Youn *et al. BioData Mining* (2025) 18:61
https://doi.org/10.1186/s13040-025-00476-3

BioData Mining

RESEARCH Open Access

Identification of severity related mutation hotspots in SARS-CoV-2 using a density-based clustering approach

Sohyun Youn[1†], Dabin Jeong[2†], Hwijun Kwon[1†], Eonyong Han[1], Sun Kim[2,3,4,5] and Inuk Jung[1*]

[†]Sohyun Youn, Dabin Jeong and Hwijun Kwon contributed equally to this work.

*Correspondence:
Inuk Jung
inukjung@knu.ac.kr
[1]School of Computer Science and Engineering, Kyungpook National University, Buk-gu, Daegu 41566, Republic of Korea
[2]Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK
[3]AIGENDRUG Co., Ltd, Seoul 1793, Republic of Korea
[4]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, Republic of Korea
[5]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, Republic of Korea

Abstract

Background The immune response to SARS-CoV-2 varies greatly among individuals yielding highly varying severity levels among the patients. While there are various methods to spot severity associated biomarkers in COVID-19 patients, we investigated highly mutated regions, or mutation hotspots, within the SARS-CoV-2 genome that correlate with patient severity levels. SARS-CoV-2 mutation hotspots were searched in the GISAID database using a density based clustering algorithm, Mutclust, that searches for loci with high mutation density and diversity.

Results Using Mutclust, 477 mutation hotspots were searched in the SARS-CoV-2 genome, of which 28 showed significant association with severity levels in a multi-omics COVID-19 cohort comprised of 387 infected patients. The patients were further stratified into moderate and severe patient groups based on the 28 severity related mutation hotspots that showed distinctive cytokine and gene expression levels in both cytokine profile and single-cell RNA-seq samples. The effect of the SARS-CoV-2 mutation hotspots on human genes was further investigated by network propagation analysis, where two mutation hotspots specific to the severe group showed association with NK cell activity. One of them showed to decrease the affinity between the viral epitope of the hotspot region and its binding HLA when compared to the non-mutated epitope.

Conclusion Genes related to the immunological function of NK cells, especially the NK cell receptor and co-activating receptor genes, were significantly dysregulated in the severe patient group in both cytokine and single-cell levels. Collectively, mutation hotspots associated with severity and their related NK cell associated gene expression regulation were identified.

Keywords SARS-CoV-2, Mutation, Cluster, Severity, Multi-omics

Background

COVID-19 is an infectious respiratory disease that resulted in more than 700 million confirmed cases and 7 million deaths worldwide. COVID-19 has placed unprecedented strain on the world's health systems due to the continuous evolution of SARS-CoV-2 and the emergence of lethal variants. As SARS-CoV-2 has spread worldwide, a large number

of mutations have occurred changing the characteristics of the virus, such as its transmissibility, ability to cause severe disease, and capacity to evade antibodies. To mitigate such burden, various studies made effort to explain the association between mutations and the characteristics of the virus, especially the severity of the illness in patients.

Studies that explain virus characteristics by their mutation profiles mainly utilize mutation frequency at the amino acid or nucleotide level, based on the assumption that frequently occurring mutations are more likely to be functionally significant [1–4]. While effective in some contexts, frequency-based approaches have several limitations. They are prone to lineage-driven biases, and they often overlook rare but diverse mutations that may signal viral adaptation or immune evasion. Although entropy-based approaches such as Shannon entropy have been introduced to measure mutation diversity [5]they too fall short when used alone, as they fail to account for the overall mutation prevalence. Functionally important mutation hotspots often arise from regions under structural or immunological selective pressure and tend to cluster in nonuniform and irregular patterns [6, 7]. However, conventional detection methods—such as fixed-length sliding windows or traditional clustering algorithms are inadequate for capturing the biological complexity of viral genomes [8].

To overcome these limitations, we propose a novel density-aware clustering algorithm named MutClust, which utilizes a mutation importance metric called the H-score. MutClust is based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN [9]) framework but introduces key improvements: Local ε adjustment based on H-score, incorporating both mutation density and biological importance, Weighted edge handling, allowing cluster boundaries to dynamically adjust using an attenuation factor, Improved detection of biologically meaningful clusters, especially in regions with diverse but moderate-frequency mutations. These enhancements enable MutClust to more accurately detect mutation hotspots by integrating both spatial and functional characteristics.

The overall flowchart of this study is illustrated in Fig. 1. We applied our method to a dataset of 224,318 SARS-CoV-2 genome sequences from GISAID and identified 477 mutation hotspots using the H-score and MutClust algorithm. Among these, certain hotspots exhibited moderate mutation frequency but high diversity and were strongly associated with disease severity. To validate their clinical relevance, we utilized a multi-omics COVID-19 cohort from the Korea Disease Control and Prevention Agency (KDCA), which included matched electronic health records (EHR), cytokine profiles, and single-cell RNA-seq data.

By introducing MutClust and the H-score, we present a novel computational framework that enables the discovery of clinically and biologically relevant mutation hotspots that may be missed by traditional methods. This approach offers a scalable and interpretable strategy for hotspot detection in highly mutagenic viral or tumor genomes. Our findings provide new insights into the relationship between viral mutation patterns and disease severity, with potential applications in vaccine design, immune monitoring, and real-time genomic surveillance.

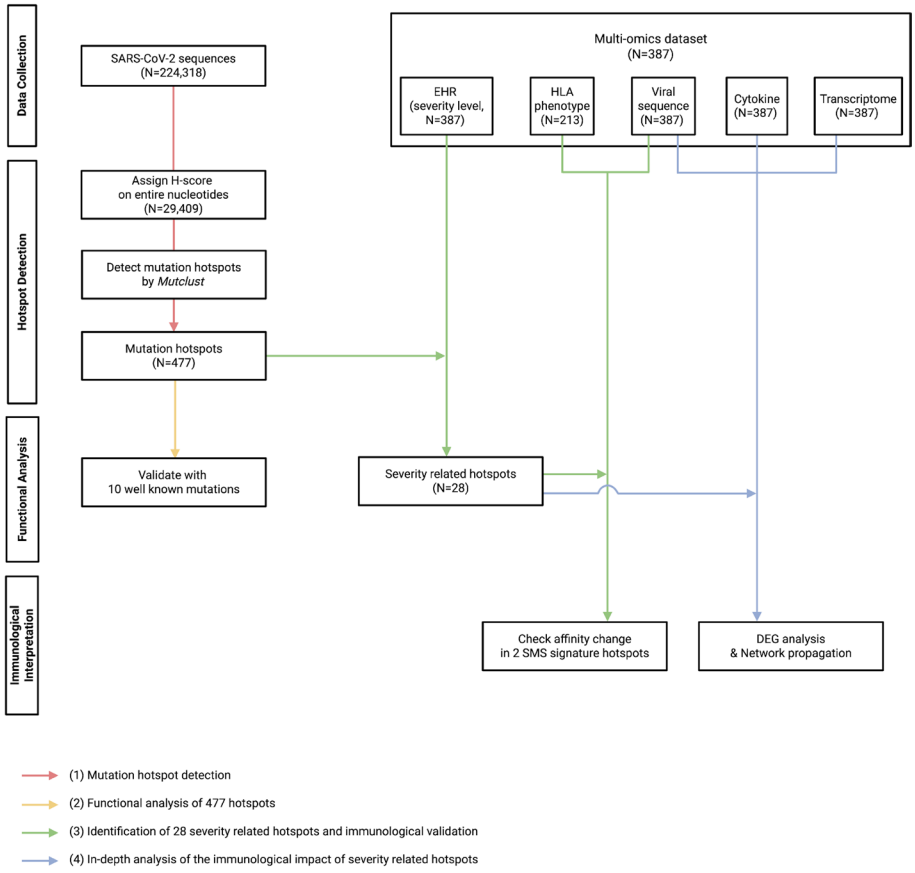

Fig. 1 Flowchart of mutation hotspot identification and severity association analysis in SARS-CoV-2. Flowchart presents the overall flowchart of this study, which is composed of four major steps: (1) Hotspot detection, (2) Functional analysis of hotspots, (3) Identification of severity-associated hotspots and immunological validation, and (4) In-depth analysis of the immunological impact of severity-associated hotspots. We collected 224,318 SARS-CoV-2 genome sequences from the GISAID database and computed the H-score for all 29,409 nucleotide positions to quantify mutation importance (Step 1). Based on the H-score, a density-based clustering algorithm (MutClust) was applied to identify 477 mutation hotspots across the viral genome. The biological relevance of these hotspots was validated by confirming the inclusion of 10 well-known functional mutations in the spike protein region (Step 2). To evaluate their clinical significance, we utilized a multi-omics dataset from 387 COVID-19 patients provided by the Korea Disease Control and Prevention Agency (KDCA) (Step 3). Statistical testing using WHO-based severity labels revealed 28 hotspots significantly associated with disease severity. To further explore immunological relevance, we identified a subgroup of patients exhibiting severe mutation signatures (SMS) and two key hotspots (c315 and c442) enriched in this group, defined as SMS signature hotspots. Using patient-specific HLA phenotype data, we assessed affinity changes in predicted epitopes spanning these hotspots and identified 8 HLA-epitope pairs with reduced binding affinity. Finally, we performed network propagation and differential expression analysis using viral genomic and RNA-seq data, revealing disruption in NK cell signaling pathways and supporting the immunopathological role of the SMS group in severe disease progression (Step 4). Furthermore, leveraging viral genomic and RNA-seq data from the same dataset, we conducted network propagation and differential expression analyses, which revealed disruption in NK cell activation and NK receptor pathways in the SMS group, offering deeper immunological insight into the mechanisms underlying severe clinical outcomes

Materials and methods

Data acquisition and processing

SARS-CoV-2 sequences were collected from the GISAID [10] database for searching mutation hotspots. Using a multi-omics COVID-19 cohort dataset [11] severity associated SARS-CoV-2 mutation hotspots were collected leveraging EHR, cytokine and scRNA-seq samples. The cohort data was collected across three hospitals in South Korea. The details and preprocessing procedures of each data is described in the following sections.

Dataset for detecting mutation hotspots

A total of 224,318 SARS-CoV-2 sequences were obtained from the GISAID database, within the period from January 2020 to November 2022. Sequence alignment was performed using MAFFT [12] v7.490 using hCoV-19/Wuhan/WIV04/2019 (GISAID Access ID: EPI_ISL_402124, GenBank Access: MN908947) as the reference genome. The length of the reference sequence is 29,903 bp. Furthermore, ambiguous nucleotides according to the IUPAC nucleotide code were excluded. Metadata information for the collected sequence data was categorized into location, clade, lineage, patient age, gender, severity, and collection date. Lineages such as B.1.1.7, AY.43, AY.122, B.1.617.2, AY.4, BA.2, AY.20, B.1.1, and others were included in the dataset, and for clustering analysis, only lineages with a sequence count of 17,000 or more were considered. Detailed information about the collected SARS-CoV-2 sequences is provided in the Supplementary Table S1.

The multi-omics COVID-19 cohort dataset

EHR and multi-omics samples of the COVID-19 cohort were collected by KDCA across three hospitals: Chungnam National University Hospital, Seoul Medical Center, and Samsung Medical Center. Clinical features, such as D-dimer, lactate dehydrogenase (LDH), C-reactive protein (CRP), neutrophil-lymphocyte ratio (NLR) and white blood cell counts, were present. Patient matched SARS-CoV-2 sequences, cytokine profiles and scRNA-seq samples are included. The dataset includes cytokine data from plasma and laboratory data (e.g., neutrophil and lymphocyte counts) from blood samples, alongside clinical data from the patients Electronic Health Records (EHRs). Cytokine profiles were collected by the Korea National Institute of Health (KNIH) using the Luminex MAGPIX system with a customized panel, following a standardized protocol. Additionally, these omics were sampled at multiple time points during for each patient during hospitalization. The number of time points and intervals for sample collection varied across the patients in the cohort. Severity classification was based on the WHO ordinal scale [13] a measurement standard used to assess the clinical status of patients based on the level of oxygen therapy required. It ranges from patients not requiring hospitalization and no oxygen therapy to those needing mechanical ventilation or higher forms of life support, measuring COVID-19 severity on a scale from 0 (i.e., uninfected) to 8 (i.e., death). Patients with a WHO score of 6 or above were categorized as severe and those scored below 6 were categorized as moderate. As a result, 345 and 42 patients were labelled as moderate and severe, respectively. Data from patients infected with SARS-CoV-2 in a total of 22 lineages were collected, including B.1.497 (155 patients), B.1.619 (58 patients), AY.69 (56 patients), B.1.619.1 (34 patients), B.1.620 (19 patients), B.1.1.7 (18 patients), among others. Among the 387 patients, 211 and 176 were male and 176 female patients, respectively. In terms of age, 258 patients were aged 60 or younger, and 129 were aged over 60. Additionally, among the 80 patients with comorbidities, 31 patients had COVID-19 severe outcome related comorbidities, including hyperlipidemia (24), hypothyroidism (5), dyslipidemia (3), pneumothorax (2), obesity (1), pulmonary artery stenosis (1), pulmonary emphysema (1), arrhythmia (1), cardiomegaly (1). The detailed statistics are provided in the Supplementary Table S2. The demographics and clinical statistics are provided in the previous research that describes the COVID-19 cohort data [11].

Computing the H-score

In this research, we aimed to search for regions with high mutation frequency and mutation diversity. The frequency of a nucleotide k at position i is denoted as f_{ik} for $k \subset \{A, T, G, C]$ and the ratio of each nucleotide j at position i is calculated as Eq.1.

$$r_{ij} = \frac{f_{ij}}{\sum_{k \in \{A,T,G,C\}} f_{ik}} \tag{1}$$

Defining the set of the mutated nucleotides at position i as M_i, the ratio of mutations occurred at position i is calculated as Eq.2.

$$P_i = \sum_{k \in M_i} r_{ik} \tag{2}$$

Nucleotide diversity is often measured using entropy, which is the level of uncertainty observing a certain nucleotide [14]. Similarly, we measured the mutation diversity at position i as shown in Eq.3.

$$E_i = \sum_{k \in M_i} \frac{r_{ik}}{P_i} \cdot log_2(\frac{r_{ik}}{Pi}) \tag{3}$$

Under the condition that a mutation occurred at position i, the mutation entropy E_i is computed using the mutation probability of each mutated nucleotide. The mutation probability and diversity of mutations at position i, which is named H-score, is obtained by Eq.4.

$$H_i = log_2(P_i \cdot E_i \cdot 100 + 1) \tag{4}$$

The mutation probability P_i and mutation entropy E_i for the entire sequence of the SARS-CoV-2 genome obtained from GISAID is visualized as a plot in Fig. 2a. A high H-score at a single locus implies a high mutation probability and mutation diversity. Figure 2b (top) shows the average $H\ score$ and the average mutation probability of the identified 477 hotspots. For each value, the 50% percentile is marked by a dotted line for reference. It was observed that mutation hotspot c315 and c442, which showed specific mutation signature in the SMS group, exhibit higher percentiles in terms of H-score (c315: 2.3%, c442: 5.2%) compared to their mutation ratio (c315: 8.4%, c442: 11.5%). This demonstrates the validity of incorporating the entropy of mutations into the mutation probability in identifying severity related mutations. As a comparison, bottom of Fig. 2b visualizes the H-score calculated using the nucleotide entropy instead of the mutation entropy. Here, we can see that c315 and c442 yield significantly lower H-scores compared to other hotspots when mutation entropy is not considered (c315: 99.8%, c442: 97.3%). This justifies the use of mutation entropy, which is calculated under the assumption of mutation occurrence, instead of nucleotide entropy, which is calculated using conventional Shannon entropy.

Density-based mutation hotspot clustering

We employed an in-house developed density-based clustering algorithm, Mutclust, to detect the mutation hotspots. The objective was to identify regions within the SARS-CoV-2 genome with enriched mutations whose mutation diversity was also high, which we refer to as mutation hotspots. This approach is particularly suitable for our data,

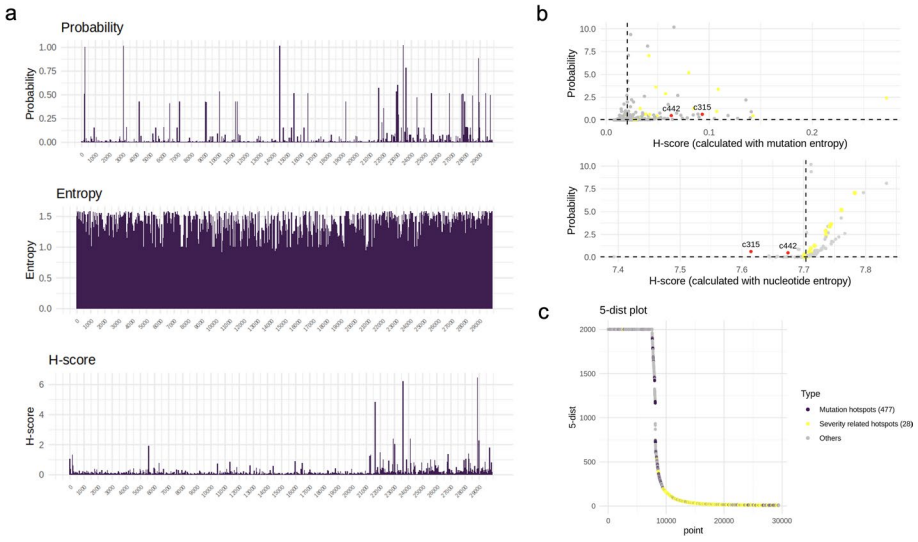

Fig. 2 Bedgraph of the SARS-CoV-2 genome sequence and parametric test of *Mutclust* algorithm. (**a**) The probability, entropy and the H-score of each nucleotide in the SARS-CoV-2 genome is shown. The x-axis represents the nucleotide indices of the SARS-CoV-2 genome. (**b**) The scatter plot of the average H-score and mutation probability of identified 477 hotspots. Here, the H-score is calculated using the mutation entropy (above) and nucleotide entropy (below). The identified 477 hotspots are marked in light gray, the identified 28 severity-related hotspots are in yellow, and hotspots c315 and c442 are highlighted in red. (**c**) The k-dist plot of the parametric test of Mutclust (k = 5). The 5-dist calculation was performed on the SARS-CoV-2 genome sequence. The 8,402 clustered nucleotides belonging to one of the 477 hotspots are marked in blue, and the 1,241 clustered nucleotides belonging to the 28 severity related hotspots are marked in yellow. The remaining nucleotides were color coded in light gray. The x-axis is the index of each nucleotide aligned according to 5-dist. The y-axis represents the 5-dist of the total nucleotides

where regions of interest are characterized by a high frequency of mutations in a one-dimensional space. The most common density-based clustering method is the Density-Based Spatial Clustering of Applications with Noise (DBSCAN [9]), which organizes database data into k-dimensional space, forming clusters such that each contains a minimum number of data points (MinPts) within a given radius (Eps). DBSCAN begins with an arbitrary point p and searches for all points density-reachable from p, forming a cluster that satisfies the core point conditions based on Eps and MinPts. However, using global values for Eps and MinPts is not appropriate for our purpose, which is not only to find clusters of densely packed mutation data points but also to identify clusters with high probability and diversity, i.e., high H-score points. To incorporate the H-score concept, we used a global value for MinPts but a local and dynamically adjusted value for Eps that reflects the H-score.

Mutation clusters are searched in two sequential steps: (1) searching for an arbitrary core point (i.e., locus) from the database that meets the core point condition, and then (2) expanding a cluster that includes the core mutation by searching for all base locations density-reachable from the core point. To create a database that satisfies the core point condition for the first step, we assigned a pre-calculated Eps, which is named Deps, to each base in the genome and identified bases whose H-score statistics within that Deps met a threshold. Here, a core point is defined as a specific mutated locus that satisfies the following four conditions: (1) the H-score of the given base is greater than 0.03, (2) the average H-score within the assigned Deps is greater than 0.01, (3) the sum of H-scores is greater than 0.05, and (4) the count of mutant bases is more than 5 within the assigned Deps. In summary, a core point locus is expected to be highly and diversely mutated,

with sufficient number of closely located neighboring mutations that are also highly and diversely mutated.

The Deps was calculated in proportion to the H-score, applying a larger Deps to locations with a higher probability of mutation diversity and a smaller Deps to lower ones. This effectively assigns a larger Deps to positions with higher H-score. This is done by applying an Eps scaling factor. Bases satisfying the core point condition were termed Core Candidate Mutations (CCMs). In the second step, to define a mutation cluster containing a CCM, the Deps value of the CCM was adjusted to include bases with a high H-score around the CCM while discarding low H-score bases. Specifically, the Deps update was performed to decrease the value if low-H-score bases were present around the CCM, thereby identifying clusters where high-H-score bases are more densely packed. If the number of mutations within the defined Deps boundaries of a cluster satisfies MinPts, the region was designated as a cluster. The cluster is then expanded to both left and right directions in the genome as long as the MinPts condition was met, which is the stopping condition of searching a single cluster. For each expansion, the Deps was scaled down by a pre-defined diminishing factor to prevent a cluster to expand too much. The final cluster is defined as a mutation hotspot. This process is repeated for each of the identified CCM across the entire SARS-CoV-2 genome, which will yield in the same number of clusters as the number of searched CCMs as defined in step 1. The initial Deps for searching CCMs is computed by Eq. 5. The diminishing Deps that is used during the cluster expansion starting from the searched CCMs is computed by Eq. 6. Here, γ and ∂ refer to the eps scaling factor and diminishing factor, respectively. Deps is diminished while increasing the distance from the CCM by 1 at each step. If the current distance exceeds the updated Deps, the diminishing process terminates, and the distance is set to Eps. The diminishing concept of Deps is detailed in Supplementary Figure S1.

$$Deps_{CCM} = \lceil \gamma \cdot H_i \rceil \tag{5}$$

$$Deps_{CCM} = Deps_{CCM} - \frac{(Deps_{CCM} - Deps_i)}{\partial} \tag{6}$$

Parametric tests were conducted to search the optimal settings for the hyperparameters MinPts, epsilon scaling factor γ, and diminishing factor ∂. MinPts defines the required minimum number of mutated loci within a hotspot, while γ and ∂ determine the local Eps value, with the purpose of the ∂ being to enhance the density of high H-score bases in the cluster by discarding low H-score bases. The parametric k-dist test for DBSCAN involves computing the distance to the k-th nearest neighbor for all points p and identifying the valley point where the cluster boundary is determined through the k-dist plot. Since Mutclust aims to discover clusters on the genome where H-score is high, we calculated the k-distance, which satisfies the core point condition of CCMs for the radius within H-score statistics and visualized the valley through the k-dist plot, naming it 5-dist due to the core point condition of CCMs being 5, which is the MinPts. The valley on the 5-dist plot was observed at 500 distance. This allows for the reasonable identification of dense and diversely mutated clusters if the 5-dist of bases belonging to the hotspot region is less than 500, thereby configuring the parameters to ensure such

condition. Consequently, γ was set to 10 and ∂ to 3. The 5-dist plot for the performed parametric test is presented in Fig. 2c.

To compare the ability to detect mutation clusters between MutClust and other density-based algorithms, we performed DBSCAN clustering using the same dataset of 29,903 nucleotide positions used in our study. While MutClust performs density-based clustering based on an H-score that incorporates both mutation percentage and entropy, we applied DBSCAN using only mutation percentage and entropy as two-dimensional input features for a fair comparison. Following the same setting as our study, the minimum number of points (MinPts) was set to 5. The optimal ε value was determined as 0.15 based on the k-distance plot with k = 5. With these parameters, DBSCAN clustering produced a total of four clusters, with the following sizes: Cluster0 (29,831; 99.76%), Cluster1 (9), Cluster2 (16), Cluster3 (10). Additionally, 37 data points (0.12%) were labeled as noise. This result indicates that DBSCAN primarily captured a single large cluster across the distribution, with the remainder either categorized as noise or forming very small clusters. Although DBSCAN is a density-based algorithm and is theoretically suitable for this type of data, it lacks local adjustment mechanisms for density variation, and thus failed to effectively separate key mutation clusters. In particular, the fixed global parameters (ε and MinPts) prevented the algorithm from detecting dense but biologically important regions that may differ in shape or scale. The results of the DBSCAN clustering have been saved in the file Supplementary Table S3. The DBSCAN clustering result figure has been provided as Supplementary Figure S2.

Selection and evaluation of severity related hotspots

To select the severity related hotspots, feature selection was conducted to differentiate patients by severity. The SelectKBest function in the python scikit-learn [15] package was used with k set to 'all' and ANOVA [16] for statistical testing. Default settings were used for other parameters. Clustering analysis was then performed using the default settings of the clustermap [17] function.

To evaluate the significance of the identified severity related mutation hotspots, bootstrap testing was conducted for each hotspot to assess its statistical significance in terms of mutation density and H-score. Random genomic regions of the same size as the clusters were selected, and within these random windows, the number of mutations, the mean H-score, and the sum of H-scores were computed. This process was repeated 1,000 times for each cluster to generate a distribution of expected values under the null hypothesis of no significant clustering. The results of the bootstrap testing were used to calculate the *p*-values, which reflect the probability of observing a cluster with equal or greater mutation density and H-score purely by chance. The Benjamini-Hochberg [18] procedure was applied to adjust the *p*-values for multiple testing.

Network propagation analysis

To investigate the divergence of immune response across patient groups characterized by the infected virus hotspots, we employed network propagation on a gene-gene interaction network. This network analysis technique simulates the effect of seed nodes, which in this context are viral genes, across the whole gene interaction network consisting of interactions among human and viral genes. For network preparation, we constructed a gene-gene interaction network based on public protein-protein interaction

database, STRING [19] and Biogrid [20]. To narrow our focus to viral genes relevant to COVID-19, we filtered the viral gene set based on the virus taxonomy, specifically targeting species within the *Coronaviridae* family. Target species include *Human coronavirus 229E*, *Human SARS coronavirus*, *Human coronavirus EMC*, *Severe acute respiratory syndrome coronavirus 2*, and *Severe acute respiratory syndrome-related coronavirus*. After species filtering, networks from STRING and Biogrid are concatenated. Given that our research focused on understanding how the human immune response against COVID-19 infection is mediated, we filtered the network using immune-related genes in mSigDB [21]. For seed preparation, we identified viral genes for each patient group. We profiled mutation count of viral hotspots for each COVID-19 patient, collecting hotspots with mutation count over 2. Utilizing SARS-CoV-2 reference genome from ENSEMBL [22] database, we assigned genes to each hotspot region if a gene has an overlap with hotspot region in reference genome. The set of all viral genes in the collected hotspots for each patient group was considered as the seed genes. In the network propagation process, given a network G = (V, E) and seed nodes $V_s \subset V$, we initially assigned node resources to V_s and then iteratively propagates node resources to the neighboring nodes in G until convergence. Propagation step at the t-th iteration can be formulated as follows, where W represents adjacency matrix of the graph. The hyperparameter α determines the ratio of retaining the initial node resources ensuring the effects of seed nodes persist during propagation, where α is set to 0.01. p_0 is a binary vector of initial node resources, where initial node resources of the seed genes are set to 1. p^t is a vector representing node resources at t-th step. After convergence, the amount of resources for each node represents the effect of seed genes on the node. We subsequently filtered the human genes exclusively to prioritize the most affected human genes given the viral genes of each cluster as seed nodes.

Differentially expressed gene analysis

Differential gene expression analysis between the two patient groups was performed using edgeR (version 3.32.1) [23]. Genes with zero read count in more than 80% of the samples were excluded, resulting in 12,537 genes, which were subject for downstream analysis. The statistical testing employed was Fisher's exact test [24]. The *p*-values were adjusted using the Benjamini-Hochberg (BH) method. Genes with a p-adjusted value of 0.05 or lower were considered significant.

Validation on influenza genome

A total of 276,910 virus sequences were downloaded from the GISAID database. Only strains with more than 10,000 sequences were included in the analysis. The strains analyzed were A-H1N1 (26,703 sequences), A-H3N2 (52,734 sequences), and B-Victoria (197,473 sequences). For A-H1N1 and A-H3N2, *Alphainfluenzavirus influenzae* were used as the reference sequence (NCBI [25] taxonomy ID: 11320), while *Betainfluenzavirus influenzae* were used for B-Victoria (NCBI taxonomy ID: 11520).

Results

A total of 224,318 SARS-CoV-2 genome sequences were collected from the Global Initiative for Sharing All Influenza Data (GISAID) database. Using the Wuhan-Hu-1 genome as the reference genome, mutation frequency and diversity were computed for each

locus. First, the ratio of mutation was computed for each nucleotide at a specific locus by dividing the frequency of mutation occurrence by the total depth, thus, obtaining the mutation occurrence ratio for each nucleotide. Second, to identify regions where a mutation occurs with high diversity, the ratio of each possible nucleotide specific mutation occurrence was computed, given the condition that a mutation has occurred at that position. The mutation probability and diversity at every locus in the SARS-CoV-2 were integrated to compute a single H-score. Regions that contain significantly high H-scores within a dense range were searched, which were labeled as mutation hotspots. Hence, mutation hotspots are defined as regions that include a sufficient number of mutations with high mutation diversity. A mutation hotspot with a high average H-score implies it includes mutations with high mutation ratio and diversity.

Mutation hotspots

Among the 29,903 nucleotides in the genome, 29,409 (98.3%) were assigned a H-score. The small portion of prefix and suffix sequences were not considered for mutation analysis due to their unlikely high sequence variance, which would not be severity related. As a result, 477 clusters with sufficiently high average H-score were detected. The searched hotspots and their corresponding H-scores across the entire sequence of the SARS-CoV-2 genome are shown in Fig. 3. Among the 477 mutation hotspots, 88 contained lineage dividing mutations. The genomic details of the 477 mutation hotspots are provided in Supplementary Table S4 and their associated lineage information is provided in Supplementary Table S5. The spike protein in SARS-CoV-2 is known to be associated with virus infection and pathogenicity due to its involvement in host ACE2 receptor binding and membrane fusion [26]. Previous meta-analysis studies showed that it was enriched with mutations [27]. A manifest of well known mutations of SARS-CoV-2 and their functions are summarized in Table 1. Interestingly, 9 out of the 10 mutations, except S943P, were found within our identified mutation hotspots. V483a is known to induce strong drug resistance [26]. E484K contributes to rapid and strong virus transmission, and E484Q is also associated with increased infectivity and transmission [28]. Q677 exists as two variants: Q677P and Q677H, contributing to host cell entry of SARS-CoV-2

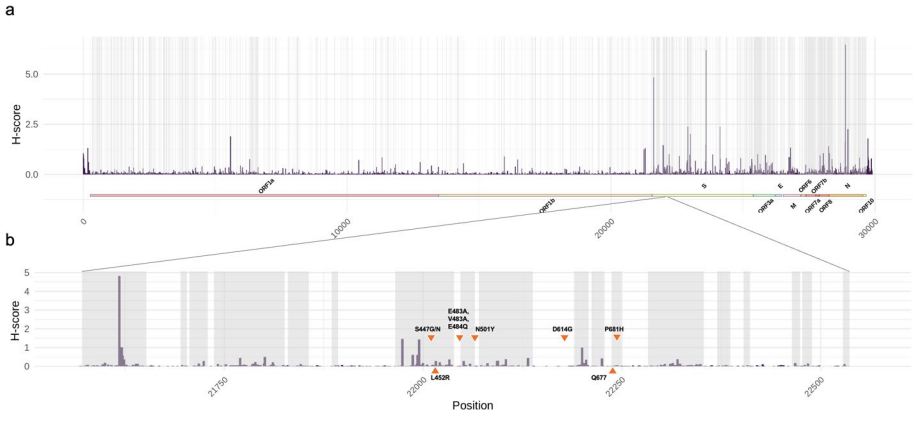

Fig. 3 The bedgraph of the H-score and identified hotspots of the SARS-CoV-2 genome. (**a**) The bedgraph of the entire SARS-CoV-2 genome sequence. The H-scores are shown as blue bars and the 477 mutation hotspots are labeled in light gray. The x-axis represents the indices of nucleotides from the first to the last of the SARS-CoV-2 genome. The corresponding gene annotations are color coded. (**b**) The bedgraph for part of S proteins and the locations of nine well-known mutations of SARS-CoV-2. The H-scores and mutation hotspot of bases 21,563 – 22,562 are shown, and 9 mutations with known functions are indicated

Table 1 A list of functional mutations in SARS-CoV-2 in association to the mutation hotspots

Mutation	Effect	Hotspot
D614G	Increased transduction in many cell types, resistant to proteolytic cleavage, 4–9 times more contagious	c346
S943P	Result of recombination of different viruses in an infected host	-
V483a	Strong drug resistance	c338
E484K	Rapid spread mutation	c338
N501Y	Increased affinity to ACE2, Related with receptor binding neutralizing antibodies	c340
L452R	Weakened antibody neutralization, increased the virus's infection ability	c334
Q677	Enabled virus to enter the human cells more easily due to its Q location	c350
P681H	Potentially causes breakdown of the disulfite bridges in and around RBD	c350
E484Q	Increased infection or spread	c338
S447G/N	Increased binding affinity for hACE2 is observed	c337

[29]. P681H was found to potentially disrupt disulfide bridges near the receptor-binding domain (RBD) [30]. Collectively, the identified hotspots showed to successfully encompass functional mutations, including the highly diverse mutation Q677. The locations of the 9 well-known mutations that were present in the set of mutation hotspot are depicted in Fig. 3.

Mutation hotspots of moderate and severe patients

To investigate the correlation between COVID-19 severity and the identified mutation hotspots, clinical data from 387 patients along with patient matched SARS-CoV-2 sequence, cytokine and single-cell RNA-seq samples were analyzed. The data is a large multi-omics dataset collected from a COVID-19 cohort in South Korea that were sampled across three hospitals (i.e., Chungnam National University Hospital, Seoul Medical Center, and Samsung Medical Center) as a national project initiated by the Korea Disease Control and Prevention Agency (KDCA). In terms of severity, 345 patients experienced moderate illness and 42 severe illness according to the World Health Organization (WHO) scale. Each mutation hotspot was tested for significant association with one of the severity patient groups. A total of 28 hotspots showed significant FDR corrected p-values, which were subject to downstream analysis. To evaluate the statistical significance of the 28 identified hotspots associated with severity, bootstrap testing was conducted for each cluster to assess the statistical significance of mutation density and H-score metrics compared to random expectations. The details of the 28 mutation hotspots are summarized in Table 2. To identify patient groups that share similar mutation profiles in respect to the 28 hotspots, hierarchical clustering was performed, which result is shown in Fig. 4a. Patients were clustered based on the number of mutations within each hotspot. Clustering divided the patients into 7 groups, each showing distinct characteristics regarding severity. Clusters with sufficient number of patients (more than 10) were used for subsequent analysis. As a result, clusters 4, 5, and 7, which included 1, 5, and 1 patients respectively, were excluded, and clusters 1, 2, 3, and 6, which comprised of 123, 186, 18, and 53 patients respectively, were selected for further analysis. The selected clusters were further categorized into two groups, that showed clear difference in severity, which we labeled as Moderate Mutation Signature (MMS) and Severe Mutation Signature (SMS) patient groups. Based on the severity, each cluster was assigned to either the MMS or SMS group. Cluster 6, with the highest proportion of severe patients, was assigned to the SMS group, while clusters 1, 2, and 3, which had a relatively higher

Table 2 The position and statistical information of the 28 severity related mutation hotspots

Gene	Hotspot	Position	Mutation ratio	*P*-adj	MMS average	SMS average
ORF1a	c22	1055–1072	0.66	0.0291	0.52	0.038
ORF1a	c90	3808–3821	0.64	0.0291	0.0	0.24
ORF1a	c118	5171–5185	0.73	0.0291	0.0	0.96
ORF1a	c123	5574–5587	0.57	0.0	0.0	0.89
ORF1a	c124	5612–5623	0.67	0.0397	0.22	0.0
ORF1a	c198	11,514–11,544	0.68	0.0291	0.30	0.96
ORF1b	c239	16,462–16,471	0.7	0.0	0.0	0.98
ORF1b	c258	18,024–18,032	0.67	0.0293	0.48	0.0
ORF1b	c298	20,675–20,688	0.5	0.0291	0.41	0.0
S	c309	21,571–21,652	0.88	0.0	0.46	1.02
S	c315	21,965–22,039	0.97	0.0	0.24	7.04
S	c319	22,212–22,228	0.82	0.0291	0.03	0.96
S	c334	22,874–22,924	0.69	0.0	0.33	1.0
S	c337	22,982–23,002	0.76	0.0	0.095	1.0
S	c350	23,580–23,616	0.84	0.0	0.16	1.0
S	c364	24,406–24,418	0.62	0.0108	0.0	0.98
ORF3a	c385	25,665–25,728	0.91	0.0132	0.003	1.07
ORF3a	c390	25,899–25,939	0.83	0.0213	0.22	0.0
M	c412	26,759–26,776	0.72	0.0	0.30	1.0
ORF7a	c429	27,453–27,696	0.94	0.0291	0.48	1.98
ORF7a, ORF7b	c431	27,727–27,770	0.93	0.0291	0.003	1.02
ORF8	c438	28,065–28,126	0.90	0.0083	0.07	0.94
ORF8,N	c442	28,230–28,404	0.79	0.0	1.23	9.04
N	c444	28,456–28,462	0.71	0.0242	0.0	1.0
N	c460	29,108–29,122	0.89	0.0291	0.0	0.32
N	c462	29,169–29,185	0.8	0.0291	0.48	0.0
N	c468	29,353–29,452	0.65	0.0159	0.034	0.98
ORF10	c474	29,613–29,633	0.84	0.0291	0.22	0.0

proportion of moderately severe patients, were assigned to the MMS group for subsequent analysis. The mutation hotspots c315 and c442 showed higher mutation counts in the SMS group compared to the other 26 hotspots and mainly contributed to the difference between MMS and SMS. Among the 28 severity-associated hotspots, c315 and c442 exhibited the largest differences in average mutation counts between SMS and MMS groups (c315: 6.80, c442: 7.81), and were the only hotspots with an average mutation count above 7 in the SMS group (c315: 7.04, c442: 9.04). These features led to their selection for downstream analyses.

Several cytokines showed significant differential expression between the two groups, with CXCL10 being an exception, indicating a possible link between mutation hotspots density, diversity and immune biomarkers. Especially, interferon γ (IFNG) and tumor necrosis factor (TNF) showed elevated levels in the SMS group. These findings suggest a correlation between the mutational landscape of the virus and the immune response in COVID-19 patients, as depicted in Fig. 4b. Elevated TNF levels, significant in COVID-19 pathophysiology for potentially triggering cytokine release syndrome (CRS) and enhancing SARS-CoV-2's interaction with ACE2 receptors, were notably higher in patients with severe COVID-19, aligning with previous observations [31]. Furthermore, clinical laboratory markers that indicate immune response activation, including lactate dehydrogenase (LDH) and the neutrophil–lymphocyte ratio (NLR). These markers showed

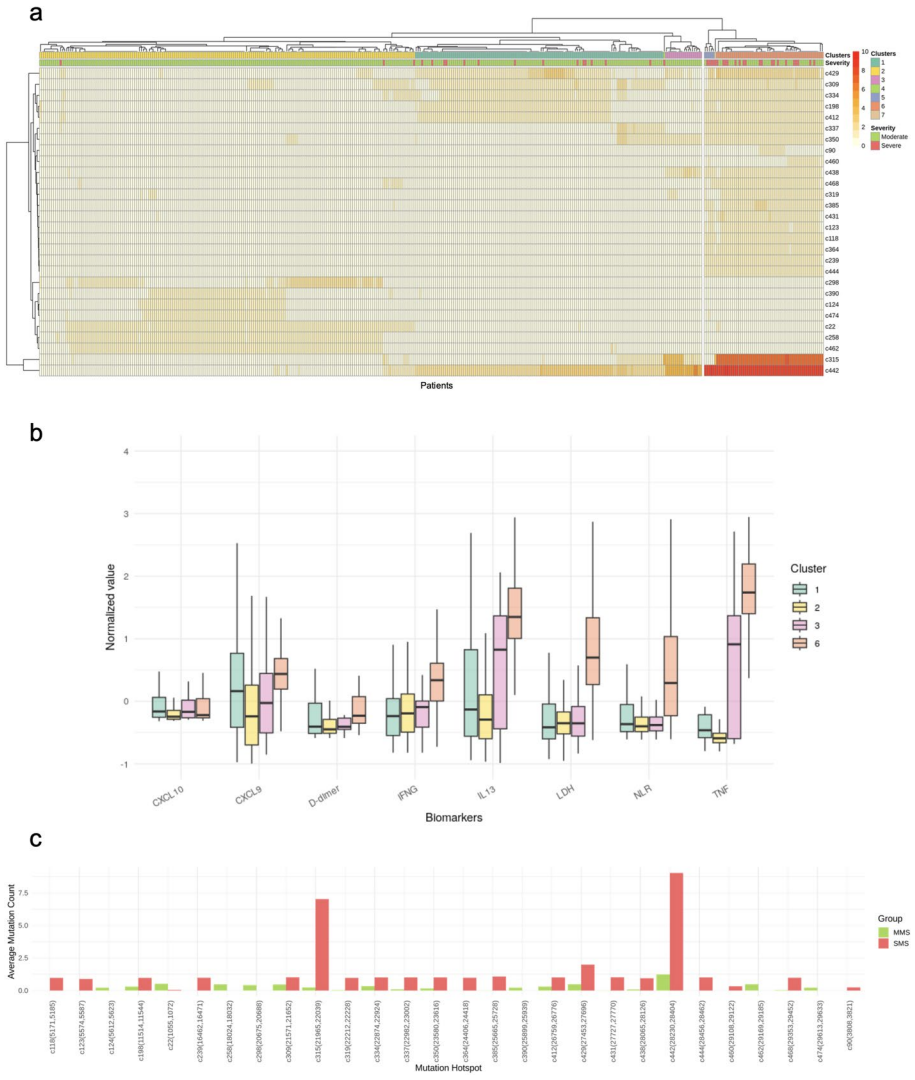

Fig. 4 Differential immune response patterns in the MMS and SMS patient groups. (**a**) The heatmap of the clustered COVID-19 patients using the 28 severity-related mutation hotspots. The rows represent the 387 patients, and the columns represent the 28 hotspots. Patients were clustered based on the number of mutations within the hotspots. The color scale indicates the number of mutations in each hotspot. (**b**) Box plots comparing cytokine levels and clinical laboratory markers between SMS (cluster 6) and MMS (cluster 1,2 and 3) patient groups. The y-axis refers to the concentration or measurement units for each marker. (**c**) Bar plot visualizes the average mutation counts across the 28 hotspots for both MMS and SMS groups. Hotspots c315 and c442 are hotspots with an average mutation count above 7 in the SMS group (c315: 7.04, c442: 9.04)

significantly higher levels in the SMS group, implying a stronger immune response or inflammation in patients with severe conditions.

Hotspots c315 and c442 were identified as key factors distinguishing SMS and MMS groups, and their potential impact on immune response warrants further investigation. Immunologically, lower HLA affinity results in shorter epitope presentation times on the APC surface [32, 33]. The presentation of epitopes by APCs can influence the screening process of T and B cells during the adaptive immune response, potentially affecting the maturation of T and B cells specific to those epitopes. To investigate how mutations in hotspots c315 and c442 impacted HLA affinity in our dataset, we utilized HLA genotype data from COVID-19 patients. Of the 387 patients, HLA data were available for 213 individuals. A total of 39 HLAs were included in the analysis, comprising the

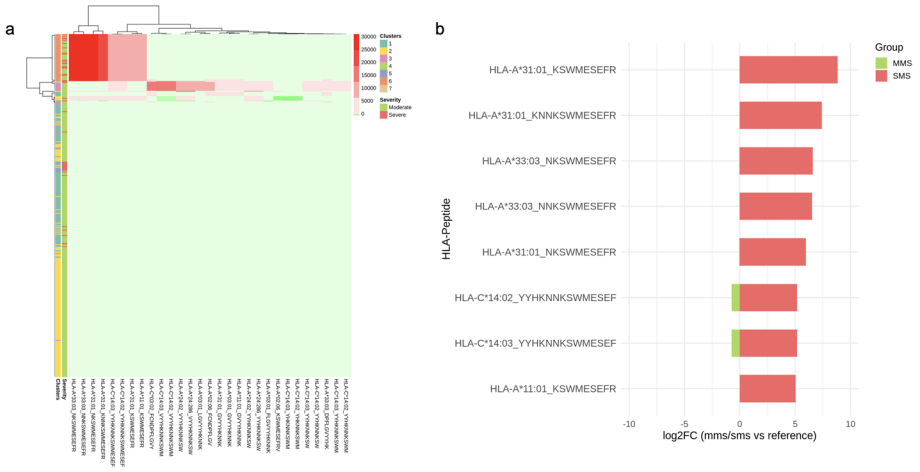

Fig. 5 Visualization of HLA–epitope affinity changes induced by mutations in c315. (**a**) Heatmap visualization of HLA–epitope pair affinities that showed mild or strong reference affinity (< 500 nM) based on epitopes derived from the reference sequence. The x-axis represents 29 HLA–epitope pairs, and the y-axis represents 387 patients. (**b**) Visualization of the top 8 HLA–epitope pairs from hotspot c315 that showed the largest decrease in affinity in the SMS group compared to the reference. Red bars indicate the log fold change (LFC) in affinity for the SMS group, while green bars represent the LFC for the MMS group

Table 3 Affinity scores (nM) and log2FoldChange for the top 8 HLA-peptide pairs from c315

Allele	Peptide	Affinity (Reference)	Affinity (MMS)	Affinity (SMS)	LFC (MMS)	LFC (SMS)
HLA-A*31:01	KSWMESEFR	17	17.5	8342	0.04	8.94
HLA-A*31:01	KNNKSWMESEFR	126	129	21,609	0.03	7.42
HLA-A*33:03	NKSWMESEFR	293	304	28,571	0.05	6.61
HLA-A*33:03	NNKSWMESEFR	340	350	31,303	0.04	6.52
HLA-A*31:01	NKSWMESEFR	402	418	25,382	0.05	5.98
HLA-C*14:03	YYHKNNKSWMESEF	227	382	8250	0.75	5.18
HLA-C*14:02	YYHKNNKSWMESEF	227	382	8250	0.75	5.18
HLA-A*11:01	KSWMESEFR	279	283	0.0293	0.02	5.05

union of the top 10 most frequent HLAs for six HLA genotypes. The affinity scores for 4,095 HLA-epitope pairs in c315 and 13,104 pairs in c442 were calculated. Among these, 29 pairs in c315 and 9 pairs in c442 exhibited mild or strong reference affinity (affinity score < 500 nM). To evaluate the impact of mutations on affinity, the differences from the reference affinity scores were analyzed across the 387 samples (Fig. 5). The results showed that mutations in c315 from the SMS group significantly reduced the affinity in 8 out of 29 HLA-epitope pairs. Notably, HLA-A11:01 and HLA-C14:02, which were mentioned in a study on a Chinese cohort as significantly associated with severe disease or worse outcomes, were among the HLAs affected [34]. Although this trend was not observed in c442, these findings suggest that mutations in c315 may contribute to severe outcomes by reducing affinity for certain HLAs. This analysis is based on potential epitope screening using in silico methods, and additional studies, such as in vitro validation, are required to establish a definitive relationship with HLA affinity. The affinity scores for top 8 pairs in c315 are presented in Table 3. The HLA phenotypes of the 213 patients used in this analysis, along with their corresponding patient-derived epitopes, are provided in the Supplementary Tables S6 and S7.

Network propagation analysis on the mutation hotspots

To investigate the gene-level response to viral infection, we prioritized human genes that are plausibly affected by the viral genes embedding one or more of the 28 mutation hotspots defined above using network propagation analysis on the STRING database [19]. Especially, we were interested on how the gene-level response differed between the moderate and severe patient groups, and thus indirectly investigating the difference in mutation hotspot compositions between the patients. The network propagation was performed for each patient group. Collectively, the question was to answer whether SARS-CoV-2 genes with mutation hotspots effected genes in the human genome and whether the effected genes showed different biological functions between the moderate and severe patient groups. The top 100 human genes with the highest propagation score were selected from the SMS and MMS groups, which were subject for gene enrichment analysis, using EnrichR [35]. The MMS group showed high enrichment in antimicrobial humoral response, as shown in Fig. 6, in which Immunoglobulin A (IgA) and Immunoglobulin M (IgM) are known to neutralize viral response [36, 37]. Natural killer (NK) cell mediated cytotoxicity was the most over-represented biological process in human genes in the SMS group. After the SARS-CoV-2 virus enters a target cell, the NLRP3 is activated leading to a cascade of reactions involving the secretion of IL-1β and IL-18 that in turns activate NK cells to secrete IFN-γ. IFN-γ has been identified as a prominent cytokine in severe COVID-19 patients [38, 39]. This suggests that SARS-CoV-2 has evolved mechanisms to inhibit and delay the induction of the type 1 IFN response. In peripheral blood of severe COVID-19 patients, elevated NK cell activation and increased presence of adaptive NK cells are reported [40].

To further validate the network propagation analysis, we conducted DEG analysis on the scRNA-seq samples across various timepoints to spot any transcriptomic difference between MMS and SMS. Especially, genes related to the NK cell compartments were of particular interest. Among the 387 patients, scRNA-seq data was available for 106 patients, from whom single-cell pseudobulk samples were obtained. The number of longitudinal samples between the patients varied from 1 to 7 timepoints, resulting in a total of 261 and 92 expression samples for the MMS and SMS groups, respectively. DEG

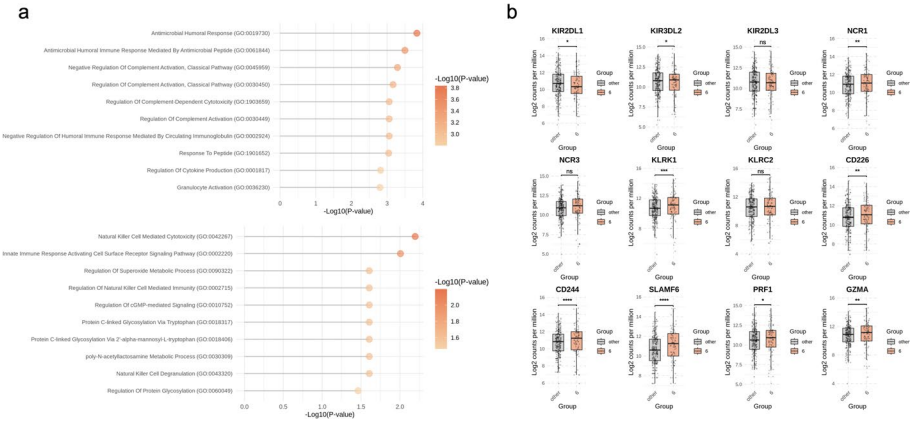

Fig. 6 GO enrichment analysis of human genes and differentially expressed genes in the MMS and SMS patient groups. (**a**) The over-represented GO terms in the MMS group and SMS group. The length of the bars represents the *p*-value ranking in ascending order, where the longest bar has the lowest *p*-value. (**b**) The RNA expression levels of for four inhibitory receptor genes, three activating receptor genes, three activating co-receptor genes, and genes for granzymes and perforin. The level of significance is indicated by asterisks (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$)

analysis was performed between the two groups resulting in 3,951 DEGs. The DEG analysis results showed that NK cells exhibited significant increases in the expression of activating receptors and activating co-receptors, while the expression of inhibitory receptors significantly decreased. NK cells play a crucial role in controlling viral infections at the innate immune stage, and their activation is regulated by a balance between activating signals triggered by ligand recognition of activating receptors and inhibitory signals mediated by recognition of major histocompatibility complex class I (MHC-I) molecules by inhibitory receptors [41]. Our findings indicated a decrease in the expression of inhibitory receptors KIR2DL1, KIR2DL3, and KIR3DL2, with significant difference in expression levels for KIR2DL1 and KIR3DL2. Furthermore, increased expression levels, but without significance, of activating receptors NCR1, NCR3, NKG2D, NKG2C, and activating co-receptors CD226, CD244, CD352 were observed, with significant expression level increase for NCR1, NKG2D, CD226, CD244, and CD352. This is characterized by all inhibitory receptors with significant p-values showing decreased expression, and all activating receptors and co-receptors with significant p-values showing increased expression. The significant enrichment of the "Innate Immune Response Activating Cell Surface Receptor Signaling Pathway" in the SMS group further supports such result. Additionally, the PRF1 and GZMA upregulated in NK cells in the SMS group, which are known to release cytotoxic granules containing granzymes and perforin or the expression of cytokines that direct the immune response in a specific direction and attract other cells to the site of infection [42]. While not in the scRNA-seq data, the cytokines IFN-γ and TNF-β were significantly upregulated in the SMS group, which also takes part in NK cell activation. Collectively, the increased expression of activating receptors and the decreased expression of inhibitory receptors in NK cells led to an increase in NK cell activation. These results suggest that severe COVID-19 outcomes may be associated with functional disruption of NK cells due to changes in the expression of activating and inhibitory receptors. The differentially expressed NK cell receptor genes and NK cell function related genes between MMS and SMS groups are shown in Fig. 6b.

Discussion

In this study, we identified regions within the SARS-CoV-2 genome where mutations with high diversity are densely located, termed mutation hotspots, and subsequently conducted clustering analysis on COVID-19 patients based on the number of mutations in these hotspots. Following network propagation analysis and DEG analysis highlighted significant differences in immune responses mediated by NK cells within severe clusters. The network propagation results showed significant changes in NK cell function in the SMS group. Also, this was coherent with the DEGs searched in the SMS group, where the expression of activating receptors and a decrease in inhibitory receptors in NK cells were increased.

The immune function of NK cells plays a critical role in controlling viral infections at the innate immune stage, and the activation of NK cells is regulated by a balance between activation and inhibitory signals [38]. An interesting aspect of our results is the significant increase in the expression of activating NK cell receptors and co-receptors, and a significant decrease in the expression of inhibitory receptors. An increase in NK cells with elevated expression of the activating receptor NKG2C and restricted expression of the inhibitory receptor KIR has been observed in severe COVID-19 patients

[39, 40]. This could be attributed to the accumulation of adaptive-like NK cells due to the increased resistance to cytokine-induced apoptosis, or indirectly due to the release of excessive pro-inflammatory cytokines [38]. This suggests that the complementary changes in the expression of activating and inhibitory receptors in NK cells may have a significant relationship with severe outcomes. Typically, when NK cells encounter infected cells, the inhibitory signals are reduced or absent, leading to the predominance of activating signals and activation of NK cells [43]. Since inappropriate activation of NK cells can pose risks to healthy cells, the balance between activation and inhibitory signals plays an important role in the normal activation of NK cells [41]. Therefore, the imbalance between these signals indicate that NK cells may have deviated from their normal activation state. Indeed, severe COVID-19 patients have been associated with an increase in NK cell population followed by high levels of cytotoxic proteins such as perforin because of excessive production of pro-inflammatory cytokines [42]. The significant increase in the expression of perforin and granzymes in severe patients agree with our DEG results.

Viruses disrupt the immune function of NK cells through various mechanisms. The viruses Human Immunodeficiency Virus (HIV-1), Human Cytomegalovirus (HCMV) and Kaposi's Sarcoma-associated Herpesvirus (KSHV) commonly suppress the expression of ligands on the surface of infected cells to evade NK cells and thus escape cell destruction using proteins such as, NEF [44] the viral glycoprotein UL142 [45] and ORF54 [44], respectively. The Human T-cell Lymphotrophic Virus (HTLV-1) uses another mechanism to escape NK cells, where it suppresses the expression of intercellular adhesion molecules 1 and 2 (ICAM-1 and ICAM-2) via p12I to hinder the attachment of NK cells to the infected cells [46].

These examples show how specific proteins encoded by viruses can affect the recognition and elimination functions of NK cells, thereby disrupting their immune functions. There is not much knowledge regarding the disruption of NK cell immune functions in SARS-CoV-2 infected cases. A known aspect is the overexpression of HLA-E in immune and stromal cells in COVID-19 patients, with the SARS-CoV-2 spike protein implicated [47]. Also, the non-structural protein 13, Nsp13, of SARS-CoV-2 encodes a peptide that forms a stable complex with HLA-E that prevents binding to the inhibitory receptor NKG2A, thereby exposing target cells to NK cells [48]. Previous findings regarding the disruption of NK cell immunological functions by virus infections have mostly focused on strategies to evade elimination by NK cells. Our results suggest that the expression of activating and inhibitory receptors in NK cells occurs in a complementary manner, potentially leading to excessive activation of NK cells.

Our results demonstrate a significant imbalance in the expression of activating and inhibitory receptors of NK cells within severe patients, further supported by the network propagation results related to the SARS-CoV-2 hotspots. Considering that MMS and SMS were formed based on the patient specific mutation hotspot profiles, we suggest that NK cells are involved in pathways contributing to severe outcomes associated with mutations identified in these hotspots, providing an additional mechanism by which SARS-CoV-2 disrupts NK cell functions. The relationship between the observed results for NK cells is shown in Fig. 7. It presents a hypothesis-driven conceptual diagram integrating correlation-level findings from mutation burden, cytokine imbalance, and transcriptomic network analysis of NK cell receptor expression. This figure is not based on

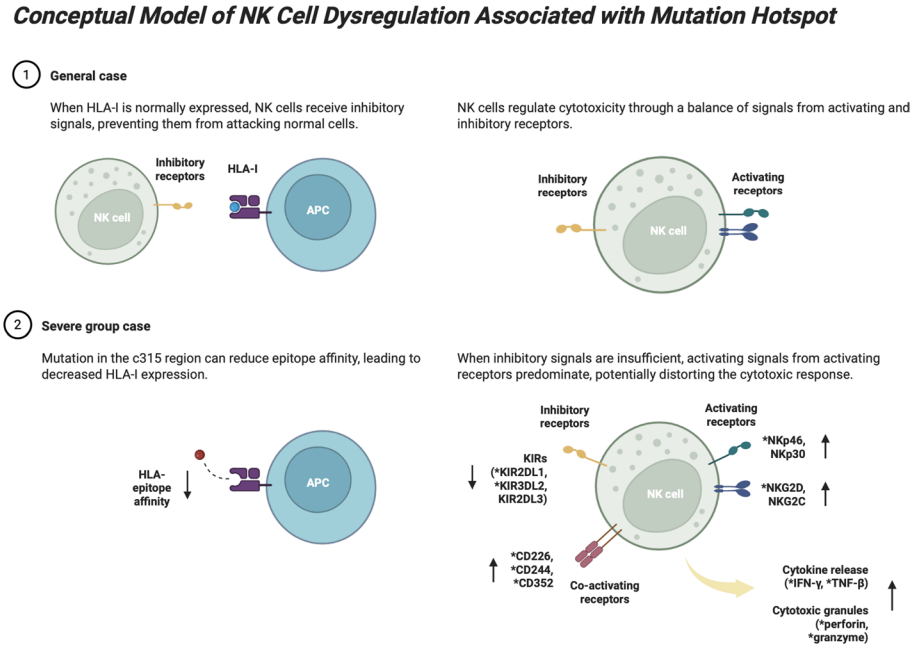

Fig. 7 Relationship between observed changes in NK cells. Diagram summarizing observed changes in NK cell receptor expression and functional alterations. This is a conceptual model based on mutation burden, cytokine imbalance, and transcriptomic network analysis

direct experimental evidence, such as HLA-binding assays, peptide processing validation, or NK cytotoxicity tests, and therefore should not be interpreted as a mechanistic pathway. Further validation using functional immunological assays is needed in future studies.

The approach of the Mutclust algorithm can be applied to two-dimensional sequence data containing position-specific frequency information. Therefore, this algorithm can also be used to identify mutation hotspots in other viruses. For validation, Mutclust was applied to the influenza virus genome to show that mutation hotspots are also present in other viruses. A total of 276,910 virus sequences from the GISAID database were collected, including the A-H1N1, A-H3N2, and B-Victoria strains. Influenza viruses are comprised of eight genome segments, each with lengths ranging from 890 to 2,341 base pairs [49]. The Mutclust algorithm was applied using the same parameters as in the SARS-CoV-2 genome analysis, and a total of 24 runs were performed (i.e., eight segments of the three strains). The analysis identified up to 39 mutation hotspots (B-Victoria, PB1) per segment. It was confirmed that some of these hotspots included mutations associated with severe outcomes and mortality. For example, the D222G mutation, included in hotspot c4 of H1N1-HA, is known to be associated with high mortality, while the T123V mutation, included in hotspot c3 of H1N1-NS1, was related to severe outcome [50]. The E627K mutation in c31 of H2N2-PB2 enables the replication of influenza viruses carrying PB2 from a poultry source to human respiratory epithelial cells [51]. Additionally, the K338R mutation in hotspot c14 of B-PB2 has been associated with increased pathogenicity of the virus [52]. These findings demonstrate that Mutclust effectively identified clinically significant hotspots in influenza. Influenza viruses have eight genome segments, approximately one-tenth the length of SARS-CoV-2. Therefore, some parameters optimized for SARS-CoV-2 may not be fully suitable for the influenza

genome, and additional parameter tuning may be required. Further research is needed for applications to influenza and other viruses. The bedgraph for the identified hotspot regions in influenza are shown in Supplementary Figure S3.

This study opens avenues for further research into the functional implications of these mutations and their role in the pathogenesis of COVID-19, potentially aiding in the development of more effective vaccines and therapeutic interventions. Also, we showed that such mutation hotspots are also present in the influenza virus, suggesting extended analysis to other viruses than COVID-19.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-025-00476-3.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
The authors of this paper thanks the Division of Healthcare and Artificial Intelligence group within the Korea National Institute of Health for constructing the valuable multi-omics COVID-19 cohort dataset.

Author contributions
I.J., S.K: conceptualization, data curation, writing and review; S.Y., H.K, D.J: Software development, investigation, data curation, visualization, writing and review; E.H.: writing, review, editing and visualization; I.J.: supervision, project administration and funding acquisition.

Funding
This work was supported by the project for Infectious Disease Medical Safety, funded by the Ministry of Health and Welfare, South Korea (grant number: RS-2022-KH124555 (HG22C0014)), the "Korea National Institute of Health" (KNIH) research project (project no. 2024-ER-0801-01). This research was also supported by a grant of the project for 'Research and Development for Enhancing Infectious Disease Response Capacity in Medical&Healthcare settings', funded by the Korea Disease Control and Prevention Agency, the Ministry of Health & Welfare, Republic of Korea (grant number : RS-2025-02307351).

Data availability
The dataset of the KNIH cohort used in this study are available in the Clinical and Omics Data Archive (CODA, https://coda.nih.go.kr) database by the accession number CODA_D23017.

Code availability
Code and related data are available in the following repository https://github.com/cobi-git/mutclust

Declarations

Competing interests
The authors declare no competing interests.

Received: 28 March 2025 / Accepted: 15 August 2025
Published online: 01 September 2025

References

1. Huang F, et al. Identifying COVID-19 severity-related SARS-CoV-2 mutation using a machine learning method. Life. 2022;12:806.
2. Kim J, Cheon S, Ahn I. NGS data vectorization, clustering, and finding key codons in SARS-CoV-2 variations. BMC Bioinformatics. 2022;23:187.
3. Harvey WT, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nat Rev Microbiol. 2021;19:409–24.
4. Goswami P, et al. SARS-CoV-2 hot-spot mutations are significantly enriched within inverted repeats and CpG Island loci. Brief Bioinform. 2021;22:1338–45.
5. Mullick B, Magar R, Jhunjhunwala A, Farimani AB. Understanding mutation hotspots for the SARS-CoV-2 spike protein using Shannon entropy and K-means clustering. Comput Biol Med. 2021;138:104915.
6. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci. 2016;73:4433–48.
7. Lusvarghi S, et al. Key substitutions in the spike protein of SARS-CoV-2 variants can predict resistance to monoclonal antibodies, but other substitutions can modify the effects. J Virol. 2022;96:e01110–01121.
8. Martinez-Ledesma, Emmanuel D, Flores, Trevino V. Computational methods for detecting cancer hotspots. Comput Struct Biotechnol J. 2020;18:3567–76.
9. Ester M, Kriegel H, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;1996:226–31.

10. Shu Y, McCauley JGISAID. Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017;22:30494.
11. Jo H-Y, et al. Establishment of the large-scale longitudinal multi-omics dataset in COVID-19 patients: data profile and biospecimen. BMB Rep. 2022;55:465.
12. Katoh K, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
13. Varghese GM, John R, Manesh A, Karthik R, Abraham O. Clinical management of COVID-19. Indian J Med Res. 2020;151:401–10.
14. Adami C. Information theory in molecular biology. Phys Life Rev. 2004;1:3–22.
15. Pedregosa F et al. Scikit-learn: machine learning in Python. The Journal of machine Learning research. 2011;12:2825–2830.
16. St L, Wold S. Analysis of variance (ANOVA). Chemometr Intell Lab Syst. 1989;6(4):259–72.
17. Waskom ML. Seaborn: statistical data visualization. J Open Source Softw. 2021;660:3021.
18. Benjamini Y, Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300.
19. Cook HV, Doncheva NT, Szklarczyk D, Von Mering C, Jensen LJ, Viruses. STRING: a virus-host protein-protein interaction database. Viruses. 2018;10:519.
20. Oughtred R, et al. The biogrid interaction database: 2019 update. Nucleic Acids Res. 2019;47:D529–41.
21. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102:15545–50.
22. Zerbino DR et al. Ensembl 2018. Nucleic acids research. 2018;46.D.:D754-D761.
23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. bioinformatics. 2010;26:139–140.
24. Fisher RA. On the interpretation of χ 2 from contingency tables, and the calculation of P. J Roy Stat Soc. 1922;85(1):87–94.
25. Schoch CL et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database. 2020;2020:baaa062.
26. Cosar B, et al. SARS-CoV-2 mutations and their viral variants. Cytokine Growth Factor Rev. 2022;63:10–22.
27. Lin L, Liu Y, Tang X, He D. The disease severity and clinical outcomes of the SARS-CoV-2 variants of concern. Front Public Health. 2021;9:775224.
28. Wise J. Covid-19: the E484K mutation and the risks it poses. British Medical Journal Publishing Group; 2021.
29. Liu Z, et al. Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. Cell Host Microbe. 2021;29:477–88. e474.
30. Maison DP, Ching LL, Shikuma CM, Nerurkar VR. Genetic characteristics and phylogeny of 969-bp S gene sequence of SARS-CoV-2 from Hawai 'i reveals the worldwide emerging P681H mutation. Hawai'i J Health Social Welf. 2021;80:52.
31. Guo Y, et al. Targeting TNF-α for COVID-19: recent advanced and controversies. Front Public Health. 2022;10:833967.
32. Nersisyan S, Zhiyanov A, Shkurnikov M, Tonevitsky A. T-CoV: a comprehensive portal of HLA-peptide interactions affected by SARS-CoV-2 mutations. Nucleic Acids Res. 2022;50:D883–7.
33. Zhang Y-H, et al. Identification of the core regulators of the HLA I-peptide binding process. Sci Rep. 2017;7:42768.
34. Migliorini, et al. Association between HLA genotypes and COVID-19 susceptibility, severity and progression: a comprehensive review of the literature. Eur J Med Res. 2021;26:1–9.
35. Kuleshov MV, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:W90–7.
36. Akbarian S, et al. The correlation between humoral immune responses and severity of clinical symptoms in COVID-19 patients. Epidemiol Infect. 2023;151:e158.
37. Feng C, et al. Protective humoral and cellular immune responses to SARS-CoV-2 persist up to 1 year after recovery. Nat Commun. 2021;12:4984.
38. Mansoor S, et al. Expression of IFN-Gamma is significantly reduced during severity of covid-19 infection in hospitalized patients. PLoS ONE. 2023;18:e0291332.
39. Chen P-K, et al. The detectable anti-interferon-γ autoantibodies in COVID-19 patients may be associated with disease severity. Virol J. 2023;20:33.
40. Maucourant C, et al. Natural killer cell immunotypes related to COVID-19 disease severity. Sci Immunol. 2020;5:eabd6832.
41. Pegram HJ, Andrews DM, Smyth MJ, Darcy PK, Kershaw MH. Activating and inhibitory receptors of natural killer cells. Immunol Cell Biol. 2011;89:216–24.
42. Letafati A, et al. Unraveling the dynamic mechanisms of natural killer cells in viral infections: insights and implications. Virol J. 2024;21:18.
43. Saresella M, et al. NK cell subpopulations and receptor expression in recovering SARS-CoV-2 infection. Mol Neurobiol. 2021;58:6111–20.
44. Madrid AS, Ganem D. Kaposi's sarcoma-associated herpesvirus ORF54/dUTPase downregulates a ligand for the NK activating receptor NKp44. J Virol. 2012;86:8693–704.
45. Chalupny NJ, Rein-Weston A, Dosch S, Cosman D. Down-regulation of the NKG2D ligand MICA by the human cytomegalovirus glycoprotein UL142. Biochem Biophys Res Commun. 2006;346:175–81.
46. Banerjee P, Feuer G, Barker E. Human T-cell leukemia virus type 1 (HTLV-1) p12I down-modulates ICAM-1 and-2 and reduces adherence of natural killer cells, thereby protecting HTLV-1-infected primary CD4 + T cells from autologous natural killer cell-mediated cytotoxicity despite the reduction of major histocompatibility complex class I molecules on infected cells. J Virol. 2007;81:9707–17.
47. Di Vito C, et al. Natural killer cells in SARS-CoV-2 infection: pathophysiology and therapeutic implications. Front Immunol. 2022;13:888248.
48. Marquardt N, et al. Unique transcriptional and protein-expression signature in human lung tissue-resident NK cells. Nat Commun. 2019;10:3841.
49. Bouvier NM, Palese P. The biology of influenza viruses. Vaccine. 2008;26:D49–53.
50. Goka E, Vallely P, Mutton K, Klapper P. Mutations associated with severity of the pandemic influenza A (H1N1) pdm09 in humans: a systematic review and meta-analysis of epidemiological evidence. Arch Virol. 2014;159:3167–83.
51. Shao W, Li X, Goraya MU, Wang S, Chen J-L. Evolution of influenza a virus by mutation and re-assortment. Int J Mol Sci. 2017;18:1650.

52. Bae J-Y, et al. A single amino acid in the polymerase acidic protein determines the pathogenicity of influenza B viruses. J Virol. 2018;92:jvi10112800259–00218.

Publisher's note