**Distribution Agreement**

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Yang Lu April 12, 2022

# Understanding the rise of SARS-CoV-2 Delta variants in Atlanta and Georgia

by

Yang Lu

Anne Piantadosi
Adviser

Biology

Anne Piantadosi

Adviser

David Civitello

Committee Member

William Kelly

Committee Member

2022

# Understanding the rise of SARS-CoV-2 Delta variants in Atlanta and Georgia

By

Yang Lu

Anne Piantadosi

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2022

Abstract

Understanding the rise of SARS-CoV-2 Delta variants in Atlanta and Georgia
By Yang Lu

SARS-CoV-2 is a positive-sense single-stranded RNA virus first identified in 2019 in Wuhan, China. This virus causes pandemics around the world. Previous studies state that Delta variants of SARS-CoV-2 are more transmissible than others. Thus, we are interested in the rationale of higher transmissibility, which may cause the predominance of a certain lineage in the whole population. We utilize data from Emory Healthcare and Grady Healthcare and data from GISAID to analyze lineage evolution over time in the Atlanta area and in the whole Georgia state. Clinical information used to research viral loads are also from Emory Healthcare and Grady Healthcare system. We use heat maps to visualize lineage data while using boxplot to visualize Ct values data. From our results, we find no significant difference for lineages distribution from March 2021 to December 2021 in the Atlanta area and the whole Georgia state. Alpha variant is predominant from March to May and start to decline in June, in which month Delta and Delta Plus start to increase and gradually predominate. Delta and Delta Plus eliminate almost all other lineages during their period. In December, Omicron emerges and Delta and Delta Plus start to decline. Other than lineage behaviors, we also find no significant difference for Ct values, which is an approximate to viral loads, between lineages of SARS-CoV-2. However, we have not controlled other factors affecting Ct values in this study such as duration of symptoms and vaccination status; further studies may include those factors. In conclusion, this study provides an understanding of why Delta and Delta Plus can predominate the population while other lineages cannot. If methods in this study can be applied to a broader scale and more factors are considered for further researches, we will achieve a better understanding on SARS-CoV-2 pandemics and on pandemic surveillance.

# Understanding the rise of SARS-CoV-2 Delta variants in Atlanta and Georgia

By

Yang Lu

Anne Piantadosi

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2022

Acknowledgements

My thanks to Anne Piantadosi, my adviser, who provided me opportunities of doing research with her and guided me to be a researcher in the field of biomedical science. In addition to my thesis advisor, David Civitello and William Kelly are very supportive through this process, for which I am grateful.

Table of Contents

List of Figures

# Understanding the rise of SARS-CoV-2 Delta variants in Atlanta and Georgia

**Authors: Yang Lu**

**Advisor: Anne Piantadosi**

## INTRODUCTION

SARS-CoV-2 was first detected in late 2019 in Wuhan, China, and later spread worldwide to become a global pandemic. It is a positive-sense single-stranded RNA virus[1] in the coronaviridae family, and it causes respiratory infection in humans. SARS-CoV-2 can mutate through replication error, and recombination of variants in the same host can increase the diversity of coranoviruses[2]. In addition, host-mediated RNA editing by APOBEC and ADAR enzymes are also believed to contribute to the diversity of SARS-CoV-2[3]. The CDC defines a variant of SARS-CoV-2 as a virus whose genome sequence contains one or more mutations, and a lineage is a group of closely related variants with a common ancestor[4]. Several variants raise significant concerns, and VOC, Variant of Concern, are defined by the CDC due to mutations causing immunological as well as viral dynamical differences. For example, lineage B.1.1.7, also known as the alpha variant, have higher transmissibility[5]. The previous study of lineage B.1.1.7, which first emerged in the U.K., inspired researchers to closely study different lineages of SARS-CoV-2 and how they rapidly evolved over time[6]. Lineage B.1.617.2, Delta variant, was first identified in December 2020 in India[7]. Delta Plus refers to several sublineages whose ancestor is Delta variant. According to definition from US CDC, Delta and its sublineages, Delta Plus, includes signature Spike mutations: T19R, (V70F*), T95I, G142D, E156-, F157-, R158G, (A222V*),

(W258L*), (K417N*), L452R, T478K, D614G, P681R, D950N. Among these mutations, K417N mutation differentiate Delta Plus from Delta [ii]. Moreover, Delta Plus have mutations also found in previous emerging lineages, such as B.1.1.7 and B.1.351[8].

Given their predominance, it is of substantial public health importance to understand the transmission and clinical characteristics of Delta and Delta Plus. They been shown to have higher transmissibility[9]. More specifically, household transmission for Delta variants is 70% more than Alpha variants. Previous studies also indicate Delta Plus appeared to be more transmissible than its parent variant, Delta[viii]. It is likely that higher viral load is one reason for higher transmissibility. A higher viral load in Delta compared to previously predominant lineages was observed before. A study in Guangzhou, China presents some clinical outcomes that Delta variants yield higher viral load and a shorter incubation period compared to previous variants[10]. The mean Ct value, which indicates viral load, of Delta is 19.62, lower than the one of Alpha, which is 21.74 in this study. However, in this case, only 159 cases are reported and they are all from a single outbreak. Thus, we are interested in investigating more in clinical outcomes of Delta and Delta Plus with our datasets.

For this study, we are interested in dynamics of Delta and Delta plus among patients in Emory and Grady Healthcare Systems in Atlanta. We define the timing of first detections of Delta and Delta Plus and compare this to what we see in Georgia as a whole. Atlanta is a city with international airport and as one of the most important cities with high density population in Georgia state. The city might contain a different pattern of lineage distribution over time compared to the pattern of a larger geographical scale, the whole Georgia state because there is

a higher possibility of viral lineages importing into Atlanta compared to other regions. We sequence SARS-CoV-2 from patient samples in the Atlanta area and download sequence data in Georgia from GISAID to analyze lineage distribution over time and rises of certain lineages through that time. In order to understand why certain lineage can become predominant while others are disappearing, we investigate whether patients with Delta and Delta Plus have higher viral loads compared to patients with other lineages.

## METHOD

Nasopharyngeal swab samples are collected from Emory Healthcare and Grady Healthcare, and those samples go through RNA extraction, DNase treatment, random primer cDNA synthesis, Nextera library construction, and finally go through Illumina sequencing to generate raw sequence data. Then, we use reference strain NC_045512 to complete reference-based viral genome assembly. We put all fasta files of sequences into the Pango lineage system and used the lineage report analyzed by Pangolin. The Pango nomenclature is a system for identifying SARS-CoV-2 lineage and Pangolin contains information and software tools helping researchers analyze sequences to obtain the lineage information based on the Pango nomenclature[11]. Pango nomenclature is first proposed in June 2020 and it is based on two principles. First, groups of infections with shared ancestry are signified, similar to the method of phylogenetic tree, and secondly, novel epidemiological events, such as appearance of virus in novel locations, a rapidly increasing number of infections, or novel phenotype noticed in cases, need to be highlighted, which is the other principle Pango nomenclature use[12]. The clinical information was gathered in

hospitals from patients' information, and we used RT-PCR Ct value from PCR test while diagnosing positive infections. PCR(RT-PCR) is considered to be the "gold standard" for the detection of SARS-CoV-2, and Ct value indicates how many cycles are done before reaching the testing threshold[13]. Patient metadata, including collection date and lineage information, in Georgia state is downloaded from GISAID, a publicly-available database.

Datasets are processed and complied as CSV files. After compiling into CSV files, we use pandas package in Python to process data and transfer case occurrence to case frequency in a month. Dates of the collection will be transformed to a specific month. Information of lineages over time and information of Ct value is visualized using the matplotlib package in Python, including heatmap and boxplot function in the package. We select "coolwarm" as the colormap of the heat map because this specific color combinations are more intuitive and easier to compare each lineage/variant. We uploaded all codes available into our Github repository[14].

There are 738 patients' metadata entries from the Emory Healthcare system and 34713 data entries from Georgia. They are used to analyze patterns of lineages over time. Due to geographical scale, data in Georgia and Atlanta can only be roughly compared. However, the patterns can be analyzed clearly within each scale, and similar patterns may occur in different geographical scales.

The matplotlib package draws heat maps in Python and several modifications are applied on the original datasets. The heat map means to improve the visualization compared to the scatter plots. Scatter plots can only present relatively limited numbers of data entries and lineage evolution over time are not intuitive. More specifically, when encountering

epidemiological real-world big data, scatter plots are limited by overlapping spots in the same days and lineage evolving trend may not be seen clearly. However, heat maps solve this problem, and patterns can be identified intuitively or quantitatively. For this study, the patterns are identified intuitively without further computation. First, using the original patient metadata, we need to convert the occurrence of each lineage into frequencies. Then, in order to show the patterns clearly, several lineages with significantly lower frequencies are filtered out. Specifically, if a lineage does not occur with at least 0.5% frequency in at least one month, this lineage will be filtered out, while others will be kept to make the heat maps.

For Ct value data, 822 patients' clinical information is included to analyze Ct value for different lineages. This data is broader than what we use for studying lineage distribution, which includes only samples with full genome sequences, and thus have a different number of cases included. Importantly, all lineages are grouped into five subgroups of variants. They are Alpha (93), Delta (614), Brazil (14), Omicron (65), and Other (36). Other than Alpha and Delta, the other three groups' Ct value data are not normally distributed in these groups. This makes the ANOVA test to test the significance of the difference between groups invalid. Thus, the Kruskal-Wallis h test is used to analyze the significance here.

## RESULTS

### Advantages of using heatmaps rather than scatter plots

At the beginning, we use scatter plots to represent our lineage data in a certain time span. However, as in Figure 1, several disadvantages of the scatter plots make the figure unintuitive

and hard to read. The scatter plot may have overlapping dots. Therefore, information is lost and the high frequencies of a certain lineage in a time spot are not available. Moreover, comparison between lineages and different geographical scales can be extremely hard because the patterns in scatter plots are difficult to describe or quantify. Thus, we introduced the heat map as an alternative visualization method and it solves the problems we meet when using scatter plots. When using heat maps, we easily observe and understand the emergences and vanishment of each lineage and we can also understand the predominance of lineages with intuition. Moreover, when data size is relatively large, scatter plots will become crowded and messy while heatmap handle the large data quite well. Figure 2 includes more than 30000 data entries but is still tidy and intuitively understandable.

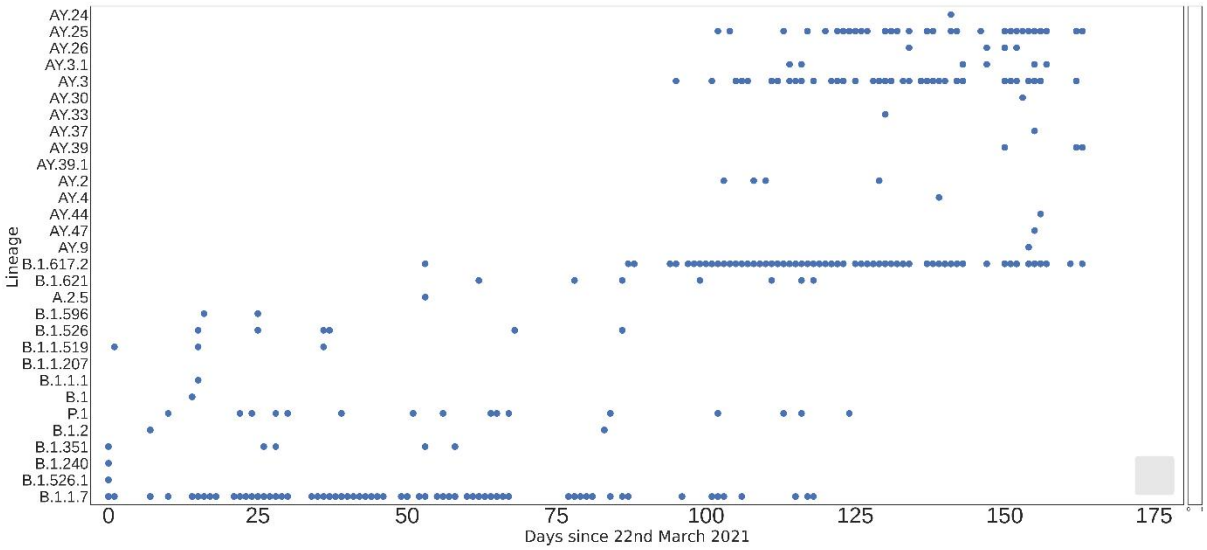


*Figure 1.* Scatter plot of lineages identified in Atlanta overtime since March 22nd. Timeline is converted from normal dates of collection to days after the March 22nd. All lineages are included in this figure,

**Lineages distribution in Georgia state**

To provide context for analysis of lineages in the Atlanta area, we examine lineages across the whole state using publicly available data from GISAID. Total 34713 sequences and data are downloaded from March 22nd 2021 to December 31st 2021. The time period covers both the rise and the decline of Delta and Delta Plus. However, to make the heat map easier to read and compare, after converting occurrence into frequencies, lineages with very low frequency (<0.5%) are filtered. As shown in Figure 2, in March, Alpha variant, which is B.1.1.7, is predominant with 63% frequencies and reaches the highest frequencies at 77% in May. It drops to 47% and starts to lose the predominance in June (47%) while Delta and Delta plus start to increase in frequencies (33%). Delta keeps relatively low frequencies (1-5%) after its emergence in April but Delta Plus begins to predominate the population of lineages in June (89%). From July to November, Delta Plus keeps its predominance (>90%), but suddenly drops (24%) in December when Omicron emerges and takes the predominance (74%).

Interestingly, many other lineages coexist with Alpha during its predominance period, but disappear or keep in very low frequencies (<0.3%) during the predominance of Delta and Delta Plus. The frequency of Alpha reaches only 77% in May, which is the highest frequency Alpha have throughout the time period of its predominance. As a comparison, Delta and Delta Plus reach a frequency of 98% in their period of predominance and maintain this level of frequencies from September to November. Though both lineages predominate (>50%) the population for several months, Delta and Delta Plus show a stronger force of predominance compared to Alpha and most of other lineages disappear during its predominance.

According to the data in the Georgia state, Delta and Delta Plus are first identified in April

and Omicron is first identified in November. The heat map of Georgia (Figure 1) provides a clear presentation of how predominance of Delta and Delta Plus starts and ends and also the behavior all other lineages.

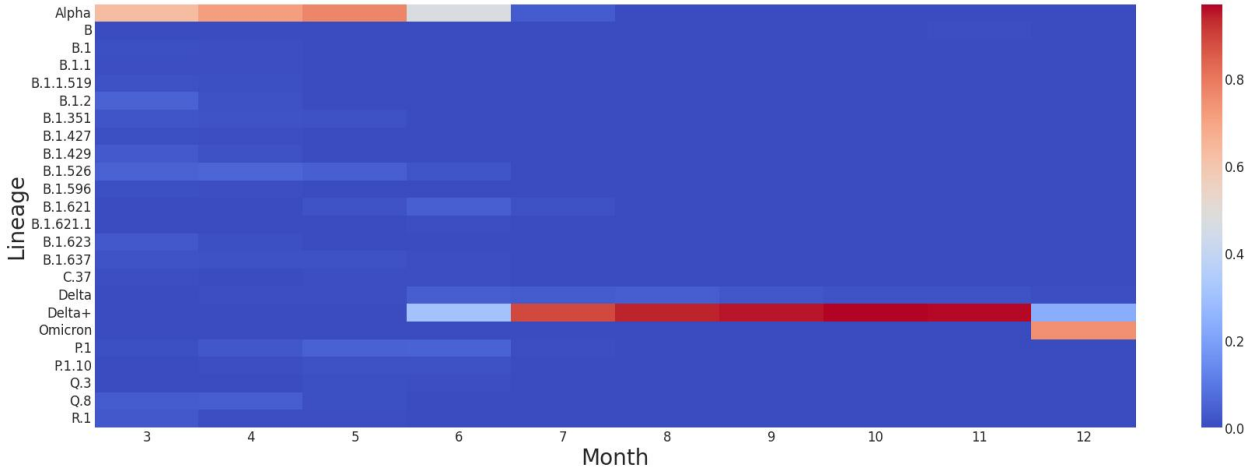


*Figure 2.* Heatmap of Lineage Distribution in Georgia. Timeline is March 2021 to December 2021 and it is represented by number in this figure. All lineages/variants are arranged alphabetically. The color bar on the right indicates that blue is lower and red is higher frequency.

**Lineages Distribution in the Atlanta area**

To specifically analyze the dynamics of Delta and Delta Plus among patients in Emory and Grady Healthcare Systems, we sequence samples from 738 patients who had well-characterized clinical data. The process of converting and filtrating is the same as it is done for the heat map in Georgia state. Alpha variants are predominant in March(70%) and its predominance persists until June(52%), and drops to 5% in July. The decreasing frequency of Alpha is followed by the rise of Delta and Delta Plus. In June, Delta and Delta plus has a 30% frequency and in July, the frequency of the combination of Delta and Delta plus is 89%. The predominance of Delta and Delta Plus persists until November. In October, the predominance of Delta and Delta plus

suddenly drops but it is just because that we only have 3 data entries in October. Other than October, the behavior of Delta and Delta Plus in the Atlanta area is highly similar to the one in Georgia state. Omicron is first identified in Atlanta in December, and it becomes predominant (88%) right after its emergence and Delta Plus drops to 12% while the original Delta lineage disappears. Omicron has a relatively stronger start in Atlanta compared to in the whole Georgia state.

Generally, other lineages have similar behaviors and patterns as the one in Georgia. Before Delta and Delta Plus's predomination, lineages other than Alpha are diverse with a range of frequencies from 2% to 9%, but during predominance of Delta and Delta Plus, most of other lineages generally have approximately 0%. Alpha, Gamma (P.1) and Mu (B.1.621) still exist (3%) in July when Delta and Delta Plus just begin their predominance, but then disappear totally(0%). In October, there is a pattern difference between Georgia and the Atlanta area, but it is just because that in October, data from Emory and Grady Healthcare System only contains 3 data entries. The number of lineages is not the same in two datasets. Naturally, data of Georgia has more data entries and has a higher number of lineages. Thus, for the heat map, map for Georgia contains more lineages than map for the Atlanta area.

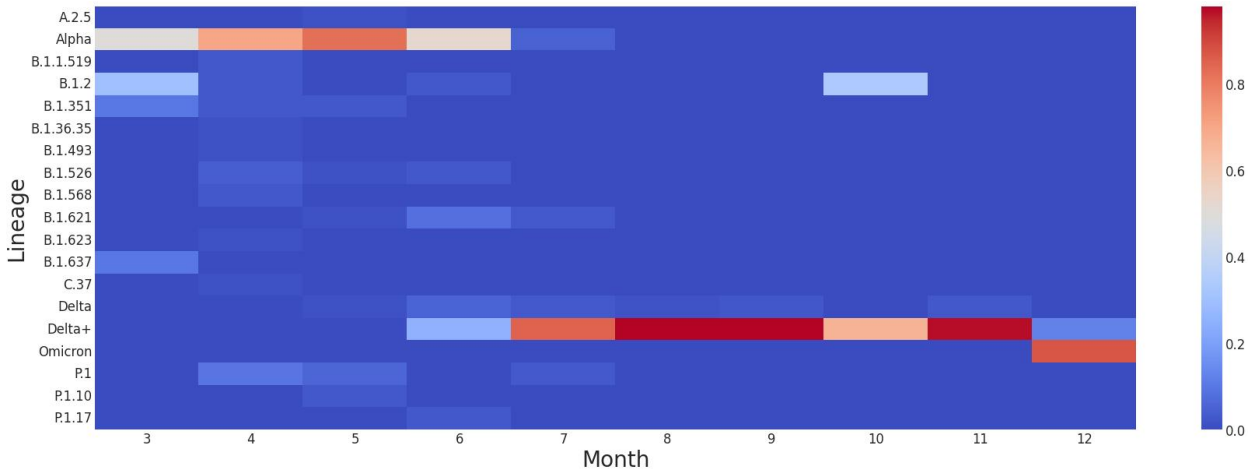


*Figure 3.* Heatmap of Lineage Distribution in the Atlanta area. This figure is similar to Figure 1 and should have a good comparison with it. All lineages/variants are arranged alphabetically. The color bar on the right indicates that blue is lower and red is higher frequency.

**Delta and Delta Plus distribution across the Atlanta area and Georgia State**

As shown above, canonical Delta lineage, B.1.617.2, is less predominant than Delta Plus during the period of predominance by Delta and Delta Plus. The period can actually be identified as a time that Delta sublineages predominate. We therefore, are interested in which sublineages are predominant. From Figure 4 and 5, Delta variant and Delta Plus, whose ancestor is Delta, are first identified in Georgia in April and then in Atlanta in May. Thus, these two figures have different timelines. More specifically, the Atlanta area first identifies Delta and its sublineages in the May 14$^{th}$ while Georgia state first identifies those lineages in the April 9$^{th}$. There is approximately one month gap between the emergence of delta variants in Georgia and the Atlanta area. Starting from June, many novel sublineages of Delta are identified in Atlanta and Georgia, but most of them are maintained in a relatively low frequency (<3%). Other than the difference of emergence time, the Delta and Delta Plus distributions across the Atlanta area and Georgia are similar. Noticeably, however, some delta sublineages, including AY.103, AY.25,

AY.3, and AY. 44, are more frequently identified in both Atlanta and Georgia state. Their frequencies maintain relatively high (>8%) from July to November during the period of predominance by Delta and Delta Plus. The canonical Delta variant, B.1.617.2, only stays in a high frequency (>66%) from April to May in Georgia state and only May in the Atlanta area. After May, B.1.617.2 decline rapidly and its sublineages predominate the population from July to November.

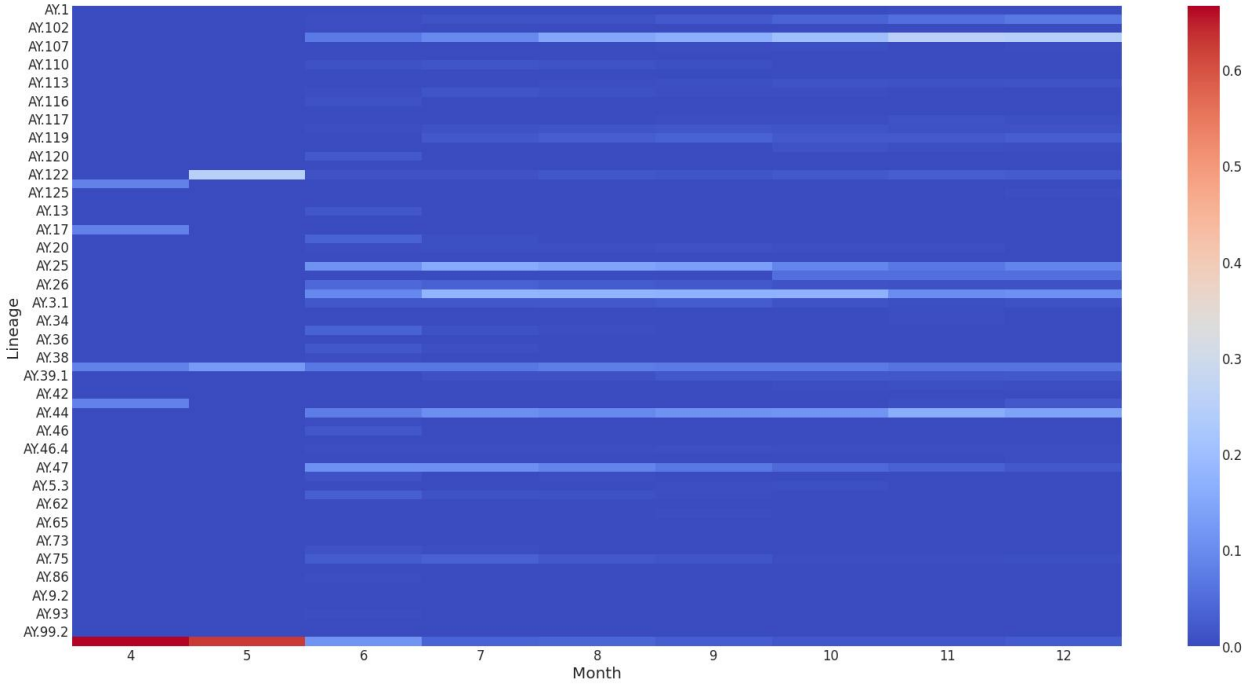


*Figure 4.* The heat map of only Delta and Delta Plus in Georgia. The time line starts from April to December 2021.

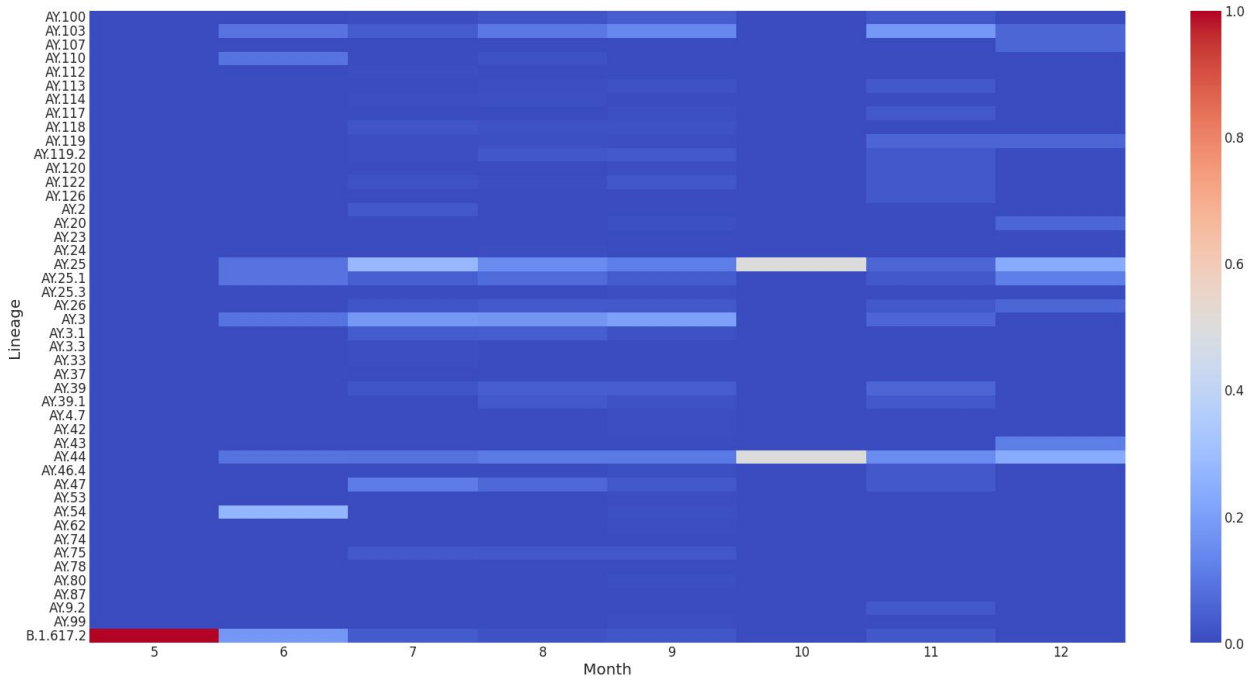


*Figure 5.* The heat map of only Delta and Delta Plus in the Atlanta area. The time line starts from May to December 2021.

**Ct Value**

We are interested in evaluating whether Delta and Delta Plus became predominant because they caused higher viral loads, which would presumably lead to higher transmissibility. Ct value is the cycle threshold of a RT-PCR test we used to test and diagnose SARS-CoV-2 infections. Ideally, a high Ct value means a low concentration of viral genes, which also means that viral loads of the sample are low. Our result is that Ct values between 5 different groups of variants are not significantly different. The range of Ct values overall is 11.3 to 34.7 for 822 data (Alpha (93), Delta (614), Brazil (14), Omicron (65), Other (36)). The mean Ct values of Alpha (21.61), Brazil (20.91), Omicron (20.38), Other (21.7) and Delta (20.98) groups may have some difference but after statistical analysis using ANOVA (p value=0.44) and Kruskal-Wallis h test (p value=0.31),

we conclude that the differences between groups are not significant.

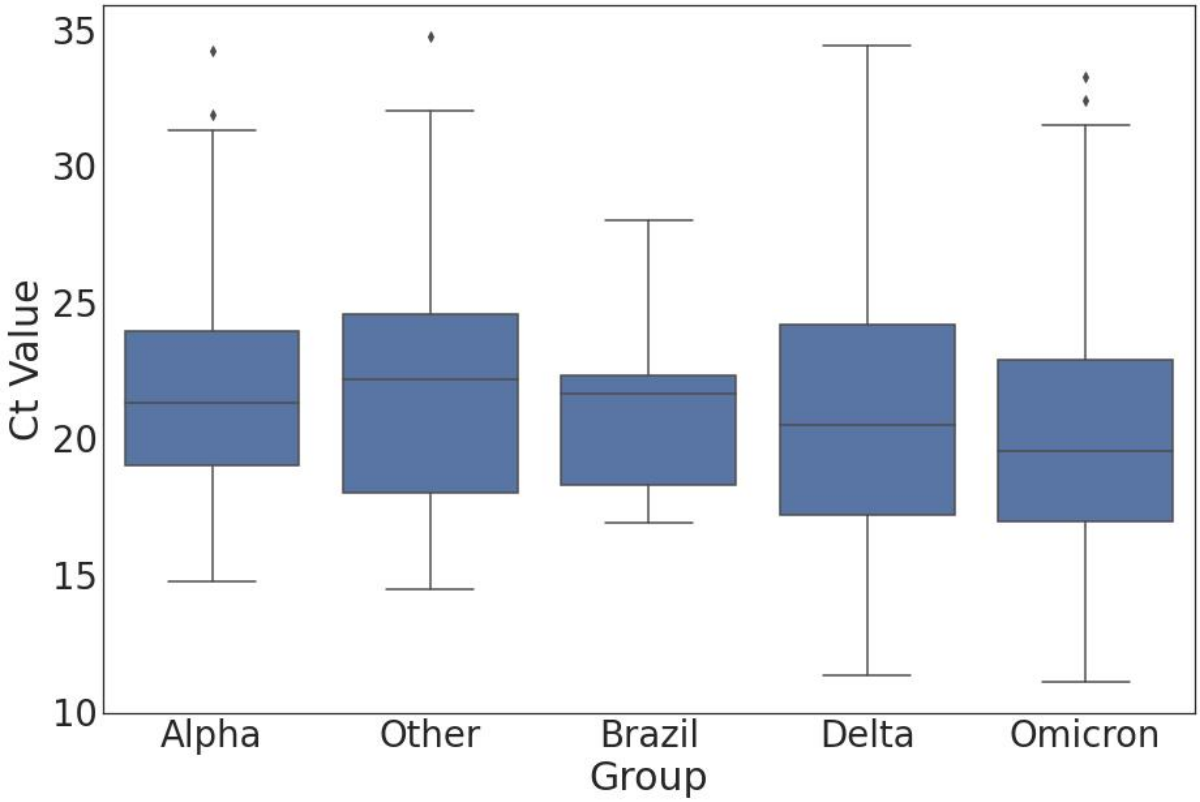


*Figure 6.* The boxplot for Ct values through different lineage groups. The mean and the distribution of Ct values for Alpha (B.1.1.7), Brazil (P.1), Delta (B.1.617.2 and all Delta Plus), Omicron (B.1.1.529, BA.1 and BA.2) and Other (all other lineages) are well shown.

We are also interested in the Ct values difference between Delta and Delta Plus. With 216 cases of Delta and 340 cases of Delta Plus, there is no significant difference between two. The range of Delta is from 12.6 to 32.6 while the range of Delta Plus is from 11.3 to 34.4. Both Delta and Delta Plus have normal distributions for Ct values, therefore we choose their means as parameters to compare. Delta's Ct value mean is 21.0 while Delta Plus's Ct value mean is 20.9. The ANOVA test (p value=0.84) indicates that Ct values for Delta and Delta Plus are not significantly different.

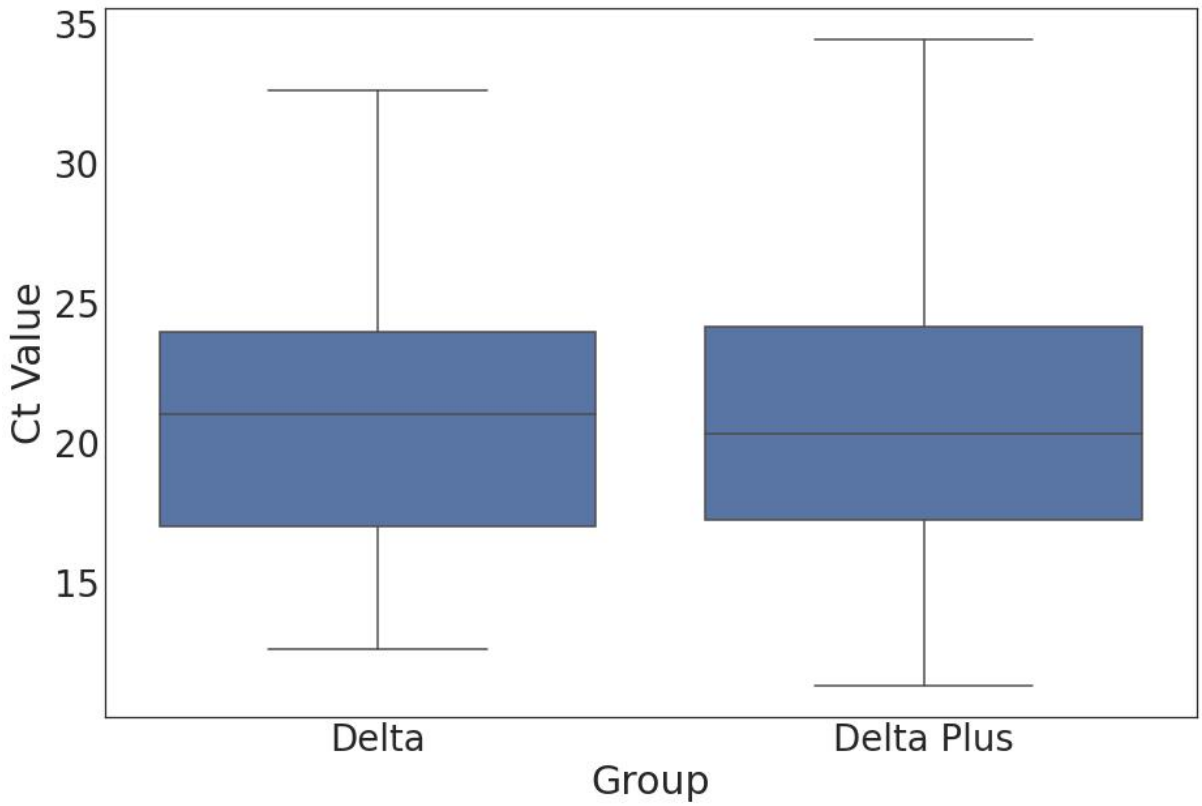


*Figure 7.* The boxplot of Ct values between the original Delta lineage, B.1.617.2, and its sublineages, Delta Plus.

## DISCUSSION

### Lineage distribution

We observe that both throughout Georgia and within Emory and Grady Healthcare patients in Atlanta, Delta and Delta Plus lineages quickly rise to predominance from July to November. Delta and Delta Plus starts to become major lineages (30%) in June and become predominant in July (88%) in just one month. From emergence to predominance, Delta and Delta Plus take 3 months. The results of lineage distribution in Georgia and the Atlanta area generally indicate similarity in lineage patterns, even if the geographical scale is different. When the Alpha variant predominates from March to June, the other lineages still exist in a relatively low frequency but

never extinct. When Delta and Delta Plus is introduced and first identified, Alpha and other lineages start to decrease and disappear. From this pattern in both Georgia state and the Atlanta area, we conclude that Delta and Delta Plus have a stronger force to predominate and eliminate other lineages. Interestingly, canonical Delta is quickly replaced by its sublineages and the predominance of Delta and Delta Plus is mainly driven by Delta Plus rather than Delta.

Figure 4 and Figure 5 reveal that delta and its sublineages are first identified in Georgia state but not in the Atlanta area. One explanation is that Atlanta's Delta variants may be imported from other places in Georgia state. Intuitively, as an international city with high international travelers, Atlanta should be exposed to delta variants more easily, but the results may indicate a different fact that Atlanta is not an origin of delta variants spread in Georgia state. The other explanation is that the result is mainly because Georgia datasets contain more samplings than Atlanta's datasets. With more samplings, the possibility of identifying one new variant is naturally higher. Though the emergences of Delta and Delta Plus in the Atlanta area and Georgia are not at the same time with an approximately 1-month gap, the behavior of its predomination is similar. Delta and Delta Plus start to predominate (>50%) both Georgia state and the Atlanta area, a smaller geographical scale, in July, and maintain their predominance until November. Due to Georgia dataset size is almost 50 times larger than the dataset of the Atlanta area, the difference of emergence may not be significant while the similarity indicates that SARS-CoV-2 lineage predominance is unrelated to geographical factors.

The datasets also include the new variant omicron, first identified in Georgia in December. It later becomes the predominant variant in the population. Similar to what we observe for the

delta and its sublineages, after omicron is introduced, delta and delta plus, start to disappear quickly.

**Ct value and Transmissbility**

In our study, there is no significant difference between Ct values of different groups, which means that the hypothesis that predominant lineages have a lower Ct value cannot be accepted. The results suggest that the predominant lineages with higher transmissibility do not have a higher viral load, which is different from Luo's[ix] results showing that Delta variants have higher viral loads than alpha variants. However, we haven't yet controlled for factors other than lineage that may affect Ct values. For example, vaccinations may increase Ct value[15], indicating a lower viral load due to partial immune protection. Also, Ct values gained from RT-PCR can change over time after infections, so we need to control the time between symptom onsets and testings. The information of vaccinations from most of patients are recorded in our datasets but have not yet been considered in this study. The data of days after infection is possible to obtain by medical record review. Thus, results in this research are preliminary and further improvement and more comprehensive analysis may yield a different result.

The Ct values studies contradicted some results given by previous research. For this study, we only analyze clinical data from Emory Healthcare System, and the results are informative and possibly biased. We also need to notice that other factors, including vaccinations and time after infections, may also affect Ct values. Further research may gather more information, include more factors and increase the data size to see if Ct values, which is the approximate of viral

loads, of predominant lineage in a time period is significantly lower than other lineages.

## CONCLUSION

With our results, we find out that the behaviors of lineage evolution over time are highly similar between a small geographical scale, the Atlanta area, and a bigger geographical scale, the whole Georgia state. The change of predominant lineages is also interesting because we notice the difference between different predominant lineage in different time period. Alpha variant's predominance starts from March to May and other lineages coexist with Alpha, while in June, Alpha starts to decline and Delta and Delta Plus starts to increase. From July to November, Delta and Delta Plus predominates and eliminate all other lineages. In December, Omicron emerges and take the position of predominance from Delta and Delta Plus. Other than analysis on lineages, we also find out that Ct values between different lineages are not significantly different. However, we are concerned that for this study, we haven't controlled for other factors such as duration of symptoms and vaccination that may affect viral loads of SARS-CoV-2.

In the process of research, we provide a novel method of visualizing the lineages change over time in this study. If possible, by changing the x-axis of the heat map into days, the heat map of lineages distribution over time can serve as a reporter for a larger population, offering daily information about lineages' predominance. This information should be able to help researchers and public health policymakers to progress on their projects. One of the other benefits of using the heat map is that it solves the problem of presenting a large amount of data

for this type of study. Inspired by this study, if we introduced a third dimension of geographical locations, we could create a 3D version of lineage distribution over time in each community.

As an insight for the future study, our study also includes the emergence of Omicron. The rise of Omicron occurs at the same time as the decline of Delta and Delta Plus, which is similar to the procedure of how Delta and Delta Plus replace Alpha. The rationale behind these phenomena is possible to be similar. If the theory of competition between lineages are correct, we may assume that Omicron owns features that make it the strongest lineage for now. However, when Omicron emerges, its only competitors are Delta and Delta Plus. Thus, another explanation is that Omicron only contain features that compete out Delta and Delta Plus. More researches[16] on Omicron are still ongoing and these studies will reveal the facts of why Omicron can be predominant over Delta Plus and also shed a light on why Delta and Delta Plus are predominant for several months.

In addition, this study provides an understanding of why Delta and Delta Plus can predominate the population while other lineages eventually disappeared. The analysis over the question of predominance should be inspiring to further researches and perhaps a solution to the question will soon be discovered. The question is urgent for us to answer because Omicron is now spreading around the world and becomes predominant lineage while Delta and Delta Plus vanish. It is irresponsible to conclude that no future new variants will emerge and thus, if we can understand the rationale of predominance, we may find methods of interfering the process and prevents us from suffering a longer period of the SARS-CoV-2 pandemic. Furthermore, this study also brings insights to public health institutes to surveil SARS-CoV-2

epidemic over time by monitoring lineage distribution instantly and continuously using the heat map. In general, if methods in this study can be applied to a broader scale and to different infectious diseases, we will achieve a better understanding on epidemiology and a critical improvement on pandemics surveillance.

## References

[1] Machhi J, Herskovitz J, Senan AM, Dutta D, Nath B, Oleynikov MD, et al. (September 2020). "The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections". Journal of Neuroimmune Pharmacology.15(3): 359–386. doi:10.1007/s11481-020-09944-5. PMC 7373339. PMID 32696264.

[2] Tao, Kaiming, et al. "The biological and clinical significance of emerging SARS-CoV-2 variants." Nature Reviews Genetics 22.12 (2021): 757-773.

[3] Simmonds, Peter. "Rampant C→ U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short-and long-term evolutionary trajectories." *Msphere* 5.3 (2020): e00408-20.

[4] Centers for Disease Control and Prevention. "SARS-CoV-2 variant classifications and definitions." (2021).

[5] Volz, E., Mishra, S., Chand, M. et al. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature 593, 266–269 (2021). https://doi.org/10.1038/s41586-021-03470-x

[6] Frampton et. al., Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study,The Lancet Infectious Diseases,Volume 21, Issue 9,2021,Pages 1246-1256,ISSN 1473-3099, https://doi.org/10.1016/S1473-3099(21)00170-5.

[7] Shiehzadegan, Shayan, et al. "Analysis of the delta variant B. 1.617. 2 COVID-19." Clinics and

Practice 11.4 (2021): 778-784.

[8] Arora, Prerna, et al. "Delta variant (B. 1.617. 2) sublineages do not show increased neutralization resistance." Cellular & molecular immunology 18.11 (2021): 2557-2559.

[9] Allen, Hester, et al. "Increased household transmission of COVID-19 cases associated with SARS-CoV-2 variant of concern B. 1.617. 2: a national case-control study." Public Heal Engl (2021).

[10] Luo, Chun Huai, et al. "Infection with the SARS-CoV-2 delta variant is associated with higher infectious virus loads compared to the alpha variant in both unvaccinated and vaccinated individuals." medRxiv (2021).

[11] O'Toole, Áine, et al. "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool." *Virus Evolution* 7.2 (2021): veab064.

[12] Rambaut, Andrew, et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology." Nature microbiology 5.11 (2020): 1403-1407.

[13] Tahamtan, Alireza, and Abdollah Ardebili. "Real-time RT-PCR in COVID-19 detection: issues affecting the results." *Expert review of molecular diagnostics* 20.5 (2020): 453-454.

[14] Lu, Y. (2022). Honor thesis code (Version 1.0.0) [Computer software]. https://github.com/ylu268/Honor_thesis_file

[15] Levine-Tiefenbrun, Matan, et al. "Initial report of decreased SARS-CoV-2 viral load after

inoculation with the BNT162b2 vaccine." Nature medicine 27.5 (2021): 790-792.

[16] Gozzi, Nicolò, et al. "Preliminary modeling estimates of the relative transmissibility and immune escape of the Omicron SARS-CoV-2 variant of concern in South Africa." *medRxiv* (2022).