Infection, Genetics and Evolution 11 (2011) 997–1005

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Research paper

Evolutionary dynamics of the influenza A pandemic (H1N1) 2009 virus with emphasis on Indian isolates: Evidence for adaptive evolution in the HA gene

Jayati Mullick [1], Sarah S. Cherian [1], Varsha A. Potdar, Mandeep S. Chadha, Akhilesh C. Mishra *

National Institute of Virology, 20-A, Ambedkar Road, Pune 411001, India

ARTICLE INFO

Article history:
Received 10 October 2010
Received in revised form 17 March 2011
Accepted 22 March 2011
Available online 30 March 2011

Keywords:
Influenza
Pandemic H1N1 (pH1N1)
Hemagglutinin (HA)
Evolutionary dynamics
Molecular clock
Selection pressure
India

ABSTRACT

The indigenous transmission of the 2009 pandemic H1N1 (pH1N1) virus in India made it as one of the major sub-types in circulation. Genetic characterization indicated that the viruses predominantly clustered in clade 7, the globally most widely circulating pH1N1 clade. It is imperative to continue monitoring the genetic make-up of the pH1N1 viruses to understand their adaptability and evolutionary dynamics in the country. We characterized 31 full genomes and 94 hemagglutinin (HA) sequences of the pH1N1 viruses from various regions of India (May 2009–October 2010). Among the newly identified mutations reported in the pH1N1 viruses that could alter the viral fitness, E374K in the HA was increasingly noted in 35 Indian isolates beyond September 2009 and its co-occurrence with D97N or V30A was also observed in the more recent isolates. Molecular clock analysis based on all Indian isolates and closely related global representatives indicated higher substitution rates ($\sim 7.1 \times 10^{-3}$ subs/site/year) when compared to an earlier report. Several independent introductions were noted within the country along with considerable evidence of indigenous evolution during the latter period of the study. The estimate for the mean age of the common ancestor of all the pandemic isolates dated to around August 2008 correlating well with the global estimate. Evidence for adaptive evolution in the HA was observed in the clade 7 isolates at the 'Ca' antigenic site that may have implications for future re-evaluation of the vaccine composition. The study thus warrants the need for continued surveillance and genetic characterization of whole genome sequences to detect any possible reassortment events that might further contribute to the viral fitness of the pH1N1 viruses.

1. Introduction

Since the first report of the Influenza A pandemic (H1N1) 2009 viruses in April 2009 from Mexico (Mossad, 2009; CDC, 2009a), the virus has spread globally. The dynamics and impact of this pandemic in different geographical regions is expected to be different. It is speculated that in densely populated low-income countries, having limited public health facilities and drug availability, it would be impossible to control the pandemic thus resulting in a greater number of cases occurring in the susceptible populations (Boni et al., 2009). The pandemic H1N1 (pH1N1) viruses possess a unique genetic combination of a triple reassortant wherein the hemagglutinin (HA), nucleoprotein (NP) and non-structural (NS) protein genes belong to the classical swine lineage, the neuraminidase (NA) and the matrix (M) protein genes descend from the Eurasian swine lineage which entered pigs from the avian hosts around 1979, while the polymerase gene segments PB1, PB2 and PA were derived from the North American triple reassortant swine lineage (Garten et al., 2009; Itoh et al., 2009; Smith et al., 2009).

The pH1N1 virus was first detected in India in May 2009 (Potdar et al., 2010) and since then outbreaks have been reported from many parts of the country. During the initial phase of the pandemic, a large number of cases were reported from major cities, though as the pandemic progressed, an increasing number of cases were being reported from rural areas and outskirts of the cities. Within a short period after the establishment of the indigenous transmission, pH1N1 became one of the predominant sub-types in the country and co-circulated with the seasonal influenza viruses (Mishra et al., 2010). Both seasonal-A and pH1N1 influenza viruses were represented in almost equal proportion in most parts of the country. As of October 2010, the total number of confirmed cases in India was about 45,101 with 2679 deaths and an average percent positivity of ~23%, (Ministry of Health and Family Welfare, 2010) Beyond October 2010, only sporadic cases of pH1N1 are being reported in the country.

* Corresponding author. Tel.: +91 20 26127301; fax: +91 20 26122669.
E-mail addresses: mullick.j@niv.co.in (J. Mullick), cherian.ss@niv.co.in (S.S. Cherian), potdar.va@niv.co.in (V.A. Potdar), chadha.ms@niv.co.in (M.S. Chadha), acm1750@rediffmail.com, mishra.ac@niv.co.in (A.C. Mishra).
[1] Both the authors contributed equally to this work.

1567-1348/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.03.015

Our earlier study on surveillance and subsequent genetic characterization of 13 isolates during the period May–November 2009 had revealed the introduction of three lineages, namely clades 5–7 in the country (Potdar et al., 2010). The predominant clade 7 was found to be the globally most widely circulating clade. Genetic sequence information of an increasingly large number of isolates from different countries since early 2009 has shown the increased presence of amino-acid polymorphisms in the outbreak sequences (Smith et al., 2009). It remained to be understood whether this was due to a burst of adaptive evolution in the new host or a relaxation of purifying selection. Considering the unabated circulation of the pH1N1 virus in several parts of India for a significant period of time, it becomes important to monitor the genetic diversity within the country to detect virulent variants and understand the evolutionary dynamics of the viruses.

In the present study we have characterized 31 full genomes and 94 HA sequences of the pH1N1 viruses from various regions of India during the period May 2009–October 2010 to gain insight into: (i) the mutational trend of residues at the signature sites and the unique changes (ii) existence of monophyletic lineages within Indian isolates (iii) the molecular clock and selection pressure. Such information may also have implications for the selection of the vaccine strains during the subsequent seasons in the country.

2. Materials and methods

2.1. Virus isolation and sequencing

The virus was isolated from the clinical materials, naso-pharyngeal/throat swabs, obtained from patients during the period May 2009–October 2010. Selection of samples (Supplementary Table 1) for the whole genome ($n = 31$) was based on the representation of major cities in India while sequencing of the HA gene ($n = 94$) majorly represented several districts in the state of Maharashtra, in proportion to the number of referred samples and subsequent pH1N1 positivity. After processing the clinical materials by the standard laboratory procedures the individual sample was inoculated in Madin Darby Canine Kidney (MDCK) cell lines for the isolation of the respective viruses. In few cases, Specific pathogen-free (SPF) embryonated White Leghorn eggs were also used. Supplementary Table 1 gives details of the isolate name, date and location of sample collection, patient information and isolate's passage history. The inoculated MDCK/eggs cell line was observed for cytopathic effects following which the tissue culture fluid/allantoic fluid was harvested. Hemagglutination (HA) and Hemagglutination inhibition (HAI) tests were performed using guinea pig RBC as described previously (Kendal et al., 1982).

RNA extraction and sequencing of the selected isolates were performed as described earlier (Ray et al., 2008). In brief, 140 μl of the isolate was used for the RNA isolation using the QIAmp viral RNA mini kit (Qiagen, Germany) according to the manufacturer's instructions. All the eight influenza gene segments from 31 isolates selected for full genomes and the HA gene segment of 94 isolates mentioned here were subjected to One-step reverse transcription polymerase chain reaction (RT-PCR) using the WHO-CDC recommended whole genome primers (CDC, 2009b). The segments were amplified using the Superscript III (Invitrogen) enzyme in tiled amplicons of 500–600 bp having a 100 bp overlap to get a minimum four fold overlap and coverage. Sequencing of the amplified DNA was carried out using Big Dye terminator V3.1 cycle sequencing ready reaction kit (ABI, USA) as described earlier (Potdar et al., 2010). The sequences of representative samples grown in both MDCK and egg were found to be identical. As the isolates have undergone low levels of passage (Supplementary Table 1) the possibility of mutations being incorporated during cell culture adaptation, is ruled out. The GenBank accession numbers of the sequences are also mentioned in Supplementary Table 1.

2.2. Phylogenetic analysis, evolutionary rate and molecular clock analyses

The HA-based phylogenetic tree was constructed based on the 94 HA gene sequences (including 13 sequences from Potdar et al., 2010) using the maximum likelihood (ML) approach in PhyML program (Guindon and Gascuel, 2003) with 500 bootstrap replicates. Phylogenetic analysis of the 31 whole genomes (including 6 genomes from Potdar et al., 2010) was also inferred using the ML method by concatenating the eight segments considering that there were no reassortments. The analysis utilized the HKY + I + G4 nucleotide substitution model and employed the SH-like branch support. In addition to the Indian sequences, global representatives were also included in the present study. The global sequences were selected using an initial dataset of 2256 whole genomes available in GenBank as on May 30, 2010 and constructing distance-based trees using the neighbour-joining (NJ) method in MEGA version 4 (Kumar et al., 2004).

For temporal phylogeny and estimation of the rates of substitution, a Bayesian Markov chain Monte Carlo (MCMC) framework as in BEAST (Drummond and Rambaut, 2007) was applied. The dataset was composed of the HA gene sequences ($n = 135$) of all the 94 Indian isolates and 41 closely related globally representative isolates, having known sampling time. All sequences were dated considering the reference point as January 1, 2009. We employed both the strict and relaxed (uncorrelated lognormal and uncorrelated exponential) molecular clocks with constant, exponential, expansion and Bayesian population demographic models in each case. The strict clock assumes a single evolutionary rate along all branches while the uncorrelated lognormal relaxed (UCLD) clock allows the evolutionary rate to vary along the branches within lognormal distributions. The best nucleotide substitution model detected by the Akaike Information Criterion (AIC) in MODELTEST 3.7 (Posada and Crandall, 1998) was found to be the TIM2 + I + G4 model. The model close to the best fit model, being TN93 in BEAST, was chosen. Three independent runs of 50 million steps were used and combined with a 10% 'burn-in'. The convergence was analysed by using Tracer 1.4 (Rambaut and Drummond, 2007) and effective sample size (ESS) values of >200 indicated sufficient level of sampling. Model comparison was done by comparing posterior and marginal likelihood values along with the Bayes factor. The maximum clade credibility tree was generated by using Tree Annotator and Fig Tree 1.2.2 (http://tree.bio.ed.ac.uk/) was used for visualization of the annotated trees. The 95% highest probability density (HPD) intervals were obtained to ascertain the uncertainty in the parameter estimates.

2.3. Selection pressure analysis

To identify the existence of positive selection pressure at individual codon sites, three likelihood procedures were used: the single-likelihood ancestor counting (SLAC) method, fixed effects likelihood (FEL) method and the more powerful random-effects likelihood (REL) method (Pond and Frost, 2005). The mean ratio of non-synonymous changes per non-synonymous site (dN) and the synonymous changes per synonymous site (dS) was obtained using SLAC which calculates inferred ancestral sequences for each internal node in a phylogeny using a codon model and then counts the synonymous and non-synonymous mutations by comparing each codon to its immediate ancestor. All analyses were carried out using the online Datamonkey facility (http://www.datamonkey.org/), incorporating the HKY model of nucleotide substitution, with the phylogenetic tree inferred using the NJ method.