ATP6V1B2 and IFI27 and their intrinsic functional genomic characteristics associated with SARS-CoV-2

Zhengjun Zhang

University of Wisconsin

January 13, 2022

Abstract

Genes functionally associated with SARS-CoV-2 and genes functionally related to COVID-19 disease can be different, whose distinction will become the first essential step for successfully fighting against the COVID-19 pandemic. Unfortunately, this first step has not been completed in all biological and medical research. Using a newly developed max-competing logistic classifier, two genes, ATP6V1B2 and IFI27, stand out to be critical in transcriptional response to SARS-CoV-2 with differential expressions derived from NP/OP swab PCR. This finding is evidenced by combining these two genes with one another gene in predicting disease status to achieve better-indicating power than existing classifiers with the same number of genes. In addition, combining these two genes with three other genes to form a five-gene classifier outperforms existing classifiers with ten or more genes. With their exceptional predicting power, these two genes can be critical in fighting against the COVID-19 pandemic as a new focus and direction. Comparing the functional effects of these genes with a five-gene classifier with 100% accuracy identified and tested from blood samples in the literature, genes and their transcriptional response and functional effects to SARS-CoV-2 and genes and their functional signature patterns to COVID-19 antibody are significantly different, which can be interpreted as the former is the point of a phenomenon, and the latter is the essence of the disease. Such significant findings can help explore the causal and pathological clue between SARS-CoV-2 and COVID-19 disease and fight against the disease with more targeted vaccines, antiviral drugs, and therapies.

Keywords: PCR sample, blood sample, COVID-19 detection, gene-gene interaction, functional effects, competing risks, computational medicine.

1 Introduction

The fluctuations in infection rates of the COVID-19 pandemic has been like sea waves, with many small ones and several big ones in the past two years. In the meantime, variants of SARS-CoV-2 have emerged and made scientists and medical practitioners on high alert all the time, and many problems have remained unanswered[1;2;3;4;5;6;7;8;9;10;11]. In addition, there have been new concerns with COVID-19 disease, e.g., SARS-CoV-2 enters the brain[12], COVID-19 vaccines complicate mammograms[13], memory loss and 'brain fog'[14], amongst others. However, these new concerns are observational and experimental outcomes, and they do not have genetic bases due to a lack of effective analytical methods to link COVID-19 to the concerned. Regarding gene expression samples, the literature didn't point out the significant difference between samples with differential expressions derived from NP/OP swab PCR and samples derived from blood samples as the majority of research work focused on individual genes' fold changes. Zhang[15] first applied an innovative algorithm to analyze 126 blood samples from COVID-19-positive and COVID-19-negative patients[16] and reported five critical genes and their competing classifiers, which led to 100% accuracy to classify all hospitalized patients, including ICU patients, to their respective groups. Zhang[17] further develops a mathematical and biological equivalence between COVID-19 and five critical genes and proves the existence of at least three genomic signature patterns and at least seven subtypes. This paper studies gene expression data drawn from NP/OP swab PCR tested samples with COVID-19 positives and negatives. Surprisingly, we find that the functional effects of those five critical genes, ABCB6, KIAA1614, MND1, SMG1, RIPK3, found in Zhang[15;17] are no longer playing a decisive role in PCR samples. At first glance, this observation seems not useful at all, or it even brings doubts about the study methodology and genomics. A careful thought confirms that this observation perfectly suggests the relationship between blood samples and PCR samples. The former stands for the essence of the disease, while the latter stands for the point of the phenomenon. Metaphorically, let's consider samples from the deep sea and samples from the shoreside. The samples from the deep sea represent the meta contents and functions of the sea, while the samples from the shoreside contain likely polluted contents from the bank. Also, the structures of the deep sea have changed along sea waves. As a result, samples from the deep sea and samples from the shoreside will provide very different information. Analogously, deep-sea samples correspond to blood samples, while shoreside samples correspond to PCR samples, which therefore explains the significant difference inferred from the study in Zhang[15;17] and this study. On the other hand, our new finding calls forth an old question: treat the symptoms, cure the root cause or both. Zhang[17] argues that the existence of a genomic signature pattern has to be solved to end the disease, i.e., it is about to cure the root cause. This paper is about treating the symptoms. These two researches reinforce each other, and both are important to current studies of the disease.

The studies[15;17;18;19;20] applied an innovative algorithm to study classifications of COVID-19 patients, breast cancer patients, colorectal cancer patients, and lung cancer patients and gained the highest accuracy (100%) among eleven different study cohorts with thousands of patients. The 100% accuracy establishes a mathematical and biological equivalence between the formed classifiers and the disease, which shows that the study method is effective, informative, and robust. These applications are advanced as they lead to new, interpretable, and insightful functional effects of genes linked to the diseases. The findings can be the key factor in achieving breakthroughs against the diseases. As a result, they shouldn't be wrongly treated as data reanalysis exercises. Due to the limitation of the existing analysis methods and the limited knowledge of the diseases, the fundamental functional effects of genes associated with the disease couldn't be quested even the truth in the collected data has existed for a long time, and the chances of discovering the truth have been wasted. Conducting new experiments, producing new data, and applying the same analysis methods just like repeating, making the same errors of finding suboptimal (even sometimes misleading) answers. This paper is using the innovative method to study differential expressions of human upper respiratory tract gene expressions from 93 COVID-19 positive patients and 141 patients having other acute respiratory illnesses with or without viral[21], and to study host gene expression among RNA-sequencing profiles of nasopharyngeal swabs from 430 individuals with SARS-CoV-2 and 54 negative controls[22]. Using the first dataset, we identify two genes, ATP6V1B2 and IFI27, critical in transcriptional response to SARS-CoV-2. The gene IFI27 was also identified in Mick et al. (2020), but was not entered their final classifiers. In the analysis of the first dataset, a combination of these two genes with RIPK3[15] can lead to an overall accuracy of 87.2%, the sensitivity of 76.3%, and specificity of 94.3%, and a combination of these two genes with one of these three genes, BTN3A1, SERTAD4, EPSTI1, can lead to an overall accuracy of 89.74%, the sensitivity of 89.25∼93.55%, and specificity of 87.24∼90.12%, which are higher than the classifiers in the literature. Using these two genes and one other gene together can easily get overall accuracies between 87.2% and 89.74%, which reveals that these two genes can be fundamental. Combining all these five genes can get to an overall accuracy of 91.88%, the sensitivity of 94.62%, and specificity of 90.08%, which are higher than the classifiers with 10 genes or more in the literature. In the analysis of the second dataset, a combination of the above five genes led to an overall accuracy of 93.39%, a sensitivity of 98.37%, and specificity of 53.70%. Many other combinations will be illustrated in Data Section. These performance results from different combinations indicate that COVID-19 can have many different variants. Different from the studies in Zhang (2021; 2021), the accuracy from any of the combinations applied to PCR gene expressions hasn't been up to 100%. There are three possible reasons, e.g., 1) samples themselves can be false positive or false negative from PCR tests; 2) sample signals were weak, and counts were inaccurate; 3) experimental conditions vary. We note that there are many zero expression values in the second dataset, which may be the reason for a low specificity.

These two critical genes ATP6V1B2 (ATPase H+ Transporting V1 Subunit B2) and IFI27 (Interferon Alpha Inducible Protein 27) had previously been reported to be associated with several diseases. For example, de novo mutation in ATP6V1B2 was found to impair lysosome acidification and cause dominant deafness-onychodystrophy syndrome[23], while IFI27 was found to discriminate between influenza and bacteria in patients with suspected respiratory infection[24], among others.

The significant differences of gene functional effects, gene-gene interactions, and gene-variants interactions between blood sampled gene expressions and PCR sampled gene expressions reveal that ATP6V1B2 and IFI27 are associated with SARS-CoV-2, which points to a new optimal direction of developing more effective vaccines and antiviral drugs. On the other hand, the functional effects of ABCB6, KIAA1614, MND1, SMG1, RIPK3 can be critical to understanding the disease.

The contribution of this paper includes: 1) signifying the genomic difference between PCR samples and blood samples (hospitalized patients); 2) identifying single digit critical genes (ATP6V1B2, IFI27, BTN3A1, SERTAD4, EPSTI1) which are a transcriptional response to SARS-CoV-2; 3) presenting interpretable functional effects of gene-gene interactions, gene-variants interactions using explicitly mathematical expressions; 4) presenting graphical tools for medical practitioners to understand the genomic signature patterns of the virus; 5) making suggestions on developing more efficient vaccines and antiviral drugs; 6) identifying potential genetic clues to other diseases due to COVID-19 infection. The remaining part of the paper is organized as follows. Section 2 briefly reviews the studying methodology. Section 3 reports the data source, analysis results, and interpretations. Finally, Section 4 concludes the study.

2 Methodology

Many medical types of research, especially gene expression data related, applied the classical logistic regression as a starting base, then together with implementations of some advanced machine learning methods. However, Teng and Zhang (2021)[25] points out that classical logistic regression can only model absolute treatments, not relative treatments, and as a result, it has led (and will lead) to many supposedly efficient trials to be wrongly concluded as inefficient. Four clinical trials, including one COVID-19 study trial, were illustrated in their paper. Their new AbRelaTEs regression model for medical data is much more advanced than the classical logistic regression as it greatly enhances interpretability and truly being personalized medicine computability. Our new study in this paper is different from AbRelaTEs as we don't deal with treatment and control, and we use a new innovative method to study the existence of functional effects of genes associated with SARS-CoV-2.

The competing risk factor classifier has been successfully applied in the literature[15;18;19;20]. This section briefly introduces necessary notations and formulas for self-contained due to different data structures used in this work. For continuous responses, the literature papers[26;27;28] deal with max-linear computing factor models and max-linear regressions with penalization. Max-logistic classifier has some connections to the logistic polytomous models but with different structures[29;30;31].

Suppose Y_i is the ith individual patient's COVID-19 status ($Y_i = 0, 2$ for COVID-19 free, $Y_i = 1$ for infected) and $X_i^{(k)} = (X_{i1}^{(k)}, X_{i2}^{(k)}, \ldots, X_{ip}^{(k)})$, $k = 1, \ldots, K$, being the gene expression values with $p = 15979, 35784$ genes in this study. Here k stands for the kth type of gene expression levels drawn based on K different biological sampling methodologies. Note that most published work set $K = 1$, and hence the supercript (k) can be dropped from the predictors. In this research paper, $K = 4$ as we have two datasets, and in the first dataset, there are other ARIs patients with other viral or non-viral. Using a logit link (or probit link, Gumbel link), we can model the risk probability $p_i^{(k)}$ of the ith person's infection status as:

$$\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) = \beta_0^{(k)} + X_i^{(k)}\beta^{(k)} \tag{1}$$

or alternatively, we write

$$p_i^{(k)} = \frac{\exp(\beta_0^{(k)} + X_i^{(k)}\beta^{(k)})}{1+\exp(\beta_0^{(k)} + X_i^{(k)}\beta^{(k)})}$$

where $\beta_0^{(k)}$ is an intercept, $X_i^{(k)}$ is a $1 \times p$ observed vector, and $\beta^{(k)}$ is a $p \times 1$ coefficient vector which characterizes the contribution of each predictor (gene in this study) to the risk.

Considering there have been several variants of SARS-COV-2 and multiple symptoms (subtypes) of COVID-19 diseases, it is natural to assume that the genomic structures of all subtypes can be different. Suppose that all subtypes of COVID-19 diseases may be related to G groups of genes

$$\Phi_{ij}^{(k)} = (X_{i,j_1}^{(k)}, X_{i,j_2}^{(k)}, \ldots, X_{i,j_{g_j}}^{(k)}), j = 1, \ldots, G, g_j \geq 0,\ k = 1, \ldots, K \tag{2}$$

where i is the ith individual in the sample, g_j is the number of genes in jth group.

The competing (risk) factor classifier is defined as

$$\log\left(\frac{p_i^{(k)}}{1-p_i^{(k)}}\right) = \max(\beta_{01}^{(k)} + \Phi_{i1}^{(k)}\beta_1^{(k)},\ \beta_{02}^{(k)} + \Phi_{i2}^{(k)}\beta_2^{(k)}, \ldots, \beta_{0G}^{(k)} + \Phi_{iG}^{(k)}\beta_G) \tag{3}$$

where $\beta_{0j}^{(k)}$'s are intercepts, $\Phi_{ij}^{(k)}$ is a $1 \times g_j$ observed vector, $\beta_j^{(k)}$ is a $g_j \times 1$ coefficient vector which characterizes the contribution of each predictor in the jth group to the risk.

Remark 1. *In (3), $p_i^{(k)}$ is mainly related to the largest component $\beta_{0j}^{(k)} + \Phi_{ij}^{(k)}\beta_j^{(k)}$, $j = 1, \ldots, G$, i.e., all components compete to take the most significant effect.*

Remark 2. *Taking $\beta_{0j}^{(k)} = -\infty$, $j = 2, \ldots, G$, (3) is reduced to the classical logistic regression, i.e., the classical logistic regression is a special case of the new classifier. Compared with blackbox machine learning methods (e.g., random forest, deep learning (convolution) neural network (DNN, CNN)) and regression tree methods, (3) shows clear patterns. Each competing risk factor forms a signature with the selected genes. The number of factors corresponds to the number of signatures, i.e., G. This model can be regarded as a bridge between linear models and more advanced (blackbox) machine learning methods. However, (3) remains the desired properties of interpretability, computability, predictability, and stability. Note that this remark is the same as Remark 1*[20].

In practice, we have to choose a threshold probability value to decide a patient's class label. Following the general trend in the literature, we set the threshold to be 0.5. As such, if $p_i^{(k)} \leq 0.5$, the ith individual is classified as disease free, otherwise the individual is classified to have the disease.

With the above established notations and the idea of quotient correlation coefficient[32], Zhang (2021)[20] introduces a new machine learning classifier, smallest subset and smallest number of signatures (S4) as

$$
\begin{aligned}
(\hat{\beta}, \hat{S}, \hat{G}) \quad = \quad \underset{\beta,\ S_j \subset S,\ j=1,2,\ldots,G}{\arg\min} & \Bigg\{ (1+\lambda_1+|S_u|)^{\sum_{k=1}^{K}\sum_{i=1}^{n}\left(I(p_i^{(k)} \leq 0.5)I(Y_i=1)+I(p_i^{(k)}>0.5)I(Y_i=0)\right)} \\
& +\lambda_2\Big(|S_u| - \frac{|S_u|+G-1}{(|S_u|+1)\times G-1}\Big)\Bigg\}
\end{aligned}
\tag{4}
$$

where $I(.)$ is an indicative function, $p_i^{(k)}$ is defined in Equation (3), $S = \{1, 2, \ldots, 15979, 35784\}$ is the index set of all genes, $S_j = \{j_{j1}, \ldots, j_{j,g_j}\}$, $j = 1, \ldots, G$ are index sets corresponding to (2), S_u is the union of $\{S_j,\ j = 1, \ldots, G\}$, $|S_u|$ is the number of elements in S_u, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are penalty parameters, and $\hat{S} = \{j_{j1}, \ldots, j_{j,g_j},\ j = 1, \ldots, \hat{G}\}$ and $\hat{G}$ are the final gene set selected in the final classifiers and the number of final signatures.

Remark 3. *The case of $K = 1$ corresponds to the classifier introduced in Zhang (2021)*[20]*. The case of $K = 1$ and $\lambda_2 = 0$ corresponds to the classifier introduced in Zhang (2021)*[15]*.*

3 Data Descriptions, Results and Interpretations

3.1 The data

Two COVID-19 datasets to be analyzed are publicly available at https://github.com/czbiohub/covid19-transcriptomics-pathogenesis-diagnostics-results[21] and as GSE152075[22]. The first dataset contains 15979 genes, 93 patients with PCR tested COVID-19 positive, 41 patients with viral acute respiratory illnesses (ARIs) and COVID-19 negative, and 100 non-viral acute respiratory illnesses (ARIs) COVID-19 negative. The second dataset contains 35784 genes, individuals with PCR confirmed SARS-CoV-2, and 54 negative controls. We note that there are many gene expression values in the second dataset being zero.

3.2 The competing factor classifiers and their resulting risk probabilities

Solving the optimization problem (4) among all genes (15979 and 35784), with different combinations, various competing classifiers can be identified. Although, as discussed in Introduction, the gene expression data used in this study were drawn from PCR samples (not blood samples), 100% accurate classifiers with a single-digit number of genes do not exist. Also, with the same accuracy (smaller than 100%), different combinations of genes can be candidate classifiers. Therefore, we report the best-performed classifiers in this subsection. After an extensive Monte Carlo search of the best combinations of genes, five genes, ATP6V1B2, IFI27, BTN3A1 (Butyrophilin Subfamily 3 Member A1), SERTAD4 (SERTA Domain Containing 4), EPSTI1 (Epithelial Stromal Interaction 1), are found to form the S4 classifiers.

Given the first dataset has three categories (COVID-19 positive, ARIs with non-SARS-CoV-2 viral, ARIs without viral), we also study the classification between COVID-19 positive and ARIs with non-SARS-CoV-2 viral, and between COVID-19 positive and ARIs without viral, which leads to $K = 4$ as stated in the prior subsection.

Table 1: First dataset: Characteristics of the top performed individual genes together with ATP6V1B2 and IFI27 to form a three-gene classifier.

Classifier	Intercept	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1	Accuracy	Sensitivity	Specificity
BTN3A1	-9.8180	-8.0116	2.1871	5.2583			88.46%	83.87%	91.49%
SERTAD4	-4.5269	-1.9712	2.1584		-7.8030		89.32%	86.02%	91.49%
EPSTI1	-7.2904	-7.2500	2.6524			4.1633	89.74%	93.55%	87.23%

Table 2: First dataset: Characteristics of RIPK3 together with ATP6V1B2 and IFI27.

Classifier	Intercept	ATP6V1B2	IFI27	RIPK3	Accuracy	Sensitivity	Specificity
RIPK3	-1.2487	-5.7586	1.3916	9.9020	87.2%	76.3%	94.3%

Note that in (3) each individual component itself is a classifier which has the following form

$$\beta_0 + \beta_1 \times \text{ATP6V1B2} + \beta_2 \times \text{IFI27} + \beta_3 \times \text{BTN3A1} + \beta_4 \times \text{SERTAD4} + \beta_5 \times \text{EPSTI1} \tag{5}$$

where $(\beta_0, \beta_1, \ldots, \beta_5)$ are coefficients. In the subsequent subsections, we use tables to present individual $(\mathrm{CF}_{i,j})$ and combined (CFmax_j) classifiers representing (5), where i is the index for classifier, and j is for dataset.

The risk probabilities of each component classifier are

$$\mathrm{P}_{i,j} = \frac{\exp\left(\mathrm{CF}_{i,j}\right)}{1+\exp\left(\mathrm{CF}_{i,j}\right)}, \ i=1,\,2,\ j=1,\,2, \tag{6}$$

and the risk probabilities based on all three component classifiers together are

$$\mathrm{Pmax}_j = \frac{\exp\left(\mathrm{CFmax}_j\right)}{1+\exp\left(\mathrm{CFmax}_j\right)}, \ j=1,\,2. \tag{7}$$

3.3 First dataset: Three-gene classifiers $(G = 1)$

Note that the results in this subsection are not from our final best-performed classifiers. We found that a combination of ATP6V1B2 and IFI27 with many other genes can lead to high accuracy classifiers. We present their performance combined with the remaining genes of the best subset of five genes in this paper and one of the five critical genes found by Zhang[15]. Tables 1 and 2 summarize the results.

In both Tables 1 and 2, we see that the coefficient signs of ATP6V1B2 and IFI27 are the same across all individual classifiers, which is a strong indication that they are truly associated with the virus. Although gene RIPK3 plays a key role in the perfect classifier identified in Zhang[15], its performance is inferior to the other three genes identified from PCR samples in this paper. This phenomenon reflects the discussions in Introduction that RIPK3 is related to the natural essence of COVID-19, while ATP6V1B2, IFI27, BTN3A1, SERTAD4, and EPSTI1 contain more information about SARS-CoV-2.

We note that for BTN3A1, its combinations with ATP6V1B2 and IFI27 can have numerous types, which also leads to the same accuracy; for SERTAD4, there are numerous combinations with ATP6V1B2 and IFI27; and the same is true for EPSTI1. The coefficients listed in Table 1 are just a particular type of coefficient. Also, for EPSTI1, we can get different sensitivities and specificities while maintaining the same accuracy. Among four genes (BTN3A1, SERTAD4, EPSTI1, and RIPK3), EPSTI1 has the best performance in Tables 1 and 2. This empirical evidence proves that ATP6V1B2 and IFI27 are at the center of genes associated with SARS-CoV-2.

Table 3: First dataset: Characteristics of the top performed five-gene classifier. CF1 and CF2 stand for the first and second individual classifier for data COVID-19 patients vs. other viral ARIs and non-viral patients.

Classifier	Intercept	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1	Accuracy	Sensitivity	Specificity
CF1	9.1930	-1.8935	1.5774		-4.3303		87.61%	81.72%	91.49%
CF2	-7.2786	-5.2993		3.2572		2.3400	86.32%	76.34%	92.91%
max{CF1, CF2}							91.88%	94.62%	90.07%

Table 4: First dataset: Characteristics of the top performed five-gene classifier. CF1 and CF2 stand for the first and second individual classifier for data COVID-19 patients vs. other viral ARIs but non-viral patients.

Classifier	Intercept	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1	Accuracy	Sensitivity	Specificity
CF1	-2.0520		3.9086	2.5578		-9.6586	70.15%	62.37%	87.80%
CF2	5.5979	-7.4352			8.3704	4.4936	76.12%	74.19%	80.49%
max{CF1, CF2}							91.04%	97.85%	75.61%

3.4 First dataset: Five-gene classifiers and the existence of variants

Our extensive Monte Carlo search leads to the best solution of the accuracy of 91.82% to the optimization problem (4) as five genes, i.e., ATP6V1B2, IFI27, BTN3A1, SERTAD4, and EPSTI1 though the solution is not unique. These five genes stand out after comparing solutions for all three categories in the first dataset. Tables 3-5 summarize the results.

Table 6 demonstrates part of patients' expression values of the five critical genes, competing classifier factors, predicted probabilities. Note that due to very relative large scales in Columns CF-1, CF-2, CFmax, they are rescaled by a division of 100 when computing the risk probabilities as very large values can result in an overflow in computation. The validity of rescaling was justified in Zhang[17].

Figure 1 presents critical gene expression levels and risk probabilities corresponding to different combinations in the first dataset and Tables 3-5. It can be seen that each plot shows a genomic signature pattern and functional effects of genes involved.

From Tables 1-5, we can immediately see that the coefficient signs associated with ATP6V1B2 are uniformly negative, which shows that increasing the expression level of ATP6V1B2 will decrease the virus (SARS-CoV-2) strength; the coefficient signs associated with IFI27 are uniformly positive, which shows that decreasing the expression level of IFI27 will decrease the virus (SARS-CoV-2) infection strength. Such functional effects of ATP6V1B2 and IFI27 can also be clearly seen in Figure 1 around origins which show the higher the IFI27 level, the higher the risk probability (yellow color); the higher the ATP6V1B2 level, the lower the risk probability (blue color). These observations show that ATP6V1B2 and IFI27 are in the circle of genes associated with SARS-COV-2. BTN3A1 appears three times in Tables 3-5 with positive coefficients, which shows decreasing the expression level of BTN3A1 will decrease the virus (SARS-CoV-2) infection strength. The coefficient signs of SERTAD4 and the coefficient signs of EPSTI1 show both positive and negative in Tables 3-5 depending on the ways of genes being combined. These phenomena explain the reason SARS-CoV-2 variants have emerged as variants can be related to different coefficient signs corresponding to genes.

Table 5: First dataset: Characteristics of the top performed five-gene classifier. CF1 and CF2 stand for the first and second individual classifier for data COVID-19 patients vs. non-viral ARIs patients.

Classifier	Intercept	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1	Accuracy	Sensitivity	Specificity
CF1	-2.2381	-7.9733		4.5448		4.7567	90.16%	81.72%	98.00%
CF2	-2.1003	-4.8036	4.0849		-9.9738		90.16%	82.80%	97.00%
max{CF1, CF2}							96.37%	95.70%	97.00%

Table 6: First dataset: Expression values of the five critical genes, competing classifier factors, predicted probabilities.

#ID	Status	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1	CF-1	CF-2	CFmax_1	P_1	P_2	P-1
e-202	0	277	604	104	158	138	-246.74	-813.52	-246.74	0.08	0.00	0.08
e-080	0	866	103	82	76	94	-1797.21	-4109.42	-1797.21	0.00	0.00	0.00
e-287	0	3127	717	271	233	151	-5789.75	-15342.15	-5789.75	0.00	0.00	0.00
e-753	1	1053	2029	766	214	819	289.20	-1175.97	289.20	0.95	0.00	0.95
e-751	1	253	1423	266	114	369	1281.12	381.87	1281.12	1.00	0.98	1.00
e-520	0	617	344	120	11	559	-664.10	-1578.02	-664.10	0.00	0.00	0.00
e-505	0	721	240	298	10	500	-1020.75	-1687.43	-1020.75	0.00	0.00	0.00
i-083	0	191	320	119	72	71	-159.48	-465.70	-159.48	0.17	0.01	0.17
e-764	0	1667	202	76	3	1232	-2841.63	-5710.78	-2841.63	0.00	0.00	0.00
e-451	0	1880	24	98	2	27	-3521.39	-9587.58	-3521.39	0.00	0.00	0.00
e-285	0	794	826	530	392	300	-1888.79	-1786.61	-1786.61	0.00	0.00	0.00
e-254	0	512	253	195	388	69	-2241.35	-1923.91	-1923.91	0.00	0.00	0.00
e-726	1	398	1395	362	96	567	1040.34	389.49	1040.34	1.00	0.98	1.00

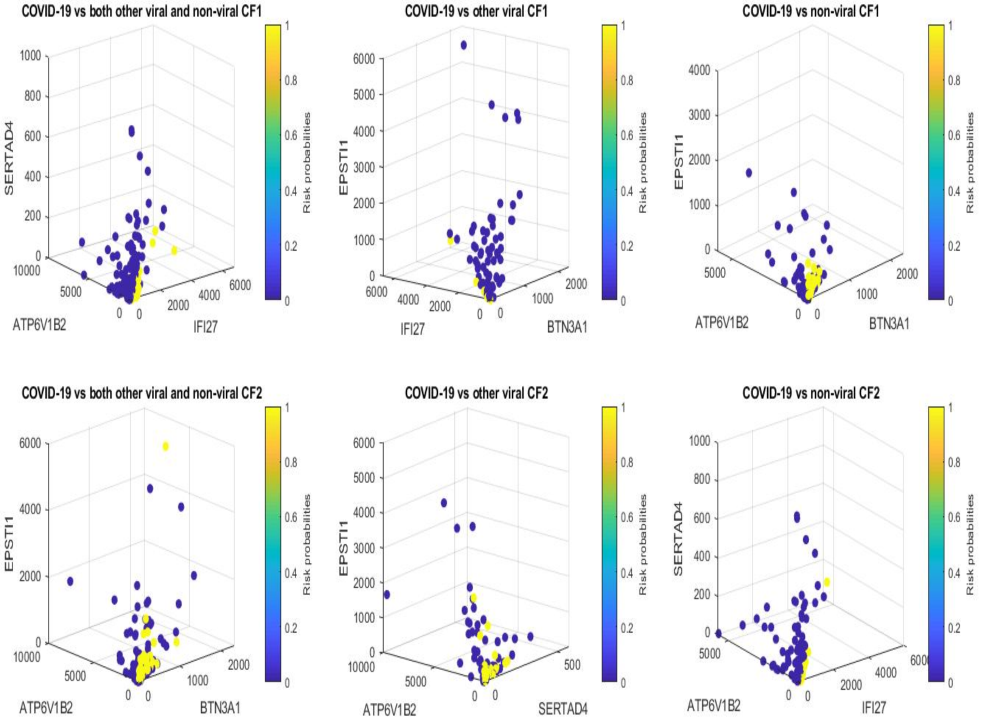

Figure 1: *COVID-19 classifiers in Tables 3-5: Visualization of gene-gene relationship and gene-risk probabilities. Note that 0.5 is the probability threshold.*

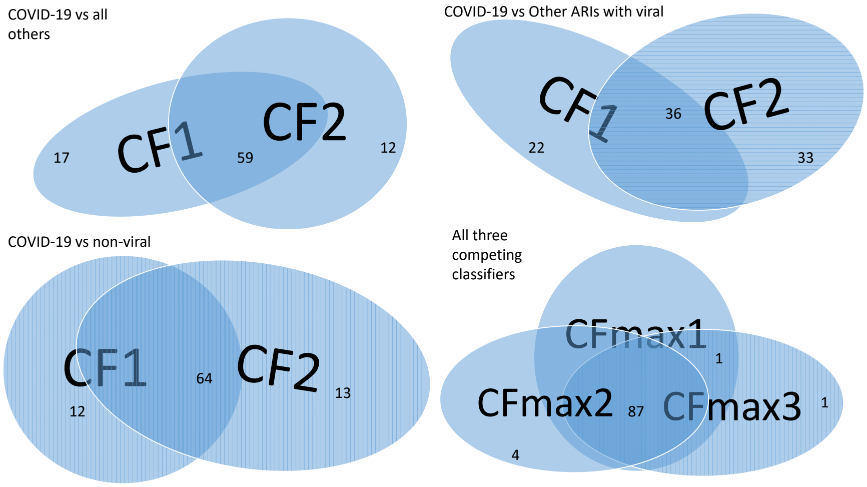

Figure 2: *Venn diagram of variants of SARS-CoV-2 (the first dataset): Top-left panel is for COVID-19 vs. all others; Top-right panel is for COVID-19 vs. other viral; Bottom-left panel is for COVID-19 vs. non viral; Bottom-right panel is for all three together.*

Figure 2 is a Venn diagram to illustrate the performance of each classifier and the combined classifier. In Venn diagram, those patients who fall in the intersections are relatively easy to be tested and confirmed positive, while for those who only fall in one category, it is relatively hard to test and confirm their status. Two individual classifiers can be explained as having two times COVID-19 tests using two different testing procedures, and with both tests being positive, the probability of infection will be higher depending on the sensitivity and the specificity of each test. Summarizing Tables 3-5 and Figure 2, mathematically speaking, SARS-CoV-2 can have $3 \times 3 \times 3 \times 4 = 108$ variants with some of them being insignificant from dominant ones while some of them being dominant and having emerged (or will emerge), where the multiplier 3 corresponds to 3 classes in one Venn diagram, and similarly, other numbers are interpreted. We note that the joint functional effects of genes are not directly observable, and the meaning of variants is defined by the joint functional effects. As a result, the variants of the virus are not directly referred to what has been known in the literature and practice.

Comparing the individual classifiers and combined classifiers among COVID-19 vs. all others, COVID-19 vs. ARIs with other viral, and COVID-19 vs. without viral, we see that the combined classifier for the case of COVID-19 vs. without viral works the best. We found some ARIs with other viral may be COVID-19 patients but not yet confirmed. If we apply the classifier in Figure 2 bottom-right panel, we can get sensitivity up to 98.94% with a slight loss of specificity.

3.5 Second dataset: Five-gene classifiers and the existence of variants

The five genes, ATP6V1B2, IFI27, BTN3A1, SERTAD4, EPSTI1, achieved superior performance in classifying patients in their respective groups. In this subsection, we test their performance in a second dataset.

Table 7: Second dataset: Characteristics of the top performed five-gene classifier. CF1 and CF2 stand for the first and second individual classifier for data COVID-19 confirmed vs. COVID-19 negative.

Classifier	Intercept	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1	Accuracy	Sensitivity	Specificity
CF1	-9.0153	8.2227	-3.6174	0.2020	-10.2465	8.4920	91.32%	95.35%	59.26%
CF2	1.7700	-7.0875	-1.7572	-8.7975	6.2012	8.7980	27.07%	18.84%	92.59%
max{CF1, CF2}							93.39%	98.37%	53.70%

Table 8: Pairwise correlation coefficients: The upper triangle is for the first dataset, and the lower triangle is for the second dataset.

	ATP6V1B2	IFI27	BTN3A1	SERTAD4	EPSTI1
ATP6V1B2	–	0.2080	0.5416	0.0510	0.5415
IFI27	0.4031	–	0.5463	0.3084	0.5616
BTN3A1	0.6900	0.3823	–	0.2500	0.7527
SERTAD4	0.3417	0.3302	0.2663	–	0.0079
EPSTI1	0.6531	0.3366	0.6562	0.1791	–

One significant difference between these two datasets is that the patients in the first study (dataset) are either COVID-19 positive or ARIs with other viral or ARIs without viral, while the patients in the second study (dataset) are PCR confirmed SARS-CoV-2 or negative controls. As a result, genes found to be critical from the first dataset can be thought of as SARS-CoV-2 specific. It turned out that those five genes are also the best subset for the second dataset. Table 7 presents the individual classifier and the combined classifier. Data are ln(raw+1) normalized.

We can see that the signs of ATP6V1B2, IFI27 in CF1 remain the same as their counterparts in Tables 1-5 while the sign of ATP6V1B2 changed in CF2. This phenomenon is not surprising as CF1 has 91.32% overall accuracy, while CF2 has only 27.07% accuracy. This table again supports our earlier claim that ATP6V1B2, IFI27 are in the circle of critical genes associated with SARS-CoV-2.

Note that individual classifiers in the second dataset involve all five genes while counterparts in the first dataset only involve three genes. This phenomenon can be explained as the patients' attributes from these two datasets are different. Next, we compute the correlations among those five genes for each dataset. Table 8 presents pairwise correlations in a matrix form in which the upper triangle is for the first dataset, and the lower triangle is for the second dataset.

Table 8 shows different correlation structures among the five genes, which shows the difference of classifiers between two datasets is reasonable.

4 Discussions

The results presented in this paper are the first to directly associate a few critical genes with SARS-CoV-2 with the best performance (relative to other subsets with the same number of genes). Furthermore, the results signify the genomic difference between PCR samples and blood samples (hospitalized patients), identify single digit critical genes (ATP6V1B2, IFI27, BTN3A1, SERTAD4, EPSTI1) which are a transcriptional response to SARS-CoV-2, interpretable functional effects of gene-gene interactions, gene-variants interactions using explicitly mathematical expressions, introduce graphical tools for medical practitioners to understand the genomic signature patterns of the virus, make suggestions on developing more efficient vaccines and antiviral drugs, and finally identify potential genetic clues to other diseases due to COVID-19 infection.

In Zhang[17], a conceptual visualization of the gene-gene relationship was created. At the top of the figure, virus variants were placed. With new findings of this paper, six signature patterns from Tables 3-5 can be used to replace those virus variants, and then a complete dynamic flow can be formed.

As discussed in Introduction, the genes identified in Zhang[17] are hypothesized to link to the root cause of COVID-19, while the genes identified in this study are the key to treat the symptoms. Based on the findings in this paper, we make the following hypotheses.

Hypothesis 1: The five genes[17] ABCB6, KIAA1614, MND1, SMG1, RIPK3 and their functional effects are the key to cure the root cause.

Hypothesis 2: The five genes ATP6V1B2, IFI27, BTN3A1, SERTAD4, EPSTI1 and their functional effects are the key to treat the symptoms.

Hypothesis 3: The gene CDC6[17] (cell division cycle 6) is a protein essential for the initiation of RNA replication.

Hypothesis 1 is based on the mathematical and biological equivalence between COVID-19 disease and the functional effects of these five genes proved in Zhang[17]. At the moment, testing Hypothesis 2 is more urgent than testing Hypothesis 1 given variants of SARS-CoV-2 have been emerging, and waves of COVID-19 have been arriving one after another. Once Hypothesis 2 is tested and confirmed, scientists can test their counterparts from animals, trace the virus origin, and find the intermediate host species of SARS-CoV-2. As to Hypothesis 3, in Zhang (2021), a combination of CDC6 and ZNF282 (Zinc Finger Protein 282) can lead to 97.62% accuracy (98% sensitivity, 96.15% specificity), which suggests the protein encoded by CDC6 is a protein essential for the initiation of RNA replication.

As mentioned in Introduction, ATP6V1B2 was found to impair lysosome acidification and cause dominant deafness-onychodystrophy syndrome[23], while IFI27 was found to discriminate between influenza and bacteria in patients with suspected respiratory infection[24]. There have been new concerns with COVID-19 disease, e.g., SARS-CoV-2 enters the brain[12], COVID-19 vaccines complicate mammograms[13], memory loss and 'brain fog'[14]. Using the findings from this paper, we may hypothesize that ATP6V1B2 can be a leading factor causing COVID-19 to brain function and ENT problems. As to IFI27, given that COVID-19 is a respiratory tract infection, it makes sense to hypothesize IFI27 is the infection's key. EPSTI1 has been found related to breast cancer and oral squamous cell carcinoma (OSCC) and lung squamous cell carcinoma (LSCC)[33], which may link COVID-19 to what has been found in mammograms complication[13]. Liang et al.[34] suggests that BTN3A1 may function as a tumor suppressor and may serve as a potential prognostic biomarker in NSCLCs and BRCAs. However, all of these findings have not been confirmed. A confirmed Hypothesis 2 may help further explore whether these genes reported in the literature are truly effective, as suggested in the literature.

Finally, with the proven existence of signature patterns associated with SARS-CoV-2 and COVID-19, variants of the disease will continue to emerge if the problems revealed by the existing signatures are not solved. We have witnessed that each time after a peak of the COVID-19 pandemic, the world saw hopes of the end of the pandemic, and the public lowered their guard; as a result, another wave (small or big) appeared. As such, we shouldn't forget the pain where the gain follows as existence determines recurrence noted by Murphy's law "Anything that can go wrong will go wrong."

Acknowledgments

To be added later.

Supplementary materials

Real data and computer outputs are in a supplementary file available online and submitted together with this paper. A Matlab® demo code for solving A in Equation (4) ($\lambda_2 = 0$) is also available.

Data availability

The datasets are publicly available. The data links are stated in Section Data Description.

Competing interests

The author declares no competing interests.

Limitations

Solving optimization problems (4) involves combinatorial optimization, integer programming, and continuous programming. The computational complexity is exceptionally high, and we haven't figured out how to define the complexity. We used an extensive Monte Carlo search method to find the best solution. However, we cannot guarantee whether additional sets of genes can also be the optimal solutions. Although we have identified functional effects by gene-gene interactions and gene-subtype (variants) interactions of the five genes, we haven't identified how gene-gene interacts with each other and their causal directions. We are working in this direction. Due to the lack of available new sampled data for new variants, it's difficult to infer the risks of variants. Finally, our results are in the field of computational biology/medicine, and they are not lab-confirmed.

References

[1] C. Rowland. Doctors and nurses want more data before championing vaccines to end the pandemic: Health systems are launching bids to assure their medical workers that vaccines will be safe and effective. *CNN*, pages November 21, 2020 at 6:00 a.m. CST, 2020. URL `https://www.washingtonpost.com/business/2020/11/21/vaccines-advocates-nurses-doctorscoronavirus/`.

[2] Ewen Callaway. The quest to find genes that drive severe covid. *Nature*, pages 346–348, 2021. ISSN 595(7867). doi: 10.1038/d41586-021-01827-w.

[3] COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of covid-19. *Nature*, pages 1476–4687, 2021. doi: 10.1038/s41586-021-03767-x. URL `https://doi.org/10.1038/s41586-021-03767-x`.

[4] E. Pairo-Castineira, S. Clohisey, L. Klaric, and et al. Genetic mechanisms of critical illness in covid-19. *Nature*, 591:92–98, 2021. URL `https://doi.org/10.1038/s41586-020-03065-y`.

[5] The RECOVERY Collaborative Group. Dexamethasone in hospitalized patients with covid-19. *New England Journal of Medicine*, 384(8):693–704, 2021. doi: 10.1056/NEJMoa2021436. URL `https://doi.org/10.1056/NEJMoa2021436`.

[6] Gillian S. Dite, Nicholas M. Murphy, and Richard Allman. Development and validation of a clinical and genetic model for predicting risk of severe covid-19. *Epidemiology and Infection*, 149:e162, 2021. doi: 10.1017/S095026882100145X.

[7] Zhang, Q. et al. Inborn errors of type i ifn immunity in patients with life-threatening covid-19. *Science*, 370:eabd4570, 2020. doi: 10.1126/science.abd4570.

[8] Bastard, P. et al. Autoantibodies against type i ifns in patients with life-threatening covid-19. *Science*, 370:eabd4585, 2020. doi: 10.1126/science.abd4585.

[9] Povysil, Gundula et al. Failure to replicate the association of rare loss-of-function variants in type i ifn immunity genes with severe covid-19. *medRxiv*, 2020. doi: 10.1101/2020.12.18.20248226. URL `https://www.medrxiv.org/content/early/2020/12/21/2020.12.18.20248226`.

[10] Kosmicki, J. A. et al. Genetic association analysis of sars-cov-2 infection in 455,838 uk biobank participants. *medRxiv*, 2020. doi: 10.1101/2020.10.28.20221804. URL `https://www.medrxiv.org/content/early/2020/11/03/2020.10.28.20221804`.

[11] Fallerini, Chiara et al. Association of toll-like receptor 7 variants with life-threatening covid-19 disease in males: findings from a nested case-control study. *eLife*, 10:e67569, mar 2021. ISSN 2050-084X. doi: 10.7554/eLife.67569. URL `https://doi.org/10.7554/eLife.67569`.

[12] Elizabeth M. Rhea, Aric F. Logsdon, Kim M. Hansen, and et al. The s1 protein of sars-cov-2 crosses the blood–brain barrier in mice. *Nature Neuroscience*, 24(3):368–378, 2021. URL `https://doi.org/10.1038/s41593-020-00771-8`.

[13] Covid-19 vaccines complicate mammograms. *Cancer Discovery*, 11(8):1868–1868, 2021. ISSN 2159-8274. doi: 10.1158/2159-8290.CD-NB2021-0366. URL `https://cancerdiscovery.aacrjournals.org/content/11/8/1868.1`.

[14] Jacqueline H. Becker, Jenny J. Lin, Molly Doernberg, Kimberly Stone, Allison Navis, Joanne R. Festa, and Juan P. Wisnivesky. Assessment of Cognitive Function in Patients After COVID-19 Infection. *JAMA Network Open*, 4(10):e2130645–e2130645, 10 2021. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2021.30645. URL `https://doi.org/10.1001/jamanetworkopen.2021.30645`.

[15] Zhengjun Zhang. Five critical genes related to seven Covid-19 subtypes: A data science discovery. *Journal of Data Science*, 19(1):142–150, 2021. URL `https://doi.org/10.6339/21-JDS1005`.

[16] Katherine A. Overmyer, Evgenia Shishkova, Ian J. Miller, Joseph Balnis, Matthew N. Bernstein, Trenton M. Peters-Clarke, Jesse G. Meyer, Qiuwen Quan, Laura K. Muehlbauer, Edna A. Trujillo, Yuchen He, Amit Chopra, Hau C. Chieng, Anupama Tiwari, Marc A. Judson, Brett Paulson, Dain R. Brademan, Yunyun Zhu, Lia R. Serrano, Vanessa Linke, Lisa A. Drake, Alejandro P. Adam, Bradford S. Schwartz, Harold A. Singer, Scott Swanson, Deane F. Mosher, Ron Stewart, Joshua J. Coon, and Ariel Jaitovich. Large-scale multi-omic analysis of covid-19 severity. *Cell Systems*, page doi.org/10.1016/j.cels.2020.10.003, 2020. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2020.10.003. URL `http://www.sciencedirect.com/science/article/pii/S2405471220303719`.

[17] Zhengjun Zhang. The existence of at least three genomic signature patterns and at least seven subtypes of covid-19 and the end of the disease. *Editor Invited Minor Revision Submitted*, waiting for editor's decision, 2021.

[18] Zhengjun Zhang. Lift the veil of breast cancers using four or fewer critical genes. *Cancer Informatics*, In press, 2022. URL `doi: 10.1177/11769351221076360`.

[19] Zhengjun Zhang, Yuqing Xu, Xiaoxing Li, Mengke Chen, Xueqin Wang, Ning Zhang, and Yongjun Liu. A perfect classifier for machine learning heterogeneous cohort studies: The puzzle and the future of colorectal cancers told by four critical genes. *Manuscript under revision*, 2021.

[20] Zhengjun Zhang. Functional effects of four or fewer critical genes linked to lung cancers and new subtypes detected by a new machine learning classifier. *Journal of Clinical Trials*, 11:S14:100001, 2021. URL `https://www.longdom.org/open-access/functional-effects-of-four-or-fewer-critical-genes-linked-to-lu`

[21] E Mick, J Kamm, A.O. Pisco, K Ratnasiri, and et al. Upper airway gene expression reveals suppressed immune responses to sars-cov-2 compared with other respiratory viruses. *Nat Commun*, 11:5854, 2020. URL `https://doi.org/10.1038/s41467-020-19587-y`.

[22] NAP Lieberman, V Peddu, H Xie, and et al. In vivo antiviral host transcriptional response to sars-cov-2 by viral load, sex, and age. *PLoS Biol.*, 18(9):e3000849, 2020.

[23] Benjamin M. Tang, Maryam Shojaei, Grant P. Parnell, and et al. A novel immune biomarker ifi27 discriminates between influenza and bacteria in patients with suspected respiratory infection. *European Respiratory Journal*, 49(:):1602098, 2017. doi: 10.1183/13993003.02098-2016.

[24] Y Yuan, J Zhang, Q Chang, and et al. De novo mutation in ATP6V1B2 impairs lysosome acidification and causes dominant deafness-onychodystrophy syndrome. *Cell Res.*, 24(11):1370–3, 2014. doi: 10.1038/cr.2014.77.

[25] Hao Yang Teng and Zhengjun Zhang. Directly and simultaneously expressing absolute and relative treatment effects in medical data models and applications. *Entropy*, 23(11), 2021. ISSN 1099-4300. doi: 10.3390/e23111517. URL `https://www.mdpi.com/1099-4300/23/11/1517`.

[26] J. Aitchison and J. A. Bennett. Polychotomous quantal response by maximum indicant. *Biometrika*, 57(2):253–262, 08 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.2.253. URL `https://doi.org/10.1093/biomet/57.2.253`.

[27] Qiurong Cui and Zhengjun Zhang. Max-linear competing factor models. *Journal of Business & Economic Statistics*, 36(1):62–74, 2018. doi: 10.1080/07350015.2015.1137761. URL `https://doi.org/10.1080/07350015.2015.1137761`.

[28] Qiurong Cui, Yuqing Xu, Zhengjun Zhang, and Vincent Chan. Max-linear regression models with regularization. *Journal of Econometrics*, 222: 579–600, 2021. doi: https://doi.org/10.1016/j.jeconom.2020.07.017. URL `http://www.sciencedirect.com/science/article/pii/S0304407620302074`.

[29] Daniel McFadden. Econometric Models for Probabilistic Choice among Products. *The Journal of Business*, 53(3):S13–29, 1980.

[30] Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, Cambridge, 1985.

[31] Jing Qin. *Discrete Data Models*, pages 249–257. Springer Singapore, Singapore, 2017. ISBN 978-981-10-4856-2. doi: 10.1007/978-981-10-4856-2-13. URL `https://doi.org/10.1007/978-981-10-4856-2-13`.

[32] Zhengjun Zhang. Quotient correlation: a sample based alternative to Pearson's correlation. *The Annals of Statistics*, 36(2):1007–1030, 2008.

[33] Mengmeng Fan, Makoto Arai, Akinobu Tawada, and et al. Contrasting functions of the epithelial-stromal interaction 1 gene, in human oral and lung squamous cell cancers. *Oncology Reports*, 47(1):5, 2022. URL `https://doi.org/10.3892/or.2021.8216`.

[34] F Liang, C Zhang, Guo H, and et al. Comprehensive analysis of btn3a1 in cancers: mining of omics data and validation in patient samples and cellular models. *FEBS Open Bio.*, 11(9):2586–2599, 2021. doi: 10.1002/2211-5463.13256.