PredictION: a predictive model to establish the performance of Oxford sequencing reads of SARS-CoV-2

David E. Valencia-Valencia[1,*], Diana Lopez-Alvarez[1,2,3,*], Nelson Rivera-Franco[1,3], Andres Castillo[1], Johan S. Piña[4], Carlos A. Pardo[5] and Beatriz Parra[3]

[1] Laboratorio de Técnicas y Análisis Ómicos—TAOLab/CiBioFi, Facultad de Ciencias Naturales y Exactas, Universidad del Valle, Cali, Valle del Cauca, Colombia

[2] Departamento de Ciencias Biológicas, Facultad de Ciencias Agropecuarias, Universidad Nacional de Colombia, Palmira, Valle del Cauca, Colombia

[3] Grupo VIREM—Virus Emergentes y Enfermedad, Escuela de Ciencias Básicas, Facultad de Salud, Universidad del Valle, Cali, Valle del Cauca, Colombia

[4] Department of Data Science, People Contact, Manizales, Caldas, Colombia

[5] Department of Neurology, Pathology, Johns Hopkins University School of Medicine, Baltimore, MD, United States of America

[*] These authors contributed equally to this work.

ABSTRACT

The optimization of resources for research in developing countries forces us to consider strategies in the wet lab that allow the reuse of molecular biology reagents to reduce costs. In this study, we used linear regression as a method for predictive modeling of coverage depth given the number of MinION reads sequenced to define the optimum number of reads necessary to obtain >200X coverage depth with a good lineage-clade assignment of SARS-CoV-2 genomes. The research aimed to create and implement a model based on machine learning algorithms to predict different variables (*e.g.*, coverage depth) given the number of MinION reads produced by Nanopore sequencing to maximize the yield of high-quality SARS-CoV-2 genomes, determine the best sequencing runtime, and to be able to reuse the flow cell with the remaining nanopores available for sequencing in a new run. The best accuracy was −0.98 according to the R squared performance metric of the models. A demo version is available at https://genomicdashboard.herokuapp.com/.

Subjects Bioinformatics, Computational Science, Data Mining and Machine Learning, Data Science, COVID-19

Keywords Oxford nanopore technologies, Genomes, Machine learning, Sequences, Linear models

Submitted 19 July 2022
Accepted 30 October 2022
Published 30 November 2022

Corresponding authors
Diana Lopez-Alvarez, dilopezal@unal.edu.co, diana.lopez.alvarez@correounivalle.edu.co
Beatriz Parra, beatriz.parra@correounivalle.edu.co

Academic editor
Brenda Oppert

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.14425**

INTRODUCTION

The Oxford Nanopore Technologies (ONT) MinION sequencing platform provides a method for high-throughput and cost-effective long-read sequencing in a portable device (*Gauthier et al., 2021*), and has become a fast and reliable tool for epidemiological surveillance of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). During the COVID-19 pandemic ONT accelerated the production of genome data worldwide

How to cite this article Valencia-Valencia DE, Lopez-Alvarez D, Rivera-Franco N, Castillo A, Piña JS, Pardo CA, Parra B. 2022. PredictION: a predictive model to establish the performance of Oxford sequencing reads of SARS-CoV-2. *PeerJ* **10**:e14425 http://doi.org/10.7717/peerj.14425

(As of October 15, 2022 there are about 13.8 million SARS-CoV-2 genomes submitted to the Global Initiative on Sharing All Influenza Data (GISAID) database), allowing for characterization of different lineages or variants and providing an essential tool for effective health policy decision making. The most widely adopted targeted amplicon approach for SARS-CoV-2 genomic sequencing is the ARTIC protocol (*Lambisia et al., 2022*).

Despite its multiple advantages for clinical and epidemiological applications, this sequencing technology is very expensive, especially for the public health systems in developing countries. The estimated costs of reagents and consumables range between $11.50 to $35.88 for one sample when calculated based on 96 samples per sequencing run (*Lambisia et al., 2022*) plus $900 per flow cell. Therefore, optimized one-time reuse of the MinION flow cells is a feasible cost-effective alternative that provides an opportunity to perform another additional experiment using the same set of barcodes with a different type of sequencing approach/targets. The idea would be to stop sequencing once enough reads are attained for optimal genome assembly, leaving a remaining number of nanopores still available for sequencing in a new experiment. In fact, Oxford Nanopore provides flow cell wash buffers and storage buffers that facilitate the storage of second-use flow cells. The implementation of machine learning algorithms can help optimize different variables involved in obtaining a good quality SARS-CoV-2 genome from a certain number of sequenced reads. Linear Regression is a statistical model well-known in Supervised Machine Learning (SML), and it is applied to establish the relationship between a dependent variable and one or more independent variables since the algorithm is trained on both input features and output labels (*Osisanwo et al., 2017*).

In addition to generating value and exploiting data potential, data science applications should allow visualization and manipulation of the results obtained from the analysis (*Verbert et al., 2014*). Therefore, we also propose an online monitoring model that provides other researchers with a helping hand for their experiments, since the Web applications enhance the software performance, availability, and scalability (*Verbert et al., 2013*). In this study, different machine learning algorithms were implemented to make predictions of the optimal number of sequenced reads needed to obtain good coverage depth (>200X) using the MinION sequencing platform and lab-scale data for SARS-CoV-2 genomes. Additionally, we designed a Machine Learning application by loading the best model into dynamic dashboards that allow users to modify and interact with the model's inputs, apply filters, and visualize graphics interactively.

MATERIALS & METHODS

Data collection and processing

We used five variables: number of sequenced reads per genome, CT (Cycle threshold) value (N2 target gene; N is the Nucleocapside protein gene and number 2 is a second specific sequence targeted within the N gene), mean coverage depth, coverage genome (percentage), and quantification cDNA (ng/µl) for a dataset (1) that included 1461 samples without CT values, and a dataset (2) with 471 samples that was a subset of dataset 1 (Table S1) that included CT values. Data were generated in amplicon-targeted sequencing experiments for SARS-CoV-2 genomes assemblies using ARTIC Network's protocol V3 (*Tyson et al., 2020*) and a MinION device. This protocol of primer sets, and amplicons is one of the most widely used SARS-CoV-2 sequencing protocols (*Tyson et al., 2020*). To understand the meaning and the predictive power of the variables we conducted exploratory analysis before modeling, as follows: (i) The number of reads data were grouped into bins and the means (and medians) of coverage depth in each bin were compared; if coverage depth was similar across bins, the number of reads would be non-predictive. (ii) We plotted the number of reads against coverage depth. These plots were generated using Python (v3.9.9) and libraries Pandas (v1.4.4), NumPy (v1.23.3), Matplotlib (v3.5.3), and Seaborn (v0.12.0).

Model building

The primary independent variable was "sequenced reads", while the remaining variables (*e.g.*, CT, cDNA) were used to describe and proceed with the design, training, testing, and evaluation of the SML model. Both datasets were scaled using the RobustScaler function from Sklearn (v1.1.1), then the datasets were split into train (75%) and test (25%) sets. For reproducibility purposes, a random seed of 27 was set for all models that used a random state.

All models were trained with both datasets to select the best performance estimator. We use Lasso (LssR), Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), and Support Vector Regressor (SVR) as SML algorithms to predict continuous-valued outputs. LssR is a linear model regression that estimates sparse coefficients based on L1 penalization (*Cherkassky & Ma, 2003*). On the other hand, both RFR and GBR are ensemble methods that combine multiple learning algorithms to get a better performance prediction. Specifically, GBR builds an additive model in a forward stage-wise fashion and each stage fits a regression tree on the negative gradient of the given loss function. RFR fits a set number of decision trees of subsamples of a dataset to improve the predictive power. Lastly, SVR is an SML algorithm that analyses data regression based on a hyperplane and support vectors (*Smola & Schölkopf, 2004*); this method considers the points that are within the decision boundary line and fit the error within a certain threshold. The best fit line is the hyperplane that optimizes classification. Model hyperparameters were modified to achieve better performance for models. LssR was trained with different values for *alpha* regularization; those values were 200 numbers on a log scale from -10 to 3. For RFR and GBR, random seeds were set at 27. For SVR default hyperparameters were used. All algorithms were implemented using Python v3.9.9 and Sklearn (v1.1.1). Additionally, to compare the models, we used R squared as our metric for model performance with k-fold cross-validation splitting the dataset into five random folds to avoid model overfitting. Additionally, we reported the average of the five scores provided by the cross-validation process.

Taking into account that all variables are continuous, a Pearson correlation coefficient was calculated as a measure of the strength of the relationship between them; this kind of analysis was carried out on both datasets to decide which features to keep or discard according to predictive power. To aid the reproducibility of this work, the code was uploaded to an open access repository on GitHub (https://github.com/TAOLabUV/PredictION) as well as selected datasets.

Performance metrics

We assessed the efficiency of the model using the coefficient of determination (R^2) (1), mean absolute error (MAE) (2), and root mean squared error (RMSE) (3). Since errors can be both positive (actual > prediction) and negative (actual < prediction), we measure the absolute value and the squared value of each error. The R^2, MAE, and RMSE are computed as follows:

$$R^2 = \frac{\left(\sum_{i=1}^{n}(\mathrm{obs}_i - \mu_{\mathrm{obs}})(\mathrm{pred}_i - \mu_{\mathrm{pred}})\right)^2}{\sum_{i=1}^{n}(\mathrm{obs}_i - \mu_{\mathrm{obs}})^2 \sum_{i=1}^{n}(\mathrm{pred}_i - \mu_{\mathrm{pred}})^2} \quad (1)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\mathrm{obs}_i - \mathrm{pred}_i| \quad (2)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{obs}_i - \mathrm{pred}_i)^2} \quad (3)$$

where n represents the total number of sampled genomes, obs_i corresponds to the number of reads measured for a specific genome, pred_i is the predicted value of reads, i represents each individual genome, and µcorresponds to the mean. The method is shown in a flow chart in Fig. 1.

Creation of the dashboard

Since the machine learning model was trained in Python, we created a dashboard developed in this language. The source code was packaged in an image using Docker (*Anderson, 2015*) containers. This image was uploaded to an open-source cloud platform as a web application for other researchers to use. This dashboard has three sliders and a textbox for "Concentration of cDNA (ng/µl)", "Coverage depth (mean)" and "Coverage per genome (percentage)" that is updated when the user changes these objects. Also, input values are displayed in a figure with all values used for model training. Finally, users can verify the value of the prediction reads together with the local explanation graph; in this way, every time a value is modified in the slider, the prediction, and the graphs are automatically updated. The application also allows the saving of the graphs obtained with each prediction thanks to interactive buttons that are displayed on each graph. The web application can be accessed at: https://genomicdashboard.herokuapp.com/.

Ethical considerations

The study was approved by the Ethics Committee of Universidad del Valle, Colombia, with code 188-020, samples and database were anonymized.

RESULTS

Data analysis

The average number of reads in the evaluated genomes was 42.58k ± 42.9k for dataset (1) and 42.81k ± 34.93k for dataset (2) (Table S1) with a positively skewed distribution in

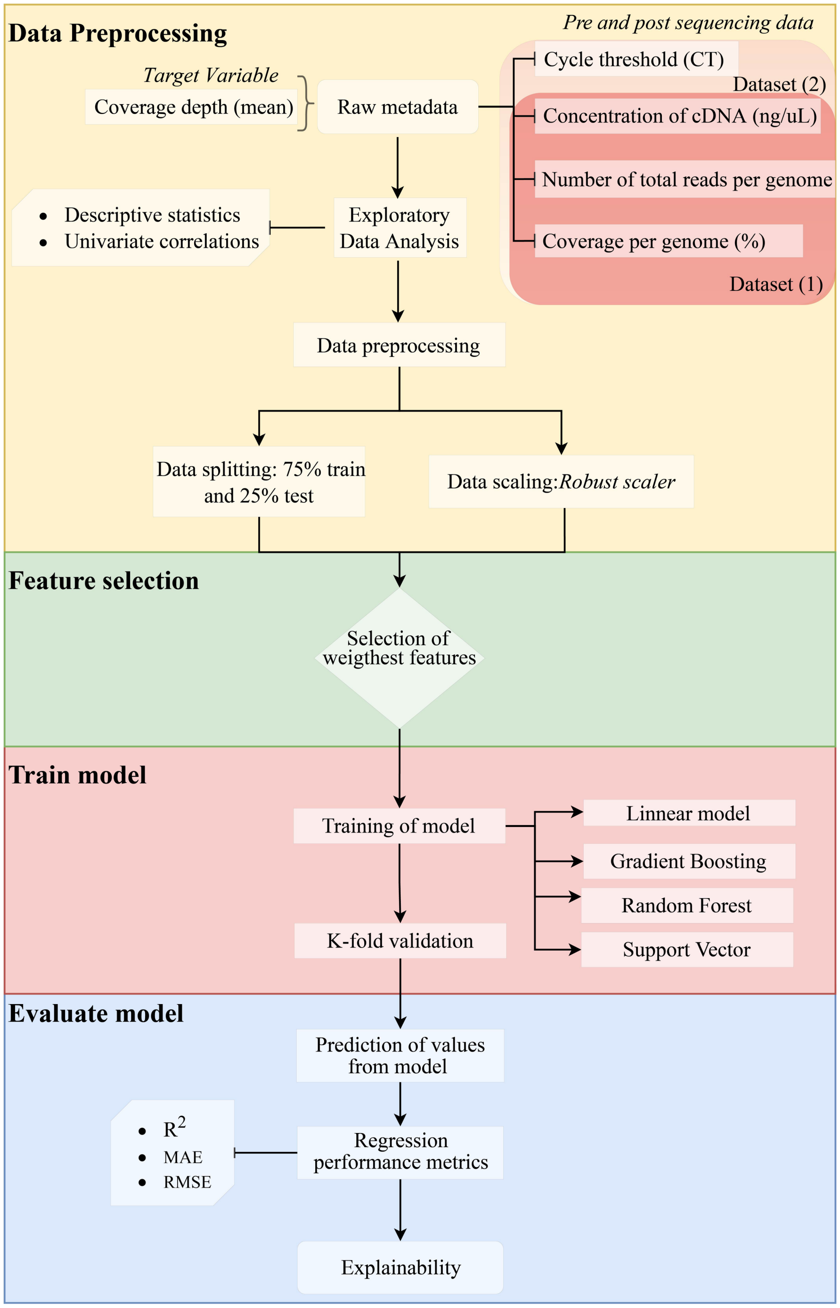

Figure 1 **Methodology flowchart for an obtained optimal model.**
Full-size DOI: 10.7717/peerj.14425/fig-1

both datasets (Figs. 2A and 2C). In dataset (1), outliers were on both sides of log e scale boxplots, mainly accumulated on the lower side, while dataset (2) has values exclusively in that space (Fig. S1). The average coverage depth per genome was 519.91 ± 563.5X and

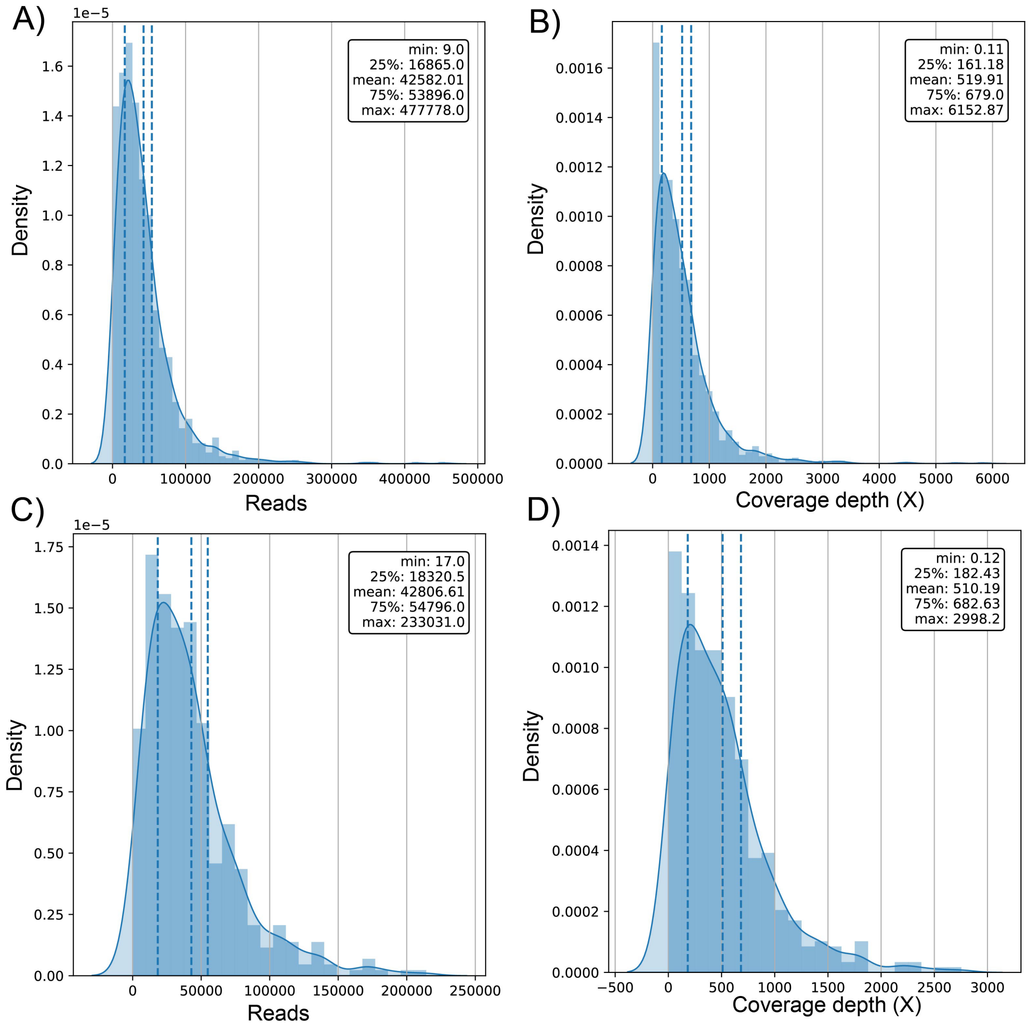

Figure 2 Histogram of distribution of sequenced reads and coverage depth in the dataset of 1461 (A and B, respectively) and 471 genomes (C and D, respectively).

Full-size DOI: 10.7717/peerj.14425/fig-2

510.2 ± 453.22X for dataset (1) and dataset (2), respectively (Table S1), with an almost identical distribution (Figs. 2B and 2D).

We analyzed the correlation between the target variable (number of reads sequenced) and the dependent variable (coverage depth) (Fig. 3). Since the curve was not flat, the feature was predictive and could be used for model construction (Figs. 3A and 3C). This trend is shown in the univariate regression of the number of reads compared to coverage depth, pointing to a strong direct relationship between these features (Figs. 3B and 3D) with a Pearson's Correlation Coefficient of 0.977 and 0.959 for dataset (1) and dataset (2), respectively (p-value <0.001 in both cases).

We found the relationship in the dataset (1): $y = 0.0129x - 30.315$; for every 100 reads added in real-time in an experiment, the expected average coverage depth of each sample

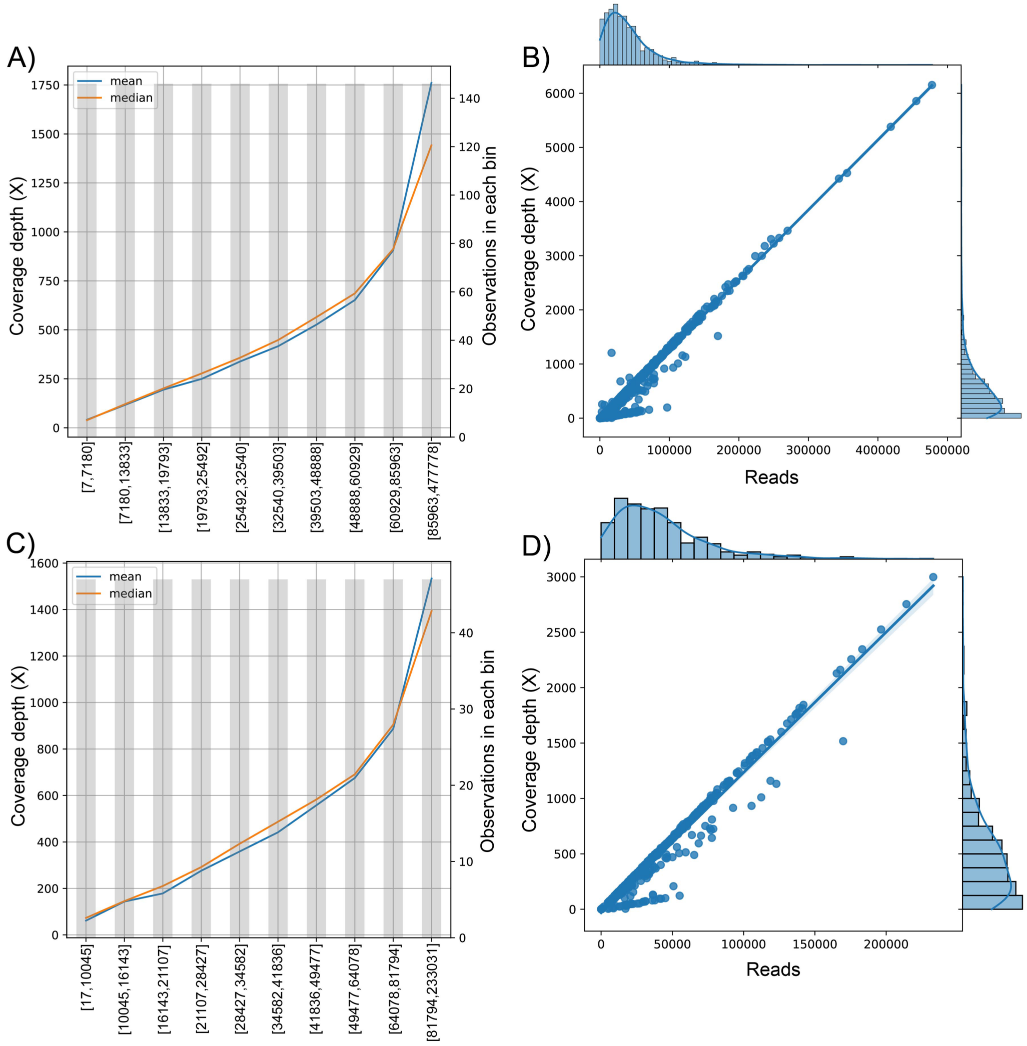

Figure 3 The behavior of the target variable (number of reads per genome) for coverage depth in the dataset 1 (A) and dataset 2 (C). Scatter plot with the distributions of the two variables in dataset 1 (B) and dataset 2 (D). The plot of sequenced reads into bins comparing the mean and median value of coverage depth data in each bin.

Full-size DOI: 10.7717/peerj.14425/fig-3

increases by 1.3X. Therefore, if we desire 200X depth for lineage assignment, we need at least 17,854 reads.

Models' performance

The performance metrics values for all models tested showed R^2 between 0.93 and 0.98 (Table 1). The ensemble models had better performance in both datasets with an R^2 of 0.9794 for GBR in dataset1 and 0.9506 for RFR in dataset2; moreover, SVR showed in both cases the lowest R^2 values (0.9646 and 0.9347, respectively) (Table 1). In the case of dataset1, GBR was followed by models LssR, RFR, then SVR. However, the performance order of the models changed when we used the smaller dataset (2), with the best being RFR, followed by GBR, LssR, then SVR. This performance pattern is similar when the average

Table 1 R^2, RMSE, and MAE performance values of dataset 1 and 2 evaluated under different models.

Dataset	Model	Average R^2	R^2	MAE	RMSE
1	LssR	0.9715	0.9741	3817.5526	6509.1245
	SVR	0.9610	0.9646	2762.9007	7607.8926
	GBR	**0.9771**	**0.9794**	2843.2006	**5804.2179**
	RFR	0.9756	0.9727	**2559.9966**	6683.7374
2	LssR	**0.9637**	0.9411	5191.3255	9082.4665
	SVR	0.9539	0.9347	**3874.7963**	9565.7193
	GBR	0.9539	0.9497	5099.3944	8393.9312
	RFR	0.9603	**0.9506**	4478.2003	**8317.3773**

Notes.

The best values for each metric are in bold.

R^2 of cross-validation is evaluated, where GBR outperforms the rest of the models with an R^2 of 0.9771 and only changes radically in dataset 2, where LssR is the highlight as the best fitting, with 0.9637, following by RFR, SVR, and GBR. RFR and SVR presented the least mean absolute error of predictions in both datasets (2560 and 3875 sequenced reads for dataset (1) and dataset (2), respectively: Table 1).

We used GBR for the construction of the predictive app of variables modulating the total number of sequenced reads. In short, the best performance models (GBR and RFR) explain 95–98% of the variance of total reads of genomes in both used datasets. On average, predictions for these models in dataset (1) and dataset (2) have a mean error of 2698 (12%) and 4779 (18%) sequenced reads, respectively.

We visualized the results of k-fold validation by plotting predicted values against the actual sequenced reads, observing that points are close to a diagonal line where predicted = real (Figs. S2 and S3). Consistent with the performance metrics, the ensemble models presented better predictive power and fitting compared to other algorithms. Figure 4 shows the high similarity in predicted reads values between models in both datasets, with SVR having more outliers compared to the rest, underestimating the true values of the target variable.

The biggest error in the test set was over 73,353 and 53,444 sequenced reads modeled by RFR for dataset (1) and dataset (2), respectively. Figs. S4 and S5 visualize the errors in both datasets by plotting predicted against the residual of each prediction; for dataset (1), most errors lie on the negative side, meaning these predictions are underestimated (composed of 61.20, 51.91, 57.92, and 60.38% of residuals in LssR, SVR, GBR, and RFR, respectively). Similarly, dataset (2) has 70.34, 53.40, 61.01, and 70.34% negative residuals, respectively. In both datasets, SVR showed the most equitable distribution of positive and negative errors. Moreover, LssR and SVR showed a slight positive skew with many values in the interval of 10,000 to 30,000 compared to ensemble models that have more randomly distributed errors, which is consistent with a more approximately normal distribution of residuals in GBR and RFR compared to LssR and SVR (Figs. S6 and S7).

Finally, the plausibility of the GBR model predictions for both datasets was evaluated. In the case of the dataset (1), the model predicts 43,089 reads were obtained (42,740 true reads), given cDNA quantification values between 14.85–24.20 ng/uL, coverage between

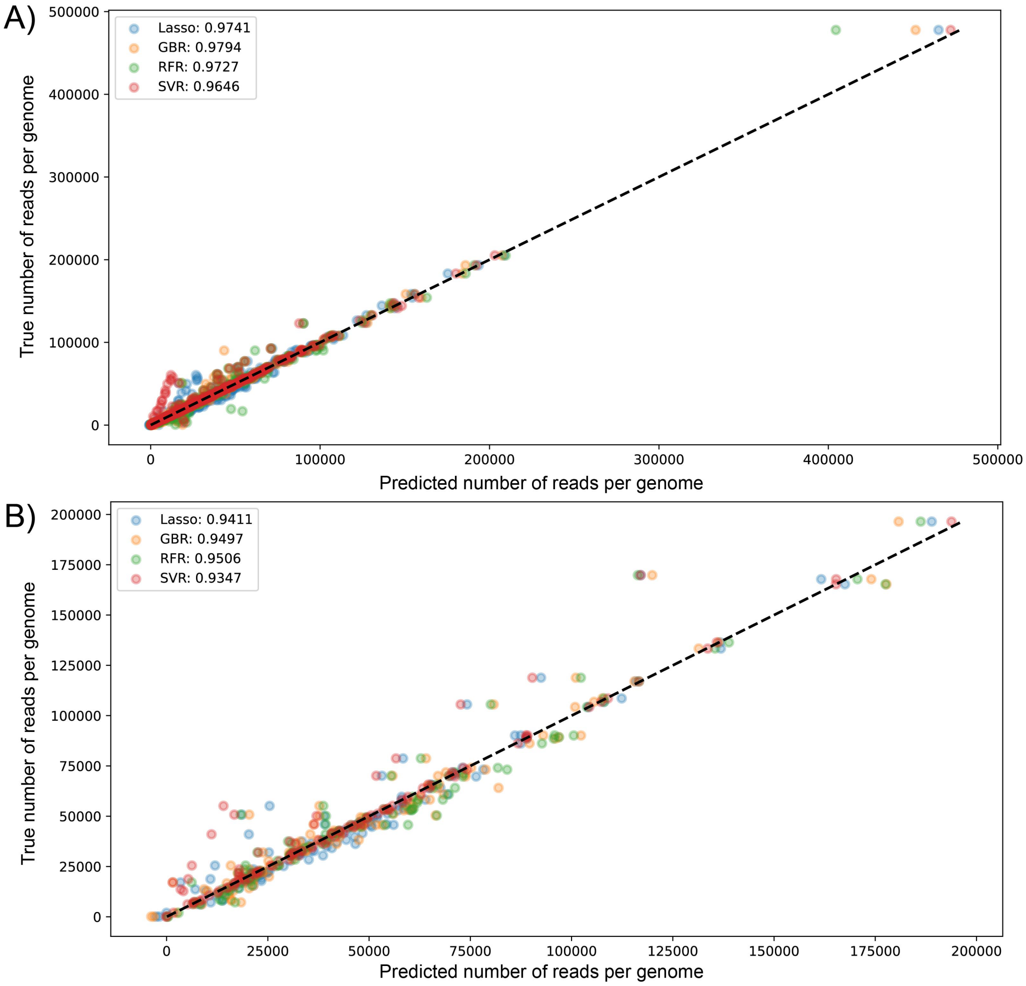

Figure 4 Scatter plot for predicted *vs* measured (true) values for sequenced reads, using linear regression with Lasso regularization, Support Vector Regressor, Gradient Boosting, and Random Forest for two datasets of 1461 (A) and 471 genomes (B). R^2 scores are equivalent to Table 1 values, obtained from test partitions of each dataset. "Average R^2" values obtained from the geometric mean of the scores of 5-iterations of cross-validation.

Full-size DOI: 10.7717/peerj.14425/fig-4

91.12–96.03, and coverage depth between 380-688X. In dataset (2), the model predicts that 78,807 reads were obtained (73,161 true reads), given N2 CT between 13-16, cDNA quantification values above 34.8 ng/uL, coverage between 91.18–95.28, and coverage depth above 643X.

DISCUSSION

The objective of this study was to test multiple SML regression algorithms to accurately estimate enough sequenced reads per SARS-CoV-2 genome to achieve an acceptable depth of coverage given different sample variables. Therefore, we compared simple linear regression with parametric models SVR and ensemble algorithms for the accuracy and robustness of the predicted target variable. The sample variables of the mean number of sequenced reads, mean coverage per genome and CT have literature-supported relationships (*Liu et al., 2022*; *Wang et al., 2011*). Both genome coverage and depth of coverage are highly dependent on the viral concentration of the sample, *Gauthier et al. (2021)* showed that RT-qPCR SARS-CoV-2 samples with CTs greater than 25 tend to have coverage less than 95%. Similarly, *Brinkmann et al. (2021)* reported that using 75,000 reads with a depth >10X, they achieved > 98% coverage of the SARS-CoV-2 genome using ONT. Consequently, including these variables was a priority to develop and create a machine learning model, from data gathering to evaluation.

Estimated performance metrics show GBR had the highest R^2 and lowest RMSE values for predicted reads in the test and training datasets, indicating this ensemble model can provide accurate depth estimations. However, this performance order changes in average R^2 of cross-validation in dataset 2, indicating that GBR (where $R^2_{test} < R^2_{training}$) does not generalize well or has poor stability in this dataset (*Li, Luan & Wu, 2020*); however, this limitation can be attributed to small sample size. On the other hand, MAE in both tested datasets was reduced in RFR and SVR models, outperforming GBR (Table 1). This pattern can be attributed to the nature of RMSE, which penalizes large gaps in the model predictions, while MAE does not (*Koutsandreas et al., 2021*). On average, RFR and SVR have more large-scale errors but fewer small-scale errors than GBR; this can be visualized in the residuals plots (Figs. S4 and S5), where GBR accumulates more small-scale errors in the interval of 10,000 to 30,000 reads, while SVR and RFR have more large-scale errors randomly distributed.

Most of the sequencing input variables used in this study are correlated. Due to this, GBR was preferred over other models that are more sensitive to collinearity, such as SVR (*Cutler et al., 2007*). Moreover, both the adaptability of ensemble models to small sample sizes along with the insensitivity to overfitting data, outliers, and less predictive input (when the depth of coverage function was removed from data set 2) were advantages (*Bellido-Jiménez, Gualda & García-Marín, 2021*; *Wang et al., 2016*).

Finally, we propose the use of Gradient Boosting Regressor for real-time monitoring of ONT MinION sequencing of SARS-CoV-2 samples. The prediction accuracy of the method should be further validated by optimizing the modeling algorithms. Therefore, the online dashboard was created using GBR to predict the mean depth of coverage and other variables from simple input (number of sequenced reads), to improve the cost-effectiveness of flow cells and ONT sequencing for laboratories with fewer resources. It would be interesting to apply the method to monitor other reactions *a priori* sequencing parameter as storage of sample time, quality metrics of cDNA, and different primer kits to verify reproducibility. This research contributes to the establishment of SARS-CoV-2 genomic surveillance management strategies for simple experiment monitoring and precise modeling methods.

CONCLUSIONS

We propose a novel method for the prediction of reads necessary to sequence SARS-CoV-2 at a sufficient coverage depth that allows database-acceptable lineage. Different machine learning models were trained; the Gradient Boosting Regression algorithm showed the best fit, explaining more than 98% of the variance of the sequenced reads, and presented an average error of 16% in the predictions. The Gradient Boosting Regressor provides a useful exploratory and predictive tool for estimating reads given *a priori* sequencing process variables such as CT (gene N) and amount of cDNA per sample and *a posteriori* sequencing quality variables coverage and mean depth of coverage per genome. The implementation of these methods will bring a reduction in sequencing costs through process optimization.

ACKNOWLEDGEMENTS

We are thankful to all the team at "Virus Emergentes y Enfermedades"(VIREM) at Universidad del Valle, Cali-Colombia, who are supporting the diagnostics of SARS-CoV-2 in Colombia, as well as different investigators of Colombia and worldwide who deposited their genomes in GISAID. We also thank the Colombian network for SARS-CoV-2 genomic surveillance led by the Colombian National Institute of Health (INS). Thanks to Diego Ceballos, CEO of People Contact Enterprise for supporting the dashboard builder and Samantha Hao at the Johns Hopkins School of Medicine for editing the English version.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The study was supported by National Institutes of Health (NIH, No. R01NS110122) to the Neuroviruses Emerging in the Americas Study (NEAS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institutes of Health: NIH, No. R01NS110122.
Neuroviruses Emerging in the Americas Study (NEAS).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- David E. Valencia-Valencia conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Diana Lopez-Alvarez conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Nelson Rivera-Franco performed the experiments, authored or reviewed drafts of the article, genome processing, and approved the final draft.
- Andres Castillo conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Johan S. Piña analyzed the data, prepared figures and/or tables, and approved the final draft.

- Carlos A. Pardo analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Beatriz Parra conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Ethics

The following information was supplied relating to ethical approvals (*i.e.*, approving body and any reference numbers):

The study was approved by the Ethics Committee of Universidad del Valle, Colombia.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub: https://github.com/TAOLabUV/PredictION.

The dataset is available at Zenodo: David E. Valencia-Valencia, Diana Carolina Lopez Alvarez, Nelson Rivera-Franco, Andres Castillo, Johan S Piña, Carlos A Pardo, & Beatriz Parra. (2022). PredictION: A predictive model to establish the performance of Oxford sequencing reads of SARS-CoV-2 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7104956.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.14425#supplemental-information.

REFERENCES

Anderson C. 2015. Docker [Software engineering]. *IEEE Software* **32**(**3**):c102–c103 DOI 10.1109/MS.2015.62.

Bellido-Jiménez JA, Gualda JE, García-Marín AP. 2021. Assessing machine learning models for gap filling daily rainfall series in a semiarid region of Spain. *Atmosphere* **12**(**9**):1158 DOI 10.3390/atmos12091158.

Brinkmann A, Ulm S-L, Uddin S, Förster S, Seifert D, Oehme R, Corty M, Schaade L, Michel J, Nitsche A. 2021. AmpliCoV: rapid whole-genome sequencing using multiplex PCR amplification and real-time Oxford nanopore MinION sequencing enables rapid variant identification of SARS-CoV-2. *Frontiers in Microbiology* **12**:651151 DOI 10.3389/fmicb.2021.651151.

Cherkassky V, Ma Y. 2003. Comparison of model selection for regression. *Neural Computation* **15**(**7**):1691–1714 DOI 10.1162/089976603321891864.

Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. Random forests for classification in ecology. *Ecology* **88**(**11**):2783–2792 DOI 10.1890/07-0539.1.

Gauthier NPG, Nelson C, Bonsall MB, Locher K, Charles M, MacDonald C, Krajden M, Chorlton SD, Manges AR. 2021. Nanopore metagenomic sequencing for detection and characterization of SARS-CoV-2 in clinical samples. *PLOS ONE* **16**(**11**):e0259712 DOI 10.1371/journal.pone.0259712.

Koutsandreas D, Spiliotis E, Petropoulos F, Assimakopoulos V. 2021. On the selection of forecasting accuracy measures. *Journal of the Operational Research Society* **73**(**3**):1–18 DOI 10.1080/01605682.2021.1892464.

Lambisia AW, Mohammed KS, Makori TO, Ndwiga L, Mburu MW, Morobe JM, Moraa EO, Musyoki J, Murunga N, Mwangi JN, Nokes DJ, Agoti CN, Ochola-Oyier LI, Githinji G. 2022. Optimization of the SARS-CoV-2 ARTIC network V4 primers and whole genome sequencing protocol. *Frontiers in Medicine* **9**:836728 DOI 10.3389/fmed.2022.836728.

Li X, Luan F, Wu Y. 2020. A comparative assessment of six machine learning models for prediction of bending force in hot strip rolling process. *Metals* **10**(**5**):685 DOI 10.3390/met10050685.

Liu H, Li J, Lin Y, Bo X, Song H, Li K, Li P, Ni M. 2022. Assessment of two-pool multiplex long-amplicon nanopore sequencing of SARS-CoV-2. *Journal of Medical Virology* **94**(**1**):327–334 DOI 10.1002/jmv.27336.

Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. 2017. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology* **48**(**3**):128–138 DOI 10.14445/22312803/IJCTT-V48P126.

Smola AJ, Schölkopf B. 2004. A tutorial on support vector regression. *Statistics and Computing* **14**(**3**):199–222 DOI 10.1023/b:stco.0000035301.49549.88.

Tyson JR, James P, Stoddart D, Sparks N, Wickenhagen A, Hall G, Choi JH, Lapointe H, Kamelian K, Smith AD, Prystajecky N, Goodfellow I, Wilson SJ, Harrigan R, Snutch TP, Loman NJ, Quick J. 2020. Improvements to the ARTIC multiplex PCR method for SARS-CoV-2 genome sequencing using nanopore. *bioRxiv: the preprint server for biology* DOI 10.1101/2020.09.04.283077.

Verbert K, Duval E, Klerkx J, Govaerts S, Santos JL. 2013. Learning analytics dashboard applications. *American Behavioral Scientist* **57**(**10**):1500–1509 DOI 10.1177/0002764213479363.

Verbert K, Govaerts S, Duval E, Santos JL, Van Assche F, Parra G, Klerkx J. 2014. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing* **18**(**6**):1499–1514 DOI 10.1007/s00779-013-0751-2.

Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, Zhou H. 2011. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics* **12**(**10**):S5 DOI 10.1186/1471-2105-12-S10-S5.

Wang LA, Zhou X, Zhu X, Dong Z, Guo W. 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal* **4**(**3**):212–219 DOI 10.1016/j.cj.2016.01.008.