Check for updates

Identifying Methylation Signatures and Rules for COVID-19 With Machine Learning Methods

Zhandong Li[1†], Zi Mei[2†], Shijian Ding[3†], Lei Chen[4], Hao Li[1], Kaiyan Feng[5], Tao Huang[6,7]* and Yu-Dong Cai[3]*

[1]College of Biological and Food Engineering, Jilin Engineering Normal University, Changchun, China, [2]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, [3]School of Life Sciences, Shanghai University, Shanghai, China, [4]College of Information Engineering, Shanghai Maritime University, Shanghai, China, [5]Department of Computer Science, Guangdong AIB Polytechnic College, Guangzhou, China, [6]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, Shanghai, China, [7]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

The occurrence of coronavirus disease 2019 (COVID-19) has become a serious challenge to global public health. Definitive and effective treatments for COVID-19 are still lacking, and targeted antiviral drugs are not available. In addition, viruses can regulate host innate immunity and antiviral processes through the epigenome to promote viral self-replication and disease progression. In this study, we first analyzed the methylation dataset of COVID-19 using the Monte Carlo feature selection method to obtain a feature list. This feature list was subjected to the incremental feature selection method combined with a decision tree algorithm to extract key biomarkers, build effective classification models and classification rules that can remarkably distinguish patients with or without COVID-19. EPSTI1, NACAP1, SHROOM3, C19ORF35, and MX1 as the essential features play important roles in the infection and immune response to novel coronavirus. The six significant rules extracted from the optimal classifier quantitatively explained the expression pattern of COVID-19. Therefore, these findings validated that our method can distinguish COVID-19 at the methylation level and provide guidance for the diagnosis and treatment of COVID-19.

Keywords: methylation, COVID-19, feature selection, decision tree, rule

INTRODUCTION

Coronavirus disease 2019 (COVID-19) was announced as a "public health emergency of international concern" by the World Health Organization (WHO) on January 30, 2020 and was assessed as a global pandemic on March 11, 2020 (Rodríguez-Morales et al., 2020; Eurosurveillance Editorial Team, 2020). The causative agent of COVID-19 is a new type of coronavirus, whose complete gene sequence is approximately 79.5% similar to that of severe respiratory syndrome coronavirus SARS-CoV. Therefore, it was named SARS-CoV-2 by the International Virus Laboratory Classification (Zhou et al., 2020; Zhu et al., 2020). SARS-CoV-2 is a group 2B ß-coronavirus, which is a linear single-stranded positive-stranded RNA virus. It is similar to other coronaviruses and consists of four structural proteins, namely, spike protein, envelope protein, membrane protein/matrix protein, and nucleocapsid protein. COVID-19 has a huge impact on global public health. According to WHO, SARS-CoV-2 had caused 156,496,592

infections and 3,264,143 deaths worldwide until May 8, 2021, and a total of 1,171,658,745 vaccine doses had been administered worldwide till May 5, 2021 (World Health Organization, 2020). However, definite and effective treatments for COVID-19 are still lacking, and no antiviral drug has been confirmed by a rigorous "randomized, double-blind, placebo-controlled" study.

As early as 1975, researchers (Holliday and Pugh, 1975; Riggs, 1975) found that in vertebrates, cytosine methylation at the CpG site can be used as a genetic marker and can be passed on to the next generation by cell division. In plants and mammals, methylation on the 5th carbon atom of cytosine residues is the most widely studied epigenetic modification. In mammals, cytosine methylation mostly exists on the CG sequence; plants have CHG and CHH methylation (H = A, C, or T) in addition to CG methylation. DNA methylation is relatively stable and can exist persistently during DNA replication. The position of DNA methylation can be determined and its relationship with gene regulation can be explored with the development of whole-genome methylation sequencing technology. The DNA methylation of the promoter region can suppress gene expression by preventing transcription factor accessibility. Moreover, the DNA methylation of the gene body can affect chromatin structure, alternative splicing, and transcription efficiency (Lorincz et al., 2004). In mammals, such as *Homo sapiens* and *Mus musculus*, DNA methylation is necessary to maintain normal embryonic development (Yin et al., 2012; Guo et al., 2014), and abnormal methylation has remarkable effects on diseases (Robertson, 2005). In addition, methylation plays an important role in the regulation of the expression of tissue-specific genes or developmental stage-dependent genes (Gehring and Henikoff, 2007).

Strong evidence showed that epigenetic markers, including histone modifications, DNA methylation, chromatin remodeling, and non-coding RNAs, affect gene expression profiles and increase individual vulnerability to virus infections (Fang et al., 2012; Menachery et al., 2014). Meanwhile, viruses have developed a complex, highly evolved, and coordinated process that can regulate the host's epigenome, control the host's innate immune and antiviral defense processes, and thus promote the powerful replication of the virus and the onset of disease (Schäfer and Baric, 2017). Circulating blood DNA methylation profiles are altered in patients with severe diseases, including severe sepsis and pediatric critical illness (Binnie et al., 2020; Güiza et al., 2020).

In this study, we obtained methylation data from 106 SARS-CoV-2-positive patients and 26 SARS-CoV-2-negative patients. Machine learning algorithms, such as Monte Carlo feature selection (MCFS) (Dramiński et al., 2007) and decision tree (DT) (Safavian and Landgrebe, 1991), were applied to identify methylation features and decision rules that clearly distinguish different cases and to build classification models with excellent performance to provide insight into the diagnosis, susceptibility, and potential pathogenesis of COVID-19.

MATERIALS AND METHODS

Datasets
We downloaded the methylation data of the 128 samples from Gene Expression Omnibus with accession number GSE174818

(Balnis et al., 2021), which contains 102 samples from patients with COVID-19 and 26 samples from patients without COVID-19. For each sample, 86,5807 methylation sites were identified by the Illumina Human Methylation EPIC platform.

Monte Carlo Feature Selection
For the investigated methylation data, features (methylation sites) were much more than sample numbers. Evidently, not all features were related to COVID-19. It is necessary to analyze all features and extract essential ones. As different feature selection methods may produce quite different results, selection of proper methods was quite essential. To our knowledge, MCFS is good at dealing with data containing few samples and large features. Thus, it was adopted in this study.

The MCFS method (Dramiński et al., 2007) is an effective and broadly adopted feature selection method that is composed of various DTs and builds various bootstrap sets with subsets of randomly selected features. First, m bootstrap sets and t feature subsets are created from the primary data set. Then, one tree is constructed for m bootstrap sets and t feature subsets. Overall, $m \times t$ DTs are created. The relative importance (RI) score for each feature can be calculated based on the resultant DTs. RI score is calculated as the frequency of a target feature in the growing DT as follows:

$$RI_f = \sum_{t=1}^{mt} (wAcc)^u IG\left(n_f(\tau)\right) \left(\frac{\text{no.in } n_f(\tau)}{\text{no.in } \tau}\right)^v, \quad (1)$$

where f indicates a feature; $wAcc$ means the weighted accuracy of the DT τ; $IG(n_f(\tau))$ refers to the information gain of node $n_f(\tau)$; no.in $n_f(\tau)$ and $no.in\,\tau$ denote the number of samples of $n_f(\pi)$ and τ, respectively; and u and v are weighting factors, which are set to 1. More significant features have higher RI values. Therefore, the features were sorted in decreasing order based on their RI values in the new feature list after MCFS processing.

The MCFS program used in this study was loaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. For convenience, the program was run using the default parameters, and u and v were set to 1.

Incremental Feature Selection
Although the MCFS method can rank features by their importance, it cannot determine which features are essential. Therefore, the incremental feature selection (IFS) method (Liu and Setiono, 1998) was used to determine the optimal number of essential features required for the classification algorithm. First, IFS yields a series of feature subsets based on step size from the list of features received from the MCFS method described above. For example, when the step size is 5, the first feature subset includes the top 5 features, the second feature subset includes the top 10 features, and so on. Afterward, IFS trains the classifier on the training samples, which contain these features on each feature subset. The best subset was determined based on the evaluation metrics of the obtained model by evaluating this classifier through 10-fold cross-validation (Kohavi, 1995; Chen et al., 2021; Liu et al., 2021; Li X. et al., 2022; Tang and Chen, 2022; Yang and Chen, 2022).

Decision Tree

DT is one of the most classic machine learning algorithms (Safavian and Landgrebe, 1991). Although it is not very powerful, even much weaker than several strong machine learning algorithms, it also has its merits. In fact, DT is a white-box model, meaning it is possible for users to understand its classification principle. This cannot be achieved for all black-box models, which is always more powerful than DT. In the field of biomedical research, such merit is quite helpful as investigators want to not only build efficient models but also obtain helpful clues to understand the complicated underlying mechanism. Accordingly, DT is widely accepted in the field of biomedical research (Zhang et al., 2021a; Zhang et al., 2021b; Huang et al., 2021; Li Z. et al., 2022; Chen et al., 2022; Ding et al., 2022). Generally, DT uses the IF–THEN format to accomplish classification or regression tasks through a tree structure. It often yields satisfactory performance at a low computational cost. In this work, we applied the Scikit-learn module in Python to build the DT classifier.

Synthetic Minority Oversampling Technique

As described in the *Datasets Section*, a considerable variation in the sample sizes of patients with COVID-19 was observed. In this case, a classifier with excellent performance is difficult to build, because the predicted results are suitable for the type with the largest sample size. Synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) was performed in the present work to address this problem. This method ensures that the number of samples in the minority class is equal to the number of samples in the majority class after processing by adding new samples to the minority class. In detail, x is a randomly selected sample in the minor class, and some samples with the same class that are closest to x are identified. Next, sample y is randomly selected from the closest samples mentioned above, and a novel sample is produced by choosing a randomly selected point between x and y in the feature space. The newly produced sample is deeply associated with x and y; thus, it has a high probability of belonging to the same class as x and y and is therefore considered to be in the same class. The above procedure executes several times until the minority class has same number of samples in the majority class.

In this study, we employed the SMOTE procedure acquired from https://github.com/scikit-learn-contrib/imbalanced-learn and directly used the default parameters.

Performance Measurement

For the binary model used in this study, its predicted results can be counted as a confusion matrix, which contains four entries: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). According to these entries, several measurements can be calculated. In this study, we adopted the following measurements: sensitivity (SN, also called recall), specificity (SP), prediction accuracy (ACC), Matthews correlation coefficient (MCC), precision and F_1-measure (Sasaki, 2007; Powers, 2011; Zhao et al., 2018; Wu and Chen, 2022). They can be computed by

$$SN = \frac{TP}{TP + FN}, \tag{2}$$

$$SP = \frac{TN}{TN + FP}, \tag{3}$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}, \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}}, \tag{5}$$

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$F_1 - measure = \frac{2 \times precision \times recall}{precision + recall}. \tag{7}$$

Among above measurements, we selected F1-measure as the key measurement as it can better reflect the stability of the model. A higher F1-measure indicates a more robust classification model.
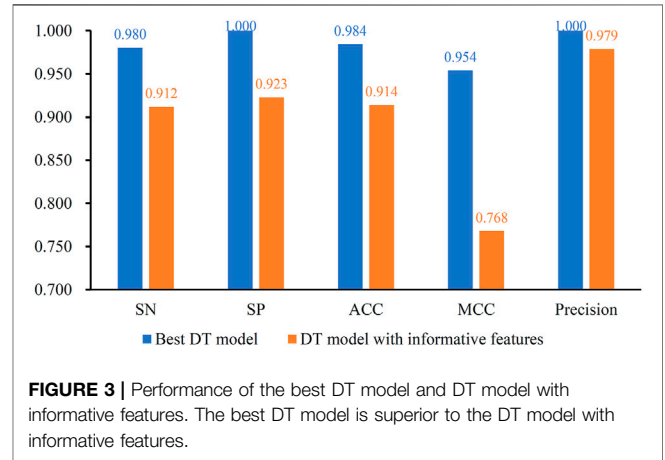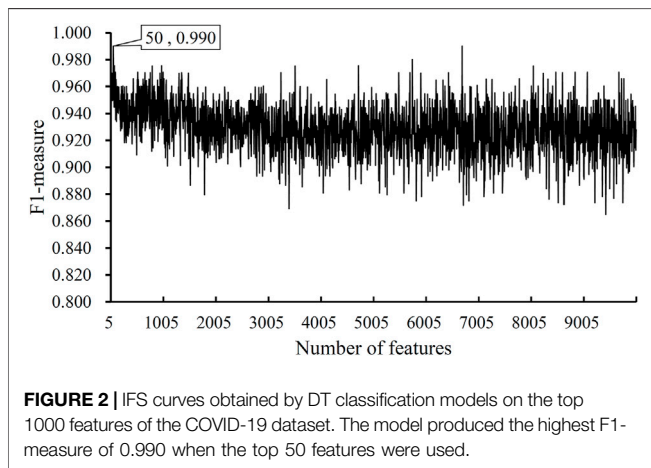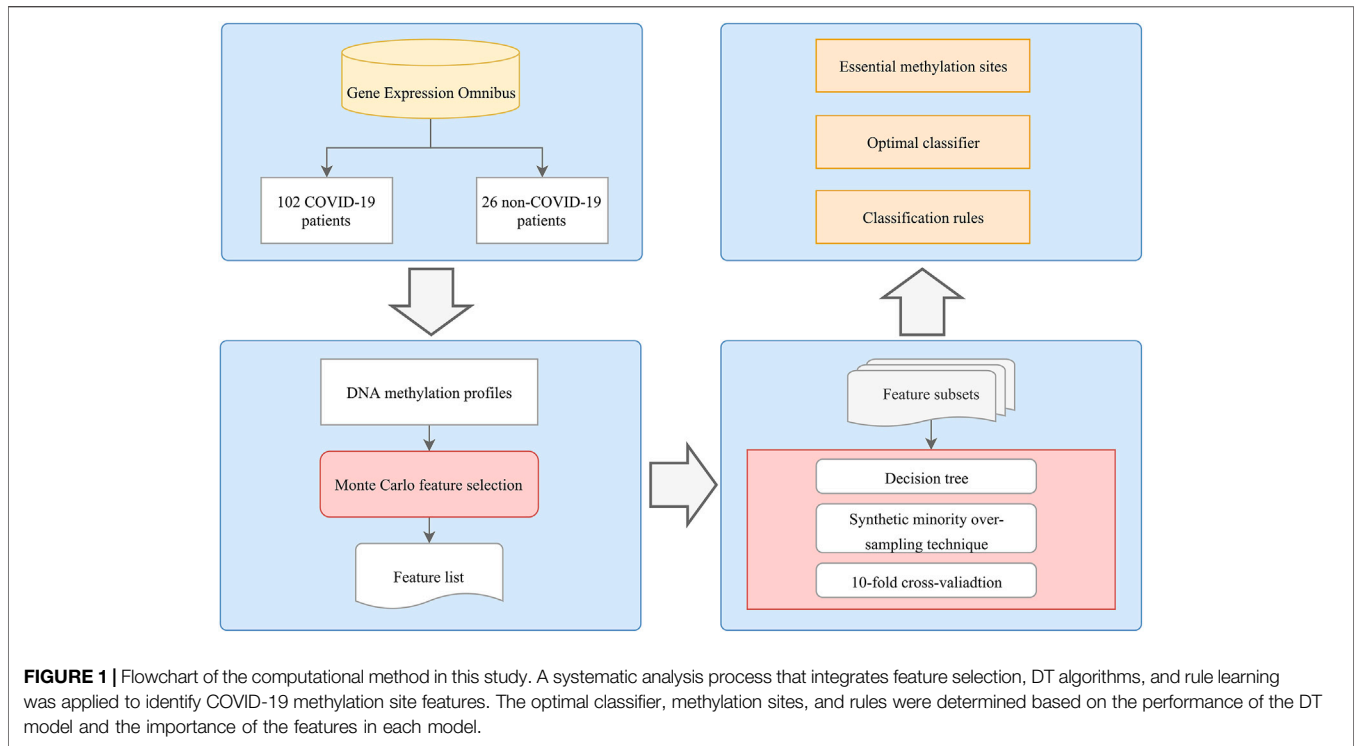
RESULTS

As shown in **Figure 1**, we applied an analysis flow to extract key features and build the classification model and rules. The results are summarized in the following sections.

Results of MCFS Method on Methylation Profiles

We employed the MCFS method to assess the importance of each feature and select key sites from the COVID-19 methylation dataset. These features were ranked in decreasing order of RI scores, and the results are presented in **Supplementary Table S1**.

Results of IFS Method With DT

After the MCFS analysis, we brought the obtained feature list into the IFS with the DT algorithm. The step size of the IFS was set to 5. Since the list was very large, it would take lots of time to consider all possible feature subsets. Furthermore, not all methylation features are related to COVID-19. Thus, only top 10,000 methylation features in the list were considered, that is, 2000 feature subsets were investigated. DT model was constructed on each feature subset and was evaluated by 10-fold cross-validation. The obtained evaluation metrics, including SN, SP, ACC, MCC, precision and F1-measure are listed in **Supplementary Table S2**. To display the DT models on different feature subsets, we plotted an IFS curve using the number of features as the X-axis and F1-measure as the Y-axis, which is shown in **Figure 2**. It can be observed that the highest F1-measure was 0.990, which was obtained by using top 50 features in the list. The other five measurements of such model are illustrated in **Figure 3**. The ACC and MCC were 0.984 and 0.954, respectively. As for SN, SP and precision, they were 0.980, 1.000 and 1.000, respectively. This result indicated that the constructed optimal DT model has a near-perfect performance and proved the effectiveness of the analysis method.

FIGURE 1 | Flowchart of the computational method in this study. A systematic analysis process that integrates feature selection, DT algorithms, and rule learning was applied to identify COVID-19 methylation site features. The optimal classifier, methylation sites, and rules were determined based on the performance of the DT model and the importance of the features in each model.

FIGURE 2 | IFS curves obtained by DT classification models on the top 1000 features of the COVID-19 dataset. The model produced the highest F1-measure of 0.990 when the top 50 features were used.

FIGURE 3 | Performance of the best DT model and DT model with informative features. The best DT model is superior to the DT model with informative features.

Comparison of DT Models With Informative Features

In this study, the IFS method was adopted to extract best features for DT and the best DT model was constructed with these features. In fact, MCFS can yield essential features, called informative features, by only analyzing the methylation data. With these informative features, a DT model can also be built. It is interesting to compare the performance of these two models.

For the methylation data, 257 informative features were obtained by MCFS. The DT model with these features was evaluated by 10-fold cross-validation. The F1-measure was 0.944, which was much lower than that of the best DT model (0.990). As for other five measurements, they are provided in

Figure 3. It can be observed that each measurement was lower than that produced by the best DT model. This indicated the superiority of the best DT model. The employment of IFS method can help to build a more efficient model.

Classification Rules

As described in the previous section, DT yielded the highest F1-measure on the COVID-19 methylation dataset when the top 50 features are used. Therefore, we applied DT to all samples using these 50 features to obtain six rules, which are provided in **Table 1**. Three rules each were related to COVID-19 and non-COVID-19. These rules clearly expressed the expression patterns of these features. These rules were described in detail in *Discussion Section*.

TABLE 1 | Rules yielded by decision tree on top 50 features.

Index	Condition	Result
Rule0	cg03753191 ≤ 0.1398 cg15959262 > 0.5931 cg17439158 > 0.5681	Patient with COVID-19
Rule1	cg03753191 > 0.1398 cg17439158 ≤ 0.6170	Patient without COVID-19
Rule2	cg03753191 ≤ 0.1398 cg15959262 ≤ 0.5931 cg08399733 ≤ 0.9130	Patient without COVID-19
Rule3	cg03753191 > 0.1398 cg17439158 > 0.6170	Patient with COVID-19
Rule4	cg03753191 ≤ 0.1398 cg15959262 > 0.5931 cg17439158 ≤ 0.5681	Patient without COVID-19
Rule5	cg03753191 ≤ 0.1398 cg15959262 ≤ 0.5931 cg08399733 > 0.9130	Patient with COVID-19

DISCUSSION

Our research is dedicated to search for pathogenic clues of SARS-CoV-2 infection based on the methylation profiles of COVID-19 in confirmed and suspected patients. Epigenetic markers, such as histone modifications, DNA methylation, chromatin remodeling, and non-coding RNAs, can affect gene expression profiles and increase individual susceptibility to the virus. For example, DNA methylation is the basis for antigen presentation and host adaptive immune response in Middle East respiratory syndrome coronavirus infection (Menachery et al., 2018). Therefore, we aimed to explore how DNA methylation influences SARS-CoV-2 infection.

Meanwhile, our research proposed a novel and creative pattern with a high distinguishing degree in confirmed and suspected COVID-19 cases through MCFS. Although the real-time polymerase chain reaction test of sputum is the gold standard for the diagnosis of COVID-19, it takes a long time to confirm the diagnosis of patients because of the high level of false negatives. Therefore, researchers conducted various methods to better identify SARS-CoV-2 infection. (Hemdan et al., 2020) constructed a novel deep learning classifier to diagnose COVID-19 through X-ray images, which are cheaper, more convenient, and accessible compared with traditional chest X-ray and computed tomography. (Siddiqui et al., 2020) focused on the correlation of temperature with suspected, confirmed, and death cases by machine learning and found that temperature presents diverse trends in most cities and cannot be the decisive factor in different cases or situations. We focused on several top features and decision rules because they have a crucial impact on the classification and discussed them further through a wide literature publication to prove that our findings are reliable and convincing.

Epithelial stromal interaction 1 (EPSTI1, probeID: cg03753191) is an interferon (IFN)-responsive gene that was originally isolated from mixed cultured human breast cancer cells and fibroblasts (Nielsen et al., 2002). This gene is located on chromosome 13q13.3; is 104.2 kb in length; contains 11 exons; and is involved in tumor cell metastasis, epithelial–mesenchymal transition, chronic inflammation, tissue reconstruction, embryonic development, and other biological processes (De Neergaard et al., 2010). EPSTI1 plays an important role in the regulation of cell apoptosis. (Capdevila-Busquets et al., 2015) confirmed by *in vitro* experiments that EPSTI1 can inhibit breast cancer cell apoptosis by interacting with caspase 8. In addition, EPSTI1 has an antiviral effect against hepatitis C virus (HCV) by

affecting the life cycle, viral replication, assembly, and release of HCV. (Meng et al., 2015) confirmed that EPSTI1 can promote the expression of protein kinase-R (PKR)-dependent genes, including *IFNβ*, *IFIT1*, *OAS1*, and *RNase L*, by activating the promoter of PKR to play an antiviral effect during the process of HCV infection. Without the involvement of IFN treatment, EPSTI1 overexpression effectively suppressed HCV replication, whereas the lack of EPSTI1 enhanced the viral activities. Current research discovered that EPSTI1 expression influences the performance of immune cells. (Kim et al., 2018) found that EPSTI1 expression is remarkably upregulated after macrophage activation with IFNγ and lipopolysaccharide. The proportion of M2-type macrophages is increased in the bone marrow-derived macrophage deficiency of EPSTI1. In EPSTI1 knockout mice, the number of M1 macrophage cells in the peritoneal cavity was significantly reduced. These findings demonstrate an important regulatory role of EPSTI1 in macrophage polarization. Therefore, EPSTI1 methylation may participate in the process of SARS-CoV-2 infection and affect inflammatory and immune function by regulating EPSTI1 expression.

NACAP1 (probeID: cg15959262) is a pseudogene of nascent polypeptide-associated complex-alpha (*NACA*). Phosphatase and tensin homolog (PTEN) pseudogene 1 (PTENP1) has been first revealed to contain microRNA response elements (MREs), which also exist in its corresponding protein-coding gene, *PTEN* (Poliseno et al., 2010). Increasing pseudogenes are found to have a similar phenomenon, that is, pseudogenes and their corresponding protein-coding genes function as competitive endogenous RNAs for binding to the same microRNAs (Lujambio and Lowe, 2012; Karreth et al., 2015). NACA encodes the *a* chain of nascent polypeptide-associated complex (NAC), which performs multiple functions, including protecting newborn peptides and regulating the translocation of new peptides into the endoplasmic reticulum and mitochondria (Rospert et al., 2002). The alpha chain of NAC alone acts as a transcriptional co-activator for developmental regulation (Yotov et al., 1998). Furthermore, NACA can regulate the conformation of Fas-associated death domain protein oligomer, which is an important mediator in the signal transduction pathway and can be activated by several members of the tumor necrosis factor (TNF) receptor family (Liguoro et al., 2003). NACA is related to neurodegenerative diseases. Patients with Alzheimer's disease and Down's syndrome have lower NACA expression levels in their brain cells (Kim et al., 2002). More importantly, the inhibition of NACA can induce the proliferation and differentiation of CD8[+] T cells and enhance cytotoxicity. (Al-

Shanti and Aldahoodi, 2006) used anti-sense technology to reduce the concentration of mRNA that translates NACA and found that CD8$^+$ T cells will differentiate and activate to a higher degree in the presence of antisense oligonucleotide chains. Compared with the control group, the lethality of CD8$^+$ T cells on target cells was enhanced.

SHROOM3 (probeID: cg17439158), a member of the Shroom family, encodes an actin-binding protein, which is important in epithelial cell shape and tissue morphogenesis (Haigo et al., 2003; Hildebrand, 2005). Shoom3 overexpression in epithelial cells induces rho kinase (Rock) recruitment and increases myosin 2 (Myo2) accumulation through phosphorylation and activation. The activation of the Rock/Myo2 signaling pathway leads to the local contraction of the actomyosin network on the top surface of the cell, which results in changes in cell morphology. Recent research has proved the indispensable role of SHROOM3 in glomerular filtration barrier integrity (Yeo et al., 2015). Forced Shroom3 expression in fawn-hooded hypertensive rat and endogenous shroom3 knockdown zebrafish improved kidney glomerular function. Moreover, multiple genome-wide association studies and *in vivo* experiments strongly demonstrated the correlation between SHROOM3 and congenital kidney disease (Khalili et al., 2016).

C19ORF35 (probeID: cg08399733), also named *PEAK3*, is a member of the New Kinase Family 3 (NKF3) that can regulate cytoskeleton stability and cell motility by binding with an adaptor protein, CrkII (Lopez et al., 2019). C19ORF35 is associated with cancer progression. C19ORF35 overexpression has been detected in various cancers, including pancreatic, breast, and colon cancers (Wang et al., 2010; Kelber et al., 2012; Fujimura et al., 2014). C19orf35 methylation is related to early carcinogenesis. According to a DNA methylation sequencing study on 12 patients with early gastric cancers (EGCs), C19orf35 is remarkably hypomethylated in the diffuse type of EGC tissue compared with adjacent corresponding non-tumor mucosal tissue (Chong et al., 2014).

IFN-induced with helicase C domain 1 (*IFIH1*, probeID: cg21060789) and IFN-induced protein 44-like (*IFI44L*, probeID: cg13452062) are IFN-stimulated genes. IFIH1, also known as melanoma differentiation-associated gene-5, is a cytoplasmic RNA receptor protein composed of 1025 amino acids. IFIH1 recognizes double-stranded RNA with a length of more than 1 kb. It is an important member of the retinoic acid-inducible gene I (RIG-I)-like receptor family, which can activate type I IFN signaling pathway and participate in the pathogenesis of a variety of autoimmune diseases. IFIH1 and IFN-β interact to activate the body's anti-tumor immune response. IFIH1 can promote type I IFN response and increase the secretion of TNF-α and IFN-β. The upregulation of IFIH1 expression may increase the effectiveness of IFN therapy (Pappas et al., 2009). IFN-β can also stimulate the upregulation of IFIH1 and RIG-I, mediate innate immune response, kill tumor cells with low neurotoxicity, and therefore inhibit tumor growth (Wu et al., 2017; Bufalieri et al., 2020). *IFI44L* is a paralog gene of *IFI44* and functions as a regulator of cell apoptosis, virus infection, and congenital immune response. The DNA methylation level of the IF144L promoter may be related to kidney damage in patients with systemic lupus erythematosus (SLE). (Zhao et al., 2016) found that the DNA methylation level of IFI44L promoter in patients with SLE

was remarkably lower than that of the normal control group. In addition, the DNA methylation level of the IFI44L promoter in patients with SLE and renal involvement was also remarkably lower than that of patients with SLE without renal involvement. IFI44L participates in the antiviral process of IFN-mediated innate immune response and is a confirmed marker of early viral infection. IFN is the earliest discovered cytokine that can inhibit viral infection and replication and is activated in the early stage of viral infection (within a few minutes to a few hours) (Zaas et al., 2009). (Henrickson et al., 2018) pointed out that when influenza virus and respiratory syncytial virus infections occur, *IFI44L* acts as an IFN-stimulated factor regulatory gene, and its expression level increases. Therefore, the mRNA expression of *IFI44L* can be used as an early indicator of virus infection. According to (Kaforou et al., 2017), in 111 bacterial-infected children less than 60 days old, the detection sensitivity of IFI44L mRNA expression is 88.8% (95% CI, 80.3–94.5%), and the specific degree is 93.7% (95% CI, 87.4–97.4%). Therefore, aberrant IFI44L methylation may occur during SARS-CoV2 infection and lead to abnormal IFI44L expression.

MX1 (probeID: cg26312951) belongs to human mycovirus resistance genes (*MX*) with biological functions, such as GTP binding and GTP enzyme activities. The two kinds of MX proteins, namely, MX1 and MX2, differ greatly in virus specificity and mechanism of action. MX1 has antiviral activity against a variety of RNA viruses and certain DNA viruses induced by type I and type II IFNs, including negative-strand RNA viruses and hepatitis B virus. MX1 is enriched in the IFN-γ and Toll-like receptor signaling pathways (Haller et al., 2015). A 2020 study detailed the expression of MX1 in 403 patients with COVID-19 and 50 patients without COVID-19 (Bizzotto et al., 2020). The expression of MX1, MX2, ACE2, and BSG/CD147 can cluster individuals with and without COVID-19 through principal component analysis, which indicated that the expression levels of MX1 and MX2 between patients with and without COVID-19 are remarkably different. MX1 can directly act on the ribonucleoprotein complex of the virus, so it has a wide range of antiviral activity. This feature has been proven to be suitable for RNA viruses and DNA viruses. (Verhelst et al., 2012) have reported that Mx1 interfering with functional viral ribonucleoprotein complex assembly led to inhibition of influenza virus and high MX1 expression can result in a better prognosis to influenza A (H1N1) pandemic in 2009. It is worth noting that GTPase activity is also positively correlated with its antiviral function. Different from MX1, the antiviral function of MX2 is limited to certain viruses, such as HIV. Although the expression of MX1 and MX2 in patients with COVID-19 were significantly higher than those in non-COVID-19 groups, MX1 shows a greater positive correlation with patients with COVID-19 and may be more specific than MX2 in response to SARS-CoV-2. Therefore, MX1 is a key responder to SARS-CoV-2 infection.

Collectively, the top identified discriminative feature genes and settled rules have a crucial role in virus infection and IFN-mediated immune response. This method demonstrates that our method is reliable and convincing. Our newly presented computational approach based on methylation profiles also provides a new perspective for exploring the mechanism of COVID-19. Furthermore, it is a new method that distinguishes confirmed and suspected COVID-19 cases and

has applicable clinical value in the differential diagnosis of patients with confirmed and suspected COVID-19.

CONCLUSION

The current study aimed to apply computational methods to extract the best biological features and decision rules from COVID-19 methylation profiles. This study has shown that the extracted optimal methylation site signatures and expression rules have been validated by previous work and are reliable and valid for distinguishing COVID-19. This study provides a new set of potential biomarkers/rules that can be used to differentiate patients with COVID-19 at the methylation level. These findings enhance our understanding of COVID-19 expression at the methylation level and could offer guidance for future studies on COVID-19.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE174818.

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. SD, LC, and KF performed the experiments. ZL, ZM, and HL analyzed the results. ZL, ZM, and SD wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.908080/full#supplementary-material

REFERENCES

Al-Shanti, N., and Aldahoodi, Z. (2006). Inhibition of Alpha Nascent Polypeptide Associated Complex Protein May Induce Proliferation, Differentiation and Enhance the Cytotoxic Activity of Human CD8+ T Cells. *J. Clin. Immunol.* 26, 457–464. doi:10.1007/s10875-006-9041-3

Balnis, J., Madrid, A., Hogan, K. J., Drake, L. A., Chieng, H. C., Tiwari, A., et al. (2021). Blood DNA Methylation and COVID-19 Outcomes. *Clin. Epigenet* 13, 118. doi:10.1186/s13148-021-01102-9

Binnie, A., Walsh, C. J., Hu, P., Dwivedi, D. J., Fox-Robichaud, A., Liaw, P. C., et al. (2020). Epigenetic Profiling in Severe Sepsis. *Crit. care Med.* 48, 142–150. doi:10.1097/ccm.0000000000004097

Bizzotto, J., Sanchis, P., Abbate, M., Lage-Vickers, S., Lavignolle, R., Toro, A., et al. (2020). SARS-CoV-2 Infection Boosts MX1 Antiviral Effector in COVID-19 Patients. *Iscience* 23, 101585. doi:10.1016/j.isci.2020.101585

Bufalieri, F., Caimano, M., Lospinoso Severini, L., Basili, I., Paglia, F., Sampirisi, L., et al. (2020). The RNA-Binding Ubiquitin Ligase MEX3A Affects Glioblastoma Tumorigenesis by Inducing Ubiquitylation and Degradation of RIG-I. *Cancers* 12, 321. doi:10.3390/cancers12020321

Capdevila-Busquets, E., Badiola, N., Arroyo, R., Alcalde, V., Soler-López, M., and Aloy, P. (2015). Breast Cancer Genes PSMC3IP and EPSTI1 Play a Role in Apoptosis Regulation. *PLoS One* 10, e0115352. doi:10.1371/journal.pone.0115352

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *jair* 16, 321–357. doi:10.1613/jair.953

Chen, L., Li, Z., Zhang, S., Zhang, Y. H., Huang, T., and Cai, Y. D. (2022). Predicting RNA 5-methylcytosine Sites by Using Essential Sequence Features and Distributions. *Biomed. Res. Int.* 2022,11 4035462. doi:10.1155/2022/4035462

Chen, W., Chen, L., and Dai, Q. (2021). iMPT-FDNPL: Identification of Membrane Protein Types with Functional Domains and a Natural Language Processing Approach. *Comput. Math. Methods Med.* 2021,10 7681497. doi:10.1155/2021/7681497

Chong, Y., Mia-Jan, K., Ryu, H., Abdul-Ghafar, J., Munkhdelger, J., Lkhagvadorj, S., et al. (2014). DNA Methylation Status of a Distinctively Different Subset of Genes Is Associated with Each Histologic Lauren Classification Subtype in Early Gastric Carcinogenesis. *Oncol. Rep.* 31, 2535–2544. doi:10.3892/or.2014.3133

De Neergaard, M., Kim, J., Villadsen, R., Fridriksdottir, A. J., Rank, F., Timmermans-Wielenga, V., et al. (2010). Epithelial-stromal Interaction 1 (EPSTI1) Substitutes for Peritumoral Fibroblasts in the Tumor Microenvironment. *Am. J. pathology* 176, 1229–1240. doi:10.2353/ajpath.2010.090648

Ding, S., Wang, D., Zhou, X., Chen, L., Feng, K., Xu, X., et al. (2022). Predicting Heart Cell Types by Using Transcriptome Profiles and a Machine Learning Method. *Life* 12, 228. doi:10.3390/life12020228

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo Feature Selection for Supervised Classification. *Bioinformatics* 24, 110–117. doi:10.1093/bioinformatics/btm486

Eurosurveillance Editorial Team (2020). Note from the Editors: World Health Organization Declares Novel Coronavirus (2019-nCoV) Sixth Public Health Emergency of International Concern. *Euro Surveill.* 25, 200131e. doi:10.2807/1560-7917.ES.2020.25.5.200131e

Fang, T. C., Schaefer, U., Mecklenbrauker, I., Stienen, A., Dewell, S., Chen, M. S., et al. (2012). Histone H3 Lysine 9 Di-methylation as an Epigenetic Signature of the Interferon Response. *J. Exp. Med.* 209, 661–669. doi:10.1084/jem.20112343

Fujimura, K., Wright, T., Strnadel, J., Kaushal, S., Metildi, C., Lowy, A. M., et al. (2014). A Hypusine-eIF5A-PEAK1 Switch Regulates the Pathogenesis of Pancreatic Cancer. *Cancer Res.* 74, 6671–6681. doi:10.1158/0008-5472.can-14-1031

Gehring, M., and Henikoff, S. (2007). DNA Methylation Dynamics in Plant Genomes. *Biochimica Biophysica Acta (BBA) - Gene Struct. Expr.* 1769, 276–286. doi:10.1016/j.bbaexp.2007.01.009

Güiza, F., Vanhorebeek, I., Verstraete, S., Verlinden, I., Derese, I., Ingels, C., et al. (2020). Effect of Early Parenteral Nutrition during Paediatric Critical Illness on DNA Methylation as a Potential Mediator of Impaired Neurocognitive Development: a Pre-planned Secondary Analysis of the PEPaNIC International Randomised Controlled Trial. *Lancet Respir. Med.* 8, 288–303. doi:10.1016/S2213-2600(20)30046-1

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., et al. (2014). The DNA Methylation Landscape of Human Early Embryos. *Nature* 511, 606–610. doi:10.1038/nature13544

Haigo, S. L., Hildebrand, J. D., Harland, R. M., and Wallingford, J. B. (2003). Shroom Induces Apical Constriction and Is Required for Hingepoint

Formation during Neural Tube Closure. *Curr. Biol.* 13, 2125–2137. doi:10.1016/j.cub.2003.11.054

Haller, O., Staeheli, P., Schwemmle, M., and Kochs, G. (2015). Mx GTPases: Dynamin-like Antiviral Machines of Innate Immunity. *Trends Microbiol.* 23, 154–163. doi:10.1016/j.tim.2014.12.003

Hemdan, E. E.-D., Shouman, M. A., and Karar, M. E. (2020). Covidx-net: A Framework of Deep Learning Classifiers to Diagnose Covid-19 in X-Ray Images. doi:10.4850/arXiv.2003.1105

Henrickson, S. E., Manne, S., Dolfi, D. V., Mansfield, K. D., Parkhouse, K., Mistry, R. D., et al. (2018). Genomic Circuitry Underlying Immunological Response to Pediatric Acute Respiratory Infection. *Cell Rep.* 22, 411–426. doi:10.1016/j.celrep.2017.12.043

Hildebrand, J. D. (2005). Shroom Regulates Epithelial Cell Shape *via* the Apical Positioning of an Actomyosin Network. *J. Cell Sci.* 118, 5191–5203. doi:10.1242/jcs.02626

Holliday, R., and Pugh, J. E. (1975). DNA Modification Mechanisms and Gene Activity during Development. *Science* 187, 226–232. doi:10.1126/science.187.4173.226

Huang, G.-H., Zhang, Y.-H., Chen, L., Li, Y., Huang, T., and Cai, Y.-D. (2021). Identifying Lung Cancer Cell Markers with Machine Learning Methods and Single-Cell RNA-Seq Data. *Life* 11, 940. doi:10.3390/life11090940

Kaforou, M., Herberg, J. A., Wright, V. J., Coin, L. J. M., and Levin, M. (2017). Diagnosis of Bacterial Infection Using a 2-transcript Host RNA Signature in Febrile Infants 60 Days or Younger. *Jama* 317, 1577–1578. doi:10.1001/jama.2017.1365

Karreth, F. A., Reschke, M., Ruocco, A., Ng, C., Chapuy, B., Léopold, V., et al. (2015). The BRAF Pseudogene Functions as a Competitive Endogenous RNA and Induces Lymphoma *In Vivo*. *Cell* 161, 319–332. doi:10.1016/j.cell.2015.02.043

Kelber, J. A., Reno, T., Kaushal, S., Metildi, C., Wright, T., Stoletov, K., et al. (2012). KRas Induces a Src/PEAK1/ErbB2 Kinase Amplification Loop that Drives Metastatic Growth and Therapy Resistance in Pancreatic Cancer. *Cancer Res.* 72, 2554–2564. doi:10.1158/0008-5472.can-11-3552

Khalili, H., Sull, A., Sarin, S., Boivin, F. J., Halabi, R., Svajger, B., et al. (2016). Developmental Origins for Kidney Disease Due to Shroom3 Deficiency. *Journal of the American Society of Nephrology* 27, 2965–2973. doi:10.1681/asn.2015060621

Kim, S. H., Shim Ki, S., and Lubec, G. (2002). Human Brain Nascent Polypeptide-Associated Complex α Subunit Is Decreased in Patients with Alzheimer's Disease and Down Syndrome. *J. Investig. Med.* 50, 293–301. doi:10.2310/6650.2002.33287

Kim, Y.-H., Lee, J.-R., and Hahn, M.-J. (2018). Regulation of Inflammatory Gene Expression in Macrophages by Epithelial-Stromal Interaction 1 (Epsti1). *Biochem. Biophysical Res. Commun.* 496, 778–783. doi:10.1016/j.bbrc.2017.12.014

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2 (Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.).doi:10.5555/1643031.1643047

Li, X., Lu, L., Lu, L., and Chen, L. (2022a). Identification of Protein Functions in Mouse with a Label Space Partition Method. *Mbe* 19, 3820–3842. doi:10.3934/mbe.2022176

Li, Z., Wang, D., Liao, H., Zhang, S., Guo, W., Chen, L., et al. (2022b). Exploring the Genomic Patterns in Human and Mouse Cerebellums *via* Single-Cell Sequencing and Machine Learning Method. *Front. Genet.* 13, 857851. doi:10.3389/fgene.2022.857851

Liu, H., Hu, B., Chen, L., and Lu, L. (2021). Identifying Protein Subcellular Location with Embedding Features Learned from Networks. *Cp* 18, 646–660. doi:10.2174/1570164617999201124142950

Liu, H., and Setiono, R. (1998). Incremental Feature Selection. *Appl. Intell.* 9, 217–230. doi:10.1023/a:1008363719778

Lopez, M. L., Lo, M., Kung, J. E., Dudkiewicz, M., Jang, G. M., Von Dollen, J., et al. (2019). PEAK3/C19orf35 Pseudokinase, a New NFK3 Kinase Family Member, Inhibits CrkII through Dimerization. *Proc. Natl. Acad. Sci. U.S.A.* 116, 15495–15504. doi:10.1073/pnas.1906360116

Lorincz, M. C., Dickerson, D. R., Schmitt, M., and Groudine, M. (2004). Intragenic DNA Methylation Alters Chromatin Structure and Elongation Efficiency in Mammalian Cells. *Nat. Struct. Mol. Biol.* 11, 1068–1075. doi:10.1038/nsmb840

Lujambio, A., and Lowe, S. W. (2012). The Microcosmos of Cancer. *Nature* 482, 347–355. doi:10.1038/nature10888

Menachery, V. D., Eisfeld, A. J., Schäfer, A., Josset, L., Sims, A. C., Proll, S., et al. (2014). Pathogenic Influenza Viruses and Coronaviruses Utilize Similar and Contrasting Approaches to Control Interferon-Stimulated Gene Responses. *MBio* 5, e01174–14. doi:10.1128/mBio.01174-14

Menachery, V. D., Schäfer, A., Burnum-Johnson, K. E., Mitchell, H. D., Eisfeld, A. J., Walters, K. B., et al. (2018). MERS-CoV and H5N1 Influenza Virus Antagonize Antigen Presentation by Altering the Epigenetic Landscape. *Proc. Natl. Acad. Sci. U. S. A.* 115, E1012. doi:10.1073/pnas.1706928115

Meng, X., Yang, D., Yu, R., and Zhu, H. (2015). EPSTI1 Is Involved in IL-28A-mediated Inhibition of HCV Infection. *Mediat. Inflamm.* 2015, 716315. doi:10.1155/2015/716315

Nielsen, H. L., Rønnov-Jessen, L., Villadsen, R., and Petersen, O. W. (2002). Identification of EPSTI1, a Novel Gene Induced by Epithelial-Stromal Interaction in Human Breast Cancer. *Genomics* 79, 703–710. doi:10.1006/geno.2002.6755

Pappas, D. J., Coppola, G., Gabatto, P. A., Gao, F., Geschwind, D. H., Oksenberg, J. R., et al. (2009). Longitudinal System-Based Analysis of Transcriptional Responses to Type I Interferons. *Physiol. genomics* 38, 362–371. doi:10.1152/physiolgenomics.00058.2009

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A Coding-independent Function of Gene and Pseudogene mRNAs Regulates Tumour Biology. *Nature* 465, 1033–1038. doi:10.1038/nature09144

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to roc., Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.* 2, 37–63.

Riggs, A. D. (1975). X Inactivation, Differentiation, and DNA Methylation. *Cytogenet Genome Res.* 14, 9–25. doi:10.1159/000130315

Robertson, K. D. (2005). DNA Methylation and Human Disease. *Nat. Rev. Genet.* 6, 597–610. doi:10.1038/nrg1655

Rodríguez-Morales, A. J., Macgregor, K., Kanagarajah, S., Patel, D., and Schlagenhauf, P. (2020). Going Global - Travel and the 2019 Novel Coronavirus. *Travel Med. Infect. Dis.* 33, 101578. doi:10.1016/j.tmaid.2020.101578

Rospert, S., Dubaquié, Y., and Gautschi, M. (2002). Nascent-polypeptide-associated Complex. *CMLS, Cell. Mol. Life Sci.* 59, 1632–1639. doi:10.1007/pl00012490

Safavian, S. R., and Landgrebe, D. (1991). "A Survey of Decision Tree Classifier Methodology,"*Systems, Man and Cybernetics*.21 doi:10.1109/21.97458

Sasaki, Y. (2007). The Truth of the F-Measure. *Teach. Tutor mater*, 1–5.

Schäfer, A., and Baric, R. S. (2017). Epigenetic Landscape during Coronavirus Infection. *Pathogens* 6, 8.

Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H. M. N., Hussain, F., Khatoon, K., et al. (2020). Correlation between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases Based on Machine Learning Analysis. *J. Pure Appl. Microbiol.* 14, 1017–1024. doi:10.22207/jpam.14.spl1.40

Stilo, R., Liguoro, D., di Jeso, B., Leonardi, A., and Vito, P. (2003). The Alpha-Chain of the Nascent Polypeptide-Associated Complex Binds to and Regulates FADD Function. *Biochem. Biophys. Res. Commun.* 303, 1034–1041. doi:10.1016/s0006-291x(03)00487-x

Tang, S., and Chen, L. (2022). iATC-NFMLP: Identifying Classes of Anatomical Therapeutic Chemicals Based on Drug Networks, Fingerprints and Multilayer Perceptron. *Curr. Bioinforma.* doi:10.2174/1574893617666220318093000

Verhelst, J., Parthoens, E., Schepens, B., Fiers, W., and Saelens, X. (2012). Interferon-inducible Protein Mx1 Inhibits Influenza Virus by Interfering with Functional Viral Ribonucleoprotein Complex Assembly. *J. Virol.* 86, 13445. doi:10.1128/jvi.01682-12

Wang, Y., Kelber, J. A., Cao, H. S. T., Cantin, G. T., Lin, R., Wang, W., et al. (2010). Pseudopodium-enriched Atypical Kinase 1 Regulates the Cytoskeleton and Cancer Progression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10920. doi:10.1073/pnas.0914776107

World Health Organization (2020). *World Health Organization Coronavirus Disease (COVID-19) Pandemic*.

Wu, Y., Wu, X., Wu, L., Wang, X., and Liu, Z. (2017). The Anticancer Functions of RIG-I-like Receptors, RIG-I and MDA5, and Their Applications in Cancer Therapy. *Transl. Res.* 190, 51–60. doi:10.1016/j.trsl.2017.08.004

Wu, Z., and Chen, L. (2022). Similarity-based Method with Multiple-Feature Sampling for Predicting Drug Side Effects. *Comput. Math. Methods Med.* 2022,13 9547317. doi:10.1155/2022/9547317

Yang, Y., and Chen, L. (2022). Identification of Drug-Disease Associations by Using Multiple Drug and Disease Networks. *Cbio* 17, 48–59. doi:10.2174/1574893616666210825115406

Yeo, N. C., O'Meara, C. C., Bonomo, J. A., Veth, K. N., Tomar, R., Flister, M. J., et al. (2015). Shroom3 Contributes to the Maintenance of the Glomerular Filtration Barrier Integrity. *Genome Res.* 25, 57–65. doi:10.1101/gr.182881.114

Yin, L.-J., Zhang, Y., Lv, P.-P., He, W.-H., Wu, Y.-T., Liu, A.-X., et al. (2012). Insufficient Maintenance DNA Methylation Is Associated with Abnormal Embryonic Development. *BMC Med.* 10, 26. doi:10.1186/1741-7015-10-26

Yotov, W. V., Moreau, A., and St-Arnaud, R. (1998). The Alpha Chain of the Nascent Polypeptide-Associated Complex Functions as a Transcriptional Coactivator. *Mol. Cell Biol.* 18, 1303–1311. doi:10.1128/mcb.18.3.1303

Zaas, A. K., Chen, M., Varkey, J., Veldman, T., Hero, A. O., Iii, Lucas, J., et al. (2009). Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans. *Cell Host Microbe* 6, 207–217. doi:10.1016/j.chom.2009.07.006

Zhang, Y.-H., Li, H., Zeng, T., Chen, L., Li, Z., Huang, T., et al. (2021a). Identifying Transcriptomic Signatures and Rules for SARS-CoV-2 Infection. *Front. Cell Dev. Biol.* 8, 627302. doi:10.3389/fcell.2020.627302

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021b). Determining Protein-Protein Functional Associations by Functional Rules Based on Gene Ontology and KEGG Pathway. *Biochimica Biophysica Acta (BBA) - Proteins Proteomics* 1869, 140621. doi:10.1016/j.bbapap.2021.140621

Zhao, M., Zhou, Y., Zhu, B., Wan, M., Jiang, T., Tan, Q., et al. (2016). IFI44L Promoter Methylation as a Blood Biomarker for Systemic Lupus Erythematosus. *Ann. Rheum. Dis.* 75, 1998–2006. doi:10.1136/annrheumdis-2015-208410

Zhao, X., Chen, L., and Lu, J. (2018). A Similarity-Based Method for Prediction of Drug Side Effects with Heterogeneous Information. *Math. Biosci.* 306, 136–144. doi:10.1016/j.mbs.2018.09.010

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733. doi:10.1056/nejmoa2001017