How fast does the SARS-Cov-2 virus really mutate in heterogeneous populations?

Vishal Rana, Eli Chien, Jianhao Peng and Olgica Milenkovic
Coordinated Science Lab and the Department of Electrical and Computer Engineering
University of Illinois, Urbana-Champaign, USA
(DRAFT)

Abstract

We introduce the problem of determining the *mutational support* of genes in the SARS-Cov-2 virus and estimating the distribution of mutations within different genes using small sample sizes that do not allow for accurate maximum likelihood estimation. The mutational support refers to the unknown number of sites mutated across all strains and individual samples of the SARS-Cov-2 genome; given the high cost and limited availability of real-time polymerase chain reaction (RT-PCR) test kits, especially in early stages of infections when only a small number of genomic samples ($\sim 1000s$) are available that do not allow for determining the exact degree of mutations in an RNA virus that comprises roughly $30,000$ nucleotides. Nevertheless, working with small sample sets is required in order to quickly predict the mutation rate of this and other viruses and get an insight into their transformational power. Furthermore, with the small number of samples available, it is hard to estimate the mutational landscape across different age/gender groups and geographical locations which may be of great importance in assessing different risk categories and factors influencing susceptibility to infection. To this end, we use our state-of-the art polynomial estimator techniques and the Good-Turing estimator to obtain estimates based on only roughly $1,000$ samples per category. Our analysis reveals an interesting finding: the mutational support appears to be statistically more significant in patients which appear to have lower infection rates and handle the exposure with milder symptoms, such as women and people of relatively young age (≤ 55).

1 Introduction

Viruses tend to mutate rapidly for a number of reasons, including highly unreliable replication of their genetic content and the need to evolve, adapt and compete with the host organism. The rate of mutation varies widely across various types of viruses and has been extensively studied in the past [1], [2]. It is known that RNA viruses tend to mutate faster than DNA viruses as RNA replication is much less accurate than DNA replication. Similarly, single stranded viruses are more likely to mutate than double stranded ones [3] due to their structural instabilities. There is also evidence to indicate the length of the viral genome is inversely correlated with the mutation rate, with shorter viruses mutating faster those having longer strands of genomic material [4]. Mutational and fitness landscapes of viruses are frequently used to determine their evolvability and potential to spread across diverse populations [5], [6], [7].

If the immune system of a host encounters viral protein from a strain it was already exposed to, its response is fast and the infected cells are efficiently eliminated. If the virus mutates at a very high rate, the host immune system may take longer to respond, giving the virus more time to replicate and spread. This phenomenon is known as *antigenic drift* [8]. It is hence widely believed that fast mutating viruses pose a greater health risk as they provide an escape mechanism not countered by the host. Nevertheless, some recent studies have shown that high mutation rates could also trigger rapid innate immune response in the host; they can also be a sign that the host is successfully fending of the infection and that the virus has to explore a significant number of changes to its genome to successfully compete with the immune system. Elevated levels of mutations can be disadvantageous to the survival of virus, at least at short time scales [9]. It hence remains an open problem to determine the exact causes for elevated mutation rates in some viruses and their correlation to clinical patient outcomes.

Despite the fact that all potential sources for viral mutations are still unknown, a large body of works reports mutational rates of viruses as indicators of their virulence and potential to cause epidemic and pandemic outbreaks [10, 11]. Almost exclusively, the estimates are based on simple counts of mutations in sequenced genomes, using a reference retrieved either from Patient 0 (the first infected individual) or more frequently, from Patient 1 (the first sequenced individual). Given the very limited number of samples (i.e., sequenced genomes) compared to the length of the genomes

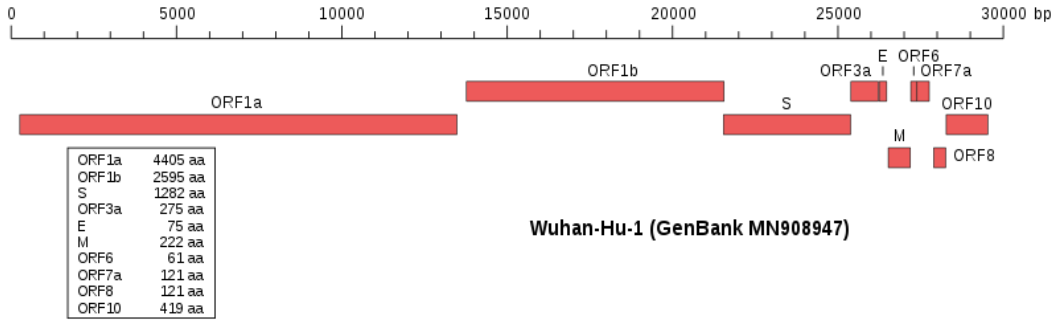


Figure 1: Organization of the SARS-Cov-2 genome, taken from Wikipedia.

(ranging from several $1,000$s to $100,000$ nucleotides) it is apparent that naive counts and the corresponding maximum likelihood estimators are inadequate for this purpose. This small sample effect is well known and extensively studied in the machine learning community [12, 13]. The distribution of the mutations and the mutational support have not been studied from an estimation perspective. In this work, we present methods for determining support of mutations and their distribution given sequencing data from a limited number of patients.

The problem of mutational support and distribution is of independent interest for future outbreaks as well. It is important to be able to come up with a mutational landscape with limited number of samples available during the very early phases of the outbreak.

2 Data and Methods

Organization of the SARS-Cov-2 genome. A breakdown of the genomic structure of SARS-Cov-2 is shown in Figure 1, and described in detail in [14]. Typically, coronaviruses have genomes including at least six open reading frames (ORFs). ORF1a and ORF1b constitute the longest component of the genomes are responsible for encoding two polypeptides, pp1a and pp1ab, which are jointly used to create a family of nsp proteins. The family includes replicase-transcriptase proteins, responsible for promoting cellular mRNA degradation and blocking the translation in host cells, thereby impairing the operation of the immune response, proofreading and scaffolding proteins, processivity clamps, as well as transmembrane proteins. The pp1a/b polypeptides are functionally combined using proteases, such as the native chymotrypsin-like protease. Viral structural proteins are encoded by the sgRNA region of CoVs, and include the spike (S), membrane (M), envelope (E), and nucleocapsid (N) proteins, as well proteins encoded by ORF 10. For RT-PCR testing and detection of Covid-19, the oligonucleotide primers and probes used should be selected from the nucleocapsid (N) gene region, per recommendation by the CDC and as provided in panels produced by IDT [15]. The latter panel is designed for specific detection of the 2019-nCoV (two primer/probe sets). As a control, additional primer/probe sets used as controls such as the human RNase P gene (RP) are included in the panel. It is hence of special interest to focus particularly on the N region of the genome, as high-rate mutations in this region may cause highly undesirable false negatives in the test outcomes.

Data acquisition. For our analyses, we used data from the GISAID EpiCoV [16] database which contains sequenced viral strains collected from patients across the world. We downloaded the data on several occasions, starting from 04-03-2020, and continuing on 04-10-2020 and 04-14-2020. The data size grew significantly during this time span and in order to observe the influence of the sample set sizes on the estimates of the mutational supports we used different sizes of samples. We filtered the genomic datasets only to include nearly-complete samples i.e., those of length $> 29,000$ nts, resulting in a total of 3511 samples in 04-03-2020, 5650 samples on 04-10-2020 and 8893 samples on 04-14-2020. We also downloaded the associated metadata. As the first step in our analysis, we used the sequence alignment software MUSCLE [17] to perform pairwise alignment of all the samples with the SARS-Cov-2 sequence of Patient 1, published under the name Wuhan-Hu-1, collected from a patient admitted to the Central Hospital of Wuhan on December 26, 2020 (GenBank accession number MN909847). Next, for each aligned pair of samples we generate a "mutation profile", a list containing the position in the reference genome in which the patient aligned to the reference has a mutation. We do not perform multiple sequence alignment in order to assess the mutation landscape as we need to analyze each patient data separately. The mutation profile lists are subsequently aggregated over all the patient samples, resulting in a mutation histogram accounting for all positions in the viral reference genome. The aggregate profiles are then partitioned according to the 11 genes they are located in on the viral genome depicted in Figure 1. The total count of mutations for each location in each gene is used as a sufficient statistics for estimating the mutational support and distribution of the mutations in each of the 11 genes. To adjust for alignment artifacts introduced by sequencing errors, dropouts and alignment gaps, we removed all gaps encountered in the prefixes and suffixes and sufficiently long gaps (> 10 nts) within the alignments. Most gaps are encountered at the 5'UTR and

Table 1: Number of samples available for different classes for data collected on three different dates.

Date	Age		Gender		Region			Total # of samples
	Above 55	Below 55	Male	Female	Asia	Europe	NA	
04-03-2020	909	1,477	1,349	1,061	510	1,695	818	3,511
04-10-2020	2,373	1,850	2,315	1,956	615	3,194	1,147	5,650
04-14-2020	3,047	3,231	3,526	2,817	636	5,890	1,774	8,893

3'UTR regions of the genome, as expected from outputs of global alignment algorithms.

As there exists a large body of evidence of stratified susceptibility and severity of symptoms across different racial, age and gender groups, we perform four types of mutational support and distribution estimates. In the first set of tests, we divided the patient mutations based on gender (Male/Female), based on age (under 55/over 55) and based on geographical locations (Asia/ North America/ Europe) and based on a combination of features for which sufficiently many samples are available, such as Male/Female, Below55/Above55, Europe. Since the number of samples per feature type may vary significantly, we performed two tests. In one test we used all samples available, while in another we adjusted for difference in sizes of the sets by subsampling the larger of the two classes to make the sample set sizes equal. The number of samples available for various classes is depicted in Table 1. For data obtained on 04-03-2020, we used all the samples available for all the classes, without balancing the class sizes. For data from 04-10-2020 and 04-14-2020, we balanced the classes by subsampling from the larger of the two classes for both age and gender. For geographical regions, we used all 615 samples from Asia and subsampled Europe and North America to 1000 samples each for 04-10-2020. Similarly, we used all 636 samples from Asia and subsampled Europe and North America to 1774 samples each for 04-14-2020 to account for differences in class sizes. It is important to point out that by performing the experiments with different sample set sizes one can compare the quality of the estimates in the early stages of epidemics and later stages when more information about individual strains of viruses becomes available. Furthermore, the methods outlined in this work apply to any other viral or bacterial dataset collection with the obvious modifications in place to account for the genetic profile of the microorganisms.

Estimation methods. The most commonly used techniques for support and distribution estimation are maximum likelihood (ML) methods which may be seen as methods that directly employ the empirical counts of the symbols to determine the quantities of interest. It is well known that ML approaches perform poorly for large alphabet sizes (supports of the distribution) when only a small number of samples from the distribution is available as they fail to take account the fact some samples have never been observed due to limited data. The problem of estimating the support of an unknown probability distribution or estimation the distribution itself in the context of small sample sets has a long history. The first line of work in this area is attributed to Laplace, as described in [18], who introduced a class of smoothed distribution estimators termed add 1 (or more generally, add constant c estimators). These estimators adjust the counts of symbols in order to account for the unseen.

The support of a discrete probability distribution is the number of symbols with positive probability. We define the *mutational support* of a virus as the total number of genomic sites mutated in any viral genome in any individual (observed or unobserved due to limited testing) compared to a reference genome, which in this case is the Patient 1 genome, the first sequenced SARS-Cov-2 genome. Our postulate is that the mutational support provides a good assessment of the overall number of mutations encountered in a virus and its strains throughout an epidemic/pandemic outbreak. Other types/definitions of mutation rates for SARS-Cov-2 have been widely reported in the literature. What is referred to as the *genomic mutation rate* is the product of the per-nucleotide site mutation rate and the genome size, and it represents the average number of mutations each offspring will have compared to the parental (or ancestral) genome. RNA viruses have a per site mutation rate that lie in the range $10^{-6} - 10^{-4}$ [19]. The mutation rate of a virus is often equated with the rate at which errors are made during replication of the viral genome. Clearly, determining the genomic mutation rate in a large carrier population appears to be a challenging task as each host will have a different mutation rate and due to the fact that distinguishing offspring and ancestors appears hard. Furthermore, replication errors may clearly not be the only mechanism behind viral mutations. The genome mutational rate for SARS-Cov-2 is estimated at roughly 2 to 3 mutations a month.

We argue that the mutational support more accurately reflects the mutational power of a virus than the mutation rate based on RNA replication analysis alone as it is obtained through a small-sample statistical analysis of a cohort of hosts. The underlying statistical approaches and methods are designed to account for unsequenced and hence unseen mutations and genomes.

To estimate the mutational support given small sample sets, we use the polynomial estimators in [20] based on the method first described in [21]. For completeness, the *regularized weighted Chebyshev estimators* are described below.

Let $P = (p_1, p_2, \ldots)$ be a discrete distribution over some finite alphabet and let $\mathbf{x}^n$ be a vector of i.i.d. samples

drawn according to the distribution P. The problem of interest is to estimate the support size, defined as $S(P) = \sum_i \mathbf{1}_{\{p_i>0\}}$. We use S instead of $S(P)$ to avoid notational clutter. An important assumption used in all estimation methods is that the minimum non-zero probability of the distribution P is greater than $\frac{1}{k}$, for some $k \in \mathbb{R}^+$, i.e., $\inf\{p \in P \,|\, p > 0\} \geq \frac{1}{k}$. We let D_k denote the space of all probability distribution satisfying $\inf\{p \in P \,|\, p > 0\} \geq \frac{1}{k}$. A sufficient statistics for $\mathbf{x}^n$ is the empirical distribution (i.e., histogram) $n = (n_1, n_2, \ldots)$, where $n_i = \sum_{j=1}^n \mathbf{1}_{\{\mathbf{x}_j=i\}}$ and $\mathbf{1}_A$ stands for the indicator function of the event A.

To determine the quality of the estimator, we use the minmax risk under normalized squared loss $R^*(k,n)$ defined as

$$R^*(k,n) = \inf_{\hat{S}} \sup_{P \in D_k} \mathbb{E}\left[\left(\frac{\hat{S}(N) - S}{k}\right)^2\right]. \tag{1}$$

We seek a support estimator $\hat{S}$ that minimizes

$$\sup_{P \in D_k} \mathbb{E}\left[\left(\frac{\hat{S}(N) - S}{k}\right)^2\right] = \sup_{P \in D_k}\left[\mathbb{E}^2\left(\frac{\hat{S}(N) - S}{k}\right) + var\left(\frac{\hat{S}(N) - S}{k}\right)\right].$$

The first term within the supremum captures the expected bias of the estimator $\hat{S}$. The second term represents the variance of the estimator $\hat{S}$. A"good" estimator is required to balance out the worst-case contributions of the bias and variance.

The Chebyshev polynomial of the first kind of degree L is defined as $T_L(x) = \cos(L \arccos(x)) = \frac{z^L + z^{-L}}{2}$, where z is the solution of the quadratic equation $z + z^{-1} = 2x$. From the definition, it is easy to see that the polynomial T_L is bounded in the interval $[-1, 1]$. Chebyshev polynomials may be scaled and shifted to lie in an interval $[l, r]$ not necessarily equal to $[-1, 1]$,

$$R_L(x) = -\frac{T_L(\frac{2x-r-l}{r-l})}{T_L(\frac{-r-l}{r-l})} \triangleq \sum_{j=0}^{L} \tilde{a}_j x^j.$$

Clearly, $R_L(0) = -1$, and $\tilde{a}_0 = -1$. The coefficients in the above expansion equal

$$\tilde{a}_j = \frac{R_L^{(j)}(0)}{j!}. \tag{2}$$

The estimator proposed in [21] takes the form $\tilde{S} = \sum_i \tilde{g}_L(N_i)$, where

$$\tilde{g}_L(j) = \begin{cases} \tilde{a}_j j! + 1, & \text{if } j \leq L, \\ 1, & \text{otherwise,} \end{cases}, \; L \triangleq \lfloor c_0 \log k \rfloor, \; [l, r] \triangleq \left[\frac{n}{k}, c_1 \log k\right]. \tag{3}$$

By introducing a regularization term and an exponential weighting factor, this estimator can be significantly improved in practice as documented in [20]. Since the estimator formulation is nontrivial, we omit its full description and refer the interested reader to the previously cited work. We only remark that the estimator termed RWC (regularized weighted Chebyshev) optimizes the regularized risk described above, while the RWC-S estimator uses a risk objective which involves a different normalization term.

By far the most frequently used method for distribution estimation is the Good-Turing estimator [12], which in a slightly modified form may be described as follows. For a sequence $\mathbf{x}^n$ of length n over an unknown finite alphabet, we let n_i denote the number of times a symbol i appears in $\mathbf{x}^n$. Furthermore, we let φ_t stand for the count of counts, i.e., the number of symbols that appear t times in $\mathbf{x}^n$. The estimator proposed in [13] combines the Good-Turing and ML estimators, the latter being used for the frequently observed symbols. For symbols that appear t times, if $\varphi_{t+1} \geq \Omega(t)$, then the Good-Turing estimate is used to determine the underlying total probability mass, otherwise, the ML estimator is used instead. More precisely, for a symbol appearing t times, if $\varphi_{t+1} \geq t$ we use the Good-Turing estimator, otherwise we use the empirical estimator. If $n_i = t$, then the probability of the symbol i is computed according to

$$p_i = \begin{cases} \frac{t}{N}, & \text{if } t > \varphi_{t+1}, \\ \frac{\varphi_{t+1}+1}{\varphi_t} \frac{t+1}{N}, & \text{otherwise,} \end{cases}$$

where N is a normalization term that ensures that the obtained values are probability masses. The term φ_{t+1} used in the Good-Turing estimator is replaced by $\varphi_{t+1} + 1$ so that every symbol has a nonzero probability.

A software implementation of Good-Turing estimators is available at:
http://crr.ugent.be/papers/A%20Python%20program%20to%20calculate%20the%20Good-Turing%20algorithm.pdf.

Modifications of the Good-Turing estimator that take sampling artifacts/errors such as community structures into account may be found in [22, 23].

Table 2: Support sizes of different age groups based on 909 samples for individuals over 55 years of age and $1,477$ samples below 55 years of age. The data was obtained from GISAID on 04-03-2020 and includes all the samples for the two categories available at the given date.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Above 55	Below 55	Above 55	Below 55	Above 55	Below 55	All Ages
ORF1a	625	764	1,280	1,544	1,209	1,454	13,203
ORF1b	276	616	570	1,301	514	1,223	8,087
S	160	218	291	420	277	375	3,822
ORF3a	55	73	103	132	92	121	828
E	14	13	23	23	23	22	228
M	34	35	58	63	54	55	669
ORF6	11	25	19	42	19	42	186
ORF7a	24	27	41	45	39	44	366
ORF8	340	340	87	344	235	343	366
N	66	110	108	197	97	172	1,260
ORF10	26	29	29	53	33	53	117

3 Results

3.1 Support Estimation

The first set of results pertains to data collected at an earlier stage of the pandemic (04-03-2020) that did not include sufficiently many samples to allow for sample set size leveling and therefore included all available samples. From Table 2, it is apparent that naive ML methods underestimate the mutational support in the ORF1a and ORF1b genes roughly two-fold and that the mutational support of both genes is roughly 10% of the total gene lengths. Note that it is interesting to observe that both the ML and RWC-S estimators indicate that the mutational support is higher in younger patients, but in this case the results may be explained by the uneven sample set sizes for the two patient categories (909 versus $1,477$). Also, the mutational support of the N region is significantly smaller, amounting to roughly 1% of the genome length for both categories. Similar results may be observed for the case that patients are partitioned according to gender, as listed in Table 3. An interesting observation is in place for the results pertaining to different geographic regions. Despite the fact that the number of available samples from Asia is smaller than that of Europe and North America, the mutational support in the ORF1a region of Asian patients is more than twice as large as that of North America. A similar result holds true for the case of ORF1b, where the European population has three times more mutations than the North American population. ORF1a encodes replicase polyproteins pp1a and pp1b, which implies that the replication machinery has undergone significantly more adaptations in Asia than North America. It has been documented that the ratio of ORF1a- and ORF1b-encoded proteins plays an important role in the replication efficiency of coronaviruses [24].

Tables 5,6,7 list the results analogue to those in Table 2,3,4 respectively, obtained using larger datasets retrieved on 04-10-2020 which allow for random subsampling that leads to equal sample set sizes for all subpopulations considered.

Based on the results of Table 5, one additional week of data collection amounting to roughly twice the samples increased the mutational support by 5% for both ORF1 and ORF1b. On the other hand, the additional data samples show that the N region of the SARC-Cov-2 genome exhibited a much more significant increase in mutations than as could be predicted from early small-set sample sizes, amounting to roughly 10% of the genome. This finding is of great significance for Covid-19 and other viral outbreak testing methods as it indicates that genomic regions used as identifiers for the virus may mutate much faster than predicted based on small preliminary sample set information and that one may have to change the primers used for testing as the disease progresses. It also suggests that, as different strata of the population exhibit different mutation rates, different primers have to be used for testing them.

The most surprising result is listed in Table 6, and pertains to the ORF1b region. As may be seen, the mutational support in the female population is $1,621$ compared to 941 in the male population, which amounts to a 8.4% difference with respect to the length of the open reading frame. To address this issue further, we performed another test the results of which are shown in Table 11 and discussed later. Geographic trends are depicted in Table 7 and follow the same trend observed in Table 4.

Tables 8,9,10 show the trends of increase for the mutation support with increased sample sizes, which in this case exceed $8,000$.

Table 3: Support sizes for different genders based on $1,349$ male and $1,061$ female samples. The data was obtained from GISAID on 04-03-2020 and includes all the samples for the two categories available at the given date.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Male	Female	Male	Female	Male	Female	Both Genders
ORF1a	854	702	1,807	1,468	1,702	1,388	13,203
ORF1b	348	594	690	1,307	640	1,234	8,087
S	225	186	447	359	405	329	3,822
ORF3a	68	61	132	111	115	99	828
E	18	10	30	18	29	18	228
M	37	36	62	68	57	60	669
ORF6	13	27	22	49	21	50	186
ORF7a	32	21	55	38	53	38	366
ORF8	340	341	344	592	343	458	366
N	96	85	165	143	146	129	1,260
ORF10	26	10	30	17	29	17	117

Table 4: Support sizes for different geographical regions based on 510 samples from Asia, $1,695$ from Europe and 818 from North America. The data was obtained from GISAID on 04-03-2020 and includes all the samples for the two categories available at the given date.

Gene	ML			RWC			RWC-S			Maximum Support
Symbol	Asia	Europe	NA	Asia	Europe	NA	Asia	Europe	NA	All Three Regions
ORF1a	770	757	397	1,645	1,558	776	1,603	1,455	720	13,203
ORF1b	279	590	205	566	1251	372	553	1,159	345	8,087
S	168	181	131	321	345	254	313	309	230	3,822
ORF3a	84	62	38	158	113	71	154	100	63	828
E	37	11	6	66	19	9	65	19	9	228
M	30	29	15	53	49	25	50	44	24	669
ORF6	2	28	5	2	46	8	2	45	7	186
ORF7a	108	38	49	215	66	90	214	65	89	366
ORF8	340	27	19	341	46	26	342	43	28	366
N	53	90	68	93	152	122	85	137	114	1,260
ORF10	10	25	9	18	28	15	17	27	14	117

Table 5: Support sizes of different age groups based on $1,850$ samples from each group. The data was retrieved from GISAID on 04-10-2020.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Above 55	Below 55	Above 55	Below 55	Above 55	Below 55	All Ages
ORF1a	996	934	2,039	1,857	1,896	1,743	13, 203
ORF1b	499	484	991	965	924	896	8,087
S	265	279	490	547	458	501	3,822
ORF3a	104	79	188	138	171	124	828
E	23	19	36	33	36	32	228
M	55	47	98	86	92	77	669
ORF6	38	26	65	43	64	41	186
ORF7a	60	31	108	50	103	49	366
ORF8	340	341	93	342	236	343	366
N	140	163	248	294	223	265	1,260
ORF10	31	28	35	49	39	50	117

Table 6: Support sizes for the two different genders based on 1956 samples for each group. The data was retrieved from GISAID on 04-10-2020.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Male	Female	Male	Female	Male	Female	Both Genders
ORF1a	1,071	1,115	2,176	2,313	2,055	2,175	13,203
ORF1b	500	804	1,013	1,721	941	1,621	8,087
S	283	293	551	562	509	519	3,822
ORF3a	114	99	216	175	190	158	828
E	24	14	37	23	36	22	228
M	52	56	87	101	82	94	669
ORF6	42	30	75	51	74	50	186
ORF7a	42	51	74	87	71	84	366
ORF8	341	342	344	345	344	345	366
N	143	162	251	282	226	259	1,260
ORF10	29	12	33	20	32	19	117

Table 7: Support size for three different geographic regions based on 615 samples from Asia and $1,000$ samples from Europe and North America each. The data was retrieved from GISAID on 04-10-2020.

Gene	ML			RWC			RWC-S			Maximum Support
Symbol	Asia	Europe	NA	Asia	Europe	NA	Asia	Europe	NA	All Regions
ORF1a	827	504	470	1,768	975	948	1,725	919	874	13,203
ORF1b	308	271	244	631	531	478	611	491	432	8,087
S	182	163	142	352	336	269	340	293	243	3,822
ORF3a	91	56	39	174	96	74	168	85	63	828
E	37	12	14	66	21	24	65	21	24	228
M	31	23	17	55	38	28	52	35	27	669
ORF6	3	48	15	3	87	26	3	86	25	186
ORF7a	109	63	51	216	118	94	214	116	93	366
ORF8	340	19	21	335	29	31	339	29	31	366
N	58	72	77	96	121	137	91	108	129	1,260
ORF10	10	26	7	18	48	10	17	48	10	117

Table 8: Support sizes for different age ranges based on $3,047$ samples in each group. The data was retrieved from GISAID on 04-14-2020. Note that the entry in the table under ORF8 marked by * corresponds to a rare scenario where our estimators produce a value smaller than that predicted by an ML estimator. This is due to severe sampling artifacts and in this case, one should choose the larger of the two estimates available.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Above 55	Below 55	Above 55	Below 55	Above 55	Below 55	All Ages
ORF1a	1,488	1,439	3,049	2,876	2,830	2,697	13,203
ORF1b	787	953	1,526	2,008	1,431	1,876	8,087
S	431	400	757	780	725	724	3,822
ORF3a	156	165	277	268	251	255	828
E	43	26	71	40	69	39	228
M	79	62	141	105	133	98	669
ORF6	53	32	93	50	90	49	186
ORF7a	86	66	133	113	131	109	366
ORF8	341	343	91*	344	239	343	366
N	201	219	356	408	322	364	1,260
ORF10	35	33	43	56	46	55	117

Table 9: Support size differences for males and females based on $2,817$ samples for each group. The data was retrieved from GISAID on 04-14-2020.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Male	Female	Male	Female	Male	Female	Both Genders
ORF1a	1,478	1,456	2,947	3,004	2,762	2,792	13,203
ORF1b	705	991	1,376	2,074	1,292	1,958	8,087
S	405	389	790	716	740	673	3822
ORF3a	169	140	272	255	262	230	828
E	30	36	47	63	45	61	228
M	67	69	107	119	103	112	669
ORF6	50	40	87	66	84	65	186
ORF7a	68	72	113	106	110	105	366
ORF8	343	342	347	345	347	345	366
N	195	204	338	348	312	327	1,260
ORF10	31	13	33	22	38	21	117

Table 10: Support sizes for different geographic regions based on 636 samples from Asia and $1,774$ each from Europe and North America. The data was retrieved from GISAID on 04-14-2020.

Gene	ML			RWC			RWC-S			Maximum Support
Symbol	Asia	Europe	NA	Asia	Europe	NA	Asia	Europe	NA	All Regions
ORF1a	835	911	804	1,799	1,857	1,608	1,752	1,746	1,509	13,203
ORF1b	316	477	403	645	920	811	624	858	727	8,087
S	188	246	209	367	478	415	353	438	375	3822
ORF3a	93	99	81	178	188	152	171	166	134	828
E	36	15	15	64	26	26	63	26	25	228
M	31	51	28	54	85	47	51	81	44	669
ORF6	3	52	21	3	92	36	3	90	33	186
ORF7a	109	66	135	216	119	273	214	116	269	366
ORF8	340	32	29	335	50	46	339	49	44	366
N	60	139	138	99	246	241	93	221	227	1,260
ORF10	11	30	10	18	36	16	17	39	16	117

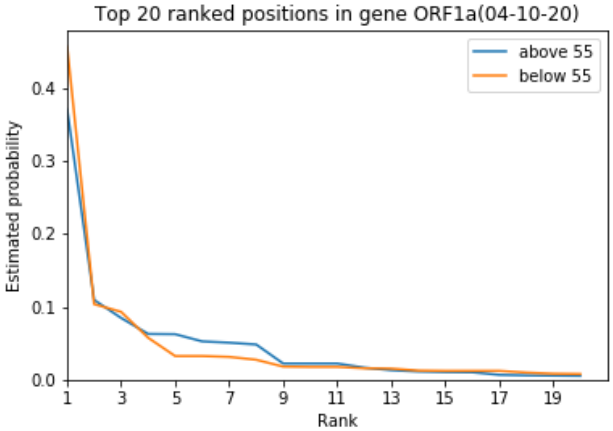
Table 11: Support size differences between males above 55 years of age and females below 55 years of age from Europe based on $1,078$ samples in each group. The data was retrieved from GISAID on 04-14-2020.

Gene	ML		RWC		RWC-S		Maximum Support
Symbol	Male, > 55	Female, < 55	Male, > 55	Female, < 55	Male, > 55	Female, < 55	Both Categories
ORF1a	588	670	1,159	1,374	1,078	1,294	13,203
ORF1b	349	553	686	1189	638	1,117	8,087
S	209	166	420	329	387	296	3,822
ORF3a	76	61	138	104	124	96	828
E	10	9	17	15	16	14	228
M	27	33	45	58	40	52	669
ORF6	15	28	25	47	24	48	186
ORF7a	31	23	54	36	52	36	366
ORF8	27	28	45	48	43	46	366
N	110	108	197	199	178	183	1,260
ORF10	27	5	28	7	33	7	117

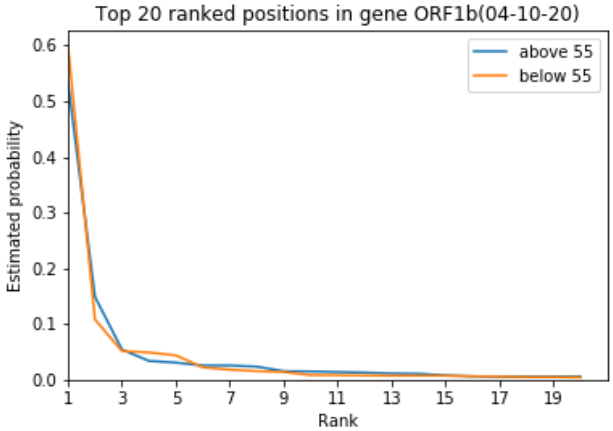
Table 11 provides results for a finer partition of test samples into two categories, one including males over 55 years of age and another females below 55 years of age, with both populations sampled from Europe. The first category has been empirically observed to be at higher risk of infection and for exhibiting more severe symptoms [25]. The important finding is that the mutational support of ORF1b is almost twice as large in the low risk population compared to the high risk population. This result may imply that the large mutational support is a result of a highly competitive virus-host interaction which forces the virus to mutate in order to gain advantage over the host's immune system.

3.2 Distribution Estimation

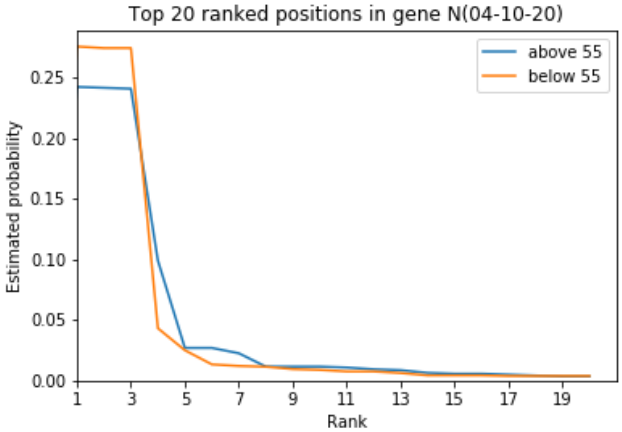
Next, we examine the distribution of mutations in the ORF1a,b and N regions of the SARS-Cov-2 virus obtained using the Good-Turing estimator and once again focusing on different population traits. As may be seen from Figures 2 and 3 there is a surprisingly small difference in the distribution of the top-20 mutated sites across different age and gender groups, except for a marked difference in the largest probability (in particular, in the N region for populations partitioned according to age and populations partitioned according to gender when including larger sample sets from 04-14-2020). This is especially the case for samples partitioned according to gender, despite the fact that the number of mutations in female subjects in the ORF1b region was close to twice as much as that in male subjects. In addition, the probability of having a mutation at the highest probability sites is significantly larger in "younger" than "older" populations. The trend remains the same for different collection dates as supported by the results in Figures 5 and 6. The situation is completely different when comparing the distributions of mutations across different geographic



(a) Mutations in the ORF1a region



(b) Mutations in the ORF1b region



(c) Mutations in the N region

Figure 2: Comparison of the distributions of mutations in adults <55 of age and adults ≥55 of age tested before 04-14-2020.

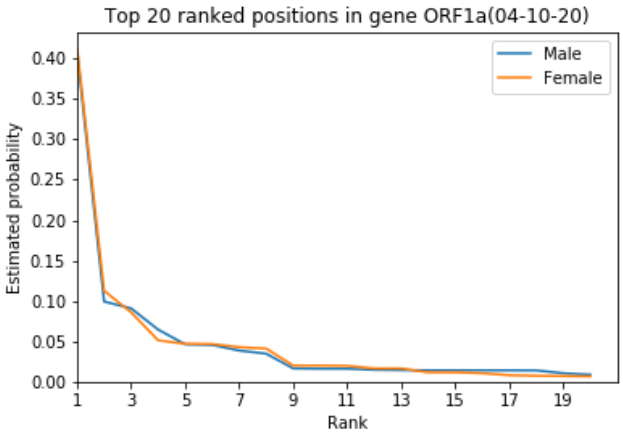
regions (Figures 4 and 7), where there are significant differences in the distributions as one would expect. To compare the distributions, we computed all three pairwise symmetric Kulback-Leibler (KL) divergences for the normalized top-20 mutation probabilities. The symmetric KL divergence between two discrete probability distributions p and q is defined as

$$D_s(p,q) = D(p||q) + D(q||p), \quad D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

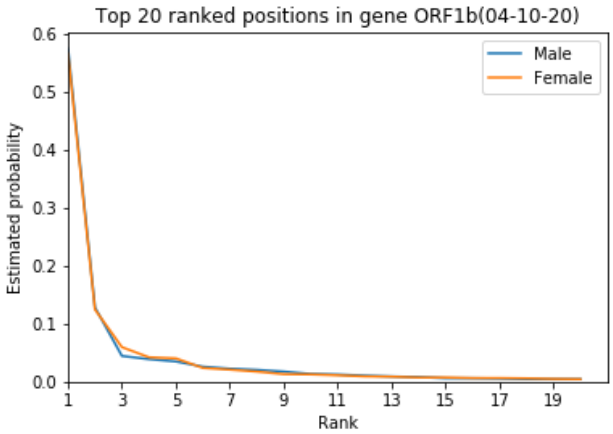
For the mutation distributions pertaining to the pairs Europe-NA, Europe-Asia and Asia-NA, the KL divergences equal 0.672, 0.316 and 0.376 (ORF1a), 0.491, 0.435 and 0.646 (ORF1b), 0.293, 1.021 and 0.303 (N), respectively, for data collected until 04-14-2020. These results indicate that the largest differences in the distributions in the ORF1a region exist between Europe and North America, while the largest differences in the ORF1b region exist between Asia and North America. For the N region, a significant difference between the distributions of mutations is observed between Europe and Asia, and at this point, these large distances do not seem to have a simple explanation.

Similarly, the corresponding KL divergences based on the samples collected until 04-10-2020 equal 0.788 (which is significantly larger than the one predicted based on data collected on 04-14-2020), 0.328 and 0.371 (ORF1a), 0.743 which is significantly larger than the one predicted based on data collected on 04-14-2020), 0.615 and $0.0.755$ (ORF1b), 0.315, 0.893 and 0.248 (N), respectively. The results for the KL divergences for the N regions suggest relatively small changes in the distribution of mutations in the N region, and larger changes in the ORF1a and ORF1b regions, which is expected.

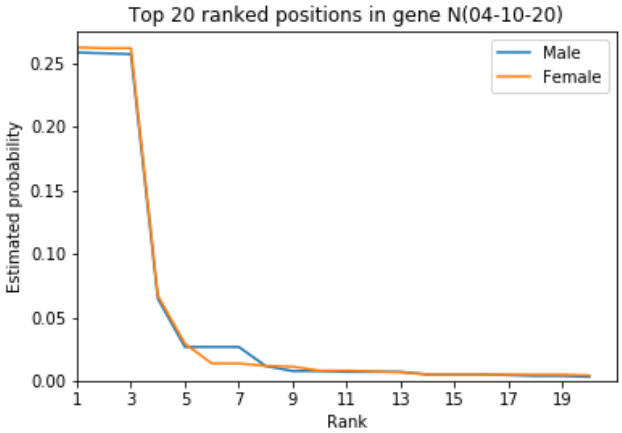
The distributions of mutations only reveal the statistical landscape of the mutation sites but not their exact locations in the genome. The actual mutated sites in the SARS-Cov-2 genomes are depicted in Figures 9,10 and 11. As can



(a) Mutations in the ORF1a region

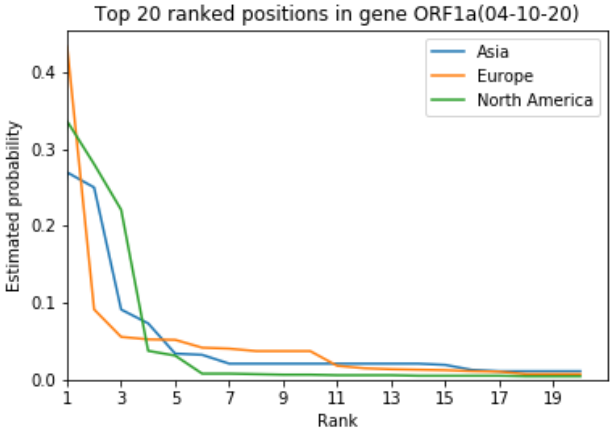


(b) Mutations in the ORF1b region

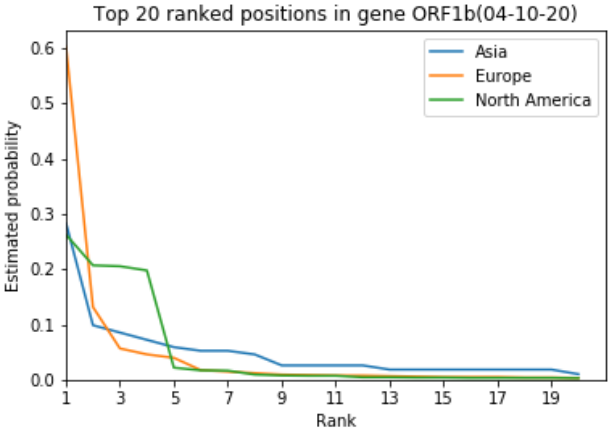


(c) Mutations in the N region

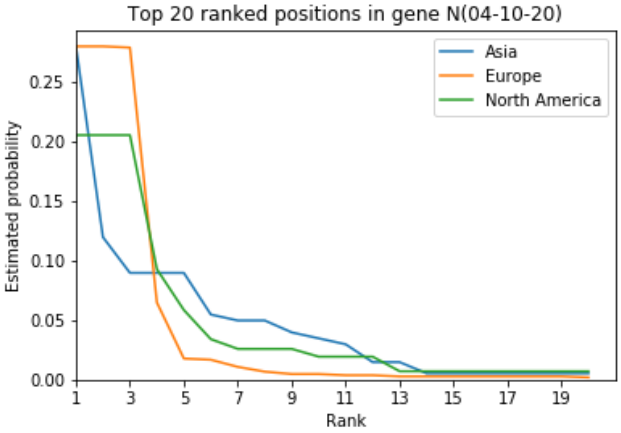
Figure 3: Comparison of the distributions of mutations in male and female test subjects tested before 04-14-2020.



(a) Mutations in the ORF1a region

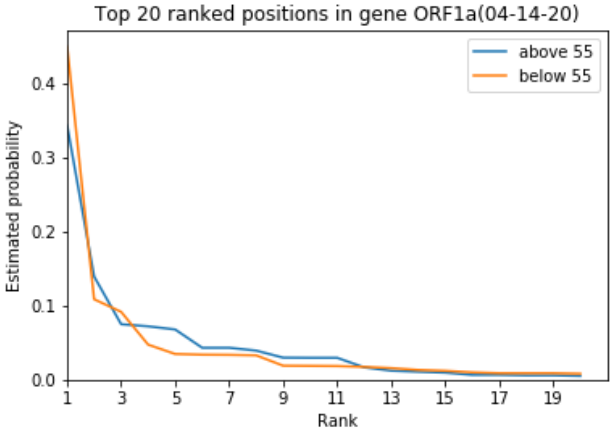


(b) Mutations in the ORF1b region

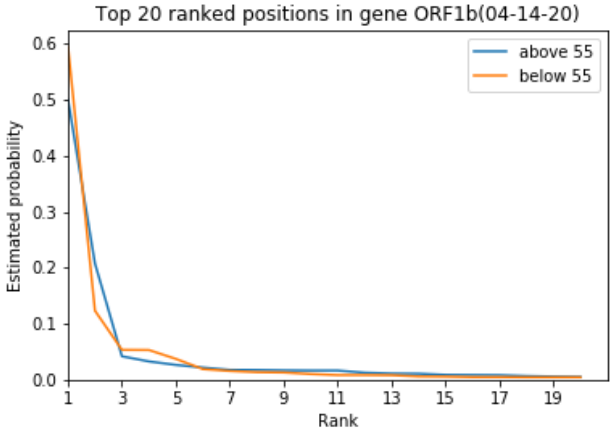


(c) Mutations in the N region

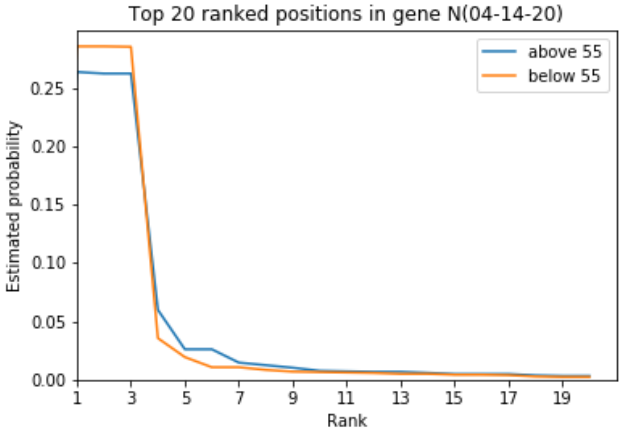
Figure 4: Differences in the distribution of mutations for different geographic regions based on subjects tested before 04-10-2020.



(a) Mutations in the ORF1a region

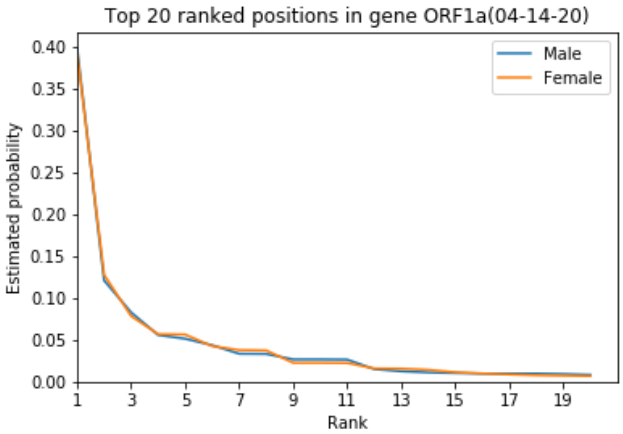


(b) Mutations in the ORF1b region

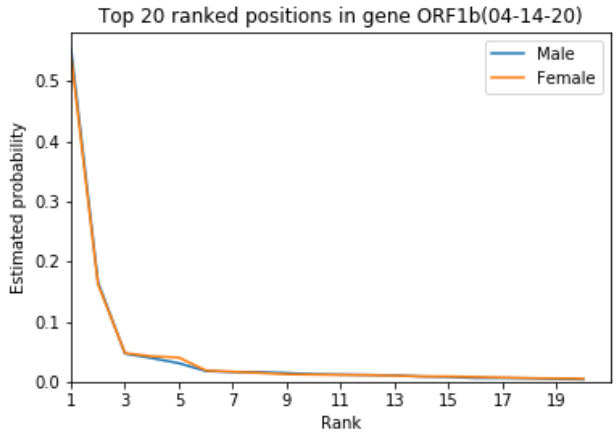


(c) Mutations in the N region

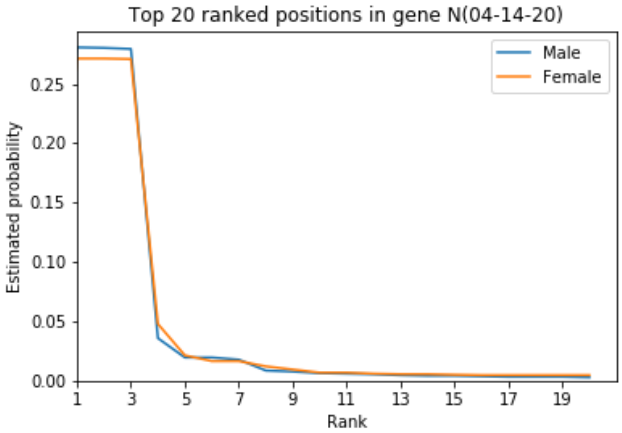
Figure 5: Comparison of the distributions of mutations in adults <55 years of age and adults ≥55 years of age tested before 04-14-2020.



(a) Mutations in the ORF1a region

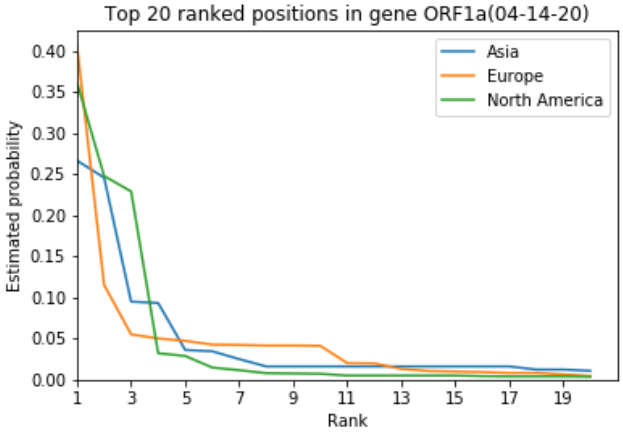


(b) Mutations in the ORF1b region

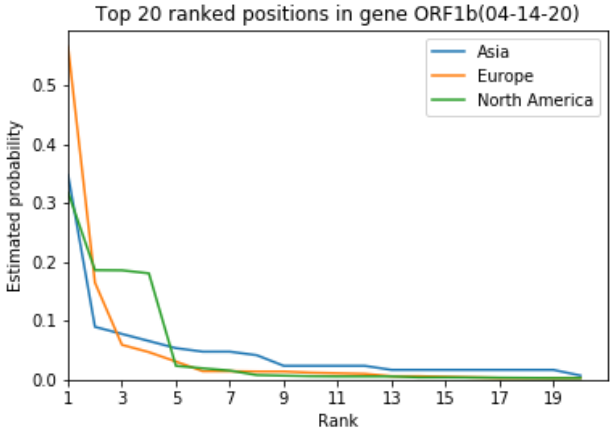


(c) Mutations in the N region

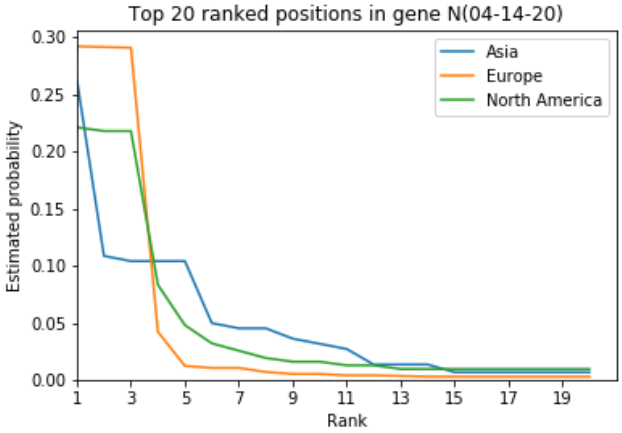
Figure 6: Comparison of the distributions of mutations in male and female test subjects tested before 04-14-2020.



(a) Mutations in the ORF1a region

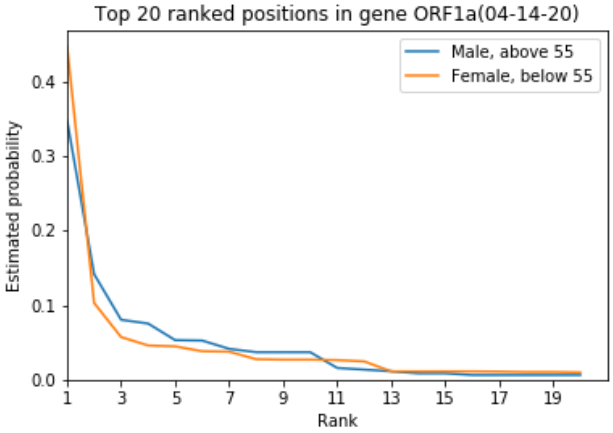


(b) Mutations in the ORF1b region

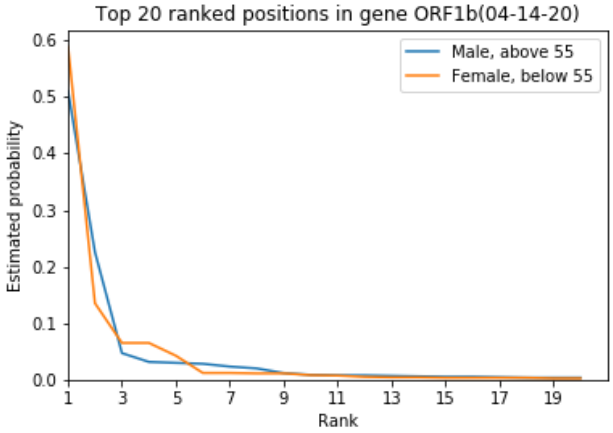


(c) Mutations in the N region

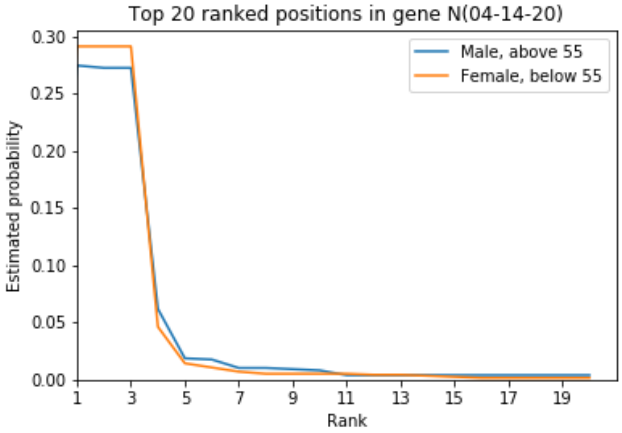
Figure 7: Comparison of the distributions of mutations in test subjects from three different continents tested before 04-14-2020.



(a) Mutations in the ORF1a region



(b) Mutations in the ORF1b region



(c) Mutations in the N region

Figure 8: Differences in the distribution of mutations between males of age $\geq$55 and females of age $<$55 from Europe tested before 04-14-2020.

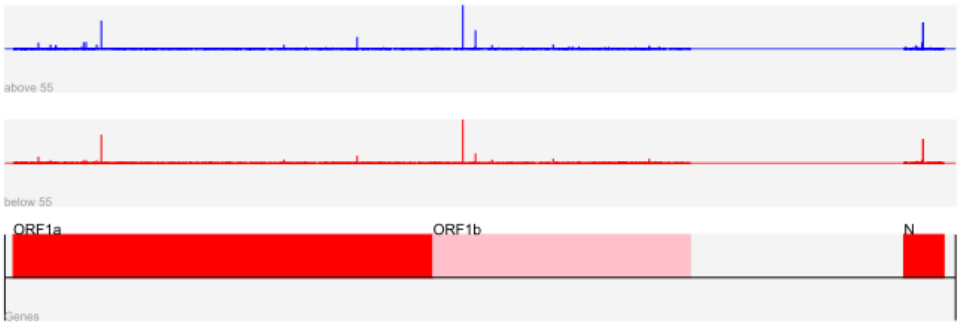


Figure 9: Positions in the SARS-Cov-2 genome with high probability of mutations in patients above the age of 55 (top) and patients below the age of 55 (bottom) based on data collected until 04-14-2020. The height of the bar is proportional to the probability of the mutation.

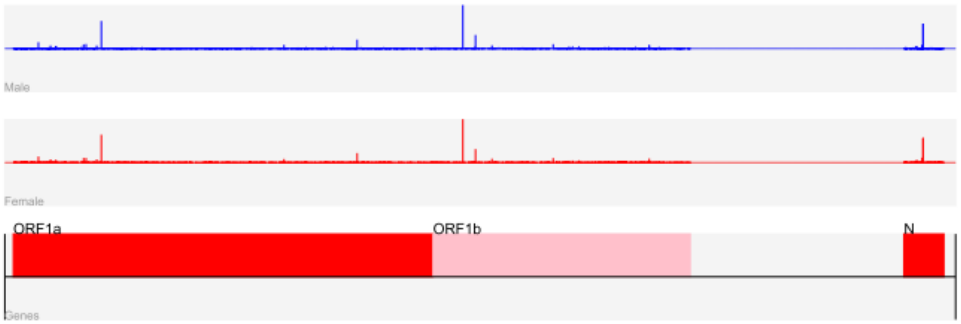


Figure 10: Positions in the SARS-Cov-2 genome with high probability of mutations in male patients (top) and female patients (bottom) based on data collected until 04-14-2020. The height of the bar is proportional to the probability of the mutation.

be seen, the locations of the mutations for the first two pairs of categories are almost identical. Nevertheless, the positional stratification of mutations is significant for patients from different continents, especially in the N region of the SARS-Cov-2 genome. The largest spread of probability mass is observed for patients in Asia which may be indicative of the larger exploration rate for mutations in the region where the outbreak originated. Another plausible explanation is that Asia, the origin of the pandemics, is in a later phase of the pandemic when compared to Europe and North America. The observation may also has an impact on the design of testing schemes which use the N region of the genome as for patients from Europe there exist only 2-3 sites with high mutation rates, with a similar trend observed for North American populations.

References

[1] R. Sanjuán, M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw, "Viral mutation rates," *Journal of virology*, vol. 84, no. 19, pp. 9733–9748, 2010.

[2] J. W. Drake and J. J. Holland, "Mutation rates among rna viruses," *Proceedings of the National Academy of Sciences*, vol. 96, no. 24, pp. 13910–13913, 1999.

[3] R. Sanjuán and P. Domingo-Calap, "Mechanisms of viral mutation," *Cellular and molecular life sciences*, vol. 73, no. 23, pp. 4433–4448, 2016.

[4] S. Duffy, L. A. Shackelton, and E. C. Holmes, "Rates of evolutionary change in viruses: patterns and determinants," *Nature Reviews Genetics*, vol. 9, no. 4, pp. 267–276, 2008.

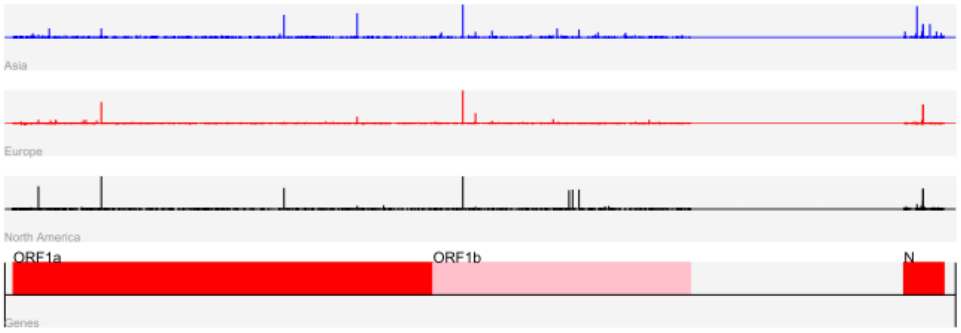


Figure 11: Positions in the SARS-Cov-2 genome with high probability of mutations in patients across three different continents collected until 04-14-2020. The height of the bar is proportional to the probability of the mutation.

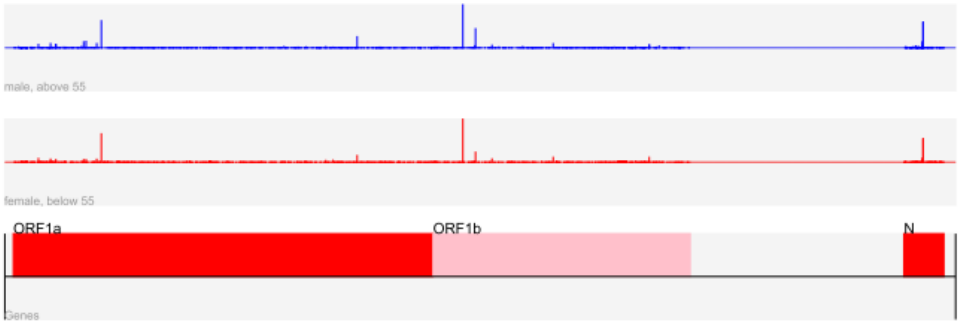


Figure 12: Positions in the SARS-Cov-2 genome with high probability of mutations in European females below the age of 55 and males above the age of 55 collected until 04-14-2020. The height of the bar is proportional to the probability of the mutation.

[5] R. Sanjuán, A. Moya, and S. F. Elena, "The distribution of fitness effects caused by single-nucleotide substitutions in an rna virus," *Proceedings of the National Academy of Sciences*, vol. 101, no. 22, pp. 8396–8401, 2004.

[6] A. Acevedo, L. Brodsky, and R. Andino, "Mutational and fitness landscapes of an rna virus revealed through population sequencing," *Nature*, vol. 505, no. 7485, pp. 686–690, 2014.

[7] C. L. Burch and L. Chao, "Evolvability of an rna virus is determined by its mutational neighbourhood," *Nature*, vol. 406, no. 6796, pp. 625–628, 2000.

[8] S. Schultz-Cherry and J. C. Jones, "Influenza vaccines: the good, the bad, and the eggs," in *Advances in virus research*, vol. 77, pp. 63–84, Elsevier, 2010.

[9] J. J. Bull, R. Sanjuan, and C. O. Wilke, "Theory of lethal mutagenesis for viruses," *Journal of virology*, vol. 81, no. 6, pp. 2930–2939, 2007.

[10] T. Hoenen, D. Safronetz, A. Groseth, K. Wollenberg, O. Koita, B. Diarra, I. Fall, F. Haidara, F. Diallo, M. Sanogo, *et al.*, "Mutation rate and genotype variation of ebola virus from mali case sequences," *Science*, vol. 348, no. 6230, pp. 117–119, 2015.

[11] R. M. Ribeiro, H. Li, S. Wang, M. B. Stoddard, G. H. Learn, B. T. Korber, T. Bhattacharya, J. Guedj, E. H. Parrish, B. H. Hahn, *et al.*, "Quantifying the diversification of hepatitis c virus (hcv) during primary infection: estimates of the in vivo mutation rate," *PLoS pathogens*, vol. 8, no. 8, 2012.

[12] W. A. Gale and G. Sampson, "Good-turing frequency estimation without tears," *Journal of quantitative linguistics*, vol. 2, no. 3, pp. 217–237, 1995.

[13] A. Orlitsky and A. T. Suresh, "Competitive distribution estimation: Why is good-turing good," in *Advances in Neural Information Processing Systems*, pp. 2143–2151, 2015.

[14] L. Mousavizadeh and S. Ghasemi, "Genotype and phenotype of covid-19: Their roles in pathogenesis," *Journal of Microbiology, Immunology and Infection*, 2020.

[15] CDC *2019 Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel, Catalog Number 2019-nCoVEUA-01 with 1000 reactions - For Emergency Use Only*, 2020.

[16] Y. Shu and J. McCauley, "Gisaid: Global initiative on sharing all influenza data–from vision to reality," *Eurosurveillance*, vol. 22, no. 13, 2017.

[17] R. C. Edgar, "Muscle: a multiple sequence alignment method with reduced time and space complexity," *BMC bioinformatics*, vol. 5, no. 1, p. 113, 2004.

[18] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Always good turing: Asymptotically optimal probability estimation," *Science*, vol. 302, no. 5644, pp. 427–431, 2003.

[19] K. M. Peck and A. S. Lauring, "Complexities of viral mutation rates," *Journal of virology*, vol. 92, no. 14, pp. e01031–17, 2018.

[20] I. Chien and O. Milenkovic, "Regularized weighted chebyshev approximations for support estimation," *arXiv preprint arXiv:1901.07506*, 2019.

[21] Y. Wu, P. Yang, *et al.*, "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *The Annals of Statistics*, vol. 47, no. 2, pp. 857–883, 2019.

[22] F. Farnoud, O. Milenkovic, and N. P. Santhanam, "Small-sample distribution estimation over sticky channels," in *2009 IEEE International Symposium on Information Theory*, pp. 1125–1129, IEEE, 2009.

[23] F. Farnoud, N. P. Santhanam, and O. Milenkovic, "Alternating markov chains for distribution estimation in the presence of errors," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 2017–2021, IEEE, 2012.

[24] E. P. Plant, R. Rakauskaitė, D. R. Taylor, and J. D. Dinman, "Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins," *Journal of virology*, vol. 84, no. 9, pp. 4330–4340, 2010.

[25] N. Y. Times, "Does covid-19 hit women and men differently? u.s. isn't keeping track," 2020.