Origin of the 1918 pandemic H1N1 influenza A virus as studied by codon usage patterns and phylogenetic analysis

DARISUREN ANHLAN,[1] NORBERT GRUNDMANN,[2] WOJCIECH MAKALOWSKI,[2] STEPHAN LUDWIG,[1] and CHRISTOPH SCHOLTISSEK[3]

[1]Institute of Molecular Virology (IMV), Centre of Molecular Biology of Inflammation (ZMBE), University of Münster, 48149 Münster, Germany
[2]Institute of Bioinformatics, University of Münster, 48149 Münster, Germany
[3]St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA

ABSTRACT

The pandemic of 1918 was caused by an H1N1 influenza A virus, which is a negative strand RNA virus; however, little is known about the nature of its direct ancestral strains. Here we applied a broad genetic and phylogenetic analysis of a wide range of influenza virus genes, in particular the *PB1* gene, to gain information about the phylogenetic relatedness of the 1918 H1N1 virus. We compared the RNA genome of the 1918 strain to many other influenza strains of different origin by several means, including relative synonymous codon usage (RSCU), effective number of codons (ENC), and phylogenetic relationship. We found that the *PB1* gene of the 1918 pandemic virus had ENC values similar to the H1N1 classical swine and human viruses, but different ENC values from avian as well as H2N2 and H3N2 human viruses. Also, according to the RSCU of the *PB1* gene, the 1918 virus grouped with all human isolates and "classical" swine H1N1 viruses. The phylogenetic studies of all eight RNA gene segments of influenza A viruses may indicate that the 1918 pandemic strain originated from a H1N1 swine virus, which itself might be derived from a H1N1 avian precursor, which was separated from the bulk of other avian viruses in toto a long time ago. The high stability of the RSCU pattern of the *PB1* gene indicated that the integrity of RNA structure is more important for influenza virus evolution than previously thought.

Keywords: negative strand RNA virus; 1918 pandemic virus; relative synonymous codon usage; RSCU patterns; effective number of codons; phylogenetic relationship

INTRODUCTION

Influenza A viruses are negative strand RNA viruses with a genome consisting of eight RNA segments encoding up to 11 viral proteins. These viruses exhibit great genetic variation, both by point mutations and by reassortment of their eight RNA segments between different isolates within the same type. This happens by coinfection in vitro or in vivo. Influenza A viruses are dangerous pathogens with the potential to cause pandemic outbreaks. The most serious pandemic in the last century, now known as the Spanish influenza, occurred in the winter of 1918/19. More than 40 million people died from an infection with an influenza A virus of the subtype H1N1. The exact origin of this pandemic influenza A virus strain is still not known. It has been suggested already in the early 1990s that at about the time of the Spanish flu an avian influenza virus had crossed the species barrier from birds to pigs and humans (Gammelin et al. 1990; Gorman et al. 1990, 1991); however, a detailed analysis only became possible after Taubenberger and colleagues isolated genetic material from viruses of the 1918 pandemic (Reid et al. 1999; Taubenberger et al. 2005). Cloning and sequencing of all eight virus genes (Taubenberger et al. 2005) revealed that these viruses are closely related to avian influenza viruses with regard to their coding sequence; however, the identity of the exact precursor of the pandemic strain still remained unclear (Taubenberger et al. 2005, 2007; Morens and Fauci 2007). Despite a variety of phylogenetic analyses performed so far, it is still a matter of debate whether the 1918 strain had crossed the species barrier from birds to humans in toto (Reid et al. 2004; Taubenberger et al. 2005; Rabadan et al. 2006) or whether the respective virus was a genetic reassortant or a recombinant strain (Fanning et al. 2002; Antonovics et al. 2006;

Reprint requests to: Stephan Ludwig, Institute of Molecular Virology (IMV), Centre of Molecular Biology of Inflammation (ZMBE), University of Münster, 48149 Münster, Germany; e-mail: ludwigs@uni-muenster.de; fax: 49-251-83-57793; or Wojciech Makalowski, Institute of Bioinformatics, University of Münster, 48149 Münster, Germany; e-mail: wojmak@uni-muenster.de; fax: 49-251-8353005.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.2395211.

Gibbs and Gibbs, 2006; Vana and Westover, 2008; Smith et al. 2009a). However, Taubenberger et al. (2005) noticed already that, when compared to avian sequences, the nucleotide sequences of the 1918 polymerase genes have more synonymous differences than expected, suggesting evolutionary distance from known avian strains.

To shed some light on these open questions, we studied a variety of influenza viruses with different origins by thorough phylogeny and codon usage.

RESULTS

RSCU patterns of influenza A viruses

The genetic code is degenerated so that 64 triplets code for only 20 amino acids and a translation stop signal. During translation all 64 triplets are used, including starts and stops. Therefore, for most amino acids there is more than one triplet available, the maximum number being six synonymous codons. Importantly, not all codons are used with equal frequencies. It seems possible to characterize a gene not only by the amino acid sequence that it codes for but also by its codon usage (Taubenberger et al. 2005). Thus, codon usage may represent a genetic tool that can be used to clarify the phylogenetic relationship of related sequences (Zhou et al. 2005). Therefore, studies of synonymous codon usage may reveal information about the molecular evolution of individual genes. We calculated the RSCU according to Sharp et al. (1986) as exemplified by the *PB1* genes of influenza A viruses. Only fully sequenced *PB1* genes of all human H1N1 viruses, including the 1918 virus, H2N2, H3N2, swine H1N1, and avian viruses of all subtypes found in GenBank were analyzed. Figure 1 shows the RSCU pattern of threonine of the *PB1* gene of influenza A viruses with respect to ACA and ACU codons. These codons were chosen because they are most abundant among the four synonymous codons of threonine. According to the results the viruses could be placed into groups, namely into human H1N1, H2N2, and H3N2, classical swine H1N1 isolates, and avian influenza viruses. Interestingly, we found a subset of avian strains that revealed a distinct RSCU pattern from the other avian viruses and the "avian-like"

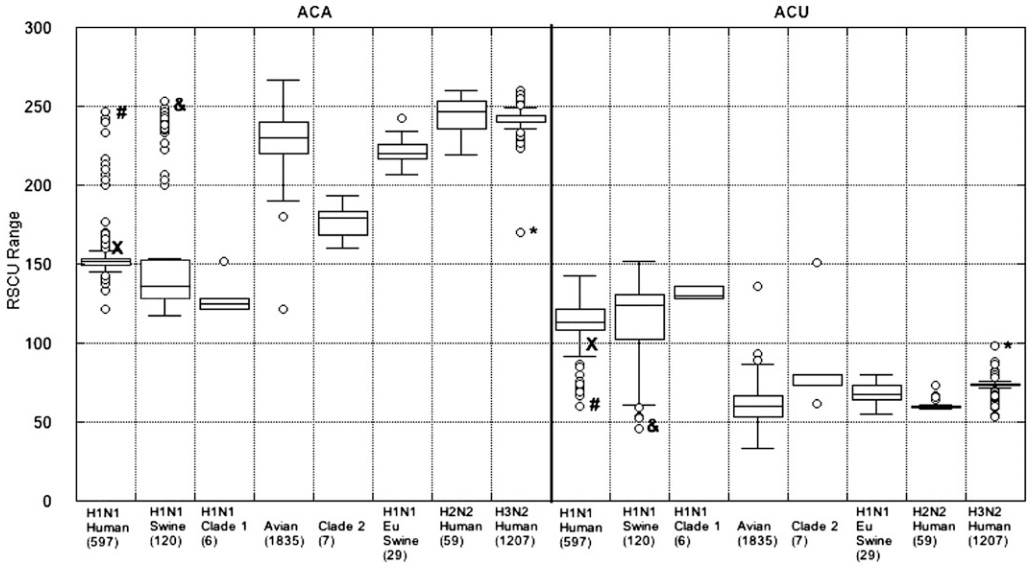

FIGURE 1. RSCU of threonine of the *PB1* gene of influenza A viruses. The pattern is divided into human H1N1, H2N2, and H3N2, "classical" and European (Eu) swine H1N1 isolates, and avian influenza viruses, including clade 1 and 2 in respect to two synonymous codons of threonine: ACA and ACU, respectively. The numbers in the chamber shows estimated *PB1* gene sequences. Only the full open reading frame (ORF) sequenced *PB1* genes of all human H1N1, H2N2, H3N2, swine H1N1, and avian viruses of all subtypes found in GenBank were analyzed. The cross (*x*) indicates the RSCU value of the 1918 virus *PB1* gene. The crosshatch (#) represents human and new swine-origin (SO) 2009 H1N1 reassortants, which had acquired their *PB1* gene from human H3N2 viruses (for review, see Scholtissek 1998; Garten et al. 2009). The ampersand (&) shows outliers of Euroasian swine H1N1 viruses (Smith et al. 2009b) and H1N1 reassortant viruses isolated from pigs in the USA (Vincent et al. 2009) and six swine H1N1 reassortants from North America (Accession numbers [Acc. No]: EU409959, EU692905, EU692906, EU692907, EU409945, and GQ150326), bearing the *PB1* gene of avian origin, respectively (see also, outliers in Fig. 2B). Clade 1 and clade 2 viruses were isolated from birds, but they are outliers of bird viruses. They have a distinct RSCU pattern. Phylogenetically, their *PB1* gene is closely related to human H1N1 viruses (Table 1; Fig. 3). The H1N1 European swine viruses (Eu) are "avian-like" swine viruses (Fig. 3; Schultz et al. 1991). The asterisk (*) presents the A/Victoria/1968 (H3N2), a reassortant strain with a human H1N1 *PB1* gene. The RSCU has been calculated according to Sharp et al. (1986). Each box encloses 50% of the data with the median value of the variable displayed as a line. The *top* and *bottom* of the box mark the data value located halfway between the median and the largest or the smallest data value and they define the limits of ±25% of the variable population. The lines extending from the *top* and *bottom* of each box mark the minimum and maximum values within the data set that fall within an acceptable range. Any value outside of this range, called an outlier, is displayed as an individual point. The outliers are defined as values, which are located outside the box by >150% of its size.

European swine viruses. These subsets called outliers of avian strains were designated clade 1 or clade 2, which are further defined below according to their phylogenetic relationship (for details, see Fig. 3 below).

Overall, the most prominent differences were observed for the ACA codon. The bulk of the human H2N2 and H3N2 viruses that contain a *PB1* gene of avian origin (Scholtissek et al. 1978; Kawaoka et al. 1989) and the avian viruses (which we define as group 2) exhibit higher RSCU values when compared to the human H1N1 viruses and classical swine H1N1 viruses (called group 1) (Table 1). Interestingly, the Brevig-Mission strain from 1918 (x) did not group with the avian strains (group 2) in this analysis (Fig. 1). Instead, this isolate was found to be related to the human and swine H1N1 strains. The outliers observed in the patterns of human H1N1 viruses represent natural reassortants, which had acquired their *PB1* gene from human H3N2 viruses. Interestingly, new swine origin (SO) 2009 H1N1 viruses, which were isolated from humans during the human swine flu outbreak of 2009, also contain the avian-derived *PB1* gene (see Fig. 1, #) (for reviews, see Scholtissek, 1998; Garten et al. 2009). The outliers of swine H1N1 viruses (see Fig. 1, &) are Euroasian "avian-like" swine H1N1 viruses (Smith et al. 2009b) which exhibit a similar RSCU range as "avian-like" swine viruses from Europe (Eu) (Schultz et al. 1991) or H1N1 viruses isolated from pigs in the USA (Vincent et al. 2009). Six of these swine H1N1 viruses are triple reassortant swine viruses (Accession numbers [Acc. No]: EU409959, EU692905, EU692906, EU692907, EU409945, and GQ150326), which obtain the *PB1* gene of avian viruses, respectively (see outliers, Fig. 2B).

The only exception observed in the pattern of the human H3N2 viruses (Fig. 1, asterisks *) is the A/Victoria/68 (H3N2), a reassortant strain with human H3N2 surface genes and the remaining six internal genes (*PB2*, *PB1*, *PA*, *NP*, *M*, and *NS*) from a human H1N1 virus. The clear discrimination in the RSCU patterns of the *PB1* gene was also apparent for the codons of glycine and lysine but not for the other amino acids (data not shown), an observation that will be discussed below. As for the other viral genes that code for internal virus proteins, we did not observe apparent differences in the RSCU patterns for threonine, glycine, and lysine. As an example, the pattern of threonine of the *PB2* gene is shown in Supplemental Figure 1A. Therefore we concentrated our studies mainly on the *PB1* gene in all our further experiments.

TABLE 1. Last base of the four threonine codons of the *PB1* gene of influenza viruses

	Strains	Acc.No.	Subtype	Selected codon positions of threonine																			
				20	21	59	123	132	141	183	196	223	243	291	301	326	400	417	434	435	528	570	662
Group 2	Ck/GermanyN/49	CY014677	H10N7	U	A	A	C	A	U	A	U	A	A	U	A	A	C	A	C	A	A	G	U
	Duck/Czechoslovakia/56	CY005821	H4N6	U	A	G	C	A	U		C	A	A	U	A	A	C	A	C	A	A	G	U
	Duck/Ukraine/1/63	CY005818	H3N8	U	A	G	C	A	U	A	C	A	A	U	A	A	A	A	A	A	A	G	U
	Quail/Italy/1117/65	CY005799	H10N8	U	A	G	C	A	U	A	C	A	A	U	A	A	A	A	A	A	A	G	U
	Duck/NZL/164/76	CY005744	H11N3	C	A	G	C	A	U	A	C	A	A	U	A	A	A	A	A	A	A	U	U
	Shearw./Austr/405/78	CY005664	H3N8	C	G	G	C	A	U	A	C	A	A	C	A	A	A	A	A	A	A	U	U
	Goose/HK/23/78	CY005588	H5N3	U	A	G	C	A	U	A	C	A	G	U	A	A	A	A	C	A	A	G	U
	Pigeon/MN/1407/81	CY005872	H1N1	C	A	A	C	A	U	G	C	A	A	U	C	A	A	A	A	G	A	G	U
	R. Turnstone/NJ/47/85	CY004823	H4N6	C	A	A	C	A	U	A	C	A	A	U	A	A	A	A	A	G	A	G	U
	Herr.Gull/NJ/406/89	CY004975	H5N3	C	A	A	C	A	U	G	C	A	A	U	A	A	A	A	A	A	A	G	U
	Hongkong/156/97*	AF036362	H5N1	C	A	A	C	A	U	A	C	A	A	U	A	A	A	G	C	A	A	U	U
	L. Gull/De/5/03	CY004426	H9N1	C	A		C	A	U	A	C	A	A	U	A	A	A	G	A	A	A	A	U
	Shorebird/DE/122/04	CY005262	H10N7	C	A	A	C	A	U	G	C	C	A	U	A	A	A	A	A	A	A	A	U
	M.duck/NY/21211-5/05	CY029847	H1N1	C	A	A	C	A	U	G	C	A	A	A	C	A	A	A	A	A	A	A	U
	Ann Arbor/6/60	M23972	H2N2	U	A	G	C	A	U	A	C	A	A	U	A	A	A	A	A	A	A	A	U
	England/10/67	AY210021	H2N2	U	A	G	C	C	U	A	C	A	A	U	A	A	A	A	A	A	A	A	U
	Udorn/72	CY009642	H3N2	C	A	A	C	A	U	A	C	A	A	U	A	A	A	G	A	A	A	G	U
	W.Australia/23/02	CY013230	H3N2	C	A	A	U	A	U	A	C	A	A	U	A	A	A	G	A	A	A	G	U
Group 1	Swine/Iowa/1976/31	M55472	H1N1	A	G	U	G	U	A	C	U	C	C	C	C	C	U	U	U		U	C	A
	Brevig Mission/1/18	DQ208310	H1N1	A	G	C	G	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	WSN/33	J02178	H1N1	A	U	C	A	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	Melbourne/35	CY009330	H1N1	A	U	C	A	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	Bel/42	CY009282	H1N1	A	U	U	A	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	Cam/46	CY009602	H1N1	A	U	C	A	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	Malaysia/54	CY009346	H1N1	A	U	C	A	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	Memphis/2/83	CY012894	H1N1	A	U	C	A	U	A	U	U	U	C	C	C	C	U	U	U	U	U	C	A
	Texas/36/91	CY012894	H1N1	A	U	C	A	U	A	U	U	C	C	C	C	C	U	U	U	U	U	C	A
	NY/222/03	CY012894	H1N1	A	U	C	A	U	A	U	U	C	C	C	C	C	U	C	U	U	U	C	A
	Duck/LA17G/87	EU871834	H3N8	C	A	U	U	C	A	U	U	C	U	C	U	C	U	U	C	U	U	C	C
	Turkey/Ontario7732/66	CY015107	H5N9	A	A	U	A	U	A	C	C	C	C	C	C	U	C	U	C	U	C	C	A
	Duck Memphis546/74	CY014692	H11N9	A	A	U	A	U	A	C	U	C	C	C	C	C	C	U	C	U	U	C	A
	Turkey/KS4880/80	EU742642	H1N1	A	G	C	G	U	A	U	U	C	C	C	C	C		U	U		U	U	A
	Turkey/IA/21089-3/92	EU743165	H1N1	A	G	C	G	U	A	U	U	C	C	C	C	C		U	U		U	U	A

Of the 60 threonine positions of the *PB1* gene, 20 were selected in which group 1 strains use a different codon when compared with group 2 strains, with rare exceptions. A blank field indicates an amino acid other than threonine at that position. Because of space limitations, only the most heterogeneous strains were selected since many strains were very similar. Only the third base of the codons was shown, since the first two bases were identical for all four synonymous codons (AC).
Red background color indicates adenine (A), green guanine (G), white cytosine (C), and blue uracil (U), respectively. Codon positions are indicated with residue 1 being the start codon in the open reading frame of the gene segment.
Asterisk (*) indicates this isolate originated from avian sources.

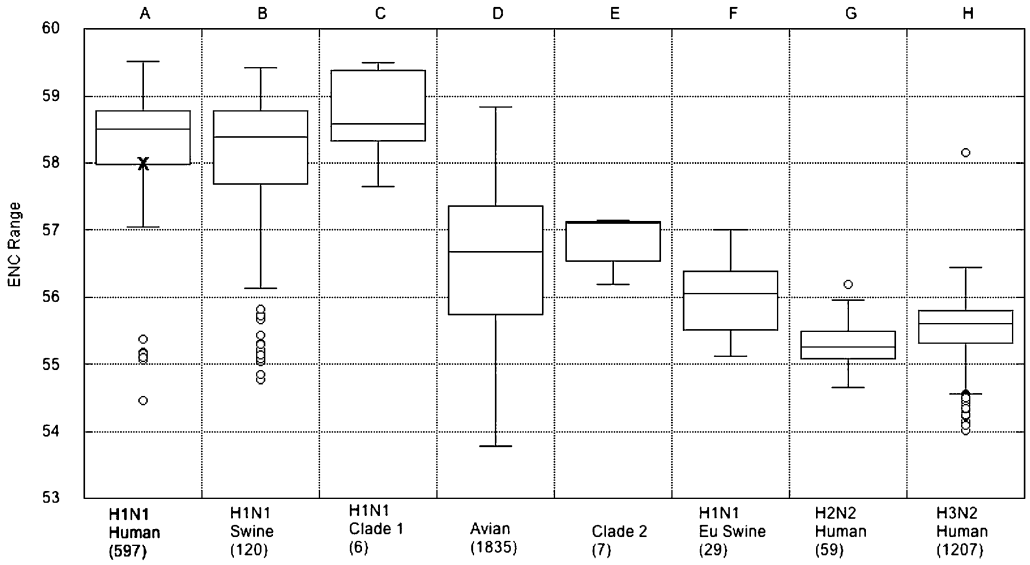

FIGURE 2. The ENC value of the *PB1* gene of influenza A viruses. The cross (*x*) indicates the ENC value of the 1918 *PB1* gene. The bulk of the *PB1* gene of human H2N2 and H3N2 viruses as well as the H1N1 European swine viruses is of avian origin. The exceptional swine or human H1N1 viruses are "avian-like" swine viruses like those of the Euroasian lineage (Fig. 3), and eight human H1N1 reassortants (Accession numbers [Acc. No]: DQ889683, CY019745, CY026417, CY021723, M38376, CY028730, CY021915, and AF342823), which contain an avian *PB1* gene, respectively. The numbers in the chamber shows estimated *PB1* gene sequences with full-length ORF. The ENC value is calculated according to Novembre (2002).

Analysis of synonymous codons along the amino acid sequence of influenza A virus genes

In a more thorough study, mapping the 60 (on average) synonymous codons of threonine along the sequence of the PB1 protein, eight positions were identified at which threonine codons were identical for all strains examined, with minor exceptions. At these positions the codon usage was highly restricted, most likely due to structural constrains of the RNA. Among the remaining 52 positions there were 32 positions where all four codons were equally used, suggesting that there is no bias for a special codon. Interestingly, there were 20 positions where the human and classical swine H1N1 viruses, including the 1918 Brevig-Mission isolate (group 1), collectively used a different codon compared to that found in group 2 viruses, comprising all the avian strains (see Table 1, above, with exceptions shown at the bottom of the table). All fully sequenced *PB1* genes of group 1 and group 2 viruses found in GenBank (Supplemental Table 2) were analyzed and almost all sequences followed this rule (Supplemental Table 1A). Similar patterns were observed with glycine (Supplemental Table 1B), and all amino acids for which more than one codon exists (data not shown). As could be expected, these clear patterns as obtained with amino acids that are coded by four or six codons were not obtained by amino acids that are coded by only two codons. With the latter amino acids the pattern was somewhat less obvious (there are not enough choices).

Interestingly, less clear patterns were observed in the other internal genes, as exemplified for glycine of the *PB2* gene (e.g., Supplemental Table 1C). An exception was the *HA* gene. Here, the human and classical swine H1N1 viruses exhibited a distinct usage pattern of the threonine (ACU) codon, which was different from the avian H1 viruses (Supplemental Fig. 1B). Here again the 1918 Brevig-Mission strain was found with the group 1 viruses.

Among the avian *PB1* genes (Fig. 1) 13 exceptions (clade 1 and 2) from the general rule were identified when analyzed by the method used in Table 1. The data for five representative strains are included at the bottom of Table 1. While these sequences appeared to be outliers in the global RSCU analysis (Fig. 1), the *PB1* gene clearly showed the pattern identified with group 1 viruses.

ENC of the *PB1* gene

We then estimated the absolute usage of synonymous codons of the *PB1* gene by determining the effective number of codons (ENC) (Fig. 2), because this measure is independent of codon number and amino acid compositions. The ENC value varies from 20 (absolutely biased, only one codon is used for each amino acid) to 61 (no bias, all synonymous codons are used equally) (Wright, 1990). This analysis revealed that the *PB1* gene of the 1918 pandemic virus had an ENC value similar to that of classical swine and human H1N1 viruses but not of avian viruses. We also found significant differences with regard to the ENC values of reassortant human H2N2 and H3N2 viruses. The "avian-like" European (Eu) (Schultz et al. 1991) and Asian swine (commonly called Euroasian) H1N1 viruses (Smith

et al. 2009b) bearing the *PB1* gene of avian origin clustered within the lower part of the avian viruses (see outliers in Fig. 2B and in box Fig. 3). These values were significantly lower when compared to the ENC values of the classical swine strains and also of the human H1N1 strains. According to these ENC values of the *PB1* gene, the pandemic virus of 1918 was most likely of swine and not of avian origin.

Phylogenetic analysis of influenza A virus genes

We extended our genetic analysis by constructing phylogenetic trees based on representative sequences from different viral origins, applying the neighbor-joining method (Saitou and Nei 1987; Tamura et al. 2007) and the maximum-likelihood method [Phylogeny Inference Package (PHYLIP), version 3.6]. Both methods gave virtually the same results and consequently only neighbor-joining trees are presented in the paper. The neighbor-joining algorithm (Saitou and Nei 1987; Tamura et al. 2007) provided statistically consistent results. Figure 3 shows a distance-based phylogenetic tree of the *PB1* gene in which the Brevig-Mission strain from 1918 clusters between the classical swine and human H1N1 viruses. The same was found by Taubenberger et al. (2005). The upper part of the tree (marked in box) includes group 2 viruses represented by avian and "avian-like" swine viruses from different regions of the world as well as by human H2N2 and H3N2 viruses, while the lower part includes group 1 viruses represented by the classical swine and human H1N1 viruses. Similar to the RSCU patterns (Fig. 1), we found a subset of avian strains that revealed a distinct evolution pattern from the other avian viruses and the "avian-like" Euroasian swine viruses (Fig. 3). In the phylogenetic tree of *PB1*, they form two separate clades that we designated clade 1 and 2. Clade 2 comprises viral isolates that do not belong to the H1N1 subtype, in contrast to isolates of clade 1. The avian clade 1 H1N1 viruses were isolated predominantly from domestic poultry. This clade also contains many recent H1N1 swine viruses (Figs. 3, 4). The clade was adjacent to the H1N1 classical swine and human virus lineages. Phylogenetic analysis of other influenza virus genes such as *NP* (Fig. 4), *PB2*, *PA*, *HA*, *NA*, *M*, and *NS* genes (Supplemental Figs. 3A-F) supports our finding that the clade 1 viruses form a cluster proximal to classical swine H1N1 and human viruses without reassortment. With respect to *NP*, *PB2*, *PA*, *NA*, *M*, and *NS* genes of the clade 2 viruses they clustered within the bulk of the avian viruses. This indicated that clade 2 viruses were avian reassortants, which had obtained a group 1 *PB1* gene from a group 1 virus. Phylogenetically all genes of the 1918 influenza virus were found to be located between clade 1 viruses and the swine/human lineages.

We have studied also two equine influenza viruses, A/equine/Tennessee/5/1986(H3N8) and A/equine/London/1416/1973(H7N7), with respect to RSCU patterns of threonine and glycine in the *PB1* gene according to Table 1 and Supplemental Table 1B. We found that ~40% of the corresponding positions were as in group 1 viruses, ~40% of the positions as in group 2 viruses, and ~20% of the positions were in none of these two (data not shown). The *PB1* genes of the representative three equine influenza viruses were placed as a separate cluster (or clade) exactly between group 1 and group 2 viruses concerning our RSCU results and the phylogeny analysis (Figs. 3, 4). An identical tree topology of the *PB1* gene of equine isolate A/equine/London/1416/1973(H7N7) was demonstrated by Taubenberger et al. (2005). Since these equine viruses belonged in this respect neither to group 1 nor to group 2, we did not study or discuss equine influenza viruses further.

DISCUSSION

The nature of the direct precursor of the 1918 H1N1 influenza virus that caused the Spanish flu was still enigmatic so far. Here we show by various genetic and phylogenetic analyses that the isolate of 1918 is in all genes closely related to early human and swine H1N1 isolates and an evolutionary distinct subgroup of avian and swine influenza viruses, designated clade 1 viruses in this study. It is interesting that viruses of this distinct clade include avian viruses that still circulated in recent times and stayed genetically stable without reassortment. The phylogenetic data as well as the ENC values presented in Figure 2 are compatible with the hypothesis that a certain swine H1N1 virus had crossed the species barrier from pigs to humans shortly before the pandemic of 1918. This original swine virus might have been of low pathogenicity, causing no severe symptoms and was therefore undetected at that time; however, it may have served as a precursor for the classical swine H1N1 virus lineage. The highly pathogenic swine virus that was detected in the pig population after the human pandemic outbreaks had started (Koen 1919; Chun 1919) is most likely not the precursor of the H1N1 classical swine virus lineage, but their precursors cocirculated during the pandemic (Smith et al. 2009a). According to our study of the ENC values (Fig. 2) the 1918 virus is rather of swine virus origin. The important new finding is that the RSCU signature of the *PB1* gene of the 1918 virus is identical to that of the classical swine and human H1N1 viruses (including clade 1 viruses), and not related to that of any of the contemporary avian viruses, although the overall sequence of the PB1 protein is avian-like. To reconcile these observations at least two interpretations are possible: i) An avian virus entered the swine population a long time ago, adapted to pigs to obtain the new RSCU signature, and was able to shift between domestic birds and pigs since they live in a close neighborhood (Hinshaw et al. 1983; Webster et al. 1992; Ludwig et al. 1994) before it entered the human population around 1918; or, ii) a specific avian

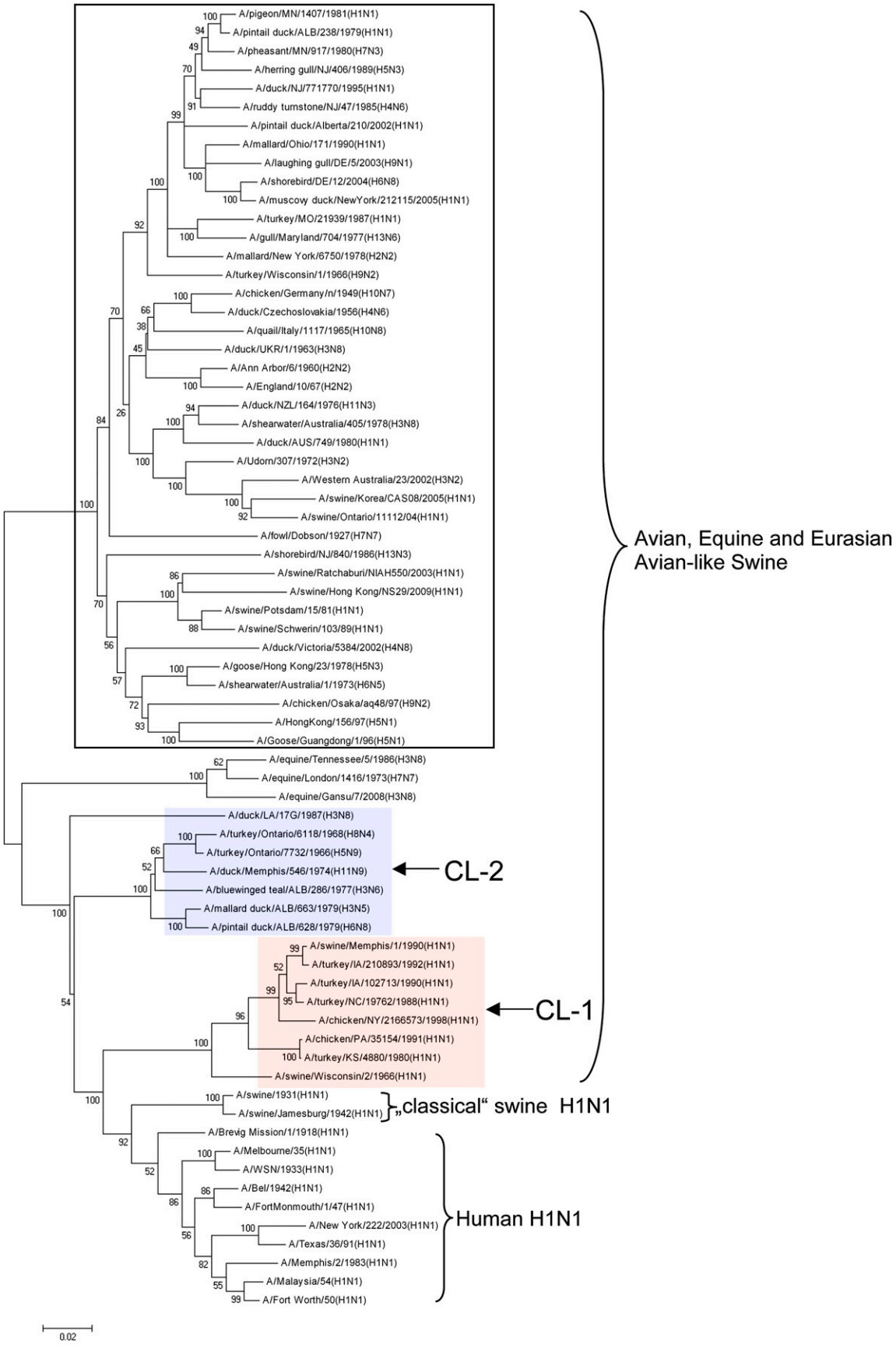

FIGURE 3. Phylogenetic tree of the *PB1* genes of representative influenza viruses. Nucleotide sequences were aligned by using MUSCLE multiple sequence alignment program with default parameters (Edgar 2004) and constructed for the dendrogram using the neighbor-joining method (MEGA 4.0) (Tamura et al. 2007). Bootstrap values were estimated based on 1000 replicates and were given for all presented branches. A distance bar scale was shown under the tree. Clades (CL) are identified with different color backgrounds and with large letters as well as arrows, respectively.

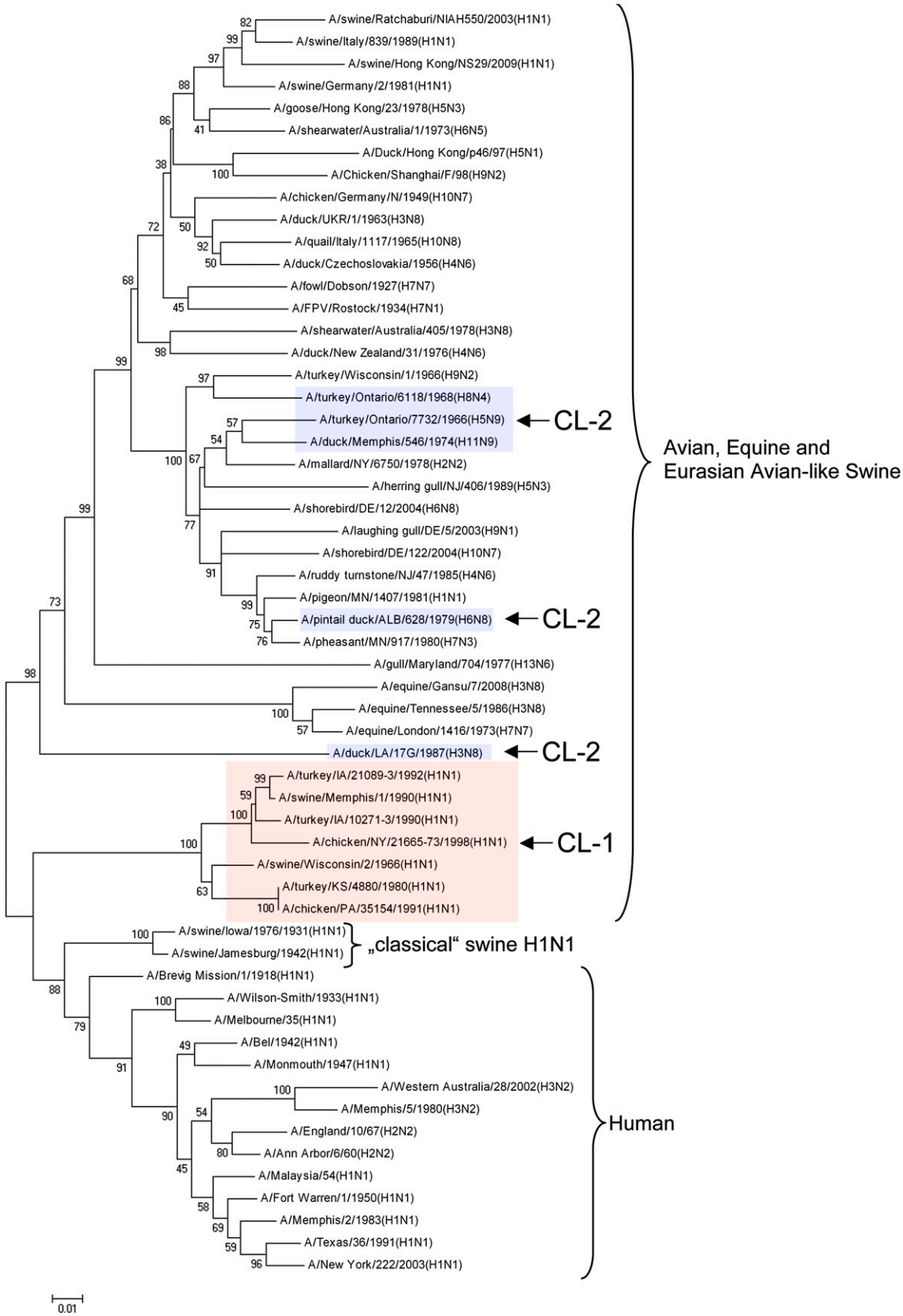

FIGURE 4. Phylogenetic tree of the *NP* genes of representative influenza viruses. Nucleotide sequences were aligned and analyzed as described in the legend for Figure 3.

virus that differed in RSCU signature from the bulk of the other avian viruses was able to enter the pig population easily, possibly a long time ago, and then disappeared as avian virus. From there it may finally have been introduced to the human population shortly before 1918. Regardless which of these interpretations is correct, the data imply that the pig is not only necessary for the creation of reassortants between avian and human influenza viruses, but also for the adaptation of avian influenza viruses to humans. This indicates that the precursor of the pandemic virus of 1918 might have been a swine virus, which remained longer in the pig population than anticipated (Rabadan et al. 2006; Greenbaum et al. 2008). The clade 1 viruses might represent descendants of these avian/swine strains again forming their own clade separate from all the other avian viruses. This splitting-off must have occurred a very long time ago, since the RSCU signature of the *PB1* gene is very different from that of the avian viruses, and it seems to have remained stable over a long time (only a few changes over a period of 70 years). These viruses seem to shift quite easily from pigs to birds. The domestic poultries (e.g., turkeys, chickens, etc.) are usually not a natural (or reservoir) host of either avian or "swine-like" avian influenza A viruses (e.g., in this respect clade 1 H1N1 avian viruses), but susceptible to infection with wild-bird–derived influenza A virus after adaptation (Taubenberger and Kash 2010). According to the phylogenetic trees, the *PB1* genes of the clade 2 viruses have undergone reassortment into the background of other avian viruses presumably from a precursor of the clade 1 viruses a long time ago, and have kept the corresponding RSCU signature almost unchanged. According to Figures 1 and 2 their *PB1* gene seems to be on the way to adapt to an avian host. Our interpretation would imply that the human H1N1 virus lineage started shortly before 1918 by crossing the species barrier from pigs to humans, again with an avian virus as an ultimate precursor.

Our observation of the presence of specific RSCU patterns in the *PB1* gene of influenza viruses was very surprising. The evolutionary mechanisms that may have led to the creation of the specific pattern observed and the reason why this can only be detected so clearly in the *PB1* gene remain unknown. Given these patterns, there is a clear difference in the secondary RNA structure while the amino acid sequence is more or less preserved. Since the codon usage of *Homo sapiens* and *Gallus gallus* is very similar, there was apparently no need for adaptation to a different tRNA content when influenza viruses crossed the avian-human species barrier. Furthermore, the codon usage in the various genes of the same virus is quite different, e.g., the main codon for threonine of the *A*-allel of the *NS* gene is either not used by the *B*-allel at all or is a rare codon, and vice versa (data not shown).

Altered codons may also result in the presence or absence of motifs for RNA methylation. Furthermore, a different secondary structure of mRNA caused by alternate codons results in a different local speed of protein synthesis and thereby causing altered protein folding (Oresic and Shalloway, 1998; Komar et al. 1999; Cortazzo et al. 2002). Assuming that the structural constrains of the *PB1* gene are very tight, only a certain codon might be allowed at the respective position of avian versus human strains. This may also explain the high evolutionary stability of this genetic pattern. Furthermore, selection of RNA segments for virus maturation might also depend on a specific secondary structure of the vRNA (Noda et al. 2006). All in all, the issue of codon usage seems to be much more important at least for influenza viruses than previously thought.

In summary, our data supports the idea that the 1918 Spanish flu influenza virus was derived from a swine virus that itself might be a descendent of a distinct avian H1N1 virus. What we can say for sure is that the 1918 H1N1 virus is not related to one of the known avian influenza strains, except the clade 1 viruses.

MATERIALS AND METHODS

Sequence data

All the sequences used in this study were retrieved from the National Center for Biotechnology Information (NCBI). We used both "Influenza Virus Resource" (Bao et al. 2008; Zaslavsky et al. 2008) and GenBank (Benson et al. 2009). A comprehensive list of all the sequences used is freely available upon request. All accession numbers (Acc. No) of the 3860 sequences of *PB1* gene used in the RSCU analysis and of all eight gene segment sequences in the phylogenetic analyses were listed in the Supplemental Material (Supplemental Table 2). Since not all of these sequences (they are too many) could be used to construct phylogenetic trees, a representative set total of 419 sequences for the genes is used (*PB2*-57, *PB1*-70, *PA*-38, *HA* [only H1N1 subtype] -31, *NP*-57, *NA* [only H1N1 subtype] -49, *M*-49, and *NS*-68 sequences, respectively). These most heterogeneous sequences were selected from different clades of avian isolates, e.g., from gulls, shearwater and shore birds, domestic and feral water and terrestrial fowl, etc., including all strains used for the RSCU analysis in Table 1, as well as the outliers that were used for the phylogenetic studies.

Codon usage analysis

Two methods were applied to estimate codon bias in influenza coding sequences: relative synonymous codon usage (Sharp et al. 1986) and effective number of codons (Novembre 2002).

RSCU is a simple method to calculate deviation from the expected (random) codon distribution that minimizes the bias from the amino acid composition. An RSCU value for a given codon is the observed frequency of that codon divided by the frequency expected under assumption of equal usage of the synonymous codons for a given amino acid and is calculated as follows:

$$RSCU_{ij} = \frac{obs_{ij}}{\frac{1}{n_i}\sum_{j-1}^{n_i} obs_{ij}}$$

where obs_{ij} is the number of occurences of the jth codon for the ith amino acid and n_i is the number of codons for the ith amino acid.

ENC was proposed by Wright (1990) and can be interpreted as the average homozygosity of codons used to code the same amino acid. However, the original Wright's has a major limitation, namely it assumes equal background nucleotide composition. Since influenza genomes have biased nucleotide contribution, we calculated ENC values using Novembre's method. ENC values were calculated as follows:

$$ENC = 2 + \frac{9}{\overline{F}_2} + \frac{1}{\overline{F}_3} + \frac{5}{\overline{F}_4} + \frac{3}{\overline{F}_6}$$

where $\overline{F}_k$ is the average of the $\overline{F}_k$ values for k-fold amino acids (Novembre 2002). The F value represents the average homozygosity for the k-fold degenerate codon group and is calculated according to the following formula:

$$\overline{F} = \frac{X_a^2 + n_a - k}{k(n_a - 1)} \text{ where } X_a^2 = \sum_{i=1}^{k} \frac{n_a(p_i - e_i)^2}{e_i}$$

and p_i is the frequency of of the ith codon, e_i is expected frequency of that codon, and n_a is the observed number of codons for ath amino acid. Please note that the ENC values range from 20 in only one codon is used for each amino acid to 61, when all codons are used equally.

In-house software was developed in Java to calculate and graphically present codon usage bias. The software is available as a web interface at http://www.compgen.uni-muenster.de/tools/sca.

Box plot presentation

Box plots were created using Kaleidagraph software (http://www.synergy.com/). Each box encloses 50% of the data with the median value of the variable displayed as a line. The top and bottom of the box mark the data value located halfway between the median and the largest or the smallest data value and they define the limits of ±25% of the variable population. The lines extending from the top and bottom of each box mark the minimum and maximum values within the data set that fall within an acceptable range. Any value outside of this range, called an outlier, is displayed as an individual point. The outliers are defined as values, which are located outside the box by >150% of its size.

Phylogenetic analysis

Selected viral sequences were aligned using MUSCLE with default parameters (Edgar 2004). Each of the eight genomic segments was anayzed separately. The aligned sequences were subjected to phylogenetic analysis using both distance and maximum likelihood methods. For the distance method, the neighbor-joining algorithm (Saitou and Nei 1987; Tamura et al. 2007) was applied as implemented in MEGA 4.0 software (Tamura et al. 2007). The Tamura-Nei substitution model and maximum composite likelihood method were used to estimate evolutionary distances between sequences and the statistical significance of inferred branches was evaluated by bootstrap using 1000 replicas. The maximum likelihood trees were calculated using the *proml* method from the PHYLIP package with default parameters (Phylogeny Inference Package version 3.6).

SUPPLEMENTAL MATERIAL

Supplemental material can be found at http://www.rnajournal.org.

ACKNOWLEDGMENTS

We would like to thank Dr. R.G. Webster, and J. Franks from the Division of Virology, St. Jude Children's Research Hospital, Memphis, TN, for the helpful discussions and suggestions, and for selecting some of the sequences of the various influenza virus strains. Furthermore, we would like to thank D. Demirov, M. Schmolke (IMV), and G. Abrusan from the Institute of Bioinformatics, University of Münster, Germany, for critical reading and helpful suggestions for the manuscript.

Received July 30, 2010; accepted October 6, 2010.

REFERENCES

Antonovics J, Hood ME, Baker CH. 2006. Molecular virology: Was the 1918 flu avian in origin? *Nature* **440:** E9. doi: 10.1038/nature04824.

Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. 2008. The influenza virus resource at the National Center for Biotechnology Information. *J Virol* **82:** 596–601.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. *Nucleic Acids Res* **37:** D26–D31. doi: 10.1093/nar/gkp1024.

Chun J. 1919. Influenza including its infection among pigs. *Nat Med J China (Peking)* **5:** 34–44.

Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, Deana A. 2002. Silent mutations affect in vivo protein folding in Escherichia coli. *Biochem Biophys Res Commun* **293:** 537–541.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32:** 1792–1797.

Fanning TG, Slemons RD, Reid AH, Janczewski TA, Dean J, Taubenberger JK. 2002. 1917 avian influenza virus sequences suggest that the 1918 pandemic virus did not acquire its hemagglutinin directly from birds. *J Virol* **76:** 7860–7862.

Gammelin M, Altmuller A, Reinhardt U, Mandler J, Harley VR, Hudson PJ, Fitch WM, Scholtissek C. 1990. Phylogenetic analysis of nucleoproteins suggests that human influenza A viruses emerged from a 19th-century avian ancestor. *Mol Biol Evol* **7:** 194–200.

Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, et al. 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325:** 197–201.

Gibbs MJ, Gibbs AJ. 2006. Molecular virology: Was the 1918 pandemic caused by a bird flu? *Nature* **440:** E8. doi: 10.1038/nature04823.

Gorman OT, Bean WJ, Kawaoka Y, Webster RG. 1990. Evolution of the nucleoprotein gene of influenza A virus. *J Virol* **64:** 1487–1497.

Gorman OT, Bean WJ, Kawaoka Y, Donatelli I, Guo YJ, Webster RG. 1991. Evolution of influenza A virus nucleoprotein genes: implications for the origins of H1N1 human and classical swine viruses. *J Virol* **65:** 3704–3714.

Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. 2008. Patterns of evolution and host gene mimicry in influenza and other RNA

viruses. *PLoS Pathog* **4:** e1000079. doi: 10.1371/journal.ppat.1000079.

Hinshaw VS, Webster RG, Bean WJ, Downie J, Senne DA. 1983. Swine influenza-like viruses in turkeys: potential source of virus for humans? *Science* **220:** 206–208.

Kawaoka Y, Krauss S, Webster RG. 1989. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *J Virol* **63:** 4603–4608.

Koen J. 1919. A practical method for field diagnosis of swine diseases. *Am J Vet Med* **14:** 468–470.

Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett* **462:** 387–391.

Ludwig S, Haustein A, Kaleta EF, Scholtissek C. 1994. Recent influenza A (H1N1) infections of pigs and turkeys in northern Europe. *Virology* **202:** 281–286.

Morens DM, Fauci AS. 2007. The 1918 influenza pandemic: insights for the 21st century. *J Infect Dis* **195:** 1018–1028.

Noda T, Sagara H, Yen A, Takada A, Kida H, Cheng RH, Kawaoka Y. 2006. Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature* **439:** 490–492.

Novembre JA. 2002. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19:** 1390–1394.

Oresic M, Shalloway D. 1998. Specific correlations between relative synonymous codon usage and protein secondary structure. *J Mol Biol* **281:** 31–48.

Rabadan R, Levine AJ, Robins H. 2006. Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. *J Virol* **80:** 11887–11891.

Reid AH, Fanning TG, Hultin JV, Taubenberger JK. 1999. Origin and evolution of the 1918 "Spanish" influenza virus hemagglutinin gene. *Proc Natl Acad Sci* **96:** 1651–1656.

Reid AH, Fanning TG, Janczewski TA, Lourens RM, Taubenberger JK. 2004. Novel origin of the 1918 pandemic influenza virus nucleoprotein gene. *J Virol* **78:** 12462–12470.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406–425.

Scholtissek C. 1998. *Genetic reassortment of human influenza viruses in nature*. Blackwell Science, Oxford, England.

Scholtissek C, Rohde W, Von Hoyningen V, Rott R. 1978. On the origin of the human influenza virus subtypes H2N2 and H3N2. *Virology* **87:** 13–20.

Schultz U, Fitch WM, Ludwig S, Mandler J, Scholtissek C. 1991. Evolution of pig influenza viruses. *Virology* **183:** 61–73.

Sharp PM, Tuohy TM, Mosurski KR. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res* **14:** 5125–5143.

Smith GJ, Bahl J, Vijaykrishna D, Zhang J, Poon LL, Chen H, Webster RG, Peiris JS, Guan Y. 2009a. Dating the emergence of pandemic influenza viruses. *Proc Natl Acad Sci* **106:** 11709–11712.

Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, et al. 2009b. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459:** 1122–1125.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24:** 1596–1599.

Taubenberger JK, Kash JC. 2010. Influenza virus evolution, host adaptation, and pandemic formation. *Cell Host Microbe* **7:** 440–451.

Taubenberger JK, Reid AH, Lourens RM, Wang R, Jin G, Fanning TG. 2005. Characterization of the 1918 influenza virus polymerase genes. *Nature* **437:** 889–893.

Taubenberger JK, Hultin JV, Morens DM. 2007. Discovery and characterization of the 1918 pandemic influenza virus in historical context. *Antivir Ther* **12:** 581–591.

Vana G, Westover KM. 2008. Origin of the 1918 Spanish influenza virus: a comparative genomic analysis. *Mol Phylogenet Evol* **47:** 1100–1110.

Vincent AL, Ma W, Lager KM, Gramer MR, Richt JA, Janke BH. 2009. Characterization of a newly emerged genetic cluster of H1N1 and H1N2 swine influenza virus in the United States. *Virus Genes* **39:** 176–185.

Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. 1992. Evolution and ecology of influenza A viruses. *Microbiol Rev* **56:** 152–179.

Wright F. 1990. The 'effective number of codons' used in a gene. *Gene* **87:** 23–29.

Zaslavsky L, Bao Y, Tatusova TA. 2008. Visualization of large influenza virus sequence datasets using adaptively aggregated trees with sampling-based subscale representation. *BMC Bioinformatics* **9:** 237. doi: 10.1186/1471-2105-9-237.

Zhou T, Gu W, Ma J, Sun X, Lu Z. 2005. Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses. *Biosystems* **81:** 77–86.