Machine learning models exploring characteristic single-nucleotide signatures in Yellow Fever Virus

Álvaro Salgado[a#*], Raquel C. de Melo-Minardi[b*], Marta Giovanetti[a,c*], Adriano Veloso[b*], Francielly Morais-Rodrigues[a], Talita Adelino[d], Ronaldo de Jesus[e], Stephane Tosta[a], Vasco Azevedo[a], Jose Lourenço[f#], Luiz Carlos J. Alcantara[a,c#]

[a] Laboratório de Genética Celular e Molecular, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[b] Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[c] Laboratório de Flavivírus, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

[d] Laboratório Central de Saúde Pública, Fundação Ezequiel Dias, Belo Horizonte, Minas Gerais, Brazil

[e] Coordenação Geral dos Laboratórios de Saúde Pública, Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, DF, Brazil

[f] Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

Running Head: Machine learning applied to YFV genomic investigation

*These authors contributed equally.

#Corresponding authors: Álvaro Salgado, alvarosalgado@ufmg.br; Luiz Carlos J Alcantara, luiz.alcantara@ioc.fiocruz.br; Jose Lourenço, jose.lourenco@zoo.ox.ac.uk

Abstract

Yellow fever virus (YFV) is the agent of the most severe mosquito-borne disease in the tropics. Recently, Brazil suffered major YFV outbreaks with a high fatality rate affecting areas where the virus has not been reported for decades, consisting of urban areas where a large number of unvaccinated people live. We developed a machine learning framework combining three different algorithms (XGBoost, random forest and regularized logistic regression). This method was applied to 56 YFV sequences from human infections and 27 from non-human primate (NHPs) infections to investigate the presence of genetic signatures possibly related to disease severity (in human related sequences) and differences in the PCR cycle threshold (Ct) values (in NHP related sequences). Our analyses reveal four non-synonymous single nucleotide variations (SNVs) on sequences from human infections, in proteins NS3 (E614D), NS4a (I69V), NS5 (R727G, V643A) and six non-synonymous SNVs on NHP sequences, in proteins E (L385F), NS1 (A171V), NS3 (I184V) and NS5 (N11S, I374V, E641D). We performed comparative protein structural analysis on these SNVs, describing possible impacts on protein function. Despite the fact that the dataset is limited in size and that this study does not consider virus-host interactions, our work highlights the use of machine learning as a versatile and fast initial approach to genomic data exploration.

Importance

Yellow fever is responsible for 29-60 thousand deaths annually in South America and Africa and is the most severe mosquito-borne disease in the tropics. Given the range of clinical outcomes and the availability of YFV genomic data, the use of machine learning analysis promises to be a powerful tool in the investigation of genetic signatures that could impact disease severity and its potential of being reintroduced in an urban transmission cycle. This can assist in the search for biomarkers of severity as well as help elucidating variations in host's Ct value. This work aims to propose a relatively fast and inexpensive computational analysis framework, which can be used as a real-time, innitial strategy associated with genomic surveillance to identify a set of single nucleotide variants putatively related to biological and clinical characteristics being observed.

Introduction

Yellow fever (YF) is an acute viral hemorrhagic disease endemic in tropical areas of Africa and Latin America. The causative agent, yellow fever virus (YFV), represents the prototypical member of the genus *Flavivirus* (family *Flaviviridae*), consisting of a single-stranded, positive-sense RNA virus, with a genome about 11,000 kb and a single open-reading frame of 10,233 nucleotides (1, 2). Disease varies from nonspecific febrile illness to a fatal hemorrhagic fever. Symptoms usually appear after an incubation period of three to six days following the bite of an infected mosquito, with a period of infection lasting several days (2–4). The World Health Organization (WHO) reports case fatality rates in the order of 15 to 50% (5). Vaccination remains the most effective YF prevention method, providing lifetime immunity in up to 99% of vaccinated people (6). Nevertheless, the burden of YF is estimated to be between 84,000 to 170,000 severe cases and 29,000 to 60,000 deaths annually (7, 8), while an estimated 35 million people remain unvaccinated in areas at risk in Brazil only (9).

YFV spreads in two different cycles: sylvatic and urban. The sylvatic transmission cycle occurs in forested areas, where the virus is endemically transmitted between several non-human primate (NHP) species. The urban transmission cycle occurs when the virus is introduced into human populations with high density and urban-dwelling mosquitoes (mainly *Aedes aegypti)* (3). Urban cycles of YFV transmission have been eradicated in Brazil since 1942 due to vaccination and vector control campaigns (10–13).

In the last decade in Brazil, however, human and NHP epizootic YF cases have been notified at places beyond the limits of regions previously considered (sylvatic) endemic for the virus (14–17). The severe impact of these recent outbreaks can be measured, in part, by its fatality rate at around 34%, higher than the general rate estimated by Monath and colleagues (4), motivating the inquiry as to what could be the possible factors contributing to such a high fatality rate, and if YFV genetic signatures could be among those factors.

Additionally, important findings in recent epidemics (11, 18) show a significant difference in the distribution of NHP Ct values, in which *Callithrix* spp. exhibit generally higher Ct values than other NHP species, do not develop fatal YFV infections similar to those reported in humans and can persist for longer, thus increasing the infectious period. The latter can be an essential factor in igniting an urban cycle of transmission, mainly due to the genus' proximity to densely urbanized areas.

On this respect, genomic and epidemiological monitoring have become an integral part of the national (Brazil) and international response to emerging and ongoing epidemics of viral infectious diseases, allowing the availability of a large amount of genomic data (19–23).

In genomic and epidemiological monitoring analysis, machine learning (ML) approaches are usually applied (24, 25), as described in works that analyze the effectiveness of large scale genome-wide association studies (GWAS), due to their capability to computationally model the relationship between combinations of single nucleotide variants, other genetic variations and environmental factors with observed outcomes (26–28).

We curated two different datasets of YFV genomes, one from human cases and the other from NHP cases. After data curation, the human dataset contained 56 YFV sequences, with 40 sequences related to infections leading to severe outcomes or death, and 16 sequences related to cases with no severe outcome. We also gathered an NHP (*Callithrix* spp.) dataset, that after curation contained 27 sequences, of which 21 were related to low Ct values (< 20) and 6 were related to high Ct values (>= 20).

We applied three different ML models to each dataset, to guarantee robustness of the ML analysis (29). We then analysed the models using SHAP (SHapley Additive explanation) (30–32) to highlihght genetic signatures. The possible biological impacts of these signatures were investigated and discussed by means of in-silico protein structural analysis coupled with literature review.

Results

Non-human primates Ct value statistical analysis

Figure 1 shows the distribution of cycle threshold (Ct) values from *Callithrix* spp. sequences, with two distinct clusters roughly around 12 and 30, with a median value of 26.1. The result of Hartigan's dip test of unimodality (33) rejected the null hypothesis of a unimodal distribution for Ct values ($p < .001$), which indicate the existence of two groups of *Callithrix* spp. Ct values.

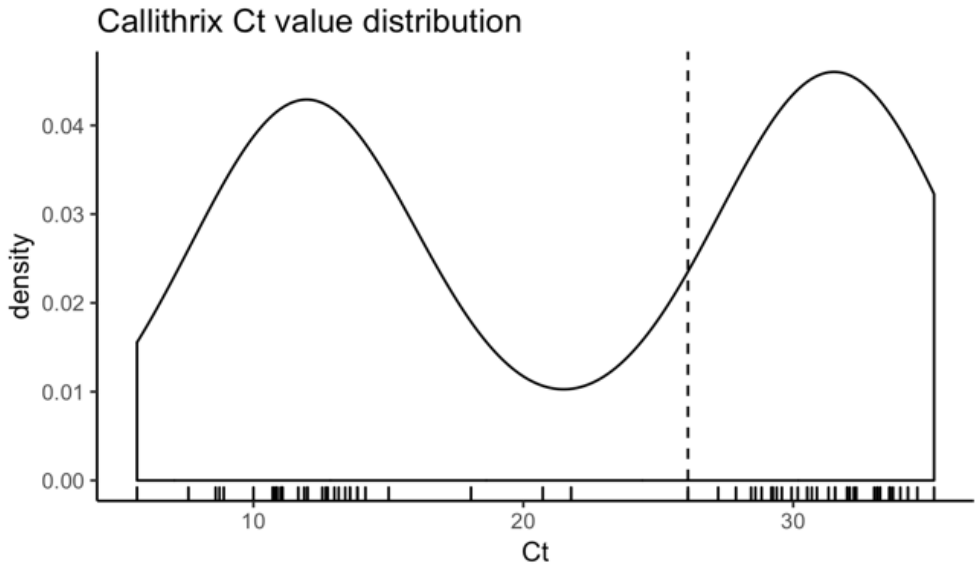

Figure 1 – Ct values associated with *Callithrix spp.* sequences. The median value shown by a dashed line. Hartigan's dip test of unimodality indicates bimodal distribution ($p < .001$).

Machine learning models' performance

Figure 2 shows the confusion matrices for the machine learning models applied. For the human dataset, the XGBoost classification model achieved an accuracy of 75% on the test set, with F-1 scores of 0.59 and 0.82 for classes 0 (not severe/not death) and 1 (severe/death), respectively. The random forest model achieved an accuracy of 82% on the test set, with F-1 scores of 0.74 and 0.86 for classes 0 and 1 respectively. The modified logistic regression model achieved an accuracy of 82% on the test set, with F-1 scores of 0.74 and 0.86 for classes 0 and 1, respectively.

For the *Callithrix* spp. dataset, the three models achieved an accuracy of 100% on the test set, with F-1 scores of 1.00 and 1.00 for classes 0 (low Ct) and 1 (high Ct) respectively.

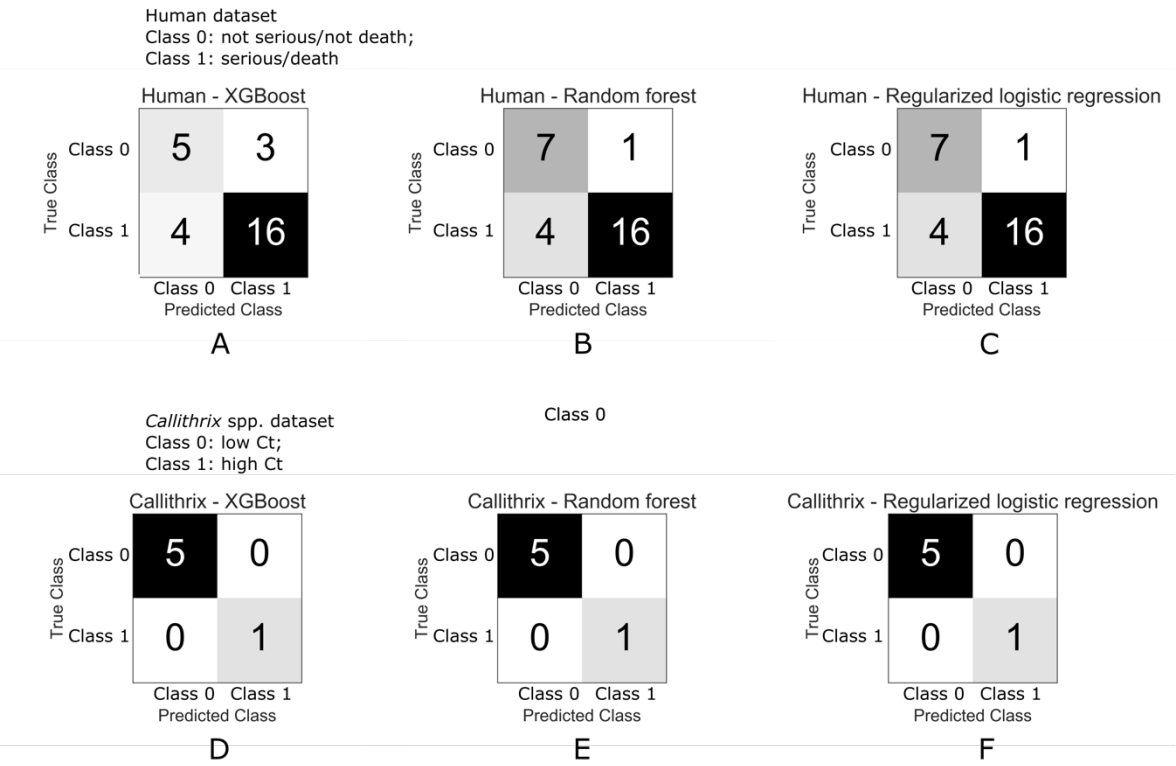

Figure 2 - Confusion matrices. Top three figures correspond to human test dataset and the bottom three figures correspond to *Callithrix spp.* test dataset, for XGBoost, random forest and regularized logistic regression respectively.

YFV genetic signatures

The machine learning methods identified the non-synonymous single nucleotide variants (SNVs) shown in

131 **Table *1***. The table displays each protein where the SNV was found, with nucleotide position on
132 YFV genome, position relative to the protein's sequence, amino acid position relative to the
133 translated protein, reference genome amino acid and corresponding codon, analyzed sequences
134 amino acid variation and corresponding codon and SNV position inside codon (1^{st}, 2^{nd} or 3^{rd}).

135 The results obtained by the analysis of human YFV sequences highlighted 4 SNV positions that
136 result in the amino acid change, and the results obtained by the analysis of *Callithrix* spp. shows 6
137 SNV positions that resulted in the amino acid change.

138
139

140 **Table 1** - YFV genetic signatures

Human dataset

Protein	nn postition	nn position on protein	aa position on protein	aa reference	codon reference	aa variation	codon variation	SNV codon position (1, 2, 3)
NS3	6412	1842	614	E	gaa	D D	gac gat	3
NS4a	6644	205	69	I	atc	V	gtc	1
NS5	9815	2179	727	R	agg	G	ggg	1
NS5	9564	1928	643	V	gtt	A	gct	2

NHP dataset

Protein	nn postition	nn position on protein	aa position on protein	aa reference	codon reference	aa variation	codon variation	SNV codon position (1, 2, 3)
NS5	8756	1120	374	I	atc	V	gtc	1
NS5	9559	1923	641	E	gaa	D D	gac gat	3
NS3	5120	550	184	I	atc	V	gtc	1
E	2126	1153	385	L	ctc	F	ttc	1
NS5	7647	11	4	N	aat	S	agt	2
NS1	2964	512	171	A	gca	V	gta	2

141

142

143 **Protein structural analysis**

144 We performed protein structural analysis for all SNVs indicated in

145 Table ***1***. The templates, their resolution, the quality of models provided from Swiss-Model (34)
146 and the changes in binding affinity and stability predicted by mCSM-NA (35) are summarized in
147

148 **Table *2***.

149

150 **Table 2** - Protein structural analysis results

	Callithrix spp. Dataset						**Human dataset**			
Protein	E	NS1	NS3	NS5			NS3	NS4a	NS5	
SNV	L385F	A171V	I184V	N11S	I374V	E641D	E614D	I69V	R727G	V643A
Template (PDB ID)	Template (6WI5) (?)	Zika virus NS1 (5K6K) (27)	Dengue virus NS3 (5YV8:A) (31)	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS3 (1YKS) (29)	No structure deposited in the PDB	Yellow fever virus NS5 (6QSN) (32)	Yellow fever virus NS5(6QSN) (32)
Resolution	1.83 Å	1.89 Å	2.5 Å	3.00 Å	3.00 Å	3.00 Å	1.80 Å	-	3.00 Å	3.00 Å
Coverage (%)		100%	99%					-		
Sequence identity		47.58%	50.57%					-		
Localization at protein	Domain III	Wing flexible loop	Linker region	MTase domain	Palm subdomain	Palm subdomain	Helicase domain	-	Thumb subdomain	Palm subdomain
Global Model Quality Estimation (GMQE)		0.79	0.76					-		
(Qualitity mean) QMEAN		-2.5	-2.22					-		
Predicted change in binding affinity	Target distant from binding sites				ΔΔG=0.003 Kcal/mol (Increased affinity)		ΔΔG=0.025 Kcal/mol (Increased affinity)	-	ΔΔG=-1.545 Kcal/mol (Reduced affinity)	ΔΔG=-0.001 Kcal/mol (Reduced affinity)
Predicted change in stability					-1.908 Kcal/mol (Destabilising)		-0.254 Kcal/mol (Destabilising)	-	-0.751 Kcal/mol (Destabilising)	-1.884 Kcal/mol (Destabilising)

151

Figure *3* shows the structural representation of proteins E, NS1, NS3 and NS5, with corresponding SNVs.

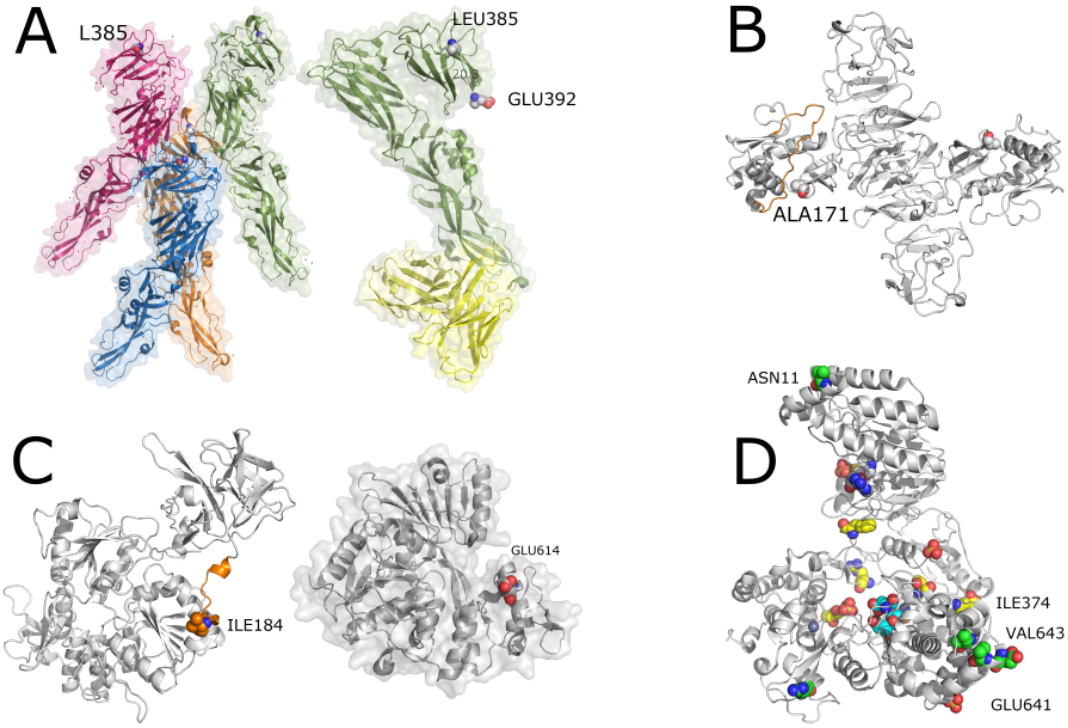

Figure 3 - **(A)** Envelope protein - **(A, left)** PDB id 6wi5 - tretamer of E protein LEU385 presented in spheres. **(A, right)** PDB id 6ivz - in green, one chain of protein E and in yellow, the light and heavy chains of monoclonal antibody 5A. Note that the L385F SNV is distant from both the binding between protein E chains and the antibody recognition site. **(B)** NS1 protein. Comparative model built with Swiss-Model and PDB id 5k6k. In orange, we depict the wing flexible loop and in spheres, SNV ALA171. **(C)** NS3 protein. **(C, left)** Comparative model built with Swiss-Model and PDB id 5yvu. In orange, we depict the interdomain linker region and in spheres, SNV ILE184. **(C, right)** GLU614 from PDB id 1yks showing that the SNV E614D is close to the cleft where the DNA binds the NS3 protein. **(D)** NS5 protein. In green, SNVs. In yellow, important conserved residues across ZIKV, DENV and WNV. In cyan, active site. In grey, ligand S-adenosyl-L-homocysteine. Sulfate and Zn íons are also represented in spheres.

E (envelope) protein

LEU385 (NHP dataset) is far from the intra-chain binding site and the antibody recognition site (Figure 4-A). LEU385 is in the Domain III (DIII) which has an immunoglobulin C domain (IgC-like) presenting a seven-stranded fold and is supposed to contain the receptor-binding site. DIII suffers a rotation and goes closer to the fusion loop (FL), bringing the C-terminal part of DIII (residue 392) close to FL. LEU385 is 20.3 angstroms far from GLU392.

NS1 protein

The SNV A171V (NHP dataset) is in close contact with a region called *wing flexible loop* (highlighted in orange in Figure 4-B) and it is also a probable glycosylation site (36).

NS3 protein

Although there is a crystallographic structure (PDB id: 1yks) of the helicase domain (37) and another containing part of NS3 complexed with NS2B (PDB id 6urv) (38) deposited in the PDB, the referred SNV occurs in the unresolved stretch. The I184V (NHP dataset) is in the linker region (shown in orange in Figure *4*-C, left). It connects protease and helicase domains and corresponds to sequence KEEGKEELQEIP that encompasses residues between 174 and 185. The SNV E614D (human dataset) occurs in the helicase domain and is located in the RNA binding cleft, shown in red on Figure *4*-C, right.

NS4a protein

Since it has multiple transmembrane hydrophobic segments, structural analysis of NS4a has been unsuccessful and, so far, there is no structure deposited in the PDB. It is still one of the least characterized proteins from YFV. It was not possible to obtain a good structural model since the best template found had a coverage of only 37% and a sequence identity of 25.53%.

NS5 protein

There is a recent YFV NS5 structure deposited on PDB (PDB id 6qsn) (39). The analyzed SNVs (shown in green in Figure *4*-D) are N11S, the only SNV in the MTase domain; I374V and E641D, both located in the palm subdomain. These three SNVs were found on the NHP dataset. Additionally, SNV V643A, located in the palm, and R727G, in the thumb subdomain, were found in the human dataset. As depicted in Figure *4*-D, they (green) are located far from the Zn and sulfate ions and the ligand (S-adenosyl-L-homocysteine) (grey). They are not close to any important / conserved mentioned residue (yellow) that interact with the nucleic acid. SNV I374V is also present in ZIKV, DENV and WNV. Position E641D varies across other viruses (K, N, R). Position V643A is also not conserved being an insertion, K or N. Position R727G is S, E or T in other flaviviruses.

Discussion

In this study, we demonstrate the potential of applying ML approaches on real-time genomic surveillance, to quickly identify genetic loci which may be of public health interest.

We find signals in multiple genetic loci and present a structural-based review on the potential impact of changes at those loci. However, the limited number of sequences analyzed demands caution when presenting the results. A large number of high-quality sequences is ideal for the application of ML analysis, especially when dealing with viruses, whose high mutation rates tend to insert many variations on its genomes. Furthermore, our analysis didn't consider virus-host interactions, such as host genome or immune system and pre-existing health conditions. In this regard, efficient host data collection, such as Electronic Health Records (EHR), are of paramount importance for a thorough investigation of clinical outcomes.

The envelope (E) protein is related to virus attachment and fusion (40). NHP dataset analysis shows SNV L395F on Domain III, a region containing an IgC-like domain and supposed to contain a receptor-binding site crucial for virion maturation (41). It is possible that this SNV could have an impact on the plasticity of this domain and affect the virus' receptor-binding site and it would be interesting to investigate this behaviour through simulations of molecular dynamics in future work.

NS1 protein is a crucial non-structural protein (36). We found a non-synonymous SNV (A171V) on NHP dataset, located near a highly flexible region on the protein, called the *wing flexible loop*, which is a probable glycosylation site (GS) (42). NS1 is also a key protein secreted by infected cells, which has the potential to interact with the adaptive immune system responses (36). In dengue infections, NS1 is known to modulate capillary leakage in severe disease and may thus have a role to play in the severity of YFV infection (43).

NS3 protein, which is composed of protease and helicase domains, has functions related to viral polyprotein processing and cleavage, viral genome replication and RNA capping (40). SNV I184V, found on NHP dataset, is in a region with probable limited functional constraints (44). SNV E614D, found on the human dataset, occurs in the NS3 helicase domain and is located in the RNA binding cleft. ASP and GLU are both negatively charged amino acids, but ASP has a shorter side chain which can cause it to lose access to the ligand. We used mCSM-NA (35) to evaluate the impact of the SNV on stability and affinity with RNA, showing a small destabilizing effect on the interaction with RNA (-0.646 Kcal/mol), which could have an impact on its function, fundamental to viral genome replication. Unfortunately, there are no structures in complex with RNA available. Models built with protein-RNA docking techniques could help elucidate if there is a significant impact of this SNV on RNA interaction.

We found one SNV on the NS4a protein (I69V, human dataset). However, a lack of current knowledge on YFV NS4a impeded us from further exploring the possible role of I69V in human

hosts. Based on the protein's proposed functions (45–47), this SNV could, in principle, affect viral replication, but such hypothesis would have to be tested by non-computational means.

As an outlier, NS5 had the highest number of identified variations - I374V, E641D, N11S, R727G, V643A - from both human and NHP YFV sequences. NS5 protein is a fundamental enzyme for viral replication because it contains an N-terminal methyltransferase domain (MTase) and a C-terminal RNA dependent RNA polymerase domain (RdRp) (48). MTase domain has important functions involved in protecting viral RNA from degradation and innate immunity response. RdRp is essential for viral RNA replication, because its activities cannot be performed by host polymerases, and is a promising target for antiviral drug development (49). Furthermore, dengue virus NS5 has been associated with immune response evasion (50). With the structural analyses, we found that none of the SNVs found has been reported to be in positions with apparent connections to protein function or structure. However, R727G on the human dataset, on the thumb subdomain of RdRp domain, shows a change in predicted affinity for RNA upon SNV occurrence. This reduction in affinity could impact polymerase function and viral replication efficiency.

In conclusion, even though the method proposed was applied on data that was already available from other sources, our study demonstrate that it is efficient and easy to replicate, making it suitable for real-time genomic surveillance, in which genetic data is analized as it is generated. This approach may help detect and inform on possible connections between ongoing genetic changes and public health in a timely manner.

Materials and Methods

Datasets

We retrieved YFV complete or near complete genome sequences from the recent Brazilian outbreaks, available on public databases (19–23), with associated epidemiological and clinical data containing relevant information regarding clinical severity and outcome for human infections samples, as well as PCR cycle threshold value (Ct) for both human and NHP samples. The alignment was made using MAFFT online (51) and was manually verified and corrected using AliView (https://ormbunkar.se/aliview/). Sequences with coverage lower than 90% were removed from the study. For ML analysis, human infection sequences were divided between not severe/not death and severe/death. *Callithrix* ssp. infection sequences were divided by low Ct (<20) and high Ct (≥ 20). After curation, the human dataset contained 56 YFV sequences, with 40 sequences related to severe/death cases, and 16 sequences related to not severe/not death cases. The NHP

(*Callithrix* spp.) dataset, after curation, contained 27 sequences, of which 21 were related to low Ct values (< 20) and 6 were related to high Ct values (>= 20).

Machine learning model adjustment

We applied three different ML models for each of two analyzed datasets, XGBoost (52, 53), random forest (54) and regularized logistic regression (55). We adjusted the XGBoost model parameters in a "grid-search cross-validation" scheme with five folds. Random forest adjustment used "out of bag" data as validation. Regularized logistic regression parameters were adjusted on a 10-fold cross-validation scheme, using their averages in the final model.

Model interpretation

Feature importance was computed using SHAP (SHapley Additive exPlanation) (30–32). We followed the author's suggestion (https://github.com/slundberg/shap/issues/397) on dealing with categorical data when using SHAP.

Protein structural analysis

We searched on Protein Data Bank (PDB) (56) for experimentally resolved structures. For those proteins that did not have structures, we looked for templates for comparative modeling with at least 30% identity. The comparative models were built with the Swiss-Model server (57). We used mCSM (35) method to predict the impact of SNVs on protein stability and interactions.

Acknowledgements

This work was supported by Decit, SCTIE, Brazilian Ministry of Health, Conselho Nacional de Desenvolvimento Científico - CNPq - (440685/2016-8 and 440856/2016-7), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES (88887.130716/2016-00, 88881.130825/2016-00 and 88887.130823/2016-00), by the European Union's Horizon 2020 Research and Innovation Programme under ZIKAlliance Grant Agreement No. 734548.

References

1. Chambers TJ, Hahn CS, Galler R, Rice CM. 1990. Flavivirus genome organization, expression, and replication. Annu Rev Microbiol 44:649–688.

2. Monath TP. 2001. Yellow fever: an update. Lancet Infect Dis 1:11–20.

3. Gardner CL, Ryman KD. 2010. Yellow Fever: A Reemerging Threat. Clinics in Laboratory Medicine 30:237–260.

4. Monath TP. 2008. Treatment of yellow fever. Antiviral Research 78:116–124.

5. 2015. WHO Report on Global Surveillance of Epidemic-prone Infectious Diseases - Yellow fever. World Health Organization. World Health Organization.

6. 2020. Yellow Fever. Pan American Health Organization / World Health Organization.

7. Paules CI, Fauci AS. 2017. Yellow Fever — Once Again on the Radar Screen in the Americas. New England Journal of Medicine 376:1397–1399.

8. 2019. Yellow fever. World Health Organization.

9. Shearer FM, Moyes CL, Pigott DM, Brady OJ, Marinho F, Deshpande A, Longbottom J, Browne AJ, Kraemer MU, O'Reilly KM, others. 2017. Global yellow fever vaccination coverage from 1970 to 2016: an adjusted retrospective analysis. The Lancet infectious diseases 17:1209–1217.

10. Consoli RA, de Oliveira RL. 1994. Principais mosquitos de importância sanitária no Brasil. SciELO-Editora FIOCRUZ.

11. Mares-Guia MAM de M, Horta MA, Romano A, Rodrigues CDS, Mendonça MCL, dos Santos CC, Torres MC, Araujo ESM, Fabri A, de Souza ER, Ribeiro ROR, Lucena FP, Junior

LCA, da Cunha RV, Nogueira RMR, Sequeira PC, de Filippis AMB. 2020. Yellow fever epizootics in non-human primates, Southeast and Northeast Brazil (2017 and 2018). Parasites & Vectors 13.

12. 2017. Yellow fever, the return of an old threat. Fiocruz.

13. Vasconcelos PF da C. 2010. Yellow fever in Brazil: thoughts and hypotheses on the emergence in previously free areas. Revista de Saúde Pública 44:1144–1149.

14. Delatorre E, Abreu FVS de, Ribeiro IP, Gómez MM, dos Santos AAC, Ferreira-de-Brito A, Neves MSAS, Bonelly I, de Miranda RM, Furtado ND, Raphael LMS, da Silva L de FF, de Castro MG, Ramos DG, Romano APM, Kallás EG, Vicente ACP, Bello G, Lourenço-de-Oliveira R, Bonaldo MC. 2019. Distinct YFV Lineages Co-circulated in the Central-Western and Southeastern Brazilian Regions From 2015 to 2018. Front Microbiol 10.

15. DIVE - Boletim Epidemiológico n° 06/2020 Situação epidemiológica da Febre Amarela em Santa Catarina (Atualizado em 10/06/2020).

16. Boletim epidemiológico da Febre Amarela no Brasil 2019/2020 | RETS - Rede Internacional de Educação de Técnicos em Saúde.

17. 2020. Febre Amarela | Secretaria da Saúde.

18. Cunha MS, da Costa AC, de Azevedo Fernandes NCC, Guerra JM, dos Santos FCP, Nogueira JS, D'Agostino LG, Komninakis SV, Witkin SS, Ressio RA, Maeda AY, Vasami FGS, Kaigawa UMA, de Azevedo LS, de Souza Facioli PA, Macedo FLL, Sabino EC, Leal É, de Souza RP. 2019. Epizootics due to Yellow Fever Virus in São Paulo State, Brazil: viral dissemination to new areas (2016–2017). Scientific Reports 9:5474.

19. Cunha M dos P, Duarte-Neto AN, Pour SZ, Ortiz-Baez AS, Černý J, Pereira BB de S, Braconi CT, Ho Y-L, Perondi B, Sztajnbok J, Alves VAF, Dolhnikoff M, Holmes EC, Saldiva PHN, Zanotto PM de A. 2019. Origin of the São Paulo Yellow Fever epidemic of 2017–2018 revealed through molecular epidemiological analysis of fatal cases. Scientific Reports 9.

20. Faria NR, Kraemer MUG, Hill SC, Jesus JG de, Aguiar RS, Iani FCM, Xavier J, Quick J, Plessis L du, Dellicour S, Thézé J, Carvalho RDO, Baele G, Wu C-H, Silveira PP, Arruda MB, Pereira MA, Pereira GC, Lourenço J, Obolski U, Abade L, Vasylyeva TI, Giovanetti M, Yi D, Weiss DJ, Wint GRW, Shearer FM, Funk S, Nikolay B, Fonseca V, Adelino TER, Oliveira M a. A, Silva MVF, Sacchetto L, Figueiredo PO, Rezende IM, Mello EM, Said RFC, Santos DA, Ferraz ML, Brito MG, Santana LF, Menezes MT, Brindeiro RM, Tanuri A, Santos FCP dos, Cunha MS, Nogueira JS, Rocco IM, Costa AC da, Komninakis SCV, Azevedo V, Chieppe AO, Araujo ESM, Mendonça MCL, Santos CC dos, Santos CD dos, Mares-Guia AM, Nogueira RMR, Sequeira PC, Abreu RG, Garcia MHO, Abreu AL, Okumoto O, Kroon EG, Albuquerque CFC de, Lewandowski K, Pullan ST, Carroll M, Oliveira T de, Sabino EC, Souza RP, Suchard MA, Lemey P, Trindade GS, Drumond BP, Filippis AMB, Loman NJ, Cauchemez S, Alcantara LCJ, Pybus OG. 2018. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science 361:894–899.

21. Giovanetti M, de Mendonça MCL, Fonseca V, Mares-Guia MA, Fabri A, Xavier J, de Jesus JG, Gräf T, dos Santos Rodrigues CD, dos Santos CC, Sampaio SA, Chalhoub FLL, de Bruycker Nogueira F, Theze J, Romano APM, Ramos DG, de Abreu AL, Oliveira WK, do Carmo Said RF, de Alburque CFC, de Oliveira T, Fernandes CA, Aguiar SF, Chieppe A, Sequeira PC, Faria NR, Cunha RV, Alcantara LCJ, de Filippis AMB. 2019. Yellow Fever Virus Reemergence and Spread in Southeast Brazil, 2016–2019. Journal of Virology 94.

22. Goes de Jesus J, Gräf T, Giovanetti M, Mares-Guia MA, Xavier J, Lima Maia M, Fonseca V, Fabri A, dos Santos RF, Mota Pereira F, Ferraz Oliveira Santos L, Reboredo de Oliveira da Silva L, Pereira Gusmão Maia Z, Gomes Cerqueira JX, Thèze J, Abade L, Cordeiro M de CS, Torquato SSC, Santana EB, de Jesus Silva NS, Dourado RSO, Alves AB, do Socorro Guedes A, da Silva Filho PM, Rodrigues Faria N, de Albuquerque CFC, de Abreu AL, Martins Romano AP, Croda J, do Carmo Said RF, Cunha GM, da Fonseca Cerqueira JM, Mello AL e S de, de Filippis AMB, Alcantara LCJ. 2020. Yellow fever transmission in non-human primates, Bahia, Northeastern Brazil. PLOS Neglected Tropical Diseases 14:e0008405.

23. Hill SC, de Souza R, Thézé J, Claro I, Aguiar RS, Abade L, Santos FCP, Cunha MS, Nogueira JS, Salles FCS, Rocco IM, Maeda AY, Vasami FGS, du Plessis L, Silveira PP, de Jesus JG, Quick J, Fernandes NCCA, Guerra JM, Réssio RA, Giovanetti M, Alcantara LCJ, Cirqueira CS, Díaz-Delgado J, Macedo FLL, Timenetsky M do CST, de Paula R, Spinola R, Telles de Deus J, Mucci LF, Tubaki RM, de Menezes RMT, Ramos PL, de Abreu AL, Cruz LN, Loman N, Dellicour S, Pybus OG, Sabino EC, Faria NR. 2020. Genomic Surveillance of Yellow Fever Virus Epizootic in São Paulo, Brazil, 2016 – 2018. PLOS Pathogens 16:e1008699.

24. Chen X, Ishwaran H. 2012. Random forests for genomic data analysis. Genomics 99:323–329.

25. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. 2010. High-Dimensional Variable Selection for Survival Data. Journal of the American Statistical Association 105:205–217.

26. Behravan H, Hartikainen JM, Tengström M, Pylkäs K, Winqvist R, Kosma V, Mannermaa A. 2018. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. Scientific Reports 8.

27. Ho DSW, Schierding W, Wake M, Saffery R, O'Sullivan J. 2019. Machine Learning SNP Based Prediction for Precision Medicine. Frontiers in Genetics 10.

28. Moore JH, Asselbergs FW, Williams SM. 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26:445–455.

29. Wolpert DH. 1996. The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation 8:1341–1390.

30. Lundberg S, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:170507874 [cs, stat].

31. Lundberg SM, Lee S-I. 2018. Consistent feature attribution for tree ensembles. arXiv:170606060 [cs, stat].

32. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, Lee S-I. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomedical Engineering 2:749–760.

33. Hartigan JA, Hartigan PM. 1985. The Dip Test of Unimodality. Ann Statist 13:70–84.

34. Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195–201.

35. Pires DEV, Ascher DB. 2017. mCSM-NA: predicting the effects of mutations on protein-nucleic acids interactions. Nucleic Acids Res 45:W241–W246.

36. Brown WC, Akey DL, Konwerski J, Tarrasch JT, Skiniotis G, Kuhn RJ, Smith JL. 2016. Extended Surface for Membrane Association in Zika Virus NS1 Structure. Nat Struct Mol Biol 23:865–867.

37. Wu J, Bera AK, Kuhn RJ, Smith JL. 2005. Structure of the Flavivirus Helicase: Implications for Catalytic Activity, Protein Interactions, and Proteolytic Processing. J Virol 79:10268–10277.

38. Noske GD, Gawriljuk VO, Fernandes RS, Furtado ND, Bonaldo MC, Oliva G, Godoy AS. 2020. Structural characterization and polymorphism analysis of the NS2B-NS3 protease from the 2017 Brazilian circulating strain of Yellow Fever virus. Biochim Biophys Acta Gen Subj 1864:129521.

39. Dubankova A, Boura E. 2019. Structure of the yellow fever NS5 protein reveals conserved drug targets shared among flaviviruses. Antiviral Research 169:104536.

40. Barrows NJ, Campos RK, Liao K-C, Prasanth KR, Soto-Acosta R, Yeh S-C, Schott-Lerner G, Pompon J, Sessions OM, Bradrick SS, Garcia-Blanco MA. 2018. Biochemistry and Molecular Biology of Flaviviruses. Chemical Reviews 118:4448–4482.

41. Lu X, Xiao H, Li S, Pang X, Song J, Liu S, Cheng H, Li Y, Wang X, Huang C, Guo T, ter Meulen J, Daffis S, Yan J, Dai L, Rao Z, Klenk H-D, Qi J, Shi Y, Gao GF. 2019. Double Lock of a Human Neutralizing and Protective Monoclonal Antibody Targeting the Yellow Fever Virus Envelope. Cell Reports 26:438-446.e5.

42. Watanabe Y, Bowden TA, Wilson IA, Crispin M. 2019. Exploitation of glycosylation in enveloped virus pathobiology. Biochimica et Biophysica Acta (BBA) - General Subjects 1863:1480–1497.

43. Puerta-Guardo H, Glasner DR, Harris E. 2016. Dengue Virus NS1 Disrupts the Endothelial Glycocalyx, Leading to Hyperpermeability. PLOS Pathogens 12:e1005738.

44. Luo D, Xu T, Hunke C, Grüber G, Vasudevan SG, Lescar J. 2008. Crystal Structure of the NS3 Protease-Helicase from Dengue Virus. Journal of Virology 82:173–183.

45. Zou J, Xie X, Wang Q-Y, Dong H, Lee MY, Kang C, Yuan Z, Shi P-Y. 2015. Characterization of Dengue Virus NS4A and NS4B Protein Interaction. Journal of Virology 89:3455–3470.

46. Miller S, Kastner S, Krijnse-Locker J, Bühler S, Bartenschlager R. 2007. The non-structural protein 4A of dengue virus is an integral membrane protein inducing membrane alterations in a 2K-regulated manner. J Biol Chem 282:8873–8882.

47. Lin M-H, Hsu H-J, Bartenschlager R, Fischer WB. 2014. Membrane undulation induced by NS4A of Dengue virus: a molecular dynamics simulation study. Journal of Biomolecular Structure and Dynamics 32:1552–1562.

48. El Sahili A, Lescar J. 2017. Dengue Virus Non-Structural Protein 5. Viruses 9.

49. Zhu W, Chen CZ, Gorshkov K, Xu M, Lo DC, Zheng W. 2020. RNA-Dependent RNA Polymerase as a Target for COVID-19 Drug Discovery. SLAS DISCOVERY: Advancing the Science of Drug Discovery 247255522094212.

50. Ashour J, Laurent-Rolle M, Shi P-Y, García-Sastre A. 2009. NS5 of Dengue Virus Mediates STAT2 Binding and Degradation. J Virol 83:5408–5418.

51. Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in Bioinformatics 20:1160–1166.

52. Friedman JH. 2001. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics 29:1189–1232.

53. Schapire RE. 2003. The Boosting Approach to Machine Learning: An Overview, p. 149–171. *In* Denison, DD, Hansen, MH, Holmes, CC, Mallick, B, Yu, B (eds.), Nonlinear Estimation and Classification. Springer New York, New York, NY.

54. Breiman L. 2001. Random forests. Machine learning 45:5–32.

55. Morais-Rodrigues F, Silvério-Machado R, Kato RB, Rodrigues DLN, Valdez-Baez J, Fonseca V, San EJ, Gomes LGR, dos Santos RG, Vinicius Canário Viana M, da Cruz Ferraz Dutra J, Teixeira Dornelles Parise M, Parise D, Campos FF, de Souza SJ, Ortega JM, Barh D, Ghosh P, Azevedo VAC, dos Santos MA. 2020. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. Gene 726:144168.

56. Goodsell DS, Zardecki C, Di Costanzo L, Duarte JM, Hudson BP, Persikova I, Segura J, Shao C, Voigt M, Westbrook JD, Young JY, Burley SK. 2020. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. Protein Sci 29:52–65.

57. Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385.