Virology 522 (2018) 122–130

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/virology

Molecular footprints of selective pressure in the neuraminidase gene of currently circulating human influenza subtypes and lineages

Vanessa Correia[a,b], Ana B. Abecasis[c], Helena Rebelo-de-Andrade[a,b,*]

[a] Infectious Diseases Department, Instituto Nacional de Saúde Doutor Ricardo Jorge, IP, Av. Padre Cruz, 1649-016 Lisbon, Portugal
[b] Host-Pathogen Interaction Unit, Research Institute for Medicines (iMed.ULisboa), Faculdade de Farmácia, Universidade de Lisboa, Av. Professor Gama Pinto, 1649-003 Lisbon, Portugal
[c] Global Health and Tropical Medicine, Instituto de Higiene e Medicina Tropical, Universidade NOVA de Lisboa, Rua da Junqueira 100, 1349-008 Lisbon, Portugal

ARTICLE INFO

Keywords:
Influenza
Neuraminidase
dN/dS ratio
Overall selection
Positively selected sites
Site-specific selective pressures

ABSTRACT

Influenza neuraminidase (NA) is under selective pressure (SP) of both host immune system and drug use. Here, we assembled large datasets of NA sequences of worldwide circulating viruses to estimate the global and site-specific SP acting on all current subtypes/lineages of human influenza NA. An overall negative SP of similar magnitude and a prevalence of negatively selected sites were observed for all subtypes/lineages. Positively selected sites varied according to the subtype/lineage, including N1-NA sites 247 and 275, N2-NA sites 148 and 151, and B/Victoria-NA site 395 associated with drug-resistance or reduced susceptibility. These results evidenced a potential role of positive selection in the low-level spread of A(H1N1)pdm09-H275Y drug-resistant viruses, and alerted for a potential higher risk of spread of a synergistic A(H1N1)pdm09 drug-resistant variant (H275Y/S247N). The positive selection detected at N2-NA sites 148 and 151 was probably an artefact from cell-culture. Overall mapping revealed six potential new druggable regions.

1. Introduction

Neuraminidase (NA) is the second major surface protein of influenza A and B viruses. It plays a key role at the final stage of infection by removing terminal sialic acid moieties from host-cell receptors and newly synthesized hemagglutinin (HA) and NA proteins, facilitating the release of progeny virions from infected cells and preventing their self-aggregation (Garman and Laver, 2005). NA also exerts its sialidase activity on the mucus layer overlaying the human airway epithelium (rich in sialic acid), cleansing the environment and facilitating the access of virus particles to host cells (Ison and Hay, 2013). As a surface antigen, NA is under constant selective pressure (SP) from the host immune system (antibody-mediated response). The selection of amino acid variants with mutations that hinder antibody recognition (antigenic drift) is one of the most important mechanisms driving the evolution of influenza A NA and, to a lesser extent, influenza B NA (Chen and Holmes, 2008). SP forces acting on virus NA may also arise from interactions with other viral genes, stability of the protein structure and adaptation to new species (Illingworth and Mustonen, 2012). Moreover, the introduction of an antiviral drug class targeting NA protein into the market in 1999 - NA inhibitors (NAIs), created an additional source of SP, changing the environment in which NA was evolving.

In spite of its role as a target for both neutralizing antibodies and antiviral drugs, NA has not been as extensively studied as HA (primary target of immune response) in regard to the SP forces acting on its coding gene. Most studies are focused on the N1 NA of recently emerging A(H1N1)pdm09 viruses and/or based on sequence datasets either comprising a small number of sequences (75–345 sequences) or covering short periods of time (1–11 years) (Arunachalam, 2013; Chen and Holmes, 2008; Espinola, 2012; Janies et al., 2010; Khandaker et al., 2013; Suzuki, 2006; Vijaykrishna et al., 2015; Westgeest et al., 2012, 2014; Wilbert, 2008). Also, none of the studies revealed a complete picture of the site-specific SP acting on NA that would allow knowing the SP acting on other codon sites (herein designated as sites) than those positively or negatively selected, particularly on the sites functionally relevant.

Investigating the SP forces acting on a gene is important for a better understanding of its evolutionary dynamics. It may also allow for the identification of new potential regions for drug targeting as amino acid substitutions at negatively selected sites are likely to be intolerable

* Correspondence to: Antiviral Resistance Lab, Research & Development Unit, Infectious Diseases Department, Instituto Nacional de Saúde Doutor Ricardo Jorge, IP, Av. Padre Cruz, 1649-016 Lisbon, Portugal.
E-mail addresses: vanessa.correia@insa-min-saude.pt (V. Correia), ana.abecasis@ihmt.unl.pt (A.B. Abecasis), h.rebelo.andrade@insa.min-saude.pt (H. Rebelo-de-Andrade).

https://doi.org/10.1016/j.virol.2018.07.002
Received 17 November 2017; Received in revised form 3 July 2018; Accepted 4 July 2018

(Suzuki, 2006). Finding alternative targets for new drugs is currently a priority in influenza given the limited repertoire of effective antivirals available. NAIs constitute the only antiviral drug class effective against currently circulating seasonal influenza viruses and only two inhibitors are approved worldwide - oseltamivir (OS) and zanamivir (ZA) (Ison and Hay, 2013). In protein antigens, positively selected sites may also be useful for identifying epitope contacting residues. Antibody epitopes in influenza NA are still very poorly characterized, particularly in type B NA in which only few amino acid residues identified through *in vitro* selection of antibody escape mutants are known (Air et al., 1990).

Here, we conducted a large-scale sequence analysis to investigate the global and site-specific SP acting on influenza NA, extending the analysis to all virus subtypes and lineages currently circulating in humans - A(H1N1)pdm09, A(H3N2), and B/Victoria (B/VIC) and B/Yamagata (B/YAM) lineage.

2. Materials and methods

2.1. Sequence datasets

For this study all potentially complete NA coding sequences of human influenza viruses circulating worldwide available at the Global Initiative on Sharing All Influenza Data (GISAID) EpiFlu™ database (http://platform.gisaid.org) (Shu and McCauley, 2017) and the National Center for Biotechnology Information (NCBI) Influenza Virus Resource (IVR) database (https://www.ncbi.nlm.nih.gov/genomes/FLU) (Bao et al., 2008) were selected. A minimum length of 1407 nucleotides (nts) was considered for A(H1N1)pdm09 and A(H3N2) NA sequences, while for influenza B lineage sequences the limit was set at 1398 nts (total length of coding regions excluding the stop codon). Databases were accessed on 7 and 30 April 2013 (A(H3N2) and A(H1N1)pdm09, respectively) and on 11 November 2013 (influenza B). A total of 14589 A(H1N1)pdm09 and 14689 A(H3N2) NA sequences were retrieved from both databases, while 2780 B/VIC- and 2137 B/YAM-lineage NA sequences were retrieved from GISAID EpiFlu™ database and 2772 influenza B NA sequences from NCBI IVR database. Sequences derived from laboratory strains and sequences repeated between the two databases, identified by the strain name, were excluded. The lack of lineage information for most influenza B NA sequences available at NCBI IVR database required a preliminary phylogenetic analysis with B/VIC and B/YAM-lineage reference sequences. Reference sequences were based on the World Health Organization (WHO) reports for the annual consultations on the composition of influenza vaccines (available at https://www.crick.ac.uk/research/worldwide-influenza-centre/annual-and-interim-reports/) and retrieved from GISAID EpiFlu™ database, including: B/Brisbane/60/2008 (EPI173276), B/Hong Kong/514/2009 (EPI243626), B/Denmark/23/2012 (EPI408048), B/Formosa/V2367/2012 (EPI406251), B/Hong Kong/3496/2012 (EPI392576), B/Johannesburg/3964/2012 (EPI406273) (B/VIC-lineage); and B/Brisbane/3/2007 (EPI463609), B/Wisconsin/01/2010 (EPI271599), B/Estonia/55669/2011 (EPI319346), B/Stockholm/12/2011 (EPI319559), B/Hong Kong/3577/2012 (EPI392590), B/Massachusetts/02/2012 (EPI376343), B/Novosibirsk/1/2012 (EPI368751) (B/YAM-lineage). The sequences exclusive of NCBI IVR database and reference sequences were aligned by Clustal W method in MEGA version 5 (MEGA5) (Tamura et al., 2011) and a maximum-likelihood (ML) phylogenetic tree was inferred as described below in the phylogenetic inference section. An influenza B lineage was then assigned to each virus sequence according to its position in the tree. To these publicly available sequences were further added unpublished sequences from viruses circulating in Portugal. Specifically, 57 A(H1N1)pdm09, 73 A(H3N2), 5 B/VIC- and 20 B/YAM-lineage NA sequences.

The NA sequences of each human influenza virus subtype or lineage were aligned by Clustal W method in MEGA5. Each alignment was then manually inspected and trimmed to include coding regions only, and sequences containing gaps or not completed and different sequences from a same virus strain, identified by the strain name, were excluded. Redundant sequences (100% similarity) and sequences containing degenerate or untranslatable nucleotides were also excluded using Jalview version 2.8 (Waterhouse et al., 2009). Due to posterior software constraints in SP analyses, a sequence similarity threshold of 99.99% was further applied to A(H3N2) and A(H1N1)pdm09 datasets to reduce the number of sequences to below 4000. This step was performed in CD-HIT-EST, using the web server CD-HIT Suite (http://weizhong-lab.ucsd.edu/cdhit_suite) (Huang et al., 2010).

Final datasets were comprised as follows: A(H1N1)pdm09 - 3428 sequences, from 2009 to 2013; A(H3N2) – 3712 sequences, from 1968 to 2013; B/VIC – 1978 sequences, from 1972 to 2013; B/YAM – 1441 sequences, from 1973 to 2013. Each dataset included sequences from all years or epidemiological weeks (A(H1N1)pdm09) within the time period covered, and from all 5 worldwide continents and 18 WHO influenza transmission zones (*i.e.* geographical area with a similar influenza transmission pattern (WHO, 2011)) (data not shown). Sequence diversity (number of base substitutions per site from a mean of all sequence pairs) was estimated by maximum composite likelihood method in MEGA5, using a standard gamma distribution for rate variation among sites and 500 bootstrap replications for estimating the associated standard error. Each dataset was also tested for evidence of recombination by genetic algorithm recombination detection method (SingleBreakpointRecomb.bf batch file) in HyPhy version 2.2 (HyPhy 2.2) (Pond et al., 2005). Since no single phylogenetic tree can accurately describe the evolutionary relationships of recombinant sequences, it is important to rule out or account for recombination as its presence can mislead inferences of selection (Noh and Marshall, 2016).

2.2. Maximum-likelihood based phylogenetic inference

ML phylogenetic trees were inferred using the PhyML 3.0 platform (Guindon et al., 2010) implemented in SeaView version 4.4.0 (Gouy et al., 2010) or, when the number of sequences was above 3000 (A(H1N1)pdm09, A(H3N2)) using RAxML version 8 (Stamatakis, 2014). The general time reversible (GTR) model assuming a proportion of invariable sites (I) and with a 4-category gamma distribution (Γ4) for rate variation among sites (GTR + I + Γ4 model) was used for tree construction. This model was determined as the best-fit agreed model of nucleotide substitution for the different types and subtypes of human influenza NA in a previous study conducted with sequences of viruses circulating in Portugal and reference sequences (Giria et al., 2012). Briefly, model selection was performed in jModelTest version 2.1.2 (Darriba et al., 2012; Guindon and Gascuel, 2003) according to Akaike's information criterion (AIC), using a 95% confidence interval (CI) and likelihood scores estimated with 11 substitution schemes. The option “best” of nearest-neighbour-interchange (NNI) and subtree-pruning-and-regrafting (SPR) rearrangement operations was selected in PhyML to optimize tree topology, using 10 random starting trees. Branch support values were estimated by approximate likelihood-ratio test (aLRT) with the nonparametric Shimodaira-Hasegawa (SH) correction (SH-aLRT) (PhyML) or by standard nonparametric bootstrap (500 replicates) (RAxML).

2.3. Selective pressure analysis

SP was measured by the ratio of the number of nonsynonymous substitutions per nonsynonymous site (dN) to the number of synonymous substitutions per synonymous site (dS). A dN/dS ratio significantly less than one indicates negative selection, $dN/dS \approx 1$ represents neutral evolution, and a dN/dS ratio significantly greater than one provides evidence of positive selection (Pond et al., 2009). SP analyses were carried out in HyPhy 2.2, using the final sequence alignment and corresponding ML phylogenetic tree as input. The analyses were preceded by the selection of the time-reversible model that