Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2023**; 28(11): 284
https://doi.org/10.31083/j.fbl2811284

Original Research

Identification of Whole-Blood DNA Methylation Signatures and Rules Associated with COVID-19 Severity

Fei Yuan[1,†], JingXin Ren[2,†], HuiPing Liao[3,†], Wei Guo[4], Lei Chen[5], KaiYan Feng[6], Tao Huang[7,8,*], Yu-Dong Cai[2,*]

[1]Department of Science and Technology, Binzhou Medical University Hospital, 256603 Binzhou, Shandong, China
[2]School of Life Sciences, Shanghai University, 200444 Shanghai, China
[3]Changping Laboratory, 102206 Beijing, China
[4]Key Laboratory of Stem Cell Biology, Shanghai Jiao Tong University School of Medicine (SJTUSM) & Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), 200031 Shanghai, China
[5]College of Information Engineering, Shanghai Maritime University, 201306 Shanghai, China
[6]Department of Computer Science, Guangdong AIB Polytechnic College, 510507 Guangzhou, Guangdong, China
[7]Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 200031 Shanghai, China
[8]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 200031 Shanghai, China
*Correspondence: tohuangtao@126.com (Tao Huang); cai_yud@126.com (Yu-Dong Cai)
[†]These authors contributed equally.
Academic Editor: Rosa Alduina
Submitted: 24 April 2023 Revised: 29 June 2023 Accepted: 25 July 2023 Published: 8 November 2023

Abstract

Background: Different severities of coronavirus disease 2019 (COVID-19) cause different levels of respiratory symptoms and systemic inflammation. DNA methylation, a heritable epigenetic process, also shows differential changes in different severities of COVID-19. DNA methylation is involved in regulating the activity of various immune cells and influences immune pathways associated with viral infections. It may also be involved in regulating the expression of genes associated with the progression of COVID-19. **Methods**: In this study, a sophisticated machine-learning workflow was designed to analyze whole-blood DNA methylation data from COVID-19 patients with different severities versus healthy controls. We aimed to understand the role of DNA methylation in the development of COVID-19. The sample set contained 101 negative controls, 360 mildly infected individuals, and 113 severely infected individuals. Each sample involved 768,067 methylation sites. Three feature-ranking algorithms (least absolute shrinkage and selection operator (LASSO), light gradient-boosting machine (LightGBM), and Monte Carlo feature selection (MCFS)) were used to rank and filter out sites highly correlated with COVID-19. Based on the obtained ranking results, a high-performance classification model was constructed by combining the feature incremental approach with four classification algorithms (decision tree (DT), k-nearest neighbor (kNN), random forest (RF), and support vector machine (SVM)). **Results**: Some essential methylation sites and decision rules were obtained. **Conclusions**: The genes (*IGSF6*, *CD38*, and *TLR2*) of some essential methylation sites were confirmed to play important roles in the immune system.

Keywords: COVID-19 severity; DNA methylation; machine learning; rules

1. Introduction

Coronavirus disease 2019 (COVID-19) is currently the most serious public health problem and has caused millions of deaths worldwide. It can be classified into mild, moderate, and severe infection categories according to the clinical manifestations of the SARS-CoV-2 infection. These manifestations include asymptomatic, fever or chills, cough, loss of taste and/or smell, muscle or body aches, nausea or vomiting, and diarrhea. Studies have shown that severe COVID-19 is associated with host genetic variation related to host immune responses to viral infection and inflammasome modulators. COVID-19 severity is closely correlated with host factors [1]. Studies have revealed the interplay of genetic and epigenetic alterations that control host responses. Epigenetic changes that regulate chromatin structure have important implications for genome stability and the maintenance of cellular homeostasis since they are related to the pathophysiology of viral infections. DNA methylation is a heritable epigenetic process, whereby methyl groups are added to the C-5 position of the DNA cytosine loop by DNA methyltransferases. DNA methylation plays a key role in gene imprinting, X inactivation, silencing of repeat elements, and transposon expression. It also participates in cell development and aging [2]. DNA cytosine methylation at the 5′-C-phosphate G-3′ (CpG) site is highly sensitive to age and environmental factors [3–6]. COVID-19 severity is associated with impaired blood cell proportions and epigenetic modification of innate immune responses [7,8]. Changes in DNA methylation in neutrophils, B lymphocytes, and $CD8^+$ T lymphocytes reg-

ulate functional pathways related to autoimmune diseases and viral defense mechanisms [9]. Epigenetic changes associated with the respiratory environment can distinguish patients with severe and mild COVID-19 from those with systemic autoimmune diseases [10]. Specific hypermethylation in mild cases shows a genetic contribution, whereas methylation quantitative trait loci are enriched in SNPs associated with environmental traits [11]. DNA methylation may affect the expression of genes that regulate COVID-19 progression, which, in this context, is a targetable process.

An epigenome-wide association study of novel coronavirus infections revealed important DNA methylation regulation processes associated with COVID-19 progression [12]. Specific differences in CpG methylation were found between patients with severe and mild diseases. The differential methylation is primarily related to the activation of the interferon signaling pathway and the overactivation of B and T lymphocytes [13–15]. Transcriptomic studies have confirmed that these pathways are associated with COVID-19 severity, while it was shown that regulation of these pathways is mediated by epigenetic changes at the promoter level of the relevant genes [16]. Moreover, epigenetic dysregulation exists in the *CD209* signaling pathway, phagocytosis pathway, and AKT signaling pathway in COVID-19 patients with specific blood cell types. *CD209* is primarily expressed in B lymphocytes and dendritic cells (DCs), and it interacts with *CD209L* in endothelial cells of SARS-CoV-2 target tissues, which may contribute to virus invasion [17]. Therefore, hypermethylation of the *CD209* signaling pathway may be related to the protective effect during SARS-CoV-2 infection. Interestingly, *EDC3* hypermethylation in severe cases may mediate the overexpression of the angiotensin-converting enzyme (*ACE*) *2* protein in COVID-19 patients, thereby worsening the infection [18]. The hypomethylation signatures associated with COVID-19 severity include interferon-related signatures and lymphocyte activation signatures. Interferon-related features are associated with systemic autoimmune disease in mild and severe cases [19]. Further analysis revealed that the enrichment of transcription-factor binding sites, which regulate the levels of cytokines (e.g., interleukin (IL)-6, IL-1α, and IL-12) and other proinflammatory cytokines, is associated with COVID-19 severity [20–24].

High-throughput sequencing data provides extensive molecular information relevant to patients with COVID-19. Our team has previously used machine learning to screen 49 key methylation sites associated with COVID-19. The degree of methylation at these sites was correlated to the age of the patient [25]. Accordingly, we aimed to further explore the COVID-19 mechanism based on whole-blood genome-wide DNA methylation profiling from 101 negative controls, 360 patients with mild infections, and 113 individuals with severe infections.

2. Materials and Methods

Fig. 1 illustrates the flow of the machine-learning method used in this study. The samples were grouped according to COVID-19 severity. The methylation sites were ranked using three feature-ranking methods. Then, the obtained feature-ranking list was fed into the incremental feature selection (IFS) framework, which contained four classification algorithms. In summary, we obtained key methylation sites that were strongly correlated with COVID-19 severity and a model that could predict the COVID-19 status of the samples. Quantitative classification rules were also summarized. This section describes the methods used in each segment.

2.1 Data

The whole-blood DNA methylation data for the 574 used samples were obtained from the GEO database, with the accession number: 179325 [26]. The samples were divided into three groups according to the COVID-19 severity and included 101 negative controls, 360 mildly infected individuals, and 113 severely infected individuals. The DNA methylation sites in the samples were considered as features, and each sample contained 768,067 methylation sites.

2.2 Feature-Ranking Algorithms

The number of DNA methylation sites involved in the dataset was large, although only a very small number of sites were significantly methylated during COVID-19 development. All sites were screened and ranked using the least absolute shrinkage and selection operator (LASSO) [27], light gradient-boosting machine (LightGBM) [28], and Monte Carlo feature selection (MCFS) [29]. A higher site was ranked corresponding to its higher degree of association with the target variable. These methods are extensively accepted in the life sciences [30–33].

2.2.1 Least Absolute Shrinkage and Selection Operator

LASSO is a regression analysis method that ranks features and performs preliminary feature screening. The method is internally designed with a penalty function that is bounded by an L1-type regular formula. The methylation-site features are assigned to independent variables, and the absolute values of the coefficients of the independent variables describe the degree of association with the target variable. By optimizing the function, the coefficients of some features become zero, indicating that they should be treated as irrelevant features. Afterward, the remaining features are ranked according to the absolute value of the coefficients. Finally, the dimensionality reduction and ranking of the data are completed.

2.2.2 LightGBM

LightGBM is based on gradient-boosting decision trees (DTs) for optimization, which consumes less memory and is suitable for high-dimensional datasets. It uses the

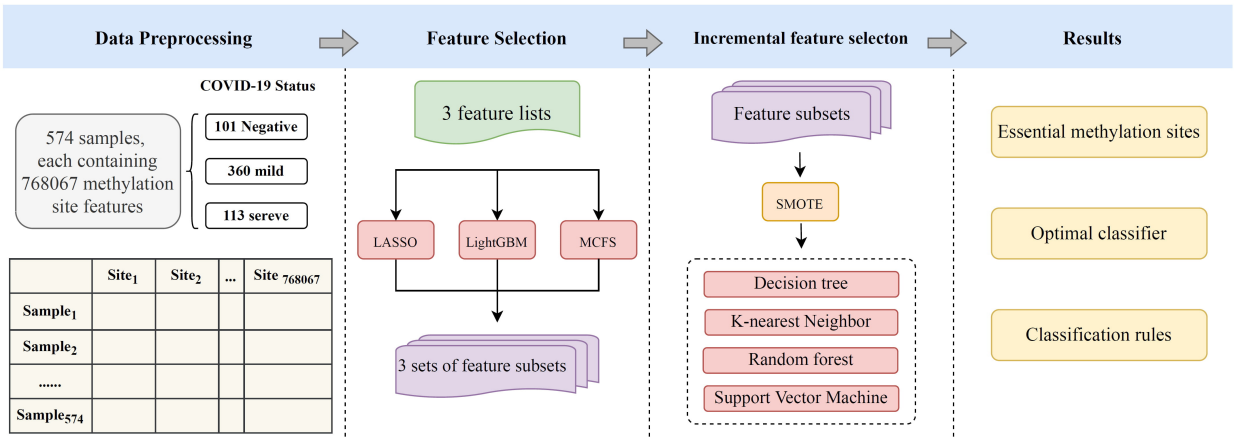

Fig. 1. Flow chart of the entire analytical process. Whole-blood DNA methylation data from patients with different COVID-19 severities were analyzed by machine-learning approach. The dataset contained 101 negative controls, 360 mildly infected individuals, and 113 severely infected individuals. The methylation-site features were analyzed by three feature-selection methods, namely, least absolute shrinkage and selection operator (LASSO), light gradient-boosting machine (LightGBM), and Monte Carlo feature selection (MCFS). The obtained feature lists were fed into the incremental feature-selection method, which combined decision tree (DT), k-nearest neighbor (kNN), random forest (RF), and support vector machine (SVM) to extract key site features and construct effective classifiers and classification rules.

gradient-based one-side sampling method to discard some samples with small gradients, and it bundles selected samples using an exclusive feature-bundling method to merge mutually exclusive features. It uses a leaf-wise strategy to construct the tree, extending only the more efficient branches. The importance of a feature is proportional to its involvement in the DT.

2.2.3 Monte Carlo Feature Selection

The MCFS method constructs a number of independent DTs. Each DT uses a different subset of features and sample data sets. The features and samples used by each tree are determined by independent random selection. For the selected sample subsets, a division was performed t times into a training subset and a testing subset. By combining the determined p feature subsets, we obtained $p \times t$ DTs. The importance of the sites was expressed using the relative importance score (RI).

$$RI_g = \sum_{\tau=1}^{p\times t} (\omega A_{CC})^u \sum_{ng(\tau)} IG(ng(\tau)) \left(\frac{no.in\ ng\ (\tau)}{no.in\ \tau} \right)^v, \quad (1)$$

where ωA_{CC} is the weighted precision of the tree, τ, under consideration; $ng\ (\tau)$ is the node of the DT whose information gain is denoted as $IG(ng\ (\tau))$; $no.in\ ng\ (\tau)$ denotes the sample size of $ng\ (\tau)$; u and v are two positive reals weighting the ωA_{CC} and the ratio $no.in\ ng\ (\tau)\ /no.in\ \tau$, respectively.

2.3 Incremental Feature Selection

The use of IFS is well-established in machine learning research [33,34]. When building a classification model, computational efficiency needs to be considered. For this purpose, we performed the IFS method to determine the optimum number of features required to build the classification model. It converted the list of methylation sites generated by the ranking method into a number of feature subsets. Depending on the set step size, the number of features contained in these subsets grew equally according to the ranking order. These subsets were used to train the downstream classification algorithm, where in reference to the performance of the obtained models, the finalized feature subset was the best feature subset, and the model at this point was the best classification model.

2.4 Synthetic Minority Oversampling Technique

Looking at the sample dataset, a difference existed in the number of samples within the three classes, whereby there were 3.6 times more samples from the mildly infected individuals than from the negative controls. Moreover, directly training the model using an uneven dataset biased the results toward the majority classes. The SMOTE method generated new samples based on known samples. All samples were projected into the high-dimensional space, and one sample was randomly selected for the minority classes. Using Euclidean distance as a metric, SMOTE determined the k-nearest neighbors to that sample in the same class. Moreover, any point on the concatenation of this sample and any of its nearest neighbors can be selected as a new sample. The above process was repeated and the new samples were added into minority classes until the number of samples in each class was balanced.

2.5 Classification Algorithm

Based on several feature subsets generated by IFS, this study used DT [35], kNN [36], RF [37], and SVM [38] algorithms to construct classification models. These algorithms have been approved by previous publications [34,39–45].

2.5.1 Decision Tree

As its name suggests, the DT algorithm constructs a tree structure that contains a root, branch, and leaf structure. The instances were inputted from the root, and the attributes of the instances were judged in the internal node. Based on the result, they were transported along the branches of the tree after several judgments until they reached the leaf structure. The leaves of the tree contained the final judgment of the algorithm on the class of the instance. As a white-box algorithm, the judgment process was transparent, thereby enabling the highly interpretable classification rules to be summarized.

2.5.2 K-Nearest Neighbor

The principle of the kNN algorithm was to determine the category of an unknown sample by comparing the distribution of known samples around a new sample in the feature space. The samples were mapped into a high-dimensional space based on the feature vectors; the k-nearest known samples for each new sample were selected, and the category labels for these nearest neighbors were referred to finally determine the category of the new sample.

2.5.3 Random Forest

RF was based on the DT algorithm. A number of independent DT models were constructed at once, and randomness was introduced into the selection of features and training samples used for judgment. Each DT constructed in this way had a different judgment process for the same sample, meaning the results often differed. For a sample, the classification results of each tree were combined, and the final classification results were obtained using the principle of majority rule.

2.5.4 Support Vector Machine

The SVM algorithm utilized a kernel function that also mapped the samples into a high-dimensional space based on feature vectors. Through an optimization function, a hyperplane was determined in the space. This hyperplane partitioned the samples of different categories in space and ensured that the margin between each category of samples was maximized at this hyperplane. The model presented the best generalization ability at this time.

2.6 Performance Evaluation

Using the IFS method and above four classification algorithms, we obtained several classification models and evaluated the performance of these models using 10-fold cross-validation [46]. The F1 measure is often used as a key metric to evaluate the performance of classification models, although it does not perform fairly enough in multiclassification problems. The current study required weighting the F1 measure.

$$Precision_i = \frac{TP_i}{TP_i + FP_i}, \tag{2}$$

$$Precision_{weighted} = \sum_{i=1}^{L} Precision_i \times w_i, \tag{3}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}, \tag{4}$$

$$Recall_{weighted} = \sum_{i=1}^{L} Recall_i \times w_i, \tag{5}$$

$$Weight\ F1 = \frac{2 \cdot Precision_{weighted} \cdot Recall_{weighted}}{Precision_{weighted} + Recall_{weighted}}, \tag{6}$$

where TP represents true positives, FP represents false positives, FN represents false negatives, i represents the category, L represents the number of classes, and w_i represents the proportion of samples categorized to the overall sample. We also used accuracy (ACC) and Matthew's correlation coefficient (MCC) [47] as references. A higher value corresponded to better model performance.

3. Results

3.1 DNA Methylation Sites Ranking and IFS

Whole-blood DNA methylation data from 574 samples, each containing 768,067 methylation-site features, were analyzed using a machine-learning workflow. LASSO, LightGBM, and MCFS were used to rank all methylation sites and three ranked lists were obtained (**Supplementary Table 1**). LightGBM discarded the redundant features, meaning only 33,537 features were output in LightGBM for ranking. The lists involved numerous site features, although only a very small number were associated with COVID-19 levels. Thus, for subsequent analyses, only the top 10,000 site features in the list were used.

The first 10,000 methylation sites in the three lists were input into the IFS framework, and the step parameter was set to 10. Each list was transformed into a subset of 1000 sites, and these subsets of methylated sites were used to construct 4 classification models, namely, DT, kNN, RF, and SVM. **Supplementary Table 2** describes the classification performance of these models. Based on the performance results, IFS curves were plotted using the number of features as the independent variable to depict the weighted F1 trend (Fig. 2). The IFS curve observation revealed that

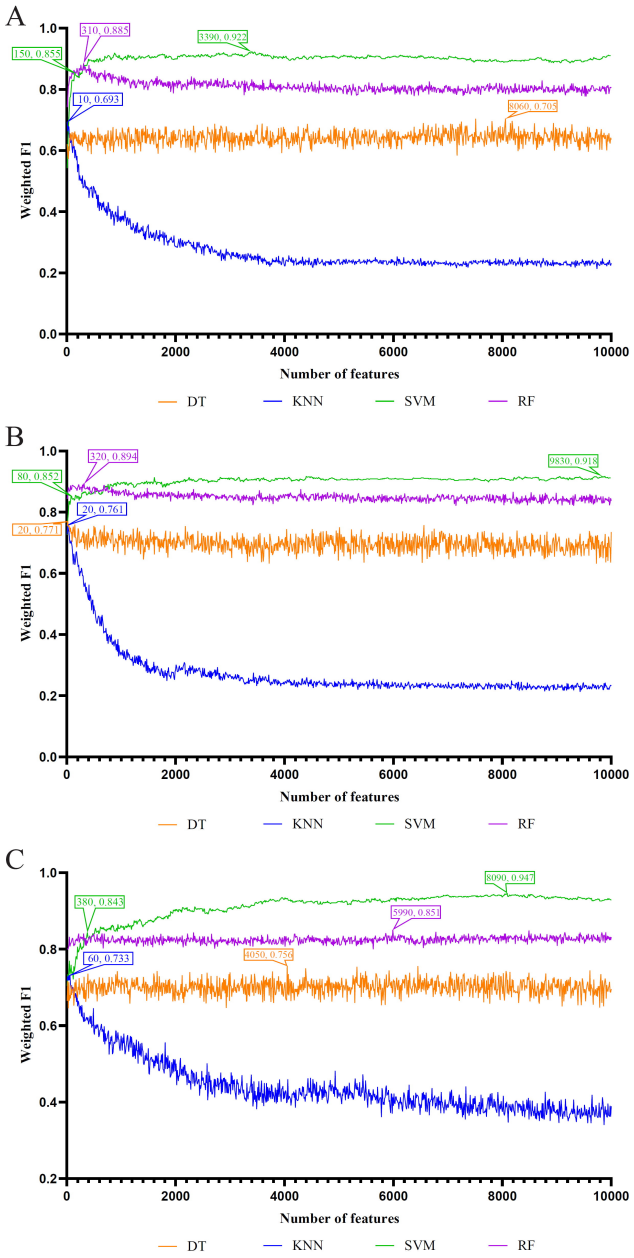

Fig. 2. IFS curves for evaluating the performance of the three classification algorithms based on the weighted F1. Four classifier models were constructed under each algorithm. (A) IFS curves based on LASSO results. (B) IFS curves based on LightGBM results. (C) IFS curves based on MCFS results. IFS, incremental feature selection.

the SVM model always had the highest performance regardless of the list being used. The best SVM model used the top 3390, 9830, and 8090 methylation sites from the Lasso, LightGBM, and MCFS lists, respectively, to provide weighted F1 values of 0.922, 0.918, and 0.947, respectively. Their ACC and MCC results were also excellent, with ACC of 0.922, 0.916, and 0.946, respectively, and MCC of 0.862, 0.857, and 0.904, respectively (Table 1).

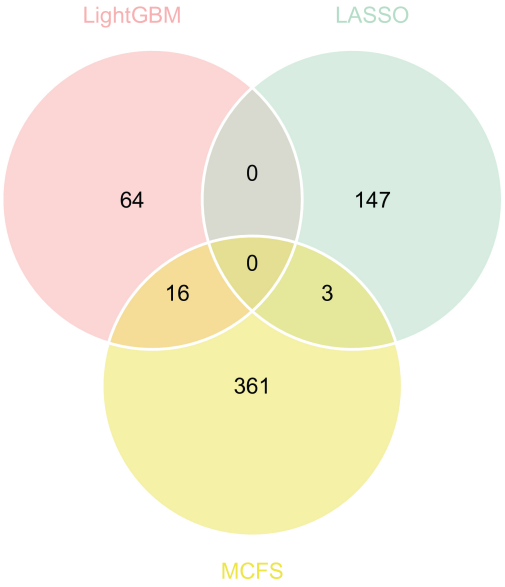

Fig. 3. Venn diagram of the most critical subset of features obtained using LASSO, LightGBM, and MCFS. The overlapping circles indicate methylation sites that are identified as the most critical features by the different ranking algorithms.

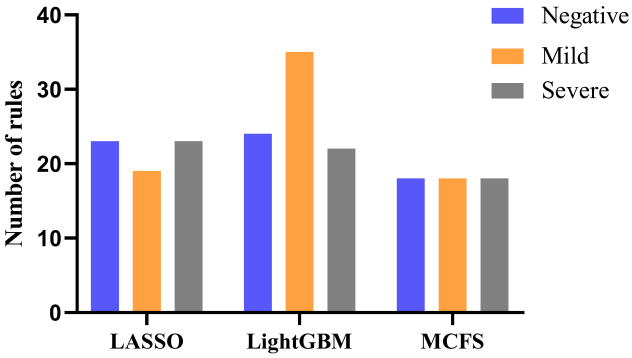

Fig. 4. The number of rules for the three classes as summarized by the DT classification model, constructed using the three ranking methods.

To extract the most critical sites among the methylation sites used by the best classifier, we observed the IFS curves of the SVM models and set the inflection points of the curves. The inflection points were 150, 80, and 380 for the LASSO, LightGBM, and MCFS lists, respectively. The performance evaluation of the model at the inflection points is displayed in Table 1. Recurring methylation sites tended to play a more important role. We calculated the intersection from the pre-inflection point sites and plotted a Wayne diagram (Fig. 3). We did not obtain methylation sites that appeared in all three lists but obtained 19 features that appeared in two lists. The specific intersection results are shown in **Supplementary Table 3**. A detailed analysis of these sites is presented later.

Table 1. Performance of SVM classification models constructed from three lists when using the best feature subset and inflection subset.

Feature ranking algorithms	Number of features	Weighted F1	MCC	ACC
LASSO	150*	0.855	0.761	0.852
	3390**	0.922	0.862	0.922
LightGBM	80*	0.852	0.753	0.848
	9830**	0.918	0.857	0.916
MCFS	380*	0.843	0.755	0.840
	8090**	0.947	0.904	0.946

* indicates the number of features at the inflection points of the IFS curve; ** indicates the number of features in the best feature subset. MCC, Matthew's correlation coefficient; ACC, accuracy.

Table 2. Essential methylation sites identified by multiple algorithms.

Methylation locus	Gene symbol	Description	Algorithm	Reference
cg02481950	*IGSF6*	chr16:21650858-21663981	LASSO, MCFS	[48–53]
cg04332373	*CD38*	chr4:15778275-15853232, CpG island	LightGBM, MCFS	[54–61]
cg03753191	*EPSTI1*	chr13:43566902, Shore region of CpG islands	LightGBM, MCFS	[62]
cg22930808	*DTX3L*	chr4:15778275-15853232, CpG island	LightGBM, MCFS	[63–67]

3.2 Classification Rules

The DT model demonstrated a transparent decision process, which helped us obtain specific classification cues. Based on these cues, we summarized three sets of quantitative classification rules, which had implications for our understanding of the biological context of these key methylation sites. **Supplementary Table 4** shows the specific classification rules. Each rule contained several parameters that represented the methylation levels of these sites. For the different lists, Fig. 4 shows the number of rules representing the different categories.

4. Discussion

This study had three groups: negative controls, mildly infected individuals, and severely infected individuals. Each algorithm presents its advantages, and taking the intersection of the results of multiple algorithms can increase the coverage and accuracy of the results better than one approach. Through the intersection of all analysis approaches, we identified some important methylation features. Combined with previous studies, summarizing the experimental evidence of the above-mentioned methylation features was necessary.

4.1 Analysis of Top Features Identified by Multiple Algorithms

As shown in Fig. 3, using 2 algorithms identified 19 methylation sites, which were generally more essential than other methylation sites. Here, we selected four for detailed analysis and these are listed in Table 2 (Ref. [48–67]).

IGSF6 (cg02481950) is located at chr16:21650858-21663981 and belongs to the immunoglobulin superfamily member 6, which is expressed in dendritic and myeloid cells [48]. Studies have shown that *IGSF6* is associated with changes in the ratio of M0 macrophages and γ-delta T cells, alongside plasma cells and monocytes in atherosclerotic plaques. *IGSF6* may be related to immune-related IFN-γ and PD-1 signaling pathways [49] and is reportedly a susceptibility gene for inflammatory bowel disease; therefore, may influence disease manifestations [50]. *IGSF6* is potentially involved in the atrial fibrillation development mechanism and is associated with the course or maintenance of autoimmune and chronic inflammatory diseases [51]. Moreover, it can act as a potential molecular marker for antigen presentation by DCs before host vaccinations have been identified. After stimulation with human papillomavirus E7 peptide (p11–20), the immune response in immature DCs (iDCs) is upregulated by *IGSF6* and other molecules. With prolonged stimulation time, the genes related to the immune response become more significantly upregulated [52]. DCs play important roles in preventing viral infection. Thus, the proportion of plasmacytoid and myeloid DC levels in serum and the IFN level decrease in patients with severe COVID-19. The impairment of DC function and interferon secretion correlates to the COVID-19 severity [53]. Our analysis showed that the methylation level of *IGSF6* increased in severe COVID-19 patients. These results suggest that the upregulation of these genes may be a marker of antigen presentation, and *IGSF6* may be a potential molecular marker of the extent of the immune response in patients with COVID-19.

CD38 (cg04332373) encodes a type II transmembrane glycoprotein that synthesizes and hydrolyzes calcium-ion-mobilized messengers. The cg04332373 methylation site is located in the shore region of the CpG island. Studies have shown that CD38 levels in white adipose tissue (WAT) and the liver increase with age. Senescent cell signals pro-

Table 3. Important methylation sites in rules.

Methylation locus	Gene symbol	Description	References
cg00197681	*TBC1D4*	chr13:76056419, CpG island	[68]
cg09623286	*TLR2*	chr4:154605468, CpG island	[69–71]

mote the accumulation of $CD38^+$ cells in WAT [54]. CD38 regulates nicotinamide dinucleotide (NAD^+) metabolism and extracellular nucleotide homeostasis, while CD38 inhibition and "NAD^+ enhancement" can help in metabolic disorders related to aging, inflammation, and tumor immune hyperplasia [55]. Increased *CD38* expression is a consequence of aging [56], which is a major factor associated with the risk of SARS-CoV-2 infection. CD38 plays an important role in viral infections, including AIDS and COVID-19 [57]. Thus, in COVID-19 patients, CD38 mediates thrombosis and bacterial phagocytosis dependent on NAD^+. CD38 promotes immune-cell migration into the site of infection through signaling [58,59], meaning *CD38* may aggravate SARS-CoV-2 infection and increase the risk of secondary bacterial infections [60,61]. Activation of CD38 and decreased NAD^+ can be considered features of aging and may be considered regulators of COVID-19 in old age. High inflammation in COVID-19 may lead to CD38 activation, especially causing severe reactions, such as tissue fibrosis and injury in the elderly [57]. Our analysis showed that *CD38* methylation levels were lower in patients with severe COVID-19 than in patients with moderate COVID-19. This evidence suggests that SARS-CoV-2 infection may reduce CD38 methylation levels, promote CD38 activation, promote immune-cell migration, and induce NAD^+-dependent bacterial phagocytosis, ultimately, leading to local tissue damage and disease risk.

cg03753191 is located at chr13:43566902 and in the Shore region of CpG islands. It corresponds to the *EPSTI1* gene and encodes a protein that promotes tumor invasion and metastasis. Studies have reported that *EPSTI1* expression levels were significantly upregulated in leukocytes and nasopharyngeal tissues of COVID-19 patients, compared to normal tissues. The downregulation of *EPSTI1* also predicted poorer clinical outcomes in patients with COVID-19, such as intensive care unit hospitalization and increased viral load. Therefore, *EPSTI1* may play an important role in antiviral immune regulation [62]. The expression profile of *EPSTI1* can classify COVID-19 patients into different groups, whereby the younger patient population exhibited a stronger antiviral immune response, higher *EPSTI1* expression level, and better clinical treatment effect. The expression level or methylation pattern of *EPSTI1* may significantly affect the COVID-19 severity by modulating antiviral immune responses.

DTX3L (cg22930808) encodes Deltex E3 ubiquitin ligase 3L, which plays a role in DNA damage repair and interferon-mediated antiviral response [63–66]. *DTX3L* plays a role in the antiviral response by mediating the "LS-48"-linked ubiquitination in the C3 protease of the cerebral myocarditis virus and human rhinovirus, thereby promoting their proteome-mediated degradation [65]. *Dtx3l*-related pathways include DNA damage and the innate immune system. The activation of the IFN response has been demonstrated to induce the adp nucleation of host proteins, which depend on PARP9 and its binding partner DTX3L. Expression of the large domain of SARS-CoV-2 nonstructural protein 3 (Nsp3) or deletion of *PARP9* or *DTX3L* had no effect on IFN signaling. *PARP9*/*DTX3L*-dependent adp nucleation is a downstream effector of the host IFN response, and the SARS-CoV-2 Nsp3 macrodomain can hydrolyze the IFN signaling end product [67]. This study has added evidence to the relationship between *DTX3L* methylation patterns and disease severity in patients with SARS-CoV-2 infection, thereby providing more reference for the study of COVID-19 progression and pathogenesis.

4.2 Analysis of Key Methylation Sites in Rules

In addition to identifying essential methylation sites by using various algorithms, we also obtained several classification rules, as listed in **Supplementary Table 4**. These rules can predict disease severity based on the methylation level of key DNA methylation sites. However, it is impossible to analyze them in detail because a huge number of rules were constructed. Herein, we focused on valuable features because they can identify important DNA methylation sites and demonstrate their important role as epigenetic-susceptibility sites in patients infected with SARS-CoV-2. We collected the scientific findings of other researchers and preliminarily summarized experimental evidence to conduct the discussion. The key methylation sites are listed in Table 3 (Ref. [68–71]).

TBC1D4 (cg00197681) belongs to TBC1 domain family member 4, and cg00197681 is located at chr13:76056419 and CpG island region. TBC1D4 encodes a rabbit-GTPase activating protein, which is involved in regulating glucose homeostasis through transporting glucose transporter 4 (GLUT4). Human genetic variants may influence SARS-CoV-2 infection and COVID-19 pathology. Researchers have applied machine-learning algorithms to predict phosphorylation-associated missense single nucleotide variants (pSNVs) and found that these variants are associated with the evolution of host defense systems [68]. Proteins with pSNVs include TBC1D4 and TRIM28, which are involved in the regulation of the viral life cycle and host antiviral response. TBC1D4 may also be associated with the dysregulation of glucose homeostasis in SARS-CoV-2 infection. In the present study, we found

that *TBC1D4* methylation levels were higher in patients with severe COVID-19 than in the normal population. Combined with previous studies, we hypothesized that SARS-CoV-2 infection induced an increase in the methylation levels of TBC1D4, thereby affecting the host antiviral immunity.

The methylation site cg09623286 is located at chr4:154605468 and the island region of CpG islands. Its related gene is *Toll-like receptor* (*TLR*) *2*, which encodes products belonging to the TLR family. As a cell surface protein, TLR2 can form dimers with other TLRs to recognize pathogen-associated molecular patterns and activate host immune responses. Activation of different TLR pathways leads to the secretion of proinflammatory factors, including IL, tumor necrosis factor-alpha, and interferon, and is involved in SARS-CoV-2 invasion and infection. However, some TLRs, such as TLR2, may play a dual role in COVID-19 infection [69]. SARS-CoV-2 infection can cause severe disease features (e.g., cytokine storm) and organ failure (e.g., testicular and germ-cell damage). Studies have shown that exposure to SARS-CoV-2 envelope proteins causes TLR2 receptor-dependent testicular cytopaplasia [70]. *TLR2* mRNA expression levels also significantly increase in patients with moderate to severe COVID-19. Moreover, the mRNA expression levels of CK-MB, ACE2, and neuroproteinase 1 receptors were positively correlated with TLR2 expression in all patients. The mRNA expression of *TLR2* was positively correlated with renal biomarkers and cardiac enzymes in severe and moderate patients. These results suggested that it may be related to COVID-19 severity [71]. The present study showed that patients with moderate COVID-19 had higher methylation levels than normal controls; however, this was not observed in severe patients. Therefore, the TLR2 methylation level may serve as an indicator of COVID-19 severity, although the specific mechanism needs further verification.

5. Conclusions

This study aimed to determine the characteristics and patterns of methylation sites associated with COVID-19 severity through computational analysis. The results showed that the identified key features were validated in published academic studies. Important features may be involved in viral immune regulation through different methylation patterns or expression levels. Additionally, antiviral immunity was associated with age, which is consistent with our previous findings. The accuracy and credibility of the methylation characteristics of important groups were improved by mutual verification of various methods. On one hand, it laid the foundation for more accurate analysis methods in the future, while, alternatively, it provided a broad reference for research on the mechanism of serious diseases, such as COVID-19.

Abbreviations

COVID-19, Coronavirus disease 2019; IFS, incremental feature selection; LASSO, least absolute shrinkage and selection operator; LightGBM, light gradient-boosting machine; MCFS, Monte Carlo feature selection; SMOTE, Synthetic Minority Oversampling Technique; DT, decision tree; kNN, k-nearest neighbor; RF, random forest; SVM, support vector machine; MCC, Matthew correlation coefficient.

Availability of Data and Materials

The datasets analyzed during the current study are available in the [Gene Expression Omnibus] repository [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179325].

Author Contributions

TH and YDC designed the research study. JXR, LC and KYF performed the research. FY, HPL, WG and FY analyzed the data. FY, JXR and HPL wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research was funded by the National Key R&D Program of China [2022YFF1203202], Strate-gic Priority Research Program of Chinese Academy of Sciences [XDA26040304, XDB38050200], the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences [202002], Shandong Provincial Natural Science Foundation [ZR2022MC072].

Conflict of Interest

The authors declare no conflict of interest. YDC is serving as Editorial Board member of this journal and TH has served as the Guest Editor of the journal. We declare that YDC and TH had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Rosa Alduina.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2811284.

References

[1] COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature. 2021; 600: 472–477.

[2] Trotman JB, Calabrese JM. How to silence an X chromosome. Nature. 2020; 578: 365–366.

[3] Kane AE, Sinclair DA. Epigenetic changes during aging and their reprogramming potential. Critical Reviews in Biochemistry and Molecular Biology. 2019; 54: 61–83.

[4] Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT, *et al*. DNA methylation aging clocks: challenges and recommendations. Genome Biology. 2019; 20: 249.

[5] Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. Cancer Research. 2016; 76: 3446–3450.

[6] McCartney DL, Min JL, Richmond RC, Lu AT, Sobczyk MK, Davies G, *et al*. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. Genome Biology. 2021; 22: 194.

[7] Bernardes JP, Mishra N, Tran F, Bahmer T, Best L, Blase JI, *et al*. Longitudinal Multi-omics Analyses Identify Responses of Megakaryocytes, Erythroid Cells, and Plasmablasts as Hallmarks of Severe COVID-19. Immunity. 2020; 53: 1296–1314.e9.

[8] Schultze JL, Aschenbrenner AC. COVID-19 and the human innate immune system. Cell. 2021; 184: 1671–1692.

[9] Wang G, Xiong Z, Yang F, Zheng X, Zong W, Li R, *et al*. Identification of COVID-19-Associated DNA Methylation Variations by Integrating Methylation Array and scRNA-Seq Data at Cell-Type Resolution. Genes. 2022; 13: 1109.

[10] Liu Y, Sawalha AH, Lu Q. COVID-19 and autoimmune diseases. Current Opinion in Rheumatology. 2021; 33: 155–162.

[11] Koh IU, Lee HJ, Hwang JY, Choi NH, Lee S. Obesity-related CpG Methylation (cg07814318) of Kruppel-like Factor-13 (KLF13) Gene with Childhood Obesity and its cis-Methylation Quantitative Loci. Scientific Reports. 2017; 7: 45368.

[12] Cao X, Li W, Wang T, Ran D, Davalos V, Planas-Serra L, *et al*. Accelerated biological aging in COVID-19 patients. Nature Communications. 2022; 13: 2135.

[13] Giamarellos-Bourboulis EJ, Netea MG, Rovina N, Akinosoglou K, Antoniadou A, Antonakos N, *et al*. Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure. Cell Host & Microbe. 2020; 27: 992–1000.e3.

[14] Mathew D, Giles JR, Baxter AE, Oldridge DA, Greenplate AR, Wu JE, *et al*. Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science. 2020; 369: eabc8511.

[15] Diao B, Wang C, Tan Y, Chen X, Liu Y, Ning L, *et al*. Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19). Frontiers in Immunology. 2020; 11: 827.

[16] Sawalha AH, Zhao M, Coit P, Lu Q. Epigenetic dysregulation of ACE2 and interferon-regulated genes might suggest increased COVID-19 susceptibility and severity in lupus patients. Clinical Immunology. 2020; 215: 108410.

[17] Amraei R, Yin W, Napoleon MA, Suder EL, Berrigan J, Zhao Q, *et al*. CD209L/L-SIGN and CD209/DC-SIGN Act as Receptors for SARS-CoV-2. ACS Central Science. 2021; 7: 1156–1165.

[18] Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, *et al*. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell. 2020; 181: 271–280.e8.

[19] Liu F, Wang B, Liu Y, Shi W, Tang X, Wang X, *et al*. Novel TYK2 Inhibitors with an *N*-(Methyl-d_3)pyridazine-3-carboxamide Skeleton for the Treatment of Autoimmune Diseases. ACS Medicinal Chemistry Letters. 2022; 13: 1730–1738.

[20] Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, *et al*. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science. 2020; 370: eabd4570.

[21] Lucas C, Wong P, Klein J, Castro TBR, Silva J, Sundaram M, *et al*. Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature. 2020; 584: 463–469.

[22] Del Valle DM, Kim-Schulze S, Huang HH, Beckmann ND, Nirenberg S, Wang B, *et al*. An inflammatory cytokine signature predicts COVID-19 severity and survival. Nature Medicine. 2020; 26: 1636–1643.

[23] Ha SD, Cho W, DeKoter RP, Kim SO. The transcription factor PU.1 mediates enhancer-promoter looping that is required for IL-1β eRNA and mRNA transcription in mouse melanoma and macrophage cell lines. The Journal of Biological Chemistry. 2019; 294: 17487–17500.

[24] Wen AY, Sakamoto KM, Miller LS. The role of the transcription factor CREB in immune function. Journal of Immunology. 2010; 185: 6413–6419.

[25] Chen L, Liao H, Huang G, Ding S, Guo W, Huang T, *et al*. Identification of DNA Methylation Signature and Rules for SARS-CoV-2 Associated with Age. Frontiers in Bioscience (Landmark Edition). 2022; 27: 204.

[26] Barturen G, Carnero-Montoro E, Martínez-Bueno M, Rojo-Rello S, Sobrino B, Porras-Perales Ó, *et al*. Whole blood DNA methylation analysis reveals respiratory environmental traits involved in COVID-19 severity following SARS-CoV-2 infection. Nature Communications. 2022; 13: 4597.

[27] Ranstam J, Cook J. LASSO regression. Journal of British Surgery. 2018; 105: 1348–1348.

[28] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, *et al*. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems. 2017; 30: 3146–3154.

[29] Dramiński M, Koronacki J. rmcfs: An R Package for Monte Carlo Feature Selection and Interdependency Discovery. Journal of Statistical Software. 2018; 85: 1–28.

[30] Li H, Huang F, Liao H, Li Z, Feng K, Huang T, *et al*. Identification of COVID-19-Specific Immune Markers Using a Machine Learning Method. Frontiers in Molecular Biosciences. 2022; 9: 952626.

[31] Li Z, Mei Z, Ding S, Chen L, Li H, Feng K, *et al*. Identifying Methylation Signatures and Rules for COVID-19 With Machine Learning Methods. Frontiers in Molecular Biosciences. 2022; 9: 908080.

[32] Lu J, Li J, Ren J, Ding S, Zeng Z, Huang T, *et al*. Functional and embedding feature analysis for pan-cancer classification. Frontiers in Oncology. 2022; 12: 979336.

[33] Li H, Zhang S, Chen L, Pan X, Li Z, Huang T, *et al*. Identifying Functions of Proteins in Mice With Functional Embedding Features. Frontiers in Genetics. 2022; 13: 909040.

[34] Huang F, Fu M, Li J, Chen L, Feng K, Huang T, *et al*. Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores. Biochimica et Biophysica Acta. Proteins and Proteomics. 2023; 1871: 140889.

[35] Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics. 1991; 21: 660–674.

[36] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967; 13: 21–27.

[37] Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, *et al*. SNP-based analysis of genetic substructure in the German population. Human Heredity. 2006; 62: 20–29.

[38] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20: 273–297.

[39] Huang F, Chen L, Guo W, Zhou X, Feng K, Huang T, *et al*. Identifying COVID-19 Severity-Related SARS-CoV-2 Mutation Us-

ing a Machine Learning Method. Life. 2022; 12: 806.
[40] Li Z, Guo W, Ding S, Chen L, Feng K, Huang T, *et al*. Identifying Key MicroRNA Signatures for Neurodegenerative Diseases With Machine Learning Methods. Frontiers in Genetics. 2022; 13: 880997.
[41] Li Z, Guo W, Ding S, Feng K, Lu L, Huang T, *et al*. Detecting Blood Methylation Signatures in Response to Childhood Cancer Radiotherapy via Machine Learning Methods. Biology. 2022; 11: 607.
[42] Huang F, Ma Q, Ren J, Li J, Wang F, Huang T, *et al*. Identification of Smoking-Associated Transcriptome Aberration in Blood with Machine Learning Methods. BioMed Research International. 2023; 2023: 5333361.
[43] Ren J, Zhang Y, Guo W, Feng K, Yuan Y, Huang T, *et al*. Identification of Genes Associated with the Impairment of Olfactory and Gustatory Functions in COVID-19 via Machine-Learning Methods. Life. 2023; 13: 798.
[44] Wu C, Chen L. A model with deep analysis on a large drug network for drug classification. Mathematical Biosciences and Engineering. 2023; 20: 383–401.
[45] Wang H, Chen L. PMPTCE-HNEA: Predicting metabolic pathway types of chemicals and enzymes with a heterogeneous network embedding algorithm. Current Bioinformatics. 2023. (online ahead of print)
[46] Kohavi R (ed.). A study of cross-validation and bootstrap for accuracy estimation and model selection. International joint Conference on artificial intelligence. Lawrence Erlbaum Associates Ltd.: Mahwah, New Jersey, USA. 1995.
[47] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta. 1975; 405: 442–451.
[48] Bates EE, Kissenpfennig A, Péronne C, Mattei MG, Fossiez F, Malissen B, *et al*. The mouse and human IGSF6 (DORA) genes map to the inflammatory bowel disease 1 locus and are embedded in an intron of a gene of unknown function. Immunogenetics. 2000; 52: 112–120.
[49] Shen Y, Xu LR, Tang X, Lin CP, Yan D, Xue S, *et al*. Identification of potential therapeutic targets for atherosclerosis by analysing the gene signature related to different immune cells and immune regulators in atheromatous plaques. BMC Medical Genomics. 2021; 14: 145.
[50] King K, Moody A, Fisher SA, Mirza MM, Cuthbert AP, Hampe J, *et al*. Genetic variation in the IGSF6 gene and lack of association with inflammatory bowel disease. European Journal of Immunogenetics. 2003; 30: 187–190.
[51] Liu L, Yu Y, Hu LL, Dong QB, Hu F, Zhu LJ, *et al*. Potential Target Genes in the Development of Atrial Fibrillation: A Comprehensive Bioinformatics Analysis. Medical Science Monitor: International Medical Journal of Experimental and Clinical Research. 2021; 27: e928366.
[52] Yang AX, Chong N, Jiang Y, Catalano J, Puri RK, Khleif SN. Molecular characterization of antigen-peptide pulsed dendritic cells: immature dendritic cells develop a distinct molecular profile when pulsed with antigen peptide. PLoS ONE. 2014; 9: e86306.
[53] Rajamanickam A, Kumar NP, Pandiaraj AN, Selvaraj N, Munisankar S, Renji RM, *et al*. Restoration of dendritic cell homeostasis and Type I/Type III interferon levels in convalescent COVID-19 individuals. BMC Immunology. 2022; 23: 51.
[54] Chini CCS, Peclat TR, Warner GM, Kashyap S, Espindola-Netto JM, de Oliveira GC, *et al*. CD38 ecto-enzyme in immune cells is induced during aging and regulates NAD^+ and NMN levels. Nature Metabolism. 2020; 2: 1284–1304.
[55] Hogan KA, Chini CCS, Chini EN. The Multi-faceted Ecto-enzyme CD38: Roles in Immunomodulation, Cancer, Aging, and Metabolic Diseases. Frontiers in Immunology. 2019; 10: 1187.
[56] Guerreiro S, Privat AL, Bressac L, Toulorge D. CD38 in Neurodegeneration and Neuroinflammation. Cells. 2020; 9: 471.
[57] Szlasa W, Czarny J, Sauer N, Rakoczy K, Szymańska N, Stecko J, *et al*. Targeting CD38 in Neoplasms and Non-Cancer Diseases. Cancers. 2022; 14: 4169.
[58] Horenstein AL, Faini AC, Malavasi F. CD38 in the age of COVID-19: a medical perspective. Physiological Reviews. 2021; 101: 1457–1486.
[59] Zeidler JD, Kashyap S, Hogan KA, Chini EN. Implications of the NADase CD38 in COVID pathophysiology. Physiological Reviews. 2022; 102: 339–341.
[60] Matalonga J, Glaria E, Bresque M, Escande C, Carbó JM, Kiefer K, *et al*. The Nuclear Receptor LXR Limits Bacterial Infection of Host Macrophages through a Mechanism that Impacts Cellular NAD Metabolism. Cell Reports. 2017; 18: 1241–1255.
[61] Partida-Sánchez S, Cockayne DA, Monard S, Jacobson EL, Oppenheimer N, Garvy B, *et al*. Cyclic ADP-ribose production by CD38 regulates intracellular calcium release, extracellular calcium influx and chemotaxis in neutrophils and is required for bacterial clearance *in vivo*. Nature Medicine. 2001; 7: 1209–1216.
[62] Dong Z, Yan Q, Cao W, Liu Z, Wang X. Identification of key molecules in COVID-19 patients significantly correlated with clinical outcomes by analyzing transcriptomic data. Frontiers in Immunology. 2022; 13: 930866.
[63] Takeyama K, Aguiar RCT, Gu L, He C, Freeman GJ, Kutok JL, *et al*. The BAL-binding protein BBAP and related Deltex family members exhibit ubiquitin-protein isopeptide ligase activity. The Journal of Biological Chemistry. 2003; 278: 21930–21937.
[64] Yan Q, Dutt S, Xu R, Graves K, Juszczynski P, Manis JP, *et al*. BBAP monoubiquitylates histone H4 at lysine 91 and selectively modulates the DNA damage response. Molecular Cell. 2009; 36: 110–120.
[65] Zhang Y, Mao D, Roswit WT, Jin X, Patel AC, Patel DA, *et al*. PARP9-DTX3L ubiquitin ligase targets host histone H2BJ and viral 3C protease to enhance interferon signaling and control viral infection. Nature Immunology. 2015; 16: 1215–1227.
[66] Yan Q, Xu R, Zhu L, Cheng X, Wang Z, Manis J, *et al*. BAL1 and its partner E3 ligase, BBAP, link Poly(ADP-ribose) activation, ubiquitylation, and double-strand DNA repair independent of ATM, MDC1, and RNF8. Molecular and Cellular Biology. 2013; 33: 845–857.
[67] Russo LC, Tomasin R, Matos IA, Manucci AC, Sowa ST, Dale K, *et al*. The SARS-CoV-2 Nsp3 macrodomain reverses PARP9/DTX3L-dependent ADP-ribosylation induced by interferon signaling. The Journal of Biological Chemistry. 2021; 297: 101041.
[68] Pellegrina D, Bahcheli AT, Krassowski M, Reimand J. Human phospho-signaling networks of SARS-CoV-2 infection are rewired by population genetic variants. Molecular Systems Biology. 2022; 18: e10823.
[69] Liu ZM, Yang MH, Yu K, Lian ZX, Deng SL. Toll-like receptor (TLRs) agonists and antagonists for COVID-19 treatments. Frontiers in Pharmacology. 2022; 13: 989664.
[70] Giannakopoulos S, Strange DP, Jiyarom B, Abdelaal O, Bradshaw AW, Nerurkar VR, *et al*. *In vitro* evidence against productive SARS-CoV-2 infection of human testicular cells: Bystander effects of infection mediate testicular injury. PLOS Pathogens. 2023; 19: e1011409.
[71] Sultan RH, Elesawy BH, Ali TM, Abdallah M, Assal HH, Ahmed AE, *et al*. Correlations between Kidney and Heart Function Bioindicators and the Expressions of Toll-Like, ACE2, and NRP-1 Receptors in COVID-19. Vaccines. 2022; 10: 1106.