Enhancing SARS-CoV-2 Lineage surveillance through the integration of a simple and direct qPCR-based protocol adaptation with established machine learning algorithms

Cleber Furtado Aksenen[1]*, Débora Maria Almeida Ferreira[1]*, Pedro Miguel Carneiro Jeronimo[1]*, Thais de Oliveira Costa[1], Ticiane Cavalcante de Souza[1], Bruna Maria Nepomuceno Sousa Lino[1], Allysson Allan de Farias[1], Fabio Miyajima[1]

1 Fiocruz Genomic Network, Oswaldo Cruz Foundation (FIOCRUZ), branch Ceara, Eusebio, Brazil

*These authors contributed equally to this article.

ABSTRACT

The emergence of the SARS-CoV-2 and continuous spread of its descendent lineages have posed unprecedented challenges to the global public healthcare system. Here we present an inclusive approach integrating genomic sequencing and qPCR-based protocols to increment monitoring of variant Omicron sublineages. Viral RNA samples were fast tracked for genomic surveillance following the detection of SARS-CoV-2 by diagnostic laboratories or public health network units in Ceara (Brazil) and analyzed using paired-end sequencing and integrative genomic analysis. Validation of a key structural variation was conducted with gel electrophoresis for the presence of a specific ORF7a deletion within the "BE.9" lineages. A simple intercalating dye-based qPCR assay protocol was tested and optimized through the repositioning primers from the ARTIC v.4.1 amplicon panel, which was able to distinguish between "BE.9" and "non-BE.9" lineages, particularly BQ.1. Three ML models were trained with the melting curve of the intercalating dye-based qPCR that enabled lineage assignment with elevated accuracy. Amongst them, the Support Vector Machine (SVM) model had the best performance and after fine-tuning showed ~96.52% (333/345) accuracy in comparison to the test dataset. The integration of these methods may allow rapid assessment of emerging variants and increment molecular surveillance strategies, especially in resource-limited settings. Our approach not only provides a cost-effective alternative to complement traditional sequencing methods but also offers a scalable analytical solution for enhanced monitoring of SARS-CoV-2 variants for other laboratories through easy-to-train ML algorithms, thus contributing to global efforts in pandemic control.

Keywords

SARS-CoV-2; Variant detection; Genomic monitoring; qPCR-based protocols; Machine Learning; Laboratory Surveillance.

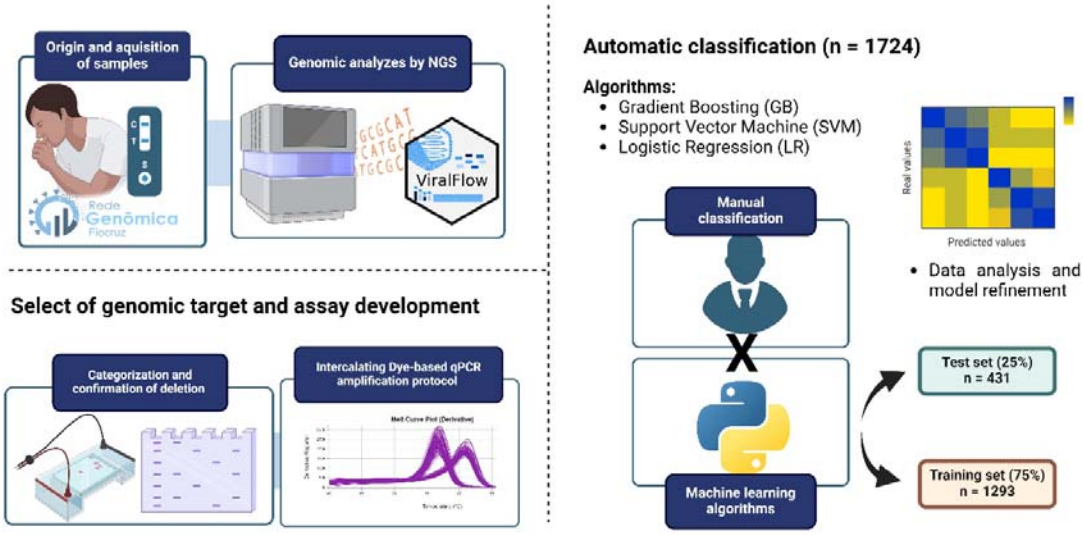


Figure 1 Intercalating dye-based qPCR protocol devised for surveillance of SARS-CoV-2 BE.9 lineages. The schematic illustrates the step-by-step process employed for the precise detection of the targeted deletion within the ORF7a gene (27,508-27,751) by analysis of amplification curves. Synapomorphy was identified by NGS, confirmed through electrophoresis of a subset of high-quality sequencing samples and compared with the amplification results. The protocol was automated using refined machine learning models for an extended set of 1,724 samples, trained through manual classification.

INTRODUCTION

The adaptability of viruses like SARS-CoV-2 through cumulative mutations denotes the dynamic interaction between pathogens and their environment. Mutations leading to structural modifications, such as insertions or deletions, are more likely to account for significant alterations in the biological behavior of the virus, ultimately fueling the emergence of variants with potential selective advantage and pathogenic profiles [1–3]**.** This adaptive mechanism has been illustrated by the emergence of SARS-CoV-2 variants, which is known for its increased infectivity due to specific amino acid substitutions [4,5]. The genetic diversity observed in RNA viruses, underscored by the continuous emergence of new mutations, highlights the evolving nature of these pathogens and the critical role of genomic surveillance in tracking these changes [3,6,7]. The swift emergence and global proliferation of the Omicron variant (B.1.1.529) of SARS-CoV-2, along with its descendant subvariants, have heightened global apprehensions because of their extensive repertoire of distinctive genetic configurations and unprecedented transmission capabilities [8,9]. Studies have illuminated the variant's ability to outpace previous strains, such as the Delta variant, in terms of spread, leading to a considerable uptick in reinfection rates, affecting even those previously vaccinated or infected [10–13]. The situation is compounded by the variant's elusive severity profile compared with its predecessors, necessitating rigorous public health interventions [14–16] . Given the dynamic nature of the virus, heightened emphasis on genomic surveillance is imperative to track and understand the emergence of new strains, enabling proactive measures to mitigate their spread and impact.

The global effort to monitor and control the spread of SARS-CoV-2 faces numerous challenges, including the economic and infrastructural disparities among countries. The challenges posed by NGS analyses, including high costs, lengthy response times, and its inaccessibility in economically disadvantaged regions, have spurred the scientific community to explore supplementary techniques [17–19]. There has been a notable change toward

integrating polymerase chain reaction (PCR) and computational algorithms into the genomic surveillance toolkit. These methods offer a more immediate and cost-effective capability for detecting specific genetic markers, thereby enhancing the efficiency and scope of pathogen surveillance efforts [20,21].

Among this landscape of innovation, the intercalating dye-based qPCR protocol has emerged as an important technique in the field of genetic surveillance. Distinguished by its capacity for real-time DNA amplification monitoring and low cost, this protocol has shown remarkable efficacy in pinpointing specific genetic markers. Its use not only marks a significant advanced strategy in the rapid identification of variants of concern (VOCs) but also in understanding the intricate dynamics of viral adaptations [22,23]. The protocols' insights into the ORF7a gene, particularly its role in immune modulation and interaction with host cells, underscore the complex interplay between viral genetics and host defenses, highlighting the importance of nuanced genetic surveillance in preparedness to the challenges of the COVID-19 pandemic and beyond [3,24,25].

The intercalating dye-based qPCR protocol is a low-cost assay technique, highly adaptable to near real time tool in the field of genomic surveillance, due to its steadfast deployment. This strategy can significantly improve both the speed and precision for target detections, proving reliable for the confirmation of key molecular signatures used for tracking the population dynamics and evolution of pathogens, such as SARS-CoV-2. Our work has highlighted the applicability of a lineage-defining genetic marker, a 244-base deletion within the ORF7a gene (27508 – 27751) characteristic marker of the Brazilian BE.9 lineage. We proposed this specific deletion could be informative and able to track in the spread of this lineage from September 2022 to May 2023 (https://gisaid.org/), underscoring the utility of a qPCR-based protocol in pinpointing the expansion of emerging variants and sublineages that pose new challenges to public health and vaccine efficacy. This seamless integration of computational analyses and a straightforward intercalating dye-based qPCR protocol represents a more direct and inclusive approach to monitoring viral evolution. It embodies the scientific community's and public health policies in engaging in rapid response measures to monitor evolving pathogen variants, ensuring that public health strategies remain robust and responsive in the ongoing battle against immune escape and SARS-CoV-2 adaptability.

RESULTS AND DISCUSSION

Integrative Genomic Analysis and Categorization. SARS-CoV-2 genomic sequences with high-quality samples (horizontal coverage exceeding 90% and vertical coverage surpassing 100x) revealed a low depth region at position 27,508 – 27,751 of the ORF7a gene for previously classified as "BE.9" **(Figure 2A)**, when compared with classified as "non-BE.9", particularly BQ.1 **(Figure 2B)**.

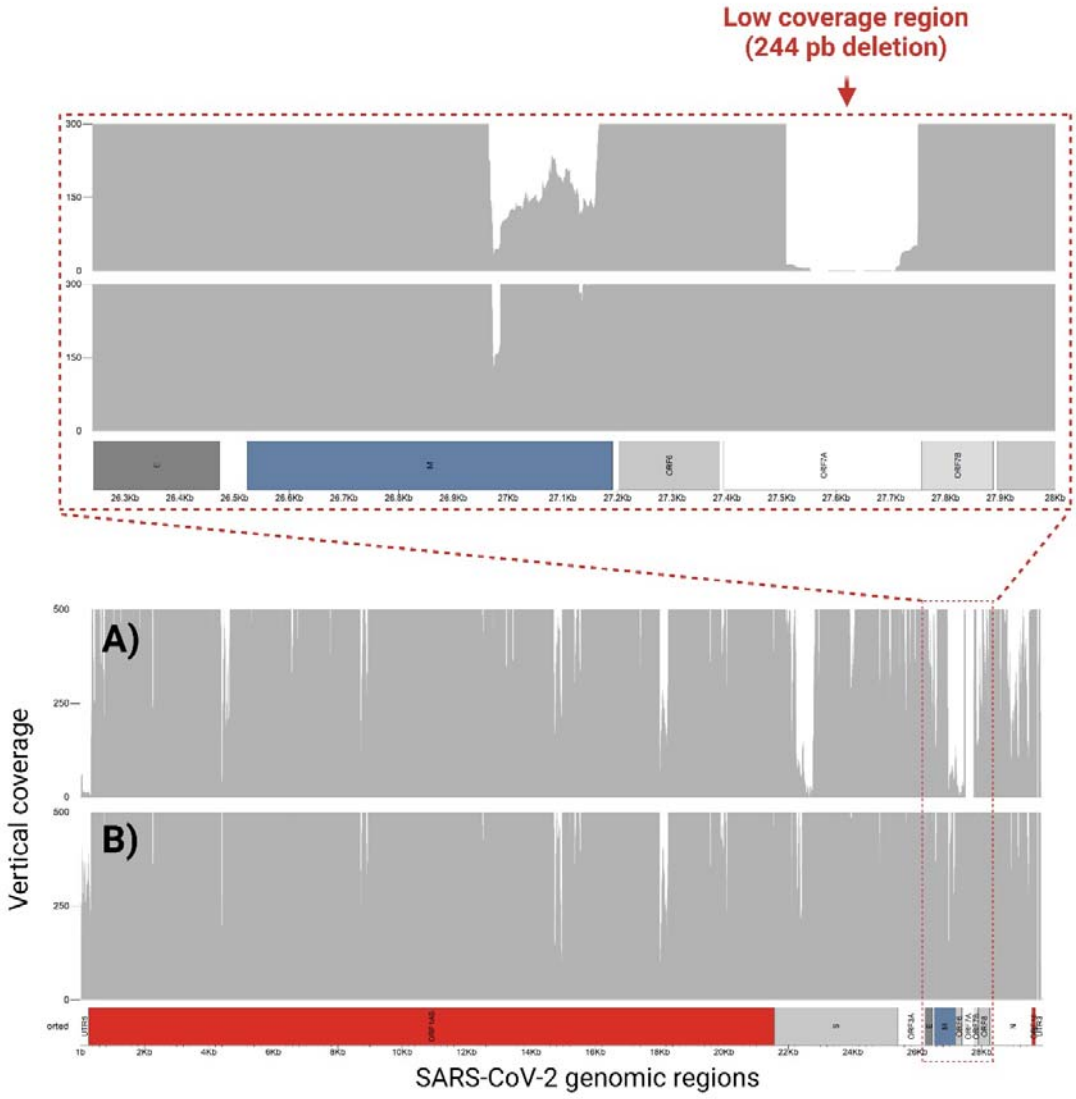


Figure 2 Genomic vertical coverage profile of SARS-CoV-2 highlighting variations. The coverage distribution across the genome shows the difference between BE.9 (A) and non-BE.9 (B) lineages in the ORF7a gene region, showing the low coverage region due to the presence of the 244-base deletion in the BE.9 samples.

The presence of extensive low-depth sequenced regions presents significant challenges to bioinformatics analyses and interpretation, undermining potentially the accurate identification of genuine evolutionary events, such as deletions. The detection of this particular structural mutation (a deletion of 244 bp) within the ORF7a gene was corroborated by routine inspection of amplified targets separated by gel electrophoresis, which endorsed it as a synapomorphic signature of the BE.9 subvariants. The detection of a characteristic band in the range of 170-200 bp (S01 to S08) was found across all BE-9 samples phylogenetically assigned by whole genome sequencing. Amongst the 'non-BE.9' samples (S09 to S16), *ORF7a* bands between 400-430 bp were consistently present, thus denoting the absence of deletion **(Figure 3)**. These distinct band patterns offer compelling evidence for the existence of genuine structural alterations between two major SARS-CoV-2 subvariants, of

independent origins, and reinforce findings from previous studies concerning the loss of genetic elements during the natural evolution of SARS-CoV-2[3]. Additionally, it also contributed to obtaining evidence regarding the applicability of a PCR amplification protocol that makes use of intercalating dye-based strategies, aiming to increase speediness and robustness of the investigations.

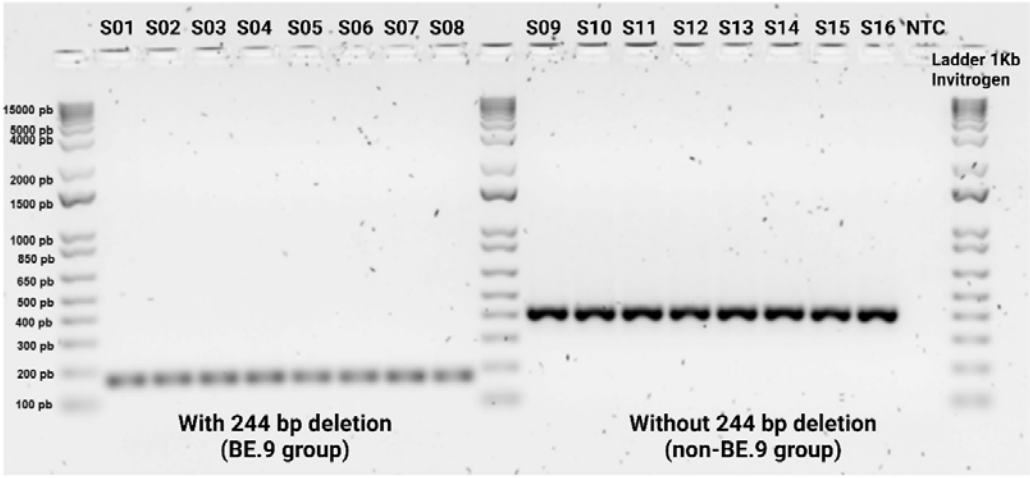


Figure 3 Agarose gel electrophoresis of amplified DNA fragments to validate the ORF7a deletion. The 'BE.9' groups (S01 to S08) show the absence of part of the band corresponding to ORF7a, located around 244 base pairs (bp), while the 'non-BE.9' groups (S09 to S16) show bands intact sections of ORF7a, between 400 and 430 bp.

Initial intercalating dye-based qPCR amplification protocol. Building upon these insights, a qPCR protocol enhanced by the integration of the intercalating dye-based assay (BRYT® Green GoTaq mastermix, Promega inc.) aimed into amplifying the region encompassing the identified 244-base pair deletion, thus providing a targeted and high-throughput method for distinguishing between 'BE.9' and 'non-BE.9' lineages, as well as a reliable, flexible and cost-effective approach [26,27].

The results obtained from the first derivative melt curves of the sixteen samples are clarity in **Figure 4**, illustrating the melt curves corresponding to the BE.9 group and the non-BE.9 group. All melt curves for the BE.9 group exhibited amplification at an average melting temperature (TM) of 76.78 ± 0.18°C, accompanied by fluorescence levels ranging between 200k and 300k at its peak. In contrast, the non-BE.9 group displayed an average melting temperature of 80.76 ± 0.24°C, with a wider range of fluorescence intensity, spanning from 200k to 400k. This confirmation, highlighted by the lower Tm for BE.9 and higher Tm for non-BE.9, aligns with prior research suggesting that longer amplicons exhibit higher melting temperatures (TM) compared to shorter ones [28]. Notably, it was observed during manual analysis and categorization of the samples that the melting curves with fluorescence levels below 100k were challenging to visualize and classify accurately. This challenge was significantly alleviated when the range was filtered to values greater than 100k, enhancing the clarity and precision of group classification.

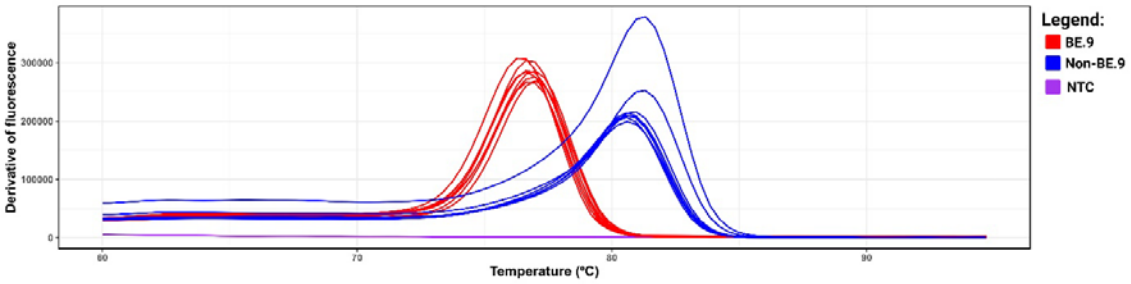


Figure 4 Dissociation curve generated from the ORF7a_244del assay, designed for BE.9 detection via RT-qPCR protocol. This assay targets the 244-base deletion in the ORF7a region of the SARS-CoV-2 genome, a defining characteristic of the BE.9 lineages. Samples attributed to the 'BE.9' designation are highlighted in red, consistently exhibiting lower Tm values (76.78 ± 0.18°C), while 'non-BE.9' samples, depicted in blue, demonstrate higher Tm values (80.76 ± 0.24°C). Notably, the negative control displayed no amplification.

The first derivative melt graphs of the sixteen samples demonstrated a distinct separation between the BE.9 and non-BE.9 groups. No instances of BE.9 were observed within the melting temperature (TM) range of the non-BE.9 groups, and reciprocally, underscoring the assay's effectiveness in distinguishing between these virus lineages. The behavior of the negative control (NTC) visualized in **Figure 4**, representing the absence of the virus, exhibited no fluorescence, indicating the absence of primer dimers or unintended products in the assay. This underscores the careful management of primers and ensures the assay's reliability by minimizing the presence of contaminating artifacts.

Machine learning algorithms and data analysis. The SVM with a linear kernel emerged as the best-performing model, surpassing Logistic Regression and Gradient Boosting **(Table 2)**.

Algorithm	Precision			Recall			F1-score			Accuracy
	BE.9	Non- BE.9	Inconclusive	BE.9	Non-BE.9	Inconclusive	BE.9	Non- BE.9	Inconclusive	
SVM	0.991	0.992	0.936	0.974	0.968	0.981	0.983	0.980	0.960	**0.974**
Logistic Regression	1.000	0.976	0.910	0.932	0.984	0.971	0.965	0.980	0.940	**0.963**
Gradient Boosting	0.983	0.976	0.961	0.983	0.984	0.952	0.983	0.980	0.957	**0.974**

Table 2 Performance Comparison of Machine Learning Algorithms. This table presents the performance metrics of machine learning algorithms tested of different groups: "BE.9", "non-BE.9", and "Inconclusive".

Optimizing the SVM parameters resulted in a tie between various hyperparameters configurations **(Supplementary table 3)**. This **table 3** compares the unoptimized version of the SVM model with the one using the settings {'C': 100, 'degree': 2, 'kernel': rbf, gamma: auto} on the evaluate set with the fine-tuned model on the test set. The accuracy demonstrated a marginal improvement, accompanied by enhanced precision, recall, and F1-score metrics for certain subsets within the BE.9, non-BE.9, or inconclusive groups when evaluating the impact of fine-tuning on the test set. This uptick in accuracy indicates the accurate classification of previously mislabeled inconclusive curves as non-BE.9. However, it

is imperative to assess these metrics on unseen data. While there was a decrease in metrics, it is probable that these values reflect the true performance on other unseen datasets.

Algorithm	Precision			Recall			F1-score			Accuracy
	BE.9	Non- BE.9	Inconclusive	BE.9	Non-BE.9	Inconclusive	BE.9	Non- BE.9	Inconclusive	
SVM Before Tuning (Evaluate set)	0.991	0.992	0.936	0.974	0.968	0.981	0.983	0.980	0.958	**0.974**
SVM After Tuning (Evaluate set)	1.000	0.969	0.980	0.992	0.992	0.962	0.996	0.980	0.971	**0.983**
SVM After Tuning (Test set)	0.989	0.957	0.955	0.949	1.000	0.939	0.968	0.978	0.947	**0.965**

Table 3 Performance Comparison before and after SVM Parameter Optimization. This table illustrates the performance comparison between the unoptimized version of the Support Vector Machine (SVM) model and the version utilizing specific hyperparameter settings {'C': 100, 'degree': 3, 'kernel': 'linear'}.

The high accuracy signifies substantial reliability when utilizing melting curve points for curve classification, automating the process. Other studies have approached diagnostic classification using derived metrics, whether through Principal Component Analysis (PCA) [30,31], metrics related to curve shape (skewness, kurtosis etc) [20], or variables associated with the technique itself (amplicon melting temperature) [32], achieving accuracies ranging from 72% to 100%. In this work, however, we chose to directly use the curve itself, transforming each point of every curve into a column, or feature, for the model, employing a simple data normalization step. This approach streamlines the model development process, ensuring simplicity without compromising accuracy. It is noteworthy that the clear distinction between curves for BE.9 and non-BE.9 classification enables this approach. By utilizing points from the curve directly, the model gains the flexibility to discern nuances indicating which individual points are more crucial for correct result classification.

The confusion matrix reveals correct classification values for the BE.9 lineage at 94.85% (n = 92), non-BE.9 at 100.00% (n = 134), and inconclusive at 93.86% (n = 107) **(Figure 5)**. Despite the high classification accuracy for SARS-CoV-2 BE.9 and non-BE.9 lineages, there is a noticeable decline in classification quality for inconclusive curves, often reflecting the subjective nature of classification by analysts. It is crucial, therefore, during the establishment of the gold standard used for model training, to clearly define each of the curves.

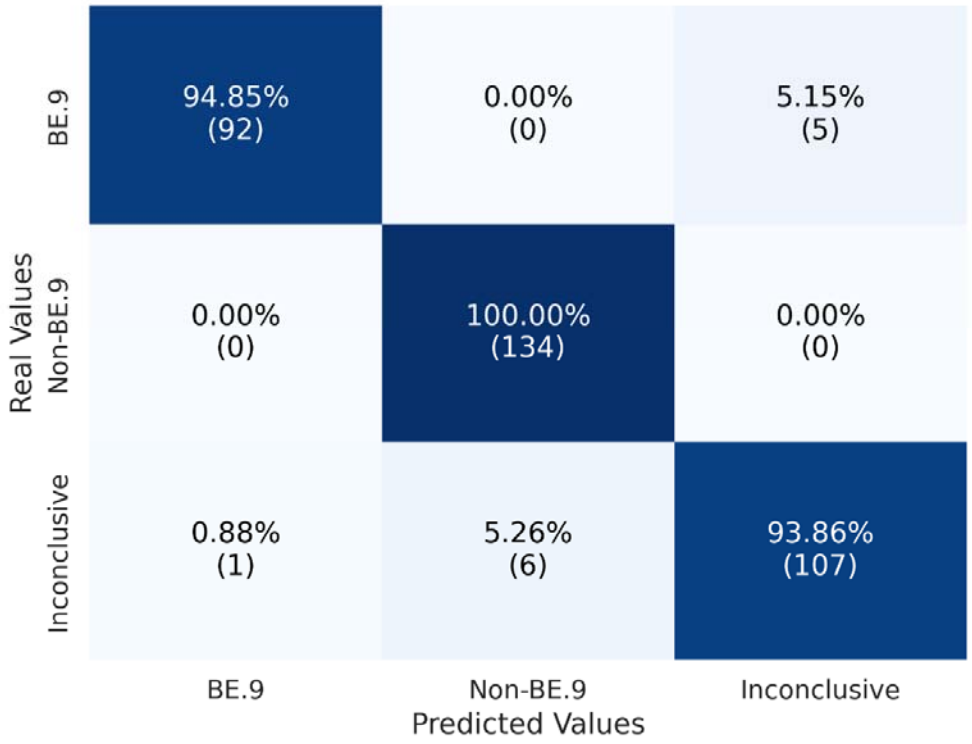


Figure 5 Melting Curve Classification Performance. The performance of melting curve classification shows the high accuracy achieved, indicating substantial reliability in automating the classification process.

The utilization of free platforms such as Google Colaboratory could contribute to the democratization and swift investigation of outbreaks of new variants in regions lacking computational power [33]. Simple modeling from minimally processed data represents an encouraging opportunity for other groups to optimize protocols, demystifying the use of machine learning algorithms in routine laboratory procedures, allowing for biological applications as already used for other purposes [34,35].

CONCLUSIONS

In conclusion, our study demonstrates the efficacy of the implemented optimized intercalating dye-based qPCR protocol combined with machine learning (ML) analysis as a powerful method for discriminating and classifying independent SARS-CoV-2 sublineages of high homology. This approach offers automated binary inference of the most probable circulating SARS-CoV-2 sublineages (BE.9 or non-BE.9), providing a valuable complement to the more complex NGS-based surveillance methods. The identification of a region of low vertical coverage in BE.9 samples, confirmed through gel electrophoresis as a genuine synapomorphy in the form of a 244 bp deletion, underscores the importance of structural genomic alterations in providing alternatives for monitoring emergence and spread of SARS-CoV-2 variants.

Moreover, the distinct melting temperature (TM) curves between 'BE.9' and 'non-BE.9' groups, along with a classification sensibility of 94.85% and 100.00%, respectively, using the SVM ML algorithm, highlight the robustness of our methodology. Despite initial challenges with "inconclusive" samples, primarily stemming from characteristics of reused rapid antigen tests, our method maintained a high classification sensibility of 93.86% for identifying such samples. These results underscore the potential of qPCR-based protocols for investigating evolutionary patterns in pathogens, with broad implications for diagnostics, surveillance, and public health interventions.

Moving forward, further research is warranted to validate and refine our method, extending its applicability to other infectious diseases and addressing any existing limitations. This will ensure its continued relevance in the dynamic landscape of infectious disease research and control. Additionally, the integration of machine learning methodology, as demonstrated in this study, enhances the analytical capabilities of generated data, ultimately optimizing lineage diagnosis.

Furthermore, exploring the potential application of non-specific intercalating dye assays for detecting and identifying various pathogens opens avenues for extending this innovative methodology of machine learning to other assays. This broader application not only enhances its utility but also reduces costs and the need for robust equipment, making it more accessible to diverse research settings. Overall, our study contributes to advancing methodologies in infectious disease research and underscores the potential of interdisciplinary approaches in combating emerging pathogens.

METHODS

Origin and acquisition of samples. The viral RNA samples were acquired through a collaborative initiative focused on genomic monitoring of SARS-CoV-2, conducted by the Fiocruz Genomic Surveillance Network—an entity under the Brazilian Ministry of Health. These samples were derived from the repurposing of rapid antigen tests conducted as part of routine clinical care, screening processes, and active surveillance for variants in hospitals and health centers in the state of Ceara, Brazil.

Integrative Genomic Analysis and Categorization. The paired-end sequencing was conducted during the routine process for genomic surveillance of SARS-CoV-2, employing the Artic v4.1 primer set (https://github.com/artic-network/artic-ncov2019/blob/master/primer_schemes/nCoV-2019/V4/SARS-CoV-2.primer.bed) in conjunction with the CovidSeq protocol, used as recommended by the manufacturer, implemented on the Illumina NextSeq 2000 platform for all samples. The raw sequencing data underwent a rigorous analysis utilizing the ViralFlow v1.0.0 workflow (https://viralflow.github.io/), which encompasses quality control, pre-processing, alignment of high-quality reads to the reference genome and genome assembly. Lineage classification was executed utilizing the Pangolin v4.3.1 [36] and Nextclade v3.0.1 [37] softwares, which facilitated the identification and annotation of genetic variations.

Sixteen high-quality sequencing samples were meticulously chosen for the validation step based on stringent criteria, ensuring horizontal coverage exceeding 90% and vertical coverage surpassing 100x. These samples were drawn from two distinct groups: the 'BE.9' group, comprising S01 to S08, and the 'non-BE.9' group (other lineages), consisting of S09 to S16, based on lineage classification generated by Pangolin. The BAM file was evaluated

using the Geneious prime software and the coverage variation throughout the genome was used to predict the 244 bp deletion present in the ORF7a gene of BE.9 group.

To confirm the presence of the anticipated ORF7a deletion, genomic DNA from each selected sample underwent 2% agarose gel electrophoresis. Electrophoresis was conducted for 4 hours at 90V, facilitating thorough separation and visualization of DNA fragments, including the targeted deletion in ORF7a. The bands were viewed using ThermoFisher iBright equipment, allowing instantaneous image generation.

Machine learning algorithms and data analysis. A total of 1,724 curves of the optimized protocol were manually analyzed and categorized based on previously established standards as 'BE.9', 'non-BE.9', or 'Inconclusive', and the curve points were served as input for model training. The total curves were separated into a matrix X, containing all points from all curves, and a vector y, containing correct classification values for each curve. The 192nd point of each curve (last column of matrix X) was removed due to 475 samples having a null value at this position. Following, the data was split into training sets (60%, n = 1034), evaluate set (20%, n = 345) and test sets (20%, n = 345). Subsequently, X values were normalized to a range of 0 to 1, crucial for unbiased training of two employed models and the X train was balanced to prevent over representative class bias.

Three machine learning algorithms were employed for data modeling: Gradient Boosting (GB), Support Vector Machine (SVM), and Logistic Regression (LR). Models were run with default parameters, except for SVM, where the 'kernel' parameter was changed to 'linear' instead of 'rbf.' The analysis was conducted using Python 3.10.12 in conjunction with the Scikit-learn 1.4.0 library (for Support Vector Machine and Logistic Regression) and XGBoost 2.0.3 package (for Gradient Boosting), all implemented within the Google Colaboratory environment. The code used for training the machine learning models is available in Supplementary Material 1.

The model, exhibiting the highest accuracy, reflecting overall correctness, underwent a grid search optimization step, exploring different parameters for fine-tuning, with a particular focus on optimizing for the accuracy parameter. The supplementary table 3 compiles the results of the grid search [38].

ASSOCIATED CONTENT

Supporting information

Supplementary Table 1 RT-qPCR Protocol volumes for Manufacturer's Protocol vs. Optimized Protocol in the BRYT Green Assay for SARS-CoV-2 Group Differentiation (.xlsx).

Supplementary Table 2 RT-qPCR Cycling for Manufacturer's Protocol vs. Optimized Protocol in the BRYT Green Assay for SARS-CoV-2 Group Differentiation (.xlsx).

Supplementary Material 1 Jupyter Notebook with the codes for training machine learning models (.ipynb).

Supplementary Table 3 Ranking of hyperparameters of SVM model (.xlsx).

REFERENCES

(1) Abulsoud, A. I.; El-Husseiny, H. M.; El-Husseiny, A. A.; El-Mahdy, H. A.; Ismail, A.; Elkhawaga, S. Y.; Khidr, E. G.; Fathi, D.; Mady, E. A.; Najda, A.; Algahtani, M.; Theyab, A.; Alsharif, K. F.; Albrakati, A.; Bayram, R.; Abdel-Daim, M. M.; Doghish, A. S. Mutations in SARS-CoV-2: Insights on Structure, Variants, Vaccines, and Biomedical Interventions. *Biomedicine and Pharmacotherapy*. Elsevier Masson s.r.l. January 1, 2023. https://doi.org/10.1016/j.biopha.2022.113977.

(2) Tomaszewski, T.; DeVries, R. S.; Dong, M.; Bhatia, G.; Norsworthy, M. D.; Zheng, X.; Caetano-Anollés, G. New Pathways of Mutational Change in SARS-CoV-2 Proteomes Involve Regions of Intrinsic Disorder Important for Virus Replication and Release. *Evolutionary Bioinformatics* **2020**, *16*. https://doi.org/10.1177/1176934320965149.

(3) Jeronimo, P. M. C.; Aksenen, C. F.; Duarte, I. O.; Lins, R. D.; Miyajima, F. Evolutionary Deletions within the SARS-CoV-2 Genome as Signature Trends for Virus Fitness and Adaptation. *J Virol* **2023**. https://doi.org/10.1128/jvi.01404-23.

(4) Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; Robertson, D. L. SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nature Reviews Microbiology*. Nature Research July 1, 2021, pp 409–424. https://doi.org/10.1038/s41579-021-00573-0.

(5) Carabelli, A. M.; Peacock, T. P.; Thorne, L. G.; Harvey, W. T.; Hughes, J.; de Silva, T. I.; Peacock, S. J.; Barclay, W. S.; de Silva, T. I.; Towers, G. J.; Robertson, D. L. SARS-CoV-2 Variant Biology: Immune Escape, Transmission and Fitness. *Nature Reviews Microbiology*. Nature Research March 1, 2023, pp 162–177. https://doi.org/10.1038/s41579-022-00841-7.

(6) Schneider, W. L.; Roossinck, M. J. Genetic Diversity in RNA Virus Quasispecies Is Controlled by Host-Virus Interactions. *J Virol* **2001**, *75* (14), 6566–6571. https://doi.org/10.1128/jvi.75.14.6566-6571.2001.

(7) Villa, T. G.; Abril, A. G.; Sánchez, S.; de Miguel, T.; Sánchez-Pérez, A. Animal and Human RNA Viruses: Genetic Variability and Ability to Overcome Vaccines. *Archives of Microbiology*. Springer Science and Business Media Deutschland GmbH March 1, 2021, pp 443–464. https://doi.org/10.1007/s00203-020-02040-5.

(8) He, X.; Hong, W.; Pan, X.; Lu, G.; Wei, X. SARS-CoV-2 Omicron Variant: Characteristics and Prevention. *MedComm*. John Wiley and Sons Inc December 1, 2021, pp 838–845. https://doi.org/10.1002/mco2.110.

(9) Fan, Y.; Li, X.; Zhang, L.; Wan, S.; Zhang, L.; Zhou, F. SARS-CoV-2 Omicron Variant: Recent Progress and Future Perspectives. *Signal Transduction and Targeted Therapy*. Springer Nature December 1, 2022. https://doi.org/10.1038/s41392-022-00997-x.

(10) Zhao, H.; Lu, L.; Peng, Z.; Chen, L. L.; Meng, X.; Zhang, C.; Ip, J. D.; Chan, W. M.; Chu, A. W. H.; Chan, K. H.; Jin, D. Y.; Chen, H.; Yuen, K. Y.; To, K. K. W. SARS-CoV-2 Omicron Variant Shows Less Efficient Replication and Fusion Activity When Compared with Delta Variant in TMPRSS2-Expressed Cells. *Emerg Microbes Infect* **2022**, *11* (1), 277–283. https://doi.org/10.1080/22221751.2021.2023329.

(11) Wrenn, J. O.; Pakala, S. B.; Vestal, G.; Shilts, M. H.; Brown, H. M.; Bowen, S. M.; Strickland, B. A.; Williams, T.; Mallal, S. A.; Jones, I. D.; Schmitz, J. E.; Self, W. H.; Das, S. R. COVID-19 Severity from Omicron and Delta SARS-CoV-2 Variants. *Influenza Other Respir Viruses* **2022**, *16* (5), 832–836. https://doi.org/10.1111/irv.12982.

(12) Backer, J. A.; Eggink, D.; Andeweg, S. P.; Veldhuijzen, I. K.; van Maarseveen, N.; Vermaas, K.; Vlaemynck, B.; Schepers, R.; van den Hof, S.; Reusken, C. B. E. M.; Wallinga, J. Shorter Serial Intervals in SARS-CoV-2 Cases with Omicron BA.1 Variant Compared with Delta Variant, the Netherlands, 13 to 26 December 2021. *Eurosurveillance* **2022**, *27* (6). https://doi.org/10.2807/1560-7917.ES.2022.27.6.2200042.

(13) Lyngse, F. P.; Mortensen, L. H.; Denwood, M. J.; Christiansen, L. E.; Møller, C. H.; Skov, R. L.; Spiess, K.; Fomsgaard, A.; Lassaunière, R.; Rasmussen, M.; Stegger, M.; Nielsen, C.; Sieber, R. N.; Cohen, A. S.; Møller, F. T.; Overvad, M.; Mølbak, K.; Krause, T. G.; Kirkeby, C. T. Household Transmission of the SARS-CoV-2 Omicron Variant in Denmark. *Nat Commun* **2022**, *13* (1). https://doi.org/10.1038/s41467-022-33328-3.

(14) Mistry, P.; Barmania, F.; Mellet, J.; Peta, K.; Strydom, A.; Viljoen, I. M.; James, W.; Gordon, S.; Pepper, M. S. SARS-CoV-2 Variants, Vaccines, and Host Immunity. *Frontiers in Immunology*. Frontiers Media S.A. January 3, 2022. https://doi.org/10.3389/fimmu.2021.809244.

(15) Mohsin, M.; Mahmud, S. Omicron SARS-CoV-2 Variant of Concern: A Review on Its Transmissibility, Immune Evasion, Reinfection, and Severity. *Medicine (United States)*. Lippincott Williams and Wilkins May 13, 2022, p E29165. https://doi.org/10.1097/MD.0000000000029165.

(16) Kim, D.; Ali, S. T.; Kim, S.; Jo, J.; Lim, J. S.; Lee, S.; Ryu, S. Estimation of Serial Interval and Reproduction Number to Quantify the Transmissibility of SARS-CoV-2 Omicron Variant in South Korea. *Viruses* **2022**, *14* (3). https://doi.org/10.3390/v14030533.

(17) Elbe, S.; Buckland-Merrett, G. Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health. *Global Challenges* **2017**, *1* (1), 33–46. https://doi.org/10.1002/gch2.1018.

(18) Inzaule, S. C.; Tessema, S. K.; Kebede, Y.; Ogwell Ouma, A. E.; Nkengasong, J. N. Genomic-Informed Pathogen Surveillance in Africa: Opportunities and Challenges. *The Lancet Infectious Diseases*. Lancet Publishing Group September 1, 2021, pp e281–e289. https://doi.org/10.1016/S1473-3099(20)30939-7.

(19) Brito, A. F.; Semenova, E.; Dudas, G.; Hassler, G. W.; Kalinich, C. C.; Kraemer, M. U. G.; Ho, J.; Tegally, H.; Githinji, G.; Agoti, C. N.; Matkin, L. E.; Whittaker, C.; Kantardjiev, T.; Korsun, N.; Stoitsova, S.; Dimitrova, R.; Trifonova, I.; Dobrinov, V.; Grigorova, L.; Stoykov, I.; Grigorova, I.; Gancheva, A.; Jennison, A.; Leong, L.; Speers, D.; Baird, R.; Cooley, L.; Kennedy, K.; de Ligt, J.; Rawlinson, W.; van Hal, S.; Williamson, D.; Singh, R.; Nathaniel-Girdharrie, S. M.; Edghill, L.; Indar, L.; St. John, J.; Gonzalez-Escobar, G.; Ramkisoon, V.; Brown-Jordan, A.; Ramjag, A.; Mohammed, N.; Foster, J. E.; Potter, I.; Greenaway-Duberry, S.; George, K.; Belmar-George, S.; Lee, J.; Bisasor-McKenzie, J.; Astwood, N.; Sealey-Thomas, R.; Laws, H.; Singh, N.; Oyinloye, A.; McMillan, P.; Hinds, A.; Nandram, N.; Parasram, R.; Khan-Mohammed, Z.; Charles, S.; Andrewin, A.;

Johnson, D.; Keizer-Beache, S.; Oura, C.; Pybus, O. G.; Faria, N. R.; Stegger, M.; Albertsen, M.; Fomsgaard, A.; Rasmussen, M.; Khouri, R.; Naveca, F.; Graf, T.; Miyajima, F.; Wallau, G.; Motta, F.; Khare, S.; Freitas, L.; Schiavina, C.; Bach, G.; Schultz, M. B.; Chew, Y. H.; Makheja, M.; Born, P.; Calegario, G.; Romano, S.; Finello, J.; Diallo, A.; Lee, R. T. C.; Xu, Y. N.; Yeo, W.; Tiruvayipati, S.; Yadahalli, S.; Wilkinson, E.; Iranzadeh, A.; Giandhari, J.; Doolabh, D.; Pillay, S.; Ramphal, U.; San, J. E.; Msomi, N.; Mlisana, K.; von Gottberg, A.; Walaza, S.; Ismail, A.; Mohale, T.; Engelbrecht, S.; Van Zyl, G.; Preiser, W.; Sigal, A.; Hardie, D.; Marais, G.; Hsiao, M.; Korsman, S.; Davies, M. A.; Tyers, L.; Mudau, I.; York, D.; Maslo, C.; Goedhals, D.; Abrahams, S.; Laguda-Akingba, O.; Alisoltani-Dehkordi, A.; Godzik, A.; Wibmer, C. K.; Martin, D.; Lessells, R. J.; Bhiman, J. N.; Williamson, C.; de Oliveira, T.; Chen, C.; Nadeau, S.; du Plessis, L.; Beckmann, C.; Redondo, M.; Kobel, O.; Noppen, C.; Seidel, S.; de Souza, N. S.; Beerenwinkel, N.; Topolsky, I.; Jablonski, P.; Fuhrmann, L.; Dreifuss, D.; Jahn, K.; Ferreira, P.; Posada-Céspedes, S.; Beisel, C.; Denes, R.; Feldkamp, M.; Nissen, I.; Santacroce, N.; Burcklen, E.; Aquino, C.; de Gouvea, A. C.; Moccia, M. D.; Grüter, S.; Sykes, T.; Opitz, L.; White, G.; Neff, L.; Popovic, D.; Patrignani, A.; Tracy, J.; Schlapbach, R.; Dermitzakis, E.; Harshman, K.; Xenarios, I.; Pegeot, H.; Cerutti, L.; Penet, D.; Stadler, T.; Howden, B. P.; Sintchenko, V.; Zuckerman, N. S.; Mor, O.; Blankenship, H. M.; de Oliveira, T.; Lin, R. T. P.; Siqueira, M. M.; Resende, P. C.; Vasconcelos, A. T. R.; Spilki, F. R.; Aguiar, R. S.; Alexiev, I.; Ivanov, I. N.; Philipova, I.; Carrington, C. V. F.; Sahadeo, N. S. D.; Branda, B.; Gurry, C.; Maurer-Stroh, S.; Naidoo, D.; von Eije, K. J.; Perkins, M. D.; van Kerkhove, M.; Hill, S. C.; Sabino, E. C.; Pybus, O. G.; Dye, C.; Bhatt, S.; Flaxman, S.; Suchard, M. A.; Grubaugh, N. D.; Baele, G.; Faria, N. R. Global Disparities in SARS-CoV-2 Genomic Surveillance. *Nat Commun* **2022**, *13* (1). https://doi.org/10.1038/s41467-022-33713-y.

(20) Godmer, A.; Bigot, J.; Giai Gianetto, Q.; Benzerara, Y.; Veziris, N.; Aubry, A.; Guitard, J.; Hennequin, C. Machine Learning to Improve the Interpretation of Intercalating Dye-Based Quantitative PCR Results. *Sci Rep* **2022**, *12* (1). https://doi.org/10.1038/s41598-022-21010-z.

(21) Langer, T.; Favarato, M.; Giudici, R.; Bassi, G.; Garberi, R.; Villa, F.; Gay, H.; Zeduri, A.; Bragagnolo, S.; Molteni, A.; Beretta, A.; Corradin, M.; Moreno, M.; Vismara, C.; Perno, C. F.; Buscema, M.; Grossi, E.; Fumagalli, R. Development of Machine Learning Models to Predict RT-PCR Results for Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in Patients with Influenza-like Symptoms Using Only Basic Clinical Data. *Scand J Trauma Resusc Emerg Med* **2020**, *28* (1). https://doi.org/10.1186/s13049-020-00808-8.

(22) Nemudryi, A.; Nemudraia, A.; Wiegand, T.; Nichols, J.; Snyder, D. T.; Hedges, J. F.; Cicha, C.; Lee, H.; Vanderwood, K. K.; Bimczok, D.; Jutila, M. A.; Wiedenheft, B. SARS-CoV-2 Genomic Surveillance Identifies Naturally Occurring Truncation of ORF7a That Limits Immune Suppression. *Cell Rep* **2021**, *35* (9). https://doi.org/10.1016/j.celrep.2021.109197.

(23) Pyke, A. T.; Nair, N.; van den Hurk, A. F.; Burtonclay, P.; Nguyen, S.; Barcelon, J.; Kistler, C.; Schlebusch, S.; McMahon, J.; Moore, F. Replication Kinetics of b.1.351 and b.1.1.7 Sars-Cov-2 Variants of Concern Including Assessment of a b.1.1.7 Mutant Carrying a Defective Orf7a Gene. *Viruses* **2021**, *13* (6). https://doi.org/10.3390/v13061087.

(24) Lucas, S.; Jones, M. S.; Kothari, S.; Madlambayan, A.; Ngo, C.; Chan, C.; Goraichuk, I. V. A 336-Nucleotide in-Frame Deletion in ORF7a Gene of SARS-CoV-2 Identified in Genomic Surveillance by next-Generation Sequencing. *Journal of Clinical Virology*. Elsevier B.V. March 1, 2022. https://doi.org/10.1016/j.jcv.2022.105105.

(25) Foster, C. S.; Rawlinson MBBS, W. D. Rapid Spread of a SARS-CoV-2 Delta Variant with a Frameshift Deletion in ORF7a. https://doi.org/10.1101/2021.08.18.21262089.

(26) Fuchs Wightman, F.; Godoy Herz, M. A.; Muñoz, J. C.; Stigliano, J. N.; Bragado, L.; Moreno, N. N.; Palavecino, M.; Servi, L.; Cabrerizo, G.; Clemente, J.; Avaro, M.; Pontoriero, A.; Benedetti, E.; Baumeister, E.; Rudolf, F.; Remes Lenicov, F.; Garcia, C.; Buggiano, V.; Kornblihtt, A. R.; Srebrow, A.; de la Mata, M.; Muñoz, M. J.; Schor, I. E.; Petrillo, E. A DNA Intercalating Dye-Based RT-QPCR Alternative to Diagnose SARS-CoV-2. *RNA Biol* **2021**, *18* (12), 2218–2225. https://doi.org/10.1080/15476286.2021.1926648.

(27) Watzinger, F.; Ebner, K.; Lion, T. Detection and Monitoring of Virus Infections by Real-Time PCR. *Molecular Aspects of Medicine*. April 2006, pp 254–298. https://doi.org/10.1016/j.mam.2005.12.001.

(28) Gudnason, H.; Dufva, M.; Bang, D. D.; Wolff, A. Comparison of Multiple DNA Dyes for Real-Time PCR: Effects of Dye Concentration and Sequence Composition on DNA Amplification and Melting Temperature. *Nucleic Acids Res* **2007**, *35* (19). https://doi.org/10.1093/nar/gkm671.

(29) Vossen, R. H. A. M.; Aten, E.; Roos, A.; Den Dunnen, J. T. High-Resolution Melting Analysis (HRMA) - More than Just Sequence Variant Screening. *Human Mutation*. June 2009, pp 860–866. https://doi.org/10.1002/humu.21019.

(30) Larios, G.; Ribeiro, M.; Arruda, C.; Oliveira, S. L.; Canassa, T.; Baker, M. J.; Marangoni, B.; Ramos, C.; Cena, C. A New Strategy for Canine Visceral Leishmaniasis Diagnosis Based on FTIR Spectroscopy and Machine Learning. *J Biophotonics* **2021**, *14* (11). https://doi.org/10.1002/jbio.202100141.

(31) Pacher, G.; Franca, T.; Lacerda, M.; Alves, N. O.; Piranda, E. M.; Arruda, C.; Cena, C. Diagnosis of Cutaneous Leishmaniasis Using FTIR Spectroscopy and Machine Learning: An Animal Model Study. *ACS Infect Dis* **2024**, *10* (2), 467–474. https://doi.org/10.1021/acsinfecdis.3c00430.

(32) Athamanolap, P.; Parekh, V.; Fraley, S. I.; Agarwal, V.; Shin, D. J.; Jacobs, M. A.; Wang, T. H.; Yang, S. Trainable High Resolution Melt Curve Machine Learning Classifier for Large-Scale Reliable Genotyping of Sequence Variants. *PLoS One* **2014**, *9* (10). https://doi.org/10.1371/journal.pone.0109094.

(33) Nakhle, F.; Harfouche, A. L. Ready, Steady, Go AI: A Practical Tutorial on Fundamentals of Artificial Intelligence and Its Applications in Phenomics Image Analysis. *Patterns*. Cell Press September 10, 2021. https://doi.org/10.1016/j.patter.2021.100323.

(34) Engelberger, F.; Galaz-Davison, P.; Bravo, G.; Rivera, M.; Ramírez-Sarmiento, C. A. Developing and Implementing Cloud-Based Tutorials That Combine Bioinformatics Software, Interactive Coding, and Visualization Exercises for Distance Learning on

Structural Bioinformatics. *J Chem Educ* **2021**, *98* (5), 1801–1807. https://doi.org/10.1021/acs.jchemed.1c00022.

(35) Carneiro, T.; Da Nobrega, R. V. M.; Nepomuceno, T.; Bian, G. Bin; De Albuquerque, V. H. C.; Filho, P. P. R. Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access* **2018**, *6*, 61677–61685. https://doi.org/10.1109/ACCESS.2018.2874767.

(36) O'Toole, Á.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J. T.; Colquhoun, R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; Yeats, C.; du Plessis, L.; Maloney, D.; Medd, N.; Attwood, S. W.; Aanensen, D. M.; Holmes, E. C.; Pybus, O. G.; Rambaut, A. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus Evol* **2021**, *7* (2). https://doi.org/10.1093/ve/veab064.

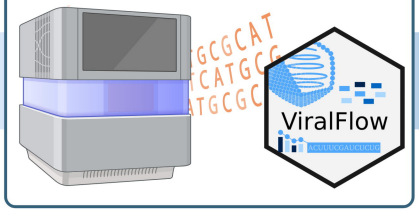
(37) Aksamentov, I.; Roemer, C.; Hodcroft, E.; Neher, R. Nextclade: Clade Assignment, Mutation Calling and Quality Control for Viral Genomes. *J Open Source Softw* **2021**, *6* (67), 3773. https://doi.org/10.21105/joss.03773.

(38) Noble, W. S. *What Is a Support Vector Machine?*; 2006; Vol. 24. https://doi.org/https://doi.org/10.1038/nbt1206-1565.

Validation step (n = 16)

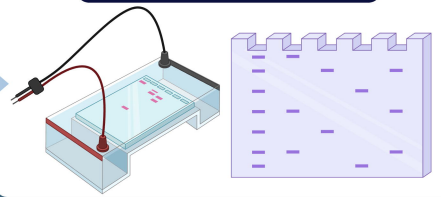
Origin and aquisition of samples

Genomic analyzes by NGS

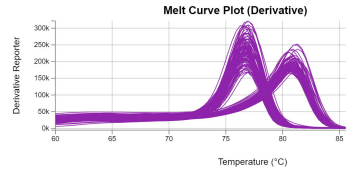


High quality samples

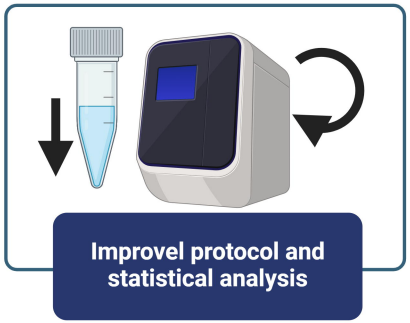
Categorization and confirmation of deletion



Intercalating Dye-based qPCR amplification protocol



Protocol optimization step (n = 70)



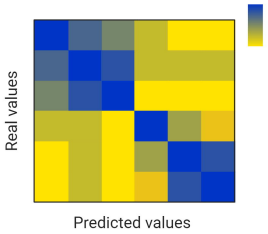
Improvel protocol and statistical analysis

- Performances estimations
  -> Kappa coefficients and
  -> Standard deviation

Automatic classification step (n = 1724)

Algorithms:

- Gradient Boosting (GB)
- Support Vector Machine (SVM)
- Logistic Regression (LR)



- Data analysis and model refinement

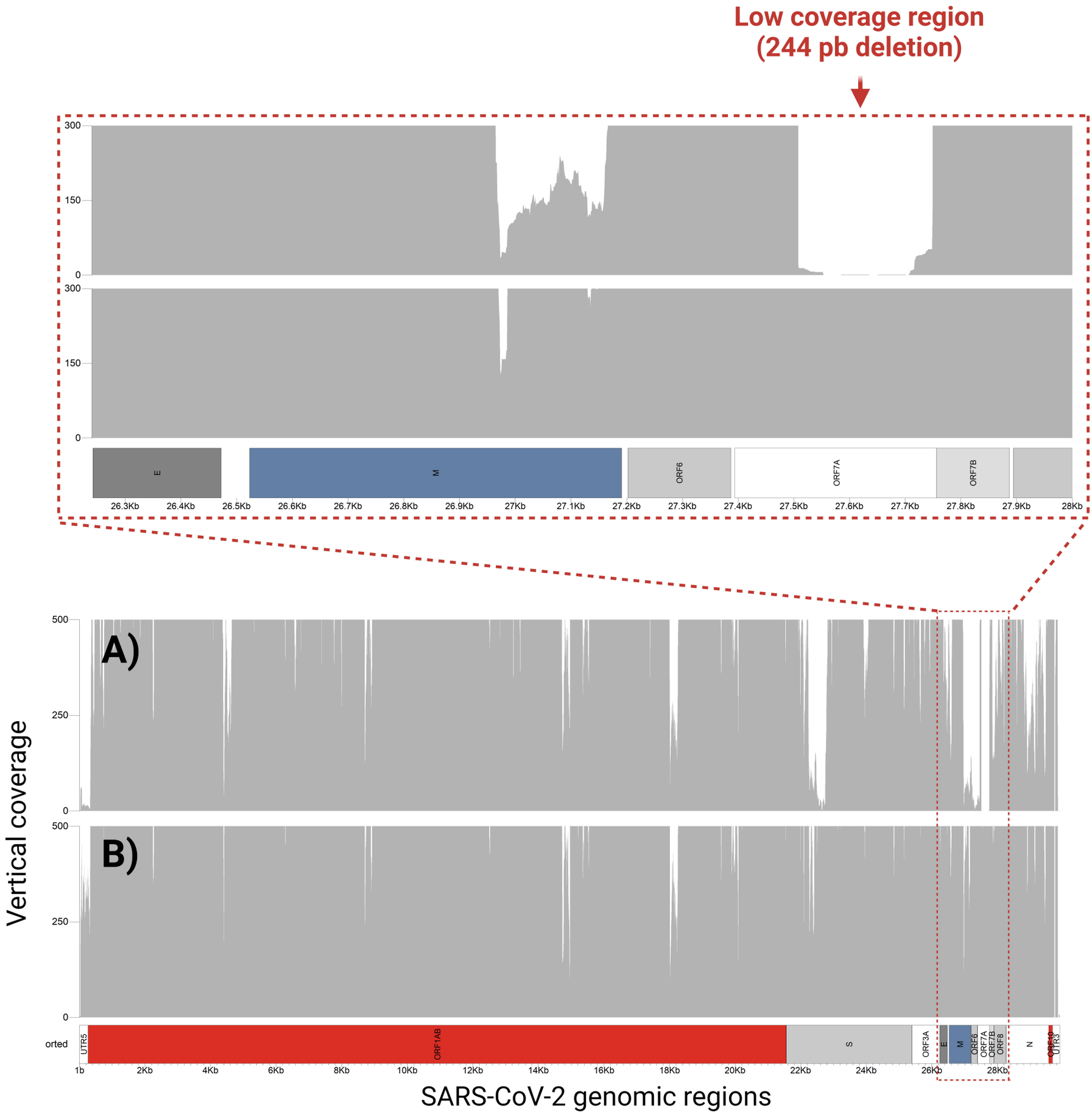
Test set (25%)
n = 431

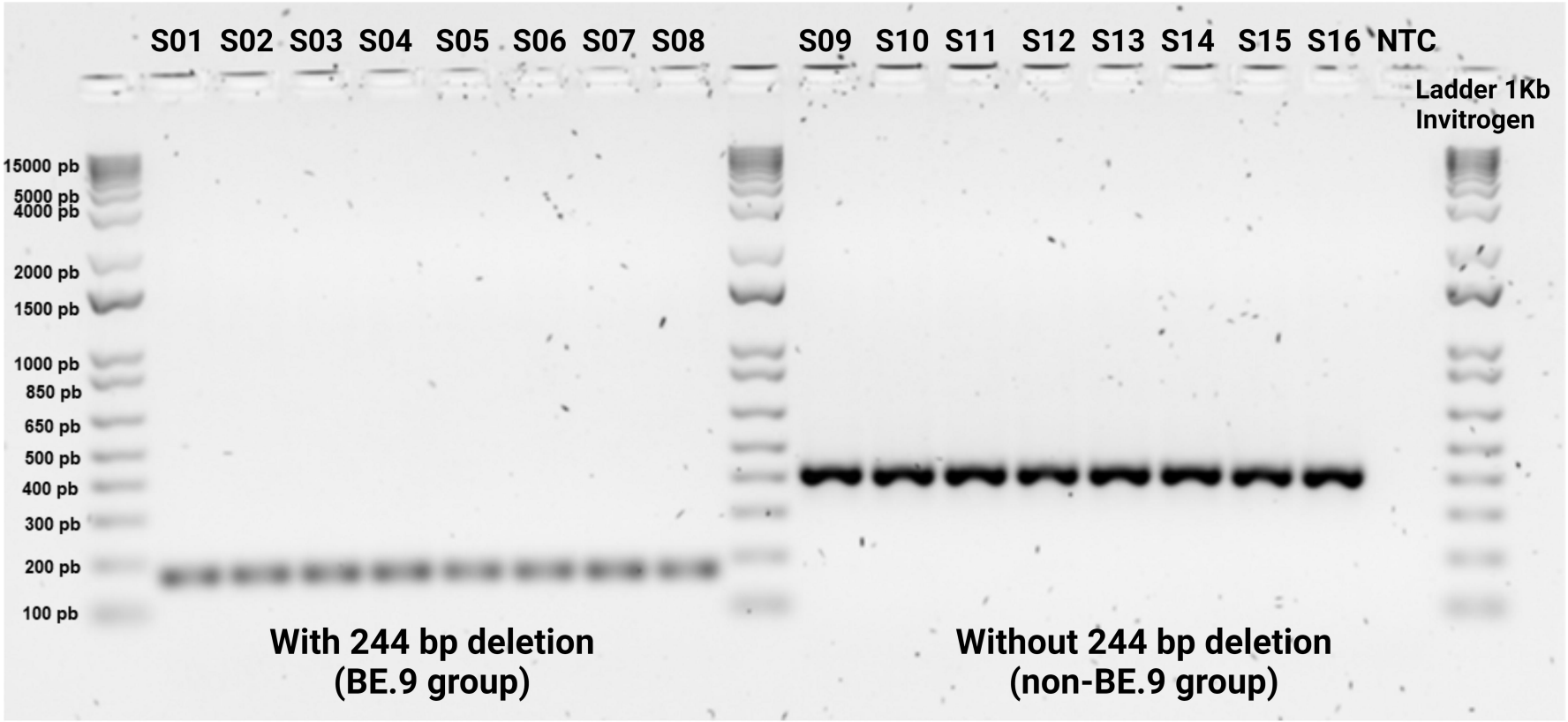
Training set (75%)
n = 1293

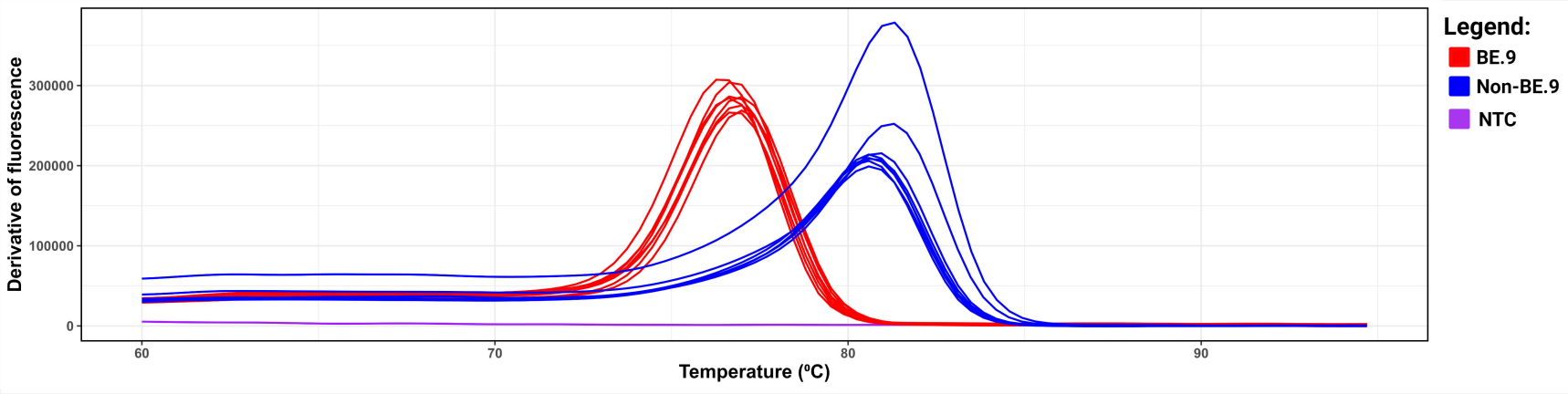
Manual classification

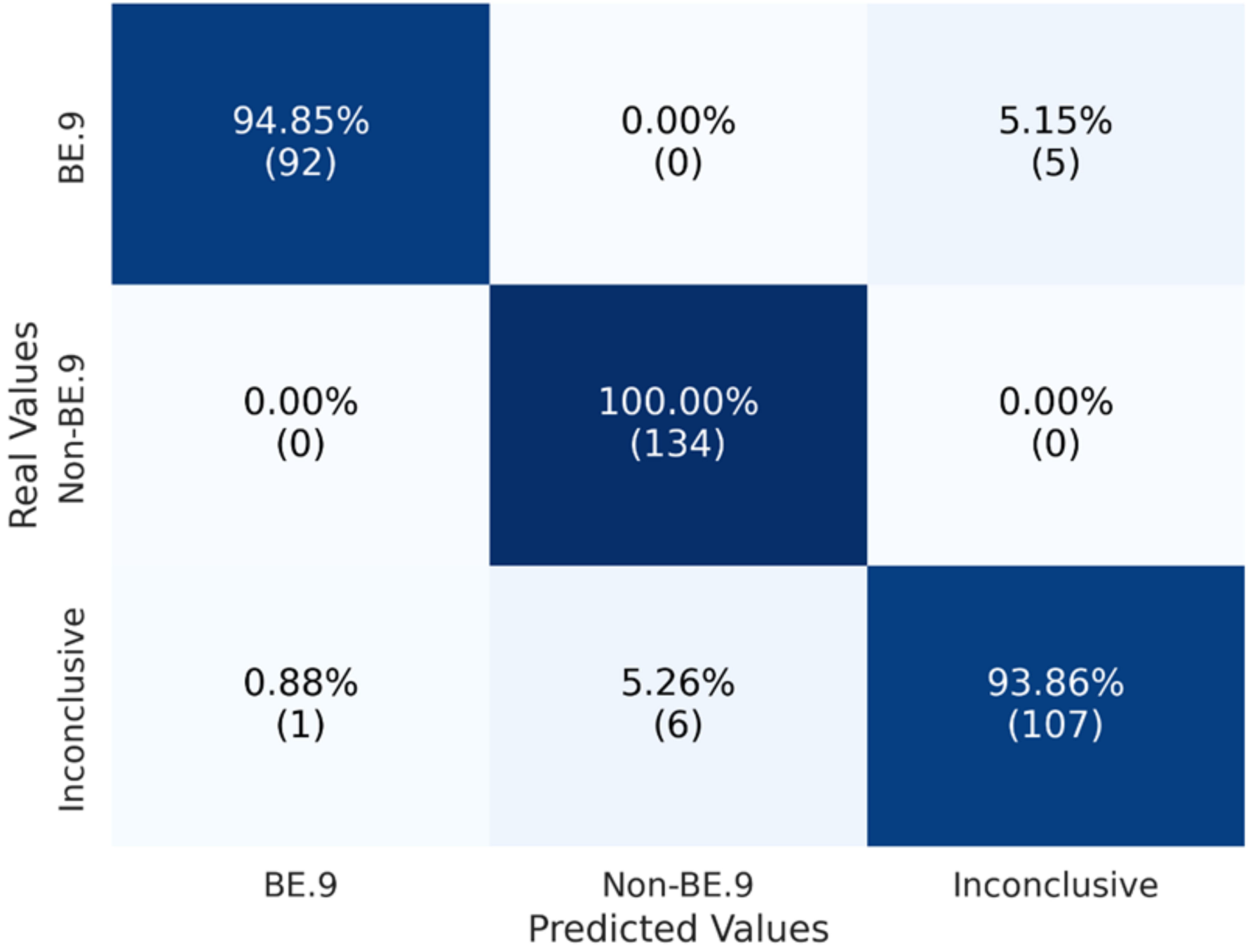
X

Machine learning algorithms


Low coverage region
(244 pb deletion)
300
150
0
E
M
ORF6
ORF7A
ORF7B
26.3Kb
26.4Kb
26.5Kb
26.6Kb
26.7Kb
26.8Kb
26.9Kb
27Kb
27.1Kb
27.2Kb
27.3Kb
27.4Kb
27.5Kb
27.6Kb
27.7Kb
27.8Kb
27.9Kb
28Kb
A)
B)
500
250
Vertical coverage
UTR5
ORF1AB
S
ORF3A
E
M
ORF6
ORF7A
ORF7B
ORF8
N
ORF10
UTR3
1b
2Kb
4Kb
6Kb
8Kb
10Kb
12Kb
14Kb
16Kb
18Kb
20Kb
22Kb
24Kb
26Kb
28Kb
SARS-CoV-2 genomic regions


S01 S02 S03 S04 S05 S06 S07 S08
S09 S10 S11 S12 S13 S14 S15 S16 NTC
Ladder 1Kb
Invitrogen
15000 pb
5000 pb
4000 pb
2000 pb
1500 pb
1000 pb
850 pb
650 pb
500 pb
400 pb
300 pb
200 pb
100 pb
With 244 bp deletion
(BE.9 group)
Without 244 bp deletion
(non-BE.9 group)


Legend:
BE.9
Non-BE.9
NTC
Derivative of fluorescence
300000
200000
100000
0
60
70
80
90
Temperature (ºC)


Real Values
BE.9
Non-BE.9
Inconclusive
94.85%
(92)
0.00%
(0)
5.15%
(5)
0.00%
(0)
100.00%
(134)
0.00%
(0)
0.88%
(1)
5.26%
(6)
93.86%
(107)
BE.9
Non-BE.9
Inconclusive
Predicted Values