보건학 석사학위논문

Analyzing Bayesian phylogeography of SARS-CoV-2 in South Korea and factors affecting their transmission

한국 SARS-CoV-2 대유행의 베이지안 계통지리학적 분석과 이에 영향을 미치는 요인

2023년 2월

서울대학교 보건대학원
보건학과 보건학전공

이 상 이

Analyzing Bayesian phylogeography of SARS-CoV-2 in South Korea and factors affecting their transmission

지도교수 조 성 일

이 논문을 보건학 석사학위논문으로 제출함

2022년 11월

서울대학교 보건대학원
보건학과 보건학전공

이 상 이

이상이의 보건학 석사학위논문을 인준함

2022년 12월

Chair	원 성 호	(Seal)
Vice Chair	권 정 훈	(Seal)
Examiner	조 성 일	(Seal)

Abstract

Following the global emergence of the Alpha variant of concern (VOC) of SARS-CoV-2 in 2019, another wave emerged due to the SARS-CoV-2 Delta variant in 2021. The AY.69 lineage, a Delta VOC, was particularly prevalent in Korea between May 2021 and January 2022, despite the synchronized implementation of vaccine programs and non-pharmaceutical interventions (NPIs), such as social distancing. Here, we used phylogeographic analysis supplemented by a generalized linear model (GLM) to determine the influence of human movement and vaccination on viral transmission. The results suggested that transmission began predominantly in the metropolitan areas of South Korea, and that total human mobility tracked by GPS using mobile phones and estimated by credit card consumption had a positively affected the occurrence of introduction events. This phylodynamic findings also supported the notion that non-vaccinated persons dominantly transmitted the virus during the study period, despite of vaccination programs that started three months before the propagation of AY.69. Therefore, our results suggest that co-implementing both NPIs and an early vaccination program would effectively reduce viral spread.

Keyword : SARS-CoV-2, phylogeography, genomic epidemiology, human movement, Phylodynamics, Generalized linear model
Student Number : 2021-27881

Table of Contents

Chapter 1. Introduction 1

1.1. Study background 1

1.2. Purposes of research 2

Chapter 2. Methods 6

2.1. Sequence data and subsampling 6

2.2. Bayesian Phylogeographic Analysis 8

2.3. Mobility Data 9

2.4. Generalized linear model of discrete trait diffusion 10

Chapter 3. Results 11

3.1. The AY.69 variant was predominant in mid-2021 in South Korea 11

3.2. AY.69 variant mostly spread from Seoul and Gyeong-gi 16

3.3. AY.69 variant mostly spread from Non-vaccinated group 20

3.4. GLM analysis of mobility data and phylogeography 23

Chapter 4. Discussion 30

Chapter 5. Conclusion 35

Bibliography 36

국문초록 43

Appendix 45

List of Tables

Table 1. Calculated cross-regional transmission values of the 680-sample dataset..... 1 2

Table 2. Calculated transmission values among vaccine status group of the 680-sample dataset..... 21

Table 3. Descriptive statistics of each predictor of GLM..... 25

Table 4. Inclusion support statistics for each dataset..... 26

List of Figures

Figure 1. The prevalence of each SARS-CoV-2 virus variant in South Korea 12

Figure 2. Comparison on sequence samples and incidence counts of each region and immune groups 14

Figure 3. The effective population size of Ay.69 during the study period and its inter-region introduction events with a phylogenetic tree 18

Figure 4. The introduction events between immune group with a phylogenetic tree and chord diagram 21

Figure 5. The results of generalized linear model of inter-region viral transmission and its predictors 27

Figure 6. The results of generalized linear model of inter-regional viral transmission and its predictors considering sample sizes 29

Chapter 1. Introduction

1.1. Study Background

As human history has gone through several pandemics, the interaction between pathogen and human population has been identified [1, 2]. This can be well utilized when applying to infectious disease epidemiology in the emergence of new disease and this field of study of identifying disease transmission pattern is called genomic epidemiology [1]. Furthermore, for better real-time pandemic surveillance and epidemic control and preparedness, phylodynamics - the combination of epidemiology with immunodynamics and phylogenetics - was coined by Grenfell in 2004 [1]. Whereas phylogenetics only focuses on estimating a tree with the most minimum evolutionary steps or the maximum likelihood tree, phylodynamics emphasizes population dynamic factors such as reproductive numbers, generation time and epidemic growth rate [3]. The coalescent theory is central concept in phylodynamics which trace back coalescent events by mutational drift and predicts effective viral population size(Ne) change [3].

However, the detailed link between spatiotemporal link of disease outbreak and intervention scenario such as non-pharmaceutical intervention (NPI) or superspreading were not

explored although they are very important factors during pandemic [4]. This problem can relatively easily be solved by phylogeographic analysis which involves sampling locations as sequence traits and builds a phylogenetic tree [4]. In addition to phylogeography, generalized linear model(GLM) framework has been applied to figure out non-genomic predictors of pandemic pattern [4].

Besides from population dynamics of infectious disease, individual contact tracing would be more informative to track all the transmission chain in the real world since the superspreading event and related transmitted people can be found out by this method [5]. It is important to import individual-level data since population immunity and new viral introduction accumulate differently within each network [6]. Therefore, demographic or measured information needs to be included in the phylodynamic analysis, and here, we adopted individual vaccine status data and sampling locations are included as traits.

1.2. Purpose of Research

The outbreak of COVID-19 caused an unprecedented pandemic, characterized by extremely prolonged periods of time and an excess of deaths globally. Because SARS-CoV-2 virus rapidly mutates and spreads widely from coast to coast and has caused multiple large outbreaks, there are limitations with the traditional epidemiological tracking of contacts by surveys [7]. Because real-time tracing is

very important for tracking the fast-changing COVID-19 pandemic, which has had clustered outbreaks around the world, genomic epidemiology may be a good complementary epidemiological tool [8]. That is why light has been shed on phylogenetic analyses recently, with the development of complex analyses for phylodynamics with spatiotemporal analysis also called a phylogeography [9].

In South Korea, the largest epidemic was driven by the sublineage SARS-CoV-2 Delta variant of concern (VOC) designated by WHO, AY.69, which was present in the large cluster of cases that occurred between May 2021 and January of 2022. The number of complete sequences for this variant worldwide is 11,296, with 11,234 found in South Korea and the 11,234 cases found in South Korea during that time, the sequence of AY.69 took up 51.85% of whole Delta variants found in South Korea [10].

During the pandemic, South Korea implemented a number of policies, including vaccination programs and social distancing [11]. The national immunization program, which aimed to immunize 70% of the population by November 2021, started with priority groups from February 26, 2021, and later expanded to younger age groups [12]. This age-based-prioritization vaccination strategy, with a high rate of vaccine distribution, promoted vaccination; with the high rate of compliance, the vaccination program resulted in a steep increase in the immunized population [13]. Indeed, the proportion of fully vaccinated individuals exceeded 70% on October 23, 2021, and exceeded 80% on October 29, 2021 [14].

Vaccination programs reduce deaths [15] and South Korea reduced the case fatality rate over the study period, which subsequently plateaued at 0.8. However, the rate of infection increased, despite vaccination coverage; unvaccinated persons predominated among cases in 2021, although the incidence of breakthrough infections gradually increased until the end of 2021 [16]. Transmission dynamics among groups of different immune status are unknown, unless contact histories are interpreted in conjunction with ongoing epidemiologic surveillance [17].

The human mobility pattern has a considerably large impact on viral spread. The importance of human factors in pandemics had been emphasized in previous phylogeographical studies, such as that of air traffic density on human immunodeficiency virus spread [18] and of freight transportation on avian influenza virus spread [19]. Those studies concluded that human mobility measured by public transportation scales partially explains spatial spread of virus more than the other factors influencing epidemics, including environmental factors such as humidity and temperature, and human factors such as railway connectivity, population, and immigration. Although viruses have different characteristics, the impact of human movement on COVID-19 warrants investigation [20].

It is important to consider all associations of government strategies with viral spread and measure their effects to assess policies and prevent future pandemics [21]. During the COVID-19 pandemic, social distancing strategies were implemented by regions

of South Korea based on local movement patterns, necessitating a high-resolution method of measuring movement, possibly using mobile telephones [22]. Individual mobility data can also be used to estimate movement in a region for phylodynamic analysis.

To enhance tracking of virus sources and transmission routes, genomic epidemiologists have used phylogeography to evaluate the geographic transmission histories of viruses [23]. By assigning distinct sites for each node, metadata, including sampling sites, are included [24]. By merging phylogeography with state-of-the-art statistical methods, phylodynamic analyses are becoming more useful [25].

Using this approach, we investigated the effects of immunization and social distancing policies on viral transmission [26]. In addition, epidemiological surveillance by genetic sequence acquisition provides a basis for state decisions to detect viral variation and source [27]. Moreover, extending phylodynamic analysis into parameterizing spatial movement rates as a generalized linear model (GLM) of potential predictors facilitates evaluation of factors linked to virus transmission and mutation, enabling evaluation of prevention policies during the phase of the pandemic dominated by Delta variants [28].

Chapter 2. Methods

2.1. Sequence data and subsampling

To estimate the effect of sample size, we included sample size as a predictor and identified its largest effect. Delta variant genetic sequences were downloaded on January 31, 2022, from the Global Initiative for Sharing All Influenza Data [29]. We used metadata provided by the Korea Disease Control and Prevention Agency to eliminate samples from inbound travelers and assigned each sequence its position in relation to data on the immune status of an infected individual. Among 10,232,901 AY.69 clade genomes from South Korea (approximately 1% of confirmed cases in the period under study) that had < 5% nucleotide uncertainty, we redesignated each genome to a global-standard Pango lineage nomenclature using Pangolin v. 1.11 (data published June 30, 2022) [30]. To avoid mixing of sequences that might be confused with AY.69, we retained sequences containing mutations associated with AY.69 (A4838G, G9431A, G16864A, and C27559T) designated by the Pango network [31]. The quality of filtered AY.69 sequences was assessed using Nextclade Quality Control, and only those of good quality (8,806 sequences) were retained [32].

To identify Korean transmission lineages, we constructed a maximum-likelihood phylogenetic tree with FastTree v. 2.1.11 using sequences from other countries. We downloaded the 1,348 Delta-

representative sequences (123 from Africa, 168 from Asia, 342 from Europe, 340 from North America, 185 from Oceania, and 190 from South America) selected from Nextstrain v.11 [33] as background sequences in the tree to prevent loss of important cladistic structures and continental distributions.

To reduce the number of sequences, we used TARDiS subsampling software [34]. TARDiS subsampling of the positional groups considered both the date of collection (ω_{td}) and the genetic relatedness (ω_{gd}), using user-defined weights set by default (ω_{td}:ω_{gd} = 0.5:0.5). Due to the large collections available, we repeated the steps with two subsampled datasets, yielding final sequence counts of 642 and 220, respectively. [27]

We chose Wuhan-Hu-1 as the reference sequence, to which every sequence was aligned and hand-trimmed to an equal length (29,409 bp) from the ORF1ab starting codon to the ORF10 stop codon, and all sequences were masked to the reference gene. Gaps shared by > 99% of the sequences were considered sequencing errors and removed manually to ensure that they were not known mutation sites in AY.69 variants, according to the pangolin reference [35]. Sequence purification was performed using Geneious Prime software v. 2021.2.2 [36].

The approximated trees were generated using FastTree v. 2.1.11 under the generalized time-reversible nucleotide substitution model, with gamma-distributed rates among sites (general time reversible + gamma [GTR+γ]) [37]. We chose only those trees in

which Korean sequences were most tightly grouped and shared the same ancestor, which had > 0.7 support values in the final dataset. Outlier sequences that had > 8.0 × 10−4 variances in the tree were checked in TempEst v. 1.5.3 with a root−to−tip regression [38], resulting in the removal of a further 38 sequences. Finally, we analyzed 220 and 642 samples.

2.2. Bayesian Phylogeographic Analysis

We constructed the phylogenetic trees on timescales using BEAST v. 1.10.4 [39]. The GTR nucleotide substitution models were selected by the Bayesian average of the site−based phylogenetic models using BModeltest [40]. The user−specified tree model that we constructed previously and Bayesian skygrid tree priors were used, and an uncorrelated relaxation−driven clock model was used to assess changes in virus population size with a flexible approach. Markov chain Monte Carlo was performed over 100 million steps and parameters and trees were sampled at 10,000 steps. The parameters were analyzed using TRACER v. 1.7.1 with 10-20% burn−in [41]. We first constructed a tree without the GLM model and used it as an initial tree for GLM. Most parameters had effective sampling sizes of > 200. The resultant log files and trees were combined using LogCombiner v. 1.10.4 [42], resulting in 32,804 parameter states and posterior trees. Time−scaled max−clade−credibility trees were generated using the TreeAnnotator [42] function in BEAST and were

visualized using FigTree v. 1.4.3 [43].

Bayesian stochastic search variable choice procedures were applied to determine the most supported transitions among discrete states using the Bayes factor test, and transmissions were played along the timeline using SPREAD3 software v. 0.9.6 [44]. A transition was identified as significant at a Bayes factor of > 6 and posterior probability of > 0.5.

2.3. Mobility Data

We prepared the datasets as coalescent tree priors. Sample size was modeled as the log-transformed count of sequences. The two measures of movement were an aggregated dataset from BC financial services company, the data of people's card consumption on entertainment, and a travel dataset based on the origin-destination movements tracked using KT mobile phones [45]. Each measure was collected from May 2021 to January 2022 and they were not strongly correlated (Appendix 3). We note that these data might not represent the entire population.

2.4. Generalized linear model of discrete trait diffusion

Before investigating predictors of spatial transmission by GLM, which is computationally and time intensive, we constructed an annotated tree without GLM and used it to explore associations between candidate covariates in the web-based application Phylogeographic Covariate Analysis (PhyCovA) [46]. In this way, several most probable predictors other than sample size were selected.

Using the selected predictors, we performed Bayesian phylogeography in discrete space and the GLM-diffusion model using BEAST v.1.10.4, including sample region, date, mobility data, and sample size as parameters. The dependent variable of the GLM was the log-transformed transition rates among 17 discrete regions according to the continuous-time Markov chain. The independent variables were log-transformed mobility data and sample size. The Bayesian model estimates the phylogenetic history, ancestral movement, and contributions of covariates simultaneously [17].

Because the sampling bias represented by larger coefficients for sample sizes will influence the other predictors' impact, we also implemented a GLM with no covariates for sample sizes. Moreover, we determined effect sizes for the covariates and their probabilities of inclusion using the Spike-and-Slide procedure

Chapter 3. Results

3.1. The AY.69 variant was predominant in mid-2021 in South Korea

According to the genetic surveillance, the AY.69 variant was first detected in Korea in the year 2021-05-14, with approximately half detected in Seoul and Gyeong-gi. (Fig. 2) It became a dominant subgroup starting from mid-July 2021, with increasing numbers of isolates through September 2021, before being detected for the last time in 2022-01-22. The fourth peak of incidences in Korea occurred during July, which is consistent with AY.69 prevalence. During this time, the South Korean government announced its highest level of social distancing. However, AY.69 variation did not exhibit decreasing trends before October, when the number of cases increased. AY.69 was partly replaced with other delta variant, AY.122.5, after 2nd vaccine shot schedule began in 2021-10-14, but when the large surge occurred in December, 2021, AY.69 gently downturned and comprised less than 50% after higher level of social distancing policy implemented. (Fig. 1)

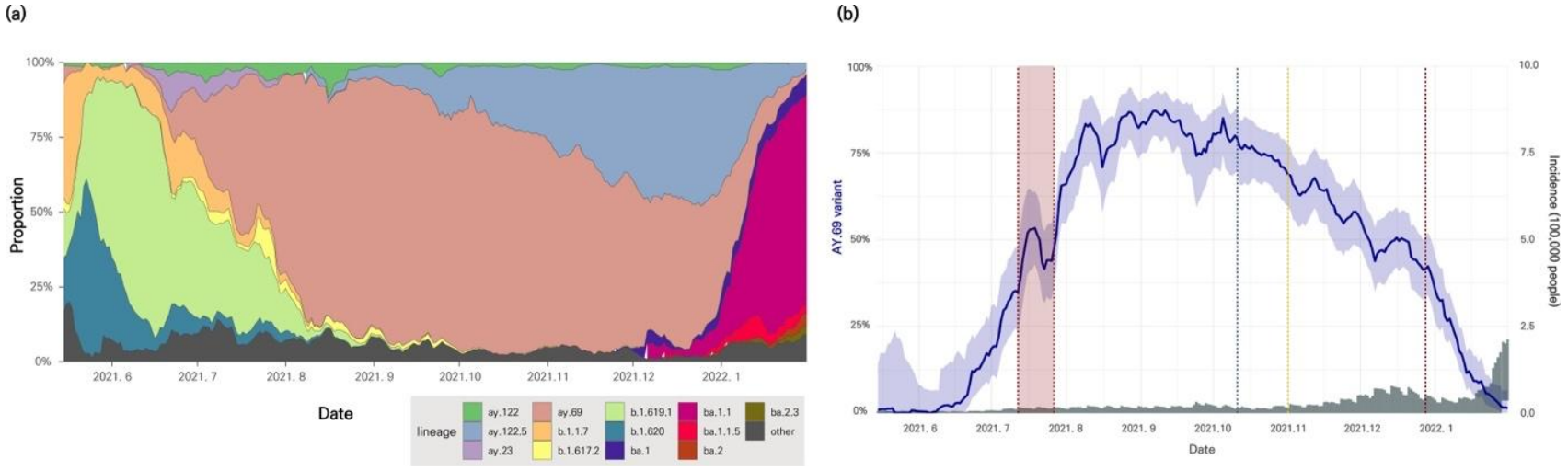


Figure 1. The prevalence of each SARS-CoV-2 virus variant in South Korea

(a) The Delta-variant pandemic in South Korea showing different Delta sub-lineages' prevalence in the overall study period. (b) The most dominant variant, AY.69, was predominant in mid- to late-2021 exceeding 75% in August. Grey shaded bars represent every COVID19 variants' incidence number/100,000 people. Red shade and dashed lines are the time when social distancing level raised (2021-07-17~2021-07-27, 2021-12-21) whereas yellow dashed line is the time point when social distancing level lowered (2021-11-01). Blue dashed line is the beginning of 2nd vaccination shot program (2021-10-14).

Because 74.77% of cases occurred in the South Korean metropolises(Seoul, Gyeong-gi and In-cheon), the sampling frequencies were concentrated in these regions (52.49%) during AY.69 prevalence periods. The geographical distribution of our sequence data was commensurate with outbreak proportion of each regions (Fig. 2).

Moreover, 64.80% of the samples were not vaccinated, 10.59% were partially vaccinated, and others were fully vaccinated. The vaccination program in South Korea started on February 26, with over 70% of the population being fully vaccinated by October 23 [47]. However, samples from partial or fully vaccinated people began to be collected in July 2021; once more than 70% of the population was fully vaccinated, the fully vaccinated (FV) group accounted for most of the genetic sequence samples. (Fig. 2)

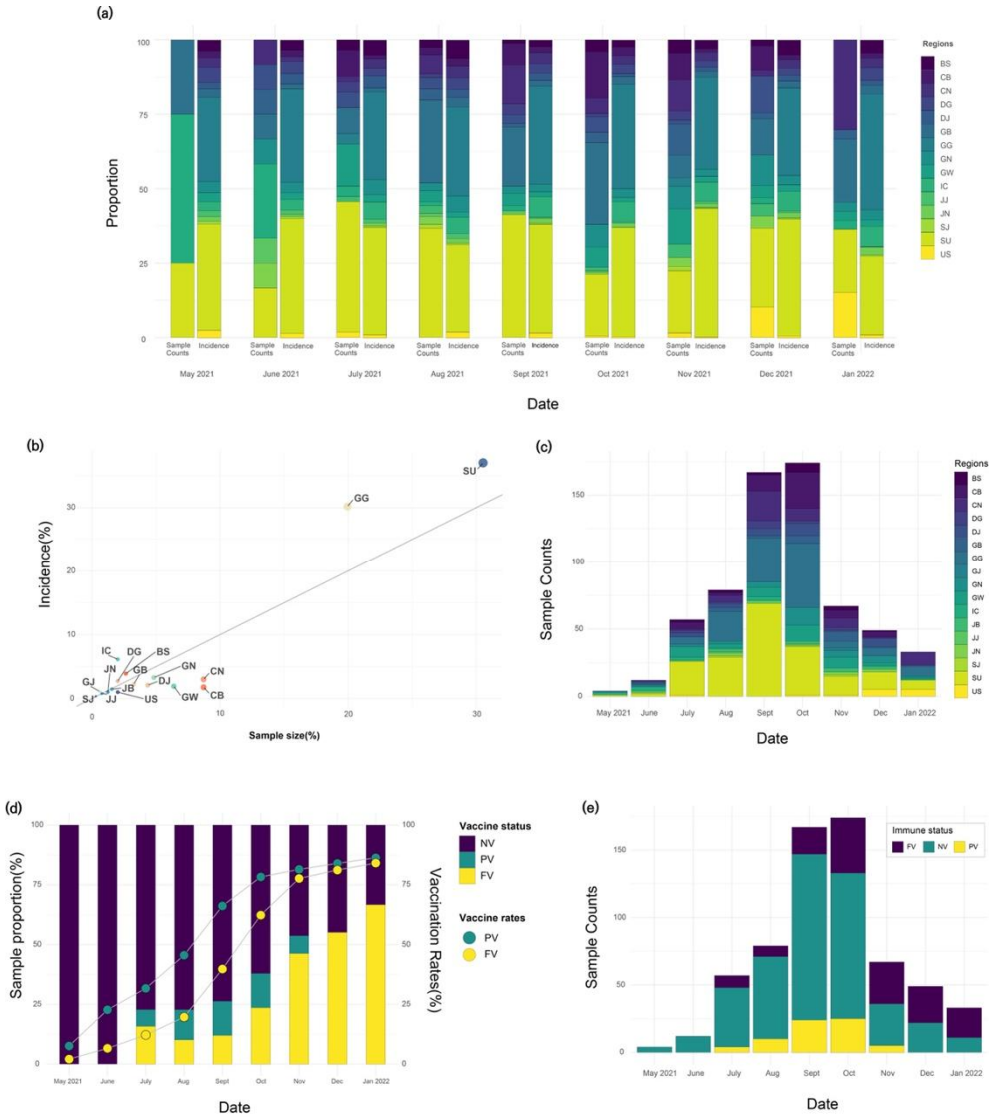


Figure 2. Comparison on sequence samples and incidence counts of each region and immune groups.

(a) Monthly sample counts and incidences by region from May 2021 to January 2022. Incheon accounted for most samples in May 2021, and Seoul and Gyeonggi) in later periods. Colors indicate regions as labelled on the right side of the figure. (b) Comparison of sample counts and incidence proportion in South Korea. Each region is labelled on the figure. (c) Sample counts for each region. Colors

indicates each region as labelled on the right side of the figure. (d) Proportions of immune groups (bar graph) with vaccination rates (line graph). Colors are labelled on the right side of the figure. (e) Sample counts for each immune status. Colors indicates each immune status as labelled on the right upper side of the figure.BS, Busan; CB, Chungbuk; CN, Chungnam; DG, Daegu; DJ, Daejeon; GB, Gyeongbuk; GN, Gyeongnam; GG, Gyeonggi; GW, Gangwon; IC, Incheon; JJ, Jeju; JN, Jeonnam; SJ, Sejong; SU, Seoul; US, Ulsan; NV, non-vaccinated group; PV, partially vaccinated group; FV, fully vaccinated group.

3.2. AY.69 variant mostly spread from Seoul and Gyeong-gi

On the other hand, it was established that the most recent common ancestor(tMRCA) for Korean AY.69 was April 22, 2021(95% height posterior density [HPD]: April 1 through May 10). This indicates that AY.69 likely occurred several weeks prior to initial detection. The estimated effective virus population size increased through mid-July, showing a plateau before the decline began in December 2021. The effective population size reached its first peak on 25 July, just prior to heightened social distancing in 27 July 2021 (Fig. 3a).

By ancestral reconstruction in Bayesian phylogenetic analyses, we identified spatiotemporal diffusion between regions in South Korea, which all the transitions between regions were measured by Markov jump counts. Based on posterior average ratios of each region's introduction, dispersal between metropolises led spatial transmission of the AY.69 variation in the 2021-05-08 timeframe, with virus mainly spreading out of Seoul during the entire period, showing 276 (95% height posterior density [HPD]: 222-322) times outfluxes out of the total 331 (95% height posterior density [HPD]: 235-438) outfluxes within the whole tree. The largest dissemination originated from Seoul then flew into Gyeong-gi(82; 95% HPD = 72-92) and spread from Seoul to Chung-nam (31; 95% HPD = 25-37) was the second largest spread. (Fig. 3b, 3c)

Table 1. Calculated cross-regional transmission values of the 680-sample dataset

	Transmitted to																	
From	BS	CB	CN	DG	DJ	GB	GG	GJ	GN	GW	IC	JB	JJ	JN	SJ	SU	US	Total
BS		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CB	0		1	0	0	1	1	0	1	2	0	0	0	0	0	1	0	7
CN	0	0		0	1	0	2	0	0	0	0	0	0	0	0	0	0	3
DG	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
DJ	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
GB	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
GG	2	4	6	1	3	2		0	3	3	2	1	1	0	0	17	2	47
GJ	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
GN	1	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	1
GW	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
IC	0	0	0	0	0	0	0	0	0	0		0	0	0	0	1	0	1
JB	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0
JJ	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
JN	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
SJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0
SU	10	25	31	11	20	14	82	4	17	26	8	9	4	7	2		7	277
US	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0
Total	13	29	38	12	24	17	85	4	21	31	10	10	5	7	2	19	9	336

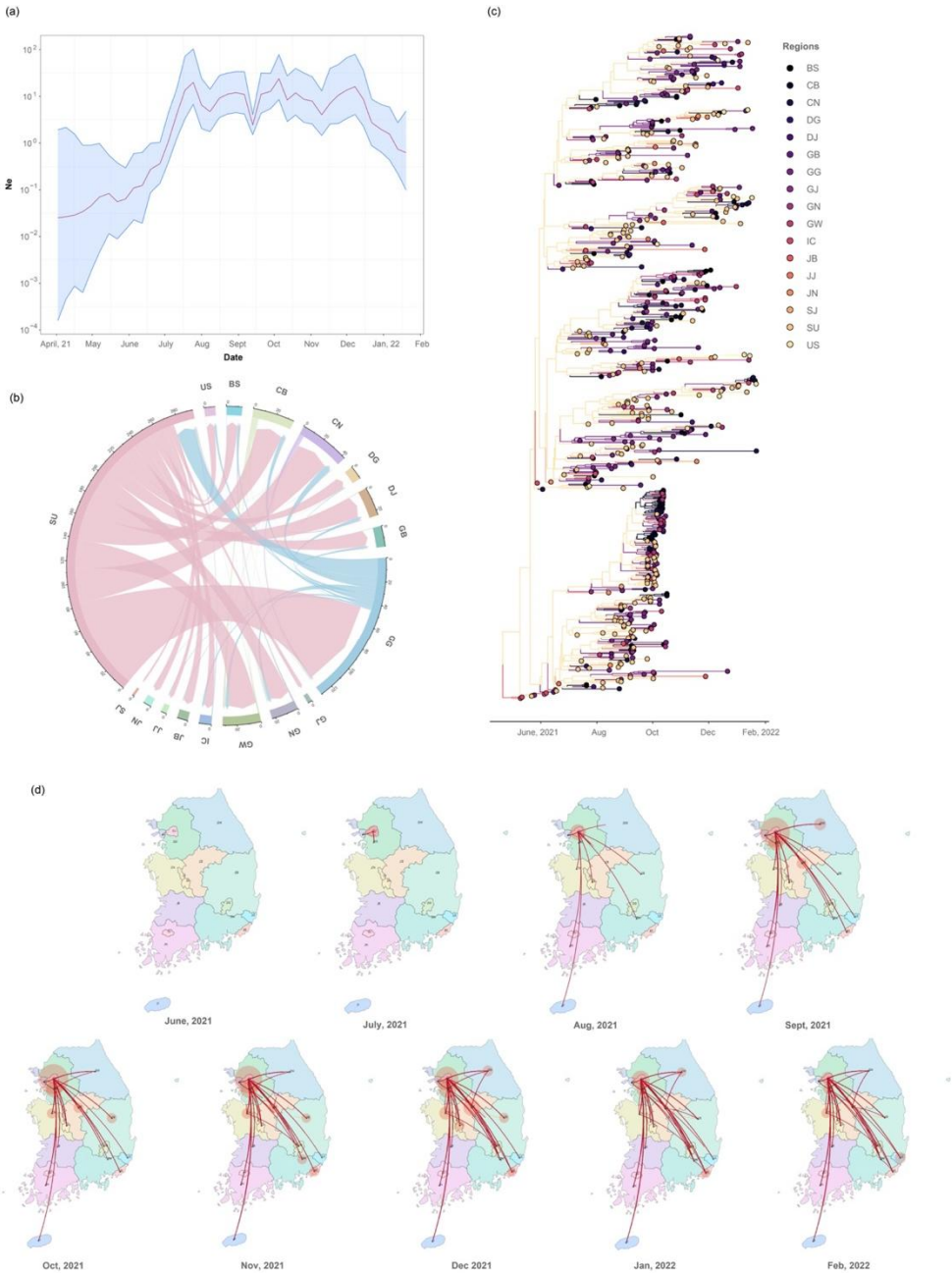


Figure 3. The effective population size of Ay.69 during the study period and its inter-region introduction events with a phylogenetic tree and SpreaD3 (Spatial Phylogenetics Reconstruction of Evolutionary Dynamics using Data-Driven Documents (D3))

(a) Bayesian skyline plot based on AY.69 variant. X-axis shows time scale in months and y-axis is the logarithmic effective population size. The red line represents the median shaded area indicates the 95% highest posterior density. (b) Chord diagram showing the introduction events between regions. Chord away from the edge represents the influx. The width or chord is relative to the number of flow. (c) Phylogenetic tree of AY.69 variant. The tree is the maximum clade credibility(MCC) summary of Bayesian inference. (d) The routes and magnitude of viral spread animation through serial months played in SpreadD3. Colours corresponds to each region in the legend. Ne, effective population size; BS, Busan; CB, Chung-buk; CN, Chung-nam; DG, Dae-gu; DJ, Dae-jeon; GB, Gyeong-buk; GN, Gyeong-nam; GG, Gyeong-gi; GW, Gang-won; IC, In-cheon; JJ, Je-ju; JN, Jeon-nam; SJ, Se-jong; SU, Seoul; US, Ul-san.

3.3. AY.69 variant mostly spread from Non-vaccinated group

Based on posterior average ratios of introductions of each immune status group, non-vaccinated (NV) groups led in the spatial expansion of the AY.69 variants in the period 2021-04-23, showing an outflux of 154 times (95% HPD = 140-168), among the total outfluxes of 171 (95% HPD = 149-195) in the entire tree. Fully-vaccinated (FV) groups represented the largest influx (100 times; 95% HPD = 90-113) since July 2021, followed by partially-vaccinated (55 times; 95% HPD = 50-59) and NV (16 times; 95% HPD = 9-23). (Fig. 4a)

The NV group was predominantly responsible for disease spread from April 27, 2021, to January 11, 2022 (Fig. 4b). Because the vaccination schedule peaked during summer (July to August 2021) in South Korea, with FV individuals exceeding 70% from October 23, 2021 [47], large virus influxes occurred from NV to FV and PV on June 6, 2021, and July 9, 2021, respectively. NV influx and outflow decreased after September 4, 2021, because of the absolute decline in non-vaccinated population; the PV group showed a similar trend after October 8, 2021 (Fig. 4c).

Table 2. Calculated transmission values among vaccine status group of the 680-sample dataset

	Transmitted to			
From	FV	NV	PV	Total
FV		16	0	16
NV	99		55	154
PV	1	0		1
Total	100	16	55	171

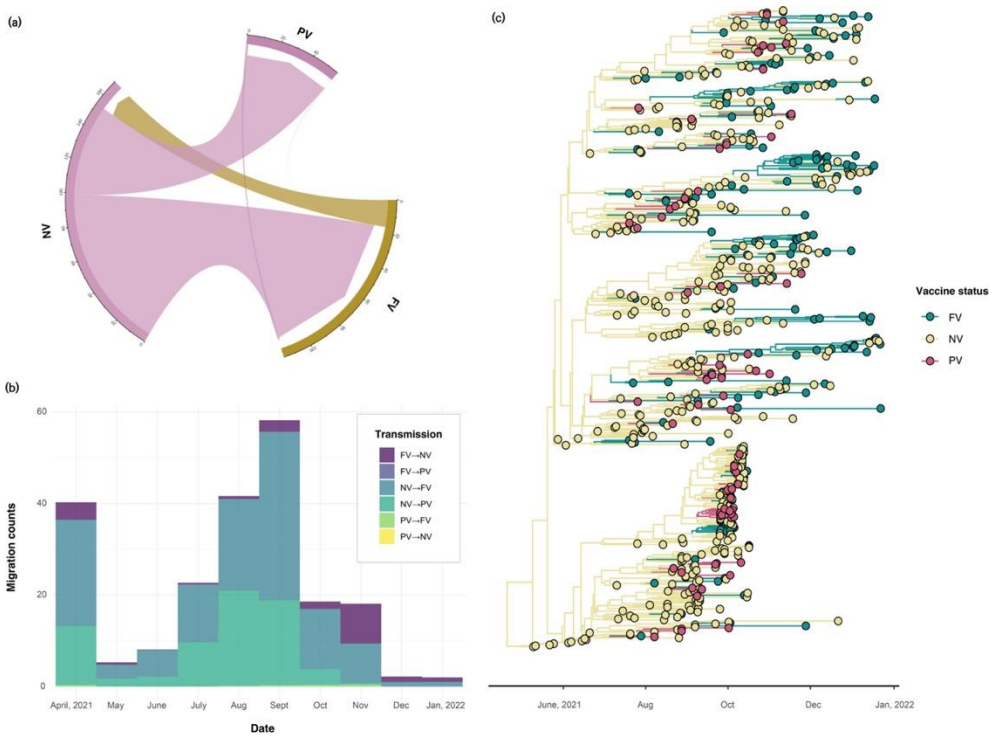


Figure 4. The introduction events between immune group with a phylogenetic tree and chord diagram

(a) Chord diagram showing the introduction events between immune groups. Chord away from the edge represents the influx. The width or chord is relative to the number of flow. (b) The number of transmission counts among immune groups for each month. (c) The

maximum clade credibility(MCC) summary tree of Bayesian inference. Colours correspond to each immune group in the legend. NV, Non-vaccinated group; PV, Partially-vaccinated group; FV, Fully-vaccinated group.

3.4. GLM analysis of mobility data and phylogeography

Because probabilistic inference is computationally demanding and time-consuming, we explored a lot of potential predictors before integrating all predictors in phylodynamic analysis [46]. (Appendix 2)

First, public transportations such as buses and trains cannot measure peoples' mobility of both inflow and outflow of Jeju-island. Beside from amount of shipment and aircraft, Jeju-island has zero transportation movement to the other regions whereas genetic sequences are collected in Jeju-island. As expected, we could find a little significant relationship between the number of passengers of intercity or express bus and the migration rates calculated in the phylogenetic tree. Even if the number of railway passengers shows the significant positive association, several regions doesn't have inter-regional railway (e.g., Incheon is connected to two regions only - Gyeonggi and Gangwon). Therefore, amount of movement would be better estimated by location of mobile phone measured by GPS.

Since a couple of previous studies suggested the relationship of inter-regional distance or population density with migration rates, we've identified their relevance by PhyCovA [48,49]. Distances and population density of disseminating regions had no association with migrations and importing regions showed insignificant association. (Appendix 2)

Based on the associations of every factor with inter-region-transmission rates in PhyCovA, we chose three most associated predictors besides from sample sizes - mobile phone mobility OD matrix, credit card data of disseminating region and vaccinated population of importing region. The selected predictors have positive relationship with the number of viral transmission events and all predictors showed low to moderate correlation. (Appendix 3)

We also considered the sample sizes of the source and destination regions separately [50]. GLM without sample size, but including vaccination and mobility variables, showed that total mobility between regions measured using mobile phones had a positive effect on viral spatial transmission in both datasets. In the 220-subsample dataset, credit card spending on entertainment, representative of movement, had a more positive effect on viral outflow than their effect in 642-subsample dataset. Bayes factor of credit card data and mobile phone mobility was both > 200 in large and small datasets. The effect size of mobile phone mobility was 0.97 and 0.902 respectively and the effect size of card mobility was 1.131 and 1.739 respectively (Fig. 5).

Table 3. Descriptive statistics of each predictor of GLM

	Unit	Mean	Median	Standard deviation	IQR
Card consumption	number of payments/day	1087.5	732.6	1075.3	462.5
Mobile phone mobility	people/day	595599.0	223248.0	1310065.1	393807.0
Vaccination	total population that completed initial vaccination	2477393.4	1552120.0	2692755.1	1225293.0
Population density	population/km^2	2122.9	788.0	3632.4	2519.0
Intercity bus	total traffic in month	18119.4	3218.4	40680.4	15532.3
Express bus	total traffic in month	10832.3	609.3	30713.2	4848.1
Distance	km	187.9	178.4	96.6	135.3
Rail	total traffic in month	36799.7	7850.0	70420.4	40874.0
Sample counts	counts	37.8	17.0	49.6	31.0

Table 4. Inclusion support statistics for each dataset

Including sample size model				
	680 samples		220 samples	
Predictor	Posterior inclusion probability	Bayes factor	Posterior inclusion probability	Bayes factor
Card consumption	0.108	0.1	0.365	0.8
Mobile phone mobility	0.304	0.1	0.015	0.0
Vaccination	0.023	0.00017	0.012	0.00072
Sample size of destination region	1	1.096	1	1.036
Sample size of origin region	0.966	2.166	0.718	1.840

Excluding sample size model				
	680 samples		220 samples	
Predictor	Posterior inclusion probability	Bayes factor	Posterior inclusion probability	Bayes factor
Card consumption	1	1.129	1	1.656
Mobile phone mobility	1	0.969	1	0.901
Vaccination	0.06	0.006	0.07378	0.011

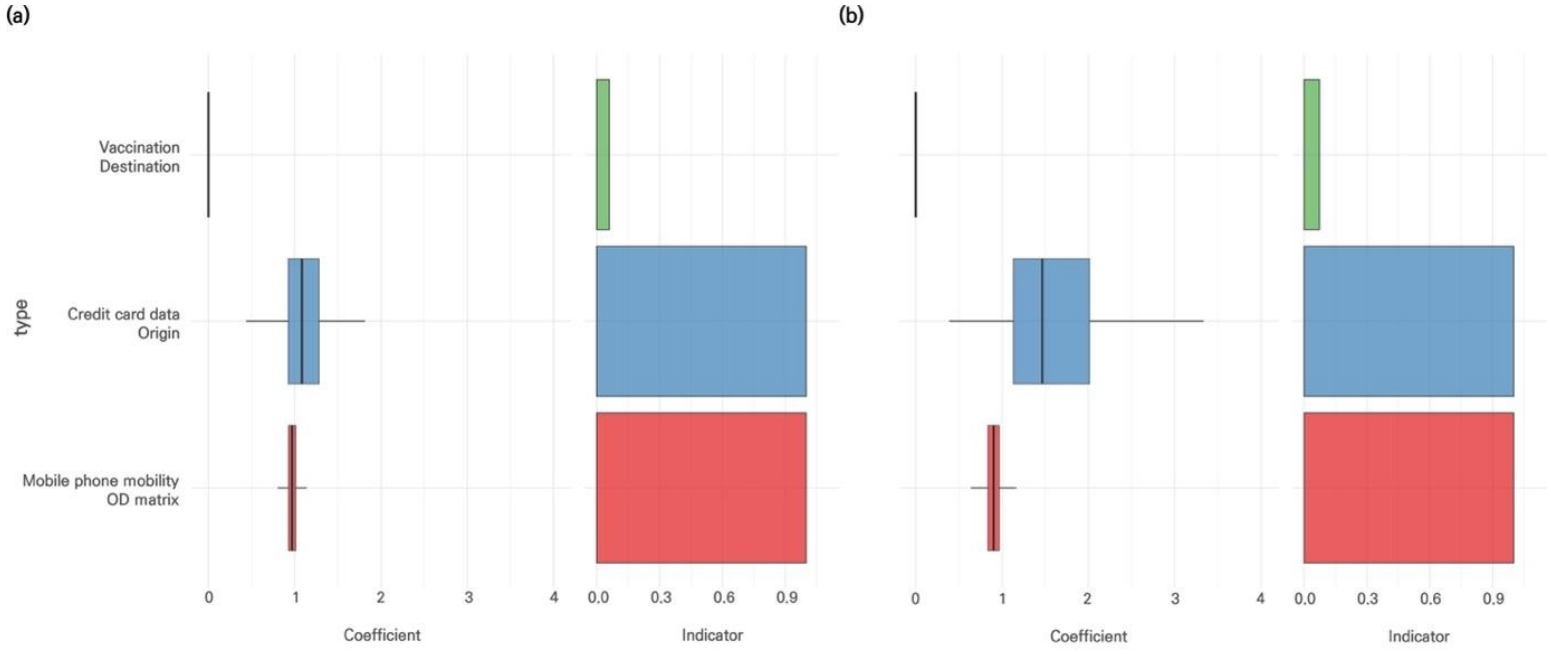


Figure 5. The results of generalized linear model of inter-region viral transmission and its predictors

The support and contribution of AY.69 variant transmission predictors. Support for each factor is represented by an inclusion probability(right) and the contribution of each factor is represented by the mean GLM coefficient on a log scale conditional on the predictor being included in the model(left). (a) GLM of 642-subsampled dataset. (b) GLM of 220-subsampled dataset.

Next, we considered sample sizes in our model. For the sample size of origin and destination regions, in the large dataset, the Bayes factors were >200 with a positive effect size 2.187 and 1.098 respectively. In the small dataset the sample size of origin regions had a Bayes factor of 17.04 with a positive effect size 1.955 and the destination regions had a Bayes factor >200 with a positive effect size 1.036. While vaccination was expected to have negative effect, it shows no relevance with viral transmission. In the large dataset, the mobile phone mobility was associated variable (BF = 2.84) with a positive effect size (EF = 0.878) and whereas the small dataset has an associated variable (BF = 3.86) with a positive effect size (EF = 0.9823). All the other mobility predictors had Bayes factors < 1. (Fig.6)

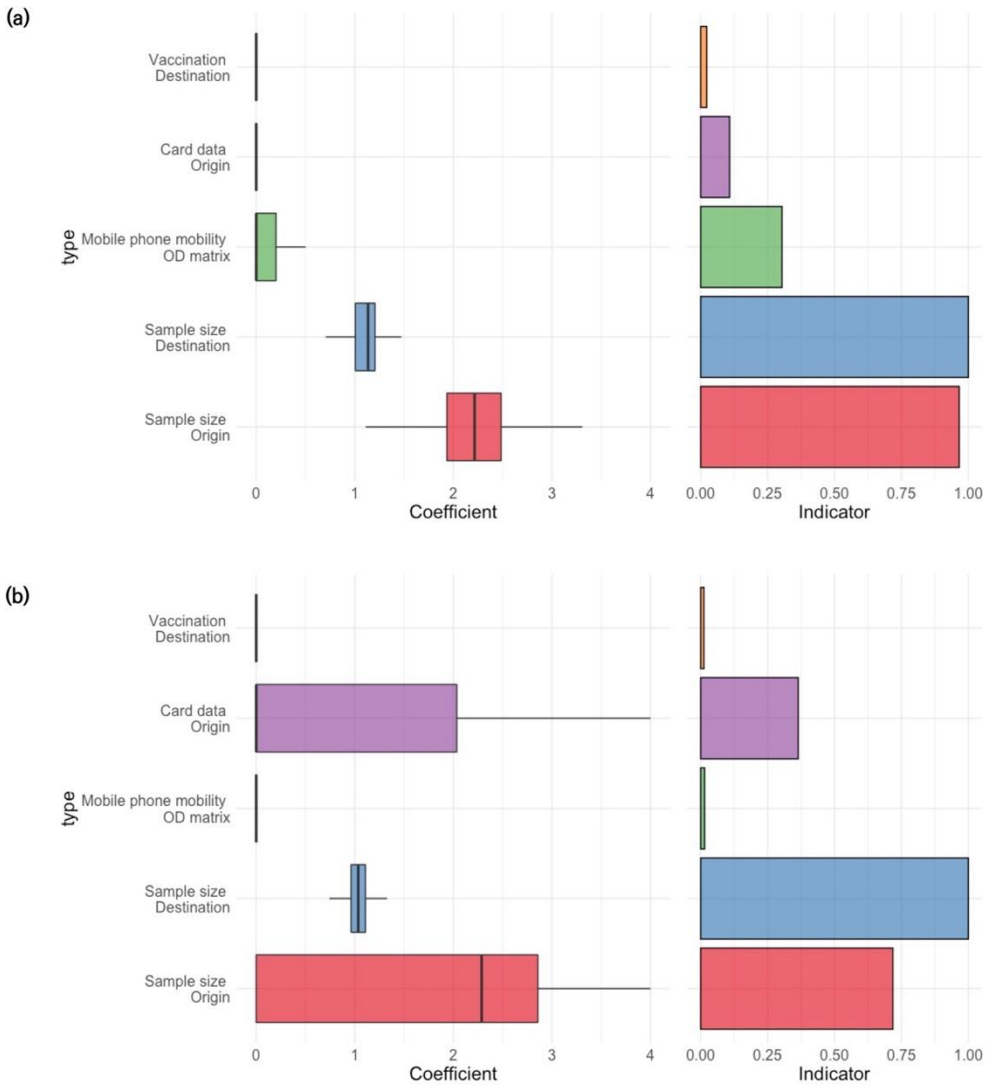


Figure 6. The results of generalized linear model of inter-regional viral transmission and its predictors considering sample sizes

The support and contribution of AY.69 variant transmission predictors. Support for each factor is represented by an inclusion probability(right) and the contribution of each factor is represented by the mean GLM coefficient on a log scale conditional on the predictor being included in the model(left)(a) GLM of 642-subsampled dataset. (b) GLM of 220-subsampled dataset.

Chapter 4. Discussion

Through this study, we could investigate transmission of COVID-19 Delta variant AY.69 in South Korea between regions and immune groups. Since this information was lack in traditional surveillance system, it is an additional significant study understanding the characteristics of disease outbreak when came into Korea and preparing for the future pandemic. According to the results, viral transmission began in late-May from metropolitan regions and the regions led viral spread in the overall period (2021.05~2022.01). Furthermore, non-vaccinated groups led the virus spread in a partially/fully vaccinated group throughout the study period even though full vaccination rate in Korea exceeded 70% in October.

Our phylodynamic analysis showed that unvaccinated individuals contributed to the initial spread of AY.69 in South Korea. In fact, the proportion of non-vaccinated South Koreans had been declining precipitously throughout September, coinciding with the spike of transmission among other groups of immune status. Even with a partial or full vaccination rate over 57% on Sept. 1 and about 24% remaining unvaccinated on Sept. 30 [51], the unvaccinated were overwhelmingly the ones who transmitted viruses. Therefore, herd immunization and its impact have been estimated by phylodynamics and found to reduce burden of disease. The effectiveness of vaccine would be better at the beginning of pandemic as our results show, then genetic diversity has reduced vaccine efficacy. [52]

According to our results from the GLM, human motility accounts for spatiotemporal transmission of SARS-Cov-2 AY.69 variant. Consistent with previous findings, human movement estimated from mobile phone had significant influence on virus spread [53], which means the large people's movements in metropolitan explain predominant spread exported from Seoul and Gyeong-gi. People's mobility has been estimated by many different data sources in the previous studies including bus or railway passengers and air traffic. [54] The mobile phone mobility, on the other hand, measures all the mobility whatever people have taken on for their transportation, which recommends that phylodynamics would better estimates disease outbreak when we consider all the movement, not a specific transportation. Other factors such as population density and distances between regions in phylogeography with GLM study have been considered in the previous studies, [19 55], but we predicted that these data would have low relevance (Appendix 2). We also measured mobility based on credit card spending on, for example, sports, movies, entertainment, and other personal interests, which were related to phylodynamic movements. Although an asymmetrical mobile-phone-detected region-to-region mobility matrix and credit card spending data have a little impact on cross-regional transmission in the present, it also shows moderate Bayes factors, leaving a door open for future findings avoiding sampling bias [45].

To sum up, implementing non-pharmaceutical intervention to

reduce people mobility would be the most effective way to flatten the curve. Even herd immunization seems to be efficient in the low prevalence to reduce transmission, vaccinated people spread virus more in the later phase with large outbreak. Furthermore, viral dissemination can be predicted incorporating with data that explains people's inter-region mobility. We investigated the possibility of adopting O/D matrix mobile phone data and credit card consumption data to be used as an informative source of epidemiological surveillance.

This phylogeography and GLM study gives us insight to predict future pandemic overcoming the limitation of current phylogenetic study in the forecasting aspect. [4] Furthermore, an epoch model can consider random and fixed effects in following study where the model can add time-homogeneous random effects, although this analysis would be computationally burden. [53] Recently, several methodologies of disease predicting by phylogenetic methods are proposed or theoretically hypothesized [56,57]. If the most relevant predictors of viral spread are accumulated by numerous studies, the most adequate model will be designed and be utilized to predict next pandemic transmission. In the meanwhile, the migration rates between regions itself acquired from our study as who acquires infection from whom (WAIFW) matrix form can be adopted in a mathematical modelling. When another demographic factor such as age is designated as a trait, a new WAIFW matrix of age groups can be used in mathematical models by

age groups instead of surveyed contact matrix which is most frequently used currently.

This study needs to be interpreted cautiously as we performed a practice of down-sampling from selected genetic sequence data. The sequencing frequency over the period studied was about an average of 0.8 percent of total detected cases in Korea [58], meaning that unsampled cases may explain unreported transition events. Even though a particular variant could be densely populated locally and the proportion in our sample was lower compared with the real incidence, absolute sample sizes were smaller at some regions than at others, thus, some transmissions may have been missed. However, as people are crowded in Metropolitan regions in South Korea, these regions recorded the high incidence rate throughout pandemic and unbalanced sample counts for each region seems to be inevitable. Indeed, the pandemic aspects in the study period were distinguished by a concentration of cases in metropolitan areas including Seoul and Gyeong-gi (67%), but these regions represented approximately 52% of total AY.69-variant pandemic cases in South Korea. (Appendix 1) Sampling bias still cannot be avoided showing considerably overwhelming effect sizes in GLM. The other method of down-sampling can be tried or putting every prominent predictor in a GLM could be tried in the following study.

Considering even less biased sampling rates, spreading from the large regions was a general trend. The results suggest that when

the new viral lineage first introduced in metropolitan regions in South Korea, this can be led to the unprecedent big surges, which a new initial detection of virus in nearby metropolitan regions allows us to predict a following larger outbreak throughout Korea peninsula in the future pandemic

Chapter 5. Conclusion

According to the phylogeographic analysis, the pandemic of SARS-CoV-2 Delta variant of concern in South Korea disseminated from metropolitan regions from May 2021 to January 2022. To investigate factors mostly affecting viral spatiotemporal spread, generalized linear model was implemented and the average amount of people's credit card consumption and total mobility measured by their mobile phones could explain a part of phylocynamics of disease outbreak. Moreover, this phylogeographic analysis showed that virus majorly spread from non-vaccinated people to partially or fully vaccinated people. Therefore, it would be possible to reduce viral spatial transmission among regions in South Korea implementing both Non-pharmaceutical intervention (NPI) and vaccination policies.

Bibliography

1 Grenfell BT. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. .

2 Grenfell BT, Dobson AP. *Ecology of infectious diseases in natural populations / edited by B.T. Grenfell, A.P. Dobson*. Cambridge University Press, 1995.

3 Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nat Rev Genet. 2022; **23**: 547–562.

4 Hill V, Ruis C, Bajaj S, Pybus OG, Kraemer MUG. Progress and challenges in virus genomic epidemiology. Trends Parasitol. 2021; **37**: 1038–1049.

5 Lee EC, Wada NI, Grabowski MK, Gurley ES, Lessler J. The engines of SARS-CoV-2 spread. *Science (1979)* 2020; **370**: 406–407.

6 Dalziel BD, Kissler S, Gog JR *et al.* Urbanization and humidity shape the intensity of influenza epidemics in U.S. cities. https://www.science.org.

7 Sachs JD, Karim SSA, Aknin L *et al.* The Lancet Commission on lessons for the future from the COVID-19 pandemic. *The Lancet* 2022. doi:https://doi.org/10.1016/S0140-6736(22)01585-9.

8 Aggarwal D, Myers R, Hamilton WL *et al.* The role of viral genomics in understanding COVID-19 outbreaks in long-term care facilities. *Lancet Microbe* 2022; **3**: e151–e158.

9 Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biol* 2009; **5**: 1000520.

10 Gangavarapu K, Latif AA, Mullen JL *et al.* Outbreak.info genomic

reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *medRxiv* 2022; : 2022.01.27.22269965.

11 Lim S, Sohn M. How to cope with emerging viral diseases: Lessons from South Korea's strategy for COVID-19, and collateral damage to cardiometabolic health. *Lancet Reg Health West Pac* 2022; : 100581.

12 Lee H, Choi EH, Park YJ, Choe YJ. Short Term Impact of Coronavirus Disease 2019 Vaccination in Children in Korea. *J Korean Med Sci* 2022; **37**: 1–6.

13 How Asia, Once a Vaccination Laggard, Is Revving Up Inoculations - The New York Times. https://www.nytimes.com/2021/09/30/business/economy/asia-covid-vaccinations.html (accessed 22 Sep2022).

14 국민들의 높은 참여와 관심으로 코로나19 예방접종 1차 접종률 80% 달성했습니다. | 카드뉴스 | 홍보자료 | 알림·자료 : 질병관리청. https://www.kdca.go.kr/gallery.es?mid=a20503010000&bid=0003&b_list=9&act=view&list_no=145369&nPage=1&vlist_no_npage=1&keyField=&keyWord=&orderby= (accessed 22 Sep2022).

15 Norheim OF. Protecting the population with immune individuals. *Nature Medicine 2020 26:6* 2020; **26**: 823–824.

16 Coronavirus Pandemic (COVID-19) - Our World in Data. https://ourworldindata.org/coronavirus (accessed 22 Sep2022).

17 Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Comput Biol* 2017; **13**: e1005697.

18 Vrancken B, Zhao B, Li X *et al.* Comparative Circulation Dynamics of the Five Main HIV Types in China. *J Virol* 2020; **94**. doi:10.1128/jvi.00683-20.

19 Lu L, Leigh Brown AJ, Lycett SJ. Quantifying predictors for the spatial diffusion of avian influenza virus in China. *BMC Evol Biol* 2017; **17**.

doi:10.1186/s12862-016-0845-3.

20 Wang P, Liu L, Nair MS *et al.* SARS-CoV-2 neutralizing antibody responses are more robust in patients with severe disease. Emerg Microbes Infect. 2020; **9**: 2091–2093.

21 Snoeijer BT, Burger M, Sun S, Dobson RJB, Folarin AA. Measuring the effect of Non-Pharmaceutical Interventions (NPIs) on mobility during the COVID-19 pandemic using global mobility data. *NPJ Digit Med* 2021; **4**. doi:10.1038/s41746-021-00451-2.

22 Gibbs H, Liu Y, Abbott S *et al.* Association between mobility, non-pharmaceutical interventions, and COVID-19 transmission in Ghana: A modelling study using mobile phone data. *PLOS Global Public Health* 2022; **2**: e0000502.

23 Bollen N, Artesi M, Durkin K *et al.* Exploiting genomic surveillance to map the spatio-temporal dispersal of SARS-CoV-2 spike mutations in Belgium across 2020. *Sci Rep* 2021; **11**: 18580–18580.

24 du Plessis L, McCrone JT, Zarebski AE *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* 2021; **371**: 708–712.

25 Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nature Reviews Genetics 2022 23:9* 2022; **23**: 547–562.

26 Duerr R, Dimartino D, Marier C *et al.* Dominance of Alpha and Iota variants in SARS-CoV-2 vaccine breakthrough infections in New York City. *J Clin Invest* 2021; **131**. doi:10.1172/JCI152702.

27 Wang Y, Chen D, Zhu C *et al.* Genetic Surveillance of Five SARS-CoV-2 Clinical Samples in Henan Province Using Nanopore Sequencing. *Front Immunol* 2022; **13**: 1412.

28 Lemey P, Rambaut A, Bedford T *et al.* Unifying viral genetics and

human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog* 2014; **10**. doi:10.1371/JOURNAL.PPAT.1003932.

29 Khare S, Gurry C, Freitas L *et al.* GISAID's Role in Pandemic Response. *China CDC Weekly, 2021, Vol 3, Issue 49, Pages: 1049-1051* 2021; **3**: 1049–1051.

30 O'Toole Á, Scher E, Underwood A *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* 2021; **7**. doi:10.1093/VE/VEAB064.

31 O'Toole Á, Hill V, Pybus OG *et al.* Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2 with grinch. *Wellcome Open Res* 2021; **6**: 121.

32 Aksamentov I, Roemer C, Hodcroft EB, Neher RA. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J Open Source Softw* 2021; **6**: 3773.

33 Hadfield J, Megill C, Bell SM *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018; **34**: 4121–4123.

34 Marini S, Mavian C, Riva A, Prosperi M, Salemi M, Magalis BR. Optimizing viral genome subsampling by genetic diversity and temporal distribution (TARDiS) for phylogenetics. *Bioinformatics* 2022; **38**: 856–860.

35 Rambaut A, Holmes EC, O'Toole Á *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology 2020 5:11* 2020; **5**: 1403–1407.

36 Kearse M, Moir R, Wilson A *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012; **28**: 1647–1649.

37 Price MN, Dehal PS, Arkin AP. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol Biol*

Evol 2009; **26**: 1641–1650.

38 Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016; **2**. doi:10.1093/VE/VEW007.

39 Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 2018; **4**. doi:10.1093/VE/VEY016.

40 Bouckaert RR, Drummond AJ. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol Biol* 2017; **17**: 1–11.

41 Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* 2018; **67**: 901–904.

42 Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007; **7**: 1–8.

43 FigTree. http://tree.bio.ed.ac.uk/software/figtree/ (accessed 23 Sep2022).

44 Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: Interactive Visualization of Spatiotemporal History and Trait Evolutionary Processes. *Mol Biol Evol* 2016; **33**: 2167–2169.

45 한국관광 데이터랩. https://datalab.visitkorea.or.kr/datalab/portal/main/getMainForm.do (accessed 22 Sep2022).

46 Blokker T, Baele G, Lemey P, Dellicour S. Phycova — a tool for exploring covariates of pathogen spread. *Virus Evol* 2022; **8**. doi:10.1093/ve/veac015.

47 보도자료 내용보기 " 전 국민 70% 접종 완료, 단계적 일상회복 발판 마련 " < 뉴스 & 이슈 < 코로나바이러스감염증-19. http://ncov.mohw.go.kr/tcmBoardView.do?brdId=3&brdGubun=31&da

taGubun=&ncvContSeq=6033&contSeq=6033&board_id=312&gubun=ALL (accessed 22 Sep2022).

48 Vrancken B, Zhao B, Li X *et al.* Comparative Circulation Dynamics of the Five Main HIV Types in China. *J Virol* 2020; **94**. doi:10.1128/jvi.00683-20.

49 He WT, Bollen N, Xu Y *et al.* Phylogeography Reveals Association between Swine Trade and the Spread of Porcine Epidemic Diarrhea Virus in China and across the World. *Mol Biol Evol* 2022; **39**. doi:10.1093/molbev/msab364.

50 He WT, Bollen N, Xu Y *et al.* Phylogeography Reveals Association between Swine Trade and the Spread of Porcine Epidemic Diarrhea Virus in China and across the World. *Mol Biol Evol* 2022; **39**. doi:10.1093/MOLBEV/MSAB364.

51 질병관리청 코로나19 백신 및 예방접종: National Center for Mental Health. https://ncv.kdca.go.kr/ (accessed 23 Sep2022).

52 Cardona-Ospina JA, Rojas-Gallardo DM, Garzón-Castaño SC, Jiménez-Posada E v., Rodríguez-Morales AJ. Phylodynamic analysis in the understanding of the current COVID-19 pandemic and its utility in vaccine and antiviral design and assessment. Hum Vaccin Immunother. 2021; **17**: 2437–2444.

53 Lemey P, Ruktanonchai N, Hong SL *et al.* Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* 2021; **595**: 713.

54 Malik O, Gong B, Moussawi A, Korniss G, Szymanski BK. Modelling epidemic spread in cities using public transportation as a proxy for generalized mobility trends. *Sci Rep* 2022; **12**. doi:10.1038/s41598-022-10234-8.

55 Bui CM, Adam DC, Njoto E, Scotch M, MacIntyre CR. Characterising routes of H5N1 and H7N9 spread in China using Bayesian phylogeographical analysis. *Emerg Microbes Infect* 2018; **7**.

doi:10.1038/s41426-018-0185-z.

56 Scrima M, Cossu AM, D'Andrea EL *et al.* Genomic Characterization of the Emerging SARS-CoV-2 Lineage in Two Districts of Campania (Italy) Using Next-Generation Sequencing. *Frontiers in Virology* 2022; **2**. doi:10.3389/fviro.2022.814114.

57 Volz E, Mishra S, Chand M *et al.* Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* 2021; **593**: 266–269.

58 Chen C, Nadeau S, Yared M *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics* 2022; **38**: 1735–1737.

국문초록

한국 SARS-CoV-2 대유행의 베이지안 계통지리학적 분석과 이에 영향을 미치는 요인

이상이

보건학과 보건학 전공

서울대학교 보건대학원

2019년에 SARS-CoV-2 알파 우려 변종(VOC, Variant of Concern)이 유행한 이후 2021년 델타 변이 중 특히 AY.69 우려 변종 바이러스가 한국의 코로나바이러스 유행을 이끌었다. AY.69는 2021년 5월부터 2022년 1월까지 백신접종 프로그램이나 사회적 거리두기와 같은 비약물적 중재정책을 도입했음에도 불구하고 한국에서 특히 큰 유행을 이끌었다. 본 연구에서는 선형회귀모델(GLM) 분석을 통해 사람들의 이동과 면역도와 바이러스의 전파와의 관계를 알아보기 위해 계통지리학적 분석을 실시하였다. 결과에 따르면 전파는 한국의 수도권 지역에서 시작되었으며, 해당 지역 사람들의 신용카드 사용량과 휴대용 이동통신기기의 GPS로 측정한 사람들의 모든 이동량이 다른 지역으로의 바이러스 유입과 관련이 있는 것으로 나타났다. 또한 본 계통역학적 연구는 한국에서는 AY.69 변이 바이러스가 유행하기 3개월 전에 백신접종 프로그램이 시작되었지만, 백신 접종을 하지 않은 사람들이 유행기간동안 바이러스 전파를 주도하였다는 것을 계통지리학적 분석을 통해 밝혔다. 따라서 본 연구는 비약물적 중재정책과

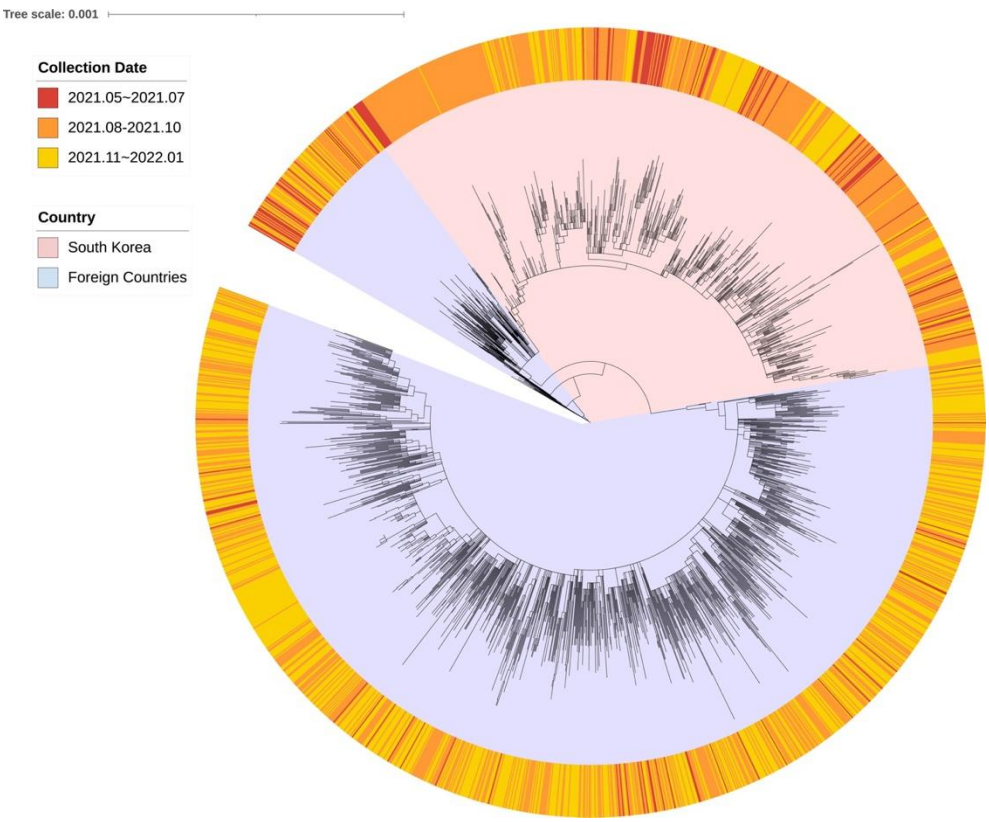
백신접종 프로그램을 동시에 실시하는 것이 바이러스의 전파를 효율적으로 막을 수 있다는 것을 제안한다.

주요어 : SARS-CoV-2, 계통지리학, 계통역학, 분자역학, BEAST, 인구이동, 일반화선형모델, 한국

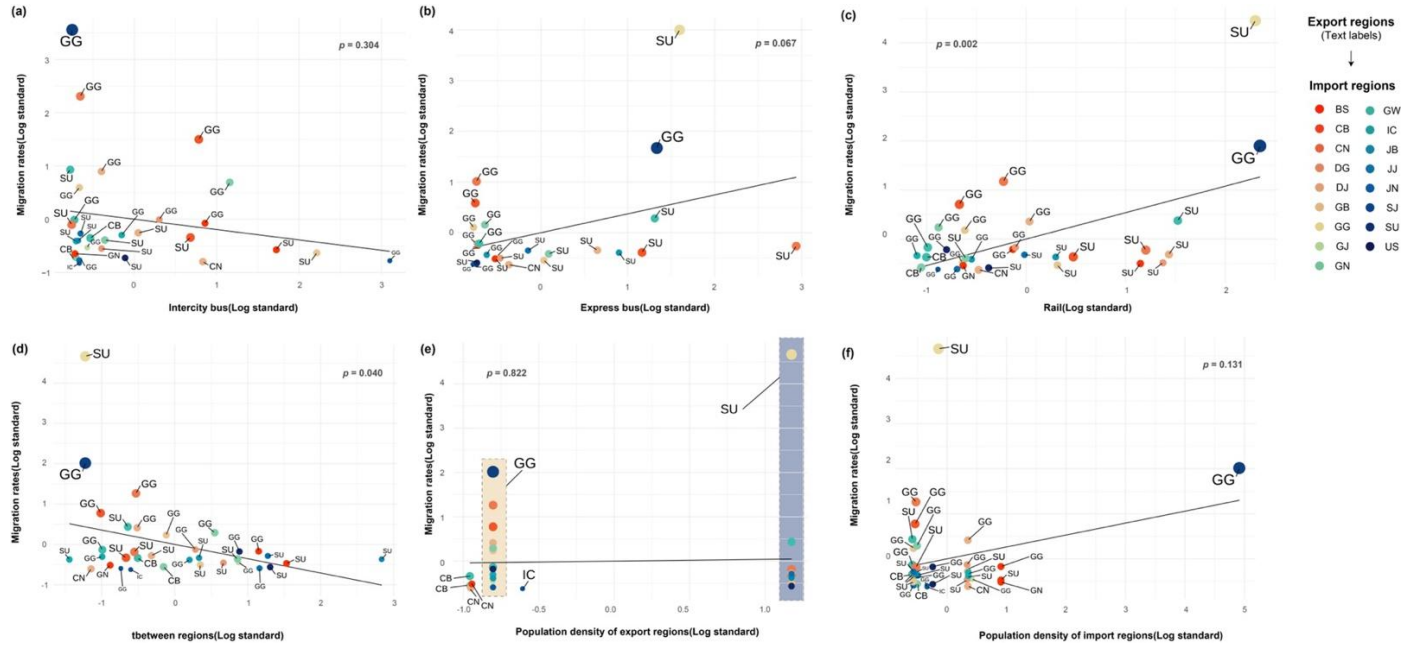
학번 : 2021-27881

Appendix

Appendix 1. A maximum likelihood phylogenetic tree formed by FastTree v.2.1.11

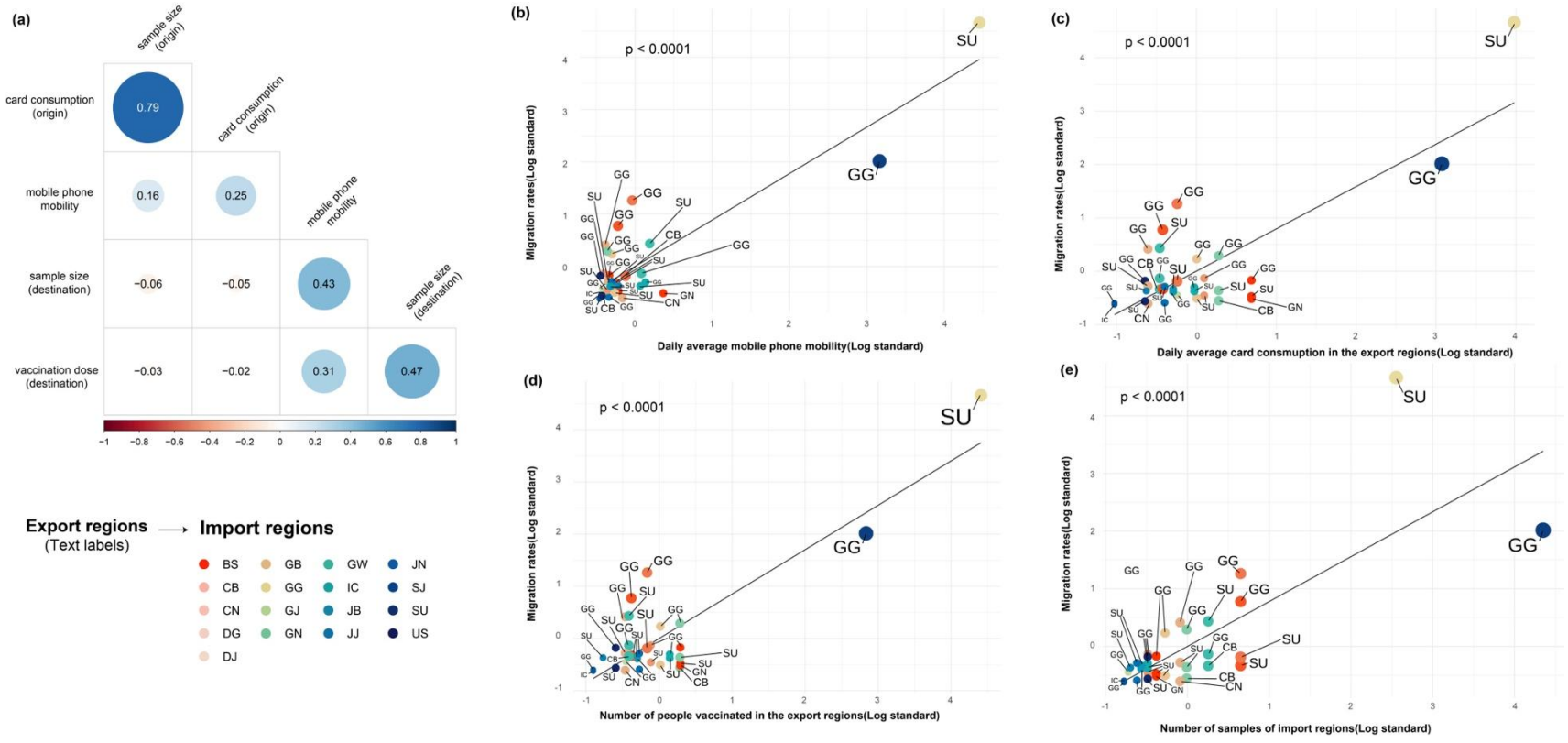


Appendix 2. Correlations of predictors in the generalized linear model and trends comparing the migration rates and predictors for each region



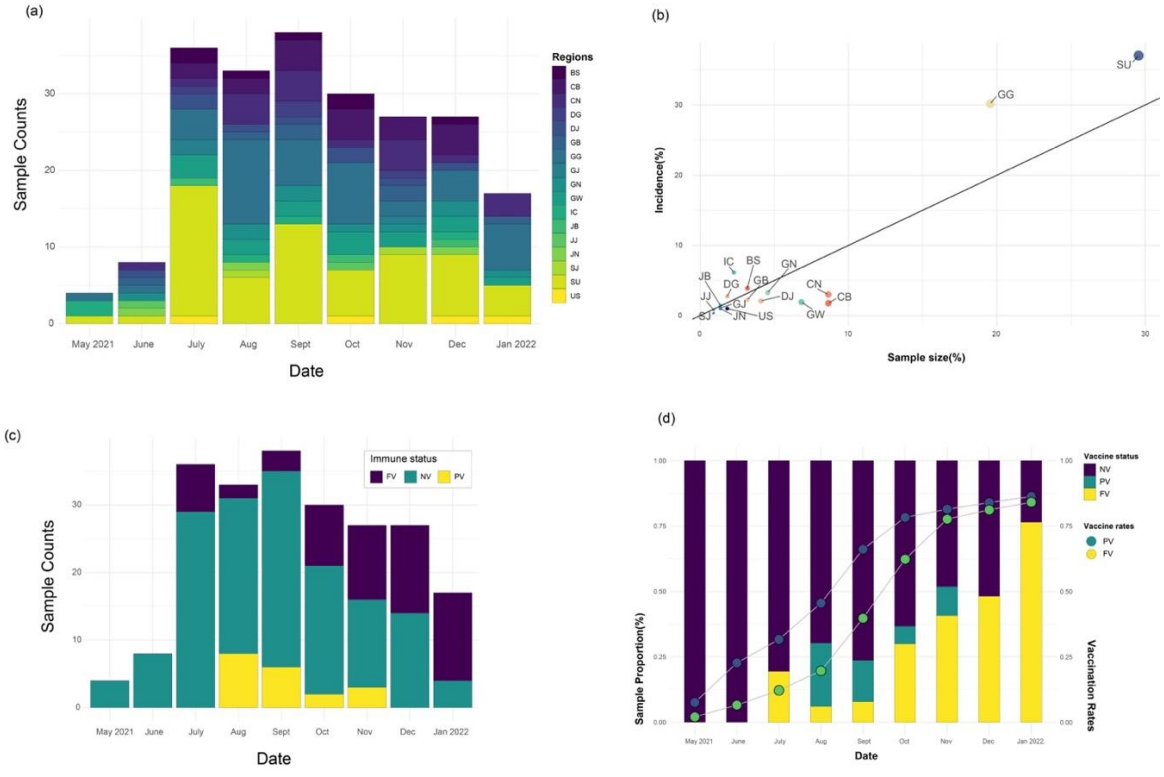
Associations with migration rates; x−axis log−scale (a) intercity bus, (b) express bus, (c) railway, (d) intercity distance, (e, f) population density. BS, Busan; CB, Chungbuk; CN, Chungnam; DG, Daegu; DJ, Daejeon; GB, Gyeongbuk; GN, Gyeongnam; GG, Gyeonggi; GW, Gangwon; IC, Incheon; JJ, Jeju; JN, Jeonnam; SJ, Sejong; SU, Seoul; US, Ulsan

Appendix 3. Correlation of each predictor used in GLM and trends comparing migration rates and each predictors of every region



(a) The correlation matrix of each predictors. Card spending and sample size were inevitably correlated since metropolitan regions have dense population spending money and being infected. (b–e) Associations with migration rates; x–axis log–scale (b) phone mobility, (c) credit card spending, (d) vaccinated population, and (e) genome sequence sample size. BS, Busan; CB, Chungbuk; CN, Chungnam; DG, Daegu; DJ, Daejeon; GB, Gyeongbuk; GN, Gyeongnam; GG, Gyeonggi; GW, Gangwon; IC, Incheon; JJ, Jeju; JN, Jeonnam; SJ, Sejong; SU, Seoul; US, Ulsan.

Appendix 4. Comparison on sequence samples and incidence counts of each region and immune groups of 220-subsampled tree



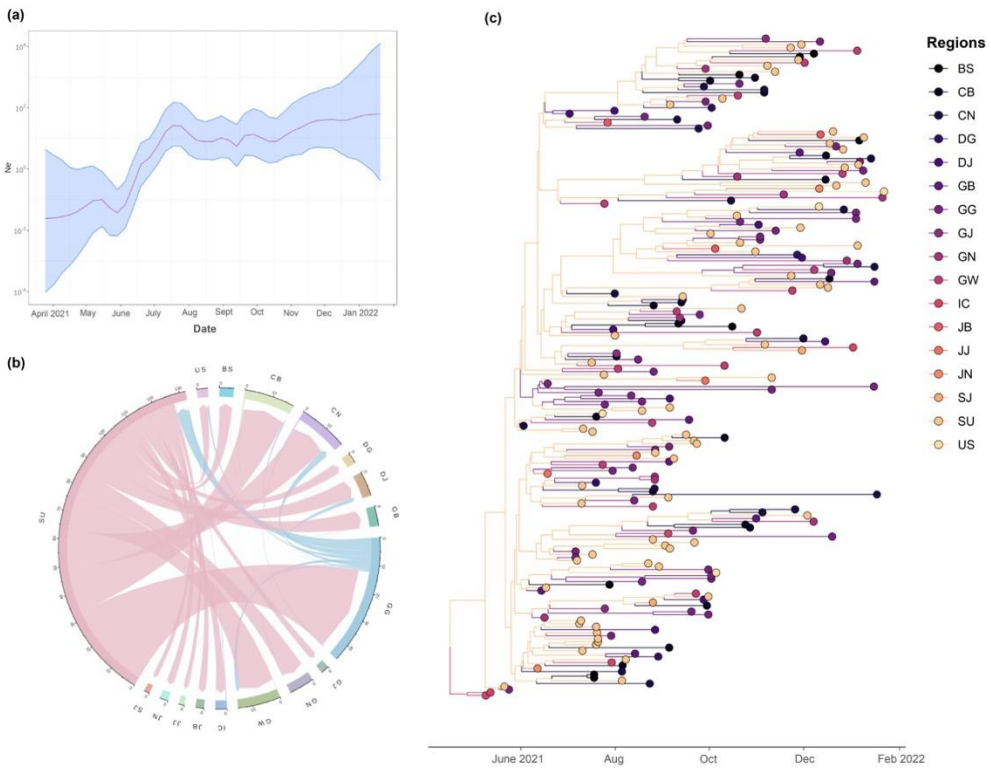
(a) Monthly sample counts and incidence proportions of each region from May 2021 to January 2022. Incheon accounted for most samples in May 2021, whereas Seoul and Gyeonggi predominated subsequently. Colors indicate regions as labelled on the right side of the figure. (b) Comparison of sample counts and incidence proportion in South Korea. Each region is labelled on the figure. (c) Sample count by immune group and vaccination rate. (d) Proportions of immune groups (bar graph) with vaccination rates (line graph). Colors are labelled on the right side of the figure. BS, Busan; CB, Chungbuk; CN, Chungnam; DG, Daegu; DJ, Daejeon; GB, Gyeongbuk; GN, Gyeongnam; GG, Gyeonggi; GW, Gangwon; IC, Incheon; JJ, Jeju; JN, Jeonnam; SJ, Sejong; SU, Seoul; US, Ulsan; NV, non−vaccinated group; PV, partially vaccinated group; FV, fully vaccinated group.

Appendix 5. Calculated cross-regional transmission values of the 220-sample dataset

Transmitted to

From	BS	CB	CN	DG	DJ	GB	GG	GJ	GN	GW	IC	JB	JJ	JN	SJ	SU	US	Total
BS		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CB	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CN	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DG	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0
DJ	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0
GB	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0
GG	0	1	2	0	2	0		0	0	1	0	0	0	0	0	5	1	12
GJ	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0
GN	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0
GW	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0
IC	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0
JB	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0
JJ	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0
JN	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0
SJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0
SU	5	16	14	4	7	7	33	3	9	14	4	3	2	3	2		3	129
US	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0
Total	5	17	16	4	9	7	33	3	9	15	4	3	2	3	2	5	4	141

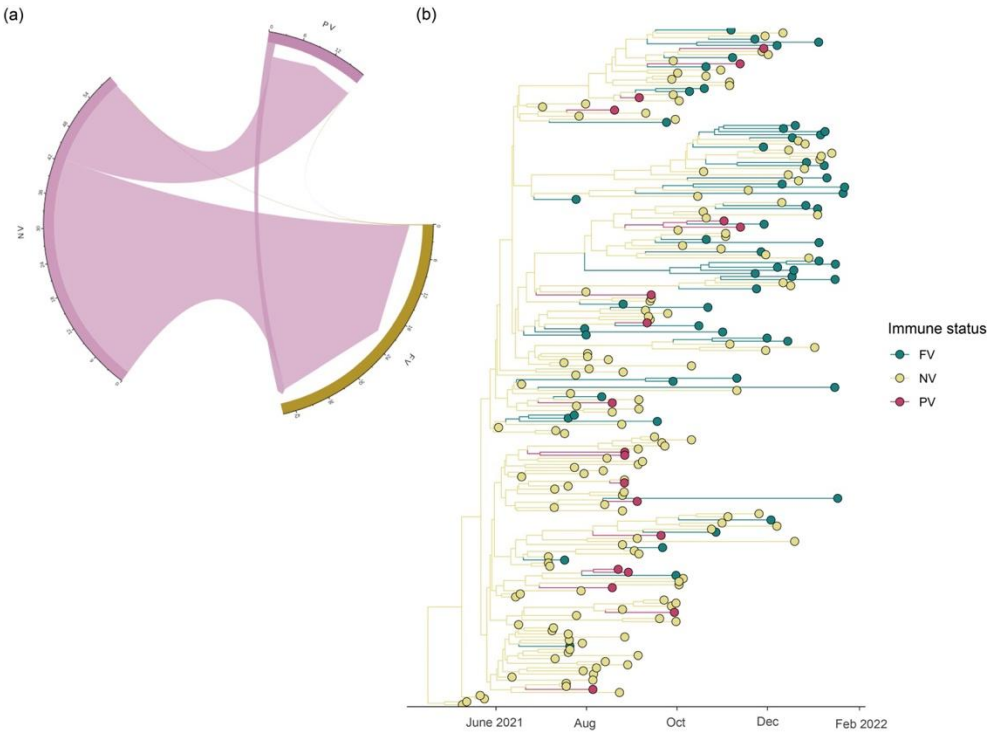
Appendix . The effective population size of Ay.69 during the study period and its inter-regonal introduction events with a phylogenetic tree of 220-subsampled tree



(a) Bayesian skyline plot based on AY.69 variant. X-axis shows time scale in months and y-axis is the logarithmic effective population size. The red line represents the median shaded area indicates the 95% highest posterior density. It was established that the most recent common ancestor(tMRCA) for AY.69 was April 18, 2021(95% height posterior density [HPD]: April 26 through May 9). The estimated effective virus population size increased through mid-July, showing a plateau thereafter. The effective population size reached its first peak on 18 July, just prior to heightened social distancing in 27 July 2021. (b) Chord diagram showing the introduction events

between regions. Chord away from the edge represents the influx. The width or chord is relative to the number of flow. Based on posterior average ratios of each region's introduction, dispersal between metropolises led spatial transmission of the AY.69 variation in the 2021-05-08 timeframe, with virus mainly spreading out of Seoul during the entire period, showing 127 times outfluxes out of the total 138 outfluxes within the whole tree, followed by Gyeong-gi (46 times). Gyeong-gi represented the largest influx (33 times) mainly from Seoul, with Chung-buk (17 times), Chung-nam (16 times) following. (c) Phylogenetic tree of AY.69 variant. The tree is the maximum clade credibility (MCC) summary of Bayesian inference. Colors are corresponds to each region in the legend. Ne, effective population size; BS, Busan; CB, Chung-buk; CN, Chung-nam; DG, Dae-gu; DJ, Dae-jeon; GB, Gyeong-buk; GN, Gyeong-nam; GG, Gyeong-gi; GW, Gang-won; IC, In-cheon; JJ, Je-ju; JN, Jeon-nam; SJ, Se-jong; SU, Seoul; US, Ul-san.

Appendix 7. The introduction events between immune group with a phylogenetic tree and chord diagram of 220-subsampled tree



(a) Chord diagram showing the introduction events between immune groups. Chord away from the edge represents the influx. The width or chord is relative to the number of flow. (b) The maximum clade credibility(MCC) summary tree of Bayesian inference. Based on posterior average ratios of introductions of each immune status group, nonvaccinated (NV) groups led in the spatial expansion of the AY.69 variants in the period 2021-04-27, showing an outflux of 40 times. Fully-vaccinated(FV) groups represented the largest inflow (25 times) since April 2021, followed by partially-vaccinated (15 times). Large virus influxes were observed from 2021-04-27, both from

NV to FV and PV. Both the outflow of NV decreased from 2021-09-13, because of the absolute decline in that compartment. We noted that the FV groups, with a relatively low sampling rate at first, had fewer viral spread, but started spreading viruses to the nonvaccinated groups starting from the end of 2021, as the majority were getting the vaccine. Colors are corresponds to each immune group in the legend. NV, Non-vaccinated group; PV, Partially-vaccinated group; FV, Fully-vaccinated group.

Appendix 8. **Calculated transmission values among vaccine status group of the 220-sample dataset**

Transmitted to

From	FV	NV	PV	Total
FV		0	0	0
NV	43		16	59
PV	2	0		2
Total	45	0	16	61