First report on the Latvian SARS-CoV-2 isolate genetic diversity

Nikita Zrelovs[1], Monta Ustinova[1], Ivars Silamiķelis[1], Līga Birzniece[1], Kaspars Megnis[1], Vita Rovīte[1], Lauma Freimane[1], Laila Silamiķele[1], Laura Ansone[1], Jānis Pjalkovskis[1], Dāvids Fridmanis[1], Marta Priedīte[2], Anastasija Caica[2], Mikus Gavars[3], Dmitrijs Perminovs[3], Jeļena Storoženko[4], Oksana Savicka[4], Elīna Dimiņa[5], Uga Dumpis[6,7], Jānis Kloviņš[1*]

[1] Latvian Biomedical Research and Study Centre, Ratsupites 1, Riga, Latvia, LV-1067

[2] Centrala Laboratorija, Ltd, Sarlotes 1B, Riga, Latvia, LV-1001

[3] E. Gulbja Laboratorija, Ltd, Brivibas gatve 366, Riga, Latvia, LV-1006

[4] Riga East University hospital, Laboratory Service, Latvian Centre of Infectious Diseases laboratory, National Microbiology Reference Laboratory, Molecular biology and virology department, Linezera 3, Riga, Latvia, LV-1006

[5] The Centre for Disease Prevention and Control (CDPC) of Latvia, Infectious Diseases Risk Analysis and Prevention Department, Infectious Diseases Surveillance and Immunization Division, Duntes 20/3, Riga, Latvia, LV-1005

[6] University of Latvia, Faculty of Medicine, Jelgavas 3, Riga, Latvia, LV-1004

[7] Pauls Stradins Clinical University Hospital, Pilsonu 13, Riga, Latvia, LV-1002

[*] **Corresponding author:** klovins@biomed.lu.lv

Abstract

Remaining a major healthcare concern with nearly 27 million confirmed cases worldwide at the time of writing, novel severe acute respiratory syndrome coronavirus - 2 (SARS-CoV-2) has caused more than 880 thousand deaths since its outbreak in China, December 2019. First case of a person testing positive for SARS-CoV-2 infection within the territory of the Republic of Latvia was registered on 2nd of March 2020, nine days prior to the pandemic declaration by WHO. Since then, more than 230 000 tests were carried out confirming a total of 1330 cases of COVID-19 in the country as of 20th of August 2020. Rapidly reacting to the spread of the infection, an ongoing sequencing campaign was started mid-March in collaboration with the local testing laboratories, with an ultimate goal in sequencing as much local viral isolates as possible, resulting in first full-length SARS-CoV-2 isolate genome sequences from the Baltics region being made publicly available in early April. With 109 viral isolates representing ~8.2% of the total COVID-19 cases in the country being completely sequenced as of today, here we provide a first report on the genetic diversity of Latvian SARS-CoV-2 isolates.

Keywords: SARS-CoV-2, hCoV-19, 2019-nCoV, genetic diversity, next-generation sequencing, COVID-19, Latvia

Introduction

Current novel coronavirus disease (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was formerly known as 2019 novel coronavirus (2019-nCoV), and is often referred to as Human coronavirus 2019 (hCoV-19), responsible for a sudden rise in pneumonia cases in Wuhan, China, late December 2019, was preventively deemed a Public Health Emergency of International Concern by WHO as early as 30th January, 2020 with only as few as 7836 cases confirmed worldwide back then. With rapidly growing number of confirmed positive cases throughout the world, SARS-CoV-2 quickly became arguably the most sequenced virus in history with nearly 55 thousands viral isolate near complete genome sequences of high quality available publicly at the time of writing at GISAID repository thanks to the unprecedented rate of collaborations between researchers and unpublished data sharing with the goal of effectively tackling the novel disease [1,2].

Genome of SARS-CoV-2

First reported genomic sequence of SARS-CoV-2 was deduced from a metagenomic RNA of bronchoalveolar lavage fluid specimen sampled from a patient who worked at Wuhan seafood market, where the epidemiological onset of human-to-human transmission of a novel zoonotic coronavirus is thought to have taken place [3], although evidence of an earlier contraction of the disease that was not associated with the seafood market has been documented, leading to the conclusion that the primary spill-over event has taken place elsewhere[4,5]. The sequence of a 29903 base long non-segmented positive-sense single-stranded RNA molecule representing complete genome of the aforementioned isolate Wuhan-Hu-1 was deposited in GenBank [6] on 5th of January, 2020 and is now known as a SARS-CoV-2 reference sequence available under accession numbers NC_045512.2 or MN908947.3.

While viral family *Coronaviridae,* that comprises α/β/Δ/γ coronavirus genera, representatives are somewhat unique in comparison with most other RNA viruses in regards to their large genome size of ~30 kb, genomic organization of individual species does not differ much among other lower taxa within the family, while boasting variable number of open reading frames. The genome of SARS-CoV-2 begins with a 265 base long 5' UTR region starting with a leader sequence followed by a 21 290 base long ORF1ab, comprising about 70% of the genome length, that translates into two polyproteins via -1 ribosomal frameshift and encodes 16 non-structural proteins (nsp1-nsp16). The remaining part of the genome comprises ORFs coding for structural and accessory proteins of unknown function, sequentially: Spike glycoprotein (S), ORF3a, envelope protein (E), membrane glycoprotein (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid phosphoprotein (N), ORF10, followed by 3' UTR ending in poly(A) tail. However, no evidence that would support the expression of SARS-CoV-2 ORF10 encoded protein of unknown function is yet found in the literature [7].

Possible origins of SARS-CoV-2

SARS-CoV-2 is the seventh zoonotic human coronavirus known up to date, and, along with SARS-CoV and MERS-CoV, is considered to be highly pathogenic and more severe

compared to other, milder symptoms causing, community-acquired human coronaviruses (HCoV-229E, HCoV-OC43, HCoV-HKU1 and HCoV-NL63) [8].

Studies on the origin of novel coronavirus have revealed that complete genomic sequence of SARS-CoV-2 suggests a more close, although not a direct parental, ancestral relationship with bat (~96% overall nucleotide homology with RaTG13[9]) and pangolin coronaviruses (up to ~92% homology, with S protein ACE2 receptor binding domain amino acid sequence being 97.4% identical to SARS-CoV-2 [10]), than to those of humans(~79% and ~50% identity to SARS-CoV and MERS-CoV, respectively [11], and, while bats are already a long-time acknowledged reservoir of SARS-CoV-like β-coronaviruses [12,13], the assumption that pangolins can serve as a natural host for CoVs has been made only recently before the emergence of SARS-CoV-2 [14,15]. Although the current risk of animal-human transmission of COVID-19 is considered low, a number of felines[16], canines[17] and minks[18] worldwide have been reported to be infected with SARS-CoV-2.

SARS-CoV-2 isolate classification

With a steadily growing number of complete hCoV-19 genome sequences, early efforts to classify novel isolates based on their genetic make-up have resulted in numerous proposals of different SARS-CoV-2 isolate classification systems [19][20] [21], some of which (e.g. PANGOLIN lineages) are complementary. However, with more than 84 000 of hCoV-19 genome sequences being available publicly as of now, ongoing efforts to aid in the classification of newly sequenced viral isolates have resulted in the general acceptance of GISAID's team developed SARS-CoV-2 major clade and lineage nomenclature system based on the specific combinations of 9 hCoV-19 genetic markers [1]. In accordance with this system, hCoV-19 isolates can be classified in at least six distinct major clades, namely: S, L (containing reference sequence Wuhan-Hu-1), V, G, GH, GR and O (other) isolate clades (Table 1).

Clade genetic markers	241 bp	3037 bp	23403 bp	8782 bp	11083 bp	25563 bp	26144 bp	28144 bp	28882 bp	Percent of isolates Worldwide (n=81963)	Percent of isolates in Europe (n=48752)	Percent of isolates in Latvia (n=109)
S	C	C	A	T	G	G	G	C	G	6.58	2.96	0.00
L	C	C	A	C	G	G	G	T	G	5.03	5.91	3.67
V	C	C	A	C	T	G	T	T	G	6.08	8.81	0.00
G	T	T	G	C	G	G	G	T	G	23.54	28.24	10.09
GH	T	T	G	C	G	T	G	T	G	22.03	10.10	30.28
GR	T	T	G	C	G	G	G	T	A	31.72	41.88	54.13
Other	X	X	X	X	X	X	X	X	X	5.02	2.10	1.83

Table 1. Major hCoV-19 clades defining genetic markers and their occurrence in Latvia, Europe and Worldwide (as of 2020-08-15). X denotes any nucleotide.

Mid-August, 2020, the most represented clades Worldwide are GR, G and GH, roughly corresponding to 31.72%, 23.54% and 22.03% of total hCoV-19 isolates, respectively. All three of these clades are characterized by C241T base substitution in 5' UTR region, C3037T silent mutation in ORF1a and missense A23403G mutation that causes Aspartic acid at position 614 of spike glycoprotein (S) to change to Glycine (S-D614G), that is associated with higher viral loads and, in turn, is hypothesized to increase the infectivity of these genotypes, with isolates bearing this mutation quickly becoming

dominant ones in various regions throughout the world [22–24]. More recent clades GR and GH are further distinguished from the ancestral G genotype by G25563T mutation resulting in position 57 of ORF3a protein to change from glutamine to histidine for clade GH, and G28882A that changes glycine at nucleocapsid phosphoprotein (N) aa position 204 to arginine for clade GR. While the exact effect of GH clade-defining G25563T change in apoptosis-inducing transmembrane ORF3a protein (Q57R) remains unknown, it does not seem to affect any of the conserved functional domains distinguishable within the protein [25,26]. Whereas, G28882A mutation associated with GR genotype is almost always a trinucleotide mutation of neighboring loci resulting in GGG to AAC change at positions 28881, 28882 and 28883, respectively. This trinucleotide mutation results in two (R203K and G204R) consecutive amino acid changes in N protein, which, in turn, might have potential implications on nucleocapsid phosphoprotein structure and/or function via reduction of conformational entropy and changes in inter-residue interactions in the proximity of the mutated amino acid positions (elaborated on in [27]). The currently estimated mutation rate of SARS-CoV-2 is around 9.86×10^{-4} to 1.85×10^{-4} substitutions per position per year [28], and, based on the isolates sequenced worldwide up to date, there is evidence that mutations in nearly every position in the genome of SARS-CoV-2 have already been documented [29].

In this study, we are reporting the first results of an ongoing massive sequencing campaign that allows us to elaborate on the genetic diversity of SARS-CoV-2 isolates from Latvian patients.

Materials and Methods

Sample management and detection of SARS-CoV-2

For viral genome analysis either oropharyngeal or nasopharyngeal swabs obtained from COVID-19 patients or already extracted RNA samples were provided to Latvian Biomedical Research and Study Centre by the three accredited diagnostic laboratories (E. Gulbis Laboratory, Central Laboratory and Latvian Centre of Infectious Diseases) covering diagnostics of all officially reported cases of SARS-CoV-2. RNA extraction from oropharyngeal and nasopharyngeal swabs and the following SARS-CoV-2 detection was performed by multiple different methods according to standard procedures of each laboratory. These included manual Trizol-based RNA extraction (TRI reagent, Sigma) and automated purification methods with STARMag 96 X 4 Universal Cartridge Kit (Seegene Inc.), NucliSENS easyMAG (bioMérieux), QIAamp 96 Virus QIAcube HT Kit (QIAGEN). The presence of SARS-CoV-2 in the purified RNA samples was estimated by, either commercial or in-house RT-QPCR (Allplex™ 2019-nCoV Assay, Seegene Inc) methods. Samples showing amplification (ct < 40) of at least one viral gene (RdRp, E, N) were considered as positive and directed to next-generation sequencing.

Next-generation sequencing

Metatranscriptome sequencing was the first-choice method for the SARS-CoV-2 genome analysis. Nevertheless, since the majority of samples showed an insufficient number of sequencing reads mapping to the SARS-CoV-2 genome and could not be reliably analyzed, targeted sequencing approaches were considered. A

methodological strategy plan was developed in order to apply the most effective next-generation sequencing method for each sample according to the quantity of SARS-CoV-2 (Supplementary figure 1). At first, RT-QPCR was repeated for each sample in order to evaluate the quantity of viral RNA with a common approach for all samples. Three SARS-CoV-2 genome-specific primer pairs and probes targeting different regions of the nucleocapsid protein (N) gene implemented in the 2019-nCoV RUO Kit (IDT), and SOLIScript® 1-step CoV Kit (Solis Biodyne) were used for the amplification. Probes N1 and N2 specifically detected SARS-CoV-2, while the N3 probe universally detected all currently recognized clade 2 and 3 viruses within the subgenus Sarbecovirus [30]. To evaluate the RNA extraction and PCR efficiency, simultaneous amplification of the human RNase P gene was performed and a control sample containing a plasmid with the SARS-CoV-2 nucleoplasmid protein gene (2019-nCoV_N_Positive Control, IDT) was added to each reaction set. The potential contamination was evaluated by a negative control (nuclease-free water instead of RNA) added to each sample set. RT-QPCR was conducted on the ViiA 7 Real-Time PCR System (Thermo Fisher Scientific), and only the samples showing amplification (ct < 40) of all three SARS-CoV-2 nucleoplasmid protein genes were further directed to metatranscriptome sequencing. Samples exhibiting poor amplification of viral genes (ct > 40 for at least one target region) were considered for one of targeted sequencing approaches: hybridization capture or amplification of SARS-CoV-2.

Metatranscriptome sequencing

In order to eliminate contaminating DNA, DNase I treatment (NEB) of RNA samples was performed, followed by rRNA depletion with MGIEasy rRNA Depletion Kit (MGI Tech Co. Ltd). Complementary DNA libraries were prepared using MGIEasy RNA Library Prep Set (MGI Tech Co. Ltd). Quantity and quality of both RNA and cDNA were evaluated using the Qubit 2.0 fluorometer and Agilent 2100 Bioanalyzer system, respectively. The presence of the SARS-CoV-2 genome was repeatedly tested in each cDNA library by Q-PCR before sequencing, using the same primers and probes (2019-nCoV_N_Positive Control, IDT) together with TaqMan™ Gene Expression Master Mix (Thermo Fisher Scientific). After multiple experimental tests, a ct value threshold of 25 was chosen for N1 and N3 probes for cDNA libraries to be forwarded to metatranscriptome sequencing (N2 probe appeared to be unstable and therefore uninformative). Metatranscriptome cDNA libraries were sequenced on the DNBSEQ-G400RS sequencing platform with DNBSEQ-G400RS High-throughput Sequencing Set (PE150) (MGI Tech Co. Ltd), obtaining at least 100 million 150-bp-paired-end sequencing reads per each sample. Those libraries that failed to pass the ct threshold (ct > 25 for N1 and N3) were directed to a targeted approach.

SARS-CoV-2 hybridization capture

One of the targeted sequencing strategies was based on the enrichment of the SARS-CoV-2 genome by hybridization probes. For cDNA library preparation TruSeq Stranded Total RNA Library Prep Gold kit and TruSeq RNA UD Indexes (Illumina) were used. The indexed cDNA libraries were enriched for the SARS-CoV-2 genome using compatible hybridization probes implemented in the myBaits Expert SARS-CoV-2 kit (Arbor Biosciences) according to manufacturers instructions. The enriched libraries were

sequenced on Illumina MiSeq system with MiSeq Reagent Kit v3 (150-cycle), obtaining at least 1 million of around 75-bp-paired-end reads per sample.

Amplification of SARS-CoV-2 genome

The second targeted approach involved multiplexed primers for amplification of the whole SARS-CoV-2 genome. QIAseq SARS-CoV-2 Primer Panel (QIAGEN) [31] was used together with QIAseq FX DNA Library Kit (QIAGEN) for cDNA library preparation. Next-generation sequencing was performed on Illumina MiSeq system with MiSeq Reagent Kit v2 (300-cycles), obtaining at least 1 million of around 150bp paired-end reads per sample.

Sequencing data quality control, variant calling and data sharing

Adapter clipping was performed with cutadapt 1.16 [32]. Subsequent read trimming was performed with fastp 0.20.0 [33] using 5 base sliding window trimming from both ends with quality threshold 20. Reads with length less than 75 bp or an average quality of less than 20 were removed. Quality controlled reads were then aligned against SARS-CoV-2 isolate Wuhan-Hu-1 reference genome (Accession number: NC_045512.2) with bowtie2 2.3.5.1 [34]. Variant calling and consensus sequence construction were implemented using bcftools 1.10.2 [35]. Average coverage for each of the genomes was calculated using samtools and in-house awk [36] scripts. Less than 1% of the missing bases were allowed for a genome to be considered successfully sequenced and missing bases were treated as reference bases from the Wuhan-Hu-1 genome. Consensus sequences of the successfully sequenced isolates were then proceeded to the manual variant quality inspection by sequence alignment map visualization in IGV [37], sequences that have passed the manual variant quality check were immediately publicly shared by deposition to GISAID database [1]. Variant annotations were performed using coronapp SARS-CoV-2 genome autoannotation web server by comparisons to reference sequence [38] and the results were summarized with the help of custom R scripts, ggplot2 R library was used for plot visualizations [39][40].

Phylogenetic reconstructions

Sequences of the Latvian SARS-CoV-2 isolates and Wuhan-Hu-1 reference sequence were aligned using Clustal-Omega v 1.2.4. [41]. Neighbor-joining tree was generated in MEGA7 [54]. Maximum likelihood phylogeny was performed using IQTREE v 2.0.6 [42] with TIM2+F+I as best fit model determined by ModelFinder [43] according to Bayesian Information Criterion (ultrafast bootstrap with 1000 replicates [44]) and assessment of temporal signal associated with the data was performed by importing resulting ML tree into TempEst v1.5.3 [45], parsing sampling dates of isolates and visualizing the root-to-tip divergence.

Bayesian phylogenetic trees were estimated using BEAST v1.10.4 [46], employing GTR nucleotide substitution model with empirical base frequencies and invariant site proportion assuming strict molecular clock. Taxon sets corresponding to samples belonging to clades L, G, GR and GH were set to be monophyletic. Coalescent

exponential growth prior (growth rate prior: Laplace with scale 100; population size prior: Lognormal with mu 1 and sigma 2) with growth rate parametrization [47,48] was selected and Markov chain Monte Carlo (MCMC) was run for 30 million states sampling log parameters and trees every 3 000 states. Tracer v 1.7.1 [49] was used for MCMC trace (log file) inspection to evaluate sufficiency of sampling (all parameters had an ESS of more than 500) and infer substitution rate and date of the most recent common ancestor estimates. To summarize Bayesian phylogenetic inference, maximum clade credibility time-scaled tree was generated in TreeAnnotator v 1.10.4 (distributed with BEAST package) using 10% of the states (3 million) as the burn-in and visualized using FigTree v 1.4.4 [50].

Results and discussion

With 1330 cumulative positive cases as of 20th of August 2020 (1093 people recovered, 204 active cases of the disease and 33 COVID-19 associated deaths), 109 hCoV-19 isolates representing ~8.2% of the total local COVID-19 cases have been completely sequenced as of today, making Latvia one of the leading countries not only in regards to the containment of the spread of COVID-19 disease, but in the number of sequenced SARS-CoV-2 isolates to the cumulative number of positive COVID-19 cases ratio as well (Figure 1).

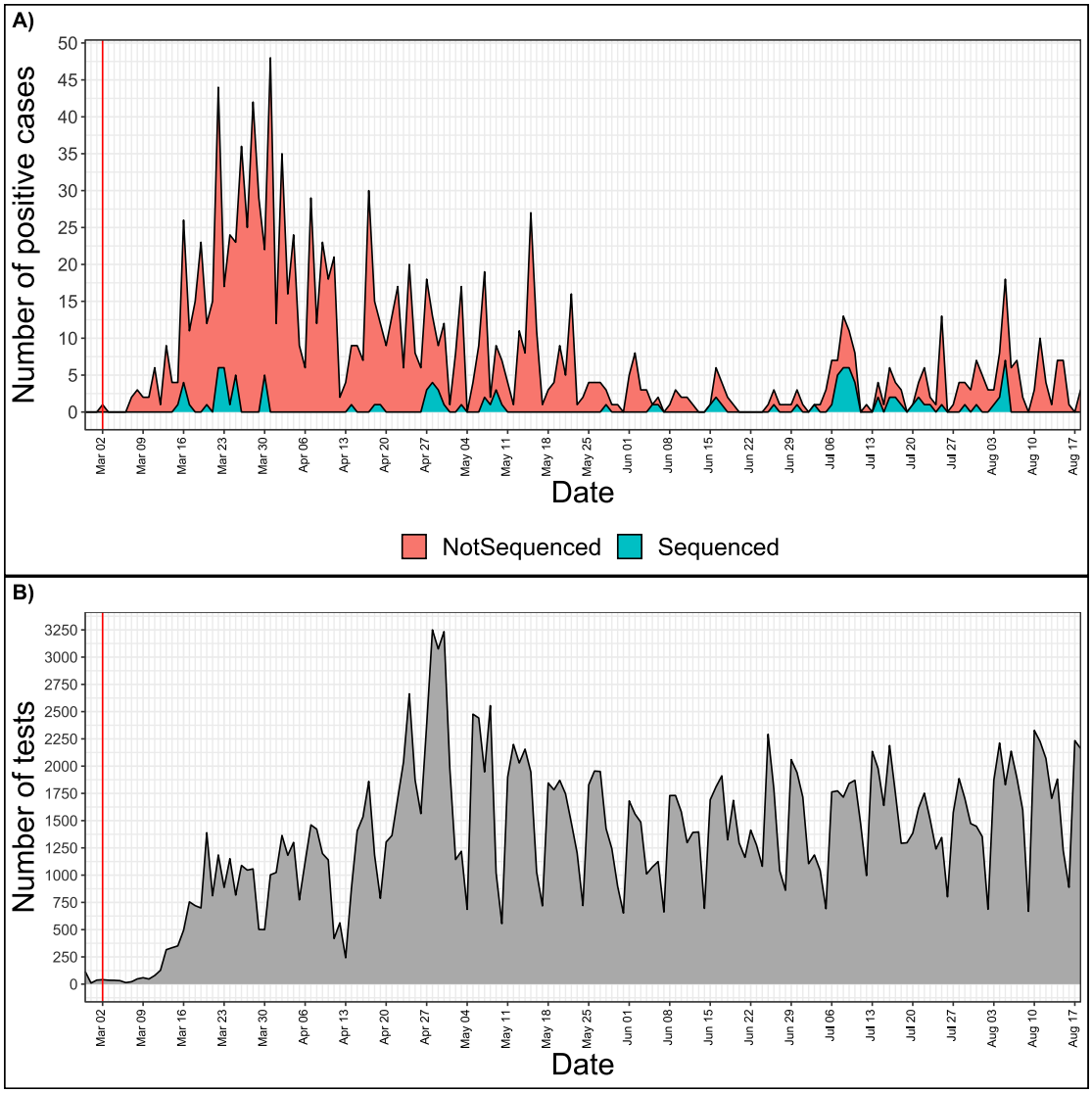


Figure 1. Daily numbers **of** positive COVID-19 cases (A) and tests performed (B) in Latvia. X axis is the same for both tiles and represents daily time series from 28th of February, 2020 to 18th of August, 2020. The red vertical line indicates the date of the first COVID-19 case registered in Latvia. A) Y value represents the total number of positive cases registered on a given day. Blue area shows the number of successfully sequenced isolates, while the red area represents the positive cases not sequenced during this study. B) Y value represents the number of tests carried out on a given date in Latvia.

Reacting to the emergence of the SARS-CoV-2 in Latvia, a high-throughput framework for SARS-CoV-2 isolate sequencing and data analysis with capabilities of near real-time tracking of the epidemiological situation in Latvia was built to aid the governmental decision-making and study the molecular epidemiology of hCoV19.

One of the challenges to obtain good quality sequences for maximal number of samples is the variable quality of input material that can be caused by highly variable viral loads, different collection, storage and RNA isolation methods. Although for the current study we did not have the information on the severity of COVID-19 symptoms for particular cases, it should be noted that the absolute majority of cases in Latvia are with low symptom severity expected to have lower concentration of virus in diagnostic samples. We therefore developed an approach to verify sample quality and select appropriate sequencing method to recover maximal available information from

existing samples ensuring cost efficacy of the process (Figure S1). According to this strategy developed during the implementation of the study, complete SARS-CoV-2 genome sequence was successfully obtained by metatranscriptome approach for 37 viral isolates, 56 samples were analyzed by amplification of SARS-CoV-2 genome with multiplexed primers, and for 16 isolates enrichment of SARS-CoV-2 genome was performed by hybridization capture method prior the sequencing.

As of now it could be cautiously speculated that the obtained results on the SARS-CoV-2 genotype distributions might be somewhat representative of a whole Baltics region, taking the geographical proximity, travel habits and mild governmental travel regulations between the Baltic states during the most of the pandemic into the account. However, the extent of similarity between the isolates circulating in different Baltic states currently cannot be reliably established due to SARS-CoV-2 isolate undersequencing in neigboring Estonia and Lithuania, and the founder effect of multiple independent (re-)introductions of different SARS-CoV-2 genotypes, as well as containment effectivity of respective COVID19 cases, in each of the countries should not be overlooked.

Distribution of sequenced virus isolates by SARS-CoV-2 clades

Major isolate clade distributions across distinct geographical regions show clear spatial differences of the epidemic (Figure 2) and a trend of “older” isolate clades L and S losing their initial prevalence to the dominance of the more recently-emerged G-associated clades (G, GH, GR) that seem to be accountable for the majority of the cases worldwide since the middle of March 2020. GR, which is the most common isolate clade in Latvia (54.13% of cases), is also a dominant clade in Europe and South America. Currently, GH still seems to be the most common isolate clade circulating throughout North America, but a rise in the number of GR isolates can be observed since the middle of May 2020. The prevalence of GR and, in particular, G clade isolates is also currently on the rise in Africa, and, to a very moderate amount in Oceania and Asia. The relatively high number of isolates not corresponding to any of the currently recognized major SARS-CoV-2 clades (dubbed “Other” or belonging to the “O” clade as of now) in Asia and Oceania makes it possible to speculate about it being indicative of either (but not mutually exclusive), poor quality of the sequences obtained or the possibility of novel clade emergence originating from these regions in the future, should their spread not be effectively contained.

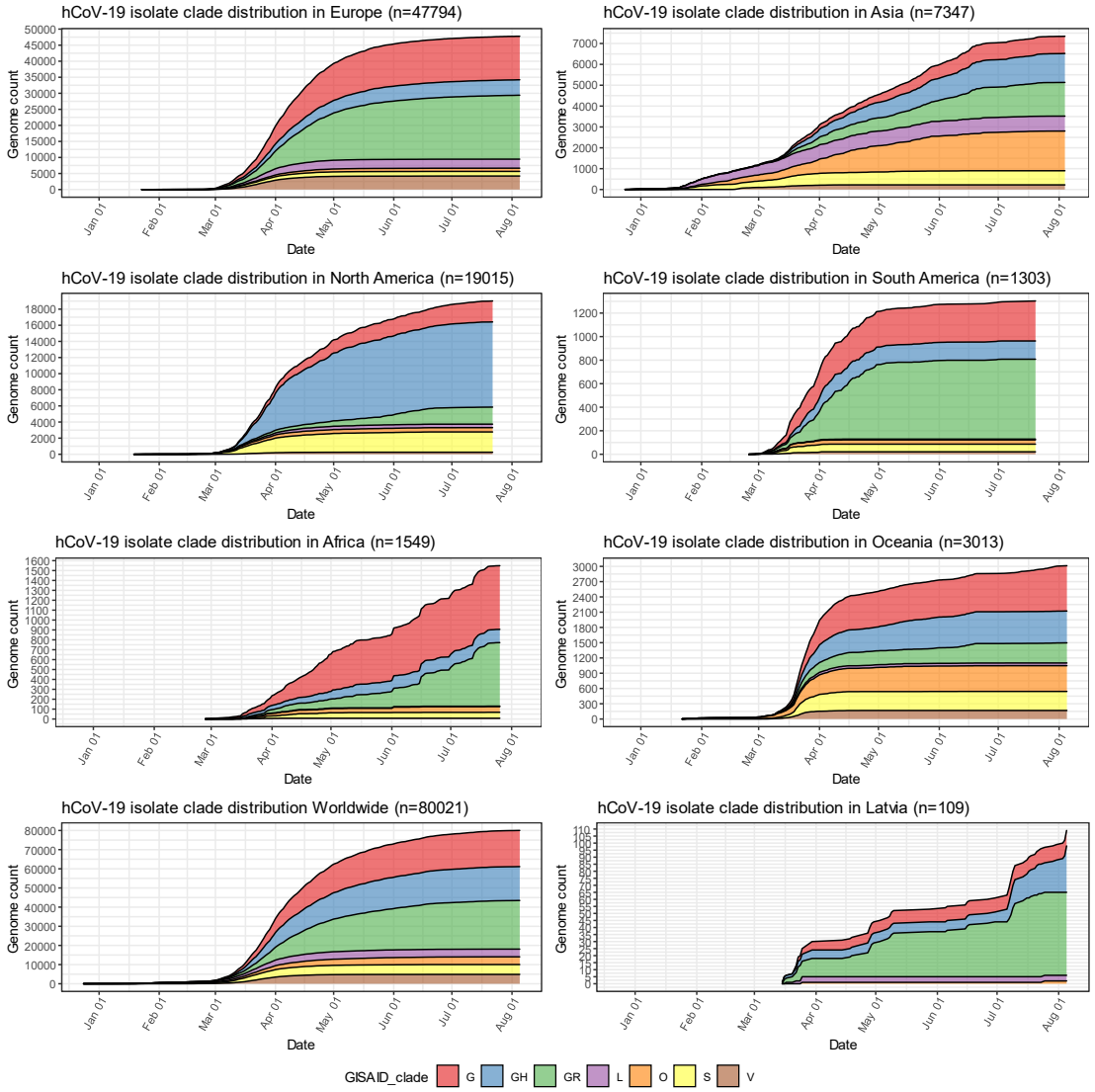


Figure 2. Distribution of sequenced hCoV-19 isolates by clades in major regions of the world, worldwide and in Latvia. Y axis depicts cumulative complete hCoV-19 genome count (with unambiguous collection date) from a particular region and has different scale within the subplots. X axis is the same for all subplots and depicts sampling time-series from 24th of December, 2019 till 5th of August, 2020.

Mutational landscape of Latvian SARS-CoV-2 isolates

After joining of the neighboring loci, among 109 local isolates, 188 different unique mutational events (125 nonsynonymous, 57 synonymous, 6 substitutions in extragenic regions, and a single deletion) that affected 186 positions of the SARS-CoV-2 genome were registered from a total of 1009 variants that were identified. 113 out of 188 distinct mutational events were registered only in one of the 109 samples, while 75 were present in two or more samples (Supplementary table 1 and supplementary figure 2). NSP3 was found to be the mature peptide most frequently affected by non-synonymous substitutions (24 distinct variants resulting in an amino acid change), followed by an N protein that had 15 non-synonymous SNVs documented among Latvian SARS-CoV-2 isolates. Among the most frequently mutated proteins, NSP2, S and NSP12b mature peptides harbored 13, 12 and 10 different amino acid altering mutations, respectively.

Based on the current coronapp web-server [38] report updated at 2020-08-14 (n=73889), most frequent mutational events worldwide are: A23403G corresponding to S:D614G, C3037T silent mutation, C14408T resulting in NSP12b:P314L, C241T extragenic substitution and GGG28881ACC trinucleotide mutation of neighboring loci resulting in N:RG203KR, G25563T – ORF3a:Q57H. All five of these mutations were also among most frequent mutational events registered in Latvian samples: 5' UTR C241T extragenic substitution that was present in 105 out of 109 sequenced genomes, while C3037T silent (NSP3:F106F) mutation, A23403G (S:D614G) and C14408T (NSP12b:P314L) were all present in 104/109 samples, GGG28881AAC trinucleotide mutation (N:RG203KR) was observed almost in half (54/109) of the samples, and 28881 position of the genome had two more variants detectable in the samples – GGGG28881AACT (N:RG203KL) quadranucleotide mutation (24/109) and G28881A (N:R203K) substitution being present in five of the samples (See figure 3 and supplementary table 1).

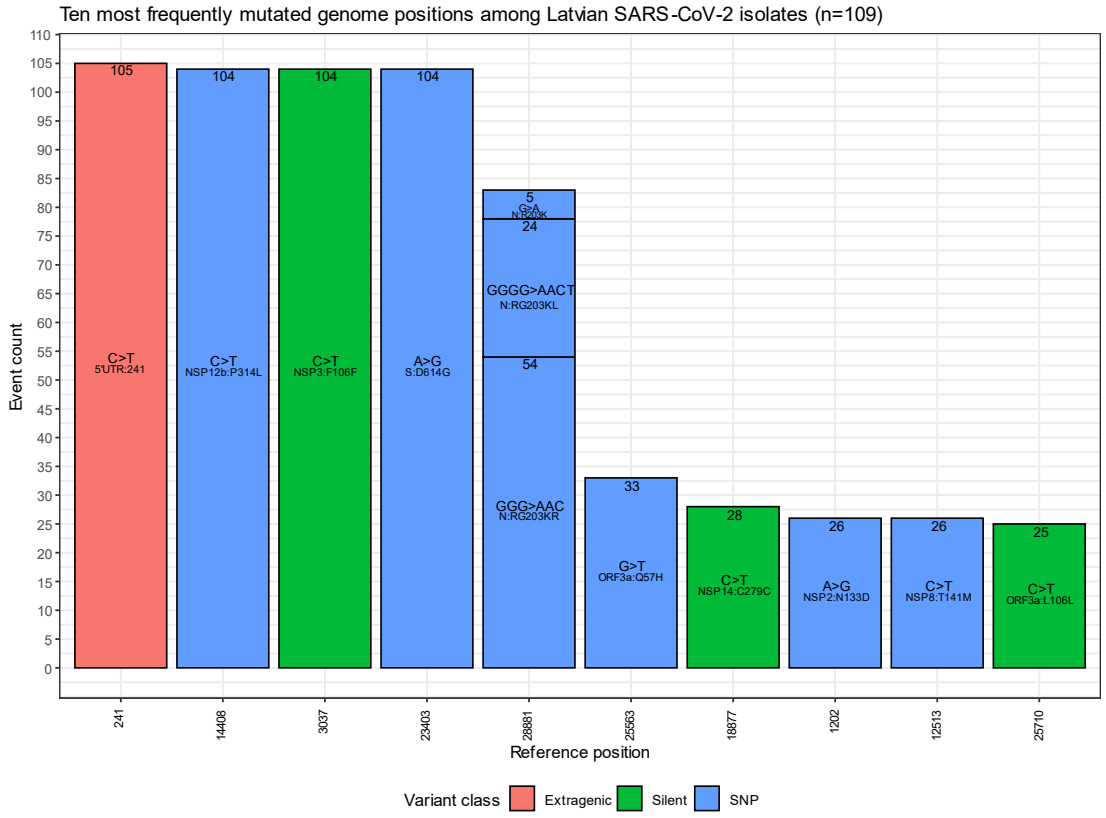


Figure 3. Ten most frequently mutated genome positions among Latvian SARS-CoV-2 isolates (n=109). X axis is discrete and shows genome position corresponding to one of the ten most frequent mutated positions (ordered in descending order). Y axis represents mutation occurrence among the samples (same as numbers within the respective upper boundary of bars) at a given position. Labels in the middle of bars represent the nucleotide change at a given position and its effect on the respective protein amino acid sequence. Color coding is based on the variant class and bars are colored according to the legend. Note: bars representing different mutations of the same locus are stacked (position 28881).

It was noted that four of out of five aforementioned mutations (with the exclusion of C14408T) are in the genome positions serving as markers for current SARS-CoV-2 isolate major clade definition and correspond to GR clade, that is the most represented clade Worldwide and hosts more than a half of the sequenced isolates in

Latvia (Figure 3 and table 1). The C14408T substitution resulting in NSP12b:P314L amino acid change has been previously reported to co-occur with C241T, C3037T and A23403G mutations [51], which is consistent with our data, where four of these SNPs were simultaneously present 104/109 of the Latvian SARS-CoV-2 isolates sequenced up to date. While no experimental evidence of C14408T substitution implications on the NSP12b (RdRp) activity is yet present, isolates bearing this variant were previously speculated to have more mutations, and elaborations about possible implications of RdRp mutations on antiviral drug resistance were made [52]. The fitness of G and G-derived strains, as denoted by the recent rise of their prevalence throughout different regions of the world, is hardly explainable only by the founder effect alone, thus highlighting the fact that further evidence on molecular and clinical implications of the most common substitutions in the genomes of currently circulating SARS-CoV-2 is urgently needed to improve the measures of containment of COVID-19 and develop effective antiviral therapies and vaccines, that would help to not only combat the present virus of immediate concern, but also be of vital importance for other coronaviruses to yet emerge.

Phylogenetic analyses

Root-to-tip regression analysis with the "best-fitting root" and "correlation" function options resulted in a correlation coefficient of the analysis being estimated at 0.6098 and a determination coefficient (R^2) equaling to 0.3718 (Supplementary Figure 3). Although having some of the sequences that diverged more or less than expected at their sampling date, the dataset had a moderate association between sequence divergence and sampling date, implying suitability for phylogenetic molecular clock analysis.

Following Bayesian phylogenetic inference, mean mutation rate derived from Latvian SARS-CoV-2 isolates was found to be 7.126×10^{-4} substitutions per site per year (5.5086×10^{-4} - 8.7086×10^{-4}, 95% highest posterior density interval), roughly corresponding to an average of 21 mutational events in genome per year (95% HPD: ~16 – ~26), and lies within mutation rate ranges predicted by other researchers. Based on the analysis, the estimated most recent common ancestor of the isolates has emerged on 18th of November, 2019 (7th October, 2019 – 25th December, 2019, 95% interval). Our molecular clock analysis further supported the more recent divergence of clades GR and GH with the most common ancestor for both clades dating back to 6th of February, 2020 (95% HPD GR: 12th of January, 2020 – 26th of February, 2020; 95% HPD GH: 12th of January, 2020 – 29th of February, 2020). The 95% HPD date ranges are consistent with the collection dates of unambiguously dated genomes belonging to clades GH and GR deposited at GISAID (accessed 2020-08-19). Earliest reported SARS-CoV-2 genome belonging to clade GH was collected on 27th of January 2020 in Cardiff, Whales (GISAID accession: EPI_ISL_474597), while earliest reported GR clade genome was collected on 16th of February 2020 in London, England (GISAID accession: EPI_ISL_466615). While the TMCRA of GR and GH ancestral clade G was dated to 25th of January, 2020 (95% HPD: 24th of December, 2019 – 20th of February, 2020) and earliest reported sequences with unambiguous collection date belonging to clade G were collected on 24th of January, 2020 in China, cities of Zhejiang and Chengdu (GISAID accessions: EPI_ISL_422425, EPI_ISL_451345).

Our phylogenetic analysis of the local isolates suggests multiple unlinked initial introductions of already divergent SARS-CoV-2 isolates to Latvia. Just two weeks after the first positive case of COVID-19 was documented in Latvia on the 2nd of March, isolates representing at least three major SARS-CoV-2 clades (L, GR and GH) were already circulating within the country corresponding to at least four epidemiologically unlinked introductions. No isolates belonging to clade L (most similar to the initial Wuhan-Hu-1 reference) were sequenced after the end of March and local circulation of clade G representatives was not detectable since early March, while clade GH and, specifically, GR isolates seem to have taken hold of the epidemic without showing signs of ceasing their proliferation within the Latvian population, however, recent reintroduction event possibility should not be ruled out due to cancelation of travel restrictions and insufficient testing of those entering the country. With more than half of the sequenced isolates belonging to the widely represented GR clade, up to this date, no isolates representing clades V or S were documented among the sequenced Latvian COVID-19 cases (Figures 2 and 4).

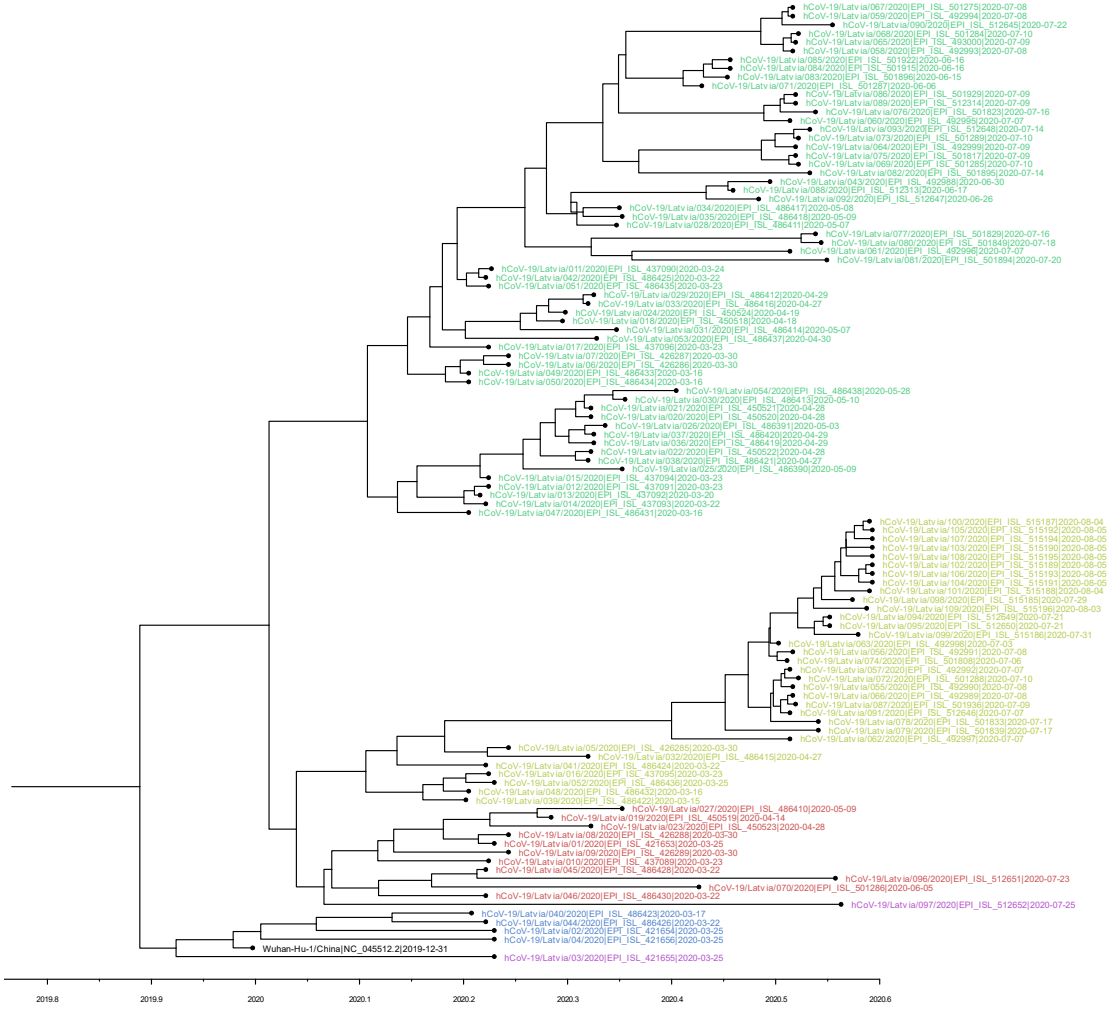


Figure 4. Maximum clade credibility tree estimated from completely sequenced Latvian isolates (n=109) and Wuhan-Hu-1. Node labels are colored according to the GISAID major clade of particular isolate, green – GR, yellow – GH, red – G, blue – L, purple – O (other), black – Wuhan-Hu-1 reference sequence. The tree is time-scaled and values next to nodes indicate node dates in years (one months is ~0.08333 of a year and one day is approximately 0.00274 of a year).

Neighbor-joining phylogenetic tree was built to more apparently infer genetic distances between the samples (Figure 5). Although of satisfactory topology, supporting major clade clustering, the tree evidently shows the possible discrepancies between the reported sampling dates and expected sequence divergence (e.g. some of the samples most divergent from the root are dated with the end of April, while some of the most recent ones are twice less divergent), which is not attributable to sequencing errors or the possibility of co-infection by two different “strains”. Identical sequences sampled within a short date range (Figure 5.) might be strongly indicative of epidemiologically linked transmission, given the relatively small daily amount of positive COVID-19 cases in Latvia even during the peaks of the disease spread.

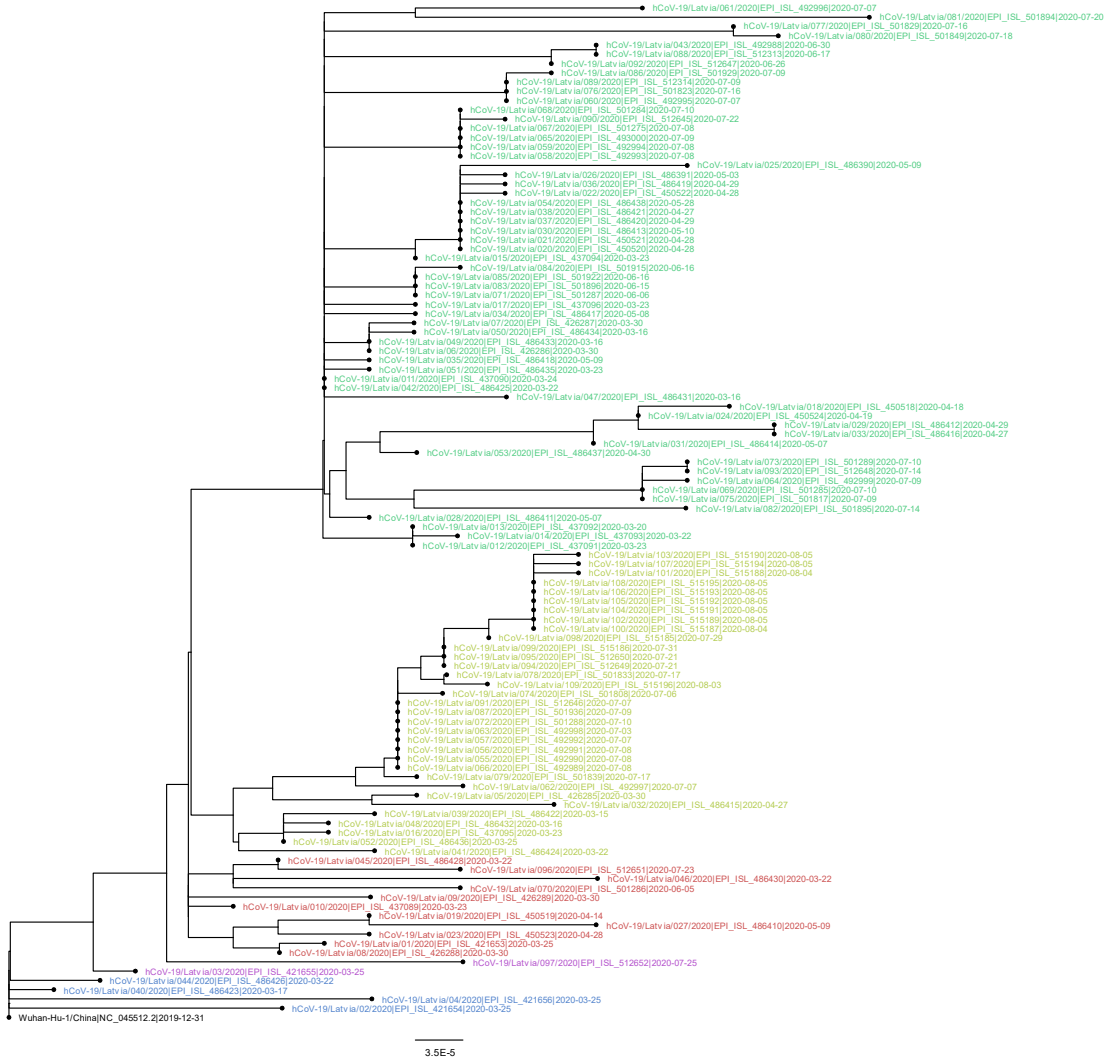


Figure 5. Evolutionary relationships of 109 sequenced Latvian and Wuhan-Hu-1 SARS-CoV-2 isolates. The evolutionary history was inferred using the Neighbor-Joining method [53]. The tree is rooted at Wuhan-Hu-1 reference sequence. The optimal tree with the sum of branch length = 0.00668471 is shown. The evolutionary distances were computed using the p-distance method [54] and are in the units of the number of base differences per sequence. The tree is drawn to scale, branch lengths correspond to nucleotide substitutions per site. The analysis involved 110 nucleotide sequences (109 Latvian SARS-CoV-2 isolates and Wuhan-Hu-1 reference sequence. There were a total of 29903 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [55]. Node labels are colored according to the GISAID major clade of particular isolate, green – GR, yellow – GH, red – G, blue – L, purple – O (other), black – Wuhan-Hu-1 reference sequence.

While providing interesting insights on the COVID-19 situation in Latvia, which might be representative of Baltics region to an extent, given the scarce amount of isolate genomes available from neighboring countries, it, however, should be noted, that the main drawback for each of the presented analyses is stemming from the available dataset - discrete early sampling with some of the dates since first positive case not being sampled at all (Figure 1). Another major drawback is the unavailability of patient/isolate epidemiological data that could be linked to the respective cases sequenced (e.g. sequence epidemiological linkage, patient travel history etc.) in the frame of this study. Inclusion of additional data and retrospective sequencing of a larger number of cases that would allow for a more in-depth analysis of epidemiological situation will be performed and published elsewhere.

In conclusion, the high-throughput framework for SARS-CoV-2 isolate sequencing and data analysis in Latvia has been built by Latvian Biomedical Research and Study Centre early on during the start of the pandemic, tested with the help of both, governmental and local private laboratory sample providers, and proposed as a pivotal tool to monitor the local outbreaks and aid in decision making. This framework allowed us to ensure the sequencing of viral isolates from almost all new cases starting from the beginning of July, 2020 with fast date delivery to the Centre for Disease Prevention and Control in Latvia allowing to link the epidemiological data to the isolates being sequenced. We believe that this framework is of vital importance for rapid implementation of the most suitable public health measures, possible transmission history deduction and viral evolution monitoring for the prevention of future epidemiological outbreaks.

Acknowledgements

The authors acknowledge the use of infrastructure provided by High Performance Computing Centre of Riga Technical University.

Funding

This research is funded by the Ministry of Education and Science, Republic of Latvia, project "Establishment of COVID-19 related biobank and integrated platform for research data in Latvia", project No. VPP-COVID-2020/1-0016, project "Multidisciplinary approach to monitor, mitigate and contain COVID19 and other future epidemics in Latvia", project No. VPP-COVID-2020/1-0008 and 04.06.2020. service contract No.23-11.3e/20/64 "Study on the epidemiology and phylogenesis of SARS-CoV-2 virus in Latvia" within the framework of ERDF project No.1.1.1.5 / 17 / I / 002 “Integrated national level measures for strengthening interest representations for research and development of Latvia as part of European Research Area”.

Ethics statement

The study was approved by the Central Medical Ethics Committee of Latvia (protocol No. 01-29.1/2429 and 01-29.1/1677).

REFERENCES

1. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance. 2017;22(13):2–4.

2. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Challenges. 2017;1(1):33–46.

3. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.

4. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med. 2020;382(13):1199–207.

5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497–506.

6. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;47(D1):D94–9.

7. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell [Internet]. 2020;181(4):914-921.e10. Available from: https://doi.org/10.1016/j.cell.2020.04.011

8. Ye ZW, Yuan S, Yuen KS, Fung SY, Chan CP, Jin DY. Zoonotic origins of human coronaviruses. Int J Biol Sci. 2020;16(10):1686–97.

9. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang R Di, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature [Internet]. 2020;579(7798):270–3. Available from: http://dx.doi.org/10.1038/s41586-020-2012-7

10. Lam TTY, Jia N, Zhang YW, Shum MHH, Jiang JF, Zhu HC, Tong YG, Shi YX, Ni XB, Liao YS, Li WJ, Jiang BG, Wei W, Yuan TT, Zheng K, Cui XM, Li J, Pei GQ, Qiang X, Cheung WYM, Li LF, Sun FF, Qin S, Huang JC, Leung GM, Holmes EC, Hu YL, Guan Y, Cao WC. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature [Internet]. 2020;583(February). Available from: http://dx.doi.org/10.1038/s41586-020-2169-0

11. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet [Internet]. 2020;395(10224):565–74. Available from: http://dx.doi.org/10.1016/S0140-6736(20)30251-8

12. Yang L, Wu Z, Ren X, Yang F, He G, Zhang J, Dong J, Sun L, Zhu Y, Du J, Zhang S, Jin Q. Novel SARS-like betacoronaviruses in bats, China, 2011. Emerg Infect Dis. 2013;19(6):989–91.

13. Epstein JH, McEachern J, Zhang J, Daszak P, Wang H, Field H, Li W, Eaton BT, Wang L-F, Yu M, Hu Z, Zhang S, Shi Z, Crameri G, Zhang H, Ren W, Smith C. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. Science (80-) [Internet]. 2005;310(5748):676–9. Available from: http://dialnet.unirioja.es/servlet/extart?codigo=1323722

14. Liu P, Jiang J-Z, Wan X-F, Hua Y, Li L, Zhou J, Wang X, Hou F, Chen J, Zou J, Chen J. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLOS Pathog [Internet]. 2020 May 14;16(5):e1008421. Available from: https://doi.org/10.1371/journal.ppat.1008421

15. Liu P, Chen W, Chen JP. Viral metagenomics revealed sendai virus and coronavirus infection of malayan pangolins (manis javanica). Viruses. 2019;11(11).

16. Halfmann PJ, Hatta M, Chiba S, Maemura T, Fan S, Takeda M, Kinoshita N, Hattori S, Sakai-Tagawa Y, Iwatsuki-Horimoto K, Imai M, Kawaoka Y. Transmission of SARS-CoV-2 in Domestic Cats. N Engl J Med. 2020;

17. Sit THC, Brackman CJ, Ip SM, Tam KWS, Law PYT, To EMW, Yu VYT, Sims LD, Tsang DNC, Chu DKW, Perera RAPM, Poon LLM, Peiris M. Infection of dogs with SARS-CoV-2. Nature [Internet]. 2020; Available from: http://dx.doi.org/10.1038/s41586-020-2334-5

18. Oreshkova N, Moelnaar RJ, Vreman S, Harders F, Munnink BBO, Van Der Honin RWH, Gerhards N, Tolsma P, Bouwstra R, Sikkema RS, Tacken MGJ, Rooij MMT De, Weesendorp E, Engelsma MY, Bruschke CJ, Smit LA, Koopman M, Van der Poel WH, Stegeman A. SARS-CoV-2 infection in farmed minks, the Netherlands, April and May 2020. Euro Surveill [Internet]. 2020;25 (23)(May):1–7. Available from: https://doi.org/10.2807/1560-7917.ES.2020.25.23.2001005

19. Emma B H, Hadfield J, Richard A N, Bedford T. Year-letter Genetic Clade Naming for SARS-CoV-2 on Nextstain.org [Internet]. 2020 [cited 2020 Jul 17]. Available from: https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming

20. Zhao Z, Sokhansanj BA. Genetic Grouping of SARS-CoV-2 Coronavirus Sequences using Informative Subtype Markers for Pandemic Spread Visualization. 2020;

21. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol [Internet]. 2020; Available from: https://doi.org/10.1038/s41564-020-0770-5

22. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Evans CM, Freeman TM, Silva TI De, Mcdanal C, Perez LG, Tang H, Whelan SP, Labranche CC, Saphire EO, Montefiori DC, Silva D. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020;

23. Lorenzo-Redondo R, Nam HH, Roberts SC, Simons LM, Jennings LJ, Qi C, Achenbach CJ, Hauser AR, Ison MG, Hultquist JF, Ozer EA. A Unique Clade of SARS-CoV-2 Viruses is Associated with Lower Viral Loads in Patient Upper Airways. medRxiv Prepr Serv Heal Sci [Internet]. 2020 May 26;2020.05.19.20107144. Available from: https://pubmed.ncbi.nlm.nih.gov/32511558

24. Hu J, He C-L, Gao Q-Z, Zhang G-J, Cao X-X, Long Q-X, Deng H-J, Huang L-Y, Chen J, Wang K, Tang N, Huang A-L. The D614G mutation of SARS-CoV-2 spike protein enhances viral infectivity and decreases neutralization sensitivity to individual convalescent sera. bioRxiv [Internet]. 2020 Jan 1;2020.06.20.161323. Available from: http://biorxiv.org/content/early/2020/06/20/2020.06.20.161323.abstract

25. Ugurel OM, Ata O, Turgut-Balik D. An updated analysis of variations in SARS-CoV-2 genome. Turkish J Biol. 2020;44(Special issue 1):157–67.

26. Ren Y, Shu T, Wu D, Mu J, Wang C, Huang M, Han Y, Zhang XY, Zhou W, Qiu Y, Zhou X. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. Cell Mol Immunol. 2020;(May):8–10.

27. Maitra A, Sarkar MC, Raheja H, Biswas NK, Chakraborti S, Singh AK, Ghosh S, Sarkar S, Patra S, Mondal RK, Ghosh T, Chatterjee A, Banu H, Majumdar A, Chinnaswamy S, Srinivasan N, Dutta S, Das S. Mutations in SARS-CoV-2 viral RNA identified in Eastern India: Possible implications for the ongoing outbreak in India and impact on viral structure and host susceptibility. J Biosci. 2020;45(1):1–18.

28. Koyama T, Platt DE, Parida L. Variant analysis of COVID-19 genomes. J Bull World Heal Organ [Internet]. 2020;2(March):1–21. Available from: https://www.researchgate.net/publication/339461351_Variant_analysis_of_COVID-19_genomes

29. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. NextStrain: Real-time tracking of pathogen evolution. Bioinformatics. 2018;34(23):4121–3.

30. Lu X, Wang L, Sakthivel SK, Whitaker B, Murray J, Kamili S, Lynch B, Malapati L, Burke SA, Harcourt J, Tamin A, Thornburg NJ, Villanueva JM, Lindstrom S. US CDC Real-Time Reverse Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis. 2020;26(8).

31. Itokawa K, Sekizuka T, Hashino M, Tanaka R, Kuroda M. Disentangling primer interactions improves SARS-CoV-2 genome sequencing by the ARTIC Network{\textquoteright}s multiplex PCR. bioRxiv [Internet]. 2020; Available from: https://www.biorxiv.org/content/early/2020/06/01/2020.03.10.985150

32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;

33. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.

34. Langmead B, Salzberg S. Bowtie2. Nat Methods [Internet]. 2013;9(4):357–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/pdf/nihms-366740.pdf

35. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;27(21):2987–93.

36. Aho A V., Kernighan BW, Weinberger PJ. Awk — a pattern scanning and processing language. Softw Pract Exp. 1979;9(4):267–79.

37. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol [Internet]. 2011 Jan;29(1):24–6. Available from: https://pubmed.ncbi.nlm.nih.gov/21221095

38. Mercatelli D, Triboli L, Fornasari E, Ray F, Giorgi F. coronapp: A Web Application to Annotate and Monitor SARS-CoV-2 Mutations [Internet]. bioRxiv; 2020. Available from: https://doi.org/10.1101/2020.05.31.124966

39. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. 2nd ed. Springer Publishing Company, Incorporated; 2009.

40. R Studio Team. R Studio. R.S. ed. http://www.rstudio.com/. 2015.

41. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7(539).

42. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;

43. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;

44. Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;

45. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;

46. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29(8):1969–73.

47. Griffiths RC, Tavaré S. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci. 1994;344(1310):403–10.

48. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002;161(3):1307–20.

49. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67(5):901–4.

50. Rambaut A. FigTree v. 1.4.4. http://tree.bio.ed.ac.uk/software/figtree/. 2018.

51. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. Genomics [Internet]. 2020;112(5):3588–96. Available from: https://doi.org/10.1016/j.ygeno.2020.04.016

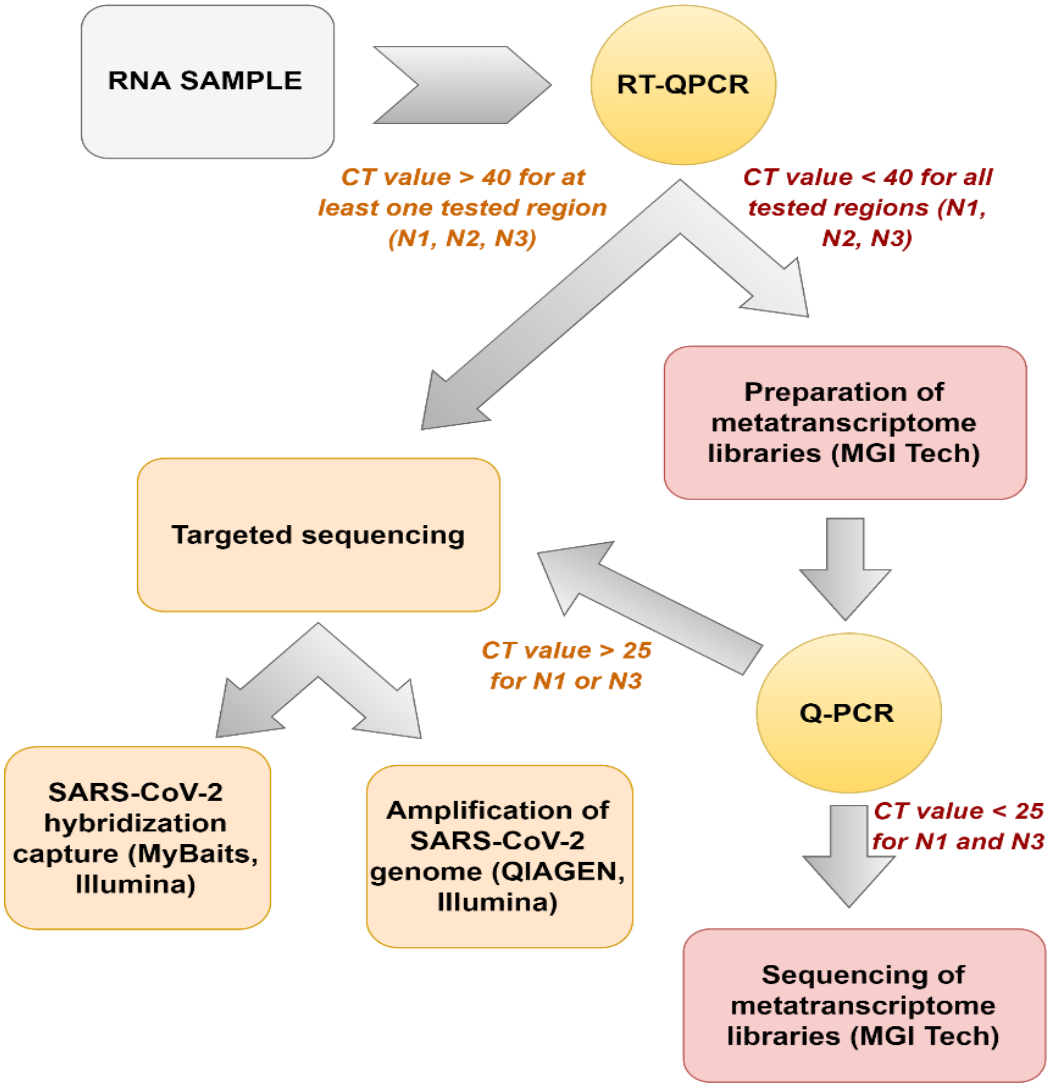
52. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D, Ippodrino R. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. J Transl Med [Internet]. 2020;18(1):1–9. Available from: https://doi.org/10.1186/s12967-020-02344-6

53. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.

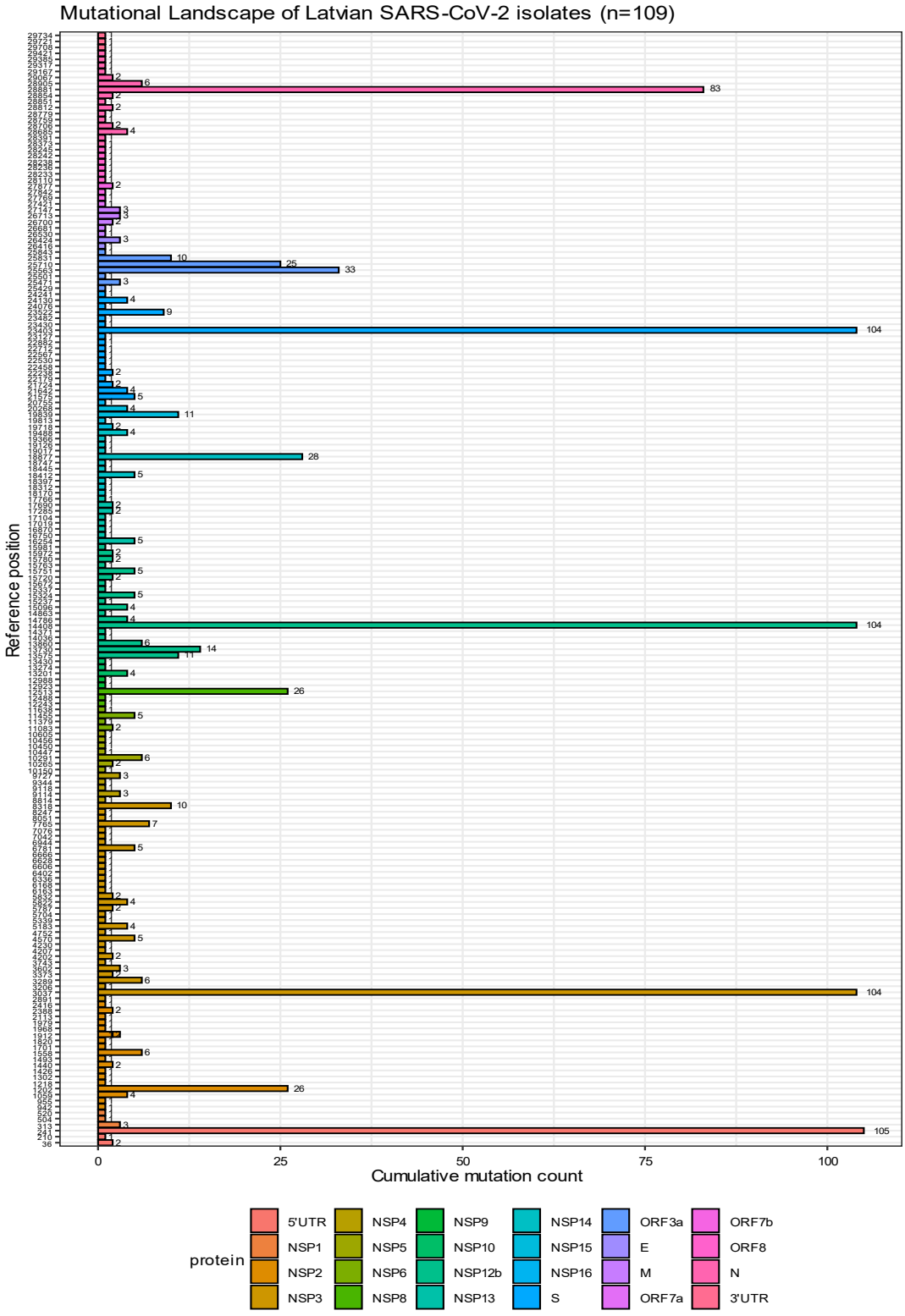
54. Nei M, Kumar S. Molecular Evolution and Phylogenetics. New York: Oxford University Press; 2000.

55. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33(7):1870–4.

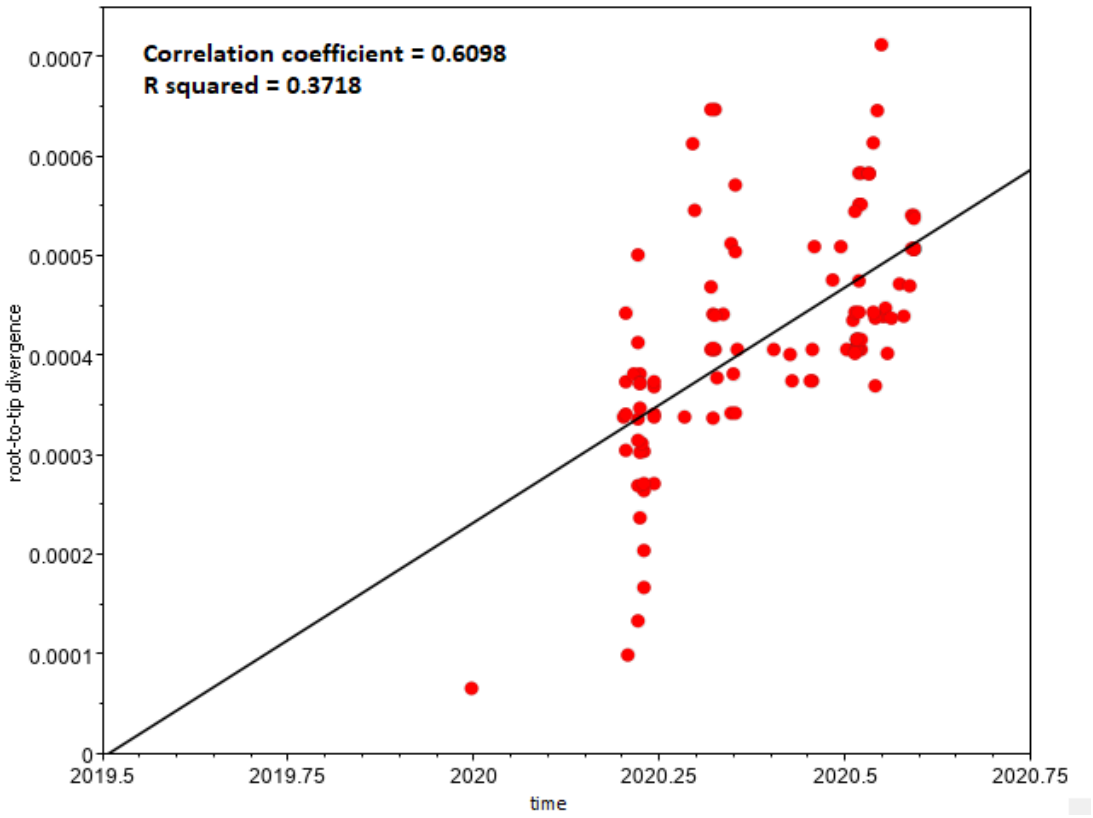
SUPPLEMENTARY MATERIALS



Supplementary figure 1. Methodological strategy plan for SARS-CoV-2 genome analysis based on different next-generation sequencing methods.



Supplementary figure 2. Mutational landscape of Latvian SARS-CoV-2 isolates. Y axis shows the mutated position of a reference SARS-CoV-2 genome. X axis shows the cumulative mutation count at a given position. Number to the right of bars indicate cumulative mutation count at a given position and bars are color-coded according to the protein that the corresponding site participates in encoding. Note, that Y axis is discrete and only positions with mutations documented in local isolates are shown.



Supplementary figure 3. Root-to-tip regression analysis of 109 Latvian SARS-CoV-2 isolates and Wuhan-Hu-1 sequence.

Supplementary table 1. Annotation and occurrence of variants identifiable among 109 Latvian SARS-CoV-2 isolates sequenced (sorted by occurrence).

Variant	Reference	Position in Genome	Variant	Variant class	Region affected	Amino acid change	Function of mature peptide	Variant frequency	Occurrence among Latvian isolates
5'UTR:241	C	241	T	Extragenic	5'UTR	N/A	N/A	105	96.33%
NSP3:F106F	C	3037	T	Silent	NSP3	F106F	Predicted phosphoesterase, papain-like proteinase	104	95.41%
NSP12b:P314L	C	14408	T	SNP	NSP12b	P314L	RNA-dependent RNA polymerase, post-ribosomal frameshift	104	95.41%
S:D614G	A	23403	G	SNP	S	D614G	Spike	104	95.41%
N:RG203KR	GGG	28881	AAC	SNP	N	RG203KR	Nucleocapsid protein	54	49.54%
ORF3a:Q57H	G	25563	T	SNP	ORF3a	Q57H	ORF3a protein	33	30.28%
NSP14:C279C	C	18877	T	Silent	NSP14	C279C	3'-to-5' exonuclease	28	25.69%
NSP2:N133D	A	1202	G	SNP	NSP2	N133D	Non-Structural protein 2	26	23.85%
NSP8:T141M	C	12513	T	SNP	NSP8	T141M	Non-Structural Protein 8	26	23.85%
ORF3a:L106L	C	25710	T	Silent	ORF3a	L106L	ORF3a protein	25	22.94%
N:R203K	G	28881	A	SNP	N	R203K	Nucleocapsid protein	24	22.02%
NSP12b:A88V	C	13730	T	SNP	NSP12b	A88V	RNA-dependent RNA polymerase, post-ribosomal frameshift	14	12.84%
NSP12b:A37S	TG	13575	CT	SNP	NSP12b	A37S	RNA-dependent RNA polymerase, post-ribosomal frameshift	11	10.09%
NSP15:N73N	T	19839	C	Silent	NSP15	N73N	endornase	11	10.09%
NSP3:P1867S	C	8318	T	SNP	NSP3	P1867S	Predicted phosphoesterase, papain-like proteinase	10	9.17%
ORF3a:L147F	C	25831	T	SNP	ORF3a	L147F	ORF3a protein	10	9.17%
S:E654K	G	23522	A	SNP	S	E654K	Spike	9	8.26%
NSP3:S1682S	C	7765	T	Silent	NSP3	S1682S	Predicted phosphoesterase, papain-like proteinase	7	6.42%
NSP2:I251M	A	1558	G	SNP	NSP2	I251M	Non-Structural protein 2	6	5.50%
NSP3:V190V	T	3289	C	Silent	NSP3	V190V	Predicted phosphoesterase, papain-like proteinase	6	5.50%
NSP5:G79G	A	10291	G	Silent	NSP5	G79G	3C-like proteinase	6	5.50%
NSP12b:D131D	C	13860	T	Silent	NSP12b	D131D	RNA-dependent RNA polymerase, post-ribosomal frameshift	6	5.50%
N:A211V	C	28905	T	SNP	N	A211V	Nucleocapsid protein	6	5.50%
NSP3:I617I	C	4570	T	Silent	NSP3	I617I	Predicted phosphoesterase, papain-like proteinase	5	4.59%
NSP3:F1354F	C	6781	T	Silent	NSP3	F1354F	Predicted phosphoesterase, papain-like proteinase	5	4.59%
NSP6:A161A	C	11455	T	Silent	NSP6	A161A	Transmembrane protein	5	4.59%
NSP12b:N619N	C	15324	T	Silent	NSP12b	N619N	RNA-dependent RNA polymerase, post-ribosomal frameshift	5	4.59%
NSP12b:A762S	G	15751	T	SNP	NSP12b	A762S	RNA-dependent RNA polymerase, post-ribosomal frameshift	5	4.59%
NSP13:V6V	T	16254	C	Silent	NSP13	V6V	Helicase	5	4.59%
NSP14:V125F	G	18412	T	SNP	NSP14	V125F	3'-to-5' exonuclease	5	4.59%
S:L5F	C	21575	T	SNP	S	L5F	Spike	5	4.59%
N:RG203KL	GGGG	28881	AACT	SNP	N	RG203KL	Nucleocapsid protein	5	4.59%
NSP2:T85I	C	1059	T	SNP	NSP2	T85I	Non-Structural protein 2	4	3.67%
NSP3:P822S	C	5183	T	SNP	NSP3	P822S	Predicted phosphoesterase, papain-like proteinase	4	3.67%
NSP3:L1035F	C	5822	T	SNP	NSP3	L1035F	Predicted phosphoesterase, papain-like proteinase	4	3.67%
NSP10:P59P	G	13201	A	Silent	NSP10	P59P	Growth-factor-like protein	4	3.67%
NSP12b:A440V	C	14786	T	SNP	NSP12b	A440V	RNA-dependent RNA polymerase, post-ribosomal frameshift	4	3.67%
NSP12b:N543N	T	15096	C	Silent	NSP12b	N543N	RNA-dependent RNA polymerase, post-ribosomal frameshift	4	3.67%
NSP14:V483V	C	19488	T	Silent	NSP14	V483V	3'-to-5' exonuclease	4	3.67%
NSP15:L216L	A	20268	G	Silent	NSP15	L216L	endornase	4	3.67%
S:A27V	C	21642	T	SNP	S	A27V	Spike	4	3.67%
S:N856N	C	24130	T	Silent	S	N856N	Spike	4	3.67%
N:A138S	G	28685	T	SNP	N	A138S	Nucleocapsid protein	4	3.67%
NSP1:L16L	C	313	T	Silent	NSP1	L16L	Leader protein	3	2.75%
NSP2:S369S	C	1912	A	Silent	NSP2	S369S	Non-Structural protein 2	3	2.75%
NSP2:S369S	C	1912	T	Silent	NSP2	S369S	Non-Structural protein 2	3	2.75%
NSP3:H295Y	C	3602	T	SNP	NSP3	H295Y	Predicted phosphoesterase, papain-like proteinase	3	2.75%
NSP4:P187H	C	9114	A	SNP	NSP4	P187H	Transmembrane protein	3	2.75%
NSP4:Y391Y	T	9727	C	Silent	NSP4	Y391Y	Transmembrane protein	3	2.75%
ORF3a:D27Y	G	25471	T	SNP	ORF3a	D27Y	ORF3a protein	3	2.75%
E:S60S	T	26424	C	Silent	E	S60S	Envelope	3	2.75%
M:C64F	G	26713	T	SNP	M	C64F	Membrane	3	2.75%
M:D209Y	G	27147	T	SNP	M	D209Y	Membrane	3	2.75%
5'UTR:36	C	36	G	Extragenic	5'UTR	N/A	N/A	2	1.83%
NSP2:G212D	G	1440	A	SNP	NSP2	G212D	Non-Structural protein 2	2	1.83%
NSP2:T528I	C	2388	T	SNP	NSP2	T528I	Non-Structural protein 2	2	1.83%
NSP3:D218E	C	3373	A	SNP	NSP3	D218E	Predicted phosphoesterase, papain-like proteinase	2	1.83%
NSP3:M494M	C	4202	T	Silent	NSP3	M494M	Predicted phosphoesterase, papain-like proteinase	2	1.83%
NSP3:S1023F	C	5787	T	SNP	NSP3	S1023F	Predicted phosphoesterase, papain-like proteinase	2	1.83%
NSP3:S1038F	C	5832	T	SNP	NSP3	S1038F	Predicted phosphoesterase, papain-like proteinase	2	1.83%
NSP5:G71S	G	10265	A	SNP	NSP5	G71S	3C-like proteinase	2	1.83%
NSP6:L37F	G	11083	T	SNP	NSP6	L37F	Transmembrane protein	2	1.83%
NSP12b:D751D	C	15720	T	Silent	NSP12b	D751D	RNA-dependent RNA polymerase, post-ribosomal frameshift	2	1.83%
NSP12b:K771K	G	15780	A	Silent	NSP12b	K771K	RNA-dependent RNA polymerase, post-ribosomal frameshift	2	1.83%
NSP12b:V835V	A	15972	G	Silent	NSP12b	V835V	RNA-dependent RNA polymerase, post-ribosomal frameshift	2	1.83%
NSP13:S350L	C	17285	T	SNP	NSP13	S350L	Helicase	2	1.83%
NSP13:S485L	C	17690	T	SNP	NSP13	S485L	Helicase	2	1.83%
NSP15:T33I	C	19718	T	SNP	NSP15	T33I	endornase	2	1.83%
S:L54L	G	21724	A	Silent	S	L54L	Spike	2	1.83%
S:L226V	T	22238	G	SNP	S	L226V	Spike	2	1.83%

M:V60L	G	26700	T	SNP	M	V60L	Membrane	2	1.83%
ORF7b:C41F	G	27877	T	SNP	ORF7b	C41F	ORF7b protein	2	1.83%
N:H145Y	C	28706	T	SNP	N	H145Y	Nucleocapsid protein	2	1.83%
N:S180I	G	28812	T	SNP	N	S180I	Nucleocapsid protein	2	1.83%
N:S194L	C	28854	T	SNP	N	S194L	Nucleocapsid protein	2	1.83%
N:T265I	C	29067	T	SNP	N	T265I	Nucleocapsid protein	2	1.83%
5'UTR:210	G	210	T	Extragenic	5'UTR	N/A	N/A	1	0.92%
NSP1:P80L	C	504	T	SNP	NSP1	P80L	Leader protein	1	0.92%
NSP1:M85I	G	520	T	SNP	NSP1	M85I	Leader protein	1	0.92%
NSP2:R46K	G	942	A	SNP	NSP2	R46K	Non-Structural protein 2	1	0.92%
NSP2:C50C	C	955	T	Silent	NSP2	C50C	Non-Structural protein 2	1	0.92%
NSP2:S138L	C	1218	T	SNP	NSP2	S138L	Non-Structural protein 2	1	0.92%
NSP2:T166I	C	1302	T	SNP	NSP2	T166I	Non-Structural protein 2	1	0.92%
NSP2:Y207Y	C	1426	T	Silent	NSP2	Y207Y	Non-Structural protein 2	1	0.92%
NSP2:V230M	G	1493	A	SNP	NSP2	V230M	Non-Structural protein 2	1	0.92%
NSP2:S299F	C	1701	T	SNP	NSP2	S299F	Non-Structural protein 2	1	0.92%
NSP2:G339S	G	1820	A	SNP	NSP2	G339S	Non-Structural protein 2	1	0.92%
NSP2:T388I	C	1968	T	SNP	NSP2	T388I	Non-Structural protein 2	1	0.92%
NSP2:G392R	G	1979	A	SNP	NSP2	G392R	Non-Structural protein 2	1	0.92%
NSP2:I436I	C	2113	T	Silent	NSP2	I436I	Non-Structural protein 2	1	0.92%
NSP2:Y537Y	C	2416	T	Silent	NSP2	Y537Y	Non-Structural protein 2	1	0.92%
NSP3:A58T	G	2891	A	SNP	NSP3	A58T	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:D163Y	G	3206	T	SNP	NSP3	D163Y	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:H342Y	C	3743	T	SNP	NSP3	H342Y	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:A496A	G	4207	T	Silent	NSP3	A496A	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:T504I	C	4230	T	SNP	NSP3	T504I	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:T678I	C	4752	T	SNP	NSP3	T678I	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:P874S	C	5339	T	SNP	NSP3	P874S	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:Q995Q	G	5704	A	Silent	NSP3	Q995Q	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:D1148E	T	6163	A	SNP	NSP3	D1148E	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:V1150A	T	6168	C	SNP	NSP3	V1150A	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:S1206L	C	6336	T	SNP	NSP3	S1206L	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:P1228L	C	6402	T	SNP	NSP3	P1228L	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:S1296F	C	6606	T	SNP	NSP3	S1296F	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:T1303T	C	6628	T	Silent	NSP3	T1303T	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:P1316L	C	6666	T	SNP	NSP3	P1316L	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:I1409V	A	6944	G	SNP	NSP3	I1409V	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:M1441I	G	7042	T	SNP	NSP3	M1441I	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:L1453M	T	7076	A	SNP	NSP3	L1453M	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:N1778D	A	8051	G	SNP	NSP3	N1778D	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP3:S1843F	C	8247	T	SNP	NSP3	S1843F	Predicted phosphoesterase, papain-like proteinase	1	0.92%
NSP4:A87V	C	8814	T	SNP	NSP4	A87V	Transmembrane protein	1	0.92%
NSP4:D188D	C	9118	T	Silent	NSP4	D188D	Transmembrane protein	1	0.92%
NSP4:L264F	C	9344	T	SNP	NSP4	L264F	Transmembrane protein	1	0.92%
NSP5:L32L	T	10150	C	Silent	NSP5	L32L	3C-like proteinase	1	0.92%
NSP5:R131R	G	10447	A	Silent	NSP5	R131R	3C-like proteinase	1	0.92%
NSP5:P132P	C	10450	T	Silent	NSP5	P132P	3C-like proteinase	1	0.92%
NSP5:F134F	C	10456	T	Silent	NSP5	F134F	3C-like proteinase	1	0.92%
NSP5:P184L	C	10605	T	SNP	NSP5	P184L	3C-like proteinase	1	0.92%
NSP6:A136V	C	11379	T	SNP	NSP6	A136V	Transmembrane protein	1	0.92%
NSP6:T222T	T	11638	C	Silent	NSP6	T222T	Transmembrane protein	1	0.92%
NSP8:R51L	G	12243	T	SNP	NSP8	R51L	Non-Structural Protein 8	1	0.92%
NSP8:P133S	C	12488	T	SNP	NSP8	P133S	Non-Structural Protein 8	1	0.92%
NSP9:P80T	C	12923	A	SNP	NSP9	P80T	ssRNA-binding protein	1	0.92%
NSP9:M101I	G	12988	T	SNP	NSP9	M101I	ssRNA-binding protein	1	0.92%
NSP10:P84S	C	13274	T	SNP	NSP10	P84S	Growth-factor-like protein	1	0.92%
NSP10:P136S	C	13430	T	SNP	NSP10	P136S	Growth-factor-like protein	1	0.92%
NSP12b:A190D	C	14036	A	SNP	NSP12b	A190D	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:A302S	G	14371	T	SNP	NSP12b	A302S	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:V466I	G	14863	A	SNP	NSP12b	V466I	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:H590H	C	15237	T	Silent	NSP12b	H590H	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:M624V	A	15337	G	SNP	NSP12b	M624V	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:E735D	G	15672	T	SNP	NSP12b	E735D	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:G765G	C	15763	T	Silent	NSP12b	G765G	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP12b:I838I	C	15981	T	Silent	NSP12b	I838I	RNA-dependent RNA polymerase, post-ribosomal frameshift	1	0.92%
NSP13:P172S	C	16750	T	SNP	NSP13	P172S	Helicase	1	0.92%
NSP13:Y211Y	C	16870	A	Silent	NSP13	Y211Y	Helicase	1	0.92%
NSP13:E261D	G	17019	T	SNP	NSP13	E261D	Helicase	1	0.92%
NSP13:H290Y	C	17104	T	SNP	NSP13	H290Y	Helicase	1	0.92%
NSP13:V510V	C	17766	T	Silent	NSP13	V510V	Helicase	1	0.92%
NSP14:G44V	G	18170	T	SNP	NSP14	G44V	3'-to-5' exonuclease	1	0.92%
NSP14:V91V	C	18312	T	Silent	NSP14	V91V	3'-to-5' exonuclease	1	0.92%
NSP14:V120L	G	18397	T	SNP	NSP14	V120L	3'-to-5' exonuclease	1	0.92%
NSP14:V136I	G	18445	A	SNP	NSP14	V136I	3'-to-5' exonuclease	1	0.92%
NSP14:V236V	C	18747	T	Silent	NSP14	V236V	3'-to-5' exonuclease	1	0.92%
NSP14:F326F	C	19017	T	Silent	NSP14	F326F	3'-to-5' exonuclease	1	0.92%
NSP14:I363V	A	19126	G	SNP	NSP14	I363V	3'-to-5' exonuclease	1	0.92%
NSP14:P443S	C	19366	T	SNP	NSP14	P443S	3'-to-5' exonuclease	1	0.92%
NSP15:P65S	C	19813	T	SNP	NSP15	P65S	endornase	1	0.92%

NSP16:S33R	A	20755	C	SNP	NSP16	S33R	2'-O-ribose methyltransferase	1	0.92%
S:K206R	A	22179	G	SNP	S	K206R	Spike	1	0.92%
S:T299I	C	22458	T	SNP	S	T299I	Spike	1	0.92%
S:T323I	C	22530	T	SNP	S	T323I	Spike	1	0.92%
S:L335L	G	22567	A	Silent	S	L335L	Spike	1	0.92%
S:P384S	C	22712	T	SNP	S	P384S	Spike	1	0.92%
S:N440K	T	22882	G	SNP	S	N440K	Spike	1	0.92%
S:A522V	C	23127	T	SNP	S	A522V	Spike	1	0.92%
S:A623V	C	23430	T	SNP	S	A623V	Spike	1	0.92%
S:S640S	T	23482	C	Silent	S	S640S	Spike	1	0.92%
S:G838G	T	24076	C	Silent	S	G838G	Spike	1	0.92%
S:A893A	A	24241	T	Silent	S	A893A	Spike	1	0.92%
ORF3a:V13L	G	25429	T	SNP	ORF3a	V13L	ORF3a protein	1	0.92%
ORF3a:I37V	A	25501	G	SNP	ORF3a	I37V	ORF3a protein	1	0.92%
ORF3a:T151A	A	25843	G	SNP	ORF3a	T151A	ORF3a protein	1	0.92%
E:V58F	G	26416	T	SNP	E	V58F	Envelope	1	0.92%
M:D3G	A	26530	G	SNP	M	D3G	Membrane	1	0.92%
M:F53F	C	26681	T	Silent	M	F53F	Membrane	1	0.92%
ORF7a:I10V	A	27421	G	SNP	ORF7a	I10V	ORF7a protein	1	0.92%
ORF7b:S5L	C	27769	T	SNP	ORF7b	S5L	ORF7b protein	1	0.92%
ORF7b:W29C	G	27842	T	SNP	ORF7b	W29C	ORF7b protein	1	0.92%
ORF8:Y73H	T	28110	C	SNP	ORF8	Y73H	ORF8 protein	1	0.92%
ORF8:V114F	G	28233	T	SNP	ORF8	V114F	ORF8 protein	1	0.92%
ORF8:R115C	C	28236	T	SNP	ORF8	R115C	ORF8 protein	1	0.92%
ORF8:V116S	TGT	28238	ATC	SNP	ORF8	V116S	ORF8 protein	1	0.92%
ORF8:V117L	G	28242	C	SNP	ORF8	V117L	ORF8 protein	1	0.92%
ORF8:L118V	T	28245	G	SNP	ORF8	L118V	ORF8 protein	1	0.92%
N:G34W	G	28373	T	SNP	N	G34W	Nucleocapsid protein	1	0.92%
N:R40C	C	28391	T	SNP	N	R40C	Nucleocapsid protein	1	0.92%
N:P162P	T	28759	C	Silent	N	P162P	Nucleocapsid protein	1	0.92%
N:K169R	A	28779	G	SNP	N	K169R	Nucleocapsid protein	1	0.92%
N:S193T	G	28851	C	SNP	N	S193T	Nucleocapsid protein	1	0.92%
N:Y298Y	C	29167	T	Silent	N	Y298Y	Nucleocapsid protein	1	0.92%
N:D348D	T	29317	C	Silent	N	D348D	Nucleocapsid protein	1	0.92%
N:D371G	A	29385	G	SNP	N	D371G	Nucleocapsid protein	1	0.92%
N:P383L	C	29421	T	SNP	N	P383L	Nucleocapsid protein	1	0.92%
3'UTR:29708	C	29708	T	Extragenic	3'UTR	N/A	N/A	1	0.92%
3'UTR:29721	C	29721	A	Extragenic	3'UTR	N/A	N/A	1	0.92%
3'UTR:29734	G	29734	C	Extragenic	3'UTR	N/A	N/A	1	0.92%

Supplementary table 2. Latvian SARS-CoV-2 isolate sequencing approaches.

GISAID isolate name	Accession ID	Collection date	Number of mutations	Average coverage	Type of NGS library	Sequencing technology
hCoV-19/Latvia/01/2020	EPI_ISL_421653	2020-03-25	7	7,138x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/02/2020	EPI_ISL_421654	2020-03-25	6	91x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/03/2020	EPI_ISL_421655	2020-03-25	3	21x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/04/2020	EPI_ISL_421656	2020-03-25	8	2,398x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/05/2020	EPI_ISL_426285	2020-03-30	9	55x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/06/2020	EPI_ISL_426286	2020-03-30	6	1,457x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/07/2020	EPI_ISL_426287	2020-03-30	7	129x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/08/2020	EPI_ISL_426288	2020-03-30	6	304x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/09/2020	EPI_ISL_426289	2020-03-30	8	45x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/010/2020	EPI_ISL_437089	2020-03-23	5	518x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/011/2020	EPI_ISL_437090	2020-03-24	5	37x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/012/2020	EPI_ISL_437091	2020-03-23	7	4,303x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/013/2020	EPI_ISL_437092	2020-03-20	7	53x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/014/2020	EPI_ISL_437093	2020-03-22	8	511x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/015/2020	EPI_ISL_437094	2020-03-23	6	669x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/016/2020	EPI_ISL_437095	2020-03-23	7	3,563x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/017/2020	EPI_ISL_437096	2020-03-23	7	220x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/018/2020	EPI_ISL_450518	2020-04-18	13	2,277x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/019/2020	EPI_ISL_450519	2020-04-14	8	6,173x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/020/2020	EPI_ISL_450520	2020-04-28	7	1,074x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/021/2020	EPI_ISL_450521	2020-04-28	7	563x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/022/2020	EPI_ISL_450522	2020-04-28	8	3,836x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/023/2020	EPI_ISL_450523	2020-04-28	8	1,363x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/024/2020	EPI_ISL_450524	2020-04-19	11	37x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/025/2020	EPI_ISL_486390	2020-05-09	12	2715x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/026/2020	EPI_ISL_486391	2020-05-03	8	3297x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/027/2020	EPI_ISL_486410	2020-05-09	13	5930x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/028/2020	EPI_ISL_486411	2020-05-07	6	191x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/029/2020	EPI_ISL_486412	2020-04-29	14	7839x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/030/2020	EPI_ISL_486413	2020-05-10	7	861x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/031/2020	EPI_ISL_486414	2020-05-07	10	79x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/032/2020	EPI_ISL_486415	2020-04-27	12	330x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/033/2020	EPI_ISL_486416	2020-04-27	14	27x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/034/2020	EPI_ISL_486417	2020-05-08	7	138x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/035/2020	EPI_ISL_486418	2020-05-09	6	604x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/036/2020	EPI_ISL_486419	2020-04-29	8	25x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/037/2020	EPI_ISL_486420	2020-04-29	7	30x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/038/2020	EPI_ISL_486421	2020-04-27	7	7480x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/039/2020	EPI_ISL_486422	2020-03-15	8	6066x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/040/2020	EPI_ISL_486423	2020-03-17	1	7129x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/041/2020	EPI_ISL_486424	2020-03-22	8	6560x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/042/2020	EPI_ISL_486425	2020-03-22	5	4274x	Enrichment by hybridization capture	Illumina, MiSeq

hCoV-19/Latvia/044/2020	EPI_ISL_486426	2020-06-30	11	2763x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/045/2020	EPI_ISL_486428	2020-03-22	2	4545x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/046/2020	EPI_ISL_486430	2020-03-22	6	4368x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/047/2020	EPI_ISL_486431	2020-03-22	13	6833x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/048/2020	EPI_ISL_486432	2020-03-16	9	6247x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/049/2020	EPI_ISL_486433	2020-03-16	7	5337x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/050/2020	EPI_ISL_486434	2020-03-16	6	6222x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/051/2020	EPI_ISL_486435	2020-03-16	7	4398x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/052/2020	EPI_ISL_486436	2020-03-23	6	846x	Enrichment by hybridization capture	Illumina, MiSeq
hCoV-19/Latvia/053/2020	EPI_ISL_486437	2020-03-25	6	43x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/054/2020	EPI_ISL_486438	2020-04-30	7	63x	Metatranscriptome	DNBSEQ-G400RS, MGI Tech Co., Ltd
hCoV-19/Latvia/043/2020	EPI_ISL_492988	2020-05-28	7	3991x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/066/2020	EPI_ISL_492989	2020-07-08	10	4404x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/055/2020	EPI_ISL_492990	2020-07-08	10	3715x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/056/2020	EPI_ISL_492991	2020-07-07	10	3801x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/057/2020	EPI_ISL_492992	2020-07-08	8	4169x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/058/2020	EPI_ISL_492993	2020-07-08	8	4322x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/059/2020	EPI_ISL_492994	2020-07-07	9	3363x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/060/2020	EPI_ISL_492995	2020-07-07	12	4163x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/061/2020	EPI_ISL_492996	2020-07-07	10	3534x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/062/2020	EPI_ISL_492997	2020-07-03	10	1334x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/063/2020	EPI_ISL_492998	2020-07-09	13	4315x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/064/2020	EPI_ISL_492999	2020-07-09	8	773x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/065/2020	EPI_ISL_493000	2020-07-08	10	4022x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/067/2020	EPI_ISL_501275	2020-07-08	8	3200x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/068/2020	EPI_ISL_501284	2020-07-10	8	3682x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/069/2020	EPI_ISL_501285	2020-07-10	12	4184x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/070/2020	EPI_ISL_501286	2020-06-05	10	3932x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/071/2020	EPI_ISL_501287	2020-06-06	7	3345x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/072/2020	EPI_ISL_501288	2020-07-10	10	4639x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/073/2020	EPI_ISL_501289	2020-07-10	13	4606x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/074/2020	EPI_ISL_501808	2020-07-06	11	5190x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/075/2020	EPI_ISL_501817	2020-07-09	12	2923x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/076/2020	EPI_ISL_501823	2020-07-16	9	3816x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/077/2020	EPI_ISL_501829	2020-07-16	14	1132x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/078/2020	EPI_ISL_501833	2020-07-17	11	4088x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/079/2020	EPI_ISL_501839	2020-07-17	9	322x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/080/2020	EPI_ISL_501849	2020-07-18	15	2729x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/081/2020	EPI_ISL_501894	2020-07-20	15	250x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/082/2020	EPI_ISL_501895	2020-07-14	13	4873x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/083/2020	EPI_ISL_501896	2020-06-15	7	5512x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/084/2020	EPI_ISL_501915	2020-06-16	8	4766x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/085/2020	EPI_ISL_501922	2020-06-16	7	3296x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/086/2020	EPI_ISL_501929	2020-07-09	10	4980x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/087/2020	EPI_ISL_501936	2020-07-09	10	4248x	Amplification by multiplexed primers	Illumina, MiSeq

hCoV-19/Latvia/088/2020	EPI_ISL_512313	2020-06-17	11	2883x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/089/2020	EPI_ISL_512314	2020-07-09	9	3332x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/090/2020	EPI_ISL_512645	2020-07-22	9	3330x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/091/2020	EPI_ISL_512646	2020-07-07	10	3593x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/092/2020	EPI_ISL_512647	2020-06-26	10	2643x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/093/2020	EPI_ISL_512648	2020-07-14	13	3664x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/094/2020	EPI_ISL_512649	2020-07-21	11	3682x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/095/2020	EPI_ISL_512650	2020-07-21	11	3745x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/096/2020	EPI_ISL_512651	2020-07-23	10	3320x	Amplification by multiplexed primers	Illumina, MiSeq
hCoV-19/Latvia/097/2020	EPI_ISL_512652	2020-07-25	9	2830x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/098/2020	EPI_ISL_515185	2020-07-29	12	4176x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/099/2020	EPI_ISL_515186	2020-07-31	11	3191x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/100/2020	EPI_ISL_515187	2020-08-04	13	4296x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/101/2020	EPI_ISL_515188	2020-08-04	14	4137x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/102/2020	EPI_ISL_515189	2020-08-05	13	3888x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/103/2020	EPI_ISL_515190	2020-08-05	14	3488x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/104/2020	EPI_ISL_515191	2020-08-05	13	4111x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/105/2020	EPI_ISL_515192	2020-08-05	13	3429x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/106/2020	EPI_ISL_515193	2020-08-05	13	4349x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/107/2020	EPI_ISL_515194	2020-08-05	14	3860x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/108/2020	EPI_ISL_515195	2020-08-05	13	3696x	Amplification by multiplexed primers	Illumina, MiSeq
hCov-19/Latvia/109/2020	EPI_ISL_515196	2020-08-03	12	1090x	Amplification by multiplexed primers	Illumina, MiSeq