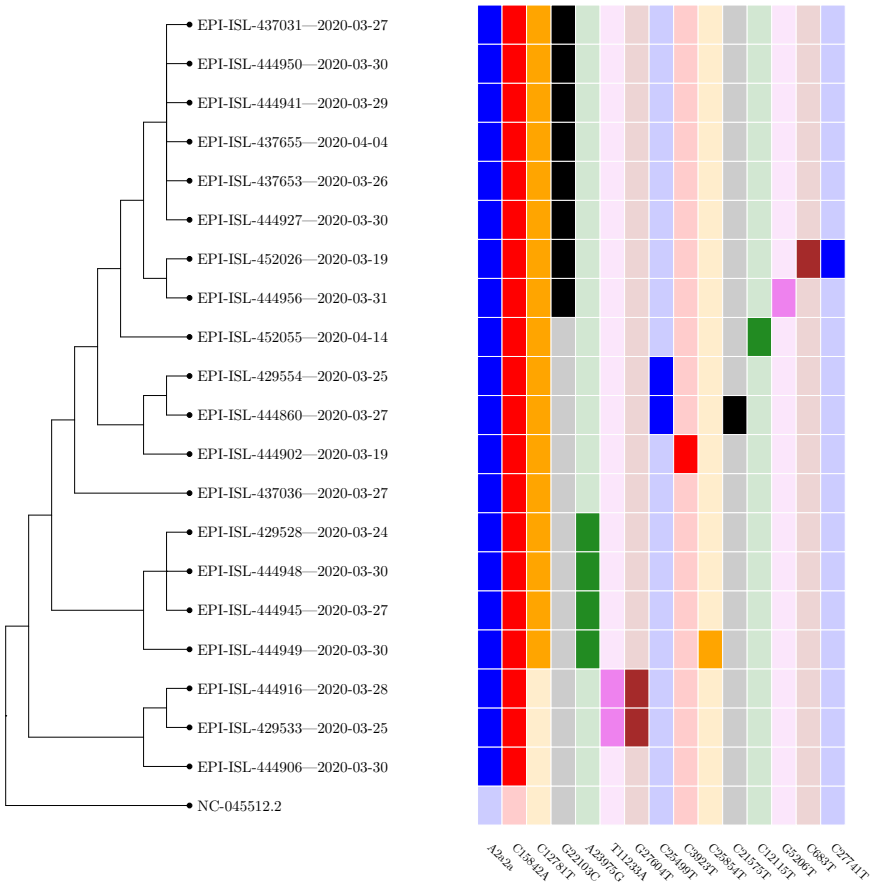


FIGURE 3. Phylogenetic tree for sequences containing haplotype C15842A. The second haplotype shown is C15842A, followed by C12781T. After that it bifurcates into G22103C and A23975G.

cooccurent with the haplotype A2a2a, 103 of which are Danish. The remaining ones are Latvian (1), Icelandic (5) and Swedish (6). The travel histories of the 5 Icelandic sequences show that 2 have traveled to Denmark, while the other cases do not contain travel information. Of the 6 Swedish sequences, the one is from Uppsala dated to March 12 while the 5 others are from Norrbotten (in the North of Sweden) dated from March 24 until April 2. The earliest Danish sequence with C1304T is from March 3. See also Figure 5 for an illustration of the international presence of the mutation.

4. DISCUSSION

4.1. **Introduction into Denmark.** On the basis of a breakdown of the haplotypes, an analysis of Icelandic travel histories, location information on the Austrian

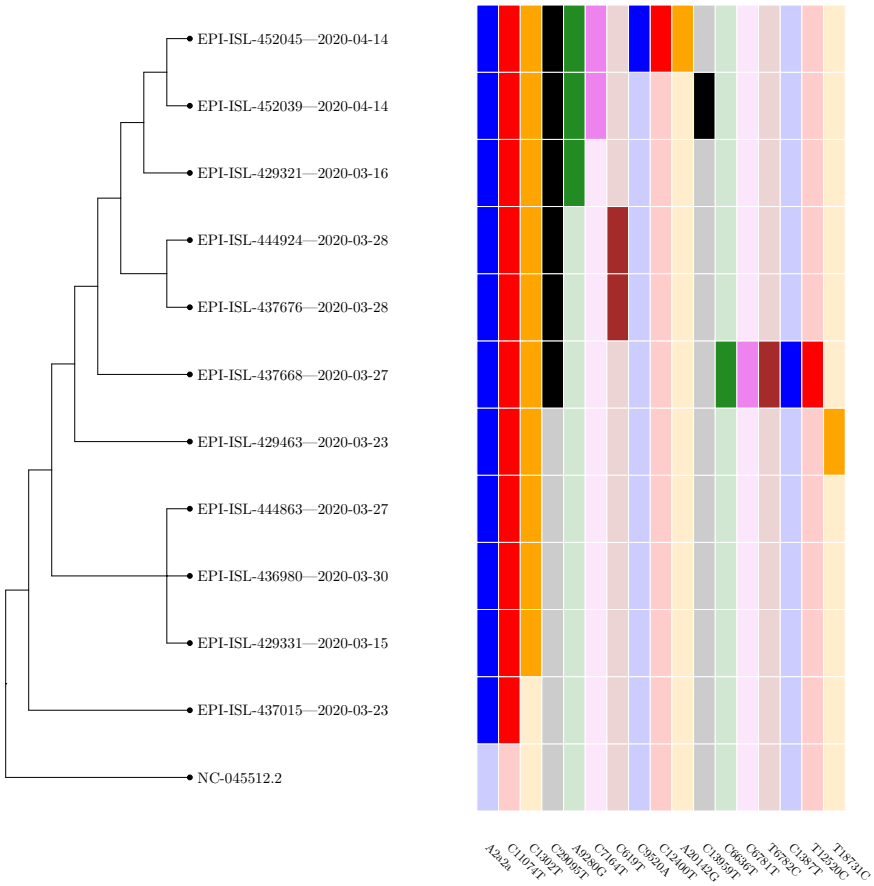


FIGURE 4. A phylogenetic tree for the Danish sequences containing haplotype C11074T with their mutations (with respect to the Wuhan reference) plotted on the right.

data and Norwegian travel information we conclude that the haplotype A2a2a is (among others) associated with ski areas in Austria. Looking further at the mutations A6825C and G15380T we conclude that they happened in Ischgl and spread from there to Denmark. Their abundance in the Danish and Austrian data relative to the total number of A2a2a is consistent with the vast majority of the sequences with haplotype A2a2a stemming from Ischgl. However, due to the abundance of the haplotype in the world, it is likely that a minor portion of the sequences with the haplotype A2a2a are not part of transmission chains through Ischgl (e.g. A2a2a + G24368T which we can associated to spread from the UK - see Subsection A.4). Hence, since about 70 % of the Danish sequences have the haplotype A2a2a it is plausible that most of the Danish cases originate from ski tourists coming back

from Austria. This is not unexpected and gives an independent cross-check of the public knowledge of travel histories [Staa].

As a very specific introduction to Denmark, we deem it likely that the 2 Danish sequences with the triple deletion and the mutation T514C stem from the Netherlands.

Finally, we see that the haplotype A2a1 based on the same metadata from Iceland, Norway and Austria can be vaguely associated to Italy, but the data is not conclusive. We see that SARS-CoV-2 likely has had multiple independent entries to Denmark (e.g. Austria, Italy, the Netherlands and the UK). This is consistent with the testing results in mid-March [Stac].

The abundance of the haplotype A2a2a associated to Ischgl in Denmark might have several different causes. First, there might be a bias in the sequencing as mainly people coming back from these areas were tested at the time of sequencing. Second, that the ski areas in Austria were declared high risk areas only on March 10, which is significantly later than the corresponding declaration for the ski areas in Northern Italy, which happened on March 2. Hence fewer might have cancelled their skiing holidays in Austria in early March than correspondingly in Italy and ski tourists coming back from Austria were not told to go into quarantine as early as tourists coming back from Northern Italy. This could also explain why the A2a2a haplotype is more common in the Danish data as opposed to the Icelandic where all ski areas in the Alps were declared high risk areas from February 29.

4.2. **Transmission chains inside Denmark.** We have listed all mutations that appear at least 3 times inside Denmark in Table 2 and correspondingly plotting those in Figure 6.We see that there are many more Danish transmission chains which one can visually spot on Figure 6. In the results we discussed two particularly interesting chains of mutations based on this plot and Figures 4 and 3. For the two chains described in the results since these mutations chains only cooccur with the haplotype A2a2a in Denmark (except of the first mutation $C1302T$) we conclude that they are transmission chains that happened inside Denmark. This shows clearly how one can track the virus mutating as it spreads inside Denmark. The longest chain that we conclude happened inside Denmark is 5 mutations long and consists of the mutations

$$\text{C1302T} \longrightarrow \text{G11074T} \longrightarrow \text{C29095T} \longrightarrow \text{A9280G} \longrightarrow \text{C7164T}.$$

These mutations took place in a period from before March, 15 to before April, 14 based on the dating of the sequences. The average mutation rate estimated in [vAR+20] of 10^{-4} nucleotides\genome\year and it would be interesting to cross check the mutation rate with the number of mutations in the Danish data.

4.3. **Transmission out of Denmark.** For the mutation C1302T, its high prevalence in Denmark compared to the rest of the world together with the travel histories of the Icelandic cases, we conclude that it has appeared in Denmark and spread from there to Sweden and to Iceland. Some reservations remain since the Swedish data in GISAID is very limited with only 164 sequences as of May 26. Further, the transmission chain described shows how the virus has spread extensively within Denmark and mutated at least 4 times after that.

Hence we see indications that the virus has mutated several times inside Denmark and spread from Denmark, as illustrated in Figure 5. We have listed and discussed the mutations that are most common in the data. Some of the mutations we see

seem to have occurred independently elsewhere, in particular in the UK, which is probably due to their large number of submitted sequences.

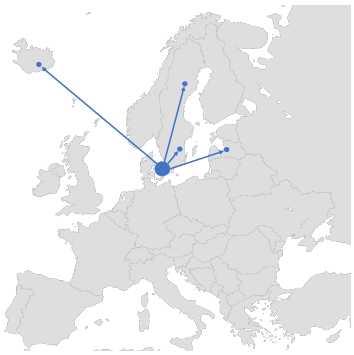


FIGURE 5. The graph shows the likely spread of the strain with haplotype C1302T from Denmark to other Northern European countries. Within Denmark, it also mutated further.

4.4. **Outlook.** The conclusions above show how efficient the combination of metadata and unique mutations can be in determining the transmission chains related to Denmark. We therefore expect that Danish metadata such as the municipality the samples originate from could be used for a more specific tracking of the transmission chains of the virus inside Denmark and to and from other countries. In particular, the genomic knowledge of transmissions could be used in conjunction with other contact tracing methods as earlier demonstrated in the Icelandic study [GHJ+20]. This may provide critical information for the decision making during the reopening, which has an increased focus on contact tracing [Sunc].

Although we here present a case study from Denmark, a similar analysis could be easily carried out for other countries with sequenced genome data. This may provide important insights in the epidemic spread of a virus disease, especially during the unique situation of an onset of a lockdown, the lockdown period and the following reopening.

5. Acknowledgements

We thank Anders Krogh, Jakob Sture Madsen and Carsten Wiuf for valuable comments. We acknowledge financial support from VILLUM FONDEN via the QMATH Centre of Excellence (Grant no. 10059). AHW thanks the VILLUM FONDEN for its support with a Villum Young Investigator Grant (Grant No. 25452). We would like to thank everyone submitting sequenced genome data to GISAID, in particular Statens Serum Institut, with whom we shared an earlier version of this manuscript, and Mads Albertsen's lab. A full list of the contributors can be found in a supplementary file.

6. Note

During completion of this work, we became aware of concurrent work, which was announced here [TV].

References

[Amt] Amt der Tiroler Landesregierung. Ischgl-Chronologie. Available at https://www.tirol.gv.at/meldungen/meldung/artikel/ischgl-chronologie/. Accessed: May 28, 2020.

[BSRS+20] S. Bernard Stoecklin, P. Rolland, Y. Silue, A. Mailles, et al. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. *Euro Surveillance*, 25(6):2000094, 2020.

[Dan] Danish Broadcasting Corporation. DR news – Gymnasium lukker efter nyt tilfælde af corona-smitte, 8 Marts 2020. Accessed: May 15, 2020.

[EBM17] S. Elbe and G. Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.

[Emb] Embætti landlæknis. Skíðasvæðið Ischgl í Austurríki í hóp skilgreindra áhættusvæða. Available at https://www.landlaeknir.is/um-embaettid/frettir/frett/item39457/Skidasvadid-Ischgl-i-Austurriki-i-hop-skilgreindra-ahattusvada. Accessed: May 28, 2020.

[GHJ+20] D. F. Gudbjartsson, A. Helgason, H. Jonsson, O. T. Magnusson, et al. Spread of SARS-CoV-2 in the Icelandic Population. *New England Journal of Medicine*, Apr. 2020.

[HMB+18] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[HWL+20] C. Huang, Y. Wang, X. Li, L. Ren, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*, 395(10223):497–506, Jan. 2020.

[IMP+20] E. Issa, G. Merhi, B. Panossian, T. Salloum, and S. Tokajian. Sars-cov-2 and orf3a: Nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems*, 5(3), 2020.

[PS19] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2019.

[R C17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[Sch11] K. P. Schliep. phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[SM17] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data – from vision to realty. *EuroSurveillance*, 22(13):30494, 2017.

[Staa] Statens Serum Institut. COVID-19 i Danmark: Epidemiologisk overvågningsrapport den 15. maj 2020. Accessed: May 15, 2020, available at https://files.ssi.dk/COVID19-overvaagningsrapport-15052020-fgg8.

[Stab] Statens Serum Institut. Epidemiologisk overvågningsrapport, 12 Maj 2020.

[Stac] Statens Serum Institut. Epidemiologisk overvågningsrapport, 13 Marts 2020.

[Stad] Statsministeriet Pressemøde. Pressemøde om COVID-19 den 10 Marts 2020. Available at http://www.stm.dk/_p_14919.html. Accessed: May 15, 2020.

[Stae] Statsministeriet Pressemøde. Pressemøde om COVID-19 den 11 Marts 2020. Available at http://www.stm.dk/_p_14920.html. Accessed: May 15, 2020.

[Staf] Statsministeriet Pressemøde. Pressemøde om COVID-19 den 6 April 2020. Available at http://www.stm.dk/_p_14938.html. Accessed: May 15, 2020.

[Suna] Sundhedsstyrelsen nyheder. COVID-19: Nye anbefalinger til borgere, 3 Marts 2020. Available at https://www.sst.dk/da/nyheder/2020/nye-anbefalinger-fra-sundhedsstyrelsen-om-at-blive-hjemme-i-to-uger. Accessed: May 15, 2020.

[Sunb] Sundhedsstyrelsen nyheder. En person, der er blevet undersøgt på Rigshospitalet, er det andet bekræftede tilfælde af COVID-19 i Danmark, 8 Februar 2020. Available at https://www.sst.dk/da/Nyheder/2020/En-person-der-er-blevet-undersoegt-paa-Rigshospitalet-er-det-andet-bekraeftede-tilfaelde-af-COVID-19. Accessed: May 15, 2020.

[Sunc] Sundhedsstyrelsen nyheder. Smitteopsporing hos nære kontakter, 14 May 2020. Available at

https://www.sst.dk/da/Nyheder/2020/Smitteopsporing-hos-naere-kontakter .

[TV] TV Avisen, DR. Ny genforskning afslører: Sådan endte Midt- og Vestjylland med at blive hotspot for corona. Available at https://www.dr.dk/nyheder/indland/ny-genforskning-afsloerer-saadan-endte-midt-og-vestjylland-med-blive-hotspot-corona#!/. 21:00, May 28, 2020.

[TV2] TV2. Dansker smittet med coronavirus, 27 Februar 2020. Available at https://nyheder.tv2.dk/samfund/2020-02-27-dansker-smittet-med-coronavirus. Accessed: May 15, 2020.

[vAR+20] L. van Dorp, M. Acman, D. Richard, L. P. Shaw, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution*, page 104351, 2020.

[WHO20a] Coronavirus disease (COVID-19) Situation Report – 11. Technical report, World Health Organization, Geneva, 2020.

[WHO20b] Coronavirus disease (COVID-19) Situation Report – 116. Technical report, World Health Organization, Geneva, 2020.

[WHO20c] Coronavirus disease (COVID-19) Situation Report – 37. Technical report, World Health Organization, Geneva, 2020.

[Wri15] E. S. Wright. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics*, 16(322), 2015.

[Wri16] E. S. Wright. Using decipher v2.0 to analyze big biological sequence data in r. *The R Journal*, 8(1):352–359, 2016.

[Yu20] G. Yu. Using ggtree to Visualize Data on Tree-Like Structures. *Current Protocols in Bioinformatics*, 69(1):e96, Mar. 2020.

Appendix A. Discussion on additional common Danish mutations

In this section, we discuss most common Danish mutations (with respect to the Wuhan reference) in addition to the ones already discussed in the main text. We refer to Table 2 for a summary of mutations occurring at least three times in the Danish data.

A.1. **Mutation C7011T.** Worldwide there are 64 sequences with the mutation C7011T, of which 38 exhibit haplotype A8. Among these 38 are 13 Danish. We suspect that the mutation occurred independently at least twice. Once seems to be from the UK and the other cluster (the one cooccurent with A8) is more unclear. The cluster with the Danish sequences contains 5 Austrian sequences. The Austrian sequences are from Innsbruck, Tyrol, Ischgl, Vienna and St. Anton. This means that 4 out of 5 are from the region of Tyrol, but they are not concentrated in Ischgl. In the overview of the Austrian haplotypes 7 the haplotype A8 is missing in 4 of the Austrian sequences, but they all have at least 1 of the 2 mutations constituting A8. .

A.2. **Mutations T8788C and G26951A.** Worldwide there are 55 sequences with the mutation T8788C and 56 with mutation G26951A and they are all Danish. On Figure 6 one can also see those two mutations are cooccurring and that they are in the beginning of several transmission chains.

A.3. **Mutation C7834T.** Worldwide there are 67 sequences with the mutation C7834T, of which 66 are with haplotype A2a. Among those are the 60 Danish ones so the mutation C7834T constitutes one most abundant Danish mutations and on Figure 6 one can see that they are starting transmission chains.

A.4. **Mutation G24368T.** Worldwide there are 163 sequences with the mutation G24368T, of which 156 are with haplotype A2a2a. Among those are the 5 Danish ones (all with haplotype A2a2a). The majority of the others are from the UK (95), Sweden (50), the Netherlands (4), Iceland (3). The 3 Icelandic samples all exhibit a travel history connected to the UK. The mutation is non-synonymous and is located in the spike protein. The mutation G24368T seems closely tied to the UK and we note that compared to the total number of sequences, a much larger fraction of the Swedish sequences is of this type (50/163) compared to the corresponding fraction of Danish sequences (5/583). This indicates that Sweden might have been infected through the UK to a larger extend than Denmark. Another point of interest is that G24368T is co-occurring with haplotype A2a2a. This is a specific example that the link between haplotype A2a2a and Austria is non-exclusive. We think that this question merits further investigation.

Appendix B. Supplementary material

Total	World 20254	Denmark 583	Sweden 164	Norway 34	France 376	Italy 85	UK 6669	Netherlands 490	Austria 248	Germany 174
A	20235	581	163	33	375	84	6646	489	247	173
A1a	1418	0	1	6	11	7	901	50	4	6
A1a1	1372	0	1	1	9	0	898	50	4	6
A1a1a	546	0	0	0	2	0	271	45	3	1
A1a1a1	42	0	0	0	0	0	0	0	0	0
A1a1a2	14	0	0	0	0	0	0	0	0	0
A1a1a3	8	0	0	0	0	0	0	0	0	0
A1a1a4	84	0	0	0	0	0	59	0	0	0
A1a1a5	27	0	0	0	1	0	15	0	0	0
A1a1b	645	0	1	0	7	0	523	2	0	5
A1a2	0	0	0	0	0	0	0	0	0	0
A2	13900	553	153	25	332	76	5024	314	211	141
A2a	13867	552	153	25	332	76	5024	314	211	126
A2a1	4363	38	43	3	26	36	2690	110	96	50
A2a1a	363	13	1	0	2	1	85	83	34	17
A2a1a1	27	0	0	0	0	0	0	0	0	0
A2a1a2	28	0	0	0	0	0	7	4	0	8
A2a1a3	12	1	0	0	0	0	8	0	0	0
A2a1b	349	4	1	0	0	0	280	2	2	2
A2a1c	230	0	1	0	1	0	153	5	4	0
A2a1d	6	0	0	0	0	0	0	0	0	0
A2a2	5343	410	70	14	185	0	421	52	66	44
A2a2a	4343	405	66	12	112	0	279	34	63	44
A2a3	943	13	3	1	1	2	425	19	22	2
A2a3a	82	0	0	0	0	0	0	0	0	0
A2a4	13	0	0	0	0	0	0	0	0	0
A2a5	195	3	3	6	1	5	63	5	0	0
A2a6	104	4	0	0	0	0	22	48	5	2
A2a7	432	0	0	0	0	0	409	0	0	0
A2a8	513	1	0	0	110	0	43	6	1	2
A2a9	136	1	1	0	2	6	51	7	2	6
A2a10	125	0	0	0	0	0	92	0	0	0
A2a11	23	0	0	0	0	0	2	19	0	0
A2a12	26	0	0	0	0	1	7	5	0	0
A2a13	149	2	0	0	0	0	131	0	0	0
A3	162	1	1	1	0	0	2	2	0	1
A6	158	2	0	0	0	0	19	99	3	1
A7	17	0	0	0	0	0	4	0	0	0
A8	310	16	3	0	0	0	189	6	5	17
A9	73	0	0	0	0	0	63	0	0	0
A10	3	0	0	0	0	0	0	0	0	0
B	2173	1	2	0	2	0	72	7	5	2
B1	1299	0	0	0	0	0	5	0	0	0
B1a	1275	0	0	0	0	0	5	0	0	0
B1a1	1249	0	0	0	0	0	5	0	0	0
B1a1a	124	0	0	0	0	0	0	0	0	0
B1a1a1	13	0	0	0	0	0	0	0	0	0
B1a1a1a	10	0	0	0	0	0	0	0	0	0
B1a1a1b	1	0	0	0	0	0	0	0	0	0
B2	46	0	0	0	0	0	0	0	0	1
B4	98	0	2	0	0	0	0	3	0	0

Table 1. Summary of haplotypes per country as per May 26, 2020. We include the haplotypes described in [GHJ+20].

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

Email address: christandl@math.ku.dk

Total	World 20254	Denmark 583	Sweden 164	Norway 34	France 376	Italy 85	UK 6669	Netherlands 490	Austria 248	Germany 174
C1302T	115	103	6	0	0	0	0	0	0	0
C7834T	73	60	0	0	0	0	4	0	0	0
G26951A	56	54	0	0	0	0	2	0	0	0
T8788C	55	53	0	0	0	0	2	0	0	0
C15842A	20	20	0	0	0	0	0	0	0	0
C12781T	37	17	0	0	0	0	0	1	0	0
G15380T	40	16	0	1	1	0	5	0	8	2
C7011T	64	13	2	0	0	0	28	1	5	0
C21707T	51	11	0	0	0	0	4	0	0	0
C11074T	48	11	0	0	0	0	19	1	0	0
C29200T	41	8	0	0	0	0	21	0	0	0
G22103C	8	8	0	0	0	0	0	0	0	0
C28253A	8	7	0	0	0	0	0	0	0	0
A6825C	10	6	0	1	0	0	1	1	1	0
C3743T	8	6	0	0	0	0	1	0	0	0
C16887T	72	5	1	0	3	0	32	0	0	0
G19086T	13	6	0	0	0	0	3	0	0	1
G24368T	163	5	50	0	1	0	95	4	0	0
T99C	4	4	0	0	0	0	0	0	0	0
C218T	8	4	0	0	0	0	1	0	0	0
A518G	7	3	0	0	0	0	0	0	0	0
T658C	26	4	0	0	0	0	1	4	0	2
C4002T	174	4	1	0	0	0	118	2	2	0
A4548G	3	3	0	0	0	0	0	0	0	0
C4668T	23	3	0	0	0	0	16	1	0	0
C6539T	8	6	0	0	1	0	1	0	0	0
C7029T	4	3	0	0	0	0	0	0	0	0
C7819T	10	3	0	0	1	0	0	0	0	0
A9280G	8	3	0	0	0	0	0	0	0	0
C9803T	16	3	0	0	1	0	1	1	0	0
C9979T	15	4	0	0	0	0	7	0	0	0
G10465A	11	5	0	0	0	0	0	0	0	0
C13536T	190	4	1	0	0	0	118	2	2	2
G15906A	11	4	0	0	0	0	0	0	0	0
C18744T	29	5	0	0	0	0	3	0	0	0
C18877T	595	3	1	2	0	0	30	9	7	0
C19263T	15	3	0	0	0	0	1	0	0	0
G20398A	7	3	0	0	0	0	0	0	0	0
G20419T	12	4	3	1	0	0	0	0	1	0
G20887A	10	4	0	0	2	0	0	0	0	1
C21742T	21	5	0	0	0	0	2	0	0	0
C21855T	18	9	0	0	0	0	1	0	0	0
T21976C	8	4	0	0	0	0	0	0	0	0
A23975G	4	4	0	0	0	0	0	0	0	0
G25785T	15	3	1	0	0	0	3	0	0	0
C28629T	20	4	0	0	0	0	1	4	1	2
G28899T	8	4	0	0	0	0	2	0	0	0
C29719T	6	3	0	0	0	0	2	1	0	0
G29734C	171	3	1	0	0	0	64	9	9	1
G29779T	13	4	0	0	0	0	5	0	0	0
C23575T	29	3	1	1	1	1	9	0	0	0
C23587T	17	3	0	1	0	0	6	0	0	0
ATGA1605A	484	9	1	0	2	0	306	108	0	1
C29095T	71	7	0	0	0	0	13	0	0	1
C7164T	12	2	0	0	0	0	5	1	0	0
C619T	21	2	0	0	0	0	0	0	0	0
C25499T	2	2	0	0	0	0	0	0	0	0

TABLE 2. Summary of mutations per country as per May 26, 2020. We include all mutations with respect to the Wuhan reference that appears at least 3 times in the Danish data. The fourth mutation from the bottom is the deletion at position 1606–1608 described in Section 3.1.3; the last three mutations are included because relevant for the discussion in Section 3.2.

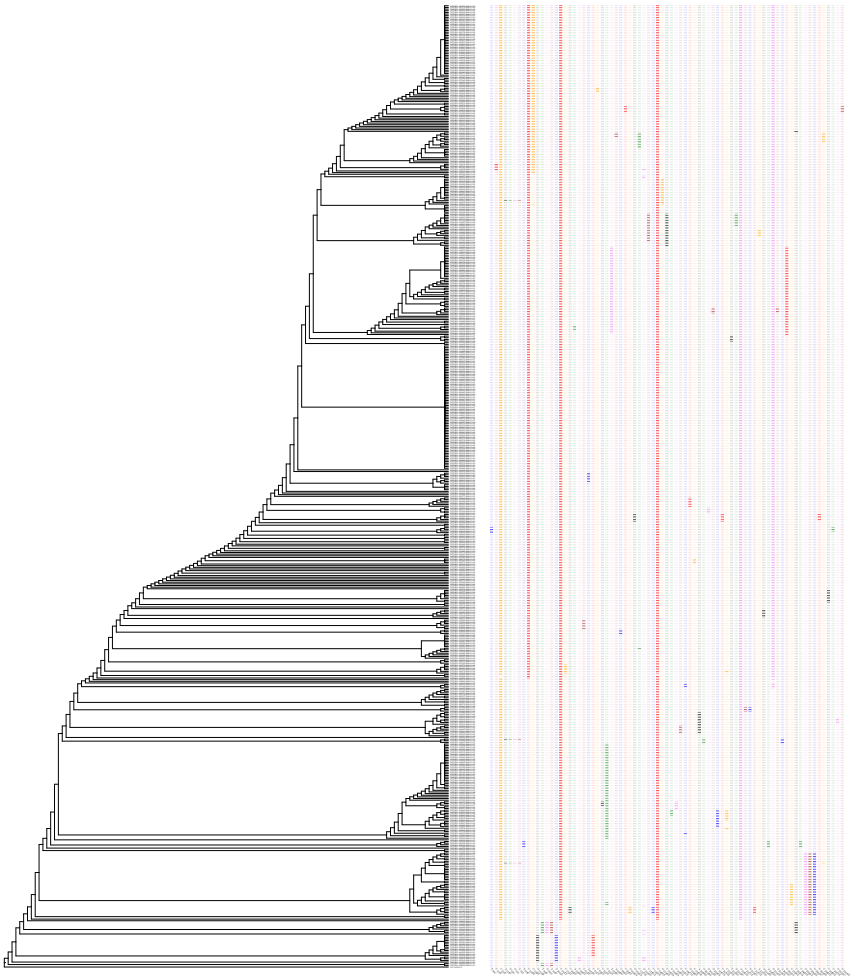


FIGURE 6. Overview of the phylogenetic tree of Denmark on the left with the corresponding haplotypes on the right. The plot shows how about 70 % of the sequences have the haplotype A2a2a (corresponding to the columns (C241T,C3037,A23403 (A2)), C14408T(A2a),G25563T(A2a2) and C1059 (A2a2a)) and that 103 have the further mutation C1302T. We see how the virus was introduced to Denmark from multiple sources and many further chains of mutations.

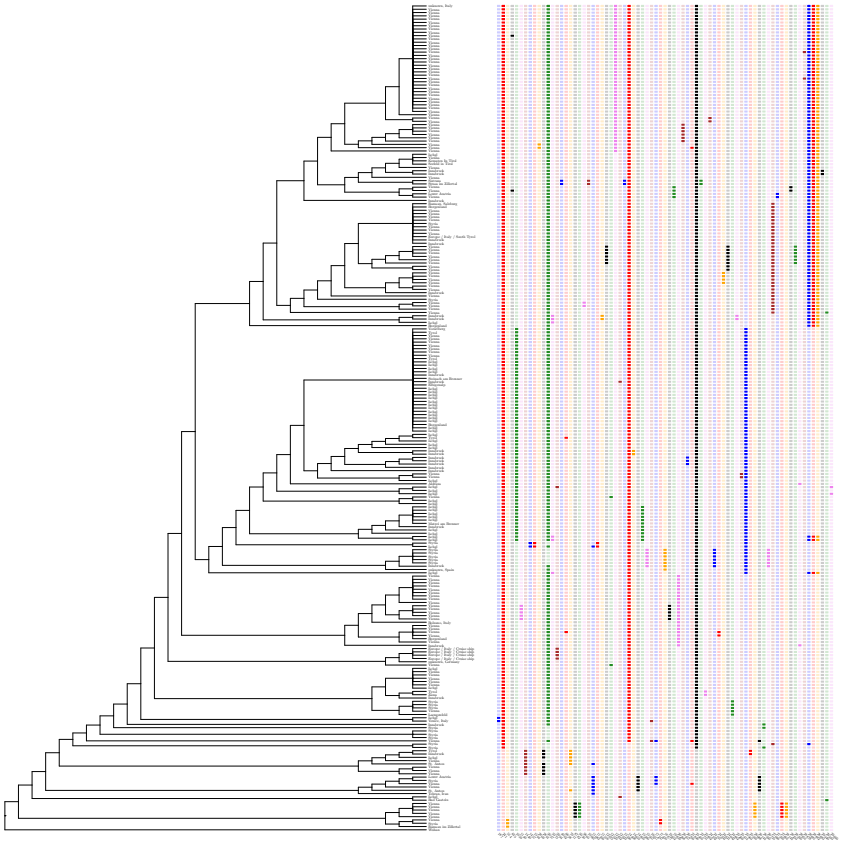


FIGURE 7. Austrian sequences with location metadata and mutations with respect to the Wuhan standard reference.