Exploring the Mutations that Comprise the Genetically Linked Haplotype of the SARS-CoV-2 Virus

Nathan Beshai, Owen Dailey, Kam Taghizadeh, Ian Wright

BIOL 368: Bioinformatics Lab

December 10th, 2020

Outline

1. The D614G Mutation is Almost Always Observed with Three Other Mutations
2. The Effects and Locations of these Four Mutations
3. The Sequences were Collected and a Sequence Alignment and Phylogenetic Tree were Generated
4. The Sequence Alignment Only Yielded One Sequence with Irregular Mutations
5. The phylogenetic tree did not show a relationship between the location collected and date collected
6. Future Directions

The D614G Mutation is Almost Always Observed with Three Other Mutations

- In March of 2020, the D614G mutation was observed globally and quickly became the dominant form of the SARS-CoV-2 virus
- The D614G mutation increases the flexibility of the SARS-CoV-2 spike protein and has the ability to influence the dynamics of the fusion peptide
  - D614G is just a part of the story
- Four mutations comprise a "genetically linked haplotype" (Korber et al., 2020)
  - C241T (UTR)
  - C3,037T (silent mutation)
  - C14,408T (RdRp)
  - A23,403G (D614G)
- Global sequences: CCCA -> TTTG
- In some cases TTCG and CCCG sequences have been observed
- Did each mutation of the genetically linked haplotype arise independently, and if so, in what order?

The Effects and Locations of these Four Mutations

C241T

- Located in the SARS-CoV-2 5′UTR
- May effect secondary RNA structure
  - Influences the rate of RNA replication (Kim et al., 2020)

C3037T

- Located in Nsp3
- Silent mutation

C14,408T

- Located in RNA-dependent RNA polymerase
- Missense mutation
  - Proline to leucine at position 323 (P323L) in RdRp protein
- Likely to have mutations in the membrane (M) and envelope (E) proteins (Ugurel et al., 2020)

A23,403G

- D614G substitution on S protein

Regions Across the World Experienced Switch to G614 Dominance at Different Times

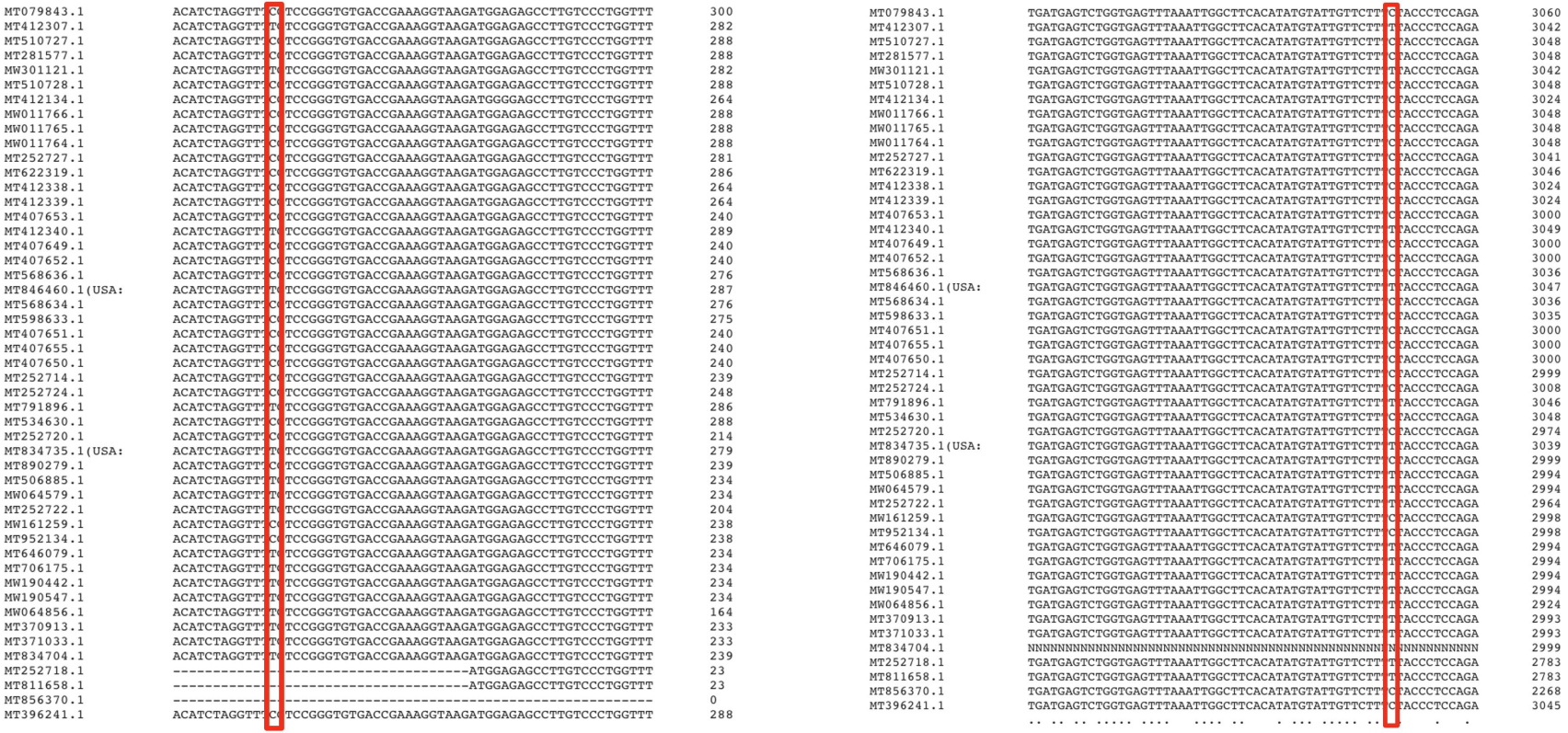
	Onset	|--Before-|		Delay	|--After--|		Last	Delta	Fisher
Location	Date	G/(G+D)	G+D	Date	G/(G+D)	G+D	Sample	G/(G+D)	p-val
Africa	Mar 13	0.870	23	Mar 27	0.974	2780	Nov 18	0.10	0.021409
Asia	Jan 28	0.009	344	Feb 11	0.658	9292	Nov 15	0.65	0.000000
Europe	Jan 29	0.235	17	Feb 12	0.942	127747	Nov 24	0.71	0.000000
North-America	Feb 28	0.051	99	Feb 20	0.911	37866	Nov 16	0.86	0.000000
Oceania	Mar 4	0.059	51	Mar 18	0.929	10408	Dec 2	0.87	0.000000
South-America	Mar 4	0.611	18	Mar 18	0.971	1364	Nov 26	0.36	0.000001

- **Frequency switchover date**
  - **Asia- Feb 11**
  - **North America- Feb 20**
- **This informed collection date periods to be considered when compiling sequences**

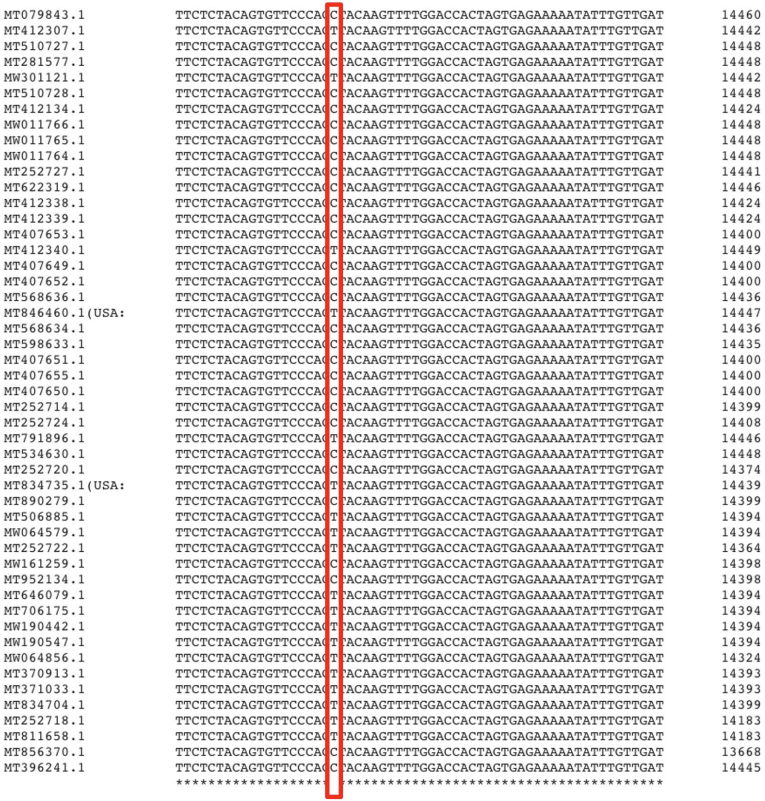
The Sequences were Collected and a Sequence Alignment and Phylogenetic Tree were Generated.

- 48 sequences were collected from the NCBI sequence and NCBI virus Sequence database.
  - 12 Sequences were collected from China; 12 before 2/11 and 12 after 2/11
  - 12 sequences were collected from the US; 12 before 2/20 and 12 after 2/20
- The sequences were aligned using EMBL clustal omega sequence aligner tool and a phylogenetic tree was generated using the same tool.
- The mutation locations were observed in all the sequences observing the:
  - C to T mutation at the 3037 nucleotide location
  - C to T mutation at the 14,408 nucleotide location
  - A to G at the 23, 403 nucleotide location
  - C to T at the 241 nucleotide location

Sequence Alignment Showing the C241T and C3037T Mutation



Sequence Alignment Showing the C14,408T and C23,403T Mutation



The Sequence Alignment Only Yielded One Sequence With Irregular Mutations

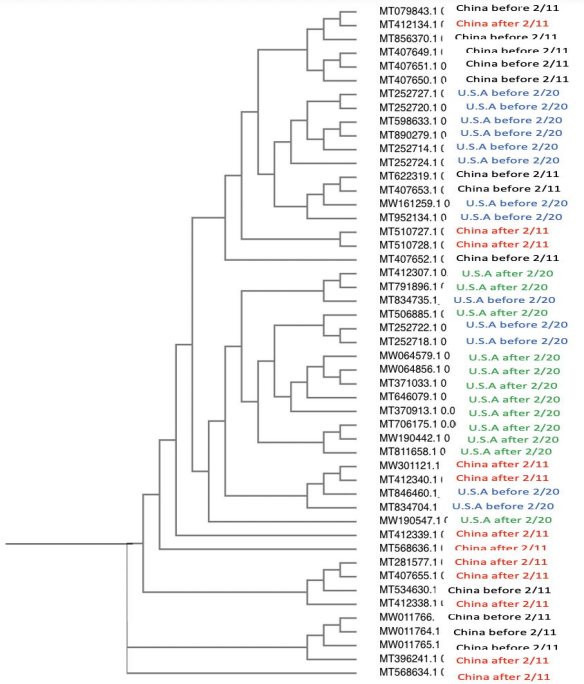
MT079843.1	CCCA	MT412307.1	TTTG	MT856370.1	NCTA
MT510727.1	CCCA	MT281577.1	TTTG		
MW301121.1	CCCA	MT412340.1	TTTG		
MT510728.1	CCCA	MT846460.1	TTTG		
MT412134.1	CCCA	MT252714.1	TTTG		
MW011766.1	CCCA	MT534630.1	TTTG		
MW011765.1	CCCA	MT834735.1	TTTG		
MW011764.1	CCCA	MT890279.1	TTTG		
MT252727.1	CCCA	MT506885.1	TTTG		
MT412338.1	CCCA	MW161259.1	TTTG		
MT412339.1	CCCA	MT646079.1	TTTG		
MT407653.1	CCCA	MT706175.1	TTTG		
MT407649.1	CCCA	MW190442.1	TTTG		
MT407652.1	CCCA	MW190547.1	TTTG		
MT568636.1	CCCA	MW064856.1	TTTG		
MT568634.1	CCCA	MT370913.1	TTTG		
MT598633.1	CCCA	MT371033.1	TTTG		
MT407651.1	CCCA	MT834704.1	TNTG		
MT407655.1	CCCA	MT252718.1	NTTG		
MT252724.1	CCCA	MT811658.1	NTTG		
MT252724.1	CCCA				
MT791896.1	CCCA				
MT252720.1	CCCA				
MW064579.1	CCCA				
MT252722.1	CCCA				
MT396241.1	CCCA				

- ~55% of the sequences were CCCA
- ~44% of the sequences were TTTG, TNTG, or NTTG
- ~1% of the sequences were NCTA
- The sequence alignment results provide support for the “genetically linked haplotype” (Korber et al., 2020)
  - All four mutations arose at once
- May need more sequences to view more irregular mutations

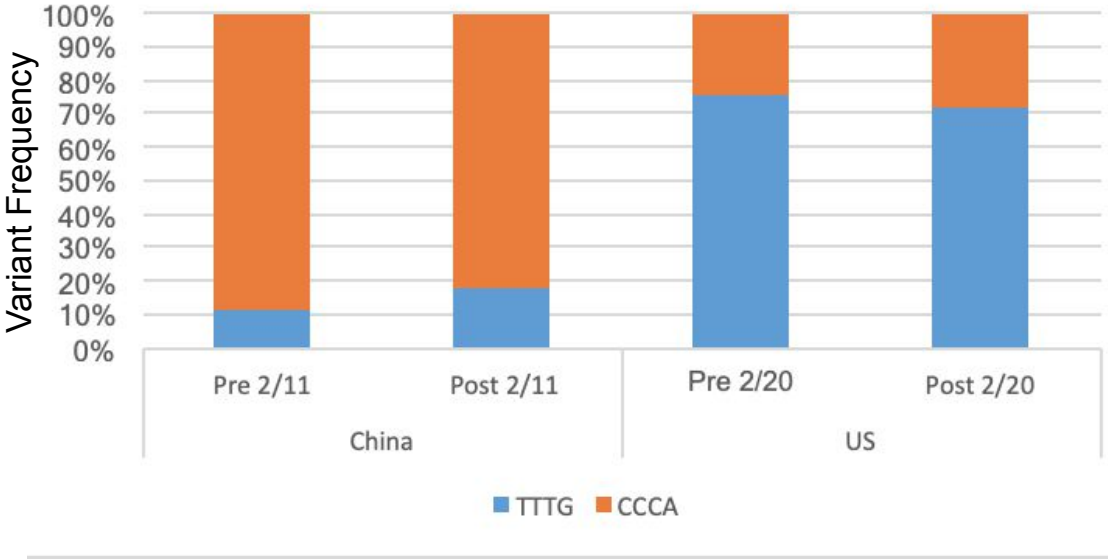
Phylogenetic Tree does not Show that Sequences collected Around the Same Time in the Same location have the Nearest Common Ancestor.

- Some sequences collected from the same area and time are grouped together.

- No clear trends are observable.



CCCA Variant is Dominant Strain in China While TTTG is Dominant in USA On Both Ends of Conversion Date



- This data is confounding when considering the frequency of G614 increases to >90% after the conversion dates
- Spreading the range of collection dates could produce more accuracy of frequencies

55 percent Sequences Collected demonstrated the D614 variant.

- Of the 48 Sequences collect, 26 of the sequences showed the D614 Mutation.
- The sequences were randomly collected and the variants were not searched for.
- The Phylogenetic Tree did not show an overall trend with sequences collected from the same area and time not sharing the most common ancestor.
- After the conversion date the G614 variant was the dominant strand.

Future Directions

- Gather more sequences and find a software that allows for the sequence alignment of a large number of nucleotides
  - Sequences spanning a large geography and time range
  - Limited by database and alignment software
- Utilize more automated databases and pipelines such as Korber et al.’s database

Acknowledgments

- We wanted to thank Dr. Dahlquist for helping us find a resourceful article and aiding us with a couple figures.

- We wanted to thank our T/A Annika Dinulos.

- We also wanted to thank the LMU library for giving us access to articles and databases and the LMU biology department.

References

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., ... & Hastie, K. M. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell, 182(4), 812-827, retrieved from https://doi.org/10.1016/j.cell.2020.06.043.

Ugurel, O. M., Ata, O., & Turgut-Balik, D. (2020). An updated analysis of variations in SARS-CoV-2 genome. Turkish journal of biology = Turk biyoloji dergisi, 44(3), 157–167. https://doi.org/10.3906/biy-2005-111

Kim, D., Lee, J. Y., Yang, J. S., Kim, J. W., Kim, V. N., & Chang, H. (2020). The architecture of SARS-CoV-2 transcriptome. *Cell*.