BMC Medical Genomics

RESEARCH Open Access

g-FLUA2H: a web-based application to study the dynamics of animal-to-human mutation transmission for influenza viruses

Muhammad Farhan Sjaugi[1†], Swan Tan[1†], Hadia Syahirah Abd Raman[1], Wan Ching Lim[1], Nik Elena Nik Mohamed[1], J Thomas August[2], Asif M Khan[1,2,3*]

From Joint 26th Genome Informatics Workshop and Asia Pacific Bioinformatics Network (APBioNet) 14th International Conference on Bioinformatics (GIW/InCoB2015)
Tokyo, Japan. 9-11 September 2015

Abstract

g-FLUA2H is a web-based application focused on the analysis of the dynamics of influenza virus animal-to-human (A2H) mutation transmissions. The application only requires the viral protein sequences from both the animal and human host populations as input datasets. The comparative analyses between the co-aligned sequences of the two viral populations is based on a sliding window approach of size nine for statistical significance and data application to the major histocompatibility complex (MHC) and T-cell receptor (TCR) immune response mechanisms. The sequences at each of the aligned overlapping nonamer positions for the respective virus hosts are classified as four patterns of characteristic diversity motifs, as a basis for quantitative analyses: (i) "index", the most prevalent sequence; (ii) "major" variant, the second most common sequence and the single most prevalent variant of the index, with at least one amino acid mutation; (iii) "minor" variants, multiple different sequences, each with an incidence (percent occurrence) less than that of the major variant; and (iv) "unique" variants, each with only one occurrence in the alignment. The diversity motifs and their incidences at each of the nonamer positions allow evaluation of the mutation transmission dynamics and selectivity of the viral sequences in relation to the animal and the human hosts. g-FLUA2H is facilitated by a grid back-end for parallel processing of large sequence datasets. The web-application is publicly available at http://bioinfo.perdanauniversity.edu.my/g-FLUA2H. It can be used for a detailed characterization of the composition and incidence of mutations present in the proteomes of influenza viruses from animal and human host populations, for a better understanding of host tropism.

Introduction

Influenza has a history as one of the world's most serious pathogens, with yearly regional infections and episodic global pandemics. Influenza viruses belong to the *Orthomyxoviridae* family of RNA viruses and occur as three major types: influenza A, infectious in birds and some mammals, including humans [1]; influenza B, a genus that infects only humans and seals [2]; and influenza C, a rare type that is known to have infected human and pigs [3].

Proliferation of influenza A is predominantly in avian hosts with very rapid mutation, resulting in a "quasispecies" [4], a vast number of viruses that are genetically related but differ in the amino acid sequences of the viral proteins. These mutations are a classic example of Darwinian evolution [5]; they occur in a random fashion and the variant viruses that have the best genetically endowed combination of efficient infection, rapid replication, and greatest survival become the dominant populations. The survival properties include virus escape from the immune responses of humans previously infected or immunized with an earlier virus strain.

Human infections by influenza A viruses commonly occur yearly, with a seasonal peak incidence [6], usually

* Correspondence: asif@perdanauniversity.edu.my
† Contributed equally
[1] Centre for Bioinformatics, Perdana University, Jalan MAEPS Perdana, 43400 Serdang, Selangor, Malaysia
Full list of author information is available at the end of the article

as a mild disease, but for some, as a more severe illness that may be fatal. A major complication is the occurrence of global pandemics resulting from the emergence of highly infectious subtypes of the virus, particularly those capable of human-to-human transmission. The history of influenza pandemics began with the H1N1 "Spanish Flu" strain of 1918-1919 that killed an estimated fifty million people [7]. This was followed by other less severe strains, the H2N2 "Asian influenza" of 1957-1958, H3N2 "Hong Kong flu" of 1968-1969, and H5N1 "bird flu" in 2006-2007, and recently the H1N1 "swine flu" of 2009-2010. It thus appears inevitable that, without means to prevent influenza infection, another pandemic will occur within the foreseeable future. It is simply a matter of chance that the mutations responsible for the infectivity and pathogenicity of a particular influenza virus in animals does not include the ability to efficiently infect humans with human-to-human transmission. If a strain comparable to that of the 1918 pandemic was to occur, the global consequences are inconceivable.

Sequence change (mutation) events can transform an animal-origin virus into a human virus, with varying levels of fitness to survive in the new host [7]. A number of mutations that facilitate the transmission from animal to human hosts have been described for the influenza A viral proteome. For example, specific amino acid substitutions can alter the host receptor binding site and specificity of hemagglutinin (HA) protein, a major determinant of host tropism, for preference from avian to human sialic acid linkages [1]. Additionally, a single amino acid change at position 627 of the viral polymerase complex subunit PB2 was found to enhance replication in human host [8]. There remains, however, the need for a comprehensive identification of sequence changes that allow or enhance human infection, with potential applications to influenza virus surveillance and possibly to prevention or treatment of human infection.

A number of computational models have been reported that discriminate between animal and human influenza A viruses based on molecular patterns in protein sequences [9-12] and genomic signatures [13]. Although these models are potentially useful for predicting interspecies transmission of influenza viruses, there is still a great difficulty in deciphering adaptation of viruses that show mixed signatures of both animal and human hosts. There remains the need for a greater understanding of the viral sequence diversity and the dynamics of sequence change, including the composition and incidence of mutations between the viral host populations.

Herein, we present g-FLUA2H, a web-based application, with a grid-backend, to analyze the dynamics of animal-to-human (A2H) mutation transmission for influenza viruses. The application only requires the viral protein sequences from both the animal and human host populations as input datasets. g-FLUA2H is facilitated by a grid back-end to manage large datasets, as there are currently more than 500,000 sequences of influenza A virus reported in the public Influenza Research Database (IRD; http://www.fludb.org/). g-FLUA2H is publicly available at http://bioinfo.perdanauniversity.edu.my/g-FLUA2H.

System description

Figure 1 provides a schematic workflow summarizing the methodology employed by g-FLUA2H. Input protein sequences, in FASTA format, can be downloaded from the publicly available Influenza Research Database (IRD; http://www.fludb.org) for all worldwide, recorded animal and human influenza viruses. There is a specific format for the FASTA file header that needs to be followed: "> Strain name information | Accession number | Protein name" where the strain name is essential and follows the influenza nomenclature by WHO in the format "Antigenic type/Host/Location of sampling/Isolate ID code/Year of sampling". There is no specific format for accession number and protein name, both of which are optional. An example of a FASTA header format is "> A/blue_winged_teal/Ohio/566/2006 | A7IRT3 | HA". Instructions on how to download sequences from the FluDB in this header format are provided on the help page of g-FLUA2H.

Both full-length and incomplete partial sequences are considered for the analyses. Partial sequences are included because they provide additional data, however, they can be a source of spurious multiple sequence alignment, which require manual editing to correct the misalignment. Duplicate sequences for each protein are removed if present. The non-redundant sequences of each protein from both the animal and human hosts of the virus are then merged for a co-alignment by use of a local copy of MAFFT (obtained from http://mafft.cbrc.jp/alignment/software/); MAFFT is one of the most accurate multiple sequence alignment methods currently available [14]. The co-alignment provides for a unified alignment position, allowing comparison between the two host populations. Users are provided with parameters to disable the removal of redundant sequences and to remove alignment positions with a high fraction of gaps (insertions or deletions), considered to be of low statistical support. Additionally, users are provided with the option to upload manually edited co-alignments as input, in TALN format, if necessary. The sequences of each host population are eventually separated from the co-alignment for a position-by-position comparative analyses.

The comparative analyses is based on a sliding window approach of size nine (1-9, 2-10, 3-11, *etc*) on the separated alignments [15-17]. Each of the aligned

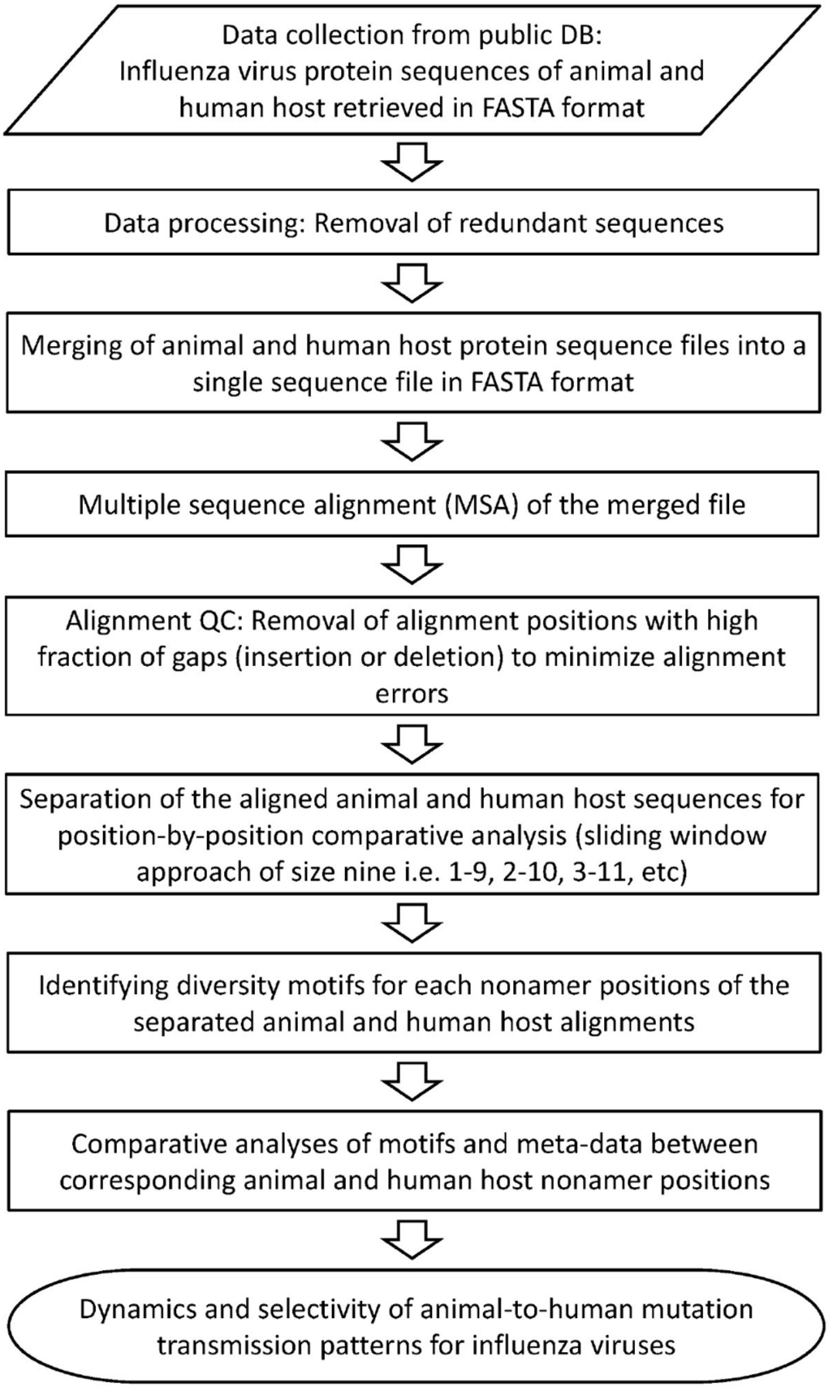

Data collection from public DB:
Influenza virus protein sequences of animal and human host retrieved in FASTA format
Data processing: Removal of redundant sequences
Merging of animal and human host protein sequence files into a single sequence file in FASTA format
Multiple sequence alignment (MSA) of the merged file
Alignment QC: Removal of alignment positions with high fraction of gaps (insertion or deletion) to minimize alignment errors
Separation of the aligned animal and human host sequences for position-by-position comparative analysis (sliding window approach of size nine i.e. 1-9, 2-10, 3-11, etc)
Identifying diversity motifs for each nonamer positions of the separated animal and human host alignments
Comparative analyses of motifs and meta-data between corresponding animal and human host nonamer positions
Dynamics and selectivity of animal-to-human mutation transmission patterns for influenza viruses

Figure 1 Schematic workflow of the methodology employed by g-FLUA2H.

overlapping nonamer (9-mer) positions represent a possible antigenic core binding domain for human leukocyte antigen (HLA; human MHC) molecules and T-cell receptors [16,17]; this assumption is based on the fact that there are large array of HLAs with different binding specificities in the human population [17]. Further, the repeated associations of each amino acid in a moving, overlapping 9-mer window can facilitate the detection of possible sequencing errors. The sequences at each of the 9-mer positions are classified as four patterns of characteristic diversity motifs that we previously defined in Hu, Y. *et al*. [18] and discussed the biological implications.

In brief, the motifs are: (i) "index", the most prevalent sequence; (ii) "major" variant, the second most common sequence and the single most prevalent variant of the index, with at least one amino acid mutation; (iii) "minor" variants, multiple different sequences, each with an incidence (percent occurrence) less than that of the major variant; and (iv) "unique" variants, each observed only once. The quantification of the diversity motifs is detailed in [18] and also described on the g-FLUA2H help page (see Section D). The diversity motifs and their incidences at each of the nonamer positions allow evaluation of the mutation transmission dynamics and selectivity of the sequences in relation to the animal or human hosts. Distinct sequences of any motif that contain gaps (-) or any of the unresolved characters, including B (asparagine or aspartic acid), J (leucine or Isoleucine), X (unspecified or unknown amino acid), and Z (glutamine or glutamic acid) are excluded from all analyses.

The output page of g-FLUA2H provides users with a dropdown menu listing all the nonamer positions to view motif data compared between the animal and human hosts of the virus. The selection of a given nonamer position presents users with a detailed information on the distinct nonamer sequences at the position for both hosts, such as their motif classifications and incidences, amino acid substitutions relative to the index, strain name of origin virus in animal and human hosts, host species of the animal host, and the geographical distribution of the human host. Advanced functionalities include multi-motif analysis and transmission selection where users can analyse for any combination of the diversity motifs and incidences between the hosts, and filter for positive (the incidence of the corresponding motif is greater in human host) or negative selections (incidence of motif is greater in animal host). Snapshots of the key features of g-FLUA2H are provided in Figure 2.

g-FLUA2H was built on top of the gUSE/WS-PGRADE framework [19-23], a grid-backend with multiple CPU cores (several hundreds). This was implemented not as a key distinguishing feature, but for practical purpose to allow ease of handling large datasets, and thus reducing computation time. The g-FLUA2H pipeline employs a combination of parallel compute intensive and distributed compute less-intensive executions (Figure 3).

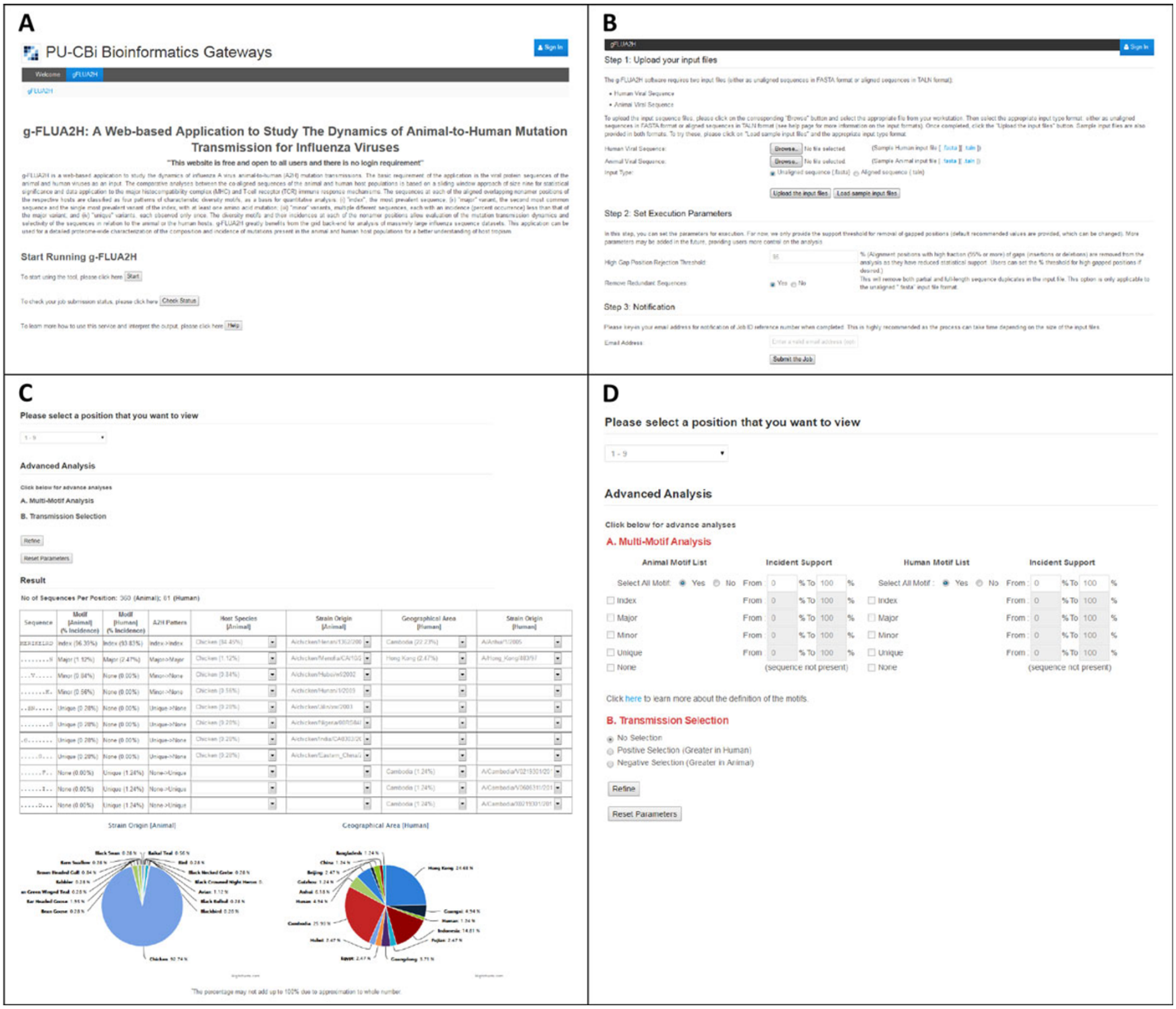
g-FLUA2H Application: animal-to-human mutation patterns of amino acid substitution E627K in influenza A H5N1 and H7N9 subtypes

The amino acid substitution E627K in PB2 of the viral polymerase complex has been reported to be crucial for host tropism. Glutamic acid (E) is found at position 627 in most of the avian strains, whereas replacing the amino acid with lysine (K) enhances viral replication in humans [8,24-28]. The dynamics of this substitution between the animal (avian) and human populations was analyzed and compared by use of g-FLUA2H for H5N1 and H7N9 subtypes, over the time period and the geographical range of the reported data in IRD (Figure 4). There were 2402 (avian: 2130, human: 272) and 127 (avian: 84, human: 43) PB2 sequences of H5N1 and H7N9 reported in IRD as of May 2015 and April 2014, respectively.

g-FLUA2H sample output for a representative nonamer position of H5N1 PB2 that contained this mutation is shown in Figure 4A. At this position (623-631; alignment numbering), there were 21 distinct nonamer sequences representing the reported populations of both avian and human viruses. Both populations contained the same index sequence AAPP**E**QSRM, present in ~75% of the 1380 avian and ~62% of the 208 human virus sequences reported in the public database for this position. Chicken (~24%) was the dominant avian source species for this shared index sequence, followed by duck (~20%, dropdown menu; number not shown in Figure 4A) and many other host species (~< 4% each). The human strains containing this sequence were isolated primarily in Indonesia (~22%) and from many other geographical regions in Asia.

The other dominant sequence AAPP**K**QSRM contained a mutation relative to the index that corresponded to the E627K substitution. This sequence was classified as the major variant of the avian index, present in ~21% of the avian sequences and corresponded to the human major variant, present in ~30% of the human sequences. Chicken (~14%) was also the dominant avian source for this sequence. The human strains containing this sequence were isolated primarily from Vietnam (~9%). Although the E627K substitution has been reported to be important for host tropism, it is clear here that a large majority of H5N1 viruses that infected human did not possess this mutation. Two of the six distinct nonamer sequences of human viruses at this position were not present in the avian viruses, suggesting human specific mutations. Similarly, as many as 15 distinct nonamer sequences appeared to be avian specific, non-existent in the reported human viruses.

In comparison, the corresponding aligned H7N9 PB2 position (623-631) contained only six distinct nonamer sequences representing both the avian and human viruses (Figure 4B), and all but two (A624D; P626R & E627K) matched to H5N1. However, unlike H5N1, the avian and human viruses did not share the index sequence. The avian index sequence AAPP**E**QSRM, present in ~94% of the 80 sequences reported in the public database for this position, was the same as the human major variant, present in ~33% of the 40 human sequences reported for this position.



A

PU-CBi Bioinformatics Gateways

gFLUA2H

g-FLUA2H: A Web-based Application to Study The Dynamics of Animal-to-Human Mutation Transmission for Influenza Viruses

"This website is free and open to all users and there is no login requirement"

g-FLUA2H is a web-based application to study the dynamics of influenza A virus animal-to-human (A2H) mutation transmissions. The basic requirement of the application is the viral protein sequences of the animal and human viruses as an input. The comparative analyses between the co-aligned sequences of the animal and human host populations is based on a sliding window approach of size nine for statistical significance and data application to the major histocompatibility complex (MHC) and T-cell receptor (TCR) immune response mechanisms. The sequences at each of the aligned overlapping nonamer positions of the respective hosts are classified as four patterns of characteristic diversity motifs, as a basis for quantitative analysis: (i) "index", the most prevalent sequence, (ii) "major" variant, the second most common sequence and the single most prevalent variant of the index, with at least one amino acid mutation; (iii) "minor" variants, multiple different sequences, each with an incidence (percent occurrence) less than that of the major variant; and (iv) "unique" variants, each observed only once. The diversity motifs and their incidences at each of the nonamer positions allow evaluation of the mutation transmission dynamics and selectivity of the sequences in relation to the animal or the human hosts. g-FLUA2H greatly benefits from the grid back-end for analysis of massively large influenza sequence datasets. This application can be used for a detailed proteome-wide characterization of the composition and incidence of mutations present in the animal and human host populations for a better understanding of host tropism.

Start Running g-FLUA2H

To start using the tool, please click here Start

To check your job submission status, please click here Check Status

To learn more how to use this service and interpret the output, please click here Help

B

gFLUA2H

Step 1: Upload your input files

The g-FLUA2H software requires two input files (either as unaligned sequences in FASTA format or aligned sequences in TALN format):

- Human Viral Sequence
- Animal Viral Sequence

To upload the input sequence files, please click on the corresponding "Browse" button and select the appropriate file from your workstation. Then select the appropriate input type format, either as unaligned sequences in FASTA format or aligned sequences in TALN format (see help page for more information on the input formats). Once completed, click the "Upload the input files" button. Sample input files are also provided in both formats. To try these, please click on "Load sample input files" and the appropriate input type format.

Human Viral Sequence: Browse... No file selected. (Sample Human input file [.fasta][.taln])
Animal Viral Sequence: Browse... No file selected. (Sample Animal input file [.fasta][.taln])
Input Type: ● Unaligned sequence (.fasta) ○ Aligned sequence (.taln)

Step 2: Set Execution Parameters

In this step, you can set the parameters for execution. For now, we only provide the support threshold for removal of gapped positions (default recommended values are provided, which can be changed). More parameters may be added in the future, providing users more control on the analysis.

High Gap Position Rejection Threshold: 95 % (Alignment positions with high fraction (95% or more) of gaps (insertions or deletions) are removed from the analysis as they have reduced statistical support. Users can set the % threshold for high gapped positions if desired.)

Remove Redundant Sequences: ● Yes ○ No This will remove both partial and full-length sequence duplicates in the input file. This option is only applicable to the unaligned ".fasta" input file format.

Step 3: Notification

Please key-in your email address for notification of Job ID reference number when completed. This is highly recommended as the process can take time depending on the size of the input files.

Email Address: Enter a valid email address (opt

C

Please select a position that you want to view

1 - 9

Advanced Analysis

Click below for advance analyses

A. Multi-Motif Analysis

B. Transmission Selection

Result

No of Sequences Per Position: 360 (Animal); 81 (Human)

Sequence	Motif [Animal] (% Incidence)	Motif [Human] (% Incidence)	A2H Pattern	Host Species [Animal]	Strain Origin [Animal]	Geographical Area [Human]	Strain Origin [Human]
MERIKELRD	Index (96.39%)	Index (93.83%)	Index->Index	Chicken (94.45%)	A/chicken/Henan/1352/200	Cambodia (22.23%)	A/Anhui/1/2005
........N	Major (1.12%)	Major (2.47%)	Major->Major	Chicken (1.12%)	A/chicken/Menofia/CA/10/	Hong Kong (2.47%)	A/Hong_Kong/483/97
...V.....	Minor (0.84%)	None (0.00%)	Minor->None	Chicken (0.84%)	A/chicken/Hubei/wi/2002		
.......K.	Minor (0.56%)	None (0.00%)	Minor->None	Chicken (0.56%)	A/chicken/Hunan/1/2009		
..SN.....	Unique (0.28%)	None (0.00%)	Unique->None	Chicken (0.28%)	A/chicken/Jilin/sw/2003		
........D	Unique (0.28%)	None (0.00%)	Unique->None	Chicken (0.28%)	A/chicken/Nigeria/08RS84/		
.G.......	Unique (0.28%)	None (0.00%)	Unique->None	Chicken (0.28%)	A/chicken/India/CA0303/20		
.....G...	Unique (0.28%)	None (0.00%)	Unique->None	Chicken (0.28%)	A/chicken/Eastern_China/		
......P..	None (0.00%)	Unique (1.24%)	None->Unique			Cambodia (1.24%)	A/Cambodia/V0219301/20
......I..	None (0.00%)	Unique (1.24%)	None->Unique			Cambodia (1.24%)	A/Cambodia/V0606311/201
.....D...	None (0.00%)	Unique (1.24%)	None->Unique			Cambodia (1.24%)	A/Cambodia/X0219301/201

Strain Origin [Animal]

Geographical Area [Human]

*The percentage may not add up to 100% due to approximation to whole number.

D

Please select a position that you want to view

1 - 9

Advanced Analysis

Click below for advance analyses

A. Multi-Motif Analysis

Animal Motif List	Incident Support	Human Motif List	Incident Support
Select All Motif: ● Yes ○ No	From: 0 % To 100 %	Select All Motif: ● Yes ○ No	From: 0 % To 100 %
□ Index	From: 0 % To 100 %	□ Index	From: 0 % To 100 %
□ Major	From: 0 % To 100 %	□ Major	From: 0 % To 100 %
□ Minor	From: 0 % To 100 %	□ Minor	From: 0 % To 100 %
□ Unique	From: 0 % To 100 %	□ Unique	From: 0 % To 100 %
□ None	(sequence not present)	□ None	(sequence not present)

Click here to learn more about the definition of the motifs.

B. Transmission Selection

● No Selection
○ Positive Selection (Greater in Human)
○ Negative Selection (Greater in Animal)

Figure 2 Snapshots of selected g-FLUA2H features. **A.** Homepage interface of g-FLUA2H, which includes a brief description of the application and links to start an analysis, check job status and get more information from the help page. **B**. Web-interface to upload input sequence files, parameters to disable the removal of redundant sequences and to remove alignment positions with a high fraction of gaps (insertions or deletions), and an option to receive email notification when the job submission is complete. **C**. A sample output page, providing users with a dropdown menu listing all the nonamer positions to view motif data compared between the animal and human hosts of the virus. The selection of a given nonamer position presents users with a detailed information on the distinct nonamer sequences at the position for both hosts, such as their motif classifications and incidences (% occurrence), amino acid substitutions relative to the index, strain name of origin virus in animal and human hosts, host species of the animal host, and the geographical distribution of the human host. **D**. Options for advanced analyses on the results, such as multi-motif analysis and transmission selection where users can analyse for any combination of the diversity motifs and incidences between the hosts, and filter for positive (the incidence of the corresponding motif is greater in human host) or negative selections (incidence of motif is greater in animal host).

Chicken (~31%) was the dominant avian source species for this sequence, followed by duck (~16%), wild duck (~14%), and many other host species (< 8% each) (drop-down menu; numbers not shown in Figure 4B). The human virus strains containing this sequence were primarily isolated from various cities in mainland China. Notably, the index sequence of the human viruses (~68%), which contained the E627K substitution (AAPP**K**QSRM), was found in the avian host population solely as a unique sequence (~1%) in the Tree sparrow species. The other four distinct nonamers were unique sequences (~1% each) and appeared to be avian specific, not present in the human viruses. There was no report of human specific mutations for H7N9 at this position, possibly due to the short history of the virus in the human host.

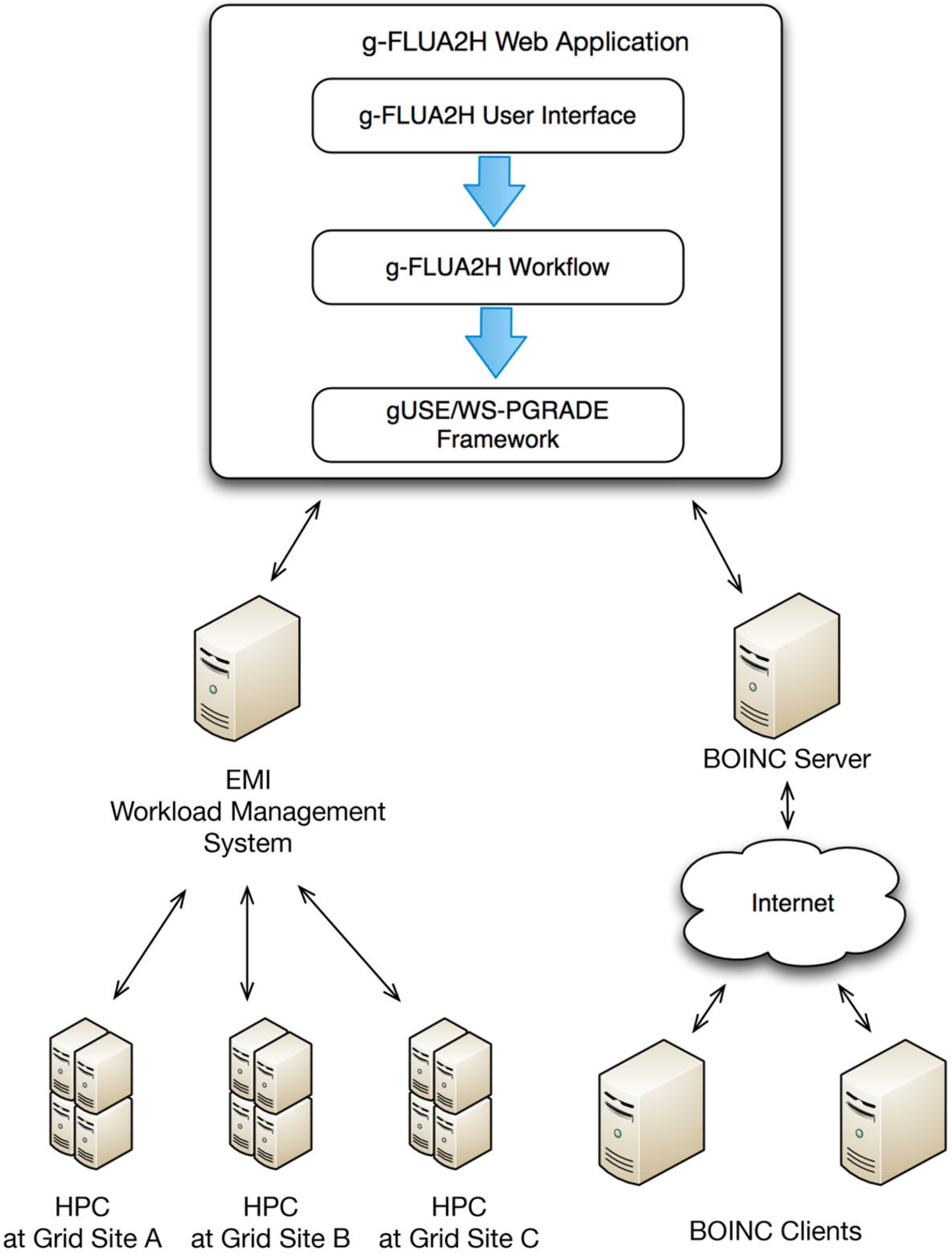
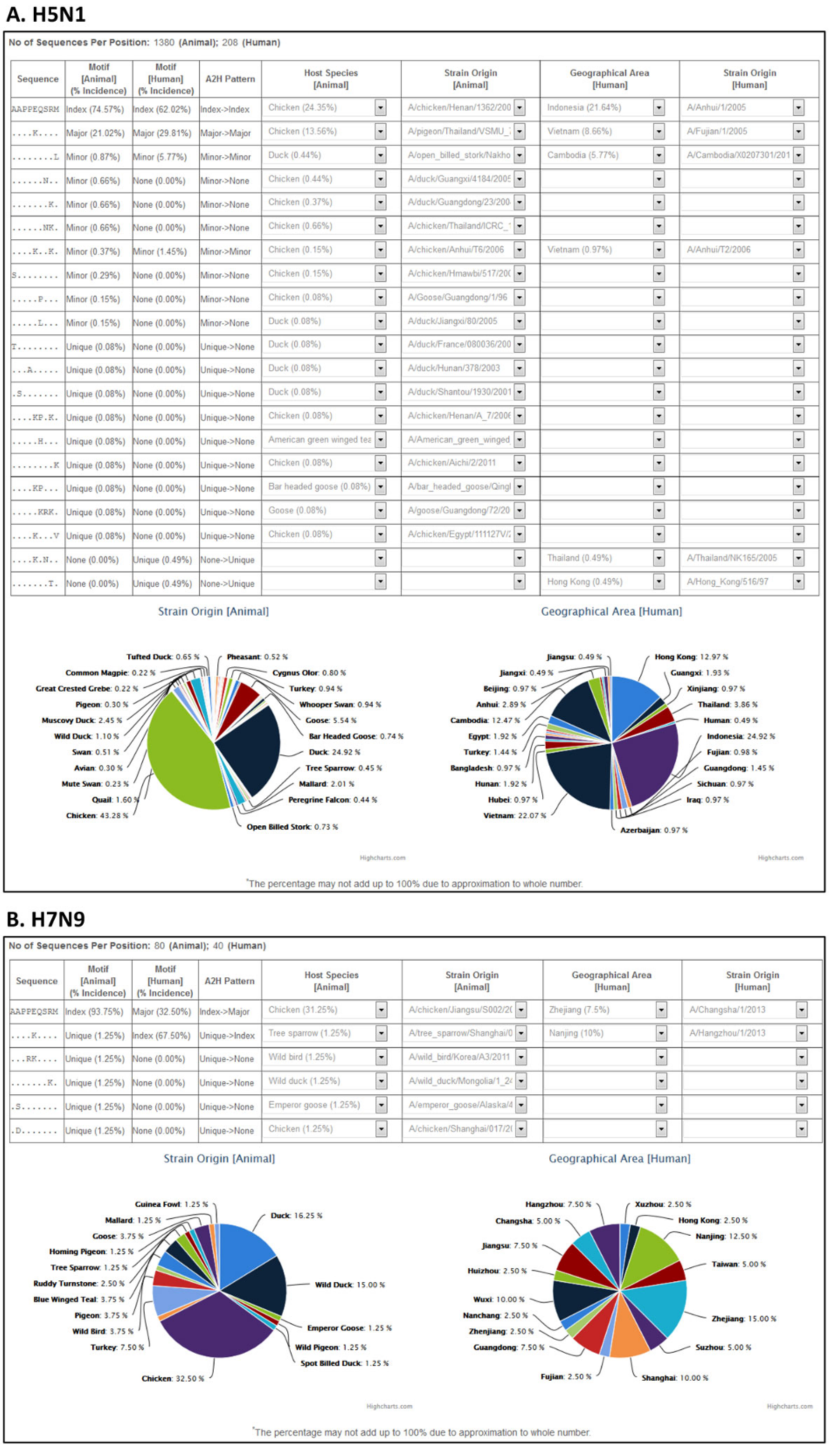


Figure 3 The grid-backend of g-FLUA2H. The user interface of the web application allows input of data which then undergoes a series of steps as part of the data processing and analysis workflow, built on top of the gUSE/WS-PGRADE framework. Depending on the computing load, the framework manages distribution of jobs. Compute intensive executions are handled by sending jobs to a computing cluster, such as HPC/Grid; in this case to the Academic Grid Malaysia Infrastructure (http://academicgrid.my), which is based on the EMI middleware (http://www.eu-emi.eu). Compute less-intensive executions are sent to the distributed Desktop Grid (http://desktopgridfederation.org), based on BOINC setup (https://boinc.berkeley.edu). Abbreviations: HPC, high performance computing; gUSE/WS-PGRADE, Grid and cloud User Support Environment/Web Services Parallel Grid Runtime and Developer Environment; EMI, European Middleware Initiative; BOINC, Berkeley Open Infrastructure for Network Computing.



A. H5N1

No of Sequences Per Position: 1380 (Animal); 208 (Human)

Sequence	Motif [Animal] (% Incidence)	Motif [Human] (% Incidence)	A2H Pattern	Host Species [Animal]	Strain Origin [Animal]	Geographical Area [Human]	Strain Origin [Human]
AAPPEQSRM	Index (74.57%)	Index (62.02%)	Index->Index	Chicken (24.35%)	A/chicken/Henan/1362/200	Indonesia (21.64%)	A/Anhui/1/2005
....K....	Major (21.02%)	Major (29.81%)	Major->Major	Chicken (13.56%)	A/pigeon/Thailand/VSMU_	Vietnam (8.66%)	A/Fujian/1/2005
........L	Minor (0.87%)	Minor (5.77%)	Minor->Minor	Duck (0.44%)	A/open_billed_stork/Nakho	Cambodia (5.77%)	A/Cambodia/X0207301/201
......N..	Minor (0.66%)	None (0.00%)	Minor->None	Chicken (0.44%)	A/duck/Guangxi/4184/2005		
.......K.	Minor (0.66%)	None (0.00%)	Minor->None	Chicken (0.37%)	A/duck/Guangdong/23/200		
......NK.	Minor (0.66%)	None (0.00%)	Minor->None	Chicken (0.66%)	A/chicken/Thailand/ICRC_		
....K..K.	Minor (0.37%)	Minor (1.45%)	Minor->Minor	Chicken (0.15%)	A/chicken/Anhui/T6/2006	Vietnam (0.97%)	A/Anhui/T2/2006
S........	Minor (0.29%)	None (0.00%)	Minor->None	Chicken (0.15%)	A/chicken/Hmawbi/517/20(		
.....P...	Minor (0.15%)	None (0.00%)	Minor->None	Chicken (0.08%)	A/Goose/Guangdong/1/96		
.....L...	Minor (0.15%)	None (0.00%)	Minor->None	Duck (0.08%)	A/duck/Jiangxi/80/2005		
T........	Unique (0.08%)	None (0.00%)	Unique->None	Duck (0.08%)	A/duck/France/080036/200		
...A.....	Unique (0.08%)	None (0.00%)	Unique->None	Duck (0.08%)	A/duck/Hunan/378/2003		
.S.......	Unique (0.08%)	None (0.00%)	Unique->None	Duck (0.08%)	A/duck/Shantou/1930/2001		
....KP.K.	Unique (0.08%)	None (0.00%)	Unique->None	Chicken (0.08%)	A/chicken/Henan/A_7/2006		
.....H...	Unique (0.08%)	None (0.00%)	Unique->None	American green winged tea	A/American_green_winged		
........K	Unique (0.08%)	None (0.00%)	Unique->None	Chicken (0.08%)	A/chicken/Aichi/2/2011		
....KP...	Unique (0.08%)	None (0.00%)	Unique->None	Bar headed goose (0.08%)	A/bar_headed_goose/Qingl		
.....KRK.	Unique (0.08%)	None (0.00%)	Unique->None	Goose (0.08%)	A/goose/Guangdong/72/20		
....K...V	Unique (0.08%)	None (0.00%)	Unique->None	Chicken (0.08%)	A/chicken/Egypt/111127V/		
....K.N..	None (0.00%)	Unique (0.49%)	None->Unique			Thailand (0.49%)	A/Thailand/NK165/2005
.......T.	None (0.00%)	Unique (0.49%)	None->Unique			Hong Kong (0.49%)	A/Hong_Kong/516/97

Strain Origin [Animal]

Geographical Area [Human]

*The percentage may not add up to 100% due to approximation to whole number.

B. H7N9

No of Sequences Per Position: 80 (Animal); 40 (Human)

Sequence	Motif [Animal] (% Incidence)	Motif [Human] (% Incidence)	A2H Pattern	Host Species [Animal]	Strain Origin [Animal]	Geographical Area [Human]	Strain Origin [Human]
AAPPEQSRM	Index (93.75%)	Major (32.50%)	Index->Major	Chicken (31.25%)	A/chicken/Jiangsu/S002/2(	Zhejiang (7.5%)	A/Changsha/1/2013
....K....	Unique (1.25%)	Index (67.50%)	Unique->Index	Tree sparrow (1.25%)	A/tree_sparrow/Shanghai/0	Nanjing (10%)	A/Hangzhou/1/2013
...RK....	Unique (1.25%)	None (0.00%)	Unique->None	Wild bird (1.25%)	A/wild_bird/Korea/A3/2011		
.......K.	Unique (1.25%)	None (0.00%)	Unique->None	Wild duck (1.25%)	A/wild_duck/Mongolia/1_2		
.S.......	Unique (1.25%)	None (0.00%)	Unique->None	Emperor goose (1.25%)	A/emperor_goose/Alaska/4		
.D.......	Unique (1.25%)	None (0.00%)	Unique->None	Chicken (1.25%)	A/chicken/Shanghai/017/2(		

Strain Origin [Animal]

Geographical Area [Human]

*The percentage may not add up to 100% due to approximation to whole number.

Figure 4 g-FLUA2H output for a representative PB2 nonamer position containing the E627K substitution, reported to be crucial for host tropism. The representative position for H5N1 (panel A) and H7N9 (panel B) are the same, 623-631. The output provides detailed information on the distinct nonamer sequences reported at the position for both animal and human hosts of the virus, such as their motif classifications and incidences (% occurrence), amino acid substitutions relative to the index, strain name of origin virus in animal and human hosts, host species of the animal host, and the geographical distribution of the human host. The dropdown menu for each of the distinct nonamer sequences shows the animal species origin, strain names of origin virus in animal and human hosts, and geographical locations of the viral isolate for the human host. The pie-charts show the frequency distribution of the animal host species origin and the geographical location of the human isolates for the nonamer position as a whole, covering all the distinct sequences.

In summary, the nonamer sequence (AAPP**E**QSRM) without the E627K substitution was conserved across both avian H5N1 and H7N9 viruses as the index sequences, with chicken as the dominant avian source species. However, the motif transmission pattern to human differed; it remained as index in human H5N1 viruses, but was a major variant in human H7N9 viruses. The sequence containing the E627K substitution (AAPP**K**QSRM) was a major variant in both the avian and human H5N1 viruses, and did not appear to be required by all human viruses. In contrast, this substitution was only seen once (a unique variant) in the reported avian H7N9 viruses, but was predominantly observed in the human viruses as the index, possibly indicating a role in host tropism. Furthermore, the analysis identified human specific mutations for H5N1, suggesting further adaptation in the human host. These results merit further investigation (Tan, S. *et al.*, manuscript in preparation), and the small number of sequences for the reported human viruses, particularly for H7N9, should be treated as a caveat, with the possibility of sampling bias.

Conclusion

There is a rich metadata of information that can be obtained from a comparative analysis of animal and human host populations of influenza viruses. The data above reveal a complex dynamic of animal to human mutation transmission patterns. g-FLUA2H provides for a detailed proteome-wide characterization of the composition and incidence of mutations present in the animal and human host populations of influenza viruses, for a better understanding of host tropism and possibly identifying human adaptation mutations that are pan-subtype or subtype-specific. This may contribute to the development of tools for surveillance of influenza viruses, and possibly to efforts for the prevention or treatment of viral infection. Given that zoonosis (human disease caused by animal pathogens) is not unique to influenza viruses, g-FLUA2H may be applicable to other zoonotic diseases, which are mostly viral origin and are emerging and reemerging, such as rabies, Ebola, Rift valley fever, and Crimean-Congo hemorrhagic fever.

Competing interests

We declare that we have no competing interests in this research.

Authors' contributions

Supervised the research: JTA MAK. Conceived and designed the web application specifications: MAK MFS ST. Implemented the web application: MFS. Contributed to the content of the web application: MFS MAK ST HSAR WCL. Tested the web-application: MFS MAK ST HSAR WCL NENM. Contributed to the writing of the paper: MAK JTA MFS ST HSAR NENM WCL.

Acknowledgements

We would like to thank Academic Grid Malaysia for providing computing resources for g-FLUA2H execution and also the International Desktop Grid Federation (IDGF) for providing assistance in porting the application to the Desktop Grid platform.

Declaration

Publication costs for this article were funded by Perdana University.
This article has been published as part of *BMC Medical Genomics* Volume 8 Supplement 4, 2015: Joint 26th Genome Informatics Workshop and 14th International Conference on Bioinformatics: Medical genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcmedgenomics/supplements/8/S4.

Authors' details

[1]Centre for Bioinformatics, Perdana University, Jalan MAEPS Perdana, 43400 Serdang, Selangor, Malaysia. [2]Department of Pharmacology and Molecular Sciences, The Johns Hopkins University School of Medicine, 725 North Wolfe Street, 21205 Baltimore, MD, USA. [3]Graduate School of Medicine, Perdana University, Jalan MAEPS Perdana, 43400 Serdang, Selangor, Malaysia.

Published: 9 December 2015

References

1. Bouvier NM, Palese P: **The biology of influenza viruses.** *Vaccine* 2008, **26.**
2. Osterhaus AD: **Influenza B Virus in Seals.** *Science* 2000, 1051-1053.
3. Yuanji G, Fengen J, Ping W: **Isolation of influenza C virus from pigs and experimental infection of pigs with influenza C virus.** *J Gen Virol* 1983, **64**:177-182.
4. Nowak MA: **What is a quasispecies?** *Trends Ecol Evol* 1992, **7**:118-21.
5. Fitch WM, Leiter JM, Li XQ, Palese P: **Positive Darwinian evolution in human influenza A viruses.** *Proc Natl Acad Sci USA* 1991, **88**:4270-4274.
6. Thompson WW, Comanor L, Shay DK: **Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease.** *J Infect Dis* 2006, **194(Suppl 2)**:S82-S91.
7. Morens DM, Fauci AS: **The 1918 influenza pandemic: insights for the 21st century.** *J Infect Dis* 2007, **195**:1018-1028.
8. Steel J, Lowen AC, Mubareka S, Palese P: **Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N.** *PLoS Pathog* 2009, **5.**
9. Miotto O, Heiny A, Tan TW, August JT, Brusic V: **Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis.** *BMC Bioinformatics* 2008, **9(Suppl 1)**:S18.
10. Eng CLP, Tong JC, Tan TW: **Predicting host tropism of influenza A virus proteins using random forest.** *BMC Med Genomics* 2014, **7(Suppl 3)**:S1-S1.
11. Qiang X, Kou Z: **Prediction of interspecies transmission for avian influenza A virus based on a back-propagation neural network.** *Math Comput Model* 2010, **52**:2060-2065.
12. Wang J, Ma C, Kou Z, Zhou Y-H, Liu H-L: **Predicting Transmission of Avian Influenza A Viruses from Avian to Human by Using Informative Physicochemical Properties.** *Int J Data Min Bioinforma* 2013, **7**:166-179.
13. Chen GW, Chang SC, Mok CK, Lo YL, Kung YN, Huang JH, Shih YH, Wang JY, Chiang C, Chen CJ, Shih SR: **Genomic signatures of human versus avian influenza A viruses.** *Emerg Infect Dis* 2006, **12**:1353-1360.
14. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: Improvements in performance and usability.** *Mol Biol Evol* 2013, **30**:772-780.
15. Heiny AT, Miotto O, Srinivasan KN, Khan AM, Zhang GL, Brusic V, Tan TW, August JT: **Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets.** *PLoS One* 2007, **2.**
16. Khan AM, Miotto O, Nascimento EJM, Srinivasan KN, Heiny AT, Zhang GL, Marques ET, Tan TW, Brusic V, Salmon J, August JT: **Conservation and variability of dengue virus proteins: Implications for vaccine design.** *PLoS Negl Trop Dis* 2008, **2.**
17. Khan A, Heiny AT, Lee K, Srinivasan KN, Tan T, August JT, Brusic V: **Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus.** *BMC Bioinformatics* 2006, **7(Suppl 5)**:S4.
18. Hu Y, Tan PT, Tan TW, August JT, Khan AM: **Dissecting the Dynamics of HIV-1 Protein Sequence Diversity.** *PLoS One* 2013, **8.**
19. Balasko A, Farkas Z, Kacsuk P: **Building Science Gateways by Utilizing The Generic WS-PGRADE/GUSE Workflow System.** *Comput Sci* 2013, **14**:2.
20. Kacsuk P, Farkas Z, Kozlovszky M, Hermann G, Balasko A, Karoczkai K, Marton I: **WS-PGRADE/gUSE generic DCI gateway framework for a large variety of user communities.** *J Grid Comput* 2012, **10**:601-630.

21. Kacsuk P: **P-GRADE portal family for grid infrastructures.** *Concurr Comput Pract Exp* 2011, **23**:235-245.
22. Kovács J, Kacsuk P, Lomaka A: **Using a private desktop grid system for accelerating drug discovery.** *Futur Gener Comput Syst* 2011, **27**:657-666.
23. Kacsuk P, Farkas Z, Hermann G: **Workflow-Level Parameter Study Support for Production Grids.** In *Computational Science and Its Applications – ICCSA 2007 SE - 74. Volume 4707.* Springer Berlin Heidelberg;Gervasi O, Gavrilova M 2007:872-885, Lecture Notes in Computer Science.
24. Cauldwell A V, Long JS, Moncorgé O, Barclay WS: **Viral determinants of influenza A virus host range.** *J Gen Virol* 2014, **95(Pt 6)**:1193-210.
25. Chen H, Bright RA, Subbarao K, Smith C, Cox NJ, Katz JM, Matsuoka Y: **Polygenic virulence factors involved in pathogenesis of 1997 Hong Kong H5N1 influenza viruses in mice.** *Virus Res* 2007, **128**:159-63.
26. Hatta M, Gao P, Halfmann P, Kawaoka Y: **Molecular basis for high virulence of Hong Kong H5N1 influenza A viruses.** *Science* 2001, **293**:1840-1842.
27. Shinya K, Hamm S, Hatta M, Ito H, Ito T, Kawaoka Y: **PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice.** *Virology* 2004, **320**:258-66.
28. Gao R, Cao B, Hu Y, Feng Z, Wang D, Hu W, Chen J, Jie Z, Qiu H, Xu K, Xu X, Lu H, Zhu W, Gao Z, Xiang N, Shen Y, He Z, Gu Y, Zhang Z, Yang Y, Zhao X, Zhou L, Li XX, Zou S, Zhang YY, Yang L, Guo J, Dong J, Li Q, Dong L, Zhu Y, Bai T, Wang S, Hao P, Yang W, Han J, Yu H, Li D, Gao GF, Wu G, Wang Y, Yuan Z, Shu Y: **Human infection with a novel avian-origin influenza A (H7N9) virus.** *The New England journal of medicine* 2013.

doi:10.1186/1755-8794-8-S4-S5
Cite this article as: Sjaugi *et al.*: **g-FLUA2H: a web-based application to study the dynamics of animal-to-human mutation transmission for influenza viruses.** *BMC Medical Genomics* 2015 **8**(Suppl 4):S5.