Infection, Genetics and Evolution 11 (2011) 437–441

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

The role of swine as "mixing vessel" for interspecies transmission of the influenza A subtype H1N1: A simultaneous Bayesian inference of phylogeny and ancestral hosts

Johanna Haß [a,1,3,*], Sebastian Matuszewski [a,2,3], Dietmar Cieslik [a], Martin Haase [b]

[a] Institute for Mathematics and Information Theory, University of Greifswald, Walther-Rathenau-Str. 47, D-17487 Greifswald, Germany
[b] Bird Observatory, Zoological Institute and Museum, University of Greifswald, Soldmannstr. 23, D-17489 Greifswald, Germany

ARTICLE INFO

Article history:
Received 18 October 2010
Received in revised form 30 November 2010
Accepted 1 December 2010
Available online 14 December 2010

Keywords:
Influenza A
H1N1
Phylogeny
Host shift
Ancestral reconstruction

ABSTRACT

Condensing the information of a total of 1551 to 469 influenza A H1N1 isolates we investigated the frequency of host shifts among bird, human and swine. Phylogenies of hemagglutinin and neuraminidase as well as ancestral host reconstructions were simultaneously inferred in a Bayesian framework. The surface proteins had to be analyzed separately because of reassortment. Also the different tree topologies indicated the different evolutionary histories of these genes. The majority of interspecies transmissions involved isolates from swine confirming the role of pigs as "mixing vessel" for the influenza A virus. This was emphasized by the investigation of host specific amino acid positions. However, the simultaneous estimation of phylogeny and ancestral states resulted in considerable ambiguity in particular at deeper nodes and at the root cautioning against overstated conclusions. Our analysis highlights the urge of intensifying influenza surveillance programs for porcine hosts.

1. Introduction

Influenza is a seasonally occurring respiratory disease in humans. Worldwide, there are between three and five million infections per year, 10% of which end fatally. The viral infection caused by the influenza virus is classified into the three types A, B, and C, with A being mainly responsible for human infections. Influenza A is further classified into subtypes referring to the combination of the two surface proteins hemagglutinin (HA) and neuraminidase (NA) (WHO, 2009).

The natural host for all influenza A subtypes are wild birds. Despite the high number of possible combinations (15 HA types × 9 NA types = 135) only H1, H2 and H3 as well as N1 and N2 have successfully established in humans (Baigent and McCauley, 2003). Most relevant for seasonal outbreaks are the subtypes H1N1 and H3N2, in temperate zones preferentially in autumn and winter (Rambaut et al., 2008; WHO, 2009).

Besides birds and humans the virus can also occur in other mammals like swine, horse, or whales, but for physiological reasons interspecies transmissions are rare. However, successful transmission between two species can rapidly lead to a dispersion of the virus of pandemic dimension due to the new host's lack of antibodies (Baigent and McCauley, 2003). The first recorded and since then most devastating pandemic was the "Spanish Flu" in 1918 with more than 40 million deaths. Induced by the same subtype – H1N1 – that circulated until the fifties, the "Russian Flu" of 1977 still has an unexplained origin (Taubenberger and Morens, 2006).

The genome of the influenza virus contains eight pieces of segmented negative-sensed RNA with 13,600 bases all together that code for 10 viral proteins. The trimeric complex of hemagglutinin is encoded on the fourth segment and is responsible for virus binding and for adsorption to the host cell (Skehel and Wiley, 2000). The receptor binding site is host specific due to its configuration at certain codon positions (Baigent and McCauley, 2003). Binding takes place with sialic acid-α-2,3-galactose or sialic acid-α-2,6-galactose (Baigent and McCauley, 2003; Shen et al., 2009). As α-2,3 bound receptors are mainly found in birds, and α-2,6 bound receptors appear in the trachea of humans, a direct

* Corresponding author. Tel.: +49 0351 4587060; fax: +49 0351 4585754.
E-mail addresses: johannahass@hotmail.de (J. Haß), sebastian.matuszewski@univie.ac.at (S. Matuszewski), cieslik@uni-greifswald.de (D. Cieslik), martin.haase@uni-greifswald.de (M. Haase).
[1] Present address: Universitätsklinikum Carl Gustav Carus, Klinik und Poliklinik für Kinder- und Jugendpsychiatrie und -psychotherapie, Fetscherstr. 74, D-01307 Dresden, Germany.
[2] Present address: Mathematics and BioScience Group, University of Vienna, Nordbergstr. 15, A-1090 Vienna.
[3] These authors contributed equally to this work.

1567-1348/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2010.12.001

transmission is rarely successful but possible – as human cases of avian flu H5N1 demonstrate – because of the existence of avian specific receptors in the lower respiratory system. Swine has both receptor types and has therefore been called the "mixing vessel" for the influenza virus because it can function as an intermediate host (Castrucci et al., 1993; Scholtissek, 1994; Kida et al., 1994; Ma et al., 2008).

Neuraminidase, which is also indispensable for virus transmission, is encoded on segment six. This homotetrameric complex is functionally tightly linked with hemagglutinin and responsible for the virus release from the host cell, eventually leading to the infection of new cells. The release is again facilitated via host specific codon positions (Tamuri et al., 2009).

Phylogenetic reconstructions have been important for our understanding of virus evolution and epidemic pathways including geographical spread and host shifts (Rambaut et al., 2008; Tamuri et al., 2009; Lemey et al., 2009). Methodologically, tree reconstructions and reconstruction of ancestral states, such as geographical locations, hosts or codon positions (Wilson et al., 1991; Ronquist, 1994; Akashi et al., 2007) have been conducted in a parsimony framework. However, parsimony reconstructions do not account for statistical uncertainty both in the reconstruction of the phylogeny as well as the inference of ancestral states (Cunningham et al., 1998; Gubareva et al., 2002; Lemey et al., 2009; Haase et al., 2010; Hovmoller et al., 2010). In addition parsimony ignores branch lengths.

Probabilistic approaches have only recently been applied in the study of viral evolution. Haase et al. (2010), for example, reconstructed host shifts and ancestral ranges of H5N1 in a likelihood framework, but still using a single "optimal" tree. A fully probabilistic Bayesian approach allowing to express mapping and phylogenetic uncertainty simultaneously has been developed by Lemey et al. (2009). Most importantly probabilistic methods avoid overconfidence in seemingly unambiguous inferences usually resulting from maximum parsimony analysis.

Here we investigate the interspecies transmission of the H1N1 influenza A virus between the three hosts human, bird and swine based on all hemagglutinin and neuraminidase sequences available from the NCBI Flu-database as of October 31, 2009. We incorporated an ancestor state reconstruction into a Bayesian framework to study host shifts and additionally examined the human and avian specific amino acid positions in porcine sequences. Central to our analyses was the question about the role of swine as "mixing vessel" for H1N1.

2. Materials and methods

2.1. Data selection

We retrieved all human, avian and porcine H1 and N1 full-length sequences available from the NCBI Flu-database as of October 31, 2009 (Bao et al., 2008; NCBI, 2009). We only excluded sequences from the pandemic (H1N1) 2009 virus, since their origin has recently been clarified (Smith et al., 2009). Sequences annotated as full-length by mistake were excluded by using an additional protein specific length criterion: minimum length of H1 = 1600 bases, minimum length of N1 = 1300 bases. The remaining data set was further reduced to those H1 and N1 sequences stemming from the same isolate resulting in 1551 pairs of sequences. In the next step isolates with redundant information with respect to host, geography and time of isolation (±3 years) were eliminated by manually comparing neighbor-joining trees of H1 and N1 reconstructed in `SplitsTree` (Huson and Bryant, 2006). Among the remaining 469 pairs of sequences (Supplementary Material S1) there were 327 human, 104 porcine, and 38 avian hosts and therefore still a bias towards human isolates especially from the USA (Supplementary Materials S2.1 and S.2.2).

2.2. Phylogenetic analysis

The 469 pairs of sequences were aligned using `mafft` (Katoh et al., 2002, 2005; Katoh and Toh, 2008) and the resulting alignments manually corrected in `jalview` (Clamp et al., 2004; Waterhouse et al., 2009) and `BioEdit` (Hall, 1997–2005). The data set was tested for recombination using the *PHI*-test implemented in `PhiPack` (Bruen et al., 2006).

Phylogenetic analysis was carried out in a Bayesian framework using the software package `BEAST` (Drummond and Rambaut, 2007). We used codon position models to model sequence evolution, as they show better performance especially for coding sequences of RNA viruses (Shapiro et al., 2006). To each codon position the best fitting evolutionary substitution model – GTR+Γ model (Tavaré, 1986) in all three cases – was determined using `jModeltest` (Guindon and Gascuel, 2003; Posada, 2008, 2009) based on the corrected AIC (Akaike, 1973, 1974). We chose a relaxed lognormal molecular clock (Drummond et al., 2006), calibrated by the dates of isolation and assumed a constant population-size coalescent process prior over the unknown phylogeny (Kingman, 2000) to avoid possible overparametrization. According to Lemey et al. (2009) the demographic prior has little influence on the ancestral state inference. The model for ancestral state reconstruction was included manually following the approach recently published by Lemey et al. (2009) with human, swine and avian constituting our state set. Based on UPGMA starting trees, we carried out several independent simulations until all parameters reached an effective sample size of at least 200. For hemagglutinin six runs comprising a total of 800 million generations and for neuraminidase four runs with a total of 600 million generations were carried out. Every 5000th generation was sampled and a burn-in of 10% was discarded. All simulations were combined, whereby the respective tree-files had to be resampled at decreased frequency resulting in a total of 10,000 trees. We built maximum clade credibility trees for both H1 and N1 using `TreeAnnotater` (Drummond and Rambaut, 2007) with branch lengths rescaled to mean posterior estimates.

3. Results

3.1. Sequence evolution

In separate analyses recombination within H1 and N1, respectively, could not be detected (*PHI*-test: $p = 0.768$ for H1 and $p = 1.0$ for N1). However, for concatenated sequences the *PHI*-test did indicate recombination ($p = 3.66 \times 10^{-13}$). Because H1 and N1 did not share a common evolutionary history we only reconstructed separate trees for each protein.

Hemagglutinin had higher substitution rates at codon positions 1 and 2, whereas at position 3 neuraminidase evolved faster (Supplementary Materials S3.1 and S3.2). Overall rates were 2.631×10^{-3} for H1 vs. 2.470×10^{-3} nucleotide substitutions per site per year for N1. However, the differences were not significant as confidence intervals overlapped (Fig. 1).

In H1 the mean of the root was estimated to 1862, 30 years older than that of N1. However, confidence intervals again did overlap (Supplementary Material S3.3).

3.2. Phylogenetic trees

General simplified tree structures are presented in Fig. 2. Complete trees are presented in Supplementary Materials S4.1 and S4.2. The general tree of H1 comprised two clades. In the first the