Statistical Detection of SARS-CoV-2 Variants based on PCR

Jingqi Zhu

September 2022

1 Introduction

SARS-CoV-2 is a single-stranded RNA virus capable of person-to-person transmission and has been identified as the cause of the global outbreak of respiratory illness starting in 2019.

SARS-CoV-2 virus is prone to mutation (insertion, deletion and substitution of nucleotide). The calling of variants is defined by the number and site of mutations based on sequencing result. Among the great number of variants, those confirmed to have detrimental changes in transmissibility, severity or immune evasion compared to the current dominant variant, or a growth rate potentially compatible with the eventual replacement of the current dominant variant are further categorised as variants of concern (VOC) by UKHSA. When there are several variants of concern circulating, for treatment purposes, it is of great importance to differentiate which variant each patient is infected with. While sequencing can be conducted on a small number of patients, RT-PCR test is far more efficient and less costly than sequencing, suggesting a potential to improve detection efficiency.

RT-PCR, namely real-time reverse transcription polymerase chain reaction, first reverse transcripts the extracted viral RNA to cDNA and then amplifies some target regions of cDNA in a designed thermal cycle. Whenever some target regions of cDNA are synthesised, the corresponding probes emit specific fluorescent signals that can accumulate. The outcome of RT-PCR are Ct values for each target region (described in Figure 1), the number of thermal cycles at which the fluorescent signal exceeds a limit. Therefore, a small Ct value indicates a high concentration of the target, while a large Ct value suggests a low concentration of the target or the target region is mutated (probes are designed based on the reference genome in Figure 2, mutations in the target region may give rise to decreased binding of the probes, therefore larger Ct values[5]).

There are a number of different RT-PCR assays used across the UK. Some are designed to identify a single target gene, while others detect multiple targets. The TaqPath PCR assay, which is a primary assay in NHS, targets three genes: S (spike) gene, N (nucleocapcid) gene and ORF1ab (open reading frame 1a

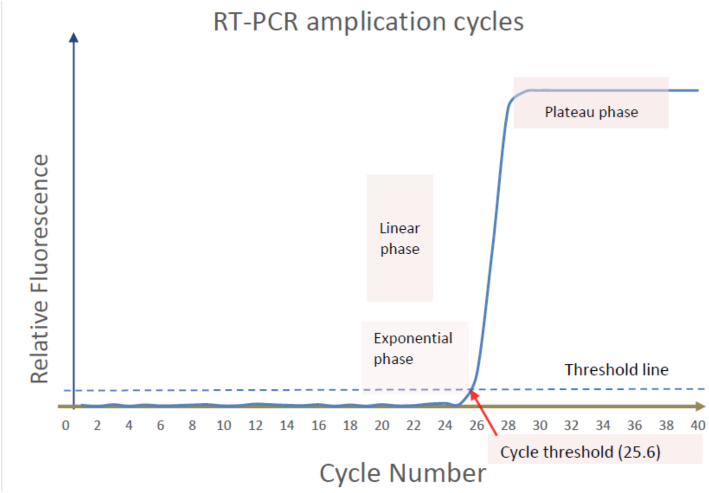
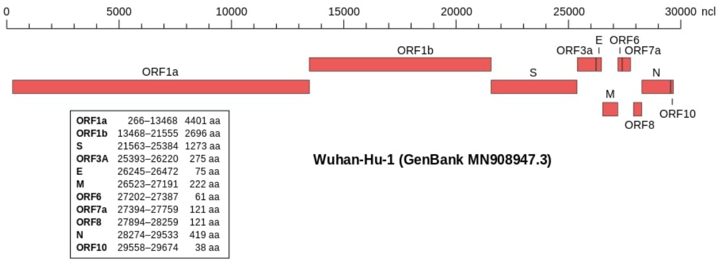


Figure 1: The outcome of RT-PCR assays are Ct values for each target region, the number of thermal cycles at which the fluorescent signal exceeds a limit[4].

and 1b) gene. The S gene is highly variable among variants. As it encodes the spike protein, a great number of mutations in S gene give rise to various structure of spike protein, which in turn lead to increased transmissibility and decreased antibody neutralization. The N gene encodes the nucleocapcid, a RNA-binding protein packaging the viral genome. The ORF1ab gene stands for ORF1a and ORF1b gene. They encode two polyproteins that are cleaved into 16 non-structural proteins, including RNA-dependent RNA polymerase (RdRp), helicase and various proteases. The N gene and ORF1ab gene are relatively conservative comparing to the S gene. Thus, they are also commonly used in other diagnostic RT-PCR assays, such as TaqPath PCR 2.0[1].



ORF1a	266–13468	4401 aa
ORF1b	13468–21555	2696 aa
S	21563–25384	1273 aa
ORF3A	25393–26220	275 aa
E	26245–26472	75 aa
M	26523–27191	222 aa
ORF6	27202–27387	61 aa
ORF7a	27394–27759	121 aa
ORF8	27894–28259	121 aa
N	28274–29533	419 aa
ORF10	29558–29674	38 aa

Figure 2: The reference genome (Wuhan-Hu-1) contains 4 structural genes: S, E, M, N and 7 non-structural genes: ORF1a, ORF1b, ORF3a, ORF6, ORF7a, ORF8, ORF10[9].

It is well studied that the 69-70 del mutation in S gene (6-nucleotide deletion, 2-amino-acid deletion) can result in the S gene target failure (nondetection of the S-gene target among samples that tested positive for both the N gene and ORF1ab gene targets). Previously, S gene target failure was predominantly observed in B.1.1.7 (Alpha) variant. Since its close ancestor B.1 and successors B.1.351 (Beta), P.1 (Gamma), B.1.617.2 (Delta) do not carry 69-70 del, in the UK, the presence of S gene target failure in TaqPath PCR assays has been used as a proxy to track B.1.1.7 cases since 2020. As time evolves, B.1.1.529 (Omicron) became the dominant circulating strain from December 2021. The B.1.1.529 can be further categorized into some sublineages, including three variants of concern up to March 2022: BA.1, BA.2, BA.3. While BA.1 and BA.3 carry 69-70 deletion, BA.2 does not have this mutation. Thus, S gene target failure is not reliable in discerning Omicron from previous strains, however it can be used to discern BA.2 from other Omicron sublineages[8].

This report aims to develop a statistical classification method for differentiating patients infected by emerging mutant and dominating wildtype strain based on TaqPath PCR data. The classification method can be cautiously applied to follow-up sampled data by updating wildtype strain and mutant strainas as SARS-CoV-2 evolves, provided that the two strains demonstrate a significant difference in their binding capacity to some probe.

2 Data Description

The whole data set contains 1000 observations from TaqPath PCR assays sampled between August 2021 and April 2022 in the UK. Each observation includes specimen date, SGTF (encoding 1 for S gene target failure, 0 for S gene target positive), cycle threshold value for S gene, ORF1ab gene, N gene and MS2 gene. The Ct value for MS2 gene acts as an internal control to test the validity of PCR assays and reduce false negative rate: If MS2 gene is tested negative or the Ct value of MS2 exceeds 40, the corresponding diagnostic result is considered to be invalid or inconclusive data[3].

Among the 1000 observations, there are no missing Ct values for ORF1ab gene, 3 missing Ct values for N gene and 436 S gene target failures. There are also 35 inconsistent records for SGTF, which means the those Ct values are 0 but SGTF are encoded 0 (SGTF is defined as nondetection of the S gene target among samples that tested Ct$\leq$30 for both the N gene and ORF1ab gene targets[2][10]). After getting rid of these, 561 observations with complete Ct values for S, ORF1ab and N genes remain. This is the data set we used for training. In accordance with UKHSA surveillance report[11], most sequenced episodes from 1 August 2021 to 1 December 2021 were B.1.617.2 (Delta), others were almost exclusively AY.4.2 (mutated Delta). Since mutations of S gene and ORF1ab genes for the two variants have little difference (AY.4.2 only has additional T95I in S gene and A1711V in ORF1ab gene), we assume that there is little difference in their probe binding ability for the three target regions, thus take both of them as one strain-B.1.617.2/AY.4.2 (Delta/Delta plus). The

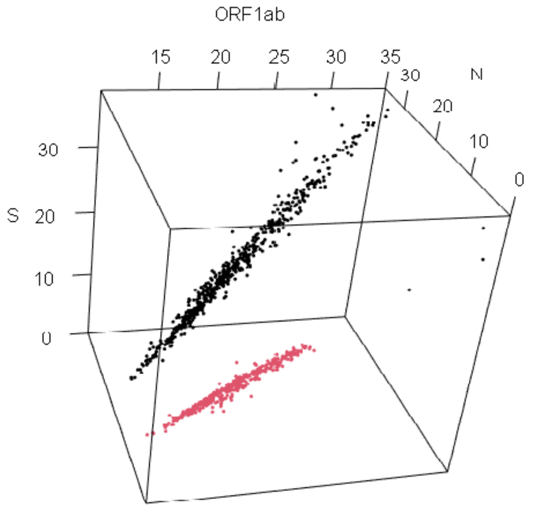


Figure 3: 3D visualisation of all consistent record. Red points denote record with S gene target failure, black points denotes record with complete Ct values.

B.1.1.529 (Omicron) was first detected in December 2021 in the UK, which carried an unprecedentedly large number of mutations in the S gene, including the 69-70 deletion. This suggests B.1.1.529 can cause S gene target failure. With rising concern for the Omicron variant, B.1.1.529 was further categorised into some sub-lineages: BA.1, BA.2. Until January 2022, the BA.1 strain has been the dominant strain and the proportion of BA.2 cases has been very low . In February 2022, BA.2 demonstrated a substantial growth advantage (75% higher than other circulating Omicron lineages in England). By the end of March, BA.2 was estimated to account for approximately 93.7% of cases in England.

The complete data set is divided into 2 data sets based on specimen date. The first data set consists of 480 observations from 1 August 2021 to 15 January 2022. The dominant strain shifted from B.1.617.2/AY.4.2 (Delta) to B.1.1.529 (Omicron) during this period of time. Since B.1.1.529 carries much more mutations than Delta in S gene including 69-70 deletion, we expect to see an increased proportion of SGTF or a two-component mixture in Ct value for S gene in this data set. The cluster with small S residuals is B.1.617.2/AY.4.2, the cluster with larger S residuals is B.1.1.529. The second data set consists of 81 observations from 16 January 2022 to 2 April 2022. This corresponds the time when BA.2 replaced the dominance of BA.1. As BA.2 does not have 69-70 deletion, the second data set is expected to present a decreasing trend for SGTF or a two-component mixture in Ct value for S gene - a cluster with relatively small Ct values for S gene for BA.2 and large Ct values for S gene for BA.1.

Table 1: Mutations in probe binding genes of primary Delta and Omicron variants from 1 August 2021 to 2 April 2022. AY.4.2 has a similar combination of mutation as B.1.617.2. Omicron sublineages carry far more mutations in S gene than Delta sublineages. BA.1 has 69-70 del, but Delta and BA.2 do not[7].

PANGO	WHO	S	ORF1ab	N
B.1.617.2	Delta	T19R,G142D,E156G,△157-158, L452R,T478K,D614G,P681R,D950N	A488S,P1228L, P1469S,V167L, T492I,T77A, P323L,G671S, P77L,A394V	D63G, R203M, G215C, D377Y
AY.4.2	Delta	T19R,T95I,G142D,E156G,△157-158, L452R,T478K,D614G,P681R,D950N	A488S,P1228L, P1469S,A1711V, V167L,T492I, T77A,P323L, G671S,P77L,A394V	D63G, R203M, G215C, D377Y
BA.1	Omicron	A67V,△69-70,T95I,G142D, △143-145,N211I,△212,R214ins, G339D,S371L,S373P,S375F,K417N, N440K,G446S,S477N,T478K,E484A, Q493R,G496S,Q498R,N501Y,Y505H, T547K,D614G,H655Y,N679K,P681H, N764K,D796Y,N856K,Q954H,N969K, L981F	K38R,S1265I, △1266,A1892T, T492I,P132H, △105-107,I189V, P323L,I42V	P13L, △31-33, R203K, G204R
BA.2	Omicron	T19I,L24S,△25-27,G142D,V213G, G339D,S371F,S373P,S375F,T376A, D405N,R408S,K417N,N440K,S477N, T478K,E484A,Q493R,Q498R,N501Y, Y505H,D614G,H655Y,N679K,P681H, N764K,D796Y,Q954H,	S135R,T24I, G489S,L264F, T327I,L438F, T492I,P132H, △106-108, P323L,R392C, I42V,T112I	P13L, △31-33, R203K, G204R, S413R

3 Methods

3.1 Observations with SGTF

For the 401 observations with S gene target failure, as B.1.1.529/BA.1 is the only lineage that carries 69-70 deletion mutation between August 2021 and March 2022, all these observations call be called by B.1.1.529/BA.1 with a high confidence. Though there were some mutated alpha still under investigation, their prevalence were too minor to take into account.

3.2 Observations with complete Ct values

For the 561 observations with complete Ct values for the three genes, we separate data according to the specimen date: The first data set accounts for 'Delta-Omicron shift'; The second data set accounts for 'BA.1-BA.2 shift'. The rationale for processing two data sets is the same: distinguish two components from a mixture and draw a decision boundary in the data. For clarity and brevity, we call the early dominating strain as wildtype and call the emerging strain as mutant[6].

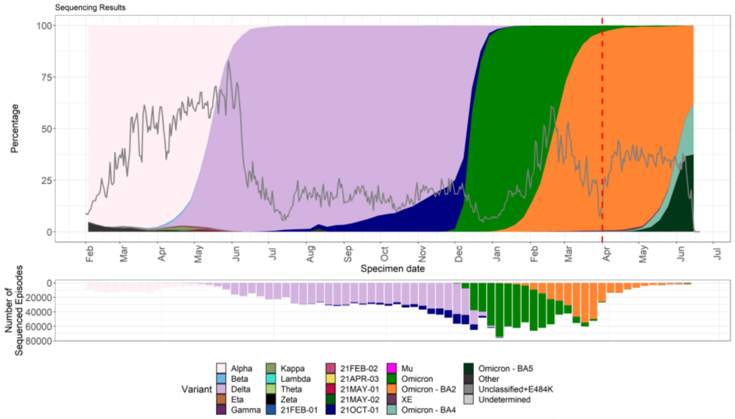


Figure 4: Percentage of sequenced episodes for different strains. Here 21OCT-01 is AY.4.2 (Delta plus), a sub-lineage of B.1.617.2 (Delta)[11].

However, there are two main challenges for this classification task: 1. How to overcome person-to-person variability in Ct values; 2. How to draw a good decision boundary.

3.2.1 Overcoming Person-to-Person Variability

For the first challenge, the raw data itself is hardly reliable to be used for the classification task. The viral load of a patient changes approximately as a Gamma distribution from two weeks before to 6 weeks after symptom onset, as illustrated in figure 6. As PCR assays can be conducted in any stage of illness, this person-to-person variability complicates the problem. It is hard to discern whether the relatively high Ct value is due to the strain with decreased binding probe ability or due to a illness stage with low viral load. To minimize this person-to-person variability, we baseline one gene's Ct using other one ore two genes' Ct as predictors. By taking advantage of the high correlation between each pair of three target Ct values, we examine the possibility of using linear regression to realize the person-to-person variability reduction. In this method, the gene whose Ct is balanced by other gene's Ct is called the surveillance gene, the gene that is used to baseline other gene's Ct is called the conditioning gene. The surveillance gene is expected to have a big difference in probe binding ability or even target failure in PCR assays between wildtype and mutant. The conditioning gene is expected to be relatively conservative in mutation between wildtype and mutant. Based on these criteria, we choose S gene as the surveillance gene and try ORF1ab gene or N gene or both of them as conditioning genes.

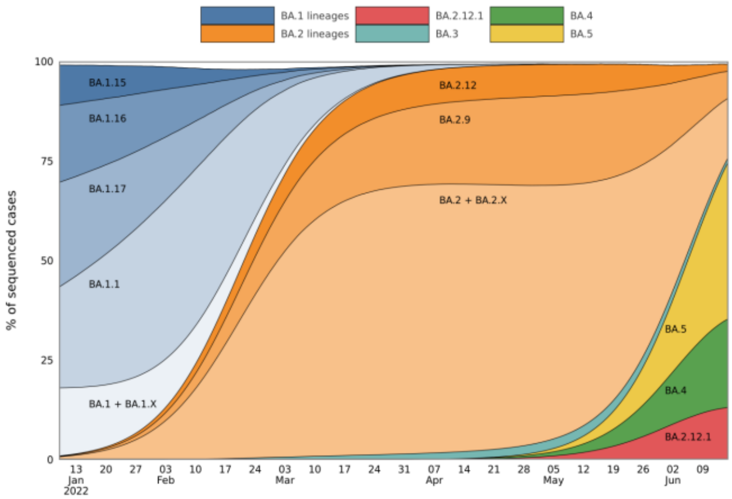


Figure 5: Prevalence of different strains after mid January 2022 built on hierarchical model[11].

3.2.2 Drawing a Decision Boundary

After overcoming person-to-person variability using linear regression on some conditioning gene's Ct value, we expect to observe some pattern for a two-component mixture, something like 2 bumps in the histogram or density plot of residuals. For the first data set corresponding 'Delta-Omicron shift', the bump with low S-residuals represents B.1.617.2/AY.4.2 and the bump with large S-residuals represents B.1.1.529. For the second data set corresponding 'BA.1-BA.2 shift', the bump with low S-residuals represents BA.2 and the bump with large S-residuals represents BA.1. Identifying which strain each observation belongs is not a trivial task, as there is an overlap in residuals between the two strains and we have to draw a decision boundary such that the error rate is minimized. A common approach for this unsupervised problem is Gaussian mixture model: Assume the distribution of residuals $f(x)$ is a mixture of two normal distributions $f_1(x)$ and $f_2(x)$ with corresponding mean μ_1, μ_2 and variance σ_1^2, σ_2^2

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) \tag{1}$$

where π_1 and π_2 are mixing weight of two components, $\pi_1 + \pi_2 = 1$.

We fit the residual data to the model and estimate parameters $\theta = (\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2)$ using the EM algorithm. The EM algorithm views the mixture problem as a missing data problem where the latent variables (group allocations) $\mathbf{k} = (k_1, \ldots, k_n)^T$ for each observations $x_1, \ldots, x_n$ are missing. $k_i = 0$ indicates the i-th observation belongs to the first component, $k_i = 1$ indicates the i-th observation belongs to the second component. Once these latent variables are

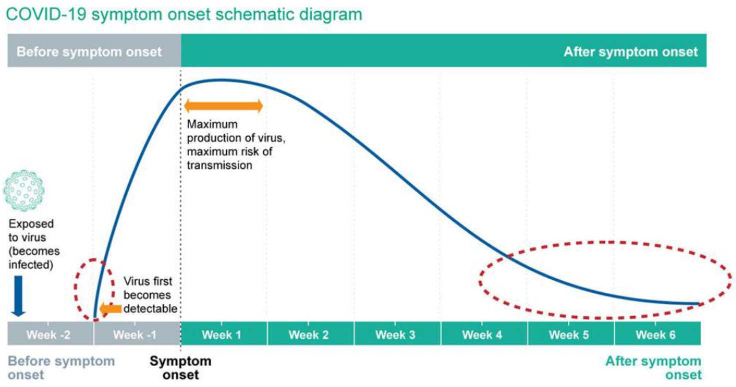

COVID-19 symptom onset schematic diagram
Before symptom onset
After symptom onset
Maximum production of virus, maximum risk of transmission
Exposed to virus (becomes infected)
Virus first becomes detectable
Week -2
Week -1
Week 1
Week 2
Week 3
Week 4
Week 5
Week 6
Before symptom onset
Symptom onset
After symptom onset

Figure 6: The viral load change in patients as the stage of illness evolves[4].

known, the complete data log-likelihood can be computed by

$$logL(\theta|\mathbf{x},\mathbf{k}) = \sum_{i=1}^{n} log(\pi_{k_i} f_{k_i}(x_i)) \tag{2}$$

The EM algorithm starts by a guess of true parameters θ and iterates between:

- Expectation step: using current estimate of θ to estimate the distribution of latent variables $\mathbf{z}$.
$$z_{ik}^{(b+1)} = \frac{\pi_k f_k(\mathbf{x_i})}{f(\mathbf{x_i})} \tag{3}$$
where $\theta^{(\mathbf{b})}$ is required in this step, $k = 0, 1$, $b \in N$.
Construct the expected complete data log-likelihood function
$$Q^{(b+1)}(\theta|\mathbf{x}) = \sum_{i=1}^{n}\sum_{k=1}^{2} z_{ik}^{(b+1)} log(\pi_k f_k(x_i)) \tag{4}$$

- Maximisation step: using current distribution of latent variables $\mathbf{z}$ to estimate θ that maximises the expected complete data log-likelihood.
$$\theta^{(b+1)} = argmax\ Q^{(b+1)}(\theta|\mathbf{x}) \tag{5}$$

The algorithm ends when $\theta^{(b)}$ converges.

Then we are able to estimate the probability of an observation x belonging to wildtype and mutant via Bayes' theorem

$$P(x\ belongs\ to\ group\ 1|x) = \frac{\pi_1 f_1(x)}{f(x)} \tag{6}$$

$$P(x\ belongs\ to\ group\ 2|x) = \frac{\pi_2 f_2(x)}{f(x)} \tag{7}$$

Any S-residual x with $P(x\ belongs\ to\ group\ 1|x) > P(x\ belongs\ to\ group\ 2|x)$ is assigned to group 1, otherwise is assigned to group 2.

4 Results

4.1 Data 1: Delta-Omicron shift

To minimize the person-to-person variability, we tried to condition S gene Ct value on other genes' Ct values and eventually proposed 3 linear regression models:

- 1. $S = \beta_0 + \beta_1 N$
- 2. $S = \beta_0 + \beta_1 ORF1ab$
- 3. $S = \beta_0 + \beta_1 N + \beta_2 ORF1ab$

It is important to note that our analysis is only based on residuals of linear regression, so regression coefficients are not of our interest. The only thing we care about is whether person-to-person variability can be reduced enough to unmask the two clusters in S gene Ct values for different strains. In this sense, the R^2 statistic and residual density plots should be scrutinized with care.

The outcome of three linear regression models for the data before 15 January 2022, as shown in Figure 7, suggests that N gene's Ct value is difficult to discern two components of the mixture, so it is not enough to baseline person-to-person variability in the data. The other two models demonstrate two bumps well in the plot and the model with only ORF1ab gene's Ct as predictor works best. Considering N gene's Ct and ORF1ab gene's Ct are highly collinear with correlation coefficient 0.9896, the N gene's Ct almost provide the same amount of information about S gene's Ct as ORF1ab's Ct. As model 3 has no significant improvement in identifying the two components of residual density plot, we favor model 2 for reducing person-to-person variability in data set 1.

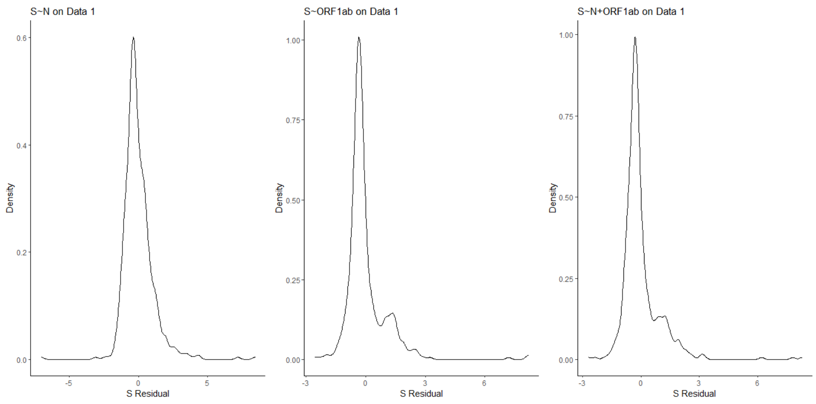


Figure 7: The S residual density plot for three linear regression models.

The analysis for residuals of the second model involves estimating parameters for two Gaussian components $\pi_1, \mu_1, \mu_2, \sigma_1, \sigma_2$ and calculating the probability of belonging to each strain. These can be implemented by mixtools package in R, a tool package applying EM algorithm to Gaussian mixture models.

The result of this analysis gives a decision boundary 0.07 and the two mixing components are shown in Figure 8. Any observation with S-residual smaller than this value is called B.1.617.2/AY.4.2 (Delta), otherwise is called B.1.1.529 (Omicron).

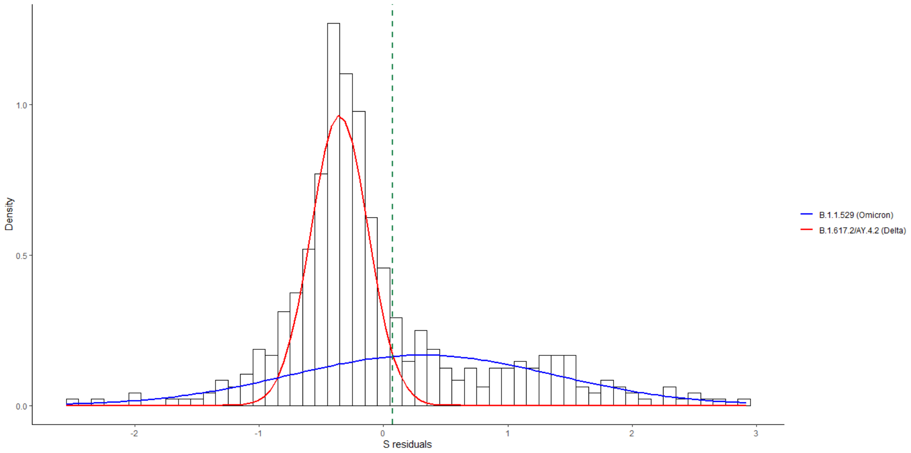


Figure 8: The two components of Gaussian mixture model, B.1.617.2/AY.4.2 in red and B.1.1.529 in blue. The green dashed line denotes the decision boundary 0.07.

4.2 Data 2: BA.1-BA.2 shift

For the second data set, similarly, we tried three different linear regression models to baseline person-to-person variability. The model with only ORF1ab gene's Ct value as predictor works best for presenting two bumps and the model with multiple predictors has no significant improvement in presenting two components. So we once again favor model 2 for reducing person-to-person variability of the second data set.

The S-residual obtained by conditioning on ORF1ab's Ct value can be taken as two normal densities, as shown in Figure 10. And the calculated decision boundary is 0.19. Any observation with S-residual smaller than this value is called BA.2, otherwise is called BA.1.

5 Discussion

5.1 How should we interpret total variability?

This classification problem is essentially a task of interpreting variability of the data. The total variability of the S gene's Ct value can be divided into two

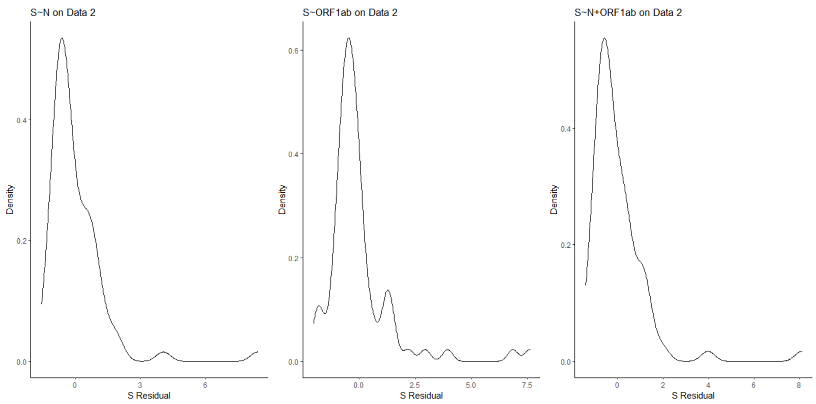


Figure 9: The S residual density plot for three linear regression models for the second data set.

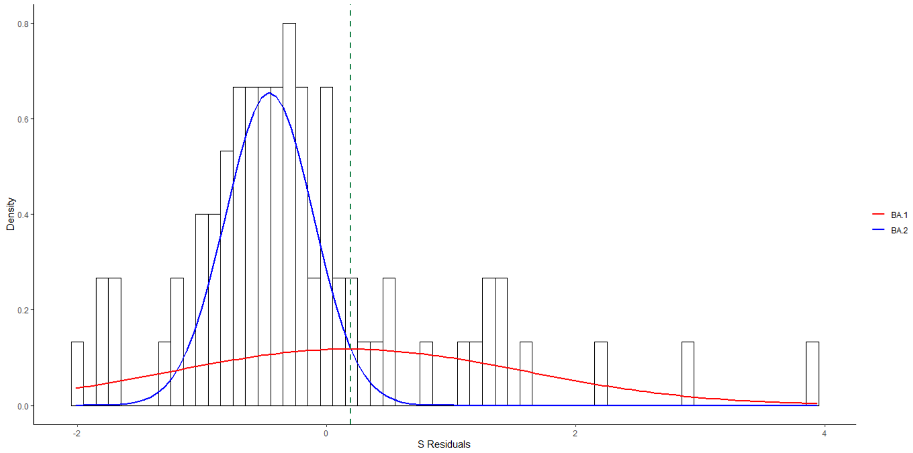


Figure 10: The two components of Gaussian mixture model, BA.1 in red and BA.2 in blue. The green dashed line denotes decision boundary 0.19.

parts: 1) Variability due to samples are collected on patients with different symptomatic phase. i.e. the person-to-person variability. This is the variability that we want to get rid of, the noise. 2) Variability due to the different probe binding ability of two strains. This is the signal, our objective for us to find patterns. In the raw data, the second type of variability is masked by the first type, so we tried linear regression with three combinations of N gene and ORF1ab gene's Ct values as control to reveal the second type of variability.

5.2 How much variability should be taken as noise?

Now the question is: how much variability in the raw data should be explained as the person-to-person variability? Or equivalently, how hard should we denoise in order to get a clear signal? If the proportion we conditioned is less than the underlying true person-to-person variability, then the second type of variability is still masked by the first type, the data is still noisy. In this case, as illustrated in the left panels of Figure 7 and 9, no patterns of the two strains can be observed. If the person-to-person variability is over-explained, then similar to overfitting, our conditioning may work too hard that even eliminate variability more than the noise itself. In this case, some signal is taken as the noise and erased. Therefore, we may not be able to observe the two components in the signal as well.

How can we obtain a model that just extracts an adequate amount of person-to-person variability from the total variability? This model selection problem is different from the usual model selection in regression analysis, as here we reverse the usual problem and consider residuals of regression as signals. With no widely accepted methods at hand, the outcome-deterministic model selection (tuning) can be a good option: Consider all possible regression models as control for person-to-person variability, and then choose the model whose residuals present the most obvious pattern of two bumps. For both data 1 and data 2 in the PCR data, we considered linear regressions with different combinations of N and ORF1ab genes' Ct as predictors. As a result, the model only using ORF1ab gene's Ct gave the most obvious pattern of two bumps for both data sets.

Since all regression methods can be used to control noise, is it necessary to consider more advanced methods other than linear regression? Though it is possible to get a better result using other regression methods to overcome the noise, we cautiously doubt the necessity of trying this. The 3d scatter plot of the data demonstrates an approximate linear relationship between the three target Ct values, so any working regression scheme used for conditioning should be close to the linear regression. As the linear regression method provides low variance and high interpretability, more complex methods are likely to separate a slightly stronger signal, but with a cost of increasing variance and losing significant interpretability.

5.3 How to separate within-group variability and between-group variability?

After using other Ct values to overcome person-to-person variability, we have S residuals, a mixture of two components. This is turned to a typical clustering problem. Now, the total variability in the S residuals can be further divided into within-group variability and between-group variability. The aim is to separate them such that the between-group variability is big enough to present two different clusters.

$$\underbrace{Var(S)}_{total} = \underbrace{Var_k(E_S(S|k))}_{between-group} + \underbrace{E_k(Var_S(S|k))}_{within-group} \tag{8}$$

where S for S-residual and $k = 0, 1$.

There are broadly two types of clustering methods: Algorithmic clustering methods and probabilistic model-based methods. The probabilistic model-based methods are based on a probabilistic model, while algorithmic methods are not (explicitly).

Specifically, the Gaussian mixture model (GMM) we used is a probabilistic model-based methods. We first parameterize both components as normal distributions and then estimate parameters. We finally calculate the probability of each S-residual belonging to each cluster. Observation are assigned according to their corresponding probabilities. Alternatively, we can also try K-means method, an algorithmic method. The k-means algorithm initializes by randomly allocating observations $x_1, x_2, \ldots, x_n$ to the two clusters with equal probability. Then the algorithm iterates between

- Estimate the group means by

  $$\mu_k = \frac{1}{n_k} \sum_{i \in G_k} x_i \tag{9}$$

  where n_k is the number of observations in group k, $C(x_i) = 0, 1$ is the allocation of x_i, $G_k = \{i | C(i) = k\}$ is the indices of observation belonging to group k.

- Observations are reassigned to the group k with the nearest mean μ_k

  $$C(x_i) = argmin|x_i - \mu_k| \tag{10}$$

The implementation of k-means on the two data sets of our PCR data gives two non-overlapping clusters for both data sets, see Figure 11. For the first data set, observations with S-residuals smaller than 0.46 is called B.1.617.2/AY.4.2 (Delta), otherwise is called B.1.1.529 (Omicron). For the second data set, observations with S-residuals smaller than 0.13 is called BA.2, otherwise is called BA.1.

The decision boundary given by k-means method is slightly different from the previous one from GMM. Theoretically, k-means and GMM can give similar result for univariate data when the mixing weight for two groups are 0.5. i.e. k-means is viewed as a restricted GMM with equal weight for two groups. So the difference in decision boundary between k-means and GMM can be reduced as the weights of two strains become evenly matched. How should we compare these two methods in the PCR example? With no additional validation set, there is no significant evidence for favoring one method against the other. In terms of computational cost, K-means is faster and efficient than GMM. On

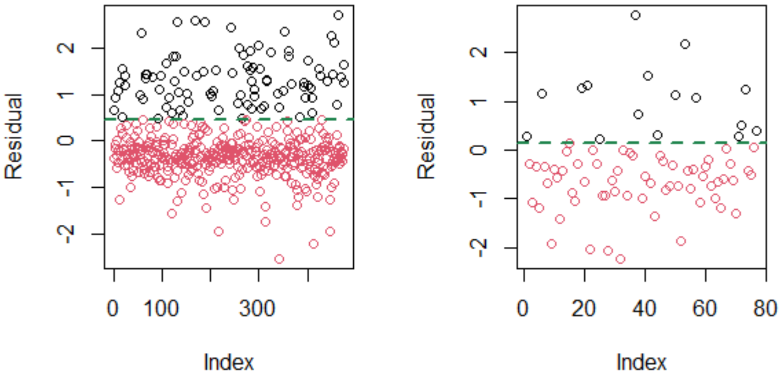


Figure 11: Left panel: two non-overlapping clusters, B.1.617.2/AY.4.2 in red, B.1.1.529 in black, the green dashed line denotes the decision boundary 0.46. Right panel: two non-overlapping clusters, BA.2 in red, BA.1 in black, the green dashed line denotes the decision boundary 0.13.

the other hand, K-mean has no explicit probabilistic interpretation and it is a blackbox, while GMM provides full probabilistic model for clustering.

References

[1] Ana Banko, Gordana Petrovic, Danijela Miljanovic, Ana Loncar, Marija Vukcevic, Dragana Despot, and Andja Cirkovic. Comparison and sensitivity evaluation of three different commercial real-time quantitative pcr kits for sars-cov-2 detection. *Viruses*, 13(7):1321, 2021.

[2] Kevin A Brown, Jonathan Gubbay, Jessica Hopkins, Samir Patel, Sarah A Buchan, Nick Daneman, and Lee William Goneau. S-gene target failure as a marker of variant b. 1.1. 7 among sars-cov-2 isolates in the greater toronto area, december 2020 to march 2021. *Jama*, 325(20):2115–2116, 2021.

[3] Jens Dreier, Melanie Stormer, and Knut Kleesiek. Use of bacteriophage ms2 as an internal control in viral reverse transcription-pcr assays. *Journal of clinical microbiology*, 43(9):4551–4557, 2005.

[4] PH England. Understanding cycle threshold (ct) in sars-cov-2 rt-pcr. a guide for health protection teams. *Public Health*

England, Available at: https://assets. publishing. service. gov. uk/government/uploads/system/uploads/attachment_data/file/926410/Understanding_Cycle_Threshold__C CoV-2_RT-PCR_. pdf, 2020.

[5] National Library for Medicine. Severe acute respiratory syndrome coronavirus 2 isolate wuhan-hu-1, complete genome. https://www.ncbi.nlm.nih.gov/nuccore/1798174254, accessed in Aug.2022.

[6] Yang Han, Yujia Sun, Jason C Hsu, Thomas House, Nick Gent, and Ian Hall. Statistical design and analysis of diagnostic tests for mutating viruses. *medRxiv*, 2021.

[7] Stanford HIVDB. Coronavirus antiviral and resistance database accessed on 8 sep. 2022. https://covdb.stanford.edu/variants/omicron/.

[8] World Health Organization et al. Methods for the detection and characterisation of sars-cov-2 variants: second update, 21 june 2022. Technical report, World Health Organization. Regional Office for Europe, 2022.

[9] Vishal Rana, Eli Chien, Jianhao Peng, and Olgica Milenkovic. How fast does the sars-cov-2 virus really mutate in heterogeneous populations? *mediRxiv. https://doi.org/10.1101/2020.04.23.20076075*, 2020.

[10] UKHSA. Omicron voc-21nov-01 (b.1.1.529) update on cases, s gene target failure and risk assessment. *SARS-CoV-2 variants of comcern and variants under investigation in England, Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/103964* 2021.

[11] UKHSA. Sars-cov-2 variants of concern and variants under investigation in england technical briefing 44. *UKHSA technical briefings, Available at: https://www.gov.uk/government/publications/investigation-of-sars-cov-2-variants-technical-briefings*, 2022.