Statistics and Applications ISSN 2452-7395 (online)
Volume 18, No. 1, 2020 (New Series), pp 253–268

# Identification of Geographic Specific SARS-Cov-2 Mutations by Random Forest Classification and Variable Selection Methods

**Manoj Kandpal and Ramana V Davuluri[1]**
*Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.*
[1]*Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA.*

Received: 06 June 2020; Revised: 28 June 2020; Accepted: 30 June 2020

**Abstract**

RNA viral genomes have very high mutations rates. As infection spreads in the host populations, different viral lineages emerge acquiring independent mutations that can lead to varied infection and death rates in different parts of the world. By application of Random Forest classification and feature selection methods, we developed an analysis pipeline for identification of geographic specific mutations and classification of different viral lineages, focusing on the missense-variants that alter the function of the encoded proteins. We applied the pipeline on publicly available SARS-CoV-2 datasets and demonstrated that the analysis pipeline accurately identified country or region-specific viral lineages and specific mutations that discriminate different lineages. The results presented here can help designing country-specific diagnostic strategies and prioritizing the mutations for functional interpretation and experimental validations.

*Key words:* Random forest; Feature selection; Classification; SARS-CoV-2; Coronavirus.

## 1. Introduction

In December 2019, researchers identified a novel coronavirus that first infected and caused coronavirus disease (COVID) in patients in Wuhan, China (Lu et al. 2020b). The virus, initially named as 2019-nCoV, was officially renamed as SARS-CoV-2 by the International Committee on Taxonomy of Viruses to indicate that it was very closely related to the SARS (Severe Acute Respiratory Syndrome Coronavirus). It infected 6,265,496 confirmed cases and caused 375,526 deaths globally as of June 1, 2020 (https://coronavirus.jhu.edu/). SARS-CoV-2 is an enveloped single-stranded RNA virus. It infects a human host by breaking into the host's cell and acquires mutations during replications in the cell. As it spreads from person to person, the accumulated mutations in the viral genomes can lead to different viral lineages. One particular type of mutations, called missense mutations, alter the amino acids encoded by the RNA sequences. For example, some missense mutations alter a protein to give growth advantage for the virus – allowing virus entry into a host cell, and others can lead to changes in the target region of a drug or antibody that acts against the virus protein (Zhao et al. 2018; Holland et al. 2020). Therefore, computational methods to prioritize specific mutations from a large set of passenger mutations and classify different lineages is of great importance for the ongoing COVID research.

Corresponding Author: Ramana V. Davuluri
Email: Ramana.Davuluri@stonybrookmedicine.edu

We developed a computational pipeline for constructing a tree based Random Forest classifier to discriminate SARS-CoV-2 lineages from different geographic regions and identify important mutations, using the rich source of existing mutational profiles and associated genomic annotations and geographic information. Here, we attempt to classify viral lineages from four geographic locations – 1) USA-New York; 2) China; 3) Europe-Spain and Italy; and 4) India. We prepared a dataset by processing publicly available mutational profiles that were curated by analyzing 20,746 SARS-CoV-2 genome sequences. These genome sequences were sequenced from infected patient samples in different countries. We systematically trained and evaluated Random Forest (RF) classifiers on subset of this dataset, using both cross validation and testing on independent test set, and selected the best performing RF classifier for the final algorithm.

## 2. Data Description

Working around the world in different countries, teams of scientists are racing to understand the virus's genetic sequences, develop treatments and vaccine candidates, and to accurately forecast future outbreaks. In this unprecedented effort, more than 30,000 SARS-CoV-2 genomes have been sequenced and submitted to public data repositories since the outbreak in December, 2019 (Colson et al. 2020; Lu et al. 2020a; Yadav et al. 2020). By aligning these genomic sequences to a reference SARS-CoV-2 genome, numerous mutation sites are identified and stored in public databases. We downloaded the following data files from 2019 Novel Coronavirus Resource at China National Center for Bioinformation (https://bigd.big.ac.cn/ncov/release_genome).

1. **VCF file** from https://bigd.big.ac.cn/ncov/variation/statistics?lang=en. File name "2019-nCoV_total.vcf". VCF (Variant Call Format) file contains meta-information lines, header lines, and then data lines (rows) each containing information about a mutation in the genome. The columns contain genotype information on samples for each position. The downloaded file contains 10,261 non-header rows (each corresponding to specific mutation in the genome) and 20,755 columns, of which first 9 columns are mutation information and the rest of the columns contain genotype information for 20,746 virus samples. Supplement Table 1 provides an example of top-ranking mutations, and their genotype information for two samples (columns 10 and 11).

2. **Variant Annotation file** from https://bigd.big.ac.cn/ncov/variation/annotation. File name "Variation Annotation.xls". This file contains the genomic annotations of the identified mutations, such as a) genomic position, b) gene name or region in which the mutation is located, c) Number of viruses with the mutation, d) Annotation type – missense, synonymous or intergenic variant, etc., e) Mutation type – SNP, insertion or deletion, etc., and f) Protein position and amino acid change, etc.

In particular, we focused our analysis of missense variants – those genomic variants that alter the encoded amino acid sequences; because study of proteins is key to understanding the viral spread and successful development of vaccines and neutralizing antibodies. We choose four countries/regions based on the wide variations in infection and death rates. The four regions are –1) USA-NY, the epicenter in the United States; 2) China, where the pandemic originated; 3) Spain and Italy, two epicenters in Europe; and 4) India, where the world's biggest coronavirus lockdown measures were strictly implemented.

## 3. Methodology and Computational Framework

This Section describes the methodology and computational processing used in this analysis. We applied advanced tree-based ensemble learning algorithm – Random Forests (Breiman 2001) for building the classification model for discriminating the virus lineages of four geographical locations. Since RF results its output in a 'black-box' model, we applied Classification and Regression Tress (CART) methodology on selected feature sets due to its key advantage in terms of interpretability (James et al. 2013).

***Random forest:*** Fandom forest is a collection of tree structured classifiers $\{h(x, \Theta_k), k = 1, \ldots\}$, where the $\Theta_k$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input *x*. For a given ensemble of classifiers, with the training set drawn at random from the distribution of the random vector *Y,X*, the margin function is defined as,

$$mg(\boldsymbol{X}, Y) = av_k I(h_k(\boldsymbol{X}) = Y) - max_{j \neq Y}\ av_k I(h_k(\boldsymbol{X}) = j),$$

where $I(.)$ is the indicator function. The confidence in the classification is directly proportional to the margin, as the margin measures the extent to which the average number of votes at $\boldsymbol{X,Y}$ for the right class exceeds the average vote for any other class. Each tree is constructed using a different bootstrap sample from the original data where about one-third (33%) of the cases are left out of the bootstrap sample and not used in the construction of the *k-th* tree. These left out samples, usually called "out-of-bag" data, is used to get a running unbiased estimate of the classification error as trees are added to the forest. Thus, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, let *j* to be the class that got most of the votes every time case/sample *n* was out-of-bag. The proportion of times that *j* is not equal to the true class of *n* averaged over all cases is the out-of-bag error estimate. At each node, only a subset of the possible predictors are used, primarily for reducing the correlation between trees and improving the accuracy of classification (Breiman and Cutler 2001).

***CART and rpart:*** CART is a tree-based algorithm that works via recursive partitioning of the training set in order to obtain multiple subsets that are closest (or as homogeneous as possible) to a given target class (Breiman 1984). At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the dependent (predicted) variable.

We describe the computational pipeline (Supplementary Figure 1) in the following two steps:

***Step 1*** (*Data processing step*): We wrote a Perl program to scan the two downloaded files for extracting the geographic information (from sample IDs) and selecting only the missense mutations with minor allele frequency greater than a certain cut-off. For each mutation site, major allele is the one with the highest count and minor allele is the one with the second highest count. In other words, *Minor Allele Frequency* (MAF) is the frequency at which the second most common allele occurs in a given population. We included only those missense variants with minor allele count greater than 10. This step prepares the data in tab-separated tabular form for statistical analysis in R programing environment.

***Step 2*** (*Variable Selection and Classification Steps*): Prior to building the final classification model using RF, we applied a variable selection algorithm (Diaz-Uriarte 2007) to select a

small set of important non-redundant mutations. Feature selection was done using an RF based algorithm, that uses hybrid approach of selecting the virus mutations (predictor variables) based on importance spectrum (similar to scree plot) and backward variable elimination (for the selection of small sets of non-redundant variables) by changing parameters related to trees and iteration. Using 80,000 initial trees and a fractional variable drop of 10%, we finally selected 42 mutations as the most discriminative variables between the four geographic regions (classes) and created the RF classifier for viral lineage prediction with high accuracy. The accuracy of the classifier was determined by RF's cross-validation analysis (out-of-bag approach) and testing on independently set aside dataset. We split the data into 80:20 ratio and classification model generated from the training set (80%) was applied to the test set (20%) to strengthen the accuracy claims. We then developed CART model for better visualization of mutation differences among classes. For developing CART model, we used the important features as selected by RF. Analyses were done using custom scripts in R and libraries including, *randomForest* and *varSelRF*. Recursive Partitioning and Regression Trees (rpart), an R implementation of the CART algorithm, is used for developing CART model in this study. *randomForest* library provides an R interface to the Fortran programs (available at https://www.stat.berkeley.edu/~breiman/RandomForests/). *varSelRF* library is used for selecting a small set of mutations while preserving Random Forest classification accuracy.

## 4. Results

We downloaded mutation profile of 10,261 mutations in 20,746 SARS-CoV-2 samples as described in Section 3. After selecting only missense variants that show variation in at least 10 samples, we retained 588 missense mutations. Further, we selected a total of 2,927 samples that correspond to four geographic locations. Data was divided into training (2,341) and testing (586) sets based on number of samples, maintaining the class ratios. In Table 1, we list the top 50 mutations observed among all the sequenced viral samples. USA-NY samples showed highest mutation rate, suggesting that coronavirus was probably circulating undetected in the NY area for quite some time. Additionally, we found that the top four most common mutations showed significantly higher mutation rates in USA-NY samples than the other geographical regions, including rest of the world samples. For example, two of the mutations – one at genomic location 23403 (A mutated to G) and the other at 25563 (G mutated to T) – alter amino acids QHD43416.1:D614G and QHD43417.1:Q57H, respectively, in the S and ORF3a proteins of the SARS-Cov-2 virus. For the virus to break into a human cell (host cell), the S protein of the virus binds to ACE2 (angiotensin converting enzyme 2) protein on the human cell surface. The D614G mutation in S protein might change the protein structure so that it binds to the target enzyme (ACE2) with different affinity than the other lineage proteins (Amin et al. 2020). Similarly, the Q57H mutation in ORF3a protein might change important functional domains linked to virulence, infectivity, ion channel formation, and virus release (Issa et al. 2020). We speculate that this may partly explain why the infection rate is much higher in USA-NY area than other geographic locations.

Next, we built a four-class classification model for discriminating the SARS-CoV-2 samples grouped according to the four geographic locations, by specifying the country/region as factor variable $\boldsymbol{y}$ and mutation profile as predictor variable matrix $\boldsymbol{X}$ (a binary matrix, where 1 and 0 denote presence and absence of the mutation respectively). The accuracy of the finally selected classifier based on cross-validation analysis (out-of-bag approach) is 85%. Table 2 shows the confusion matrix for the final model and Figure 1 shows corresponding AUC. The robustness of developed model was then confirmed on the test data with high accuracy. Table 3 shows evaluation metrics of the model on training and test data. USA-NY and China classes

have shown the best classification accuracy, followed by Italy-Spain class. The least accuracy was observed for India class, which could be due to small sample size of that class. However, we speculate that the misclassification of 26.8% and 10.3% of India class into Italy-Spain and USA classes, respectively, might be due the viral samples from the infected patients who traveled back to India from those geographic regions and not due to local spread of the virus.

Class 4 (India) and Class 2 (China) showed better Specificity and PPV and lower FDR than the other two classes. DOR, ratio of the odds of positivity was also higher for Indian and Chinese lineages than the other two geographical regions. DOR depends significantly on the sensitivity and specificity of a test. A test with high specificity and sensitivity with low rate of false positives and false negatives has high DOR. A diagnostic odds ratio of 1 is similar to an AUC of 0.5 and represents an uninformative test. Higher values for both reflect usefulness of the classification model. Youden index analysis is useful in finding the optimal cutoff value. The value provides the best tradeoff between sensitivity and specificity and is highest for Class 2. F-score, which combines precision with recall is a good measure of goodness of model and shows high value for the current model predictions. Similarly, greater the value of positive likelihood ratio (PLR) for a particular test, the more likely a positive test result be a true positive. A good classifier should have high PLR and low Negative Likelihood Ratio (NLR). Matthews correlation coefficient (MCC), initially developed for binary classifier, considers true and false positives and negatives and is regarded as a balanced measure, which can be used even if the classes are of very different sizes. Optimized precision, a novel metrics used to negate the detrimental effects of using Precision (P) for performance evaluation of unbalanced data, also shows high values for all classes. The evaluation results from training model clearly indicate that the algorithm effectively distinguished the samples from the four regions based on the mutation profile of SARS-CoV-2. Results from testing the classifiers on test data agree with the cross-validation results and support the accurate predictive performance of the classification model.

The results from CART analysis are shown in Table 4 and Figure 2. Although, the CART model is not as accurate as the random forest model, it provided a better visualization of the associations between viral samples/lineages in different geographic regions and the mutation patterns. For example, the final CART model (Figure 3) found that the presence of mutations at genomic locations 1059_C and 17747_C, but not at genomic locations 14408_C, 13730_C, 9477_T and 11083_G classified majority of USA-NY samples from the rest in one branch of the tree. Similarly, mutation at genomic location 13730_C, but not at 1059_C, 14408_C classified majority of India samples in one branch of the tree. Interestingly, most of the missense mutations in the model alter the amino acids encoded by gene *orf1ab*. This gene encodes 16 non-structural poly-proteins (Nsp1-Nsp16) of viral RNA synthesis complex (Kirchdoerfer and Ward 2019). We expect that these results will help prioritization of select mutations, and study of their effect on SARS-Cov-2 and Human protein-protein interactions through focused experimental validations.

**Table 1: Top 50 Mutations in the four geographic regions and the rest of the world.** Column 2 – Genomic location of the mutation and the reference allele; Column 3 – Gene location of the missense mutation; Columns 4-8 – Percentage of samples mutation observed in each geographic location; Column 9 – Range (difference of highest and lowest mutation percentages among the four geographic locations).

| | | | Percentage of mutated samples/isolates in (Total number of samples in the parentheses) | | | | | |
|---|---|---|---|---|---|---|---|---|
| Mutation rank | Genomic location of the mutation (Ref Allele) | Gene name | Rest of the World (17816) | USA-NY (1243) | China (656) | Italy, Spain (534) | India (494) | Range |
| 1 | 23403 (A) | S | 74.45 | **93.24** | 5.34 | 65.92 | 62.35 | 87.91 |
| 2 | 14408 (C) | orf1ab | 74.25 | **93.24** | 3.66 | 64.42 | 62.35 | 89.58 |
| 3 | 25563 (G) | ORF3a | 28.37 | **86.32** | 0.91 | 1.69 | 23.08 | 85.41 |
| 4 | 1059 (C) | orf1ab | 21.57 | **79.00** | 0.91 | 1.31 | 0.61 | 78.40 |
| 5 | 28881 (G) | N | 24.36 | 4.02 | 1.52 | 17.04 | 12.15 | 15.52 |
| 6 | 28882 (G) | N | 24.33 | 4.10 | 1.22 | 16.67 | 12.15 | 15.45 |
| 7 | 28883 (G) | N | 24.29 | 4.10 | 1.22 | 16.85 | 12.15 | 15.63 |
| 8 | 28144 (T) | ORF8 | 10.22 | 2.65 | 31.10 | 27.53 | 4.25 | 28.44 |
| 9 | 11083 (G) | orf1ab | 9.65 | 4.67 | 5.95 | 6.55 | **34.01** | 29.34 |
| 10 | 17858 (A) | orf1ab | 6.93 | 1.21 | 0.00 | 0.00 | 0.20 | 1.21 |
| 11 | 17747 (C) | orf1ab | 6.87 | 1.21 | 0.00 | 0.00 | 0.20 | 1.21 |
| 12 | 26144 (G) | ORF3a | 5.96 | 3.78 | 5.03 | 4.87 | 0.61 | 4.42 |
| 13 | 27964 (C) | ORF8 | 3.09 | 0.56 | 0.00 | 0.00 | 0.00 | 0.56 |
| 14 | 2558 (C) | orf1ab | 2.23 | 0.48 | 0.15 | 0.00 | 0.61 | 0.61 |
| 15 | 28854 (C) | N | 1.82 | 0.00 | 1.37 | 0.00 | **14.17** | 14.17 |
| 16 | 13730 (C) | orf1ab | 1.36 | 0.08 | 0.46 | 0.19 | **30.16** | 30.08 |
| 17 | 28311 (C) | N | 1.40 | 0.00 | 0.46 | 0.19 | **28.95** | 28.95 |
| 18 | 27046 (C) | M | 2.09 | 0.16 | 0.00 | 1.12 | 0.00 | 1.12 |
| 19 | 2480 (A) | orf1ab | 2.07 | 0.48 | 0.15 | 0.00 | 0.40 | 0.48 |
| 20 | 6312 (C) | orf1ab | 1.25 | 0.00 | 0.46 | 0.19 | **28.14** | 28.14 |
| 21 | 11916 (C) | orf1ab | 0.74 | 17.62 | 0.00 | 0.00 | 0.00 | 17.62 |
| 22 | 10097 (G) | orf1ab | 1.96 | 0.08 | 0.00 | 0.00 | 0.00 | 0.08 |
| 23 | 25429 (G) | ORF3a | 1.80 | 0.08 | 0.00 | 0.19 | 0.20 | 0.20 |
| 24 | 28077 (G) | ORF8 | 1.55 | 0.72 | 1.52 | 0.00 | 0.81 | 1.52 |
| 25 | 1440 (G) | orf1ab | 1.48 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 26 | 2891 (G) | orf1ab | 1.46 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 27 | 26530 (A) | M | 1.32 | 0.24 | 0.00 | 1.12 | 0.61 | 1.12 |
| 28 | 18998 (C) | orf1ab | 0.34 | 14.32 | 0.00 | 0.00 | 0.00 | 14.32 |
| 29 | 3177 (C) | orf1ab | 1.27 | 0.40 | 0.15 | 0.00 | 0.00 | 0.40 |
| 30 | 490 (T) | orf1ab | 1.22 | 0.40 | 0.15 | 0.00 | 0.00 | 0.40 |
| 31 | 28863 (C) | N | 0.63 | 0.24 | 0.15 | 19.66 | 0.20 | 19.51 |
| 32 | 1397 (G) | orf1ab | 0.98 | 0.16 | 2.90 | 0.94 | 3.85 | 3.69 |
| 33 | 9477 (T) | orf1ab | 0.59 | 0.24 | 0.15 | 20.04 | 0.20 | 19.89 |
| 34 | 18736 (T) | orf1ab | 1.19 | 0.40 | 0.00 | 0.00 | 0.00 | 0.40 |
| 35 | 25979 (G) | ORF3a | 0.61 | 0.16 | 0.15 | 19.10 | 0.20 | 18.95 |
| 36 | 11109 (C) | orf1ab | 1.20 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 37 | 6310 (C) | orf1ab | 0.82 | 0.00 | 0.00 | 0.00 | 9.51 | 9.51 |
| 38 | 4002 (C) | orf1ab | 1.03 | 0.08 | 0.00 | 0.00 | 0.00 | 0.08 |
| 39 | 28836 (C) | N | 0.95 | 0.16 | 0.00 | 0.00 | 0.00 | 0.16 |
| 40 | 13862 (C) | orf1ab | 0.94 | 0.00 | 0.00 | 0.00 | 0.40 | 0.40 |
| 41 | 24368 (G) | S | 0.88 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 42 | 21575 (C) | S | 0.69 | 1.13 | 0.00 | 0.37 | 0.61 | 1.13 |
| 43 | 28878 (G) | N | 0.55 | 0.72 | 1.22 | 0.75 | 3.85 | 3.12 |
| 44 | 16289 (C) | orf1ab | 0.72 | 0.16 | 0.00 | 0.19 | 0.00 | 0.19 |
| 45 | 25688 (C) | ORF3a | 0.68 | 0.32 | 0.00 | 0.19 | 0.00 | 0.32 |
| 46 | 10323 (A) | orf1ab | 0.68 | 0.08 | 0.30 | 0.19 | 0.00 | 0.30 |
| 47 | 10798 (C) | orf1ab | 0.70 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 48 | 25350 (C) | S | 0.66 | 0.00 | 0.00 | 0.00 | 0.20 | 0.20 |
| 49 | 28580 (G) | N | 0.66 | 0.00 | 0.00 | 0.19 | 0.00 | 0.19 |
| 50 | 1302 (C) | orf1ab | 0.66 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

While the Random Forest classification method used here is considered as a "black box" method, with no interpretable classification model, the method provides useful information, such as variable importance. One of the measures of variable importance in Random Forest method is the mean decrease in accuracy, calculated using the out-of-bag sample. The difference between the prediction accuracy on the untouched out-of-bag sample and that on the out-of-bag sample permuted on one predictor variable is averaged over all trees in the forest and normalized by the standard error. This gives the mean decrease in accuracy of that particular predictor variable which has been permuted. Figure 3 shows the list of feature variables ranked according to mean decrease in accuracy of classification. It is interesting to note that the mutations in genes orf1ab, ORF3A and S genes rank among the most discriminative variables from mean decrease in accuracy graph (Figure 3).

**Table 2: Confusion matrix of the prediction results of fitted model on Training set**

| | | **True Class** – Number of real samples in each class | | | | |
|---|---|---|---|---|---|---|
| | | USA-NY (Class 1) | China (Class 2) | Italy, Spain (Class 3) | India (Class 4) | Total |
| **Predicted Class** – Number of predicted samples in each class | USA-NY | 907 (91.25%) | 15 (1.51%) | 72 (7.24%) | 0 (0.00%) | 994 |
| | China | 9 (1.71%) | 497 (94.67%) | 16 (3.05%) | 3 (0.57%) | 525 |
| | Italy, Spain | 20 (4.68%) | 42 (9.84%) | 362 (84.78%) | 3 (0.70%) | 427 |
| | India | 41 (10.38%) | 23 (5.82%) | 106 (26.84%) | 225 (56.96%) | 395 |
| | Total | 977 | 577 | 556 | 231 | |

**Table 3: Evaluation metrics of the four-class classification model based on cross-validation and independent test data. Bold numbers represent best performance among classes in training and test set.** NPV – Negative Predictive Value; FDR – False Detection Rate; FNR – False Negative Rate; DRO – Diagnostic Odds Ratio; PLR – Positive Likelihood Ratio; NLR – Negative Likelihood Ratio; MCC – Matthews correlation coefficient.

| Metric | Cross validation | | | | Test data | | | |
|---|---|---|---|---|---|---|---|---|
| | USA (Class 1) | China (Class 2) | Europe (Class 3) | India (Class 4) | USA (Class 1) | China (Class 2) | Europe (Class 3) | India (Class 4) |
| Balanced accuracy | 0.93 | **0.95** | 0.87 | 0.78 | 0.91 | **0.94** | 0.83 | 0.82 |
| Sensitivity or Recall | 0.91 | **0.95** | 0.85 | 0.57 | 0.91 | **0.95** | 0.74 | 0.65 |
| Specificity | 0.94 | 0.95 | 0.89 | **1.00** | 0.92 | 0.94 | 0.91 | **0.99** |
| PPV or Precision | 0.93 | 0.86 | 0.65 | **0.97** | 0.90 | 0.83 | 0.66 | **0.94** |
| NPV | 0.93 | **0.98** | 0.96 | 0.91 | 0.92 | **0.98** | 0.94 | 0.92 |
| FDR | 0.07 | 0.14 | 0.35 | **0.03** | 0.10 | 0.17 | 0.34 | **0.06** |
| FNR | 0.09 | **0.05** | 0.15 | 0.43 | 0.09 | **0.05** | 0.26 | 0.35 |
| False Omi-ssion Rate | 0.07 | **0.02** | 0.04 | 0.09 | 0.08 | **0.02** | 0.06 | 0.08 |
| False Positive Rate | 0.06 | 0.05 | 0.11 | **0.00** | 0.08 | 0.06 | 0.09 | **0.01** |
| DRO | 161.44 | 331.48 | 46.76 | **389.56** | 109.32 | **261.46** | 29.20 | 196.11 |
| Youden's Index | 0.85 | **0.90** | 0.74 | 0.57 | 0.83 | **0.88** | 0.65 | 0.64 |
| Geometric Mean | 0.93 | **0.95** | 0.87 | 0.75 | 0.91 | **0.94** | 0.82 | 0.80 |
| F-score (beta 0.5) | **0.93** | 0.88 | 0.68 | 0.85 | **0.90** | 0.85 | 0.68 | 0.86 |
| F-score (beta 1) | **0.92** | 0.90 | 0.74 | 0.72 | **0.91** | 0.89 | 0.70 | 0.77 |
| F-score (beta 2) | 0.92 | **0.93** | 0.80 | 0.62 | **0.91** | 0.92 | 0.72 | 0.69 |
| PLR | 15.04 | 18.63 | 7.97 | **168.23** | 11.01 | 14.92 | 8.38 | **69.98** |
| NLR | 0.09 | **0.06** | 0.17 | 0.43 | 0.10 | **0.06** | 0.29 | 0.36 |
| MCC | 0.85 | **0.87** | 0.67 | 0.71 | 0.82 | **0.85** | 0.62 | 0.74 |
| Markedness | 0.85 | **0.87** | 0.67 | 0.71 | 0.82 | **0.85** | 0.62 | 0.74 |
| Optimization Precision | 0.84 | **0.85** | 0.82 | 0.58 | **0.84** | **0.84** | 0.74 | 0.63 |

**Table 4: Confusion matrix of the prediction results using CART on training set**

| | | **True Class** – Number of real samples in each class | | | | |
|---|---|---|---|---|---|---|
| | | USA-NY (Class 1) | China (Class 2) | Italy, Spain (Class 3) | India (Class 4) | Total |
| **Predicted Class** – Number of predicted samples in each class | USA-NY | 914 (91.95%) | 9 (0.91%) | 65 (6.54%) | 6 (0.60%) | 994 |
| | China | 10 (1.90%) | 489 (93.14%) | 16 (3.05%) | 10 (1.90%) | 525 |
| | Italy, Spain | 22 (5.15%) | 38 (8.90%) | 363 (85.01%) | 4 (0.94%) | 427 |
| | India | 41 (10.38%) | 10 (2.53%) | 115 (29.11%) | 229 (57.97%) | 395 |
| | Total | 987 | 546 | 559 | 249 | |

## 5. Conclusions

In theory, accumulation of mutations could make a virus more infectious or deadly, or vice versa, but the vast majority of mutations do not affect a virus's performance. While some mutations lead to more virulent and lethal strains, other mutations make the virus less infectious and less lethal in the populations. Computational methods that effectively integrate genomic profiles to identify and prioritize important genomic features and classify different groups of samples are valuable tools for Bioinformatics researchers. SARS-CoV-2 related research is rapidly evolving with numerous publications. Phylogenetic methods have been applied to SARS-CoV-2 genome sequences to construct the phylogenetic trees (clusters of closely related lineages) and predict future global hot spots of disease transmission and surge (Forster et al. 2020). Similarly, analysis pipelines are being developed for analysis of SARS-CoV-2 genomes to facilitate identification of novel mutations (Pachetti et al. 2020) and for functional annotations of mutations in specific gene regions, for example, nonsynonymous mutations in the ORF3a protein (Issa et al. 2020). Here, we have developed a complementary computational pipeline based on Random Forest based classification methods to identify a subset of missense mutations that can classify groups of virus lineages. It was previously reported, based on analysis of 220 genomic sequences, that the mutations located at positions 2891, 3036, 14408, 23403 and 28881 positions were predominantly observed in Europe, whereas those located at positions 17746, 17857 and 18060 were exclusively present in North America (Pachetti et al. 2020). However, we found that the top-ranking mutations located at positions 14408 and 23403 were most frequent in USA-NY samples than the rest of the geographical regions. We believe that this contradictory result could be due to much bigger sample size and small geographic regions in our analysis. Our findings suggest that the virus is evolving locally, and presence of small geographic region-specific strains that could be accurately classified by different mutational patterns.

Random Forest based algorithms have been successfully applied in various genomic analysis studies. For example, we have earlier used an integrative modeling approach that combines CART (Breiman 1984) and Random Forest to classify different estrogen receptor alpha responsive promoters (Cheng et al. 2006) and SMAD target promoters (Qin et al. 2009) with reasonably good classification accuracy and reduced instability (Qin et al. 2009). Although the main goal in classification is to build a model with minimal misclassification error in cross-validation, in these applications we are equally interested in identifying

biologically important features, such as genomic mutations or single nucleotide polymorphisms, for future experimental prioritization. The computational pipeline presented here will help the discovery of geographic specific SARS-CoV-2 mutations for further computational modeling and experimental validations and help in the interpretation of their functional effects.

**Acknowledgements**

This work was supported by the National Library of Medicine of the NIH [R01LM011297 to RD]. We thank the reviewer and editor for their suggestions and thoughtful comments, which substantially helped the revised version.

**References**

Amin, M., Sorour, M. K. and Kasry, A. (2020). Comparing the Binding Interactions in the Receptor Binding Domains of SARS-CoV-2 and SARS-CoV. *Journal of Physical Chemistry Letters*, **11**, 4897-4900. doi:10.1021/acs.jpclett.0c01064.

Breiman, L. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, California.

Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5-32.

Breiman, L. and Cutler, A. (2001). Random Forests . https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

Cheng, A. S., Jin, V. X., Fan, M., Smith, L. T. et al. (2006). Combinatorial analysis of transcription factor partners reveals recruitment of c-MYC to estrogen receptor-alpha responsive promoters. *Molecular Cell*, **21**, 393-404.

Colson, P., Lagier, J. C., Baudoin, J. P., Bou, Khalil J., La Scola, B. and Raoult, D. (2020). Ultrarapid diagnosis, microscope imaging, genome sequencing, and culture isolation of SARS-CoV-2. *European Journal of Clinical Microbiology and Infectious Diseases*, 1-3. doi:10.1007/s10096-020-03869-w.

Diaz-Uriarte, R. (2007). GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics* **8**, 328.

Forster, P., Forster, L., Renfrew, C. and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences USA*, **117**, 9241-9243.

Holland, L. A., Kaelin, E. A., Maqsood, R., Estifanos, B., Wu, L. I., Varsani, A., Halden, R. U., Hogue, B. G., Scotch, M. and Lim, E. S. (2020). An 81 nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (Jan-Mar 2020). *Journal of Virology*. doi:10.1128/JVI.00711-20.

Issa, E., Merhi, G., Panossian, B., Salloum, T. and Tokajian, S. (2020). SARS-CoV-2 and ORF3a: Nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems*, **5**.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning*: *With Applications in R*. Springer, New York.

Kirchdoerfer, R. N. and Ward, A. B. (2019). Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications*, **10**, 2342.

Lu, I. N., Muller, C. P. and He, F. Q. (2020a). Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies. *Virus Research*, **283**, 197963.

Lu, R., Zhao, X., Li, J., Niu, P. et al. (2020b). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, **395**, 565-574.

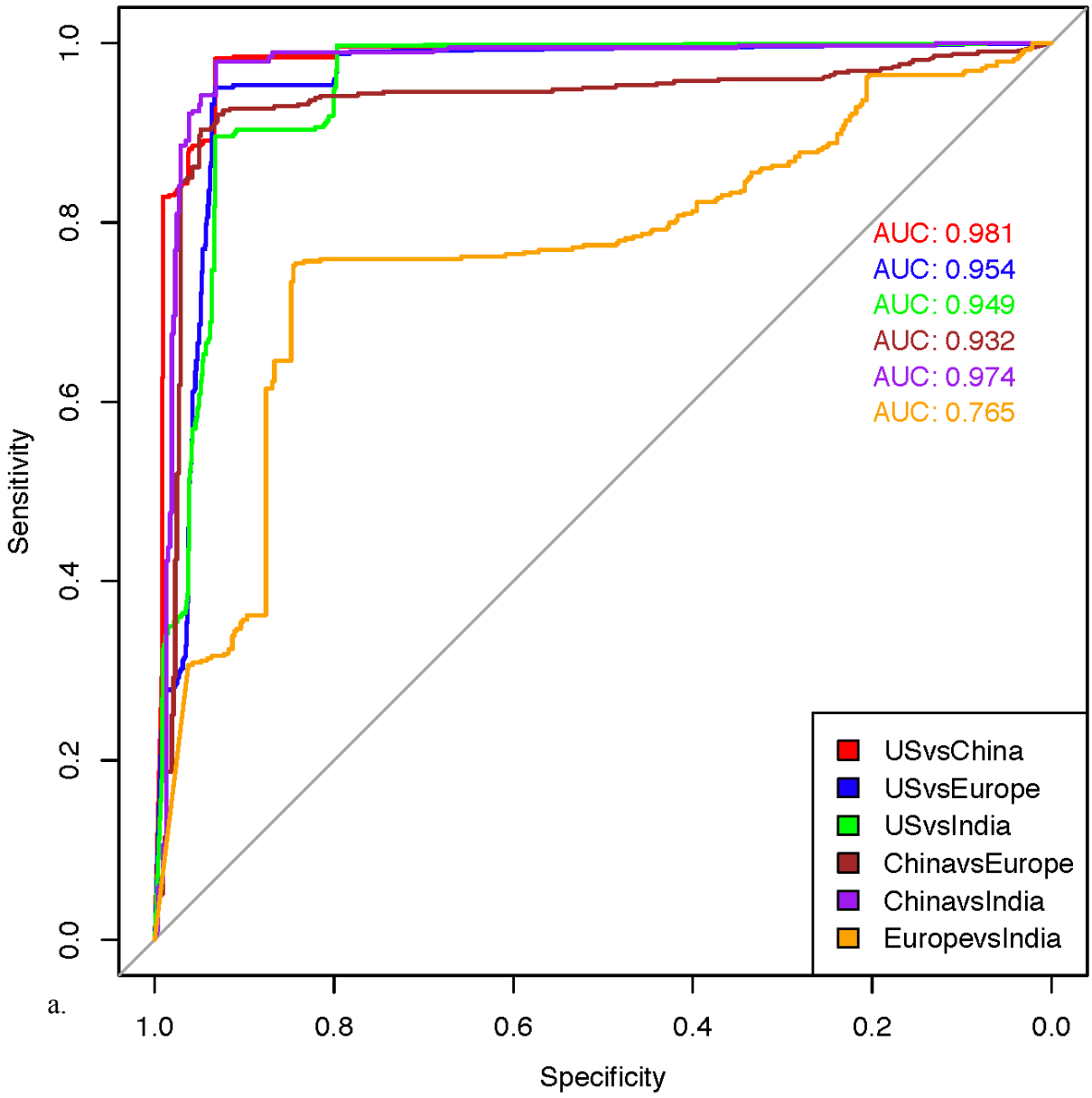
Pachetti, M., Marini, B., Benedetti, F., Giudici, F. et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*, **18**, 179.

Qin, H., Chan, M. W., Liyanarachchi, S., Balch, C. et al. (2009). An integrative ChIP-chip and gene expression profiling to model SMAD regulatory modules. *BMC System Biology*, **3**, 73.

Yadav, P.D., Potdar, V.A., Choudhary, M. L., Nyayanit, D. A. et al. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian Journal of Medical Research*, **151**, 200-209.

Zhao, X., Sehgal, M., Hou, Z., Cheng, J. et al. (2018). Identification of residues controlling restriction versus enhancing activities of IFITM proteins on entry of human coronaviruses. *Journal of Virology*, 92.

**Figure 1: ROC curve between classes for (a) training set** USA-NY (Class 1); China (Class 2); Italy, Spain (Class 3); India (Class 4)




a.

**Figure 2: Model features and their importance**

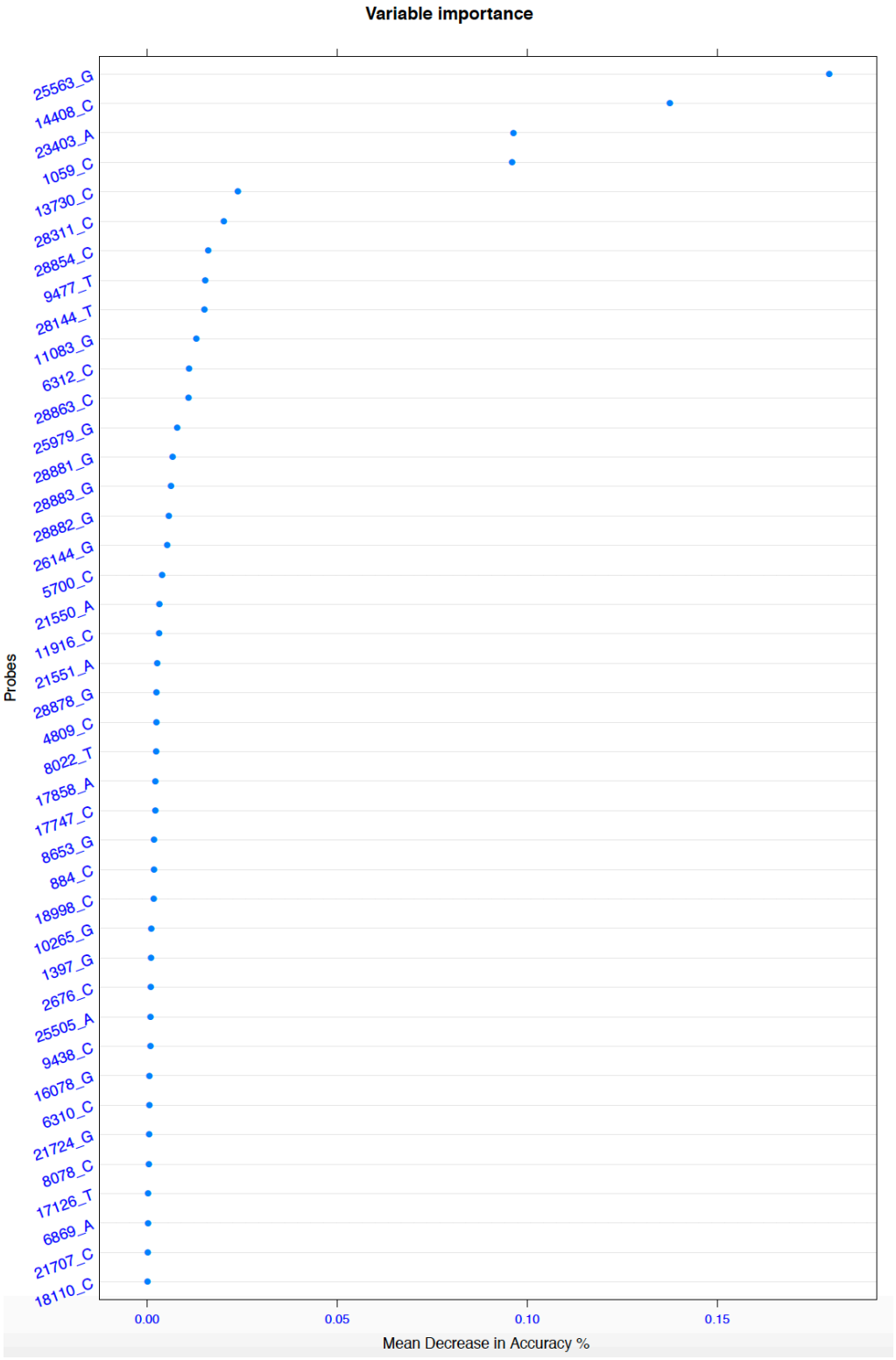
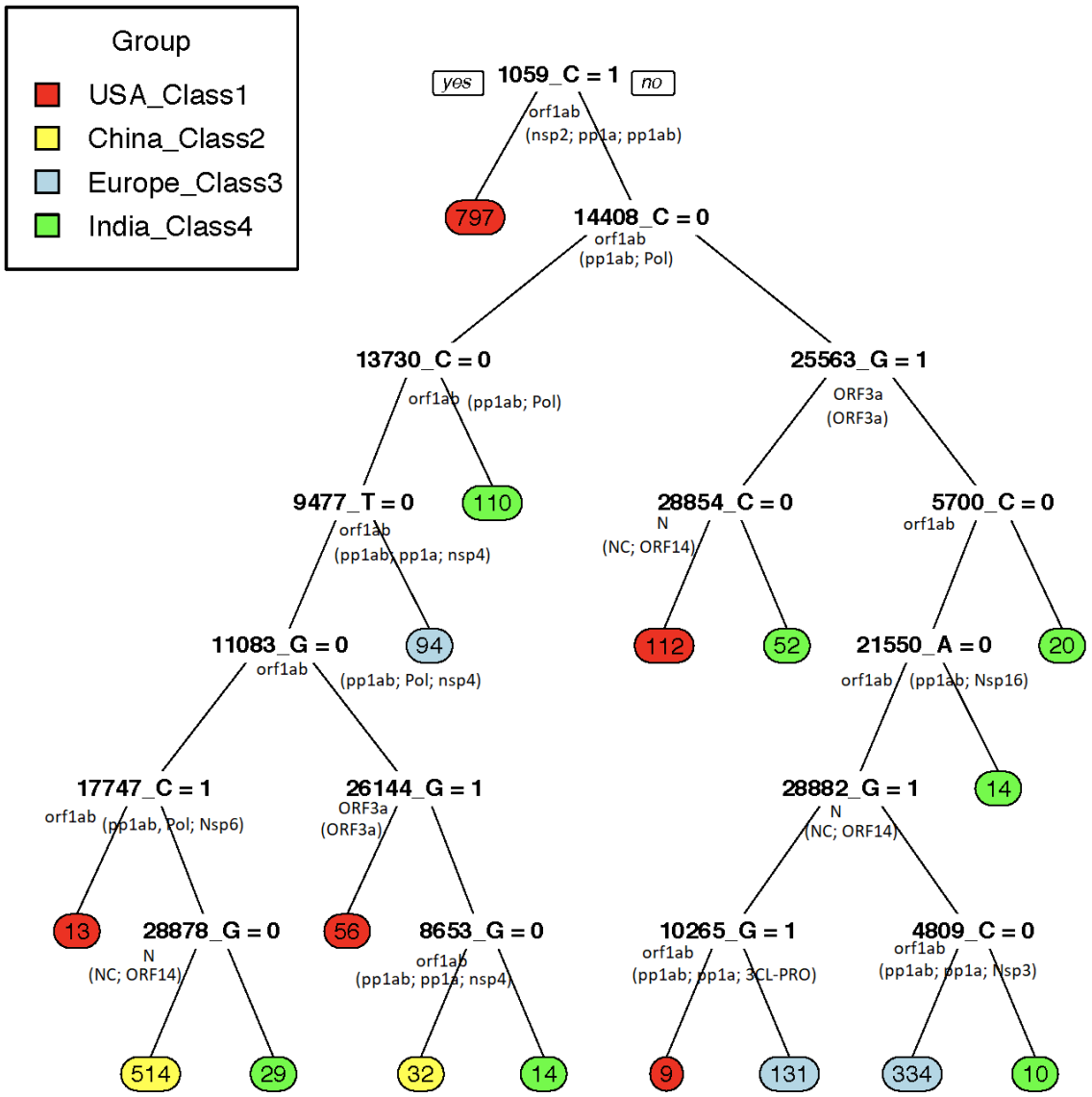


**Figure 3: Pruned Tree representation of CART model, generated using 42 features selected by Random forest feature selection method.** The gene name and UniProt Protein Products or Polypeptide Chains (in parentheses) in which the mutation is located is mentioned at the bottom of each mutation in the tree.



**Supplement Table 1: List of top ranking (top 10) mutations.** First four rows are header lines; fifth line is for column headings. Columns 1 to 9 provide mutation information, such as the chromosome (CHROM), genomic position (POS), unique identifier (ID), reference allele (REF), alternative alleles (ALT) identified in different lineages, sequence quality score (QUAL), filtering out (FILT) criteria for low quality mutations, any information (INFO) and format of the mutation, GT – Genotype. Genotype data are given for two samples, one for USA and the other from India. Missing information is denoted by period "." symbol. If more than one alternative alleles exist, those are comma-separated in ALT column. The nucleotide symbols in REF and ALT columns are: A – Adenine; C – Cytosine; G – Guanine; T – Thymine; R – G or A (purine); Y – C or T (pyrimidine); K – G or T; M – A or C; S – G or C; W – A or T; B – G or T or C; D – G or A or T; H – A or C or T; V – G or C or A.

##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##reference=file:///xtdisk/apod/licp/Virus/ref/2019-nCoV.fa
##contig=<ID=2019-nCoV,length=29903>

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | 2019-nCoV/USA-AZ1/2020 | SARS-CoV-2/human/IND/GMCKN318/2020 |
|---|---|---|---|---|---|---|---|---|---|---|
| 2019-nCoV | 23403 | 2019-nCoV_23403 | A | R,G | . | . | . | GT | 0 | 2 |
| 2019-nCoV | 14408 | 2019-nCoV_14408 | C | A,T,Y | . | . | . | GT | 0 | 2 |
| 2019-nCoV | 3037 | 2019-nCoV_3037 | C | T,Y | . | . | . | GT | 0 | 1 |
| 2019-nCoV | 241 | 2019-nCoV_241 | C | T,Y | . | . | . | GT | 0 | 1 |
| 2019-nCoV | 25563 | 2019-nCoV_25563 | G | T,C,R,K | . | . | . | GT | 0 | 0 |
| 2019-nCoV | 1059 | 2019-nCoV_1059 | C | T,Y | . | . | . | GT | 0 | 0 |
| 2019-nCoV | 28881 | 2019-nCoV_28881 | G | A,T,R | . | . | . | GT | 0 | 0 |
| 2019-nCoV | 28882 | 2019-nCoV_28882 | G | A,T,R | . | . | . | GT | 0 | 0 |
| 2019-nCoV | 28883 | 2019-nCoV_28883 | G | A,S,C | . | . | . | GT | 0 | 0 |
| 2019-nCoV | 8782 | 2019-nCoV_8782 | C | T,Y | . | . | . | GT | 1 | 0 |

**Supplementary Figure 1: Flowchart of the computational frame-work**