Detection of Phenotype-Related Mutations of COVID-19 via the Whole Genomic Data

Jinxiong Lv, Shikui Tu, and Lei Xu

Abstract—The coronavirus disease 2019 (COVID-19) epidemic continues to spread rapidly around the world and nearly 20 millions people are infected. This paper utilises both single-locus analysis and joint-SNPs analysis for detection of significant single nucleotide polymorphisms (SNPs) in the phenotypes of symptomatic versus asymptomatic, the early collection time versus the late collection time, the old versus the young, and the male versus the female. Also, this paper analyses the relationship between any two SNPs via linkage disequilibrium analysis, and visualises the patterns of cumulative mutations of SNPs over collection time. The results are in three folds. First, the SNP which locates at the nucleotide position 4321 is found to be an independent significant locus associated with all the first three phenotypes. Moreover, 12 significant SNPs are found in the first two studies. Second, gene orf1ab containing SNP-4321 is detected to be significantly associated with the first three phenotypes, and the three genes S, ORF3a, and N, are detected to be significant in the first two phenotypes. Third, some of the detected genes or SNPs are related to the SARS-COV-2 as supported by literature survey, which indicates that the results here may be helpful for further investigation.

Index Terms—COVID-19, detection of mutations, whole genome data, asymptomatic, collection time, age, gender, single-locus analysis, joint-SNPs analysis, linkage disequilibrium analysis

1 Introduction

In late December 2019, the coronavirus disease 2019 (COVID-19) outbreak was identified in Wuhan, China. Then the virus spreads rapidly in other countries resulting in a big global concern. The COVID-19 has caused 722,285 deaths and the number of infected patients is 19,462,112 by August 9, 2020.[1] Furthermore, the reproductive number of COVID-19 is higher compared to severe acute respiratory syndrome (SARS) coronavirus [1], which indicates that it is difficult to control the spread of the COVID-19, especially when the asymptomatic cases exist.

The asymptomatic patients have no clinical symptoms including fever, cough, fatigue, poor appetite, diarrhea, and headache at admission [2]. Some might also show normal CT image [3]. As a result, asymptomatic carriers can only be identified by laboratory to be positive for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) through nucleic acid testing. It has been shown that the asymptomatic cases can also be transmission resources, especially in family clusters [3], [4], [5]. A broadened SARS-CoV-2 testing including asymptomatic persons has been argued for effective epidemic prevention [6]. Therefore, it is critically important to investigate phenotypic factors associated with COVID-19.

1. https://covid19.who.int/

• *The authors are with the Center for Cognitive Machines and Computational Health (CMaCH), Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: {lvjinxiong, tushikui, leixu}@sjtu.edu.cn.*

Manuscript received 15 May 2020; revised 30 Nov. 2020; accepted 3 Jan. 2021. Date of publication 8 Jan. 2021; date of current version 6 Aug. 2021. (Corresponding authors: Lei Xu and Shikui Tu.) Digital Object Identifier no. 10.1109/TCBB.2021.3049836

In addition to the phenotype of whether displaying symptoms, there is a clear difference in incidence rate between male and female per 100,000 people ($P < 0.001$) in COVID-19 [7]. The adjusted case fatality rates (CFR) in male patients (4.45 percent) is three times more than that in female patients (1.25 percent) [7]. Also, various age groups have different clinical features. The COVID-19 more severely affects older patients with comorbidities [8]. Patients who are 60 years or older suffer from a much more excessive adjusted CFR of 5.30 percent [7]. Moreover, the SARS-COV-2 evolves over time and the time can be regarded as a phenotype as well. The collection time for SARS-COV-2 refers to the time when the samples (swab, sputum and so on) are collected. Although the above studies have been made on phenotypes, however, the existing phenotype-related works are conducted from the perspective of clinical research, not on the genetic mutations.

Existing efforts on genetic mutations of SARS-COV-2 have been made from the aspect of evolution for SARS-COV-2. After construction of phylogenetic tree which describes the path through evolutionary time from a common ancestor to different descendants, all of sequences are divided into two categories based on the single nucleotide polymorphisms (SNPs) positions 8517 and 27641 [9]. Specifically, sequences in group 1 have thymine at 8517 and cytosine at 27641 while the sequences in group 2 have cytosine at 8517 and thymine at 27641. The unique SNPs in nsp13, nsp14, nsp15, nsp16 (present in ORF1b polyprotein region) and S-Protein were identified in 10 American sequences [10]. The SNPs at location 8782 (orf1ab: T8517C, synonymous) and 28144 (ORF8: C251T, S84L) show significant linkage disequilibrium and they can be utilized to classify the 103 sequences into two groups, i.e., S group and L group [11], where the L type evolves from the S type. To our best knowledge, there are still lack of investigations on phenotype-related mutations of SARS-COV-2 based on genomic data.

In this paper, we present a combined single-locus analysis and joint-SNP analysis to detect significant SNPs associated with four phenotypes, which are the symptomatic status versus the asymptomaic status, the early collection time versus the late collection time, the elder versus the young, and the male versus the female. We collected 101 genomic sequences of the SARS-COV-2, and the locations of the hosts are all in Japan. After quality control, we conducted SNP calling procedure. The collection time and age were binarized according to their median values. To be able to analyze the SNPs of low minor allele frequency (MAF $\leq$ 0.05), we also employed the joint-SNPs analysis method by considering SNPs within a gene as a whole. The P values were corrected for multiple testings according to the Benjamini and Hochberg (BH) procedure [12]. Our analysis provides a list of significant SNPs for the four phenotypes, and a part of the list are related to the virus supported by the literature. Specially, the SNP locating at the nucleotide position 4321 is associated with all the first three phenotypes and there are 12 significant SNPs related to the first two phenotypes. In the joint-SNVs analysis, We find that gene orf1ab containing SNP-4321 can be detected to be significantly associated with the first three phenotypes, and the three genes including S, ORF3a, and N are detected significant in first two phenotypes. Furthermore, linkage disequilibrium was done for the correlation between the SNPs, and mutations of all SNPs over collection time were visualized to show the cumulative patterns of SNPs over time.

2 Material and Methods

2.1 Datasets

The whole-genome sequence data for SARS-COV-2 in Japan comes from GISAID database[2] and we collected 101 sequences published on April 6, 2020. According to the submission information, some sequences contain gaps or unknown nucleotides (denoted by NNNs) when comparing to the reference sequence. For quality control, the sequences with gaps or 1.18 percent NNNs were discarded, and 95 sequences were left. Of the 95 sequences, 67 sequences come from Diamond Princess cruise ship and 8 sequences belong to the Nile River cruise ship. The length of sequences ranges from 29,632 to 29,903. The reference sequence (NC_045512) comes from Genbank[3] and its length is 29,903. The details of the 95 sequences are summarized in Table 1.

TABLE 1
Detailed Information of SARS-COV-2-Related Sequence Data

Case Type		Num/Time
Symptom	Symptomatic	28
	Asymptomatic	58
	Unknown	9
Collection Time	Start	2020/1/23
	End	2020/3/20
	Median	2020/2/16
Age	$\leq$ 60yrs	12
	$>$ 60yrs	8
	Unknown	75
Gender	Male	7
	Female	13
	Unknown	75

2.2 SNP Calling Process

We utilized the bwa software (*bwa mem*) [13] to conduct the alignment of 95 sequences using NC_045512 as the reference. Subsequently, the samtools software [14] was used to convert the sam file into bam file, which is input file format for bcftools [14]. We used *bcftools mpileup* to conduct SNP calling and obtained 76 SNPs. The Indels and mutations on multiallelic were not taken into account and 74 SNPs were left. Moreover, 4 SNPs with missing value were not considered. Finally, the output file of bcftools was converted into Plink format by vcftools for the following SNP analysis [15]. These SNPs were assigned identifications based on basic position. For instance, SNP which locates at position 1234 is named after SNP-1234.

2. https://platform.gisaid.org/epi3/frontend#56d9bc
3. https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs

2.3 SNP Analysis

We utilized the Plink software [16] to conduct the quality control and single locus analysis for SNP. The SNPs which have missing values were filtered out and 70 high-quality SNPs were left. Then, the case-control association analysis was performed by Plink (*–assoc*). This command is based on contingency table tests. We assume that the number of minor alleles in case population is a, the number of minor alleles in control population is b while c and d is for major allele in case and control population, respectively. The corresponding Chi-square statistic is:

$$\chi^2 = \frac{n(ad-bc)}{(a+b)(a+c)(b+d)(c+d)}, \tag{1}$$

where the $n = a+b+c+d$ and the statistic χ^2 under the null hypothesis follows a Chi-square distribution with one degree of freedom. The linkage disequilibrium (LD) plot representing the linear relationship between any two of SNPs was produced by Haploview software [17].

Also, we conducted joint-SNPs analysis to take into account the existence of rare variants. As is known, in univariate hypothesis tests, a test statistic s is derived from the sample to quantify, within observed data, behaviours that would distinguish the null hypothesis H_0 from the alternative one. P value is defined as the probability of s equal to or more extreme than $\tilde{s}$ which is the evaluation of s at the observed data. Different from univariate hypothesis tests, the multivariate hypothesis tests extend the scale statistic s into a vector statistic $\mathbf{s}$. Recently proposed in [18], the boundary-based test (BBT) is an effective multivariate test method by testing whether a separable plane is existed between the case population and control population. The separating plane can be sought in the original data space by classification methods, or the boundary between rejection region and acceptance region is ascertained after calculating the statistics from the original data space. The latter one, called statistics-space BBT or S-BBT, has been shown to achieve the strongest detection power compared with other related methods on the simulation and real-world data in

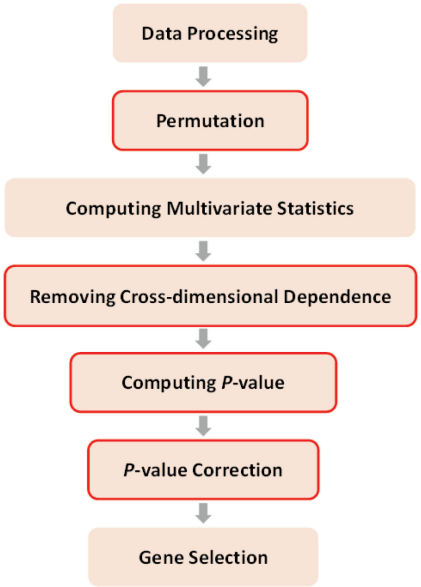


Fig. 1. The flow chart of statistic-space boundary-based test.

our previous works [19], [20], [21]. The advantages of S-BBT come from tackling two limitations in existing methods, i.e., no considerations on either the relationship between the dimensions or the direction for each component [18], [21]. Thus, we adopt S-BBT as our joint-SNPs analysis method. An overview of the S-BBT pipeline was given in Fig. 1.

It consists of four major components [18]:

- The test statistic $\tilde{\mathbf{s}}$ is computed from the observed data, and then used to determine the rejection domain $\Gamma(\tilde{\mathbf{s}})$ in the statistics space as follows:

$$\Gamma(\tilde{\mathbf{s}}) = \{\mathbf{s} : (\mathbf{s} - \tilde{\mathbf{s}})^T \mathbf{sign}(\tilde{\mathbf{s}}) > \mathbf{0}\}, \quad (2)$$

  where the $\mathbf{sign}(\mathbf{s}) = [sign[s_1], \ldots, sign[s_m]]^T$ with $sign[u] = \frac{u}{|u|}$ and m is the the number of dimensions.
- The P values are calculated by the permutation test as given by Eq. (65) in [18]. Here, the number of permutation times is set to 1000.
- Principal components analysis (PCA) is employed to remove the second order cross-dimensional dependence among the vector $\tilde{\mathbf{s}}$ so that the lattice taxonomy of tests can be conducted with different dimensions of rejection for computing P values [22].
- The posterior version of P values is computed for reducing background disturbance and it can be defined as the probability that the false alarm rate of randomly permuted samples is equal to or smaller than the one on the original samples (see Eq. (90)$\sim$(93) in [18]).

The S-BBT can be regarded as implementing the multivariate hypothesis test via an integration of multiple univariate hypothesis tests so that different dimensions and their combinations are taken into account. Moreover, both the scale and direction of the multivariate statistic $\tilde{\mathbf{s}}$ which indicates the difference between case and control population are also considered. With these benefits, S-BBT achieves the strongest detection power compared with other related methods.

TABLE 2
Results of Single Locus Analysis in the Symptomatic versus Asymptomatic Study

SNP ID	Position	Minor	Major	P	P_{BH}
SNP-11083	11083	G	T	6.56E-22	4.59E-20
SNP-241	241	T	C	1.81E-19	6.34E-18
SNP-3037	3037	T	C	1.81E-19	4.22E-18
SNP-14408	14408	T	C	1.81E-19	3.17E-18
SNP-23403	23403	G	A	1.81E-19	2.53E-18
SNP-28881	28881	A	G	4.87E-13	5.68E-12
SNP-28882	28882	A	G	4.87E-13	4.87E-12
SNP-28883	28883	C	G	4.87E-13	4.26E-12
SNP-18877	18877	T	C	2.74E-06	2.13E-05
SNP-25563	25563	T	G	2.74E-06	1.92E-05
SNP-23248	23248	T	C	3.33E-04	2.12E-03
SNP-29635	29635	T	C	3.32E-03	1.94E-02
SNP-4321	4321	T	C	3.59E-03	1.93E-02
SNP-18656	18656	T	C	6.68E-03	3.34E-02
SNP-3025	3025	T	G	4.06E-02	1.89E-01
SNP-11557	11557	T	G	4.06E-02	1.78E-01
SNP-15324	15324	T	C	4.06E-02	1.67E-01
SNP-15907	15907	A	G	4.06E-02	1.58E-01
SNP-25810	25810	G	C	4.06E-02	1.50E-01
SNP-26642	26642	T	C	4.06E-02	1.42E-01
SNP-29227	29227	T	G	4.06E-02	1.35E-01
SNP-29229	29229	A	G	4.06E-02	1.29E-01
SNP-29303	29303	T	C	4.06E-02	1.24E-01

3 Results

3.1 Single Locus Analysis

In this section, we take the symptom, collection time, age and gender into consideration for the case-control study and detect significant SNPs. The threshold for significance is set to 0.05 and we conduct the Benjamini and Hochberg (BH) correction [12] for reference.

The existence of asymptomatic patients shows that the clinical symptoms and radiological abnormalities are not the essential components of SARS-CoV-2 infection, resulting in difficulty in the detection and isolation of infected cases [23]. Here, we intend to uncover the SNPs which differentiates the symptomatic or the asymptomatic. The results were shown in Table 2.

There are 23 significant SNPs ($P \leq 0.05$), and 14 out of 23 passes the BH correction ($P_{BH} \leq 0.05$). The SNP-11083 (T $\rightarrow$ G) which locates at 11083 has the smallest P value. The SNP-241, SNP-3037 and SNP-14408 share common mutation patterns (C $\rightarrow$ T) and their P values are 1.81E-19. The P value of SNP-23403 is also 1.81E-19.

The collection time indicates time of onset for patients to some degree. The sample EPI_ISL_414511 does not have accurate collection date and we filter it out. The median date is February 16, 2020. The number of samples collected until that is 66, and the number is 28 after that. The results were shown in Table 3. There are 16 significant SNPs ($P \leq 0.05$), 14 out of 16 are still significant after BH correction for multiple testings ($P_{BH} \leq 0.05$). Four SNPs, including SNP-241, SNP-3037, SNP-14408, and SNP-23403, have the smallest P value. Of them, the SNP-241, SNP-3037 and SNP-14408 share common mutation patterns (C $\rightarrow$ T) while $A \rightarrow G$ for SNP-23403. The SNP-28881 and SNP-28882 share common mutation patterns and their P values are equal to 1.90E-07.

TABLE 3
Results of Single Locus Analysis Before and After 2020/2/16

SNP ID	Position	Minor	Major	P	P_{BH}
SNP-241	241	T	C	3.33E-12	2.33E-10
SNP-3037	3037	T	C	3.33E-12	1.17E-10
SNP-14408	14408	T	C	3.33E-12	7.77E-11
SNP-23403	23403	G	A	3.33E-12	5.83E-11
SNP-28881	28881	A	G	1.90E-07	2.66E-06
SNP-28882	28882	A	G	1.90E-07	2.22E-06
SNP-28883	28883	C	G	1.90E-07	1.90E-06
SNP-11083	11083	G	T	6.71E-05	5.87E-04
SNP-18877	18877	T	C	6.96E-04	5.41E-03
SNP-25563	25563	T	G	6.96E-04	4.87E-03
SNP-2662	2662	T	C	5.60E-03	3.56E-02
SNP-28144	28144	C	T	5.60E-03	3.27E-02
SNP-29095	29095	T	C	5.60E-03	3.02E-02
SNP-23248	23248	T	C	9.38E-03	4.69E-02
SNP-4321	4321	T	C	3.49E-02	1.63E-01
SNP-5845	5845	T	A	3.49E-02	1.53E-01

Both SNP-23403 and SNP-23248 can be detected in considering symptom and collection time and they locate in Spike (S) gene. Point mutations in a murine coronavirus spike protein can result in increased virulence through instability of the viral machinery and alter viral to cell membrane fusion [24]. So both of them might have effects on virulence of the virus. The SNP-23403 encodes the SARS-CoV-2 S protein variant D614G which increases the infectivity [25], [26]. Furthermore, the G614 variant is more resistant to cleavage in vitro and in human cells [27].

Age is one basic factor related to the death of COVID-19 [7], [28], and we detect SNPs significantly associated with age by dividing patients into two groups, i.e., the young and the old. Of the 95 collected sequences, the age information of 75 sequences is missing so that the size of the remained sample is 20. Since the sample size is small, only one SNP which locates at 4321 stands out to be significant ($P \leq 0.05$). Extra data may be needed to verify its significance, because it is not significant after the BH correction. Table 4 lists the top 10 SNPs.

As given in [7], gender is related to the fatality rate and incidence rate by COVID-19. The adjusted case fatality rate of men is three times higher than that of women. Also, the incidence rate of males is higher than that of females. Among the collected 95 sequences, there are 20 sequences

TABLE 4
Top 10 SNPs of Single Locus Analysis Between the Young and the Old

SNP ID	Position	Minor	Major	P	P_{BH}
SNP-4321	4321	T	C	0.01977	1.00E+00
SNP-3025	3025	T	G	0.1087	1.00E+00
SNP-11083	11083	G	T	0.1087	1.00E+00
SNP-26642	26642	T	C	0.1087	1.00E+00
SNP-29229	29229	A	G	0.1087	1.00E+00
SNP-29635	29635	T	C	0.1087	1.00E+00
SNP-11557	11557	T	G	0.1894	1.00E+00
SNP-15324	15324	T	C	0.1894	1.00E+00
SNP-25810	25810	G	C	0.1894	1.00E+00
SNP-29303	29303	T	C	0.1894	1.00E+00

TABLE 5
Top 10 SNPs of Single Locus Analysis Between the Male and the Female

SNP ID	Position	Minor	Major	χ^2	P
SNP-3025	3025	T	G	1.134	0.287
SNP-11083	11083	G	T	1.134	0.287
SNP-11557	11557	T	G	1.134	0.287
SNP-15324	15324	T	C	1.134	0.287
SNP-25810	25810	G	C	1.134	0.287
SNP-26642	26642	T	C	1.134	0.287
SNP-29229	29229	A	G	1.134	0.287
SNP-29303	29303	T	C	1.134	0.287
SNP-29635	29635	T	C	1.134	0.287
SNP-241	241	T	C	0.4396	0.5073

containing information about gender. We conducted single locus analysis about gender and no significant SNP was detected ($P \leq 0.05$). Still, we list the top 10 SNPs in Table 5 for future reference.

It might be attributed to two reasons. First, most sequences (75 out of 95) have missing value on gender. Second, sequence diversity of SARS-COV-2 is not associated with gender of hosts.

According to the results in Tables 2, 3, and 4, several SNPs are significant in more than one phenotypes. As demonstrated in the Venn plot by Fig. 2, 12 SNPs are significantly associated with both asymptomatic status and the sample collection time, and 5 of 12 SNPs, i.e., SNP-241, SNP-3037 and SNP-14408 not only share common mutation patterns, but also SNP-28881 and SNP-28882. Especially, we find that the SNP-4321 is significantly associated with three phenotypes including symptom, collection time, and age, and it might deserve further experimental investigation. The SNP-4321 locates in NSP10 gene which can be translated into nsp10 (nonstructural proteins 10) protein. The nsp7-nsp10 are crucial cofactors of replicative enzymes and contribute to the emerging nsp interactome [29]. Furthermore, the nsp10 works as part of the coronavirus capping machinery [29].

3.2 Patterns of Mutations at the 70 SNPs Over the Sample Collection Time

The virus would evolve over time, and it is important to study the patterns of mutation frequencies at the 70 SNPs over the sample collection time. We calculate the numbers of mutations occurred at the 70 SNPs, and visualize the dynamic patterns of the average number of mutations along with the collection time in Fig. 3.

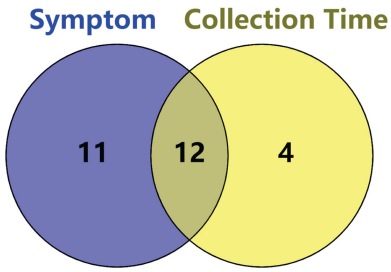


Fig. 2. The Venn plot of significant SNPs associated with asymptomatic status and sample collection time.

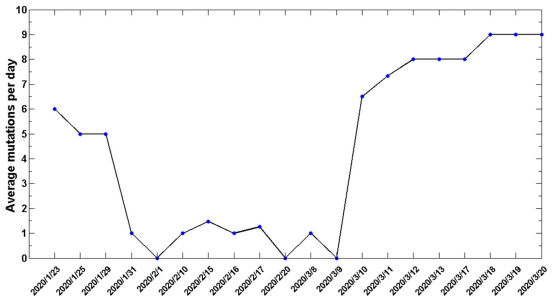


Fig. 3. The mean of the times of mutations at 70 SNPs over time.

It can be observed that the average number of mutations first decreases to zero in Feb. 01, 2020, stays at a low level, and then increases to a level higher than ever before from Mar. 09, 2020. Such sudden growth deserves further investigation on the COVID-19 pandemic at that time.

Furthermore, we show the cumulative numbers of mutations at SNPs over the sample collection time by a heatmap in Fig. 4.

Only the SNPs with more than one mutation are shown, and the SNPs are sorted in descending order according to their accumulative number of mutations. The SNP-11083 has the most number of mutations, while the cumulative distributions of mutations at SNP-241, SNP-3037, SNP-14408 and SNP-23403, are the same and equal to 17 finally. The SNP-29635 obtains 14 mutations on Feb. 17, 2020 and stays at the same level after that time. On February 10, 2020, the first two patients denoted by EPI_ISL_412968 and EPI_ISL_412969 obtained one mutation on SNP-29635, respectively. Then, on February 15, 2020, one symptomatic patient (EPI_ISL_416569) of the 19 collected samples was found to have one mutation on SNP-29635. On February 16, 2020, 5 out of 21 samples have mutations on SNP-29635 and all of five are asymptomatic patients. On February 17, 2020, 6 asymptomatic patients were found to have mutations. Altogether, it could be noticed that the patients with mutations for SNP-29635 are mostly asymptomatic. The SNP-28881, SNP-28882 and SNP-28883 do not have any mutations until Mar. 12, 2020 and accumulate mutations to 10 in the same way.

3.3 Linkage Disequilibrium Analysis

In this section, the correlation of 70 SNPs is shown in Fig. 5.

The SNP-4321 has a weak correlation with other SNPs ($R^2 \leq 0.38$), which indicates that it is an independent locus associated with asymptomatic status, collection time and age. There is strong correlation between SNP-8782 and other three SNPs including SNP-2662, SNP-28144, SNP-29095 ($R^2 = 0.79$). It is not a surprise to see that the correlation coefficients of SNPs which share the same patterns of accumulative number of mutations are equal to one. They are (SNP-241, SNP-3037, SNP-14408, SNP-23403), (SNP-28881, SNP-28882, SNP-28883), (SNP-254, SNP-22104), (SNP-973, SNP-9141, SNP-26062), (SNP-2662, SNP-28144, SNP-29095), (SNP-3025, SNP-26642), (SNP-4201, SNP-25000, SNP-25244), (SNP-8004, SNP-21917), (SNP-8139, SNP-26211), (SNP-8323, SNP-25658), (SNP-9141, SNP-26062), (SNP-11310, SNP-24819, SNP-29674), (SNP-12964, SNP-29324), (SNP-15324, SNP-25810, SNP-29303), (SNP-18877, SNP-25563), (SNP-20762, SNP-24797, SNP-29463), (SNP-21575, SNP-29592). Among them, the distance between SNP-2662 and SNP-29095 is equal to 26,433 bp while the length of SARS-COV-2 virus is 29,903.

In order to figure out whether these highly correlated and similar patterns of mutations are involved in symptomatic or asymptomatic patients, correlation coefficients of them were also calculated for the two populations. On the one hand, the (SNP-214, SNP-3037, SNP-14408, SNP-23403), (SNP-3025, SNP-26642), (SNP-18877, SNP-25563) and (SNP-28881, SNP-28882, SNP-28883) are involved in symptomatic infection in patients. On the other hand, the (SNP-254, SNP-22104), (SNP-2662, SNP-28144, SNP-29095), (SNP-4201, SNP-25000, SNP-25244), (SNP-8004, SNP-21917), (SNP-8139, SNP-26211), (SNP-8323, SNP-25658), (SNP-11310, SNP-24819, SNP-29674), (SNP-12964, SNP-29324), (SNP-20762, SNP-24797, SNP-29463) and (SNP-21575, SNP-29592) are involved in asymptomatic patients.

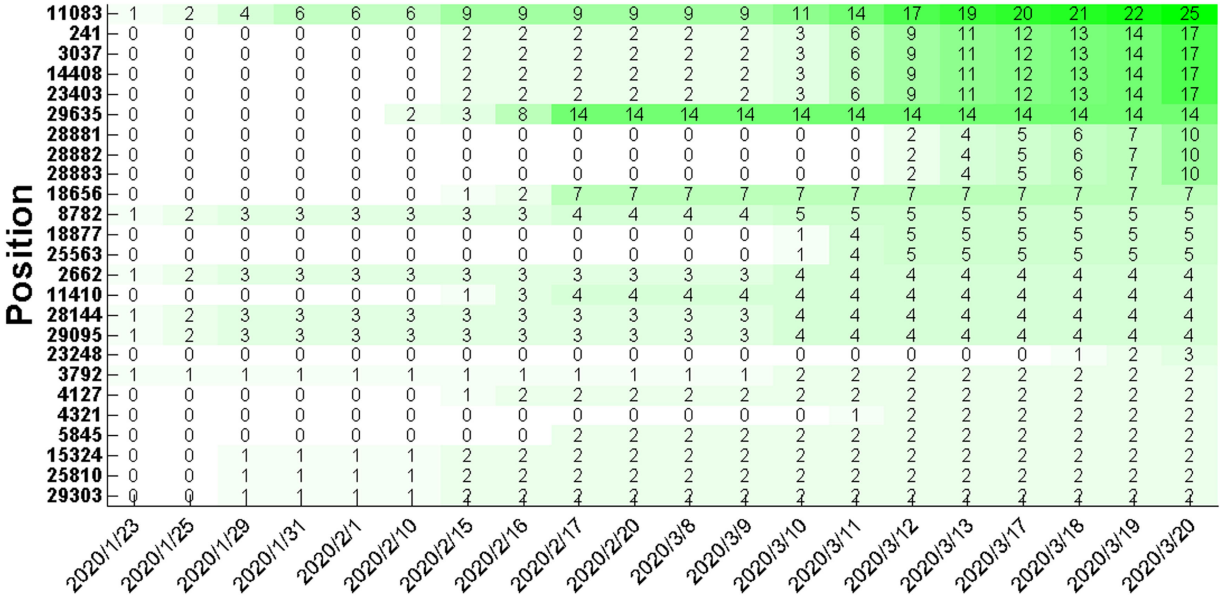


Fig. 4. The cumulative number of mutations over time.

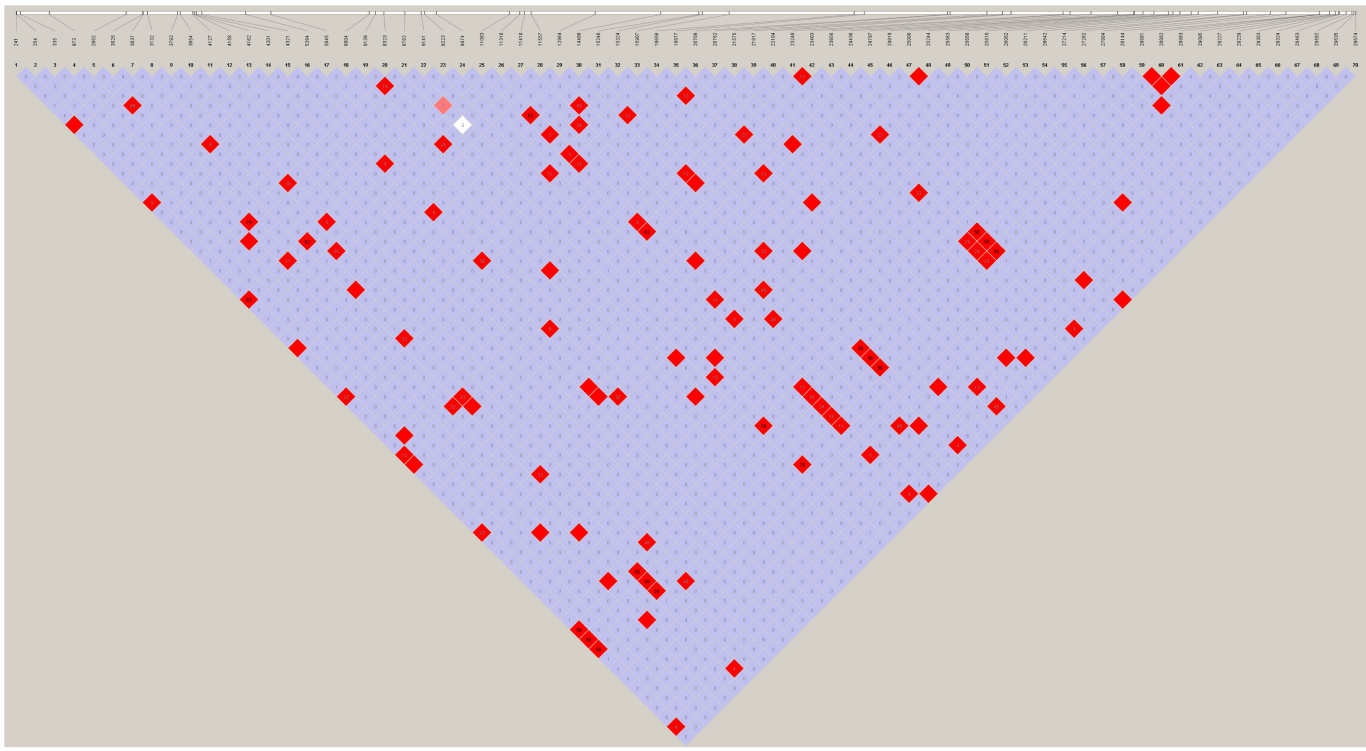


Fig. 5. LD plot for 70 SNPs. Correlation coefficient R^2 of any two of 70 SNPs are shown and the values equals R^2 times 100.

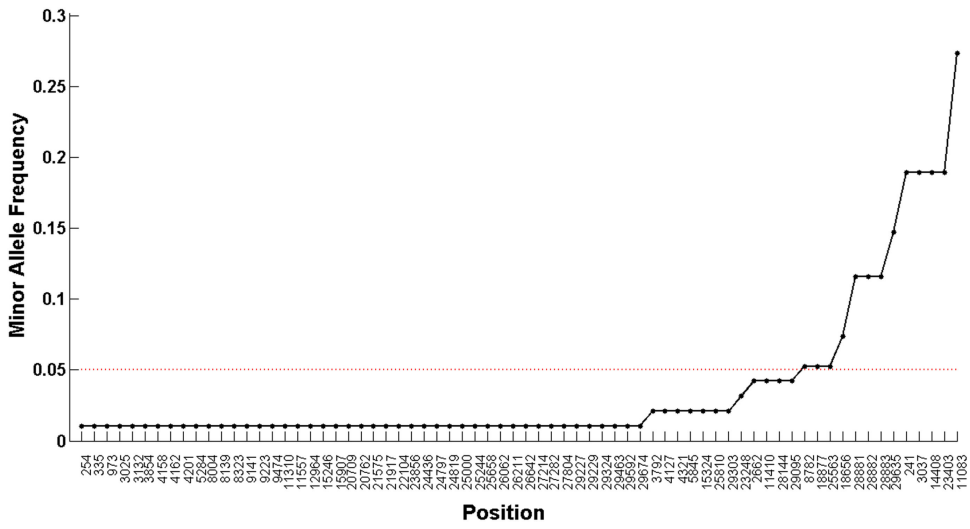


Fig. 6. The minor allele frequency of 70 SNPs.

3.4 Joint-SNPs Analysis

As shown by Fig. 6, the majority of 70 SNPs are rare variants with Minor Allele Frequency (MAF) ≤ 0.05.

Interesting rare variants have difficulty in being detected by single-locus analysis owing to the insufficient sample size. Instead, the joint-SNPs analysis combines SNPs which locate in same unit (e.g., gene, exon and so on) as one computation unit, resulting in higher detection power [30]. Here, we combine multiple SNPs in same genes, and conduct the joint-SNPs analysis by S-BBT under four kinds of phenotypes. The P values calculated by S-BBT and corresponding BH-adjusted P values were given in Table 6.

It can be observed that the gene orf1ab containing SNP-4321 is detected to be associated with three phenotypes, i.e., asymptomatic status, collection time, and gender. In considering BH-adjusted P values of orf1ab, it can be detected in two phenotypes including symptom and collection time while it cannot be detected in collection time by single locus analysis. The gene orf1ab encodes 16 nonstructural proteins (Nsps) which promote cellular mRNA degradation and block host cell translation [31]. It can also be noticed that three genes, i.e., S, ORF3a and N, achieved significant P values in considering of different levels of symptom and collection time. It should be noted that the spike (S) protein encoded by S gene is responsible for the distinctive spike

TABLE 6
Results of Joint-SNPs Anlaysis by S-BBT

Gene Symbol	Position (start:end)	$P_{Symptom}$	$P_{Collection\ time}$	P_{Gender}	P_{Age}
5′UTR	1:265	3.43E-12(8.00E-11)	3.74E-09(2.62E-07)	4.61E-01(1.00E+00)	1.00E+00(1.00E+00)
orf1ab	266:21555	6.51E-14(4.56E-12)	5.17E-07(1.21E-05)	2.49E-03(1.74E-01)	1.00E+00(1.00E+00)
S	21563:25384	5.60E-13(1.96E-11)	1.53E-08(5.36E-07)	5.90E-01(1.00E+00)	1.00E+00(1.00E+00)
ORF3a	25393:26220	1.39E-05(1.95E-04)	3.17E-04(4.44E-03)	3.07E-01(1.00E+00)	1.00E+00(1.00E+00)
E	26245:26472	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)
M	26523:27191	1.27E-01(1.00E+00)	4.67E-01(1.00E+00)	2.30E-01(1.00E+00)	1.00E+00(1.00E+00)
ORF6	26523:27191	1.00E+00(1.00E+00)	1.91E-01(1.00E+00)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)
ORF7a	27202:27387	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)
ORF7b	27394:27759	1.00E+00(1.00E+00)	4.42E-01(1.00E+00)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)
ORF8	27894:28259	1.00E+00(1.00E+00)	8.98E-03(1.05E-01)	1.00E+00(1.00E+00)	1.00E+00(1.00E+00)
N	28274:29533	2.35E-09(4.11E-08)	2.56E-06(4.48E-05)	8.10E-01(1.00E+00)	6.37E-01(1.00E+00)
ORF10	29558:29674	7.97E-03(9.30E-02)	7.06E-01(1.00E+00)	2.48E-01(1.00E+00)	1.00E+00(1.00E+00)

structure on the surface of SARS-COV-2 [31]. The S gene is the most variable region of the genome [32]. Moreover, the ORF3a protein is a hypothetical protein which is similar to SARS3a protein in SARS-COV and it plays an important role in mediating immune evasion and favoring viral spread [33]. Gene N can be translated into nucleocapsid protein (N) which can cause inflammation of the lungs in SARS-COV [34]. The 5′ UTR region with the SNP-241 is significant in considering of symptom and collection time.

4 Conclusion and Discussion

In this paper, we have presented an investigation on phenotype-related mutations of SARS-COV-2 via not only single-locus analysis but also joint-SNPs analysis to improve the statistical detection power. Four phenotypes were considered, including asymptomatic status, sample collection time, age, and gender. We have used 95 public virus genome sequences collected from hosts in Japan, and called 70 SNPs when comparing to the reference genome. Through the SNP analysis, we find 23 SNPs that are significantly associated with the asymptomatic status which is one of major difficult factors to control the spread of COVID-19. Also, there are 16 SNPs significantly associated with sample collection time, and one SNP is associated with age. No significant SNPs can be detected related to gender. Indicated by linkage disequilibrium analysis, the SNP-4321 is an independent SNP associated with all three phenotypes, and gene orf1ab, which contains SNP-4321 and promotes celluar mRNA degradation and blocks host cell translation, can be further detected by joint-SNPs analysis. Three genes, S, ORF3a, and N, are found to be associated with the asymtomatic status and the sample collection time. Literature evidences suggest that the three genes are related to SARS-COV-2, indicating that our analysis is reliable and may provide helpful candidate SNPs or genes for future experimental study on COVID-19. Our study on phenotype-related mutations for COVID-19 is just a beginning, and further investigations can be considered in the future. First, we only collect 101 samples for reducing the impact of ethnic group on calculated results. It is expected that the calculated results would be more reliable when more samples are available. Second, only four phenotypes are considered in this study and we consider more phenotypes for better understanding of the pathogenicity of SARS-COV-2. Finally, multiple omics data can be integrated for reliable results.

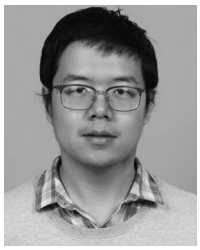
Acknowledgments

This work was supported by National Science and Technology Innovation 2030 Major Project (2018AAA0100700) of the Ministry of Science and Technology of China, and the National Natural Science Foundation of China (NSFC 61802256), as well as ZhiYuan Chair Professorship Start-up Grant (WF220103010) from Shanghai Jiao Tong University.

References

[1] Y. Liu, A. A. Gayle, A. Wildersmith, and J. Rocklov, "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *J. Travel Med.*, vol. 27, no. 2, pp. 1–4, 2020.
[2] Y. Wang, Y. Liu, L. Liu, X. Wang, N. Luo, and L. Li, "Clinical outcomes in 55 patients with severe acute respiratory syndrome Coronavirus 2 who were asymptomatic at hospital admission in shenzhen, China," *J. Infect. Dis.*, vol. 221, pp. 1770–1774, 2020.
[3] Z. Hu *et al.*, "Clinical characteristics of 24 asymptomatic infections with COVID-19 screened among close contacts in nanjing, china," *Sci. China Life Sci.*, vol. 63, pp. 706–711, 2020.
[4] X. Pan *et al.*, "Asymptomatic cases in a family cluster with SARS-CoV-2 infection," *Lancet Infect. Dis.*, vol. 20, pp. 410–411, 2020.
[5] J. F. W. Chan *et al.*, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster," *Lancet*, vol. 395, no. 10223, pp. 514–523, 2020.
[6] M. Gandhi, D. S. Yokoe, and D. V. Havlir, "Asymptomatic transmission, the achilles' heel of current strategies to control COVID-19," *New Eng. J. Med.*, vol. 382, pp. 2158–2160, 2020.
[7] Y. Yang *et al.*, "Epidemiological and clinical features of the 2019 novel coronavirus outbreak in china," *medRxiv*, 2020.
[8] N. Zhang *et al.*, "Recent advances in the detection of respiratory virus infection in humans," *J. Med. Virol.*, vol. 92, no. 4, pp. 408–417, 2020.
[9] A. Wu *et al.*, "Mutations, recombination and insertion in the evolution of 2019-nCoV," *bioRxiv*, 2020.
[10] R. Kumar *et al.*, "Comparative genomic analysis of rapidly evolving SARS CoV-2 viruses reveal mosaic pattern of phylogeographical distribution," *bioRxiv*, 2020.
[11] X. Tang *et al.*, "On the origin and continuing evolution of SARS-CoV-2," *Nat. Sci. Rev.*, vol. 7, pp. 1012–1023, 2020.
[12] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Series B-Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.
[13] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[14] H. Li *et al.*, "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.
[15] P. Danecek *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, 2011.
[16] S. Purcell *et al.*, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Amer. J. Hum. Genet.*, vol. 81, no. 13, pp. 559–575, 2011.
[17] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: Analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, no. 2, pp. 263–265, 2005.
[18] L. Xu, "Bi-linear matrix-variate analyses, integrative hypothesis tests, and case-control studies," *Appl. Inform.*, vol. 2, no. 1, 2015, Art. no. 4.
[19] J. Lv, H. Huang, R. Chen, and L. Xu, "A comparison study on multivariate methods for joint-SNVs association analysis," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2016, pp. 1771–1776.
[20] J. Lv, H. Huang, R. Chen, and L. Xu, "Comparative studies on multivariate tests for joint-SNVs analysis and detection for bipolar disorder susceptibility genes," *Int. J. Data Mining Bioinf.*, vol. 17, no. 4, pp. 341–358, 2017.
[21] J. Lv, S. Tu, and L. Xu, "A comparative study of joint-SNVs analysis methods and detection of susceptibility genes for gastric cancer in korean population," in *Proc. 7th Int. Conf. Intell. Sci. Big Data Eng.*, 2017, pp. 619–630.
[22] L. Xu, "A new multivariate test formulation: Theory, implementation, and applications to genomescale sequencing and expression," *Appl. Inform.*, vol. 3, no. 1, pp. 1–23, 2016.
[23] Z. Ling *et al.*, "Asymptomatic SARs-CoV-2 infected patients with persistent negative ct findings," *Eur. J. Radiol.*, vol. 126, pp. 108 956–108 956, 2020.
[24] A. Brufsky, "Distinct viral clades of SARS-CoV-2: Implications for modeling of viral spread," *J. Med. Virol.*, vol. 92, pp. 1386–1390, 2020.
[25] L. Yurkovetskiy *et al.*, "Structural and functional analysis of the D614G SARS-CoV-2 Spike protein variant," *Cell*, vol. 183, no. 3, pp. 739–751, 2020.
[26] L. Zhang *et al.*, "SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity," *Nat. Commun.*, vol. 11, no. 6013, pp. 1–9, 2020.
[27] Z. Daniloski, X. Guo, and N. E. Sanjana, "The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types," *bioRxiv*, 2020.
[28] W.-J. Guan *et al.*, "Clinical characteristics of 2019 novel coronavirus infection in china," *medRxiv*, 2020.
[29] E. Snijder, E. Decroly, and J. Ziebuhr, "Chapter three - the nonstructural proteins directing coronavirus RNA synthesis and processing," in *Coronaviruses*, J. Ziebuhr, Ed. Cambridge, MA, USA: Academic Press, 2016, pp. 59–126.
[30] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin, "Rare-variant association analysis: Study designs and statistical tests," *Amer. J. Hum. Genet.*, vol. 95, no. 1, pp. 5–23, 2014.
[31] A. R. Fehr and S. Perlman, "Coronaviruses: An overview of their replication and pathogenesis," *Methods Mol. Biol.*, vol. 1282, pp. 1–23, 2015.
[32] P. Zhou *et al.*, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, pp. 270–273, 2020.
[33] E. Issa, G. Merhi, B. Panossian, T. Salloum, and S. Tokajian, "SARS-CoV-2 and ORF3a: Non-synonymous mutations and polyproline regions," *bioRxiv*, 2020.
[34] X. Yan, Q. Hao, Y. Mu, K. A. Timani, L. Ye, Y. Zhu, and J. Wu, "Nucleocapsid protein of SARS-CoV activates the expression of cyclooxygenase-2 by binding directly to regulatory elements for nuclear factor-kappa B and CCAAT/enhancer binding protein," *Int. J. Biochem. Cell Biol.*, vol. 38, no. 8, pp. 1417–1428, 2006.



Jinxiong Lv received the bachelor's degree in computer science and technology from the Sichuan University of China, in 2015. He is currently working toward the PhD degree in the Department of Computer Science and Engineering and is also a member of Center for Cognitive Machines and Computational Health (CMaCH) in Shanghai Jiao Tong University. His research interests include statistical learning and bioinformatics.

Shikui Tu received the BSc degree in applied mathematics from Peking University, in 2006, and the PhD degree in computer science from the Chinese University of Hong Kong, in 2012. He is a tenure-track associate professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University (SJTU), and also is the academic assistant to the head of Center for Cognitive Machines and Computational Health (CMaCH). Before he joined SJTU, he was a postdoctoral associate with UMass Worcester between Dec. 2012 and Jan. 2017. His research interests include machine learning, and bioinformatics. He has published more than 40 academic papers in top conferences and journals, including Science, Cell, NAR, etc. with high impact factors.

Lei Xu (Fellow, IEEE) Zhiyuan chair professor of Computer Science and Engineering Department, Shanghai Jiao Tong University (SJTU); director of SJTU SEIEE Centre for Cognitive Machines and Computational Health (CMaCH); Chief Scientist of SJTU AI Research Institute, Chief Scientist of SJTU-Sensetime Research Institute; Director of Neural Computation Research Centre in Brain and Intelligence Science-Technology Institute, Zhang Jiang Lab.; Emeritus professor of Computer Science and Engineering, Chinese University of Hong Kong. Conducted researches in several areas of Artificial Intelligence (AI) for more than 40 years, such as neural networks, machine learning, pattern recognition, bioinformatics, and computational finance. Published about 400 papers (including more than 130 Journal papers), given dozens keynote /invited lectures on various international conferences. His influential contributions on Randomized Hough Transform (RHT), RPCL learning, classifier combination, mixture of experts and EM algorithm, nonlinear Hebbian and ICA learning, LMSER bidirectional learning, and Bayesian Ying-Yang (BYY) harmony learned are well known and followed by many subsequent studies. Served as EIC and associate editors of several academic journals, e.g., including Neural Networks (1995–2016), the *IEEE Transactions Neural Networks* (1994–98). Taken various roles in academic societies, e.g., INNS Governing Board (2001–03), Fellow committee of the *IEEE Computational Intelligence Society* (2006–07), and the EURASC Scientific Committee (2014–17). Received several national and international academic awards, e.g., including 1993 National Nature Science Award, 1995 Leadership Award from International Neural Networks Society (INNS) and 2006 APNNA Outstanding Achievement Award. Elected to fellow of IEEE in 2001 of International Association for Pattern Recognition in 2002 and of European Academy of Sciences (EURASC) in 2003.