frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 17 March 2023
DOI 10.3389/fmicb.2023.1138674

OPEN ACCESS

EDITED BY
Lihong Peng,
Hunan University of Technology,
China

REVIEWED BY
Nizhuan Wang,
ShanghaiTech University,
China
Qi Dai,
Zhejiang Sci-Tech University,
China

*CORRESPONDENCE
Tao Huang
✉ tohuangtao@126.com
Yu-Dong Cai
✉ cai_yud@126.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 06 January 2023
ACCEPTED 01 March 2023
PUBLISHED 17 March 2023

CITATION
Li J, Ren J, Liao H, Guo W, Feng K, Huang T and Cai Y-D (2023) Identification of dynamic gene expression profiles during sequential vaccination with ChAdOx1/BNT162b2 using machine learning methods.
Front. Microbiol. 14:1138674.
doi: 10.3389/fmicb.2023.1138674

Identification of dynamic gene expression profiles during sequential vaccination with ChAdOx1/BNT162b2 using machine learning methods

Jing Li[1†], JingXin Ren[2†], HuiPing Liao[3†], Wei Guo[4], KaiYan Feng[5], Tao Huang[6,7*] and Yu-Dong Cai[2*]

[1]School of Computer Science, Baicheng Normal University, Baicheng, Jilin, China, [2]School of Life Sciences, Shanghai University, Shanghai, China, [3]Changping Laboratory, Beijing, China, [4]Key Laboratory of Stem Cell Biology, Shanghai Jiao Tong University School of Medicine (SJTUSM) and Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai, China, [5]Department of Computer Science, Guangdong AIB Polytechnic College, Guangzhou, China, [6]CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Science, 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

To date, COVID-19 remains a serious global public health problem. Vaccination against SARS-CoV-2 has been adopted by many countries as an effective coping strategy. The strength of the body's immune response in the face of viral infection correlates with the number of vaccinations and the duration of vaccination. In this study, we aimed to identify specific genes that may trigger and control the immune response to COVID-19 under different vaccination scenarios. A machine learning-based approach was designed to analyze the blood transcriptomes of 161 individuals who were classified into six groups according to the dose and timing of inoculations, including I-D0, I-D2-4, I-D7 (day 0, days 2–4, and day 7 after the first dose of ChAdOx1, respectively) and II-D0, II-D1-4, II-D7-10 (day 0, days 1–4, and days 7–10 after the second dose of BNT162b2, respectively). Each sample was represented by the expression levels of 26,364 genes. The first dose was ChAdOx1, whereas the second dose was mainly BNT162b2 (Only four individuals received a second dose of ChAdOx1). The groups were deemed as labels and genes were considered as features. Several machine learning algorithms were employed to analyze such classification problem. In detail, five feature ranking algorithms (Lasso, LightGBM, MCFS, mRMR, and PFI) were first applied to evaluate the importance of each gene feature, resulting in five feature lists. Then, the lists were put into incremental feature selection method with four classification algorithms to extract essential genes, classification rules and build optimal classifiers. The essential genes, namely, *NRF2*, *RPRD1B*, *NEU3*, *SMC5*, and *TPX2*, have been previously associated with immune response. This study also summarized expression rules that describe different vaccination scenarios to help determine the molecular mechanism of vaccine-induced antiviral immunity.

KEYWORDS

SARS-CoV-2, vaccination, immune response, machine learning, blood transcriptome

1. Introduction

Coronavirus disease-19 (COVID-19) is a pandemic infectious disease that is currently affecting many people in approximately 200 countries around the world. It is caused by acute respiratory syndrome coronavirus-2 (SARS-CoV-2), a highly pathogenic coronavirus that belongs to the subfamily Coronaviridae. The SARS-CoV-2 genome contains a variety of structural and nonstructural proteins. The rapid rate at which the virus mutates and spreads has created enormous challenges for prevention and control efforts. Currently, vaccination against SARS-CoV-2 is accepted as an effective strategy against COVID-19 (Folegatti et al., 2020; Amano et al., 2022), with two or more doses giving better protection than one dose alone. The risk of death from COVID-19 varies widely in different countries and may be related to factors such as vaccination rate and number of vaccinations (Masic et al., 2020).

When the body receives the first dose of the COVID-19 vaccine (basic immunization injection), it recognizes viral-specific antigens and produces antibodies and memory cells against SARS-CoV-2. However, the amount of antibodies produced by the primary immune response is much lower than the level required to resist viral invasion. Early clinical trials showed that with just one dose (initial exposure), the body's resistance to SARS-CoV-2 is very low at about 50%. Therefore, a second vaccine dose and a booster shot have been recommended after a period of time (3–4 weeks). When exposed to the same antigen twice, the memory cells that have been generated in the human body respond rapidly, producing sufficient antibodies and a strong secondary immune response. Therefore, two doses of vaccination are more effective for protection. The ChAdOx1 nCoV-19 (AZD1222) vaccine is constructed from a replication-defective simian adenovirus vector encoding the spike (S) protein of SARS-CoV-2. Clinical trials have shown that the ChAdOx1 vaccine is 74% protective against symptomatic COVID-19 (Cross et al., 2003). Meanwhile, BNT162b2, also known as the Pfizer-BioNTech COVID-19 vaccine, is a messenger RNA (mRNA) vaccine that has been approved by the US FDA for the prevention of COVID-19 caused by the SARS-CoV-2 Beta coronavirus. A heterologous ChAdOx1-S-nCoV-19 and BNT162b2 vaccination combination provides better protection against severe SARS-CoV-2 infection in a real-world observational study ($n = 13{,}121$). Studies have shown that T-cell responses following ChAdOx1 vaccination were higher than those elicited by BNT162b2. Meanwhile, T-cell responses elicited by BNT162b2 booster doses were enhanced in different vaccination strategies. Both homologous and heterologous vaccinations were able to induce progressively increased frequencies of CD4 and CD8 T cells. However, the heterologous combination elicited stronger CD4 T-cell responses; CD8 T-cell responses were also progressively stronger after the booster dose (Pozzetto et al., 2021). The tolerability and safety profile of BNT162b2 at 30 μg administered as a 2-dose regimen are favorable. In participants who received only one ChAdOx1 dose, antibodies against the SARS-CoV-2 spike protein peaked at day 28 (median 157 ELISA units [EU]); on day 56, the median was 119 EU. Among participants who received the booster dose, the median antibody at day 56 was 639 EU (Folegatti et al., 2020). Studies have demonstrated the efficacy of a two-dose regimen of the BNT162b2 vaccine (Mizrahi et al., 2021).

An increasing number of studies have confirmed that high-throughput sequencing data information can provide important guidance for revealing the pathogenic mechanism of diseases and tackling various medical problems (Dai et al., 2018; Kong et al., 2020; Yang et al., 2020, 2022). Our team has long been working on using machine learning analysis methods to screen for disease-related signatures and explain their pathogenic mechanisms. We divided the data on 161 people vaccinated against COVID-19 into six groups according to the injection and vaccination time, aiming to further explore changes in blood gene expression after different doses, especially the molecular characteristics of antiviral immunity. A variety of algorithms were used to analyze gene expression information on vaccines from different vaccinations. The algorithms included feature ranking algorithms, such as least absolute shrinkage and selection operator (Lasso) (Tibshirani, 2011), light gradient-boosting machine (LightGBM) (Ke et al., 2017), Monte Carlo feature selection (MCFS) (Draminski et al., 2007), max-relevance and min-redundancy (mRMR) (Peng et al., 2005), and permutation feature importance (PFI) (Fisher et al., 2019), as well as classification algorithms, such as decision tree (DT) (Safavian and Landgrebe, 1991), random forest (RF) (Breiman, 2001), K-nearest neighbor (KNN) (Cover and Hart, 1967), and support vector machine (SVM) (Cortes and Vapnik, 1995). Based on feature ranking algorithms, gene feature lists were obtained, which were subjected to incremental feature selection (IFS) method (Liu and Setiono, 1998), incorporating four classification algorithms, for extracting essential genes, classification rules, and build optimal classifiers. This study revealed that blood gene expression changed after the initial immunization and booster vaccination. A number of important genes (e.g., *NRF2*, *RPRD1B*, *NEU3*, *SMC5*, and *TPX2*) may be closely related to the antiviral immunity induced by vaccines. These findings are helpful for understanding the importance of vaccination and boosting injections by revealing the effects of different injections on the expression of immune-related molecules in the host and by providing a reference for viral immune intervention strategies for COVID-19.

2. Materials and methods

The workflow of the machine learning framework is shown in Figure 1. The samples were grouped according to the number of inoculations and inoculation time. The genes were subsequently ranked using five methods and further processed by IFS method with four classification algorithms. By observing the performance of the classifiers, a number of key genes and summarized quantitative classification rules were identified. Last, the key genes were functionally enriched to determine the biological processes involved in their action. The methods used are described in detail in this section.

2.1. Data

Blood transcriptome data from 161 individuals were obtained from the GEO database under the registration number GSE201533 (Lee et al., 2022a). We divided the vaccinees into two groups: I for the first COVID-19 vaccination dose and II for the second dose. For the first group, three subsets were included: I-D0, I-D2-4, and I-D7, meaning day 0, days 2–4, and day 7 after the first dose of ChAdOx1, respectively. There were also three subsets in the second group, say II-D0, II-D1-4, II-D7-10, meaning the day 0, days 1–4, and days 7–10 after the second dose of BNT162b2, respectively. Four of the vaccinees received a second

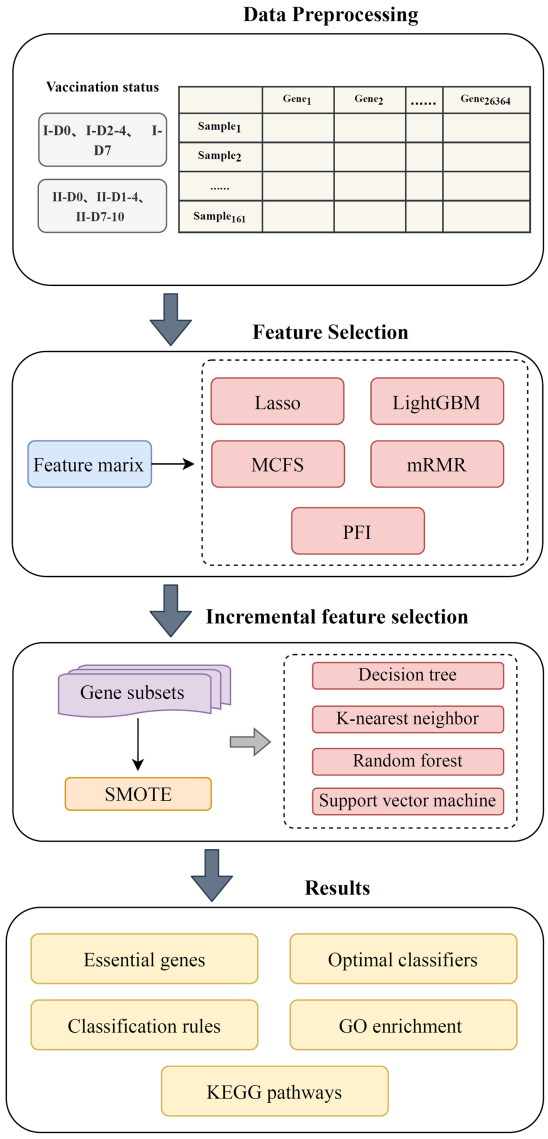

FIGURE 1
Flow chart of the entire analysis process. The blood transcriptome data of 161 vaccinees with different COVID-19 vaccination was investigated. Each vaccinee was represented by 26,364 gene expression levels. Five feature ranking algorithms (Lasso, LightGBM, MCFS, mRMR, and PFI) were used to rank gene features according to their importance. Subsequently, these lists were fed into incremental feature selection method, which contained four classification algorithms, to extract essential genes, classification rules, and build optimal classifiers.

dose of ChAdOx1. Table 1 shows the number of samples in each subset. Each sample was represented by 26,364 gene expression levels, which were deemed as features in this study. The six subsets (I-D0, I-D2-4, I-D7, II-D0, II-D1-4, and II-D7-10) were termed as labels. The current study was conducted by deeply investigating such classification problem.

2.2. Feature ranking algorithms

Lots of features were used to represent each sample. Evidently, some were important and others were useless. It was necessary to extract important features. To date, several feature analysis methods have been proposed, which can evaluate the importance of features. The selection of such method is a challenge problem as each method has its own merits and defects. Generally, one method can only output a part of essential features. Thus, it was beneficial to employ multiple methods, thereby providing a more complete picture on essential features. Here, five algorithms, namely, Lasso (Tibshirani, 2011), LightGBM (Ke et al., 2017), MCFS (Dramiński et al., 2007), mRMR (Peng et al., 2005), and PFI (Fisher et al., 2019), were employed to rank genes according to their importance. These algorithms have been frequently applied to solve many life science problems (Zhao et al., 2018; Ren et al., 2022; Li et al., 2022a,b,c; Huang et al., 2023a,b).

2.2.1. Least absolute shrinkage and selection operator

Based on the nonnegative garrote proposed by Breiman (1995), Robert Tibshirani first proposed the Lasso algorithm in 1996 (Tibshirani, 2011). The algorithm proposes a first-order penalty function containing regularized formulas, where each feature is regarded as an independent variable in the function. The coefficients of the features are then obtained by solving the optimization function. The absolute value of a coefficient indicates the degree of correlation of each feature to the target dependent variable. To achieve data compression and reduce overfitting, the algorithm regularizes the coefficients of some variables while setting some to zero to eliminate the features that tend to contribute less to the follow-up prediction. Accordingly, the algorithm can rank features according to the absolute values of their coefficients. In present study, the Lasso program in Scikit-learn (Pedregosa et al., 2011) was adopted, which was executed using default parameters.

2.2.2. Light gradient-boosting machine

LightGBM (Ke et al., 2017) is based on the gradient-boosting decision tree framework and introduces gradient one-sided sampling, exclusive feature bundling, histogram algorithm, and leaf-wise growth strategy. It enables data slicing, bundling, and dimensionality reduction and ultimately reduces computational cost while improving prediction accuracy. The importance of each feature is determined by the number of trees that the feature participates in building: the higher the participation, the higher the importance. Thus, features can be ranked in a list with decreasing order of this number. The current study used the LightGBM program obtained from.[1] For convenience, it was performed using default parameters.

2.2.3. Monte Carlo feature selection

Monte Carlo feature selection was originally developed by Dramiński et al. (2007). The algorithm selects some features randomly and repeatedly to obtain p feature subsets. Each feature subset is then divided into a training set and a test set t times, and t trees are constructed. Thus, $p \times t$ trees are obtained. The importance of features can be evaluated by their contributions to building these trees and is defined as the relative importance (RI) score, which is calculated as follows:

1 https://lightgbm.readthedocs.io/en/latest/

TABLE 1 Sample sizes of six vaccination status.

Index	Vaccination status		Sample size
1	I-D0	(Day 0 after the first dose)	37
2	I-D2-4	(Day2 2–4 after the first dose)	36
3	I-D7	(Day 7 after the first dose)	37
4	II-D0	(Day 0 after the second dose)	17
5	II-D1-4	(Days 1–4 after the second dose)	18
6	II-D7-10	(Days 7–10 after the second dose)	16

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

where ωA_{CC} is the weighted precision of the tree τ under consideration, $ng\left(\tau\right)$ is a node of the tree whose information gain is denoted as $IG\left(ng\left(\tau\right)\right)$, and $no.in\ ng\left(\tau\right)\left(no.in\ \tau\right)$ denotes the sample size of $ng\left(\tau\right)\left(\tau\right)$. u and v are two positive numbers weighting the ωA_{CC} and the ratio $no.in\ ng\left(\tau\right)/no.in\ \tau$, respectively. To execute MCFS, we downloaded its program from.[2] Default parameters were used.

2.2.4. Max-relevance and Min-redundancy

The mRMR method was proposed by Peng et al. (2005) in 2005. It screens features based on their correlation with the target variable and the redundancy between features. The correlation and redundancy can be calculated from the mutual information between features or target variables. The tradeoff of correlation and redundancy is used to evaluate the importance of features. At each round, one feature with the maximum correlation to target variables and minimum redundancy to features in the current list is selected and appended to the current list. Here, we used the mRMR program sourced from.[3] It was executed with default parameters.

2.2.5. Permutation feature importance

The PFI for RFs was first introduced in 2001 by Breiman (2001) and was later extended to any fitted estimator for features by Fisher et al. (2019). The idea is relatively simple. If a feature is important, the prediction error will further increase after the feature's values are shuffled. If a feature is not important, shuffling its values does not increase the prediction error. The PFI program used in this study was retrieved from scikit-learn (Pedregosa et al., 2011), which was executed with default parameters.

Above five algorithms were applied to the blood transcriptome data one by one. Each algorithm produced one feature list. For easy descriptions, the generated lists were called Lasso, LightGBM, MCFS, mRMR and PFI feature lists.

2.3. Incremental feature selection

When the feature list contains an excessive number of features, it is not suitable for direct use in building prediction models. In this study, the IFS (Liu and Setiono, 1998) method was used to extract the best subset of features. From the feature list, a series of feature subsets can be constructed. Each subset includes 10 more features than the previous subset in the order of the list. These feature subsets were then fed to one classification algorithm to build the classifier. The performance of these classifiers was evaluated by 10-fold cross-validation. Lastly, the best classifier can be obtained, which was termed as the optimal classifier. The feature subset for constructing this classifier was called the optimal feature subset.

2.4. Synthetic minority oversampling technique

According to Table 1, some classes (e.g., I-D0) contained much more samples than other classes (e.g., II-D7-10). The dataset was imbalanced. The results of the classifier would have preferences for the majority class when the number of samples from different categories differs significantly. This study used synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) to balance the dataset. For each class with a small number of samples, a sample is random chosen. Then its k nearest neighbors in the same class are identified by Euclidean distance. A neighbor is randomly selected. A new sample is then randomly generated by linearly interpolating the randomly chosen sample and the selected nearest neighbor. New samples are continuously generated until such class contains samples as many as those in the largest class. The SMOTE package reported in[4] was used in this study. Default settings were adopted.

2.5. Classification algorithms for building classifiers

Four classification algorithms were used in the IFS approach. Key genes were then screened based on the performance of the constructed classifiers.

2 http://www.ipipan.eu/staff/m.draminski/mcfs.html

3 http://home.penglab.com/proj/mRMR/

4 https://github.com/scikit-learn-contrib/imbalanced-learn

2.5.1. Decision tree

The DT algorithm (Safavian and Landgrebe, 1991) constructs a tree-like structure in which instances are judged in each internal node of the tree. Starting from the root node, all samples are assigned to different classes through continuous judgments. Each tree branch contains clues to the classification of instances and thus provides interpretable classification rules that underlie the understanding of biological mechanisms. In this study, we used the CART classification tree algorithm with node ranking using the Gini coefficient.

2.5.2. Random forest

In the RF algorithm for classification, a judgment is completed by constructing DTs based on different training sets and then combining their results to make predictions (Breiman, 2001; Wang et al., 2021; Ran et al., 2022; Tang and Chen, 2022; Wu and Chen, 2023). The training set with the same number of samples in the input dataset is repeatedly sampled to generate numerous new training sets. Each new training set is then used to build a new DT, and an ensemble of DTs is constructed. Given a new instance, each DT makes a prediction. Predictions taken from all DTs are combined to reach a final decision.

2.5.3. K-nearest neighbor

In KNN (Cover and Hart, 1967), new samples are predicted by comparing each with samples with known labels (training samples) and determining the k-nearest neighbors. Subsequently, the class of a new sample is determined by voting according to the classes of the k-nearest neighbors. In this study, the distance was defined as the Minkowski distance.

2.5.4. Support vector machine

The SVM algorithm (Cortes and Vapnik, 1995; Wang and Chen, 2022; Wang and Chen, 2023) utilizes a kernel function that maps the attributes of the instances, i.e., the feature vectors, into a higher-dimensional space and attempts to find a separating hyperplane. This hyperplane partitions the instances by class and ensures that the margin between the two categories is maximum. This method is generally to have good generalization.

We adopted public packages in scikit-learn (Pedregosa et al., 2011) to implement above four classification algorithms. All packages were performed using default parameters.

2.6. Performance evaluation

In the multi-class classification problem, weighted F1 is an important measurement to evaluate the performance of the classifier. It is obtained by calculating and integrating the F1-measure values of different classes based on the proportion of the samples in each class. It is known that F1-measure is an integrated measurement combining precision and recall, which can be computed by

$$\mathrm{Pr}ecision_i = \frac{TP_i}{TP_i + FP_i}, \tag{2}$$

$$\mathrm{Re}call_i = \frac{TP_i}{TP_i + FN_i}, \tag{3}$$

$$F1-measure_i = \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}, \tag{4}$$

where i represents the index of class, TP represents true positive, FP represents false positive, and FN represents false negative. Then, weighted F1 can be calculated by

$$Weighted\ F1 = \sum_{i=1}^{L} w_i \times F1-measure_i, \tag{5}$$

where L represents the number of classes and w_i represents the proportion of samples in the i-th class to overall samples. Here, weighted F1 was selected as the major measurement.

In addition, overall accuracy (ACC) and Matthew correlation coefficient (MCC) (Matthews, 1975) are also widely used to assess the quality of classifiers. ACC is defined as the proportion of correctly predicted samples to all samples. MCC is a balanced measurement, which is more objective than ACC when the dataset is imbalanced. For the calculation of MCC, two matrices X and Y must be constructed first, which store the one-hot representation of true and predicted class of each sample. Then, MCC can be computed by

$$MCC = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}} \tag{6}$$

where $\mathrm{cov}(X,Y)$ denotes the correlation coefficient of X and Y.

2.7. Functional enrichment analysis

Using the IFS method, we can obtain the best subset of features under different rankings. To clarify the biological processes behind genes in these subsets, thereby uncovering their relationship with antiviral immunity, this study used gene ontology (GO) enrichment analysis to discover the role of the genes and applied Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis to identify the underlying pathways. ClusterProfiler package (Wu et al., 2021) in R was used to perform GO and KEGG enrichment analyses.

3. Results

3.1. Results of feature ranking

To evaluate the importance of features from multiple aspects. Five feature ranking algorithms were employed, which were applied to the blood transcriptome data one by one. As a result, five feature lists, named Lasso, LightGBM, MCFS, mRMR and PFI feature lists, were obtained, which are provided in Supplementary Table S1. Table 2 shows the top 10 genes in each list. It can be observed that top genes in different lists were very different, meaning that the

TABLE 2 The top 10 features in five feature lists.

Index	Lasso feature list	LightGBM feature list	MCFS feature list	mRMR feature list	PFI feature list
1	CENPF	RPRD1B	NEU3	FAM98B	SLC16A14
2	NDUFB9	ITM2C	C2	TSSK4	THRAP3
3	BRCA2	HSP90B1	SMC5	CSF1R	STAC3
4	LOC102031319	TK1	ZFC3H1	TOP1	ATF5
5	SSBP1	LPAR3	GLS2	NEU3	RAD51
6	PDP1	CENPF	NFE2L2	UBE2H	CDC45
7	LINC01089	TPX2	C1QC	ATP6V1E1	GABPB1
8	C2orf16	ITGAE	SDC1	SRPRB	CTNNBL1
9	ID2	SPATA24	CAV1	ZNF672	ARHGAP42
10	LINC00630	GTSE1	SNORA2B	CUL3	PSME2

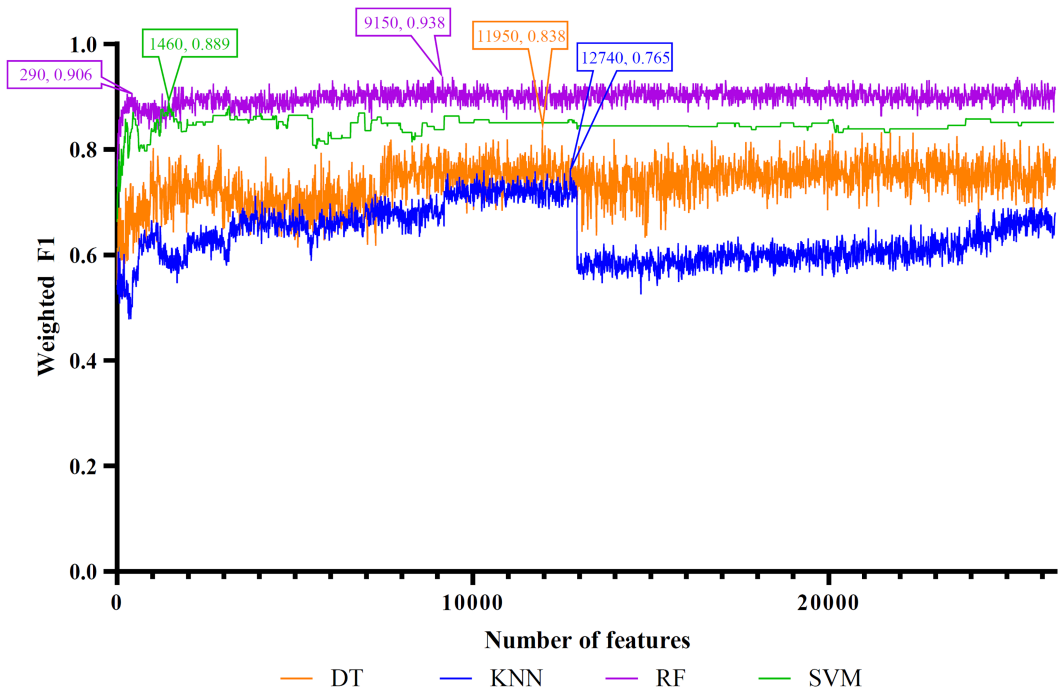

FIGURE 2
IFS curves of four classification algorithms on Lasso feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.838, 0.765, 0.938, and 0.889 when top 11,950, 12,740, 9,150, and 1,460 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.906) when top 290 features were used.

importance of one feature was quite different under the evaluation of different methods. Usage of different methods can provide more opportunities to discover more essential features.

3.2. Results of incremental feature selection

Five feature lists were subjected to the IFS method one by one. From each feature list, a series of feature subsets with step ten were constructed. On each subset, one classifier was built for each of four classification algorithms (DT, KNN, RF, and SVM). When constructing the classifiers, the dataset was processed by SMOTE to tackle the imbalanced problem. All classifiers were evaluated by 10-fold cross-validation. The evaluation results were counted as weighted F1, ACC, and MCC, which are provided in Supplementary Table S2. Weighted F1was selected as the major measurement. Thus, several IFS curves were plotted for different classification algorithms and feature lists, as shown in Figures 2–6, in which weighted F1 was set as Y-axis and number of features was defined as *X*-axis.

For the Lasso feature list, the IFS curves of four classification algorithms are illustrated in Figure 2. It can be observed that when top 11,950, 12,740, 9,150 and 1,460 features were adopted,

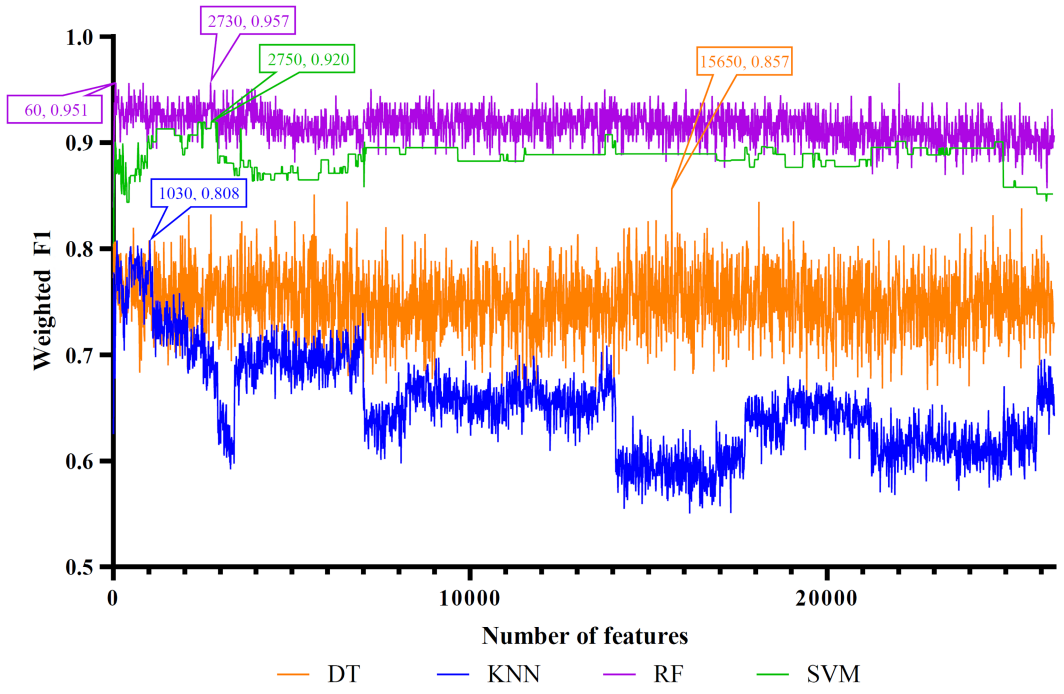

FIGURE 3
IFS curves of four classification algorithms on LightGBM feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.857, 0.808, 0.957, and 0.920 when top 15,650, 1,030, 2,730, and 2,750 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.951) when top 60 features were used.

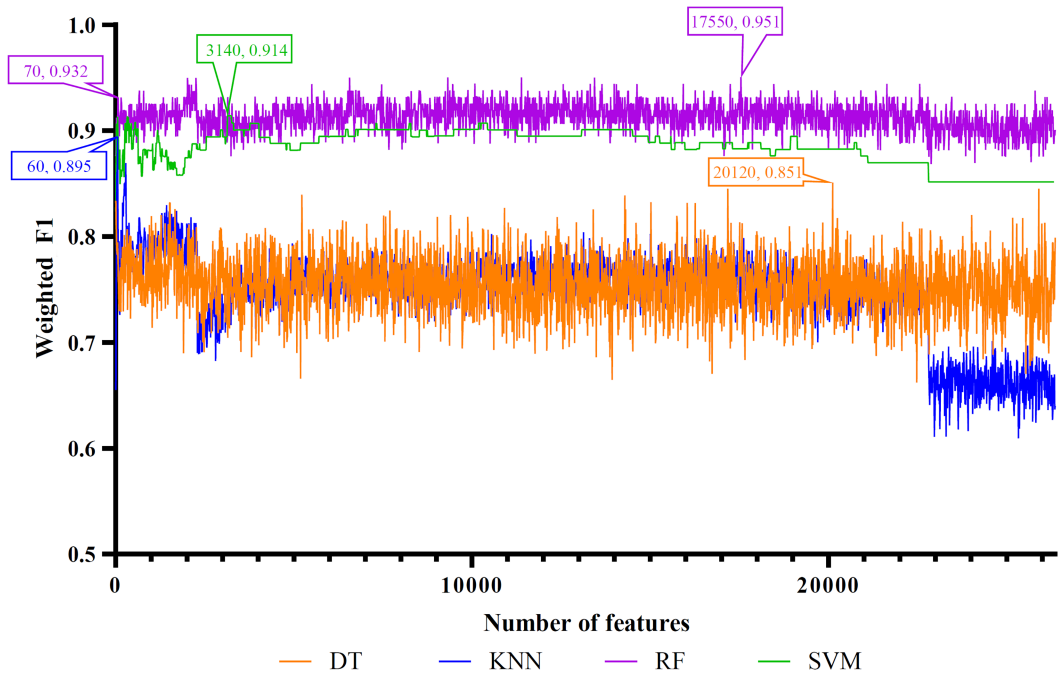

FIGURE 4
IFS curves of four classification algorithms on MCFS feature list. DT, KNN, RF and SVM yielded the highest weighted F1 values of 0.851, 0.895, 0.951, and 0.914 when top 20,120, 60, 17,550, and 3,140 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.932) when top 70 features were used.

four algorithms yielded the highest weighted F1 values of 0.838, 0.765, 0.938, and 0.889, respectively. Thus, the optimal DT, KNN, RF, and SVM classifiers can be built using these features. The ACC and MCC values of these classifiers are listed in Table 3. Evidently, the optimal RF classifier was best among these optimal classifiers.

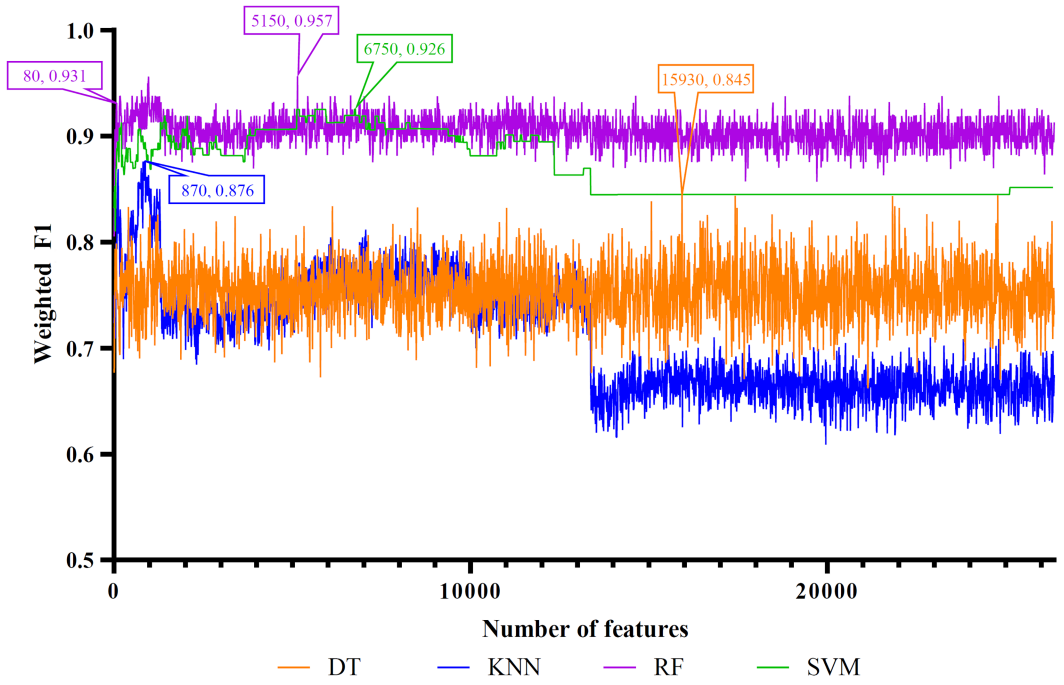

FIGURE 5
IFS curves of four classification algorithms on mRMR feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.845, 0.876, 0.957, and 0.926 when top 15,930, 870, 5,150, and 6,750 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.931) when top 80 features were used.

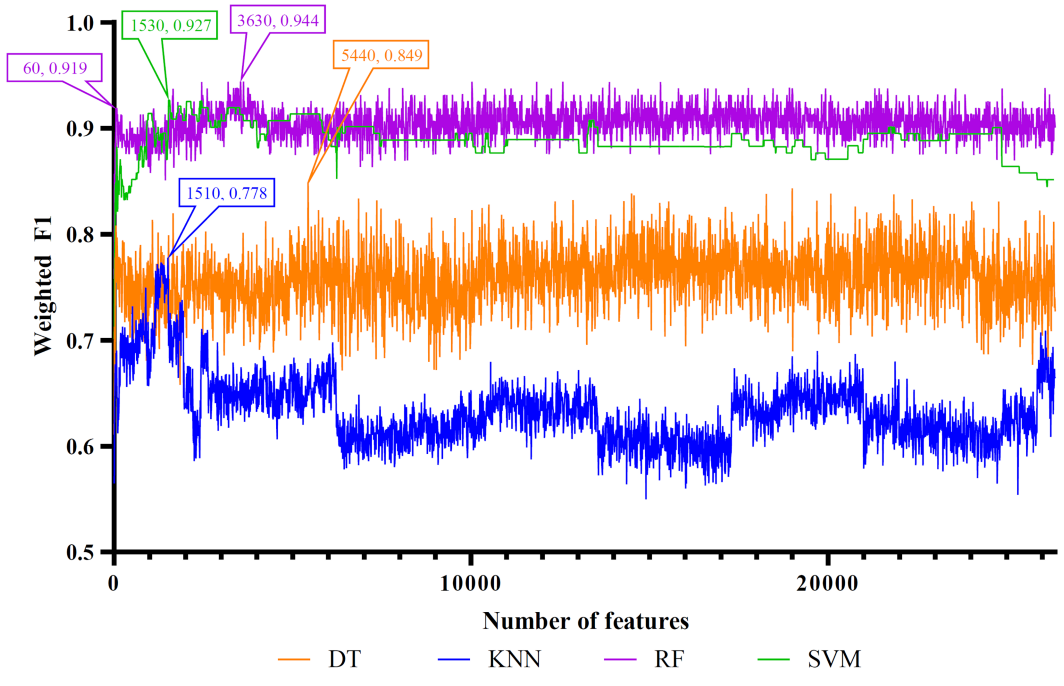

FIGURE 6
IFS curves of four classification algorithms on PFI feature list. DT, KNN, RF, and SVM yielded the highest weighted F1 values of 0.849, 0.778, 0.944, and 0.927 when top 5,440, 1,510, 3,630, and 1,530 features were adopted, respectively. RF can yield quite high performance (weighted F1=0.919) when top 60 features were used.

For the LightGBM feature list, Figure 3 shows the IFS curves of four classification algorithms. The optimal DT/KNN/RF/SVM classifier can be built using top 15,650/1030/2730/2750 features in this list. Their ACC, MCC, and weighted F1 values are listed in Table 3. Clearly, RF still provided the best performance as the optimal RF classifier yielded the highest weighted F1 of 0.957.

TABLE 3 Performance of the optimal classifiers based on different classification algorithms and feature lists.

Feature list	Classification algorithm	Number of features	Weighted F1	MCC	ACC
Lasso feature list	Decision tree	11,950	0.838	0.801	0.839
	K-nearest neighbor	12,740	0.765	0.722	0.770
	Random forest	9,150	0.938	0.924	0.938
	Support vector machine	1,460	0.889	0.863	0.888
LightGBM feature list	Decision tree	15,650	0.857	0.825	0.857
	K-nearest neighbor	1,030	0.808	0.764	0.807
	Random forest	2,730	0.957	0.947	0.957
	Support vector machine	2,750	0.920	0.901	0.919
MCFS feature list	Decision tree	20,120	0.851	0.817	0.851
	K-nearest neighbor	60	0.895	0.870	0.894
	Random forest	17,550	0.951	0.939	0.950
	Support vector machine	3,140	0.914	0.894	0.913
mRMR feature list	Decision tree	15,930	0.845	0.809	0.845
	K-nearest neighbor	870	0.876	0.847	0.876
	Random forest	5,150	0.957	0.947	0.957
	Support vector machine	6,750	0.926	0.908	0.925
PFI feature list	Decision tree	5,440	0.849	0.817	0.851
	K-nearest neighbor	1,510	0.778	0.734	0.783
	Random forest	3,630	0.944	0.932	0.944
	Support vector machine	1,530	0.927	0.909	0.925

TABLE 4 Performance of feasible classifiers on different feature list.

Feature list	Classification algorithm	Number of features	Weighted F1	MCC	ACC
Lasso feature list	Random forest	290	0.906	0.886	0.907
LightGBM feature list	Random forest	60	0.951	0.939	0.950
MCFS feature list	Random forest	70	0.932	0.916	0.932
mRMR feature list	Random forest	80	0.931	0.916	0.932
PFI feature list	Random forest	60	0.919	0.901	0.919

As for the rest three feature lists, the IFS curves are shown in Figures 4–6. The optimal DT/KNN/RF/SVM classifier can be set up on each feature list. The numbers of top features used in these classifiers are listed in Table 3, where the performance of these classifiers is also provided. Similar to the results on the Lasso and LightGBM feature lists, the optimal RF classifier was also better than other three optimal classifiers on each feature list.

To make full use of the utility of five algorithms, the best features should be extracted from each feature list, thereby obtaining the latent essential gene features. As mentioned above, the optimal RF classifier was best for each feature list. Thus, the features used in these classifiers can be picked up as important candidates. However, such feature numbers (9,150 for Lasso feature list, 2,730 for LightGBM feature list, 17,750 for MCFS feature list, 5,150 for mRMR feature list, 3,630 for PFI feature list) were too large to make detailed analyses. In view of this, we tried to find out another RF classifier, which adopted much less features and provided a little lower performance than the optimal RF classifier, on each feature list. By carefully checking the IFS results on RF on each feature list, such RF classifiers adopted the top 290 features in the Lasso feature list, top 60 features in the LightGBM feature list, top 70 features in the MCFS feature list, top 80 features in the mRMR feature list, and top 60 features in the PFI feature list. The corresponding points have been marked on the IFS curves of RF, as illustrated in Figures 2–6. The detailed performance of these RF classifiers is listed in Table 4. It can be observed that their performance was still quite high, the weighted F1 values were all higher than 0.900. Compared with the weighted F1 yielded by the optimal RF classifier on the same feature list, this RF classifier provided a little lower weighted F1. However, their efficiencies were sharply improved because much less features were involved. This indicated the extreme importance of features used in these RF classifiers. For easy descriptions, these RF classifiers were called feasible RF classifiers. Furthermore, the performance of the feasible RF classifier on one feature list was generally better than the optimal DT/KNN/SVM classifier on the same feature list, further confirming the importance of features in the feasible RF classifiers. To clear show the relationship between the feature sets used in five feasible RF classifiers, a Venn diagram was plotted, as shown in Figure 7. The detailed results of the intersection are shown in Supplementary Table S3.

10.3389/fmicb.2023.1138674

Some gene features occurred in multiple subsets, meaning that they were deemed to be important by multiple feature ranking algorithms. They may have strong associations with antiviral immunity. Some of them would be discussed in detail in the subsequent sections.

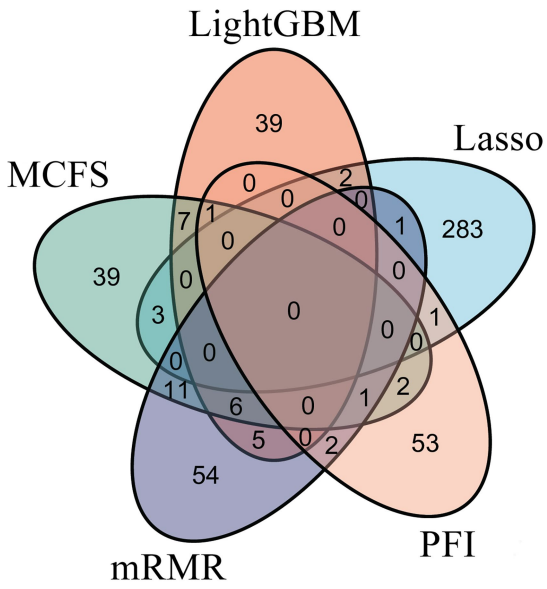

FIGURE 7
Venn diagram of the feature sets used to construct feasible random forest classifiers on five feature lists that were obtained by Lasso, LightGBM, MCFS, mRMR, and PFI, respectively. The overlapping circles indicated genes that occurred in multiple sets. These genes were deemed to be important by multiple feature ranking algorithms.

3.3. Classification rules

Although the performance of DT was much lower than RF and SVM according to the IFS results on five feature lists, DT has an exclusive merit as it is a white-box algorithm. It can provide quantitative rules that can be interpreted to aid in the analysis. On the Lasso, LightGBM, MCFS, mRMR, and RF feature lists, the optimal DT classifier adopted the first 11,950, 15,650, 20,120, 15,930, and 5,440 gene features. Based on the samples represented by these features, five trees were obtained, from which five groups of classification rules can be extracted. Supplementary Table S4 shows these classification rule groups. Some conditions in major rules would be discussed in detail later.

3.4. Enrichment analysis

Five feature sets used to construct five feasible RF classifiers were combined into one set. To uncover the underlying biological meanings behind gene features in such set, the enrichment analysis was conducted on these genes. Figure 8 visualizes top five GO terms in three GO clusters and top five pathways. The GO terms, such as thioester and fatty acid metabolic processes, were enriched, along with peroxisomes and some terms related to metabolism and transport. KEGG enriched pathways included fatty acid biosynthesis, catabolism, and metabolism. Thioesters can be directly involved in the immune response as carriers of antigen presentation and thioesterified fatty acids or other lipid products can be involved in the regulation of immune cells as signaling molecules. Their metabolism is inseparable from the peroxisome.

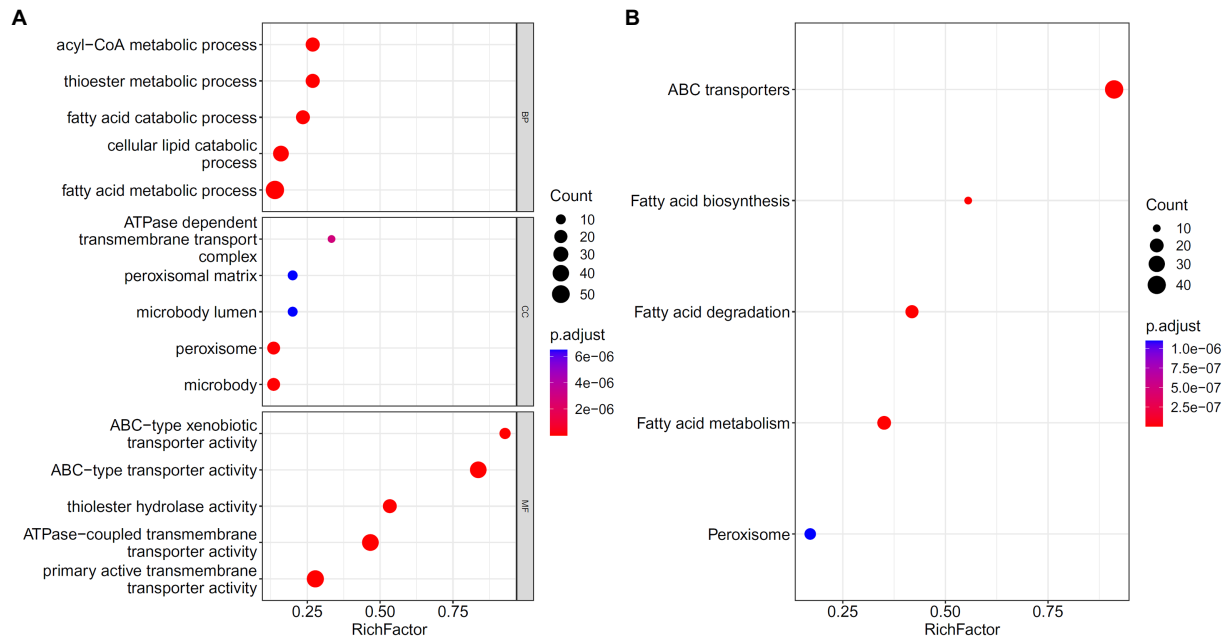

FIGURE 8
Gene ontology (GO) and KEGG pathway enrichment analysis on the union of five feature sets used to construct feasible random forest classifiers. The FDR<0.05 criterion was used to filter GO terms and KEGG pathways. The top five significant GO terms in three GO clusters **(A)** and top five KEGG pathways **(B)** were shown.

4. Discussion

As listed in "Results", some essential genes and classification rules were discovered. As they can be strongly related to the response to vaccination in antitumor viral immunity, they were discussed in this section. We collected the scientific findings of other researchers and initially summarized the experimental evidence of the aforementioned genes and rules, proving the accuracy of the findings.

4.1. Analysis of essential conditions in rules

Five rule groups were discovered as listed in Supplementary Table S4. As each rule contained multiple gene features and thresholds on expression levels, it was not easy to confirm the special pattern expressed by each rule through existing publications. Thus, we divided each rule into multiple conditions and analyzed the reasonability of some essential conditions. If the conditions used the same gene and same expression trend, they were deemed to be identical. The occurrence number of each condition in five rule groups was counted, which represented how many feature ranking methods identified the condition to be important. Some representative conditions with such numbers larger than two were discussed.

4.1.1. Analysis of conditions identified *via* four methods

IFI27 occurred in four rule groups, including rule groups on Lasso, LightGBM, mRMR, and MCFS feature lists. The study found that the expression levels of antiviral-related genes such as *IFI27* decreased during the vaccinations. This result is consistent with the dynamically enhanced inflammatory response in vaccinated individuals. *IFI27* is considered a biomarker with high sensitivity and specificity (AUC > 0.85) (Wang et al., 2022). Vaccination can improve the body's ability to fight viruses. Our analysis results show that the expression level of *IFI27* gradually increased within 2–4 days of the first injection and decreased 7 days after vaccination. However, after the second injection, the expression level of *IFI27* gradually increased within 1–4 days after the injection. Compared with the first injection, some patients had the fastest response times earlier than the first injection. The expression level of *IFI27* decreased 7–10 days after vaccination. The peak duration of the second injection is speculated to be longer than that of the first injection. The antiviral immune-related molecular mechanism of IFI27 has been reported. As a common interferon (IFN)-stimulated gene, *IFI27* encodes a mitochondrial protein that is normally induced by IFN to express and function in most responding cells. It may regulate apoptosis through the stability of mitochondrial membrane, thereby affecting immune response (Cheriyath et al., 2011). In addition, IFI27 can inhibit viral DNA replication and gene expression (Ullah et al., 2021). *In vitro* studies have shown that IFI27 is up-regulated in plasmacytoid dendritic cells, which are antigen-presenting cells sensitive to viral infection (Tang et al., 2017). Transcriptome results showed that vaccinated patients had significantly attenuated IFN responses compared to unvaccinated Omicron and Alpha-infected patients, represented by IFI27, which controls antiviral responses (Lee et al., 2022b). The results of RNA sequencing data analysis showed that macrophages in the blood of SARS-CoV-2-infected patients released a large number of IFNs, activated mitochondrial *IFI27* expression, and disrupted energy metabolism in immune cells, ultimately aggravating viral immune evasion and replication (Duan et al., 2022). Based on existing research reports and our analysis, we speculate that after vaccination, the release of IFN increases, which promotes an increase in mitochondrial protein IFI27, inhibits SARS-CoV-2 replication and gene expression, and enhances antiviral immunity. In addition, after two vaccine doses, some people's antiviral immunity takes effect earlier than after the first dose, and vaccine efficacy lasts longer. Therefore, *IFI27* may be used as a biomarker for antiviral immunity of vaccines.

4.1.2. Analysis of conditions identified *via* three methods

Syndecan-1 (*SDC1*) and *small nuclear ribonucleoprotein polypeptide G* (*SNRPG*) were found in rule groups on LightGBM, mRMR, and MCFS feature lists. *SDC1* encodes a transmembrane (type I) heparan sulfate proteoglycan protein that belongs to the syndecan proteoglycan family. As a component of glycocalyx (GAC), SDC1 plays an important role in cell proliferation, cell migration, and other processes through extracellular matrix protein receptors (Reszegi et al., 2022). SDC1 was found to be elevated in COVID-19 patients (Goonewardena et al., 2021). SDC1 may contribute to early risk stratification of staged diseases such as COVID-19 and provide a pathobiological reference (Goonewardena et al., 2021). Studies have confirmed that patients infected with COVID-19 can produce inflammation-induced degradation of the GAC layer of endothelial cells, and SDC1 can be used as an important parameter to assess GAC damage (Vollenberg et al., 2021). High levels of SDC1 may cause more severe endothelial damage and inflammation (Zhang et al., 2021). Molecular experiments demonstrate that *SDC1* acts as a target gene of miR-10a-5p during porcine hemagglutinating encephalomyelitis virus (PHEV) infection and is involved in host defense mechanisms. Decreased expression levels of *SDC1* lead to reduced viral replication, and downstream inhibition of *SDC1* exerts an antiviral effect in PHEV-induced disease (Hu et al., 2020). Transcriptome analysis showed that the expression level of *SDC1* increased only 7 days after the first dose of vaccination. After the second dose, the expression level remained low. On the one hand, this low level may help prevent endothelial damage and severe inflammatory response. On the other hand, it may inhibit viral replication and facilitate a more efficient antibody production.

SNRPG is a protein-coding gene involved in the formation of the U1, U2, U4, and U5 small nuclear ribonucleoprotein complexes. Related pathways include SARS-CoV-2 infection and gene expression.[5] Studies have shown that SNRPG-related risk models are associated with infiltration of immune cells such as T cells and M2 macrophages (Liu et al., 2022). The specific mechanism between *SNRPG* and SARS-CoV-2 infection is limited. Transcriptome analysis showed that the *SNRPG* expression level was high on the day of the first vaccine injection, whereas the expression level was lower on the day of the second vaccine injection. The low *SNRPG* level continued until day 10 after vaccination. The obvious differences in *SNRPG* levels after different injections suggest that the gene can be regarded as an indicator of the effectiveness of vaccination. However, the molecular mechanism needs to be further explored.

5 https://pathcards.genecards.org/Card/sars-cov-2_infection?queryString=SNRPG

4.1.3. Analysis of conditions identified *via* two methods

Rules found in two methods included *TPX2, CCDC28A, FAM227B, PKN2-AS1, NEK2, USP46, C22orf15, SLC20A1, TMSB15A, C2,* and *ZFC3H1*. Some of these genes are associated with antiviral immunity. For example, *TPX2* (microtubule nucleation factor) is a gene whose encoded product is involved in the activation of protein kinase activity, DNA damage, gene transcription, and other physiological processes. PPI network analysis from STRING revealed that as a hub gene, *TPX2* may be a novel COVID-19 intervention target and biomarker (Hasan et al., 2022). As one of the antigen components of a multivalent recombinant fusion protein prophylactic vaccine (rBmHAXT), *TPX2* can promote the production of high titers of antigen-specific antibodies and their isotypes. Animals vaccinated with the *TPX2* antigen secreted higher levels of blood IFN-γ and showed better immune protection compared with unvaccinated animals (Khatri et al., 2018). Studies have shown that *TPX2* can activate Aurora A kinase (AURKA), which is involved in cell cycle regulation. *TPX2* overexpression enhanced cell proliferation and migration (Zou et al., 2018). The *TPX2* gene may be a potential target for diagnosis and prognosis in patients already infected with hepatitis B virus (HVB) (Ji et al., 2020). Transcriptome data analysis showed that *TPX2* expression levels increased within 7–10 days after the patients received the second vaccine dose. This is consistent with activation of IFN-induced responses, increased transcripts of specific IGHV clones, and a trend toward memory B cell enrichment (Lee et al., 2022a). *TPX2* may be related to antiviral immunity caused by different doses. However, the correlation and mechanism of action need to be further verified.

4.2. Top features identified *via* multiple methods

On the basis of the features identified by the five feature ranking algorithms (Figure 7), an intersection of results obtained by multiple methods (≥3) was selected as important candidates. We summarized the evidence for some vital gene features, listed in Table 5, based on the broad studies shown below.

NFE2-like bZip transcription factor 2 (*NRF2*), also called *NFE2L2*, encodes a cap'n'collar (CNC) transcription factor and belongs to the small family of basic leucine zipper (bZIP) proteins (Khan et al., 2021). *NRF2* can bind to antioxidant response elements and participate in the transcription of downstream target genes. Thus, it plays an important role in physiological processes such as cellular redox, tissue damage, and metabolic homeostasis. The encoded protein of *NRF2* is involved in various injury and inflammatory responses involving class I MHC-mediated antigen presentation and KEAP1-NFE2L2 pathway, among others. *NRF2* contributes to GSH metabolism and stress response and is associated with the pro-inflammatory effects of SARS-CoV-2 in host cells (Galli et al., 2022). The protein synthesis of SARS-CoV-2 may increase Cys and activate endoplasmic reticulum stress of transcription factors, which ultimately promotes changes in cellular oxidation, cellular metabolism, and GSH transmembrane flux (Galli et al., 2022). Importantly, *NRF2* activation has been shown to benefit respiratory infections in various animal models (Muchtaridi et al., 2022). *NRF2* exerts anti-inflammatory effects by inhibiting pro-inflammatory genes such as *IL6* and *IL1B* (Huang et al., 2022). *NRF2* induces the expression of genes that promote specificity of macrophages such as the macrophage receptor, which is responsible for bacterial phagocytosis (Schaefer et al., 2022), and the cluster of differentiation gene 36 (CD36), which resists viral infection (Hillier et al., 2022). *NRF2* Activation is involved in inflammatory cascade (Jayakumar et al., 2022), regulation of innate immune responses, and antiviral cytosolic DNA sensing. *NRF2* inhibits pro-inflammatory signaling pathways such as TNF-α signaling and is involved in regulating the innate immune response during sepsis. *NRF2* increases susceptibility to DNA virus infection by inhibiting the expression of the adaptor protein STING1, thereby inhibiting antiviral cytosolic DNA sensing (Olagnier et al., 2018). After SARS-CoV-2 infection, *NRF2* is activated and restricts the release of pro-inflammatory cytokines by inhibiting IRF3 dimerization. In addition, *NRF2* inhibits the replication of SARS-CoV-2 and other viruses through a type I IFN-independent pathway (Olagnier et al., 2020).

Regulation of nuclear pre-mRNA domain containing 1B (*RPRD1B*), also named cell-cycle-related and expression-elevated protein in tumor (*CREPT*) or *C20ORF77*, is located on chromosome 20q11 and can bind to RNA polymerase on the cyclin D1 gene, resulting in the formation of a cyclin D1 ring structure, which can promote transcription (Lu et al., 2012; Wang et al., 2014). *RPRD1B* can also participate in the transcription of genes related to the Wnt/β-catenin signaling pathway (Wu et al., 2010). GO annotation results showed that *RPRD1B* can bind to the RNA polymerase II complex and play a role in pathways such as TCR signaling and T-cell activation. The mRNA and protein expression of *RPRD1*B in patients under 50 years old were significantly different from those in patients over 50 years of age. *RPRD1B* expression levels correlate with human papillomavirus infection and may be affected by age (Wen et al., 2021). The expression level of *RPRD1B* in peripheral blood T cells of psoriasis, lichen planus (LP), and atopic dermatitis (AD) was found higher than that of healthy subjects. *RPRD1B* is involved in the pathogenesis of inflammatory diseases by regulating the transcription of genes such as *IL-4*, *RGS16*, and *CD30* (Li et al., 2013). Our analysis showed that the *RPRD1B* expression level changed in patients who received different vaccinations. Combined with existing evidence, we speculate that *RPRD1B* uses T cells as a carrier to play a role in antiviral immunity.

Neuraminidase 3 (*NEU3*) is a protein-encoding gene whose product is located in the plasma membrane and belongs to the glycohydrolase family. Its activity is specific to gangliosides and may be involved in gangliosides in lipid bilayer adjustment. Pathways associated with *NEU3* include protein metabolism and glycosphingolipid metabolism. It can directly interact with signaling receptors such as EGFR to regulate transmembrane signaling (Wada et al., 2007; Mozzi et al., 2015). Sialidase activity in human polymorphonuclear leukocytes plays a key role in infection and inflammatory responses (Cross et al., 2003; Sakarya et al., 2004). Sialidase activity is determined by membrane-associated sialidase (*NEU3*), which promotes cell adhesion and cell proliferation.

TABLE 5 Essential genes identified by three feature ranking algorithms.

Index	Gene symbol	Description
1	RPRD1B	Regulation of nuclear pre-mRNA domain containing 1B
2	NFE2L2	NFE2-like bZip transcription factor 2
3	SMC5	Structural maintenance of chromosome 5
4	NEU3	Neuraminidase 3

Combined with existing evidence, our results indicate that after vaccination, the body produces antibodies against SARS-CoV-2 that regulate the host immune response by affecting the activity of *NEU3*.

The encoded product of structural maintenance of chromosome 5 (*SMC5*) has ATP-binding activity and is involved in physiological processes such as DNA recombination, cellular senescence, protein metabolism, and transport of mature mRNAs. In addition, SMC*5* can bind to SMC6, participate in the repair of DNA double-strand breaks through homologous recombination, and prevent the transcription of free DNA such as circular virus DNA genomes (Decorsière et al., 2016). Proteomic analysis revealed that Epstein–Barr virus infection disrupts the adhesion proteins SMC5/6, thereby affecting DNA damage repair. In the absence of the involucrin protein BNRF1, SMC5/6 interferes with the formation and encapsidation of viral replication compartments (RCs), ultimately affecting viral lytic replication. SMC5/6 may act as intrinsic immunosensors and restriction factors of human herpes virus RC in viral infectious diseases (Yiu et al., 2022). The SMC5/6 complex compresses viral chromatin to silence gene expression; thus, its depletion enhances viral expression. The SMC5/6 complex also functions in immunosurveillance of extrachromosomal DNA (Dupont et al., 2021). As an intrinsic antiviral restriction factor, Smc5/6, when localized to nuclear domain 10 (ND10) in primary human hepatocytes, inhibits HBV transcription without inducing an innate immune response (Niu et al., 2017). We screened SMC5 signatures in populations vaccinated with different doses. The results suggest that SMC5 may serve as an indicator of vaccine effectiveness.

5. Conclusion

The purpose of this study was to analyze the blood transcriptome in response to different numbers and timing of vaccinations through a variety of machine learning algorithms. It also aimed to identify antiviral immunity-related molecules in different vaccinated populations. The feature intersection of multiple analysis methods reflects the effects of different vaccinations on host gene expression. The analysis results showed that the key gene features were highly consistent with existing research conclusions, which helped us to further clarify the possible mechanisms of these genes. The important antiviral immune characteristics obtained in this study will help in understanding the differences in mechanisms of action of different vaccinations and provide a reference for targeted COVID-19 intervention and for optimization of vaccine strategies.

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=GSE201533.

Author contributions

TH and Y-DC designed the study. JL, WG, and KF performed the experiments. JR and HL analyzed the results. JL, JR, and HL wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

Funding

This research was supported by the National Key R&D Program of China [2022YFF1203202], Strategic 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], and Shandong Provincial Natural Science Foundation [ZR2022MC072].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1138674/full#supplementary-material

SUPPLEMENTARY TABLE S1
Feature lists obtained using Lasso, LightGBM, MCFS, mRMR and PFI.

SUPPLEMENTARY TABLE S2
Performance of IFS with different classification algorithms.

SUPPLEMENTARY TABLE S3
Intersection of feature sets used to construct feasible random forest classifiers on five feature lists.

SUPPLEMENTARY TABLE S4
Classification rules generated by the optimal DT classifier on different feature lists.

References

Amano, M., Otsu, S., Maeda, K., Uemura, Y., Shimizu, Y., Omata, K., et al. (2022). Neutralization activity of sera/IgG preparations from fully BNT162b2 vaccinated individuals against SARS-CoV-2 alpha, Beta, gamma, Delta, and kappa variants. *Sci. Rep.* 12:13524. doi: 10.1038/s41598-022-17071-9

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37, 373–384. doi: 10.1080/00401706.1995.10484371

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Cheriyath, V., Leaman, D. W., and Borden, E. C. (2011). Emerging roles of FAM14 family members (G1P3/ISG 6-16 and ISG12/IFI27) in innate immunity and cancer. *J. Interf. Cytokine Res.* 31, 173–181. doi: 10.1089/jir.2010.0105

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Cross, A. S., Sakarya, S., Rifat, S., Held, T. K., Drysdale, B. E., Grange, P. A., et al. (2003). Recruitment of murine neutrophils in vivo through endogenous sialidase activity. *J. Biol. Chem.* 278, 4112–4120. doi: 10.1074/jbc.M207591200

Dai, Q., Bao, C., Hai, Y., Ma, S., Zhou, T., Wang, C., et al. (2018). MTGIpick allows robust identification of genomic islands from a single genome. *Brief. Bioinform.* 19, 361–373. doi: 10.1093/bib/bbw118

Decorsière, A., Mueller, H., Van Breugel, P. C., Abdul, F., Gerossier, L., Beran, R. K., et al. (2016). Hepatitis B virus X protein identifies the Smc5/6 complex as a host restriction factor. *Nature* 531, 386–389. doi: 10.1038/nature17170

Dramiński, 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

Duan, C., Ma, R., Zeng, X., Chen, B., Hou, D., Liu, R., et al. (2022). SARS-CoV-2 achieves immune escape by destroying mitochondrial quality: comprehensive analysis of the cellular landscapes of lung and blood specimens from patients with COVID-19. *Front. Immunol.* 13:946731. doi: 10.3389/fimmu.2022.946731

Dupont, L., Bloor, S., Williamson, J. C., Cuesta, S. M., Shah, R., Teixeira-Silva, A., et al. (2021). The SMC5/6 complex compacts and silences unintegrated HIV-1 DNA and is antagonized by Vpr. *Cell Host Microbe* 29, 792–805.e6. doi: 10.1016/j.chom.2021.03.001

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: learning a Variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81.

Folegatti, P. M., Ewer, K. J., Aley, P. K., Angus, B., Becker, S., Belij-Rammerstorfer, S., et al. (2020). Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet* 396, 467–478. doi: 10.1016/S0140-6736(20)31604-4

Galli, F. A.-O. X., Marcantonini, G., Giustarini, D. A.-O., Albertini, M. A.-O., Migni, A., Zatini, L., et al. (2022). How aging and oxidative stress influence the cytopathic and inflammatory effects of SARS-CoV-2 infection: the role of cellular glutathione and cysteine metabolism. *Antioxidants (Basel)* 11:1366. doi: 10.3390/antiox11071366

Goonewardena, S. N., Grushko, O. G., Wells, J., Herty, L., Rosenson, R. S., Haus, J. M., et al. (2021). Immune-mediated Glycocalyx remodeling in hospitalized COVID-19 patients. *Cardiovasc. Drugs Ther.*, 1–7. doi: 10.1007/s10557-021-07288-7

Hasan, M. I., Rahman, M. H., Islam, M. B., Islam, M. Z., Hossain, M. A., and Moni, M. A. (2022). Systems biology and bioinformatics approach to identify blood based signatures molecules and drug targets of patient with COVID-19. *Inform. Med. Unlocked* 28:100840. doi: 10.1016/j.imu.2021.100840

Hillier, J. A.-O., Allcott, G. J., Guest, L. A., Heaselgrave, W., Tonks, A. A.-O., Conway, M. A.-O., et al. (2022). The BCAT1 CXXC motif provides protection against ROS in acute myeloid Leukaemia cells. *Antioxidants (Basel)* 11:683. doi: 10.3390/antiox11040683

Hu, S., Li, Z., Lan, Y., Guan, J., Zhao, K., Chu, D., et al. (2020). MiR-10a-5p-mediated Syndecan 1 suppression restricts porcine Hemagglutinating encephalomyelitis virus replication. *Front. Microbiol.* 11:105. doi: 10.3389/fmicb.2020.00105

Huang, F., Fu, M., Li, J., Chen, L., Feng, K., Huang, T., et al. (2023a). Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores. *BBA-Proteins and Proteomics* 1871:140889. doi: 10.1016/j.bbapap.2023.140889

Huang, F., Ma, Q., Ren, J., Li, J., Wang, F., Huang, T., et al. (2023b). Identification of smoking associated transcriptome aberration in blood with machine learning methods. *Biol. Med. Res. Int.* 2023:5333361. doi: 10.1155/2023/5333361

Huang, J., Zhang, Z., Hao, C., Qiu, Y., Tan, R., Liu, J., et al. (2022). Identifying drug-induced liver injury associated with inflammation-drug and drug-drug interactions in pharmacologic treatments for COVID-19 by bioinformatics and system biology analyses: the role of Pregnane X receptor. *Front. Pharmacol.* 13:804189. doi: 10.3389/fphar.2022.804189

Jayakumar, T., Huang, C. J., Yen, T. A.-O., Hsia, C. A.-O., Sheu, J. A.-O., Bhavan, P. A.-O., et al. (2022). Activation of Nrf2 by Esculetin mitigates inflammatory responses through suppression of NF-κB signaling Cascade in RAW 264.7 cells. *Molecules* 27:5143. doi: 10.3390/molecules27165143

Ji, Y., Yin, Y., and Zhang, W. (2020). Integrated Bioinformatic analysis identifies networks and promising biomarkers for hepatitis B virus-related hepatocellular carcinoma. *Int. J. Genom.* 2020, 1–18. doi: 10.1155/2020/2061024

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree, in Proceedings of the 31st International Conference on Neural Information Processing Systems. (Long Beach, California, USA: Curran Associates Inc.).

Khan, H., Patel, S., and Majumdar, A. (2021). Role of NRF2 and Sirtuin activators in COVID-19. *Clin. Immunol.* 233:108879. doi: 10.1016/j.clim.2021.108879

Khatri, V., Chauhan, N., Vishnoi, K., Von Gegerfelt, A., Gittens, C., and Kalyanasundaram, R. (2018). Prospects of developing a prophylactic vaccine against human lymphatic filariasis - evaluation of protection in non-human primates. *Int. J. Parasitol.* 48, 773–783. doi: 10.1016/j.ijpara.2018.04.002

Kong, R., Xu, X., Liu, X., He, P., Zhang, M. Q., and Dai, Q. (2020). 2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome. *BMC Bioinform.* 21:159. doi: 10.1186/s12859-020-3501-2

Lee, H. K., Go, J., Sung, H., Kim, S. W., Walter, M., Knabl, L., et al. (2022a). Heterologous ChAdOx1-BNT162b2 vaccination in Korean cohort induces robust immune and antibody responses that includes omicron. *iScience* 25:104473 doi: 10.1016/j.isci.2022.104473

Lee, H. K., Knabl, L., Walter, M., Knabl, L. Sr., Dai, Y., Fussl, M., et al. (2022b). Prior vaccination exceeds prior infection in eliciting innate and humoral immune responses in omicron infected outpatients. *Front. Immunol.* 13:916686. doi: 10.3389/fimmu.2022.916686

Li, Z., Guo, W., Ding, S., Chen, L., Feng, K., Huang, T., et al. (2022b). Identifying key MicroRNA signatures for neurodegenerative diseases with machine learning methods. *Front. Genet.* 13:880997. doi: 10.3389/fgene.2022.880997

Li, H., Huang, F., Liao, H., Li, Z., Feng, K., Huang, T., et al. (2022a). Identification of COVID-19-specific immune markers using a machine learning method. *Front. Mol. Biosci.* 9:952626. doi: 10.3389/fmolb.2022.952626

Li, X., Li, J., Yang, Y., Hou, R., Liu, R., Zhao, X., et al. (2013). Differential gene expression in peripheral blood T cells from patients with psoriasis, lichen planus, and atopic dermatitis. *J. Am. Acad. Dermatol.* 69, e235–e243. doi: 10.1016/j.jaad.2013.06.030

Li, Z., Mei, Z., Ding, S., Chen, L., Li, H., Feng, K., et al. (2022c). Identifying methylation signatures and rules for COVID-19 with machine learning methods. *Front. Mol. Biosci.* 9:908080. doi: 10.3389/fmolb.2022.908080

Liu, J., Gu, L., Zhang, D., and Li, W. (2022). Determining the prognostic value of spliceosome-related genes in hepatocellular carcinoma patients. *Front. Mol. Biosci.* 9:759792. doi: 10.3389/fmolb.2022.759792

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230. doi: 10.1023/A:1008363719778

Lu, D., Wu, Y., Wang, Y., Ren, F., Wang, D., Su, F., et al. (2012). CREPT accelerates tumorigenesis by regulating the transcription of cell-cycle-related genes. *Cancer Cell* 21, 92–104. doi: 10.1016/j.ccr.2011.12.016

Masic, I., Naser, N., and Zildzic, M. (2020). Public health aspects of COVID-19 infection with focus on cardiovascular diseases. *Mast. Sociomed.* 32, 71–76. doi: 10.5455/msm.2020.32.71-76

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-protein. *Structure* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Mizrahi, B., Lotan, R., Kalkstein, N., Peretz, A., Perez, G., Ben-Tov, A., et al. (2021). Correlation of SARS-CoV-2-breakthrough infections to time-from-vaccine. *Nat. Commun.* 12:6379. doi: 10.1038/s41467-021-26672-3

Mozzi, A., Forcella, M., Riva, A., Difrancesco, C., Molinari, F., Martin, V., et al. (2015). NEU3 activity enhances EGFR activation without affecting EGFR expression and acts on its sialylation levels. *Glycobiology* 25, 855–868. doi: 10.1093/glycob/cwv026

Muchtaridi, M. A.-O., Amirah, S. R., Harmonis, J. A., and Ikram, E. A.-O. (2022). Role of nuclear factor erythroid 2 (Nrf2) in the recovery of Long COVID-19 using natural antioxidants: A systematic review. *Antioxidants (Basel)* 11:1551. doi: 10.3390/antiox11081551

Niu, C., Livingston, C. M., Li, L., Beran, R. K., Daffis, S., Ramakrishnan, D., et al. (2017). The Smc5/6 complex restricts HBV when localized to ND10 without inducing an innate immune response and is counteracted by the HBV X protein shortly after infection. *PLoS One* 12:e0169648. doi: 10.1371/journal.pone.0169648

Olagnier, D., Brandtoft, A. A.-O., Gunderstofte, C., Villadsen, N. L., Krapp, C., Thielke, A. L., et al. (2018). Nrf2 negatively regulates STING indicating a link between antiviral sensing and metabolic reprogramming. *Nat. Commun.* 9:3506. doi: 10.1038/s41467-018-05861-7

Olagnier, D. A.-O., Farahani, E., Thyrsted, J., Blay-Cadanet, J., Herengt, A., Idorn, M. A.-O., et al. (2020). SARS-CoV2-mediated suppression of NRF2-signaling reveals potent antiviral and anti-inflammatory activity of 4-octyl-itaconate and dimethyl fumarate. *Nat. Commun.* 11:4938. doi: 10.1038/s41467-020-18764-3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Pozzetto, B., Legros, V., Djebali, S., Barateau, V., Guibert, N., Villard, M., et al. (2021). Immunogenicity and efficacy of heterologous ChAdOx1-BNT162b2 vaccination. *Nature* 600, 701–706. doi: 10.1038/s41586-021-04120-y

Ran, B., Chen, L., Li, M., Han, Y., and Dai, Q. (2022). Drug-drug interactions prediction using fingerprint only. *Comput. Math. Methods Med.* 2022, 1–14. doi: 10.1155/2022/7818480

Ren, J., Zhou, X., Guo, W., Feng, K., Huang, T., and Vcai, Y.-D. (2022). Identification of methylation signatures and rules for sarcoma subtypes by machine learning methods. *Biomed. Res. Int.* 2022, 1–11. doi: 10.1155/2022/5297235

Reszegi, A., Tatrai, P., Regos, E., Kovalszky, I., and Baghy, K. (2022). Syndecan-1 in liver pathophysiology. *Am. J. Physiol. Cell Physiol.* 323, C289–C294. doi: 10.1152/ajpcell.00039.2022

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* 21, 660–674. doi: 10.1109/21.97458

Sakarya, S., Rifat, S., Zhou, J., Bannerman, D. D., Stamatos, N. M., Cross, A. S., et al. (2004). Mobilization of neutrophil sialidase activity desialylates the pulmonary vascular endothelial surface and increases resting neutrophil adhesion to and migration across the endothelium. *Glycobiology* 14, 481–494. doi: 10.1093/glycob/cwh065

Schaefer, R. E. M., Callahan, R. C., Atif, S. M., Orlicky, D. J., Cartwright, I. M., Fontenot, A. P., et al. (2022). Disruption of monocyte-macrophage differentiation and trafficking by a heme analog during active inflammation. *Mucosal Immunol.* 15, 244–256. doi: 10.1038/s41385-021-00474-8

Tang, S., and Chen, L. (2022). iATC-NFMLP: identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron. *Curr. Bioinforma.* 17, 814–824. doi: 10.2174/1574893617666220318093000

Tang, B. M., Shojaei, M., Parnell, G. P., Huang, S., Nalos, M., et al. (2017). A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection. *Eur. Respir. J.* 49:1602098. doi: 10.1183/13993003.02098-2016

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *J. R. Stat. Soc.* 73, 273–282. doi: 10.1111/j.1467-9868.2011.00771.x

Ullah, H., Sajid, M., Yan, K., Feng, J., He, M., Shereen, M. A., et al. (2021). Antiviral activity of interferon alpha-inducible protein 27 against hepatitis B virus gene expression and replication. *Front. Microbiol.* 12:656353. doi: 10.3389/fmicb.2021.656353

Vollenberg, R. A.-O., Tepasse, P. A.-O., Ochs, K. A.-O., Floer, M., Strauss, M., Rennebaum, F., et al. (2021). Indications of persistent Glycocalyx damage in convalescent COVID-19 patients: A prospective multicenter Study and hypothesis. *Viruses* 13:2324. doi: 10.3390/v13112324

Wada, T., Hata, K., Yamaguchi, K., Shiozaki, K., Koseki, K., Moriya, S., et al. (2007). A crucial role of plasma membrane-associated sialidase in the survival of human cancer cells. *Oncogene* 26, 2483–2490. doi: 10.1038/sj.onc.1210341

Wang, R., and Chen, L. (2022). Identification of human protein subcellular location with multiple networks. *Curr. Proteom.* 19, 344–356. doi: 10.2174/1570164619666220531113704

Wang, H., and Chen, L. (2023). PMPTCE-HNEA: Predicting metabolic pathway types of chemicals and enzymes with a heterogeneous network embedding algorithm. *Curr. Bioinform.* doi: 10.2174/1574893618666230224121633

Wang, Y., Li, J., Zhang, L., Sun, H. X., Zhang, Z., Xu, J., et al. (2022). Plasma cell-free RNA characteristics in COVID-19 patients. *Genome Res.* 32, 228–241. doi: 10.1101/gr.276175.121

Wang, Y., Qiu, H., Hu, W., Li, S., and Yu, J. (2014). RPRD1B promotes tumor growth by accelerating the cell cycle in endometrial cancer. *Oncol. Rep.* 31, 1389–1395. doi: 10.3892/or.2014.2990

Wang, Y., Xu, Y., Yang, Z., Liu, X., and Dai, Q. (2021). Using recursive feature selection with random Forest to improve protein structural class prediction for low-similarity sequences. *Comput. Math. Methods Med.* 2021, 1–9. doi: 10.1155/2021/5529389

Wen, N., Bian, L., Gon, J., and Meng, Y.-A. (2021). RPRD1B is a potentially molecular target for diagnosis and prevention of human papillomavirus E6/E7 infection-induced cervical cancer: A case-control study. *Asia Pac. J. Clin. Oncol.* 17, 230–237. doi: 10.1111/ajco.13439

Wu, C., and Chen, L. (2023). A model with deep analysis on a large drug network for drug classification. *Math. Biosci. Eng.* 20, 383–401. doi: 10.3934/mbe.2023018

Wu, Y., Fau-Yang, X., Fau-Wang, Y., Fau-Ren, F., Fau-Liu, H., Fau-Zhai, Y., et al. (2010). p15RS attenuates Wnt/β-catenin signaling by disrupting {beta}-catenin·TCF4 interaction. *J. Biol. Chem.* 285, 34621–34631. doi: 10.1074/jbc.M110.148791

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovations* 2:100141. doi: 10.1016/j.xinn.2021.100141

Yang, S., Wang, Y., Chen, Y., and Dai, Q. (2020). MASQC: next generation sequencing assists third generation sequencing for quality control in N6-Methyladenine DNA identification. *Front. Genet.* 11:269. doi: 10.3389/fgene.2020.00269

Yang, Z., Yi, W., Tao, J., Liu, X., Zhang, M.Q., Chen, G., et al. (2022). HPVMD-C: a disease-based mutation database of human papillomavirus in China. Database (Oxford) 2022, baac018. doi: 10.1093/database/baac018

Yiu, S. P. T., Guo, R., Zerbe, C., Weekes, M. P., and Gewurz, B. E. (2022). Epstein-Barr virus BNRF1 destabilizes SMC5/6 cohesin complexes to evade its restriction of replication compartments. *Cell Rep.* 38:110411. doi: 10.1016/j.celrep.2022.110411

Zhang, D., Li, L., Chen, Y., Ma, J., Yang, Y., Aodeng, S., et al. (2021). Syndecan-1, an indicator of endothelial glycocalyx degradation, predicts outcome of patients admitted to an ICU with COVID-19. *Mol. Med.* 27:151. doi: 10.1186/s10020-021-00412-1

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

Zou, J., Huang, R. Y., Jiang, F. N., Chen, D. X., Wang, C., Han, Z. D., et al. (2018). Overexpression of TPX2 is associated with progression and prognosis of prostate cancer. *Oncol. Lett.* 16, 2823–2832. doi: 10.3892/ol.2018.9016