A comprehensive analysis of reassortment in influenza A virus

U. Chandimal de Silva[1,2,*], Hokuto Tanaka[1,‡], Shota Nakamura[1,3], Naohisa Goto[1,3] and Teruo Yasunaga[1,3]

[1]Department of Genome Informatics and [3]Department of Infection Metagenomics, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamada-oka, Suita City, Osaka 565-0871, Japan
[2]World Premier International Immunology Frontier Research Centre, Osaka University, Osaka 565-0871, Japan
*Author for correspondence (chandi@biken.osaka-u.ac.jp)
‡Present address: Kansai Division, Mitsubishi Space Software Co., Ltd, Hyogo 661-0001, Japan

Biology Open 1, 385–390
doi: 10.1242/bio.2012281

Summary

Genetic reassortment plays a vital role in the evolution of the influenza virus and has historically been linked with the emergence of pandemic strains. Reassortment is believed to occur when a single host - typically swine - is simultaneously infected with multiple influenza strains. The reassorted viral strains with novel gene combinations tend to easily evade the immune system in other host species, satisfying the basic requirements of a virus with pandemic potential. Therefore, it is vital to continuously monitor the genetic content of circulating influenza strains and keep an eye out for new reassortants. We present a new approach to identify reassortants from large data sets of influenza whole genome nucleotide sequences and report the results of the first ever comprehensive search for reassortants of all published influenza A genomic data. 35 of the 52 well supported candidate reassortants we found are reported here for the first time while our analysis method offers new insight that enables us to draw a more detailed picture of the origin of some of the previously reported reassortants. A disproportionately high number (13/52) of the candidate reassortants found were the result of the introduction of novel hemagglutinin and/or neuraminidase genes into a previously circulating virus. The method described in this paper may contribute towards automating the task of routinely searching for reassortants among newly sequenced strains.

Key words: Influenza A, Reassortment, Phylogeny, Matrix, Neighbourhood

Introduction

The influenza A virus has caused substantial morbidity and mortality in humans as well as livestock. Many severe influenza epidemics have been documented in the past 300 years of world history, while four pandemics and numerous other epidemics have occurred in the past century (Kilbourne, 2006; Potter, 2001; Zimmer and Burke, 2009). Furthermore, each year seasonal influenza results in a considerable death toll worldwide. Spanish Flu, the first pandemic of the past century, killed around 50 million people and earned the honorary title “Mother of all pandemics” (Taubenberger and Morens, 2006). The evolutionary origin of the virus responsible for this calamity has always been subject to controversy and is still not fully resolved (Antonovics et al., 2006; Reid and Taubenberger, 2003; Taubenberger, 2006).

It has been long established that genetic reassortment of the eight RNA segments that constitute the influenza A virus may produce novel viruses in nature (Desselberger et al., 1978). This happens when two or more viruses coinfect the same cell and plays a very important role in the long term evolution of the virus as well as in the making of global influenza pandemics. In particular, reassortment events in pigs may give birth to novel viruses with pandemic potential in humans (Castrucci et al., 1993; Ma et al., 2008; Ma et al., 2009). In fact, all known pandemics in recent human history have been attributed as results of reassortment of genes between two or more distinct viruses, with concrete evidence in support of this fact for all except the 1918 Spanish Flu pandemic (Kawaoka et al., 1989; Lindstrom et al., 2004; Scholtissek et al., 1978; Smith et al., 2009a). This was only possible due to the availability of genetic information of contemporary and precursor viral isolates. Furthermore, there is independent multiple evidence that the pandemic H1N1/2009 virus has reassorted again in swine, possibly giving it a chance to escape herd immunity in humans (Ducatez et al., 2011; Vijaykrishna et al., 2010). Isolated human infections of such viruses have also been reported recently (Centers for Disease Control and Prevention (CDC), 2011).

There is little argument about the necessity of continuously monitoring circulating influenza viruses for the possible emergence of new reassortant influenza A viruses. With the number of full genome influenza sequences in publicly available databases growing exponentially in recent years (http://www.ncbi.nlm.nih.gov/genomes/FLU/growth.html), there is little concern about the availability of data as well. Furthermore, with astronomic advances in sequencing technology within the last few years (Schuster, 2008), it will not be long before it becomes not only possible, but also the most reasonable thing to do, to sequence each individual viral genome we would encounter or have access to. Computing infrastructure development would be the need of the hour when such a situation unfolds, as a lot of data would be of no use without ways and means to interpret them efficiently and accurately.

However, very few attempts have been made to automate the identification of reassortment in the influenza virus. A recent paper (Nagarajan and Kingsford, 2011) described an interesting method to identify reassortments which employs a graph mining technique to find topological incongruities between collections of Markov chain Monte Carlo-sampled trees from different segments. It seemed to be technically sound and may, by itself, even prove to be sufficiently robust in handling very large data sets. However, the computational cost of phylogeny reconstruction (MCMC-sampling in this case) is a formidable obstacle towards using phylogeny dependent methods for identifying reassortments from very large data sets. We aim to overcome this problem by formulating a phylogeny independent method which uses only the nucleotide distance matrices as input.

Virologists would usually check for reassortment in their full genome sequences by doing a homology search on a public influenza sequence database or by reconstructing the phylogenies with reference sequences for each of the segments (Holmes et al., 2005; Karasin et al., 2000; Lindstrom et al., 2004). If the most homologous strains do not match across all segments, the possibility of the query sequence belonging to a reassorted virus is considered high. Our algorithm is based on the same principle, but we focus on the 'neighbours' on the phylogenetic tree of each segment rather than doing a homology search. The *neighbourhood* is defined as a fixed number of closest neighbours of a given strain on a given segment phylogeny. The number of common elements in the neighbourhoods of the same taxa on two different segments denotes whether the two segments originated from a single parent or not. A very low common neighborhood size is very often a sign of reassortment.

Materials and Methods

Definition of reassortment

While genetic reassortment is a subject of wide interest in influenza research, it has not been clearly defined so far to our knowledge. We do not plan to give a rigourous mathematical definition here, but a solid image of what amounts to reassortment is imperative before setting out to detect them. Any influenza virus that has at least one pair of segments (out of all possible combinations of two out of eight) such that each segment is clearly derived from one of two distinct parents is what we call a *reassortant* - the direct product of a reassortment event - in this paper. It must be noted that, for all practical purposes, a direct progeny of such a reassortant may not be distinguishable from the original strain under this definition, particularly if the original strain has not been sampled. Thus, it is practically impossible to fully ascertain when or where the reassortment event took place, unless there is sufficient clinical data to establish the viral transmission pathways surrounding the event. Here, we base our results solely on nucleotide sequence data, and refrain from making any distinction between strains with near identical sequences or reassortment patterns.

Overview of the algorithm

For any given flu genome sequence, the only straightforward practical test for reassortment is to find for each segment the 'closest' strains out of a reference set and compare the results across segments. For each segment, the set of r closest strains (including self) is called the *r-neighbourhood* for that segment (Fig. 1). When the r-neighbourhoods of two segments are compared, the common elements in the two sets form the *common neighborhood* and its number of elements is referred to as the *common neighborhood size*. If this size is one, the two segments are deemed to have different ancestry; if it is more than one but comparatively smaller than r, there is still a high chance of different ancestry. We test all combinations of segment pairs and determine if a given strain is a reassortant or not. Highest sensitivity would be achieved by our method when all known complete genomic sequences are included in the reference set. We shall make a data set of all available complete genome sequences, and sequentially test each genome sequence using the rest of the set as the reference.

Preparation of nucleotide sequence data

All available full genome influenza A nucleotide sequences (as at 24th June 2011) were downloaded from the FTP servers of Influenza Virus Resource at NCBI (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html). Duplicate sequences were removed within each segment, and incomplete genomes were further removed to have 9284 sequences for each segment (see supplementary material Table S1 for a complete list). The coding regions for M1, M2, NS1 and NS2 proteins were treated as four individual segments (segment 7–10 respectively) for ease of enumeration. Both coding regions were used for these two segments for added confidence. The coding region of PB1-F2 was left out due to its relatively short length and lack of annotation or presence in some lineages. For each of the 10 segments, multiple alignment was performed by MAFFT and the results were further enhanced by manual editing. Phylogenetic analysis of each major coding region of the eight RNA segments (M2 and NS2 are not included) was performed as a separate study (de Silva et al., 2012).

In order to minimize sampling bias, highly similar genome sequences (>97% in HA and NA; >98% in all other segments simultaneously) were removed, while retaining one genome to represent each class of such genome sequences. This resulted in a final dataset of 1670 representative sequences, on which all further analysis was performed.

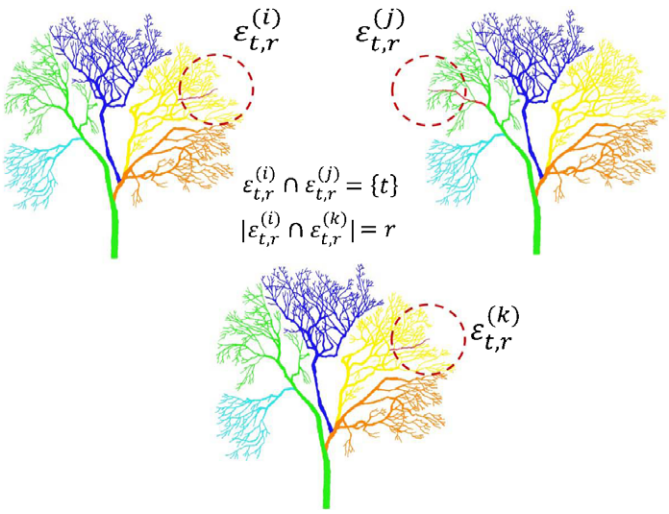

Fig. 1. Neighbourhoods of strain *t* in segments *i*, *j* and *k*. *t* falls on different lineages in segments *i* and *j* and have no common elements in their neighbourhoods except for strain *t* itself. Conversely, *t* has identical roots in segments *i* and *k*, giving rise to identical *r*-neighbourhoods.

Algorithm

The algorithm is described in detail along with definitions of the concepts involved. It is implemented by a Ruby script using Bioruby (Goto et al., 2010), which is available from the authors upon request. All calculations were performed on a Debian Linux server powered by 2 quad-core Intel Xeon X7560 / 2.27 GHz Processors with 264 GB RAM. The run time was approximately 16 hrs for the current data.

Step 0. Calculation of genetic distances

The genetic distances between all pairs of nucleotide sequences are calculated by Phylip dnadist version 3.67 using Jukes Cantor method and the respective distance matrices for each of the segments are consequently used in the analysis as described below.

Step 1. Determination of the common neighbourhoods

For strain t, let its r- neighbourhood in segment $i=(1,..,10)$ be defined as $\varepsilon_{t,r}^{(i)} = \{s \in \Omega | d^{(i)}(t,s) \le d^{(i)}(t,s_{t,r})\}$ where Ω denotes the complete set of viral strains and $d^{(i)}$ is the distance measure on segment i, and $d^{(i)}(t,s_{t,1}) \le d^{(i)}(t,s_{t,2}) \le \cdots \le d^{(i)}(t,s_{t,n})$ where $n=|\Omega|$ and all the $s_{t,*}$ are distinct. In other words, it is the set of r nearest strains to strain t with respect to the genetic distance calculated using the nucleotide sequence data from the i'th segment.

Moreover, for strain t, let $M_t = \left(m_{ij}^t\right)$ be its common neighborhood size matrix defined by

$$m_{ij}^t = \left|\varepsilon_{t,r}^{(i)} \cap \varepsilon_{t,r}^{(j)}\right| \text{ for } i,j = 1,\ldots,10.$$

M_t is calculated for all strains, which in turn is examined for signs of reassortment. m^t_{ij} is the i,j- common neighbourhood size of strain t, which is the number of common elements in its r - neighbourhoods in segments i and j.

Parameter tweaking

The smaller the values of the $m^t_{ij}(1 \le i,j \le 10)$, the stronger the argument for being a reassortant, but where to draw the line is not trivial. We experimented with different numbers and arrived at two separate search criteria, biased more towards sensitivity than specificity. The neighbourhood size r is 80 (approximately 5% of the number of taxa) in what follows.

Step 2. Reassortment search

Let $min_{ti} = \min_{1 \le j \le 10} \left\{ m^t_{ij} \right\}$,

and $max_{ti} = \max_{\substack{1 \le j \le 10 \\ j \ne i}} \left\{ m^t_{ij} \right\}$,

For each strain t, each row of its common neighbourhood matrix is inspected to see if it satisfies one or more of the following conditions.

If $min_{ti} \le 0.1r$, and

1) if $max_{ti} \le 0.1r$, t is a product of single segment reassortment.(S)

2) if $max_{ti} > min_{ti} + 0.8r$, t is a product of multiple segment reassortment.(M)

Here, "single segment reassortment" refers to the event where one segment is uniquely derived from a parent virus different from those of the other segments, while "multiple segment reassortment" refers to the event where more than one segment are derived from a common parent different from those of the other segments. A typical reassortment event may be a mixture of one or several reassortments of these two types.

Results and Discussion

Our common neighbourhood approach easily picked up reassortants whose parent sequences were sufficiently distant and where the reassortment had not become fixed in the population. The common neighbourhood matrix of A/Swine/Italy/1521/98(H1N2) is a good example (Table 1). In segment 4(HA), it has only two elements in common between its segment 4 neighbourhood and most of the other neighbourhoods, which is a good example of a clean reassortment event. Segment 6(NA) is very similar in its relationship with the other segments, with their biggest common neighbourhoods having eleven elements. The rest of the segments have a fairly high number of common 'neighbours' amongst themselves. Altogether, this virus has three parents: two of them contributing one segment each and the last contributing the remaining segments.

A comprehensive search of all available full genome sequence data (9284 strains represented by 1670 genome sequences) resulted in 280 hits (see supplementary material Table S2 for a complete list). These represented a total of 3086 influenza A strains in the original set, with the pandemic H1N1/2009 virus responsible for 2636 of them. Due to space constraints, we chose to show just 52 of them with the highest confidence by limiting to hits with 12 or more very small elements (size ≤2) in their respective common neighbourhood size matrices (Table 2). Some of the reassortants that we found have already been reported, in which case we have shown the reference or commented about the annotation. In some other cases, the sequences have been published in a journal, but the reassortment has not been explicitly declared. Altogether, 35 of the total of 52 reassortants are reported here for the first time, to the best of our knowledge.

A/California/04/09, the reference strain for the pandemic H1N1/2009 virus, was easily picked up by the algorithm notwithstanding the huge sampling bias, while its reassortment pattern (Smith et al., 2009a) was subsequently determined correctly (Table 3).

Of particular interest was an individual segment's propensity to reassort and acquire genetic information from a parent unique to itself or at most common to one more segment. The 7-1(eg:aaabaaaa) and 6-2(eg:aaababaa) reassortment patterns are the most typical of this kind. Segment 4 and 6, which code the HA and NA genes, tend to reassort in this way very often (13/52 instances).

Using our algorithm, we were able to identify further breakdowns in the ancestry of known reassortants. In A/Swine/Ontario/53518/03 (Karasin et al., 2006), for example, we found that PB2 - as well as the previously reported PB1 - was derived from a unique parent of its own. In the 2005 triple reassortant H3N2 viruses from Canada (Olsen et al., 2006), we found that the PB1 gene was of a lineage distinct from that of the other polymerase genes and close to that of HA. Moreover, as a by-product of our analysis, we found that these swine viral sequences were very similar to A/turkey/Ontario/31232/2005(H3N2), a contemporary avian virus from the same region, strongly suggesting cross species transmission. This finding was only possible owing to our comprehensive dataset spanning all hosts.

Similarly, it is quite trivial to find influenza sequences "frozen" in time. A/USSR/90/1977(H1N1), one of the first H1N1 isolates after its re-emergence in 1977 (Zimmer and Burke, 2009) after a 20 year lapse, happened to possess a genomic sequence very similar to that of A/Roma/1949(H1N1).

Furthermore, it proved to be powerful enough to analyse complex reassortment patterns within closely related sequences, when an appropriate data set is used. For example, the predicted

Table 1. Common neighbourhood size matrix of A/Swine/Italy/1521/98(H1N2).

	PB2	PB1	PA	HA	NP	NA	M1	M2	NS1	NS2	min	max	sum
PB2	80	56	55	2	51	10	52	46	42	40	2	56	354
PB1	56	80	57	2	61	10	55	49	50	41	2	61	381
PA	55	57	80	2	48	10	53	49	40	40	2	57	354
HA	2	2	2	80	2	8	8	5	2	2	2	8	33
NP	51	61	48	2	80	7	52	46	45	38	2	61	350
NA	10	10	10	8	7	80	11	11	6	6	6	11	79
M1	52	55	53	8	52	11	80	58	43	41	8	58	373
M2	46	49	49	5	46	11	58	80	38	42	5	58	344
NS1	42	50	40	2	45	6	43	38	80	43	2	50	309
NS2	40	41	40	2	38	6	41	42	43	80	2	43	293

The first ten columns make up the common neighbourhood matrix, where each entry gives the number of common elements in the two neighbourhoods of the given strain from the respective segments. The last three columns give the minimum, maximum, and sum of each row excluding the diagonal entry.

Table 2. List of predicted reassortant strains with strong confidence. All strains with a reference have been reported previously unless otherwise noted. Strains that do not have a reference or a specific remark have not been reported to date, to the best of our knowledge.

Year	Strain	Subtype	#nbds	Type	Ref	Comments
Avian						
1966	A/equine/Lexington/1/1966	H7N7	12	M		
1966	A/turkey/Ontario/6213/1966	mixed	14	M		Genotype:H3,5,6N1
1968	A/turkey/Wisconsin/1/1968	H5N9	12	S		
1976	A/pintail duck/ALB/86/1976	H3N2	18	M	(Obenauer et al., 2006)	Reassortment not mentioned
1976	A/mallard duck/ALB/57/1976	H5N2	16	M	(Obenauer et al., 2006)	Reassortment not mentioned
1979	A/pintail duck/ALB/628/1979	H6N8	18	S	(Obenauer et al., 2006)	Reassortment not mentioned
1982	A/blue-winged teal/ALB/685/1982	H6N4	16	S	(Obenauer et al., 2006)	Reassortment not mentioned
1987	A/turkey/MO/21939/1987	H1N1	18	S		
1988	A/mallard duck/ALB/321/1988	H9N2	12	S	(Hatchette et al., 2004; Obenauer et al., 2006)	Suggests reassortment
1992	A/duck/Nanchang/1681/1992	H3N8	14	M	(Obenauer et al., 2006)	Reassortment not mentioned
1994	A/duck/NY/13152-13/1994	H1N1	16	S		
1996	A/Goose/Guangdong/1/96	H5N1	-	M	(Li et al., 2004; Mukhtar et al., 2007; Xu et al., 1999)	Progenitor of HPAI
1999	A/duck/Shimane/188/1999	H1N1	34	S		
2000	A/mallard/Netherlands/02/2000	H10N7	18	S		
2001	A/duck/Hokkaido/120/2001	H6N2	18	M		
2001	A/chicken/Kobe/aq26/2001	H9N2	16	M	(Mase et al., 2007)	Reassortment not mentioned
2002	A/mallard/Alberta/149/2002	H2N4	14	S	(Hatchette et al., 2004)	Reassortment not mentioned
2002	A/environment/Hong Kong/674.15/2002	H5N1	16	S		
2004	A/chicken/Henan/01/2004	H5N1	14	S		
2005	A/domestic green-winged teal/Hunan/79/2005	H5N1	18	M	(Chen et al., 2009)	
2005	A/teal/Italy/3931-38/2005	H5N2	16	S		
2005	A/common murre/Oregon/19497-004/2005	H9N5	16	M		
2005	A/whooper swan/Mongolia/232/2005	H12N3	20	M	(Spackman et al., 2009)	Reassortment not mentioned
2006	A/mallard/Netherlands/30/2006	H1N4	12	M		
2006	A/pekin duck/California/P30/2006	H4N2	22	S		
2006	A/mallard/Pennsylvania/454069-12/2006	H5N4	14	M		
2006	A/northern shoveler/Netherlands/1/2006	H8N4	12	M		
2006	A/chicken/Pakistan/UDL-01/2006	H9N2	16	M	(Iqbal et al., 2009)	
2006	A/sanderling/Delaware/449/2006	H9N2	16	S		
2006	A/shorebird/Delaware/249/2006	H9N2	12	S		
2007	A/chicken/Laos/P0130/2007	H5N1	16	M	(Boltz et al., 2010)	
2007	A/environment/Hunan/5-32/2007	H5N1	26	M		Annotation implies reassortment
2007	A/little egret/Hong Kong/8550/2007	H5N1	18	M	(Smith et al., 2009b)	
2007	A/garganey/Altai/1213/2007	H5N2	18	S		
2007	A/shorebird/Delaware/554/2007	H9N1	14	S		
2007	A/chicken/Hubei/C1/2007	H9N2	14	M	(Wu et al., 2008)	
2007	A/mallard/Maryland/798/2007	mixed	22	S		Serotype: H9N1; Genotype:H3,5,9N1
2007	A/Eurasian wigeon/Netherlands/4/2007	H10N1	16	M		
2007	A/mallard/Netherlands/17/2007	H11N8	14	M		
2008	A/environment/Hunan/6-69/2008	H5N1	24	M		Annotation implies reassortment
2008	A/peregrine falcon/Hong Kong/2142/2008	H5N1	26	M	(Smith et al., 2009b)	
2008	A/northern shoveler/Interior Alaska/8BM3470/2008	H9N2	12	S		
2009	A/mallard/Hokkaido/24/2009	H5N1	32	M	(Yamamoto et al., 2011)	Suggests reassortment
2009	A/goose/Czech Republic/1848-K9/2009	H7N9	12	S		
Swine						
1998	A/swine/Italy/1521/98	H1N2	12	S	(Marozin et al., 2002)	Reassortment not mentioned
2002	A/swine/Guangdong/102/2002	H3N2	20	M		
2004	A/swine/Guangxi/wz/2004	H5N1	14	M		
2004	A/swine/Guangdong/wxl/2004	H9N2	22	M		
2005	A/swine/Shandong/3/2005	H3N2	20	M		
2008	A/swine/Shandong/1123/2008	H1N1	16	M	(Lu et al., 2010)	
Human						
1976	A/New Jersey/1976	H1N1	12	M	(Gaydos et al., 2006)	1976 Fort Dix flu outbreak strain
Others						
2010	A/canine/Korea/1/2010	H3N1	14	M		Seq. annotated as reassortant

Table 3. Reassortment patterns of selected swine and human strains.

	Strain	Subtype	PB2	PB1	PA	HA	NP	NA	M	NS
Swine										
	A/swine/Italy/1521/98	H1N2	a	a	a	b	a	c	a	a
	A/swine/Guangdong/102/2002	H3N2	a	a	a	a	a	a	b	a
	A/swine/Guangxi/wz/2004	H5N1	a	a	a	a	b	a	a	a
	A/swine/Guangdong/wxl/2004	H9N2	a	a	b	a	a	a	c	a
	A/swine/Shandong/3/2005	H3N2	a	b	b	b	b	b	b	c
	A/swine/Shandong/1123/2008	H1N1	a	a	a	a	a	a	a	b
Human										
	A/New Jersey/1976	H1N1	a	b	a	b	a	b	a	a
	A/California/04/2009	H1N1	a	a	a	a	a	b	b	a

reassortment patterns for Clade B (A/New York/32/2003, A/New York/198/2003 and A/New York/199/2003), Clade C (A/New York/52/2004 and A/New York/59/2004) and A/New York/11/2003 from a comprehensive phylogenetic study of 156 complete genomes of H3N2 influenza A collected between 1999 and 2004 from New York (Karasin et al., 2000) are a perfect match with the patterns that were previously inferred by examining their phylogenies (data not shown).

Sample bias is a major confounding factor in molecular evolutionary analyses, particularly so in our reassortment search. The number of isolates available from the first half of the 20th century is very scarce, making it difficult to determine evolutionary lineages. This is exacerbated by our fixed neighbourhood size of 80, which is too big for sparsely sampled lineages. We actually did not have any hits from that era.

In a preliminary analysis with a smaller data set, the oldest influenza A strain in the database, A/Brevig Mission/1/1918, was picked up by our algorithm, in spite of the fact that, by definition, its ancestry and reassortment history could not be directly determined by the available data. This is a result of our neighbourhoods consisting of both ancestors and descendants, when only ancestors define a given strain's reassortment history. This would potentially pose problems in highly reassortment driven lineages. For example, A/Goose/Guangdong/1/96 (Gs/Gd/1/96), the precursor of the recent HPAI H5N1 lineage in Asia (Li et al., 2004) has passed various combinations of its gene segments to a few generations of multiple reassortants, which did adversely affect our grouping of its own segments by ancestry. Direct descendants could also negatively affect the output when the reassorted genes get fixed in the population. Conversely, such direct descendants of reassortant strains may be wrongly selected as reassortants themselves.

The ideal solution for this problem is to have only ancestors in the neighbourhoods. However, it is not possible to distinguish ancestors from descendants from our distance matrices alone. It would require the construction of all the phylogenies with additional assumptions about the relative rate of evolution on each branch.

Minor topological and distance inconsistencies may occur across segments in phylogenies even without a reassortment event, due to stochastic errors and limitations in distance estimation methods. We need to allow for such minor inconsistencies so that our algorithm does not wrongly pick up strains that are in fact not reassortants. To this end, we must avoid too small a neighbourhood size, thereby allowing movements upto a certain degree to occur without being considered as results of reassortment. Too large a neighbourhood size would, on the other hand, not detect small movements that are actually reassortment driven and may give distorted results when the immediate surroundings are sparsely sampled. After much deliberation, we decided to use a neighbourhood size of approximately 5% of the data set, which seemed to work reasonably well. Perhaps, a neighbourhood size that varies across lineages by sampling density would be a potential improvement to our algorithm.

The property of common ancestry should be transitive over all segments in order to group the segments by ancestry without confusion. (ie. if i and j have common ancestry, and j and k have common ancestry, then i and k should also have common ancestry). Nevertheless, many of our results do not satisfy this criteria, which is no wonder given the fact that we use a common cutoff value for all segment combinations and all lineages. Hence, we have had to assume transitivity in some cases before assigning the ancestry of each segment. (ie. we assume i and k have common ancestry even if the common neighbourhood size falls below the cutoff, provided there exists a segment j that has common ancestry with both of them).

We have tried to reconstruct the phylogenies for our data using MrBayes (Huelsenbeck and Ronquist, 2001) as described in the GiRaF paper (Nagarajan and Kingsford, 2011) and found the computation time till convergence with sufficient mixing to be at least in the order of months per segment on a single processor machine (data not shown). It seems inevitable that we would have to settle for phylogeny independent methods at current processor speeds, when doing comprehensive analyses of influenza genomic data. One such earlier method (Rabadan et al., 2008) seemed to perform well in detecting reassortants within lineages, but no comprehensive study has been undertaken to date using this method.

In this paper, we demonstrate our algorithm using a comprehensive complete genome data set, and strive to find the reassortants within that data set while using the same data for reference. The same algorithm may be used to check any given new influenza A strain with a complete genome sequence for reassortment. If this algorithm is to be used for that purpose, it is imperative that the reference data set is always maintained up to date. We believe this method could be efficiently utilized for rapidly testing high throughput sequence data if the need arises.

Acknowledgements

The authors gratefully acknowledge an open access fee waiver from the publisher for this work. U. C. de S. is thankful for the support from his employer, Immunology Frontier Research

Centre, Osaka University, and N. S. is grateful for funding from the Japan Initiative for Global Research Network on Infectious Diseases (J-GRID).

Competing Interests

The authors declare no competing interests.

References

Antonovics, J., Hood, M. E. and Baker, C. H. (2006). Molecular virology: was the 1918 flu avian in origin? *Nature* **440**, E9; discussion E9-E10.

Boltz, D. A., Douangngeun, B., Phommachanh, P., Sinthasak, S., Mondry, R., Obert, C., Seiler, P., Keating, R., Suzuki, Y., Hiramatsu, H. et al. (2010). Emergence of H5N1 avian influenza viruses with reduced sensitivity to neuraminidase inhibitors and novel reassortants in Lao People's Democratic Republic. *J. Gen. Virol.* **91**, 949-959.

Castrucci, M. R., Donatelli, I., Sidoli, L., Barigazzi, G., Kawaoka, Y. and Webster, R. G. (1993). Genetic reassortment between avian and human influenza A viruses in Italian pigs. *Virology* **193**, 503-506.

Centers for Disease Control and Prevention (CDC). (2011). Swine-Origin Influenza A (H3N2) Virus Infection in Two Children--Indiana and Pennsylvania, July-August 2011. *MMWR Morb. Mortal. Wkly Rep.* **60**, 1213-1215.

Chen, J., Yang, Z., Chen, Q., Liu, X., Fang, F., Chang, H., Li, D. and Chen, Z. (2009). Characterization of H5N1 influenza A viruses isolated from domestic green-winged teal. *Virus Genes* **38**, 66-73.

de Silva, U. C., Tanaka, H., Nakamura, S., Yamashita, A., Goto, N. and Yasunaga, T. (2012). Large scale phylogenetic analysis of influenza A genomes until the 2009 H1N1 pandemic. *Knol* **44**, [available from: http://knol.google.com/k/u-chandimal-de-silva/large-scale-phylogenetic-analysis-of/1n8xmwx3y3wxf/1].

Desselberger, U., Nakajima, K., Alfino, P., Pedersen, F. S., Haseltine, W. A., Hannoun, C. and Palese, P. (1978). Biochemical evidence that "new" influenza virus strains in nature may arise by recombination (reassortment). *Proc. Natl. Acad. Sci. USA* **75**, 3341-3345.

Ducatez, M. F., Hause, B., Stigger-Rosser, E., Darnell, D., Corzo, C., Juleen, K., Simonson, R., Brockwell-Staats, C., Rubrum, A., Wang, D. et al. (2011). Multiple Reassortment between Pandemic (H1N1) 2009 and Endemic Influenza Viruses in Pigs, United States. *Emerg. Infect. Dis.* **17**, 1624-1629.

Gaydos, J. C., Top, F. H., Jr, Hodder, R. A. and Russell, P. K. (2006). Swine influenza a outbreak, Fort Dix, New Jersey, 1976. *Emerg. Infect. Dis.* **12**, 23-28.

Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J. and Katayama, T. (2010). BioRuby: bioinformatics software for the Ruby programming language. *Bioinformatics* **26**, 2617-2619.

Hatchette, T. F., Walker, D., Johnson, C., Baker, A., Pryor, S. P. and Webster, R. G. (2004). Influenza A viruses in feral Canadian ducks: extensive reassortment in nature. *J. Gen. Virol.* **85**, 2327-2337.

Holmes, E. C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B. T., Salzberg, S. L., Fraser, C. M., Lipman, D. J. et al. (2005). Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biol.* **3**, e300.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755.

Iqbal, M., Yaqub, T., Reddy, K. and McCauley, J. W. (2009). Novel genotypes of H9N2 influenza A viruses isolated from poultry in Pakistan containing NS genes similar to highly pathogenic H7N3 and H5N1 viruses. *PLoS ONE* **4**, e5788.

Karasin, A. I., Schutten, M. M., Cooper, L. A., Smith, C. B., Subbarao, K., Anderson, G. A., Carman, S. and Olsen, C. W. (2000). Genetic characterization of H3N2 influenza viruses isolated from pigs in North America, 1977-1999: evidence for wholly human and reassortant virus genotypes. *Virus Res.* **68**, 71-85.

Karasin, A. I., Carman, S. and Olsen, C. W. (2006). Identification of human H1N2 and human-swine reassortant H1N2 and H1N1 influenza A viruses among pigs in Ontario, Canada (2003 to 2005). *J. Clin. Microbiol.* **44**, 1123-1126.

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

Kilbourne, E. D. (2006). Influenza pandemics of the 20th century. *Emerg. Infect. Dis.* **12**, 9-14.

Li, K. S., Guan, Y., Wang, J., Smith, G. J., Xu, K. M., Duan, L., Rahardjo, A. P., Puthavathana, P., Buranathai, C., Nguyen, T. D. et al. (2004). Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia. *Nature* **430**, 209-213.

Lindstrom, S. E., Cox, N. J. and Klimov, A. (2004). Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957-1972: evidence for genetic divergence and multiple reassortment events. *Virology* **328**, 101-119.

Lu, L., Yin, Y., Sun, Z., Gao, L., Gao, G. F., Liu, S., Sun, L. and Liu, W. (2010). Genetic correlation between current circulating H1N1 swine and human influenza viruses. *J. Clin. Virol.* **49**, 186-191.

Ma, W., Kahn, R. E. and Richt, J. A. (2008). The pig as a mixing vessel for influenza viruses: Human and veterinary implications. *J. Mol. Genet. Med.* **3**, 158-166.

Ma, W., Lager, K. M., Vincent, A. L., Janke, B. H., Gramer, M. R. and Richt, J. A. (2009). The Role of Swine in the Generation of Novel Influenza Viruses. *Zoonoses Public Health*. **56**, 326-337.

Marozin, S., Gregory, V., Cameron, K., Bennett, M., Valette, M., Aymard, M., Foni, E., Barigazzi, G., Lin, Y. and Hay, A. (2002). Antigenic and genetic diversity among swine influenza A H1N1 and H1N2 viruses in Europe. *J. Gen. Virol.* **83**, 735-745.

Mase, M., Eto, M., Imai, K., Tsukamoto, K. and Yamaguchi, S. (2007). Characterization of H9N2 influenza A viruses isolated from chicken products imported into Japan from China. *Epidemiol. Infect.* **135**, 386-391.

Mukhtar, M. M., Rasool, S. T., Song, D., Zhu, C., Hao, Q., Zhu, Y., Hao, Q., Zhu, Y. and Wu, J. (2007). Origin of highly pathogenic H5N1 avian influenza virus in China and genetic characterization of donor and recipient viruses. *J. Gen. Virol.* **88**, 3094-3099.

Nagarajan, N. and Kingsford, C. (2011). GiRaF: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Res*. **39**, e34.

Obenauer, J. C., Denson, J., Mehta, P. K., Su, X., Mukatira, S., Finkelstein, D. B., Xu, X., Wang, J., Ma, J., Fan, Y. et al. (2006). Large-scale sequence analysis of avian influenza isolates. *Science* **311**, 1576-1580.

Olsen, C. W., Karasin, A. I., Carman, S., Li, Y., Bastien, N., Ojkic, D., Alves, D., Charbonneau, G., Henning, B. M., Low, D. E. et al. (2006). Triple reassortant H3N2 influenza A viruses, Canada, 2005. *Emerg. Infect. Dis.* **12**, 1132-1135.

Potter, C. W. (2001). A history of influenza. *J. Appl. Microbiol.* **91**, 572-579.

Rabadan, R., Levine, A. J. and Krasnitz, M. (2008). Non-random reassortment in human influenza A viruses. *Influenza Other Respi. Viruses* **2**, 9-22.

Reid, A. H. and Taubenberger, J. K. (2003). The origin of the 1918 pandemic influenza virus: a continuing enigma. *J. Gen. Virol.* **84**, 2285-2292.

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

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nat. Methods* **5**, 16-18.

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S. et al. (2009a). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122-1125.

Smith, G. J., Vijaykrishna, D., Ellis, T. M., Dyrting, K. C., Leung, Y. H., Bahl, J., Wong, C. W., Kai, H., Chow, M. K., Duan, L. et al. (2009b). Characterization of avian influenza viruses A (H5N1) from wild birds, Hong Kong, 2004-2008. *Emerg. Infect. Dis.* **15**, 402-407.

Spackman, E., Swayne, D. E., Gilbert, M., Joly, D. O., Karesh, W. B., Suarez, D. L., Sodnomdarjaa, R., Dulam, P. and Cardona, C. (2009). Characterization of low pathogenicity avian influenza viruses isolated from wild birds in Mongolia 2005 through 2007. *Virol. J.* **6**, 190.

Taubenberger, J. K. (2006). The origin and virulence of the 1918 "Spanish" influenza virus. *Proc. Am. Philos. Soc.* **150**, 86-112.

Taubenberger, J. K. and Morens, D. M. (2006). 1918 Influenza: the mother of all pandemics. *Emerg. Infect. Dis.* **12**, 15-22.

Vijaykrishna, D., Poon, L. L., Zhu, H. C., Ma, S. K., Li, O. T., Cheung, C. L., Smith, G. J., Peiris, J. S. and Guan, Y. (2010). Reassortment of pandemic H1N1/2009 influenza A virus in swine. *Science* **328**, 1529.

Wu, R., Sui, Z. W., Zhang, H. B., Chen, Q. J., Liang, W. W., Yang, K. L., Xiong, Z. L., Liu, Z. W., Chen, Z. and Xu, D. P. (2008). Characterization of a pathogenic H9N2 influenza A virus isolated from central China in 2007. *Arch. Virol.* **153**, 1549-1555.

Xu, X., Subbarao, Cox, J, N. and Guo, Y. (1999). Genetic characterization of the pathogenic influenza A/Goose/Guangdong/1/96 (H5N1) virus: similarity of its hemagglutinin gene to those of H5N1 viruses from the 1997 outbreaks in Hong Kong. *Virology* **261**, 15-19.

Yamamoto, N., Sakoda, Y., Motoshima, M., Yoshino, F., Soda, K., Okamatsu, M. and Kida, H. (2011). Characterization of a non-pathogenic H5N1 influenza virus isolated from a migratory duck flying from Siberia in Hokkaido, Japan, in October 2009. *Virol. J.* **8**, 65.

Zimmer, S. M. and Burke, D. S. (2009). Historical perspective--Emergence of influenza A (H1N1) viruses. *N. Engl. J. Med.* **361**, 279-285.