Virology 524 (2018) 182–191

Contents lists available at ScienceDirect

Virology

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

Design and validation of a universal influenza virus enrichment probe set and its utility in deep sequence analysis of primary cloacal swab surveillance samples of wild birds

Yongli Xiao[a,*], Jacqueline M. Nolting[b], Zong-Mei Sheng[a], Tyler Bristol[a], Li Qi[a], Andrew S. Bowman[b], Jeffery K. Taubenberger[a]

[a] Viral Pathogenesis and Evolution Section, Laboratory of Infectious Diseases, NIH/NIAID, 33 North Drive MSC 3203, Bethesda, MD 20892-3203, USA
[b] Department of Veterinary Preventive Medicine, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

Keywords:
Influenza
Universal influenza enrichment
Deep sequencing
Mixed avian influenza infection

ABSTRACT

Influenza virus infections in humans and animals are major public health concerns. In the current study, a set of universal influenza enrichment probes was developed to increase the sensitivity of sequence-based virus detection and characterization for all influenza viruses. This universal influenza enrichment probe set contains 46,953 120nt RNA biotin-labeled probes designed based on all available influenza viral sequences and it can be used to enrich for influenza sequences without prior knowledge of type or subtype. Marked enrichment was demonstrated in influenza A/H1N1, influenza B, and H1-to-H16 hemagglutinin plasmids spiked into human DNA and in cultured influenza A/H2N1 virus. Furthermore, enrichment effects and mixed influenza A virus infections were revealed in wild bird cloacal swab samples. Therefore, this universal influenza virus enrichment probe system can capture and enrich influenza viral sequences selectively and effectively in different samples, especially ones with degraded RNA or containing low amount of influenza RNA.

1. Introduction

Influenza viruses have a negative-sense, single-stranded, segmented RNA genome, and are enveloped viruses of the family Orthomyxoviridae (Wright et al., 2007). They infect many warm blooded avian and mammalian species, including humans, and infections caused by these viruses have great public health, animal health, and economic significance. Influenza A viruses (IAV) are extremely diverse, both genetically and antigenically, and widely distributed in wild avian species globally, in domestic avian and mammalian species, and in humans. There are two major surface glycoproteins, hemagglutinin and neuraminindase with multiple subtypes. Currently, 18 hemagglutinin (HA) subtypes and 11 neuraminidase (NA) subtypes have been described, of which 16 HA and 9 NA subtypes are frequently found in avian species. Of the 144 possible HA-NA subtype combinations possible, at least 131 have been characterized in strains isolated from birds in NCBI influenza virus database (Bao et al., 2008). The two recently described IAV subtypes (H17N10 and H18N11) have only been isolated in fruit bats (Wu et al., 2014). Aquatic birds likely serve as the predominant natural reservoir for all other known subtypes and probably are the ultimate source of all human pandemic IAV strains (Webster et al., 1992).

Annual epidemics of influenza vary in their impact, but up to 56,000 people die annually from influenza in the United States (https://www.cdc.gov/flu/about/disease/2015–16.htm), and up to 650,000 annual influenza fatalities occur globally (http://www.who.int/mediacentre/factsheets/fs211/en/). Novel pandemic viruses occur sporadically in history. There have been four pandemics in the last century (Morens et al., 2009) and the 1918 H1N1 pandemic killed approximately 50 million people globally (Iuliano et al., 2018; Johnson and Mueller, 2002). Epizootic outbreaks of avian- and swine-origin IAV strains poses a risk for future pandemics. Type B and C influenza viruses are adapted to and isolated almost exclusively from humans, although influenza B viruses have been isolated from seals and influenza C viruses have been isolated from pigs and dogs (Wright et al., 2007).

Therefore, the continuous research on surveillance, rapid diagnosis, transmission, pathogenesis, and vaccinology of influenza is essential to prevent and mitigate its impact.

One critical aspect of influenza research is to detect and classify influenza viruses as to type, subtype, and genotype, especially for newly emerging viral variants in different samples, such as from wild birds, domestic animals, and human patients for the purposes of surveillance,

* Corresponding author.
E-mail address: yongli.xiao@nih.gov (Y. Xiao).

https://doi.org/10.1016/j.virol.2018.08.021
Received 10 May 2018; Received in revised form 21 August 2018; Accepted 28 August 2018
0042-6822/ Published by Elsevier Inc.

prevention, and treatment. A second important aspect to understanding the ecobiology of IAV, in birds especially, is the detection of mixed IAV infections in wild birds, which appear to be playing a key role in the natural history of these viruses. However, this task is extremely time and resource demanding. If multiple strains of avian IAV are present in the original sample, one strain of virus may outgrow others in embryonated chicken eggs, and other strains go undetected. Subtyping of avian IAV samples are classically determined by HA-and NA-inhibition tests (HIT and NIT) using reference antibodies and antigens (Lee et al., 2006). Therefore, over the years, the frequency of mixed IAV in wild birds has likely been greatly underestimated (Dugan et al., 2008; Wang et al., 2008).

Technological advances allow new methods to be developed to identify and characterize influenza virus isolates from a variety of sample types, including methods based on culture, antibody binding, serological assays, nucleic acid amplification strategies, and nucleic acid sequencing methodologies (Vemula et al., 2016). The latest and most comprehensive, but still costly, approach is the incorporation of high throughput sequencing technology. First introduced in 2005, 'high-throughput' DNA sequencers, which can determine megabases of DNA sequence per run (Service, 2006), have evolved dramatically, increasing sequencing capacity by a factor of 100–1000 (Goodwin et al., 2016). They have been widely used in influenza research studies, including detecting IAV and norovirus infections in patients (Nakamura et al., 2009), uncovering mixed infection with 2009 pandemic influenza A viruses (Ghedin et al., 2011), high throughput sequencing of influenza B viruses (Zhou et al., 2014), evaluating genetic stability of influenza vaccine viruses (Laassri et al., 2015), revealing antigenic variants at low frequencies in IAV-infected patients (Dinis et al., 2016), and high-throughput identification of influenza A/H3N2 virus antigenic drift variants (Mishin et al., 2017). However, because of the limited amount of viral RNA in typical clinical samples, all of these studies employed virus-specific primers and virus-specific PCR amplification strategies to enrich for target influenza sequences.

PCR-introduced errors have been emphasized more and more in next generation sequencing recently. It has been show that most commonly used PCR enzymes, including high fidelity enzymes, all have the error rates at 10^{-5} to 10^{-6} point mutations/bp/duplication (McInerney et al., 2014). Besides well-characterized polymerase base substitution errors, other sources of error were found to be equally prevalent, including PCR-mediated recombination, template-switching, and DNA damage introduced during temperature cycling (Potapov and Ong, 2017). In fact, Primer ID method was developed to distinguish PCR-introduced errors from real single nucleotide polymorphisms (SNPs) that occurred during virus evolution (Zhou et al., 2015). Another challenge using PCR to enrich cDNA derived from influenza virus RNA is that the isolated influenza RNA may have been degraded (Wang et al., 2008) and will be thus difficult to amplify by influenza universal primers requiring full-length segments as templates (Hoffmann et al., 2001). In many surveillance situations, subtype-specific primers cannot be used because the type(s) or subtype(s) of influenza virus in the sample is unknown.

Therefore, this study designed a set of universal influenza probes for enrichment of any influenza A, B, or C virus sequences by hybridization capture. These hybridization capture methods were first used to enrich sequence targets from the human genome. The designed oligonucleotide probes that capture the sequencing targets are attached to a solid phase support (Albert et al., 2007; Hodges et al., 2007; Okou et al., 2007). Subsequently, enrichment of the target sequences was performed in solution (Gnirke et al., 2009; Tewhey et al., 2009). Recently, a virome capture sequencing platform was developed for vertebrate viruses (Briese et al., 2015). Currently, many commercial companies provide different design and capture methods (Bodi et al., 2013; Garcia-Garcia et al., 2016; Zhou et al., 2009). In this study, we designed an enrichment probe set specific for all the influenza strains. Utilizing all available influenza A, B, and C virus sequences, a clustered unique data set for designing universal influenza enrichment probes was obtained. Using these universal influenza enrichment probes, we found significant enrichment of IAV and influenza B virus (IBV) RNA in control experiments. Application of the methods to cloacal swab samples collected from wild bird field surveillance, revealed IAV sequences and subtypes in these samples that were not detected by traditional methods.

2. Results

2.1. Design influenza universal enrichment probes

All the influenza A, B, and C virus sequences were downloaded from the NCBI influenza database (ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/) on 11/18/2015. A total of 408,140 influenza sequences were obtained. Among them, there were 390,301 (95.6%) 'clean' sequences and 17,839 (4.4%) 'non-clean' sequences (containing ambiguous non A, T, G, C bases). First, the clean sequences were collapsed into a unique set of 277,949 sequences using FASTX-Toolkit version 0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/) as the clean sequence set. For the sequences with ambiguous bases (17,839), if the total length of the ambiguous bases together in a sequence was longer than 10% of the total length of the sequence, they were discarded. Consequently, 649 sequences were discarded. The remainder of the 17,190 sequences were clustered at the 90% identity level by cd-hit-v4.3 (Li and Godzik, 2006) and resulted in 492 sequences. These sequences were split if they contained 8 Ns continuously and the separated sequences were retained if longer than 20 bp, which generated 649 sequences as the non-clean sequence set. Subsequently, the clean sequence set (277,949) and the non-clean sequence set (649) were combined and clustered together at the 90% identity level by cd-hit-v4.3 (Li and Godzik, 2006). This process generated a total of 905 sequences with 823 sequences derived from the clean sequence set and 82 sequences from the non-clean sequence set. After aligning these 82 sequences from the non-clean sequence set against 823 sequences from the clean sequence set, we found that 78 out of 82 sequences had high sequence homology, with the lowest percent identity at 82.24%. Therefore, these 78 sequences were eliminated from the dataset. Consequently, the final data set contained 823 sequences from the clean sequence set and 4 sequences from non-clean sequence set. The combined sequences (827) were clustered together at the 90% identity level by cd-hit-v4.3 (Li and Godzik, 2006) again and used to generate the final sequences (825). The process of generation of the final sequence data set from downloaded all influenza sequences is shown in Fig. 1. The final sequences (825) were used to make enrichment RNA oligonucleotide probes (Agilent Technologies, Santa Clara, CA) at 5X tiling resulting in 46,953 120nt RNA biotin-labeled universal influenza enrichment probes. All the enrichment probe sequences are listed in Supplemental Table 1.

2.2. In silico evaluation of enrichment probes

Based on phylogenetic analysis, IAV HA subtypes can be categorized into two major clades, group 1, which contains H1, H2, H5, H6, H8, H9, H11, H12, H13, H16, H17, and H18, and group 2 contains H3, H4, H7, H10, H14, and H15 (Air, 1981; Dugan et al., 2008; Joyce et al., 2016; Nobusawa et al., 1991). In an initial screen, the 1918 H1N1 pandemic strain (A/Brevig Mission/1/1918 (H1N1)) (group 1) and a 2013 H7N9 epizootic strain (A/Hangzhou/1/2013 (H7N9)) (group 2) were used to evaluate the probes *in silico*, in which the enrichment probes were aligned using the blastn program (with percent identity as 90 and e-value as 0.001 cutoff) (Altschul et al., 1990) against both viral genomes. Table 1 shows the coverage of each segment from the two different viruses provided by the probes and Fig. 2 shows the graphic coverage of the corresponding HA segments. As can be seen, the enrichment probes provided 100% coverage of all the segments from both IAV strains. In