Variants in nucleocapsid protein and endoRNase are found to associate with severe COVID-19 in a case-control study in Washington State, USA

Lue Ping Zhao (lzhao@fhcrc.org)
Fred Hutchinson Cancer Research Center

Pavitra Roychoudhury
Fred Hutchinson Cancer Research Center

Keith Jerome
Fred Hutchinson Cancer Research Center

Peter Gilbert
Fred Hutchinson Cancer Research Center

Joshua Schiffer
Fred Hutchinson Cancer Research Center

Terry Lybrand
Vanderbilt University

Thomas Payne
University of Washington Medical Center

April K Randhawa
Fred Hutchinson Cancer Research Center

Sara E Thiebaud
Fred Hutchinson Cancer Research Center

Margaret Mills
Fred Hutchinson Cancer Research Center

Alexander Greninger
Fred Hutchinson Cancer Research Center

Chul-Woo Pyo
Fred Hutchinson Cancer Research Center

Ruihan Wang
Fred Hutchinson Cancer Research Center

Renyu Li
Fred Hutchinson Cancer Research Center

Alexander S Thomas
Fred Hutchinson Cancer Research Center

Brandon M Norris

Fred Hutchinson Cancer Research Center

Wyatt C Nelson

Fred Hutchinson Cancer Research Center

Daniel Geraghty

Fred Hutchinson Cancer Research Center

Research Article

Keywords: haplotype, mutation, severe COVID-19, single nucleotide variant (SNV), transitory variant, variant, SARS-CoV-2

Posted Date: October 25th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-888049/v2

Abstract

SARS-CoV-2 is spreading worldwide with continuously evolving variants, some of which occur in the Spike protein and appear to increase the viral transmissibility. However, variants that cause severe COVID-19 or lead to other breakthroughs have not been well characterized. To discover such viral variants, we assembled a cohort of 683 COVID-19 patients; 388 inpatients ("cases") and 295 outpatients ("controls") from April to August 2020 using electronically captured COVID test request forms and sequenced their viral genomes. To improve the analytic power, we accessed 7,137 viral sequences in Washington State to filter out viral single nucleotide variants (SNVs) that did not have significant expansions over the collection period. Applying this filter led to the identification of 53 SNVs that were statistically significant, of which 13 SNVs each had 3 or more variant copies in the discovery cohort. Correlating these selected SNVs with case/control status, eight SNVs were found to significantly associate with inpatient status (q-values<0.01). Using temporal synchrony, we identified a four SNV-haplotype (t19839-g28881-g28882-g28883) which was significantly associated with case/control status (Fisher's exact $p=2.84*10^{-11}$) that appeared in April 2020, peaked in June, and persisted into January 2021. This association was replicated (OR=5.46, p-value=$4.71*10^{-12}$) in an independent cohort of 964 COVID-19 patients (June 1, 2020 to March 31, 2021). The haplotype included a synonymous change N73N in endoRNase, and three non-synonymous changes coding residues R203K, R203S and G204R in the nucleocapsid protein. This discovery points to the potential functional role of the nucleocapsid protein in triggering "cytokine storms" and severe COVID-19 that led to hospitalization. The study further emphasizes a need for tracking and analyzing viral sequences in correlations with clinical status.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), initially reported in Wuhan, Hubei, People's Republic of China[1], is the causal pathogen for the coronavirus disease (COVID-19), causing over 3 million fatalities worldwide as of April 15, 2021 (covid19.who.int). In the United States, COVID-19 has infected 32 million people and claimed 569,556 lives as of this reporting date (covid.cdc.gov). In Washington State, where the first COVID-19 patient in the US was reported on January 19, 2020, a total of 5,407 patients have died, among 383,000 known infections (www.doh.wa.gov/Emergencies/COVID19). Like other viruses, SARS-CoV-2 accumulates mutations with each cycle of replication known as single nucleotide variants (SNVs). Based on mutational frequencies or associated phenotypes, a viral strain with one or more such SNVs are referred to as variants, and variants meeting specific criteria can be classified as either Variants of Interest (VOI), Variants of Concern (VOC) or Variants of High Consequence (VOHC) by the Centers for Disease Control and Prevention (CDC) (https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html). Currently, three lineages are classified as VOI (B.1.526-Iota, B.1.525-Eta and P.2-Zeta), and five lineages as VOC (B.1.1.7-Alpha, P.1-Gamma, B.1.351-Beta, B.1.427-Epsilon, B.1.429-Epsilon, and, recently, B.1.617-Kappa/Delta), because of their elevated transmissibility or impact on neutralization[2,3]. In particular, the delta variant (B.1.617.2) was initially reported in India and was found to spread throughout the world with substantial transmissibility

(https://www.cdc.gov/coronavirus/2019-ncov/variants/variant.html). At this point, no VOHC has been declared by CDC.

The timely identification of VOHC is essential for the public health response against COVID-19 and requires viral genomic research directly connected to appropriately acquired clinical phenotypes in the clinically relevant setting. Here our primary interest was to discover viral SNVs that associate with the severity of COVID-19, one important phenotype of any VOHC. We report results identifying SNVs that associated with severe COVID-19, through a discovery case-control study of COVID-19 patients whose viral genomes were sequenced and hospitalization status (inpatient versus outpatient), together with demographic data, were retrieved from a database of COVID testing request forms. To improve the analytic power, we utilized a large collection of viral sequences from 7137 Washington residents deposited in GISAID and selected those SNVs that showed significant and substantial expansions from January 2020 to 2021. By correlating these identified SNVs with hospitalization status in the discovery study, we identified those SNVs that associated with hospitalization status, results that were further replicated in an independent cohort.

Results

SNVs with significant and substantial expansions

The SARS-COV-2 accumulates mutations during replication, potentially generating new strains, some of which have undergone substantial expansions in the population either because of super-spreader events or due to functional changes resulting in elevated transmissibility. Consequently, here we considered an SNV to be of interest if it had a statistically significant expansion in the study period and further its average proportion of mutations in the last three months exceeded 10%, to focus on prevalent variants. To identify such SNVs, we utilized 7137 viral genomes that had been generated from laboratories in Washington state and deposited to GISAID and that had been aligned and subject to quality control (see **Materials and Methods**). Comparing all viral genome sequences to the reference sequence led to counts of mutations at each nucleotide in the genome shown in Figure 1A. This analysis showed the Spike protein had two common SNVs with 300 or more copies and the nucleocapsid (N) protein demonstrated several SNVs with 300 or more occurrences. Since over 90% of positions in the 30 Kb genome had fewer than 3 mutations, we focused on the remaining 2516 nucleotides for possible significant expansions. A non-linear logistic regression model to regress the binary indicator for mutant type at each selected nucleotide on the collection time was then employed. In essence, this model fitted “locally averaged proportions of mutants” throughout the study period. If an SNV had gone through expansion, its locally averaged proportion would increase over time, as the mutant type became more common in the population.

This temporal expansion, deviating from random fluctuation, was quantified by a non-parametrically estimated function and its statistical significance was quantified by the p-value. To account for multiple comparisons, p-values were then converted to false positive error rates, i.e., q-values. An SNV was

deemed to have a significant expansion if the q-value was less than 0.01. Meanwhile, to focus on those pertinent SNVs that emerged or maintained their dominance in the population, fitted models were used to compute locally averaged proportions daily, and calculated the maximum proportion in last three months, denoting them as Pmax. An SNV was deemed to have a substantial expansion if the Pmax exceeded 10%. Figure 1B shows q-values and Pmax from 2516 nucleotides, in which 53 nucleotides met the established threshold values (q-value<0.01 and Pmax>0.10) (supplementary Table S1). Noticeably, four coding SNVs and 1 non-coding SNVs were identified in the Spike protein. Equally important to note was that SNVs with substantial expansions occurred in other ORFs/genes as well.

variants in Nucleocapsid, endoRNase and ORF3a have significant associations with severe COVID-19

The central goal of this study was to correlate identified SNVs with the hospitalization status in the discovery case-control samples in Washington state. The discovery cohort included 295 outpatients (controls) and 388 inpatients (cases), on which SARS-CoV-2 genome sequences were obtained (Table 1). Given constraints on the assembly of nasal swab samples and extraction of clinical data from the operational database, the case-control study took available samples from COVID-19 patients, while attempting to balance cases and controls for sex, age and collection times. As a result, there are some imbalances in collection time for outpatients from March to June and for inpatient from March to August.

To ensure the robustness of the association analysis on individual SNVs, we limited our analysis to 13 SNVs that had ten or more mutants observed (Table 2). Correlating these SNVs with case/control status through an unadjusted logistic regression model, we estimated coefficient (log odds ratio), standard error, Z value, p-value and q-value. For two SNVs with zero occurrences among outpatients, we performed the Fisher's exact test, where the logistic regression was not appropriate. On the far-right panel, we present results from the adjusted logistic regression analysis. For readability, we highlighted q-value if it was less than 0.01 and used green and red to correspond to positive and negative Z-value, corresponding the susceptibility and resistance to severe COVID-19.

Strikingly, three SNVs (g28881, g28882, g28883) in the nucleocapsid had significant associations with hospitalization status. These SNVs were found in perfect linkage disequilibrium and denoted as haplotype g28881/2/3, significantly elevated the risk of hospitalization (OR=5.81 and 6.55, $q=1.47*10^{-5}$ and $4.15*10^{-6}$ in the unadjusted and adjusted analysis, respectively). Further, the SNV c28854, also in the nucleocapsid, was not observed among outpatients and was found to be highly associated with hospitalization ($p=2.03*10^{-11}$) by the Fisher's exact test, given the logistic regression could not deal with the extreme association.

Two SNVs (t19838 and a20268) in endoRNase were found to have significant associations. The SNV t19838 mutant was absent among outpatients and was found to be significantly associated with

hospitalization status by the Fisher's exact test ($p=1.09*10^{-11}$). Similarly, the SNV a20268 was found to associate with the risk of hospitalization in both unadjusted and adjusted analysis (OR=38.47 and 42.95, $q=6.85*10^{-4}$ and $4.67*10^{-4}$, respectively). Two remaining SNVs (c1059, g25563) were found to have resistant association with severe COVID-19 (OR=0.46 and 0.45, $p=8.06*10^{-6}$ and $7.18*10^{-6}$, respectively). Similar resistant associations were observed for the adjusted analysis. It is of interest to note that a single SNV a23403, coding for the well-known D614G, appeared to have no association with the severe COVID-19 in unadjusted or adjusted analysis (p=0.39 and 0.29, respectively).

SNVs in nucleocapsid and endoRNase have synchronized expansion patterns

By the selection threshold values, six discovered SNVs were expected to have significant and substantial expansion during the study period (Figure 1C). The SNV haplotype g28881/2/3 started to expand prior to day 100, peaked around day 150, declined to below 10% around day 320, and re-emerged in January 2021 (black line). Following a similar temporal pattern, SNV t19839 followed a synchronized pattern with g28881/2/3, with a slightly later incline and earlier decline (red line). Similarly, two SNVs (a20268, c28854) in enodRNase and Nucleocapsid, respectively, appeared to have a synchronized expansion pattern; proportions started to rise around day 150 and reached a plateau after day 200 (blue dash and dotted lines, respectively). Finally, both SNVs (c1059, g25563) in nsp2 of ORF2ab and ORF3a, respectively, were synchronizing well, expanding from day 30, then contracting and expanding again towards the end of the study period. Such synchronies may imply that these variants share the same mutational histories, and thus the same haplotypes.

SNV-Haplotype (t19839-g28881-g28882-g28883) associates with Severe COVID

Due to the haploidic nature of the viral genome, the presence of multiple SNVs in a single patient implies that they form a single haplotype. Focusing on four SNVs that were synchronized temporally (t19839, g28881, g28882, g28883), their haplotypic frequencies across outpatients and inpatients in the discovery case-control study were tabulated (Table 3). The reference haplotype "tggg" had frequencies of 287 (97%) and 334 (86%) copies in outpatients and inpatients, respectively, while the haplotype "taac" with a single mutation was observed 8 (3%) and 11 (3%) in outpatients and inpatients. Interestingly, the haplotype "caac" was absent among outpatients completely, while it was observed 43 times (11%) among inpatients. The application of Fisher's exact analysis suggested that this SNV haplotype was significantly associated with the severe COVID-19 ($p\text{-value}=2.84*10^{-11}$).

Repeating the same haplotype tabulation in the replication case-control study yielded corresponding haplotype frequencies. Other than including a rare haplotype "tag", the replication analysis showed

largely comparable frequencies of three SNV haplotypes “tggg”, “taac” and “caac”. This haplotypic association with severe COVID-19 was replicated with high confidence ($p=2.21*10^{-10}$).

Applying the logistic regression of hospitalization status over this SNV haplotype, haplotypic association with severe COVID-19, with “tggg” as the reference haplotype was evaluated (Table 4). Treating “tggg” as the reference, we effectively set its coefficient to zero (OR=1). The mutant haplotype “caac” was found to have a significantly elevated risk of severe COVID-19 (OR=3.69, $p=3.44*10^{-10}$), without adjusting any covariates. After adjusting for sex, age and a potential non-linear effect of collection time, the “caac” association was improved further (OR=5.46, $p=4.71*10^{-12}$). Note that male and older age tended to increase risk of severe COVID-19 from the adjusted analysis, while the collection time appeared to have a significant curvature, i.e., risk of severe COVID-19 was relatively lower in the middle of the study period.

Using temporal synchrony, we considered the haplotypic association of t20268-c28854 with hospitalization (Table 3). In the discovery set, the mutant “gt” was absent in outpatients, and was observed 10% among all inpatients. As a result, this haplotype was found to have a significant association ($p=4.56*10^{-11}$) in the discovery set. However, the replication analysis provided a support for this association with a marginal significance (p=0.05), given 22% of outpatients carried this haplotype in comparison with 18% of inpatients. The discovered association of c1059-g25563 was replicated also with marginal significance (p=0.07).

dynamic Expansion of SNV haplotype (t19839-g28881-g28882-g28883) in Washington State

The SNV haplotype t19839-g28881-g28882-g28883 has a group of seven relatively uncommon haplotypes (cagg, cgac, cggc, taaa, tagc, tag, ttgg) with fewer than 5 copies, known as rare haplotypes, and has four other relatively common haplotypes tggg/0 with 5462 copies, cggg/1 with 29 copies, taac/3 with 434 copies and caac/4 with 1201 copies, in which /# indicates the number of mutants in the haplotype. Tabulating these haplotypes over collection time by months, Figure 2 shows that the reference haplotype tggg (gray) dominated over all months. The mutant haplotype “caac” (green) appeared in April, peaked in June, declined to a relatively low level, and appeared to rise again in January, 2021. The other mutant haplotype taac (red) was relatively steady throughout the year, since its appearance from April, 2020. In Washington state, the reference haplotype “tggg” had a frequency of 76%, while the mutant “caac” had a frequency of 17% (Table 3).

Classification of The haplotype (t19839c-g28881a-g28882a-g28883c)

All of Washington viral genomes obtained from GISAID were classified by nextstrains, GISAID-clade, and lineage. To assess the relationship between the haplotype and nextstrain classification, we tabulated

their cross-table frequencies (Table 5). All 1201 carriers of “caac” haplotype were classified to 20B, while a few carriers of “tggg” were classified to 20B. Similarly, carriers of “caac” belonged to the clade GR, as did “taac”, while no carriers of the reference haplotype were assigned to the clade. Finally, with respect to the assigned lineage, 80% of “caac” carriers were assigned to the lineage B.1.1.291, 7% to B.1.1.290, 6% to B.1.1, in addition to several sporadic assignments to, mostly, B.1.1 (Table 6). In contrast, only 11% of the carriers of the reference haplotype “tggg” were assigned to the B.1.371 lineage.

Discussion Of Washington State

This investigation utilized 7,137 viral sequences obtained from Washington state from January 19, 2020 through January 31, 2021 identifying 53 SNVs that had significant expansions and maximum proportions of mutations in the last three months exceeding 10%. Through a discovery case-control study, this study discovered six SNVs associating with the increased risk of severe COVID-19, while two SNVs associated with resistance. Among these six SNVs, four nucleotides (c28854, g28881, g28882, g28883) were non-synonymous and code residues S194L, R203K, R203S and G204R, respectively, in Nucleocapsid, and two nucleotides (t19839, a20268) in endoRNase of orf1ab encoded synonymous changes (N73N and L216L, respectively). Interestingly, t19839 appears to have a comparable expansion process to that of g28881, g28882, g28883, and the combined haplotype has a significant association with severe COVID-19 ($p=2.84*10^{-11}$ and $2.21*10^{-10}$ in the discovery and replication studies, respectively). Conversely, the risk association of (a20268, c28854) was discovered ($p=4.56*10^{-11}$) but was only marginally replicated ($p=0.05$).

The non-synonymous mutations R203K, R203S and G204R in the nucleocapsid protein all occur in the flexible linker region between the N-terminal RNA-binding domain and the C-terminal dimerization domain, and this linker segment is not resolved in any reported cryo-EM or x-ray structures.[4] However, small-angle X-ray scattering (SAXS) experiments revealed that the full-length nucleocapsid protein has a much larger radius of gyration than would be expected for a 99kDa globular protein, indicating that the flexible linker region is relatively extended in solution[4]. Consistent with the low-resolution conformational ensemble results from the SAXS experiments, recent single-molecule Förster resonance energy transfer (FRET) and fluorescence correlation spectroscopy experiments demonstrated that the linker region is highly flexible, with rapid interconversion between two general conformational populations[5], Together, these two experimental studies show that the linker region undergoes rapid conformational transitions but is generally extended, thus minimizing direct interactions of the well-structured RNA binding and dimerization domains. The mutations R203S and G204R are non-conservative and even the R->K mutation (R203K) is often observed to function as a non-conservative substitution in many cases, due to the different size of the R versus K residues and the notably different chemical features of the side-chain guanidinium group (arginine) versus the primary amine (lysine). Thus, we hypothesize that these mutations may influence disease severity by altering linker region flexibility and dynamics, which would likely alter nucleocapsid function. We also note that the linker region is involved in RNA binding interactions, so at least some of the linker region mutations might impact non-specific RNA binding.

The main variant in the endoRNase, N73N, is synonymous which suggests testable hypotheses concerning potential impact on virus fitness or function. It is well documented that the translation kinetics for synonymous codons are often different, and this can have an impact on co-translational protein folding kinetics, yielding proteins with identical primary sequence but different conformations and properties.[5] For example, Kimchi-Sarfaty et al demonstrated that P-glycoprotein expression using different synonymous codons yielded product with identical amino acid sequence but different substrate specificities that was attributed to differing P-glycoprotein conformations.[6] More recently, Hu et al showed that use of synonymous codons in heterologous expression of anti-IgE single chains in *E. coli* yielded scFv molecules with identical sequence but altered solubilities and antigen-binding affinities.[7] Thus, these synonymous mutations may lead to an endoRNase with improved function and/or properties due to alternate protein conformations. It is also possible that the synonymous mutations may impart a competitive advantage simply by resulting in enhanced translational kinetics for the endoRNase.

Many mutations in Spike protein that are correlated with increased transmission and/or severity exhibit "predictable" attributes. Specifically, such mutations are non-synonymous and occur in functionally important regions of the Spike protein where they may logically be anticipated to impact ACE2 receptor binding, alter neutralizing antibody recognition sites, or affect function via modulation of Spike protein flexibility. It is worth noting that the a23403 (D614G), the only SNV in the spike protein, was found not to associate with the severity of COVID-19 ($p=0.29$), implying Spike protein may have limited role in disease severity.

An interesting and important finding was that all SNVs associated with hospitalization status were located in endoRNase or Nucleocapsid, but not in Spike protein. This observation led us to postulate that while Spike protein is essential for the transmission of the virus mediated by its binding to the angiotensin converting enzyme 2 (ACE2)[8,9], it may play a diminished role in triggering autoimmune responses that lead to a "cytokine storm". Instead, the presence of new mutants in Nucleocapsid may accelerate replication of the virus [10], and endoRNase and Helicase may be responsible for initiating the secondary immune responses.

The results suggest that the viral genomes deposited in the GISAID are useful for filtering out SNVs with limited temporal patterns, allowing purpose-driven association analysis with clinical outcome data to have a sufficient power to discover phenotype-associated SNVs without sacrificing powers to correct unnecessary comparisons/tests. Furthermore, this exercise also supports a hybrid design that integrates GISAID with the purpose-driven study, given GISAID includes sequences from the State surveillance program are representative of the Washington study population. Direct access electronic health records of those COVID-19 patients whose sequences have been deposited to GISAID could significantly enhance analytic rigor and findings. In essence, such a hybrid design can be viewed as a two-stage design, which has been shown to be highly efficient and multiple statistical methodologies have been developed to extract maximum and unbiased association results[11-16].

There are limitations to our study that are worthy of note. One limitation of this study is that use of hospitalization status as a proxy for disease severity may lead to misclassification, since inpatients may be hospitalized for reasons other than COVID-19. Conversely, some patients with severe COVID-19 may not be hospitalized or their hospitalization may not be reported at the time of testing because it occurs later. However, misclassification errors tend to dilute association results[17]. Thus, the true magnitude of discovered and replicated association with the haplotype "caac" of t19839-g28881-g28882-g28883 may be even greater than estimated here, if the severity of COVID-19 could be clinically adjudicated. Another limitation was that it was not possible to completely match our discovery and replication case-control studies could not match age, sex and collection time, partly due to challenges facing research studies relying on the operational database and biospecimen during the pandemic. To address this issue, we applied the logistic regression model to evaluate viral genetic associations, while adjusting for these potential confounding variables.

Successful development and implementation of COVID vaccines will certainly curtail this pandemic, but infections among unvaccinated people, and to a lesser extent vaccinated, in and outside of the USA is near certain to generate new variants in the coming years. While ongoing efforts are continuously monitoring potential new variants that arise through real time genomic sequencing, what is lacking is a two-staged study accessing electronic health records and identifying their vaccine status. Through this strategy, viral genome sequences could be correlated with COVID-19 and vaccine related clinical outcomes, allowing for the real time identification of new variants of high consequence.

Materials And Methods

Patient biospecimen and data

This study was approved by the Human Subject Review Committee at Fred Hutchinson Cancer Research Center (IRB#6007-2043) and by the University of Washington Institutional Review Board (STUDY00000408). The current study includes: 1) a discovery case-control study of 683 COVID-19 patients (March-August 2020), and 2) a replication case-control study of 964 patients (June 2020-March 2021) from healthcare organizations in Washington State. All subjects were de-identified, and deidentified nucleic acid samples were extracted from leftover nasal swabs and were used for viral sequencing. Deidentified demographic and healthcare-related information were extracted from electronic forms of COVID testing requests, including sex, age, and collection times (Table 1).

Treating hospitalization status as a proxy for severity of Covid19, this study used a case-control design with inpatients as cases and outpatients as controls. In the discovery study, we attempted to match inpatients and outpatients' sex and age by frequency as much as possible, subject to the availability of nasal swab samples. All viral sequences were deposited to Genebank (accession numbers MW593154-MW593926). The replication study included all available patients whose sequences were obtained and deposited to GISAID (https://www.gisaid.org).

This study had also downloaded viral genome sequences that have been deposited to GISAID (https://www.gisaid.org) from all Washington laboratories. On the downloaded dataset, we aligned all sequences against the reference genome, performing quality control, eliminated 3 samples of poor sequence quality, and removed 5' and3' end sequences of variable lengths. We used submission dates as a proxy for collection time, and used their classification by Nextstrains[18], clade by GISAID (https://www.gisaid.org) and lineage by PANGO [19].

Samples, RNA extraction, and PCR

Patient samples were obtained and tested according to local and CDC guidelines. The University Washington (UW) Virology Division Laboratory is CLIA-certified and CAP-accredited and was one of the first academic labs in the US to offer clinical testing for SARS-CoV-2. UW Virology uses lab-developed RT-PCR tests based on either the CDC N1 and N2 or the WHO E/RdRp primer/probe sets, and FDA Emergency Use Authorization tests from Hologic (Panther Fusion), and Roche (Cobas 6800).[20-28]

Nasopharyngeal swabs were collected in either viral transport medium (VTM) or phosphate-buffered saline (PBS). Total nucleic acid was extracted from 200 µl of VTM/PBS sample and eluted into 50 µl of buffer using MagNA Pure 96 DNA and viral NA SV Kit on MagNA Pure 96 instrument (Roche). Nucleic acids were then used for genotyping.

Amplicon-Based Sequencing for Discovery samples

A commercially available ScisGo®-COVID-19 kit (Scisco Genetics Inc., Seattle WA) employing an amplicon-based sequencing by synthesis approach was used to determine sequences from SARS-CoV-2 positive samples obtained from the local testing site. The approach mirrors a system previously developed for HLA and KIR typing [29,30] using a two-stage amplicon-based PCR for locus amplification and sample barcoding and substituting two primer sets, each independently yielding non-overlapping SARS-CoV-2 amplicon sequences of ~400 bp. The combined derivative data spans the complete SARS-CoV-2 genome including de novo sequencing of all primer binding sites excepting the two primers at the extreme 5' and 3' ends. Briefly, after conversion of total nucleic acid into cDNA using the Invitrogen SuperScript IV First Strand Synthesis System (Thermo Fisher, Bothell, WA) the samples were sequentially applied to stage 1 (S1) and stage 2 (S2) PCR amplification according the manufacturer supplied protocol. After amplification, the reactions were combined, purified, and applied to a MiSeq using Illumina Version 2 chemistry with 500-cycle, paired-end sequencing (Illumina, San Diego, CA). Data assembly and analysis was performed using Sciscloud™ (Scisco Genetics Inc., Seattle WA) computational tools adapted specifically to assemble SARS-CoV-2 genomic sequences derivative from the ScisGo®-COVID-19 kit. Access to all software for data transfer and analysis was included as a component of the kit and made available through a web browser. All discovery cohort samples were sequenced using the ScisGo® approach and are accessible in genbank under accession numbers MW593154-MW593926. All

other samples from UW Virology were sequenced using either metagenomic or amplicon-based approaches using the Illumina COVIDseq Test (Illumina, San Diego, CA) and the Swift Biosciences' Normalase amplicon SARS-CoV-2 panel (Swift Biosciences, Ann Arbor, MI) as previously described[31].

All biospecimen collections and processing have been detailed in the Human Subject Research protocol, and have been reviewed and approved by the Human Subject Research Committee. With respect to statistical analyses (below), we use robust and reproducible statistical procedures that have been well-documented in the statistical literature and have been implemented in commonly accepted statistical software packages in R.

A Non-Parametric logistic regression model

To model dynamic expansions and contractions of individual SNVs over time, we applied a non-parametric logistic regression model, a member of the generalized additive model (GAM), regressing a binary SNV indicators over collection times[32-34]. After fitting the model, we obtained p-value that quantifies the non-linear dynamics of the mutation proportion and computed the fitted values as locally weight-averaged mutation proportion daily throughout the year. The maximum proportion in the last three month, denoted as Pmax, was computed to indicate if the mutation proportion had expanded and reached a substantial level in the end of the study period. To correct multiple comparisons, the false positive error rates (q-values) were computed from p-values. An SNV was selected if the q-value threshold (<0.01) indicated statistically significant dynamics and if the Pmax threshold (>10%) suggested a substantial expansion.

Imputing Missing Nucleotides

Due to nature of sequencing technologies, a small fraction of nucleotides were untyped, and were coded as "n". Given high linkage disequilibrium across all SNVs, we assembled a panel of polymorphic nucleotides that had no missing values and had not been selected into SNVs of interest, and treated them as an "imputation base". Fusing one SNV with those in the imputation base, we computed their haplotype frequencies, and used their haplotype frequencies to compute posterior probabilities to impute missing nucleotides, in the same way as imputing single nucleotide polymorphisms [35].

Logistic Regression and Statistics

Treating a binary indicator of 1 and 0 for inpatient and outpatient, respectively, as an outcome, the logistic regression model regresses on SNVs or their haplotypes, to generate association statistics: estimated coefficient, standard error, Z-score, p-value and q-value. For SNVs with zero frequencies in either outpatients or inpatients, we performed Fisher's exact test, instead of logistic regression model.

Fisher exact test produced the exact p-values. Confounders (age, sex, collection time) were included into the logistic regression model as an adjusted analysis.

Declarations

Acknowledgments:

Funding:

National Institutes of Health grant R01-GM129325

National Institutes of Health/National Institute of Allergy and Infectious Diseases grant UM1 AI068635

National Institutes of Health/National Institute of Allergy and Infectious Diseases contract BAA-NIAID-DAIT-75N93019R00020

Author contributions:

Conceptualization: LPZ, DEG, PR, KRJ, PBG, JTS, TL, THP

Methodology: LPZ, PR, CWP, DEG

Investigation: LPZ, DEG, PR, KRJ

Visualization: LPZ, DEG, KRJ

Funding acquisition: LPZ, PBG, DEG

Formal analysis: LPZ, PR

Data curation: DEG, PR, KRJ, PBG, JTS, AR, MM, AG, CWP, RW, RL, AT, BN, WCN, ST

Supervision: LPZ, DEG

Writing – original draft: LPZ, PR, DEG, PBG, TL, JTS

Writing – review & editing: LPZ, DEG, PR, KRJ, PBG, JTS, TL, THP

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: All sequence data analyzed here are publicly available at GSIAD (https://www.gisaid.org/) and Genebank (https://www.ncbi.nlm.nih.gov/genbank).

References

1. Coronaviridae Study Group of the International Committee on Taxonomy. of, V. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol*, **5**, 536–544 https://doi.org/10.1038/s41564-020-0695-z (2020).
2. Deng, X. *et al.* Transmission, infectivity, and antibody neutralization of an emerging SARS-CoV-2 variant in California carrying a L452R spike protein mutation. *medRxiv*, https://doi.org/10.1101/2021.03.07.21252647 (2021).
3. Davies, N. G. *et al.* Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England., https://doi.org/10.1126/science.abg3055 (2021).
4. Zeng, W. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem Biophys Res Commun*, **527**, 618–623 https://doi.org/10.1016/j.bbrc.2020.04.136 (2020).
5. Mitra, S., Ray, S. K. & R, B. Synonymous codons influencing gene expression in organisms.*Res Rep Biochem***6** (2016).
6. Kimchi-Sarfaty, C. *et al.* A "silent" polymorphism in the MDR1 gene changes substrate specificity., **315**, 525–528 https://doi.org/10.1126/science.1135308 (2007).
7. Hu, S., Wang, M., Cai, G. & He, M. Genetic code-guided protein synthesis and folding in Escherichia coli. *J Biol Chem*, **288**, 30855–30861 https://doi.org/10.1074/jbc.M113.467977 (2013).
8. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**, 270–273 https://doi.org/10.1038/s41586-020-2012-7 (2020).
9. Letko, M., Marzi, A. & Munster, V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol*, **5**, 562–569 https://doi.org/10.1038/s41564-020-0688-y (2020).
10. Savastano, A., Ibanez de Opakua, A., Rankovic, M. & Zweckstetter, M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nat Commun*, **11**, 6041 https://doi.org/10.1038/s41467-020-19843-1 (2020).
11. Breslow, N. E. & Cain, K. C. Logistic regression for two-stage case-control data., **75**, 11–20 (1988).
12. White, K. C. & Ramus, D. L. Two-stage impression technique for overdentures. *J Prosthet Dent*, **61**, 452–457 (1989).
13. Zhao, L. P. & Lipsitz, S. R. Designs and analysis of two-stage studies. *Stat Med*, **11**, 769–782 (1992).
14. Whittemore, A. S.1–29(1995).
15. Lin, D. Y. Evaluating statistical significance in two-stage genomewide association studies. *Am J Hum Genet*, **78**, 505–509 https://doi.org/10.1086/500812 (2006).
16. Zuo, Y. & Kang, G. A mixed two-stage method for detecting interactions in genomewide association studies. *J Theor Biol*, **262**, 576–583 https://doi.org/10.1016/j.jtbi.2009.10.029 (2010).
17. Yi, G. Y., Delaigle, A. & Gustafson, P. Handbook of measurement error. First edition. edn(CRC Press, 2022).
18. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution., **34**, 4121–4123 https://doi.org/10.1093/bioinformatics/bty407 (2018).

19. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*, **5**, 1403–1407 https://doi.org/10.1038/s41564-020-0770-5 (2020).
20. Roxby, A. C. *et al.* Detection of SARS-CoV-2 Among Residents and Staff Members of an Independent and Assisted Living Community for Older Adults - Seattle, Washington, 2020. *MMWR Morb Mortal Wkly Rep*, **69**, 416–418 https://doi.org/10.15585/mmwr.mm6914e2 (2020).
21. Pettit, S. D. *et al.* 'All In': A Pragmatic Framework for COVID-19 Testing and Action on a Global Scale. *EMBO Mol Med*, https://doi.org/10.15252/emmm.202012634 (2020).
22. Perchetti, G. A. *et al.* Validation of SARS-CoV-2 detection across multiple specimen types. *J Clin Virol*, **104438**, https://doi.org/10.1016/j.jcv.2020.104438 (2020).
23. Peddu, V. *et al.* Metagenomic analysis reveals clinical SARS-CoV-2 infection and bacterial or viral superinfection and colonization. *Clin Chem*, https://doi.org/10.1093/clinchem/hvaa106 (2020).
24. Nalla, A. K. *et al.* Comparative Performance of SARS-CoV-2 Detection Assays using Seven Different Primer/Probe Sets and One Assay Kit. *J Clin Microbiol*, https://doi.org/10.1128/JCM.00557-20 (2020).
25. Lieberman, J. A. *et al.* Comparison of Commercially Available and Laboratory Developed Assays for in vitro Detection of SARS-CoV-2 in Clinical Laboratories. *J Clin Microbiol*, https://doi.org/10.1128/JCM.00821-20 (2020).
26. Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States., https://doi.org/10.1016/j.cell.2020.04.021 (2020).
27. Bryan, A. *et al.* Performance Characteristics of the Abbott Architect SARS-CoV-2 IgG Assay and Seroprevalence in Boise, Idaho. *J Clin Microbiol*, https://doi.org/10.1128/JCM.00941-20 (2020).
28. Bhatraju, P. K. *et al.* Covid-19 in Critically Ill Patients in the Seattle Region - Case Series. *N Engl J Med*, https://doi.org/10.1056/NEJMoa2004500 (2020).
29. Nelson, W. C. *et al.* An integrated genotyping approach for HLA and other complex genetic systems. *Hum Immunol*, **76**, 928–938 https://doi.org/10.1016/j.humimm.2015.05.001 (2015).
30. Smith, A. G. *et al.* Comparison of sequence-specific oligonucleotide probe vs next generation sequencing for HLA-A, B, C, DRB1, DRB3/B4/B5, DQA1, DQB1, DPA1, and DPB1 typing: Toward single-pass high-resolution HLA typing in support of solid organ and hematopoietic cell transplant programs. *HLA* **94**, 296-306, doi:10.1111/tan.13619 (2019).
31. Addetia, A. *et al.* Sensitive Recovery of Complete SARS-CoV-2 Genomes from Clinical Samples by Use of Swift Biosciences' SARS-CoV-2 Multiplex Amplicon Sequencing Panel. *J Clin Microbiol*, **59**, https://doi.org/10.1128/JCM.02226-20 (2020).
32. Hastie, T. & Tibshirani, R. Generalized Additive Models. *Statistical Science*, **1**, 297–318 (1991).
33. Schwartz, J. Nonparametric smoothing in the analysis of air pollution and respiratory illness. *The Canadian Journal of Statistics*, **22**, 471–487 (1994).
34. Zhao, L. P. *et al.* Control costs, enhance quality, and increase revenue in three top general public hospitals in Beijing, China. *PLoS ONE [Electronic Resource]*, **8**, e72166

https://doi.org/10.1371/journal.pone.0072166 (2013).

35. Li, S., Khalid, N., Carlson, C. & Zhao, L. P. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics*, **4**, 513–522 (2003).

Tables 1-6

Tables 1-6 are available in the Supplementary Files section.

Figures

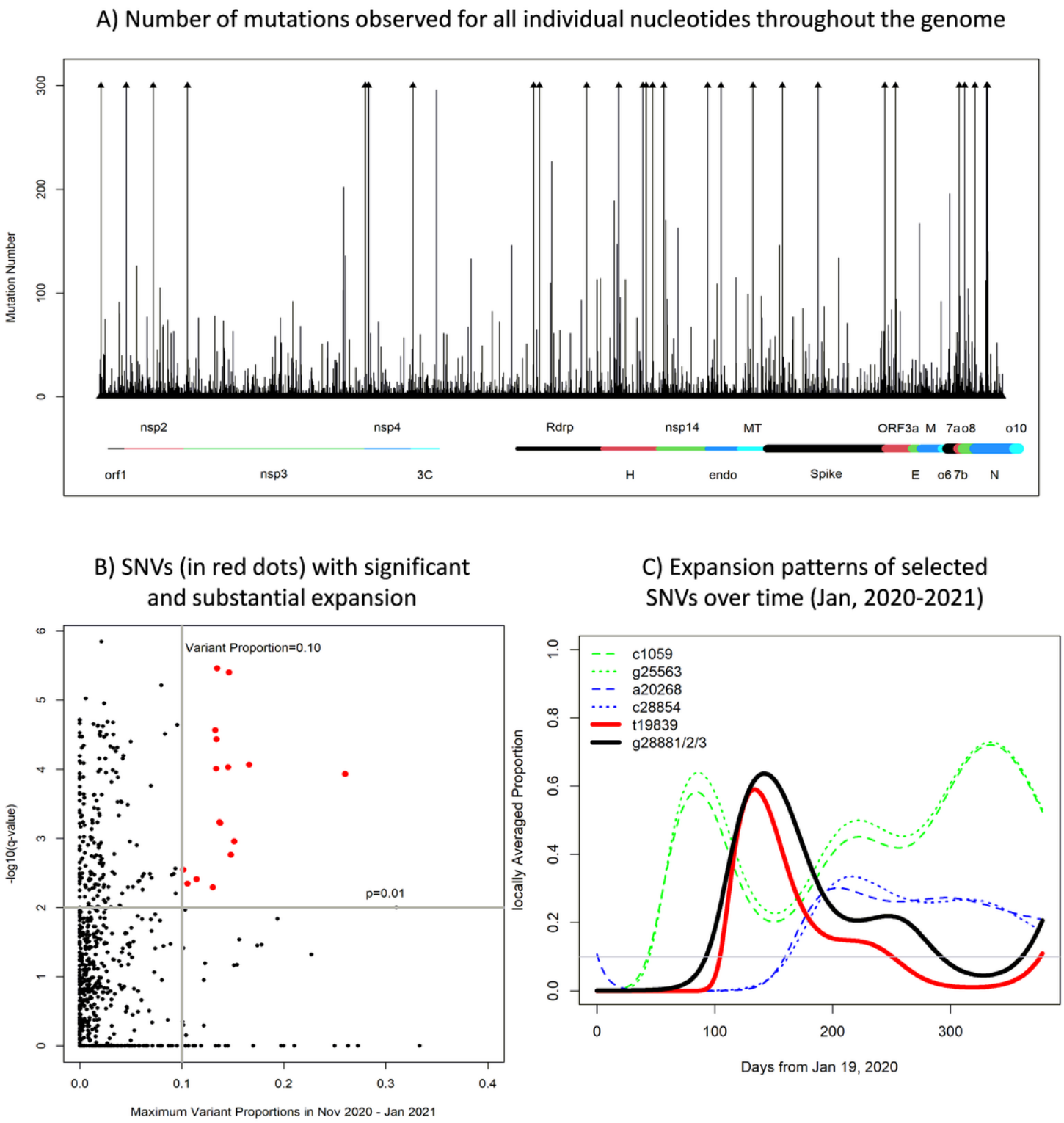

Figure 1

A) Results from counting mutational numbers per nucleotides throughout the viral genome, in which upper arrow indicates observed counts greater than 300 and the viral genome with its genesis is annotated below the figure; B) q-values and maximum values of variant proportions in November, December and January 2021 are obtained from fitting generalized linear models to all SNVs; C)

Computed locally averaged variant proportions over time from fitting the generalized linear model to eight selected SNVs.

Fig 2. Evolving haplotype frequencies of SNV haplotype (t19839-g28881-g28882-g28883) over January 2020-2021 in Washington

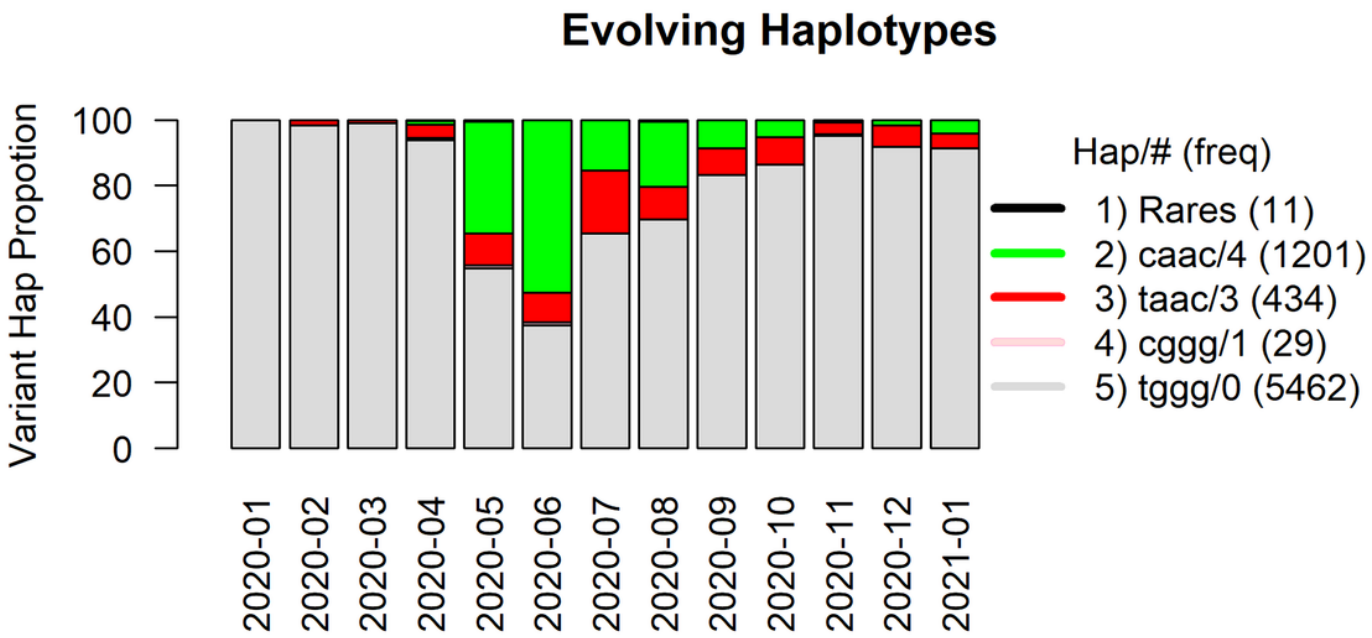

Figure 2

Evolving haplotype frequencies of SNV haplotype (t19839-g28881-g28882-g28883) over January 2020-2021 in Washington

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx
- supplementaryTableS1.xlsx