WILEY-BLACKWELL

Cladistics 27 (2011) 61–66

Cladistics

10.1111/j.1096-0031.2010.00314.x

The Supramap project: linking pathogen genomes with geography to fight emergent infectious diseases

Daniel A. Janies[a,*], Travis Treseder[a], Boyan Alexandrov[a], Farhat Habib[b], Jennifer J. Chen[a], Renato Ferreira[c], Ümit Çatalyürek[a], Andrés Varón[d,e] and Ward C. Wheeler[d]

[a]*Department of Biomedical Informatics, The Ohio State University, College of Medicine, Columbus, OH 43210, USA;* [b]*Indian Institute of Science Education and Research (IISER) Garware Circle, Sutarwadi, Pashan Pune, Maharashtra 411021, India;* [c]*Universidade Federal de Minas Gerais, Departamento de Ciência da Computação, Belo Horizonte, MG, Brazil ;*[d]*Division of Invertebrate Zoology, The American Museum of Natural History, New York, NY 10024, USA;* [e]*Computer Science Department, The Graduate Center, The City University of New York, New York, NY 10016, USA*

Accepted 23 February 2010

Abstract

Novel pathogens have the potential to become critical issues of national security, public health and economic welfare. As demonstrated by the response to Severe Acute Respiratory Syndrome (SARS) and influenza, genomic sequencing has become an important method for diagnosing agents of infectious disease. Despite the value of genomic sequences in characterizing novel pathogens, raw data on their own do not provide the information needed by public health officials and researchers. One must integrate knowledge of the genomes of pathogens with host biology and geography to understand the etiology of epidemics. To these ends, we have created an application called Supramap (http://supramap.osu.edu) to put information on the spread of pathogens and key mutations across time, space and various hosts into a geographic information system (GIS). To build this application, we created a web service for integrated sequence alignment and phylogenetic analysis as well as methods to describe the tree, mutations, and host shifts in Keyhole Markup Language (KML). We apply the application to 239 sequences of the polymerase basic 2 (PB2) gene of recent isolates of avian influenza (H5N1). We map a mutation, glutamic acid to lysine at position 627 in the PB2 protein (E627K), in H5N1 influenza that allows for increased replication of the virus in mammals. We use a statistical test to support the hypothesis of a correlation of E627K mutations with avian-mammalian host shifts but reject the hypothesis that lineages with E627K are moving westward. Data, instructions for use, and visualizations are included as supplemental materials at: http://supramap.osu.edu/sm/supramap/publications.

We have created a web-based workflow application, Supramap (http://supramap.osu.edu). Using a web browser, a user inputs text files containing sequence and or phenotypic data, latitude and longitude coordinates, and (optionally) a date of isolation for each strain. Our application then executes a workflow that entails integrated sequence alignment and phylogenetic analysis, computation of character changes (e.g., mutations and host shifts), and geographical projection of the tree on a computing cluster. Once the analyses are complete, the user can download a phylogenetic layer expressed in KML file and view the file with a Geographic Information System (GIS). The user can use the phylogenetic layer to visualize several aspects of pathogen evolution including: spread of lineages, mutations, shifts among hosts, and phenotypic changes over geography and time. We illustrate the use of the system with a case study on H5N1 and discuss use of visualization in conjunction with statistical validation.

Other tree projection efforts

Supramap is superficially similar to other efforts for projecting phylogenetic trees in GIS, such as

*Corresponding author:
E-mail address: Daniel.Janies@osumc.edu

MigraPhyla (HoDac et al., 2007) and Geophylobuilder (Kidd and Liu, 2008). There are also papers that use GIS to study the spread of pathogens but do not provide accompanying applications (Lemey et al., 2009). Some have released applications that project trees but have not published research case studies (e.g., Piel, 2007; Maddison and Maddison, 2009). GenGIS combines data visualizations, including trees and gene frequency data, with R scripts (Parks et al., 2009).

In contrast to these efforts, the Supramap web application provides a complete workflow including parallel direct optimization of raw data into trees and implied alignments, tree projection into a Keyhole Markup Language (KML) file, and character diagnosis with presentation of mutation or other character change data implied by the phylogeny for each node of the tree. The diagnosis functionality is similar to a user performing character mapping on a traditional cladogram using the ''Trace Character History'' feature in Mesquite or ''Map Characters'' feature in TNT (Goloboff et al., 2008). In the case of Supramap, character changes are optimized on the projected tree. Thus, Supramap enables character evolution studies in biogeographic and temporal contexts. Below, we include a case study for character evolution in hypothesis driven research on the spread of key mutations in lineages of infectious diseases (Janies et al., 2007, 2008; Hill et al., 2009).

A client, web service and computing power

Supramap is distinct from all the efforts described above in that it is both a client and a web service that is instantiated on a computing cluster freely available for use by the scientific community. In this respect, Supramap resembles the RAxML web-server (Stamatakis et al., 2008; http://phylobench.vital-it.ch/raxml-bb/) or SWAMI (Rifaieh et al., 2007; http://www.phylo.org/portal2) in that a phylogenetic search application is spawned via a web interface to a computing cluster. RAxML requires the user to upload a precomputed alignment. In contrast, Supramap does not require an alignment (although it can use pre-aligned data if the user prefers). A typical Supramap run performs integrated alignment and phylogenetic analysis as implemented in POY (Varón et al., 2009). SWAMI supports POY but not the geographical projection capabilities of Supramap.

Pathogen-centric disease surveillance

Many services for syndromic surveillance of infectious diseases are available. These services scrape, filter, and map news or search data from the Internet on the occurrence of disease symptoms (Brownstein et al., 2008; Ginsberg et al., 2008). Supramap complements these services. Syndromic surveillance is patient-centric whereas Supramap is pathogen-centric. As Supramap focuses on genomic and geographic data, it allows the user to understand the key mutations and pathogen phenotypes as diseases spread over time, geography, and among hosts, including animals and humans.

Like syndromic data, pathogen genomic data are a key source of medical intelligence that aid public health officials in controlling an outbreak of disease. For example, in the early days of the SARS epidemic, the cases were reported as atypical pneumonia, which often indicates a bacterial agent. Only through serological, microscopic and nucleotide sequence characterization was it understood that SARS was caused by a previously undiscovered coronavirus (CoV) (Ksiazek et al., 2003). Furthermore, it was not until the full genome of SARS-CoV was sequenced, shared and compared via phylogenetics to related coronaviruses from a variety of animals that the zoonotic origins of SARS-CoV could be put in the evolutionary and geographic context necessary for disease surveillance (Guan et al., 2003; Marra et al., 2003; Rota et al., 2003; Janies et al., 2008).

In this report, we focus on the implementation and use of Supramap to track the evolution of phenotypes and mutations in sequence data derived from H5N1 influenza over time, hosts, and geography. Supramap has been used to visualize the spread of pathogens carrying key mutations that confer the ability of the pathogen to replicate in novel host species or to resist drugs (Janies et al., 2007, 2008; Hill et al., 2009). Studies in natural history and anthropology are also underway with Supramap.

Methods

Architecture

The Supramap server consists of two main elements. The first is a web service created via a JBOSS application server (http://www.jboss.org). The web service executes POY processes and generates KML files on the cluster. The second is a client written in Ruby using the Ruby on Rails® framework (http://www.rubyonrails.org). A Phusion Passenger™ (http://www.modrails.com) server supports the client. Both elements use MySQL® (http://www.mysql.com) as the data store.

User interface

We have created a web interface where users can upload data files, name projects, and organize sets of data files into jobs to be executed. The user may study one or many loci and set up various types of analyses by choosing from uploaded files. We post a detailed manual: http://supramap.osu.edu/sm/supramap/tutorials. Once the user starts a job, the Supramap

system will perform a phylogenetic analysis using dynamic homology (integrated alignment and tree search; Wheeler et al., 2006), generate a tree and KML file, and present statistics on the run. Statistics include the optimal tree cost found during the search and the number of times this tree cost was hit. If the user prefers, multiple sequence alignment and or tree search can be pre-computed locally with edit costs, search heuristics and optimality criterion specified by the user. Then the user can upload aligned data and or a tree and use the Supramap web service for character diagnosis and tree projection resulting in a KML file.

Phylogenetic analyses

Phylogenetic analyses using dynamic homology are performed in parallel on a cluster made of personal computer components running LINUX. POY uses MPI v1.0 for parallel execution (http://www.mpi-forum.org/docs/mpi-11-html/mpi-report.html). At present, the default script for POY in the Supramap web service includes a simple set of key commands (discussed below). However, the user is free to vary the commands and their arguments in their own local runs of POY. A stand-alone binary of POY can be enabled for Supramap by compiling from source code with a plug in (http://supramap.osu.edu/sm/supramap/tutorials#section2).

Key commands include:

`transform (tcm:(1,1))` These commands set the edit costs for ancestor-descendent changes in nucleotide bases.

`search(max_time:0:0:3, memory:gb:2)` These commands implement numerous tree search heuristics including Wagner building, branch swapping, tree fusing, and ratcheting (Varón et al., 2009). These commands set a stopping rule at a wall clock time limit (e.g., 3 min) and amount of memory (e.g., 2 gigabytes). POY will use the computing resources allocated to it within the limits set by the search commands to find the best heuristic length tree.

`select(best:1)` If more than one tree are found at heuristic minimum length, this command randomly chooses a tree from the pool of trees. In some cases, many heuristic minimum length trees are implied by the data. In the web service, we select one best tree at random for display using this command. If the user compiles their own binary of POY from source, then the user can explore results implied by multiple trees. A more thorough exploration of the use of multiple trees in phylogeography is available in Hovmöller et al., (2010) (http://routemap.osu.edu).

`transform(static_approx)` Using these commands, dynamic homology characters are transformed into static homology characters akin to a multiple alignment (Wheeler, 2005; Varón et al., 2009).

`report("result.kml", kml:(supramap, "lat-longdata.csv")` These commands perform the geographic projection of the tree, diagnose the apomorphies (e.g. mutations and host shifts implied by the tree), and create the output as a KML file. The KML file can be viewed with a variety of online mapping services, GIS, and virtual globes.

`report("result.tre", trees)` These commands export the best tree in nested parenthesis format, suitable for viewing with flat tree viewers.

`report("result.stats", treestats)` These commands export the heuristic length the best tree found and the number of times this length was hit during the search.

Geographical projection of the tree

The algorithm that generates the KML file computes three data points: altitude, longitude, and latitude for each node of the tree. The altitude of each node is calculated by multiplying its tree height by a constant (Janies et al., 2007). The root node is assigned the highest altitude, internal nodes are lower, and terminal nodes have an altitude of zero.

Leaf nodes of the tree represent observed taxa (e.g., viral isolates), whose GIS coordinates are found in the CSV file. The algorithm must compute the latitude and longitude of the internal nodes, which represent common ancestors. In a post order traversal of the tree, each node's latitude and longitude equals the calculated geographical midpoint of its two children. Once the geographical data have been computed for the tree, the algorithm generates the KML tags, includes apomorphy data in the pop-up windows of the KML, and writes a KML file that is viewable with a GIS (Fig. 1).

Output

A heuristic, minimum length tree found by the POY search is presented to the user as a text file with taxa in nested parentheses. This file can be viewed with a flat tree viewer such as TreeView (Page, 2001). The results of the tree projected onto the earth are presented to the user as a KML file suitable for viewing with a variety of software including Google Earth™ (http://earth.google.com), ArcGIS Explorer™ (http://www.esri.com/software/arcgis/explorer/index.html), NASA Worldwind (http://worldwind.arc.nasa.gov), and Google Maps™ (http://maps.google.com).

Data for the case study

We used 239 nucleotide sequences for the polymerase basic 2 genomic segment of recent isolates of H5N1 influenza. These isolates represent a sample of the diversity of lineages spreading westward from China

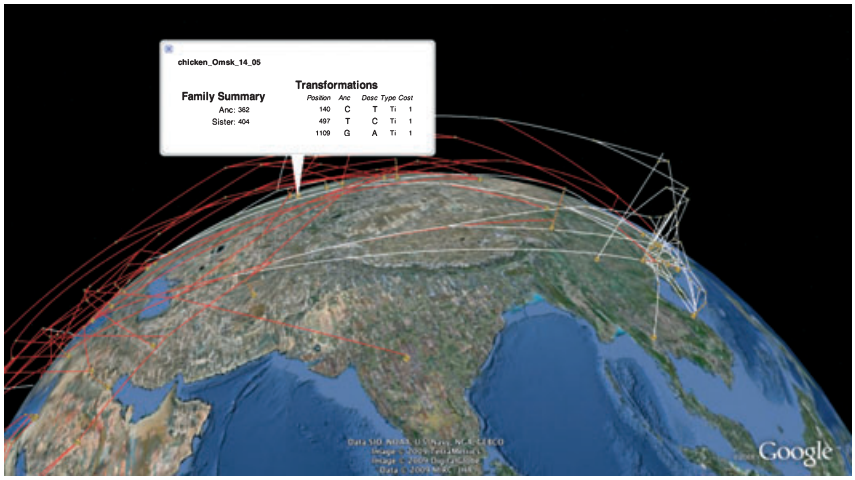

Fig. 1. Screen capture of a phylogeny of avian influenza (H5N1) projected into Google Earth™ by Supramap. This tree depicts the westward spread of H5N1. Red colour for tree branches indicate a genotype of lysine (K) at amino acid position 627 in the PB2 protein of the avian influenza virus (H5N1), which confers increase replication in mammals. White colour for tree branches indicate a genotype of glutamic acid (E), the wild type for H5N1. Colour changes along branches indicate a mutation. Mutations at each node can be viewed in pop-up windows.

across Russia, the Middle East, Africa, and Europe. We chose the strain A/goose/Yunnan/3315/2005 as an outgroup. This strain was discovered to be the most closely related isolate to the root of the westward clade in our analyses of datasets that contained much more background data, including many isolates from China, Japan, Korea and South East Asia ranging back to 1996.

For character evolution studies we chose two characters. One character was chosen to illustrate host data (i.e. avian, feline, human, mustelid, or an environmental isolate) and another character was chosen for amino acid genotypes (i.e., glutamic acid or lysine) for position 627 of polymerase basic protein 2 (PB2). When the PB2 protein contains lysine at position 627, the influenza virus is able replicate in the lower temperature (relative to avian) of the mammalian respiratory tract (Tarendeau et al., 2008).

Results

The phylogenetic analysis found a single best tree of length 1528 steps. The tree in the visualization suggests that H5N1 lineages carrying lysine-627 in PB2 are spreading westward (Fig. 1). For raw data, results, and an interactive KML file, see: http://supramap.osu.edu/sm/supramap/publications.

We used the concentrated changes test (CCT; Maddison, 1990) to provide an independent check on two hypotheses generated by the visualization and tree. The CCT is a measure of covariation between two binary characters on a phylogenetic tree. The CCT was calculated using MacClade (Maddison and Maddison, 2003). As the tree and the number of changes are large, we used a simulation of 500 000 iterations. We examined covariation in shifts from avian to mammalian hosts and mutations from glutamic acid (E) to lysine (K) at position 627 of the PB2 protein (E627K). We also examined covariation of the spread of the virus west of 100° east longitude with E627K mutations.

In the tree resulting from analysis of 239 PB2 sequences, there are eight shifts from avian to mammal hosts for H5N1 influenza and eight E627K mutations. Two of the eight avian to mammal host shifts are coincident with E627K mutations. We found a correlation (CCT = 0.0059) between the shift from avian to mammalian hosts and the E627K mutation.

The westward movement of the virus is less well correlated with E627K (CCT = 0.0289). The virus is highly mobile eastward and westward. There are 16 east-to-west transitions and two of these transitions are coincident with E627K mutations. This leads to the higher CCT value.

As we conducted two tests, we adjusted for multiple testing using a Bonferroni correction resulting in a cutoff of 0.025 for statistical significance. Based on this cutoff, avian to mammal host shifts in H5N1 are significantly correlated with E627K mutations. Although close to the cutoff, the westward spread of the virus is not significantly correlated with E627K mutations.

Discussion

These results for E627K are consistent with those of Janies et al. (2007). The correlation between avian-mammalian host shifts and the E627K mutation remains clear in phylogenetic work as well as in research on laboratory animals (Subbarao et al., 1993). The visualization suggests that H5N1 is mutating at position

627 while it is spreading westward. This is an intriguing hypothesis, but as shown in Janies et al. (2007), cannot be supported statistically. We highlight this case study because it underscores the point that visualizations should be treated with caution. Although visualization is a great tool for inspiration of hypotheses and communication of ideas, results should be checked for statistical significance. We choose to use the concentrated changes test as it explicitly takes into account the interrelationships of the organisms whereas other methods do not (Grupe et al., 2001). Felsenstein (1985) argues that due to common ancestry, organisms cannot be treated as statistically independent entities. The interrelationships between organisms can be obtained via phylogenetic analysis and should not be ignored. As data are updated frequently, our easy-to-use workflow enables researchers to periodically rerun phylogenetic analyses and retest hypotheses on the spread of zoonotic pathogens such as H5N1.

Performance

The Supramap workflow initially depended on using Extensible Stylesheet Language (XSL) transformations (XSLT) to filter and convert results from POY to KML. This information includes the tree structure and character changes associated with each branch in nested elements.

However, XSLT processing is memory intensive. Trees with large numbers of isolates and character changes can exceed the memory available and prevent the process from completing XSLT.

To address the limitations of XSLT, we have implemented plug-ins for POY. The plug-in contains functions that perform the KML generation inside POY. Thus the user can complete the entire Supramap workflow using POY as a stand-alone application. No XSLT steps are needed using plug-ins. The data are kept in memory that is allocated to the POY process and used therein for calculating and generating the KML visualization.

A stand-alone application

The use of a workflow that is entirely contained within POY has the disadvantage over the web-based application in that the user has to learn the command line syntax of POY and use their local computing resources. However, an important issue in research on emergent infectious diseases is data security. One solution to this problem is that users can run a binary of POY enabled via plug-ins for KML generation on local computers without the use of a remote cluster and without transmission of their data outside of their organization. We include instructions for: the web and stand-alone implementations of Supramap and POY, projection of precomputed trees, and colouring tree branches in KML at this web page: http://supramap.osu.edu/sm/supramap/tutorials.

Sharing of expertise, computational resources and results in kind with data

We hope that our tools will complement data sharing efforts such as NIH's GenBank (http://ncbi.nlm.nih.gov) and The Global Initiative on Sharing All Influenza Data (http://gisaid.org). In Supramap, we share powerful software tools and computing via an easy-to-use interface. Our overarching goal is to build communities of researchers and public health officials to leverage combined strengths and expertise to fight infectious diseases. Once produced and shared, KML encoded maps can be layered in a GIS. Layering pathogen phylogenies with other data such as host distributions, transit systems and environmental conditions will provide a common analytical framework in which a multidisciplinary team can work to understand the origin and spread of emergent diseases.

Acknowledgements

We acknowledge that this material is based upon work supported by, or in part by, the US Army Research Laboratory and Office under grant number W911NF-05-1-0271. This work was also Funded by grants from the Research on Research Program of the Ohio State University (OSU) and the Google.org Fund of the Tides Foundation. We acknowledge the support of the American Museum of Natural History and The Department of Biomedical Informatics of OSU. We thank the Medical Center Information Services team of OSU and the Ohio Supercomputer Center for hosting computing clusters used in this study. We thank Joe Camoriano, Ben Bays, and Earle Holland of OSU who produced the instructive videos on http://supramap.osu.edu. We thank Jon Studer for proofreading. We are grateful to Jori Hardman and Manirupa Das for recent maintenance of http://supramap.osu.edu. We thank two anonymous peer reviewers for their constructive criticisms and the editors for guiding the review process.

References

Brownstein, J., Freifeld, C., Reis, B., Mandl, K., 2008. Surveillance sans frontières: internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 5, e151.

Felsenstein, J., 1985. Phylogenies and the comparative method. Am. Nat. 125, 1–15.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., Brilliant, L., 2008. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Grupe, A., Germer, S., Usuka, J., Aud, J., Belknap, J., Klein, R., Ahluwalia, M., Higuchi, R., Peltz, G., 2001. In silico mapping of complex disease-related traits in mice. Science 292, 1915–1918.

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J., Butt, K.M., Wong, K.L., Chan, K.W., Lim, W., Shortridge, K.F., Yuen, K.Y., Peiris, J.S., Poon, L.L., 2003. Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science 302, 276–278.

Hill, A., Guralnick, R., Wilson, M., Habib, F., Janies, D., 2009. Evolution of drug resistance in multiple distinct lineages of H5N1 avian influenza. Infect. Genet. Evol. 9, 169–178.

HoDac, H., Fitch, W.M., Lathrop, R.H., Wallace, R.G., 2007. MigraPhyla: statistical analysis of migration events through a phylogeny. Version 1.0b. http://pd.bio.uci.edu/ee/WallaceR/MigraPhyla.html.

Hovmöller, R., Alexandrov, B., Hardman, J., Janies, D., 2010. Tracking the geographic spread of avian influenza (H5N1) with multiple phylogenetic trees. Cladistics 26, 1–13.

Janies, D., Hill, A., Guralnick, R., Habib, F., Waltari, E., Wheeler, W.C., 2007. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). Syst. Biol. 56, 321–329.

Janies, D., Habib, F., Alexandrov, B., Hill, A., Pol, D., 2008. Evolution of genomes, host shifts and the geographic spread of SARS-CoV and related coronaviruses. Cladistics 24, 111–130.

Kidd, D., Liu, X., 2008. GEOPHYLOBUILDER 1.0: an ARCGIS extension for creating 'geophylogenies'. Mol Ecol Resour 8, 88–91.

Ksiazek, T., Erdman, D., Goldsmith, C., Zaki, S., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J., Lim, W., Rollin, P., Dowell, S., Ling, A., Humphrey, C., Shieh, W., Guarner, J., Paddock, C., Rota, P., Fields, B., DeRisi, J., Yang, J., Cox., N., Hughes, J., LeDuc, J., Bellini, W., Anderson, L., for the SARS Working Group., 2003. A novel coronavirus associated with Severe Acute Respiratory Syndrome. N. Eng. J. Med. 348, 1953–1966.

Lemey, P., Rambaut, A., Drummond, A., Suchard, M., 2009. Bayesian phylogeography finds its roots. PLoS Comput. Biol. 5, e1000520.

Maddison, W., 1990. A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? Evolution 44, 539–557.

Maddison, D., Maddison, W., 2003. MacClade. http://www.macclade.org.

Maddison, D., Maddison, W., 2009. Cartographer module for Mesquite. http://mesquiteproject.org/packages/cartographer/.

Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., Cloutier, A., Coughlin, S., Freeman, D., Girn, N., Griffith, O., Leach, S., Mayo, M., McDonald, H., Montgomery, S., Pandoh, P., Petrescu, A., Robertson, A., Schein, J., Siddiqui, A., Smailus, D., Stott, J., Yang, G., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R., Krajden, M., Petric, M., Skowronski, D., Upton, C., Roper, R., 2003. The genome sequence of the SARS-associated coronavirus. Science 300, 1399–1404.

Page, R., 2001. TreeView. http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Parks, D.H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J., Brooks, S., Beiko, R., 2009. GenGIS: a geospatial information system for genomic data. Genome Res. 19, 1896–1904.

Piel, W., 2007. Experimental Google Earth Phylogenetic Tree Server. http://www.treebase.org/getrees/

Rifaieh, R., Unwin, R., Carver, J., Miller, M., 2007. SWAMI: integrating biological databases and analysis tools within user friendly environment. In: Cohen-Boulakia, S., Tannen, V. (Eds.), Data Integration in the Life Sciences 2007, Lecture Notes in Bioinformatics (LNBI), 4544, pp.48–58.

Rota, P.A., Oberste, M.S., Monroe, S.S., Nix, W.A., Campagnoli, R., Icenogle, J.P., Peñaranda, S., Bankamp, B., Maher, K., Chen, M., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J.L., Chen, Q., Wang, D., Erdman, D.D., Peret, T.C.T., Burns, C., Ksiazek, T.G., Rollin, P.E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Günther, S., Osterhaus, A.D.M.E., Drosten, C., Pallansch, M.A., Anderson, L.J., Bellini, W.J., 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394–1399.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–771.

Subbarao, E.K., London, W., Murphy, B., 1993. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. J. Virol. 67, 1761–1764.

Tarendeau, F., Crepin, T., Guilligay, D., Ruigrok, R., Cusack, S., Hart, D., 2008. Host determinant residue lysine 627 lies on the surface of a discrete, folded domain of influenza virus polymerase PB2 subunit. PLoS Pathog. 4, e1000136.

Varón, A., Vinh, L., Wheeler, W.C., 2009. POY version 4: phylogenetic analysis using dynamic homologies. Cladistics 26, 72–85.

Wheeler, W.C., 2005. Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. Cladistics 19, 261–268.

Wheeler, W.C., Aagesen, L., Arango, C.P., Faivonvich, J., Grant, T., D'Haese, C., Janies, D.A., Smith, W.L., Varón, A., Giribet, G., 2006. Dynamic Homology and Phylogenetic Systematics: a Unified Approach Using POY. American Museum of Natural History, New York.