

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

4 Cleber Furtado Aksenén^{1*}, Débora Maria Almeida Ferreira^{1*}, Pedro Miguel Carneiro
5 Jeronimo^{1*}, Thais de Oliveira Costa¹, Ticiane Cavalcante de Souza¹, Bruna Maria
6 Nepomuceno Sousa Lino¹, Allysson Allan de Farias¹, Fabio Miyajima¹

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

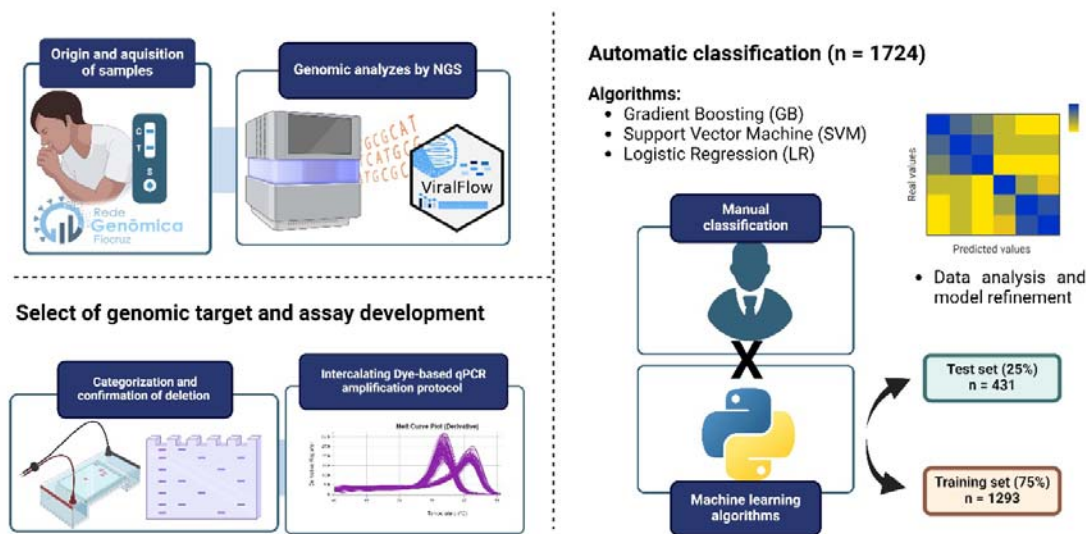
8 *These authors contributed equally to this article.

9 **ABSTRACT**

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

32 **Keywords**

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



35

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

43 INTRODUCTION

44 The adaptability of viruses like SARS-CoV-2 through cumulative mutations denotes the
45 dynamic interaction between pathogens and their environment. Mutations leading to
46 structural modifications, such as insertions or deletions, are more likely to account for
47 significant alterations in the biological behavior of the virus, ultimately fueling the emergence
48 of variants with potential selective advantage and pathogenic profiles¹⁻³. This adaptive
49 mechanism has been illustrated by the emergence of SARS-CoV-2 variants, which is known
50 for its increased infectivity due to specific amino acid substitutions^{4,5}. The genetic diversity
51 observed in RNA viruses, underscored by the continuous emergence of new mutations,
52 highlights the evolving nature of these pathogens and the critical role of genomic
53 surveillance in tracking these changes^{3,6,7}. The swift emergence and global proliferation of
54 the Omicron variant (B.1.1.529) of SARS-CoV-2, along with its descendant subvariants,
55 have heightened global apprehensions because of their extensive repertoire of distinctive
56 genetic configurations and unprecedented transmission capabilities^{8,9}. Studies have
57 illuminated the variant's ability to outpace previous strains, such as the Delta variant, in
58 terms of spread, leading to a considerable uptick in reinfection rates, affecting even those
59 previously vaccinated or infected¹⁰⁻¹³. The situation is compounded by the variant's elusive
60 severity profile compared with its predecessors, necessitating rigorous public health
61 interventions¹⁴⁻¹⁶. Given the dynamic nature of the virus, heightened emphasis on genomic
62 surveillance is imperative to track and understand the emergence of new strains, enabling
63 proactive measures to mitigate their spread and impact.

64 The global effort to monitor and control the spread of SARS-CoV-2 faces numerous
65 challenges, including the economic and infrastructural disparities among countries. The
66 challenges posed by NGS analyses, including high costs, lengthy response times, and its
67 inaccessibility in economically disadvantaged regions, have spurred the scientific community
68 to explore supplementary techniques¹⁷⁻¹⁹. There has been a notable change toward

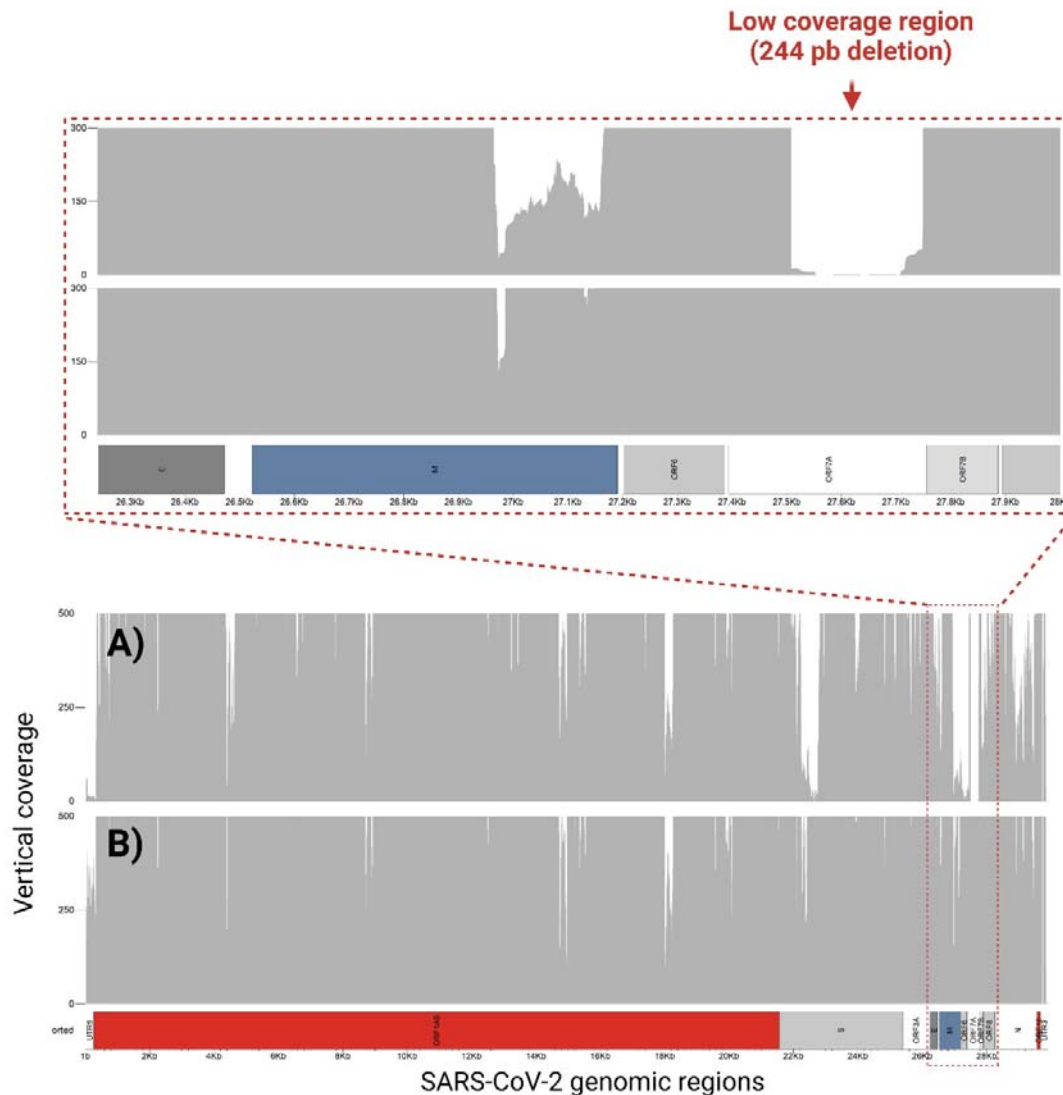
69 integrating polymerase chain reaction (PCR) and computational algorithms into the genomic
70 surveillance toolkit. These methods offer a more immediate and cost-effective capability for
71 detecting specific genetic markers, thereby enhancing the efficiency and scope of pathogen
72 surveillance efforts^{20,21}.

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

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

99 **RESULTS AND DISCUSSION**

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

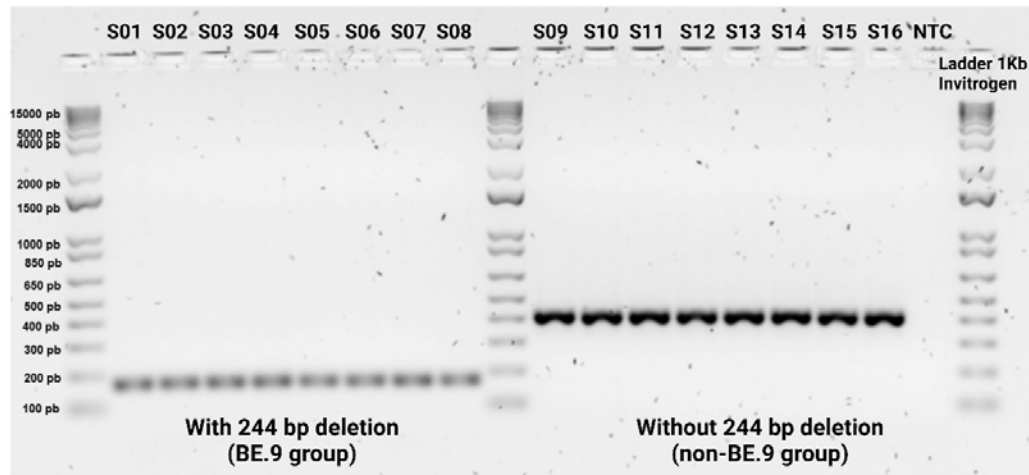


105

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

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

121 independent origins, and reinforce findings from previous studies concerning the loss of
122 genetic elements during the natural evolution of SARS-CoV-2³. Additionally, it also
123 contributed to obtaining evidence regarding the applicability of a PCR amplification protocol
124 that makes use of intercalating dye-based strategies, aiming to increase speediness and
125 robustness of the investigations.

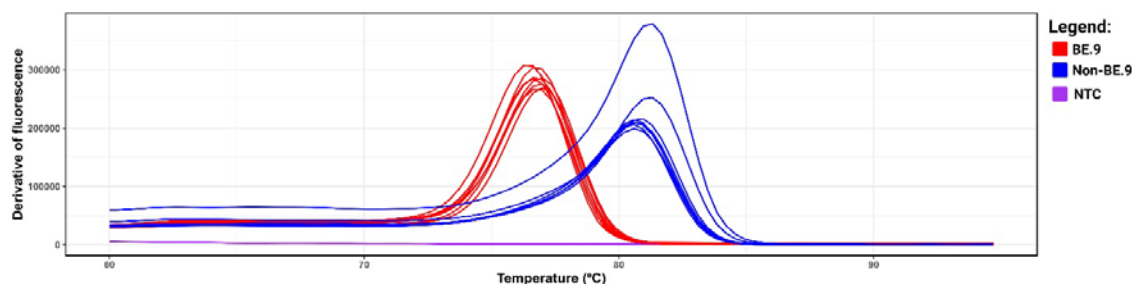


126

127 **Figure 3 Agarose gel electrophoresis of amplified DNA fragments to validate the ORF7a**
128 **deletion.** The 'BE.9' groups (S01 to S08) show the absence of part of the band corresponding to
129 ORF7a, located around 244 base pairs (bp), while the 'non-BE.9' groups (S09 to S16) show bands
130 intact sections of ORF7a, between 400 and 430 bp.
131

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

138 The results obtained from the first derivative melt curves of the sixteen samples are clarity in
139 **Figure 4**, illustrating the melt curves corresponding to the BE.9 group and the non-BE.9
140 group. All melt curves for the BE.9 group exhibited amplification at an average melting
141 temperature (T_m) of 76.78 ± 0.18°C, accompanied by fluorescence levels ranging between
142 200k and 300k at its peak. In contrast, the non-BE.9 group displayed an average melting
143 temperature of 80.76 ± 0.24°C, with a wider range of fluorescence intensity, spanning from
144 200k to 400k. This confirmation, highlighted by the lower T_m for BE.9 and higher T_m for non-
145 BE.9, aligns with prior research suggesting that longer amplicons exhibit higher melting
146 temperatures (T_m) compared to shorter ones²⁸. Notably, it was observed during manual
147 analysis and categorization of the samples that the melting curves with fluorescence levels
148 below 100k were challenging to visualize and classify accurately. This challenge was
149 significantly alleviated when the range was filtered to values greater than 100k, enhancing
150 the clarity and precision of group classification.



151

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

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

166 **Machine learning algorithms and data analysis.** The SVM with a linear kernel
 167 emerged as the best-performing model, surpassing Logistic Regression and Gradient
 168 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

169

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

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

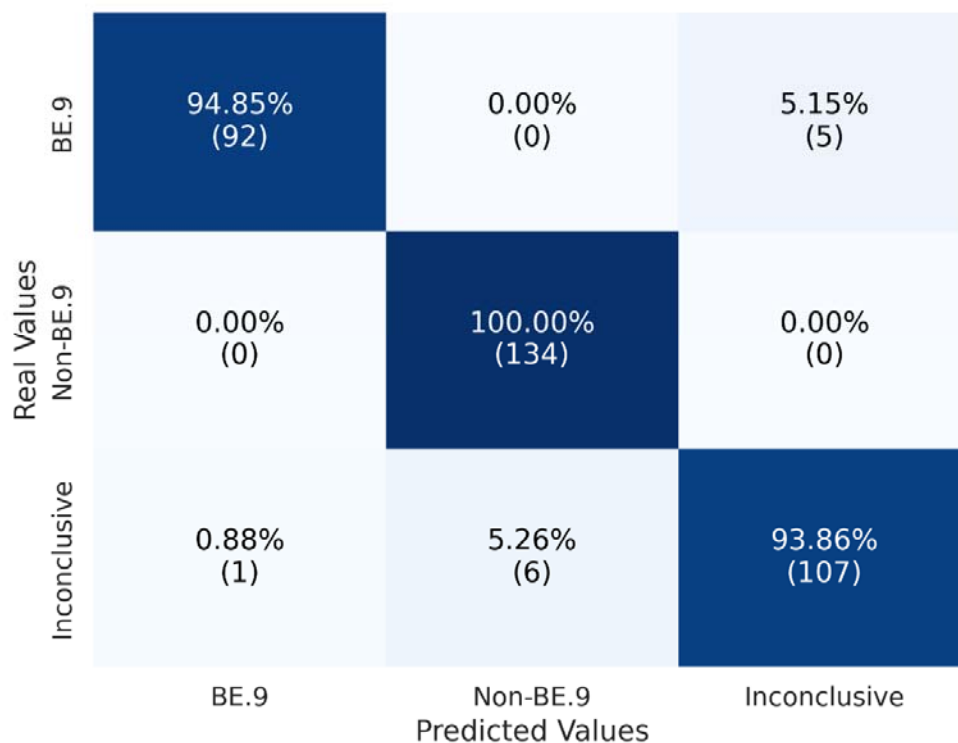
181 is imperative to assess these metrics on unseen data. While there was a decrease in
 182 metrics, it is probable that these values reflect the true performance on other unseen
 183 datasets.

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

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

187 The high accuracy signifies substantial reliability when utilizing melting curve points for curve
 188 classification, automating the process. Other studies have approached diagnostic
 189 classification using derived metrics, whether through Principal Component Analysis (PCA)
 190 ^{30,31}, metrics related to curve shape (skewness, kurtosis etc) ²⁰, or variables associated with
 191 the technique itself (amplicon melting temperature) ³², achieving accuracies ranging from
 192 72% to 100%. In this work, however, we chose to directly use the curve itself, transforming
 193 each point of every curve into a column, or feature, for the model, employing a simple data
 194 normalization step. This approach streamlines the model development process, ensuring
 195 simplicity without compromising accuracy. It is noteworthy that the clear distinction between
 196 curves for BE.9 and non-BE.9 classification enables this approach. By utilizing points from
 197 the curve directly, the model gains the flexibility to discern nuances indicating which
 198 individual points are more crucial for correct result classification.

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



206

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

210 The utilization of free platforms such as Google Colaboratory could contribute to the
211 democratization and swift investigation of outbreaks of new variants in regions lacking
212 computational power³³. Simple modeling from minimally processed data represents an
213 encouraging opportunity for other groups to optimize protocols, demystifying the use of
214 machine learning algorithms in routine laboratory procedures, allowing for biological
215 applications as already used for other purposes^{34,35}.

216 CONCLUSIONS

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

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

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

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

248 **METHODS**

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

255 **Integrative Genomic Analysis and Categorization.** The paired-end sequencing was
256 conducted during the routine process for genomic surveillance of SARS-CoV-2, employing
257 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
258 conjunction with the CovidSeq protocol, used as recommended by the manufacturer,
259 implemented on the Illumina NextSeq 2000 platform for all samples. The raw sequencing
260 data underwent a rigorous analysis utilizing the ViralFlow v1.0.0 workflow
261 (<https://viralflow.github.io/>), which encompasses quality control, pre-processing, alignment of
262 high-quality reads to the reference genome and genome assembly. Lineage classification
263 was executed utilizing the Pangolin v4.3.1³⁶ and Nextclade v3.0.1³⁷ softwares, which
264 facilitated the identification and annotation of genetic variations.
265

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

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

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

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

288 Three machine learning algorithms were employed for data modeling: Gradient Boosting
289 (GB), Support Vector Machine (SVM), and Logistic Regression (LR). Models were run with
290 default parameters, except for SVM, where the 'kernel' parameter was changed to 'linear'
291 instead of 'rbf.' The analysis was conducted using Python 3.10.12 in conjunction with the
292 Scikit-learn 1.4.0 library (for Support Vector Machine and Logistic Regression) and XGBoost
293 2.0.3 package (for Gradient Boosting), all implemented within the Google Colaboratory
294 environment. The code used for training the machine learning models is available in
295 Supplementary Material 1.

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

300 **ASSOCIATED CONTENT**

301 **Supporting information**

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

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

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

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

REFERENCES

- 310
311
- 312 (1) Abulsoud, A. I.; El-Husseiny, H. M.; El-Husseiny, A. A.; El-Mahdy, H. A.; Ismail, A.;
313 Elkhawaga, S. Y.; Khidr, E. G.; Fathi, D.; Mady, E. A.; Najda, A.; Algahtani, M.; Theyab,
314 A.; Alsharif, K. F.; Albrakati, A.; Bayram, R.; Abdel-Daim, M. M.; Doghish, A. S.
315 Mutations in SARS-CoV-2: Insights on Structure, Variants, Vaccines, and Biomedical
316 Interventions. *Biomedicine and Pharmacotherapy*. Elsevier Masson s.r.l. January 1,
317 2023. <https://doi.org/10.1016/j.biopha.2022.113977>.
- 318 (2) Tomaszewski, T.; DeVries, R. S.; Dong, M.; Bhatia, G.; Norsworthy, M. D.; Zheng, X.;
319 Caetano-Anollés, G. New Pathways of Mutational Change in SARS-CoV-2 Proteomes
320 Involve Regions of Intrinsic Disorder Important for Virus Replication and Release.
321 *Evolutionary Bioinformatics* **2020**, *16*. <https://doi.org/10.1177/1176934320965149>.
- 322 (3) Jeronimo, P. M. C.; Aksenon, C. F.; Duarte, I. O.; Lins, R. D.; Miyajima, F. Evolutionary
323 Deletions within the SARS-CoV-2 Genome as Signature Trends for Virus Fitness and
324 Adaptation. *J Virol* **2023**. <https://doi.org/10.1128/jvi.01404-23>.
- 325 (4) Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E.
326 M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; Robertson, D. L. SARS-CoV-2
327 Variants, Spike Mutations and Immune Escape. *Nature Reviews Microbiology*. Nature
328 Research July 1, 2021, pp 409–424. <https://doi.org/10.1038/s41579-021-00573-0>.
- 329 (5) Carabelli, A. M.; Peacock, T. P.; Thorne, L. G.; Harvey, W. T.; Hughes, J.; de Silva, T. I.;
330 Peacock, S. J.; Barclay, W. S.; de Silva, T. I.; Towers, G. J.; Robertson, D. L. SARS-CoV-2
331 Variant Biology: Immune Escape, Transmission and Fitness. *Nature Reviews*
332 *Microbiology*. Nature Research March 1, 2023, pp 162–177.
333 <https://doi.org/10.1038/s41579-022-00841-7>.
- 334 (6) Schneider, W. L.; Roossinck, M. J. Genetic Diversity in RNA Virus Quasispecies Is
335 Controlled by Host-Virus Interactions. *J Virol* **2001**, *75* (14), 6566–6571.
336 <https://doi.org/10.1128/jvi.75.14.6566-6571.2001>.
- 337 (7) Villa, T. G.; Abril, A. G.; Sánchez, S.; de Miguel, T.; Sánchez-Pérez, A. Animal and
338 Human RNA Viruses: Genetic Variability and Ability to Overcome Vaccines. *Archives of*
339 *Microbiology*. Springer Science and Business Media Deutschland GmbH March 1,
340 2021, pp 443–464. <https://doi.org/10.1007/s00203-020-02040-5>.
- 341 (8) He, X.; Hong, W.; Pan, X.; Lu, G.; Wei, X. SARS-CoV-2 Omicron Variant: Characteristics
342 and Prevention. *MedComm*. John Wiley and Sons Inc December 1, 2021, pp 838–845.
343 <https://doi.org/10.1002/mco2.110>.
- 344 (9) Fan, Y.; Li, X.; Zhang, L.; Wan, S.; Zhang, L.; Zhou, F. SARS-CoV-2 Omicron Variant:
345 Recent Progress and Future Perspectives. *Signal Transduction and Targeted Therapy*.
346 Springer Nature December 1, 2022. <https://doi.org/10.1038/s41392-022-00997-x>.
- 347 (10) Zhao, H.; Lu, L.; Peng, Z.; Chen, L. L.; Meng, X.; Zhang, C.; Ip, J. D.; Chan, W. M.; Chu, A.
348 W. H.; Chan, K. H.; Jin, D. Y.; Chen, H.; Yuen, K. Y.; To, K. K. W. SARS-CoV-2 Omicron
349 Variant Shows Less Efficient Replication and Fusion Activity When Compared with
350 Delta Variant in TMPRSS2-Expressed Cells. *Emerg Microbes Infect* **2022**, *11* (1), 277–
351 283. <https://doi.org/10.1080/22221751.2021.2023329>.

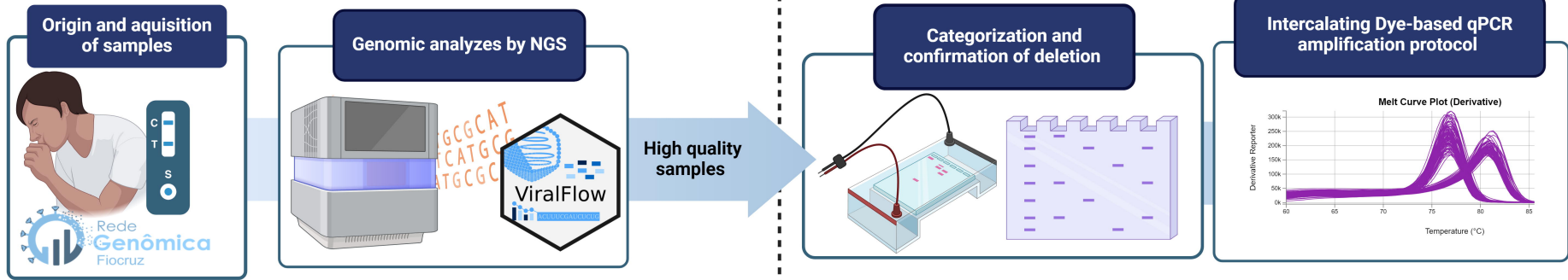
- 352 (11) Wrenn, J. O.; Pakala, S. B.; Vestal, G.; Shilts, M. H.; Brown, H. M.; Bowen, S. M.;
353 Strickland, B. A.; Williams, T.; Mallal, S. A.; Jones, I. D.; Schmitz, J. E.; Self, W. H.; Das, S.
354 R. COVID-19 Severity from Omicron and Delta SARS-CoV-2 Variants. *Influenza Other*
355 *Respir Viruses* **2022**, *16* (5), 832–836. <https://doi.org/10.1111/irv.12982>.
- 356 (12) Backer, J. A.; Eggink, D.; Andeweg, S. P.; Veldhuijzen, I. K.; van Maarseveen, N.;
357 Vermaas, K.; Vlaemyneck, B.; Schepers, R.; van den Hof, S.; Reusken, C. B. E. M.;
358 Wallinga, J. Shorter Serial Intervals in SARS-CoV-2 Cases with Omicron BA.1 Variant
359 Compared with Delta Variant, the Netherlands, 13 to 26 December 2021.
360 *Eurosurveillance* **2022**, *27* (6). [https://doi.org/10.2807/1560-](https://doi.org/10.2807/1560-7917.ES.2022.27.6.2200042)
361 [7917.ES.2022.27.6.2200042](https://doi.org/10.2807/1560-7917.ES.2022.27.6.2200042).
- 362 (13) Lyngse, F. P.; Mortensen, L. H.; Denwood, M. J.; Christiansen, L. E.; Møller, C. H.; Skov,
363 R. L.; Spiess, K.; Fomsgaard, A.; Lassaunière, R.; Rasmussen, M.; Stegger, M.; Nielsen,
364 C.; Sieber, R. N.; Cohen, A. S.; Møller, F. T.; Overvad, M.; Mølbak, K.; Krause, T. G.;
365 Kirkeby, C. T. Household Transmission of the SARS-CoV-2 Omicron Variant in Denmark.
366 *Nat Commun* **2022**, *13* (1). <https://doi.org/10.1038/s41467-022-33328-3>.
- 367 (14) Mistry, P.; Barmania, F.; Mellet, J.; Peta, K.; Strydom, A.; Viljoen, I. M.; James, W.;
368 Gordon, S.; Pepper, M. S. SARS-CoV-2 Variants, Vaccines, and Host Immunity. *Frontiers*
369 *in Immunology*. Frontiers Media S.A. January 3, 2022.
370 <https://doi.org/10.3389/fimmu.2021.809244>.
- 371 (15) Mohsin, M.; Mahmud, S. Omicron SARS-CoV-2 Variant of Concern: A Review on Its
372 Transmissibility, Immune Evasion, Reinfection, and Severity. *Medicine (United States)*.
373 Lippincott Williams and Wilkins May 13, 2022, p E29165.
374 <https://doi.org/10.1097/MD.00000000000029165>.
- 375 (16) Kim, D.; Ali, S. T.; Kim, S.; Jo, J.; Lim, J. S.; Lee, S.; Ryu, S. Estimation of Serial Interval
376 and Reproduction Number to Quantify the Transmissibility of SARS-CoV-2 Omicron
377 Variant in South Korea. *Viruses* **2022**, *14* (3). <https://doi.org/10.3390/v14030533>.
- 378 (17) Elbe, S.; Buckland-Merrett, G. Data, Disease and Diplomacy: GISAID's Innovative
379 Contribution to Global Health. *Global Challenges* **2017**, *1* (1), 33–46.
380 <https://doi.org/10.1002/gch2.1018>.
- 381 (18) Inzaule, S. C.; Tessema, S. K.; Kebede, Y.; Ogwel Ouma, A. E.; Nkengasong, J. N.
382 Genomic-Informed Pathogen Surveillance in Africa: Opportunities and Challenges. *The*
383 *Lancet Infectious Diseases*. Lancet Publishing Group September 1, 2021, pp e281–
384 e289. [https://doi.org/10.1016/S1473-3099\(20\)30939-7](https://doi.org/10.1016/S1473-3099(20)30939-7).
- 385 (19) Brito, A. F.; Semenova, E.; Dudas, G.; Hassler, G. W.; Kalinich, C. C.; Kraemer, M. U. G.;
386 Ho, J.; Tegally, H.; Githinji, G.; Agoti, C. N.; Matkin, L. E.; Whittaker, C.; Kantardjiev, T.;
387 Korsun, N.; Stoitsova, S.; Dimitrova, R.; Trifonova, I.; Dobrinov, V.; Grigorova, L.;
388 Stoykov, I.; Grigorova, I.; Gancheva, A.; Jennison, A.; Leong, L.; Speers, D.; Baird, R.;
389 Cooley, L.; Kennedy, K.; de Ligt, J.; Rawlinson, W.; van Hal, S.; Williamson, D.; Singh, R.;
390 Nathaniel-Girdharie, S. M.; Edghill, L.; Indar, L.; St. John, J.; Gonzalez-Escobar, G.;
391 Ramkisoorn, V.; Brown-Jordan, A.; Ramjag, A.; Mohammed, N.; Foster, J. E.; Potter, I.;
392 Greenaway-Duberry, S.; George, K.; Belmar-George, S.; Lee, J.; Bisasor-McKenzie, J.;
393 Astwood, N.; Sealey-Thomas, R.; Laws, H.; Singh, N.; Oyinloye, A.; McMillan, P.; Hinds,
394 A.; Nandram, N.; Parasram, R.; Khan-Mohammed, Z.; Charles, S.; Andrewin, A.;

- 395 Johnson, D.; Keizer-Beache, S.; Oura, C.; Pybus, O. G.; Faria, N. R.; Stegger, M.;
396 Albertsen, M.; Fomsgaard, A.; Rasmussen, M.; Khouri, R.; Naveca, F.; Graf, T.;
397 Miyajima, F.; Wallau, G.; Motta, F.; Khare, S.; Freitas, L.; Schiavina, C.; Bach, G.;
398 Schultz, M. B.; Chew, Y. H.; Makheja, M.; Born, P.; Calegario, G.; Romano, S.; Finello, J.;
399 Diallo, A.; Lee, R. T. C.; Xu, Y. N.; Yeo, W.; Tiruvayipati, S.; Yadahalli, S.; Wilkinson, E.;
400 Iranzadeh, A.; Giandhari, J.; Doolabh, D.; Pillay, S.; Ramphal, U.; San, J. E.; Msomi, N.;
401 Mlisana, K.; von Gottberg, A.; Walaza, S.; Ismail, A.; Mohale, T.; Engelbrecht, S.; Van
402 Zyl, G.; Preiser, W.; Sigal, A.; Hardie, D.; Marais, G.; Hsiao, M.; Korsman, S.; Davies, M.
403 A.; Tyers, L.; Mudau, I.; York, D.; Maslo, C.; Goedhals, D.; Abrahams, S.; Laguda-
404 Akingba, O.; Alisoltani-Dehkordi, A.; Godzik, A.; Wibmer, C. K.; Martin, D.; Lessells, R.
405 J.; Bhiman, J. N.; Williamson, C.; de Oliveira, T.; Chen, C.; Nadeau, S.; du Plessis, L.;
406 Beckmann, C.; Redondo, M.; Kobel, O.; Noppen, C.; Seidel, S.; de Souza, N. S.;
407 Beerenwinkel, N.; Topolsky, I.; Jablonski, P.; Fuhrmann, L.; Dreifuss, D.; Jahn, K.;
408 Ferreira, P.; Posada-Céspedes, S.; Beisel, C.; Denes, R.; Feldkamp, M.; Nissen, I.;
409 Santacrose, N.; Burcklen, E.; Aquino, C.; de Gouvea, A. C.; Moccia, M. D.; Grüter, S.;
410 Sykes, T.; Opitz, L.; White, G.; Neff, L.; Popovic, D.; Patrignani, A.; Tracy, J.; Schlapbach,
411 R.; Dermitzakis, E.; Harshman, K.; Xenarios, I.; Pegeot, H.; Cerutti, L.; Penet, D.; Stadler,
412 T.; Howden, B. P.; Sintchenko, V.; Zuckerman, N. S.; Mor, O.; Blankenship, H. M.; de
413 Oliveira, T.; Lin, R. T. P.; Siqueira, M. M.; Resende, P. C.; Vasconcelos, A. T. R.; Spilki, F.
414 R.; Aguiar, R. S.; Alexiev, I.; Ivanov, I. N.; Philipova, I.; Carrington, C. V. F.; Sahadeo, N.
415 S. D.; Branda, B.; Gurry, C.; Maurer-Stroh, S.; Naidoo, D.; von Eije, K. J.; Perkins, M. D.;
416 van Kerkhove, M.; Hill, S. C.; Sabino, E. C.; Pybus, O. G.; Dye, C.; Bhatt, S.; Flaxman, S.;
417 Suchard, M. A.; Grubaugh, N. D.; Baele, G.; Faria, N. R. Global Disparities in SARS-CoV-
418 2 Genomic Surveillance. *Nat Commun* **2022**, *13* (1). <https://doi.org/10.1038/s41467-022-33713-y>.
419
- 420 (20) Godmer, A.; Bigot, J.; Giai Gianetto, Q.; Benzerara, Y.; Veziris, N.; Aubry, A.; Guitard, J.;
421 Hennequin, C. Machine Learning to Improve the Interpretation of Intercalating Dye-
422 Based Quantitative PCR Results. *Sci Rep* **2022**, *12* (1). <https://doi.org/10.1038/s41598-022-21010-z>.
423
- 424 (21) Langer, T.; Favarato, M.; Giudici, R.; Bassi, G.; Garberi, R.; Villa, F.; Gay, H.; Zeduri, A.;
425 Bragagnolo, S.; Molteni, A.; Beretta, A.; Corradin, M.; Moreno, M.; Vismara, C.; Perno,
426 C. F.; Buscema, M.; Grossi, E.; Fumagalli, R. Development of Machine Learning Models
427 to Predict RT-PCR Results for Severe Acute Respiratory Syndrome Coronavirus 2
428 (SARS-CoV-2) in Patients with Influenza-like Symptoms Using Only Basic Clinical Data.
429 *Scand J Trauma Resusc Emerg Med* **2020**, *28* (1). <https://doi.org/10.1186/s13049-020-00808-8>.
430
- 431 (22) Nemudryi, A.; Nemudraia, A.; Wiegand, T.; Nichols, J.; Snyder, D. T.; Hedges, J. F.;
432 Cicha, C.; Lee, H.; Vanderwood, K. K.; Bimczok, D.; Jutila, M. A.; Wiedenheft, B. SARS-
433 CoV-2 Genomic Surveillance Identifies Naturally Occurring Truncation of ORF7a That
434 Limits Immune Suppression. *Cell Rep* **2021**, *35* (9).
435 <https://doi.org/10.1016/j.celrep.2021.109197>.
- 436 (23) Pyke, A. T.; Nair, N.; van den Hurk, A. F.; Burtonclay, P.; Nguyen, S.; Barcelon, J.;
437 Kistler, C.; Schlebusch, S.; McMahon, J.; Moore, F. Replication Kinetics of b.1.351 and
438 b.1.1.7 Sars-Cov-2 Variants of Concern Including Assessment of a b.1.1.7 Mutant
439 Carrying a Defective Orf7a Gene. *Viruses* **2021**, *13* (6).
440 <https://doi.org/10.3390/v13061087>.

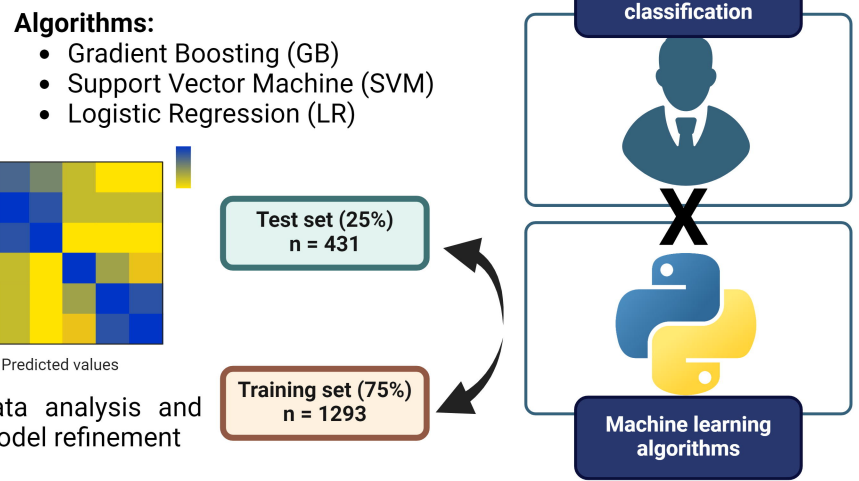
- 441 (24) Lucas, S.; Jones, M. S.; Kothari, S.; Madlambayan, A.; Ngo, C.; Chan, C.; Goraichuk, I. V.
442 A 336-Nucleotide in-Frame Deletion in ORF7a Gene of SARS-CoV-2 Identified in
443 Genomic Surveillance by next-Generation Sequencing. *Journal of Clinical Virology*.
444 Elsevier B.V. March 1, 2022. <https://doi.org/10.1016/j.jcv.2022.105105>.
- 445 (25) Foster, C. S.; Rawlinson MBBS, W. D. Rapid Spread of a SARS-CoV-2 Delta Variant with
446 a Frameshift Deletion in ORF7a. <https://doi.org/10.1101/2021.08.18.21262089>.
- 447 (26) Fuchs Wightman, F.; Godoy Herz, M. A.; Muñoz, J. C.; Stigliano, J. N.; Bragado, L.;
448 Moreno, N. N.; Palavecino, M.; Servi, L.; Cabrerizo, G.; Clemente, J.; Avaro, M.;
449 Pontoriero, A.; Benedetti, E.; Baumeister, E.; Rudolf, F.; Remes Lenicov, F.; Garcia, C.;
450 Buggiano, V.; Kornblihtt, A. R.; Srebrow, A.; de la Mata, M.; Muñoz, M. J.; Schor, I. E.;
451 Petrillo, E. A DNA Intercalating Dye-Based RT-QPCR Alternative to Diagnose SARS-CoV-
452 2. *RNA Biol* **2021**, *18* (12), 2218–2225.
453 <https://doi.org/10.1080/15476286.2021.1926648>.
- 454 (27) Watzinger, F.; Ebner, K.; Lion, T. Detection and Monitoring of Virus Infections by Real-
455 Time PCR. *Molecular Aspects of Medicine*. April 2006, pp 254–298.
456 <https://doi.org/10.1016/j.mam.2005.12.001>.
- 457 (28) Gudnason, H.; Dufva, M.; Bang, D. D.; Wolff, A. Comparison of Multiple DNA Dyes for
458 Real-Time PCR: Effects of Dye Concentration and Sequence Composition on DNA
459 Amplification and Melting Temperature. *Nucleic Acids Res* **2007**, *35* (19).
460 <https://doi.org/10.1093/nar/gkm671>.
- 461 (29) Vossen, R. H. A. M.; Aten, E.; Roos, A.; Den Dunnen, J. T. High-Resolution Melting
462 Analysis (HRMA) - More than Just Sequence Variant Screening. *Human Mutation*. June
463 2009, pp 860–866. <https://doi.org/10.1002/humu.21019>.
- 464 (30) Larios, G.; Ribeiro, M.; Arruda, C.; Oliveira, S. L.; Canassa, T.; Baker, M. J.; Marangoni,
465 B.; Ramos, C.; Cena, C. A New Strategy for Canine Visceral Leishmaniasis Diagnosis
466 Based on FTIR Spectroscopy and Machine Learning. *J Biophotonics* **2021**, *14* (11).
467 <https://doi.org/10.1002/jbio.202100141>.
- 468 (31) Pacher, G.; Franca, T.; Lacerda, M.; Alves, N. O.; Piranda, E. M.; Arruda, C.; Cena, C.
469 Diagnosis of Cutaneous Leishmaniasis Using FTIR Spectroscopy and Machine Learning:
470 An Animal Model Study. *ACS Infect Dis* **2024**, *10* (2), 467–474.
471 <https://doi.org/10.1021/acsinfecdis.3c00430>.
- 472 (32) Athamanolap, P.; Parekh, V.; Fraley, S. I.; Agarwal, V.; Shin, D. J.; Jacobs, M. A.; Wang,
473 T. H.; Yang, S. Trainable High Resolution Melt Curve Machine Learning Classifier for
474 Large-Scale Reliable Genotyping of Sequence Variants. *PLoS One* **2014**, *9* (10).
475 <https://doi.org/10.1371/journal.pone.0109094>.
- 476 (33) Nakhle, F.; Harfouche, A. L. Ready, Steady, Go AI: A Practical Tutorial on Fundamentals
477 of Artificial Intelligence and Its Applications in Phenomics Image Analysis. *Patterns*.
478 Cell Press September 10, 2021. <https://doi.org/10.1016/j.patter.2021.100323>.
- 479 (34) Engelberger, F.; Galaz-Davison, P.; Bravo, G.; Rivera, M.; Ramírez-Sarmiento, C. A.
480 Developing and Implementing Cloud-Based Tutorials That Combine Bioinformatics
481 Software, Interactive Coding, and Visualization Exercises for Distance Learning on

- 482 Structural Bioinformatics. *J Chem Educ* **2021**, *98* (5), 1801–1807.
483 <https://doi.org/10.1021/acs.jchemed.1c00022>.
- 484 (35) Carneiro, T.; Da Nobrega, R. V. M.; Nepomuceno, T.; Bian, G. Bin; De Albuquerque, V.
485 H. C.; Filho, P. P. R. Performance Analysis of Google Colaboratory as a Tool for
486 Accelerating Deep Learning Applications. *IEEE Access* **2018**, *6*, 61677–61685.
487 <https://doi.org/10.1109/ACCESS.2018.2874767>.
- 488 (36) O’Toole, Á.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J. T.; Colquhoun,
489 R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; Yeats, C.; du Plessis, L.; Maloney, D.; Medd, N.;
490 Attwood, S. W.; Aanensen, D. M.; Holmes, E. C.; Pybus, O. G.; Rambaut, A. Assignment
491 of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus*
492 *Evol* **2021**, *7* (2). <https://doi.org/10.1093/ve/veab064>.
- 493 (37) Aksamentov, I.; Roemer, C.; Hodcroft, E.; Neher, R. Nextclade: Clade Assignment,
494 Mutation Calling and Quality Control for Viral Genomes. *J Open Source Softw* **2021**, *6*
495 (67), 3773. <https://doi.org/10.21105/joss.03773>.
- 496 (38) Noble, W. S. *What Is a Support Vector Machine?*; 2006; Vol. 24.
497 <https://doi.org/https://doi.org/10.1038/nbt1206-1565>.

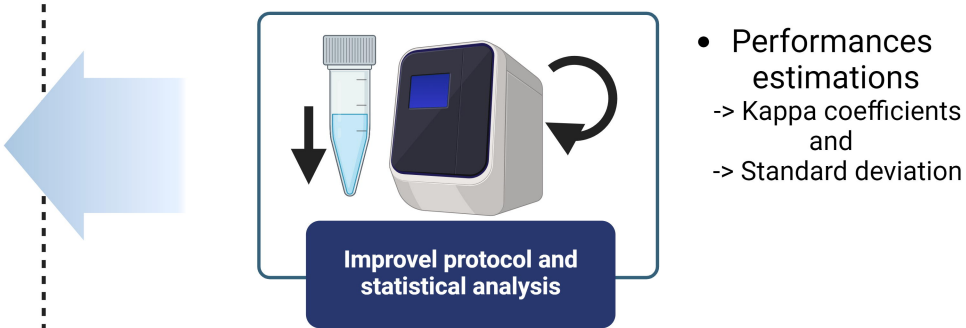
Validation step (n = 16)



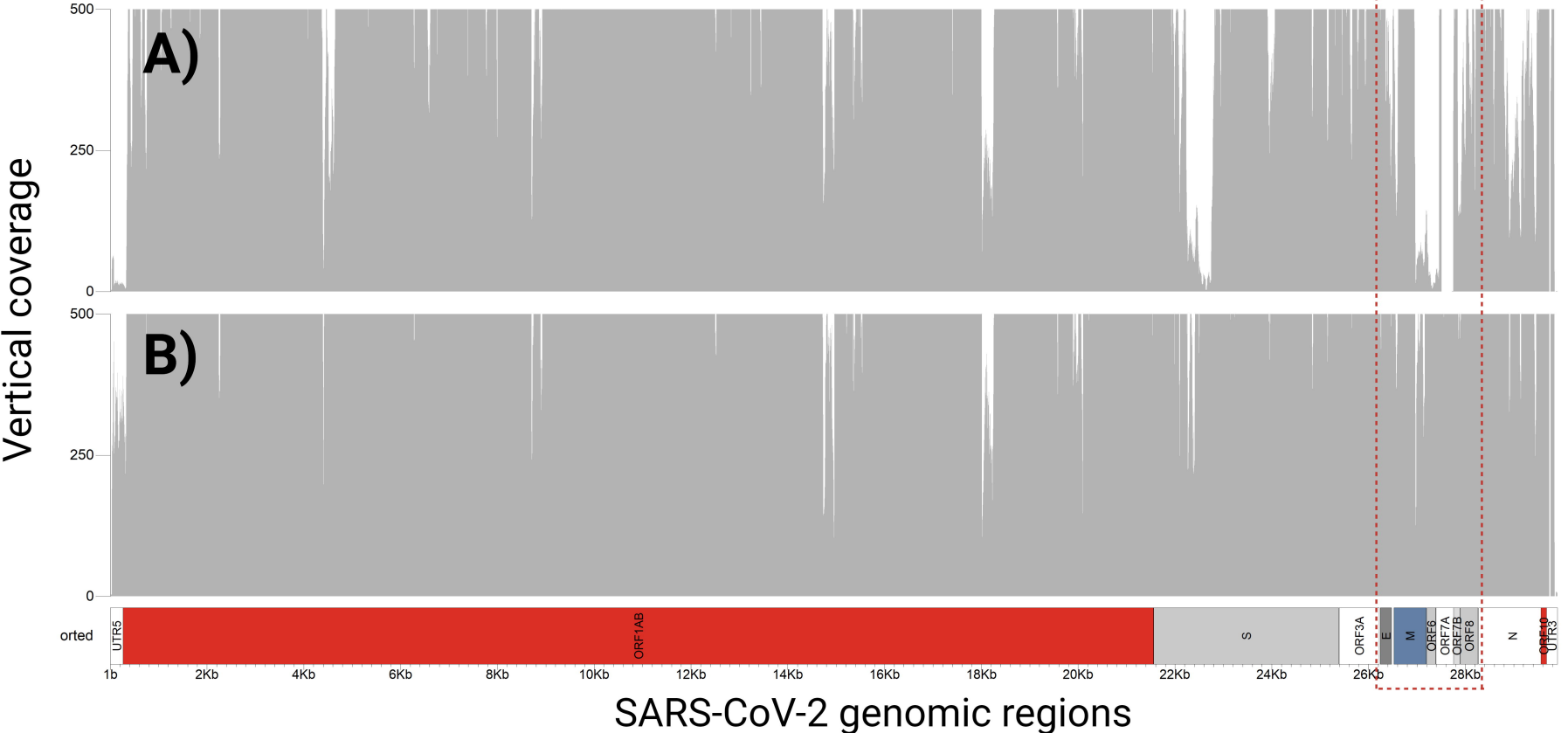
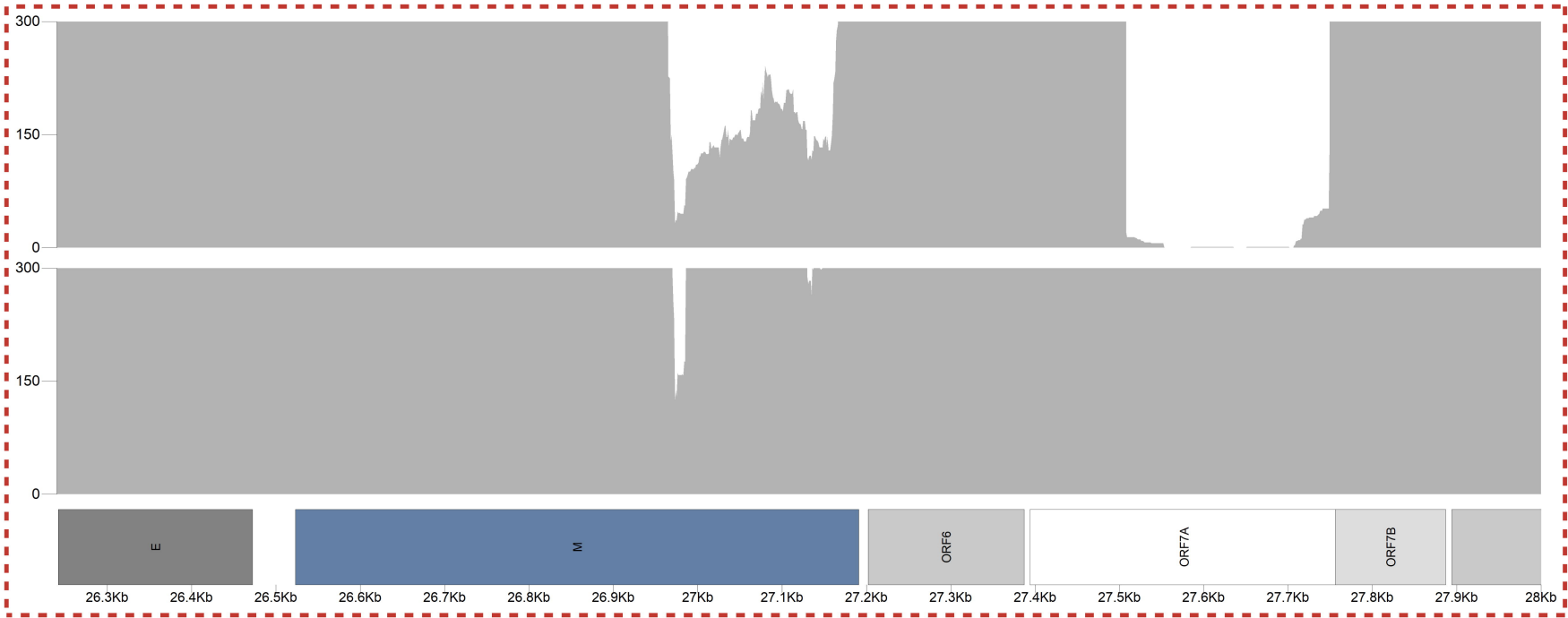
Automatic classification step (n = 1724)

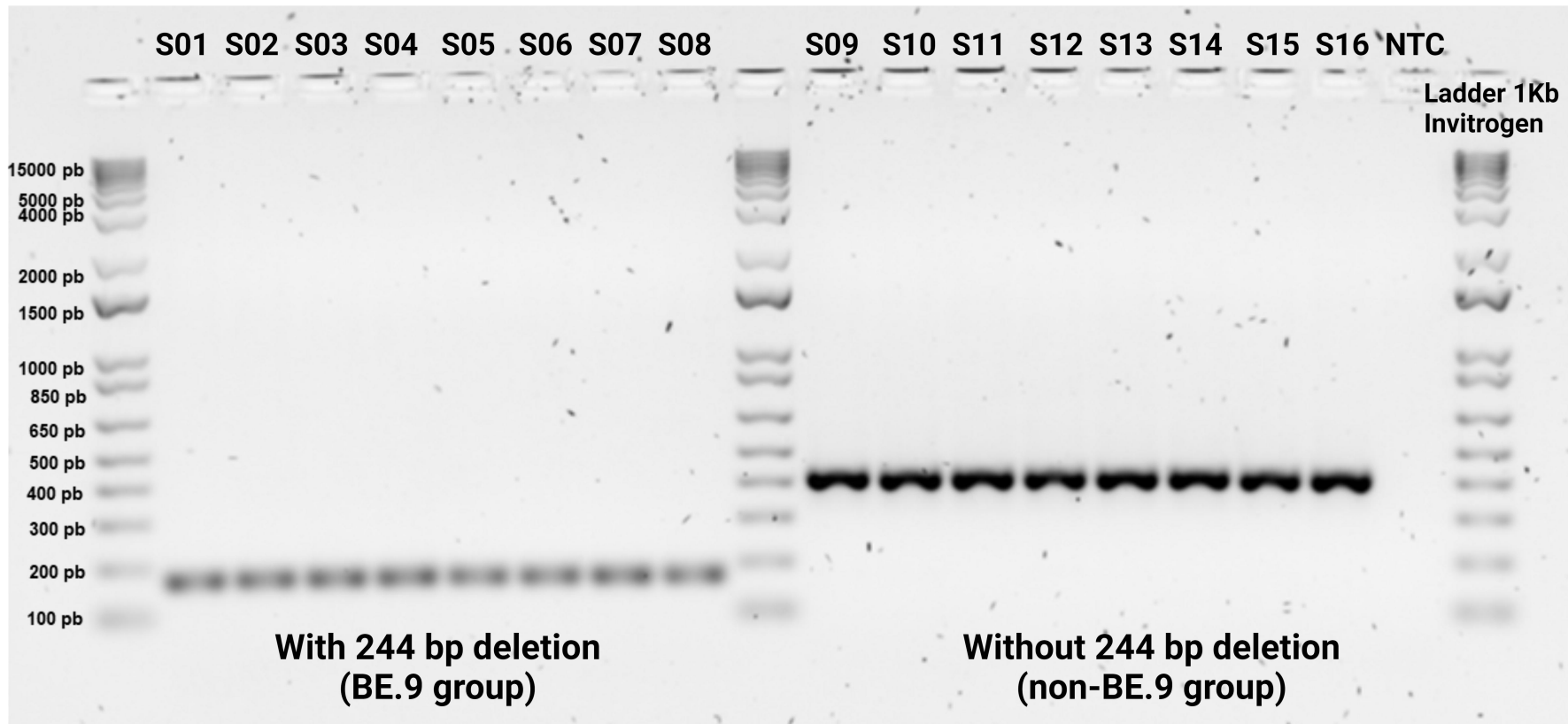


Protocol optimization step (n = 70)

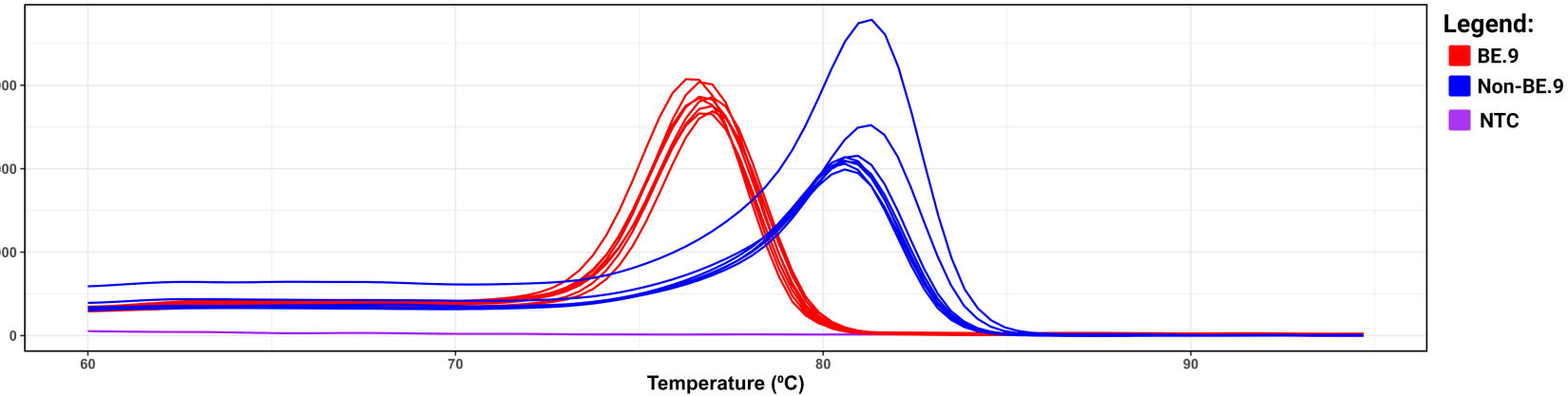


Low coverage region
(244 pb deletion)





Derivative of fluorescence



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