

# Clustering and Identification of SARS-CoV-2 Mutations Associated with Clinical Severity

Jingxue Feng<sup>\*</sup>, Jie Wang<sup>†</sup>, Jiarui Zhang<sup>‡</sup>, and Liangliang Wang<sup>§</sup>

*Department of Statistics and Actuarial Science, Simon Fraser University*

January 2021

## Abstract

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus emerged in December 2019 in China that causes the outbreak of COVID-19 worldwide. The genetic cluster analysis of SARS-CoV-2 variants is crucial to characterize the virus and has been widely studied. However, the existing genetic clustering methods are merely based on whole genome sequencing data without giving consideration to clinical features. In our work, with the involvement of both genome sequencing data and clinical data, we developed a model-based clustering method to group SARS-CoV-2 mutations that share similar relationship to the clinical features, with a focus in disease severity. Parameters in the model are estimated via Bayesian inference that takes model uncertainty into account. SARS-CoV-2 mutations are finally classified into three interpretable clusters. One cluster is strongly associated with moderately severe and severe patients, one cluster is moderately associated with disease severity, and the third cluster is strongly associated with the asymptomatic patients. Our analysis facilitates the process of identifying clusters of SARS-CoV-2 mutations, and simultaneously provides insights into the association between mutations and clinical features.

## 1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel coronavirus emerged in Wuhan, China at the end of year 2019. With a high person-to-person transmission rate, the coronavirus caused the outbreak of coronavirus not only in China, but also in other countries across the world. Symptoms of this coronavirus disease (COVID-19) include but not limited to fever, headache, cough, chills, difficult breathing, etc. Different infected individuals may have different symptoms based on their physical conditions. World Health Organization (WHO) officially announces COVID-19 outbreak a pandemic in March 2020. As of January 2021, over 90 million confirmed cases have been reported worldwide, including around 1.95 million deaths (<https://covid19.who.int/>).

---

<sup>\*</sup>jingxuef@sfu.ca

<sup>†</sup>wangjiew@sfu.ca

<sup>‡</sup>jiaruiz@sfu.ca

<sup>§</sup>lwa68@sfu.ca

Jingxue Feng, Jie Wang, and Jiarui Zhang contribute equally to this report.

Recently, the evolution of SARS-CoV-2 coronavirus has been widely studied. Zhou et al. (2020) showed that SARS-CoV-2 has a 96% genome sequence similarity with a bat coronavirus, indicating that the original natural host of coronavirus might be bat. The phylogenetic analysis of SARS-CoV-2 genomes (Forster et al., 2020; Yang et al., 2020; Laamarti et al., 2020) suggested to cluster genomes by certain variants. Forster et al. (2020) used 160 complete human SARS-Cov-2 genomes to identify three central variants, labeled as type A, B and C. Type A is the ancestral viral genomes. Type B is derived from A by mutations C<sup>8782</sup>T and T<sup>28144</sup>C. Type C differs from type B by carrying the mutation G<sup>26144</sup>T. Whereas, Mavian et al. (2020) later criticized their work in terms of small sample size and incorrect phylogenetic network root. Forster et al.’s genetic clustering result was later confirmed by Yang et al. (2020) via phylodynamic analysis, where four super-spreader clusters were identified based on 247 SARS-CoV-2 genomes collected worldwide. Yang et al. (2020) reported that the first cluster carries mutations C<sup>8782</sup>T and T<sup>28144</sup>C; the second cluster carries the mutation G<sup>26144</sup>T; the third cluster carries the mutation G<sup>11083</sup>T; the fourth cluster carries the mutations C<sup>241</sup>T, C<sup>3037</sup>T and A<sup>23403</sup>G. According to 3,067 SARS-CoV-2 genomes collected worldwide, Laamarti et al. (2020) then constructed phylogenetic trees to reveal the two major clades that carry the mutations F<sup>3606</sup>G and D<sup>614</sup>G on protein separately. Besides phylogenetic studies, the frequencies of 16 common mutations were used to classify 28 countries into 3 clusters through hierarchical clustering, and fatality rates were demonstrated based on clusters (Toyoshima et al., 2020). One common disadvantage of these studies is that some important clinical features are not considered into the process of clustering. Moreover, as more relevant data become available, the sample data used in these current studies is not large enough to be persuasive.

We propose a model-based clustering technique to group the SARS-CoV-2 mutations based on the relationship between mutations and clinical features. The proposed method is related to Qin and Self (2006), in which the clustering of regression model (CORM) is used to cluster genes with similar relationship to the covariates. The CORM method incorporates the cluster membership per gene and assumes the genes within the same cluster share the same regression coefficients. Motivated by the idea of CORM, we propose a model-based clustering method, called clustering of logistic regression model (CLRM), to cluster the SARS-CoV-2 mutations. We perform genetic clustering by analyzing more than 8,000 observed SARS-CoV-2 genome sequences and the corresponding clinical features of the hosts, and finally group the most common mutations identified from global genome sequencing data by considering the relationship to the clinical features. Instead of fitting the model via an expectation–maximization algorithm (Qin and Self, 2006), the no-U-turn sampling (NUTS) algorithm (Hoffman and Gelman, 2014), one type of the Monte carlo Markov Chain (MCMC) methods, is implemented to estimate parameters in the CLRM model. The uncertainties of models and parameters are taken into full account while using the Bayesian approach. Our work is beneficial for recovering the group of SARS-CoV-2 mutations and provide insights into how the SARS-CoV-2 mutations are associated with clinical features, especially with disease severity. We identify three clusters of mutations that are associated the severity of COVID-19 differently. One cluster is strongly associated with moderately severe and severe patients, one cluster is moderately associated with disease severity, and the third cluster is strongly associated with the asymptomatic patients.

The advantages of our proposed method are summarized as follows. First, the proposed model can incorporate both SARS-CoV-2 genome data and patients’ clinical data in a systematic way. Therefore, it can provide more accurate results than first clustering only based on the genome data and then relating the clusters to the clinical features (Toyoshima et al., 2020). Second, our model controls confounding variables when considering the association between mutations and disease severity. Third, the obtained clusters are easily interpretable. Our model can quantify the association between mutations and disease severity using odds ratios of being in one cluster for two groups

of patients with different COVID-19 severity levels. Fourth, our Bayesian statistical inference can provide uncertainty estimates besides the point estimation.

The rest of report is organized as follows. In Section 2, we describe our COVID-19 global data resources. In Section 3, we introduce the CLRM method as well as the Bayesian inference of parameters in the model. In Section 4, the CLRM method is applied to our global data set and some meaningful results are presented. The conclusion and discussion about this work are provided in Section 5.

## 2 Data Description

Our data are composed of two parts: genome sequence data and clinical data. All data were downloaded from the GISAID platform (Shu and McCauley, 2017). The GISAID Initiative (<https://www.gisaid.org/>) provides open-access to the novel coronavirus responsible for COVID-19. We will illustrate the two major parts of data in the following subsections.

### 2.1 Extract mutations from genomes

A total of 9,355 complete SARS-CoV-2 genomes ( $> 29,000$  base pairs) with Nov 30, 2020 as cut-off date were downloaded from the GISAID database. The FASTA file of complete reference SARS-CoV-2 genome is downloaded from NCBI Genbank ([https://www.ncbi.nlm.nih.gov/nucleotide/NC\\_045512?%3Fdb=nucleotide](https://www.ncbi.nlm.nih.gov/nucleotide/NC_045512?%3Fdb=nucleotide)). The length of the reference SARS-CoV2 genome is 29,903 base pairs. The relative variants on genomes are identified using the tool Snippy (<https://github.com/tseemann/snippy>). Specifically, the variant calling process identifies the differences between the genome sample from a patient and the known reference genome. The differences could involve single nucleotide polymorphisms (SNPs) and insertions/deletions. These genetic differences will be regarded as mutations.

After examining the 9,355 patients' genomes and clinical data, we removed 1,078 patients' records due to missing information on sex, age and patient status. Out of the 8,277 SARS-CoV-2 genomic sequences after data-cleaning process, Snippy identifies 8,877 mutation events in total compared to the Wuhan-Hu-1 reference genome. The distribution of the genome sequence length is shown in Figure A.2 in Appendices. In order to avoid rare-event bias in logistic regression, we use 5% of the number of SARS-CoV-2 genomic sequences as the cutoff to select the most frequent 19 mutations for further cluster analysis. The selected mutations with corresponding counts and types are shown in Figure 1. The mutations are written in the form "nucleotide(s) in the reference + position in the reference sequence + alternate nucleotide(s) in the individual genomic sequence". Mutation types are classified by the SNP transitions ( $A \leftrightarrow G$  or  $C \leftrightarrow T$ ) and SNP transversions ( $A/G \leftrightarrow T/C$ ). Two unusual tri-nucleotide mutations are also observed in our global data. According to the table in Figure 1, the single-nucleotide transitions are the major worldwide mutation events, which highly agrees with those findings for the SARS-CoV-2 mutations (Mercatelli and Giorgi, 2020). Interestingly, while classifying the mutations based on super-spreader clusters defined by Yang et al., a majority of cases belongs to the 4-th cluster, which implies the frequent co-occurrence of mutations  $C^{241}T$ ,  $C^{3037}T$  and  $A^{23403}G$  in our global genome sequencing data. Readers may refer to Figure A.1 in Appendices for more details.

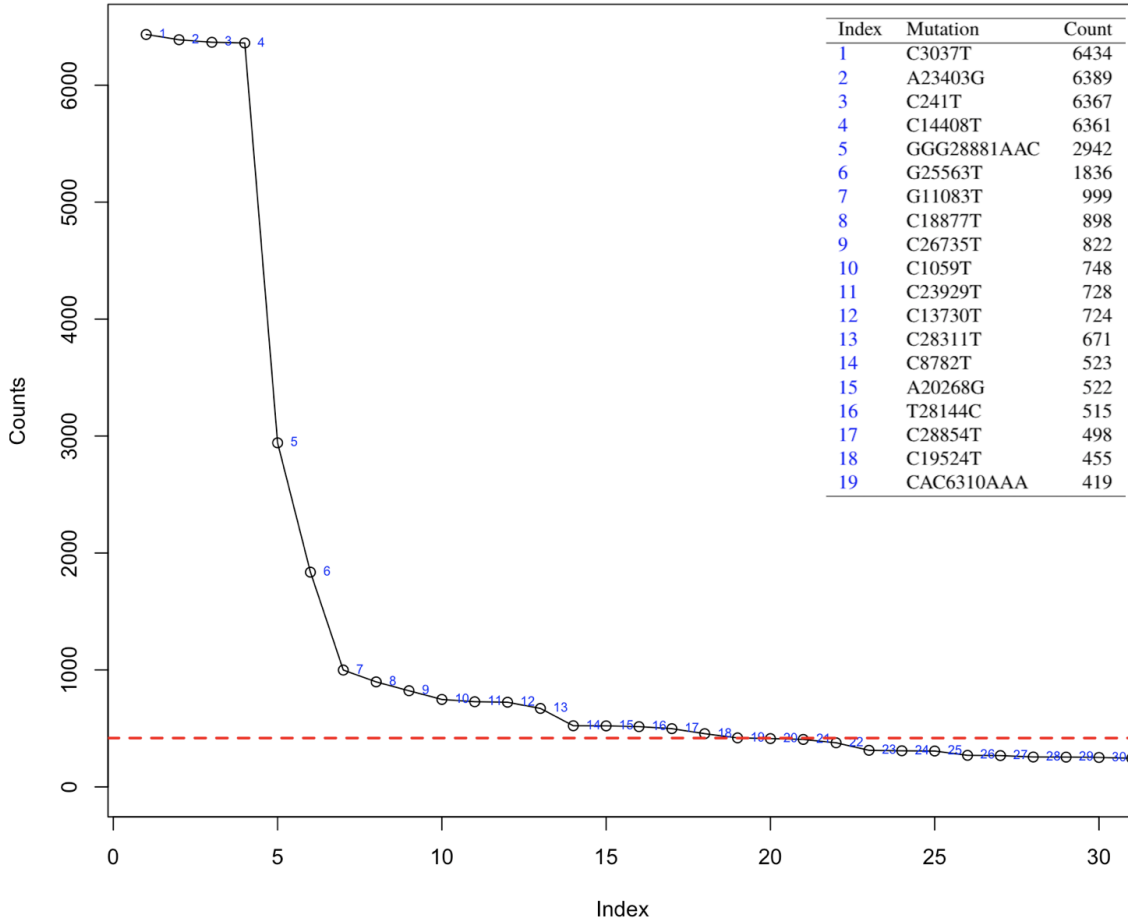


Figure 1: The top 30 frequency top mutations in Global data. Note that the red dashed line indicates the 5% cut-off of all mutations. The top right of the plot summarizes the top 19 mutations obtained from 8,277 sequenced SARS-CoV-2 genomes.

## 2.2 Clinical data

The clinical data provide information on each patient’s location, sex, age, status and so on. We selected sex, age and patient status as the covariates of interest. In the original clinical data, patient status contains several categories, such as Deceased, Recovered, Mild, Moderate, and Severe. We chose to group similar categories. Specifically, the raw data contains more than 20 categories since each lab’s wordings of the patient status are defined differently. Therefore, we further grouped the original categories into 4 main categories. The 4 categories are ordered as follows based on the disease severity: “asymptomatic”, “others”, “recovered”, “severe / deceased”, where “others” mainly includes quarantined, mild and moderate cases, and “recovered” is related to moderately severe cases. For the cleaned global data, Figure 2 displays the number of cases in each patient status category and each continent at the global level. It can be seen that the proportion of asymptomatic cases is extremely small (1.39%), which is mainly contributed by Asia and Europe as shown in Figure 3. Although a certain number of complete genome sequences are collected from Oceania, only a few cases have known patient status and survived through our data cleaning process. Therefore, from Figure 2 (b), we see a valley for the count in Oceania.

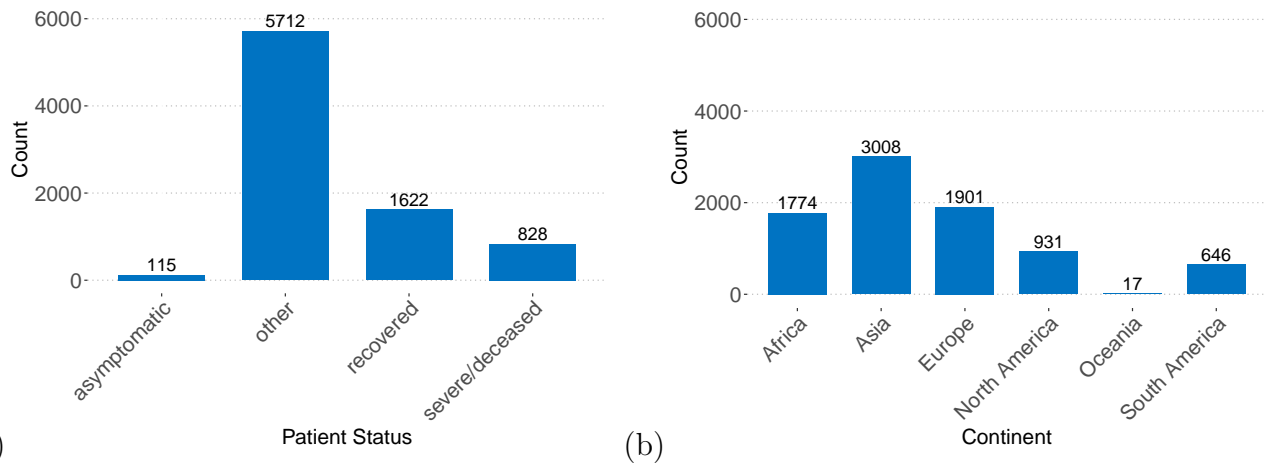


Figure 2: (a) Number of cases in each patient status category at global level; (b) Number of cases in each continent at global level.

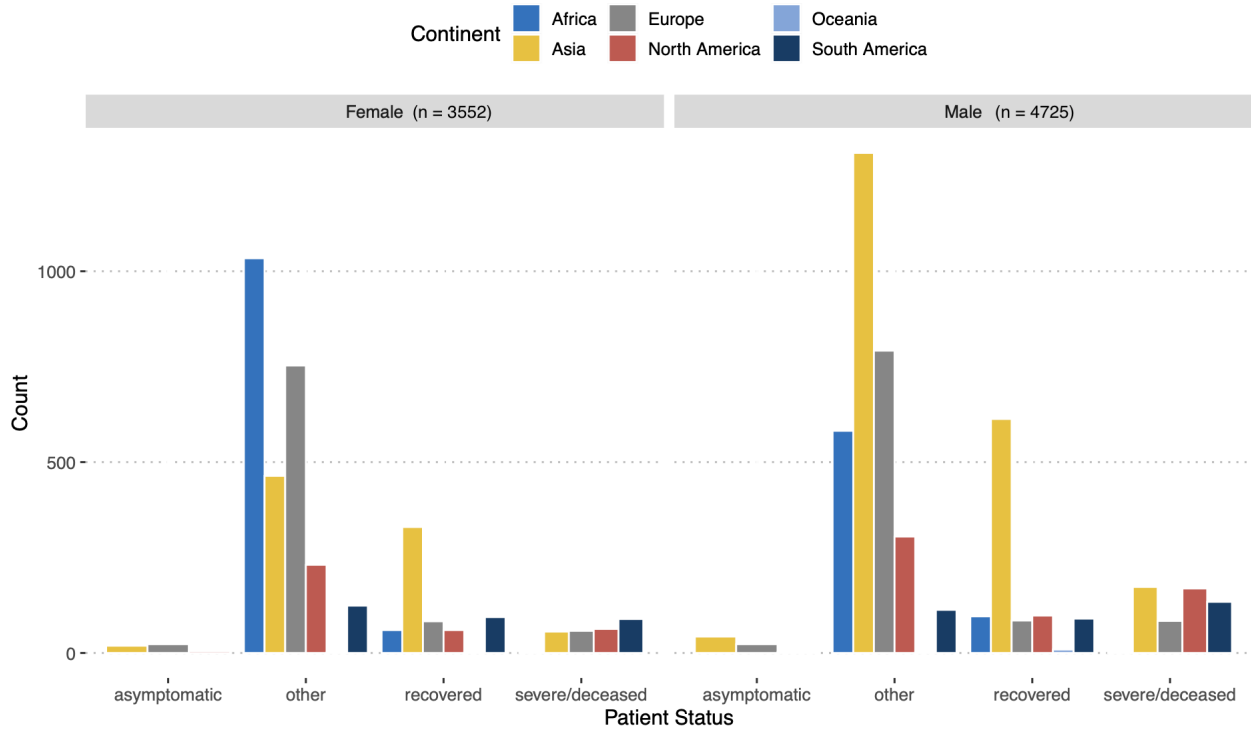


Figure 3: The distribution of patient status by continent and sex in 8,277 observations.

### 3 The Clustering of Logistic Regression Model

The CLRM is a model-based clustering method for clustering mutations based on their association with the covariates of interest. In this case, the covariates are the clinical data illustrating the characteristics of patients, including age, sex, and patient status. Particularly, we are interested in the relationship between the most common SARS-CoV-2 mutations and the patient status while controlling the main confounding variables age and sex.

From herein, we introduce some notations for the general framework of CLRM. Suppose the data consists of  $N$  observations and  $D$  covariates. Let  $M$  denote the number of top mutations detected from all the genome sequences. The relationship between top  $M$  mutations and  $D$  covariates are of our interest. Let  $\mathbf{y}_{mn}$  ( $M \times N$ ) denote the indicator matrix using binary values (0 or 1) to indicate the existence of mutation  $m$  on  $n$ -th individual genome sequence, and  $\mathbf{x}$  ( $N \times (D+1)$ ) be the design matrix in the logistic regression. Assume there are  $K$  clusters. Let  $\boldsymbol{\beta}_k$  ( $(D+1) \times 1$ ) denote the logistic regression coefficients + intercept for mutations in cluster  $k$ . Denote the cluster membership for mutation  $m$  as  $Z_m$ , where  $Z_m \in \{1, \dots, K\}$ . The logistic regression underlying the CLRM method can be written as

$$\log \left( \frac{\pi_{mn}}{1 - \pi_{mn}} \middle| Z_m = k, \mathbf{x}_n \right) = \mathbf{x}_n^\top \boldsymbol{\beta}_k, \quad (1)$$

where  $\mathbf{x}_n = [1, x_{n1}, x_{n2}, \dots, x_{nD}]^\top$ , and  $\pi_{mn}$  is the probability of having the  $m$ -th mutation on the  $n$ -th genome sequence when this mutation belongs to the  $k$ -th cluster. Let the latent variable  $\mathbf{Z} = [Z_1, \dots, Z_M]^\top$  represent the cluster memberships for all mutations, and  $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_K^\top]^\top$  represent the coefficient matrix consists of coefficients in each cluster. The likelihood function of the unknown parameters are

$$p(\mathbf{y}|\boldsymbol{\beta}, \mathbf{Z}) = \prod_{n=1}^N \prod_{m=1}^M [\text{Bernoulli}(y_{mn}|\pi_{mn}, Z_m)] = \prod_{n=1}^N \prod_{m=1}^M [\pi_{mn}^{y_{mn}} (1 - \pi_{mn})^{1-y_{mn}}], \quad (2)$$

where

$$\pi_{mn} = \frac{\exp(\mathbf{x}_n^\top \boldsymbol{\beta}_{Z_m})}{1 + \exp(\mathbf{x}_n^\top \boldsymbol{\beta}_{Z_m})}. \quad (3)$$

Our aim is to perform Bayesian statistical inference in CLRM conditional on some observations  $\mathbf{y}_{mn}$ , treating both the latent variable  $\mathbf{Z}$  and the parameter  $\boldsymbol{\beta}$  as unknowns. The full Bayesian inference with MCMC sampling is performed in Stan (Carpenter et al., 2017). Since it is not able to sample the discrete  $Z_m$  directly in Stan, this discrete latent variable is marginalized across clusters

$$y_{mn}|\pi_{mn}, \boldsymbol{\lambda}_m \sim \sum_{k=1}^K \lambda_{mk} \text{Bernoulli}(y_{mn}|\pi_{mn}, Z_m = k), \quad (4)$$

where  $\boldsymbol{\lambda}_m = [\lambda_{m1}, \dots, \lambda_{mK}]^\top$  with  $\lambda_{mk} = P(Z_m = k)$ , which is the probability that the mutation  $m$  belongs to cluster  $k$ . The likelihood function in Equation (2) can be rewritten as

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda}) = \prod_{n=1}^N \prod_{m=1}^M [\text{Bernoulli}(y_{mn}|\pi_{mn}, \boldsymbol{\lambda}_m)] = \prod_{n=1}^N \prod_{m=1}^M \sum_{k=1}^K \lambda_{mk} [\pi_{mn}^{y_{mn}} (1 - \pi_{mn})^{1-y_{mn}}]. \quad (5)$$

For this model, the parameters of interest are  $\boldsymbol{\beta}$  and  $\boldsymbol{\lambda}$ . We ascribe a prior density  $p(\boldsymbol{\beta})$  to  $\boldsymbol{\beta}$  and a prior density  $p(\boldsymbol{\lambda})$  to  $\boldsymbol{\lambda}$ . So Bayesian inference depends on the following joint density

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})p(\boldsymbol{\beta})p(\boldsymbol{\lambda}),$$

where  $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\lambda})$  can be factorized as in Equation (5). The posterior of  $Z_m$  can be computed via Bayes's rule via the following formula

$$p(Z_m = k|\mathbf{y}_m, \boldsymbol{\lambda}_m) = \frac{p(\mathbf{y}_m|Z_m = k)p(Z_m = k|\boldsymbol{\lambda}_m)}{\sum_{k'=1}^K p(\mathbf{y}_m|Z_m = k')p(Z_m = k'|\boldsymbol{\lambda}_m)}, \quad (6)$$

where  $\mathbf{y}_m = (y_{m1}, \dots, y_{mN})^\top$ .

## 4 Data analysis

We constructed a data analysis pipeline to enable exploration of potentially interesting mutation clusters from SARS-CoV-2 sequences and clinical data. Figure 4 shows the pipeline schematic in our analysis. In summary, Section 2 covers the process of filtering sequences and identifying mutations. Out of the 8,877 mutation events identified in Snippy, we selected the top 19 mutations with counts greater than the cutoff. We then applied the methodology introduced in Section 3 to these top 19 mutations and clinical data to obtain the clusters of mutations and coefficients.

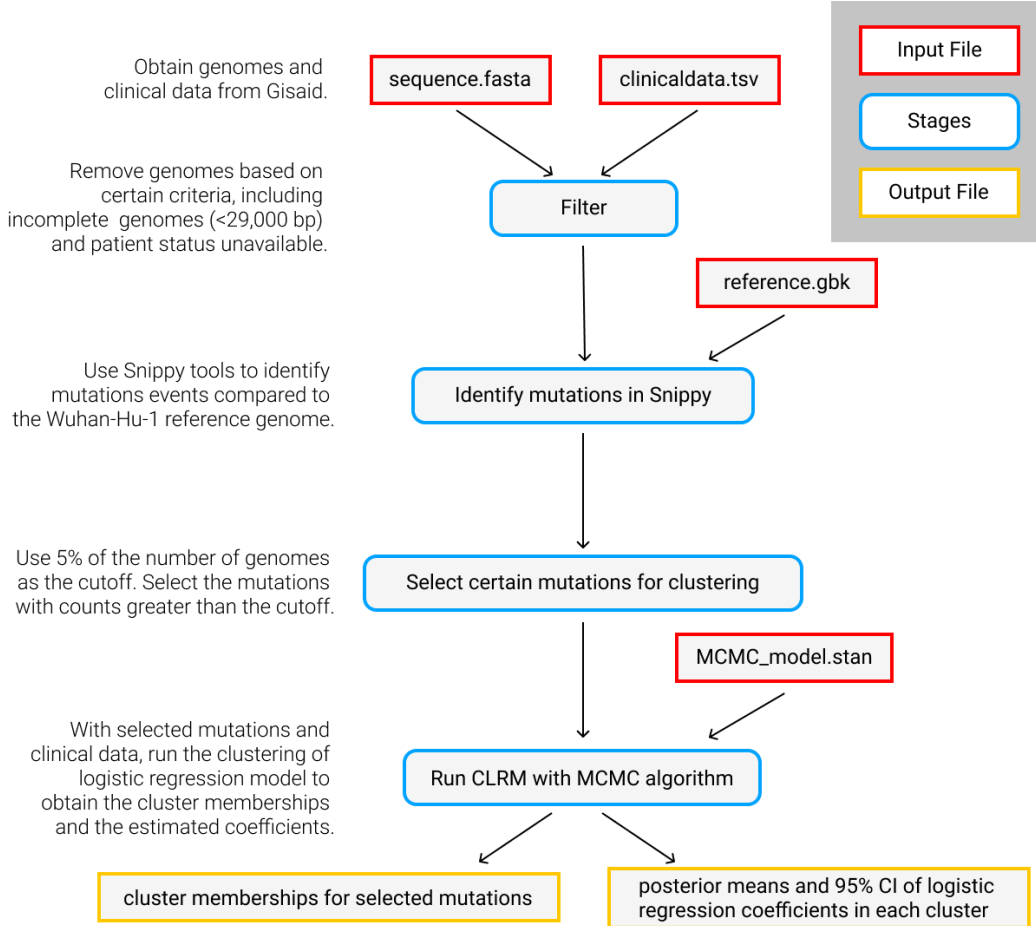


Figure 4: The pipeline of the clustering analysis of the COVID-19 data.

Here we discuss our choices of prior distributions on  $\beta$  and  $\lambda$ . The initial values for  $\beta$  (coefficients + intercept) were selected according to the following procedure: first, we obtained the maximum likelihood estimators (MLEs) of  $\beta$  for each mutation; second, the K-means clustering algorithm was performed on those MLEs to get an initial clustering; within each cluster of mutations, we ran the logistic regression again using the response variables and covariates as being in this cluster to get the initial values for  $\beta$ . Let  $\beta_{MLE,k}$  denote the vector of starting values for  $\beta$  in cluster  $k$ . We assumed a multivariate normal prior on  $\beta_k$  with mean vector  $\hat{\beta}_{MLE,k}$  and variance-covariance matrix  $\sigma^2 \mathbb{I}_{D+1}$ , where  $\sigma^2 = 0.5$ . As for the prior of  $\lambda$ , we assumed  $\lambda_m \sim \text{Dirichlet}(\alpha)$ , independently for  $m = 1, \dots, M$ . We further assumed the hyperprior  $\alpha$  to be an all-ones vector of length  $K$ .

To determine the number of clusters, we run our model with  $K = 2, 3, 4, \dots$ . Each time, we look at the estimated  $\beta$ . If the estimates of these coefficients for two clusters are quite similar, we

stop increasing the number of clusters. The no-U-turn sampling (NUTS) algorithm (Hoffman and Gelman, 2014) was executed by using 5,000 iterations and a burn-in of 2,500 iterations for 3 chains to ensure that the Markov chain has converged. As a result of all attempts,  $K = 3$  is selected in this study:

- Cluster 1: C<sup>3037</sup>T, A<sup>23403</sup>G, C<sup>241</sup>T, and C<sup>14408</sup>T;
- Cluster 2: G<sup>11083</sup>T, C<sup>18877</sup>T, C<sup>26735</sup>T, and C<sup>8782</sup>T;
- Cluster 3: GGG<sup>28881</sup>AAC, G<sup>25563</sup>T, C<sup>1059</sup>T, C<sup>23929</sup>T, C<sup>13730</sup>T, C<sup>28311</sup>T, A<sup>20268</sup>G, T<sup>28144</sup>C, C<sup>28854</sup>T, C<sup>19524</sup>T, and CAC<sup>6310</sup>AAA.

The logistic regression coefficients  $\beta$  are converted into odds ratios through the natural exponential function for easier interpretations. Table 1 shows a summary of estimated odds ratios and the corresponding 95% credible intervals (CI) resulting from the MCMC for each cluster. To identify the mutations that are associated with the clinical covariates, we examined the 95% CIs of these estimated odds ratios to see if they contain 1 or not. If the 95% CI does not cover 1, we conclude this covariate is associated with the mutations in this cluster. We use the magnitude of the odds ratio to study the strength of the association. An odds ratio of 10 or above suggests a strong association. We will interpret each of the three identified clusters of mutations in terms of their associations with the patient status, age and sex.

		Cluster 1			Cluster 2			Cluster 3		
		Posterior mean	95 % LCI	95 % UCI	Posterior mean	95 % LCI	95 % UCI	Posterior mean	95 % LCI	95 % UCI
Age		1.043	1.034	1.054	0.363	0.179	0.636	0.382	0.183	0.650
Sex	Level = "male"	0.038	0.023	0.063	0.721	0.306	1.619	1.084	0.485	2.452
Patient status	Level = "others"	2.586	1.507	4.331	1.081	0.468	2.533	0.145	0.060	0.330
	Level = "recovered"	15.358	8.227	28.516	2.549	1.095	5.587	0.120	0.050	0.291
	Level = "severe/deceased"	36.779	18.030	77.744	3.453	1.281	9.476	0.063	0.024	0.166

Table 1: Summary of estimated odds ratio and 95% CI resulting from the MCMC algorithm for each cluster. Note that "95% LCI" and "95% UCI" refer to the lower and upper bounds of 95% CI.

From Table 1, the mutations in Cluster 1 are strongly associated with moderately severe and severe patients. More specifically, the odds of having the mutations in Cluster 1 for patients with mild or moderate symptoms is 2.6 (1.5, 4.3) times of the odds of having such mutations for asymptomatic patients; the odds ratio of having these mutations for recovered patients versus the asymptomatic patients is 15.4 (8.2, 28.5); the odds ratio of having these mutations for severe/deceased patients versus asymptomatic patients is 36.8 (18.0, 77.7). Moreover, sex is also strongly associated with the mutations in Cluster 1. The estimated odds of female patients who having the mutations in the first cluster, holding other covariates constant (i.e., same age and patient status), are 26.5 (15.9, 42.8) times the estimated odds of that for male patients. Furthermore, age is not strongly associated with mutations in this cluster. The odds ratio of being in this cluster associated with a five year increase in age is 1.24 (1.18, 1.23).

Cluster 2 has a strong association with age, a moderate association with disease severity, and no association with sex. The odds of being in Cluster 2 for severe/deceased patients is 3.5 (1.23, 9.5) times of the odds for asymptomatic patients. The odds ratio of being in this cluster associated with a five year decrease in age is 158.0 (9.6, 5467.9).

Cluster 3 has a strong association with the asymptomatic patients, a strong association with age, and no association with sex. The odds of being in Cluster 3 for asymptomatic patients is 15.8 (6.0, 41.7) times of the odds for severe/deceased patients, and 8.4 (3.4, 20.0) times of the odds for moderately severe (recovered) patients. The odds ratio of being in this cluster associated with a five year decrease in age is 122.8 (8.6, 4807.7).

For each continent, we calculated the percentage from counts of cases carrying any of the mutations within each cluster based on estimated cluster memberships. The pie chart map shown in Figure 5 illustrates the proportions of counts being in each cluster for each continent. From Figure 5, all of the continents except Oceania display a similar pattern in the pie chart. For Europe, Africa, North America, and South America, the proportions of cases carrying mutations in Cluster 1 and Cluster 3 are high, indicating that the mutations in Cluster 1 and Cluster 3 are prevalently present in these four continents. For Asia, the percentage of cases carrying mutations in Cluster 3 is relatively higher than that of Cluster 1 and 2. From the estimated results of within-cluster coefficients, we realized that the mutations in Cluster 1 and Cluster 3 are strongly associated with the patient status, which may bring insights into the transmission of SARS-CoV-2 in relevant locations. For Oceania, it is hard to draw a valid conclusion due to the limited number of cases in our analysis.

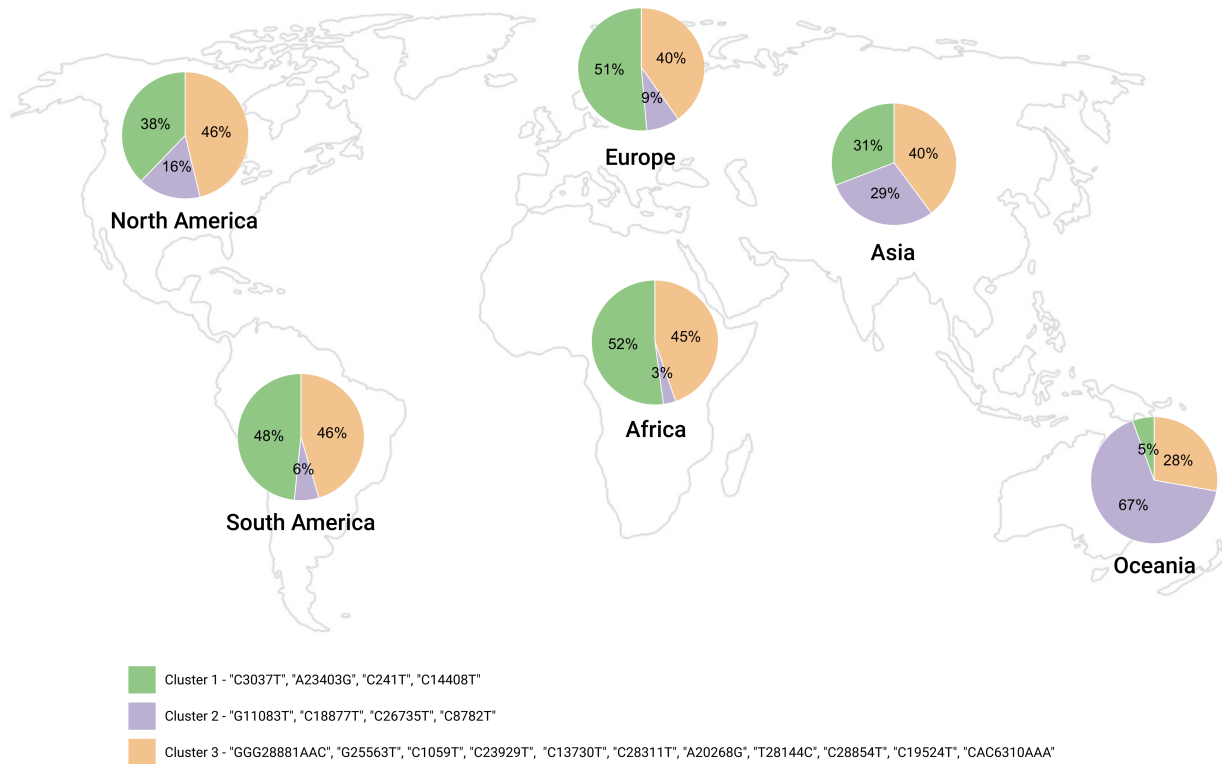


Figure 5: The percentage of counts being in the clusters of mutations for each continent.

## 5 Conclusion and discussion

We have developed a model-based clustering method, named the clustering of logistic regression model (CLRM), to group the binary responses based on their relationships with the covariates. The proposed method is applied to the COVID-19 data to find clusters of SARS-CoV-2 mutations that

are associated with clinical features of hosts. More specifically, we consider the top 19 most common mutations from a total of 8,877 mutations, and the clinical features of our interest include age, sex, and disease severity.

As a result of our study, three clustered memberships of mutations were obtained by the CLRM. The three obtained clusters of mutations have a clear interpretation in terms of their association with the patients' severity level of COVID-19. Mutations in Cluster 1 are strongly associated with moderately severe and severe patients; Cluster 2 has a moderate association with disease severity; Cluster 3 have a strong association with the asymptomatic patients. Particularly, the mutations (i.e., A<sup>23403</sup>G, C<sup>241</sup>T, C<sup>3037</sup>T, C<sup>14408</sup>T) in Cluster 1 are widespread in four continents like Africa, Europe, North America, and South America. Also, these mutations are known as the four most frequent mutations in SARS-Cov-2 genomes (Mercatelli and Giorgi, 2020). Additional information can be referred to Figure A.3 in Appendix.

Note our proposed CLRM is essentially a type of mixture models, which may suffer from the so-called label switching problem when a Bayesian approach is used to do parameter estimation and clustering. The label switching problem is caused by symmetry in the likelihood of the model parameters (Stephens, 2000). When symmetric prior distributions are used, both the likelihood and posterior are invariant to permutations of the cluster memberships. In a Bayesian context, this invariance makes the MCMC samples suffer from the non-identifiable problem. To deal with this label switching problem, we used asymmetric prior distributions and carefully chose initialization to help the MCMC chains to stay at one local mode. In particular, we assigned a set of specific starting values for  $\beta$  obtained from the K-means clustering algorithm when initializing the MCMC chains. We also gave an asymmetric prior to control the chains that focus on the local modes. Our current solution to this problem can be improved by exploring advanced probabilistic and deterministic relabeling algorithms implemented by Papastamoulis (2015).

The usefulness of our proposed method, demonstrated in this report, indicates the importance of collecting more clinical data of higher quality. Although we can obtain a total of 229,422 complete SARS-CoV-2 genomes within the time interval from GISAID, only around 4% (9,355 cases) have entries of patient status available. The percentage of cases with known patient status in Oceania is as low as 0.1%, whereas the percentages in Africa, Asia, and South America are around 30%. The imbalanced sample sizes across the continent add uncertainties into the continent-based interpretation. In terms of the clinical data quality, the raw data contains a certain number of groups for patient status, and the definition of each group varies considerably lab by lab. We pre-processed the clinical data by manually grouping patient status into four major categories. The category "others" includes the largest number of cases, which can be further split into a few meaningful subcategories to draw a more accurate conclusion.

Our current methodology of the CLRM can be improved in several directions as future work. First, we can incorporate the number of clusters  $K$  into the current model by replacing the Dirichlet distribution for the parameter of probabilities of being in  $K$  clusters with a Dirichlet process. Second, we can extend the current model to be a time-varying one to describe the dynamics of time-dependent mutations. Third, given SARS-CoV-2's fast evolutionary rate and rapid growth of cases, random emergence of new mutations is entirely expected, even in a relatively short time frame. An efficient online clustering method that can adapt the latest data to the current analysis result without starting from scratch will be an exciting topic for further research.

# A Appendices

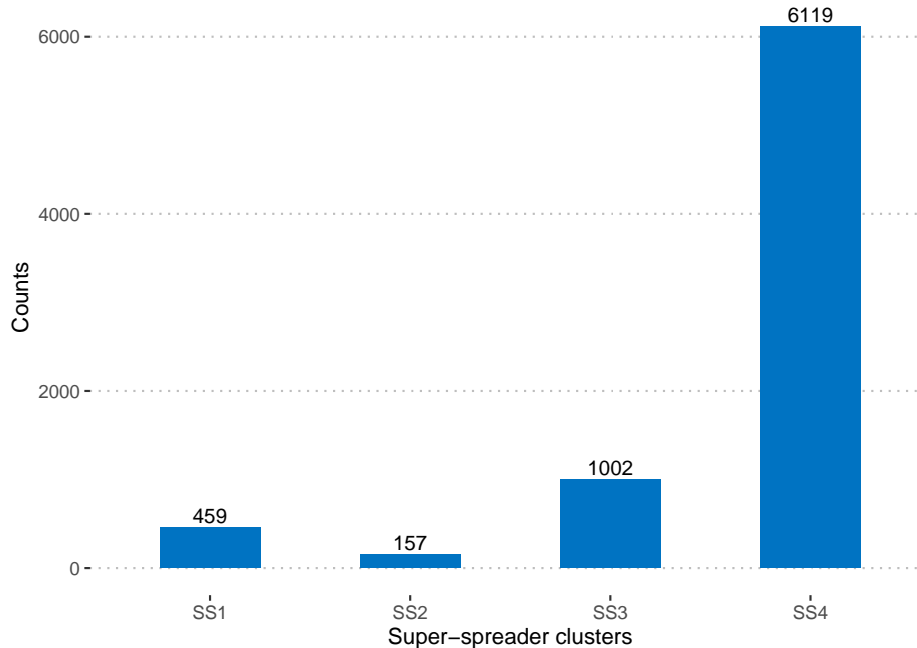


Figure A.1: Motivated by the super-spreader clusters in Yang et al.’s paper, super-spreader cluster 1 (SS1) carried signature mutations  $C^{8782}T$  and  $T^{28144}C$ ; super-spreader cluster 2 (SS2) carried the signature mutation  $G^{26144}T$ ; super-spreader cluster 3 (SS3) carried the signature mutation  $G^{11083}T$ ; super-spreader cluster 4 (SS4) carried the signature mutation  $C^{241}T$ ,  $C^{3037}T$  and  $A^{23403}G$ . We classifies 8,277 genomes from global data according to these mutations defined in distinct super-spreader clusters. The number of genomes belong to each cluster is displayed in Figure A.1. An interesting finding is that a majority of patients belong to the 4-th super-spreader cluster. SS4 contributed to the early pandemic in Europe as of March 2020, then transmitted to other parts of the world (Yang et al., 2020). It suggests that most reported cases of COVID-19 worldwide might have been in Europe.

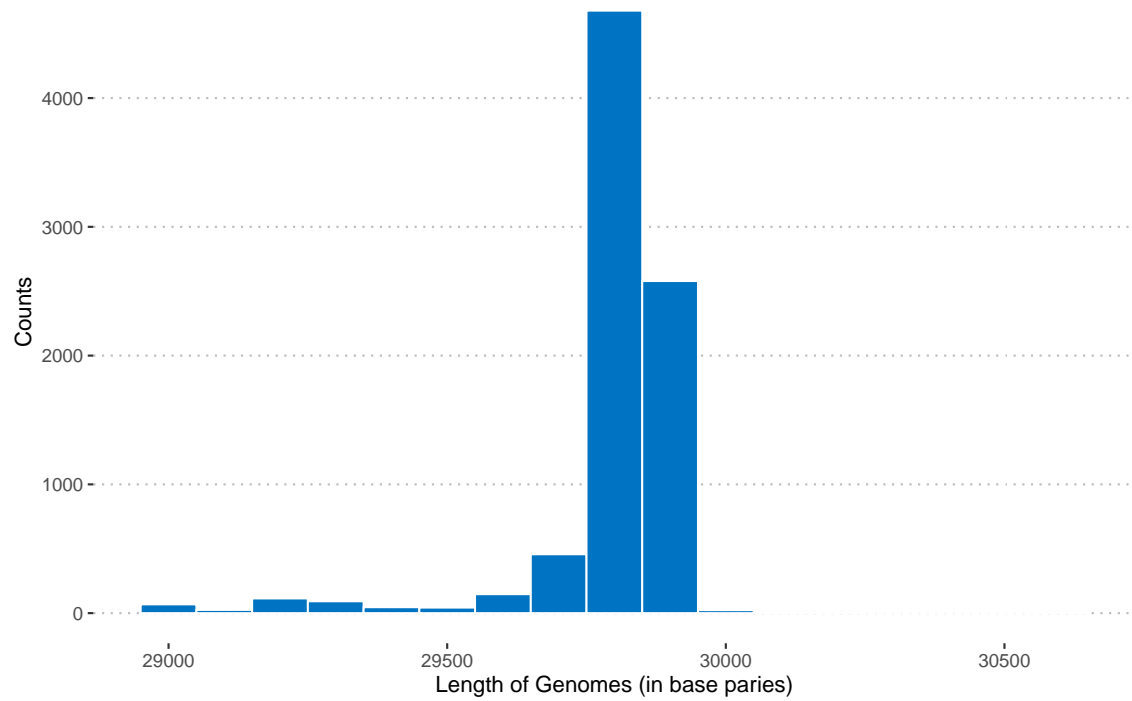


Figure A.2: Distribution of length of genome sequences collected globally. It can be seen that all genome sequences have  $> 29,000$  base pairs. Most observed genome sequences are 29,700 to 29,900 base pairs long.

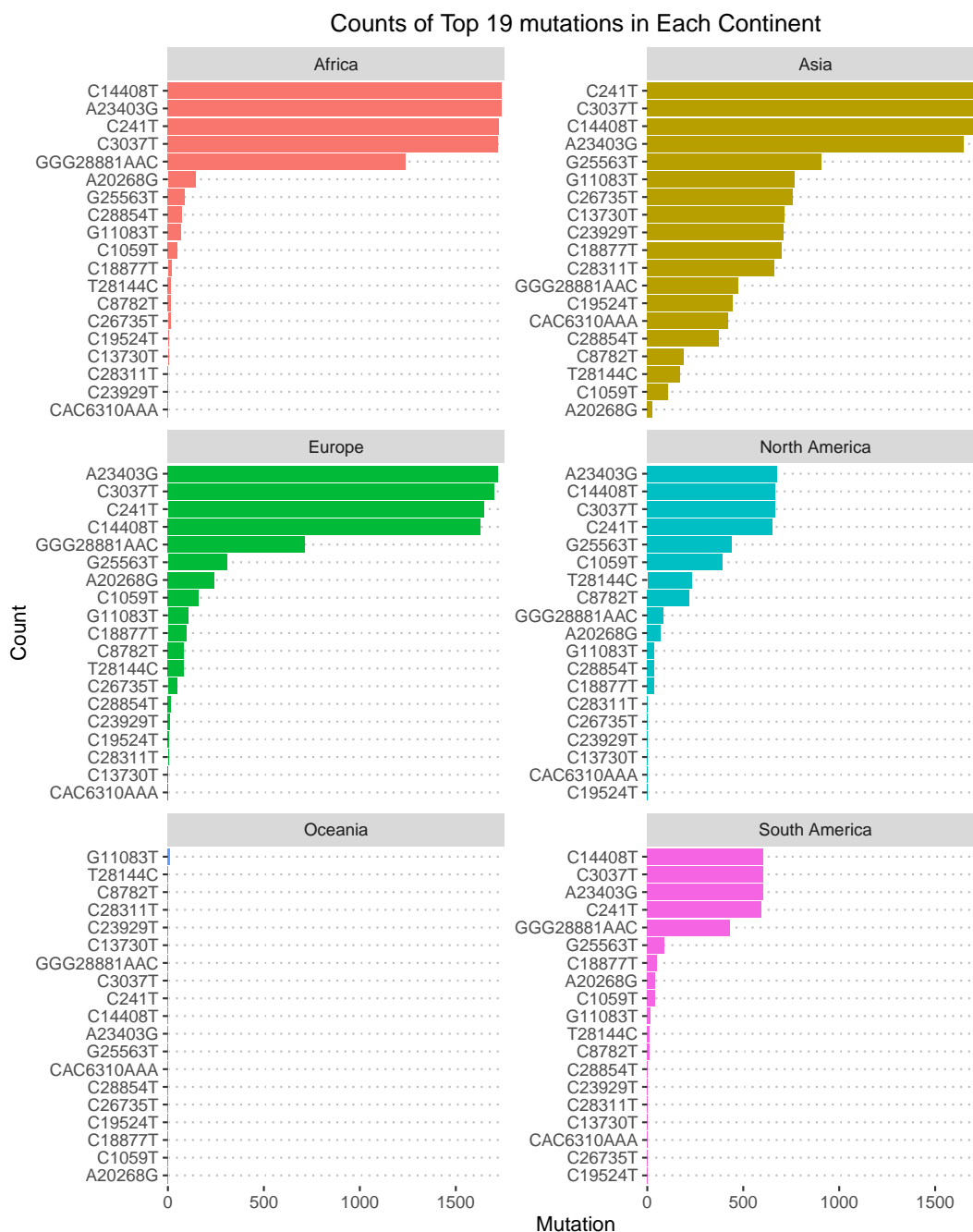


Figure A.3: The continent-based distribution of the most common 19 SARS-CoV-2 mutations identified in our study. It shows that the four mutations A<sup>23403</sup>G, C<sup>241</sup>T, C<sup>3037</sup>T, C<sup>14408</sup>T are the top four in Africa, Asia, Europe, North America, and South America. These four mutations are not identified as the top ones in Oceania because of small number of observations (17 samples) in this continent. The unusual tri-nucleotide mutation GGG<sup>28881</sup>AAC, observed as the 5-th most common mutation event globally, was mostly found in Africa and Europe.

## References

- B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.
- P. Forster, L. Forster, C. Renfrew, and M. Forster. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Sciences*, 117(17):9241–9243, 2020.
- M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.
- M. Laamarti, T. Alouane, S. Kartti, M. Chemaou-Elfihri, M. Hakmi, A. Essabbar, M. Laamart, H. Hlali, L. Allam, N. El Hafidi, et al. Large scale genomic analysis of 3067 sars-cov-2 genomes reveals a clonal geodistribution and a rich genetic variations of hotspots mutations. *bioRxiv*, 2020.
- C. Mavian, S. K. Pond, S. Marini, B. R. Magalis, A.-M. Vandamme, S. Dellicour, S. V. Scarpino, C. Houldcroft, J. Villabona-Arenas, T. K. Paisie, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of sars-cov-2 infections unreliable. *Proceedings of the National Academy of Sciences*, 117(23):12522–12523, 2020.
- D. Mercatelli and F. M. Giorgi. Geographic and genomic distribution of sars-cov-2 mutations. 2020.
- P. Papastamoulis. label.switching: An r package for dealing with the label switching problem in mcmc outputs. *Journal of statistical software*, 69(1):1–24, 2015. ISSN 1548-7660.
- L.-X. Qin and S. G. Self. The clustering of regression models method with applications in gene expression data. *Biometrics*, 62(2):526–533, 2006.
- Y. Shu and J. McCauley. Gisaid: Global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance*, 22(13):30494, 2017.
- M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.
- Y. Toyoshima, K. Nemoto, S. Matsumoto, Y. Nakamura, and K. Kiyotani. Sars-cov-2 genomic variations associated with mortality rate of covid-19. *Journal of human genetics*, 65(12):1075–1082, 2020.
- X. Yang, N. Dong, E. W.-C. Chan, and S. Chen. Genetic cluster analysis of sars-cov-2 and the identification of those responsible for the major outbreaks in various countries. *Emerging Microbes & Infections*, 9(1):1287–1299, 2020.
- P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273, 2020.